PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2007, 34(6): 573~579
www.pibb.ac.cn

蛋白质亚细胞定位的生物信息学研究 *

张 松 黄 波 夏学峰 孙之荣 **
(清华大学生物信息与系统生物学研究所，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 细胞中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. 蛋白质的亚细胞定位是蛋白质功能研究的重要方面，也是生物信息学中的热点问题，数据库的构建和亚细胞定位分析及预测加速了蛋白质结构和功能的研究.

关键词 亚细胞定位，生物信息学，数据库，预测
学科分类号 Q6

随着人类基因组计划的实施和推进，生命科学研究已进入了后基因组时代，研究的重点已经集中到功能基因组学上. 蛋白质的亚细胞定位是蛋白质组学研究的重要信息，也是蛋白质功能研究的重要方面. 目前更是提出了定位组(localizome)来大规模研究蛋白质的亚细胞定位.

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，如细胞核、高尔基体、内质网、线粒体、胞浆和细胞膜等. 蛋白质在核糖体中合成后经蛋白质分选信号引导后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. 另外，蛋白质在细胞里不是静止不动的，它们在细胞里常常通过在不同亚细胞环境里的运动发挥作用. 例如细胞周期的调控过程、细胞的信号转导和转录调控，都依赖于蛋白质空间位置的变化和运动. 成熟蛋白质必须在特定的细胞部位才能发挥其生物学功能，细胞内的区域化分布能够影响到蛋白质折叠、聚合以及转录后修饰的过程，对细胞功能产生深远的影响. 了解蛋白质的亚细胞定位信息，可以为我们推断蛋白质的生物学功能提供必要的帮助，同时对蛋白质的其他研究如相互作用、进化等也能提供必要的信息. 反过来，对同一亚细胞区域的蛋白质功能的研究也有利于更为深刻地理解该亚细胞结构.

蛋白质亚细胞定位信息的日渐重要，除了传统的亚细胞分离技术外，融合绿色荧光蛋白[1]、质谱[2]和同位素亲和标签[3]等实验技术提供了一些比较精确的亚细胞定位数据. 但是，这些技术多是昂贵、耗时的，并且重复性也比较差，目前来看，单纯依靠这些实验技术来研究是不现实的. 近年来，生物信息学在这方面开展了广泛的研究并且取得一系列很有意义的成果，数据库的构建和亚细胞定位分析及预测加速了蛋白质结构和功能的研究. 一方面，生物信息学研究可以对大规模的实验数据进行分析和提取生物学信息，同时可以根据现有数据对一些目前还未知的蛋白质做出预测；另一方面，不断增长的亚细胞定位数据也可以用来验证并改进预测结果.

1 亚细胞相关的数据库

除了一些综合数据库(SWISS-PROT[4]、MIPS[5]等)和模式生物数据库收录有部分蛋白质亚细胞信息，目前出现了一些专门的亚细胞定位数据库，如LOCATE[6]、DBSubloc[7]等. 这些数据库的构建主要基于计算机预测、大规模实验和文献挖掘技术. LOCATE 是一个小鼠蛋白质亚细胞定位数据库，

*国家自然科学基金(90408019)和国家重点基础研究发展计划(973)(2003CB715900)资助项目.
** 通讯联系人.
Tel: 010-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn
收稿日期：2006-11-21，接受日期：2007-01-28

主要是高通量芯片数据和 1 700 多篇文献挖掘的亚细胞信息，同时在该数据库中还有些哺乳动物的Ⅱ型跨膜蛋白的亚细胞定位数据[8]. 最近 Kislinger 等[9]又基于质谱注释了 4 768 个小鼠蛋白质的亚细胞定位，其中 3 274 个是高可靠性的. Kumar 等[10]收录了 2 744 个实验所得的啤酒酵母蛋白亚细胞定位数据，2006 年 Li 等[11]分析了拟南芥的 1 300 个蛋白质的亚细胞定位并构建了数据库. 还有一些针对单个细胞器的数据库，如 MitoProteome[12]等. 另外一些亚细胞定位数据库则主要收录预测数据或者同时收录经过实验验证和计算预测的数据. LOCtarget[13]和 LOC3D[14]都是基于计算机预测的亚细胞数据库. PSORTdb[15]则同时收录了经过实验验证和计算预测的亚细胞定位信息. DBSubloc[7]中的亚细胞定位注释主要来自综合数据库、模式生物基因组和文献挖掘，收录了超过 60 000 条蛋白质记录. Organelle DB[16]是一个真核定位数据库，其亚细胞分类名称采用的是 GO 中的注释标准. PA- GOSUB[17]收录了 10 种模式生物 107 000 多个蛋白质，并预测了每个蛋白质的生物学功能和亚细胞定位. 表 1 中列出了一些专门的亚细胞定位数据库及网址. 目前各种蛋白质数据库中蛋白质的亚细胞定位信息还不是很完善，相信随着一系列定位实验技术的出现和成熟、计算机预测精度的提高，更多的高质量亚细胞定位数据库会出现，为我们提供更好的数据.

Table 1 Databases of subcellular localization

表 1 亚细胞定位数据库

数据库	网址
LOCATE[6, 8]	http://locate.imb.uq.edu.au
小鼠亚细胞数据库[9]	http://tap.med.utoronto.ca/~mts/
酵母亚细胞数据库[10]	http://ygac.med.yale.edu
拟南芥亚细胞数据库[11]	http://aztec.stanford.edu/gfp/
MitoProteome[12]	http://www. mitoproteome.org/
LOCtarget[13]	http://www.rostlab.org/services/LOCtarget/
LOC3D[14]	http://cubic.bioc.columbia.edu/db/LOC3d/
PSORTdb[15]	http://db.psort.org/
DBSubloc[7]	http://www.bioinfo.tsinghua.edu.cn
Organelle DB[16]	http://organelledb.lsi.umich.edu
PA-GOSUB[17]	http://www.cs.ualberta.ca/~bioinfo/PA/GOSUB

2 亚细胞定位分析

对于日渐增长的亚细胞数据，数据的分析显得越来越重要，从中找到亚细胞定位的生物学规律并确定蛋白质功能才是我们真正关心的问题. 分析和亚细胞定位相关的蛋白质序列特征可以为计算预测提供相关特征信息，是亚细胞定位预测的基础，到目前为止出现了很多分析工具. Nair 等[18]基于序列比对等分析了各亚细胞器内的序列保守性，Cokol 等[19]从文献中收集了 91 个经过实验验证的核定位信号(NLSs)片段，对核内蛋白质进行了分析，此外，蛋白质的跨膜 α 螺旋[20]、Ⅰ型信号肽[21]和 β 桶结 构[22]等都进行了分析.

对于亚细胞定位的分析并不仅仅局限于其序列特征，Huh 等[1]将细胞中亚细胞器分为 22 类，然后对裂殖酵母中的蛋白质进行亚细胞定位. 他们结合了转录、遗传和蛋白质相互作用的数据，从亚细胞器层面上分析了转录调控等细胞功能.

2006 年德国马普研究所的 Foster 等[23]开发出了蛋白质相互关系分析法(PCP)将小鼠肝脏中 1 404 个蛋白质定位到了 10 个亚细胞器，进行了全面系统的分析. 他们研究发现 39%的蛋白质可以定位到 2 个或 2 个以上的亚细胞器中. 事实上，有些蛋白质在细胞内各个亚细胞器中是动态存在的，不同的亚细胞中产生不同的细胞功能. 另外，在信号转导过程中，信号通路因子、转录因子等可能在不同的时间出现在不同的细胞器内. 这种时间和空间上的跨度使得大量的蛋白质可能具有多亚细胞定位，这

些蛋白质对于我们理解细胞器的相关作用、蛋白质调控过程和信号通路等具有重要作用，目前对于这些蛋白质的研究还比较少.

3 亚细胞定位预测

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，已经出现了很多种预测工具，预测精度不断提高，取得了大量的研究成果. 一般来说，亚细胞定位预测的过程包括如下几个步骤：a. 数据集的建立，抽取出一个高质量的亚细胞定位数据集并分为训练集和测试集；b. 从这些蛋白质数据中抽取出特征信息向量；c. 选择合适的算法，根据前面的特征信息向量作出预测；d. 用检验数据集对预测结果进行评价. 总的来说，这些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出之后转化成高维的特征向量，作为预测的输入. 第二，算法的设计与实现，根据提取的特征向量集，利用有效的算法预测蛋白质的亚细胞定位. 现有预测算法中，统计学和机器学习方法使用的最为广泛. 算法是影响亚细胞预测精度的重要因素.

3.1 蛋白质信息的提取

蛋白质信息的提取是亚细胞定位预测的基础，体现了亚细胞定位的生物学内涵. 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，包括分选信号、序列、结构域特征和残基的理化性质等等，目前所采用的特征向量的提取基本上都是基于某一特征或几个特征的综合. 根据各类方法中抽取特征信息的不同，大致分为以下 4 种：蛋白质分选信号，氨基酸性质与组成，其他特征信息，以及几种特征信息的组合.

合成的蛋白质必须要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. N 端分选信号包括信号肽、线粒体引导肽、叶绿体运输肽和核定位信号等. Nakai 等[24]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开始关注. Emanuelsson 等[25]开发出了 ChloroP 来预测叶绿体运输肽， SignalP[21]则是专门用来识别信号肽. Nakai 小组[26]在 1999 年又发表了亚细胞定位预测软件 PSORT. TargetP[27]则是利用蛋白质的 N 端序列来进行亚细胞定位预测. 这种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，理论上预测精确度比较高. 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，效果并不是很好.

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. 蛋白质一般由 20 种氨基酸组成，氨基酸组成将 20 种氨基酸在蛋白质序列中出现的频率抽取出来作为一个 20 维的向量. 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是 Nakashima 等[28]，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. Cedano 等[29]的 ProtLock 预测算法将蛋白质亚细胞器扩大为 5 类. Reinhardt 等[30]又基于人工神经网络进一步对 4 种真核生物和 3 种原核生物蛋白质进行了分类预测，其后 Yuan[31]和 Hua[32]等根据氨基酸组成分别基于不同的算法来进行蛋白质亚细胞预测，也取得不错的预测效果.

氨基酸组成特征信息虽然提取比较方便，但是没有考虑到序列的顺序以及氨基酸残基，于是很多算法开始改进，形成了很多衍生算法. Fujiwara 等[33]采用了氨基酸组成和序列的残基顺序相结合的方法，并用不同的算法来处理. Park 等[34]用多种氨基酸序列作为特征信息，对 12 类亚细胞器进行了定位. 除了氨基酸序列之外，氨基酸的一些物理化学性质也被整合到序列或者氨基酸组成中来了. 以 Chou 小组[35]为主，他们根据氨基酸之间的物理化学距离提出了准序列顺序(Quasi-sequence-order)，并且他们将氨基酸的亲水性、疏水性等物理化学性质整合到氨基酸组成中，定义为伪氨基酸组成(pseudo-amino acid composition).

除了这些常用的蛋白质特征信息之外，很多人开始关心并引入一些其他特征信息，如功能域组成、结构、GO 注释等等. Chou 等[36]将功能域信息作为特征向量来进行亚细胞定位预测，随后他们又结合了 GO 注释进行预测并取得了不错的效果. Marcotte 等[37]采用了同源蛋白的系统发育信息. Nair 等[13]充分利用了 PDB 的结构信息和多序列比对的进化信息. Scott 等[38]则基于蛋白质的模体(motif)信息发展出了 PSLT. 除此以外，文本信息[39]等也作为特征信息来提高亚细胞定位预测性能.

将多种特征向量结合起来已经成为亚细胞定位

预测中最为普遍的一种方法. 蛋白质在细胞内的定位并不只是和某一种特征信息相关，大家发现，单纯依靠某一种特征向量进行预测已经很难再有突破. 这种混合提取特征信息来建模的方法使得信息的输入更加完备，显著地提高了预测能力，并且使得我们更好地理解蛋白质的亚细胞定位与其序列、结构、物化性质和功能之间的关系. Gardy 等[40]提出的 PSORT-B 将氨基酸组成、N 端分选信号和模体信息等一起作为特征信息来预测革兰氏阴性菌蛋白的细胞定位. Bhasin 等[41]将二肽组成和 PSI-BLAST 的输出信息一起构建了一个 458 维的向量作为输入来预测真核蛋白. 近年来，这种混合信息的提取已经逐渐成为了一种趋势，如 MultiLoc [42] 和 GNBSL[43]等等. 表 2 列出了一些常用预测服务器及网址.

Table 2 Frequently-used subcellular localization prediction online servers

表 2 常用的亚细胞定位预测在线服务

数据库	网址
ChloroP[25]	http://www.cbs.dtu.dk/services/ChloroP/
SignalP[21]	http://www.cbs.dtu.dk/services/SignalP/
NNPSL[30]	http://predict.sanger.ac.uk/nnpsl
PSORT[26]	http://psort.ims.u-tokyo.ac.jp/
PSORT-B[40]	http://www.psort.org/psortb/
TargetP[27]	http://www.cbs.dtu.dk/services/TargetP/
SubLoc[32]	http://www.bioinfo.tsinghua.edu.cn/SubLoc/
PLOC[34]	http://www.genome.ad.jp/SIT/ploc.html
PSLT[38]	www.mcb.mcgill.ca/~hera/PSLT
ESLpred[41]	http://www.imtech.res.in/raghava/eslpred/
MultiLoc[42]	http://www-bs.informatik.uni-tuebingen.de/Services/MultiLoc/
GNBSL[43]	http://166.111.24.5/webtools/GNBSL/index.htm
Cello[44]	http://cello.life.nctu.edu.tw/
Proteome Analyst[45]	http://www.cs.ualberta.ca/%7Ebioinfo/PA/Sub/index.html

3.2 预测算法

算法的设计与实现是亚细胞定位预测中最重要的一个步骤，也是生物信息学研究的有力工具. Nakai 等[24]最先使用“if-then”规则构建了一个专家系统来进行预测，Cedano 等[29]对蛋白质的细胞定位和氨基酸组成做了相关性分析，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，机器学习方法的基本思想是根据已有生物数据发现有意义的生物学知识或者规律，通过推理、模型匹配或样本学习从中自动学习知识和理论，然后利用这些规律去对未知数据进行预测. 最近邻法、神经网络、隐 Markov 模型、支持向量机和贝叶斯网络等都是亚细胞定位预测中常用的机器学习算法.

神经网络是模拟人的神经元信息处理过程的一种算法，它具有很强的鲁棒性和容错性，可以学习和自适应不确定的系统，所以很早就被 Reinhardt 等[30]应用到亚细胞定位预测中来，随后很快得到了广泛的使用[25, 27].

支持向量机是一种基于统计学习理论的分类技术. 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类面，将样本分为两类，并且使得分类误差率最小. Hua 等[32]在 2001 年首先开始用支持向量机来进行亚细胞定位预测，在真核生物中总的预测精度为 79.4%，原核生物中达 91.4%，其出色的预测效果使其很快就成为使用得最为普遍的一种算法.

其他一些算法如最近邻法[46]、隐马尔可夫模型[21]、贝叶斯网络[38]等等都取得了不错的效果. 但是随着预测精度的不断提高，将多种算法结合起来进行预测逐渐成为目前亚细胞定位预测的趋势，用不同的算法处理不同的特征信息或者综合多种算法进行多级预测，都取得更高的预测精度. Fujiwara 等[33]用神经网络方法描述蛋白质序列的氨基酸组

成，用隐马尔可夫模型描述序列的残基序列，取得了在植物中 86%、在非植物中 91%的预测精度. GNBSL[43]是联合了概率神经网络和支持向量机，并采用了多级算法.

3.3 预测性能评估

对于亚细胞定位预测，在采用不同的特征信息和算法进行预测后，一个很重要的步骤就是要客观地评价一下其性能. 有些方法可能会采用一些自己的评价标准，但是基本上都是基于下面 4 个统计参数：真阳性数目(*TP*)、假阳性数目(*FP*)、真阴性数目(*TN*)和假阴性数目(*FN*). 对于训练集，最常用的方法是 *n* 轮交叉验证(*n*-fold cross-validation)，在该评估过程中，所得样本集被随机分为 *n* 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 *n*-1 个子集作为训练集，这样通过 *n* 轮测试后，取其平均值作为总的分类性能，大部分的预测方法都采用了这种评估方法.

作为 *n*-fold cross validation 的延伸，留一交叉验证(leave-one-out cross-validation，LOOCV)每次取出数据集中的一条蛋白质序列作为测试样本，而剩余的蛋白质序列作为训练集对测试样本的亚细胞进行定位预测，直到所有样本序列都被测试一遍为止. LOOCV 的缺点是计算成本高，比较费时间，但是其结果更加严格可靠，已经在很多方法中得到了应用[31, 32].

此外，Matthew 相关系数(Matthew correlation coefficient，*MCC*)，敏感性(*Sensitivity*)和特异性(*Specificity*)等都是常用的几个性能指标.

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

$$Sensitivity=\frac{TP}{TP+FN}$$

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

TP：真阳性数目，*TN*：真阴性数目，*FP*：假阳性数目，*FN*：假阴性数目.

4 展望与结语

亚细胞定位的生物信息学研究作为亚细胞蛋白质组学实验研究的补充，相互验证与促进. 经过这么多年的发展，取得了很大的成果，甚至有些预测精度已经超过了目前的有些实验技术[47]. 但是从生物学的角度来看，还是有一些方面需要继续研究：a. 目前各个数据库中的亚细胞定位注释并不是非常统一，给大规模分析预测带来一定的困难. b. 对分选信号的理解还不是很透彻，需要更进一步地理解其生物学意义. c. 有些蛋白质在细胞内并不是固定在某一个亚细胞器内，如有些转录因子等，它们在细胞内具有更加重要的作用，并且数量不少[23]，目前对这类蛋白质研究还比较少. d. 对于蛋白质功能与亚细胞定位之间的关系理解还不够深入. 蛋白质亚细胞定位研究是生物信息学中的一个热点问题，其最终目的还是为了更好地理解蛋白质的功能和生物学意义，随着细胞蛋白质组学的发展，将会迎来更多的挑战并取得更大的成就.

参 考 文 献

1 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

2 Dunkley T P, Watson R, Griffin J L, *et al*. Localization of organelle proteins by isotope tagging (LOPIT). Mol Cell Proteomics, 2004, **3** (11): 1128～1134

3 Jiang X S, Dai J, Sheng Q H, *et al*. A comparative proteomic strategy for subcellular proteome research: ICAT approach coupled with bioinformatics prediction to ascertain rat liver mitochondrial proteins and indication of mitochondrial localization for catalase. Mol Cell Proteomics, 2005, **4** (1): 12～34

4 Boeckmann B, Bairoch A, Apweiler R, *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res, 2003, **31** (1): 365～370

5 Mewes H W, Albermann K, Heumann K, *et al*. MIPS: a database for protein sequences, homology data and yeast genome information. Nucleic Acids Res, 1997, **25** (1): 28～30

6 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

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

8 Aturaliya R N, Fink J L, Davis M J, *et al*. Subcellular localization of mammalian type II membrane proteins. Traffic, 2006, **7** (5): 613～625

9 Kislinger T, Cox B, Kannan A, *et al*. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell, 2006, **125** (1): 173～186

10 Kumar A, Agarwal S, Heyman J A, *et al*. Subcellular localization of the yeast proteome. Genes Dev, 2002, **16** (6): 707～719

11 Li S, Ehrhardt D W, Rhee S Y. Systematic analysis of Arabidopsis organelles and a protein localization database for facilitating fluorescent tagging of full-length Arabidopsis proteins. Plant Physiol, 2006, **141** (2): 527～539

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 Nair R, Rost B. LOCnet and LOCtarget: sub-cellular localization for

structural genomics targets. Nucleic Acids Res, 2004, **32** (Web server issue): W517～521

14 Nair R, Rost B. LOC3D: annotate sub-cellular localization for protein structures. Nucleic Acids Res, 2003, **31** (13): 3337～3340

15 Rey S, Acab M, Gardy J L, *et al*. PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res, 2005, **33** (Database issue): D164～168,

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

17 Lu P, Szafron D, Greiner R, *et al*. PA-GOSUB: a searchable database of model organism protein sequences with their predicted gene ontology molecular function and subcellular localization. Nucleic Acids Res, 2005, **33** (Database issue): D147～153

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

19 Cokol M, Nair R, Rost B. Finding nuclear localization signals. EMBO Rep, 2000, **1** (5): 411～415

20 Krogh A, Larsson B, von Heijne G, *et al*. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol, 2001, **305** (3): 567～580

21 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

22 Bigelow H R, Petrey D S, Liu J, *et al*. Predicting transmembrane beta-barrels in proteomes. Nucleic Acids Res, 2004, **32** (8): 2566～2577

23 Foster L J, de Hoog C L, Zhang Y, *et al*. A mammalian organelle map by protein correlation profiling. Cell, 2006, **125** (1): 187～199

24 Nakai K, Kanehisa M. Expert system for predicting protein localization sites in gram-negative bacteria. Proteins, 1991, **11** (2): 95～110

25 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

26 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

27 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

28 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

29 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

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

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

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

33 Fujiwara Y, Asogawa M. Prediction of subcellular localizations using amino acid composition and order. Genome Inform, 2001, **12**: 103～112

34 Park K J, Kanehisa M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics, 2003, **19** (13): 1656～1663

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

36 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

37 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

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

39 Nair R, Rost B. Inferring sub-cellular localization through automated lexical analysis. Bioinformatics, 2002, **18** (Suppl 1): S78～86

40 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

41 Bhasin M, Raghava G P. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res, 2004, **32** (Web server issue): W414～419

42 Hoglund A, Donnes P, Blum T, *et al*. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics, 2006, **22** (10): 1158～1165

43 Guo J, Lin Y, Liu X. GNBSL: A new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics, 2006, **6** (19): 5099～5105

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. Protein Sci, 2004, **13** (5): 1402～1406

45 Lu Z, Szafron D, Greiner R, *et al*. Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics, 2004, **20** (4): 547～556

46 Huang Y, Li Y. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics, 2004, **20** (1): 21～28

47 Rey S, Gardy J L, Brinkman F S. Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. BMC Genomics, 2005, **6**: 162

Bioinformatics Research in Subcellular Localization of Protein*

ZHANG Song, HUANG Bo, XIA Xue-Feng, SUN Zhi-Rong**

(*Institute of Bioinformatics and System Biology, Ministry of Education Key Laboratory of Bioinformatics, Department of Biological Science and Biotechnology, Tsinghua University, Beijing* 100084, China)

Abstract Protein is transported to specific subcellular localization after it is synthesized in cells, which is crucial to its function. Inaccurate destination will have great impact on cellular function or even life. Protein subcellular localization, which is one of the important areas in protein function research, is the hot issue in bioinformatics. Databases, analyses and prediction of subcellular localization accelerate the research of protein structure and function.

Key words subcellular localization, bioinformatics, database, prediction

---

*This work was supported by grants from The National Natural Science Fundation of China (90408019) and National Basic Research Program of China (2003CB715900).

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

Received: November 21, 2006 Accepted: January 28, 2007