

基于TAP分子亲和力模型预测MHC I类分子提呈短肽的免疫原性*

盛浩^{1)**} 卢玉峰²⁾ 包永明³⁾

(¹⁾ 辽宁石油化工大学理学院, 抚顺 113001; (²⁾ 大连理工大学数学科学学院, 大连 116023;

(³⁾ 大连理工大学生命科学与技术学院, 大连 116023)

摘要 主要组织相容性复合体 (major histocompatibility complex, MHC) I类分子亲和力模型有助于筛选出候选短肽, 为实验确定可以与MHC I类分子形成复合物从而激活细胞毒性T细胞的短肽提供帮助; 抗原处理相关转运体 (transporter associated with antigen processing, TAP) 分子亲和力模型也可用于筛选候选短肽; 如何更好利用两类亲和力模型筛选出候选短肽、这两类亲和力模型对短肽选择性的异同以及这种选择性异同的生物学机制尚不清楚. 本文重新整理了TAP亲和力测试集, 使训练样本达到699个; 使用核函数平衡矩阵算法建立的TAP亲和力模型预测精度高于同类算法, 5折交叉检验Pearson相关系数0.89; 结合HLA A3分子亲和力模型、TAP亲和力模型显著提高免疫原性短肽预测精度, AUC值从~0.82上升到0.87. 提高的原因是两个亲和力模型对第2、9位置“最佳”氨基酸不同偏好性和这种不同偏好的互补性造成的. TAP分子与短肽模拟对接结果反映了这个结论. TAP亲和力可在线预测 (<http://www.bilologymaths.top/mbtwo/major.aspx>).

关键词 免疫原性, 细胞毒性T细胞, MHC I类分子, 抗原处理相关转运体, 核函数平衡矩阵方法

中图分类号 R34, R35, R392

DOI: 10.16476/j.pibb.2019.0013

免疫治疗是一个涉及多个步骤的复杂工程, 具有抗原性的短肽搜寻处于重要地位. 其中一类需要搜寻的短肽来源于受病毒、细菌感染的机体细胞或肿瘤细胞. 在这些细胞胞浆内, 蛋白质被泛素化后会被蛋白酶体水解, 其中的一些水解产物短肽会有选择性地被抗原处理相关转运体 (the transporter associated with antigen processing, TAP) 转运到内质网里面与新组装的主要组织相容性复合体 (major histocompatibility complex, MHC) I类分子结合. 如果形成的MHC I类分子/短肽复合物能够与宿主体内的细胞毒性T淋巴细胞 (cytotoxic T lymphocyte, CTL) 的T细胞受体 (T cell receptor, TCR) 结合, 这个短肽就具有抗原性 (antigenicity); 如果MHC I类分子/短肽复合物与TCR结合后成功激活CTL对短肽提呈细胞的杀伤作用, 这个短肽就具有免疫原性 (immunogenicity). 一方面由于CTL只需MHC I限制性信号, 即MHC I类分子/短肽复合物, 而不需

要协同刺激信号即可完成对提呈细胞的杀伤作用; 另一方面大部分体细胞都表达MHC I类分子, 可以为CTL提供MHC I限制性信号, 因此具有免疫原性的MHC I限制性短肽搜寻对于利用CTL治疗疾病具有重大的医学意义. 本文中短肽的免疫原性指的是MHC I类分子限制性短肽所具有的免疫原性.

TAP分子参与了MHC I类分子限制性抗原加工提呈流程. 在TAP缺陷细胞中, 大多数MHC I类分子的肽结合槽是空的或者没有被正确装载, 原因是短肽无法进入内质网^[1]. TAP分子是一种腺苷三磷酸结合盒转运载体蛋白 (ATP-binding cassette, ABC). 这类转运蛋白在腺苷三磷酸水解条件下可

* 国家自然科学基金基金 (61602228), 辽宁省自然科学基金 (L2017LQN03) 和辽宁石油化工大学启动基金 (1100140109) 资助项目.

** 通讯联系人.

Tel: 15898372769, E-mail: shenghao@lnpu.edu.cn

收稿日期: 2019-05-08, 接受日期: 2019-12-26

以跨膜转运底物. TAP分子是由两个半转运蛋白单体形成的异二聚体. 这两个半转运蛋白单体由同源的MHC基因编译, 称为TAP1分子和TAP2分子. TAP1分子和TAP2分子的结构相似, 都包括核苷酸结合域、跨膜结构域和tapasin结合结构域. 其中的跨膜结构域负责底物的跨膜捕捉和释放. TAP分子对底物的转运机制处于猜想阶段^[2-3].

MHCI类分子、TAP分子与短肽结合都有选择性, 研究人员可以用已获得的MHCI类分子、TAP分子与某些短肽的亲合力数据训练一些算法^[4-8], 利用得到的模型搜寻具有免疫原性的短肽. 具有免疫原性短肽的预筛选主要使用MHCI类分子亲和力模型.

虽然MHCI类分子、TAP分子与短肽的结合都有选择性, 这种选择性是有差异性的, 因此MHCI类分子亲和力模型和TAP亲和力模型结合起来成为一个模型用于筛选具有免疫原性的短肽可能是有意义的. 已经有研究人员把这两种预测方法结合起来, 结合后的模型相对于单独的预测模型, 预测结果确实有提高^[5].

但是预测结果提高的原因, 该文并没有给出解释. 本文以此为出发点, 重复该实验. 同时由于线性模型可以直接反映每个位置氨基酸对结合亲和力的影响, 本文通过分析线性的MHCI类分子亲和力模型、TAP亲和力模型参数, 尝试区分短肽的MHCI类分子亲和力、TAP亲和力与免疫原性的差别. 短肽的高MHCI类分子亲和力和免疫原性的差别会导致MHCI类分子亲和力模型筛选具有免疫原性的短肽时精度产生系统性误差. 已有的实验结果提示了两者之间的差异, 主要体现在两方面:

a. MHCI类分子/短肽复合物的免疫原性偏低. 结核分枝杆菌的免疫实验表明, 只有不到两成的HLA A2类、HLA A3类和HLA B7类分子与短肽的复合物(8成人含有这些HLA基因)成功激活CTL细胞的免疫应答反应^[9]. b. 低MHCI类分子亲和力的短肽, 与MHCI类分子形成复合体后成功激活CTL细胞的免疫应答反应^[10].

另外, 本文尝试回答MHCI类分子的选择性与TAP的选择性差异, 如何提高了它们的双重选择对具有免疫原性短肽的预测精确率. MHCI类分子亲和力、TAP亲和力模型的预测精确度如果都是70%, 如果两个模型的相关性为0, 两个模型结合的预测精确度是49%, 不仅没有升高, 反而明显降低.

2 材料与方法

2.1 TAP分子亲和力模型相关数据集的建立

本文使用的与TAP分子作用的短肽数据除了部分来源于已有的文献^[5, 7], 也有部分数据搜集自MHCBN数据库^[11], MHCBN数据库是不断更新的, 这样保证了TAP分子亲和力数据集的即时性. 截至2018年12月12日, MHCBN数据库包含1 052条与TAP分子作用的短肽数据, 短肽长度为7~25个氨基酸. 本文把其中长度为9个氨基酸的730条短肽数据和文献^[5, 7]中的数据合并; 本文在EXCEL表中打开合并后的数据, 对表示短肽的字母串按照升序排列, 然后对空白单元格G1赋值“=IF(COUNTIF(A1:A\$2000, "*"&A1&"")>1, "duplicate", "not")”, 出现“+”后下拉, 这样本文找出全部重复数据. 有的短肽对应不同 IC_{50} 值的多条数据时, 本文使用这些数据的平均值. 最终得到699条与TAP分子作用的短肽数据. 短肽的结合亲和力表示为 $-\lg(\text{relative } IC_{50})$, 其中参照短肽“RRYNASTEL”的 IC_{50} 值为400 nmol/L. 短肽的结合亲和力值越大, 短肽能被TAP分子转运的可能性越高.

本文的研究范围是两类HLA超型A2、A3类分子与短肽结合的亲和力以及HLA超型A2、A3限制性短肽的免疫原性. 超型A2、A3的配体特异性信息取自于实验数据最多的两个分型HLA-A 2.1分子和HLA-A 3.1分子, 这些数据下载自IEDB数据库^[12]. IEDB数据库里的短肽数据有的包含了 IC_{50} 值, 有的只给出了定性的性质. 具有免疫原性的HLA-A 2.1分子和HLA-A 3.1分子限制性短肽从数据库IEDB^[12]、SYFPEITHI^[13]、肿瘤抗原数据库^[14]和Los Alamos HIV数据库^[15]提取. 这些短肽形成的相应的MHCI分子/短肽复合体已经被证实可以引起CTL细胞的免疫应答, 作为识别免疫原性能力测试集的正样本. 本文同时假定不可以和HLA分子结合的短肽不具有免疫原性, 这些短肽作为识别免疫原性能力测试集的负样本.

IEDB数据库包含7 314个与HLA-A 2.1分子作用的短肽数据. 本文首先删除其中重复的1 545条数据. 剩下的数据中有772条包含了 IC_{50} 值, 以这772条数据作为训练集、以(1)式中的 E_i 作为引导函数, 训练线性核函数KSMM, 这样得到了HLA-A 2.1分子亲和力模型(简称为KSMM-LINEAR_{2.1}). 下面开始A2限制性、识别免疫原性

能力测试集 (称为 A2-immunogenic) 的构建: IEDB 数据库中除了重复的和作为训练集使用的, 剩下的 4 701 条不与 HLA-A 2.1 分子结合的短肽作为测试集 A2-immunogenic 的负样本; 测试集 A2-immunogenic 的正样本是 863 条 HLA-A 2.1 限制性的、具有免疫原性的短肽; 负样本集和正样本集重复的短肽有 27 条, 在负样本集中删除; 这样最终的测试集 A2-immunogenic 包含 4 674 条负样本、863 条正样本. 本文另建立 HLA-A 2.1 亲和力测试集 IEDB-0201, 包含 4 701 条负样本 (非结合肽) 和 296 条正样本 (结合肽).

IEDB 数据库包含 5 400 个与 HLA-A 3.1 分子作用的短肽数据. 本文首先删除其中重复的 1 159 条数据. 剩下的数据中有 594 条包含了 IC_{50} 值, 以这 594 条数据作为训练集、以 (1) 式中的 E_1 作为引导函数, 训练线性核函数 KSMM, 这样得到了 HLA-A 3.1 分子亲和力模型 (简称为 KSMM-LINEAR_{3.1}). 下面开始 A3 限制性、识别免疫原性能力测试集 (称为 A3-immunogenic) 的构建: IEDB 数据库中除了重复的和作为训练集使用的, 剩下的 2 712 条不与 HLA-A 3.1 分子结合的短肽作为测试集 A3-immunogenic 的负样本; 测试集 A3-immunogenic 的正样本是 276 条 HLA-A 3.1 限制性的、具有免疫原性的短肽; 负样本集和正样本集重复的短肽有 9 个, 在负样本集中删除; 这样最终的测试集 A3-immunogenic 包含 2 703 条负样本、276 条正样本. 本文另建立 HLA-A 3.1 亲和力测试集 IEDB-0301, 包含 2 712 条负样本 (非结合肽) 和 347 条正样本 (结合肽).

2.2 KSMM算法简介

训练集 $T = \left\{ (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \right\}$, 其

中的 $x_i = [x_{i1}, x_{i2}, \dots, x_{in}]$ 是对氨基酸稀疏编码的向量, y_i 是属性值. KSMM 算法包含两步, 第一步是最小化 E_1 来寻找一个向量 w ,

$$E_1 = \sum_i |c_1 \cdot K_1(x_i, w) + b - y_i| + \lambda \cdot \|w\|^2 \quad (1)$$

其中的 b , c_1 , λ 是可以调节大小的参数, K_1 表示核函数, 本文使用线性 (linear) 和径向基 (radial based kernel function, Rbf) 核函数.

除了短肽每个位置氨基酸对短肽亲和力会产生影响, 短肽不同位置氨基酸的相互作用也有可能对短肽的亲和力会产生影响. 本文对这些影响简化处理, 考虑短肽两个不同位置氨基酸的相互作用对亲

和力的影响, 称为配对. 对于长度为 9 的短肽, 需要估计 14 400 个配对. 训练集中出现次数太少的配对不具有统计学意义, 本文设定的配对出现次数最小值是 10, 对于 TAP 与短肽的亲和力训练集, 一共有 126 个配对出现次数不小于 10. 每个样本表示为一个 126 维向量 x_i' , 若配对出现在样本中, 则 x_i' 的对应分量赋值 1; 否则 x_i' 的对应分量赋值 0. 若 w 最优解 w^* 和相应的 c_1^* 和 b^* , 最小化 E_2 寻找配对系数向量 w' 最优解

$$E_2 = \sum_i |c_1^* \cdot K_1(x_i, w^*) + c_2 \cdot K_2(x_i', w') + b^* - y_i| + \lambda' \cdot \|w'\|^2 \quad (2)$$

其中的 c_2 , λ' 是可以调节大小的参数, K_2 表示核函数, 本文使用线性和径向基核函数.

2.3 模拟对接软件的使用

为了尽量弥补单一软件模拟分子对接的缺陷和不足, 本文尝试了 3 种蛋白质与短肽对接软件, Autodock Vina^[16]、Ledock^[17] 和 Surflex-Dock^[18]. 为了调试这些软件和证明本文对所有软件的使用和参数设置是正确的, 本文选取了含有 HLA-A2/peptide 复合体的晶体结构文件 (pdb code: 1hhk), 分别使用这 3 种软件把短肽 (c 链) 对接回 HLA-A2 分子 (a 链). 如果软件预测的短肽构象相对于其原始构象的均方根偏差 (Root mean square deviation, RMSD) (所有原子) 不大于 2.2Å, 说明参数设置正确, 而且软件预测的短肽构象精确度较好^[17]. 对于 Surflex-Dock, 本文直接参考它给出的短肽预测构象相对于其原始构象的 similarity 值, 如果 similarity 值大于 0.8, 表明预测结果较好.

使用 Autodock Vina、Ledock 的默认参数, 设置合适大小的结合位点盒子, 设置 Autodock vina、Ledock 给出的最大配体构象数是 20. 对于 Surflex-Dock, 受体蛋白质力场使用 “Kollman All-Atom”、电场使用 “Amber”、配体力场使用 “Tripos”、电场使用 “Gasteriger-Hackel”、对接参数 Threshold=0.5、Bloat=0.

本文使用 Autodock Vina、Ledock 预测的短肽构象 RMSD 分别为 1.76Å、2.19Å; 本文使用 Surflex-Dock 给出的短肽构象 similarity 值为 0.82. 说明本文对这些软件的使用是正确的. 所有测试结果数据见于 <http://www.bilologymaths.top/mbtwo/supplement.aspx>.

3 结 果

3.1 基于KSMM算法的TAP亲和力模型

本文把699个TAP配体数据等分成5份,依次取出其中1份作为测试集,剩下4份作为训练集. KSMM计算的结果和已知的亲和力比较得到 Pearson 相关系数 (relevant Pearson's correlation

coefficient, R_p). 表1列出的 R_p 是KSMM在5个测试集上得到的 R_p 平均值. 为了获得最优KSMM模型,需要对KSMM模型所有参数在一定范围内试验性取值,表1列出了这些参数的选值范围和预测表现最佳时的最优值. 由表1最后1列,基于径向基的KSMM (以下简称KSMM-RBF_{TAP}, $R_p=0.89$) 比线性KSMM (以下简称KSMM-LINEAR_{TAP}, $R_p=0.88$) 预测准确度要稍好一些.

Table 1 The performance of KSMM with different Kernel functions

Function type	Kernel function parameter			Scale parameter		Penalty parameter			R_p
	Range	BV (K_1)	BV (K_2)	C_1	C_2	Range	BV (λ)	BV (λ')	
Rbf	$10^{-4} \sim 10^4$	40	40	2.00	0.40	$10^{-2} \sim 10^6$	0.5	0.4	0.89
Linear				1.00	1.00	$10^{-2} \sim 10^6$	1.00	5.00	0.88

BV: Best Value; R_p : relevant Pearson's correlation coefficient.

表2比较了不同方法^[4-8]和KSMM-RBF_{TAP}在不同的TAP亲和力训练集、测试集上预测结果的Spearman相关系数 (Spearman's correlation coefficient, R_s). 其中, KSMM-RBF_{TAP}计算结果的 R_s 值由Origin8.0获得. 表2第1列列出的方法里面CM^[8]、ADM^[5]和TAPRED^[6]的训练集分别是163条、435条和435条数据,测试集含有178条数据. HMM^[4]和TAPREG^[7]的结果是在493条、613条数据集上交叉验证得到的. 从表2可以看出, KSMM-RBF_{TAP}和TAPREG取得的 R_s 都为0.89, 是最高的,要高于CM、ADM、TAPRED和HMM的 R_s , 考虑到KSMM-RBF_{TAP}的训练集比TAPREG的大,本文使用KSMM-RBF_{TAP}得到未知短肽的TAP亲和力预测值可靠性更高.

Table 2 Spearman's correlation coefficients between observed TAP affinities and predictive TAP affinities of different methods

Method	R_s	Reference
KSMM-RBF _{TAP}	0.89	This study
CM	0.87	[8]
HMM	0.88	[4]
TAPREG	0.89	[6]
ADM	0.74	[5]
TAPPRED	0.67	[6]

R_s : Spearman's correlation coefficient.

3.2 KSMM-RBF_{TAP}在测试集A2-immunogenic、IEDB-0301和A3-immunogenic的预测表现

首先,本文比较了KSMM-RBF_{TAP}和KSMM-

LINEAR_{2,1}识别A2限制性免疫原性短肽的能力. 如图1a所示,在测试集A2-immunogenic上,KSMM-LINEAR_{2,1}的区分度相当高, AUC值为0.873; KSMM-RBF_{TAP}没有任何区分度, AUC值为0.491, 相当于随机预测. 本文后面关于KSMM-RBF_{TAP}及KSMM-LINEAR_{TAP}的计算结果是在660个TAP配体亲和力数据集上训练KSMM得到的.

其次,本文比较了KSMM-RBF_{TAP}和KSMM-LINEAR_{3,1}识别A3限制性结合肽的能力. 如图1b所示,在测试集IEDB-0301上, KSMM-RBF_{TAP}的区分度一般, AUC值为0.672; KSMM-LINEAR_{3,1}的区分度很高, AUC值为0.918.

最后,本文比较了KSMM-RBF_{TAP}和KSMM-LINEAR_{3,1}识别A3限制性免疫原性肽的能力. 如图1c所示,在测试集A3-immunogenic上, KSMM-RBF_{TAP}和KSMM-LINEAR_{3,1}区分度相当, AUC值分别为0.842和0.798.

KSMM-RBF_{TAP}在测试集A2-immunogenic上没有显示出区分度,不在该测试集进行联合预测.

KSMM-RBF_{TAP}在测试集IEDB-0301和A3-immunogenic显示出了明显的区分度,本文在这两个测试集上结合KSMM-RBF_{TAP}和KSMM-LINEAR_{3,1}. 本文选择KSMM-RBF_{TAP}在TAP亲和力测试集上特异性 (特异性=预测正确的负样本数/预测为负样本总数) 84.4%和敏感度 (敏感度=预测正确的正样本数/预测为正样本总数) 90.2%时对应的阈值. 相对而言,使用敏感度约90%时对应的阈值,可以显著增加特异性而使敏感度略有下降. 对于TAP预测值低于阈值的短肽, KSMM-LINEAR_{3,1}

预测值加上-2.6即为联合预测值; 因为KSMM-LINEAR_{3,1}预测短肽亲和力取值范围从-2.6到2.6, 这保证过滤掉的短肽得到一个负分数. 其余短肽联合预测分数是KSMM-LINEAR_{3,1}预测分数.

在测试集IEDB-0301上, 联合预测模型AUC值分别为0.831, 区分度比KSMM-LINEAR_{3,1}低(图1b). 在A3-immunogenic测试集上, 联合预测

模型有最佳的区分度, AUC值为0.873(图1c). 在A3-immunogenic上, 在绘制AUC曲线时计算敏感性的整个区间内, 联合预测模型相对于KSMM-LINEAR_{3,1}特异性增加了39.3%(图1c). 在IEDB-0301上, 绘制AUC曲线时计算敏感性的整个区间内, 联合预测模型相对于KSMM-LINEAR_{3,1}特异性减少了36.3%(图1b).

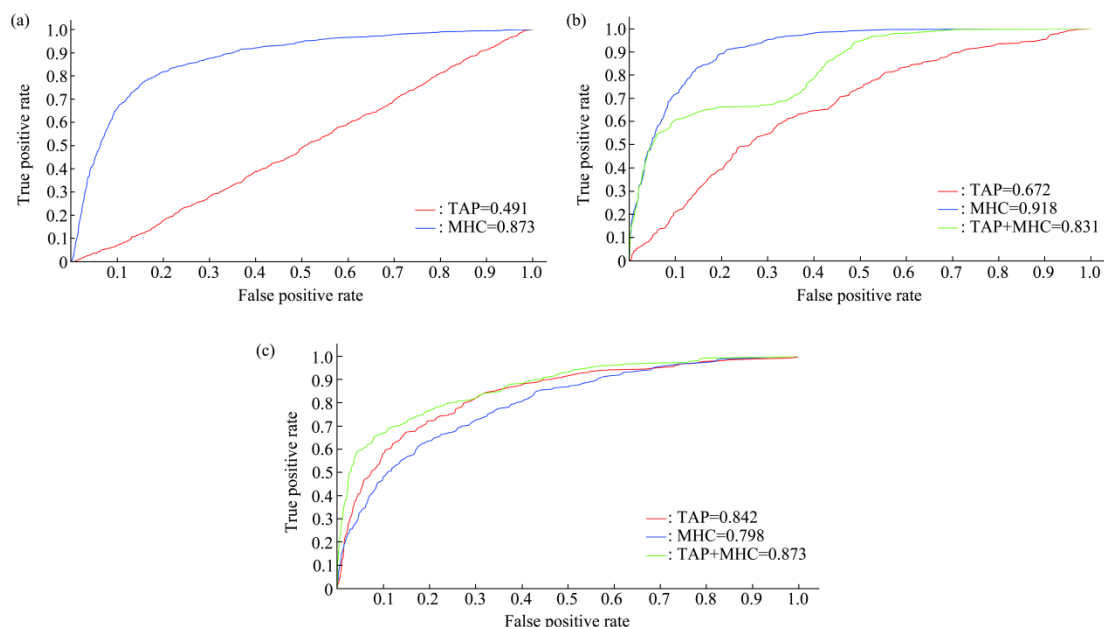


Fig. 1 ROC analysis curves for KSMM-RBF_{TAP} on A2-immunogenic(a), IEDB-0301(b) and A3-immunogenic(c)

3.3 KSMM-LINEAR_{TAP}、KSMM-LINEAR_{3,1}和KSMM-LINEAR_{2,1}线性模型的权重系数

本文使用线性核函数、以(1)式中的 E_i 作为引导函数, 训练KSMM算法得到最优解 w^* , 这个 w^* 各分量和短肽9个残基位置20个氨基酸有一一对应关系: 即 w^* 每个分量的正负和绝对值大小分别表示受体分子对配体每个位置20种氨基酸的偏好性和这种偏好性的强弱. 所以本文通过研究KSMM-LINEAR_{TAP}、KSMM-LINEAR_{2,1}和KSMM-LINEAR_{3,1}的 w^* 各分量来研究TAP蛋白、HLA-A 2.1分子和HLA-A 3.1分子对短肽的选择特异性, 这些 w^* 由图2a~c给出. w^* 各分量的绝对值对应每个位置每个氨基酸所表示符号的高度. 每个分量的符号用氨基酸字符的正置或倒置表示. 每个氨基酸

的理化性质用符号的颜色表示: 黑色表示中性疏水性氨基酸、蓝色表示碱性氨基酸、绿色表示中性极性氨基酸、红色表示酸性氨基酸.

3.4 TAP亲和力预测服务

本文对使用不同核函数的TAP亲和力预测模型KSMM-LINEAR_{TAP}、KSMM-RBF_{TAP}提供了预测服务, 提交页面如图3所示. 如果用户选择预测类型是“peptide”, 程序可以预测长度范围在6~40个氨基酸短肽的TAP亲和力; 如果用户选择预测类型是“protein”, 程序会依次对蛋白质序列依次截取长度为9个连续氨基酸的短肽, 然后分别预测. 用户最好在IE浏览器7及以上版本打开, 其他浏览器有不能正常打开的情况, 设置为兼容模式浏览一般可以正常运行. 网址是 <http://www.bilologymaths.top/mbtwo/major.aspx>.

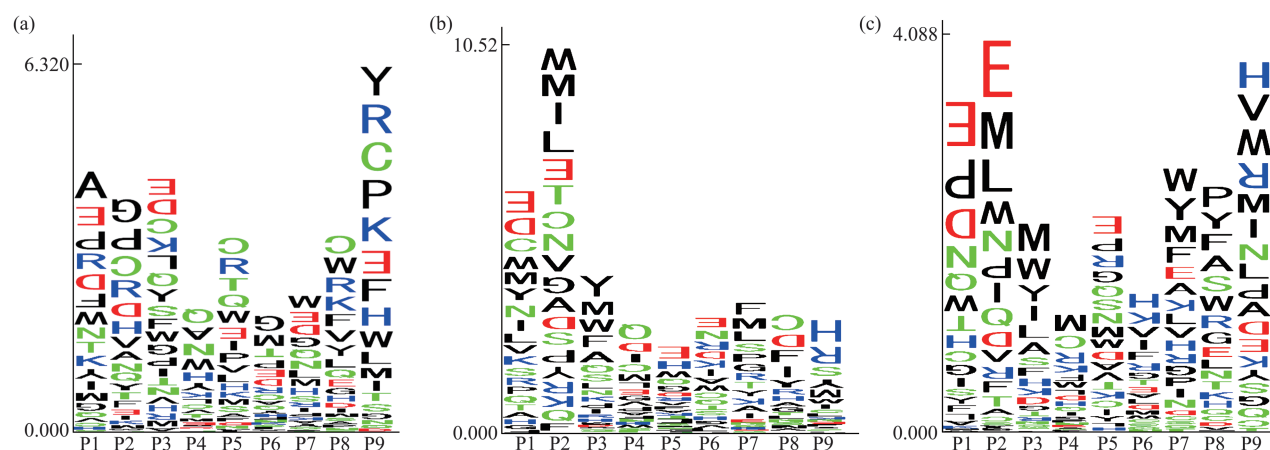


Fig. 2 The weight coefficients of amino acids for KSMM-LINEAR_{TAP}, KSMM-LINEAR_{3.1} and KSMM-LINEAR_{2.1}

(a) TAP; (b) HLA-A 3.1; (c) HLA-A2.1.

SUBMISSION

Fig. 3 Online submission page of TAP binding models

4 讨 论

4.1 线性模型 KSMM-LINEAR_{TAP}、KSMM-LINEAR_{3.1}和KSMM-LINEAR_{2.1}的准确性

KSMM-LINEAR_{TAP}在TAP亲和力测试集上5折交叉检验的 R_p 是0.88(表1); KSMM-LINEAR_{3.1}在HLA-A 3.1分子亲和力测试集IEDB-0301的AUC值为0.918(图1b); KSMM-LINEAR_{2.1}在HLA-A 2.1分子亲和力测试集IEDB-0201的AUC值为0.919; 以上结果表明经过训练得到的线性的KSMM模型可以准确预测TAP或MHC I分子与短肽的亲和力, 这些线性模型的 w^* 每个分量的正负和绝对值大小可以准确反映TAP分子或MHC I分子对短肽每个位置每种氨基酸的偏好性和这种偏好性的强弱(图

2a, 2b和2c). 各个线性模型的 w^* 的值可以在下列网址找到: <http://www.bilologymaths.top/mbtwo/supplement.aspx>.

4.2 具有免疫原性A2、A3限制性抗原肽的序列特征

Frahm等^[19]根据大量已有的A2、A3限制性短肽与T细胞反应数据, 提出了“最佳”A2、A3限制性抗原肽需要满足一定的条件: 当长度为9个氨基酸的短肽与A2、A3分子短肽结合槽时, 短肽的第2、9位置如果是一些特定种类的氨基酸, 形成的MHC I/短肽分子复合物最大可能具有免疫原性. 表3、表4列出了“最佳”A2、A3限制性抗原肽在这两个位置具有的氨基酸种类.

Table 3 Best defined A2-restricted epitopes and components of w^* in KSMM-LINEAR_{2,1} and KSMM-LINEAR_{TAP}

Best defined A2-restricted epitopes	P1—P2—P3—P4—P5—P6—P7—P8—P9	
	L	L
	M	V
Components of $\bar{w}^*$ in KSMM-LINEAR _{2,1}	L 0.42	L 0.20
	M 0.48	V 0.34
	Ave 0.07	Ave -0.05
	L -0.3	L 0.27
	M -0.08	V 0.0

Table 4 Best defined A3-restricted epitopes and the components of w^* in KSMM-LINEAR_{3,1} and KSMM-LINEAR_{TAP}

Best defined A3-restricted epitopes	P1—P2—P3—P4—P5—P6—P7—P8—P9	
	L	K
	V	Y
	M	F
	I	R
	A	
	S	
	T	
Components of $\bar{w}^*$ in KSMM-LINEAR _{3,1}	L 0.74	K -0.07
	V 0.58	Y -0.29
	M 0.74	F 0.07
	I 0.75	R -0.6
	A 0.52	
	S 0.45	
	T 0.70	
	Ave -0.01	Ave -0.04
	L -0.3	K 0.54
	V 0.26	Y 0.64
	M -0.08	F 0.51
	I 0.09	R 0.64
	A 0.2	
	S 0.01	
	T 0.15	

4.3 HLA-A 2.1分子、HLA-A 3.1分子和TAP分子对A2、A3限制性结合肽“最佳”氨基酸的偏好性

表3中, KSMM-LINEAR_{2,1}的 w^* 对应第2位置“最佳”氨基酸(L/M)分量为同一位置所有分量平均值的(6~7倍); 对应第9位置“最佳”氨基酸(L/V)分量为同一位置所有分量平均值的-5~-7倍, 且这些分量均为正值, 这显示HLA-A 2.1分子对结合短肽在这两位置的“最佳”氨基酸偏好性明显; KSMM-LINEAR_{TAP}的 w^* 对应第2位置的A2限

制性“最佳”氨基酸分量数值为负的, KSMM-LINEAR_{TAP}的 w^* 对应第9位置的A2限制性“最佳”氨基酸分量相对于分量平均值没有很大倍数关系, 这显示TAP分子对这两位置的A2限制性“最佳”氨基酸没有明显偏好性.

表4中, KSMM-LINEAR_{3,1}的 w^* 对应第2位置“最佳”氨基酸(L/V/M/I/A/S/T)对应分量平均值为0.63, 最小值为0.45; 而同一位置所有分量平均值为-0.01; 对应第9位置“最佳”氨基酸(K/Y/F/R)分量均为负值, 或者绝对值较小. 这显示HLA-

A 3.1 分子对结合短肽在第2位置“最佳”氨基酸偏好性很强，而HLA-A 3.1分子对结合短肽在第9位置“最佳”氨基酸偏好性较弱。KSMM-LINEAR_{TAP}的 w^* 对应第2位置“最佳”氨基酸分量数值有正值也有负值，且相对于分量平均值没有很大倍数关系，显示TAP分子对第2位置A3限制性“最佳”氨基酸没有明显偏好性；KSMM-LINEAR_{TAP}的 w^* 对应第9位置“最佳”氨基酸分量是所有分量平均值的约2.5倍，且均为正值，显示TAP分子对第9位置A3限制性“最佳”氨基酸有明显偏好性。

4.4 TAP分子与HLA-A 3.1分子对短肽的选择性具有互补性

由图1b可以看到，TAP分子与HLA-A 3.1分子的双选择性比任何一个分子单独的选择性对免疫原性短肽识别精度要高，AUC值从~0.82上升到0.87。提升的原因通过比较表4中KSMM-LINEAR_{3.1}、KSMM-LINEAR_{TAP}的 w^* 对应第2、9位置A3限制性“最佳”氨基酸偏好性差异可以很容易区分出来：KSMM-LINEAR_{3.1}对第2位置A3限制性“最佳”氨基酸偏好性明显，而KSMM-LINEAR_{TAP}对第9位置A3限制性“最佳”氨基酸偏好性明显。

4.5 TAP蛋白与短肽结合位点的确定

PDB文件5u1d给出了TAP蛋白的分子结构^[20]，包含了TAP1链、TAP2链和单疱疹病毒蛋白ICP47。ICP47蛋白含有55个氨基酸，其与TAP蛋白的结合使TAP蛋白失去转运多肽的能力。一般

情况下，ICP47蛋白会优先与TAP蛋白质内“最黏的部分”、用来转运短肽的结合位点结合。本文根据这一假设来确定短肽结合位点：本文首先统计了任何一个原子与ICP47蛋白任何一个原子距离小于5Å的所有TAP蛋白残基，见表5；由于表5列出的TAP蛋白残基都处于TAP2链一侧，本文确定的两个可能的TAP蛋白短肽结合位点也都位于TAP2链一侧，如图4所示；两个可能的短肽结合位点构成残基见表5；本文使用对接软件进行试验性对接，发现TAP蛋白的第2个可能的短肽结合位点和短肽基本没有亲和力，由此确定第1个短肽结合位点是TAP蛋白主要的短肽结合位点。

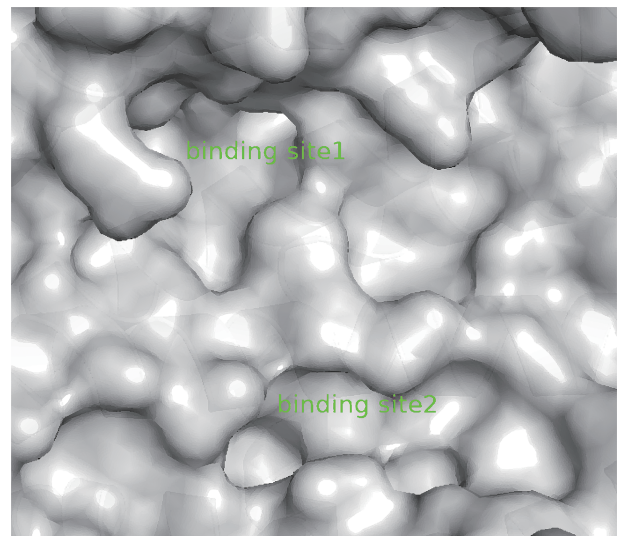


Fig. 4 Possible peptide-binding sites of TAP

Table 5 Possible component parts of peptide-binding sites of TAP

Residues in the peptide-binding site 1	Residues in the peptide-binding site 2	All residues of TAP less than 5Å away from ICP47
GLN350 [#] , LEU352 [#] , GLU353 [#] , VAL354 [#] , GLN355 [#] , VAL356 [#] , ARG357 [#] , GLU358 [#] , SER359 [#] , LEU360 [#] , ALA361 [#] , ILE397 [#] , LYS398 [#] , LEU400 [#] , ASN401 [#] , GLN402 [#] , GLU404 [#] , ALA405 [#] , TYR408 [#] , ASP148 [#] , LEU151 [#] , LEU152 [#] , LEU159 [#] , ARG210 [#] , GLY211 [#] , GLY212 [#] , CYS213 [#] , PHE214 [#] , THR215 [#] , TYR216 [#] , THR217 [#] , MET218 [#] , SER219 [#] , ARG220 [#] , ILE221 [#] , ASN222 [#] , LEU223 [#] , ILE225 [#] , ARG226 [#] , THR258 [#] , SER261 [#] , ASN262 [#] , LEU264 [#] , PRO265 [#] , LEU266 [#] , ALA268 [#] , ASN269 [#]	LEU259 [#] , TRP263 [#] , ASN267 [#] , LEU271 [#] , LEU426 [#] , TYR428 [#] , ILE429 [#] , MEY433 [#] , ASN436 [#]	SER286 [#] , ASP297 [#] , SER298 [#] , GLU301 [#] , LEU305 [#] , TYR309 [#] , TYR349 [#] , GLN350 [#] , GLU353 [#] , ARG357 [#] , GLN365 [#] , ILE368 [#] , SER372 [#] , PRO375 [#] , GLU404 [#] , TYR408 [#] , GLN453 [#] , GLN456 [#] , GLU459 [#] , ARG467 [#] , GLU166 [#] , ARG210 [#] , PHE214 [#] , THR215 [#] , MET218 [#] , LEU238 [#] , PHE241 [#] , GLN242 [#] , THR244 [#] , LYS245 [#] , THR246 [#] , GLY247 [#] , GLU248 [#] , LEU249 [#] , ASN250 [#] , SER251 [#] , ARG252 [#] , SER254 [#] , SER255 [#] , ASP256 [#] , THR258 [#] , LEU259 [#] , ASN262 [#] , PRO265 [#] , LEU266 [#] , ASN267 [#] , ASN269 [#] , VAL270 [#] , ARG273 [#] , HIS314 [#] , ARG373 [#] , LEU377 [#] , ARG380 [#] , ARG381 [#] , HIS384 [#] , GLU417 [#] , SER421 [#] , TYR422 [#] , THR425 [#] , TYR428 [#] , ILE429 [#] , ASP432 [#] , SER435 [#] , ASN436 [#] , LYS442 [#] , TYR477 [#] , TYR520 [#]

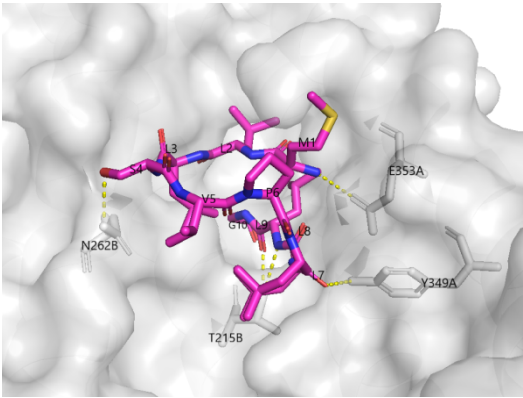
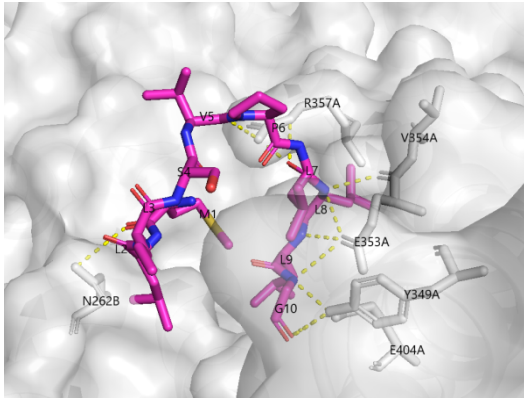
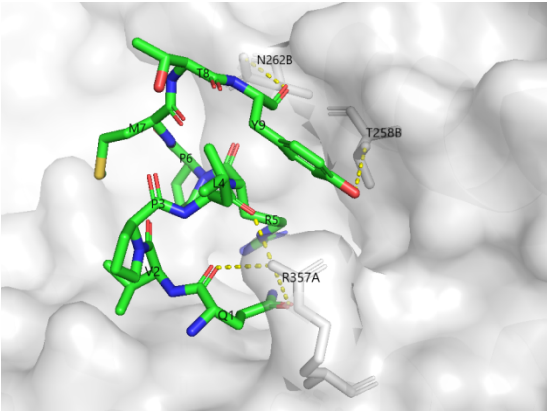
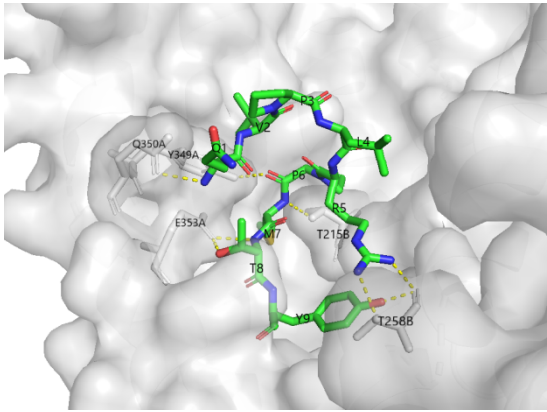
[#] TAP1; [#] TAP2.

4.6 TAP蛋白与短肽的模拟对接

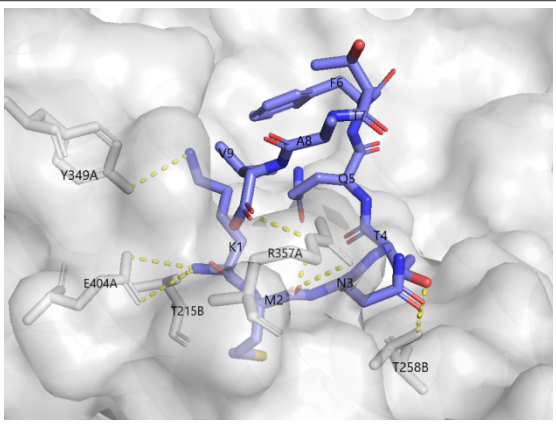
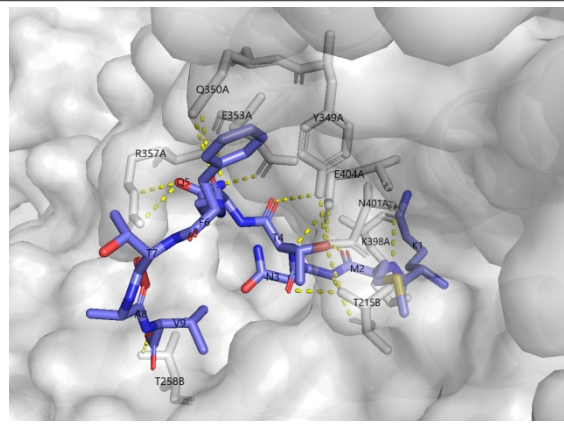
根据已经找到的第1个短肽结合位点, 本文分别利用3种对接软件对TAP蛋白、3条短肽分别进行模拟对接; 本文使用的参数和设置与测试对接时相同. 3条短肽经KSMM-RBFTAP测试, 有足够大的TAP分子亲和力. TAP蛋白晶体结构文件取自pdb文件5u1d; 3条短肽晶体结构文件分别取自包含MHC分子/短肽复合物的pdb文件2clr、3lr2和5h94. 表6列出了使用Autodock vina和Ledock实现

的TAP蛋白与3条短肽的模拟对接结果, 其中的黄色虚线表示可能发生的氢键作用. 每次配体和TAP蛋白模拟对接, Autodock vina和Ledock都会生成多个配体构象. 本文最优构象选取原则是前两个能量最低的构象中形成氢键较多的那个. 表6中对接结果表明, 短肽容易在碳端和氮端和TAP蛋白形成氢键作用. Autodock vina、Ledock和Surflex-Dock的对接结果见于<http://www.bilologymaths.top/mbtwo/supplement.aspx>.

Table 6 TAP–ligand docking predictions by Autodock vina and Ledock

Autodock vina	Ledock
	
Ligand, MLLSVPLLLG; Extracted from PDB, 2clr; Predictive binding affinity to TAP, 0.82(bindners)	Ligand, MLLSVPLLLG; Extracted from PDB, 2clr; Predictive binding affinity to TAP, 0.82(bindners)
Predictive score by vina, −2.7 kcal/mol;	Predictive score by Ledock, −2.74 kcal/mol;
Selected pose in the index, the second;	Selected pose in the index, the second;
Residues forming possible hydrogen bonds with TAP, M1, L2, L7, G10	Residues forming possible hydrogen bonds with TAP, M1, V5, L7, L8, L9, G10
	
Ligand, QVPLRPMTY; Extracted from PDB, 3lr2; Predictive binding affinity to TAP, 0.48(bindners)	Ligand, QVPLRPMTY; Extracted from PDB, 3lr2; Predictive binding affinity to TAP, 0.48(bindners)
Predictive score by vina, −4.36 kcal/mol;	Predictive score by Ledock, −2.89 kcal/mol;
Selected pose in the index, the first;	Selected pose in the index, the second;
Residues forming possible hydrogen bonds with TAP, Q1, L4, Y9	Residues forming possible hydrogen bonds with TAP, Q1, P6, M7, T8, Y9

Continued to Table 6

Autodock vina	Ledock
	
Ligand, KMNTQFTAV; Extracted from PDB, 5h94; Predictive binding affinity to TAP, 0.7(bindners)	
Predictive score by vina, -3.2 kcal/mol;	Predictive score by Ledock, -3.36 kcal/mol;
Selected pose in the index, the second;	Selected pose in the index, the second;
Residues forming possible hydrogen bonds with TAP, K1, M2, N3, T4, V9	Residues forming possible hydrogen bonds with TAP, K1, M2, N3, T4, Q5, V9

5 结 论

本文建立类TAP亲和力模型, 预测精确度比同类算法好, TAP亲和力可在线预测: <http://www.bilologymaths.top/mbtwo/major.aspx>. 利用线性模型的 w^* 分析了HLA-A 2.1分子、HLA-A 3.1分子和TAP分子对短肽P2、P9位置的选择性如何影响提呈短肽的免疫原性, 最后对TAP分子和短肽进行了模拟对接。

参 考 文 献

- [1] Smith K D, Lutz C T. Peptide-dependent expression of HLA-B7 on antigen processing-deficient T2 cells. *J Immunol*, 1996, **156**(10): 3755-3764
- [2] Hinz A, Tampé R. ABC transporters and immunity: mechanism of self-defense. *Biochemistry*, 2012, **51**(25): 4981-4989
- [3] Abele R, Tampé R. The ABCs of immunology: structure and function of TAP, the transporter associated with antigen processing. *Physiology (Bethesda)*, 2004, **19**(4): 216-224
- [4] Zhang G L, Petrovsky N, Kwoh C K, *et al.* PredTAP: a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome Res*, 2006, **2**: 3
- [5] Doytchinova I, Hemsley S, Flower D R. Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatics evaluation. *J Immunol*, 2004, **173**(11): 6813-6819
- [6] Bhasin M, Raghava G P. Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci*, 2004, **13**(3): 596-607
- [7] Díez-Rivero C M, Chenlo B, Zuluaga P, *et al.* Quantitative modeling of peptide binding to TAP using support vector machine. *Proteins*, 2010, **78**(1): 63-72
- [8] Peters B, Bulik S, Tampe R, *et al.* Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J Immunol*, 2003, **171**(4): 1741-1749
- [9] Tang S T, van Meijgaarden K E, Caccamo N, *et al.* Genome-based in silico identification of new *Mycobacterium tuberculosis* antigens activating polyfunctional CD8⁺ T cells in human tuberculosis. *J Immunol*, 2011, **186**(2): 1068-1080
- [10] Mukherjee G, Chaparro R J, Schloss J, *et al.* Glucagon-reactive islet-infiltrating CD8 T cells in NOD mice. *Immunology*. 2015, **144**(4): 631-640
- [11] Lata S, Bhasin M, Raghava G P. MHCBN 4.0: A database of MHC/TAP binding peptides and T-cell epitopes. *BMC Res Notes*, 2009, **2**(1): 61-66
- [12] Vita R, Zarebski L, Greenbaum J A, *et al.* The Immune Epitope Database 2.0. *Nucleic Acids Res*, 2010, **38**(Database issue): D854-862
- [13] Rammensee H, Bachmann J, Emmerich N P, *et al.* SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 1999, **50**(3-4): 213-219
- [14] Vigneron N, Stroobant V, Eynde V, *et al.* Database of T cell-defined human tumor antigens: the 2013 update. *Cancer Immunol*, 2013, **13**(3): 15-20
- [15] Kuiken C, Thurmond J, Dimitrijevic M, *et al.* The LANL hemorrhagic fever virus database, a new platform for analyzing biothreat viruses. *Nucleic Acids Res*, 2012, **40**(Database issue): D587-592

- [16] Trott O, Olson A J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 2010, **31**(2): 455-461
- [17] Zhao H, Caflisch A. Discovery of ZAP70 inhibitors by high-throughput docking into a conformation of its kinase domain generated by molecular dynamics. *Bioorg Med Chem Lett*, 2013, **23**(20): 5721-5726
- [18] Jain A N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem*, 2003, **46**(4): 499-511
- [19] Frahm N, Linde C, Brander C. Identification of HIV-derived, HLA class I restricted CTL epitopes: Insights into TCR repertoire, CTL escape and viral fitness. *HIV Molecular Immunology*, 2006: 3-28
- [20] Oldham M L, Grigorieff N, Chen J. Structure of the transporter associated with antigen processing trapped by herpes simplex virus. *eLife*, 2016, **5**: e21829

Immunogenicity Prediction of The Peptides Presented by MHC I Molecules Based on The TAP Binding Affinity Model*

SHENG Hao^{1)**}, LU Yu-Feng²⁾, BAO Yong-Ming³⁾

⁽¹⁾College of Sciences, Liaoning Shihua University, Fushun 113001, China;

²⁾School of Mathematical Sciences, Dalian University of Technology, Dalian 116023, China;

³⁾School of Life Science and Biotechnology, Dalian University of Technology, Dalian 116023, China)

Abstract The major histocompatibility complex (MHC)I binding affinity models have contributed to screen the candidate peptides, and have assisted the experiments in determining the peptides that can form complexes with MHC I molecules to activate cytotoxic T cells. The transporter associated with antigen processing (TAP) binding models could also be used for screening the candidate peptides. How to make the best of the two types of binding affinity models for screening out the candidate peptides, the similarities and differences between the selectivity of TAP and MHC I molecules in peptides and the biological mechanism of that similarities and differences, these three questions remains obscure. Herein, we rearranged the TAP binding test set, increasing its size to 699. The established TAP binding model based on Kernel-function stabilized matrix method (KSMM) had a higher prediction accuracy than that of the state of the art, achieving a relevant correlation coefficient of 0.89 on a 5-fold cross-validation. The integrative prediction of HLA-A3 affinity and TAP affinity models remarkably improved the discrimination accuracy, with the AUC value increasing from ~0.82 to 0.87. This improvement is due to the different preferences of the two types of affinity models for the best defined amino acids on 2nd and 9th positions, as well as to the complementarity of such different preferences. The results of TAP-peptide docking also supported this conclusion. The TAP model is available online: <http://www.bilologymaths.top/mbtwo/major.aspx>.

Key words immunogenicity, cytotoxic T lymphocytes, MHC I, TAP, Kernel-function stabilized matrix method

DOI: 10.16476/j.pibb.2019.0013

* This work was supported by grants from The National Natural Science Foundation of China (61602228), the Natural Science Foundation of Liaoning (L2017LQN03) and Scientific Research Foundation of Liaoning Shihua University (1100140109).

** Corresponding author.

Tel: 15898372769, E-mail: shenghao@lnpu.edu.cn

Received: May 8, 2019 Accepted: December 26, 2019