[1]南重汉,邹凌云·基于分组重量编码和特征选择技术预测外膜蛋白[J]·第三军医大学学报,2013,35(13):1366-1370.

Nan Chonghan,Zou Lingyun.Prediction of outer membrane proteins with encoding based on grouped weight and feature selection techniques[J].J Third Mil Med Univ,2013,35(13):1366-1370.

点击复制

基于分组重量编码和特征选择技术预测外膜蛋白(PDF)

Title: Prediction of outer membrane proteins with encoding based on grouped weight and feature selection techniques

作者: 南重汉; 邹凌云
解放军第3医院传染结核病防治所；第三军医大学基础医学部微生物学教研室，重庆市微生物工程实验室

Author(s): Nan Chonghan; Zou Lingyun

Institute of Infectious TB Prevention, No.3 Hospital of PLA, Baoji, Shaanxi Province, 721006; Department of Microbiology, Chongqing Laboratory of Microbial Engineering, College of Basic Medical Sciences, Bioinformatics Center, Third Military Medical University, Chongqing, 400038, China

关键词: 外膜蛋白; 分组重量编码; 特征选择; 支持向量机

Keywords: outer membrane protein; encoding based on grouped weight; feature selection; support vector machine

分类号: Q51;Q755;Q78

文献标志码: A

摘要: 目的　建立新的机器学习模型，从蛋白质数据集或全基因组蛋白质序列中预测外膜蛋白。　方法　采用分组重量编码和氨基酸组成计算蛋白质序列特征，采用F-score方法反向选择特征，采用支持向量机算法建立分类模型，在1 087 条蛋白质序列构成的数据集上进行测试，评价预测模型的敏感性、特异性和预测精度，在多个细菌的全基因组蛋白质中预测外膜蛋白。　结果　新的序列组合编码方法与SVM相结合，区分外膜蛋白和α螺旋跨膜蛋白、球状蛋白、非外膜蛋白的准确度分别达到94.7%、96.4%和94.6%，经特征选择之后，分类准确度分别提高到95.7%、96.9%和95.9%，且在基因组数据集中的预测结果与已知事实相符度高。　结论　该方法预测准确度优于其他基于序列特征的预测方法，可用于在基因组序列中预测和筛选新的外膜蛋白。

Abstract: Objective　To establish a new machine learning model, and to predict outer membrane proteins (OMPs) from various protein datasets and genomic protein sequences using the methods with better accuracy than existing methods. Methods　For protein sequences, combined features were constructed by encoding based on grouped weights and computing amino acid compositions, and the F-score technique was used to obtain backward selection subsets, which

were then inputted to a classifier based on support vector machine algorithms (SVM). The classification model was tested on a dataset including 1 087 protein sequences for evaluation of sensitivity, specificity and overall accuracy, and was also trained and performed to predict OMPs in multiple bacterial genomes.   Results   Through combing the new sequence encoding methods with a SVM-based classifier, our model discriminated OMPs from α-helical transmembrane proteins, globular proteins and non-OMPs with cross-validated accuracy of 94.7%, 96.4% and 94.6%, respectively, and the accuracy severally increased to 95.7%, 96.9% and 95.9% after the feature selection. The whole-genomic prediction also fit the known facts reasonably.   Conclusion   The presented model shows better accuracy than other methods based on sequence features in the literature, and can be used for screening OMPs in genomic sequences.

参考文献/REFERENCES:

南重汉，邹凌云. 基于分组重量编码和特征选择技术预测外膜蛋白[J].第三军医大学学报,2013,35(13):1366-1370.

相似文献/REFERENCES:

[1]葛迪,郭刚,毛旭虎,等.幽门螺杆菌UreB414-OMP18-LTB融合蛋白的构建及免疫原性初步研究[J].第三军医大学学报,2008,30(12):1107.

GE Di,GUO Gang,MAO Xu-hu,et al.Construction and immunogenicity of a fusion vaccine candidate of UreB-OMP18-LTB from H. pylori[J].J Third Mil Med Univ,2008,30(13):1107.