DOI: 10.13376/j.cbls/2014034
文章编号：1004-0374(2014)03-0219-09

鲁志 教授

鲁志实验室的研究涉及生物信息学和基因组学领域，主要研究生命信息是如何被编码在 DNA 和 RNA 的结构和序列中，以及它们如何来调控整个生物系统。本课题组致力于应用已有的分析手段和开发新的算法，以发现和诠释新的非编码 RNA 基因的结构和功能；运用高通量测序技术在基因组层面上去挖掘关于结构、调控等方面的新知识；最终，我们希望把科研成果运用到人类疾病的研究和治疗中。

实验室尤其侧重植物基因组和癌症基因组的相关研究。主要以生物信息学的方法，结合最新一代的大规模测序技术，研究植物基因组、人类基因组和模式生物基因组里的新型非编码基因的结构特征。

非编码RNA的生物信息学研究方法：RNA结构预测及其应用

张浩文，杨禹丞，鲁　志*
（清华大学生命科学学院生物信息学教育部重点实验室，北京 100084）

摘　要：近 10 多年来的研究逐步揭示了 RNA 的各种生物学功能。RNA 不仅是信息从 DNA 传递到蛋白质的中间体，还直接参与基因沉默、表观遗传学修饰等生物学过程。单链的 RNA 在体内通过碱基配对折叠成一定的二级结构。介绍了现在预测 RNA 二级结构的主要算法及其应用，其中包括基于热力学、同源比对和统计学习的各种算法，以及如何引入实验数据辅助预测。二级结构预测算法被广泛用于寻找 RNA 功能单元和预测新非编码 RNA 等各种问题。如何利用高通量实验数据帮助结构预测，探索长非编码 RNA 功能，研究 RNA 与蛋白质相互作用，是 RNA 二级结构预测算法和应用的一些前沿方向。

关键词：RNA 二级结构预测；自由能；共突变；非编码 RNA
中图分类号：Q811.4；Q52　　**文献标志码**：A

From sequence to structure: RNA secondary structure prediction methods and the applications

ZHANG Hao-Wen, YANG Yu-Cheng, LU Zhi*
(MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China)

Abstract: Researches in past 10 years uncovered various biological functions of RNA. Besides as the intermediate to transfer information from DNA to protein, RNA has been shown involved in gene silencing, epigenetic modifications and many other processes. RNAs are transcribed as single strand, fold into secondary structures by

收稿日期：2013-10-12
基金项目：国家自然科学基金青年基金项目(31100601)；国家自然科学基金面上项目(31271402)；国家重点基础研究发展计划(“973”项目)(2012CB316503)；国家高技术研究发展计划(“863”项目)(2014AA021100)；IBM基金；Bayer基金；罗氏基金
***通信作者**：E-mail: zhilu@tsinghua.edu.cn

base pairing. Here different algorithms of RNA secondary structure prediction and their applications are reviewed, including thermodynamic algorithms, homologous alignments comparison, statistical learning and incorporating experimental data into prediction model. RNA secondary structure prediction algorithms are widely used in detection of RNA functional elements, identification of new non-coding RNAs and other researches. Challenges such as determination of secondary structures transcriptome wide under the help of high-throughput sequencing, assist to long non-coding RNA function discoveries and study of protein-RNA interactions will draw more attentions in this field.

Key words: RNA secondary structure prediction; free energy; co-variation; noncoding RNA

细胞环境中 RNA 常折叠形成二级或 3D 结构，影响 RNA 的降解 [1]、翻译起始 [2] 等。除了编码蛋白质的 mRNA，细胞中还有很多非编码 RNA (non-coding RNA, ncRNA) 不翻译成蛋白质，但仍行使各种生物功能。有研究表明，基因组有超过 93% 的部分在不同细胞中表达 [3]。非编码 RNA 的功能与其结构密切相关 [4]。正是由于 RNA 结构有着重要的生物学意义，研究人员希望得到准确的 RNA 结构。相比于获取一级序列，实验手段 (X 射线晶体衍射或磁共振) 确定 RNA 二级和 3D 结构不仅花费高、难度大，而且对很多 RNA 分子并不可行，这导致了已知的一级序列信息与二级、3D 结构信息之间形成了巨大的知识鸿沟。有鉴于此，过去 30 年中，生物信息学家发展出各种预测 RNA 结构的方法。

1 二级结构预测算法

RNA 二级结构由碱基配对定义 (图 1A)，除主要是沃森 - 克里克配对 (A-U，G-C) 外，还会有一些非常规配对 (G-U，A-A)。RNA 二级结构预测算法 (表 1) 可大致分为热力学模型、同源比对模型和统计学习模型。以下将分别对 3 种 RNA 二级结构预测算法进行综述。

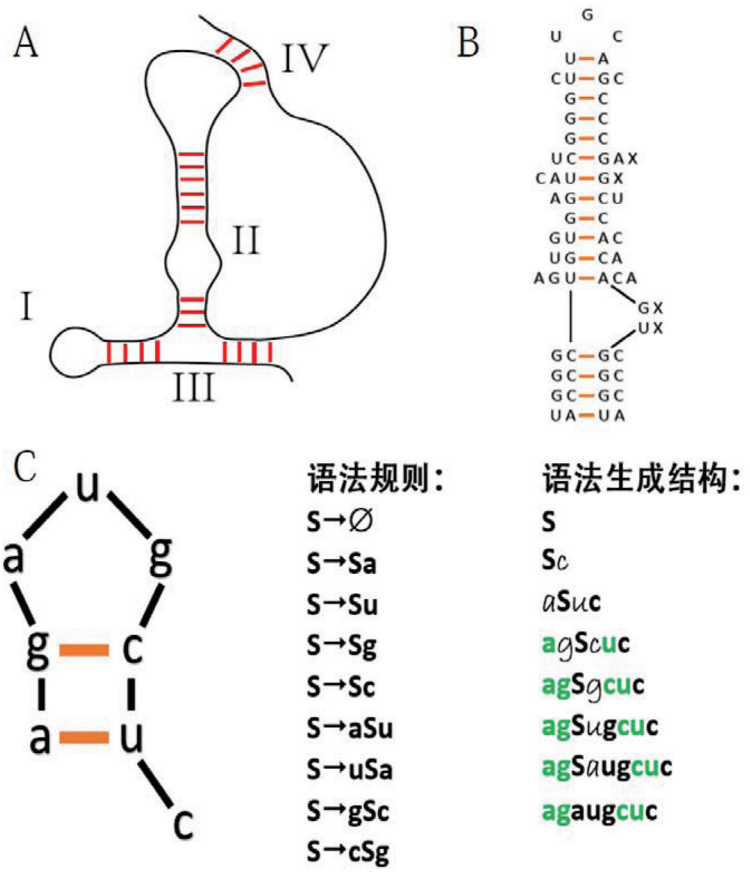
1.1 热力学模型

使用最小自由能方法根据 RNA 序列预测二级结构是相当常用的手段。对于 700 bp 以下的 RNA，已知的碱基对中有 40%~70% 可以被正确预测 [5]。假设 RNA 分子服从热力学规律，则自由能较低的结构出现的概率更大：

$U \leftrightarrow F$，$[F]/[U] = e^{-\Delta G/RT}$

此处，ΔG 为 RNA 从解链状态折叠到特定结构的自由能变，我们可以利用计算机搜索得到自由能最低的结构，作为预测结果。通过组合数学可以估计出，一段长度为 n 个碱基的 RNA 分子可以折叠出 1.8^n 种可能的二级结构 [6]；对长度仅为 100 个碱基的 RNA，这个数字就超过了 10^{25}，穷举所有可能的结构远远超出了现今计算机的计算能力。为了在合理的时间内获得结果，搜索最小自由能结构时会用到最近邻假设，并结合动态规划算法降低搜索复杂度。在最近邻假设中，若两个碱基互相配对，它们之间的序列会自行折叠成自由能最低的结构，不再受外部序列的影响。早期的最小化自由能算法较为粗糙，即简单假设自由能与氢键数目呈负相关，最大化氢键数目从而预测最小自由能结构 [6]。之后的研究对自由能的计算式和参数不断改进，准确率也逐渐提高。Turner2004 是当前一组广泛使用的热力学最近邻模型参数 [7]。此外，动态规划算法不仅能找到自由能最低的“最优”结构，还可以根据需要返回自由能较低的“次优结构”[8]，而且最小自由能方法还允许人为加入一些限制条件，把不合理的结构排除在搜索范围之外 [8]。这些方法可以辅助我们克服模型和参数误差带来的错误。

无论是在细胞环境还是溶液环境中，RNA 的折叠应该是一个动态过程。我们也希望对一条 RNA 所有可能的结构进行汇总，得到任意两个碱基配对的概率，或是一个局部结构出现的概率。在热力学中，可以通过计算一类结构的配分函数，除以系统的配分函数，估计出这类结构在系统中出现的概率。1990 年，McCaskill 等 [9] 提出了使用动态规划算法快速计算配分函数，使计算 RNA 中每两个碱基的配对概率成为了可能。我们可以从 3 个角度利用这样的概率估计。第一，可以根据这个概率，对最小自由能或其他方法预测出的结构进行标注，从而知道这个结构中哪一部分基本准确，哪一部分不太确定。Mathews[10] 的工作显示，利用配分函数计算出的高概率 (>90%) 的碱基对，有很高的可能性 (约 83%) 存在于真实结构中。第二，利用配分函数计算出的碱基配对概率可以被用来计算结构。简单假设一个结构与配分函数计算出的概率越符合，则这个结构就越接近真实结构。MEA 算法利用动态规



A：二级结构示意图。图中罗马数字表示二级结构中不同的motif，分别是，I：茎环结构(stem-loop)；II：内部环(internal-loop)，III：多分枝内部环(multibranch-loop)；IV：假结(pseudoknot)。B：RNA同源序列配对碱基共突变。结构来自于R2 retrotransposon[38]。C：语言学上下文无关语法(context free grammar, CFG)示意图。此语法生成规则允许字符向两端延伸，即N→αMβ，α、β可以是零个、一个或连续几个碱基，一个语法包含一系列生成规则。一个语法表征一类RNA结构，图中是可以表征所有RNA二级结构的一个语法，手写体表示模型新生成碱基，绿色表示配对碱基。语言学模型详见教材Biological sequence analysis[39]。

图1 二级结构及基本算法图示

表1 常用的RNA二级结构预测软件

软件包名	网址	执行算法	参考文献
Vienna RNA Package	http://rna.tbi.univie.ac.at/	最小化自由能，计算配分函数和估计碱基配对概率，优化期望准确度，计算重心结构，同源序列结构预测	[8]
RNAstructure	http://rna.urmc.rochester.edu/RNAstructure.html	最小化自由能，计算配分函数和估计碱基配对概率，随机取样结构，优化期望准确度，同源序列结构预测	[34]
CONTRAfold	http://contra.stanford.edu/contrafold/	语言学模型，服务器端已经完成训练，可直接预测结构	[19]
Foldalign/FoldalignM	http://foldalign.ku.dk/software/index.html	自动比对序列和预测结构	[35]
CentroidFold	http://www.ncrna.org/centroidfold/	单序列或同源序列计算重心结构	[36]

表内软件均支持服务器端结构预测，算法和软件选取可参考引文[37]。

划去优化期望准确度 (maximize expected accuracy) 预测结构 [11]。与真实结构比对发现，MEA 算法获得的结构的准确率略高过最小自由能算法获得的结构。第三，依照概率对给定 RNA 的可能结构进行取样 [12]。取样的优势使我们可以很方便地对这些样本进行后续分析，如通过对取样的结构进行聚类，

分析结构可能会有的亚型。为了表述方便，与聚类中所有其他结构距离之和最小的一个结构被定义为重心结构 (centroid structure)，代表其所在的聚类。最大聚类的重心结构通常十分接近真实结构，有预测算法寻找最大聚类重心从而预测结构 [13]。

1.2 同源比对模型

RNA 的生物功能与其结构密切相关，重要的 RNA 结构会在进化中体现出保守性。在结构保守的位置，配对碱基的序列突变会呈现出相关性，如在不影响配对的情况下同时发生突变 (如 G-C 突变成 A-U 或 C-G)(图 1B)。因此，通过从同源序列中寻找共突变的碱基对，我们可以进行结构预测。相比于热力学模型，同源比对方法的一个优势是直接反映出 RNA 在细胞中的结构状态，但需要额外提供同源序列作为输入。同源比对方法可以进一步划分为人为指导和自动比对。人为指导获得的结构准确率极高，在核糖体 ITS (internal transcribed spacers) 结构的预测中，所预测的碱基对有 95% 的置信 [14]。自动比对预测包括先比对后折叠、先折叠后比对和边折叠边比对。在 RNAalifold 中，序列比对直接获得的多重序列被看成一个碱基位有多种赋值的 RNA 序列，在碱基形成配对时，所有的赋值都参与自由能贡献 [15]。利用修改后的自由能规则，仍然使用最小自由能算法，可以有效地对同源序列进行结构预测，这种先比对后折叠的算法速度快，近似等同于折叠一条序列所需的时间。TurboFold 则是边折叠边比对，通过迭代，利用同源序列不断修正碱基配对概率 [16]。M 条同源序列中的每条序列在每次迭代时都被用配分函数计算了一次碱基配对概率，其他序列的配对概率会对当前序列配分函数的计算进行修正。这一方法的优势是并不要求每条同源序列的结构都严格保守，同源序列的选取更加灵活。上述两算法均由成熟的程序实现并易于使用。

1.3 统计学习模型

现在，已知结构的 RNA 越来越多。RNA STRAND 收集了可信的 4 666 个 RNA 序列和二级结构，这些二级结构来自晶体衍射、磁共振或 RNA 同源序列比对 [17]。统计学习模型从已知的结构进行学习，从而进行结构预测。统计学习模型在预测准确率上达到或超过最小自由能法，但统计学习模型不可避免会对训练数据集过学习，即对训练数据集和相似数据集的预测效果超过普遍预测效果，因此，普遍情况下统计学习模型的准确率仍待确定。利用训练数据集帮助预测主要分为两种思路。第一，假设 RNA 折叠仍旧服从热力学规律，通过数据集对热力学参数进行重新估计。Andronescu 等 [18] 的工作用到了两种不同的优化目标：Constraint Generation，在限制已知结构是最小自由能结构的条件下，使 RNA 折叠的自由能变与实验测定尽可能接近；Boltzmann Likelihood，通过 Boltzmann 分布，估计数据库中每条 RNA 的已知结构出现的概率，并最大化似然函数 (所有真实结构出现概率相乘取对数)。优化后的两组热力学参数的预测性能在测试集上均超过了优化前的热力学参数。第二，不再假设 RNA 折叠服从热力学规律，而是认为真实结构来自概率模型的一个采样，通过最大化似然函数估计模型参数。在最近邻假设下，语言学中的随机上下文无关语法 (SCFG) 非常适合作为描述 RNA 结构的概率模型 (图 1C)。在 CONTRAFold 这一工作中，使用了类似于 SCFG 的语言学模型，预测性能在其测试集上也超过了最小自由能方法 [19]。

1.4 引入实验数据的二级结构预测

除了晶体衍射和磁共振，化学标记 (chemical mapping) 也是重要的检测 RNA 二级结构的实验手段。使用这种方法时，待测定结构的 RNA 用小分子化学物质或是 RNA 酶处理，包括 DMS、CMCT、S1 nuclease、RNase V1 等，它们倾向性地标记或剪切单链 (或双链) RNA，从而达到对结构的标记和检测 [4]。不同于晶体衍射，化学标记实验容易操作，但准确度不够，这就对后续的数据处理和预测提出了要求。实验数据可用作限制条件，对热力学模型或是概率模型进行修正。最直接的做法是，根据实验数据确定出一些配对的碱基或是单链碱基，将它们硬性限制住，再用最小自由能算法预测结构 [20]，但从实验数据获取确信的结构信息并不简单。另一种方法是将 SHAPE 实验数据作为非硬性限制条件加入到自由能的计算式中 [5]，预测最小自由能结构。利用这种方法，并结合进化信息，已成功预测出 HIV-1 的基因组结构 [21]。另一种策略对概率模型进行修正，或是进行大量结构采样，或是对碱基配对概率进行修正，两者都会利用到配分函数。SeqFold 根据 Boltzmann 分布取样出大量结构，聚类后，选取与实验数据最接近的聚类重心作为预测结果 [22]。RNApbfold 则假设 RNA 折叠中真实自由能变是自由能模型加上一个扰动，这个扰动应该尽可能的小，而且扰动后重新估计的碱基配对概率需与实验结果尽可能的接近 [23]。当然，此类算法的效果本身会受实验数据质量的影响。

2　二级结构预测的应用

2.1　预测功能单元

RNA 的二级结构广泛影响各类 RNA 的各种生物学过程。行使特定功能的蛋白质及一些 RNA 也会特异性结合有结构特征的单元，执行其生物学功能 [4]。寻找、预测这些结构单元的工作经常会用到结构预测的方法，部分原因是 DNA 及 RNA 的序列十分容易获取。按照问题的不同，本文把寻找功能单元的工作进一步分为两类。第一类，已知功能单元应该具有的一些结构特征，希望通过这些特征进行搜索。siRNA 的预测是典型的例子。mRNA 的 UTR 的某一局部折叠成稳定的二级结构会影响 siRNA 的结合。OligoWalk 算法预测这些结构解链的自由能，作为特征之一，再预测 siRNA 沉默 mRNA 的效果，达到 78.6% 的有效概率 [24]。第二类，有一些序列被确定与某种生物功能相关，希望从中找出一些具有保守结构的单元，但对功能单元本身并不了解。RNApromo 通过计算手段，从快速降解、慢速降解的 mRNA 中寻找结构特征。首先，对 UTR 区域进行结构预测，枚举出大量局部特征。这些特征按照所包含的序列数进行排序，选取包含序列最多的几个特征，分别用训练语言学规则 (SCFG) 表征这些特征，最终得到的这些语言学规则就是结构特征 [25]。

2.2　预测新非编码RNA

二级结构预测也被广泛应用在寻找新的非编码 RNA。在假设基因组随机序列比非编码 RNA 的折叠自由能变更低的条件下，我们可以利用二级结构预测算法搜索基因组，找到可能包含非编码 RNA 的候选区域，而这一条件本身的合理性有统计结果支持 [26]。不过，仅仅使用这一条件还不足以精确区分非编码 RNA 和背景 [26]。二级结构可以与各种 RNA 测序和芯片数据及蛋白质序列保守性数据整合在一起，使用机器学习手段，在全基因组水平上对非编码 RNA 进行预测 [27]，这种方法可以十分有效地区分编码序列 (CDS)、内含子 (intron)、非编码 RNA 及基因间片段。

2.3　其他应用

二级结构的预测还被广泛应用在序列比对、3D 结构预测、结构设计等问题中。传统的序列比对只针对 RNA 的一级序列，对二级结构保守性的研究可以为进化问题带来新的视角。核糖体 ITS 经常被用来估计进化关系，它们与核糖体 RNA 来自于同一前体，但功能不像核糖体 RNA 那么重要，所以选择压力较小，突变更为丰富，因此，很适合小时间尺度的物种进化关系的估计。尽管看不出很直观的选择压力且序列非常多变，但对 ITS2 的研究显示，其二级结构在真核物种都有共通的结构 [28]。RNA 的 3D 结构模拟要难于蛋白质，一来核苷酸分子的转动更加灵活，二来缺乏很好的近似力场。所以，给定适当的初始结构非常重要。有些 3D 结构的预测算法会把二级结构作为其初始结构来开展后续的分子模拟 [29]。

3　展望

RNA 的序列到结构，序列、结构到功能，还存在很多未解决的问题和可探索的领域。

结构预测本身的前沿问题，除了进一步提高二级结构预测算法的速度和准确率外，还包括利用和整合各种实验数据，利用日益增长的数据库，从二级结构预测 3D 结构、假结 (pseudoknot) 等。值得一提的是，Kertesz 等 [20] 使用 RNase V1 (倾向于降解双链) 和 S1 nuclease (倾向于切断单链) 处理酵母转录组并测序，第一次在全转录组层面对 RNA 二级结构进行了实验测定。如何有效地把这样的高通量数据整合进预测算法会是将来的一大挑战。

最近，长非编码 RNA (long noncoding RNA) 的研究方兴未艾。作为真核生物中存在的、大于 200 nt 且不编码蛋白质的一类 RNA，它们被发现参与 X 染色体失活、基因沉默等调控过程 [30]。有证据显示，在基因沉默过程中，长非编码 RNA 折叠成双茎环和其他结构，从而招募多梳状复合体 (polycomb complex)，进而执行功能 [4]。在长非编码 RNA 的研究中，结构与功能的对应关系有待揭示，结构预测作为确定结构的重要一环自然不可或缺。

另外，细胞中的 RNA 很少单独存在，它常被 RNA 结合蛋白 (RNA binding protein) 覆盖和包裹。很多重要的生物过程，包括 microRNA 的成熟、RNA 的剪切、翻译的起始都会受到蛋白质结合和 RNA 结构的共同影响。通过交联免疫沉淀配合高通量测序 (CLIP-seq)，可以对一个蛋白质或一组蛋白质的 RNA 结合谱进行快速测定，极大地推动了 RNA 与蛋白质作用的研究 [31]。在此基础上发展出的 PAR-CLIP 和 iCLIP 进一步提高了测定 RNA 蛋白质相互作用的精度 [31]。日益丰富的 RNA 与蛋白质相互作用数据将会为结构预测带来大量潜在的应用和新的方向。一方面，研究显示蛋白质的结合会

影响 RNA 的折叠 [32]；另一方面蛋白质结合的信息可用来反推 RNA 的结构。RNA 结合蛋白更倾向于结合在 RNA 的单链区 [33]，但我们对此还知之甚少。可以预见，在不久的将来，结构预测和蛋白质结合数据共同分析的工作会大量出现。

[参 考 文 献]

[1] Goodarzzi H, Najafabadi HS, Oikonomou P, et al. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature, 2012, 485(7397): 264-8

[2] Parsyan A, Svitkin Y, Shahbazian D, et al. mRNA helicases: the tacticians of translational control. Nat Rev Mol Cell Biol, 2011, 12(4): 235-45

[3] Biirrneey E, Stamatoyannopoulos JA, Dutta A, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 2007, 447(7146): 799-816

[4] Wan Y, Kertesz M, Spitale RC, et al. Understanding the transcriptome through RNA structure. Nat Rev Genet, 2011, 12(9): 641-55

[5] Deigan KE, Li TW, Mathews DH, et al. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci USA, 2009, 106(1): 97-102

[6] Zuker M, Sankoff D. RNA secondary structures and their prediction. Bull Mathem Biol, 1984, 46(4): 591-621

[7] Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res, 2010, 38(Suppl 1): D280-2

[8] Gruber AR, Lorenz R, Bernhart SH, et al. The Vienna RNA websuite. Nucleic Acids Res, 2008, 36(Suppl 2): W70-W4

[9] Mccaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers, 1990, 29(6-7): 1105-19

[10] Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA, 2004, 10(8): 1178-90

[11] Lu ZJ, Gloor JW, Mathews DH. Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA, 2009, 15(10): 1805-13

[12] Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res, 2003, 31(24): 7280-301

[13] Hamada M, Kiryu H, Sato K, et al. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics, 2009, 25(4): 465-73

[14] Goertzen LR, Cannone JJ, Gutell R, et al. ITS secondary structure derived from comparative analysis: implications for sequence alignment and phylogeny of the Asteraceae. Mol Phylogenet Evol, 2003, 29(2): 216-34

[15] Bernhart SH, Hofacker IL, Will S, et al. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics, 2008, 9: 474

[16] Harmanci A, Sharma G, Mathews D. TurboFold: iterative probabilistic estimation of secondary structures for multiple RNA sequences. BMC Bioinformatics, 2011, 12: 108

[17] Andronescu M, Bereg V, Hoos HH, et al. RNA STRAND: the RNA secondary structure and statistical analysis database. BMC Bioinformatics, 2008, 9: 340

[18] Andronescu M, CondonA, Hoos HH, et al. Computational approaches for RNA energy parameter estimation. RNA, 2010, 16(12): 2304-18

[19] Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics, 2006, 22(14): e90-e8

[20] Kertesz M, Wan Y, Mazor E, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature, 2010, 467(7311): 103-7

[21] Watts JM, Dang KK, Gorelick RJ, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature, 2009, 460(7256): 711-6

[22] Ouyang Z, Snyder MP, Chang HY. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. Genome Res, 2013, 23(2): 377-87

[23] Washietl S, Hofacker IL, Stadler PF, et al. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. Nucleic Acids Res, 2012, 40(10): 4261-72

[24] Lu ZJ, Mathews DH. OligoWalk: an online siRNA design tool utilizing hybridization thermodynamics. Nucleic Acids Res, 2008, 36(Suppl 2): W104-W8

[25] Rabani M, Kertesz M, Segal E. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. Proc Natl Acad Sci USA, 2008, 105(39): 14885-90

[26] Uzilov AV, Keegan JM, Mathews DH. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. BMC Bioinformatics, 2006, 7: 173

[27] Lu ZJ, Yip KY, Wang G, et al. Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Res, 2011, 21(2): 276-85

[28] Schultz J, Maisel S, Gerlach D, et al. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. RNA, 2005, 11(4): 361-4

[29] Ding F, Sharma S, Chalasani P, et al. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA, 2008, 14(6): 1164-73

[30] Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem, 2012, 81: 145-66

[31] Konig J, Zarnack K, Luscombe NM, et al. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet, 2012, 13(2): 77-83

[32] Karaduman R, Fabrizio P, Hartmuth K, et al. RNA

structure and RNA–protein interactions in purified yeast U6 snRNPs. J Mol Biol, 2006, 356(5): 1248-62
[33] Iwakiri J, Kameda T, Asai K, et al. Analysis of base-pairing probabilities of RNA molecules involved in protein-RNA interactions. Bioinformatics, 2013, 29(20): 2524-8
[34] Reuter JS, Mathews DH. RNA structure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics, 2010, 11: 129
[35] Torarinsson E, Havgaard JH, Gorodkin J. Multiple structural alignment and clustering of RNA sequences. Bioinformatics, 2007, 23(8): 926-32
[36] Sato K, Hamada M, Asai K, et al. CENTROIDFOLD: a web server for RNA secondary structure prediction. Nucleic Acids Res, 2009, 37(Suppl 2): W277-W80
[37] Seetin MG, Mathews DH. RNA structure prediction: an overview of methods. Methods Mol Biol, 2012, 905: 99-122
[38] Kierzek E, Christensen SM, Eickbush TH, et al. Secondary structures for 5' regions of R2 retrotransposon RNAs reveal a novel conserved pseudoknot and regions that evolve under different constraints. J Mol Biol, 2009, 390(3): 428-42
[39] Durbin R, Eddy S, Krogh A, et al, Biological sequence analysis: probabilistic models of proteins and nucleic acids [M]. New York: Cambridge university press, 1998

鲁志报告讨论

施蕴渝（中国科学技术大学生命科学学院）

Q：如果你发现它跟 H3K36 结合，能预测它是怎么样的 ncRNA ？

A：对。

Q：假设有一个蛋白质，它是不是特异性地结合在这个 H3K36 上，可以预测吗？预测出来可能有什么样的意义，可能跟它结合吗？

A：可以预测出来，这个蛋白质是和 H3K36 结合的。这个和我刚才说的是两码事，H3K36 有一个很强的信号，然后你可以看，在基础位置上面，有个 long ncRNA。您说的这个可能是分子层面的，这个分子和那个分子有很强的 binding。但是在基因组层面上，如果有这个信号的话，在这个位置就有个 long ncRNA 表达，但是跟您刚才说的那个蛋白质有没有结合，现在还是不知道的，可能得做 CLIP。

王恩多（中国科学院上海生命科学研究院生物化学与细胞生物学研究所）

Q：我想问一个问题，就是你说的蛋白质与 RNA，A 到 G 的突变引起 RNA 二级结构的变化。你有没有想，它到底跟哪个蛋白质结合，有没有拿这个东西去真正做一下。就是说，开始的那个结构在肝脏组织里跟某个蛋白质结合，而现在变成这个结构，就不能结合了，有没有这样的考虑。另外一个问题是，RNA 的折叠与很多金属离子有关系。金属离子，如 Mg^{2+}，对二级结构的折叠非常重要。那怕是同一个序列，不同的金属离子对二级结构的影响可能都会不同，这方面你是怎么考虑的。

A：您这个问题问的很好！我先回答第二个问题。关于金属离子，我们现在用的一个简单的替代就是用生理盐水去模拟生理状态，我们考虑的是这种标准状态下的结构预测，即在一种标准的离子情况下的结构。的确，在真实的情况中，可能离子浓度变了，的确会发生一些变化。现在这个阶段是没有考虑到的，可能对大部分预测是没有太大影响，但对某个特别的 motif 可能会有特别的影响。因为我是做生物信息的，高通量在我看来是一大片的，对一两个现象的确没法解释。还有一个就是，跟蛋白质结合，我觉得这个非常好。这个结果我们也刚刚做出来，我们需要验证结构是否真的发生变化。如果有真的发生变化，我们应该去考察和蛋白质的结合。之前我们没有想到，可能我们也没有能力去做，可能需要和大家合作去做 CLIP，然后做质谱。

邵宁生（军事医学科学院）

Q：我想接着王老师的那句话问一下，对于这个 RNA 生前预测，我们原来也做过 RNA、DNA。确实，这个现象很意外。我的问题有点类似王老师，是不是在自由能最低的时候，理论上的结构就是功能的结构。有 Mg^{2+} 和没有 Mg^{2+}，对结构有没有功能影响很大。就是说，如果没有 Mg^{2+}，结构再怎么预测，肯定没结合。光用生理盐水，肯定是不对的。没有 Mg^{2+}，这个结构就没有功能。所以它不是一两个现象，是普遍现象。这个方面，我们现在也面临着这个问题，我们用 RNA structure 和 RNA fold 一一 show 了好几个结果，如果我挑自由能最低的，往往不是我结合的最好的这个，特别让我困惑。我不知道该怎么挑。按传统，挑自由

能最低的，但是往往都不对，这个问题怎么解决是一个特别重要的问题。我觉得在功能状态下，特别是蛋白质结合有 Zn^{2+}，但没有 Mg^{2+}。对于这个结构，请你们生物信息学家给我们一个具体的指导。选这个自由能最低是对的，但验证肯定不对；也不是完全不对，有时候对。

王恩多

Q：tRNA 的功能也跟 Mg^{2+} 非常密切，多了也不行，少了也不行，反正有一个水平。因此，这个结构预测在生理盐水下可能跟细胞里面不一样。细胞里面不是均一的，它们有微环境，微环境有很多的离子，所以预测最终还是得用实验验证。

邵宁生

Q：特别是做 RNA 的，Mg^{2+}，金属离子做实验不能少，少了它反正实验好像有点问题。

A：我觉得这个问题挺大的，不是生物信息学一种方法可以解决的。首先，做化学的得能猜测下去，有个离子有关系。然后我们算法上可以做一些改进，就是我们后来提出了不再预测一个结构，而是给个概率，就是这个 base pairing 的概率可能性是 50% 还是 70%，有可能在生理状态是转换的，50% 是这种，50% 是那种，我们现在可以预测。还有一个是算法上的，你可以加上 evolution，evolution 本身是看序列的 pattern，它已经不再受限于任何化学的东西。就是说，如果这个东西非常保守，你看这个序列，这个地方在一个物种是 AU 配对，在另外的物种 A 变成 C，U 也变成 G 的话，那么这个结构可能是很强的 base pairing，可以部分解决这个问题。最后回答一下，在这种情况下，我们给个数字，70% 的我们预测的单链的是对的；80% 的，如果你有 evolution 的信息的话，是对的；如果你有 chemical profiling 的方法，95% 是对的。

施蕴渝

Q：现在你的那个 pseudo-tRNA 的结构的预测到底怎么回事？

A：大概 100 个以下，预测精度达到 20 Å。可以把轮廓说出来。给个序列，一下子把结构模拟出来。

Q：那个 pseudo-tRNA 是零预测背景的吗？

A：那个比较容易预测。

Q：你没有验证？

A：没有验证，都是预测的。

王恩多

Q：这是 70 多个吧？

A：对，70 多个。我只测试轮廓，本来是 L 型的 (tRNA)，现在变成树型的，这个大概能区分出来。

施蕴渝

Q：你通过预测得到这么多的 ncRNA 的话，有没有从基因组的位置上估计一下。它们的功能是就近调控上下游编码基因的表达，比如说抑制基因的表达，还是说它们自身的 RNA 因为在 chromosome 的另外一个地方序列比较接近，你有没有统计一下？

A: 在植物里面我们有做，我们会挑出一些 cis-regulatory 的 ncRNA 出来，然后看它跟它旁边那个基因，有时候可能是相互抑制的。会挑出一些来，但并不多。很多是 intergenic，还有很多是，包括 anti-sense 研究很多，像你所说的那个 cis-regulatory ncRNA，前面的 promoter 区，一般定义 1K 到 5K 之内，会有些关系，也的确做了一些预测。我们现在在跟别人合作，做一些验证。是不是有这个共定位，是不是真的抑制。

陈润生（中国科学院生物物理研究所）

Q：我是主席，我先说两句，我也是搞生物信息学的。第一个，对邵教授说的那个东西，我想讲一下，从生物信息学，就是说结构模拟这个理论方法来讲，其实它自身存在着一些根本性的缺陷。这里有两个问题，一个是物理模型。基于理论的方法，原则来讲，是用现在的原子物理的方法，那么是严格解量子力学的那套方程。但是现在看来不行的，对于多粒子体系，计算复杂度太高，所以这个势函数，这个物理模型是近似的，近似的模拟要得到精确的解释不可能的。这是一个根本的不足。第二个就是便利性的问题。你要得到全局的能量最小，这是个 NP hard 的问题，实际上是逐渐逼近，不可能的。所以我的理解是，实际上，我们的系统工作的真实系统是在这样一个大的势能面上的某些局部的最小值，是这样的东西，这是我理解

的，也是我了解的生物信息学对什么样的结构是具有生物活性的。如果真要从理论上来讲，得到的真正全局的能量绩效，我觉得这都不只是实验，是很难理解的，作为理论也计算不到。我这里愿意来做这点解释。另外，我问鲁志教授一个问题，在结构模拟当中，在二级结构上，RNA 的结构跟蛋白质很不同，关键非常麻烦的是假节，假节的计算具有极高的计算复杂度，一般来讲，一般要比整个二级结构的模拟要高几个数量级，所以我不知道假节，这是二级向三级过渡的一个关键，在这个问题上，你有些什么进展或者有些什么考虑。

A：这不是我做的。是我的 PhD 老板最近刚发的 PNAS，他将假节和 chemical profiling 结合起来，的确可以预测一些假节，也可以把 chemical profiling 的方法整合进来，预测一些假节。老实说，我在这个上面做的工作并不多，因为只有 2%~3% 才是假节，所以我基本没有往这个方向继续做了。但是它的确很重要，不是说它数量少就不重要，三级结构很大程度上就决定于假节。

王恩多

Q：我再问一个小问题，就是刚才一张幻灯片 (tRNA、pseudo-tRNA) 上，pseudo-tRNA 的 D-stem 似乎是没有了，类似于 D 环的那个，它的核苷酸的序列，比方说它的 D 还有没有，还有整个的相当于 D 环那个地方的序列有没有大的变化，还是只是少了一个 D 茎 (D-stem)。

A：王老师说的很对。

Q：D loop 全跟 tRNA 一样吗？

A：你说在序列上？

Q：我只是讲序列。

A：我还真没看，这个 base pair 就没有了。

Q：就是那个 D20 没有了，是吧，大的结构看起来是 D 茎没有了。

A：是。

Q：细看环也没有了。D 环上的其他序列不知道怎么样？

A：我还真没比对过。