

基于 Monte Carlo 局部增强的多模态优化算法

陈先跑 张贵军 秦传庆 郝小虎

(浙江工业大学信息工程学院 杭州 310023)

摘 要 高维构象空间搜索是蛋白质结构从头预测领域中一个亟需解决的关键问题。基于差分进化算法框架,提出了一种多模态蛋白构象空间优化算法。算法建立基于蛋白质空间特征向量的相似性测度指标,采用排挤更新策略,避免算法早熟,对蛋白质构象空间模态进行全局搜索;设计基于 Monte Carlo 局部搜索的片段组装方法,实现模态增强过程,有效平衡算法的收敛速度和种群多样性。采用 Rosetta 粗粒度能量模型,针对 5 种测试蛋白的实验结果表明, Monte Carlo 局部增强和蛋白质特征向量的相似性测度能够有效地提高算法的性能,与 Baker 小组和 Shehu 小组的研究成果相比,提出的算法能够达到较高的预测精度,并得到一系列的亚稳态稳定结构。

关键词 多模态,蛋白质结构从头预测,排挤差分进化算法,蛋白质结构特征向量,片段组装

中图法分类号 TP301.6 文献标识码 A

Local Monte Carlo Search Approach to Multimodal Problem in Protein Conformation Space Optimization

CHEN Xian-pao ZHANG Gui-jun QIN Chuan-qing HAO Xiao-hu

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract We elucidated the native structure of a protein molecule from its sequence of amino acids. A problem known as de novo structure prediction, is a long standing challenge in molecular biology. High dimensional conformational space search is the key issue of protein structure prediction that is needed to be solved. Based on differential evolution algorithm framework, we proposed a multimodal protein conformational space optimization algorithm to address the multiple-minima problem in decoy sampling for de novo structure prediction. Algorithm builds the index of similarity measure that is based on the vectors of features of proteins, using exclusion strategy to implement global search. Local minimum search strategy with fragment assembly is able to avoid premature convergence, and can balance the convergence rate and the diversity of the population. A greedy search maps a child conformation to its nearest local minimum, and the molecular fragment replacement technique and differential evolution algorithm help child jump out of local minimum, thus the algorithm can get better search ability. Using Rosetta coarse-grained energy model, results show that the additional mini-mization and the exclusion strategy based on conformation space are key to obtaining a diverse ensemble of decoys. Compared with Baker research team and Shehu research team, the proposed algorithm can achieve better prediction accuracy.

Keywords Multimodal, De novo structure prediction, Crowding differential evolution, Vector of protein structure feature, Fragment replacement

以计算机为工具,运用适当算法,从氨基酸序列出发直接预测蛋白质的结构,是当前分子生物学中一种主要的研究课题^[1-3]。基于 Anfinsen 提出的热力学假说^[4],蛋白质结构预测方法通常划分为以下 3 类:针对高相似序列的同源建模方法,针对较低相似序列的折叠识别方法^[5],以及不依赖模板而利用物理学原理直接进行计算的从头预测方法。从头预测方法有助于揭示蛋白质折叠机理,并且具有普遍性,是一种重要的结构预测方法。

已有的从头预测方法显示,在 Anfinsen 蛋白质一级结构决定其空间结构的理论^[2]前提下,要获得实用模型就必须考虑 3 个因素^[1,6]:(1)是否有一个精确的数学模型准确、真实地

反映氨基酸残基间的相互作用;(2)是否有一种能够在有限时间内找到全局最优结构的高效算法;(3)算法得到的全局稳定结构可能并不满足实际需求,算法在得到全局稳定结构的同时还要尽可能地得到局部稳定结构。蛋白质高维构象空间优化是蛋白质从头预测中的关键问题^[7,8],受当前计算能力的限制,粗粒度力场模型在蛋白质构象空间优化领域得到广泛应用,然而力场模型的不精确性造成计算所得的全局最优解可能并不对应蛋白质物理实验测定结构^[9]。设计一种多模态构象空间优化算法来解决粗粒度模型下的采样问题显得极为重要。

遗传算法(GA)^[10-13]、分子动力学模拟(MD)^[14-16]、Monte

本文受国家自然科学基金(61075062,61379020),浙江省自然科学基金(LY13F030008),浙江省科技厅公益项目(2014C33088),浙江省重中之重学科开放基金(20120811),杭州市产学研合作项目(20131631E31)资助。

陈先跑(1992—),男,主要研究方向为智能优化计算、生物信息学;张贵军(1974—),男,博士,教授,主要研究方向为智能信息处理、全局优化理论及算法设计、生物信息学,E-mail:zgj@zjut.edu.cn(通信作者)。

Carlo(MC)^[17-19]、构象空间退火(CSA)^[20-22]以及构象树指导搜索(CTGE)^[23,24]等随机优化算法在构象空间优化领域得到广泛应用,然而单纯的算法依旧不能有效地解决构象空间优化问题。近年来,基于知识的片段组装技术被广泛应用于构象空间的优化。2005年Baker基于Monte Carlo算法,辅以片段组装技术和能量极小化构象更新技术,其平均预测精度达 $1.6\text{\AA}$ ^[25];2009年Shehu提出构象树Monte Carlo片段组装搜索方法^[23];2011年Lee将动态片段组装技术引入构象空间退火中,得到CASP8中14个目标蛋白质的结构^[18];同年Shehu基于增强片段库进一步对算法作了改进^[24];2012年Zhang基于副本交换算法,利用片段组装技术开发出了国际领先的QUARK蛋白从头预测服务器^[19];2013年,Saleh等人提出一种基于局部极值点搜索的群体进化算法^[26]。

针对构象空间优化这个多模态问题,本文提出一种多模态构象空间优化算法(Fragment USR Monte Carlo Differential Evolution, FUMDE)。在基本差分进化算法的基础上,算法采用蛋白质特征向量相似性测度,首次将基于蛋白质特征向量空间的排挤策略引入多模态算法,结合局部Monte Carlo片段组装优化方法,对5种测试蛋白质进行实验。结果表明,本算法不仅能够达到较高的预测精度,而且能够得到一系列的亚稳态稳定结构。

1 结构片段库组装

片段库构建过程首先通过PISCES服务器^[27]以sequence similarity $\leq 30\%$, resolution $\leq 3.0\text{\AA}$ 和R-factor ≤ 0.3 为参数,对现有的蛋白质数据库搜索选择得到非冗余蛋白质子集,然后将得到的蛋白质子集中的蛋白质链分解成片段长度为L的片段,最后根据Rosetta片段计分函数^[28]从这些小片段中挑选出一部分构成查询序列位置特定结构片段库。构建片段库过程使用序列对比工具PSI-BLAST、二级结构预测服务器PsiPred、二级结构预测服务器Jufo、二级结构预测服务器SAM等工具对序列以及二级结构评分,根据配额筛选所需序列以及结构,以保持片段库的多样性。

蛋白质构象空间的高维特性、能量模型的不确定性,使得片段组装技术成为从头预测蛋白质结构的重要方法。片段组装过程指将优化目标蛋白质的某个片段与从蛋白质片段库中随机选择的相应位置的片段进行替换,即3种二面角(ϕ 、 φ 、 ω)的替换。首先在优化目标蛋白质上随机选取一个氨基酸 i ,即确定需要替换的蛋白质片段 $[i, i+L-1]$,L为蛋白质片段的长度,然后随机从片段库中选择 $3L$ 个匹配的二面角(ϕ 、 φ 、 ω)进行替换。蛋白质的空间结构具有一定的层次性和规律性,许多序列同源性较低的蛋白质也存在与目标蛋白质具有相关性的结构片段。片段组装技术弥补了同源建模方法必须使用具有很高同源性蛋白质作为模板的缺陷,利用现有的资源构建出合理的蛋白质结构模型,借鉴同源建模思想,片段组装能够有效地减小构象搜索空间,避免进化算法采样过程的盲目性,并且降低能量函数对局部作用的敏感性,结合差分进化算法的进化特性,可提高整体种群个体的质量,进而提高构象空间的采样效率。Kolodny研究团队研究表明,片段长度越长,则越需要一个大的蛋白质数据来保持构象的多样性,片段长度越短,则越容易得到新型的优良结构^[29]。综合考虑,本文L取3。图1是L取3时的片段组装示意图。

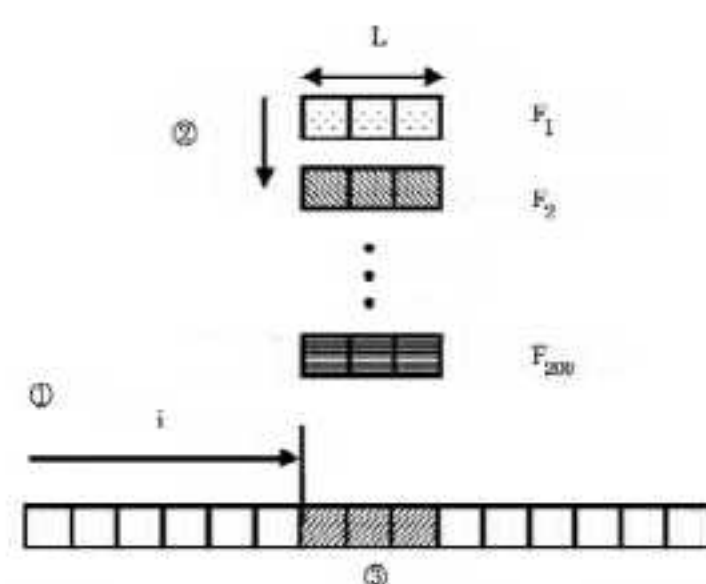


图1 片段组装示意图

2 Rosetta 粗粒度力场模型

理想的蛋白质结构预测方法是基于能量极小化的理论计算方法。因为蛋白质是大分子的复杂体系,其构象空间的自由度的数量级高达 $10^4 \sim 10^6$,如何降低计算复杂度,减小计算代价,成为解决蛋白质预测问题的关键。Rosetta粗粒度模型在保持重要结构信息的前提下,明确N、C、C_α和O 4种原子的信息,将每个侧链等效成一个位于质心位置的伪原子,有效减小了计算量。

Rosetta粗粒度力场模型不同于依赖于原子三维坐标的经验势能函数,是一种基于知识的力场模型。Rosetta能量函数是独立加权计算的线性和。采用的Rosetta粗粒度力场模型的能量函数表示形式如下^[30,31]:

$$E_{\text{protein}} = W_{\text{repulsion}} E_{\text{repulsion}} + W_{\text{attraction}} E_{\text{attraction}} + W_{\text{solvation}} E_{\text{solvation}} + W_{\text{bb-sc hb}} E_{\text{bb-sc hb}} + W_{\text{bb-bb hb}} E_{\text{bb-bb hb}} + W_{\text{sc-sc hb}} E_{\text{sc-sc hb}} + W_{\text{pair}} E_{\text{pair}} + W_{\text{dunbrack}} E_{\text{dunbrack}} + W_{\text{rama}} E_{\text{rama}} + W_{\text{reference}} E_{\text{reference}} \quad (1)$$

Rosetta从头预测算法分4个过程,每个过程采用不同的能量函数配置。本文采用Rosetta从头预测算法的第4过程所采用的配置——Score3配置,此配置考虑了10种不同的能量项,相对于其它3种配置更加详细。具体参数的配置参考文献^[30]。

3 基于蛋白质特征向量的相似性测度

蛋白质能量曲面上拥有大量的局部极小值,基本的差分进化算法极易陷入局部陷阱,设计一种排挤策略保持种群的多样性就显得极为重要。传统的排挤策略主要基于二面角及均方根偏差指标,以两蛋白质骨架结构的均方根偏差距离最小为目标,寻求最优的刚体转动平移参数来获得最好对应关系,用均方根偏差距离评价相似性,根据相似性排挤个体。但在刚体转动过程,刚体的二面角可能会发生未知的变化,导致比对结果的错误,且均方根偏差距离不是严格数学意义下的距离,不能满足三角不等式且依赖于比较结构的尺度,另外,均方根偏差计算代价很大。

基于蛋白质特征向量的相似性测度采用超快速形状特征识别方法^[32](Ultrafast Shape Recognition, USR),摒弃通过旋转平移寻到对应关系的做法,基于蛋白质的特征向量对蛋白质进行比对。基于Rosetta的笛卡尔坐标系,首先求得蛋白质分子中原子质心的坐标,然后求得离质心距离最近的原子的坐标(CTD);求得离CTD距离最近的原子的坐标(CST);求得离CTD距离最远的原子的坐标(FCT),求得离FCT距离最远的原子的坐标(FTF)。分别计算蛋白质分子粗粒度骨架模型中所有原子与4个特征点CTD、CST、FCT以及FTF的平均距离、平均距离的方差以及平均距离的偏差3个统计

量,即将每个蛋白质抽象为 12 维特征向量,通过归一化处理,根据式(2)将蛋白质之间的相似度用一个分值表示。

$$S_q = \frac{1}{(1 + \frac{1}{12} \sum_{i=1}^{12} |M_i^q - M_i^j|)} \quad (2)$$

式中, $\vec{M} = (\mu_1^{cd}, \mu_2^{cd}, \mu_3^{cd}, \mu_1^{ct}, \mu_2^{ct}, \mu_3^{ct}, \mu_1^{ft}, \mu_2^{ft}, \mu_3^{ft}, \mu_1^{cf}, \mu_2^{cf}, \mu_3^{cf}, \mu_1^{cf}, \mu_2^{cf}, \mu_3^{cf})$, 其中 μ_1, μ_2, μ_3 分别为蛋白质分子粗粒度模型中所有原子与 CTD、CST、FCT、FTF 4 个特征点的平均距离、平均距离的方差、平均距离的偏差。 S_q 为一个分数结果, $S_q \in [0, 1]$ 。

在差分进化算法的框架下,选择替换只在相似的优化目标之间进行,可以有效保持种群的多样性,从而弥补算法易过早收敛于局部极值点的缺陷。基于蛋白质特征向量的相似度测度能够大量降低模型维数,可有效解决估计空间复杂度问题。

4 算法描述

4.1 差分进化算法

1995 年 Price 和 Storn 提出了一种基于群体的启发式全局优化算法,即差分进化算法(DE),它具有高效、鲁棒的特性,可以求解非线性不可微连续的函数,并成功地应用到了多个科学领域。1996 年,在日本举行的 ICEO 国际会议上,众多实验结果证明 DE 是一种除确定性优化算法外收敛最快的群体进化算法^[33]。同时研究结果还表明,DE 类型的算法比 GA、PSO 类型的算法具有更好的全局搜索能力和局部增强能力。

差分进化算法不仅具有较强的全局搜索能力,还具有简单、通用和可并行处理等特点。但是在使用差分进化这种群体优化算法解决多模态优化问题时,由于其贪婪特性较强,算法只能收敛到全局最优解,因而丢失了众多局部极值解;其次,问题模型的复杂性也造成这些算法极易陷入某个局部解。单种优化算法总存在一些不可避免的缺点,如果将两种或多种优化算法融合到一起或在一种优化算法中引入其他优化算法的思想,则可以有效地扬长避短,既能发挥某种优化算法的优点,又能弥补其缺点,从而提高算法的各项性能。

4.2 局部 Monte Carlo 片段组装优化方法

由于力场模型的不精确性,求解优化目标的一个全局最优解或者多个局部最优解是蛋白质构象空间优化算法的重要目标。然而算法的收敛速度与种群多样性往往是相互矛盾的。采用局部 Monte Carlo 片段组装优化方法^[26],实现模态增强过程,结合基于排挤策略的差分进化算法的全局搜索特性跨越能量障碍,能够有效平衡算法收敛速度与种群的多样性,从而达到增强算法整体搜索能力的目的。局部 Monte Carlo 片段组装优化方法如下所示。

局部 Monte Carlo 片段组装优化方法

针对种群中测试个体 C_{newest} , 执行以下操作:

1. while reject-counter $\leq$ reject-times do
 - 1.1 对 C_{newest} 进行一次片段组装操作,得到一个中间变量 C_{tmp} 。
 - 1.2 如果 $f(C_{newest}) > f(C_{tmp})$, 则 $C_{newest} = C_{tmp}$, reject-counter = 0; 否则 reject-counter++。
2. end while

注:(1) reject-times 为优化目标蛋白质的序列长度;
(2) 初始的 reject-counter = 0。

如上面算法所示,局部 Monte Carlo 片段组装优化方法具体为:对一个优化目标进行一次片段组装,判断此优化目标的能量是否降低,如果能量降低,则用片段组装后产生的新个体取代原来的个体;否则,忽略此次片段组装,重新对优化目

标进行片段组装,再进行判断。算法不断迭代,终止条件为优化目标被连续拒绝 N 次(N 为优化目标蛋白质的序列长度),即优化目标已经达到或者近似达到局部极值点。贪婪的局部搜索结合差分进化算法可快速找到最优解。

4.3 基于 Monte Carlo 局部增强的构象空间优化算法

Rosetta 从头预测方法首先从已知的蛋白质结构的数据库中选择与目标蛋白质相关的蛋白质片段,然后随机组合这些蛋白质片段,形成一个粗粒度的蛋白质构象集,最后通过 Monte Carlo 模拟退火法将侧链的构象添加到目标蛋白质的骨干链上。Rosetta 从头预测方法作为一种国际领先的蛋白质结构预测方法,在历届 CASP 大赛取得了相当不错的成绩。其寻找最低能量形状的过程大致如下:

1. 从没有任何折叠的氨基酸序列开始。
2. 移动序列中的一部分,产生一个新的结构。
3. 通过能量模型计算新结构的能量。
4. 判断能量的变化来决定是否保留这次移动(否则淘汰)。
5. 迭代步骤 2—步骤 4,直到链中每一部分都得到足够多次数的移动。

上面的过程称为一条算法轨迹,每条轨迹的最终结果就是一个预测对应的结构。Rosetta 会保存每条轨迹中找到的最低能量结构。每条轨迹都是唯一的,因为每次尝试的移动方向都是随机决定的。

本文提出的蛋白质结构预测算法 FUMDE,针对 Rosetta 粗粒度力场模型,在片段组装的基础上,采用排挤更新策略,实现模态增强,结合差分进化算法强大的全局搜索能力对构象空间进行优化。算法 FUMDE 的流程描述如下:

FUMDE 算法

1. $t \leftarrow 0$
2. 初始化种群,从蛋白质片段库中随机选取片段产生 popSize 个种群个体 P_{int} ,并设置算法参数:种群大小 popSize,蛋白质序列长度 Length(即优化问题的维数),算法的迭代次数 T ,蛋白质片段的长度 L 。
3. 根据评分函数 f 计算每个种群个体的函数值大小,并记录。
4. while not termination condition do
 - 4.1 for 种群 P_{int} 中每个个体 P_i do
 - 4.1.1 设 $i = 1$, 其中 $i \in \{1, 2, 3, \dots, \text{popSize}\}$ 。
 - 4.1.2 其中 $P_{origin} = P_i$ 。
 - 4.1.3 随机生成正整数 rand1, rand2, rand3; rand1 $\neq$ rand2 且 $\in \{1, 2, \dots, \text{Length}\}$, 其中 rand3 $\in \{1, 2, 3, \dots, \text{popSize}\}$ 。
 - 4.1.4 针对个体 P_i 做变异操作,其中 $j \in \{\min(\text{rand1}, \text{rand2}), \dots, \max(\text{rand1}, \text{rand2})\}$ 。
 - a. 令 $P_{origin}.phi(j) \leftarrow P_{rand3}.phi(j)$;
 - b. 令 $P_{origin}.psi(j) \leftarrow P_{rand3}.psi(j)$;
 - c. 令 $P_{origin}.omega(j) \leftarrow P_{rand3}.omega(j)$ 。
 - 4.2 通过变异操作得到个体 C_{new} 。
 - 4.3 对个体 C_{new} 片段组装操作得到 C_{newest} 。
 - 4.4 对个体 C_{newest} 局部 Monte Carlo 增强操作得到 C_{new} 。
 - 4.4 对所得到的 C_{new} 执行选择操作,计算 C_{new} 个体与种群中的各个个体的相似性测度,找出与 C_{new} 空间最相似的个体 P_{sim} ,若 $f(C_{new}) > f(P_{sim})$, 则 C_{new} 替换 P_{sim} ; 否则保持种群不变。
 - 4.5 end for
5. 判断是否达到算法的终止条件(算法迭代执行 T 次),如若未达到,则 $t \leftarrow t + 1$, 转至第 1 步继续循环执行算法。
6. end while

- 注:(1)步骤 4.1.3 中随机数 $rand1, rand2, rand3$ 的选取中, $rand1 \neq rand2, rand3 \neq i$ (步骤 4.1 中的 i 值)
- (2)步骤 4.1.4 中氨基酸 j 值的大小在 $rand1$ 和 $rand2$ 之间。
- (3)步骤 4.1.4 中变异操作将 P_{origin} 的氨基酸 j 所对应的二面角 ϕ, ψ, ω 替换为 P_{rand3} 的相同位置所对应的二面角。
- (4)步骤 4.4 的局部增强操作即算法描述中 4.2 所描述的算法。

5 实验结果分析

本文采用 5 种蛋白质来测试算法的有效性。如表 1 所列, 1ENH 是一种 DNA 结合蛋白; 1GB1 是一种免疫球蛋白结合蛋白; 2JUI 是一种连接酶; 1GYZ 是一种核糖体蛋白; 4ICB 是一种钙结合蛋白。测试蛋白质下载于蛋白质数据库 (<http://www.rcsb.org/pdb/home/>)。算法运行于 Intel Core i3 CPU, 内存为 4GB 的 PC。算法实现语言为 C++。为了验证 FUMDE 算法的有效性, 文中对比了两种同类的算法, 即 FUDE 算法和 FDE 算法。FDE 算法采用片段组装方法, FUDE 算法采用相似性测度排挤更新策略以及片段组装方法。3 种算法采用相同的参数, 种群大小 $popSize=200$, 算法终止条件为循环迭代 50000 次。算法独立运行 20 次。

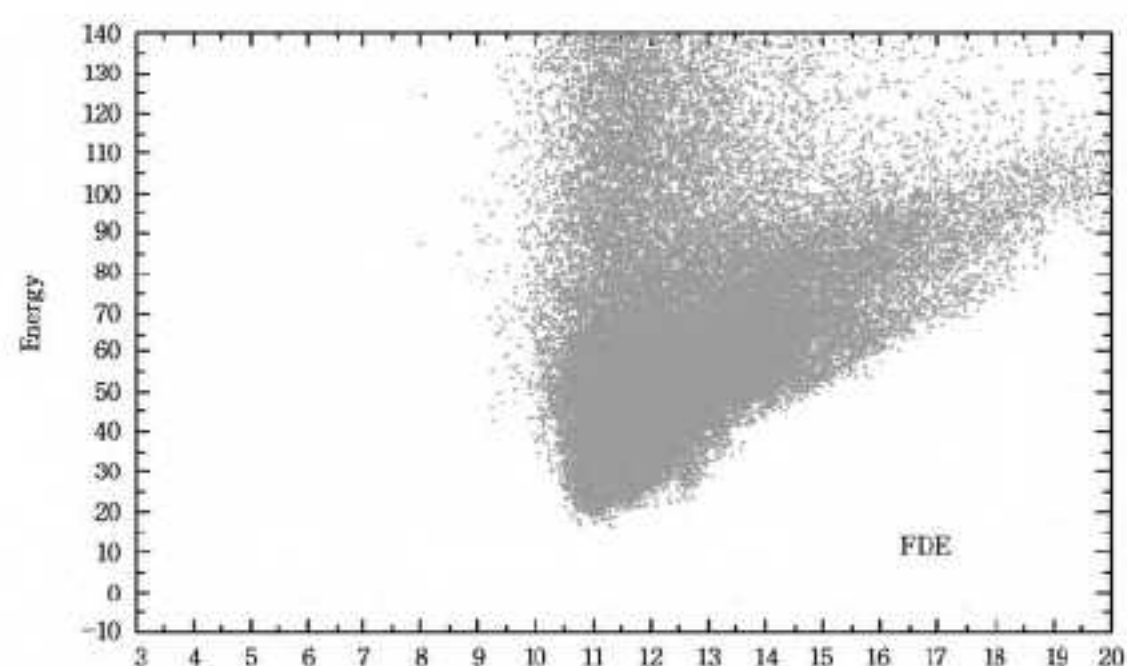
表 1 实验结果对比

PDB ID	Length	Fold	Minimum C_{α} (RMSD(Å))				
			FDE	FUDE	FUMDE	Shehu	Rosetta
1ENH	54	α	3.02	3.02	1.94	4.25	2.27
1GB1	56	α/β	6.99	6.02	3.10	6.89	1.56
2JUI	56	α	4.07	3.34	2.82	4.50	4.82
1GYZ	60	α	2.77	2.69	1.88	5.11	3.68
4ICB	76	α	3.74	3.49	2.91	5.76	3.23

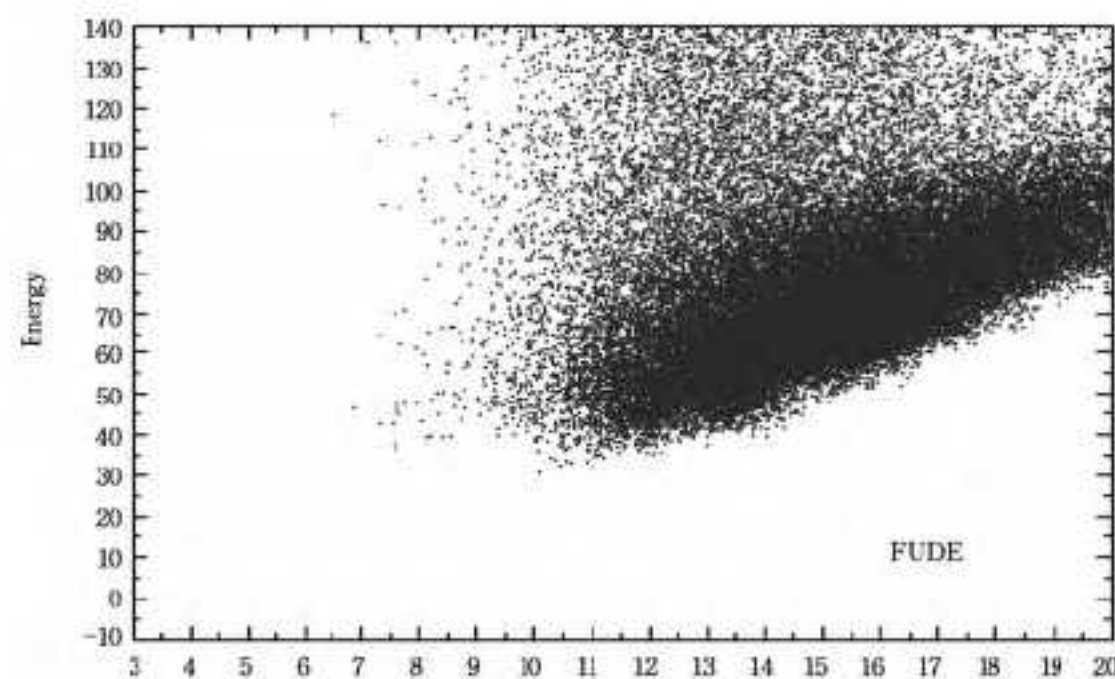
表 1 列出了 3 种算法独立运行 20 次所得到的实验结果, 并且与 Baker 小组和 Shehu 小组^[23,24]的研究成果(部分 Shehu 小组未测试蛋白质, 实现 Shehu 小组算法对其进行测试)进行了对比。对比本文提出的 3 种算法, 3 种算法均能得到精度较高的结果, 其中 FUMDE 算法能够达到相当高的预测精度。由于蛋白质 1GB1 有多种二级结构, FDE 算法与 FUDE 算法对于蛋白质 1GB1 的效果比其它 4 种蛋白质差, 这说明对于结构复杂的蛋白质, FDE 算法与 FUDE 算法的适应性较差, 不如 FUMDE 算法。对比 FDE 算法与 FUDE 算法, FUDE 算法效果比 FDE 算法略好, 这说明排挤策略在一定程度上改善了算法的搜索能力。对比 FUDE 算法与 FUMDE 算法, FUMDE 算法的效果相比于 FUDE 算法有较大改善, 这说明局部 Monte Carlo 片段组装优化方法能够有效平衡算法的收敛速度和种群多样性, 从而达到提高预测精度的目的。对比本文的结果与 Baker 小组的结果, 除了蛋白质 1GB1, FUMDE 算法略优于 Rosetta 从头预测算法; 对比与 Shehu 小组的结果, FUMDE 算法优于 Shehu 小组算法。Baker 小组的 Rosetta 从头预测算法采用片段组装技术与 Monte Carlo 优化方法, 算法复杂度远大于 FUMDE 算法, 而且 Rosetta 从头预测算法忽略个体与个体之间的联系, 计算所得的结果具有独立性, 算法尽管能达到较高的预测精度, 但也具有一定的不可靠性^[34]。FUMDE 算法在差分进化算法框架下, 有效利用个体与个体之间的重要信息, 能够达到较高的预测精度, 并得到一系列的亚稳态稳定结构。

图 2 是蛋白质 1GB1 算法效果图。FDE 算法得到的种群

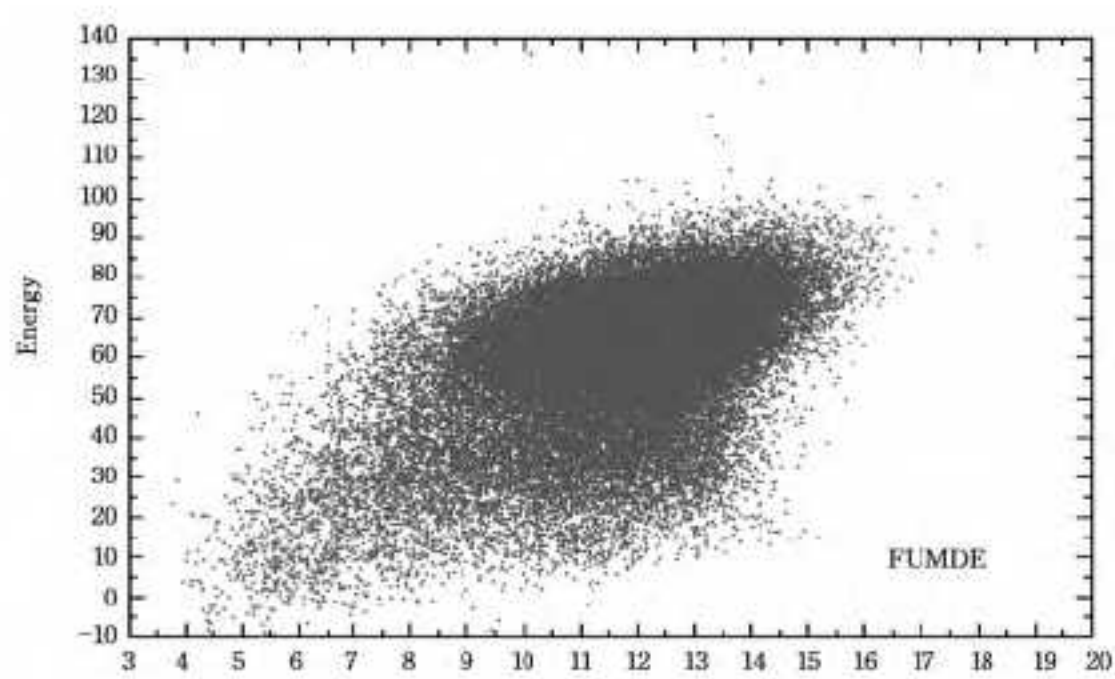
在坐标轴上呈单峰状, C_{α} 的平均均方根偏差为 12.67Å。可以得出, 算法收敛于某个局部极值点, 体现了普通差分进化算法早熟的弊端, 算法虽然在运行多次的情况下有时能够得到较好的结果, 但总体并不可靠。当面对多模态优化问题时, 需要定位优化问题的多个最优解, 普通差分进化算法的变异操作效果并不理想。FUDE 算法得到的种群相对于 FDE 的, 在坐标轴上整体左移且分布均匀。可以得出, 基于蛋白质特征向量空间的排挤策略在一定种群规模下可以有效维持种群多样性。作为多模态优化算法, 种群的多样性是评价多模态优化算法的重要指标, 可见 FUDE 算法相对于 FDE 算法有一定优越性。相比之下, FUMDE 算法得到的种群个体更接近坐标原点, 种群个体具有更低的均方根偏差和能量值。



(a) C_{α} RMSD to the Native Structure(Å)



(b) C_{α} RMSD to the Native Structure(Å)



(c) C_{α} RMSD to the Native Structure(Å)

图 2 蛋白质 1GB1 3 种算法效果

这表明, FUMDE 算法的效果相对于前两种算法有很大的改善, 局部增强以及排挤策略是算法性能增强的主要原因。

图 3 是 5 种测试蛋白质的算法效果对比图。右上角的点是 FDE 算法所得的种群个体, 中间的点是 FUDE 算法所得的种群个体, 左下角的是 FUMDE 算法所得的种群个体。整体来看, FUMDE 算法最接近于坐标轴原点, 算法效果最好。蛋白质 4ICB 算法效果图中, 出现突出的单峰, 表示这个单峰的能量更低但整体均方根偏差却相反。这表明, 由于力场模型

的不精确性,能量值大小并不是评判蛋白质结构是否稳定的唯一标准。FUMDE 算法能够得到一系列稳定的蛋白质结构,能够较好地解决构象空间优化这个多模态优化问题。

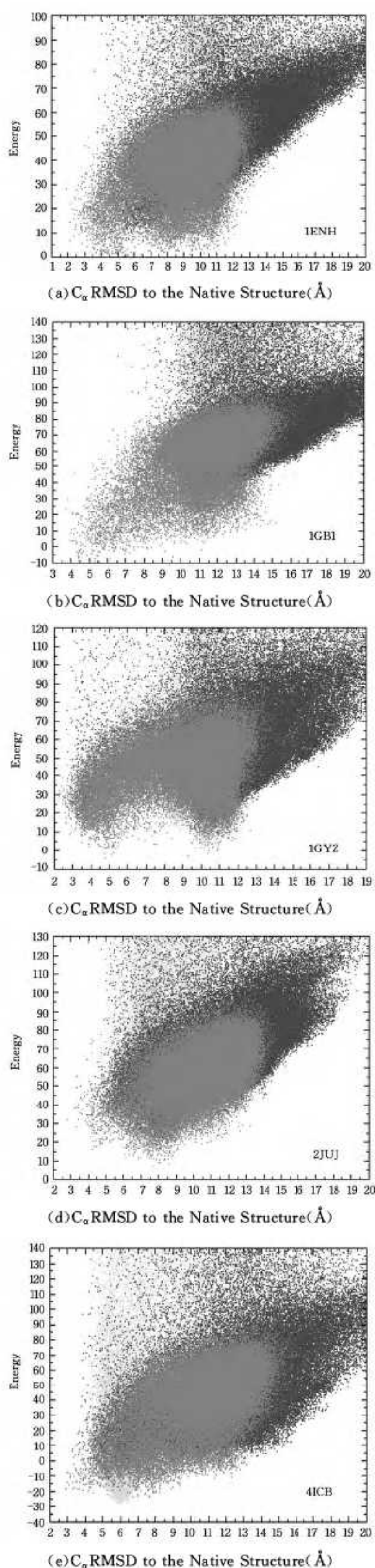


图 3 5 种测试蛋白质的算法效果

图 4 是 FUMDE 算法所得的最优稳定结构及两个次优结

构与物理实验结构的对比图。从图中可以发现,算法所得的最优稳定结构及两个次优结构与物理实验结构具有相当高的契合度,算法所得的 α 螺旋结构基本与物理实验所得结果吻合, β 转角的走势基本与物理实验所得结果相同,可见算法能够较好地预测蛋白质的二级结构以及三级结构。

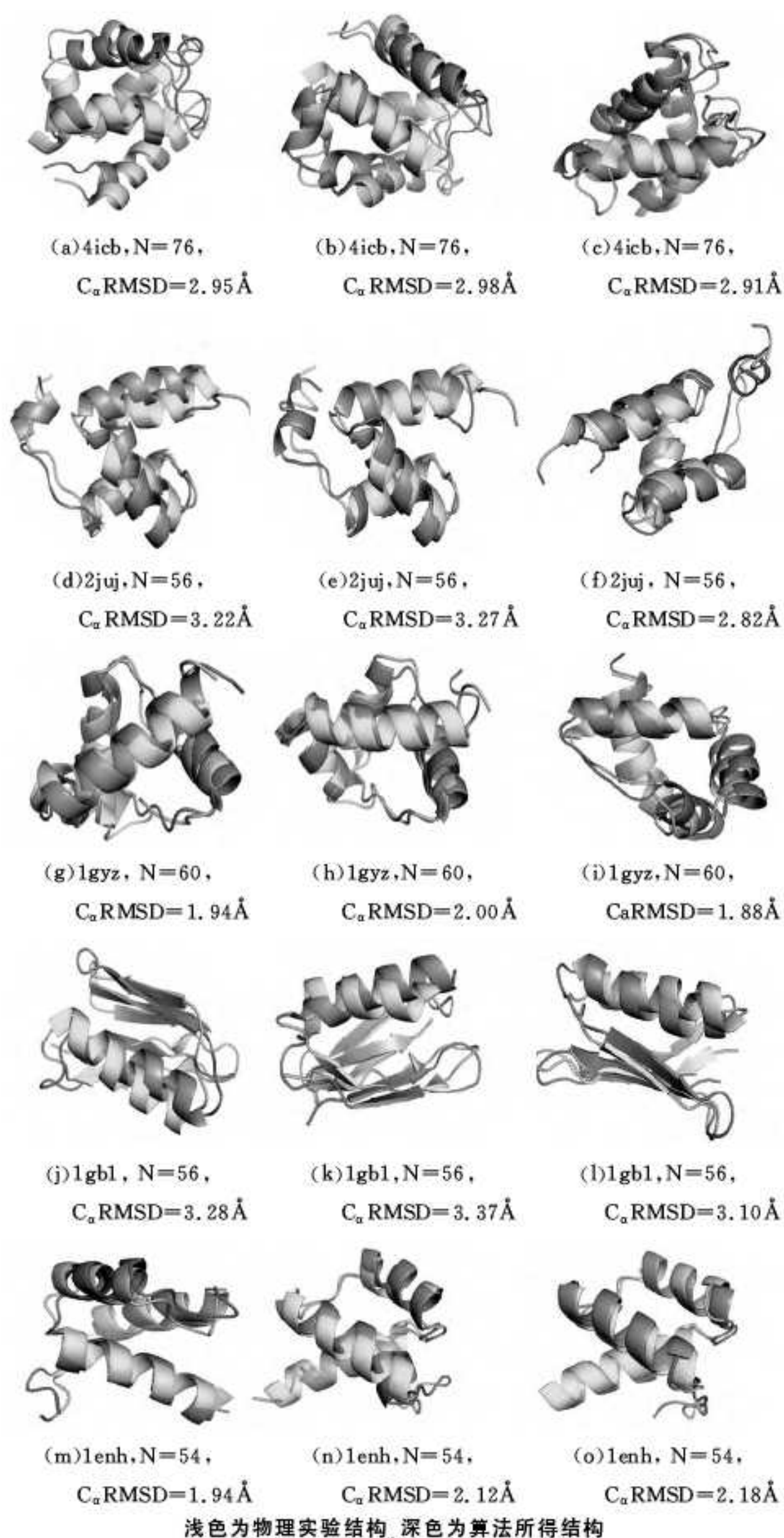


图 4 算法所得最优结构以及两个次优结构与实验结构的对比

结束语 本文提出了一种基于 Monte Carlo 局部增强的多模态构象空间优化算法,其采用蛋白质的粗粒度表示方法与片段组装技术。结果表明, FDE 算法因其早熟特性,通常会陷入局部极值点;FUDE 算法即使在保持种群多样性的情况下采用了片段组装方法降低了构象空间的复杂度,但搜索效率依然低;FUMDE 采用局部增强法,可有效从一个局部极值点跳到另一个能量更低的局部极值点,能够达到较高的预测精度并得到一系列亚稳态稳定结构,能够较好地解决构象空间优化这个多模态优化问题。通过与物理测定实验结果对比可发现,由于力场模型的不精确性,能量值并不能很好地代表蛋白质结构的稳定性,基于 Monte Carlo 片段组装优化方法的局部增强以及基于蛋白质特征向量空间的排挤策略是提

高采样效率的关键。我们可以从结果中挑选符合要求的结构进行侧链组装,以得到更高精度的蛋白质稳定结构。接下来将改进片段库以及基于蛋白质特征向量空间的排挤策略,同时选取更多的蛋白质,来验证所得结论。

参 考 文 献

- [1] Dill K A, MacCallum J L. The protein-folding problem, 50 years on[J]. *Science*, 2012, 338(6110): 1042-1046
- [2] Collins F, Patrinos A, Jordan E, et al. New goals for the US Human Genome Project, 1998-2003 [J]. *Science*, 282(5389): 682-689
- [3] Mitra P, Shultis D, Zhang Y. EvoDesign: de novo protein design based on structural and evolutionary profiles[J]. *Nucleic acids research*, 2013, 41(W1): W273-W280
- [4] Anfinsen C B. Principles that govern the folding of protein chains[J]. *Science*, 1973, 181(4096): 223-230
- [5] 黄俊峰, 段鹏, 吴文言. 基于模板的蛋白质结构预测[J]. *生物物理学报*, 2011, 27(1): 28-37
- [6] Baker D, Sali A. Protein structure prediction and structural genomics[J]. *Science*, 2001, 294(5540): 93-96
- [7] Bradley P, Misura K M S, Baker D. Toward high-resolution de novo structure prediction for small proteins[J]. *Science*, 2005, 309(5742): 1868-1871
- [8] Kim D E, Blum B, Bradley P, et al. Sampling Bottlenecks in De novo Protein Structure Prediction[J]. *Journal of molecular biology*, 2009, 393(1): 249-260
- [9] Saleh S, Olson B, Shehu A. A population-based evolutionary algorithm for sampling minima in the protein energy surface[C]// 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops(BIBMW). IEEE, 2012: 64-71
- [10] Hoque M T, Chetty M, Lewis A, et al. Twin removal in genetic algorithms for protein structure prediction using low-resolution model[J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2011, 8(1): 234-245
- [11] Tantar A A, Melab N, Talbi E G, et al. A parallel hybrid genetic algorithm for protein structure prediction on the computational grid[J]. *Future Generation Computer Systems*, 2007, 23(3): 398-409
- [12] Islam M K, Chetty M. Clustered memetic algorithm with local heuristics for ab initio protein structure prediction[J]. *IEEE Transactions on Evolutionary Computation*, 2013, 17(4): 558-576
- [13] Custódio F L, Barbosa H J C, Dardenne L E. A multiple minima genetic algorithm for protein structure prediction[J]. *Applied Soft Computing*, 2014, 15: 88-99
- [14] Duan Y, Kollman P A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution[J]. *Science*, 1998, 282(5389): 740-744
- [15] Lindorff-Larsen K, Trbovic N, Maragakis P, et al. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation[J]. *Journal of the American Chemical Society*, 2012, 134(8): 3787-3791
- [16] Scheraga H A, Khalili M, Liwo A. Protein-folding dynamics: overview of molecular simulation techniques[J]. *Annu. Rev. Phys. Chem.*, 2007, 58: 57-83
- [17] Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding[J]. *Proteins: Structure, Function, and Bioinformatics*, 2002, 48(2): 192-201
- [18] Lee J, Lee J, Sasaki T N, et al. De novo protein structure prediction by dynamic fragment assembly and conformational space annealing[J]. *Proteins: Structure, Function, and Bioinformatics*, 2011, 79(8): 2403-2417
- [19] Xu D, Zhang Y. Toward optimal fragment generations for ab initio protein structure assembly[J]. *Proteins: Structure, Function, and Bioinformatics*, 2013, 81(2): 229-239
- [20] Dotu I, Cebrian M, Van Hentenryck P, et al. On lattice protein structure prediction revisited[J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2011, 8(6): 1620-1632
- [21] Tyka M D, Jung K, Baker D. Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers[J]. *Journal of computational chemistry*, 2012, 33(31): 2483-2491
- [22] Joo K, Lee J, Sim S, et al. Protein structure modeling for CASP10 by multiple layers of global optimization[J]. *Proteins: Structure, Function, and Bioinformatics*, 2014, 82(S2): 188-195
- [23] Shehu A. An Ab-initio tree-based exploration to enhance sampling of low-energy protein conformations[C]// *Robotics: Science and Systems*. 2009: 241-248
- [24] Olson B, Molloy K, Shehu A. In search of the protein native state with a probabilistic sampling approach[J]. *Journal of bioinformatics and computational biology*, 2011, 9(03): 383-398
- [25] Bradley P, Misura K M S, Baker D. Toward high-resolution de novo structure prediction for small proteins[J]. *Science*, 2005, 309(5742): 1868-1871
- [26] Saleh S, Olson B, Shehu A. A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction[J]. *BMC structural biology*, 2013, 13(Suppl 1): S4
- [27] Wang G, Dunbrack R L. PISCES: a protein sequence culling server[J]. *Bioinformatics*, 2003, 19(12): 1589-1591
- [28] Gront D, Kulp D W, Vernon R M, et al. Generalized fragment picking in Rosetta: design, protocols and applications[J]. *PloS one*, 2011, 6(8): e23294
- [29] Kolodny R, Koehl P, Guibas L, et al. Small libraries of protein fragments model native protein structures accurately[J]. *Journal of molecular biology*, 2002, 323(2): 297-307
- [30] Handl J, Knowles J, Vernon R, et al. The dual role of fragments in fragment-assembly methods for de novo protein structure prediction[J]. *Proteins: Structure, Function, and Bioinformatics*, 2012, 80(2): 490-504
- [31] Renfrew P D, Choi E J, Bonneau R, et al. Incorporation of non-canonical amino acids into Rosetta and use in computational protein-peptide interface design[J]. *PloS one*, 2012, 7(3): e32637
- [32] Ballester P J, Richards W G. Ultrafast shape recognition for similarity search in molecular databases[J]. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 2007, 463(2081): 1307-1321
- [33] Storn R. Differential evolution design of an R-filter [C]// *Proceedings of IEEE International Conference on Evolutionary Computation*, 1996. Nagoya, 1996: 268-173
- [34] Leaver-Fay A, Tyka M, Lewis S M, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules[J]. *Methods Enzymol*, 2011, 487: 545-574