蛋白质结构型的识别方法*

李晓琴 罗辽复**
(内蒙古大学物理系，呼和浩特 010021)

摘要 给出了α型、β型、α/β型、多域型蛋白质二级结构主序列六联体的分布规律. 提出了根据蛋白质二级结构主序列对蛋白质结构型进行识别 (分类) 的方法. 以蛋白质二级结构主序列三联体为参数，利用 Mahalanobis 距离方法对上述 4 种结构型的蛋白质进行识别，分类的总体准确率为 81%；以二级结构主序列中六联体的频数构成蛋白质结构的多样性源，利用多样性增量极小化对上述 4 种结构型进行识别，分类的总体准确率为 83%. 同时也给出了对紧结构域的识别途径.

关键词 蛋白质结构型，二级结构序列，多样性指标
学科分类号 Q61

蛋白质结构型的确定主要有两种方法：一是根据蛋白质的二级结构含量[1]；二是根据真实的蛋白质拓扑结构图[2]，SCOP 库就是这样得到的. 根据我们的分析，有些蛋白质是很难从二级结构含量来确定其结构型的[3,4]，SCOP 库的结构类区分也存在问题[3]. 文献 [5] 中引进了紧结构域的概念，并利用紧结构域定义蛋白质的 5 种结构型，本文是上述工作的继续.

如何在结构型定义[5]的基础上给出一个合适的识别方法？本文从蛋白质二级结构主序列出发，利用多样性指标法和 Mahalanobis 距离方法对蛋白质结构型进行分类.

1 蛋白质二级结构主序列六联体的统计分布

对文献 [5] 中 S+ P 集的 1 130 个蛋白质 (除ζ类外) 进行统计，求出每一蛋白质二级结构主序列所含各种六联体的频次，然后对一定结构型的全部蛋白质求和，并进行归一化 (归一到 100)，得到结果列于表 1. 不难看到，每类蛋白质二级结构主序列所用的六联体都有特殊的倾向性，例如α类蛋白质倾向使用 αααααα，占 70.2%.

Table 1 Hexa-structures and frequencies in four classes of proteins

α	β	α/β	multi domain
HHHHHH(70.2) EEHHHH(3.5)	EEEEEE(30.2) EEEEHE(5.4)	EHEHEH(16.8) HEHEHE(14.7)	HHHHHH(6.4) EHEHEH(5.2)
EHHHHH(3.2) HHHEEH(2.9)	EEEEEH(5.3) EEEHEE(5.1)	HEHEHH(5.0) HEHHEH(4.0)	HEHEHE(4.8) EEEEEE(4.6)
HHEEHH(2.7) HEEHHH(2.7)	EHEEEE(4.9) EEHEEE(4.8)	HHEHEH(4.0) EHEHHE(3.7)	EHHHHH(2.5) HEHEHH(2.4)
HHHHEE(2.3) HHHHHE(2.0)	HEEEEE(4.6) HEEEEH(2.4)	EHEHEE(3.4) EHHEHE(3.4)	HHEHEH(2.3) HHHHHE(2.0)
HEHHEH(0.9) HHEHHH(0.9)	EHEEHE(2.2) EEHEEH(1.9)	EEHEHE(3.3) HEEHEH(2.7)	EHEHHE(1.9) HEHHEH(1.9)
EHHEHH(0.8) HEHHHH(0.7)	HEEHEE(1.9) HHEEEE(1.8)	HEHEEH(2.6) EHEEHE(2.4)	EHEHEE(1.9) HEHHHH(1.9)
HHHHEH(0.5) HHEHHE(0.5)	EEEEHH(1.8) EEEHEH(1.5)	EHEHHH(1.8) HEHEEE(1.6)	EEHEHE(1.9) EHEEEE(1.8)
HEEEHH(0.5) EEEHHH(0.5)	HEHEEE(1.5) EHEHEE(1.4)	EEEHEH(1.6) EHEEEH(1.5)	EEHHHH(1.8) HEEEEE(1.8)
HHHEEE(0.3) HHHEHE(0.3)	EEHEHE(1.4) EEHHEE(1.3)	HEEEHE(1.5) EHHEHH(1.3)	EHEHHH(1.7) HHHHEE(1.7)
EHHHEE(0.3) HEHHEE(0.3)	EHEEEH(1.3) HEEEHE(1.3)	HHEHHE(1.3) HHHEHE(1.2)	EEEEHE(1.6) EEEEEH(1.6)
EHEHHH(0.3) HEHEHH(0.3)	EHHEEE(1.3) EEEHHE(1.1)	HEHHHE(1.1) EHHHEH(1.1)	HHHHEH(1.6) EEEHEH(1.6)
EEHHEH(0.3) HHEEEH(0.3)			HHHEHE(1.6)

HHHHHH means 6-mer αααααα, EEEEEE means 6-mer ββββββ, etc. The number in bracket indicates the normalized frequency (in 100) of the 6-mer. The 6-mers with too small frequencies in each class have been neglected.

* 国家自然科学基金资助项目 (39960023，90103030).
** 通讯联系人. Tel: 0471-4992676，E-mail: lfluo@mail.imu.edu.cn
收稿日期：2002-06-21，接受日期：2002-08-02

2 蛋白质结构型的 Mahalanobis 距离分类法

2.1 Mahalanobis 方法

对于任意一个蛋白质，可以用 7 个独立的二级结构三联体（ααα、ααβ、αβα、αββ、βαα、βαβ、ββα）在该蛋白质中出现频率构成的向量 μ 来表示该蛋白质. 记第 i 种结构型 π_i（i = α, β, α/β, multi-domain）的均值向量为 $\mu^{(i)}$，协方差阵为 $\Sigma_{(i)}$.

$$\sum_{(i)} = \begin{bmatrix} S^{(i)}_{11} & S^{(i)}_{12} & \cdots & S^{(i)}_{1n} \\ S^{(i)}_{21} & S^{(i)}_{21} & \cdots & S^{(i)}_{21} \\ \hline S^{(i)}_{n1} & S^{(i)}_{n2} & \cdots & S^{(i)}_{nn} \end{bmatrix} \qquad \mu^{(i)} = \begin{bmatrix} \overline{x^{(i)}_1} \\ \overline{x^{(i)}_2} \\ \cdots \\ \overline{x^{(i)}_n} \end{bmatrix} \qquad n = 7, \quad \overline{x^{(i)}_k} = \frac{1}{N^{(i)}} \sum_{j=1}^{N^{(i)}} x^{(i)}_{jk}$$

$$S^{(i)}_{kl} = \frac{1}{N^{(i)}} \sum_{m=1}^{N^{(i)}} (x^{(i)}_{mk} - \overline{x^{(i)}_k})(x^{(i)}_{ml} - \overline{x^{(i)}_l}) \qquad (k, l = 1, \cdots, n = 7)$$

$x^{(i)}_{jk}$ 为 i 类第 j 个蛋白质的第 k 个独立的二级结构三联体在蛋白质 j 中出现的频率，$N^{(i)}$ 为 i 类蛋白质总数.

任意一个蛋白质 x = [x_1, x_2, …x_7]^T（x_1~x_7 分别代表 7 个独立的三联体出现的频率）到结构型 π_i 的马氏距离定义为:

$$d^2(x, \pi_i) = (x - \mu^{(i)})^T \sum_{(I)}{}^{-1} (x - \mu^{(i)})$$

若 $d^2(x, \pi_i) = \min \{ d^2(x, \pi_\lambda) \mid \lambda = \alpha, \beta, \alpha/\beta, \text{multi-domain} \}$，则判别蛋白质 x 属于结构型 π_i.

2.2 分类结果

利用 Mahalanobis 方法，在每一结构型中随机抽取 10 组、每组 150 个蛋白质分别作为训练集，用来计算 Σ_i 和 $\mu^{(i)}$，10 次识别平均成功率及标准偏差见表 2 (1)，即表 2 中的第 3 列. 以全部蛋白质作训练集，算出 Σ_i 和 $\mu^{(i)}$，分类结果见表 2 (2)，即表 2 中的第 4 列.

Table 2 The prediction results by use of Mahalanobis method

Protein class	Protein number	Prediction accuracy (1)	Prediction accuracy (2)
α	262	89.9±1.4	90.85
β	408	89.4±3.5	87.01
α/β	191	76.2±2.5	74.35
Multi-domain	269	68.9±4.8	68.77
Total	1130	81.1±3.1	81.41

3 蛋白质结构型的多样性指标分类法

Laxton[6] 于 1978 年提出多样性（diversity）及多样性指标（measure of diversity）的概念，它对生物学的研究具有重要意义，在生物信息分类中得到广泛应用. 将多样性指标用于蛋白质结构型分类取得了比较好的结果[7, 8]. 文献 [7] 和 [8] 主要是利用二级结构参数，如 N_α、N_β、$N_{\beta\alpha\beta}$、$N_{(\beta\alpha\beta)}$ 等构成多样性源，进行结构型分类. 本文将以蛋白质二级结构主序列中六联体出现的频次 N_i（i = 1 …64）为参数构成多样性源，对蛋白质结构型进行分类.

3.1 多样性增量方法

对于给定的蛋白质二级结构主序列，令 N_i（i = 1 …64）为二级结构主序列六联体出现的频次，其多样性源可以表示为 X: [N_1, N_2 …N_{64}]，多样性指标定义为:

$$D(X) = D(N_1, N_2 \cdots N_{64}) = N \lg N - \sum_{i=1}^{64} N_i \lg N_i$$

$$\text{其中 } N = \sum_{i=1}^{64} N_i$$

结构型 λ（λ 为 α, β, α/β, multi-domain）的标准多样性源定义为 X^λ: [N^λ_1, N^λ_2 …N^λ_{64}]，其中 N^λ_i 为第 i 种六联体在结构型 λ 的全部蛋白质二级结构主序列中出现的总频次. 多样性指标为[9]:

$$D(X^\lambda) = N^\lambda \lg N^\lambda - \sum_{i=1}^{64} N_i{}^\lambda \lg N_i{}^\lambda$$

$$\text{其中 } N^\lambda = \sum_{i=1}^{64} N_i{}^\lambda$$

任意给定未知结构型的蛋白质 X 与标准多样性源 X^λ 的多样性增量为[9]:

$$\Delta(X^\lambda, X) = D(X^\lambda + X) - D(X^\lambda) - D(X)$$

D (X^λ + X) 为 X 和 X^λ 合并成的多样性源 X^λ + X 的多样性指标，其中多样性源 X^λ + X 为: [N^λ_1 + N_1, N^λ_2 + N_2, …, N^λ_{64} + N_{64}]. 可以证明[9]:

$$0 \leqslant \Delta(X^\lambda, X) \leqslant D(N^\lambda, N) = (N^\lambda + N)\lg(N^\lambda + N) - N^\lambda \lg N^\lambda - N \lg N$$

定义相对多样性增量为: $I(X^{\lambda}, X) = \Delta(X^{\lambda}, X) / D(N^{\lambda}, N)$.

蛋白质 X 的结构型就由这 4 个多样性增量 $\Delta(X^{\lambda}, X)$ 或相对多样性增量 $I(X^{\lambda}, X)$ 的最小值所决定.

3.2 结果

蛋白质结构型的识别分三步进行: 第一步利用多样性增量 $\Delta(X^{\lambda}, X)$ 的最小值首先将 α 型、β 型蛋白质区分出来; 第二步利用相对多样性增量 $I(X^{\lambda}, X)$ 的最小值, 将第一步中分在 α/β 型和多域型的蛋白质重新区分 (而第一步已分在 α 型和 β 型的蛋白质保持不动); 第三步, 对于二级结构主序列长度小于 6 的蛋白质, 以 8 个二级结构三联体为参数构成多样性指标和多样性源, 利用多样性增量 $\Delta(X^{\lambda}, X)$ 的最小值来识别它们的结构型.

为使结构型 λ 的标准多样性源更具有普适性, 不依赖给定的数据库, 同时使结构型的分类结果更加可信, 我们采用在特定结构型中随机进行 5 组抽样, 每组抽取 120 个蛋白质构成该结构型的 5 组标准多样性源, 5 次分类平均成功率及标准偏差见表 3 (1), 利用全部样本构造的标准多样性源进行自洽性分类, 结果见表 3 (2).

Table 3 Prediction on structural classes by use of diversity measurement

(1) Prediction accuracy/%

	Protein number	α	β	α/β	Multi-Domain	Tot
Step 1	1058	84.1±3.4	84.0±2.5	89.0±2.3	69.8±3.0	81.3±0.7
Step 2	1058	84.1±3.4	84.0±2.5	85.1±3.9	73.0±2.0	81.4±0.7
Step 3	1130	87.7±2.6	83.7±2.4	85.1±3.9	73.0±2.0	82.3±0.6

(2) Prediction accuracy/%

	Protein number	α	β	α/β	Multi Domain	Tot
Step 1	1058	86.7	84.3	90.6	72.9	83.0
Step 2	1058	86.7	84.3	83.2	77.3	82.8
Step 3	1130	89.7	84.1	83.2	77.3	83.6

4 讨 论

由表 1 可见: α 类蛋白质倾向使用 αααααα, 占 70.2%; β 类蛋白质倾向使用 ββββββ, 占 30.2%; α/β 类倾向使用 βαβαβα 或 αβαβαβ, 占 31.5%. 这三种基本模式的频繁出现表明了 α 和 β 二级结构的特殊成团现象. 它们都有各自的结构学基础, 这是蛋白质结构中最值得注意的现象[3,10]. 基于此, 每类蛋白质二级结构主序列所用的六联体也都有特殊的倾向性, 在 64 种可能存在的六联体中只有少数几种是频繁出现的, 这是本文两种分类方法能获得成功的基础. 在资料库扩大过程中, 只要 α 和 β 结构成团规律不变, 这两种分类方法就能保持较高的成功率.

从二级结构主序列出发进行分类, 成功率可达 80% 以上. 需要说明的是: 尽管二级结构本身还有一个预测问题, 还没有很好解决, 但是本文方法的出发点只是二级结构主序列, 它并不要求准确给定每一残基的二级结构. 而确定 α 和 β 的结构段落是相对容易的, 它具有对二级结构分界点等细节不敏感的鲁棒性. 因此, 从二级结构主序列出发的分类途径在实践上也是可行的.

紧结构域在蛋白质结构和功能中起着重要作用. 若干紧结构域的组装是蛋白质工程的重要途径之一[11]. 所以, 由序列出发识别一个蛋白质是单域还是多域, 以及是何种单域, 便是亟待解决的问题. 本文所识别的结构型是在紧结构域基础上定义的, 因此对结构型的识别方法, 也适于对紧结构域进行识别.

将本文两种分类方法和文献 [5] 的结构型识别规则三者合在一起进行比较, 总体分类成功率相近, 都在 80% 以上. 其中以参数 $N_{(\alpha)}$, $N_{(\beta)}$ 等为基础的分类方法成功率稍高, 达 86%. 但从逻辑简单性考虑, Mahalanobis 距离方法和多样性指标

法更具优势. 多样性指标 D (X) 实际上就是熵, 和熵只差一个常因子; 多样性增量 Δ (X^{λ}, X) 实际上就是 X 和 X^{λ} 合并后的熵增量. 用熵增量最小来判断一个蛋白质的归属是很自然的. 但我们发现单用 Δ (X^{λ}, X) 来分类, 对于 multi-domain 类不利, 一些本应归属多域的划成了单域. 这是由于冗余结构的存在, 有些单域和多域本来就难区分. 加用相对多样性增量 I (X^{λ}, X) 后, 就可弥补此缺陷. 因此, 我们采取了二者并用的方案. 由于多样性指标法中包含 Δ (X^{λ}, X) 和 I (X^{λ}, X) 两个尺度, 用它进行分类更具灵活性, 可以适用于像蛋白质这种均一性较低的系统的分类.

参 考 文 献

1 Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem, 1986, **88** (2): 153~ 162

2 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

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

4 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 Biophysica, 1998, **14** (4): 730~ 736

5 李晓琴, 罗辽复. 蛋白质结构型的定义和识别. 生物化学与生物物理进展, 2002, **29** (1): 124~ 127
Li X Q, Luo L F. Prog Biochem Biophys, 2002, **29** (1): 124~ 127

6 Laxton R R. The measure of diversity. J Theor Biol, 1978, **71** (1): 51~ 67

7 Li Q Z, Lu Z Q. The prediction of the structural class of protein: application of the measure of diversity. J Theor Biol, 2001, **213** (4): 493~ 502

8 贾孟文. 蛋白质结构预测的研究和三核苷重复序列的研究: [学位论文]. 呼和浩特: 内蒙古大学, 2001
Jia M W. The study of protein structural class prediction and the study of TRS bending and flexibility: [Master Thesis]. Hohhot: Inner Mongolia University, 2001

9 徐克学. 生物数学. 北京: 科学出版社, 1999. 277
Xu K X. Biomathematics. Beijing: Science Press, 1999. 277

10 Efimov A V. Structural trees for protein superfamilies. Proteins: Structure, Function and Genetics, 1997, **28** (3): 241~ 260

11 康德君, 许根俊. 蛋白质结构与功能中的结构域. 生物化学与生物物理进展, 1997, **24** (6): 482~ 486
Kang D J, Xu G J. Prog Biochem Biophys, 1997, **24** (6): 482~ 486

The Recognition of Protein Structural Class*

LI Xiao-Qin, LUO Liao-Fu**

(Department of Physics, Inner Mongolia University, Hohhot 010021, China)

Abstract The distribution of hexa-structures in secondary structure sequences of different classes of proteins has been found. Based on this, two methods for the recognition of the structural class of a protein are proposed. The first is the method of Mahalanobis distance which is based on the frequencies of tri-structures in secondary structure sequence. The second is the method of diversity measure which is based on the frequencies of hexa-structures that are regarded as the source of diversity. The prediction has been done in a set of 1 130 proteins of four classes, namely α-class, β-class, α/β-class and multi-domain protein. The successful rates for two recognitions are about 81% and 83% respectively. The method introduced here also gives an approach to predict the compact structural domain of proteins.

Key words protein structural class, secondary structure sequence, diversity measure

* This work was supported by grants from The National Natural Sciences Foundation of China (39960023 and 90103030).

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

Received: June 21, 2002 Accepted: August 2, 2002