Hans Journal of Computational Biology 计算生物学, 2011, 1, 1-3
http://dx.doi.org/10.12677/hjcb.2011.11001 Published Online September 2011 (http://www.hanspub.org/journal/hjcb/)

Recognition of Subcellular Localization of Proteins Using of Sequences Fusion

Yun Jia
Department of Physics Experiment, School of Basic Science Inner Mongolia University of Technology, Hohhot
Email: yunbao2004haijun@163.com
Received: Sep. 18th, 2011; revised: Sep. 27th, 2011; accepted: Sep. 29th, 2011.

Abstract: Functional annotation of unknown proteins is a major goal in proteomics. A key annotation is the prediction of a protein's subcellular localization. We used the method of Increment of Diversity with Quadratic Discriminant analysis (IDQD) to predict subcellular localization of proteins which are recognized by the four plant categories and three non-plant and obtained accuracy 87.4(±0.5)% and 91.2(±0.2)%, respectively in 5-fold cross-validation test. Our result is better than comparable existing methods.

Keywords: Subcellular Localization; *F*-Value; Quadratic Discriminant Analysis

基于序列关联的蛋白质亚细胞定位识别

贾　芸
内蒙古工业大学理学院物理实验中心，呼和浩特
Email: yunbao2004haijun@163.com
收稿日期：2011 年 9 月 18 日；修回日期：2011 年 9 月 27 日；录用日期：2011 年 9 月 29 日

摘　要：对未知蛋白的功能注释是蛋白质组学的主要目标。一个关键的注释是蛋白质亚细胞定位的预测。应用基于序列关联的二次判别分析方法进行蛋白质亚细胞定位预测，对 4 个植物定位类型进行 5-fold 交叉检验。

关键词：亚细胞定位；*F* 值；二次判别分析

1. 引言

在后基因组时代随着蛋白质序列雪崩式的被测出，各种基于序列信息的方法被用于预测蛋白质亚细胞定位识别[1,2]。本文作者在研究生学习期间工作的基础上继续引入了 F 值参量结合多样性增量进行二次判别分析(IDQD)[3]方法对蛋白质亚细胞定位进行预测获得了一系列结果。

2. 数据集与方法

2.1. 数据集

本文使用了与 TargetP[4] 相同的数据集(http://www.cbs.dtu.dk/services/TargetP)。应用了植物类数据共 4 类 940 个蛋白质序列，包括叶绿体(chloroplast transit peptide, cTP)，线粒体(mitochondrial targeting peptide, mTP)，分泌途径(secretory pathway signal peptide, SP)和其它(other, OT)等 4 个类别。SP 类由内质网(endoplasmic reticulum, ER)，细胞外(extracellular space, EX)，高尔基体(golgi apparatus, GO)，溶酶体(lysosome, LY)，质膜(plasma membrane, PM)和液泡(vacuole, VA)等类别组成.OT 类由细胞质和细胞核蛋白组成.各类别的蛋白序列数详细情况见表 1。

Table 1. The number of protein sequences listed for each dataset according to localization
表 1. 依据亚细胞定位分类序列数据表

Set	The number of protein sequences			
	cTP	mTP	SP	OT
Plant	141	368	269	162

*基金项目：内蒙古工业大学校基金(ZS201124)。

2.2. 方法

2.2.1. *F* 值

F 值[5]的概念是在研究 DNA 序列间的统计关联时提出的，我们引用至蛋白质序列。令序列中氨基酸 *i* 的出现概率为 p_i(*i* = A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y)，一对氨基酸二联体 *i*, *j* 出现在相邻位点的联合概率为 $P_{i,j}$，

$$P_{i,j} = p_i p_{j|i} \tag{1}$$

$p_{j|i}$ 为单氨酸 *i* 后出现单氨酸 *j* 的条件概率。引入信息熵

$$H = -\sum_i p_i \log_2^{p_i} \tag{2}$$

和一阶信息冗余

$$D_1 = H_{\max} - H = 2 + \sum_i p_i \log_2^{p_i} \tag{3}$$

D_1 描述序列中氨基酸分布相对于随机等概率分布的偏离。引入马尔可夫熵 H_M (平均条件熵)，

$$H_M = -\sum p_i \sum_i p_{j|i^*} \log_2^{p_{j|i^*}} \tag{4}$$

和二阶信息冗余 D_2，

$$D_2 = H - H_M = 2H + \sum_{i,j} p_{ij} \log_2^{p_{ij}} \tag{5}$$

D_2 描述序列中氨基酸关联相对于独立序列的偏离。引入两条氨基酸序列 *A*、*B* 的关联，

$$F = \frac{2D_2(AB) - \left\{ \frac{n_A}{n_A + n_B} D_2(A) + \frac{n_B}{n_A + n_B} D_2(B) \right\}}{\left\{ \frac{n_A}{n_A + n_B} D_2(A) + \frac{n_B}{n_A + n_B} D_2(B) \right\}} \tag{6}$$

这里 n_A、n_B 分别表示氨基酸序列 *A*、*B* 的长度。

2.2.2. 信息的多样性增量(ID)处理

一般的，由已知的知识或经验可以从不同的侧面对样本提取多组特征信息。设一个样品的一组特征信息的某种分布可以由一个高维特征向量来表示，向量元素用整数表示。例如一条氨基酸序列可以用一个 400 维的向量表示，其中的向量元素为紧邻二联体在序列中出现频次的一个分布。第 *l* 个样本的第 *k* 组特征的特征向量为 X_k^l，该向量的第 *i* 个元素为 n_{ki}^l，则这组特征的多样性量定义为(1)式[6,7]。通常一组特征不是用样品本身就能表示清楚，而必须通过和标准样品(称为标准源)的比较来确定，亦即由样品特征的多样性分布和标准源特征的多样性分布的比较来确定。这时可以像(3)式那样通过定义待测样品与标准样品之间的多样性增量来度量它们的差异[7]。

$$\begin{aligned} D_k\left(\boldsymbol{X}_{\boldsymbol{k}}^{\boldsymbol{l}}\right) &= D_k\left(n_{k1}^l, n_{k2}^l, \cdots, n_{kd}^l\right) \\ &= N_k^l \log_2 N_k^l - \sum_{i=1}^{d} n_{ki}^l \log_2 n_{ki}^l \\ &\left(N_k^l = \sum_i n_{ki}^l\right) \end{aligned} \tag{7}$$

这里，*d* 为特征向量 $\boldsymbol{X}_{\boldsymbol{k}}^{\boldsymbol{l}}$ 的维数。

进一步 *s* 个标准样品(训练集)的这组特征多样性量为

$$\begin{aligned} D_k\left(\boldsymbol{S}_{\boldsymbol{k}}\right) &= D_k\left(m_{k1}, m_{k2}, \cdots, m_{kd}\right) \\ &= M_k \log_2 M_k - \sum_{i=1}^{d} m_{ki} \log_2 m_{ki} \\ &\left(M_k = \sum_i m_{ki}\right) \end{aligned} \tag{8}$$

这里，$\boldsymbol{S}_{\boldsymbol{k}}$ 为标准特征向量，$m_{ki} = \sum_{l=1}^{s} n_{ki}^l$。则待测样品与标准样品之间第 *k* 组特征的多样性增量定义为：

$$\mathrm{ID}_{\mathrm{k}}\left(\boldsymbol{X}_{\boldsymbol{k}}^{\boldsymbol{l}}, \boldsymbol{S}_{\boldsymbol{k}}\right) = D_k\left(\boldsymbol{X}_{\boldsymbol{k}}^{\boldsymbol{l}} + \boldsymbol{S}_{\boldsymbol{k}}\right) - D_k\left(\boldsymbol{X}_{\boldsymbol{k}}^{\boldsymbol{l}}\right) - D_k\left(\boldsymbol{S}_{\boldsymbol{k}}\right) \tag{9}$$

ID 表征了样品 *X* 和标准源信息参数分布的差异性，它提供了样品 *X* 特征的数量表示.

2.2.3. 信息的二次判别函数(QD)整合

对于一个 *c* 分类问题，假设对一个样品我们可以提取它的 *r* 组特征，构成一个 *r* 维的判别向量，则二次判别函数由下面的(4)式给出。

$$g_i(R) = \ln P_i - \frac{1}{2}\delta_i - \frac{1}{2}\ln\left|\Sigma_i\right| \tag{10}$$

$$\delta_i = (R - \mu_i)^T \Sigma_i^{-1} (R - \mu_i) \qquad (i = 1,2,\cdots,c)$$

其中 P_i 为第 *i* 类别的训练集的样品总数，$\boldsymbol{\mu}_{\boldsymbol{i}}$ 是第 *i* 类别的训练集中 *R* 的平均向量，δ_i 是第 *i* 类别的 *R* 与 $\boldsymbol{\mu}_{\boldsymbol{i}}$ 之间的马氏距离，$\boldsymbol{\Sigma}_{\boldsymbol{i}}$ 是第 *i* 类别的 $r \times r$ 维协方差矩阵，$\left|\Sigma_i\right|$ 是矩阵 $\boldsymbol{\Sigma}_{\boldsymbol{i}}$ 的行列式值。(4)式由 Bayes 理论导出。文献[8]和[9]的 IDQD 算法均是针对两分类问题的。在两分类问题时，分类是在 ξ 空间完成的，这时最佳分类域值可以由经验确定，往往域值 ξ_0 选择不是 0[8,9]。将 IDQD 算法应用于蛋白质定位或蛋白质二级结构预测这样的多分类问题时，判别规则要做一些小

的调整。多分类问题分类决策规则由文献[10]改为(11)式：

$$g_k(R) = \max\left(g_1(R), g_2(R), g_3(R), g_4(R)\right) \quad (11)$$

对于植物用(11)。即对每一类样品 R 都可以通过计算二次判别函数 $g_k(R)$，如果 $g_k(R)$ 最大，则 R 归为第 k 类。

2.2.4. 参数选取

基于 N 端序列的结构特征，本文选取了计算两条序列紧邻二联体关联 F 值和 N 端信号特征[3]，计算了节选不同氨基酸片断的结果见表 2；此外选取了计算两条序列间隔两个氨基酸的次次紧邻二联体关联 F 值和 N 端信号特征进行计算，见表 3；最后计算了紧邻 F 值和次次紧邻 F 值及 N 端作多样性增量的 QD 整合，见表 4。

3. 结果和讨论

3.1. 结果

作为对预测算法的预测能力的检验，我们采取了 5-fold 交叉检验，结果分别见表 5~7。

3.2. 讨论

由上述结果可知加入 F 值参数后预测结果没有显著降低和明显提高，但降低了参数种类，说明 F 值还是有意义的参数，以后将试图添加诸如结构类信息。

Table 2. The parameter of *F* value, ID value selection
表 2. *F* 值参数、ID 参数的选取

F 值参数	(1,50)	(1,50)	(1,30)	(1,50)
ID 参数	(2,40)	(2,20)	(2,20)	(2,80)

Table 3. The parameter of *F* value, ID value selection
表 3. *F* 值参数、ID 参数的选取

F 值参数	(1,50)	(1,40)	(1,40)	(1,40)
ID 参数	(2,90)	(2,30)	(2,70)	(2,130)

Table 4. The parameter of *F* value, ID value selection
表 4. *F* 值参数、ID 参数的选取

F 值参数(紧邻)	(2,40)
F 值参数(次次紧邻)	(2,40)
ID 参数	(1,40)

Table 5. Table 2 corresponding forecast result
表 5. 表 2 对应预测结果

	cTP	mTP	SP	OT
cTP	89/94/96/93			
mTP		330/326/324/332		
SP			240/239/233/237	
OT				119/119/113/123

Table 6. Table 3 corresponding forecast result
表6. 表3对应预测结果

	cTP	mTP	SP	OT
cTP	89/85/92/94			
mTP		339/333/332/329		
SP			238/243/240/242	
OT				126/130/128/128

Table 7. Table 4 corresponding forecast result
表 7. 表 4 对应预测结果

	cTP	mTP	SP	OT
cTP	57			
mTP		323		
SP			226	
OT				92

4. 致谢

感谢内蒙古工业大学校基金的资助。感谢吕军教授的悉心指导。

参考文献 (References)

[1] K. Nakai. Protein sorting signals and prediction of subcellular localization. Advances in Protein Chemistry, 2000, 54: 277-344.
[2] K. C. Chou, H. B. Shen. Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry, 2007, 370(1): 1-16.
[3] 贾芸, 赵巨东, 吕军. 基于*N*端信号的蛋白质亚细胞定位识别[J]. 内蒙古工业大学学报(自然门科学版), 2008, 27(2): 81-87.
[4] O. Emanuelsson, H. Niclsen, S. Brunak, et al. Predicting subcellular localization of proteins using amino acid terminal amino acid sequence. Journal of Molecular Biology, 2000, 300(4): 1005-1016.
[5] 罗辽复. 生命进化的物理观[M]. 上海: 上海科学技术出版社, 2000: 169-189.
[6] R. R. Laxton. The measure of diversity. Journal of theoretical biology, 1978, 70(1): 51-67.
[7] 徐克学. 生物数学[M]. 北京: 科学出版社, 1999: 278-286.
[8] L. R. Zhang, L. F. Luo. Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Research, 2003, 31(21): 6214-6220.
[9] 吕军, 罗辽复. 人类 PolII 启动子的识别[J]. 生物化学与生物物理进展, 2005, 32: 1185-1191.
[10] 边肇祺, 张学工等. 模式识别[M]. 北京: 清华大学出版社, 2004: 9-42.

