预测并表征蛋白质三维结构

曾浩

2024 年 1 月

摘要

学习有效的蛋白质表征在生物学中的各种任务中至关重要，例如预测蛋白质的核酸结合位点。蛋白质的功能由其结构决定。使用蛋白质的结构表征可以更好地预测蛋白质与核酸的相互作用。现有的蛋白质语言模型通常在大量未标记的氨基酸序列上预训练，然后在下游任务中使用一些标记数据对模型进行微调。在这份报告中，我复现了一个基于蛋白质三维结构的蛋白质语言模型 GearNet，它可以对蛋白质的三维结构进行表征。由于实验方法获取的核酸结合蛋白三维结构处于结合状态，而结合导致的蛋白质构象变化会影响蛋白质核酸结合位点的预测准确率，因此我使用了 ESMFold 方法来预测蛋白质的三维结构，并使用 GearNet 对预测的蛋白质三维结构进行表征，以应用到后续的蛋白质核酸结合位点的预测任务中。

关键词：蛋白质语言模型；蛋白质三维结构；深度学习

1 引言

预测并表征蛋白质的三维结构是预测蛋白质-核酸相互作用的前置工作。蛋白质-核酸相互作用在细胞活动中起着至关重要的作用，例如控制转录和翻译、DNA 修复、剪接、细胞凋亡和应激反应的介导。核酸与蛋白质的相互作用主要有两种方式：一种是蛋白质与 DNA 的结合，另一种是蛋白质与 RNA 的结合。核酸结合蛋白（NBP）的突变会影响基因表达并导致疾病。例如，TDP-43 是一种重要的 RNA 结合蛋白，其基因突变可能导致渐冻症。湿实验方法可用于预测蛋白质的核酸结合残基，但这些方法具有挑战性且成本高昂。使用计算方法来预测蛋白质核酸结合残基更加高效，尤其是使用深度学习的方法。

蛋白质的功能由其三维结构决定，引入蛋白质三维结构特征的深度学习方法可以更加有效地预测蛋白质的核酸结合残基。为了将蛋白质的三维结构嵌入到深度学习模型中，我们首先要对蛋白质的三维结构特征进行表征。现有的蛋白质语言模型很多都是基于氨基酸序列的，无法对蛋白质的三维结构进行表征。我复现了一个基于蛋白质三维结构的蛋白质语言模型 GearNet [7]，GearNet 是一种简单而有效的基于结构的蛋白质编码器，它通过添加不同类型的顺序或结构边来编码蛋白质结构的空间信息，然后在蛋白质残基图上执行关系消息传递，这可以通过边缘消息传递机制进一步增强。GearNet 在多个基准测试中实现了非常强大的性能。

现有的预测蛋白质核酸结合位点的方法大多使用来自 PDB 的核酸结合蛋白三维结构来预训练模型。这些结构都是通过实验方法获取的处于与核酸结合状态的蛋白质结构，与未结

合状态下的蛋白质结构相比，其空间构象发生了变化。使用这种存在构象变化的蛋白质三维结构来训练核酸结合位点预测模型，可能会影响模型的预测性能。为了避免这一问题，我使用 ESMFold [3] 方法来预测蛋白质的三维结构，避免了构象变化带来的影响。

2 相关工作

现有的方法基于多种特征来学习蛋白质表征。比如，有的方法基于蛋白质的氨基酸序列来学习蛋白质表征 [4,5]，有的方法基于多序列对比来学习表征 [1]，也有的方法基于蛋白质的三维结构来学习表征 [2]。这些方法的共同点是通过学习蛋白质的表征来优化下游任务，例如预测蛋白质的功能位点。在这些方法中，基于特征的方法一般且理应比基于序列的方法取得更好的效果。除了基于结构的蛋白质编码器外，现有的基于结构的预训练蛋白质语言模型不多，它们使用了对比学习和自预测等方法来预训练模型。

3 本文方法

3.1 构建和更新蛋白质图

构建蛋白质图，首先将蛋白质的每一个残基添加到图中作为图的节点，然后根据蛋白质序列和三维结构来添加边。在 GearNet 中，一共有三种边：序列边、k-近邻边、半径边。序列边的添加以残基在氨基酸序列中的距离来判断，如果距离等于 1 或者-1 则在蛋白质图中相应的残基之间添加边。每个残基与在三维空间中距离它最近的 k 个残基之间添加 k-近邻边。以每个残基中心，添加空间中指定半径内的残基之间的半径边。蛋白质图中每一个残基和每一条边都包含了蛋白质的特征信息。

构建蛋白质图之后，GearNet 通过图卷积来更新每个节点特征，通过边缘消息传递机制来更新每条边的特征。

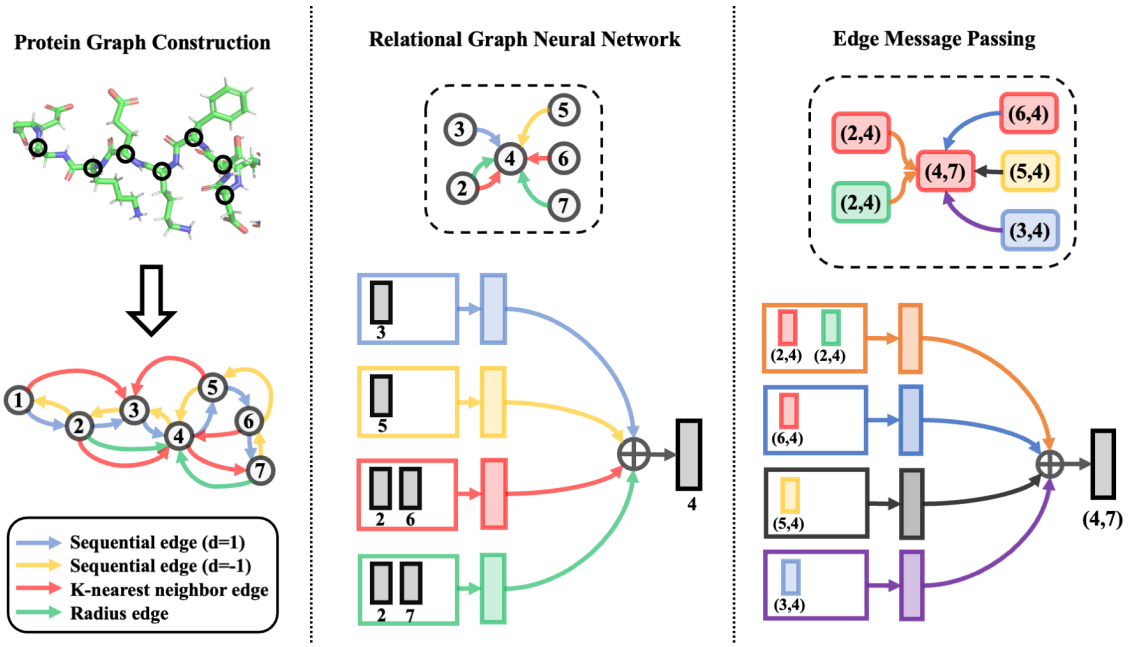

图 1. 方法示意图

3.2 预训练方法

为了捕获蛋白质结构中的生化和空间信息，GearNet 使用了多视图对比学习和自预测方法来预训练模型。多视图对比学习使用在潜在空间中定义的相似性测量，生物相关的子结构彼此嵌入得很近，而不相关的子结构则被映射得很远。多视图的构建使用了子序列裁剪和子空间裁剪。自预测方法就是在给定蛋白质结构的其余部分的情况下预测蛋白质的一部分，本文对单残基、残基对、三联体和四联体进行掩模预测。

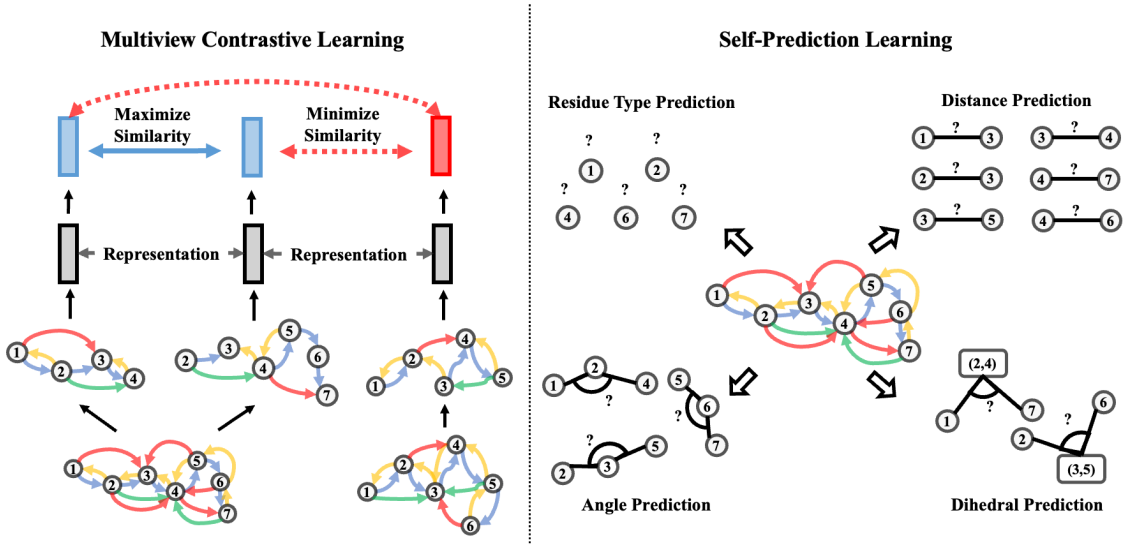

图 2. 预训练方法

4 复现细节

4.1 与已有开源代码对比

GearNet 模型可以对输入的蛋白质结构提取表征，但开源的代码中只有预训练模型以及下游任务实验的部分，并没有提供从蛋白质结构到其表征这样一个端到端的功能。我研究了开源的代码以及 GearNet 所使用 TorchProtein 库，复现了 GearNet 的代码，并实现了端到端的功能，使得对 GearNet 感兴趣的人可以直接使用 GearNet 来表征指定的蛋白质结构，以便应用到下游任务中去。

4.2 实验环境搭建

参考论文的 GiThub 仓库。

4.3 创新点

为 GearNet 实现端到端功能。

5 实验结果分析

我进行了 GearNet 的下游任务实验，取得和论文中一致的效果。我还对端到端的代码进行了实验，使用 GearNet 表征 ESMFold 预测的蛋白质结构。

论文结果

Method	EC	GO-BP	GO-MF	GO-CC
GearNet	0.730	0.356	0.503	0.414
GearNet-Edge	0.810	0.403	0.580	0.450
Multiview Contrast	0.874	0.490	0.654	0.488
Residue Type Prediction	0.843	0.430	0.604	0.465
Distance Prediction	0.839	0.448	0.616	0.464
Angle Prediction	0.853	0.458	0.625	0.473
Dihedral Prediction	0.859	0.458	0.626	0.465
ESM_GearNet	0.883	0.491	0.677	0.501
ESM_GearNet (Multiview Contrast)	0.894	0.516	0.684	0.5016

复现结果

```
19:47:21   EnzymeCommission(
  #sample: 18716
  #task: 538
)
19:47:21   #train: 15170, #valid: 1686, #test: 1860


08:54:30   Evaluate on valid
08:54:36   ------------------------------
08:54:36   auprc@micro: 0.668008
08:54:36   f1_max: 0.697332
08:54:36   >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
08:54:36   Evaluate on test
08:54:42   ------------------------------
08:54:42   auprc@micro: 0.735329
08:54:42   f1_max: 0.763647
```

图 3. 下游任务实验

```
{'graph_feature': tensor([[ 4.9261e+02,  6.8545e+03,  4.8785e+01, -4.5283e+01,  1.5796e+02,
          1.5615e+02,  2.0509e+03,  2.7874e+03, -3.9668e+01, -9.2138e+00,
          3.5193e+03,  1.0343e+02, -5.6242e+01,  3.5078e+01, -3.6075e+00,
          8.9514e+01, -6.5108e+01,  2.3447e+03,  7.4705e+02,  2.1710e+03,
         -2.4376e+01, -6.0209e+01,  4.0445e+01,  4.5065e+03,  2.3635e+03,
         -4.9230e+01,  8.5986e+01,  1.7613e+03,  3.8344e+01, -4.9066e+01,
         -4.2833e+01,  6.8866e+02,  8.5891e+02,  3.5105e+02,  4.7401e+02,
         -4.0533e+01,  1.2276e+03]], requires_grad=True), 'node_feature': tensor([[ 6.4718e+00,  7.8028e+01, -6.2828e-01,  ..., -1.0318e+00,
          1.4035e+00,  3.2575e+01],
        [ 5.3059e+00,  8.7798e+01, -2.7554e-02,  ...,  2.1169e+00,
         -1.0736e+00,  1.8017e+01],
        [ 6.7863e+00,  8.6318e+01,  2.3593e-01,  ...,  4.0522e+00,
         -1.2262e+00,  9.7375e+00],
        ...,
        [ 7.1816e+00,  8.7887e+01, -5.6047e-01,  ..., -5.4005e-01,
         -6.8287e-01,  1.6242e+01],
        [ 6.8304e+00,  6.3908e+01, -2.2534e-01,  ..., -1.0318e+00,
         -4.8421e-01,  2.2518e+01],
        [ 6.7954e+00,  7.6520e+01, -6.2958e-01,  ...,  3.7911e-01,
          3.7662e+00,  3.1793e+01]], requires_grad=True)}
```

图 4. 对预测结构进行表征

6 总结与展望

在本次复现工作，我复现了 GearNet 的下游任务实验，取得了和论文中一致的效果。我使用 ESMFold 预测了 GraphBind [6] 数据集中的蛋白质的三维结构，并通过我实现的端到端代码使用预训练的 GearNet 模型表征了这些预测结构。这次复现工作的不足之处在于，受硬件性能限制，我并没有成功使用大量的蛋白质结构数据的预训练出一个通用的 GearNet 模型，而是使用了作者提供的模型状态字典。未来，我将把使用 GearNet 学习的预测结构表征应用到预测蛋白质核酸结合位点的任务中去。

参考文献

[1] Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, and George M Church. Low-n protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

[2] Pedro Hermosilla, Marco Schäfer, Matěj Lang, Gloria Fackelmann, Pere Pau Vázquez, Barbora Kozlíková, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *arXiv preprint arXiv:2007.06252*, 2020.

[3] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[4] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[5] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[6] Ying Xia, Chun-Qiu Xia, Xiaoyong Pan, and Hong-Bin Shen. Graphbind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research*, 49(9):e51–e51, 2021.

[7] Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*, 2022.