



## 研究报告

## Research Report

# 真核生物蛋白质亚细胞定位的计算预测：一个经验报告

John Meinken<sup>1</sup>, Jack Min<sup>2</sup>

<sup>1</sup> 扬斯敦州立大学计算机科学与信息学院, 扬斯敦, OH44555, 美国

<sup>2</sup> 扬斯敦州立大学生物科学学院应用化学生物学研究中心, 扬斯敦, OH44555, 美国

✉ 通讯作者: xmin@ysu.edu; ✉ 作者

计算分子生物, 2012 年, 第 1 卷, 第 8 篇 doi: 10.5376/cmb.cn.2012.01.0008

收稿日期: 2012 年 10 月 01 日

接受日期: 2012 年 10 月 01 日

发表日期: 2012 年 10 月 01 日

本文首次发表在 *Computational Molecular Biology* 上。现依据版权所有人授权的许可协议, 采用 *Creative Commons Attribution License* 协议对其进行授权, 再次发表与传播。只要对原作有恰当的引用, 版权所有人允许并同意第三方无条件的使用与传播。 建议最佳引用格式:

引用格式:

Meinken and Min, 2012, *Computational Prediction of Protein Subcellular Locations in Eukaryotes: an Experience Report*, *Computational Molecular Biology*, Vol.2, No.1 1-7 (doi: 10.5376/cmb.2012.02.0001)

**摘要** 在真核生物中蛋白质亚细胞位置的计算预测有利于实验设计和蛋白质组分析。我们提供了一个关于的计算工具最新进展, 和一些我们评估这些工具的经验简述。计算工具可以相对准确地预测经典分泌蛋白质组, 从而来预测分泌信号肽的存在及消除跨膜蛋白和内质网 (endoplasmic reticulum, ER) 蛋白。通过差异结合 SignalP, Phobius, WoLFPSORT 和 TargetP 来识别真核生物在不同领域的分泌信号肽, TMHMM 来消除跨膜蛋白和 PS-Scan 来消除内质网蛋白的协议, 显著地提高了分泌蛋白预测精度。我们的评估表明, 目前用于预测其他的亚细胞位置, 包括线粒体或叶绿体定位的计算工具仍然需要改进。

**关键词** 真核生物; 蛋白质亚细胞定位预测; 蛋白质分泌组; 计算预测

## Computational Prediction of Protein Subcellular Locations in Eukaryotes: an Experience Report

John Meinken<sup>1</sup>, Jack Min<sup>2</sup>

<sup>1</sup> Department of Computer Science and Information Systems, Youngstown State University, Youngstown, OH 44555, USA

<sup>2</sup> Center for Applied Chemical Biology, Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555, USA

✉ Corresponding author, xmin@ysu.edu ✉ Authors

**Abstract** Computational prediction of protein subcellular locations in eukaryotes facilitates experimental design and proteome analysis. We provide a short review on recent development of computational tools and our experience in evaluating some of these tools. Classical secretomes can be relatively accurately predicted using computational tools to predict existence of a secretory signal peptide and to remove transmembrane proteins and endoplasmic reticulum (ER) proteins. The protocols of differentially combining SignalP, Phobius, WoLFPSORT, and TargetP for identifying a secretory signal peptide in different kingdom of eukaryotes, with TMHMM for removing transmembrane proteins and PS-Scan for removing ER proteins significantly improve the secretome prediction accuracies. Our evaluation showed that current computational tools for predicting other subcellular locations, including mitochondrial or chloroplast localization, still need to be improved.

**Keywords** Eukaryotes; Protein subcellular location; Secretome; Computational prediction

真核细胞有一个复杂的内膜系统, 除了独立的细胞器结构如线粒体和叶绿体。这些亚细胞结构包括细胞核、内质网 (ER), 高尔基体、溶酶体、过氧化物酶体、液泡、细胞骨架、细胞溶质、线粒体、叶绿体及细胞膜。这些亚细胞膜的封闭结构由膜和内部空间如内质网腔组成。质膜外, 细胞壁和细胞外基质和空间也是细胞活动的重要场所。

真核细胞合成成千上万种不同的蛋白质。例如, 通常被称为面包酵母的酿酒酵母 (*Saccharomyces cerevisiae*), 其基因组相对较小, 为 12 MB, 编码大约 5 000~6 000 种不同的蛋白质。游离在细胞质中的或粗面内质网上的核糖体合成由核基因组编码的蛋白质。但是这些蛋白质需要被转移到一个或多个特定的亚细胞位置才能发挥其生物学作用, 这一



过程称为蛋白质的定位或分类。识别蛋白质的亚细胞位置的实验方法已经被广泛利用, 包括细胞器的隔离, 绿色荧光标记蛋白等(Heazlewood et al., 2005)。实验证实一些信号定位肽决定蛋白质的亚细胞位置(Blobel and Dobberstein, 1975)。因此, 蛋白质的定位被认为是由目标区域的蛋白质的物理和化学性质决定。这些目标区域可以从蛋白质定位的氨基酸序列确定。最近已经开发了一些计算工具来预测真核细胞的蛋白质的亚细胞位置。Nakai and Horton (2007)全面综述了亚细胞定位预测的计算方法和工具。我们对这方面的最新进展进行了一个简述, 并讨论了在我们的研究经验基础上未来发展面临的挑战。

### 分泌信号肽和蛋白质的预测

分泌蛋白质组这个术语是用以指有机体中一整套在细胞外分泌的蛋白质, 包括细胞壁、细胞外基质和细胞外空间。最近, 人们已作出许多努力来识别这些分泌蛋白质组, 因为这些蛋白在环保工业和生物医学领域有应用潜力(Lum and Min, 2011; Makridakis and Vlahou, 2010)。例如, 真菌分泌蛋白质组常包含有分泌细胞外的酶来分解生物聚合物, 这些生物聚合物在生物燃料生产的应用潜力大(Lum and Min, 2011)。人类分泌蛋白质组发挥着重要的生物学作用, 如胰岛素, 并且为新的生物标记的发掘, 如用于癌症诊断提供了有用的信息(Makridakis and Vlahou, 2010)。

分泌蛋白质组由两种蛋白质组成: 经典和非经典分泌蛋白。一个典型的经典分泌蛋白含有位于 N 端的分泌信号肽, 且在其序列中不含其它定位信号(Emanuelsson et al., 2007)。分泌信号肽指示核糖体靠近粗面内质网, 从而完成一个含有信号肽的蛋白质的合成。分泌信号肽, 通常是 15~30 个氨基酸长, 在整个膜的转运过程中被裂解(Von Heijne, 1990)。应明确的一个基本概念是, 并不是所有含有分泌信号肽的蛋白质都是分泌的。一些文献报道的预测的分泌蛋白质组是完全基于分泌信号肽预测的存在, 导致在蛋白质组中分泌蛋白数量的高估。分泌蛋白只占进入内质网分泌途径的蛋白质的一部分, 如含有信号肽并且进入内质网的蛋白质还包括的粗面内质网, 光面内质网、高尔基体、溶酶体、胞内体与质膜内的固有成分。

目前分泌性信号肽预测的常用工具包括

SignalP 3.0 (Bendtsen et al., 2004b), SignalP 4.0 (Petersen et al., 2011), Phobius (Käll et al., 2004; 2007), TargetP (Emanuelsson et al., 2000)和 PrediSi (<http://www.predisi.de/>) (Hiller et al., 2004)。此外, WoLFPSORT 和 MultiLoc2 也可用于分泌性蛋白的预测(Horton et al., 2007; Blum et al., 2009)。SignalP 4.0 的精密度超过 SignalP 3.0, 具有较高的跨膜预测改进的特性。然而, SignalP 3.0 在信号肽的切割位点上的预测比 SignalP 4.0 更准确。在 SignalP 3.0/4.0 和 PrediSi 中, N 端肽的默认长度是 70 残基, 因此当使用默认的切割参数时, 具有长信号肽 (>70 个氨基酸) 的蛋白质序列不能预测。Phobius 也是一个相对准确的信号肽预测工具, 结合了跨膜拓扑结构和信号肽预测。

总体而言, 上述工具的信号肽预测的准确性是在一般用途上是可以接受的。然而, 我们对这些工具的最新评估显示, 经典分泌性蛋白的预测精度可以通过结合多个工具显著改善, 主要是由于在预测特异性上的提升(Min, 2010)。此外, 加入 TMHMM 去除跨膜蛋白和 PS-Scan (从 Scan Prosite 上下载的单机版)去除内质网存留的蛋白(Prosite: PS00014, 内质网靶向序列)显着提高了蛋白质组预测的准确性(Min, 2010)。我们的评价也表明不同的工具在处理真核生物有机体中不同区域产生的蛋白质数据的强度不同。我们提出以下真核细胞在不同区域的分泌蛋白预测的协议: 在真菌中使用 SignalP/WoLFPSORT/Phobius, 动物使用 Phobius/WoLFPSORT/TargetP, 植物使用 SignalP/Phobius/TargetP, 原生生物使用 SignalP/Phobius/TargetP/WoLFPSORT。当使用两个或两个以上的工具时, 信号肽预测的特异性显著增加。此外, TMHMM 和 PS-Scan 应该用于所有真核分泌性蛋白质的预测。

只有一个工具, SecretomeP, 适用于哺乳动物和细菌有机体的非经典分泌性蛋白预测(<http://www.cbs.dtu.dk/services/SecretomeP/>) (Bendtsen et al., 2004a)。在植物中, 大约 50% 的分泌性蛋白估计为非经典分泌性蛋白, 即无领导分泌蛋白 (LSPs) (Agrawal et al., 2010), 当然植物特有的非经典分泌蛋白质组预测需要植物特有的预测工具或方法。

### 多个亚细胞位置的预测因子

TargetP 通过区分叶绿体转运肽 (cTP、植物),



线粒体靶向肽 (mTP) 和分泌途径的信号肽来预测真核细胞蛋白质的亚细胞位置, (Emanuelsson et al., 2007)。除了真菌蛋白的数据集, 通过 TargetP 和 SignalP, TMHMM 以及 PS-Scan 结合, 增加了其它的真核生物分泌蛋白组数据集的预测精度。用于预测多个亚细胞位置的其他广泛使用的工具是 WoLFPSORT 和 MultiLoc2。WoLFPSORT 预计 12 个亚细胞位置包括叶绿体、细胞质、细胞骨架、内质网、细胞外、高尔基体、溶酶体、线粒体、细胞核、过氧化物酶体、质膜和液泡膜(Horton et al., 2007)。MultiLoc2 为动物和真菌预测 9 个亚细胞定位和 10 个植物亚细胞定位(Blum et al., 2009)。Chou and Shen(2008)开发了一系列的 Web 服务器, 称为 Cell-PLoc, 其中包括 6 个不同的服务器。这些服务器能够预测各种生物包括病毒、细菌、植物、人类蛋白质或一般的真核生物中, 高达 22 个的蛋白质亚细胞定位。然而, Cell-PLoc 的一系列服务器只能处理每次提交的单一序列, 并没有能够使用的单机

工具, 阻碍了我们进一步评估这些工具的精度。

而开发的植物分泌蛋白组的知识库 ( PlantSecKB ) 目前是公开的 (<http://proteomics.yzu.edu/secretomes/plant.php>), 我们通过使用一组从 UniProtKB Swiss-Prot 数据集获得的植物蛋白比较了 TargetP, WoLFPSORT 和 MultiLoc2 的预测精度。具有多个亚细胞位置或标记为“片段”, 或在亚细胞定位注释上有术语“相似”或“可能”或“预测”的蛋白质已被去除。共选择了 6908 个已经被注释的亚细胞位置的蛋白质。结果如表 1 所示。如果我们忽略小于 100 个阳性条目的亚细胞定位预测, 我们的评价表明, 用三个工具进行分泌蛋白的预测比其他亚细胞位置的预测是相对更准确的。TargetP 在分泌蛋白预测上比其他两个工具更准确。三个工具的对所有其他亚细胞定位预测马修斯相关系数 (MCC) (Matthews, 1975)值均低于 50%。因此, 这些亚细胞位置的植物蛋白的预测精度确实需要改善。

表 1 不同的工具比较植物蛋白的亚细胞位置预测精度

Table 1 Comparison of prediction accuracies of plant protein subcellular locations by different tools

	总阳性 Total positives	总阴性 Total negatives	TargetP			WoLFPSORT			MultiLoc2		
			Sn (%)	Sp (%)	MCC (%)	Sn (%)	Sp (%)	MCC (%)	Sn (%)	Sp (%)	MCC (%)
分泌物 Secreted	263	6645	76.8	98.7	72.3	43.7	99.6	58.0	51.3	98.6	53.7
线粒体 Mitochondrial	402	6506	61.4	77.5	21.1	33.3	96.2	30.4	60.9	84.2	27.3
叶绿体 Chloroplast	4918	1990	28.2	90.7	20.4	28.2	83.9	12.7	19.2	98.6	23.2
内质网 ER	87	6821	-	-	-	9.2	99.9	24.5	29.9	98.4	23.0
细胞质 Cytosol	23	6885	-	-	-	52.2	65.6	2.2	95.7	51.5	5.4
高尔基体 Golgi Apparatus	54	6854	-	-	-	0.0	99.9	-0.2	24.1	98.5	15.5
过氧化物酶体 Peroxisome	52	6856	-	-	-	15.4	99.3	14.3	42.3	98.1	23.9
细胞核 Nucleus	788	6120	-	-	-	80.3	78.1	41.1	26.8	93.3	22.4
质膜 Plasma Membrane	14	6894	-	-	-	21.4	97.9	6.1	7.1	99.4	3.9
液泡 Vacuole	121	6787	-	-	-	14.0	99.5	21.2	5.0	99.8	11.6

注: Sn: 敏感性; Sp: 特异性; MCC: 马修斯相关系数

Note: Sn: Sensitivity; Sp: Specificity; MCC: Matthews' correlation coefficient



表 2 公布的蛋白质亚细胞定位预测工具的集合

Table 2 A collection of published protein subcellular localization prediction tool

工具名称 Tool name	位置或蛋白质的功能预测 Locations or protein features predicted	生物类 Organism categories	出版的参考文献 Publication citation
TargetP	出版的引文 Extracellular, mitochondrial, chloroplast	非植物、植物 Non-plant, plant	Emanuelsson et al., 2000
TMHMM	跨膜螺旋 Transmembrane helices	任何 Any	Krogh et al., 2001
ScanProsite (PS-Scan)	内质网保留信号 ER retention signal	任何 Any	de Castro et al., 2006
SecretomeP 2.0	非经典即非信号肽引发的蛋白质分泌 Non-classical i.e. not signal peptide triggered protein secretion	革兰氏阳性菌、革兰氏阴性细菌、哺乳动物 Gram-positive bacteria, gram-negative bacteria, mammal	Bendtsen et al., 2004a
Phobius	信号肽和跨膜结构 Signal peptide and transmembrane topology	任何 Any	Käll et al., 2007
WoLF PSORT	>5个位置 > 5 locations	真菌、植物、动物 Fungi, plant, animal	Horton et al., 2007
PRED-LIPO	Lipoprotein signal peptides	革兰氏阴性菌 Gram-positive bacteria	Bagos et al., 2008
ProLoc-GO	>5个位置 > 5 locations	人类和真核生物 Human and eukaryotes	Huang et al., 2008
KnowPredsite	>5个位置 > 5 locations	原核生物和真核生物 Prokaryotes and eukaryotes	Lin et al., 2009
MultiLoc2	>5个位置 > 5 locations	动物、真菌、植物 Animal, fungal, plant	Blum et al., 2009
PRED-SIGNAL	信号肽 Signal peptides	古生菌 Archaea	Bagos et al., 2009
RSLPred	细胞质、叶绿体、线粒体、细胞核 Chloroplast, cytoplasm, mitochondria, nucleus	仅稻类 <i>Oryza sativa</i> only	Kaundal and Raghava, 2009
SherLoc2	>5个位置 > 5 locations	动物、真菌、植物 Animal, fungi, plant	Briesemeister et al., 2009
Cell-PLoc 2.0	>5个位置 > 5 locations	真核生物、人类、植物、病毒、革兰氏阳性菌、革兰氏阴性菌 Eukaryote, human, plant, virus, gram-positive bacteria, gram-negative bacteria	Chou and Shen, 2010
CoBaltDB	>5个位置 > 5 locations	原核生物 Prokaryotes	Goudenège et al., 2010
PSORTb	>5个位置 > 5 locations	革兰氏阳性菌、革兰氏阴性菌 Gram-positive and gram-negative bacteria	Yu et al., 2010
SCLPred	细胞质、线粒体、细胞核、分泌物、叶绿体 Cytoplasm, mitochondrion, nucleus, secretory, chloroplast	动物、真菌、植物 Animals, plants, fungi	Mooney et al., 2011
SignalP 4.0	信号肽 Signal peptide	真核生物、革兰氏阳性菌、革兰氏阴性菌 Eukaryotic, gram-positive bacteria, gram-negative bacteria	Petersen et al., 2011
SlocX	>5个位置 > 5 locations	仅拟南芥 <i>Aribdopsis thaliana</i> only	Ryngajillo et al., 2011

总体来说, WoLFPSORT 和 MultiLoc2 利用其序列为基础的预测方法的预测精度差异不显著。MultiLoc2 纳入系统发育谱, 基因本体论术语, 其在报道中的表现大大优于其他动植物蛋白的预测方法(Blum et al., 2009)。然而, 它的精确性不能被公平测试, 因为我们的数据都有基因本体论注释。

此外, 我们还发现, MultiLoc2 在数据处理时比 WoLFPSORT 大约慢 500 倍, 在进行数据库开发的数据处理时我们没有使用 multiloc2。

### 其他计算工具

表 2 列出了亚细胞定位预测工具及其相关出版





物的集合。

所有这些工具的网站链接, 都可以在我们的网络服务器发现 (<http://proteomics.ysu.edu/tools/subcell.html>)。这不是一个详尽的清单, 但重点列出了本文中讨论的工具, 以及自 2008 年以来发布的最新工具。我们目前从 SignalP 3.0, SignalP 4.0, TMHMM, Phobius, TargetP, WoLFPSORT, PS-Scan 和 FragAnchor 收集的预测知识库如上所述。

一些工具只能进行单一的亚细胞定位或识别单一的蛋白质特性的存在预测 (如一个信号肽)。然后也有更全面的可以预测更多位置工具, 并且也可以采用多种计算方法的结合。最近几年乎是更趋向于更全面的工具。我们收集了 2008 年以来发表的这些工具, 十五分之十二能够进行四个或更多的亚细胞定位预测。

随着这么多已经可以预测的各种亚细胞位置的工具有出现, 人们可能会问, 我们将分析结果从多个工具组合成一个数据库的方法是否仍然是相关的。我们相信我们的工作能在这方面做出一些有价值的贡献。首先, 从多个预测的数据相结合, 往往会产生比单一的预测更准确的结果。这个准则已经在我们对于分泌蛋白质组的具体工作中表现 (Min, 2010), 它也是一个被广泛认可的统计概念。此外, 数据库可以在预测工具不能用时使用。对于大多数的预测工具, 分析是在请求后进行。用户必须提前知道他们感兴趣的蛋白质, 才可以得到分析结果。有了我们的数据库, 用户可以在另一个方向上工作。他们可以开始一个亚细胞定位和获得他们感兴趣的物种的符合这些标准蛋白质列表。

此外, 能执行相同任务的如此多的工具的发展对研究人员造成了一个困境, 谁必须选择他们将使用哪些工具。这需要进行一个测试, 比较不同的工具并确定他们的相对优势和劣势。也许一些工具对于植物表现更好, 而其他的工具对于细菌表现更好。一些工具可能对于特定的亚细胞定位有更好的特异性, 而其他则可能有更好的灵敏度。我们的知识库可以为进行这样丰富的数据集比较服务。在这项工作中, 我们利用 TargetP, WoLFPSORT 和 MultiLoc2 对植物蛋白的预测比较了精确度。我们需要进行更多的工作来继续进行这些类型的比较

研究, 从而提高全蛋白质组学蛋白的亚细胞定位在未来的预测精度。

## 致谢

本研究由 Ohio 植物生物技术协会和扬斯敦州立大学的支持 (YSU) 研究委员会 XJM 支持。

## 参考文献

- Agrawal G.K., Jwa N.S., Lebrun M.H., Job D., and Rakwal R., 2010, Plant secretome: unlocking secrets of the secreted proteins, *Proteomics*, 10: 799-827
- Bagos P.G., Tsirigos K.D., Plessas S. K., Liakopoulos T. D., and Hamodrakas S. J., 2009, Prediction of signal peptides in archaea, *Protein engineering, design & selection: PEDS*, 22(1): 27-35
- Bagos P.G., Tsirigos K.D., Liakopoulos T.D., and Hamodrakas S.J., 2008, Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model, *J. proteome res.*, 7(12): 5082-5093
- Bendtsen J.D., Jensen L.J., Blom N., Von Heijne G., and Brunak S., 2004a, Feature based prediction of non-classical and leaderless protein secretion, *Protein Eng. Des. Sel.*, 17(4): 349-356
- Bendtsen J.D., Nielsen H., Von Heijne G., and Brunak S., 2004b, Improved prediction of signal peptides: SignalP 3.0, *J. Mol. Biol.*, 340: 783-795
- Blobel G., and Dobberstein B., 1975, Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma, *J. Cell Biol.*, 67: 835-851
- Blum T., Briesemeister S., and Kohlbacher O., 2009, MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction, *BMC Bioinformatics*, 10: 274
- Briesemeister S., Blum T., Brady S., Lam Y., Kohlbacher, O., and Shatkay H., 2009, SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins, *J. proteome res.*, 8(11): 5363-5366
- Chou K.C., and Shen H.B., 2008, Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms, *Nat. protoc.*, 3(2): 153-162
- Chou K., and Shen H., 2010, Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms, *Natural Science*, 2: 1090-1103, doi: 10.4236/ns.2010.210136
- de Castro E., Sigrist C.J., Gattiker A., Bulliard V., Langendijk-Genevaux P.S., Gasteiger E., Bairoch A., and Hulo N., 2006, ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins, *Nucleic Acids Res.*,



- 34(Web Server issue): W362-365
- Emanuelsson O., Brunak S., Von Heijne G., and Nielsen H., 2007, Locating proteins in the cell using TargetP, SignalP and related tools, *Nat. Protoc.*, 2: 953-971
- Emanuelsson O., Nielsen H., Brunak S., and Von Heijne G., 2000, Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, *J. mol. Biol.*, 300(4): 1005-1016
- Goudenège D., Avner S., Lucchetti-Miganeh C., and Barloy-Hubler F., 2010, CoBaltDB: Complete bacterial and archaeal orfomes subcellular localization database and associated resources, *BMC microbiol.*, 10: 88
- Heazlewood J.L., Tonti-Filippini J., Verboom R.E., and Millar A.H., 2005, Combining experimental and predicted datasets for determination of the subcellular location of proteins in Arabidopsis, *Plant Physiol.*, 139(2): 598-609
- Hiller K., Grote A., Scheer M., Münch R., and Jahn D., 2004, PrediSi: prediction of signal peptides and their cleavage positions, *Nucleic Acids Res.*, 32(Web Server issue): W375-379
- Horton P., Park K.J., Obayashi T., Fujita N., Harada H., Adams-Collier C.J., and Nakai K., 2007, WoLF PSORT: protein localization predictor, *Nucleic acids res.*, 35(Web Server issue): W585-587
- Huang W.L., Tung C.W., Ho S.W., Hwang S.-F., and Ho S.Y., 2008, ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization, *BMC bioinformatics*, 9: 80
- Käll L., Krogh A., and Sonnhammer E.L., 2004, A combined transmembrane topology and signal peptide prediction method, *J. Mol. Biol.*, 338: 1027-1036
- Käll L., Krogh A., and Sonnhammer E.L.L., 2007, Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server, *Nucleic acids res.*, 35(Web Server issue): W429-432
- Kaundal R., and Raghava G.P.S., 2009, RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information, *Proteomics*, 9(9): 2324-2342
- Krogh A., Larsson B., von Heijne G., and Sonnhammer E.L., 2001, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J. mol. Biol.*, 305(3): 567-580
- Lin H.N., Chen C.T., Sung T.Y., Ho S.Y., and Hsu W.L., 2009, Protein subcellular localization prediction of eukaryotes using a knowledge-based approach, *BMC bioinformatics*, 10(Suppl 15): S8
- Lum G., and Min X.J., 2011, FunSecKB: the fungal secretome knowledgebase, *Database-the Journal of Biological Databases and Curation*, Vol. 2011, doi: 10.1093/database/bar001
- Makridakis M., and Vlahou A., 2010, Secretome proteomics for discovery of cancer biomarkers, *J. proteomics*, 73(12): 2291-2305
- Matthews B.W., 1975, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim. Biophys. Acta*, 405: 442-451
- Min X.J., 2010, Evaluation of computational methods for secreted protein prediction in different eukaryotes, *J. Proteomics Bioinform*, 3: 143-147
- Mooney C., Wang Y.H., and Pollastri G., 2011, SCLpred: protein subcellular localization prediction by N-to-1 neural networks, *Bioinformatics*, 27(20): 2812-2819
- Nakai K., and Horton P., 2007, Computational prediction of subcellular localization, *Methods Mol. Biol.*, 390: 429-466
- Petersen T.N., Brunak S., von Heijne G., and Nielsen H., 2011, SignalP 4.0: discriminating signal peptides from transmembrane regions, *Nature methods*, 8(10): 785-786
- Ryngajllo M., Childs L., Lohse M., Giorgi F.M., Lude A., Selbig J., and Usadel B., 2011, SLocX: predicting subcellular localization of Arabidopsis proteins leveraging gene expression data, *Frontiers plant sci.*, 2: 43
- Von Heijne G., 1990, The signal peptide, *J. Membr. Biol.*, 115: 195-201
- Yu N.Y., Wagner J.R., Laird M.R., Melli G., Rey S., Lo R., Dao P., Sahinalp S.C., Ester M., Foster L.J., and Brinkman F.S.L., 2010, PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, *Bioinformatics*, 26(13): 1608-1615