基于副本交换的局部增强差分进化蛋白质结构从头预测方法

李章维 郝小虎 张贵军
(浙江工业大学信息工程学院 杭州 310023)

摘 要 针对蛋白质高维构象空间搜索问题,提出一种基于副本交换的局部增强差分进化蛋白质结构从头预测方法(RLDE)。首先,采用基于知识的Rosetta粗粒度能量模型显著降低构象空间优化变量维数;其次,引入基于片段库知识的片段组装技术进一步减小构象搜索空间,有效避免搜索过程中的熵效应;此外,在每个副本层设置构象种群,采用差分进化算法对种群进行更新,然后利用Monte Carlo算法对种群做局部增强,以此得到全局和部分局部最优构象。综上,RLDE利用差分进化算法较强的全局搜索能力可以对构象空间进行有效的全局搜索;借助Monte Carlo算法局部搜索性能对构象空间局部极小区域进行更为充分的采样;副本交换策略保证了副本层中种群的多样性,同时能够增强算法跳出局部极小的能力,从而使得算法对构象空间的搜索能力进一步增强。15个目标蛋白测试结果表明,所提方法能够有效地对构象空间采样,得到高精度的近天然态蛋白质构象。

关键词 从头预测,蛋白质结构预测,副本交换,Monte Carlo,片段组装,差分进化算法
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.05.038

Replica Exchange Based Local Enhanced Differential Evolution Searching Method in Ab-initio Protein Structure Prediction

LI Zhang-wei HAO Xiao-hu ZHANG Gui-jun
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract To address the searching problem in high-dimensional protein conformational space, a replica exchange based local enhanced differential evolution searching method in ab-initio protein structure prediction (RLDE) was proposed. In this paper, the knowledge-based coarse-grained energy model, Rosetta, was employed to considerably reduce the optimal variable dimension of protein conformational space; the knowledge-based fragment assembly technique was introduced to further reduce the dimension of protein conformational space. Thus the entropy effect caused by searching in high-dimensionality conformational space could be avoided. Additionally, a conformation population was put into every replica layer, differential evolution algorithm was adopted to update the population in each layer, and the updated populations were then enhanced by Monte Carlo method. As a consequence, the global optimal conformation and a series of metastable conformations were generated. In conclusion, RLDE can effectively search the global conformational space through the strong global searching ability of differential evolution algorithm. The well local searching performance of Monte Carlo is also employed to sample the local minimum area adequately. Replica exchange strategy ensures the diversity of population in replica layers, and the capacity of algorithm to jump out of local minimum is enhanced as well, thereby makes the searching ability further heightened. Test results of 15 target proteins show that the proposed method can generated high-resolution near-native protein conformations by searching the conformational space effectively.

Keywords Ab-initio prediction, Protein structure prediction, Replica exchange, Monte Carlo, Fragment assembly, Differential evolution algorithm

蛋白质分子在生物细胞化学反应过程中起着至关重要的作用,它们的结构和生物活性状态对于理解和治愈由多种由蛋白质结构改变而引起的疾病具有重要的意义,因此获得蛋白质的三维结构就显得尤为重要[1]。目前,蛋白质结构数据库[1]中通过各种实验手段测定的天然态结构的蛋白质数目与UniProtKB/TrE MBL数据库[2]中已知一级结构的蛋白质数

[1] http://www.rcsb.org

[2] http://www.ebi.ac.uk/uniprot/TrEMBLstats

到稿日期:2016-04-01 返修日期:2016-08-01 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)资助。

李章维(1967-),男,博士,副教授,CCF会员,主要研究方向为智能信息处理,E-mail:lzw@zjut.edu.cn;郝小虎(1990-),男,博士生,主要研究方向为智能信息处理、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,E-mail:zgj@zjut.edu.cn(通信作者)。

目差距悬殊，据 2016 年 3 月统计，前者仅占0.186%；并且实验手段测定蛋白质三维结构难度大、耗时长、花费高，成为制约生命科学、计算生物学等学科发展的重要因素。因此，根据 Anfinsen 原则[2]，以计算机为工具，设计有效的算法，直接从氨基酸序列出发预测蛋白质的结构，成为近几十年来计算生物学领域中一个重要的研究课题。

蛋白质结构预测问题自提出的几十年来，不断有新的方法被提出。总体来讲，蛋白质结构预测方法可以分为以下 4 类[3]：1)不基于蛋白质数据库信息的第一性原则预测方法；2)基于蛋白质数据库信息的第一性原则预测方法；3)基于折叠识别和穿线的预测方法；4)比较建模和序列比对的预测方法。它们在不同的条件下适用，如：序列同源性大于 50%时，采用后两种方法可以达到 1 Å 左右的预测精度，而在序列同源性小于 30%的情况下其效果并不理想[4-5]；而对于序列同源性小于 20%的情况，前两种方法(也称为从头预测方法)则是唯一的选择[6]，并且从生物学意义上来讲，从头预测方法有助于揭示蛋白质的折叠机理，进而全面阐释生物学中心法则中第二遗传密码的理论部分。

以上方法也不可避免地存在一些缺陷：第一性原则预测方法不使用任何先验知识，仅从蛋白质氨基酸序列出发预测蛋白质三维结构，面临着庞大的构象搜索空间，并且随着序列长度的增加，空间搜索维数将呈指数增长；基于蛋白质数据库的第一性原则预测方法使用已知结构的蛋白质信息，利用片段比对来组装蛋白质三维结构，有效减小了搜索空间的维数，但是仍然面临着较为巨大的搜索空间，并且如何避免片段置换引起的局部势能增加也是一个具有挑战性的问题；基于模板的预测方法尽管在某些情况下能够取得较高的预测精度，但是由于使用了模板信息，使得其只能够预测与 PDB 库中已知结构相似性较高的蛋白质，不利于发现新的折叠类型和结构，同时基于模板的预测方法也不利于了解蛋白质折叠机理，即蛋白质是如何在瞬间从无序状态折叠成为具有特定功能的三维结构，并在折叠过程中精细地保持平衡[7]。

针对上述问题，作为普适性更强的从头预测方法被广泛采用。尤其是在 CASP(Critical Assessment of Techniques for Protein Structure Prediction)竞赛的推动下，不同的蛋白质结构从头预测方法被提出，在广阔的构象空间进行搜索，执行能量极小化过程获得蛋白质三维预测结构。如：遗传算法(GA)[8-11]、差分进化算法(DE)[12-14]、分子动力学模拟(MD)[15-17]、蒙特卡洛方法(MC)[18-20]、构象空间退火(CSA)[21-23]、副本交换方法(REM)[24-27]、格点系统搜索(SGS)[28-30]、分支定界(BB)[31-33]、构象树指导搜索(CTGE)[34-36]等。此外，Rosetta[37]，QUARK[38]从头预测方法在历届 CASP 赛事中表现不俗，已经成为当今国际领先的从头预测服务器。

尽管很多行之有效的构象空间搜索方法不断地被提出，但是蛋白质构象空间的高维特性和能量表面的崎岖性，使得算法在如此广阔的构象空间中进行搜索时仍然面临巨大的困难，往往会陷入局部极值解而造成算法早熟收敛[39-40]的问题；同时，对构象空间的采样不够充分也是一个值得关注的方面。因此，对构象空间进行有效的降维处理，设计更为有效的搜索算法对构象空间进行充分的采样，是解决蛋白质三维结构预测瓶颈问题的途径[41]。本文提出一种基于副本交换的局部增强差分进化蛋白质结构从头预测方法(RLDE)。其采用基于知识的粗粒度能量模型 Rosetta Score3 和片段组装技术显著降低构象空间的维数，而又不失去主要的结构信息，有效避免了搜索过程中的熵效应；与常规的副本交换方法不同，RLDE 用构象种群取代副本层中单独的构象，增加了构象空间中采样的多样性；采用全局搜索性能较好的差分进化算法对种群进行更新，然后利用 Monte Carlo 算法的局部搜索特性对种群做局部增强，副本交换的策略同时可以增强算法跳出局部极值解的能力，从而得到全局和一系列局部最优构象。

1 理论分析

1.1 粗粒度能量模型

蛋白质折叠问题本质上是典型的多尺度问题[42]。自从 Levitt 及 Warshel 于 1975 年建立蛋白质分子粗粒度能量模型以来，陆续有研究者建立了一系列从粗粒度到全原子尺度的多尺度能量模型[43-44]，即利用粗粒度能量模型快速优化得到“诱饵构象”，然后进一步基于精度更高的全原子能量模型对其进行修正，得到精度更高的近天然态构象。

RLDE 依据粗粒度蛋白质表达模型，采用基于知识的 Rosetta Score3 粗粒度能量模型。粗粒度蛋白质表达模型在不丢失氨基酸序列重要结构信息的前提下只保留 N,C,Cα,O 这些主链骨干原子及侧链替代原子，用一系列设置为 $-120°$ 到 $120°$ 的二面角 φ(N—Cα)，ψ(Cα—C)表示氨基酸链，有效减小了计算空间的复杂度[45]，降低了构象空间优化变量维数，为计算带来了方便。Rosetta Score3 粗粒度能量模型是 10 种不同能量项的独立加权线性和[46-47]，是一种基于知识的能量函数，它隐式地体现了形成蛋白质天然结构的内在作用，计算成本较低，而且非常有效[48]。

1.2 副本交换

副本交换是一种有效的构象空间采样方法，尤其是在构象空间存在大量高能势垒的情况下，副本交换表现出比常规 MD 和 MC 方法更强的构象空间搜索能力[49]。副本交换的基本思想是：N 个体系相同的副本在不同温度下执行常规 MD 或 MC 方法，每隔一段时间(或者一定步数)之后，根据设定的交换概率依次对不同温度下的副本个体进行交换。这样可以实现在相应的温度下以“任意步长”搜索构象，进而在整个能量空间实现以“任意步长”对构象进行搜索。在典型的计算机模拟时间尺度下，高温系统通常能够采样到更大的相空间(系统所有可能状态的空间)，而低温系统虽然在相空间局部区域进行精确采样，但可能陷入局部极值解[50]。而副本交换通过交换不同温度间的完整构象，在低温体系引入高温体系的模拟可以确保低温体系也能够采集到目标温度区域具有代表性的样本，从而达到对构象空间进行充分采样、增强算法跳出局部极值区域能力的目的。副本交换方法尽管表现出空间采样的优势，但是由于采用常规的 MD/MC 方法进行模拟时，在构象空间中搜索形成的是单独的 Markov 链，采样得到的构象的多样性比较有限，并且其性能取决于所选用力场模型的精确性[50]。

与现有文献中描述的副本交换方法不同，RLDE 在副本交换中引入种群的概念，将单独的 MD/MC 模拟个体用种群

代替，以增加搜索过程中采样的多样性；采用全局搜索性能较好的 DE 算法代替 MD/MC 模拟过程，对不同温度下的副本种群进行更新，进一步增强算法的全局搜索能力；同时，借用 MC 方法的局部搜索性能对更新的种群进行局部增强，从而得到一系列局部最优的亚稳态构象；当不同温度下的副本种群更新之后，按照概率 $P(A\leftarrow B)=\exp((1/kT_B-1/kT_A)\cdot(E_B-E_A))$ 对相邻温度层之间的副本种群进行交换。其中 A,B 是要进行交换的两个构象；T_A,T_B 为 A,B 所在副本层的温度；E_A,E_B 为 A,B 的能量值；k 为玻尔兹曼常数。副本交换仅在相邻温度层之间进行，是因为 1）副本交换的接受概率会随着温度差值的增加急剧下降，上述概率公式也表明了这一点；2）这样可以降低算法的复杂度，有利于搜索过程的加速，而又不会对算法的搜索性能造成太大的损失。

1.3　片段组装技术

鉴于蛋白质构象空间的高维特性、模型的多尺度特性及不精确性，片段组装技术的应用成为蛋白质结构从头预测的一种重要手段。片段组装技术首先将预测序列划分成若干连续区段，然后通过序列比对寻找局部拟合已知蛋白质结构的片段，与优化目标蛋白质的指定片段进行替换，即 3 种二面角 ϕ,ψ,ω 的替换，最后组装成完整的目标结构。片段组装有 3 个关键过程：1）选择开始插入的位置 i；2）从片段库中选择用于该位置组装的片段；3）从选定的片段中选择插入长度 L，并替换目标序列中的相应片段。

通过片段组装，一方面可以极大地减小搜索空间，从而显著地提高计算速度；另一方面，由于蛋白质三维结构具有一定的层次性和规律性，许多序列同源性很低的远亲蛋白质也存在相似的结构片段折叠模式，因此通过片段组装可以构建更为合理的蛋白质三维结构。在现有技术水平的条件下，采用单纯的优化技术，目前只能得到 5～20 左右长度的多肽链公认稳定构象；而引入片段组装，则可以得到 150 序列长度的高精度预测结构[20]。

1.4　基于片段组装的 Monte Carlo 局部增强方法

DE 是一种基于群体的启发式全局优化算法，它是除确定性优化算法外收敛最快的群体进化算法[51]。DE 不仅具有较强的全局搜索能力，还具有简单、通用、可并行处理等特点。但是，由于较强的贪婪特性，其在求解多模态函数时往往会使得算法只收敛到全局最优解，而丢失了众多局部极值解。与 DE 相对应，MC 具有较强的局部搜索能力，广泛应用于蛋白质构象优化问题。MC 算法首先产生一个随机构象，对构象结构进行改变，生成一个新的构象，通过计算两个构象的能量差值，按照设定温度下的 Boltzmann 概率接收新产生的构象，迭代地运行该过程，直至得到能量最低的构象。假定算法各态遍历产生的 Markov 过程将收敛到正则分布[52]。

RLDE 采用副本交换的策略来搜索构象空间，而又不同于常规副本交换方法：在每一个温度层下使用 DE 算法对副本种群进行更新，并调用 MC 方法对种群进行局部增强，然后做副本交换操作。在每一个温度层进行的操作称为基于片段组装的 Monte Carlo 局部增强。MC 算法良好的局部搜索性能使得算法能够得到局部能量最低的亚稳态构象，与具有较强全局搜索能力的 DE 算法相结合，可以对构象空间进行更为有效的搜索采样。

1.5　RLDE 算法

RLDE 采用基于知识的粗粒度能量模型 Rosetta Score3 降低构象空间维数，采用片段组装技术可以极大地减小搜索空间，提高算法收敛速度，并且能够有效地避免搜索过程中的熵效应；与常规的副本交换方法不同，用构象种群取代副本层中单独的构象，增加了构象空间中采样的多样性；采用全局搜索性能优良的差分进化算法对种群进行更新，然后利用 Monte Carlo 算法的局部搜索特性对种群做局部增强，副本交换的策略同时可以增强算法跳出局部极值解的能力，从而得到全局和一系列局部最优构象。

RLDE 算法描述下。

算法 1　RLDE 算法

1. Input ← 序列信息
2. 参数设置：副本层数 RE，副本层温度参数 kT，种群大小 popSize，算法的迭代次数 iter，交叉因子 CR，片段长度 L
3. 种群初始化：在每个副本层生成 popSize 个种群个体 P_{init}
4. Try：
   - 4.1 for l in range(0,RE)：
     - 4.1.1 for 个体 P_i in 种群 P_{int}：
       - 4.1.1.1 令 $P_{target}=P_i$，其中 i∈{1,2,3,…,popSize}
       - 4.1.1.2 随机生成正整数 randl，rand2，rand3，其中 randl∈{1,2,3,…, popSize}，$randl\neq i$，$rand2\neq rand3$，rand2，rand3∈{1,2,…,Length}
       - 4.1.1.3 针对个体 P_j 做变异操作，其中 j=randl。令 a=min(rand2,rand3)，b=max(rand2,rand3)，k∈[a,b]
       - 4.1.1.4 for k in range(a,b)：
         - a. 令 P_{target}. phi(k)←P_j. phi(k)
         - b. 令 P_{target}. psi(k)←P_j. psi(k)
         - c. 令 P_{target}. omega(k)←P_j. omega(k)
       - 4.1.1.5 end for
       - 4.1.1.6 通过变异（片段组装）得到测试个体 P_{trial}
       - 4.1.1.7 生成随机数 rand4 和 rand5，rand4∈(0,1)，rand5∈(1,Length)
       - 4.1.1.8 根据下式执行交叉过程：

$$P_{trial}=\begin{cases}P_{trial,rand5}\leftarrow P_{target,rand5}, & \text{if (rand4}\leqslant\text{CR)}\\ P_{trial,rand5}, & \text{otherwise}\end{cases}$$

       - 4.1.1.9 计算 P_{target} 和 P_{trial} 的能量：$E(P_{targe})$ 和 $E(P_{trial})$
       - 4.1.1.10 若 $E(P_{target})>E(P_{trial})$，则用 P_{trial} 替换 P_{target}，否则保持种群不变
     - 4.1.2 end for
     - 4.1.3 得到更新种群 P_{update}
     - 4.1.4 for 个体 P_i in 种群 P_{update}：
       - 4.1.4.1 调用 MC 方法对个体做局部增强
       - 4.1.4.2 计算增强过程中产生的构象的能量 E(MC)
       - 4.1.4.3 若 $E(P_i)>E(MC)$，则更新种群，否则保持种群不变
     - 4.1.5 得到局部增强后的种群 $P_{enhance}$
     - 4.1.6 end for
   - 4.2 end for
   - 4.3 for m in range(0,popSize)：
     - 4.3.1 for n in range(0,RE)：
       - 4.3.1.1 在第 n，n+1 个副本层中各随机选择一个个体 P_A,P_B
       - 4.3.1.2 计算 P_A,P_B 的能量 E_A,E_B
       - 4.3.1.3 根据以下公式进行副本交换：

$$P(A\leftarrow B)=\exp((1/kT_B-1/kT_A)\cdot(E_B-E_A))$$

4.3.2 end for

4.4 end for

4.5 if end condition

4.6 Yes, goto 5, No, return 4

5. Finally:

5.1 end

1.6 算法复杂度分析

基本 DE 算法的时间复杂度为 $O(NP \cdot D \cdot G_{max})$[53]，其中 NP，D，G_{max} 分别是种群规模、优化变量维数和迭代次数。RLDE 算法相对于基本 DE 算法增加的计算复杂度主要来自副本交换和局部增强过程，根据算法描述可以计算如下：步骤 4.1.4 对种群个体进行局部增强，时间复杂度为 $O(NP \cdot N_{MC} \cdot G_{max} \cdot N_{RE} + NP \cdot G_{max} \cdot N_{RE} + NP \cdot G_{max} \cdot N_{RE})$，其中 N_{MC} 为调用 MC 进行局部增强的次数，N_{RE} 为副本层数；步骤 4.3 执行副本交换操作，时间复杂度为 $O(3NP \cdot N_{RE} \cdot G_{max})$。因此，RLDE 所增加的时间复杂度为 $O(NP \cdot N_{MC} \cdot G_{max} \cdot N_{RE} + 5NP \cdot G_{max} \cdot N_{RE})$。由此可以看出，RLDE 并没有显著增加计算的复杂度。

2 实验和结果

2.1 实验环境及测试材料

在蛋白质结构预测中，评价一种预测方法优劣的标准是其通过计算捕获蛋白质天然态结构的能力[54]。基于 Rosetta 平台，采用 Python 语言实现 RLDE 算法，选取折叠类型为 α，β，α/β 且序列长度从 32 到 106 的 15 种蛋白质对 RLDE 进行测试，并与 Rosetta 算法进行比较。这些测试蛋白分别是：1VII，1ENH，2JUJ，1GYZ，2MU2，1AIL，4ICB，2EZK，3GWL，2MRF，1FD4，1GB1，1AOY，2MIT，1I6C，它们从蛋白质结构 PDB 库[1]下载得到。选取这些蛋白质做测试是因为它们在生物体内有着重要的作用，并且它们的三级结构已经由实验方法测定，被广泛用于蛋白测试。相关测试蛋白的信息如表 1 所列。

片段库通过序列比对的方法获得，通过 PISCES 服务器[2]，以分辨率小于 2Å、同源性小于 30% 为参数，从筛选得到的 8096 条序列中进行比对搜索，在查询序列的每个片段位置产生200个得分最高的片段，将这些片段的二面角等信息记录下来，构成片段库文件。在执行片段组装时，片段库中的片段将被随机选择，ϕ，ψ，ω 这 3 个二面角将替换目标序列中相应位置的二面角，从而得到新的构象用于构象空间中执行搜索过程。

RLDE 的运行环境为搭载 Inter Core i7 处理器，16GB 运行内存的 64 位 Windows7 操作系统。算法参数设置如下：$RE=8$，$kT=[0.67, 0.72, 0.95, 1.14, 1.36, 1.63, 1.95, 2.33]$，$popSize=30$，$iter=10000$，$CR=0.5$，$L=3$，算法独立运行 30 次。

表 1 测试蛋白信息及测试结果

No	PDB ID	Len	Fold	Min RMSD/Å		Mean±std RMSD/Å	Replica Exchange Rate/%
				RLDE	Rosetta		
1	1VII	36	α	1.51	1.86	1.65±0.07	17.92
2	1ENH	54	α	1.24	2.01	1.47±0.16	18.62
3	2JUJ	56	α	2.86	4.13	3.34±0.19	18.08
4	1GYZ	60	α	1.68	2.14	1.96±0.17	17.59
5	2MU2	71	α	2.01	3.37	2.15±0.08	16.53
6	1AIL	73	α	3.31	5.11	3.95±0.38	15.53
7	4ICB	76	α	2.44	4.21	2.63±0.14	15.52
8	2EZK	99	α	3.02	5.09	3.57±0.43	15.70
9	3GWL	106	α	4.11	7.49	5.21±0.56	13.41
10	2MRF	33	α/β	1.82	2.53	1.95±0.10	20.09
11	1FD4	41	α/β	3.06	5.45	4.27±0.58	16.28
12	1GB1	56	α/β	3.05	6.27	3.84±0.39	12.86
13	1AOY	78	α/β	3.07	5.25	3.91±0.48	14.48
14	2MIT	32	β	3.98	5.05	4.68±0.26	16.18
15	1I6C	39	β	3.46	4.21	3.79±0.24	16.20

2.2 实验结果

部分蛋白的测试结果如图 1 所示，图中横坐标为预测结构和实验方法测定结构的相似度指标 RMSD 值，单位为 Å；纵坐标为计算得到构象的能量分值，颜色较浅的点表示 RLDE 搜索过程中产生的所有构象，颜色较深的点表示采用 Rosetta 算法计算产生的构象，Rosetta 算法根据文献[37]实现。部分蛋白质预测结构和实验方法测定结构的三维相似性对比如图 2 所示，图中颜色由深到浅的结构分别表示实验室测定结构、采用 RLDE 算法所得到的预测结构、采用 Rosetta 算法得到的预测结构。测试蛋白信息及由不同算法计算得到的预测结构和实验方法测定的结构的最小相似度指标 RMSD 值如表 1 所列。表 1 中最后两列分别表示 RLDE 算法搜索得到构象的平均值标准差、副本交换率(副本交换次数占总构象数的百分比)。

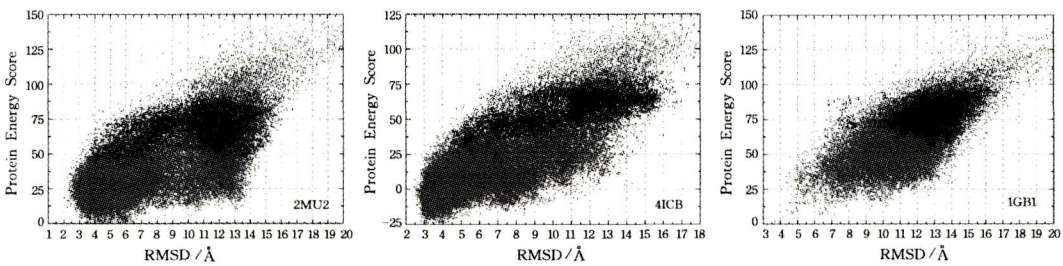

图 1 算法搜索过程示意图

1) http://www.rcsb.org

2) http://dunbrack.fccc.edu/PISCES.php

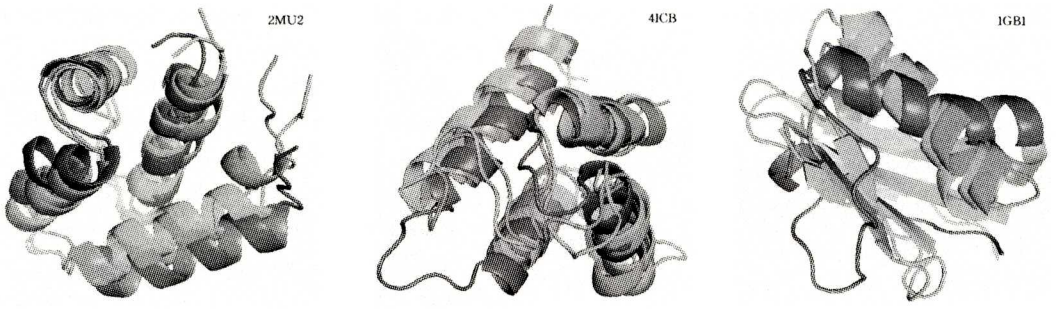

图 2　预测结构和实验方法测定结构的三维相似性对比图

3　分析与讨论

实验结果中，RLDE 算法预测得到最高精度的蛋白质是 1ENH，达到了 1.24Å，在所有测试蛋白上平均精度达到了 2.71Å，尤其是折叠类型为 α、长度小于 100 的蛋白，平均预测精度达到了 2.25Å。图 1 表明，RLDE 算法较 Rosetta 算法有更好的全局和局部搜索能力：RLDE 算法能够在更为广阔的构象空间中采样，并且逐步趋向能量更低的区域；而 Rosetta 算法的搜索区域明显较小，且对低能量区域的采样明显不足。图 2 清晰地展示了采用 RLDE 算法和 Rosetta 算法得到的蛋白质三维结构与实验室测定的结构之间的相似程度。总体而言，RLDE 算法计算得到的结构与实验测定的结构比 Rosetta 算法计算得到的结构重叠区域更多，更为相似。对于绝大多数 α 螺旋结构，RLDE 算法能够很好地探测组装得到；而对于 β 折叠结构和 loop 环区结构，RLDE 算法也能够找到局部拟合的片段，但整体组装效果并不是非常理想。

针对 15 个测试蛋白，RLDE 算法得到了较高精度的近天然态构象，但还需要对 RLDE 算法的执行性能进行验证。在常规副本交换方法中，为使副本交换以最优性能执行，副本层温度值分布的设定是一个需要考虑的问题[24]。经过反复测试验证，RLDE 采用如下温度分布：$kT=[0.67,0.72,0.95,1.14,1.36,1.63,1.95,2.33]$，这些值依据模拟退火策略的设置呈指数分布，这样的分布是最优温度分布[55]。此外，还需要考虑两个因素：副本层是否足够；最高温度是否足够高以确保搜索过程不会陷入局部极值。验证以上两个因素是否满足，即验证副本交换概率是否足够高（通常要大于0.1）。由表 1 中所列数据可知，在所选测试蛋白上，副本交换率均大于 10%，但又不会过高（大于 0.3）使得副本交换过于频繁从而达不到该温度下的平衡状态，造成空间构象采样不当[56]；对于温度是否足够高以确保搜索过程不会陷入局部极小值，并没有一种直观的验证方法，但是从测试结果图中可以看到，搜索过程随着迭代次数的增加逐步向低能量区域逼近，并没有陷入局部极小值而不能跨越能量势垒，因此可以验证：副本层最高温度的设置是足够高的。综上，所设计的 RLDE 算法在副本交换环节按照最优性能运行。

以上测试结果表明，所提 RLDE 算法能够在构象空间进行充分的采样，有效地计算得到蛋白质近天然态构象。图 3 所示为搜索过程中得到的构象 RMSD 值分布情况的统计，线条较粗的形表示 RLDE 算法，线条较细的表示 Rosetta 算法。从图中可以看出，RLDE 算法采样得到的构象 RMSD 值整体小于 Rosetta 算法的，这表明 RLDE 算法比 Rosetta 算法有更强、更有效的采样能力。但是，在少数几个测试蛋白上，RLDE 算法没有表现出比 Rosetta 更有效的采样能力，如图 3 所示的测试蛋白 1VII，它们的预测精度在两种方法上没有明显的差距，一方面由于序列本身长度较短，采用这两种不同的预测方法并不会有十分明显的差别；另一方面，力场模型的不精确性也可能导致两种预测方法在特定的蛋白上没有明显的差异。

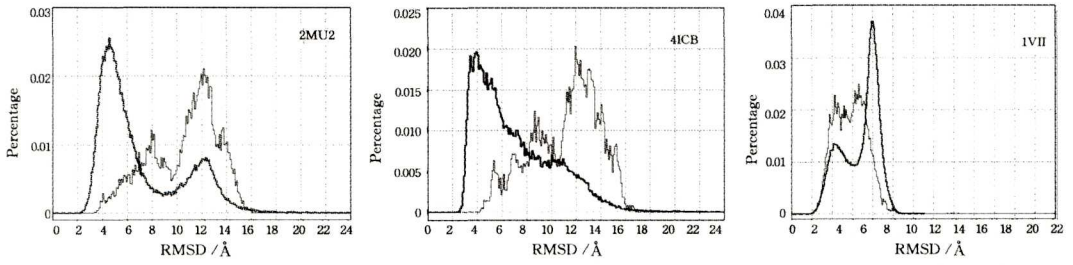

图 3　RMSD 分布柱状图

综上所述，RLDE 算法在所有测试蛋白上的整体预测结果精度较高，但是对于 β 折叠结构和 loop 环区结构的预测能力仍然稍显不足。造成这种结果的因素有很多：首先，一个重要的原因是，折叠类型为 β 和 loop 的蛋白质原子间作用力多为远程作用力，这可能导致力场模型不够精确，使得所构建的能量函数不够精确；其次，片段库的质量问题以及选择的片段长度和插入长度也是其中一种原因；最后，构象空间能量曲面的粗糙性也可能使得在搜索过程中陷入局部极小值。

结束语　本文提出一种基于副本交换的局部增强差分进化蛋白质结构从头预测方法（RLDE）。采用基于知识的粗粒度能量模型 Rosetta Score3 和片段组装技术显著降低了构象空间的维数，同时又不失去主要的结构信息，有效避免了搜索过程中的熵效应；用构象种群取代副本层中单独的构象，增加了构象空间中采样的多样性；采用全局搜索性能优良的差分

进化算法对种群进行更新，然后利用 Monte Carlo 算法的局部搜索特性对种群做局部增强，副本交换的策略同时可以增强算法跳出局部极值解的能力，从而得到全局和一系列局部最优构象。15 个蛋白质测试的结果表明，所提方法能够有效地对构象空间进行采样，得到高精度的蛋白质近天然态构象。

此外，新的预测方法的提出，有助于检验能量力场模型的准确性，并进一步对其进行修正。如图 1 所示，并非所有能量最低的构象所对应的 RMSD 值都最小，这就说明了力场模型仍然不够精确。在下一步的研究中，将进一步设计更为有效的预测算法，并尝试对所使用的能量力场模型进行修正，以期得到更好的预测结果。

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] ANFINSEN C B. Principles that govern the folding of protein chains [J]. Science, 1973, 181(96): 223-230.

[3] DORNA M, SILVAB M B, BURIOLA L S, et al. 3-dimensional protein structure prediction: Methods and computational strategies [J]. Computational Biology and Chemistry, 2014, 53(B): 51-276.

[4] KRYSHTAFOVYCH A, FIDELIS K, MOULT J. CASP10 results compared to those of previous CASP experiments [J]. Proteins: Structure, Function, and Bioinformatics, 2014, 82(S2): 164-174.

[5] BAKER D, SALI A. Protein structure prediction and structural genomics [J]. Science, 2001, 294(5540): 93-96.

[6] BELIAKOV G, LIMA K F. Challenges of continuous global optimization in molecular structure prediction [J]. European Journal of Operational Research, 2007, 181(3): 1198-1213.

[7] LEE J, WU S, ZHANG Y. From Protein Structure to Function with Bioinformatics [M]. Berlin: Springer Netherlands, 2009: 3-25.

[8] TANTA 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.

[9] 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.

[10] 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): 58-576.

[11] 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(2): 88-99.

[12] STORN R, PRICE K. Differential Evolution-A Simple and Efficient Heuristic for global Optimization over Continuous Spaces [J]. Journal of global optimization, 1997, 11(4): 341-359.

[13] ZOU D X, WU J H, GAO L Q, et al. A modified differential evolution algorithm for unconstrained optimization problems [J]. Neurocomputing, 2013, 120(11): 469-481.

[14] CASCIATI, SARA. Differential evolution approach to reliability-oriented optimal design [J]. Probabilistic Engineering Mechanics, 2014, 36(4): 72-80.

[15] 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.

[16] SCHERAGE H A, KHALIL I, LIWO A. Protein-folding dynamics: overview of molecular simulation techniques [J]. Annual Review of Physical Chemistry, 2007, 58(1): 57-83.

[17] LINDORFF-LARSEN K, TRBOVICN, 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.

[18] 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.

[19] SHEN Y, PICORD G, GUYON F, et al. Detecting protein candidate fragments using a structural alphabet profile comparison approach [J]. PloS one, 2013, 8(11): e80493.

[20] 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.

[21] DOTU I, CEBRIA M, VAN H P, et al. On Lattice Protein Structure Prediction Revisited [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(6): 1620-1632.

[22] 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.

[23] 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.

[24] SUGITA Y, OKAMOTO Y. Replica-exchange molecular dynamics method for protein folding [J]. Chemical Physics Letters, 1999, 314(1/2): 141-151.

[25] SUGITA Y, OKAMOTO Y. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape [J]. Chemical Physics Letters, 2000, 329(3): 261-270.

[26] CZAPLEWSKI C, KALINOWSKI S, LIWO A, et al. Application of Multiplexed Replica Exchange Molecular Dynamics to the UNRES Force Field: Tests with alpha and alpha+beta Proteins [J]. Journal of Chemical Theory and Computation, 2009, 3(5): 627-640.

[27] HANSMANN U H E. Parallel tempering algorithm for conformational studies of biological molecules [J]. Chemical Physics Letters, 1997, 281(1): 140-150.

[28] GÜNTERTA P, BILLETERA M, OHLENSCHLÄGERB O, et al. Conformational analysis of protein and nucleic acid fragments

with the new grid search algorithm FOUND [J]. Journal of Biomolecular NMR, 1998, 12(4): 543-548.

[29] PETRELLA R J. A versatile method for systematic conformational searches: Application to CheY [J]. Journal of computational chemistry, 2011, 32(11): 2369-2385.

[30] SCHERAGE H A. Some approaches to the multiple-minima problem in the calculation of polypeptide and protein structures [J]. International Journal of Quantum Chemistry, 1992, 42(5): 1529-1536.

[31] ADJIMAN C S, DALLWIG S, FLOUDAS C A, et al. A global optimization method, αBB, for general twice-differentiable constrained NLPs-I. Theoretical advances-II. Application of theory and test problems [J]. Computers and Chemical Engineering, 1998, 22(9): 1137-1158.

[32] KLEPEIS J L, PIEJA M J, FLOUDAS C A. Hybrid global optimization algorithms for protein structure prediction: Alternating hybrids [J]. Biophysical Journal, 2003, 84(2): 869-882.

[33] FLOUDAS C A, FUNG H K, MCALLISTER S R, et al. Advances in protein structure prediction and de novo protein design: A review [J]. Chemical Engineering Science, 2006, 61(3): 966-988.

[34] 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(3): 383-398.

[35] OLSON B, SHEHU A. Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface[J]. Proteome Science, 2012, 10(S1): S5.

[36] MOLLOY K, SALEH S, SHEHU A. Probabilistic Search and Energy Guidance for Biased Decoy Sampling in Ab-initio Protein Structure Prediction [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013, 10(5): 1162-1175.

[37] KEAVER-FAY A, TYKA M, LEWIS S M, et al. ROSETTA3: an objectoriented software suite for the simulation and design of macromolecules [J]. Methods in Enzymology, 2011, 487: 545-574.

[38] XU D, ZHANG Y. Ab-initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field [J]. Proteins: Structure, Function, and Bioinformatics, 2012, 80(7): 1715-1735.

[39] STOEAN C, PREUSS M, STOEAN R, et al. Multimodal Optimization by Means of a Topological Species Conservation Algorithm [J]. Evolutionary Computation, 2010, 14(6): 842-864.

[40] ZHANG Y S, HAO Z F, HUANG H. Global convergence and premature convergence of two-membered evolution strategy [J]. Journal of Computer Research and Development, 2014, 54(4): 754-761. (in Chinese)
张宇山，郝志峰，黄翰. 二元进化策略的全局收敛与早熟收敛 [J]. 计算机研究与发展，2014，54(4)：754-761.

[41] 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.

[42] KMIECIK S, JAMROZ M, KOLINSKI A. Multiscale Approaches to Protein Modeling [M]. Berlin: Springer Science, 2011: 281-293.

[43] BRADLEY P, MISURA K M, BAKER D. Toward high-resolution de novo structure prediction for small proteins [J]. Science, 2005, 309(5742): 1868-1871.

[44] LIWO A, KHALILI M, SCHERAGE H A. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains [J]. PNAS, 2005, 102(7): 2362-2367.

[45] 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(S1): S4.

[46] KUHLMAN B, BAKER D. Native protein sequences are close to optimal for their structures [J]. Proceedings of the National Academy of Sciences of the Unitized States of America, 2000, 97(19): 10383-10388.

[47] KORTEMME T, MOROZOV A V, BAKER D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes [J]. Journal of molecular biology, 2003, 326(4): 1239-1259.

[48] HUANG E S, SAMUDRALA R, PARK B H. Scoring Functions for ab initio Protein Structure Prediction [J]. Methods in Molecular Biology, 2000, 143: 223-245.

[49] EARL D J, DEEM M W. Parallel tempering: Theory, applications, and new perspectives [J]. Physical Chemistry Chemical Physics, 2005, 7(23): 3910-3916.

[50] LIAO C Y, ZHOU J. Replica Exchange Molecular Dynamics Simulations on the Folding of Trpzip4 β-Hairpin [J]. ACTA CHIMICA SINICA, 2013, 71: 593-601. (in Chinese)
廖晨伊，周健. β发卡多肽 Trpzip4 折叠的副本交换分子动力学模拟 [J]. 化学学报，2013，71：593-601.

[51] STORN R. Differential evolution design of an IIR-filter in Evolutionary Computation [C]// Proceedings of IEEE International Conference on Evolutionary Computation (ICEC'96). New York: IEEE, 1996: 268-273.

[52] BERG B A, NEUHAUS T. Multicationical ens EMBLE-A new approach to simulate1ST-order Phase-Transitions [J]. Physics Review Letter, 1992, 68(1): 9-12.

[53] YU W J, SHEN M, CHEN W N, et al. Differential evolution with two-level parameter adaptation [J]. IEEE Transactions on Cybernetics, 2014, 44(7): 1080-1099.

[54] DING F, TSAO D, NIE H, et al. Ab initio folding of proteins using all-atom discrete molecular dynamics [J]. Structure, 2008, 16(7): 1010-1018.

[55] OKAMOTO Y, FUKUGITA M, NAKAZAWA T, et al. Alpha-helix folding by Monte Carlo simulated annealing in isolated C-peptide of ribonuclease A[J]. Protein Eng., 1991, 4(6): 639-686.

[56] ZHANG W, WU C, DUAN Y. Convergence of replica exchange molecular dynamics [J]. Journal of Chemical Physics, 2005, 123(15): 154105-154113.