第50卷第2期
2010年3月

大连理工大学学报
Journal of Dalian University of Technology

Vol. 50, No. 2
Mar. 2010

工程力学

文章编号：1000-8608(2010)02-0157-05

一种采用知识打分函数的分子对接方法

王希诚*[1]，赵晓宇[2]，康　玲[3]，李洪林[4]

（1.大连理工大学 工业装备结构分析国家重点实验室，辽宁 大连 116024；
2.大连理工大学 工程力学系，辽宁 大连 116024；
3.大连理工大学 计算机科学与技术学院，辽宁 大连 116024；
4.中国科学院 上海药物研究所，上海 201203）

摘要：分子对接是计算机辅助药物分子优化设计的重要组成部分.引入了一种通过计算原子对间距离来评价结合自由能的知识打分函数，其构造方法与平均力势能函数相似.同时采用基于信息熵的多种群自适应遗传算法，形成一种新型的分子对接程序KGAsDock.给出与著名的分子对接程序DOCK6.1的对接结果比较，数值实验表明，该算法在不降低计算效率的前提下，提高了对接的精度.

关键词：分子对接；知识打分函数；遗传算法；优化模型

中图分类号：TP18　　**文献标志码**：A

0　引　言

随着计算机技术的高速发展及其在各个领域的广泛应用，计算机辅助药物设计已经成为创新药物研究的一种新方法和技术.分子对接作为基于受体药物设计的重要方法之一，已经成为可靠、相对廉价的用于先导化合物发现的一种重要手段.分子对接包括3个相互关联的部分：结合位点的识别、有效的构象优化方法及打分函数.20世纪80年代，Kuntz等[1]发展了模拟小分子与生物大分子结合三维结构及其强度的计算方法——分子对接(molecular docking)方法，并开发了第一个分子对接程序DOCK.此后，为得到精确的结合构象和正确地预测活性，各种构象优化方法及打分函数应运而生.

本文通过Boltzmann规则将原子间距离的概率分布转化为与距离有关的蛋白质-配体原子对间作用能的知识打分函数，将其与基于信息熵的多种群自适应遗传算法相结合，形成有效的分子对接程序，用于计算配体与蛋白质的结合能；并与DOCK6.1对接结果相比较，以证明其有效性.

1　基于知识打分的分子对接模型

目前，可以用于分子对接及虚拟筛选的结合自由能评价方法，大致上可以分为基于力场、基于经验及基于知识的3类打分函数.基于力场的打分函数多采用AMBER和CHARMM力场的非键相互作用部分，将蛋白质受体-配体的结合自由能近似为范德华力与静电力相互作用的加和，DOCK4[2](最新版本为DOCK6.1)、GAsDock[3]等对接程序均采用力场打分函数作为分子对接的评价标准.经验打分函数认为结合自由能可以通过多项不同作用的加和来解释，权系数可以通过已知结合能的蛋白质-配体的训练集获得.知识打分函数通过已知的受体-配体结构，利用Boltzmann规则[4]将原子间距离的概率分布转化为与距离有关的受体-配体原子对间的作用能，并将结合过程中具有复杂相关性而又很难明确建模的结合效应隐含进去.本文采用类似经典打分PMF[5](potentials of mean force)的构造方法，从包含2 422个复合物的训练集中确定了17种蛋白质受体原子类型(详见表1)及25种配体原子类型(详见表2)，

收稿日期：2007-10-24；　**修回日期**：2010-01-14.
基金项目：国家自然科学基金资助项目(10772042)；“八六三”国家高技术研究发展计划资助项目(2006AA01124).
作者简介：王希诚*(1946-)，男，教授，博士生导师，E-mail：guixum@dlut.edu.cn.

表 1 蛋白质原子类型

Tab. 1 Protein atom type

类型	类型说明
CF	非极性的脂肪族碳(例如 C_β)
CP	极性 sp^2 或 sp^3 上不与碳或氢相邻的碳原子(例如 C_α)
cF	非极性芳香族的碳原子
cP	极性芳香族的碳原子
CO	与带负电荷的氧原子相邻的碳原子
CN	与带正电荷的氮原子相邻的碳原子
NR	平面环结构中的氮原子(例如 HIS ND、HIS NE)
ND	作为氢键给体的氮原子(例如 backbone N、TRP NE、ASN ND)
NC	氮正离子
OA	作为氢键受体的氧原子(例如 backbone O、ASN OD、GLN OE)
OC	带负电荷的氧原子
OD	作为氢键给体的氧原子(例如 TYR OH、SER OG、THR OG)
OW	水分子中的氧原子
SA	作为氢键受体的硫原子(例如 MET SD)
SD	作为氢键给体的硫原子(例如 CYS SG)
HH	蛋白质中的氢原子
ME	蛋白质中的金属原子

表 2 配体原子类型

Tab. 2 Ligand atom type

类型	类型说明
C. 3	sp^3 杂化的碳原子
C. 2	sp^2 杂化的碳原子
C. 1	sp 杂化的碳原子
C. ar	芳香族碳原子
C. cat	碳正离子
N. 3	sp^3 杂化的氮原子
N. 2	sp^2 杂化的氮原子
N. 1	sp 杂化的氮原子
N. ar	芳香族氮原子
N. am	氨基氮原子
N. pl3	硝基氮原子
N. 4	sp^3 杂化的带正电荷的氮原子
O. 3	sp^3 杂化的氧原子
O. 2	sp^2 杂化的氧原子
O. co2	羧基氧原子
S. 3	sp^3 杂化的硫原子
S. 2	sp^2 杂化的硫原子
S. O	亚砜硫原子
S. o2	砜硫原子
P. 3	sp^3 杂化的磷原子
Cl	氯原子
F	氟原子
Br	溴原子
H	氢原子
V	其他原子

通过 Boltzmann 规则得到了不同类型原子对在各个距离上的作用能,并且通过体积修正项将结合过程中的疏水作用及熵变隐含进去,其表达式如下:

$$\Delta A_b = \sum_{\substack{k_1 \\ r<\bar{r}}} A_{ij}(r) \tag{1}$$

$$A_{ij}(r) = -kT \ln\left[f^j_{vol\_corr}(r)\frac{\rho^{ij}_s(r)}{\rho^{ij}_b}\right] \tag{2}$$

式中:$A_{ij}(r)$ 为 i 类型受体原子与 j 类型配体原子在距离 r 上的能量值;k_1 为复合物训练集中所有距离 $r<\bar{r}$ 的原子对的数量,$\bar{r}$ 是在该训练集中定义的一个截断距离(本文取 12 nm);k 为 Boltzmann 常数;T 为热力学温度;$\rho^{ij}_s(r)$ 及 ρ^{ij}_b 分别为某一距离半径$(r, r+\Delta r)$的球壳上及距离半径为 R 的球上的 ij 类型原子对的原子类型数密度;$f^j_{vol\_corr}(r)$ 为隐含熵变及疏水作用的体积修正项;最终的能量值 ΔA_b 为将所有原子对距离 $r<\bar{r}$ 的原子对能量值加和得到. 关于各个距离上的原子类型数密度及体积修正项的计算表达如下:

$$\rho^{ij}_b = \sum_{k_1} \frac{n^{ij}_b}{V(R)};$$

$$\rho^{ij}_s(r) = \rho^{ij}_{(r,r+\Delta r)}(r) = \sum_{k_1} \frac{n^{ij}_{(r,r+\Delta r)}(r)}{V_{(r,r+\Delta r)}(r)} \tag{3}$$

$$f^j_{vol\_corr}(r) = \frac{\rho^{kj}_b/(\rho^{kj}_b + \rho^{lj}_b)}{\rho^{kj}(r)/(\rho^{kj}(r)+\rho^{lj}(r))} \tag{4}$$

式中:$n^{ij}_{(r+\Delta r)}(r)$ 及 n^{ij}_b 分别表示距离半径$(r, r+\Delta r)$的球壳上及距离半径为 R 的球上的 ij 类型原子对的个数;$V(R)$ 及 $V_{(r,r+\Delta r)}(r)$ 分别表示不同半径的球体及球壳的体积;$\rho^{kj}(r)$ 及 ρ^{kj}_b 表示与 j 类型的配体原子距离分别为 r 及 R 的所有蛋白质受体原子数密度;而 $\rho^{lj}(r)$ 及 ρ^{lj}_b 表示与 j 类型的配体原子距离分别为 r 及 R 的所有配体原子数密度. 体积修正项 $f^j_{vol\_corr}(r)$ 通过计算配体原子周围蛋白质原子所占比例,将疏水作用及熵变效应隐含进去,从而对打分函数做出进一步的修正.

2 分子半柔性对接优化模型

本文采用只考虑小分子柔性的半柔性对接优化模型,包括小分子平动、转动及旋转键在内的一系列变化. 优化对接模型为

$$\begin{aligned} &\min \quad f(\boldsymbol{x}) \\ &\text{s.t.} \quad g_k(\boldsymbol{x}) \leqslant 0; \quad k = 1, 2, \cdots, r \end{aligned} \tag{5}$$

式中:$\boldsymbol{x} = (T_x \quad T_y \quad T_z \quad R_x \quad R_y \quad R_z \quad T_{b1} \quad T_{b2} \quad \cdots \quad T_{bn})^T$,其中 T_x、T_y、T_z、R_x、R_y、R_z 是配体分子的几何中心及旋转度,对应于配体分子的取向,$T_{b1}, T_{b2}, \cdots, T_{bn}$ 是配体分子的可旋转键,

描述配体分子的构象信息，n 为可旋转键数目. 目标函数 $f(\boldsymbol{x})$ 选取上述知识型打分函数.

3　基于信息熵的多种群自适应遗传算法

本文在采用带有空间收缩的多种群遗传算法[6]的基础上同时加入了自适应策略，将其与知识打分函数结合用于寻找分子对接过程中的低能构象，用信息熵控制最优解搜索空间的收缩，并用空间收缩的尺度作为算法停止的判据，进化过程中添加了最优保留策略，从而确保了算法的全局收敛性.

对于多约束优化问题(5)，可利用评价约束函数 PEC 及精准惩罚函数法将其转化为序列无约束优化问题：

$$\min\ \varphi_{\psi}(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{\alpha}{\psi}\ln\left\{1 + \sum_{i=1}^{m}\exp\left[\psi g_i(\boldsymbol{x})\right]\right\} \tag{6}$$

式中：α 为惩罚因子，α 只要大于一个阈值就可以使问题的解位于可行域内；ψ 的取值一般为$[10^3, 10^5]$，这种方法针对所有约束按“松”与“紧”自动调整惩罚力度，能够有效地处理约束，计算效率较高. 对于遗传算法，需要将上式转化为无约束最大化问题：

$$\max F(\boldsymbol{x}) = C - \varphi_{\psi}(\boldsymbol{x}) \tag{7}$$

式中：C 是一个大的正数以确保 $F(\boldsymbol{x})$ 在计算过程中为正值，式(7) 就是本文采用的演化设计模型，$F(\boldsymbol{x})$ 为适应值函数.

将通讯论中的信息熵理论引入优化方法中，构造基于信息熵控制的遗传演化模型如下：

$$\begin{cases} \min\left[-\sum_{m=1}^{M} p_m F(\boldsymbol{x})\right] \\ \min H = -\sum_{m=1}^{M} p_m \ln p_m \\ \text{s.t.}\ \sum_{m=1}^{M} p_m = 1;\ p_m \in [0,1] \\ g_k(\boldsymbol{x}) \leqslant 0;\ k = 1,2,\cdots,q \end{cases} \tag{8}$$

式中：M 为种群个数，通过定义最优解落在第 m 个种群的概率 $p_m(m = 1,2,\cdots,M)$，从而引入信息熵 H 以衡量最优解落于某一种群的不确定性. 初始时，$p_m = 1/M, m = 1,2,\cdots,M$，$H$ 取最大值；随着优化的进行，遗传迭代解将逐步逼近最优解，p_m 及 H 都将随之变化，当在某一种群取到最优解时，不确定性为零，熵 H 取极小值，从而得到原问题(7) 的最优解. 信息熵的介入有助于加快进化过程.

在本文算法中，还将遗传算法中的交叉概率及变异概率作为设计变量参与优化，这种自适应策略，可以有效防止过早收敛问题的发生，同时提高了算法的搜索速度，保持了种群的多样性，从而大大降低了人为因素对优化算法的影响.

这种基于信息熵的多种群自适应遗传算法，引入了种群竞争机制及交叉、变异概率的自适应策略，并用信息熵控制空间收缩，提高了遗传迭代的效率，算法稳定可靠，具有较强的全局寻优能力，收敛速度也有较大的提高.

4　结果与讨论

本文将知识打分函数与优化算法相结合，开发了新的分子对接程序. 为测试程序的有效性，选取乙酰胆碱酯酶抑制剂(AChE)、凝血酶抑制剂(thrombin-MQPA)及 HIV 蛋白酶抑制剂 3 种晶体复合物，进行晶体结构复原，并与广泛应用的分子对接程序 DOCK(Kuntz 研究组推出的最新版本 DOCK6.1)在能量得分、均方根偏差和对接所消耗的计算机时间方面进行了比较，得到了较为满意的结果.

4.1　凝血酶抑制剂晶体结构复原

凝血酶是与血液凝固有关的重要蛋白酶之一，它能水解 L-精氨酸的肽、酰胺和酯类. 临床表明，凝血酶抑制剂对血液栓塞、外伤出血等与血液凝固相关的疾病具有较好的疗效. 本文选取凝血酶(PDB：1ETR)复合物中配体 MQI 与其受体进行分子对接，其对接结果如表 3 及图 1 所示，表中 *energy* 为能量得分，单位 kJ/mol；*RMSD* 为晶体结构的均方根偏差，单位 nm；*time* 为对接时间，单位 s.

表 3　1ETR：对接结果与 DOCK 6.1 的比较

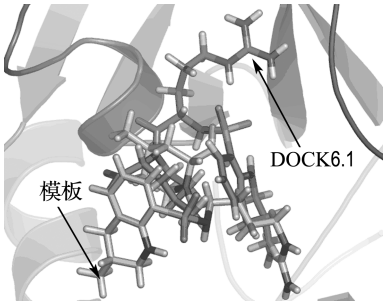
Tab. 3　1ETR: Comparisons of the docking results with DOCK 6.1

应用程序	*energy*/(kJ·mol^{-1})	*RMSD*/nm	*time*/s
DOCK 6.1	−259.41	0.513	15 242
本文	−1 846.31	0.025	2 523

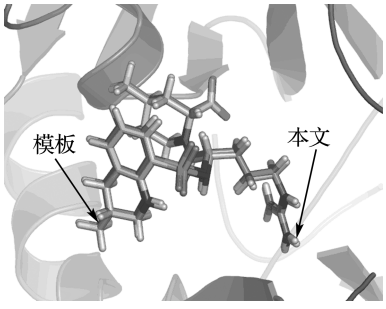
4.2　乙酰胆碱酯酶抑制剂(AChE)晶体结构复原

老年痴呆症 (alzheimer's disease，AD) 是一种多因异质性疾病，伴有认知及行为障碍，多发生在 65 岁以上的老年人群. 由于 AD 的病因病机尚未明确，目前对该病尚无特效药物. 目前从血液和

脑脊液中发现一些具有诊断和鉴别诊断意义的生化指标,有望成为 AD 早期诊断极有价值的指标,乙酰胆碱酯酶就是其中之一.



(a) DOCK6.1 对接结果与晶体结构比较



(b) 本文对接结果与晶体结构比较

图 1 1ETR:对接结果与晶体结构的比较

Fig. 1 1ETR: Comparisons of the docking results with crystal structure

本文运用改进的方法,对乙酰胆碱酯酶(PDB:1EVE)抑制剂晶体复合物中配体 E20 与其受体作对接,与 DOCK6.1 的对接结果相比较,结果如表 4 及图 2 所示.

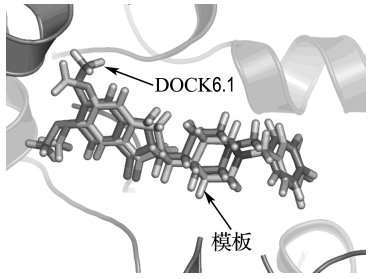
表 4 1EVE:对接结果与 DOCK 6.1 的比较

Tab. 4 1EVE: Comparison of the docking results with DOCK 6.1

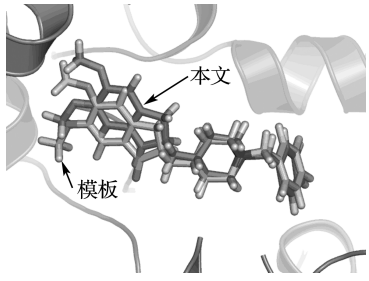
应用程序	*energy*/(kJ·mol^{-1})	*RMSD*/nm	*time*/s
DOCK6.1	−204.15	0.121	18 208
本文	−2 771.84	0.093	2 684

4.3 HIV 蛋白酶抑制剂晶体结构复原

人类免疫缺陷病毒(HIV)是艾滋病的主要致病因,针对艾滋病的化学药物治疗中 HIV 蛋白酶抑制剂发挥了重要作用.它通过抑制 HIV 在复制后期的构造蛋白、调节蛋白的功能,使其成为无外膜蛋白构造且不具感染力的病毒,从而达到治疗作用.本文选取 HIV 蛋白酶中的一种(PDB:1QBS),将其与配体 DMP 进行对接,其晶体结构与对接结果如表 5 及图 3 所示.



(a) DOCK6.1



(b) 本文

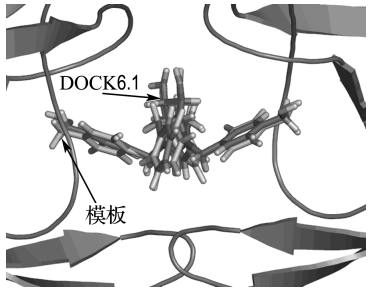
图 2 1EVE:对接的最优构象与晶体结构的比较

Fig. 2 1EVE: Comparisons of the optimal docking conformations with crystal structure

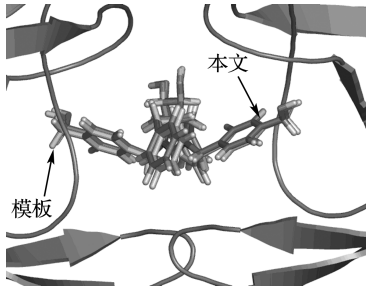
表 5 1QBS:对接结果与 DOCK6.1 的比较

Tab. 5 1QBS: Comparison of docking results with DOCK6.1

应用程序	*energy*/(kJ·mol^{-1})	*RMSD*/nm	*time*/s
DOCK6.1	−251.72	0.916	21 648
本文	−2 141.29	0.035	2 932



(a) DOCK6.1



(b) 本文

图 3 1QBS:对接的最优构象与晶体结构的比较

Fig. 3 1QBS: Comparisons of the optimal docking conformations with crystal structure

由上述 3 个实例可以看出，对于活性位点形成氢键或结合位点存在疏水性口袋的复合物（如 1ETR、1QBS），本文的方法精度远好于 DOCK6.1. 这是由于本文采取的打分函数并不单纯以力场作为衡量能量的标准，而是通过将原子对间的距离分布转化为受体与配体分子间的结合能，从而将难以用公式显性表达的氢键、疏水等结合过程中的力隐含在概率分布中，因而得到了更好的结果. 同时，3 个实例均表明，本文的方法在保证精度的前提下，效率优于 DOCK6.1 的结果.

5　结　语

打分函数的选取与搜索算法的改进是分子对接过程中较为重要的两个部分. 本文在传统对接程序 DOCK 的基础上，采用基于原子间概率分布的知识打分函数替代了基于力场的打分函数；同时采用基于信息熵的多种群自适应遗传算法，发展出一种新型对接程序 KGAsDock，通过算例证明该方法在保证效率的前提下，提高了计算的精度，得到了较为满意的结果.

参考文献：

[1] 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**(12):269-288

[2] EWING T J, MAKINO S, SKILLMAN A G, *et al*. DOCK4.0:Search strategies for automated molecular docking of flexible molecule databases [J]. **Journal of Computer-aided Molecular Design**, 2001, **15**(5): 411-428

[3] LI Hong-lin, LI Chun-lian, GUI Chun-shan, *et al*. GAsDock: a new approach for rapid flexible docking based on an improved multi-population genetic algorithm [J]. **Bioorganic & Medicinal Chemistry Letters**, 2004, **14**(18):4671-4676

[4] SIPPL M J. Boltzmann′s principle, knowledge-based meanfields and protein folding. An approach to the computational determination of protein structures [J]. **Journal of Computer-aided Molecular Design**, 1993, **7**(4):473-501

[5] MUEGGE I. PMF scoring revisited [J]. **Journal of Medicinal Chemistry**, 2006, **49**(20):5895-5902

[6] 李纯莲，王希诚，赵金城，等. 一种基于信息熵的多种群遗传算法 [J]. 大连理工大学学报，2004，**44**(4):589-593
(LI Chun-lian, WANG Xi-cheng, ZHAO Jin-cheng, *et al*. An information entropy-based multi-population genetic algorithm [J]. **Journal of Dalian University of Technology**, 2004, **44**(4):589-593)

A molecular docking method using knowledge scoring function

WANG Xi-cheng[*1], ZHAO Xiao-yu[2], KANG Ling[3], LI Hong-lin[4]

(1. State Key Laboratory of Structural Analysis for Industrial Equipment, Dalian University of Technology, Dalian 116024, China;
2. Department of Engineering Mechanics, Dalian University of Technology, Dalian 116024, China;
3. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China;
4. Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China)

Abstract: Molecular docking plays an important role in computer-aided drug molecular optimal design. A knowledge scoring function for estimating the binding free energy by means of distances between atom pairs is introduced, whose formula is similar to that of potential of mean force (PMF). Based on the knowledge scoring function, and combining an improved multi-population adaptive genetic algorithm based on information entropy, a new docking program KGAsDock is developed. The comparing with famous docking program DOCK6.1 is given, and the numerical results show that the method can improve the docking accuracy considerably without reducing computing efficiency.

Key words: molecular docking; knowledge scoring function; genetic algorithm; optimization model