中国科学: 物理学 力学 天文学 2023 年 第 53 卷 第 9 期: 290018
SCIENTIA SINICA Physica, Mechanica & Astronomica physcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
进 展 华中师范大学百廿校庆专辑

RNA-蛋白质复合物结构预测研究进展

曾成伟, 赵蕴杰*

华中师范大学物理科学与技术学院, 生物物理研究所, 武汉 430079
*联系人, E-mail: yjzhaowh@mail.ccnu.edu.cn

收稿日期: 2023-02-04; 接受日期: 2023-03-09; 网络出版日期: 2023-08-18
国家自然科学基金(编号: 12175081)和中央高校基本科研业务费(编号: CCNU22QN004, CCNU22JC009, KJ02502022-0450)资助项目

摘要 RNA-蛋白质复合物是生命体中参与细胞活动的重要生物大分子, 相关复合物三级结构的测定对理解细胞活动过程和相关疾病的治疗有重要帮助. 然而, 实验测定RNA-蛋白质复合物分子结构需要大量的时间和财力, 高分辨率的复合物结构测定依然面临重大挑战. RNA-蛋白质复合物结构预测是目前软物质物理和生物物理学领域最具挑战性和影响力的研究方向之一. 本文首先综述了RNA-蛋白质复合物结构预测的研究进展与挑战, 随后讨论了如何优化复合物结构采样和结构评估, 最后对如何利用生物分子大数据和深度学习模型进一步优化RNA-蛋白质复合物预测方法做了展望.

关键词 RNA-蛋白质结构预测, 结构采样, 结构评估, 深度学习

PACS: 87.14.E-, 87.14.gn, 87.15.bg, 88.10.gc

1 引言

RNA-蛋白质复合物是生命体中重要的调控分子, 广泛参与生命体中的基因调控、DNA修复、RNA剪切、免疫反应和蛋白质合成等生物学功能[1–7]. RNA-蛋白质复合物结构异常还与人类的神经系统疾病和癌症密切相关, 是靶向基因治疗的潜在药物标靶[8–12]. 深入理解RNA-蛋白质的生物学功能需要知道其精确的三级结构[13–15]. 然而, RNA自身结构具有较大的柔性特征, 通过X射线晶体衍射、NMR或冷冻电镜解析RNA-蛋白质分子结构需要花费大量的时间和财力[16,17]. 截至2022年11月11日, 核酸数据库(NDB)中仅有2251个RNA-蛋白质复合物结构, RNA-蛋白质复合物的三级结构数量增长十分缓慢[18,19]. 鉴于实验测定RNA-蛋白质复合物结构的复杂性与局限性, 通过理论预测RNA-蛋白质复合物结构是目前迫在眉睫的需求[20,21]. 本文首先回顾了RNA-蛋白质复合物结构预测的国内外研究进展和瓶颈问题, 随后讨论了结合位点预测优化结构采样和口袋拓扑信息优化结构评估的可行性, 最后讨论了如何利用分子大数据和深度学习模型进一步改进RNA-蛋白质复合物结构预测.

2 RNA-蛋白质复合物结构预测的研究进展与挑战

RNA-蛋白质复合物结构预测大多数通过分子对

引用格式: 曾成伟, 赵蕴杰. RNA-蛋白质复合物结构预测研究进展. 中国科学: 物理学 力学 天文学, 2023, 53: 290018
Zeng C W, Zhao Y J. Advances in RNA-protein structure prediction (in Chinese). Sci Sin-Phys Mech Astron, 2023, 53: 290018, doi: 10.1360/SSPMA-2023-0045

© 2023 《中国科学》杂志社

接算法实现, 主要分为两个步骤: (1) 结构采样, 配体在受体结构表面通过旋转平移搜索生成待评估的复合物结构; (2) 结构评估, 构建能量函数对采样结构进行排名并筛选近天然态结构[22–24]. 图1为RNA-蛋白质结构预测的标志性进展, 将按照复合物结构采样和结构评估两个方面展开讨论.

2.1 复合物结构采样

RNA-蛋白质复合物对接方法的精度低于蛋白质-蛋白质复合物结构预测方法, 一方面由于大多数方法最初是为蛋白质-蛋白质对接而开发的[25–30], 并未充分考虑RNA的结构特征. 另一方面由于RNA结构具有柔性, 在对接过程中将产生一定的构象变化[23,31]. RNA-蛋白质复合物对接采样方法可以分为刚性对接和柔性对接两类(表1). 在RNA-蛋白质复合物形成的过程中, RNA和蛋白质都有可能在骨架或侧链上发生构象变化. 刚性对接不考虑RNA或蛋白质的构象变化, 将两个单体视为刚性球体进行对接. 柔性对接可以考虑对接过程中微小构象变化, 产生更加接近天然态的复合物结构.

3dRPC和NPDock为针对RNA-蛋白质复合物开发的刚性对接算法. 3dRPC[23]针对RNA-蛋白质相互作用界面的几何匹配、静电互补性、核苷酸碱基与带电氨基酸芳香环相互作用特征, 生成三维正交网格模型进行旋转平移, 通过快速傅里叶变换(FFT)加速空间构象搜索, 从而快速生成大量的预测结构. NPDock首先利用几何匹配模型将单体结构视为刚体进行旋转平移, 生成大量的复合物结构[22]. 随后使用针对RNA-蛋白质的评估函数DARS-RNP与QUASI-RNP[31]进行结构评估. 该方法进一步对得分最高的诱饵结构进行聚类, 选择三个最大聚类的代表结构为候选结构. 最后, 再通过蒙特卡罗模拟退火优化RNA-蛋白质相互作用, 从而得到最终的预测结构. 刚性对接可以在较短的时间内生成成千上万的候选结构, 但没有考虑复合物的柔性变化, 较难得到精确的RNA-蛋白质近天然态结构.

HDOCK, RNP-denovo, P3DOCK和MPRDock为柔性对接算法. HDOCK[33]可以同时输入结构与序列信息, 通过HHSuite[37]对PDB序列数据库进行序列相似性搜索, 利用同源序列的相似结构进行单体结构建模来考虑蛋白质的柔性, 最后进行分子对接得到候选复合物结构. P3DOCK[36]是一种基于模板对接算法PRIME[38,39]和刚性对接算法3dRPC[23]的组合对接协议, 首先通过PRIME和3dRPC分别生成诱饵结构, 进一步整合诱饵结构并聚类得到最终的预测结构. MPRDock[35]通过多种蛋白质构象的集合来考虑对接中蛋白质的柔性. 在对接过程中, MPRDock将RNA与集合中的每个蛋白质结构对接, 然后进行聚类分析从而得到候选复合物结构. RNP-denovo[34]在分子对接的同时进行RNA折叠, 考虑在蛋白质影响下RNA折叠的构象变化. RNP-denovo的折叠过程通过Rosetta实现,

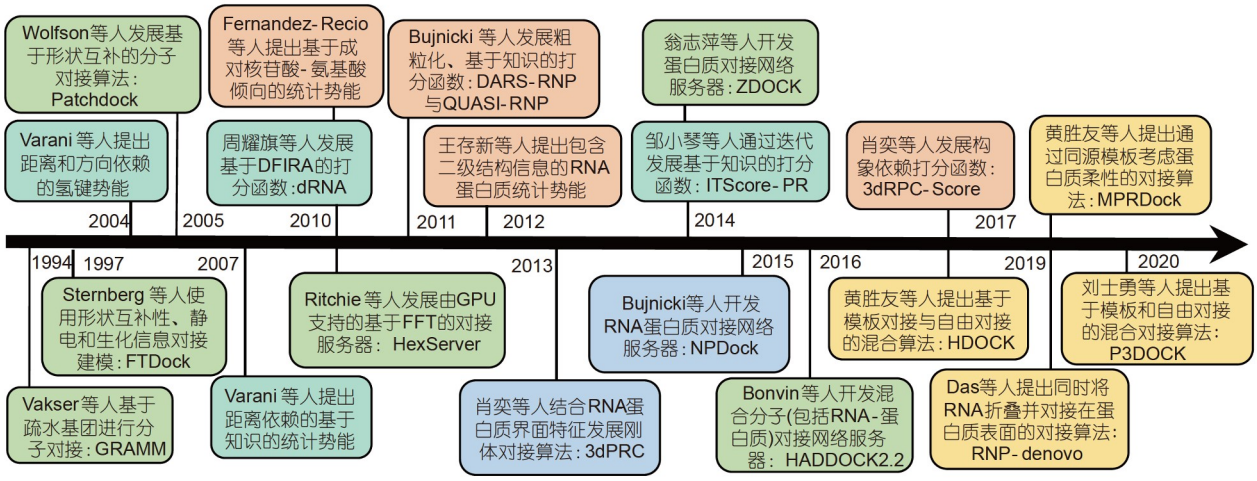


图 1 (网络版彩图) RNA-蛋白质复合物结构预测发展时间轴. 蛋白质-蛋白质对接程序拓展方法(绿色)、刚性对接方法(蓝色)、柔性对接方法(黄色)、基于残基层次的能量评估函数(红色)和基于原子层次的能量评估函数(青色)

Figure 1 (Color online) Timeline of RNA-protein complex structure prediction development: the prediction methods developed from protein-protein docking (green), rigid docking (blue), flexible docking (yellow), residue-level scoring function (red), and atomic-level scoring function (cyan).

表 1 RNA-蛋白质复合物对接方法

Table 1 The RNA-protein complex docking methods

对接程序	专为RNA-蛋白质对接设计	对接方式	柔性考虑方式	可用性		参考文献
				网络服务器	独立程序	
GRAMM	×	刚性	N/A	√	√	[25]
FTDock	×	刚性	N/A	×	√	[26]
Patchdock	×	刚性	N/A	√	√	[27]
HexServer	×	刚性	N/A	√	√	[28]
3dRPC	√	刚性	N/A	√	√	[23,24,32]
ZDOCK	×	刚性	N/A	√	√	[29]
NPdock	√	刚性	N/A	√	×	[22]
HADDOCK	×	柔性	RNA/蛋白质结构建模采样	√	√	[30]
HDOCK	√	柔性	蛋白质结构建模采样	√	×	[33]
RNP-denovo	√	柔性	RNA动力学结构采样	×	√	[34]
MPRDock	√	柔性	蛋白质结构建模采样	√	×	[35]
P3DOCK	√	柔性	RNA/蛋白质结构建模采样	√	×	[36]

通过改进后的FARNA算法计算RNA-蛋白质相互作用合理性[40].

RNA-蛋白质复合物结构采样计算中的瓶颈问题在于单体结构在结合前后会发生构象变化, RNA单体的结构变化更为明显, 分子对接过程中需要充分考虑RNA结构的柔性. 目前有一些方法考虑了蛋白质在对接过程中的柔性, 而如何考虑RNA的柔性依然是尚未解决的难题. 另外, 基于相互作用位点的分子对接可以极大提高采样效率, 得到更加精确的近天然态结构. 综合考虑RNA在结合过程中复杂的构象变化成为RNA-蛋白质复合物结构采样的关键.

2.2 复合物结构评估

RNA-蛋白质复合物结构评估方法聚焦于对对接采样产生的所有结构进行打分排名, 筛选出排名靠前的结构作为近天然态结构. 表2为RNA-蛋白质复合物结构评估的代表性方法, 残基水平上粗粒化的打分函数和原子水平上全原子的打分函数都主要分为基于力场的和基于知识的两大类. 基于力场的打分函数通常计算几个与物理相关能量项的加权和, 如分子间静电力项、范德华力项、氢键项和去溶剂化项等. 基于知识的打分函数通过反玻尔兹曼关系将距离依赖的成对接触概率分布转化为统计势能函数, 这也是目前使用最多的方法.

基于知识的评分方法主要分为倾向和参考态两类. 倾向是在RNA-蛋白质相互作用的情况下, 某些化学结构在相互作用界面上发生的趋势的度量. 而参考状态问题一直是构建成对、距离依赖的统计势能的关键所在, 成对核苷酸-氨基酸在距离上的独立分布可以通过参考状态的有效定义进行修正, 如平均参考态、准化学近似参考态与有限理想气体参考态等. 基于知识的评分方法都依赖于反玻尔兹曼分布构建统计势能函数:

$$E=\sum_{ij}E_{ij}=-RT\sum_{ij}\ln(P_{ij}), \tag{1}$$

其中, E表示总能量, E_{ij}表示成对接触i-j (下文将用i-j表示核苷酸-氨基酸对, p-q表示原子对)的能量, R为摩尔气体常数, T为温度, P_{ij}表示成对接触i-j的概率分布.

巴塞罗那超级计算中心(BSC) Fernandez-Recio团队[42]开发了基于核苷酸-氨基酸成对距离依赖的粗粒化统计势:

$$P_{ij}^{I}=\frac{N_{ij}^{I}/\sum_{ij}N_{ij}^{I}}{\left(N_{i}^{S}/\sum_{i}N_{i}^{S}\right)\times\left(N_{j}^{S}/\sum_{j}N_{j}^{S}\right)}, \tag{2}$$

其中, N_{ij}^{I}为界面处核苷酸i与氨基酸j的统计对数量, $\sum_{ij}N_{ij}^{I}$为界面处核苷酸与氨基酸的统计对总数, 而N_{i}^{S}

表 2 RNA-蛋白质复合物构象评估函数

Table 2 RNA-protein complex scoring functions

打分函数	方法	特点	类型	独立于对接程序使用	参考文献
Varani等人的氢键统计势能	基于力场的函数	氢键统计势能	全原子	×	[13]
Varani等人的基于知识的统计势能	基于知识的函数	原子对距离依赖	全原子	×	[41]
Fernandez-Recio等人的统计势能	基于知识的函数	核苷酸-氨基酸成对倾向	粗粒化	×	[42]
DRNA	基于知识的函数	体积分数校正DFIRE参考态	全原子	×	[14]
DARS-RNP与QUASI-RNP	基于知识的函数	准化学近似和诱饵参考态	粗粒化	√	[31]
王存新等人的统计势能	基于知识的函数	二级结构与 核苷酸-氨基酸成对倾向	粗粒化	×	[43]
ITScore-PR	基于知识的函数	原子对距离依赖	全原子	√	[44]
3dRPC-Score	基于知识的函数	核苷酸-氨基酸成对倾向	粗粒化	√	[24]

与N_j^S分别表示表面上核苷酸i与氨基酸j的数量, $\sum_i N_i^S$与$\sum_j N_j^S$分别表示表面上核苷酸与氨基酸的总数, 通过反玻尔兹曼分布得到核苷酸i与氨基酸j对的统计势能:

$$E_{ij} = -RT\ln\left(P_{ij}^I\right). \tag{3}$$

相互作用界面处所有氨基酸与核苷酸对的统计势能的总和作为给定RNA-蛋白质复合物的能量分数:

$$E = \sum_{ij} E_{ij} = -RT\sum_{ij}\ln\left(P_{ij}^I\right). \tag{4}$$

国际分子与细胞生物学研究所Tuszynska和Bujnicki[31]发展了基于粗粒化模型的统计势评估函数: DARS-RNP (以诱饵结构作为参考态)和QUASI-RNP (基于准化学势参考态). DARS-RNP和QUASI-RNP根据氨基酸大小, 采用联合原子模型来表示氨基酸, 并由两个分别位于磷酸基团和核糖中的联合原子来表示核苷酸主链, 而嘧啶和嘌呤则分别用一个原子和两个原子表示. DARS-RNP和QUASI-RNP的统计势函数也是基于反玻尔兹曼统计生成, 这两个评估函数的总能量E由基于距离的能量项(E_d)、基于角度的能量项(E_a)、基于位点的能量项(E_s)和碰撞惩罚项(E_p)构成:

$$E = E_d + E_a + E_s + E_p, \tag{5}$$

其中, E_d, E_a, E_s和E_p被赋予相同的权重, 相互作用能量由下式计算生成:

$$E_{pq}(r) = -RT\ln\frac{N_{\rm obs}(p,q,r)}{N_{\rm exp}(p,q,r)}, \tag{6}$$

其中, $E_{pq}(r)$表示E_d, E_a, E_s和E_p中的任一能量, $N_{\rm obs}(p,q,r)$是在训练集中统计观察到的距离角度比为r时联合原子p, q的统计对数量. 而$N_{\rm exp}(p,q,r)$是在参考状态下距离角度比为r时的预期统计对数量. 基于准化学近似参考态的QUASI-RNP由残基的摩尔分数所计算:

$$N_{\rm exp}(p,q,r) = X_p * X_q * N_{\rm obs}(r), \tag{7}$$

其中, X_p和X_q分别为联合原子p和q的摩尔分数, 而$N_{\rm obs}(r)$指在距离角度比为r时所有统计对的数量. DARS-RNP中的$N_{\rm exp}(p,q,r)$是一组诱饵结构中联合原子p和q之间的归一化统计对数量, 因此可以看作随机模型.

华中科技大学肖奕教授团队[24]开发的3dRPC-Score认为仅用距离与角度来评估核苷酸-氨基酸对之间的能量不准确, 还需要考虑核苷酸-氨基酸对之间的相对距离和方向. 3dRPC-Score以核苷酸-氨基酸对的构象作为统计变量, 通过核苷酸-氨基酸对之间的相对RMSD (均方根偏差)进行分类, 认为同一类中的核苷酸-氨基酸对拥有相同的能量, 统计势能表示为

$$E_{ij}(C) = -\ln\frac{P_{ij}(C)}{P_i P_j * P_v}, \tag{8}$$

其中, $P_{ij}(C)$指C类中核苷酸i与氨基酸j对的统计概率, $P_i\left(P_j\right)$指相互作用界面中核苷酸i (氨基酸j)的概率, P_v是理想状态下C类核苷酸-氨基酸对在整个构象空间中出现的概率. 其中界面是否含有核苷酸(氨基酸)是通过溶液可及表面积确定. 若RNA (蛋白质)单体与复

合物中的核苷酸(氨基酸)的溶液可及表面积不同, 则认为该核苷酸(氨基酸)位于相互作用界面上. 在理想状态下, 每一类核苷酸-氨基酸对在构象空间中出现的概率都相同, 因此,

$$E_{ij}(C) = -\ln\frac{P_{ij}(C)}{P_iP_j} + \text{constant}, \tag{9}$$

其中, constant = $\ln P_v$. 而北京工业大学王存新团队[43]提出的统计势能纳入了RNA与蛋白质二级结构偏好信息, 考虑界面处不同二级结构中核苷酸与氨基酸对的差异. 该方法在残基水平上使用基于粗粒化模型的函数考虑核苷酸-氨基酸的成对相互作用, 对构象变化不敏感. 因此, 与大多数粗粒化模型一样, 该方法很难考虑由于柔性引起的微小构象变化, 在精细化结构的辨别效果不好.

基于全原子的结构评估函数在近天然态结构的评估中优于粗粒化模型. 华盛顿大学Varani团队[13]于2004年开发了基于RNA-蛋白质氢键相互作用的统计势能. 然而氢键只代表RNA-蛋白质界面中复杂且多样相互作用中的一种, 因此Varani团队[41]于2007年开发了全原子、距离依赖的统计势能. 深圳湾实验室周耀旗研究员[14]利用有限理想气体参考态(DFIRE)构建了距离依赖的评估函数DRNA, RNA-蛋白质相互作用能量函数$E_{pq}(r)$为

$$E_{pq}(r) = \begin{cases} -\eta\ln\dfrac{N_{\text{obs}}(p,q,r)}{\left(\dfrac{f_p^{\nu}(r)f_q^{\nu}(r)}{f_p^{\nu}(r_{\text{cut}})f_q^{\nu}(r_{\text{cut}})}\right)^{\beta}\dfrac{r^{\alpha}\Delta r}{r_{\text{cut}}^{\alpha}\Delta r_{\text{cut}}}N_{\text{obs}}^{lc}(p,q,r_{\text{cut}})}, \\ \quad r < r_{\text{cut}}, \\ 0, r \geqslant r_{\text{cut}}, \end{cases} \tag{10}$$

其中, 体积分数因子 $f_p^{\nu}(r) = \sum_q N_{\text{obs}}^{\text{Protein-RNA}}(p,q,r)/\sum_q N_{\text{obs}}^{\text{All}}(p,q,r)$, $N_{\text{obs}}(p,q,r)$为在给定结构数据库中观察到距离r处p-q原子对数量, p-q原子对距离r按Δr_{cut}=0.5 Å截断, r_{cut}=15 Å为相互作用最大截断距离. 而α是由r^{α}与有限蛋白质球体中理想气体点的实际距离拟合确定, 体积校正采用β=0.5的精确值.

密苏里大学哥伦比亚分校邹小琴团队[44]于2014年发展了基于迭代模型的统计势ITScore-PR, 通过比较训练集中诱饵原子对与天然态结构原子对之间的差异来计算统计势

$$E_{pq}^{(n+1)}(r) = E_{pq}^{(n)}(r) + \Delta E_{pq}^{(n)}(r), \tag{11}$$

$$\Delta E_{pq}^{n}(r) = \frac{1}{2}k_{\text{B}}T\left[g_{pq}^{(n)}(r) - g_{pq}^{\text{obs}}(r)\right], \tag{12}$$

其中, n为迭代次数, $g_{pq}^{(n)}(r)$和$g_{pq}^{\text{obs}}(r)$分别表示基于$E_{pq}^{(n)}(r)$和天然态结构所计算的径向分布函数. 迭代过程通过式(11)和(12)进行, 直到训练集中的所有天然态结构都可以与诱饵结构区分开来.

RNA-蛋白质复合物结构评估在刚性对接的测试集中取得了重大成功, 在柔性对接测试集中的表现依然有待提高. 基于粗粒化和全原子的统计势模型仅考虑RNA-蛋白质复合物静态结构的相互作用, 并未考虑RNA在结合前后的构象变化与复杂的多体相互作用. RNA结合口袋以及氢键网络等RNA-蛋白质复合物结构特征已经表明原子或残基间非线性多体相互作用对复合物有不可忽视的作用[45–48]. 如何考虑RNA分子结合前后的构象变化与RNA-蛋白质间多体相互作用是目前RNA-蛋白质结构评估的瓶颈问题.

3 RNA-蛋白质结构预测的思考与展望

RNA-蛋白质复合物结构预测在刚性对接和较为简单的复合物测试中结果较好, 但在柔性对接和复杂的复合物测试中表现依然有待提高. 下面对如何进一步优化复合物结构预测中的结构采样与结构评估过程, 提高结构预测精度进行探讨.

3.1 基于RNA相互作用位点预测的结构采样优化

RNA相互作用位点信息可以缩小复合物构象搜索空间, 提高结构采样效率和预测精度. 本课题组于2018年提出了基于复杂网络模型的RNA相互作用位点预测方法RBind[49]. 该方法将RNA结构中的核苷酸定义为网络中的节点, 核苷酸之间的非共价相互作用距离小于8 Å定义为网络中的边, 通过计算网络中节点的接近度(全局特征)和度(局部特征)来确定相互作用位点. 网络中接近度的计算公式为

$$C(x) = \frac{n-1}{\sum d(x,y)}, \tag{13}$$

其中, $d(x,y)$是节点x与其他任意节点y之间的最短路径

的距离. 当节点的接近度和度都高于相应的截断阈值时, 该节点为RNA与配体结合的相互作用位点. RBind在RNA-配体测试集中的平均预测精度为0.82, RNA-蛋白质测试集的平均预测精度为0.63, 可以有效鉴定RNA的相互作用位点, 有较高的预测精度和灵敏度. RNA在结合蛋白质的过程中有一定的构象变化, 通过RNA结构建模采样了RNA的柔性结构, 评估了柔性结构对RNA相互作用位点预测的影响. 结果表明, RNA结构的构象变化较小时, RBind对RNA相互作用位点进行了可靠预测. 如果RNA结构的构象变化较大, RBind的准确性会相应下降. RBind的假阳性预测位点大多数位于结合口袋附近, 会减小假阳性预测位点对复合物结构采样精度的影响.

对RNA相互作用位点用于指导复合物结构采样优化进行了可行性分析(图2). 例如, tRNA-蛋白质复合物结构(PDB ID: 2ZUE)为火球菌精氨酸-tRNA合成酶与tRNA (Arg)和ATP类似物(ANP)复合的晶体结构. 该复合物结构为柔性采样问题, 蛋白质来自结构2ZUE中的A链, RNA来自结构2ZUF的B链. 利用3dRPC进行结构采样并生成1000个复合物结构, 其中排名前50的采样结构的RMSD平均为17.26 Å. RBind对复合物中tRNA结构的相互作用位点预测结果为14个正确相互作用位点(G906, G907, U908, A909, C912, U913, A914, G915, C916, A921, G923, G924, C948, C949)与一个假阳性相互作用位点(G946), 预测精度为93.0%. 加入RNA相互作用预测位点信息后, 复合物结构采样排名前50结构的RMSD平均为9.57 Å. 结果表明, RNA相互作用位点预测对RNA-蛋白质复合物结构采样的效率有较大提升, 提高了44.55%. 同时, 在对接过程中可以在相互作用位点附近进行局部增强采样, 获得更多的近天然态结构, 也可以在相互作用位点附近采样到足够多的结构后提前中止采样, 从而加速构象采样.

3.2 基于RNA口袋拓扑信息的结构评估

RNA口袋拓扑信息可以描述RNA结合蛋白质的关键作用位点, 提高RNA-蛋白质结构评估精度. 本课

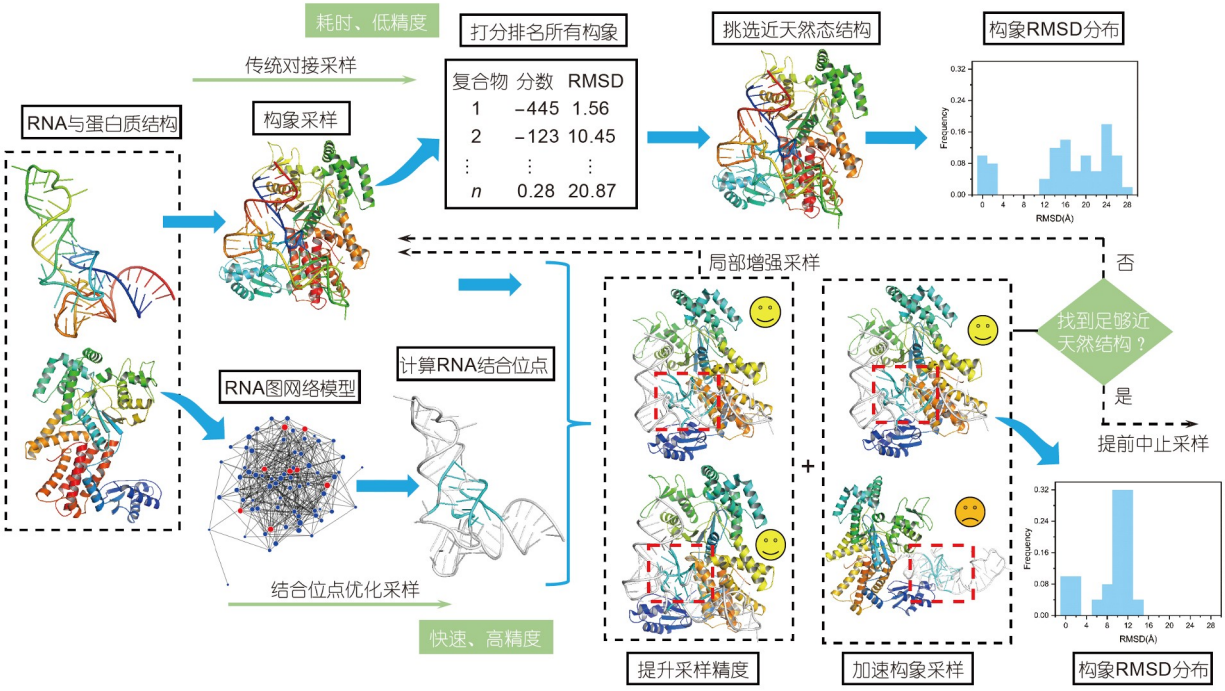


图 2 (网络版彩图)结合位点优化构象采样示意图. 传统的RNA-蛋白质复合物对接方法通过旋转平移采样复合物结构, 采样效率与精度较低. RBind可以先预测RNA结构的结合位点, 加入结合位点的约束信息可以加快RNA-蛋白质复合物对接的采样效率与提高采样精度

Figure 2 (Color online) Optimizing structure sampling by RNA binding sites. The traditional RNA-protein complex docking method generates complex structures through rotation and translation, resulting in low sampling efficiency and accuracy. RBind first predicts the RNA binding sites and then generates complex structures by binding site constraint information, improving the sampling efficiency and RNA-protein complex sampling accuracy.

题组于2022年发展了基于RNA口袋拓扑信息的RNA-蛋白质相互作用资源数据库RPpocket[50]，该方法从核苷酸-氨基酸相互作用、二级结构相互作用和拓扑结构三个维度系统讨论分析了RNA-蛋白质的结构特征(图3(a)–(d)). RPpocket首先分析了RNA-蛋白质相互作用界面上的片段序列，发现了11个RNA与44个蛋白质序列片段出现较多，其中“CG”和“KR”片段分别在RNA与蛋白质中出现频率最高. RNA中核苷酸“G”为双环侧链的化学结构，更容易与氨基酸发生相互作用，在相互作用的界面分布(31.14%)明显高于C (26.41%)，A (21.42%)和U (21.03%). RNA与蛋白质结合过程中首先通过静电相互作用远程识别分子，然后再利用短程相互作用(疏水或氢键等)优化和固定RNA-蛋白质复合物结构. 在二级结构相互作用方面，茎区是RNA二级结构中最常见的相互作用单元，其次是发卡环、内环、单链、凸起和多分支环结构. 蛋白质的二级结构中最常见的相互作用单元为环区，其次是螺旋区和β片区.

本课题组进一步发展了RPDescriptor定量计算RNA口袋的拓扑结构特征[51]. 该方法首先将RNA口袋置于三维空间网格, 长a (Å)、宽b (Å)、高c (Å)的三维空间被划分为n (n=a×b×c)个大小为1 Å的网格, 每个网格$F(i, j, k)$由1或0表示, RNA口袋结构由值为1的网格构成. 然后根据主转动惯量的归一化比值(NPRs)得到形状描述符rpd_1与rpd_2，定量计算RNA口袋空间拓扑形状:

$$rpd_1=\frac{I_{11}}{I_{33}},\ rpd_2=\frac{I_{22}}{I_{33}}, \tag{14}$$

其中, I_{11}, I_{22}, I_{33}分别为RNA口袋坐标归一化转动惯量. 通过投影rpd_1与rpd_2到二维平面定量计算RNA口袋拓扑形状，按照到几何中心的分布将RNA口袋分为: 球状、圆盘状与杆状. 并通过形状相似度分数对RNA口袋定量分类:

$$\begin{aligned} s_1&=rpd_1+rpd_2-1,\\ s_2&=2-2\times rpd_2,\\ s_3&=rpd_2-rpd_1, \end{aligned} \tag{15}$$

其中，s_1，s_2和s_3分别表示RNA口袋的球状、圆盘状与杆状程度. 在RNA-蛋白质复合物的统计中, RNA结合口袋的平均体积与表面积分别为1607.03 Å^{3}与

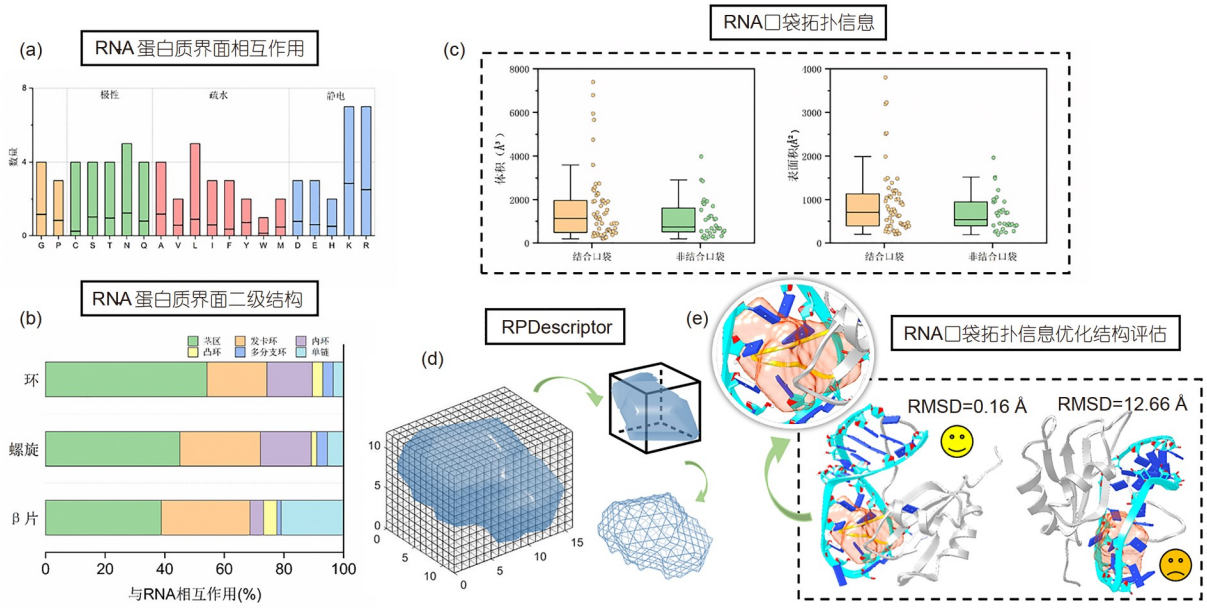


图 3 (网络版彩图) RNA口袋拓扑信息优化结构评估示意图. (a)–(c) RPpocket[50]统计分析了RNA-蛋白质复合物结构中核苷酸-氨基酸相互作用、二级结构相互作用和口袋拓扑结构信息. (d) RPDescriptor可以定量计算RNA口袋结构的拓扑构象. (e) RNA口袋拓扑信息可以优化RNA-蛋白质复合物结构评估, 提高结构预测精度

Figure 3 (Color online) Optimizing structure evaluation by RNA pocket topology information. (a)–(c) RPpocket [50] statistically analyzes the nucleotide-amino acid interaction, secondary structure interaction, and pocket topology information of RNA-protein complex structure. (d) RPDescriptor can quantitatively calculate the topological conformations of RNA pockets. (e) RNA pocket topology information can optimize the structure evaluation of RNA-protein complexes and improve structure prediction accuracy.

923.77 Å^2, RNA非结合口袋的平均体积与表面积分别为1122.91 Å^{3}与678.94 Å^2. RNA结构较为柔性, 在结合过程中容易受到蛋白质结构的影响并被拉伸.

RNA口袋拓扑信息对RNA-蛋白质复合物结构评估有重要的意义. 例如, microRNA 20b前置体与蛋白质Rbfox RRM会结合形成复合物结构(PDB ID: 2N82). 用3dRPC对该结构采样并生成100个复合物结构. 在3dRPC-Score的结构评估中, 排名前5的结构中有3个近天然态结构, 另外两个结构(RMSD分别为12.66与13.49 Å)被错误识别为近天然态结构, 这两个复合物结构的蛋白质未结合在RNA的口袋中. 如果加上RNA口袋拓扑信息约束, 可以采样到精确的近天然态结构(RMSD为0.16 Å), 有效去除假阳性预测结果(图3(e)).

3.3 基于深度学习的RNA-蛋白质复合物结构评估

近几年来, 深度学习被广泛应用于生物大分子结构预测的生物物理问题, 尤其在蛋白质结构预测和RNA结构评估上取得了重大突破. 例如, DeepMind团队[52]开发了“端到端”的神经网络模型Alphafold2, 该模型采用了图网络加上注意力机制的深度学习策略, 利用序列共进化分析将蛋白质的结构信息整合到深度学习算法中, 在CASP14 (第十四届国际蛋白质结构预测技术评估大赛)中达到了92.4%的成功率, 极大推动了蛋白质结构预测的进步, 被誉为“2022十大科技突破”之一. 美国斯坦福大学Das团队[53]发展了RNA结构评估模型ARES, 该方法利用卷积神经网络对18个已知RNA结构的原子坐标和化学元素类型等信息进行训练, 可以有效识别RNA空间结构. ARES首先利用FARFAR2对RNA结构盲测挑战中的分子进行重新建模, 再在候选结构中进行评估筛选, 测试结果优于基于统计势能的评分函数. 武汉大学谭志杰团队[54]于2022年发展了基于统计势能的RNA单体结构评估方法rsRNASP, 测试集结果表明传统的基于统计势能的评分函数(包括rsRNASP, 3dRNAscore和DFIRE-RNA等)在大多数测试集中的结果明显优于ARES. ARES的训练集与测试集都来自于FARFAR2, 可能预测结果具有较强的偏好性. 南京大学王炜团队[55]发展了基于卷积神经网络的RNA结构评估模型RNA3DCNN, 该方法将RNA中的每个核苷酸及其周围环境转化为由体素组成的3D图像, 尺寸大小32 Å×32 Å×32 Å的盒子中的每个网格由其中原子的占有数、质量与电荷的总和所表示, 利用卷积神经网络学习RNA中每个核苷酸的空间特征. 结果表明, 深度学习可以有效捕获RNA与蛋白质单体的三维空间结构特征, 在蛋白质结构预测与RNA结构评估中取得惊人成功.

RNA-蛋白质复合物结构评估和蛋白质或RNA单体分子不同, 依然面临重大挑战. (1) 结构数据库中RNA-蛋白质复合物的三级结构数据仍然十分有限[18,19,56], 较难通过深度学习进行大数据分析. Alphafold2[52]在蛋白质结构预测的成功在于庞大的蛋白质结构训练集, 目前很难拓展到RNA-蛋白质复合物结构预测. (2) RNA-蛋白质复合物界面特征更加复杂. RNA主要通过静电相互作用、碱基堆积相互作用、范德华相互作用与氢键等相互作用与蛋白质结合. RNA-蛋白质界面的原子堆积比蛋白质-蛋白质复合物界面的原子堆积更松散[23]. RNA分子主链上每个磷酸基团都带有一个电子单位的负电荷, 将导致带正电氨基酸更易出现在RNA-蛋白质界面[42,46,57]. 同时核苷酸的碱基将与带电氨基酸芳香环在相互作用界面处发生π-π堆积作用[48]. 蛋白质表面形状(凹陷或凸起)与RNA碱基配对(成对或未成对的核苷酸)之间也有选择偏好性, 例如凹陷的蛋白质表面更喜欢界面处未配对的核苷酸而不是成对核苷酸[58]. (3) RNA分子存在较大柔性, 在与蛋白质结合过程中会发生较大的构象变化. 例如转录因子NF-κB二聚体体系在结合前后RNA的RMSD为5.4 Å[59]. RNA-蛋白质在识别过程中的构象变化将导致结构评估函数很难精确确定近天然态复合物结构. 最近, 我们课题组发展了基于深度学习的RNA-蛋白质复合物结构评估模型DRPScore[60], 该方法聚焦于RNA-蛋白质相互作用界面, 通过4D卷积神经网络模型不仅可以有效学习原子质量、原子电荷、相互作用类型(静电相互作用、氢键等)和相互作用距离的局部特征, 还可以学习RNA-蛋白质复合物二级结构相互作用(RNA为配对、发卡环和内环等; 蛋白质为螺旋和beta折叠单元)等全局特征. DRPScore在不同柔性程度的测试集上进行了广泛评估, 在RNA-蛋白质复合物结构预测的刚性对接测试集中成功率为91.67%, 在柔性对接测试集中的成功率为56.14%, 仍有较大的提升空间. 我们相信随着实验结构的增加和深度学习模型的发展, 基于深度学习的RNA-蛋白质结构评估在未来几年里会得到较大发展.

4 结语

本文综述了RNA-蛋白质复合物结构预测的研究进展与挑战，讨论了结合位点预测优化结构采样和口袋拓扑结构优化结构评估的可行性，最后对基于深度学习模型的RNA-蛋白质复合物结构评估进行了展望. RNA-蛋白质复合物结构预测是目前软物质物理和生物物理学领域最具挑战性和影响力的研究方向之一. 随着计算预测精度的不断提高，结构预测方法已经逐渐成为实验确定RNA-蛋白质结构的替代手段，成为分子生物学家越来越青睐的工具. 虽然RNA-蛋白质复合物结构预测在刚性对接上较为成功，柔性对接和复杂的复合物预测依然面临挑战. 数据库中RNA-蛋白质结构数据较少、RNA-蛋白质界面复杂以及RNA的结构柔性制约了其复合物结构预测的精度. 我们相信随着实验结构与计算模型的不断发展，RNA-蛋白质复合物结构预测方法将在近几年里获得颠覆性的突破.

参考文献

1 Chung C S, Tseng C K, Lai Y H, et al. Dynamic protein-RNA interactions in mediating splicing catalysis. Nucl Acids Res, 2019, 47: 899–910
2 Glisovic T, Bachorik J L, Yong J, et al. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett, 2008, 582: 1977–1986
3 Licatalosi D D, Darnell R B. RNA processing and its regulation: Global insights into biological networks. Nat Rev Genet, 2010, 11: 75–87
4 Lunde B M, Moore C, Varani G. RNA-binding proteins: Modular design for efficient function. Nat Rev Mol Cell Biol, 2007, 8: 479–490
5 Mittal N, Roy N, Babu M M, et al. Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. Proc Natl Acad Sci USA, 2009, 106: 20300–20305
6 Müller-McNicoll M, Neugebauer K M. How cells get the message: Dynamic assembly and function of mRNA-protein complexes. Nat Rev Genet, 2013, 14: 275–287
7 Liu X Y, Tsubota K, Yu Y, et al. A numerical method to predict the membrane tension distribution of spreading cells based on the reconstruction of focal adhesions. Sci China-Phys Mech Astron, 2022, 65: 264612
8 Lukong K E, Chang K, Khandjian E W, et al. RNA-binding proteins in human genetic disease. Trends Genet, 2008, 24: 416–425
9 Cooper T A, Wan L, Dreyfuss G. RNA and disease. Cell, 2009, 136: 777–793
10 Modic M, Ule J, Sibley C R. CLIPing the brain: Studies of protein-RNA interactions important for neurodegenerative disorders. Mol Cell Neurosci, 2013, 56: 429–435
11 De Conti L, Baralle M, Buratti E. Neurodegeneration and RNA-binding proteins. WIREs RNA, 2017, 8: e1394
12 Khalil A M, Rinn J L. RNA-protein interactions in human health and disease. Semin Cell Dev Biol, 2011, 22: 359–365
13 Chen Y, Kortemme T, Robertson T, et al. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. Nucl Acids Res, 2004, 32: 5147–5162
14 Zhao H, Yang Y, Zhou Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucl Acids Res, 2011, 39: 3017–3025
15 Zhao H, Yang Y, Zhou Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. RNA Biol, 2011, 8: 988–996
16 Ke A. Crystallization of RNA and RNA-protein complexes. Methods, 2004, 34: 408–414
17 Khatter H, Myasnikov A G, Natchiar S K, et al. Structure of the human 80S ribosome. Nature, 2015, 520: 640–645
18 Berman H M, Olson W K, Beveridge D L, et al. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J, 1992, 63: 751–759
19 Coimbatore Narayanan B, Westbrook J, Ghosh S, et al. The Nucleic Acid Database: New features and capabilities. Nucl Acids Res, 2014, 42: D114–D122
20 Arnautova Y A, Abagyan R, Totrov M. Protein-RNA docking using ICM. J Chem Theor Comput, 2018, 14: 4971–4984
21 Pérez-Cano L, Romero-Durana M, Fernández-Recio J. Structural and energy determinants in protein-RNA docking. Methods, 2017, 118-119: 163–170
22 Tuszynska I, Magnus M, Jonak K, et al. NPDock: A web server for protein-nucleic acid docking. Nucl Acids Res, 2015, 43: W425–W430

23 Huang Y, Liu S, Guo D, et al. A novel protocol for three-dimensional structure prediction of RNA-protein complexes. Sci Rep, 2013, 3: 1887

24 Li H, Huang Y, Xiao Y. A pair-conformation-dependent scoring function for evaluating 3D RNA-protein complex structures. PLoS ONE, 2017, 12: e0174662

25 Tovchigrechko A, Vakser I A. GRAMM-X public web server for protein-protein docking. Nucl Acids Res, 2006, 34: W310–W314

26 Gabb H A, Jackson R M, Sternberg M J E. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol, 1997, 272: 106–120

27 Schneidman-Duhovny D, Inbar Y, Nussinov R, et al. PatchDock and SymmDock: Servers for rigid and symmetric docking. Nucl Acids Res, 2005, 33: W363–W367

28 Macindoe G, Mavridis L, Venkatraman V, et al. HexServer: An FFT-based protein docking server powered by graphics processors. Nucl Acids Res, 2010, 38: W445–W449

29 Pierce B G, Wiehe K, Hwang H, et al. ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics, 2014, 30: 1771–1773

30 van Zundert G C P, Rodrigues J P G L M, Trellet M, et al. The HADDOCK2.2 web server: User-friendly integrative modeling of biomolecular complexes. J Mol Biol, 2016, 428: 720–725

31 Tuszynska I, Bujnicki J M. DARS-RNP and QUASI-RNP: New statistical potentials for protein-RNA docking. BMC Bioinf, 2011, 12: 348

32 Huang Y, Li H, Xiao Y. Using 3dRPC for RNA-protein complex structure prediction. Biophys Rep, 2016, 2: 95–99

33 Yan Y, Zhang D, Zhou P, et al. HDOCK: A web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucl Acids Res, 2017, 45: W365–W373

34 Kappel K, Das R. Sampling native-like structures of RNA-protein complexes through rosetta folding and docking. Structure, 2019, 27: 140–151. e5

35 He J, Tao H, Huang S Y. Protein-ensemble-RNA docking by efficient consideration of protein flexibility through homology models. Bioinformatics, 2019, 35: 4994–5002

36 Zheng J, Hong X, Xie J, et al. P3DOCK: A protein-RNA docking webserver based on template-based and template-free docking. Bioinformatics, 2020, 36: 96–103

37 Remmert M, Biegert A, Hauser A, et al. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods, 2011, 9: 173–175

38 Zheng J, Kundrotas P J, Vakser I A, et al. Template-based modeling of protein-RNA interactions. PLoS Comput Biol, 2016, 12: e1005120

39 Zheng J, Xie J, Hong X, et al. RMalign: An RNA structural alignment tool based on a novel scoring function RMscore. BMC Genomics, 2019, 20: 276

40 Cheng C Y, Chou F C, Das R. Modeling complex RNA tertiary folds with Rosetta. Methods Enzymol, 2015, 553: 35–64

41 Zheng S, Robertson T A, Varani G. A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. FEBS J, 2007, 274: 6378–6391

42 Perez-Cano L, Solernou A, Pons C, et al. Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. In: Biocomputing 2010: Proceedings of the Pacific Symposium. Kamuela, 2010. 293–301

43 Li C H, Cao L B, Su J G, et al. A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. Proteins, 2012, 80: 14–24

44 Huang S Y, Zou X. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. Nucl Acids Res, 2014, 42: e55

45 Morozova N, Allers J, Myers J, et al. Protein-RNA interactions: Exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. Bioinformatics, 2006, 22: 2746–2752

46 Jones S. Protein-RNA interactions: A structural analysis. Nucl Acids Res, 2001, 29: 943–954

47 Treger M, Westhof E. Statistical analysis of atomic contacts at RNA-protein interfaces. J Mol Recognit, 2001, 14: 199–214

48 Allers J, Shamoo Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. J Mol Biol, 2001, 311: 75–86

49 Wang K, Jian Y, Wang H, et al. RBind: Computational network method to predict RNA binding sites. Bioinformatics, 2018, 34: 3131–3136

50 Yang R, Liu H, Yang L, et al. RPpocket: An RNA-protein intuitive database with RNA pocket topology resources. Int J Mol Sci, 2022, 23: 6903

51 Zhou T, Wang H, Zeng C, et al. RPocket: An intuitive database of RNA pocket topology information with RNA-ligand data resources. BMC

BioInf, 2021, 22: 428
52 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589
53 Townshend R J L, Eismann S, Watkins A M, et al. Geometric deep learning of RNA structure. Science, 2021, 373: 1047–1051
54 Tan Y L, Wang X, Shi Y Z, et al. rsRNASP: A residue-separation-based statistical potential for RNA 3D structure evaluation. Biophys J, 2022, 121: 142–156
55 Li J, Zhu W, Wang J, et al. RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. PLoS Comput Biol, 2018, 14: e1006514
56 Berman H M. The protein data bank. Nucl Acids Res, 2000, 28: 235–242
57 Bahadur R P, Zacharias M, Janin J. Dissecting protein-RNA recognition sites. Nucl Acids Res, 2008, 36: 2705–2716
58 Iwakiri J, Tateishi H, Chakraborty A, et al. Dissecting the protein-RNA interface: The role of protein surface shapes and RNA secondary structures in protein-RNA recognition. Nucl Acids Res, 2012, 40: 3299–3306
59 Pérez-Cano L, Fernández-Recio J. Optimal protein-RNA area, OPRA: A propensity-based method to identify RNA-binding sites on proteins. Proteins, 2010, 78: 25–35
60 Zeng C, Jian Y, Vosoughi S, et al. Evaluating native-like structures of RNA-protein complexes through the deep learning method. Nat Commun, 2023, 14: 1060

Advances in RNA-protein structure prediction

ZENG ChengWei & ZHAO YunJie*

Institute of Biophysics and College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China

RNA-protein complexes are crucial biomolecules in living organisms. The determination of their tertiary structure is helpful for understanding the cellular activity processes and RNA-related diseases. However, the experimental determination of the RNA-protein structure consumes a considerable amount of time and financial resources, while its high-resolution determination remains a complex problem. Therefore, the prediction of the RNA-protein structure is considered a challenging research topic in soft matter physics and biophysics. In this paper, we first review the research progress and challenges in predicting an RNA-protein structure. Then, we list and discuss methods to optimize the complex structure sampling and evaluation to improve the prediction.

RNA-protein structure prediction, structure sampling, structure evaluation, deep learning

PACS: 87.14.E-, 87.14.gn, 87.15.bg, 88.10.gc

doi: 10.1360/SSPMA-2023-0045