基于序列的蛋白质二面角预测研究综述

郑美丽[1],朱 琪[1],张步忠[1,2]
(1. 安庆师范大学 计算机与信息学院,安徽 安庆 246013;
2. 苏州大学 江苏省计算机信息处理技术重点实验室,江苏 苏州 215006)

摘 要:蛋白质结构决定功能,蛋白质主链上 N-C 键和 C-C 键形成的二面角(φ,ψ)对蛋白质三维结构和空间构象起着重要作用。从蛋白质一级序列出发,预测骨架二面角可以加速对低能结构构象空间的有效采样,大大推进三维结构预测,可作为生物实验的有效快速辅助手段。随着蛋白质生物样本数据增多和计算性能提升,近年来,深度学习方法广泛应用到蛋白质二面角预测。介绍了蛋白质残基的主要特征表示、计算方法对二面角预测处理、评价标准和常用数据集等;对近几年来的基于深度学习模型诸多研究工作进行系统归纳与整理,从网络结构设计、输入特征表示、模型泛化性能等方面进行总结,并对比分析各算法特点及存在的问题。在此基础上,对其未来研究发展方向与应用前景进行了展望。

关键词:骨架二面角;机器学习;深度学习;序列表征;蛋白质结构
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2023)03-0001-08
doi:10.3969/j.issn.1673-629X.2023.03.001

Review of Sequence-based Protein Dihedral Angle Prediction Research

ZHENG Mei-li[1], ZHU Qi[1], ZHANG Bu-zhong[1,2]
(1. School of Computer and Information, Anqing Normal University, Anqing 246013, China;
2. Jiangsu Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China)

Abstract: Protein structure determines function, and the dihedral angles formed by N-C and C-C bonds in the protein backbone play an important role in protein three-dimensional structure and spatial conformation. Starting from the protein primary sequence, prediction of backbone dihedral angle can accelerate the effective sampling of low-energy structural conformational space and greatly advance the 3D structure prediction, which can be used as an effective and rapid aid for biological experiments. With more protein biological samples and improved computational performance, deep learning methods have been widely applied to protein dihedral angle prediction in recent years. To deeply understand this work, what are detailed introduced are feature representations of protein residues, computational methods for dihedral angle processing, evaluation metrics and common datasets. And the recent research progress based on deep learning is rigorously reviewed in terms of network structure design, input feature representation, model generalization performance, etc. The effectiveness and shortcoming of each algorithm are also compared and analyzed. Upon above analysis, the future research field and application prospect are presented.

Key words: backbone dihedral angle; machine learning; deep learning; sequence characterization; protein structure

0 引 言

蛋白质是由氨基酸缩水链接成的一种有机复合物,一个氨基酸残基的基本构成有中心 Cα 原子、氨基(-NH_2)、羧基(-COOH)、氢键(-H)和侧链 R 基团。蛋白质三维结构中,骨架主链上的二面角很大程度上反映了三维构象。一个氨基酸残基通常对应两个二面角(首尾残基除外)[1],φ(phi)和 ψ(psi),范围在-180°至 180°之间。围绕 N-Ca 的 C-N-Ca-C 原子构成 φ 二面角,围绕 Ca-C 键的 N-Ca-C-N 构成 ψ 二面角,实例如图 1 所示。蛋白质主链二面角(φ,ψ)是蛋白质结构的一部分,研究蛋白质二面角对于蛋白质功能的研究具有重要的意义。诺贝尔奖获得者

收稿日期:2022-06-05 **修回日期**:2022-10-09
基金项目:安徽省高校优秀人才支持计划项目(gxyq2020029);教育部中国高校产学研创新基金-新一代信息技术创新项目(2019ITA01046);安庆师范大学 2021 年度研究生学术创新项目(2021yjsXSCX014)
作者简介:郑美丽(1998-),女,硕士研究生,研究方向为生物信息学;通讯作者:张步忠(1980-),男,博士,教授,CCF 会员(33714M),研究方向为生物信息学。

Anfinsen[2]的实验表明,蛋白质的结构信息包括二面角蕴含于其序列之中,从而表明从序列出发进行蛋白质二面角预测是可行的。

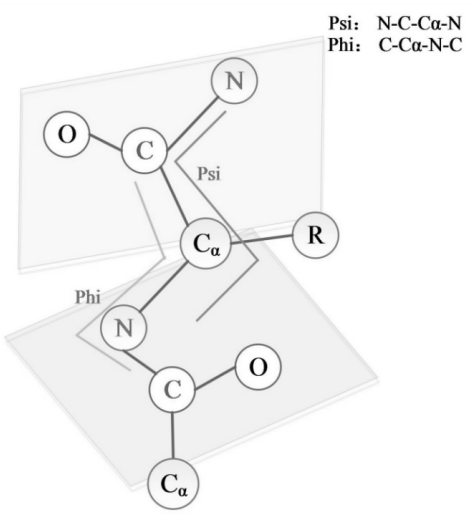

图1 蛋白质二面角

骨架二面角作为一种重要的结构约束,它对主链构象变化的影响远大于键长和键角,在蛋白质结构预测的空间进行采样以研究蛋白质折叠和细化中起着关键作用,准确预测骨架二面角可以加速对低能结构构象空间的有效采样,大大推进三级结构预测。

1 特征表示和数据集

1.1 序列表示

蛋白质序列残基的表示,主要有位置特异性评分矩阵[3](PSSM)、隐马尔可夫模型打分矩阵[4](HMM)、物理化学性质[5](PP)、蛋白质二级结构[6](SS)、溶剂可及性[7](SA)、序列编码[8](SC)等。

1.1.1 位置特异性评分矩阵

序列进化信息对揭示蛋白质结构和功能非常重要。多序列比对方法 PSI-BLAST[9]产生的位置特异性评分矩阵能揭示序列进化信息,被广泛应用在蛋白质相关的生物信息学中。如式(1)所示,PSSM 是形如 $L\times20$ 的矩阵,其中 L 是蛋白质序列长度,行是序列中残基,列是现有20种氨基酸。$P(i\rightarrow j)$ 表示序列中第 i 个残基突变为第 j 种氨基酸残基的概率。对于将 PSSM 数据作为序列表示,还需进行归一化处理。

$$\mathrm{PSSM}=\begin{bmatrix}P_{1\rightarrow1}P_{1\rightarrow2}\cdots P_{1\rightarrow20}\\P_{2\rightarrow1}P_{2\rightarrow2}\cdots P_{2\rightarrow20}\\\cdots\\P_{L\rightarrow1}P_{L\rightarrow2}\cdots P_{L\rightarrow20}\end{bmatrix}\tag{1}$$

1.1.2 隐马尔可夫模型评分矩阵

在多序列比对中,HH-suite[10]套件中的 HHblits 基于其专用格式的多序列数据库,通过聚类 UniProt 或者 NR 库,将序列长度对齐性80%以上、相似度20%以上的序列聚集,并生成对应的 HMM 特征文件。通过 HHsearch/HHblits 生成的 HMM 格式中,HMM 数据部分表示的是该位置残基向对应残基变异的发生概率,用-1 000 * lb(frequency)表示成正整数,"*"表示零。随后的一行是10种转移概率。在作为数据特征表示时,可取前30列。

1.1.3 氨基酸的物理化学特性

蛋白质其物理化学性质一部分与氨基酸相似,一部分在特定环境下具有特定的性质。在蛋白质结构特性预测中,常用的氨基酸理化性质有:空间参数、极化率、体积、疏水性、等电点、螺旋概率和片概率,具体取值参见文献[5]。

1.1.4 序列编码

蛋白质一级序列是字母编码,对一级序列常用0-1编码表示,多数用21或22维向量的正交编码。由于该编码形式只有一个非零向量,不利于梯度优化类算法值更新,Zhang[11]采用自编码器方式将0-1稀疏向量映射到稠密向量,计算方法如式(2),用 h 表示新的编码。

$$\begin{aligned}h&=\sigma(W_1x+b_1)\\\hat{x}&=\sigma(W_2h+b_2)\end{aligned}\tag{2}$$

1.1.5 其 它

蛋白质结构决定功能,描述其空间特性的二级结构、溶剂可及表面积、残基接触图[12](Contact Map,CM)、多序列比对信息(MMseqs2)等也应用到了二面角预测中。OPUS-TASS 和 OPUS-TASS2 方法还使用独特的 PSP[13](Potential Based On Side Chain Packing)特征。为方便对比,表1列出了近年来的典型预测算法的特征表示。

1.2 输 出

蛋白质二面角预测值表示主要有:数值(已归一化),二面角的正弦和余弦函数值。SPINE X、Real-SPINE、Real-SPINE 2.0、Real-SPINE 3.0、DANGLE 等方法输出都是将二面角进行归一化。SPIDER 2、SPIDER 3、SPIDER3-Single、DeepRIN、RaptorX-Angle、SPOT-1D、ProteinUnet、CRRNN2、SPOT-1D-single 等通过输出二面角的正弦、余弦函数来消除角度周期性,再通过公式 $\alpha=\tan^{-1}[\sin\alpha/\cos\alpha]$ 还原二面角。DESTRUCT、ANGLOR 则是直接输出二面角的角度。

直接预测二面角角度难度较大,多数模型没有采用该方式;直接以归一化的形式输出,则无法兼顾到角度的周期性,如360°和0°归一化值为1和0。以三角函数值的形式输出,则多数没有考虑到 $\sin^2\hat{y}+\cos^2\hat{y}=1$ 的约束特性。

表 1　预测方法的输入特征

方法	PSSM	SS	PSFM	HMM	SA	PP	CM	PSP	SC	MM seqs2
DESTRUCT[14]	√	√								
ANGLOR[15]	√	√			√					
SPINE X[16]	√	√			√	√				
TANGLE[17]	√	√			√					
Real-SPINE[18]	√					√				
Real-SPINE 2.0[19]										
Real-SPINE 3.0[20]	√	√				√				
SPIDER 2[21]	√					√				
SPIDER 3[22]	√			√		√				
SPIDER 3-Single[23]									√	
RaptorX-Angle[24]	√	√	√		√					
DeepRIN[25]	√	√		√		√				
CRRNN2[26]	√			√					√	
SPOT-1D[27]	√			√		√	√			
NetSurfP-2.0[28]				√		√			√	√
ProteinUnet[29]	√			√		√			√	
SPOT-1D-Single[30]										
OPUS-TASS[31]	√			√		√		√		
OPUS-TASS2[32]	√			√		√		√		

1.3　评价指标

二面角预测评价标准有皮尔逊相关系数(PCC)、平均绝对误差(MAE)、均方根误差(RMSE)等。在进行蛋白质二面角预测评价时,对预测值 P' 和真实值 E 之间的差值通常先按式(3)将二面角进行角度变换,其中 P' 是预测二面角的原始值。PCC、MAE、RMSE 分别通过式(4)~(6)计算。

$$P=\begin{cases}P', \text{if}\ |P'-E|\leqslant 180^\circ\\ P'+360^\circ, \text{if}\ P'-E<-180^\circ\\ P'-360^\circ, \text{if}\ P'-E>180^\circ\end{cases} \tag{3}$$

$$\text{PCC}=\frac{1}{N-1}\sum_{i=1}^{n}\left(\frac{P-\bar{P}}{S_p}\right)\left(\frac{E-\bar{E}}{S_E}\right) \tag{4}$$

其中,$\bar{P}$、$\bar{E}$ 是 P、E 的均值,S_p、S_E 是 P、E 的标准差。

$$\text{MAE}=\frac{1}{N}\sum_{i=1}^{N}|E-P| \tag{5}$$

$$\text{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(E-P)^2} \tag{6}$$

1.4　数据集

1.4.1　训练集

训练集多数来自 PISCES CullPDB[33] 挑选的数据集,序列之间的相似度一般低于30%。也常从 PDB 蛋白质数据库中直接抽取序列并筛选用作训练集。

1.4.2　测试集

对模型泛化性能进行测试用的数据集,多数来自一些公开数据集,如 CB513、CASP(10-14)数据集等。另一种是算法提出者给出的测试集合,如 TEST2016、TEST2018[27]、CASP-FM56[34]等。表 2 和表 3 分别给出了近年来的预测方法对 (φ,ψ) 二面角的预测性能,测试数据用 CASP 公共数据集。“FM”表示数据集中是无模板序列,预测难度更大。对比指标用 PCC 和 MAE,“-”表示指标缺失。

表 2　φ 角在各测试集的评价指标

	casp10	casp11	casp12	CASP-FM(56)	CASP12(FM)	CASP13(FM)	CASP14(FM)
方法	MAE (PCC)	MAE (PCC)	MAE (PCC)	MAE	MAE	MAE	MAE
ANGLOR	-	25.69(-)	25.79(-)	-	-	-	-
SPINE-X	-	20.7(-)	24.85(-)	-	-	-	-
Spider2	25.83 (0.766)	26.06 (0.756)	26.85 (0.758)	-	-	-	-

续表2

	casp10	casp11	casp12	CASP-FM(56)	CASP12(FM)	CASP13(FM)	CASP14(FM)
Spider3	20.41 (0.831)	21.38 (0.815)	20.59 (0.819)	–	–	–	–
RaptorX-Angle	21.19 (0.818)	20 (0.798)	20.69 (0.788)	–	25.03	–	–
DeepRIN	17.39 (0.867)	18.97 (0.849)	20.21 (0.828)	–	–	–	–
NetSurfP-2.0	18.03 (0.862)	19.63 (0.839)	20.44 (0.828)	20.53	22.47	22.689 5	22.62
CRRNN2	16.56 (0.879)	18.49 (0.851)	19.14 (0.842)	–	–	–	–
SPOT-1D	–	–	18.9(–)	19.39	21.844	20.238	23.19
OPUS-TASS	–	–	18.08(–)	18.85	21.9	–	21.91
SPIDER3-Single	–	–	–	–	26.168	25.315	–
OPUS-TASS2-76D	–	–	–	18.58	–	–	21.79
OPUS-TASS2	–	–	–	17.94	–	–	20.53
ProteinUnet	–	–	–	–	25.942	25.036	–
SPOT-1D-Single	–	–	–	–	25.426	25.125	–

表3 ψ角在各测试集的评价指标

	casp10	casp11	casp12	CASP-FM(56)	CASP12(FM)	CASP13(FM)	CASP14(FM)
方法	MAE (PCC)	MAE (PCC)	MAE (PCC)	MAE	MAE	MAE	MAE
ANGLOR	–	46.17(–)	47.37(–)	–	–	–	–
SPINE-X	–	34.6(–)	46.57(–)	–	–	–	–
Spider2	49.18 (0.693)	48.13 (0.699)	50.59 (0.688)	–	–	–	–
Spi der3	31.17 (0.824)	33.05 (0.805)	35.67 (0.783)	–	–	–	–
RaptorX-Angle	34.85 (0.8)	30.14 (0.788)	31.6 (0.834)	–	44.62	–	–
DeepRIN	25.79 (0.865)	27.96 (0.855)	31.39 (0.834)	–	–	–	–
NetSurfP-2.0	26.16 (0.869)	28.68 (0.854)	30.88 (0.843)	31.94	35.892	34.38	40.54
CRRNN2	24.02 (0.878)	26.68 (0.864)	28.9 (0.853)	–	–	–	–
SPOT-1D	–	–	27.46(–)	30.1	33.962	31.114	43.98
OPUS-TASS	–	–	25.98(–)	28	33.63	–	38.93
SPIDER3-Single	–	–	–	–	47.462	46.164	–
OPUS-TASS2-76D	–	–	–	26.91	–	–	38.65
OPUS-TASS2	–	–	–	25.17	–	–	33.48
ProteinUnet	–	–	–	–	46.527	46.884	–
SPOT-1D-Single	–	–	–	–	43.457	44.022	–

ψ 角的预测难度高于 φ 角。从预测结果看,自 Spider 3 起的深度学习方法均取得了更优性能,一方面是训练数据更多,另外一方面参数更多的深度学习方法得到充分训练后,泛化性能更好。特别地,OPUS-TASS2 性能最好,其次分别是 OPUS-TASS 和 SPOT-1D。这三个方法都是参数规模大的集成模型,SPOT-1D 和 OPUS-TASS 混合使用长短期记忆网络[35](Long-Short Term Memory,LSTM)、卷积神经网络[36](Convolutional Neural Networks,CNN)和残差网络[37](Residual Neural Network,ResNet);OPUS-TASS2 进一步融合 Transformer[38]。SPOT-1D 和 OPUS-TASS2 都将 Contact map 作为输入。

2　二面角预测

二面角预测可归类为回归问题,随着蛋白质已知结构的数据集增多,机器学习方法已应用到该问题中。

2.1　传统机器学习方法

二面角预测引入到计算领域最初是作为辅助手段提升二级结构预测性能。2000 年 Bystroff[39] 提出用隐马尔可夫模型预测二面角构象,将 (φ,ψ) 映射到 10 个区域和 1 个顺式肽结构。2004 年,Kuang 等[40]用支持向量机和神经网络预测二面角构象,将 (φ,ψ) 映射到(A,B,G,E)四个区域。

2005 年 Wood 等[14]提出 DESTRUCT 方法,用级联反馈输入策略构建三个神经网络,预测二级结构和 ψ 二面角,但其关注点依然在二级结构。2008 年,Wu 和 Zhang[15]提出 ANGLOR 方法,用 PSSM 和计算软件预测的二级结构、溶剂可及性作为输入,用神经网络预测 φ 角、支持向量机预测 ψ 角。2009 年 Shen 等[41]提出 TALOS+模型,利用两级前馈神经网络预测 (φ,ψ) 二面角。

Zhou 课题组用神经网络方法提出了 Real-SPINE[18]、Real-SPINE 2.0[19]、Real-SPINE 3.0[20]模型。Real-SPINE 集成了两个神经网络并对各自输出取均值,用于预测溶剂可及性和 ψ 角。Real-SPINE 2.0 第一次用计算方法预测 φ 角和 (φ,ψ) 真实值,该方法同样使用神经网络模型,为了消除角度的周期性影响,将 (φ,ψ) 调整在-180°和 180°之间。Real-SPINE 3.0 预测 (φ,ψ) 二面角和溶剂可及表面积,模型两层隐藏层,每层 101 个节点,输入窗口 41 个残基。为克服梯度后传局限于小网络的问题,在梯度传递时附加监督因子(Guiding Factor) g_{jk}^{1} , g_{kl}^{2} , g_{lm}^{3} 分别对应两层隐藏层和输出层。

2012 年 Song 等[17]提出 TANGLE,用两阶段的支持向量回归(SVR)策略预测骨架二面角。TANGLE 不依赖已知结构信息,输入是 PSSM 和 PSIPred 软件预测的二级结构、Scratch 预测的溶剂可及性、Disopred2 预测的固有无序信息、序列长度以及序列权重,输入到第一阶段 SVR 后,预测结果再输入第二阶段的 SVR。Song 还综合验证了输入窗口尺寸和输入特征组合对性能的影响。TANGLE 完全从一级序列出发以及两阶段训练、逐步求精的策略,对后续研究工作有深刻的影响。

2010 年,Cheung 等[42]提出 DANGLE 模型,用贝叶斯生成模型预测二面角。利用残基键化学位移特性,将 (φ,ψ) 二面角映射到 36×36 的 10°方格拉氏空间(Ramachandran Space)图中。利用贝叶斯公式计算查询散布模式(Query Scatter Pattern,QSP)下的 (φ,ψ) 概率。

限于计算资源和已知蛋白质三维结构不多等因素,传统计算方法性能有待改进。但给本问题研究提供了参考意义。

2.2　深度学习方法

近年来,深度学习方法由于其良好的泛化性能,广泛应用在多个领域。2014 年,Lyons 等[43]用深度稀疏自编码器堆叠的模型 SPIDER 预测骨架 Cα 上的 θ 和 τ 二面角。2015 年 Heffernan 等[21]提出的 SPIDER2 可以预测蛋白质二级结构、二面角和溶剂可及表面积。SPIDER2 模型结构和 SPIDER 一样,在训练时进行三次迭代,上一次训练结果作为下一次迭代训练的输入。实验表明迭代式训练能有效提升模型泛化性能。

2017 年,Li 等[44]将受限的玻尔兹曼机(RBM)与深度神经网络结合,设计出了深度递归 RBM(DReRBM)模型。在受限的玻尔兹曼机的基础上,将上一次的输出 h_{t-1} 作为本次输入,充分拟合了蛋白质序列上下游环境。DReRBM 模型由输入层、隐藏层和输出层组成,多个 RBM 堆栈在其中,以一种逐步的方式训练,一个训练过的 RBM 的隐藏数据作为可见的输入数据馈送给下一个 RBM,模型梯度计算通过吉布斯采样完成。

2017 年,Heffernan 等[22]提出 SPIDER3 模型预测蛋白质二级结构、溶剂可及性、接触图和二面角。SPIDER3 由两层双向长短期记忆网络[45](BLSTM)构建。SPIDER3 和 SPIDER2 一样,采用迭代训练的策略,上一轮输出作为下一轮输入,共迭代了 4 次。SPIDER3 分别训练了回归和分类模型。BLSTM 能记忆前向和后向两个方向的时序输入信息,较好地拟合了蛋白质序列残基和左右上下文相关的特性,能够学习距离较远和距离较近的序列内的依赖关系。

卷积神经网络[36]结合残差网络[37],通过网络层加深提升长范围特征感知。2018 年,Gao 等[24]提出了 RaptorX-Angle 模型。RaptorX-Angle 中堆叠了多个

残差块。结合K-means算法,首先,从训练数据中生成一组(φ,ψ)的聚类,从中可以得到每个聚类的分布;然后,利用深度学习方法对离散标签进行预测;最后,通过混合经验聚类及其预测概率来预测实际二面角值。

另一典型CNN结构模型是Fang等[25]在2018年提出的DeepRIN。DeepRIN结合Inception ResNet,构建残差Inception块,并堆叠两层残差Inception块。DeepRIN用小窗口卷积来提高网络的计算效率。DeepRIN使用9 000条训练数据,并将每条序列长度对齐到700,输入不再使用窗口形式。

鉴于RNN在长范围特征获取的优势、CNN局部特性获取和ResNet便捷残差传递的特点,多数模型结合三者用于预测二面角。2019年Klausen等[28]提出了NetSurfP-2.0模型,预测残基的溶剂可及性、二级结构、无序蛋白和骨架二面角。NetSurfP-2.0输入是HMM特征和序列编码,分别经过32个卷积核129、257的CNN后,输入两层BLSTM网络,BLSTM单向1 024个单元。NetSurfP-2.0使用了10 337条训练序列,其参数规模也达到了3 400万,过多网络参数给训练和预测带来了不便。

2019年,Kim等[46]提出使用生成对抗网络(GAN)进行二面角预测,训练了GAN的鉴别器来估计密度,但模型的显式密度不易处理。因此,引入噪声对比估计(Noise-Contrastive Estimation,NCE)来估计非归一化统计模型的归一化常数,即引入了噪声对比估计生成对抗网络(NCE-GAN),通过从已知分布(如噪声对比估计)中输入噪声样本,并为鉴别器添加相应的类,从而实现生成对抗网络的显式密度估计。

2019年,Hanson等[27]提出SPOT-1D模型,预测蛋白质二级结构、二面角、溶剂可及性和残基接触数(Contact Number)。SPOT-1D的输入在PSSM、HMM特征和理化性质的基础上,将预测的接触图作为输入改进模型泛化性能。SPOT-1D利用BLSTM和ResNet混合模型的集成来识别和传播整个序列的短期和长期依赖,SPOT-1D由9个网络结构的模型集成。SPOT-1D训练集包含10 029条序列,分别训练了分类和回归两类模型,总模型文件大小10 GB左右。SPOT-1D在多个任务上均取得较好性能,但其庞大的模型不利于在生物领域应用开展。

2020年,Xu等[31]提出的OPUS-TASS性能比SPOT-1D更好。OPUS-TASS输入为PSSM、HMM、理化性质和19位PSP,分别送到5层CNN网络、2层Transformer网络(编码部分),两部分合并得到228维的数据再送给4层BLSTM网络。OPUS-TASS分别集成7个模型用于分类和回归预测。OPUS-TASS模型文件3.7 GB,比SPOT-1D要小,但依然对实际应用的资源要求较高。

2018年,Heffernan等提出了仅使用序列信息的SPIDER3-Single[23]模型,SPIDER3-Single的网络结构和训练方法与SPIDER3类似,不同的是仅用了20维序列编码作为输入。但SPIDER3-Single模型泛化性能和SPIDER3相比还是有较大差距。2021年Kotowski等[29]提出ProteinUnet模型,ProteinUnet输入和SPIDER3-Single一样,但大幅度提升了预测性能。

2021年,Singh等[30]同样提出了面向单序列输入的SPOT-1D-Single模型。SPOT-1D-Single集成了三个不同的结构模型,也分别面向分类和回归训练。SPOT-1D-Single使用39 120条训练序列。SPOT-1D-Single泛化性能不如SPOT-1D,但已超越ProteinUnet,并已有了一定的实用价值。

3 展 望

通过上述分析,计算方法特别是深度学习预测二面角,取得了较好进展,但预测性能依然有提升空间,针对相关研究,可以从以下几个方面进行思考:

(1)单序列输入更方便生物学人员使用。现有模型多依赖PSSM、HMM等多序列比对信息,对非专业人员要求高。仅有序列编码信息的单序列模型,对生物学人员更友好,而单序列输入模型性能还有待提高。

(2)需要设计对计算资源依赖更少的模型。在泛化性能一致时,轻量级模型更方便用户使用,推理时对计算资源依赖更少。

(3)可将多个问题联合解决。二面角、二级结构、接触图等蛋白质结构层面问题,相互依赖。如二级结构预测性能提升,同样能推动二面角预测性能。将预测得到的接触图等信息作为模型输入,同样也能提升二面角、二级结构预测性能。

(4)训练样本依然偏少。截至到2022年4月,PDB数据库中通过生物实验手段解析的已知蛋白质结构10.66万条。这些数据无法支持类似BERT大规模模型训练,需要设计对序列特性捕获更好的深度学习模型。

4 结束语

蛋白质骨架二面角是蛋白质结构的重要属性,高精度地预测蛋白质骨架二面角以加速对三维结构构象空间的有效采样,对蛋白质三级结构预测具有重要意义。该文对蛋白质骨架二面角预测算法的发展和领域内最新研究进行了综述,从序列表征、输出、数据集、结构框架等方面介绍算法。同时,对当前二面角预测存在的问题进行了思考。

参考文献:

[1] 龙施洋. 蛋白质主链二面角相关性研究[D]. 长春:吉林大学,2017.

[2] 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 of the United States of America,1961,47(9):1309-1314.

[3] ALTSCHUL S F,MADDEN T L,SCHÄFFER A 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.

[4] SODING J. Protein homology detection by HMM-HMM comparison[J]. Bioinformatics,2005,21(7):951-960.

[5] MEILER J,MULLER M,ZEIDLER A,et al. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks[J]. Molecular Modeling Annual,2001,7(9):360-369.

[6] KABSCH W,SANDER C. Dictionary of protein secondary structure: pattern recognition of hydrogen - bonded and geometrical features[J]. Biopolymers:Original Research on Biomolecules,1983,22(12):2577-2637.

[7] LEE B,RICHARDS F M. The interpretation of protein structures: estimation of static accessibility[J]. Journal of Molecular Biology,1971,55(3):379-400.

[8] ZHOU J,TROYANSKAYA O G. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction[C]//31st international conference on machine learning. Beijing: International Machine Learning Society,2014:1121-1129.

[9] ALTSCHUL S F,GERTZ E M,AGARWALA R,et al. PSI-BLAST pseudocounts and the minimum description length principle[J]. Nucleic Acids Research,2009,37(3):815-824.

[10] JOHANNES S. Protein homology detection by HMM - HMM comparison[J]. Bioinformatics,2005(7):951-960.

[11] ZHANG Z,XIAO J,WU J,et al. ParaAT:a parallel tool for constructing multiple protein-coding DNA alignments[J]. Biochemical and Biophysical Research Communications,2012,419(4):779-781.

[12] HANSON J,PALIWAL K,LITFIN T,et al. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks[J]. Bioinformatics,2018,34(23):4039-4045.

[13] LU M,DOUSIS A D,MA J. OPUS-PSP:an orientation-dependent statistical all-atom potential derived from side-chain packing[J]. Journal of Molecular Biology,2008,376(1):288-301.

[14] WOOD M J,HIRST J D. Protein secondary structure prediction with dihedral angles[J]. Proteins,2005,59(3):476-481.

[15] WU S,ZHANG Y. Anglor:a composite machine-learning algorithm for protein backbone torsion angle prediction[J]. PLoS One,2008,3(10):e34000.

[16] FARAGGI E,ZHANG T,YANG Y,et al. SPINE X:improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles[J]. Journal of Computational Chemistry,2012,33(3):259-267.

[17] SONG J,TAN H,WANG M,et al. TANGLE:two-level support vector regression approach for protein backbone torsion angle prediction from primary sequences[J]. PLoS One,2012,7(2):e30361.

[18] DOR O,ZHOU Y. Real-SPINE:an integrated system of neural networks for real-value prediction of protein structural properties[J]. Proteins,2007,68(1):76-81.

[19] XUE B,DOR O,FARAGGI E,et al. Real-value prediction of backbone torsion angles[J]. Proteins,2008,72(1):427-433.

[20] FARAGGI E,XUE B,ZHOU Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network[J]. Proteins,2009,74(4):847-856.

[21] HEFFERNAN R,PALIWAL K,LYONS J,et al. Improving prediction of secondary structure, local backbone angles,and solvent accessible surface area of proteins by iterative deep learning[J]. Scientific Reports,2015,5(1):1-11.

[22] HEFFERNAN R,YANG Y,PALIWAL K,et al. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility[J]. Bioinformatics,2017,33(18):2842-2849.

[23] HEFFERNAN R,PALIWAL K,LYONS J,et al. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning[J]. Journal of Computational Chemistry,2018,39(26):2210-2216.

[24] GAO Y,WANG S,DENG M,et al. RaptorX-angle:real-value prediction of protein backbone dihedral angles through a hybrid method of clustering and deep learning[J]. BMC Bioinformatics,2018,19(4):73-84.

[25] FANG C,SHANG Y,XU D. Prediction of protein backbone torsion angles using deep residual inception neural networks[J]. IEEE/ACM Transactions on Computational Biology And Bioinformatics,2018,16(3):1020-1028.

[26] ZHANG B,LI J,QUAN L,et al. Multi-task deep learning for concurrent prediction of protein structural properties[J/OL]. (2021-02-04). https://www.biorxiv.org/content/10.1101/2021.02.04.429840v1.full.pdf.

[27] HANSON J, PALIWAL K, LITFIN T, et al. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks[J]. Bioinformatics, 2019, 35(14): 2403-2410.

[28] KLAUSEN M S, JESPERSEN M C, NIELSEN H, et al. NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning[J]. Proteins, 2019, 87(6): 520-527.

[29] KOTOWSKI K, SMOLARCZYK T, ROTERMAN-KONIECZNA I, et al. ProteinUnet-an efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures[J]. Journal of Computational Chemistry, 2021, 42(1): 50-59.

[30] SINGH J, LITFIN T, PALIWAL K, et al. SPOT-1D-Single: improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning[J]. Bioinformatics, 2021, 37(20): 3464-3472.

[31] XU G, WANG Q, MA J. OPUS-TASS: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks[J]. Bioinformatics, 2020, 36(20): 5021-5026.

[32] XU G, WANG Q, MA J. OPUS-X: an open-source toolkit for protein torsion angles, secondary structure, solvent accessibility, contact map predictions and 3D folding[J]. Bioinformatics, 2022, 38(1): 108-114.

[33] WANG G, JR R L D. PISCES: a protein sequence culling server[J]. Bioinformatics, 2003, 19(12): 1589-1591.

[34] UDDIN M R, MAHBUB S, RAHMAN M S, et al. SAINT: self-attention augmented inception-inside-inception network improves protein secondary structure prediction[J]. Bioinformatics, 2020, 36(17): 4599-4608.

[35] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[36] GU J, WANG Z, KUEN J, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354-377.

[37] TARG S, ALMEIDA D, LYMAN K. Resnet in resnet: generalizing residual architectures[J/OL]. (2016-03-25). https://arxiv.org/pdf/1603.08029.pdf.

[38] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[39] BYSTROFF C, THORSSON V, BAKER D. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins[J]. Journal of Molecular Biology, 2000, 301(1): 173-190.

[40] KUANG R, LESLIE C S, YANG A S. Protein backbone angle prediction with machine learning approaches[J]. Bioinformatics, 2004, 20(10): 1612-1621.

[41] SHEN Y, DELAGLIO F, CORNILESCU G, et al. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts[J]. Journal of Biomolecular NMR, 2009, 44(4): 213-223.

[42] CHEUNG M S, MAGUIRE M L, STEVENS T J, et al. DANGLE: a Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure[J]. Journal of Magnetic Resonance, 2010, 202(2): 223-233.

[43] LYONS J, DEHZANGI A, HEFFERNAN R, et al. Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network[J]. Journal of Computational Chemistry, 2014, 35(28): 2040-2046.

[44] LI H, HOU J, ADHIKARI B, et al. Deep learning methods for protein torsion angle prediction[J]. BMC Bioinformatics, 2017, 18(1): 1-13.

[45] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[46] KIM H. Dihedral angle prediction using generative adversarial networks[J/OL]. (2018-03-29). https://arxiv.org/pdf/1803.10996.pdf.