

如何通过实验的方法检验单分子混合物假说

唐宇

中国科学院上海有机化学研究所 生命过程小分子调控全国重点实验室

Email: tangyu@sioc.ac.cn

ORCID ID: [0000-0002-4272-2234](https://orcid.org/0000-0002-4272-2234)

摘要: 如何通过实验方法验证“单分子混合物”这一假说? 是单分子混合物科学发展过程中不可避免的一个关键问题, 本文提出了一种通过实验检验来推断一个分子聚集体样品有多大的可能性呈现“绝大多数分子的结构都各不相同”的“近单分子混合态”的方法, 以及如何通过实验检验呈现“单分子混合物”状态的分子聚集体与相应的纯态分子聚集体的性质差异的具体方法, 为解决这一问题提供了可能的方案。

关键词: 单分子混合物, 单分子混合态, 实验检验方法

化学科学史上一些著名假说或理论, 例如原子-分子假说、碳四面体结构假说、聚合物假说、非经典碳正离子假说等, 在最初提出的时候, 都难以通过确凿的实验证据获得证实, 以至于在很长的时间里引起了激烈的争论和质疑, 然而, 随着研究工具的发展, 最终它们获得了证实, 并展现出了巨大的价值, 成为现代化学学科的基础理论。具体到聚合物假说而言, 虽然已经提出超过一个世纪, 并已获得极为广泛的应用, 但直到最近, 聚合物假说才通过直接的实验观测方法获得真正的证实¹。

复杂研究对象, 由于缺乏合适的实验研究工具, 难以得到有效研究, 这一情形在当今的科学界普遍存在。在研究的初始阶段, 首先采用数学建模-数学分析的方法对一类新的研究对象进行理论上的研究, 是一种有效的途径。对于“单分子混合物”这一极为复杂的全新化学研究对象^{2,3}, 采用数学的方法进行理论研究是我们目前可以做到的。尤其是, 统计学、概率论思想对于理解这一类化学体系帮助极大。随着概念的确立和研究的深入, 一个不可避免关键问题出现了: 如何通过实验方法验证“单分子混合态”这一假说?

这一问题又可分为两个方面: 1) 如何通过实验的方法检验一个分子聚集体样品中的分子结构都各不相同? 2) 如何通过实验的方法检验呈现“单分子混合物”状态的分子聚集体与相应的纯态分子聚集体的性质差异? 下面对这两个问题分别进行讨论。

1. 如何通过实验的方法检验一个分子聚集体样品中的分子结构都各不相同?

与“纯态”只是一个理想状态, 无法实际获得一样, 绝对的“单分子混合态”也很可能是一个理想状态, 只能无限接近, 无法实际达到, 但在不远的未来我们确实有可能通过实验检验的方法推断一个分子聚集体样品有多大的可能性呈现“绝大多数分子的结构都各不相同”的“近单分子混合态”。

对于包含巨大数目分子的分子聚集体而言，想要证明其中的绝大部分分子的结构各不相同，初看起来似乎是一个“几乎不可能”完成的任务，然而，采用数学建模和概率论的思维方式，我们可以通过随机取样检测的方法推测样品中分子呈现“绝大部分分子的结构都不相同”的“近单分子混合态”的概率。具体方法如下。

考虑如下数学模型：有一堆小球，每个小球有一个编号。编号可能重复，重复的编号表示多个球有同一个编号。现在我们从这堆小球里随机抽取一定量的小球，发现它们的编号各不同（即没有重复的编号）。在这一观察结果基础上，我们要推断：至少有 99% 的球，其编号在整个堆中是唯一的，并希望这个推断正确的概率大于或等于 99%。问：最少抽取多少小球？

在 DeepSeek 辅助下，我对上述模型进行了计算，下面是详细的计算过程。

设总体中有 N 个球（ N 很大），每个球有一个编号，编号可能重复。记

- U 为仅出现一次的编号对应的球的总数（即唯一球数），
- $R = N - U$ 为至少出现两次的编号对应的球的总数（重复球数）。

从总体中随机不放回地抽取 n 个球，观察到它们的编号各不相同。

要求：基于该观察，推断 $U \geq 0.99N$ （即 $R \leq 0.01N$ ）这一结论正确的概率至少为 99%。

求满足条件的最小 n 。

1. 统计推断框架

这是一个假设检验与置信推断相结合的问题。

考虑原假设

$$H_0: R > 0.01N$$

若在原假设下，观察到“ n 个球编号各不相同”的概率 $P(A | H_0)$ 很小，则当实际观察到 A 时，可以拒绝 H_0 ，从而接受

$$H_1: R \leq 0.01N$$

要求犯错误的概率（即 H_0 为真却拒绝它）不超过 1%。

因此需要：

$$\max_{R > 0.01N} P(n \text{ 个球编号各不相同} | R) \leq 0.01$$

在所有的 $R > 0.01N$ 的总体配置中，取使得 $P(A | R)$ 最大的情形，若该最大概率 ≤ 0.01 ，则对一切 $R > 0.01N$ 均成立。

2. 最坏情况构造

固定 R ，求 $P(A | R)$ 的上界。

编号各不相同的必要条件是：抽到的 n 个球来自 n 个不同的编号。

为了使 $P(A | R)$ 尽可能大，应让重复编号中的球在抽样中尽可能不造成编号重复。

这要求对每个重复编号，一次抽样中不会抽到其两个球。

在给定 R 下，最有利于 $P(A | R)$ 的配置是：

- 所有重复球集中在尽可能少的编号上,且每个这样的编号恰有 2 个球(因为若多于 2, 同一编号被抽到至少 2 个的概率更大,从而降低 $P(A)$)。

设 $r = R/2$ 为双球编号的个数(假定 R 为偶数,不影响渐近分析)。

则总体中有

- $u = N - R$ 个唯一球,
- $r = R/2$ 个双球编号, 每个有 2 个球。

总编号种类数为 $u + r$ 。

3. 概率表达式

在随机无放回抽样中, 事件 A 的概率为:

从 u 个唯一编号与 r 个双球编号中, 选出 n 个不同的编号种类,

并且对每个选中的双球编号, 从该编号的 2 个球中任选 1 个。

因此

$$P(A) = \frac{\sum_{i=0}^{\min(n,r)} \binom{r}{i} \binom{u}{n-i} 2^i}{\binom{N}{n}}$$

其中 i 是从双球编号中选中的种类数, $n - i$ 是从唯一编号中选中的种类数。

为求上界, 考虑最坏情况 $R = 0.01N$ (最大可能重复比例), 则

$$u = 0.99N, \quad r = 0.005N.$$

4. 简化与不等式

由于 N 很大, 可用近似:

$$\binom{u}{n-i} \approx \frac{u^{n-i}}{(n-i)!}, \quad \binom{r}{i} \approx \frac{r^i}{i!}, \quad \binom{N}{n} \approx \frac{N^n}{n!}$$

于是

$$P(A) \approx \frac{\sum_{i=0}^n \frac{r^i}{i!} \cdot \frac{u^{n-i}}{(n-i)!} \cdot 2^i}{\frac{N^n}{n!}}$$

乘以 $n!/N^n$ 得

$$P(A) \approx \sum_{i=0}^n \binom{n}{i} \left(\frac{2r}{N}\right)^i \left(\frac{u}{N}\right)^{n-i}$$

即

$$P(A) \approx \left(\frac{u}{N} + \frac{2r}{N}\right)^n = \left(\frac{u + 2r}{N}\right)^n$$

代入 $u = 0.99N$, $2r = 0.01N$, 得

$$P(A) \approx (0.99 + 0.01)^n = 1$$

此近似过粗, 因为它忽略了“不能同时选中双球编号中的两个球”这一约束。

更精确地, 对于每个双球编号, 在一次抽样中其两个球不被同时选中的概率近似

为

$$1 - \binom{n}{2} \cdot \frac{1}{N^2} \quad (\text{对一个大数据号})$$

对所有 r 个双球编号联合, 得

$$\begin{aligned} P(A) &\approx \left(\frac{u}{N}\right)^n \cdot \left(1 - \frac{\binom{n}{2}}{N^2}\right)^r \\ &\approx (0.99)^n \cdot \left(1 - \frac{n(n-1)}{2N^2}\right)^{0.005N} \end{aligned}$$

当 N 很大时, 第二项 $\rightarrow 1$, 因此

$$P(A) \lesssim (0.99)^n$$

5. 临界不等式

为保证对一切 $R > 0.01N$ 有 $P(A) \leq 0.01$, 只需

$$(0.99)^n \leq 0.01$$

取对数:

$$n \ln 0.99 \leq \ln 0.01$$

$$n \geq \frac{\ln 0.01}{\ln 0.99}$$

计算:

$$\ln 0.01 = -4.605170, \quad \ln 0.99 \approx -0.0100503$$

$$n \geq \frac{-4.605170}{-0.0100503} \approx 458.2$$

因此最小的整数 n 为 459。

但考虑离散性与实际抽样中不放回造成的概率略小于 $(0.99)^n$, 为保证严格满足 0.01 概率界, 通常取

$$n = 460$$

由此我们可以得出结论, 当从总体中随机不放回地抽取 $n = 460$ 个球, 并观察到它们的编号全部不同时, 可以以至少 99% 的置信度推断: 总体中至少有 99% 的球的编号是唯一的。

从这一模型可以看到, 对应于一个包含极大数目分子的分子聚集体而言, 从中随机抽取单个分子进行“单分子结构鉴定”, 记录其结构, 重复 460 次测试, 得到 460 个结构, 如果观察到这些结构各不相同, 那么我们就可以以至少 99% 的置信度推断: 这一分子聚集体中至少有 99% 的分子的结构是唯一的。由此, 我们就得到了一种通过实验检验来推断一个分子聚集体样品有多大的可能性呈现“绝大多数分子的结构都各不相同”的“近单分子混合态”的方法, 这为解决“如何通过实验方法检验单分子混合态假说”这一问题提供了一种可能的方案。

值得一提的是, 通过大量采样观察来理解一个重要化学研究体系的完整规律的方法在化学领域是一个获得认可的极为有效的方法, 最近的 Kotov 等人报道的通过大量采样观察的方法研究纳米颗粒的演化规律的工作就是一个很好的例子⁴。

虽然目前直接检测单分子精确结构的技术尚不成熟,但单分子结构检测及单分子测序技术都已经有了很大的发展^{5,6},在不远的将来,聚合物单分子结构的精确检测很可能会取得突破,届时,“单分子混合态”假说将有望通过实验方法获得验证,从而为单分子混合物科学的发展提供坚实的理论基础。

2. 如何通过实验的方法检验呈现“单分子混合物”状态的分子聚集体与相应的纯态分子聚集体的性质差异?

要研究这一问题,首先需要建立一个合适的模型体系,这个模型体系最好来源于现实中的聚合物体系,合成方法及性质研究方法都比较成熟,便于实验的进行。经过详细的文献调研,我发现全乙酰化纤维素多糖是一个合适的模型体系,下面对该模型体系进行详细的介绍。

三乙酰化纤维素⁷⁻⁹,有成熟的合成方法,广泛而重要的用途,对其物理性质的研究也很充分,是极好的单分子混合物假说测试的“原型材料”,然而,天然来源的纤维素制成的三乙酰纤维素的糖链长度不均一,无法作为底物直接用于测试本文所提出的“等分子量单分子混合物”,以之为原型,我设计了一个基于全己酰化纤维 128 聚糖的单分子混合物模型,如图 1 所示。

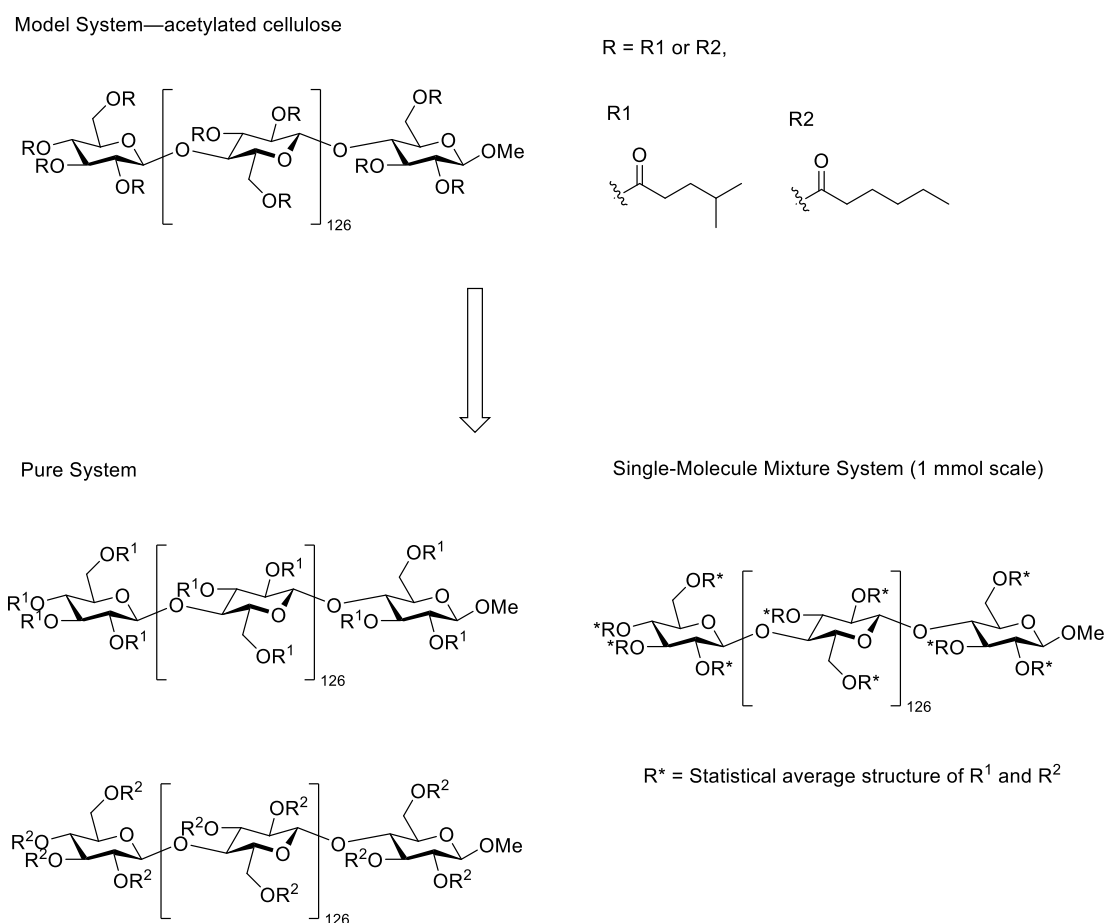


图 1. 一个基于全己酰化纤维 128 聚糖的单分子混合物模型

首先,采用俞飏课题组发展的迭代指数增长策略来高效组装 128 聚糖¹⁰,得到无保护的纤维 128 聚糖,以此为原料,进行全己酰化反应(如图中 R¹, R² 所

示的两种己酰基), 制备不同取代类型的对应于纯态和单分子混合物状态的两种体系。在两种纯态结构中, 糖上的酰基为单一的正己酰基(R^1)或异己酰基(R^2), 在单分子混合物态中, 糖上的酰基为统计平均分布的两种己酰基(R^*), 按照我之前提出的计算方法可知, 在 1 mmol 的制备规模下, 该分子聚集体呈现典型的“单分子混合物”状态。

在解决了模型的构建和样品的制备问题后, 我们接下来可以详细考察“纯态”和“单分子混合态”两种形态的分子聚集体的性质差异。可以考察的物理性质类型包括: 熔点、溶解度、旋光度、折光率、结晶行为等, 在实际应用方面, 包含统计平均结构的 R^* 的单分子混合物形态的全己酰化纤维素, 有可能呈现出独特的性质, 应用于手性膜分离等领域。

通过这一例子, 我详细展示了如何通过实验的方法检验单分子混合态分相较于纯态分子聚集体的独特性质。事实上, 作为一种全新形态的分子聚集体, 一旦通过实验揭示出其不同寻常的独特性质, 就有很大希望将其应用于广泛的领域, 实现多种多样的功能, 我甚至可以大胆预测, 单分子混合物未来的应用前景不亚于目前最前沿的功能材料如 MOF, 碳纳米管, 分子器件等。

期待这一天的早日到来!

声明: 在本研究公式推导的准备过程中, 作者使用了 DeepSeek (DeepSeek-V4 版本) 辅助完成数学推导验证和符号规范化。使用时间为 2026 年 5 月 19 日, 输入的关键提示包括: “有一堆小球, 每个小球有一个编号。编号可能重复, 重复的编号表示多个球有同一个编号。现在我们从这堆小球里随机抽取一定量的小球, 发现它们的编号各不同 (即没有重复的编号)。在这一观察结果基础上, 我们要推断: 至少有 99% 的球, 其编号在整个堆中是唯一的, 并希望这个推断正确的概率大于或等于 99%。问: 最少抽取多少小球? ”。作者已对 AI 生成的全部推导内容进行了逐行审查与验证, 并对最终论文内容的准确性承担全部责任。

参考文献

1. Liu, Y.; Vancso, G. J. Polymer single chain imaging, molecular forces, and nanoscale processes by Atomic Force Microscopy: The ultimate proof of the macromolecular hypothesis. *Prog. Polym. Sci.* **2020**, *104*, 101232
2. Tang, Y. Single-Molecule Mixture: A Concept in Polymer Science. *Int. J. Mol. Sci.* **2024**, *25*, 7571.
3. Tang, Y. Micro-Nonuniformity and Single-Molecule Mixture Science: An Introduction. *Preprints* **2026**, 2026041387.
4. Hallstrom, J.; Pan, P.; Sia, J.; Bae, S.; Qian, D.; Qian, C.; Liu, S.; Yao, L.; Truskett, T. M.; Milliron, D. J.; Chen, Q.; Mao, X.; Bogdan, P.; Kotov, N. A. Decoding

collective dynamics and complexity in nanoparticle assemblies using graph theory. *Science*, 2026, 392, eaeb5134. DOI:10.1126/science.aeb5134

5. Ying, YL., Yang, CN., Liu, W. *et al.* Understanding single-molecule reactions using nanopore-based techniques. *Nat. Chem.* 2025, **17**, 1450–1461.
6. Ren, M.; Qin, F.; Liu, Y.; Liu, D.; Lopes, R. P.; Astruc, D.; Liang, L. Single-molecule resolution of the conformation of polymers and dendrimers with solid-state nanopores. *Talanta*, **2025**, 286, 127544
7. Wallner, G.; Hausner, R.; Hegedys, H.; Schobermayr, H.; Lang, R. Application Demonstration and Performance of a Cellulose Triacetate Polymer Film based Transparent Insulation Wall Heating System. *Sol. Energy* **2006**, 80, 1410–1416.
8. Lee, J.; Nguyen, D.; Lee, S.; Kim, H.; Ahn, B.; Lee, H.; Kim, H. Cellulose Triacetate-based Polymer Gel Electrolytes. *J. Appl. Polym. Sci.* **2010**, 115, 32–36.
9. Selvakumar, M.; Bhat, D. LiClO₄ doped Cellulose Acetate as Biodegradable Polymer Electrolyte for Supercapacitors. *J. Appl. Polym. Sci.* **2008**, 110, 594–602.
10. Zhu, Q.; Shen, Z.; Chiodo, F.; Nicolardi, S.; Molinaro, A.; Silipo, A.; Yu, B. Chemical synthesis of glycans up to a 128-mer relevant to the *O*-antigen of *Bacteroides vulgatus*. *Nat Commun* **2020**, 11, 4142.