基于模板匹配与塔式分解的蛋白质结构域分类*

施建宇[a], 张艳宁[b]
(西北工业大学 a. 生命学院; b. 计算机学院, 西安 710072)

摘 要: 首先构造结构域的距离矩阵灰度图像;其次建立典型二级结构的距离函数,并分析所呈现的灰度模式;然后基于模板匹配和塔式分解,提出了结构域特征;最后在结构类和折叠子两个层次实施结构域分类。本方法在第一种验证策略的分类精度分别为 90.7% 和 74.6%,使用第二种验证策略的为 93.8% 和 78.7%。相比其他方法,具有更高分类精度和更低的特征维数,说明本方法更有效。

关键词: 结构域; 距离矩阵; 模板匹配; 塔式分解; 图像处理; 分类
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2012)06-2081-04
doi:10.3969/j.issn.1001-3695.2012.06.020

Structural classification of protein domain based on template match and pyramid decomposition

SHI Jian-yu[a], ZHANG Yan-ning[b]
(*a. School of Life Sciences, b. School of Computer Science & Technology, Northwestern Polytechnical University, Xi'an* 710072, *China*)

Abstract: The classification of structural domain is one of important approaches which contribute to explore the mechanism of folding and the relationship of protein structure and its biological function. First, this paper mapped spatial structure of protein domain into C_α-C_α distance matrix which could be further regarded as gray texture image. Next, it modeled two distance functions for α helix and β strand/sheet by considering their geometrical properties, and used to find their gray patterns in distance matrix image respectively. After that, it applied the techniques of spatial template match and pyramid decomposition to present the composition feature of α helix and β strand and the multi-scale topology feature of β sheet respectively. Furthermore, in terms of the hierarchy of structural classification of proteins(SCOP), performed domain classifications on structural class and fold levels respectively and compared with other methods. Finally, the results of domain classification show that the proposed method achieves the accuracies 90.7% and 74.6% in the first validation strategy, and 93.8% and 78.7% in the second validation strategy respectively. The comparison with other methods validates the presented method can perform domain classification effectively and outperform its competitors with both the higher classified accuracy and the more compacted dimension of feature vector.

Key words: structural domain; distance matrix; template match; pyramid decomposition; image processing; classification

0 引言

在蛋白质的二级结构和三级结构之间存在的过渡结构——结构域,不仅具有特定的空间组织方式,而且还具有很明确的生物功能,是结构—功能的基本单元[1]。通常一个蛋白质包含一个或多个结构域,蛋白质结构特性由结构域决定,结构域可进一步细分为结构类、折叠子、超家族和家族四个层次[2]。前两个按结构特性分类,主要涉及二级结构的组成与拓扑,后两个分别与功能和进化分类,主要涉及序列同源性分析。

蛋白质结构域的结构分类在了解典型折叠模式的简单性和规律性的物理原因以及结构预测等研究方面具有重要作用。常用的结构域分类数据库或者依赖于人类专家的手工分类[2],或者使用计算量巨大的结构比对方法[3]。显然,结构表示方法是结构域分类的关键。当前的结构表示方法可分为三大类[4]:a)基于空间原子分布[5];b)基于拓扑结构[6];c)基于几何形状[7,8]。方法 a)利用同一类蛋白质的原子空间分布存在一定的相似性,作为提取蛋白质结构特征,所得到空间分布特征通常是基于统计原子坐标的计算,但无法辨识二级结构细节及其之间的拓扑关系。方法 b)主要将蛋白质结构用另一种图形来表示,然后利用相关的图工具进行结构特征分析,但在子结构作为节点的自动获取上尚存在问题,并且通常运行效率较低,对于复杂蛋白质结构的自动辨识常常无法得到满意的结果。方法 c)从蛋白质骨架元素的几何位置或者元素之间的距离来提取蛋白质空间结构全局和细节的特征,如近似骨架曲线和距离矩阵[8]等方法。由于距离矩阵不仅具有更低的计算复杂度,而且它蕴涵了除蛋白质结构的特性之外,可以重构三维结构的足够信息且包含了丰富的结构信息,更重要的是可以反向重构出完整结构[9],此外其数值具有空间结构平移和旋转不变性。

本文将距离矩阵映射为灰度图像,结合二级结构的几何特性,利用模板匹配和塔式分解,提出了一种新的蛋白质结构特征提取方法,所提取特征具有较低维数,并能够充分反映蛋白

收稿日期: 2011-11-28; **修回日期**: 2011-12-30 **基金项目**: 国家自然科学基金资助项目(60872145);博士后科学基金特别资助项目(201104682);香江学者计划资助项目;西北工业大学基础研究项目(JC201164);西北工业大学翱翔之星计划资助项目

作者简介: 施建宇(1977-),男,江西鄱阳人,副教授,博士,主要研究方向为图像处理、模式识别、生物信息学等(jianyushi@nwpu.edu.cn);张艳宁(1968-),女,教授,博士,主要研究方向为计算机视觉、图像处理、模式识别.

质二级结构组成信息和二级结构拓扑信息,且获得了更好的结构域结构分类。

1 二级结构灰度模式分析

1.1 距离矩阵

设蛋白质结构域由 N 个残基组成,则其距离矩阵可表示为 $DM=\{\mathrm{dm}_{p,q}=\mathrm{dist}(\mathrm{Coor}_{\alpha,p},\mathrm{Coor}_{\alpha,q})\}$。其中:$1\leqslant p,q\leqslant N$;$\mathrm{Coor}_{\alpha,n}$ 是第 n 个残基的 C_α 原子的三维坐标向量;dist 表示欧氏距离。可将距离矩阵视做一种纹理图像,即将每一个矩阵元素对应为一个图像像素,每个元素值被映射为对应像素的灰度值[10]。

蛋白质结构域由多个二级结构以某种拓扑方式连接而成。常见的典型二级结构有 α 螺旋和 β 叠片,其中 β 叠片通常成对出现,根据其连接方式可分为反平行和平行连接。根据二级结构的组成,如 α 螺旋、β 叠片组成以及 β 叠片对的几何走向来分类,常见的蛋白质结构类可分为全 α、全 β、α/β 和 $\alpha+\beta$。折叠子的种类较多,同种折叠子具有相同的折叠模式,即其中的二级结构单元通常具有相同的排列和拓扑结构[2]。

本文首先根据距离矩阵图像建立典型二级结构的灰度模式,然后在此基础上提取结构域的结构特征。

1.2 α 螺旋的距离模式

由于 α 螺旋是周期性结构,所以可从其中的任意一个氨基酸残基的 C_α 原子出发来计算它与螺旋结构中其他所有 C_α 原子之间的距离。假设作为出发点 C_α 原子的编号为 1,则编号为 n 的 C_α 原子到编号为 1 的 C_α 原子之间的距离可定义如下:

$$L^{\alpha}_{1\to n}=\sqrt{r^2(1-2\cos((n-1)\cdot\theta))+((n-1)\cdot d)^2} \quad (1)$$

其中:r 表示螺旋半径;d 表示相邻 C_α 原子的螺旋中轴方向距离(螺距);θ 表示相邻 C_α 原子的螺旋转角。这些参数的数值可从教科书上查询获得:$r=2.3$Å,$d=1.5$Å,$\theta=100°$,Å 表示长度单位埃。

根据上述公式,可以在距离矩阵中找到 α 螺旋对应的纹理模式。图 1 给出了一个蛋白质结构域 α 螺旋纹理模式示例,图中颜色越深,表示该像素的值越小。

1.3 单股 β 叠片的距离模式

β 叠片由氢键连接的两个单股组成,从侧面看,β 叠片的单股结构是周期性折线结构。图 2 为 β 叠片的单股中多个氨基酸残基的 C_α 原子间距示意图,其中黑色圆点表示 C_α 原子。

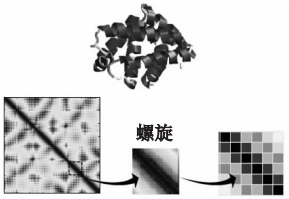

图1 结构域(d1hbga_)的空间结构及其α螺旋在距离矩阵图像中的灰度模式

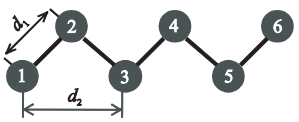

图2 单股β叠片的结构几何示意图

设出发点 C_α 原子的编号为 1,以 d_1 表示相邻两个 C_α 原子之间的距离,d_2 表示编号相隔为 2 的两个 C_α 原子之间的距离,则第一个 C_α 原子到其他 C_α 原子的距离可定义为

$$L^{\beta}_{1\to n}=\begin{cases}\dfrac{d_2\times(n-1)}{2} & n\text{ 为奇数}\\ \sqrt{\left(\dfrac{d_2}{2}\times(n-1)\right)^2+d_1^2-\left(\dfrac{d_2}{2}\right)^2} & n\text{ 为偶数}\end{cases} \quad (2)$$

其中:$d_1=3.8$Å,$d_2=6$Å,$n\geqslant 1$ 是 C_α 原子编号。

图 3 显示了单股 β 叠片对应的纹理模式,图中颜色越深,表示该像素的值越小。

1.4 反平行与平行 β 叠片的距离模式

反平行 β 叠片由两条走向反向平行的股构成,平行 β 叠片由两条走向平行的股构成,编号顺序以蛋白质氨基酸序列的 N 端为起始点,C 端为终止点,如图 4 所示。设 β 叠片结构中第一个氨基酸残基的 C_α 原子编号等于 p,最后一个氨基酸残基的 C_α 原子编号等于 q,其中 $q>p+k$,k 是单股 β 叠片中参与氢键连接的氨基酸残基数目。

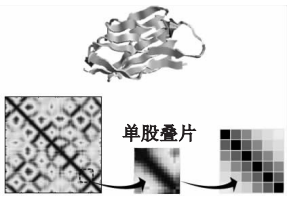

图3 结构域(d1neua_)的空间结构及其单股β叠片在距离矩阵图像中的灰度模式

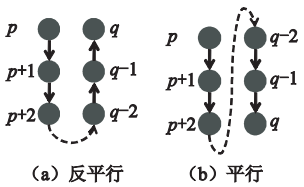

图4 β叠片对的拓扑结构示意图

由图 4 可以得出,组成反平行 β 叠片的两个股之间,所有对应位置残基对的序号之和均等于 $p+q$,所以反平行 β 叠片在图像中表现为在垂直于主对角线的方向上存在灰度相同或相近的直线状区域(边缘);组成平行 β 叠片的两个股之间,所有对应位置残基对的序号之差等于 $q-p-k+1$,所以在距离矩阵图像中表现为在平行于主对角线的方向上存在灰度相同或相近的直线状区域(边缘)。图 5 显示了结构域 d1neua_的反平行与平行 β 叠片的纹理模式,图中颜色越深,表示该像素的值越小。

2 特征提取

2.1 二级结构单元组成特征

因为距离矩阵图像是按主对角线对称的,所以只对图像的上三角区域进行分析即可。由于像素 (p,q) 的灰度值 $f(p,q)$ 对应的是蛋白质中第 p 个 C_α 原子到第 q 个 C_α 原子的距离 $dm_{p,q}$,因此只需在一定的图像区域内进行搜索二级结构的灰度模式。

设蛋白质由 N 个残基组成,则它的图像 ROI(region of interest)可定义为

$$\mathrm{ROI}=\{f(p,q)\mid q\in[p,p+\delta)\}\quad p=1,2,\cdots,N \quad (3)$$

其中:$\delta=5$。图 6 为 ROI 的示意图。

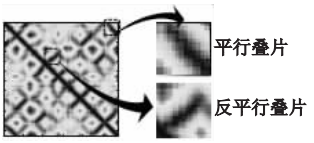

图5 反平行与平行β叠片在距离矩阵图像中的灰度模式(结构域d1neua_)

图6 计算二级结构组成特征的ROI

对 α 螺旋和单股 β 叠片可以分别建立像素模板:

$$T_\lambda=\begin{bmatrix}L^{\lambda}_{1\to1} & L^{\lambda}_{1\to2} & L^{\lambda}_{1\to3} & L^{\lambda}_{1\to4} & L^{\lambda}_{1\to5}\\ & L^{\lambda}_{1\to1} & L^{\lambda}_{1\to2} & L^{\lambda}_{1\to3} & L^{\lambda}_{1\to4}\\ & & L^{\lambda}_{1\to1} & L^{\lambda}_{1\to2} & L^{\lambda}_{1\to3}\\ & & & L^{\lambda}_{1\to1} & L^{\lambda}_{1\to2}\\ & & & & L^{\lambda}_{1\to1}\end{bmatrix},\lambda=\{\alpha,\beta\} \quad (4)$$

对 ROI 区域进行模板匹配,游走方向为距离矩阵的主对角线,这样可以确定每一个像素对应的二级结构单元类型,则二级结构组成特征 P_λ 可由下式确定:

$$P_\lambda = \frac{\#\lambda}{N}, \lambda = \{\alpha, \beta\} \tag{5}$$

其中:#λ 为距离矩阵主对角线上属于 α 螺旋或单股 β 叠片的像素数目,N 为距离矩阵的边长。

2.2　多尺度 β 叠片拓扑组成特征

距离矩阵的对角线附近包含了二级结构的类型和位置信息,而其上三角部分则包含了蛋白质空间结构的 β 叠片结构是由哪两端股形成的信息。这正反映了结构域所包含的 β 叠片排列方式,对蛋白质折叠子结构分类来说至关重要。所以,这里将距离矩阵的上三角区域作为 ROI,并应用多尺度分析思想对其进行塔式分解。首先定义图像的 ROI 区域:

$$\mathrm{ROI} = \{f(p,q) \mid q \in [p+\delta, N)\} \quad p = 1, 2, \cdots, N \tag{6}$$

图 7 为 ROI 及其三级塔式分解示意图。

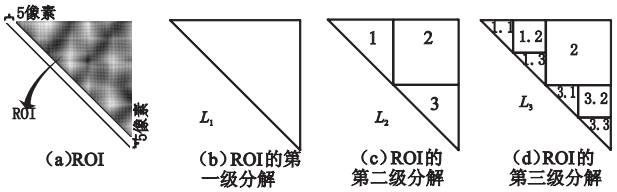

图7　距离矩阵图像塔式级分解

首先对 ROI 进行阈值滤波,定义如下:

$$B_j^i(p,q) = \begin{cases} 1 & \text{if } \Delta < f(p,q) < \Delta + 2 \\ 0 & f(p,q) \in R_j^i \end{cases} \tag{7}$$

其中:$\Delta = 5$, R_j^i 表示第 i 级分解第 j 个区域。则 β 叠片拓扑组成特征可定义为

$$P_\rho^{i,j} = \frac{\sum B_j^i(p,q)}{\#R_j^i} \tag{8}$$

其中:$P_\rho^{i,j}$ 表示第 i 级分解中第 j 个区域 R_j^i 的拓扑特征,$\#R_j^i$ 表示第 i 级分解第 j 个区域的像素总数。

3　实验与分析

实验分别在结构类和折叠子两个分类层次实施,分类器使用支持向量机[11],其核函数选用径向基函数,共采用三种验证策略:a)对训练集实施 10 交叉验证(10-CV),记做 Trn;b)使用训练集来训练分类器模型,使用测试集作独立测试,记做 Ind;b)将训练集与独立测试集合并成一个数据集进行 10 交叉验证,记做 All。整个实验分为三个部分执行。

3.1　实验 1

第一个实验分别使用 α 螺旋和单股 β 叠片的单元组成特征(简称单元组成特征,记做 AB),以及多尺度 β 叠片拓扑组成特征(简称拓扑组成特征)进行,其中尺度分解级别分别为 1、2 和 3,分别记做 L1、L2 和 L3,总共进行了 24 组分类实验,分类结果如表 1 所示。

表 1　使用单元组成特征和拓扑组成特征的分类精度

单特征	维数	结构类/%			折叠子/%		
		Trn	Ind	All	Trn	Ind	All
AB	2	86.3	86.2	86.5	39.9	47.3	44.6
L1	1	82.1	74.8	78.1	25.6	28.8	27.4
L2	3	74.8	74.0	76.9	51.8	55.6	58.2
L3	7	82.1	81.0	86.0	62.3	68.1	74.5

由表 1 可以看出,对于结构类,AB 特征在三种验证策略中所取得的分类识别率均最高,这符合结构类的类别与二级结构单元含量相关这一事实;对于折叠子,随着分解尺度增加,拓扑组成特征的识别率大幅度增加,L3 特征在所有分解级别中识别率最好。此外,除了 L1 之外,其他尺度级拓扑组成特征的识别率均优于 AB 特征。这说明拓扑组成信息对折叠子识别来说比单元组成信息更为重要,这符合折叠子的生物物理意义。

3.2　实验 2

在第二个实验中,将单元组成特征与拓扑组成特征组合之后进行实验,尺度分解级别和验证策略均与上一个实验相同,总共进行了 18 组分类实验,结果如表 2 所示。

表 2　使用组合特征的分类精度

组合特征	维数	结构类/%			折叠子/%		
		Trn	Ind	All	Trn	Ind	All
AB + L1	3	90.1	90.7	90.1	40.9	51.4	50.3
AB + L2	5	91.4	91.2	91.3	60.4	64.2	67.5
AB + L3	9	93.6	90.7	93.8	68.4	74.6	78.7

通过对比表 1 可以看出,对于结构类而言,组合特征要好于对应的最好的单种特征。例如,AB + L1 比 AB 分别增加了 3.8%、4.4% 和 3.6%。同样,对于折叠子而言,组合特征也优于对应的最好的单种特征。例如,AB + L3 比 L3 分别增加了 6.1%、6.5% 和 4.2%。此外,无论是对于结构类还是折叠子层次,随着分解尺度的增加,组合特征的分类精度均逐步增加。

3.3　实验 3

为了验证本文方法的有效性,使用组合特征(AB + L3),分别与多个使用同一个数据库的文献进行了对比。根据文献中使用的测试方法,可分为两大类:一类使用 Ind 验证策略进行分类评估[12~16];另一类则使用 All 验证策略来评估分类结果[17,18]。表 3 和 4 分别列出了两类测试方法的对比结果。

表 3　使用 Ind 验证策略的方法对比

方法	特征维数	结构类/%	折叠子/%
文献[12]	125	N/A	56.5
文献[13]	125	80.52	58.18
文献[15]	125	N/A	61.04
文献[14]	1007	83.6	65.5
文献[16]	1007	87.0	69.6
本文	9	90.7	74.6

表 4　使用 All 验证策略的方法对比

方法	特征维数	结构类/%	折叠子/%
文献[17]	125	84	74
文献[18]	183	N/A	78
本文	9	93.8	78.7

对比实验结果表明,相比多个其他方法,本文方法不仅能够获得最高的分类精度,而且拥有比其他方法低 1 ~ 2 个数量级维数的特征。

4　结束语

本文将蛋白质折叠子三维空间结构映射成为二维距离矩

阵,并将该矩阵视做灰度图像;在此基础上,构建了二级结构 α 螺旋、单股 β 叠片的 C_α-C_α 原子距离函数,分析了反平行和平行 β 叠片对的结构,并分别给出了它们在距离矩阵图像中对应的灰度模式。基于灰度模板匹配,构建了二级结构单元组成的特征;应用塔式分解,构建了多尺度二级结构拓扑组成特征。

本文进一步分析了这两种特征在结构类和折叠层次分类的作用和效果,而且还讨论了采用不同分解级别拓扑组成特征的分类效果,并与多个结构域分类方法进行了对比。结果表明,本文方法是一种有效的蛋白质结构域分类方法。

参考文献:

[1] 谢雪英,李鑫,曹晨. 基于复杂网络的蛋白质结构域组进化分析[J]. 生物物理学报,2010,26(12):1145-1153.

[2] ANDREEVA A, HOWORTH D, JOHN-MARC C, *et al*. Data growth and its impact on the SCOP database: new developments[J]. Nucleic Acids Research,2008,36(Suppl 1):D419-D425.

[3] ALISON L C, IAN S, TONY L, *et al*. The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies[J]. Nucleic Acids Research,2009,37(Suppl 1):D310-D314.

[4] NANNI L, SHI Jian-yu, BRAHNAM S, *et al*. Protein classification using texture descriptors extracted from the protein backbone image [J]. Journal of Theoretical Biology,2010,264(3):1024-1032.

[5] DARAS P, ZARPALAS D, AXENOPOULOS A, *et al*. Three-dimensional shape-structure comparison method for protein classification [J]. IEEE Trans on Computational Biology and Bioinformatics, 2006,3(3):193-207.

[6] DOKHOLYAN N V, LI L, DING F, *et al*. Topological determinants of protein folding[J]. Proceedings National Academy of Sciences of the United States of America, 2002, 99(13): 8637-8641.

[7] KOTLOVYI V, NICHOLS W, TEN E L. Protein structural alignment for detection of maximally conserved regions [J]. Biophysical Chemistry,2003,105(2-3):595-608.

[8] CHOI I G, KWON J, KIM S H. Local feature frequency profile: a method to measure structural similarity in proteins[J]. Proceedings National Academy of Sciences of the United States of America, 2004,101(11):3797-3802.

[9] TIMOTHY H, IRWIN K, GORDON C. The theory and practice of distance geometry[J]. Bulletin of Mathematical Biology,1983,45(5):665-720.

[10] 施建宇, 张艳宁. 使用图像特征构建快速有效的蛋白质折叠识别方法[J]. 生物物理学报,2009,25(2):106-116.

[11] HSU C W, LIN C J. A comparison of methods for multi-class support vector machines[J]. IEEE Trans on Neural Networks, 2002, 13(2):415-425.

[12] DING C, DUBCHAK I. Multi-class protein fold recognition using support vector machines and neural networks[J]. Bioinformatics, 2001,17(4):349-358.

[13] CHINNASAMY A, SUNG W K, MITTAL A. Protein structure and fold prediction using tree-augmented naive Bayesian classifier[J]. Journal of Bioinformatics and Computational Biology, 2005, 3(4):803-820.

[14] HUANG C D, LIN Chin-teng, PAL N R. Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification[J]. IEEE Trans on NanoBioscience,2003,2(4):221-232.

[15] 施建宇,潘泉,张绍武,等. 基于支持向量机融合网络的蛋白质折叠子识别研究[J]. 生物化学与生物物理进展,2006,33(2):155-162.

[16] LIN Ken-lin, LIN C Y, HUANG C D, *et al*. Feature selection and combination criteria for improving accuracy in protein structure prediction[J]. IEEE Trans on NanoBioscience,2007,6(2):186-196.

[17] MARSOLO K, PARTHASARATHY S, DING C. A multi-level approach to SCOP fold recognition[C]//Proc of the 5th IEEE Symposium on Bioinformatics and Bioengineering. Washington DC: IEEE Computer Society,2005:57-64.

[18] MARSOLO K, PARTHASARATHY S. Alternate representation of distance matrices for characterization of protein structure[C]//Proc of the 5th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2005: 298-305.