计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.181202395

基于接触图残基对距离约束的蛋白质结构预测算法

谢腾宇[1]　周晓根[2]　胡　俊[1]　张贵军[1]
1 浙江工业大学信息工程学院　杭州 310023
2 密西根大学计算医学与生物信息学系　安娜堡 MI45108
(xty@zjut.edu.cn)

摘　要　从头预测是蛋白质结构建模的一种重要方法,该方法的研究有助于人类理解蛋白质功能,从而进行药物设计和疾病治疗。为了提高预测精度,文中提出了基于接触图残基对距离约束的蛋白质结构预测算法(CDPSP)。基于进化算法框架,CDPSP将构象空间采样分为探索和增强两个阶段。在探索阶段,设计基于残基对距离的变异与选择策略,即根据接触图的接触概率选择残基对,并通过片段组装技术对所选择的残基对的邻近区域进行变异;将残基对距离离散化为多个区域并为其分配期望概率,根据期望概率确定是否选择变异的构象,从而增加种群的多样性。在增强阶段,利用基于接触图信息的评分指标,结合能量函数,衡量构象的质量,从而选择较优的构象,达到增强CDPSP近天然态区域采样能力的效果。为了验证所提算法的性能,通过CASP12中的10个FM组目标蛋白质对其进行了测试,并将其与一些先进算法进行比较。实验结果表明,CDPSP可以预测得到精度较高的蛋白质三维结构模型。

关键词:蛋白质结构预测;从头预测;残基对距离;接触图;进化算法;片段组装
中图法分类号　TP301.6

Contact Map-based Residue-pair Distances Restrained Protein Structure Prediction Algorithm

XIE Teng-yu[1],ZHOU Xiao-gen[2],HU Jun[1] and ZHANG Gui-jun[1]
1 College of Information Engineering,Zhejiang University of Technology,Hangzhou 310023,China
2 Department of Computational Medicine and Bioinformatics,University of Michigan,Ann Arbor,MI 45108,USA

Abstract　De novo prediction is an important method for protein structure modeling. Research of the method contributes to humanity's understanding of protein functions to conduct drug design and disease treatments. In order to improve the accuracy of prediction,contact map-based residue-pair distances restrained protein structure prediction algorithm (CDPSP) was proposed. Based on the framework of evolutionary algorithm,CDPSP was used to sample conformational space,which was divided into exploration and exploitation stages. In the exploration stage,the strategies of mutation and selection were designed on the basis of the distances of residue-pair,which can increase the diversity of the population. In detail,a residue-pair was chosen according to the contact probability of contact map and the mutation was conducted through fragment assembly technique on the adjacent region of the residue-pair. The selection of mutated conformation was determined by the expected probability distributed through the discretization of residue-pair distances. In the exploitation stage,the contact-based score and energy function were used to evaluate the quality of conformations in search of good conformations,which can enhance the sampling ability of CDPSP in near-native region. In order to verify the performance of the proposed algorithm,CDPSP is tested on 10 targets in the FM group of CASP12 and compared with advanced algorithms. The test results show that CDPSP can predict more accurate protein tertiary structure models.

Keywords　Protein structure prediction,De novo prediction,Distances of residue-pair,Contact map,Evolutionary algorithm,Fragment assembly

1　引言

从氨基酸序列出发,预测蛋白质结构,即"第二遗传密码",是一个尚未解决的问题[1]。该问题的研究在理解蛋白质功能、疾病诊断、药物设计等方面具有重要意义,且有助于缩小已知蛋白质序列数目与结构数目之间的差距。国内已有多位研究学者对该领域的发展现状进行了总结和回顾[2-4],该问题是国内学术界的研究热点。

到稿日期:2018-12-24　返修日期:2019-03-26　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61773346);浙江省自然科学重点基金(LZ20F030002)
This work was supported by the National Natural Science Foundation of China (61773346) and Zhejiang Provincial Science Foundation of China (LZ20F030002).
通信作者:张贵军(zgj@zjut.edu.cn)

从头预测方法是蛋白质结构建模的一类重要方法[5-6]。在历届 CASP 竞赛中，从头预测方法的预测精度逐渐提升[7-10]。该方法无须实验解析结构，基于 Anfinsen 准则[11]构建能量函数，直接从伸展的肽链开始预测三维结构；在能量函数的引导下，进行构象空间的搜索，最后从构象池中选择近天然态模型作为预测结果。由于蛋白质构象空间随氨基酸序列的增长呈指数增加，残基序列长度超过 120 的蛋白质预测精度仍然无法满足实际要求[12]，因此，如何高效地搜索高维构象空间，是从头预测方法中的一个关键问题[13-14]。

现有文献中，有许多构象空间采样算法被用于蛋白质结构的从头预测，包括蒙特卡洛算法[15-17]、分子动力学模拟[18-20]、进化算法[21-31]、构象空间退火[32]、副本交换[16,33-34]、构象树搜索[35]等。此外，片段组装技术[36-38]作为一种提高搜索效率的手段，在以上算法中被广泛应用。Rosetta[37]和 QUARK[39]均采用了片段组装技术，这两种方法在历届 CASP 竞赛中表现突出[40-43]。Rosetta[37]从头预测协议中，在蒙特卡洛算法的框架下，分阶段使用不同长度的片段库和多尺度的能量函数进行高效的构象空间搜索。QUARK[39]则利用神经网络预测距离谱等多种结构特征，然后采用基于特征约束的副本交换蒙特卡洛算法进行蛋白质结构的预测。

残基接触图对提升从头预测方法的精度做出了巨大贡献[44-45]。FRAGFOLD[46]结合片段组装技术和接触图进行构象空间的搜索。Flib-Coevo[47]利用接触图约束产生高质量的片段库。SCDE[48]在差分进化算法的框架下，利用接触图和二级结构提出了两种选择策略。此外，残基对距离对于提高蛋白质结构预测的精度至关重要。Google 旗下的 DeepMind 开发了蛋白质结构预测系统 AlphaFold[49]，在 CASP13 竞赛中取得了显著的进步；该方法利用神经网络预测残基对间距离的离散分布，并用其来评估构象的优劣。DeepCDpred[50]通过深度神经网络预测残基间的距离和接触图，将其用于从构象池中选择近天然态构象。

为了提高从头预测方法的精度，本文提出了一种基于接触图残基对距离约束的蛋白质结构预测算法(CDPSP)。在进化算法的框架下，该算法分探索和增强两个阶段来进行蛋白质结构的预测。在探索阶段，设计了基于残基对距离的变异和选择策略，以增加群体的多样性。在增强阶段，利用基于接触图的评分指标辅助构象空间采样，以提升局部构象区域的搜索能力。本文以 CASP12 的 10 个 FM(Free Modeling)组目标蛋白作为测试蛋白，并通过与参赛的多种从头预测方法进行对比，来验证所提算法的有效性。

2 相关工作

2.1 Rosetta 能量函数

Rosetta 的从头预测协议 ClassicAbinitio[37]利用已知的蛋白质结构进行构象统计，得到基于知识的能量函数。Rosetta 从头预测协议采用粗粒度模型表达蛋白质结构(见图 1)，仅保留主链原子及侧链虚拟中心；采用 Rosetta 能量函数 score3 来评价以粗粒度模型表达的蛋白质结构。该能量函数是多种能量项的线性加权和，主要包含溶剂可及性、残基对相互作用、氢键作用、旋转半径、C_β 密度、范德华力等[37]。

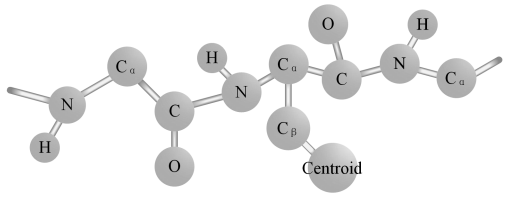


图 1 粗粒度表达模型

Fig. 1 Coarse-grained representation

2.2 接触图

接触图中包含残基对间的接触信息。如果一对残基在三维结构中是接近的，那么认为这对残基接触。在许多情况下，当残基间的 Cβ-Cβ(甘氨酸 Cα-Cα)距离小于 8 Å 时，认为残基对是接触的。此外，预测的接触图中包含残基对的接触概率，其表示接触的可能性。本文通过 RaptorX-Contact 获得目标蛋白的接触图。RaptorX-Contact[51-53]分析多家族共同进化信息，利用卷积神经网络预测蛋白质接触图。蛋白质残基对接触信息的有效利用，可以显著提高从头预测方法的精度[54]。此外，考虑到残基对的接触概率以及接触信息的冗余，本文利用接触概率在前 L(称为“top L”，L 是目标蛋白的氨基酸序列长度)的残基进行蛋白质结构预测。

3 CDPSP 算法

CDPSP 算法分为探索和增强两个阶段，如图 2 所示。在探索阶段，为了尽可能保留群体的多样性，设计了基于残基对距离的变异与选择策略；此外，设计了基于接触图的评分指标，结合能量函数进行构象空间的优化。在增强阶段，按概率选择能量函数和基于接触图的评分指标，进行构象空间的搜索；选择能量函数的概率随着进化过程的进行逐渐增大。最后，通过聚类工具 SPICKER[55]对构象进行聚类，输出最终的预测结果。

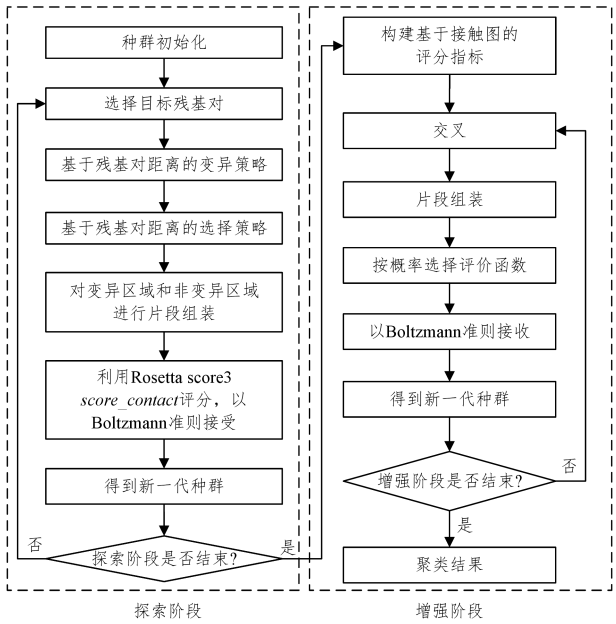


图 2 CDPSP 流程框图

Fig. 2 Flowchart of CDPSP

3.1 基于残基对距离的变异与选择策略

利用 RaptorX-Contact 预测的 top L 对残基接触概率进行有偏变异，选择潜在构象，使得种群的残基对在满足接触概率的条件下尽可能均匀分布，从而增加群体的多样性。

3.1.1 变异策略

在探索阶段，期望种群的残基对距离在接触图的约束下均匀分布。以 top L 对残基的接触概率为适应度，轮盘赌选择残基对 C_{ij}。对选择的残基对 C_{ij}（i 和 j 是残基序号）的邻近区域进行片段组装，变异范围（Mutation range）由 i，j 和 f（f 是片段长度）确定。具体选择变异区域 MR_i 的方式如式(1)所示：

$$MR_i=\begin{cases}[1,\frac{4}{3}f+1], & \text{if } i<\frac{2}{3}f\\ [L-\frac{4}{3}f,L], & \text{if } i>L-\frac{2}{3}f\\ [i-\frac{2}{3}f,i+\frac{2}{3}f], & \text{其他}\end{cases} \tag{1}$$

其中，$i=1,\cdots,L$。MR_j 以同样的方式选取。

3.1.2 选择策略

DPSP 将残基对间的距离 d_{ij} 离散化为 B 个不间断区域，并根据当前残基对 C_{ij} 的接触概率 p_{ij} 给 B 个区域等概率地分配期望概率。如图 3 所示，当 $d_{ij}<5$Å 时，将[3.8,5)作为第一个区域，3.8 Å 是两个序列相邻残基的距离；当 $5\leqslant d_{ij}<20$ Å 时，每个区域长度为 1 Å；当 $d_{ij}\geqslant 20$ Å 时，将[20，3.8·|$i-j$|]作为最后一个区域，其中 3.8·|$i-j$| 是残基对间结构处于伸展状态时的残基对距离。在当前划分的情况下，区域个数 $B=17$。当 $d_{ij}<8$ Å 时，每个区域的概率 $P_1=p_{ij}/4$；当 $d_{ij}\geqslant 8$ Å 时，$P_2=(1-p_{ij})/13$。

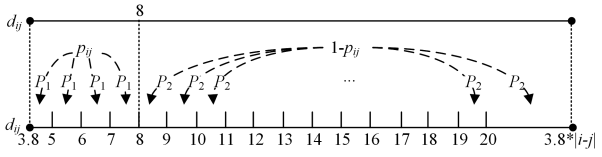


图 3 残基对距离离散化

Fig. 3 Discretization of distance of residue pairs

根据 B 个不间断区域的期望概率，利用轮盘赌的方法选择期望达到的变异区域。若变异后的构象向期望区域靠近，则接受该构象。

3.2 基于接触图的评分指标

基于接触图的评分指标 *score_contact* 如式(2)所示：

$$score\_contact=\begin{cases}\sum_{i,j}(1-p_{ij})+\frac{(2\cdot p_{ij}-1)}{1+e^{8-d_{ij}}}, & \text{if } d_{ij}<20\\ \sum_{i,j}(1-p_{ij})+\frac{(2\cdot p_{ij}-1)}{1+e^{8-20}}+e^{\frac{d_{ij}}{20}}-e, & \text{其他}\end{cases} \tag{2}$$

随着残基对距离的增加，对应残基对的评分总体呈递增趋势；此外，考虑到残基对接触概率，设计了在 8 Å 处跃变的 Sigmod 函数项。该指标是将 top L 对残基对的评分求和，并以此作为整个构象的评分 *score_contact*。图 4 给出了一对残基的评分在不同概率下与残基对距离的对应关系。在探索阶段和增强阶段，CDPSP 利用接触图建立评分指标，辅助构象空间搜索。

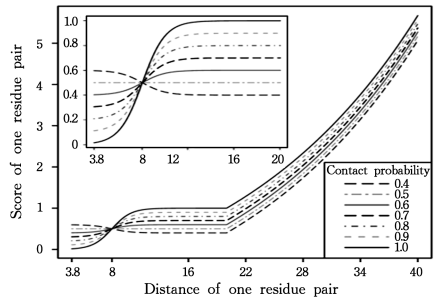


图 4 残基对的评分

Fig. 4 Score of residue-pair

3.3 算法描述

3.3.1 探索阶段

1）种群初始化：输入目标蛋白的序列、接触图的 top L 对残基接触信息、片段库（通过 Robetta 在线服务器获得），生成 NP 个伸展态构象，并将其作为初始种群 $T_{g_1}=\{T_1,T_2,\cdots,T_{NP}\}$，探索阶段的当前迭代次数 $g_1=0$。

2）将种群中第 k 个个体 T_k 作为目标个体，$k\in\{1,\cdots,NP\}$，并根据残基对接触概率，以轮盘赌的方式选择目标残基对 C_{ij}。

3）根据目标残基对 C_{ij} 的概率 p_{ij}，将残基对距离离散化为 B 个区域，并为其分配期望概率 P_1 与 P_2；以期望概率为适应度，轮盘赌选择期望区域 b，$b\in\{1,\cdots,B\}$。

4）根据式(1)确定变异区域，此时 $f=9$；并对 T_k 的变异区域进行一次片段组装，得到测试个体 T_k'。

5）分别计算 T_k 与 T_k' 的残基对 C_{ij} 间的距离 d_{ij} 和 d'_{ij}。判断 d'_{ij} 是否到达期望区域 b，若到达，则接受 T_k'，即 $T_k=T_k'$，执行下一步；否则判断 d'_{ij} 是否距离期望区域 b 的中心更近，若是则接受 T_k' 并执行下一步，否则拒绝 T_k'。返回步骤 4），直至达到最大连续拒绝次数 R，执行下一步。

6）对 T_k 的变异区域进行片段组装，片段长度 $f=3$，并采用玻尔兹曼准则接受测试个体，评分函数为 score3 和基于接触图的评分指标 *score_contact*。重复当前步骤 R_S 次。

7）对 T_k 的非变异区域进行片段组装，片段长度 $f=9$，并根据玻尔兹曼准则接受测试个体，评分函数为能量函数 score3 和基于接触图的评分指标 *score_contact*。重复当前步骤 R_L 次。

8）以种群每个个体作为目标个体，执行步骤 2）－步骤 7），得到新一代种群，$g_1=g_1+1$。

9）重复执行步骤 8），直至达到最大进化次数 G_{max1}，探索阶段结束。

3.3.2 增强阶段

1）将探索阶段的最后一代种群作为增强阶段的初始种群 T_{g_2}，当前代数 $g_2=0$。

2）将种群中的第 k 个个体 T_k 作为目标个体，并从种群中随机选择另一个个体 T_r。

3）按交叉概率 CR 对 T_k 和 T_r 进行局部片段互换，片段长度 $f=3$，记交叉后的 T_k 为 T_t。

4）对 T_t 进行片段组装，片段长度 $f=9$，得到变异个体 T_m。

5)计算当前代数与增强阶段的最大代数 G_{max2} 的比率，即 $r=g_2/G_{max2}$。

6)选择评分函数：产生[0,1]之间的随机数 s，若 $s<r\cdot Q$，则选择能量函数 score3，否则选择 *score_contact*。其中，Q 是对 *score_contact* 的置信比率。

7)根据上一步选择的评分函数评价 T_k 和 T_m，并根据玻尔兹曼准则接受变异个体。重复步骤 4)－步骤 7)R_M 次。

8)以种群中的每个个体为目标个体，执行步骤 2)－步骤 7)，得到新一代种群，$g_2=g_2+1$。

9)重复执行步骤 8)，直至达到最大进化次数 G_{max2}。

10)利用聚类工具 SPICKER[55] 对增强阶段玻尔兹曼准则接受的构象进行聚类，将最终得到的 5 个模型作为预测结果。

4 测试蛋白和实验设置

本文以 CASP12 竞赛 FM 组的 10 个目标蛋白进行测试，序列长度为 69～246，折叠类型包括 α，β，α/β，如表 1 所列。

CDPSP 算法的实验参数设置如下：种群规模 $NP=100$，残基对距离离散化区域的个数 $B=17$，探索阶段的最大代数 $G_{max1}=100$，增强阶段的最大代数 $G_{max2}=300$，最大连续拒绝次数 $R=100$；探索阶段连续片段的组装次数 $R_S=100$，$R_L=30$，交叉概率 $CR=0.5$；增强阶段连续片段的组装次数 $R_M=150$；*score_contact* 置信比率 $Q=0.7$。利用 Robetta[56] 得到去同源的片段库，片段长度分别为 3 和 9。

利用 RaptorX-Contact 预测目标蛋白的接触图，所有目标蛋白的天然态结构是 CASP12 发布的实验测定结构。此外，将 CDPSP 算法与参赛的 3 种预测方法进行比较，以对该算法的预测性能进行验证。3 种比较方法分别为 BAKER-ROSETTASERVER，QUARK 和 RaptorX-Contact。以上 3 种方法的预测结果均从 CASP12 中获得。

本文采用两种被广泛使用的结构指标来评价预测结构与实验测定结构的相似度：RMSD(均方根偏差)和 TM-score。RMSD 是两个结构经过最优刚体结构比对后 Cα 原子间的平均距离。TM-score[57-58] 也是比对结构间的整体拓扑结构相似度的指标，其取值范围为(0,1]。

5 实验结果

为了验证探索阶段的作用，仅执行 CDPSP 算法的增强阶段(记为 CDPSP-O)，种群初始化和探索阶段的初始化过程一致；为了验证基于接触图的评分指标的有效性，CDPSP-E 将 CDPSP 算法中所有基于接触图的评分换为能量函数。CDPSP，CDPSP-O，CDPSP-E 的预测结果如表 1 所列。表 1 中，第一列 Targets 表示目标蛋白的参赛编号，第二列 Length 表示目标蛋白的序列长度，第三列 Type 表示折叠类型，第四列 RMSD-average 和第五列 TM-score-average 分别是对应方法预测的 5 个模型的 RMSD 平均值和 TM-score 平均值。

表 1 CDPSP 算法的预测结果

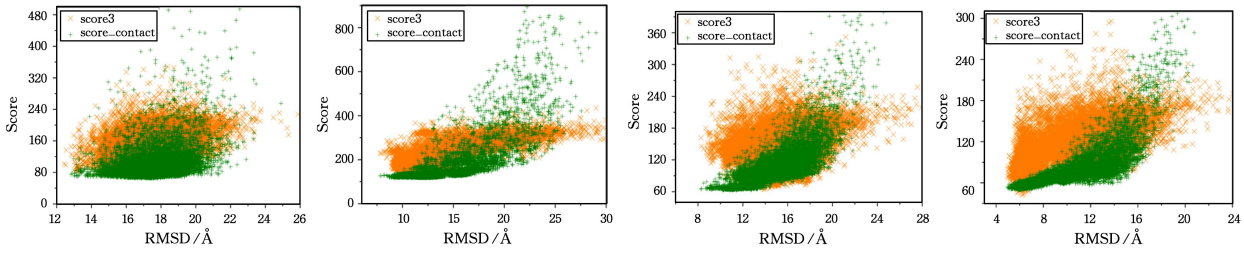
Table 1 Results of CDPSP

Targets	Length	Type	RMSD-average(Å)			TM-score-average		
			CDPSP	CDPSP-E	CDPSP-O	CDPSP	CDPSP-E	CDPSP-O
T0859-D1	129	α/β	15.80	16.59	**15.69**	**0.22**	0.17	0.19
T0862-D1	101	α	12.74	13.73	**8.72**	**0.44**	**0.44**	0.43
T0863-D1	193	α	**11.19**	18.50	14.11	**0.36**	0.20	0.32
T0864-D1	246	β	19.82	23.49	**19.12**	0.18	0.16	**0.21**
T0866-D1	104	β	13.32	12.86	**11.34**	0.26	0.27	**0.35**
T0868-D1	116	α/β	**6.62**	11.49	6.74	0.46	0.28	**0.50**
T0869-D1	104	α/β	**5.93**	11.86	8.72	**0.45**	0.29	0.37
T0870-D1	123	α	11.65	13.64	**9.07**	0.37	0.23	**0.42**
T0886-D1	69	β	12.40	12.59	**11.62**	0.22	0.10	**0.25**
T0886-D2	127	α/β	**13.55**	17.38	14.12	**0.30**	0.22	0.26

注：表中加粗数据表示在对应指标下，相应算法的预测精度最高

如表 1 所列，CDPSP 和 CDPSP-O 的预测精度总体比 CDPSP-E 高，说明基于接触图的评分指标有助于提高构象空间的搜索能力。从折叠类型的角度分析，CDPSP 算法对 1 个 α 蛋白(T0863-D1)、2 个 α/β 蛋白(T0869-D1、T0886-D2)的预测精度最高，而 CDPSP-O 对 1 个 α 蛋白(T0870-D1)和 3 个 β 蛋白(T0864-D1，T0866-D1，T0886-D1)的预测能力更好。

图 5 为 score3(或 score_contact)与 RMSD 的散点图，从中可以看出评分指标对近天然态区域搜索能力的影响。



(a)T0859-D1 (b)T0863-D1 (c)T0866-D1 (d)T0869-D1

图 5 CDPSP 散点图

Fig. 5 CDPSP plots

以 T0859-D1，T0863-D1，T0866-D1 和 T0869-D1 为例，图中数据均是在增强阶段按玻尔兹曼标准接受的构象的评分和 RMSD。对于 T0869-D1，score_contact 与 RMSD 的相关性明显高于能量函数与 RMSD 的相关性，从而使得算法能够更快搜索到近天然态区域，从而得到精度更高的预测结果。对于 T0866-D1，score_contact 与 RMSD 的相关性稍弱，但仍能够辅助能量函数区分构象的优劣。对于 T0859-D1 和 T0863-D1，能量函数和基于接触图的评分指标与 RMSD 的相关性均较弱，导致预测结果精度不够高。如表 2 所列，分别对 CDPSP-E，CDPSP-O 的 RMSD-average 值与 CDPSP 进行成对 t 检验，得到 p-value。

从表 2 中看出，CDPSP-E 与 CDPSP 的 p-value 值为 0.0065，小于 0.05，说明 CDPSP 显著优于 CDPSP-E；而 CDPSP-O 与 CDPSP 的 p-value 值为 0.5990，两者的差异不显著。

表 2 成对 t 检验结果

Table 2 Results of paired t-test

CDPSP vs	CDPSP-E	CDPSP-O
p-value	0.0065	0.5990
Significanc	+	−

6 对比分析

为了进一步验证 CDPSP 算法的预测能力，本节将该算法与 BAKER-ROSETTASERVER，QUARK 和 RaptorX-Contact 进行比较。下中分别用 BR，QK 和 RX 表示以上 3 种算法。对比结果如表 3 所列，RMSD-average 与 TM-score-average 的含义与表 1 一致，RMSD-min 表示 5 个模型中 RMSD 的最小值，TM-score-max 表示 5 个模型中 TM-score 的最大值。

表 3 各算法的对比结果

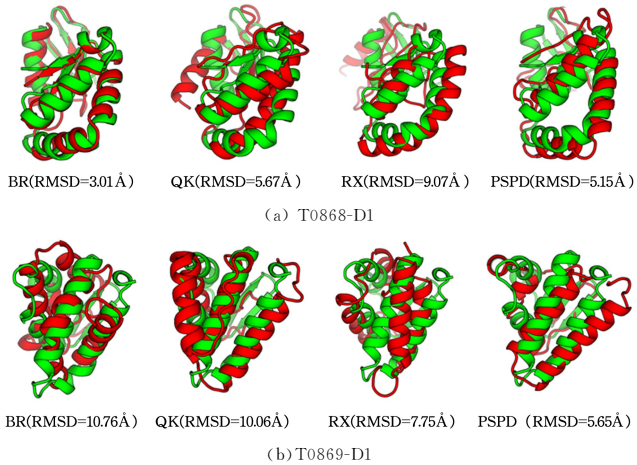
Table 3 Comparison results of algorithms

Targets	RMSD-average(Å)				RMSD-min(Å)				TM-score-average				TM-score-max			
	BR*	QK*	RX*	CDPSP	BR*	QK*	RX*	CDPSP	BR*	QK*	RX*	CDPSP	BR*	QK*	RX*	CDPSP
T0859-D1	18.32	16.03	16.93	**15.80**	17.05	15.22	16.27	**14.80**	0.22	**0.25**	0.22	0.22	0.27	**0.28**	0.24	0.24
T0862-D1	13.50	12.90	15.72	**12.74**	9.86	**5.55**	13.33	9.90	0.40	0.42	0.29	**0.44**	0.49	**0.54**	0.34	0.46
T0863-D1	18.95	17.44	21.40	**11.19**	16.20	13.68	17.59	**9.96**	0.25	0.26	0.22	**0.36**	0.34	0.30	0.26	**0.39**
T0864-D1	19.58	20.22	**15.79**	19.82	17.76	16.11	**11.22**	17.77	**0.36**	0.24	0.33	0.18	0.40	0.36	**0.41**	0.20
T0866-D1	**3.64**	7.74	8.04	13.32	**3.14**	4.23	5.09	10.74	**0.75**	0.50	0.41	0.26	**0.81**	0.63	0.52	0.28
T0868-D1	**6.07**	9.84	10.43	6.62	**3.01**	5.67	9.07	5.15	**0.59**	0.41	0.32	0.46	**0.80**	0.47	0.38	0.51
T0869-D1	12.18	13.88	9.64	**5.93**	10.76	10.06	7.75	**5.65**	0.34	0.29	0.40	**0.45**	0.37	0.34	0.45	**0.49**
T0870-D1	11.67	12.50	11.74	**11.65**	**9.59**	10.41	10.73	11.04	0.35	0.31	0.32	**0.37**	**0.41**	0.39	0.33	0.38
T0886-D1	20.11	15.52	**7.55**	12.40	15.20	11.07	**5.45**	10.33	0.20	0.29	**0.32**	0.22	0.27	0.35	**0.41**	0.28
T0886-D2	10.20	**10.08**	12.04	13.55	8.14	**6.41**	10.65	10.62	**0.47**	0.43	0.36	0.30	**0.54**	0.49	0.45	0.30

注：表中加粗数据表示在对应指标下相应算法的预测精度最高，* 表示对应数据来源于 http://predictioncenter.org/casp12/index.cgi

从 RMSD-average 的角度进行分析，对 5 个目标蛋白（T0859-D1，T0862-D1，T0863-D1，T0869-D1 和 T0870-D1），CDPSP 的预测精度比其他 3 种方法高，尤其是 T0869-D1，CDPSP 的预测精度有了显著提高。如第 5 节中的分析，基于接触图的评分指标使得 CDPSP 算法能快速找到更接近天然态构象的区域，并在近天然态区域进行了充分的搜索，使得预测精度有了极大的提高。从 TM-score-average 角度进行考虑，BAKER-ROSETTASERVER 和 CDPSP 算法的预测精度最高的蛋白质均有 4 个，其中 T0866-D1，BAKER-ROSETTASERVER 的 TM-score-average 达到 0.75。从 RMSD-min 角度分析，BAKER-ROSETTASERVER 对 T0866-D1 和 T0868-D1 的预测精度能够达到 3.5Å 以下；相比于其他 3 种算法，CDPSP 对 T0863-D1 和 T0869-D1 的预测精度有明显改进。在 TM-score-max 方面，CDPSP 的表现没有 RMSD 评价指标中的表现好，这可能是由于探索阶段根据期望区域接受了拓扑结构比较差的变异个体，而在增强阶段没有足够的能力使得整体的拓扑结构更优。

图 6 是 4 种方法预测 T0868-D1 和 T0869-D1 的最优模型（5 个模型中 RMSD 最小的模型）与实验测定结构的三维比对图（绿色代表实验解析结构，红色代表对应方法的最优模型）。对于 T0868-D1，BAKER-ROSETTASERVER，QUARK 和 CDPSP 的整体拓扑结构更接近于天然态结构。对于 T0869-D1，CDPSP 的预测结果与天然态结构更加接近，而其他 3 种方法的预测结果较差，整体拓扑结构与天然态结构差异较大。



(a) T0868-D1

(b) T0869-D1

图 6 CDPSP 预测结构（红色）与实验测定结构（绿色）的比对（电子版为彩色）

Fig. 6 Structural comparisons between predicted models (red) of CDPSP and experimental structures (green)

结束语 本文提出了基于接触图残基对距离约束的蛋白质结构预测算法。该算法在进化算法的框架下，设计了基于残基对距离离散概率分布的有偏变异与选择策略，从而对构象空间进行更充分的探索；并设计了基于接触图的评分指标，以辅助构象空间的优化，从而增强对近天然态区域的搜索。实验结果表明，CDPSP 具有较好的预测性能，是一种有效的蛋白质结构从头预测方法。然而，该从头预测算法中，残基对距离的离散分布仅考虑了接触图的信息，并且采取随机变异

的方式希望达到残基对期望区域。该分布的精细化将有利于提高算法的近天然态区域搜索能力;针对期望区域设计有偏变异策略将有利于增强算法的搜索效率。未来将考虑利用上述因素对从头预测方法进行改进,从而提高预测能力。

参 考 文 献

[1] KOLATA G. Trying to crack the second half of the genetic code [J]. Science, 1986, 233: 1037-1040.

[2] WANG C, ZHU J W, ZHANG H C, et al. A Survey on Algorithms for Protein Tertiary Structure Prediction [J]. Chinese Journal of Computers, 2018, 41(4): 760-779.

[3] DENG H Y, JIA Y, ZHANG Y. Protein structure prediction [J]. Acta Physica Sinica, 2016, 65(17): 169-179.

[4] MA B G. Protein Folding Prediction [J]. Chinese Science Bulletin, 2016, 61(24): 2670-2680.

[5] DILL K A, MACCALLUM J L. The protein-folding problem, 50 years on [J]. Science, 2012, 338(6110): 1042-1046.

[6] ZHANG Y. Protein structure prediction: when is it useful? [J]. Current Opinion in Structural Biology, 2009, 19(2): 145-155.

[7] MOULT J, FIDELIS K, KRYSHTAFOVYCH A, et al. Critical assessment of methods of protein structure prediction (CASP)-Round XII [J]. Proteins: Structure, Function, and Bioinformatics, 2018, 86 (Suppl 1): 7-15.

[8] MOULT J, FIDELIS K, KRYSHTAFOVYCH A, et al. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI [J]. Proteins: Structure, Function, and Bioinformatics, 2016, 84(Suppl 1): 4-14.

[9] MOULT J, FIDELIS K, KRYSHTAFOVYCH A, et al. Critical assessment of methods of protein structure prediction (CASP)--round x [J]. Proteins: Structure, Function, and Bioinformatics, 2014, 82(Suppl 2): 1-6.

[10] KEASAR C, MCGUFFIN L J, WALLNER B, et al. An analysis and evaluation of the WeFold collaborative for protein structure prediction and its pipelines in CASP11 and CASP12 [J]. Scientific Reports, 2018, 8(1): 9939.

[11] ANFINSEN C B, HABER E, SELA M, et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain [J]. Proceedings of the National Academy of Sciences, 1961, 47(9): 1309-1314.

[12] LEE J, FREDDOLINO P L, ZHANG Y. Ab Initio Protein Structure Prediction [M] // From Protein Structure to Function with Bioinformatics. Netherlands, Dordrecht: Springer, 2017: 3-35.

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

[14] 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.

[15] LI Z, SCHERAGA H A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding [J]. Proceedings of the National Academy of Sciences, 1987, 84(19): 6611-6615.

[16] KIHARA D, LU H, KOLINSKI A, et al. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints [J]. Proceedings of the National Academy of Sciences, 2001, 98(18): 10125-10130.

[17] LEE J. New Monte Carlo algorithm: Entropic sampling [J]. Physical Review Letters, 1993, 71(2): 211-214.

[18] PIANA S, LINDORFF-LARSEN K, SHAW D E. Atomic-level description of ubiquitin folding [J]. Proceedings of the National Academy of Sciences, 2013, 110(15): 5915-5920.

[19] LINDORFF-LARSEN K, MARAGAKIS P, PIANA S, et al. Picosecond to Millisecond Structural Dynamics in Human Ubiquitin [J]. Journal of Physical Chemistry B, 2016, 120(33): 8313-8320.

[20] PEARLMAN D A, CASE D A, CALDWELL J W, et al. Amber, a Package of Computer-Programs for Applying Molecular Mechanics, Normal-Mode Analysis, Molecular-Dynamics and Free-Energy Calculations to Simulate the Structural and Energetic Properties of Molecules [J]. Computer Physics Communications, 1995, 91(1/2/3): 1-41.

[21] CLAUSEN R, SHEHU A. A multiscale hybrid evolutionary algorithm to obtain sample-based representations of multi-basin protein energy landscapes [C] // Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM, 2014: 269-278.

[22] GARZA-FABRE M, KANDATHIL S M, HANDL J, et al. Generating, Maintaining, and Exploiting Diversity in a Memetic Algorithm for Protein Structure Prediction [J]. Evolutionary Computation, 2016, 24(4): 577-607.

[23] HAO X H, ZHANG G J, ZHOU X G. Conformational Space Sampling Method Using Multi-Subpopulation Differential Evolution for De novo Protein Structure Prediction [J]. IEEE Transactions on Nanobioscience, 2017, 16(7): 618-633.

[24] HAO X H, ZHANG G J, ZHOU X G, et al. A Novel Method Using Abstract Convex Underestimation in Ab-Initio Protein Structure Prediction for Guiding Search in Conformational Feature Space [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2016, 13(5): 887-900.

[25] HAO X H, ZHANG G J, ZHOU X G. Guiding exploration in conformational feature space with Lipschitz underestimation for ab-initio protein structure prediction [J]. Computational Biology and Chemistry, 2018, 73: 105-119.

[26] ZHOU X G, ZHANG G J. Differential Evolution With Underestimation-Based Multimutation Strategy [J]. IEEE Transactions on Cybernetics, 2018, PP(99): 1-12.

[27] ZHANG G J, ZHOU X G, YU X F, et al. Enhancing Protein Conformational Space Sampling Using Distance Profile-Guided Differential Evolution [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 14(6): 1288-1301.

[28] ZHOU X G, ZHANG G J, HAO X H, et al. Enhanced differential evolution using local Lipschitz underestimate strategy for computationally expensive optimization problems [J]. Applied Soft Computing, 2016, 48: 169-181.

[29] ZHOU X G, ZHANG G J. Abstract Convex Underestimation Assisted Multistage Differential Evolution [J]. IEEE Transactions on Cybernetics, 2017, 47(9): 2730-2741.

[30] ZHOU X G, ZHANG G J, HAO X H, et al. A novel differential evolution algorithm using local abstract convex underestimate strategy for global optimization [J]. Computers & Operations Research, 2016, 75(11): 132-149.

[31] RAKHSHANI H，IDOUMGHAR L，LEPAGNOT J，et al. Speed up differential evolution for computationally expensive protein structure prediction problems[J/OL]. https://doi.org/10.1016/j.swevo.2019.01.009.

[32] LEE J，SCHERAGA H A，RACKOVSKY S. New optimization method for conformational energy calculations on polypeptides：Conformational space annealing [J]. Journal of Computational Chemistry，1997，18(9)：1222-1232.

[33] ZHANG Y. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10 [J]. Proteins：Structure，Function，and Bioinformatics，2014，82：175-187.

[34] LI Z W，HAO X H，ZHANG G J. Replica Exchange Based Local Enhanced Differential Evolution Searching Method in Ab-initio Protein Structure Prediction [J]. Computer Science，2017，44(5)：211-217.

[35] SHEHU A，OLSON B. Guiding the Search for Native-like Protein Conformations with an Ab-initio Tree-based Exploration [J]. International Journal of Robotics Research，2010，29(8)：1106-1127.

[36] ROY A，KUCUKURAL A，ZHANG Y. I-TASSER：a unified platform for automated protein structure and function prediction [J]. Nature Protocols，2010，5(4)：725-738.

[37] ROHL C A，STRAUSS C E M，MISURA K M S，et al. Protein structure prediction using rosetta [J]. Methods in Enzymology，2004，383：66-93.

[38] HAO X H，ZHANG G J，ZHOU X G，et al. Protein Conformational Space Optimization Algorithm Based on Fragment-assembly[J]. Computer Science，2015，42(3)：237-240.

[39] 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.

[40] KC D B. Recent advances in sequence-based protein structure prediction [J]. Briefings in bioinformatics，2016，18(6)：1021-1032.

[41] ZHANG C，MORTUZA S M，HE B，et al. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12 [J]. Proteins：Structure，Function，and Bioinformatics，2018，86(Suppl 1)：136-151.

[42] ZHANG W X，YANG J Y，HE B J，et al. Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11 [J]. Proteins：Structure，Function，and Bioinformatics，2016，84：76-86.

[43] OVCHINNIKOV S，PARK H，KIM D E，et al. Protein structure prediction using Rosetta in CASP12 [J]. Proteins：Structure，Function，and Bioinformatics，2018，86(Suppl 1)：113-121.

[44] MOULT J，FIDELIS K，KRYSHTAFOVYCH A，et al. Critical assessment of methods of protein structure prediction (CASP) Round XII [J]. Proteins：Structure，Function，and Bioinformatics 2018，86：7-15.

[45] ADHIKARI B，BHATTACHARYA D，CAO R，et al. CONFOLD：Residue-residue contact-guided ab initio protein folding [J]. Proteins：Structure，Function，and Bioinformatics，2015，83(8)：1436-1449.

[46] JONES D T. Predicting novel protein folds by using FRAGFOLD [J]. Proteins：Structure，Function，and Bioinformatics，2001(Suppl 5)：127-132.

[47] DE OLIVEIRA S H P，DEANE C M. Combining co-evolution and secondary structure prediction to improve fragment library generation [J]. Bioinformatics，2018，34(13)：2219-2227.

[48] ZHANG G J，MA L F，WANG X Q，et al. Secondary Structure and Contact Guided Differential Evolution for Protein Structure Prediction [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics，2018，PP(99)：1-1.

[49] DeepMind. AlphaFold：Using AI for scientific discovery [EB/OL]. (2018-12-02). https://deepmind.com/blog/alphafold/.

[50] JI S，ORUC T，MEAD L，et al. DeepCDpred：Inter-residue Distance and Contact Prediction for Improved Prediction of Protein Structure [J]. PLOS ONE，2019，14(1)：e0205214.

[51] WANG S，LI W，ZHANG R，et al. CoinFold：a web server for protein contact prediction and contact-assisted protein folding [J]. Nucleic Acids Research，2016，44(W1)：W361-W366.

[52] WANG S，SUN S Q，LI Z，et al. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model [J]. Plos Computational Biology，2017，13(1)：e1005324.

[53] MA J Z，WANG S，WANG Z Y，et al. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning [J]. Bioinformatics，2015，31(21)：3506-3513.

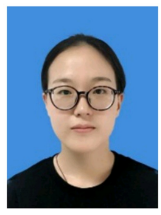
[54] ABRIATA L A，TAMO G E，MONASTYRSKYY B，et al. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods [J]. Proteins：Structure，Function，and Bioinformatics，2018，86：97-112.

[55] ZHANG Y，SKOLNICK J. SPICKER：A clustering approach to identify near-native protein folds [J]. Journal of Computational Chemistry，2004，25(6)：865-871.

[56] CHIVIAN D，KIM D E，MALMSTROM L，et al. Automated prediction of CASP-5 structures using the Robetta server [J]. Proteins：Structure，Function，and Bioinformatics，2003，53：524-533.

[57] ZHANG Y，SKOLNICK J. Scoring function for automated assessment of protein structure template quality [J]. Proteins：Structure，Function，and Bioinformatics，2004，57(4)：702-710.

[58] XU J，ZHANG Y. How significant is a protein structure similarity with TM-score= 0.5? [J]. Bioinformatics，2010，26(7)：889-895.



XIE Teng-yu，born in 1993，postgraduate. Her main research interests include intelligent information processing，optimization theory and algorithm design and bioinformatics.

ZHANG Gui-jun，born in 1974，Ph.D，professor，Ph.D supervisor，is a member of China Computer Federation (CCF). His main research interests include intelligent information processing，optimization theory and algorithm design and bioinformatics.