蛋白质亚细胞定位预测的机器学习方法

张树波[1,3] 赖剑煌[2,3]
(中山大学数学与计算科学学院 广州 510275)[1] (中山大学信息技术与科学学院 广州 510275)[2]
(广东省信息安全技术重点实验室 广州 510275)[3]

摘 要 蛋白质亚细胞定位与其功能密切相关。蛋白质在细胞中的正确定位是细胞系统高度有序运转的前提保障。研究细胞中蛋白质定位的机制和规律,预测蛋白质的亚细胞定位,对于了解蛋白质的性质和功能,了解蛋白质之间的相互作用,探索生命的规律和奥秘具有重要意义。基于机器学习方法的蛋白质亚细胞定位预测是生物信息学研究的热点之一。从数据集的建立、蛋白质序列特征刻画和蛋白质亚细胞定位预测算法3个方面,总结和评述了在过去十几年里机器学习方法在蛋白质亚细胞定位研究中的应用情况和取得的成果,分析了机器学习方法在蛋白质亚细胞定位预测方面存在的问题和面临的挑战,指出了蛋白质亚细胞定位研究的主要方向。

关键词 亚细胞定位,生物信息学,机器学习,分类器,特征提取
中图法分类号 TP3-05 **文献标识码** A

Machine Learning-based Prediction of Subcellular Localization for Protein

ZHANG Shu-bo[1,3] LAI Jian-huang[2,3]
(School of Mathmatics and Computational Science, Sun Yat-sen University, Guangzhou 510275, China) [1]
(School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510275, China) [2]
(Province Key Laboratory for Technology of Security Information, Guangzhou 510275, China) [3]

Abstract Subcellular localization of protein is closely related to its function, properly localization of protein in the cell is the precondition for the cell system to operate orderly, Probing into the mechanism and principle of protein sorting and predicting its subcellular localization can provide insight into the protein's properties and functional annotation of proteins, and it is significant meaningful to apprehend the nature of life. Predicting the subcelllular localization of protein has become a hot research field in the recent year. The development of constructing datasets, representing protein sequences, as well as classification algorithm in the area of protein subcellular localization prediction were reviewed and commented, and the challenge of machine learning methods faced in this field was then pointed out, in the end, perspectives in this realm were also proposed.

Keywords Subcellular localization, Bioinformatics, Machine learning, Classifier, Feature extraction

1 引言

蛋白质的亚细胞定位与其功能密切相关,同时是维持高度有序的细胞系统正常运转的保障,研究蛋白质的亚细胞定位,有助于了解蛋白质性质和功能,了解蛋白质之间相互作用和调控机制,为发明新药物提供参考信息。

随着人类在基因组学和蛋白质组学方面研究的不断深入,生物数据和生物序列迅速增长,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[1],预测精度不理想[2],已经无法满足生命科学研究的需要。大量原始数据的积累为生物信息学的发展提供了基本的素材,也为机器学习方法在生命科学领域的应用提供了舞台。近年来,基于机器学习方法的蛋白质亚细胞定位预测逐渐成为生物信息学的一个研究热点。经过多年的努力,机器学习方法在蛋白质亚细胞定位预测方面已取得一定的成果,已经开发出不少性能优良的蛋白质亚细胞定位预测系统并公布出来。表1列出了一些蛋白质亚细胞定位预测系统。

表1 蛋白质亚细胞定位在线服务系统

预测系统	网址	物种	亚细胞数量	参考文献
PSORT家庭	http://psort.nibb.ac.jp/	多种物种	3—12	[3-5]
TargetP家庭	http://www.cbs.dtu.dk/services/TargetP/	多种物种	2—5	[6-9]
NNPSL	http://www.sanger.ac.uk. Teams/faculty/hubbard/	原核/真核	3—4	[10]
DBSubloc	http://www.bioinfo.tsinghua.edu.cn/	原核/真核	3—4	[11]
Hum-mPLoc	http://22.120.37.186/bioinf/hum-multi/	人类	14	[12]

到稿日期:2008-05-15 本文受国家自然科学基金(No. 60675016, 60633030)资助。
张树波 博士研究生,研究方向为模式识别、生物信息学,E-mail:treaton@tom.com;**赖剑煌** 教授,博士生导师,研究方向为模式识别、数字图像处理。

Plant-PLoc	http://22.120.37.186/bioinf/plant/	植物	11	[13]
Virus-PLoc	http://22.120.37.186/bioinf/virus/	病毒	7	[14]
LOCSVMPSI	http://bioinformatics.uste.edu.cn/LOCSVMPSI/	真核	4—12	[15]

建立一个典型的基于机器学习方法的蛋白质亚细胞定位系统包括如下3个基本步骤：(1)建立数据集；(2)刻画蛋白质序列特征；(3)设计分类器。本文围绕这3个环节进行综述，探索机器学习方法在蛋白质亚细胞定位研究上面临的问题和今后发展的方向。第2节给出了蛋白质亚细胞定位研究所采用的主要数据集；第3节讨论了蛋白质序列特征的刻画方法；第4节评述了蛋白质亚细胞定位研究的一些主要算法以及分类器的评价标准；最后指出了蛋白质亚细胞定位研究的发展方向。

2 数据集的建立

研究蛋白质亚细胞定位的数据集基本来自SWISS-PROT数据库。该数据库建于1986年，是目前世界上存储蛋白质序列最主要的一级数据库之一。SWISS-PROT数据库中的蛋白质序列基本上是研究人员提交的原始数据，数据的质量没有保证。利用这个数据库研究蛋白质的亚细胞定位时，需要对其中的数据进行筛选。通常的筛选标准有：(1)物种类型。如果只研究某个特定物种的蛋白质亚细胞定位，则只挑选该物种的相关蛋白质序列。(2)亚细胞类型。在构建数据集时，需要知道每个蛋白质序列所在的亚细胞位置，所以只有包含明确的亚细胞定位信息的序列才被选入数据集中。(3)序列长度。长度太短的蛋白质序列很可能是碎片，不是完整的蛋白质序列，通常不会被选入数据集。(4)数据冗余度。为了消除样本之间的同源性对分类结果带来的偏差，在筛选过程中需要去掉一定的冗余度。早期的数据集对同源性的要求相对较低，通常只要数据集中序列的同源性小于90%即可。随着研究的深入，数据集的同源性要求越来越高，如Chou[17]在研究人类蛋白质亚细胞定位时，要求序列的同源性<25%。(5)样本量。考虑到样本的统计意义，需要排除掉包含样本量太少的亚细胞类别。

除了利用SWISS-PROT数据库外，研究人员根据不同的研究目的，建立了自己的蛋白质亚细胞信息数据集，包括PSORT家族、TargetP家族数据集等。早期的数据集比较简单，只包含两类蛋白质，数据集中蛋白质种类早期的只有2种。近年来，随着研究的不断深入，蛋白质序列数据集越来越复杂，目前最复杂的数据集是酵母蛋白质序列数据集，包含22种亚细胞蛋白质。表2列出了一些主要的数据集。

表2 蛋白质亚细胞定位数据集

数据集名称	网址	建立时间	参考文献
酵母数据集	http://www.yeastgfp.ucsf.edu	2003	[2]
PSORT家庭	http://psort.nibb.ac.jp/	1991	[3-5]
TargetP家庭	http://www.cbs.dtu.dk/services/	1997	[6-9]
NNPSL	http://www.sanger.ac.uk/Teams/facult/hubbard/	1998	[10]
Plant-PLoc	http://202.120.37.186/bioinf/plant	2007	[13]
UniProt	http://www.ebi.ac.uk/swissprot/index.html	1986	[16]
Hum-mPLoc	http://202.120.37.186/bioinf/hum-mult	2006	[17]
Organelle DB	http://organelledb.lsi.umich.edu/	2005	[18]
LOCATE	http://locate.imb.uq.edu.au/	2006	[19]

3 蛋白质序列特征的刻画

3.1 蛋白质亚细胞定位数据的特点

蛋白质亚细胞定位研究的基本数据是蛋白质序列。蛋白质序列是由20种常见的氨基酸残基以共价键形式相互作用构成的，它们与一般的数据不同，具有如下的特点：

1)蛋白质序列不是由数值型数据组成，而是用字符串表示；

2)蛋白质序列由20个字母组成，每个字母代表1种氨基酸；

3)蛋白质序列的长度不等，有的序列长度只有几十个氨基酸，有的包含几千个氨基酸；

4)蛋白质序列中的每个字母不是一般意义上的字母，它们具有特殊的生物学含义。

3.2 蛋白质特征的分析和提取

蛋白质序列特征的分析和刻画是亚细胞定位研究的重要内容。由于蛋白质序列具有上述特点，在分析蛋白质序列的特征时，不能仅仅将其作为一般的字符串进行处理，而需要考虑其物理化学特性和生物学意义。目前在蛋白质亚细胞定位研究的机器学习方法中，用来刻画蛋白质序列的特征主要有：

(1)蛋白质序列的N端信息

一般认为蛋白质在合成的过程中，其N端包含一些特殊的子序列(也称分选信号)，这些信号能够指导新合成的蛋白质分选到特定的亚细胞中。已经知道的蛋白质N端分选信号包括信号肽、线粒体转移肽、叶绿体运输肽、核定位信号、类囊体腔转移肽和过氧化物酶体定位信号等[20]。1991年，Nakai和Kanehisa建立了第一个亚细胞器定位预测系统[27]，在这个系统中利用了蛋白质的N端信号序列。随后，用N端分选信号刻画蛋白质的特征逐渐引起人们的重视。Emanuelsson等人[6]利用N端序列信息预测叶绿体运输肽，并开发了基于N端分选信号的蛋白质亚细胞定位预测方法[7,8]；Nielsen[9,21]用N端序列预测信号肽；Claros等人[22]用N端序列预测线粒体转移肽。最近，有学者[23,24]考虑用N端序列的氨基酸成分来刻画蛋白质序列。这种信息的有效性取决于蛋白质序列完整性，一旦蛋白质序列的N端信号不完整或者丢失，那它就失效。

(2)蛋白质序列的氨基酸组分(包括n-肽组分)信息

氨基酸组分是蛋白质序列最简单的特征。自从Nakai和Kanehisa[25]发现细胞内外的蛋白质中氨基酸的组分存在明显差异，并首先用氨基酸组分预测细胞内外蛋白质之后，蛋白质序列的氨基酸组分信息被广泛应用于亚细胞定位研究中。Reinhardt和Hubbard[10]基于氨基酸组分构造了预测蛋白质亚细胞定位第一个人工神经网络。Hua和Sun[11]利用氨基酸成分构造了第一个基于SVM的预测系统。Yuan[28]利用氨基酸成分信息构造了基于Markov链模型的预测方法。由于氨基酸组分的信息是一种全局信息，无法有效利用序列的顺序和位置信息，氨基酸组分的特征被推广到氨基酸二肽组分特征。Huang和Li[29]提出了基于氨基酸二肽组分的预测

方法。Yu 等人[30]更是提出了基于 n-肽信息的预测方法。氨基酸组分(或者 n-肽组分)反映的是蛋白质序列的整体信息，无法刻画序列的局部和顺序信息，同时这种信息只是将蛋白质序列看成一般的字符序列，没有考虑序列上氨基酸的物理化学性质，具有一定的局限性。

(3)蛋白质的功能域信息

蛋白质序列在长期的进化过程中，某些特定位点上的氨基酸残基具有高度的保守性。这些特定的位点联系着蛋白质特定的生物学功能，在生物学上被称为蛋白质的功能 motif。这些功能 motif 具有特异性，可以有效地刻画蛋白质序列。Horton[5]、Chou[31] 和 Scott[32] 等人将蛋白质序列上的功能 motif 信息用于预测亚细胞定位。蛋白质功能域信息具有很高的特异性，这种方法可靠性较高。但是，为了确定蛋白质序列中特定的功能域，要求功能域数据库必须包含足够多的功能域条目，否则无法确定序列中的功能域信息。

(4)序列比对信息

蛋白质的比对就是找出两条序列之间的一种最佳匹配。这种匹配对应一个数值型输出结果，用来度量序列之间的相似程度，这种相似程度的度量可以直接用于预测蛋白质的亚细胞定位。BLAST 是一个常用的序列比对算法，可用于蛋白质序列相似性计算。PSI-BLAST 是 BLAST 的一个改进版本，用于计算同源性较低的序列之间的相似性。Xie[15]、Bhasin[33]、Nair[34] 和 Guo[35] 等人将蛋白质序列的比对信息用于蛋白质亚细胞定位预测。比对方法既考虑序列的位置信息，也考虑序列中氨基酸相互替换的生物学含义，是一种比较有效的信息。

(5)GO 注释信息

由于蛋白质必须在特定的亚细胞中通过与其它的蛋白质进行相互作用才能执行特定的生物学功能，所以蛋白质的功能与它所处的亚细胞位置密切相关。如果知道了蛋白质的功能信息，就可以知道它所处的亚细胞位置。基因本体论(Gene Ontology，简称 GO)是一个公认的基因功能注释标准化项目，包括了分子功能、生物学过程和细胞组件 3 种基本的信息。Chou 在一系列的工作中[13,17,36-38]提出了将蛋白质序列的功能注释信息用于蛋白质亚细胞定位预测。与功能域信息一样，这种信息的有效性取决于 GO 数据库的完善程度。

(6)氨基酸物理化学性质

蛋白质序列是由氨基酸残基构成的，序列中氨基酸残基的物理化学性质从根本上决定了蛋白质序列的整体物理化学性质，因此氨基酸残基的物理化学性质是描述蛋白质序列的一种重要信息。Chou [39]利用氨基酸的亲水性、疏水性和分子量构造伪氨基酸成分信息，结合 20 种氨基酸成分信息用于亚细胞定位预测。天津大学张春霆院士[40-42]提出了利用氨基酸的疏水性指标构造蛋白质序列的拟序列阶信息，结合 20 种氨基酸成分信息用于亚细胞定位预测。目前这类特征对蛋白质的刻画能力有限，经常需要结合其它特征使用，这可能是蛋白质分子结构非常复杂导致的。

除了上面提到信息之外，Nair 等人[43]利用文本挖掘技术，从文献中挖掘蛋白质定位的相关信息。Drawid 等人[44]利用基因表达数据。Marcotte 等人[45]采用了蛋白质的系统发育信息。

由于不同的特征从不同的角度刻画蛋白质序列，目前还没有一种特征能够很好地刻画蛋白质的亚细胞定位特征，单独利用某种特征难以在预测效果上取得大的突破。近年来很多学者更加倾向于融合多种特征来刻画蛋白质序列，结果表明这种做法是有效的。

4 分类器的设计

4.1 识别算法

识别算法是蛋白质亚细胞定位预测中另一个关键因素。一个性能优良的预测系统，不但要求使用的特征能够充分反映识别问题的本质，同时需要高效、稳健的识别算法。在蛋白质亚细胞定位预测方面，主要算法包括如下 5 类：

(1)基于简单选择判别规则的方法

Nakai 和 Kanehisa[27]在预测革兰氏阴性菌蛋白质亚细胞定位时，基于实验观察结果得出了一些“if-then”形式的决策判别规则。这虽然是一种最简单的分类算法，但是它却是机器学习方法在蛋白质亚细胞定位研究的雏形。

(2)基于距离度量的近邻方法

这种方法根据某种距离度量方法来度量样本之间的相似程度，两个样本之间的距离越近，则说明它们更加可能在相同的细胞器中出现。Nakashima 和 Nishikawa[25] 采用了基于欧式距离的最近邻方法；Cedano 等人[26] 采用的是基于 Mahalanobis 距离的最近邻方法；Horton 等人[3-5] 则将最近邻方法做了推广，将 k-近邻方法用于他们的预测方法中；Huang 等人[29]在他们的研究中进一步推广了 k-近邻方法，采用了 k 模糊-近邻方法。

(3)基于人工神经网络的方法

人工神经网络在模式识别的很多领域中已经得到了成功的运用，它良好的容错性和鲁棒性广为人们青睐。Reinhardt 和 Hubbard[10] 构造了第一个用于预测蛋白质亚细胞定位的 BP 神经网络，Emanuelsson[6,7] 等人也采用了人工神经网络方法，清华大学孙之荣教授的研究小组[46]将概率神经网络用于预测蛋白质的亚细胞定位。

(4)基于马尔可夫模型的方法

Markov 模型是一种概率论模型，一般用来描述随机变量状态之间的转移概率。它在生物序列分析和基因识别方面具有广泛的应用，例如用于寻找新基因和识别开放阅读框。在蛋白质亚细胞定位预测方面，Yuan [28]构造了基于 Markov 链模型的预测方法，Bendtsen 等人[9]进一步将隐马尔可夫模型用于他们的预测系统。

(5)基于支持向量机(SVM)的方法

上世纪 90 年代中期以来，基于统计学习理论的支持向量机(Support Vector Machine－SVM)[47] 的出现，使机器学习研究进入一个崭新的发展阶段。支持向量机的目标是寻找样本空间的一个最优分类面，使得不同类别样本之间的间隔(margin)最大化，从而达到最佳的推广能力。它的优势还在于可以通过构造合适的核函数有效地解决非线性分类问题。自从 Hua 和 Sun[11] 首次将 SVM 用于蛋白质亚细胞定位预测以来，越来越多的学者[15,30,33]倾向于选择 SVM 作为蛋白质亚细胞定位的分类器。

除了上面提到的分类算法之外，Scott[32] 和 Drawid[44] 提出了基于贝叶斯网络的预测方法，Chou[48]将协方差判别函数用于蛋白质亚细胞定位预测。

值得注意的是，在上述分类方法中，支持向量机由于具有良好的推广能力而受到人们的广泛重视，现在很多学者将支持向量机作为这一研究领域的首选分类器。同时，由于数据集的复杂程度不断增加，单一分类器很难再取得新的进展，人们已经逐渐将精力集中到如何构建更加高效的集成分类器上来。

4.2 分类器性能评价方法

4.2.1 测试方法

一个分类器性能的评价结果与训练和测试过程中样本采集的方法有密切关系，不同的测试方法得到的评价结果的可信程度大不相同。在蛋白质亚细胞定位研究中，最经常用的测试方法是5重交叉验证法，留一法也有一定的应用[11,15,28]。

4.2.2 评价指标

评价分类器的性能是分类器设计的一个重要方面，在评价蛋白质亚细胞定位分类器的性能时经常使用如下4个指标：

(1)总体准确率

总体准确率是被正确识别样本占总体的比例：

$$Totalaccuracy=\frac{1}{N}\sum_{i=1}^{k}tp(i) \tag{1}$$

其中 $tp(i)$ 是第 i 类样本被正确识别的数量(称真阳数)，N 是测试样本的数量，k 是样本的类别数。

(2)敏感性指标

敏感性指标是指每类样本中被正确识别的比例：

$$Sensitivity(i)=\frac{tp(i)}{tp(i)+fn(i)} \tag{2}$$

其中 $fn(i)$ 是指第 i 类样本中被错误识别的数量(称假阴数)。

(3)特异性指标

特异性指标是指被判别为第 i 类的样本中真正属于第 i 类的比例：

$$Specificity(i)=\frac{tp(i)}{tp(i)+fp(i)} \tag{3}$$

其中 $fp(i)$ 是指被错误判别为第 i 类的样本数量(称假阳数)。

(4)Matthews 相关系数 MCC

$$MCC(i)=\frac{tp(i)\times tn(i)-fp(i)\times fn(i)}{\sqrt{(tp(i)+fn(i))\times(tp(i)+fp(i))\times(tn(i)+fp(i))\times(tn(i)+fn(i))}} \tag{4}$$

其中 $tn(i)$ 是非第 i 类样本被正确判别的数量(称真阴数)。MCC 指标取值0至1，取值越高说明分类器的性能越好，当 MCC 取1时，所有样本均被正确识别；当 MCC 取0时，分类器的判别效果与随机指派的结果一样，这样的分类器是最差的。

结束语 经过了十几年的努力，机器学习方法在蛋白质亚细胞定位预测方面取得了显著的进步。这主要体现在如下4个方面：(1)数据的复杂程度不断增加。早期的数据集只包含细胞内外两种不同蛋白质，接着包含3种、4种蛋白质，最近的预测方法使用的数据集包含22种亚细胞蛋白质；在序列的同源性方面，早期的数据集中序列的一致性通常小于90%，最近的多数数据集中序列的一致性小于25%。(2)用来刻画蛋白质序列的信息越来越丰富。早期的预测方法通常只使用蛋白质序列的某一种特征，随着数据复杂程度的增加，单一特征已经无法有效地将不同类别的蛋白质区分开来，因此基于多种特征组合的方法逐渐成为改进识别效果的重要出路。(3)识别算法越来越复杂。从早期基于简单分支算法的预测方法，到神经网络和支持向量机的使用，最近很多分类器采用集成算法。这些健壮分类器的构建是克服数据复杂度增加带来困难的有效方法。(4)预测的精度不断改善。早期的预测精度多数在70%～85%之间，现在的很多预测方法能够达到90%左右的预测精度。总的来说，机器学习方法在很大程度上弥补了实验研究方法的不足，在成本和效率方面比实验方法具有明显的优势，在某些情况下取得的预测精度已经超过了传统的实验方法。

尽管机器学习方法在蛋白质亚细胞定位预测方面取得了一定的成绩，但是它在这一领域的应用还是面临着一些困难：(1)预测结果的可理解性不够，机器学习的原理与真实的生物学机制没有很好地联系起来，有的预测方法缺乏可靠的生物学知识的支持，难以得到生物学家的认可。(2)蛋白质定位与特定的细胞环境密切相关，目前还没有一种机器学习方法将细胞环境的影响因素考虑进来。(3)有的蛋白质在不同的生命过程会出现在不同的细胞器中，这样的蛋白质对生物体来说非常重要，它们可能具有一些特殊的生物功能，可是很多预测方法只考虑一个蛋白质属于一个细胞器的情况，没有考虑属于多个细胞器的情况。(4)细胞内有的蛋白质峰度较高，有的峰度较低，导致样本集中的样本严重不均衡，给分类器带来一定的不良影响，目前的研究很少考虑如何解决这种问题。

今后机器学习方法在蛋白质亚细胞定位预测方面的研究应该重点解决如下3个方面的问题：(1)建立更加符合真实生物现象的数据集，使得数据集更加符合蛋白质在细胞中不同细胞器上的分布情况，包括蛋白质的分类更加细化、将属于多种亚细胞的蛋白质样本包括进数据集中、每个蛋白质序列的注释更加明确。(2)紧密联系蛋白质的合成、分选以及蛋白质相互作用的机制，进一步研究更加有效刻画蛋白质序列的特征。(3)从机器学习的角度着重研究多类标、非平衡样本的模式识别方法，提高机器学习理论在蛋白质亚细胞定位预测方面解决实际问题的能力。

随着生命科学研究的发展，人类对蛋白质亚细胞定位的生物学机理将了解得更加全面和深入，对蛋白质序列的刻画将更加准确。同时，随着机器学习理论的发展，新的机器学习方法和技术也将被用于蛋白质亚细胞定位预测中，蛋白质定位预测将取得一些新的突破。

参 考 文 献

[1] Lee K Y, Kim D W, et al. PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. Nucleic Acids Res., 2006, 34(17): 4655-4666

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

[3] Horton P, Nakai K. Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier. Intelligent Systems for Molecular Biology, 1997, 5: 147-152

[4] 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

[5] Horton P, et al. WoLF PSORT: protein localization predictor.

Nucleic Acids Res. ,2007,35:W585-W587
[6] Emanuelsson O, Nielsen H, von Heijne G. ChloroP: a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci, 1999, 8 (5): 978-984
[7] 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
[8] Emanuelsson O, Brunak S, von Heijne G, et al. Locating proteins in the cell using TargetP, SignalP and related tools. Nature Protocols, 2007, 2: 953-971
[9] Bendtsen J D, Nielsen H, von Heijne G, et al. Improved prediction of signal peptides: SignalP 3. 0. J Mol Biol, 2004, 340(4): 783-795
[10] Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. , 1998, 26 (9): 2230-2236
[11] Hua S J, Sun Z R. Support Vector Machine Approach for Protein Subcellular Location Prediction. Bioinformatics, 2001, 17: 721-728
[12] Shen H B, Chou K C. Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem. Biophys. Res. Commun, 2007, 355: 1006-1011
[13] Chou K C, Shen H B. Large-Scale Plant Protein Subcellular Location Prediction. Journal of Cellular Biochemistry, 2007, 100: 665-678
[14] Shen H B, Chou K C. Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virusinfected cells. Biopolymers, 2007, 85: 233-240
[15] Xie D, Li A, Wang M, et al. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res. , 2005, 33: W105-W110
[16] UniProtKBSwissProt. 2007. http://www. ebi. ac. uk/swissprot/
[17] Chou K C, Shen H B. Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. Biochemical and Biophysical Research Communications, 2006, 347: 150-157
[18] Wiwatwattana N, Kumar A. Organelle DB: a cross-species database of protein localization and function. Nucleic Acids Res. , 2005, 33: D598-D604
[19] Fink J L, Aturaliya R N, Davis M J, et al. LOCATE: a mouse protein subcellular localization database. Nucleic Acids Res. , 2006, 34: D213-D217
[20] Emanuelsson O. Predicting protein subcellular localisation from amino acid sequence information. Briefings in Bioinformatics, 2002, 3(4): 361-376
[21] Nielsen H, et al. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int. J. Neural Syst. , 1997, 8: 581-599
[22] Claros M G. MitoProt: a Macintosh application for studying mitochondrial proteins. Comput. Appl. Biosci. , 1995, 11: 441-447
[23] Li Y F, Liu J. Predicting subcellular Localization of Proteins Using Support Vector Machine with N-Terminal Amino Acid Composition. ADMD, 2005: 618-625
[24] Matsuda S, Vert J P, Saigo H, et al. A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Sci. , 2005, 14: 2804-2813
[25] 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
[26] 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
[27] Nakai K, Kanehisa M. Expert system for predicting protein localization sites in gram-negative bacteria. Proteins, 1991, 11(2): 95-110
[28] Yuan Z. Prediction of protein subcellular locations using Markov chain models. FEBS Lett, 1999, 451 (1): 23-26
[29] Huang Y, Li Y D. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics, 2004, 20 (1): 21-28
[30] Yu C S, Lin C J, Hwang J K. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. , 2004, 13 (5): 1402-1406
[31] Chou K C, Cai Y D. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem, 2002, 277 (48): 45765-45769
[32] Scott M S, Thomas D Y, Hallett M T. Predicting subcellular localization via protein motif co-occurrence. Genome Res. , 2004, 14(10A): 1957-1966
[33] Bhasin M, Raghava G P. ESLpred: SVM-based method for subcellular localization of eukaryotic protein using dipeptide composition and PSI-BLAST. Nucleic Acids Res. , 2004, 32: W414-W419
[34] Nair R, Rost B. Better Prediction of Sub-Cellular Localization by combining Evolutionary and Structural Information. Proteins: Struct. Fuct. Gen. , 2003, 53: 917-930
[35] Guo J, Lin Y L. TSSub: eukaryotic protein subcellular localization by extracting features from profiles. Bioinformatics, 2006, 22(14): 1784-1785
[36] 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: 743-747
[37] 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
[38] Chou K C, Shen H B. Euk-mPLoc: A Fusion Classifier for Large-scale Eukaryotic Protein Subcellular Location Prediction by Incorporating Multiple Sites. Journal of Proteome Research, 2007, 6: 1728-1734
[39] Chou K C, Elrod D W. Using Discriminant Function for Prediction of Subcellular Location of Prokaryotic Proteins. Bliochemical and Biophysical Research Communications, 1998, 252: 63-68
[40] Feng Z P, Zhang C T. Prediction of the membrane protein types based on the hydrophobic indices. J. Protein Chem. 2000, 19: 269-275
[41] Feng Z P, Zhang C T. Prediction of the subcellular location of prokaryotic proteins based on the hydrophobic index of the amino acids. Int. J. Biol. Macromol, 2001, 14: 255-261
[42] Feng Z P. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers, 2001, 58: 491-499

(下转第 49 页)

号，对子载波和功率进行联合优化，每分配一个载波给用户时，立刻优化用户的功率分配，有效提高了系统性能。

2.3 信道状态和队列状态信息的重要性比较

文献[22]的研究显示，信道状态信息和队列信息对调度决策的重要性随着业务负荷的变化而变化：当到达率小时，队列状态信息对调度策略和系统性能影响较大；而当到达率较大时，信道状态信息在调度策略中起支配作用，对系统性能影响较大。另外，队列状态还得到了其他应用，例如队列意识的上行链路带宽分配[23]、队列意识的速率控制[23]等。

结束语 随着无线多媒体业务的广泛应用，日益增长的业务需求与有限带宽资源之间的矛盾日渐突出。资源调度对保证业务的 QoS 需求，提高无线频谱效率发挥着重要作用。本文对无线跨层资源调度技术研究的近期发展情况进行了概述。值得进一步研究的问题：

(1)联合流量预测、信道状态信息预测和队列状态信息的跨层队列调度算法。信道状态和业务流量的随机动态变化以及传播时延使调度器所获得的系统状态信息往往是不准确的，影响了调度的准确性和系统性能。利用流量预测和信道状态预测机制，减少状态信息的不准确性，提高调度的准确性，增强系统性能。

(2)基于模糊逻辑的跨层队列调度算法。由于业务流量、队列状态和信道状态等信息的不准确性，很难做出正确、准确的调度决策。模糊逻辑利用模糊集和模糊推理的方法，能够根据不完整、不准确的输入信息做出有效决策。研究基于模糊逻辑的跨层队列调度算法，减少状态信息不准确所带来的影响。

参 考 文 献

[1] Bhagwat P, Krishna A, Tripathi S. Enhancing throughput over wirelessLAN using channel state dependent packet scheduling// Proc. IN FOCOM96. Mar. 1996:1133-1140

[2] Ng T S E, Stoica I, Zhang H. Packet fair queuing algorithms for wireless networks with location-dependent errors// Proc. INFOCOM98. Mar. 1998:1103-1111

[3] Lu Songwu, Bharghavan V, Srikant R. Fair Scheduling in Wireless Packet Networks . Networking IEEE/ACM Transactions, 1999,7(4)

[4] Liu X. Opportunistic scheduling in wireless communication networks. Ph. D. dissertation. Purdue University, 2002

[5] Liu X, Chong E K P, Shroff N B. A framework for opportunistic scheduling in wireless networks. Computer Networks, 2003, 41(4):451-474

[6] Rhee J H, Kim T H, Kim D K. A Wireless Fair Scheduling Algorithm for 1x HRPD System. IEEE, 2001:743-746

[7] Liu X, Chong E K P, Shroff N B. Opportunistic transmission scheduling with resource sharing constraints in wireless networks. Selected Areas in Communications, IEEE Journal, 2001, 19:2053-2064

[8] Zhang Zhi, He Ying, Chong E KP. Opportunistic downlink scheduling for multiuser ofdm systems // WCNC2005. March 2005:1205-1212

[9] Liu Yonghe, Knightly E. Opportunistic fair scheduling over multiple wireless channels// INFOCOM 2003. 2003,2:1106- 1115

[10] Lee Jang-Won, Mazumdar R R, Shroff N B. Opportunistic power scheduling for dynamic multi-server wireless systems. Wireless Communications, IEEE Transactions on, 2006, 5(6):1506-1515

[11] Kulkarni S, Rosenberg C. Opportunistic Scheduling: Generalizations to Include Multiple Constraints, Multiple Interfaces, and Short Term Fairness. Springer Wireless Networks, 2005, 11(5): 557-569

[12] Liao Dan, Li Lemin, Xu Shizhong, et al. Opportunistic Scheduling with Multiple QoS Constraints in Wireless Multiservice Networks// WCNC2007. March 2007:1525-1531

[13] Andrews M, Kumaran K, Ramanan K, et al. Providing quality of service over a shared wireless link. IEEE Communication Magazine, 2001

[14] Shakkottai S, Stolyar A. Scheduling algorithms for a mixture of real-time and non-real-time data in HDR// Proc. of International Teletraffic Congress (ITC). 2001

[15] 赵新胜，鞠涛，尤肖虎. 一种适用于 B3G 移动通信系统下行共享信道的调度算法[J]. 电子学报，2005，33(7)：1173-1176

[16] Mehrjoo M, Shen Xuemin, Naik K. A joint channel and queue-aware scheduling for IEEE 802. 16 wireless metropolitan area networks// WCNC2007. Mar. 2007

[17] Huang Jinri, Niu Zhisheng. Buffer-aware and Traffic-dependent Packet Scheduling in Wireless OFDM Networks// WCNC2007. Mar. 2007

[18] Song Guocong. Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels // WCNC2004. Mar. 2004

[19] 陈巍，曹志刚，樊平毅，等. 基于信道和队列状态信息的跨层最优功率分配[J]. 通信学报，2007，28(8)

[20] Yang Xaing, Yum Tak - Shing P. Minimal waiting time assignment of subcarriers and power for ofdma system// WCNC2007. Mar. 2007

[21] Joint subcarrier and power allocation in channel-aware queue-aware scheduling for multiuser ofdm. IEEE Trans. on Wireless Communication, 2007, 6(9)

[22] Somsak. Resource allocation in ofdma with time-varying channel and bursty arrivals. IEEE Communications Letters, 2007, 11(9)

[23] Niyato D, Hossain E. Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE802. 16 broadband wireless networks. IEEE Trans. on mobile Computing, 2006, 5(6):668-679

(上接第 33 页)

[43] Nair R, Rost B. Inferring sub-cellular localization through automated lexical analysis. Bioinformatics, 2002, 18(Suppl): S78-S86

[44] Drawid A, Gerstein M. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. J. Mol. Biol., 2000, 301:1059-1075

[45] 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

[46] Guo J, Lin Y L, Sun Z R. A Novel Method for Protein Subcellular Localization Based on Boosting and Probabilistic Neural Network// Asia-Pacific Bioinformatics Conference (APBC2004). 2004

[47] Vapnik V. The Nature of Statistical Learning Theory. New York: Spring-Verlag, 1995

[48] Chou K C, Elrod D W. Using Discriminant Function for Prediction of Subcellular Location of Prokaryotic Proteins. Biochemical and Biophysical Research Communications, 1998, 252:63-68