第×卷 第×期 集成技术 Vol. X No. X
××××年××月 JOURNAL OF INTEGRATION TECHNOLOGY ××. ××××
引文格式
夏辰亮，唐乾元. 基于 AlphaFold 数据库分析蛋白质进化中的统计规律 [J].集成技术, 2024,?(?):??

Citing format
Xia CL, Tang QY. Uncovering the Statistical Trends of Protein Evolution with AlphaFold Database[J]. Journal of Integration Technology,2024,?(?):??

基于 AlphaFold 数据库分析蛋白质进化中的统计规律

夏辰亮[1]，唐乾元[2]

[1]（三江学院数理部，南京 210012）

[2]（香港浸会大学物理系，香港九龙塘 999077）

摘要：由 DeepMind 开发的 AlphaFold 在蛋白质结构预测领域取得了前所未有的巨大突破，对生命科学的研究产生了革命性的影响。基于大规模的结构预测，AlphaFold 结构预测数据库得以建立，它包含超过 2 亿种蛋白，并覆盖了数十种物种的完整蛋白质组。这篇综述介绍了在“后 AlphaFold 时代”利用统计物理方法研究蛋白质进化问题的一些最新进展。传统的蛋白质进化研究往往关注同一个家族的蛋白质序列或者结构（微观视角），而随着 AlphaFold 预测的海量蛋白质结构的出现，研究者可以把视角扩展到大量蛋白质的集合，甚至是直接对比不同物种体内的全部蛋白质，从中挖掘统计趋势（宏观视角）。基于 AlphaFold 数据库，通过对比 40 多种模式生物体内相似链长的蛋白质，研究者发现了蛋白质分子进化中的统计规律。随着物种复杂度的提高，蛋白质结构将趋向于更高的柔性和模块化程度，蛋白质序列将趋向于出现更显著的亲疏水片段分隔，蛋白质的功能专一性也不断提高。这些基于 AlphaFold 的统计研究在分子进化和物种进化之间建立了联系，有助于我们理解生物复杂性的演化。

关键词：AlphaFold，蛋白质，进化，蛋白质动力学，简正模分析，统计物理

中图分类号： **文献标志码 A doi: 10.12146/j.issn.2095-3135.20230912001**

Uncovering the Statistical Trends of Protein Evolution with AlphaFold Database

XIA Chenliang[1], TANG Qianyuan[2]

[1] (Department of Mathematics and Physics, Sanjiang University, Nanjing 210012, P.R.China)
[2](Department of Physics, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR, P.R.China)
Corresponding Author: TANG Qianyuan. Email: tangqy@hkbu.edu.hk

来稿日期：2023/9/12 修回日期：2023/10/30
基金项目：江苏省高等学校自然科学研究项目(22KJD14005)；香港研究资助局杰出青年学者计划（ECS-22302723）
作者简介：夏辰亮，博士，讲师，研究方向为蛋白质动力学的统计物理研究；唐乾元（通讯作者），博士，助理教授，研究方向为统计物理在生命科学中的应用，E-mail：tangqy@hkbu.edu.hk

Abstract: AlphaFold, which is developed by DeepMind, has made amazing advances in predicting protein structures for life sciences research. Using the vast structural predictions made possible by AlphaFold, a database of over 200 million proteins has been established. Such a database covers the complete proteomes of many organsims. This review outlines the most recent progresses in exploring protein evolution using statistical physical methods based on the AlphaFold database. Traditional protein evolution research often concentrates on the sequences or structures of proteins within the same family, using a narrow microscopic approach. With the new emergence of extensive protein structure predictions by AlphaFold, whereas, scientists can expand their horizons to include vast assortments of proteins to make parallels with all proteins in different species and extract statistical trends through macroscopic observation. By comparing the proteins with similar chain lengths in over 40 model organisms, the statistical trends in protein evolution is discovered. For organisms with higher complexity, their constituent proteins present larger radii of gyration, higher flexibility, and higher segregation of hydrophobic and hydrophilic residues in both spatial and sequence. It is also validated by statistical physics analysis that higher organismal complexity correlates with higher functional specialization of constituent proteins. The findings in these studies connect molecular evolution to organism evolution, contributing to the understanding of the origin and evolution of lives.

Key words: AlphaFold; Protein; Evolution; Protein dynamics; Normal mode analysis; Statistical physics

Funding: This project is supported by The Natural Science Foundation of the Jiangsu Higher Education Institutions of China (22KJD14005) and Early Career Scheme (No. 22302723) from Research Grants Council of Hong Kong.

飞速发展的人工智能（AI）技术正在改变人们的生活，在我们身边，已经出现了许多由AI预测或生成的各种大数据，如对话、图像、音频、视频等等。而随着AI日益成为科学家的重要工具，由AI预测或生成的各种科学数据也在不断产生，对这些源于AI的大数据进行深度分析和处理已经成为科学研究中的重大挑战[1-4]。随着AI预测或生成数据的增加，许多此前无法开展的大规模统计分析现在也有了实现的可能，科学家们需要新的分析视角、统计方法和算法来解读和分析这些数据。

在生命科学领域，AI技术最为成功的应用之一要属AlphaFold。2021年，*Science*和*Nature*杂志不约而同地将“年度十大科学突破”颁给了由DeepMind开发的AI蛋白质结构预测系统AlphaFold[5-7]。AlphaFold是一种基于机器学习的蛋白质结构预测模型，基于蛋白质的氨基酸序列就能以极高的精度预测蛋白质的天然态结构。该系统的第二代AlphaFold 2（以下简称为AF2）在2020年12月举行的第14届蛋白质结构预测大赛中取得了前所未有的压倒性成功，目前已经成为分子生物学研究者们不可缺少的工具。

1 AlphaFold 简介和基本原理

1.1 蛋白质结构预测方法简介

蛋白质的三维结构预测是计算生物学和化学中最具挑战性的问题之一，50 多年来一直困扰着科学家们[8]。尽管科学家们使用了包括核磁共振[9]、 X 射线晶体学[10]、和冷冻电镜[11]等多种方法，但也只确定了大约 20 万个蛋白质的结构（截止到 2022 年），占蛋白质总数的不到 0.1%。自从 1961 年，Anfinsen 提出了关于蛋白质折叠的著名理论“蛋白质的三维结构仅由其氨基酸序列决定” [12]，此后科学家们一直在探索从氨基酸序列直接预测蛋白质三维结构的算法[13-18]，发展出了同源建模、从头建模和机器学习等方法。同源建模方法假设相似的序列编码着相似的三维结构[19-20]，因此对于给定的目标氨基酸序列，可以先从结构已知的蛋白质数据库中寻找它的同源序列，然后以同源蛋白质的三维结构为模版来搭建待预测序列的结构；而从头建模是一种基于“第一性原理”[21]的蛋白质结构预测方法，该方法根据目标序列的原子坐标建立一个自由能的能量函数，然后在整个构象空间搜索并计算不同构象的能量，最后将能量最低的构象作为最终的预测结构[22-23]。近年来，随着深度学习技术的发展，各种基于机器学习的方法逐渐成为了结构预测领域的主流，AF2 正是其中的重要代表。除此以外，美国华盛顿大学 David Baker 团队开发的 RoseTTAFold[24-25]和美国 Meta 公司团队基于大语言模型开发的 ESMFold[26]也都用到了机器学习的方法。

1.2 AlphaFold 2 工作原理简介

AF2 的结构预测主要基于蛋白质中不同位点的共进化（coevolution）信息。所谓“共进化”，指的是蛋白质序列中的两个位点在进化中的关联性。在进化过程中，这种突变关联有助于蛋白质维持结构和功能上的稳定性。如图 1（A）所示，在大量的同源蛋白序列中，如果其中的第 i 和第 j 个氨基酸残基总是一起发生突变，这说明第 i 和 j 个氨基酸残基在空间上较为靠近。基于这一思路，研究者们发展了一系列算法，基于多序列比对（Multiple Sequence Alignment，简写为 MSA），提取同源序列的突变关联，推断残基间的相互作用和距离约束[27-29]，并以此为基础进行蛋白质结构预测。这种思路也是 AF2 进行结构预测的理论基础。尽管同样用到了大量的同源序列，但这种方法与同源建模的思路是完全不同的。

图 1（B）展示了 AF2 的架构和工作原理，它主要包括四个部分：输入模块、编码模块、解码模块和循环[6]。首先是输入模块，输入给定序列后，AF2 用这个序列搜索序列数据库，搜到与该序列相似的同源序列后，生成一个包含了共进化信息的矩阵“MSA 表示”，这是一个三维张量，其大小为 $s \times r \times c$，其中，s 是 MSA 中的蛋白质序列总数、r 是输入序列的氨基酸数量，c 是氨基酸的通道数，接下来对输入序列中的氨基酸两两之间进行结构约束建模，在结构数据库中搜索氨基酸之间的结构信息，生成一个包含了结构约束信息的矩阵“残基对表示”，这也是一个三维张量，其大小为 $r \times r \times c$，r 和 c 的含义同上。第二个模块是编码模块，被称为 Evoformer，它是基于自注意力机制的 Transformer 的一种变形。Transformer 是近年在人工智能领域备受关注的一类深度学习框架，它能够从输入信息不同位置对之间提取信息，因此特别适合处理文本类、时序性的信息[30]。Evoformer 由 48 个块（block）组成，其中每个块都包含了多层神经网络，不同的块分别负责进行自注意力建模、多层感知器、外积平均、残基对之间的三角注意力建模等。上一步得到的两个矩阵“MSA 表示”和“残基对表示”放入 Evoformer 模块中处理，值得一提的是，Evoformer 可以让“MSA

表示”和“残基对表示”互相提供信息彼此优化。第三个模块是解码模块，将上一步处理后的“MSA 表示”和“残基对表示”解码成蛋白质三维结构，即构成蛋白质的各原子以欧几里得坐标(x,y,z)表示的空间坐标。最后 AF2 还要将该过程进行 3 次循环迭代，每次循环用上一次的输出作为本次的输入，进一步修正和提高预测的精度。

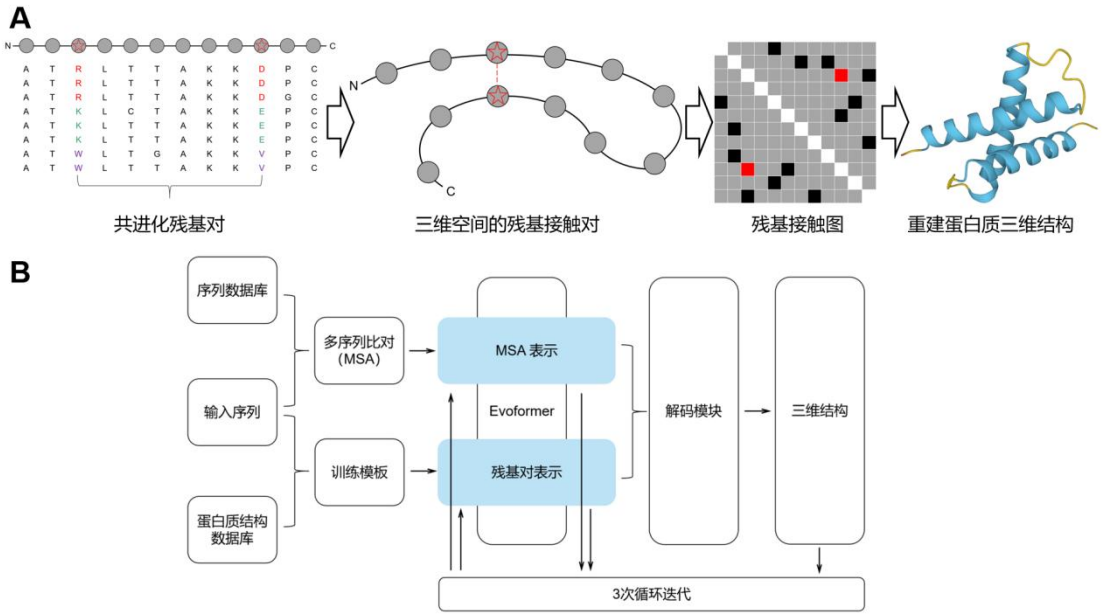

图 1 AlphaFold 2 蛋白质结构预测的（A）基本原理与（B）基本架构示意图，图 B 参考了 AF2 的原始论文[6]。

Fig. 1 The schematic illustration of the (A) working mechanism and (B) model architecture of AlphaFold 2 protein structure prediction. Figure B refers to the original paper of AF2[6].

2 AlphaFold 2 的应用

2.1 AlphaFold 2 在生命医学中的广泛应用

尽管 AF2 的诞生距今只有短短几年的时间，但它已经对生物医学领域的研究产生了重要影响，涌现出了许许多多的应用研究[31]。在结构生物学领域，AF2 结构预测可以结合 X 射线晶体衍射技术[32]和冷冻电镜技术测定蛋白质的结构[33-34]；在蛋白质设计领域，AF2 所提供的结构预测可以作为重要的参考[35-37]；在药物研发领域，AF2 可以被用于计算机药物筛选[38-39]；在生物信息学领域，从 AF2 所预测的蛋白质结构中可以提取出许多关键的特征，辅助蛋白质功能的预测[40-41]。除此以外，AF2 还广泛地用于蛋白质与配体相互作用的预测[42-43]和蛋白质—蛋白质相互作用的预测[44-46]。

2.2 AlphaFold 蛋白质结构数据库

从 2021 年开始，AlphaFold 发布了自己的蛋白质结构数据库（AlphaFold Protein Structure Database，以下简称 AFDB），其中包含了从细菌、古细菌、单细胞和多细胞真核生物到人类等在内的许多物种的完整蛋白质组[47-48]。目前，AFDB 数据库还在不断扩大，

在 2022 年 7 月底的更新中，AFDB 已经扩充到包含约 2 亿个预测的蛋白质结构。AFDB 展现出了巨大的应用前景，包括利用其进行大规模的药物筛选等。在 AFDB 的基础上，研究者们可以对特定类型的蛋白质或者蛋白质的特定属性展开研究，例如：对别构蛋白的分析[49]、对钙调蛋白的结构进行跨物种分析[50]、对罕见病相关的蛋白进行分析[51]、研究蛋白质的水溶性质[52]等等。值得一提的是，AFDB 不仅能帮助科学家们解决医学和生命科学中的关键问题，在进化研究中也显示出了新的可能性[53-57]。本文接下来将主要介绍基于 AFDB 展开蛋白质结构进化的相关研究。

3 利用 AlphaFold 数据库分析蛋白质结构进化中的统计规律

3.1 研究进化的宏观与微观视角

在经典的分子进化研究中，序列分析是最重要的研究工具，即通过比较不同蛋白质所对应的基因或氨基酸序列，计算其相似性并构建进化树，然而，基于序列的分析方法不足以从宏观的角度刻画物种进化与分子进化之间的相互关系。AI 结构预测方法的发展使得基于蛋白质结构及其内禀动力学的大规模进化分析成为可能，这些分析将有助于我们更完整地获取蛋白质进化的图像。

近期，Barrio-Hernandez 等人开发了一种基于结构对齐的聚类算法，能够根据蛋白质形状的相似性，快速比较 AFDB 中的每个结构。他们使用该算法在 AFDB 中识别了 230 多万个形状相似的蛋白质簇（cluster）[57]。进化分析表明，大多数簇的起源是古老的，但其中 4% 似乎是物种特异性的，这可能意味着新基因的诞生。这项研究开始探索蛋白质“宇宙”中的未知“星系”，并且已经发现了一些令人惊讶的联系。例如，人类和其他复杂生物体用来检测病毒 DNA 并引发快速免疫攻击的一种蛋白质与来自单细胞细菌和古细菌的蛋白质处于一个簇中，这种联系此前并未被其它文献所报道。

除了上述基于结构聚类的分析方法之外，也可以利用 AFDB，从宏观与微观之间的联系的角度来分析蛋白质进化。从宏观尺度来看，在漫长的进化过程中，生命的复杂性呈现随时间不断增加的趋势，例如从原核生物到真核生物，从单细胞生物到多细胞生物等。在微观尺度下，与生物体的复杂化并行发生着另一种进化过程，那就是分子进化，即作为生物体基本构件的蛋白质分子也在不断进化。进化的这种宏观（生物体）与微观（蛋白质分子）视角之间是否存在某种联系？直观上来看，某种特定的蛋白质分子的进化不一定遵循物种进化的路径，然而，如果把视角扩展到大量蛋白质的集合、甚至是一个生物体内的全部蛋白质，或许能从中挖掘出某些集体特征，反映出与生物体复杂性相一致的统计趋势。

上述这种宏观与微观之间的联系与经典的统计物理问题类似：从微观出发，观察气体分子的运动，会发现其运动杂乱无章，看似毫无秩序；若是切换到宏观视角，将整个系统用少数几个热力学量（如压强、温度等）来描述，则能发现系统某种整体的演化趋势。如果能从大量的微观个体的演化（即蛋白质的进化）中提取出与系统宏观演化方向（即物种演化）相一致的趋势，就能对生命的起源和进化问题有全新的认识。不过在很长的时间里，由于已知蛋白质结构仍然非常有限，难以真正讨论物种体内蛋白质整体进化趋势。AI 的发展为研究者们提供了全新的强大工具，让上述研究思路能够真正得以实现。

我们基于 AFDB 发展了一套基于物种全蛋白质组结构预测的进化分析方法，对不同生物体内的全部蛋白质进行统计性的研究，而不是只关注特定的蛋白质家族[56]。在研究中，我们从序列、结构、残基接触的拓扑、蛋白质天然态动力学等角度出发，揭示了随着物种复杂度的提高，物种体内的蛋白质呈现出的整体进化趋势。

3.2 结构对比分析

我们首先对不同生物体内、链长相近的蛋白质分子的结构进行了对比分析。尽管选取的这些蛋白质链长接近，但在不同生物体内，这些蛋白质分子的回转半径（radius of gyration）分布却非常不同（如图 2A 所示）。例如，在大肠杆菌（*E. coli*）体内链长约为 250 个氨基酸的蛋白质，回转半径的平均值大约为 20 Å，而在人类（*Homo sapiens*）体内相近链长的蛋白质，其平均回转半径却接近 30 Å。这一结果显示，在两个复杂度差异巨大的物种体内，即使是链长接近的蛋白质分子，其半径分布也有显著的统计差异。回转半径上的这种差别主要与蛋白质结构中结构涨落较大的柔性片段（即无规卷曲结构）相关，因此，该结果还表明对于链长相近的蛋白质，人体内的蛋白质比大肠杆菌体内的蛋白质有更高的柔性。

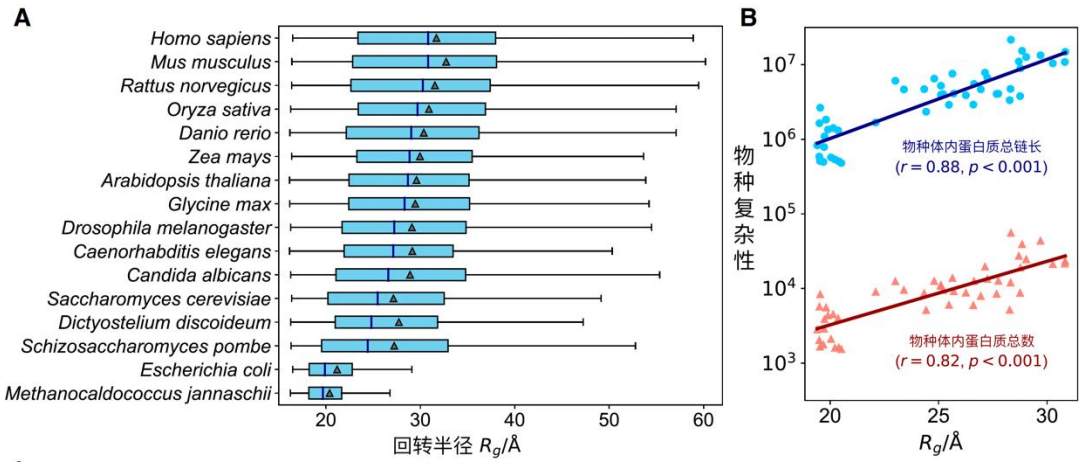

图 2 基于 AFDB 的蛋白质结构对比统计分析。（A）不同物种体内链长 $N \approx 250$ 个氨基酸 ($225 \leq N \leq 275$) 的蛋白质的回转半径分布箱型图。（B）物种复杂度与生物体内一定链长的蛋白质的平均回转半径正相关。

Fig. 2 Statistical analysis of protein structure based on AFDB. (A) For the proteins from different model organisms with similar chain lengths $N \approx 250$ ($225 \leq N \leq 275$), the distributions of the radii of gyration R_g are shown. (B) The measures of complexity are positively correlated with the average radii of gyration R_g of proteins.

对不同物种体内的链长相近的蛋白质结构进行统计，会发现一个粗略的相关关系：随着物种复杂度的提高，该物种体内的蛋白质的回转半径相应地会出现增大的趋势。需要注意的上，这里涉及到了“物种复杂度”的概念，尽管复杂性的数学定义仍有争议[58-59]，在实际操作中，生物学家们往往会对生物体的复杂性引入不同的衡量标准，例如生物体内的各种细胞类型的总数、基因组大小、蛋白质组大小等等。这些度量分别侧重于生物复杂度的

不同层面，它们相互之间往往也是相互关联的[60-61]。在本研究中，我们基于蛋白质组的数据，引入了：（1）一个生物体内所有的蛋白质种类数，以及（2）该物种体内各种不同蛋白质的总链长作为生物复杂性的度量。如图 2B 所示，这两种生物复杂度的度量都与一定链长的蛋白质的回转半径成正比，证明随着物种复杂度的提高，其体内的蛋白质表现出更高的柔性。二级结构分析还显示，这种柔性的差异主要是由无规卷曲片段的长度所带来的，而进一步的统计分析还证明，对于其它链长的蛋白质，相关的统计趋势依然存在，这些结果验证了生物体复杂度与其体内蛋白质的平均柔性成正相关关系的结论。

3.3 拓扑与序列

要更深入地分析蛋白质的结构，除了对蛋白质的二级结构、三级结构进行分析以外，也可以将蛋白质视为氨基酸残基在空间中相互靠近接触而形成的网络，用网络拓扑分析的方法来分析蛋白质的性质。在残基接触网络（residue contact network）中，每个节点代表一个氨基酸残基，如果两个氨基酸残基在空间中的距离小于一定的阈值，则被视作存在连边。我们对该网络的许多拓扑性质进行了分析，其中与蛋白质的物理性质最为相关的度量是网络的同配性（assortativity）。在一个复杂网络上，如果那些度数（连边数）较大的节点倾向于和度数同样较大的节点相连接，那么这样的网络就是同配的。蛋白质的残基接触网络是高度同配的，这是因为构成蛋白质的氨基酸残基可以被分为“亲水”和“疏水”两类，疏水氨基酸残基往往被包埋在蛋白质的内部，形成紧密的堆积，而亲水氨基酸残基则暴露在蛋白质的表面，甚至可能形成高度柔性的卷曲（如图 3A 右所示）。

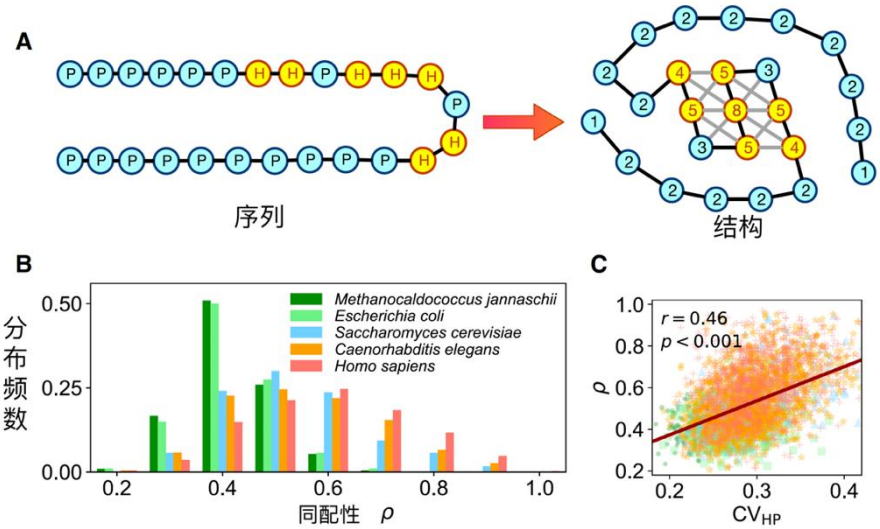

图 3 基于 AFDB 的蛋白质残基接触拓扑与序列特征统计分析。（A）蛋白质的氨基酸序列（HP 模型）决定了其折叠态的氨基酸残基接触网络。其中，图中的 H 与 P 分别代表疏水和亲水（极性）残基，右图节点中心的数字代表其度数，右图中残基接触网络表现出较高的同配性。（B）不同物种（詹氏古细菌、大肠杆菌、酵母、线虫及人类）体内约为 250 个氨基酸的蛋白质残基接触网络的同配性分布。（C）蛋白质残基接触网络的同配性与序列亲疏水分隔度量 CV_{HP} 之间的相关关系。

Fig. 3 Statistical analysis of the features of protein residue contact topology and sequence based

on AFDB. (A) The amino acid sequence of a protein (left) determines the assortative residue contact network (right). In the left subplot, hydrophobic (H) and hydrophilic (P) amino acid residues are represented as the nodes. In the right subplot, the numbers represent the node's degree (number of connections with other nodes). When protein folds into the native structure, the hydrophobic residues tend to aggregate into a densely connected hydrophobic core. (B) For proteins with similar chain length $N \approx 250$ in the five selected organisms (*M.jannaschii*, *E.coli*, *S.cerevisiae*, *C.elegans*, and *H. sapiens*), the histogram of residue contact network assortativity. (C) The scattering plot and the fitted trend line of hydropathy variation CV_{HP} versus residue contact network assortativity.

对 AF2 预测的蛋白质结构进行统计，如图 3（B）所示，对于链长相近的蛋白，在更为复杂的生物体内的蛋白质的残基接触网络显示出更高的平均同配性（例如人体蛋白的残基网络同配性高于大肠杆菌体内蛋白）。该结果与上一节讨论的统计趋势也是自洽的，因为高同配性的残基接触网络使亲水和疏水氨基酸残基在空间上产生了分隔，导致蛋白质中疏水区域产生了更为紧密的堆积，而亲水区域则更为暴露，在蛋白质的表面附近出现了更多高度柔性的无序片段，蛋白质的回转半径也因此增加。

蛋白质残基接触网络同配性的进化趋势也可以从蛋白质的序列进化中得到进一步的验证。研究发现，亲水和疏水氨基酸残基在三维空间中的分隔，与其在序列上的分隔是相关的。为了描述氨基酸序列亲疏水分隔，我们引用了度量 CV_{HP}，它的计算方法如下：首先得到蛋白质各氨基酸残基疏水指数[62]序列，接着对该序列进行长度为 7 的窗口平滑后得到滤波后的疏水指数序列，最后计算该滤波序列变异系数（方差除以均值）。值得注意的是，疏水指数和平滑窗口长度的选取尽管会对数值略有影响，但并不影响最终结论。如图 3（C）所示，蛋白质残基接触网络的同配性与 CV_{HP} 之间存在正相关的关系。换言之，如果蛋白质的序列出现了大段的连续亲水或者连续疏水氨基酸，该蛋白在折叠中将更容易形成高度同配的残基接触网络。研究证明，随着物种复杂度的增加，序列中亲疏水氨基酸的分隔的确有逐步提升的趋势。需要特别指出的是，上述的序列分析完全不依赖于 AF2 的结构预测，而序列分析中所揭示的统计趋势又可以在很大程度上支持结构和拓扑分析的结果。这表明，本研究所观察到的蛋白质进化的统计趋势并非是由结构预测方法所带来的系统偏差，而的确是反映了某种自然趋势。

4 从结构到功能

4.1 理论基础

蛋白质的生物功能是由其结构及其内禀的动力学所决定的。前文所讨论的序列、拓扑和结构变化毫无疑问会影响蛋白质的生物动力学，从而影响相应蛋白的生物学功能。生物系统的"动力学"和"进化"是生命科学研究中两个核心主题。其中，动力学通常涉及的是那些细胞内快速发生的生物化学反应，正是这些反应维持着生命的日常活动，让生物适应环境。而生物的进化建立在随机突变和自然选择的基础之上，需要经年累代的积累。也就是

说，动力学和进化涉及到了生命过程中两个相差巨大的时间尺度。然而，许多研究都表明，动力学和进化这两个看起来相差巨大的过程之间存在着很高的相似性。

有趣的是，在前期研究中，我们发现，尽管蛋白质的动力学和进化本身都涉及到非常高的自由度（组成蛋白质的成百上千个氨基酸都会发生热涨落，也都可以在进化中发生突变），但是通过主成分分析等降维方法，可以将蛋白质的动力学和进化分别约化到较低的维度[63-64]。在这种低维描述中，蛋白质动力学和进化各自的低维子空间依然是高度吻合的。换言之，蛋白质的动力学和进化都体现出鲜明的“准低维”特征，蛋白质功能运动中的构象变化被限制在低维空间，与之相应的，蛋白质的结构进化也主要被限制在这个方向上。蛋白质分子本身的结构模块化对这种低维特征有贡献。“模块化”所刻画的是蛋白质分子中模块内部的残基对接触数占全部接触数的比例，当蛋白质分子的模块化增强时，模块内部的残基接触数占比越高，模块间残基接触数占比越低，蛋白质此时将表现出显著的模块间相对运动。这一结果更深刻地揭示了动力学和进化之间对应关系的起源。

4.2 蛋白质功能进化的统计趋势

随着物种复杂度的提升，物种体内蛋白质的生物功能会产生怎样的统计趋势呢？为便于对海量 AF2 预测结构进行统计分析，我们将蛋白质的运动简化为天然态（能量最低结构）附近的振动。在实践中，可以用弹性网络模型（elastic network model，以下简称 ENM）来描述蛋白质的天然态动力学[65-68]。ENM 又可以分为各向同性的高斯网络模型[65]（Gaussian network model，以下简称 GNM）和各向异性网络模型[68]。本研究主要采用的模型即为 GNM。如图 4 所示，在 GNM 中，构成蛋白质的基本单元（氨基酸残基）被描述为一系列的节点，这些节点以相应氨基酸残基的α-碳原子的坐标所表示，节点之间的连边由蛋白质的天然态结构中的残基接触所决定，当两个节点的距离小于给定的截断距离 R_C 时（通常 R_C 取为 0.7 ~ 1.5 nm），两个节点被视为以弹簧链接，弹簧的弹性系数可以取为固定的常数或者根据氨基酸相互作用强度而选取。节点与弹簧一起构成了弹性网络，这样，蛋白质的振动问题就变成了力学中求解耦合振子的振动模式的经典问题，该问题可以用简正模分析（normal mode analysis）的方法求解[69]。

在 GNM 中，网络的拓扑结构可以由该网络的拉普拉斯矩阵（graph Laplacian）L 描述，该矩阵的各特征值正比于相应振动模式的频率的平方，而与这些特征值相对应的特征向量则描述了相应振动模式的基本形态。在矩阵 L 的特征值谱中，较小的特征值反映的是氨基酸残基低频、大尺度的整体运动，而较大的特征值反映的则是高频的局域运动。蛋白质天然态动力学中的主要运动模式（主成分）主要由矩阵中 L 最小的那些非零特征值所决定，这些运动模式与蛋白质动力学中的长程关联密切相关[70-71]。

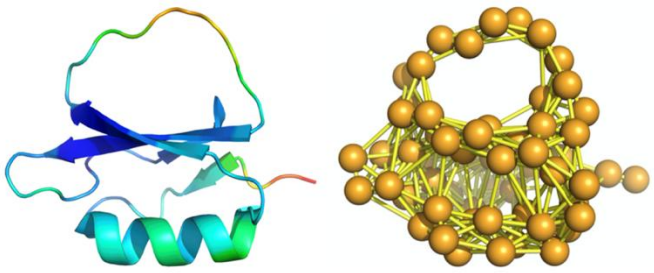

图 4 蛋白质的天然态结构（左）与其所对应的弹性网络模型（右）示意图。

Fig 4. The cartoon graph of a protein's native structure (left) and the corresponding schematic diagram of the elastic network model (right).

对不同物种体内的蛋白质进行简正模分析，可以发现，随着物种复杂度的提升，蛋白质平衡态运动中的主成分比例会发生相应的变化。例如同样链长的蛋白质，在大肠杆菌中，它运动的第 1 主成分和第 2 主成分之间的相对大小较为接近，而在人体中，它的第 1 主成分（对应于第一低频模式）和第 2 主成分（对应于第二低频模式）之间会有较大的区别（如图 5A 所示）。进一步的分析发现，随着物种由简单到复杂，其体内的蛋白质分子的动力学会出现“降维”的趋势，即运动的第 1 主成分会与第 2 主成分之间拉开越来越大的差距，第 2 主成分会与第 3 主成分之间拉开越来越大的差距（图 5B），以此类推。在理论生物学的其它研究中，也观察到类似的“进化降维”现象[72-74]，这种降维趋势使蛋白质特定的功能运动模式变得更加突出（如图 5C 所示）。在复杂度更高的生物体内，有更多蛋白质倾向于沿着特定的主成分方向发生功能运动，这种特定的主成分方向往往与特定的功能有关。这些结果表明，随着物种从简单到复杂，构成生物体的蛋白质呈现出从“通用”到“专用”的统计趋势，高复杂度的生物体内更可能出现高度功能专业化的蛋白质。

蛋白质的“功能专业化”和生物体的复杂性之间的统计相关性与此前大量生物化学实验观察结果一致。许多研究表明，祖先酶往往具有更高的混杂性（promiscuity）[75-79]，即它们不仅可以催化主反应，还具有催化副反应的能力，利用祖先序列重建的方法，有助于设计具有高热稳定性和高混杂性的酶；而在进化中较晚产生的基因或蛋白则表现出较高的无序性[80-81]。

值得一提的是，祖先酶较高的热稳定性及混杂性与祖先物种的低复杂性是相匹配的。复杂性低的生物体的基因组相对较小，其体内所包含的酶的种类也较少。尽管基因组规模小，但高混杂性的酶可以帮助这些生物体实现各种生命活动。相反，较大的基因组可以编码更多的蛋白质，能够发挥高度专业化的功能，应对更复杂和多样化的细胞环境[82]。蛋白质的专业化和多样化使它们能够在更复杂和多样化的细胞环境中发挥作用。因此，复杂的生物体可以更有效地发挥其生物功能，获得适应复杂和多样化的外部环境的可塑性。

生物体的复杂性和组成蛋白质的功能专业化之间的兼容性不是生物体系的某种特例，而是复杂系统中具有普适性的某种规律，即：复杂系统的整体和部分之间是相协调的。当一个系统变得更加复杂时，其组成部分或元素在进化中也会发生属性的改变（例如变得更加可塑或模块化）。不过需要强调的是，本文所讨论的各种“趋势”都是统计性的，因此，具体到每一种蛋白质分子，在定向进化和设计的过程中，都需要具体问题具体分析。在实际应用中，也完全有可能找到一些酶，它们具有更高的柔性，同时也有较高的混杂性。

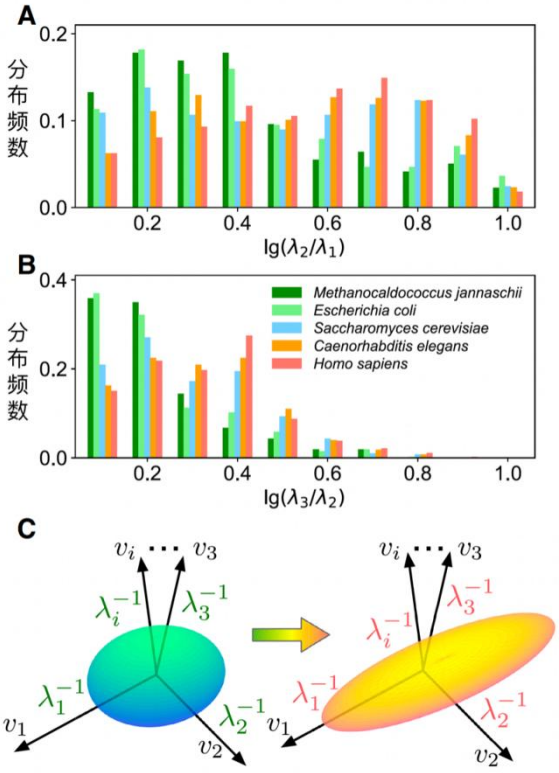

图 5 蛋白质动力学进化的统计趋势。不同物种（詹氏古细菌、大肠杆菌、酵母、线虫及人类）体内链长约为 250 个氨基酸的蛋白质，它们（A）第一和第二低频模式之间特征值的差距与（B）第二和第三低频模式之间特征值的差距。（C）随着物种复杂度的提高，生物体内蛋白质动力学中相应主成分的比例会发生变化，蛋白质结构变化的空间会出现“降维”的趋势。

Fig 5. The statistical trend in the evolution of protein dynamics. For proteins with similar chain lengths $N \approx 250$ in the five selected organisms (*M.jannaschii*, *E.coli*, *S.cerevisiae*, *C.elegans*, and *H. sapiens*), the distributions of the eigengaps (A) between the first and second low-frequency modes, and (B) between the second and third low-frequency modes. (C) As the organismal complexity increases, the proportion of corresponding principal components in protein dynamics changes and there will be a trend of dimensional reduction in the conformational spaces.

5 总结与展望

综上所述，通过利用统计物理方法对 AFDB 中来源于不同物种的蛋白进行对比分析，可以展示出蛋白质进化中的统计趋势，即：随着生物体向更高的复杂性进化，其体内的蛋白质在统计意义上倾向于向更高的灵活性、更高的结构多样性进化，分子本身的功能专一性也在不断增强。本研究中综合采用了多种统计物理方法，包括但不限于关联分析、标度

分析、蛋白质运动模式分析等。这些方法的运用不仅可以帮助从高维的蛋白质结构数据中提取低维的特征（关键突变、关键运动模式等），也能帮助找到适用于不同物种、不同尺寸的蛋白质分子的普适规律，大大提高模型的可解释性。AI 大模型（如 AF2 等）拥有较高的预测精度，而物理模型则具有较高的可解释性，综合这两种模型将有助于精准预测与定量理解序列变化对蛋白质功能的影响。

尽管 AF2 在蛋白质结构预测领域已经取得了重大的成功，但毫无疑问，作为一种基于 AI 深度学习的预测方法，AF2 在技术上有一定的局限性，如深度学习模型的可解释性仍然较低、模型训练成本较高、对较大蛋白质的结构预测的计算时间较长等。除此以外，从分子生物学的角度来看，AF2 的蛋白质预测结构是静止状态的，而蛋白质在发挥功能时是高度动态的，许多重要的生理和病理蛋白质在不同的活性状态下具有非常微妙的构象变化，并且与细胞内外的其他蛋白质结合后，也会产生结构变化。仅仅凭借 AF2 给出的单一结构，还不足以涵盖蛋白质的结构多样性。例如，有研究表明，用 AF2 预测的血红蛋白结构恰好处在实验测得的氧分子结合态和解离态结构之间，这既显示出 AF2 结构预测的准确性，又说明其所预测的静态结构与蛋白质的动力学之间仍有一定的距离[83]。其次，对于近年来越来越受到研究者们关注的内禀无序蛋白（Intrinsic disordered proteins, 简称 IDP），AF2 的结构预测也有重要的缺陷，难以准确预测它们的形态、动力学和相互作用[84-87]。此外，AF2 在预测蛋白质突变、修饰以及配体结合对蛋白质构象的变化方面也有一定的局限性，研究表明，针对单个突变对蛋白质稳定性和功能的影响，AF2 的输出指标与蛋白质稳定性或荧光的变化之间的相关性非常弱，这说明 AF2 目前还不能胜任蛋白质折叠领域中较复杂的问题[88]。尽管目前也有一些研究表明 AF2 能在一定程度上预测突变对蛋白质三维结构的影响[89]，但 AF2 仍然无法应用于分析各种不同类型的突变[90-92]或者磷酸化、乙酰化、甲基化等翻译后修饰[93-95]对蛋白质结构所造成的影响。另外，由于 AF2 的输入信息仅包含蛋白质的序列，因此其无法预测别构调节[96-97]等蛋白质与其他配体相互作用下的结构变化。随着 AI 技术的进一步发展，在未来基于 AI 预测的蛋白质结构的蛋白质组分析，与其他类型的生物信息（如蛋白质与蛋白质的相互作用网络、蛋白质的表达水平、进化速度等）相整合，必将为我们对细胞和生物体的行为和进化提供全新的见解。

参考文献

[1] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects [J]. Science, 2015, 349(6245): 255-260.

[2] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.

[3] Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering [J]. Science, 2018, 361(6400): 360-365.

[4] 夏伟,李慧云. 基于深度强化学习的自动驾驶策略学习方法 [J]. 集成技术, 2017, 6(3): 29-40.

Xia W, Li H. Training Method of Automatic Driving Strategy Based on Deep Reinforcement Learning [J]. Journal of Integration Technology, 2017, 6(3): 29-40.

[5] Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning [J]. Nature, 2020, 577(7792): 706-710.

[6] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873): 583-589.
[7] Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer [J]. biorxiv, 2021, 2021.10. 04.463034.
[8] Dill KA, MacCallum JL. The protein-folding problem, 50 years on [J]. Science, 2012, 338(6110): 1042-1046.
[9] Wüthrich K. Protein structure determination in solution by NMR spectroscopy[J]. Journal of Biological Chemistry, 1990, 265(36): 22059-22062.
[10] Shi Y. A glimpse of structural biology through X-ray crystallography [J]. Cell, 2014, 159(5): 995-1014.
[11] Earl LA, Falconieri V, Milne JL, et al. Cryo-EM: beyond the microscope [J]. Current Opinion in Structural Biology, 2017, 46: 71-78.
[12] Anfinsen CB, Haber E, Sela M, et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain [J]. Proceedings of the National Academy of Sciences of the United States of America, 1961, 47(9): 1309-1314.
[13] Torrisi M, Pollastri G, Le Q. Deep learning methods in protein structure prediction [J]. Computational and Structural Biotechnology Journal, 2020, 18: 1301-1310.
[14] Kuhlman B, Bradley P. Advances in protein structure prediction and design [J]. Nature Reviews Molecular Cell Biology, 2019, 20(11): 681-697.
[15] AlQuraishi M. Machine learning in protein structure prediction [J]. Current Opinion in Chemical Biology, 2021, 65: 1-8.
[16] Jisna VA, Jayaraj PB. Protein structure prediction: conventional and deep learning perspectives [J]. Protein Journal, 2021, 40(4): 522-544.
[17] Pearce R, Zhang Y. Toward the solution of the protein structure prediction problem [J]. Journal of Biological Chemistry, 2021, 297(1): 100870.
[18] Gao W, Mahajan SP, Sulam J, et al. Deep learning in protein structural modeling and design [J]. Patterns, 2020, 1(9): 100142.
[19] Al-Lazikani B, Jung J, Xiang Z, et al. Protein structure prediction [J]. Current Opinion in Chemical Biology, 2001, 5(1): 51-56.
[20] Xiang Z. Advances in homology protein structure modeling [J]. Current Protein & Peptide Science, 2006, 7(3): 217-227.
[21] Abriata LA, Dal Pearo M. State-of-the-art web services for de novo protein structure prediction [J]. Briefings in Bioinformatics, 2021, 22(3): bbaa139.
[22] Bradley P, Misura KMS, Baker D. Toward high-resolution de novo structure prediction for small proteins [J]. Science, 2005, 309(5742): 1868-1871.
[23] Zhao KL, Liu J, Zhou XG, et al. MMpred: a distance-assisted multimodal conformation sampling for de novo protein structure prediction [J]. Bioinformatics, 2021, 37(23): 4350-4356.
[24] Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network [J]. Science, 2021, 373(6557): 871-876.
[25] Humphreys IR, Pei J, Baek M, et al. Computed structures of core eukaryotic protein complexes [J]. Science, 2021, 374(6573): eabm4805.
[26] Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model [J]. Science, 2023, 379(6637): 1123-1130.
[27] Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue coevolution

captures native contacts across many protein families [J]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(49): E1293-E1301.
[28] Uguzzoni G, John Lovis S, Oteri F, et al. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis [J]. Proceedings of the National Academy of Sciences of the United States of America, 2017, 114(13): E2662-E2671.
[29] Zerihun MB, Pucci F, Peter EK, et al. pydca v1.0: a comprehensive software for direct coupling analysis of RNA and protein sequences [J]. Bioinformatics, 2020, 36(7): 2264-2265.
[30] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Advances in Neural Information Processing Systems 30, 2017: 5998-6008.
[31] Yang Z, Zeng X, Zhao Y, et al. AlphaFold2 and its applications in the fields of biology and medicine [J]. Signal Transduction and Targeted Therapy, 2023, 8(1): 115.
[32] Hu L, Salmen W, Sankaran B, et al. Novel fold of rotavirus glycan-binding domain predicted by AlphaFold2 and determined by X-ray crystallography [J]. Communications Biology, 2022, 5(1): 419.
[33] Hutin S, Ling WL, Tarbouriech N, et al. The Vaccinia virus DNA helicase structure from combined singleparticle cryo-electron microscopy and AlphaFold2 prediction [J]. Viruses, 2022, 14(10): 2206.
[34] Jin Y, Fyfe PK, Gardner S, et al. Structural insights into the assembly and activation of the IL-27 signaling complex [J]. EMBO Reports, 2022, 23(10): e55450.
[35] Anishchenko I, Pellock SJ, Chidyausiku, TM, et al. De novo protein design by deep network hallucination [J]. Nature, 2021, 600(7889): 547-552.
[36] Norn C, Wicky BIM, Juergens D, et al. Protein sequence design by conformational landscape optimization [J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(11): e2017228118.
[37] Goverde CA, Wolf B, Khakzad H, et al. De novo protein design by inversion of the AlphaFold structure prediction network [J]. Protein Science, 2023, 32(6): e4653.
[38] Zhang Y, Vass M, Shi D, et al. Benchmarking Refined and Unrefined AlphaFold2 Structures for Hit Discovery [J]. Chemrxiv, 2022, https://chemrxiv.org/engage/chemrxiv/article-details/62b41f0c0bbbc117477285a4.
[39] Friesner, RA, Banks JL, Murphy RB, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy [J]. Journal of Medicinal Chemistry, 2004, 47(7): 1739-1749.
[40] Ma W, Zhang S, Li Z, et al. Enhancing protein function prediction performance by utilizing AlphaFold-predicted protein structures [J]. Journal of Chemical Information and Modeling, 2022, 62(17): 4008-4017.
[41] Rappoport D, Jinich A. Enzyme Substrate Prediction from Three-Dimensional Feature Representations Using Space-Filling Curves [J]. Journal of Chemical Information and Modeling, 2023, 63(5): 1637-1648.
[42] Wong F, Krishnan A, Zheng EJ, et al. Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery [J]. Molecular Systems Biology, 2022, 18(9): e11081.
[43] Liang M, Chen X, Zhu C, et al. Identification of a novel substrate motif of yeast separase and deciphering the recognition specificity using AlphaFold2 and molecular dynamics simulation [J]. Biochemical and Biophysical Research Communications, 2022, 620: 173-179.

[44] Athanasios A, Charalampos V, Vasileios T, et al. Protein-protein interaction (PPI) network: recent advances in drug discovery [J]. Current Drug Metabolism, 2017, 18(1): 5-10.
[45] Rabbani G, Baig MH, Ahmad K, et al. Protein-protein interactions and their role in various diseases and their Prediction Techniques [J]. Current Protein & Peptide Science, 2018, 19(10): 948-957.
[46] Gómez-Marín E, Posavec-Marjanović M, Zarzuela L, et al. The high mobility group protein HMG20A cooperates with the histone reader PHF14 to modulate TGFβ and Hippo pathways [J]. Nucleic Acids Research, 2022, 50(17): 9838-9857.
[47] Varadi M, Velankar S. The impact of AlphaFold Protein Structure Database on the fields of life sciences [J]. Proteomics, 2023, 23(17): e2200128.
[48] Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models [J]. Nucleic Acids Research, 2022, 50(D1): D439-D444.
[49] Nussinov R, Zhang M, Liu Y, et al. AlphaFold, Artificial Intelligence (AI), and Allostery [J]. Journal of Physial Chemistry B, 2022, 126(34): 6372-6383.
[50] Amani K, Shivnauth V, Castroverde CDM. CBP60-DB: An AlphaFold-predicted plant kingdom-wide database of the CALMODULIN-BINDING PROTEIN 60 protein family with a novel structural clustering algorithm [J]. Plant Direct, 2023, 7(7): e509.
[51] Sebastiano MR, Ermondi G, Hadano S, et al. AI-based protein structure databases have the potential to accelerate rare diseases research: AlphaFoldDB and the case of IAHSP/Alsin [J]. Drug Discovery Today, 2022, 27(6): 1652-1660.
[52] Brookes E, Rocco M. A database of calculated solution parameters for the AlphaFold predicted protein structures [J]. Scientific Reports, 2022, 12(1): 7349.
[53] Burnim AA, Xu D, Spence MA, et al. Analysis of insertions and extensions in the functional evolution of the ribonucleotide reductase family [J]. Protein Science, 2022, 31(12): e4483.
[54] Kolesnik MV, Fedorova I, Karneyeva KA, et al. Type III CRISPR-Cas systems: deciphering the most complex prokaryotic immune system [J]. Biochemistry, 2021, 86(10): 1301-1314.
[55] Alvarez-Carreño C, Penev PI, Petrov AS, et al. Fold evolution before LUCA: common ancestry of SH3 domains and OB domains [J]. Molecular Biology and Evolution, 2021, 38(11): 5134-5143.
[56] Tang QY, Ren W, Wang J, et al. The statistical trends of protein evolution: a lesson from AlphaFold database [J]. Molecular Biology and Evolution, 2022, 39(10): msac197.
[57] Barrio-Hernandez I, Yeo J, Jänes J, et al. Clustering predicted structures at the scale of the known protein universe [J]. Nature, 2023, 622(7983): 637-645.
[58] Lloyd S. Measures of complexity: a nonexhaustive list [J]. IEEE Control Systems Magazine, 2001, 21(4): 7-8.
[59] Liu Y, Mathis C, Bajczyk MD, et al. Exploring and mapping chemical space with molecular assembly trees [J]. Science Advances, 2021, 7(39): eabj2465.
[60] Markov AV, Anisimov VA, Korotayev AV. Relationship between genome size and organismal complexity in the lineage leading from prokaryotes to mammals [J]. Paleontological Journal, 2010, 44: 363-373.
[61] Niklas KJ, Cobb ED, Dunker AK. The number of cell types, information content, and the evolution of complex multicellularity [J]. Acta Societatis Botanicorum Poloniae, 2014, 83(4): 337-347.

[62] Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods [J]. Journal of Theoretical Biology, 1968, 21(2): 170-201.
[63] Tang QY, Hatakeyama TS, Kaneko K. Functional Sensitivity and Mutational Robustness of Proteins [J]. Physical Review Research, 2020, 2(3): 033452.
[64] Tang QY, Kaneko K. Dynamics-Evolution Correspondence in Protein Structures [J]. Physical Review Letters, 2021, 127: 098103.
[65] Haliloglu T, Bahar I, Erman B. Gaussian dynamics of folded proteins [J]. Physical Review Letters, 1997, 79: 3090-3093.
[66] Bahar I, Atilgan AR, Demirel MC, et al. Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability [J]. Physical Review Letters, 1998, 80: 2733-2736.
[67] Bahar I, Lezon TR, Yang LW, et al. Global dynamics of proteins: bridging between structure and function [J]. Annual Review of Biophysics, 2010, 39: 23-42.
[68] Atilgan AR, Durell SR, Jernigan RL, et al. Anisotropy of fluctuation dynamics of proteins with an elastic network model [J]. Biophysical Journal, 2001, 80(1): 505-515.
[69] Case DA. Normal mode analysis of protein dynamics [J]. Current Opinion in Structural Biology, 1994, 4(2): 285-290.
[70] Tang QY, Zhang YY, Wang J, et al. Critical fluctuations in the native state of proteins [J]. Physical Review Letters, 2017, 118(8): 088102.
[71] Tang QY, Kaneko K. Long-range correlation in protein dynamics: Confirmation by structural data and normal mode analysis [J]. PLoS Computational Biology, 2020, 16(2): e1007670.
[72] Furusawa C, Kaneko K. Formation of dominant mode by evolution in biological systems [J]. Physical Review E, 2018, 97(4-1): 042410.
[73] Sato TU, Kaneko K. Evolutionary dimension reduction in phenotypic space [J]. Physical Review Research, 2020, 2: 013197.
[74] Sakata A, Kaneko K. Dimensional reduction in evolving spinglass model: correlation of phenotypic responses to environmental and mutational changes [J]. Physical Review Letters, 2020, 124(21): 218101.
[75] O'Loughlin TL, Patrick WM, Matsumura I. Natural history as a predictor of protein evolvability [J]. Protein Engineering Design & Selection, 2006, 19(10): 439-442.
[76] Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective [J]. Annual Review of Biochemistry, 2010, 79: 471-505.
[77] Modi T, Risso VA, Martinez-Rodriguez S, et al. Hinge-shift mechanism as a protein design principle for the evolution of β-lactamases from substrate promiscuity to specificity [J]. Nature Communications, 2021, 12(1): 1852.
[78] Campbell EC, Correy GJ, Mabbitt PD, et al. Laboratory evolution of protein conformational dynamics [J]. Current Opinion in Structural Biology, 2018, 50: 49-57.
[79] van Loo B, Bayer CD, Fischer G, et al. Balancing Specificity and Promiscuity in Enzyme Evolution: Multidimensional Activity Transitions in the Alkaline Phosphatase Superfamily [J]. Journal of the American Chemical Society, 2019, 141(1): 370-387.
[80] Wilson BA, Foy SG, Neme R, et al. Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth [J]. Nature Ecology & Evolution, 2017, 1(6): 0146-146.

[81] Foy SG, Wilson BA, Bertram J, et al. A Shift in Aggregation Avoidance Strategy Marks a Long-Term Direction to Protein Evolution [J]. Genetics, 2019, 211(4): 1345-1355.
[82] Sharov AA. Genome increase as a clock for the origin and evolution of life [J]. Biology Direct, 2006, 1:17.
[83] Buel G R, Walters K J. Can AlphaFold2 predict the impact of missense mutations on structure? [J]. Nature Structural & Molecular Biology, 2022, 29(1): 1-2.
[84] Ward JJ, Sodhi JS, McGuffin LJ, et al. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life [J]. Journal of Molecular Biology, 2004, 337(3): 635-645.
[85] Peng Z, Mizianty MJ, Kurgan L. Genome-scale prediction of proteins with long intrinsically disordered regions [J]. Proteins, 2014, 82(1): 145-158.
[86] Liu Y, Wang X, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction [J]. Briefings in Bioinformatics, 2019, 20(1): 330-346.
[87] Miskei M, Horvath A, Vendruscolo M, et al. Sequence-based prediction of Fuzzy protein interactions [J]. Journal of Molecular Biology, 2020, 432(7): 2289-2303.
[88] Pak MA, Markhieva KA, Novikova MS, et al. Using AlphaFold to predict the impact of single mutations on protein stability and function [J]. PLoS One, 2023, 18(3): e0282689.
[89] McBride JM, Polev K, Abdirasulov A, et al. AlphaFold2 can predict single-mutation effects on structure and phenotype [J]. bioRxiv, 2022, https://doi.org/10.48550/arXiv.2204.06860.
[90] Reynisdottir T, Anderson KJ, Boukas L, et al. Missense variants causing Wiedemann-Steiner syndrome preferentially occur in the KMT2A-CXXC domain and are accurately classified using AlphaFold2 [J]. PLoS Genetics, 2022, 18(6): e1010278.
[91] Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? [J]. Nature Structural & Molecular Biology, 2022, 29(1): 1-2.
[92] Pak MA, Markhieva KA, Novikova MS, et al. Using AlphaFold to predict the impact of single mutations on protein stability and function [J]. PLoS One, 2023, 18(3): e0282689.
[93] Tolsma TO, Hansen JC. Post-translational modifications and chromatin dynamics [J]. Essays in Biochemistry, 2019, 63(1): 89-96.
[94] Samaržija I. Post-translational modifications that drive prostate cancer progression [J]. Biomolecules, 2021, 11(2): 247.
[95] Salas-Lloret D, González-Prieto R. Insights in post-translational modifications: ubiquitin and SUMO [J]. International Journal of Molecular Sciences, 2022, 23(6): 3281.
[96] Li W, Wolynes PG, Takada S. Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins [J]. Proceedings of the National Academy of Sciences of the United States of America, 2011, 108(9): 3504-3509.
[97] Li W, Wang W, Takada S. Energy landscape views for interplays among folding, binding, and allostery of calmodulin domains [J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(29): 10550-10555.