一种基于最优局部信息融合的蛋白质亚细胞定位预测方法*

张树波[1]，赖剑煌[2]，何建国[3]
（1. 中山大学数学与计算科学学院，广东 广州 5120275；
2. 中山大学信息技术与科学学院，广东 广州 5120275；
3. 中山大学生命科学技术学院，广东 广州 5120275）

摘 要： 基于蛋白质的合成及分选机制，提出了一种新的蛋白质亚细胞定位预测方法。先采用遍历搜索技术，找出各种亚细胞蛋白质序列分选信号和成熟蛋白质之间的最佳分割位点，把蛋白质序列分为两条子序列，计算这两条子序列中的氨基酸组份并将它们融合起来作为整条蛋白质序列的特征，然后构造用于识别每类蛋白质的最佳子分类器，再根据最大化原则组建集成分类器。在NNPSL数据集上，采用5重交叉验证方法对本文方法进行测试，原核和真核两个蛋白质序列子集分别取得94.1%和87.5%的总体预测精度。同时，此方法在一些蛋白质序列中找到的分割位点与真实生物现象相吻合，能为预测蛋白质序列的剪切位点提供参考信息。

关键词： 亚细胞定位 *N* 端分选信号 成熟蛋白质 支持向量机 分割位点
中图分类号： TP391.4 **文献标识码：** A **文章编号：** 0529-6579（2008）06-0016-06

蛋白质的亚细胞定位与其功能密切相关。预测蛋白质在亚细胞中的定位，有助于推断基因和蛋白质的功能，了解细胞内蛋白质之间相互作用和调控机制，同时也能为新药物发明提供参考信息。

随着人类在蛋白质组学领域的研究不断深入，单纯以传统的实验方法研究蛋白质定位，由于成本高、实验时间长，预测精度不理想，已经无法满足生命科学研究的需要，因此，基于生物信息学的蛋白质亚细胞定位预测方法逐渐引起人们的关注。蛋白质亚细胞定位的生物信息学方法大概可以归为6类：（1）基于蛋白质序列 *N* 端分选信号的方法[1-2]。这类方法是基于蛋白质序列的 *N* - 端或 *C* - 端存在特殊的分选信号。（2）基于蛋白质序列中氨基酸组份的方法[3-5]。由于基于序列氨基酸组份的信息是一种全局信息，无法有效刻画序列的顺序和位置信息，Guo 等[6]提出了基于伪氨基酸组份和带间隔的氨基酸组份的预测方法，Matsuda 等[5]更是将蛋白质序列分成 *N* 端、中间和 *C* 端三部分，然后从这三个部分提取氨基酸组份等信息，用于亚细胞定位预测。（3）基于功能域的方法[7-8]。由于蛋白质的亚细胞定位与功能密切相关，而它们的功能通常与一些特定的功能 motif 联系，因此蛋白质序列上的功能 motif 信息可以用来预测其亚细胞定位。（4）基于比对信息的方法[9-10]。比对算法可以用于发现序列间同源性较远的关系，这种同源关系可以用作蛋白质序列的相似性度量，Xie 等[9]将 PSI - BLAST 计算得到的信息用于各自的预测系统中，最近 Tamura 和 Akutsu[10]更是提出了一种基于子序列比对的预测方法。（5）基于 GO 分子功能注释信息的方法[11-12]。Chou 等人将蛋白质序列的功能注释（Gene Ontology，简称 GO）信息用于蛋白质亚细胞定位预测。（6）基于氨基酸的物理化学性质的方法[13]。Chou 等[13]将氨基酸的亲水性、疏水性和分子量三种指标用于刻画蛋白质序列的特征。

在亚细胞定位研究的生物信息学方法中，用来刻画蛋白质序列的各种主要特征，除了上述的第一和第五类之外，用来刻画蛋白质序列的特征与亚细胞定位没有直接的联系，难以对亚细胞定位给出生物学解释。本文基于蛋白质分选的生物现象，提出了一种具有合理生物解释的蛋白质亚细胞定位预测方法。先找出每种亚细胞上蛋白质序列 *N* 端分选信号和成熟蛋白质的最佳分割位点，将一条蛋白质序列分割为两部分，分别计算它们的氨基酸组份，并将这两种局部组份信息融合起来作为预测蛋白质亚细胞定位的特征，然后设计用于识别该类蛋白质的子分类器，再根据最大化原则组建集成分类器。在公共数据集上的实验结果表明，本文方法能够有

* **收稿日期：** 2008 - 04 - 11
基金项目： 国家自然科学基金资助项目（60675016，60633030）
作者简介： 张树波（1971年生），男，博士生；**通讯联系人：** 赖剑煌；E-mail：stsljh@mail.sysu.edu.cn

效改进蛋白质亚细胞定位预测的效果，同时，在真核蛋白质上找出的最佳分割位点与真实的生物现象相符合，这能够为预测蛋白质的剪切位点提供有用参考信息。

1　数据集与蛋白质序列特征提取

1.1　数据集的选取

为了与已有的蛋白质预测方法进行比较，本文选用了 Reinhardt 和 Hbnard 建立的数据集 NNPSL[3]，这个数据集包括原核生物蛋白质和真核生物蛋白质两个子集，其中原核生物蛋白质子集包含细胞质蛋白质、细胞外蛋白质和周质蛋白质三类，共有 997 条蛋白质序列；真核生物蛋白质子集包含细胞质蛋白质、细胞外蛋白质、线粒体蛋白质和细胞核蛋白质四类，共有 2 427 条蛋白质序列。

本文为了确保蛋白质序列 N 端分选信号完整性，在原来数据集的基础上，将不是以甲硫氨酸（Met）残基开头的蛋白质序列剔除掉；同时为了消除碎片序列和保证分割位点搜索所需的最小长度（原因详见最优分割位点搜索部分），长度小于 70 个氨基酸残基的序列也被筛选掉。这样最终使用的数据集包含原核蛋白质序列 973 条，真核蛋白质序列 2 345 条。表 1 详细列出各类蛋白质序列的数量。

表 1　NNPSL 数据集中各种蛋白质序列的数量

Tab. 1　Number of different kind of proteins in NNPSLdatasets

蛋白质	真核蛋白质数量		蛋白质	原核蛋白质数量	
类型	NNPSL	本文	类型	NNPSL	本文
细胞质蛋白	684	663	细胞质蛋白	688	680
细胞外蛋白	325	304	细胞外蛋白	107	103
线粒体蛋白	321	318	周质蛋白	202	190
细胞核蛋白	1097	1060			
合计	2427	2345	合计	997	973

1.2　蛋白质序列分析

蛋白质的分选机制[14]表明，在蛋白质的合成过程中，蛋白质前体中往往包含着某些特定的分选信号，这种分选信号能够指导蛋白质前体分选到特定的细胞器中，当它们到达相应的细胞器之后，分选信号被剪切掉，剩下的序列被进一步加工和修饰成为成熟蛋白质，发挥特定的生物学功能。因此，一个完整的蛋白质序列可以从两个角度来刻画：一是与蛋白质分选相关的信息，这种信息通常存在蛋白质前体的 N 端；另一种是与蛋白质功能相关的信息，这种信息存在于成熟的蛋白质序列中。本文基于这种机制，在蛋白质序列中搜索合适的分割位点，将蛋白质序列分割为 N 端分选信号和成熟蛋白质两个部分，提取这两种蛋白质子序列的特征并将它们融合起来，作为整条蛋白质序列的特征。

1.3　蛋白质序列特征提取

由于不同类型蛋白质的 N 端具有不同的分选信号，这些分选信号中氨基酸组份存在差异[15]，这说明蛋白质 N 端子序列中氨基酸组份可以作为分选信号的特征。同时为了与其它基于氨基酸组份的亚细胞定位预测方法比较，本文用氨基酸组份作为蛋白质子序列的特征。

氨基酸组份是指蛋白质序列中，20 种常见氨基酸残基分别所占的比例，通过将蛋白质序列映射为一个 20 维的向量，其中每个元素代表蛋白质序列中一种氨基酸的含量，则蛋白质序列的特征可以表示为：

$$f = (f_1, f_2, \cdots, f_{20}) \tag{1}$$

其中：$f_i = n_i \Big/ \sum_{j=1}^{20} n_j$，$n_i$ 代表序列中第 i 种氨基酸残基的数量，n_j 代表序列中第 j 种氨基酸残基的数量。

对一条给定的蛋白质序列，从最优分割位点将其分开并分别计算这两条子序列中的氨基酸组份，得到两个 20 维的特征向量，然后将它们融合成一个 40 维的向量（Nf_1，Nf_2，…，Nf_{20}，Mf_1，Mf_2，…，Mf_{20}），作为整条蛋白质序列的特征。

1.4　蛋白质序列最佳分割位点的选择

由于不同分选信号长度不同[15]，所以不同亚细胞上的蛋白质中 N 端信号和成熟蛋白质的分割位点是不同的。为了找出每种蛋白质序列上的合适分割位点，我们综合考虑各种分选信号的大致长度[15]，假定蛋白质序列中 N 端分选信号的最小长度为 10 氨基酸残基，最大长度为 70 个氨基酸残基。在蛋白质序列 N 端 10 – 70 氨基酸残基的范围内遍历搜索每一个位置，分别以这些位置为分割位点将蛋白质序列分为两个部分，然后计算这两个子序列中的氨基酸组份，并将它们融合起来构成整条蛋白质序列的特征，再设计和训练相应的分类器。

利用这种方法，对每种亚细胞上的蛋白质，需要设计 61 个二类分类器，每个分类器对应一个分割位点，通过比较各分类器的性能，分类效果最好的分类器对应的分割位点，就是这种细胞器上蛋白质 N 端信号和成熟蛋白质的最佳分割位点。

2　分类器的设计

2.1　子分类器的设计

由前面的方法，我们可以在每种亚细胞蛋白质上找到一个最优的分割位点，从该位置将蛋白质序列分割开来并构造一个 40 维的融合特征向量，根

据这个特征向量可以明显地将该类蛋白质与其它的蛋白质区分开来。因此对于每类蛋白质，我们可以构造一个两类分类器用于有效判别该类蛋白质与其它蛋白质。

支持向量机（SVM）是基于统计学习和最优化理论，以结构风险最小化为准则的机器学习技术，它的目标是在样本空间找出最优分类超平面，使得各类样本之间的间隔（margin）最大化。支持向量机具有良好的推广能力，在很多机器学习领域得到成功的应用，近年来在蛋白质亚细胞定位预测方面也有很多成功的应用，为了与一些经典的方法进行比较，我们采用SVM做为分类器。实验时选用SVM软件包LIBSVM2.85。

由于径向基函数在很多情况下具有良好的分类效果，本文统一选用径向基函数作为核函数：

$$K(v,u) = \exp(-\gamma\|v-u\|^2) \qquad (2)$$

其中 v 和 u 分别是由最优局部信息融合后得到的两条蛋白质序列对应特征向量。γ 是可调整的核参数，用遍历的方法在 $-50 \sim 50$ 之间对每个数值进行筛选。

2.2 集成分类器的设计

根据前面的方法，对每个数据集中的每一类蛋白质，我们可以找到一个最佳的SVM分类器，这个分类器的输出是 $-1 \sim 1$ 之间的数，表示一个蛋白质序列属于该类的可能性大小。这样，原核蛋白质数据集包括三种蛋白质，需要训练出三个最优子分类器，真核蛋白质数据集各包括四种蛋白质，需要训练出四个最优子分类器。

值得注意的是，尽管每个子分类器能够有效地将这种蛋白质与其它蛋白质区分开来，但是它只能判别一个未知的蛋白质序列属于某类或不是，即这样的分类器只能解决两类的分类问题，无法解决本文的多类问题。所以，为了识别一个新的测试样本，需要将各子分类器结合起来，组建一个集成分类器。构建集成分类器时，先根据搜索得到的各类蛋白质上的最佳分割位点，将蛋白质序列分割为 N 端和成熟蛋白质两部分，计算用于刻画整条蛋白质序列的40维特征向量，再将这些特征输入相应的子分类器中，得到的输出结果表示蛋白质序列属于对应亚细胞上的蛋白质的可能性大小，通过比较各子分类器的输出值，采用最大化原则，将该蛋白质序列判给输出值最大的子分类器对应的亚细胞。图1是原核生物蛋白质亚细胞定位预测的集成分类器流程图。其中分割位点1是找到的细胞质蛋白质最佳分割位点，分割位点2是找到的细胞外蛋白质最佳分割位点，分割位点3是找到的周质蛋白最佳分割位点。

3 实验结果与分析

为了评估本文方法的性能和与其它的方法进行比较，我们采用5重交叉验证的方法和总体预测精度、单类预测精度、MCC系数三个评价指标[4]对本文方法的实验结果进行评估。

3.1 蛋白质序列分割位点搜索的结果与分析

在搜索各种蛋白质序列的最佳分割位点的过程中，我们发现支持向量机的参数 C 在1 000左右时，各分类器的性能较好，而且在1 000附近变化对实验结果没有明显的影响，因此在本文的实验中 C 值统一取1 000。图2(a)是当 γ 取19时真核细胞中线粒体蛋白质在不同分割点分类器的识别率。由图可见，在N端27个氨基酸残基处，线粒体与其它蛋白质区分的效果最好，因此这个位点就是线粒体蛋白质上N端信号和成熟蛋白质分割的最佳位点；图2（b）是 γ 取12时原核生物细胞周质蛋白质在不同分割点分类器的识别率。从图中可以看出，原核生物细胞中周质蛋白与其它蛋白质区分效果最好的位置是N端33个氨基酸残基处，因此这个位置是它的最佳分割位点。表2详细列出了不同参数条件下，NNPSL数据集中各种亚细胞中蛋白质的最佳分割位点及相应子分类器的识别结果。

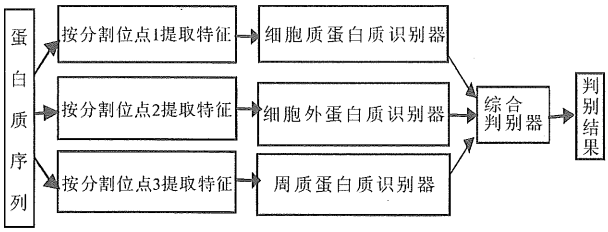

图1 原核生物蛋白质分类器

Fig. 1 The hybrid classifier for prokaryotic proteins

在真核生物数据集上的搜索结果中我们发现两个有趣的现象：（1）线粒体蛋白质的最佳分割位点为27，落在线粒体转移肽（mTP）大致长度范围（25～45个氨基酸残基[15]）范围内，这表明我们在这种蛋白质上找到的最佳分割位点符合实际的生物现象；（2）细胞外蛋白质的最佳分割位点是18，落在信号肽（SP）的长度范围（20～30个氨基酸残基[15]）附近，与真实的生物现象比较接近。

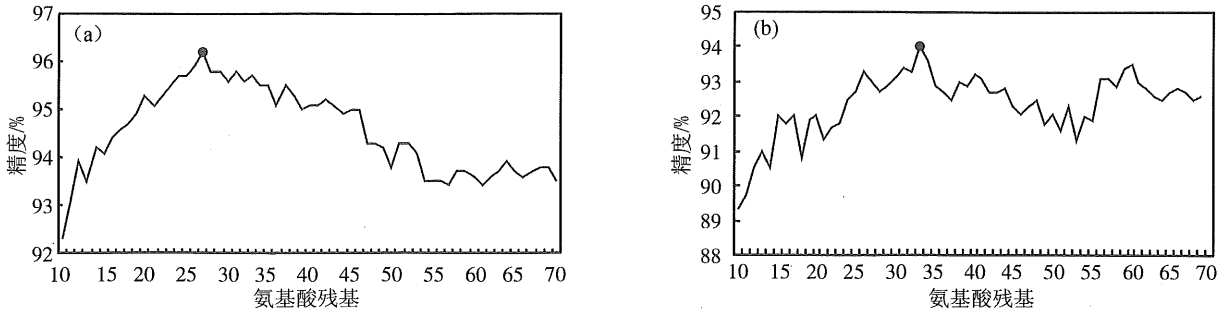

图 2　不同分割位点的识别效果

Fig. 2　The effect of classifier at different splice sits

表 2　NNPSL 数据集中各种亚细胞上蛋白质的最佳分割位点及对应子分类器的性能

Tab. 2　Optimal splice sites and corresponding classifiers for proteins in NNPSL datasets

真核生物				原核生物			
蛋白质类型	γ 值	最佳分割位点	总体精度/%	蛋白质类型	γ 值	最佳分割位点	总体精度/%
细胞质蛋白	42	25	89.7	细胞质蛋白	28	36	96.7
细胞外蛋白	16	18	97.5	细胞外蛋白	18	35	96.2
线粒体蛋白	19	27	96.4	周质蛋白	12	33	94.4
细胞核蛋白	38	41	90.7				

3.2　真核生物数据集上的测试结果与分析

为了客观地评价本文提出方法的效果，我们将本文的实验结果与在 NNPSL 的一些方法的结果进行比较，这些方法有的是采用氨基酸成份作为序列的特征，有的是采用支持向量机作为分类器，有的是将序列分成 N 端、中间和 C 端三部分，与本文的方法存在一定的可比性。

表 3 列出了本文方法在真核生物蛋白质数据集上预测结果，同时也列出了一些典型方法在该数据集上的测试结果。从表中可以看出，与基于氨基酸组份的神经网络[3]、支持向量机[4,6]方法相比，本文预测结果的各项指标均有明显的改善，与同是基于氨基酸组份和 SVM 分类器[4]的方法相比，本文的总体预测精度提高了 8% 左右，细胞外、线粒体和细胞核三种亚细胞上蛋白质的预测精度也显著提高了，尤其是线粒体蛋白质的预测精度提高了将近 30%。与同本文类似的将蛋白质序列分成 N 端、中间和 C 端三部分的方法[5]相比，虽然本文只是采用最简单的氨基酸组份作为特征，提取的特征向量维数只有 40 维，比该方法少 144 维，但是总体准确率已经超过它，单类蛋白质预测精度等各项指标与之大致相当，某些指标超过它。

这些说明将蛋白质序列分为 N 端分选子序列和成熟蛋白质子序列，更加符合真实的生物现象，有效地反映蛋白质亚细胞定位的本质，用这种方法提取的蛋白质序列特征包含更多适合蛋白质亚细胞定位预测的信息，有利于提高预测的精度。

表 3　不同预测方法在真核生物数据集上的性能比较[1)]

Tab. 3　The performance comparison of different methods on eukaryotic dataset

预测方法	总体	细胞质		细胞外		线粒体		细胞核	
	精度/%	精度/%	MCC	精度/%	MCC	精度/%	MCC	精度/%	MCC
Reinhardt et al. [3]	66	55	-	75	-	61	-	72	-
Hua et al. [4]	79.4	76.9	0.64	80.0	0.78	56.7	0.58	87.4	0.75
Guo et al. [6]	86.9	85.8	0.77	85.9	0.89	65.4	0.72	94.2	0.85
Matsuda et al. [5]	87.1	82.5	0.75	89.2	0.90	81.9	0.81	90.9	0.82
本文方法	87.5	79.0	0.76	88.3	0.92	83.8	0.85	93.0	0.80

1）Markov 链模型、氨基酸组份 + SVM 使用的是 Jackknife 测试方法，其余采用交叉验证方法。

3.3　原核生物数据集上的测试结果与分析

表 4 列出不同预测方法在原核生物蛋白质数据集上的预测结果，从表中可见，SVM 方法在性能上具有一定的优势，总体识别率比其它方法要高一些。就同是基于 SVM 方法而言，尽管本文的预测方法比其它两种方法略有改进，但是这种改进却没

有像真核生物那样明显，这可能受两个因素的影响：（1）原核生物本身细胞器分化不明显有关。由于原核细胞器结构简单，细胞内各区室分化不明显，细胞内蛋白质的定位不是十分明确，提供的与蛋白质定位相关的信息有限，找到的最佳分割位点不是很明显；（2）各类蛋白质的样本量严重不均衡对预测结果带来一些不利的影响。如果引进更加稳健的分类算法，或者将蛋白质序列的C端信息也用作蛋白质序列的特征，识别的效果可能会有所改进。

表4 不同预测方法在原核生物数据集上的性能比较[1)]

Tab. 4 The performance Comparison of different methods on prokaryotic dataset

预测方法	总体	细胞质		细胞周质		细胞外	
	精度/%	精度/%	MCC	精度/%	MCC	精度/%	MCC
Reinhardt et al. [3]	81	80	-	85	-	77	-
Hua et al. [4]	91.4	97.5	0.86	78.7	0.78	75.7	0.77
Guo et al. [6]	92.0	99.0	0.89	77.6	0.78	75.7	0.79
Matsuda et al. [5]	91.7	98.5	0.87	77.7	0.86	73.7	0.78
本文方法	94.1	98.9	0.90	83.7	0.86	81	0.87

1）Markov链模型、氨基酸组份+SVM使用的是Jackknife测试方法，其余采用交叉验证方法。

4 结 论

本文根据蛋白质合成和分选的生物学机制，提出一种蛋白质亚细胞定位预测的新方法：在刻画蛋白质序列方面，采用遍历搜索技术，找出每种亚细胞上蛋白质的N端分选信号和成熟蛋白质之间的最佳分割位点，以该位点为界，将蛋白质序列分为前后两条子序列，前面的子序列刻画蛋白质的分选信息，后面的子序列刻画蛋白质的功能信息；在设计分类器时，先根据找到的最佳分割位点设计用于识别每类蛋白质的子分类器，再根据最大化原则组建集成分类器，用于预测未知蛋白质序列的亚细胞定位。与其它方法相比，本文方法建立在蛋白质分选的生物学机理之上，更加符合真实的生物现象，取得更高的预测精度，为蛋白质亚细胞定位研究提供了一种新的思路；同时，在真核生物线粒体和细胞外蛋白质上找到的最佳分割位点，与真实的生物现象相吻合，这能够为预测蛋白质的剪切位点提供有益的参考信息。

参考文献：

[1] HOGLUND A, DONNES P, BLUM T, et al. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition [J]. Bioinformatics, 2006, 22(10): 1158 - 1165.

[2] PETSALAKI E I, BAGOS P G, LITOU Z I, et al. PredSL: A tool for the N-terminal sequence-based prediction of protein subcellular location [J]. Genomics Proteomics Bioinformatics, 2006, 4: 48 - 55.

[3] REINHARDT A, HUBBARD T. Using neural networks for prediction of the subcellular location of proteins [J]. Nucleic Acids Res, 1998, 26(9): 2230 - 2236.

[4] HUA S J, SUN Z R. Support vector machine approach for protein subcellular location prediction [J]. Bioinformatics, 2001, 17: 721 - 728.

[5] MATSUDA S, VERT J P, SAIGO H, et al. A novel representation of protein sequences for prediction of subcellular location using support vector machines [J]. Protein Sci, 2005, 14: 2804 - 2813.

[6] GUO J, LIN Y, SUN Z. A novel method for protein subcellular localization: Combining residue-couple model and SVM[C]. Proceedings of the 3rd Asia-Pacific Bioinformatics Conference, Singapore, 2005, 117 - 129.

[7] CHOU K C, CAI Y D. Using functional domain composition and support vector machines for prediction of protein subcellular location [J]. J Biol Chem, 2002, 277 (48): 45765 - 45769.

[8] SCOTT M S, THOMAS D Y, HALLETT M T. Predicting subcellular localization via protein motif co-occurrence [J]. Genome Res, 2004, 14: 1957 - 1966.

[9] XIE D, LI A, WANG M. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST [J]. Nucleic Acids Res, 2005, 33: 105 - 110.

[10] TAMURA T, AKUTSU T. Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition [J]. BMC Bioinformatics, 2007, 8: 466.

[11] CHOU K C, CAI Y D. Prediction of protein subcellular locations by GO-FunD-PseAA predictor [J]. Biochem Biophys Res Commun, 2004, 320(4): 1236 - 1239.

[12] CHOU K C, SHEN H B. Large-scale plant protein subcellular location prediction [J]. J Cell Biochem, 2007,

100(3)：665－678.

[13] CHOU K C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect [J]. Biochem Biophys Res Commun,2000,278 (2):477－483.

[14] SCOTT D E, MICHAEL N H, THOMAS I S. A mechanism of protein localization: the signal hypothesis and bacteria [J]. J Cell Biol, 1980,86: 701－711.

[15] EMANUELESSON O. Predicting protein subcellular localization from amino acid sequences information [J]. Briefings in Bioinformatics,2002,3:361－376.

A Novel Approach for Prediction of Protein Subcellular Localization Using Optimal Local Information

ZHANG Shu-bo[1], *LAI Jian-huang*[2], *HE Jian-guo*[3]

(1. School of Mathematics and Computational science, Sun Yat-sen University, Guangzhou 510275, China;
2. School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510275, China;
3. School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Prediction of protein subcellular localization can help infer the function of proteins and apply insight into the interaction between proteins. A novel approach based on the sorting mechanism of proteins, is proposed for predicting subcellular localization of proteins. An optimal splice site is found through iterative searching technique to divide the sequence into sorting signal and mature protein subsequence for each kind of proteins. When designing the classifier, a sub-classifier is built to discriminate each kind of protein from the rest, these sub-classifiers are then combined into an ensemble classifier to predict the subcellular localization of unknown proteins. Through five-fold cross-validation tests on NNPSL datasets and TargetP datasets, overall accuracies of 94.1% and 87.5% are obtained for prokaryotic and eukaryotic proteins respectively, as for TargetP datasets, the overall accuracies are 90.2% and 93.9% for plant and non-plant proteins respectively. Meanwhile, the optimal splice sites found in this paper are coincided with the biological facts in most of kinds protein, this can help predict the cleavage sites of proteins.

Key words: subcellular localization; N-terminal sorting signal; mature protein; support vector machine; splice site