DOI:10.13376/j.cbls/2010.11.003

文章编号：1004-0374(2010)12-1129-09

对称蛋白质序列与结构关系研究

肖 奕*，冯建辉，黄延昭
(华中科技大学物理学院生物分子物理与模拟研究组，武汉 430074)

摘 要：进化的观点认为，蛋白质结构的对称性是基因复制和融合的结果，但是由于在长期进化过程中的氨基酸突变，绝大多数现有的蛋白质序列都失去了这种直观的重复性特征。该文简要地回顾了国际上发展的寻找蛋白质序列中重复片段的方法，重点介绍了作者自己提出的分析蛋白质序列和结构对称性的方法以及在蛋白质对称结构形成机理方面的初步工作，并系统分析了各类对称折叠子的序列与结构关系，发现它们的序列都具有隐含的与结构相同的对称性，或者说序列的对称性决定结构的对称性。

关键词：结构对称性；序列对称性；相似矩阵；基因复制；进化
中图分类号：Q51　　**文献标识码：**A

Studies of sequence-structure relations of symmetric proteins

XIAO Yi*, FENG Jian-hui, HUANG Yan-zhao
(Biomolecular Physics and Modeling Group, Department of Physics, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: From the view of evolution, the symmetry of protein structure is the result of gene duplication and fusion. However, due to the mutation in evolution, most of proteins have lost their repetitive signals in sequences. In this paper we will briefly review the methods of detecting repeats in protein sequences. In particular, we shall introduce our methods of detecting sequence and structure symmetries of proteins as well as our studies in the mechanism of the formation of symmetric structures of proteins. We investigated the sequence-structure of different types of folds and found their hidden sequence symmetries as the structures or sequence symmetry encodes structural symmetry.

Key words: protein; structural symmetry; sequence symmetry; similarity matrix; gene duplication; evolution

许多蛋白质分子的三级结构呈现出明显对称性(严格讲是准对称性)，例如，在前十类常见的蛋白质折叠子(Fold)中有六类具有对称性[1]。因此，我们自然会问：蛋白质形成这种对称性的生物学和物理意义是什么；它们是如何进化而来的。图 1 是最常见的一种蛋白质折叠子(β/α)-barrel。除个别例外，这个家族的所有成员都是酶[2]，其活性位点位于连接中心八条链 C 端的 loop 围成的区域[3]。关于这些$(\beta/\alpha)_8$-barrel 酶是如何进化而来的已有大量研究，有些研究倾向于它们是通过趋同进化演变成一个稳定的折叠子，而另外一些则认为它们是由同一个祖先趋异进化的结果[2]。一般认为$(\beta/\alpha)_8$-barrel 是典型的单结构域蛋白[1]，但是实验证明，把它拆成片段后，这些片段仍能自动聚集形成稳定有活性的酶，这表明它可能由几个亚结构单元组成[4,5]。来自海栖热袍菌(*Thermotoga maritime*)的咪唑甘油磷酸合成酶(imidazoleglycerol phosphate synthase)是组氨酸合成中的双酶复合物。在原核生物中，它由 HisF

收稿日期：2010-07-22；修回日期：2010-10-28
基金项目：国家自然科学基金项目(30525037，30870678)
* 通讯作者：E-mail: yxiao@mail.hust.edu.cn

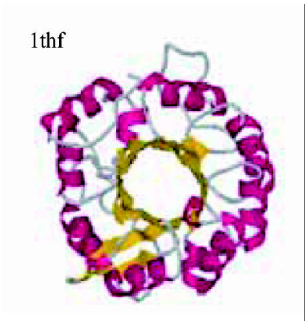

图1 海栖热袍菌的HisF(PDB ID: 1thf)的三级结构是(b/a)$_8$-barrel

(合成酶亚单元)和HisH(谷氨酸酶亚单元)按1:1构成[6]。X射线晶体结构显示海栖热袍菌的HisF的结构(PDB ID: 1thf)是$(\beta/\alpha)_8$-barrel[7]。如果把催化关键残基(Asp 11 和 Asp 130)在序列上对齐，发现HisF由两个可以重合的亚结构组成(HisF-N 和 HisF-C)[6]。HisF-N由N端的四个(β/α)单元组成，而HisF-C由C端的四个(β/α)单元组成，也就是HisF的结构显示两重对称性。因此，人们假定HisF是由$(\beta/\alpha)_4$半桶结构通过基因复制和融合而来的，也就是说，HisF的三级结构是一个具有二重对称性的$(\beta/\alpha)_8$-barrel折叠子。在组氨酸合成中HisF前面的一个酶，N'-((5'-phosphoribosyl)-formimino)-5-aminoimidazol-4-carboxamid ribonucleotide (ProFAR) isomerase (HisA)，也有同样的特性。为了验证这个假设，Lang等[7]和Höcker等[8]把HisF-N和HisF-C在大肠杆菌中单独纯化和表征，发现它们都可以独立折叠成HisF中的天然构象，但是没有催化活性，也就是说单独的HisF-N或HisF-C都没有HisF的功能。但是，如果在体内共表达或在体外共折叠，它们可以聚集成一个有催化活性的HisF-NC复合物，因此完整的HisF具有了HisF-N和HisF-C都没有的新功能。这说明$(\beta/\alpha)_8$-barrel可能是由半桶形状的祖先通过复制和融合进化而来的。实际上，很多证据显示现有的具有复杂结构的蛋白质分子可能是通过基因复制(重复)和融合进化而来的[9]。

关于蛋白质结构对称性(重复性)是同型多聚体通过基因复制和融合进化的假设要追回到20世纪70年代McLachlan的研究工作。他在十年间发表了一系列文章研究了从纤维蛋白、单结构域蛋白到多结构域蛋白基于基因复制的进化问题[10-18]。他和后来许多研究者[19]都认为基因复制不仅是多结构域蛋白进化的重要机制，也是单个结构域本身进化的重要途径。由于长期进化过程中大量的氨基酸突变，对称蛋白质分子中由这种基因复制产生的直观重复特征在序列上大多丢失，现在只能在结构水平直接看到。前面提到的前十种常见的蛋白质折叠子中六种具有内部结构对称性[1]，它们分别是four-helix bundle、ferredoxin、β-trefoil、$(\beta/\alpha)_8$-barrel、jelly-roll和immunoglobulin (Ig)折叠子。其他典型的对称折叠子还有β-propeller和β-sandwich等。具有up-and-down拓扑的Four-helix bundle类折叠子现在仍然有以同型四聚体和二聚体的形式出现，提供了通过基因复制进化的直接证据。然而，大部分的对称折叠子没有发现同型多聚体，而且在ferredoxin、β-trefoil和$(\beta/\alpha)_8$-barrel折叠子中也只是极少数成员的氨基酸序列中可以看到明显的重复子序列，Ig fold和Jelly-roll折叠子用传统的方法则没有发现任何序列具有内部重复性[20]。

另一方面，Anfinsen著名的核糖核酸酶变性和复性实验研究表明[10]，每个蛋白质的形状是由它的氨基酸序列决定的。氨基酸序列中包含的这种形状信息称为第二遗传密码。如果蛋白质的三级结构由一级结构决定，那么对称蛋白质的氨基酸序列应该编码三级结构的这种对称性，但这些蛋白质序列的氨基酸排列表面看似近似随机，这是一个矛盾。进一步，不同的氨基酸序列又是如何编码相同的对称性结构。因此，要证明对称折叠子是通过基因复制和融合而来的，首先需要揭示它们的氨基酸序列所隐含的重复性或对称性(关于序列对称性的定义见下节)。

1 蛋白质序列对称性分析方法

人们提出了不同的方法探测蛋白质序列中的重复片段[21-31]，最早的方法是Gibbs和Mcintyre[21]在1970年提出的Dotplot以及后来Junier和Pagni[22]在2000年提供的在线服务软件Dotlet。这两种方法基本原理一致，是用二维的点图来显示蛋白质序列内部的重复片断：横轴和纵轴放上同一条蛋白质序列，在残基相同的位置处打上一个点，对角线表示自身对齐，而平行于对角线的点线就表示蛋白质序列中的重复片段。由于该方法原理简单，算法容易实现，在很多大型蛋白质分析软件中都集成了该算法或改进的版本，如Antheprot[23]。更完善的方法是基于序列比对的方法，其中有Heger和Holm[24]提出的RADAR(Rapid Automatic Detection and Align-

ment of Repeats)算法、George 和 Heringa[25]提出的 REPRO(REPeat PROtein)和Szklarczyk和Heringa[26]提出的TRUST(Tracking Repeats Using Significance and Transitivity)算法、Gruber 等[27]提出的 REPPER (REPeats and their PERiodicities)方法[27]和Soding等[28]提出的 HHREP(de novo protein REPpeat detection by HMM-HMM comparison)方法。另外一类方法是基于 Eckmann 等[29]1987 年提出的重现图方法(Recurrence Plots)。经过了二十多年的发展，重现图方法及其衍生方法在生物序列重复性分析方面也得到广泛应用[30]。2005 年在德国、2007 年在意大利已成功举办了两次 RP 方法国际学术研讨会。另外，还有基于傅立叶变换的方法，如 Turutina 等[31]提出的准周期算法。特别是，Rackovsky[19]用傅立叶变换方法分析了表征蛋白质构型的序列信号，发现 TIM barrel 和 Ig 折叠子的氨基酸序列存在表征不同折叠子的特征。

然而，以上方法大多是从同源的角度，而不是从编码结构的方面来确定蛋白质序列中的重复片段，因此一般给出的重复片段比较短，也就是重复片段长度总和远小于序列的总长。而与蛋白质结构对称性对应的重复片段其长度总和应覆盖整个序列。为了区别起见，我们把这种与结构对称性对应的序列重复性称为序列对称性。为了寻找蛋白质序列的对称性，我们基于两个蛋白质，如果它们序列的氨基酸一致性超过 25%，它们的结构就相似的事实[32]，提出了一种分析蛋白质序列隐含对称性的方法[33]。我们假定，蛋白质分子中的两条子序列，如果它们序列上氨基酸一致性在 25%以上就认为它们相似。换句话说，如果两条子序列的氨基酸一致性超过 25%，就认为它们具有相似的三级结构。这种方式定义的“相似性”看起来和日常观念不符合，但是从蛋白质序列决定结构的角度来说则是合理的，因为两个蛋白质结构相同并不需要它们的序列完全一样。具体地，对一条长度为 N 的蛋白质序列，考虑其中任意一条长度为 d 的子序列 $X_i = x_i x_{i+1} \ldots x_{i+d-1} (1 \le i \le N-d+1)$，并在剩余的$(N-d)$条长度相同的子序列中找出有多少和它相似，也就是有多少和它的氨基酸一致性超过 25%。我们把该数目用 $S(d, i)$表示。对不同的 d 进行这种统计，就得到相似矩阵 S。相似矩阵 S 能够直观显示序列的对称性(图 2 左图)。为了利用所有局部序列相似信息来确定序列的对称性，我们进一步引进了相似矩阵 S 的 Pearson 关联分析，也就是比较相似矩阵 S 中各个子列矩阵之间的关联强度。例如，如果蛋白质序列的前半段和后半段为重复片段，子列矩阵 S_1 和 S_2 间的关联系数会远远大于其他子列矩阵间的关联系数，其中子列矩阵 S_1 是 S 的 1 到 N/2 列，S_2 是 N/2+1 到 N 列。图 2 右图就是左图相似矩阵 S 的 Pearson 关联系数图。Pearson 关联图能够精确确定对称片段的位置和对称度，并且由于子列矩阵包含不同长度和不同起点氨基酸的信息，因此能够显著提高分析的灵敏度(图 2 右图)。这使得我们成功地应用于一些序列相似度很低或很复杂的蛋白质序列的对称性分析，例如 Ig fold。

图 2 给出了前面实验研究的具有$(\beta/\alpha)_8$-barrel折

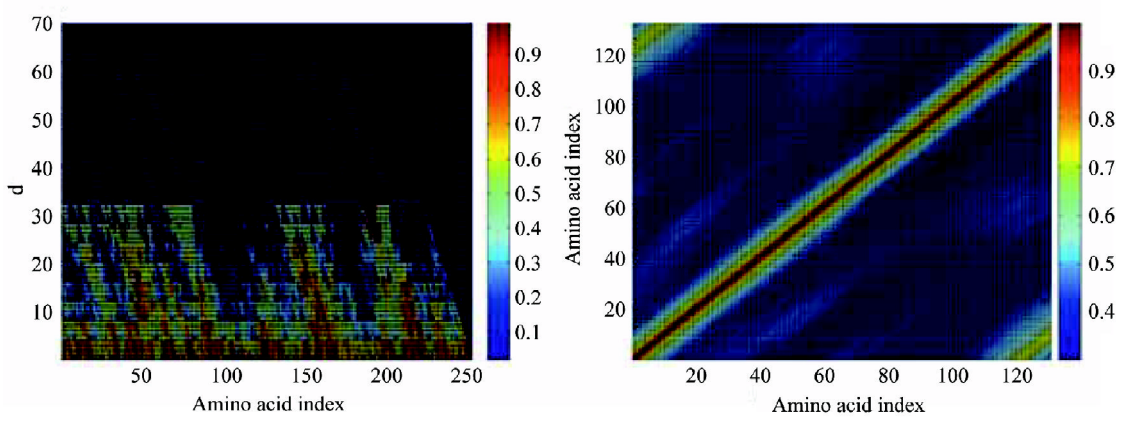

图2　HisF序列的相似矩阵S和它的Pearson关联系数图

左图纵坐标 d 为序列片段长度。关联系数大小由彩条颜色表示，该图显示片段 1~126 和 122~247 之间有很强的关联，Pearson 关联系数 r 达到 0.7476，可以认为它们相似

叠子的 HisF 序列的相似矩阵 S 和它的各种可能子列矩阵之间的关联系数图[34]。可以看出子序列 1~126 和子序列 122~247 具有很强的关联，因此可以认为它们是重复片段：

```
1
MLAKRIIACLDVKDGRVVKGSNFENLRDSGDPVELGKFYSEIGIDELVFLDITASVEKRKTMLEL
SQAVVVAIDAKRVDGEFMVFTYSGKKNTGILLRDWVVEVEKRGAGEILLTSIDRDGTKSGYDTEM
122
66
VEKVAEQIDIPFTVGGGIHDFETASELILRGADKVSINTAAVENPSLITQIAQTFGSQAVV
IRFVRPLTTLPIIASGGAGKMEHFLEAFLAGADAALAASVFHFREIDVRELKEYLKKHGVN
187
```

这和基于结构的序列比对给出的结果几乎完全一致[7,8]。图 2 显示序列有明显的 2 重对称性。而别的方法(如 Radar 和 Trust)只能给出比较短的重复片段。例如，Radar 只找到下面的重复片段：

```
43
GIDELVFLDITASVEKRKTMLELVEKVAEQIDIPFTVGGGIHDFETASELILRGAD
GAGEILLTSIDRDGTKSGYDTEMIRFVRPLTTLPIIASGGAGKMEHFLEAFLAGAD
164
```

这表明我们的方法能够探测与结构相关的氨基酸序列隐含的对称性。为了证明这种隐含对称性的普遍性，我们对 SCOP 数据库中各种类型的蛋白质对称结构进行了系统的分析，典型的有 α-helix boundle 类、β-Trefoil 类、β-Propeller 类、β-barrel 类、β-Prism 类、β sandwich、Ig fold、Jelly roll 等[32-40]。这些结构类的序列都存在对称性，而且与结构的对称性一致。图 3 是其中一类对称蛋白质的序列对称性分析结果。

特别是我们揭示了 Ig 和 Jelly roll 折叠子序列的对称性，它们是目前空间结构对称而用传统方法没有找到任何序列对称性的两种主要对称结构类。例如，对具有 Ig 折叠子结构的 Fab NC10 的 Ig 的 kappa L 链(1al4L)[34]，图 4 中的相似矩阵显示它的序列具有二重对称性，Pearson 关联分析则更清楚地显示出子序列 1~47 和子序列 55~101 间有很强的关联，Pearson r 为 0.7382：

```
1
DIELTQTTSSLSASLGDRVTISCRASQDISNYLNWYQQNPDGTVKLLIYY
HSEVPSRFSGSGSGTDYSLTISNLEQEDIATYFCQQDFTLPFTFGGGTAA
55
```

但 Radar 和 Trust 方法都没有发现这些重复子序列。原因可能是这些方法基于序列同源性，所以只能探测很相似的子序列。我们的方法只要求序列相似性不小于 25%，因为我们是要找那些与结构有关的重复子序列。更重要的是，我们方法中的 Pearson 关联分析考虑了不同长度相似片段的信息，因此是一种 profile-profile 的比较，能够探测到低相似性的重复片段。

我们的工作一方面说明了蛋白质的序列对称性编码其结构对称性的普遍性，特别是成功地确定了 Ig fold 和 Jelly-roll fold 序列的对称性。另一方面也说明了具有同一种对称结构的蛋白质，虽然它们序列不同(相似性小于 30%)，但都具有相同的序列对称性。这在对称性的层面上解释了为什么不同蛋白质序列具有相同的三级结构。

由于我们提出的序列对称性判断方法原理的一般性，我们把它推广到了蛋白质结构对称性的定量判定[41]。目前判定分子结构的对称性大多还仅依赖于眼睛，不仅人为性强，而且对复杂的分子无法判断。基于序列对称性分析中的相似矩阵和 Pearson 关联分析方法的思想，我们只要用子结构相似性分析替代子序列相似性分析，就能够定量地判断蛋白质结构对称性。子结构相似性用 α 碳原子坐标间的均方偏差来度量。因此，我们的方法的一个优点是

可以用相同的方法来分析序列和结构的对称性。

2　蛋白质结构对称性的形成机制

序列对称性与结构对称性之间存在普遍的对应关系只是“是什么”和“怎么样”的问题，更重要的是要弄清楚“为什么”，这才能真正理解序列对称性如何决定结构的对称性，也就是其物理机

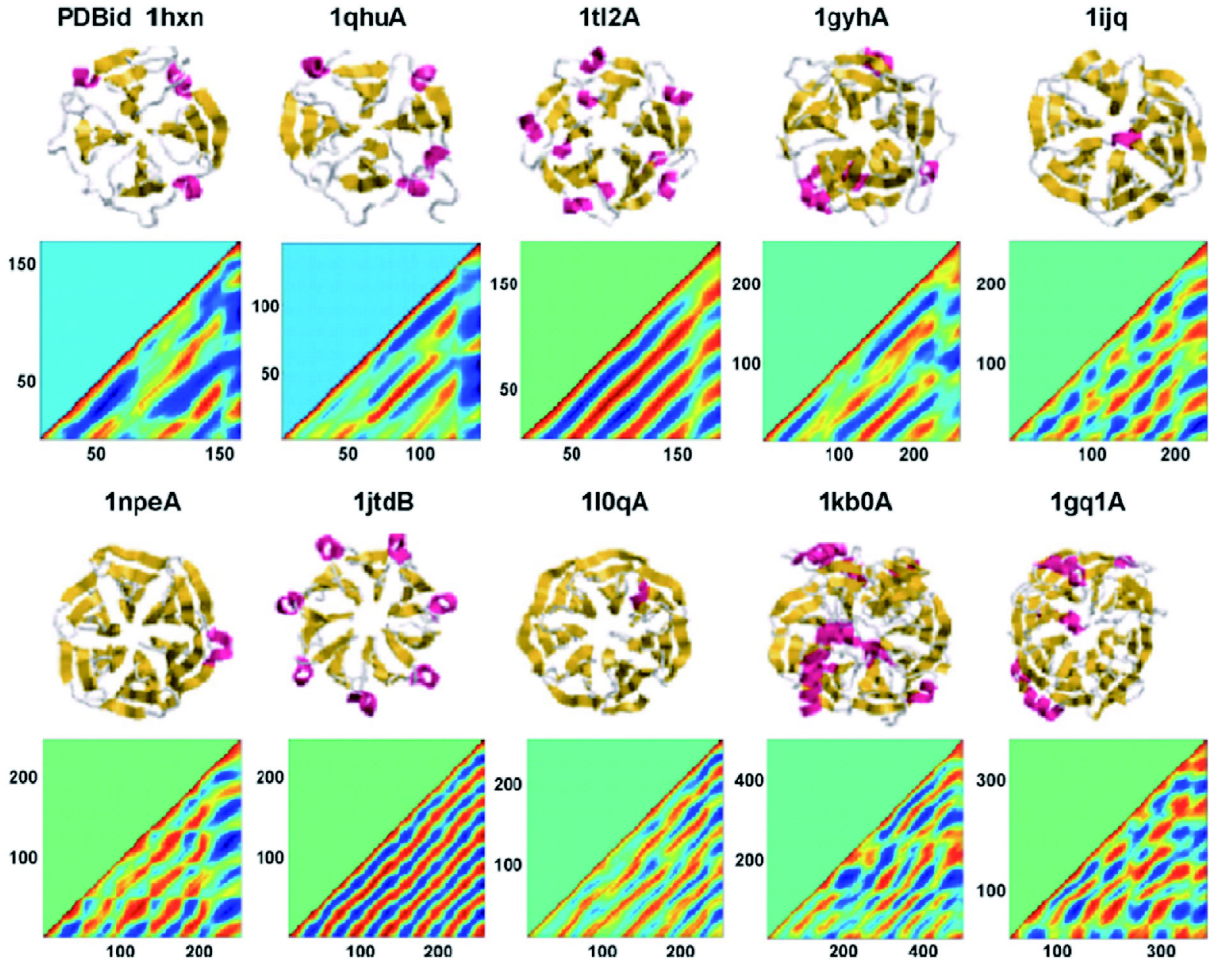

图3　Propeller类典型蛋白质的结构(上)和Pearson关联系数图(下)

注：各图坐标表示的量同图 2

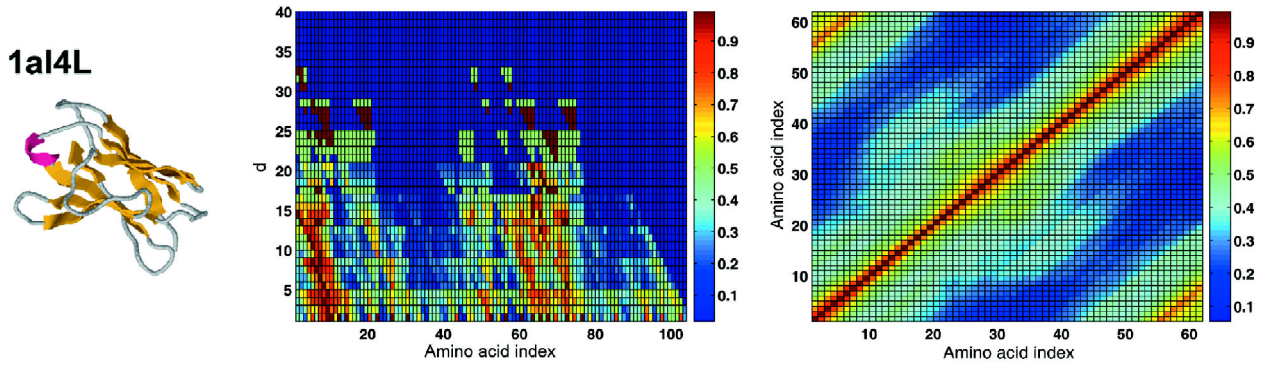

图4　Ig fold蛋白质序列结构与序列对称性

左图为一典型结构(PDB ID: 1al4)，中图为其序列的相似矩阵(纵坐标 d 为序列片段长度)，右图为 Pearson 关联系数图。利用 Pearson 关联图很容易确定对称片段的位置和相似程度。关联系数大小由彩条颜色表示。该图显示片段 1~47 和 55~101 之间有很强的关联，Pearson 关联系数 r 达到 0.7382。

制。只有这样我们也才能完全解决序列编码结构问题和结构预测问题，并用于指导新蛋白质的设计。上述研究表明，如果定义两条蛋白质序列氨基酸一致性超过25%，它们就相似，就可以显示大多数蛋白质序列隐含的对称性。因此，我们可以假定蛋白质对称结构的形成是序列上对称分布的少数氨基残基起关键作用。我们提出了一种直接从能量角度来定义残基间相互作用的方法来确定这些关键氨基酸[42]，并研究了典型的具有7重对称的propeller结构域。它的7个对称子结构的序列长度都约为40个氨基酸(称为WD repeats)，它们有约15% (5~6个) 的氨基酸残基相同且呈对称分布。计算结果显示这些氨基酸都具有很强的平均相互作用能，确实表明它们在蛋白质结构稳定中起重要作用。对β-Trefoil类结构域的分析也表明存在对称分布的关键氨基酸，它们与其他氨基酸有很强的相互作用，在对称结构的稳定中起关键作用(图5)。另一方面，为了研究非关键氨基酸的作用，我们也研究对称蛋白质的序列对称性和内部残基相互作用对称性与结构对称性的关联性(图6)。结果发现，内部残基相互作用对称性与结构对称性关联性相对更强。这意味着非关键氨基酸(子序列中不一致的氨基酸)对形成对称结构也有贡献。因此，氨基酸序列在对称结构形成中的作用还有待进一步的研究。

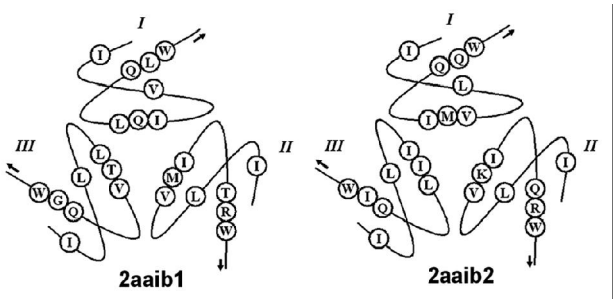

图5 Ricin Toxin B (PDB ID: 2aaib)二个b-Trefoil结构域的关键氨基酸在序列和结构中对称性分布示意图
三个trfoil单元顺时针排列。箭头表示beta链的方向。

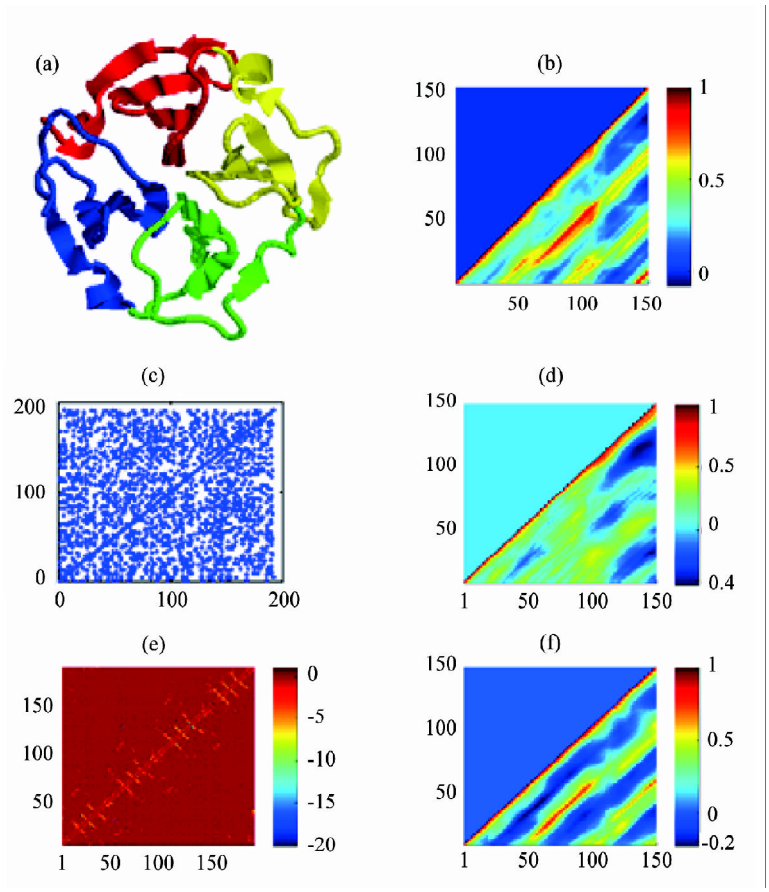

图6 Four-blade beta-propeller蛋白质(PDB ID: 1pexA)
(a)，结构卡通图；(b)，结构相似矩阵的Pearson关联系数图；(c)，序列dotplot；(d)，序列相似矩阵的Pearson关联系数图；(e)，内部残基相互作用图；(f)，内部残基相互作用相似矩阵的Pearson关联系数图。(b)-(f)图的横坐标和纵坐标都是氨基酸序号。

我们更进一步的研究发现[43]，蛋白质结构的对称性不仅仅是由其本身的序列决定的，还和其他蛋白质分子或结构域的相互作用有关，蛋白质与外界的相互作用会影响其序列的对称性。虽然蛋白序列是决定结构的主要因素，环境效应同样重要，这符合 Anfinsen 的蛋白质天然结构是热力学自由能最小假设的原义。系统分析 β-Trefoil 结构域序列的对称性表明，虽然所有序列都显示三重对称性，但不是所有序列的对称程度都一样。我们认为这和每个结构域是否与其他结构域或蛋白质相互作用有关。通过分析这些结构域是否和其他结构域或蛋白质相互作用及其强度，发现序列的不对称程度是和蛋白质分子与外界相互作用的数目成正比。因此，我们认为，序列和与外界的相互作用两者的共同作用可能使蛋白质结构对称性在进化中保持不变。这些结果的一个重要启示是在分析蛋白质序列与结构关系中，氨基酸残基间相互作用的模式和分布是联系序列与结构的关键因素，也就是说序列与结构需要通过相互作用来建立真正的联系，这可能是序列与结构关系的本质。

对称蛋白的每个对称单元如果原来就是一个基因，那么它们有可能是古老蛋白质分子，有可能独立地折叠。因此，在通过序列对称性分析得到序列的对称单元后，要证明其是由基因复制进化而来，需要研究这些片段的可折叠性。除实验验证外，分子模拟也是很好的补充，也可以给出有意义的结果。一般来讲，对称蛋白质分子及其对称子结构都比较大，全原子分子动力学方法模拟其折叠过程还十分困难。因此我们利用粗粒化的联合残基模型研究了一个由 120 个氨基酸组成的 6 螺旋蛋白质(Ku86 的 C 端结构域，PDB ID: 1q2z) 的折叠[44]。序列对称性分析表明，这个蛋白质可以看成是由两个三螺旋组成的二重对称性的蛋白。模拟过程显示该蛋白质的折叠分成三个步骤：首先，局部螺旋的形成；其次，链两端的螺旋分别聚集形成比较紧凑结构；再次，这两端聚集的结构再装配成天然的整体结构(图 7)。这意味着该蛋白质的两个半段可以独立折叠。通过单独模拟两个半段的折叠，发现它们确实可以分别折叠，而且折叠过程和整体折叠时相同。这表明该蛋白质有可能是三螺旋束通过基因复制和融合进化而来的。

3　蛋白质形成对称结构的意义

不仅许多蛋白质分子本身形成对称结构，而且

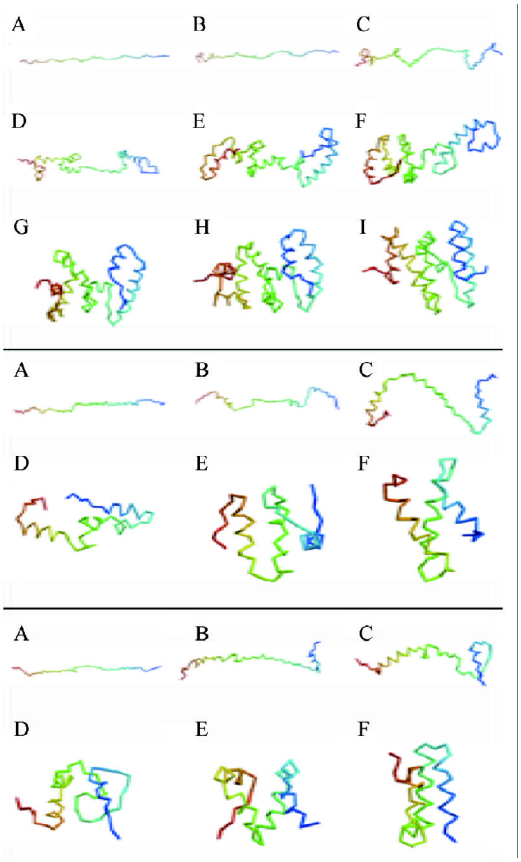

图7　六螺旋蛋白1q2z(上)、它的前半段(中)和后半段(下)的折叠过程

大多数蛋白质分子具有生物活性时是处于具有某种对称性的多聚态[45]。为什么要形成对称的结构，目前有以下基本假说。(1)对称态的能量可能最低，因此蛋白质分子更稳定[46,47]；对称结构单元间有最大的接触面可以使对称蛋白质能量比非对称的更低。蛋白质分子对接研究也显示，复合物最低能量态是对称态[48]。(2)对称结构折叠过程势垒较少，折叠的自由能面比较光滑[47]；*E. Coli* 蛋白质平均的聚集度是 4，很少以单体的方式存在。蛋白质最常见的复合体是相同单体形成的有一个旋转对称轴的二聚体，另外还有三聚体、四聚体和六聚体以及极少的很长的多聚体[49]。(3)重复是构建和设计新功能蛋白质或形成结合新配体位点的最简单的方法。实际上，在蛋白质工程中人们也越来越对重复和对称蛋白(例如ankyrin repeat和leucine-rich repeat)感兴趣[50]。简单来说，对称蛋白质分子的优点是基因编码简单，折叠或装配容易，稳定性高。因此，认识自然界构建对称蛋白质分子结构的策略，能够帮助我们有效地设计具有新功能的蛋白质分子。

对蛋白质对称结构形成机制的研究目前只能说刚刚开始，我们也只做了初步的工作。比如我们还不清楚对称蛋白折叠(自组装)的机制：对称单元是先后折叠还是同时折叠；对称结构的形成是由少数几个关键氨基酸决定还是由所有氨基酸共同决定；对称蛋白质的动力学的细节以及与功能的关系、进化的过程、对称单元的可折叠性等，还有待进一步研究。因此，要认识对称蛋白质的生物意义和物理意义以及进化过程还需要更深入、全面的研究。

致谢：感谢参加该文研究工作的李明锋、何毅、陈长军、周睿、许瑞珍、王晓春、纪晓峰、刘秀华、杨艾红、陈寒林、沈小娟和彭东海等。

[参 考 文 献]

[1] Salem GM, Hutchinson EG, Orengo CA, et al. Correlation of observed fold frequency with the occurrence of local structural motifs. J Mol Biol, 1999, 287: 969-81

[2] Pujadas G, Palau J. TIM barrel fold: structural, functional and evolutionary characteristics in natural and designed molecules. Biol Bratislava, 1999, 54: 231-54

[3] Higgins W, Fairwell T, Miles EW. An active proteolytic derivative of the a subunit of tryptophan synthase: identification of the site of cleavage and characterization of the fragments. Biochemistry, 1979, 18: 4827-35

[4] Eder J, Kirschner K. Stable substructures of eightfold β/α-barrel proteins: fragment complementation of phosphoribosylanthranilate isomerase. Biochemistry, 1992, 31: 3617-25

[5] Bertolaet BL, Knowles JR. Complementation of fragments of triosephosphate isomerase defined by exon boundaries. Biochemistry, 1995, 34: 5736-43

[6] Thoma R, Schwander M, Liebl W, et al. A histidine gene cluster of the hyperthermophile Thermotoga maritima: sequence analysis and evolutionary significance. Extremophiles, 1998, 2: 379-89

[7] Lang D, Thoma R, Henn-Sax M, et al. Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. Science, 2000, 289: 1546-50

[8] Höcker B, Beismann-Driemeyer S, Hettwer S, et al. Dissection of a (β/α)8-barrel enzyme into two folded halves. Nat Struct Biol, 2001, 8: 32-6

[9] Söding J, Lupas AN. More than the sum of their parts: on the evolution of proteins from peptides. Bioessays, 2003, 25: 837-46

[10] McLachlan AD. Repeating sequences and gene duplication in proteins. J Mol Biol, 1972, 64: 417-37

[11] McLachlan AD. Gene duplication and the origin of repetitive protein structures. Cold Spring Harb Symp Quant Biol, 1987, 17: 411-20

[12] McLachlan AD. Evidence for gene duplication in collagen. J Mol Biol, 1976, 107: 159-74

[13] McLachlan AD, Stewart M, Smillie LB. Sequence repeats in α atropomyosin. J Mol Biol, 1975, 98: 281-91

[14] McLachlan AD. Repeated helical pattern in apolipoprotein-A-I. Nature, 1977, 267: 465-66

[15] McLachlan AD. Analysis of gene duplication repeats in the myosin rod. J Mol Biol, 1983, 169: 15-30

[16] McLachlan AD. Three-fold structural pattern in the soybean trypsin inhibitor (Kunitz). J Mol Biol, 1979, 133: 557-63

[17] McLachlan AD. Repeated folding pattern in copper-zinc superoxide dismutase. Nature, 1980, 285: 267-68

[18] McLachlan AD, Bloomer AC, Butler PJ. Structural repeats and evolution of obacco mosaic virus coat protein and RNA. J Mol Biol, 1980, 136: 203-24

[19] Rackovsky S. "Hidden" sequence periodicities and protein architecture. Proc Natl Acad Sci USA, 1998, 95: 8580-4

[20] Anfinsen CB. Principles that govern the folding of protein chains. Science, 1973, 181: 223-30

[21] Gibbs AJ, Mcintyre GA. The diagram, a method for comparing sequences its use with amino scid and nucleotide sequences. Eur J Biochem, 1970, 16: 1-11

[22] Junier T, Pagni M. Dotlet: diagonal plots in a Web browser. Bioinformatics, 2000, 16: 178-79

[23] Deleage G, Combet C, Blanchet C, et al. ANTHEPROT: An integrated protein sequence analysis software with client/server capabili. Comput Biol Med, 2001, 31: 259-67

[24] Heger A, Holm L. Rapid automatic detection and alignment of repeats in protein sequences. Proteins, 2000, 41: 224-37

[25] George RA, Heringa J. The REPRO server: finding protein internal sequence repeats through the web. Trends Biochem Sci, 2000, 25: 515-7

[26] Szklarczyk R, Heringa J. Tracking repeats using significance and transitivity. Bioinformatics, 2004, 20: 311-7

[27] Gruber M, Söding J, Lupas AN. REPPER- repeats and their periodicities in fibrous proteins. Nucleic Acids Res, 2005, 33: W239-43

[28] Soding J, Remmert M, Biegert A. HHrep: *de novo* protein repeat detection and the origin of TIM barrels. Nucleic Acids Res, 2006, 34: W137-42

[29] Eckmann JP, Oliffson Kamphorst S, Rull D. Recurrence plots of dynamical systems. Europhys Lett, 1987, 5: 973-7

[30] Giuliani A, Benigni R, Zbilut JP, et al. Nonlinear signal analysis methods in the elucidation of protein sequence-structure relationships. Chem Rev, 2002, 102: 1471-91

[31] Turutina VP, Laskin AA, Kudryashov NA, et al. Identification of amino acid latent periodicity within 94 protein families. J Comput Biol, 2006, 13: 946-64

[32] Sweet RM, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J Mol Biol, 1983, 171: 479-88

[33] Xu R, Xiao Y. A common sequence-associated physicochemical feature for proteins of β-trefoil family. Comput Biol Chem, 2005, 29: 79-82

[34] Huang YZ, Xiao Y. Detection of gene duplication signals of Ig folds from their amino acid sequences. Proteins, 2007, 68: 267-72

[35] Wang XC, Huang YZ, Xiao Y. Structural-symmetry-related

sequence patterns of the proteins of β-propeller family. J Mol Graph Model, 2008, 26: 829-33
[36] Ji XF, Chen HL, Xiao Y. Hidden symmetries in the primary sequences of β-barrel family. Comput Biol Chem, 2007, 31: 61-3
[37] Huang YZ, Li MF, Xiao Y. Nonlinear analysis of sequence repeats of multi-domain proteins. Chaos Solitons Fractals, 2007, 34: 782-6
[38] Li MF, Huang YZ, Xiao Y. Nonlinear correlations of protein sequences and symmetries of their structures. Chinses Phys Lett, 2005, 22: 1006-9
[39] Li MF, Huang YZ, Xu RZ, et al. Nonlinear analysis of sequence symmetry of β-trefoil family proteins. Chaos Solitons Fractals, 2005, 25: 491-7
[40] Xu RZ, Li MF, Chen HL, et al. A symmetry-related sequence-structure relation of proteins. Chn Sci Bull, 2005, 50: 536-8
[41] Chen HL, Huang YZ, Xiao Y. A simple method of identifying symmetric substructures of proteins. Comput Biol Chem, 2009, 33: 100-7
[42] Chen CJ, Li L, Xiao Y. Identification of key residues in proteins by using their physical characters. Phys Rev E, 2006, 73: 041926-1-7
[43] Li M F, Huang YZ, Xiao Y. Effects of external interactions on protein sequence-structure relations of β-trefoil fold. Proteins, 2008, 72: 1161-70
[44] He Y, Zhou R, Huang Y, et al. Foldable subunits of helix protein. Comput Biol Chem, 2009, 33: 325-8
[45] Goodsell DS, Olson AJ. Structural symmetry and protein function. Annu Rev Biophys Biomol Struct, 2000, 29: 105-53
[46] Blundell TL, Srinivasan N. Symmetry, stability, and dynamics of multidomain and multi- component protein systems. Proc Natl Acad Sci USA, 1996, 93: 14243-8
[47] Wolynes PG. Symmetry and the energy landscapes of biomolecules. Proc Natl Acad Sci USA, 1996, 93: 14249-55
[48] Andre I, Strauss CE, Kaplan DB, et al. Emergence of symmetry in homo- oligomeric biological assemblies. Proc Natl Acad Sci USA, 2008, 105: 16148-52
[49] Levy ED, Boeri Erba E, Robinson CV, et al. Assembly reflects evolution of protein complexes. Nature, 2008, 453: 1262-5
[50] Forrer P, Binz HK, Stumpp MT, et al. Consensus design of repeat proteins. Chem Biol Chem, 2004, 5: 183-9