

生物信息学在免疫学上的应用

黄文晖 综述 纪志梁* 审阅 (厦门大学生命科学院, 福建 厦门 361000)

[摘要] 生物信息学的快速发展为免疫学的研究提供了新的手段, 如果运用得当, 能够大大加快研究进程, 降低研究费用, 缩短研究周期。兹从免疫学的常用数据库、免疫基因组以及 MHC 和多肽结合预测等 3 个方面, 对生物信息学在免疫学中的应用进行了综述。

[关键词] 生物信息学; 免疫基因组; 免疫信息学; 肽类疫苗; 免疫数据库

[中图分类号] R318.04 **[文献标识码]** A

免疫学在经历了一个世纪的从经验到实验的研究后, 进入一个量化的、基于基因组的时代。像生命科学的其他领域那样, 高通量技术已把免疫学从假说时代带入了数据驱动的研究时代。大量更新的基因组和蛋白质组数据, 在为后基因组时代的免疫学研究提供了丰富资源的同时, 伴随而来的是对海量繁芜数据的困惑。因此, 发展生物信息体系来管理和分析这些数据以提供线索将成为现代免疫学研究的一种必然趋势。我们将对目前生物信息学在免疫学中的应用进行简单的概述。

1 免疫学常用数据库

在过去的几年中, 数据库对于生物医学研究的作用日益显著。目前, 用于免疫学研究的专门数据库仍十分有限, 其中较大型的包括免疫多态数据库 (immunopolymorphism database, IPD, 网址: <http://www.ebi.ac.uk/ipd>) 和国际免疫遗传学数据库 (the international immunogenetics database, MGT, 网址: <http://imgt.cines.fr>)。

IPD 是一组研究免疫基因多态性的专家数据库集合。IPD 项目最初是由 ANRI (Anthony Nolan Research Institute) 的 HLA 信息组所创建^[1], 目前包括 IPD-KIR、IPD-MHC、IPD-HPA 及 IPD-ESTDAB 等 4 个子数据库可供查询。其中, IPD-KIR 序列数据库可提供人类杀伤细胞免疫球蛋白样受体 (KIR) 的完整数据, 包括从 17 个 KIR 基因衍生而来的 89 个 KIR 等位基因 (alleles)^[2]。除此之外, 它还提供了不断更新的命名法 (nomenclature) 和序列比对信息^[3]。IPD-MHC 序列数据库可提供人类以外物种的 MHC 基因序列的数据, 当前主要收集了 4 类

不同物种的 MHC 基因信息, 其中有牛科动物 (Cattle)、犬科动物 (Canines)、猫科动物 (Felines) 以及非人类灵长类动物 (non human primates)。IPD-HPA 数据库收集了人类血小板抗原的基因信息, 包括 24 个血小板特异性异型抗原的等位基因及双等位基因系统 (b allelic system) 组^[4], 并含有异型抗原蛋白质数据、遗传信息、等位基因频率及非 HPA 抗原等 4 张表格可供查询。IPD-ESTDAB (欧洲可搜寻肿瘤细胞系的数据) 提供了一个能够在线查找 HLA 类型、具有免疫特征的肿瘤细胞数据库, 包括从欧洲、澳洲和美国收集的 100 个转移黑色素瘤细胞系^[3], 并可提供 3 种不同的查询方式: 即按照主要搜索决定子 (primary search determinants)、所有搜索决定子 (all search determinants) 和标记词典 (dictionary of markers) 查询。

MGT 在 Ig TcR、MHC 以及人类和其他脊椎动物免疫系统相关蛋白的收集和整理方面具有独到之处, 在医疗研究领域、兽医研究、后天性免疫反应的基因组多样性研究、抗体工程的生物技术以及医学疗法 (器官移植、免疫疗法、疫苗学) 等方面均具有重要的应用前景^[5]。MGT 由序列数据库、基因组数据库和三维结构数据库组成, 其中序列数据库又包括 LIGM-DB、MHC-DB 及 PRIMER-DB 3 个子数据库: LIGM-DB 包含 150 个物种的 Ig 和 TcR 的信息, 2004 年 8 月 10 日止已收集了 86 740 项数据; MHC-DB 包含了 MHC-HLA、-NHR、-DLA 及 -FLA 等信息; PRIMER-DB 包含 11 个物种的 Ig 和 TcR, 2004 年 10 月 28 日止已收集了 1 827 项数据。基因组数据库包含了人类和鼠类的 Ig 和 TcR 基因, 2004 年 9 月 23 日止已收集了 1 378 个基因和 2 205 个等位基因的信息; 三维结构数据库包括 Ig TcR 和 MHC 基因和等位基因的鉴定, 2004 年 9 月 2 日止已收集了 809 项数据。MGT 还内置了丰富的分析工具, 使用户无需运用专业的分析软件, 只要在该网站上就能对数据进行分析整理, 包括序列分析工具 (V-QUEST、Junction Analysis、Allele-A lignment、Phylogenetic)、基因组分析工具 (GeneSearch、GeneView、LocusView) 和三维结构分析工具 (GeneInfo、StructuralQuery)。

除了上述两个较大型的免疫数据库外, 还有一些小型的数据库, 它们收集了可提供免疫学特定方向的信息。如 SYFPEITHI (<http://www.syfpeithi.de/scripts/MHCServer.dll/home.htm>) 和 JENPEP (<http://www.jenner.ac.uk/jenpep2/>) 就是 2 个具有代表性的专门提供 MHC 和肽结合数据的数据库。SYFPEITHI 中包括 2 000 个 MHC 配基以及超过 200 个的结合基序信息。HIV 分子免疫数据库 (HIV Molecule Immunology Database <http://www.hiv.lanl.gov/content/immunology/index.htm>)

收稿日期: 2005- 06- 20 修回日期: 2005- 08- 19

基金项目: 国家自然科学基金资助项目 (No. 30400573)

作者简介: 黄文晖 (1979-), 男, 福建福州人, 硕士生。

Tel: (0592) 2182897; Email: wenhui_h@bioinf.xmu.edu.cn

* Corresponding author. Email: app@bioinf.xmu.edu.cn

mV) 包含 HIV-1 特异性细胞毒性 T 细胞和辅助性 T 细胞表位及抗体结合位点的信息, 有超过 400 个 T 细胞表位和 300 个抗体结合位点可供查询。

2 生物信息学辅助免疫基因组的研究

物种测序工作的逐步展开给免疫学研究者打开了一个新的视野, 在这个基因组的世界里, 蕴含着许多鲜为人知的秘密。在人类和许多其他哺乳动物中, 接近 5% 的基因参与免疫功能, 其中包括免疫球蛋白超家族 (IgSF)、外源凝集素 (lectins)、细胞因子和其受体 (cytokines and their receptors)、趋化因子 (chemokines)、整合蛋白 (integrins)、MHC-I 和-II 类分子、先天性免疫受体 (innate immune receptors) 和补体蛋白等。这些免疫相关基因以多基因、基因簇和基因多态性等方式存在于基因组中, 反映了他们在免疫防御中所扮演的角色。运用比较基因组学的方法, 分析不同物种间免疫基因的演化过程, 可为深入研究某个免疫相关基因或者基因家族提供很好的线索, 在进化分析上也有其积极的意义。

2.1 免疫系统多基因和基因多态性现象的比较基因组分析

早在 1970 年, Ohno 就指出, 现代基因 (modern gene) 是祖先基因 (ancestral gene) 复制的后代。复制提供了一种保留的方式, 即只保留一份其自身有用的拷贝, 而其他的复制子则可通过一系列自由变化的方式 (如突变、重组等) 而获得新的功能^[6]。这种多基因现象是许多免疫系统基因的共同特征, 如白介素家族和人类 4Mbp 的 MHC 区域等^[7]。这些家族中包含多拷贝的彼此相关的复制子, 他们在染色体上紧密排列。如 NK 细胞受体家族在人类染色体 19q 位置上的顺式复制可产生 50 个包括 KIR、LLR 和 CD66 位点的相关序列片段, 其中 KR 位点呈严密连锁状排列^[8]。Annalise 等^[9]利用生物信息的方法比较了 KR 基因簇, 构建了其两个单体型 (haplotypes) 的完整复制模型以及相应的进化事件, 发现 KR 基因的迅速扩增可能是由多轮复制的单个或者多个基因叠加而成; 他们还发现, HLA-I 类分子 α 区内的单个或多个基因片段的串联重复复制不仅相似而且伴随有基因内的重组, 因此, 认为人类的 KR 和 HLA 基因是由同一个祖先进化而来。

在对小鼠基因组序列和人类基因组序列的比较中, 发现与免疫功能相关的基因相对于其他基因来说, 在进化上通常具有很大的变动性^[10], 免疫相关基因在进化上的高变动性是免疫分子在长期不断与入侵免疫系统的病原体作斗争的结果。免疫系统面临强烈、间歇的选择压力, 导致了序列的多态性 (如 MHC 中等位基因的不同) 和频繁的基因复制及遗失事件的发生。有报道称, MHC 中每个基因位点上大约有 500 种不同的等位基因^[11]。MHC 对于免疫应答的启动具有重要的作用且能直接和病原体相互作用, 所以在选择压力下促进了 MHC 的多态性。目前, 针对 MHC 多态性的产生存在两种假说: 对 MHC 异型接合宿主偏爱的选择假说 (selection favoring MHC heterozygous hosts) 以及由于宿主和病原体共进化而

获得的罕见 MHC 等位基因的选择假说 (selection for rare MHC alleles by host-pathogen co-evolution)^[12]。MHC 异型接合宿主偏爱的选择假说认为 MHC 异型接合宿主相比 MHC 同型接合宿主能提呈更多不同的抗原肽, 结果比后者能识别并抵抗更多病原体; 由于宿主和病原体共进化而获得的罕见 MHC 等位基因的选择假说则认为病原体通常能够逃脱最常见 MHC 分子 (the most common MHC molecules) 的提呈, 因此在选择压力下, 宿主罕见 MHC 等位基因的生成频率将不断增加, 而常见 MHC 等位基因的生成频率将逐步减少, 在这个过程中导致动态多态性 (dynamic polymorphism) 的产生, 换言之, 生成罕见 MHC 分子的宿主具有最高的适应度 (highest fitness)。

在用常规实验手段难以证实的情况下, 可通过计算机模拟宿主和病原体的进化过程, 研究这两种假说在导致 MHC 多态性的可能性中的作用中发现异型接合优势本身仅仅轻微地提高了 MHC 的多态性, 而宿主和病原体共同进化的说法则可以解释在每个 MHC 位点上有超过 50 个等位基因的现象^[13]。

2.2 免疫相关基因的鉴定

20 世纪末分子免疫学的兴起, 克隆技术已成为发现新的免疫分子并对其进行功能分析的首选技术。然而, 由于缺少足够的线索, 新的免疫基因的发现举步维艰, 且带有一定的随机性。近几年来, 应用生物信息学的手段从 EST 数据库中寻找免疫同源分子以发现新免疫分子的技术不断成熟, 极大地推动了基础免疫学和临床免疫学的发展。David 等^[14]利用英国曼彻斯特科技大学、诺丁汉大学等提供的鸡的 EST 序列数据库进行了一系列包括聚簇分析、EST 扫描、多重序列比对及系统发生树构造等的分析处理后, 鉴定出在鸡的 TLR 途径中超过半数的免疫分子, 并且以此为基准, 构建了鸡 TLR 免疫球蛋白家族相互作用的网络路径图。Yannick 等^[15]从牡蛎中提取了 1 142 个 EST, 经过同源性分析后, 发现 54% 的序列包含假定功能 (putative function) 的基因, 并将其中的免疫基因根据预测的功能分为 4 组, 分别是蛋白酶和蛋白酶抑制因子组、黏性蛋白组、压力蛋白组及信号转导组, 并鉴定出各组中相应的蛋白。

3 MHC 和抗原肽结合的预测

MHC 对抗原肽的识别与结合是启动适应性免疫应答的关键步骤。因此, MHC-肽的相互作用对于开发抗传染的药物、变态反应及自身免疫病的免疫治疗, 以及癌症疫苗的研制来说都是至关重要的^[16]。传统的优势肽筛选合成法不仅难度高、花费大, 而且效率低^[17]。随着多种 MHC 分子晶体结构的获得和能与 MHC 分子结合的大量内源性肽序列的阐明, 为利用基于生物信息学的模型来预测 MHC 和抗原肽的结合, 并由此筛选理想的表位成为可能。以下简单介绍目前常见的几种利用生物信息学来预测抗原肽和 MHC 相结合的方法, 包括: 结合基序法、量化基序法、人工智能法及人工神经网络 (artificial neural networks ANN)、藏匿马尔科夫模型 (hidden markov models HMM) 和分子建模法。

3.1 结合基序法 结合基序是指当抗原肽与其特异性 MHC 分子 (specific MHC molecule) 相结合时, 经常性地出现在结合位点的氨基酸残基。这些氨基酸残基通常称为锚点氨基酸残基, 而与特异性 MHC 分子相结合的位置称为锚点位置 (anchor positions)。肽结合到 MHC 上主要由初级和二级锚点的氨基酸残基所介导, 如果能依据一级锚点和二级锚点的相对作用来打分, 就可以形成一个评估抗原肽结合能力的粗略模型^[18]。进一步, 如果将这种模型做成经验数据库对于预测表位是很有用的, 如在 SYFPEPE 数据库中就包含了超过 250 个 MHC 的结合基序^[19]。然而, 作为一种是与非的两态性确定性方法 (deterministic method), 结合基序法虽然易于理解和使用, 但精确度不高。在比较了用肽结合实验 (peptide binding assay) 从人类乳头状瘤上得到的大鼠 class II 细胞表位和用结合基序法预测的表位之后, 发现只有 40% 的 MHC 结合肽和预测的结果相一致^[20]。

3.2 量化基序法 量化基序法是对结合基序法的简单和有效的延伸。在抗原肽与 MHC-I 类分子的结合中, 抗原肽的侧链效应 (锚定、抑制或中性) 主要依赖于相应氨基酸残基在抗原肽中所处的位置, 而受邻近氨基酸残基的影响较小。因此, 在一定范围内可以假设: 在一个抗原肽序列中, 每个氨基酸残基相互独立地以一定的结合能影响该肽与 MHC 分子的亲和力, 即非依赖性侧链结合 (independent binding of side chains, IBS), 而抗原肽与 MHC 分子的结合是该肽中每个位置上的氨基酸残基结合能的综合, 通过对大量多肽或肽库与某型 MHC 分子的结合实验和统计学分析, 可得到反映该型 MHC 分子结合特异性的量化矩阵, 进而借助数学和计算机科学编制成程序即可用于表位的预测。该方法已经被应用于 MHC-I 类分子^[21]、HLA-DR II 类分子^[22]、从实验得来的肽结合数据中以及组合肽库的定点扫描^[23]。如 TEPITOPE 软件在预测人类 HLA-DR 时, 就包含了许多量化矩阵算法来标识随机的结合多肽^[24]。量化基序法的主要缺点是缺乏通用性, 因为 IBS 假设忽略了抗原肽可能形成的三维结构对结合的影响, 而且偏向于那些使用了矩阵系数的肽段。

3.3 人工智能法 “量化基序法”的前提是 IBS 假设, 但事实上已证明, 抗原肽中不同位置的氨基酸残基的侧链在与 MHC 分子结合的过程中是互相影响的, 因此, 要使表位的预测更为准确, 就必须将整个多肽的序列考虑在内, 这就涉及到大量序列依赖性以及对与氨基酸残基相关数据的处理, 必须借助于人工智能。人工智能法中最常用的莫过于 ANN 和 HMM。用大量已知可结合或非结合的多肽信息“训练”ANN, 可使其“学习”其中隐含的规律, 进而将这些规律泛化, 对某个未知肽的结合性作出判断, 以达到预测的目的。大量 MHC-II 类分子结合肽经过 ANN 分析后, 其阳性预测率可达 78%。HMM 也是一种捕捉数据复杂关系的方法, 特别适合于生物序列的建模。一个重要的特色是, 其能在单个模型中表达不同长度的肽段序列。Ryuji 等^[25]使用 HMM 预测了复杂

度较高的肽与 MHC-II 类分子的结合, 在比较了 8 种 HLA-DR1、-DR2、-DR4、-DR7、-DR15、DR17、-DR51 及 -DQ2 和相应肽的结合后, 发现精确率平均可以达到 87%, 并且能够区分与 DR 和 DQ 结合的肽。

3.4 分子建模法 分子建模法可提供分子间相互作用的特定 3 维结构和反应的详细情况, 可以更准确地看出活性中心 (active site) 的各个氨基酸残基与 MHC 分子对应的情况。如果已知 MHC 分子的晶体结构以及肽和 MHC 相互作用的结合数据, 就可以利用这些数据为模板来构建他们在三维结构中的作用模式。基于分子建模的虚拟筛选技术, 为解决高通量筛选提供了一个新思路。这一方法可通过一个特别设计的评分函数 (scoring function) 来“训练”MHC 肽链数据^[26], 以及把肽链自动对位在 MHC 分子上并与其相结合。定量结构活性关系 (quantitative structure-activity relationship, QSAR) 为最早提出的虚拟筛选方法, 包括平面 (2D QSAR) 和三维 (3D QSAR) 模式。2D QSAR 仅仅基于肽链的序列 (一级结构), 所以又叫加成性方法 (additive method)^[27], 即每一个取代物的生物活性都有其恒定的作用, 而与结构上的差异无关, 最后才是相邻侧链间的作用。如 Doytchinova 等^[28]用 2D QSAR 筛选出了一组高亲和力的 HLA-A2 结合肽。3D QSAR 是根据生物活性分子与受体结合时 3 维结构的性质, 用计算机仿真分析一系列已知可不同程度地与蛋白质或酶结合能力的化学分子及各个分子排列重叠的构形, 来找出其共同性, 进而确定与蛋白质或酶结合的结构活性关系, 这样就可以确定化学分子重要作用力之间的立体关系。此种方法比较典型的应用, 有 MSA-QSAR (分子形状分析-定量结构活性关系) 和 CoMFA (比较分子场分析)。

4 结语

无论从免疫系统组成成分的多样性 (如免疫球蛋白、淋巴细胞受体和细胞因子等), 还是从调控网络的复杂性 (如调控网络中蛋白质的相互作用) 来看, 如果仅仅依靠传统的实验手段进行免疫学研究, 不仅费时费力, 而且很难取得很大的进展。

生物信息学是一门交叉学科, 它把统计分析和计算模拟等一系列计算科学的技术引入生命科学的领域, 利用数据挖掘等信息学的原理, 大大加快了新基因发现和标记速度, 从而为后基因组时代的海量序列数据的处理和分析提供了有效的解决方法。

数据库是生物信息学中的重要组成部分, 也是进一步进行生物信息学分析的基础。通用数据库 (如 GenBank、EMBL、DDBJ 和 SWISS-PROT 等) 由于涵盖的范围太广, 往往不能满足特定领域的研究者便捷查询的需求。对于免疫学来说, 现有可供查询的专业数据库种类较少, 且存在命名规则混乱、收集的数据量不大, 以及更新缓慢等问题, 在一定程度上制约了分子免疫学的高速发展。因此, 如何建立一个专供免疫

学应用的, 结构合理、命名规范、数据完整、更新及时, 而且易于操作、库容量大的数据库, 将成为今后生物信息工作者和免疫学研究工作者所共同面临的一大挑战。

免疫基因组是生物经过亿万年前逐渐进化而成的适合于自身免疫防御的一套完整而严谨的结构, 作为生物信息学分支的比较基因组学就是分析这种变化过程的有效工具。值得一提的是, 由于 EST 序列相对于物种全基因组来说要短的多, 且易于克隆和测定, 又去除了内含子并具有组织特异性, 因此, 利用生物信息学的方法从 EST 数据库中鉴定同源免疫基因乃是一种多、快、好、省的技术。

T 细胞表位的筛选对于疫苗开发和免疫疗法来说, 都具有重要的作用。据统计, 全球疫苗市场每年的消费量约为 50 亿美元。随着许多病原体基因组序列的测定以及抗原蛋白的鉴定工作的开展, 计算机预测必将成为鉴定免疫反应靶点的首要一环。如在混栖 T 细胞表位 (promiscuous T-cell epitopes) 的鉴定中, 由于混栖肽 (promiscuous peptides) 可有效地作用于大范围的人群, 因而成为疫苗设计和免疫疗法的良好靶点。

参考文献:

[1] Brusic V, Zeleznikow J, Petrovsky N. Molecular immunology databases and data repositories [J]. *Immunol Meth*, 2000, 238: 17-28

[2] Garcia C, Robinson J, Guehlein IA, et al. Human KIR sequences [J]. *Immunogenetics*, 2003, 55: 227-239

[3] James R, Matthew J, Peter S, et al. IPD-the immunopolymorphism database [J]. *Nucleic Acids Res*, 2005, 33: D523-526

[4] Metcalfe P, Watkins N, Ouweland W, et al. The influence of human platelet antigen match on the success of allogeneic peripheral blood progenitor cell transplantation following a reduced-intensity conditioning regimen [J]. *Vox Sanguinis*, 2003, 85: 240-245

[5] Lefranc MP. MGT-ONTOLOGY and MGT databases: tools and Web resources for immunogenetics and immunoinformatics [J]. *Mol Immunol*, 2004, 40: 647-660

[6] Ohno S. Evolution by gene duplication [M]. Springer, Berlin, 1970.

[7] Trötschel J. Genetic and functional relationships between MHC and NK receptor genes [J]. *Immunity*, 2001, 15: 363-374.

[8] Trötschel J, Barten R, Haude A, et al. The genomic context of natural killer receptor extended gene families [J]. *Immunol Rev*, 2001, 181: 20-38

[9] Annalisse M. Comparative genomic analysis: diversity and evolution of the KIR haplotypes A and B [J]. *Gene*, 2004, 335: 121-131

[10] Waterston R. Initial sequencing and comparative analysis of the mouse genome [J]. *Nature*, 2002, 420: 520-562

[11] Marsh SGE, Parham P, Barber LD. The HLA Factsbook [M]. Academic Press, San Diego, 2000

[12] Kelley J, Walter L, Trötschel J. Comparative genomic of major histocompatibility complexes [J]. *Immunogenetics*, 2004, 56: 683-695.

[13] Borghans JA, Belman JB, De Boer R J. MHC polymorphism under host-pathogen coevolution [J]. *Immunogenetics*, 2004, 55: 732-739.

[14] Davil J, Lynn AT, Lloyd, et al. In silico identification of components of the Toll-like receptor (TLR) signaling pathway in clustered chicken expressed sequence tags (ESTs) [J]. *Veterinary Immunology Immunopathol*, 2003, 93: 177-184

[15] Yannick G. Immune gene discovery by expressed sequence tags generated from hemocytes of the bacterin-challenged oyster, *Crassostrea gigas* [J]. *Gene*, 2003, 303: 139-145.

[16] Mattner F, Fleitmann JK, Lingnau K, et al. Vaccination with poly-L-arginine as immunostimulant for peptide vaccines: induction of potent and long-lasting T-cell responses against cancer antigens [J]. *Cancer Res*, 2002, 62: 1477-1480

[17] 王莉, 吴玉章, 林治华, 等. 分子模拟在筛选 HLA 2A211 高亲和性 MART21 CTL 表位中的应用研究 [J]. 第三军医大学学报, 2002, 24: 1166-1168.

[18] Falk K. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules [J]. *Nature*, 1991, 351: 290-296.

[19] Rammensee HG. SYFPEITHI database for MHC ligands and peptide motifs [J]. *Immunogenetics*, 1999, 50: 213-219

[20] Stauss H J. Induction of cytotoxic T lymphocytes with peptides in vitro: identification of candidate T-cell epitopes in human papilloma virus [J]. *Proc Natl Acad Sci USA*, 1992, 89: 7871-7875

[21] Singh H, Raghava GP. ProPred: prediction of promiscuous MHC Class-I binding sites [J]. *Bioinformatics*, 2003, 19: 1009-1014

[22] Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites [J]. *Bioinformatics*, 2001, 17: 1236-1237

[23] Udaka K. An automated prediction of MHC class II binding peptides based on positional scanning with peptide libraries [J]. *Immunogenetics*, 2000, 51: 816-826

[24] Sumioto T. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices [J]. *Nat Biotechnol*, 1999, 17: 555-561

[25] Ryuji K, Hiroyuki H, Hiroyuki H, et al. Hidden Markov model-based approach as the first screening of binding peptides that interact with MHC class II molecules [J]. *Enzyme Microbial Technology*, 2003, 33: 472-481.

[26] Logean A. Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions [J]. *Mol Chem Lett*, 2001, 11: 675-679

[27] Doytchinova IA. Additive method for the prediction of protein-peptide binding affinity: Application to the MHC class I molecule HLA-A*0201 [J]. *J Proteome Res*, 2002, 1: 263-272

[28] Doytchinova IA, Walsh VA, Jones NA, et al. Coupling in silico and in vitro analysis of peptide-MHC binding: a bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes [J]. *J Immunol*, 2004, 172: 7495-7502