PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2015, 42(12): 1136~1143
www.pibb.ac.cn

基于 Blast-GO 的蛋白质亚线粒体定位预测 *

曩 毅 梅含雪 赵 燕 侯宝妍 赵志远 樊国梁 **
(内蒙古大学物理科学与技术学院，呼和浩特 010021)

摘要 本文建立了一个最新的蛋白质亚线粒体定位数据集，包含 4 个亚线粒体定位的 1 293 条序列，结合基因本体(GO)信息和同源信息对线粒体蛋白质进行特征提取，利用支持向量机算法建立分类器，经 Jackknife 检验，对于 4 个亚线粒体位置的总体预测准确率为 93.27%，其中 3 个亚线粒体位置的总体预测准确率为 94.73%.

关键词 亚线粒体定位，基因本体，同源信息，支持向量机
学科分类号 Q61
DOI: 10.16476/j.pibb.2015.0190

线粒体存在于大多数真核细胞的细胞器中，具有两层膜结构：内膜和外膜．线粒体外膜是细胞的边界，线粒体内膜包裹着线粒体基质，两层膜结构之间的部分被称为线粒体膜间隙．线粒体与多种细胞生理活动有关，例如有氧呼吸、克雷布斯循环(柠檬酸循环)、细胞凋亡[1]和细胞的离子稳态[2]．人类的许多疾病与线粒体直接相关，例如阿尔兹海默病[3]、帕金森综合症[4]、Ⅱ型糖尿病[5]．蛋白质是细胞生命活动的主要承担者，但是蛋白质并不是集中地分布在细胞质的一个位置上，因此了解线粒体蛋白质在细胞中的位置对于了解线粒体的生理过程以及与线粒体相关的疾病具有重要意义．线粒体蛋白质的亚线粒体位置可以通过生物学实验来确定，但是实验时间长，费用昂贵．

在过去的 30 年中，出现了许多预测蛋白质亚细胞定位的计算模型和方法，有多种方法提取蛋白质序列的特征信息来预测蛋白的亚细胞定位，如 N 端信号肽[6-7]、氨基酸组分[8-11]、伪氨基酸组分[12-13]、二肽组分[14-15]、功能域信息[16-17]、基因本体(GeneOntology, GO)[16, 18]信息等．一些机器学习算法也被用来预测蛋白质的亚细胞定位，如马尔科夫链算法[19]、判别函数算法[20-21]、支持向量机算法[12, 22-24]、神经网络算法[25-26]、优化 K 近邻算法(OET-KNN)[27]、融合 K 近邻算法(FUZZY-KNN)[14]和分类融合算法[23-24, 26]等．

目前，已提出的亚线粒体蛋白质定位预测算法主要有：SUBMITO [28]、GP-Loc [29]、Predict_subMITO [30]、Shi [31]、Fan [32]、SubMito-PSPCP[33]和 TetraMito[34]．第一代亚线粒体蛋白质定位预测算法 SUBMITO 对于内膜的预测准确度为 85.5%，基质的预测准确度为 94.5%，外膜的预测准确度为 51.2%，平均正确率为 85.20%；GP-Loc 算法对于内膜的预测准确度为 83.21%，基质的预测准确度为 97.24%，外膜的预测准确度为 78.05%，平均正确率为 89.00%；Predict_subMITO 算法对于内膜的预测准确度为 91.80%，基质的预测准确度为 96.40%，外膜的预测准确度为 66.10%，平均正确率为 89.70%；Shi 的算法对于内膜的预测准确度为 91.6%，基质的预测准确度为 97.93%，外膜的预测准确度为 82.93%，平均正确率为 93.38%；Fan 的算法对于内膜的预测准确度为 96.1%，基质的预测准确度为 93.9%，外膜的预测准确度为 86.9%，平均正确率为 93.57%；SubMito-PSPCP 算法对于内膜的预测准确度为 98.6%，基质的预测准确度为 93.9%，外膜的预测

* 国家自然科学基金(61461038)，内蒙古自治区自然科学基金(2013MS0504)，内蒙古自治区高等学校科学研究项目(NJZY13014)，内蒙古大学高层次人才引进科研项目(135147)和内蒙古大学大学生创新创业训练计划项目(201412155)资助.
** 通讯联系人. Tel: 0471-4992958
E-mail: fanguoliang@imu.edu.cn, eeguoliangfan@sina.com
收稿日期：2015-06-25，接受日期：2015-09-21

准确度为 70.7%，平均正确率为 93.1%；TetraMito 算法对于内膜的预测准确度为 100%，基质的预测准确度为 96.6%，外膜的预测准确度为 65.9%，平均正确率为 94.0%．在目前已提出的亚线粒体蛋白质定位预测算法中，使用的最新数据集是 Fan 在 2010 年建立的包含 1 105 条亚线粒体蛋白的数据集，且由于数据量缺乏，以往的亚线粒体蛋白质定位预测算法中，只将蛋白质定位到 3 个亚线粒体位置上，即线粒体内膜、线粒体基质和线粒体外膜．

本文通过检索 SWISS-PROT 数据库，首次建立一个最新的亚线粒体蛋白质定位数据集，包含有 4 个亚线粒体位置：线粒体内膜、线粒体基质、线粒体外膜及线粒体膜间隙．通过结合蛋白质的同源信息和 GO 信息对蛋白质序列进行特征提取，使用支持向量机建立模型，预测了线粒体蛋白质的 4 个亚线粒体定位，利用 Jackknife 交叉验证，总体预测准确率达到 93.27%．

1 材料与方法

1.1 数据集

通过检索 SWISS-PROT (release 2014_11 - November 26, 2014)[35]数据库建立了一个本地亚线粒体蛋白质定位数据集．由于 SWISS-PROT 数据库是 UNIPROT 的一部分，使用 UNIPROT 的高级检索功能，选择“Subcellular location”、“Subcellular location [CC]”、“Subcellular location term”，“Evidence”选择“Any assertion method”，“Type”选择“Any”，在“Term”中分别使用关键词“mitochondrial inner membrane”，“mitochondrial matrix”，“mitochondrial outer membrane”和“mitochondrial intermembrane”搜索得到 4 个亚线粒体蛋白质定位数据．为了保证数据的可靠性，需要对得到的蛋白质序列数据进行验证和筛选：去除“unreviewed”类型的蛋白质序列；去除中含有不明确氨基酸的蛋白质序列，例如“X”，“B”，“Z”；去除“fragment”类型的蛋白质序列；去除定位于多个亚线粒体位置的蛋白质序列；序列的长度均大于 15 个氨基酸长度．考虑到序列间过高的相似度会对预测结果真实性产生影响，同时保证训练集包含足够的数据，我们使用 CD-HIT[36]去除相似度过高的序列，使得数据集内的任意两个序列的相似度小于 40%．

最后，共获得 1 293 条定位到 4 个亚线粒体位置上的蛋白质序列，包括 793 条线粒体内膜蛋白质序列，287 条线粒体基质蛋白质序列，172 条线粒体外膜蛋白质序列和 41 条线粒体膜间隙蛋白质序列．

1.2 同源-GO 信息库

基因本体(gene ontology，GO)信息是一类描述基因和基因产物的注释信息[37]，包含生物过程、分子功能和细胞组分 3 个部分，每一种 GO 信息都有一个唯一 ID 与之对应，并使用 GO term 进行描述，如 GO:0005743 代表“线粒体内膜”这个细胞组分．在目前的蛋白质亚细胞定位预测研究中，GO 信息已经被作为一种特征信息，用于蛋白质序列的特征表示[16, 38-41]．Du[42]使用了低相似度序列的 GO 信息来对蛋白质亚线粒体定位进行预测，取得了较好的预测结果．现有的基于 GO 信息的研究方法可分为两类：a．使用目标序列自身的 GO 信息结合氨基酸组成等信息对目标序列进行特征表示；b．借用目标序列的同源序列 GO 信息来扩充目标序列的 GO 信息．本文将预测的蛋白质序列称为目标序列，由于目标序列自身的 GO 信息可能不完备，我们建立了一个同源 GO 信息库，使用同源序列的 GO 信息对目标序列进行特征表示．同源 GO 信息库的建立基于 SWISSPROT 数据库，由于 GO 信息不完备的同源蛋白质序列无法对目标序列的特征信息起到扩充作用，这类蛋白质序列在同源 GO 信息库中没有意义，因此我们去除 SSWISS-PROT 数据库中 GO 信息不完备的蛋白质序列，保证得到的数据库里每条序列至少含有一条 GO 信息．

1.3 GO 基建立

对目标蛋白质序列的特征信息进行有效提取是预测的重要部分，本文使用 GO 向量作为特征参数对目标序列进行数学表征．在使用 GO 信息建立蛋白质序列的特征参数时，首先需要建立一个标准的 GO 基，然后把每一条序列的同源序列的 GO 信息在 GO 基上进行映射，就可以得到序列的特征参数．自从伪氨基酸组分(PseAAC)[13]被提出用来表示蛋白质的序列特征信息，该方法已广泛应用于蛋白质序列相关问题的研究．我们利用 PseAAC 的一般形式把 GO 基 B 表示为：

$$B=[\psi_1, \psi_2, \cdots, \psi_u] \quad (1)$$

其中，u 表示 GO 基的维数，ψ_u 表示 GO 基的第 u 个坐标 GO ID．因此，每一条目标蛋白质序列根据自己的 GO ID 均可以映射到 u 维的 GO 基上，组成 GO 向量特征参数．

可以看出一条蛋白质序列信息完全由 GO 基决

定，因此 GO 基的选择对于预测结果起决定性作用．本文使用 PSI-Blast[43]工具对建立的亚线粒体蛋白质定位数据集在同源 GO 信息库中进行同源性检索，设置每条亚线粒体蛋白质序列输出最多 11 条同源序列，设定 *E*-value 小于 0.00001．由于目标亚线粒体蛋白质序列存在于检索后的同源 GO 信息库中，需要去除目标亚线粒体蛋白质序列，这样得到每条目标亚线粒体蛋白质序列的 10 条同源序列的 GO 信息．1 293 条定位到 4 个亚线粒体位置上的蛋白质序列共产生 12 930 条同源序列．

GO 基的选择应使得 GO 向量能够反映所有目标蛋白质序列的同源 GO 信息．最简单的办法是把 12 930 条同源序列的 GO ID 去重，然后按照 GO ID 从小到大的顺序建立 GO 基．但是，由于 GO 基维数太大，而序列的 GO ID 较少，会使得序列的 GO 信息在 GO 基映射后的许多特征分量为零．因此，选择出现频率较高的一些 GO ID 作为基本的 GO 基，可以在最大程度上反映目标蛋白质序列的特征信息，同时保证最后产生的 GO 向量的维数不会太大．但是，仍然有少许目标蛋白质序列的同源 GO 信息在使用这些基本的 GO 基时，各个特征分量为零，考虑到这些目标蛋白质序列的特殊性，分析它们的同源 GO 信息，再对基本 GO 基进行补充，使得所有具有同源 GO 信息的目标蛋白质序列能够完整地使用 GO 向量表示．补充后的 GO 基表示为：

$$B'=[\psi_1, \psi_2, \cdots, \psi_j, \cdots \psi_u; \phi_1, \phi_2, \cdots, \phi_k, \cdots, \phi_v] \quad (2)$$

其中，ψ_j 表示基本 GO 基对应的特征值，ϕ_k 表示补充 GO 基对应的特征值．

为了确定 GO 基，首先需要计算数据集中每个 GO ID 出现的频率．本文使用了两种方法来确定 GO ID 出现的频率(f)：

a．依据目标蛋白质序列计数的频率(f_p)

数据集中定位在不同亚线粒体位置的目标序列的数量不同，因此不能简单地使用同源 GO 信息中 GO ID 出现的次数作为选择 GO 基的标准．我们将 GO ID 出现的频率表示为：

$$f_P=\frac{N_i}{N_x}(x=1, 2, 3, 4) \quad (3)$$

其中 N_x 表示定位在亚线粒体 x 位置处的目标蛋白质序列的总条数，N_i 表示第 i 个 GO ID 出现的总次数．

b．依据 GO ID 总数的频率(f_G)

定位在不同亚线粒体位置的同源 GO 信息中 GO ID 出现的总数也有差异，将 GO ID 出现的频率表示为

$$f_G=\frac{N_i}{G_x}(x=1, 2, 3, 4) \quad (4)$$

其中 G_x 表示在定位于亚线粒体 x 位置的目标蛋白质序列的同源 GO 信息中，所有 GO ID 出现的总次数，N_i 表示第 i 个 GO ID 出现的总次数．

可以依据上述两种方式建立 GO 基，对于预测结果的影响见表 1．

1.4 GO 向量映射

特征向量中每个分量对应的特征值是目标蛋白质序列特征信息的直接体现，特征值的确定对于预测模型至关重要．要把目标序列转换为特征向量，这就需要把目标序列的 10 个同源序列的 GO ID 映射到建立好的 GO 基 B' 上．本文使用两种方法来映射每个 GO ID 特征值：

a．0 或 1

对目标蛋白质序列 P^i 的 10 个同源序列的 GO ID 信息在 GO 基上映射，当与 GO 基的某一分量 ψ_j 或 ϕ_k 相同时，则将目标序列向量中与 GO 基对应位置的特征值记为 1，否则为 0，不计重复次数，即：

$$P^i=[a_1^i, a_2^i, \cdots, a_j^i, \cdots a_{u+v}^i], a_j^i=\begin{cases}1 & \text{GO-ID hit}\\ 0 & \text{Otherwise}\end{cases} \quad (5)$$

其中 P^i 为第 i 条目标序列的 GO 向量，a_j^i 表示第 j 维特征分量的特征值．

b．频次信息 n

考虑到同一 GO ID 在不同的序列中出现的频次不同，如果把频次信息 n 加入特征向量，则目标蛋白质序列 P^i 可以表示为：

$$P^i=[n_1^i, n_2^i, \cdots, n_j^i, \cdots n_{u+v}^i] \quad (6)$$

其中 n_j^i 表示第 i 条目标蛋白质序列 P^i 对应的 10 个同源序列中第 j 维 GO ID 出现的次数．

本文所使用的 GO 向量建立的理论方法简单，计算速度快，容易得到表示蛋白质序列的特征向量，在后续的机器学习算法能够得到很好的分类效果．并且，GO 基一旦确定，任何新的序列都可以映射到 GO 基上，产生特征向量．Du[42]通过 Blast 建立一个 GO 的超集，然后通过一个超表达 GO 打分函数来进行分类，尽管结果比较理想，但是理论比较复杂．

1.5 支持向量机(SVM)

支持向量机(SVM)是一种基于统计学习理论的

机器学习算法[44]，基于 SVM 的机器学习算法已广泛应用于蛋白质亚细胞定位预测、膜蛋白类型预测、蛋白质结构分类预测等研究中[45-52]. SVM 首先将目标向量数据映射到一个或多个高维空间，称为向量空间，然后在向量空间中建立超平面，通过寻找间隔最大的超平面作为分界面，目标数据可以在向量空间中被这些超平面分类.

本文使用 lib-svm[53]软件包来预测蛋白质的亚线粒体定位，使用径向基函数(RBF)作为核函数. 对于多分类系统，SVM 采用一对一方法，构建 $k\times(k-1)/2$ 个分类器，再采用投票方法来决定目标数据的类别. 采用网格搜索方法来寻找合适的惩罚因子常量 C 和 RBF 核函数参量 γ.

2 结果与讨论

2.1 评价方法

在统计预测中，通常用来检验分类器有效性的方法主要有独立数据集检验、抽样检验和 Jackknife 检验[54]. 根据 Chou 等[38, 55]的研究，在这三种方法中，Jackknife 检验被认为是最客观的[56-58]，对于给定的一个标准数据集，Jackknife 总是趋向一个唯一的预测结果. 因此，Jackknife 被研究者广泛地应用于各种分类器的检验中[59-60]. 在 Jackknife 检验过程中，数据集中的每一条目标蛋白质均会被隔离出来作为测试集，剩下的作为训练集，利用 SVM 进行训练并预测.

本文使用 Jackknife 检验来验证分类结果，预测结果还通常使用以下几种方法作为评价标准：敏感性(SN)、特异性(SP)、平均预测准确率(ACC)、Mathew 相关系数(MCC)[32, 49-50, 57].

2.2 预测结果

经过计算，依据目标蛋白质序列计数的频率选择 GO 基时，u=250，v=4，即选择 250 个基本 GO 基和 4 个补充 GO 基；依据 GO ID 总数的频率选择 GO 基时，u=300，v=4，即选择 300 个基本 GO 基和 4 个补充 GO 基. 然后再结合两种特征值的确定方法，建立了 4 个分类器. 图 1 显示了目标蛋白质序列被预测到 1 个亚线粒体位置上流程，4 个分类器经过 Jackknife 检验的平均预测准确率(ACC)结果见表 1.

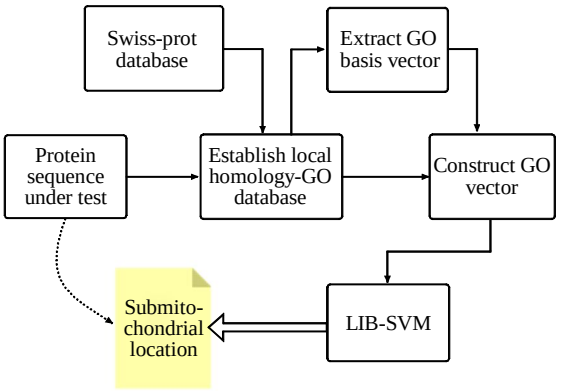

Fig. 1 The framework of the model for predicting protein submitochondria locations

Table 1 The predictive accuracy for different classifiers

Features	GO base	
	Based on f_P	Based on f_C
0 or 1	93.27%	93.12%
Counts(n)	92.81%	93.12%

从表中可以看出，预测模型在使用基于目标蛋白质序列计数的频率选择 GO 基以及选择 0 或 1 表示特征值的方法组合时，得到最高总体预测准确率为 93.27%. 表 2 显示了该方法在不同亚线粒体位置的具体预测准确率，其中预测模型对于线粒体内膜的预测准确率 94.35%，对于线粒体基质的预测准确率为 98.17%，对于线粒体外膜的预测准确率为 88.57%，对于线粒体膜间隙的预测准确率为 97.29%. 其中，使用 RBF 作为核函数，经网格搜索得到的最优参数为 C=32，γ=0.031 25.

Table 2 The predictive accuracy for optimized methods

Submitochondria locations	*TP*	*TN*	*FP*	*FN*	*ACC*/%	*MCC*
Inner membrane	763	457	43	30	94.354	0.881
Matrix	501	520	14	5	98.173	0.964
Outer membrane	108	140	25	7	88.571	0.776
Intermembrane space	127	124	3	4	97.287	0.946
Overall predictive accuracy/%			93.27			

2.3 与其他方法对比

为了评估模型的预测能力，使用数据集中定位在线粒体内膜、基质和外膜 3 类亚线粒体位置的目标蛋白质作为训练集对本模型进行训练和 Jackknife 检验，并与 Submito、GP-Loc、Predict_subMITO 和 Fan 的方法进行对比，表 3 列出了本方法与其他预测方法的对比情况，可以看出本方法在 3 类亚线粒体位置预测中，获得了 94.73%的总体预测准确率，线粒体外膜的预测准确率为 91.54%，高于其他预测模型.

Table 3 The comparison of the predictive accuracy with other methods

Submitochondria locations	ours		Submto		GP-loc		Predict_subMITO		Fan	
	ACC/%	*MCC*	*ACC*/%	*MCC*	*ACC*/%	*MCC*	*ACC*/%	*MCC*	*ACC*/%	*MCC*
Inner membrane	94.35	0.88	85.5	0.79	83.21	0.8	91.8	0.79	94.7	0.91
Matrix	99.21	0.98	94.5	0.77	97.24	0.85	96.4	0.79	99.3	0.96
Outer membrane	91.54	0.83	51.2	0.64	78.05	0.77	66.1	0.63	80.5	0.84
Overall predictive accuracy/%	94.73		85.2		89		89.7		94.95	

2.4 讨论

在蛋白质亚线粒体定位预测算法中，使用 GO 信息可以极大地提高预测结果的准确度. 在以往的亚细胞定位预测中也有学者把 GO 信息作为蛋白质的特征向量使用[16, 38-41]，这些算法是直接针对目标蛋白质序列的 GO 信息进行表示的. 因为 GO 信息包含亚细胞定位信息，所以直接使用虽然能够把预测精度提高到接近 100%，但是在方法上不合理，相当于使用已有的定位信息预测定位，这就要求把 GO 定位信息去除，用 GO 信息的生物学过程和分子功能做特征向量[32]，这种方法的缺点是大量蛋白质的 GO 信息不全，导致预测准确率下降.

在本文中，没有直接使用目标序列的 GO 信息，而是使用了和目标蛋白质序列同源的蛋白质序列的 GO 信息，一方面由于序列数量大幅度增加，使得 GO 信息相对较完整，即使目标序列没有 GO 注释，也可以依靠和它同源的 10 个序列的 GO 信息进行描述. 另外，由于没有直接使用目标序列的 GO，同源序列的 GO 信息中亚细胞定位的 GO信息也可以用来建立特征向量，这在一定程度上对预测精度的提高有很大帮助. 该算法如果结合其他方法[49-50, 57]，在预测精度，特别是外膜的预测上还能有所提升.

3 结 论

本文建立了一个较新的蛋白质亚线粒体定位数据集和一个包含 GO 信息的同源 GO 信息库，使用 GO 信息和同源信息构建蛋白质的特征向量，经 Jackknife 检验，该模型在线粒体内膜、基质、外膜以及膜间隙 4 个亚线粒体位置预测中取得了 93.27%的总体准确率，线粒体外膜的准确率较以往研究方法都有所提高.

比较以往的数据集，在我们构建的蛋白质亚线粒体定位数据集中加入了膜间隙定位数据，在数据方面比较完备. 膜间隙在细胞活动中的主要功能是进行氧化磷酸化，含有众多生化反应底物、可溶性的酶和辅助因子等，膜间隙蛋白质在诱导细胞凋亡中具有重要作用. 亚线粒体定位数据的完备以及预测精度的提高可以为进一步理解细胞活动以及内部的生物化学等过程具有重要的意义. 值得注意的是，在一条蛋白质序列的 GO 信息中往往会出现多个具有不同分子功能和生物学过程的注释信息，类似于氨基酸二肽组合，此类信息组合对于蛋白质的功能可能具有重要作用，在将来的研究中有待于进一步挖掘.

参 考 文 献

[1] Green D R, Reed J C. Mitochondria and apoptosis. Science, 1998, **281**(5381): 1309-1312

[2] Jassem W, Fuggle S V, Rela M, *et al*. The role of mitochondria in ischemia/reperfusion injury. Transplantation, 2002, **73**(4): 493-499

[3] Hutchin T, Cortopassi G. A mitochondrial DNA clone is associated with increased risk for Alzheimer disease. Proc Natl Acad Sci USA, 1995, **92**(15): 6892-6895

[4] Orth M, Schapira A H. Mitochondrial involvement in Parkinson's disease. Neurochem Int, 2002, **40**(6): 533-541

[5] Gerbitz K D, Gempel K, Brdiczka D. Mitochondria and diabetes. Genetic, biochemical, and clinical implications of the cellular energy circuit. Diabetes, 1996, **45**(2): 113-126

[6] Emanuelsson O, Nielsen H, Brunak S, *et al*. Predicting subcellular

localization of proteins based on their N-terminal amino acid sequence. J Mol Biol, 2000, **300**(4): 1005−1016

[7] Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci, 1999, **24**(1): 34−36

[8] Andrade M A, O'donoghue S I, Rost B. Adaptation of protein surfaces to subcellular location. J Mol Biol, 1998, **276**(2): 517−525

[9] Cedano J, Aloy P, Perez-Pons J A, *et al*. Relation between amino acid composition and cellular location of proteins. J Mol Biol, 1997, **266**(3): 594−600

[10] Cui Q, Jiang T, Liu B, *et al*. Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. BMC Bioinformatics, 2004, **5**(9): 1−7

[11] Zhou G P, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins, 2003, **50**(1): 44−48

[12] Cai Y D, Liu X J, Xu X B, *et al*. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem, 2002, **84**(2): 343−348

[13] Chou K C. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins, 2001, **43**(3): 246−255

[14] Huang Y, Li Y. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics, 2004, **20**(1): 21−28

[15] Park K J, Kanehisa M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics, 2003, **19**(13): 1656−1663

[16] Chou K C, Cai Y D. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochem Biophys Res Commun, 2004, **320**(4): 1236−1239

[17] Guda C, Subramaniam S. pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. Bioinformatics, 2005, **21**(21): 3963−3969

[18] Chou K C, Cai Y D. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem Biophys Res Commun, 2003, **311**(3): 743−747

[19] Yuan Z. Prediction of protein subcellular locations using Markov chain models. FEBS Lett, 1999, **451**(1): 23−26

[20] Chou K C, Elrod D W. Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem Biophys Res Commun, 1998, **252**(1): 63−68

[21] Chou K C, Elrod D W. Protein subcellular location prediction. Protein Eng, 1999, **12**(2): 107−118

[22] Cai Y D, Liu X J, Xu X B, *et al*. Support vector machines for prediction of protein subcellular location. Mol Cell Biol Res Commun, 2000, **4**(4): 230−233

[23] Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics, 2001, **17** (8): 721−728

[24] Sarda D, Chua G H, Li K B, *et al*. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinformatics, 2005, **6**(4): 152−156

[25] Cai Y D, Chou K C. Using neural networks for prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol Cell Biol Res Commun, 2000, **4**(3): 172−173

[26] Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res, 1998, **26**(9): 2230−2236

[27] Chou K C, Shen H B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. J Proteome Res, 2006, **5**(8): 1888−1897

[28] Du P, Li Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics, 2006, **7**(6): 597−605

[29] Nanni L, Lumini A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids, 2008, **34**(4): 653−660

[30] Zeng Y H, Guo Y Z, Xiao R Q, *et al*. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J Theor Biol, 2009, **259**(2): 366−372

[31] Shi S P, Qiu J D, Sun X Y, *et al*. Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction. Biochim Biophys Acta, 2011, **1813**(3): 424−430

[32] Fan G L, Li Q Z. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. Amino Acids, 2012, **43** (2): 545−555

[33] Du P, Yu Y. SubMito-PSPCP: predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions. Biomed Res Int, 2013, **2013**(8): 263829−263829

[34] Lin H, Chen W, Yuan L F, *et al*. Using over-represented tetrapeptides to predict protein submitochondria locations. Acta Biotheor, 2013, **61**(2): 259−268

[35] Uniprot C. The universal protein resource (UniProt). Nucleic Acids Res, 2008, **36**(Database issue): D190−195

[36] Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics, 2001, **17**(3): 282−283

[37] Harris M A, Clark J, Ireland A, *et al*. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res, 2004, **32**(Database issue): D258−261

[38] Chou K C, Shen H B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat Protoc, 2008, **3**(2): 153−162

[39] Chou K C, Shen H B. Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PLoS One, 2010, **5**(6): e11335

[40] Chou K C, Shen H B. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS One, 2010, **5**(4): e9931

[41] Fyshe A, Liu Y, Szafron D, *et al*. Improving subcellular localization prediction using text classification and the gene ontology.

Bioinformatics, 2008, **24**(21): 2512−2517

[42] Du P. Predicting protein submitochondrial locations by incorporating gene ontology annotations of low similarity sequences. 中国科技论文在线, 2013, http://www.paper.edu.cn/releasepaper/content/201308-236

[43] Schaffer A A, Aravind L, Madden T L, *et al*. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res, 2001, **29**(14): 2994−3005

[44] Vapnik V N. An overview of statistical learning theory. IEEE Trans Neural Netw, 1999, **10**(5): 988−999

[45] Cai Y D, Feng K Y, Li Y X, *et al*. Support vector machine for predicting alpha-turn types. Peptides, 2003, **24**(4): 629−630

[46] Cai Y D, Lin S L, Chou K C. Support vector machines for prediction of protein signal sequences and their cleavage sites. Peptides, 2003, **24**(1): 159−161

[47] Cai Y D, Ricardo P W, Jen C H, *et al*. Application of SVM to predict membrane protein types. J Theor Biol, 2004, **226** (4): 373−376

[48] Cai Y D, Zhou G P, Chou K C. Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J, 2003, **84**(5): 3257−3263

[49] Fan G L, Li Q Z. Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. J Theor Biol, 2012, **304**(2012): 88−95

[50] Fan G L, Liu Y L, Zuo Y C, *et al*. acACS: improving the prediction accuracy of protein subcellular locations and protein classification by incorporating the average chemical shifts composition. Scientific World Journal, 2014, **2014**(2014): 864135

[51] 黄志华, 李明泓, 马原野, 等. 事件诱发电位信号分类的时空特征提取方法. 生物化学与生物物理进展, 2011, **38**(9): 866−871
Huang H Z, Li M H, Ma Y Y, *et al*. Prog Biochem Biophys, 2011, **38**(9): 866−871

[52] 王 宏, 曲晓莉, 赵 研, 等. 基于表达及网络拓扑结构挖掘动脉粥样硬化风险疾病基因. 生物化学与生物物理进展, 2010, **37**(8): 916−922
Wang H, Qu X L, Zhao Y, *et al*. Prog Biochem Biophys, 2010, **37**(8): 916−922

[53] Chang C C, Lin C J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol, 2011, **2**(27): 27

[54] Chou K C, Zhang C T. Prediction of protein structural classes. Crit Rev Biochem Mol Biol, 1995, **30**(4): 275−349

[55] Chou K C, Shen H B. Recent progress in protein subcellular location prediction. Anal Biochem, 2007, **370**(1): 1−16

[56] Feng Z P. An overview on predicting the subcellular location of a protein. In Silico Biol, 2002, **2**(3): 291−303

[57] Fan G L, Li Q Z. Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition. J Theor Biol, 2013, **334**(2013): 45−51

[58] Zhou X B, Chen C, Li Z C, *et al*. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol, 2007, **248**(3): 546−551

[59] Lin H, Ding H, Guo F B, *et al*. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett, 2008, **15**(7): 739−744

[60] Zhang G Y, Li H C, Gao J Q, *et al*. Predicting lipase types by improved Chou's pseudo-amino acid composition. Protein Pept Lett, 2008, **15**(10): 1132−1137

Predicting Proteins Submitochondria Locations Using Blast-GO*

NANG Yi, MEI Han-Xue, ZHAO Yan, HOU Bao-Yan, ZHAO Zhi-Yuan, FAN Guo-Liang**

(*Department of Physics, College of Sciences and Technology, Inner Mongolia University, Huhhot* 010021, China)

Abstract In this study, a novel protein submitochondia locations dataset was constructed which contained 1 293 proteins classified into four kinds of submitochondria locations. The GO information and homologous information was extracted to combine the feature vectors of proteins and the Supported Vector Machine algorithm was used to construct the classifier. As a result, by using the Jackknife Cross-Validation, an accuracy of 93.27% for four kinds of protein submitochondria locations and that of 94.73% for three kinds of protein submitochondria locations was obtained. Especially, the predictive accuracy for outer membrane of protein submitochondia locations was enhanced than previous methods. The data set of protein submitochondia locations constructed by ours has the intermembrane proteins compared to old ones. The intermembrane proteins have important functions in protein apoptosis. The integrity of data set and the improvement of prediction accuracy can help to understand the cell activity and internal biochemical process.

Key words submitochondria location, gene ontology, homologous information, Support Vector Machine

DOI: 10.16476/j.pibb.2015.0190

*This work was supported by grants from The National Natural Science Foundation of China (61461038), The Scientific Research Program at Universities of Inner Mongolia Autonomous Region of China (NJZY13014),The Natural Science Foundation of Inner Mongolia Autonomous Region of China (2013MS0504), The Program of Higher-level Talents of Inner Mongolia University (135147) and The Students Innovation Training Program of the Inner Mongolia University(201412155).

**Corresponding author.

Tel: 86-471-4992958, E-mail: fanguoliang@imu.edu.cn, eeguoliangfan@sina.com

Received: June 25, 2015 Accepted: September 21, 2015