Hans Journal of Data Mining 数据挖掘, 2013, 3, 6-11
http://dx.doi.org/10.12677/hjdm.2013.31002 Published Online January 2013 (http://www.hanspub.org/journal/hjdm.html)

Extreme Learning Machine for Protein Subcellular Localization from Primary Sequence*

Feng Shi, Hong Chen, Huijuan Xiong#

College of Science, Huazhong Agricultural University, Wuhan
Email: #xiongdou1231@gmail.com

Received: Sep. 28th, 2012; revised: Oct. 26th, 2012; accepted: Nov. 4th, 2012

Abstract: Predicting protein subcellular localization from primary sequence is crucial to genome annotation, protein function prediction, drug discovery and etc. Extreme learning machine is an attractive learning method in recent years. This paper explores the potential of extreme learning machine for protein subcellular localization prediction. For this, a new feature selection strategy is established first. By utilizing the feature selection strategy, each primary sequence can be expressed as a 25-dimensional numerical vector. Furthermore, some numerical comparisons of Support Vector Machine with new features, Extreme Learning Machine with new features and another existing Support Vector Machine method with Pseudo amino acid composition features are given on 852 mycobcterial proteins data. The data arises from Swiss-Prot 48 database and belongs to four different classes. Results of five cross-validation for 852 protein sequences show that ELM with new features achieves the best accuracy. It achieves 97.2% accuracy, SVM with new features obtains 96.4% accuracy and SVM with Pseudo amino acid composition features displays 95.2% accuracy.

Keywords: Protein Subcellular Localization; Extreme Learning Machine; Homologous Protein

基于一级序列预测蛋白质亚细胞定位的超级学习机方法*

石　峰，陈　洪，熊慧娟#

华中农业大学理学院，武汉
Email: #xiongdou1231@gmail.com

收稿日期：2012 年 9 月 28 日；修回日期：2012 年 10 月 26 日；录用日期：2012 年 11 月 4 日

摘　要：蛋白质一级序列的亚细胞定位在基因组注释、蛋白质功能预测、药物发现等领域起着重要作用。超级学习机是近年来新兴的机器学习方法。本文探讨了超级学习机在蛋白质亚细胞定位预测中的潜力。为此，我们首先给出了一种新的特征提取策略，将每个蛋白质一级序列表示成 25 维的数值向量。在此基础上，我们将 852 组分枝杆菌蛋白质数据分别用基于新特征的支持向量机方法、基于新特征的超级学习机方法和已有的基于伪氨基酸组成特征的支持向量机方法做数值试验。这 852 组数据从 Swiss-Prot 48 数据库中选取，分属于四个不同种类。通过在这些数据上做五折交叉数值比较发现，基于新特征提取策略的超级学习机方法的准确率最高，达到了 97.2%，超过基于新特征的支持向量机方法的 96.4%的准确率以及基于伪氨基酸组成特征的支持向量机方法的 95.2%的准确率。

关键词：蛋白质亚细胞定位；超级学习机；同源蛋白质

1. 引言

蛋白质亚细胞定位与蛋白质的结构与功能密切相关，真核细胞中，在细胞质中合成的蛋白质必须处于特定的亚细胞区域内(如细胞核、线粒体、细胞质等)

*资助信息：本文由国家自然科学基金(编号：11001092)及中央部属高校专项基金(编号：2011QC064)支持。
#通讯作者。

才能发挥其功能。因此，研究蛋白质的亚细胞定位是研究蛋白质功能的一个重要的手段。早期关于蛋白质亚细胞定位的工作多集中于实验方法。该方法能较为精确的进行亚细胞定位。但近年来，随着蛋白质测序手段的发展，数据库中蛋白质序列的数量迅猛地增加，序列和结构的数量差别也越来越大，实验方法花费过大且速度过慢，无法满足需要。越来越多的研究工作集中于考虑有效的计算方法进行蛋白质亚细胞定位预测。目前，通过计算方法来预测蛋白质亚细胞定位已是生物信息学领域被广泛研究的重要课题[1]。

1.1. 相关研究

在过去的一些年里，基于蛋白质序列信息预测蛋白质亚细胞定位的计算方法取得了很大的进展。总结起来，这些方法的大致思想都是先将序列表示成固定长度的数值向量，进而采用机器学习的一些技巧如支持向量机、人工神经网络、k-近邻等方法进行预测[2-8]。

在这些不同的计算方法中，如何有效的提取序列特征是影响预测准确率的关键因素之一。氨基酸组分是蛋白质序列最简单的特征，自 Nakai 和 Kanehisa 首先发现细胞内外的蛋白质中氨基酸的组分存在明显差别，并用氨基酸组分信息预测内外蛋白质之后[9]，蛋白质的氨基酸组分信息被广泛用于亚细胞定位的研究中。Chou 等人在此基础上，进一步的将组分的顺序信息考虑进来，给出了伪氨基酸组分特征[2]。除氨基酸组分特征之外，还有些附加信息如：如序列同源性、基因组功能注释、序列同源性等等都可作为序列特征。一般而言，这些附加特征对预测一些特定蛋白质的亚细胞定位有很好的改进作用。特别是序列的同源性特征，很多用于评价预测方法优劣的蛋白质标准测试数据通常具有高度的同源性。如果高度同源的蛋白质的功能相近或相同，那么他们的亚细胞定位一定有相似性。因此，蛋白质的同源相似性对于预测亚细胞定位也是极为有益的。但是，由于数据数量和质量的原因，目前利用同源相似性做定位预测的文献并不常见。

基于不同的特征，可以采用不同的算法进行预测，在这之中，统计学方法和机器学习方法是比较常用的方法。与传统的花费大且效率低的实验方法相比，诸如神经网络、支持向量机这样的机器学习方法能得到较为满意的预测效果，并且计算花费更少。虽然神经网络和支持向量机已经在已有的机器学习方法中占主要地位，但他们在学习速度和人工干预等方面还有很大的待提高空间。

超级学习机(Extreme Learning Machine, ELM，也称作单隐层前向回馈神经网络)是近年来新兴的机器学习方法。该方法的输入权重可随机选定[10]。与常规的神经网络和支持向量机方法相比，ELM 方法具备更少的计算量和更高可伸缩性的同时，数值实验时能达到与另两种方法相当的准确率[10,11]。最近，对 ELM 方法的理论分析和应用研究已经在计算智能和机器学习相关领域受到了广泛关注。已经有一些文献考虑对各种不同的变形 ELM 模型给出相应的优化算法以提高求解效率。基于 ELM 算法解决一些应用问题如函数逼近、分类问题、回归问题等工作也被越来越多的人研究[10-12]。但是将该方法用于生物信息相关领域的应用工作还很少。

1.2. 本文的工作

本文重点探讨 ELM 方法在蛋白质亚细胞定位预测中的应用潜力。我们以分枝杆菌蛋白质序列(mycobacterial protein)的测试数据为例来说明 ELM 方法在亚细胞定位预测中的效率。分枝杆菌的病源菌以其能导致多种肺结核疾病而闻名，成功预测该病菌的亚细胞定位在生物学上和病理学上都有重要意义。文献[6]首先考虑了该病菌的亚细胞定位的预测。他们利用支持向量机方法对 852 条分枝杆菌蛋白质序列的亚细胞定位进行预测。通过在四种亚细胞上用五折交叉验证，最终得到了最高准确率为 86.8%。文献[13]基于伪氨基酸组分特征(Pseduo Amino Acid Composition, PseAAC)，给出支持向量机方法和结合马氏判别分析的离散增量方法(Increment of Diversity combined with Modified Mahalanobis Discriminant, IDQD)两种方法，进一步改进了文献[6]的准确率。

本文主要考虑 ELM 方法预测 852 条分枝杆菌蛋白质序列的效率。基于文献[14]中关于序列相似性与亚细胞定位的一致性的密切联系的研究，我们首先给出蛋白质序列的新的特征提取策略。新策略提取蛋白质一级序列中的 20 个氨基酸组分特征，并额外提取一个表示序列同源程度的 5 维数值向量，最终将每个序列表示成一个 25 维的数值向量。基于该特征提取策略，将 ELM 方法用于训练和预测。为证明新特征

提取策略的优势，进一步将新方法(ELM-AAC-H)与已有的伪氨基酸组成特征下的 SVM 方法(SVM-Pse-AAC)[13]以及基于新特征的 SVM 方法(SVM-AAC-H)做数值比较。在分属四种不同类别的 852 组蛋白质序列数据上的数值结果显示：新特征提取策略下的 ELM 方法得到了最高的准确率。

2. 数据与方法

2.1. 数据描述

我们选用亚细胞定位预测的一个公用测试数据集——RH 数据集进行数值比较。该数据集是 Reinhardt 和 Hubbard 等人以 Swiss-Prot 数据库 33.0 版本的数据为基础，选出有明确亚细胞定位注释的蛋白质序列建立起来。我们的实验采用 Swiss-Prot 48 的数据，找出 852 组有亚细胞定位注释的真核蛋白质序列数据，这其中包含 340 条胞质定位蛋白(Cytoplasmic)、402 条积分膜蛋白(Integral membranes, Integra-membrane)，50 条分泌腺蛋白(Secretory)以及 60 条附着在膜的脂质锚蛋白(Proteins attached to the membrane by a lipid anchor, Membrane-attached)。进一步的，可用 CD-HIT 程序分析核对序列的一致性(该程序的介绍请见文献[15])。通过该程序，在 852 条蛋白质序列中，34.6%的序列的一致性不低于 90%，12.6%的序列一致性在 80%至 90%之间，6.9%的序列一致性在 70%至 80%之间，3.4%的序列一致性在 60%至 70%之间，1.6%的序列一致性在 50%至 60%之间，0.6%的序列的一致性在 40%至 50%之间，40.3%的序列的一致性不超过 40%。出于可行性考虑，我们首先采用少量数据进行测试比较，我们先采用具有 80%的一致性的数据做数据比较，然后在整个数据集上做测试。

2.2. 数据的特征表示

我们首先提取蛋白质序列中的 20 种氨基酸的组分特征。每个组分特征是该氨基酸在序列中的百分比，即：所有的 20 个氨基酸组分特征可用如下公式计算：

$$\lambda_i = \frac{\text{序列中第 } i \text{种氨基酸的总数}}{\text{序列中氨基酸的总数}}, i = 1, \cdots, 20 \quad (2.1)$$

除了 20 个氨基酸组成特征之外，蛋白质序列的相似程度也与亚细胞定位密切相关。在进化过程中，同源的蛋白质通常具有相同或相似的功能，它们的亚细胞定位也具有相似性。文献[16]对此做了广泛研究，该文献表明蛋白质序列同源相似程度越高(在一定的阈值之上)，它们出现在同一个亚细胞器中的可能性就越大。我们尝试给出 5 个用于核定蛋白质序列同源性的特征与 20 个氨基酸组成特征一起作为输入信息，以得到更高的预测准确率。

我们对整个训练蛋白质序列用序列比对软件 Blast 做同源相似性比对(关于该软件的具体介绍见文献[17])。我们先初始给定阈值 e = 0.1，0.01，0.001，最后选用 0.001 为最终阈值，因为基于该阈值给出的特征向量算得了最高的准确率。对给定的阈值，利用 Blast 将每条序列跟四类数据集中的训练序列做同源比对，e 值小于给定阈值的蛋白质被认为是同源蛋白。把所有同源蛋白统计起来，每个亚细胞都能得到一个打分，这样一共能得到 4 个亚细胞的分值。这四个分值构成了序列的四个特征。另外，给出第 5 个特征用于判断是否找到同源蛋白，如果在训练集中找到同源蛋白，该特征取为 1，不然值为 0。由此，每条蛋白质序列可以被表示成如下数值向量：

$$\boldsymbol{p} = \left[p_1, \cdots, p_{20}, p_{20+1}, \cdots, p_{20+5} \right]$$

各分量如下计算：

$$p_i = \begin{cases} \lambda_i, & 1 \le i \le 20 \\ \dfrac{\mu_j}{\sum_{j=1}^{4} \mu_j}, & 21 \le i \le 24, j = i - 20 \\ s, & i = 25 \end{cases}$$

这里 λ_i 如(2.1)式计算得到，μ_j 为第 j 个亚细胞蛋白质的分数，$s \in \{0,1\}$ 为判断是否找到同源蛋白的标记特征。下面不妨以 Membrane-attached 类别中的第 51 条序列为例来说明一下特征提取的大致过程，该序列名为“amla_51”，序列的总长度为 220，具体信息如下：

“MINVQAKPAAAASLAAIAIAFLAGCSSTKPVSQD
TSPKPATSPAAPVTTAAMADPAADLIGRGCAQYAA
QNPTGPGSVAGMAQDPVATAASNNPMLSTLTSALS
GKLNPDVNLVDTLNGGEYTVFAPTNAAFDKLPAA
TIDQLKTDAKLLSSILTYHVIAGQASPSRIDGTHQT
LQGADLTVGARDDLMVNNAGLVCGGVHTANATV

YMIDTVLMPPAQ”对该序列，我们首先将 20 个氨基酸组分按照“AVLWIFPMSTCQGHNRKDEY”的顺序统计各组分在整个序列中的百分比，如“A”在序列中出现 41 次，该组分的百分比为 41/220。进一步的，剔除该序列，利用 Blast 软件统计四个类别中剩余序列与该序列同源的序列条数信息。经过序列比对发现，在 Membrane-attached 类别剩余的 59 条序列中有一条与“amla_51”同源的序列，在 Secretory 中有两条与之同源的序列，其余两类中都没有与之同源的序列。基于此，该序列向量表示的最后 5 个分量为“1/3, 0, 0, 2/3, 1”，这里最后一个分量 1 表示在数据集中找到了同源序列。基于如上步骤，我们得到序列“amla_51”的 25 维向量表示为“0.1864，0.0682，0.0864，0，0.0455，0.0136，0.0773，0.0318，0.0682，0.0955，0.0136，0.0455，0.0727，0.0136，0.0500，0.0136，0.0318，0.0636，0.0045，0.0182，0.3333，0，0，0.6667，1”。

2.3. 超级学习机方法

超级学习机方法(extreme learning machine, ELM)是近年较为热门的机器学习方法，该方法最初被作为单隐层前向回馈神经网络给出(single-hidden-layer feedforward neural networks, SLFNs)，进而被推广到一般的广义 SLFNs(见[8,10-12]等参考文献)。该方法的大致思想如下：

给定样本集 $\{x_i, y_i\}_{i=1}^N$，$x_i \in R^d$，$y_i \in \{1,-1\}$，ELM 方法的目标是希望找到如下决策函数：

$$f_L(x) = h(x)\beta$$

这里参数 $\beta = [\beta_1, \cdots, \beta_L]^T$ 为 L 个隐层结点与输出结点 $h_1(x), \cdots, h_L(x)$ 之间的权重向量，该参数向量通过最小化训练误差及输出权重，即解如下优化问题得到：

$$\text{Minimize: } \|\boldsymbol{H}\beta - T\|^2 + \frac{C}{2}\|\beta\|^2$$

C 为罚参数，$\boldsymbol{H}$ 为单隐层输出矩阵：

$$\boldsymbol{H} = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} h_1(x_1) & \cdots & h_L(x_1) \\ \vdots & \vdots & \vdots \\ h_1(x_N) & \cdots & h_L(x_N) \end{bmatrix}$$

输出结点 $h_1(x), \cdots, h_L(x)$ 如文献[10]所示，事先选定为满足 ELM 一致逼近定理的分片非线性连续函数，本文的数值实验中，我们从如下几种函数中选择，最后取准确率最高的：

1) Sigmoid 函数

$$h(x) = \frac{1}{1+\exp(1-(a^T x + b))}$$

2) Hard 极限函数

$$h(x) = \begin{cases} 1, & \text{如果} a^T x + b \geq 0 \\ 0, & \text{不然} \end{cases}$$

3) Gaussian 函数

$$h(x) = \exp(-b\|x - a\|^2)$$

上面三个函数中，参数 $\{a_i, b_i\}_{i=1}^l$ 分别在[0,1]区间内按均匀分布随机产生得到。

3. 数值结果与比较

3.1. 相关评价指标的定义

方法的预测能力通过测试数据上的相关指标评价得到，具体包括准确率(accuracy, Acc.)、Matthew’s 相关系数(Matthew’s correlation coefficient, MCC.)、灵敏度(sensitivity, Sens.)和精度(precision, Prec.)。这些评价指标的具体计算方法如表 1 所示，表一中的 TP 表示被正确识别的正类点数，FN 表示为正类但被识别为负类的错分点数，TN 表示被正确识别的负类点数，FP 表示为负类但被识别为正类的错分点数。

3.2. 算法的数值比较

文献[13]通过提取蛋白质序列的伪氨基酸组成特征(Pseudo amino acid composition features, SVM-

Table 1. Criteria employed in this paper
表 1. 本文使用的评价指标

评价指标	缩写	计算公式
精度	Prec.	$\frac{TP}{TP+FP}$
灵敏度	Sens.	$\frac{TP}{TP+FN}$
准确率	Acc.	$\frac{TP+TN}{TN+FN+TP+FP}$
Matthew’s 相关系数	MCC.	$\frac{TP*TN-FP*FN}{(TN+FN)(TP+FP)(TP+FN)(TN+FP)}$

PseAA)，进而采用支持向量机方法对 852 条真核蛋白质序列的亚细胞定位进行了预测。我们将本文的新方法(ELM-AAC-H)与文献[13]的方法(SVM-PseAAC)作比较。为揭示 ELM 方法的效率，我们还将本文的 ELM-AAC-H 方法与基于本文特征提取策略的支持向量机方法(SVM-AAC-H)做了比较。

所有方法在 Matlab 7.0 平台上数值实现。ELM 方法所需的隐层结点的个数在 $\{10,12,\cdots,100\}$ 当中分别选定测试，最后保留准确率最高的参数。隐层输出函数如本文的 2.3 部分里介绍的方式选取，该方法的具体代码可在 http://www.ntu.edu.sg/home/egbhuang 上下载。SVM 方法由软件 OSUsvm 实现，该软件是 SVM 方法较为成熟的软件 LIBSVM 的 Matlab 版本[8]。在 SVM 方法中，效益参数 C 在 $\{10,12,\cdots,100\}$ 中选取，核参数在 $\{1,2,\cdots,10\}$ 中选取。

注意到序列相似度对预测准确率有一定的影响，高度相似的数据可能会导致方法的过估计。为分析真核蛋白质亚细胞定位中序列一致性与预测准确率的关系，我们先只比较具有 80%序列一致性的不同数据值，然后再对整个数据集做数值比较。

852 条真核蛋白质序列分属于四个不同的亚细胞器，因此，这些序列的亚细胞定位预测问题本质是一个四类多分类问题。对一个 *k*-分类问题，常用“一对一”或“一对多”策略进行处理，将多分类问题转化为若干个二分类问题。“一对一”策略将每个类别两两比较，对每组测试数据的归属类别进行投票，投票数最多的那个类别即为测试数据的所属类别。用该策略处理 *k*-分类问题，最后需要构造 $C_k^2=\frac{k(k-1)}{2}$ 个分类器。“一对多策略”需要构造 *k* 个分类器，第 *i* 个分类器将所有训练数据中属于第 *i* 类的归为正类，其他剩余样本归属于负类。

本文采用“一对多”策略，构造 4 个二分类的分类器。对每个分类器，采用五折交叉验证做比对。保留准确率最高的结果，如表 2 和表 3 所示。为评价各个计算方法的计算花费，我们额外给出了表 4，列出每个方法的大致计算花费。

4. 结论

为更好的进行亚细胞定位预测，本文将与亚细胞定位密切相关的蛋白质序列的同源性特征考虑进来，给出了蛋白质一级序列的一个新特征提取策略，并进一步的探讨了近年比较流行的 ELM 方法在蛋白质亚细胞定位预测中的潜力。从表 2 和表 3 的结果来看，对具有 80%一致性的蛋白质序列和整个 852 条真核蛋白质序列分别作亚细胞定位预测时，采用 SVM 方法及新特征提取策略的准确率分别为 87.1%和 96.4%，要优于 SVM-PseAAC 方法的 82.2%和 95.2%。这说明新特征提取方法是行之有效的。在给出的三种方法中，ELM-AAC-H 方法所得的准确率最高，这说明 ELM 方法在解决亚细胞定位预测等生物信息挖掘领域中的实际问题具备一定潜力。

在计算花费比较的表 4 中，ELM 方法每步迭代要求解的线性方程组是结构特殊的严格正定矩阵。在实际计算时，采用一定的技巧可将它转换为一个依赖于

Table 2. Results for four protein data sets with 80% identity
表 2. 具有 80%一致性的蛋白质数据的计算结果

数据	方法(Prec., MCC(%))		
	SVM-PseAAC	SVM-AAC-H	ELM-AAC-H
Cytoplasmic	83.8, 74.5	89.9, 80.4	90.3, 81.3
Integra-membrane	84.5, 71.5	92.1, 85.2	92.1, 84.3
Secretory	69.2, 47.9	66.7, 72.3	69.4, 77.2
Membrane-attached	64.9, 62.7	60.0, 65.4	64.0, 66.5
平均准确率(Acc, %)	82.2	87.1	87.8

Table 3. Comparison results of total data sets
表 3. 整个数据集上的比较结果

数据	方法(Sens., %)		
	SVM-PseAAC	SVM-AAC-H	ELM-AAC-H
Cytoplasmic	96.8	90.0	96.7
Integra-membrane	94.8	98.8	98.2
Secretory	85.7	84.0	88.0
Membrane-attached	96.7	96.7	97.6
平均准确率(Acc, %)	95.2	96.4	97.2

Table 4. Computational cost of each method
表 4. 三种方法的计算花费比较

方法	每步迭代要求的线性方程组的系数矩阵	用 SMW 公式计算的花费	迭代次数
SVM-PseAAC	$DK(A,A^T)D, D\in R^{N\times N}, A\in R^{28\times N}$	$O(N^3)$	≥1
SVM-AAC-H	$DK(A,A^T)D, D\in R^{N\times N}, A\in R^{25\times N}$	$O(N^3)$	≥1
ELM-AAC-H	$\frac{I}{C}+HH^T, H\in R^{N\times L}$	$O(L^3)$	1

注：*L* 为隐层结点个数，*N* 为样本点个数。

结点个数 L 的 $L \times L$ 阶矩阵的求逆，而传统的 SVM 方法实际计算时每步迭代求的是一个跟样本点个数 N 相关的 $N \times N$ 阶矩阵的逆。从这个角度来说，ELM 方法的花费更多的依赖于隐层结点个数，当实际问题的数据量 N 远远超出结点个数 L 时，该方法的计算花费有其竞争力。从另一方面来说，由于算法依赖于隐层结点个数，当隐层结点个数过多时，算法的计算量会显著增加，但若隐层结点个数选择的过少，算法效率又会受到影响。如何选择合适的参数 L 在计算花费尽量少的同时获得尽量高的准确率，是算法后续可以探讨的问题之一。

参考文献 (References)

[1] T. Blum, S. Briesemeister and O. Kohlbacher. MultiLoc2: Integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinformatics, 2009, 10: 274.

[2] K. C. Chou, H.-B. Shen. Review: Recent progresses in protein subcellular localization prediction. Analytical Biochemistry, 2007, 370: 1-16.

[3] R. Casadio, P. L. Martelli and A. Pierleoni. The prediction of protein subcellular localization from sequence: A shortcut to functional genome annotation. Briefings in Functional Genomic Proteomic, 2008, 7(1): 63-73.

[4] K. C. Chou, H. B. Shen. 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.

[5] A. Garg, M. Bhasin and G. P. Raghava. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. Journal of Biological Chemistry, 2005, 280: 14427-14432.

[6] M. Rashid, S. Saha and G. P. S. Raghava. Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics, 2007, 8(1): 337.

[7] K.-C. Chou, Z.-C. Wu and X. Xiao. iLoc-Euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. Plos ONE, 2011, 6(3): e18258.

[8] C. C. Chang, C. J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Internet Systems and Technology, 2011, 2: 1-27.

[9] H. Nakashima, K. Nishikawa. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. Journal of Molecular Biology, 1994, 238(1): 54-61.

[10] G.-B. Huang, D.-H. Wang and Y. Lan. Extreme learning machines: A survey. International Journal of Machine Learning and Cybernetics, 2011, 2(2): 107-122.

[11] G.-B. Huang, Q.-Y. Zhu and C.-K. Siew. Extreme learning machine: Theory and applications. Neurocomputing, 2006, 70: 489-501.

[12] G.-B. Huang, H.-M. Zhou, X.-J. Ding and R. Zhang. Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man & Cybernetics-Part B: Cybernetics, 2012, 42(2): 513-529.

[13] H. Lin, H. Ding, F.-B. Guo, Y.-A. Zhang and J. Huang. Predicting subcellular localization of mycobaterial proteins by using Chow's pseudo amino acid composition. Protein & Peptide Letters, 2008, 15(7): 739-744.

[14] R. Nair, B. Rost. Sequence conserved for subcellular localization. Protein Science, 2002, 11(12): 2836-2847.

[15] Z. Lei, Y. Dai. Assessing protein similarity with gene ontology and its use in subnuclear localization prediction. BMC Bioinformatics, 2006, 7: 491.

[16] S. Mei, W. Fei and S. Zhou. Gene ontology based transfer learning for protein subcellular localization. BMC Bioinformatics, 2011, 12: 44.

[17] S. F. Altschul, T. L. Madden, A. A. Schaffer, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 1997, 25(17): 3389-3402.

