中国科学 B辑:化学 2008 年 第 38 卷 第 9 期: 808 ~ 815
www.scichina.com chem.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

蛋白酶体对抗原蛋白酶切特异性的研究

宋哲[①], 刘涛[①], 焦春波[①], 刘伟[①②*], 朱鸣华[①], 王晓刚[①]

① 大连理工大学高科技研究院, 大连 116023;
② 大连理工大学物理与光电工程学院, 大连 116023
* 联系人, E-mail: jchjys@dlut.edu.cn

收稿日期: 2008-03-11; 接受日期: 2008-05-30

摘要 在 MHC I 类分子结合的抗原表位加工呈递途径中, 泛素-蛋白酶体系统对抗原蛋白的降解发挥着重要作用. 为了进一步研究蛋白酶体的酶切特异性, 采用偏最小二乘方法(partial least squares method, PLS)建立了蛋白酶体的酶切位点预测模型, 预测准确度为 82.8%; 由样本数据相应氨基酸对酶切位点形成的权重系数, 得出蛋白酶体酶切位点及其两侧区域氨基酸的裂解特异性, 它反映了蛋白酶体对抗原蛋白酶切的相互作用信息, 也表明蛋白酶体对抗原蛋白酶切处理不是随机的, 而是有一定模式和选择性的.

关键词
蛋白酶体
抗原表位
偏最小二乘法

抗原表位是指病原微生物中能够引起免疫应答和免疫反应的一簇特殊化学基团, 也称为抗原肽. 在MHC I类分子结合的抗原表位加工呈递途径中, 真核细胞中的泛素-蛋白酶体系统对抗原蛋白发挥着重要的酶切作用和降解功能, 其过程是靶蛋白首先以共价键形式联结多个泛素(ubiquitin)分子, 形成靶蛋白多聚泛素链, 即靶蛋白泛素化(ubiquitination), 然后再被送到 26S蛋白酶体中消化降解[1]. 26S蛋白酶体(proteasome)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依赖性蛋白水解酶复合体. 19S RP调节颗粒的功能是识别泛素化的靶蛋白并在其进入 20S CP催化颗粒前对其进行去泛素化、去折叠和移位[2]. 20S CP是 26S蛋白酶体的催化核心, 它是由 4 个环堆砌形成的一个圆桶状结构, 其中两侧外环每个环是由α_1~α_7 7 个亚基组成, 2 个内环每个环是由β_1~β_7 7 个亚基组成, 4 个环的中央形成一个狭窄的内腔[3]. α环上形成的窄口是底物进入位于β环上的催化中心的通道, 该窄口一般被α亚基上的N端所封闭, 阻止胞内非目的靶蛋白进入 20S CP内被降解破坏. 20S CP催化颗粒与 19S RP调节颗粒的结合导致α亚基构象改变, 并开启底物通道, α亚基能促使底物进入 20S CP的水解中心, 只有进入蛋白酶体内部的靶蛋白才能够被水解. β亚基N端苏氨酸残基是蛋白酶体活性位点的中心, 但是不同的β亚基有不同的蛋白酶活性, 分别具有能够使底物中大多数肽键断裂的能力, 从而消化降解底物, 释放出多肽碎片, 同时解离出泛素分子使其重新参与降解[4]. 文献上已有报道依据相关蛋白酶体裂解产物的实验数据, 对蛋白酶体酶切抗原蛋白的酶切位点进行理论预测, 如PAProC[5,6] 用酵母和人的20S蛋白酶体的体外酶切数据作为训练集, 以进化算法训练一个单层人工神经网络模型来预测蛋白酶体酶切位点; MAPPP[7,8] 是包含蛋白酶体酶切位点预测程序FragPredict和MHC分子结合配体预测程序的一个软件包, FragPredict程序是基于实验上统计分析得出蛋白酶体酶切基序并结合 20S蛋白酶体

动力学模型来预测蛋白酶体酶切位点; NetChop[9] 用20S蛋白酶体的体外酶切数据及MHC-I类分子配体数据作为训练集, 以人工神经网络模型来预测蛋白酶体酶切位点; 蛋白酶体裂解底物实质上是蛋白质与蛋白质的相互作用, 上述文献[5~9]研究工作并没有从蛋白酶体结构方面进行分析和讨论, 所以预测蛋白酶体酶切位点的准确性和可靠性还有待进一步提高. 本文采用偏最小二乘方法(partial least squares method, PLS)对蛋白酶体的酶切位点裂解特异性进行研究, 从蛋白酶体裂解产物的实验数据中, 提取尽量多的有用信息并去除噪声, 并试图结合蛋白酶体结构信息得出有益结果, 这有助于人们对MHC I类分子结合抗原表位加工提呈过程的深入了解, 对人们利用蛋白酶体影响进行肿瘤疫苗设计都是具有一定指导意义的.

1 模型与方法

1.1 裂解样本数据集和非裂解样本数据集

在MHC I类分子结合的抗原表位加工呈递过程中, 蛋白酶体主要负责准确酶切抗原表位羧基末端, 而抗原表位氨基末端是有胞内其他酶来进一步修剪形成的[4]. 在抗原表位的源蛋白氨基酸序列中, 若酶切是在P1位置氨基酸肽键上发生的, 则P1点位置的两侧氨基酸序列是以(PL…P2P1| P1’P2’…PL’)方式表示, 酶切位点是以符号“|”表示, 确定裂解样本或非裂解样本数据中的氨基酸序列窗口大小 (amino acids windows size, windows size)是在源蛋白氨基酸序列中进行的. 本文主要考察抗原表位羧基端的酶切位点及其两侧区域氨基酸的裂解特异性, 所以裂解样本数据(正样本)是以HLA-I类分子配体羧基端作为裂解位点(P1)获取其两侧氨基酸序列; 而非裂解样本数据(负样本)只是以HLA-I类分子配体内的中间位点为非裂解位点(P1)获取其两侧氨基酸序列. Goldberg等人[10]证实蛋白酶体即可以产生抗原表位又可以同时破坏摧毁该表位. 因此, 本文观点是不能简单地将抗原表位内部的所有位点都作为非裂解点. 而应该是将抗原表位内部的位点都作为次要裂解点. 相对而言, 抗原表位的中间位点较其内部的其他位点似乎应具有更小的裂解概率.

本文从AntiJen数据库[11]http://www.jenner.ac.uk/AntiJen获得 4915 个HLA I类分子的配体, 这些配体来源于 307 个人类蛋白质并且与 22 个HLA-A分子、21 个HLA-B分子和 3 个HLA-C分子相关. 去除重复的、氨基酸序列长度大于 12 或小于 8 的配体, 剩余1160 个HLA I类分子的配体, 最后得到裂解样本 1160, 非裂解样本 1160.

1.2 预测模型的建立

蛋白酶体裂解抗原蛋白除了与酶切位点氨基酸有关, 还与酶切位点“|”两侧氨基酸序列有关[12]. 本文设定在第k个样本数据氨基酸序列中, 每个位置出现的氨基酸对酶切位点 “|” 的影响可用下式描述

$$Y_k = \text{const} + \sum_{i=1}^{W} C_i \,, \tag{1}$$

式中 const 为常数, 表示此样本氨基酸的主链对酶切位点“|”形成的贡献, $\sum_{i=1}^{W} C_i$ 为样本中每个氨基酸残基 Pi 对酶切位点“|”形成的贡献和, W 为样本数据氨基酸序列窗口大小(windows size), W=2L. 若样本为裂解样本, 则裂解值 Y_k=1, 否则, 裂解值 Y_k=0.

在样本数据氨基酸序列中, 每个位置可能出现 20 种氨基酸残基中的任意一种, 所以可用 20 个变量对应样本数据氨基酸序列中每个位置可能出现的 20 种氨基酸残基, 即

$$C_i = \sum_{j=1}^{20} \alpha_{ij} g_{ij} \,, \quad (i = 1, 2, \cdots, W) \,, \tag{2}$$

其中 g_{ij} 是样本数据氨基酸序列中第 i 个位置上, 第 j 种氨基酸对应的变量. 若第 i 个位置是第 j 种氨基酸, 则该氨基酸对应的 g_{ij} 为 1, 否则为 0; α_{ij} 为相应氨基酸残基对酶切位点“|”形成的贡献, 也是相应氨基酸残基对酶切位点“|”形成的权重系数. 将(2)式代入到(1)式中可得

$$Y_k = \text{const} + \sum_{i=1}^{W} \sum_{j=1}^{20} \alpha_{ij} g_{ij} \,, \tag{3}$$

用变量 x (取值范围为 1 或 0)代替 g_{ij}, 则(3)式变成

$$Y_k = \text{const} + \sum_{i=1}^{20 \times W} \alpha_i \, \mathrm{x}_i \,. \tag{4}$$

若有 N 个样本数据, 样本氨基酸序列窗口大小

W=2L 及其相应的裂解值 Y_k, 则有多元线性方程组

$$\begin{bmatrix} y_1 \\ \vdots \\ y_k \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_k \\ \vdots \\ c_N \end{bmatrix} + \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1\mathrm{m}} \\ \vdots & \vdots & \vdots & \vdots \\ x_{k1} & x_{k2} & \ddots & x_{k\mathrm{m}} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{N\mathrm{m}} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_j \\ \vdots \\ \alpha_m \end{bmatrix}, \quad (5)$$

式(5)中 y_k 是第 k 个样本数据对应的裂解值, x_{kj} 为第 k 个样本数据相应位置氨基酸残基对应的第 j 个变量, α_j 为相应氨基酸残基对酶切位点 “|” 形成的权重系数, 且 $k=1,\cdots,N; j=1,\cdots,m;\ m=20\times W$. N 为样本数据的个数, m 为每个样本数据所有位置氨基酸残基对应的变量个数.

对n个样本组成的样本数据阵, 一般来说, 矩阵X中的各列向量并不是相互独立的, 而且矩阵X和Y存在相关性, 那么矩阵X和Y所包含的数据信息存在冗余. 偏最小二乘法是通过线性重组自变量, 产生一组相互正交且与因变量相关的新的自变量, 建立因变量与新自变量的线性关系, 从而能够达到提取重要原始变量信息并且去除噪声的目的. 本文采用文献[13~15] PLS方法自编程序求解(5)式, 结果得预测模型为

$$y = X\alpha = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_m x_m . \quad (6)$$

1.3 模型预测能力评估

为了评估模型的预测能力, 引入以下几个参量: 敏感度(SE), 特异性(SP), 裂解精确度(PPV), 非裂解精确度(NPV), 准确度(AC)和相关系数(CC)[16], 其中,

$$\mathrm{SE}=\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}\times 100,\ \mathrm{SP}=\frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FP})}\times 100,$$

$$\mathrm{PPV}=\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}\times 100,\ \mathrm{NPV}=\frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FN})}\times 100,$$

$$\mathrm{AC}=\frac{\mathrm{TP}+\mathrm{TN}}{(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN})}\times 100,$$

$$\mathrm{CC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FN}\times\mathrm{FP}}{\sqrt{(\mathrm{TN}+\mathrm{FN})(\mathrm{FN}+\mathrm{TP})(\mathrm{TP}+\mathrm{FP})(\mathrm{FP}+\mathrm{TN})}}. \quad (7)$$

(7)式中的 TP, FP, TN, FN 分别表示预测出的真裂解样本数、假裂解样本数、真的非裂解样本数、假的非裂解样本数. SE 表示正确预测裂解样本的百分比, SP 表示正确预测非裂解样本的百分比, AC 和 CC 为预测裂解样本和非裂解样本的评价指标.

2 结果

2.1 蛋白酶体裂解位点预测模型的性能

在样本数据氨基酸序列窗口大小为 4~28 AAs时, 本文采用k-折交叉验证(k-fold cross-validation)方法(设定k=10)研究蛋白酶体裂解位点预测模型的性能. 设定一个模型参数——阈值(threshold), 将预测得到的y值与阈值比较, 若y值大于阈值, 则认为测试样本为裂解样本; 反之, 则认为测试样本为非裂解样本. 当模型取不同的阈值时, 其预测能力也不同. 只有当敏感度与特异性相等(或最接近)的时候, 模型预测出的结果才是最好的, 可信度才最高[17]. 当样本数据氨基酸序列窗口大小为 20 AAs时, 模型的预测准确度82.8%为最好, 见表 1 所示.

表 1 蛋白酶体裂解位点模型的预测能力 a)

序列窗口大小	阈值	SE/%	SP/%	PPV/%	NPV/%	AC/%	CC
4	0.59	73.45	73.71	73.64	73.52	73.58	0.4716
6	0.58	74.05	74.05	74.05	74.05	74.05	0.4811
8	0.515	80.60	80.52	80.53	80.59	80.56	0.6112
10	0.515	80.43	80.69	80.64	80.48	80.56	0.6112
12	0.515	80.00	80.00	80.00	80.00	80.00	0.6
14	0.505	80.69	80.26	80.34	80.61	80.47	0.6095
16	0.515	81.29	82.07	81.93	81.44	81.68	0.6336
18	0.515	82.59	82.24	82.30	82.53	82.41	0.6483
20	**0.505**	**83.02**	**82.59**	**82.66**	**82.94**	**82.80**	**0.656**
22	0.505	82.76	82.59	82.62	82.73	82.67	0.6534
24	0.51	82.59	82.50	82.52	82.57	82.54	0.6509
26	0.51	82.07	82.24	82.21	82.10	82.16	0.6431
28	0.51	82.93	82.59	82.65	82.87	82.76	0.6552

a) SE: 敏感度; SP: 特异性; PPV: 裂解精确度; NPV: 非裂解精确度; AC: 准确度; CC: 相关系数

2.2 权重系数

当样本数据氨基酸序列窗口大小为 20 AAs时, 用文献[13,14] PLS方法求解得到(6)式中的权重系数α值, 见图 1 所示. 在图 1 的样本数据氨基酸序列P10~P10’ 位置中, 每个位置氨基酸符号的高度表示了该氨基酸对裂解位点“|”形成的权重系数α值大小, 如氨基酸Hisα值大于氨基酸Leuα值, 氨基酸符号的正置或倒置表示相应α值为正值或负值; 每个位置全部氨基酸符号堆积的总高度表示了该位置所有氨基酸对裂解位点“|”形成的权重系数总和的绝对值大小. 20 种氨基酸被分成酸性、碱性、疏水性和中性 4 类, 其中酸性氨基酸为红色, 碱性氨基酸为蓝色, 疏水性氨基酸为黑色, 中性氨基酸为绿色.

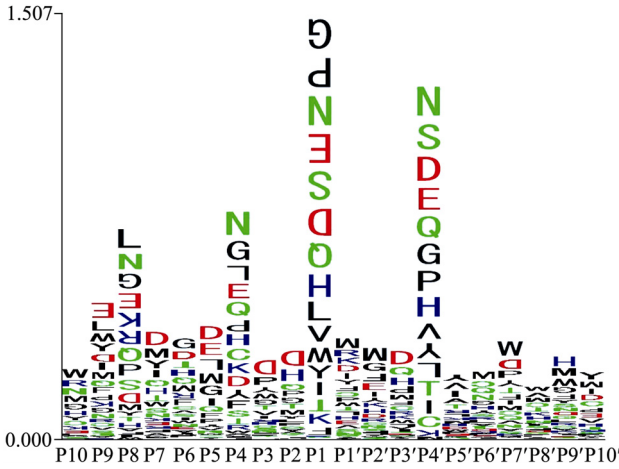

图 1 样本数据序列窗口大小为 20 时每个位置氨基酸对裂解位点形成的权重系数

2.3 不同预测模型在相同检验集下的性能比较

采用文献[16]提供的样本数据测试集, 将本文开发的预测蛋白酶体酶切位点软件(PLS_dut)与其他预测软件[5~9]进行比较, 结果本模型预测的性能指标是最好, 见表 2. 文献[16]提供了PAProC[5,6], MAPPP[7,8], NetChop[9]预测软件的性能指标.

表 2 本模型与其他预测模型的性能比较

预测模型	*N*	敏感度	特异性	CC
PAProC[16]	217	45.6	30.0	−0.25
FragPredict[16]	231	83.5	16.5	0.00
NetChop1.0[16]	231	39.8	46.3	−0.14
NetChop2.0[16]	231	73.6	42.4	0.16
PLS_dut	231	76.2	61.9	0.38

3 讨论

在 20S CP内起裂解催化作用的亚基主要是β_1, β_2, β_5, 当β亚基的*N*端前导序列在 20S CP装配过程中被切除后, 苏氨酸Thr1 残基被暴露出来, Thr1 是酶的活性位点, 分别存在于β环的内表面, 使β亚基具有类似的丝氨酸蛋白酶的催化作用[4]. 哺乳动物蛋白酶体 2 个β内环共有 6 个β亚基的活性位点, 这些活性位点能够单独作用裂解底物蛋白分子的肽键, 也可能 2 个活性位点共同作用裂解底物蛋白分子的肽键[18]. 由于进入蛋白酶体的靶蛋白是完全去折叠的(fully unfolded)[3], 依据文献“生物分子尺度”观点[18]: 蛋白酶体活性位点的距离确定了裂解产物片段的不同长度, 本文计算了哺乳动物 20S蛋白酶体活性位点相互间的距离如表 3 所示, 实际计算是以蛋白酶体的三维空间结构(PDB code：1IRU)β亚基的活性位点苏氨酸Thr1 的Oγ原子之间的距离为参考. 在表 3 中, 蛋白酶体 2 个活性位点相互间的距离为 2.8~6.6 nm. 根据Coux等人[3]给出的去折叠蛋白质肽链长度与氨基酸序列长度的对应关系, 裂解产物片段2.8 nm肽链长度相当于 7~8 AAs, 裂解产物片段 6.6 nm肽链长度相当于 16~19 AAs. Kisselev等人[19]通过实验方法得知: 哺乳动物的蛋白酶体裂解产物片段的长度范围为 3~22 AAs, 约 70%裂解产物片段的长度太短(<7 AAs)而不适合提呈到内质网中, 只有不足 15%裂解产物片段的长度为 8~9 AAs. 可见表 3 中蛋白酶体的活性位点的距离与裂解产物片段的长度之间不形成一一对应的关系. 那么, 又如何解释大部分蛋白酶体裂解产物片段的长度是<10 AAs的实验事实? Dick等人[20]通过实验认为：蛋白酶体的 2 个活性位点共同作用产生的裂解产物, 有可能作为中间产物经过释放-捕获

表 3 哺乳动物 20S 蛋白酶体中活性位点相互间的距离[a)]

PDB ID	β	(Thr1)β_1	(Thr1)β_2	(Thr1)β_5
1IRU.pdb	(Thr1)β_1		2.8 nm	6.2 nm
	(Thr1)β_2	2.8 nm		6.3 nm
	(Thr1)β_5	6.2 nm	6.3 nm	
PDB ID	β β^*	(Thr1)β_1	(Thr1)β_2	(Thr1)β_5
1IRU.pdb	(Thr1)β^*_1	2.9 nm	5.0 nm	5.8 nm
	(Thr1)β^*_2	5.0 nm	6.6 nm	4.1 nm
	(Thr1)β^*_5	5.8 nm	4.1 nm	4.9 nm

a) β 和 β^*为哺乳动物 20S 蛋白酶体中不同 β 环的亚基

途径被蛋白酶体再次裂解, 使得大部分蛋白酶体裂解产物片段的长度是<10 AAs. 由表 1 可知, 当样本数据氨基酸序列窗口大小为 20 AAs时, 预测准确度为 82.8%, 它是能够包含大多数裂解产物片段的氨基酸序列信息. 另外, 本模型的预测能力也优于文献[5~9]提供的指标, 见表 2.

从图 1 中可以知道: 样本数据氨基酸序列每个位置氨基酸符号高度表示了该氨基酸对裂解位点“|”形成的权重系数α值大小, 每个位置全部氨基酸符号堆积的总高度表示了该位置所有氨基酸对裂解位点“|”形成的权重系数总和的绝对值大小. 在裂解位点“|”及其两侧的 P1, P4', P4, P8 和 P9 位置上氨基酸具有显著的裂解特异性, 这表明蛋白酶体对靶蛋白的酶切处理不是随机的,而是有一定模式和选择性的.

为了解释图 1 中裂解位点“|”及其两侧的氨基酸序列窗口显示的裂解特异性, 本文以哺乳动物 20S 蛋白酶体的三维空间结构(PDB code: 1IRU)为例, 考察了活性位点苏氨酸Thr1“非键接触”的近邻氨基酸, 其中定义“非键接触”的 2 个氨基酸中至少有 2 个原子之间的距离小于 0.4 nm[21], 见表 4 所示. 由表 4 可知, 3 个亚基β_1, β_2和β_5活性位点周围的氨基酸中, β_1亚基近邻氨基酸中有 3 个碱性、带正电(Arg19, Lys33和Arg45)和 2 个酸性、带负电的(Asp17 和Asp167); β_2亚基近邻氨基酸中有 1 个碱性、带正电的(Lys33)和 2 个酸性、带负电的(Asp17 和Asp166); β_5亚基近邻氨基酸中有 2 个碱性、带正电的(Arg19 和Lys33)和 2 个酸性、带负电的(Asp17 和Asp167). 可以简单地认为: β_1亚基活性位点周围的氨基酸呈正电性, 易于裂解底物中酸性带负电的氨基酸的肽键; β_2亚基活性位点周围的氨基酸呈负电性, 易于裂解底物中碱性带正电的氨基酸的肽键; β_5亚基活性位点周围的氨基酸呈中性. 文献[3]证实: β_1亚基能够识别底物中的酸性氨基酸残基,并切断该残基后的肽键; β_2亚基能够识别底物中的碱性氨基酸残基,并切断该残基后的肽键; β_5亚基能够识别底物中大的疏水性氨基酸残基,并切断该残基后的肽键. 可见, 蛋白酶体中β_1, β_2和β_5亚基能够裂解底物中不同氨基酸的肽键, 是与各自活性位点中心苏氨酸Thr1 的“非键接触”近邻氨基酸特性有关. 图 1 中裂解位点“|”一侧的P1 位置上氨基酸裂解特异性也反映了每个β亚基活性位点近邻氨基酸的偏好特性.

由图 1 中可知, 在样本氨基酸序列 P1 位置上出现正值权重系数大的是疏水性氨基酸 Leu, Val, Tyr, Ile 等, 它们有利于 20S 蛋白酶体对底物的裂解. P1 位置上出现负值权重系数绝对值大的是亲水性氨基酸 Asn, Ser, Gln 和 Thr, 酸性氨基酸 Glu, Asp 和碱性氨基酸 His, 它们可能不利于 20S 蛋白酶体对底物的裂解. 考察哺乳动物 20S 蛋白酶体的三维空间结构(PDB

表 4 哺乳动物 20S 蛋白酶体中与活性位点苏氨酸 Thr1“非键接触”的近邻氨基酸[a)]

β_1 活性位点苏氨酸 Thr1 的近邻氨基酸		β_2 活性位点苏氨酸 Thr1 的近邻氨基酸		β_5 活性位点苏氨酸 Thr1 的近邻氨基酸	
β_1 亚基	Thr2	β_2 亚基	Thr2	β_5 亚基	Thr2
	Ile3		Ile3		Thr3
	Asp17		Asp17		Asp17
	Arg19		Lys33		Arg19
	Lys33		Ala46		Lys33
	Arg45		Met127		Met45
	Ser46		Gly128		Gly129
	Gly129		Ser129		Ser130
	Ser130		Gly130		Gly131
	Gly131		Asp166		Asp167
	Asp167		Gly168		Tyr169
	Ser169		Ser169		Ser170
	Ser170				

a) 带正电、碱性的氨基酸用蓝色字体表示, 带负电、酸性的氨基酸用红色字体表示, 中性氨基酸用黑色字体表示

code: 1IRU), 可知β亚基活性位点中心苏氨酸Thr1 的Oγ原子是处于β亚基内腔空穴(the inner cavity wall)的中心, 疏水性氨基酸的残基易进入到该内腔空穴中, 使得β亚基苏氨酸Thr1 的Oγ原子能够直接作用到底物蛋白主链的肽键上, 并与其周围的氨基酸共同作用完成酶切靶蛋白. P1 位置疏水性氨基酸Leu, Val的肽键在被 20S蛋白酶体酶切后, 形成多肽片段的C端, 这也与氨基酸Leu, Val为HLA-A2 分子配体在Pocket F 的初级锚点[22]事实相符合. 从图 1 可以看出, P1 位置上出现负值权重系数绝对值大的是疏水性氨基酸Gly和Pro, 它们的疏水性很弱, 不易进入β亚基活性位点苏氨酸Thr1 的Oγ原子处的内腔空穴中; Nussbaum等人[23]通过实验得出: 酵母原生 20S蛋白酶体裂解底物时, 对于底物P1 位置, 氨基酸Leu是最重要的, 而Gly不利于裂解. 由此可见, 本文结果与Nussbaum等人[23]的实验结果基本一致. 因此, 疏水性氨基酸在底物P1 位置容易被β亚基的活性位点识别, 有利于活性位点裂解底物; 反之, 亲水性氨基酸在底物P1 位置不容易被β亚基的活性位点识别, 不利于活性位点裂解底物.

由图 1 中可知, 在样本氨基酸序列P4'位置上出现正值权重系数大的是亲水性氨基酸Asn, Ser, Gln, Thr, 酸性氨基酸Asp, Glu, 碱性氨基酸His和弱疏水性氨基酸Gly, Pro, 它们可能有利于 20S蛋白酶体对底物的裂解. P4'位置上出现负值权重系数绝对值大是疏水性氨基酸Val, Tyr, Leu和Ile, 它们可能不利于20S蛋白酶体对底物的裂解. 有意思的是: 在裂解位点"|"两侧的P1 和P4'位置氨基酸显示裂解特异性的结果恰好相反. 蛋白酶体β亚基中具有酶切活性的苏氨酸Thr1 作用靶点是P1 位置氨基酸的肽键, P4'位置的氨基酸特异性又如何解释? Wenzel等人[18]认为: 底物蛋白能够结合于蛋白酶体α环和β环界面处. 我们计算了哺乳动物 20S蛋白酶体结构(PDB code: 1IRU)β 环的 3 个亚基 β_1, β_2, β_5的苏氨酸 Thr1 的 Oγ 原子与α环原子最短距离的平均值为 2.03 nm(2.17, 1.86, 2.07 nm), 相当于裂解位点"|"至 P4'位置的去折叠肽链长度, 大概为 4~5 AAs. P4'位置氨基酸裂解特异性可能反映了蛋白酶体α、β 环界面处氨基酸与底物 P4'位置氨基酸相互作用特性.

由图 1 中可知, 在样本氨基酸序列 P4 位置上出现正值权重系数大是亲水性氨基酸 Asn, Gln, 弱疏水性氨基酸 Gly, 酸性氨基酸 Glu 和 Asp、碱性氨基酸 His 和 Lys 及疏水性氨基酸 Phe, 它们可能有利于 20S蛋白酶体对底物的裂解. P4 位置上出现负值权重系数绝对值大是疏水性氨基酸 Leu, Pro 和 Tyr, 它们可能不利于 20S 蛋白酶体对底物的裂解. Nussbaum 等人[23]通过实验得出: 酵母原生 20S蛋白酶体裂解底物时, P4 位置氨基酸Pro是最重要的. 而本文的结果与此实验结果相反, 这可能是由于不同生物的 20S蛋白酶体的结构差异造成的. 因为哺乳动物和酵母 20S蛋白酶体的α亚基存在较明显的结构差异, 而它们的β亚基也存在结构差异[24]. 本文所采用裂解和非裂解样本是由HLA I类分子结合抗原肽扩展而来, P4 位置氨基酸特异性有可能是抗原肽的Pocket C的结合特异性信息, 在文献[13~15]中我们未发现Pocket C具有很强的结合特异性, 那么本文得出的P4 位置的氨基酸特异性又如何解释? 考察哺乳动物 20S蛋白酶体结构(PDB code: 1IRU)可知: 蛋白酶体的上层β_1亚基活性位点的苏氨酸Thr1 的Oγ原子与下层β环中的β^*_1, β^*_7亚基原子的最近距离分别为 1.62 和 1.08 nm; 蛋白酶体的上层β_2 亚基活性位点的苏氨酸Thr1 的Oγ原子与下层β环中的β^*_7, β^*_6亚基原子的最近距离分别为 1.49 和 1.00 nm; 蛋白酶体的上层β_5亚基活性位点的苏氨酸Thr1 的Oγ原子与下层β环中的β^*_4, β^*_3亚基原子的最近距离分别为 1.56 和 1.09 nm; 蛋白酶体上层β环活性位点与下层β环交界面处原子的距离为 1.00~1.62 nm, 相当于裂解位点"|"至P4 位置的去折叠肽链长度, 大概为 3~4 AAs, P4 位置氨基酸特异性应该是反映了蛋白酶体裂解特异性的信息. P4 位置氨基酸裂解特异性可能反映了蛋白酶体上下两层β环界面处氨基酸与P4 位置氨基酸相互作用特性.

由图 1 中可知, 在样本氨基酸序列 P8 位置上出现正值权重系数大是疏水性氨基酸 Leu, Pro, 它们可能有利于 20S 蛋白酶体对底物的裂解. P8 位置上出现负值权重系数绝对值大是亲水性氨基酸 Asn, Gln, Ser, 弱疏水性氨基酸 Gly, 酸性氨基酸 Glu, Asp, 碱性氨基酸 Lys, Arg, 它们可能不利于 20S 蛋白酶体对底物的裂解. P8 位置的氨基酸特异性又如何解释? 由表 3

可知：2 个不同层β环中β_1亚基之间的距离为 2.9 nm, 大约为去折叠肽链 7~8 AAs长度, P8 位置与P1 位置是在裂解位点“|”同一侧, 在去折叠的蛋白质肽链上也相差 8 AAs, P8 位置与P1 位置的氨基酸特异性有些相似, 有可能是蛋白酶体不同层β环中 2 个β_1亚基活性位点共同作用底物显现出的裂解特异性. 氨基酸Leu又是抗原肽结合在HLA-A2 类分子Pocket B的初级锚点[21], P8 位置上氨基酸裂解特异性可能确定了抗原肽在HLA I类分子Pocket B初级锚点的结合特异性.

由图 1 中可知, 在样本氨基酸序列P9 位置上出现正值权重系数大是疏水性氨基酸Leu, Tyr, Ile和Met, 它们可能有利于 20S蛋白酶体对底物的裂解. P9 位置上出现负值权重系数绝对值大是碱性氨基酸Glu, Asp, 疏水性氨基酸Trp, 它们可能不利于 20S蛋白酶体对底物的裂解. 在HLA-A*0201 分子结合抗原肽中, 疏水性氨基酸Tyr 作为锚点出现在Pocket A中频率最大[13], P9 位置上氨基酸裂解特异性可能确定了抗原肽在HLA I类分子Pocket A锚点的结合特异性.

综上所述, 预测蛋白酶体裂解位点模型所选取样本数据氨基酸序列 Windows Size 为 20 是可行的合理的, 裂解位点“|”及其近邻位置氨基酸显示了明显的裂解特异性, 疏水性氨基酸在 P1 位置容易被 β 亚基的活性位点识别, 有利于活性位点裂解底物; P4’和 P4 位置氨基酸的裂解特异性反映了蛋白酶体裂解底物的相互作用信息.

4 结论

(1) 利用 PLS 方法建立的蛋白酶体裂解位点预测模型是可行合理的, 预测准确度为 82.8%.

(2) 蛋白酶体裂解抗原蛋白时, 裂解位点“|”及其近邻位置氨基酸显示了明显的裂解特异性, 疏水性氨基酸在 P1 位置容易被 β 亚基的活性位点识别, 有利于活性位点裂解底物; P4’和 P4 位置氨基酸的裂解特异性反映了蛋白酶体裂解底物的相互作用信息.

参考文献

1 Adams J. The proteasome: structure, function, and role in the cell. Cancer Treat Rev, 2003, 29(1): 3—9[DOI]

2 Smith D, Benaroudj N, Goldberg A L. Proteasomes and their associated ATPases: A destructive combination. J Structural Biol, 2006, 156: 72—83

3 Coux O, Tanaka K, Goldberg A L. Structure and functions of the 20S and 26S proteasomes. Annu Rev Biochem, 1996, 65: 801—847 [DOI]

4 Heinemeyer W, Ramos P C, Dohmen R J. The ultimate nanoscale mincer: assembly, structure and active sites of the 20S proteasome core. Cell Mol Life Sci, 2004, 61: 1562—1578[DOI]

5 Kuttler C, Nussbaum A K, Dick T P, Rammensee H G, Schild H, Hadeler K P. An algorithm for the prediction of proteasomal cleavages. J Mol Biol, 2000, 298(3): 417—429[DOI]

6 Nussbaum A K, Kuttler C, Hadeler K P, Rammensee H G, Schild H. PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. Immunogenetics, 2001, 53(2): 87—94[DOI]

7 Holzhutter H G, Frommel C, Kloetzel P M. A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20S proteasome. J Mol Biol, 1999, 286(4): 1251—1265[DOI]

8 Holzhutter H G, Kloetzel P M. A kinetic model of vertebrate 20S proteasome accounting for the generation of major proteolytic fragments from oligomeric peptide substrates. Biophysical J, 2000, 79: 1196—1205

9 Kesmir C, Nussbaum A K, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. Protein Eng, 2002, 15(4): 287—296[DOI]

10 Goldberg A L, Cascio P, Saric T, Rock K L. The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides. Mol Immunol, 2002, 39(3-4): 147—164[DOI]

11 Blythe M J, Doytchinova I A, Flower D R. JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics, 2002, 18: 434—439[DOI]

12 Altuvia Y, Margalit H. Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage

mechanism. J Mol Biol, 2000, 295: 879—890[DOI]
13 宋哲, 刘涛, 刘伟, 朱鸣华, 王晓钢. 抗原肽与 MHC 分子相互作用的 QSAR 模型研究. 物理化学学报, 2007, 23(2): 198—205
14 宋哲, 刘涛, 王雪莹, 刘伟. PLS 方法应用于 T 细胞表位定量构效关系的研究. 免疫学杂志, 2007, 23(2): 166—171
15 刘涛, 宋哲, 刘伟, 王雪颖, 邱晓明. 基于改进的人工神经网络方法预测 CTL 表位. 大连理工大学学报, 2007, 47(4): 473—478
16 Saxova P, Buus S, Brunak S, Kesmir C. Predicting proteasomal cleavage sites: a comparison of available methods. Int Immunol, 2003, 15(7): 781—787[DOI]
17 Bhasin M, Raghava G P. Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine, 2004, 22: 3195—3204[DOI]
18 Wenzel T, Eckerskorn C, Lottspeich F, Baumeister W. Existence of a molecular ruler in proteasomes suggested by analysis of degradation products. FEBS Lett, 1994, 349: 205—209[DOI]
19 Kisselev A F, Akopian T N, Woo K M, Goldberg A L. The sizes of peptides generated from protein by mammalian 26 and 20 S proteasomes. Implications for understanding the degradative mechanism and antigen presentation. J Biol Chem, 1999, 274(6): 3363—3371[DOI]
20 Dick L R, Moomaw C R, DeMartino G N, Slanghter C A. Degradation of oxidized insulin B chain by the multiproteinase complex macropain(proteasome). Biochemistry, 1991, 30: 2725—2734[DOI]
21 Altuvia Y, Margalit H. A structure-based approach for prediction of MHC-binding peptides. Methods, 2004, 34(4): 454—459[DOI]
22 Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee H G. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature, 1991, 351: 290—296[DOI]
23 Nussbaum A K, Dick T P, Keilholz W, Schirle M Stevanovis S, Dietz K, Heinemeyer W, Groll M, Wolf D N, Huber R, Rammensee H G, Schild H. Cleavage motifs of the yeast 20S proteasome β subunits deduced from digests of enolase 1. Proc Natl Acad Sci USA, 1998, 95(21): 12504—12509[DOI]
24 Unno M, Mizushima Y, Morimoto Y, Tomisugi Y, Tanaka K, Yasuoka N, Tsukihara T. The structure of the mammalian 20S proteasome at 2.75 Å resolution. Structure, 2002, 10: 609—618[DOI]