南開大學
Nankai University

专题三 蛋白质结构预测

Protein Structure Prediction

生信课题组

允公允能 日新月異

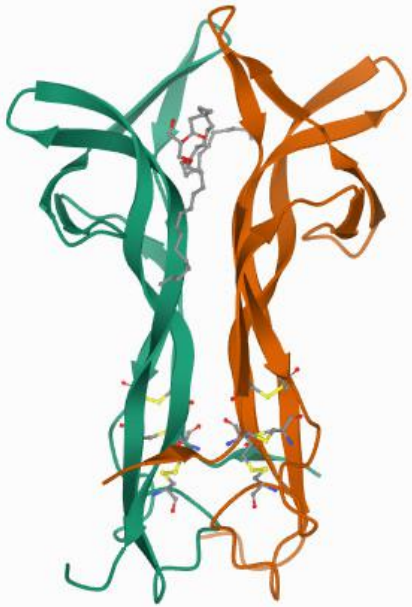
南開大學
Nankai University

1 Part One

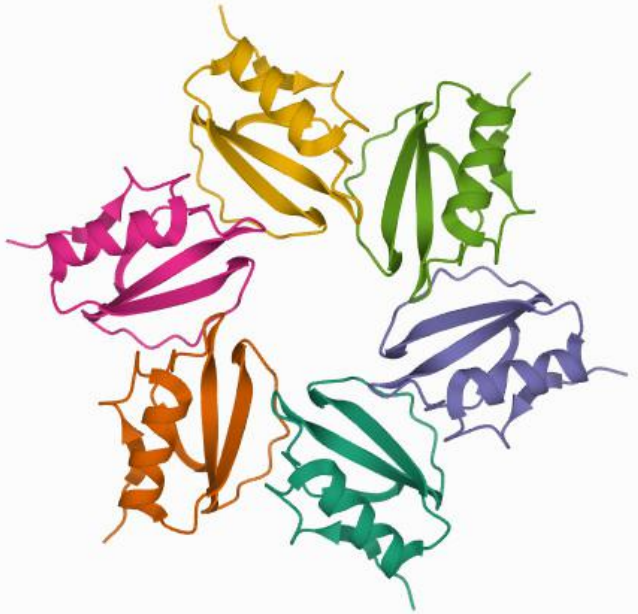
蛋白质结构之美

The Beauty of Protein Structure

欣赏蛋白质结构之美

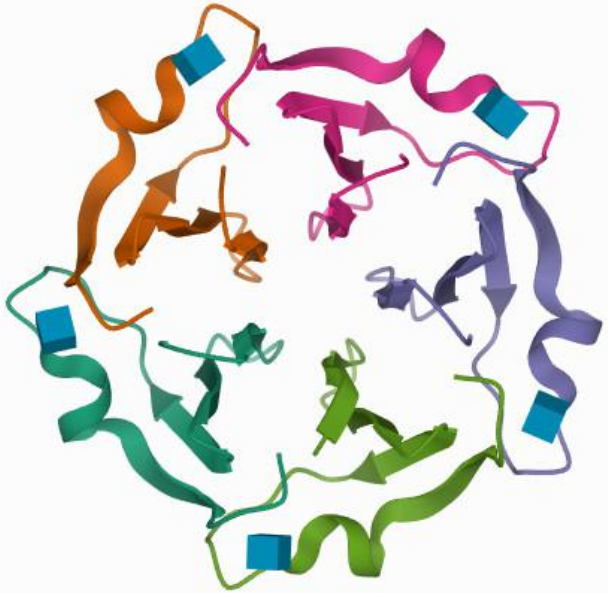


神经营养素-神经生长因子NGF
PDB ID: 4EC7

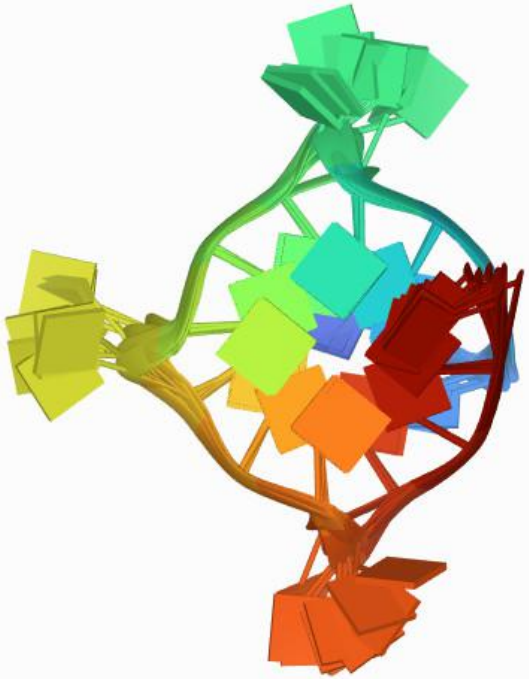


大麦蛋白酶抑制剂CI-2
PDB ID: 2CI2

欣赏蛋白质结构之美

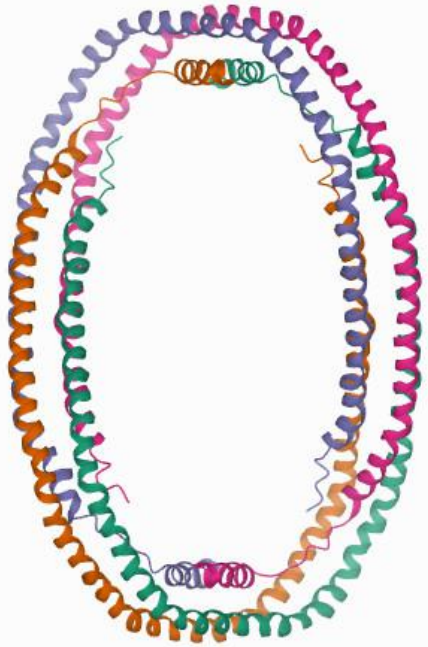


由折叠约束形成的对称蛋白质的从头进化
PDB ID: 5C2N

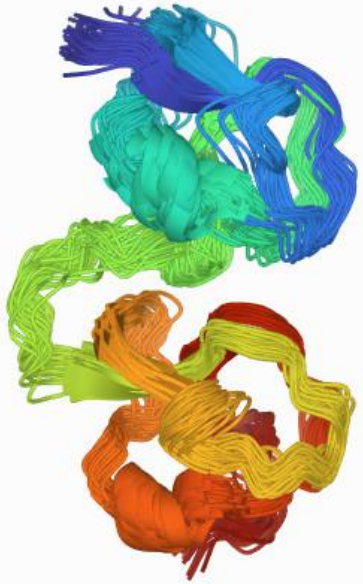


DNA poly-G 延伸中的平行链 G-四链体
PDB ID: 2MB2

欣赏蛋白质结构之美

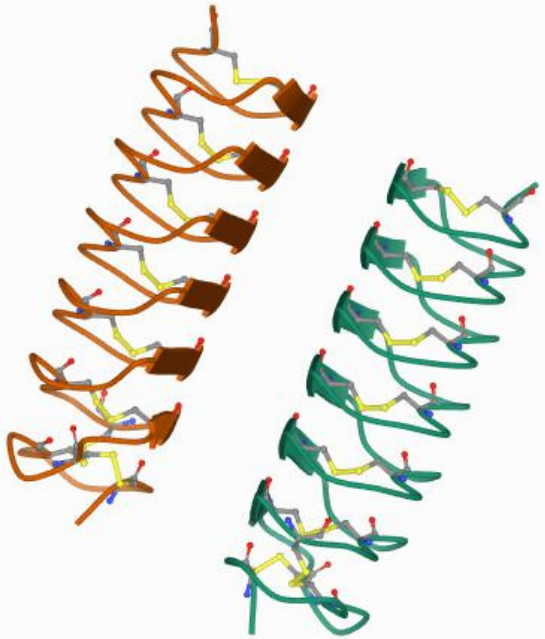


人载脂蛋白 A-I
PDB ID: 1AV1

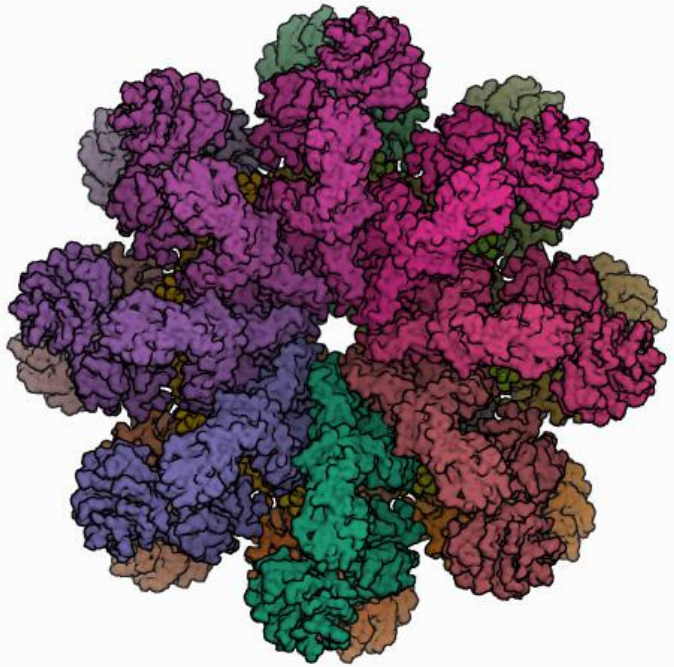


分子内二聚体抗冻蛋白RD3
PDB ID: 1C89

欣赏蛋白质结构之美



甲虫黄粉虫防冻蛋白
PDB ID: 1EZG



黑腹果蝇的凋亡小体
PDB ID: 3J9K

南開大學
Nankai University

2 Part Two

AI对本领域的影响

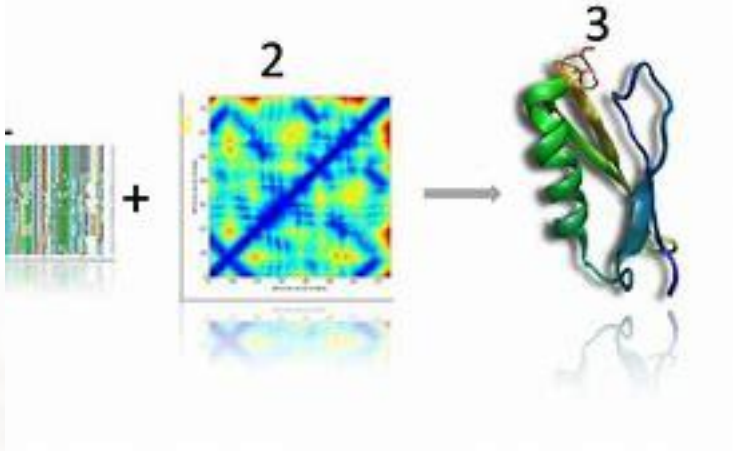
The Impact of AI on this Field

AlphaFold2

过去半个多世纪，人类一共解析了5万多个人源蛋白质的结构，人类蛋白质组里大约17%的氨基酸已有结构信息，而**AlphaFold2**预测的结构将这一数字从**17%**提高到**58%**。

它带来的在**生命科学**各分支领域的**革命**，将在今后几年到十几年中逐渐显现出来。

蛋白质结构预测是生物学的重要“圣杯”，也是**人工智能**落子生命科学领域最炙手可热的研究之一。



蛋白质折叠示例

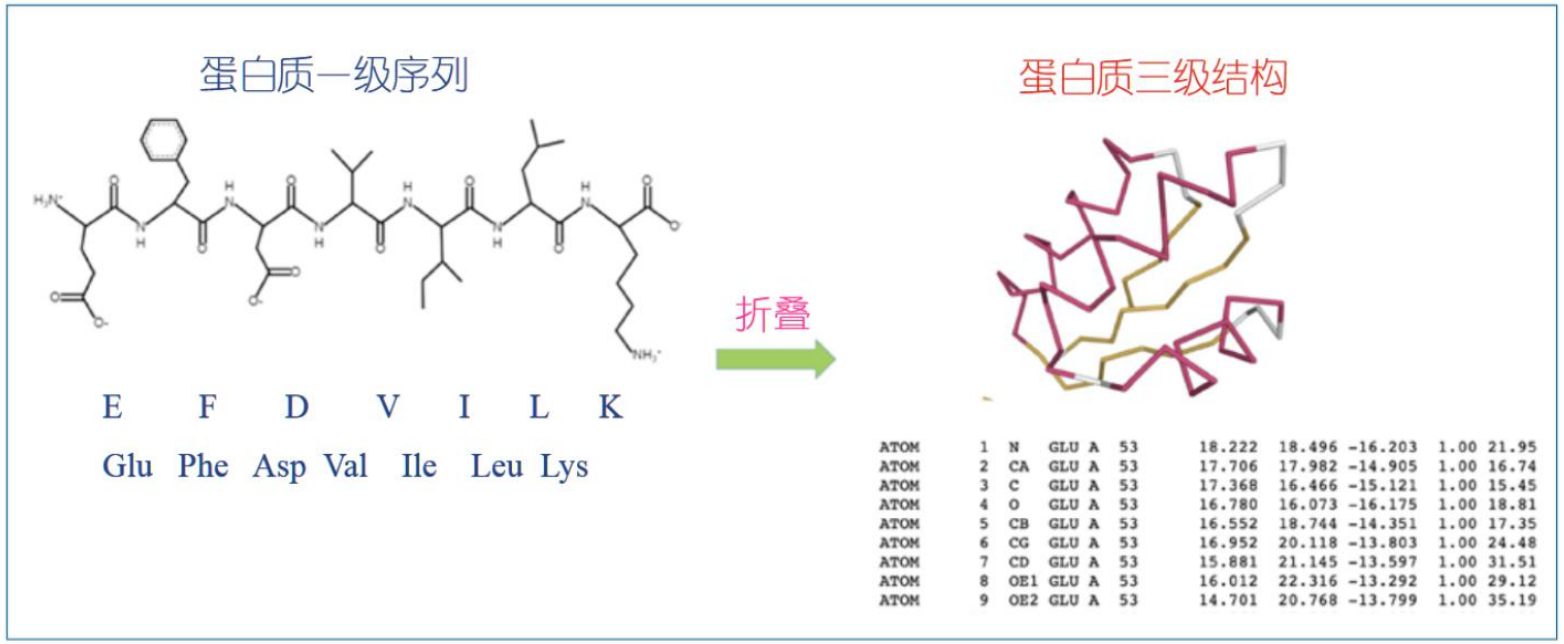


图1　蛋白质折叠示例。多个氨基酸依靠肽键连接成一条长链，此图表示蛋白质1ctf的7个氨基酸脱水后形成的长链以及折叠成的三级结构

蛋白质结构预测方法

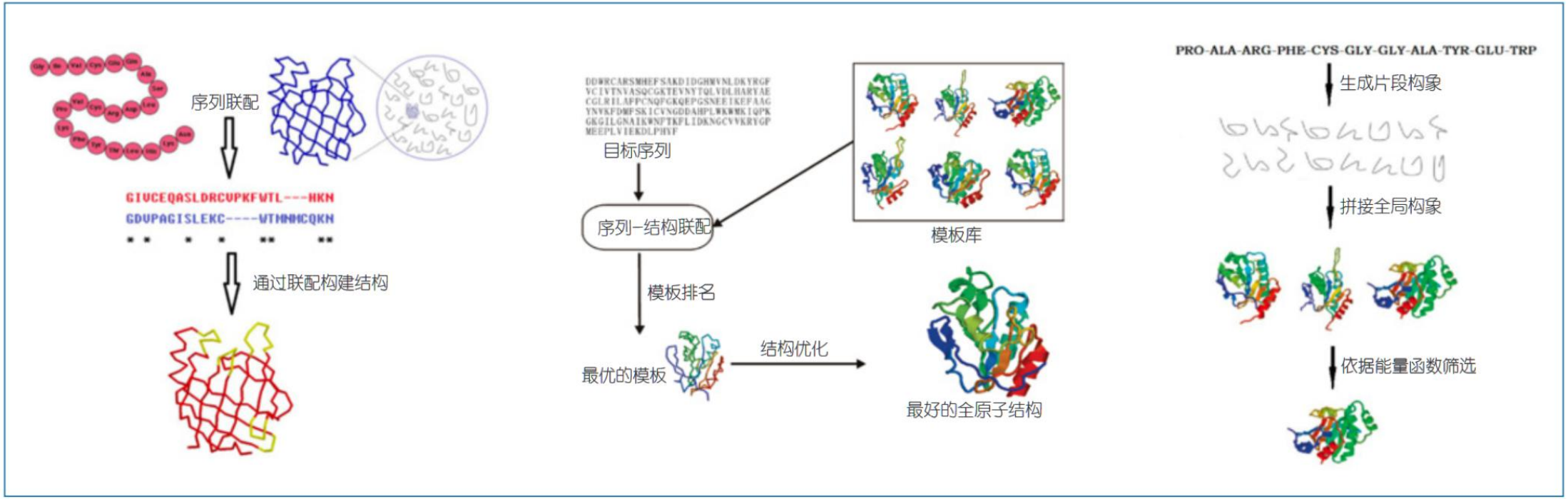


图2 蛋白质结构预测方法分类示意。左：同源建模法；中：归范法；右：从头预测法

- 有模板建模法。
  - 同源建模法；归范法，就是之前的 threading 方法
- 从头预测法（无模板）：

有模板建模法

- 同源建模法： 理论依据：如果两个蛋白质的序列比较相似，则其结构也有很大可能比较相似。
- 归范法：寻找与目标序列具有同一结构折叠类型（fold） 的蛋白质，关键步骤是 “序列-结构” 比较计算，以获得最可能的序列-结构联配。

- 一般来说，归范法，比同源法有更精确的结果。

原因， 归范法能充分利用模板的结构信息，例如，残基见相互作用，溶剂可及性等。

从头预测法（无模板法）

- 从头预测法核心思想： 寻找目标蛋白质能量最小的构象。从头预测的基本原理： 系统的稳定状态通常是自由能最小的状态。

蛋白质结构预测发展史

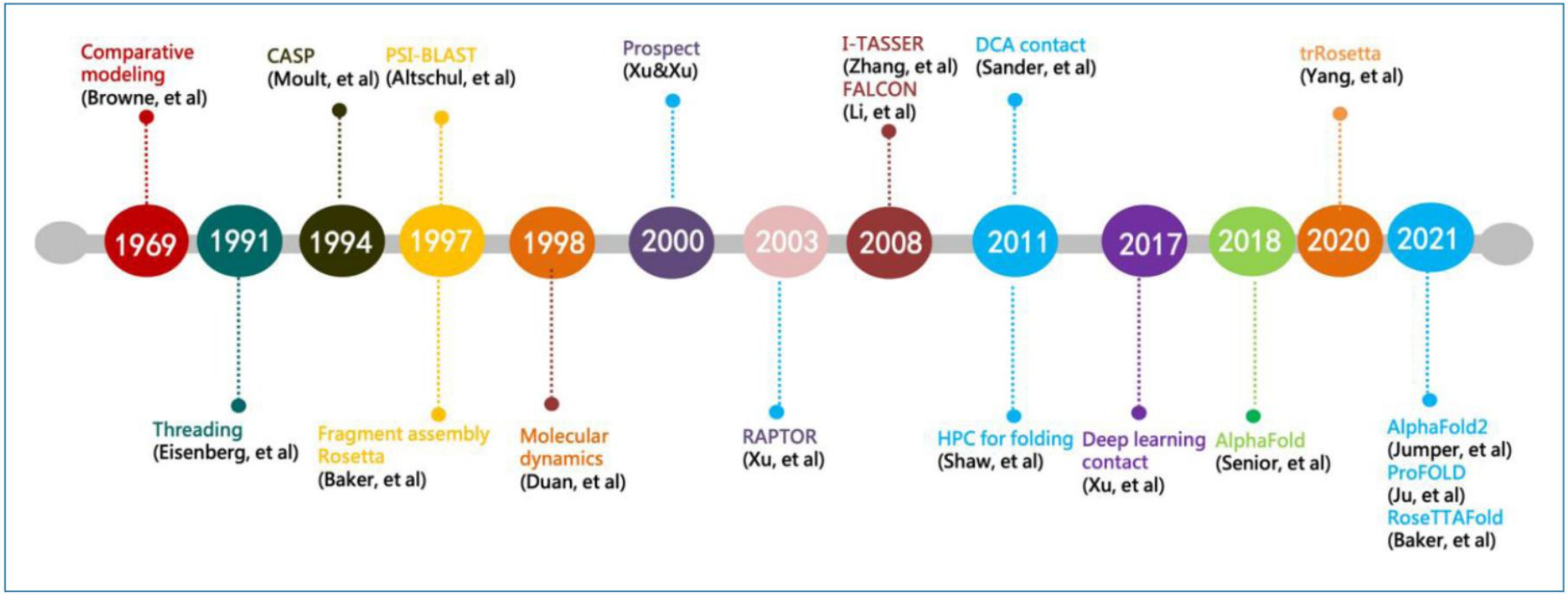
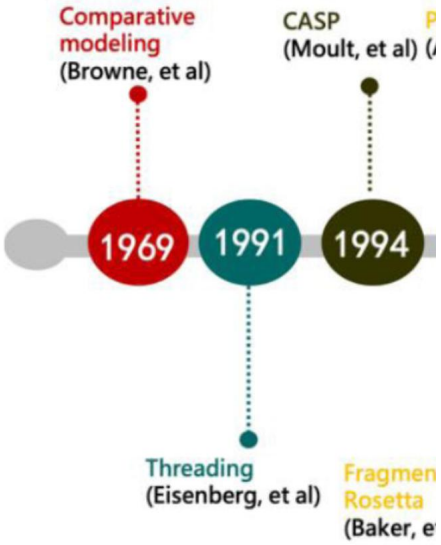
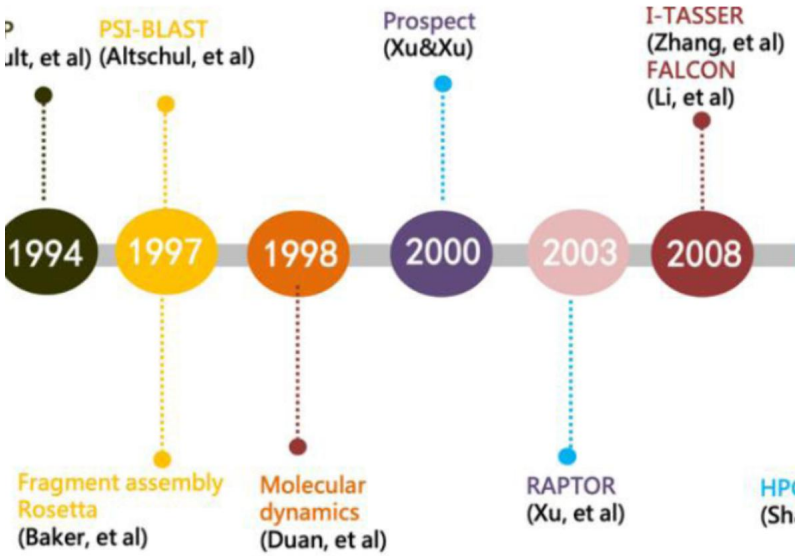


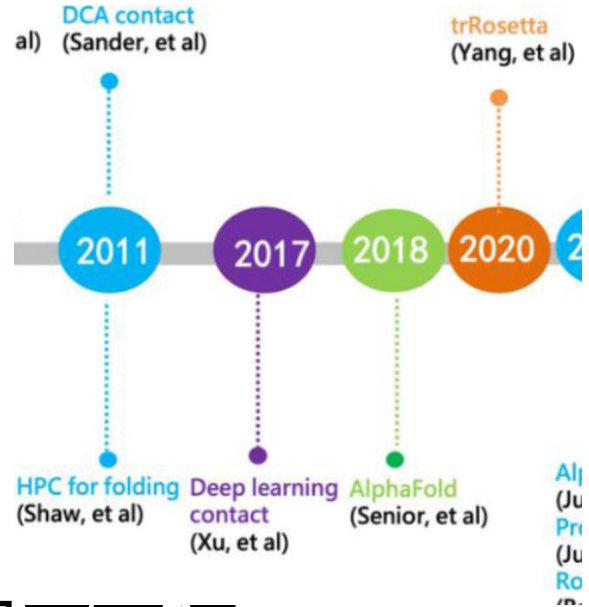
图3　蛋白质结构预测关键技术里程碑



- 1. 1969 年，布朗（Browne）等首次提出“比较 建模”策略，即借助具有充分序列相似度的蛋白质 模板进行预测；
- 2. 1991 年，艾森伯格（Eisenberg）等首次提出了“归范法”策略，其核心思想是“序列 - 结构”比对， 核心概念是“局部微环境”（local environment）；
- 3. 1994 年，约翰·莫尔特（John Moult）等组 织了 CASP 竞赛，采用盲测策略（blind test），客观 评价预测算法的性能，显著促进了结构预测的发展；



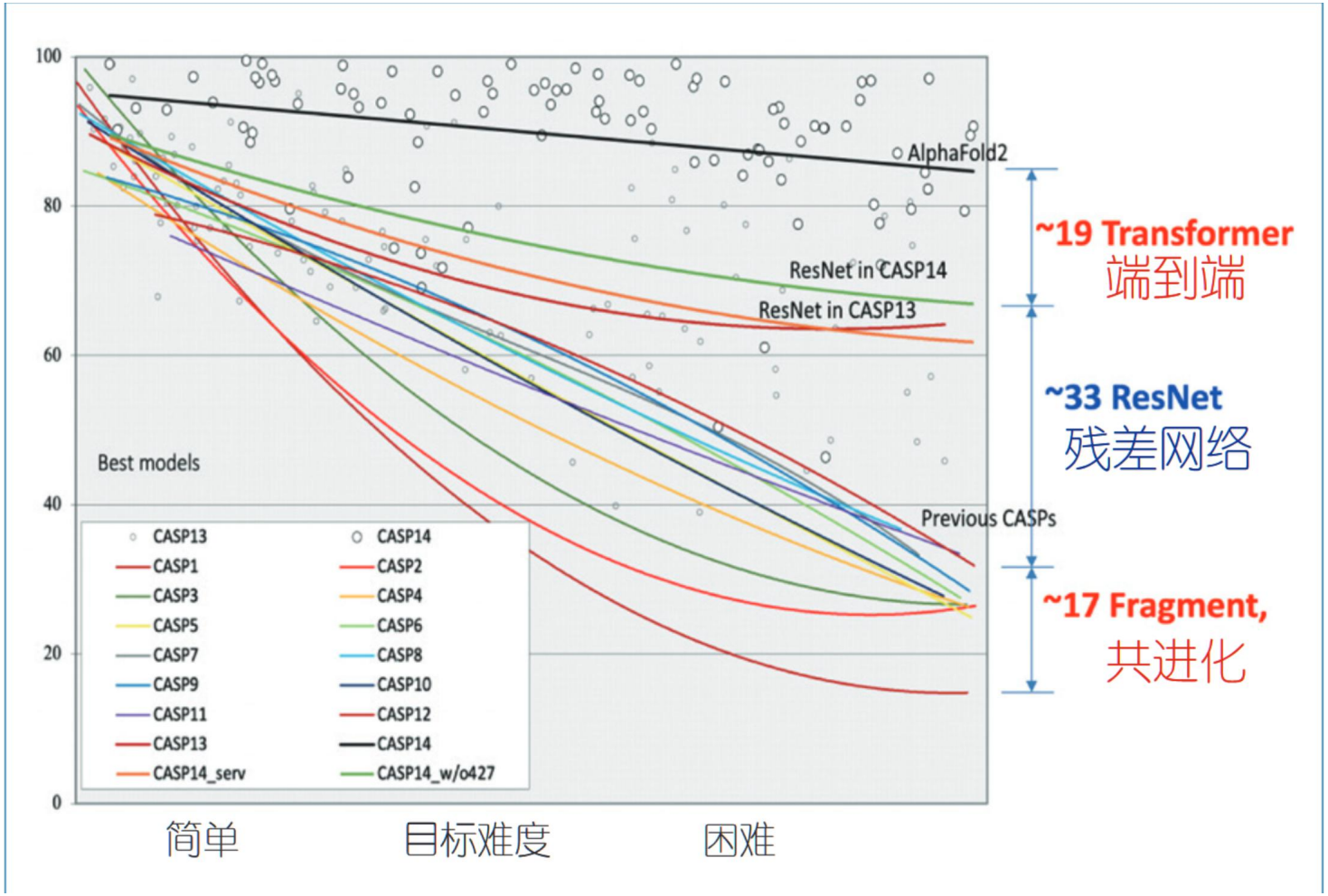
- 4. 1997 年，阿尔丘尔（Altschul）等开发了 PSI-Blast 工具，“位置特 异性”打分矩阵；
- 5. 1997 年，贝克（Baker）等开发了蛋白质结 构从头预测算法和软件 Rosetta，其核心思想是“结 构片段拼接”（fragment assembly），采用局部结 构隐式地表示“能量函数中的精细细节”，并采用 Monte Carlo 策略搜索出能量足够小的构象；
- 6. 2008 年，Zhang 等提出了 I-TASSER，基于“归 范法”得到的局部结构片段进行组装，其核心思想 是“不等长局部结构片段”以及“依据模板估计残 基间距离”，之后 D. Xu 等进一步开发了 QUARK； Li 等提出了“片段隐马尔科夫模型”（fragment-HMM），进行二面角采样，并采用类似“原始 - 对偶”的优化 策略，逐次迭代改进，获得了较高精度的预测结构 [24]



- 7. 2011 年，克里斯·桑德（Chris Sander）等 提出了平均成本法（DCA）策略，从测量系统分析 （MSA）中预测残基接触，其核心思想是“去除共 变中的传递性”；
- 8. 2017 年，J. Xu 等引入深度学习技术，基于 CCMPred 预测结果和一维信息学习出残基接触模 式，显著提高了残基接触的预测精度；
- 9. 2020 年，Yang 等提出了 trRosetta 算法，利 用残基间距离构建能量函数，进而计算出能量足够 小的构象。

- 10. 2021 年，Ju 等提出了新型神 经网络架构 CopulaNet，弥补了协方差 矩阵的信息丢失缺陷，显著提高了残 基间距离的预测精度，据此开发的预 测软件 ProFOLD 性能超过了 AlphaFold；DeepMind 公司继 AlphaFold 之 后，提出了 AlphaFold2，采用 3D 旋 转等变网络，实现了“端到端”的结结构预测。

历届CASP比赛各算法比较



导致技术进步的关键技术：

（1）基于结构片 段拼接的预测方法；

（2）基于共进化信息预测残基 间接触；

（3）采用 ResNet 预测残基间距离；

（4）采 用 Transformer 预测残基间距离，以及端到端的结构 预测技术。

生物学家、物理学家如何看

- 郑伟谋等将对蛋白质折叠过程的认识概括成四句话：“**精英绑架，层次折叠；强弱搭配，弱是必须**”，
- 其大意为：蛋白质折叠过程中，起主导作用的氨基酸并不多，而折叠过程是按照“先局部起始，再全局调整”的层次进行的；在蛋白质中，有些残基携带强结构信号，有些残基携带弱结构信号。值得强调的是：虽然蛋白质总是能够自发折叠成固定的天然态构象，但是这并不意味着所有残基携带的结构信息是同等重要的，恰恰相反，弱结构信号的残基是必不可少的，否则会妨碍蛋白质形成全局最优的构象

数学家、统计学家、计算机专家

- 统计学家：统计学家一上手就是建模：产生数据的模型是什么？数据的分布是什么？
- 数学家和计算机科学家多是按“能量最小化”的思路做蛋白质结构预测，把蛋白质结构预测问题形式化成一个最优化问题：定义能量函数，设计最小化能量函数的优化算法；或者直接最小化预测结构与真实结构（native structure）之间的差异。
- 近年来，采用深度学习技术直接“学”出蛋白质结构，成为这一波革命性进展的核心。靠“数据垒出经验分布”，而不是“靠人工经验定义分布”。

研究范式

- 观点一： 只要规律确实存在（ 1973 年安芬森（ Anfinsen）发现“结构信息蕴含于序列之中”）、数据足够多（迄今有 17 万个已知蛋白质结构、 22亿条非冗余序列）、规律足够简单或者能够简化，深度学习技术就可能借助这些规律进行预测 。
- 观点二： 一个成功的深度神经网络架构设计是建立在足够深刻的生物学洞察的基础之上的。
- 观点三： 理论研究与工程之间的界限日趋模糊，可以用工程的手段做基础研究。
- 观点四： 如何从深度学习的结果中获得知识，是值得思考的关键问题。

深度学习可在计算生物学领域广泛应用

这个里程碑事件令结构生物学家们感慨，自己用价值1000万美元的电镜努力了好几年得出的结果，Alphafold2竟然一下就算出来了。

“依我之见，这是人工智能对科学领域最大的一次贡献，也是人类在21世纪取得的最重要的科学突破之一。”

——生物物理学家、西湖大学校长施一公

参考资料

https://m.gmw.cn/baijia/2021-12/13/35376162.html

卜东波，大数据时代的生物信息学研究范式嬗变-以蛋白质结构预测为例

南開大學
Nankai University

感谢助教任钰同学的帮助！

制作：生信课题组

允公允能 日新月異