高分子链在膜结构中势能面和位置分布的预测

刘闵 {191240030@smail.nju.edu.cn}

ABSTRACT

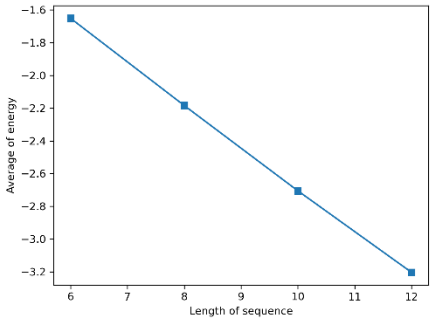
生物膜由磷脂分子组成，磷脂分子具有一个亲水头部和疏水尾部，组成生物膜时亲水、疏水部分会分别相互靠近，因此会形成双分子膜和球状膜两种结构。高分子链可以粗粒化为由亲水单元和疏水单元组成的长链，因此在生物膜的不同区域，高分子链具有不同的自由能 F 和内能 U。能量越高，高分子链处在该区域的概率就越低，反之亦然。本文借助不同长度高分子链的实验数据并结合神经网络方法，预测高分子链在生物膜不同区域的能量及概率。本文提出的神经网络，基于对数据一些简单且合理的假设，简化了问题，并在数据集上取得了良好的预测效果。

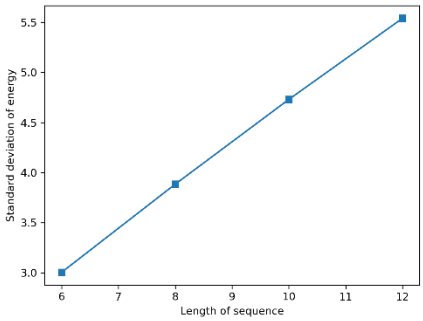
1. Experiments

1.1 Observations

由于数据中的能量和概率有很强的物理意义，首先对数据特征的分析。

对于能量的预测，数据集中提供的信息仅包括序列的长度、序列中疏水和亲水结构单元的分布情况、Z-distance。分析能量数据的均值和标准差，可以发现能量和序列的长度间存在很强的线性关系。观察亲水单元占比相同的序列，可发现其能量分布很相似，为简化问题，在能量预测的任务中可以忽略序列的排布顺序，仅考虑疏水和亲水单元的比例。最后，Z-distance 也是必须考虑的因素。



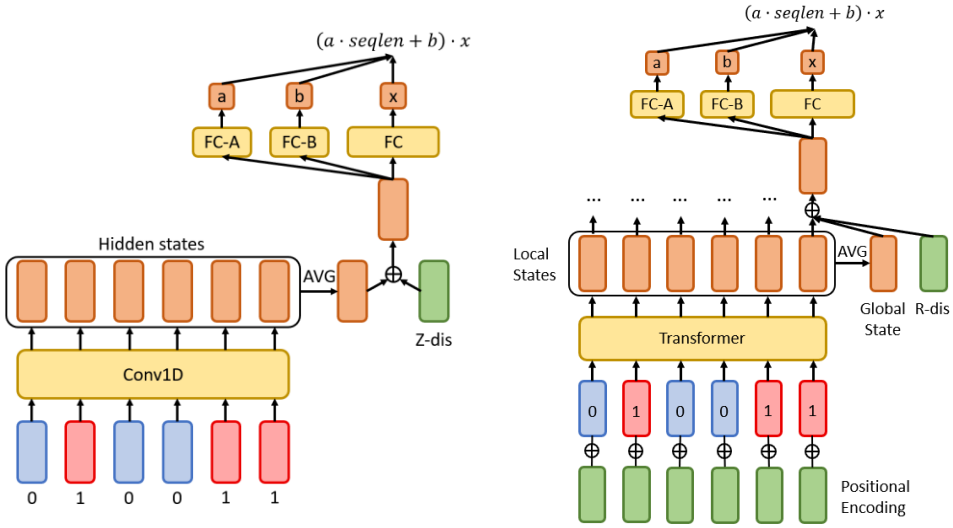


对于概率的预测，我们需要得到每个结构单元在不同 R 上的概率，并计算确定的 R 上结构单元概率的均值。观察概率数据的均值和标准差，也能发现与序列长度间存在一定的线性关系。与能量预测不同的是，概率的预测需要对每个结构单元分别进行，且由于相邻单元应出现在相近的位置，每个单元的预测结果会受到周围单元的影响，在预测过程中我们需要捕捉序列中不同单元间的相互影响。

1.2 Model Structure

本文提出两个模型，分别用于能量的预测和分布的预测。

如下左图为用于预测能量的模型。输入序列首先通过一个维度为 d_model 的 embedding 转化为向量，接着对每个向量分别使用 1 维的卷积网络，得到维数为 d_model 的 Hidden States。由于模型不需要考虑序列中结构顺序的影响，只考虑 0、1 的占比情况，模型将 Hidden States 取平均值，然后与 Z-distance 得到的 embedding 相加。接着，对相加后的相连通过三个全连接层、BatchNorm 以及 ReLU 组合得到的多层网络，分别得到维数为 1 的 a、b、x。通过计算$a \cdot seqlen + b$得到预测值 x 的缩放因子，用于考虑序列长度对能量的影响。



右图用于每个结构单元在球状膜各个区域的概率预测。模型在输入部分同样采用了 embedding 处理序列，并为各个结构单元加上了维数为 d_model 的 Positional Encoding，用于在 Transformer 中捕捉到序列的顺序。要注意的是，由于 Positional Encoding 最大长度为 20，但训练数据中最大为 12，训练过程中需要在[0, 20-seqlen]中随机选取一个位置作为 Positional Encoding 的起始点。随后用 Transformer 处理输入，与 Conv1D 不同的是，Transformer 会通过 attention 机制捕捉到序列中结构单元之间的相互影响。Transformer 得到的 Hidden States 为每个结构单元的 Local State，为考虑整个序列的影响，对 Hidden States 计算均值得到了一个 Global State。为计算每个结构单元出现在位置 R 的概率，需要将 Global State、对应的 Local State 以及 R-distance 得到的 embedding 相加，得到一个 d_model 维的向量，然后和左图的模型相同，通过三个多层网络计算 a, b 和 x，最后计算出该计算结构单元在 R 处的概率。

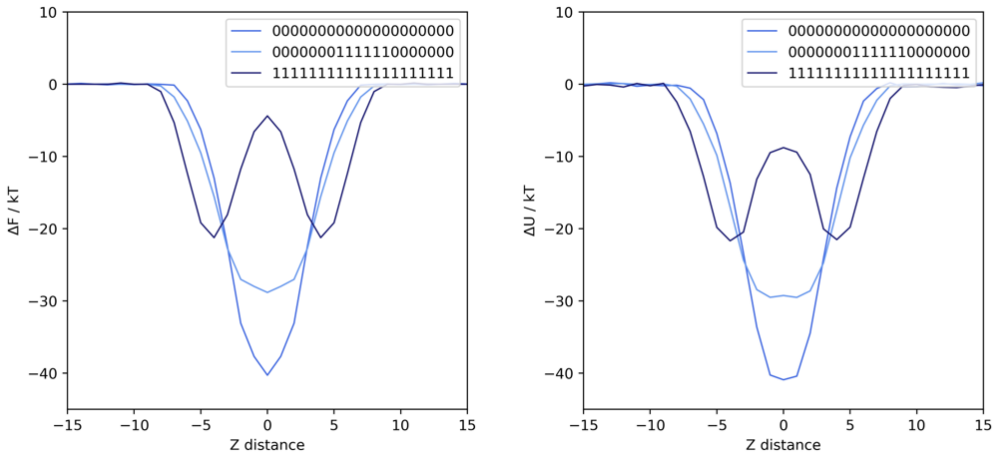
1.3 Training

对每个模型，实验中设置 batch size 为 128，d_model 为 16。学习率通过指数衰减的方式从 1e-2 降到 1e-4，共训练 100 轮。优化器采用 Adam，损失函数为

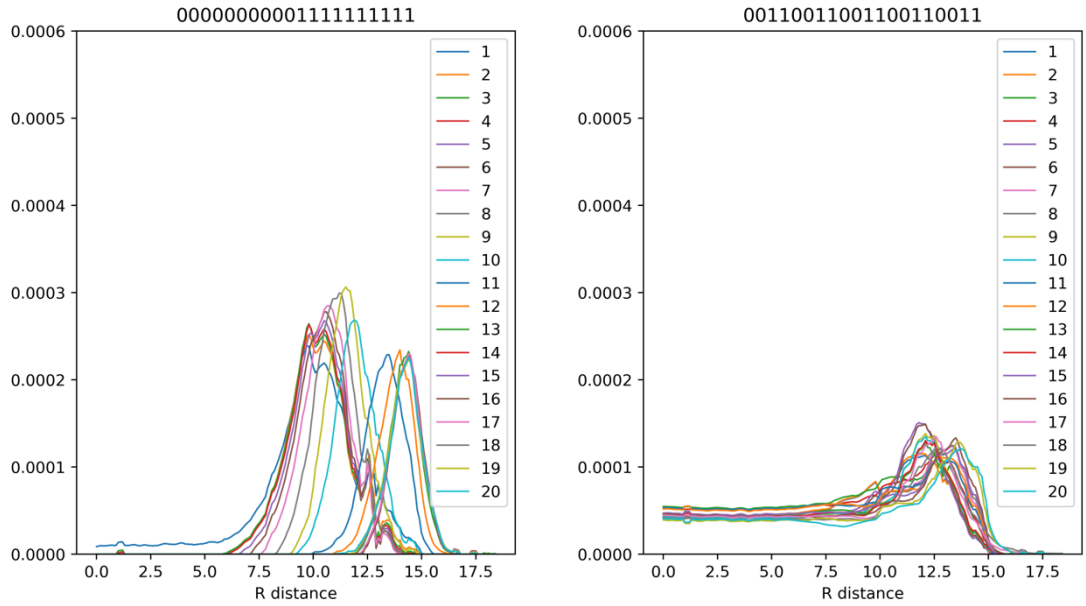
MSELoss。在调参过程中，使用长度为 6、8、10 的数据作为训练集，长度为 12 的数据作为验证集。确定参数后，采用长度为 6、8、10、12 的数据作为训练集，在长度为 20 的序列上进行预测，不再设验证集。

1.4 Experiment Results

在使用长度为 12 的序列作为验证集时，验证集的 loss 已经和训练集的 loss 在同一个水平。下图为预测长度为 20 的序列得到的自由能与内能。



预测每个结构单元在 R 处概率的难度要大不少，但是模型也取得了很好的表现。观察下图的结果发现，预测结果存在噪声，曲线不够平滑，但是反映了概率分布的规律，并且基本满足相邻的结构单元概率分布相近的要求。



2. Discussion

本文提出了两个模型，分别用于预测高分子链在膜结构中势能面和位置分布，并在数据集上取得了很好的表现。此外，本文提出的模型较为轻量化，训练效率较高，并可以进一步扩大模型规模以取得更好的表现。针对可能出现的预测噪声较大的问题，可以结合预测结果的物理意义，对数据进行平滑化的处理。

本文的实验过程中比较有启发的发现是，神经网络的结构设计过程中，将数据的特性纳入考虑会比较有帮助。设计预测能量的模型时，可以发现能量与序列长度呈现出一个明显的线性关系。起初的方案是通过将序列长度作为一个参数加入输入中，让模型自己学到长度对结果的影响，但是这种方法效果并不好。后来通过预测 a 和 b，用以计算不同序列长度下得缩放因子，可以大幅提高预测的准确度。这种做法本质上是将领域知识融入到模型的设计中，调整模型的归纳偏置。尽管近年来很多深度学习的研究旨在找到更为通用的模型结构和学习方法，很多时候我们还是可以发现针对某项任务设计的精巧结构能够比更一般的方法取得更好的表现。