Statistical Analysis of Hairpins and BasePairs in RNA Secondary Structures*

BU AJ · KAUERS Manuel · ZEILBERGER Doron

DOI:
Received: x x 20xx / Revised: x x 20xx

Abstract We derive precise asymptotic expressions for the expectations, variances, covariance, and quite a few further mixed moments for the number of hairpins and the number of basepairs in RNA secondary structures, and give convincing evidence that the central-scaled distribution of the pair of random variables (hairpins, basepairs) tends in distribution to the bi-variate normal distribution with correlation $\sqrt{5\sqrt{5}-11}/2 = 0.2123322205\ldots$

Keywords Experimental Mathematics, Enumerative Combinatorics, Computer Algebra, Automated Asymptotics

Dedication: This article is in fond memory of Kequan Ding, a very deep, original, and versatile mathematician, whose many interests included RNA secondary structures [GD].

1 Introduction

In a seminal paper [W], Michael Waterman presented a combinatorial framework for RNA secondary structures. A secondary structure is a vertex-labeled graph on n vertices obeying certain conditions. See [W], [GD], and [CRU]. These have several *structure elements*, the most important one being *basepair* and another one is *hairpin*. We refer to [CRU] (readily available on-line) for the definitions.

The following theorem is due to Waterman [W]:

BU AJ
Department of Mathematics, Rutgers University (New Brunswick), Hill Center-Busch Campus, 110 Frelinghuysen Rd., Piscataway, NJ 08854-8019, USA. Email: `ab1854 at math dot rutgers dot edu`
KAUERS Manuel
Institute for Algebra, Johannes Kepler Universität, Altenbergerstraße 69 A-4040 Linz, Austria. Email: `manuel dot kauers at jku dot at`
ZEILBERGER Doron
Department of Mathematics, Rutgers University (New Brunswick), Hill Center-Busch Campus, 110 Frelinghuysen Rd., Piscataway, NJ 08854-8019, USA. Email: `doronzeil at gmail dot com`
*The second author was supported by the Austrian FWF under Grant Nos 10.55776/PAT8258123, 10.55776/I6130, and 10.55776/PAT9952223.
◇ *This paper was recommended for publication by Editor .*

Theorem 1 Let $a(n)$ be the number of RNA secondary structures on n vertices. Then

$$\sum_{n=0}^{\infty} a(n)X^n = \frac{-2X^4 + X^3 + 2X - 1 + \sqrt{X^6 - 4X^5 + 4X^4 - 2X^3 + 4X^2 - 4X + 1}}{2X^2\,(X-1)} \quad .$$

Using symbolic computation, one of us [B], proved the following generalization.

Theorem 2 Let $A(n,i,j)$ be the number of RNA secondary structures on n vertices, with exactly i hairpins and j basepairs. Let's define the tri-variate generating function $F(X;x,z)$

$$F(X;x,z) := \sum_{n=0}^{\infty}\sum_{i=0}^{n}\sum_{j=0}^{n} A(n,i,j)\, X^n\, x^i\, z^j \quad ,$$

viewed as a formal power series in X with coefficients that are polynomials in x and z. We have the following *hairy* formula

$$F(X;x,z) \;=\; \frac{-2X^4z + X^3xz + X^2z - X^2 + 2X - 1 + \sqrt{R}}{2X^2z\,(X-1)} \quad ,$$

where

$$R = X^6x^2z^2 - 2X^5x\,z^2 - 2X^5xz + 4X^4xz + X^4z^2 - 2X^4z - 2X^3xz$$
$$+ X^4 + 4X^3z - 4X^3 - 2X^2z + 6X^2 - 4X + 1 \quad .$$

In this *methodological* article (the method is much more important than the actual results!), we will establish the following theorem.

Theorem 3 For a secondary structure s on n vertices, let $X_n(s)$ and $Z_n(s)$ be the number of hairpins and number of basepairs, respectively, of s. We have the following asymptotic expressions

$$\mathrm{E}[X_n] = (1 - \frac{2}{\sqrt{5}})\,n \cdot \left(1 + O(\frac{1}{n})\right) \;=\; (0.1055728092\ldots)\cdot n \cdot \left(1 + O(\frac{1}{n})\right) \quad .$$

$$\mathrm{E}[Z_n] = \frac{5-\sqrt{5}}{10}\,n \cdot \left(1 + O(\frac{1}{n})\right) \;=\; (0.2763932023\ldots)\cdot n \cdot \left(1 + O(\frac{1}{n})\right) \quad .$$

$$\mathrm{Var}(X_n) = (2 - \frac{22}{5\sqrt{5}})\,n \cdot \left(1 + O(\frac{1}{n})\right) \;=\; (0.032260180\ldots)\cdot n \cdot \left(1 + O(\frac{1}{n})\right) \quad .$$

$$\mathrm{Var}(Z_n) = \frac{1}{10\sqrt{5}}\,n \cdot \left(1 + O(\frac{1}{n})\right) \;=\; (0.04472135954\ldots)\cdot n \cdot \left(1 + O(\frac{1}{n})\right) \quad .$$

$$\mathrm{Corr}(X_n, Z_n) = \frac{\sqrt{5\sqrt{5}-11}}{2} \cdot \left(1 + O(\frac{1}{n})\right) \;=\; (0.2123322161\ldots)\cdot \left(1 + O(\frac{1}{n})\right) \quad .$$

More precisely, we have:

$$\begin{aligned}
\mathrm{E}[X_n] &= (1-\frac{2}{5}\sqrt{5})n\cdot\Big(1+(\frac{7}{4}+\frac{11}{20}\sqrt{5})n^{-1}+(\frac{243}{160}+\frac{99}{160}\sqrt{5})n^{-2}+(\frac{339}{160}+\frac{1029}{800}\sqrt{5})n^{-3}\\
&\quad+(\frac{11917563}{819200}+\frac{5400687}{819200}\sqrt{5})n^{-4}+\mathrm{O}(n^{-5})\Big)\\
\mathrm{E}[Z_n] &= (\frac{1}{2}-\frac{1}{10}\sqrt{5})n\cdot\Big(1+(-\frac{5}{8}-\frac{13}{40}\sqrt{5})n^{-1}+(\frac{21}{320}+\frac{21}{320}\sqrt{5})n^{-2}\\
&\quad+(-\frac{93}{320}-\frac{177}{1600}\sqrt{5})n^{-3}+(-\frac{13887249}{1638400}-\frac{6272793}{1638400}\sqrt{5})n^{-4}+\mathrm{O}(n^{-5})\Big)\\
\mathrm{Var}[X_n] &= (2-\frac{22}{25}\sqrt{5})n\cdot\Big(1+(\frac{1}{16}+\frac{23}{80}\sqrt{5})n^{-1}+(-\frac{651}{160}-\frac{177}{64}\sqrt{5})n^{-2}\\
&\quad+(-\frac{1208783}{81920}-\frac{3216693}{409600}\sqrt{5})n^{-3}+\mathrm{O}(n^{-4})\Big)\\
\mathrm{Var}[Z_n] &= (0+\frac{1}{50}\sqrt{5})n\cdot\Big(1+(1+\frac{1}{10}\sqrt{5})n^{-1}-\frac{261}{160}n^{-2}\\
&\quad+(-\frac{27179}{2560}-\frac{915879}{204800}\sqrt{5})n^{-3}+\mathrm{O}(n^{-4})\Big)\\
\mathrm{Corr}[X_n,Z_n] &= \frac{1}{2}\sqrt{5\sqrt{5}-11}\cdot\Big(1+(\frac{15}{32}+\frac{13}{32}\sqrt{5})n^{-1}+(\frac{4469}{1024}+\frac{1345}{1024}\sqrt{5})n^{-2}\\
&\quad+(\frac{1766711}{32768}+\frac{801983}{32768}\sqrt{5})n^{-3}+\mathrm{O}(n^{-4})\Big)
\end{aligned}$$

We have lots of evidence that the following conjecture is true, and one of us (DZ) is offering a donation of $100 to the OEIS for its rigorous proof.

Conjecture The centralized-scaled version of the pair (X_n, Z_n), namely

$$\left(\frac{X_n-\mathrm{E}[X_n]}{\sqrt{\mathrm{Var}(X_n)}},\frac{Z_n-\mathrm{E}[Z_n]}{\sqrt{\mathrm{Var}(Z_n)}}\right) \quad ,$$

tends, in distribution, to the bi-variate normal distribution with covariance $c=\frac{\sqrt{5\sqrt{5}-11}}{2}$, whose probability density function (pdf) is:

$$\frac{e^{-\frac{1}{2}x^2-\frac{1}{2}y^2+cxy}\sqrt{1-c^2}}{2\pi} \quad .$$

2 Reminders about using Symbolic Computations to do Statistics

Given a finite set S, equipped with some random variable, R, $s\to R(s)$, the *expectation*, μ, is, of course

$$\mu=\mathrm{E}[R]:=\frac{1}{|S|}\left(\sum_{s\in S}R(s)\right) \quad ,$$

where, as usual, $|S|$ denotes the number of elements in S. More generally, the i-th moment of R, is the average of the i-th power:

$$\mathrm{E}[R^i]:=\frac{1}{|S|}\left(\sum_{s\in S}(R(s))^i\right) \quad .$$

More useful, are the **central moments**

$$\mathrm{E}[(R-\mu)^i] := \frac{1}{|S|}\left(\sum_{s\in S}(R(s)-\mu)^i\right) \quad ,$$

the most important being the *second*, aka *variance*. Using the binomial theorem, once you know μ, you can derive the central moments from the straight moments.

A convenient way to record the data regarding the random variable R defined on a finite (and even infinite) set, is via its *weight-enumerator*, aka *generating function*, pick any (formal) variable name, say x, and define:

$$f_R(x) := \frac{1}{|S|}\sum_{s\in S} x^{R(s)} \quad .$$

In terms of $f_R(x)$ we have

$$\mu = \mathrm{E}[R] = f_R'(1) = (x\frac{d}{dx})f_R(x)\Big|_{x=1} \quad .$$

More generally, the i-th moment is

$$m_i = \mathrm{E}[R^i] = (x\frac{d}{dx})^i f_R(x)\Big|_{x=1} \quad .$$

Once you know the expectation, μ, and the *variance*, (aka second central moment), $\sigma^2 = E[(R-\mu)^2]$, (where σ, the square-root of the variance, is called the *standard deviation*), you know the two most important numbers, what it is 'on average', and how spread-out it is. But it is still nice to know more, and then one forms the *scaled-central distribution*

$$\overline{R}(s) := \frac{R(s)-\mu}{\sigma} \quad ,$$

whose expectation is 0 and variance is 1. Its higher moments are expressible in terms of the moments of the original random variable R, and hence a knowledge of the weight-enumerator, $f_R(x)$, is useful.

In combinatorics, we often have an *infinite* family of sets S_n, naturally indexed by non-negative integers, and a natural random variable defined on them, with a 'nice' sequence of weight-enumerators $f_n(x)$. For example for the set of subsets of $\{1,\dots,n\}$ with *weight* 'cardinality', we have the closed-form expression, $f_n(x) = (1+x)^n$. If not so lucky, we often have, in combinatorics, that the **grand generating function**,

$$F(X;x) := \sum_{n=0}^{\infty} f_n(x)X^n \quad ,$$

has some explicit expression in X and x, or otherwise, in many combinatorial families, satisfies some *algebraic equation*

$$\mathcal{P}(F,X,x) \,=\, 0 \quad ,$$

where $\mathcal{P}$ is some specific polynomial in its arguments.

If we have an 'explicit' (or even 'implicit') expression for the grand-generating function, that is a certain formal power series in X with coefficients that are polynomials in x, we can get generating functions for the numerical sequences of the moments.

Of course, we also need to handle the cardinality of the family of sets, let's call is $a_0(n)$, whose generating function is

$$\sum_{n=0}^{\infty} a_0(n)\, X^n = F(X;1) \quad .$$

Defining

$$\sum_{n=0}^{\infty} a_i(n)\, X^n = (x\frac{d}{dx})^i F(X;x)\Big|_{x=1} \quad ,$$

the i-th moment is then $\frac{a_i(n)}{a_0(n)}$.

Alas, it is generally not possible to have closed-form expressions for $a_i(n)$ and not even for $a_0(n)$, but one can have the next-best thing, *asymptotics*. If the grand-generating function is *algebraic*, and especially if it is quadratic, it is relatively easy to find the *leading* asymptotics of the coefficients (See [KP], Section 6.5, and the classic [FS]). Alas, for the variance and higher central moments, because of cancellations, we need the more challenging problem of having higher-order asymptotics for these quantities.

The good news is that at least for quadratic algebraic formal power series, this can be done effectively.

3 Pairs of random variables

Often our combinatorial set is equipped with several random variables, and in addition to wanting to find the asymptotics for the individual expectations, means, and higher moments, we'd love to know how they *interact*. If we have two random variables, $R_1(s)$, and $R_2(s)$, then the most important quantity is the *covariance*

$$\mathrm{Cov}(R_1, R_2) := \mathrm{E}[(R_1 - \mu_1)(R_2 - \mu_2)] = \mathrm{E}[R_1 R_2] - \mathrm{E}[R_1]\,\mathrm{E}[R_2] \quad ,$$

and more generally, the *central mixed moments*

$$m_{i,j} := \mathrm{E}[(R_1 - \mu_1)^i (R_2 - \mu_2)^j] \quad ,$$

that by using the binomial theorem can be expressed in terms of the moments $\{\mathrm{E}[R_1^i R_2^j]\}$. Even more important than the covariance is the *correlation* (that is always between -1 and 1):

$$\mathrm{Corr}(R_1, R_2) := \frac{\mathrm{Cov}(R_1, R_2)}{\sqrt{\mathrm{Var}(R_1)\,\mathrm{Var}(R_2)}} \quad .$$

If you define the bi-variate weight-enumerator

$$f(x_1, x_2) := \sum_{s \in S} x_1^{R_1(s)} x_2^{R_2(s)} \quad ,$$

then again all the relevant quantities can be obtained by repeatedly applying $x_1 \frac{d}{dx_1}$ and/or $x_2 \frac{d}{dx_2}$ and at the end of the day plugging-in $x_1 = 1, x_2 = 1$.

Going back to combinatorial families. You have an infinite sequence of related sets, e.g. RNA secondary structures on n vertices, and two natural random variables defined on them, e.g. 'number of hairpins' and 'number of basepairs', and you have the grand-generating function

$$F(X; x_1, x_2) = \sum_{n=0}^{\infty} f_n(x_1, x_2)\, X^n \quad .$$

By applying $(x_1 \frac{d}{dx_1})^i (x_2 \frac{d}{dx_2})^j$, and then plugging-in $x_1 = 1, x_2 = 1$, we get formal powers series in X, with *numerical* coefficients, for the (i, j)-mixed moment, and then and use the algorithms to find higher-order asymptotics for the needed quantities, in particular for the covariance, and from there, the correlation.

4 Let's Keep it Simple: Finding the asymptotics approximately

Once *symbolic* computation gave us explicit expressions for the relevant generating functions, we can get lots and lots of coefficients, and since we know from *general nonsense*, (we are dealing with the *Schützenberger ansatz* (algebraic formal power series)), that the asymptotics, in each case, must be of the form

$$\mu^n n^\theta \cdot (C_0 + \frac{C_1}{n} + \frac{C_2}{n^2} + \dots) \quad ,$$

for *some* numbers μ, θ, and $C_0, C_1, C_2, \dots$.

Both μ and θ are easily found, and even C_0 is relatively easy (see [KP], [FS]), but C_1, C_2, ..., are harder. But by using *numerics* you can do *regression* and get good estimates, that for all practical purposes are *good enough*.

(Please note that this μ has nothing to do with the former μ. Here we bow to two traditions. In statistics, μ denotes the expectation (alias mean, alias average), while in Statistical Physics, μ denotes the so-called *connective constant*, which is the exponential rate of growth. Let us also mention that θ is called the *critical exponent*).

This is implemented in the Maple package `RNAstat.txt` available from:

`http://sites.math.rutgers.edu/~zeilberg/tokhniot/RNAstat.txt` .

For a detailed numeric statistical analysis see the output file:

`http://sites.math.rutgers.edu/~zeilberg/tokhniot/oRNAstat1.txt` .

But it is still nice to get the relevant constants *exactly*, in terms of algebraic numbers. While these can sometimes be *guessed* using, e.g. Maple's *identify* command (based on PSLQ and LLL algorithms), these are not rigorous, and even then, it is often not even possible to get a reasonable guess.

In the next section we will describe how we proved Theorem 3.

5 Finding the Exact Asymptotics

Every algebraic power series satisfies a linear differential equation with polynomial coefficients, and every such differential equation can be easily translated into a linear recurrence equation with polynomial coefficients for the coefficient sequence of the series.

Given such a recurrence equation, it is possible to determine a basis of its solution space in terms of asymptotic series, see [WZ], [K11], [K23] for how this is done in principle, and the examples below for how this is done in practice.

For example, the algebraic power series

$$F(X) = \frac{1}{\sqrt{1-2X}}$$

satisfies the differential equation

$$(1-2X)\,F'(X) - F(X) = 0 \quad ,$$

and its coefficient sequence, f_n, satisfies the recurrence

$$(n+1)\,f_{n+1} - (2n+1)\,f_n = 0 \quad .$$

This recurrence has the asymptotic series solution

$$2^n\,n^{-1/2}\,(1 - \frac{1}{8}n^{-1} + \frac{1}{128}n^{-2} + \frac{5}{1024}\,n^{-3} + \cdots) \quad ,$$

and, since the recurrence has order one, the space of all solutions consists of all constant multiples of this series.

Using the Mathematica package described in [K11], the asymptotic series can be computed as follows:

In[1]:= **Asymptotics[(n + 1)f[n + 1]** − **(2n + 1)f[n], f[n], Order** → **3]**

Out[1]= $\{\frac{2^n\left(1+\frac{5}{1024n^3}+\frac{1}{128n^2}-\frac{1}{8n}\right)}{\sqrt{n}}\}$

In general, an algebraic power series leads to a recurrence of higher order, and in that case, there will be several linearly independent asymptotic series solutions. The program will return a basis of the solution space. Fortunately, in many cases (at least in all cases considered in this paper), one of the series solutions dominates all the others, and the asymptotics of the coefficient series is simply a constant multiple of the series solution with the fastest growth rate.

So for a given algebraic power series $F(X)$ whose coefficient series grows like

$$C\,\mu^n\,n^\theta(1 + \frac{C_1}{n} + \frac{C_2}{n^2} + \cdots) \quad ,$$

we can find μ, θ, and $C_1, C_2, \ldots$ by first translating the algebraic equation into a differential equation, then translating the differential equation into a recurrence, and then computing the dominating asymptotic series solution of this recurrence.

It remains to determine the multiplicative constant C.

For sequences defined by arbitrary linear recurrences with polynomial coefficients, it is not known in general how to identify this constant exactly, but since the coefficient sequences of algebraic power series are more special, this can be done. Following the advice of [FS], we should view the power series as an analytic function rather than merely as a formal object. The asymptotics of the series coefficients at the origin are then determined by the local behaviour of the function near the singularity that is closest to the origin.

More precisely, if ζ is the singularity closest to the origin, and near this singularity (approached from the right direction) we have

$$F(X) \sim A\,(1 - \frac{X}{\zeta})^{k/r}$$

for some rational number k/r (not a positive integer), then the asymptotics of the series coefficients of $F(X)$ is given by

$$f_n \sim \frac{A}{\Gamma(-k/r)}\,\zeta^{-n}\,n^{-1-k/r} \qquad (n \to \infty) \quad .$$

Once more, we refer to [FS] for the full story, and to the example below for how to do things in practice.

In the example discussed earlier, we have $\zeta = 1/2$ and

$$F(X) = \frac{1}{\sqrt{1-2X}} = 1\,(1 - \frac{X}{1/2})^{-1/2} \quad ,$$

which confirms the terms 2^n and $n^{-1/2}$ that we found before, and in addition reveals $C = 1/\Gamma(1/2) = 1/\sqrt{\pi}$. The code of [K11] only handles recurrence equations but not algebraic functions, so we wrote some additional code, available at

`http://sites.math.rutgers.edu/~zeilberg/tokhniot/AlgAsympt.m` ,

for performing this computation, as follows:

In[2]:= **AlgebraicAsymptotics[y^2(1 − 2x) − 1, x, y, 0]**
Out[2]:= $-\dfrac{2^n}{\sqrt{n}\sqrt{\pi}}$.

Every singularity of an algebraic function is either a pole or a branch point, and both can be easily computed from the minimal polynomial of the function. It is also easy to determine which of them is closest to the origin, at least for algebraic functions of degree two.

It is less easy to pick the right branch when determining the first term of the series expansion at the singularity, and since the program happend to pick the wrong one, the result it produced is off by a factor of -1. For equations of degree two, that's the only thing that can go wrong, and if it happens, it is easy to detect and correct manually. For higher degrees, things would become more delicate, but fortunately, we do not need to worry about this issue here because we are only concerned with algebraic functions of degree two.

For obtaining the higher order terms of the asymptotic expansion, our code first uses Koutschan's package [K10] to transform the given algebraic equation into a differential equation, and then the differential equation into a recurrence. Then it uses the package from [K11] to determine the asymptotic expansion, as described above. For example:

In[3]:= **AlgebraicAsymptotics[y^2(1 − 2x) − 1, x, y, 3]**

$$\text{Out[3]}= -\frac{2^n\left(1+\frac{5}{1024n^3}+\frac{1}{128n^2}-\frac{1}{8n}\right)}{\sqrt{n}\sqrt{\pi}}$$

The results reported in earlier sections were obtained in the same way. As all these computations are performed with exact algebraic numbers, they establish a rigorous proof of Theorem 3.

References

[B] AJ Bu, *A Maple package for computing generating functions enumerating RNA secondary structures*. The Maple package is here:
`https://sites.math.rutgers.edu/~zeilberg/tokhniot/RNA.txt` .
The input file `https://sites.math.rutgers.edu/~zeilberg/tokhniot/inRNA1.txt`
produces the following output file:
`https://sites.math.rutgers.edu/~zeilberg/tokhniot/oRNA1.txt`

[CRU] Sang Kwan Choi, Chaiho Rim, and Hwajin Um, *Narayana number, Chebyshev polynomial, and Motzkin path on RNA abstract shapes*, in: *2017 Matrix Annals*, edited by David R. Wood et. al, Springer, 2019.
`https://sites.math.rutgers.edu/~zeilberg/akherim/choi2019.pdf`

[FS] Phillipe Flajolet and Robert Sedgewick, "Analytic Combinatorics", Cambridge University Press, 2009. Freely available on-line:
`https://ac.cs.princeton.edu/home/AC.pdf`

[GD] Shile Gao, and Kequan Ding, *A graphical criterion of planarity for RNA secondary structures with pseudoknots in Rivas-Eddy class*, Theoretical Computer Science **395** (2008), 47-56.
`https://www.sciencedirect.com/science/article/pii/S0304397507006858`

[K10] Christoph Koutschan, *HolonomicFunctions (User's Guide)*, Technical Report 10-01, RISC-Linz. (2010).
`http://www.risc.uni-linz.ac.at/research/combinat/software/HolonomicFunctions/`

[K11] Manuel Kauers, *A Mathematica Package for Computing Asymptotic Expansions of Solutions of P-Finite Recurrence Equations*, Technical Report 11-04, RISC-Linz. (2011).
`https://www3.risc.jku.at/research/combinat/software/ergosum/RISC/Asymptotics.html`

[K23] Manuel Kauers, *D-Finite Functions*, Springer, 2023.
`https://link.springer.com/book/10.1007/978-3-031-34652-1`

[KP] Manuel Kauers and Peter Paule, *The Concrete Tetrahedron*, Springer, 2011.
`https://link.springer.com/book/10.1007/978-3-7091-0445-3`

[Sl] Neil A. J. Sloane, *The On-Line Encyclopedia of Integer Sequences® (OEIS)*, `https://oeis.org/`

[W] Michael S. Waterman, *Secondary structure of single-stranded nucleic acids*, Adv. Math. Supp. Stud. **1** (1978), 167–212. `https://dornsife.usc.edu/msw/wp-content/uploads/sites/236/2023/09/msw-026.pdf`

[WZ] Jet Wimp and Doron Zeilberger, *Resurrecting the Asymptotics of Linear Recurrences*, J. Math. Analysis and Appl. **111** (1985), 162–176. `https://sites.math.rutgers.edu/~zeilberg/mamarimY/WimpZeilberger.pdf`