计算机科学
COMPUTER SCIENCE

河海图结构蛋白质数据集及预测模型

魏想想, 孟朝晖

引用本文

魏想想, 孟朝晖. 河海图结构蛋白质数据集及预测模型[J]. 计算机科学, 2024, 51(8): 117-123.

WEI Xiangxiang, MENG Zhaohui. Hohai Graphic Protein Data Bank and Prediction Model[J]. Computer Science, 2024, 51(8): 117-123.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100014

河海图结构蛋白质数据集及预测模型

魏想想　孟朝晖
河海大学计算机与软件学院　南京 211106
(221307030006@hhu.edu.cn)

摘　要　蛋白质是一种具有空间结构的物质。蛋白质结构预测的主要目标是从已有的大规模的蛋白质数据集中提取有效的信息,从而预测自然界中蛋白质的结构。目前蛋白质结构预测实验存在的一个问题是,缺少能够进一步反映出蛋白质空间结构特征的数据集。当前主流的 PDB 蛋白质数据集虽然是经过实验测得,但没有利用到蛋白质的空间特征,而且存在掺杂核酸数据和部分数据不完整的问题。针对以上问题,从蛋白质的空间结构角度来研究蛋白质的预测。在原始 PDB 数据集的基础上,提出了河海图结构蛋白质数据集(Hohai Graphic Protein Data Bank,HohaiGPDB)。该数据集以图结构为基础,表达出了蛋白质的空间结构特征。基于传统 Transformer 网络模型对新的数据集进行了相关的蛋白质结构预测实验,在 HohaiGPDB 数据集上的预测准确率可以达到 59.38%,证明了 HohaiGPDB 数据集的研究价值。HohaiGPDB 数据集可以作为蛋白质相关研究的通用数据集。

关键词:河海图结构蛋白质数据集;蛋白质空间结构;蛋白质结构预测;Transformer 模型
中图分类号　TP391

Hohai Graphic Protein Data Bank and Prediction Model

WEI Xiangxiang and MENG Zhaohui
School of Computer and Software,Hohai University,Nanjing 211106,China

Abstract　Protein is a kind of substance with spatial structure. The main goal of protein structure prediction is to extract effective information from existing large-scale protein datasets,so as to predict the structure of proteins in nature. At present,one of the problems in protein structure prediction experiments is the lack of data sets that can further reflect the spatial structure of proteins. Although the current mainstream PDB(protein data bank) is experimentally measured,it does not utilize the spatial characteristics of proteins,and there are problems of doping nucleic acid data and partial data is incomplete. In view of the above problems,this paper studies the prediction of protein from the perspective of spatial structure. Based on the original PDB,the Hohai graphic protein data bank is proposed. The dataset expresses the spatial structure characteristics of proteins based on the graph structure. Based on the traditional Transformer network model,relevant protein structure prediction experiments are carried out on the new dataset,and the prediction accuracy of HohaiGPDB could reach 59.38%,which proves the research value of HohaiGPDB. The HohaiGPDB could be used as a general data set for protein-related studies.

Keywords　Hohai graphic protein data bank,Protein spatial structure,Protein structure prediction,Transformer model

1　引言

蛋白质在自然界中是一种空间的动态的物质。了解蛋白质的结构和功能生命的本质和过程,利用科学手段对蛋白质的结构和功能进行预测和分析,有助于深入了解生物体的生理机制。随着生物技术的不断发展,大量的蛋白质数据不断涌现,为蛋白质相关研究提供了宝贵的资源。然而,如何有效地分析和利用这些数据,仍然是蛋白质相关领域的一个挑战。

蛋白质数据库(PDB)[1]产生于 1971 年,是一种全球共享的结构生物学数据库。数据集中包含了来自不同物种的蛋白质和核酸的结构信息,这些结构信息是通过实验测定的。UniProt 数据库[2]主要包含的是氨基酸序列的数据,截至 2023 年 6 月,已有超过 2.4 亿个氨基酸序列被存入数据库,而只有大约 20 万个实验确定的蛋白质结构被存入蛋白质数据库(PDB)[3]。为了发掘未知蛋白质的潜在价值,预测其三维结构就显得至关重要。DeepMind 的端到端模型 AlphaFold2,已被证明了具有预测许多未知蛋白质的三维结构的能力。DeepMind 和 EMBL's European Bioinformatics Institute(EMBL-EBI)构建了 AlphaFold 蛋白结构数据库(AlphaFold DB)[4],已经发布了超过 2 亿个蛋白质结构[3],但其中的蛋白质结构都是根据已知蛋白质序列预测得来的,并非直接通过实验测定。

到稿日期:2023-11-01　返修日期:2024-03-05
通信作者:孟朝晖(mengzhaohui@hhu.edu.cn)

近年来，深度学习领域的快速发展为蛋白质结构预测提供了新的机会。大规模蛋白质语言模型[5](Protein Language Model，PLM)在蛋白质预测任务中取得了很大的成绩，其中最具代表性的 AlphaFold2[6]是一个具有革命性的人工智能蛋白质模型，其在 CASP14 蛋白质结构预测任务上达到了原子级别的预测准确度。AlphaFold2 的突破，为其在生物医学研究的应用创造了条件[37]。端到端蛋白质结构预测方法利用深度学习技术对氨基酸序列的三维结构进行预测，端到端网络模型关注的主要是输入序列和输出结构之间的关系[8]。起初，端到端的方法的预测效果并不明显[9]，但 AlphaFold2 在 CASP14 取得突破后，证明了端到端深度学习架构的可行性[10-11]。AlQuraishi 提出的递归几何网络(RGN)[12]是用于蛋白质结构预测的端到端深度学习架构的开创性尝试之一。RGN 是一个端到端可微模型，通过微分原语优化输入到输出。在 RGN 被提出的同时，Ingraham 等提出了端到端可微模型神经能量建模与优化(NEMO)[13]。NEMO 通过预测蛋白质的空间特征，利用朗之万动力学和基于这些特征的原子插补网络推断输入序列的原子坐标[3]。DeepMind 的端到端模型 AlphaFold2 已经被证明具有预测许多蛋白质三维结构的能力，人工智能在蛋白质结构预测领域取得了显著的成就。但其局限性也很明显，在预测目标结构时，PDB 中是否含有目标结构的同源物对 AlphaFold2 的预测准确率有显著影响[14]，对孤儿蛋白的预测精度仍然有限[3,15-16]。

Transformer 网络模型[17]是一种基于自注意力机制的深度学习模型，它由编码器和解码器组成，每个部分都包含多个子层，其中一个是多头注意力层，用来学习输入或输出序列内部或之间的关系。蛋白质数据集利用 Transformer 网络模型，可以有效地提取残基所处的全局和局部序列环境特征，这有助于预测蛋白质的结构。其次，模型利用自注意力机制[17]学习残基中原子的相互作用信息，这有助于理解蛋白质序列中不同位置之间的依赖关系。

基于以上分析，对于蛋白质预测问题，在深度学习领域，传统的方法都是学习数据集中蛋白质结构中的特征来达到预测的目的，并取得了很可观的成绩。本文从蛋白质物质的空间结构角度考虑，将蛋白质的空间结构特征在数据集上表达出来，建立了新的数据集；并将数据集与 Transformer 网络模型结合进行了相关实验。

本文的贡献包括：1)以新的方法思路来研究蛋白质结构预测问题，即从蛋白质空间结构的角度考虑，建立了一个以原始 PDB 为基础，在特有的局部坐标系下能够表达出蛋白质的空间结构及其他完整特征的河海图结构蛋白质数据集(HohaiGPDB)；2)对 HohaiGPDB 数据集进行数据处理，将其与 Transformer 网络模型相结合，构建了一个蛋白质结构数据的预测模型，有效地利用新数据集中的空间特征对蛋白质结构做预测，并达到了一个良好的预测准确率，从而证明了新数据集的价值。

2 相关工作

本章主要从蛋白质数据集和蛋白质数据的预测模型这两个方面来介绍相关工作。

PDB(Protein Data Bank)，是一个专门收集和存储蛋白质及核酸三维结构资料的数据库。PDB 长期以来都是蛋白质问题相关的主流数据集，但其中的数据内容比较复杂和多样化[18]。其中包含了来自不同物种的蛋白质和核酸的结构信息，这些结构信息通过实验测定，在原始数据中还掺杂着核酸的数据，并且还有部分蛋白质数据不完整，这给数据处理和解析带来了一定的困难，且没有将蛋白质中更深层次的特征挖掘出来。

蛋白质模型预测相关领域已经有了近 60 年的发展历史[19]，随着计算机技术和生物信息学技术的不断进步，近年来蛋白质预测领域取得了一些突破性的进展[20]。如今，基于人工智能的蛋白质结构预测方法已经成为了研究热点，其中最引人注目的是深度学习算法的应用[21]。利用深度学习算法可以从大量数据中提取关键特征，并利用这些特征来预测蛋白质的结构。

DeepContact[22]是基于深度学习的蛋白质接触预测模型，在模型网络上结合卷积神经网络 CNN[23]和 RNN[23]，预测蛋白质相互作用界面上的接触残基，其预测的准确性有待提高。DeepContact 方法虽然可以预测蛋白质相互作用界面上的接触残基，但预测的准确性并不总是能够达到实际应用的需求。RGN 模型[13]是一种创新的蛋白质预测模型，能够通过氨基酸序列来预测蛋白质的三维结构，输入的氨基酸序列包括氨基酸类型、概率分布和位置索引信息，通过循环神经网络来处理序列数据。作为一个端到端的模型，其整个学习过程是可微分的，有助于模型的优化和训练。但其所输入的数据并不能体现蛋白质本身的空间结构，且模型在处理数据时并行化处理的效率较低。

基于上述研究，本文的模型在数据上有所创新。已有模型都是仅以蛋白质表面特征为基础进行学习，本文所提预测模型突出了蛋白质的结构特征，从而能更好地理解蛋白质的结构。在蛋白质结构预测任务中，序列的依赖性非常重要。本文所采用的 Transformer 模型网络以自注意力机制为基础，可从蛋白质序列中有效地捕捉残基所处的全局和局部序列环境的空间特征，通过自注意力机制学习残基及原子间的相互作用信息，从而提高预测结果的准确率。

3 河海图结构蛋白质数据集

本章主要介绍河海图结构数据集的来源、特点和数据结构。蛋白质本质上是一种空间结构，提出 HohaiGPDB 蛋白质数据集的目的就是将蛋白质的空间空间结构展现出来，并利用深度学习方法进行研究，从一种新的角度去解决蛋白质预测问题。

PDB 蛋白质数据是一种全球共享的结构生物学数据库[1]。本文在原始 PDB 文件的基础上将数据特征做了进一步

的处理，在新的命名规范下得到了更易处理的 JSON 类型的蛋白质数据库。

规范的数据集能够更全面地表达出物质结构的特征。在介绍完整的数据集之前，需先了解本数据集所规定的蛋白质氨基酸的表达方案。

3.1 氨基酸的表达方案

蛋白质氨基酸残基有 20 种，但在实际的蛋白质肽链中，首端(N 端)与尾端(C 端)的氨基酸残基与肽链中间的同名残基在具体分子结构上是有差别的。新的命名系统采用 60 个基本残基的编码方案，区分首端、尾端与中间的残基，残基名依然采用标准的 3 字母名，在前面加上“N_”“C_”“R_”，形成 5 字母残基编码。图 1 给出了肽链上 3 种不同位置的残基。

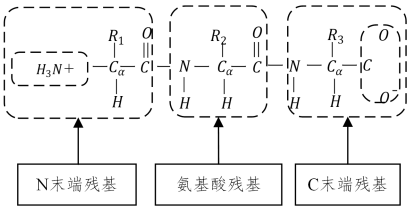

图 1 肽链和残基示意图

Fig. 1 Schematic diagram of peptide chains and residues

下面解释数据集中各个标志的含义。

1)原子名

四字母的名字为原子名称。这个方案由 PDB 数据库的原子名称改进而来，其中第二个字母为元素名(氮碳氧氢硫 NCOHS)；第三个字母主要表示氨基酸残基侧链编号，如 A-alpha，B-beta，G-gamma，D-delta，E-epsilon，Z-zeta，H-eta。第一个字母和第四个字母为同类项编号，规则不是很一致，视具体残基而定，遵循“同一个残基中不能有重复和歧义”的原则。

2)共价键及键长

原子之间的实线为共价键。原则上来说，对于一个完整的蛋白质肽链，这些共价键是蛋白质肽链能够成型的基础。键长指原子间的距离，以埃(10^{-10} m)为单位。真实生物活体中的蛋白质，其键长是个动态变化的值，但可以简化为一个固定值。

3)原子局部立体关系

分子式表达是平面的，但实际上，每个蛋白质氨基酸残基也是立体结构的，其中每个原子都可以被看作是一个局部的立体中心点，其周边的原子位于中心原子的立体周围。具体的位置关系主要有两种：一种是四面体结构，比如 0CA0 及其周围的 4 个原子(0HA0，0C00，0CB0，0N00)；另一种是(大致的)平面结构，比如 R_ARG 中的 0NE0 及其周围的 3 个原子(0CD0，0HE0，0CZ0)。

4)扭转角

扭转角是针对 4 个点形成的空间结构定义的一种角度，如图 2 所示。图中有 i，j，k，l 4 个空间点(不共面)，其中 i，j，k 3 个点确定一个平面，j，k，l3 个点确定另一个平面，两个平面之间的夹角φ_{ijkl}就是空间四点扭转角，扭转角的主体是中间的边(j--k 边)，但是 4 个点缺一不可。

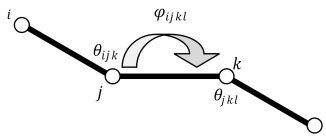

图 2 扭转角

Fig. 2 Torsion angle

扭转角定义的有效性：在残基图中的实线共价键上，部分有扭转角(标记为 φ，ω，χ_1 和 ψ 等)，部分没有。将整个蛋白质肽链抽象成一个图，图中的结点为原子，边为共价键，那么，在图中边缘的边不能成为四点空间结构的中间边(比如图 3 中 R_ALA 的 0CB0－2HB0)，在蛋白质结构中就不会形成扭转角；而具有四点结构的中间边就可以形成扭转角，比如 R_ALA 中的四点结构 0C00-0CA0-0CB0-1HB0，其扭转角为 χ_1。

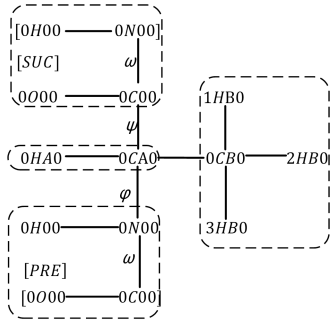

图 3 残基 R_ALA

Fig. 3 Residue R_ALA

扭转角定义的唯一性：以 R_ALA 中的 χ_1 扭转角为例，实际上可以有 9 种以 0CA0-0CB0 为中间边的四点空间结构。那么以哪一个为准呢？在数据集的系统中，按照主链优先的原则，对每个扭转角的 4 个关联点及其次序有系统的规定。

3.2 蛋白质肽链的半边图表达

在新数据集中增添了半边的数据结构，下面介绍肽链中半边图的表达以及本数据集独有的基于图结构的局部坐标系。将蛋白质肽链抽象表达为一个图，原子为图的结点，共价键为边。进一步，将每个边一分为二，称为半边(HALFLINK)，两个半边构成一个边。这样，系统的图就是一个由基本元素即顶点和半边组成的图，边就不再是基本元素了。在这样的图结构下，半边可以认为是属于顶点的，每个顶点具有若干个半边，这些半边与其他顶点的半边结合在一起，形成相应的边。

这种基于半边图的表达方式，可以将每个顶点(原子)看作一个局部中心，而其各个半边分别指向周边的其他原子，那么其他原子的位置就可以表达为以该中心原子为原点的局部坐标系下的局部点，而且这种局部坐标可以具有统一的表达方案，就是本系统独立开发的局部极坐标系。

1)原子局部坐标系数据

每个原子数据都是以自己为原点，附带其周围若干原子局部极坐标位置的数据包。

2)共价键的键长和扭转角

每个边都有键长，并且大部分边都具有扭转角。将以上

两种数据结合在一起，可以唯一确定一个蛋白质的立体结构（即每个原子的三维坐标）。反过来，如果有完整的蛋白质肽链的三维坐标，则可以计算出所有原子的局部坐标系数据和所有的键长和扭转角。这个结果具有唯一性，即如果一个蛋白质肽链立体结构是唯一确定的，其笛卡尔系下的三维坐标可能是不一样的（蛋白质肽链会平移和旋转），但计算出的原子局部坐标和扭转角是唯一确定的。

3.3 数据集介绍

河海图结构蛋白质数据集一共有 278807 个 JSON 文件，每个文件表达一个蛋白质肽链，就是一条完整的蛋白质氨基酸链形成大分子的每个原子坐标数据。这些坐标值数据均由生物物理实验方法测定，主要为 x-ray[24] 方法和 NMR 方法[25]。

本数据集的命名规范建立在上文所介绍的蛋白质氨基酸的表达方案基础上。下面以文件 1viiA00000.json 为例，具体介绍各类数据的意义。

1）文件名规范

1vii 为蛋白质的名字，A 为链的名字（一个蛋白质可以包含一个或多个氨基酸肽链），00000 为模型的编号（模型指实验测定数据的不同结果），json 为扩展名，文件数据符合 JSON 格式规范。

2）文件数据格式

原始的 PDB 数据文件中，不是每个原子都有坐标值，没有坐标值的原子不罗列；在本文的 HohaiGPDB 数据文件中，每个原子都在列，是否有坐标用 vld 标识。

蛋白质数据文件的原子名也是与残基名合并表达的，比如 N_GLU_0N00，表示首端 GLU 残基的 0N00 原子。之所以这样表示，是因为不同残基的原子（比如 N_GLU_0N00 与 N_ALA_0N00），即使是相同名称的原子，但其物理化学性质是不同的。

HohaiGPDB 数据文件的原子名也是独特的（四字符定长），与原始 PDB 数据原子名（不定长）不同，四字符中，第二个为元素名（氮碳氧氢硫 NCOHS），其他为编号。

每个 JSON 文件包含 7 个键，分别是 GRAPHNAME，RESIDUENUMBER，VERTEXNUMBER，以及另外 4 个列表（数组）RESIDUELIST，VERTEXLIST，HALFLINKLIST，LINKLIST。

GRAPHNAME 的值表示氨基酸肽链的名称，这里和 JSON 文件的名称相同，命名规范和文件名的命名规范相同。

RESIDUENUMBER 的值表示此氨基酸肽链中残基的个数。

VERTEXNUMBER 的值表示此氨基酸肽链中原子的总个数。

RESIDUELIST 表示的是氨基酸残基名列表，此列表中包含了 RESIDUENUMBER 条数据，每条数据只包含残基序数和残基名称。

VERTEXLIST 表示的是氨基酸肽链中原子的列表，此列表中包含了 RESIDUENUMBER 条数据。

HALFLINKLIST 表示的是氨基酸肽链中半边的列表。半边的数量是边的数量的二分之一。列表中除了半边的基本信息外，还包含半边的宿主原子、目标原子和目标原子相对于宿主原子的两种位置坐标（极坐标和直角坐标）表示和所属的边。

LINKLIST 表示的是氨基酸肽链中边的列表，列表里边的数量是半边的数量的两倍，列表中的每条数据包含每条边及其两个原子的相关信息，包括键长和扭转角的角度信息。由于不是每条边都有扭转角，故其中也包含着扭转角有效性 vld 的信息。

4 数据处理与实验

本章将会介绍使用 Transformer 网络模型的原因；为了构建预测模型，如何对新数据集进行处理；怎样将 HohaiGPDB 数据集与模型相结合；以及所采用整个模型的结构和模型实验的结果。

在深度学习领域，有许多优秀的传统模型网络。基于研究的问题和所使用的实验手段，本文决定采用 Transformer 网络模型。采用端对端的方法，将蛋白质的预测问题转换成数据序列的预测，更加注重序列间全局各个位置的关系，蛋白质的特征也保存在全局序列中的关系内。Transformer 网络模型适用于序列问题，且更适用于捕捉序列中上下文元素间的关系，从而得到序列所包含的空间结构特征。

由于实验的目标是预测蛋白质的结构，因此可以对原子的位置进行预测。而每个原子在空间中的位置，可以由原子间的键长和扭转角推算得到。也可以理解为本文利用蛋白质局部坐标的方式替换了全局直角坐标的方式来表达出蛋白质的结构，所以可以通过得到原子间的键长和扭转角度数来预测蛋白质的结构。

4.1 数据处理

4.1.1 数据切片

如图 4 所示，模型的第一步就是要提取 HohaiGPDB 文件中的有效数据作为模型的输入。本文所采取的模型网络为传统的 Transformer 模型。传统网络模型并不直接适配于蛋白质数据，因为传统的 Transformer 模型主要应用在自然语言处理领域，一般的自然语言序列长度并不会达到像蛋白质数据那样长，即原始的模型并不适用于序列长度过大的数据样本。HohaiGPDB 数据集的每个 JSON 文件中的 LINKLIST 中的数据的元素个数基本都在 2000 以上，直接对每个 JSON 文件粗暴套用模型将无法发挥所选择的传统模型中神经网络的优势，故对于 HohaiGPDB 数据集的预处理，本文打算采用切片的形式，无论是输入序列还是预测序列，都控制在一个切片所设置的长度范围内。为了得到输入和输出序列，从 HohaiGPDB 数据集中的每个 JSON 文件中的 LINKLIST 中提取数据，所对应的输入是两个原子名称序列 $V_n=\{V_{n1},V_{n2},\cdots,V_{n2L}\}$ 映射为键长和扭转角目标位置序列 $P=\{P_{l1},P_{t1},P_{l2},P_{t2},\cdots,P_{lL},P_{tL}\}$。原始的两个序列长度均为 $2L$，L 表示切片的长度。V_n 是原子的名称，P_l 表示键长，P_t 表示扭转角的度数。

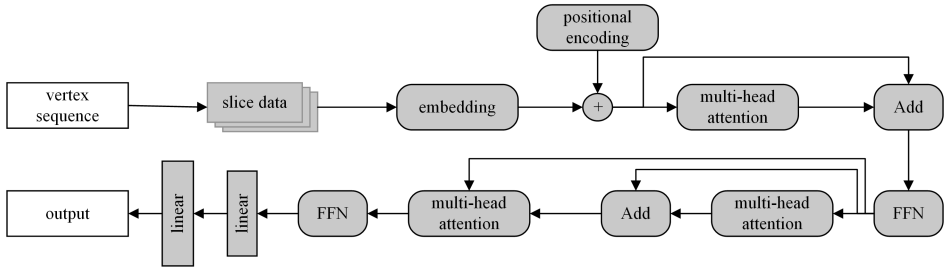

图 4　蛋白质数据序列预测网络结构

Fig. 4　Structure of protein data sequence prediction network

每个蛋白质 JOSN 文件中 LINKLIST 中数据的数量不固定，最后一个切片大概率不会凑足 L 个切片长度。将未凑足的输入和输出部分尾部补齐，即在模型训练时将提取的原子名称、键长和扭转角列表转换成数字序列之后，长度未凑足的地方补零。但在进行归一化操作时，在 softmax 函数（见式(1)）中，e 的 0 次方等于 1，补齐的部分可能会对结果产生一定的影响，这里做一个掩码操作[17]，对补齐的部分设置了一个很大的偏置负数，以在进行归一化操作时保证在切片中所补齐数据的无效性。

$$\sigma(z_i)=\frac{e^{z_i}}{\sum_{j=1}^{K}e^{z_j}} \tag{1}$$

4.1.2　词汇表

为了能够识别所提取的输入和输出数据，需要制作输入和输出词汇表来帮助输入和输出序列向量化。在制作数据集时，将所有的原子名称保存在 HohaiGPDB 数据集之外的一个.dict 文件中，将所有的名称提取出来放入一个列表中作为输入词汇表。HohaiGPDB 数据集中原子的总个数是 1 037。输出词汇表就是长度和角度的数值范围，在数据集内长度和扭转角的数学数值都精确到了小数点后 3 位；而在数据的预处理中，为了提高模型的效率，降低了输出词汇表的精确度，只将其数值部分保留到了小数点后 1 位，键长和扭转角的共同的取值范围在 0.0 到 360.0 之间，输出词汇表就是 0.0 到 360.0 共 3 601 个输出数值。

4.2　模型架构

4.2.1　网络结构

传统的 Transformer 模型，其结构完全基于自注意力机制，这样可以使得模型在进行预测时充分学习不同的位置长度以及扭转角特征，提高蛋白质局部模型的位置特征学习能力。得到整个模型的输入序列之后，就可将其加入模型的处理中。

首先，通过切片操作已经把数据集中需要用到的输入和输出数据提取出来，还需要在输入序列部分加入开始标识符 S 和结束标识符 E，这些标识符也被放进相应的词汇表中。根据所作的输入词汇表，将原子名称映射成相应的数字序列（式(2)）。初始化序列编码的大小为 $[src\_size, d\_model]$，src_size 为词汇表中所有元素的数量，d_model 为字向量的维度。然后，将每个原子名称映射成词向量表示的 $\boldsymbol{X}_i$。位置编码的过程如式(3)和式(4)所示，其中 pos 表示元素的位置，i 表示元素向量维度的序号，取值范围是 $[0, embedding\_dimmension/2)$，然后与位置编码 $\boldsymbol{P}_i$ 相加，得到处理后的嵌入词向量 $\boldsymbol{E}_i$。

$$src_{meb}=embedding(src\_size, d\_model) \tag{2}$$

$$\boldsymbol{PE}(pos,2i)=\sin(pos/10\,000^{2i/d_{model}}) \tag{3}$$

$$\boldsymbol{PE}(pos,2i+1)=\cos(pos/10\,000^{2i/d_{model}}) \tag{4}$$

$$\boldsymbol{E}_i=\boldsymbol{X}_i+\boldsymbol{PE}_i \tag{5}$$

接下来将以 $\boldsymbol{E}_i$ 词向量这种形式作为输入进入 Transformer 的编码器部分，由 6 个编码器层组成，在每个编码层中进行编码器部分的自注意力机制计算。然后通过式(6)计算自注意力分数 $attention$。$\boldsymbol{E}_i$ 经过线性变换后得到编码器中的 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 矩阵。d_k 是矩阵的维度。然后将 $attention$ 和 $\boldsymbol{V}$ 矩阵相乘后的矩阵经过一个线性变换后作为输出，再经过一个全连接前馈子层，其结果编码器输出向量 $\boldsymbol{enc}_{out}$。

$$attention=\mathrm{softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot\boldsymbol{V} \tag{6}$$

$$\boldsymbol{enc}_{out}=LN(\boldsymbol{E}_i+EnAttn(\boldsymbol{E}_i)+LN(\boldsymbol{E}_i+FFN(\boldsymbol{E}_i)) \tag{7}$$

得到了输出向量 $\boldsymbol{enc}_{out}$ 和自注意力计算分数 $attention$ 之后，进入解码器层。下面是 Transformer 的解码器部分。将带有序列初始符的原子序列 $\boldsymbol{dec}_{input}$（即右移一位的输入序列）作为解码器的一个输入，经过一个自注意力机制得到 $\boldsymbol{dec}_{midout}$，同时编码器的输出隐藏层 $\boldsymbol{enc}_{out}$ 进入解码器的下一个自注意力机制；然后根据 $\boldsymbol{dec}_{outputs}$ 和前一个时间状态生成的输出来生成预测结果 $\boldsymbol{dec}_{out}$，进而生成解码器的输出隐藏层状态 $\boldsymbol{dec}_{output}$。Transformer 解码器也是由 6 个解码层组成，每一层又包含着如下的子层：多头注意力子层、编码层、解码器注意力子层和全连接前馈子层。这些子层均包含着残差连接和归一化。其具体的表达式如下：

$$\boldsymbol{dec}_{outputs}=decoder(\boldsymbol{dec}_{input},\boldsymbol{enc}_{out}) \tag{8}$$

$$\boldsymbol{dec}_{midout}=LN(DecAttn(\boldsymbol{dec}_{in})+\boldsymbol{dec}_{in}) \tag{9}$$

$$\boldsymbol{dec}_{out}=LN(EDAtn(\boldsymbol{dec}_{midout},\boldsymbol{enc}_{out})+\boldsymbol{dec}_{midout}) \tag{10}$$

$$\boldsymbol{dec}_{output}=LN(\boldsymbol{dec}_{out}+FFN(LN(\boldsymbol{dec}_{out})) \tag{11}$$

最后，解码器的输出隐藏状态 $\boldsymbol{dec}_{output}$ 经过一个线性变换和归一化操作之后生成键长和扭转角序列的概率分布，输出过程如式(12)所示。每个原子预测得到出扭转角或键长的大小，当预测得到结束符时，就表示已经生成了完整的输出序列。

$$\boldsymbol{P}_i=\mathrm{softmax}(Linear(\boldsymbol{dec}_{outputs})) \tag{12}$$

键长和扭转角在预测阶段采用的是贪婪解码策略，即根据每个步骤的最优选择来生成输出序列。在每个解码步骤中，通过贪婪解码器[17]选择具有最高概率的输出符作为当前步骤的输出，这种方法可以快速生成输出序列，并更新模型内部状态以继续生成下一个输出元素。

4.2.2 参数设置

为了提供足够的表示能力并平衡计算效率，元素嵌入向量的维度均设置为512，子注意力机制中向量的维度均为64，通过传统的网络模型的经验以及所进行的实验对比，在多头子注意力机制中共设置了8个Head，编码器和解码器的层数都设置为6。

4.3 实验与结果

4.3.1 数据集

河海图结构蛋白质数据集，是本文所提出的全新的蛋白质数据集，一共有278807个JSON文件。由于数据集比较庞大，本文只选取了其中部分完整的数据集，但所选取的数据集容量足够表现出完整蛋白质的特征；并且对小部分达不到要求容量的切片采取了补数0和掩码的处理。对数据集采用不同的切片大小进行实验，通过对比得出切片大小对实验结果的影响。

4.3.2 实验说明

本文实验平台为Ubuntu 16.04操作系统，代码均在PyTorch深度学习框架下实现，模型的训练使用了一块NVIDIA GeForce RTX 2080 Ti GPU显卡，在训练时提取HohaiGPDB数据集中的部分氨基酸肽链的JSON文件，对每个文件进行切片，然后添加到了序列列表中。通过对比实验，在训练模型时将批大小(batch size)设置为32，键长和扭转角预测阶段使用交叉熵损失函数(Cross Entropy Loss)和随机梯度下降(Stochastic Gradient Descent，SGD)优化，全局学习率设置为0.001，使用线性衰减对学习率进行动态设置。

4.3.3 评价指标

由于本文使用了全新的数据集和预测信息，数据集的精确度又较高，同时考虑到蛋白质结构的实际因素，本文认为所预测结果在一定的误差范围内是可以接受的。因此将预测的输出序列与正确的输出序列中的数据依次对比，扭转角的度数之差在10度以内，键长之差在0.1埃(10^{-10} m)范围内，将所预测的结果都视为正确。通过此种方式分别对每一个切片求得准确率，最后所有切片的准确率相加再求平均值就得到了最终的平均准确率。准确率的计算式如式(13)所示：

$$Accu=\frac{CTA}{TA} \tag{13}$$

其中，CTA(Correct Total Amount)表示在正确结果误差区间内正确的总数，TA(Total Amount)表示所预测的总数。

4.3.4 实验结果

在此实验环境下，根据本研究所采用的评估策略，HohaiGPDB数据集上的最佳预测准确率可以达到59.38%。此结果表明了本数据集的重要研究价值。

同时，为了验证切片大小对实验结果的影响，对所用数据集采用了不同容量的切片策略来测试结果。

表1 不同切片大小结果的比较

Table 1 Result comparison of different sections

Section size	*Accuracy*/%
150	57.36
100	59.38
50	55.53

从表1中可以看出，在所设置的切片大小为100时，测试数据集的最佳准确率为59.38%；当切片大小减小为50时，测试数据集的最佳准确率为55.53%；切片大小为150时，准确率为57.26%。对比结果可以表明，切片的大小对实验结果没有决定性的影响。

结束语 本文从一个新的角度研究蛋白质预测问题，在原始的PDB蛋白质数据集的基础上提出了以局部坐标为基础，能够表达出蛋白质空间结构的HohaiGPDB数据集。相对于当前流行的蛋白质数据集，HohaiGPDB更能表明各种氨基酸肽链、各个原子和共价键之间的空间特征，命名规范相较于原始的PDB数据也更加全面。同时，其为今后深入研究蛋白质模型结构提供了更全面的数据集。为了证明所提数据集的价值，利用Transformer模型对其进行了相关的预测实验。通过对新数据集进行分片处理，使得数据更适配于原始的模型网络，对所提供的蛋白质数据进行训练，其在测试集上可以得到良好的预测准确率。实验结果证明了本文所采用思路的可行性。

HohaiGPDB数据库的规模庞大，在实验中并没有利用全部规模的数据量和数据中的全部特征(如半边的空间特征、半边与边和原子之间的序列关系等)。对数据集中其他未利用的特征进行进一步发掘，以及利用本数据集进行其他蛋白质研究，将是未来工作的重点；同时，对蛋白质动态特征进行探索也是今后研究的另外一个思路。

参 考 文 献

[1] BERMAN H M, BATTISTUZ T, BHAT T N, et al. The protein data bank[J]. Acta Crystallographica Section D: Biological Crystallography, 2002, 58(6): 899-907.

[2] BATEMAN A, MARTIN M J, ORCHARD S, et al. UniProt: the universal protein knowledgebase in 2023[J]. Nucleic Acids Research, 2022, 51(D1): D523-D531.

[3] PENG C X, LIANG F, XIA Y H, et al. Recent Advances and Challenges in Protein Structure Prediction[J]. Journal of Chemical Information and Modeling, 2023, 64(1): 76-95.

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

[5] CHEN B, CHENG X, GENG Y, et al. xtrimopglm: Unified 100b-scale pre-trained transformer for deciphering the language of protein[J]. arXiv: 2401.06199v1, 2024.

[6] BRYANT P, POZZATI G, ELOFSSON A. Improved prediction of protein-protein interactions using AlphaFold2[J]. Nature Communications, 2022, 13(1): 1265.

[7] AKDEL M, PIRES D E V, PARDO E P, et al. A structural biology community assessment of AlphaFold2 applications[J]. Nature Structural & Molecular Biology, 2022, 29(11): 1056-1067.

[8] JISNA V A, JAYARAJ P B. Protein structure prediction: conventional and deep learning perspectives[J]. The Protein Journal, 2021, 40(4): 522-544.

[9] PEARCE R, ZHANG Y. Toward the solution of the protein structure prediction problem[J]. Journal of Biological Chemis-

try,2021,297(1).

[10] KANDATHIL S M,GREENER J G,LAU A M,et al. Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins[J]. Proceedings of the National Academy of Sciences,2022,119(4):e2113348119.

[11] WEISSENOW K,HEINZINGER M,STEINEGGER M,et al. Ultra-fast protein structure prediction to capture effects of sequence variation in mutation movies[J]. arXiv: 2022. 11. 14. 516473v2,2022.

[12] ALQURAISHI M. End-to-end differentiable learning of protein structure[J]. Cell Systems,2019,8(4):292-301. e3.

[13] INGRAHAM J,RIESSELMAN A,SANDER C,et al. Learning protein structure with a differentiable simulator[C]// International Conference on Learning Representations. 2018.

[14] JONES D T,THORNTON J M. The impact of AlphaFold2 one year on[J]. Nature Methods,2022,19(1):15-20.

[15] WANG W,PENG Z,YANG J. Single-sequence protein structure prediction using supervised transformer protein language models [J]. Nature Computational Science,2022,2(12):804-814.

[16] LIN Z,AKIN H,RAO R,et al. Evolutionary-scale prediction of atomic-level protein structure with a language model [J]. Science,2023,379(6637):1123-1130.

[17] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J]. arXiv:1706. 03762,2017.

[18] BERMAN H M. The protein data bank:a historical perspective [J]. Acta Crystallographica Section A:Foundations of Crystallography,2008,64(1):88-95.

[19] AL-LAZIKANI B,JUNG J,ANG Z,et al. Protein structure prediction[J]. Current Opinion in Chemical Biology, 2001,5(1): 51-56.

[20] PHAN H K,DANG T H. Protein structure prediction using Deep Learning[R]. VNU University of Engineering and Technology,2018.

[21] TORRISI M,POLLASTRI G,LE Q. Deep learning methods in protein structure prediction[J]. Computational and Structural Biotechnology Journal,2020,18:1301-1310.

[22] SKWARK M J,RAIMONDI D,MICHEL M,et al. Improved contact predictions using the recognition of protein like contact patterns[J]. PLoS Computational Biology, 2014, 10(11): e1003889.

[23] LECUN Y,BENGIO Y,HINTON G. Deep learning[J]. Nature, 2015,521(7553):436-444.

[24] SRIVASTAVA A,NAGAI T,et al. Role of computational methods in going beyond X-ray crystallography to explore protein structure and dynamics[J]. International Journal of Molecular Sciences,2018,19(11):3401.

[25] BILLETER M,WAGNER G,WÜTHRICH K. Solution NMR structure determination of proteins revisited[J]. Journal of Biomolecular NMR,2008,42:155-158.

WEI Xiangxiang, born in 1999, master. His main research interests include artificial intelligence and neural network.

MENG Zhaohui, born in 1968, associate professor. His main research interests include neural network and artificial intelligence.

(责任编辑:柯颖)