RNA 折叠结构预测算法与计算复杂性

刘振栋 著

科学出版社

RNA 折叠结构预测算法与计算复杂性

刘振栋 著

科学出版社

北京

内 容 简 介

本书介绍了RNA结构特征、假结表示模型和最小自由能量原理等内容。重点介绍了包含假结的RNA折叠结构预测启发式算法、限制性RNA结构预测问题的精确算法、多项式时间近似方案、近似算法等问题。分析其时间复杂度和空间复杂度，比较其特异性和敏感性。针对NP难问题，设计了预测包含假结的$1+\varepsilon(\varepsilon>0)$多项式时间近似方案，分析了包含假结的RNA折叠结构，剖析连续堆叠和假结的结构特性，提出求解最大堆叠数的近似算法。针对RNA折叠结构盆跳图的结构和性质进行解析，对几种类型加以分析、分类，并对类型之间的转换进行阐述。最后对RNA折叠结构的应用加以简单的介绍。

本书可作为生物信息学或计算生物学专业高等院校本科生及研究生参考用书，也可作为相关领域学者及兴趣爱好者的参考用书。

图书在版编目（CIP）数据

RNA折叠结构预测算法与计算复杂性 / 刘振栋著. —北京：科学出版社，2019.10

ISBN 978-7-03-062144-3

Ⅰ. ①R… Ⅱ. ①刘… Ⅲ. ①核糖核酸-分子结构-预测-计算复杂性-研究 Ⅳ. ①Q522

中国版本图书馆CIP数据核字（2019）第180945号

责任编辑：赵艳春 / 责任校对：王萌萌

责任印制：吴兆东 / 封面设计：蓝 正

科学出版社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

北京捷迅佳彩印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2019年10月第 一 版 开本：720×1000 B5

2020年 1 月第二次印刷 印张：7 1/2

字数：151 000

定价：99.00元

（如有印装质量问题，我社负责调换）

前　言

核糖核酸(ribonucleic acid, RNA)作为生物大分子具有十分重要的生物学功能，RNA 结构预测是计算分子生物学的基本课题之一，也是当今国际研究的热点。作者从 2000 年开始研究 RNA 折叠结构预测算法及其计算复杂性方面的内容，重点研究 RNA 折叠结构预测启发式算法，包含假结的 RNA 结构预测算法、限制性 RNA 结构预测问题的精确算法、多项式时间近似方案、近似算法等。

RNA 三级结构是比较稳定的结构，而预测 RNA 三级结构需先预测 RNA 二级结构。预测 RNA 二级结构方法主要有序列对比方法和最小自由能量方法。序列对比分析法预测 RNA 二级结构，是通过将在不同生物有机体中起相同生物功能的一级结构进行比对得到 RNA 碱基序列的二级结构。许多生物有机体 RNA 分子的同源序列不易得到，需要耗费大量人力，因而序列对比方法的预测效率较低，利用最小自由能量方法来预测 RNA 二级结构是广泛采用的方法之一。

连续基对构成堆叠，基对的交叉形成假结点，茎区的交叉构成假结结构，目前现有的预测含假结的 RNA 二级结构的算法，对较大的 RNA 分子计算很困难。基于茎区组合来寻找 RNA 优化结构成为包含假结 RNA 结构预测的重要方法，Benedeti 等提出基于茎区组合的能量集合算法来预测 RNA 二级结构，Ruan 等提出基于茎区的启发式算法来预测包含假结的 RNA 二级结构，其时间复杂度为 $O(n^4)$，空间复杂度为 $O(n^2)$。根据 RNA 假结表示模型，基于 RNA 茎区结构相对稳定的特征和最小自由能量原理，本书提出预测含假结的 RNA 二级结构的启发式算法，时间复杂度为 $O(n^3)$，空间复杂度为 $O(n^2)$，通过 RNA 假结库实验表明，该算法有较好的预测特异性和敏感性。

连续堆叠可构成茎区，针对基于茎区的 RNA 优化结构，将序列划分为长度不大于 $t(t>2)$的子序列，计算出由长度不大于 t 的子序列构成的最优结构作为整个序列的近似结构，设计出预测任意假结的 $1+\varepsilon(\varepsilon>0)$多项式时间近似方案。

堆叠最大化问题也是近年来人们十分关注的含假结 RNA 二级结构预测问题。在平面 RNA 二级结构中，允许假结的存在使计算最大堆叠数问题成为 NP(non-deterministic polynomial)难的，Ieong 等提出了最大堆叠基对数问题，设计了带任意假结的 RNA 二级结构预测近似算法，分别设计出平面二级结构的近似算法和普通二级结构的近似算法，并且证明了平面 RNA 二级结构中求含假结的最大堆叠数问题也是 NP 难的。本书分析包含假结的 RNA 二级结构，剖析连续堆

叠对和假结的结构特性，分析求解最大堆叠数的近似算法，设计其近似性能比为3，并给出证明，讨论最大堆叠数问题的计算复杂性，并且可以预测更复杂的假结。本书给出 RNA 碱基序列 S 和正整数 h 的最大堆叠数问题，通过把三划分匹配这一 NP 难问题规约到该问题，判断在平面 RNA 二级结构中是否可能存在大于等于 h 的最大堆叠数，从而证明在平面 RNA 二级结构中含假结的最大堆叠数问题也是 NP 难的。

本书的主要成果为：

(1) 基于最小自由能量的 RNA 结构的表示建模是 RNA 结构预测的关键。对于假结而言，可分为平面假结和非平面假结，假结可形成嵌套或并列结构，由两个茎区结构可形成嵌套假结，由内环和凸起可构成平面假结，平面假结经常出现在 RNA 分子中，交叉假结也存在于 RNA 中。茎区在 RNA 结构稳定性中承担着重要作用，基于茎区的交叉可形成假结的特性，可利用茎区结构建立关于假结的表示模型。在 PseudoBase 假结数据库中，大部分为平面假结，也包含少量的非平面假结。本书通过设计启发函数、用恰当的假结表示建模来预测 RNA 假结结构，取得了较好的效果。

根据 RNA 假结表示模型，基于最小自由能量原理，本书设计预测任意平面假结和非平面假结的启发式算法，通过在 PseudoBase、Rfam 等 RNA 假结数据库中的实验验证表明，算法的预测敏感性特异性和预测准确度均有所提高，其时间复杂度为 $O(n^3)$，空间复杂度为 $O(n^2)$。

(2) 一般来说，连续的堆叠可形成茎区结构，茎区结构可使 RNA 结构能量降低，结构更稳定。通过茎区的组合优化特性来预测 RNA 优化结构是我们采用的重要方法，茎区之间可形成并列结构、嵌套结构和交叉结构。含有交叉结构即包含假结，假结的存在使 RNA 结构预测变得复杂，是问题难解的重要因素，使得设计多项式时间算法变得异常困难，设计该问题的近似算法或近似方案成为处理该问题的重要手段。

本书针对基于茎区的 RNA 优化结构，把 RNA 碱基序列用短茎进行划分，计算出由长度不大于 t 的茎区构成的结构作为整个序列的近似结构，重新分析预测任意假结的 $1+\varepsilon(\varepsilon>0)$多项式时间近似方案。

(3) 在 RNA 碱基序列中，连续的两个碱基对可构成堆叠，从堆叠的角度看，多个连续碱基对可形成连续堆叠，连续堆叠中堆叠的个数越多，则 RNA 结构越稳定。在 RNA 结构预测中，包含假结的计算最大堆叠数问题也是 NP 难的，针对该类问题，如果设计不出多项式时间精确算法，不如退而求其次，通过其内在特性的深入分析，设计求解该类问题的多项式时间近似算法。分析其近似性能比，尝试降低近似比，指导该问题的求解。

针对连续堆叠对的结构特性，本书重新分析 RNA 二级结构最大堆叠数问题，

通过在 RNA 折叠结构中查找连续堆叠，并对内在特性加以剖析，分析计算最大堆叠数的近似算法，设计其近似性能比为 3，并给出近似性能比的证明。

(4) 本书剖析 RNA 折叠结构盆跳图的结构和性质，对包含假结的盆跳图几种类型加以分析、分类，并对类型之间的转换进行阐述。

本书是作者及所在课题组多年来研究成果的总结，主要内容来自作者的博士学位论文及所主持的国家自然科学基金面上项目、山东省自然科学基金面上项目等项目的研究成果。

本书的研究工作得到了清华大学戴琼海院士、山东大学朱大铭教授、美国加州大学洛杉矶分校 Gang Li 教授、哈佛大学 Jun S Liu 教授的指导。本书得到了作者主持的国家自然科学基金面上项目(基金号：61672328)的资助。在此一并感谢。

希望读者能从书中得到有益的启示。由于作者的水平有限，难免存在不足之处，恳请读者批评指正。

刘振栋

2019 年 3 月于济南

目　录

第1章 绪 论

1.1 背 景

1985年，美国科学家提出了人类基因组计划(Human Genome Project, HGP)[1]，随着21世纪初人类基因组测序的完成，人类进入后基因时代。诺贝尔奖获得者Dulbecco在癌症研究的过程中，在*Science*上发表了关于人类基因组测序的论文，推动了人类基因组计划的实施，也极大地促进了计算生物学和生物信息学的发展。近年来，有关RNA的研究，引起了国内外众多学者的关注，如何利用现有技术与方法预测其他生物有机体的DNA基因组和RNA结构是当今世界的研究热点。有关RNA的研究已经多年被*Science*列入世界十大科技进展，这充分地说明了RNA的研究在当前计算生物学领域的核心地位。自由能量是衡量RNA结构稳定性的重要指标，一般观点认为，RNA各结构单元的自由能量值越小，则堆叠力使RNA结构越稳定。

RNA折叠结构预测中很多问题都是NP难的，而NP难问题是世界七大数学难题之一。如果设计不出精确算法，不如去设计其多项式时间近似算法，去指导该类问题的生物应用。

包含假结的RNA结构预测问题被证明是NP难问题[2]，预测RNA结构的本质就是找出序列的各个位点之间形成的配对关系，茎环结构是构成RNA碱基序列的重要结构单元，而A—U、C—G碱基对占据茎环的主要基对，有十几种A—U、C—G碱基对组合，其中11种邻位组合与G—U错配有关[3]。随着对环结构热动力学研究的深入和实验手段的提高，自由能量参数得以进一步修正。Turner等[3,4]对RNA二级结构中自由能量参数加以改进，使其正确率进一步提高。独立结构单元模型和最邻近邻居模型是RNA的能量模型的两种典型代表，最邻近邻居模型中结构单元中堆叠与环是由一对最邻近碱基对决定的，RNA分子的自由能量来源于堆叠和环的贡献。但不包括假结，假结是稳定RNA结构的重要的三级结构单元，吸引着越来越多的学者对其进行预测研究。如毒菌RNA的假结结构由Pleij等于1985年加以预测，1998年，Kolk等证实了假结的存在。一些RNA分子不编码蛋白称为非编码RNA(non-coding RNA, ncRNA)。大量例子证明ncRNA突变与很多疾病有关，突变可能影响了RNA的结构。为了统计突变与疾病的关联原因，必须解析RNA结构。RNA也是生命中最基本、最基础的分子之一，

人类有必要了解 RNA 的结构，以便理解它们的功能。在各种 RNA 分子中，假结具有调节、催化、构造等多样化重要功能，在探索生命科学中的现象和规律中具有重要意义[5,6]。

基因是生命的蓝图，蛋白质是生命的机器。来自于四种字符字母表(A，T(U)，C，G)的核酸序列中蕴藏着生命的信息，而蛋白质则执行着生物体内各种重要的工作，如生物化学反应的催化、营养物质的输运、生长和分化控制、生物信号的识别和传递等。蛋白质序列由相应的核酸序列所决定，核酸是由核苷酸聚合而成的高分子化合物。核酸中储存着生命体的全部遗传信息，是所有生物遗传信息的携带者。根据核苷酸分子中戊糖的类型，将核酸分为脱氧核糖核酸(Deoxyribonucleic Acid, DNA)和核糖核酸两大类，RNA 是由多个核苷酸聚合成的一种单链高分子化合物。生物体内共有三种 RNA，即信使核糖核酸(mRNA)、核糖体核糖核酸(rRNA)和转运核糖核酸(tRNA)。DNA 是遗传信息的载体，遗传信息的作用通常由蛋白质的功能来实现，但 DNA 并非蛋白质合成的直接模板，合成蛋白质的模板是 RNA。正常细胞遗传信息的流向是

DNA→(转录为)RNA→(翻译成)蛋白质

针对包含假结的 RNA 结构预测问题被证明是 NP 难问题，针对包含假结的 RNA 结构预测近似算法及计算复杂性相关问题本书进行了研究。该问题是 RNA 结构预测问题中的典型代表，RNA 结构预测问题来源于 RNA 编码的秘密。除 RNA 的一级结构用实验的方法来测定外，其二级和三级结构目前用实验的方法测定十分困难，为获取 RNA 结构功能信息，获知生物分子的生物学功能，通过计算方法来预测 RNA 结构成为计算生物学领域一个重要的课题和研究热点。

在多项式时间可解的问题得到普遍研究之后，RNA 结构预测 NP 难问题的近似算法研究成为算法分析与设计这一计算机科学的经典领域中的活跃分支[7, 8]。关于 RNA 结构预测算法问题近似理论的研究吸引了来自世界各地著名大学和研究机构的众多专家学者，是近似算法领域中显著的热点。本书定位于 RNA 假结结构预测算法这一专题，开展近似算法和不可近似性的研究，具有较高的学术价值。本书的研究内容是通过给出有效的近似算法、参数化算法，来求解包含假结的 RNA 结构预测问题，该问题在理论上被证明是 NP 难问题，且对若干研究问题获得新结果，在近似算法设计和分析中提出新思想、新观点。本书的研究结果体现在两个层面：一方面，将有助于近似算法的研究，为算法与计算复杂性近似理论的发展做出贡献；另一方面，项目的实施将促进近似理论在计算实践中的应用。RNA 结构决定 RNA 功能，RNA 结构预测算法和技术的改进，为探索 RNA 结构与功能在生命活动中的机理提供新的途径和方法。RNA 结构预测的研究，为寻找非编码 RNA 基因，为 RNA 病毒和靶向核糖体药物研制提供了新思路、新方法，对揭开 RNA 编码的秘密，探索生命起源、进化具有重要意义。

1.2　国内外研究现状

RNA 在遗传信息从 DNA 表达为蛋白质的过程中起转录作用(图 1.1)，RNA 结构预测，特别是 RNA 二级结构预测研究是当今学术界的研究热点，其中也存在预测准确度不高、特异性与敏感性需要改进、预测算法时空复杂度不理想等问题。本书的研究，有望解决 RNA 结构预测算法中存在的前沿问题，也为生物医学研究提供了理论和技术指导。

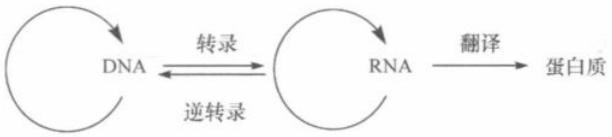

图 1.1　DNA、RNA 与蛋白质关系的中心法则

RNA 分子一般是线状单链结构，然而 RNA 分子的某些区域可自身回折，进行碱基互补配对而形成局部双螺旋结构。RNA 双螺旋中，一般是 A 与 U 配对、G 与 C 配对，但存在非标准配对，如 G 与 U 错配对。RNA 分子中的双螺旋与 A 型 DNA 双螺旋相似，而非互补区则膨胀形成凸出(bulge)或者环(loop)，短的双螺旋区域和环可形成发夹结构(hairpin)，发夹结构是 RNA 中最普通的二级结构形式，二级结构进一步折叠形成三级结构，RNA 分子只有在具有三级结构时才有活性。RNA 也能与蛋白质形成核蛋白复合物，RNA 的四级结构是 RNA 与蛋白质的相互作用，RNA 结构预测是计算生物学与生物信息学的典型问题。

算法及计算复杂性研究始终是计算机科学与技术的一个研究热点，算法往往来源于实际问题的抽象。计算生物学就是从实际生物问题中抽象出的计算模型，并且通过设计有效的算法而研究生物现象的一门学科。算法及复杂性的研究极大地促进了计算生物学和生物信息学的发展，也为生物医学发展提供了理论和技术支持，如基因组重组问题、RNA 二级结构、三级结构预测问题等。2000 年，世界著名计算机科学家，美籍华人姚期智因伪随机数生成、密码学与通信复杂度等计算理论方面的杰出贡献获得图灵奖，包括以图灵奖获得者 Cook、Karp、Shamir、Hopcroft、Tarjan、Hartmanis、Stearns、Blum 为代表的一批世界级理论计算机科学家，以创造性的工作推动着算法研究不断深入发展，吸引着一大批算法爱好者，而算法及计算复杂性在计算生物学或生物信息学方面的研究，特别是对 RNA 结构预测研究是当今世界研究的热点。

Hochbaum 在其著作 *Approximation Algorithms for NP-hard Problems*(《NP 难问题的近似算法》)中开展了典型 NP 难问题近似算法的深入研究。近似算法为解

决含假结的 RNA 结构预测这一 NP 难问题提供了本质方法，对于 NP 难类问题都有其本身的组合结构，从 RNA 折叠结构预测问题求解和预测算法入手，其算法设计的精髓就是发现 RNA 结构预测问题的组合性质。NP 难问题的最优解有时是不可奢求的，本书尝试在多项式时间内设计出近似算法来求可行次优解，用近似解与最优解比值的上界(或下界)，来衡量算法近似性能比所求解的质量。本书致力于 RNA 折叠结构研究，特别是 NP 难包含假结的 RNA 二级结构、三级结构预测算法的研究。在包含假结的 RNA 结构预测中，设计计算最大堆叠数的近似算法，探求 NP 难问题的可近似性上界(或下界)及近似难度结果。

包含假结的 RNA 二级结构预测问题被证明是 NP 完全问题，本书针对包含假结的 RNA 折叠结构预测算法及计算复杂性这一专题开展研究，该专题是 RNA 结构预测问题中的典型代表。RNA 结构预测问题来源于 RNA 编码的秘密，除 RNA 的一级结构用实验的方法来测定外，其二级和三级结构目前用实验的方法测定十分困难，本书在 RNA 结构功能信息获取、RNA 分子的生物学功能获知(如 2014 年在非洲肆虐的埃博拉 RNA 病毒生物学功能获知)、非编码 RNA 基因寻找、RNA 编码秘密揭开、RNA 病毒机理解密和靶向核糖核酸药物研制、RNA 分子机理探索方面取得了突破，所以本书取得的成果在探索生命起源和生命进化过程中具有十分重要的意义。因此，通过计算方法来预测 RNA 结构成为国际计算生物学、生物信息学领域一个重要课题和研究热点。

不同于 DNA 的双螺旋结构，RNA 是单链结构，RNA 碱基序列中包含 A、C、G、U 四种碱基(图 1.2)。影响 RNA 折叠结构稳定性的因素有很多，通常用碱基配对所需要的自由能量来衡量，并且自由能量越小，RNA 结构越稳定。

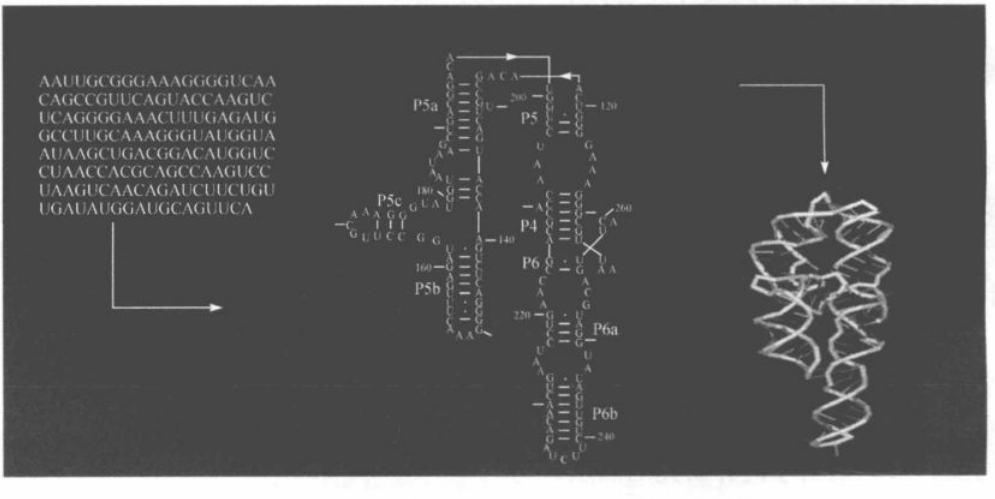

图 1.2　RNA 二级结构和三级结构

预测 RNA 折叠结构的本质就是找出 RNA 碱基序列的各个位点之间形成的配

对关系。RNA 碱基序列中大多数是仅包含 A—U、C—G 基对的茎环结构，常见的 RNA 茎环的邻位基对有十余种可能的组合，有趣的是 G—U 错配在 RNA 碱基序列中也经常发现，也有十几种邻位组合涉及 G—U 错配。环结构对 RNA 结构稳定性承担重要角色，目前对环的热动力学研究相对较少，其自由能量参数可见相关文献，随着寡核苷酸合成技术的提高，合成了大量不同的寡核苷酸链用于实验测定，使自由能量参数的正确率进一步提高。Turner 等对改进的参数做了系统的总结，成为目前普遍采用的参数。RNA 能量模型有结构单元间的近邻相互作用模型和独立结构单元模型，而最邻近邻居模型是一种独立结构单元模型的特殊情况，该结构单元中堆叠与环是由最邻近碱基对决定的，RNA 分子的自由能量主要来源于堆叠和环的贡献，但不包括假结。

更有趣的是假结的存在使 RNA 折叠结构复杂化，假结(pseudoknot)是 RNA 分子中最广泛的三级结构单元，是非常复杂和稳定的 RNA 结构。假结在不同的 RNA 分子中具有调节、催化、构造等重要功能，承担着多样化的角色，在探索生命科学中的现象和规律中具有重要意义，包含假结的 RNA 结构预测是目前 RNA 结构预测研究的难点和关键点。1985 年，Pleij 等预测了几种毒菌 RNA 的假结结构，并由 Kolk 等于 1998 年予以证实，假结作为一种复杂而稳定的 RNA 折叠结构，广泛存在于生物细胞中，也是实现其生物功能的重要因素，如端粒酶的 RNA 分子活性就取决于其假结结构，有些病毒细胞利用假结结构模仿宿主 tRNA 分子从而入侵生物体。1990 年，Pleij 等利用数学方法提出了 14 种理论上存在的 RNA 假结结构，并利用弧图进行了分析和归纳。假结种类繁多、结构复杂，在 RNA 结构研究早期，包含假结的 RNA 结构预测就已被证明为 NP 难问题，经过多年来对 RNA 结构预测算法的不断改进完善，目前也仅有发夹环与单链形成的 H 型假结可被精确预测。有关含假结的 RNA 结构预测算法近似理论与技术的研究吸引了来自世界各地知名大学和研究机构的众多专家学者的关注，是近似算法领域研究中的热点之一。在多项式时间可解的问题得到普遍研究之后，包含假结 RNA 折叠结构预测的 NP 难问题的近似算法研究成为算法分析与设计这一计算机科学经典领域中的活跃分支。

南加利福尼亚大学 Waterman 开创了生物信息学和计算生物学的先河。美国的 McMahon 首先提出运用计算机技术预测 RNA 二级结构，1994 年，Walter 等对同轴堆叠在 RNA 折叠中的作用进行了研究，困难的是不少 RNA 结构中还包含一种特殊而又重要的非嵌套结构——假结，假结的出现破坏了动态规划算法所依赖的 RNA 结构的嵌套子结构性质，Rivas 等[9]提出的 PKNOTS 算法可预测包含任意假结的 RNA 二级结构，其时间复杂度和空间复杂度分别为 $O(n^6)$和 $O(n^4)$，该算法通过限制假结的类型(只允许出现几种比较简单的假结)来预测 RNA 带假结的二级结构，较高的时间复杂度和空间复杂度严重制约了算法所能处理问题的规模，

从而使带假结的 RNA 二级结构预测成为一个难题，实际上包含假结的 RNA 二级结构预测问题是 NPC 问题。RNA 假结结构预测在国际上受到高度重视，是 RNA 结构预测领域中显著的热点，关于假结参数一般为非假结参数乘以系数 $g(0.83)$作为补偿[10]，这些参数值一部分由实验结果计算得到，另外一些参数为理论估计值。Zuker 等[11]设计的 MFOLD 算法中，最邻近邻居模型中热力学能量参数由动态规划算法来计算，Nixon 等[12]对 mRNA 假结进行研究，提出移码突变的 mRNA 解决方案，2003 年，Ieong 等[13]设计了包含任意假结的计算最大堆叠数问题，提出近似性能比为 1/3 的近似算法。2004 年，Lyngsø [14]分析了基于堆叠简单模型的假结复杂性，证明了假结预测的难解性，设计了计算优化堆叠数的递归算法，但其时间复杂度为 $O(n^{81})$，空间复杂度为 $O(n^{80})$。Ruan 等[15,16]也对 RNA 假结进行了研究，分别提出了包含假结的环匹配算法和启发式算法，Huang 等[17]对 RNA 假结结构的预测敏感性进行了研究，Han[18]提出了包含假结的 RNA 结构队列算法。

2011 年美国罗切斯特大学的 Ellaousov 提出了包含假结的 RNA 二级结构快速预测算法，该算法时间复杂度为 $O(n^2)$，该算法的预测准确度为 69.3%，但预测长度不超过 700 的核苷酸，预测精度和预测长度仍需改进。如果一个茎区的形成能使 RNA 假结结构更稳定，则这个茎区更有可能先形成，而衡量 RNA 结构稳定的这个指标就是自由能，因此算法采用自由能作为评估和衡量候选茎区的标准，假结可用茎环结构来表示，以此设计相关 RNA 假结结构预测近似算法，相关研究成果申请者已发表于国内外学术期刊。

2013 年，加拿大麦吉尔大学计算机科学学院的 Reinharz 等利用抽样方法和加权样本技术提出了加权样本算法，对 RNA 二级结构加以预测，取得了较好效果。研究者深入分析了含假结的 RNA 折叠结构内部特性，基于最大堆叠数和能量最小原理，提出了预测包含假结的 RNA 结构预测算法。

近年来，国内许多学者投入到 RNA 结构预测，特别是 RNA 二级结构预测中来。20 世纪末期，清华大学自动化系李衍达和张学工在国内率先致力于生物信息学方面、计算生物学方面的研究，清华大学自动化系李梢、汪小我也在基因调控分析与建模、复杂疾病计算分析等方面取得了若干研究成果。国际著名计算生物学家、生物信息学专家吉林大学长江学者徐鹰长期致力于癌症生物信息学、微生物信息学和结构生物信息学等相关领域的研究，在生物通路和网络的计算方法与模型研究、比较基因组分析、蛋白质结构预测与建模等方面做出了重要的和公认的贡献。中南大学陈建二、王建新、李敏利用参数化算法等理论与技术在生物信息计算领域内进行了深入系统的研究，取得了具有领先水平的理论成果。近年来，国内许多学者也投入 RNA 结构预测研究中，特别是 RNA 二级结构预测中来。中国科学院计算技术研究所孙凝晖指导的徐琳等提出一种对动态规划矩阵采用分块技术的细粒度并行算法，对面向 FPGA 的 RNA 二级结构进行预测[19,20]，提高了

算法效率，但没有考虑假结。中国科学院陈翔等根据 RNA 折叠的特点，提出了一种启发式搜索算法来预测带假结的 RNA 二级结构，以茎区结构为基本搜索单元，该算法能降低搜索 RNA 二级结构的时间复杂度[21]。吉林大学刘元宁等[22]补充提出了 14 种类型的 RNA 假结结构，并使用一种改进的 RNA 平面结构表示法——弧图，求解基于茎区组合的 RNA 二级结构。刘元宁等[22]根据 RNA 折叠的特点，提出了一种启发式搜索算法来预测带假结的 RNA 二级结构，该算法以 RNA 的茎为基本单元，采用启发式搜索策略在茎的组合空间中搜索自由能最小并且出现频率最高的 RNA 二级结构，该算法能降低搜索 RNA 二级结构的时间复杂度。

Gupta 等[23,24]在求解 Rent-or-Buy 问题时，将算法博弈论的费用分摊(cost-sharing)方法应用到近似算法的设计与分析中，相关研究成果分别发表于理论计算机科学国际顶级会议(Foundations of Computer Science, FOCS)和国际著名期刊 *Journal of the ACM*。费用分摊方法在近似算法中的应用得到了许多著名学者的关注，近似算法的不可近似性成为近似算法领域中的一个新的热点，近似算法及随机算法的去随机化技术，也为 RNA 假结结构预测提供了新思路、新方法[25-27]。有趣的是，把 RNA 序列碱基看作点，两碱基若配对则画一条线段，若线段之间有交叉，则表明存在假结，这样把 RNA 假结结构优化问题转化为图问题或网络问题，利用近似算法、随机算法理论和技术，使设计 NP 难 RNA 假结结构预测近似算法和证明问题的可近似性下界成为可能。香港大学 Wong 对含复杂假结的 RNA 结构进行了研究，提出了 RNA 结构比对方法，主要来判断 ncRNA，并且在超过 350 个 ncRNA 家族中实验，证明该算法是有效的。Wong 等[28-30]也设计了包含简单假结的 RNA 结构比对算法，其时间复杂度为 $O(mn^3)$，并设计含嵌套假结的 RNA 结构比对算法，时间复杂度为 $O(mn^4)$。

西安电子科技大学高琳等近年来在生物网络大数据的建模、分析及应用，以及长非编码 RNA 识别与功能分析方面取得了丰硕的成果。哈尔滨工业大学郭茂祖等利用贝叶斯网络结合不同算法来预测小 RNA，提高了预测的敏感性和特异性。2015 年，山东大学李国君联合吉林大学、美国阿肯色州立大学、佐治亚大学等的研究人员提出了一种新的 RNA 转录组组装方法——Bridger，为两种大众组装方法——基于参考序列的 Cufflinks 和从头组装方法 Trinity 之间搭建了一种桥梁关系，其研究成果发表在国际著名学术杂志 *Genome Biology* 上。

Liu 等[31,32]对包含假结的 RNA 折叠结构的内在性质进行了深入研究，提出了预测 RNA 假结结构多项式时间近似方案(Polynomial Time Approximation Scheme, PTAS)，并设计了近似性能比为 2 的近似算法，Liu 等[33]的研究成果发表于 2014 年 SCI 期刊 *International Journal of Sensor Networks*(影响因子 1.386)及 2014 年 SCI 期刊 *International Journal of Pattern Recognition and Artificial Intelligence*。包含假结的 RNA 结构预测中有些近似算法仍有改进余地，例如，求解含任意假结的平

面和非平面最大堆叠数问题近似算法、最大堆叠基对数问题近似算法等，如何证明该类问题的可近似性下界，如何降低近似性比，这些挑战性的工作会激发我们极大的热情。针对 NP 难 RNA 结构预测近似算法中预测效率的提高、不可近似性的证明问题，这也极具有挑战性。

美国华盛顿大学的 Andronescu 等[34-37]对 RNA 折叠中的最邻近邻居的参数进行了研究，提出了用已知 RNA 序列库来确定参数值的新方法。2015 年，美国芝加哥大学的数学和计算机科学家 László[38]对图同构问题(graph isomorphism problem)宣称找到了一个拟多项式时间的算法，意味着可同时对两个网络系统进行计算，这一创造性成果可使生物计算网络更加简单，也为我们进行 RNA 结构网络计算提供了新思路。2015 年，Sarah 等[39]研究了 HIV-1 的 RNA 结构包装信号，对 HIV-1 研究做出了贡献。

近年来，关于基于盆跳图(basin hopping graph, BHG)的包含假结 RNA 折叠结构研究提供了一个新思路，对于包含假结的几种类型加以研究，并对类型之间的转换加以阐述。

2016 年，Vu 等[40]对单细胞 RNA 序列的数据加以分析，建立了 Beta Poisson 模型。

2017 年，Gómez-Schiavon 等[41]在单细胞单分子 RNA 中，提出用 Monte Carlo 方式，建立模型参数来研究 BayFish 的贝叶斯后验概率。

2017 年，宾夕法尼亚大学 Wang 等[42]研究了在单细胞碱基序列中，基因表达的去卷积化问题。

在研究基因表达过程中，对于剪接体的 RNA 加工酶识别的序列，剪接内含子。内含子几乎总是以碱基 G—U 开始，以 A—G 终止，可是需要剪接位点周围的附加序列来提供其足够的特异性。

RNA 是生命中最基本的分子之一，我们需要深入地了解 RNA 的折叠结构，了解它们是如何运作的。许多疾病都与 RNA 的突变相关，这些突变可能影响了 RNA 的结构。为了统计分析突变与疾病的相关性，也需要分析 RNA 折叠结构。bpRNA 数据库能够解析 RNA 结构，包括包含假结点的 RNA，应该对所有环结构、茎区结构和假结点给出一个客观的、精确的、易于解释的描述。还可以得到每一种 RNA 折叠结构特征的位置、序列和碱基对，bpRNA 数据库可以使我们大规模地研究 RNA 折叠结构。

bpRNA 是一个元数据库，在没有以自动化的方式确定所有的结构特征之前，提供了一个颜色编码的地图，显示所测 RNA 折叠结构的位置。使用注释也使我们能够对 RNA 折叠结构形成统计趋势，并可为人工智能、机器学习、深度学习算法打开一扇门，使其以崭新的方式预测 RNA 折叠结构。

研究单细胞的全局基因表达图谱，包括 RNA 表达图谱，有助于解剖隐藏的细胞群体中的异质性问题。与使用大块 RNA 样品，仅提供不同组成细胞的虚拟