References

[1] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Advances in Mathematics; Supplementary Volume*, 1:167–212, 1978.

[2] P. R. Stein and M. S. Waterman. On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Mathematics*, 26:261–272, 1978.

[3] W. R. Schmidtt and M. S. Waterman. Linear trees and RNA secondary structure. To appear, *Discrete Applied Mathematics.*

[4] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):491–621, 1984.

[5] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics.* Addison-Wesley, 1989.

and, in particular,

$$\log t_1^{1,2}(n) \sim 2n \log \phi \qquad (n \to \infty).$$

Theorem 5 also implies that a typical depth-1, loop length ≥ 2 structure has order

$$\frac{n}{2}\left(1 - \frac{1}{\sqrt{1 + 4\sqrt{\gamma}}}\right)(1 + o(1)).$$

In particular, for $\gamma = 1$ a typical $(d = 1; l \geq 2)$ structure has approximately $0.276n$ matching base pairs, while for uniformly distributed RNA sequences ($\gamma = 1/4$) a typical $(d = 1; l \geq 2)$ structure has approximately $0.211n$ matching base pairs.

Case 3: $l \geq 3$ — structures with matching pairs separated by at least two elements.

In this case (11) gives

$$\begin{aligned} P(z,w)^r &= \sum_h \binom{-r}{h}(-1)^h z^h \left(1 - \gamma w z \frac{(1 - z^{l-1})}{(1 - z)}\right)^h \\ &= \sum_{h,i} \binom{r + h - 1}{h}\binom{h}{i}(-1)^i \gamma^i w^i z^{h+i} z^{h+i} \frac{(1 - z^{l-1})^i}{(1 - z)^i} \\ &= \sum_{h,i,j,k} (-1)^{i+k} \binom{r + h - 1}{h}\binom{h}{i}\binom{i + j - 1}{j}\binom{i}{k} \gamma^i w^i z^{h+i+j+k(l-1)} . \end{aligned}$$

Thus,

$$H(z,w) = \sum_{g,h,i,j,k} (-1)^{i+k} C_g \binom{2g + h}{h}\binom{h}{i}\binom{i}{k}\binom{i + j - 1}{j} \gamma^{g+i} w^{g+i} z^{2g+h+i+j+k(l-1)} .$$

Substituting $n \leftarrow 2g + h + i + j + k(l - 1)$, $m \leftarrow g + i$ in turn, obtain

$$\begin{aligned} t_\gamma^{1,l}(n,m) &= \gamma^m \sum_{g,h,k} (-1)^{m-g+k} \frac{1}{g+1} \binom{2g + h}{k, m - g - k, h - m + g, g, g} \\ &\quad \times \binom{n - 2g - h - kl - 1}{m - g - 1} . \end{aligned}$$

It does not seem likely that the expression above can be further simplified to obtain a simple closed form for general values of l; for such cases we have to be satisfied with the compact form (12) for the generating function.

Acknowledgements

Mick Noordewier introduced us to the problem in a seminar at AT&T Bell Laboratories and Martin Farach was helpful in pointing out some of the biological issues involved. Thanks to Shao Fang for showing us how the method of rational certification of binomial identities works, and Mike Waterman for acquainting us with the literature and sending us a preprint of joint work with W. R. Schmidtt. The second named author also thanks Aaron Wyner for facilitating a visit to Bell Labs during which this problem was investigated.

Theorem 5 *The expected number $t_\gamma^{1,2}(n,m)$ of depth-1, order-m, loop length $l \geq 2$ structures on $\{1,\ldots,n\}$ that are valid for a random sequence with match probability γ is maximized when m attains a value*

$$m^* = \frac{1}{2}\left(1 - \frac{1}{\sqrt{1+4\sqrt{\gamma}}}\right) n + O(1),$$

and its value then is

$$t_\gamma^{1,2}(n,m^*) = \frac{(1+4\sqrt{\gamma})}{\pi\left(1+2\sqrt{\gamma}-\sqrt{1+4\sqrt{\gamma}}\right)}\,\frac{1}{n^2}\left(\frac{1+\sqrt{1+4\sqrt{\gamma}}}{2}\right)^{2n}\left(1+O\Big(\frac{1}{n}\Big)\right).$$

An interpolative approach similar to that for the case $l = 1$ can be used to determine the $O(1)$ term in m^*. This will require determining the roots of a fourth-degree polynomial.

When $\gamma = 1$ we have the interesting result

$$t_1^{1,2}(n,m^*) = \frac{5}{\pi\left(3-\sqrt{5}\right)}\,\frac{\phi^{2n}}{n^2}\left(1+O\Big(\frac{1}{n}\Big)\right) = \Theta\left(\frac{\phi^{2n}}{n^2}\right),$$

where $\phi = (1+\sqrt{5})/2 \approx 1.618$ is the golden ratio.

As for the depth-1, arbitrary loop length $l \geq 1$ case, we can represent $t_\gamma^{1,2}(n) = \sum_{m\geq 0} t_\gamma^{1,2}(n,m)$ as a simple hypergeometric. Using Assertion 5, write

$$t_\gamma^{1,2}(n) = 1 + \sum_{m\geq 1}\frac{1}{m}\binom{n-m}{m+1}\binom{n-m-1}{m-1}\gamma^m \overset{\text{def}}{=} \sum_{m\geq 0} q_m.$$

The term ratios

$$\frac{q_{m+1}}{q_m} = \frac{(m-\frac{1}{2}n+1)(m-\frac{1}{2}n+\frac{1}{2})(m-\frac{1}{2}n+\frac{1}{2})(m-\frac{1}{2}n)}{(m+2)(m-n+1)(m-n)}\,\frac{16\gamma}{(m+1)}$$

are rational functions of m and $q_0 = 1$. Comparing with the standard form (14), we have hence shown the following identity.

Assertion 7 *For any match probability γ and sequence length n,*

$$t_\gamma^{1,2}(n) = {}_4\mathrm{F}_3\left(\begin{matrix} -\frac{1}{2}n+1,\ -\frac{1}{2}n+\frac{1}{2},\ -\frac{1}{2}n+\frac{1}{2},\ -\frac{1}{2}n \\ 2,\ -n+1,\ -n\end{matrix}\,\middle|\,16\gamma\right),$$

While simple closed forms do not exist for $t_\gamma^{1,2}(n)$ for general values of γ, an application of Theorem 5 shows that

$$\log t_\gamma^{1,2}(n) \sim 2n\,\log\left(\frac{1+\sqrt{1+4\sqrt{\gamma}}}{2}\right) \qquad (n\to\infty),$$

so that the term ratios are rational functions of h. Comparing with the standard form (14), it follows that

$$R = r_0\,{}_2\mathrm{F}_1\!\left(\begin{matrix} n-m+1,\ -m \\ 2 \end{matrix} \,\middle|\, 1 \right).$$

Using Lemma 1 and the expression for r_0, we finally obtain

$$\begin{aligned} R &= (-1)^m \binom{n-m}{n-2m} \frac{(-n+2m)^{\underline{m}}}{(m+1)^{\underline{m}}}\,\gamma^m \\ &= \frac{(n-m)!}{(n-2m)!m!}\,\frac{(n-2m)^{\overline{m}}}{(m+1)!}\,\gamma^m \\ &= \frac{(n-m)!}{m!(n-2m)!}\,\frac{(n-m-1)!}{(m+1)!(n-2m-1)!}\,\gamma^m. \end{aligned}$$

Simple algebraic manipulations complete the proof. ▮

Schmidtt and Waterman [3] have also recently demonstrated Assertion 6 (for $\gamma = 1$). Their approach, in sharp contradistinction to the straightforward combinatorial attack on the generating function espoused here, involves an ingenious, non-intuitive transformation of the problem to a combinatorial problem on linear trees for which there is a known solution.

For fixed n, we can now find the value of m for which $t_\gamma^{1,2}(n,m)$ is maximized. As for the depth-1, arbitrary loop length $l \geq 1$ case, the maximum will occur when m is proportional to n. Accordingly, suppose $m = an$ where $a = \Theta(1)$. Stirling's approximation applied to (18) then gives

$$\begin{aligned} t_\gamma^{1,2}(n,m) &= \frac{1}{2\pi a^2 n^2}\,2^{2n\left[h(a,a,1-2a)-h(a)+a\,\log\sqrt{\gamma}\right]}\left(1+O\left(\frac{1}{n}\right)\right) \\ &= \frac{1}{2\pi a^2 n^2}\,2^{2n\left(h(2a)-h(a)+a(2+\log\sqrt{\gamma})\right)}\left(1+O\left(\frac{1}{n}\right)\right). \end{aligned}$$

By differentiation, the exponent

$$\theta(a) = 2n(h(2a) - h(a) + a(2 + \log\sqrt{\gamma})),$$

achieves its maximum when a attains a unique value a^* in the range $[0, 0.2764)$ given by

$$a^* = \frac{1}{2}\left(1 - \frac{1}{\sqrt{1+4\sqrt{\gamma}}}\right).$$

Simple algebra shows then that the exponent maximum is given by

$$\theta(a^*) = 2n\,\log\frac{(1-a^*)}{(1-2a^*)} = 2n\Big(-1 + \log(1+\sqrt{1+4\sqrt{\gamma}})\Big).$$

This argument has neglected the nonexponential part of $t_\gamma^{1,2}(n,m)$. As before, this can affect the location of the maximum by at most a constant term. Thus, we have shown:

and

$$\frac{1}{2y}\left(1-\sqrt{1-4y}\right)=\sum_{h\geq 0}\frac{1}{h+1}\binom{2h}{h}y^h,$$

we now have

$$\begin{aligned}B(x) &= \frac{1}{(1+x)}\sum_{h\geq 0}\binom{2h}{h}\frac{1}{h+1}\left[\frac{x}{(1+x)^2}\right]^h\\ &= \frac{1}{(1+x)}\frac{(1+x)^2}{2x}\left(1-\sqrt{1-\frac{4x}{(1+x)^2}}\right)\\ &= \frac{(1+x)}{2x}\left(1-\frac{1-x}{1+x}\right)\\ &= 1\ .\end{aligned}$$

The claim is proved. ▮

To complete verification of the boundary conditions, consider

$$\begin{aligned}t_\gamma^{1,2}(n,0) &= \sum_{h=0}^{m}(-1)^{m-h}\binom{n-m+h}{h,h,m-h,n-2m}\frac{\gamma^m}{h+1}\bigg|_{m=0}\\ &= \binom{n}{n} = \begin{cases}1 & \text{for } n\geq 0,\\ 0 & \text{for } n<0,\end{cases}\end{aligned}$$

as was to be verified.

Finally, consider $n\geq 1$, $m\geq 1$. Equation (17) admits of further simplification.

Assertion 6 *For any match probability γ,*

$$t_\gamma^{1,2}(n,m)=\frac{1}{m}\binom{n-m}{m+1}\binom{n-m-1}{m-1}\gamma^m \qquad (n\geq 1;\ m\geq 1)\ . \tag{18}$$

PROOF: Write $R=t_\gamma^{1,2}(n,m)$ for simplicity. Using (17), set

$$R=\sum_{h\geq 0}(-1)^{m-h}\binom{n-m+h}{h,h,m-h,n-2m}\frac{1}{h+1}\overset{\text{def}}{=}\sum_{h\geq 0}r_h.$$

Assume, without loss of generality, that $1\leq m\leq n/2$. Then

$$r_0=(-1)^m\binom{n-m}{n-2m}\gamma^m$$

is nonzero; further, for $h\geq 0$,

$$\frac{r_{h+1}}{r_h}=\frac{(h+n-m+1)(h-m)}{(h+2)}\frac{1}{(h+1)},$$

where the second equation obtains from the successive replacements $m \leftarrow h+j$ and $n \leftarrow m+h+i$. Using the simple identity

$$\binom{-2h-1}{n-m-h} = (-1)^{-(n-m-h)}\binom{n-m+h}{n-m-h},$$

we can now compare (10) and (16) to obtain

$$\begin{aligned} t_\gamma^{1,2}(n,m) &= \sum_{h\geq 0}(-1)^{m-h}\binom{n-m+h}{n-m-h}\binom{n-m-h}{m-h}\binom{2h}{h}\frac{\gamma^m}{h+1} \\ &= \sum_{h\geq 0}(-1)^{m-h}\binom{n-m+h}{h,h,m-h,n-2m}\frac{\gamma^m}{h+1}. \end{aligned} \tag{17}$$

The multinomial coefficients in the summand will be nonzero only when the following conditions hold simultaneously: $0 \leq h \leq m \leq n/2$. In particular, $t_\gamma^{1,2}(n,m) = 0$ when $m < 0$ or $n < 2m$, as required by the boundary conditions. The case $n = 2m$ is of interest as the boundary conditions require

$$t_\gamma^{1,2}(2m,m) = \begin{cases} 1 & \text{if } m = 0, \\ 0 & \text{if } m \neq 0. \end{cases}$$

For notational expedience, define

$$\beta_m = t_\gamma^{1,2}(2m,m) = \gamma^m \sum_{h\geq 0}(-1)^{m-h}\binom{m+h}{h,h,m-h}\frac{1}{h+1}.$$

The following lemma shows that the boundary conditions are indeed satisfied.

Lemma 2

$$\sum_{h\geq 0}(-1)^{m-h}\binom{m+h}{h,h,m-h}\frac{1}{h+1} = \begin{cases} 1 & \textit{for } m = 0, \\ 0 & \textit{for } m \neq 0. \end{cases}$$

PROOF: The case $m < 0$ is obvious as $m - h < 0$ $(h \geq 0)$. Suffices hence to show $\beta_m = 0$ $(m > 0)$. Consider the generating function $B(x) = \sum_m \beta_m x^m$. Substituting for β_m, we have

$$\begin{aligned} B(x) &= \sum_{h\geq 0}(-1)^h\binom{2h}{h}\frac{1}{h+1}\sum_{m\geq 0}(-1)^m\binom{m+h}{2h}x^m \\ &= \sum_{h\geq 0}\binom{2h}{h}\frac{x^{-h}}{h+1}\sum_{\tau\geq 0}\binom{\tau}{2h}(-x)^\tau, \end{aligned}$$

the first equation following by replacing the multinomial by a product of binomials, and the second equation resulting from the substitution $\tau \leftarrow m+h$. Recalling the simple identities

$$\frac{y^t}{(1-y)^{t+1}} = \sum_{\tau\geq 0}\binom{\tau}{t}y^\tau \qquad (\text{integer } t \geq 0),$$

in accordance with the results of the direct argument leading to Assertion 1. Now

$$t_\gamma^{1,1}(n) = \sum_{m\geq 0} t_\gamma^{1,1}(n,m) = \sum_{m\geq 0} \binom{n}{m,m,n-2m}\frac{\gamma^m}{m+1}$$

is the total number of depth-1 structures on $\{1,\ldots,n\}$. (See Assertion 3 for an alternative form.) Writing

$$t_\gamma^{1,1}(n) \stackrel{\text{def}}{=} \sum_{m\geq 0} s_m$$

it is easy to see that the sum is a hypergeometric series. Indeed, $s_0 = 1$, and for $m \geq 0$ the term ratios

$$\frac{s_{m+1}}{s_m} = \frac{(m-\frac{1}{2}n+\frac{1}{2})\,(m-\frac{1}{2}n)}{(m+2)}\,\frac{4\gamma}{(m+1)}$$

are rational functions of m. The following assertion now follows by comparison with the standard form (14) for the term ratios.

Assertion 5 *For any match probability γ and sequence length n,*

$$t_\gamma^{1,1}(n) = {}_2\mathrm{F}_1\left(\begin{matrix} -\frac{1}{2}n+\frac{1}{2},\ -\frac{1}{2}n \\ 2 \end{matrix}\,\middle|\, 4\gamma\right).$$

Note that for $n > 1$—the cases $n = 0$ and $n = 1$ are trivial—either $-n/2$ or $-(n-1)/2$ is a negative integer. Thus, when $\gamma = 1/4$, Lemma 1 applies and the above hypergeometric series yields the simple closed form of Assertion 2. For general values of γ, however, the hypergeometric does not yield a simple closed form.

Case 2: $l = 2$ — structures where adjacent elements cannot be matched, i.e., any nested set of matched parentheses must have at least one unmatched node in the center:

$$(((\cdot\cdot)))((\cdot)) \ \text{ etc.}$$

From (11) we have

$$\begin{aligned}
P(z,w)^r &= (1-z+\gamma w z^2)^{-r} \\
&= \sum_i \binom{-r}{i}(-1)^i z^i (1-\gamma w z)^i \\
&= \sum_i \sum_j (-1)^{i+j}\binom{-r}{i}\binom{i}{j}\gamma^j z^{i+j} w^j\ .
\end{aligned}$$

Substituting in (15) gives

$$\begin{aligned}
H(z,w) &= \sum_h \sum_i \sum_j (-1)^{i+j} C_h \binom{-2h-1}{i}\binom{i}{j}\gamma^{h+j} z^{2h+i+j} w^{h+j} \\
&= \sum_n \sum_m \sum_h (-1)^{n-2h} C_h \binom{-2h-1}{n-m-h}\binom{n-m-h}{m-h}\gamma^n z^n w^m\ , \qquad (16)
\end{aligned}$$

i.e., the term ratios f_{h+1}/f_h are rational functions of h. Comparing (13) with the standard form (14), we hence obtain

$$S = s_0\ {}_2\mathrm{F}_1\left(\begin{matrix} a,\ -m \\ c \end{matrix} \middle| 1 \right) = \binom{c+m-1}{m}\ {}_2\mathrm{F}_1\left(\begin{matrix} a,\ -m \\ c \end{matrix} \middle| 1 \right).$$

But, by Vandermonde's convolution,

$$S = \binom{c+m-a-1}{m}.$$

Equating the two forms for S gives the desired result. ∎

Extend the definition of Catalan numbers, given in (1) to all integers h by

$$C_h = \begin{cases} 0 & \text{if } h < 0, \\ \frac{1}{h+1}\binom{2h}{h} & \text{if } h \geq 0. \end{cases}$$

For general x, we then have the familiar identity

$$\frac{1}{2}\left(1-\sqrt{1-4x}\right) = \frac{1}{2}\left[1 - \sum_{h\geq 0}\binom{1/2}{h}(-1)^h 2^{2h} x^h\right] = \sum_h C_h x^{h+1}\ ,$$

where we've again invoked (8). Substituting in (12) we obtain

$$H(z,w) = \sum_h C_h \gamma^h w^h z^{2h} P(z,w)^{2h+1}\ . \tag{15}$$

Case 1: $l = 1$ — structures with no loop constraints.

In (11), $P(z,w) = 1/(1-z)$. It follows therefore that

$$P(z,w)^r = (1-z)^{-r} = \sum_i \binom{-r}{i}(-z)^i = \sum_i \binom{r+i-1}{i} z^i\ .$$

Substituting in (15), we now have

$$H(z,w) = \sum_h \sum_i C_h \binom{2h+i}{i} \gamma^h z^{2h+i} w^h = \sum_n \sum_m C_m \binom{n}{n-2m} \gamma^m z^n w^m\ ,$$

where the second equation follows with the substitutions $n \leftarrow 2h+i$ and $m \leftarrow h$. Comparing term by term with (10) we have

$$t_\gamma^{1,1}(n,m) = \binom{n}{m,m,n-2m} \frac{\gamma^m}{m+1} \qquad (n \geq 0;\ m \geq 0),$$

The boundary conditions specify which root of the quadratic is permissible. We hence have

$$H(z,w) = \frac{1}{2\gamma w z^2 P(z,w)} \left[1 - \sqrt{1 - 4\gamma w z^2 P(z,w)^2}\right] . \tag{12}$$

The above expression for the generating function H is nice and compact, and for general values of l this may be all we can hope for. For special values of l, however, we can go further. Before we proceed, let us detour through a hypergeometric identity.

Lemma 1 *Let a be any real number, m a positive integer, and c any real number which is not zero or a negative integer. Then*

$${}_2\mathrm{F}_1\left(\begin{array}{c} a,\ -m \\ c \end{array} \middle| 1 \right) = \frac{(c+m-a-1)^{\underline{m}}}{(c+m-1)^{\underline{m}}}.$$

PROOF: Let us begin with the familiar convolution identity ("Vandermonde's convolution")

$$\sum_{h \ge 0} \binom{r}{h} \binom{s}{m-h} = \binom{r+s}{m}$$

valid for any real r and s and integer m. (It is easy to verify equality above by multiplying both sides by x^m and summing over all m to obtain the same generating function $(1+x)^{r+s}$ for both sides.) Anticipating later algebraic simplification, let $a = -r$ and $c = s - m + 1$, and write the sum on the left-hand side as

$$S = \sum_{h \ge 0} \binom{-a}{h} \binom{c+m-1}{m-h} \overset{\text{def}}{=} \sum_{h \ge 0} s_h.$$

We then have

$$s_0 = \binom{c+m-1}{m},$$

and for $h \ge 0$, we have the term ratios

$$\frac{s_{h+1}}{s_h} = \frac{\binom{-a}{h+1}\binom{c+m-1}{m-h-1}}{\binom{-a}{h}\binom{c+m-1}{m-h}} = \frac{(h+a)(h-m)}{(h+c)} \frac{1}{(h+1)}. \tag{13}$$

The term ratios s_{h+1}/s_h are hence rational functions of h; it follows that S can be expressed in terms of a hypergeometric series (cf. Graham, Knuth, and Patashnik [5], for instance). It is easy to verify that for a general hypergeometric series

$${}_m\mathrm{F}_n\left(\begin{array}{c} a_1,\ \dots,\ a_m \\ b_1,\ \dots,\ b_n \end{array} \middle| z \right) = \sum_{h \ge 0} \frac{a_1^{\overline{h}} \dots a_m^{\overline{h}}}{b_1^{\overline{h}} \dots b_n^{\overline{h}}} \frac{z^h}{h!} \overset{\text{def}}{=} \sum_{h \ge 0} f_h,$$

we have $f_0 = 1$, while the term ratios f_{h+1}/f_h satisfy

$$\frac{f_{h+1}}{f_h} = \frac{(h+a_1)\dots(h+a_m)}{(h+b_1)\dots(h+b_n)} \frac{z}{(h+1)}, \tag{14}$$

The boundary conditions allow us to extend the summation over j to range from $-\infty$ to $+\infty$ as one or the other term in the product is identically zero outside the range indicated above. Similarly, the upper limit on the sum over i can be extended to $+\infty$. Therefore,

$$\begin{aligned} t_\gamma^{1,l}(n,m) &= t_\gamma^{1,l}(n-1,m) + \gamma \sum_{i\geq l}\sum_j t_\gamma^{1,l}(i-1,j)t_\gamma^{1,l}(n-1-i,m-1-j) \\ &= t_\gamma^{1,l}(n-1,m) + \gamma \sum_{i\geq l-1}\sum_j t_\gamma^{1,l}(i,j)t_\gamma^{1,l}(n-2-i,m-1-j) \\ &= t_\gamma^{1,l}(n-1,m) + \gamma \sum_{i}\sum_j t_\gamma^{1,l}(i,j)t_\gamma^{1,l}(n-2-i,m-1-j) \\ &\qquad -\gamma \sum_{i\leq l-2}\sum_j t_\gamma^{1,l}(i,j)t_\gamma^{1,l}(n-2-i,m-1-j)\ . \end{aligned}$$

An examination of the boundary conditions shows that this recurrence applies whenever $-\infty < m < \infty$ and $n \neq 0$. A final application of the boundary conditions hence yields

$$\begin{aligned} t_\gamma^{1,l}(n,m) &= t_\gamma^{1,l}(n-1,m) + \gamma \sum_{i}\sum_j t_\gamma^{1,l}(i,j)t_\gamma^{1,l}(n-2-i,m-1-j) \\ &\qquad -\gamma \sum_{i=0}^{l-2} t_\gamma^{1,l}(n-2-i,m-1) \qquad (n\neq 0;\ -\infty < m < \infty). \end{aligned}$$

Define the ordinary bivariate generating function:

$$H(z,w) \overset{\text{def}}{=} \sum_n\sum_m t_\gamma^{1,l}(n,m)z^n w^m\ . \tag{10}$$

Multiplying both sides of the above recurrence by $z^n w^m$ and summing over the allowed ranges of n and m yields

$$H(z,w) - \sum_m t_\gamma^{1,l}(0,m)w^m = zH(z,w) + \gamma z^2 w H(z,w)^2 - \gamma w z^2 H(z,w)\sum_{i=0}^{l-2} z^i\ .$$

Thus H is obtained as a root of the quadratic

$$\gamma w z^2 H(z,w)^2 - \left(1 - z + \gamma w z^2 \sum_{i=0}^{l-2} z^i\right) H(z,w) + 1 = 0\ .$$

Define

$$P(z,w) = \frac{1}{1 - z + \gamma w z^2 \sum_{i=0}^{l-2} z^i} = \frac{1-z}{1 - 2z + (1+\gamma w)z^2 - \gamma w z^{l+1}}\ . \tag{11}$$

Then

$$\gamma w z^2 H(z,w)^2 - \frac{H(z,w)}{P(z,w)} + 1 = 0\ .$$

$$\times \sum_{\substack{m_1,m_1',\ldots,m_d,m_d' \\ \sum(m_i+m_i')\le n}} \binom{n+d}{m_1,\ m_1',\ldots,m_d,\ m_d',\ n+d-\sum_{i=1}^{d}(m_i+m_i')} \gamma^{\frac{1}{2}\sum_{i=1}^{d}(m_i+m_i')}$$

$$= \frac{\left(1+2d\sqrt{\gamma}\right)^{n+d}}{\gamma^{d/2}\prod_{i=1}^{d}(n+i)} .$$

where the last equality follows from the multinomial theorem.

We have therefore shown:

Theorem 4 *For any match probability $\gamma \in (0,1]$, and sequence length n, the total number of structures of depth $\le d$ is bounded by*

$$\frac{\left(1+2d\sqrt{\gamma}\right)^{n+2d+1/2}}{(2\pi\gamma)^d n^{2d}}\left(1+O\left(\frac{1}{n}\right)\right) \le t_\gamma^{d,1}(n) \le \frac{\left(1+2d\sqrt{\gamma}\right)^{n+d}}{\gamma^{d/2}\prod_{i=1}^{d}(n+i)} .$$

In particular, $\log t_\gamma^{d,1}(n) \sim n\,\log(1+2d\sqrt{\gamma})$ *as* $n\to\infty$.

Thus, the typical depth $\le d$, loop length $\ge l$ structure has

$$\frac{n}{d\left(2+\frac{1}{d\sqrt{\gamma}}\right)}(1+o(1))$$

matching base pairs in each of d colors. In particular, when $d=2$ and $\gamma=1$ the typical structure has approximately $n/5$ matching base pairs in each of two colors; for uniformly distributed RNA sequences ($\gamma=1/4$) with $d=2$ a typical structure has approximately $n/6$ matching base pairs in each of two colors.

5 Structures of depth-1, order-m, and loop length $\ge l$

We now turn our attention to $t_\gamma^{1,l}(n,m)$, the expected number of depth-1, order-m structures on $\{1,\ldots,n\}$ with loop length $\ge l$. A generatingfunctionological[1] approach is indicated.

We begin by setting up the boundary conditions for a recurrence. We have

$$t_\gamma^{1,l}(n,m) = \begin{cases} 0 & \text{for } (m<0) \text{ or } (m=0 \text{ and } n<0) \text{ or } (m>0 \text{ and } n<2m+l-1), \\ 1 & \text{for } (m=0 \text{ and } n\ge 0). \end{cases}$$

Arguments similar to those before show that for all $m\ge 1$ and $n\ge 2m+l-1$ we have the recurrence:

$$\begin{aligned} t_\gamma^{1,l}(n,m) &= t_\gamma^{1,l}(n-1,m) \\ &\quad + \gamma\sum_{i=l}^{n-1}\ \sum_{j=\max(0,m-1-\lfloor(n-i-l)/2\rfloor)}^{\min(m-1,\lfloor(i-l)/2\rfloor)} t_\gamma^{1,l}(i-1,j)\,t_\gamma^{1,l}(n-1-i,m-1-j) . \end{aligned}$$

---

[1] Coined by H. Wilf.

For any given large m, this expression is maximized when each m_i is $\frac{m}{d}$, i.e., when $a_i = \frac{a}{d}$. For large n, these values can be well approximated by actual m_i's:

$$t_\gamma^{d,1}(n,\mathbf{m}^*) = \frac{d^{2d}}{(2\pi)^d (an)^{2d}\sqrt{1-2a}}\, 2^{n\left(h(2a)+2a\left(1+\log\sqrt{\gamma}+\log d\right)\right)}\left(1+O\Big(\frac{1}{n}\Big)\right) .$$

According to (3), the exponent on the right-hand side is maximized when $a = m/n$ takes the value

$$a^* = \frac{1}{2+\frac{1}{d\sqrt{\gamma}}}.$$

As argued earlier, the nonexponential part can affect the estimate of $\mathbf{m}^*$ only by a constant, we hence have:

Theorem 3 *For any match probability $\gamma \in (0,1]$, fixed depth d, and fixed sequence length n, $t_\gamma^{d,1}(n,\mathbf{m})$ is maximized when $\mathbf{m}$ attains the value*

$$\mathbf{m}^* = \left(\frac{n}{d\left(2+\frac{1}{d\sqrt{\gamma}}\right)} + O(1)\right)(1,\ldots,1) ,$$

and its value then is

$$t_\gamma^{d,1}(n,\mathbf{m}^*) = \frac{1}{(2\pi\gamma)^d n^{2d}}\left(1+2d\sqrt{\gamma}\right)^{n+2d+1/2}\left(1+O\Big(\frac{1}{n}\Big)\right) .$$

This estimate can be used to bound $t_\gamma^{d,1}(n)$, the expected number of depth-d structures on $\{1,\ldots,n\}$ valid for a sequence chosen according to a probability distribution with match probability γ. On the one hand,

$$t_\gamma^{d,1}(n) \ge t_\gamma^{d,1}(n,\mathbf{m}^*) = \frac{1}{(2\pi\gamma)^d n^{2d}}\left(1+2d\sqrt{\gamma}\right)^{n+2d+1/2}\left(1+O\Big(\frac{1}{n}\Big)\right) .$$

On the other hand,

$$\begin{aligned}
t_\gamma^{d,1}(n) &= \sum_{\substack{m_1,\ldots,m_d\\ \sum m_i \le n/2}} \left(\prod_{i=1}^{d}\frac{1}{m_i+1}\right)\binom{n}{m_1,\ m_1,\ldots,m_d,\ m_d,\ n-2\sum_{i=1}^d m_i}\gamma^{\sum_{i=1}^d m_i}\\
&= \frac{1}{\prod_{i=1}^d (n+i)}\\
&\quad\times \sum_{\substack{m_1,\ldots,m_d\\ \sum m_i \le n/2}} \binom{n+d}{m_1,\ m_1+1,\ldots,m_d,\ m_d+1,\ n+d-2\sum_{i=1}^d m_i}\gamma^{\sum_{i=1}^d m_i}\\
&\le \frac{1}{\gamma^{d/2}\prod_{i=1}^d (n+i)}
\end{aligned}$$

From Theorems 1 and 2 we can now argue that a typical depth-1, arbitrary loop length $l \geq 1$ structure has order

$$\frac{n}{2+\frac{1}{\sqrt{\gamma}}}(1+o(1)).$$

In particular, if $\gamma = 1$ the typical depth-1, $l \geq 1$ structure has roughly $n/3$ matching base pairs, while for uniformly distributed RNA sequences ($\gamma = 1/4$) the typical depth-1, $l \geq 1$ structure has roughly $n/4$ matching base pairs.

4 Structures of depth $\leq d$

For structures of depth d larger than 1, we can imagine that each structure pair is colored with one of d given colors so that no two pairs of the same color intersect. Let $\mathbf{m} \stackrel{\text{def}}{=} (m_1, \ldots, m_d)$ and define $t_\gamma^{d,1}(n, \mathbf{m})$ to be the expected number of depth $\leq d$, arbitrary loop length $l \geq 1$ structures on $\{1, \ldots, n\}$ that are (1) valid for a random sequence selected according to a probability distribution with match-probability γ, and (2) for $1 \leq i \leq d$ have m_i (nonintersecting) pairs colored with the ith color. Throughout this section we assume that d is fixed. Define

$$m \stackrel{\text{def}}{=} \sum_{i=1}^{d} m_i$$

to be the total number of matches, i.e., the order of the structure. Arguments similar to those used in Section 2 yield

$$t_\gamma^{d,1}(n, \mathbf{m}) = \binom{n}{2m_1, \ldots, 2m_d,\, n-2m} \left(\prod_{i=1}^{d} \frac{1}{m_i+1} \binom{2m_i}{m_i} \right) \gamma^m .$$

We hence have the following

Assertion 4 *Let* $\mathbf{m} = (m_1, \ldots, m_d)$ *where the* m_i *are integers summing to* m. *Then*

$$t_\gamma^{d,1}(n, \mathbf{m}) = \left(\prod_{i=1}^{d} \frac{1}{m_i+1} \right) \binom{n}{m_1, m_1, \ldots, m_d, m_d, n-2m} \gamma^m .$$

To maximize $t_\gamma^{d,1}(n, \mathbf{m})$ and determine the value $\mathbf{m}^*$ of $\mathbf{m}$ that attains it, we first approximate $t_\gamma^{d,1}(n, \mathbf{m})$ for m_i's which are proportional to m. In consonance with earlier usage, for each i set $m_i = a_i n$, and let $m = an$. (Of course, $a = \sum_{i=1}^{d} a_i$.) Here, as before, $a = \Theta(1)$ is bounded above (away from 1/2) and below (away from 0). Using Stirling's approximation formula we get

$$\begin{aligned} t_\gamma^{d,1}(n, \mathbf{m}) &= \frac{1}{(2\pi)^d \sqrt{1-2a}\, n^2 \prod_{i=1}^{d} a_i^2}\, 2^{n(h(a_1, a_1, \ldots, a_d, a_d, 1-2a) + a \log \gamma)} \left(1 + O\left(\frac{1}{n}\right)\right) \\ &= \frac{1}{(2\pi)^d \sqrt{1-2a}\, n^2 \prod_{i=1}^{d} a_i^2}\, 2^{n\left(h(2a) + 2a\left(1 + \log\sqrt{\gamma} + h(\frac{a_1}{a}, \ldots, \frac{a_d}{a})\right)\right)} \left(1 + O\left(\frac{1}{n}\right)\right) . \end{aligned}$$

Then,

$$\Big|\sum\nolimits^{(1)}\Big| \;\leq\; n|c_{n+2-\log n}| = O\left(n\frac{(1+2\sqrt{\gamma})^n|\Lambda|^{-\log n}}{n^{3/2}(\log n)^{3/2}}\right) = O\left(\frac{(1+2\sqrt{\gamma})^n}{n^{3/2}(\log n)^{3/2}}\right) ,$$

and

$$\begin{aligned}
\sum\nolimits^{(2)} &= \sum_{i=0}^{\log n} c_{n+2-i} \;=\; \sum_{i=0}^{\log n}\frac{(-1)^{n-1}}{2\gamma}\binom{1/2}{n+2-i}\binom{1/2}{i}(1+2\sqrt{\gamma})^{n+2}\Lambda^{-i}\\
&= \frac{(1+2\sqrt{\gamma})^{n+2}}{4\gamma\pi^{1/2}n^{3/2}}\left(1+O\left(\frac{\log n}{n}\right)\right)\sum_{i=0}^{\log n}\binom{1/2}{i}(-\Lambda)^{-i}\\
&= \frac{(1+2\sqrt{\gamma})^{n+2}}{4\gamma\pi^{1/2}n^{3/2}}\left(1+O\left(\frac{\log n}{n}\right)\right)\\
&\qquad\left(1-\frac{1}{\Lambda}\right)^{1/2}\left(1-\left(1-\frac{1}{\Lambda}\right)^{-1/2}\sum_{i>\log n}\binom{1/2}{i}(-\Lambda)^{-i}\right)\\
&= \frac{(1+2\sqrt{\gamma})^{n+2}}{4\gamma\pi^{1/2}n^{3/2}}\left(1-\frac{1}{\Lambda}\right)^{1/2}\left(1+O\left(\frac{\log n}{n}\right)\right)\left(1+O\left(\frac{1}{(\log n)^{3/2}}\right)\right) .
\end{aligned}$$

Also

$$\left(1-\frac{1}{\Lambda}\right)^{1/2} = \left(1-\frac{1-2\sqrt{\gamma}}{1+2\sqrt{\gamma}}\right)^{1/2} = \left(\frac{4\sqrt{\gamma}}{1+2\sqrt{\gamma}}\right)^{1/2} = \frac{2\gamma^{1/4}}{(1+2\sqrt{\gamma})^{1/2}} .$$

Thus,

$$t_\gamma(n) = \frac{(1+2\sqrt{\gamma})^{n+3/2}}{2\gamma^{3/4}\pi^{1/2}n^{3/2}}\left(1+O\left(\frac{1}{(\log n)^{3/2}}\right)\right) .$$

We have therefore shown:

Theorem 2 *For any match probability* $\gamma \in (0,1]$*,*

$$t_\gamma(n) = \frac{(1+2\sqrt{\gamma})^{n+3/2}}{2\gamma^{3/4}\pi^{1/2}n^{3/2}}\left(1+O\left(\frac{1}{(\log n)^{3/2}}\right)\right)$$

(this includes the case $\gamma = 1/4$ *where (6) gives better error terms). In particular, for* $\gamma = 1$*,*

$$t_1(n) = \frac{3^{n+3/2}}{2\pi^{1/2}n^{3/2}}\left(1+O\left(\frac{1}{(\log n)^{3/2}}\right)\right)$$

is the number of structures on $\{1,\ldots,n\}$*.*

Stein and Waterman [2] directly estimated $t_1(n)$, the maximum number of depth-1 structures on $\{1,\ldots,n\}$, using a "folk theorem" in combinatorics.

Assertion 3 *For any match probability $\gamma \in (0,1]$ and sequence length n,*

$$t_\gamma(n) = \frac{(-1)^{n-1}}{2\gamma} \sum_{i=0}^{n+2} \binom{1/2}{i} \binom{1/2}{n+2-i} (1+2\sqrt{\gamma})^i (1-2\sqrt{\gamma})^{n-i+2}.$$

Note that the expression for $t_\gamma(n)$ also subsumes the case $\gamma = 1/4$.

Now write $t_\gamma(n)$ in the form

$$t_\gamma(n) = c_0 + c_1 + \cdots + c_{n+2}\ ,$$

where

$$c_i \stackrel{\text{def}}{=} \frac{(-1)^{n-1}}{2\gamma} \binom{1/2}{i} \binom{1/2}{n+2-i} (1+2\sqrt{\gamma})^i (1-2\sqrt{\gamma})^{n-i+2}\ .$$

Define

$$\Lambda = \frac{1+2\sqrt{\gamma}}{1-2\sqrt{\gamma}}\ .$$

By simple algebra,

$$\frac{|c_{i+1}|}{|c_i|} = |\Lambda| \frac{\left(1+\frac{3}{2n-2i+1}\right)}{\left(1+\frac{3}{2i-1}\right)}\ .$$

It is hence clear that $|c_{i+1}|/|\Lambda c_i|$ monotonically increases with i. As $|\Lambda| > 1$, it follows that for some integer i_0, the sequence $\{|c_i|\}$ decreases monotonically for $0 \le i < i_0$ and increases monotonically for $i_0 \le i \le n+2$; (i_0 can be 0, as when $|\Lambda| > 4$). Note, however, that as $|\Lambda| > 1$, it suffices for $|c_{i+1}| > |c_i|$ that $|\Lambda|/(2n-2i+1) > 1/(2i-1)$, or $i > \frac{2n+|\Lambda|+1}{2(|\Lambda|+1)}$. Hence,

$$i_0 \le \frac{2n+|\Lambda|+1}{2(|\Lambda|+1)} < \frac{n}{2}\ .$$

Henceforth in this section, let $\log \equiv \log_{|\Lambda|}$. Then,

$$|c_0| = \left| \frac{1}{2\gamma} \binom{1/2}{n+2} (1-2\sqrt{\gamma})^{n+2} \right| = O(|1-2\sqrt{\gamma}|^n)$$

is exponentially subdominant with respect to

$$|c_{n+2-\log n}| \sim \frac{(1+2\sqrt{\gamma})^{n+2-\log n} |1-2\sqrt{\gamma}|^{\log n}}{8\gamma\pi n^{3/2} (\log n)^{3/2}} = \Omega(|1+2\sqrt{\gamma}|^n)\ .$$

Therefore all terms $c_0, \ldots, c_{n+1-\log n}$ are dominated by $|c_{n+2-\log n}|$.

Write

$$t_\gamma(n) = \sum_{i=0}^{n+2} c_i = \sum_{i=0}^{n+1-\log n} c_i + \sum_{i=n+2-\log n}^{n+2} c_i \stackrel{\text{def}}{=} \sum\nolimits^{(1)} + \sum\nolimits^{(2)}\ .$$

RNA sequences. The boundary conditions eliminate $H_+(s)$, leaving

$$\begin{aligned} H_-(s) &= \frac{2}{s^2}[1-s-(1-2s)^{1/2}] \\ &= \frac{2}{s^2}\left[1-s-\sum_{i\geq 0}\binom{1/2}{i}(-2s)^i\right] \\ &= \sum_{n\geq 0}(-1)^{n-1}\binom{1/2}{n+2}2^{n+3}s^n \ . \end{aligned}$$

Using the identity

$$\binom{1/2}{k} = \frac{(-1)^{k-1}}{k2^{2k-1}}\binom{2k-2}{k-1} \ , \tag{8}$$

we have

$$t_{1/4}(n) = (-1)^{n-1}\binom{1/2}{n+2}2^{n+3} = \frac{1}{(n+2)2^n}\binom{2n+2}{n+1} = 2^{-n}C_{n+1} \ ,$$

in agreement with (6).

Case 2: $\gamma \notin \{0, 1/4\}$.

Let $\alpha \overset{\text{def}}{=} 1+2\sqrt{\gamma}$ and $\beta \overset{\text{def}}{=} 1-2\sqrt{\gamma}$. Have

$$\begin{aligned} (1-\alpha s)^{1/2}(1-\beta s)^{1/2} &= \sum_{i\geq 0}\sum_{j\geq 0}(-1)^{i+j}\binom{1/2}{i}\binom{1/2}{j}\alpha^i\beta^j s^{i+j} \\ &= \sum_{h\geq 0}\left[(-1)^h\sum_{j=0}^{h}\binom{1/2}{j}\binom{1/2}{h-j}\alpha^j\beta^{h-j}\right]s^h \\ &\overset{\text{def}}{=} \sum_{h\geq 0} d_h s^h \ . \end{aligned}$$

Then

$$d_0 = \binom{1/2}{0}\binom{1/2}{0}\alpha^0\beta^0 = 1 \ ,$$

and

$$d_1 = -\binom{1/2}{0}\binom{1/2}{1}\alpha^0\beta^1 - \binom{1/2}{1}\binom{1/2}{0}\alpha^1\beta^0 = -\frac{\alpha+\beta}{2} = -1 \ .$$

Again, the boundary conditions eliminate $H_+(s)$ leaving

$$H_-(s) = \frac{1-s-\sum_{h\geq 0} d_h s^h}{2\gamma s^2} = -\sum_{h\geq 2}\frac{d_h s^h}{2\gamma s^2} = -\sum_{n\geq 0}\frac{d_{n+2}}{2\gamma}s^n \ . \tag{9}$$

Equating corresponding coefficients in the expressions (7) and (9), we have the following

The two bounds differ by a factor proportional to n. To eliminate this discrepancy, we use generating functions.

We begin by setting up the base for a recurrence:

$$t_\gamma(n) = \begin{cases} 0 & \text{if } n < 0, \\ 1 & \text{if } n = 0 \text{ or } n = 1. \end{cases}$$

The extension of the domain of $t_\gamma(n)$ to the negative integers accords later notational simplicity; when n is 0 or 1, only the empty structure is permissible, hence $t_\gamma(0) = t_\gamma(1) = 1$. We can now set up the recurrence

$$t_\gamma(n) = t_\gamma(n-1) + \gamma \sum_{i=1}^{n-1} t_\gamma(i-1) t_\gamma(n-i-1) \qquad (n \geq 2).$$

The first summand represents the expected number of structures when no match contains 1, and the ith term in the second summand represents the expected number of structures when 1 is matched to $i+1$, and relies on the fact that, as the structure has depth 1, this match partitions such a structure into two independent structures, one on $\{2, \ldots, i\}$, the other on $\{i+2, \ldots, n\}$, hence the expectation of the product is the product of the expectations.

Let

$$H(s) = \sum_n t_\gamma(n) s^n \tag{7}$$

be the generating function of $\{t_\gamma(n)\}$. Then

$$\gamma s^2 H(s)^2 + (s-1)H(s) + 1 = 0 \ .$$

Trivial Case 0: $\gamma = 0$ — no matches can be made.

We then have

$$H(s) = \frac{1}{1-s} = 1 + s + s^2 + \cdots$$

implying

$$t_0(n) = 1 \ .$$

Without loss of generality now assume $\gamma \neq 0$. Then

$$\begin{aligned} H_\pm(s) &= \frac{1 - s \pm \sqrt{(1-s)^2 - 4\gamma s^2}}{2\gamma s^2} \\ &= \frac{1 - s \pm [1 - (1 + 2\sqrt{\gamma})s]^{1/2} [1 - (1 - 2\sqrt{\gamma})s]^{1/2}}{2\gamma s^2} \ . \end{aligned}$$

Case 1: $\gamma = 1/4$. As mentioned earlier, this corresponds to uniformly distributed

In the special case where $\gamma = \frac{1}{4}$, which corresponds to uniformly distributed RNA sequences, we can determine $t_\gamma(n)$ exactly.

$$\begin{aligned} t_{1/4}(n) &= \frac{1}{n+1}\sum_{m=0}^{\lfloor n/2\rfloor}\binom{n+1}{m,\, m+1,\, n-2m}\left(\frac{1}{4}\right)^m \\ &= \frac{2^{-n}}{n+1}\sum_{m=0}^{\lfloor n/2\rfloor}\binom{n+1}{m,\, m+1,\, n-2m}2^{n-2m}\ . \end{aligned}$$

To calculate the sum, consider the set of $\binom{2n+2}{n}$ binary sequences of length $2n+2$, consisting of n zeros and $n+2$ ones. The parsing of such a sequence into $n+1$ pairs of consecutive bits contains m both-zero-pairs, $m+1$ both-one-pairs, and $n-2m$ mixed-pairs, where $0 \le m \le \lceil n/2\rceil$. For example, the 10-bit ($n = 4$) sequence 0111100011 is parsed into $01, 11, 10, 00, 11$, hence contains one both-zero-pair, two both-one-pairs, and two mixed-pairs. The number of sequences yielding a given m is $\binom{n+1}{m,\, m+1,\, n-2m}2^{n-2m}$. Therefore,

$$\sum_{m=0}^{\lfloor n/2\rfloor}\binom{n+1}{m,\, m+1,\, n-2m}2^{n-2m} = \binom{2n+2}{n}\ ,$$

implying that

$$t_{1/4}(n) = \frac{2^{-n}}{n+1}\binom{2n+2}{n} = \frac{2^{-n}}{n+2}\binom{2n+2}{n+1}\ .$$

With C_{n+1} denoting the $(n+1)$th Catalan number, we hence have the following particularly simple expression for $t_{1/4}(n)$:

Assertion 2 *For any sequence length n, when the match probability $\gamma = 1/4$, we have*

$$t_{1/4}(n) = 2^{-n}C_{n+1}\ . \tag{6}$$

For $\gamma \ne \frac{1}{4}$, we could use the results of the last section only to obtain a crude estimate. Bounding the right-hand side of (5), we obtain

$$\begin{aligned} t_\gamma(n) &\le \frac{1}{\sqrt{\gamma}(n+1)}\sum_{\substack{m,m'\ge 0\\ m+m'\le n+1}}\binom{n+1}{m,\, m',\, n+1-m-m'}\gamma^{\frac{1}{2}(m+m')} \\ &= \frac{1}{\sqrt{\gamma}(n+1)}(1+2\sqrt{\gamma})^{n+1}\ , \end{aligned}$$

where the last equality follows from the multinomial theorem.

On the other hand, Theorem 1 implies that

$$t_\gamma(n) \ge t_\gamma(n, m^*) = \frac{1}{2\pi\gamma n^2}(1+2\sqrt{\gamma})^{n+5/2}\left(1+O\left(\frac{1}{n}\right)\right)\ .$$

interpolate $R(m)$. Solving for $R(x^*) = 1$, we have,

$$(4\gamma - 1)(x^*)^2 - (4\gamma n - 2\gamma + 3)x^* + (\gamma n^2 - \gamma n - 2) = 0.$$

Suppose $\gamma \neq 1/4$. Then,

$$\begin{aligned} x^* &= \frac{4\gamma n - 2\gamma + 3 \pm \sqrt{4\gamma n^2 + 20\gamma n + 4\gamma^2 + 20\gamma + 1}}{8\gamma - 2} \\ &= \frac{4\gamma n - 2\gamma + 3 \pm 2\sqrt{\gamma} n\left(1 + \frac{5}{2n}\right)}{8\gamma - 2} + O\left(\frac{1}{n}\right). \end{aligned}$$

The + solution falls outside the range $\{0, \ldots, \lfloor n/2 \rfloor\}$ (both for $\gamma < 1/4$ and for $\gamma > 1/4$), hence

$$\begin{aligned} x^* &= \left(\frac{4\gamma - 2\sqrt{\gamma}}{8\gamma - 2}\right) n + \frac{3 - 2\gamma - 5\sqrt{\gamma}}{8\gamma - 2} + O\left(\frac{1}{n}\right) \\ &= \frac{n + 2.5}{2 + \frac{1}{\sqrt{\gamma}}} - 1.5 + O\left(\frac{1}{n}\right), \end{aligned}$$

in accord with the heuristic calculation in (4). (It is easy to verify that this also subsumes the case $\gamma = 1/4$.) We therefore have:

Theorem 1 *The expected number $t_\gamma(n, m)$ of depth-1, order-m, loop length $l \geq 1$ structures on $\{1, \ldots, n\}$ that are valid for a random sequence with match probability γ is maximized when m attains the value*

$$m^* = \left\lceil \frac{n + 2.5}{2 + \frac{1}{\sqrt{\gamma}}} - 1.5 + O\left(\frac{1}{n}\right) \right\rceil,$$

and its value then is

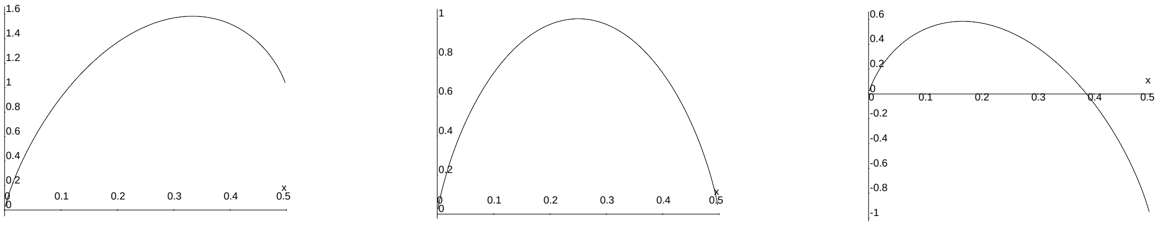
$$t_\gamma(n, m^*) = \frac{1}{2\pi\gamma n^2}\left(1 + 2\sqrt{\gamma}\right)^{n+5/2}\left(1 + O\left(\frac{1}{n}\right)\right).$$

3 Structures of depth-1

Now consider $t_\gamma(n)$, the expected number of depth-1 structures of arbitrary order and loop length on $\{1, \ldots, n\}$ that are valid for a random sequence selected according to a probability distribution with match-probability γ. Using results derived in the last section, we express $t_\gamma(n)$ as a simple sum and determine its value when $\gamma = 1/4$. We employ generating functions to evaluate $t_\gamma(n)$ for $\gamma \neq 1/4$.

Using Assertion 1 we have

$$t_\gamma(n) = \sum_{m=0}^{\lfloor n/2 \rfloor} t_\gamma(n, m) = \frac{1}{n+1} \sum_{m=0}^{\lfloor n/2 \rfloor} \binom{n+1}{m,\ m+1,\ n-2m} \gamma^m. \tag{5}$$



$\gamma = 1$ $\gamma = 1/4$ $\gamma = 1/16$

Figure 2: Exponential part of $t_\gamma(n,m)$ vs. m/n for various values of γ.

Differentiation shows that for any y,

$$\max_{0\le x\le 1}\{h(x)+xy\} = y + \log(1+2^{-y}) \ ,$$

and that this value is achieved when x takes a value $x^* = \frac{1}{1+2^{-y}}$. It follows that

$$\max_{0\le x\le 1}\{h(x)+(1+\log\sqrt{\gamma})x\} = \log(1+2\sqrt{\gamma}) \tag{3}$$

and that this value is achieved when x takes the value $x^* = 1/\Big(1+\frac{1}{2\sqrt{\gamma}}\Big)$.

Hence, the exponent of $t_\gamma(n,m)$, given in (2), is maximized when m attains the value $n/(2+\frac{1}{\sqrt{\gamma}})$, and the exponent then is $n\log(1+2\sqrt{\gamma})$. The preceding derivation however ignored the non-exponential part of $t_\gamma(n,m)$. This, however, can affect the location of the maximum of $t_\gamma(n,m)$ by at most a constant. Thus, $t_\gamma(n,m)$ is maximized when m attains a value

$$m^* = \frac{n}{2+\frac{1}{\sqrt{\gamma}}} + O(1) \ , \tag{4}$$

and

$$\log t_\gamma(n,m^*) \sim n\log(1+2\sqrt{\gamma}) \qquad (n\to\infty) \, .$$

In order to determine the $O(1)$ term in m^*, let

$$R(m) \stackrel{\text{def}}{=} \frac{t_\gamma(n,m+1)}{t_\gamma(n,m)} = \frac{(n-2m-1)(n-2m)}{(m+1)(m+2)}\gamma$$

be the ratio between successive values of $t_\gamma(n,m)$ for $m \in \{0,\ldots,\lfloor n/2\rfloor - 1\}$. It is simple to see that $R(m)$ decreases with m, hence $t_\gamma(n,m)$ is unimodal, achieving its maximum value at the first integer m^* satisfying $R(m^*) \le 1$.

Let

$$R(x) \stackrel{\text{def}}{=} \frac{(n-2x-1)(n-2x)}{(x+1)(x+2)}\gamma$$

$$
\begin{aligned}
&= \sum_{\sigma\in T(n,m)} \sum_{s\in\mathcal{B}^n} \mathbf{P}(s)1(\sigma \text{ is valid for } s \text{ under } \mathcal{R})\\
&= \sum_{\sigma\in T(n,m)} \mathbf{P}(\sigma \text{ is valid for a random string under } \mathcal{R})\\
&= \sum_{\sigma\in T(n,m)} \gamma^m\\
&= |T(n,m)|\gamma^m.
\end{aligned}
$$

We have hence obtained the following simple expression for $t_\gamma(n,m)$.

Assertion 1 *For any match probability $\gamma \in (0,1]$, sequence length n, and order m,*

$$t_\gamma(n,m) = \frac{1}{m+1}\binom{n}{m,\, m,\, n-2m}\gamma^m \;.$$

We now approximate $t_\gamma(n,m)$ for large n and m. Applying Stirling's approximation formula

$$n! = \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n\Big(1+O\Big(\frac{1}{n}\Big)\Big)$$

to the assertion, we see that for m proportional to n,

$$
\begin{aligned}
t_\gamma(n,m) &= \frac{1}{2\pi m(m+1)}\sqrt{\frac{n}{n-2m}}\,\frac{n^n}{m^m m^m (n-2m)^{n-2m}}\,\gamma^m\Big(1+O\Big(\frac{1}{n}\Big)\Big)\\
&= \frac{1}{2\pi m(m+1)}\sqrt{\frac{n}{n-2m}}\,2^{nh\left(\frac{m}{n},\frac{m}{n},\frac{n-2m}{n}\right)+m\log\gamma}\Big(1+O\Big(\frac{1}{n}\Big)\Big).
\end{aligned}
$$

We hence have the following insight into the growth pattern of $t_\gamma(n,m)$ as n increases and the relative order $m/n = a$ remains fixed. In fact, suppose m is proportional to n, i.e., $m = an$ with $a = \Theta(1)$ bounded above (away from 1/2) and below (away from 0). Then

$$t_\gamma(n,m) = \frac{1}{2\pi a^2 n^2\sqrt{1-2a}}\,2^{n\left(h(2a)+2a(1+\log\sqrt{\gamma})\right)}\Big(1+O\Big(\frac{1}{n}\Big)\Big)\;.$$

Thus, when the relative order $m/n = a$ is fixed, $t_\gamma(n,m)$ grows exponentially with

$$n(h(2a)+2a(1+\log\sqrt{\gamma}))\;. \tag{2}$$

Figure 2 shows how this exponent varies with m/n for $\gamma = 1$ (all base-pairs match), for $\gamma = 1/4$ (uniformly distributed RNA sequences), and for $\gamma = 1/16$. Note that when $\gamma < 1/4$, for large relative orders the exponent is negative, hence the expected number of structures decreases with n. We now determine the order having the maximal expected number of structures. We first do so informally, concentrating only on the exponent, then perform a precise calculation.

2 Structures of depth-1, given order, and arbitrary loop length

Recall that $t_\gamma(n,m)$ is the expected number of depth-1, order-m, arbitrary loop length $l \geq 1$ structures on $\{1,\ldots,n\}$ valid for a sequence chosen according to a probability distribution with match probability γ. We first calculate $t_\gamma(n,m)$, then approximate it for large n and m.

A sequence of parentheses is *well formed* if it contains the same number of left and right parentheses, and when scanned from left to right, the number of right parentheses never exceeds the number of left parentheses. The number of well-formed sequences of $2m$ parentheses is well known to be the mth Catalan number

$$C_m \stackrel{\mathrm{def}}{=} \frac{1}{m+1}\binom{2m}{m} . \tag{1}$$

An order-m structure on $\{1,\ldots,2m\}$ can be mapped into a well-formed sequence of $2m$ parentheses by setting the ith parenthesis to be a left parenthesis if i is the smaller integer in its structure pair, and making it a right parenthesis if i is the larger integer in its pair. This mapping is surjective but not injective. However, when restricted to depth-1, arbitrary loop length structures, i.e., to structures in $T(2m,m)$, the mapping can be shown to be a bijection. For example, the structures $\{(1,6),(2,3),(4,5)\}$ and $\{(1,3),(2,6),(4,5)\}$, both map into $(()())$, however, $\{(1,6),(2,3),(4,5)\}$ is the only depth-1 structure mapped to $(()())$. Therefore,

$$|T(2m,m)| = C_m .$$

Each structure in $T(n,m)$ corresponds to a $2m$-element subset of $\{1,\ldots,n\}$ (consisting of the elements included in structure pairs) and to an order-m structure on $\{1,\ldots,2m\}$. Hence,

$$|T(n,m)| = \binom{n}{2m} C_m = \frac{1}{m+1}\binom{n}{m,\ m,\ n-2m} = \frac{1}{n+1}\binom{n+1}{m,\ m+1,\ n-2m} .$$

For the expected value, let $1(\text{statement})$ be the indicator function that is 1 if 'statement' is true and is 0 otherwise. We noted that the probability that a given structure pair matches for a random sequence s under $\mathcal{R}$ is the match probability γ. By independence, the probability that a given structure of order m is valid for s under $\mathcal{R}$ is γ^m. Therefore,

$$\begin{aligned} t_\gamma(n,m) &= \sum_{s\in\mathcal{B}^n} \mathbf{P}(s)|T_{\mathcal{R}}(s,m)| \\ &= \sum_{s\in\mathcal{B}^n} \mathbf{P}(s) \sum_{\sigma\in T(n,m)} 1(\sigma \text{ is valid for } s \text{ under } \mathcal{R}) \end{aligned}$$

In particular, when $\gamma = 1$ we obtain the interesting result

$$t_1^{1,2}(n, m^*) = \Theta\left(\frac{\phi^{2n}}{n^2}\right),$$

where $\phi = (1+\sqrt{5})/2 \approx 1.618$ is the golden ratio. As for the $d = 1$, $l = 1$ case, for general values of γ we show that $t_\gamma^{1,2}(n)$ has the hypergeometric representation

$$t_\gamma^{1,2}(n) = {}_4\mathrm{F}_3\left(\begin{matrix} -\frac{1}{2}n+1,\ -\frac{1}{2}n+\frac{1}{2},\ -\frac{1}{2}n+\frac{1}{2},\ -\frac{1}{2}n \\ 2,\ -n+1,\ -n \end{matrix} \,\middle|\, 16\gamma \right),$$

and show that

$$t_\gamma^{1,2}(n) = O\left(\frac{1}{n}\left(\frac{1+\sqrt{1+4\sqrt{\gamma}}}{2}\right)^{2n}\right).$$

Again, when $\gamma = 1$, $t_1^{1,2}(n)$ has the very alluring dependence on the golden ratio:

$$\log t_1^{1,2}(n) \sim 2n \log \phi \qquad (n \to \infty).$$

For general l, a series solution for $t_\gamma^{1,l}(n, m)$ (and hence for $t_\gamma^{1,l}(n)$ also) is readily obtained from the generating function. It does not appear likely, however, that the series solution can be resolved into a simple closed form for general l.

is obtained when there are about $\frac{\sqrt{\gamma}}{1+2d\sqrt{\gamma}}n$ pairs of each color. We then approximate the total expected number $t_\gamma^{d,1}(n)$ of depth $\le d$ structures of arbitrary order and loop length, and show that

$$t_\gamma^{d,1}(n) = O\left(\frac{(1+2d\sqrt{\gamma})^n}{n^d}\right).$$

Finally, in Section 5 we consider structures of depth-1, order-m, and loop length $\ge l$. Setting up a recurrence for $t_\gamma^{1,l}(n,m)$, we obtain the following simple closed form expression for the bivariate generating function

$$H(z,w) = \sum_n \sum_m t_\gamma^{1,l}(n,m) z^n w^m = \frac{1}{2\gamma w z^2 P(z,w)} \left[1 - \sqrt{1 - 4\gamma w z^2 P(z,w)^2}\right],$$

where

$$P(z,w) = \frac{1-z}{1-2z+(1+\gamma w)z^2 - \gamma w z^{l+1}}.$$

The particular case $l = 2$, namely structures where adjacent elements cannot be matched, is of interest. An examination of the generating function yields the interesting result

$$t_\gamma^{1,2}(n,m) = \frac{1}{m}\binom{n-m}{m+1}\binom{n-m-1}{m-1}\gamma^m.$$

The particularization of this result when $\gamma = 1$ has also been obtained recently by Schmidtt and Waterman [3] using an ingenious, non-intuitive argument whose key element is the replacement of the problem at hand by an equivalent combinatorial problem on linear trees which has a known solution. The approach outlined here in contradistinction is a straightforward combinatorial attack on the generating function.

For large n and m proportional to n, i.e., $m = an$ with $a = \Theta(1)$ bounded above (away from 1/2) and below (away from 0), approximations show that

$$t_\gamma^{1,2}(n,m) = \frac{1}{2\pi a^2 n^2} 2^{2n\left(h(2a)-h(a)+a(2+\log\sqrt{\gamma})\right)}\left(1+O\left(\frac{1}{n}\right)\right).$$

Carrying the analysis further, we show that, for fixed n, $t_\gamma^{1,2}(n,m)$ attains its maximum value when m takes a value m^* given by

$$m^* = \frac{1}{2}\left(1 - \frac{1}{\sqrt{1+4\sqrt{\gamma}}}\right)n + O(1),$$

and show that this maximum value is

$$t_\gamma^{1,2}(n,m^*) = \frac{(1+4\sqrt{\gamma})}{\pi\left(1+2\sqrt{\gamma}-\sqrt{1+4\sqrt{\gamma}}\right)}\frac{1}{n^2}\left(\frac{1+\sqrt{1+4\sqrt{\gamma}}}{2}\right)^{2n}\left(1+O\left(\frac{1}{n}\right)\right).$$

For large n and for m proportional to n, i.e., $m = an$ with $a = \Theta(1)$ bounded above (away from $1/2$) and below (away from 0), approximations show that

$$t_\gamma(n,m) = \frac{1}{2\pi a^2 n^2 \sqrt{1-2a}}\, 2^{n\left(h(2a)+2a(1+\log\sqrt{\gamma})\right)} \left(1 + O\Big(\frac{1}{n}\Big)\right) .$$

We prove that, for fixed n, $t_\gamma(n,m)$ is maximized when m takes a value m^* given by

$$m^* = \left\lceil \frac{n+2.5}{2+\frac{1}{\sqrt{\gamma}}} - 1.5 + O\Big(\frac{1}{n}\Big) \right\rceil$$

and that its value then is

$$t_\gamma(n,m^*) = \frac{1}{2\pi\gamma n^2}\,(1+2\sqrt{\gamma})^{n+2.5} \left(1 + O\Big(\frac{1}{n}\Big)\right) .$$

In Section 3 we consider the expected number $t_\gamma(n)$ of depth-1 structures on $\{1,\ldots,n\}$ having arbitrary order and loop length. For $\gamma = 1/4$, which corresponds to uniformly-distributed RNA sequences, we use direct calculations to show that

$$t_{1/4}(n) = \frac{2^{-n}}{n+2}\binom{2n+2}{n+1} .$$

For general values of γ we use generating functions to show that $t_\gamma(n)$ has the hypergeometric representation

$$t_\gamma(n) = {}_2\mathrm{F}_1\left(\begin{matrix} -\frac{1}{2}n+\frac{1}{2},\ -\frac{1}{2}n \\ 2 \end{matrix} \,\middle|\, 4\gamma \right) ,$$

and prove that

$$t_\gamma(n) = \frac{(1+2\sqrt{\gamma})^{n+3/2}}{2\gamma^{3/4}\pi^{1/2}n^{3/2}}\left(1 + O\bigg(\frac{1}{(\log n)^{3/2}}\bigg)\right) .$$

In particular,

$$t_1(n) = \frac{3^{n+3/2}}{2\pi^{1/2}n^{3/2}}\left(1 + O\bigg(\frac{1}{(\log n)^{3/2}}\bigg)\right)$$

is the number of structures on $\{1,\ldots,n\}$ determined previously by Stein and Waterman [2].

In Section 4 we address structures of depth $\leq d$ where d is any fixed integer. Each pair in such a structure can be viewed as being colored in one of d colors so that no two pairs of the same color intersect. We approximate the number of structures on $\{1,\ldots,n\}$ that have arbitrary loop length and a given number of pairs of each color. We show that the largest number of structures,

$$\frac{1}{(2\pi\gamma)^d n^{2d}}(1+2d\sqrt{\gamma})^{n+2d+1/2}\left(1 + O\Big(\frac{1}{n}\Big)\right) ,$$

The rising factorial $a^{\overline{k}}$ is defined for every real a and integer $k \geq 0$ by

$$a^{\overline{k}} = a(a+1)\dots(a+k-1).$$

Similarly, the falling factorial $a^{\underline{k}}$ is defined for real a and integer $k \geq 0$ by

$$a^{\underline{k}} = a(a-1)\dots(a-k+1).$$

(In accordance with the usual custom that a product over a null set is 1, we set $a^{\overline{0}} = a^{\underline{0}} = 1$.) The binomial coefficients $\binom{r}{k}$ are defined for every real r and integer k by

$$\binom{r}{k} = \begin{cases} 0 & \text{if } k < 0, \\ \frac{r^{\underline{k}}}{k!} & \text{if } k \geq 0. \end{cases}$$

We will also have recourse to multinomial coefficients: if $m_1, \dots, m_j$ are integers, $\sum_{i=1}^{j} m_i = m$, define

$$\binom{m}{m_1, \dots, m_j} = \begin{cases} 0 & \text{if any } m_i < 0, \\ \frac{m!}{m_1! \dots m_j!} & \text{if all } m_i \geq 0. \end{cases}$$

The general hypergeometric series with m upper parameters and n lower parameters is formally defined by

$${}_m\mathrm{F}_n\left(\begin{matrix} a_1, \dots, a_m \\ b_1, \dots, b_n \end{matrix} \middle| z \right) = \sum_{k \geq 0} \frac{a_1^{\overline{k}} \dots a_m^{\overline{k}}}{b_1^{\overline{k}} \dots b_n^{\overline{k}}} \frac{z^k}{k!}.$$

None of the lower parameters b_i can be zero or a negative integer, but other than that the upper parameters a_i and the lower parameters b_i can be arbitrary.

All logarithms (with one clearly indicated exception) are to base 2. Finally, if $\{\alpha_1, \dots, \alpha_k\}$ is a discrete probability distribution, define the binary entropy function by

$$h(\alpha_1, \dots, \alpha_k) = -\sum_{i=1}^{k} \alpha_i \log \alpha_i,$$

with the usual convention $0 \log 0 = 0$. In accordance with customary usage, for a Bernoulli probability distribution $\{\alpha, 1-\alpha\}$ we write $h(\alpha)$ instead of $h(\alpha, 1-\alpha)$.

Overview of Results In Section 2 we determine an exact expression for $t_\gamma(n, m)$, the expected number of structures on $\{1, \dots, n\}$ that have depth-1, order-m, and arbitrary loop length $l \geq 1$ that are valid for a randomly chosen sequence s with match probability γ. We show that

$$t_\gamma(n, m) = \frac{1}{m+1} \binom{n}{m,\ m,\ n-2m} \gamma^m .$$

s if (s_i, s_j) is a matching base-pair. A structure on $\{1, \ldots, n\}$ is *valid for s* if all its pairs match in s. In the visualization above, all chords connect indices corresponding to matching base pairs.

For RNA molecules, $\mathcal{B} = \{A, C, G, U\}$ and $\mathcal{R}$ is the symmetric relation $\{(A, U), (U, A), (C, G), (G, C)\}$. The pair $(1, 2)$ matches in the sequence $s = UAGC$ while the pair $(1, 3)$ does not. The order-2, depth-1, loop-length-1, structure $\{(1, 2), (3, 4)\}$ is valid for s whereas the order-2, depth-2, loop-length-2, structure $\{(1, 3), (2, 4)\}$ is not. Secondary RNA structures have been frequently modeled as structures of depth 1 or 2 and minimum loop length 4 or 5 on long sequences.

Let $T^{d,l}(n, m)$ be the set of structures on $\{1, \ldots, n\}$ that have order m, depth $\le d$, and loop-length $\ge l$. Note the $T^{d,l}(n, m)$ is independent of $\mathcal{R}$. Let $T^{d,l}_{\mathcal{R}}(s, m)$ be the set of structures in $T^{d,l}(n, m)$ that are valid for a sequence s of length n (under $\mathcal{R}$) and let

$$T^{d,l}_{\mathcal{R}}(s) = \bigcup_{m=0}^{\lfloor n/2 \rfloor} T^{d,l}_{\mathcal{R}}(s, m) ,$$

be the set of structures of depth $\le d$, loop length $\ge l$, and arbitrary order, that are valid for s (under $\mathcal{R}$).

We are mostly interested in the expected values of $|T^{d,l}_{\mathcal{R}}(s, m)|$ and $|T^{d,l}_{\mathcal{R}}(s)|$ when the sequence s is formed randomly—each element s_i is chosen independently according to some probability distribution $\mathbf{P}$ over $\mathcal{B}$. The *match probability* of the distribution $\mathbf{P}$ with respect to $\mathcal{R}$ is

$$\gamma \stackrel{\text{def}}{=} \sum_{(a,b) \in \mathcal{R}} \mathbf{P}(a)\mathbf{P}(b) ,$$

the probability that two elements of $\mathcal{B}$, chosen independently according to $\mathbf{P}$ are a matching base pair. By construction of s, it is also the the probability that a given structure pair matches in a sequence s chosen randomly according to $\mathbf{P}$. We will see that the expected values of $|T^{d,l}_{\mathcal{R}}(s, m)|$ and $|T^{d,l}_{\mathcal{R}}(s)|$ depend on $\mathbf{P}$ and $\mathcal{R}$ only through γ, and denote them by $t^{d,l}_{\gamma}(n, m)$ and $t^{d,l}_{\gamma}(n)$ respectively. By linearity of expectations,

$$t^{d,l}_{\gamma}(n) = \sum_{m=0}^{\lfloor n/2 \rfloor} t^{d,l}_{\gamma}(n, m) .$$

We are mostly interested in structures of depth 1 and loop length ≥ 1. To simplify notation, we omit the superscripts d and l if both are 1. For example, $T^{d,1}(n, m)$ is the set of order-m structures on $\{1, \ldots, n\}$ that have depth $\le d$ and arbitrary loop length; $T^{1,l}_{\mathcal{R}}(s, m)$ is the set of depth-1, order-m structures on $\{1, \ldots, n\}$ that have loop length $\ge l$ and are valid for the sequence s; and $t_{\gamma}(n, m)$ is the expected number of structures on $\{1, \ldots, n\}$ that have depth-1, order-m, arbitrary loop length $l \ge 1$, and are valid for a randomly chosen sequence s.

Notational Conventions Define $0^0 = 1$. When explicit limits are absent in sums of the form $\sum_k$, it is to be understood that the summation variable (here k) ranges over all integer values in the range $-\infty$ to $+\infty$.

secondary structures possible when any two bases can bond to each other. Zuker and Sankof [4] refined some of the calculations in [1, 2] by imposing the constraint that only certain base pairs can bond to each other and calculating the expected number of secondary structures when the sequence of bases (the primary structure) is chosen randomly.

Continuing in this line of research, we consider the expected number of structures under various constraints on the type and the number of bonds allowed. As in [4], we assume that the sequence of bases is chosen randomly. We begin with a mathematical description.

A *structure* S *on* $\{1, \ldots, n\}$ is a collection of disjoint *pairs* (i, j) where $1 \leq i < j \leq n$. The *order* of S is the number of pairs it contains. Its *loop length* is the smallest difference between the elements of a pair. Two pairs (i, j) and (i', j') *intersect* if $(i' - i) \cdot (j' - j) > 0$. The *depth* of S is the smallest integer d such that S can be partitioned into d structures, each having the property that no two of its pairs intersect.

To "visualize" S, place the integers from 1 to n in increasing order on a circle and for every pair $(i, j) \in S$ draw the corresponding chord. By definition, the number of chords is the order of S, each chord connects two distinct integers, and no integer is connected to more than one chord. The loop length of S is the minimum difference between the integers at the endpoints of a chord. Two pairs intersect if their corresponding chords cross each other. S is of depth 1 if no two chords cross. The structure S is of depth $\leq d$ if the chords can be drawn in d colors so that no two chords of the same color cross. Figure 1 illustrates the circles corresponding to two structures. Note that a structure of depth 2 corresponds to a planar graph on a circle: draw the edges of one colored set inside the circle, and the others outside.

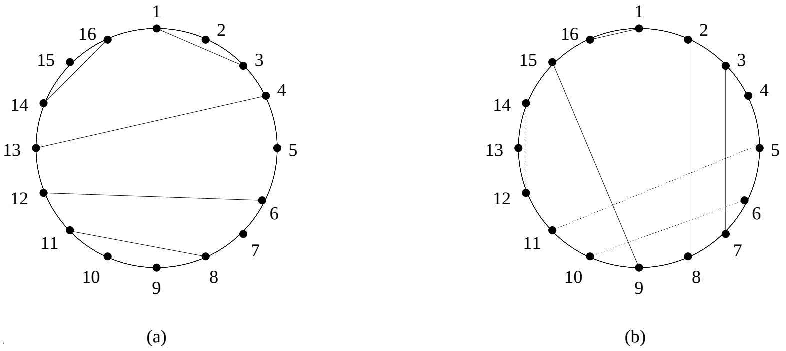


Figure 1: Two structures of loop-length 2 on $\{1, \ldots, 16\}$. Structure (a) has depth 1 and, as indicated by the solid and dotted lines, structure (b) has depth 2.

Throughout the paper, we assume a fixed set $\mathcal{B}$ of *bases* and a symmetric relation $\mathcal{R}$ of *matching base-pairs*. Namely, $\mathcal{R} \subseteq \mathcal{B} \times \mathcal{B}$, and $(b, b') \in \mathcal{R}$ implies $(b', b) \in \mathcal{R}$. Let $s \overset{\text{def}}{=} s_1 \ldots s_n$ be a sequence of n bases. A pair (i, j) of distinct indices *matches in*

On edge-colored interior planar graphs on a circle and the expected number of RNA secondary structures

Alon Orlitsky* Santosh S. Venkatesh†

First Draft: 19 November 1992
Second Draft: 6 January 1993

INDEX TERMS: *Planar graphs, ribonucleic acid (RNA), secondary structure, generating functions, hypergeometric series.*

Abstract

Using a mathematical model for an RNA molecule as a family of disjoint edge-colored interior planar graphs on a circle, we determine the expected number of secondary RNA structures that can form under various assumptions on the type and number of ribonucleotide bonds.

1 Introduction

A ribonucleic acid (RNA) molecule consists of a sequence of ribonucleotides. Each ribonucleotide contains one of four bases—adenine, cytosine, guanine, and uracil, commonly denoted by A, C, G, and U, respectively. As the molecule forms, chemical bonds join certain adenine and uracil nucleotide pairs and certain cytosine and guanine nucleotide pairs. These bonds constrain the three-dimensional configuration of the molecule which in turn influences its function.

The sequence of bases in an RNA molecule is referred to as its *primary structure*, the set of bonded nucleotide pairs is the molecule's *secondary structure*, and the three-dimensional configuration is the *tertiary structure* of the molecule. Waterman [1], Stein and Waterman [2], and Schmidtt and Waterman [3] have studied the number of

---

*AT&T Bell Laboratories, 600 Mountain Av., Murray Hill, NJ 07974.

†Department of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104; work on this paper was largely done while the author was visiting AT&T Bell Laboratories, Murray Hill, NJ 07974.