中国科学 C 辑:生命科学 2009 年 第 39 卷 第 10 期: 963 ~ 973
www.scichina.com life.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

论 文

蛋白质复合物结构预测的集成分子对接方法

龚新奇†, 刘斌†, 常珊, 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

† 同等贡献

* 联系人, E-mail: cxwang@bjut.edu.cn

收稿日期: 2009-09-01; 接受日期: 2009-09-22

国家自然科学基金(批准号: 30670497 和 20773006)、教育部博士点基金(批准号: 200800050003)和国家重点基础研究发展计划(批准号: 2009CB930200)资助项目

摘要 蛋白质分子间相互作用与识别是当前生命科学研究的热点, 分子对接方法是研究这一问题的有效手段. 为了推进分子对接方法的发展, 欧洲生物信息学中心组织了国际蛋白质复合物结构预测(CAPRI)竞赛. 通过参加 CAPRI 竞赛, 逐步摸索出了一套用于蛋白质复合物结构预测的集成蛋白质-蛋白质分子对接方法 HoDock, 它包括结合位点预测、初始复合物结构采集、精细复合物结构采集、结构成簇和打分排序以及最终复合物结构挑选等主要步骤. 本文以最近的 CAPRI Target 39 为例, 具体说明该方法的主要步骤和应用. 该方法在 CAPRI Target 39 竞赛中取得了比较好的结果, 预测结构 Model 10 是所有参赛小组提交的 366 个结构中仅有的 3 个正确结构之一, 其配体均方根偏差(L_Rmsd)为 0.25 nm. 在对接过程中, 首先用理论预测和实验信息相结合的方法来寻找蛋白质结合位点残基, 确认 CAPRI Target 39 A 链的 A31TRP 和 A191HIS, B 链的 B512ARG 和 B531ARG 为可能结合位点残基. 同时, 用 ZDock 程序做不依赖结合位点的初步全局刚性对接. 然后, 根据结合位点信息进行初步局部刚性对接, 从全局和局部对接中挑出了 11 个初始对接复合物结构. 进而, 用改进的 RosettaDock 程序做精细位置约束对接, 并对每组对接中打分排序前 200 的结构进行成簇聚类. 最后, 综合分析打分、成簇和结合位点三方面的信息, 得到 10 个蛋白质复合物结构. 竞赛结果表明, A191HIS, B512ARG 和 B531ARG 三个结合位点残基预测正确, 提交的 10 个蛋白质复合物结构中有 5 个复合物受体-配体界面残基预测成功率较高. 与其他参赛小组的对接结果比较, 表明 HoDock 方法具有一定优势. 这些结果说明我们提出的集成分子对接方法有助于提高蛋白质复合物结构预测的准确率.

关键词
蛋白质复合物
结构预测
结合位点
分子对接
CAPRI

蛋白质-蛋白质相互作用是细胞内多种生理活动的基础, 如基因的复制、转录、翻译和细胞周期调控、信号转导、免疫反应等都是以蛋白质分子间相互作用为纽带进行的. 在蛋白质组学研究中, 随着酵母双杂交等实验技术的广泛应用, 多种模式生物的蛋白质相互作用图谱已经完成[1~4]. 蛋白质复合物结构的确定是理解蛋白质-蛋白质相互作用机理和功能的前提. 与蛋白质单体结构数据及蛋白质相互作用图谱数据相比, 目前已知结构的蛋白质复合物非常有限[5], 原因在于蛋白质复合物结构的实验解析难度很大. 因此, 如何从蛋白质单体结构出发正确预测蛋白质复合物的三维结构已成为目前国际上关注的热点问题[6,7].

引用格式: 龚新奇, 刘斌, 李春华, 等. 蛋白质复合物结构预测的集成分子对接方法. 中国科学 C 辑: 生命科学, 2009, 39: 963—973

用分子对接(molecular docking)技术, 可以由两个蛋白质单体结构出发, 来预测其三维复合物结构. 近年来, 随着计算机的处理能力不断增强和人们对蛋白质-蛋白质相互作用的了解逐步深入, 分子对接方法得到了越来越多的重视[8,9]. 为了推动分子对接技术的发展, 2001 年欧洲生物信息学研究所(European Bioinformatics Institute, EBI)开始举办蛋白质复合物结构预测竞赛(Critical Assessment of Predicted Interactions, CAPRI). 在竞赛中, 竞赛组委会选取尚未发表实验结构数据的蛋白质复合物为竞赛内容, 要求参赛者在规定时间内, 从蛋白质单体结构出发, 用分子对接方法对蛋白质复合物结构进行预测, 通过网络提交 10 个预测结果[10~12].

近年来蛋白质复合物结构预测在范围和精度上都有了较大进展. 从CAPRI竞赛委员会提供的预测题目能够看出, 可以预测的蛋白质体系越来越多样, 由抗原-抗体、酶-抑制剂等发展到基因转录和翻译、信号转导、RNA和蛋白质加工、膜转运等多种生物过程中的蛋白质复合物体系[6,11]. 由构象变化较小的结合态(bound)对接发展到构象变化较大的非结合态(unbound)对接. 有些竞赛内容其中一个单体只给出氨基酸序列, 需要直接从氨基酸序列出发, 通过同源结构预测进行该单体三维结构预测, 然后再与另一给出三维结构的单体对接生成复合物三维结构. 在预测内容越来越多样的同时, 蛋白质复合物结构预测的精度也有了较大提高, 好的预测结果与天然结构的均方根偏差Rmsd(root mean square deviation)可达到X-射线晶体学方法的分辨率范围[11~13].

在CAPRI竞赛和实验需求的推动下, 分子对接过程中涉及的几个关键步骤, 比如结合位点预测、复合物结构采集、柔性处理和打分函数等都有一些新方法出现. 结合位点预测方法通过分析已知蛋白质复合物体系界面氨基酸残基与非界面的表面氨基酸残基之间的物理化学等性质的不同[14], 利用线性回归[15]、神经网络[16]等数值拟合方法或贝叶斯网络[17]、隐马尔科夫模型[18]等方法来预测结合位点残基. 有一些公开的结合位点预测网络服务器可以免费使用, 例如SPPIDER[16], InterProSurf[19], Meta-PPISP[20]和PPI-Pred[21]等. Zhou和Qin[22]综述了最近的结合位点预测方法进展. 复合物结构采集应用到的算法主要有快速傅立叶变换(fast Fourier transform, FFT)、遗传算法(genetic algorithm, GA)和蒙特卡罗(Monte Carlo, CA)方法等. 目前的蛋白质复合物结构预测中应用比较好的有采用快速傅立叶变换的ZDock[23]和ClusPro[24], 还有采用蒙特卡罗方法的RosettaDock[25]等. 合理敏感的打分函数是分子对接中成功挑选出近天然结构的重要基础. 打分函数一般分为三类: 基于物理的能量函数、经验函数和基于知识的函数[25]. 基于物理的能量函数利用“热力学主方程”进行自由能计算. 经验函数则考虑了多种因素的贡献, 如残基成对偏好性、几何互补性及静电、氢键、疏水相互作用能等, 以适当的权重把各项组合起来. 基于知识的函数是用统计方法分析已有的蛋白质结构数据库得到的, 诸如残基-残基接触势和原子-原子接触势等[24]. 分子对接中蛋白质柔性处理目前做得比较好的有三种方法: 多构象叠落、转子库(rotamer libarary)和折叠树(fold tree)模型. 多构象叠落方法是用蛋白质单体的多个构象进行对接, 然后综合起来挑选对接结果[26]. 侧链转子库方法是利用统计得到的氨基酸残基侧链的可能构象来考虑侧链的柔性[25,27]. 折叠树模型尝试局部考虑蛋白质主链的柔性, 把蛋白质结构中原子与原子之间的连接按照树状形式表示, 在对接中可只转动有柔性的分枝上的原子[28,29].

近年来蛋白质复合物结构预测方法有了一些进展, 逐渐产生了以Zdock[23], RosettaDock[25]和ClusPro[24]为代表的对接方法. 本研究小组在过去的竞赛中也采取了一些策略来改进对接算法, 例如针对蛋白质复合物不同种类的分类打分函数[30]、多构象跌落考虑结合位点部位柔性[26]和利用蛋白质单体内氢键的包埋特性来过滤不好的对接结构[13]等, 这些方法在不同程度上提高了预测成功率. 但是, 面对体系差异多、实验信息少、构象变化大和结合自由能难以准确计算等现实困难和挑战, 一般的分子对接方法都存在着各自的局限性. 它们在分子对接的某些步骤可以做得较好, 但因为其他步骤不好而导致最后的预测结果与天然复合物结构的偏差较大[11,12]. 比如, Zdock的刚性对接效果较好, 但它缺乏结合位点预测和柔性处理; RosettaDock的柔性处理和精细打分函数较好, 但它缺乏结合位点预测, 并且计算缓慢; ClusPro的打分函数较好, 但它只是利用Zdock等刚性程序采样,

缺乏结合位点预测和柔性处理.

在最近的CAPRI竞赛中, 为了提高分子对接各个步骤的准确性, 本实验发展了一套集成对接方法, 命名为HoDock(Holistic Dock). 本文以2009年1月29日~2月8日举行的第17届CAPRI竞赛中Target 39为例, 详细分析了该方法在蛋白质复合物结构预测中的过程和结果. 在这次预测竞赛中, 总共有 37 个参赛小组, 提交了 366 个结构, 只有 3 个结构达到了CAPRI组委会制定的"好结构"的标准[6]. 本实验室提交的结构Model 10是3个好结构之一, 其配体主链原子均方根偏差(root mean square deviation of the ligand, L_Rmsd)为 0.25 nm, 被评为中等(medium)好结构(http://www.ebi.ac.uk/msd-srv/capri/round17/round17.html).

1 体系与方法

1.1 Target 39 蛋白质体系简介

Target 39 包括 A 和 B 两条链, A 链有 357 个氨基酸残基, B 链有 98 个氨基酸残基. 它的单体结构由加拿大多伦多大学 Hee-Won Park 教授提供(http://www.ebi.ac.uk/msd-srv/capri/round17/round17.html). 在CAPRI竞赛中, 竞赛委员会提供给参赛者两条单链A和B的三维结构, 要求参赛者从这两条单链结构出发来预测它们的复合物结构. 如图 1 所示, (A)为受体A链, (B)为配体B链. A链包含3个结构域, B链是一个单独的结构域, 图中还用CPK模型显示了理论预测所得到的结合位点残基的信息. 通过查阅文献发现, A链是蛋白质centaurin-alpha 1, 又称 3,4,5-三磷酸磷脂酰肌醇(phosphatidylinositol 3,4,5-trisphosphate, PIP3) 结合蛋白, 是一种在神经系统中高表达的ADP核糖基化因子(ADP-ribosylation factor, ARF)激活蛋白. Centaurin-alpha 1蛋白由3部分组成: N端ARF激活结构域(残基 1~126)、中间PH结构域(残基 130~230)、C-端PH结构域(残基253~357), 其中后两个PH结构域用于结合PIP3[31,32]. B链为KIF13B蛋白的叉状(forkhead associated, FHA)结构域, KIF13B从属于驱动蛋白超家族(kinesin superfamily, KIF), 是含有FHA结构域的最大的一类蛋白[33,34]. FHA结构域主要由β-片层构成, β-片层间由无规卷曲连接, 形似叉状, 故得名叉状结构域, 是由Hofmann和Bucher于 1995 年发现的[35]. FHA结构域的β3~β8及其之间的无规卷曲是保守区域. FHA结构域的作用是参与磷酸化依赖的蛋白质相互作用, 如调控与其相互作用蛋白的定位、转录激活等[34].

1.2 对接方法

集成分子对接方法的基本思想如下: 第一步, 预测结合位点. 综合理论预测和实验文献的信息, 给出

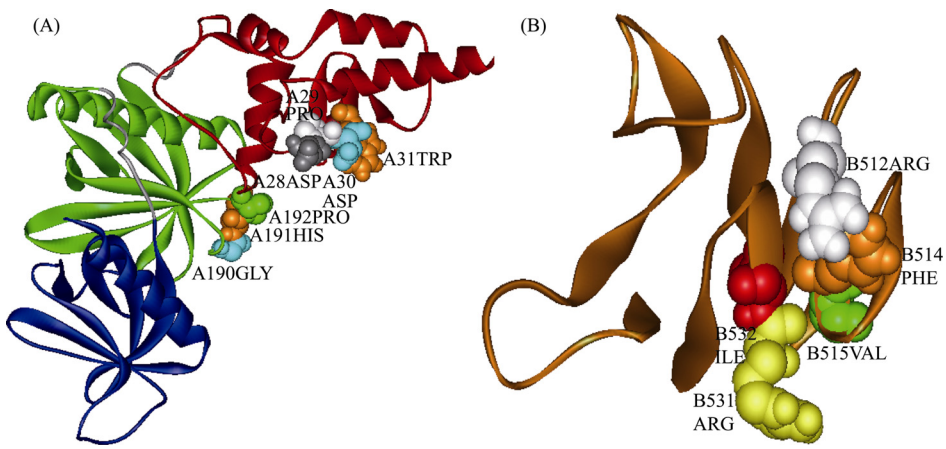

图 1 CAPRI Target 39 受体和配体的三维结构

(A) 受体包含 3 个结构域: N 端 ARF 激活结构域(红色)、中间 PH 结构域(绿色)、C 端 PH 结构域(蓝色). 理论预测的结合位点残基有两块: A190GLY, A191HIS, A192PRO 和 A28ASP, A29PRO, A30ASP, A31TRP; (B) 配体上理论预测的结合位点残基有两块: B512ARG, B514PHE, B515VAL 和 B531ARG, B532ILE. 二级结构用刚性飘带(rigid ribbon)模型表示, 其中预测的结合位点残基用 CPK 模型显示, 残基上标注有链序号、残基序号及残基名称

蛋白质单体中可能参与形成复合物界面的结合位点氨基酸残基. 本研究小组开发了一个结合位点残基预测方法, 将另文发表. 这里简要介绍它的基本思想, 蛋白质表面上同时满足以下 3 个特征的残基块可能是结合位点区域: 残基所形成的单体内部主链氢键暴露在水溶液中; 残基在蛋白质的功能慢运动模式中起铰链连接作用; 残基所在块内密集程度高. 在用理论方法预测的同时, 也从实验文献中寻找与结合位点有关的生物化学信息, 用理论预测与实验信息相结合的方法来确认结合位点. 第二步, 采集初始复合物结构. 在采集复合物结构过程中, 通过两种方式利用第一步给出的结合位点残基. 一种是用这些残基约束结构采集的范围, 缩小采样空间. 这种方式可以减少计算量, 但也可能受到不准确的结合位点信息的误导. 另一种方式是在全局采样后, 把结合位点信息与能量打分等综合起来挑出正确结构. 通过这两种方式挑出少数可能正确的复合物结构. 第三步, 采集精细复合物结构. 在上一步结果的基础上做限制搜索范围的更加精细的局部对接, 在搜索的同时考虑结合界面上侧链原子的柔性, 同时对保存下来的满足过滤条件的结构进行打分. 由于目前的打分函数还不能十分准确的描述蛋白质-蛋白质相互作用的结合自由能, 因此还不够完善, 还需要与其他信息结合起来挑选复合物结构[11,36]. 第四步, 根据上面生成的复合物结构之间的相似程度, 对它们进行成簇聚类. 一般认为簇越大, 即其中相似的结构越多, 则这个簇更有可能包含近天然结构[37]. 最后, 综合打分、成簇和结合位点信息挑出 10 个结构作为最终结果提交给CAPRI竞赛委员会.

根据以上给出的对接方法的 4 个步骤, 图 2 给出了集成分子对接方法 HoDock 的流程图, 下面根据图 2 说明各个步骤具体的做法.

(1) 结合位点残基的确定. 从理论预测和文献检索分析两方面来综合确定结合位点残基. 根据 CAPRI 竞赛委员会提供的结构文件中的蛋白质一级结构、生物学名称等信息去检索相关文献和数据库, 分析同源蛋白质的功能及蛋白质-蛋白质相互作用位点信息, 从中挖掘出对复合物结构预测有价值的结合位点残基和相互作用模式信息. 同时, 采用上面提到的理论预测方法找出A链和B链的可能的结合位点残基. 由于 CAPRI 竞赛的内容越来越难, 若涉及的蛋白质体系还没有人做过专门系统的实验分析, 或者是虽然做过实验但结果还未发表, 也就找不到较好的实验信息, 这时只能通过理论预测来确定结合位点残基. 一般, 从来自不同的实验文献中的信息与理论预测中获得的结合位点残基并不一定一致. 所以, 在工作中总结出一个基本原则, 若满足以下 3 个条件中任一个的残基都可能被确定为结合位点残基: 在实验文献和理论预测中都确认的位点; 理论预测中没有但实验文献特别强调的位点; 实验文献中没有但理论预测时排序靠前的位点.

(2) 初步刚性对接复合物结构的采集. 通过刚性对接初步采集复合物结构, 包括位置约束局部刚性对接和全局刚性对接, 从中挑出满足结合位点信息并且打分较好的结构作为可能的初步对接结果. 用 ZDock 2.3 对接程序[38]进行刚性对接, 让受体A固定不动, 配体B转动, 每次单独对接都采集 2000 个复合物结构. 在位置约束局部对接中, 把那些从结合位点残基信息中确认不在界面上的残基排除掉, 就可

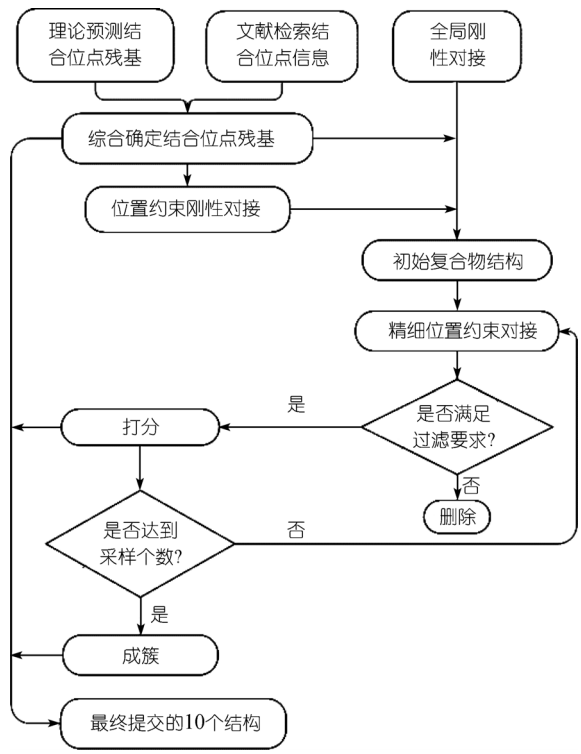

图 2　集成分子对接方法 HoDock 的流程图

以更快地采集满足结合位点条件的结构. 这里采用CAPRI的评估标准来定义结合界面残基: 如果受体中一个残基的任一非氢原子(C, N, O, S)与配体中一个残基的任一非氢原子间的距离小于0.5 nm, 那么这两个残基都是界面残基[11]. 在全局刚性对接中受体A和配体B的起始位置是随机的, 而且不排除任何残基在界面上的可能性. 然后, 对局部刚性对接和全局刚性对接两种采样得到的所有结构进行打分排序并结合位点信息来挑选初始结构. 每一种可能的结合位点情况都至少要有一个结构被选入初始结果, 那些打分比较好的情况可以有 2 或 3 个结构被挑选出来.

(3) 考虑精细位置约束和侧链柔性的对接复合物结构的采集. 在该对接过程中, 采用本研究提出的过滤标准把错误的结构尽早过滤掉, 以提高采样效率[13]. 具体是由上一步得到的复合物结构作为对接的初始结构, 采用改进的RosettaDock[13,39]程序做局部微扰对接, 每组对接都采集 10000 个复合物结构. 限制配体分子相对受体分子的运动距离和旋转角度, 这里使用RosettaDock的3个默认微扰参数, 法线方向范围n取 0.3 nm, 平行方向范围p取 0.8 nm, 旋转范围r取 8°. 考虑到一般天然蛋白质三维结构中的主链氢键都被疏水基团包埋, 在本实验改进的RosettaDock程序中, 引入了一个可提高采样效率的过滤条件, 把不满足主链氢键被疏水基团完全包埋的结构提前排除掉[40].

(4) 打分和成簇. 采用RosettaDock的多项打分函数[39]对生成的所有结构进行排序. 然后对每组对接采集的结构分别成簇聚类, 成簇算法的基本思想是以 0.25 nm作为主链原子均方根偏差(Root Mean Square Deviation, RMSD)的截断距离, 把RMSD小于0.25 nm的结构归为同一簇, 并对每组对接采集的结构按照簇中所含结构的数目从多到少排序.

(5) 挑选最终复合物结构. 综合考虑结合位点信息、打分值和成簇情况挑选出 10 个近天然结构. 分析所有结构的结合界面, 挑出符合结合位点信息的结构. 找到它们所对应的簇, 并在每组对接采集的结构中按照簇的大小排序. 最后, 从排在前面的簇中挑出打分值最高的结构作为最终结构. 由于每组对接代表了不同的结合位点情况, 所以至少要从每组对接中挑选出 1 个结构, 对于打分值和成簇效果都比较好的组可挑选出 2 个或者更多结构进入最后提交的10 个结构中.

2 结果与讨论

2.1 A 链和 B 链的结合位点残基

Venkateswarlu等人[41]发现, A链蛋白centaurin-alpha 1 的N端结构域可与包含B链蛋白的KIF13B分子结合, 这一相互作用可使centaurin-alpha 1 集中于细胞中特定的膜区域, 使其ARF激活功能受到调控. Durocher等人[33]发现, FHA结构域β3/4, β4/5, β6/7 间无规卷曲的ARG70, ARG83 与Ser85, Asn107 等保守残基偏向与其他蛋白相互作用, 考虑到残基ARG的界面活泼特性[42~44], 由实验信息推测同为FHA结构域的B链的B512ARG和B531ARG为结合位点 残基.

理论预测对受体 A 链和配体 B 链分别给出了可能的结合位点残基块(图 1). A 链最可能的第一块结合位点包含 3 个残基(A190GLY, A191HIS 和 A192PRO), 第二块结合位点包含 4 个残基(A28ASP, A29PRO, A30ASP 和 A31TRP). B 链第一块包括 3 个残基(B512 ARG, B514PHE 和 B515VAL), 第二块包括 2 个残基(B531ARG 和 B532ILE).

综合分析实验信息和理论预测的结果, 可以确定出可能的结合位点残基. 对于A链, 预测得到的第一个结合位点块虽然没有文献信息支持, 但从预测结果来看, 它排序靠前, 最后确定其中带正电荷的A191HIS作为结合位点残基, 关于残基HIS在蛋白质相互作用中的活性已有文献报道[45,46]. CAPRI竞赛结果也表明A191HIS确实在正确结构的界面上. 图 3 显示了预测的正确结构Model 10, 包含A链受体和B链配体, 它与天然晶体结构的L_Rmsd为 0.25 nm. 图 3 还标出了 3 个正确的结合位点残基. 由图可见界面上包含A191HIS位点, 实际上还包含第一个结合位点块的所有其他残基(图中未显示). 预测得到的第二个结合位点块与实验文献信息相符. 有研究表明, 溶剂可接近面积大的残基TRP倾向于参与蛋白质分子间的相互作用[47,48], 所以本实验把A31TRP选为可能的结合位点残基. 但竞赛结果表明, A31TRP并不在Target 39 的复合物界面上, 所以按照A31TRP为结合位点的

预测结构都是错误的. 对于B链, 由于实验信息中的两个残基B512ARG和B531ARG都在理论预测得到的两个结合位点块中, 所以就直接把其作为结合位点残基, 竞赛结果表明这个判断是正确的, 这两个残基都在本研究预测的正确结构Model 10 的界面上.

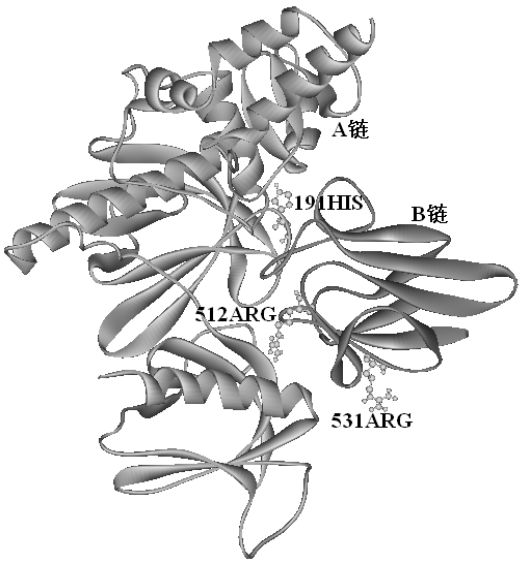

图 3　CAPRI Target 39 复合物结构预测结果 Model 10

它与天然晶体结构的 L_Rmsd 为 0.25 nm, 3 个预测正确的结合位点残基 A191HIS, B512ARG, B531ARG 用球棍模型显示

2.2　初步刚性对接

首先, 用 ZDock 进行初步全局刚性对接, 从全局对接生成的 2000 个复合物结构中挑选满足结合位点信息的结构. 表 1 给出了包含 4 个结合位点的对接复合物结构编号. 从表中可以看出, 没有一个对接复合物结构的界面同时包含 4 个位点, 包含 3 个结合位点残基的只有 3 号结构, A191HIS 和 B512ARG 都在界面上的有 3, 24 号结构, A191HIS 和 B531ARG 都在界面上的只有 3 号结构, A31TRP 和 B512ARG 都在界面上的有 8, 22 号结构, B512ARG 和 B531ARG 都在界面上的有 1, 3 号结构, 所以选定 1, 3, 8, 22 和 24 号结构作为下一步 RosettaDock 对接的起始结构. 然后, 按照同样的方法从局部对接结果中挑选出界面上包含结合位点的 6 个结构, 总共挑出了 11 个结构作为 RosettaDock 精细对接的起始结构.

表 1　界面含有预测结合位点残基的初步对接复合物结构

预测的结合位点残基	初步对接复合物结构编号[a)]
A191HIS	3,4,6,11,12,18,20,24,28
A31TRP	8,22
B512ARG	1,3,8,22,23,24,26,27,29,30
B531ARG	1,3,5,14

a) 由ZDock对接给出

2.3　精细对接

采用改进的RosettaDock程序对这 11 个初始结构分别进行局部微扰对接[39], 这 11 组对接分别被标记为T39Rd1, T39Rd2,…… , T39Rd11. 由于竞赛时间和计算条件的限制, 不可能进行大规模的采样, 这里每组对接采集 10000 个结构, 总共采集了 110000 个结构. 在该过程中通过打分函数中的范德华排斥项找出原子接触太多的对接结构, 并提前将其排除掉, 最终得到的对接结构用RosettaDock的多项打分函数进行打分排序.

2.4　打分排序和成簇

把每组对接生成的 10000 个结构按照打分结果排序, 对前面的 200 个结构(Top 200)成簇聚类. 一般第一簇比较大, 后面的簇比较小. 根据不同的对接情况, 200 个结构能聚成 10~30 个簇, 本研究只从最大簇中挑选结构. 表 2 给出了 11 组对接的 Top 200 成簇后最大簇所含有的结构数目以及该簇中打分排序在第一位的那个结构的打分值.

2.5　最终提交的 10 个复合物结构

分析发现, 从 11 组对接得到的 11 个最大簇里的结构虽然相互迥异, 但它们都至少含有一个结合位点残基在结合界面上, 所以从每个最大簇中挑出打分值最好的那个结构, 总共得到 11 个结构. 由于 T39Rd6 与 T39Rd1 中打分值最高的结构进行几何叠落后 RMSD 不足 0.1 nm, 所以保留了打分值稍好的 T39Rd1 中的结构, 而删除了 T39Rd6 的结构, 剩下的 10 个结构就是最后提交的结构. 由于 CAPRI 复合物结构预测竞赛主要关心是否能预测出正确结构, 而且最终提交的结构综合考虑了打分、成簇和结合位点信息, 所以本实验在结构提交时并没有严格按照打分高低的顺序排列, 而是按照对接的前后顺序挑出结构写入提交的结果文件中.

表 3 列出了提交的 10 个结构 Model 1(M1)~

Model 10(M10)的分析结果. 可以看出, 每个结构都至少包含 2 个结合位点残基在界面上, 其中 M4 和 M10 的界面上还包含有 3 个位点残基(A191HIS, B512ARG 和 B531ARG), 而且 M10 的 L_Rmsd 只有 0.25 nm, 这表明 3 个结合位点残基的预测是正确的(图 3). 但是, 同样包含这 3 个残基在界面上的 M4 却是错误结构, 表明 M4 中受体与配体的结合方位有偏差.

对于配体均方根偏差 L_Rmsd, 提交的 10 个结构中只有 M10 一个结构达到了 CAPRI 可接受(acceptable)结构的标准(<1 nm), 其他结构的 L_Rmsd 都超过了 2 nm, M4 的 L_Rmsd 是 2.3 nm, M3 和 M9 两个结构则超过了 5 nm. 这些结果表明, 即使界面残基预测正确, 若受体和配体之间的结合方位稍微变动, 原子水平的 RMSD 就会差别很大. Target 39 这个体系比较大, 受体有 357 个残基, 配体有 98 个残基, 而 M10 与晶体结构的 L_Rmsd 只有 0.25 nm, 表明 M10 与晶体结构中受体与配体的结合方位基本一样. 由于大多数 X-ray 晶体结构的分辨率是在 0.1~0.3 nm 之间, 而 M10 与晶体结构的偏差是 0.25 nm, 这表明 M10 结构基本达到了晶体结构的准确程度.

表 3 最后 3 列数据是 CAPRI 组委会对本研究组提交的 10 个结构的评价结果. L_Rmsd 是配体均方根偏差, 它是把预测复合物结构和天然晶体复合物结构中的受体叠落后, 计算两个配体的主链原子之间的几何偏差结果. 配体结合界面残基预测正确率(fIR_Ligand)是预测复合物结构里配体界面上的正确残基数除以天然复合物结构里配体界面上的残基数的百分比. 受体结合界面残基预测正确率(fIR_ Receptor)是受体的界面预测成功率, 计算方法与 fIR_ Ligand 相同.

由表 3 还可看出, 有 4 个结构(M1, M2, M7 和 M10)的 fIR_Ligand 大于 70%, 表明这些结构中的配体基本上以正确的表面区域与受体结合. 特别是

表 2 每组精细对接 Top200 成簇后最大簇包含的结构数目及该簇中排序第一的打分值

组别	簇大小[a)]	打分[b)]	组别	簇大小[a)]	打分[b)]
T39Rd1	69	–760.61	T39Rd7	92	–762.65
T39Rd2	38	–761.66	T39Rd8	78	–762.51
T39Rd3	67	–759.46	T39Rd9	71	–762.38
T39Rd4	163	–761.11	T39Rd10	104	–761.13
T39Rd5	42	–760.68	T39Rd11	65	–760.93
T39Rd6	37	–760.42			

a) 最大簇包含的结构数目; b) 该簇中排序第一的打分值

表 3 提交的 10 个复合物结构的相关分析结果

编号	包含的预测结合位点残基	L_Rmsd/nm	fIR_Ligand[a)]	fIR_Receptor[b)]
M1	A191HIS,B512ARG,B531ARG	2.39	0.72	0.62
M2	B512ARG,B531ARG	2.32	0.72	0.38
M3	A31TRP,B512ARG	5.04	0.33	0
M4	B512ARG,B531ARG	2.30	0.67	0.48
M5	A31TRP,B512ARG	2.29	0.28	0.24
M6	A191HIS,B512ARG	2.20	0.44	0.38
M7	B512ARG,B531ARG	2.36	0.72	0.38
M8	A31TRP,B512ARG	2.87	0.17	0.19
M9	A31TRP,B512ARG	5.41	0.11	0
M10	A191HIS,B512ARG,B531ARG	0.25	1.00	0.91

a) fIR_Ligand 为配体结合界面残基预测正确率; b) fIR_Receptor 为受体结合界面残基预测正确率

M10 的配体界面预测成功率达到 100%, 更进一步证明本研究预测的配体界面结合位点残基是完全正确的. 有 3 个结构(M5, M8 和 M9)的配体界面预测成功率不到 30%, 虽然它们的界面上都有 1 个正确结合位点残基 B512ARG, 但它们却以错误的方位与受体结合.

M10 的 fIR_Receptor 达到了 91%, 表明预测的受体 A 链的结合位点残基 A191HIS(由于在边缘, 所以不能确定一定在结合界面上)很可能在天然复合物的界面上. 另一个结合残基 A31TRP 位于 4 个结构(M3, M5, M8 和 M9)的界面上, 但这些结构的受体界面预测成功率很低, 特别是 M3 和 M9 的受体界面没有成功预测到任何结合位点残基, 这表明先前预测 A31TRP 在界面上是错误的. 由于实验文献只给出了受体中包含 A31TRP 的 N 端结构域参与相互作用, 没有给出其他两个结构域的信息, 而这里的结合位点残基 A191HIS 位于中间结构域上, 所以理论预测弥补了实验文献的不足.

综上所述, 提交的 10 个结构中有 5 个结构(M1, M2, M4, M7 和 M10)的 fIR_Ligand 和 fIR_Receptor 数值比较高, 特别是 M1 和 M10 的受体/配体界面预测成功率分别达到了 0.72/0.62 和 1.00/0.91, 表明这些结构中的结合位点预测准确程度比较高, 同时也说明结合位点预测对于蛋白质复合物结构预测至关重要.

表 4 列出了HoDock与 3 个常用的对接方法(ZDock[23], RosettaDock[25]和ClusPro[24])在CAPRI Target 39 竞赛中所提交的 10 个结构的评价结果 L_Rmsd, fIR_Ligand (fIR_L)和fIR_Receptor (fIR_R)的数值. 可以看出, HoDock和ClusPro的Model 10 的 L_Rmsd都小于 0.5 nm, 分别为 0.25 和 0.23 nm, 达到了CAPRI竞赛组委会规定的“中等”好结构的标准. RosettaDock和ZDock的最小L_Rmsd分别为 3.09 和 4.82 nm, 其对接结果都是错误的. 从配体结合界面残基预测正确率fIR_Ligand可以看出, HoDock的最高值 1 和最低值 0.11 都比ClusPro的相应值 0.95 和 0 高. 而 RosettaDock, 只有Model 8 结构的fIR_Ligand为 0.33, 是部分正确的, 其他 9 个结构则完全错误. ZDock的所有 10 个结构的fIR_Ligand都是 0, 配体结合界面预测完全错误. 对于fIR_Receptor, HoDock的最高值是 0.91, ClusPro的最高值是 1, 而RosettaDock和ZDock的最高值分别为 0.78 和 0.94, 所以这 4 种方法对受体结合界面的预测比配体好. 在本实验室以往的CAPRI竞赛中, 有时也会预测对了一个或两个单体的结合位点, 但由于它们相对的结合方位不正确, 最终导致结果错误. 在CAPRI Target 39 的应用中, HoDock比ZDock和RosettaDock结果好的原因可能在于本实验室不仅综合了ZDock和RosettaDock的优

2.6 HoDock 与其他对接方法的结果比较

表 4 HoDock 与其他对接方法的结果比较(L_Rmsd 的单位为 nm)

编号	HoDock			ClusPro			RosettaDock			ZDock		
	L_Rmsd	fIR_L[a)]	fIR_R[b)]	L_Rmsd	fIR_L	fIR_R	L_Rmsd	fIR_L	fIR_R	L_Rmsd	fIR_L	fIR_R
M1	2.39	0.72	0.62	5.81	0	0.61	5.75	0	0.33	5.23	0	0.50
M2	2.32	0.72	0.38	2.81	0.48	0.44	5.03	0	0.5	5.06	0	0.94
M3	5.04	0.33	0.00	5.66	0	0.72	6.36	0	0.33	4.82	0	0.94
M4	2.30	0.67	0.48	5.99	0	0.39	6.47	0	0.28	4.83	0	0.50
M5	2.29	0.28	0.24	2.27	0.76	0.33	5.70	0	0.5	5.47	0	0.22
M6	2.20	0.44	0.38	5.64	0	0.39	5.30	0	0.06	5.70	0	0.11
M7	2.36	0.72	0.38	2.60	0.19	0.61	5.66	0	0	5.01	0	0
M8	2.87	0.17	0.19	2.44	0.29	0.17	3.09	0.33	0.78	5.40	0	0
M9	5.41	0.11	0.00	2.39	0.76	0.94	6.74	0	0.17	5.27	0	0.11

M10	0.25	1.00	0.91	0.23	0.95	1	4.81	0	0.61	5.33	0	0.22

a) fIR_L 是 fIR_Ligand 的简写, 表示配体结合界面残基预测正确率; b) fIR_R 是 fIR_Receptor 的简写, 表示受体结合界面残基预测正确率

点, 而且在对接前预测了结合位点, 从而在复合物结构搜索和打分挑选中提高了计算效率和准确性.

由于 CAPRI 基本集中了目前最好的分子对接方法, 加上严格的组织和评价, 它的结果能够说明各种方法的优劣. 所以竞赛成绩表明本研究的对接方法有一定的优势.

在 CAPRI Target 39 竞赛中, 获得 3 个正确结构的小组, 除了 ClusPro 小组与本小组外, 还有美国 Boston 大学 Sandor Vajda 教授小组, 由于他们以往使用的方法也是 ClusPro, 这次的方法还没有公开报道, 所以在表 4 中未列出他们的结果.

从评估结果可以看出, 本实验的对接方法 HoDock 的准确率仍然需要进一步提高. 在提交的 10 个结构中只有 1 个正确, 其他 9 个结构都错误. 赛后, 本实验室对当初采集的所有结构进行了统一分析, 发现大部分结构的结合方位与 Model 10 偏差较大, 说明还要提高结合位点预测的精度. 而且, 由于打分函数区分度不够, Model 10 的结构打分值也不是最高的.

2.7 如何区分界面上结合位点残基正确而受体与配体方位稍有差别的结构?

如上所述, 结合位点预测比较好的 5 个结构中只有 1 个结构 M10 是正确的, 其他 4 个结构都是错误的. 那么, 怎样才能在对接过程中准确区分这种结合位点基本正确而受体与配体方位却有差别的结构, 从而更有效地挑出近天然结构, 提高对接的成功率呢?

在 CAPRI 竞赛结果公布之后, 本实验室尝试通过成簇和打分两个条件来同时对那些大簇中满足结合位点信息的结构进行区分. 图 4 显示了提交的 10 个结构的打分与各个结构所在簇大小的关系. 从图中可以看出, M10 所在的簇含有最多对接复合物结构, M8 所在簇含有结构最少. 综合考虑打分和成簇两个因素可增加 M10 被选中的机会, 从而提高对接的成功率. 今后本实验室将对以往对接产生的所有结构做进一步的分析来验证这个方法的可行性.

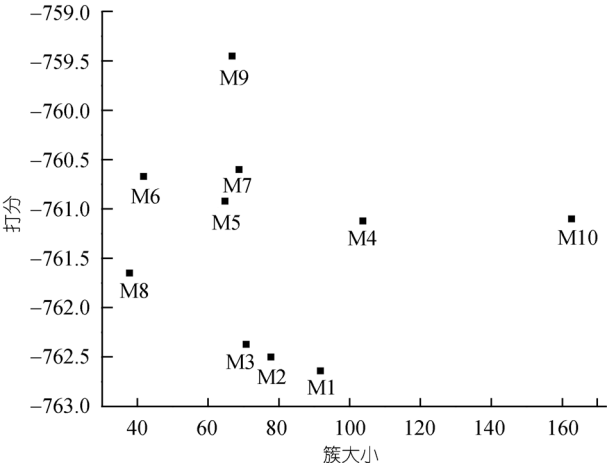

图 4 提交的 10 个复合物结构所在簇的大小与它的打分的二元关系

3 结论

本文提出了用于蛋白质复合物结构预测的集成分子对接方法, 该方法包括结合位点预测、初步刚性对接、精细局部对接、打分和成簇等主要步骤, 最后综合打分、成簇和结合位点信息挑出最佳结构作为复合物结构预测结果. CAPRI Target 39 复合物结构预测结果表明, 提交的最好结构 M10 的 L_Rmsd 只有 0.25 nm, 是仅有的 3 个正确预测结果之一, 这充分说明本实验室提出的集成分子对接方法是行之有效的蛋白质复合物结构预测方法. 由于目前的分子对接方法特别是打分函数还有待完善, 理论预测方法和分析文献实验信息相互结合补充可更好地找准结合位点, 从而缩小复合物结构搜索空间并且引导受体与配体以正确的方位结合. 初步、精细两步对接可以提高对接效率, 更快地找到近天然复合物结构. 打分和成簇相互补充可有利于挑选出正确的复合物结构. 另外, 受体与配体的结合方位是至关重要的, 只有受体与配体的结合位点和结合方位同时正确, 才能保证结构预测的准确率. 这提醒人们要更深入地协调用好打分和成簇两个评价标准, 来选出最好的近天然结构, 同时还要去进一步挖掘更多的结合位点以及它们参与的相互作用模式等更深层的信息.

参考文献

1 Ito T, Chiba T, Ozawa R, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA, 2001, 98: 4569—4574[DOI]

2 Giot L, Bader J S, Brouwer C, et al. A protein interaction map of *Drosophila melanogaster*. Science, 2003, 302: 1727—1736[DOI]

3 Li S, Armstrong C, Bertin N, et al. A map of the interactome network of the metazoan *C. elegans*. Science, 2004, 303: 540—543[DOI]

4 Rual J F, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature, 2005, 437: 1173—1178[DOI]

5 Hwang H, Pierce B, Mintseris J, et al. Protein-protein docking benchmark version 3.0. Proteins, 2008, 73: 705—709[DOI]

6 Janin J, Henrick K, Moult J, et al. CAPRI: a critical assessment of PRedicted interactions. Proteins, 2003, 52: 2—9[DOI]

7 Keskin O, Gursoy A, Ma B, et al. Principles of protein-protein interactions: what are the preferred ways for proteins to interact? Chem Rev, 2008, 108: 1225—1244[DOI]

8 Wodak S J, Janin J. Computer analysis of protein-protein interaction. J Mol Biol, 1978, 124: 323—342[DOI]

9 Jiang F, Kim S H. "Soft docking": matching of molecular surface cubes. J Mol Biol, 1991, 219: 79—102[DOI]

10 李春华, 马晓慧, 陈慰祖, 等. 蛋白质-蛋白质分子对接方法研究进展. 生物化学与生物物理进展, 2006, 33: 616—621

11 Lensink M F, Méndez R, Wodak S J. Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins, 2007, 69: 704—718[DOI]

12 Méndez 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]

13 Gong X Q, Chang S, Zhang Q H, et al. A filter enhanced sampling and combinatorial scoring study for protein docking in CAPRI. Proteins, 2007, 69: 859—865[DOI]

14 高莹, 来鲁华. 蛋白质-蛋白质相互作用界面统计分析. 物理化学学报, 2004, 20: 676—679

15 Irina K, Levon B, Eugene R, et al. PIER: protein interface recognition for structural proteomics. Proteins, 2007, 67: 400—417[DOI]

16 Porollo J, Meller J. Prediction-based fingerprints of protein-protein interactions. Proteins, 2007, 66: 630—645[DOI]

17 Neuvirth H, Heinemann U, Birnbaum D, et al. ProMateus-an open research approach to protein-binding sites analysis. Nucleic Acids Res, 2007, 35: W543—548[DOI]

18 Friedrich T, Pils B, Dandekar T, et al. Modelling interaction sites in protein domains with interaction profile hidden Markov models. Bioinformatics, 2006, 22: 2851—2857[DOI]

19 Negi S S, Schein C H, Oezguen N, et al. InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics, 2007, 23: 3397—3399[DOI]

20 Qin S, Zhou H X. meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics, 2007, 23: 3386—3387[DOI]

21 Bradford J R, Westhead D R. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics, 2005, 21: 1487—1494[DOI]

22 Zhou H X, Qin S. Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics, 2007, 23: 2203—2209[DOI]

23 Wiehe K, Pierce B, Mintseris J, et al. ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. Proteins, 2005, 60: 207—213[DOI]

24 Comeau S R, Vajda S, Camacho C J. Performance of the first protein docking server ClusPro in CAPRI rounds 3-5. Proteins, 2005, 60: 239—244[DOI]

25 Schueler-Furman O, Wang C, Baker D. Progress in protein-protein docking: atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. Proteins, 2005, 60: 187—194[DOI]

26 Ma X H, Li C H, Shen L Z, et al. Biologically enhanced sampling geometric docking and backbone flexibility treatment with multiconformational superposition. Proteins, 2005, 60: 319—323[DOI]

27 Nabuurs S B, Wagener M, de Vlieg J. A flexible approach to induced fit docking. J Med Chem, 2007, 50: 6507—6518[DOI]

28 Zhao Y, Stoffler D, Sanner M. Hierarchical and multi-resolution representation of protein flexibility. Bioinformatics, 2006, 22: 2768—2774[DOI]

29 Wang C, Bradley P, Baker D. Protein-protein docking with backbone flexibility. J Mol Biol, 2007, 373: 503—519[DOI]

30 Li C H, Ma X H, Chen W Z, et al. A protein-protein docking algorithm dependent on the type of complexes. Protein Eng, 2003, 16:

265—269[DOI]
31 Tanaka K, Imajoh O S, Sawada T, et al. A target of phosphatidylinositol 3,4,5-trisphosphate with a zinc finger motif similar to that of the ADP-ribosylation-factor GTPase-activating protein and two pleckstrin homology domains. Eur J Biochem, 1997, 245: 512—519[DOI]
32 Venkateswarlu K, Oatey P B, Tavare J M, et al. Identification of centaurin-alpha1 as a potential *in vivo* phosphatidylinositol 3,4,5-trisphosphate-binding protein that is functionally homologous to the yeast ADP-ribosylation factor (ARF) GTPase-activating protein, Gcs1. Biochem J, 1999, 340: 359—363[DOI]
33 Durocher D, Taylor I A, Sarbassova D, et al. The molecular basis of FHA domain: phosphopeptide binding specificity and implications for phosphodependent signaling mechanisms. Mol Cell, 2000, 6: 1169—1182[DOI]
34 Durocher D, Jackson S P. The FHA domain. FEBS Lett, 2002, 513: 58—66[DOI]
35 Hofmann K, Bucher P. The FHA domain: a putative nuclear signaling domain found in protein kinases and transcription factors. Trends Biochem Sci, 1995, 20: 347—349[DOI]
36 Shen L, Li C, Ma X, et al. Scoring function for the other-type protein complexes. Acta Phys Chim Sin, 2006, 22: 622—626[DOI]
37 Stephan L, Yang Z. Identification of near-native structures by clustering protein docking conformations. Proteins, 2007, 68: 187—194[DOI]
38 Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins, 2003, 52: 80—87[DOI]
39 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]
40 Fernandez A, Scheraga H A. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. Proc Natl Acad Sci USA, 2003, 100: 113—118[DOI]
41 Venkateswarlu K, Hanada T, Chishti A H. Centaurin-α1 interacts directly with kinesin motor protein KIF13B. J Cell Sci, 2005, 118: 2471—2484[DOI]
42 Jandu S K, Ray S, Brooks L, et al. Role of arginine 67 in the stabilization of chymotrypsin inhibitor 2: examination of amide proton exchange rates and denaturation thermodynamics of an engineered protein. Biochemistry, 1990, 29: 6264—6269[DOI]
43 Moreira I S, Fernandes P A, Ramos M J. Hot spots——a review of the protein-protein interface determinant amino-acid residues. Proteins, 2007, 68: 803—812[DOI]
44 Bogan A A, Thorn K S. Anatomy of hot spots in protein interfaces. J Mol Biol, 1998, 280: 1—9[DOI]
45 Schneider E L, Marletta M A. Heme binding to the histidine-rich protein II from *Plasmodium falciparum*. Biochemistry, 2005, 44: 979—986[DOI]
46 Guney E, Tuncbag N, Keskin O, et al. HotSprint: database of computational hot spots in protein interfaces. Nucleic Acids Res, 2008, 36: D662—D666[DOI]
47 Ma B, Pan Y, Gunasekaran K, et al. The contribution of the Trp/Met/Phe residues to physical interactions of p53 with cellular proteins. Phys Biol, 2005, 2: S56—S66[DOI]
48 Negi S, Braun W. Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces. J Mol Model, 2007, 13: 1157—1167[DOI]