# 计算机在蛋白质三维结构预测中的应用

物理 4+4 1801　胡喜平　U201811966

https://hxp.plus/

2020 年 12 月 3 日

## 1　Abstract 引言

Proteins are basic units of life, consisting of amino acids. As far as our human know, there are about 20 different types of amino acids associates with life, the in-sequence amino acids folds automatically into proteins with unique structure and function. Although the chemists have revealed us the forces between every two of the amino acids, but as for a protein, there are thousands of amino acids. Whenever a new amino acids joins, each of the previous ones changes position, the whole protein's structure changes, so the question becomes difficult and may not be solved by simply calculating the forces between each of the two amino acids, even if we have compute clouds consists of thousands of compute nodes. So we need algorithms to simplify our compute in predicting structure of proteins with given amino acid sequences.

蛋白质是生命活动的基本单元，它由氨基酸组成。而目前人类已知的构成生命体中蛋白质的氨基酸一共只有 20 余种，这些氨基酸按照顺序拼接在一起形成肽链，肽链自动折叠成具有三维空间结构的蛋白质。而氨基酸分子之间的相互作用，分子化学已经能够解释。但是对于成百上千的氨基酸按照顺序形成的蛋白质，每多一个氨基酸，肽链形成的空间结构就会改变，不同氨基酸之间链接时的相对方向和距离就会改变，这个问题就会变得无比复杂，即使人类的云计算集群有上千台计算节点，依旧不能处理这样的问题。因此需要一些算法来简化分子预测时的运算。最新新兴起的人工智能与机器学习技术被应用在了已知氨基酸序列的蛋白质结构预测中。

## 2　用计算机研究蛋白质折叠的意义

蛋白质的折叠一直是困扰物理学家和生物学家很久的问题。现在依旧没有解决的问题主要有三个方面：为什么特定氨基酸序列能折叠成特定的三维蛋白质空间结构；为什么蛋白质折叠的速度比我们预想的要快，以及如何根据氨基酸序列得到折叠成的蛋白质空间结构、如何从已知的蛋白质结构反推出它的氨基酸序列。

而计算机能解决的问题就是通过计算模拟来研究特定的氨基酸序列能折叠成什么样的三维空间结构。而且用计算机来模拟蛋白质的折叠对于实验和新药的研制有重大的意义。就像计算机芯片的设计一样，在芯片设计好之后要在计算机上模拟来检验芯片是否能达到预期的功能，以及效果是否符合预期、有没有潜在的问题，之后才会去制造一个样品芯片来进一步测试。芯片的制造很贵而且需要很多时间，因此先用计算机模拟尽可能发现问题再造实际的样品去测试能大大提升工作效率并节约经费。药物的设计

也是先根据我们想要的蛋白质空间结构来设计氨基酸序列，之后计算机模拟来初步检验设计是否达到了预期的效果，之后再实际制造样品来测试。

# 3 计算机模拟预测蛋白质结构的算法

## 3.1 Protein Data Bank

虽然分子化学已经揭示了氨基酸之间相互作用，但是我们不可能让计算机来计算每个氨基酸之间的相对作用，至少以现在的计算机的计算能力，这个想法是不现实的。所以现在的算法都是要基于人类已经用原子力显微镜等实验仪器做实验测出的蛋白质空间结构和它们氨基酸序列的对应关系。这些已知的蛋白质空间结构和对应的氨基酸序列是成百上千的科学家测出来并上传到数据库公开共享给所有人使用的。以下介绍的算法也是必须依赖数据库，即 Protein Data Bank。

## 3.2 Temple-based Modeling

Temple-based Modeling 算法指的是将需要预测的氨基酸序列在数据库中搜索相似的序列，然后找到相似的序列对应的空间结构。数据库的容量随着人类做实验的次数而增加，这个算法就会越精准。但是随着数据库越来越大，查找速度就会越来越慢，必须要有一个算法来尽可能加快查找速度。

在云计算中，我们部署 Redis 或者 Memcached 集群来更快地查找数据，数据库集群分布式地部署在不同的服务器上，集群中的主从节点相互协调，实现数据的快速查找和数据库整体的稳定。Redis 和 Memcached 为了查找迅速都把数据存储在服务器的内存而不是硬盘上，同时，在算法方面，都应用的 Hash Table 查找算法。服务器存储相应的 key 和 value，每一个 key 对应一个 value。对于每一个 key，存储的时候计算它的 Hash 值，Hash 值作为索引存在数据库的相应位置。之后查询到这个 key 的时候，计算 Hash，直接在 Hash 索引的位置找到对应的 value。

同时，氨基酸和蛋白质空间结构也是像 key 和 value 那样的一一对应关系，也存储在数据库中，那当然可以用相似的算法来进行快速查找，HHblits（HMM-HMM–based lightning-fast iterative sequence search）算法是目前搜索相似序列比较快的算法，其中里面的 “Hash” 叫做 “Hidden Markov Models”，即 HMM，“key” 就是目标序列，“value” 就是空间结构。HHblits 算法先将要查询的序列转化成 HMM，这个步骤和计算机 Hash Table 查找算法中，将 key 通过 Hash 算法转换成索引有相似之处。之后通过查找 HMM 数据库，也就是 “Hash 表”，找出相似的 HMM，进而得到相似的序列。之后递归循环，对查找到的相似序列转换成 HMM，查询 HMM 数据库，得到更多相似的序列。

但是这个查找虽然类似于 Hash Table，但并不完全是 Hash Table，否则不直接用现成的 Redis 而研究这个就是浪费时间。Hash Table 只能用于精准的匹配，不能用来搜索相似但是不相等的数据。而 HHblits 算法计算了 “Hash”，也就是 HMM 后，能根据 HMM 找出相似的序列的 HMM。

## 3.3 fragment-assembly

fragment-assembly 是指将目标序列分解成小的、相互重叠的片段，之后在数据库里搜索相应的序列并应用结构，进行预测。其中比较历史悠久的有 FRAGFOLD 算法。FRAGFOLD 算法先对目标序列进行分解，先在目标序列中选择一些 “超二级结构”，其中 “超二级结构” 由蛋白质数据库里面两到三个 “二级结构” 定义。选择好目标序列的片段后，开始模拟折叠。首先随机选取片段来组成随机的结构，每随机选取一个氨基酸片段，就把它加入到序列中，当所有的片段都加上去以后开始检验这个序列组成的结构

是不是有效的，如果检测到存在两个原子距离过进，那就是无效序列，从头开始继续尝试，直到得到有效的序列。

## 4 深度学习对计算机模拟算法的进一步优化

深度学习是现在人工智能的先进技术，它指的是通过构建计算机神经网络，然后向神经网络中投喂大量的数据，以此来训练神经网络，让神经网络在庞大的数据中发现数据之间的关联性。并且之后在输入未知的数据，计算机通过之前学习到的数据之间的关联，预测出结果。

在这方面最新的成功 AlphaFold 中，AlphaFold 的神经网络学习蛋白质数据库，之后神经网络训练对于蛋白质中势能的预测，之后根据梯度下降的算法得到蛋白质势能最小的空间结构，即蛋白质最有可能的空间结构。

## 5 前景与展望

虽然蛋白质的序列预测问题已经存在了 50 年以上了，50 年来人类的科技有了很大的进步，尤其是计算机技术的进步。从 1969 年末互联网的诞生，1970 年 Unix 元年，到后来 1991 年 Linux 内核诞生，再到现在 Linux 云计算集群，不仅单台计算机的计算能力在逐渐提高，而且 OpenStack 云计算和 Redis 数据库集群技术使得成百上千台计算机可以同时进行云计算、云存储，组成更强大的计算能力。但是即使是这样依旧不能解决蛋白质结构预测这种问题，即使是最先进的 AlphaFold，在最新的 CASP 挑战赛中也只能做对 2/3 的题目。

计算机运算能力可以使得这个问题更准确地被解决，蛋白质数据库的容量也能提高预测的准确度，但是一味地购买更多的服务器来并入云计算集群来增加计算能力，或者夜以继日地做实验丰富蛋白质数据库，都不如研发一个更精准的预测算法对解决蛋白质序列预测问题更有帮助。毕竟算法的更新是革命性的突破。

## 参考文献


[1] HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment , Michael Remmert, Andreas Biegert, Andreas Hauser & Johannes Söding , nature methods,VOL.9 NO.2,FEBRUARY 2012

[2] Predicting Novel Protein Folds by Using FRAGFOLD , David T. Jones , Department of Biological Sciences, Brunel University, Uxbridge, Middlesex, United Kingdom

[3] AlphaFold: Improved protein structure prediction using potentials from deep learning , Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick , Laurent Sifre, Tim Green1, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, Demis Hassabis , DeepMind, London, UK, The Francis Crick Institute, London, UK, University College London, London, UK,