文章编号:1001 - 2486(2007)02 - 0091 - 03

基于分组重量编码的蛋白质同源寡聚体分类研究*

张振慧[1],王正华[2],王勇献[2]
(1. 国防科技大学 理学院,湖南 长沙 410073;
2. 国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:基于一种新的特征提取方法——分组重量编码(Encoding on the basis of Grouped Weight,简记为EBGW),采用组分耦合算法作为分类器,从蛋白质一级序列出发对四类同源寡聚体蛋白进行分类研究。结果表明,在 Jackknife 检验下,基于分组重量编码的分类方法总体分类精度达到 70.92%,比基于氨基酸组成和加权伪氨基酸成分特征提取方法分别提高 20.28 和 7.53 个百分点,说明分组重量编码对于蛋白质同源寡聚体分类是一种高效的特征提取方法。

关键词:分组重量编码;同源寡聚体;组分耦合算法
中图分类号:Q617 **文献标识码**:B

Classification of Protein Homo-Oligomers with Encoding on the Basis of Grouped Weight for Protein Sequence

ZHANG Zhen-hui[1], WANG Zheng-hua[2], WANG Yong-xian[2]
(1. College of Science, National Univ. of Defense Technology, Changsha 410073, China;
2. College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: The homo-dimer, homo-trimer, homo-tetramer, homo-hexamer of proteins were classified with a new encoding method—encoding on the basis of grouped weight (EBGW) and component-couple algorithm. It was found that the overall classification accuracy by Jackknife test was 70.92%, which was 20.28 and 7.53 percentile higher than that of the amino acid composition and that of the weight pseudo-amino acid composition method with the same classifying algorithm on the same data set. The results indicate that EBGW method has reached a satisfying performance despite its simplicity.

Key words: encoding on the basis of grouped weight; homo-oligomer; component-coupled algorithm

1958 年,Bernal[1] 首次提出蛋白质四级结构的概念,四级结构是蛋白质一级结构、二级结构和三级结构的延伸,是指寡聚蛋白质中亚基的种类、数目、空间排布以及亚基之间的相互作用。利用蛋白质一级结构序列信息预测蛋白质空间结构是当前研究的热点内容[2]。目前对于蛋白质同源寡聚体分类的研究较少。Garian[3] 用决策树和简单 Binning function 特征提取方法对蛋白质的同源二聚体和非同源二聚体进行了分类研究;Chou 等[4] 使用组分耦合算法和伪氨基酸组成方法对蛋白质四级结构的七种分类进行了研究;张邵武等[5-10] 使用支持向量机和组分耦合方法,利用加权自相关函数、伪氨基酸组成成分方法和氨基酸组成分布方法等对四类同源寡聚体蛋白进行分类研究。

本文提出了分组重量编码(EBGW)特征提取方法,结合组分耦合算法对蛋白质同源二聚体、同源三聚体、同源四聚体和同源六聚体进行分类研究。

1 理论与方法

1.1 数据库

我们采用张邵武[8] 等建立的数据集。该数据集从 Swiss-Prot[11] 数据库中挑选 1568 条蛋白质序列,

* 收稿日期:2006 - 11 - 05
作者简介:张振慧(1979—),女,博士生。

其中 914 条同源二聚体(2EM)序列、139 条同源三聚体(3EM)序列、407 条同源四聚体(4EM)序列和 108 条同源六聚体(5EM)序列。

1.2 分组重量编码(EBGW)

考虑氨基酸的疏水性、电价性质等特性可以将 20 种氨基酸划分为四大类[12]:中性非极性 $C1=\{G,A,V,L,I,M,P,F,W\}$、中性极性 $C2=\{Q,N,S,T,Y,C\}$、酸性 $C3=\{D,E\}$ 和碱性氨基酸 $C4=\{H,K,R\}$。将这四种分类两两归并就有三种组合方式:$\{C1,C2\}$ 和 $\{C3,C4\}$、$\{C1,C3\}$ 和 $\{C2,C4\}$、$\{C1,C4\}$ 和 $\{C3,C2\}$。

定义 1(特征序列) 设 $A=a_1a_2\cdots a_n$ 是一条蛋白质序列,我们通过如下定义的映射 Φ_i 将蛋白质序列 A 约化为三条二进制 0 - 1 序列 $\Phi_i(A)=\Phi_i(a_1)\cdots\Phi_i(a_n)$:

$$若\ a_j\in\{C1,C2\},\Phi_1(a_j)=1\ 若\ a_j\in\{C3,C4\},\Phi_1(a_j)=0 \tag{1}$$

$$若\ a_j\in\{C1,C3\},\Phi_2(a_j)=1\ 若\ a_j\in\{C2,C4\},\Phi_2(a_j)=0 \tag{2}$$

$$若\ a_j\in\{C1,C4\},\Phi_3(a_j)=1\ 若\ a_j\in\{C2,C3\},\Phi_3(a_j)=0 \tag{3}$$

记 $H^i=\Phi_i(A)=h_1^ih_2^i\cdots h_n^i(i=1,2,3)$,称 H^1、H^2、H^3 为 A 的特征序列。

定义 2(重量) 特征序列 $H=h_1h_2\cdots h_n$ 的重量定义为该序列中 1 出现的次数。

定义 3(正规重量) 特征序列 $H=h_1h_2\cdots h_n$ 的正规重量定义为该序列中 1 出现的频率,记为 $w(n)$,即 $w(n)=p/n$,其中 p 是序列 H 的重量。

定义 4(EBGW 编码) $H=h_1h_2\cdots h_n$ 是长度为 n 的特征序列,取定正整数 L,我们可以将特征序列 H 划分为 L 条长度递增的子序列,记为 $H(\ kn/L\rfloor)$,每条子序列的长度为 $kn/L\rfloor$,其中 $\cdot\rfloor$ 表示取整运算。记 $w(\ kn/L\rfloor)$ 为子序列的正规重量,则得到一个 L 维的向量 $W=[w(\ n/L\rfloor),w(\ 2n/L\rfloor),\cdots,w(\ Ln/L\rfloor)]$,称由 H 到 W 的转变过程为特征序列 H 的 EBGW 编码。对于一条蛋白质序列 $A=a_1a_2\cdots a_n$,依据其三条特征序列可以将其转换为 $3L$ 维向量,记为 $X=[x_1,x_2,\cdots,x_{3L}]$,称 X 为蛋白质序列 A 的 EBGW 编码。

下面采用组分耦合算法,利用 EBGW 编码进行蛋白质同源寡聚体的分类研究。组分耦合算法的具体算法和设计可以参阅文献[3,6]。

2 实验结果与讨论

2.1 实验结果

采用 Jackknife 检验对分类结果进行评价,精度评估分别采用总体分类精度 Q_3 和每类样本的分类精度 $Q_3(i)$。结果列于表 1 中,其中 $L=24$。

由表 1 可以看出,二聚体的分类最为准确,分类精度达到 81.18%;四聚体的分类次之,分类精度为 69.29%;而六聚体的分类效果最差,分类精度仅为 25.00%。结合每类样本的样本总数,我们发现,分类精度与样本数成正比。

表 1 Jack-knife 检验结果

Tab. 1 Classification results by Jack-knife test

类型	总数	2EM	3EM	4EM	6EM	$Q_3(i)$
2EM	914	742	1	171	0	81.18%
3EM	139	68	61	10	0	43.88%
4EM	407	125	0	282	0	69.29%
6EM	108	63	0	18	27	25.00%
Q_3						70.92%

同时也发现,二聚体蛋白大多数被错误分类为四聚体蛋白,而四聚体蛋白大多数被错误分类为二聚体蛋白,三聚体和六聚体蛋白之间不会被错误分类。这为进一步研究同源寡聚体蛋白的结构和功能提供了有用的信息。

2.2 分组数对分类系统的影响

图1显示了分组数的变化对总体分类精度的影响。可以清晰地看到,随着分组数 L 的增加,总体分类精度持续增长,当 $L=24$ 时达到最高。实验中我们发现,当 $L>25$ 时,分类算法发散。这是由于随着分组组数的增加,零元素(噪声)逐渐增多造成的。因此,当采用组分耦合算法时,我们取 $L=24$。

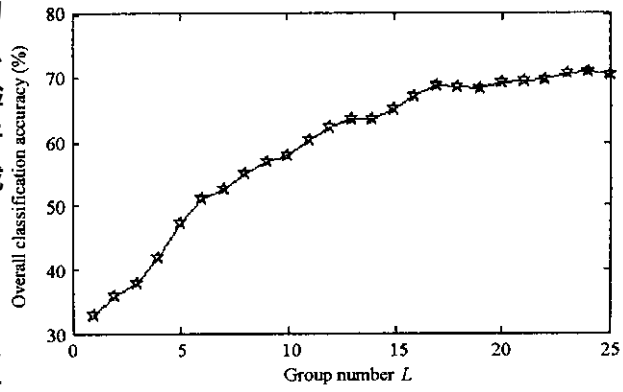

图1 总体分类精度随分组数 L 的变化图

Fig. 1 Distribution of overall classification accuracy with L

2.3 三种特征提取方法对比

将基于组分耦合算法的 EBGW 方法与 AAC[8] 和 PILV[9] 方法进行比较分析,结果见表2。由表2可以看出,EBGW 方法的总体分类精度比 AAC 方法提高 20.28%,比 PILV 方法提高 7.53%。针对二聚体蛋白,EBGW 方法的分类精度比 AAC 和 PILV 方法分别提高 38.51 和 21.44 个百分点;但针对三聚体、四聚体和六聚体蛋白,EBGW 方法的分类精度与 AAC 和 PILV 方法相比都有不同程度的下降。这表示不同特征提取方法的能力是互补的,而 EBGW 方法比较适用于大样本数据集。

表2 三种编码方法的比较

Tab. 2 Comparison results with three methods

方法	2EM	3EM	4EM	6EM	总分类精度
AAC[8]	42.67	62.59	65.36	47.22	50.64
PILV[9]	59.74	59.71	77.15	47.22	63.39
EBGW	81.18	43.88	69.29	25.00	70.92

3 结论

本文依据氨基酸的物化特性,利用物理学中“粗粒化”和“分组”的思想,构建了蛋白质序列的 EBGW 特征提取方法,并结合组分耦合算法进行寡聚蛋白质分类研究。对标准集的 Jackknife 检验分类准确性达到 70.92%。实验结果表明,EBGW 方法的计算简单,包含的信息全面,能够有效地提取蛋白质序列中蕴含的结构信息,进行寡聚蛋白质分类以及蛋白质领域相关问题的预测研究。

参考文献:

[1] Klotz I M, Darnall D W, Langerman N R. The Protein[M]. New York: Academic Presess, 1975.
[2] 王正华,王勇献. 后基因组时代生物信息学的新进展[J]. 国防科技大学学报, 2003, 25: 1-6.
[3] Robert G. Prediction of Quaternary Structure from Primary Structure[J]. Bioinformatics, 2001, 17: 551-556.
[4] Chou K C, Cai Y D. Predicting Protein Quaternary Structure by Pseudo Amino Acid Composition[J]. Proteins, 2003, 53: 282-289.
[5] Zhang S W, Pan Q, Zhang H C, et al. Classification of Protein Quaternary Structure with Support Vector Machine[J]. Bioinformatics, 2003, 19: 2390-2396.
[6] Zhang S W, Pan Q, Zhang H C, et al. Support Vector Machines for Predicting Protein Homo-Oligomers by Incorporating Pseudo-Amino Acid Composition[J]. Internet Electronic Journal of Molecular Design, 2003, 2: 392-402.
[7] 张邵武,潘泉,张洪才,张云龙,王海瑜. 基于支持向量机和贝叶斯方法的蛋白质四级结构分类研究[J]. 生物物理学报, 2003, 19: 171-175.
[8] 张邵武,潘泉,陈润生,张洪才. 基于支持向量机的蛋白质同源寡聚体分类研究[J]. 生物化学与生物物理进展, 2003, 30: 879-883.
[9] 张绍武. 基于支持向量机的蛋白质分类研究[D]. 西安:西北工业大学, 2003.
[10] 施建宇,潘泉,张邵武,程咏梅. 基于氨基酸组成分布的蛋白质同源寡聚体分类研究[J]. 生物物理学报, 2006, 22: 49-56.
[11] Bairoch A, Apeweiler R. The Swiss-Prot Protein Sequence Data Bank and Its Supplement TrEMBL[J]. Nucleic Acids Research, 2000, 25: 31-36.
[12] 林钧材,杨康成. 生物化学(第三版)[M]. 沈阳:辽宁科学技术出版社, 1996.