4.1运用神经网络学习突变蛋白的局部结构功能关系

EVmutation[1]和 DeepSequence[2]等无监督学习方法在进化相关蛋白质序列的大比对上进行训练。这些方法可以模拟蛋白质家族的天然功能，但它们无法预测不受长期进化选择约束的特定蛋白质特性。AlphaFold基本实现将蛋白质序列放置在正确的能量最小值中[3]，但缺乏将深度学习应用到相反的问题：结构如何限制给定位点允许的蛋白质。这将限制有害突变的采样，加速靶向诱变和蛋白质工程工作。

为了解决以上问题，Sam Gelman团队新提出了一个有监督的深度学习框架，训练有监督的神经网络学习序列到功能的映射函数。[Anastasiya V. Kulikova](https://link.springer.com/article/10.1007/s10867-021-09593-6" \l "auth-Anastasiya_V_-Kulikova-Aff1)团队使用3DCNN对此进行研究，研究了从局部化学环境（微环境）预测隐蔽残基的CNN模型能否良好预测野生型残基和进化上分化的同源物中的残基，并关注那些高准确率预测的氨基酸分布。

4.1.1关于蛋白质的特征工程及模型训练

在Sam Gelman等人的研究中，他们对蛋白质序列进行编码，编码信息包含了每个氨基酸在每个位置的物理和化学性质。分别采用One hot编码用于标记特定位置的氨基酸，以及AAindex编码捕获氨基酸的物理性质和化学性质，并采用PCA将特征向量降到19维，最后将两个编码连接。团队共测试了线性回归和全连接、数列卷积、图卷积三种神经网络。卷积神经网络中，卷积层可以学习识别β链中常见的极性和非极性氨基酸，使网络能评估整个输入序列的β链倾向，并将这些信息与蛋白质功能关联。

而在[Anastasiya V. Kulikova](https://link.springer.com/article/10.1007/s10867-021-09593-6" \l "auth-Anastasiya_V_-Kulikova-Aff1)的团队中，他们建立了一个蛋白质的三维模型。具体来说，我们在蛋白质数据集中随机取样残基，以创建一个反映每种氨基酸自然丰度的微环境数据集。对每个蛋白质链，我们最多选取50个残基，并且选取的残基数不超过总数的50%[4]，以免模型偏向那些分子量大的蛋白质。利用这些元数据，我们生成了微环境的体素化表示（4D 张量），该表示以α指定残基的碳为中心，并相对于主链定向，使得侧链沿+z轴定向。体素化表示由 3D 空间 （x、y 、z ） 和 7 个辅助通道组成。辅助通道编码有关体素中存在的原子性质（C、H、O、N、S）以及部分电荷和溶剂可及表面积的信息。之后通过两对3D卷积层提取特征，再进入分类块，得到隐蔽残基处20个氨基酸的概率向量

4.1.2 模型评估

[Anastasiya V. Kulikova](https://link.springer.com/article/10.1007/s10867-021-09593-6" \l "auth-Anastasiya_V_-Kulikova-Aff1)的团队在包含130个结构的独立数据集PSICOV[4]数据集上评估他们的模型,野生型序列的预测准确率通常很高，平均达到60%.将生化相似的氨基酸分组后，预测氨基酸组别的能力高于预测特定氨基酸的能力达到71%。接下来，团队询问了该网络在多序列比对 （MSA） 中预测位点的共有氨基酸的能力。这种预测的可能程度取决于给定位点周围的微环境在同源结构中的保守程度。有40%的概率预测正确的共有氨基酸，55%的概率预测正确的类别。

4.1.3 成果及应用

Sam Gelman团队运用他们的模型进行了以下研究：

1） 采用统一流行近似和投影[5]，将网络中的连接层可视化成蛋白质的三维结构（图1[6]），并根据其得分给蛋白质着色，从而了解哪些序列位置对蛋白质的影响最大，了解哪些区域是突变不耐受的。

2 ）在野生型蛋白基础上设计新蛋白质。运用训练的模型对那些相对野生型仅有局部变化的蛋白质进行评估，运用线性回归和三种神经网络四个模型联合打分,这里采用随机重启爬山算法，最大化最低预测功能分数。

4.1.4 讨论与展望

上述研究结果进一步确认了氨基酸与其所处的局部生化环境之间存在着紧密联系的观点，并表明蛋白质中的氨基酸所受的限制条件会随着蛋白质的进化过程而发生改变。模型的错误预测也许为蛋白质工程提供了一种新的视角，即能够识别出那些可能适合进行突变的位置。具体而言，当模型“自信”地预测出与现有野生型不同的氨基酸时，这大概率意味着该位置具有较高的突变潜力，而模型错误预测的特定氨基酸则可能是实现结构稳定或功能增强突变的理想选择。这一发现不仅加深了我们对蛋白质进化机制的理解，还为未来的研究提供了宝贵的指导方向。

参考文献

1. T. A. Hopf et al., Mutation effects predicted from sequence co-variation. Nat. Biotechnol. 35, 128–135 (2017). [Crossref](https://doi.org/10.1038/nbt.3769). [PubMed](https://pubmed.ncbi.nlm.nih.gov/28092658/).
2. A. J. Riesselman, J. B. Ingraham, D. S. Marks, Deep generative models of genetic variation capture the effects of mutations. Nat. Methods 15, 816–822 (2018). [Crossref](https://doi.org/10.1038/s41592-018-0138-4). [PubMed](https://pubmed.ncbi.nlm.nih.gov/30250057/).
3. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589
4. Jones, D.T., Buchan, D.W.A., Cozzetto, D., Pontil, M.: PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28, 184–190 (2011)
5. L. McInnes, J. Healy, UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv [Preprint] (2020). <https://arxiv.org/abs/1802.03426> (Accessed 18 September 2020).
6. Sam Gelman, Sarah A. Fahlberg, Pete Heinzelman, Philip A. Romero & Anthony Gitter. [Proceedings of the National Academy of Sciences](https://doi.org/10.1073/pnas.2104878118) • [code](https://github.com/gitter-lab/nn4dms) • Nov 2021
7. **Learning the local landscape of protein structures with convolutional neural networks.** Anastasiya V. Kulikova, Daniel J. Diaz, James M. Loy, Andrew D. Ellington & Claus O. Wilke. [Journal of Biological Physics 47.4 (2021)](https://link.springer.com/article/10.1007/s10867-021-09593-6)

多年来，预测由单点突变引起的蛋白质稳定性变化一直是一个难题，吸引了众多研究人员的兴趣。蛋白质稳定性极易收到外界环境和基因突变的扰动，即使是微小的扰动，如单点突变，也可以使得活性蛋白走向无功能、错误折叠或者聚合的形式。准确预测蛋白质热稳定性对于药物开发、蛋白质进化分析和酶合成等生物化学领域的应用至关重要。

由于蛋白质分子成的任何变化都可以被视作化学反应，因此热稳定性可以被视作突变发生的趋势，通常表示为一对野生和突变蛋白的吉布斯自由能之差[1]：

\Delta\Delta G = \Delta G\_{\text{M}} - \Delta G\_{\text{W}}

其中M表示突变型，W表示野生型，较高的\Delta\Delta G意味着更高的热稳定性[2]。

本文下面介绍一个自监督模型Pythia[3]，它可以用于零样本预测蛋白质突变后的自由能变化。

2.1 模型基本原理

在假设蛋白质在未折叠状态下的能量在很大程度上不受突变影响的前提下，通过公式推导，我们可以得到自由能变化与野生型、突变型蛋白质氨基酸在所有Rotamer构象中的概率之和的正比关系：

\Delta\Delta G \propto -\ln \frac{P\_{\text{MUT}}}{P\_{\text{WT}}

基于以上结果，Sun的团队使用一种图神经网络架构，以蛋白质的局部结构作为输入，该结构被表示为一个k近邻图（k-NN graph），节点代表氨基酸残基，边则根据Cα原子之间的欧几里得距离来定义。每个节点的特征包括氨基酸类型以及三个二面角（φ, ψ, ω）,边特征包含主链原子间的距离、序列位置信息及链信息。之后使用了消息传递神经网络（MPNN）[4]，使用基于注意力的消息传递和读出功能进行定制。在注意力消息传递层（AMPL）的每一层中，顶点表示使用 Attention 块进行更新，并与 edge 连接表示来获得消息表示。模型的训练任务是预测中心节点正确的氨基酸类型。通过这种方式，Pythia能够解码给定蛋白质中残基之间的内在模式，从而精确地预测突变的影响。

2.2 Pythia模型的可解释性：

由于 Pythia 采用了注意力机制，我们可以利用该模型用于研究它是否确实成功地捕获了蛋白质内复杂的交互。结果显示 Pythia 对突变对的注意力得分较高，这表明 Pythia 对突变体结构的敏感性，并且可以有效地捕捉突变残基和周围环境之间的重要关系。

2.3模型评估：

Pythia在预测相关性和计算速度方面都超过了现有的深度学习方法和其他传统方法,相比传统的基于力场的方法，Pythia实现了105倍的加速。

Pythia成功应用于预测柠檬烯环氧化物水解酶（LEH）的有效热稳定突变，并展示了其在探索2600万高质量蛋白质结构上的潜力。

参考文献

1. Nobuhiko Tokuriki and Dan S Tawfik. Stability effects of mutations and protein evolvability.Current opinion in structural biology, 19(5):596–604, 2009.
2. Angel L Pey, François Stricher, Luis Serrano, and Aurora Martinez. Predicted effects of missense mutations on native-state stability account for phenotypic outcome in phenylketonuria, a paradigm of misfolding diseases. The American Journal of Human Genetics, 81(5):1006–1024,2007.
3. **Structure-based self-supervised learning enables ultrafast prediction of stability changes upon mutation at the protein universe scale**  
   Jinyuan Sun, Tong Zhu, Yinglu Cui, Bian Wu
4. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message

passing for quantum chemistry. In *Proc. 34th International Conference on Machine*

*Learning* Vol. 70, 1263-1272 (JMLR.org, 2017).