



蛋白质相互作用的网络分析

常珊, 龚新奇, 焦雄, 李春华, 陈慰祖, 王存新*

北京工业大学生命科学与生物工程学院, 北京 100124

* 联系人, E-mail: cxwang@bjut.edu.cn

2009-03-18 收稿, 2009-06-22 接受

国家自然科学基金(批准号: 20773006, 30670497)和高等学校博士学科点专项科研基金(编号: 200800050003)资助项目

摘要 建立了蛋白质复合物的残基网络, 其中蛋白质的残基被视为节点, 残基之间的接触视为连接. 复合物残基网络可以分成两种类型, 即疏水和亲水残基网络. 分析网络参量发现, 正确结合的复合物比错误结合的结构具有更高的界面度加和值和更低的网络特征路径长度. 这些性质反映出正确结合的复合物结构具有更好的几何或残基类型互补, 同时正确的结合模式对于保证天然蛋白质复合物的特征路径长度具有重要作用. 此外, 两个基于网络的打分项被建立, 它们能够很好地考虑复合物整体形状和残基类型互补特性. 将基于网络的打分项与其他打分项进行组合, 提出了一个新的多项打分函数 HPNCscore, 它能够改进 RosettaDock 组合打分函数的区分能力超过 12%. 这些研究将加深我们对蛋白质-蛋白质结合机制的了解.

关键词

界面

特征路径长度

残基网络

打分函数

蛋白质-蛋白质相互作用和识别在分子生物学中发挥着重要作用. 在蛋白质-蛋白质相互作用机理和对接预测研究中, 蛋白质拓扑结构被认为是很重要的因素. 蛋白质结合界面通常由疏水残基内核周围环绕的亲水残基组成^[1,2]; 蛋白质表面的空穴部位与小分子或蛋白质配体的结合部位相关^[3~5]; 蛋白质的整体形状特征与配体结合模式也具有一定关联^[6,7]. 分子对接技术是预测蛋白质-蛋白质复合物结构的有效方法. 在对接研究中, 许多打分方法是基于几何互补的^[8~10]. 总之, 蛋白质的拓扑结构揭示了其结合的机制^[11,12]. 因此, 分析蛋白质复合物的拓扑结构将为蛋白质-蛋白质相互作用的研究提供启示.

最近, 小世界网络方法已经成为研究蛋白质结构-功能关系的一个有力工具. 许多理论研究发现, 蛋白质残基网络具有小世界性^[13~17], 即具有大的聚集系数和小的特征路径长度. 一些应用方面的研究集中于单体蛋白质结构, 如预测蛋白质折叠中的关键残基^[13,18]或定义蛋白质结构中的功能残基^[19], 还

发现残基的波动性与平均最短路径相关^[20]以及其他一些蛋白质折叠方面的机制^[21~23]. 此外, 一些小组也研究了蛋白质-蛋白质复合物网络, 定义蛋白质-蛋白质相互作用的关键残基或热点残基^[14,24~26], 揭示配体结合过程中网络的桥连效应^[27], 以及分析蛋白质-蛋白质结合过程中的构象变化机理^[28].

残基网络的研究已经对蛋白质结构进行了系统和深入的解释. 然而, 先前的复合物残基网络研究集中于已经正确结合的蛋白质复合物构象, 即晶体结构. 实际上, 在对接过程中, 许多复合物结构将产生, 只有很少的对接构象是正确的构象. 这些正确构象和那些不正确的构象有着完全不同的结构, 对接程序需要将它们从错误结构中区分出来. 由于不同的复合物结构将导致不同的残基网络拓扑, 不正确的复合物结构应该有区别于正确结构不同的网络特征. 因此, 蛋白质的网络表示能够给蛋白质-蛋白质对接一些有用的信息.

研究发现疏水-疏水和极性-极性相互作用对蛋

引用格式: 常珊, 龚新奇, 焦雄, 等. 蛋白质相互作用的网络分析. 科学通报, 2009, 54: 3795~3803

Chang S, Gong X Q, Jiao X, et al. Network analysis of protein-protein interaction. Chinese Sci Bull, doi: 10.1007/s11434-009-0742-x

白质-蛋白质对接具有重要的作用, 而疏水-极性相互作用却不明显^[29]. 因此, 在本工作中, 分别研究了蛋白质-蛋白质复合物的疏水氨基酸网络和亲水氨基酸网络, 并从网络的观点分析了蛋白质-蛋白质相互作用的机制. 此外, 两个基于网络参量的打分项被提出. 这些打分项不依赖于能量函数, 仅仅需要测定网络参数. 通过这些网络参量, 复合物的整体形状特征和残基类型互补能够被考虑到打分项中. 因此, 可以将这些基于网络的打分项应用于蛋白质-蛋白质对接的研究中.

1 体系和方法

1.1 数据集和构象

数据集包含来自 Benchmark 2.0^[30]的 42 个二体复合物结构. 为了考察两个单体网络的重连效应, 仅有单链的单体蛋白质结构被选择. 因此, 数据集包含 18 个酶/抑制剂类型和 24 个 others 类型. 抗原/抗体类型由于具有抗体互补决定簇, 被排除出我们的数据集. 这些复合物的氨基酸数量从 126 到 915. 为了评价打分函数的正确性, 对每一个体系, 采用 RosettaDock1.0^[31]程序分别产生 1000 个 bound 结构和 unbound 结构.

1.2 定义界面残基

溶剂可及表面面积(ASA)被用来定义界面残基. 计算采用 NACCESS 程序^[32]. 表面残基(残基位于蛋白质表面)被定义为相对溶剂可接近面积至少达到其最大表面积(残基所有能被溶剂接触到的表面积)25%的残基. 界面残基被定义为复合物结构中的溶剂可接近表面积比单体中的减少了 $1 \text{ \AA}^2$ 以上的表面残基^[33].

1.3 蛋白质复合物网络和网络参量

20 种氨基酸被分成两类^[34], 疏水残基包括 Ile, Leu, Val, Phe, Met, Trp, Cys, Tyr, Pro 和 Ala, 亲水残基包括 Gly, Lys, Thr, Ser, Gln, Asn, Glu, Asp, Arg 和 His. 当蛋白质结合时, 残基类型互补(疏水-疏水和极性-极性相互作用)在蛋白质相互作用中起重要作用. 蛋白质复合物结构被分成两个无向网, 这样能够更好地描述残基类型的互补性. 一个是疏水氨基酸网络, 疏水残基作为节点, 残基间的原子接触是边, 另一个是亲水氨基酸网络, 亲水残基作为节点, 残基间的原子接触是边. 原子接触被定义为两个不同残基之间任何两个原子的距离小于 $5 \text{ \AA}$ ^[35]. 因此, 邻接矩

阵元素 A_{ij} 被写成

$$A_{ij} = \begin{cases} 1, & \text{残基 } i \text{ 与残基 } j \text{ 之间有原子接触,} \\ 0, & \text{否则.} \end{cases} \quad (1)$$

任意节点 i 的度被定义为节点 i 的邻居个数, 可表示为

$$K_i = \sum_{j=1}^{N_v} A_{ij}, \quad (2)$$

其中 N_v 是节点的数量.

任意节点 i 的聚集系数是对局部内聚性的一种测量. 可表示为

$$C_i = \frac{n_i}{K_i(K_i - 1)/2}, \quad (3)$$

其中 n_i 是节点 i 邻居实际形成的边数. 网络的聚集系数 C 是所有节点聚集系数 C_i 的平均值. 可表示为

$$C = \frac{1}{N_v} \sum_i C_i, \quad (4)$$

其中 N_v 是节点的数量.

特征路径长度 L 是图中所有节点对的平均最短路径. 可表示为

$$L = \frac{1}{N_p} \sum_{j>i} l_{ij}, \quad (5)$$

其中 N_p 表示图中节点对的数量, l_{ij} 是节点 i 和 j 之间的最短路径^[13].

1.4 打分项和打分函数

为了区分不同的复合物构象, 两个基于网络的打分项被定义. 疏水相互作用项 S_{hnscore} 和极性相互作用项 S_{pnscore} 分别计算如下:

$$S_{\text{hnscore}} = -\frac{\sum_i^{N_i^h} K_i^h}{L^h} \quad (6)$$

和

$$S_{\text{pnscore}} = -\frac{\sum_i^{N_i^p} K_i^p}{L^p}, \quad (7)$$

其中 K_i^h 是疏水网络界面残基 i 的度值, N_i^h 是界面残基数, L^h 是特征路径长度. 类似的, K_i^p 是亲水网络界面残基 i 的度值, N_i^p 是界面残基数, L^p 是特征路径长度.

为了与通常的组合打分函数比较, 组合网络打分项与 RosettaDock 打分函数的其他能量项^[31], 得到一个新的组合打分函数 HPNCscore(hydrophobic and

polar networks combined scoring function). 采用 logistic 回归来拟合权重, 使正确的构象与其他构象最大化分离开. $L_RMSD \leq 4.0 \text{ \AA}$ 的构象被定义为正确的构象. 训练集包括 22 个 bound 的对接结构, 氨基酸数量从 300 到 500. 其他的 20 个 bound 对接结构和所有的 42 个 unbound 对接结构用来进行测试. 回归结果列于表 1. 打分函数每一项的显著性通过它的权重体现. 此外, 每一项权重的 z -value 反映出该项在打分函数中的贡献. 首先, RosettaDock 能量项和网络项同时进行回归. 回归结果显示, 短程静电吸引和短程静电排斥的 z -value 不高. 同时, 增加的两个基于网络的打分函数项能够直观地考虑整个几何和残基类型的匹配情况. 因此, 去除这两项重新拟合权重, 打分函数 HPNCscore 可以写成这些项的线性组合:

$$S_{\text{hpncscore}} = w_{\text{rep}} S_{\text{rep}} + w_{\text{atr}} S_{\text{atr}} + w_{\text{sol}} S_{\text{sol}} + w_{\text{hbbsc}} S_{\text{hbbsc}} + w_{\text{hbbs}} S_{\text{hbbs}} + w_{\text{dun}} S_{\text{dun}} + w_{\text{pair}} S_{\text{pair}} + w_{\text{sasa}} S_{\text{sasa}} + w_{\text{elec}}^{\text{lr-rep}} S_{\text{elec}}^{\text{lr-rep}} + w_{\text{elec}}^{\text{lr-atr}} S_{\text{elec}}^{\text{lr-atr}} + w_{\text{hnscore}} S_{\text{hnscore}} + w_{\text{pnscore}} S_{\text{pnscore}}, \quad (8)$$

其中 S_{rep} 是范德华排斥项, S_{atr} 是范德华吸引项, S_{sol} 是溶剂化项, S_{hbbsc} 是侧链氢键项, S_{hbbs} 是主链氢键项, S_{dun} 是转子概率项, S_{pair} 是残基成对能量项, S_{sasa} 是基于表面面积的溶剂化项, $S_{\text{elec}}^{\text{lr-rep}}$ 是长程静电排斥项,

$S_{\text{elec}}^{\text{lr-atr}}$ 是长程静电吸引项, S_{hnscore} 是疏水网络项和 S_{pnscore} 是极性网络项. HPNCscore 与 RosettaDock 有相同的项数, 可以进行比较.

1.5 打分函数评价标准

采用一些评价指标来评估打分函数的区分能力, 如前 10 个构象中最好结构的 L_RMSD , 第一个 hit 结构的排序位置和前 10 位结构中的 hit 数量. L_RMSD 是受体叠落在晶体结构上得到的配体主链原子(N, C, CA, O)的均方根偏差. 在 CAPRI 实验中, 高预测结果的范围是 $L_RMSD \leq 1.0 \text{ \AA}$, 中等预测结果是 $1.0 \text{ \AA} < L_RMSD \leq 5.0 \text{ \AA}$, 可接受预测结果 $5.0 \text{ \AA} < L_RMSD \leq 10.0 \text{ \AA}$ ^[36]. 通常, 考虑不低于中等预测结果的构象为 hit 结构^[31,37]. Unbound 对接在结合过程中主链构象将发生大范围的变化, 因此它们的 L_RMSD 通常比 bound 对接要大. Bound 结构比 unbound 结构更容易对接, 能够获得更多的近天然结构. 因此, hit 结构(或近天然结构)被定义为 bound 结构 $L_RMSD \leq 4.0 \text{ \AA}$ 和 unbound 结构 $L_RMSD \leq 5.0 \text{ \AA}$.

2 结果和讨论

2.1 残基网络性质分析

首先, 我们分析两种复合物残基网络的网络性质. 42 个生物多样性的蛋白质复合物结构的网络

表 1 打分函数 HPNCscore 的权重

打分项	回归 1		回归 2	
	权重	z -value	权重	z -value
范德华排斥项	0.016	11.604	0.016	11.489
范德华吸引项	0.126	17.577	0.126	17.544
溶剂化项	0.061	5.820	0.063	6.031
侧链氢键项	0.187	18.629	0.187	18.688
主链氢键项	0.386	17.462	0.388	17.565
转子概率项	0.088	13.871	0.088	13.906
残基成对能量项	0.218	8.073	0.214	8.377
表面面积溶剂化项	0.168	15.011	0.170	15.138
简单静电项				
短程排斥项	-0.005	-1.126	-a)	-
短程吸引项	0.026	1.633	-	-
长程排斥项	0.037	3.740	0.036	3.689
长程吸引项	0.112	7.342	0.118	7.921
网络打分项				
疏水网络项	0.097	10.563	0.097	10.579
极性网络项	0.045	4.899	0.044	4.795

a) “-” 表示该项没有被包含进回归中

参数被分析, 结果与前人的研究一致^[15](图 1). 蛋白质复合物的疏水和亲水网络均具有大的聚集系数和小的特征路径长度. 当与相同节点和平均连接数量的随机网络和规则网络进行比较时, 它们展示出小世界的网络特征.

接着, 比较正确对接和不正确对接结构的复合物网络属性. 在对接过程中, 许多构象将产生, 其中正确对接的复合物构象数量 and 不正确结构相比是很少的. 为了进行公平的比较, 抽取正确对接的复合物构象(1740 个结构)和相同数量随机选择的不正确结构. 图 2 展示了界面残基度的加和值和特征路径长度的分布情况. 度是一个局部网络参量, 它反映残基的局部环境或结构. 因为单体网络的重连, 界面残基的度值增加, 内部残基和表面残基度值变化并不明显. 那么, 我们仅需要考虑界面残基度值的加和. 而特征路径长度是一个全局拓扑参量, 它能够反映网络结构的连通情况. 因为不同的复合物有不同的残基数, 网络参量的数值会受蛋白质大小的影响. 为了排除蛋白质大小的影响, 界面残基度的加和值和网络特征路径长度被标准化, 即计算每个蛋白质结构偏离平均值的标准偏差. 如图 2 所示, 不正确的对接结构($L_RMSD > 5.0 \text{ \AA}$)有低的界面残基度加和值, 高的特征路径长度. 相反, 正确的对接结构($L_RMSD \leq 4.0 \text{ \AA}$)则表现出高的界面残基度加和值, 低的特征路径长度. 界面残基的度加和值高表明界面的几何互补性好, 特征路径长度低表示网络的连通性好. 正确对接结构和错误对接结构网络特征的不同是与它们特有的结合模式有关的.

图 3 展示了蛋白质-蛋白质结合模式的示意图. 蛋

白质结合模式更倾向于选择大的结合界面^[7]. 如图 3(a), 不正确的对接结构通常具有小的结合界面. 那么, 两个单体蛋白质间的接触数少, 则界面度的加和值低. 相反, 如图 3(b)所示, 正确对接结构在两个单体蛋白质间有更多的接触形成. 从能量的观点看, 这样的结合模式更加稳定. 同时, 从网络的观点看, 蛋白质结合可以看成单体网络的重新连接. 模式(b)相比模式(a)将在两个单体网络间建立更多的连通路程, 这样更适合两个单体网络间信息的传递. 同时, 由于增加了连通路程, 网络的联通性被改进, 特征路径长度将降低. 因此, 打分项被设计成能够区分正确和不正确对接结构的形式(见方法). 通过分析晶体结构, 先前的研究已经发现中心保守残基或热点残基在保持蛋白质复合物的特征路径长度中起着关键作用^[14,24,25]. 然而, 通过比较正确对接结构和不正确的对接结构, 我们的研究发现正确的结合模式比错误的结合模式能够具有低的特征路径长度和更好的网络联通性. 因此, 正确的结合模式是保证天然蛋白质复合物特征路径长度的重要前提.

2.2 两个打分函数比较

对打分研究而言^[38,39], 一个好的打分函数应该是多项组成的, 不同的打分项能够为打分函数提供不同的贡献. 打分函数加入疏水和亲水网络打分项, 区分效果可能会有所改进. HPNCscore 是一个多项打分函数, 它组合了基于网络的打分项和其他能量项. 这里, 我们比较多项打分函数 RosettaDock 和 HPNCscore 的区分能力. 比较结果见表 2, 更细节的信息被列在表 3 和 4 中.

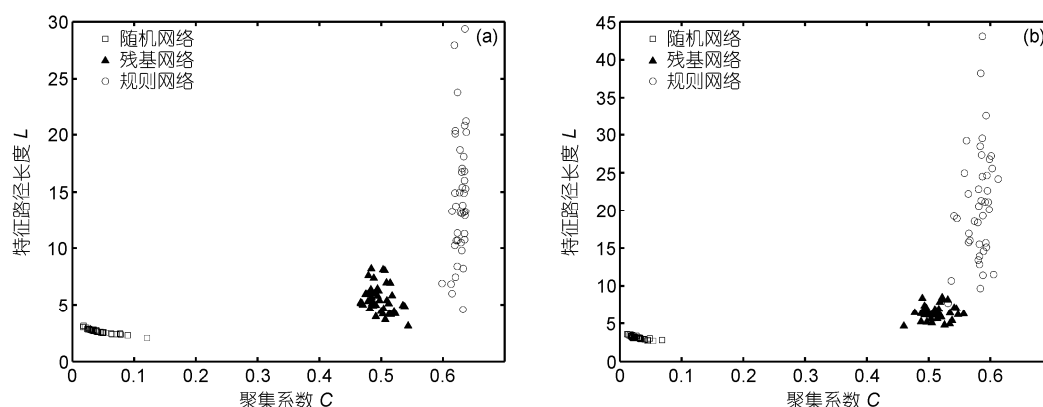


图 1 42 个蛋白质复合物的聚集系数和特征路径分布

(a) 疏水网络, (b) 亲水网络. 实心三角形表示 42 个蛋白质的氨基酸网络, 空心方形和空心圆分别表示具有相同节点和平均连接的随机网络和规则网络

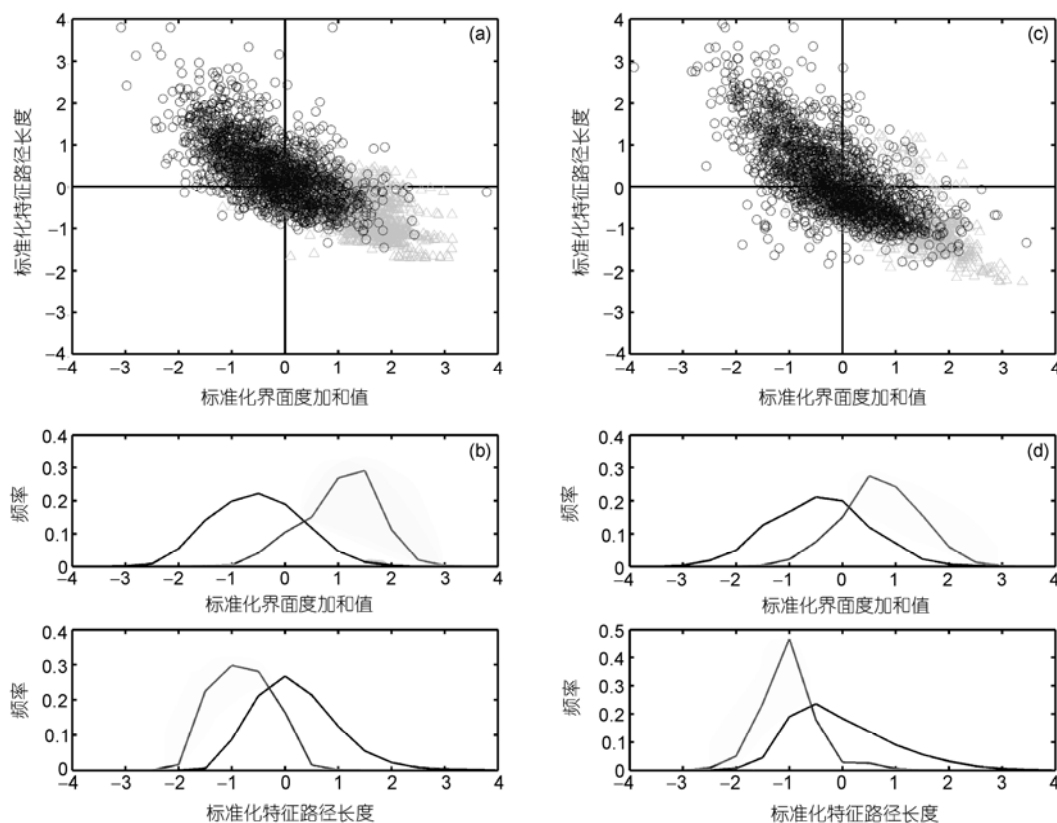


图2 42个蛋白质复合物1740个对接结构的界面残基度加和值和特征路径长度的分布情况

(a)和(b)疏水网络, (c)和(d)亲水网络. (a)和(c)中, 不正确的结构($L_RMSD > 5.0 \text{ \AA}$)为深色圆圈, 正确的结构($L_RMSD \leq 4.0 \text{ \AA}$)为浅色三角. (b)和(d)中, 不正确的结构($L_RMSD > 5.0 \text{ \AA}$)为深色线条, 正确的结构($L_RMSD \leq 4.0 \text{ \AA}$)为浅色线条

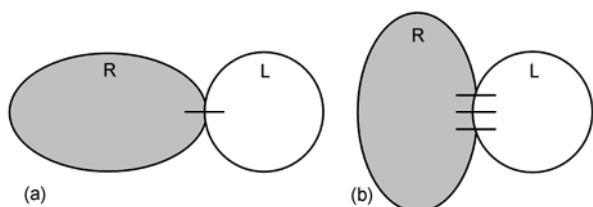


图3 蛋白质-蛋白质相互作用模式的示意图

(a) 不倾向的结合模式, 低的界面残基度加和值和高的特征路径;
(b) 更倾向的结合模式, 高的界面残基度加和值和低的特征路径.
R表示受体(receptor), L表示配体(ligand)

Bound 对接的打分结果见表3. 对酶/抑制剂类型, 18个体系中有16个, 采用 HPNCscore 排序的前10个构象最好的 L_RMSD 不大于 RosettaDock 打分函数的排序结果. 有16个体系, HPNCscore 能够给出更好的或相同的 hit 结构的排序位置. 用 RosettaDock 打分函数, 只有16个体系能够将 hit 结构排入前10的构象, 12个体系能够把 hit 结构排在第一的位置. 而采用 HPNCscore, 有16个体系能够将 hit 结构排入前10

的构象, 13个体系能够把 hit 结构排在第一的位置. 对于其他的两个体系(1EWY 和 2PCC), 两个打分函数均难以在前10位的构象中找到 hit 结构. 然而, HPNCscore 能够改进第一个 hit 结构的排序位置, 从48提升到22(1EWY), 以及从75提升到56(2PCC).

对 others 类型, 24个体系中有19个, 采用 HPNCscore 排序的前10个构象最好的 L_RMSD 不大于 RosettaDock 打分函数的排序结果. 所有的24个体系, HPNCscore 都能够给出更好的或相同的 hit 结构的排序位置. 用 RosettaDock 打分函数, 只有17个体系能够将 hit 结构排入前10的构象, 13个体系能够把 hit 结构排在第一的位置. 而采用 HPNCscore, 有20个体系能够将 hit 结构排入前10的构象, 15个体系能够把 hit 结构排在第一的位置. 对于3个体系(1AK4, 1FQJ 和 1KAC), RosettaDock 打分函数不能在前10位的构象中找到 hit 结构, 而 HPNCscore 分别找到了7, 8和1个 hit 结构. 对4个体系(1BUH, 1E96, 1GHQ 和 1QA9), HPNCscore 能够改进第一个 hit 结构的排序

表 2 打分函数的区分能力比较

	Bound		Unbound	
	RosettaDock	HPNCscore	RosettaDock	HPNCscore
酶/抑制剂类型				
Hit ^{a)}	16	16	17	20
Hit ranked 1st ^{b)}	12	13	13	15
others 类型				
Hit	12	14	7	9
Hit ranked 1st	8	10	3	4
Total hits ^{c)}	116/135/251 ^{d)}	148/152/300	105	125

a) 第 1 个 hit 结构(Bound 对接 L_RMSD $\leq$ 4.0 Å 和 Unbound 对接 L_RMSD $\leq$ 5.0 Å)被排在前 10 位的体系数量; b) Hit 结构被排在第一的体系数量; c) 打分函数找到的总的 hit 结构数量; d) 第 1 个数值表示 22 个训练蛋白体系的 hit 数量, 第 2 个数值表示 20 个测试蛋白体系的 hit 数量, 第 3 个数值表示 42 个蛋白体系总的 hit 数量

表 3 Bound 对接 RosettaDock 和 HPNCscore 区分能力比较

PDB 编号	RosettaDock			HPNCscore			PDB 编号	RosettaDock			HPNCscore		
	Best L_RMSD ^{a)}	Rank of 1st hit ^{b)}	Num of hits ^{c)}	Best L_RMSD	Rank of 1st hit	Num of hits		Best L_RMSD	Rank of 1st hit	Num of hits	Best L_RMSD	Rank of 1st hit	Num of hits
酶/抑制剂类型(18)							1F34	0.26	1	10	0.26	1	10
1ACB	0.77	1	10	0.59	1	10	1MAH*	0.66	1	10	0.63	1	10
1AVX	1.51	2	6	1.26	1	10	1PPE*	0.4	1	10	0.4	1	10
1AY7*	2.65	8	1	2.65	2	3	1TMQ*	0.38	1	10	0.38	1	10
1BVN*	0.44	1	10	0.44	1	10	1UDI	0.34	1	10	0.34	1	10
1CGI	1.11	1	10	1.03	1	10	2PCC	9.02	75	0	8.96	56	0
1D6R*	1.16	9	1	0.74	1	10	2SIC	0.58	1	10	0.58	1	10
1DFJ*	2.25	1	6	2.25	2	5	2SNI	0.2	1	10	0.2	1	10
1E6E*	0.87	1	10	1.43	2	3	7CEI*	1.59	8	2	0.75	1	7
1EWY	4.57	48	0	5.06	22	0							
Others 类型(24)							1HE1	0.86	1	9	0.29	1	10
1AK4	16.21	39	0	0.39	1	7	1HE8*	3.77	4	1	2.76	4	2
1ATN*	1.01	1	10	0.7	1	10	1I2M*	0.65	1	10	1.12	1	9
1B6C	0.96	1	10	1.09	1	10	1IBR*	1.34	1	9	1.34	1	9
1BUH	13.22	69	0	12.9	22	0	1KAC	4.29	11	0	1.98	6	1
1E96	13.51	242	0	4.41	51	0	1KTZ*	0.73	1	9	0.88	1	10
1FC2*	2.11	1	8	1.56	1	9	1KXP*	1.13	1	7	1.13	1	9
1FQ1	2.93	4	2	2.55	2	8	1M10	0.79	1	8	0.79	1	10
1FQJ	6.96	44	0	0.66	1	8	1QA9*	6.77	541	0	9.71	13	0
1GCQ*	0.62	7	2	0.59	2	8	1SBB	3.68	4	1	1.46	3	5
1GHQ	8.26	424	0	7.35	51	0	1WQ1	0.6	1	10	0.62	1	9
1GRN	0.4	1	10	0.38	1	10	2BTF*	0.35	1	9	0.35	1	8
1H1V*	0.81	1	10	0.81	1	10							

a) 前 10 个构象中最好结构的 L_RMSD; b) 第 1 个 hit 结构的排序位置; c) 前 10 位结构中的 hit 数量; “*” 表示测试蛋白质

位置, 从 69, 242, 424, 541 分别提升到 22, 51, 51, 13. 采用 RosettaDock 打分函数在前 10 个构象中找到的所有 hit 结构数量是 251, 而用 HPNCscore, 数量提升到 300.

回归是基于 bound 结构进行的训练, 因此, un-

bound 结构可以作为测试集. Unbound 对接结构的打分比较见表 4. 对酶/抑制剂类型, 18 个体系中有 10 个, 采用 HPNCscore 排序的前 10 个构象最好的 L_RMSD 不大于 RosettaDock 打分函数的排序结果. 有 17 个体系, HPNCscore 能够给出更好的或相同的 hit 结构的排序

表 4 Unbound 对接 RosettaDock 和 HPNCscore 区分能力比较^{a)}

PDB 编号	RosettaDock			HPNCscore			PDB 编号	RosettaDock			HPNCscore		
	Best L_RMSD	Rank of 1st hit	Num of hits	Best L_RMSD	Rank of 1st hit	Num of hits		Best L_RMSD	Rank of 1st hit	Num of hits	Best L_RMSD	Rank of 1st hit	Num of hits
酶/抑制剂类型(18)							1F34	5.94	21	0	3.63	7	2
1ACB	12.11	174	0	5.64	112	0	1MAH	2.65	4	5	2.43	2	5
1AVX	6.07	46	0	7.05	44	0	1PPE	1.45	1	10	1.9	1	10
1AY7	4.75	5	1	2.38	4	3	1TMQ	2.69	1	8	2.65	1	7
1BVN	3.03	1	6	4.05	1	2	1UDI	1.3	1	10	1.46	1	9
1CGI	2.77	1	10	2.77	1	10	2PCC	9.41	62	0	9.25	145	0
1D6R	5.74	30	0	4.76	8	2	2SIC	3.15	1	6	2.88	1	7
1DFJ	2.38	1	3	2.38	1	4	2SNI	3.8	3	2	1.98	1	7
1E6E	1.32	1	8	1.32	1	8	7CEI	1.37	3	7	1.37	1	9
1EWY	6.47	53	0	6.35	15	0							
Others 类型(24)							1HE1	2.78	1	9	2.76	3	7
1AK4	14.45	724	0	14.39	460	0	1HE8	7.85	297	0	6.96	24	0
1ATN	8.69	187	0	11.08	337	0	1I2M	8.6	116	0	9.92	157	0
1B6C	8.89	33	0	10.8	83	0	1IBR	6.98	—	0	13.15	—	0
1BUH	9.62	33	0	8.78	24	0	1KAC	4.13	4	5	2.76	3	5
1E96	14.91	217	0	8.67	211	0	1KTZ	1.82	3	3	1.82	1	5
1FC2	14.86	80	0	5.33	37	0	1KXP	1.41	1	3	1.41	1	7
1FQ1	6.71	97	0	3.87	9	1	1M10	13.36	750	0	10.99	527	0
1FQJ	6.41	376	0	5.28	135	0	1QA9	8.41	183	0	7.24	47	0
1GCQ	16.21	43	0	1.08	4	4	1SBB	3.17	6	1	1.68	1	7
1GHQ	8.98	—	0	8.98	—	0	1WQ1	3.66	1	6	3.24	3	3
1GRN	11.14	173	0	10.68	26	0	2BTF	4.13	3	2	4.89	1	1
1H1V	19.66	—	0	19.97	—	0							

a) “—” 表示在采样过程中, 该体系没有采集到 L_RMSD < 5.0 Å 的对接结构

位置. 用 RosettaDock 打分函数, 只有 12 个体系能够将 hit 结构排入前 10 的构象, 8 个体系能够把 hit 结构排在第一的位置. 而采用 HPNCscore, 有 14 个体系能够将 hit 结构排入前 10 的构象, 10 个体系能够把 hit 结构排在第一的位置. 对 2 个体系(1D6R 和 1F34), RosettaDock 打分函数不能在前 10 位的构象中找到 hit 结构, 而 HPNCscore 分别找到了 2 个 hit 结构. 对 3 个体系(1ACB, 1AVX 和 1EWY), HPNCscore 能够改进第一个 hit 结构的排序位置, 从 174, 46, 53 分别提升到 112, 44, 15.

对 others 体系, 1GHQ, 1H1V 和 1IBR, 没有 L_RMSD < 5.0 Å 的对接结构在采样过程中被采集到. 其中, 2 个体系(1H1V 和 1IBR)在 Benchmark 2.0^[30]中是困难对接类型. 这些困难对接类型的复合物结构有较大的构象变化, 因此很难采集到近天然结构. 那么, 在 others 类型中, 只有 21 个蛋白质体系有近天然的对接结构. 而 21 个体系中有 14 个, 采用 HPNCscore 排序的前 10 个构象最好的 L_RMSD 不大于

RosettaDock 打分函数的排序结果. 有 16 个体系, HPNCscore 能够给出更好的或相同的 hit 结构的排序位置. 用 RosettaDock 打分函数, 只有 7 个体系能够将 hit 结构排入前 10 的构象, 3 个体系能够把 hit 结构排在第一的位置. 而采用 HPNCscore, 有 9 个体系能够将 hit 结构排入前 10 的构象, 4 个体系能够把 hit 结构排在第一的位置. 对于 2 个体系(1FQ1 和 1GCQ), RosettaDock 打分函数不能在前 10 位的构象中找到 hit 结构, 而 HPNCscore 分别找到了 1 和 4 个 hit 结构. 对 9 个体系(1AK4, 1BUH, 1E96, 1FC2, 1FQJ, 1GRN, 1HE8, 1M10 和 1QA9), HPNCscore 能够改进第一个 hit 结构的排序位置. 采用 RosettaDock 打分函数在前 10 个构象中找到的所有 hit 结构数量是 105, 而用 HPNCscore, 数量提升到 125.

从 bound 和 unbound 对接结构的比较中发现, HPNCscore 显示出比 RosettaDock 打分函数相对有效的区分能力. 然而, 为验证其可靠的扩展能力, 函数需要在测试集上具有好的表现能力. 因此, 需要更多

地关注 20 个测试蛋白质和 42 个 unbound 结构. 如表 2 所示, 采用 RosettaDock 打分函数在测试蛋白质上, 总的 hit 数是 135, 而 HPNCscore 是 152, 数量增加了 12.6%. 同时, 采用 RosettaDock 打分函数在 unbound 结构上, 总的 hit 数是 105, 而 HPNCscore 是 125, 数量增加了 19.0%. 因此, HPNCscore 在区分能力上改进 RosettaDock 打分函数超过 12%.

我们选择了 4 个对接体系(pdb 编号: 1ACB, 1D6R, 1AK4, 1FQ1), 进一步分析网络打分项的改进效果. 如图 4(a)~(d)所示, RosettaDock 打分函数排序前 50 个结构的质心分布很分散. 4 个体系前 50 个结构最大的 L_RMSD 分别为 20.93, 20.20, 31.91 和 23.62 Å. 相反, 加入网络参量, HPNCscore 排序前 50 个结构的质心分布更集中, 而且位置都接近正确的质心位置(见图 4(e)~(h)). 4 个体系前 50 个结构最大的 L_RMSD 分别为 2.17, 5.98, 19.36 和 7.85 Å. HPNCscore 排序的前 50 个结构的结合模式更类似于图 3 中的模式(b). 这进一步验证了我们关于蛋白质结合的网络观点. 值得注意的是体系 1AK4, 晶体结构的界面面积是 1029 Å², 这比大多数稳定结合的蛋白质-蛋白质复合物界面面积小^[40]. 但由于残基类型互补性好^[41], 正确结构比不正确对接结构建立了更多有效的连通路程. 因而, HPNCscore 也显示出比 RosettaDock 更优的打

分表现. 我们的结果反映出蛋白质-蛋白质结合有一些明显的网络特征, 网络参量较好地描述了复合物的整体性状和残基类型互补的性质. 这一工作能够被应用于蛋白质-蛋白质相互作用和结构预测的研究.

3 结论

因为复合物界面的特征互补, 蛋白质-蛋白质相互作用可以被视为特殊的网络重连. 本工作分别研究和分析了蛋白质-蛋白质复合物的疏水和亲水残基网络. 网络方法能够考虑蛋白质-蛋白质复合物的全局拓扑特征, 从而更好地解释了蛋白质-蛋白质相互作用的结构机制. 研究发现, 正确结合的复合物构象比错误结合的具有更高的界面残基度加和值和更低的特征路径长度. 根据蛋白质-蛋白质相互作用的网络机制, 设计了两个基于网络的打分项, 它们由两个基本的网络参量组成, 即界面残基的度加和值和特征路径长度. 组合基于网络的打分项和其他能量项, 一个多项打分函数 HPNCscore 被提出. 与 RosettaDock 打分函数比较结果表明, HPNCscore 能够改进区分能力超过 12%. 本工作将增进我们对于蛋白质-蛋白质相互作用拓扑学机理的了解, 并可能在将来用于蛋白质设计.

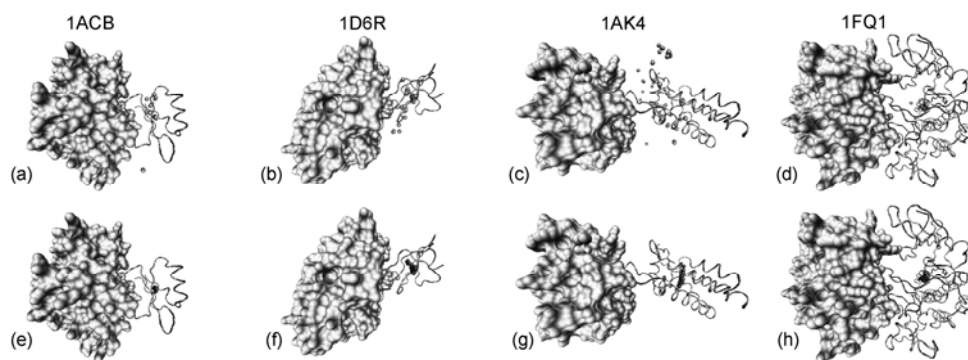


图 4 4 个对接体系(pdb 编号: 1ACB, 1D6R, 1AK4, 1FQ1), RosettaDock((a)~(d))和 HPNCscore((e)~(h)) 排序的前 50 个结构的几何中心分布

浅色球表示 RosettaDock 打分排序的前 50 个结构的几何中心, 深色球表示 HPNCscore 打分排序的结果. 晶体结构中受体被表示成实心表面, 配体结构被表示成线条形式

参考文献

- 1 Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. *Proteins*, 2002, 47: 334—343[doi]
- 2 Bahadur R P, Chakrabarti P, Rodier F, et al. Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 2003, 53: 708—719[doi]
- 3 Ho C M W, Marshall G R. Cavity search: An algorithm for the isolation and display of cavity-like binding regions. *J Comput Aided Mol*

- Des, 1990, 4: 337—354[\[doi\]](#)
- 4 Noyal M, Honig B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins*, 2006, 63: 892—906[\[doi\]](#)
- 5 Binkowski T A, Joachimiak A, Liang J. Protein surface analysis for function annotation in high-throughput structural genomics pipeline. *Protein Sci*, 2005, 14: 2972—2981[\[doi\]](#)
- 6 Foote J, Raman A. A relation between the principal axes of inertia and ligand binding. *Proc Natl Acad Sci USA*, 2000, 97: 978—983[\[doi\]](#)
- 7 Nicola G, Vakser I A. A simple shape characteristic of protein-protein recognition. *Bioinformatics*, 2007, 23: 789—792[\[doi\]](#)
- 8 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: 2195—2199[\[doi\]](#)
- 9 Walls P H, Sternberg M J E. New algorithm to model protein-protein recognition based on surface complementarity: Applications to antibody-antigen docking. *J Mol Biol*, 1992, 228: 277—297[\[doi\]](#)
- 10 Norel R, Lin S L, Wolfson H J, et al. Shape complementarity at protein-protein interfaces. *Biopolymers*, 1994, 34: 933—940[\[doi\]](#)
- 11 Levy Y, Wolynes P G, Onuchic J N. Protein topology determines binding mechanism. *Proc Natl Acad Sci USA*, 2004, 101: 511—516[\[doi\]](#)
- 12 Levy Y, Cho S S, Onuchic J N, et al. A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. *J Mol Biol*, 2005, 346: 1121—1145[\[doi\]](#)
- 13 Vendruscolo M, Dokholyan N V, Paci E, et al. Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E*, 2002, 65: 061910[\[doi\]](#)
- 14 del Sol A, Fujihashi H, O'Meara P. Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, 2005, 21: 1311—1315[\[doi\]](#)
- 15 Kundu S. Amino acid network within protein. *Physica A*, 2005, 346: 104—109[\[doi\]](#)
- 16 Aftabuddin M, Kundu S. Weighted and unweighted network of amino acids within protein. *Physica A*, 2006, 369: 895—904[\[doi\]](#)
- 17 Aftabuddin M, Kundu S. Hydrophobic, hydrophilic and charged amino acid networks within protein. *Biophys J*, 2007, 93: 225—231[\[doi\]](#)
- 18 Li J, Wang J, Wang W. Identifying folding nucleus based on residue contact networks of proteins. *Proteins*, 2008, 71: 1899—1907[\[doi\]](#)
- 19 Amitai G, Shemesh A, Sitbon E, et al. Network analysis of protein structures identifies functional residues. *J Mol Biol*, 2004, 344: 1135—1146[\[doi\]](#)
- 20 Atilgan A R, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. *Biophys J*, 2004, 86: 85—91[\[doi\]](#)
- 21 Jiao X, Chang S, Li C H, et al. Construction and application of the weighted amino acid network based on energy. *Phys Rev E*, 2007, 75: 051903[\[doi\]](#)
- 22 Dokholyan N V, Li L, Ding F, et al. Topological determinants of protein folding. *Proc Natl Acad Sci USA*, 2002, 99: 8637—8641[\[doi\]](#)
- 23 Chang S, Jiao X, Gong X Q, et al. Evolving model of amino acid networks. *Phys Rev E*, 2008, 77: 061920[\[doi\]](#)
- 24 del Sol A, Fujihashi H, Amoros D, et al. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol Syst Biol*, 2006, 2: 2006.0019
- 25 del Sol A, O'Meara P. Small-world network approach to identify key residues in protein-protein interaction. *Proteins*, 2005, 58: 672—682[\[doi\]](#)
- 26 Brinda K V, Vishveshwara S. Oligomeric protein structure networks: Insights into protein-protein interactions. *BMC Bioinformatics*, 2005, 6: 296—310[\[doi\]](#)
- 27 Hu Z J, Bowen D, Southerland W M, et al. Ligand binding and circular permutation modify residue interaction network in DHFR. *PLoS Comput Biol*, 2007, 3: 1097—1107
- 28 Daily M D, Upadhyaya T J, Gray J J. Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins*, 2008, 71: 455—466[\[doi\]](#)
- 29 Berchanski A, Shapira B, Eisenstein M. Hydrophobic complementarity in protein-protein docking. *Proteins*, 2004, 56: 130—142[\[doi\]](#)
- 30 Mintseris J, Wiehe K, Pierce B, et al. Protein-protein docking benchmark 2.0: An update. *Proteins*, 2005, 60: 214—216[\[doi\]](#)
- 31 Gray J J, Moughon S, Wang C, et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*, 2003, 331: 281—299[\[doi\]](#)
- 32 Hubbard S J, Thornton J M. NACCESS Computer Program. London: University College London, 1993
- 33 Jones S, Thornton J M. Principles of protein-protein interactions. *Proc Natl Acad Sci USA*, 1996, 93: 13—20[\[doi\]](#)
- 34 Sun S J, Brem R, Chan H S, et al. Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng*, 1995, 8: 1205—1213[\[doi\]](#)
- 35 Greene L H, Higman V A. Uncovering network systems within protein structures. *J Mol Biol*, 2003, 334: 781—791[\[doi\]](#)
- 36 Mendez R, Leplae R, De Maria L, et al. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins*, 2003, 52: 51—67[\[doi\]](#)
- 37 Camacho C J, Gatchell D W, Kimura S R, et al. Scoring docked conformations generated by rigid-body protein-protein docking. *Proteins*, 2000, 40: 525—537[\[doi\]](#)
- 38 Mohan V, Gibbs A C, Cummings M D, et al. Docking: Successes and challenges. *Curr Pharm Des*, 2005, 11: 323—333[\[doi\]](#)
- 39 Mendez R, Leplae R, Lensink M F, et al. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 2005, 60: 150—169[\[doi\]](#)
- 40 Janin J, Chothia C. The structure of protein-protein recognition sites. *J Biol Chem*, 1990, 265: 16027—16030
- 41 Gamble T R, Vajdos F F, Yoo S H, et al. Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid. *Cell*, 1996, 87: 1285—1294[\[doi\]](#)