计算机科学
COMPUTER SCIENCE

远程模板检测算法及其在蛋白质结构预测中的应用

梁方，徐旭瑶，赵凯龙，赵炫锋，张贵军

引用本文

梁方，徐旭瑶，赵凯龙，赵炫锋，张贵军. 远程模板检测算法及其在蛋白质结构预测中的应用[J]. 计算机科学，2024，51(6A)：230600225-7.

LIANG Fang, XU Xuyao, ZHAO Kailong, ZHAO Xuanfeng, ZHANG Guijun. Remote Template Detection Algorithm and Its Application in Protein Structure Prediction [J]. Computer Science, 2024, 51(6A): 230600225-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600225

远程模板检测算法及其在蛋白质结构预测中的应用

梁 方 徐旭瑶 赵凯龙 赵炫锋 张贵军
浙江工业大学信息工程学院 杭州 310023
(zgj@zjut.edu.cn)

摘 要 在从传统力场驱动的蛋白质结构预测到当前数据驱动的AI结构建模的发展历程中,蛋白质结构模板检测是蛋白质结构预测中的关键环节,如何检测高精度蛋白质结构远程模板对提升结构的预测精度具有重要的研究意义。该研究提出了一种基于自适应特征向量提取的远程同源模板检测算法ASEalign。首先,采用多特征信息融合的深度学习技术预测蛋白质接触图;然后,设计了融合接触图、二级结构、序列谱谱比对和溶剂可及性等多维度特征打分函数,并通过自适应地提取接触图矩阵中的特征值和特征向量进行模板比对;最后,将检测出的高质量模板输入AlphaFold2中进行结构建模。在135个蛋白质的测试集上的结果表明,ASEalign相于主流的模板检测算法HHsearch精度提升了11.5%;同时,结构建模的精度优于AlphaFold2。

关键词:模板检测;模板建模;接触图预测;深度学习;二级结构
中图分类号 TP389

Remote Template Detection Algorithm and Its Application in Protein Structure Prediction

LIANG Fang,XU Xuyao,ZHAO Kailong,ZHAO Xuanfeng and ZHANG Guijun
School of Information Engineering,Zhejiang University of Technology,Hangzhou 310023,China

Abstract In the development process from traditional force field-driven protein structure prediction to current data-driven AI structure modeling,protein structure template detection is a key module in protein structure prediction,and how to detect high-precision protein structure remote templates is important to improve the prediction accuracy of structures. In this paper,a remote homology template detection algorithm ASEalign based on adaptive eigenvector extraction is proposed. Firstly,a deep learning technique of multi-feature information fusion is used to predict protein contact maps. Then,a multi-dimensional feature scoring function is designed to fuse contact maps,secondary structures,sequence profiles-profiles alignment and solvent accessibility,and the eigenvalue and eigenvector in the contact map matrix extracted by adaptive template alignment is performed. Finally,the detected high-quality templates are input to AlphaFold2 for structural modeling. Results on the test set of 135 proteins indicate that,compared to HHsearch,ASEalign improves the accuracy by 11.5%. Meanwhile,its accuracy of modeled structure is better than that of AlphaFold2.

Keywords Template detection,Template modeling,Contact map prediction,Deep learning,Secondary structure

1 引言

蛋白质在生命体的基础生物学活动中扮演着重要角色,是生物体内最重要的分子之一,它们在细胞内执行许多关键的功能,包括催化化学反应、传递信号、提供结构支持和运输分子等[1-2]。通过计算机技术预测蛋白质三维结构可为药物设计、疾病诊断和精准治疗提供重要的理论基础[3]。

模板对于蛋白质结构预测是非常重要的。传统的蛋白质结构预测包括模板建模(Template-based Modeling)和从头预测(Ab Initio Modeling)两种方法[4-5]。模板建模方法是利用已知的蛋白质结构作为模板来预测目标蛋白质的结构[6]。通过将目标蛋白质序列与已知结构的库进行比对,找到相似的模板结构,并将模板的结构信息映射到目标蛋白质上,生成初始的结构模型。常见的模板建模方法包括SWISSMODEL[7],RosettaCM[8]和MODELLER[9]。这些方法都依赖于已知的蛋白质结构作为模板,通过比对、拼接、约束等方式来预测目标蛋白质的结构。蛋白质结构从头预测法是在没有合适模板的情况下,从目标蛋白质的序列出发,通过搜索蛋白质构象空间中的最低自由能状态,来寻找最稳定的结构[10]。常见的蛋白质结构从头预测方法包括D-I-TASSER[6],Rosetta[11]和MMpred[12]。这些方法虽然没有显式地使用模板,但它们都使用了基于模板结构构建的片段信息[13]。这些

基金项目:国家自然科学基金(62173304);国家重点研发计划(2019YFE0126100)
This work was supported by the National Natural Science Foundation of China(62173304)and National Key Research and Development Program of China(2019YFE0126100).
通信作者:张贵军(zgj@zjut.edu.cn)

片段信息可以提供有关蛋白质的结构特征和构象约束。

除了传统的建模方法，基于机器学习的蛋白质预测技术也十分依赖于模板结构[14]。基于机器学习的蛋白质预测方法通常通过训练模型来学习蛋白质序列和结构之间的对应关系[6,15]。这些方法可以利用已知的蛋白质结构作为训练数据，提取结构特征和序列特征来建立预测模型[15]。此外，还可以结合传统的模板建模方法，通过将模板建模和机器学习技术相结合，综合利用多种信息源来进行结构预测，包括 AlphaFold2[16]，RoseTTAFold[17] 和 trRosetTaX[18]。这些方法将模板结构作为输入特征之一，与其他结构特征和序列特征一起输入机器学习模型中进行训练和预测，提高预测的准确性和泛化能力[19]。因此，开发一种高效的方法来识别高质量的远程同源模板对蛋白质结构预测是至关重要的。

总体来讲，现有的模板检测方法包括 3 类，分别是序列-序列比对、谱-谱比对和穿线法。序列-序列比对的模板检测方法包括 BLAST 系列[20]。BLAST 是将目标蛋白质序列与数据库中的已知结构蛋白质序列进行比对搜索，根据相似性得分和期望值来评估比对结果。PSI-BLAST[21] 是 BLAST 的一个改进版本，它利用迭代比对的方式增强序列相似性搜索的灵敏度。谱-谱比对的模板检测方法包括 HHsearch[22]，MUSTER[23] 和 SPARKS-X[24] 等。这些方法使用了序列谱信息、二级结构、溶剂可及性和扭转角来构建评分函数，相比序列-序列比对，提高了模板搜索的精度和覆盖率。穿线法包括 EigenTHREADER[25] 和 CEthreader[26] 等。这类方法首先预测目标序列的残基接触图并转化为特征向量表示，然后通过特征向量分解和动态规划算法的结合，可以在结构空间中寻找最佳的结构对齐。除此之外，元穿线法 LOMETS 系列[27-29] 集成了各类的模板检测方法，利用多种模板检测方法的优势，提高了蛋白质结构预测的准确性和覆盖范围。虽然这些方法在一定程度上提升了模板检测的精度，但仍然有很大的提升空间。

本文设计了一种基于特征向量自适应提取比对的远程同源模板检测算法（Adaptively Selected Eigenvector alignment，ASEalign）。其采用多特征信息融合的深度学习技术预测蛋白质接触图，并设计了多维度特征打分函数，通过自适应地提取接触图矩阵中的特征值和特征向量进行模板比对，提升了模板检测的效率和准确性。通过将检测出的模板输入 AlphaFold2 中进行结构建模，提升了 AlphaFold2 的模型精度，进一步验证了模板结构对蛋白质建模的重要性。

2 方法

远程同源模板检测算法 ASEalign 的流程图如图 1 所示。从序列出发，通过 HHblits[30] 搜索 UniRef30[31] 和 BFD 库生成 MSAs 并从中提取一维和二维特征，通过水平条带化将一维特征和二维特征进行组合，得到一个 $L\times L\times 490$ 的张量，然后将特征张量输入自注意力机制模块和卷积残差模块中预测出残基接触图。基于接触图、二级结构、序列谱谱比对和溶剂可及性设计的打分函数，自适应选取接触图矩阵中的特征值和特征向量与 PAcluter80[32] 模板库进行比对，检测出最终的模板结构。最后使用 AlphaFold2 预测器进行模板结构建模。

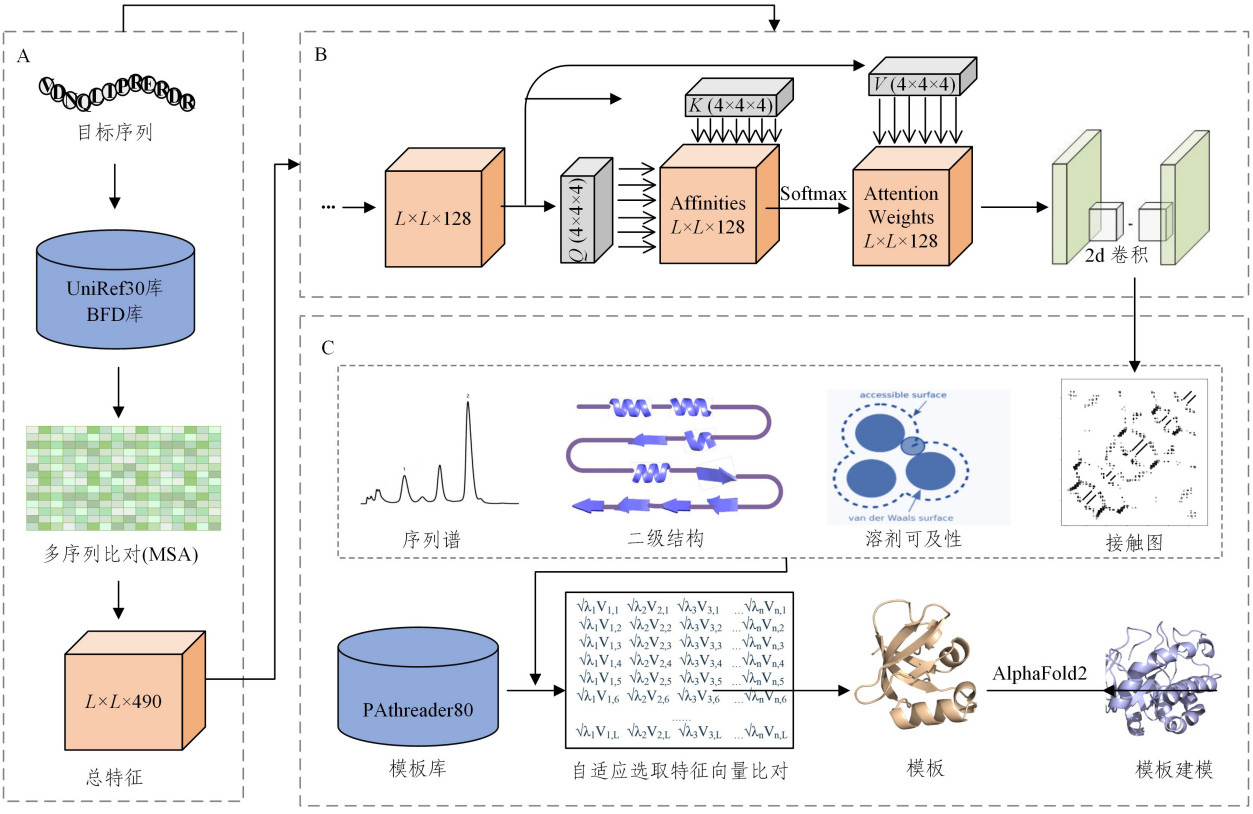


图 1 ASEalign 流程图

Fig. 1 Flowchart of ASEalign

2.1　蛋白质接触图预测

2.1.1　训练集和测试集的构建

PDB库[1)]收集了所有已被生物实验方法测定的蛋白质结构,提供了丰富的模板结构信息[33-34]。本文使用PDB数据库构建了用于蛋白质接触图预测的训练集。首先从PDB数据库中挑选由X射线衍射解析出的蛋白质,然后筛选出序列长度在40～500之间且分辨率小于2.5Å的蛋白质作为候选数据。使用CD-HIT工具[35]对收集到的蛋白质进行聚类,以30%的序列相似度为阈值去除冗余,最终得到13980条蛋白质序列作为训练集,其中95%用于训练,5%用于验证。

为了更加客观地评价方法的性能,本文从SCOPe 2.07[36]数据库中构建了测试集[37]。首先通过CD-HIT工具以30%的序列相似度为阈值去除冗余,提取到11198条蛋白质序列;然后选取长度在50～500之间且分辨率小于1.5Å的蛋白质作为候选数据;再一次使用CD-HIT工具以30%的序列相似度对训练集进行去冗余处理,最终得到135个序列作为测试集。

2.1.2　特征提取

首先为训练集中的每一条蛋白质序列搜索MSA。利用HHblits,将其E-value的阈值设置为0.001,序列覆盖率设置为至少50%,在UniRef30和BFD为训练集的蛋白质序列生成MSAs,并且从该目标序列以及MSAs中提取蛋白序列的一维特征和二维特征。一维特征包括序列频率谱、溶剂可及性信息、二级结构信息;二维特征包括协方差矩阵、CCMpred耦合分数[38]、残基对接触势能、去除背景噪声的互信息。

序列频率谱:表示残基在MSAs中出现的频率。不同的频率表征了氨基酸在蛋白质序列中的位置和特性。

溶剂可及性信息:指氨基酸在溶液中受到水分子包围的程度。溶剂可及性反映了氨基酸在溶剂中亲疏水性所引起的相互作用力。

二级结构信息:蛋白质的折叠过程中形成的局部稳定的α螺旋、β折叠等结构单元。

协方差矩阵:是通过MSA中的残基频率计算得到的,描述了蛋白质序列中任意两列残基之间的边缘分布和联合分布之间的相关性。其对角线元素表示残基的方差,非对角线元素表示不同残基之间的协方差,反映了结构中的相互作用。

CCMpred耦合分数:CCMpred描述残基间的非线性关系,通过马尔可夫随机场(Markov Random Field,MRF)学习MSA数据的生成模型,根据残基对接触势能的Frobenius范数应用平均乘积修正来消除传递相互作用。

残基对接触势能:不同氨基酸残基之间的接触势能,主要包括范德华力、氢键作用、静电吸引力等相互作用力引起的势能变化。

去除噪声的互信息:互信息用来表征MSA中固定位置残基在共进化过程中的共变程度。为了消除进化压力对氨基酸出现频率的影响,利用氨基酸背景频率修正互信息中的边缘频率分布,引入去除噪声的互信息。

2.1.3　网络搭建

采用4层注意力机制以及128个卷积残差块的网络模型。注意力机制可以对不同的残基特征进行加权学习和融合,使得模型能够更加关注那些对预测接触关系最为重要的特征,捕捉到不同残基的长程依赖关系。卷积神经网络通过卷积以滑动一个卷积核的形式来提取蛋白质特征,利用池化操作对特征图进行降维,再利用全连接层将卷积层和池化层提取的特征进行组合,最后引入非线性因素拟合更复杂的数据。多头注意力机制和卷积神经网络能够自动学习和提取蛋白质序列中的残基间进化关系,并输出其映射到的接触图。两者相互协作,更加全面地描述蛋白质的特征,提高了模型的泛化能力。

首先把提取的一维特征和二维特征进行组合,得到一个$L\times L\times 490$的特征张量。通过二维实例归一化层对特征数据进行规范化处理后,得到$L\times L\times 128$的特征张量,输入两层自注意力机制模块中,使用4头注意力同时关注多个空间特征,防止单一注意力机制将信息集中于自身。然后将得到的$L\times L\times 128$张量输入卷积模块。卷积模块包含128个残差块,每个残差块包含一个二维卷积层(卷积核为3 * 3)、一个批量归一化(BP)层、一个指数线性单元(ELU)激活层、一个dropout层(dropout rate 20%)和一个二维卷积层(卷积核在1,2,4,8,16核之间交替膨胀)。

2.1.4　模型训练

网络模型是用于预测目标蛋白质残基之间的接触图,残基之间的是否接触定义为残基C_β原子(甘氨酸为C_α原子)之间的距离是否小于等于8Å。在训练阶段,从训练集中选择95%的蛋白质用于训练,5%的蛋白质用于验证。为了防止训练过程中模型会形成记忆,在每一轮训练之前都会打乱蛋白质的顺序。训练时在序列距离大于4的残差对上计算损失,并使用预测和真实接触点之间的二元交叉熵作为损失函数,该网络模型的训练共迭代了50个epoch。

2.2　自适应接触矩阵特征分解

自适应接触矩阵特征分解是一种自适应选取接触矩阵的特征值和特征向量进行模板比对的算法。当矩阵的特征值较大时,特征空间所包含的信息量也较大。通过特征向量相关性最大的几个特征向量来近似接触矩阵,减少了比对次数。自适应特征向量比对方法能够根据特征值、特征向量灵活调整其贡献率,在保证算法精度的同时,提高了算法的效率。与固定特征值数量的模板比对算法相比,自适应选取特征向量的模板比对方法能保留较长序列蛋白质更多的特征值,具有较大的优势。因此,用两组少量特征向量的全局对齐计算两个接触图的重叠程度,能有效地提高匹配的准确度和鲁棒性。

接触矩阵$\boldsymbol{M}$是对角线均为0的实对称矩阵,有L个特征值和相对应的特征向量,根据谱图理论可知:

$$\boldsymbol{M}=V\Delta V^{-1}=V\Delta V^{T} \tag{1}$$

其中,λ_i表示第i个特征值,$i\in\{1,2,\cdots,L\}$以及$\vec{V}_i=[v_{1,i},\cdots,v_{L,i}]^T$表示中对应的特征向量。将特征值按递减顺序排序为$\lambda_1\geqslant\lambda_2\geqslant\lambda_3\geqslant\cdots\geqslant\lambda_i\geqslant\cdots\geqslant\lambda_L$。根据式(2)、式(3),对接触矩阵全部特征值总和的贡献比例进行如下计算:

$$C_i=\lambda_i/\sum_{i=1}^{L}\lambda_i\times 100\% \tag{2}$$

$$C_1+C_2+C_3+\cdots+C_i+\cdots+C_k\geqslant C \tag{3}$$

其中,C表示λ_i的贡献率。将贡献率累加计算,为使得接触矩阵贡献率之和刚好大于阈值的第一个特征值的索引值,$k\in\{1,2,\cdots,L\}$。其中,C是0.5的阈值。利用筛选出的k个特征值及其对应特征向量重塑接触矩阵得到接触矩阵的近似表达如式(4)所示:

1) http://www.rcsb.org/

$$\boldsymbol{M}\approx\sum_{i=1}^{k}\lambda_i\vec{V}_l*\vec{V}_l^{\mathrm{T}}$$
$$=\begin{bmatrix} v_{1,1} & \cdots & v_{1,k} & \cdots & 0\\ v_{2,1} & & v_{2,k} & \cdots & 0\\ \vdots & & \ddots & & \vdots\\ \vdots & & \vdots & & \vdots\\ v_{L,1} & \cdots & v_{L,k} & \cdots & 0 \end{bmatrix}\begin{bmatrix} \lambda_1 & \cdots & 0\\ 0 & & 0\\ \vdots & & \vdots\\ & \lambda_k & \\ 0 & \cdots & 0 \end{bmatrix}\begin{bmatrix} v_{1,1} & \cdots & v_{1,L}\\ \vdots & \cdots & \vdots\\ v_{k,1} & \ddots & v_{k,L}\\ \vdots & \vdots & \vdots\\ 0 & \cdots & 0 \end{bmatrix} \tag{4}$$
$$=\begin{bmatrix} \sqrt{\lambda_1}v_{1,1} & \cdots\sqrt{\lambda_k}v_{1,k} & \cdots & 0\\ \sqrt{\lambda_1}v_{2,1} & \sqrt{\lambda_k}v_{2,k} & \cdots & 0\\ \vdots & \ddots & & \vdots\\ \vdots & \vdots & & \vdots\\ \sqrt{\lambda_1}v_{L,1} & \cdots\sqrt{\lambda_k}v_{L,k} & \cdots & 0 \end{bmatrix}\begin{bmatrix} \sqrt{\lambda_1}v_{1,1} & \cdots\sqrt{\lambda_k}v_{1,k} & \cdots & 0\\ \sqrt{\lambda_1}v_{2,1} & \sqrt{\lambda_k}v_{2,k} & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ \vdots & \vdots & & \\ \sqrt{\lambda_1}v_{L,1} & \cdots\sqrt{\lambda_k}v_{L,k} & \cdots & 0 \end{bmatrix}^{\mathrm{T}}$$

则接触矩阵 $\boldsymbol{M}$ 残基 i 和 j 之间的接触信息可以近似为式(5)：

$$M_{i,j}\approx(\sqrt{\lambda_1}v_{i,1},\sqrt{\lambda_2}v_{i,2},\cdots,\sqrt{\lambda_k}v_{i,k},0,\cdots,0)*(\sqrt{\lambda_1}v_{j,1},\sqrt{\lambda_2}v_{j,2},\cdots,\sqrt{\lambda_k}v_{j,k},0,\cdots,0)^{\mathrm{T}} \tag{5}$$

其中，$[v_{i,1}\ \cdots\ v_{i,k}\cdots\ 0]$表示自适应选出 i 位置的特征向量，$i\in\{1,2,\cdots,L\}$，$[v_{j,1}\ \cdots\ v_{j,k}\ \cdots\ 0]$表示自适应选出位置 j 的特征向量，$j\in\{1,2,\cdots,L\}$。

2.3 模板检测打分函数设计

不同于传统的仅利用序列信息进行模板比对，本文根据多维度特征设计了更加精准的打分函数，通过多特征融合的评分函数来提高模板库筛选的准确性。使用动态规划算法进行序列-模板对齐打分。

首先，将目标蛋白 p_1 和模板蛋白 p_2 中的残基分别表示为 $R_1=\{1,\cdots,n\}$；$R_2=\{1,\cdots,m\}$。接触图分别是 $M^{P_1}\in\{0,1\}^{n\times n}$ 和 $M^{P_2}\in\{0,1\}^{m\times m}$。目标蛋白的第 i 个残基与模板蛋白的第 j 个残基接触匹配得分的计算式如(6)所示：

$$E(i,j)=w_1*E_c(i,j)+w_2*\sum_{r=1}^{R}E_r(i,j) \tag{6}$$

其中，$R=3$，$E_1(i,j)$，$E_2(i,j)$和$E_3(i,j)$分别表示序列谱谱比对得分、二级结构得分和溶剂可及性得分。详细介绍及参数可参见文献[39]。$E_c(i,j)$表示目标蛋白的第 i 个残基与模板蛋白的第 j 个残基的通过接触匹配的得分，计算式如式(7)和式(8)所示：

$$E_c(i,j)=E_{\mathrm{con}}(i,j)+E_{\mathrm{gap}}(i,j) \tag{7}$$

$$E_{\mathrm{con}}(i,j)=\vec{U}_n\vec{V}_m=\sum_{\substack{g=1\\ j\neq 0}}^{\min(k,t)}\sqrt{\lambda_g^{P_1}}v_{i,g}\sqrt{\lambda_g^{P_2}}u_{j,g} \tag{8}$$

其中，$\vec{U}_n$，$\vec{V}_m$分别表示 p_1 和 p_2 中的第 i 和第 j 个残基与其他残基形成接触的特征向量；k 和 t 分别表示自适应选取后特征向量的个数，$1\leqslant k\leqslant n$，$1\leqslant t\leqslant m$；$E_{\mathrm{gap}}(i,j)$是比对过程中加入 Gap 的惩罚分数。

3 结果分析

3.1 模板检测精度比较与分析

为了检验 ASEalign 方法的预测性能，本文在 135 个蛋白质上进行了模板检测性能的测试，并与基于隐马尔可夫的谱-谱比对方法 HHsearch 以及基于接触图比对的 EigenThreader进行了结果对比。使用 TM-score 评估了模板的精度，如图 2 所示。它考虑了两个结构的全局拓扑特征。TM-score 的取值范围在 0～1 之间，值越接近 1 表示两个结构越相似，值越接近 0 表示两个结构越不相似。ASEalign 在 135 个测试蛋白上检测模板的 TM-score 均值为 0.695，比 HHsearch (0.623)高出 11.5%，比 EigenThreader(0.636)高出 9.2%。当 TM-score ≥ 0.5 时，模板结构与天然蛋白的拓扑结构非常相似。统计两个算法检测的模板的 TM-score ≥ 0.5 的数量，ASEalign 有 110 个，占总测试集的 85%。这表明 ASEalign 的模板检测性能有了显著的提升。在 135 个测试蛋白质中，ASEalign 有 85 个蛋白质的模板检测结果好于 HHsearch。其余 50 个模板检测结果比 HHsearch 差，主要是因为 ASEalign 只搜索了 PAcluster80 的质心结构，这提升了模板检测的速度，但同时也损失了一部分模板的精度。

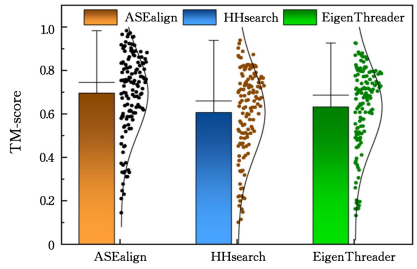
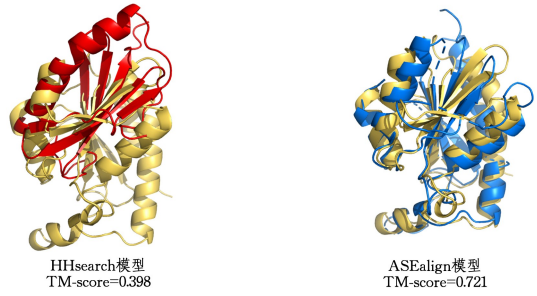


图 2 ASEalign 与 HHsearch 和 EigenThreader 检测模板的 TM-score 箱型图

Fig. 2 TM-score box plots of templates detected by ASEalign, HHsearch and EigenThreader

以肽基-tRNA 水解酶(PDB ID：6J93_A)为例，如图 3 所示。



注：红色和蓝色分别是 HHsearch 和 ASEalign 检测的模板，黄色是天然结构。

图 3 ASEalign 和 HHsearch 在肽基-tRNA 水解酶上检测出的模板与天然结构的比较

Fig. 3 Comparison of templates detected by ASEalign and HHsearch on peptidyl-trna hydrolases with native structure

本文对 ASEalign 和 HHsearch 检测的模板进行了详细的比较与分析。肽基-tRNA 水解酶是一种细菌酶，可裂解肽基-tRNA 或 *N*-酰基-氨酰-tRNA 以产生游离肽或 *N*-酰基-氨基酸和 tRNA。肽基-tRNA 水解酶是一个单一的 α/β 球状结构，具有 7 个 β 链，形成一个扭曲的中心 β 折叠，被 6 个螺旋包围。对于 HHsearch，它检测出的模板精度为 0.398，仅仅包含了中心的 β 折叠和一个 α 螺旋，没有识别出具有完整结构区域的模板。而 ASEalign 检测出的模板精度为 0.721，该模板基本覆盖了整个目标蛋白，结果显著好于 HHsearch。这

是因为该目标蛋白的 MSAs 数量仅有 653 条，HHsearch 基于少量的 MSAs 提取的隐马尔可夫谱信息是有限的，这降低了 HHsearch 的模板精度。而 ASEalign 使用同样数量的 MSA 信息，通过机器学习预测出了准确的接触图，通过提取特征向量进行比对检测出了精度更高的模板。这表明 ASEalign 基于接触图比对的模板检测能够比 HHsearch 基于谱-谱比对的模板检测获得更精确的模板结构。

为了进一步检验 ASEalign 远程同源模板的检测性能，本文在一周的 CAMEO 数据集(2023/08/19)上进行了远程模板(去除了大于等于 30%序列相似的同源模板)检测并与 HHpred 进行了比较，如图 4 所示，共包含了 16 个蛋白质。ASEalign 检测的模板的平均 TM-score 为 0.796，比 HHpred 的平均 TM-score 高出 5%。该结果再次表明 ASEalign 对远程同源模板的检测性能优异。

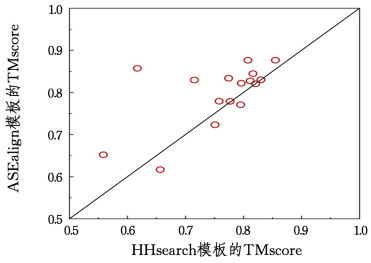


图 4　ASEalign 与 HHpred 在 CAMEO 数据集上检测的远程模板的比较

Fig. 4　Comparison of remote templates detected by ASEalign and HHpred on CAMEO dataset

3.2　自适应选取特征向量的结果分析

为了提高模板检测的准确性并加快模板检测速度，采用了自适应选取特征向量的策略进行模板比对，而不是固定数目的特征向量。特征向量的数目是根据特征值贡献率来选择的。特征值贡献率是接触矩阵每个特征值与所有特征值总和的比值。本文分别对 0.3～1.0 的特征值贡献率进行了实验，如图 5 所示。

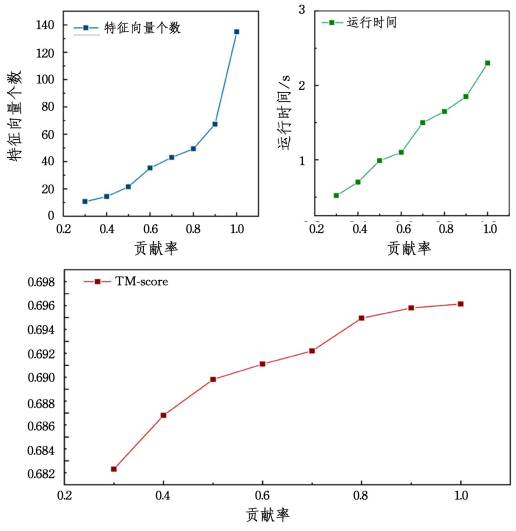


图 5　特征值贡献率与特征向量个数、运行时间和 TM-score 的关系

Fig. 5　Relationship between eigenvalue contribution rate and the number of eigenvectors, running time and TM-score

随着贡献率的逐步提高，特征向量的个数和模板精度(TM-score)都逐步增加。这表明越多的特征向量所包含的信息越丰富，将有助于提升模板检测的精度。然而，当贡献率增加时，模板检测的时间也随之增加。本文选择了特征值贡献率为 0.8 作为最终比对方案，既在 0.8 特征值贡献率阈值下，采用自适应选取前 80%的特征值及其对应的特征向量用于模板比对。在这个阈值下，可以保证特征向量比对数量在趋势激增之前，既保留了包含主要接触信息的特征值，避免浪费非必要的特征比对，同时减少了模板检测所运行的时间。对于长度不同的蛋白自适应地选取不同特征数量的信息进行比对，提高了模板检测的精度和效率。

为了检验自适应选择特征向量的有效性，本文与 CEthreader 做了进一步的比较，如表 1 所列。CEthreader 采用固定数量为 7 的特征向量进行模板比对；ASEalign 采用自适应选取特征向量比对的策略，在所有目标蛋白上选取的特征向量平均个数为 49。结果显示，ASEalign 的平均 TM-score 为 0.695，比 CEthreader 高出了 8.4%。这是因为在一些较长的蛋白中，固定的特征值数量可能会导致搜索模板时丢失一些额外的信息，从而降低模板比对的精度。相比之下，ASEalign 自适应选取的特征值基本覆盖了蛋白质残基间的接触信息，因此在保证搜索效率的前提下提升了模板检测的精度。

表 1　ASEalign 和 CEthreader 的模板检测结果

Table 1　Results of template detected by ASEalign and CEthreader

方法	平均特征向量个数	TM-score
CEthreader	7	0.641
ASEalign	49	0.695

3.3　模板增强的 AlphaFold2 建模

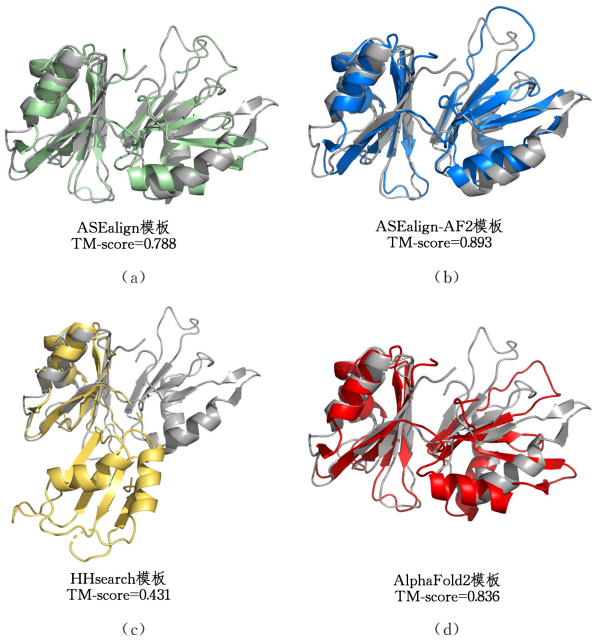
为了进一步验证 ASEalign 检测的模板的有效性，本文通过模板建模与 AlphaFold2 进行了比较。AlphaFold2 使用 HHsearch 搜索的前 4 个模板进行建模。本文用 ASEalign 检测的模板替换掉 HHsearch 的模板，并使用 AlphaFold2 进行了建模，并将方法命名为 ASEalign-AF2，结果如表 2 所列。AlphaFold2 预测的模型平均 TM-score 为 0.834，通过 ASEalign-AF2 模板增强的模板建模精度提升到了 0.839，较初始模型提升了 0.6%。在所有目标蛋白质中，其中有 76 个的结果比初始模型效果更好，占比 58%。在 RMSD 的比较上，经过模板增强后的 AlphaFold2 模型的平均 RMSD 值从 5.42Å 下降到了 5.14Å。这表明通过模板增强的建模确实可以提高 AlphaFold2 的预测准确性。

表 2　ASEalign-AF2 和 AlphaFold2 的建模结果

Table 2　Modeling results of ASEalign-AF2 and AlphaFold2

Model	RMSD	TM-score
AlphaFold2	5.42Å	0.834
ASEalign-AF2	5.14Å	0.839

以两域蛋白质 6J93_A 为例，本文对 ASEalign-AF2 和 AlphaFold2 的建模结果进行了分析，如图 6 所示。ASEalign-AF2 的建模精度为 0.893，AlphaFold2 的建模精度为 0.836。ASEalign-AF2 的模型精度显著高于 AlphaFold2，这主要得益于 ASEalign 提供了准确的模板信息。ASEalign 检测的模板结构的 TM-score 为 0.788，而 HHsearch 检测的模板结构的 TM-score 为 0.431。从图 6 可以看出，HHsearch 检测的模板在单域上的精度较好，而域间的方向是不正确的。ASEalign 检测的模板提供了正确的域方向，因此得到了较高精度的模型。这表明 ASEalign 可以提供准确的模板结构信息用于模板建模。



注：(a)和(b)表示 ASEalign 检测的模板(绿色)以及 ASEalign-AF2 建模(蓝色)的结果；(c)和(d)表示 AlphaFold2 中使用的 HHsearch 检测的模板(黄色)以及 AlphaFold2 建模(红色)的结果，灰色是天然结构。

图 6 在示例蛋白质 6J93_A 上 ASEalign 和 AlphaFold2 的模板和建模结果

Fig. 6 Templates and modeling results for ASEalign and AlphaFold2 on example protein 6J93_A

结束语 蛋白质的结构对于其功能和相互作用至关重要，因此准确地预测蛋白质的结构是理解其功能和设计新药物的关键一步。模板结构为蛋白质的预测提供了一个框架，可以指导目标蛋白质的结构建模。尤其是对于没有同源序列的孤儿蛋白，它的结构预测将十分依赖于远程同源模板。因此，检测出合适的模板结构对于预测高精度的蛋白质结构非常关键。

本文设计了一种基于蛋白质接触图自适应选取特征向量比对的远程同源模板检测算法 ASEalign。从序列出发，通过 HHblits 搜索 MSAs 并从中提取一维和二维特征，通过水平条带化将一维特征和二维特征进行组合，得到一个 $L\times L\times 490$ 的张量，然后将特征张量输入自注意力机制模块和卷积残差模块中预测出残基接触图。基于设计的多维度特征打分函数，自适应地选取接触图矩阵中的特征值和特征向量与 PAcluter80 模板库进行比对，检测出最终的模板结构并使用 AlphaFold2 预测器进行模板增强的结构建模。在 135 个蛋白的测试集上的结果表明，ASEalign 相比主流的模板检测算法 HHsearch 精度提升了 11.5%，并通过模板增强的蛋白质结构建模提升了 AlphaFold2 的单体模型精度。这表明 ASEalign 检测远程同源模板的性能优于目前主流的模板检测算法。

尽管 AI 技术的发展给单域蛋白质结构的预测带来了巨大的进步，但多域蛋白质建模以及复合物的组装仍然存在很大的挑战，这些都离不开模板提供的结构信息[40]。除此之外，模板检测在蛋白质功能注释、药物设计以及生物工程和蛋白质工程等领域都具有重要的用途[19,41]。它们提供了有关蛋白质结构和功能的宝贵信息，为科学研究和应用开发提供了基础和指导。

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] CHEUNG M S, CHAVEZ L L, ONUCHIC J N. The energy landscape for protein folding and possible connections to function[J]. Polymer, 2004, 45(2): 547-555.

[3] CARLSON H A. Protein flexibility is an important component of structure-based drug discovery[J]. Current Pharmaceutical Design, 2002, 8(17): 1571-1578.

[4] 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, 2016, 84(Suppl 1): 4-14.

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

[6] ZHOU X, ZHENG W, LI Y, et al. I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction[J]. Nature Protocols, 2022, 17(10): 2326-2353.

[7] SCHWEDE T, KOPP J, GUEXN, et al. SWISS-MODEL: an automated protein homology-modeling server[J]. Nucleic Acids Research, 2003, 31(13): 3381-3385.

[8] SONG Y, DIMAIO F, WANG R Y, et al. High-resolution comparative modeling with RosettaCM[J]. Structure(London, England: 1993), 2013, 21(10): 1735-1742.

[9] WEBB B, SALI A. Comparative Protein Structure Modeling Using MODELLER[J]. Current Protocols in Bioinformatics, 2016, 54: 5.6.1-5.6.37.

[10] XIA Y H, PENG C X, ZHOUX G, et al. A sequential niche multimodal conformational sampling algorithm for protein structure prediction[J]. Bioinformatics(Oxford, England), 2021, 37(23): 4357-4365.

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

[12] 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(Oxford, England), 2021, 37(23): 4350-4356.

[13] FENG Q, HOU M, LIU J, et al. Construct a variable-length fragment library for de novo protein structure prediction[J]. Briefings in Bioinformatics, 2022, 23(3): bbac086.

[14] XIE T Y, ZHOU X G, HU J, et al. Contact Map-based Residue-pair Distances Restrained Protein Structure Prediction Algorithm[J]. Computer Science, 2020, 47(1): 59-65.

[15] ABRIATA L A, TAMÒ G E, DAL PERARO M. A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments[J]. Proteins: Structure, Function, Bioinformatics, 2019, 87(12): 1100-1112.

[16] JUMPER J, EVANS R, PRITZEL A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[17] BAEK M, DIMAIO F, ANISHCHENKO I, et al. Accurate prediction of protein structures and interactions using a three-track

neural network[J]. Science,2021,373(6557):871-876.
[18] SU H,WANG W,DU Z,et al. Improved Protein Structure Prediction Using a New Multi-Scale Network and Homologous Templates[J]. Advanced Science(Weinheim,Baden-Wurttemberg,Germany),2021,8(24):e2102592.
[19] JONES D T,THORNTON J M. The impact of AlphaFold2 one year on[J]. Nature methods,2022,19(1):15-20.
[20] ALTSCHUL S F,MADDEN T L,SCHÄFFERA A,et al. Gapped BLAST and PSI-BLAST:a new generation of protein database search programs[J]. Nucleic Acids Research,1997,25(17):3389-3402.
[21] ALTSCHUL S F,KOONIN E V. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases[J]. Trends in Biochemical Sciences,1998,23(11):444-447.
[22] SÖDING J. Protein homology detection by HMM-HMM comparison[J]. Bioinformatics(Oxford,England),2005,21(7):951-960.
[23] WU S,ZHANG Y. MUSTER:Improving protein sequence profile-profile alignments by using multiple sources of structure information[J]. Proteins,2008,72(2):547-556.
[24] YANG Y,FARAGGI E,ZHAO H,et al. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates[J]. Bioinformatics(Oxford,England),2011,27(15):2076-2082.
[25] BUCHAN D W A,JONES D T. EigenTHREADER:analogous protein fold recognition by efficient contact map threading[J]. Bioinformatics(Oxford,England),2017,33(17):2684-2690.
[26] ZHENG W,WUYUN Q,LI Y,et al. Detecting distant-homology protein structures by aligning deep neural-network based contact maps[J]. PLoS Computational Biology,2019,15(10):e1007411.
[27] WU S,ZHANG Y. LOMETS:a local meta-threading-server for protein structure prediction[J]. Nucleic Acids Research,2007,35(10):3375-3382.
[28] ZHENG W,ZHANG C,WU Y ,et al. LOMETS2:improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins[J]. Nucleic Acids Research,2019,47(W1):W429-W436.
[29] ZHENG W,QI Q G,WU Y ,et al. LOMETS3:integrating deep learning and profile alignment for advanced protein template recognition and function annotation[J]. Nucleic Acids Research,2022,50(W1):W454-W464.
[30] REMMERT M,BIEGERT A,HAUSERA,et al. HHblits:lightning-fast iterative protein sequence searching by HMM-HMM alignment[J]. Nature Methods,2012,9(2):173-175.
[31] MIRDITA M,VON DEN DRIESCH L,GALIEZ C,et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments[J]. Nucleic Acids Research,2017,45(D1):D170-D176.
[32] ZHAO K,XIA Y,ZHANG F,et al. Protein structure and folding pathway prediction based on remote homologs recognition using PAthreader[J]. Communications Biology,2023,6(1):243.
[33] THORNTON J M,LASKOWSKI R A,BORKAKOTIN. AlphaFold heralds a data-driven revolution in biology and medicine[J]. Nature Medicine,2021,27(10):1666-1669.
[34] TUNYASUVUNAKOOL K,ADLER J,WU Z,et al. Highly accurate protein structure prediction for the human proteome[J]. Nature,2021,596(7873):590-596.
[35] FU L,NIU B,ZHU Z,et al. CD-HIT:accelerated for clustering the next-generation sequencing data[J]. Bioinformatics(Oxford,England),2012,28(23):3150-3152.
[36] FOX N K,BRENNER S E,CHANDONIA J-M. SCOPe:Structural Classification of Proteins—extended,integrating SCOP and ASTRAL data and classification of new structures[J]. Nucleic Acids Research,2014,42(D1):D304-D309.
[37] LI Z W,X L Q,ZHOU X G,et al. Multimodal Optimization Algorithm for Protein Conformation Space[J]. Computer Science,2020,47(7):161-165.
[38] SEEMAYER S,GRUBER M,SÖDING J. CCMpred—fast and precise prediction of protein residue-residue contacts from correlated mutations[J]. Bioinformatics(Oxford,England),2014,30(21):3128-3130.
[39] DU Z,PAN S,WUQ,et al. CATHER:a novel threading algorithm with predicted contacts[J]. Bioinformatics,2020,36(7):2119-2125.
[40] SKOLNICK J,GAO M,ZHOU H,et al. AlphaFold 2:why it works and its implications for understanding the relationships of protein sequence,structure,and function[J]. Journal of Chemical Information,2021,61(10):4827-4831.
[41] CONNELL K B,MILLER E J,MARQUSEE S. The folding trajectory of RNase H is dominated by its topology and not local stability:a protein engineering study of variants that fold via two-state and three-state mechanisms[J]. Journal of Molecular Biology,2009,391(2):450-460.

LIANG Fang,born in 1999,research assistant. Her main research interests include intelligent information processing,optimization theory and algorithm design and bioinformatics.

ZHANG Guijun,born in 1974,Ph. D,professor,Ph.D supervisor,is a member of CCF(No. 50785G). His main research interests include intelligent information processing,optimization theory and algorithm design and bioinformatics.