生物物理学报(2006) 22(4), 275-282

PDF文档

利用分组重量编码预测细胞凋亡蛋白的亚细胞位置

张振慧[1]、王正华*[1]、王勇献[2]
1 国防科技大学理学院博士生队
2 国防科技大学计算机学院并行与分布处理国家重点实验室

从氨基酸的物化特性出发，利用物理学中“粗粒化”和“分组”的思想，提出了一种新的蛋白质序列特征提取方法——分组重量编码方法（Encoding Based on 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细胞凋亡蛋白的亚细胞定位进行研究。采用Zhou和Doctor使用的数据集，Re-substitution和Jackknife检验总体预测精度分别为98.0%和85.7%，比基于氨基酸组成和组分耦合算法的总体预测精度提高了7.2%和13.2%；采用陈颖丽和李前忠使用的数据集，Re-substitution和Jackknife检验总体预测精度分别为94.0%和80.1%，比基于二肽组成和离散增量算法的总体预测精度提高了5.9%和2.0%。实验结果表明蛋白质序列的分组重量编码对于细胞凋亡蛋白的定位研究是一种高效的特征提取方法。

Prediction of the Subcellular Location of Apoptosis Proteins with Encoding Based on Grouped Weight for Protein Sequence

Apoptosis proteins have a central role in the development and homeostasis of an organism. These proteins are very important for understanding the mechanism of programmed cell death. Based on the idea of coarse-grained description and grouping in physics, a new encoding method with grouped weight for protein sequence is presented, and applied to apoptosis protein subcellular location prediction associated with component-coupled algorithm. The average rate of correct recognition is 98.0% in Re-substitution test and 85.7% in Jackknife test for standard set of 98 proteins. For the same training dataset and the same predictive algorithm, the overall predictive accuracy of our method for the Re-substitution and Jackknife test is 7.2% and 13.2% higher than the accuracy based only on the amino-acid composition. The average rate of correct recognition is 94.0% in Re-substitution test and 80.1% in Jackknife test for standard set of 151 proteins, which is 5.9 and 2.0 percentile higher than that of method based on bipeptide composition and the algorithm of measure of diversity. The experiment results show that the encoding method is efficient to extract the structure information implicated in protein sequence and EBGW method has reach a satisfied performance despite its simplicity.

关键词

分组重量编码(Encoding Based on Grouped Weight); 凋亡蛋白(Apoptosis protein); 组分耦合算法(Component-coupled algorithm); 亚细胞定位预测(Subcellular location prediction)