研究论文

DOI:10.3969/j.issn. 1673-1689.2017.03.011

词袋模型在蛋白质亚细胞定位预测中的应用

赵 南[1]， 张 梁[2]， 薛 卫[*1]， 王雄飞[1]， 任守纲[1]

(1. 南京农业大学 信息科学技术学院,江苏 南京 210095;2. 江南大学 粮食发酵工艺与技术国家工程实验室,江苏 无锡 214122)

摘要：运用词袋模型结合传统的蛋白质特征提取算法提取蛋白质序列特征,采用 K-means 算法构建字典,计算获得蛋白质序列的词袋特征,最终将提取的特征值送入 SVM 多类分类器,对数据集中蛋白质的亚细胞位置进行预测,在一定程度上提高了亚细胞定位预测的准确率。

关键词：词袋模型;K-means ;支持向量机;亚细胞定位预测

中图分类号:TP 391.4　**文献标志码**:A　**文章编号**:1673—1689(2017)03—0296—06

Application of Bag of Words Model in the Prediction of Protein Subcellular Location

ZHAO Nan[1]， *ZHANG Liang*[2]， *XUE Wei*[*1]， *WANG Xiongfei*[1]， *REN Shougang*[1]

(1. School of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China; 2. National Engineering Laboratory for Cereal Fermention Technology, Jiangnan University, Wuxi 214122, China)

Abstract: Predecessors have done a lot of work in the feature extraction of protein and subcellular localization prediction. Previous studies showed that prediction accuracy obtained by traditional feature extraction algorithm is low. In order to improve accuracy, bag of words model combined with traditional protein features extraction algorithm is used to extract feature of protein sequence in this study. Firstly, K-means algorithm is used to construct feature dictionary. Then bag of words features of protein sequences are counted by dictionary.Finally extracted feature is inputted into SVM classifier to forecast the protein subcellular location. Results showed that predictionaccuracy of subcellular localization has been improved.

Keywords: bag of words model, K-means, support vector machine, subcellular localization prediction

人类对生命科学的研究因计算机技术的蓬勃发展发生了巨大变化，自从进入后基因组时代,人类获得了大规模的核酸和蛋白质序列数据,借助先进高效的计算机自动化数据处理技术[1]从这些海量数据中挖掘有效信息成为必然趋势。国内外学者在以往的研究中,主要采用数学方法描述提取的蛋白

收稿日期：2015-03-10

基金项目：中央高校基本科研业务费专项资金项目(KYZ201668);江苏省自然科学基金项目(BK2012363,BK2011153);江苏省博士后科研计划项目(1302038B)。

***通信作者**：薛 卫(1979—),男,江苏南通人,理学博士,副教授,硕士研究生导师,主要从事生物信息、模式识别方面的研究。E-mail:xwsky@njau.edu.cn

引用本文：赵南,张梁,薛卫,等. 词袋模型在蛋白质亚细胞定位预测中的应用[J]. 食品与生物技术学报,2017,36(03):296-301.

质序列特征信息,用高维的特征向量表示蛋白质序列,然后设计使用高效的分类器进行预测分析。

目前,用于蛋白质序列特征提取的算法主要包括:氨基酸组成(AAC)、氨基酸的物化特性、二肽及多肽组成、伪氨基酸组成(PseAAC)以及不同特征的融合等[2-6]。如 Lin 等[4]的蛋白质亚细胞定位预测研究采用了四肽信息;杨会芳等[5]在预测蛋白质亚细胞定位中采用了分段伪氨基酸的特征提取方法;Gao 等[6]通过寻找蛋白质不同结构与物化特性的最佳组合来区分外膜蛋白。同时,在预测算法的设计方面国内外研究者开展了大量工作,统计学和机器学习方法在已有的预测算法中得到了充分应用,如陈颖丽等[7]在 6 类细胞凋亡蛋白的亚细胞定位研究中使用了离散增量结合支持向量机的方法;还有基于人工神经网络、马尔可夫模型和贝叶斯网络等的分类预测方法[8-9]。

总结前人研究成果不难发现,单纯采用传统的蛋白质序列特征提取算法如 AAC 等,进行特征提取并送入分类器进行定位预测的准确率偏低。为了改善这一问题,作者引入词袋模型(Bag of Words Model,简称 BOW 模型),BOW 模型源自文档处理领域,也被广泛应用于图像分类方法中。不考虑语法和词序,收集所有文档中出现过的单词,形成一本字典,然后统计获得文档中出现过的单词及其出现的频率[10],将文档表示成高维的向量。作者使用词袋模型完成序列信息的提取,实验证明结合使用 BOW 模型与传统序列特征提取算法 AAC 和 PseAAC 完成蛋白质序列特征的提取,并使用支持向量机分类方法进行定位预测,能有效提高识别精度。

1 材料与方法

1.1 数据集

采用两个凋亡蛋白数据集,第一个数据集由 Zhou 和 Doctor[11]构建,该数据集包含 98 条凋亡蛋白质序列,分为四个亚细胞定位类别,分别是 43 个细胞质蛋白、30 个膜蛋白、13 个线粒体蛋白和 12 个其它类蛋白;第二个数据集是由 Chen 和 Li[12]构建,该数据集包含 317 条蛋白质序列,总共有 6 个亚细胞定位类别,分别是 112 个细胞质蛋白、55 个膜蛋白、34 个线粒体蛋白、17 个分泌蛋白、52 个细胞核蛋白和 47 个内质网蛋白。这两个数据集的蛋白质序列均从 SWISS-PROT 数据库获得。

1.2 蛋白质序列的词袋特征

BOW 模型描述文档的方法是用 D 表示一个存在的文档集合,由 M 个文档组成,提取 M 个文档中出现过的单词,假设不同的单词个数为 N,由这 N 个单词构成字典,则每一个文档都可以被表示成一个 N 维的向量[13]。同理,一个蛋白数据集包含若干条蛋白质序列,连续选取每一条蛋白质序列的若干个片段,称这样的片段为序列单词,分别采用传统的序列特征提取算法 AAC 和 PseAAC 统计序列单词的氨基酸组分信息和位置信息,用向量表示,称这样的向量为序列单词特征;然后采用 K-means 聚类算法对所有的序列单词特征进行聚类分析,聚类分析之后所得到的所有聚类中心的集合,称为字典,字典的大小由聚类中心的个数 k 决定,所有的序列单词特征将映射到字典中的各个聚类中心;逐一统计每一条蛋白质序列属于各个聚类中心的序列单词个数,从而绘制出每一条蛋白质序列的序列单词直方图,计算各个聚类中心上序列单词个数占该条蛋白质序列序列单词总数的比例即可得到蛋白质序列的词袋特征,则每一条蛋白质序列都可以用一个 k 维向量来表示。此方法主要分为 5 个步骤:

1) 分割数据集中所有的蛋白质序列产生若干个序列单词;

2) 提取序列单词的序列单词特征;

3) 对序列单词特征进行聚类分析,获得字典,字典大小为聚类中心个数 k;

4) 经聚类分析后序列单词特征被映射到字典中的各个聚类中心,统计每一条蛋白质序列属于各个聚类中心的序列单词个数,获得蛋白质序列的序列单词直方图;

5) 对每一条蛋白质序列计算各个聚类中心上序列单词个数占该条蛋白质序列序列单词总数的比例,从而获得蛋白质序列的词袋特征,每一条蛋白质序列被表示成一个 k 维的向量。

词袋特征提取过程见图 1。

1.2.1 序列单词特征提取 提取特征前对蛋白质序列进行分割处理,分割蛋白质序列可采用均匀分割和滑动窗口分割。均匀分割法是把每条蛋白质序列均匀分割为多个序列单词,得到的大量序列单词的集合构成构建字典的基础。滑动窗口方法则每间

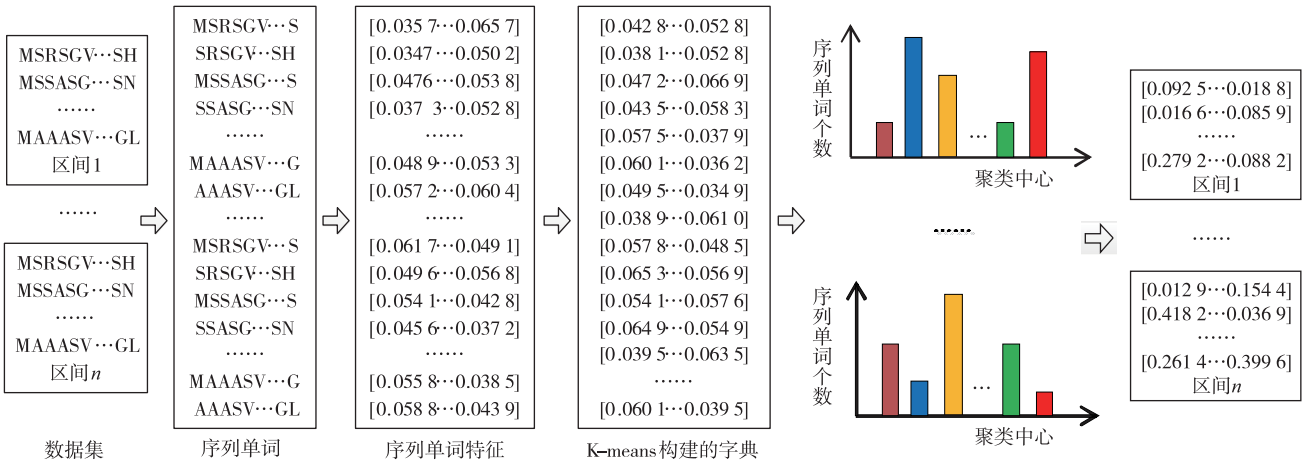

图 1　词袋特征提取过程

Fig. 1　Bag of words feature extraction process

隔一定数量截取窗口内的蛋白质序列片段作为一个序列单词,设定不同的间隔字符个数和窗口大小可以得到不同长度的序列单词。

主要采用滑动窗口分割法,从序列的 N 端到 C 端每次滑动间隔固定为 1，窗口大小决定序列单词的长度,选取方法如下:

$$L=\mathrm{Min}\{L_1,L_2,\cdots,L_n\},\frac{L}{2}\leqslant d\leqslant L,d\in Z \tag{1}$$

其中 $L_1,L_2,\cdots,L_n$ 表示数据集中所有蛋白质序列的长度,L 为数据集中最短蛋白质序列的长度,d 为滑动窗口大小，即序列单词长度在 $\frac{L}{2}$ 与 L 之间选取。

分割后统计序列单词的组分信息和位置信息,运用 BOW 模型结合已有的 AAC 和 PseAAC 算法,采用两种统计方法，分别称为 BOW_AAC 和 BOW_PseAAC。

设序列单词 P 为:

$$p=R_1R_2R_3R_4R_5\cdots R_L \tag{2}$$

其中 $R_1R_2R_3R_4R_5$ 表示序列单词 P 的第一到第五个氨基酸残基，以此类推,R_L 表示序列单词 P 的最后一个氨基酸残基。

1)BOW_AAC 序列单词特征提取:P 的氨基酸组分信息定义如公式(3)[2]所示:

$$p=[f_1f_2\cdots f_{20}]^{\mathrm{T}} \tag{3}$$

$f_1f_2\cdots f_{20}$ 的计算用公式(4)求解:

$$f_u=\frac{1}{N}\sum_{i=1}^{L}R_i,R_i=\begin{Bmatrix}1,\mathrm{If}\ R_i=A(u)\\0,\mathrm{If}\ R_i\neq A(u)\end{Bmatrix} \tag{4}$$

其中,$f_u(u=1,2,3,\cdots,20)$表示 20 种氨基酸在序列单词中出现的频率,L 表示一个序列单词的长度,N 表示一个序列单词包含的所有氨基酸残基的总数目,$A(u)$表示序号 u 所对应的氨基酸残基。经过统计计算，所有的序列单词都可以用一个 20 维的向量表示,从而获得所有蛋白质序列的序列单词特征。

2)BOW_PseAAC 序列单词特征提取:假设序列单词有 L 个氨基酸残基,表示同公式(2),任意一个氨基酸残基在同一个序列单词中与其他氨基酸残基存在不同程度的相关作用,用序列相关因子定义氨基酸残基之间的相关性[14],定义如公式(5)[15]所示:

$$\begin{Bmatrix}\theta_1=\frac{1}{L-1}\sum_{i=1}^{L-1}C_{i,i+1}\\\theta_2=\frac{1}{L-2}\sum_{i=1}^{L-2}C_{i,i+2}\\\theta_3=\frac{1}{L-3}\sum_{i=1}^{L-3}C_{i,i+3}\\\cdots\\\theta_\lambda=\frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}C_{i,i+\lambda}\end{Bmatrix} \tag{5}$$

其中,θ_1 表示第一级相关因子，反映序列单词中相邻两个氨基酸残基之间的相关性;θ_2 表示第二级相关因子,反映序列单词中每间隔一个氨基酸残基的两个氨基酸残基之间的相关性;θ_3 表示第三级相关因子,反映序列单词中每间隔两个残基的两个氨基酸残基之间的相关性;以此类推。$C_{i,j}$ 是根据氨

基酸残基的疏水性、亲水性和侧链分子量构建的相关函数,定义如公式(6)[15]所示:

$$C_{i,j}=\frac{1}{3}\{[H_1(R_j)-H_1(R_i)]^2+[H_2(R_j)-H_2(R_j)]^2+[M(R_j)-M(R_i)]^2\} \quad (6)$$

其中,$H_1(R_j)$表示 R_j 的疏水性值,$H_1(R_i)$表示 R_i 的疏水性值;$H_2(R_j)$表示 R_j 的亲水性值,$H_2(R_i)$表示 R_i 的亲水性值;$M(R_j)$表示 R_j 的侧链原子量,$M(R_i)$表示 R_i 的侧链原子量。然后序列单词特征可表示为:

$$Z=[x_1\cdots x_{20}\cdots x_{20+\lambda}]^T \quad (7)$$

其中

$$x_u=\begin{cases}\dfrac{f_u}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{\lambda}\theta_j},(1\leqslant u\leqslant 20)\\ \dfrac{\omega\theta_{u-20}}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{\lambda}\theta_j},(20+1\leqslant u\leqslant 20+\lambda)\end{cases}(\lambda<L)[15] \quad (8)$$

λ 表示选取的相关因子类型数目,f_i 表示序列单词中第 i 种氨基酸出现的频率,w 表示序列顺序效应的权重因子,θ_j 表示序列单词中第 j 级序列相关因子。

1.2.2 构建字典 得到序列单词特征之后,下一步即是对这些特征值进行处理,用 K-means 聚类算法构建字典,聚类中心的个数即为字典的大小。核心思想是按照类内方差和最小的原则将 n 个序列单词特征值分为指定的 k 类,k 的选取方法为:

$$k=20+x,0\leqslant x\leqslant 500,x\in Z \quad (9)$$

即聚类中心个数从 20 开始逐一递增选取,结合序列单词长度 d 的选取,可以找到一组(d,k)使获得的词袋特征具有最高的识别精度。而类内方差和最小的定义如公式(10)[16]所示:

$$V=\sum_{i=1}^{k}\sum_{x_j\in S_i}(x_j-\mu_i)^2 \quad (10)$$

其中,$S_i(i=1,2,\cdots,k)$表示聚类中心位置是 μ_i 的第 i 个聚类类别,x_j 为属于聚类类别 S_i 的特征值。利用 K-means 聚类算法构建字典的过程描述如下:

输入:DS:n 个序列单词特征值组成的数据集合,k:聚类中心的个数。

输出:k 个聚类中心的集合即字典。

算法:

1)从 DS 中任意选取 k 个序列单词特征值作为初始聚类中心;

2)计算每个序列单词特征值与各聚类中心的距离,按照最近距离原则将 n 个特征值分配到以 k 个初始中心为代表的聚类类别中;

3)根据步骤 2 得到的结果对新产生的 k 个类别进行中心计算,得到新的聚类中心;

4)重复步骤 2~3,直至达到终止条件,如聚类中心不再变化或者已达到最大迭代次数等。

1.3 支持向量机

支持向量机(SVM)拥有坚实的理论基础,并且数学模型简单明了,在解决高维模式识别问题中具有泛化能力强、分类效率高等优点[17]。借助林智仁等开发设计的 LIBSVM 工具箱用一对一法构造 SVM 多类分类器,为任意两类样本设计一个 SVM,当存在一个未知样本需要分类时,它的类别取得票最多的那个类别。基于这样的 SVM 分类实验,在提取出蛋白质序列的词袋特征之后,主要是选取最佳惩罚参数 c 和核函数参数 g 的问题,作者通过交叉验证选择最佳参数,调用工具箱中的 SVMcgForClass 函数将 c 和 g 划分网格进行搜索,最佳参数是达到最高验证分类准确率时最小参数 c 对应的那组 c 和 g,如果存在多组 g 对应最小参数 c,则最佳参数是搜索到的第一组 c 和 g。然后将训练样本(C_i,y_i)送入分类器,向量 C_i 表示第 i 组训练样本的词袋特征值,y_i 表示该条蛋白质序列所对应的亚细胞位置,最后送入测试样本并统计预测结果。

2 结果与讨论

为了检验方法的预测性能,采用 Jackknife 检验,每次仅从数据集中选取一条蛋白质序列构成测试集,训练集由剩余的蛋白质序列构成,测试次数等于数据集的大小,这种检验方法具有最小的任意性,是一种客观有效的交叉验证方法[18]。最后将本文方法 BOW_AAC_SVM 和 BOW_PseAAC_SVM 在 98 和 317 数据集上的预测结果列于表 1-2。为了方便比较,将运用传统蛋白质序列特征提取算法氨基酸组成(AAC)和伪氨基酸组成(PseAAC)进行特征提取并送入 SVM 分类器得到的预测成功率一并列出,如表中 AAC_SVM 和 PseAAC_SVM 两行所示,同时在表 1 的第一行列出了 G.P.ZHOU 和 K.DOCTOR[11]利用氨基酸组成提取特征值以及采用 Jackknife 进行检验的实验结果。

从表 1 可以看出,在 98 数据集上直接采用

AAC、PseAAC 特征提取算法的总体预测精度分别是 80.2%和 83.3%,用 BOW 模型结合 AAC、PseAAC 提取的特征值的总体识别精度达到了 90.6%和 91.7%,分别提高了 10.4%和 8.4%,对于每一个亚细胞类,也都有不同程度的提高,在传统方法预测成功率较低的 Mitochondrial 和 Other 亚细胞类上最高提升了 23%~25%, 尤其在最后一个亚细胞类上将 AAC_CCA 方法的预测成功率由 25%提高到了 83.3%。通过表 2 的比较发现,运用 BOW 模型的总体预测精度也比传统方法高出 6.7%和 6.9%, 在各个亚细胞类上也都有不同程度的提高, 在 Nuclear 亚细胞类上分别提升了 15.7%和 11.8%,在 Secreted 上比传统方法高出 23.6%。

表 1 98 数据集结果比较

Table 1 Comparison of the results of 98 data sets

算法	Jackknife 检验/%				
	Cyto-plasmic	Mem-brane	Mito-chondrial	other	Ac/%
AAC_CCA[11]	97.7	73.3	30.8	25	72.5
AAC_SVM	90.7	82.1	61.5	58.3	80.2
PseAAC_SVM	93	85.7	69.2	58.3	83.3
BOW_AAC_SVM	97.7	89.3	84.6	75.0	90.6
BOW_PseAAC_SVM	97.7	92.9	76.9	83.3	91.7

表 2 317 数据集结果比较

Table 2 Comparison of the results of 317 data sets

算法	Jackknife 检验/%						
	Cyto	Memb	Mito	Secr	Nucl	Endo	Ac/%
AAC_SVM	90.2	81.8	70.6	52.9	70.6	89.4	81.3
PseAAC_SVM	91.1	83.6	73.5	58.8	72.5	85.1	82.3
BOW_AAC_SVM	94.6	85.5	70.6	76.5	86.3	93.6	88.0
BOW_PseAAC_SVM	94.6	87.3	82.4	82.4	84.3	91.5	89.2

3 结 语

作者引入词袋模型应用于蛋白质亚细胞定位预测中,主要技术包括:蛋白质序列分割——滑动窗口法,用来获得大量序列单词的集合,作为构建字典的基础; 序列单词特征提取——BOW_AAC 与 BOW_PseAAC,运用词袋模型结合传统的蛋白质特征提取算法统计蛋白质序列的氨基酸组分信息和位置信息;构建字典——Kmeans 算法,对所有的序列单词特征进行聚类分析处理,再通过统计计算获得蛋白质序列的词袋特征; 亚细胞定位预测——SVM 多类分类器,对数据集中蛋白的亚细胞位置进行预测。预测准确率较传统的蛋白质序列特征提取算法有所提升,最高达到了 91.7%,尤其在传统方法预测准确率较低的亚细胞类上识别精度明显提高,如在 98 数据集 other 这一亚细胞分类上,预测成功率提高了 25%,在 317 数据集 Secreted 这一亚细胞分类上,预测成功率也提高了 20%以上,对准确预测未知蛋白质的亚细胞位置具有重要作用。此次在特征提取方面做了研究工作并取得了一些成果,接下来将在滑动窗口大小和聚类中心个数的选取方法上做一些改进,并尝试在预测算法设计方面做一些工作,重点关注集成学习以及深度学习等。

参考文献:

[1] QIAO Shanping,YAN Baoqiang. The research review of protein subcellular localization prediction[J]. **Application Research of Computers**,2014,31(2):321-327.(in Chinese)

[2] CHOU Kuochen. Some remarks on protein attribute prediction and pseudo amino acid composition[J]. **Journal of Theoretical Biology**,2011,273(1):236-247.

[3] FAN Guoliang,LI Qianzhong. Predictingprotein submitochondrialocations by combining different descriptors into the general form of Chou's pseudo amino acid composition[J]. **Amino Acids**,2012,43(2):545-555.

[4] LIN Hao,CHEN Wei,YUAN Lufeng,et al. Using over-represented tetrapeptides to predict protein submitochondria locations[J]. **Acta Biotheoretica**,2013,61(2):259-268.

[5] YANG Huifang,CHENG Yongmei,ZHANG Shaowu,et al.Based on the pseudo amino acid composition feature extraction

method to predict protein subcellular localization[J]. **Acta Biophysica Sinica**,2008,24(3):232-238.(in Chinese)
[6] GAO Qingbin,YE Xiaofei,JIN Zhichao,et al. Improving discrimination of outer membrane proteins by fusing different forms of pseudo amino acid composition[J]. **Analytical Biochemistry**,2009,398(1):52-59.
[7] CHEN Yingli,LI Qianzhong,YANG Keli,et al. Based on the discrete incremental support vector machine method of apoptosis protein subcellular location prediction[J]. **Acta Biophysica Sinica**,2007,23(3):192-198.(in Chinese)
[8] ZOU Lingyun,WANG Zhengzhi,HUANG Jiaomin. Prediction of subcellular localization of eukaryotic proteins using position-specific profiles and neural network with weighted inputs[J]. **Journal of Genetics and Genomics**,2007,34(12):1080-1087.
[9] ZHANG Shubo,LAI Jianhuang. Machine learning-based prediction of subcellular localization for protein[J]. **Computer Science**,2009,36(4):29-33,49.(in Chinese)
[10] ZHAO Chunhui,WANG Ying,Masahide KANEKO. An optimized method for image classification based on bag of words model [J]. **Journal of Electronics & Information Technology**,2012,34(9):2064-2070.(in Chinese)
[11] ZHOU Guoping,DOCTOR Kutbuddin. Subcellular location prediction of apoptosis proteins[J]. **Proteins**,2002,50(1):44-48.
[12] CHEN Yingli,LI Qianzhong. Prediction of the subcellular location of apoptosis proteins[J]. **Journal of Theoretical Biology**,2006,245(4):775-783.
[13] YANG Quan,PENG Jinye. Chinese sign language recognition research using SIFT-BoW and depth image information [J]. **Computer Science**,2014,41(2):302-307.(in Chinese)
[14] MA Junwei,GAO Xinzhong,ZHANG Jie. Study on the sequence encoding method of protein subcellular location prediction[J]. **Computer Science**,2012,39(11A):283-287,312.(in Chinese)
[15] CHOU Kuochen. Prediction of protein cellular attributes using pseudo-amino acid composition[J]. **Proteins**,2001,43(3):246-255.
[16] LEI Xiaofeng,XIE Kunqing,LIN Fan. An efficient clustering algorithm based on local optimality of K-Means[J]. **Journal of Software**,2008,19(7):1683-1692.(in Chinese)
[17] GU Yaxiang,DING Shifei. Advances of support vector machines[J]. **Computer Science**,2011,38(2):14-17.(in Chinese)
[18] WANG Wei,ZHENG Xiaoqi,DOU Yongchao,et al. Prediction of protein subcellular location using optimal cleavage site[J]. **Bioinformatics**,2011,9(2):171-175,180.(in Chinese)

会 议 消 息

会议名称(中文):首届微生物药物学学术研讨

所属学科:病毒与免疫学 ,药学

开始日期:2017-04-21　　结束日期:2017-04-23

所在城市:重庆市　渝中区　　具体地点:西南大学

主办单位:中国微生物学会分子微生物学及生物工程专业委员会、重庆市微生物学会

协办单位:重庆市免疫学会、重庆市科学技术协会、中国科学院合成生物学重点实验室、微生物代谢国家重点实验室、浙江大学药物生物技术研究所、微生物资源前期开发国家重点实验室、微生物技术国家重点实验室、农业微生物国家重点实验室、中国科学院热带海洋生物资源与生态重点实验室

承办单位:西南大学三峡库区生态环境与生物资源省部共建国家重点实验室培育基地、西南大学生命科学学院/药学院现代生物医药研究所

联系人:廖国建　　联系电话:13594017530　　E-MAIL:gjliao@swu.edu.cn

会议网站:http://csm.im.ac.cn/templates/team/introduction.aspx?nodeid=9&page=ContentPage&contentid=4562