ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 4,Jul. 2021,pp. 629－638
DOI:10. 16337/j. 1004-9037. 2021. 04. 001

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

# 基于片段组装的蛋白质结构预测方法综述

张贵军，刘　俊，赵凯龙

（浙江工业大学信息工程学院，杭州 310012）

**摘　要：**蛋白质三维结构决定了其特殊的生物功能，蛋白质三维结构对蛋白质功能研究、疾病的诊断与治疗、创新药物研发都有着重要的科学意义。利用计算机技术从氨基酸序列预测蛋白质三维结构是获取蛋白质三维结构的有效方法。片段组装是一种广泛采用的蛋白质结构预测技术，它将连续的构象空间优化问题转换成离散的实验片段组合优化问题，从而有效地减小了构象搜索空间。首先介绍了片段组装技术；其次总结了基于片段组装的蛋白质结构预测的发展历程，并对部分具有代表性的方法进行了简要阐述；然后介绍了蛋白质结构预测研究中常用的数据库和评价指标，并比较了不同预测方法的性能；最后分析并指出了当前基于片段组装的蛋白质结构预测方法所存在的挑战性问题，并对该领域未来的研究方向进行了展望。

**关键词：**蛋白质结构预测；片段组装；进化算法

**中图分类号：**TP391　　**文献标志码：**A

## Review of Protein Structure Prediction Methods Based on Fragment Assembly

ZHANG Guijun, LIU Jun, ZHAO Kailong

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310012, China)

**Abstract:** The 3D structure of protein determines its special biological function. The 3D structure of protein has important scientific significance for protein function research, disease diagnosis and treatment, and innovative drug research and development. It is an effective method to predict protein 3D structure from amino acid sequence by computer. Fragment assembly is a widely used technique for protein structure prediction, which can effectively reduce the conformational search space by converting continuous conformational space optimization into discrete experimental fragment combination optimization. This paper first introduces the technology of fragment assembly. Next, the development of protein structure prediction based on fragment assembly is summarized, and some typical prediction methods are briefly described. The commonly used databases and evaluation indexes in protein structure prediction are then demonstrated, and the performance of the representative prediction methods is compared. Finally, we analyse and point out the challenges of the current protein structure prediction methods based on fragment assembly, and look forward to the future research directions in this field.

**Key words:** protein structure prediction; fragment assembly; evolutionary algorithm

---

**基金项目：**国家自然科学基金(61773346)资助项目；浙江省自然科学基金重点项目(LZ20F030002)资助项目。
**收稿日期：**2021-05-14；**修订日期：**2021-07-01

## 引　言

蛋白质是生命活动的主要承担者，几乎支撑着生命的所有功能，细胞内发生的大部分反应都依赖于蛋白质。蛋白质的功能取决于其独特的三维结构，也就是常说的“结构决定功能”。随着2003年人类基因组计划宣布完成[1]，由DNA或RNA转译为蛋白质氨基酸序列的第一遗传密码已被破解，然而蛋白质序列折叠成特定的三维结构才能够执行其特定的功能。蛋白质序列如何折叠形成独特的三维结构仍然是未解之谜[2]。《Science》杂志在纪念创刊125周年之际，把“能否预测蛋白质折叠？”列为21世纪125个科学前沿问题之一[3]。因此，对蛋白质折叠过程的深入研究，对于直接、准确地分析蛋白质的生物学功能和解释各种生命活动现象至关重要，将为相关疾病的诊断与治疗、创新药物研发奠定基础。

目前，主要通过X射线衍射、核磁共振和冷冻电镜等生物实验手段来测定蛋白质的三维结构，这些方法不仅费钱费力，而且周期长，导致已测定蛋白质结构的数量远远低于已测定蛋白质的序列数量。2021年4月最新统计数据显示，UniProtKB/TrEMBL数据库中共存储蛋白质序列214 406 399条（数据来源于http://www.ebi.ac.uk/uniprot/TrEMBLstats），其中177 426条序列结构被实验测定（数据来源于http://www1.rcsb.org/stats/growth/growth-released-structures），仅占序列总数的0.083%，而且这一差距仍然在不断增加。显然，实验测定方法无法满足高效获取蛋白质结构的需求。

在理论研究和实际应用双重需求的推动下，依据Anfinsen准则[4]，通过计算机技术根据氨基酸序列预测三维结构的蛋白质结构预测取得了蓬勃发展。CASP竞赛是由美国科学家Moult发起的蛋白质结构预测技术关键评估活动，能够客观地反映蛋白质结构预测领域发展的最新技术水平，是蛋白质结构预测领域的奥林匹克竞赛[5]。CASP竞赛每两年举行一次，自1994年创办至今已举办14届。CASP根据目标蛋白预测难易程度分为基于模板（Template-based modeling，TBM）和无模板（Free modeling，FM）两类建模方法[6]。一般来讲，TBM方法中目标蛋白可以从PDB（Protein data bank）结构数据库中检测到同源模板，建模精度基本能够达到实验测定水平[5]；然而，由于无法获得同源模板，FM类目标蛋白必须采用从头预测方法，成为CASP中最具挑战、也是最受关注的一类研究问题。能量模型的复杂性和构象空间采样瓶颈是限制从头预测方法发展的主要原因[7]。

从头蛋白质结构预测不受限于模板信息，能够正确预测具有未发现的整体拓扑结构的蛋白质结构，一直受到生物信息学领域和进化计算社区的高度关注。片段组装技术在从头蛋白质结构预测领域应用广泛，事实证明片段组装方法是最有前景的蛋白质结构预测方法之一[6,8]。本文结合国内外研究现状以及本课题组开展的一些研究工作，针对基于片段组装的蛋白质结构从头预测方法的研究进展进行分析和综述。

## 1　片段组装

由于蛋白质构象空间的高维特性，在巨大的构象空间中进行采样是不合适的。片段组装技术利用已测定蛋白质结构的局部信息，将每一个残基的二面角约束在一组离散值内，从而极大地缩小了构象搜索空间[9-10]。在蛋白质结构预测中，片段组装技术分为3个步骤：首先，随机在目标序列上选择一个包含若干个（一般为3个或9个）连续残基的插入窗口；然后，从该窗口对应的片段库中随机选择一个片段替换该窗口对应的片段；最后，采用能量函数计算片段替换前后构象的能量差值，并根据Metropolis准则判断是否保留片段组装后的构象[9]。图1为长度为$L$的蛋白质进行3残基片段组装的示意图。

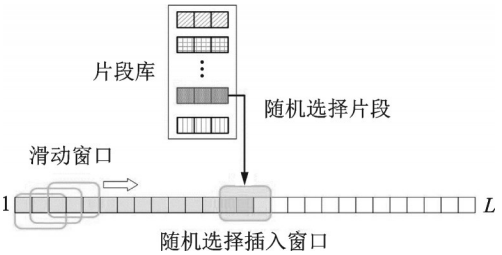


图1　片段组装示意图

Fig.1　Schematic diagram of fragment assembly

片段组装利用PDB数据库中已测定蛋白质结构的短且连续的片段信息,在基于知识力场构建的能量函数的引导下,不断组合向天然态折叠,既利用了已知蛋白质的结构信息,同时避免了同源建模方法高度依赖模板质量的缺陷[11]。

## 2 国内外研究现状

蛋白质结构预测一直受到计算生物学领域和计算智能社区的高度关注,是一个前沿研究课题[12]。1994年,Bowie和Eisenberg首次从PDB中提取序列长度为9的小片段来组装形成一个新的三维结构[13]。此后的二十多年间,片段组装成为广泛使用的从头蛋白质结构预测方法。虽然自从2016年CASP12深度学习在蛋白质残基接触/距离预测取得重大突破后,基于距离约束的几何优化方法逐渐占据主导地位[14-15],但从CASP12至CASP14(2016—2020年)的结果中可以发现片段组装方法仍然是最具竞争力的从头蛋白质结构预测方法之一[8,16]。国内外研究学者针对基于片段组装的结构预测做了大量深入研究[17-18],本文将从经典的片段组装结构预测方法、基于进化算法的片段组装方法和残基接触距离辅助的片段组装方法这3方面进行介绍。

### 2.1 经典片段组装方法

华盛顿大学Baker实验室开发的Rosetta[11,19]是较早采用片段组装技术的从头蛋白质结构预测方法。在Rosetta中,已知结构的短片段通过蒙特卡洛策略组装,以产生类似天然的蛋白质构象。Rosetta通过能量力场来描绘蛋白质折叠过程中不同状态的构象,根据热力学假说,天然态的蛋白质结构对应于自由能最低的构象,通过最小化构象能量获取近天然态构象。由于蛋白质构象空间极其复杂,为了提高采样效率,通常采用Rosetta低分辨率能量函数score3来减小自由度,同时保留重要信息。能量函数score3由10个能量项组成,反映原子排斥、氨基酸倾向、残基环境、残基对相互作用、二级结构元素之间的相互作用、密度和紧致性等,其定义为[11]

$$E_{\text{score3}} = w_{\text{vdw}}E_{\text{vdw}} + w_{\text{cenpack}}E_{\text{cenpack}} + w_{\text{pair}}E_{\text{pair}} + w_{\text{env}}E_{\text{env}} + w_{\text{cbeta}}E_{\text{cbeta}} + w_{\text{rg}}E_{\text{rg}} + w_{\text{hs}}E_{\text{hs}} + w_{\text{ss}}E_{\text{ss}} + w_{\text{rsigma}}E_{\text{rsigma}} + w_{\text{sheet}}E_{\text{sheet}} \tag{1}$$

Rosetta片段组装折叠模拟主要分为4个阶段,在每个阶段采用不同的能量函数,每个能量项的权重逐渐增加。在Rosetta的前3个阶段使用残基数目为9的片段执行片段组装,实现大规模的构象空间探索,在第4阶段使用残基数目为3的片段来更精细地调整构象拓扑结构。Rosetta的每个阶段执行大量的片段插入,并根据片段插入情况动态调整温度因子。当片段插入连续失败150次,通过提高温度因子来降低构象接受的条件,从而提高片段插入成功率;当片段插入成功后,将温度因子恢复为初始值。为了生成可靠的蛋白质模型,通常需要运行成千上万次的片段组装折叠模拟最终生成最低能量模型,这是一个极其耗时且消耗计算代价的过程。

密西根大学张阳实验室开发的QUARK[20-21]是另一个优秀的基于片段组装的从头蛋白质结构预测方法。QUARK使用的片段长度为1至20个残基,采用基于知识的复合力场引导的副本交换蒙特卡洛来从片段组装全长结构模型。为了便于力场的发展和搜索引擎的设计,QUARK采用半简化模型,用主干原子和侧链质心来表示蛋白质残基。对于查询序列,首先通过神经网络预测各种结构特征。然后通过副本交换蒙特卡洛模拟,将无缝穿线生成的小片段组装起来,从而生成全局折叠。QUARK设计了包含11个能量项的复合力场来引导构象搜索,总能量的计算公式为[20]

$$E_{\text{tot}} = E_{\text{prm}} + w_1E_{\text{prs}} + w_2E_{\text{ev}} + w_3E_{\text{hb}} + w_4E_{\text{sa}} + w_5E_{\text{dh}} + w_6E_{\text{dp}} + w_7E_{\text{rg}} + w_8E_{\text{bab}} + w_9E_{\text{hp}} + w_{10}E_{\text{bp}} \tag{2}$$

式中:$E_{\text{prm}}$、$E_{\text{prs}}$和$E_{\text{ev}}$为原子级能量项,分别表示主链原子对势能、侧链中心成对势能和排除体积;$E_{\text{hb}}$、$E_{\text{sa}}$、$E_{\text{dh}}$和$E_{\text{dp}}$为残基级能量项,分别表示氢键作用力、溶剂可及性、主链扭转角势能和基于片段的距离谱能

量；$E_{rg}$、$E_{bab}$、$E_{hp}$和$E_{bp}$为拓扑级的能量项，分别表示回转半径、$\beta$-$\alpha$-$\beta$惩罚项、$\alpha$-$\alpha$能量项和$\beta$对能量项。

QUARK设计了11个局部构象运动来增强算法的采样能力，这些局部运动分为残基级、片段级、拓扑级3个层次，在40个平行副本中运行蒙特卡洛模拟。虽然在低温下的模拟可以探测到较低能量的构象，但很容易陷入到局部能量盆地中。副本交换的目的是利用高温副本模拟帮助低温副本跳出局部低能源盆地。因此，对于交换每一对相邻的副本，保持高接受率是必要的。每个副本在每个周期内单独运行，其中将根据Metropolis准则尝试$30L^{1/2}$（$L$是蛋白质长度）次局部运动。在一个运行周期完成后，将尝试在每两个相邻副本之间进行互换操作，交换它们的诱饵构象。互换操作也遵循Metropolis准则。与Rosetta单纯的片段替换相比，QUARK模拟包含了自由链结构的复合运动和结构之间的片段替换。这些技术极大地提高了构象搜索的灵活性和效率。

除Rosetta和QUARK之外，FRAGFOLD[22]、SCRARTCH[23]、PROFESY[24]等一系列方法都属于早期典型的基于片段组装的从头蛋白质结构预测方法。

### 2.2 基于进化算法的片段组装方法

进化算法[25-26]是一种基于自然选择和遗传变异等生物进化机制的全局性搜索算法，是研究蛋白质构象优化的一类重要方法。进化算法通过交叉和变异算子以及选择策略来模拟生物进化过程，提高算法的可靠性。进化算法在蛋白质结构预测领域应用广泛，总体实现流程如下：①通过随机片段组装生成包含若干个构象的初始种群；②对种群中的父代构象进行交叉和变异操作，生成子代构象；③计算子代构象的能量，通过选择策略判断是否用子代构象替换父代构象；④迭代步骤②和③，直到满足终止条件。由于蛋白质的高维特性，能量景观中存在着大量的局部能量陷阱，蒙特卡洛算法极易陷入局部能量陷阱，使算法早熟。在进化算法的框架下通过片段组装来预测蛋白质的三维结构，无须重复运行大量独立轨迹，能够实现种群中构象的信息交互，从而提升算法的采样效率和预测精度。

Garza-Fabre等在Rosetta片段组装协议的基础上提出多阶段模因算法RMA（Rosetta-based memetic algorithm）[27]。RMA分为4个阶段，每个阶段都是基于标准Rosetta片段组装的相应阶段设计的。第1阶段，利用Rosetta第1阶段的随机片段组装进行种群初始化，得到一组多样化的初始构象。在第2、3和4阶段，首先将Rosetta相应阶段作为局部搜索更新种群中的构象；然后将种群中的构象进行两两配对作为父代构象，利用预测的二级结构信息设计了基于loop区域残基的重组和突变遗传算子，通过对每对父代构象执行遗传算子操作生成子代构象；最后在生存选择环节，通过同时考虑构象的能量和多样性从父代和子代构象中选择较优构象构建新的种群。日本理化学研究所的Zhang研究小组提出了基于统计学原理的随机优化算法EDA（Estimation of distribution algorithm）的蛋白质结构预测方法EdaFold[28]，通过统计构象搜索过程中的进化信息来指导当前构象搜索，以生成优秀的新构象。基于全原子力场模型，在EdaFold算法基础上，该研究小组进一步提出了$EdaFold_{AA}$方法[29]和基于聚类变异更新策略的$EdaFold_{C}$方法[17]。Baker团队在Rosetta基础上开发了迭代的杂化协议，整个迭代过程由进化算法指导，在每次迭代时将杂化作为变异或交叉操作，并控制构象的多样性以防止快速收敛[19]，此外，根据预测的接触图与PDB中已知结构的接触图对齐以进行折叠识别，并利用接触图比对工具map_align来挑选不连续的片段，进一步整合宏基因组数据，为614个目前结构未知的蛋白质家族生成模型[30]。

本课题组在基于进化算法的片段组装方面进行了深入研究。由于蛋白质的高维特性，需要搜索的构象空间过于庞大，传统的片段组装方法通常分为多个阶段来搜索构象空间。针对不同蛋白质的阶段切换问题，本课题组提出了包含探索和增强两阶段的群体蛋白质结构预测算法PAIE[18]，旨在通过基于熵的阶段切换策略和基于扭转角分布的选择策略来克服当前多阶段算法的局限性，确保适当搜索构象空间并进一步增强算法的探索能力。此外，根据探索阶段构象的扭转角分布设计了一种选择策略，并将其应用于增强阶段。针对能量模型的不精确性，提出了一种基于距离谱引导的差分进化算法DP-

DE[31]。在DPDE中,设计了一种基于距离谱的选择策略来指导构象空间采样,除能量外,将残基-残基距离作为一种辅助构象评估指标,以补偿能量函数的不准确性,并基于距离分布设计了一个距离接受概率,用于选择构象。当试验构象的能量低于目标构象的能量时,试验构象直接被下一代接受。否则,首先计算从片段库中提取的残基-残基距离与从构象中所有残基对的距离分布图获得的预测残基-残基距离之间的平均距离误差。如果试验构象在平均距离误差方面优于目标构象,则计算试验构象的距离接受概率,并根据距离接受概率接受试验构象。该策略保留了具有更高能量但更合理结构的构象。通过使用基于距离谱的选择策略引导采样,提高了算法逃逸局部能量陷阱的能力和搜索效率。在距离谱研究的基础上,进一步提出了一种基于距离特征的两阶段蛋白质结构预测优化算法TDFO[32]。通过二分K-means算法提取距离谱中的特征信息用于构建构象相似性评估指标,并设计了基于构象相似度的选择策略引导构象采样,在一定程度上降低了不精确的能量模型的影响,同时提高了采样过程中构象的多样性;此外,根据算法的不同阶段提出了两种变异算子,并设计了一种状态估计模型以实现不同搜索阶段的平衡。

### 2.3 残基间接触距离辅助的片段组装方法

自CASP12以来,基于深度学习的蛋白残基间接触(contact)预测和距离(distance)预测取得了重大进展,使得结构预测精度显著提升[5,14]。蛋白质的多序列比对蕴含着序列的进化信息,根据残基对的共变特性可以推断出它们在空间中的位置关系(是否接触或距离),研究表明远程contact对结构预测非常有帮助,而残基间距离分布为蛋白质折叠提供了更加丰富和细粒度的约束信息。残基接触和距离预测的成功也进一步推动了基于片段组装蛋白质结构预测的发展。

早在2014年,Jones团队就发现了将基于片段组装的折叠算法FRAGFOLD与残基间接触预测方法PSICOV相结合的潜在好处[33]。将PSICOV预测的残基间接触作为能量项添加到FRAGFOLD现有的能量函数中,并通过模拟退火将超二级结构片段和长度固定的短片段组装成三维结构。结果证明,使用残基间接触的FRAGFOLD的预测精度得到了显著提升。在2016年的CASP12中,张阳课题组将基于序列预测的残基间接触约束信息加入I-TASSER和QUARK中,使得QUARK在FM目标的前5个预测模型中最好模型的平均精度提高了37%[8];在CASP13中,张阳团队发布了C-I-TASSER和C-QUARK,将残基间接触信息进一步优化为一个新的接触势能项,与包括基于穿线的距离约束和基于固有知识(物理势能)在内的其他能量项相平衡,以指导结构组装模拟折叠目标蛋白[16];2020年11月召开的CASP14会议摘要显示,基于深度学习预测的残基间距离和扭转角也被整合到I-TASSER和QUARK之中,以进一步提升结构预测精度。

本课题组在片段组装的基础上,结合残基间接触距离信息,提出了一些有效的采样策略和优化方法来提升蛋白质结构预测的精度和效率。利用残基间接触和二级结构信息,设计了基于二级结构和残基-残基接触的选择策略来引导构象采样,分别用于提高算法在构象空间中探索近天然二级结构区域和合理结构的能力;此外,还设计了一个概率函数来平衡这两种选择策略;实验结果表明,该方案可以提高近天然态结构的采样能力[34]。在前期距离谱辅助片段组装研究的基础上,引入了残基间接触约束,提出了一种残基接触和距离谱耦合的结构预测方法CoDiFold[35]。设计了两个基于残基接触和距离谱的能量项,并将其融合到Rosetta低分辨率能量函数score3中;由两个残基接触联合的接触能量项用来约束构象,在基于接触的距离谱能量项中,利用接触信息来减弱或增强距离谱的约束;两个能量项的设计是为了缓解低分辨率能量函数的不准确性,提高模型能量与预测精度的相关性;针对搜索过程中容易陷入局部极小值的问题,设计了3种不同的变异策略,以提高开发和勘探的性质。针对结构灵活的蛋白质结构loop区域,提出了一种基于全局探索和loop扰动的残基接触辅助的从头蛋白质结构预测方法CGLFold[10]。使用过滤后的残基间接触信息构建选择模型指导构象采样。在全局探索阶段,利用片

段重组和片段组装大规模构象空间并生成近似天然态的拓扑结构;在loop扰动阶段,设计了loop区域特定的局部扰动模型,并通过差分进化算求解扰动量,进一步提高构象的精度。实验结果表明,loop扰动可以对拓扑进行微小的调整,这种微小的调整不断累积最终产生可观的增益,进而显著提高预测模型的精度。由于能量力场的不完善,计算蛋白质折叠模拟中的数学最优解并不总是对应于最优结构,传统的构象采样算法难以跨越高能障碍物,容易陷入局部盆地。针对该问题,在最新的研究中课题组提出了两种多模态蛋白质结构预测方法[36-37]。实验结果显示,通过多模态优化算法可以有效避免采样浪费或者采样不充足,并且能够采样到更多具有多样性的近天然态构象,从而显著提升结构预测的效率和精度。

随着残基接触和距离预测精度的不断提升,基于几何优化的蛋白质结构建模方法得到了广泛应用。这类方法没有采用片段组装等精巧的折叠方法,而是利用预测的contact或distance构建几何约束,通过CNS或梯度下降能量极小化协议生成结构模型。CONFOLD[38]、RaptorX[39]和DMPfold[40]等方法将contact或distance以及其他约束送入CNS中来生成模型;AlphaFold[41]和trRosetta[42]等方法将预测的distance分布转化成蛋白质特定的统计势能函数,并与经典的能量函数相结合,通过梯度下降能量极小化协议生成模型。

## 3 蛋白质结构预测实验评测

### 3.1 相关的蛋白质数据库

PDB[43]蛋白质结构数据库由美国Brookhaven国家实验室于1971年创建,由结构生物信息学研究合作组织维护,是最全的结构数据库,收录了通过实验方法测定的蛋白质结构。PDB数据库中收集了蛋白质、多糖、核酸和病毒等生物大分子的三维结构数据,这些数据可通过X射线单晶衍射、核磁共振和电子衍射等实验方法测定。通过互联网信息门户和可下载的数据档案可以访问大型生物分子(蛋白质、DNA和RNA)的三维结构数据。

UniProt[44]数据库是收录信息最全面的蛋白质序列数据库,主要包括UniParc序列归档库、UniProtKB蛋白质知识库和UniRef序列参考库。UniProtKB知识库包含了蛋白质的序列数据和大量注释信息,分为由人工审阅和注释的Swiss-Prot数据库和计算分析的TrEMBL数据库;UniRef数据库按照序列相似度将UniProtKB和UniParc中的序列分为UniRef100、UniRef90和UniRef50三个数据集,可显著减小数据库大小,从而加快序列搜索的速度。

### 3.2 评价指标

均方根偏差(Root mean square deviation, RMSD)和TM-score[45]是两种常用的计算目标结构与参考结构相似度的评价指标。RMSD表示两个蛋白质结构经过结构的刚体旋转平移后计算原子间的平均距离,以Å为单位,1 Å=$10^{-10}$ m,RMSD值越小,表明两个结构越相似。通常主要考虑主链上Cα原子间的RMSD。假设对于某个目标蛋白,考虑预测蛋白质模型$P$与实验测定结构$P'$的$n$个原子,RMSD计算公式为

$$\mathrm{RMSD}(P,P')=\sqrt{\frac{1}{n}\sum_{i=1}^{n}((P_{ix}-P'_{ix})^2+(P_{iy}-P'_{iy})^2+(P_{iz}-P'_{iz})^2)} \tag{3}$$

式中:$(P_{ix},P_{iy},P_{iz})$和$(P'_{ix},P'_{iy},P'_{iz})$分别表示模型$P$和结构$P'$第$i$个原子的三维坐标。

TM-score也是通过刚体旋转平移比对结构的相似度。不同于RMSD的是,结构的局部差异对TM-score的影响较小。TM-score的大小不受蛋白质序列长度的影响,取值在(0,1]之间,其计算公式为[45]

$$\text{TM-score}=\max\left[\frac{1}{L_{\text{target}}}\sum_{i=1}^{L_{\text{aligned}}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{\text{target}})}\right)^2}\right] \tag{4}$$

式中:$L_{\text{target}}$为目标蛋白的序列长度;$L_{\text{aligned}}$为两个结构对齐区域的长度;$d_0$为距离归一化参数,$d_0(L_{\text{target}})=1.24\sqrt[3]{L_{\text{target}}-15}-1.8$;$d_i$为第$i$个残基对间的距离。两个结构越相似,它们之间的TM-score越大;当TM-score≥0.5时,表明两个结构的拓扑形状大致相同[46]。

### 3.3 几种常见基于片段组装的蛋白质结构预测方法的性能分析与比较

为了真实反映近几年基于片段组装的蛋白质结构预测方法的性能,本节根据最新的基于片段组装的结构预测相关论文进行了方法描述,并对论文中的实验结果进行性能分析与比较。

CGLFold[10]是一种基于全局探索和loop扰动的残基接触辅助的从头蛋白质结构预测方法。在全局探索阶段,利用片段重组和片段组装大规模构象空间并生成近似天然态的拓扑结构;在loop扰动阶段,设计了loop区域特定的局部扰动模型,并通过差分进化算求解扰动量,进一步提高构象的精度。如图2所示,loop扰动可以对拓扑进行微小的调整,这种微小的调整不断累积最终产生可观的增益,进而显著提高预测模型的精度。

MMpred[36]是一种distance辅助的多模态优化采样方法。如图3所示,在蛋白质结构预测过程中,能量模型的不准确性导致数学上的最优解不一定对应于天然态结构,而次优解或局部极小值解可能与之对应。MMpred包括模态探测、模态维持和模态增强3个阶段。在模态探测阶段,通过结构相似性快速评估模型来控制种群的多样性,在不同的低能量盆地中生成具有多样性的构象;在模态维持阶段,通过自适应聚类算法对种群进行划分,并调节蒙特卡罗模拟退火温度来实现模态的融合;在模态增强阶段,使用贪婪搜索策略加快模态收敛速度,并利用预测的残基间距离信息设计构象评分模型指导构象选择。

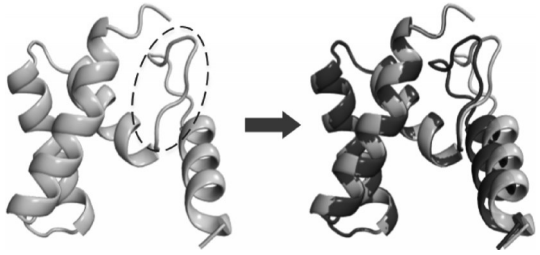

图2 loop扰动的示意图

Fig.2 Schematic diagram of loop perturbation

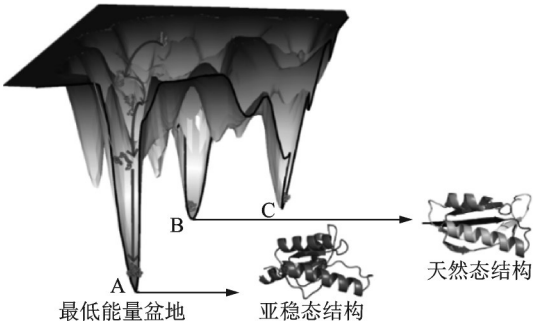


图3 蛋白质折叠的能量景观示意图

Fig.3 Schematic diagram of protein-folding energy landscape

表1给出了CGLFold、QUARK、BAKKER-ROSETTASERVER、MULTICOM_CLUSTER和RaptorX-Contact在14个CASP13的FM目标上的预测精度[10]。QUARK、BAKKER-ROSETTASERVER、MULTICOM_CLUSTER和RaptorX-Contact是CASP13中4种先进的服务器方法,其中QUARK在FM目标蛋白的搜索服务器组中排名第一。QUARK和CGLFold均是基于片段组装开发的算法,可以发现QUARK和CGLFold在这14个FM蛋白上取得了更高的平均预测精度。

表2给出了MMpred和Rosetta-d(distance约束的Rosetta片段组装方法)在320个非冗余基准测试蛋白上的平均预测结果[36]。MMpred与Rosetta-d使用了相同的片段库、distance约束和能量函数。可以发现在相同条件下,MMpred取得了更高的预测精度。

**表1 CGLFold、C-QURK、MULTICOM_CLUSTER、BAKER-ROSETTASERVER和RaptorX-Contact在14个CASP13的FM目标上的预测结果比较**[10]

**Table 1 Prediction results comparison of CGLFold, C-QUARK, MULTICOM _CLUSTER, BAKER-ROSETTASERVER, and RaptorX-Contact on the 14 FM targets of CASP13**[10]

| 预测方法 | TM-score |
|---|---|
| CGLFold | 0.49 |
| QUARK | 0.51 |
| MULTICOM_CLUSTER | 0.39 |
| BAKER-ROSETTASERVER | 0.48 |
| RaptorX-Contact | 0.48 |

**表2 MMpred和Rosetta-d（距离约束的Rosetta）在320个基准测试蛋白上的平均预测结果**[36]

**Table 2 Average prediction results of MMpred and Rosetta-d (Rosetta with distance constraints) on 320 benchmark proteins**[36]

| 预测方法 | 第一个模型 TM-score | 最优模型 TM-score |
|---|---|---|
| MMpred | 0.667 | 0.691 |
| Rosetta-d | 0.537 | 0.558 |

## 4 结束语

蛋白质三维结构的测定对疾病研究、诊断医疗和药物设计等有着重要的作用。然而，利用生物实验方法测定蛋白质结构耗时费力，代价极高。以计算机技术为手段实现蛋白质结构从头预测得到广泛关注。片段组装作为一种有效的插件式蛋白质构象空间优化技术，在蒙特卡洛构象优化算法中得到了广泛的应用。然而随着基于深度学习的残基间距离预测精度的不断提升，越来越多的方法直接采用几何优化方法来快速生成三维结构。为了进一步提升基于片段组装的蛋白质结构预测的性能，以下几个方面的研究方向是潜在的突破口。

（1）从已有研究成果来看，对于基于片段组装的蛋白质结构预测方法而言，构象空间采样仍然是一个瓶颈问题，尤其是随着蛋白质长度的增加构象空间呈几何倍数扩大。因此，设计高效的采样策略是提高算法效率和预测精度的关键之一。此外，片段组装将连续的二面角优化问题转换成了离散的实验局部结构的组合优化问题，虽然有效缩小了构象搜索，但也导致极有可能无法搜索最优解，并且随着蛋白质长度的增加这种影响会不断累计扩大。因此，如果能设计一个连续的二面角优化策略与离散的片段组装形成互补，将有望弥补片段组装这一固有缺陷。

（2）蛋白质能量模型不仅崎岖复杂，其构象搜索空间也十分庞大，这使得现有方法极易收敛到局部极值解。另外，即使搜索到全局最优解，能量模型的不准确性使得最优解不一定是最稳定的天然结构。进化计算社区的多模态优化方法，不仅能够发现全局最优解，而且可以获得更多样化的次优解，从而缓解能量模型的不准确性，提高搜索算法本身的稳定性（比如，全局最优解不一定对应于天然结构，某一个次优解可能更接近稳定的天然结构）。因此，基于群体的多模态优化方法是提高预测精度的重要保障。

（3）深度学习技术在蛋白质残基间距离预测中的成功应用使得蛋白质结构预测的精度取得了突破性进展，基于几何约束的能量极小化方法逐渐成为主流。然而，片段组装仍然具有其独特优势，既利用了已知蛋白质结构信息，又避免了同源建模方法高度依赖模板质量的缺陷，这使得片段组装方法能够正确预测具有未发现的整体拓扑结构的蛋白质结构。如果能够针对精细的残基间距离信息设计具有针对性的搜索算法，或是将能量极小化协议引入到构象采样过程中形成互补，可能会推动基于片段组装和基于几何约束能量极小化方法的进一步发展。

**参考文献:**


[1] 苏青, 陈广仁, 齐志红. 中国具有重大影响的50项科技事件(上)[J]. 科技导报, 2008, 26(13): 19-28.
SU Qing, CHEN Guangren, QI Zhihong. 50 china's influential science and technology events(I)[J]. Science & Technology Review, 2008, 26(13): 19-28.

[2] KOLATA G. Trying to crack the second half of the genetic code[J]. Science, 1986, 233(4768): 1037-1039.

[3] LISTED N. So much more to know[J]. Science, 2005, 309(5731): 78-102.

[4] ANFINSEN C. Principles that govern the folding of protein chains[J]. Science, 1973, 181(4096): 223-230.

[5] KRYSHTAFOVYCH A, SCHWEDE T, TOPF M, et al. Critical assessment of methods of protein structure prediction (CASP)-Round XIII[J]. Proteins: Structure Function and Bioinformatics, 2019, 87(12): 1011-1020.

[6] 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(S1): 1-14.

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

[8] ZHANG C X, MORTUZA S M, HE B J, 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, 2017, 86(S10): 136-151.

[9] KUHLMAN B, BRADLEY P. Advances in protein structure prediction and design[J]. Nature Reviews Molecular Cell Biology, 2019, 20(11): 681-697.

[10] LIU J, ZHOU X G, ZHANG Y, et al. CGLFold: A contact-assisted de novo protein structure prediction using global exploration and loop perturbation sampling algorithm[J]. Bioinformatics, 2020, 36(8): 2443-2450.

[11] ROHL C A, STRAUSS C, MISURA K, et al. Protein structure prediction using Rosetta[J]. Methods in Enzymology, 2004, 383: 66-93.

[12] 邓海游, 贾亚, 张阳. 蛋白质结构预测[J]. 物理学报, 2016, 65(17): 178701.
DENG Haiyou, JIA Ya, ZHANG Yang. Protein structure prediction[J]. Acta Physica Sinica, 2016, 65(17): 178701.

[13] BOWIE J U, EISENBERG D. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function[J]. Proceedings of the National Academy of Sciences of the United States of America, 1994, 91(10): 4436-4440.

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

[15] SENIOR A W, EVANS R, JUMPER J, et al. Improved protein structure prediction using potentials from deep learning[J]. Nature, 2020, 577(7792): 706-710.

[16] ZHENG W, LI Y, ZHANG C X, et al. Deep-learning contact-map guided protein structure prediction in CASP13[J]. Proteins: Structure, Function, and Bioinformatics, 2019, 87(3): 1149-1164.

[17] SIMONCINI D, SCHIEX T, ZHANG K. Balancing exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction[J]. Proteins-Structure Function & Bioinformatics, 2017, 85(5): 852-858.

[18] ZHANG G J, XIE T Y, ZHOU X G, et al. Protein structure prediction using population-based algorithm guided by information entropy[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(2): 697-707.

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

[20] DONG X, 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.

[21] 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.

[22] JONES D T. Predicting novel protein folds by using FRAGFOLD[J]. Proteins: Structure, Function, and Bioinformatics, 2001, 45(S5): 127-132.

[23] CHENG J L, RANDALL A Z, SWEREDOSKI M J, et al. SCRATCH: A protein structure and structural feature prediction server[J]. Nucleic Acids Research, 2005, 33: 72-76.

[24] LEE J, KIM S Y, JOO K, et al. Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing[J]. Proteins: Structure Function & Bioinformatics, 2010, 56(4): 704-714.

[25] STORN R. Differential evolution—A simple and efficient heuristic for global optimization over continuous space[J]. Journal of Global Optimization, 1997, 11(4): 341-359.

[26] 常珊, 陆旭峰, 王峰. 蛋白质-配体分子对接中构象搜索方法[J]. 数据采集与处理, 2018, 33(4): 586-594.
CHANG Shan, LU Xufeng, WANG Feng. Review of conformational searching method for protein-ligand molecular docking

[J]. Journal of Data Acquisition and Processing, 2018, 33(4): 586-594.
[27] 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.
[28] SIMONCINI D, BE RENGER F, SHRESTHA R, et al. A probabilistic fragment-based protein structure prediction algorithm [J]. PLoS ONE, 2012, 7(10): e38799.
[29] DAVID S, ZHANG K, ZHANG Y. Efficient sampling in fragment-based protein structure prediction using an estimation of distribution algorithm[J]. PLoS ONE, 2013, 8(7): e68954.
[30] OVCHINNIKOV S, PARK H, VARGHESE N, et al. Protein structure determination using metagenome sequence data[J]. Science, 2017, 355(6322): 294-298.
[31] 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, 2016, 14(6): 1288-1301.
[32] ZHANG G J, WANG X Q, MA L F, et al. Two-stage distance feature-based optimization algorithm for de novo protein structure prediction[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, 17(6): 2119-2130.
[33] KOSCIOLEK T, JONES D T. De novo structure prediction of globular proteins aided by sequence variation-derived contacts [J]. PloS ONE, 2014, 9(3): e92197.
[34] 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, 17(3): 1068-1081.
[35] PENG C X, ZHOU X G, ZHANG G J. De novo protein structure prediction by coupling contact with distance profile[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020. DOI: 10.1109/TCBB.2020. 3000758.
[36] ZHAO K L, LIU J, ZHOU X G, et al. MMpred: A distance-assisted multimodal conformation sampling for de novo protein structure prediction[J]. Bioinformatics, 2021. DOI: 10.1093/bioniformatics/btab484.
[37] XIA Y H, PENG C X, ZHOU X G, et al. A sequential niche multimodal conformation sampling algorithm for protein structure prediction[J]. Bioinformatics, 2020. DOI: 10.1093/bioinformatics/btab500.
[38] 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.
[39] XU J. Distance-based protein folding powered by deep learning[J]. Proceedings of the National Academy of Sciences, 2019, 116 (34): 16856-16865.
[40] GREENER J G, KANDATHIL S M, JONES D T. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints[J]. Nature Communications, 2019, 10(1): 1-13.
[41] SENIOR A W, EVANS R, JUMPER J, et al. Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction(CASP13)[J]. Proteins: Structure, Function, and Bioinformatics, 2019, 87 (12): 1141-1148.
[42] YANG J, ANISHCHENKO I, PARK H, et al. Improved protein structure prediction using predicted interresidue orientations [J]. Proceedings of the National Academy of Sciences, 2020, 117(3): 1496-1503.
[43] ROSE P W, BI C, BLUHM W F, et al. The RCSB protein data bank: New resources for research and education[J]. Nucleic Acids Research, 2013, 41: 475-482.
[44] 罗静初. UniProt蛋白质数据库简介[J]. 生物信息学, 2019, 17(3): 131-144.
LUO Jingchu. A brief introduction to UniProt[J]. Chinese Journal of Bioinformatics, 2019, 17(3): 131-144.
[45] 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.
[46] XU J, ZHANG Y. How significant is a protein structure similarity with TM-score=0.5[J]. Bioinformatics, 2010, 26(7): 889-895.

**作者简介:**

**张贵军**(1974-),**通信作者**,男,博士,教授,博士生导师,研究方向:结构生物信息学、计算智能与机器学习,E-mail: zgj@zjut. edu. cn。

**刘俊**(1994-),男,博士研究生,研究方向:结构生物信息学、计算智能与机器学习。

**赵凯龙**(1995-),男,博士研究生,研究方向:结构生物信息学、计算智能与机器学习。

(编辑:王静)