一种基于 GO 分子功能信息的序列编码辅助 预测亚细胞定位方法

沈金城 张松 孙之荣*

(清华大学生物科学与技术系, 教育部生物信息学重点实验室, 北京 100084. * 联系人, E-mail: sunzhr@mail.tsinghua.edu.cn)

摘要 亚细胞定位与蛋白质的功能紧密相关. 本研究提出了一种新的编码方式, 试图根据那些与亚细胞定位相关的蛋白质分子功能信息, 结合氨基酸二肽组成以提高亚细胞定位预测的性能. 结果显示, 在使用支持向量机和 RH 数据集的测试中, 该方法的预测精度有了显著的提高, 同时对于蛋白质序列 N 端的不准确性也具有一定的鲁棒性.

关键词 亚细胞定位 GO 分子功能 氨基酸对组成 支持向量机

亚细胞定位作为蛋白质的一项重要属性, 与蛋白质的生物学功能有着紧密的联系, 对蛋白质亚细胞定位的了解为其进一步的纯化及功能研究都提供了重要的依据. 然而, 相对目前基因组和蛋白质组学研究产生的海量资料以及迅速增长的序列数据, 用实验方法^[1]确定的蛋白质亚细胞定位注释数量远不能满足研究的需要. 鉴于需求与实验得到的定位注释之间的巨大空白, 基于或部分地基于序列信息的亚细胞定位预测方法近年来得到了飞速的发展, 一些方法取得了较好的预测结果^[2], 为功能基因组注释工作提供了一个重要的、可信赖的工具. 实际上, 近年来一些新建立的功能数据库都已开始正式地收录预测得到的亚细胞定位信息^[3,4]. 而且随着一些领域蛋白功能研究的不断深入, 出现了很多特定问题, 如凋亡蛋白^[5]、核蛋白亚定位^[6]、线粒体亚定位^[7]等的亚细胞分类预测, 使亚细胞定位预测成为了生物信息学与实验生物学结合的最为紧密的领域之一.

应用在这一领域中众多生物信息学方法的基本思路大致可分为两类: () 用蛋白质 N 端的氨基酸序列(即蛋白分选信号序列)进行分选^[8]. 这些方法往往对于蛋白质测序数据的完整性, 尤其是 N 端的完整性有很高的要求, 高通量测序数据在很多情况下是不能满足这一条件的. () 用全序列的氨基酸或 *n*-肽(*n*-氨基酸残基)组成的编码进行预测的方法. 相对来说, 这一类方法在鲁棒性(robustness)方面具有一定的优势, 而且作为近几年应用较多的编码方式, 也得到了较好的预测效果. 然而, 随着功能基因组学研究的

迅速发展, 对亚细胞定位预测的精度也提出了更高的要求, 与之相应也就出现了很多改进编码方式使更多关联信息得到有效利用的尝试, 如基于 PSI-BLAST 的进化信息^[9]、基于氨基酸的理化性质与序列顺序信息的拟氨基酸组成信息^[10]等, 都在一定程度上弥补了单纯使用氨基酸组成信息对于蛋白序列描述的不完整性, 但是这些性质与亚细胞定位的关系往往是比较间接的. 本研究把与亚细胞定位关系更为直接的蛋白质的功能信息编码应用于定位的预测中.

功能信息与亚细胞的定位有着非常紧密的联系. 例如, 如果知道某新蛋白的功能与氧化磷酸化有关, 那么在细胞中它就很可能是位于线粒体之内. 然而目前这方面的研究却非常有限, 由于现有的功能注释大多是使用自然语言描述的, 这就给用计算机识别及处理这些数据带来了困难. 最简单的例子就是两个完全不同的词汇, 可能都是描述同一种功能的, 而计算机就很难将这些完全不同的描述联系起来. 对于这样的问题, 使用一种人和计算机都容易处理的标准化的词汇和关键字表描述在目前不失为一种好的处理方法, 它允许程序对于相同意义注释之间至少在一定程度上直接比较. 基因本体论(gene ontology, GO)项目就是这方面较为公认的标准化数据源之一, 它提供了一个将功能信息进行批量化和自动化处理的平台. Chou 和 Cai^[11]就曾使用 GO 的描述系统将蛋白的某些特定的、基于结构预测的结构域信息加入到亚细胞定位预测中. 在参考了一些基于序

列的功能预测方法^[12]的基础上,本研究结合GO组织提供的分子功能注释系统,将基于序列的功能编码从某几种特征结构域拓展到建立更为直接的整个蛋白与功能信息之间的关系,以服务于定位预测.在公用数据集上的测试表明,与单独使用氨基酸组成信息的预测相比,该信息的加入使预测精度有较显著的提高,这也为该编码方式服务于其他的基于序列的分类预测问题提供了一种启示.

1 材料和方法

() 数据集的选取. 为了便于与以往预测算法的比较,本研究选择在一个亚细胞定位预测的公用标准数据集RH (Reinhardt and Hubbard)上进行测试.该数据集是Reinhardt 和 Hubbard 以SWISSPROT 33.0数据库为基础,筛选出其中有明确的亚细胞定位注释的蛋白质条目而建立的^[13].由于跨膜序列的特殊性,RH不包括跨膜螺旋区蛋白质的数据.关于此类蛋白质的定位,一些相关的文献^[14]已经有了比较准确的预测.同时,植物蛋白质也被从该数据集中剔除.最终的数据集包括了 997 个原核生物蛋白条目(含 668 条胞质定位(cytoplasmic)蛋白、107 条胞外定位(extracellular)蛋白及 202 条周质定位(periplasmic)蛋白)和 2427 个真核生物蛋白条目(包含 684 条胞质定位(cytoplasm)蛋白、325 条胞外定位(extracellular)蛋白、321 条线粒体定位(mitochondrial)蛋白以及 1097 条核定位(nuclear)蛋白).在这一领域中很多有重要意义的方法都曾经在该数据集上进行过测试,故可以将本研究的蛋白预测性能与其他的研究进行比较,以便对本研究进行评估.

() 编码信息的提取. 需说明的是,GO功能描述系统包括 3 个子类,即分子功能(molecular function)、生物学途径(biological process)和细胞组件(cellular component).本研究选择分子功能注释作为功能描述语言,而不是表面上看来和亚细胞定位关系更为直接的GO细胞组件注释,因为在GO的注释说明中已明确指出,其细胞组件注释有相当部分是人工根据分子功能注释推测得到的,分子功能注释在GO数据中往往来源更为直接和可靠^[15].

对于GO中的功能注释与亚细胞定位的关系,Chou和Cai等人^[16]的研究表明,GO中的功能信息确实有助于提高亚细胞定位预测的精度.在此基础上,Lu和Hunter^[15]进一步应用统计学方法证明了GO中的分子功能注释中确实部分地包含了亚细胞定位的

信息,他们还根据计算量化地给出了IG (information gain)值,用来描述关于某一性质的知识(这里即分子功能注释)对了解另一性质(这里即亚细胞定位信息)贡献的大小.对于一个确定的分子功能注释,IG值越大,说明确定蛋白是否具有该功能对于我们预测该蛋白的亚细胞定位帮助越大.本研究在亚细胞定位相关的GO分子功能注释中选取了IG值大于等于0.001 的共 194 个注释条目,作为我们的目标功能,表1列出了我们所选取的所有目标功能及其相应的IG值.

对于一个目标序列,首先使用BLASTPGP程序(<http://www.ncbi.nlm.nih.gov/blast/>)在GO注释蛋白数据库(共 52509 条序列)(版本号 go_200604_seqdb,参见<http://www.godatabase.org/dev>)中进行同源比对,在得到的与目标序列显著性同源且具有GO分子功能注释的序列子集中(E 值(E -value)阈值为10.0),按照我们的194个目标功能以及同源性优先的原则搜索在具有该功能或该功能子功能的同源序列中同源性最高的序列(我们称之为关系序列),然后根据这个关系序列,对目标蛋白与目标功能的关系进行编码,用四维向量描述每一个目标蛋白对目标功能的一对一关系:

(1) 比特分值(bit score). 比对中得到的目标序列与关系序列同源性的比特分值,表示了目标序列与关系序列的同源性.为了便于向量的归一化,这里使用比特分值的常用对数值.

(2) E 值. 比对中得到的目标序列与关系序列同源性的 E 值.为了便于向量的归一化,这里使用 E 值的负常用对数值.

(3) 深度值. 关系序列GO分子功能注释中与目标功能最接近的注释在GO分子功能注释系统结构中的位置相差的层数,表征BLAST得到的同源序列与目标功能之间的关系.这里的“最接近”的含义是:BLAST得到的关系序列的功能注释与目标功能相同,此时深度值为0;或者关系序列的GO分子功能注释中有一个或多个是目标功能的子功能,此时深度值即为该子功能在GO注释系统结构中所处的层数值与目标功能注释的层数值的差值.如果关系序列包含多个目标功能的子功能,或者同一对注释在GO层次结构中同时位于多条路径上,并因此同时拥有多个深度值,则在得到的几个深度值中取其最小值.需要注意的是,由于GO注释系统的层次结构

并非标准的单向无环图，所以这里所用到的两个层数值是特指使这两个功能注释处在层次结构的同一分支上，且关系序列的功能注释是目标功能注释的一个子类的途径(图 1)。对于这一途径的确认借助 GO 组织提供的 GO-Perl 工具包(<http://www.godatabase.org/dev/pod/go-perl.html#authors>)进行。

表 1 选取用来预测亚细胞定位的目标GO功能注释及其IG值^[15]

IG 值	GO 分子功能注释号
0.047	GO:0003677
0.024	GO:0005179; GO:0003676
0.022	GO:0003700
0.016	GO:0008270
0.015	GO:0004129; GO:0003735
0.013	GO:0008009
0.011	GO:0008083
0.01	GO:0016491
0.008	GO:0008121; GO:0005524
0.007	GO:0005488
0.006	GO:0008248; GO:0004872; GO:0003824
0.005	GO:0008373; GO:0008026; GO:0005509; GO:0005215; GO:0005125; GO:0004386; GO:0004222; GO:0003779
0.004	GO:0008408; GO:0008200; GO:0005489; GO:0004879; GO:0004867; GO:0004497; GO:0003887; GO:0003707
0.003	GO:0015036; GO:0008308; GO:0005515; GO:0005386; GO:0005198; GO:0004721; GO:0004295; GO:0004263; GO:0004197; GO:0003931; GO:0003930; GO:0003928; GO:0003925; GO:0003774; GO:0003723; GO:0000166
0.002	GO:0046872; GO:0018445; GO:0016874; GO:0015085; GO:0008417; GO:0008415; GO:0008375; GO:0008312; GO:0008237; GO:0008107; GO:0005550; GO:0005529; GO:0005525; GO:0005506; GO:0005388; GO:0005351; GO:0005319; GO:0005262; GO:0005261; GO:0005216; GO:0005164; GO:0005135; GO:0004812; GO:0004725; GO:0004713; GO:0004674; GO:0004672; GO:0004601; GO:0004451; GO:0004377; GO:0004348; GO:0004289; GO:0004250; GO:0004096; GO:0003997; GO:0003899; GO:0003895; GO:0003777; GO:0003697; GO:0003682; GO:0001509; GO:0000104; GO:0000062
0.001	GO:0045735; GO:0030598; GO:0030528; GO:0030492; GO:0020037; GO:0016984; GO:0016851; GO:0016836; GO:0016829; GO:0016787; GO:0016740; GO:0016615; GO:0016564; GO:0016228; GO:0015930; GO:0015450; GO:0009374; GO:0008964; GO:0008943; GO:0008757; GO:0008649; GO:0008565; GO:0008547; GO:0008478; GO:0008318; GO:0008289; GO:0008199; GO:0008168; GO:0008094; GO:0008092; GO:0005554; GO:0005549; GO:0005544; GO:0005540; GO:0005507; GO:0005496; GO:0005201; GO:0005200; GO:0005183; GO:0005173; GO:0005160; GO:0005144; GO:0005143; GO:0005140; GO:0005139; GO:0005138; GO:0005137; GO:0005136; GO:0005134; GO:0005133; GO:0005128; GO:0005096; GO:0005044; GO:0005006; GO:0005003; GO:0004888; GO:0004871; GO:0004859; GO:0004853; GO:0004840; GO:0004831; GO:0004829; GO:0004824; GO:0004815; GO:0004813; GO:0004807; GO:0004802; GO:0004801; GO:0004748; GO:0004743; GO:0004736; GO:0004714; GO:0004667; GO:0004655; GO:0004634; GO:0004623; GO:0004518; GO:0004516; GO:0004459; GO:0004364; GO:0004362; GO:0004350; GO:0004349; GO:0004347; GO:0004345; GO:0004332; GO:0004252; GO:0004198; GO:0004182; GO:0004177; GO:0004158; GO:0004152; GO:0004148; GO:0004104; GO:0004089; GO:0003994; GO:0003993; GO:0003990; GO:0003908; GO:0003862; GO:0003754; GO:0003746; GO:0003743

图 1 关系功能注释与目标功能注释的可能关系
“关系功能”即关系序列的功能注释。(a)表示关系功能与目标功能相同，(b)~(d)表示关系功能注释为目标功能注释的子功能

(4) 可靠性参量. GO 注释蛋白数据库中有对每一个序列的每一个功能注释来源的说明代码(表 2), 分别表示了得到相关注释的根据. 由于不同的来源会影响到注释的可靠性, 同时也就会影响到预测结果, 因此我们对于不同来源的注释作了如下处理: 首先, 不考虑来源为 ND 和 NR 的功能注释, 即如果在 BLAST 结果中得到的序列的分子功能被标记为 ND 或 NR, 则该序列不能成为关系序列. 对于其他来源的注释, 则在此维度中进行粗略的半定量描述, 这里定义

$$\text{可靠性参量} = \begin{cases} 1 & \text{IEA} \\ 0.5 & \text{非ND, NR, IEA来源序列} \end{cases}$$

由此, 对每一个目标序列我们可以得到一个对应的 776 维的功能表征向量.

表 2 GO 功能注释来源说明代码

缩写	含义
IC	由专家推断得出
IDA	由直接的实验得出
IEA	由电子注释方法推断得出
IEP	由表达模式推断得出
IGC	由其基因组中的位置信息推断得出
IGI	由遗传学上的相互作用信息推断得出
IMP	由突变体表型推断得出
IPI	由物理相互作用推断得出
ISS	由序列或结构相似性推断得出
NAS	由作者不明的文献推断得出
ND	无相关生物学证据
RCA	由核实过的计算分析结果得出
TAS	由作者信息明确的文献推断得出
NR	无记录

其次, 我们将相邻位置上不同氨基酸对的出现频率编码为 400 维的向量(即氨基酸对组成), 将它与功能编码合并, 这样对于每一个目标序列我们都将其编码为一个 $400 + 776 = 1176$ 维的向量.

() 分类器的选择. 为了便于和以往算法, 特别是那些单纯基于氨基酸组成的算法进行比较, 本研究使用基于氨基酸组成的亚细胞预测工作中应用

较为广泛的支持向量机算法作为分类器. 这是从线性可分情况下的最优分类面发展而来的一种比较新的机器学习算法, 近年来被广泛应用于各类模式识别问题中. 其基本思想是, 将样本变换到高维希尔伯特空间(Hilbert Space)以寻找所谓的最优分割超平面(optimal separating hyper-plane, OSH), 将不同类别的样本有效分开, 并且使训练样本中离最优分割超平面最近的样本点到该分割面的距离最大化. 支持向量机方法具有坚实的统计学习理论基础, 在实际应用中常常对于很多传统的机器学习方法显示出一定的优势, 这里我们使用由 Chih-Chung Chang 和 Chih-Jen Lin 开发的 LIBSVM 2.8.1 软件包作为分类器. 该软件可以在 <http://www.csie.ntu.edu.tw/~cjlin/libsvm/> 下载.

() 结果的评估与优化. 为了评价功能信息对于亚细胞定位预测的作用, 并将其与单纯利用氨基酸组成进行的预测进行比较, 本研究需要对于预测效果进行评估, 这里我们采用对于这一类问题应用比较广泛和权威的 5 维交叉验证(5-fold cross validation).

预测精度的评价指标采用目前这一领域中通用的若干指标参量, 包括总体预测精度(the overall prediction accuracy)、每一类的预测精度(the accuracy for each class)^[16] 及 Matthew 相关系数(Matthew's correlation coefficients, MCC)^[17].

为了达到较好的分类效果, 本研究利用 LIBSVM 2.8.1 软件包提供的多分类功能, 选用 RBF 核函数, 根据上述的评价指标对分类器参数进行格点式优化筛选, 其中参数 γ 的筛选范围为 $2^{-10} \sim 2^{10}$ 、公比为 2 的等比数列, 罚分 c 的范围为 $10^{-5} \sim 10^5$ 、公比为 10 的等比数列. 表 3 列出了优化后各分类器所选用的参数.

2 结果与讨论

表 4 列出了本研究方法在 RH 数据集上测试的结果, 同时表中也列出了一些比较有影响且在 RH 数据

表 3 SVM 分类器的参数优化

		γ	c	总精度
真核生物	全序列	0.03125	10000	85.8
	N 端切除 10 残基	0.015625	100	85.0
	N 端切除 20 残基	0.03125	10	84.5
	N 端切除 30 残基	0.015625	10000	84.2
	N 端切除 40 残基	0.03125	10	83.8
原核生物	全序列	https://engine.scichina.com/doi/10.1360/csb2007-52-13-1536		92.1

表 4 不同预测算法在 RH 数据集上的预测比较

亚细胞器		人工神经网络模型		马尔可夫模型		氨基酸对组成+SVM		氨基酸对组成+功能编码+SVM	
		预测精确度(%)	MCC	预测精确度(%)	MCC	预测精确度(%)	MCC	预测精确度(%)	MCC
原核生物数据集	细胞质	80	—	93.6	0.83	97.5	0.86	95.4	0.89
	周质	85	—	79.7	0.69	78.7	0.78	92.9	0.78
	细胞外	77	—	77.6	0.77	75.7	0.77	95.9	0.78
	总体精度	81	—	89.1	—	91.4	—	92.1	—
真核生物数据集	细胞质	55	—	78.1	0.60	76.9	0.64	88.6	0.72
	细胞外	75	—	62.2	0.63	80.0	0.78	96.3	0.85
	线粒体	61	—	69.2	0.53	56.7	0.58	92.6	0.66
	细胞核	72	—	74.1	0.68	87.4	0.75	94.1	0.88
	总体精度	66	—	73.0	—	79.4	—	85.8	—

表 5 氨基酸对组成与功能共编码方法在 N 端切除后 RH 真核生物数据集上的预测结果

N 端氨基酸切除数	总体精度	细胞质		细胞外		线粒体		细胞核	
	预测精确度(%)	预测精确度(%)	MCC	预测精确度(%)	MCC	预测精确度(%)	MCC	预测精确度(%)	MCC
0	85.8	88.6	0.72	96.3	0.85	92.6	0.66	94.1	0.88
10	85.0	87.8	0.71	95.9	0.83	91.6	0.61	94.5	0.89
20	84.5	87.6	0.70	95.5	0.81	92.0	0.63	93.8	0.88
30	84.2	87.4	0.70	96.2	0.84	91.6	0.60	93.3	0.87
40	83.8	86.8	0.69	96.2	0.83	91.1	0.58	93.4	0.87

集上取得了较好结果的工作, 包括人工神经网络^[13]、马尔可夫模型^[18]及使用全序列氨基酸组成的支持向量机^[19]的预测表现。

从表 4 可以看出, 对真核生物数据集, 与其他几种预测方法相比, 本文的预测结果无论是从总体的精度, 还是从各亚细胞器单独的预测精度来看都有了较大幅度的提高, 尤其是在与同样使用支持向量机作为分类器的Subloc^[19]算法的比较中, 这说明了我们的氨基酸对组成和功能编码相结合的编码方式, 比单独的氨基酸组成编码包含了更多的对于亚细胞定位有帮助的信息, 表明这种编码方式确实具有一定的优势。需要指出的是, 对于原核生物数据集, 本研究预测精度的提高相对来说不是很显著, 尤其是在和同类的Subloc工作比较中。在相关文献的检索中也没有找到近年来在该原核数据集上预测效果提高较大的文献, 推测可能与原核生物序列结构相对简单、对于基于序列的预测方法精度的上升空间有限有关。实际上, 近两年来的亚细胞预测工作已经把重点放在了如何提高真核蛋白的预测精度上^[16,9,20]。

对蛋白质全序列进行编码分类预测的方法应该对于序列的不完整性具有较好的鲁棒性, 为此我们模拟了在高通量测序过程中比较容易发生的 N 端序列不完整的情形, 来测试本方法在这方面的灵敏度。

对数据集中的每条序列都进行处理, 去掉其 N 端的前 n 个残基(这里分别进行了 $n = 10, 20, 30, 40$ 共 4 次实验), 然后使用切除 N 端后的数据集进行训练和预测。考虑上一节讨论的问题, 本次实验在真核生物数据集中进行, 其结果列于表 5 中。

通过比较可以看出, 即使在 N 端缺失 40 个残基的情况下, 预测精度的下降也不很明显(总精度只下降了 2.0%), 说明该算法对于序列的不完整性具有较好的鲁棒性。

以上结果初步表明, 相对于传统的单纯基于序列组成的编码方式, 本研究提出的氨基酸对组成与功能共编码, 可以较为有效地提高蛋白质亚细胞定位的预测精度。这种综合考虑序列信息、分子功能和亚细胞定位的关系的思想, 也为功能基因组研究, 特别是更多的基于序列的分类预测问题提供了有益的参考。

参 考 文 献

1 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. *Proc Int Conf Intell Syst Mol Biol*, 2000, 8: 251—259

2 Donnes P, Hoglund A. Predicting protein subcellular localization: Past, present, and future. *Genom Proteom Bioinformat*, 2004, 2(4): 209—215

doi/10.1360/scsb.2007-52-13-1536

- 3 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—D168[DOI]
- 4 Heazlewood J L, Verboom R E, Tonti-Filippini J, et al. SUBA: The *Arabidopsis* subcellular database. *Nucleic Acids Res*, 2007, 35(Database issue): D213—D218[DOI]
- 5 Zhou G P, Doctor K. Subcellular location prediction of apoptosis proteins. *Proteins*, 2003, 50(1): 44—48[DOI]
- 6 Lei Z, Dai Y. An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics*, 2005, 6: 291[DOI]
- 7 Du P, Li Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinformatics*, 2006, 7: 518[DOI]
- 8 Nakai K, Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 1992, 14(4): 897—911[DOI]
- 9 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—W419[DOI]
- 10 Chou K C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, 2001, 43(3): 246—255[DOI]
- 11 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[DOI]
- 12 Vinayagam A, del Val C, Schubert F, et al. GOPET: A tool for automated predictions of Gene Ontology terms. *BMC Bioinformatics*, 2006, 7: 161[DOI]
- 13 Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res*, 1998, 26(9): 2230—2236[DOI]
- 14 Rost B, Fariselli P, Casadio R. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci*, 1996, 5(8): 1704—1718
- 15 Lu Z, Hunter L. Go molecular function terms are predictive of subcellular localization. *Pac Symp Biocomput*, 2005, 10: 151—161
- 16 Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol*, 1993, 232: 584—599[DOI]
- 17 Matthews B W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, 1975, 405: 442—451
- 18 Yuan Z. Prediction of protein subcellular locations using Markov chain models. *FEBS Lett*, 1999, 451(1): 23—26[DOI]
- 19 Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 2001, 17(8): 721—728[DOI]
- 20 Guda C, Subramaniam S. pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics*, 2005, 21(21): 3963—3969[DOI]