PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2008, 35(5): 531~535
www.pibb.ac.cn

基于序列保守性和蛋白质相互作用的真核蛋白质亚细胞定位预测 *

张 松 夏学峰 沈金城 孙之荣 **
(清华大学生物科学与技术系，教育部生物信息学重点实验室，生物膜和膜生物技术国家重点实验室，北京 100084)

摘要 蛋白质的亚细胞定位是进行蛋白质功能研究的重要信息．蛋白质合成后被转运到特定的细胞器中，只有转运到正确的部位才能参与细胞的各种生命活动，有效地发挥功能．尝试了将保守序列及蛋白质相互作用数据的编码信息结合传统的氨基酸组成编码，采用支持向量机进行蛋白质亚细胞定位预测，在真核生物中 5 轮交叉验证精度达到 91.8 %，得到了显著的提高．

关键词 亚细胞定位，氨基酸组成，序列保守性，蛋白质相互作用，支持向量机
学科分类号 Q6

生物体细胞是一个高度有序的结构，胞内根据空间分布和功能不同，可以分成不同细胞器或细胞区域．蛋白质合成后经蛋白质分选信号引导，被转运到特定的细胞器中，部分蛋白质则被分泌到细胞外或留在细胞质中，只有转运到正确部位才能参与细胞的各种生命活动，如果定位发生偏差，将会对细胞功能甚至生命产生重大影响．

随着蛋白质亚细胞定位信息的日渐重要，实验技术提供了一些比较精确的亚细胞定位数据．相对于蛋白质组学和功能组学的飞速发展，其定位注释远不能满足目前的研究需要．近年来，生物信息学在这方面开展了广泛的研究并且取得了一系列有意义的成果，亚细胞定位分析及预测加速了蛋白质结构和功能的研究．

蛋白质的亚细胞定位预测一直是生物信息学研究的重点问题．到目前为止，预测精度不断提高，取得了大量的研究成果[1]．预测方法的不同之处主要存在于两方面：第一，蛋白质信息的提取．主要是指将蛋白质相关特征信息提取之后转化成高维的特征向量，作为预测的输入．第二，根据提取的特征向量集，利用有效算法预测蛋白质的亚细胞定位，其中统计学和机器学习方法使用得最为广泛．现有的预测方法中精度的提高主要集中在蛋白质信息的提取上．蛋白质在合成过程中被分选到特定的亚细胞器中发挥生物学功能，很大程度上是由蛋白质的特征所决定的，包括分选信号、序列、结构域特征和残基的理化性质等等．预测过程中所采用的特征向量基本上都是基于某一特征或几个特征的综合．Nakai 等[2]首先利用 N 端分选信号对蛋白质亚细胞定位进行预测，建立了革兰氏阴性菌和真核细胞蛋白质定位预测系统，随后多个研究小组开始对其关注，出现了一系列文献．这种信息提取方法对于基因 5′区或者蛋白质 N 端序列的提取随意性较大，因此预测性能很大程度上依赖于基因 5′区或者蛋白质 N 端序列的选择．氨基酸组成是一种最基本的序列特征，也是亚细胞定位预测中使用得最为普遍的一种蛋白质特征信息．Nakashima 等[3]是最早注意到并开始使用氨基酸组成来预测蛋白质亚细胞定位的，他们根据氨基酸组成提出了一种分类算法，可以区分出胞内或者胞外蛋白质．Reinhardt 等[4]又基于人工神经网络进一步对 4 种真核生物和 3 种原核生物蛋白质的亚细胞进行了分类预测．随后还形成了一些衍生算法，如采用多种氨基酸序列

* 国家重点基础研究发展计划(973)(2003CB715900)，国家高技术研究发展计划(863)(2006AA020403)和国家自然科学基金(30770498)资助项目．
** 通讯联系人．
Tel: 010-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn
收稿日期：2007-09-03，接受日期：2007-11-06

作为特征信息的方法．除此之外，氨基酸的物理化学性质[5]、GO 注释[6]、同源蛋白的系统发育信息[7]和模体(motif)信息[8]等都被用来进行预测.

很早就有文献研究[9]表明：蛋白质序列相似程度越高，它们就越可能出现在同一个亚细胞器中．这在生物学意义上也是比较直观的，在进化过程中，同源蛋白具有相同或者相似的功能，所以其亚细胞定位也具有相似性．此外，蛋白质相互作用信息也能够给亚细胞定位预测带来一些重要线索，如果两个蛋白质有相互作用，那它们有很大可能出现在同一个亚细胞中[10]．但是，由于数据数量和质量的问题，这两种信息一直很少被用来进行亚细胞定位预测．本文中，我们尝试了将这两种信息结合常用的氨基酸组成作为输入信息，采用支持向量机作为分类器来进行预测，取得了非常好的效果，5 轮交叉验证总精度达到 91.8 %.

1 数据和方法

1.1 数据集

1.1.1 训练集．为了便于与以往预测算法的比较，选择在亚细胞定位预测的一个公用标准数据集——RH 数据集上进行测试．该数据集是 Reinhardt 和 Hubbard 以 SWISSPROT 33.0 数据库中的数据作为基础，筛选出其中有明确的亚细胞定位注释的蛋白质条目而建立的[4]．同时，植物蛋白质也被从该数据集中剔除．本文中的训练集选用其中的真核生物蛋白，包含 684 条胞质定位(cytoplasm)蛋白，325 条胞外定位(extracellular)蛋白，321 条线粒体定位(mitochondrial)蛋白，以及 1 097 条核定位(nuclear)蛋白.

1.1.2 亚细胞定位数据集．为了使用同源蛋白质的亚细胞定位数据，必须构建一个已知亚细胞定位的蛋白质数据集．目前已经有一些发表的亚细胞定位数据库，为了尽量保证注释的准确性，我们选择了 LOCATE [11]、MitoProteome [12]、Organelle DB [13]、DBSubLoc[14]以及 Swiss-Prot 综合数据库中的经过实验验证的真核生物蛋白质的亚细胞定位注释，将这些数据综合成一个较为完备的亚细胞定位数据集.

1.1.3 蛋白质相互作用数据集．目前较为常用的蛋白质相互作用数据库有 DIP[15]、MINT[16]和 BIND[17]，我们将这三个数据库和 Swiss-Prot 中的相互作用数据整合成一个数据集，去除掉冗余的以及自我相互作用的记录，得到一个共有 123 852 条相互作用记录的数据集.

1.2 方法

1.2.1 编码信息的提取.

为了得到具有保守序列的同源蛋白信息，我们将测试集中每一个亚细胞的蛋白质分别与构建的亚细胞定位数据集进行 Blast 同源比对，E 值小于 0.001 的蛋白质被认为是具有序列保守性的同源蛋白．把所有同源蛋白的亚细胞定位都进行统计，每一个亚细胞都会得到一个打分，这样共得到 4 个亚细胞的分值．第 5 维向量是判断参数，如果没有找到同源蛋白，则值为 1，否则为 0．这样从同源蛋白质数据集中我们得到了一个 5 维向量.

$$Score = \frac{N(i)}{N}$$

$N(i)$是所有同源蛋白中注释为第 i 个亚细胞的个数，N 是总的同源蛋白的个数.

对于蛋白质相互作用信息，首先在我们构建的蛋白质相互作用数据库中通过 Blast 找到同源蛋白，然后采用上面同样的方式统计所有与其同源蛋白有相互作用的蛋白质的亚细胞定位，得到一个 4 维向量．由于蛋白质相互作用目前数据量还相对较少，我们同样引进一个判断参数，如果蛋白质在该数据集中找不到同源蛋白或者其相互作用蛋白质没有亚细胞定位注释，那么第 5 维向量值为 1，否则为 0，这样也得到一个 5 维向量.

最后将传统的蛋白质组成编码得到的 20 维向量同上面得到的两个 5 维向量合并，得到一个 30 维的向量，以此作为支持向量机的输入.

1.2.2 分类器的选择．近年来统计学和机器学习等模式识别方法在预测算法中得到了广泛应用，最近邻法、神经网络、隐 Markov 模型、支持向量机和贝叶斯网络等机器学习算法都被用来进行亚细胞定位预测，本文中我们采用最为常用的是支持向量机作为分类器．它在蛋白质训练集中的高维特征向量空间中找到一个最优分割面，将不同类别的样本有效分开，并且使训练样本中离最优分割平面最近的样本点到该分割面的距离最大化．Hua 等[18]在 2001 年首先开始用支持向量机来进行亚细胞定位预测，取得了不错的精度．这里我们使用由 Chih-Chung Chang 和 Chih-Jen Lin 开发的 LIBSVM 2.8.1 软件包作为分类器．该软件可以在 http://www.csie.ntu.edu.tw/～cjlin/libsvm/ 下载.

1.2.3 结果的评估.

为了评价本文中采用的方法对于亚细胞定位预测的作用，以及与其他预测方法进行比较，我们采

用目前普遍认可的 5 轮交叉验证来评估预测结果．在评估过程中，样本集被随机分为 5 个均匀且不交叉的子集，在每次测试的时候，用其中的一个子集作为测试集，其余 4 个子集作为训练集，这样通过 5 轮测试后，取其平均值作为总的分类性能．

对于预测精度的评价指标，我们使用了目前这一领域中通用的指标，包括总体预测精度(the total prediction accuracy)，每一类的预测精度(the accuracy for each class)，Matthew 相关系数 (Matthew's correlation coefficients，*MCC*)．

$$Total\ accuracy = \frac{\sum_{i=1}^{k} TP(i)}{N}$$

$$Accuracy = \frac{TP}{TP + FP}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FP) \times (TN+FN)}}$$

TP：真阳性数目，*TN*：真阴性数目，*FP*：假阳性数目，*FN*：假阴性数目，*TP*(*i*)：第 *i* 个亚细胞中的真阳性数目，*N*：总的蛋白质数目，*k*：亚细胞种类数(这里共有四类)．

2 结　果

表 1 列出了我们的方法在 RH 数据集上测试的结果，同时表中也列出了一些比较有影响和在 RH 数据集上取得了较好结果的工作，包括人工神经网络[4]、马尔可夫模型[19]及使用全序列氨基酸组成的支持向量机[18]在 RH 数据集上测试的结果作为对比．表 2 列出了不同的编码信息提取方法的预测结果．

Table 1　Prediction results comparisons of different methods on RH eukaryotic dataset

Subcellular localization	ANN method		Markov method		AA composition+SVM method		Method in this article	
	Accuracy/%	*MCC*	Accuracy/%	*MCC*	Accuracy/%	*MCC*	Accuracy/%	*MCC*
Cytoplasmic	55	–	78.1	0.60	76.9	0.64	84.3	0.81
Extra-cellular	75	–	62.2	0.63	80.0	0.78	95.7	0.95
Mitochondrial	61	–	69.2	0.53	56.7	0.58	84.8	0.80
Nuclear	72	–	74.1	0.68	87.4	0.75	97.2	0.94
Total accuracy	66	–	73.0	–	79.4	–	91.8	–

Table 2　Prediction results comparisons of different feature inputs

Subcellular localization	AA composition		AA composition+ sequence homolog		AA composition+ protein-protein interaction		AA composition+sequence homolog+ protein-protein interaction	
	Accuracy/%	*MCC*	Accuracy/%	*MCC*	Accuracy/%	*MCC*	Accuracy/%	*MCC*
Cytoplasmic	76.9	0.64	72.3	0.70	75.4	0.72	84.3	0.81
Extra-cellular	80.0	0.78	92.7	0.90	89.1	0.87	95.7	0.94
Mitochondrial	56.7	0.58	83.2	0.62	75.2	0.65	84.5	0.79
Nuclear	87.4	0.75	96.1	0.92	93.8	0.86	97.2	0.94
Total accuracy	79.4	–	86.5	–	85.3	–	91.8	–

从以上结果可见，与其他几种预测方法相比，本文的预测结果无论是从总体的精度，还是各亚细胞器单独的预测精度来看都有了较大幅度的提高，尤其是与我们以前的工作 Subloc[18]算法比较，同样采用支持向量机作为分类器，加入了序列保守信息和蛋白质相互作用信息之后，总的预测精度提高超过 12 %．表 2 的结果表明：将序列保守信息、蛋白质相互作用信息和氨基酸组成结合作为编码信息的预测结果要明显好于单独用氨基酸组成或者与序列保守信息、蛋白质相互作用信息的单独组合．

3 讨　论

在选取同源蛋白时，我们选用 Blast 的 *E* 值为 0.001 作为阈值，这并非一个随机取值，取得过高会减少同源蛋白的数量，过低则会将大量非同源蛋白划分进来，带进噪声．经测试，取 0.001 的时候

预测结果最好.

采用氨基酸组成、序列保守信息和蛋白质相互作用信息相结合的编码方式进行亚细胞定位预测，取得了总体 92 %的精度．事实上，这个精度还能进一步提高．由于实验方法的误差等，目前的亚细胞定位和蛋白质相互作用数据假阳性比较严重，这肯定会给预测带来一定的噪声，从而影响预测精度．随着高精度的亚细胞定位和相互作用注释越来越多，其预测结果能够得到进一步的提高．

本文中的预测方法在真核生物蛋白质中取得了非常好的效果，但是要应用到原核生物蛋白质中还困难，其主要原因在于数据量的缺乏，目前，原核生物中经过实验验证的亚细胞定位和蛋白质相互作用数据都远比真核生物少，这使得蛋白质相互作用数据集成为一个稀疏集，较大程度地影响预测精度．但是随着原核生物蛋白质亚细胞定位和相互作用注释的逐步深入，此方法也能够应用到原核生物中并取得不错的结果．

总的来说，本文结合了氨基酸组成、同源蛋白保守序列和蛋白质相互作用作为信息输入，采用支持向量机进行亚细胞定位预测取得了非常高的精度，比起其他方法有了很大的突破，也为功能基因组学研究提供了一种新思路．

参 考 文 献

1 Zhang S, Huang B, Xia X, *et al*. Bioinformatics research in subcellular localization of protein. Prog Biochem Biophys, 2007, **34** (6): 573～579

2 Gardy J L, Spencer C, Wang K, *et al*. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res, 2003, **31**(13): 3613～3617

3 Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol, 1994, **238**(1): 54～61

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

5 Chou K C. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun, 2000, **278**(2): 477～483

6 Cai Y D, Chou K C. Predicting 22 protein localizations in budding yeast. Biochem Biophys Res Commun, 2004, **323**(2): 425～428

7 Marcotte E M, Xenarios I, van Der Bliek A M, *et al*. Localizing proteins in the cell from their phylogenetic profiles. Proc Natl Acad Sci USA, 2000, **97**(22): 12115～12120

8 Scott M S, Thomas D Y, Hallett M T. Predicting subcellular localization via protein motif co-occurrence. Genome Res, 2004, **14** (10A): 1957～1966

9 Nair R, Rost B. Sequence conserved for subcellular localization. Protein Sci, 2002, **11**(12): 2836～2847

10 Huh W K, Falvo J V, Gerke L C, *et al*. Global analysis of protein localization in budding yeast. Nature, 2003, **425**(6959): 686～691

11 Fink J L, Aturaliya R N, Davis M J, *et al*. LOCATE: a mouse protein subcellular localization database. Nucleic Acids Res, 2006, **34**(Database issue): D213～217

12 Cotter D, Guda P, Fahy, E, *et al*. MitoProteome: mitochondrial protein sequence database and annotation system. Nucleic Acids Res, 2004, **32**(Database issue): D463～467

13 Wiwatwattana N, Kumar A. Organelle DB: a cross-species database of protein localization and function. Nucleic Acids Res, 2005, **33** (Database issue): D598～604

14 Guo T, Hua S, Ji X, *et al*. DBSubLoc: database of protein subcellular localization. Nucleic Acids Res, 2004, **32**(Database issue): D122～124

15 Xenarios I, Rice D W, Salwinski L, *et al*. DIP: the database of interacting proteins. Nucleic Acids Res, 2000, **28**(1): 289～291

16 Zanzoni A, Montecchi-Palazzi L, Quondam M, *et al*. MINT: a molecular INTeraction database. FEBS Lett, 2002, **513**(1): 135～140

17 Bader G D, Betel D, Hogue C W. BIND: the biomolecular interaction network database. Nucleic Acids Res, 2003, **31** (1): 248～250

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

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

Eukaryotic Protein Subcellular Localization Prediction Based on Sequence Conservation and Protein-Protein Interaction*

ZHANG Song, XIA Xue-Feng, SHEN Jin-Cheng, SUN Zhi-Rong**

(*Institute of Bioinformatics and System Biology, MOE Key Laboratory of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Science and Biotechnology, Tsinghua University, Beijing* 100084, China)

Abstract Subcellular localization is a key characteristic of protein functional research. Proteins are transported to specific compartment after they are synthesized in cells. They can take part in the cell activity and function efficiently when in correct subcellular location. Sequence homolog, protein-protein interaction information and traditional amino acid composition are combined as input parameters of support vector machine (SVM) to predict eukaryotic protein subcellular localization. The total accuracy of 5-fold cross validation is 91.8%, which is higher than other methods.

Key words subcellular localization, amino acid composition, sequence conservation, protein-protein interaction, support vector machine

---

*This work was supported by a grant from National Basic Research Program of China (2003CB715900), Hi-Tech Research and Development Program of China (2006AA020403) and The National Natural Science Foundation of China (30770498).

**Corresponding author.

Tel: 86-10-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn

Received: September 3, 2007 Accepted: November 6, 2007