doi:10.3969/j.issn.1672-5565.2014.03.01

基于 pose 共享的蛋白质-配体构象并行搜索算法

杨 伟[1],吕 强[2*]
(1.苏州大学计算机科学与技术学院,江苏 苏州 215006;
2.苏州大学江苏省计算机信息处理技术重点实验室,江苏 苏州 215006)

摘 要:RosettaLigand 使用多次启动对接协议的方式对蛋白质-配体复合物构象空间进行采样,在串行或并行的构象搜索实例之间并不共享采样信息。因此并行对接与串行对接相比仅仅是增加了对接的速度,并不能改善对接的性能。我们对 Rosetta 3.4 版中的 RosettaLigand 算法进行了修改,在并行的对接实例之间共享采样信息,以实现多个对接实例协同优化采样进程。在一个包含 11 个目标的测试集合上进行的测试表明,共享采样信息在大多数对接实验中显著地提高了近天然构象在候选结构集合中的比例,同时还降低了整个候选结构集合的平均能量。

关键词:分子对接;构象搜索;pose 共享;RosettaLigand
中图分类号:TH133; TP183 **文献标志码**:A **文章编号**:1672-5565(2014)-03-157-05

Protein-Ligand conformation parallel search method based on pose sharing

YANG Wei[1], LÜ Qiang[2*]
(1.*Department of Computer Science and Technology, Soochow University, Suzhou* 215006, *China*;
2.*Jiangsu Provincial Key Laboratory for Information Processing Technology, Soochow University, Suzhou* 215006, *China*)

Abstract: RosettaLigand samples the protein-ligand complex conformation space using multi-restart docking protocol. The sampling information is not shared between different docking instances. Thus parallel docking is just faster but not necessarily better than the serial version. We modified the original RoesttaLigand algorithm in Rosetta v3.4, making the sampling information shared between all docking instances to improve the sampling process. Test results on a 11-target simulation set showed that sharing sampling could increase the proportion of natural-like structures in candidate structures and lower the average energy of candidate structures in most cases.

Keywords: Molecular docking; Conformation search; Pose sharing; RosettaLigand

随着 X 射线衍射以及 NMR 等技术的发展,越来越多的蛋白质的三维结构被测定出来,使基于受体结构的计算机辅助药物设计更具现实意义。分子对接(Molecular docking)方法是基于受体结构的药物设计(Structure-based drug design,SBDD)中的重要方法之一。分子对接是指两个或多个分子通过几何匹配和能量匹配相互识别的过程,通过分子对接可以从已有药物分子库筛选出有希望的先导化合物,避免了繁琐的化学合成实验过程,缩短了药物研发周期,降低了研发成本[1]。分子对接还被用于重新评估已知药物,发现已知药物新的适应症[2]。近年来,科研人员把注意力转向天然药物,希望从中药中开发出更为安全有效的新药,中药成分复杂,分子对接成为科学阐述中药药效物质基础和作用机理成为中药研究的重要工具[3]。分子对接操作就是寻找配体与受体结合在受体活性位点处的低能构象的过程。对接算法的性能依赖于两个因素:能量函数的精度和复合物构象搜索算法的性能。分子对接所涉及的搜索空间非常巨大,即便对柔性小分子,粗略估计其搜索空间至少含 10^{30} 个解,要从中找出低能构象必须借助于各种优化算法。目前,已经有许多优化算法用来解决分子对接问题,典型的模拟退

收稿日期:2014-04-01;**修回日期**:2014-04-11.
基金项目:国家自然科学基金(61170125)资助。
作者简介:杨伟,男,硕士研究生,研究方向:生物信息计算;E-mail:20104227043@suda.edu.cn.
*通信作者:吕强,男,教授,博士生导师;研究方向:生物信息计算、元启发搜索、并行分布计算;E-mail:qiang@suda.edu.cn.

火算法、遗传算法、禁忌搜索算法、蒙特卡罗方法以及这些算法的各种修正变种[4-6]。

常用的分子对接模拟软件有 AUTODOCK[10]，RosettaLigand[7]，GLIDE[12]，DOCK[13]，FlexX[14]，GOLD[15]等。Rosetta 是华盛顿大学开发的开源生物大分子建模软件包，其中包含用于对蛋白质和核酸进行结构预测、设计和重建模的工具，在 Rosetta 中，分子的结构用 pose 数据结构表示[4]。RosettaLigand 是其中利用模拟退火算法进行蛋白质-配体柔性对接的程序工具。本文的研究基于 Rosetta v3.4 版，下文中提及 RosettaLigand 处均指 Rosetta v3.4 中的 RosettaLigand 对接程序。Rosetta 源代码可通过 http://www.rosettacommons.org/网站获取。

1 材料与方法

RosettaLigand 的对接算法流程如图 1 所示[6]。

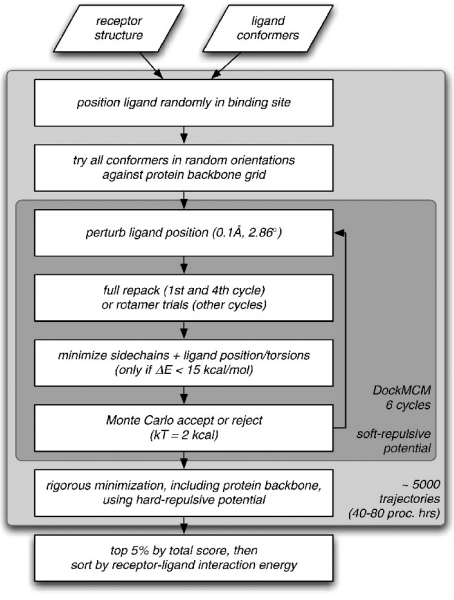

图 1 RosettaLigand 对接算法

Fig.1 Original dock algorithm in RosettaLigand

在 Rosetta 框架内，RosettaLigand 对接协议被重复启动 N 次(串行或并行)，生成指定数量的侯选结构，称为 decoys。在这次 N 次轨迹中，RosettaLigand 对接实例之间并不共享采样信息，彼此之间完全独立。我们认为在多次对接实例之间共享 pose 信息可以帮助 RosettaLigand 更好的对接蛋白质与配体。RosettaLigand 搜索蛋白质-配体复合物构象的过程本质是在一个在 Rosetta 全原子能量函数的指导下，在能量地形图上搜索的过程。由于蛋白质-配体复合物自由度非常大，而且已知能量函数并不足够精确，所以不存在全局最优结构的精确表达式，通常对接实验都会生成数目巨大的侯选结构，以期能够采样到尽可能多的近天然结构，然后通过基于能量若者机器学习的挑选方法将近天然结构挑选出来[9]。所以对接程序的搜索构象的能力依赖两个因素，一是搜索的广度，二是搜索的深度和充分性。RosettaLigand 通过运行多次对接轨迹来解决搜索的广度问题，我们现在希望通过在多个轨迹之间共享采样信息来改善 RosettaLigand 的构象采样的深度和充分性问题，如图 2a，每一个点代表一次对接过程，在不共享采样信息的情况下，每个对接实例只能在其附近空间进行不充分的构象搜索，而在图 2b 中，大量的搜索实例某个进行吸引到构象空间最能较低处，并对此处进行充分的采样。

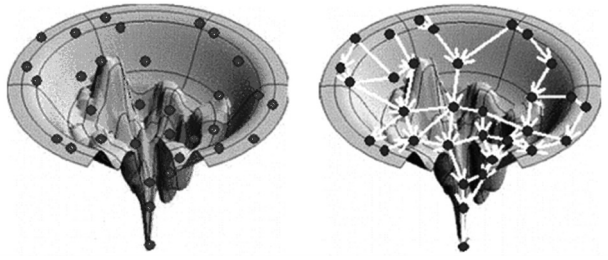

图 2 共享与非共享采样信息搜索示意图

Fig.2 Searching pattern: sampling sharing vs non-sharing

RosettaLigand 在对接的 DockMCM 阶段使用 MonteCarlo 判断接受或拒绝当前采样。在经过修改的 RosettaLigand 算法中使用 MonteCarlo 判断接受或拒绝采样时，首先将其他对接实例以文件形式共享的最佳采样结果读入，与当前采样结果进行比较，而不是与 RosttaLigand 内置算法中实现的与本实例已采样的最佳结果比较。经过修改的 DockCMC 算法流程如图 3 所示：

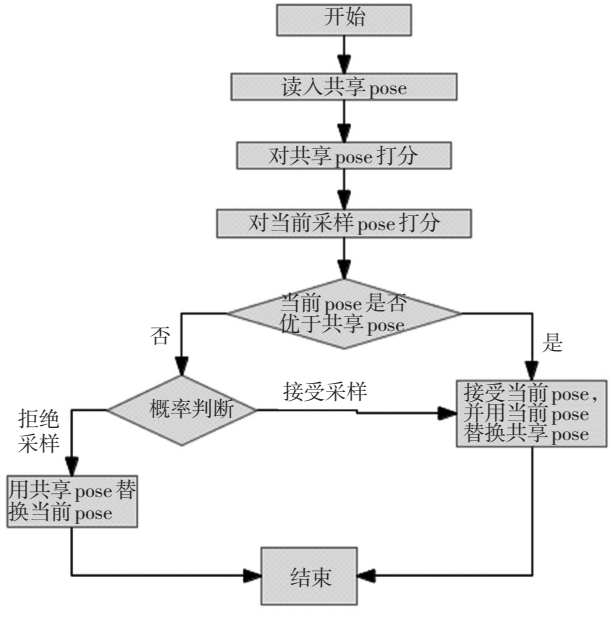

图 3 RosettaLigand DockCMC 的改进流程

Fig.3 Modification to original RosettaLigand DockMCM

RosettaLigand 原始对接算法在生成一个候选结构的采样过程中保有一个 last_accepted_pose 对象,用于存储上一次模拟退火接受的采样结果。在每一次新的采样后,将当前采样结果与此 last_accepted_pose 比较:

如果当前采样结果优于 last_accepted_pose,则接受此次采样,并用此采样结果替换 last_accepted_pose 对象;如果如果当前采样结果不如 last_accepted_pose,则按概率接受或拒绝当前采样结果,接受时同样用当前采样结果替换 last_accepted_pose,拒绝时用 last_accepted_pose 做为下次采样的起点。

在共享 pose 的 RosettaLigand 对接算法中,同样会保有一个 last_accepted_pose 对象,与原始的 RosettaLigand 算法不同,在每一次新的采样结束后,采样结果不是与上一次接受的采样结果比较,而是与共享的采样结果进行比较:如果当前采样被接受,则用当前采样结果替换共享的采样结果;如果当前采样被拒绝,则用共享的采样结果做为下一次采样的起点。

修改过的采样接受判定算法如下:

```
1:Input:current_pose, last_accpeted_pose,
score_function
2:Output:None
3:read in shared pose
4:score1=score_function(cureent_pose)
5:score2=score_function(last_accepted_pose)
6:boltz_factor=score2-score1
7:probability=exp(min(40.0,max(-40.0,
boltz_factor)))*proposal_density_ratio
8:if(probability<1)
9:if(uniform_rand()>=probability)
10:current_pose=last_accepted_pose
11:end if
12:end if
13:last_accepted_pose=current_pose
```

本文从 meiler 数据集[8]中选择 11 个自对接的算例进行实验,这 11 个目标是 1AQ1,1DBJ,1DM2,1NJA,1NJE,1PB9,1PBQ,1Y1M,2AYR,2PRG 和 4TIM。

配体的异构体库使用 OMEGA2[11]生成,实验组为 16 个并行进程以共享采样信息方式生成5 000 个候选结构,对照组使用 Rosetta 平台的内置的 MPI 框架使用 16 个进程生成 5 000 个候选结构。计算平台为 16 核 SMP Linux 集群。

2　实验结果

图 4 所示为 11 个目标的候选结构中 Ligand-RMSD 落在 0~1 和 1~2 两个区间的累积直方图。从图中可以看出,1AQ1、1DBJ、1PBQ、1Y1M、2AYR 这 5 个目标,共享 pose 算法极大的提高了侯选结构集合中 Ligand-RMSD 小于 2.0 的比例(Ligand-RMSD 小于 2.0 的构象被认为是近天然构象);1NJA 和 1NJE 这两个目标,共享 pose 算法和 RosettaLigand 内置算法表现都很差;而 1DM2、1PB9、2PRG、4TIM 这 4 个目标,虽然共享 pose 算法不如 RosettaLigand 算法,但性能相近。表 1统计了 11 个目标的候选结构中 Ligand-RMSD 小于 2.0 的构象的数量、构象最低能量以及实验运行时间。从表中可以看出,除了目标 1Y1M,对于其他所有目标,共享 pose 算法都采样到了比对照组更低的能量,而计算时间平均仅增加了不到 10%。

表 1　对照组与实验组候选结构集合中 Ligand-RMSD 小于 2.0 的数量、最低能量值以及运行时间统计

Table 1　Comparison of number of decoys with lrmsd<2.0, lowest energy and duration between experimental and control group

PDB	lrmsd<2		最低能量		运行时间	
1AQ1	60	2 539	-839.6	-866.3	20 h 35 min	22 h 19 min
1DBJ	60	355	-57.8	-83.5	13 h 10 min	15 h 4 min
1DM2	127	49	-795.7	-814.0	17 h 44 min	17 h 52 min
1NJA	0	0	-880.9	-891.5	24 h 46 min	26 h 23 min
1NJE	7	0	-868.9	-875.4	23 h 14 min	25 h 5 min
1PN9	445	185	-581.0	-634.6	15 h 37 min	18 h 37 min
1PBQ	139	1 210	-682.7	-754.7	16 h 15 min	19 h 12 min
1Y1M	478	3 256	-843.7	-581.0	17 h 14 min	19 h 19 min
2AYR	145	4 654	-857.4	-880.7	23 h 17 min	23 h 51 min
2PRG	17	7	-880.2	-892.4	23 h 13 min	24 h 12 min
4TIM	52	31	-775.2	-791.0	18 h 52 min	20 h 4 min

注:深色为实验组。

Notes:Dark as the experimental group.

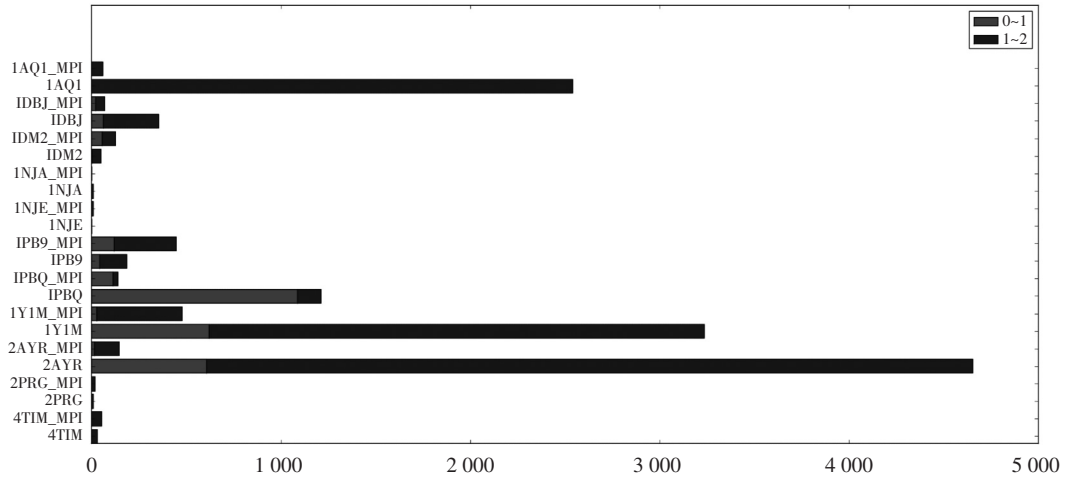

图 4 候选结构集合中 Ligand-RMSD 落在 0~1 和 1~2 两个区间的累积直方图

Fig.4 Stacked histogram of decoys with lrmsd fallen in interval 0~1 and 1~2

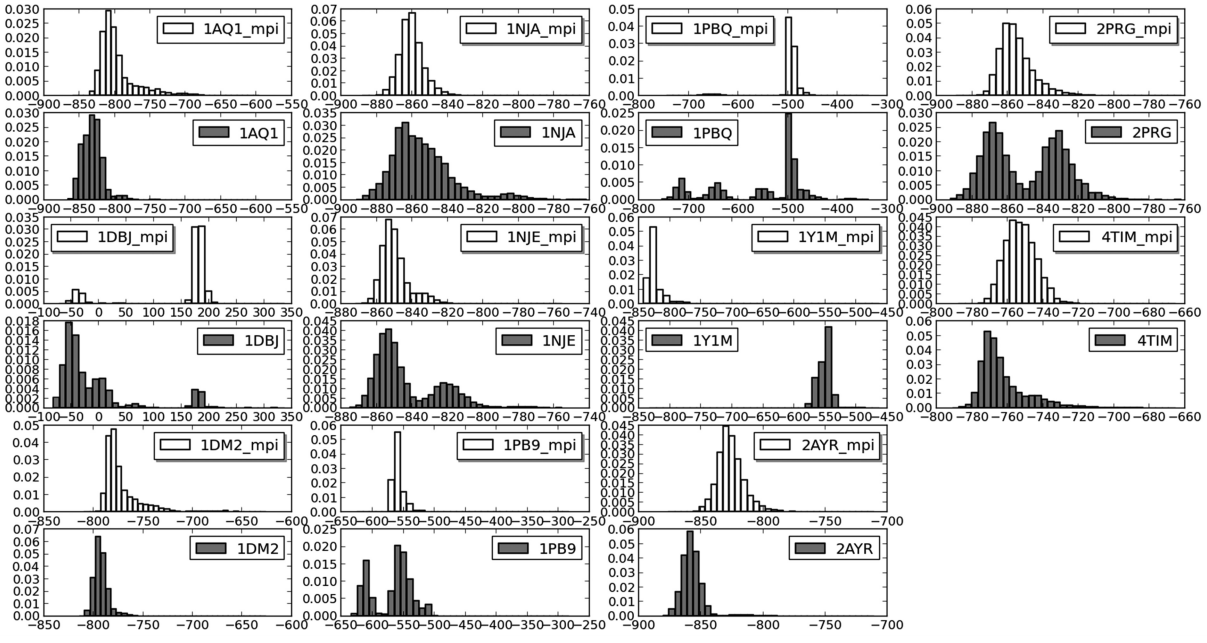

图 5 11 个目标候选结构总体能量值分布直方图

Fig.5 Histogram of Rosetta energy of all decoys of all 11 targets

注:白色为实验组,灰色为对照组。

Notes: white for experimental group, grey for control group.

图 5 所示为 11 个目标的 5 000 个侯选结构的能量分布的直方图。从图中可以看出,除 1Y1M 外,对大多数目标来说,与 RosettaLigand 内置算法相比,共享 pose 算法使降低了侯选结构整体的平均能量。

3 结 论

RosettaLigand 以多次随机启动方式来解决构象搜索的广度问题,此方式类似于在崎岖的能量地形图上随机分布起始点,并在每一个起始点附近使用模拟退火进行构象搜索,搜索的性能依赖随机启动的次数与随机程度。我们的方法在 N 个并行的对接实例之间共享采样信息,当某个进程采样到一个近天然构象时,其他的对接实例会被吸引到全局最优点附近,并对该处进行充分采样,这就是在目标 1AQ1、1DBJ、1PBQ、1Y1M、2AYR 中看到的情况,近天然构象的数量远远超过对照组。并且,由于每个对接实例在每一次采样后都与会 N 个实例当前所采样到的能量最低的构象进行比较,所以绝大多数实验组生成的候选结构集合的平均能量低于对照组的平均能量。当然,我们也看到,目标 1DM2、1PB9、2PRG、4TIM 的实验组采样到的近天然构象数量低

于对照组,分析其原因是共享信息一方面强化了并行对接实例在能量地形图上某处的采样能力,但它同时另一方面也使得陷入局部最小的问题更加严重,原本在独立采样的情况下可能采样到近天然结构的实例可能被误导到非近天然区域。共享采样信息是有用的,但要避免陷入局部最小。如同样启动 N 个并行对接实例,在每个对接实例采样都收敛之后,输出采样结果,再比较当前 N 个实例采样所得结果,以能量最低的结果当作 N 个并行对接实例下一轮对接的起始结构,以此循环。这样在保证搜索广度的同时,又可兼顾搜索的深度。更进一步的,在共享整体 pose 的前提下,搜索的广度不可能超出非共享条件下搜索的广度,原因是无论如何选择共享的时机与策略,所共享的采样信息都是当前 N 个并行对接实例可能采样到的构象。所以,为了拓宽搜索的广度,以共享采样信息为基础,结合并行对接实例自身的采样结果来构造超出当前所有对接实例本次对接轨迹可能覆盖的构象空间的构象是一条可行的道路。

参考文献(References)

[1] KAR S, ROY K. How far can virtual screening take us in drug discovery[J]. Expert Opinion on Drug Discovery, 2013, 8(3):245-261.

[2] MA D L, CHAN D S H, LEUNG C H. Drug repositioning by structure-based virtual screening[J]. Chemical Society Reviews, 2013, 42(5): 2130-2141.

[3] 任洁,魏静.分子对接技术在中药研究中的应用[J].中国中医药信息杂志 ,2014,(1):123-125.
REN Jie, WEI Jing. Application of molecular docking techniques in Chinese herbal medicine [J]. Chinese Journal of Information on Traditional Chinese Medicine, 2014,(1):123-125.

[4] KAUFMANN K W, LEMMON G H, DELUCA S L, et al. Practically useful: What the Rosetta protein modeling suite can do for you [J].Biochemistry, 2006, 49: 2987-2998.

[5] 刘敏,曾涛,徐开阔,等.一种基于免疫遗传算法的分子对接构象搜索策略[J].计算机研究与发展,2009,46(z2):597-601.
LIU Min, ZENG Tao, XU Kaikuo, et al. A molecular docking's conformational search strategy based on immune genetic algorithm [J].Journal of Computer Research and Development, 2009,46(z2):597-601.

[6] 李纯莲,王希诚,赵金城,等.药物分子对接中的构象搜索策略[J].计算机与应用化学,2004,21(2):201-205.
LI Chunlian, WANG Xicheng, ZHAO Jincheng,et al. Drug molecular docking design based on optimal conformation search [J]. Computers and Applied Chemistry,2004,21(2): 201-205.

[7] DAVIS I W, BAKER D. RosettaLigand docking with full ligand and receptor flexibility [J]. Journal of molecular biology, 2009, 385: 381-392.

[8] MEILER J, Baker D. ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility [J]. Proteins, 2006, 65, 538-584.

[9] 杨凌云, 吕强. 一种基于 SVR 的分辨近天然 GPCR —配体构象的方法[J].生物信息学,2011,9(2):167-170.
YANG Lingyu, LÜ Qiang. A SVR-Based method for indentifying near-native GPCR-Ligand confromation decoys [J]. China Journal of Bioinformatics, 2011, 09(2):167-170.

[10] MORRIS G M, GOODSELL D S, HALLIDAY D S, et al. Automated docking using a lamarckian genetic algorithm and and empirical binding free energy function [J]. J Comp Chem,1998,19:1639-1662.

[11] HAWKINS P C, SKILLMAN A G, WARREN G L, et al. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and cambridge structural database [J]. Journal of Chemical Information and Modeling, 2010,50 (4): 572-584.

[12] FRIESNER R A, MURPHY R B, REPASKY M P, et al. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-Ligand complexes [J]. J. Med. Chem., 2006, 49, 6177-6196.

[13] LANG P T, BROZELL S R, MUKHERJEE S, et al. DOCK 6: combining techniques to model RNA-small molecule complexes[J]. RNA ,2009,15(6): 1219-1230.

[14] RAREY M, KRAMER B, LENGAUER T, et al. A fast flexible docking method using an incremental construction algorithm[J]. J. Mol. Biol,1996, 261: 470-489.

[15] JONES G, WILLETT P, GLEN R C,et al. Development and validation of a genetic algorithm for flexible docking [J]. J. Mol. Biol., 1997, 267: 727-748.