



构建药物-靶标亲合性打分函数的评价方法体系

刘志海, 李婕, 韩莉, 李嫣*, 王任小*

生命有机化学国家重点实验室; 生命的化学基础协同创新中心; 中国科学院上海有机化学研究所, 上海 200032

*通讯作者, E-mail: kathyli@mail.sioc.ac.cn; wangrx@mail.sioc.ac.cn

收稿日期: 2015-07-06; 接受日期: 2015-08-03; 网络版发表日期: 2015-08-21

doi: 10.1360/N032015-00115

摘要 定量计算药物-靶标亲合性一直是分子靶向药物设计中的核心问题之一. 亲合性打分函数适用范围宽、计算速度快且精度良好, 特别适合与分子对接等方法联用, 在药物研发中得到了广泛的应用. 经过 20 多年的发展, 目前存在数量繁多的打分函数, 这些方法难免良莠不齐. 建立客观公正的针对打分函数性能的评价体系, 不仅可以帮助用户做出合理的选择, 也为发展打分函数的理论研究人员提供依据, 因此是打分函数研究领域中的一个基础问题. 近年来, 我们逐步发展了评价药物-靶标亲合性打分函数性能的方法体系 (comparative assessment of scoring functions, CASF), 为解决该问题提供了一种可行的方案. 我们提出了一整套衡量打分函数性能的基本指标, 为每种指标设计了相应的定量方法, 并确立了将打分过程与构象采样过程分开处理的原则. 我们的方法体系中所使用的测试集, 源自对 PDBbind-CN 数据库中蛋白-配体复合物样本的层层筛选, 质量较高并且具备结构多样性的特点. 与前人的类似工作相比, 我们的方法体系具有较强的系统性和先进性, 受到了国内外同行的广泛认可和应用. 本文介绍了我们建立的 CASF 方法体系的设计思想、获得的重要研究成果以及学术影响.

关键词

药物-靶标亲合性
打分函数
评价体系
CASF
PDBbind-CN
分子靶向药物设计

1 引言

大多数药物通过与特定的生物靶标分子相结合而发挥调控作用. 因此, 发展分子靶向药物是现代药物研发中的主流策略. 在分子靶向药物的研发过程中, 化合物与生物靶标之间的亲合性(binding affinity)是用来筛选先导化合物最重要的依据. 如果可以利用计算方法来准确预测药物与靶标之间的亲合性, 人们就可以实现计算机上的虚拟筛选, 大大提高发现先导化合物的效率, 并且也可以用于指导后续的先导化合物改造等目的. 因此, 发展定量计算药物-靶标亲合性的理论方法, 一直是药物分子设计中的核心问题之一^[1,2].

在用于计算药物-靶标亲合性的理论方法中, 亲合性打分函数(scoring function)的计算速度快、精度良好、适用面广, 特别适合与分子对接等方法联用, 因而是最流行的一类计算方法, 在虚拟筛选等实际应用中取得了许多的成功案例^[3,4]. 自 20 世纪 90 年代初以来, 研究和发展打分函数成为一个非常活跃、竞争激烈的领域. 据我们的不完全统计, 文献中公开报道的打分函数已经有上百种之多. 目前存在的打分函数可以划分为四大类, 即基于物理模型的方法、经验打分函数、统计势函数以及基于描述符的机器学习模型^[5]. 面对如此繁多的打分函数, 人们需要建立一套合理的评价体系(benchmark), 来对各种打分函数的性能进行客观的对比. 依靠此评价体系给出的结

果, 用户可以在可用的打分函数中做出合理的选择, 应用在实际的药物研发工作中. 另外, 准确预测药物-靶标亲合性仍然是一个远未解决的问题, 人们需要更加可靠的打分函数. 对于发展打分函数的理论工作者而言, 同样也需要对目前打分函数的优点和弱点做到心中有数, 才能为未来研制更好的打分函数提供指导.

随着打分函数研究的不断推进, 事实上人们也一直在尝试建立这样的评价体系. 在 21 世纪初, 一系列关于对比评价分子对接方法以及打分函数的工作得以公开发表^[6-12]. 例如, Bissantz 等^[6]在 10 个胸苷激酶复合物和 10 个雌激素受体复合物上分别对 3 种分子对接程序(DOCK、FlexX 和 GOLD)以及 7 种打分函数进行了对比测试; Ferrara 等^[11]选用 189 个蛋白-配体复合物对 9 种打分函数进行了测试; Wang 等^[12]选用 100 个蛋白-配体复合物作为测试集, 综合评价了 11 种常见打分函数预测结合模式及重现亲合性实验数据的能力. 这些早期的评价工作由于采用不同的测试集以及不同的评价标准, 所得出的结论没有普遍适用性, 甚至出现如下情况: 某项评价工作得出“打分函数 A 优于 B”的结论, 而在另外一项评价工作中却认为“打分函数 B 优于 A”, 令人无所适从.

另外, 这些早期工作所采用的方法也存在比较明显的缺陷, 归纳起来主要有以下两个方面:

首先, 评价方法的设计不合理. 打分函数经常与分子对接软件联合应用. 这些早期的评价工作通常把分子对接软件简单地作为“黑箱”来运用, 考察分子对接软件重现已知蛋白-配体结合模式的能力, 或者考察亲合性打分数值与亲合性实验数据是否具有相关性等方面. 但是分子对接操作涉及复杂的构象搜索采样过程, 分子对接软件通常使用多个参数来控制此过程. 如果参数设置不同, 则构象采样的结果会受很大影响. 因此, 分子对接软件的最终输出结果是由它的构象搜索算法、内置打分函数以及一些其他因素(如输入结构文件的处理方法)共同决定的, 不能单独反映打分函数的性能. 只有将计算亲合性的过程(“打分过程”)与分子对接中的构象搜索过程剥离开来, 才能客观真实地评价打分函数的基本性能.

其次, 评价打分函数使用的测试集不具有代表性, 选用的样本质量不高. 早期的工作通常会采用一些已知三维结构和亲合性数据的蛋白-配体复合物构成测试集. 这里有两种可能的情况: (1) 测试集中包

括类型较少的靶蛋白(通常少于 10 种)形成的复合物, 每种靶蛋白有若干个与配体分子匹配形成的复合物. 基于这样的测试集所得出的结果, 理论上局限于测试集中所包括的特定靶标体系, 其结论无法推广到其他靶标体系. (2) 测试集的构建偏重包括结构多样性的靶蛋白-配体复合物. 但是所选用的复合物不是对蛋白质数据库(Protein Data Bank, PDB)^[13]中可用的样本进行系统挑选的结果, 而是源自于作者的日常随机积累. 这样的测试集显然不具备充分的代表性, 而且样本的冗余度无法得到控制, 即测试集中某类型的复合物要大大多于或少于其他的样本. 同样重要的是, 这些早期工作中所选用的蛋白-配体复合物样本, 其中很多在三维结构或亲合性实验数据方面存在各种问题^[14-16]. 测试集中存在的这些质量问题, 势必对最终的评价结果造成不确定性.

针对上述问题, 本课题组近年来经过不断努力, 建立起一整套评价药物-靶标亲合性打分函数性能的方法体系(comparative assessment of scoring functions, CASF), 较之前人的类似工作有了较大提高, 得到了国内外同行的广泛认可, 逐渐成为打分函数研究领域一项主流的评价体系. 下文将介绍该方法体系的设计思想、研究成果以及学术影响.

2 评价亲合性打分函数性能的方法体系

我们设计的 CASF 方法体系包括 3 个要素: (1) 一个高质量的测试集, 包括精选出来的 195 个结构多样性的蛋白-配体复合物; (2) 衡量打分函数性能的 4 种基本指标以及相应的评价方法; (3) 评价对象包括主流药物分子设计商业软件中的 20 种打分函数. 如果把评价打分函数性能比喻为一场体育竞赛, 那么测试集就相当于竞赛场地, 评价方法就相当于竞赛规则, 而作为评价对象的打分函数就相当于参赛选手. 我们关于 CASF 方法体系的工作最早在 2009 年发表^[17], 以后又经过不断的改进和完善. 下文将以我们最新的发表工作 CASF-2013 为例^[18,19], 分别介绍这 3 个要素.

2.1 构建高质量的测试集

我们认为, 用于评价打分函数的一个高质量的测试集必须满足以下几个基本要求: (1) 测试集中的蛋白分子和配体分子必须具有多样性, 以获得更有

普适性的结论, 如果测试集中所包含的蛋白分子大多是药物靶标, 而且配体分子大多是“类药性”强的有机小分子则更具有价值; (2) 测试集中的蛋白-配体复合物样本必须具有高质量的晶体结构和可靠的亲合性实验数据; (3) 测试集要包括足够多的样本, 以保证最终的评价结果具有显著的统计意义, 但同时需要对测试集中的样本进行冗余控制, 以保证每种靶蛋白形成的复合物对最终统计结果的贡献大致均等.

本课题组在过去 10 年间, 建立并不断发展了 PDBbind-CN 数据库, 为挑选 CASF 方法体系中所需的测试集提供了绝佳的素材. 该数据库在国际上首次致力于系统地收集 PDB 中各类蛋白-配体复合物的三维结构以及亲合性实验数据, 提供结构信息和物理化学性质之间的联系, 可以为各类分子识别的理论研究提供知识基础. 该数据库自 2007 年 11 月起公开发布(<http://www.pdbbind-cn.org>), 目前已经拥有来自世界 60 余国的超过 2700 名注册用户, 其中许多用户来自著名的大学、研究所以及医药公司. PDBbind-CN 数据库的早期版本只考虑 PDB 中由小分子配体与蛋白分子形成的复合物. 自 2008 年开始, 除了小分子配体-蛋白复合物, 我们也开始将蛋白-蛋白复合物、蛋白-核酸复合物等生物大分子形成的复合物包括在内, 从而形成了对 PDB 内容的完整覆盖. 另外, 我们坚持每年对 PDBbind-CN 数据库进行定期更新, 该数据库收录的数据规模因而随着 PDB 的增长而扩大. 最新的 2015 版中将收录 PDB 中约 14600 个复合物的亲合性实验数据(表 1). 关于 PDBbind-CN 数据库更多的详细情况, 读者可以参考我们近期发表的研究论文^[20].

为了构建 CASF 方法体系中所需的测试集, 我们以 PDBbind-CN 数据库中收录的蛋白-配体复合物为基础, 从中挑选高质量的样本. 所采用的主要挑选标

准为: (1) 对三维结构的质量要求. 复合物晶体结构的总体分辨率 $\leq 2.50 \text{ \AA}$, R 因子 ≤ 0.25 ; 配体分子结构不完整或者蛋白结合口袋附近残基缺失的复合物不予考虑; 共价结合的复合物不予考虑; 配体与蛋白分子之间存在明显碰撞(重原子间距离 $< 2.0 \text{ \AA}$)的复合物不予考虑. (2) 对亲合性数据的质量要求. 仅考虑亲合性数据为解离常数(K_d)或抑制常数(K_i)的复合物, 亲合性数据为半数抑制浓度(IC_{50})的复合物不予考虑; 亲合性数据极高($K_d/K_i < 1 \text{ pmol/L}$)或极低($K_d/K_i > 10 \text{ mmol/L}$)的复合物不予考虑; 亲合性数据不是确定值的复合物(如 $K_i > 10 \text{ }\mu\text{mol/L}$ 或 $K_i \approx 1 \text{ pmol/L}$)不予考虑; 用于亲合性数据测试的蛋白分子必须与复合物晶体结构中呈现的蛋白分子保持一致(如生物来源、亚型、突变、剪裁修饰等); 同理, 用于亲合性数据测试的配体分子必须与复合物晶体结构中呈现的配体分子具有相同的结构. (3) 对复合物组成的要求. 仅考虑蛋白与配体分子以 1:1 比例结合的二元复合物; 为了方便使用分子模拟软件处理, 配体分子结构中不可以含有金属原子或非常见元素(如 B、Si、As), 且配体分子周围 $5 \text{ \AA}$ 以内不可以有任何非天然氨基酸残基; 配体分子的分子量必须小于 1000; 配体分子包埋面积百分比小于 15%的复合物不予考虑.

以上一套筛选标准大大高于他人早期工作中所采用的挑选样本的标准. 经以上标准筛选所获得的复合物, 即为“精选集”(表 1). 以 PDBbind 2013 版为例, 我们从 8300 余个蛋白-配体复合物中, 挑选出 2959 个复合物样本, 入选比例不足 36%. 该比例折射出我们在挑选样本中所执行的高标准.

但是“精选集”并不能直接拿来作为 CASF 方法体系中需要的测试集, 一是因为该集合样本数过多, 需要极大的计算量; 二是因为该集合中存在比较严重的样本冗余, 即某些特定蛋白(如常见的药物靶标)

表 1 PDBbind-CN 数据库各版本的基本情况

版本	PDB 中收录的结构总数	PDBbind 识别的复合物总数	已知亲合性实验数据的复合物总数	精选集	核心集
2007	40876	11822	3124	1300	195
2008	48092	18211	4300	1401	210
2009	55118	23284	5678	1741	219
2010	62387	26434	6772	2061	231
2011	70224	30259	7986	2455	216
2012	78235	34180	9308	2897	201
2013	87085	38918	10776	2959	195
2014	96592	44569	12995	3446	195
2015	105183	48821	14620	待定	待定

形成的复合物样本数大大高于其他蛋白形成的复合物。为了解决这个问题, 我们进一步对“精选集”中所有的复合物进行序列分析, 以蛋白序列相似度 $\geq 90\%$ 为阈值对其进行聚类。聚类的结果是每一类中通常包含由某种蛋白和配体分子组成的若干个复合物。然后我们从每一类中分别挑选出 3 个代表复合物, 它们分别具有该类中“最低”、“中等”和“最高”三个水平的亲合性数据。每两个复合物之间的亲合性数据, 其差异不得小于 10 倍。这样的挑选规则既控制了数据集的规模, 又最大限度地保存了样本的多样性。对于所有挑选出的代表性复合物, 还要进行晶体结构电子密度图的人工检查, 配体分子以及蛋白结合口袋附近残基的结构与电子密度匹配明显不合理的复合物予以剔除。经过层层挑选的最终结果即为“核心集”, 也就是 CASF 方法体系中所采用的测试集。

在本课题组^[18,19]最近发表的 CASF-2013 工作中, 所采用的测试集为 PDBbind 2013 版的核心集, 包含 65 类蛋白, 共 $65 \times 3 = 195$ 个蛋白-配体复合物。根据 DrugBank 等数据库中的信息, 其中有 33 类蛋白(51%)

是已有上市药物的成熟药物靶标, 另有 18 类蛋白(27%)是潜在药物靶标, 有若干候选药物正在不同阶段的临床试验中^[21]。另外, 我们计算了该数据集中配体分子的一些“类药性”描述符, 其中分子量、氢键给体数目、氢键受体数目和 $\log P$ 值的分布情况如图 1 所示。从图 1 中可以看出, 该数据集中绝大部分配体分子的分子量小于 500, 其他性质的分布也显示大部分配体属于“类药性”的小分子化合物。我们在构建该数据集时, 并没有刻意地限定合格复合物中的蛋白分子和配体分子必须具有“类药性”, 但实际上最终大部分样本满足了这一条件: 78% 的蛋白是成熟或潜在的药物靶标, 82% 的配体分子符合 Lipinski 的“五规则”^[22]。这表明该数据集用于测试计算药物-靶标亲合性的打分函数是非常合适的。

2.2 设计合理的评价方法

在他人完成的早期评价工作中, 对打分函数的性能缺乏清晰、系统的描述方法。我们考虑到打分函数主要有 3 种实际应用方式: (1) 用于分子对接, 预

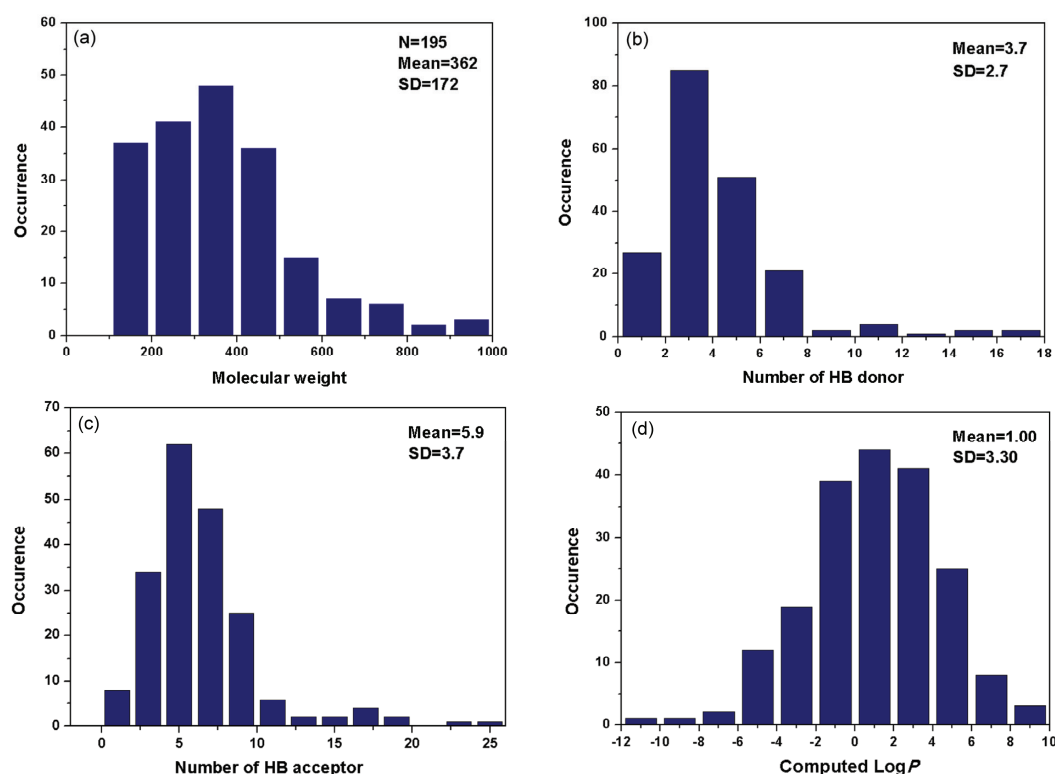


图 1 CASF-2013 所使用的测试集中包含有 195 种小分子配体。这些分子 4 个类药性描述符的分布情况。(a) 分子量; (b) 氢键给体数目; (c) 氢键受体数目; (d) $\log P$ 计算值。图中 Mean 表示平均值, SD 表示标准偏差^[18]

测蛋白与配体分子的结合模式; (2) 用于计算已知结构的蛋白-配体复合物的亲合性数据; (3) 用于虚拟筛选, 从化合物库中筛选出活性分子. 根据这 3 种应用方式, 本课题组^[18,19]设计了 4 种基本指标来衡量打分函数的性能, 分别为“打分能力”(scoring power)、“排序能力”(ranking power)、“对接能力”(docking power)和“筛选能力”(screening power), 并且为每种指标设计了相应的定量评价方法.

“打分能力”是指打分函数重现多样性的蛋白-配体复合物的亲合性实验数据的能力. 该指标可以通过计算亲合性打分数值与实验值的 Pearson 线性相关系数 R_p 和标准偏差 SD 来衡量. R_p 值越大, SD 值越小, 则表明打分函数的打分能力越好. 而“排序能力”是指打分函数对同一靶蛋白形成的复合物按照其亲合性高低进行正确排序的能力. 在我们设计的测试集中, 每一类靶蛋白都有亲合性强度在“低”、“中”和“高”三个水平的复合物. 如果按照打分函数的计算值可以给出与实验数据顺序一致的排序结果, 则认为打分函数在该靶蛋白上能够正确排序. 然后统计打分函数在测试集中所有靶蛋白上的成功率百分比, 用来表征其排序能力. “打分能力”与“排序能力”都用来考察打分函数的计算值与亲合性实验数据之间的相关性. 但是“排序能力”与“打分能力”不同, 该指标并不要求复合物的亲合性计算值与实验值之间形成线性相关, 只要求排序正确即可. 另外, “排序能力”只针对同一靶蛋白形成的复合物系列, 而“打分能力”则针对任意类型蛋白-配体复合物.

“对接能力”是指打分函数从配体分子的各种可能结合构象(即“假目标”)中识别出正确者(即晶体结构中配体分子的结合构象)的能力. 正如本文引言中所讨论的, 把构象采样过程和打分函数的计算过程分开, 才能获得对打分函数性能的准确评价, 这正是我们的 CASF 方法体系中的关键点之一. 我们的策略是采用多种分子对接软件(GOLD、Surflex 和 MOE-Dock)以及多种参数设置, 对于测试集中的每一个配体分子获得其尽可能全面的结合构象的集合, 然后从中挑选出至多 100 个代表性构象作为假目标集. 然后用打分函数来计算假目标集中的每个构象, 如果得分最高的构象与真实结合构象之间的 RMSD 值小于设定阈值(如 2.0 Å), 则认为打分函数在该复合物体系上预测成功. 然后统计打分函数在测试集中所有复合物上的成功率, 用来表征该打分函数的对接

能力.

“筛选能力”衡量打分函数从一组随机化合物中识别出真正能够与给定靶蛋白结合的配体分子的能力. 我们使用的测试集中包含 65 种靶蛋白, 每种靶蛋白对应 3 个已知亲合性数据的配体分子, 这 3 个配体分子就作为该靶蛋白的阳性样本, 测试集中其他 $195-3=192$ 个配体分子则默认为不能与该蛋白结合, 作为阴性样本. 然后考察对于每种靶蛋白, 打分函数是否可以从全部 195 个配体分子中识别出 3 个阳性样本. 为此需要产生全部 195 个配体分子在全部 65 种靶蛋白上的结合构象, 即交叉分子对接(cross-docking). 每个小分子配体与每种靶蛋白之间的结合构象集合也从 3 种分子对接软件的输出结果中挑选, 并通过聚类分析挑选出能量最低的若干构象作为代表. 对于受测试的每种打分函数, 该构象集合中打分最高的构象即作为其预测的结合构象, 而该构象的亲合性打分值则作为该分子预测的亲合性. 对于每种靶蛋白, 所有 195 个配体分子都按照亲合性的计算值从高到低排序, 就可以计算每种靶蛋白上的虚拟筛选富集因子(enrichment factor). 然后对所有 65 种靶蛋白的计算结果取平均值, 用来定量评价打分函数的筛选能力. 另外, 我们还统计每种靶蛋白上已知亲合性最高的阳性化合物排名位于 Top 1%、Top 5%和 Top 10%的成功率, 用来评价打分函数的筛选能力.

2.3 挑选流行的打分函数作为评价对象

目前存在的打分函数种类繁多, 显然无法全部纳入到我们的评价工作中. 考虑到在药物研发工作中的普及程度, 我们选择几种主流分子模拟商业软件中的打分函数作为评价对象. 它们分别是: 美国 Accelrys 公司出品的 Discovery Studio (v3.5)软件中的 5 种打分函数 LigScore^[23]、PLP^[24]、Jain^[25]、PMF^[26,27]和 LUDI^[28]; 美国 CERTARA 公司出品的 SYBYL (v8.1)软件中的 4 种打分函数 G-Score^[29]、PMF-Score^[26,27]、D-Score^[30]和 ChemScore^[31,32]; 德国 Schrödinger 公司出品的 Schrödinger (v2014)软件包中的 GlideScore-SP 和 GlideScore-XP^[33,34]; 英国 CCDC 公司出品的分子对接软件 GOLD (v5.2)中的 4 种打分函数 ASP^[35]、ChemScore^[31,32]、ChemPLP^[36]和 GoldScore^[29]; 加拿大 CCG 公司出品的 MOE (v2013)软件中的 4 种打分函数 London dG、ASE、Affinity dG 和 Alpha HB. 本课题组^[37]前期工作中发展的 X-Score 打分函数也被

纳入该评价工作中。

以上 20 种打分函数涵盖了基于分子力场的方法(GoldScore/G-Score、D-Score)、经验打分函数(LigScore、PLP、Jain、LUDI、ChemScore、GlideScore、ChemPLP、London dG、ASE、Affinity dG、Alpha HB、X-Score)以及基于统计势函数的模型(PMF/PMF-Score、ASP) 3 大类方法。另外,我们还引入了一个简单的打分函数作为评价体系中的参照模型。该模型采用配体分子与蛋白结合后的溶剂可及表面积的变化值(ΔSAS)作为唯一描述符,来衡量蛋白-配体相互作用的强弱。引入参照模型的意义在于:在任意一项测试中,如果某打分函数的表现甚至还不如这样一个最简单的模型,则认为它的表现不合格。这样的做法比人为规定一个指标(如规定成功率为 50%为及格线)要更加合理。

3 对若干流行打分函数进行评价的结果

我们的 CASF-2013 工作包括了 21 种打分函数作为测试对象。需要强调的是,我们该项工作的首要意义在于建立了比较完善的评价打分函数性能的方法

体系(即制定了竞赛规则)。同时,我们给出的若干流行打分函数的评价结果(即具体选手的竞赛成绩),对使用这些打分函数的广大用户具有明确的参考价值。本课题组曾于 2009 年发表了 CASF 方法体系对 16 种打分函数的评价结果(以下简称 CASF-2007)^[17];之后,于 2014 年再次发表了经过改进的 CASF 方法体系对 21 种打分函数的评价结果(以下简称 CASF-2013)^[18,19]。下文中,我们将结合这两次工作的评价结果进行分析和讨论。

3.1 “打分能力”的测试结果

CASF-2007 和 CASF-2013 使用了不同的测试集,但是对于打分能力的评价方法完全相同。从两次评价的结果来看,参与测试的各种打分函数的排序基本保持一致(表 2)。其中, X-Score^{HM} 的打分能力最好,其计算值与亲合性实验数据的相关系数超过 0.6,其后依次是 ChemScore@SYBYL、ChemPLP@GOLD、PLP1@DS 和 G-Score@SYBYL。另外一个有趣的发现是,除了 X-Score^{HM} 之外,没有一种打分函数在此项测试中可以优于作为参照模型的 ΔSAS 。这个发现

表 2 21 种打分函数在 CASF-2013 与 CASF-2007 中“打分能力”的测试结果

打分函数 ^{a)}	CASF-2013			CASF-2007		
	<i>N</i> ^{b)}	<i>R</i> ^{c)}	<i>SD</i> ^{d)}	<i>N</i>	<i>R</i>	<i>SD</i>
X-Score ^{HM}	195	0.614	1.78	195	0.644	1.83
ΔSAS	195	0.606	1.79	— ^{e)}	—	—
ChemScore@SYBYL	195	0.592	1.82	195	0.555	1.98
ChemPLP@GOLD	195	0.579	1.84	—	—	—
PLP1@DS	195	0.568	1.86	195	0.545	2.00
G-Score@SYBYL	195	0.558	1.87	195	0.492	2.08
ASP@GOLD	195	0.556	1.88	193	0.534	2.02
ASE@MOE	195	0.544	1.89	—	—	—
ChemScore@GOLD	189	0.536	1.90	178	0.441	2.15
D-Score@SYBYL	195	0.526	1.92	195	0.392	2.19
Alpha-HB@MOE	195	0.511	1.94	—	—	—
LUDI3@DS	195	0.487	1.97	195	0.487	2.09
GoldScore@GOLD	189	0.483	1.97	169	0.295	2.29
Affinity-dG@MOE	195	0.482	1.98	—	—	—
LigScore2@DS	190	0.456	2.02	193	0.464	2.12
GlideScore-SP	169	0.452	2.03	—	—	—
Jain@DS	191	0.408	2.05	189	0.316	2.24
PMF@DS	194	0.364	2.11	193	0.445	2.14
GlideScore-XP	164	0.277	2.18	178	0.457	2.14
London-dG@MOE	195	0.242	2.19	—	—	—
PMF@SYBYL	191	0.221	2.20	190	0.268	2.29

a) 各打分函数按照 CASF-2013 的评价结果从高到低排序; b) 每种打分函数能获得合理的计算值的复合物数目; c) 亲合性计算值和实验数据之间的 Pearson 相关系数; d) 亲合性计算值和实验数据线性拟合的标准偏差,单位为 $\log K_d$; e) CASF-2007 中未包括的打分函数

对发展打分函数的理论工作者提出了严峻的挑战.

3.2 “排序能力”的测试结果

对于“排序能力”的测试, CASF-2007 和 CASF-2013 也采用了完全相同的评价方法. 在 CASF-2013 中, 排序能力排名前五的打分函数分别是 X-Score^{HM}、ChemPLP@GOLD、PLP2@DS、GoldScore @GOLD 和 ChemScore@SYBYL (表 3). 这些排名靠前的打分函数在“打分能力”测试中也基本排在前列. 这样的结果是符合逻辑的: 打分能力好的打分函数也应该具有好的排序能力. 另外, 此项测试的结果显示, 有近半数的打分函数其排序能力超过参照模型 Δ SAS, 有 10 种打分函数的排序成功率超过 50%. 就“打分能力”和“排序能力”两个指标的设计而言, “打分能力”是比“排序能力”更高的要求, 这也得到了我们的测试结果的支持.

3.3 “对接能力”的测试结果

CASF-2013 中各打分函数“对接能力”的测试结果如图 2 所示. 我们定义, 当最佳得分构象与晶体结构中的配体分子结合构象之间的 RMSD 值小于 2.0 Å 时, 就认为打分函数在此复合物上的预测是成功的. 如果仅考虑得分最高的对接构象, 有近一半的打分函数的预测成功率超过 70%, 其中 ChemPLP@GOLD

表 3 21 种打分函数在 CASF-2013 与 CASF-2007 中“排序能力”的测试结果

打分函数 ^{a)}	排序成功率(%)	
	CASF-2013	CASF-2007
X-Score ^{HM}	58.5	58.5
ChemPLP@GOLD	58.5	— ^{b)}
PLP2@DS	55.4	53.8
GoldScore@GOLD	55.4	23.1
ChemScore@SYBYL	53.8	47.7
Affinity-dG@MOE	53.8	—
LigScore1@DS	52.3	35.4
Alpha-HB@MOE	52.3	—
G-Score@SYBYL	52.3	46.2
LUDI2@DS	52.3	43.1
D-Score@SYBYL	49.2	46.2
Δ SAS	49.2	—
ASP@GOLD	47.7	43.1
ChemScore@GOLD	46.2	36.9
PMF@DS	44.7	41.5
PMF@SYBYL	43.1	38.5
London-dG@MOE	43.1	—
Jain@DS	41.5	41.5
ASE@MOE	40.0	—
GlideScore-XP	35.4	33.8
GlideScore-SP	26.2	—

a) 各打分函数按照在 CASF-2013 中的评价结果从高到低排序; b) CASF-2007 中未包括的打分函数

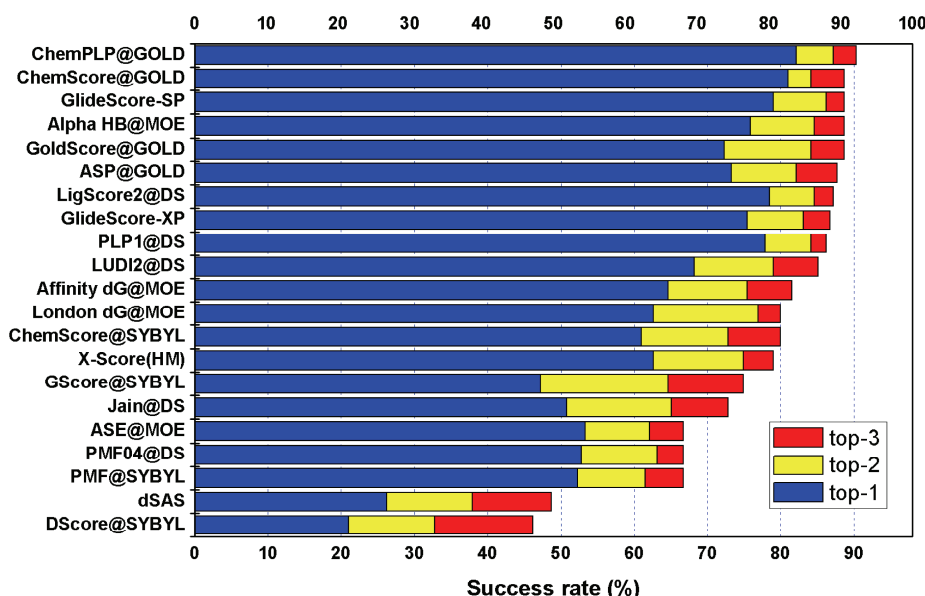


图 2 21 种打分函数在 CASF-2013 中“对接能力”的测试结果. 蓝色、黄色和红色条带分别表示考虑得分最高的一个、前两个或前三个对接构象情况下的预测成功率^[19]

的成功率甚至超过 80%。其后依次是 ChemScore@GOLD、GlideScore-SP、LigScore2@DS 和 PLP1@DS。显而易见, 相对于“打分能力”或“排序能力”而言, 现在的打分函数在“对接能力”方面更接近人们的期望值。这也解释了分子对接方法为何可以取得广泛的成功。从图 2 中也可以看出, 如果考察得分最高的前两个或前三个构象, 则各打分函数的成功率普遍会提高 5%~15%, 甚至更多。该发现说明, 在分子对接的实际应用中, 保留数个高得分的对接构象是一个推荐采用的策略。另外, 我们发现, 作为参照模型的 Δ SAS 在此项测试中的表现非常差, 说明仅凭分子间的接触面积大小无法鉴别出配体分子正确的结合构象。

3.4 “筛选能力”的测试结果

筛选能力是 CASF-2013 中新引入的评价指标, 主要采用富集因子来衡量打分函数将阳性化合物富集在化合物排序前端的能力。富集因子越高, 则表明打分函数在虚拟筛选中鉴别出阳性化合物的成功率越高。从此次评价结果来看(表 4), GlideScore-SP 在 Top 1% 水平的富集因子将近 20, 在所有打分函数中最高。其后依次是 ChemScore@GOLD、GlideScore-XP、LigScore2@DS 和 ChemPLP@GOLD, 它们的平均富集因子也在 10 以上。除了使用富集因子, 我们还采用在不同水平上筛选出最好阳性化合物的成功率来评价各打分函数的筛选能力。成功率排名前 5 名的打分函数分别是 GlideScore-SP、GlideScore-XP、ChemScore@GOLD、LigScore2@DS 与 ChemPLP@GOLD, 与以富集因子为指标时排名前 5 名的打分函数完全一致。可见这两种指标均能正确反映打分函数的筛选能力。另外一个明显的规律是: “筛选能力”强的打分函数其“对接能力”也强, 这也是非常合乎逻辑的。最后, 作为参照模型的 Δ SAS 在此项测试中的表现最差, 说明仅考虑分子间的接触面积大小也无法完成筛选活性化合物的目标。

综合以上评价结果可以得出一个基本结论: 衡量打分函数性能的四项基本指标可以分为两组, 即“打分能力”和“排序能力”为一组, “对接能力”和“筛选能力”为一组。“打分能力”好的打分函数通常也具有较好的“排序能力”, 而“对接能力”好的打分函数通常也具有较好的“筛选能力”。在实际应用中, 用户可以选择对接能力/筛选能力好的打分函数(如 GlideScore-

表 4 21 种打分函数在 CASF-2013 中“筛选能力”的测试结果^[19]

打分函数 ^{a)}	不同水平下的平均富集因子		
	Top 1%	Top 5%	Top 10%
GlideScore-SP	19.54	6.27	4.14
ChemScore@GOLD	18.90	6.83	4.08
GlideScore-XP	16.81	6.02	4.07
LigScore2@DS	15.90	6.23	3.51
ChemPLP@GOLD	14.28	5.88	4.31
LUDI1@DS	12.53	4.28	2.80
ASP@GOLD	12.36	6.23	3.79
Affinity-dG@MOE	8.21	4.15	3.19
London-dG@MOE	8.08	3.36	2.51
GoldScore@GOLD	7.95	4.52	3.16
PLP1@DS	6.92	4.28	3.04
Jain@DS	5.90	2.51	1.80
PMF@SYBYL	5.38	2.21	1.90
ChemScore@SYBYL	5.26	2.38	2.18
Alpha-HB@MOE	4.87	3.23	1.32
PMF04@DS	4.87	2.87	2.63
ASE@MOE	4.36	2.35	1.59
X-Score ^{HM}	2.31	2.14	1.41
D-Score@SYBYL	2.31	1.79	1.46
G-Score@SYBYL	1.92	1.26	1.44
Δ SAS	1.41	1.28	1.12

a) 所有打分函数按照在 Top 1% 水平下的富集因子值从高到低排序

SP)用于产生配体分子的对接构象(docking), 然后采用打分能力/排序能力好的打分函数(如 X-Score^{HM})进行重新排序(re-ranking)。这种分级联用策略可以帮助人们提高发现活性先导化合物的成功率。事实上许多公开报道的虚拟筛选成功案例已采用了这种策略, 我们得出的评价结果为这种策略提供了支持证据。另外, 我们的评价结果也显示, 某些打分函数表现出良好的综合性能, 其几个基本指标都表现良好, 如 ChemPLP@GOLD; 与之相反, 有些打分函数在各个基本指标上均表现不佳, 属于设计失败的作品。用户在药物研发的实际应用中选择打分函数时, 可以参考我们 CASF 方法体系给出的评价结果。

4 CASF 方法体系的学术影响

与前人完成的打分函数评价工作相比, 我们所设计的 CASF 方法体系具有较强的先进性和系统性, 其思路得到了同行的广泛认可。2009 年本课题组^[17]在 *Journal of Chemical Information and Modeling* 杂志上发表了 CASF-2007 工作。该论文的评审人认为:

“该项工作在评价方案的设计、测试集的选取、参与评价的打分函数多样性几个方面均超越了前人的工作,是目前为止关于打分函数评价最系统的工作.....该工作中建立的方法显然将成为其他工作所仿效的标准”. 该论文自 2009 年发表以来已被引用 140 余次,并且被国内外同行多次应用.

例如,欧洲生物信息学研究中心的 Ballester 等^[38,39]采用 PDBbind 精选集(v2007)为训练集,以蛋白-配体原子作用对为描述符,使用随机森林分析方法构建了 RF-Score 打分函数. 通过与 CASF-2007 提供的评价结果对比,认为 RF-Score 具有比较好的打分能力. 同时, Ballester 等这样评价我们的工作:“CASF 方法体系是评价打分函数事实上的标准”. 诺华公司的 Kramer 等^[40]利用 CASF-2007 中提供的测试集,提出了检验打分函数性能的“Leave-cluster-out cross-validation”新策略. 他们认为,只有在 PDBbind 数据集上表现良好的打分函数,才有可能在一个结构多样性的测试集上获得可靠的结果. 德国马尔堡大学的 Klebe 课题组^[41]使用剑桥晶体结构数据库(CSD)中的信息为基础,发展了基于统计势的打分函数 DSX. 他们使用 CASF-2007 体系测试了 DSX 的打分能力、排序能力以及对接能力,认为 DSX 取得了很好的对接能力以及较好的打分能力和排序能力. 四川大学杨胜勇课题组^[42]开发了基于多种分子描述符的打分函数 ID-Score, 采用 PDBbind 精选集(v2011)作为训练集,通过支持向量机回归方法挑选出关键的分子描述符,构建打分函数模型. 他们采用 CASF-2007 体系对 ID-Score 进行了测试,认为其打分能力明显优于其他打分函数. Wang 等^[43]在 AutoDock4 分子对接软件中采用的基于分子力场的打分函数的基础上,引入量子化学计算得出的部分电荷数值,在 CASF-2007 体系上进行了测试,认为新的模型与 AutoDock4 中的打分函数相比,在对接能力以及打分能力方面均有提高.

另外,我们在前文中曾提到,目前存在的打分函数种类繁多,在我们的工作中不可能包括数目太多的打分函数作为评价对象. 为了推广 CASF 方法体系,我们采用了开放式的“server-client”策略. 具体而言,

本课题组的角色即为“server”,负责维护和发展 CASF 方法体系,其他应用 CASF 方法体系的课题组即为“client”. CASF 方法体系中用到的所有原始数据文件,包括测试集中所有复合物的三维结构、亲合性实验数据以及每对配体分子-靶标蛋白的对接构象集合,均可从 PDBbind-CN 数据库的网站下载(<http://www.pdbbind-cn.org>). 其他研究打分函数的课题组可以下载 CASF 的完整数据包,遵循我们的评价方法对他们自己手头的打分函数进行评价. 由于采用相同的测试集,遵循统一的评价方法,这样获得的结果不仅可以与我们公开发表的结果进行横向比较,不同课题组之间也可以互相进行比较. 这种策略有效地减少了打分函数研究领域以往存在的“自说自话”的现象,是我们的工作受到同行欢迎的一个重要原因. 我们调研了 *Journal of Chemical Information and Modeling* 杂志在 2010~2014 年间发表的所有 1400 余篇论文,共计报道了 30 项新的打分函数,而其中有 16 项工作(53%)应用了我们 CASF 方法体系或 PDBbind-CN 数据库提供的数据集. 这个比例体现了我们的工作 in 打分函数研究领域形成的广泛影响力.

5 结论

药物-靶标亲合性打分函数在分子靶向药物研发中发挥着重要的作用. 经过 20 多年的发展,目前存在类型繁多的打分函数,其中难免良莠不齐. 建立客观公正的评价体系不仅可以帮助打分函数的用户做出合理的选择,也为发展打分函数的研究人员提供参考依据,是打分函数研究领域中的一个基础问题. 我们建立的 CASF 方法体系为解决这个问题提供了一个方案,受到了同行的广泛认可. 未来,我们将进一步完善我们的评价方法,不断提高测试集的规模和质量,并尽可能覆盖近年来出现的新类型打分函数,不定期发布新的评价结果. 同时,我们也在着手为适用于计算蛋白-蛋白和蛋白-核酸相互作用的打分函数或分子对接方法建立类似的评价体系,进一步拓展我们的工作.

参考文献

- 1 Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular

- receptors. *Angew Chem Int Ed*, 2002, 41: 2644–2676
- 2 Gilson MK, Zhou HX. Calculation of protein-ligand binding affinities. *Annu Rev Biophys Biomol Struct*, 2007, 36: 21–42
 - 3 Schulz-Gasch T, Stahl M. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discov Today Tech*, 2004, 1: 231–239
 - 4 Rajamani R, Good AC. Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development. *Curr Opin Drug Discov Develop*, 2007, 10: 308–315
 - 5 Liu J, Wang RX. Classification of current scoring functions. *J Chem Inf Model*, 2015, 55: 475–482
 - 6 Bissantz C, Folkers G, Rognan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem*, 2000, 43: 4759–4767
 - 7 Bursulaya B, Totrov M, Abagyan R, Brooks C. Comparative study of several algorithms for flexible ligand docking. *J Comput-Aided Mol Des*, 2003, 17: 755–763
 - 8 Chen H, Lyne PD, Giordanetto F, Lovell T, Li J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J Chem Inf Model*, 2006, 46: 401–415
 - 9 Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. *J Med Chem*, 2006, 49: 5912–5931
 - 10 Xing L, Hodgkin E, Liu Q, Sedlock D. Evaluation and application of multiple scoring functions for a virtual screening experiment. *J Comput-Aided Mol Des*, 2004, 18: 333–344
 - 11 Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL. Assessing scoring functions for protein-ligand interactions. *J Med Chem*, 2004, 47: 3032–3047
 - 12 Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem*, 2003, 46: 2287–2303
 - 13 Berman HM, Henrick K, Nakamura H. Announcing the worldwide protein data bank. *Nat Struct Biol*, 2003, 10: 98
 - 14 Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem*, 2007, 50: 726–741
 - 15 Warren GL, Do TD, Kelley BP, Nicholls A, Warren SD. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discov Today*, 2012, 17: 1270–1281
 - 16 Søndergaard CR, Garrett AE, Carstensen T, Pollastri G, Nielsen JE. Structural artifacts in protein-ligand X-ray structures: implications for the development of docking scoring functions. *J Med Chem*, 2009, 52: 5673–5684
 - 17 Cheng T, Li X, Li Y, Liu Z, Wang R. Comparative assessment of scoring functions on a diverse test set. *J Chem Inf Model*, 2009, 49: 1079–1093
 - 18 Li Y, Liu Z, Li J, Han L, Liu J, Zhao Z, Wang R. Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J Chem Inf Model*, 2014, 54: 1700–1716
 - 19 Li Y, Han L, Liu Z, Wang R. Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. *J Chem Inf Model*, 2014, 54: 1717–1736
 - 20 Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, Nie W, Liu Y, Wang R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 2015, 31: 405–412
 - 21 Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. DrugBank 3.0: a comprehensive resource for “omics” research on drugs. *Nucleic Acids Res*, 2011, 39: D1035–D1041
 - 22 Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliver Rev*, 1997, 23: 3–25
 - 23 Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M. LigScore: a novel scoring function for predicting binding affinities. *J Mol Graph Model*, 2005, 23: 395–407
 - 24 Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DR, Fogel LJ, Freer ST. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem Biol*, 1995, 2: 317–324
 - 25 Jain AN. Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J Comput-Aided Mol Des*, 1996, 10: 427–440
 - 26 Muegge I. Effect of ligand volume correction on PMF scoring. *J Comput Chem*, 2001, 22: 418–425
 - 27 Muegge I. A knowledge-based scoring function for protein-ligand interactions: probing the reference state. *Perspect Drug Discov Des*, 2000, 20: 99–114

- 28 Böhm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput-Aided Mol Des*, 1994, 8: 243–256
- 29 Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, 1997, 267: 727–748
- 30 Ewing TJA, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput-Aided Mol Des*, 2001, 15: 411–428
- 31 Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput-Aided Mol Des*, 1997, 11: 425–445
- 32 Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD. Flexible docking using tabu search and an empirical estimate of binding affinity. *Proteins: Struct, Funct, Genet*, 1998, 33: 367–382
- 33 Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem*, 2006, 49: 6177–6196
- 34 Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*, 2004, 47: 1739–1749
- 35 Mooij WTM, Verdonk ML. General and targeted statistical potentials for protein-ligand interactions. *Proteins: Struct, Funct, Bioinf*, 2005, 61: 272–287
- 36 Korb O, Stutzle T, Exner TE. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model*, 2009, 49: 84–96
- 37 Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput-Aided Mol Des*, 2002, 16: 11–26
- 38 Li H, Leung KS, Wong MH, Ballester PJ. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: cyscore as a case study. *BMC Bioinformatics*, 2014, 15: 291
- 39 Ballester PJ, Schreyer A, Blundell TL. Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *J Chem Inf Model*, 2014, 54: 944–955
- 40 Kramer C, Gedeck P. Leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets. *J Chem Inf Model*, 2010, 50: 1961–1969
- 41 Neudert G, Klebe G. DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes. *J Chem Inf Model*, 2011, 51: 2731–2745
- 42 Li GB, Yang LL, Wang WJ, Li LL, Yang SY. Comprehensive set of descriptors related to protein-ligand interactions. *J Chem Inf Model*, 2013, 53: 592–600
- 43 Wang JC, Lin JH, Chen CM, Perryman AL, Olson AJ. Robust scoring functions for protein-ligand interactions with quantum chemical charge models. *J Chem Inf Model*, 2011, 51: 2528–2537

Development of a high-quality benchmark for scoring function assessment

Zhihai Liu, Jie Li, Li Han, Yan Li^{*}, Renxiao Wang^{*}

State Key Laboratory of Bioorganic and Natural Products Chemistry; Collaborative Innovation Center of Chemistry for Life Sciences; Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China

^{*}Corresponding authors (email: kathyli@mail.sioc.ac.cn; wangrx@mail.sioc.ac.cn)

Abstract: Quantitative computation of drug-target binding affinity has been one of the key problems in molecular-targeted drug design. Scoring functions are fast, robust, and fairly accurate and thus are widely applied in drug design. Not surprisingly, a large number of scoring functions have been developed over the past twenty years. The performance of these methods is certainly not equal. An objective benchmark for assessing scoring functions is thus desired by the users of scoring function for making wise choice among available options. It is also desired by the developers of scoring function for obtaining useful guidance. In recent years, we have established a benchmark, namely Comparative Assessment Scoring Function (CASF), to tackle this critical problem. We proposed a set of indicators for measuring the performance of scoring function and developed evaluation methods accordingly. Very importantly, the scoring process was separated from the conformational sampling process in order to evaluate scoring functions on a common ground. The test set used in CASF, which consisted of non-redundant protein-ligand complexes with high-quality crystal structure and reliable binding data, was carefully selected from the PDBbind-CN database. Compare to the similar benchmarks published by other researchers, our CASF benchmark is more advanced, which has been applied by other researchers since published. We describe the basic methods, evaluation results, and international impact of our CASF benchmark in this review article.

Keywords: drug-target binding affinity, scoring function, benchmark, CASF, PDBbind-CN, molecular-targeted drug design