

# 蛋白质-配体分子对接中构象搜索方法

常 珊<sup>1</sup> 陆旭峰<sup>1</sup> 王 峰<sup>2</sup>

(1. 江苏理工学院电气信息工程学院生物信息与医药工程研究所, 常州, 213001; 2. 常州大学信息科学与工程学院, 常州, 213164)

**摘 要:** 分子对接是研究蛋白质-配体分子间相互作用与识别的有效方法。分子间的相互作用过程中形成的近天然构象是结合自由能极低的构象, 快速且准确搜索能量极低的构象对于蛋白质-配体分子对接至关重要。本文回顾了蛋白质-配体分子对接中主要的构象搜索算法, 包括快速穷举搜索、启发式搜索和其他搜索方法, 并列出了采用不同搜索算法的代表性分子对接软件。其次, 介绍了蛋白质-配体分子对接的国际评估实验、常用的测试标准库和评价的重要指标。最后, 分析并指出了当前蛋白质-配体对接构象搜索方法所存在的主要问题, 并对未来的工作进行了展望。

**关键词:** 蛋白质-配体相互作用; 构象搜索方法; 分子对接

**中图分类号:** TP301      **文献标志码:** A

## Review of Conformational Searching Method for Protein-Ligand Molecular Docking

Chang Shan<sup>1</sup>, Lu Xufeng<sup>1</sup>, Wang Feng<sup>2</sup>

(1. Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou, 213001, China; 2. School of Information Science & Engineering, Changzhou University, Changzhou, 213164, China)

**Abstract:** Molecular docking technology is an effective approach to protein-ligand interactions and recognition. The near-native conformations generated in the interaction process are the lowest energy conformations, so it is pivotal for the protein-ligand docking to searching these conformations quickly and accurately. Here we reviewed the main conformational searching methods in protein-ligand docking, including the exhaustive searching, heuristic searching and other searching strategies. Meanwhile, we listed the representative docking programs by using these different searching strategies. Furthermore, we introduced the international assessment of predicted interaction, the widely used testing benchmark and the evaluation indexes in protein-ligand docking. Finally, we pointed out the existing major problems of the search strategies in protein-ligand docking and gave a prospect for such field in the near future.

**Key words:** protein-ligand interaction; conformational searching method; molecular docking

**基金项目:** NSFC-广东联合基金(第二期)超级计算科学应用研究专项资助项目; 国家自然科学基金(11647146, 81603152)资助项目; 江苏省六大人才高峰(2016-XYDXXJS-020)资助项目; 江苏省产学研前瞻(BY2016030-06)资助项目; 常州市应用基础研究计划(CJ20160016)资助项目; 江苏省研究生科研与实践创新计划(SJCX17\_0747)资助项目; 江苏省研究生科研与实践创新计划(SJCX17\_0748)资助项目。

**收稿日期:** 2018-06-08; **修订日期:** 2018-06-16

## 引 言

生物体内很多重要的生理过程,如细胞的周期调控、合成代谢、信号转导以及遗传信息的传递等,都有赖于蛋白质和配体分子的相互作用和识别<sup>[1-3]</sup>。分子对接是研究蛋白质与配体分子间相互作用与识别的理论方法,对生命活动的分子机理研究、生物分子复合物结构预测和靶向药物筛选等都具有重要的意义<sup>[4-6]</sup>。其中,创新药物开发是分子对接的重要应用领域。创新药物研究具有重大的社会效益和经济效益,而新药开发却是一件耗资巨大且效率较低的工作,一般平均要筛选十多万种以上的化合物才能得到一种新药。目前,已知的化合物分子数量已达到数千万量级,使药物开发具备了大数据的基础。但依靠传统药理学和实验模型测试成千上万个化合物,开发的时间成本和资金成本巨大。分子对接则可以快速高效地对千万量级的化合物分子进行计算筛选,精准发现有潜力的备选新分子,因此可以有效推进新药研发的进程。

分子对接被用于研究分子间相互作用,经典的热力学认为,蛋白质和配体分子间的相互作用是系统热力学平衡的过程,相互作用所形成的复合物结构应该是结合自由能最低的构象<sup>[7]</sup>。因此,一方面分子对接计算需要采用数学模型或函数尽量准确的计算结合自由能。另一方面,还需要发展有效的搜索算法快速找到自由能极低的构象。分子对接中的构象搜索是极其复杂的问题。假设受体结构固定不动,配体相对受体运动,在刚性对接的前提下,至少需要搜索平移和旋转 6 个自由度。如果进一步考虑受体或配体的分子柔性,则搜索空间将更复杂。

蛋白质-配体分子对接构象搜索方法主要包括快速穷举搜索和启发式搜索两大类。如果配体为生物大分子,如蛋白质、RNA 或 DNA,则相互作用的区域可以出现在分子表面任何位置(图 1(a)),因此往往需要进行全局搜索,可采用快速穷举搜索遍历各种位置,或采用启发式算法进行近似的全局搜索。如果配体为小分子化合物,分子对接则往往有一定的结合口袋(图 1(b)),可以限定构象空间的搜索范围,通常会采用启发式算法进行限定区域的搜索。

本文将结合国内外研究进展以及本课题组开展的一些研究工作,就蛋白质和配体分子对接中构象搜索方法的进展情况进行分析和综述。

## 1 快速穷举搜索方法

穷举搜索是当处理的问题找不到明确解决方法和规律时,逐一枚举此问题的各种可能情况,通过一定的评判原则从中挑选出符合要求的情况作为问题的候选解。当配体为生物大分子时,蛋白质受体表面的任何区域都有可能成为相互作用界面,因此,尽可能穷举所有构象并进行评价成为蛋白质-配体分子对接中的常用的搜索方法。

### 1.1 快速傅里叶变换方法

蛋白质-配体分子对接穷举搜索中,应用最广泛的是 Katchalski-Katzir 等<sup>[8]</sup>提出的快速傅里叶变换

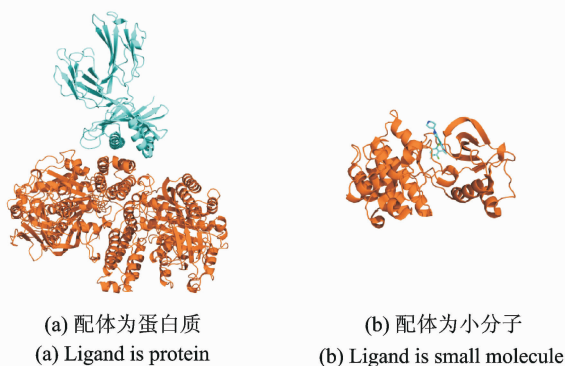


图 1 蛋白质-配体相互作用界面(蛋白质为橘黄色,配体为蓝色)

Fig. 1 Interface of protein-ligand interaction (Protein is the colored orange and ligand is the colored blue)

(Fast Fourier transform, FFT)方法。该方法首先将蛋白质分子和配体分子表示为三维格点数据,评判原则如分子几何或能量的匹配程度,采用相关函数来定量地描述,具体评判公式可表示为

$$c_{\alpha,\beta,\gamma} = \sum_{l=1}^N \sum_{m=1}^N \sum_{n=1}^N a_{l,m,n} \cdot b_{l+\alpha,m+\beta,n+\gamma} \quad (1)$$

式中: $a$ 和 $b$ 分别是蛋白质 $a$ 和配体 $b$ 的三维格点数据,可以表示几何、静电或统计势等各种格点形式。以静电计算为例,假如 $a$ 为蛋白质 $a$ 附近的静电场格点, $b$ 为配体 $b$ 的静电荷分布格点,则相关函数 $c$ 表示两者之间的静电相互作用能。其中 $\alpha,\beta,\gamma$ 分别是笛卡尔坐标系下,配体 $b$ 质心在3个方向上相对蛋白质 $a$ 质心平移的格点数,即 $(\alpha,\beta,\gamma)$ 为配体 $b$ 相对蛋白质 $a$ 的平移向量。 $N$ 是三维格点最大的格点值。直接计算相关函数的计算量较大,以式(1)为例,乘法计算量正比于 $N^3$ ,如果 $\alpha,\beta,\gamma$ 遍历所有的格点位置,则整个平移搜索的计算量为 $N^6$ 。相关函数的计算可以采用FFT方法来加速,具体的计算步骤如下:

(1)对蛋白质三维格点 $a$ 进行快速傅里叶变换并求复共轭操作

$$A_{\alpha,\beta,\gamma}^* = \{\text{FFT}(a_{l,m,n})\}^* \quad (2)$$

式中: $\alpha,\beta,\gamma \in \{1,2,\dots,N\}$ , $*$ 表示复共轭,快速傅里叶变换的乘法计算量是正比于 $N^3 \ln(N^3)$ 。

(2)对配体格点 $b$ 进行快速傅里叶变换

$$B_{\alpha,\beta,\gamma} = \text{FFT}(b_{l,m,n}) \quad (3)$$

(3)对 $A_{\alpha,\beta,\gamma}^*$ 和 $B_{\alpha,\beta,\gamma}$ 进行点乘计算

$$C_{\alpha,\beta,\gamma} = A_{\alpha,\beta,\gamma}^* \cdot B_{\alpha,\beta,\gamma} \quad (4)$$

式中点乘的计算量正比于 $N^3$ 。

(4)对 $C_{\alpha,\beta,\gamma}$ 进行快速傅里叶逆变换(Inverse fast Fourier transform, IFFT),从而得到相关函数 $c$

$$c_{\alpha,\beta,\gamma} = \text{IFFT}(C_{\alpha,\beta,\gamma}) \quad (5)$$

快速傅里叶逆变换的乘法计算量正比于 $N^3 \ln(N^3)$ ,因此通过上述计算,将构象搜索原本 $o(N^6)$ 的总计算量降低到 $o(N^3 \ln(N^3))$ 。然而,这里仅穷举和评价了配体分子 $b$ 平移的所有情况,但还需要进一步考虑配体分子旋转的自由度。因此,可根据欧拉角旋转配体 $b$ ,在每个角度下重复步骤(2)–(5),搜索和评价配体的匹配程度,并最终挑出符合要求的近天然复合物构象。

FFT算法因其高效性而被广泛应用于快速穷举的计算几何匹配,静电相互作用和原子对偏好性。大多数蛋白质-配体分子对接程序都采用该方法,如3D-Dock<sup>[9]</sup>,ClusPro<sup>[10]</sup>,ZDock<sup>[11]</sup>,FTDock<sup>[8]</sup>,DOT<sup>[12]</sup>,GRAMM<sup>[13]</sup>和MolFit<sup>[14]</sup>等程序。

## 1.2 球面极坐标傅里叶变换方法

球面极坐标傅里叶变换(Spherical polar Fourier, SPF)和快速傅里叶变换方法其本质一致,都是穷举采样构象空间。不同之处在于,球面极坐标采用正交球面极基函数的三维多项式展开来取代直接计算传统的三维笛卡尔坐标。故其具有的优势为:旋转配体操作可以通过变换其扩展系数来实现,极大降低了需要循环搜索旋转角度的计算量。甚至可以把旋转自由度也表示成系数,进行五维傅里叶变换。Hex<sup>[15]</sup>和FMFT<sup>[16]</sup>等程序主要采用球面极坐标傅里叶变换进行构象采样。

## 2 启发式搜索算法

启发式搜索算法是把对接体系中配体分子的平移和旋转操作先随机进行编码,并根据能量评分对操作后的配体构象进行优化和取舍,最终找到能量最低的配体分子构象。目前蛋白质-配体分子对接中常用的启发式搜索算法包括,蒙特卡洛(Monte Carlo, MC)算法<sup>[17]</sup>、遗传算法(Genetic algorithm, GA)<sup>[18]</sup>以及群智能(Swarm intelligence, SI)算法<sup>[19]</sup>。

## 2.1 蒙特卡洛算法

MC 算法有较好的物理学理论支持,因此在分子对接中应用也非常广泛。该方法首先在构象空间中随机地采样,并评估采样的函数值,通过函数值高低进行取舍,保留已得到的最优解作为最终解。MC 算法不受研究问题搜索空间结构和分布的影响,在采样数趋近无穷时,理论上会收敛到全局最优解。该方法的物理基础与分子对接一致,即寻找能量更低的状态。由于不可能达到无穷次的采样,故采用重要性采样的方法,即 Metropolis 准则<sup>[17]</sup>进行有偏好的采样。具体方法描述如下:

以配体分子的某个构象作为初始状态  $\sigma$ ,通过随机平移和旋转产生新的状态  $\sigma'$ 。计算两个状态的能量  $E(\sigma')$  和  $E(\sigma)$ 。如果新状态的能量更低,则有理由将其保留下来。反之,则需要判断,不能因为  $\sigma'$  状态的能量高就直接抛弃,这样会忽略了热运动的影响,并容易陷入局部极小。可通过计算 Boltzmann 因子的比值来判断,计算公式为

$$r = \exp\left(\frac{E(\sigma) - E(\sigma')}{kT}\right) \quad (6)$$

式中: $T$  是绝对温度, $k$  是 Boltzmann 常数。由于  $r$  在  $[0, 1]$  之间,且能量越接近, $r$  值越接近 1。故可以采用随机发生器产生  $[0, 1]$  之间的随机数  $\xi$ 。通过比较  $r$  和  $\xi$  的大小,判断是否保留  $\sigma'$ 。如果两个状态能量很接近,则  $r$  值接近 1,此时有很大概率会保留  $\sigma'$ ;反之,如果新状态  $\sigma'$  的能量过高,则被保留的概率将会很低。

MC 算法在配体构象空间中随机采样,且不受构象空间结构和分布的影响,是一种通用的搜索方法。但该方法在实际使用时,可能需要较长的计算时间才能给出较好的解。RosettaDock 程序<sup>[20, 21]</sup>和 GLIDE<sup>[22, 23]</sup>程序就是采用 MC 算法进行全空间构象搜索,特别是 RosettaDock 程序,其生物大分子对接往往需要运行较长的时间。

## 2.2 遗传算法

GA 算法将复杂优化问题视为生物的遗传和进化过程,先随机生成问题的解,然后进行交叉和突变操作,通过适应度选择使候选解不断优化,最终收敛于局部最优解甚至全局最优解。GA 算法最早由 Holland 在 1975 年提出<sup>[18]</sup>。该算法中,若干个体构成一个种群,个体采用二进制数据表示。先随机产生初始种群,即问题的初始解。然后通过交叉操作和突变操作来产生新的个体,通过适应度函数对新个体进行评价。根据轮盘赌的方式对新个体进行选择,适应度越高的则被选中的概率也越高,因此个体会逐步逼近适应度更高的情况,从而得到最优解。GA 算法被广泛应用于分子对接构象搜索中,蛋白质分子一般固定不动,而配体相对于蛋白质分子的平移和旋转操作作为个体的二进制数据,将蛋白质和配体相互作用的能量函数作为适应度函数。分子对接中,能量越低的构象对应的个体适应度越高,从而挑选出能量更低的复合物构象。

GA 算法在分子对接搜索中的主要步骤是:(1)先产生初始种群,种群中的个体表示配体的平移和旋转操作;(2)平移和旋转配体构象,把对配体构象的操作,如平移和旋转进行二进制编码,对二进制数据进行交叉操作或者突变操作,即产生新的平移和旋转位置;(3)选择操作。用能量函数来评价配体分子的适应度,采用轮盘赌的方式对新个体进行选择,适应度好的个体有更大的概率被选择,最终保证种群的个数不变;(4)判断搜索是否已收敛,可以采用循环的步数或种群的能量函数得分情况判断收敛,如循环步数达到阈值则结束,或新种群与旧种群的能量函数分值差距在一定的阈值之内,也可认为收敛了。

GA 算法的主要特点是不需要确定的规则,可以直接对个体对象进行操作,能自适应地调整搜索方向,不存在求导和函数连续性的限定,并具有良好的全局寻优能力。作为一种启发式寻优算法,GA 算法以其简单通用、高效且易于并行处理等显著特点,在生物信息学各领域得到了广泛应用,已成为蛋白

质-配体分子对接中最重要的构象搜索算法之一。GOLD<sup>[24]</sup>和 AutoDock<sup>[25]</sup>等著名的分子对接程序就是采用 GA 算法进行构象搜索,并取得了较好的对接结果。

### 2.3 群智能算法

SI 算法源于对自然界中的群体行为进行研究,包括蚁群、鱼群、鸟群、细菌群和其他微生物群等<sup>[26]</sup>。SI 系统由一群相互间交互和与周围环境交互的个体构成,并模拟群体自组织的行为。SI 算法相对于 GA 算法来说搜索效率更高、全局寻优能力更强、收敛速度也更快,因而该方法很快被引入到分子对接的构象搜索应用中。

分子对接的构象搜索问题,SI 算法中粒子群的应用较多,因此主要以粒子群为例介绍 SI 算法在分子对接构象搜索中的应用。在分子对接构象搜索中,目标搜索区域中每个粒子的位置代表问题的潜在可行解,即配体的旋转或平移操作,每个粒子根据周围粒子情况调整自己的位置和速度。能量函数作为适应度函数,评估每个粒子的当前位置。在整个过程中,每个粒子会根据其找到的最好位置,即个体最优解  $p_{best}$ ,以及所有粒子的  $p_{best}$  中最好的位置,即种群最优个体  $g_{best}$ ,来更新其速度和位置。粒子速度及位置更新公式为

$$\begin{aligned} \chi_i(t+1) &= \chi_i(t) + V_i(t+1) \\ V_i(t+1) &= \omega \cdot V_i(t) + c_1 \cdot r_1 \cdot (p_{best}(t) - \chi_i(t)) + c_2 \cdot r_2 \cdot (g_{best}(t) - \chi_i(t)) \end{aligned} \quad (7)$$

式中: $r_1, r_2$  为  $[0, 1]$  之间的随机数; $c_1, c_2$  为常数。SwarmDock<sup>[27]</sup>就是采用粒子群算法进行全空间构象搜索。本课题组对 Autodock 3.0 程序中全局搜索的 GA 算法采用蚁群算法进行了改写<sup>[28]</sup>,提升了程序的搜索效率和寻优能力。

## 3 其他搜索方法

除了上述两大类构象搜索算法,还有其他一些课题组提出了不同于上述两类方法的搜索策略用于构象搜索。

### 3.1 基于图像识别的构象搜索算法

将蛋白质-配体分子识别看成三维图形的识别问题,一些课题组采用图像识别中的算法<sup>[29]</sup>,如几何哈希方法。该算法先对配体分子进行预处理,对配体分子的表面关键点建立几何哈希表。然后进行分子识别,将蛋白质的表面特征与配体分子进行关键点匹配,评价配体不同方位的匹配程度,从而确定配体的构象。该方法主要针对分子表面特征的匹配,避免了分子对接中最耗时的平移和旋转操作,因而有效提升了搜索效率,采用该方法的代表程序主要有 LZerD<sup>[30]</sup>和 PatchDock<sup>[31]</sup>等。

### 3.2 基于分子动力学的构象搜索方法

分子动力学是模拟蛋白质和配体相互作用的有效计算方法,但由于其需要计算原子间的多种相互作用,并模拟动力学运动过程,因此极为耗时且容易陷入局部极小,在分子对接构象搜索中不太常用。HADDOCK<sup>[32]</sup>程序将位点约束和分子动力学模拟相结合来进行分子对接。该方法利用预测的或实验提供的位点信息,建立了模糊的相互作用约束(Ambiguous interaction restraints, AIRs),并将其作为蛋白质与配体与关键位点相互作用的位置约束来加速分子动力学构象采样的效率,从而克服常规分子动力学的局限性。

### 3.3 基于片段组合的构象搜索方法

在考虑配体分子柔性构象产生的时候,一些课题组采用片段组合的方法进行构象搜索。该方法将配体分子看成多个刚性片段组成,刚性片段通过可旋转键连接。通过不同角度的组合来产生大量柔性配体构象,采用均方根偏差判断构象的多样性,并选取代表性构象保留。采用该方法的代表程序主要有

Dock 程序<sup>[33]</sup>。

此外,在分子对接最终还需要进行构象的局部位置进行调整和优化。在局部构象调整和优化上,还有一些课题组采用了确定性优化搜索方法,如 Autodock Vina 采用拟牛顿法<sup>[34]</sup>,ITScore-PP 采用单纯型法<sup>[35]</sup>等。

综上,蛋白质-配体分子对接的方法很多,表 1 列出了目前应用较为广泛的蛋白质-配体分子对接软件,并对其配体类型和构象搜索方法进行了简单介绍。

表 1 应用较为广泛的蛋白质与其配体对接程序

Tab. 1 Widely-used docking softwares between protein and its ligands

软件名称	配体类型	构象搜索算法	研发单位
ZDock <sup>[11]</sup>	蛋白质	快速傅里叶变换	波士顿大学
FTDock <sup>[8]</sup>	蛋白质	快速傅里叶变换	魏茨曼科学研究所
ClusPro <sup>[10]</sup>	蛋白质	快速傅里叶变换	波士顿大学
3D-Dock <sup>[9]</sup>	蛋白质	快速傅里叶变换	英国癌症研究基金会
Hex <sup>[15]</sup>	蛋白质	球面极坐标傅里叶变换	法国国家信息与自动化研究所
DARWIN <sup>[36]</sup>	蛋白质	遗传算法	宾夕法尼亚大学
RosettaDock <sup>[20, 21]</sup>	蛋白质	蒙特卡洛算法	华盛顿大学
SwarmDock <sup>[27]</sup>	蛋白质	粒子群算法	弗朗西斯·克里克研究所
DOT <sup>[12]</sup>	蛋白质/DNA	快速傅里叶变换	加州大学圣迭戈分校
HADDOCK <sup>[32]</sup>	蛋白质/DNA	分子动力学	荷兰乌特勒支大学
LZerD <sup>[30]</sup>	蛋白质	几何哈希方法	普渡大学
GRAMM <sup>[13]</sup>	蛋白质/小分子	快速傅里叶变换	堪萨斯大学
ICM-DOCK <sup>[37]</sup>	蛋白质/小分子	蒙特卡洛算法	Molsoft 公司
AutoDock <sup>[25]</sup>	小分子	拉马克遗传算法	美国斯克利普斯研究所
GLIDE <sup>[22, 23]</sup>	小分子/DNA	蒙特卡洛算法	Schrödinger 公司
GOLD <sup>[24]</sup>	小分子	遗传算法	剑桥晶体数据中心
Affinity <sup>[38]</sup>	小分子	蒙特卡洛算法和分子动力学	Accelrys 公司
Dock <sup>[33]</sup>	蛋白质/小分子/DNA	片段生长法	加州大学旧金山分校

## 4 分子对接构象搜索的实验评测

### 4.1 国际评估实验

为了推动蛋白质分子对接技术的发展,2001 年欧洲生物信息学研究所开始举办生物大分子相互作用预测评估实验(Critical assessment of predicted interaction, CAPRI)<sup>[39-41]</sup>。截至目前,共进行了 45 轮实验。在该评估实验中,组委会首先选取尚未发表实验结构数据的蛋白质复合物作为预测内容,要求国际上参加实验的小组在规定时间内,从蛋白质单体结构出发,用理论方法对蛋白质复合物结构进行预测,并通过网络提交预测结果。随后,CAPRI 评估小组会将这些结构与实验结构进行对比,来评估理论预测方法的准确性,并在 CAPRI 网上公布。大约每隔 2~3 年,召开一次 CAPRI 评估与研讨会。目前已成功举办了 5 届,它们分别是于 2002,2004,2007,2009,2013 和 2016 年在法国、意大利、加拿大、西班牙、荷兰和以色列举办的会议。

### 4.2 分子对接评估数据库

美国波士顿大学的研究小组建立的数据集是目前应用最广泛的分子对接评估数据集,已经发展到第 5 版了<sup>[42]</sup>。数据集中有 230 个蛋白质复合物结构,并根据结合前后的构象变化分成 151 个刚性(Rig-

id-body)体系,45个中等难度(Medium difficult)体系,34个最难(Most difficult)体系。每个复合物结构有其相应的自由态单体(Unbound)结构。因此各小组可以测试自己的构象搜索方法,并比较构象变化对搜索算法的影响。

### 4.3 评估指标

配体均方根偏差(L\_RMSD)是评价分子对接预测的复合物构象与实验解析的复合物结构相似程度的重要指标。L\_RMSD为预测结构与实验结构中的受体叠落后,配体间主链原子(N,C,CA,O)位置的均方根偏差。CAPRI评估小组将预测结构分为高精度、中等精度、可接受和错误结果4种类型。高精度预测结果的要求是 $L\_RMSD \leq 0.1 \text{ nm}$ ,中等精度预测结果是 $0.1 \text{ nm} < L\_RMSD \leq 0.5 \text{ nm}$ ,可接受预测结果为 $0.5 \text{ nm} < L\_RMSD \leq 1.0 \text{ nm}$ , $L\_RMSD > 1.0 \text{ nm}$ 的结构为错误结构。

## 5 结束语

蛋白质-配体分子对接一方面需要搜索能量低的构象,另一方面需要在尽量短的时间内搜索到各种可能的情况,因此快速有效的搜索算法是分子对接中重要的研究领域,迫切需要采用新的理论和计算方法提升现有程序的计算效率。为进一步提升构象采样的准确性和计算效率,以下几个方面可能将是未来的研究和发展方向:

(1)蛋白质柔性采样方法。柔性依然是制约分子对接的一个技术瓶颈,蛋白质与配体相互作用的过程中会发生构象变化,考虑蛋白质主链的柔性会增加搜索的构象空间。可以考虑的方法包括精确确定蛋白质的柔性部位、缩小柔性搜索的范围,并发展快速有效的构象搜索算法。或者采用弹性网络以及图论等方法计算蛋白质的运动模式,作为新的自由度进行构象搜索和优化。另一方面,考虑蛋白质的柔性同时会增加构象数量,既有正确构象也引入了错误构象,因此还需要发展更为精确的能量函数,有效区分柔性对接产生的复合物构象。

(2)基于模板的采样方法。对于蛋白质等生物大分子对接情况,直接采用快速穷举算法,尽管可以快速搜索整个构象空间,但同时也引入了较多错误的构象,为区分正确构象提供了难度。借鉴蛋白质折叠的思想,可以搜索蛋白质数据库中已有的结构模板,如果受体和配体的结构模板恰好有相应的复合物结构,则无需搜索构象空间,可直接根据搜索到的复合物结构进行对接构象模建或局部优化。即使没有完全一致的复合物结构,如果有类似的相互作用界面结构,也能为搜索提供有益的信息,减少构象空间的搜索范围。

(3)面向新硬件平台上的构象采样方法。计算机处理器的发展日新月异,特别是图形处理单元(GPU)的计算能力快速提高,许多研究都根据新处理器改进原有算法。可以针对该处理器高效并行处理的特点,发展基于GPU的构象采样加速方法来提升采样效率。另外,量子计算机等全新的计算技术进一步发展,将很可能颠覆复杂优化问题的计算框架和模型,因此也需要针对其特点建立相应的分子对接构象搜索方法。

### 参考文献:

- [1] Huttlin E L, Bruckner R J, Paulo J A, et al. Architecture of the human interactome defines protein communities and disease networks[J]. *Nature*, 2017,545:505-509.
- [2] Li X H, Chavali P L, Babu M M. Capturing dynamic protein interactions[J]. *Science*, 2018,359(6380):1105-1106.
- [3] 於东军,朱一亨,胡俊. 识别蛋白质配体绑定残基的生物计算方法综述[J]. *数据采集与处理*,2018,33(2):195-206.  
Yu Dongjun, Zhu Yiheng, Hu Jun. An overview of biocomputing methods of targeting protein-ligand binding residues[J]. *Journal of Data Acquisition and Processing*, 2018,33(2):195-206.
- [4] Vakser I A. Protein-protein docking: From interaction to interactome[J]. *Biophysical Journal*, 2014,107(8):1785-1793.
- [5] Villoutreix B O, Bastard K, Sperandio O, et al. In silico-in vitro screening of protein-protein interactions: Towards the next

generation of therapeutics[J]. *Current Pharmaceutical Biotechnology*, 2008,9(2):103-122.

- [6] Xue L C, Dobbs D, Bonvin A, et al. Computational prediction of protein interfaces: A review of data driven methods[J]. *Febs Letters*, 2015, 589(23):3516-3526.
- [7] 王存新, 常珊, 龚新奇, 等. 蛋白质-蛋白质分子对接中打分函数研究进展[J]. *物理化学学报*, 2012,28(4):751-758.  
Wang Cunxin, Chang Shan, Gong Xinqi, et al. Progress in the scoring functions of protein-protein docking[J]. *Acta Physico-Chimica Sinica*, 2012,28(4):751-758.
- [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[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 1992,89(6):2195-2199.
- [9] Aloy P, Querol E, Aviles F X, et al. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking[J]. *Journal of Molecular Biology*, 2001,311(2):395-408.
- [10] Comeau S R, Gatchell D W, Vajda S, et al. ClusPro: An automated docking and discrimination method for the prediction of protein complexes[J]. *Bioinformatics*, 2004,20(1):45-50.
- [11] Chen R, Li L, Weng Z P. ZDock: An initial-stage protein-docking algorithm[J]. *Proteins-Structure Function and Genetics*, 2003,52(1):80-87.
- [12] Mandell J G, Roberts V A, Pique M E, et al. Protein docking using continuum electrostatics and geometric fit[J]. *Protein Engineering*, 2001,14(2):105-113.
- [13] Vakser I A. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex[J]. *Proteins-Structure Function and Genetics*, 1997:226-230.
- [14] Berchanski A, Shapira B, Eisenstein M. Hydrophobic complementarity in protein-protein docking[J]. *Proteins-Structure Function and Bioinformatics*, 2004,56(1):130-142.
- [15] Macindoe G, Mavridis L, Venkatraman V, et al. HexServer: An FFT-based protein docking server powered by graphics processors[J]. *Nucleic Acids Research*, 2010, 38(Web Server Issue):W445-W449.
- [16] Padhorny D, Kazennov A, Zerbe B S, et al. Protein-protein docking by fast generalized Fourier transforms on 5D rotational manifolds[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2016,113(30):E4286-E4293.
- [17] Metropolis N, Rosenbluth A W, Rosenbluth M N, et al. Equations of state calculations by fast computing machines[J]. *Journal of Chemical Physics*, 1953,21(6):1087-1092.
- [18] Holland J H. *Adaptation in natural and artificial system*[M]. Ann Arbor: the University of Michigan Press, 1975.
- [19] Kennedy J, Eberhart R C, Shi Y. *Swarm intelligence*[M]. San Francisco: Morgan Kaufman, 2001.
- [20] Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein-protein docking[J]. *Protein Science*, 2005,14(5):1328-1339.
- [21] Gray J J, Moughon S, Wang C, et al. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations[J]. *Journal of Molecular Biology*, 2003,331(1):281-299.
- [22] Halgren T A, Murphy R B, Friesner R A, et al. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening[J]. *Journal of Medicinal Chemistry*, 2004,47(7):1750-1759.
- [23] Friesner R A, Banks J L, Murphy R B, et al. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy[J]. *Journal of Medicinal Chemistry*, 2004,47(7):1739-1749.
- [24] Jones G, Willett P, Glen R C, et al. Development and validation of a genetic algorithm for flexible docking[J]. *Journal of Molecular Biology*, 1997,267(3):727-748.
- [25] Morris G M, Goodsell D S, Halliday R S, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function[J]. *Journal of Computational Chemistry*, 1998,19(14):1639-1662.
- [26] 刘金勇, 郑慧辉, 陆慧娟. 基于聚类和微粒群优化的基因选择方法[J]. *数据采集与处理*, 2014,29(1):83-89.  
Liu Jinyong, Zheng Enhui, Lu Huijuan. Gene selection based on clustering method and particle swarm optimization[J]. *Journal of Data Acquisition and Processing*, 2014,29(1):83-89.
- [27] Torchala M, Moal I H, Chaleil R A G, et al. SwarmDock: A server for flexible protein-protein docking[J]. *Bioinformatics*, 2013,29(6):807-809.
- [28] 常珊, 田绪红, 陈琰. 基于蚁群算法的分子对接搜索方法研究[J]. *乐山师范学院学报*, 2010,25(5):45-47.  
Chang Shan, Tian Xuhong, Chen Yan. Research on molecular docking search method based on ant colony algorithm[J].



*Journal of Leshan Teachers College*, 2010, 25(5):45-47.

- [29] Fischer D, Lin S L, Wolfson H L, et al. A geometry-based suite of molecular docking processes[J]. *Journal of Molecular Biology*, 1995, 248(2):459-477.
- [30] Venkatraman V, Yang Y F D, Sael L, et al. Protein-protein docking using region-based 3D Zernike descriptors[J]. *BMC Bioinformatics*, 2009, 10:407.
- [31] Schneidman-Duhovny D, Inbar Y, Nussinov R, et al. PatchDock and SymmDock: Servers for rigid and symmetric docking [J]. *Nucl Acids Res*, 2005, 33(Suppl\_2):W363-367.
- [32] Dominguez C, Boelens R, Bonvin A. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information[J]. *Journal of the American Chemical Society*, 2003, 125(7):1731-1737.
- [33] Kuntz I D, Blaney J M, Oatley S J, et al. A geometric approach to macromolecule-ligand interactions[J]. *Journal of Molecular Biology*, 1982, 161(2):269-288.
- [34] Trott O, Olson A J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading[J]. *Journal of Computational Chemistry*, 2010, 31(2):455-461.
- [35] Huang S Y, Zou X. An iterative knowledge-based scoring function for protein-protein recognition[J]. *Proteins: Structure, Function, and Bioinformatics*, 2008, 72(2):557-579.
- [36] Taylor J S, Burnett R M. DARWIN: A program for docking flexible molecules[J]. *Proteins-Structure Function and Genetics*, 2000, 41(2):173-191.
- [37] Fernandez-Recio J, Totrov M, Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes [J]. *Journal of Molecular Biology*, 2004, 335(3):843-865.
- [38] Luty B A, Wasserman Z R, Stouten P F W, et al. A molecular mechanics grid method for evaluation of ligand-receptor interactions[J]. *Journal of Computational Chemistry*, 1995, 16(4):454-464.
- [39] Janin J. Protein-protein docking tested in blind predictions: The CAPRI experiment[J]. *Molecular Biosystems*, 2010, 6(12):2351-2362.
- [40] Janin J, Henrick K, Moult J, et al. CAPRI: A critical assessment of predicted interactions[J]. *Proteins-Structure Function and Bioinformatics*, 2003, 52(1):2-9.
- [41] Lensink M F, Wodak S J. Docking, scoring, and affinity prediction in CAPRI[J]. *Proteins-Structure Function and Bioinformatics*, 2013, 81(12):2082-2095.
- [42] Vreven T, Moal I H, Vangone A, et al. Updates to the integrated protein-protein interaction benchmarks: Docking benchmark Version 5 and affinity benchmark Version 2 [J]. *Journal of Molecular Biology*, 2015, 427(19):3031-3041.

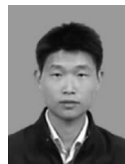
#### 作者简介:



常珊 (1982-), 男, 教授, 工学博士, 研究方向: 生物信息学和并行计算, E-mail: schang@jsut.edu.cn.



陆旭峰 (1993-), 男, 硕士研究生, 研究方向: 高性能计算方法, E-mail: xflu0913@163.com.



王峰 (1982-), 男, 工学博士, 高级工程师, 研究方向: 智能算法和生物信息学, E-mail: wfeng@cczu.edu.cn.

(编辑: 张彤)