技术交底书-工艺类

**发明人姓名：李进 联系方式： 18030866526**

|  |
| --- |
| 1. **发明创造名称：**   一种基于向量丛改进Alphafold2的蛋白质结构预测方法以及系统 |
| **2．背景技术及现有技术的缺陷和不足**：  背景技术：  从Anfinsen 实验得出结论，蛋白质一级结构包含了三维结构折叠所需要的全部信息，蛋白质的一级结构决定三维结构，三维结构决定功能，相似的一级结构具有相似的功能：相似的序列，相似的结构，相同的功能。这是应用氨基酸残基序列预测蛋白质三维结构和功能的理论基础，这也是生物技术特别是计算生物学研究的核心课题和前沿。但是目前面临不是困难，首先数据库庞大，以NCBI、PDB、Uniprot、interpro、DSEMR、BioXFinder、Pfam等蛋白库以及相关的专业数据库为代表的数据库特别庞大，数据量在TM以上，而且还有专业文献库和期刊数据库等，而且每年在快速增长。其次，目前数据库的已知蛋白质结构的数据数量不够，截止2023年才测出了20来万个蛋白质结构，而且数据本身的聚类不够，导致浪费庞大的计算资源。为了解决上述问题，建立了Blast模型，运用马尔科夫链的原理开发的算法不能解决上述问题。特别是近年来大模型兴起之后，AI算法深入应用到这个领域，国内出现了一些代表性的模型比如alphaFold2、ESMFold等算法模型，虽然取得了一些进展，但是还是不够，需要不断地完善和改进。  全球科学家目前改进方向两个：  一个是几何改进，几个改进主要是推断蛋白质结构的空间约束，比如残基之间的接触、残基的距离、残基的方向、残基氢键等。这就要把物理和化学的基础理论应用进来，优化的方法比如基于蒙特卡洛和梯度下降的折叠，对较低的能量构象进行采样。代表性的工具包括：trRosetta，trRosettaX，trRosettaX2，D-I-TASSER，D-QUARK、GDFold等。该方法主要有三部分组成：MSA encoder， coevolution aggregator，distance estimator。ProFOLD将CopulaNet估计的距离传到势能函数中，发现了具有最低势能的三级结构构象。虽然上述模型预测的残基间的空间约束(接触/距离)发挥了关键作用，许多高性能方法可以消除这些约束，但是考虑的还不周全。  二是端到端的结构预测改进，应用不同的测量扩展数据库用来多序列比对MSA信息，更加有效的使用模板，有的改进结构生成多个模型，或者做组合。比如Tencent的AI实验室开发的tFold，通过基于多MSA集成的方法，充分利用多个蛋白质序列数据库中的共演化信息，取得了最佳的性能。但是还是有欠缺。  现有技术的缺陷和不足：  1.alphaFold2的思路是基于氨基酸序列在蛋白质库中找到同源MSA蛋白质和结构，代入改造的Transfer，来预测预测三维结构，其缺陷首先是没有考虑三维结构的四级结构分类，总体描述三维结构，其次没有考虑构成四级结构的各个层级的化学键参数以及R基的各个性能基团对四级结构折叠成型的影响，导致其对比已有蛋白质库的的蛋白质准确度高，但是对于库中没有的未知的结构准确率低。其根本缺陷是把从氨基酸信息到蛋白质的四级结构以及功能看成了线性空间，其实是流形空间，远比线性空间复杂。  2.ESMFold的思路是把氨基酸序列用氨基酸缩写的字母替代，用蛋白质语言的方式，这样更是丢掉了更多的基础信息，不找MSA同源蛋白序列和结构，代入改造的Transformer，在MGnify90蛋白质数据库有6.17亿个蛋白质序列中训练模型，用于UniRef50 来验证，其减少结构和相似度的后果就是提高了速度，牺牲了准确率。  3.在alphaFold2基础上改进的两个方向，缺乏用统一的数学工具来考虑根本的问题，要通过更加适合的数学理论工具表示氨基酸、蛋白质的信息，不要停留在线性空间。  针对前面现有技术的所有缺点，本发明采用基于向量丛改进Alphafold2的蛋白质结构方法以及系统解决上述问题。 |
| **3．具体的技术方案描述：**  本部分所提供的内容涉及专利申请文件中最重要的部分，越详细越好。  基于基于向量丛改进Alphafold2的蛋白质结构预测系统，要引进新的数学工具，把生物体的氨基酸、蛋白质的结构功能所有信息放在流形空间里，真实的生物世界不是线性化的空间能够完全表示的，同时兼顾计算机只能计算线数字0 和1的数据，还是需要把流形空间局部同胚映射到欧式空间。所以，需要找到流形空间的同胚于欧式空间的理论工具，在流形空间可以用向量丛和流形张量理论刚好符合上述的要求，所以用来处理氨基酸、蛋白质的问题。  根据Anfinsen 实验结论，蛋白质一级结构包含了三维结构折叠所需要的全部信息，把20种氨基酸的构建一个流形基空间M，按照mRNA模板逆映射到向量丛E1（表示蛋白质的一级结构），再由向量丛E1光滑映射到向量丛E2（表示蛋白质的二级结构和超二级机构），再由向量丛E2光滑映射到向量丛E3（表示蛋白质的三级结构），再由向量丛E1、E3光滑映射到向量丛E4（表示蛋白质的四级结构），向量丛Ei（i=1、2、3、4）同胚于线性空间Rn。（说明：此处无参考文献）  采用基于PDB数据库中20万条有氨基酸序列、结构功能等信息的同源预测方法，这种方法是目前准确率较高的技术路线。Anfinsen 实验得出结论：蛋白质一级结构包含了三维结构折叠所需要的全部信息，这是蛋白质三维结构预测的理论基础。本发明包括应用CT-AI聚类和特征提取工具把蛋白质库进行分类建立更小的序列、结构、理化性质等特征数据集，然后输入模型得到预测的结构和功能，具体步骤如下：  步骤1：构建数学模型    步骤2：数据库聚类特征提取并建立子特征数据集  步骤3：得到MSA、结构、理化性质多个模板张量  步骤4：Ctford模型的编码  步骤5：Ctford模型的解码  步骤6：转换成三维结构并循环优化  图1：Ctfold蛋白质结构功能预测方法以及系统流程图  步骤一：构建数学模型。  通前文的思路，把20种氨基酸的构建一个流形基空间M，按照mRNA模板逆映射到向量丛E1（表示蛋白质的一级结构），再由向量丛E1光滑映射到向量丛E2（表示蛋白质的二级结构和超二级机构），再由向量丛E2光滑映射到向量丛E3（表示蛋白质的三级结构），再由向量丛E1、E3光滑映射到向量丛E4（表示蛋白质的四级结构），向量丛Ei（i=1、2、3、4）同胚于线性空间Rn。  步骤二：数据库聚类特征提取并建立子特征数据集。  收集NCBI、DSSP蛋白质二级结构参数库、PDB蛋白质三维结构库、SWISS-3DIMAGE三维结构库、FSSP已知空间结构的蛋白质家族库、Pfam蛋白质家族和结构域库、SCOP蛋白质分类数据库、SWISS-PROT蛋白质序列数据库、PIR蛋白质序列数据库、PROSITE蛋白质功能位点库、SWISS-MODEL从序列模建结构、Uniprot、interpro、DSEMR、BioXFinder、Pfam等蛋白库。上述数据库的格式和存储信息解读得到子特征数据集：同源氨基酸序列数据集、四级结构相似数据集、理化性质相似数据集。    步骤三：得到MSA、结构、理化性质模板张量。  把步骤一的数学模型用来表示在上述数据库的基础上按照同源序列、四级结构、氨基酸的理化性质进行聚类和特征提取并得到各个子特征数据集中应用预训练模型，分别得到同源序列张量、四级结构张量、理化性质张量作为模板。  步骤四：Ctford模型的编码。  把上述的多个张量作为输入编码进行同源比对、结构比对、理化性质比对学习，编码器循环调整上述的多个张量参数，学习调整后得到新的多个张量作为解码的输入。  步骤四五：Ctford模型的解码。  把上述的多个张量作为输入解码进行同源调整、结构调整、理化性质调整学习，解码器循环调整上述的多个张量参数，调整后得到新的多个张量作为下一步映射三维线性空间的的输入。  步骤六：转换成三维结构并进行循环优化。  把上述输入映射转换成蛋白质的三维结构，根据损失函数优化，并循环多次调整，输出端到端的三维结构。 |
| **4．本发明创造的优点：**  针对上述“现有技术的缺陷和不足”，本发明有如下优点：  一、本发明方法通过基于向量丛改进Alphafold2，吸收了目前全球的几何优化和同源理化两个方向优化的优点。  二、本发明采用预聚类处理数据集，减少算力节约成本，兼顾了准率和算力速度。  三、本发明引进了新的数学工具-向量丛，是一种创新。 |
| **5．具体实施方式及附图：**  步骤一：构建数学模型，具体模型的说明示意如图2所示。  1710041226504  图2 数学原理示意图  步骤1.1：根据氨基酸、和蛋白质的表示的要求选择流形空间的向量丛工具，M流形空间是底空间，作为向量丛空间Ei的基底，向量丛可以做如下运算：   1. 直和运算 2. 数量乘法运算 3. 存在对偶空间 4. 张量积运算   满足上述运算规则就可以做內积点乘，可以直接用于多种模型作为输入。  定义流形空间的局部坐标卡₯（p,Ψ）,映射到Rn空间通过一般线性群变换GL（n），  其中，P为₯的任意一点，Φ3Φ2Φ1π-1（P）如上图所示的丛空间映射，最后通过一般线性群变换到欧式空间。  步骤1.2：流形上的张量是一个偏线性函数θ：TP\*M \*……\*TP\*M X TPM \*……\* TPM  s个  r个  为点流形点P处的（r,s）型张量。  其中，TP\*M表示流形点P处的对偶张量，TPM表示流形点P处的张量。  张量也满足上述步骤1.1的运算规则，可用于后续的数据分析和处理。  步骤二：数据库聚类特征提取并建立子特征数据集。  步骤2.1：收集下载本领域的科研文献数据、发明专利数据等并建立文献数据库，部署在本地服务器。  收集NCBI、DSSP蛋白质二级结构参数库、PDB蛋白质三维结构库、SWISS-3DIMAGE三维结构库、FSSP已知空间结构的蛋白质家族库、Pfam蛋白质家族和结构域库、SCOP蛋白质分类数据库、SWISS-PROT蛋白质序列数据库、PIR蛋白质序列数据库、PROSITE蛋白质功能位点库、SWISS-MODEL从序列模建结构库、Uniprot、interpro、DSEMR、BioXFinder等蛋白库。  步骤2.2：根据需要用数据库工具和特征提取工具模型把数据分类为，以蛋白质PDB空间数据库和蛋白质序列数据库Uniprot；在Uniprot找到同源序列、然后再PDB数据库找到同源序列后找到对应的四级结构、理化性质、功能等子特征库。  因上述数据库的格式差异，数据库的详细信息解读得到子特征数据集，同源氨基酸序列数据集、四级结构相似数据集、理化性质相似数据集。  特别是结构数据集，按照一级肽链的序列，对比同源数据，为了提高精确和范围，需要用比对模型在上述的几个库中建立多个同源的MSA数据集库，比如按照同源性相似30%、50%、70%、90%分为几个子库。二级结构的结构同样的按照8种二级结构的分类提取同源的数据集和结构库以及理化性质的库。介于二级结构和三级结构之间的超二级结构一样的处理方式。三级结构和四级结构类似的处理方式。  步骤三：得到MSA、结构、理化性质模板张量。  步骤2.1：基于流形空间的向量丛和张量理论，用流形M空间基底空间的三组基：氨基酸序列基、氨基酸序列结构基、氨基酸各个基团的理化性质基的向量丛基底，用来表示表示Ei向量丛的同源氨基酸张量、第i级三维结构张量、第i级理化性质张量。  步骤2.2：在上述模型建立后，把第二步建立的各个特征数据集代入数学模型预训练得到同源模板张量、对应第i级三维结构模板张量、对应的第i级理化性质模板张量。    步骤四：Ctfold模型的编码。  步骤3.1：Ctfold模型的总架构如下图3所示。    图3 Ctfold模型总架构示意图    步骤4.1：步骤三应用库预训练了多个模板数据集，输入氨基酸序列，跟前面步骤的数据库建立的同源数据特征子数据集、理化特征子数据集、蛋白质结构特征子数据集，输出模型变为是个对应的四级结构的四个张量。  步骤4.2：模型核心的编码部分的核心就是改进的Transformer，因为数学模型流形空间的向量丛和张量直接满足四个运算而且是同胚于欧式空间，所有就不采用AF2的分为行和列分别跑四个注意力的方式，而是分别在模型的用四层的Multi-attention跑四个张量的，套用60层block，输出三个四级张量作为下一层解码的输入。  步骤五：Ctfold模型的解码。  步骤5.1：把步骤4的输出的张量作为输入，分别更新四级结构的结构，和按照理化性质作为约束条件，比如各个层级的势能最小、各个层级的化学键的稳定性最好，更新一级结构的主链骨架、按照二级结构8种构成方式的氢键稳定性更新结构、按照三级结构的6种化学键稳定更新等，经过16个block的优化循环更新参数。  步骤5.2：输出的更新的流形空间张量作为输出。  步骤六：转换成三维结构并进行循环优化。  步骤6.1：把上一步输出的流形空间的张量投影到三维空间结构数据，以及相关功能的数据。  步骤6.2：计算损失函数，求最小损失函数的作为优化参数，并进行不断地循环，设计循环5轮，再次从编码解码的过程。 |