doi:10.3969/j.issn.1672-5565.2014.04.07

蛋白质亚叶绿体和亚线粒体定位预测研究进展

王星支,李凤敏*,王晓茜
(内蒙古农业大学理学院,内蒙古 呼和浩特 010018)

摘 要: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,因此蛋白质的功能与其亚细胞定位有着密切的联系,通过确定蛋白质在细胞中的位置可以获取蛋白质功能和结构的信息。在近二十年中,蛋白质亚细胞定位预测算法研究已经取得很大的成绩,在此基础上,蛋白质在细胞器内亚结构的定位预测研究,如对蛋白质亚线粒体和亚叶绿体定位的研究成为更深层次的问题,本文简要介绍国内外在蛋白质亚叶绿体和亚线粒体定位预测方面的研究进展。

关键词:亚线粒体;亚叶绿体;定位;预测;进展
中图分类号:Q61 **文献标志码**:A **文章编号**:1672-5565(2014)-04-276-05

Progress in predicting protein subchloroplast and submitochondria locations

WANG Xingzhi, LI Fengmin, WANG Xiaoqian
(*College of Science, Inner Mongolia Agriculture University, Hohhot* 010018, *China*)

Abstract: Protein is transported to specific organelles after synthesizing and only transferred to the right position can it participate in all kinds of life activities of cells and function effectively. So the function of protein is closely related with its subcellular location. The information of protein function and structure can be obtained by predicting the subcellular location. Much progress have achieved in predicting protein subcellular location in recent decades. Based on these, the research of sub-subcellular locations, such as subchloroplast and submitochondria, has become a deeply level problem. In this paper, the progress in predicting protein subchloroplast and submitochondria locations has been introduced domestic and foreign.

Keywords: Subchloroplast; Submitochondria; Location; Prediction; Progress

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,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[1]。对于一个给定的蛋白质,知道其亚细胞位置就可以了解蛋白质的生物功能,而且还能解释不同的蛋白质和其它分子之间的相互作用,特定的位置可以改变蛋白质生理状态的变化,而且在不同的环境下可以扮演不同的角色,与其他大分子的相互作用具有重要的意义[2]。蛋白质亚线粒体和亚叶绿体定位的研究是更深层次的研究课题,为探究蛋白质的结构功能提供了更多的信息。

蛋白质亚细胞位置可以用实验方法来确定,如X射线晶体衍射,电子显微镜,核磁共振等[3],然而这些方法通常花费大量的金钱和时间,而且在实验过程中还会遇到很多难以解决的问题。近二十年来,随着蛋白质序列的日益增多,利用理论模型来预测其亚细胞定位方便快捷,成本较低,因此生物信息学已经成为生命科学中的带头学科。

叶绿体是只存在于绿色植物和真核藻类亚细胞可以进行光合作用的细胞器,了解蛋白质的亚叶绿体定位有助于了解蛋白质的功能[4]。叶绿体可以分为四个区域:基质、类囊体腔、类囊体膜和被膜。线粒体在能量代谢过程中扮演重要的角色,如氧化磷酸化,氨基酸代谢,脂肪酸氧化。线粒体参与诸如

收稿日期:2014-08-22;**修回日期**:2014-11-10.
基金项目:国家自然科学基金项目(31360206)资助,内蒙古自治区人才开发基金资助。
作者简介:王星支,女,硕士研究生,研究方向:理论生物物理;E-mail: 731258945@qq.com.
*通信作者:李凤敏,女,教授,硕士生导师,研究方向:理论生物物理,生物信息学;;E-mail:lfmbs@126.com.

细胞分化、调节体内钙,铁离子平衡、细胞信息传递和细胞凋亡和生长等过程。线粒体可以分为三个区域即:线粒体内膜、线粒体外膜、线粒体基质。功能异常的线粒体会导致能量代谢障碍,进而导致一系列相互作用的损伤状态。许多的疾病和线粒体有关,如常见的多因子的紊乱[5],帕金森氏症,糖尿病等。因此了解蛋白质亚线粒体定位能够进一步了解蛋白质功能 ,同时可以为由线粒体缺陷引起的疾病进行辅助药物设计提供帮助。

1　蛋白质亚叶绿体和亚线粒体定位预测

对蛋白质亚叶绿体和亚线粒体定位预测遵循以下的步骤:

(1)数据集的建立;

(2)从这些蛋白质数据中抽取特征信息向量;

(3)选择合适的算法,根据前面的特征信息向量做出预测;

(4)对预测结果进行评价。

1.1　蛋白质亚叶绿体和亚线粒体定位的相关数据集

随着机器学习方法的不断进步,不断创新,蛋白质亚细胞定位的预测算法研究有了新的进展,蛋白质序列数据库更加完善,在过去的二十年里,蛋白质序列的数量以指数增长,从 2002 年起,UniProt 数据库的蛋白质序列数量每 2 年翻一番,UniProt 是信息最丰富、资源最广的蛋白质数据库,它整合 Swiss-Prot、TrEMBL 和 PIR-PSD 三大数据库的数据而成。它的数据主要来自于基因组测序项目完成后,后续获得的蛋白质序列,它包含了大量来自文献的蛋白质的生物功能的信息。

除了这些综合数据库(UniProt、PIR 和 MIPS 等),目前出现了一些专门的亚亚细胞定位数据库,例如 2006 年首次由 Du and Li 提出的对线粒体蛋白质在亚线粒体位置预测的方法 SUBMITO[6]。它所建立的原始数据集 SUBMITO 共包含 737 条线粒体蛋白质,该数据库中的蛋白质经过 CD-HIT 处理得到包含了 317 条蛋白质亚线粒体序列。在近几年来的研究中,研究者利用不同的处理数据方式得到了不同的线粒体数据集,如 M317,M983[7],M399[8],M1105[9],M495[10]等。

目前,对蛋白质亚叶绿体定位的理论预测选取的数据集有以下几种:由 Du 和 Li 创建的 SubChlo 里的原始数据库有 736 条亚叶绿体蛋白质序列[11],如果用 80%,60%,40%的 CD-HIT 百分比去处理原始数据集会得到相应的 S80,S60,S40 数据集[12]。

1.2　蛋白质特征信息的提取

蛋白质特征信息的提取是蛋白质亚细胞定位预测以及亚亚细胞定位预测的基础。蛋白质在合成过程中被分选到特定的亚细胞器中发挥出生物学功能,很大程度上是由蛋白质的序列特征决定的,序列特征大致可以分为蛋白质分选信号,氨基酸性质与组成,功能域信息等。

合成的蛋白质必须定向地转运到特定细胞器中,一个重要的原因就是蛋白质中包含了各种不同的分选信号,一种信号序列决定了特定蛋白的转运方向,可以被细胞器上的分选受体特异性识别。例如蛋白质序列的 N 端分选信号的方法[13],因为氨基酸序列中 N 端或者 C 端存在特殊的分选信号,这种特征信息提取方法较多地利用了蛋白质的生物学分选过程信息。但是实际上对于基因 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,预测效果并不是很好。

氨基酸组成是一种最基本的氨基酸组分信息[14]。氨基酸组分信息即蛋白质的一级结构信息,氨基酸是组成蛋白质的基本结构单位,蛋白质由 20 种氨基酸组成。氨基酸组分是将 20 种氨基酸在蛋白质序列中出现的频数抽取出来作为一个 20 维的向量。

Huang 和 Li 提出二肽组分的预测方法是利用相邻的两个氨基酸残基出现的频数来对蛋白质序列进行描述,将蛋白质序列映射为一个 400 维的特征向量[15]。Yu 等人提出了 n 肽组分信息的预测方法[16]。

Chou 小组在氨基酸组成信息的基础之上[17],提出了物理化学性质信息,根据不同的物理化学特征,氨基酸可以分为亲疏水性 6 类、极性 4 类、酸碱性 3 类以及 R 基 6 类的化学结构,这样可以将氨基酸的 20 维序列信息大大降低,减少其复杂性。Chou 等又将氨基酸的亲水性、疏水性等物理化学性质整合到氨基酸组分中,定义为伪氨基酸组分,伪氨基酸组分信息表达在非完全失去蛋白质序列秩序信息的离散模型中的蛋白质序列特征。Chou 等建立了伪氨基酸信息服务器 PseAAC,该服务器包括 6 种特征信息,即(疏水性、亲水性、分子质量、碱性,酸性,等电点)在使用时可以任意选择其中的一种特征信息,或者随意组合。在许多的研究中都选用伪氨基酸组分信息作为特征信息[18]。

基因本体(Gene Ontology,GO)是一个在生物信息学领域广泛使用的本体,对基因和蛋白质功能进行限定和描述[19]。GO 数据库的建立主要包括生物

过程,分子功能和细胞组分三个分支。基因注释与功能分类是功能基因组学和计算系统生物学的重要基础。目前它已经成为各种数据库的基因产品注释的标准化工具,并被用于不同基因序列预测。但是由于此方法对数据库功能注释信息的完善程度要求比较高,如果数据库中没有足够的功能域或基因注释条目,那么将无法准确确定蛋白质的亚细胞定位。

1.3 预测算法

成功的分类算法应该是能够正确,高效的将不同蛋白质的亚细胞位置分开。在这方面主要的机器学习算法有近邻方法、人工神经网络方法、支持向量机方法 、马尔可夫模型方法、贝叶斯网络组合耦合算法、马尔可夫模型希尔伯特变换等,此外还有离散增量算法。目前,在蛋白质亚细胞定位研究领域,常用的预测算法有人工神经网络、支持向量机、最近邻算法和随机森林算法等。

1.4 预测性能评估

在统计预测中,以下三种评价方法常用于检查一种预测算法的预测精度,它们分别是自洽检验(Self-consistency),交叉检验(Cross-validation)和留一交叉检验(Jacknife)。留一交叉检验是交叉检验的一种特殊情况,其原理就是训练集中的每一个蛋白质轮流作为待测蛋白,其余蛋白质作为训练样本对其进行测试,是目前较为客观的检验方法之一。

预测结果主要用以下指标来衡量:

敏感性(sensitivity 即 Sn):每类样本中被正确识别的比例。特异性(specificity 即 Sp):表示分类预测中每个类别预测结果的可信度。(3)预测成功率(accuracy 即 Acc):表示预测的总体成功率。马修相关系数(Matthews's Correlation Coefficient 即 MCC):一个整体评价指标,反应预测的综合能力,它们定义如下:

$$Sn = \frac{TP}{TP + FN} \tag{1}$$

$$Sp = \frac{TN}{TN + FP} \tag{2}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

其中,TP 是第 i 类样本中被预测正确的数目 ,FP 是第 i 类样本被错误的判别为其他类别的数目,FN 是非第 i 类样本但被预测为第 i 类样本的数目,TN 是非第 i 类样本中被预测正确的样本数目。

2 不同方法的预测结果

2.1 对亚线粒体蛋白的预测结果

亚线粒体原始数据集选用 CD-HIT 程序去除相似性大于 40%的蛋白质,还有通过其他不同的标准选取得到不同的线粒体蛋白质数据集。以下是目前国内外常用的几个数据集及采用的预测方法。

数据集 M983 是 Du and Yu 从 UniProt 数据库中提取出来的数据集,该数据库包含 661 条内在膜蛋白、177 条基质蛋白和 145 条外膜蛋白。通过选取物理化学信息与伪氨基酸信息结合,利用 SVM 预测蛋白质亚线粒体位置得到了 89.01%的预测成功率。

数据集 M317 由 Du and Li 于 2006 年构建,初始数据选自 Swiss-Prot,该数据库共有 317 条蛋白,其中 131 条内在膜蛋白,145 条基质蛋白和 41 条外膜蛋白。表 1 是国内外研究人员对 M317 数据集的研究结果。

表 1 M317 数据集的研究结果

Table 1 The predicted results of M317 database

数据集	作者(年)	方法	预测准确率
M317	Du and Li (2006)	SubMito	85.2%
M317	Nanni and Lumini (2008)	GP-Loc	89.0%
M317	Shi et al (2011)	Subldent	93.4%
M317	Zakeri et al (2011)	SVM	94.7%
M317	Suyu Mei(2012)	MK-TLM	99.7%
M317	Fan and Li (2012)	SVM	94.9%
M317	Lin Hao (2013)	TetraMito	94.0%

数据集 M399 由 Zeng 等人 2009 年构建,该数据库包含 399 条蛋白,其中有 171 条内在膜蛋白,166 条基质蛋白和 62 条外膜蛋白,对该数据集的研究结果见表 2。

表 2 数据集 M399 的研究结果

Table 2 The predicted results of M399 database

数据集	作者(年)	方法	预测准确率
M399	Zeng et al. (2009)	SVM	89.7%
M399	Lin Hao (2013)	TetraMito	94.7%

数据集 M1105 由 fan and Li 于 2012 年构建,该数据库共有 1 105 条蛋白,其中有 589 条内在膜蛋白、280 条基质蛋白和 236 条外膜蛋白,对数据集 M1105 的研究情况见表 3。

数据集 M495 由 Lin Hao 等人于 2013 年构建,该数据库包含 254 条内在膜蛋白,132 条基质白和 109

条外在膜蛋白,对该数据库的预测成功率为89.7%。

表 3　数据集 M1105 的研究结果

Table 3　The predicted results ofM1105 database

数据集	作者(年)	方法	预测准确率
M1105	Fan and Li (2012)	SVM	89.7%
M1105	Lin Hao (2013)	TetraMito	94.7%

2.2　对亚叶绿体蛋白的预测结果

由 Du 和 Li 于 2009 年所建立的 SubChlo 亚叶绿体数据集 Raw-736,该数据库共有 736 条叶绿体蛋白质, 其中 71 条基质蛋白,60 条类囊体腔蛋白,516 条类囊体腔蛋白,89 条被膜蛋白。表 4 是利用四种方法对 Raw-736 数据库的预测结果。

表 4　对 Raw-736 数据库的预测结果

Table 4　The predicted results of Raw-736 database

方法	胞内位置				预测成功率
	基质	类囊体腔	类囊体膜	被膜	
subchlo	78.90%	55.00%	96.10%	84.40%	89.70%
subldent	84.04%	77.22%	98.90%	91.11%	94.75%
SubChlo-GO	93.0%	95.0%	95.3%	91.0%	94.60%
SCLAP	85.92%	82.22%	100%	100%	97.96%

亚叶绿体数据集 S60-261 来自于 SubChlo,包含了 49 条基质蛋白,44 条类囊体腔蛋白,129 条类囊体膜蛋白,39 条被膜蛋白。表 5 是用五种方法对 S60-261 数据库的预测结果。

表 5　对 S60-261 数据库的预测结果

Table 5　The predicted results of S60-261 database

方法	胞内位置				预测成功率
	基质	类囊体腔	类囊体膜	被膜	
subchlo	67.4%	43.2%	83.7%	40.0%	67.2%
chloroRF	57.1%	38.6%	87.5%	47.5%	67.4%
subldent	85.7%	64.4%	98.2%	80.0%	89.3%
SubChlo-GO	89.8%	88.6%	93.0%	71.8%	88.50%
SCLAP	91.3%	85.0%	72.5%	95.2%	89.3%

亚叶绿体数据集 S60-253 由 Yan 和 Hu 于 2012 年构建[20],该数据库包含了 46 条基质蛋白,40 条类囊体腔蛋白,127 条类囊体膜蛋白,40 条被膜蛋白。利用 BS_KNN 方法对该数据库的总预测准确率为 75.9%。

亚叶绿体数据集 S60-259 由 Lin Hao 于 2013 年构建[21],该数据库包含了 60 条基质蛋白,19 条类囊体腔蛋白,103 条类囊体膜蛋白,77 条被膜蛋白。通过选取三肽特征信息利用 SVM 方法对该数据库的总预测准确率为 88.03%。

2013 年 Yuan 和 Huang 通过设定不同的 CD-HIT 百分比得到 S40、S60、S80 数据库,表 6 是利用 PSSM 和 OET-KNN 方法对 S40、S60、S80 数据库进行预测得到的预测结果。

表 6　对 S40、S60 和 S80 数据库的预测结果

Table 6　The predicted results of S40,S60 and S80 database

数据集	胞内位置				预测成功率
	基质	类囊体腔	类囊体膜	被膜	
数据集 S80	82.00%	68.97%	92.18%	87.63%	89.08%
数据集 S60	66.09%	60.005	87.47%	81.99%	81.29%
数据集 S40	58.95%	55.56%	87.47%	72.40%	71.11%

3　展望与结语

本文介绍了国内外亚叶绿体蛋白和亚线粒体蛋白定位预测方面的研究进展,可以看到蛋白质亚线粒体和亚叶绿体定位理论研究作为实验研究的补充已经取得了一些成果,但还存在一些问题需要进一步解决。(1)不同的理论模型所选的蛋白质特征信

息参数不同,适用的数据库也不同,而且当数据库增大或蛋白质序列相似性降低时,预测结果也会降低。(2)有些蛋白质在线粒体和叶绿体内定位不是固定的,具有多定位,这些蛋白质的生物学功能更加重要,但目前对这些多定位蛋白质的研究很少。总之,随着蛋白质亚线粒体和亚叶绿体定位数据库的不断丰富和完善,对于蛋白质亚线粒体和亚叶绿体定位预测,无论是从蛋白质特征信息提取还是预测模型的改善都需要进行深入系统的研究。

参考文献(References)

[1] 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进展, 2007, 34(6): 573-579.
ZHANG Song, HUANG Bo, XIA Xuefeng, et al. Bioinformatics research in subcellular location of protein[J]. Progress in Biochemistry and Biophysics, 2007, 34(6): 573-579.

[2] DU P, LI T, WANG X. Recent progress in predicting protein sub-subcellular locations[J]. Expert Review of Proteomics, 2011, 8(3): 391-404.

[3] MURPHY R F, BOLAND M V, VELLISTE M. Towards a systematics for protein subcellular location: Quantitative description of protein localization patterns and automated analysis of fluorescence microscope images[C]. ISMB, 2000: 251-259.

[4] DU P, LI T, WANG X, et al. SubChlo-GO: Predicting protein subchloroplast locations with weighted gene ontology scores[J]. Current Bioinformatics, 2013, 8(2): 193-199.

[5] SHI S P, QIU J D, SUN X Y, et al. Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction[J]. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research, 2011, 1813(3): 424-430.

[6] DU P, LI Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence [J]. BMC Bioinformatics, 2006, 7(1): 518.

[7] DU P, YU Y. SubMito-PSPCP: Predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions[J]. BioMed Research International, 2013, 2013: 1-7.

[8] ZENG Y, GUO Y, XIAO R, et al. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach[J]. Journal of Theoretical Biology, 2009, 259(2): 366-372.

[9] FAN G L, LI Q Z. 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.

[10] LIN H, CHEN W, YUAN L F, et al. Using over-represented tetrapeptides to predict protein submitochondria locations[J]. Acta Biotheoretica, 2013, 61(2): 259-268.

[11] DU P, CAO S, LI Y. SubChlo: Predicting protein subchloroplast locations with pseudo-amino acidcomposition and the evidence-theoretic K-nearest neighbor (ET-KNN) algorithm[J]. Journal of Theoretical Biology, 2009, 261(2): 330-335.

[12] HUANG C, YUAN J Q. Predicting protein subchloroplast locations with single and multiple sites via three different modes of Chou's pseudo amino acid compositions.[J]. Journal of Theoretical Biology, 2013, 335(2): 205-212.

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

[14] 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.

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

[16] 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.

[17] CHOU K C. Prediction of protein cellular attributes using pseudo-amino acid composition[J]. Proteins: Structure, Function, and Bioinformatics, 2001, 43(3): 246-255.

[18] LI L, YU S, XIAO W, et al. Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach[J]. Biochimie, 2014, 104: 100-107.

[19] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene ontology: tool for the unification of biology [J]. Nature Genetics, 2000, 25(1): 25-29.

[20] HU J, YAN X. BS-KNN: An effective algorithm for predicting protein subchloroplast localization[J]. Evolutionary Bioinformatics Online, 2012, 8: 79.

[21] LIN H, DING C, YUAN L F, et al. Predicting subchloroplast locations of proteins based on the general form of Chou's pseudo amino acid composition: approached from optimaltripeptide composition [J]. International Journal of Biomathematics, 2013, 6(02).