基于三元残基组合对的蛋白质相互作用研究

肖　薇[1, 2†], 何增辉[2†], 李诗良[2], 李洪林[1, 2*]

(华东理工大学 1. 信息科学与工程学院, 2. 药学院, 上海市新药设计重点实验室, 上海 200237)

摘要: 蛋白质−蛋白质相互作用的作用机制对生命科学研究有着重要意义。目前已有的方法多偏向于氨基酸残基的偏好性研究, 并没有给出对应的残基组合的空间信息, 而这些空间信息对设计蛋白质−蛋白质相互作用至关重要。通过深入挖掘已有的蛋白质相互作用模式, 并提炼残基相互作用对的偏好和相对位置信息, 本文提出了一种全新的既能表征三元残基组合的偏好, 又能给出三元残基组合对的空间信息的“三棱柱”模型。该模型主要从偏好因子、氨基酸组成和蛋白质二级结构分布等多个方面对三元残基组合对进行分析。此外, 还将该模型应用于 PD-1/PD-L2 蛋白质的界面研究。通过分析 PD-1/PD-L2 蛋白质的界面残基组合对与预测残基组合对在组成和空间信息上的差异, 给出了具体的残基突变建议, 从而为蛋白质−蛋白质相互作用的设计提供了一种新的方法。

关键词: 蛋白质−蛋白质相互作用; “三棱柱”模型; 三元残基组合对

中图分类号: R916　　**文献标识码**: A　　**文章编号**: 0513-4870 (2017) 10-1578-09

Study of protein interaction based on the triplets combination pairs of the residue groups

XIAO Wei[1, 2†], HE Zeng-hui[2†], LI Shi-liang[2], LI Hong-lin[1, 2*]

(*1. School of Information Science and Engineering, 2. Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China*)

Abstract: The protein-protein interactions play an important role in life science. At present, many methods are developed with preferences of the amino acid residues, which do not offer the relative spatial information for the residue groups. However, the spatial information for the residue groups is important in the design of the protein-protein interactions. We proposed a new model, which is named ‘tri-prism’ model, by deep mining the existing protein-protein interaction patterns and refining the preference and the relative spatial information for the combination pairs of the residue groups. The model not only provided the preferences, but also offered the relative spatial information for the triplets combination pairs of the residue groups. The model was able to analyze the triplets combination pairs of the residue groups based on the preference factor, amino acid composition, and protein secondary structure. The model was applied to the interface of the PD-1/PD-L2 protein. According to the diversity characters of the composition and the spatial information between the combination pairs of the residue groups at the interface of the PD-1/PD-L2 protein and the predicted ones, we put forward the suggestions for the mutations of the residues, which offered a new view in the study of protein-protein interactions.

Key words: protein-protein interaction; ‘tri-prism’ model; the triplets combination pairs of the residue group

收稿日期: 2017-05-02;　修回日期: 2017-06-23.

基金项目: 国家重点研发计划资助课题 (2016YFA0502304); 中央高校基本科研业务费 (222201717024).

†共同第一作者.

*通讯作者 Tel / Fax: 86-21-64250213, E-mail: hlli@ecust.edu.cn

DOI: 10.16438/j.0513-4870.2017-0419

蛋白质−蛋白质相互作用 (protein-protein interactions, PPI) 在很多生命活动中起着重要作用, 例如细胞调节、信号转导、蛋白质合成与分泌、核酸复制、病毒包装及免疫反应[1−3]等。而蛋白质−蛋白质相互作用界面性质的研究是其特异性识别的前提。氨基酸作为蛋白质−蛋白质相互作用界面的基本组成单元, 其含有丰富的种类和侧链朝向信息。使用基于氨基酸残基的方法可以方便地探究蛋白结合界面的相互作用模式。该模式有助于解释蛋白质−蛋白质识别的分子机制, 并且可应用于蛋白质工程、药物设计等多个研究领域。

早期研究者主要采用数学和统计学的方法来研究蛋白界面上残基对的偏好[4, 5]。这种研究方法对蛋白界面有非常直白的解释, 但它不能提供除了残基对之外更多的信息, 也无法解释蛋白质复合物结构与其生物功能之间的联系。随后, 人们将注意力集中到热点残基上, 即蛋白结合界面上对蛋白质及其配体的结合有重大影响的关键残基[6, 7]。例如 Tuncbag 等[8]发展的用 3 种特征来预测热点残基的算法, 该方法预测的热点残基大致符合结合界面数据库[9] (binding interface database, BID) 中实验确认的热点残基。Kozakov 等[10]描述了一种通过计算溶剂映射来预测蛋白结合界面成药性热点残基的方法, 利用 FTMAP 算法[11]在整个蛋白界面进行全局搜索来寻找有利于特定探针结合的区域, 不同探针的重叠域被定义为共同位点, 最大的共同位点被预测为蛋白结合界面最重要的位点。此外, 还有一些预测热点残基的计算方法不断涌现出来, 如 Robetta[12]、HotPoint[13]、KFC[14]、PRICE[15]和 PCRPi-W[16]等。

随着对热点残基研究的日益加深, 人们发现热点残基并不是孤立的发挥作用。Bogan 和 Thorn[17]发现热点残基一般被对结合能贡献较小的残基包围着, 并包埋于界面中间。和其他界面残基相比, Keskin 等[18, 19]发现热点残基更趋于保守, 且与周围残基有着协同作用, 这使得热点残基对结合能的贡献远高于其他界面残基。此外, 蛋白质−蛋白质相互作用界面的特征几何模型研究也日益增加。Mintz 等[20]发展了一种挖掘蛋白相互作用界面上相似的相互作用模型的方法。该方法将每个残基的功能基团简化为一些“伪中心”, 即将蛋白质−蛋白质结合界面简化为一系列“伪中心”, 并使用 I2I-SiteEngine 算法[21]进行结合界面的几何比对。该研究发现, 高匹配得分的蛋白质−蛋白质结合界面拥有相似的相互作用模型。Li 等[22]提出一种“四面体”模型可用于蛋白相互作用表面上热点残基的预测、蛋白功能等研究。

目前已有的方法多偏向于残基的偏好性研究, 并没有给出对应的残基组合间的空间信息, 而这些空间信息对蛋白质−蛋白质相互作用界面的设计至关重要。通过深入挖掘已有的蛋白质−蛋白质相互作用模式, 提炼出残基组合对的偏好和相对位置信息, 作者提出了一种全新的既能表征三元残基组的偏好性, 又能给出三元残基组合对空间信息的“三棱柱”模型。通过将该模型应用于 PD-1/PD-L2 蛋白界面的研究, 主要分析了蛋白质−蛋白质界面残基组合对与预测残基组合对在组成和空间信息上的差异, 并给出了具体的残基突变建议。实验结果表明, 本研究给出的残基突变建议可以增强 PD-1/PD-L2 之间的结合亲和力。

材料与方法

数据集准备　截止 2014 年 10 月, 作者从蛋白质数据库 (protein data bank, PDB) 中下载了 104 371 个蛋白质晶体结构。为了提取有效的蛋白质−蛋白质结合界面, 基于 Tsai[21]方法采用如下标准对蛋白质晶体结构进行过滤: ① 保留分辨率小于等于 2.5 Å, 且非 NMR 或 EM 技术获得的蛋白质晶体结构; ② 筛除非标准氨基酸或者结构信息不完整的氨基酸; ③ 筛除残基数少于 10 的蛋白质结合界面; ④ 有效的蛋白质−蛋白质结合界面由两个独立的蛋白分子的两条肽链组成; ⑤若两个蛋白分子的重原子对间距离小于它们的范德华半径 (van der Waals radii) 之和再加上 0.5 Å 的偏差, 那么这两个重原子所在的氨基酸就被认为是界面残基, 所有界面残基的集合形成结合界面。通过上述筛选标准, 最终获得了 6 715 个蛋白质复合物晶体结构留作训练数据。

对于这些训练数据, 将结构中非同源肽链数目大于 3 的蛋白结构排除; 若非同源肽链数目为 3, 保留其中界面残基数目最多的两条链视为相互作用链。这样, 共有 6 405 个蛋白质−蛋白质相互作用界面被筛选出来。此外, 为了确保在蛋白结合界面能够有效地提取到三元残基组合对, 限定蛋白结合界面上每条相互作用链上的界面残基数目不少于 6 个。最终保留了 6 122 个蛋白质−蛋白质相互作用界面, 见支撑材料中表 S1。表 S1 包含蛋白质的 PDB 编号及两条相互作用链的名称, 例如, (2PVG-AB) 表示蛋白−蛋白相互作用界面由 PDB 编号为 2PVG 的晶体结构的 A 链和 B 链构成。

“三棱柱”模型　在数据集准备中, 已经提取

了 6 122 个蛋白结合界面，并保存其中的相互作用链和界面残基信息。每条链上的界面残基可能和多个界面残基相邻，通过选取和它距离最近的两个界面残基组成三元残基组合，提出了一种“三棱柱”模型(图 1)。每个三元残基组合对由 2 个三元残基组合组成。其中每个三元残基组合由 3 个氨基酸残基组成；每个氨基酸的主链碳原子到侧链特征中心的空间向量构成了“三棱柱”的三条“棱”(A、B、C)。而三元残基组合的 3 个残基向量起点和终点位置的几何中心 (o1、o2、o3、o4) 分别构成了中心向量的起点和终点。该模型不但可以表征三元残基组合的偏好性，同时还提供三元残基组合对之间的相对空间信息，包括距离和角度等。而这些信息可用于进一步预测蛋白质相互作用等研究。

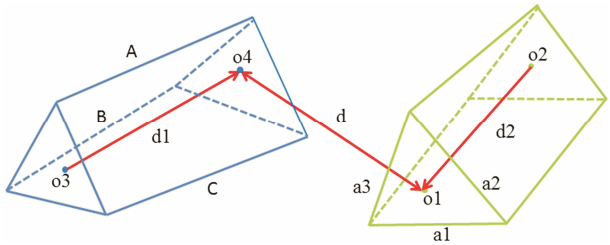

Figure 1　Architecture of the “Tri-prism” model.　A, B and C stand for the vector of the carbon atom of main chain to the characteristics of the side chain, respectively.　o1 and o3 stand for the center of the carbon atom of the main chain, respectively; o2 and o4 stand for the center of the characteristics of the side chain, respectively.　*d*1 and *d*2, respectively, stand for the vector length of the center of the carbon atom of the main chain to the center of the characteristics of the side chain in two different chains.　*d* represents the distance between the two vector *d*1 and *d*2.　*a*1, *a*2, *a*3, respectively, stand for the lengths of three sides of the carbon atom of main chain

三元残基组合的确定　三元残基组合为蛋白质界面残基与其距离最近残基的组合。为了量化界面残基间的距离，对每一个氨基酸残基都用一个确切的特征中心来表征。这个特征中心可能由主链或侧链上的某个特征原子表示，也可能由侧链上的环的几何中心表示。两个特征中心之间的距离代表两个残基之间的距离。通过该方法对 6 122 个蛋白结合界面上的残基进行三元残基组合的提取，保留其中 3 个残基向量都朝向界面的三元残基组合。此外，根据每种氨基酸残基的侧链长度和性质的不同，将 20 种标准氨基酸分为 9 大类[22] (表 1)，三元残基组合及三元残基组合对也将采用此类别进行表征。

三元残基组合对的确定　对 6 122 个蛋白结合界面上的三元残基组合采用如下步骤进行配对：每条链上的三元残基组合和与它相互作用链上距离最近的三元残基组合组成三元残基组合对。考虑到与某个三元残基组合相互作用链上最近的组合可能存在一个或者多个，因此，同一个三元残基组合可以组成一个或者多个三元残基组合对。对于这些三元残基组合对，主要计算下列相对空间信息：① 三元残基组合主链中心 o1 到对应三元残基组合侧链中心 o4 之间的距离 *d*；② 两个中心向量的模 *d*1、*d*2；③ 两个中心向量的夹角 *θ*；④ 三元残基组合主链三角形的三边边长 *a*1、*a*2、*a*3；⑤ 三元残基组合侧链三角形中心 o4 到对应的三元残基组合主链三角形 3 个顶点之间的距离 *b*1、*b*2、*b*3。

Table 1　Definition of the feature centers on 20 amino acid residues

Type	Description	Residue	Location of the feature center
0	Residues with a small side chain	Gly (G)	*α*-Carbon atom on the backbone
		Ala (A)	*β*-Carbon atom on the side chain
1	Residues with a bulky hydrophobic side chain	Val (V)	*β*-Carbon atom on the side chain
		Leu (L)	*γ*-Carbon atom on the side chain
		Ile (I)	Geometric center of *β*, *γ*1, *γ*2, and *δ*-carbon atoms on the side chain
		Cys (C)	*γ*-Sulfur atom on the side chain
		Met (M)	*δ*-Sulfur atom on the side chain
2	Residues with an aromatic side chain	Phe (F)	Geometric center of the phenyl ring
		Tyr (Y)	Geometric center of the phenol ring
		Trp (W)	Geometric center of the indole ring
3	Residues with a side chain that has a hydroxyl group	Ser (S)	*γ*-Oxygen atom on the side chain
		Thr (T)	*γ*-Oxygen atom on the side chain
4	Residues with a side chain that has a carboxylic acid group	Asp (D)	*γ*-Carbon atom on the side chain
		Glu (E)	*δ*-Carbon atom on the side chain
5	Residues with a side chain that has an amide group	Asn (N)	*γ*-Carbon atom on the side chain
		Gln (Q)	*δ*-Carbon atom on the side chain
6	Residues with a side chain that has a positively charged group	Arg (R)	*ζ*-Carbon atom on the side chain
		Lys (K)	*ζ*-Nitrogen atom on the side chain
7	All histidine residues	His (H)	Geometric center of the imidazole ring
8	All proline residues	Pro (P)	Geometric center of the pyrrole ring

这些三元残基组合对的空间信息存在一些明显的极大值或者极小值, 可能是由于部分三元残基组合在另一条链附近没有找到配对组合, 导致最终与之配对的组合距离过远; 或者是因为有些残基位于界面边缘, 造成三元残基组合对的两个中心向量的夹角过小或距离过近。上述情况虽然只为少数, 但会给分析预测造成一定的干扰。因此, 设置如下条件对三元残基组合对进行过滤: 1) $d \leqslant 8$Å; 2) $\theta \geqslant 60°$; 3) 2Å $\leqslant a1$ ($a2$, $a3$) $\leqslant$ 10Å; 4) 4Å $\leqslant b1$ ($b2$, $b3$) $\leqslant$ 12Å。

结果

1　三元残基组合对的偏好性

通过采用“三棱柱”模型, 从这些蛋白质数据集中共找到 17 541 个三元残基组合对。根据三元残基组合对中氨基酸所属类别 (表 1) 对三元残基组合 (对) 进行编号, 例如, 三元残基组合 (SER,PHE,LEU) 被编号为 (1,2,3) (组合内部编号无对应先后顺序), 而三元残基组合对 (SER,PHE,LEU)−(ASN,ILE,ILE) 被编号为 (1,2,3−1,1,5)。这些三元残基组合对为蛋白质−蛋白质相互作用的研究提供了预测和参考依据。

1.1　偏好因子　为了表征三元残基组合间的偏好性, 引入偏好因子 (preference factor, PF), 计算公式如式 (1) 所示:

$$\mathrm{PF}=\frac{N_{\mathrm{combination}}\Big/\sum_i N_{\mathrm{combination}}}{\left(N_{\mathrm{pattern\_a}}\Big/\sum_j N_{\mathrm{pattern}}\right)*\left(N_{\mathrm{pattern\_b}}\Big/\sum_j N_{\mathrm{pattern}}\right)} \quad (1)$$

其中, $N_{\mathrm{combination}}/\sum_i N_{\mathrm{combination}}$ 为数据集中该类三元残基组合对出现的概率, $N_{\mathrm{pattern\_a}}/\sum_j N_{\mathrm{pattern}}$, $N_{\mathrm{pattern\_b}}/\sum_j N_{\mathrm{pattern}}$ 分别为三元残基组合对中三元残基组合 a 和三元残基组合 b 出现的概率。

根据公式 (1) 得到每个三元残基组合对的 PF 值 (详细数据见支撑材料中表 S2), 其分布如图 2 所示。其中, PF 值 13.6 对应的三元残基组合对数目是其分布中的一个较高峰值, 可以用来将 PF 值的分布大致划分为 2 个部分。当 PF 值小于 13.6 时, 三元残基组合对的出现频次呈现一个先上升后下降的接近正态分布的“头部”。而当 PF 大于 13.6 时, 三元残基组合对的出现频次呈现出一个长长的“尾部”, 并且这个尾部包含一些出现频次较高的峰值。对于大部分的三元残基组合对, PF 值都小于 13.6。同时, 也存在少数 PF 值超过 100.0 的高 PF 值的三元残基组合对。当然, PF 值越高, 表明该三元残基组合成对的偏好性越强, 这也就说明存在部分三元残基组合对具有很强的选择偏好性。

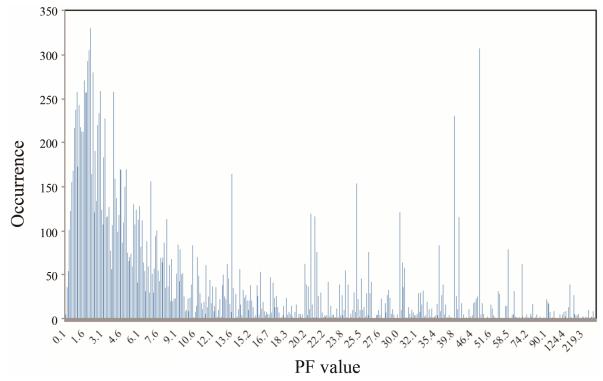

Figure 2　Distribution of preference factor (PF) values for the triplets combination pairs of the residue groups

1.2　高 PF 值的三元残基组合对　对于 PF 值分布中最大的三元残基组合对, 即组合对编号为 (2,7,7−7,7,7) 的三元残基组合对, 其 PF 值高达 2 863.8。这不仅远高于整体 PF 的平均值, 也远高于排序第二 (PF 值为 692.4) 的三元残基组合对。这两个最高的三元残基组合对均来源于同一个蛋白质晶体结构 (PDB 编号: 4GGF, 表 2)。该晶体结构 4GGF 为钙结合蛋白, 是一种由 S100A8/S100A9 组成的异源二聚体, 其在宿主应对病原菌感染的“营养免疫”(nutritional immunity) 中发挥作用。钙结合蛋白通过高亲和力结合相互作用界面的锰离子和锌离子, 从而使细菌缺乏该营养物质而死亡[23]。其中, 该蛋白的 6 个组氨酸 (HIS) 与锰离子的螯合作用 (图 3) 是其发挥功能的重要原因。表 2 的两个三元残基组合对中的 HIS 均来源于这 6 个与锰离子有螯合作用的 HIS。由于锰离子使 (2,7,7) 和 (7,7,7) 这两种不常见的三元残基组合发生特异性结合, 造成这个三元残基组合对 (2,7,7−7,7,7) 的 PF 值非常高。

根据 PF 值的分布可知, 仅有 0.2% 的三元残基组合对的 PF 值大于 200.0。而这些三元残基组合对几乎都含有 HIS (氨基酸类别 7) 或 PRO (氨基酸类别 8),

Table 2　The highest PF values of the triplets combination pairs of the residue groups

No.	PDB code	Chain 1	The triplets combination pairs of the residue group	Chain 2	The triplets combination pairs of the residue group
1	4GGF	A	(H-17, F-26, H-27)	C	(H-105, H-91, H-95)
2	4GGF	A	(H-17, F-26, H-27)	C	(H-103, H-105, H-95)

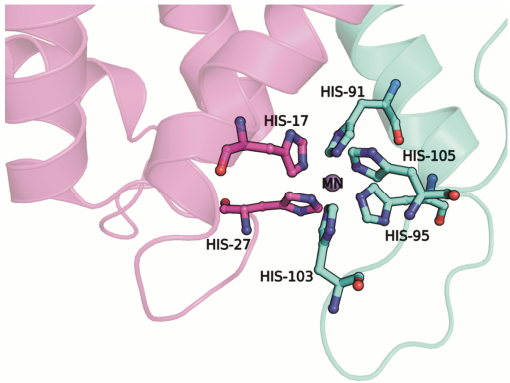

Figure 3 Six histidine residues chelate a Mn ion in the crystal structure of 4GGF

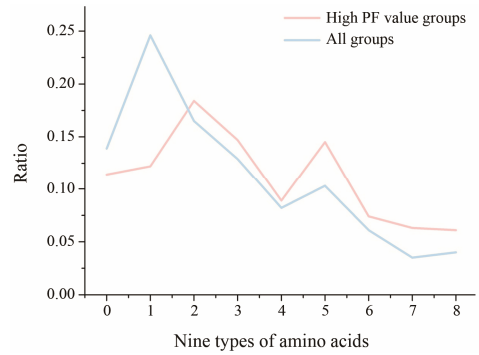

Figure 4 Comparison of the combination of nine types of amino acids between the high PF value groups and the overall groups

例如 (2,7,7-7,7,7)、(0,8,8-6,6,8)、(4,7,7-3,4,7)、(7,8,8-4,6,8) 等。这些高 PF 值的三元残基组合对大多是因为形成特殊结构造成异常高的选择性。而大部分三元残基组合对的 PF 值都在 0.0 到 200.0 范围内。为了将偏好性高的组合挑选出来，把 PF 值分布中处于“尾部”，即 PF 值大于 13.6 的三元残基组合对，视为高 PF 值组，并将之与整体组进行比较，对比结果如表 3 所示。和整体组相比，高 PF 值组的三元残基组合对具有较小 d 值和较大的 $d1$、$d2$ 和 θ 值，PF 值的均值也远大于整体 PF 值的平均水平。这就说明，与整体相比，高 PF 值组可能含有更多的长链氨基酸，而这些长链氨基酸具有比短链氨基酸更强的疏水相互作用或者极性相互作用，这也可能是该组别 PF 值高的原因。

Table 3 Comparison of the high PF values groups against the overall groups

Category	d	$d1$	$d2$	θ	Mean value of PF
The high PF values groups	5.012	2.362	2.546	128.2	39.4
The overall groups	5.457	2.202	2.286	123.0	13.9

2 三元残基组合对的氨基酸组成分析

为了探究三元残基组合对如何利用氨基酸残基介导蛋白质−蛋白质相互作用，统计了蛋白质−蛋白质相互作用界面上的 9 类氨基酸组成，并与高 PF 值组的氨基酸组成进行对比。从图 4 可以看出，与整体相比，高 PF 值组的三元残基组合对中含有较多 5、6、7 和 8 这 4 类氨基酸和较少的 0、1 类氨基酸。而根据表 1 中氨基酸的分类得知，0、1 类氨基酸是短链疏水氨基酸，而 5、6、7、8 类氨基酸更多的是长链亲水氨基酸。这也进一步验证了，高 PF 值组的三元残基组合对较整体的 $d1$、$d2$ 值更大，d 值更小，更为形象的说是，中心向量更长，距离更近。此外，5、6、7 类氨基酸带有电荷，因此，含有这 4 类氨基酸的三元残基组合对具有更强的亲和力和较高的互补性和特异性，也就表现为 PF 值更高。而整体组包含较多的 1、2 类氨基酸，这类氨基酸更多表现为疏水相互作用，具有可替代性更高和偏好性不明显的特性，也即表现为 PF 值偏低。

同样地，对高 PF 值组和整体组的三元残基组合对中 20 种标准氨基酸的组成进行统计 (图 5)，有些结果并不出乎意料。首先，无论是在高 PF 值组还是整体组，GLY 含量都非常丰富。这是因为 GLY 在 loop、helix 等蛋白质二级结构中都占据较大比例[24]。此外，PHE 等疏水性氨基酸也具有较高含量，这些氨基酸通常被认为参与介导蛋白质−蛋白质相互作用。特别地，PHE 的含量在高 PF 值组和整体组中都很高，这也从侧面说明三元残基组合很可能利用 PHE 来识别其对应的三元残基组合，类似作用的还有 SER。与在整体组的表现一般相比，PRO 在高 PF 值组的含量大约为 50%，这与其较多地出现在 loop 结构中扮演蛋白质二级结构破坏者的角色有关。调查显示，26% 的界面残基属于 α 螺旋结构，24% 的界面残基属于 β 折叠结构，而 50% 的界面残基在 loop 结构中[25]。所以推测含有 PRO 的三元残基组合对可能利用 loop 二级结构来介导蛋白质−蛋白质相互作用。高 PF 值组比整体组含有更多的带电荷氨基酸来维持三元残基组合之间相互作用。而对于 ARG、GLU 和 LYS 这 3 类氨基酸，其含量在高 PF 值组与整体组中的差别不大。但 HIS 在高 PF 值组的含量远高于整体组，这不仅与其具有带电荷氨基酸所共有的极性相互作用，同时它还可以同时作为氢键供体和氢键受体。这就使 HIS 可发生极性相互作用的范围更广，因此其在高 PF 值组的含量会比整体组的高。

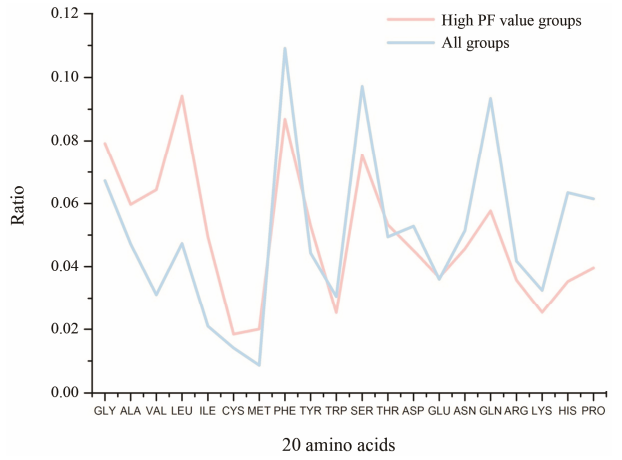

Figure 5　Comparison of the combination of 20 amino acids in the triplets combination pairs of the residue groups between the high PF value groups and the overall groups

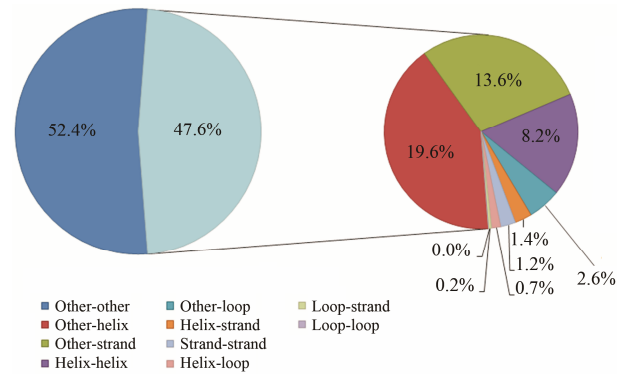

Figure 6　Composition of the secondary structure of the triplets combination pairs of the residue groups. The Pie chart in light blue on the left side represents the protein secondary structure, and the other one in dark blue represents the protein non-secondary structure. The pie chart on the right side represents the detailed distribution for the protein secondary structure

3　三元残基组合对的二级结构分析

蛋白质的二级结构主要分为螺旋 (helix)、折叠 (strand)、环 (loop) 和其他 (other) 4 种类型。如果三元残基组合中有两个氨基酸残基在同一个二级结构片段上，那么就认为该三元残基组合在这个二级结构片段上。通过使用 STRIDE[26]对三元残基组合对的蛋白质二级结构片段进行识别与统计，结果如图 6 所示。其中，图 6 左侧饼图中浅蓝色块表示含有经典的蛋白质二级结构，可以看到 47.6% 的三元残基组合对至少有一个经典的蛋白质二级结构 (helix、strand 和 loop)。而细分这些蛋白质二级结构 (图 6) 可以看出，有一个螺旋结构的三元残基组合对 (图 6 中“other-helix”) 占所有含二级结构的三元残基组合对的 19.6%，也就说明了螺旋结构在蛋白质二级结构中的含量优势明显。此外，有一个折叠结构的三元残基组合对 (图 6 中“other-strand”) 占全部经典蛋白质二级结构的 13.6%；而有一个 loop 结构 (图 6 中“other-loop”) 的含量仅为 2.6%，这也进一步说明 loop 结构更多是通过一个或两个残基来介导蛋白相互作用。相对于仅有一个蛋白质二级结构的三元残基组合对，同时含有两个蛋白质二级结构的三元残基组合对 (“helix-strand”、“strand-strand”、“helix-loop”、“loop-strand”和“loop-loop”) 要少得多。其中，“helix-helix”组合最多，为 8.2%；而“loop-loop”组合一个也没有，这可能和 loop 结构柔性较大，不利于形成稳定的构象有关。

应用

1　背景介绍

程序性死亡受体 1 (PD-1) 是 CD28/B7 家族中的一员，其在免疫反应负调节中起重要作用[27]。PD-1 与配体 (PD-L1[28]、PD-L2[29]) 结合之后可抑制下游的 NF-κB 转录，从而使肿瘤细胞得以躲避机体的免疫监控和杀伤，也就是说，阻断 PD-1/PD-L2 相互作用能增强机体内源性抗肿瘤免疫效应[30−32]。

目前，人源 PD-1 与配体的复合物结构并未解析出来，其作用机制也尚不清楚。由于人源和鼠源 PD-1 配体的高相似性，研究鼠源 PD-1 与其配体 PD-L2 的作用机制对解释人源 PD-1 与配体的作用方式有一定的参考价值。Stanley 课题组[33]将配体 PD-L2 上的一些残基 (E-28、W-110、D-111、Y-112、K-113 和 Y-114) 进行突变，来探究结合界面的关键残基。其研究发现，D-111 和 K-113 突变能完全破坏与 PD-1 的结合，W-110 突变或删除则大幅削弱了与 PD-1 的结合，Y-114 突变也能很大程度上减少与 PD-1 的结合亲和力等。因此，将这些突变后对结合亲和力产生影响的残基视为关键残基。而对蛋白 PD-1 上的残基 (K-45、I-93、P-97、K-98、A-99、I-101、E-103) 进行突变，发现其中部分残基 (K-45、I-93、I-101 和 E-103) 的突变能够显著降低与 PD-1 和 PD-L2 的结合亲和力，而 A99 的突变反而增强了结合亲和力，因此这些对突变产生影响的残基可视为关键残基。而 P-97 和 K-98 的突变对亲和力的影响较弱。此外，非关键残基 N-51 和 Y-121 的突变对结合几乎没有影响，可视为最佳突变残基。

2　方法

2.1　三元残基组合对的预测步骤　在“材料与方法”部分，从所有训练数据集的蛋白质−蛋白质相互作用界面中共找到 17 541 个三元残基组合对。基于这

些三元残基组合对, 对 PD-1 的三元残基组合进行预测, 包括三元残基组合偏好预测和空间位置预测。其中, 偏好预测给出与 PD-1 配对最常见的 3 类组合和最大 PF 值偏好组合; 空间位置预测则由与 PD-1 配对中最大 PF 值的组合的空间信息给出。具体步骤如下:

首先, 根据统计的三元残基组合对给出与目标三元残基组合配对数目最多的前 3 类组合, 它们的偏好性可能并不强, 但在现有组合对中出现次数最多, 可以作为参考; 其次, 给出与之配对的 PF 值最大的组合及相应的 PF 值。若同一 PF 值下有多个三元残基组合, 均列出; 最后, 给出目标三元残基组合的预测空间信息: ① 给定组合的 3 个残基侧链特征中心到预测配对组合主链三角形的几何中心。相比于侧链, 主链原子更加稳定, 构象变化更小, 这样就得到三维空间的第一个点, 这个点将作为“锚”, 在此基础上进一步构建“三棱柱”; ② 给出预测配对组合中 3 个残基主链碳原子所构成的三角形, 这样可以确定预测组合 3 个残基主链碳原子的相对距离, 便于将预测三元残基组合主链碳原子固定在特定的二级结构骨架上; ③ 给出侧链空间信息的 4 个参数, d、$d1$、$d2$ 和 θ, 即中心向量的长度、角度和距离, 用于约束侧链构象。

2.2 三元残基组合对的预测结果及分析 从蛋白质数据库中下载鼠源 PD-1/PD-L2 蛋白复合物晶体结构 (PDB 编号: 3BP5), 其中 A 链为 PD-1 蛋白, 由 117 个氨基酸组成; B 链为 PD-L2 蛋白, 由 202 个氨基酸组成。用内部脚本提取 PD-1 配体结合表面的三元残基组合, 共 8 个 (表 4)。其中关键残基 (K-45、I-93 和 I-101) 分别出现在表 4 第 1、5 和 6 号组合中, 仅关键残基 E-103 没有包含在所有提取的三元残基组合中。此外, 每个三元残基组合都含有根据蛋白晶体结构预测形成的结合位点的残基 (表 4 中加 a 标注的残基), 这些残基可能和关键残基协同作用增强蛋白界面的相互作用。同样可以看到, 4、5、6 号三元残基组合中还包含 A-99, 即突变后反而增强蛋白结合亲和力的残基。表 4 同样还给出了每个 PD-1 表面的三元残基组合通过“三棱柱”模型预测得到的最大 PF 值的三元残基组合及其 PF 值。

将 PD-1/PD-L2 蛋白表面实际存在的三元残基组合对提取出来, 如表 5 所示。通过与“三棱柱”模型预测的三元残基组合对 (表 4) 进行对比发现, 表 5 中 3 个 PD-1 表面三元残基组合分别出现在表 4 的第 1、2、8 号三元残基组合对中。

Table 5 The triplets combination pairs of the residue groups on the binding interface of PD-1/PD-L2. [b]Stands for the residues that formed the binding site and predicted by the crystal structure in third triplets combination pairs of the residue groups

No.	The triplets combination pairs of the residue groups on the surface of PD-1	PF value	The triplets combination pairs of the residue groups on the surface of PD-L2
1	(N-33, N-35, K-45)	80.9	(T-22, D-111, K-113)
2	(N-33, N-35, G-91)	9.7	(Y-112, K-113, Y-114)
3	(M-31[b], S-50[b], N-51)	1.1	(G-107[b], A-108[b], A-109[b])

对于表 5 中的第 1 号组合 (N-33, N-35, K-45), 在 PD-1/PD-L2 蛋白界面上找到对应的三元残基组合为 (T-22, D-111, K-113), 与预测的三元残基组合 (D, K, T) 一致。PD-1/PD-L2 蛋白界面上的这个三元残基组合对共包含 3 个关键残基 (PD-1 上的 K-45 和 PD-L2 上的 D-111、K-113), 其 PF 值 80.9 高于 PF 的整体平均水平, 说明这个三元残基组合对具有明显的偏好性。

对于表 5 中的第 2 号组合 (N-33, N-35, G-91), 预测与之相互作用的组合为 (D, R, N), 属于 (4, 5, 6) 类三元残基组合。而 PD-1/PD-L2 蛋白界面上找到对应的组合为 (Y-112, K-113, Y-114), 即 (Y, K, Y), 属于 (2, 2, 6) 类的三元残基组合。其中 PD-L2 上的 K-113 和 Y-114 属于关键残基。而 Y-112 属于突变后对蛋白结合没有影响的最佳突变残基。若将 Y-112 突

Table 4 The triplets combination pairs of the residue groups for the crystal structure of 3BP5. [a]Stands for the residues that formed the binding site and predicted by the crystal structure

No.	The triplets combination pairs of the residue groups on the surface of PD-1	PF value	The triplets combination pairs of the residue groups for predicting and their abbreviations	
1	(N-33[a], N-35[a], K-45)	80.9	(ASP, LYS, THR)	(D, K, T)
2	(N-33[a], N-35[a], G-91)	18.5	(ASP, ARG, ASN)	(D, R, N)
3	(S-40[a], N-41, Q-42[a])	11.6	(LEU, PRO, THR)	(L, P, T)
4	(L-95[a], H-96, A-99)	10.0	(ARG, ASP, THR)	(R, D, T)
5	(K-100, I-93, A-99)	7.9	(LEU, ASP, VAL)	(L, D, V)
6	(I-101, I-93, A-99)	7.5	(GLU, VAL, LEU)	(E, V, L)
7	(M-31[a], S-50[a], L-9 [a])	5.8	(ILE, PRO, GLN)	(I, P, Q)
8	(M-31[a], S-50[a], N-51)	1.3	(HIS, THR, THR)	(H, T, T)

Table 6　Comparison of the details between the third triplets combination pairs of the residue groups and the mutation ones

Category	The name of the triplets combination pairs of the residue groups	PF value	d	$d1$	$d2$	θ
The mean space information for all triplets combination pairs of the residue groups in the dataset	–	13.9	5.457	2.202	2.286	123.0
The third of the triplets combination pairs of the residue groups	(1,3,5−0,0,0)	1.1	4.836	2.948	0.731	92.5
The mutated triplets combination pairs of the residue groups for advice	(0,1,3−0,0,0)	11.8	3.261	1.593	0.886	155.0

变为其他氨基酸, 与 PD-1 表面三元残基组合组成三元残基组合对, 这些三元残基组合对的 PF 值都小于 9.7, 这也说明了保留 Y-112 有利于维持该蛋白的结合亲和力。

对于表 5 中的第 3 号组合对 (M-31, S-50, N-51)−(G-107, A-108, A-109) 属于 (1,3,5−0,0,0) 类三元残基组合对, 其 PF 值仅为 1.1, 偏好性不明显。这个三元残基组合对含有 5 个根据晶体结构预测形成的结合位点的残基 (表 5 中 b 标注的残基), 而 PD-1 上的 N-51 属于突变后对蛋白结合没有影响的最佳突变残基。将属于第 5 类氨基酸的 N-51 突变成其他类氨基酸, 其中三元残基组合对 (0,1,3−0,0,0) 的 PF 值为 11.8, 较未改变的组合对 (1,3,5−0,0,0) 的 PF 值 1.1 有比较明显的提升。可能的原因是 N-51 为带负电的极性氨基酸, 其互补的三元残基组合中的氨基酸都是短链疏水性氨基酸, 这样更有利于增强其蛋白结合亲和力。因此, 预测将 PD-1 上的 N-51 突变为 0 类氨基酸 (GLY、ALA), 这样有利于增强 PD-1/PD-L2 蛋白复合物的结合亲和力。以数据集中所有组合对的平均空间信息为参考 (表 6), 对比第 3 号组合对 (1,3,5−0,0,0) 与突变建议组合对 (0,1,3−0,0,0) 可以看出, 突变建议组合对的 d 值更小, $d1$ 和 $d2$ 差值更小, θ 值也更接近总体平均值。这样的形状更接近于正规的“三棱柱”模型, 也可能更有利于两个三元残基组合正面接触。综上所述, 无论是从氨基酸性质的角度还是从模型形状的角度都支持这样的猜测。进一步对比将 PD-1 上的 N-51 突变为 GLY 或 ALA, 发现将 N-51 突变为 GLY 将创造一个更有利于残基间接触的疏水环境, 从而增强 PD-1/PD-L2 之间的结合亲和力。

小结

针对蛋白质−蛋白质相互作用界面是如何进行特异性识别这一问题, 人们经历了残基对−关键残基−蛋白质−蛋白质相互作用模型等一系列的探索。而蛋白质−蛋白质相互作用模型既能提供多种参考信息, 也能结合关键残基来研究特定残基, 因此受到研究者的广泛关注。

作者提出了一种全新的模型, 即“三棱柱”模型, 来探究蛋白质−蛋白质之间的相互作用。根据设定的过滤条件, 从蛋白数据库中挑选出 6 122 个蛋白晶体结构, 并组成 17 541 个三元残基组合对。采用 PF 值来表征三元残基组合之间的选择偏好性, 其中以 PF 值 13.6 (PF 值分布中的一个较高峰值) 为阈值, 小于该阈值的组合对出现的频次呈现一个先上升后下降的近似正态分布, 而大于该阈值的组合对出现的频次呈现一个长长的“尾部”。对于极少的 PF 值大于 200.0 的组合对, 它们大多会形成特殊的结构, 这是造成 PF 值异常高的原因。除此之外, 通过统计三元残基组合对中氨基酸的类别和组成, 可以探究三元残基组合对是如何通过各种氨基酸来介导蛋白的相互作用。最后, 通过统计三元组合对在二级结构上的分布, 发现螺旋结构明显多于其他二级结构, 这也进一步验证了螺旋结构在蛋白相互作用设计中的重要性。为了验证“三棱柱”模型的可靠性, 将统计得到的数据和实验方法应用于 PD-1/PD-L2 蛋白相互作用研究。实验结果表明, 将蛋白界面上的 51 号天冬酰胺突变为甘氨酸可以增强 PD-1/PD-L2 之间的结合亲和力。这也进一步说明了本文基于统计数据提出的“三棱柱”模型不仅能够针对性地给出蛋白相互作用界面的残基突变建议, 还能给出氨基酸的相对空间信息, 这对设计蛋白相互作用具有一定的参考价值。后续研究中, 还将不断改进该模型的预测准确性, 以适用于更多的应用实例。

References

[1] Mészáros B, Tompa P, Simon I, et al. Molecular principles of the interactions of disordered proteins [J]. J Mol Biol, 2007, 372: 549−561.

[2] Franzosa EA, Xia Y. Structural principles within the human-virus protein-protein interaction network [J]. Proc Natl Acad Sci U S A, 2011, 108: 10538−10543.

[3] Headd JJ, Ban YA, Brown P, et al. Protein-protein interfaces: properties, preferences, and projections [J]. J Proteome Res,

2007, 6: 2576−2586.

[4] Yan C, Wu F, Jernigan RL, et al. Characterization of protein-protein interfaces [J]. Protein J, 2008, 27: 59 −70.

[5] Ofran Y, Rost B. Analysing six types of protein-protein interfaces [J]. J Mol Biol, 2003, 325: 377 −387.

[6] Kenneth MorrowJ, Zhang S. Computational prediction of protein hot spot residues [J]. Curr Pharm Design, 2012, 18: 1255−1265.

[7] Grosdidier S, Fernandez-Recio J. Protein-protein docking and hot-spot prediction for drug discovery [J]. Curr Pharm Design, 2012, 18: 4607−4618.

[8] Tuncbag N, Gursoy A, Keskin O. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy [J]. Bioinformatics, 2009, 25: 1513 −1520.

[9] Fischer TB, Arunachalam KV, Bailey D, et al. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces [J]. Bioinformatics, 2003, 19: 1453 −1454.

[10] Kozakov D, Hall DR, Chuang GY, et al. Structural conservation of druggable hot spots in protein-protein interfaces [J]. Proc Natl Acad Sci U S A, 2011, 108: 13528 −13533.

[11] Brenke R, Kozakov D, Chuang GY, et al. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques [J]. Bioinformatics, 2009, 25: 621−627.

[12] Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server [J]. Nucleic Acids Res, 2004, 32: 526−531.

[13] Tuncbag N, Keskin O, Gursoy A. HotPoint: hot spot prediction server for protein interfaces [J]. Nucleic Acids Res, 2010, 38: 402−406.

[14] Darnell SJ, LeGault L, Mitchell JC. KFC server: interactive forecasting of protein interaction hot spots [J]. Nucleic Acids Res, 2008, 36: 265−269.

[15] Guharoy M, Pal A, Dasgupta M, et al. PRICE (PRotein Interface Conservation and Energetics): a server for the analysis of protein-protein interfaces [J]. J Struct Funct Genomics, 2011, 12: 33−41.

[16] Segura Mora J, Assi SA, Fernandez-Fuentes N. Presaging critical residues in protein interfaces-web server (PCRPi-W): a web server to chart hot spots in protein interfaces [J]. PLoS One, 2010, 5: e12352.

[17] Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces [J]. J Mol Biol, 1998, 280: 1 −9.

[18] Hu J, Zhang X, Liu X, et al. Prediction of hot regions in protein-protein interaction by combining density-based incremental clustering with feature-based classification [J]. Comput Biol Med, 2015, 61: 127 −137.

[19] Keskin O, Ma B, Rogale K, et al. Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach [J]. Phys Biol, 2005, 2: 24 −35.

[20] Mintz S, Shulman-Peleg A, Wolfson HJ, et al. Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions [J]. Proteins, 2005, 61: 6−20.

[21] Keskin O, Tsai CJ, Wolfson H, et al. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications [J]. Protein Sci, 2004, 13: 1043 −1055.

[22] Li Y, Liu Z, Han L, et al. Mining the characteristic interaction patterns on protein-protein binding interfaces [J]. J Chem Inf Modeling, 2013, 53: 2437−2447.

[23] Damo SM, Kehl-Fie TE, Sugitani N, et al. Molecular basis for manganese sequestration by calprotectin and roles in the innate immune response to invading bacterial pathogens [J]. Proc Natl Acad Sci U S A, 2013, 110: 3841 −3846.

[24] Gavenonis J, Sheneman BA, Siegert TR, et al. Comprehensive analysis of loops at protein-protein interfaces for macrocycle design [J]. Nat Chem Biol, 2014, 10: 716 −722.

[25] Guharoy M, Chakrabarti P. Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions [J]. Bioinformatics, 2007, 23: 1909−1918.

[26] Frishman D, Argos P. Knowledge-based protein secondary structure assignment [J]. Proteins, 1995, 23: 566 −579.

[27] Greenwald RJ, Freeman GJ, Sharpe AH. The B7 family revisited [J]. Annu Rev Immunol, 2005, 23: 515 −548.

[28] Freeman GJ, Long AJ, Iwai Y, et al. Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation [J]. J Exp Med, 2000, 192: 1027−1034.

[29] Latchman Y, Wood CR, Chernova T, et al. PD-L2 is a second ligand for PD-1 and inhibits T cell activation [J]. Nat Immunol, 2001, 2: 261−268.

[30] Okazaki T, Honjo T. The PD-1-PD-L pathway in immunological tolerance [J]. Trends Immunol, 2006, 27: 195 −201.

[31] Zang X, Allison JP. The B7 family and cancer therapy: costimulation and coinhibition [J]. Clin Cancer Res, 2007, 13: 5271−5279.

[32] Sharpe AH, Wherry EJ, Ahmed R, et al. The function of programmed cell death 1 and its ligands in regulating autoimmunity and infection [J]. Nat Immunol, 2007, 8: 239 −245.

[33] Lazar-Molnar E,Yan Q, Cao E, et al. Crystal structure of the complex between programmed death-1 (PD-1) and its ligand PD-L2 [J]. Proc Natl Acad Sci U S A, 2008, 105: 10483 −10488.