科学通报 2023 年 第 68 卷 第 22 期: 2943 ~ 2950
观 点
《中国科学》杂志社
SCIENCE CHINA PRESS

“后AlphaFold时代”的蛋白质折叠问题

胡昕炜[1,2], 王志珍[1,2], 王磊[1,2*]

1. 中国科学院生物物理研究所, 生物大分子国家重点实验室, 北京 100101;
2. 中国科学院大学生命科学学院, 北京 100049
* 联系人, E-mail: wanglei@ibp.ac.cn

以AlphaFold为代表的使用深度学习人工智能(artificial intelligence, AI)的蛋白质结构预测方法, 已经改变了结构生物学乃至整个生命科学领域. 生命科学的研究似乎已进入“后AlphaFold时代”. 我们都知道, 蛋白质折叠过程是在活细胞中进行的, 但我们对细胞复杂环境内蛋白质折叠过程的理解还非常少. 本文回顾了蛋白质折叠早期的研究历史, 特别是中国科学家的贡献, 总结了体外和体内的蛋白质折叠问题. 体外的经典蛋白质折叠问题指多肽链如何折叠成具有三维结构的蛋白质分子. 自蛋白质折叠的热力学假说被提出以来, 近些年人们建立了几种现象学模型来描述蛋白质折叠过程中的结构变化途径. 体内的蛋白质折叠指细胞内新生肽链从核糖体上合成出来到最终成熟为有活性的功能蛋白质, 包括新生肽链的折叠、修饰、转运、跨膜、组装、分泌等过程, 甚至还包括蛋白质的降解, 这远比体外蛋白质折叠问题复杂. 进入“后AlphaFold”时代, 我们需要打破传统学科界限, 通过跨领域的深度交叉融合来全面地解析蛋白质折叠问题.

胡昕炜 中国科学院生物物理研究所在读博士研究生, 目前从事内质网稳态与应激方向的研究.

王磊 中国科学院生物物理研究所研究员, 博士生导师, 中国科学院大学岗位教授. 主要从事内质网蛋白质氧化折叠、内质网稳态与人类健康等方面的研究.

1 悬而未决的蛋白质折叠问题

如今, AlphaFold等机器学习算法可以于几秒钟内在原子分辨率预测蛋白质结构[1]. 其本质其实也还是依据Anfinsen原理, 在积累至今60多年间测定的蛋白质结构数据上进行预测. 目前, 几乎地球上所有已知的蛋白质结构都已被预测. 科学家甚至能根据需要从头设计自然界不存在的具有新型拓扑结构的有功能的蛋白质[2~4]. 生命科学的研究似乎已进入“后AlphaFold时代”. AlphaFold的发展给生命科学带来了新的可能, 自然也带来了新的问题和挑战[5,6]. AI从已知蛋白质结构中提取样式, 完全绕过了折叠过程, 就好像通过快进到最后一幕而不看前两小时来解读一部电影. 正如Rose等人[7]最近在*Proc Natl Acad Sci USA*发表的评论中所说, “我们知道蛋白质折叠的结果, 但不知道其中发生了什么”.

简单地说, 蛋白质折叠是一个物理化学过程, 其中包含一系列非共价相互作用的变化, 受到一系列物理和化学因素的影响. 打个比方, AlphaFold等工具提供了一份详尽的蛋白质语言句子列表, 而我们现在对其中受物理和化学定律所支配的语法和逻辑还不甚掌握[7]. Moore等人[6]最近在*Science*评论, “解决蛋白质折叠问题意味着从基于物理和化学的第一性原理开始, 根据氨基酸序列精确预测结构.” 然而, 事情并非这么简单, 蛋白质折叠过程是在活细胞中进行的, 细胞内复杂的环境对蛋白质折叠过程的影响是不可忽视的, 而我们对此的理解还非常少. 事实上, 体内存在一个十分复杂并被精准调控的蛋白质稳态网络, 控制蛋白质的合成、折叠、运输和降解[8]. 蛋白质稳态失衡会造成蛋白质错误折叠和许多疾病. 从一个更广阔的视角出发, 不同领域的科学家们对后AlphaFold时代有着不同的期望: 结构生物学家希望AI能辅助预测和解析蛋白质复合物的结构和不同生命活动状态下的构象[9]; 细胞生物学家期待AI能描绘细胞内蛋白质复杂多

样的定位及功能动态[10]; 物理化学家则更痴迷于多肽链折叠成蛋白质三维结构背后的原理[7]; 更多的人们则期望AI能给包括神经退行性疾病等蛋白质错误折叠导致的疾病带来治愈的希望. 显然, 我们在“后AlphaFold”时代必然会对蛋白质折叠问题有突破性认识.

2 蛋白质折叠问题的早期研究

关于蛋白质折叠最早的热力学研究可以追溯到1931年, 中国科学家吴宪在*Chin J Physiol*上发表了世界上第一个蛋白质变性理论, 提出天然可溶蛋白质具有一种紧密的构型(今天谓之构象), 这种构型由分子内的次级键维持. 蛋白质的这种次级键一旦被物理、化学的力破坏, 天然构型就会被打开, 肽链则由有规律的折叠而变为无序、松散的形式, 即发生了变性[11]. 此后的一段时间, 科学家的精力更多集中于研究遗传信息如何从DNA传递到蛋白质. 正是有了蛋白质变性理论作为基础, 20世纪60年代, Anfinsen等人[12]发现还原和变性的牛胰核糖核酸酶A在溶液中能自发地重新折叠成有活力的蛋白质. 根据这一发现, 他们认识到蛋白质的构象仅由其氨基酸序列决定, 多肽链折叠成天然的三维构象是一个自发的过程, 这也就是我们常说的“一级结构决定高级结构”; 他们同时提出了蛋白质折叠的热力学假说, 即“正常生理状态下天然蛋白质的三维结构是整个系统吉布斯自由能的最低状态”[13]. 几乎在同一时期, 1958年, 中国启动牛胰岛素人工合成工作. 胰岛素是由A链和B链两条肽链通过两个链间二硫键连接起来的蛋白质分子, 其中A链还含有一个链内二硫键. 邹承鲁领导的小组在人工全合成工作中首先解决了胰岛素A、B链二硫键拆合问题, 为确定合成路线奠定了坚实的基础, 并提出了“天然胰岛素的结构是所有AB异构物中最稳定的结构之一”的重要结论[14]. 这无疑是对蛋白质氧化折叠乃至蛋白质折叠问题的重大贡献. 值得注意的是, Anfinsen在实验中还发现牛胰核糖核酸酶A重折叠的产量与其浓度等因素有关[12], 而胰岛素拆合实验中邹承鲁找到胰岛素活力恢复需要特定的A/B链比例、温度和pH, 这说明尽管蛋白质的一级结构决定高级结构, 但其能否形成(或者说能形成多少)天然构象的产物还和环境因素相关, 也意味着蛋白质折叠还有动力学问题.

显然, 具有一级序列的多肽链折叠成蛋白质天然构象的过程包含着一系列原子间相互作用的变化和能量变化, 这个过程是否遵循一定的规律呢? 1969年, Levinthal[15]提出如果多肽链的折叠是一个随机过程, 那么一个含有150个氨基酸残基的多肽链就具有10^{300}种可能的构象, 而实际上多肽链折叠成具有天然构象的蛋白质只需要数秒, 这样推算多肽链折叠成天然构象至多也只能尝试10^{8}种构象, 即所谓的“Levinthal悖论”(Levinthal's paradox). 近些年来, 人们已经通过各种实验手段鉴定到很多蛋白质折叠过程的中间体, 说明蛋白质折叠是遵循一定的途径发生的.

3 体外的蛋白质折叠问题

经典的蛋白质折叠问题, 即多肽链如何折叠成具有三维结构的蛋白质分子, 多是用体外实验进行研究的, 包括热力学和动力学问题. Anfinsen[13]提出的蛋白质折叠热力学假说提供了一个看待蛋白质折叠问题的角度. 沿着这个方向, 1995年, Wolynes等人[16]首次系统地阐述了能量景观模型. 该模型认为, 多肽链的折叠是由热波动随机驱动的, 它以能量倾向的方式穿过自由能阱形成天然构象. 自由能阱的宽度代表了构象熵, 即蛋白质分子可以采取的构象, 其深度则代表了整个系统的自由能. 未折叠蛋白位于能阱顶部, 具有很高的构象熵和自由能. 在折叠过程中会形成具有局部自由能极小值的折叠中间体, 天然蛋白质则具有最低的自由能位于能阱底部. 当部分折叠的蛋白质在折叠过程中自身或与其他蛋白质发生非天然接触时, 可能会形成稳定的相互作用并进一步发生聚集, 这类蛋白质聚集体甚至会具有比天然态更低的自由能[17](图1). 显然, 多肽链可以通过多种途径折叠成天然构象, 目前的实验也表明许多蛋白质的确存在多条折叠途径, 一旦折叠条件发生改变其折叠途径也会发生变化, 这种多途径可选的特性为蛋白质在不同环境条件下正确折叠提供了鲁棒性[18].

自Anfinsen开创了蛋白质折叠研究的时代以来, 蛋白质结构测定、预测和折叠动力学的研究取得了许多显著的成绩. 根据近些年来对蛋白质折叠动力学的实验和计算机模拟研究, 人们提出了几种现象学模型来描述折叠过程中的结构变化途径. 一是框架模型. 该模型认为折叠过程中先有局部的二级结构形成, 它们再装配成三级结构[19,20]. 该模型又进一步发展出“折叠子”的概念, 即被称为“折叠子”的二级结构单元在折叠过程中将按照特定的顺序形成三级结构[21]. 二是成核/成核-凝聚模型. 在成核模型中首先部分片段折叠成核, 剩余部分进一步围绕核形成三维结构[22,23]. 成核-凝聚模型则认为在过渡态会先形成一个不稳定的核, 随后二级结构和三级结构的凝聚来稳定整个蛋白质的三维结构[24]. 三是疏水坍缩模型. 该模型认为疏水残基的聚集导致了多肽链起始坍缩, 提供了更小的构象熵, 有助于蛋白质进一步形成天然结构[25,26]. 上述三种模型并非相互排斥, 蛋白质在折叠过程中可能会使用其中的一种或多种机制, 选择哪种机制折叠不仅依赖于蛋白质的一级序列, 还与其溶剂条件有关[18].

目前, 人们已经在试管中研究了许多蛋白质在折叠过程中的构象变化, 包括未折叠态、中间态和天然态之间的变化以及各种状态内部的动态变化[18]. 除蛋白质本身的浓度外, 溶剂的温度、pH、黏度、极性等因素也会影响蛋白质的变性和折叠. 体外实验还发现当折叠过程中需要跨越很大的能障时, 不同构象的蛋白质会同时发生构象变化以跨越很高的活化能, 说明折叠过程中还伴随着不同折叠状态分子之间的协同作用. 此外, 测定蛋白质折叠过程中的焓变、熵变、自

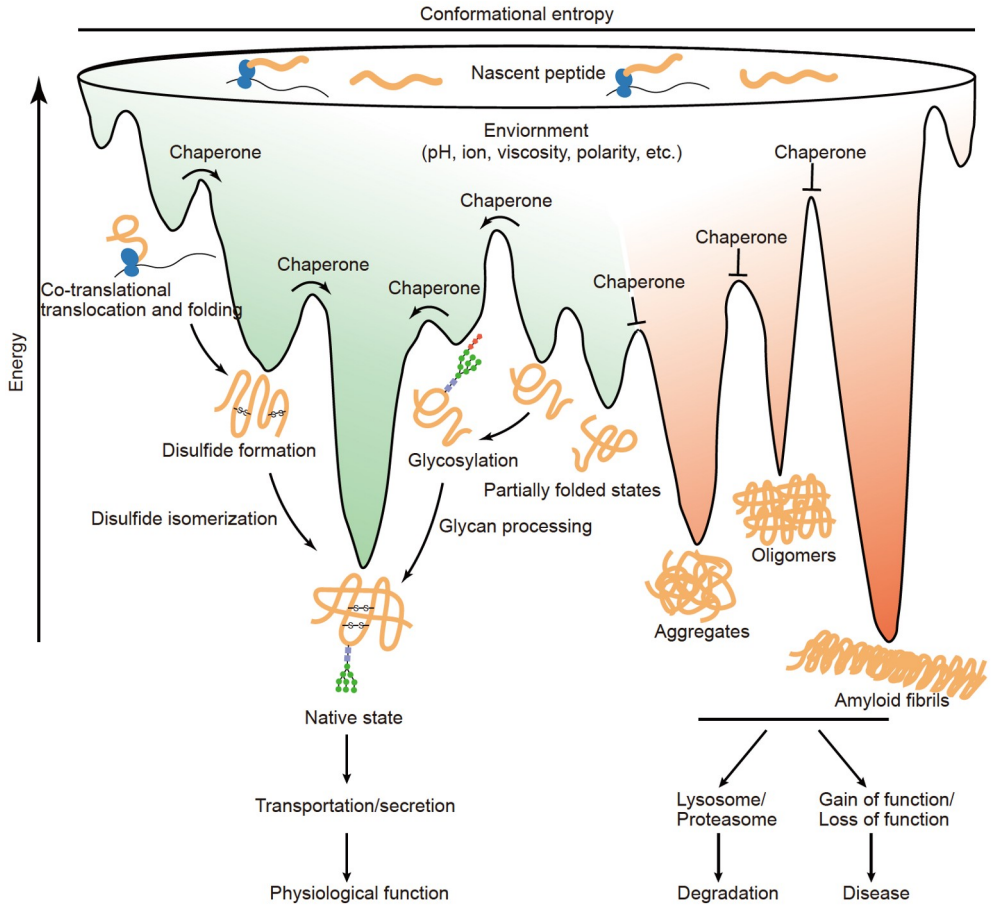

图 1 蛋白质折叠的能量景观模型. 蛋白质折叠是一个由其氨基酸残基的序列决定的自由能逐渐降低的过程, 这一过程既受到细胞内环境因素(如pH、离子、黏度、极性等)的影响, 也受到共翻译转运、翻译后修饰(包括二硫键形成、糖基化修饰等)以及分子伴侣等的共同调控(绿色部分). 错误折叠和聚集的蛋白质也具有很低的自由能, 可以被细胞降解维持其蛋白质稳态; 当正确折叠蛋白质无法产生或错误折叠蛋白质无法清除时将导致疾病发生(红色部分)

Figure 1 The energy landscape of protein folding. Protein folding is determined by the sequence of its amino acid residues, and a process of decreasing free energy. This process is affected not only by cellular environment (such as pH, ion, viscosity, polarity, etc.), but also by co-translational translocation, post-translational modifications (including disulfide bond formation, glycosylation, etc.) and regulated by molecular chaperones (green part). Misfolded proteins and protein aggregates with low free energy will be degraded by cells to maintain proteostasis. Disease occurs when correctly folded proteins cannot be produced (loss of function) or misfolded proteins cannot be cleared (gain of function) (red part)

由能变及热容变等参数对于理解蛋白质折叠的本质至关重要. 目前, 在体外已经可以精确测量蛋白质折叠过程中这些热力学参数的变化, 为上述蛋白质折叠的理论模型提供了新证据[27].

4 细胞内的蛋白质折叠问题

细胞内的蛋白质折叠指新生肽链从核糖体上合成出来到最终成熟为有活性的功能蛋白质. 由于细胞环境的多样性和复杂性, 研究细胞内的蛋白质折叠问题更加具有挑战性. 20世纪80年代, 邹承鲁[28]就提出细胞内新生肽链的折叠是一个共翻译和翻译后折叠并存的过程, 即核糖体合成的新生肽链存在边翻译边折叠的现象, 已经合成的肽链构象还会持续受到新合成肽链的影响, 当多肽链完成翻译后还会发生进一步的修饰或调整, 最终形成天然构象的有活力的蛋白质. 值得一提的是, 1991年中国实施首批30个国家基础性研究重大项目计划(“攀登计划”), 邹承鲁承担了其中的“新生肽链及蛋白质折叠的研究”项目. 2000年, 邹承鲁又详细地论述了“第二遗传密码”, 即遗传信息从蛋白质中氨基酸序列到其空间结构之间传递的规律, 并提出其具有简并性(不同序列对应相同结构)、多义性(类似序列具有不同结构)和全局性(局部序列改变往往影响整体结构)等特征. 同时, 他也提出需要在细胞内考虑蛋白质的折叠问题[29].

目前, 已有很多实验证据表明细胞内蛋白质存在共翻译折叠现象, Christodoulou等人[30]基于此提出了共翻译折叠的能量景观模型. 翻译过程中多肽链的构象熵随其延伸而增加, 这个过程与其寻找自由能更低的折叠态的过程形成竞争, 因

此比较蛋白质的翻译速率和折叠速率非常重要. 对于一些多结构域蛋白质, 其翻译速率大于折叠速率, 折叠发生在新生肽链延伸到相当长度之后, 在动力学上更容易形成部分折叠的中间体(也是自由能阱中的局部极小值), 因此也更容易出现错误折叠. 而对于折叠速率大于翻译速率的新生肽链, 它们在翻译过程中就从N端开始折叠了, 因此比较容易形成天然构象. 翻译速率的调控信息既存在于RNA中(相应密码子的tRNA浓度、mRNA的修饰和二级结构等), 也存在于新生肽链自身, 例如arrest peptides等. 同时, 越来越多的研究表明核糖体对新生肽链的折叠也存在着调控作用[31].

细胞中的蛋白质折叠还包括蛋白质的修饰、转运、跨膜、组装、分泌等过程, 甚至还包括蛋白质的降解. 这些过程中蛋白质构象的动态变化同样也受到细胞内各种因素(pH、离子、黏度、极性等)的共同调节, 还受到与其他生物大分子相互作用的影响, 这些因素都有可能影响蛋白质的折叠途径. 举个例子, 对于大部分膜蛋白和分泌蛋白, 它们需要在内质网经历信号肽剪切和二硫键的形成[32]. 对于含二硫键蛋白质在内质网共翻译转运过程的折叠, 二硫键对单结构域蛋白质和多结构域蛋白质折叠的贡献是不相同的. 单结构域蛋白质的折叠可能早于二硫键的形成[33]; 而对多结构域蛋白质, 先翻译产生的肽段可能会形成带有非天然二硫键的中间体, 它们在随后翻译部分的作用下经异构反应形成天然二硫键并最终完成折叠[34]. 对糖蛋白质来说, 糖链在新生肽链向内质网转位时修饰在特定残基上, 然后才能折叠成有功能的蛋白质. 从动力学角度来看, 糖基化能帮助多肽链形成天然的分子内相互作用, 减少蛋白质在折叠过程中的动态变化, 促进了蛋白质的折叠; 从热力学角度看, 糖基化升高了未折叠状态蛋白质的自由能, 降低了天然构象蛋白质的自由能, 从而帮助蛋白质的折叠[35](图1).

蛋白质的天然构象和聚集体都是自由能低的状态, 因此当蛋白质形成部分折叠的中间体时, 它们既有可能进一步折叠形成天然构象, 也有可能形成蛋白质错误折叠和聚集. 细胞内是一个非常拥挤的环境, 如果细胞处于蛋白质合成非常旺盛的状态, 或者由于遗传和环境的影响, 蛋白质折叠错误的可能性会增加. 错误折叠和聚集的蛋白质会导致一系列loss of function疾病, 如囊性纤维化、戈谢病等; 或gain of function疾病, 如帕金森病、亨廷顿舞蹈病等[8]. 因此, 细胞内进化出一类被称为分子伴侣(molecular chaperone)的蛋白质帮助其他蛋白质的折叠. 自1987年Ellis[36]正式提出“分子伴侣”的概念以来, 其定义不断被完善为“一大类不同的蛋白质, 具有共同的帮助其他大分子结构的非共价折叠、解折叠、组装、解组装的性质, 但不是这些结构在发挥其正常生物功能时的永久组成部分”[37]. 从本质上说, 这是发现了蛋白质的一种新的功能. 分子伴侣中研究最多的是热休克蛋白(heat shock protein)家族. 从作用方式上看, 分子伴侣蛋白通过疏水和静电相互作用与部分折叠的多肽链结合, 以反复迭代的方式完成与多肽链的结合与释放从而避免聚集物的产生, 促进天然构象的形成. 需要强调的是, 包括Ellis在1991年提出的Anfinsen cage模型在内, 几十年来人们公认分子伴侣并不主动帮助蛋白质的折叠, 而是创造蛋白质自发折叠的条件[38]. 从能量角度上看, 分子伴侣发挥作用的理论机制仍存在争议, 有模型认为分子伴侣能使肽链以更加无序的自由能更高的状态释放, 从而使其有更多的机会沿着正确折叠的轨道下落; 也有模型认为分子伴侣发挥功能主要是因为其隔绝了多肽链间的相互作用, 避免其进入错误折叠或聚集的轨道下落, 从而帮助了蛋白质折叠[39](图1). 最新的研究发现分子伴侣素家族的TRiC能够一步一步主动指导微管蛋白的折叠, 这为蛋白质折叠途径和分子伴侣作用机制提供了新的认识[40].

5 对“后AlphaFold时代”的展望

由于生物体本身的高度复杂性, 生命科学的发展尤其是理论方面相对于物理学或化学有所滞后. 不过, 蛋白质折叠领域的研究从观察数据到模式识别再到建立理论模型, 似乎正沿着数个世纪以来科学进步通常遵循的路径. 蛋白质结构数据库(Protein Data Bank)储存了半个多世纪以来解析得到的20余万种蛋白质结构, 深度学习AI则在模式识别方面取得了突破并预测出超2亿个蛋白质结构. 这种科学范式演进的下一步仍在进行中, 我们似乎已经看到了解译蛋白质折叠密码的曙光[7].

与此同时, 我们不得不考虑一个重要的问题, 即蛋白质折叠实际上恰恰是在活细胞中进行的, 如上面讨论的, 蛋白质折叠问题涉及生命活动的不同时间、空间维度, 在不同的细胞环境下, 与细胞内各种分子相互作用并被调控的动态过程. 当前, 研究细胞内的蛋白质折叠问题在实验分析和理论分析方面仍存在一系列困难. 比如, 尽管目前已经能够在细胞环境中研究蛋白质的折叠过程[41], 或是观察蛋白质正确折叠、错误折叠乃至聚集的状态[42~44], 但高时空分辨率地研究蛋白质在细胞内折叠的动态过程仍然是巨大的挑战. 又如, AlphaFold最显著的局限是缺乏蛋白质-蛋白质、蛋白质-配体以及翻译后修饰的信息, 也不能有效预测单点突变对结构造成的影响. 再比如, Anton3分子动力学模拟超级计算机能够实现以每天超过 100 μs的速度模拟一百万个原子[45], 但在活细胞的真实时间和空间尺度上模拟蛋白质折叠图景所需要的算力仍然是难以想象的. 因此, 进入“后AlphaFold”时代, 我们更需要打破传统学科界限, 通过跨领域的深度交叉融合来全面地解析蛋白质折叠问题. AI辅助的结构解析能给细胞内蛋白质全景图的绘制提供强大助力, 而深入揭示细胞内影响蛋白质折叠和功能的各种调节因素又会给AI的训练提升提供更多的反馈信息, 最终我们将全方位描绘细胞内蛋白质动态图景. 随之而来的将会是对蛋白质折叠背后的物理化学原理更深入的理解, 这对于准确预测蛋白质结构, 理性设计乃至生产有特定功能的新型蛋白质, 有效治疗蛋白质折叠异

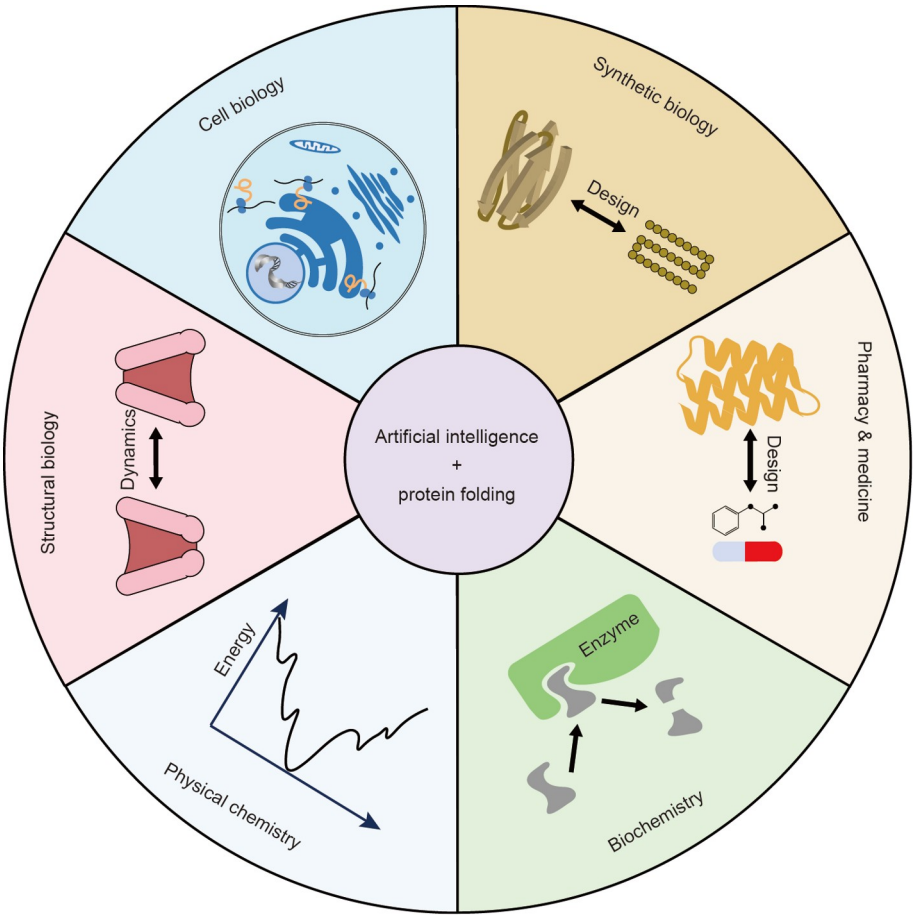

图 2 “后AlphaFold时代”对蛋白质折叠问题的深入理解有望给生命科学各领域带来新的突破
Figure 2 The in-depth understanding of protein folding problem in “post-AlphaFold era” is expected to bring new breakthroughs in various fields of life sciences

常相关疾病等具有重要意义. 最近, 利用机器学习算法理解人群中不同蛋白质突变体导致的疾病表型已初现端倪[46], 未来深度学习AI将不仅仅用于蛋白质静态结构的预测, 还有望实现对细胞内蛋白质动态的预测, 帮助我们更好地理解蛋白质稳态网络的调控机制, 促进精准药物的智能设计. 从科学发展的角度看, 一个成功的蛋白质折叠理论必将为地球上以蛋白质为基础的生命的活动、机制、功能和起源提供更深刻的理解(图2).

致谢 感谢中国科学院稳定支持基础研究领域青年团队计划(YSBR-075)、中国科学院战略性先导科技专项(B类)(XDB37020303)和中国科学院青年创新促进会(Y202028)资助. 谨以此文纪念邹承鲁先生诞辰100周年.

推荐阅读文献

1 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589
2 Huang B, Xu Y, Hu X, et al. A backbone-centred energy function of neural networks for protein design. Nature, 2022, 602: 523–528
3 Madani A, Krause B, Greene E R, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol, 2023, doi: 10.1038/s41587-022-01618-2
4 Yeh A H W, Norn C, Kipnis Y, et al. De novo design of luciferases using deep learning. Nature, 2023, 614: 774–780
5 Callaway E. What’s next for AlphaFold and the AI protein-folding revolution. Nature, 2022, 604: 234–238
6 Moore P B, Hendrickson W A, Henderson R, et al. The protein-folding problem: Not yet solved. Science, 2022, 375: 507

7 Chen S J, Hassan M, Jernigan R L, et al. Protein folds vs. protein folding: Differing questions, different challenges. Proc Natl Acad Sci USA, 2023, 120: e2214423119
8 Balch W E, Morimoto R I, Dillin A, et al. Adapting proteostasis for disease intervention. Science, 2008, 319: 916–919
9 Lutomski C A, El-Baba T J, Robinson C V, et al. The next decade of protein structure. Cell, 2022, 185: 2617–2620
10 Kobayashi H, Cheveralls K C, Leonetti M D, et al. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. Nat Methods, 2022, 19: 995–1003
11 Wu H. Studies on denaturation of proteins. XIII. A theory of denaturation. Chin J Physiol, 1931, 5: 321–344
12 Anfinsen C B, Haber E, Sela M, et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci USA, 1961, 47: 1309–1314
13 Anfinsen C B. Principles that govern the folding of protein chains. Science, 1973, 181: 223–230
14 Du Y C, Chang Y S, Lu Z X, et al. Resynthesis of insulin from pure A and B chains (in Chinese). Acta Biochim Biophys Sin, 1961, 1: 13–25 [杜雨苍, 张友尚, 鲁子贤, 等. 从胰岛素A及B链重合成胰岛素. 生物化学与生物物理学报, 1961, 1: 13–25]
15 Levinthal C. How to fold graciously. In: DeBrunner J T P, Munck E, eds. Proceedings of the Mossbauer Spectroscopy in Biological Systems. Urbana: University of Illinois Press, 1969. 22–24
16 Bryngelson J D, Onuchic J N, Socci N D, et al. Funnels, pathways, and the energy landscape of protein folding: A synthesis. Proteins, 1995, 21: 167–195
17 Clark P. Protein folding in the cell: Reshaping the folding funnel. Trends Biochem Sci, 2004, 29: 527–534
18 Bhatia S, Udgaonkar J B. Heterogeneity in protein folding and unfolding reactions. Chem Rev, 2022, 122: 8911–8935
19 Kim P S, Baldwin R L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. Annu Rev Biochem, 1982, 51: 459–489
20 Udgaonkar J B, Baldwin R L. NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A. Nature, 1988, 335: 694–699
21 Englander S W, Mayne L. The nature of protein folding pathways. Proc Natl Acad Sci USA, 2014, 111: 15873–15880
22 Wetlaufer D B. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci USA, 1973, 70: 697–701
23 Tsong T Y, Baldwin R L, McPhie P, et al. A sequential model of nucleation-dependent protein folding: Kinetic studies of ribonuclease A. J Mol Biol, 1972, 63: 453–469
24 Itzhaki L S, Otzen D E, Fersht A R. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. J Mol Biol, 1995, 254: 260–288
25 Dill K A. Theory for the folding and stability of globular proteins. Biochemistry, 1985, 24: 1501–1509
26 Agashe V R, Shastry M C R, Udgaonkar J B. Initial hydrophobic collapse in the folding of barstar. Nature, 1995, 377: 754–757
27 Rico-Pasto M, Zaltron A, Davis S J, et al. Molten globule–like transition state of protein barnase measured with calorimetric force spectroscopy. Proc Natl Acad Sci USA, 2022, 119: e2112382119
28 Tsou C. Folding of the nascent peptide chain into a biologically active protein. Biochemistry, 1988, 27: 1809–1812
29 Tsou C L. The second genetic code (in Chinese). Chin Sci Bull, 2000, 45: 1681–1687 [邹承鲁. 第二遗传密码. 科学通报, 2000, 45: 1681–1687]
30 Waudby C A, Dobson C M, Christodoulou J. Nature and regulation of protein folding on the ribosome. Trends Biochem Sci, 2019, 44: 914–926
31 Chan S H S, Włodarski T, Streit J O, et al. The ribosome stabilizes partially folded intermediates of a nascent multi-domain protein. Nat Chem, 2022, 14: 1165–1173
32 Wang L, Wang C. Oxidative protein folding fidelity and redoxtasis in the endoplasmic reticulum. Trends Biochem Sci, 2023, 48: 40–52
33 Robinson P J, Pringle M A, Woolhead C A, et al. Folding of a single domain protein entering the endoplasmic reticulum precedes disulfide formation. J Biol Chem, 2017, 292: 6978–6986
34 Kadokura H, Dazai Y, Fukuda Y, et al. Observing the nonvectorial yet cotranslational folding of a multidomain protein, LDL receptor, in the ER of mammalian cells. Proc Natl Acad Sci USA, 2020, 117: 16401–16408
35 Jayaprakash N G, Surolia A. Role of glycosylation in nucleating protein folding and stability. Biochem J, 2017, 474: 2333–2347
36 Ellis J. Proteins as molecular chaperones. Nature, 1987, 328: 378–379
37 Ellis R J. Assembly chaperones: A perspective. Phil Trans R Soc B, 2013, 368: 20110398
38 Ellis R J. Revisiting the Anfinsen cage. Fold Des, 1995, 1: R9–R15
39 Balchin D, Hayer-Hartl M, Hartl F U. *In vivo* aspects of protein folding and quality control. Science, 2016, 353: aac4354
40 Gestaut D, Zhao Y, Park J, et al. Structural visualization of the tubulin folding pathway directed by human chaperonin TRiC/CCT. Cell, 2022, 185: 4770–4787.e20
41 Dhar A, Girdhar K, Singh D, et al. Protein stability and folding kinetics in the nucleus and endoplasmic reticulum of eucaryotic cells. Biophys J, 2011, 101: 421–430

42 Gupta R, Kasturi P, Bracher A, et al. Firefly luciferase mutants as sensors of proteome stress. Nat Methods, 2011, 8: 879–884
43 Blumenstock S, Schulz-Trieglaff E K, Voelkl K, et al. Fluc-EGFP reporter mice reveal differential alterations of neuronal proteostasis in aging and disease. EMBO J, 2021, 40: e107260
44 Tang S, Wang W, Zhang X. Direct visualization and profiling of protein misfolding and aggregation in live cells. Curr Opin Chem Biol, 2021, 64: 116–123
45 Shaw D E, Adams P J, Azaria A, et al. Anton 3: Twenty microseconds of molecular dynamics simulation before lunch. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2021. 1–11
46 Wang C, Balch W E. Bridging genomics to phenomics at atomic resolution through variation spatial profiling. Cell Rep, 2018, 24: 2013–2028.e6

Summary for “‘后AlphaFold时代’的蛋白质折叠问题”

The protein folding problem in the “post-AlphaFold era”

Xinwei Hu[1,2], Chih-chen Wang[1,2] & Lei Wang[1,2*]

[1] *National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;*
[2] *College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China*
* Corresponding author, E-mail: wanglei@ibp.ac.cn

Currently, the protein structure prediction method, using deep learning artificial intelligence (AI) represented by AlphaFold, can predict protein structures at atomic resolution within seconds. To date, nearly every known protein structure on Earth has been predicted. Scientists can even design functional proteins *de novo* with novel topologies that do not exist in nature. AlphaFold has changed structural biology and the entire field of life sciences. Life science research seems to have entered the post-AlphaFold era. Although the development of AlphaFold has created new possibilities in life sciences, it comes with new problems and challenges. Protein folding is a physicochemical process that involves changes in a series of non-covalent interactions, which are affected by a range of physical and chemical factors. However, protein folding occurs in living cells, and our understanding of the process within the complex cellular environment remains limited. Notably, there is a very complex and precisely regulated proteostasis network *in vivo*, the imbalance of which results in protein misfolding and associated conditions, such as neurodegenerative diseases. Scientists in different fields have different expectations of the post-AlphaFold era. Herein, we review the early history of research on protein folding, particularly the contributions of Chinese scientists. We summarize protein folding problems *in vitro* and *in vivo*. The classic protein folding problem *in vitro* refers to how polypeptide chains fold into three-dimensional protein molecules. Following the thermodynamic hypothesis of protein folding proposed by Anfinsen, several phenomenological models have been constructed to describe alternative pathways through which a structure can undergo protein folding. The energy landscape theory has provided a framework for understanding the protein folding problem at a quantitative level. *In vivo* protein folding refers to the intracellular synthesis of nascent peptide chains from amino acid linking on ribosomes to the final maturation into functional proteins, involving protein folding, modification, transmembrane transport, assembly, secretion, and even degradation, and is considerably complex than the *in vitro* counterpart. The dynamic changes of protein conformation within the cellular environment are affected by various factors and regulated by interactions with other biomacromolecules. Thus, protein folding *in vivo* is a highly dynamic process affected by the organism’s different temporal and spatial dimensions. As we enter the post-AlphaFold era, we should break the boundaries of traditional disciplines and comprehensively analyze the protein folding problem using a multidisciplinary approach. In the future, deep learning AI will not only be used to predict the static structure of proteins but is also expected to realize the prediction of protein dynamics in cells, which will facilitate better understanding of the regulatory mechanisms of proteostasis networks and promote the intelligent design of precision drugs. From the perspective of scientific development, a successful protein folding theory may provide a deep understanding of the activities, mechanisms, functions, and origins of protein-based life on Earth.

AlphaFold, artificial intelligence, energy landscape, protein folding, proteostasis, structure

doi: 10.1360/TB-2023-0233