基于粗集的T细胞表位预测方法*)

曾 安[1] 潘 丹[2] 郑启伦[3] 彭 宏[3]
(广东工业大学计算机学院 广州 510006)[1] (中国移动通信集团广东有限公司 广州 510100)[2]
(华南理工大学计算机科学与工程学院 广州 510640)[3]

摘 要 T细胞表位预测技术对于减少实验合成重叠肽、研究病原体与机体作用的免疫机制以及深入理解T细胞介导的免疫特异性均有重要意义。为增强T细胞表位预测模型的可理解性,本文在通过肽的预处理构建出存储等长肽段的决策表之后,设计出了一种基于粗集的T细胞表位预测方法。该方法由基于信息熵的属性约简完备算法和基于锚点知识的属性值顺序约简改进算法共同组成。基于HLA-DR4 (B1 * 0401)编码的MHC II类分子结合肽的实验数据表明,在预测精度与传统神经网络方法大致相当的基础上,本文方法可以提取出用于帮助专家理解MHC分子与抗原肽结合机理的产生式规则。

关键词 T细胞表位预测,粗集,规则获取

T Cell Epitope Prediction Approach Based on Rough Set Theory

ZENG An[1] PAN Dan[2] ZHENG Qi-Lun[3] PENG Hong[3]
(Faculty of Computer, Guangdong Univ. of Tech., Guangzhou 510006)[1]
(China Mobile Group Guangdong Co., Ltd., Guangzhou 510100)[2]
(College of Comp. Eng. & Sci., South China Univ. of Tech., Guangzhou 510640)[3]

Abstract Predicting which peptides can bind to a specific MHC molecule is indispensable to minimizing the number of peptides required to synthesize, to the research on the interaction mechanics between infector and organism, and especially to helping understand the specificity of T-cell mediated immunity. In order to enhance the understandability of existing T cell epitope prediction methods based on machine learning, we firstly construct a decision table comprising the nonamers by peptide preprocessing. And then we propose a T Cell epitope prediction approach based on rough set theory, which consists of the complete attribute reduction algorithm based on information entropy and the renovated version for orderly attribute value reduction algorithm combined with expert knowledge of binding motifs. Finally, with the help of the approach, a comprehensible rule set with strong generalization ability to predict the peptides that bind to HLA-DR4(B1 * 0401)is acquired.

Keywords T cell epitope prediction, Rough set, Rule acquisition

1 引言

目前,大量研究表明:病原体感染、肿瘤发生发展、自身免疫性疾病的发生发展和组织器官移植排斥都与T细胞抗原识别和活化异常或偏离相关[1]。T细胞抗原表位就是抗原蛋白中经抗原递呈细胞(APC)处理由MHC分子递呈给TCR的多肽片断。而T细胞表位预测是指借助于计算机的海量数据处理能力,从数百万的蛋白质里找出既能与特定的MHC分子结合,又能与特定TCR结合的抗原肽,然后在此基础上通过生物实验判断找到的抗原肽能否使得T细胞活化;若能,则可确定该抗原肽为T细胞抗原表位。由于抗原提呈给TCR是一个极为复杂的过程(其中包括蛋白酶的消化、抗原提呈转运体(TAP)的转运等),因此在现阶段,T细胞表位预测技术的研究仅限于抗原肽与MHC分子的结合这一环节。本文的研究工作也是围绕抗原肽与MHC分子的结合性展开的。

迄今为止,T细胞表位预测技术的研究方法主要分为4类[2]:1)基于基序的方法。该方法主要应用于MHC I类分子。但由于不是所有的结合肽都含有明确的基序(exact motif),因而产生了较低的预测精度。2)基于量化矩阵的方法。其假设前提是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分子结合的亲和力。而事实上,氨基酸残基间的影响关系是一个复杂的非线性问题,这与其假设前提不符,从而降低了预测精度。3)基于结构的方法。此类方法的特点在于可以对大量尚未进行结合分析的MHC等位基因进行研究,从而可能获得有价值的结果。然而,由于这类方法常需要较详细的相关结构信息,而目前得到的MHC分子的三维分子模型很少且运算量大,较耗时,这在一定程度上阻碍了它们的广泛应用。4)基于机器学习的方法。此类方法由于处理了复杂的非线性模式,具有较强的推广能力、自适应和自学习能力,从而能明显提高预测的准确性。例如:1997年Gulukota等[3]针对MHC分子HLA-A * 0201等位基因,比较了量化矩阵和ANNs等模型的性能,发现ANNs具有较优越的预测性能;1998年Brusic等[4]采用遗

*)国家自然科学基金重点项目(No. 30230350)、广东工业大学博士启动基金项目(No. 063001)。曾 安 讲师,博士,主要研究方向:智能信息处理、数据挖掘、生物信息学等;潘 丹 博士,高级工程师;郑启伦 教授,博导;彭 宏 教授,博导。

传算法(GA)和 ANNs 来预测 MHC II 类分子 HLA-DRB1 的结合肽；2002 年 Kun[5] 等综合比较了基于基序、量化矩阵、ANNs 和 HMMs 等几种模型的预测性能；Pierre 等[6] 利用支持向量机(SVM)来预测 MHC I 类分子结合肽。

然而，由于目前基于机器学习的 T 细胞表位预测模型主要集中于 ANNs、HMM 等的研究上，因此生物领域专家很难直观理解蕴涵在这些训练成功的预测模型中的知识，这对于深入理解 T 细胞介导的免疫特异性以及帮助生物学专家理解其自身的推理过程均造成了巨大的障碍。而基于粗集(Rough Sets)理论的知识发现方法能够有效地提取出以规则形式表达的知识，且在数据分析过程中可以不需要借助任何外界信息或先验知识。这样，对于人类认识尚少的领域——T 细胞表位预测问题，该方法更能发挥其独特优点。于是，本文试图开发一种基于粗集的 T 细胞表位预测方法，以获得理解性和推广能力均较强的规则集。

2 粗集理论简介

粗集理论是 1982 年由 Pawlak[7] 提出的一种处理不确定和不精确信息的数学理论，它能有效地分析不精确、不一致、不完整等各种不完备的信息。粗集理论的基本思想是建立在以下假定之上的，即对于论域里的每一个对象，都能找到某些信息与它相关联。若两个对象具有相同的信息，或根据已有的信息不能够将其划分开，则它们是不可区分的，也称是不可分辨的。这种不可分辨关系是粗集理论的数学基础。

一般地，一个知识表达系统 S 可以表示为 $S=\langle U,R,V,f\rangle$。这里，U 是一个论域，是全体对象的集合；$R=C\cup D$ 是属性集合，子集 C 和 D 分别称为条件属性集和决策属性集；$V=\bigcup_{r\in R}V_r$ 是属性值的集合，V_r 表示属性 $r\in R$ 的属性值范围，即属性 r 的值域；$f:U\times R\to V$ 是一个信息函数，它指定 U 中每一个对象的每一个属性的属性值[8]。

决策表则是一类特殊而重要的知识表达系统，它是指当情况满足某些条件时，应当采取怎样的决策(行动)。多数决策问题可以用决策表来表达。若决策表是相容的，则意味着决策表中不存在所有条件属性值完全相同但决策属性值不同的记录，即满足

$$\forall x_s,x_t\in U,\forall c_i\in C,c_i(x_s)=c_i(x_t)\Rightarrow d(x_s)=d(x_t)$$

对于每个属性子集 $B\subseteq R$，我们定义一个不可分辨二元关系 $IND(B)$，即

$$IND(B)=\{(x,y)\mid(x,y)\in U^2,\forall b\in B(b(x)=b(y))\}$$

显然，$IND(B)$ 是一个等价关系，且 $IND(B)=\bigcap_{b\in B}IND(\{b\})$。

在不可分辨定义的基础上，我们可给出如下几个定义[8]：

定义 1 上近似集和下近似集的形式化定义

给定知识表达系统 $S=\langle U,R,V,f\rangle$，对于每个子集 $X\subseteq U$ 和不可分辨关系 B，X 的上近似集和下近似集分别可以由 B 的基本集定义如下：

$$B_-(X)=\bigcup\{Y_i\mid(Y_i\in U\mid IND(B)\wedge Y_i\subseteq X)\}$$

$$B^-(X)=\bigcup\{Y_i\mid(Y_i\in U\mid IND(B)\wedge Y_i\cap X\neq\Phi)\}$$

这里，$U\mid IND(B)=\{X\mid X\subseteq U\wedge\forall x\forall y\forall b(b(b(x)=b(y)))\}$ 是不可分辨关系 B 对 U 的划分，也是论域 U 的 B 基本集的集合。

定义 2 边界、正域和负域的定义

集合 $BN_B(X)=B^-(X)\setminus B_-(X)$ 称为 X 的 B 边界；$POS_B(X)=B_-(X)$ 称为 X 的 B 正域；$NEG_B(X)=U\setminus B_-(X)$ 称为 X 的 B 负域。

$B_-(X)$ 是根据知识 B(属性子集 B)，U 中所有一定能归入集合 X 的元素构成的集合。$B^-(X)$ 是根据知识 B，U 中所有一定能和可能归入集合 X 的元素构成的集合。$BN_B(X)$ 是根据知识，U 中既不能肯定归入集合 X，又不能肯定归入集合 $\overline{X}$ 的元素构成的集合。正域 $POS_B(X)$ 是根据知识，U 中所有一定能归入集合 X 的元素构成的集合。负域 $NEG_B(X)$ 是根据知识，U 中所有不能确定一定能归入集合 X 的元素构成的集合。边界域 $BN_B(X)$ 是某种意义上论域的不确定域，边界域中的元素既不能肯定地属于集合 X，也不能肯定地属于集合 $\overline{X}$。

定义 3 Q 的 P 正域的定义

设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，Q 的 P 正域记为 $POS_P(Q)$，定义为 $POS_P(Q)=\bigcup_{v\in U/Q}P_-(X)$。

定义 4 P 中 Q(不)可省略的定义

设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，若

$POS_P(Q)=POS_{(P\setminus\{r\})}(Q)$，则称 r 为 P 中相对于 Q 可省略的，简称 P 中 Q 可省略的；否则称 r 为 P 中相对于 Q 不可省略的。

定义 5 P 相对于 Q 独立的定义

设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，若 P 中的每一 r 都是 P 中 Q 不可省略的，则称 P 相对于 Q 独立的。

定义 6 约简的定义

设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，若 P 的 Q 独立子集 $S\subset P$，有 $POS_S(Q)=POS_P(Q)$，则称 S 为 P 的 Q 约简。

定义 7 核(核属性)的定义

设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，$RED_Q(P)$ 为 P 的所有 Q 约简关系簇，则 P 的 Q 核 $CORE_Q(P)$ 定义为 $CORE_Q(P)=\bigcap RED_Q(P)$。

基于粗集理论的知识获取方法主要指的是：在保持决策表决策属性和条件属性之间的依赖关系不发生变化的前提下对决策表进行约简，包括属性约简和属性值约简(这两者有时是截然分开的，而有时则是合而为一的)。

属性约简是粗集理论中的一个核心部分，其目标是从条件属性集合中发现部分必要的条件属性，使得基于这部分条件属性形成的相对于决策属性的分类与所有条件属性所形成的相对于决策属性的分类一致。由于属性约简后获得的属性子集中所包含的所有属性的全部取值都被保留下来了，但是这其中仍然包含有冗余，因此为了从基于属性约简后的决策表中获取简单明了且推广能力(泛化能力)强的规则，我们还必须进行属性值约简。

3 系统原型

基于粗集的 T 细胞表位预测系统的建立主要包含 3 个步骤：1)对肽进行预处理，将不等长的肽序转变为等长的肽序；2)构造一个存储等长肽段的决策表；3)利用本文提出的基于粗集的 T 细胞表位预测获取易于领域专家理解的规则集。

3.1 肽的预处理

从 MHC 分子洗脱下来的肽段，尤其是和 MHC II 类分子相结合的多肽，其长度变化较大(10 到 30 个氨基酸)。然

而，MHC II类分子结合肽的核心区约为13个氨基酸残基，同时进一步的研究表明，其中仅由9个氨基酸所组成的肽段（简称为九肽）实为与MHC II类分子结合所必需的。对于九肽而言，其中一个主要锚定残基又是结合的关键，而其他几个次级锚定残基则对其结合能力产生一定影响。因此，肽的预处理目的是在固定一个初级锚点后，寻找优化的联配矩阵，将不等长的肽序变为等长的肽序，以尽量获取保留了初级锚点位置上所允许的氨基酸的九肽。具体做法是[4]：将非结合肽分解为多个九肽时，只要某个九肽的第一个位置上的氨基酸满足初级锚点的要求，则该九肽对应于预处理后的肽段，其亲和力类别与原始肽所属亲和力类别一致。这样，每个非结合肽可能分解成多个九肽。而对于结合肽（高亲和力、中亲和力和低亲和力的结合肽的总称）而言，首先固定好第一个位置上的氨基酸（即为满足初级锚点的要求），找出所有可能的候选九肽；然后根据联配矩阵为每一候选九肽打分，得分最高的候选九肽才对应于初始肽的预处理结果，其亲和力类别即为原肽段的亲和力类别。这样，每个结合肽就分解成为一个九肽。

3.2 决策表的构造

决策表中每一条记录表达一条九肽及其对应的亲和力类别（决策属性）。在此决策表中，条件属性的个数为180（9个位置上的20种可能的氨基酸），其值域为{0,1}，其中“1”表示在对应位置上出现了某个氨基酸，而“0”则表示该位置上没有出现该氨基酸；决策属性的个数为1，对应于MHC II类分子结合肽的亲和力的类别，其值域为{0,1,2,3}，其中0代表非结合肽；1代表低亲和力的结合肽；2代表中亲和力的结合肽；而3代表高亲和力的结合肽。这里，20个氨基酸的顺序被设定为：ACDEFGHIKLMNPQRSTVWY。于是，若以九肽FRPIQTVLS为例，该九肽则可表示为：

000010000000000000000000000000000010000000000000000010000000000000010000000000000000000000000010000000000
000000000000010000000000000000000010000000001000000000000000000000000010000。部分决策表见表1。

表1中，条件属性名中的字母表示氨基酸编码，数字表示该氨基酸在九肽中的位置。如：条件属性名A1代表九聚物中的第一个位置是否为丙氨酸。若条件属性A1的值为0，则表示该九聚物中第一个位置未出现丙氨酸；反之，则出现。类似地，若条件属性Y9的值为1，则表示该九聚物中第9个位置出现了酪氨酸。

表1 部分决策表

编号	条件属性																						类别
	A1	C1	D1	E1	F1	G1	H1	I1	K1	L1	M1	N1	P1	Q1	R1	S1	T1	V1	W1	Y1	…	Y9	
1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	…	0	3
2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	…	0	2
3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	…	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	…	0	0
…	…	…	…	…		…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

3.3 基于粗集的T细胞表位预测方法

基于粗集的T细胞表位预测方法由两个算法组成：一个是基于信息熵的属性约简完备算法，另一个是结合锚点知识的属性值顺序约简算法。

3.3.1 基于信息熵的属性约简完备算法

直观上，由上述方法构造的决策表会包含冗余的条件属性，这既会干扰人们作出正确的决策，又会浪费存储空间。而经过属性约简后的决策表易集中于最关键的条件属性，从而有表1部分决策表利于解决多属性的决策问题。于是，本文在文[9]的基础上，以信息熵作为刻画属性重要性的度量，提出了一基于信息熵的属性约简完备算法。该算法能在保持决策表分辨能力不变的前提下获取不带任何冗余属性的属性约简。

假定决策表 $T=(U,C,D)$，论域 U 是记录的集合；C 和 D 分别表示条件属性和决策属性的集合。条件属性集 $B(B\subseteq C)$ 的熵被定义为

$$H(B)=-\sum_{i=1}^{n}p(X_i)\log(p(X_i)) \tag{1}$$

其中，$p(X_i)=|X_i|/|U|$。$|*|$ 表示集合 $*$ 的势。

条件属性集 $B(B\subseteq C,U/\mathrm{Ind}(B)=\{X_1,X_2,\cdots,X_n\})$ 相对于决策属性集 D $(U/\mathrm{Ind}(D)=\{Y_1,Y_2,\cdots,Y_m\})$ 的条件熵的定义见式(2)[9]。

$$H(D|B)=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j|X_i)\log(p(Y_j|X_i)) \tag{2}$$

其中，$p(Y_j|X_i)=|Y_j\cap X_i|/|X_i|$，$1\leqslant i\leqslant n$，$1\leqslant j\leqslant m$。

属性 a 相对于条件属性集 $B(B\subseteq C)$ 的重要度[9]按式(3)表示。

$$Sgf(a,B,D)=H(D|B)-H(D|B\cup\{a\}) \tag{3}$$

假定决策表 $T=(U,A,C,D)$，基于信息熵的属性约简完备算法流程如下所示：

输入：决策表 T；

输出：属性约简 R；

1）计算条件熵 $H(D|C)$；

2）按文[10]计算相对差异比较表 $RDCT$，设 $RDCT$ 为 DD，假定 R 为决策表 T 约简后的属性集；

3）对 DD 中各列进行横向相加，结果设为 CC（CC 是个列向量）；

4）找出 CC 中最小值所在的行 $\{r_1,r_2,\cdots,r_k\}$；

5）如果 CC 中最小值为1，找出 $\{r_1,r_2,\cdots,r_k\}$ 这些行中元素1所在的列 $\{c_{r1},c_{r2},\cdots,c_k\}$，并且初始化 R 为对应于 $\{c_{r1},c_{r2},c_k\}$ 的属性子集；否则，初始化 R 为空；

6）计算条件熵 $H(D|R)$。如果 R 为空，则设 $H(D|R)=H(D)$；

7）判断是否 $H(D|R)=H(D|C)$，如果是，转8）；否则重复步[1]到[3]，直到 $H(D|R)=H(D|C)$；

[1]设 $E=C-R$；

[2]根据式(3)计算每一个属性 $a(a\in E)$ 相对于 R 的重要度 $SGF(a,R,D)$；

[3]找出最大值 $Sgf(a,R,D)$，并且设 $R=R\cup\{a\}$；

8）调用文[10]中的RJ算法，判断 R 是否是一个约简；

9）如果 R 是个约简，算法结束；否则，利用RJ算法求取

可能冗余的属性，逐个删除可能冗余属性中具有最小 $Sgf(a, R, D)$ 的属性直到 R 是一个约简，算法终止。

其中，相对差异比较表的构造和 RJ 算法的详细计算步骤见文[10]。

3.3.2 结合锚点知识的属性值顺序约简算法

在属性约简之后，就进入了属性值约简阶段，以获得规则集。正如文[10]中所述：属性值可分为 3 类：第一类是指删除该值后决策表的分类能力发生变化，属于必不可少的；第二类是指删除该值后决策表的分类能力不变且无重复记录产生，是可以删除的；第三类是指删除该值后决策表的分类能力不变且有重复记录产生，也是可以删除的。我们可以根据这样的分类结果对不同类型的属性值赋予不同的优先级别和采取相应的判断顺序：第一类属性值的优先级别最高，不需做任何判断，将其全部保留；第二类属性值的优先级别次之，需要判断第二类属性值中是否有需要保留的；第三类属性值的优先级别最低，需要判断第三类属性值中是否有需要保留的。

为了提高计算效率并改善其效果，本算法与文[10] OAVRA 算法不同的是：不是先对决策表中所有记录的属性值进行分类，而是先从决策表中 1 次抽取 1 条记录，然后对该条记录的属性值进行分类，根据分类结果获得一条规则，最后将决策表中能利用该规则作出决策的记录从决策表中删除，这样依次类推，直至整个决策表被删空，并且尤其重要的是，本算法充分利用了领域知识，将锚点知识与该算法有机地联系起来了，其具体步骤如下：

输入：约简后的决策表和某等位基因编码的锚点知识；

输出：规则集 R；

1)从约简后的决策表中一次抽取一条记录。

2)对抽取到的记录进行属性值的分类。

3)根据分类结果，首先判断是否能够利用它的所有的第一类属性值来作出正确的决策；如果不可以，则从第二类属性值中一次一个地先补充该肽中锚点位置的氨基酸所对应的属性值；若第二类属性值抽取完后，还不能作出正确的决策，就从第三类属性值中开始一次一个地把该肽中锚点位置的氨基酸所对应的属性值补充进来，直至能够作出正确的决策为止。

4)将在当前记录中为作出正确决策而使用到的属性和属性值以及决策属性值保存到规则集 R 中。

5)将决策表中能利用当前规则作出正确决策的记录删除，产生新的决策表。

6)重复 1)～5)，直到决策表为空。

集合 R 中的规则是有优先级别的，即越先产生出的规则，也就是顺序号越小的规则，优先级越高。因此，当使用规则集做决策时，可以按照优先级从高到低的顺序扫描规则，一旦发现某条规则能够适用，则停止扫描。接着，就可以利用该条规则来做决策。当然，如果学习时，按照另外一种顺序逐条学习训练用表中的记录，那么规则集中的规则也将发生变化，而且，当利用得到的规则集来对未学习过的数据进行决策时，不同的规则集得出的结论可能有所不同。

借助于上述算法获得的规则是容易被生物学专家理解的，即对于那些九肽，如果哪些位置出现了某些氨基酸，并且哪些位置未出现某些氨基酸，就可以判断该九肽的亲和力等级。这些规则有助于生物学专家将其注意力集中于某些很可能的关键模式上，并便于生物学专家通过对这些很可能的关键模式的验证和分析来进一步理解蕴含于其中的免疫学机理。

4 实验结果分析

本文的实验数据由 Vladimir Brusic 教授提供。该数据集由基因 HLA-DR4 (B1 * 0401)编码的 MHC II 类分子结合肽构成，共 650 条。肽的长度介于 9 到 27 个氨基酸之间。根据 SYFPEITHI[11]软件，可以获取 HLA-DR(B1 * 0401)结合肽的初级锚点。在固定第一个位置为 F,Y,W,I,L,V 或 M 后，按前述的肽的预处理方法进行预处理。这样，共获取了 915 条九肽。在去掉一些不确定或未知亲和力的九肽后，将剩余 764 条九肽构成一决策表。其中，非结合肽为 553 条，低、中和高亲和力结合肽分别为 49,46 和 116 条。该决策表的条件属性个数和决策属性个数分别为 180 和 1。本文采用四折分层交叉验证采样方法把该决策表分为两部分，即训练部分和检测部分。

首先调用基于信息熵的属性约简完备算法来获取属性约简；在约简后的决策表的基础上，然后调用结合锚点知识的属性值顺序约简算法来获取带顺序的规则集，最后利用检测样本来进行验证，验证结果见表 2。

表 2 本文方法的检测结果

实验序号	非结合肽(%)	低亲和力(%)	中亲和力(%)	高亲和力(%)	平均(%)
1	12.30	79.59	67.39	37.07	23.69
2	10.00	73.47	63.04	50.86	23.43
3	15.37	61.22	69.57	43.10	25.79
4	12.66	69.39	78.26	39.66	24.35
5	13.02	81.63	71.74	37.07	24.61
6	13.56	69.39	63.04	47.41	25.26
7	12.30	59.18	63.04	46.55	23.56
8	14.29	73.47	58.70	34.48	23.82
平均(%)	12.95	70.92	66.85	42.03	24.31

表 3 CRIA 算法的检测结果

实验序号	非结合肽(%)	低亲和力(%)	中亲和力(%)	高亲和力(%)	平均(%)
1	21.70	75.51	76.09	58.62	34.03
2	22.24	79.59	84.78	48.27	33.64
3	21.34	71.43	71.74	52.59	32.33
4	22.24	79.59	76.08	55.17	34.16
5	21.16	75.51	80.44	52.59	32.98
6	21.34	77.55	76.09	56.90	33.64
7	22.42	85.71	76.09	55.17	34.69
8	20.98	79.59	71.74	53.45	32.72
平均(%)	21.68	78.06	76.63	54.10	33.52

为了比较，本文利用了两个不同的方法来处理同一个数据集。第一个简称为CRIA算法，包含了文[12]的属性约简算法和文[13]的属性值约简算法，其试验结果见表3；第二个是采用误差反传前馈神经网络(BPNN)，其网络模型为180-4-1，即输入层、隐含层和输出层节点数分别为180、4和1，学习算法采用反向传播算法，激活函数为Sigmoid函数，学习率和动量因子分别为0.2和0.9。试验结果见表4。

从列举在表2、3和4中的检测结果可看出，表2中的平均误判率(约24.31%)很接近于表4中的平均误判率(约22.43%)，并且远远低于表3中的平均误判率(约33.52%)。这说明了，与CRIA算法相比，由本文提出的基于粗集的T细胞表位预测方法得到的规则集具有较强的泛化能力。与BPNN算法的"黑箱性"相比，本文方法能在保证预测性能的前提下，提取出易于专家理解的产生式规则。

表4 BPNN算法的检测结果

实验序号	非结合肽(%)	低亲和力(%)	中亲和力(%)	高亲和力(%)	平均(%)
1	11.03	63.27	73.91	44.83	23.30
2	10.67	77.55	78.26	37.93	23.17
3	9.77	83.67	78.26	36.21	22.64
4	9.77	85.71	80.44	34.48	22.64
5	9.77	69.39	78.26	41.38	22.51
6	9.95	81.63	73.91	32.76	21.86
7	9.22	77.55	73.91	37.07	21.73
8	8.50	79.59	80.44	36.21	21.60
平均(%)	9.83	77.30	77.17	37.61	22.43

结论 为了解决现有的基于神经网络的T细胞表位预测模型所固有的"黑箱性"问题，本文巧妙地将T细胞表位预测领域知识融入到基于粗集理论的知识获取方法中，提出了基于粗集的T细胞表位预测方法。实验结果表明：本文提出的方法具有比CRIA方法更强的泛化能力；而与基于神经网络的T细胞表位预测方法相比，本文方法能在大致不降低预测精度的前提下，获取易于专家理解的产生式规则。这些规则有助于生物学专家将其注意力集中于某些很可能的关键模式上，并便于生物学专家通过对这些很可能的关键模式的验证和分析来进一步理解蕴含于其中的免疫学机理。

参考文献

1 陈慰峰. 医学免疫学. 北京：第三版. 人民卫生出版社，2000

2 Markus S, Toni W, Stefan S. Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens[J]. Journal of Immunological Methods, 2001, 257: 1～16

3 Gulukota K, Sidney J, Sette A, et al. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. Journal of Molecular Biology, 1997, 26: 1258～1267

4 Brusic V, George R, Margo H, et al. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network[J]. Bioinformatics, 1998, 14: 121～130

5 Kun Y, Petrovsky N, Schonbach C, et al. Methods for prediction of peptide binding to MHC molecules: a comparative study. Mol Med, 2002, 8(3): 137～48

6 Pierre D, Arne E. Prediction of MHC class I binding peptides using SVMHC. BMC Bioinformatics, 2002, 3(1): 25

7 Pawlak Z. Rough sets. International Journal of Information and Computer Sciences, 1982, 11: 341～356

8 王国胤. Rough集理论与知识获取. 第1版. 西安交通大学出版社，2001, 17, 118～119

9 王国胤，于洪，杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报，2002, 25(7): 759～766

10 Dan P, Zh Qi-Lun, An Z, H Jing-Song. A novel self-optimizing approach for knowledge acquisition[J]. IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems and Humans, 2002, 32: 505～514

11 Rammensee H, Bachmann J, Emmerich NP, et al. SYFPEITHI: Database for MHC ligands and peptide motifs[J]. Immunogenetics, 1999, 50: 213～219

12 吴福保，李奇，宋文忠. 基于粗集理论知识表达系统的一种归纳学习方法. 控制理论与决策，1999, 14(3): 206～211

13 Pawlak Z, Slowinski R. Rough set approach to multi-attribute decision analysis. European Journal of Operational Reaserch, 1994, 72: 443～459

(上接第209页)

表3 检测结果比较

算法	离群点数	正确率	误检率	漏检率	运行时间(秒)
SOF	4	100%	0%	0%	4.3
SLOM	3	75%	25%	25%	6.1
SLZ	3	75%	25%	25%	4.1

在SLOF的计算中取$\delta=0.01$，计算结果如表2，如果δ取更小，slof的顺序不变，其值改变，且随着δ变小，彼此的偏差拉大，因此从这种角度看δ越小越好，但也扩大了SLOF的取值范围。

结论 基于SLOF的算法充分考虑了空间数据的特点，根据空间关系确定空间邻居，减少了用户指定参数，用计算邻域距离和空间局部离群系数的方法解决了空间自相关性和异质性问题，并通过对离群系数按降序进行排序，解决了离群点的判断问题。通过与相关算法SLZ和SLOM算法的比较，可以看出SLOF算法在检测精度、计算效率和对用户的依赖性方面均优于其它算法，尤其是在检测的正确性上更为明显。未来通过对SLOF算法的改进，将其应用到基于时间与空间约束的时空离群点的检测中。

参考文献

1 Han Jiawei, Micheline K. Data mining: concepts and techniques [M]. San Francisco: Morgan Kaufmann Publishers, 2001

2 Hawkins D. Identification of Outliers[M]. London: Chapman and Hall, 1980

3 Shekhar S, Lu C T, Zhang P. A Unified Approach to Spatial Outliers Detection[J]. GeoInformatica, 2003, 7(2): 139～166

4 Sanjay C, Sun Pei. SLOM: a new measure for local spatial outliers[J]. Knowledge and Information Systems, 2006, 9(4): 412～429

5 Shekhar S, Chawla S. A Tour of Spatial Databases[M]. Upper Saddle River, N. J.: Prentice Hall, 2003

6 Xue Anrong, Ju Shiguang. Algorithm for Spatial Outlier Detection Based on Outlying Degree[C]. In: Proceedings of the WCICA 2006, Dalian: IEEE Press, 12(7): 6005-9