PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2020,47(7):634~644
www.pibb.ac.cn

蛋白质–RNA序列结构界面偏好性及用于对接打分统计势的构建*

陆 林 刘 洋 李春华**
（北京工业大学环境与生命学部，北京 100124）

摘要 本文对来自PDB（Protein Data Bank）数据库的蛋白质-RNA复合物结构构建了非冗余非核糖体数据库（694个结构），并对此数据库统计了蛋白质和RNA序列及二级结构的界面偏好性. 结果发现，蛋白质β折叠、3_{10}-helix和RNA未配对核苷酸，尤其是未配对中空间排列不规整的核苷酸，具有显著的界面偏好性. 据此，对二级结构进行归类，建立了考虑序列和二级结构信息的60×12氨基酸-核苷酸成对偏好势，并将其作为打分函数用于蛋白质-RNA对接中近天然结构的筛选. 结果表明，该60×12统计势的打分成功率为65.77%，优于考虑蛋白质或RNA二级结构信息的统计势，及我们小组之前在251个结构上构建的60×8*统计势. 该工作有助于加深对蛋白质-RNA特异性识别的理解，可推动复合物结构预测的进展.

关键词 蛋白质-RNA相互作用，分子对接，统计势，界面偏好，二级结构
中图分类号 Q61，Q71 **DOI**：10.16476/j.pibb.2020.0004

蛋白质-RNA特异性识别和相互作用在生物体细胞生命活动（如基因的表达调控与修复、蛋白质的合成等）过程中都发挥着极其重要的作用[1]. 其相互作用的异常经常会导致很多疾病的发生，如神经系统的自身免疫疾病、炎症、血管疾病和肿瘤等[2]，因此蛋白质-RNA相互作用是当前生命科学领域的研究热点. 复合物三维结构为理解分子间相互作用提供了丰富的信息. 实验方法获取蛋白质-RNA三维结构数据非常困难，因此发展可靠的理论方法来预测复合物结构是亟待解决的问题[3-4]. 分子对接是用来预测复合物结构的主要方法之一，其过程主要包括两步：结合模式搜索和打分排序[5-7]. 打分函数负责对结合模式搜索获得的结构进行评价和排序，目的是挑选出近天然结构. 打分函数的合理设计对成功的结构预测至关重要.

目前打分函数主要分为基于知识、经验和物理能量的3种类型，其中基于知识的统计势因其计算量小鲁棒性高而备受关注[8]. 2010年，Pérez-CanoL等[9]基于282个非冗余蛋白质-RNA复合物结构统计得到了仅考虑序列信息的氨基酸-核苷酸成对偏好势. 2011年，Tuszynska和Bujnicki[10]提出了QUASI-RNP（quasi-chemical potential）和DARS-RNP（decoys as the reference state potential）两种基于联合原子模型的统计势，其主要区别在于参考态的选取不同. 之后我们课题组[11]构建了考虑分子二级结构信息的60×8*氨基酸-核苷酸成对偏好势，并在此基础上组合物理能量项发展了加权组合的打分函数. 该组合打分函数获得了好于前几种方法的打分成功率[12]，并发现长程静电势和考虑分子二级结构信息的统计势在其中起到了关键作用. Re等[13]也再次证明了分子二级结构对蛋白质-RNA特异性识别有重要贡献. 目前研究人员也在试图发现蛋白质上结合RNA的序列和结构特征[14-15]. 随着蛋白质-RNA复合物结构数据的增多，我们希望深入系统地展开对序列和结构界面偏好性的研究，并在此基础上发展出更加有效的统计势以用于蛋白质-RNA分子对接复合物结构的预测.

* 国家自然科学基金（31971180，11474013）和北京市自然科学基金（4152011）资助项目.
** 通讯联系人.
Tel.：010-67392001，E-mail：chunhuali@bjut.edu.cn
收稿日期：2020-01-05，接受日期：2020-06-09

本文构建了新的蛋白质-RNA非冗余数据库，统计分析了序列及二级结构的界面偏好性，发现RNA未配对碱基空间排列规整与否具有界面偏好性的显著差异．据此构建了60×12氨基酸-核苷酸成对统计偏好势，发现该统计势较我们之前构建的60×8*统计势在分子对接打分中有更好的执行．

1　数据与方法

1.1　非冗余非核糖体蛋白质–RNA数据库的建立

下载了截止到2017年7月PDB（Protein Data Bank）数据库（http：//www.rcsb.org/pdb）中所有的蛋白质-RNA复合物结构，共2 022个．去掉740个核糖体后剩余1 282个结构．去除的原因有两个：一是此类结构属于永久性结合的复合物，不同于其他结构[16]；二是该类复合物中分子识别通常不具有特异性，其蛋白质常常与RNA的双链部分结合，形成非特异性骨架相互作用来维持复合物的稳定[17-18]．之后对数据进行去冗余处理，将复合物中蛋白质和RNA序列相似性都分别高于70%和90%的复合物归为一簇．序列相似性分析采用blast++软件（https：//blast.ncbi.nlm.nih.gov/Blast.cgi）．簇的代表结构为X-ray解析出的高分辨率结构；如果没有X-ray结构，则选择用最高磁场强度获得的NMR（nuclear magnetic resonance）结构，并选择其中与其他结构平均RMSD（root mean square deviation）最小的结构．另外去掉了复合物中RNA核苷酸数少于5的结构，因为它们很可能是随机结合形成的结构．最终获得了非冗余非核糖体数据库．

1.2　分子序列及二级结构界面偏好和统计势的计算

1.2.1　蛋白质二级结构的识别

常用的蛋白质二级结构识别软件DSSP[19]将二级结构分为8种类型：3_{10}-helix（G）、α-helix（H）、π-helix（I）、turn（T）、β-ladder（E）、β-bridge（B）、bend（S）和其他结构（‘’）（下文以字母“M”表示）．由于DSSP在区分3_{10}-helix和π-helix时常常不准确，容易将π-helix指认为3_{10}-helix，因此对3种helix的指认我们采用2015年Cao等[20]开发的专门用于识别helix结构的软件，对其他二级结构则仍采用DSSP指认的结果．

1.2.2　RNA二级结构的识别

采用程序3DNA[21]来识别RNA二级结构．RNA二级结构主要是对核苷酸配对情况的区分，核苷酸配对被分为3种类型：沃森-克里克配对（WC paired，R类）、非沃森-克里克配对（non-WC paired，Q类）和未配对（unpaired）．进一步对未配对核苷酸根据其与近邻核苷酸间空间排列规整性关系将其细分成两类：不规整（U类）和规整（V类）核苷酸．规整与不规整的指认步骤如下：首先计算RNA未配对区域上所有相邻核苷酸碱基间的张角O1-P-O2（采用程序X3dnr完成）[22]（图1），其中P为连接两个序列上相邻核苷酸的磷酸基团上的P原子，O1/O2为标准碱基参考框架最优叠落到当前相邻碱基平面后参考框架坐标原点的位置[23]，该角能反映相邻核苷酸碱基在空间伸展方向的偏离程度[22]；然后找出未配对区域上该张角大于85°的核苷酸近邻对；最后统计这些对间的核苷酸数目，当该数目小于5时定义此对及其间的核苷酸为不规则U类，其余核苷酸为V类．

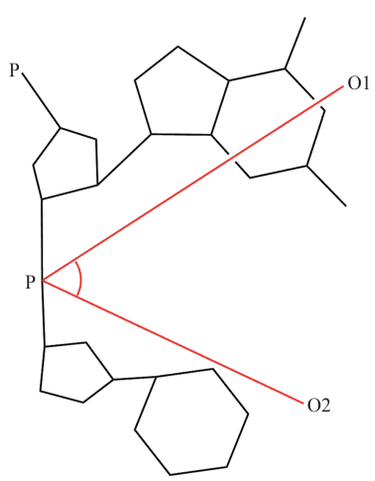

Fig. 1　Definition of O1–P–O2 angle between neighboring nucleotides along sequence

P is the intermediate phosphorus atom between the neighboring nucleotides, and O1 and O2 are coordinate origins of the two corresponding base reference frames which are superimposed on the RNA dinucleotide[22].

1.2.3　计算序列/二级结构的界面偏好性

蛋白质和RNA序列/二级结构的界面偏好性可用下式计算：

$$P_i = \frac{N_i^I \Big/ \sum_i N_i^I}{N_i^S \Big/ \sum_i N_i^S} \tag{1}$$

其中，P_i表示i类/位于i类二级结构的氨基酸或核苷酸的界面偏好性，N_i^I表示界面上i类/位于i类二级结构的氨基酸或核苷酸的数量，$\sum_i N_i^I$表示界面上所有氨基酸或核苷酸的总数，N_i^S表示表面上i类/位于i类二级结构的氨基酸或核苷酸的数量，$\sum_i N_i^S$表示表面上所有氨基酸或核苷酸的总数．注意这里的表面不包含界面．Pi大于1说明i类氨基酸或核苷酸/二级结构倾向于出现在界面上．

需要补充的是，我们用溶剂可及表面积（solvent accessible surface area，SASA）来定义表面氨基酸和核苷酸，其计算用程序NACCESS（http：//wolf.bms.umist.ac.uk/naccess）来完成．表面核苷酸的SASA> 0.1 Å^2；表面氨基酸的SASA > 5%×Si Å^2，其中Si表示氨基酸在三肽下（GLY-X-GLY，X表示该氨基酸）的SASA，i代表20种氨基酸之一，Si来源于Topham和Smith[24]在2015年的统计结果．另外，界面氨基酸/核苷酸定义为蛋白质/RNA的氨基酸/核苷酸上任一原子到与其结合的RNA/蛋白质的任一原子的距离小于4 Å的氨基酸/核苷酸．

1.3　氨基酸–核苷酸成对偏好势的计算

首先计算氨基酸-核苷酸界面成对偏好性，其公式如下：

$$P^I_{ai-bj} = \frac{N^I_{ai-bj} \Big/ \sum_{aibj} N^I_{ai-bj}}{\left(N^s_a \Big/ \sum_a N^s_a\right) \times \left(N^s_i \Big/ \sum_i N^s_i\right) \times \left(N^s_b \Big/ \sum_b N^s_b\right) \times \left(N^s_j \Big/ \sum_j N^s_j\right)} \tag{2}$$

其中，N^I_{ai-bj}表示在界面处蛋白质中出现在i类二级结构的a类氨基酸与RNA中出现在j类二级结构的b类核苷酸配对的个数，$\sum_{aibj} N^I_{ai-bj}$表示界面上氨基酸-核苷酸的总对数，N^S_a和N^S_b分别表示表面上a类氨基酸和b类核苷酸的个数，$\sum_a N^s_a$和$\sum_b N^s_b$分别表示表面上所有氨基酸和核苷酸的个数，N^S_i和N^S_j分别表示表面上出现在蛋白质的i类二级结构中的氨基酸和出现在RNA的j类二级结构中的核苷酸的个数，$\sum_i N^s_i$和$\sum_j N^s_j$分别表示表面上所有氨基酸和核苷酸的个数．注意这里的表面不包含界面．氨基酸-核苷酸成对定义为：氨基酸上任一原子到核苷酸上任一原子的距离小于4 Å的配对．需要指出的是这里我们根据分子二级结构的界面偏好性对其进行了整合和归类，见结果部分．

根据玻尔兹曼分布原理，可以将偏好性转换为相应的统计势：

$$\Delta G = -RT\ln(P) \tag{3}$$

其中，ΔG为复合物中氨基酸-核苷酸形成配对对蛋白质-RNA相互作用能量的贡献，R和T分别表示气体常数和绝对温度，这里RT = 0.59 kcal/mol．当P>1时，△G<0，说明此种倾向于配对的氨基酸和核苷酸在界面处形成配对有助于复合物的形成和稳定．

1.4　成对统计偏好势在对接打分中的应用

将构建的统计势作为打分函数用于蛋白质-RNA分子对接近天然结构的挑选．对每个对接体系，用上述构建的氨基酸-核苷酸成对统计势对其结合模式进行打分评价：

$$Score_i = \sum \Delta G \tag{4}$$

用打分成功率来衡量打分函数的有效性．打分成功率定义为在保留一定数目的对接结构中找到的近天然结构的数目占对接得到的所有近天然结构数目的比例．近天然结构定义为与天然结构的配体均方根偏差L_RMSD < 10 Å的结合模式．

1.5　对接体系

对接打分测试集来源于Huang等[25]构建的数据集合，共包括72个蛋白质-RNA复合物结构．移除其中4个核糖体结构（1DFU_P：MN、1F7Y_A：B、1G1X_FH：IJ、1MMS_A：C）后剩余68个非核糖体结构．对接方法采用常用的FTDock[26]，对每个体系产生10 000个几何互补性最好的对接结构，参数取默认值．为了避免打分排序中的随机性，去掉10 000个结构中近天然结构数目少于3的体系，剩下35个体系（表1）．

Table 1　35 protein−RNA complexes docking scoring tests

Complex[1)]	Description	Protein PDB	RNA PDB
1C0A_A：B	ASPartyltRNAsynthetase/ASPartyltRNA	1IL2_A	1EFW_C
1E8O_CD：E	Signal recognition particle protein/7SL RNA	1E8O_AB	1RY1_E
1F7U_A：B	ARGinyl-tRNAsynthetase/tRNA（ARG）	1BS2_A	1F7V_B
1FFY_A：T	IsoLEUcyl-tRNAsynthetase/IsoLEUcyl-tRNA	1QU3_A	1QU2_T
1GAX_B：D	Valyl-tRNAsynthetase/tRNA（Val）	1GAX_A	1IVS_C
1HQ1_A：B	Signal recognition particle protein/4.5S RNA domain IV	3LQX_A	1DUL_B
1JBS_A：C	Restrictocin/Sarcin/Ricin domain RNA	1JBR_A	1JBT_C
1K8W_A：B	tRNAPseudouridine Synthase B/T Stem-Loop RNA	1R3F_A	1ZL3_B
1KOG_CD：K	Threonyl-tRNAsynthetase/Threonyl-tRNAsynthetase mRNA	1EVL_AB	1KOG_I
1LNG_A：B	Signal recognition particle protein/7S.S srp RNA	3NDB_A	2V3C_M
1OOA_B：D	Nuclear factor NF-kappa-B p105 subunit/RNA aptamer	1NFK_A	2JWV_A
1R3E_A：CDE	tRNApseudouridine synthase B/a stem-loop RNA	1ZE2_A	1R3E_CDE
1SJ3_P：R	Small nuclear ribonucleoprotein A/ Precursor form of the Hepatitis Delta virus ribozyme	1M5O_C	1VC7_B
2ANR_A：B	Neuro-oncological ventral antigen 1/RNA aptamer hairpins	2ANR_A	2ANN_B
2AZ0_AB：CD	B2 protein/double-stranded RNA（dsRNA）	2B9Z_AB	2AZ2_CD
2CSX_B：D	METhionyl-tRNAsynthetase/tRNA（MET）	2CSX_A	2CT8_C
2CZJ_E：F	SsrA-binding protein/tmRNA	1WJX_A	2CZJ_B
2QUX_DE：F	Coat protein/a viral RNA	2QUD_AB	2QUX_C
2RFK_A：DE	Probable tRNApseudouridine synthase B/Guide RNA 1，Guide RNA 2	3LWR_A	3HJY_CD
2XDB_A：G	a protein toxin（ToxN）/a specific RNA antitoxin（ToxI）	2XD0_A	2XDB_G
2ZM5_A：C	tRNA delta（2）-isopentenylpyrophosphate transferase/tRNA（PHE）	3FOZ_A	2ZXU_C
2ZUE_A：B	ARGinyl-tRNAsynthetase/tRNA-ARG	2ZUE_A	2ZUF_B
3DD2_H：B	Thrombin heavy chain/an RNA aptamer	1GJ5_H	3DD2_B
3EPH_A：E	tRNAisopentenyltransferase/tRNA	3EPK_A	3EPJ_E
3FOZ_A：C	tRNA delta（2）-isopentenylpyrophosphate transferase/tRNA（PHE）	2ZXU_A	2ZM5_C
3LRR_A：CD	Probable ATP-dependent RNA helicase DDX58/double-stranded RNA	3LRN_A	3LRR_CD
3OL9_M：NO	Polymerase/Positive-strand RNA	3OL6_A	3OLB_BC
3OVB_A：C	CCA-Adding Enzyme/tRNA	3OV7_A	3OUY_C
1S03_H：A	30S ribosomal protein S8/spc Operon mRNA	3OFO_H	1S03_B
1T0K_B：CD	60S ribosomal protein L30/mRNA	3O58_Z	1T0K_CD
2BH2_A：C	23S rRNA （uracil-5-）-METhyltransferaseRumA/23S ribosomal RNA fragment	1UWV_A	2BH2_D
3LWR_ABC：DE	Pseudouridine synthase Cbf5，Ribosome biogenesis protein Nop10， 50S ribosomal protein L7Ae/H/ACA RNA	3LWP_ABC	3HJW_DE
3MOJ_B：A	ATP-dependent RNA helicase dbpA/23S ribosomal RNA fragment	2G0C_A	3MOJ_A
3FTF_A：CD	DiMEThyladenosine transferase/16S rRNA fragment	3FTD_A	3FTE_CD
2HW8_A：B	50S ribosomal protein L1/mRNA	1AD2_A	1ZHO_B

[1)] PDB codes of protein-RNA complexes with protein and RNA chains separated with "：".

2　结果与讨论

2.1　蛋白质−RNA非冗余非核糖体数据库统计

根据数据库建立的步骤，我们一共获得了694个核糖体的蛋白质-RNA复合物结构．相比于2012年构建的数据库（251个结构），本工作构建的复合物结构数量增加了近2倍．另外在之前的工作中，RNA的同源定义为序列相似性大于95%，而

在本工作中这一数字为90%，冗余去除更加严格.

2.2　蛋白质和RNA序列和二级结构的界面偏好性分析

从本文构建的蛋白质-RNA复合物的非冗余数据库中，根据公式（1）我们提取了蛋白质和RNA的序列和二级结构的界面偏好性，下面对其分别进行分析.

2.2.1　序列的界面偏好性分析

对蛋白质和RNA，分别获得了氨基酸和核苷酸的界面偏好性（图2）. 对蛋白质中的氨基酸，按照疏水、亲水和带电氨基酸的分类对其进行分析：a. 8种疏水氨基酸（ALA、VAL、LEU、ILE、PRO、MET、PHE、TRP）中，除MET（1.09）和两个芳香族氨基酸PHE（1.32）和TRP（1.11）外，其他都不倾向出现在界面上，这与Jones等[27]研究结果相符. 之前的工作显示，尽管MET、PHE和TRP的界面偏好性相对较高，但都没有大于1.0[11]. 芳香族氨基酸常常与RNA碱基间形成堆叠作用，因此高界面偏好是合理的. b. 7种亲水不带电的氨基酸（TYR、GLY、SER、THR、CYS、ASN、GLN）中，TYR（1.51）、CYS（1.88）、ASN（1.23）和GLN（1.14）明显偏好出现在界面上，其他残基的偏好性在1附近，这与我们之前的工作结果类似. 芳香族氨基酸TYR侧链上带有苯环和羟基，容易与核苷酸碱基形成堆叠和氢键作用，因此偏好性较高. CYS表现较强的偏好性，数值上略高于之前的统计结果. 另外，很多研究发现ASN偏好出现在界面上[28-30]，而GLN偏好性一般，有时表现出不偏好[27]. 这里的结果是两残基都具有界面偏好性，这与Pérez-Cano等[9]和我们之前的研究相符. 两残基的结构和物理化学性质相似，尽管不带电但都属于亲水氨基酸，可能应该具有相似的界面偏好性. c. 带正电的3种氨基酸（LYS、ARG、HIS）都偏好出现在界面上，其中ARG（2.04）的偏好性最高，位于20种残基之首，HIS（1.45）、LYS（1.23）界面偏好性次之. 因为RNA带负电，所以带正电的残基偏好界面很容易理解. d. 带负电的2种氨基酸ASP（0.57）和GLU（0.38）都不偏好界面，在20种残基中具有最低的偏好性. 我们之前的结果显示，GLU偏好性最低，与这里的结果一致，但ASP不是次低. 从带电性来讲，这里的结果可能更合理，这与统计数据量的增加有直接关系.

就RNA序列的界面偏好性而言，核苷酸U（1.40）和A（1.08）都偏好界面，而C（0.91）和G（0.79）不偏好. 从定性上讲我们之前的结果也是这样的，只是A与U的界面偏好性类似，没有这里差距明显. 核苷酸U（不同于DNA中的核苷酸）具有最强的界面偏好性，可能有进化意义[31].

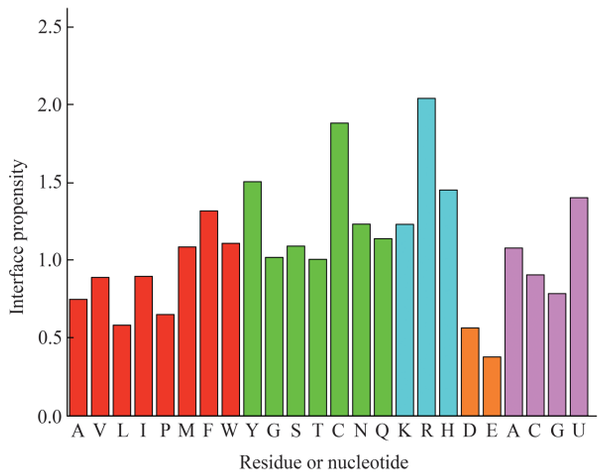

Fig. 2　Interface preferences of protein and RNA sequences

Amino acids are shown in the order of the hydrophobic (A-W), uncharged hydrophilic (Y-Q), positively charged (K-H), and negatively charged (D-E) amino acids. Nucleotides are shown in the last (A-U).

2.2.2　二级结构界面偏好性分析

图3给出了蛋白质和RNA的二级结构单元的界面偏好性. 就蛋白质二级结构的界面偏好性而言，如图3a所示：β-ladder（1.31）偏好性最强，其次是β-bridge（1.25）和3_{10}-helix（1.22），顺序上与Susan等[27]的结果相符. 我们之前的结果显示，这三者的界面偏好性都大于1.0，其中3_{10}-helix高于前两者. 就螺旋结构来说，α-helix、3_{10}-helix和π-helix三者之间的主要差异在结构盘绕的紧密度上，π-helix一圈4.1个氨基酸、α-helix 3.6个，而3_{10}-helix则有3个，所以后者缠绕更加紧密. 由于3_{10}-helix结构上的紧凑性，它更容易插入RNA的沟槽中形成相互作用，这也是Draper等[32]发现的蛋白质-RNA两种结合模式中一种常见的类型. 这两种结合模式分别是β折叠结合（蛋白质用β折叠结构作为结合界面与单链RNA相互作用）和沟槽结合（蛋白质某些二级结构特异性识别RNA双链沟槽的形状和序列）. 这也解释了β折叠和3_{10}-helix其界面偏好性高的原因.

如图3b给出了RNA二级结构的界面偏好性.

明显看出未配对核苷酸具有强的界面偏好性．有研究表明，在蛋白质-RNA的主要结合模式之一β折叠结合模式中，β折叠常常识别未配对RNA核苷酸[27]．未配对核苷酸柔性较大，容易调整结构与蛋白质发生诱导契合作用，另外未配对碱基有能力与蛋白质形成更多氢键，这可能是未配对核苷酸具有强界面偏好的原因．另外，需要指出的是这里比之前的工作更进一步，我们对未配对核苷酸提出了空间结构排列规整性特征，将未配对核苷酸分成不规整类U和规整类V．结果发现二者的界面偏好性差异非常显著，U类（2.95）比V类（1.81）高出1.14．进一步统计分析发现，694个复合物中有367个其RNA表面出现了U类结构，其中351个的U类结构出现在界面上，比例高达95.6%. 未配对中U类核苷酸的碱基具有更多的伸展方向，相比排列规整核苷酸具有更大的可塑性，前者更容易调整结构与蛋白质发生诱导契合作用，进一步加强结合，因此前者比后者具有更大的界面偏好性．U类二级结构对蛋白质-RNA特异性识别具有重要意义，我们将在统计势中考虑这一信息．

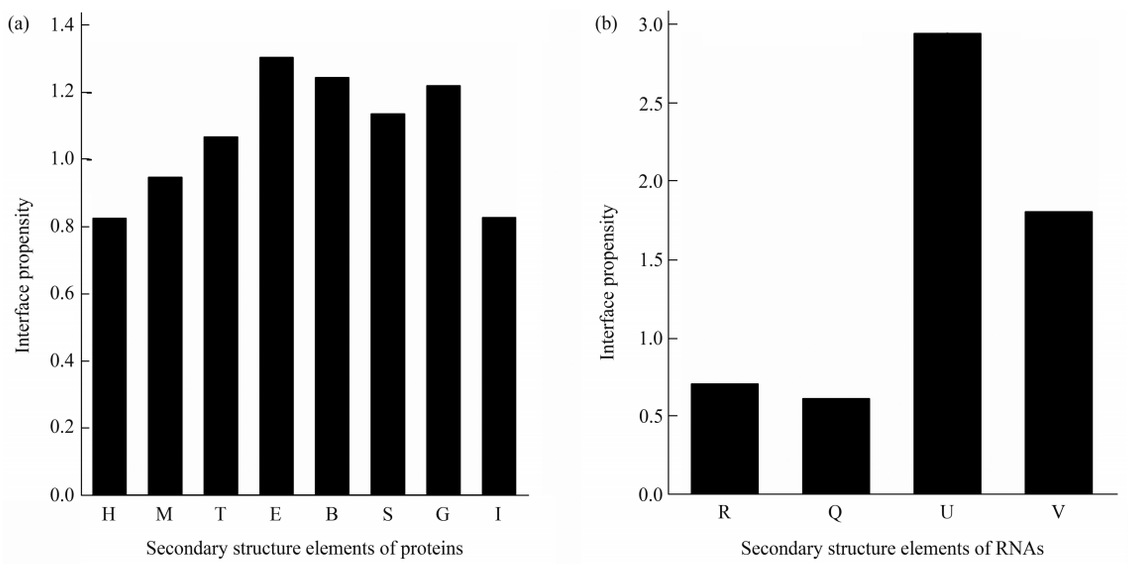

Fig. 3　Interface preferences of secondary structure elements in proteins and RNAs

H：α-helix，M：Unclassified，T：Turn，E：β-ladder，B：β-bridge，S：bend，G：3_{10}-helix，I：π-helix；R：WC paired nucleotides，Q：non-WC paired nucleotides，U：Unpaired nucleotides with irregularly arranged bases，V：Unpaired nucleotides with regularly arranged bases.

2.3　氨基酸–核苷酸成对统计偏好势分析

根据蛋白质二级结构界面偏好性，将其分为3类：X（H、M、I，P<1），Y（T、S，P≈1），Z（B、E、G，P>1）；根据RNA二级结构界面偏好性，将其也分为3类：P（R、Q，P<1），U（U，P>2），V（V，1<P<2）．

在此分类基础上，构建了20×4（只考虑序列信息）、20×12（考虑序列及RNA二级结构信息）和60×12（考虑序列及蛋白质和RNA的二级结构信息）成对统计势，其中前两个显示在图4中．由图4a可以看出，不同氨基酸与相同核苷酸配对能量有较大差异，最大差异变化范围为1.28（ARG-U：-0.75 ~ GLU-U：0.53）；而对核苷酸来说，不同核苷酸与相同氨基酸配对的能量差异较小，最大差异变化范围为0.67（VAL-U：-0.2 ~ VAL-G：0.47）；Arg-U（-0.75）具有最负的成对能量．从图4b可以看出，对核苷酸按二级结构分类后，它们与同种氨基酸配对能量差异变大了，变为2.16（PHE-U_U：-1.49 ~ PHE-C_P：0.67）．这一变化体现了在能量上可以对氨基酸-核苷酸配对具有更细致的区分，这有利于更好地评价蛋白质-RNA对接结构．进一步分析发现，处于P类二级结构中的4种核苷酸几乎都不倾向出现在界面上（与氨基酸的配对能量为正）；处于U类或V类中的核苷酸几乎都偏好（与氨基酸的配对能量为负），且前者比后者更偏好，这进一步说明RNA中的U类核苷酸对蛋白质-RNA特异性识别的重要性．

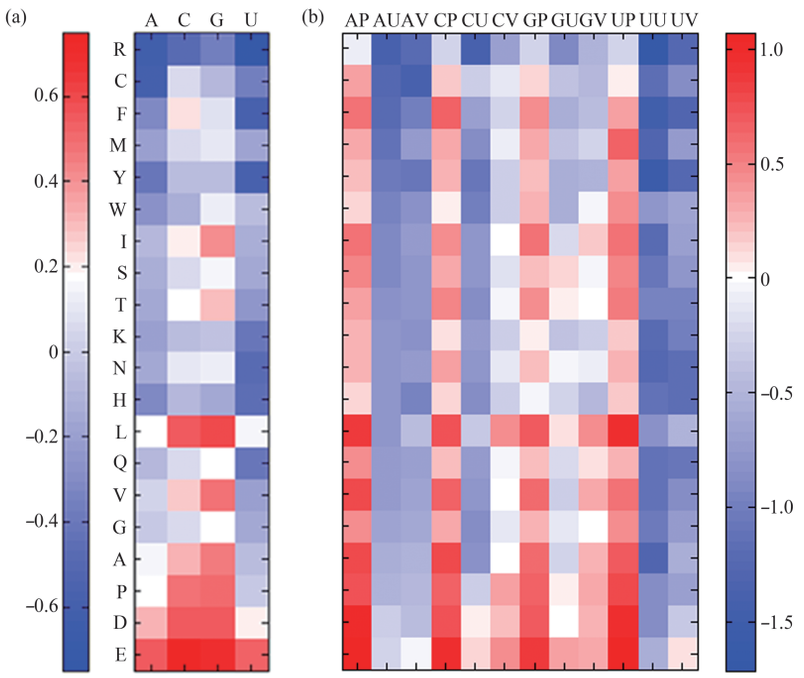

Fig. 4　Pairwise residue−nucleotide statistical potentials

Pairwise residue-nucleotide potentials：（a）20×4 potential，（b）20×12 potential. Energy bar is also shown. AP，AU and AV denote the adenines in secondary structure classes P，U and V respectively，and the others have the similar meanings.

2.4　不同统计势的打分结果

为了分别考察蛋白质和RNA二级结构信息的考虑对统计势执行效果的影响，在构建60×12统计势的同时，还构建了20×12（考虑RNA二级结构信息）和60×4（考虑蛋白质二级结构信息）统计势．为了检测RNA未配对核苷酸排列规整性考虑对统计势的影响，进一步构建了20×8统计势（考虑RNA二级结构，但这里二级结构分为两类：配对（R、Q，P<1）和未配对（U，V，P>1））来与20×12统计势进行比较．

需要说明的是本文60×12统计势与我们之前在251个复合物数据库上构建的60×8*统计势，除了所用数据库不同外，最大的不同在于：前者对RNA未配对核苷酸进一步细分成规整类和非规整类，在后者中未细分；其次前者将蛋白质的β-ladder和β-bridge两种β折叠二级结构分到P>1偏好的一类，而后者将其分到P=1的一类．根据以上残基和二级结构界面偏好性的分析，这三方面的改进都应该是有利于统计势区分对接近天然结构的，下面将对这些统计势进行测试．

将60×12统计势用于对35个蛋白质-RNA（表1）分子对接的打分测试，同时也进行了对60×4、20×12和20×8统计势的测试．为了与之前的工作进行比较，还用60×8*统计势进行了对接打分．对每个体系用FTDock进行分子对接，产生10 000个几何互补性最好的对接结构．之后用统计势对其进行打分排序，对前2 000个结构进行打分成功率的计算，结果如图5所示．

从图5可以看出，当保留2 000个结构时，分别考虑蛋白质和RNA二级结构信息后的60×4和20×12统计势的成功率分别为62.26%和65.23%．综合考虑两者的60×12统计势的成功率最高，为65.77%，比60×4统计势高出3.51%，比20×12统计势高出0.54%．尽管0.54%的优势并不是很明显，但从曲线总体的走势看，60×12统计势还是有优势的．这个结果也说明了RNA二级结构信息的考虑较蛋白质二级结构信息的考虑更加重要，这一致于我们之前的结果[11]．另一方面，20×12统计势（65.23%）比20×8统计势（52.02%）的成功率高出13.21%，这说明相比于RNA二级结构分为配对和未配对2类，将RNA二级结构分为3类更有利于提高打分成功率，也就是说本文对RNA未配对核苷酸排列规整性的进一步区分对提升蛋白质-RNA对接打分成功率具有重要意义．另外，与之前在251

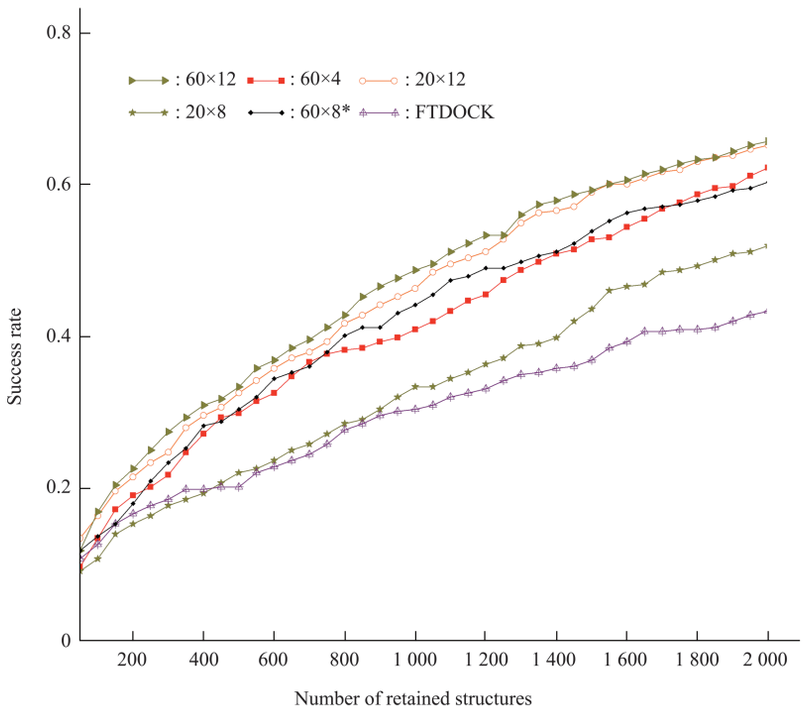

Fig. 5　Success rates of scoring for different statistical potentials

个结构上构建的60×8*（60.38%）统计势相比，本文构建的60×12统计势（65.77%）也具有明显的优势，这说明数据库的扩大及二级结构重要特征的挖掘对统计势构建的合理性有显著的改进．我们还计算了FTDock几何互补性的打分成功率，为43.40%，这说明仅依靠几何匹配程度区分近天然结构还是不够的，如果能与统计势进行有效组合相信其成功率会进一步得到提升，这个工作目前正在进行中．

下面以1OOA_B：D和2BH2_A：C体系为例，图6给出了用60×12统计势打分获得的第一个近天然结构（分别排在第7位和第3位）与实验结构的比较．由图6a和b可以看出，这两个体系主要分别属于蛋白质-RNA相互作用中的3_{10}－helix和β折叠

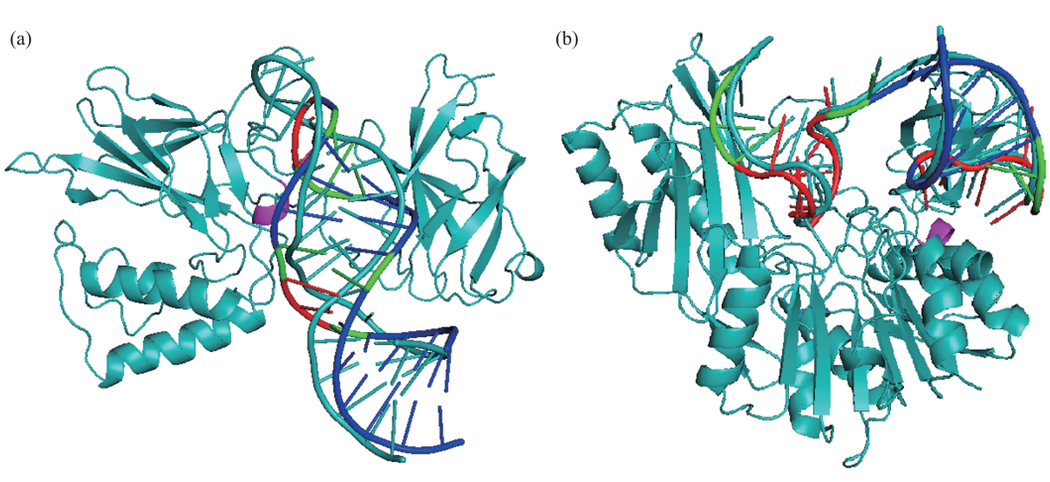

Fig. 6　Structure superposition（with proteins superimposed together）of the near native structure with the best ranking obtained by 60×12 potential and the experimental one for protein−RNA systems 1OOA_B：D（a）and 2BH2_A：C（b）

Native structure is in light blue（protein interface 3_{10}-helix in pink），and the RNA in the near native structure is colored with U，V and P types of nucleotides in red，green and blue respectively.

识别模式．另外可以明显看出，在2BH2_A：C体系中，RNA未配对U类（不规整类）核苷酸全部出现在结合界面上，V类（规整类）核苷酸中只有部分在界面上，这进一步说明60×12统计势中将RNA未配对核苷酸分成规整和非规整类的合理性和重要意义．

对统计势的获得需要说明的一点是，以下的做法是更加合理的．从提取统计势的非冗余数据库中去掉与测试复合物在一定序列相似性以上的复合物结构，再抽取统计偏好势以用于测试复合物的对接打分．如果此种情况下打分成功率高，更能说明打分函数的鲁棒性．从目前的结果来看，以20×4的统计势为例，3种带正电的氨基酸ARG、LYS和HIS与4种核苷酸配对的能量都是负的，有利于结合；2种带负电的氨基酸GLU和ASP与4种核苷酸配对的能量都是正的，不利于结合．考虑到RNA的电负性，这是合理的．这一结果也说明，从大量非冗余复合物（694）中提取的统计势在一定程度上很好地反应了氨基酸-核苷酸物理相互作用的偏好．事实上，由于该非冗余复合物界面的氨基酸-核苷酸配对数目非常大，即使去除几个复合物结构来重新获得统计势，新的统计势也几乎不会受到影响，这样打分结果也不会有大的变化．

3 结　论

本工作构建了新的蛋白质-RNA复合物非冗余非核糖体数据库（694个结构），并在此基础上统计了蛋白质中氨基酸和RNA中核苷酸，以及二级结构的界面偏好性．发现芳香族（PHE、TYR、TRP）和酰胺类（ASN、GLN）氨基酸，以及所有带正电的氨基酸（ARG、LYS、HIS）都偏好出现在界面上；RNA中核苷酸A和U偏好界面．蛋白质二级结构中两种β折叠（β-bridge、β-sheet）和3_{10}-helix具有显著的界面偏好性． RNA二级结构中未配对，尤其是其中不规整类核苷酸具有高的界面偏好．根据二级结构界面偏好，对其进行归类，构建了考虑序列和二级结构信息的60×12氨基酸-核苷酸成对偏好统计势，并将其与仅考虑蛋白质或RNA二级结构信息的60×4和20×12统计势进行了比较．结果发现，60×12统计势打分效果最好，其成功率为65.77%，比60×4和20×12统计势分别高出3.51%和0.54%，这说明统计势中分子二级结构，尤其是RNA二级结构信息的考虑可明显提高其在分子对接中筛选近天然结构的成功率，也暗示了分子二级结构对蛋白质-RNA特异性识别的重要贡献．20×12统计势比20×8统计势成功率高出13.21%，说明进一步区分RNA未配对核苷酸排列规整性结构特征对蛋白质-RNA结构预测有重要作用．另外，60×12统计势也好于基于251个结构构建的60×8*统计势，说明数据库的扩大及分子二级结构特征的挖掘对统计势的合理构建具有重要意义．本文的研究可促进蛋白质-RNA分子对接方法研究的发展，有助于加深对蛋白质-RNA特异性识别机制的理解．

参 考 文 献

[1] Glisovic T, Bachorik J L, Yong J, *et al*. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett, 2008, **582**(14): 1977-1986

[2] Lukong K E, Chang K W, Khandjian E W, *et al*. RNA-binding proteins in human genetic disease. Trends Genet, 2008, **24**(8): 416-425

[3] Perez-Cano L, Romero-Durana M, Fernandez-Recio J. Structural and energy determinants in protein-RNA docking. Methods, 2017, **118-119**: 163-170

[4] Guilhot-Gaudeffroy A, Froidevaux C, Aze J, *et al*. Protein-RNA complexes and efficient automatic docking: expanding RosettaDock possibilities. PLoS One, 2014,**9**(9):e108928

[5] Tuszynska I, Magnus M, Jonak K, *et al*. NPDock: a web server for protein-nucleic acid docking. Nucleic Acids Res, 2015, **43**(W1): W425-W430

[6] 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. Nucleic Acids Res, 2017,**45**(W1):W365-W373

[7] Chauvot D B I, de Vries S J, Zacharias M. Binding site identification and flexible docking of single stranded RNA to proteins using a fragment-based approach. PLoS Comput Biol, 2016,**12**(1):e1004697

[8] Li H, Huang Y, Xiao Y. A pair-conformation-dependent scoring function for evaluating 3D RNA-protein complex structures. PLoS One, 2017,**12**(3):e174662

[9] Perez-Cano L, Solernou A, Pons C, *et al*. Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. Pac Symp Biocomput, 2010:293-301

[10] Tuszynska I, Bujnicki J M. DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. BMC Bioinformatics, 2011,**12**:348

[11] 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**(1):14-24

[12] Zhang Z, Lu L, Zhang Y, *et al*. A combinatorial scoring function for protein-RNA docking. Proteins, 2017,**85**(4):741-752

[13] Re A, Joshi T, Kulberkyte E, *et al*. RNA-protein interactions: an

overview. Methods Mol Biol, 2014,1097:491-521

[14] El-Manzalawy Y, Abbas M, Malluhi Q, *et al*. FastRNABindR: fast and accurate prediction of protein-RNA interface residues. PLoS One, 2016,**11**(7):e158445

[15] Walia R R, El-Manzalawy Y, Honavar V G, *et al*. Sequence-based prediction of RNA-binding residues in proteins. Methods Mol Biol, 2017,1484:205-235

[16] Bahadur R P, Zacharias M, Janin J. Dissecting protein-RNA recognition sites. Nucleic Acids Res, 2008,**36**(8):2705-2716

[17] Ellis J J, Broom M, Jones S. Protein-RNA interactions: structural analysis and functional classes. Proteins, 2007,**66**(4):903-911

[18] Allers J, Shamoo Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. J Mol Biol, 2001, **311**(1):75-86

[19] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 1983,**22**(12):2577-2637

[20] Cao C, Xu S, Wang L. An algorithm for protein helix assignment using helix geometry. PLoS One, 2015,**10**(7): e129674

[21] Lu X J, Olson W K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res, 2003,**31**(17):5108-5121

[22] Lu X J, Bussemaker H J, Olson W K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. Nucleic Acids Res, 2015,**43**(21):e142

[23] Olson W K, Bansal M, Burley S K, *et al*. A standard reference frame for the description of nucleic acid base-pair geometry. J Mol Biol, 2001,**313**(1):229-237

[24] Topham C M, Smith J C. Tri-peptide reference structures for the calculation of relative solvent accessible surface area in protein amino acid residues. Comput Biol Chem, 2015,54:33-43

[25] Huang S Y, Zou X. A nonredundant structure dataset for benchmarking protein-RNA computational docking. J Comput Chem, 2013,**34**(4):311-318

[26] Katchalski-Katzir E, Shariv I, Eisenstein M, *et al*. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA, 1992,**89**(6):2195-2199

[27] Jones S, Daley D T, Luscombe N M, *et al*. Protein-RNA interactions: a structural analysis. Nucleic Acids Res, 2001,**29**(4): 943-954

[28] Jeong E, Kim H, Lee S W, *et al*. Discovering the interaction propensities of amino acids and nucleotides from protein-RNA complexes. Mol Cells, 2003,**16**(2):161-167

[29] Kim H, Jeong E, Lee S W, *et al*. Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. FEBS Lett, 2003,**552**(2-3):231-239

[30] Treger M, Westhof E. Statistical analysis of atomic contacts at RNA-protein interfaces. J Mol Recognit, 2001,**14**(4):199-214

[31] Vertessy B G, Toth J. Keeping uracil out of DNA: physiological role, structure and catalytic mechanism of dUTPases. Acc Chem Res, 2009,**42**(1):97-106

[32] Draper D E. Themes in RNA-protein recognition. Journal of Molecular Biology, 1999,**293**(2):255-270

Preferences of Sequence and Structure for Protein−RNA Interfaces and Its Application in Scoring Potential Construction for Docking*

LU Lin, LIU Yang, LI Chun-Hua**

(*Faculty of Environmental and Life Sciences, Beijing University of Technology, Beijing* 100124, China)

Abstract　We constructed a non-redundant non-ribosomal protein-RNA interface dataset (including 694 structures) from the Protein Data Bank (PDB). The interface preferences of amino acids, nucleotides and the secondary structure elements of protein and RNA were computed based on the dataset. The results show that β-ladder, β-bridge and 3_{10}-helix of proteins and the unpaired nucleotides of RNA, especially those irregularly arranged nucleotides have remarkably high interface propensities. Based on these, we classified the secondary structure elements, constructed the 60×12 amino acid-nucleotide pairwise potential, and used it as a scoring function in protein-RNA docking to select the near native structures. The results show the 60×12 pairwise potential has a scoring success rate of 65.77%, better than those of the pairwise potentials with secondary structure information of protein or RNA considered, as well as better than that of our previously constructed 60×8* potential based on the 251 protein-RNA complex structures. This work is helpful for strengthening the understanding of protein-RNA specific interactions and can advance the progress of protein-RNA complex structure prediction.

Key words　molecular docking, statistical potential, interface preference, secondary structure

DOI：10.16476/j.pibb.2020.0004

---

* This work was supported by grants from The National Natural Science Foundation of China（31971180，11474013）and Beijing Natural Science Foundation（4152011）.

** Corresponding author.

Tel.：+86-10-67392001，E-mail：chunhuali@bjut.edu.cn

Received：January 5，2020　Accepted：June 9，2020