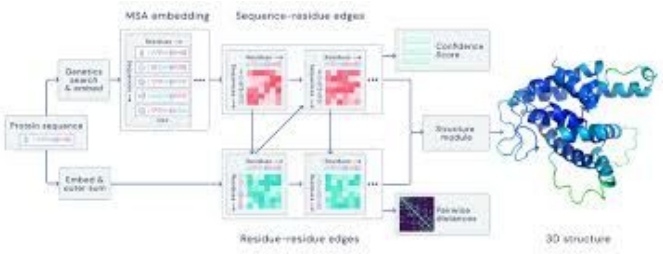
如何看待 AlphaFold 在蛋白质结构预测领域的成功？ - 知乎

https://www.zhihu.com/question/304484648/answer/544915994

郭昊天生物学话题下的优秀答主 Sat Jul, 24 20:22

更新，alphaFold出了第二版，基本上就是把第一版掀翻重做了。相关讨论大家可以移步到：

AlphaFold2 解决了蛋白质结构问题吗？DeepMind 解决这项生物学五十年难题有何重大意义？ www.zhihu.com



蛋白质可能是维持生命运动最重要的物质。自Anfinsen提出蛋白质的高级空间结构由蛋白质的氨基酸序列决定（因此获得1972年诺奖）后，人们慢慢就开始寻找一种蛋白质结构预测算法，可以精确地从蛋白质的氨基酸序列，由计算机预测出其复杂的空间结构，甚至最终由结构决定其功能。这不但是生物信息学，也是整个生物学中的一个重要的圣杯。

CASP，Critical Assessment of protein Structure Prediction，是一个国际性的比赛，相当于蛋白质结构预测界的世界杯，从94年开始每两年一届，今年是第13届。

我就围绕现在第13届CASP比赛有的材料谈一谈整个新闻吧。

先来搬运一下：
本届CASP13的所有摘要，AlphaFold参赛队伍为A7D，在第11页：http://predictioncenter.org/casp13/doc/CASP13_Abstracts.pdf

DeepMind自己的通讯稿：AlphaFold: Using AI for scientific discovery | DeepMind

本届CASP13的结果：Results - CASP13

首先不要听风就是雨

新闻媒体（尤其是DeepMind自己）不会告诉你，事实上，包括AlphaFold在内，今年CASP比赛，前五名都是深度学习和传统算法混合的方法

（DeepMind作为一个公司，有打广告的需求，像google爸爸展示自己，这个其实也怨不得他，但是你再讲一遍，等于你也有责任）

实际上今年的前五名都是深度学习和传统算法混合：

第一名 AlphaFold = CNN+Rosetta

第二名 C-I-TASSER = CNN+I-TASSER

第三名 Multicom = Deep learning + DNCON2 (CNN) + 一堆各种以前的包

第四名 C-QUARK = CNN + QUARK

第五名 C-I-TASSER Server版

那么，DeepMind的优势在哪呢？

有钱有TPU！

所以正确的新闻标题是：

硬件的胜利 | 蛋白结构预测，DeepMind的土豪金AlphaFold，大胜没钱的研究机构的其他深度学习算法

BTW，上一届CASP12就已经有一堆深度学习算法了，然而被传统算法吊打的。说白了，这还是算力进步和多次实践的结果，毕竟idea is cheap。

这不是人们第一次把深度学习使用在蛋白质结构上。今年有一个叫做DeepSF的文章发在bioinformatics上，根据蛋白质序列预测结构，不过更简单一些，仅仅是mapping。还有只做氨基酸间距离，也就是接触距离矩阵contact matrix的，最近的结果是DNCON2。今年还发表了很多使用强化学习进行蛋白质结构和功能设计的文章在BioRixv上了，很多很多，我这里就不列了。

其实这个AlphaFold的框架还是挺经典而且简单的……这个近年来我看完abstract最快明白了大概怎么实现的算法。。。。

他们一共设计了三个不同的方法。

方法1：

1. 先从一维的氨基酸序列生成一个二维的接触距离矩阵（contact matrix），表述两个氨基酸之间的距离。
2. 把蛋白质分割成几个结构域domain——domain内部的相互接触很强，但是domain内的分子和其之外的相对较弱。
3. 预测蛋白质骨架的折叠角度
4. 然后，根据蛋白质骨架的折叠角度，把结构域切割成，一系列有重叠的9个氨基酸残基为单位的短肽，分别预测，再组装到一起，预测整个结构域的结构。
5. 把折叠好的结构域组装到一起。

一般这里面的每一步都需要一定内部的评分系统，才知道哪个结果更好，太差的就扔掉，不用继续往下算了——和剪枝异曲同工。

实际上，这就是个最简单的从头预测（*ab initio*）方法的框架，不用深度学习，用生物物理知识，也是这么做的。

方法2：

在方法1的基础上，不进行“切割成小的短肽分别预测”，而是直接预测整个结构域

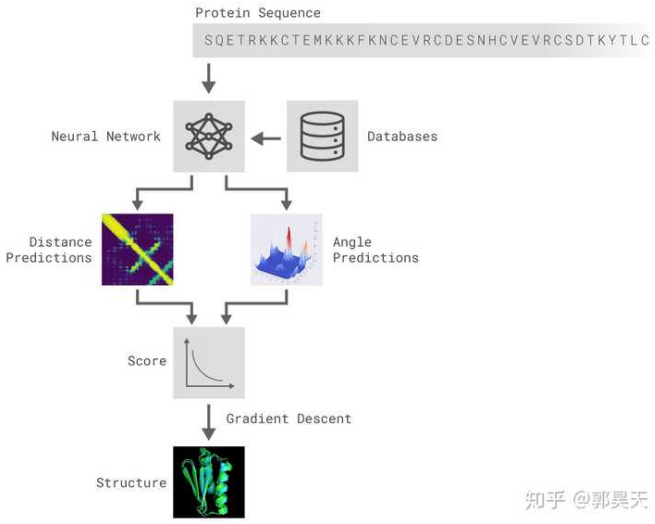
方法3：

连“结构域分割”都不做，三步走：

1. 从一维的氨基酸序列生成一个二维的contact matrix
2. 预测蛋白质骨架的折叠角度
3. 梯度下降出结果

方法2和方法3，我理解都是基于AlphaZero上尝到的甜头，抛弃人类的专家知识，只要DeepMind算力够强，就能大力出奇迹。

那方法三的训练方式就基本上就是AlphaFold自己宣传里用的这张图啦，有点过度宣传的嫌疑。在他们提交的作业里面，**只有**T0975一个使用了方法三得到的结果。



那么在哪里使用了深度学习呢？可以确定的有三个步骤。

第一个用到深度学习的地方是二维contact matrix预测。这里使用一个卷积网络CNN把一维的氨基酸序列，展成接触距离矩阵contact matrix。

训练这个接触距离矩阵网络，首先进行了BLAST序列比对，然后通过序列比对得到的特征，进行预测。这个网络深度应该非常高。但其实浅层的也可以做，比如今年早些时候的DNCON2。

这个网络会给自己也先估个分，看看靠不靠谱（likelihood），特别不靠谱的就扔了算了吧。

第二个是在预测蛋白质骨架结构的时候，要描绘每个肽键平面之间的二面角torsion angle，驻波的文章 @desmond 好像把这个漏了。这一步直接调了以前的一个图像生成的方法：A Recurrent Neural Network For Image Generation。这一步我倒是觉得挺有意思的，值得深究一下。

第三个是评估网络，还记得吧，整个算法里面，每一步都需要评分系统来确定哪些预测是好的。这个评估网络也是一个CNN。输入评估网络的是第一个CNN生成的contact matrix，序列比对产生的特征，还有结构的几何结构，等等。

有趣的是，有一个地方明确写了没有使用神经网络，就是根据contact matrix和torsion angle生成最终的蛋白质结构，以及利用短肽组装成结构域，利用结构域组装成全长蛋白结构的这几步，用的是传统的生物物理启发的模拟退火（simulated annealing）的方法。这个方法简单讲，就是在计算机里面模拟，把大分子放在开水里，逐渐降温，分子结构怎么变为最稳定的状态的一个方法，广泛地应用在各种最优化问题当中。

而在这种基本规则比较简单，但是实际能产生的情况又很复杂的（一个例子就是围棋啦），强化学习显然是一个很好的替代方案。但是靠强化学习发家致富的DeepMind却没有使用这个方法。

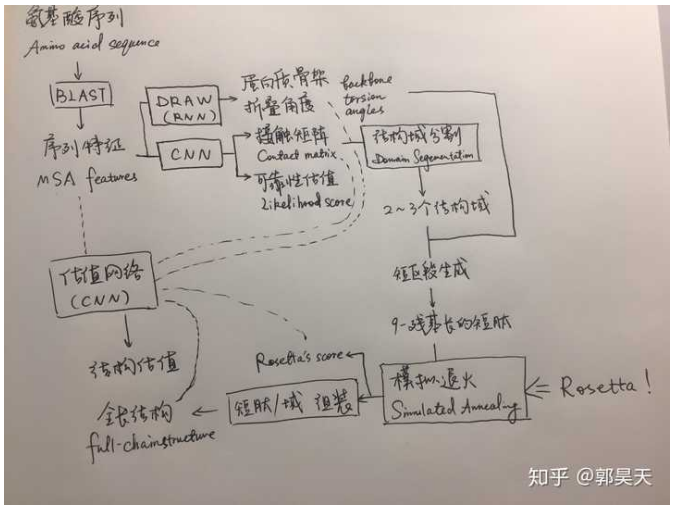
而且根据他们的测试，最好的方法似乎是使用传统的Rosetta?

Rosetta强无敌。。。

Rosetta是20年前华盛顿大学的David Baker Lab开发的一款也是可以从头预测蛋白质结构的软件（非常有名的Foldit这个众包蛋白质设计游戏也是他们家开发的），2005年期间搞了一个利用大家的电脑休息时间运行Rosetta的Rosetta@home ——比众包挖比特币的想法领先好几年……经过20年的发展，目前应该是全世界受众最广的蛋白质结构预测软件之一。

自己随便写一个模拟退火，不如Rosetta那是肯定的。但是强化学习在这个问题上不好用，并没有模拟退火好？那这个还是挺有趣的，期待看到相关结果。

A7D的这个abstract写的有点小乱，我这里先灵魂画师整理一下，步骤最繁杂的方法一的结构大致如下，希望到时DeepMind的文章出来不要打脸。。。



我们快速地过一遍。氨基酸序列经过BLAST，产生序列特征。

第一个CNN根据序列特征预测产生接触距离矩阵，contact matrix，并产生一个可靠性估值。

一个叫做DRAW的RNN，根据序列特征，预测蛋白质骨架的折叠角度backbone torsion angle。

利用contact matrix，我们可以进行结构域分割。

在每个结构域内部，我们可以进一步地，利用torsion angle，把一个大的结构域分割成很多（可重叠）的9个氨基酸残基长的短肽。

每个短肽经过已有的Rosetta fast relax传统算法模拟退火，得到一个短肽结构和Rosetta的评估。

短肽结构进一步进行模拟退火组装炒成结构域，结构域再组装成全长蛋白质的结构。（虽然没说，但是这里很可能也是用Rosetta做的）

从第一步到最后一步，每一步的输出，都被扔进一个叫做估值网络的CNN里面，一起打一个分。

训练：

第一个接触距离预测网络CNN可以单独训练，使用PDB的数据进行训练，使其生成的最高可靠性估值的contact matrix尽可能接近真实的结果。

DRAW这个RNN网络也可以单独训练，类似的，使用PDB的数据进行训练，使其生成的最高可靠性估值的torsion angle尽可能接近真实的实验结果。

最后训练估值网络其实也可以单独训练，对特定的氨基酸序列，blast产生的序列特征是不变的，但是后面的contact matrix、torsion angle可以从瞎搞的到正确的弄一串，然后分别灌进后续的程序和Rosetta，产生从乱七八糟到正确的一串结构。这一串结构和真实的实验结果一比对，就可以计算一个Z-score。而估值网路的输出应该尽可能接近Z-score。

预测：

扔进一个蛋白质序列，生成一个结构，一个评分。不满意就在生成一个。通过梯度下降找到有最好估值的结构。

其实预测的这个过程，应该相当于，在不动估值网络的情况下，把在PDB上预训练的接触距离预测网络和DRAW，针对特定的蛋白质，再进一步训练，从而得到最好的结构预测。

目前来看，这个方法一，最靠近传统框架，而且基本上后半程就完全是传统算法，效果其实会更好一些。

AlphaFold成功了吗？

其实距离所谓生物学的圣杯还远得很……为什么呢。首先要考虑一个问题：

什么时候预测结果能可信？

上个世纪生物信息学刚刚发展开始的时候，人们大概就定下了一些标准。

总体而言，一个非常宽松的标准是——你预测得到的结构，和真实实验测量得到的结构（ground truth），不同的概率应该非常小，Z-score应该在1~24这个范围，而平均值应该在11。

What should the Z-score of native protein structures be?

今天AlphaFold的平均值刚过1……所以还远得很。

AlphaFold的深度学习+模拟退火的方法是一条正确路径吗？

从逻辑上，AlphaFold的三种方法，虽然完全抛弃生物物理学知识的方法三也可以运行（需要人类手动选择），但是最接近传统的方法一最好。

那么是不是我们可以更退一步，把方法一中的一部分再替回传统方法呢？

或者把哪些传统软件中已经使用到的方法加入进来，比如说基于同源蛋白质的建模，和基于相似结构蛋白质的建模，能更进一步地提高准确率呢？

当然，**把深度学习引入蛋白质结构预测是大势所趋**，没有道理不用，也没有道理不好用。

以一个特定的算力，一定存在一个很好的处于平衡点的算法，混合了深度学习和基于人类知识的传统方法。对于正常人的算力而言，DeepMind这一套肯定不是那个平衡点。对于DeepMind而言，现在的方法一就是那个平衡点吗，也不见得。

如果我们再有DeepMind现在算力的100倍……没那么多数据喂电脑……

前五名都使用了深度学习技术，其他加入了深度学习的结构预测模型也很好

今年参赛的队伍中其实很多都是用了CNN的方法。开头就提到了前五名其实都用了深度学习技术。比如说拿到了第二名的密西根大学的Yang Zhang团队，就在I-TASSER的基础上，进行了优化得到C-I-TASSER。他们其中一步和AlphaFold很像，也是用ResPRE这个CNN生成contact matrix。而且I-TASSER本身就可以说是传统算法集大成了，在过去的几届CASP，包括今年，都垄断了Server的第一名。

相较而言，C-I-TASSER传统的比例更高一点。

那么这个C-I-TASSER的方法就一定不如AlphaFold吗?

AlphaFold的平均Z-score是1.1684，而C-I-TASSER的平均Z-score是1.0392。

这个差别其实相当微弱。前者相当于，预测结构和真实相符的概率为87.9%，而后者则相当于85.1%，也就是差不到3%的概率。

考虑DeepMind的计算力，这个边际效应太小了。你家是几千个TPU，人家在大学的顶多能分到几百个GPU，这都是我往大了估计……

如果使用DeepMind的资源，重新训练C-I-TASSER，或许，能够得到比AlphaFold更好的结果也未可知。

AlphaFold相对第2-5名的一个硬伤——预测准确率不稳定，疑似过拟合

要实现蛋白结构预测这个圣杯，最重要的是算法的输出结果的准确性很稳定。

为什么要搞算法？就是因为很多情况下做实验的成本太高了，所以人们希望能够在做极少量的实验之后，能通过算法解决大多数问题。

那么如果算法的准确性忽高忽低，那就坏了，我们怎么知道你这次是准还是不准？？

最后还是得老老实实回去做实验。

所以CASP目前使用Z-score的平均值进行排序，还是有一些缺陷的。

AlphaFold在43个参赛蛋白中，有25个拿到最佳模型，技压群雄。

但是AlphaFold预测不好的那些模型也很多，而且是真的很差很差。

T0966-D1是一个很好的例子，大家可以看专栏文章里面贴图[阿尔法狗再下一城｜蛋白结构预测AlphaFold大胜传统人类模型](#)。独树一帜地处于预测对了的算法和完全预测不对的算法中间，也就是基本瞎蒙。

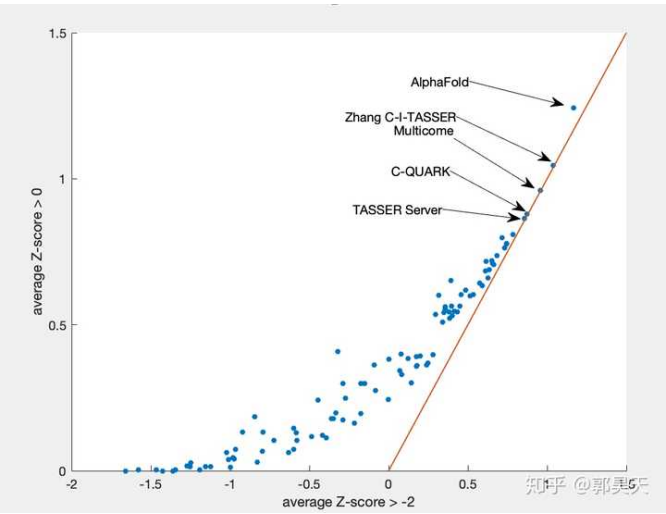
我这里，就从统计值上给大家展示一下为什么这么说。

之前说了，Z-score表示的，是有多大概率预测的结构和真实结构是同一个结构。这个Z-score越大越好。

CASP13统计了两种Z-score的平均值：一个是把所有预测结果里Z-score>-2的挑出来算一个Z-score的平均值，另一个是Z-score>0时的平均值。

那么你会希望这两种算法的结果越接近越好。当两者完全一致时，说明这个算法就没有得到Z-score<0这么差的结果的时候。

上图：这里红线标注的是两个平均值相等的情况，也就是边界，所有的数据点只能在其上方的区域，或落在线上。



可以看到第二名C-I-Tasser，第三名Multicome，第四名QUARK，第五名Tasser server，这4个模型，都贴在红线上。这说明第二到第五名，很少有Z-score掉到0以下的情况。而第一名AlphaFold却偏离了这条线不少，甚至比第六名（是个正八经的传统算法）还多。

说明一个问题，AlphaFold虽然有的结果预测得特别好，但是还有不少结果预测得特别烂。

当然，具体都有哪些结果AlphaFold预测得不好，还有具体分析一下，才能评估下一步如何优化它。

出现这样的问题，很可能是因为AlphaFold算力过剩导致过拟合了。毕竟AlphaFold和AlphaZero有很大的区别，是它本质上是一个监督学习。而PDB上的蛋白质结构的数目，和蛋白质整体的复杂度20^N相比，实在还是太太太小了，非常容易过拟合。这导致产生的模型在一些“见过”的蛋白质会表现得很好，而在陌生的蛋白质上就会表现奇差。

而考虑CASP每年用来测试的蛋白质并不多，这次AlphaFold有多大可能是撞大运赶上“见过”的蛋白质居多呢？概率虽然不大，但是也很难排除，这个要看数据。

我建议希望使用的同学，最好把今年CASP13的结果都过一遍，光看一个Z-score的平均值是不够的。

以上