氨基酸广义疏水标度(GH-scale)用于 HLA-A*0201 限制性 CTL 表位定量预测

周 鹏[①②] 李志良[①②*] 田菲菲[①] 张梦军[③]

(① 重庆大学化学化工学院, 重庆 400044; ② 化学生物传感与计量学国家重点实验室, 长沙 410082; ③ 第三军医大学医学检验系, 重庆 400040. * 联系人, E-mail: zlli2662@163.com)

摘要 将天然氨基酸 149 个疏水性质经主成分分析得到了一种新氨基酸描述子——氨基酸广义疏水标度(GH-scale). 用GH-scale结合遗传偏最小二乘(GPLS)算法对 152 个HLA-A*0201 限制性CTL表位进行定量构效关系(QSAR)研究. 所建模型拟合及交叉检验复相关系数分别为 $R^2_{cum} = 0.813$和$Q^2 = 0.725$. 研究表明, 疏水作用在CTL表位与HLA-A*0201 结合过程中扮演极其重要角色, 而锚定残基是该类作用发生最显著的部位.

关键词 广义疏水标度 **HLA-A*0201** **CTL** 表位 定量构效关系 遗传偏最小二乘

人类主要组织相容性复合体(MHC)又称为白细胞抗原(HLA)体系, 共分为 3 类, 其中Ⅰ型分子包括HLA-A, HLA-B和HLA-C, 广泛存在于各种组织细胞中. 在Ⅰ型HLA中, 位于HLA-A基因座 0201 型等位基因在人群中分布极为广泛[1], 其表达产物HLA-A*0201 在病毒[2]和肿瘤[3]抗原呈递过程中起非常重要的作用. 通常与HLA-A*0201 发生结合的细胞毒性T淋巴细胞(CTL)表位长度被限制为 9±1 个氨基酸残基, 其第 2 和 9 位残基为锚定残基, 负责与HLA-A*0201 键合[4]. 研究发现, 除第 2 和 9 位点外, 第 1, 3 和 7 位残基在结合过程中也起着重要作用, 从而被称为第二锚定残基[5]. 事实上, 一段由水解酶切割产生蛋白质序列能被HLA-A*0201 识别并有效提呈的必要条件除出现特定锚定残基外, 非锚定残基的性质也很大程度上影响着该过程的进行. 由此可见, CTL表位与HLA-A*0201 结合涉及众多复杂物理化学因素, 尽管至今已有大量肽段及亲和活性被合成和测试, 但驱动抗原肽-MHC复合物相互作用的本质仍未阐明. 近年来, 随着肽库和计算机技术发展, 人们开始求助于计算机模拟途径对抗原肽-MHC复合物结合过程进行理论和统计分析, 并在CTL表位预测方面取得长足进步: 从简单基序、延展基序发展到量化基序, 同时由于各种智能算法如遗传算法、神经网络和隐马尔可夫等方法引入使得预测结果有较大提高[6~8]. 纵观目前各类算法在CTL表位鉴定方面的应用, 不难发现大多数预测还处于定性或半定量阶段, 其结果假阳性较高, 同时由于方法本身的局限性, 使其在基于结构疫苗改造和设计上很难发挥较大作用.

研究表明, 疏水作用在肽/蛋白质复合体系形成并维持其空间构象中扮演着极为重要的角色[9,10]. 疏水作用与氢键作用、静电作用、范德华作用同属非键作用, 但人们对后 3 种相互作用研究已比较深入, 已有得到普遍认可和广泛使用的原子水平势函数表达式. 而疏水作用研究则进展缓慢, 至今仍是一个热点和难点. 由于疏水作用是一种间接熵效应, 其确切势能函数及作用形式还很难被规范化地表达出来, 因此对其恰当描述可通过经验途径来实现. 自从Tanford等人[11]为疏水作用存在提供实验数据以来, 至今已有数百种与氨基酸残基水溶性相关参数被相继提出. 这些实验测量或理论估算疏水参数在很大程度上包含各类残基在不同情况下或相同情况不同方面所表现出的疏水效应, 具有一定实用价值. 然而由于这些参数的多样性和复杂性使得实际使用过程中显得非常不便, 且从统计学角度来看由于数据间含有大量重叠和干扰信息使得具体应用包含极大不确定性因素. 鉴于此, 本文收集了 149 种文献和数据库所报道的氨基酸疏水性指标及残基溶解状态参数, 并通过经典多维数据处理技术主成分分析(PCA)对原始变量信息压缩和提取, 得到一种氨基酸残基综合疏水指数——氨基酸广义疏水标度(GH-scale). 尝试使用GH-scale研究 152 个HLA-A*0201 限制性CTL表位定量构效关系(QSAR), 结果表明, CTL表位与HLA-A*0201 结合在很大程度上受疏水驱动并在锚

定残基部位发生显著键合; 同时还发现在某些残基位点疏水效应并不明显, 可能涉及较多其他物化因素. 该结论为HLA-A*0201 分子对抗原肽提呈机理认识及相关疫苗结构的改造和设计提供了重要参考依据.

1 原理和方法

(ⅰ) 氨基酸广义疏水标度. 处于溶液环境中的氨基酸残基与水分子发生相互作用会表现出多种效应, 这些效应部分可能直接体现在残基的亲水疏水性上, 如球蛋白中亲水残基处于蛋白质表面, 而疏水残基则趋向于包埋在内部; 某些则是从氨基酸残基的其他性质中间接体现出来, 如电离度、等电性、溶解自由能、空间构象、序列柔性等. 为了综合考虑处于溶液状态氨基酸残基的疏水性指标及残基溶解状态参数, 收集了 149 种有关参数. 它们主要反映氨基酸残基以下的一些疏水信息: 溶解自由能变化、分配系数、色谱保留指数、疏水矩、残基埋藏度、溶剂可及范围等; 同时还包括部分与残基溶解状态有关指数如等电点、整体柔性、极化效应等. 可以看到这些参数并不仅是单纯意义上残基疏水能力, 而涉及残基疏水效应有关各个方面, 故这里统称为氨基酸广义疏水性质.

由于不同性质间可能存在较大信息重叠且直接使用 149 个参数来表征序列中单个残基将造成应用过于复杂、干扰因素太多等问题, 因此结合经典多维数据处理技术主成分分析(PCA)进行变量维数压缩提取. 先按列对原始变量矩阵$X_{20\times149}$自定标处理, 继而采用PCA得到 12 个显著主成分, 其累积解释方差为 95.03%; 分别解释原始变量矩阵 52.53%, 9.36%, 7.39%, 6.51%, 3.49%, 3.37%, 3.03%, 2.37%, 2.00%, 1.81%, 1.67%和 1.50%方差. 这 12 个显著主成分得分是由氨基酸 149 个原始变量值与每个主成分得分系数乘积计算而来, 当使用这 12 个得分矢量来替代原始变量矩阵时仅损失 4.97%信息. 本文称 20 种氨基酸的 12 个显著主成分为氨基酸广义疏水标度(GH-scale), 该标度提取了氨基酸残基处于溶液环境的相关性质信息, 其具体数值参见表 1. 图 1 是 20 个天然氨基酸在前两个主成分上得分散点图, 该图包含 149 个原始变量 61.89%信息. 从中可看到氨基酸分布特征具很强规律性: 异亮氨酸(Ile)、亮氨酸(Leu)、缬氨酸(Val)、苯丙氨酸(Phe)等具强疏水性氨基酸集中在该图右边; 精氨酸(Arg)、赖氨酸(Lys)、天冬氨酸(Asp)、谷酰氨酸(Gln)等强亲水性氨基酸趋向于分布在图左边; 其他如甘氨酸(Gly)、脯氨酸(Pro)、酪氨酸(Tyr)中性氨基酸则处两者间.

表 1 氨基酸广义疏水性标度值

氨基酸	GH_1	GH_2	GH_3	GH_4	GH_5	GH_6	GH_7	GH_8	GH_9	GH_{10}	GH_{11}	GH_{12}
Ala A	1.568	–3.738	–1.933	1.285	1.280	1.471	1.923	–1.156	–0.624	–1.522	–0.887	0.863
Arg R	–11.806	9.935	–5.359	–3.279	–1.087	–0.199	4.223	0.449	–0.065	0.458	0.434	–1.109
Asn N	–8.009	–0.717	0.269	1.081	1.450	–2.801	–2.900	1.552	0.568	–2.482	1.777	–1.496
Asp D	–10.748	–2.635	6.278	–4.884	1.392	–0.987	1.463	–0.434	1.088	2.284	2.019	2.941
Cys C	6.565	–4.088	–3.663	–7.783	–4.217	2.498	–1.578	1.302	1.392	–1.320	0.458	0.062
Gln Q	–8.514	–0.483	0.127	0.728	–0.220	–0.182	–2.265	–0.805	2.106	–0.867	–1.533	–1.387
Glu E	–10.406	–3.013	5.833	–3.097	1.976	2.163	1.249	0.117	–2.131	–0.646	–1.603	–3.303
Gly G	–1.898	–5.684	–4.217	3.547	1.765	–0.399	3.216	2.199	0.761	0.594	–1.405	0.249
His H	–1.835	1.990	–1.436	–1.840	–0.069	–2.456	–2.543	–0.931	–5.102	–1.111	–0.827	2.116
Ile I	13.124	0.808	0.998	0.988	1.962	0.627	0.469	3.577	–0.133	0.164	3.108	–0.468
Leu L	11.454	0.092	–0.278	1.276	0.548	1.811	–0.322	–2.362	–0.306	2.208	–0.591	–1.032
Lys K	–11.794	4.118	–1.872	3.603	2.566	5.347	–3.318	–0.041	0.600	0.569	0.818	1.527
Met M	9.954	0.232	–1.628	–2.617	1.441	0.400	–1.497	–3.585	1.118	1.686	–0.070	–0.765
Phe F	12.502	2.755	0.834	0.317	1.668	–3.145	0.065	–1.229	–0.084	–0.368	1.672	–1.508
Pro P	–2.142	–1.032	3.440	5.468	–6.327	1.918	1.494	–1.620	–1.457	–0.361	2.273	–0.402
Ser S	–5.949	–3.304	–2.798	1.730	–1.260	–2.497	–0.974	–0.508	1.221	–0.395	0.353	0.625
Thr T	–3.829	–2.440	–1.886	1.710	–0.940	–3.473	0.823	–1.214	0.282	1.893	–0.463	0.034
Trp W	9.461	4.383	4.168	0.631	0.476	0.123	2.312	–1.028	2.277	–3.803	–1.799	2.107
Tyr Y	3.241	4.247	4.441	1.179	–2.920	–1.338	–2.182	3.329	0.892	2.414	–2.916	0.022
Val V	9.062	–1.429	–1.319	–0.042	0.515	1.119	0.340	2.386	–2.406	0.605	–0.816	0.925

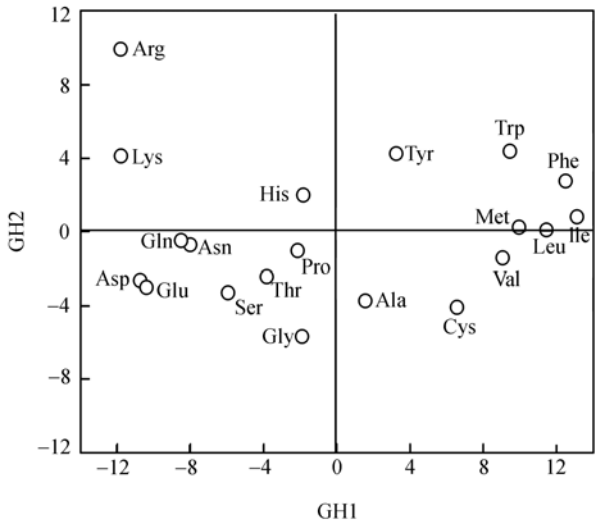

图 1 氨基酸在前两个主成分上的得分分布散点图

(ⅱ) HLA-A*0201 限制性 CTL 表位实验数据. 采用的 152 个人类HLA-A*0201 限制性CTL表位皆为 9 肽, 其具自由的N和C末端, 取自文献[12]. 抗原肽(CTL表位)氨基酸序列及与HLA-A*0201 亲和性 pIC50 参见文献[12], 其IC_{50}为不同剂量待测肽与 0.5 nmol/L放射性标记HBVc18227 (FLPSDYEPSV) CTL 表位(对照)肽/HLA-A*0201 复合物在室温下共孵育 2 h, 测定待测肽序列将对照肽/HLA-A*0201 复合物中 50 %对照肽的置换浓度. 图 2(a)展示Garboczi等人[13]采用X射线衍射方法测得的抗原肽序列LLFGYPVYV与HLA-A*0201 所形成复合物三维晶体结构(PDB码: 1AO7). 从该图中可清楚地看到由两个α螺旋和一个β片层组成HLA-A*0201 肽结合沟槽, 而抗原肽则嵌于其中呈舒展状态. 图 2(b)为从复合物中剥离出抗原肽立体结构, 该分子所有残基皆为反式构象, 从而使得侧链间距达到最大, 未出现明显扭曲现象. 通过上述分析可知处于结合状态抗原肽立体结构受 HLA-A*0201 影响较小, 位于低能构象; 而决定抗原肽与 HLA-A*0201 结合强弱关键在于抗原肽残基与附近HLA-A*0201残基作用大小, 其自身残基间相互影响较弱. 由于受体 HLA-A*0201 可认为不变, 因而抗原肽序列各位点残基差异导致它们间亲和活性不同.

(ⅲ) CTL表位亲和活性QSAR建模. Tropsha等人[14]的研究结果显示, 交叉检验Q^2值与模型预测能力并没明显直接关系, 对模型预测能力评价只能通过外部样本即测试集来进行. 由此将 152 个抗原肽样本集随机划分为一含 102 个样本的训练集和一含 50 个样本的测试集, 并用测试集对训练集所建模型验证. 对一组肽类似物, 每个位置上残基疏水特征可由 12 个GH-scale描述子所表征. 当使用GH-scale对抗原九肽表征时, 共产生 108 个GH-scale描述子, 我们用$V_{1\sim108}$来分别表示, 其中$V_{1\sim12}$依次为位置 1 上 12 个 GH-scale, $V_{13\sim24}$依次为位置 2 上 12 个GH-scale, 以此类推. 由于抗原肽不同残基及用于描述同一残基不同变量对亲和性贡献不同, 因此在建立QSAR模型前需作变量筛选以期提高模型质量并降低模型复杂度. 考虑到遗传算法是一种对复杂组合优化问题具很强全局搜索能力的非数值智能优化算法, 并已在许多大规模变量挑选中得到较好应用, 同时鉴于采用GH-scale对九肽表征将产生众多分子描述子, 通常变量筛选法难得到最佳变量子集, 故用遗传偏最小二乘(GPLS)技术用于确定QSAR模型变量组成. 该过程由 Gaot_Toolbox V5.0 与 PLS_Toolbox V3.0 基于 Matlab 7.0 环境实现.

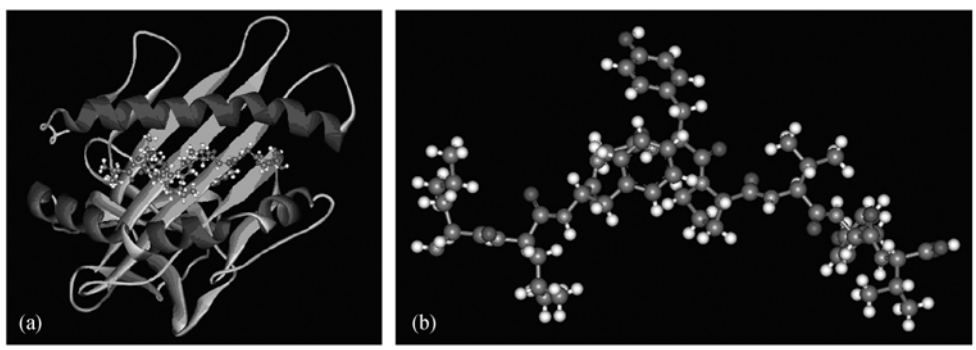

图 2 抗原肽/HLA-A*0201 复合物三维晶体结构及从中剥离抗原肽立体式构象

(a) 利用 X 晶体衍射技术测得的 LLFGYPVYV/HLA-A*0201 复合物的三维晶体结构; (b) 从复合物中剥离出的抗原肽分子立体结构

2 结果及分析

用化学计量学软件包Simca-p 10.0 对建模结果作深入数据挖掘，所得模型相关统计量为：R^2_{cum} = 0.813, Q^2 = 0.725, RMSEC = 0.375. 经分析表明, PLS 模型采用 4 个主成分累积解释筛选所得 20 个变量 82.56%方差，特别是第一个主成分，其解释绝大多数方差值(51.34%)，故用前两主成分绘制 102 个训练集样本得分布点图(图 3). 可看到不同活性样本有规律地分布在前两主成分空间上，活性低的样本主要集中在图左下角，而活性高的则集中在图右上方，活性中等样本居于两者之间. 绝大多数样本都落在该图 95%置信度Hotelling T^2椭圆置信圈内，仅 2 号样本点超出该范围，观察发现该样本是惟一在第 2 位锚定残基处出现半胱氨酸(Cys)的抗原肽，而半胱氨酸侧链存在巯基，在正常生理环境下有一定极性，表现出部分亲水特征，而通常该位点锚定残基为非极性亮氨酸(Leu)，由此可见该样本异常是由关键残基性质特殊造成. 由于锚定残基不利改变将会直接影响到抗原肽对HLA-A*0201 亲和力，从图 3 中的活性分布趋势也暗示该序列具较低活性，上述结论与实验数据一致. 模型引入 20 个变量呈不均匀状态分散在抗原九肽 9 个位点上，没缺失现象，说明每个位点残基类型均对抗原肽亲和活性有所贡献. 单从每个位点参与建模的GH-scale描述子数目来说，第 5 和 6 号位点所占个数最少，分别仅有一个GH-scale描述子被入选. 事实上，HLA-A*0201 限制性CTL表位除第 2 和 9 位为最关键锚定残基外，第 1, 3 和 7 位也对抗原肽结合受体过程有着较显著影响，被称为第二锚定残基[5]. 由此可见，第 5, 6 位点在模型所占比重偏低是因其对抗原肽亲和活性贡献较少缘故. 进一步从PLS载荷图中(图 4)可看到，在第一个主成分上，第 2, 9 位点所有 GH-scale描述子(2 号: V_{13}, V_{17}, V_{20}, V_{22}; 9 号: V_{105}, V_{108})载荷贡献绝对值都大于 0.3，而在第二个主成分上这些变量同样具有较大载荷(>0.2)，表明锚定残基疏水性直接影响抗原肽与HLA-A*0201 亲和性. 另外，变量V_{28}和V_{73}在第一个主成分上也具较大载荷(>0.3)，分别代表序列第 3 和 7 位残基. 从图 5 中 102 个训练集样本观测与计算值相关情况可以发现 第 2 和 102 号误差很大，为模型异常点. 其中 2 号样为正向误差，异常原因显然是由于第 2 位点半胱氨酸残基特殊性所致，前已做过讨论. 由于其锚定残基发生不利改变，从而导致抗原肽对受体亲和性大为降低，模型虽然也做出相应判断，但计算结果仍未能正确模拟真实活性，从而表现出较大正向偏差. 另外，102 号样为负向偏差，该抗原肽在所有样本中具最大亲和活性，可认为模型计算结果出现较大负向误差情况可能如下: (ⅰ) 实验结果偏高; (ⅱ) 所选模型不合理; (ⅲ) 样本特殊性所致. 很多情况下异常点常被忽略而被排除在模型外. 但这样做存在很大风险即往往将一些有价值信息遗漏. 本研究采取谨慎态度将这两个样本保留在模型中. 使用 PLS 模型对 50 个测试集抗原肽进行预测，从图 5 中可直观看出，预测结果与实验值非常接近，复相关系数 R^2_{pred} =0.760，均

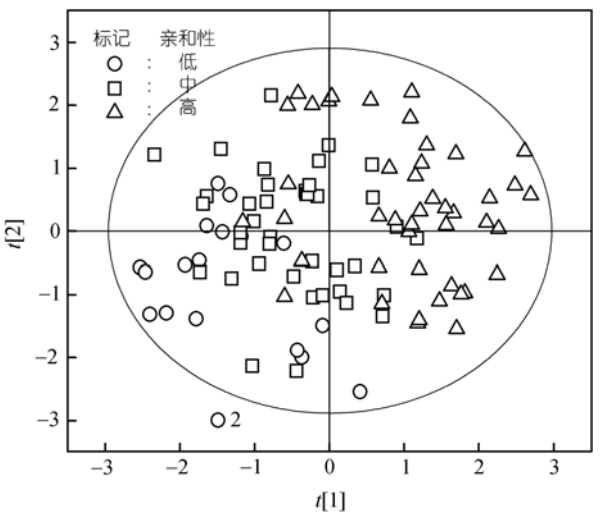

图 3 PLS 模型中 102 个训练集样本在前两个主成分上的得分分布散点图

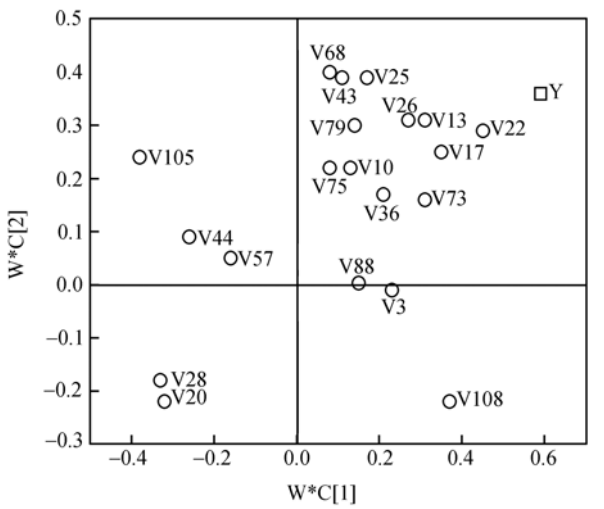

图 4 PLS 模型中 20 个 GH-scale 描述子对前两个主成分的载荷贡献分布情况

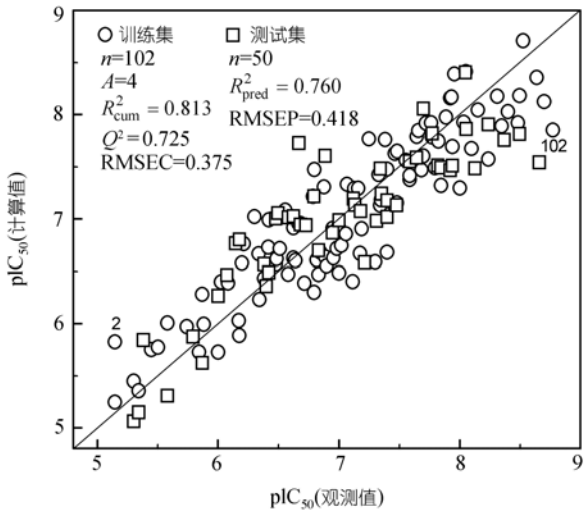

图 5 152 个抗原肽亲和活性的模型计算值与实验观测值相关情况

方根误差 RMSEP = 0.418. 该结果证实本模型具较强的泛化能力, 其对表位鉴定和疫苗改造具有一定指导意义.

3 结语

分子免疫学和结构免疫学的快速发展越来越要求人们从分子水平上认识抗原结构和免疫原性关系. 业已知道蛋白质抗原并非通过整体发挥其功能, 而是通过表位来体现其特异性. 因此研究表位结构与性质功能间的关系对加快疫苗开发有着十分重要的意义. 现已发展了多种表位预测算法, 然而这些方法并没有或很少涉及结构与活性间内在关系, 因此也限制其应用的可拓展性. 本研究从氨基酸疏水性质入手, 通过主成分分析技术得到了一种新氨基酸残基综合疏水指标: 氨基酸广义疏水标度(GH-scale). 基于 GH-scale 将定量构效关系理论和方法引入到 HLA-A*0201CTL 表位定量预测中, 取得较好结果; 研究表明疏水作用在 CTL 表位与 HLA-A*0201 结合过程扮演着极其重要角色, 特别是抗原肽锚定残基与受体键合在很大程度上取决于该类残基疏水性质, 而其他物化因素对非锚定残基贡献较大.

致谢 本工作为化学生物传感与计量学国家重点实验室(湖南大学)基金(批准号: 2005012)、霍英东基金(批准号: 98-8-7)、重庆市应用基础基金(编号: 01-3-6)、教育部“春晖计划”启动基金(编号: 99-4-4+37)、重大自主创新基金(批准号: 03-0506+04-0909)及第三军医大学青年教师自主创新基金(批准号: 05-5-28)资助项目.

参 考 文 献

1 Bodmer J. World distribution of HLA alleles and implications for disease. Ciba Found Symp, 1996, 197: 233—253
2 McMichael A J, Parham P, Brodsky F M, et al. Influenza virus-specific cytotoxic T lymphocytes recognize HLA-molecules. Blocking by monoclonal anti-HLA antibodies. J Exp Med, 1980, 152(2): 195—203
3 Schendel D J, Gansbacher B, Oberneder R, et al. Tumor-specific lysis of human renal cell carcinomas by tumor-Infiltrating lymphocytes. Ⅰ. HLA-A2-restricted recognition of autologous and allogeneic tumor lines. J Immunol, 1993, 151: 4209—4220
4 Falk K, Rötzschke O, Stefanovic S, et al. Allele specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature, 1991, 351: 290—296[DOI]
5 Ruppert J, Sidney J, Celis E, et al. Prominent role of secondary anchor residues in peptide binding to HLA-A*0201 molecules. Cell, 1993, 74: 929—937[DOI]
6 Odunsi K, Ganesan T. Motif analysis of HLA class Ⅱ molecules that determine the HPV associated risk of cervical carcinogenesis. Int Mol Med, 2001, 8(4): 405—412
7 Brusic V, Rudy G, Honeyman G, et al. Prediction of MHC class Ⅱ-binding peptides using an evolutional algorithm and artificial neural network. Bioinformatics, 1998, 14(2): 121—130[DOI]
8 Honeyman M C, Brusic V, Stone N L, et al. Neural network-based prediction of candidate T-cell epitopes. Nat Biotechnol, 1998, 16(10): 966—969[DOI]
9 Tanford C. The hydrophobic effect: formation of micelles and biological membranes. New York: Wiley, 1980
10 Tanford C. How protein chemists learned about the hydrophobic factor. Protein Sci, 1997, 6: 1358—1366
11 Tanford C. The hydrophobic effect and the organization of living matter. Science, 1978, 200: 1012—1018
12 Doytchinova I A, Flower D R. Toward the quantitative prediction of T-Cell epitopes: CoMFA and CoMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. J Med Chem, 2001, 44: 3572—3581[DOI]
13 Garboczi D N, Ghosh P, Utz U, et al. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature, 1996, 384: 134—141[DOI]
14 Tropsha A, Gramatica P, Gombar V K. The importance of being earnest: validation is the absolute essential for successful application and inerpretation of QSPR models. QSAR & Comb Sci, 2003, 22: 69—77[DOI]

(2005-12-05 收稿, 2006-03-21 接受)