蛋白质亚细胞定位预测研究综述*

乔善平[1a,2]，闫宝强[1b]

(1. 山东师范大学 a. 管理科学与工程学院；b. 数学科学学院，济南 250014；2. 济南大学 信息科学与工程学院，济南 250022)

摘　要： 蛋白质亚细胞定位预测对于确定蛋白质功能、揭示分子交互机理、理解复杂生理过程和设计药物靶标等方面都有很大的促进作用。随着后基因组时代中蛋白质序列数据的指数增长，研究基于机器学习的计算性蛋白质亚细胞定位预测方法变得越来越重要。为了能够把握该问题的研究状况，从数据集构建、蛋白质特征提取与表示、预测算法设计、算法测试和 Web 服务的建立等五个方面对蛋白质亚细胞定位预测的研究进行了综述。指出了目前该研究领域需要解决的核心问题及难点问题，分析了当前研究中出现的一些新情况，并对将来的研究方向和研究重点进行了展望。

关键词： 蛋白质亚细胞定位预测；特征表示；算法设计；算法测试；Web 服务器

中图分类号： TP301.6　　**文献标志码：** A　　**文章编号：** 1001-3695(2014)02-0321-07

doi:10.3969/j.issn.1001-3695.2014.02.001

Review of protein subcellular localization prediction

QIAO Shan-ping[1a,2], YAN Bao-qiang[1b]

(1. *a. School of Management Science & Engineering, b. School of Mathematical Science, Shandong Normal University, Jinan* 250014, *China*; 2. *School of Information Science & Engineering, University of Jinan, Jinan* 250022, *China*)

Abstract: Protein subcellular location prediction can promote scientists to determinate protein functions, to reveal how and in what kind of cellular environments proteins interact with each other and with other molecules, to understand the intricate pathways that regulate biological processes at the cellular level, and to discover the drug targets. With the exponential increase of protein sequences in the post genomic age, the machine learning based protein subcellular location predicting methods are becoming more and more important. In order to grasp the research focuses on this field, this paper reviewed the research status of this problem from the following five aspects, dataset construction, protein features representation, predicting algorithm design, algorithm testing and Web server building. It pointed out the core and difficult parts in solving this problem, analyzed the problems that had recently arisen, and prospected the research focuses in this field.

Key words: protein subcellular localization prediction; feature representation; algorithm design; algorithm test; Web server

0　引言

细胞是生命活动的基本单位，它由执行不同机体功能的称为亚细胞的各部分组成，如细胞膜、细胞核、线粒体、高尔基体、内质网等。亚细胞功能是由位于其中的蛋白质执行的，蛋白质所在的亚细胞称为蛋白质的亚细胞位置[1]。蛋白质必须转运到其应在的亚细胞位置上才能正确行使其功能，否则就会出现机体功能紊乱，进而产生各种疾病等现象。所以蛋白质的亚细胞位置是蛋白质最重要的属性之一，它有助于确定蛋白质功能、揭示分子交互机理、理解复杂生理过程和开发药物靶标等方面的研究[2]。

确定一条蛋白质的亚细胞位置称为蛋白质亚细胞定位[3]。蛋白质亚细胞定位的传统方法是通过生物化学实验，如 X 射线晶体衍射、电子显微镜、核磁共振等方法进行测定[4]。实验方法精确度高，但费时耗力、代价昂贵，而且对难于结晶的蛋白质来说，实验方法不再有效。随着人类基因组计划的完成，现代生物学的研究进入了后基因组时代，大规模高通量测序技术的发展使得蛋白质序列数量呈指数增长。面对这些海量数据，仅仅依靠人工实验测定的方法已不能满足现实需要，必须借助于先进高效的计算机自动化数据处理技术才能得以实现。因此，蛋白质亚细胞定位预测问题的研究就成为后基因组时代的一项重要研究课题[5]。

细胞生物学研究表明，蛋白质序列、结构和功能之间关系密切，而蛋白质的相关信息都隐藏在其序列当中。因此，为了帮助有关人员更好地开展研究工作，有必要从蛋白质序列出发，探索使用数学知识和机器学习理论等设计有关算法对新发现蛋白质的亚细胞位置进行预测，从而为进一步研究其结构和功能等其他方面提供线索。20 世纪 90 年代初期，Nakai 等人[6,7]利用实验结果建立了由“if-then”规则组成的知识库并开发了专家系统，利用蛋白质序列信息对革兰氏阴性细菌蛋白质的 4 种亚细胞位置、动物细胞蛋白质的 14 种亚细胞位置和植物细胞蛋白质的 17 种亚细胞位置进行了预测，由此揭开了基于计算性方法的蛋白质亚细胞定位预测研究工作。在过去的二十多年中，出现了大量的蛋白质亚细胞定位预测方法并取得

收稿日期： 2013-07-01；**修回日期：** 2013-08-21　　**基金项目：** 国家自然科学基金资助项目(61070130)；济南大学博士基金资助项目(XBS1318)

作者简介： 乔善平(1971-)，男，山东梁山人，副教授，博士研究生，主要研究方向为智能计算、生物信息学(ise_qiaosp@ujn.edu.cn)；闫宝强(1966-)，男，教授，博导，博士，主要研究方向为微分方程与动力系统.

了较大的进展[8~15]。本文重点关注基于计算性方法的蛋白质亚细胞定位预测研究的进展情况。

蛋白质亚细胞定位本质上属于生物数据中的模式分类问题。在解决该问题的过程中,主要涉及以下五个方面的内容[15]:蛋白质样本数据集的构建;蛋白质样本的特征提取与表示;预测算法的设计;算法的测试度量;Web服务器的建立。

1 数据集的构建

数据集是预测算法学习和测试的数据基础,它直接影响到算法学习的效果和测试结果,因此数据集的构建是一个基本且重要的问题。由于蛋白质序列数据有着多模态、多标记、动态、海量、不等长、高维和高冗余等特点,数据集的构建是一项非常精细且繁重的工作。在构建数据集时应考虑的主要因素有蛋白质序列的条数、蛋白质亚细胞位点的数目、是否涉及多位点蛋白质、是按照基因组还是特定物种构建、是否在蛋白质亚细胞水平上构建、数据集中蛋白质序列间同源性大小的控制等。这些因素对蛋白质样本的特征提取和表示、预测算法的设计都有较大的影响。数据集中的蛋白质序列主要来源于Swiss-Prot[16]和其他一些专门数据库如NPD(nuclear protein database)[17]和PPDB(plant proteomics database)[18]等。通过使用搜索工具[19]和过滤程序[20,21]完成数据集的构建工作。

目前主要有三种数据集构建方法:a)直接使用有关文献中已构建好的数据集,这种方法可以节省时间,也便于进行性能比较,但往往不能反映最新数据;b)在有关文献中已构建好的数据集基础上进行一定的筛选和添加,以达到某些特定的要求,这种方法也能够节省时间且能反映一定的最新数据,但与已有的方法不便于进行客观比较;c)构建全新的数据集,这种方法能够反映最新的数据情况,属于开创性工作,好的数据集能够带动和促进相关研究工作的进一步深入。比较典型的数据集构建情况可参见有关文献[22~24]。

2 特征提取与样本表示

氨基酸是构成蛋白质的基本单位,一条蛋白质从形式上可表示如下:

$$P = R_1 R_2 \cdots R_n \tag{1}$$

其中:P表示一条蛋白质序列,$R_i(i=1,2,\cdots,n)$表示蛋白质序列P中的各个氨基酸。

描述蛋白质的特征有很多,如理化特征、序列特征、进化特征等,它们对预测算法的设计和预测结果都有直接关系。过多的特征会造成维数灾难[25],降低算法的效率;特征太少又可能会丢失一些重要信息,影响预测结果。所以,如何提取有效特征进行融合并建立一种合适的表示方法和相似性度量标准是一个需要深入研究的核心问题。目前使用的方法主要分为基于序列信息和基于注释信息两大类,另外也出现了其他的一些表示方法以及多特征的融合方法。

2.1 基于序列信息

基于蛋白质氨基酸序列信息的表示方法可以分为基于序列同源性、基于类选信号和基于氨基酸组成三种。

1)基于序列同源性

根据进化原理,这种方法[26]主要使用一些相似性搜索工具[19]在蛋白质序列数据库中查找与待测蛋白质序列同源性较高的蛋白质序列,根据搜索结果进行判定。该方法具有明确的生物学基础,但受到数据库中蛋白质序列的限制,对某些待测蛋白质并不能找到同源性较高的蛋白质序列,此时这种方法就不再有效。

2)基于类选信号

根据细胞生物学相关知识,蛋白质N端或C端区域中大约有15~70个不等的氨基酸构成靶信号,它们可以引导蛋白质完成亚细胞定位任务。所以利用类选信号进行蛋白质亚细胞定位预测[27~31]也具有较明确的生物学基础。

3)基于氨基酸组成

蛋白质属性与氨基酸之间有着密切联系,所以利用蛋白质序列中各氨基酸的组成情况表示蛋白质样本也同样具有生物学基础,这种方法也是目前使用最为广泛的方法之一。根据氨基酸不同的组成情况,主要有以下几种表示方法。

(1)AAC 氨基酸组成(amino acid composition,AAC)是1986年由Nakashima等人[32]在研究蛋白质折叠类问题时最先提出的。AAC的向量表示形式为

$$X = [f_1, f_2, \cdots, f_{20}]^{\mathrm{T}} \tag{2}$$

其中:$f_i(i=1,2,\cdots,20)$代表在蛋白质X中20种原生氨基酸出现的归一化频率。1994年,Nakashima等人[33]将AAC用于蛋白质亚细胞定位预测并取得了一定的效果;1995年,Chou[34]对AAC的表示形式进行了简化,由20维减少到19维,并证明了两者是等价的;1997年,Cedano等人[35]进一步验证了AAC与蛋白质亚细胞定位存在一定的联系。在接下来的几年中,AAC得到了较为广泛的应用[36]。但这种方法的最大缺点就是完全丢失了序列中氨基酸的顺序信息,造成预测精度不会很高。2000年,Chou[37]开始考虑加入序列顺序因素,预测效果有了较大的提升,并阐明了顺序信息所起的作用。

(2)PseAAC 2001年,Chou[38]提出了伪氨基酸组成(pseudo amino acid composition,PseAAC)模型,PseAAC的基本表示形式为

$$X = [x_1 \quad \cdots \quad x_{20} \quad x_{20+1} \quad \cdots \quad x_{20+\lambda}]^{\mathrm{T}} \tag{3}$$

PseAAC模型提出以后,得到了非常广泛的应用,并在PseAAC的基础上通过加入反映氨基酸特性的各种信息,出现了PseAAC的很多变形[39]。鉴于这种形势,Shen等人[40]于2008年建立了称为PseAAC的服务器,用于产生蛋白质的几种伪氨基酸组成。最近Du和Cao等人又开发了分别称为PseAAC-Builder[41]和Propy[42]的程序作为PseAAC的补充,用于计算各种形式的PseAAC。

(3)k-peptide 出现最早的是二肽组成[43],即两个氨基酸构成一个对,对于20种原生氨基酸来说,这种形式是一个400维的向量。后来又出现了不同形式的多态组成[44],可以在一定程度上弥补顺序信息缺失的不足。最近有文献[45]使用四肽信息进行了蛋白质亚细胞定位的预测研究。

2.2 基于注释信息

随着蛋白质注释工作的发展和不断深入,出现了很多的蛋白质注释数据库,并促进了基于注释信息的蛋白质亚细胞定位预测的研究,目前主要有以下几种方法。

1)FunD

2002年,Chou等人[46]开始使用功能域(functional domain,FunD)注释信息进行蛋白质亚细胞定位预测的研究。该方法首先根据功能域数据库[47]中已注释的功能域数目确定特征向量的维数,然后使用搜索工具[19]进行功能域搜索,在设定的重要性阈值下根据每个功能域的命中情况构建特征向量。

2)GO

基因本体(gene ontology,GO)[48]方法是基于GO数据库的一种特征提取方法。首先根据所用数据库包含的GO条目确

定特征向量的维数,然后根据设定的阈值使用搜索工具[19]在 Swiss-Prot 数据库中搜索待测蛋白质的同源序列构成待测蛋白质的同源序列集合,接下来对同源序列集合中的每条蛋白质序列依据它们的访问序列号在 GO 数据库中进行 GO 条目搜索,最后根据每个 GO 条目的命中情况构建特征向量。

3)其他方法

依据物种进化理论和灰色理论,可以使用灰色位置特异性矩阵(position-specific scoring matrix,PSSM)[49]进行蛋白质样本的表示。李凤敏和赵禹等人[50,51]用离散增量方法对蛋白质的特征进行提取和表示。2008 年,Lee 等人[52]利用蛋白质交互作用(protein-protein interaction,PPI)进行了蛋白质亚细胞定位预测研究,其结果表明,使用 PPI 信息可以显著提高预测结果。近年来,随着网络构建的快速发展,如何利用 PPI 信息进行蛋白质亚细胞定位预测[53~55]已成为一个热点研究课题;也出现了一些使用 motif[56]和文本信息[57]研究蛋白质亚细胞定位预测的文献。

总之,到目前为止已出现的蛋白质样本的特征提取和表示方法已有很多。本质上,该问题重点研究如何构建有效的特征向量,以便更好地识别不同类别的蛋白质,为提高算法效率和提升预测结果奠定基础。目前大多采用多维向量的形式进行表示,这种方法简单且计算方便;采用非线性表示形式,如树和图等也许是值得探索的一种形式。另外,将不同的特征进行融合[58]也是目前的主要研究方法。

3　预测算法设计

由于蛋白质序列数量巨大且隐藏的信息很难揭示,所以对预测算法的性能要求很高。如何设计具有高通量、高准确率和高精度特性的预测算法就成了另一个需要深入研究的核心问题。在过去的 20 多年中出现了大量的预测算法,从单一的算法设计与应用到目前的集成机器学习研究;从单位点预测到现在的多位点多标记学习,预测的准确度不断提高,应用范围逐步扩大。但是,现存的算法还远不足以解决蛋白质亚细胞定位问题,这方面还需要更加深入的研究,需要有更大的突破。融入智能计算的集成学习[59,60]将是未来的研究重点。下面介绍几个有代表性的方法。

3.1　最近邻方法

最近邻算法(nearest neighbor,NN)[61]是一种简单、直观且有效的分类算法。它首先计算出待测样本与各训练样本之间的一种距离度量,然后找出与待测样本距离最近的训练样本,将待测样本归入该训练样本的类别。后来又出现了 K-最近邻(K-nearest neighbor,KNN)[62]、模糊 KNN[63]、OET-KNN[64]和其他形式的 KNN[65]。

KNN 是根据设定的 K 值从训练样本中选择出与测试样本距离最近的 K 个样本,然后统计这 K 个样本中分别属于每一类的样本数,将待测样本归入样本数最多的类别。模糊 KNN 是基于模糊集合理论,赋予待测样本一个相对于某个类的隶属度,隶属度的大小代表了该待测样本属于此类的可能性大小,隶属度对待测样本的判决有重要的参考价值。OET-KNN 是一种基于 Dempster-Shafer 理论的模式分类方法,其主要步骤如下:

a)依据某种度量找出待测样本的 K 个最近邻样本。

b)根据公式 $M(P_i,S_u)=a_0\mathrm{e}^{-r_u^2D^2(P_i,P)}$ $(i=1,2,\cdots,K;u=1,2,\cdots,M)$ 计算每个样本为待测样本所提供的属于每一类的证据值。其中 P 为待测样本,P_i 为 P 的第 i 个近邻,S_u 为第 u 个类别,a_0 一般取 0.95,r_u 是与每个类别相关的参数,$D(P_i,P)$ 表示两个样本之间的距离。在 OET-KNN 中对 r_u 进行优化是该算法的主要贡献。

c)根据公式 $M(P,S_u)=\oplus_{i=1}^{N}M(P_i,S_u)$ 计算 K 个样本证据值的正交和。

d)根据公式 $u=\arg\max\{M(P,S_j)\}$ $(j=1,2,\cdots,M)$ 得到待测样本所属的类别。

3.2　人工神经网络

人工神经网络(artificial neural network,ANN)[66]是一种应用类似于大脑神经突触连接结构进行信息处理的数学模型,是由大量神经元相互连接构成的非线性、自适应信息处理系统。每个节点代表一种特定的输出函数,称为激励函数;每两个节点间的连接代表对于通过该连接信号的加权值,称之为权重。网络输出依据网络连接方式、权重值和激励函数的不同而不同。Reinhardt 等人[67]最早使用 ANN 进行蛋白质亚细胞定位预测;后来又出现了其他一些基于 ANN 的蛋白质亚细胞定位预测方法[27]。在 ANN 的应用早期,BP 神经网络、概率神经网络和 SOM 使用较多,后来 RBF 神经网络应用较多。目前在该问题的研究中尚未出现使用分数阶神经网络的文献,这是值得研究的一个问题。使用神经网络时,首先确定网络的结构(层数、每层的节点数、使用的激励函数和连接方式等)。通常输入层的节点个数由样本的特征向量维数确定,而输出层的节点个数可由数据集中的类别数确定,也可以构造单输出的神经网络,然后利用集成的方法进行预测。利用训练集中的样本来训练神经网络,通过优化某种损失函数来确定各个连接上的权值,从而确定整个网络中的参数。最后使用训练好的神经网络对未见示例进行预测。

3.3　支持向量机

Cortes 和 Vapnik 等人于 1995 年首先提出了支持向量机(support vector machine,SVM)[68]方法,它在解决小样本、非线性及高维模式识别中表现出许多特有的优势,并能够推广应用到函数拟合等其他机器学习问题中。SVM 是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的,SVM 的关键在于核函数,低维空间向量集通常难以划分,解决的方法是将它们映射到高维空间。但这个办法带来的困难就是计算复杂度的增加,而核函数正好巧妙地解决了这个问题。近年来,使用 SVM 研究蛋白质亚细胞定位的文献[44,58,59,69~72]大量涌现。

SVM 是一种二类分类方法,为了能够使用 SVM 解决多类分类问题,通常有三种解决策略:a)一对其余策略,该策略根据类别数(设为 M)分别构造 M 个 SVM,并为每个 SVM 构造相应的训练数据集,每个 SVM 分别用于判别是否属于某一类,然后将 M 个 SVM 集成起来完成对未见示例的预测,但可能会出现分类重叠或不可分类现象;b)一对一策略,该策略根据类别数(设为 M)分别构造 $M(M-1)/2$ 个 SVM,每个 SVM 分别用于判别是属于两类中的哪一类,最后统计每个类别的得票数,以票数最多的类别作为预测类别;c)有向无环图策略,也称为 DAG-SVM,该策略训练时与一对一策略相同,但在判别时需要构造一个有向无环图,从根节点出发,根据每个节点的输出决定下一步应该到达的节点,而最底部节点的输出即为判别结果,但这种方法可能会出现错误向下累积的现象。

3.4　集成机器学习

1990 年,Schapire 证明了“如果一个概念是弱可学习的,充要条件是它是强可学习的”这样一个关键定理,由此奠定了集成机器学习的理论基础。这个定理说明多个弱分类器可以集成为一个强分类器。集成学习是一种机器学习范式,它使用多

个学习器来解决同一个问题,可以有效提高学习系统的泛化能力,因此成为国际机器学习界的研究热点。现在集成学习已经成功应用于Web信息过滤、生物特征识别[73~75]、计算机辅助医疗诊断等众多领域,然而集成学习技术还不成熟,集成学习的研究还存在着大量未解决的问题,从集成学习的实际应用情况来看,也远未达到人们所期待的水平。具有多个位点的蛋白质数量不断增加为预测增加了难度。如何更好地预测多位点蛋白质的亚细胞位置,出现了多标记学习,该方法目前已有许多应用[48,70,76~84]。

集成学习主要涉及三方面的内容:

a)个体分类器的设计。个体分类器是集成学习的基础,它们直接影响着集成学习的学习效果。1995年,Krogh和Vedelsby证明了集成的泛化误差是由组成集成的个体学习器的平均泛化误差和平均差异度之差所决定的。为了得到好的集成,就需要在尽可能提高个体学习器学习精度的同时,尽可能增大个体学习器之间的差异。目前有基于训练集处理、基于特征集处理和基于算法等构造方式。

b)个体分类器的集成。研究如何将训练好的各个分类器进行集成,如AdaBoost、随机森林和ECOC等方法。

c)输出结果的合成。主要使用加权投票法、贝叶斯方法、ECOC方法、D-S证据理论等实现对各个体分类器预测结果的合成。

3.5 其他方法

贝叶斯方法[9,61,85]、隐马尔可夫模型(hidden Markov model,HMM)[55,72,86,87]、CDA(covariant discriminant algorithm,CDA)[37,88,89]、高斯过程[79]和小波变换[90]等在蛋白质亚细胞定位预测中都有一些应用。

4 算法测试

为了验证预测算法的效能,需要对算法进行严格的测试和性能评价。对算法的测试主要涉及到测试方法和衡量指标两方面的内容。

4.1 测试方法

在确定了数据集、特征向量和预测算法的基础上,需要训练学习机(预测器)和验证学习机的泛化能力。一般包括以下步骤:

a)确定训练集。用于训练学习机,根据设定的优化目标,通过训练集中的样本进行寻优,从而确定学习机中的各项参数值。

b)确定测试集。用于测试学习机,将测试集中的每个样本都由学习机进行测试,然后统计预测结果并计算各种性能指标。

c)评价学习机。根据计算得到的有关性能指标对预测算法进行评价,并进行适当的参数调整和改进学习机。

d)确定学习机。通过评价获得最优的参数值和结构等要素,从而确定最终的学习机模型。

目前使用的测试验证方法主要有自身一致性测试(self-consistency test)、独立数据集测试(independent dataset test)、二次抽样测试(sub-sampling test)和刀切法(jackknife test)四种[91]。

在自身一致性测试中,训练数据集和测试数据集使用的是同一个数据集,这种方法容易造成类别偏见。

独立数据集测试是将整个数据集分为训练数据集和测试数据集,这两个集合的并集为全集,而交集为空集。训练数据集专门用于算法的训练和学习,然后用测试数据集对算法进行验证测试。这种方法存在两个子集如何进行划分的问题,当样本较多时,会有很多的划分方法,具有较大的主观性,每种方法都直接影响到对算法的验证测试结果。

二次抽样测试一般采用K次交叉验证的方式,将整个数据集划分为互不相交的K个子集,对每个子集都执行以下操作:以当前子集为测试数据集,以剩余子集的并集为训练数据集,依次进行训练和测试,共进行K次。这种方法有着与独立数据集测试同样的主观性弊端,而且子集的划分可能空间更大。但从整个过程来看,由于该方法进行的训练和测试次数较多,比独立数据集测试更能说明问题。

Jackknife测试是目前使用最多、也是最为客观的测试方法。该方法将含有N个样本的数据集划分为N个互不相交的子集,每个子集中仅含有一个样本。训练和测试过程与K次交叉验证相同,不同之处体现在子集的划分规则,这种情况下的子集划分结果只有一种。在这种测试方法中,每个样本轮流做测试,其余样本作训练,具有很强的客观性,是当前普遍采用和认可的验证测试方法。

本质上,这几种方法都是对数据集进行一定的划分来对算法进行交叉验证。自身一致性测试属于0次划分,独立数据集测试属于二次划分,二次抽样测试属于K次划分,而Jackknife测试则是N次划分(N为数据集样本数量)。

4.2 衡量指标

在测试中如何衡量算法的性能和预测效果,需要有相应的指标作为指导。总的说来,可以分为基于样本的评价指标和基于类别的评价指标两大类。基于样本的评价指标首先衡量预测算法在单个测试样本上的预测效果,然后计算其在整个测试集上的均值作为最终结果。基于类别的评价指标首先衡量预测算法在单个类别上的分类效果,然后计算其在所有类别上的均值作为最终结果。

在进行具体的指标计算之前,需要先计算出以下四个基本统计量:

a)真正(true positive,TP),对应于被分类模型正确预测为正类的正样本数。

b)假正(false positive,FP),对应于被分类模型错误预测为正类的负样本数。

c)真负(true negative,TN),对应于被分类模型正确预测为负类的负样本数。

d)假负(false negative,FN),对应于被分类模型错误预测为负类的正样本数。

在此基础上,可以计算下列主要的度量指标:

a)总体正确率ACC

$$\mathrm{ACC}=\frac{TP+TN}{(TP+FP+TN+FN)}$$

b)敏感度SN

$$\mathrm{SN}=\frac{TP}{TP+FN}$$

c)特异性SP

$$\mathrm{SP}=\frac{TN}{TN+FP}$$

d)精度PV

$$\mathrm{PV}=\frac{TP}{TP+FP}$$

e)马氏相关系数MCC

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

f)受试者操作特性曲线ROC

ROC(receiver operating characteristic curve,ROC)又称为感受性曲线,是用来显示分类器真正率和假正率间折中的一种图形化方法。在一个 ROC 曲线中,真正率沿 y 轴绘制,假正率显示在 x 轴上。沿着曲线的每个点对应于一个分类器归纳的模型。实际应用中一般采用 AUC(area under the roc curve),通常 AUC 的值在 0.5 ~ 1.0,较大的 AUC 代表了较好的性能。

5　Web 服务器的建立

一个预测算法经过严格的测试被验证为具有较好的预测结果以后,应该考虑借助于 Internet 的力量建立一个 Web 服务器,使更多的研究人员能够通过网络获得待测蛋白质的预测结果,以便为他们的研究工作提供帮助。目前已经有大量的蛋白质亚细胞定位预测服务器能够提供相应的 Web 服务,表 1 列出了比较有影响的一些预测服务器的基本情况。

表 1　重要的蛋白质亚细胞定位预测服务器一览表

名称	网址	文献
Cell-PLoc 2.0	http://www.csbio.sjtu.edu.cn/bioinf/Cell-PLoc-2/	[23]
TargetP	http://www.cbs.dtu.dk/services/TargetP/	[27]
WoLF PSORT	http://www.wolfpsort.org	[28]
Yloc	http://www.multiloc.org/Yloc	[29]
PSORTb v3.0	http://www.psort.org/psortb	[30]
TetraMito	http://lin.uestc.edu.cn/server/TetraMito	[44]
iLoc-Animal	http://www.jci-bioinfo.cn/iLoc-Animal	[48]
MemLoci	http://mu2py.biocomp.unibo.it/memloci	[69]
mGOASVM	http://bioinfo.eie.polyu.edu.hk/mGoaSvmServer/mGOASVM.html	[70]
EuLoc	http://euloc.mbc.nctu.edu.tw/	[72]
Virus-ECC-mPLoc	http://levis.tongji.edu.cn:8080/bioinfo/Virus-ECC-mPLoc/	[83]

6　结束语

蛋白质亚细胞定位是生物信息学领域中一项重要的研究课题,海量的蛋白质序列数据对传统的生物数据分析方法提出了新的挑战。针对这种形势,研究如何利用计算机技术实现高通量的基于计算的蛋白质亚细胞定位预测方法是非常必要的。在利用基于数据集的机器学习方法研究蛋白质亚细胞定位过程中,数据集的构建是一个基本问题,特征提取、表示和相似性度量以及预测算法的设计是两个非常关键的核心问题。

随着多位点蛋白质数量的增加和机器学习研究的不断深入,研究利用多标记[92,93]集成机器学习并融入智能计算方法进行蛋白质亚细胞定位预测将是未来的发展趋势。同时,随着蛋白质亚细胞[94]水平上数据的积累,在该水平上的定位预测研究也将得到进一步的发展。

诺贝尔奖获得者 Gilbert 在 1991 年曾指出:传统生物学解决问题的方式是实验的;现在,基于全部基因都将知晓,并以电子可操作的方式驻留在数据库中,新的生物学研究模式的出发点应是理论的。一个科学家将从理论推测出发,然后再回到实验中去追踪或验证这些理论假设。

参考文献:

[1] 韩榕. 细胞生物学[M]. 北京:科学出版社,2011:55-106.

[2] 陈铭. 生物信息学[M]. 北京:科学出版社,2012:1-10.

[3] 叶子弘. 生物信息学[M]. 杭州:浙江大学出版社,2011:179-223.

[4] MURPHY R F,BOLAND M V,VELLISTE M. Towards a systematics for protein subcelluar location:quantitative description of protein localization patterns and automated analysis of fluorescence microscope images[C]//Proc of International Conference on Intelligent Systems for Molecular Biology. 2000:251-259.

[5] 张松,黄波,夏学峰,等. 蛋白质亚细胞定位的生物信息学研究[J]. 生物化学与生物物理学进展,2007,34(6):573-579.

[6] NAKAI K,KANEHISA M. Expert system for predicting protein localization sites in gram-negative bacteria[J]. Proteins,1991,11(2):95-110.

[7] NAKAI K,KANEHISA M. A knowledge base for predicting protein localization sites in eukaryotic cells[J]. Genomics,1992,14(4):897-911.

[8] NAKAI K. Protein sorting signals and prediction of subcellular localization[J]. Advances in Protein Chemistry,2000,54:277-344.

[9] FENG Zhi-ping. An overview on predicting the subcellular location of a protein[J]. In Silico Biology,2002,2(3):291-303.

[10] DÖNNES P,HÖGLUND A. Predicting protein subcellular localization:past,present,and future[J]. Genomics Proteomics Bioinformatics,2004,2(4):209-215.

[11] GARDY J L,BRINKMAN F S. Methods for predicting bacterial protein subcellular localization[J]. Nature Review Microbiology,2006,4(10):741-751.

[12] CHOU Kuo-chen,SHEN Hong-bin. Recent progress in protein subcellular location prediction[J]. Analytical Biochemistry,2007,370(1):1-16.

[13] SHEN Hong-bin,YANG Jie,CHOU Kuo-chen. Methodology development for predicting subcellular localization and other attributes of proteins[J]. Expert Review of Proteomics,2007,4(4):453-463.

[14] IMAI K,NAKAI K. Prediction of subcellular locations of proteins:where to proceed?[J]. Proteomics,2010,10(22):3970-3983.

[15] CHOU Kuo-chen. Some remarks on protein attribute prediction and pseudo amino acid composition[J]. Journal of Theoretical Biology,2011,273(1):236-247.

[16] UNIPROT CONSORTIUM. The universal protein resource (UniProt) in 2010[J]. Nucleic Acids Research,2010,38(database issue):D142-D148.

[17] DELLAIRE G,FARRALL R,BICKMORE W A. The nuclear protein database(NPD):sub nuclear localization and functional annotation of the nuclear proteome[J]. Nucleic Acids Research,2003,31(1):328-330.

[18] ALONSO J M,ECKER J R. Moving forward in reverse:genetic technologies to enable genome-wide phenomic screens in Arabidopsis[J]. Nature Review Genetics,2006,7(7):524-536.

[19] 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[J]. Nucleic Acids Research,2001,29(14):2994-3005.

[20] WANG Guo-li,Jr DUNBRACK R L. PISCES:recent improvements to a PDB sequence culling server[J]. Nucleic Acids Research,2005,33(Web server issue):W94-W98.

[21] HUANG Ying,NIU Bei-fang,GAO Ying,*et al*. CD-HIT suite:a Web server for clustering and comparing biological sequences[J]. Bioinformatics,2010,26(5):680-682.

[22] CHOU Kuo-chen,SHEN Hong-bin. Cell-PLoc:a package of Web servers for predicting subcellular localization of proteins in various organisms[J]. Nature Protocols,2008,3(2):153-162.

[23] CHOU Kuo-chen,SHEN Hong-bin. Cell-PLoc 2.0:an improved package of Web-servers for predicting subcellular localization of proteins in various organisms[J]. Scientific Research,2010,2(10):1090-1103.

[24] 曹隽喆,顾宏,贺建军. 一种新的蛋白质亚细胞定位预测训练集构造方法[J]. 大连理工大学学报,2012,52(6):884-889.

[25] WANG Tong, YANG Jie, SHEN Hong-bin, *et al*. Predicting membrane protein types by the LLDA algorithm[J]. Protein and Peptide Letters, 2008, 15(9): 915-921.

[26] MAK Man-wei, GUO Jian, KUNG Sun-yuan. PairProSVM: protein subcellular localization based on local pairwise profile alignment and SVM[J]. IEEE/ACM Trans on Computational Biology and Bioinformatics, 2008, 5(3): 416-422.

[27] EMANUELSSON O, NIELSEN H, BRUNAK S, *et al*. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence[J]. Journal of Molecular Biology, 2000, 300(4): 1005-1016.

[28] HORTON P, PARK K J, OBAYASHI T, *et al*. WoLF PSORT: protein subcellulur localization predictor[J]. Nucleic Acids Research, 2007, 35(Web server issue): W585-W587.

[29] BRIESEMEISTER S, RAHNENFÜHRER J, KOHLBACHER O. Going from where to why-interpretable prediction of protein subcellular localization[J]. Bioinformatics, 2010, 26(9): 1232-1238.

[30] YU N Y, WAGNER J R, LAIRD M R, *et al*. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes[J]. Bioinformatics, 2010, 26(13): 1608-1615.

[31] TARDIF M, ATTEIA A, SPECHT M, *et al*. PredAlgo: a new subcellular localization prediction tool dedicated to green algae[J]. Molecular Biology Evolution, 2012, 29(12): 3625-3639.

[32] NAKASHIMA H, NISHIKAWA K, OOI T. The folding type of a protein is relevant to the amino acid composition[J]. Journal of Biochemistry, 1986, 99(1): 153-162.

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

[34] CHOU Kuo-chen. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space[J]. Proteins, 1995, 21(4): 319-344.

[35] CEDANO J, ALOY P, PEREZ-PONS J A, *et al*. Relation between amino acid composition and cellular location of proteins[J]. Journal of Molecular Biology, 1997, 266(3): 594-600.

[36] CHOU Kuo-chen. Prediction of protein structural classes and subcellular locations[J]. Current Protein & Peptide Science, 2000, 1(2): 171-208.

[37] CHOU Kuo-chen. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect[J]. Biochemical and Biophysical Research Communication, 2000, 278(2): 477-483.

[38] CHOU Kuo-chen. Prediction of protein cellular attributes using pseudo-amino acid composition[J]. Proteins, 2001, 43(3): 246-255.

[39] CHOU Kuo-chen. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology[J]. Current Proteomics, 2009, 6(4): 262-274.

[40] SHEN Hong-bin, CHOU Kuo-chen. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition[J]. Analytical Biochemistry, 2008, 373(2): 386-388.

[41] DU Pu-feng, WANG Xin, XU Chao, *et al*. PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions[J]. Analytical Biochemistry, 2012, 425(2): 117-119.

[42] CAO Dong-sheng, XU Qing-song, LIANG Yi-zeng. Propy: a tool to generate various modes of Chou's PseAAC[J]. Bioinformatics, 2013, 29(7): 960-962

[43] WU C, WHITSON G, MCLARTY J, *et al*. Protein classification artificial neural system[J]. Protein Science, 1992, 1(5): 667-677.

[44] 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[J]. Protein Science, 2004, 13(5): 1402-1406.

[45] LIN Hao, CHEN Wei, YUAN Lu-feng, *et al*. Using over-represented tetrapeptides to predict protein submitochondria locations[J]. Acta Biotheoretica, 2013, 61(2): 259-268.

[46] CHOU Kuo-chen, CAI Yu-dong. Using functional domain composition and support vector machines for prediction of protein subcellular location[J]. Journal of Biological Chemistry, 2002, 277(48): 45765-45769.

[47] MARCHLER-BAUER A, ANDERSON J B, DERBYSHIRE M K, *et al*. CDD: a conserved domain database for interactive domain family analysis[J]. Nucleic Acids Research, 2007, 35(database issue): D237-D240.

[48] CHOU Kuo-chen, CAI Yu-dong. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology[J]. Biochemical and Biophysical Research Communications, 2003, 311(3): 743-747.

[49] LIN Wei-zhong, FANG Jian-an, XIAO Xuan, *et al*. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins[J]. Molecular BioSystems, 2013, 9(4): 634-644.

[50] 李凤敏,李前忠,林昊. 基于离散增量和协变判别函数识别蛋白质亚核定位[J]. 内蒙古大学学报:自然科学版, 2008, 39(1): 56-60.

[51] 赵禹,赵巨东,姚龙. 用离散增量结合支持向量机方法预测蛋白质亚细胞定位[J]. 生物信息学, 2010, 8(3): 237-239, 244.

[52] LEE K, CHUANG Han-yu, BEYER A, *et al*. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species[J]. Nucleic Acids Research, 2008, 36(20): e136.

[53] SHIN C J, WONG S, DAVIS M J, *et al*. Protein-protein interaction as a predictor of subcellular location[J]. BMC Systems Biology, 2009, 3(1): 28.

[54] HU Le-le, FENG Kai-yan, CAI Yu-dong, *et al*. Using protein-protein interaction network information to predict the subcellular locations of proteins in budding yeast[J]. Protein and Peptide Letters, 2012, 19(6): 644-651.

[55] JIANG J Q, WU Mao-ying. Predicting multiplex subcellular localization of proteins using protein-protein interaction network: a comparative study[J]. BMC Bioinformatics, 2012, 13(suppl 10): S20.

[56] HU Yin-xia, LI Tong-hua, SUN Jiang-ming, *et al*. Predicting gram-positive bacterial protein subcellular localization based on localization motifs[J]. Journal of Theoretical Biology, 2012, 308(9): 135-140.

[57] SHATKAY H, HÖGLUND A, BRADY S, *et al*. SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data[J]. Bioinformatics, 2007, 23(11): 1410-1417.

[58] FAN Guo-liang, LI Qian-zhong. Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition[J]. Amino Acids, 2012, 43(2): 545-555.

[59] KANDASWAMY K K, PUGALENTHI G, MÖLLER S, *et al*. Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition[J]. Protein and Peptide Letters, 2010, 17(12): 1473-1479.

[60] GARCIA-LOPEZ S, JARAMILLO-GARZON J A, CASTELLANOS-

DOMINGUEZ G. Improving the prediction of sub-cellular locations of proteins with a particle swarm optimization-based boosting strategy [C]//Proc of International Conference on IEEE Engineering in Medicine and Biology Society. 2012,6313-6316.

[61] COVER T M, HART P E. Nearest neighbour pattern classification [J]. IEEE Trans on Information, Theory, 1967, 13(1): 21-27.

[62] HORTON P, NAKAI K. Better prediction of protein cellular localization sites with the K-nearest neighbors classifier[C]//Proc of International Conference on Intelligent Systems for Molecular Biology. 1997: 147-152.

[63] HUANG Ying, LI Yan-da. Prediction of protein subcellular locations using fuzzy k-NN method[J]. Bioinformatics, 2004, 20(1): 21-28.

[64] CHOU Kuo-chen, SHEN Hong-bin. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers[J]. Journal of Proteome Research, 2006, 5(8): 1888-1897.

[65] CHOU Kuo-chen, WU Zhi-cheng, XIAO Xuan. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites[J]. Molecular BioSystems, 2012, 8(2): 629-641.

[66] ENGELBRECHT A P. 计算智能导论[M]. 谭营, 译. 2 版. 北京: 清华大学出版社, 2010: 3-313.

[67] REINHARDT A, HUBBARD T. Using neural networks for prediction of the subcellular location of proteins[J]. Nucleic Acids Research, 1998, 26(9): 2230-2236.

[68] VAPNIK V N. 统计学习理论的本质[M]. 张学工, 译. 北京: 清华大学出版社, 2000: 96-118.

[69] PIERLEONI A, MARTELLI P L, CASADIO R. MemLoci: predicting subcellular localization of membrane proteins in eukaryotes [J]. Bioinformatics, 2011, 27(9): 1224-1230.

[70] WAN Shi-biao, MAK Man-wei, KUNG Sun-yuan. mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines[J]. BMC Bioinformatics, 2012, 13(11): 290.

[71] HAN Guo-sheng, YU Zu-guo, ANH V, *et al*. An ensemble method for predicting subnuclear localizations from primary protein structures [J]. PLoS One, 2013, 8(2): e57225.

[72] CHANG Tzu-hao, WU Li-ching, LEE T Y, *et al*. EuLoc: a Web-server for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC[J]. Journal of Computer-Aided Molecular Design, 2013, 27(1): 91-103.

[73] NIU Bing, JIN Yu-huan, FENG Kai-yan, *et al*. Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins [J]. Molecular Diversity, 2008, 12(1): 41-45.

[74] LIN Jie, WANG Yan. Using a novel AdaBoost algorithm and Chou's Pseudo amino acid composition for predicting protein subcellular localization[J]. Protein and Peptide Letters, 2011, 18(12): 1219-1225.

[75] SARAVANAN V, LAKSHMI PT. SCLAP: an adaptive boosting method for predicting subchloroplast localization of plant proteins[J]. Omics, 2013, 17(2): 106-115.

[76] CHOU Kuo-chen, WU Zhi-cheng, XIAO Xuan. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins[J]. PLoS One, 2011, 6(3): e18258.

[77] WU Zhi-cheng, XIAO Xuan, CHOU Kuo-chen. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites[J]. Molecular BioSystems, 2011, 7(12): 3287-3297.

[78] XIAO Xuan, WU Zhi-cheng, CHOU Kuo-chen. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites[J]. PLoS One, 6(6): e20592.

[79] HE Jian-jun, GU Hong, LIU Wen-qi. Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites[J]. PLoS One, 2012, 7(6): e37155.

[80] CAO Jun-zhe, LIU Wen-qi, GU Hong. Predicting viral protein subcellular localization with Chou's pseudo amino acid composition and imbalance-weighted multi-label K-nearest neighbor algorithm[J]. Protein and Peptide Letters, 2012, 19(11): 1163-1169.

[81] MEI Su-yu. Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning[J]. Journal of Theoretical Biology, 2012, 310(10): 80-87.

[82] WANG Xiao, LI Guo-zheng. A multi-label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins [J]. PLoS One, 2012, 7(5): e36317.

[83] WANG Xiao, LI Guo-zheng, LU Wen-cong. Virus-ECC-mPLoc: a multi-label predictor for predicting the subcellular localization of virus proteins with both single and multiple sites based on a general form of Chou's pseudo amino acid composition[J]. Protein and Peptide Letters, 2013, 20(3): 309-317.

[84] WANG Xiao, LI Guo-zheng. Multi-label learning via random label selection for protein subcellular multi-locations prediction[J]. IEEE/ACM Trans on Computational Biology Bioinformmatics, 2013, 10(2): 436-446.

[85] KING B R, VURAL S, PANDEY S, *et al*. ngLOC: software and Web server for predicting protein subcellular localization in prokaryotes and eukaryotes[J]. BMC Research Notes, 2012, 5: 351.

[86] YUAN Zheng. Prediction of protein subcellular locations using Markov chain models[J]. FEBS Letters, 1999, 451(1): 23-26.

[87] EMANUELSSON O, BRUNAK S, Von HEIJNE G, *et al*. Locating proteins in the cell using TargetP, SignalP and related tools[J]. Nature Protocols, 2007, 2(4): 953-971.

[88] CHOU Kuo-chen, ELROD D W. Protein subcellular location prediction[J]. Protein Engineering, 1999, 12(2): 107-118.

[89] KHAN A, MAJID A, HAYAT M. CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition[J]. Computational Biology and Chemistry, 2011, 35(4): 218-229.

[90] LIANG Ru-ping, HUANG Shu-yun, SHI Shao-ping, *et al*. A novel algorithm combining support vector machine with the discrete wavelet transform for the prediction of protein subcellular localization[J]. Computers in Biology and Medicine, 2012, 42(2): 180-187.

[91] CHOU Kuo-chen, ZHANG Chun-ting. Prediction of protein structural classes[J]. Critical Reviewsin Biochemistry and Molecular Biology, 1995, 30(4): 275-349.

[92] CHOU Kuo-chen. Some remarks on predicting multi-label attributes in molecular biosystems[J]. Molecular BioSystems, 2013, 9(6): 1092-1100.

[93] DU Pu-feng, XU Chao. Predicting multisite protein subcellular locations: progress and challenges[J]. Expert Review of Proteomics, 2013, 10(3): 227-237.

[94] DU Pu-feng, LI Ting-ting, WANG Xin. Recent progress in predicting protein sub-subcellular locations [J]. Expert Review of Proteomics, 2011, 8(3): 391-404.