

蛋白质三维结构的形成和进化

王 大 成

(中国科学院生物物理研究所)

现在, 我们对于蛋白质的生物合成已经有较多的了解, 但是, 对于蛋白质如何形成特征的三维结构却所知很少。由于可取独特的三维构象是蛋白质的基本属性和重要特点, 使得这个问题具有重要的理论和实践意义。另一方面, 在生命的起源与进化中, 蛋白质的起源和进化占有重要地位; 通过蛋白质的进化来更深入和更精确地了解生命的进化, 是一条十分有效的途径。近二十年来, 由于蛋白质结构知识的积累和丰富, 使我们对这两方面的问题都有了比过去深入得多的了解, 对诸如分类和进化这样古老的生物学领域产生了深刻的影响, 同时成为象蛋白质结构预测这类课题的良好生长点。在这里, 蛋白质一级结构知识发挥了重要作用。因限于篇幅, 本文重点介绍与三维结构有关的问题。

一级结构决定高级结构

——蛋白质折叠的自发性

在细胞内, 蛋白质的合成是在核蛋白体上, 通过从肽链的N端开始, 逐个加上由遗传密码控制的序列上的氨基酸来完成的。(图1) 这样合成的是一线性的具有开放构象的多肽链。它

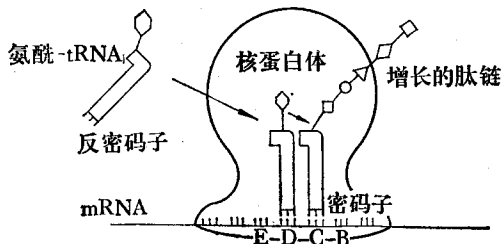


图1 多肽链在核蛋白体上的合成

如何转变成天然蛋白质的特征三维结构? 这是今日生化领域内的基本问题之一, 通常称之为蛋白质的折叠。所谓蛋白质折叠 (protein folding) 是指一个新合成的 (或变性的) 多肽链的开放构象全部转化为天然蛋白质的特征三维构象的过程。

关于蛋白质折叠的研究, 到目前为止已经肯定的是: 折叠是在自然条件下自发地进行的。所有实验事实都证明, 在给定环境中 (溶剂、pH、离子强度、温度以及其他成份的存在), 天然蛋白质是一种相对于所有自由度来说, 总体系统吉布斯自由能极小的状态, 即天然蛋白质的构象是由其组成原子间的相互作用, 亦即由其氨基酸顺序决定的。简洁地说, 只要环境条件合适, 多肽链的一级结构唯一地决定蛋白质的高级结构。

蛋白质折叠自发性的主要实验证据, 是关于体外变性蛋白在适当条件下可以自行复活的研究。所谓复活, 是指变性蛋白质重新恢复变性前的天然三维结构状态, 从而使丧失了活性重新恢复过来。许多年前就知道, 蛋白质可以经历可逆变性。我国著名生化学家吴宪首先指出, 蛋白质的变性是由于其空间结构的破坏, 但早期关于这方面的知识都是很表面和粗糙的。六十年代以来, 美国的 Anfinsen 小组用牛胰核糖核酸酶 (RNase) 作材料, 对蛋白质被变性试剂还原后的重新折叠进行了系统和精细的研究。他开始这一工作时, 遗传密码的本质尚未真正了解, 还不知道蛋白质进行正确的折叠是否还需要专门的遗传“指令”。他抓住问题的中心开展了全面研究, 这就是: 酶的活性和结

构是完全地和唯一地决定于它的一级结构还是需要一些其它信息?他以所获实验结果为基础,结合核酸酶的互补性重组研究,建立了关于蛋白质折叠的一系列概念,提供了蛋白质折叠自发性的令人信服的证据。他也因此获得 1972 年诺贝尔奖。

RNase 含有 124 个氨基酸,4 对—S—S—键,其二 硫键的配对方式提供了一个定量估计重组和复活的指标。一般说来,含有 $2n$ 个—SH 的多肽链形成 j 个—S—S—的可能组合数由下式给出

$$N_{2n}^j = \frac{(2n)!}{2^j(2n-2j)!j!}$$

对 RNase 来说, $n = 4, j = 4$, 配对方式就有 105 种。因此,考察在蛋白质折叠过程中—S—S—的配对情况,就可以对折叠的正确性有一个精确的概念。将 RNase 放在 8M 脲和大量 β -巯基乙醇中,在温和的碱性条件下经受变性,这时其四对—S—S—还原,一系列物化参数表明多肽链变为无规卷曲。然后,透析去除脲,在有 O_2 存在情况下,—S—S—重新形成,酶的活性接近完全恢复,复活蛋白与天然 RNase 有同样的结晶和相同的 X-射线衍射花样。同时,肽图谱证明,—S—S—的位置与原来完全一样。这表明,酶不仅可以自发折叠,而且仅选定了 105 种—S—S—可能配对方式中的一种。这意味着,正确的肽链折叠和—S—S—配对方式,是蛋白质在溶液中最稳定的排布,三级结构的获得仅为热力学所控制。为了进一步检查这一设想,Anfinsen 等人做了一个很有趣的实验。他们将

还原的(—SH 态)RNase 在 8M 脲中重氧化,使 8 个巯基全部氧化成二硫键以后,产物只有 1% 的酶活力。经验查,原来—S—S—配对是错的,产物是“错乱型”RNase。在透析去除脲后,将这种错乱型 RNase 放入水溶液,再加痕量的 β -巯基乙醇,它又会转变成天然的 RNase。在这里, β -巯基乙醇的存在量并不足以使酶的一—S—S—转变成相当量的一SH,而是一种促进分子二硫键重排的催化剂。这一过程如图 2 所示。关于变性蛋白的复活,还有其它许多实验研究,其结果在基本的方面都与 RNase 的一致。所有这些实验研究,已使下述结论为人们普遍接受:在给定环境中,蛋白质完整的三维结构是由其氨基酸顺序唯一地决定的。

人工合成蛋白质的成功进一步证实了折叠的自发性。1965 年我国科学工作者首次用化学的方法全合成了具有全部生物活力的结晶牛胰岛素。自那时以来又有许多蛋白质用人工方法合成了。人工合成蛋白质,就是将氨基酸按它在天然蛋白质中的顺序,用化学方法连接起来。在特定的条件下,它具有全部生物活力,并可按天然方式结晶,表明有正确氨基酸序列的多肽链可以自发地形成天然的空间构象,而无须另外的控制信息。

此外,关于蛋白质生物合成的研究至今没有发现任何专门控制和指导折叠的“指令”或类似的东西。这从另一个侧面支持一级结构决定高级结构的结论。

关于蛋白质折叠的机理,如从一维肽链转化为三维折叠的方式和过程,决定一个专一的三维结构所必须的作用性质和大小等,目前正在许多实验和理论研究,但至今肯定下来的东西并不多。

蛋白质的进化

蛋白质形成的另一头,是它的历史的起源和进化。近些年来,用比较氨基酸的顺序来建立进化树,探测蛋白质的结构和进化关系,已成为一个普遍的方法。现在,由于大量蛋白质晶体结构的测定,有可能将这种比较扩展到许多

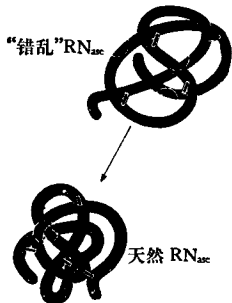


图 2 “错乱型”核糖核酸酶的自发转换
矩形连接代表—S—S—键,网状区域表示活性中心

蛋白质的三维结构，给予我们关于蛋白质进化的更深入的知识。过去的物种进化和分类学都以生物的形态学和解剖学为基础，蛋白质结构知识的应用将它推进到分子水平。尽管其中的许多问题尚有待进一步的实践和积累来解决，但其已经得到的成果和展现的前景是十分令人注目的。对那些用别的方法难以研究的问题，如微生物的进化、代谢途径的进化，这种研究是特别有意义的。

早期蛋白质晶体学的一个惊人发现是，肌红蛋白与血红蛋白亚基结构极其类似。以后，有许多不同种属来源的蛋白质结构得到测定，如鸡和人的溶菌酶、马、东方狐鲑、金枪鱼的细胞色素 C，人、马、昆虫、八目鳗、血寄生虫和羽扇豆根瘤菌的血红蛋白，以及抹香鲸、海豹、金枪鱼的肌红蛋白等等，它们都表现出三级结构的极大类似性。图 3 是从细菌到哺乳动物马共六个种属肌红蛋白和血红蛋白亚基的结构，可以明显看到，它们基本上是相同的。此外，还有 α 和 γ 胰凝乳蛋白酶、弹性硬蛋白酶和胰蛋白酶这样一些丝氨酸属水解酶，其三级结构也表现出明显的类似。在每一族蛋白质中，氨基酸顺序都有改变，有插入、缺失，有的变化量竟高达 80%，但是三维结构的类似性仍然保持着。这表明它们各自都是从共同的祖先分化而来，而且显然，在长期进化历程里，氨基酸顺序的变异比三级结构更快。这种进化方式是趋异性进化 (divergent evolution)。趋异进化的事实指出，在关系微弱的蛋白质进化比较中，来自结构同源性的信息一定是更加基本的。

在细菌丝氨酸属蛋白酶——枯草杆菌蛋白酶与胰凝乳蛋白酶之间，表现出完全不同的另一种结构类似。这两种酶的活性位置，无论就催化和结合底物所包含的氨基酸残基，还是就其相应原子的空间排布来看，几乎都是完全一样的。但化学研究表明，这两种酶完全不相关，而且除活性位置外，这两个酶分子总的三维折叠是完全不相同的。这表明，哺乳动物和细菌经由不同的进化途径，在完全不同的分子骨架中，发展出相同的催化机构来，这是一种趋同性

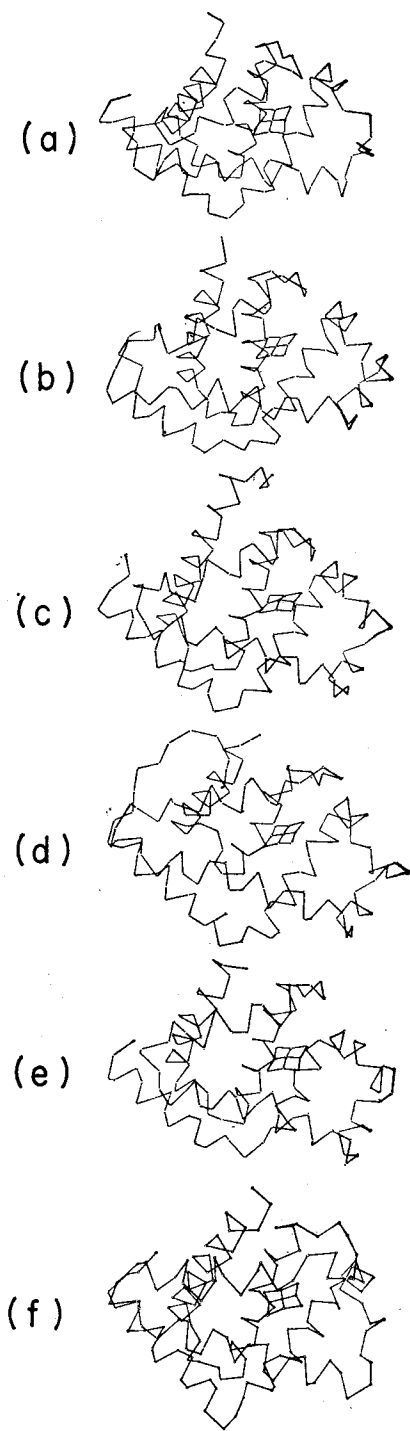


图 3 不同种属血红蛋白蛋白结构的类似性
(a) 马高铁血红蛋白的 α 链 (b) 马高铁血红蛋白的 β 链 (c) 抹香鲸肌红蛋白 (d) 八目鳗血红蛋白 (e) chironomus 血红蛋白 (f) glycera 血红蛋白

进化 (convergent evolution) 的明显例证。
仔细的比较还可以给出一些进化历史的信

息。如比较细菌细胞色素 C_2 与马细胞色素 C 的结构(特别在血红素区域),发现它们很类似。这表明,尽管系统发育树显示出细菌与高等生物在 26 亿年前就分开了,但哺乳动物的亚细胞器线粒体(细胞色素 C 位于其内)与细菌有共同的起源。据估算,导致产生丝氨酸蛋白酶的分化的基因复制,发生在 13 亿年前。

更一般的结构类似性,表现在脱氢酶系列中。结构已知的所有脱氢酶(乳酸脱氢酶, S-苹果酸脱氢酶, D-甘油醛-3-磷酸脱氢酶, 肝乙醇脱氢酶)都有一个共同的可与烟酰胺腺嘌呤二核苷酸结合的结构域,称为“核苷酸结合域”,具有相同的三维结构,如图 4 所示;而分子的另一半由所谓“催化域”组成,它们对不同的脱氢酶是很不相同的。这就是上述各种脱氢酶都有结合 CoI 的共同性,而各自又有不同的专一性的结构基础。同时,在糖酵解途径中的一些磷酸化酶(磷酸甘油酸激酶、腺苷酸激酶、己糖激酶),也都有类似的核苷酸结合域存在。这显示出它们有共同的起源。但是,在完全没有结合核苷酸能力的枯草杆菌蛋白酶和磷酸甘油酸变位酶中,也有类似的结构域存在。这里存在的是进化关系还是热力学原因? 这是一个值得研讨的十分有趣的问题。

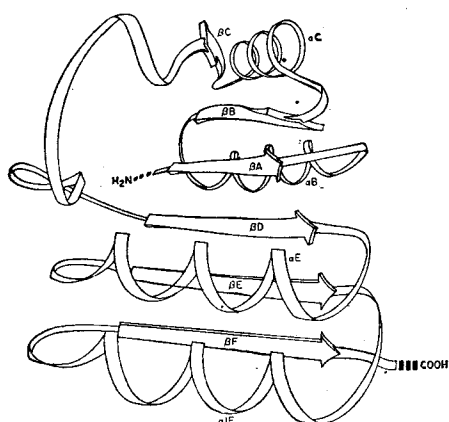


图 4 脱氢酶的“核苷酸结合域”

是否可将蛋白质按结构族分类? 结构族是否存在有限的基本类型? 以及由蛋白质追溯到基因的进化,这些十分有意义的问题,都有赖于三维结构知识的进一步丰富和研究工作的不断

深化来回答。

从氨基酸顺序预测蛋白质结构

一方面,由于折叠自发性的研究证实了蛋白质三级结构的所有信息都寓于氨基酸顺序中;另一方面,大量三维结构的阐明逐步揭示出蛋白质结构的一些规律,使人们已经掌握了一些结构规则可供运用;于是,近年来在蛋白质结构研究中兴起了一个新领域:从氨基酸顺序预测蛋白质的三维结构。

蛋白质结构预测的主要困难在于,即使数量不太大的氨基酸组成的蛋白质,其可能构象的数量也常常达到天文数字。如何在这极大数量的可能构象里求得天然结构? 例如,一个由 100 个氨基酸残基组成的肽链的折叠,至少需要调节 300 个参数,即使每个参数限定只有二个值,也会有 2^{300} 多种可能的构象。如果随机搜索其能量最稳定的构象,一次尝试只用 10^{-13} 秒,将所有可能构象搜索一遍所需的时间也远远超过地球的年龄。从运算能力来说,有人估计,将全球现有的计算机联合起来,也不敷用。另一方面,生化研究指出,在体内情况完全不是这样,合成多肽链的自发折叠只需在 10^{-1} — 10^{-3} 秒内完成。显然,必然存在某种动力学过程赋予肽链折叠以一些确定的途径,以极大地限制可能构象的数目。

所以,预测蛋白质结构既要从热力学观点考虑,又要研究动力学过程。目前,热力学考虑是通过各种粗细程度不等的能量计算来实现的。动力学实验数据很难获得,迄今只给出了 N 端折叠、预先成核等少数原则。由于上述两方面都存在一些基本困难,现在发展了许多基于已知结构资料的统计方法。它们对大量已知结构进行统计处理,求得一些统计规律,用于预测。

从氨基酸预测二级结构的工作最多,成效也较显著。这方面已经提出的方法在 15 种以上,概括说来可分两类。一是统计法。在顺序和三维结构都已知的蛋白质中,每种氨基酸都处于一类二级结构中。利用实际观测到的分

布, 导出特定氨基酸类型采用某种二级结构的概率参数, 如 C_α (螺旋)、 C_β (折迭)、 C_i (回折), 然后用于预测。二, 有少数方法不依靠统计数据, 而以立体化学标准为基础。如 Lim 考虑侧链的疏水性和大小, 推出可以形成 α 螺旋和 β 链的残基的有利类型。前一类方法只能考虑短程相互作用, 但应用简便, 较有成效; 后一类方法较复杂, 但因为强调疏水和亲水残基的重要性, 在一定程度上考虑了长程相互作用, 所以也很受重视。二级结构的预测结果是大量的, 但在较长时间里只引起人们很小的兴趣。这是因为, 它们要么测定的是已知结构, 只能是“马后炮”; 要么在处理结构时将其拆得很零散, 缺少实用价值。1975 年对腺苷酸激酶二级结构的预测取得了很好的结果。图 5 中的阴影区表示预测有二级结构出现的地方, 上面的矩形方框代表实测结构中出现二级结构的地方。从图上可见, 预测与随后的结构测定有很好的吻合。这一成功大大增加了人们对预测结构的信心。但是, 用同样的方法去预测 T_4 噬菌体溶菌酶的结构, 成效却不大, 说明方法并未成熟到具有一般指导性的地步。

对于三级结构, 已有从进化关系预测成功的实例。如, 基于 α -乳清蛋白的氨基酸顺序与鸡蛋白溶菌酶近似, 已从后者的晶体结构正确地预测出 α -乳清蛋白的肽链折叠; 肌钙蛋白 C 的结构, 连同其对称性要求, 也已从肌浆蛋白的晶体结构预测出来。这种途径虽有效, 但仅限

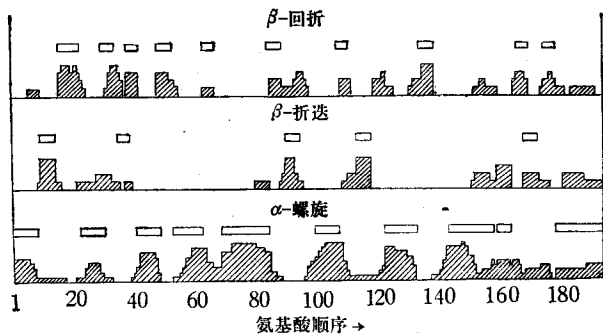


图 5 腺苷酸激酶二级结构的预测
阴影部分是预测有二级结构的区域, 上排的矩形表示实测结构中有二级结构存在的区域

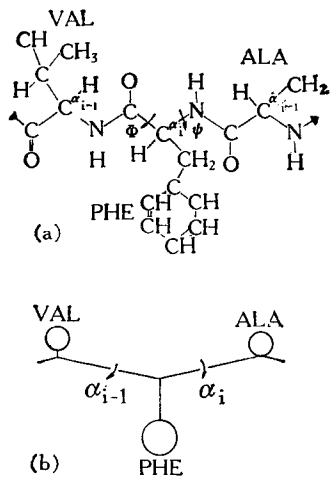


图 6 蛋白质结构的简化方案

(a) 是肽链中的两个氨基酸残基 Ala 和 Val;
(b) 是将结构简化后的模式

于进化上相关的蛋白质。通过一般的方法预测一定蛋白质的三级结构, 只是最近的事。目前, 限于一些结构已知的小蛋白质, 而且途径也极不相同。现行各种方法都对蛋白质结构加以极大的简化, 将每一氨基酸残基用两个中心来代表, 即 C_α 原子和在侧链质心处的一个球; 相互作用只考虑这些中心, 并仅允许在绕相邻 C_α 原子联线的扭角 α 有一个自由度, 如图 6 所示。这些简化是限于现有方法和技术水平, 当然这给预测结果带来很大的限制。这些方法已经用于下列各蛋白质的结构预测: 胰蛋白酶抑制剂 (58 个残基, 一段含 10 个残基的 α 螺旋)、肌浆蛋白 (108 个残基, 6 段 α 螺旋), 红氧还蛋白 (53 个残基, 没有螺旋)。可以用预测结构的坐标偏离实际观测结构的坐标的均方根误差 (r. m. s.) 来衡量这些方法的成效。上述三个蛋白质结构预测的 r. m. s. 值, 胰蛋白酶抑制剂为 $5.7 \text{ \AA}$, 肌浆蛋白为 $8.5 \text{ \AA}$ 以下, 红氧还蛋白为 $4-6.5 \text{ \AA}$ 。此外, 对肌红蛋白, 烟草花叶病毒外壳蛋白以及免疫球蛋白结构域中的 β 夹层的预测, 都取得了相当的成功。英国 Blundell 最近对一完全未知的松弛肽 (relaxin) 结构进行了预测, 原子坐标已存入蛋白质数据库, 当然其正确性有待证实。

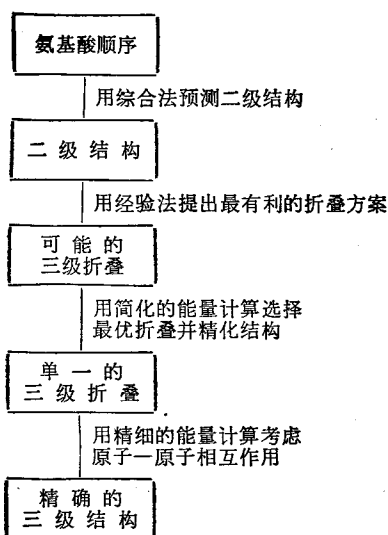


图7 从氨基酸顺序预测蛋白质三级结构的可能途径

尽管单一的预测方法还会发展和改进，但是，最有前途的还是各种可用方法的综合运用。图7示出了这类综合法的一个概貌，表明如何

从氨基酸顺序开始，通过二级结构的分析和三级结构的预测，逐步进到简化的、精细的能量计算，求得精确的天然结构。当然，离开一般地、完善地预测蛋白质的三维结构还有遥远的路程。但重要的是，第一步已经迈出。由蛋白质结构的复杂性引起的根本性障碍已被冲破，许多理论的和实验的研究者正满怀希望进行活跃的探索。有人认为，蛋白质的结构预测比地震预测的难度更大，前途未可乐观。但试回想一下，25年前，利用核酸和蛋白质有限的结构化学知识成功预测了DNA的双螺旋结构和蛋白质的 α 螺旋与 β 折迭结构，为分子生物奠定了一块基石，那么就完全有理由相信，在未来的25年里，运用今后积累起来的蛋白质三维结构知识，从氨基酸顺序预测蛋白质结构取得突破性进展是完全可能的。

科技消息

最近一次西德细胞生物学会会议上，美国Yale的科学家报道，他们将外源病毒DNA注射到小鼠胚胎，从而改变了小鼠的遗传性。新受精的鼠卵接受病毒基因，并且永久地渗入到遗传物质中，因而渗入的病毒基因，在胚胎发育过程中同样可以控制蛋白质合成。科

学家预言，如果这个技术与“试管婴儿”——解决体外受精后着床技术结合起来，则解决人类遗传病将有可能实现。

摘自 New. Sci. 87, (1218), 80

基因库和“密密”库

“密密”(Meme) = Mimemes——模仿的单位——意指的单位“密密”可以从通过通讯过程复制它们自己从一个脑到另一个脑。“密密”信息同样是化学编码

的。最近纽约大学学院研究表明，似乎也与儿茶酸胺族有关，但作用过程完全不一样。

摘自 Sci. 80, 1 (6), '80

更正

应作者要求，作如下更正：

一、《核酸蛋白检测仪》(1980年，第4期)一文附录中滤色液的合成与配方工作，有王尔文同志参加，应在文中予以致谢。

二、《用离子交换树脂法制备药用弹性蛋白酶》(1980年，第5期)一文中“10公斤胰脏平均可得48.6克药用弹性蛋白酶”，应改为“4.86克”。