Biophysics 生物物理学, 2022, 10(4), 47-54
Published Online November 2022 in Hans. http://www.hanspub.org/journal/biphy
https://doi.org/10.12677/biphy.2022.104006

Hans汉斯

蛋白质-DNA复合物中残基界面偏好性分析及在识别界面中的应用

杨　爽，李春华*

北京工业大学环境与生命学部，北京

收稿日期：2022年4月1日；录用日期：2022年11月16日；发布日期：2022年11月23日

摘　要

蛋白质-DNA识别在生物过程中起着重要作用，其结合是由序列特异性识别和结构特征共同影响的。为了研究残基类型和蛋白质二级结构对结合的贡献，本文构建了一个新的非冗余蛋白质-DNA复合物数据库，其中包含1545个结构。经过统计分析发现，残基和二级结构类型对蛋白质与DNA结合有很大贡献，二级结构中π-helix和β-ladder是最偏好界面的类型。对蛋白质二级结构按界面偏好进行分类，构建了60 × 4氨基酸－核苷酸成对界面偏好性。从该偏好性中获得氨基酸界面偏好性，并探讨了将该信息用于预测蛋白质-DNA结合界面的可能性，研究对象为对接基准数据集中的47个复合物体系。结果发现成对界面偏好性信息可以将87.23%的体系的真实界面打分排在所有表面区域的前10%。这说明本文构建的60 × 4氨基酸－核苷酸成对界面偏好性很好地反映了蛋白质-DNA的界面识别，对界面和复合物结构预测具有重要意义。

关键词

蛋白质-DNA相互作用，结合界面，二级结构，界面偏好性

Analysis of Residue Interface Preference in Protein-DNA Complexes and Its Application in Recognition of Binding Interface

Shuang Yang, Chunhua Li*

Faculty of Environment and Life, Beijing University of Technology, Beijing

Received: Apr. 1st, 2022; accepted: Nov. 16th, 2022; published: Nov. 23rd, 2022

*通讯作者 Email: chunhuali@bjut.edu.cn

文章引用: 杨爽, 李春华. 蛋白质-DNA 复合物中残基界面偏好性分析及在识别界面中的应用[J]. 生物物理学, 2022, 10(4): 47-54. DOI: 10.12677/biphy.2022.104006

Abstract

Protein-DNA recognition plays an important role in biological processes, and its binding is influenced by sequence specific recognition and structural characteristics. To investigate the contribution of residue types and protein secondary structure elements to binding, a new non-redundant protein-DNA database with 1545 complex structures was constructed. Statistical analysis reveals that protein residue and secondary structure types have significant contributions to its binding with DNA. Among the secondary structures, π-helix and β-ladder have the highest preferences. We classified the protein secondary structures according to their interface preferences, and constructed the 60 × 4 amino acid-nucleotide pairwise interface preferences. The amino acid interface preferences obtained from the pairwise ones were used to explore the possibility of predicting protein-DNA binding interfaces for 47 complex systems from the docking benchmark dataset. The result shows that the pairwise interface preferences can rank the real interfaces in the top 10% of all surface patches for 87.23% of all cases. These results indicate that the 60 × 4 amino acid-nucleotide pairwise interface preferences constructed by us can well reflect protein-DNA recognition, which is of great significance for interface and complex structure predictions.

Keywords

Protein-DNA Interaction, Binding Interface, Secondary Structure, Interface Preference

Copyright © 2022 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

蛋白质-DNA 相互作用在基因表达调控等多种生物过程中发挥着重要作用[1]。深入了解蛋白质-DNA 相互作用有助于推动基因调控网络的研究和基于结构的药物设计[2]。近几十年来，X 射线晶体学和核磁共振波谱等实验方法已经产生了大量蛋白质-DNA 复合物结构，然而还有很多有待确定[3] [4]。另外，实验方法解析复合物结构成本太高，因此当下急需可靠的理论计算方法来预测复合物结构[5]。

分子对接是目前用来预测复合物结构的一个重要方法[6]。目前蛋白质-DNA 分子对接的研究主要集中在打分函数上[7] [8]，即如何准确筛选出对接采样中的近天然结构。打分方法中，基于知识的统计势已被证明可以有效地区分出近天然结构[8] [9] [10] [11]。2008 年，Skolnick 小组构建了氨基酸残基与 DNA 官能团的成对统计偏好[7]，将 DNA 核苷酸分为磷酸、五碳糖和嘧啶/咪唑三个部分，将其应用于蛋白质-DNA 相互作用预测。2009 年，Zhou 小组在距离尺度的有限理想气体参考态(DFIRE)基础上提出了蛋白质-DNA 的统计能量函数[8]，用于识别对接近天然结构和预测结合亲和性。该方法统计了 212 个蛋白质-DNA 复合物界面上一定距离范围内蛋白质原子与 DNA 原子成对的界面偏好，根据玻尔兹曼分布转化为势能构建打分函数。2015 年，Tuszynska 等人[9]在开发的 NPDock 对接工具中，将构建的 QUASI-DNP 统计势[10]与 DFIRE 统计势[8]、Varani 组的统计势[11]相结合，组合了一套“meta-potential”用于筛选蛋白质-DNA 分子对接中的近天然结构，证明了统计势在评价复合物相互作用和预测复合物结构当中的重要作用。目前，大部分蛋白质-DNA 统计偏好的研究是仅仅考虑界面原子或残基的成对信息[8] [9] [10]，

没有考虑蛋白质二级结构对界面识别的影响。Li 小组在 2011 和 2017 年提出了蛋白质-RNA 的 60 × 8 [12]和 60 × 12 统计势[13]，其中加入了蛋白质和 RNA 的二级结构信息，结果表明加入结构信息后预测近天然结构的能力显著提高。在蛋白质-DNA 复合物中，二级结构也具有不同的界面偏好[14]。由此，我们推测对蛋白质-DNA 复合物，考虑二级结构特征的氨基酸－核苷酸偏好性能够更好地评价复合物间的相互作用。另外，如果该偏好信息可以用于蛋白质-DNA 界面预测，那么这必将有助于复合物结构预测。

在本工作中，我们对新构建的非冗余蛋白质-DNA 数据集，统计了复合物中蛋白质二级结构的界面偏好性，结合氨基酸和核苷酸的界面偏好信息，构建了 60 × 4 氨基酸－核苷酸成对偏好性。之后，将从该偏好性获得的氨基酸界面偏好性应用于蛋白质-DNA 复合物界面区域的识别，获得了好的结果。

2. 数据与方法

2.1. 研究体系

本文下载了截止到 2020 年 10 月 NDB (Nucleic acid Database) [15]数据库(http://ndbserver.rutgers.edu)中所有的蛋白质-DNA 复合物结构，数量为 5387 个。去掉其中包含 RNA 或 FANA 等核酸类似物的复合物，仅保留天然蛋白质-DNA 的复合物。使用序列比对工具 CDHIT 软件[16]比较复合物间的序列相似性，将蛋白质序列相似性和 DNA 序列相似性分别高于 70%和 90%的复合物归为一类，每一类中保留代表结构。此外，还去掉了 DNA 中核苷酸数量少于 5 的复合物结构，因为这些 DNA 很可能是与蛋白质随机结合的。经过上述处理，构建了非冗余蛋白质-DNA 数据库，其中包含 1545 个复合物结构。

用于界面预测的研究体系来自 2008 年 Marc van Dijk 等人构建的蛋白质-DNA 对接基准数据集[17]，该数据集包含 47 个蛋白质-DNA 复合物。通过比较蛋白质真实界面和表面区域的氨基酸平均界面偏好性来识别结合区域。

2.2. 复合物氨基酸－核苷酸成对及界面和表面的定义

成对氨基酸－核苷酸定义为：氨基酸中任一重原子与核苷酸中任一重原子间距离小于 5 Å 的配对。界面氨基酸/核苷酸为所有参与成对的氨基酸/核苷酸，所有界面氨基酸/核苷酸构成了蛋白质/DNA 结合界面。表面核苷酸定义为溶剂可及表面积(Solvent Accessible Surface Area, SASA)大于 0.1 Å^2 的核苷酸；表面氨基酸残基定义标准为其 SASA > 5% S_i Å^2，Si 表示氨基酸 *i* 在三肽下(即 Ala-*i*-Ala 环境下) [18]的 SASA，*i* 为氨基酸类型。溶剂可及表面积使用 NACCESS 计算(http://www.bioinf.manchester.ac.uk/naccess/)。表面区域为表面氨基酸/核苷酸构成的区域，非界面表面为表面中不包含界面氨基酸/核苷酸的部分。

2.3. 蛋白质二级结构的界面偏好性

蛋白质二级结构分为 8 类：3_{10}-helix (G)，α-helix (H)，π-helix (I)，turn (T)，β-ladder (E)，β-bridge (B)，bend (S)，不确定的二级结构(M)，用 DSSP 软件[19]进行计算。定义二级结构的界面偏好性：

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

其中，N_i^I 表示 *i* 类二级结构类型的氨基酸残基出现在界面的数量，$\sum_i N_i^I$ 表示界面上所有八种二级结构类型氨基酸残基的总数；N_i^S 表示 *i* 类二级结构类型的氨基酸残基出现在蛋白质非界面表面的数量，$\sum_i N_i^S$ 表示蛋白质非界面表面上残基总数。由公式(1)可知，当二级结构界面偏好性 $P_i > 1$ 时，说明 *i* 类二级结构类型倾向于出现在界面上。

2.4. 氨基酸－核苷酸成对偏好性

对蛋白质二级结构进行分类，计算氨基酸－核苷酸的成对偏好性：

$$P_{ai\text{-}b}^{I}=\frac{N_{ai\text{-}b}^{I}\Big/\sum_{aib}N_{ai\text{-}b}^{I}}{\left(N_{ai}^{S}\Big/\sum_{ai}N_{ai}^{S}\right)\times\left(N_{b}^{S}\Big/\sum_{b}N_{b}^{S}\right)} \tag{2}$$

其中，a 为 20 种氨基酸类型，i 为蛋白质二级结构分类，b 为 4 种核苷酸类型；$N_{ai\text{-}b}^{I}$ 表示 i 类二级结构中的 a 类氨基酸与 b 类核苷酸成对在界面上个数，$\sum_{aib}N_{ai\text{-}b}^{I}$ 表示界面上所有类型氨基酸－核苷酸成对的总数量；N_{ai}^{S} 表示 i 类二级结构中的 a 类氨基酸出现在非界面表面的个数；N_{b}^{S} 和 $\sum_{b}N_{b}^{S}$ 分别表示非界面表面上 b 类核苷酸的个数和所有核苷酸的总个数。由公式(2)可知，当 $P_{ai-b}^{I}>1$ 时，说明 ai-b 氨基酸核苷酸成对更容易出现在复合物界面上。

2.5. 复合物氨基酸－核苷酸成对及界面和表面的定义

对于蛋白质，依次以每一个表面残基(以其 Cα 原子表示)为中心，选择固定半径 20 Å，同时使用向量约束[20]，产生表面区域。向量约束是指：计算每个表面残基的溶剂向量(从其最近的 10 个残基的几何中心到该残基的 Cα 原子的向量，方向指向溶剂一侧)，在生成表面区域时，若表面残基的溶剂向量与其到中心残基的向量夹角小于 110 度，则该残基认为属于中心残基周围的表面区域，反之则从该表面区域中去除该残基[21]。添加向量约束的目的是，使生成的表面区域内的残基与中心残基都在蛋白质的同一侧，避免表面区域中包含蛋白质另一侧的表面残基。

氨基酸的界面偏好性使用与四种核苷酸界面成对偏好的平均值表示。通过统计计算出真实界面和每一个表面区域的平均氨基酸界面偏好，并按其降序排列，统计真实界面相对于该体系所有表面区域的排名。

3. 结果与讨论

3.1. 蛋白质和 DNA 残基及二级结构的界面偏好性分析

本文从亲水性、疏水性和带电性分析了氨基酸和核苷酸的界面偏好性，如图 1(a)。在 8 种疏水氨基酸中，仅氨基酸 Trp 具有很高的界面偏好性，其余氨基酸皆不倾向于出现在结合界面上。在 7 种极性不带电的氨基酸中，Tyr、Thr 和 Ser 的界面偏好最高，它们已经被发现是蛋白质-DNA 复合物中常见的关键残基[22]，其结构中的羟基可以使它们与核苷酸形成氢键相互作用[23]。带电氨基酸中，正电氨基酸(Arg、Lys 和 His)有明显的界面偏好，尤其是 Arg 的界面偏好最高，负电氨基酸(Glu 和 Asp)最不倾向于出现在界面上。在蛋白质-DNA 复合物当中，DNA 主链原子和氨基酸残基形成接触更多，因此静电互补使得正电氨基酸的界面偏好比在蛋白质－蛋白质及蛋白质-RNA 复合物中更高[22]。在 4 种核苷酸中，核苷酸 G 和 T 的界面偏好相对较高，这与它们参与特异性识别有关，核苷酸 G 的结构使得它能够与氨基酸侧链形成更多的氢键[24]。

此外，本文分析了蛋白质二级结构的界面偏好性，结果如图 1(b)所示。从图中可以看出，π-helix 和 β-ladder 的界面偏好性最高，而 3_{10}-helix 和 turn 最不倾向于出现在结合界面上。有研究表明，二级结构的界面偏好和 DNA 结合特异性有关[25]。螺旋是转录因子界面上常见的二级结构，而折叠区域倾向于出现在限制性内切酶等酶类的界面上，使得氨基酸侧链通过 DNA 沟槽与碱基形成氢键相互作用。另一方面，片层结构有利于特异性识别区域保持高度序列保守性[26]。

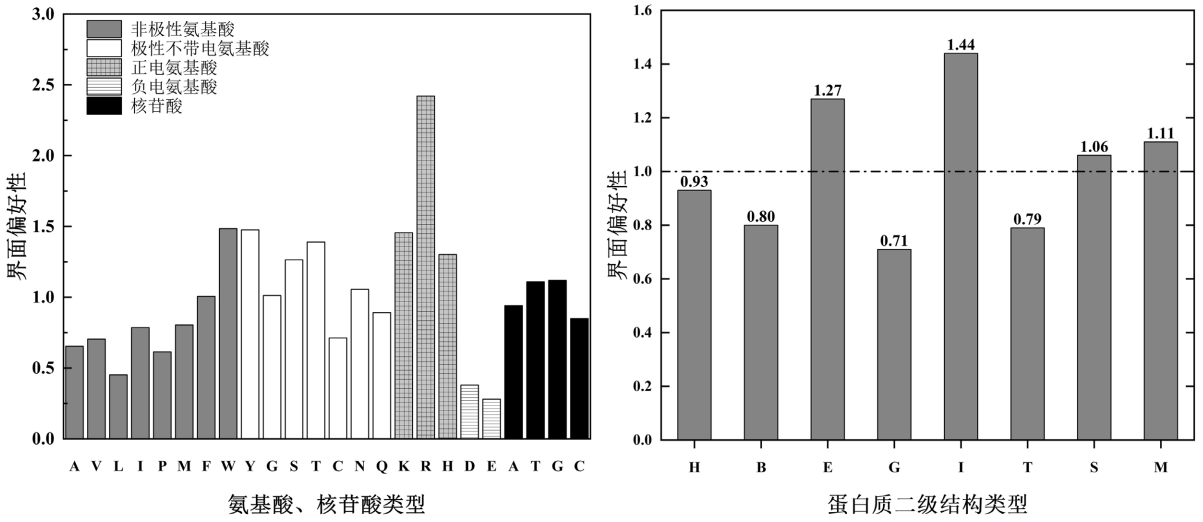


Figure 1. Interface preferences of protein-DNA residue types and secondary structure elements. (a) Interface preferences of amino acids and nucleotides. (b) Interface preferences of protein secondary structure elements
图 1. 蛋白质-DNA 残基和二级结构界面偏好性。(a) 氨基酸和核苷酸界面偏好性。(b) 蛋白质二级结构界面偏好性

3.2. 氨基酸－核苷酸成对界面偏好性分析

本文根据蛋白质二级结构界面偏好性，将其分为 3 类：X (B, G, T; P < 1)，Y (H, S, M; P ≈ 1)，Z (E, I; P > 1)。本文在此基础上，构建了考虑蛋白质二级结构的 60 × 4 氨基酸－核苷酸成对界面偏好。

在所有类型氨基酸和核苷酸的界面成对中，Arg_Z-G (氨基酸_蛋白质二级结构类型－核苷酸)成对的界面偏好最高，为 4.22；Glu_X-A 成对的界面偏好最低，为 0.15。氨基酸 Arg 成对的界面偏好最高，但在所有类型 Arg-核苷酸的成对中依然存在很大差异，其中 Arg_X-C 成对的界面偏好(1.81)相对较低，这可能与蛋白质二级结构类型差异和核苷酸特异性识别有关。在蛋白质二级结构类型当中，通常二级结构类型 X 成对的界面偏好最低，二级结构为 Z 时的成对界面偏好更高。其中也存在一些特殊情况，比如在 Tyr-核苷酸成对中，Tyr_Y-G 和 Tyr_Y-T 的界面偏好最高，分别为 1.83 和 1.82。Y 类型二级结构的界面偏好要更高，可能是因为 Tyr 更倾向于通过 α 螺旋结构参与和 DNA 碱基的特异性识别，这有助于形成更多的非键相互作用。

3.3. 氨基酸界面偏好用以区分真实界面与表面区域

本文使用 60 × 4 界面偏好来区分真实界面和表面区域，统计了每个体系真实界面在对应所有表面区域中的平均氨基酸界面偏好排名，结果如图 2 所示。在总共 47 个复合物的体系中，有 26 个体系真实界面的平均界面倾向打分能够排在前 1%，真实界面排名前 10%的体系占到了所有体系的 87.23% (41/47)。使用 Z 检验验证了真实界面内的平均氨基酸界面偏好与表面区域相比具有显著的统计学差异。对于数据集中的 47 个复合物体系，97.87% (46/47)的体系中真实界面具有明显的氨基酸界面偏好特征(Z < 1.64，即真实界面的平均氨基酸界面偏好不小于 95%的表面区域)，说明 60 × 4 界面偏好能够很好地区分真实界面和其他表面区域。

图 3 中以 1HJC_A:BC 体系为例，将平均氨基酸界面偏好打分最高的表面区域和真实界面进行了比较。从图中可以看出，打分最高的表面区域与真实界面大小基本一致，且该表面区域包括了真实界面中 90%以上的氨基酸残基。结果再次证明本文构建的 60 × 4 界面偏好可以准确预测结合界面。

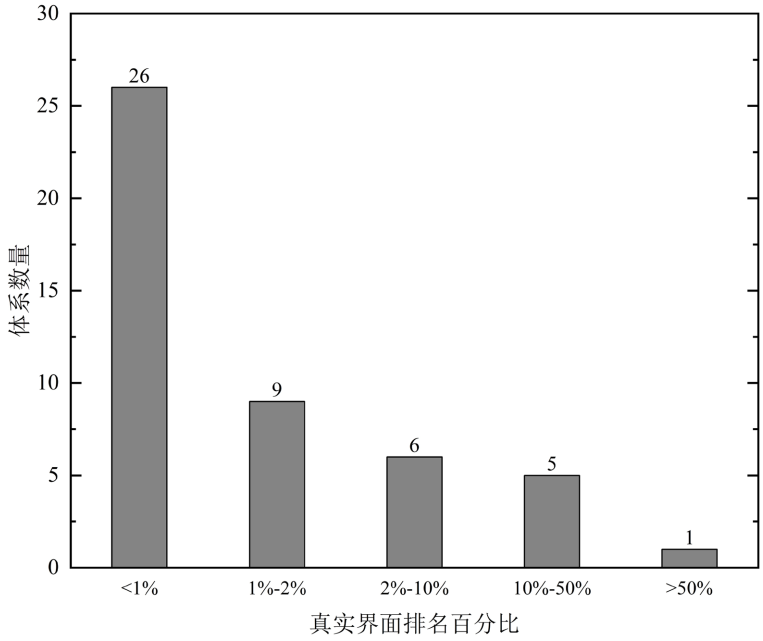
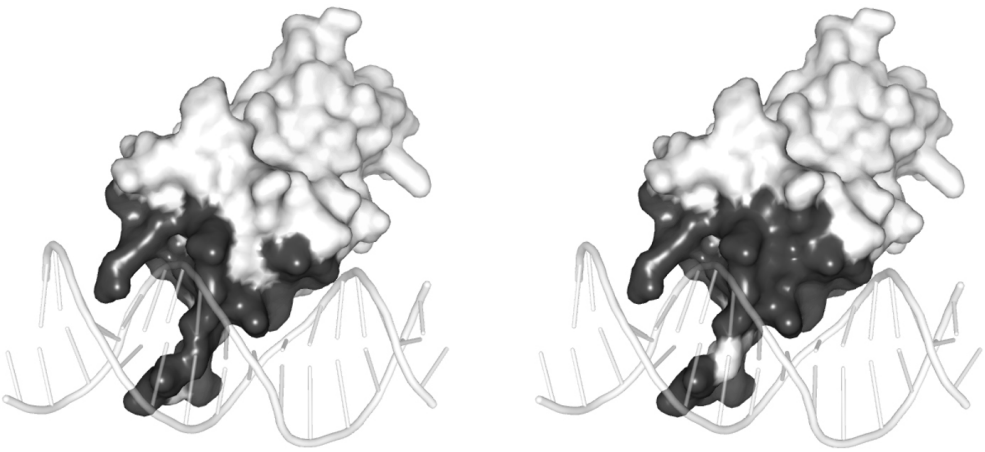


Figure 2. Ranking distribution of the real interface relative to all surface patches according to the average interface preference of amino acids
图 2. 根据平均氨基酸界面偏好对真实界面在所有表面区域的排序分布



(a) (b)

Figure 3. Comparison of the real interface and the best surface patch evaluated with 60 × 4 interface preference. (a) The real interface of the complex. (b) The surface patch with the best score. Interface and surface residues are shown in dark
图 3. 1HJC_A:BC 体系中真实结合界面与 60 × 4 界面偏好打分最高的表面区域比较。(a) 复合物真实界面。(b) 打分最高的表面区域。图中界面和表面残基均用深色表示

4. 结论

研究蛋白质-DNA 复合物中残基和二级结构的界面偏好性，对了解蛋白质-DNA 识别具有重大意义。本文提取了蛋白质二级结构特征，提出了蛋白质-DNA 60 × 4 氨基酸－核苷酸成对偏好性。我们发现 π-helix 和 β-ladder 是界面偏好最高的二级结构类型，而且在此二级结构类型中的氨基酸 Arg 与核苷酸 G 的界面成对偏好最高。分析结果表明 60 × 4 成对偏好性可以体现不同二级结构中氨基酸界面偏好的差异，

很好地反映了复合物间特异性识别规律。使用该界面偏好性信息识别复合物的真实界面，在 87.23%的体系中真实界面打分排在所有表面区域的前 10%。该工作证明了成对偏好性在预测蛋白质-DNA 结合界面中的能力，未来有望用于构建分子对接打分函数，为蛋白质-DNA 复合物结构预测及相关药物设计提供帮助。

基金项目

国家自然科学基金项目(31971180)。

参考文献

[1] Luscombe, N.M., Austin, S.E., Berman, H.M., *et al*. (2000) An Overview of the Structures of Protein-DNA Complexes. *Genome Biology*, **1**, S1. https://doi.org/10.1186/gb-2000-1-1-reviews001

[2] Corona, R.I., Sudarshan, S., Aluru, S., *et al*. (2018) An SVM-Based Method for Assessment of Transcription Factor-DNA Complex Models. *BMC Bioinformatics*, **19**, Article No. 506. https://doi.org/10.1186/s12859-018-2538-y

[3] Berman, H.M., Westbrook, J., Feng, Z., *et al*. (2002) The Nucleic Acid Database. *Acta Crystallographica*. *Section D, Biological Crystallography*, **58**, 889-898. https://doi.org/10.1107/S0907444902003487

[4] Berman, H.M., Westbrook, J., Feng, Z., *et al*. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235-242. https://doi.org/10.1093/nar/28.1.235

[5] Qin, S. and Zhou, H.X. (2011) Structural Models of Protein-DNA Complexes Based on Interface Prediction and Docking. *Current Protein and Peptide Science*, **12**, 531-539. https://doi.org/10.2174/138920311796957694

[6] Steven, A.C. and Baumeister, W. (2008) The Future Is Hybrid. *Journal of Structural Biology*, **163**, 186-195. https://doi.org/10.2174/138920311796957694

[7] Parisien, M., Freed, K.F. and Sosnick, T.R. (2012) On Docking, Scoring and Assessing Protein-DNA Complexes in a Rigid-Body Framework. *PLOS ONE*, **7**, e32647. https://doi.org/10.1371/journal.pone.0032647

[8] Xu, B., Yang, Y., Liang, H., *et al*. (2009) An All-Atom Knowledge-Based Energy Function for Protein-DNA Threading, Docking Decoy Discrimination, and Prediction of Transcription-Factor Binding Profiles. *Proteins*, **76**, 718-730. https://doi.org/10.1002/prot.22384

[9] Tuszynska, I., Magnus, M., Jonak, K., *et al*. (2015) NPDock: A Web Server for Protein-Nucleic Acid Docking. *Nucleic Acids Research*, **43**, W425-W430. https://doi.org/10.1093/nar/gkv493

[10] Tuszynska, I. and Bujnicki, J.M. (2011) DARS-RNP and QUASI-RNP: New Statistical Potentials for Protein-RNA Docking. *BMC Bioinformatics*, **12**, Article No. 348. https://doi.org/10.1186/1471-2105-12-348

[11] Robertson, T.A. and Varani, G. (2007) An All-Atom, Distance-Dependent Scoring Function for the Prediction of Protein-DNA Interactions from Structure. *Proteins*, **66**, 359-374. https://doi.org/10.1002/prot.21162

[12] Li, C.H., Cao, L.B., Su, J.G., *et al*. (2012) A New Residue-Nucleotide Propensity Potential with Structural Information Considered for Discriminating Protein-RNA Docking Decoys. *Proteins*, **80**, 14-24. https://doi.org/10.1002/prot.23117

[13] 陆林, 刘洋, 李春华. 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[J]. 生物化学与生物物理进展, 2020, 47(7): 634-644.

[14] Pabo, C.O. and Sauer, R.T. (1984) Protein-DNA Recognition. *Annual Review of Biochemistry*, **53**, 293-321. https://doi.org/10.1146/annurev.bi.53.070184.001453

[15] Coimbatore, N.B., Westbrook, J., Ghosh, S., *et al*. (2014) The Nucleic Acid Database: New Features and Capabilities. *Nucleic Acids Research*, **42**, D114-D122. https://doi.org/10.1093/nar/gkt980

[16] Li, W. and Godzik, A. (2006) Cd-hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics*, **22**, 1658-1659. https://doi.org/10.1093/bioinformatics/btl158

[17] van Dijk, M. and Bonvin, A.M. (2008) A Protein-DNA Docking Benchmark. *Nucleic Acids Research*, **36**, e88. https://doi.org/10.1093/nar/gkn386

[18] Lee, B. and Richards, F.M. (1971) The Interpretation of Protein Structures: Estimation of Static Accessibility. *Journal of Molecular Biology*, **55**, 379-400. https://doi.org/10.1016/0022-2836(71)90324-X

[19] Kabsch, W. and Sander, C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577-2637. https://doi.org/10.1002/bip.360221211

[20] Jones, S. and Thornton, J.M. (1997) Analysis of Protein-Protein Interaction Sites Using Surface Patches. *Journal of Molecular Biology*, **272**, 121-132. https://doi.org/10.1006/jmbi.1997.1234

[21] Yang, Z., Deng, X., Liu, Y., *et al*. (2020) Analyses on Clustering of the Conserved Residues at Protein-RNA Interfaces and Its Application in Binding Site Identification. *BMC Bioinformatics*, **21**, Article No. 57. https://doi.org/10.1186/s12859-020-3398-9

[22] Kulandaisamy, A., Srivastava, A., Nagarajan, R., *et al*. (2018) Dissecting and Analyzing Key Residues in Protein-DNA Complexes. *Journal of Molecular Recognition*, **31**, e2692. https://doi.org/10.1002/jmr.2692

[23] Nadassy, K., Wodak, S.J. and Janin, J. (1999) Structural Features of Protein-Nucleic Acid Recognition Sites. *Biochemistry*, **38**, 1999-2017. https://doi.org/10.1021/bi982362d

[24] Bahadur, R.P., Zacharias, M. and Janin, J. (2008) Dissecting Protein-RNA Recognition Sites. *Nucleic Acids Research*, **36**, 2705-2716. https://doi.org/10.1093/nar/gkn102

[25] Corona, R.I. and Guo, J.T. (2016) Statistical Analysis of Structural Determinants for Protein-DNA-Binding Specificity. *Proteins*, **84**, 1147-1161. https://doi.org/10.1002/prot.25061

[26] Lin, M. and Guo, J.T. (2019) New Insights into Protein-DNA Binding Specificity from Hydrogen Bond Based Comparative Study. *Nucleic Acids Research*, **47**, 11103-11113. https://doi.org/10.1093/nar/gkz963