

基于改进的人工神经网络方法预测 CTL表位

刘涛*, 宋哲, 刘伟, 王雪颖, 邱晓明

(大连理工大学 物理系, 辽宁 大连 116024)

摘要: 为简化细胞病毒 T 细胞 (cytotoxicity T lymphocyte, CTL) 表位鉴定方法, 应用改进的人工神经网络方法定量研究了短肽与 MHC (major histocompatibility complex) 分子结合亲合力的关系, 并建立了 CTL 表位的预测模型, 得到了预测模型最优性能参数. 用此模型对短肽与 HLA-A* 0201 分子结合的 805 个预测样本进行了预测, 预测准确度达到 73.8%. 对来自黑色素 MAGE-2 的短肽与 MHC 分子的结合亲合力也进行了预测, 结果较好.

关键词: 人工神经网络; BP 网络; CTL 表位; MHC-短肽复合物
中图分类号: Q811.4 **文献标识码:** A

0 引言

近年来肿瘤特异性免疫治疗越来越受到人们的关注, 人们对特异性杀伤肿瘤细胞过程也有了进一步的了解, 其过程是肿瘤抗原蛋白在抗原提呈细胞 (antigen-presenting cells, APC) 的胞浆中被蛋白酶体 (proteasome) 裂解成短肽片段 (短肽长度一般为 8~11), 由转运相关蛋白 (transporter associated with antigen processing, TAP) 负责完成将蛋白酶体裂解产生的短肽片段从胞浆转运至内质网腔, 短肽在内质网中, 与新生成的主要组织相容性复合体 (major histocompatibility complex class I, MHC I 类) 分子 T 重链 β_2 微球蛋白结合, 形成的 MHC-短肽复合体被提呈到 APC 细胞表面, 与 T 细胞表面受体 (T cell receptor, TCR) 特异性结合, 使得 CTL 细胞开始活化、增殖、分化, 从而对肿瘤细胞进行特异性杀伤. 通常将能够启动免疫应答的一个抗原短肽称为 T 细胞表位.

由于 CTL 细胞表位是 CTL 细胞激活的关键, 其选取对于 CTL 细胞的激活起着十分重要的作用. 实验上选取 T 细胞表位是采用先合成大量的交叠肽, 再通过后续淋巴细胞实验进行选取. 这种方法不仅费时费力, 费用大, 而且预测 T 细胞表位的效率也低. 随着生物信息学和计算机技术

的发展, 许多预测 CTL 表位的理论方法应运而生, 主要包括 6 大类, 即 ① 神经分类器: 人工神经网络; ② 统计学分类器: 量化矩阵、权值基序、隐马尔可夫模型; ③ 结构分类器: 结合基序、权值基序; ④ 基于启发式的方法: 序列相似法; ⑤ 计算机建模: 分子建模; ⑥ 结合法: 分子建模与隐马尔可夫模型, 其中每种方法都处于发展完善中. 目前, 对 T 细胞识别抗原短肽的机理还不是很清楚, 各种预测理论方法主要讨论的都是抗原短肽与 MHC 分子的结合, 因为抗原短肽与 MHC 的结合力越强, 成为表位的可能性就越大^[1]. 免疫学实验上用 IC_{50} 值来标记抗原短肽与 MHC 分子结合力的大小, 每一个抗原短肽的氨基酸序列都对应一个 IC_{50} 值. 本文基于人工神经网络方法研制 CTL 表位预测软件, 在认真总结国内外有关该方面研究成果的基础上, 对预测模型参数进行研究及改进, 通过对已知样本 (由以前实验所累积的数据中选取) 训练, 获得具有预测能力的权值序列, 并利用此权值序列来预测抗原短肽与 MHC 分子结合情况, 以达到对 CTL 表位初步筛选, 提高实验成功率的目的.

1 材料与方法

1.1 预测模型的数学原型

本模型应用的数学原型是 BP 网络, 它采用

光滑活化函数 $g(x)$, 具有一个或多个隐层, 相邻两层之间通过权值全连接. 其基本流程是: 逐地将训练样本 $\{\mathbf{x}^j\}_{j=1}^J \subset \mathbf{R}^N$ 从输入层送入网络, 样本信息在网络中逐层向前流动, 在输出层获得相应的实际输出 $\{\mathbf{Y}^j\}_{j=1}^J \subset \mathbf{R}^M$, 并由实际输出 $\{\mathbf{Y}^j\}_{j=1}^J \subset \mathbf{R}^M$ 和理想输出 $\{\mathbf{O}^j\}_{j=1}^J \subset \mathbf{R}^M$ 计算出误差 E , 此过程网络称为“前传网络”. 而后利用所得的误差去调整网络的权值, 即是网络的学习过程. 这次网络的权值修改过程是由网络的输出层向输入层逐层修改的, 即所谓的误差反向传播. 然后再将样本逐一地从输入层送入, 再次计算实际输出和误差, 如此反复直到误差满足事先设定的条件.

误差函数如下:

$$E = \frac{1}{2} \sum_{j=1}^J |\mathbf{O}^j - \mathbf{Y}^j|^2$$

若求得的误差值不满足事先设定的条件, 就要调整权值, 一个常用的方法是梯度下降法. 取当前权值的改变量为

$$\Delta w = -Z \frac{\partial E}{\partial w}$$

其基本思想是将线性不可分的正负样本信息投射到高维特征空间中使其变为线性可分, 并找到分割函数, 此后就可以利用此函数去预测未知样本信息.

1.2 预测模型建立

图 1 是长度为 9 的短肽 (nonamer), 从图中可见每个短肽上有 9 个氨基酸残基位置 (P1~P9), 每个位置上可能出现 20 种氨基酸残基中的任意一种. 由实验得知拥有不同氨基酸序列的短肽与 MHC 分子的实验结合值 (即 IC_{50} 或 PIC_{50} 值) 也不相同, 即短肽的氨基酸序列与实验结果有一一对应关系, 且研究表明同一个短肽中的各个位置的氨基酸存在相互影响, 显然这是一种非线性问题. 而这种非线性问题可以用多层带权值的神经元来解决, 所以便将人工神经网络的方法引入表位预测领域中来.

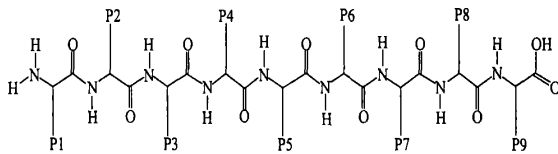


图 1 短肽的结构

Fig. 1 The structure of nonamer peptide

一般的神经网络方法都有易陷入误差曲面的局部极小和收敛慢的缺点, 对此本文做了两点优化改进:

(1) 附加动量法

在反向传播方法的基础上, 在每一个权值的变化上加上一个正比于前次权值变化的量, 并且根据反向传播法产生新的权值变化, 即

$$\Delta w(t+1) = - (1 - T) Z \frac{\partial E(t)}{\partial w} + T \Delta w(t)$$

当增加了动量项后, 促使权值的调整向着误差曲面底部的平均方向; 当网络权值进入误差曲面底部的平坦区时, $\partial E(t) / \partial w$ 将变很小. 于是, $\Delta w(t+1) = T \Delta w(t)$, 从而防止 $\Delta w(t+1) = 0$ 的出现, 有助于使网络从误差曲面的局部最小值中跳出. 同时, 为防止修正的权值在误差中导致太大的增长, T 的选取依如下规则

$$T = \begin{cases} 0, & \text{当 } E(t+1) > E(t) \times 1.04 \\ T, & \text{当 } E(t+1) \leq E(t) \times 1.04 \end{cases}$$

(2) 自适应学习率

对于特定问题, 通常是凭经验或实验设定一个适当的学习率 Z , 但即便如此, 对训练初期效果较好的学习率, 往往不适用于训练后期. 因此, 有必要自动地调整学习率, 采用的调整公式如下

$$Z(t+1) = \begin{cases} 1.05Z(t); & \text{当 } E(t+1) < E(t) \\ 0.7Z(t); & \text{当 } E(t+1) > E(t) \times 1.04 \\ Z(t); & \text{其他情形} \end{cases}$$

依照规则, 初始学习率 $Z(0)$ 的选取范围可以有很大的随意性.

本文是针对 HLA-A* 0201 分子结合长度为 9 的短肽的研究 (对于其他 MHC I 类分子结合其他长度的短肽本模型同样适用, 只需改变模型参数, 例如与 8 个氨基酸组成的短肽的结合问题, 则只需将下面的肽链长度参数改为 8 即可). 本文改进后的 BP 网络如图 2 所示, 首先将短肽扩展成 $9 \times 20 + 1 = 181$ 个神经元, 即网络的输入层, 其中 20 代表在短肽中每个位置氨基酸出现的 20 种

可能性,出现哪种氨基酸,哪个对应数值就是 1,其余 19 个则为 0;额外的那个神经元是网络偏移量 θ ;隐层有 $9+1=10$ 个神经元. 其中 9 个神经元代表短肽的 9 个位置,它们体现了输入层所有位置的氨基酸与 MHC I 类分子结合时对该位置的作用,额外的那个神经元仍然是偏移量 θ ;输出层有 1 个神经元,代表短肽与 HLA-A* 0201 分子结合亲和力,用 PIC_{50} 值表示 ($PIC_{50} = \lg(1/IC_{50})$)^[21].

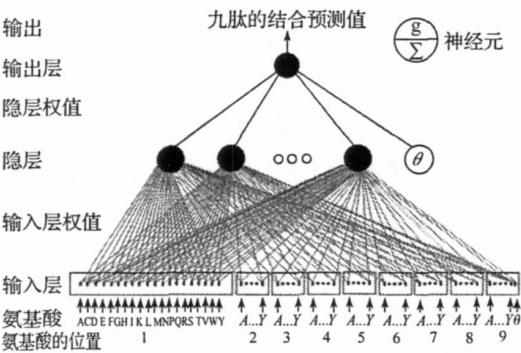


图 2 模型的结构
Fig. 2 Architecture of the model

1.3 模型参数的设定 (针对 HLA-A* 0201 结合短肽的预测)

肽链的长度 $P=9$. 网络偏移量 $\theta=-1.0$, 给每一层神经元增加一个偏移量可以加快收敛速度并且对其分类能力也有很大的改善. 本文也与没有偏移量的同类模型进行了比较,结果发现有偏移量的这种设置法效果更佳,例如文献 [3] 是将偏移量设定为 0 (即没有它),其敏感度、特异性、准确度分别为 73.2 71.2 72.2;而本文这种设定法的敏感度、特异性、准确度为 72.6 75.0 73.8,显然这种方法更优.

隐层数目 $n=1$, 本文对多个隐层也进行了研究,发现多个隐层是不必要的,反倒使网络本身的结构复杂化,计算量也大幅度提高.

初始学习率 $Z=0.01$, 前面已经提到初始学习率的设定有很大的随意性,但初始值较小的学习率还是更为稳妥的.

活化函数参数 $U=0.01$, U 是活化函数的图像曲线的平滑度. U 越小,曲线越平滑,与学习率一起决定收敛速度.

权值修改方式: 在线方式,采用这种方式是由于本文研究的样本数很大,若采用批方式,训练

次数会远远大于在线方式,而且抖动问题过多. 且在线梯度法可以看做是梯度法的一种随机扰动,有助于跳出局部极小.

目标误差 $E=0.000\,000\,5$, 即当误差 E 小于 0.000 000 5 时停止训练. 本文对更小的误差设定也作了研究,结果运行时间不仅大大增长,且还出现了过拟合现象.

阈值为 5, 阈值就是网络中用于判断结合肽和非结合肽的分界线, 本文对 0 到 10 范围内间隔 0.1 的所有阈值进行了研究, 当阈值选取为 5 时敏感度和特异性几乎相等.

动量因子 $T=0.9$, 这是对训练时收敛速度和抖动问题的综合考虑, 当设定值小于 0.9 时收敛过慢, 当设定值大于 0.9 时抖动问题过多, 这个值的设定跟网络的编写有关, 不同编码基于这两个因素的考虑会采用不同的动量因子.

1.4 模型样本的训练集和预测集

本研究用到的样本分为两部分: 训练集和预测集. 样本主要来自 MHCBN 数据库和此领域的相关表位预测的文献^[24-61]. 作为训练集的样本是这些数据中所有具有确定的结合值 (是指 IC_{50} 或 PIC_{50} 值) 的样本 (共有 320 个), 其余的作为预测集 (共有 805 个, 其中包括 753 个结合肽和 52 个非结合肽).

1.5 模型评价指标

为了评价本模型, 引进评估参数: 敏感度 S_{en} , 特异性 S_{pe} , 正预测度 P_{pv} , 负预测度 N_{pv} , 准确度 A_{ac} 以及相关系数 C ^[61].

$$S_{en} = \frac{N_{TP}}{N_{TP} + N_{FN}} \times 100\%,$$
$$S_{pe} = \frac{N_{TN}}{N_{TN} + N_{FP}} \times 100\%,$$
$$P_{pv} = \frac{N_{TP}}{N_{TP} + N_{FP}} \times 100\%,$$
$$N_{pv} = \frac{N_{TN}}{N_{TN} + N_{FN}} \times 100\%,$$
$$A_{ac} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{TN} + N_{FN}} \times 100\%,$$
$$C = \frac{N_{TP} \times N_{TN} - N_{FN} \times N_{FP}}{(N_{TN} + N_{FN})(N_{FN} + N_{TP})(N_{TP} + N_{FP})(N_{FP} + N_{TN})} \times 100\%$$

式中: N_{TR} N_{FR} N_{TN} N_{FN} 分别表示预测出正确的结合肽、错误的结合肽、正确的非结合肽、错误的非结合肽.

2 结果与讨论

图 3 为学习率 η 随训练步数 n 的变化情况. 从图 3 可以看出, 学习率总是自适应地调整, 到达一个峰值后, 就迅速地降到一个很小值, 然后再变大. 在学习过程中, 一开始学习率峰值较小, 随后学习率的峰值逐渐呈现非线性增长趋势. 这也说明采用自适应学习率比传统 BP 网络具有如下优点: 一开始采用较小的学习率可以避免网络震荡, 保证进入一个正确吸引域; 然后较大的学习率可以成倍地减少学习时间, 加速网络的收敛过程, 而传统的 BP 网络是难以两者兼顾的.

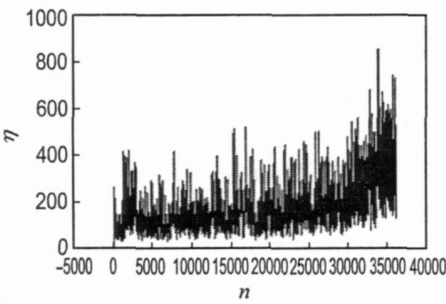


图 3 学习率曲线

Fig. 3 Curves of learning rate

应用自行研制的预测 CTL 表位的软件对 805 个预测样本进行预测 (未经任何手段的挑选), 并对 0 到 10 之间的所有阈值进行了测定. 所得数据如表 1 所示.

表 1 中用黑体标记的是本模型最终选定的阈值, 在该阈值下模型具有最为接近的敏感度和特异性, 并且具有较高的准确率和较大相关系数. 需要说明的是不同训练样本和不同编码神经网络的模型会拥有不同的阈值, 但对于测量结果不会造成影响, 只要选取的是敏感度和特异性最接近的就可以了. 从表 1 中可以看出要获得高敏感度时, 可以调低阈值; 但应注意此时是以降低特异性为代价的. 同理, 也可以调高阈值来获得高特异性. 对于阈值 T 的选取本文也搜集了一些国内外同行的结果, 分别是文献 [6]: $T(PIC_{50}) = 4 IC_{50} = 100 \mu m$; 文献 [7]: $T(PIC_{50}) = 4.3 IC_{50} = 50 \mu m$; 文献 [8-9]: $T(PIC_{50}) = 5 IC_{50} = 10 \mu m$; 文献 [3]: $T(PIC_{50}) = 5.1 IC_{50} = 8 \mu m$; 文献 [3-8]: $T(PIC_{50}) = 7 IC_{50} = 0.1 \mu m$.

表 1 改进的神经网络预测 CTL 表位模型的性能

Tab. 1 The performance of modified ANN-based prediction method at various thresholds

T	S_{en}	S_{pe}	N_{pv}	P_{pw}	A_{cc}	C
1.0	99.2	1.9	70.7	50.3	50.6	4.9
1.5	98.1	3.8	67.4	50.5	51.0	6.0
2.0	96.1	7.7	66.6	51.0	51.9	8.2
2.5	93.4	7.7	53.7	50.3	50.5	2.0
3.0	89.5	15.4	59.5	51.4	52.4	7.3
3.5	86.1	28.8	67.4	54.7	57.5	18.2
4.0	81.5	48.1	72.3	61.1	64.8	31.4
4.5	77.4	55.8	71.2	63.6	66.6	34.0
5.0	72.6	75.0	73.3	74.4	73.8	47.7
5.5	64.8	78.8	69.1	75.4	71.8	44.1
6.0	54.7	84.6	65.1	78.1	69.7	41.2
6.5	43.3	88.5	60.9	79.0	65.9	35.6
7.0	29.0	94.2	57.0	83.4	61.6	30.6
7.5	14.9	98.1	53.5	88.6	56.5	23.3
8.0	7.8	100	52.0	100	53.9	20.2
8.5	2.3	100	50.6	100	51.1	10.7
9.0	0.1	100	50.0	100	50.1	2.6

为了检测本模型的性能,作者应用本模型对黑色素 MAGE-2的蛋白质氨基酸序列(见图 4,来自 SWISS-PORT数据库)进行了预测,因为恶性肿瘤一般均有黑色素 MAGE的表达.通过基序法和上面建立神经网络模型的结合进行预测,表 2中的 P 值给出的是模型预测各个短肽与 MHC

分子结合的 PIC_{50} 值,所列出的短肽序列为预测值最高前 16个,即是该模型预测出的待定表位.可喜的是其中短肽 KMVELVHFL (112-120)、KIWEELSM L (220-228)、FLWGPRALI (271-279) 已被实验验证了(详见表 2),由此可以看出此方法的效率是令人满意的.



图 4 黑色素 MAGE-2的蛋白质氨基酸序列

Fig. 4 Amino acid sequence of melanin MAGE-2 protein

表 2 神经网络模型结合基序法预测来自黑色素 MAGE-2的短肽

Tab. 2 Melanin MAGE-2 binders predicted by ANN combined with motif prediction method

排序	位点	序列	P	文献
1	45-53	LV EV T LGEV	9. 051	
2	108-116	AISRKM V EL	9. 671	
3	112-120	KM V ELV HFL	8. 367	[10]
4	113-121	MV ELV HFLL	7. 007	
5	132-140	K AEM LESV L	7. 463	
6	159-167	QLV FGI EVV	8. 495	
7	174-182	HL YILV T CL	7. 775	
8	176-184	YILV TCLGL	7. 643	
9	194-202	QVM PKTGLL	8. 530	
10	201-209	LL IIV L AII	8. 054	
11	202-210	LIIV LAIIA	7. 742	
12	220-228	KIW EELSM L	9. 488	[11]
13	237-245	SV FAHPRKL	7. 712	
14	271-279	FLWGPRALI	7. 579	[11]
15	278-286	LIETSYV KV	9. 532	
16	283-291	YV KV IHHTL	9. 192	

3 结 语

本文建立的神经网络模型是一种较好的表位预测模型,它能较准确地预测未知样本抗原肽与 MHC分子结合的亲和力. 应该注意的是,人工神经网络方法预测 CTL表位是建立在大量的实验数据基础上的. 虽然现今的实验数据的确有了明显的增加,可是在这些数据中具有精确的 IC_{50} 或 PIC_{50} 值的却占很少一部分,而且结合肽的数据量和非结合肽的数据量仍然存在较大的不平衡性,这都将延缓人工神经网络方法展现其优越性. 不过随着 CTL表位数据库的建立和完善,相信这一点一定会在众多科学家的共同努力下不久就得以解决.

参考文献:

[1] HATTOTUWAGAMA C K, GUAN P, DOYTCHINOVA I A, *et al.* Quantitative online prediction of peptide binding to the major histocompatibility complex [J]. *J Mol Graphics and Modeling*. 2004, 22(3): 195-207

- [2] BHASIN M, SINGH H, RAGHAV A G P S. MHCBN: a comprehensive database of MHC binding and non-binding peptides [J]. **Bioinformatics**, 2003, **19**(5): 665-666
- [3] BHASIN M, RAGHAV A G P S. Prediction of CTL epitopes using QM, SVM and ANN techniques [J]. **Vaccine**, 2004, **22**(23-24): 3195-3204
- [4] DOYTCHINOV A 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]. **J Med Chem**, 2001, **44**(22): 3572-3581
- [5] ZHAO B, MATHURA V S, RAJASEGER G, *et al.* A novel MHCp binding prediction model [J]. **Human Immunology**, 2003, **64**(12): 1123-1143
- [6] KESSLER J H, BEEKMAN N J, BRES-VELOEMANS S A, *et al.* Efficient identification of novel HLA-A* 0201-presented cytotoxic T lymphocyte epitopes in the widely expressed tumor antigen PRAME by proteasome-mediated digestion analysis [J]. **J Exp Med**, 2001, **193**(1): 73-88
- [7] GULUKOTA K, SIDNEY J, SETTE A, *et al.* Two complementary methods for predicting peptides binding major histocompatibility complex molecules [J]. **J Mol Biol**, 1997, **267**(5): 1258-1267
- [8] KESMIR C, NUSSBAUM A K, SCHILD H, *et al.* Prediction of proteasome cleavage motifs by neural networks [J]. **Protein Eng**, 2002, **15**(4): 287-296
- [9] HONEYMAN M C, BRUSIC V, STONE N L, *et al.* Neural network-based prediction of candidate T-cell epitopes [J]. **Nature Biotechnol**, 1998, **16**(10): 966-969
- [10] VISSEREN M J, VAN DER BURG S H, VAN DER VOORT E I, *et al.* Identification of HLA-A* 0201-restricted CTL epitopes encoded by the tumor-specific MAGE-2 gene product [J]. **Int J Canc**, 1997, **73**(1): 125-130
- [11] GENG Miao, WU Yu-zhang, JIA Zheng-cai. Identification of novel HLA-A2-restricted CTL epitopes derived from tumor antigen MAGE-2 [J]. **Acta Academiae Medicinae Militaris Tertiae**, 2002, **24**(10): 1156-1158

Prediction of CTL epitopes based on modified artificial neural network

LIU Tao^{*}, SONG Zhe, LIU Wei, WANG Xue-ying, QIU Xiao-ming

(Dept. of Phys., Dalian Univ. of Technol., Dalian 116024, China)

Abstract To simplify the identification method of cytotoxicity T lymphocyte (CTL) epitopes, the binding affinity relationship between peptide and MHC (major histocompatibility complex) molecular was quantitatively studied with the modified artificial neural networks method. The model of predicting the epitopes of CTL was established, and the optimum parameters for the prediction model were obtained. The binding affinity prediction to the 805 peptides samples of HLA-A* 0201 molecular was carried on with this model, and the accuracy of prediction was 73.8%. The affinity between nonamer of melanin MAGE-2 and MHC is predicted, and the results are reliable.

Key words artificial neural network; BP networks; CTL epitopes; MHC-peptides compound