• 研究论文 •

基于 VHSE 结构表征的蛋白酶体酶切位点预测及酶切特异性研究

谢江安[a] 梅 虎*,[a,b] 吕 娟[b] 潘显超[b] 王 青[b] 张亚兰[b]

([a]生物流变科学与技术教育部重点实验室 重庆大学 重庆 400044)

([b]重庆大学生物工程学院 重庆 400044)

摘要 泛素-蛋白酶体在真核生物的抗原呈递、细胞周期调控和转录因子激活等生理过程中发挥着极为重要的作用, 其核心就是蛋白酶体对底物的选择性酶切作用, 因此对选择性酶切位点的预测一直是计算生物学的一个重点研究内容. 针对现有酶切位点预测方法的非线性和物理意义不明确等问题, 借鉴定量构效关系研究方法, 采用基于氨基酸物理化学性质的描述子——VHSE (Principal component score vector of hydrophobic, steric, and electronic properties)对收集的 2650 个 MHC-I 配体的源蛋白序列进行了结构表征, 在此基础上利用支持向量机建立了蛋白酶体酶切位点的预测模型, 其最优线性模型的灵敏度(Sensitivity)、特异性(Specificity)、接受者操作特征曲线下面积(area under receiver operating characteristics curve, AUC)和马休斯相关系数(Matthews coefficient of correlation, MCC)分别为 90.18%, 69.63%, 0.8797 和 0.6131. 模型分析结果表明: 影响酶切位点选择性的氨基酸性质由大到小依次为: 疏水性、电性和立体特征; P9, P8, P4, P1, P3', P4'和 P5'位氨基酸对酶切位点的选择有重要影响, 研究亦显示酶切位点上游 P1 位和下游 P1'～P5'的"疏水势差"有利于蛋白酶体的切割作用.

关键词 蛋白酶体; MHC-I 配体; VHSE; 支持向量机; 酶切位点

Studies on the Prediction of Selective Cleavage Sites and Cleavage Profile of Proteasome Using VHSE Amino Acid Descriptor

Xie, Jiangan[a] Mei, Hu*,[a,b] Lü, Juan[b] Pan, Xianchao[b] Wang, Qing[b] Zhang, Yalan[b]

([a] *Key Laboratory of Biorheological Science and Technology* (*Ministry of Education*), *Chongqing University*, *Chongqing 400044*)

([b] *College of Bioengineering*, *Chongqing University*, *Chongqing 400044*)

Abstract The ubiquitin-proteasome system (UPS) plays a critical role in proteolysis and degradation in many physiological processes of the eukaryotes, such as antigen presentation, cell cycle regulation, and transcription factors activation. Recently, due to the importance of selective substrate cleavage of proteasome in the UPS, the cleavage site prediction has attracted considerable interest in computational biology. However, the existing methods are mostly based on nonlinear models with little physicochemical meanings. In this paper, VHSE (Principal component score vector of hydrophobic, steric, and electronic properties), a novel set of amino acid descriptors, was used to characterize the source proteins of 2650 natural MHC class I ligands. Based on the structural descriptions of the amino acids adjacent to the cleavage site, support vector machine (SVM) was then employed to establish the prediction models using linear and RBF kernel functions. A linear SVM model with high prediction capability was obtained, of which the sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and the Matthews correlation coefficient

* E-mail: meihu@cqu.edu.cn; Tel: 023-65112677

Received June 4, 2011; revised September 4, 2011; accepted October 26, 2011.

Project supported by Chongqing Key Natural Science Foundation (No. 2009BA5068), and the Fundamental Research Funds for the Central Universities (No. 11231177).

重庆市自然科学基金重点项目(CSTC, 2009BA5068)和重庆大学中央高校基本科研业务费科研专项(CDJXS, 11231177)资助项目.

(MCC) were 0.9018, 0.6963, 0.8797 and 0.6131, respectively. The results showed that the hydrophobic, electronic, and steric properties of the amino acids adjacent to the cleavage site are closely related to the selective substrate cleavage, especially for those at the positions of P9, P8, P4, P1, P3', P4', and P5'. The results also showed that hydrophobic potential difference between P1 position and P1'～P5' positions may benefit the cleavage process of the proteasome.

Keywords proteasome; MHC-I ligand; VHSE; SVM; cleavage site

泛素-蛋白酶体系统(Ubiquitins-Proteasome Systems, UPS)是真核细胞中具高效性和高度选择性的蛋白质降解体系, 其主要功能是降解半衰期较短和错误折叠的蛋白质, 维持机体正常的蛋白质代谢平衡, 同时参与细胞周期和基因表达等一系列重要生理过程的调控. 在哺乳动物中, UPS 另一个重要功能是酶切降解病原微生物的表位蛋白, 生成的多肽片段经抗原提呈途径呈递到细胞表面, 供细胞毒性 T 细胞(cytotoxicy T lymphocyte, CTL)识别, 进而激发特异性免疫反应[1].

常见的蛋白酶体是 26S 蛋白酶体, 它是由一个 20S 催化颗粒(catalytic particle, CP)和两个 19S 调节颗粒(regulatory particle, RP)组成的 ATP (adenosine triposphate)依赖性蛋白水解酶复合体[2]. 其酶切产物长度集中在 3～25 个残基范围内, 不同的生理过程以及不同的组织细胞内, 产物长度存在差异, 例如, 在抗原呈递过程中, 蛋白酶体主要酶切生成长度在 8～12 个残基的抗原肽. 通过蛋白酶体酶切作用主要产生抗原肽的 C 端, 而抗原肽 N-端通常在胞浆或内质网中形成[3].

由于蛋白酶体切割是生成抗原肽的一个关键环节, 因此对酶切位点的预测一直是计算免疫学一个重点研究内容. 同时通过酶切位点预测模型也是研究蛋白酶体选择性酶切机理的一个重要途径. 目前, 蛋白酶体酶切位点预测方法主要有 PAProC[4], MAPPP[5]和 NetChop[6]等.

PAProC 采用的是人和酵母的 20S 蛋白酶体的体外酶切数据, 通过随机的爬山算法(hillclimbing)确定每个氨基酸在不同位置对酶切的影响, 继而建立人工神经网络模型来预测蛋白酶体酶切位点; MAPPP 软件包中的 FragPredict 也是基于体外的酶切数据, 经统计分析得到蛋白酶体酶切基序后, 结合动力学模型来预测蛋白酶体酶切位点. 该方法首次将影响酶切的时间因素也考虑其中, 并可直接预测生成的多肽片段; NetChop 则采用 MHC-I 配体数据, 建立多层人工神经网络模型来预测蛋白酶体酶切位点, 这也是目前预测较为准确的一种方法. 然而, 由于蛋白酶体酶解机理的复杂性以及蛋白质结构的多样性, 酶切位点预测方法的准确性还有待提高. 同时由于目前大多方法采用了非线性的神经网络等模型, 这也给模型的解释带来了极大的困难.

针对上述情况, 本文基于 2650 个 MHC-I 配体及其源蛋白序列信息, 应用定量构效关系(Quantitative structural-activity relationship, QSAR)研究方法, 将酶切位点邻近氨基酸残基的结构特征应用到酶切位点的预测研究中, 采用支持向量机(Support Vector Machine, SVM)[7]建立了蛋白酶体酶切位点的线性和非线性预测模型, 模型的灵敏度(sensitivity)和特异性(specificity)均达到 90%和 70%的水平. 在此基础上, 通过线性模型深入分析了酶切位点邻近氨基酸残基的结构特征与选择性偏好. 本文模型不仅具有较高的预测准确性, 同时具可解释性. 该方法对蛋白酶体酶切位点预测以及选择性酶切机理研究均具有重要的理论意义和参考价值.

1 原理和方法

1.1 数据来源与处理

从 AntiJen 数据库[8]中收集整理得到 7324 个 MHC-I 配体, 涉及约 230 种 MHC-I 等位基因. 将上述配体与 SWISS-PROT 数据库[9]中源蛋白质序列进行关联后, 去除重复序列、无源蛋白序列和源蛋白登录号不可用的配体, 最后得到 3148 个配体及其源蛋白氨基酸序列. 以 MHC-I 配体羧基端(P1)为酶切位点[10], 左右扩展各 14 个氨基酸残基, 得到长度为 28 个氨基酸残基的酶切样本; 同时以配体中间位置为非酶切位点(P1), 左右各扩展 14 个氨基酸残基, 得到长度为 28 个氨基酸残基的阴性样本. 最终整理得到 2650 个酶切样本, 2522 个阴性样本. 本文中, 酶切位点两侧的氨基酸位置以(P14…P1 | P1'…P14')方式表示, 符号“|”表示酶切之处.

从 2650 个酶切样本中随机选取 1325 个样本, 从 2522 个阴性样本中随机选取 1261 个样本, 共计 2586 个样本组成训练集, 余下 2586 个样本组成预测集.

1.2 结构表征

从历年文献和数据库中, 我们共搜集整理到 20 种天然氨基酸的 200 余种物理化学性质. 根据疏水性、立体和电性特征将其分类和筛选后共得到 50 种性质, 其中疏水性质 18 个, 立体性质 17 个, 电性性质 15 个. 在

此基础上, 分别对这 3 类性质进行主成分分析, 对于 18 个疏水性质、17 个立体性质和 15 个电性性质, 前 2、2 和 4 个主成分分别累计解释原始数据矩阵 74.33%, 78.68%和 79.97%的方差. 据此有理由认为: 这 8 个主成分已经能够表征各自原始数据矩阵中绝大多数的信息, 所以用 8 个主成分得分矢量来替代原始数据并作为氨基酸结构描述子时, 原始数据的信息损失相对较少. 为了方便起见, 我们将该矢量称之为 VHSE (principal component score vector of hydrophobic, steric, and electronic properties)[11]. 其中 $VHSE_1$ 和 $VHSE_2$ 代表氨基酸的疏水性特征; $VHSE_3$ 和 $VHSE_4$ 代表氨基酸的立体特征; $VHSE_5$～$VHSE_8$ 则代表氨基酸的电性特征(表 1).

以一个二肽结构为例, 其第一个位点的氨基酸残基可由对应氨基酸的 8 个 VHSE 描述子表征, 同样第二个位点也由对应的 8 个 VHSE 表征, 这样产生的 16 个有序 VHSE 结构变量(2×8=16)可表征该二肽分子结构特征. 以此类推, 长度为 n 的肽分子可被 n×8 个描述变量所表征. 在本文中, 为了避免产生过多的变量, 仅选取疏水性、立体和电性性质的第一个主成分, 即 $VHSE_1$, $VHSE_3$ 和 $VHSE_5$ 用于样本的结构表征, 因此每个样本可用 84 (28×3=84)个 VHSE 结构参数进行表征.

1.3 SVM 建模

SVM[7,12,13]起初是用于解决线性可分情况下两类样本的分类问题, 其核心思想是找到一个最优分类超平面, 使两类样本的分类间隔(margin)最大化. 对于非线性问题, 首先经一个非线性映射 Φ, 将样本映射到一个高维特征空间, 然后用线性方法来解决. 高维映射经核函数: $K(x,x_i)=\Phi(x)\bullet\Phi(x_i)$实现. 因 SVM 引入核函数, 故可有效避免维数灾难及计算复杂性等问题. 目前常用核函数主要有:

线性核(linear kernel): $K(x,x_i)=x\bullet x_i$ (1)

多项式核(polynomial kernel):
$$K(x,x_i)=(\alpha_1 x\cdot x_i+\alpha_2)^p \quad (2)$$

径向基核(RBF(radial basis function) kernel):
$$K(x,x_i)=\exp(-\gamma\|x-x_i\|^2) \quad (3)$$

二次神经网络核(sigmoid kernel):
$$K(x,x_i)=\tanh(\alpha_1 x\cdot x_i+\alpha_2) \quad (4)$$

表 1 20 种氨基酸的 VHSE 描述子

Table 1 VHSE scales for 20 natural amino acids

AA		$VHSE_1$	$VHSE_2$	$VHSE_3$	$VHSE_4$	$VHSE_5$	$VHSE_6$	$VHSE_7$	$VHSE_8$
Ala	A	0.15	−1.11	−1.35	−0.92	0.02	−0.91	0.36	−0.48
Arg	R	−1.47	1.45	1.24	1.27	1.55	1.47	1.30	0.83
Asn	N	−0.99	0.00	−0.37	0.69	−0.55	0.85	0.73	−0.80
Asp	D	−1.15	0.67	−0.41	−0.01	−2.68	1.31	0.03	0.56
Cys	C	0.18	−1.67	−0.46	−0.21	0.00	1.20	−1.61	−0.19
Gln	Q	−0.96	0.12	0.18	0.16	0.09	0.42	−0.20	−0.41
Glu	E	−1.18	0.40	0.10	0.36	−2.16	−0.17	0.91	0.02
Gly	G	−0.20	−1.53	−2.63	2.28	−0.53	−1.18	2.01	−1.34
His	H	−0.43	−0.25	0.37	0.19	0.51	1.28	0.93	0.65
Ile	I	1.27	−0.14	0.30	−1.80	0.30	−1.61	−0.16	−0.13
Leu	L	1.36	0.07	0.26	−0.80	0.22	−1.37	0.08	−0.62
Lys	K	−1.17	0.70	0.70	0.80	1.64	0.67	1.63	0.13
Met	M	1.01	−0.53	0.43	0.00	0.23	0.10	−0.86	−0.68
Phe	F	1.52	0.61	0.96	−0.16	0.25	0.28	−1.33	−0.20
Pro	P	0.22	−0.17	−0.50	0.05	−0.01	−1.34	−0.19	3.56
Ser	S	−0.67	−0.86	−1.07	−0.41	−0.32	0.27	−0.64	0.11
Thr	T	−0.34	−0.51	−0.55	−1.06	0.01	−0.01	−0.79	0.39
Trp	W	1.50	2.06	1.79	0.75	0.75	−0.13	−1.06	−0.85
Tyr	Y	0.61	1.60	1.17	0.73	0.53	0.25	−0.96	−0.52
Val	V	0.76	−0.92	0.17	−1.91	0.22	−1.40	−0.24	−0.03

参数选择对 SVM 建模成败至关重要, 其中, 惩罚系数(punish coefficient) C 是一个调控参数, 用于调节最大间隔和最小化训练误差间的平衡; 不敏感损失函数(insensitive loss function)中 ε 主要控制不敏感带的宽度, 影响支持向量的多少; 参数 γ 则控制 RBF 核 SVM 的泛化处理能力[14].

根据相关研究结果和作者已有经验, 采用 RBF 核进行非线性建模的结果常常优于其它非线性核的结果, 因此在本文中, 我们仅以线性核和 RBF 核为代表来进行 SVM 分类建模研究. 酶切样本记作 1, 阴性样本记作−1, 根据 10 折交叉验证法(10-fold cross-validation)确定模型的 C, ε, γ 的最优值. 模型质量采用准确度(accuracy, Acc)、精确度(precision, Pre)、灵敏度(sensitivity, Sen)、特异性(specificity, Spe)、马休斯相关系数(Matthews coefficient of correlation, MCC)、接受者操作特征曲线下面积(area under receiver operating characteristics curve, AUC)、富集度(enrichment, ER)以及 F-measure 进行综合评价. 其中:

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TP}+\mathrm{TN}+\mathrm{FN}}\times 100\% \tag{5}$$

$$\mathrm{Sen}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\% \tag{6}$$

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times 100\% \tag{7}$$

$$\mathrm{Spe}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\times 100\% \tag{8}$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FN}\times\mathrm{FP}}{\sqrt{(\mathrm{TN}+\mathrm{FN})(\mathrm{FN}+\mathrm{TP})(\mathrm{TP}+\mathrm{FP})(\mathrm{FP}+\mathrm{TN})}} \tag{9}$$

$$\mathrm{ER}=\frac{\mathrm{TP}(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN})}{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})} \tag{10}$$

$$\text{F-measure}=\frac{2\mathrm{Pre}\times\mathrm{Sen}}{\mathrm{Pre}+\mathrm{Sen}} \tag{11}$$

上式中 TP 和 TN 分别代表测试集中正确预测的酶切样本数和阴性样本数; FP 和 FN 分别代表错误预测的酶切样本数和阴性样本数.

2　结果与分析

2.1　模型评价

SVM 共计产生了 10 个酶切位点预测模型(表 2), 其中 model 1～5 为线性核分类模型, model 6～10 为径向基核分类模型. 可以看出 10 个模型的 AUC 和 MCC 均大于 0.8 和 0.6, 且 Sen 和 Spe 也都达到 90%和 70%的水平, 表明模型具较强的预测能力. 其中, 线性核 SVM 以模型 4 为最优; 径向基核 SVM 以模型 9 为最优. 采用模型 4 和模型 9 分别对 2586 个预测集样本的预测结果参见表 3. 从外部预测结果看, 模型 4 和 9 亦具有较优的表现. 总的来看, 径向基核和线性核分类模型结果相当, 无显著差异. 考虑到线性模型具有更好的可解释性, 最终选择了模型 4 为最优模型, 其模型参数 C=11.3315, ε=0.2875.

Saxova 等[15]曾用 231 个 T 细胞表位和 MHC-I 配体对 PAProC, MAPPP 和 NetChop 3 种方法进行了预测并对预测结果做了比较研究. 在此我们也用本文模型对上述样本进行了预测, 结果如表 4 所示. 可以看出本文建立的 SVM-VHSE 模型预测结果明显优于其它 4 种模型. 除预测方法不同外, 本文模型也得益于较大的训练集样

表 2　支持向量机建模结果及模型参数

Table 2　The performance and parameters of SVM models

Model	Acc/%	Sen/%	Spe/%	MCC	AUC	Pre/%	ER	F-measure/%	C	ε	γ
1	79.85	89.13	70.11	0.6382	0.8781	76.31	1.4893	81.95	10.0000	0.2437	
2	79.70	89.36	69.55	0.6063	0.8743	76.01	1.4835	81.89	3.6788	0.1600	
3	79.81	91.47	67.66	0.6101	0.8803	75.10	1.4658	82.30	12.8403	0.3632	
4[a]	80.16	90.18	69.63	0.6131	0.8797	75.97	1.4828	82.32	11.3315	0.2875	
5	80.05	89.20	70.42	0.6087	0.8784	76.29	1.4890	82.07	10.6449	0.2388	
6	79.97	89.96	69.47	0.6090	0.8803	76.12	1.4857	82.17	10.0000	0.2610	0.5000
7	80.08	89.88	69.79	0.6109	0.8802	76.21	1.4873	82.25	3.6378	0.2675	1.3591
8	81.09	88.68	73.11	0.6270	0.8834	77.85	1.5193	82.76	21.1700	0.2385	1.0000
9[b]	81.71	90.95	72.00	0.6430	0.8889	77.67	1.5162	83.63	25.5359	0.3517	1.3591
10	80.97	89.36	72.16	0.6261	0.8840	77.34	1.5095	82.78	25.5359	0.3031	1.0000

[a] 线性核分类模型中的最优模型; [b] 径向基核分类模型中的最优模型.

表 3 模型 4 和模型 9 的外部预测性能表现

Table 3 The predictive performance of the model 4 and the model 9

Model	Acc/%	Sen/%	Spe/%	MCC	AUC	Pre/%	ER	F-measure/%
4	75.25	89.81	59.95	0.5236	0.8300	70.21	1.3702	78.81
9	73.98	90.72	56.38	0.5037	0.8310	68.61	1.3390	78.13

表 4 本文模型与其他预测模型的性能比较

Table 4 The predictive performances of this model compared with other models

Method	N	Sen/%	Spe/%	MCC
PAProC	217	45.6	30	−0.25
FragPredict	231	83.5	16.5	0
NetChop1.0	231	39.8	46.3	−0.14
NetChop2.0	231	73.6	42.4	0.16
SVM-VHSE	231	89.4	51.9	0.45

本数, 以NetChop2.0为例, 其建模所采用的MHC-I配体数为 1110 个, 而本文采用的 MHC-I 配体数为 1325 个.

2.2 蛋白酶体酶切特异性分析

蛋白酶体对底物的酶切处理不是随机的, 而是有一定模式和选择性的. Nussbaum 等[16]研究认为: 对 P1 位残基的酶切选择偏好与其上下游各 5 个残基有一定的相关性; Altuvai 等[10] 则认为底物序列每个位置上的残基对酶切的选择特异性都有相互独立、大小不一但可以累加的贡献. 基于本文的线性 SVM 预测模型, 我们就酶切位点“|”上下游各 14 个氨基酸残基的疏水、立体和静电特性进行了分析, 同时考察了上述特征对选择性酶切贡献的权重大小(图 1).

从图 1 可以看出, 对酶切位点选择性贡献的氨基酸性质由大到小依次是疏水性、电性和立体特征, 其中权重系数绝对值大于 2 的变量涉及以下位点: P9, P8, P4, P1, P3', P4'和 P5'.

由图 1 可知, P1 位氨基酸的疏水性($VHSE_1$)对酶切的影响最大. 当 P1 位为疏水性氨基酸如 Phe, Trp, Leu, Ile, Val 和 Tyr 等, 其可能更易被蛋白酶体所酶切. P1 位优先选择疏水性氨基酸意味着有利于生成 C 端为疏水性残基的酶切产物, 这也决定了最后产生的抗原肽 9 号位锚定残基的性质. 先前研究[16,17]发现: 蛋白酶切产物抗原肽C端(P1位)出现频率较高的氨基酸主要集中在疏水性氨基酸 Leu, Ile, Val, Thr 和 Ala; 且 Leu|Lys 可能是蛋白酶体酶切底物时优先选择的一个位点[18]. 通过比较酶切位点上下游氨基酸残基的疏水性权重, 我们发现 P1 位与 P1'至 P5'位存在大的“疏水势差”对酶切有利, 尤其是当 P5'为亲水性强的氨基酸时更利于蛋白酶体的酶切作用. 除疏水性外, P1 位氨基酸的电性特征和立体特

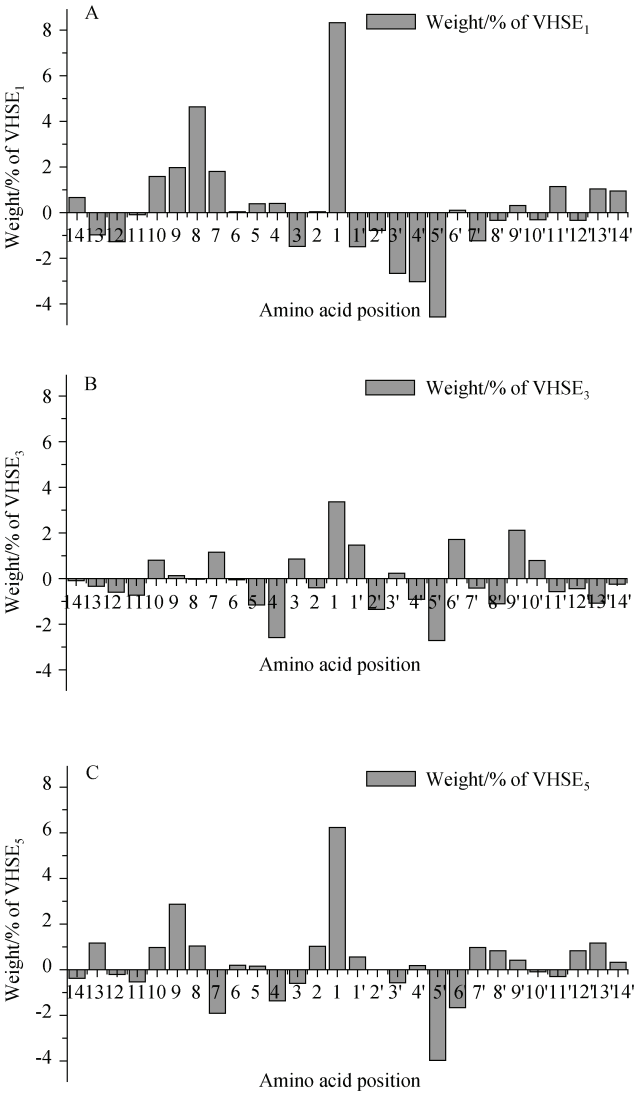

图 1 酶切位点“|”上下游氨基酸残基的物理特性对模型贡献的权重

A: 疏水特性; B: 立体特性; C: 静电特性

Figure 1 The weight coefficients of properties of amino acids adjacent to the cleavage site

A: hydrophobic property; B: steric property; C: electronic property

性对酶切也有很大的影响, 即正电性氨基酸如 Lys, Arg 和 His, 或大体积的氨基酸如 Arg, Phe 和 Trp 对酶切有利.

相比疏水和电性性质, P4 位的立体几何性质($VHSE_3$)对酶切的影响较大, 其权重值为−3.02, 表明该位不宜出现大体积的氨基酸残基, 如 Trp, Arg, Tyr,

Phe 和 Lys. Nussbaum 等[16]研究发现: 酵母原生菌 20S 蛋白酶体酶切底物时, P4 位出现 Pro 对酶切最有利. 由表 1 可知 Pro 的 $VHSE_3$ 值为－0.5, 空间位阻较小.

P8 和 P9 位对酶切的贡献较为近似, 均偏好疏水性或正电性的氨基酸残基. 尤其以 P8 位的疏水性对酶切的贡献最大, 因此 P8 位上出现 Ile, Leu, Met 和 Phe 可能更为有利. Falk 等[17]研究发现 Leu 在 P8 位的出现频率较高, 这与本文研究结果基本一致.

如前所述, 在酶切位点“|”下游 P3', P4'和 P5'位的疏水性($VHSE_1$)对酶切影响较大, 其权重系数均为负值. 表明这 3 个位置上出现疏水性氨基酸可能不利于蛋白酶体对底物的酶切; 反之, 亲水性氨基酸如: Asp, Glu, His 和 Lys 则有利于蛋白酶体对底物的酶切. 酶切位点两侧的疏水性差异可能反映了酶切位点周边氨基酸环境的差异性.

最后, 我们总结出 P9, P8, P4, P1, P3', P4'和 P5'位的氨基酸选择偏好, 结果参见表 5, 以便读者作进一步分析.

3 结论

本文构建了一种基于 VHSE 结构表征的蛋白酶体酶切预测方法, 应用该方法对 2650 个酶切样本建立了线性 SVM 预测模型, 取得了较优的预测结果. 与已有方法相比, 本模型具有样本容量大、模型物理意义明确和可解释性等特点. 此外, 由于模型全部采用细胞内 MHC-I 配体数据, 因此更合适用于抗原表位的预测研究. 模型分析结果表明: 影响酶切位点选择性的氨基酸性质由大到小依次为: 疏水性、电性和立体特征; P9, P8, P4, P1, P3', P4'和 P5'位氨基酸对酶切位点的选择有重要影响, 尤其是 P1, P8 和 P5'的疏水性质影响更为显著. 同时我们也发现, 酶切位点上下游的“疏水势差”有利于酶切的进行. 此外, P1 和 P5'的电性和立体几何性质对酶切亦有重要影响. 由于蛋白酶体结构的复杂性和底物的多样性, 因此还需进一步结合蛋白酶体的三维结构进行更为深入的分析.

表 5 酶切特异性位点氨基酸选择偏好 [a]

Table 5 Selective cleavage profiles at AA positions of substrate

Amino acid position	Property	Preferred amino acids	Contribution to cleavage
P9	Positive charge	K, R, H	＋
P8	Hydrophobic	F, W, L, I, M, V, Y	＋
P4	Steric	W, R, Y, F, K	－
P1	Hydrophobic, Positive charge, Bulky	F, W, L, I, V, Y, K, R, H	＋
P3'	Hydrophobic	F, W, L, I, M, V, Y	－
P4'	Hydrophobic	F, W, L, I, M, V, Y	－
P5'	Hydrophobic, Positive charge, Bulky	F, W, L, I, V, Y, K, R, H	－

[a] ＋, positive contribution; －, negative contribution.

References

1 Kloetzel, P. M. *Biochim. Biophys. Acta, Mol. Cell Res.* **2004**, *1695*, 225.

2 Nath, D.; Shadan, S. *Nature* (*London, U. K.*) **2009**, *458*, 421.

3 Lévy, F.; Burri, L.; Morel, S.; Peitrequin, A. L.; Lévy, N.; Bachi, A.; Hellman, U.; Van den Eynde, B. J.; Servis, C. *J. Immunol.* **2002**, *169*, 4161.

4 (a) Kuttler, C.; Nussbaum, A. K.; Dick, T. P.; Rammensee, H. G.; Schild, H.; Hadeler, K. P. *J. Mol. Biol.* **2000**, *301*, 229.
(b) Nussbaum, A. K.; Kuttler, C.; Hadeler, K. P.; Rammensee, H. G.; Schild, H. *Immunogenetics* **2001**, *53*, 87.

5 Holzhutter, H. G.; Frommel, C.; Kloetzel, P. M. *J. Mol. Biol.* **1999**, *286*, 1251.

6 Kesimir, C.; Nussbaum, A. K.; Schild, H.; Detours, V.; Brunak, S. *Protein Eng.* **2002**, *15*, 287.

7 Burges, C. J. C. *Data Min. Knowl. Disc.* **1998**, *2*, 121.

8 Blythe, M. J.; Doytchinova, I. A.; Flower, D. R. *Bioinformatics* **2002**, *18*, 434.

9 Schneider, M.; Lane, L.; Boutet, E.; Lieberherr, D.; Tognolli, M.; Bougueleret, L.; Bairoch, A. *J. Proteomics* **2009**, *72*, 567.

10 Altuvia, Y.; Margalit, H. *J. Mol. Biol.* **2000**, *295*, 879.

11 Mei, H.; Liao, Z. H.; Zhou, Y.; Li, S. Z. *Pept. Sci.* **2005**, *80*, 775.

12 Sanchez, V. D. *Neurocomputing* **2003**, *55*, 5.

13 Hofmann, T.; Schölkopf, B.; Smola, A. J. *Ann. Math. Stat.* **2008**, *36*, 1171.

14 Pardo, M.; Sberveglieri, G. *Sens. Actuators, B* **2005**, *107*, 730.

15 Saxova, P.; Buus, S.; Brunak, S.; Kesmir, C. *Int. Immunol.* **2003**, *15*, 781.

16 Nussbaum, A. K.; Dick, T. P.; Keilholz, W.; Schirle, M.; Stevanovic, S.; Dietz, K.; Heinemeyer, W.; Groll, M.; Wolf, D. H.; Huber, R.; Rammensee, H. G.; Schild, H. *Proc. Natl.*

Acad. *Sci*. *U*. *S*. *A*. **1998**, *95*, 12504.

17 Falk, K.; Rotzschke, O.; Stevanovic, S.; Jung, G.; Rammensee, H. G. *Nature* (*London*, *U*. *K*.) **1991**, *351*, 290.

18 Strehl, B.; Textoris-Taube, K.; Jakel, S.; Voigt, A.; Henklein, P.; Steinhoff, U.; Kloetzel, P. M.; Kuckelkorn, U. *J*. *Biol*. *Chem*. **2008**, *283*, 17891.

(A1106042 Zhao, X.)