arXiv:2602.19255v1 [math.CO] 22 Feb 2026

统计 分析 RNA 二级结构中的发夹和碱基对

AJ BU、*Manuel KAUERS* 和 *Doron ZEILBERGER*

谨以此文缅怀 *Kequan Ding*（*1955-2025*）

摘要：我们推导出 RNA 二级结构中发夹数量和碱基对数量的期望、方差、协方差以及许多其他混合矩的精确渐近表达式，并提供有力证据表明，随机变量对（发夹、碱基对）的中心缩放分布趋于相关系数为 $\sqrt{5}\sqrt{5-11/2} = 0.2123322205\ldots$ 的二元正态分布

献辞： 本文谨以深情缅怀 Kequan Ding。他是一位思想深邃、富有原创性且多才多艺的数学家，其众多研究兴趣包括 RNA 二级结构 [GD]。

引言：在一篇开创性论文中， [W], Michael Waterman 提出了一个用于 RNA 二级结构的组合框架。二级结构是一个满足特定条件、具有 n 个顶点的顶点标记图。参见 [W], [DG], 和 [CRU]。这些结构包含若干结构元素，其中最重要的是碱基对 ，另一个则是发夹。有关定义，请参阅 [CRU] （可轻松在线获取）。

以下定理归功于 Waterman [W]:

定理 1：设 $a(n)$ 表示具有 n 个顶点的 RNA 二级结构数量。那么

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

通过符号计算，我们中的一位作者 [B], 证明了以下推广。

定理 2：设 $A(n,i,j)$ 表示具有 n 个顶点，且恰有 i 个发夹和 j 个碱基对的 RNA 二级结构数量。定义三变量生成函数 $F(X;\,x,z)$

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

将其视为关于 X 的形式幂级数，其系数是关于 x 和 z 的多项式。我们得到如下毛茸茸的公式

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

其中

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

在这篇方法论文章中（方法远比实际结果重要！），我们将证明以下定理。

定理 3：对于一个具有 n 个顶点的二级结构 s，设 $X_n(s)$ 和 $Z_n(s)$ 分别表示 s中的发夹数和碱基对数。我们有以下渐近表达式

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

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

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

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

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

更确切地说，我们有：

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

我们有大量证据表明，以下猜想是正确的；我们中的一位（DZ）愿意向 OEIS 捐赠 100 美元，以奖励其严格证明。

猜想：数对 (X_n, Z_n) 的中心化缩放版本，即

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

依分布收敛至协方差为 $c = \frac{\sqrt{5}\sqrt{5}-11}{2}$ 的二元正态分布，其概率密度函数（pdf）为:

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

使用符号计算进行统计学计算的提示

给定一个有限集 S，并赋予其某个随机变量 R, $s \to R(s)$，其期望 μ 自然是

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

其中，照例 $|S|$ 表示 S 中元素的个数。更一般地， R 的 i 阶矩是 i 次幂的平均值:

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

更有用的是**中心矩**

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

其中最重要的是二阶，又称方差。利用二项式定理，知道 μ 后，就可以从原始矩推导出中心矩。

记录定义在有限（甚至无限）集合上的随机变量 R 相关数据的一种简便方式，是使用它的权重枚举式，也就是生成函数。任取一个（形式）变量名，例如 x，并定义:

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

用 $f_R(x)$表示，我们有

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

更一般地， i阶矩为

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

一旦知道了期望值 μ 和方差（也称二阶中心矩） $\sigma^2 = E[(R-\mu)^2]$,（其中 σ 是方差的平方根，称为标准差），就掌握了两个最重要的数值：它的“平均值”以及它的离散程度。不过，了解更多信息也很有帮助，因此可以构造缩放中心分布

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

其期望为 0，方差为 1。它的高阶矩可以用原随机变量 R 的矩表示，因此，了解权重枚举式 $f_R(x)$ 很有用。

在组合学中，我们常常有一个由非负整数自然索引的无限集合族 S_n，以及定义在这些集合上的一个自然随机变量，其权重枚举式序列$f_n(x)$具有良好的形式。例如，对于 $\{1, \ldots, n\}$ 的子集集合，以基数作为“权重”时，我们有闭式表达式 $f_n(x) = (1+x)^n$。如果没有这么幸运，在组合学中，我们往往会发现**总生成函数**

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

在 X 和 x 中有某个显式表达式；或者，在许多组合族中，它满足某个代数方程

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

其中， $\mathcal{P}$ 是其自变量的某个特定多项式。

如果我们有总生成函数的“显式”（甚至“隐式”）表达式，也就是一个关于 X 的形式幂级数，其系数是关于 x 的多项式，那么我们就能得到矩的数值序列的生成函数。

当然，我们还需要处理集合族的基数，我们称之为 $a_0(n)$，其生成函数为

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

定义

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

则第 i 阶矩为 $\frac{a_i(n)}{a_0(n)}$。

遗憾的是，通常无法为 $a_i(n)$ 求得闭式表达式，甚至连 $a_0(n)$ 也不行。不过，我们至少还能退而求其次，求出渐近结果。如果总生成函数是代数的，尤其是二次的，那么求出系数的主导 渐近结果相对容易（参见 [KP], 第 6.5 节，以及经典文献 [FS]）。遗憾的是，对于方差和高阶中心矩，由于存在抵消，我们需要解决更具挑战性的问题：求出这些量的高阶渐近结果。

好消息是，至少对于二次代数形式幂级数，这个问题可以有效地解决。

随机变量对

我们的组合集合通常带有多个随机变量。除了想求出各个变量的期望、均值和高阶矩的渐近结果之外，我们还想知道它们如何相互作用。如果有两个随机变量， $R_1(s)$ 和 $R_2(s)$，那么最重要的量就是协方差

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

更一般地说，是中心混合矩

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

它可以通过二项式定理表示为矩 $\{\mathrm{E}[R_1^i R_2^j]\}$ 的形式。比协方差更重要的是相关系数 （它总是介于 -1 和 1 之间）：

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

如果定义双变量权重枚举生成函数

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

那么，同样可以通过反复应用 $x_1 \frac{d}{dx_1}$ 和/或 $x_2 \frac{d}{dx_2}$，最后代入 $x_1 = 1$， $x_2 = 1$，得到所有相关量。—

回到组合族。我们有一个相关集合的无限序列，例如定义在 n 个顶点上的 RNA 二级结构；还有两个自然定义在这些集合上的随机变量，例如“发夹数量”和“碱基对数量”；以及总生成函数

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

应用 $(x_1 \frac{d}{dx_1})^i (x_2 \frac{d}{dx_2})^j$，再代入 $x_1 = 1$， $x_2 = 1$，我们便得到关于 X 的形式幂级数，其系数为数值，对应于 $(i,\ j)$ 混合矩；然后使用算法求出所需量的高阶渐近，尤其是协方差，并由此得到相关系数。

化繁为简：近似求解渐近

一旦符号计算给出了相关生成函数的显式表达式，我们就能得到大量系数；而且由于我们从一般理论中知道，（我们处理的是

the $Sch\ddot{}$*utzenberger* 方法（代数形式幂级数），每种情况下的渐近形式都必然是

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

对于一些数值 μ、 θ和 C_0、 C_1、 C_2，.....。

μ 和 θ 都很容易找到，甚至 C_0 也相对容易（参见 [KP], [FS]），但 C_1、 C_2、.....则更难。不过，使用数值方法可以进行回归，并得到很好的估计值；就实际用途而言，这些估计已经足够好。

（请注意，这个 μ 与前面提到的 μ毫无关系。这里我们遵循两种传统。在统计学中， μ 表示期望（也称均值、平均值）；而在统计物理学中，μ 表示所谓的连接常数，即指数增长率。还要提一下， θ 称为临界指数。）

此功能已在 Maple 程序包 RNAstat.txt 中实现，可从以下网址获取:

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

有关详细的数值统计分析，请参阅输出文件:

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

不过，若能用代数数的形式精确地得到相关常数，还是很不错的。虽然有时可以通过 Maple 的 $identify$ 命令（基于 PSLQ 和 LLL 算法）来猜测 这些常数，但这种方法并不严谨，而且即便如此，往往也无法得到合理的猜测。

下一节将介绍我们如何证明定理 3。

求精确渐近行为

每个代数幂级数都满足一个多项式系数的线性微分方程，而每个这样的微分方程都可以轻松地转化为一个关于该级数系数序列的多项式系数线性递推方程。

给定这样的递推方程，就可以用渐近级数的形式确定其解空间的一组基，关于具体做法请参见 [WZ], [K11], [K23] 。

例如，代数幂级数

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

满足该微分方程

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

且其系数序列 f_n 满足递推关系

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

该递推关系具有渐近级数解

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

并且，由于该递推关系的阶为一，全部解所构成的空间由该级数的所有常数倍组成。

一般而言，一个代数幂级数会导出更高阶的递推关系；在这种情况下，将存在若干个线性无关的渐近级数解。幸运的是，在许多情形中（至少在本文所考虑的所有情形中），其中一个级数解支配其余所有解，而系数级数的渐近不过是增长率最快的那个级数解的常数倍。

因此，对于给定的代数幂级数 $F(X)$，其系数级数的增长形如

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

我们可以求出 μ、 θ 和 $C_1, C_2, \ldots$，方法是先将代数方程转化为微分方程，再将微分方程转化为递推关系，然后计算该递推关系的主导渐近级数解。

尚需确定乘法常数 C。

对于由任意多项式系数线性递推关系定义的序列，一般来说，我们还不知道如何精确确定这个常数；但由于代数幂级数的系数序列具有更特殊的性质，因此可以做到这一点。按照 [FS]，的建议，我们应将幂级数视为解析函数，而不只是形式对象。于是，级数系数在原点处的渐近行为，便由函数在距离原点最近的奇点附近的局部行为决定。

更确切地说，如果 ζ 是距离原点最近的奇点，并且在这个奇点附近（从右侧方向趋近）我们有

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

对于某个有理数 k/r （不是正整数），那么 $F(X)$ 的级数系数的渐近行为由下式给出

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

关于完整内容，请再次参阅 [FS] 。

在前面讨论的示例中，我们有 $\zeta = 1/2$ 和

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

这证实了我们先前找到的项 2^n 和 $n^{-1/2}$ ，此外还揭示了 $C = 1/\Gamma(1/2) = 1/\sqrt{\pi}$。

代数函数的每个奇点不是极点就是分支点，而且两者都可以根据该函数的最小多项式轻松计算出来。至少对于二次代数函数，很容易确定其中哪一个离原点最近。对于次数更高的函数，情况会复杂一些，因为我们必须确保没有在错误的分支上进行计算。这里无需担心这个问题，因为我们只关注二次代数函数。

确定离原点最近的奇点后，只需计算代数函数在该点处级数展开的首项，即可获得所需数据。所有这些计算都可以使用精确代数数完成，从而为定理 3 建立严格的证明。用于这些计算的 Mathematica 代码如下：

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

参考文献

[B]AJ Bu，用于计算 *RNA* 二级结构计数生成函数的 *Maple* 软件包。Maple 软件包见：
https://sites.math.rutgers.edu/~zeilberg/tokhniot/RNA.txt 。输入文件见：
https://sites.math.rutgers.edu/~zeilberg/tokhniot/inRNA1.txt ，该文件生成以下输出文件：
https://sites.math.rutgers.edu/~zeilberg/tokhniot/oRNA1.txt 。

[CRU] Sang Kwan Choi、Chaiho Rim 和 Hwajin Um，*RNA* 抽象形状中的 *Narayana* 数、切比雪夫多项式和 *Motzkin* 路径，载于：*2017* 年《矩阵年鉴》，David R. Wood 等编，施普林格，2019 年。

https://sites.math.rutgers.edu/~zeilberg/akherim/choi2019.pdf .

[FS]Phillipe Flajolet 与 Robert Sedgewick，《解析组合学》，剑桥大学出版社，2009。可在线免费获取：
https://ac.cs.princeton.edu/home/AC.pdf 。

[GD] Shile Gao 与 Kequan Ding，*Rivas-Eddy* 类中带假结的 *RNA* 二级结构平面性的图形准则，《理论计算机科学》**395**（2008），47-56。
`https://www.sciencedirect.com/science/article/pii/S0304397507006858` 。

[K11] Manuel Kauers，用于计算 *P*-有限递推方程解的渐近展开的 *Mathematica* 程序包，技术报告 11-04，RISC-Linz。（2011）。

[K23]Manuel Kauers, 《*D*-有限函数》，施普林格，2023年。[KP] Manuel Kauers 和 Peter Paule, 《具体四面体》，施普林格，2011年。[Sl] Neil A. J. Sloane, 《在线整数数列百科全书（*OEIS*）》，https://oeis.org/。[W]M.S. Waterman, 《单链核酸的二级结构》，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 和 Doron Zeilberger, 《重现线性递推的渐近行为》，J. Math. Analysis and Appl. 111 (1985), 162–176. https://sites.math.rutgers.edu/~zeilberg/mamarimY/WimpZeilberger.pdf.

AJ Bu，罗格斯大学数学系（新不伦瑞克），HillCenter-Busch 校区，110 Frelinghuysen Rd.，皮斯卡塔韦，NJ 08854-8019，美国。电子邮箱：ab1854 at math dot rutgers dot edu

Manuel Kauers，代数研究所，约翰内斯·开普勒林茨大学¨，Altenbergerstraße 69，A-4040 林茨，奥地利。电子邮件：manuel dot kauers at jku dot at

Doron ZEILBERGER，罗格斯大学（新不伦瑞克）数学系，Hill Center–Busch 校区，110 Frelinghuysen Rd.，皮斯卡塔韦，NJ 08854-8019，美国。电子邮件：ab1854 at math dot rutgers dot edu

撰写日期：**2026年2月20日**。