蛋白质结构型的定义和识别*

李晓琴 罗辽复**
(内蒙古大学理工学院，呼和浩特 010021)

摘要 提出紧结构域的概念，由二级结构序列中一段或几段连续的α螺旋和β折叠构成的空间紧密堆集的最大折叠体称为紧结构域. 利用3种紧结构域 (α域，β域和α/β域) 定义球蛋白的5种结构型：α型蛋白，β型蛋白，α/β型蛋白，多域蛋白和ζ型蛋白. 将1 261个代表性的蛋白质 (1 022家族) 进行分类，并和SCOP库的分类做了比较. 进行了删去序列冗余的分析. 在此基础上提出结构型的预测方案，成功率在82%～85%.

关键词 蛋白质的结构型，二级结构序列，紧结构域，序列冗余，预测
学科分类号 Q61

蛋白质结构型的正确识别是蛋白质三维结构预测问题中必须首先解决的第一步. 利用氨基酸组分预测结构型曾经获得了重要进展[1]，但最近的工作表明：结果对资料库的依赖性较大，随着资料库扩大，4种结构型预测的准确率上限只有60%[2]. 这个预测是以SCOP库中的分类为基础的. 我们曾经指出，以二级结构含量为基础的结构型分类存在问题[3,4]. 曾对500个蛋白质通过计算机图形的分析直接给出结构型；并提出了从规则结构片段数出发进行预测的方案，可以达到90%的成功率[4]. 我们认为，首先要解决的是结构型的定义问题，然后才能设计出好的预测方案，并正确地估计预测结果. 本文在文献[4]的基础上，将蛋白质资料扩大到1 261个，全面分析它们的结构的计算机图形，提出紧结构域的概念，给出5种结构型的定义，然后给出适当的识别规则，这些规则是文献[4]中预测规则的扩充.

1 资料来源

SCOP库 (SCOP database 1.50 release, 29 Feb, 2000) 中每个家族 (不同家族的同源性<30%) 取一个蛋白质，在6类 (allα, allβ, α/β, α+β, multi-domain, small protein) 中共计1 249个序列[5]，结构资料取自PDB光盘 (January 2000, release # 91)，二者相匹配的共1 022个蛋白质，这里未作任何人为的舍弃. 1 022个蛋白质中的22个蛋白质在PDB库中有原子坐标的信息资料，但无二级结构资料，为保留这22个蛋白质，其二级结构资料取自DSSP库. 文献[4]中研究的500个蛋白质涉及261个家族，它们已作为另外的集合 (称为P集) 处理. 除去这些家族，剩下761个家族的761个蛋白质是本文重点研究的资料集 (称为S集). 它们在各结构类中的分布见表1.

Table 1 761 proteins classified into five structural classes

Class	In agreement with SCOP	Different from SCOP	Classified into several classes in SCOP	Total number
α	110	9	2	119
β	105	133	9	238
α/β	95	9	3	104
Multi-domain[1)]	55	139	87	194
ζ	64	42	0	106

[1)] Protein numbers in five structural classes and their comparison with scop.

2 结构型的定义

2.1 紧结构域 (α域，β域和α/β域)

对于一个蛋白质，由α螺旋和β折叠构成的序列称为该蛋白质的二级结构序列. 由α螺旋和β折叠形成的经常出现于多种无关蛋白质折叠中较为固定的组合称为模体. 最常见的模体有 $(\alpha)_n$ (α螺旋重复 n 次)，$(\beta)_n$ (β折叠重复 n 次) 和 $(\beta\alpha\beta)_n = \beta\alpha\beta\alpha\beta\ldots\beta\alpha\beta$ (n 个βαβ重叠串联).

二级结构序列中一段或几段连续的由α螺旋和

* 国家自然科学基金资助项目 (39960023).
** 通讯联系人.
Tel: 0471-4992676, E-mail: lfluo@nmg2.imu.edu.cn
收稿日期：2001-05-14，接受日期：2001-06-28

β折叠构成的空间紧密堆集的最大折叠体称为紧结构域 (compact structural domain, CSD). 紧结构域中的一部分也是紧密堆集的, 但不把这个部分称为紧结构域. 一个结构域 (domain) 如果在空间堆集不够紧密, 也不称为紧结构域. 紧密堆集的一组α螺旋和β折叠一般具有这样的性质: 对于其中任两个长度分别为 a_1 和 a_2 的相邻规则二级结构, 找不到一个方向使它们在此方向的投影范围大于 0.90 (a_1+ a_2). 紧结构域内一般存在包络面, 包络面内没有冗余结构 (参阅 **4.1** 节), 也就是说, 冗余结构安排在包络面的外侧. 紧结构域有 3 种, α域, β域和α/β域. 它们的具体定义如下:

α/β域由二级结构序列中连续的 3 个或 3 个以上的βαβ单元组成, 空间相邻的β折叠平行排列, βαβ之间满足右手螺旋关系[4].

a. βαβ单元中可以允许多插入一个短α (长度小于 6 残基), 即如βααβ; 两个连续的βαβ中可以允许插入β夹, 即如βαββαβ.

b. βαβ连续组合中β折叠平行排布, 其α螺旋和相邻β折叠属同一结构域, 不再参与其他结构域的形成.

c. α/β域要求包含至少 3 个连续的βαβ单元, 这和文献 [6] 基本一致. 该文有的α/β结构中包含βαβ单元数不足 3 个, 但考察这些结构, 皆不满足β折叠平行排布的条件, 所以我们不把它们归入α/β域.

由二级结构序列中连续的 4 个或 4 个以上β折叠参与形成空间紧凑的反平行片状或筒状结构称为β域[4,6].

a. 片状结构比较复杂. 有的片状结构虽然总体上相邻β折叠反平行, 但允许个别β折叠平行排列, 它们一般出现于βαβ中, 这里的α是β折叠状结构的冗余, 不再参与α域的形成. β折叠状结构还可以发生变形: 如对折, 对角卷曲等.

b. 片状结构中存在一种 "β螺旋", 第一第二第三个β折叠形成三角, 第四第五第六个β折叠分别与第一第二第三个β折叠平行. 这种平行结构也归入β域.

c. 由 4 个以下β折叠构成的片状结构不是β类蛋白质的标准结构单元, 同时这种片状结构在其他结构型中也可作为冗余出现, 因此 4 个以下β折叠构成的片状结构不作为β域. 我们规定β域必须至少包含 4 个β折叠, 这和文献 [6] 的结构树一致.

由二级结构序列中连续的 3 个或 3 个以上α螺旋参与形成空间上相对紧凑的结构, 称为α域.

a. 3 个α螺旋构成的紧密结构中必须包含 2 个非平行 (垂直或反平行的) 排列的α螺旋, 因为 3 个连续的平行α螺旋结构上不紧密. 这 2 个非平行α螺旋可看做很多紧密结构的 "根", 在文献 [6] 中称为α-corner.

b. α域要求有 3 个或 3 个以上α螺旋参与. 2 个α片段反平行排列或垂直排列不构成α域, 这是因为考虑到此结构在其他类蛋白质中频繁出现, 并不是α类蛋白质所特有的; 另外, 它的出现不会对这些非α域产生破坏性影响, 可以看作非α域的结构冗余.

c. 4 个或 4 个以下的均较短 (长度小于 8 残基) 的连续α螺旋即使排列紧密, 也不作为独立的α域, 若出现于其他域中, 也是一种结构冗余.

上面给出了 3 种紧结构域的定义, 显然它们正和 3 种常见的模体 $(\alpha)_n$, $(\beta)_n$ 和 $(\beta\alpha\beta)_n$ 相对应. 下面再作几点补充说明. a. 形成紧结构域的规则二级结构片段α和β在序列中可以紧邻, 也可以不紧邻. 一般α域、筒状β结构域 (和一小部分片状β结构域)、和α/β域主要由二级结构序列中紧邻的α和β构成, 而大部分片状β结构域并不要求全部皆序列紧邻的规则二级结构片段构成. b. 若β片状结构中空间相邻的β既有平行分布 (处于βαβ中, 这是α/β结构域的特征) 又有反平行分布 (这是β结构域的特征), 以优势结构来决定此β片状结构属何种结构域. c. 如前所述, 紧结构域中的结构冗余一般分布在包络面外, 但仍属于该结构域. 一个规则二级结构不能同时属于两个紧结构域. d. β结构是靠两股β折叠间的氢键 (或一β折叠内部残基之间的氢键) 相互作用形成的, 而两个α螺旋之间就没有这种氢键作用 (尽管螺旋内部残基之间有氢键作用). 由于规则片段的这种不同的形成机制, 使β折叠较α螺旋更容易结构紧邻而形成结构域. β域的形成要优于α域的形成, 我们在划分结构型分类时应特别注意这一点.

2.2 结构型定义

球蛋白有 5 种结构型, 如下定义: α型蛋白质由一个或几个α域构成; β型蛋白质由一个或几个β域构成; α/β型蛋白质由一个或几个α/β域构成; 多域蛋白质为含有 2 种或 2 种以上不同结构域的蛋白质, 有α+ β, α+ α/β, β+ α/β, α+ β+ α/β等几种情形; ζ型蛋白质为不含任何一种结构域的蛋白质. 上述定义和文献 [4] 的结构型分类 (α, β,

α/β, α+ β 和 ζ) 是一致的, 只是文献 [4] 中 α+ β 型比这里的多域蛋白质范围较窄, 按本文分类, P 集中多域蛋白质为 75 个, 而文献 [4] 中 α+ β 型蛋白质为 56 个.

2.3 结构型分类

据上述定义将资料集 (S 集) 中 761 个球蛋白进行分类, 结果如表 1 所示. 表 1 中也给出了和 SCOP 库分类的比较. 其中, Multi-domain 类和 SCOP 库中的 α+ β (一部分) 和 Multi-domain 两类相应.

2.4 和 SCOP 库分类的比较

从表 1 我们看到 SCOP 库分类和本文分类有一些差异, 主要是:

a. 本文的 β 类蛋白质, SCOP 库分在 α+ β 中的 如 2VIK、1TBD_ , 考察其结构, 只有一个域, β 域, α 未构成域, 故应分在 β 类. SCOP 库分在 α/β 中的, 如 1DMUA、1BXDA、1IGRA, 考察其结构, 空间相邻的 β 以反平行分布为主, 或虽 β 平行, 但并非由于 βαβ 模体, 故应分在 β 类. SCOP 库分在 multi-domain 中的, 如 1MML, 考察其结构, 多个域属于同一种域, 应分在 β 类. SCOP 库分在 ζ 类的, 如 1AFP_ 、1YUA_ , 考察其结构, 实际有 β 域存在, 应分在 β 类. 还有 SCOP 库分在几个类的, 如 1AQT_ 、1GRJ_ , 也应分在 β 类. 以上诸情形中第一种占多数.

需要特别指出的是, SCOP 库中 α+ β 类的若干蛋白质本文都分到了 β 类. 我们认为, 尽管它们可能包含了一定数量的 α, 但从真实的空间结构来看, 都和 SCOP 库 β 类的 β 片状结构蛋白相似, 所不同的只是前者多几个冗余的 α, 后者少几个冗余的 α, 二者应归为一类.

b. 本文的 α 类蛋白质, SCOP 库分在几个类的, 如 1SFE_ 、1QSAA, 实际只有一个域应分在 α 类. SCOP 库分在 α+ β 中的, 如 1CHKA、16VPA, 考察其结构. β 片未构成域. 只有 α 域. 应分在 α 类.

c. 本文的 α/β 类蛋白质, SCOP 库分在几个类的, 如 1B30B、1PSDA、1FSZ_ , 实际只有 α/β 域, 应分在 α/β 类. SCOP 库分在 α+ β 中的, 如 1AK0_ , 考察其结构, 空间相邻的 β 以平行分布为主, 应分在 α/β 类.

本文的多域蛋白对应 SCOP 库的多域和 α+ β, 另 SCOP 库分在几个类的也自然属于此类. 此外还有一些差别: SCOP 分在 α 类的, 如 1OPC_ 、1ADT_ , 考察其结构, 除 α 域外还有 β 域, 应属多域蛋白. SCOP 库分在 β 类的, 如 1LXA_ 、2AHJB, 除 β 域外还有 α 域, 应属多域蛋白. SCOP 库分在 α/β 类的, 如 1FIY_ 、7REQA, 除 α/β 域外还有 α 域; 如 1D8CA, 除 α/β 域外还有 β 域和 α 域, 皆应属多域蛋白质.

d. 本文的 ζ 类蛋白质相当于 SCOP 的小蛋白质, 但有一部分不同, 42 个蛋白质 SCOP 库分在小蛋白质外的其他类中, 考察其结构, 实际无域存在, 应属 ζ 类.

3 结构型的识别

3.1 二级结构序列的冗余分析

定义标准二级结构序列: α 类蛋白质的标准二级结构序列为全 α 结构, 即 ααα ...α; β 类蛋白质的标准二级结构序列为全 β 结构, 即 βββ ...β; α/β 类蛋白质的标准二级结构序列为全 βαβ 结构, 即 βαβαβαβ ...αβ. 标准二级结构序列以外的 α 和 β 单元称为序列冗余. 标准二级结构序列中插入的序列冗余数称为冗余数量. 如 α 类蛋白质的二级结构序列 βαααaββααβααβααβ, 冗余数量为 6. 序列冗余一般都是结构冗余 (即蛋白质紧结构域中的冗余部分), 只有个别例外. 如 α/β 类蛋白质中 βααβ 片段的一个 α 是序列冗余, 但可能并非结构冗余. 一个蛋白质的二级结构序列可以通过删去序列冗余向上述三种标准二级结构序列靠拢. 令向全 α 结构靠拢需删去的序列冗余数量为 n_1, 向全 β 结构靠拢需删去的序列冗余数量为 n_2, 向全 βαβ 结构靠拢需删去的序列冗余数量为 n_3. 考虑到 β 域的形成要优于 α 域的形成, 令 n_1 用等效的 $a\,n_1$ 代替 ($a > 1$, 取 $a = 1.5$). 比较 $a\,n_1$, n_2 和 n_3 的大小, 最小者所对应的结构判定为这个蛋白质的结构型, 如 $a\,n_1$ 最小, 此蛋白质的结构型判定为 α. 这个方法称为去冗余法. 显然, 去冗余法只对 α 类, β 类和 α/β 类这三类蛋白质有效. 对 S 集的 761 个蛋白质使用这个方法, α 类蛋白质判断正确的占 90.8%, β 类蛋白质判断正确的占 90.8%, α/β 类蛋白质判断正确的占 91.4%. 对 S 集和 P 集的全部蛋白质使用这个方法, 三类蛋白质的判断正确率分别为 95.0%, 93.1% 和 93.7%. 误差在 10% 以下, 说明本文的结构型分类有很好的序列基础, 从二级结构序列上就能对这三种结构型作出基本正确的判断. 这 10% 的误差主要由于序列冗余数量 (n_1, n_2 和 n_3) 不能完全反映它对结构的影响, 冗余数量最

小的不一定对结构型的影响也最小 (例如, 除了冗余 α 和 β 的数量外, α 和 β 的长度也对结构有一定影响); 误差也来自结构冗余不全是序列冗余.

3.2 结构型的识别规则

我们建议结构型的预测规则如下: 令 N_{α}, N_{β}, $N_{\beta\alpha\beta}$ 分别为二级结构序列中 α, β, βαβ 的数量, $N_{(\alpha)}$, $N_{(\beta)}$, $N_{(\beta\alpha\beta)}$ 分别为二级结构序列中 α, β, βαβ 连续出现的数量 (若不只一段连续, 计数最长的一段).

当 $N_{(\alpha)}$, $N_{(\beta)}$, $N_{(\beta\alpha\beta)}$ 中二项取大值, 即 $N_{(\alpha)} \geqslant 5$, $N_{(\beta)} \geqslant 5$ 或 $N_{(\alpha)} \geqslant 5$, $N_{(\beta\alpha\beta)} \geqslant 4$ 或 $N_{(\beta)} \geqslant 5$, $N_{(\beta\alpha\beta)} \geqslant 4$, 为多域蛋白质; 当 $N_{(\alpha)}$, $N_{(\beta)}$, $N_{(\beta\alpha\beta)}$ 中一项取大值, 二项取小值, $N_{(\alpha)} \geqslant 5$, $N_{(\beta)} < 5$, $N_{(\beta\alpha\beta)} < 4$, 为 α 类蛋白质 (其中 $N_{(\beta)} \geqslant 4$ 为多域蛋白质, 其余为 α 类蛋白质); $N_{(\beta\alpha\beta)} \geqslant 4$, $N_{(\alpha)} < 5$, $N_{(\beta)} < 5$, 为 α/β 类蛋白质; $N_{(\beta)} \geqslant 5$, $N_{(\alpha)} < 5$, $N_{(\beta\alpha\beta)} < 4$, 为 β 类蛋白质; 当 $N_{(\alpha)}$, $N_{(\beta)}$, $N_{(\beta\alpha\beta)}$ 三项皆取小值, 即 $N_{(\alpha)} < 5$, $N_{(\beta)} < 5$, $N_{(\beta\alpha\beta)} < 4$, 参考 N_{α}, N_{β}, $N_{\beta\alpha\beta}$ 取值, 由文献 [4] 中给出的规则确定.

此规则是在原有预测规则[4]的基础上加入 $N_{(\alpha)}$、$N_{(\beta)}$、$N_{(\beta\alpha\beta)}$ 参数修改而成, 这些参数和紧结构域的形成有密切关系. 对 S 集的 761 个蛋白质使用这个规则进行预测, α 类蛋白质的预测正确率为 91.6%, β 类蛋白质预测正确率为 83.6%, α/β 类蛋白质预测正确率为 82.7%, 多域蛋白质预测正确率为 77.3%, ζ 类蛋白质预测正确率为 76.4%, 总的预测正确率为 82.1%. 对 S 集和 P 集的全部蛋白质使用这个规则进行预测, 5 种结构型的预测正确率分别为 96.2%、88.9%、88.0%、73.6% 和 75.6%, 总的预测正确率为 85.6%.

参 考 文 献

1 Chou K C, Zhang C T. Prediction of protein structural classes. Crit Rev Biochem Mol Biol, 1995, **30** (4): 275~ 349

2 Wang Z X, Yuan Z. How good is prediction of protein structural class by the component - coupled method? Proteins, 2000, **38** (2): 165~ 175

3 李晓琴, 罗辽复. 氨基酸组成聚类蛋白质结构型和结构型的预测. 生物物理学报, 1998, **14** (4): 730~ 736
Li X Q, Luo L F. Acta Biophys Sinica, 1998, **14** (4): 730~ 736

4 Luo L F, Li X Q. Recognition and architecture of the framework structure of protein. Proteins, 2000, **39** (1): 9~ 25

5 Conte L L, Ailey B, Hubbard T J P, *et al*. Scop: a structural classification of proteins database. Nucl Acids Res, 2000, **28** (1): 257~ 259

6 Efimov A V. Structural trees for protein superfamilies. Proteins, 1997, **28** (3): 241~ 260

The Definition and Recognition of Protein Structural Class*

LI Xiao-Qin, LUO Liao-Fu**

(Laboratory of Theoretical Biophysics, Inner Mongolia University, Hohhot 010021, China)

Abstract The concept of compact structural domain (CSD) is proposed which means a maximal compactly-packing-fold composed of α helices and β sheets in secondary structure sequence. There are three kinds of CSDs, namely α domain, β domain and α/β domain. Then five classes of protein are defined. The mainly-α protein is constructed from one or several α domains, the mainly-β protein is constructed from one or several β domains, the α/β protein is constructed from one or several α/β domains, the multi-domain protein is constructed from two or more kinds of CSDs, and the ζ-class protein does not contain any CSD. A database of 1 261 globular proteins is classified into five classes. The classification is compared with SCOP's. The analysis of redundancy-deletion has been done. A prediction rule on structural class is given which is the generalization of previous work. The successful rate of the prediction is higher than 82%.

Key words protein structural class, secondary structure sequence, compact structural domain, sequence redundancy, prediction

* This work was supported by a grant from the National Natural Sciences Foundation of China (39960023).

** Corresponding author. Tel: 86-471-4992676, E-mail: lfluo@nmg2.imu.edu.cn

Received: May 14, 2001 Accepted: June 28, 2001