[Review] doi: 10.3866/PKU.WHXB201202022 www.whxb.pku.edu.cn

蛋白质-蛋白质分子对接中打分函数研究进展

王存新[1,*,§] 常 珊[2,§] 龚新奇[3] 杨 峰[1] 李春华[1] 陈慰祖[1]
([1]北京工业大学生命科学与生物工程学院, 北京 100124; [2]华南农业大学信息学院, 广州 510642;
[3]清华大学生命科学学院, 北京 100084)

摘要: 分子对接是研究分子间相互作用与识别的有效方法. 其中, 用于近天然构象挑选的打分函数的合理设计对于对接中复合物结构的成功预测至关重要. 本文回顾了蛋白质-蛋白质分子对接组合打分函数中一些主要打分项, 包括几何互补项、界面接触面积、范德华相互作用能、静电相互作用能以及统计成对偏好势等打分项的计算方法. 结合本研究小组的工作, 介绍了目前普遍使用的打分方案以及利用与结合位点有关的信息进行结构筛选的几种策略, 比较并总结了常用打分函数的特点. 最后, 分析并指出了当前蛋白质-蛋白质对接打分函数所存在的主要问题, 并对未来的工作进行了展望.

关键词: 蛋白质-蛋白质分子对接; 打分函数; 打分策略; 结合位点信息
中图分类号: O641

Progress in the Scoring Functions of Protein-Protein Docking

WANG Cun-Xin[1,*,§] CHANG Shan[2,§] GONG Xin-Qi[3] YANG Feng[1]
LI Chun-Hua[1] CHEN Wei-Zu[1]
([1]*College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100124, P. R. China*;
[2]*College of Informatics, South China Agricultural University, Guangzhou 510642, P. R. China*;
[3]*School of Life Sciences, Tsinghua University, Beijing 100084, P. R. China*)

Abstract: Molecular docking technology is an effective approach for prediction of intermolecular interactions and recognition. The design of a scoring function for selecting near-native structures is very important for successful prediction of complex structures. In this article, the main computational methods for scoring items in protein-protein docking, such as geometric complementarity, contact area, van der Waals′ interaction, electrostatic interaction, and statistical pair propensity potential, are reviewed. Including our work, we introduce commonly used scoring schemes and some strategies in screening decoys based on the information for protein binding sites. The characteristic scoring functions in the commonly used docking programs are compared and summarized. The major problems in the existing scoring function in protein-protein docking are discussed along with prospect for future research.

Key Words: Protein-protein docking; Scoring function; Scoring scheme; Binding site information

1 引 言

继人类基因组计划测序完成之后, 以蛋白质间相互作用与识别为核心内容的蛋白质组学(proteomics)的研究迅速兴起, 成为国际上的研究热

Received: November 22, 2011; Revised: January 12, 2012; Published on Web: February 2, 2012.
*Corresponding author. Email: cxwang@bjut.edu.cn; Tel: +86-10-67392724.
§These authors contribute equally to this work.
The project was supported by the National Natural Science Foundation of China (10974008, 31171267), Specialized Research Fund for the Doctoral Program of Higher Education, China (200800050003), and International Science & Technology Cooperation Program of China (2010DFA31710).
国家自然科学基金(10974008, 31171267), 教育部博士点基金项目(200800050003)和科技部国际合作项目(2010DFA31710)资助

点.[1−3] 蛋白质-蛋白质对接方法是研究蛋白质间相互作用与识别的有效的计算机模拟方法, 该方法不仅可揭示分子间的识别机理和蛋白质的结构−功能关系, 而且可带动相关领域的应用研究, 如蛋白质工程分子设计和计算机辅助药物设计方面的研究.[4−7]

蛋白质-蛋白质分子对接是已知受体和配体分子的三维结构, 或者是已知一个单体的三维结构和另一个单体的一级序列, 通过计算机模拟来预测其复合物的结合模式和三维结构. 目前的蛋白质分子对接方法主要包括以下四个步骤: (1) 构建受体和配体分子的三维模型; (2) 充分采集受体-配体各种可能的结合模式; (3) 初步过滤掉一些明显不可能的结合模式; (4) 从剩余的对接结构中筛选出与复合物天然结构相类似的近天然结构. 从热力学的观点来看, 蛋白质分子间的识别和相互作用是一个热力学平衡的过程, 其形成的稳定复合物构象是结合自由能最低的构象. 因此, 采用结合自由能来评价通过对接得到的复合物结构无疑是最直接的方式. 液相环境下, 受体与配体形成复合物的过程中结合自由能的变化ΔG_{bind}包括以下几项:

$$\Delta G_{\text{bind}}=\Delta H_{\text{gas}}-T\Delta S-\Delta G^{\text{R}}_{\text{solv}}-\Delta G^{\text{L}}_{\text{solv}}+\Delta G^{\text{R-L}}_{\text{solv}} \quad (1)$$

上式中 R 和 L 分别代表受体和配体分子, ΔH_{gas}为气相下分子对接过程中 R 和 L 的焓变, $\Delta G^{\text{R}}_{\text{solv}}$、$\Delta G^{\text{L}}_{\text{solv}}$和$\Delta G^{\text{R-L}}_{\text{solv}}$分别为液相下受体、配体和复合物分子的溶剂化自由能变化, ΔS表示对接过程中的熵变, T为体系的绝对温度. 由于使用公式(1)计算熵变需要耗费大量的机时, 且分子对接模拟过程将产生上百万个结合态结构, 为了加快计算速度, 常采用简化的方法计算结合自由能,[8] 这在蛋白质-蛋白质分子对接中称之为打分函数.

为推动蛋白质-蛋白质相互作用研究和分子对接方法的发展, 欧洲生物信息学中心举办了蛋白质-蛋白质复合物结构预测竞赛(critical assessment of prediction of interactions, CAPRI).[9−11] 自 2001 年开始, 迄今已成功进行了 23 届国际蛋白质-蛋白质对接竞赛, 大大促进了蛋白质-蛋白质分子对接方法的研究进展. 目前, 国际上相关领域的许多小组都在发展自己的打分函数, 然而, 从历次 CAPRI 竞赛结果看, 参赛小组提交预测结构的排序结果仍然在一定程度上缺乏可靠性. 因此, 进一步改进对接方法和打分函数仍然是一个重要的目标. 从 2005 年 CAPRI 第 8 轮竞赛开始, CAPRI 组委会增设了与结构预测竞赛平行的打分预测竞赛, 充分显示了发展准确快速的打分方法的迫切性和重要性. 本文将结合我们小组的研究工作, 就蛋白质-蛋白质分子对接打分函数的国内外进展情况进行系统的综述.

2 传统的打分参量

传统的打分参量或打分项是蛋白质-蛋白质对接打分函数设计的基础, 打分函数的性能也主要体现在对于打分参量的选取和优化上. 目前常用的打分项主要包括: 几何互补项、界面接触面积、范德华和静电项以及统计成对偏好势.

2.1 几何互补项

根据“锁钥模型”, 配体与受体分子会发生类似钥匙和锁的识别关系, 这种识别关系主要依赖两者的几何匹配. 实验结果发现, 绝大多数蛋白质复合物配体与受体的相互作用界面上具有明显的几何互补特征. 因此, 在蛋白质-蛋白质对接方法研究的早期, 几何互补性在复合物结构评价中占有至关重要的地位. 一些早期对接算法的打分函数实际上仅包含几何互补项. 大量的对接模拟结果显示, 几何互补打分用于结合态分子对接(bound docking)的结构评价效果很好, 但用于自由态分子对接(unbound docking)的效果却并不理想, 主要原因在于, 对接模拟中没有真实地考虑在结合过程中复合物所发生的构象变化. 此外, 并非所有的近天然结构都比错误结构具有更好的几何互补性. 因此, 现在的打分函数经常是综合考虑几何互补、能量互补等因素来筛选结构. 但总体来看, 几何互补性仍然是重要的对接结构评价指标. 由于计算该项不太耗时, 人们经常把它作为初步打分来预先排除掉一些不合理的结合模式, 以减少下一步用精细而耗时的打分函数来评估结构的工作量.

几何互补性的计算方法随分子模型的不同而多样化. Katchalski-Katzir 等[12]在他们发展的 FTDock 打分函数中, 将受体和配体分子投影到三维空间的网格中, 定义两个离散函数来描述分子的空间构型, 进而将几何互补性表示为两个离散函数的相关性, 其表达式为

$$E_{\alpha,\beta,\gamma}=\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}\vec{a}_{l,m,n}\vec{b}_{l+\alpha,m+\beta,n+\gamma} \quad (2)$$

其中 $\vec{a}$ 和 $\vec{b}$ 分别是受体和配体的投影离散函数, N 是格点数, α、β、γ分别是配体质心在三个坐标轴上偏离受体质心的格点步数. 用快速傅里叶变换(FFT)算法来加速相关函数的计算, 从而提高了采样效率.

由于FFT算法具有显著的高效性, 现在已广泛应用在一系列蛋白质-蛋白质分子对接程序中, 如ZDOCK,[13] DOT[14]和3D-Dock[15]程序等.

另一个被广泛应用的几何互补性算法是几何哈希方法, 这是来源于图像识别中的算法.[16] 该算法包括预处理和识别两个过程. 在预处理阶段, 依据配体上的关键位点建立几何哈希表; 在识别阶段, 将受体的特征与配体进行匹配, 根据匹配程度来确定配体的方位. 该方法避免了对接采样过程中分子的平移和旋转操作, 从而提高了计算速度. 采用该方法的代表程序主要有PatchDock[17]和LZerD[18]等. 国内江凡小组[19]在其格点模型的对接程序SOFT-DOCK中亦采取了几何互补性打分函数, 并在CAPRI中取得了较好的结果.

虽然几何互补的匹配方法有所区别, 但对其好坏的评价标准还是比较统一的: 即对分子间表面的接触给予奖励; 对分子间内部的交叠给予罚分. 另外, 蛋白质分子在结合过程中往往会发生构象变化, 该变化可以通过部分考虑软化分子表面的方式, 以减少对分子间一定程度的内部交叠的罚分. 我们小组[20]考虑了蛋白质分子表面具有较长侧链的氨基酸残基 (Arg、Lys、Asp、Glu和Met) 的柔性. 在搭建分子模型时, 适当减小了蛋白质分子表面上这五种氨基酸残基的半径, 从而使它们在对接中与其它残基之间具有一定程度的可交叠性, 该方法对自由态分子对接的结果有所改善.

2.2 界面接触面积

界面接触面积是一个重要的结构评价指标. Janin等[21]统计了大量蛋白质-蛋白质复合物的界面面积, 发现其大小在12−16 nm^{2}的范围之间变化. Gardiner等[22]已将该数据用于对接算法中, 他们假设复合物界面近似于球面, 然后估计出12−16 nm^{2}球面所对应的球直径, 并利用该直径参量来筛选近天然结构. Kuntz等[23]的研究发现, 结合能的绝对值开始会随着分子间接触面积的增加而增加, 但是当结合能达到一定数值之后, 其值将不再随界面面积增加而有明显的变化. 这说明了将界面接触面积用作打分参量的合理性. 在实际应用中, 界面接触面积还常被用来衡量复合物界面的疏水效应. 疏水效应在蛋白质-蛋白质结合过程中起着重要的作用, 而且疏水区域往往对应着蛋白质的结合位点. 溶剂可接近表面积常被视为界面接触面积, 许多打分函数都是用溶剂可接近表面积来计算蛋白质-蛋白质结合自由能中的溶剂化自由能.[24−26] Xiao等[27]在其发展的ASPDock对接方法中, 利用原子水化参数(ASP)模型来计算去水化自由能并用于打分, 获得了比几何互补性打分更好的效果.

2.3 范德华与静电相互作用

范德华与静电相互作用在蛋白质-蛋白质相互作用与识别中起着至关重要的作用, 可用理论公式对其相互作用进行定量计算. 范德华相互作用可采用6-12形式的Lennard-Jones势来描述, 即

$$E_{\mathrm{vdW}}=\varepsilon_{ij}\left[\left(\frac{r_{m,ij}}{r_{ij}}\right)^{12}-2\left(\frac{r_{m,ij}}{r_{ij}}\right)^{6}\right] \tag{3}$$

其中r_{ij}为原子i和j之间的距离, $r_{m,ij}$为范德华半径之和, ε_{ij}是势阱深乘积的平方根. Baker等[25]在Rosetta-Dock程序中就采用了这一函数形式, 并将范德华相互作用分成范德华排斥项和吸引项. 为了使范德华能不至于在原子间距离太近时产生过大的数值, 他们还对该函数形式进行了特殊的平滑处理.

静电相互作用可以通过多种不同的方法和程序来计算. 一些分子模拟软件包, 如Delphi,[28] GRASP[29]和UHBD[30]等, 通过求解泊松-玻尔兹曼方程来计算静电相互作用, 但计算速度相对较慢. 分子对接模拟往往采用简单快速的静电计算方法, 如静电库仑势:

$$E_{\mathrm{ele}}=\frac{q_i q_j}{4\pi\varepsilon_{\mathrm{r}} r_{ij}} \tag{4}$$

其中, ε_{r}为介电常数, r_{ij}为原子i和j之间的距离, q_i、q_j分别为原子i和j的电荷. BiGGER算法[31]采用的是点电荷库仑相互作用势, 原子的电荷参数来自Amber力场;[32] 在3D-Dock,[15] DOT[14]和ZDock[13]对接方法中, 静电势变为两个离散函数的相关性的形式, 并通过快速傅里叶变换方法来加速静电能的计算; RosettaDock程序[25]对静电项进行了细致的拆分, 分为短程静电吸引、短程静电排斥、长程静电吸引和长程静电排斥四项.

2.4 统计成对偏好势

统计成对势是一个纯粹的经验势. 为了获得该统计势, 首先要建立蛋白质-蛋白质复合物非冗余结构数据库, 然后统计界面上各种氨基酸(或原子)的成对偏好性, 最后根据玻尔兹曼关系导出氨基酸(或原子)的统计成对势. 该偏好势通常采用复合物界面上某氨基酸(或原子)实际成对出现的概率除以某一参考态下的期望概率. 不同的成对偏好势间的差异主要体现在粗粒化程度、参考态的选取以及用于统

计的数据集的不同. Moont 等[33]的残基成对势函数为:

$$\mathrm{Score}_{i,j}=\mathrm{Score}_{j,i}=\lg(c_{i,j}/e_{i,j}) \quad (5)$$

其中 $c_{i,j}$ 定义为残基 i 和 j 的 C_β 原子间距离在给定截断半径之内的接触数量, $e_{i,j}$ 为期望的接触对数量. 每一个接触对的分值被认为是该配对发生可能性的一个统计度量. 因为这里是将接触对发生概率取对数值, 所以一个构象所有出现的可能性为单个接触对得分的加和.

当不同类型原子 i 与 j 间的距离在给定的截断半径之内时, 则认为这两个原子相互接触并存在原子水平的接触势. Weng、[34,35] Vajda[36] 及 Zhou[37] 小组分别统计了原子成对偏好势, 并将其用于分子对接的初步打分. 为了减少全局搜索时成对偏好势的计算量, 常采用主成分分析方法来处理原始的成对偏好势, 一般采用2–4个主成分便可以得到较好的结果.

3 打分策略

打分策略对于打分函数的计算速度和准确性有着重要的影响,[38] 目前常用的打分策略主要有两大类: 一类是组合打分, 另一类是多阶段打分.

3.1 组合打分

通过一个单独的算法来有效区分正确与错误的结合模式是相当困难的. 因此, 大多数对接方法都采用多种打分函数或函数项的组合来筛选对接结构, 往往可以得到较好的预测结果. Charifson 等[39] 在多个体系上进行打分测试的结果表明: 相对于单一的打分函数, 整合了13种打分函数的综合打分策略, 即选取各种打分函数获得的排在前面的一定数目结构的交集, 可以明显地提高对酶的有效或非有效抑制剂的区分效果. 其原因在于组合打分策略能够有效地排除那些在一种打分中所获得的假阳性结果, 使其在保留的结构中近天然结构的比率显著提高. Terp 等[40] 通过多元统计分析, 整合8种不同的打分函数获得的结果来评价对接结构, 发现可以改进对蛋白质-配体复合物结构的预测能力, 并且能够定量计算出蛋白质-配体的结合能.

我们小组也提出了两个组合打分函数, 一个是针对 Others 类型蛋白质复合物的组合打分函数 ComScore,[41,42] 另一个是包含了基于复杂网络参量打分项的多项组合打分函数 HPNCscore.[43,44] 组合打分函数 ComScore 由原子接触势、范德华和静电相互作用能组成, 采用多元线性回归的方法拟合各项权重. 用 ComScore 打分函数对 CAPRI 比赛的 benchmark 1.0 [45] 中的 17 个 Others 类复合物进行打分测试, 结果表明该打分函数能够体现 Others 类型复合物的相互作用特征, 具备一定的从所采集的大量对接结构中筛选获得近天然结构的能力. 我们应用组合打分函数 ComScore, 曾在 CAPRI T25 体系的打分竞赛中获得第一名.[42] 另外, 我们小组还设计了基于复杂网络参量的打分项, 并将其与其他能量项组合, 提出了一个新的多项组合打分函数 HPNCscore. 该打分函数能够将 RosettaDock 组合打分函数的区分能力提高 12%,[43,44] 并在最近的 CAPRI T35、T37、T40 和 T41 体系的打分比赛中, 获得了比较好的预测结果.[46] 在 T35 体系打分比赛中，我们小组找到了唯一的近天然结构。在 T37、T40 和 T41 体系的打分比赛中，我们最后提交的 10 个打分结果，近天然结构数也都超过了 6 个.

3.2 多阶段打分

采用多阶段打分是目前打分策略的另一个重要的发展趋势. 先采用低分辨的打分函数对结构进行粗筛, 排除不可能的结构, 这样做可为后续精细的打分计算节省计算时间. 如 3D-Dock 程序先采用静电相互作用打分进行过滤, 然后采用几何互补打分进行排序, 最后采用残基成对偏好势进行进一步的打分排序.[15] Camacho 等[47] 使用了一种两阶段的打分算法: 在第一阶段, 用去溶剂化自由能和静电能来初筛对接结构; 在第二阶段, 首先对保留下来的结构进行能量优化, 然后使用包含静电能、溶剂化能和范德华能量项的半经验结合自由能函数来挑选近天然结构. Kollman 等[48] 将分子动力学模拟与分子力学/泊松-波尔兹曼表面积方法(MM/PBSA)相结合, 对 HIV-1 逆转录酶-抑制剂对接模式进行打分预测, 获得了较好的结果, 预测结构与晶体结构的配体均方根偏差为 0.11 nm.

4 结合位点信息的利用

应用受体或配体结合位点信息能极大地减少可能的对接结构的数量,[49−51] 以提高复合物结构预测的成功率. 在刚性对接中, 全空间结合模式的采样可以在合理的时间内完成, 但是对接产生的结构是数以百万计的, 从目前的结构预测水平看, 在不知道任何结合位点信息的情况下, 复合物结构预测的成功率是相当低的. 在柔性对接中, 尤其是将柔性对接应用于药物分子设计和筛选时, 采用精细的全

空间采样方案通常太耗计算机时. 掌握和利用受体结合位点信息, 不仅可减少对接模拟计算时间, 更重要的是可以提高复合物结构的预测成功率.[46]

4.1 结合位点的实验信息

活性位点残基的实验信息可以通过定点突变或者核磁共振滴定(NMR titration)实验获得. 在挑选对接结构时, 热点残基, 即对结合能贡献较大的残基, 往往被视为重要的生物学信息. Hu 等[52]对蛋白质-蛋白质复合物界面进行相似性成簇, 得到了 11 个界面家族(interface family), 发现可以通过对家族内部界面间三维结构的叠落来识别保守的热点残基. 该方法所得到的热点残基信息可以应用于分子对接的研究, 有助于提高复合物结构预测的准确性. NMR 实验也可以获得有关蛋白质结合位点的信息, HADDOCK 对接方法[53]成功地利用了这一信息, 建立了模糊的相互作用约束(AIRs), 并将其作为配体与受体关键位点相互作用的位置约束来筛选对接结构. 由于考虑了准确的结合位点信息, HADDOCK 对接方法在最近的 CAPRI 复合物结构预测比赛中取得了优异的成绩.[54]

我们小组在利用结合位点信息来提高结构预测成功率方面也进行了一些尝试. 鉴于在有些情况下, 人们只知道蛋白质结合区域的信息, 而不能进一步具体到结合位点残基. 针对这一情况, 我们发展了 BESDock 对接方法.[55] 该方法在构象搜索中采用 FFT 算法; 为了利用实验信息, 调整了分子三维网格的赋值, 将结合部位一定区域以外的格点赋值为 0, 对其内部格点仍采用传统赋值. 在采样中, 只有内部格点对几何互补打分有贡献, 外部格点则无贡献. 这样就会更多的采集到受体或配体结合部位附近的复合物结构, 从而大大提高了采样的有效性, 并且发现最优对接结构的配体均方根偏差也有明显降低.

4.2 结合位点的理论预测

在实验获得的结合位点信息缺乏的情况下, 用理论方法预测结合位点也是一种常用的手段.[56] 基于序列分析方法是最常用的结合位点预测方法. 一般认为, 蛋白质界面残基要发挥其生物学功能, 在序列上应该具有较高的保守性. Hu 等[52]指出, 与其他功能类复合物相比, 酶-抑制剂复合物的活性位点残基要比其他位置上的残基更加保守. 基于序列进化性分析的进化轨迹方法(evolutionary trace)[57]已成功应用到蛋白质结合位点预测, 并能获得较为准确的位点信息.

另外, 蛋白质表面的一些特性, 如物理化学性质、几何特征和带电特性等也可用于结合位点的预测. 一些小组利用这些信息, 采用机器学习方法, 如神经网络算法、支持向量机(SVM)等, 对蛋白质结合位点残基进行预测, 也得到了较好的结果.[56,58,59]

5 经典蛋白质-蛋白质分子对接软件中打分函数的设计

表 1 列出了目前国际上常用的蛋白质-蛋白质分子对接方法[13−15,17,24,25,31,53,60]所采用的打分函数的主要特点. 这里我们选择三个具有代表性的对接软件: ZDOCK, RosettaDock 和 HADDOCK, 并对其打分函数的组成和设计进行介绍.

5.1 ZDOCK 软件

ZDOCK 软件由美国波士顿大学的研究小组[13]开发完成. ZDOCK 采用 FFT 方法进行刚性对接, 对接结构采用几何互补、去溶剂化能和静电相互作用进行粗略的打分筛选. 为了更准确地评价打分结果, 后续发展的 ZRANK[61]采用了更精确地打分方法, 其打分函数的表达式为

$$\begin{aligned}\text{Score}=&w_{\text{vdW\_a}}\cdot E_{\text{vdW\_a}}+w_{\text{vdW\_r}}\cdot E_{\text{vdW\_r}}+w_{\text{elec\_sra}}\cdot E_{\text{elec\_sra}}+\\&w_{\text{elec\_srr}}\cdot E_{\text{elec\_srr}}+w_{\text{elec\_lra}}\cdot E_{\text{elec\_lra}}+w_{\text{elec\_lrr}}\cdot E_{\text{elec\_lrr}}+\\&w_{\text{ds}}\cdot E_{\text{ds}}\end{aligned}\tag{6}$$

其中 $E_{\text{vdW\_a}}$ 和 $E_{\text{vdW\_r}}$ 为范德华吸引和排斥能项, $E_{\text{elec\_sra}}$ 和 $E_{\text{elec\_srr}}$ 为短程静电吸引和排斥能, $E_{\text{elec\_lra}}$ 和 $E_{\text{elec\_lrr}}$ 为长程静电吸引和排斥能, E_{ds} 为去溶剂化能. 相应的权重参数 $w_{\text{vdW\_a}}$=1.0, $w_{\text{vdW\_r}}$=0.009, $w_{\text{elec\_sra}}$=0.31, $w_{\text{elec\_srr}}$=0.34, $w_{\text{elec\_lra}}$=0.44, $w_{\text{elec\_lrr}}$=0.50, w_{ds}=1.02. 在打分后期, 还可以采用 RDOCK[62]对排在前 2000 位的对接结构进行进一步能量优化, 以消除原子交叠.

与 ZDOCK 类似的方法还有 3D-Dock,[15] DOT,[14] BiGGER,[31] PatchDock[17]等程序, 这些程序由于采用了 FFT 方法或其他几何互补计算方法, 都能够进行全空间搜索和快速打分评价, 在不确定结合部位信息的情况下往往也能够取得较好的结果. 但这类程序在搜索的初始阶段无法考虑蛋白质结构的柔性, 如果蛋白质对接过程中的柔性较大, 会影响最后的打分结果.

5.2 RosettaDock 软件

RosettaDock 软件由美国华盛顿大学 Baker 教授小组[25,60]开发完成. RosettaDock 程序采用蒙特卡罗算法优化分子结构, 包括侧链包被、刚性最小化和

表1 常用蛋白质分子对接程序打分方法比较

Table 1 Comparison among the scoring functions of the usually used protein-protein docking methods

Software	Scoring terms	Scoring scheme	Extra information	Institution
3D-Dock[15]	geometric complementarity, electrostatic and residue pair potential	multistage scoring	filtering with interface residue pairs	BioMol. Mod. Can. Res. UK (Sternberg)
ZDOCK[13]	geometric complementarity, electrostatic, van der Waals, and residue pair potential	multistage and combined scoring	choosing residues to block from or force into the binding site	Boston Univ. (Weng)
BiGGER[31]	geometric complementarity, electrostatic and desolvation energy	multistage and combined scoring	none	BioTecnol, S.A. (Palma/Moura)
RosettaDock[25,60]	van der Waals, desolvation, hydrogen bonding, electrostatic and residue pair potential	multistage and combined scoring	filtering with binding site constraints	Johns Hopkins Univ. (Gray)
DOT[14]	geometric complementarity and electrostatic potential	combined scoring	none	S. Diego Super-comput. Cen. (Mandell)
HADDOCK[53]	van der Waals, electrostatic potential, buried surface area, desolvation energy, and ambiguous interaction restraints	combined scoring	NMR information	Utrecht Univ. (Bonvin)
ICM-DOCK[24]	van der Waals, electrostatic, solvation, hydrogen bonding, and hydrophobicity potential	multistage and combined scoring	none	MolSoft LLC (Abagyan)
PatchDock[17]	geometric complementarity and atomic desolvation energy	multistage scoring	binding site residues	Tel Aviv Univ. (Wolfson)

最终评分过程, 不同的阶段采取不同的打分函数进行评价. 其打分函数的表达式为:

$$\begin{aligned}\text{Score}=&w_{\text{atr}}\cdot E_{\text{atr}}+w_{\text{rep}}\cdot E_{\text{rep}}+w_{\text{sol}}\cdot E_{\text{sol}}+w_{\text{sasa}}\cdot E_{\text{sasa}}+\\&w_{\text{hb}}\cdot E_{\text{hb}}+w_{\text{dun}}\cdot E_{\text{dun}}+w_{\text{pair}}\cdot E_{\text{pair}}+w_{\text{elec}}^{\text{sr-rep}}\cdot E_{\text{elec}}^{\text{sr-rep}}+\\&w_{\text{elec}}^{\text{sr-atr}}\cdot E_{\text{elec}}^{\text{sr-atr}}+w_{\text{elec}}^{\text{lr-rep}}\cdot E_{\text{elec}}^{\text{lr-rep}}+w_{\text{elec}}^{\text{lr-atr}}\cdot E_{\text{elec}}^{\text{lr-atr}}\end{aligned}\quad(7)$$

上式中 E_{atr} 和 E_{rep} 为范德华吸引和排斥项, E_{sol} 为隐含溶剂化能, E_{sasa} 为基于表面面积的溶剂化能, E_{hb} 为氢键评分, E_{dun} 为转角概率项, E_{pair} 为残基成对势, $E_{\text{elec}}^{\text{sr-rep}}$ 和 $E_{\text{elec}}^{\text{sr-atr}}$ 分别为短程静电吸引和排斥项, $E_{\text{elec}}^{\text{lr-rep}}$ 和 $E_{\text{elec}}^{\text{lr-atr}}$ 分别为长程静电吸引和排斥项. 具体的权重参数见表2. 与RosettaDock程序类似的方法有ICM-Dock[24]等程序, 这些程序都采用了更为精确、复杂的能量函数, 通过蒙特卡罗等算法能够进行局部搜索和侧链优化. 在对接方位比较确定的情况下, 往往能够取得较好的对接效果, 如Weng等[63]也采用了RosettaDock的评分方法对ZDOCK程序对接后的结构进行局部调整. 但如果完全没有任何结合位点信息, 这些方法可能需要搜索较长时间, 或者容易陷入局部极小.

5.3 HADDOCK软件

表2 RosettaDock打分函数的权重

Table 2 Weights used in the RosettaDock scoring function

Scoring term	Weight		
	side-chain packing	rigid-body minimization	discrimination
repulsive van der Waals	0.800	0.338	0.080
attractive van der Waals	0.800	0.338	0.338
Gaussian solvent-exclusion	0.800	0.279	0.279
surface area solvation	–	–	0.344
hydrogen bonding	2.100	0.441	0.441
rotamer probability	0.790	0.069	0.069
residue pair probability	0.660	0.164	0.164
short-range repulsive	–	0.025	0.025
short-range attractive	–	0.025	0.025
long-range repulsive	–	0.098	0.098
long-range attractive	–	0.002	0.002

The last four scoring terms belong to simple electrostatics. '–' indicates that the term is not included in the scoring function.

表3 HADDOCK打分函数的权重参数

Table 3 Weights used in the HADDOCK scoring function

Scoring term	Weight		
	rigid body docking	semi-flexible refinement	explicit solvent refinement
van der Waals	0.01	1.00	1.00
electrostatic	1.00	1.00	0.20
ambiguous interaction restraints	0.01	0.10	0.10
buried surface area	−0.01	−0.01	–
desolvation energy	1.00	1.00	1.00

'–' indicates that the term is not included in the scoring function.

HADDOCK软件由荷兰乌特勒支大学Bonvin教授的研究小组[53]开发完成. HADDOCK程序将能量优化和分子动力学模拟相结合来进行分子对接. 首先通过刚性能量优化和半柔性模拟退火进行构象搜索, 然后采用显含水的分子动力学模拟进行进一步的结构改进. 打分函数的表达式为:

$$\text{Score}=w_{\text{vdW}}\cdot E_{\text{vdW}}+w_{\text{elec}}\cdot E_{\text{elec}}+w_{\text{AIR}}\cdot E_{\text{AIR}}+ w_{\text{BSA}}\cdot A_{\text{BSA}}+w_{\text{desolv}}\cdot E_{\text{desolv}} \quad (8)$$

其中E_{vdW}为范德华项, E_{elec}为静电相互作用, E_{AIR}为模糊相互作用约束项, A_{BSA}为包埋表面面积, E_{desolv}为去溶剂化能. 权重参数在表3中列出.

HADDOCK程序的特点是在打分项中引入位点约束信息(即AIR项), 还采用精确的显含水的分子动力学模拟进行结构优化. 由于过程中考虑了实验信息和水的影响, 在实验信息确定的对接测试中, 可以快速向正确结构收敛, 并得到非常准确的复合物结构. 但如果实验信息缺乏, 打分效果将会受到影响.

6 将来的发展方向

从目前的国内外研究进展看, 今后蛋白质-蛋白质分子对接打分函数的发展方向将主要集中在两个方面: 首先是提高打分函数的精确性. 一些对接算法所采用的打分函数基于经验的势函数, 将一个多粒子体系的相互作用, 近似为用原子间的成对相互作用之和来表示. 因此, 用半经验力场虽然可以处理包含数万个原子的体系, 但结果仍然具有一定的近似性. 由于实际的生物分子存在于水溶液环境下, 因此在结构评价中正确考虑溶剂及离子效应, 发展更加真实地描述蛋白质分子结合过程的打分函数是下一步努力的目标.[64,65] 打分函数发展的另一个方向是提高计算的效率, 如主成分分析等降维方法已被用于打分函数的计算, 可明显提高计算效率.[5] 另外, 随着计算机处理器, 特别是图形处理单元(GPU)计算能力的快速提高, 可以针对该处理器的特点, 发展基于GPU的并行计算方法来加速打分计算.[66−68] 目前, 国内外许多小组正在该领域不断开展积极的探索和研究工作, 相信在不久的将来, 蛋白质-蛋白质分子对接打分函数方面的研究一定会取得更大的进展, 相应的复合物结构预测的成功率也将会不断提高.

References

(1) Pearson, H. *Nature* **2008,** *452*, 920.

(2) Hummon, A. B.; Richmond, T. A.; Verleyen, P.; Baggerman, G.; Huybrechts, J.; Ewing, M. A.; Vierstraete, E.; Rodriguez-Zas, S. L.; Schoofs, L.; Robinson, G. E.; Sweedler, J. V. *Science* **2006,** *314*, 647.

(3) Phizicky, E.; Bastiaens, P. I. H.; Zhu, H.; Snyder, M.; Fields, S. *Nature* **2003,** *422*, 208.

(4) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. *J. Comput. Chem.* **2010,** *31*, 317.

(5) Ritchie, D. W. *Curr. Protein Pept. Sci.* **2008,** *9*, 1.

(6) Gray, J. J. *Curr. Opin. Struct. Biol.* **2006,** *16*, 183.

(7) Li, C. H.; Ma, X. H.; Chen, W. Z.; Wang, C. X. *Progress in Biochemistry and Biophysics* **2006,** *33*, 616. [李春华, 马晓慧, 陈慰祖, 王存新. 生物化学与生物物理进展, **2006,** *33*, 616.]

(8) Li, X. D.; Hou, T. J.; Xu, X. J. *Acta Physico-Chimica Sinica* **2005,** *21*, 504. [李旭东, 侯廷军, 徐筱杰. 物理化学学报, **2005,** *21*, 504.]

(9) Janin, J.; Henrick, K.; Moult, J.; Ten Eyck, L.; Sternberg, M. J. E.; Vajda, S.; Vasker, I.; Wodak, S. J. *Proteins* **2003,** *52*, 2.

(10) Mendez, R.; Leplae, R.; Lensink, M. F.; Wodak, S. J. *Proteins* **2005,** *60*, 150.

(11) Janin, J. *Mol. Biosyst.* **2010,** *6*, 2351.

(12) Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. *Proc. Natl. Acad. Sci. U. S. A.* **1992,** *89*, 2195.

(13) Chen, R.; Li, L.; Weng, Z. P. *Proteins* **2003,** *52*, 80.

(14) Mandell, J. G.; Roberts, V. A.; Pique, M. E.; Kotlovyi, V.; Mitchell, J. C.; Nelson, E.; Tsigelny, I.; Eyck, L. F. T. *Protein Eng.* **2001,** *14*, 105.

(15) Aloy, P.; Querol, E.; Aviles, F. X.; Sternberg, M. J. E. *J. Mol.*

Biol. **2001,** *311*, 395.
(16) Fischer, D.; Lin, S. L.; Wolfson, H. L.; Nussinov, R. *J. Mol. Biol.* **1995,** *248*, 459.
(17) Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. *Nucl. Acids Res.* **2005,** *33*, W363.
(18) Venkatraman, V.; Yang, Y. F. D.; Sael, L.; Kihara, D. *BMC Bioinformatics* **2009,** *10*, 407.
(19) Li, N.; Sun, Z.; Jiang, F. *Proteins* **2007,** *69*, 801.
(20) Li, C. H.; Ma, X. H.; Chen, W. Z.; Wang, C. X. *Proteins* **2003,** *52*, 47.
(21) Chakrabarti, P.; Janin, J. *Proteins* **2002,** *47*, 334.
(22) Gardiner, E. J.; Willett, P.; Artymiuk, P. J. *Proteins* **2003,** *52*, 10.
(23) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. *Proc. Natl. Acad. Sci. U. S. A.* **1999,** *96*, 9997.
(24) Fernandez-Recio, J.; Totrov, M.; Abagyan, R. *J. Mol. Biol.* **2004,** *335*, 843.
(25) Gray, J. J.; Moughon, S.; Wang, C.; Schueler-Furman, O.; Kuhlman, B.; Rohl, C. A.; Baker, D. *J. Mol. Biol.* **2003,** *331*, 281.
(26) Hou, T. J.; Xu, X. J. *Acta Physico-Chimica Sinica* **2002,** *18*, 1052. [侯廷军, 徐筱杰. 物理化学学报, **2002,** *18*, 1052.]
(27) Li, L.; Guo, D.; Huang, Y.; Liu, S.; Xiao, Y. *BMC Bioinformatics* **2011,** *12*, 36.
(28) Anthony, N.; Barry, H. *J. Comput. Chem.* **1991,** *12*, 435.
(29) Nicholls, A.; Sharp, K. A.; Honig, B. *Proteins* **1991,** *11*, 281.
(30) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. *Comput. Phys. Commun.* **1995,** *91*, 57.
(31) Palma, P. N.; Krippahl, L.; Wampler, J. E.; Moura, J. J. G. *Proteins* **2000,** *39*, 372.
(32) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005,** *26*, 1668.
(33) Moont, G.; Gabb, H. A.; Sternberg, M. J. E. *Proteins* **1999,** *35*, 364.
(34) Mintseris, J.; Pierce, B.; Wiehe, K.; Anderson, R.; Chen, R.; Weng, Z. P. *Proteins* **2007,** *69*, 511.
(35) Vreven, T.; Hwang, H.; Weng, Z. P. *Protein Sci.* **2011,** *20*, 1576.
(36) Kozakov, D.; Brenke, R.; Comeau, S. R.; Vajda, S. *Proteins* **2006,** *65*, 392.
(37) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. *J. Med. Chem.* **2005,** *48*, 2325.
(38) Vajda, S.; Kozakov, D. *Curr. Opin. Struct. Biol.* **2009,** *19*, 164.
(39) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. *J. Med. Chem.* **1999,** *42*, 5100.
(40) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jorgensen, F. S. *J. Med. Chem.* **2001,** *44*, 2333.
(41) Li, C. H.; Ma, X. H.; Shen, L. Z.; Chang, S.; Chen, W. Z.; Wang, C. X. *Biophys. Chem.* **2007,** *129*, 1.
(42) Gong, X. Q.; Chang, S.; Zhang, Q. H.; Li, C. H.; Shen, L. Z.; Ma, X. H.; Wang, M. H.; Liu, B.; He, H. Q.; Chen, W. Z.; Wang, C. X. *Proteins* **2007,** *69*, 859.
(43) Chang, S.; Gong, X. Q.; Jiao, X.; Li, C. H.; Chen, W. Z.; Wang, C. X. *Chin. Sci. Bull.* **2010,** *55*, 814.
(44) Chang, S.; Jiao, X.; Li, C. H.; Gong, X. Q.; Chen, W. Z.; Wang, C. X. *Biophys. Chem.* **2008,** *134*, 111.
(45) Chen, R.; Mintseris, J.; Janin, J.; Weng, Z. P. *Proteins* **2003,** *52*, 88.
(46) Gong, X.; Wang, P.; Yang, F.; Chang, S.; Liu, B.; He, H.; Cao, L.; Xu, X.; Li, C.; Chen, W.; Wang, C. *Proteins* **2010,** *78*, 3150.
(47) Camacho, C. J.; Gatchell, D. W.; Kimura, S. R.; Vajda, S. *Proteins* **2000,** *40*, 525.
(48) Wang, J.; Morin, P.; Wang, W.; Kollman, P. A. *J. Am. Chem. Soc.* **2001,** *123*, 5221.
(49) Bai, H. J.; Lai, L. H. *Acta Physico-Chimica Sinica* **2010,** *26*, 1988. [白红军, 来鲁华. 物理化学学报, **2010,** *26*, 1988.]
(50) Zhang, C.; Liu, S.; Zhou, Y. Q. *Proteins* **2005,** *60*, 314.
(51) Heuser, P.; Bau, D.; Benkert, P.; Schomburg, D. *Proteins* **2005,** *61*, 1059.
(52) Hu, Z. J.; Ma, B. Y.; Wolfson, H.; Nussinov, R. *Proteins* **2000,** *39*, 331.
(53) Dominguez, C.; Boelens, R.; Bonvin, A. *J. Am. Chem. Soc.* **2003,** *125*, 1731.
(54) Lensink, M. F.; Wodak, S. J. *Proteins* **2010,** *78*, 3073.
(55) Ma, X. H.; Li, C. H.; Shen, L. Z.; Gong, X. Q.; Chen, W. Z.; Wang, C. X. *Proteins* **2005,** *60*, 319.
(56) Zhou, H. X.; Qin, S. B. *Bioinformatics* **2007,** *23*, 2203.
(57) Lichtarge, O.; Bourne, H. R.; Cohen, F. E. *J. Mol. Biol.* **1996,** *257*, 342.
(58) Bradford, J. R.; Westhead, D. R. *Bioinformatics* **2005,** *21*, 1487.
(59) Gutteridge, A.; Bartlett, G. J.; Thornton, J. M. *J. Mol. Biol.* **2003,** *330*, 719.
(60) Wang, C.; Schueler-Furman, O.; Baker, D. *Protein Sci.* **2005,** *14*, 1328.
(61) Pierce, B.; Weng, Z. *Proteins* **2007,** *67*, 1078.
(62) Li, L.; Chen, R.; Weng, Z. *Proteins* **2003,** *53*, 693.
(63) Pierce, B.; Weng, Z. *Proteins* **2008,** *72*, 270.
(64) Bonvin, A. M. *Curr. Opin. Struct. Biol.* **2006,** *16*, 194.
(65) Andrusier, N.; Mashiach, E.; Nussinov, R.; Wolfson, H. J. *Proteins* **2008,** *73*, 271.
(66) Sukhwani, B.; Herbordt, M. C. *IET Comput. Digit. Tech.* **2010,** *4*, 184.
(67) Dynerman, D.; Butzlaff, E.; Mitchell, J. C. *J. Comput. Biol.* **2009,** *16*, 523.
(68) Korb, O.; Stutzle, T.; Exner, T. E. *J. Chem Inf. Model.* **2011,** *51*, 865.