用于未知病毒检测的免疫识别模型和算法研究*)

鲍欣龙　马建辉　罗文坚　曹先彬　王煦法
(中国科学技术大学计算机科学技术系　合肥230027)

摘　要　现有的反病毒技术难以识别与处理新的未知病毒。本文借鉴生物免疫系统识别未知病毒的机制，以非我识别机制为基础，进一步抽取免疫进化学习机制和阳性/阴性选择机制，提出了一种新的检测器和自我均自适应变化的免疫识别模型和算法。文中给出了算法的详细实现步骤，并针对几种实际病毒进行了检测实验。测试实验结果表明该算法能够检测到未知病毒，具有很好的应用前景。

关键词　未知病毒检测，人工免疫，非选择，进化学习，阳性/阴性选择

Research on Immune Recognition Model and Algorithm for Unknown Viruses Detection

BAO Xin-Long, MA Jian-Hui, LUO Wen-Jian, CAO Xian-Bin, WANG Xu-Fa
(Department of Computer Science and Technology, University of Science and Technology of China, Hefei 230027)

Abstract　Lacking the ability of recognizing and processing the unknown viruses is one of the major limitations of common anti-virus technologies. Inspired by the vertebrate immune system, this paper puts forward and accomplished a novel immune recognition model and algorithm. Referencing the immune evolutionary learning mechanic and positive/negative selection mechanic, this model and algorithm make the detectors and "self" to change by self-adaptation. This paper gives out the detailed implementation of this algorithm. Experiments on several real viruses prove that this algorithm has a good ability on recognizing unknown viruses and wide prospective application fields.

Keywords　Unknown viruses detection, Artificial immunity, Negative selection, Evolutionary learning, Positive/negative selection

随着 Internet 的迅速发展和普及，计算机病毒也进入了一个新时期[1]。现有的计算机反病毒技术主要是基于特征码匹配的方法，难以识别与处理未知的新病毒，导致很多新病毒都是在造成大规模破坏之后才被发现。因此，在当前新病毒层出不穷的环境下，探索和研究未知病毒检测的新思想和新方法具有重要的意义。

生物免疫系统识别未知病毒的机制为解决这一问题提供了新的思路。虽然相关研究才刚刚开始，但已经表现出了诱人的应用前景。非我识别机制是这种方法的核心，它体现了生物免疫系统根据自我来识别非我的本质。当前国际上比较通用的方法是美国新墨西哥大学的 S. Forrest 教授根据 T 细胞介导的细胞免疫机制提出的一个简单的非选择(NS: Negative Selection)算法[2,3]，其核心思想是从随机生成的串中除去和自我相匹配的部分，用剩下的串(称为检测器)当作非我和待检测的串匹配。Forrest 教授在其论文中用数学方法证明了这样做是有效的，但其前提是检测器集完备和自我数据完整[3]。检测器集不完备导致检测率降低，自我不完整导致误报率增高。

然而，在实际应用中，特别是在计算机这种复杂多变的环境下，一方面由于非我空间非常大，难于获得完备的检测器集；另一方面，要获得完整的自我数据也非常困难。甚至由于计算机程序的动态性，自我本身都在动态变化，这就更增添了获取完整自我的难度。总之，如何获取尽可能完备的检测器集和完整的自我，一直是国际上的研究难点。

本文深入考察了生物免疫系统中的抗体生成过程，首先抽取其进化学习机制，结合基因库[4]进化学习机制提出了检测器集自适应变化的免疫识别模型和算法；然后进一步抽取了 T 细胞分化发育过程中的阳性/阴性选择机制，将该机制和 NS 算法及基因库进化机制结合在一起，提出一个检测器集和自我均自适应变化的免疫识别模型和算法，并给出了算法的具体实现方案。这样，该算法不仅使得检测器集能够根据实际情况自适应变化，而且使得自我能够在阳性/阴性选择机制的调节下实现动态增长，在一定程度上弥补了检测器集难于完备和自我无法获取完整的缺陷。对测试集中选定的几种实际病毒的检测实验结果也表明这种改进是合理且有效的。

本文内容安排如下：第1节简介模型和算法的生物免疫基础，第2节介绍自适应免疫识别模型以及相应的算法步骤，第3节是算法实现及实验结果，最后是结束语。

1　生物免疫基础

1.1　生物免疫识别的基本机制

免疫系统最重要的功能[5]就是对自我和非我抗原的识别。根据著名免疫学家 Burnet 提出的克隆选择学说[6]，人体内约存在着10^5～10^7具有免疫活性的细胞克隆，每一克隆细胞都具有其特异的、能与相应抗原决定簇起反应的受体。但处于未成熟阶段的免疫细胞接受自身抗原的刺激，使得自身反

*)基金项目：国家自然科学基金(69971022)和国家863基金(2002AA142130)资助项目。鲍欣龙　硕士生，研究兴趣是人工免疫算法，网络安全；马建辉　博士生，研究兴趣是人工免疫算法，网络安全；罗文坚　博士后，讲师，研究兴趣是人工免疫算法，网络安全；曹先彬　博士，副教授，研究兴趣是计算智能，多主体系统，网络安全；王煦法　教授，博导，研究兴趣是计算智能，智能信息处理。

应性细胞克隆在早期即被淘汰杀死，故发育成熟的免疫系统对自身抗原耐受，而只对外界抗原产生反应。外界抗原进入后，机体根据免疫细胞表面受体和抗原表位间耦合的紧密程度来选择出高匹配度的那部分免疫细胞，使之激活并增殖，产生专门针对该抗原的抗体，引起适当的免疫反应，最终将其清除。图1是一个基于克隆选择学说的免疫抽象模型。

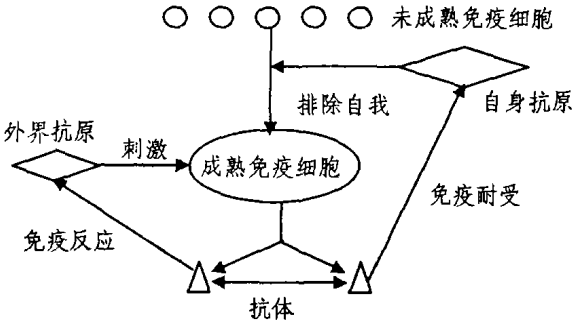

图1　基于克隆选择学说的免疫抽象模型

这里需要特别指出的是，为了快速生成高效的免疫细胞（或抗体），生物免疫系统在免疫细胞生成阶段通过免疫进化学习，主要是基因库进化机制，增加好的基因的选择概率，使得新生成的免疫细胞能够更好地识别抗原。

1.2　阴性/阳性选择机制

T细胞是在胸腺中分化发育成熟的，而阴性/阳性选择机制则是T细胞生成过程中的一个重要机制。其主要过程如下[7]：

首先，从胚肝和骨髓发展出来的未成熟T细胞在多种趋化因子的作用下进入胸腺，发展为表面不能表达CD4和CD8分子的细胞。不表达通常称作“阴性”，表达称作“阳性”，这种细胞就称为“双阴性”，记作DN(double negative)。然后，DN细胞在多种细胞因子的影响下增殖，并经过基因重排，成为既表达CD4又表达CD8的所谓“双阳性”细胞，记作DP(double positive)。DP最终才发展为仅表达一种分子的成熟T细胞。

在胸腺皮质区，若DP的表面受体能够识别的基质细胞是单核吞噬细胞等表达的MHC分子和自身抗原的复合物，则DP细胞发生程序性死亡(programmed cell death，PCD)，不能识别者才能存活与增殖。这一过程称作阳性/阴性选择。

现代免疫学认为，阳性/阴性选择何时发生决定于TCR和抗原复合物的总亲和力(avidity)。总亲和力与TCR和抗原复合物的内在亲和力(affinity)及这两种物质的浓度有关，等于它们的乘积。总亲和力在较低水平时导致阳性选择，达到较高水平则导致阴性选择。

2　自适应免疫识别模型和算法

2.1　检测器集自适应变化的免疫识别模型和算法

基于Burnet克隆选择模型，抽取其免疫细胞的高效生成机制，即基因库进化，我们提出了一个检测器集自适应变化的免疫识别模型。该模型将正常程序的行为对应于自我，检测器对应于免疫细胞，未成熟检测器通过非选择使得其中和自我匹配的部分凋亡，剩下的成为成熟检测器。待检测程序的行为对应于外界抗原，它通过匹配来刺激成熟检测器并使之产生分化，再经过免疫进化来调节检测器的分布。该模型框架如图2所示。

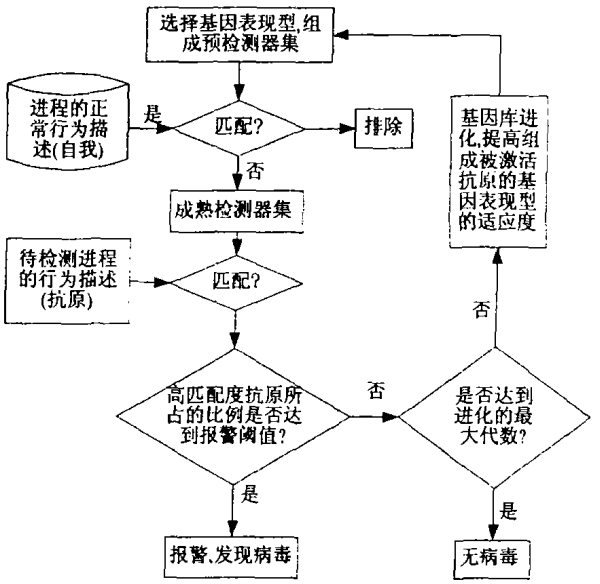

图2　检测器集自适应变化的免疫识别模型

如图2中，该模型主要包括NS算法[2]和基因库进化算法[4,8]，前者的主要作用是过滤自我数据，后者的主要作用是在保持多样性前提下根据抗原与检测器的匹配程度来调节检测器的生成及分布。

结合以上所述的生物免疫识别机制，下面给出相应的检测器集自适应变化的免疫识别算法。具体步骤如下(记为算法2.1)：

(1) 获取正常进程的操作序列，作为自我。

(2) 生成成熟检测器集。通过概率在基因库中选择基因表现型组成预检测器，再使用非我选择算法从中除去自我。最终生成的检测器集对应于生物免疫中的免疫细胞，可以用于检测病毒的异常行为。

(3) 将待检测进程的操作序列与检测器集相匹配。这些操作序列对应于生物免疫中的抗原，根据匹配程度超过激活阈值的抗原所占的比例来判断是否有病毒。若比例达到一定的阈值则报警；否则将被激活的检测器根据匹配程度提高其危险度值，反馈回基因库。

(4) 基因库进化。析构上一步反馈给基因库的检测器，根据其危险度来提高相应基因表现型的适应度。

(5) 转(2)，继续进化，直至超过最大进化代数。若激活抗原比例一直没有超过报警阈值则可判定为无病毒。

至此，本文针对检测器集的完备性，提出了一个用于计算机反病毒的检测器集自适应变化的免疫识别算法。

2.2　自我和检测器集均自适应变化的免疫识别模型和算法

引言中已经提到，在计算机这个复杂的环境下，自我是很难获取完整的，这就要求相应的识别模型和算法能够实现自我数据动态更新。而在现有模型和算法中，包括2.1小节的算法，其自我均是静态的。这就使得一部分自我没有被包括在内，这一部分自我就和外界抗原一样会被检测器所识别，导致了系统误报率的增加。

为此，我们进一步借鉴1.2小节中提到的阳性/阴性选择，引入了动态自我的概念，对2.1小节的模型和算法进行了改进。阳性/阴性选择的发生依据是检测器与抗原的“亲和力”，即它们的匹配程度。对于和抗原匹配程度高的检测器，通过将其反馈回基因库而提高了它的基因在下一代检测器集中的浓

度，即发生了阳性选择；对于和抗原匹配程度低的检测器，将其加入到自我中，本质上也就是降低了它的基因在下一代检测器集中的浓度，即发生了阴性选择。同时，为了确保自我的安全性，我们规定了只有和自我的相似程度达到一定阈值T_{self}的检测器才能被加入到自我中。图3是改进后的自我和检测器均自适应变化的免疫识别模型。

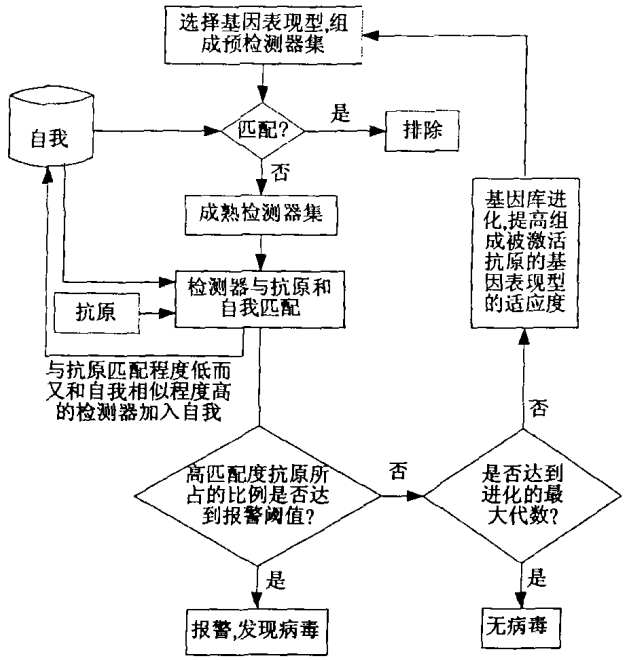

图3　自我和检测器均自适应变化的免疫识别模型

对应图3，下面给出改进后的自我和检测器均自适应变化的免疫识别算法。具体步骤如下(记为算法2.2)：

(1)获取正常进程的操作序列，作为初始自我。

(2)将基因库进化算法生成的预检测器集与自我相匹配，从中排除与自我相同的检测器。

(3)将上一步生成的检测器与抗原匹配：匹配程度超过阈值T_1的检测器反馈回基因库，提高组成它的基因的适应度；匹配程度低于阈值T_2($T_1>T_2$)的检测器再和自我匹配，和自我的相似程度超过阈值T_{self}的检测器加入到自我中。

一般NS算法的生物免疫基础是T细胞的分化发育过程，而阳性/阴性选择本身就是T细胞分化发育过程中的一种浓度调节机制，因此，将阳性/阴性选择的思想引入到NS算法中从免疫机制上说是合理的。上述算法步骤的前两步仍然保留了算法2.1的具体步骤，第三步则对应于阳性/阴性选择的思想，其目的是希望自我集能够在检测的过程中自行扩充，尽可能地在保证自我安全的前提下将初始时没有获得的自我加入到系统的自我集中，以此来提高系统的检测效果。

3　算法实现和实验结果分析

3.1　算法实现

下面我们给出了第2节算法具体实现过程中的关键问题及解决方案。

(1)程序动态"行为"的定义。为了验证模型和算法的实际效果，首先要定义相应的程序"行为"模式，以区别自我和非我。本文选择程序运行时所作的文件操作作为程序的"行为"定义，因为病毒的感染和破坏离不开对文件的读写操作，而且它所做的操作往往是正常程序不会去做的。我们选择了文件操作的三个参数：操作名，操作对象和操作结果；并对其进行了编码以方便操作。

需要注意的是，单个的文件操作是没有任何意义的，程序的动态行为蕴含在文件操作序列的时序之中。为了体现时序性，我们用定长的滑动窗口处理文件操作序列，得到了如图4所示的结构。其中，每一行对应一个文件操作，相邻的行对应相邻的文件操作。为了便于匹配，我们把自我、抗原和检测器都使用图4所示的结构来定义。

操作名1	操作对象1	操作结果1	危险度1
操作名1	操作对象2	操作结果2	危险度2
⋮	⋮	⋮	⋮
操作名n	操作对象n	操作结果n	危险度n

图4　检测器的结构

(2)确定检测器集的规模。群体规模的确定受进化操作中选择操作的影响很大。由模式定理可以知道，检测器集的规模越大，它作为一个群体的多样性就越高，算法陷入局部最优的危险就越小。所以，从群体多样性出发，检测器的规模应该较大。但是，检测器的规模过大又会严重影响算法的性能。由于已经引入了检测器自适应变化策略，本文将检测器规模定在10^2数量级上。

(3)抗原与检测器的匹配。抗原与检测器的匹配在对应行之间进行：操作名域先比较，若相等则继续比较后两个域，操作名不一样时不再比较后两个域，这是因为对于文件操作来说，操作名是最根本的，操作名不同的两个文件操作是完全不同的，不管它的对象和结果怎样。

在检测器的结构中我们加入了一个"危险度"域，用于记录抗原与检测器匹配的结果。每一行的三个域都有各自的危险度权值，匹配成功则将权值加入该行的危险度域中。只有当所有匹配完成后各行危险度域的总和超过了一个预先定义的刺激阈值的时候，才能认为该检测器被这个抗原激活了，这对应于生物免疫细胞表面受体对抗原表位的识别过程中的亲和度概念。也只有被激活的检测器才能被反馈回基因库，对下一代检测器的生成产生影响。

(4)基因库的结构。我们根据检测器的结构定义了相应的基因库结构。检测器中的每个位置对应与基因库中的一个次级库(即对于长度为n的检测器，基因库中有$3*n$个次级库)，每个次级库中包括了一系列预定义好的基因表现型及其相应的适应度。基因库将反馈的检测器拆开，根每个位置上的内容来调节相应次级库中该表现型的适应度，每个次级库都是独立进化的。组成预检测器的时候，也是独立的从每个次级库中使用轮盘赌的方法选择一个基因表现型，按照位置关系拼接起来，成为一个完整的预检测器。

以上给出了算法2.1和算法2.2在具体实现时各种关键问题的实现方案。在算法2.1和算法2.2的基础上，结合以上具体实现方案，就可以设计出针对未知病毒识别的系统。

3.2　实验及结果分析

(1)实验1：验证算法2.1的有效性

为了检验算法2.1的效果，我们应用算法2.1对几种有代表性的病毒进行了多次实验检测，结果如表1所示。从检测效果来看，算法2.1对于未知病毒有着良好的识别率(平均70%)

(下转第94页)

SPINE 既允许使用系统自身提供的标准 PROXY，也允许用户定义自己的 PROXY。用户可以通过自定义的 PROXY 实现对象预装载等功能，这可以显著地提高系统的性能。

结论 SPINE 是一个为 Virtual Helpdesk 系统而开发的专用永久对象管理器。它允许同时存在一个对象的多个拷贝，并通过乐观锁来解决对对象的访问冲突；它使用三层对象设计模式把对象应用层、对象管理层和数据访问层的代码都封装在一个类之中；它还通过代理设计模式实现了关联对象的延迟装载。这些算法和设计模式可以为类似系统的设计提供借鉴。

参 考 文 献

1 Tesch T, Volz M. A Lightweight Object Manager for Group-Aware Applications: [GMD Report 47]. 1999
2 Coldwey J, Keller W. Multilayer Class. 2001
3 Keller W. Persistence Options for Object-Oriented Programs. 2004
4 Keller W. Mapping Objects to Tables - a Pattern Language. In: Proc. EuroPLoP 1997
5 Keller W. Object/Relational Access Layers - a Roadmap, Missing Links and More Patterns. 1998
6 Ambler S W. The Design of a Robust Persistence Layer For Relational Databases. White paper, 2000

（上接第76页）

和较低的误报率（平均18%）。

表1 验证算法2.1的有效性

病毒全名	W97M.NSI.C	O97M.Tristate.C	W97M.Chaos.A	VBS.Haptime.A@mm
病毒类型	宏病毒	宏病毒	宏病毒	邮件蠕虫
检测器集大小	50	50	50	10
刺激阈值	22	22	21	16
进化最大代数	1500	1500	1300	1500
检测率(10次)	80%	40%	90%	70%
误报率(10次)	20%	20%	10%	20%

注：病毒全名依据 Norton AntiVirus Virus Definitions[9]。

(2)实验2：验证算法2.2的有效性

为了检验算法2.2的改进效果，我们分别用算法2.1和算法2.2对选定的几种有代表性的宏病毒进行了多次检测实验。

实验相关参数设置如下：初始自我集大小都是8807；每代生成成熟检测器的数量为500；进化的最大代数为500代。

表2 验证算法2.2的有效性

病毒全名	算法2.1	算法2.2	自我集扩充后的大小(平均)
W97M.Aliv	40%	80%	19214
W97M.OutlookWorm.Gen	60%	90%	16443
W97M.Class.A.Gen	50%	70%	23248
W97M.Kolop	0%	40%	23924

从表2可以看出，算法2.2的检测率相对于算法2.1有一定程度的提高，这也验证了算法2.2中所作的改进是有效的。

(3)实验3：算法2.1和算法2.2的检测速度比较

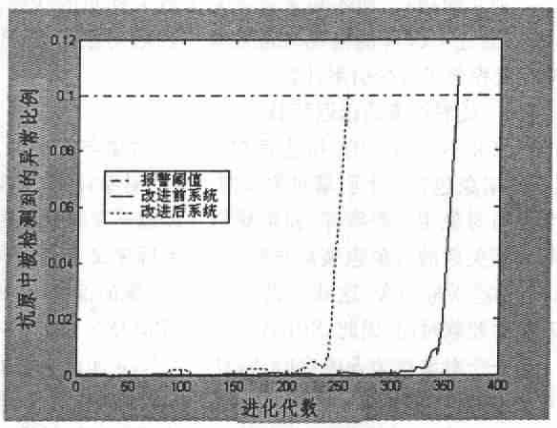

图5 W97M.Nort.A.Gen 病毒在改进前后系统中的检测曲线

本实验将考察了算法2.1和算法2.2的对这些病毒的检测速度。图5给出了算法2.1和算法2.2对其中一个宏病毒，即W97M.Nort.A.Gen 的检测曲线。

图5中改进前系统检测曲线是算法2.1的检测曲线，而改进后系统检测曲线是算法2.2的检测曲线。从图中可以看出，对于病毒 W97M.Nort.A.Gen，算法2.1需要361代的进化才能检测到，而算法2.2只需要257代。这说明了算法2.2中所做的改进使得病毒检测速度有较大的提高。

结束语 本文基于生物免疫中进化学习和阳性/阴性选择机制，提出了一种用于未知病毒检测的检测器和自我均自适应变化的免疫识别模型和算法，用以克服非我空间庞大和自我无法完整获取对检测率和误报率所产生的负面影响。通过对测试集中的几种实际病毒的检测实验，证明了该模型和算法是合理且有效的。

从本文的实践也可以看出，生物免疫中仍然有许多可以借鉴的机制等待我们去发掘，特别是对于人工免疫与网络安全研究中所遇到的问题，回到生物免疫中去寻找答案往往是直接而有效的。当然，基于人工免疫的反病毒研究还处于起步阶段，还有很多工作值得进一步研究。本文的方法也有一些应深入展开的内容，如基于大样本集的实验测试等，需要在以后的工作中逐步深入。

参 考 文 献

1 Chess D. The Future of Viruses on the Internet. http://www.research.ibm.com/antivirus/SciPapers/Chess/Future.html
2 Forrest S, Hofmeyr S, Somayaji A. Computer Immunology. Communications of the ACM, 1997, 40(10): 88～96
3 Forrest S, Perelson A S, Allen L, Cherukuri R. Self-nonself Discrimination in a Computer. In: Proc. of the 1994 IEEE Symposium on Research in Security and Privacy, Los Alamitos, CA. IEEE Computer Society Press, 1994
4 Kim J, Bentley P. An Artificial Immune Model for Network Intrusion Detection. In: Proc. of 7th European Congress on Intelligent Techniques and Soft Computing (EUFIT'99), Aachen, Germany, 1999
5 龙振洲. 医学免疫学. 人民卫生出版社(第二版), 1996
6 Burnet F M. The Clonal Selection Theory of Acquired Immunity. Vanderbilt University Press, Nashville TN, 1959
7 漆安慎, 杜婵英. 免疫的非线性模型. 上海科技教育出版社, 1998. 74～76
8 Luo Wenjian, Zhang Sihai, Liang Wen, Cao Xianbin, Wang Xufa. NIDS Research Based on Artificial Immunology. Journal of University of Science and Technology of China, 32(5): 530～541
9 http://www.symantec.com