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？ - 知乎

https://www.zhihu.com/question/432774098/answer/1605967409

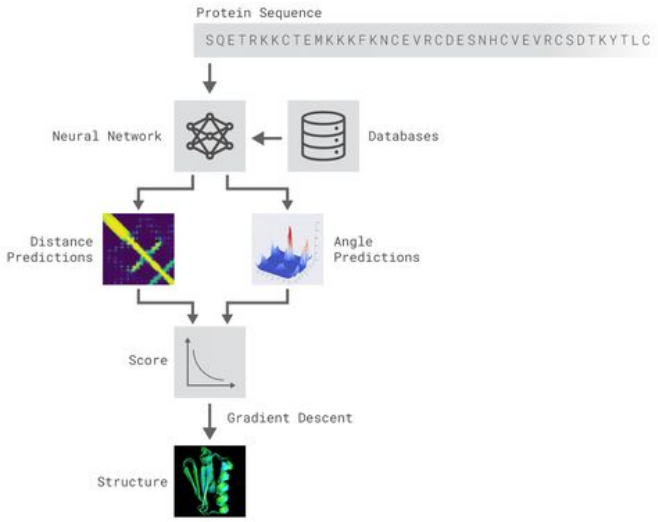
郭昊天生物学话题下的优秀答主 Sat Jul, 24 22:23

目录

谢邀 @Mon1st @SScream @李翛然

首先要挖一下坟。同属DeepMind的上一届Alphafold团队，也做出了不小的突破。但是主要体现在算力突破，方法论上并没有相对学术界实现很大的改变。当年我也很恨铁不成钢地理性批判过一番：

如何看待 AlphaFold 在蛋白质结构预测领域的成功？ www.zhihu.com



这一次，Alphafold v2终于拿出DeepMind该有的样子，全面碾压了来自学术界和产业界的竞争对手。（心疼Baker老师，显著性地slay全场之后，做了alphafold的背景板）

具体的文章还没出来，我也不讨论太多细节了。我想就一些形而上的东西，跟大家分享一些我的看法。

1. End-to-end的胜利

过往的很多科研界的探索，包括上次Alphafold 1.0，都仍然是处于一种严重路径依赖的状态。基本上框架都是相似的，只不过是试图用CNN或者RNN等机器学习方法，对已有的环节进行优化。

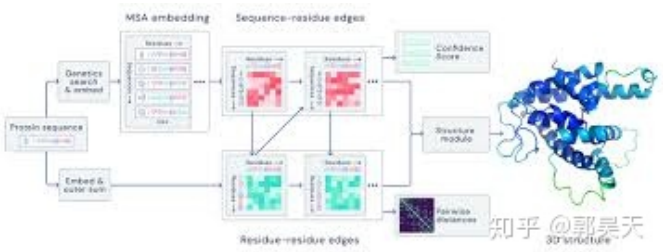
然而过去AlphaGo到AlphaZero已经展现了一个很好的例子，算力不足时产生的种种经验和方法论，在算力足够的情况下往往不是最优解。

那在这种情况下，不如就先从End-to-End做起，暴力迭代。

2. 怎么做End-to-End

简单的从Sequence-to-structure/function，也有不少1D CNN的例子啦。普遍跑出来效果渣渣。大概率不是算力的问题，而是数据量不够，导致neural network提不出feature来。

那么怎么做representation，怎么做数据增强就是最重要的问题了。而这里面，是需要设计者懂一点生物的。



一个高糊版的路线图

Alphafold 2.0 现在看起来是选用的MSA和pairwise distance的方法。这个大概率应该是目前我们能够理解而且也能够实践的最优解之一了。

Embed到2D Matrix上，是一个非常简单的trick，但是能提高的效率会非常显著。我自己以前做RNA的deep learning的时候也小尝了一下甜头，模型训练对算力和数据量的需求会大幅下降（打个广告）：

[Seq2DFunc: 2-dimensional convolutional neural network on graph representation of synthetic sequences from massive-throughput assaywww.biorxiv.org]

那么后人，可以在此基础之上，添加更多的限制，或者尝试不同的排列组合embedding，进行炼丹炒菜。

早期的基础探索，从0到1，很多是学术界淌出来的。但是一旦开始横向应用、深度拓展、多学科交叉，这种指数增长和快速变革，只能是研究型企业做的更好。（此处打个广告，有志向加入一家IT化的合成生物学平台型研究企业的，请私信联系我）

无他，企业的本质是效率工具。而学术界为了在一定程度上维护基础研究的创造性，产生了很多结构性的问题，导致其不可避免的周期长，效率低。所以在我国，产学研一体化是势在必行的。

至于其他大厂为什么被学术界吊打了……emmm……可能是因为团队过于CS了，生物不够强。所以专业的人做专业的事情是很重要的，如何组建、组织、管理一个学科交叉的队伍也不是件那么容易的事情。

4. AlphaFold能替代生物实验吗?

Alphafold很大程度上能替代简单蛋白质的结构生物学了。如果有人说只有实验才是真的，计算都是虚的。那么这些人你也不用和他们argue，他们大概率只会搬砖，实验科学的底子就没学好，不懂什么叫随机误差和系统误差。

当然结构生物学不会因此就没落了。还是有很多更复杂的结构等着解析。还有很多CNS等着发。

整体而言，alphafold不能替代实验。因为生物学实验中，功能才是最重要的。结构只是一个中间变量，重要但是也没那么重要。这里面还有大量的动力学问题是亟待解决的。

5. 生物科技的未来——效率工具和数据最重要

未来，生命科学如何发展。过去40-50年生命科学的大爆发式增长，起搏器说白了就那么几样：DNA测序、限制性内切酶、PCR、GFP、二代测序、ZFN-TALEN-CRISPR。这些毫无例外都是效率工具的革命。

AlphaFold本质上做的也是一个计算型的效率工具。

有哪个生命科学的理论革命推动了整个领域的发展吗？很难想到这样的例子。并不是因为理论不重要，而是因为：理论的提出需要少数天才的短暂灵感，是不可持续的；理论的接受需要科研界的更新换代，是非常漫长的。即使是进化论和现代遗传学这样的例子，其力量的显现也要几十年后。

那么效率工具带来什么呢？数据激增。集邮变得更简单了，可以集的邮也变得更多了。

集邮要变得有用，必须有计划的、成规模的、大批量地去做。

最近5年，大家对于AI+bio的各种尝试，已经很明显地说明了一个问题：生物数据质量和数量还远远不够。

如何产出大量的有效数据，如何快速高效地分析数据，是推动生物科技的核心。过去20年，这个事情主要是illumina在推动的。今天illumina是什么级别的公司，大家也看得见，几乎没有任何一家研究机构或者企业，在测序界会是illumina的对手。而未来20年生物科技，可以预见大量的产业创新，都将集中在这上面。

这样，我们离21世纪，也就不远了。