·综述·

深度学习在蛋白质结构预测中的应用及启示*

王天尧 李剑锋**

(复旦大学高分子科学系 聚合物分子工程国家重点实验室 上海 200433)

摘 要 蛋白质结构预测通常指借助计算机计算模拟方法从氨基酸序列推断其三维空间结构. 而空间结构决定其生理功能，故结构预测问题尤为重要. 基于单纯物理学的预测仅能应对较短蛋白质且精度不高. 而基于数据驱动和生物信息学的方法近十多年备受重视. 本文主要回顾近十多年来深度学习在蛋白质预测领域的应用，重点介绍Deepmind团队的AlphaFold方法，此方法预测在单域蛋白质达到了中低分辨率实验精度，一定程度上解决了困扰人们五十多年的蛋白质结构预测难题.

关键词 蛋白质折叠，深度学习，神经网络，结构预测

天然蛋白质通过调节一维氨基酸序列信息，能够精准地制备具有特殊的三维空间结构的蛋白质分子，实现特定的生理功能. 而蛋白质结构预测希望代替大自然通过各种方法从一维序列信息推断其三维空间结构. 蛋白质结构预测问题提出至今已困扰我们五十多年[1~3].

自然条件下，蛋白质总能在生物学相关时间尺度内迅速而准确地折叠到有限的几种(大多数情形仅一种)三维空间结构[1~5]. 这是一种在分子的随机热运动下蛋白质大分子发生构象变化折叠到自由能较低的结构，而这种稳定的空间结构被称作为蛋白质天然状态(native state).

通常认为蛋白质折叠的驱动力包括以下几种[2~5]：氢键作用、分子间的范德华相互作用、残基骨架扭转角的选择性、静电作用、非极性基团的厌水相互作用和构象熵. 上述驱动力可被统一地描述为“力场”或势能函数. 此势能函数也被称为蛋白质折叠能量全景图(protein-folding energy landscape). 而统计热力学研究表明，此全景图呈漏斗形[2,6~10]. 大部分未折叠构象形成了高能量地势较缓的平原；而少数折叠构象形成能量低且地势陡峭的漏斗底部.

Anfinsen热力学假设[3]提出：折叠结构信息蕴含于能量景观地形中，且天然态对应于自由能全局最小值. 基于此假设的算法构成了计算模拟利用势能函数进行蛋白质折叠预测的基础. 现实中蛋白质构象能量景观是复杂高维曲面，存在大量局部极小值，以前人们曾认为这些极小值会使得最终折叠成天然状态所需时间远长于目前观测时长.

因而产生了著名的Levinthal佯谬[8,11]. 一方面，若假定蛋白质在各个构象停留时长相等，则会发现其通过随机搜索方式折叠到天然态所需时间会随序列长度指数增长；而另一方面，生命体系中蛋白质总能非常快地找到能量最低的天然态. 因此，存在矛盾. 事实上，人们发现蛋白质会先近程地折叠成若干稳定的二级结构，然后再进一步折叠成全局结构，此分而治之(Divide and Conquer)的方法极大地缩短了搜索时长[2]. 另外在解决Levinthal佯谬过程中，简化的HP蛋白质格子模型起到了重要的作用[2,6,8,9].

理解了快速折叠的原理不代表解决了蛋白质预测问题.

在传统的蛋白质折叠预测中，人们通常经过构造或选择力场，从某非天然态出发，用各种动

* 材料基因组研究专题；2021-12-28收稿，2022-01-28录用，2022-03-17网络出版；国家自然科学基金(基金号21973018, 21534002)资助项目.

** 通讯联系人，E-mail: lijf@fudan.edu.cn

doi: 10.11777/j.issn1000-3304.2021.21401

力学计算或模拟方法(例如分子动力学模拟)演化其构象，直至能量达到全局最小[1~3]. 但传统预测方法会随着残基数目增加计算量迅速上升，事实上传统方法对大多蛋白质结构预测都无能为力[12].

此困境一度让蛋白质折叠预测领域的人们绝望. 因此，人们不再依赖基于纯粹物理机制的方法，而是采用结合数据驱动的方式[13,14]. 最近十多年，这种结合数据驱动的方法随着深度学习在2012年的兴起而愈受重视. 直至近3年，AlphaFold[12,15]的突然崛起，特别是AlphaFold 2预测蛋白质的高准确性甚至让许多人相信蛋白质折叠预测难题将被解决[14].

本文主要给非生命科学领域读者介绍深度学习方法在蛋白质结构预测领域的应用. 将选讲几个主要进展，特别将重点介绍AlphaFold[12,15]. 根据受众特点，本文将在下一节列举蛋白质结构预测的必要知识. 然后，介绍一些深度学习相关的知识. 紧接着介绍几种主要的预测方法，最后介绍AlphaFold[12,15]的基本思路，以及本文作者在此方向的贡献[16].

1 蛋白质结构预测的基础知识

1.1 位置特异性打分矩阵PSSM

位置特异性打分矩阵(position-specific scoring matrix, PSSM)或位置权重矩阵(position weight matrix, PWM)[17]是蛋白质及生物信息学里非常重要的统计量. 它主要衡量了不同氨基酸(或核酸)在蛋白质(或DNA)上某个特定序列位置上出现的概率. 在一些机器学习预测蛋白质的二级结构类型时[18~25]，常会将PSSM作为网络的输入. 但注意PSSM只包含残基绝对位置属性的信息，不包含不同残基配对关联信息.

图1(a)以DNA为例，给出了统计PSSM矩阵的示意流程. 首先给定一个序列库(例如针对基因库的所有DNA数据，或蛋白质库里所有可能的序列)，图中给出了由10个假想DNA序列组成的DNA库；然后统计不同的核酸在特定位置出现的频次矩阵(position frequency matrix, PFM)；再根据PFM得到位置概率矩阵(position probability matrix, PPM)；最后根据图中公式算出位置权重矩阵PWM.

1.2 多重序列比对MSA

目前大多蛋白质结构预测的深度学习算法的输入中都有多重序列比对信息(multiple sequence alignment, MSA)[12,15,26~39].

序列比对(sequence alignment)主要任务是针对查询序列(query sequence)从数据库中，用基因信息学的方法找到进化树上尽可能同源的序列，然后根据变异的氨基酸的相似程度，按照特定规

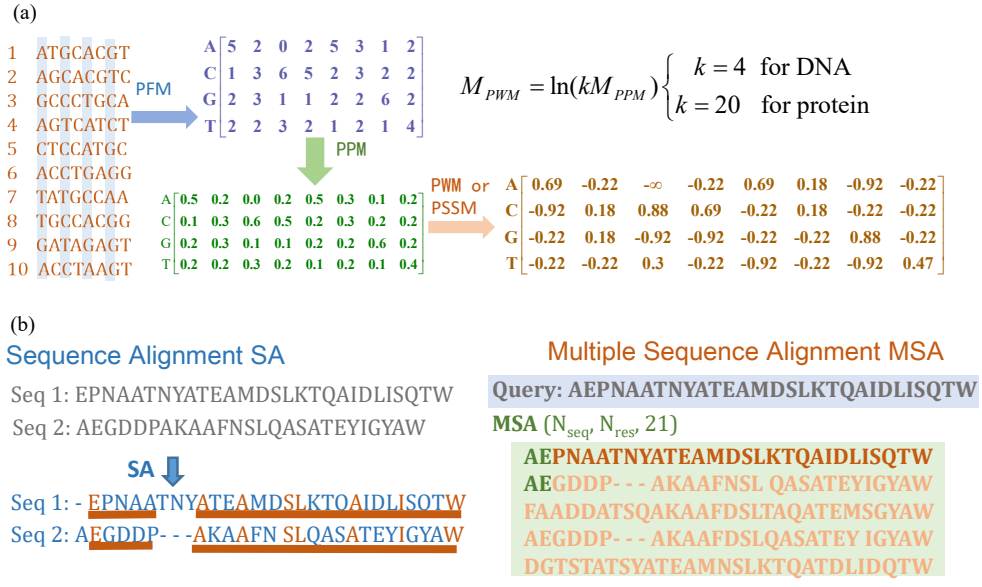


Fig. 1 Illustration of (a) position-specific scoring matrix (PSSM) and (b) multiple sequence alignment (MSA). (a) In this illustrative example, PSSM is computed using the formula given in the top right corner based on a DNA database consisting of ten DNA sequences. (b) Sequence alignment (SA) is trying to match the fragment pairs from the two given sequences as much as possible. In the alignment, inserting gaps "-" is allowed. Multiple sequence alignment (MSA) is SA on multiple sequences.

则来给该序列与查询序列的相似度打分.

某个序列的变异包括对序列中特定片段的插入、删除和替换. 相对于查询序列，当库里的蛋白质序列变异很少时，则两者相似度高.

当变异多时，还需根据进化同源的特点分类对变异片段进行进一步分析. 变异的氨基酸片段可分为保守片段(功能及化学特性相同)、半保守片段(功能及化学特性相近)和非保守片段(化学特性相差甚远). 显然，若保守片段越多，表明与查询序列越接近.

比对的目标是通过恰当地插入空片段(gap)，使得插入空片段后的2个序列尽量相似(如图1(b)左图所示). 比对的方法有许多[27]，例如动态规划(dynamic programming) 和点阵法 (dot-matrix method).

用上述比对方法对若干个给定的序列与查询序列进行比对就称为多重序列比对(multiple sequence alignment, MSA). 通常可用软件ClustalW, MAFFT, ClustalOmega以及MUSCLE等算法程序对多个序列进行MSA比对[40~52].

而在蛋白质预测中，通常会针对输入的蛋白质序列，从蛋白质数据库中找到与给定序列相近的若干个序列，然后再将这些MSA作为神经网络的输入. 此信息相比于PSSM包含了更为丰富的信息. 可从MSA中看出目标序列大致从哪些序列变异而来. 在深度学习中，MSA数据维度为(N_{seq}, N_{res}, 21)，其中N_{seq}为MSA包含序列的数目，N_{res}为目标序列的长度，21用于分辨20种氨基酸和gap“-”的热点表征(有时可能为22或23).

1.3　接触图与距离图

如图2所示，图2(b)与2(c)是一个HP蛋白质模型结构[39]的接触图(Contact Map)与距离图(Distogram). 其中接触图中只有2个残基接触时，才有值(黑)；而Distogram灰度值对应于两残基的距离，当距离大于截断阈值时，灰度为0(白色).

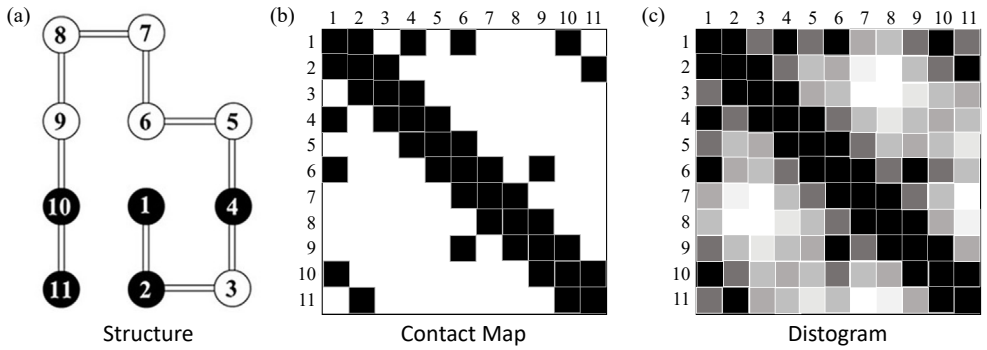


Fig. 2 Illustration of contact map and distogram. (a) A typical structure of a given HP protein. (b) The contact of the (c) structure where the black square indicates the matrix element corresponding to two contact residues. (c) The distogram of the (a) structure where the greyness indicates the distance between two residues.

同一序列中不同残基间的接触与否或距离是非常重要的信息，它基本蕴含了蛋白质骨架的三维结构所有的信息. 而且这个信息相比于纯粹的结构三维坐标信息有2个优势：(1) 具有旋转平移不变性，而三维坐标会随着蛋白质的旋转或平移而改变；(2) 表达更简洁及更易标准化. 因为存在关联变异(correlated mutation)现象，有些接触的两氨基酸会同时变异以保证变异后仍接触，故接触图或距离图信息就显得相当重要[35,36,39].

基于上述原因，在最近的深度学习预测蛋白质结构的实践中[12,15,53~62]，大多都会采用此信息去提高预测准确性或预测给定蛋白质的Contact Map或Distogram.

1.4　蛋白质数据库PDB

目前最著名的蛋白质数据库为PDB[63],即Protein Data Bank，收藏了约1×10^5多条蛋白质的三维结构数据. 这些结构由X射线、NMR或电子显微镜等方法获得.

1.5　CASP竞赛

Critical Assessment of Protein Structure Prediction (CASP)[64,65]是蛋白质结构预测科学共同体举办的两年一次的竞赛，每次竞赛优胜者的水平基本代表了当前世界结构预测的最高水准(benchmark progress). 在每次竞赛中，举办方会

给出若干个已知结构但未曾公开的蛋白质序列，参赛团队在规定时间内提交各自的结构预测结果，同时不限制预测方法. CASP自1994年以来共举办了14届，其中最近2届的第一名皆来自deepmind的AlphaFold算法.

1.6 模版建模得分TM Score

之前，人们通常用距离均方差root mean squared deviation (RMSD)衡量2个分子构象的接近程度. 但现在模版建模得分template modelling score被认为是更准确的衡量方式[66]. 其表达式如下：

$$\mathrm{TM}\left(\{\mathbf{r}\},\{\mathbf{r}^{\mathrm{true}}\}\right)=\max_{\mathrm{all}\,M}\frac{1}{n}\sum_{i=1}^{n}f(\left\|\mathbf{r}_i-M\mathbf{r}_i^{\mathrm{true}}\right\|)$$

其中

$$f(d)=\frac{1}{1+\left(\dfrac{d}{d_0(n)}\right)^2}$$

$$d_0(n)\approx 1.24\sqrt[3]{n-15}-1.8$$

式中n为蛋白质的残基数，M为旋转平移矩阵. 上式表达的含义是将预测得到的结构与各种旋转平移操作后的真实结构进行比较，取最相近(极大)的那个作为最后的分值.

显然TM score在0~1之间，分数越高表明越准确. 通常认为当TM>0.5时，预测与真实之间的折叠基本一致[56]；而对同一蛋白质，NMR与X射线测出结构之间的TM分数为0.807 ± 0.107左右. 所以，可认为当TM分数>0.8时，预测的结果已经完全正确.

而AlphaFold2(AF2)近2/3的预测结果达到中低分辨率的实验精度[12]. 也即AF2几乎解决了单域蛋白质折叠预测问题[14].

由于多域蛋白质各功能域之间可以相对独立地移动旋转，在评估多域蛋白质结构相似性上，局域距离差异性测试(local distance difference test)是一个比TM分数更佳的评分方式. lDDT不同于TM，不依赖于骨架α碳原子的重叠，能够不受功能域间位移的影响，更加有效地评估结构之间的局域相似性[67].

1.7 深度学习原理与常用神经网络模型

本小节仅罗列结构预测涉及到的深度学习技术及原理, 具体请参考相关文献[68].

神经网络(neural network)可抽象成一个函数$y=f(x;w)$，它关联了2组信息数据x与y(比如蛋白质的序列x及其结构y)，分别称为网络输入与输出；w为网络的参数. 神经网络训练的目标是为了找到恰当的w使得网络能够根据x准确地预测y.

普适近似原理(universal approximation theorem)[69]表明单隐藏层的神经网络，只要其激活函数为非线性且神经元数目足够多，便可无限精确近似任意非线性映射. 普适近似原理表明NN可用于拟合任意未知关联.

神经网络设计要点：考察待预测的量y与哪些量有关联，即找出哪些信息可足够推导出y，然后将这些信息与y之间架接合适的神经网络便可. 信息间的关联如果能用现有知识进行关联就用现有知识将其关联；未知关联用神经网络代替.

神经网络选择需要考虑输入输出信息数据特点，目前结构预测中常用的网络结构主要有下面几种.

残差网络(resnet)[70]的基本思想是不断地将未处理过的信息直接复制并叠加到下面几层由网络抽取出的特征上去. 残差网络于2015年提出，后来被广泛运用于图像处理中.

基于自注意力机制的transformer[28]近几年备受人工智能领域喜爱，它几乎完全取代循环神经网络[68]，其基本思想是从不同位置对之间提取信息，适合处理文本类、时序性的信息，不过近年也常用于图像处理. AlphaFold2[12]中大量使用了自注意力机制.

2 传统蛋白质结构预测

传统的蛋白质结构预测方法[14]主要基于以下2种模型：基于模板的方法(template-based method, TBM)[29~38]和无模板方法(template-free method, TFM)[72~77]；当然，有些方法介于这2种方法之间. 通常全局模板是指直接从PDB数据库[63]获取的实验测定的蛋白质三维(骨架)结构，而无模板方法是指没有采用全局模板的方法.

2.1 基于模板的方法TBM

TBM方法[14,29~38]大致步骤如图3所示，通常可分为以下几步. 第一步，通过数据库检索，得到目标蛋白质的一组同源性序列(MSA)，并根据MSA获得1个或多个折叠结构模板. 第二步，比对目标序列和模板对应序列，两序列一致的片段

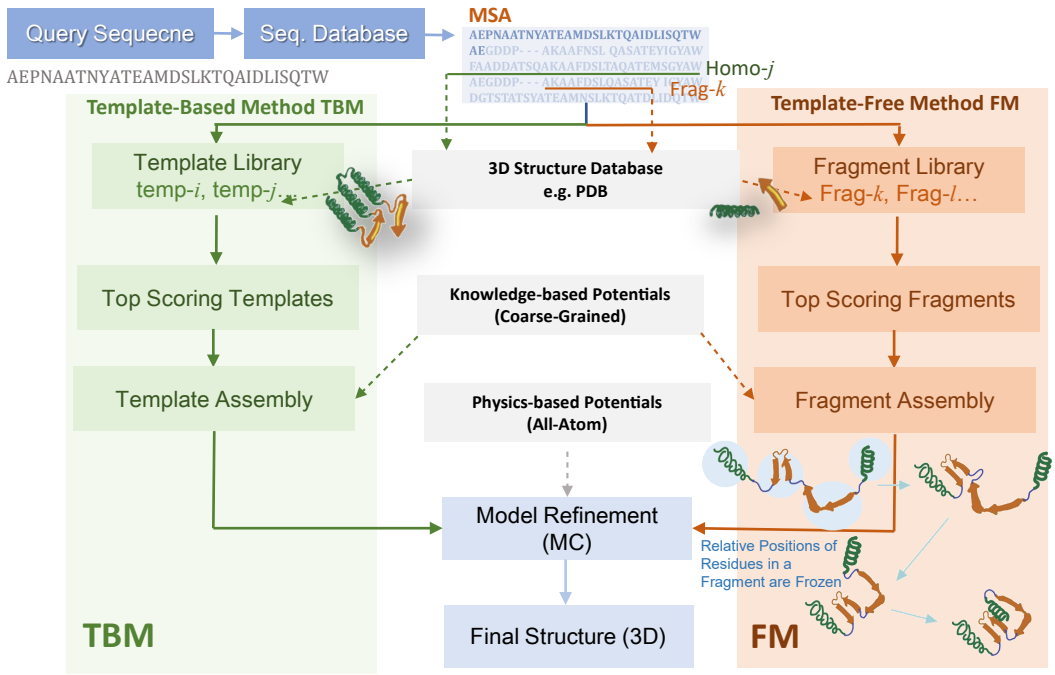


Fig. 3 Illustration of basic strategies of template-based method (TBM) and template-free method (FM).

直接使用模板的对应折叠结构. 第三步，对于目标序列与模板对应序列不一致的区域，采用碎片组装、优化算法或是数据库方法等单独预测. 当然通常最后还会用诸如分子动力学的优化方法进行模型的精细化(model refinement)，以优化全局结构[13,14]. 历史上，TBM方法[14]可以细分成comparative modelling (CM)和threading 2种方法[29~38]. 其中在CM中，模板与目标序列的同源性较近.

2.2　无模板方法FM

无模板方法[72~77]的流程见图3右半侧，FM从蛋白质数据库中依MSA比对结果找到一些片段的结构并将其放入片段库中，然后找到评分较高的片段结构拼成初始结构[72~74]，接着采用FM里非常重要的片段组装(fragment assembly)方法[72]，大致冻结片段的结构并以片段结构为单元来演化全局结构，比如可根据粗粒化的势能函数用梯度下降化进行能量优化.

3　深度学习方法

3.1　残基接触对的预测

人们发现在蛋白质变异过程中经常出现关联变异(correlated mutation)的现象：一条蛋白质链内若发生变异，总是2个氨基酸成对地变异；因为演化压力会迫使蛋白质维持一致构型，原本接触的氨基酸对在变异过程中继续保持接触，可以避免其形状发生剧烈变化. 因此，这就使得残基接触对(inter-residue contact map)的信息极为重要[56~61].

早期有许多传统方法致力于预测残基接触对. 处理该问题的早期算法，倾向于以一次一对的形式、孤立地预测每个接触对是否可能. 由于忽视了蛋白质包含的全局信息：一个残基对是否接触受到序列中其他残基的影响，早期算法陷入了困境，预测效果糟糕. 而之后研究者提出了充分利用全局信息的预测方法，例如基于Markov随机场模型MRF的direct coupling method (DCA)[58~61]，在残基接触预测上获得了突破性的成就.

深度神经网络在预测残基接触对问题上，也表现出了异常优异的性能，有时甚至还直接被用于预测键角等信息. 这些预测特征均可作为约束，辅助指导无模板方法.

比如，Raptor X-Contact深度学习模型[39]将Contact Map的预测当成图片分割任务来对待，Raptor X-Contact所采用的方法也被其他方法，如ResPRE[54]所采纳. ResPRE[54]采用了图片识别领域非常著名的残差网络(Resnet)模块[70]，残差网络的重要思想是不断地将网络前面的信息直接复制到网络后面.

而AlphaFold1[15]又将Contact Map拓展成距

离直方图(distogram)预测，基于此，它在2018年CASP13的比赛中获得了巨大成功.

3.2 AlphaFold

2020年的CASP14的比赛中，AlphaFold2 (AF2)[12]取得了骄人的成绩. 对来自89个域(domain)实验测得的蛋白质结构，AlphFold2在88个域TM分数>0.5，59个域分数>0.914. 前者意味着预测结果与答案之间折叠基本一致. NMR、X射线晶体学测出的一组112个单域蛋白质，序列相同率大于95%. NMR与X射线测出的结构之间的TM值为0.807±0.107. 这说明AlphFold2的近60%的预测达到中低分辨率的实验精度. 也就是说AlphFold2几乎解决了单域蛋白质折叠预测问题[14].

AlphaFold2深度学习模型的结构简图如图4所示，具体参考文献[12]. 它分别借助了基因同源信息和蛋白质结构数据库模板信息. 如图所示，根据同源信息，可得到序列比对信息MSA，通过同源搜索得到与输入序列同源相近的(s−1)条序列和输入序列一起放到MSA数组里，再通过线性神经网络变换得到MSA表征，此表征的维度为(s，r，c)，其中r为蛋白质的序列长度，c为表征的特征数(通道数). MSA表征包含了输入序列与其他同源序列间的关系.

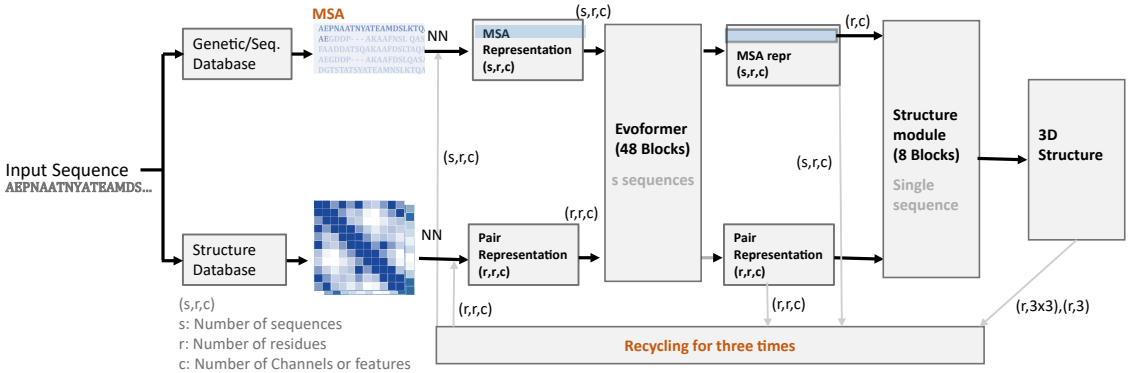


Fig. 4 Sketch of the AlphaFold2 model. Detailed description is referred to Ref.[12].

而另一输入通道中，主要输入与MSA相对应的序列的结构残基对距离信息以及扭转角的信息. 在具体输入时，AF2将距离对长度划分成64个离散块(64 bins)，并将其转化为概率的形式，故对应数组形状为(s，r，r，64)，取值为0~1. 注意配对表征中，只包含了MSA除输入序列之外的某个序列自己结构信息，不同序列之间并没有进行信息的关联.

然后再将MSA表征与配对表征输入一个称为Evoformer的模块，此模块主要将MSA的信息(同源性差异)与结构信息整合起来，最后得到输入序列的MSA表征与输入序列的配对表征. 此时，输入序列的配对表征同时将演化信息与其他模板结构信息有机地融合在了一起. Evoformer主要利用了自注意力机制来实现上述信息整合.

而下一个结构模块structure module主要的功能是将Evoformer预测的配对表征展开成三维空间结构，同时亦承担一定的预测调整功能. 此模块的结构大致如图5所示. 一条蛋白质骨架结构可想象成一系列三角形的叠加，三角形的中心相当各个残基α碳的坐标，三角形平面本身代表N-α-C-C构成的三角形. 这样，此骨架可由2个数组表示，数组形状分别为(r，3×3)和(r，3)，分别表示每个三角形取向与位置.

初始时，假设所有氨基酸都在原点，然后将此初始骨架与配对表征输入结构模块，由于配对表征存有距离对及取向信息，故可通过一个称为不变点注意力神经网络模块将其初步还原成展开的骨架结构，紧接着再加入侧链原子从而得到全原子的三维结构.

如图4所示，最后再将中间输出的MSA信息、配对信息和3D结构信息重新叠加输入到Evoformer，如此反复迭代3次，最终到预测结果.

因为PDB中只有大约1×10^5多个的序列有对应的三维结构数据. 而在big fantastic database (BFD)蛋白质序列数据有多达2,204,359,010个序列，虽然这些序列并不一定有对应的三维结构信息(无标签)，但self-distillation dataset的训练技巧可以将这些无答案的题目作为作业进行训练，自己提高预测准确度，AlphaFold2用此扩大训练集

并进一步提高了预测准确度.

后来有诸多研究团队对AlphaFold2进行了拓展与提升. 例如：Baker团队[78]的RoseTTAFold发展了三通路神经网络(three-track neural network)，对AlphaFold2只包括1D序列信息和2D距离图信息的两通路神经网络模型进行了拓展，引入了3D结构通路道网络模块；高毅勤团队的MindSpore算法[79]对AlphaFold2的计算速度进行了较大的提升.

3.3　最简单的蛋白质模型的预测

真实蛋白质结构预测无论从训练数据准备还是模型构建及训练都极其复杂. 因此，人们希望找一个简单的蛋白质模型，以便能快速地试验他们的想法. 就如手写数字识别(对应数据集为MNIST)[80]对于图像识别一样，所有的方法都会用MNIST数据集先来检验其有效性.

而HP蛋白质模型就是这样的模型[2,6,8,9]. 它仅有2类氨基酸H和P，其中H代表厌水型氨基酸，P代表亲水型.

我们基于此HP模型，提出了一个强关联神经网络[16]，如图6所示，此神经网络有2个核心要素，一是不同于传统的向量表征，它采用一个小的神经网络来代表每个氨基酸，每种氨基酸都用一个神经网络来表征，不同氨基酸对应的网络的权重亦不同，而相同的氨基酸共享网络权重；二是它有一个自洽循环通路，这样可使得输出的信息(环境)与氨基酸的属性发生强关联.

该研究发现与传统向量表征方法相比，强关联网络极大提升了预测准确性，提高了约20个百分点.

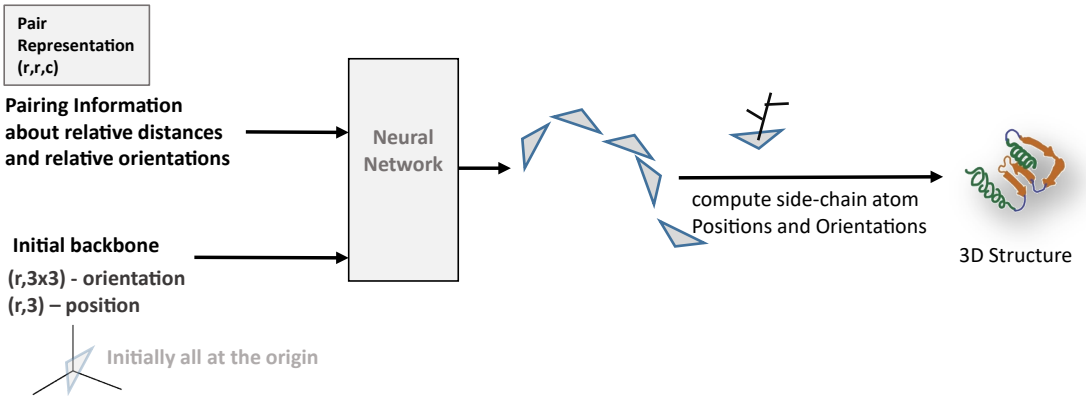


Fig. 5　Illustration of how the pairing information is transformed into the 3D structure using neural networks in AlphaFold2[12].

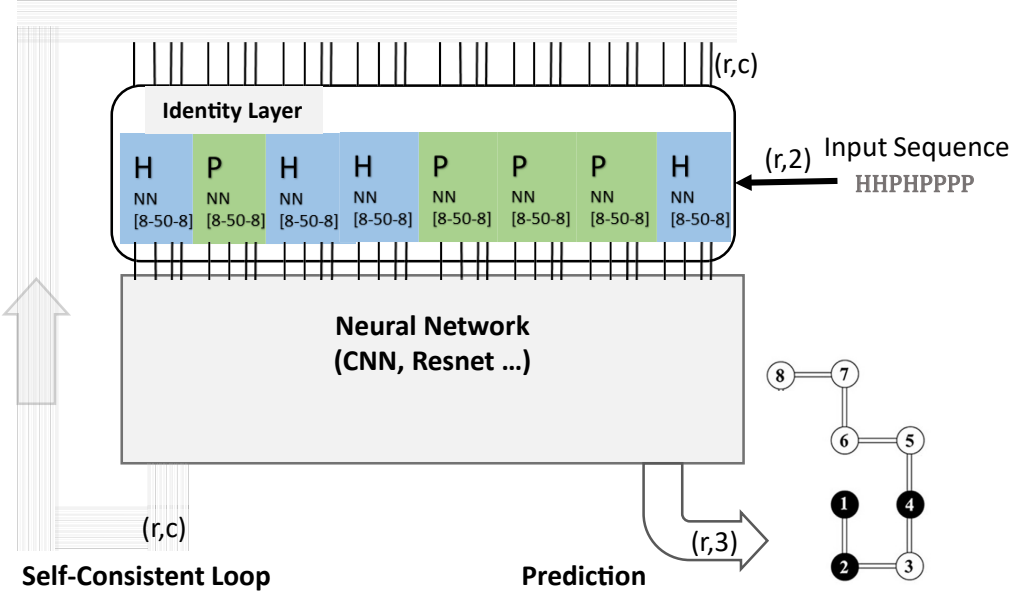


Fig. 6　Architecture of the strongly-correlated neural network (SCN) where *r* indicates number of residues and *c* indicates number of features or channels.

4 展望与启示

最近十多年深度学习在蛋白质结构预测中取得了巨大了成就，它的杰出代表AlphaFold2[12]几乎解决半个世纪前提出的蛋白质结构预测难题[12]；由于其预测结果达到了中低分辨率的实验精度，几乎等于说AlphaFold2的预测可以直接代替有些蛋白质结构分析实验，而对于通常200多个氨基酸组成的蛋白质，AlphaFold2通常在普通GPU上只需几分钟便能得到其结构，这对于以后的生物制药等领域将有巨大影响.

而另一方面，高分子材料基因组计划仍然在进行中. 因为普通高分子的组成不像蛋白质序列那样，有确定的组成单元以及较为单一明确的目标，因此难度更大. 但深度学习在蛋白质结构预测中的成功经验仍然对高分子材料基因组计划有一定的启发：

首先，它有一个标准化的结构数据库PDB. 高分子材料基因组计划或许也需要构建类似的数据库，难点在于制定统一的数据标准. 即如何准确、完整、简洁地表征高分子链，加工条件及性能.

其次，蛋白质结构预测有一个权威的CASP竞赛，CASP极大地推进了结构预测算法的演进. 在材料基因组计划中可参照CASP，建立相应的标准化竞赛.

再次，AlphaFold2充分利用了当前深度学习领域的各种先进算法，并不拘泥于某种特定算法. 这启发我们解决问题时需要以问题为导向，而非以方法为导向.

最后，AlphaFold2中将Distogram信息用神经网络转化成分子结构坐标的方法可推广至其他结构预测的问题中，当然也可用于高分子的结构预测.

作者简介：李剑锋，男，1980年生. 1999~2010年于复旦大学高分子科学系获得学士、硕士、博士学位；2007~2009年在加拿大McMaster大学公派出国留学生；2012~2013年复旦大学高分子系讲师，2013~2019年复旦大学高分子系副教授. 2019年至今，复旦大学高分子系教授. 主要从事高分子缠结理论、机器学习在高分子物理中的应用、非平衡热力学方法、大脑理论模型构建等方面研究.

REFERENCES

1 Lumry R, Eyring H. J Phys Chem, 1954, 58(2): 110−120
2 Dill K A, MacCallum J L. Science, 2012, 338(6110): 1042−1046
3 Anfinsen C B. Science, 1973, 181(4096): 223−230
4 Pauling L, Corey R B, Branson H R. Proc Natl Acad Sci USA, 1951, 37: 206−212
5 Dill K A. Biochem, 1990, 29(31): 7133−7155
6 Leopold P E, Montal M, Onuchic J N. Proc Natl Acad Sci USA, 1992, 89: 8721−8725
7 Bryngelson J D, Onuchic J N, Socci N D, Wolynes P G. Proteins, 1995, 21: 167−195
8 Dill K A, Chan H S. Nat Struct Biol, 1997, 4: 10−19
9 Dill K A. Biochem, 1985, 24: 1501−1509
10 Bryngelson J D, Wolynes P G. Proc Natl Acad Sci USA, 1987, 84: 7524−7528
11 Karplus M. Fold Des, 1997, 2: S69−S75
12 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. Nature, 2021, 596: 583−594
13 Kuhlman B, Bradley P. Nat Rev Mol Cell Biol, 2019, 20: 681−697
14 Pearce R, Zhang Y. Curr Opin Struct Biol, 2021, 68: 194−207

15 Senior A W, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson A W R, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones D T, Silver D, Kavukcuoglu K, Hassabis D. Nature, 2020, 577: 706−727
16 Li J F, Zhang H D, Chen J Z Y. Phys Rev Lett, 2019, 123: 108002
17 Stormo G D, Schneider T D, Gold L, Ehrenfeucht A. Nucleic Acids Res, 1982, 10(9): 2997−3011
18 Fang C, Shang Y, Xu D. Proteins, 2018, 86(5): 592−598
19 Jiang Q, Jin X, Lee S J, Yao S W. J Mol Graph Model, 2017 76: 379−402
20 Wang J, Zhao F, Peng J, Xu J B. Proteomics, 2019, 11 (19): 3786−3792
21 Botelho S, Simas G, Silveira P. Lect Notes Comput Sci , 2006, 4634
22 Kountouris P, Hirst J D. BMC Bioinf, 2009, 10 (1): 437−450
23 Bouziane H, Messabih B, Chouarfia A. Soft Comput, 2015, 19(6):1663−1678
24 Bouziane H, Messabih B, Chouarfia A. Evol Bioinform, 2011, 7(7): 171
25 Kountouris P, Agathocleous M, Promponas V J, Chritodoulou G, Hadjicostas S, Vassiliades V, Christodoulou C. ACM Trans Comput Biol Bioinf, 2012, 9(3): 731−739
26 Zhang C X, Zheng W, Mortuza S M, Li Y, Zhang Y. Bioinformatics, 2020, 36(7): 2105−2112
27 Thompson JD, Linard B, Lecompte O, Poch O. PLoS One, 2011, 6(3): e18093
28 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I. Attention is all you need. In: 31st Conference on Neural Information Processing Systems. Long Beach: NIPS, 2017. 1−7
29 Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Proteins, 2019, 87(12): 1011−1020
30 Roy A, Kucukural A, Zhang Y. Nat Protoc, 2010, 5(4): 725−738
31 Zhang Y, Skolnick J. Proc Natl Acad Sci USA, 2004, 101(20): 7594−7599
32 Song Y F, DiMaio F, Wang R Y R, Kim D, Miles C, Brunette T J, Thompson J, Baker D. Structure, 2013, 21: 1735−1742
33 Sali A, Blundell T L. J Mol Biol, 1993, 234: 779−815
34 Zhang J, Zhang Y. PLoS One, 2010, 5(10): 315386
35 Soding J. Bioinformatics, 2005, 21(7): 951−960
36 Xu J. Proc Natl Acad Sci USA, 2019, 116(34): 16856−16865
37 Wu S T, Zhang Y. Proteins, 2008, 72(2): 547−556
38 Altschul S F, Madden T L, Schaffer A A, Zhang J H, Zhang Z, Miller W, Lipman D J. Nucleic Acids Res, 1997, 25(17): 3389−3402
39 Wang S, Sun S Q, Li Z, Zhang R Y, Xu J B. PLoS Comput Biol, 2017, 13(1): e1005324
40 Thompson J D, Higgins D G, Gibson T J. Nucleic Acids Res, 1994, 22(22): 4673−4680
41 Notredame C, Higgins D G, Heringa J. J Mol Biol, 2000, 302(1): 205−217
42 Katoh K, Misawa K, Kuma K, Miyata T. Nucleic Acids Res, 2002, 30(14): 3059−3066
43 Edgar R C. Nucleic Acids Res, 2004, 32(5): 1792−1797
44 Larkin M A, Blackshields G, Brown N P, Chenna R, McGettigan P A, McWilliam H, Valentin F, Wallace I M, Wilm A, Lopez R, Thompson J D, Gibson T J, Higgins D G. Bioinformatics, 2007, 23(21): 2947−2948
45 Sievers F, Wilm A, Dineen D, Gibson T J, Karplus K, Li W Z, Lopez R, McWilliam H, Remmert M, Soding J, Thompson J D, Higgins D G. Mol Syst Biol, 2011, 7(1): 539
46 Katoh K, Standley D M. Mol Biol Evol, 2013, 30(4): 772−780
47 Sievers F, Higgins D G. Curr Protoc Bioinformatics, 2014, 48(1): 3.11.1−3.11.16
48 Katoh K, Rozewicki J, Yamada K D. Brief Bioinform, 2019, 20(4): 1160−1166
49 Dill K A. Biochemistry, 1985, 24(6): 1501−1509
50 Gobel U, Sander C, Schneider R, Valencia A. Proteins, 1994, 18: 309−317
51 Kass I, Horovitz A. Proteins, 2002, 48: 611−617
52 Baldassi C, Zamparo M, Feinaucer C, Procaccini A, Zecchina R, Weight M, Pagnani A. PloS One, 2014, 9: 392721
53 Ekeberg M, Lovkvist C, Lan Y L, Weight M, Aurell E. Phys Rev E, 2013, 87: 012707
54 Li Y, Hu J, Zhang C X, Yu D J, Zhang Y. Bioinformatics, 2019, 35(22): 4647−4655
55 Sun H P, Huang Y, Wang X F, Zhang Y, Shen H B. Proteins, 2015, 83(3): 485−496
56 Atchley W R, Wollenberg K R, Fitch W M, Terhalle W, Dress A W. Mol Biol Evol, 2000, 17(1): 164−178
57 Fodor A A, Aldrich R W. Proteins, 2004, 56: 211−221
58 Weight M, White R A, Szurmant H, Hoch J A, Hwa T. Proc Natl Acad Sci USA, 2009, 106(1): 67−72
59 Morcos F, Pagnani A, Lunt B, et al. Proc Natl Acad Sci USA, 2011, 108(49): E1293−E1301

60 Balkrishnan S, Kamisetty H, Carbonell J G, Lee S I, Langmead C J. Proteins, 2011, 79(4): 1061−1078
61 Jones D T, Buchan D W A, Cozzetto D, Pontil M. Bioinformatics, 2012, 28(2): 184−190
62 Marks D S, Colwell L J, Sheridan R, Hopf T A, Pagnani A, Zecchina R, Sander C. PloS One, 2011, 6(12): e28766
63 Berman H M, Battistuz T, Bhat T N, Bluhm W F, Bourne P E, Burkhardt K, Feng Z, Gilliland G L, Iype L, Jain S. Acta Crystallogr D, 2002, 58(6): 899−907
64 Hou J, Wu T Q, Cao R Z, Cheng J L. Proteins, 2019, 87(12): 1165−1178
65 Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Proteins, 2019, 87(12): 1011−1020
66 Xu J R, Zhang Y. Bioinformatics, 2010, 26(7): 889−895
67 Mariani V, Biasini M, Barbato A, Schwede T. Bioinformatics, 2013, 29(21): 2722−2728
68 Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press, 2016. 96−152
69 Hornik K, Tinchcombe M, White H. Neural Netw, 1989, 2: 359−366
70 He K M, Zhang X Y, Ren S Q, Sun J. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Lasvegas, Nevada: IEEE, 2016. 770−778
71 Zheng W, Zhang C X, Wuyun Q Q, Pearce R, Li Y, Zhang Y. Nucleic Acids Res, 2019, 47: W429−W436
72 Jones D T, McGuffin L J. Proteins, 2003, 53: 480−485
73 Simons K T, Kooperberg C, Huang E, Baker D. J Mol Biol, 1997, 268: 209−225
74 Xu D, Zhang Y. Proteins, 2012, 80(7): 1715−1735
75 Jones D T. J Mol Biol, 1999, 292: 195−202
76 Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H, Teller E. J Chem Phys, 1953, 21(6): 1087−1092
77 Bowie J U, Eisenberg D. Proc Natl Acad Sci USA, 1994, 91: 4436−4440
78 Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee G R, Wang J, Cong Q, Kinch L N, Schaeffer R D, Millan C, Park H, Adams C, Glassman C R, DeGiovanni A, Pereira J H, Rodrigues A V, van Dijk A A, Ebrecht A C, Opperman D J, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy M K, Dalwadi U, Yip C K, Burke J E, Garcia K C, Grishin N V, Adams P D, Read R J, Baker D. Science, 2021, 373: 871−876
79 Chen L. Deep Learning and Practice with MindSpore. Singapore: Springer Nature, 2021. 17−60
80 LeCun Y, Bottou L, Bengio Y, Haffner P. Proc IEEE, 1998, 86(11): 2278−2324

Review

Application of Deep Learning in Protein Structure Prediction and Its Inspirations

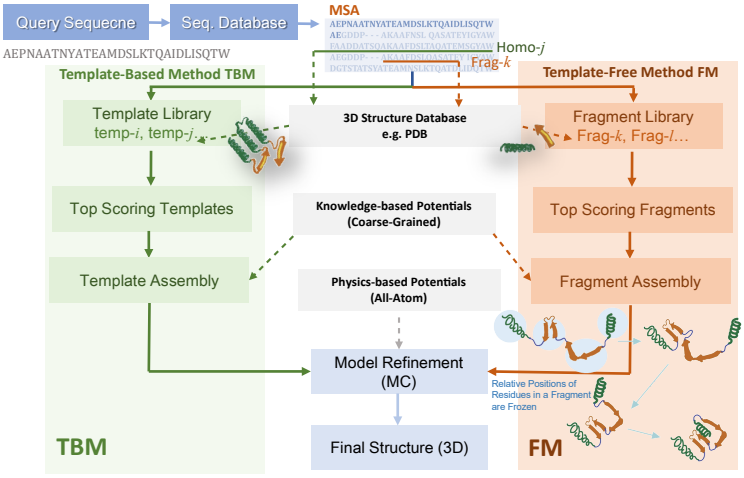
Tian-yao Wang, Jian-feng Li*

(*Department of Macromolecular Science, State Key Laboratory of Macromolecular Engineering of Polymers, Fudan University, Shanghai 200433*)

Abstract The goal of protein structure prediction is to determine, usually based on computer simulations or calculations, the three dimensional structure from a given amino acid sequence. Protein structure prediction is important since the 3D protein structure will further determine its biological functions. Nevertheless, traditional prediction method based on physics can only effectively deal with short proteins with the low accuracy. In the past decade, data-driven methods and methods based on genetic knowledge have become popular. This review covers several important developments about the deep-learning methods on protein prediction in the past ten years. Considering the education background of readers of the journal, we will first present a self-consistent but concise introduction about the prerequisite concepts and methods, related with genetic information and basis of deep learning, to understand the deep-learning methods for the protein structure prediction. The prerequisite concepts and methods include position-specific scoring matrix (PSSM), multiple sequence alignment (MSA), contact map, distogram, protein data bank (PDB), critical assessment of protein structure prediction (CASP),

* Corresponding author: Jian-feng Li, E-mail: Lijf@fudan.edu.cn

template modelling score (TM score), universal approximation theorem and several important types of neural network closely related with protein structure prediction. Then, we compared the two most popular methods employed in the data-driven protein structure prediction, template-based method and template-free method. As for the deep-learning methods, the AlphaFold method from Deepmind will be specially discussed, which has achieved the prediction accuracy comparable to median or low experimental accuracy that even rendered some people to think that it has resolved the protein structure prediction problem to some extent. Nevertheless, all the above methods are too “overwhelming” and not friendly for beginners. Therefore, this review also introduced a simplest structure prediction problem, HP protein prediction problem, together with the corresponding deep-learning solution, strongly-correlated neural network, to novices in this area.



Keywords Protein folding, Deep learning, Neural network, Structural prediction