动态克隆选择算法有如下几个步骤:

1)初始化:随机生成一个属性串(免疫细胞)的群体。

2)群体循环:对每一个抗原:①选择那些与抗原具有更高亲和力的细胞进行检测;②变异产生新的免疫细胞,遗传上代免疫细胞的特性,同时经自体耐受后,成长为成熟免疫细胞,参与免疫计算,如不能经过自体耐受,则将该免疫细胞去掉,并把其特性作为自体加入自体集合;③亲和力计算:计算每个变异后的细胞与抗原的亲和力。

3)循环:重复步骤2),直到一个给定的收敛标准被满足。

2.2　对动态克隆选择算法的改进

按照动态克隆选择算法,变异产生的免疫细胞可能与该家族病毒的亲和力较小,既不能达到其他家族病毒的亲和力,也不能识别本家族的病毒,导致免疫计算效率降低。为此,在遗传变异中引入可控变异方法,提高识别本家族病毒的效率,引入随机变异方法,提高识别其他家族病毒的效率。另外,新病毒往往与抗体的亲和力值比二次应答的值低,而导致被误认为是自体,发生错误肯定,引入疑似病毒库,进行深入判定,能有效降低错误肯定率。

2.2.1　可控变异

可控变异发生在记忆检测器库中,新加入的记忆检测器,代表当前系统被感染的病毒特征,且有相同属性值 ab_memory. age = 0。可控变异方法:$\forall ab \in \{|x.\ age = 0 \cap x \in \{记记检测器\}\}$,依据亲和力计算方法,必然与某抗原具有 r 个连续相同字符,在抗体变异的产生子代抗体中,该 r 个相同位置相同字符保持不变,其余 $N-r$ 个字符按照随机函数 random(ab) 在字符空间 Σ 中随机变化。这样就确保了子代抗体至少具有与父代抗体相同的抗原亲和力,而更高亲和力的子代抗体进入记忆检测器库中,子代抗体具有属性值 ab_memory. age = 1 ,原父代抗体进化为死亡。可以证明在可控变异中,抗体与相应抗原的亲和力满足关系:$f(x_{子}, y) \geqslant f(x_{父}, y)$。

2.2.2　随机变异

随机变异发生在成熟检测器(也叫常用检测器 Custom Detector)库中,具体方法:$\forall$ ma_dete $\in \{xx|xx.\ age = T_{ma} - 1 \cap x \in \{成熟检测器\}\}$ 的 N 个字符都在字符空间 Σ 中按照 random(ab) 随机选取,其中 T_{ma}, 表示成熟检测器生命周期。这样变异后的子代抗体与父代抗体有较大差异,确保免疫系统的多样性,扩大了抗体在亲和力变化较宽的范围内搜索,越过局部价值点,可能寻找到更高的亲和力点[4]。

2.2.3　建立疑似病毒库,减少肯定错误和否定错误

在病毒检测过程中,必须防范两类错误:肯定错误和否定错误[5]。肯定错误指:检测器误把病毒当成自体,发生漏检,记为 FP;否定错误是指:检测器误把自体当成病毒,发生误检,记为 FNP。为了降低这两类错误,文章提出疑似病毒的概念。当一个貌似病毒的文件(最大亲和力 $f(x,y)$ 满足:$A\_match > f(x,y) > A_{疑}$)提交给免疫系统时,检测器匹配的结果可能使检测器不能直接判断为病毒,也没有足够证据表明其为正常文件而加入自体库,此时,可以将该文件作为疑似病毒,交给经可控变异和随机变异的检测器再检测,作出准确判断。

2.3　改进后病毒检测算法

改进后的病毒检测算法如图1所示。其中,Mut 是变异次数,Visit 是病毒重复感染文件被记忆检测器识别的次数,M_{max} 是常用检测器进化为记忆检测器前识别病毒次数的阈值,Used 是记忆检测器识别病毒的次数,M_{max-ma} 是最大变异次数,Current 是常用检测器库的当前代数。

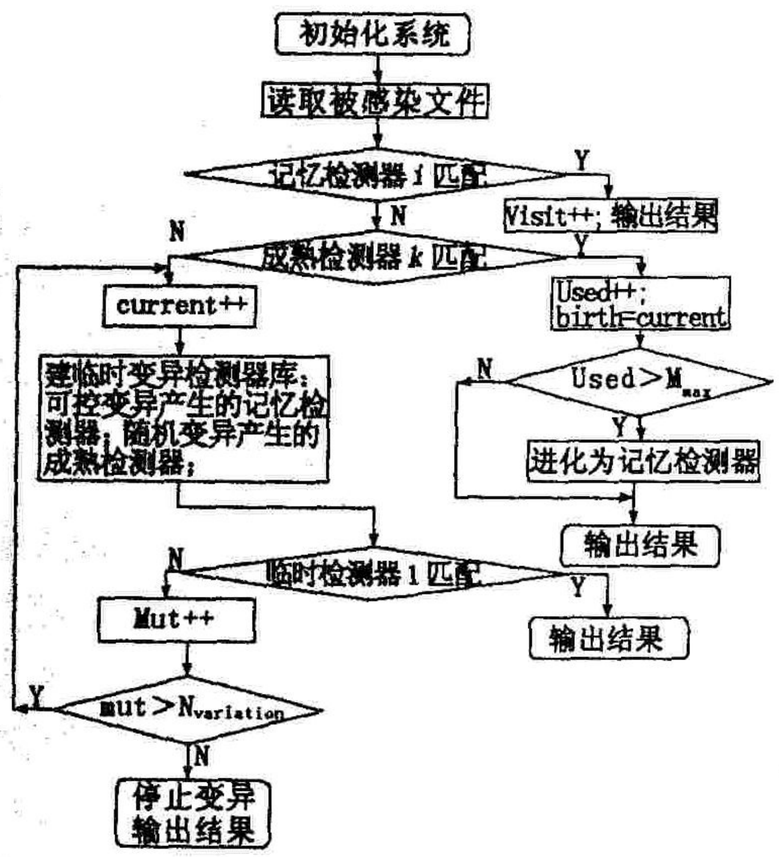

图1　改进的病毒检测算法图

2.4　检测器自体耐受

一般,检测器有两种产生方式:系统初始化时候,受病毒样本刺激产生和父代检测器(包括记忆检测器和常用检测器)变异产生。新检测器自体耐受过程:任意新检测器 x 与动态自体库中的自体抗原 y,如果有:$f(x,y) \geqslant A\_tolerance$,则 x 死亡,否则,x 耐受成功。其中, $A\_tolerance$ 为耐受阈值。

2.5 检测器生命周期

新检测器自体耐受成功,即成长为成熟检测器. 一般来说,检测器的生命周期如图 2 所示。其中,Used 表示检测器识别抗原的次数, T_{max} 表示成熟检测器进化为记忆检测器识别抗原的次数阈值, T_{ma} 表示成熟检测器生命周期, T_{evol} 表示记忆检测器不被使用而进化为成熟检测器的时间阈值,如果记忆检测器不被使用的时间超过该阈值,表明相应的病毒较长时间不活跃,记忆检测器将进化为成熟检测器,进入成熟检测器的生命过程, T_{memo} 表示记忆检测器进化为死亡的时间阈值。

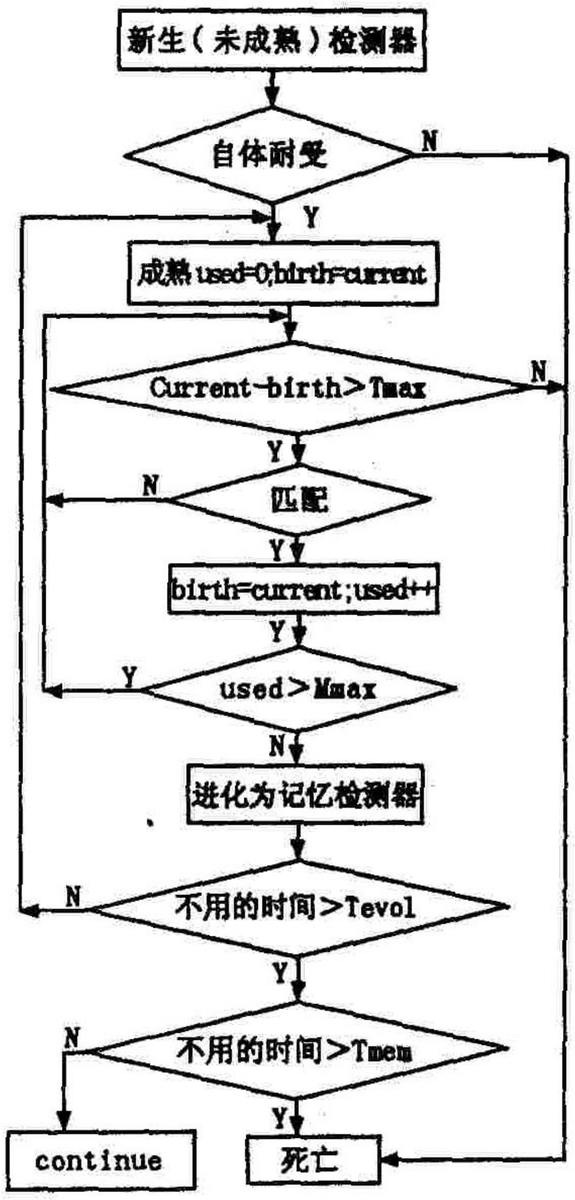

图 2 检测器生命周期图

3 仿真实验结果及分析

本实验以感染病毒的文件为非自体,以 48 个家族病毒(包含变形病毒,共 282 个病毒)的特征码为抗原。系统初始化时,由 32 个家族病毒感染文件,系统通过学习识别病毒文件,获得了一定数量的抗体和自体。病毒检测系统的主要参数设置为: $A\_match = 0.9$、$A\_tolerance = 0.9$、$M_{max} = 10$、$T_{ma} = 550$、$N_{variation} = 10\ 000$、$T_{memo} = 1\ 000\ 000$、$T_{evol} = 80\ 000$。

3.1 二次应答

本次实验采用用 BaBy、HV、2850、DataCrime、Customs 和 Vienna 等 20 种病毒检查系统二次应答,结果自动存放在 E:/testfile/testResult/memery_detector_set 文件中,该文件显示的记忆检测器使用的次数表示二次应答的次数。实验表明:系统具有非常优秀的二次应答能力,而且误判率和误别率均为 0。表 1 记录了对 7 种病毒的二次应答结果. 其中"0"(BaBy_0)表示该病毒特征码没有变异,由记忆检测器来识别。

表 1 对已知病毒的识别能力

病毒名称	Customs_0	DataCrime_0	BaBy_0	2850_0	HV_0	Vienna_0	Jerusalem_0
匹配值	100%	100%	100%	100%	100%	100%	100%

3.2 初次应答

初次应答是系统对新病毒的反应,该指标表示系统的自适应性。图 3 表示了系统对一些新病毒的初次应答,从图 3 可以看出,系统当前正在运行变异的第 49 代,我们可以看到第 48 代完整的情况,随机变异产生了两个检测器:

92e2edb4e8bf310600da4991

3521212ed0ca92e2b41ab42a

并且输出了两个检测器自体耐受中最大的匹配值。随后,在第 48 代这两个检测器被加入到成熟检测器库中,成为具有免疫功能的免疫细胞。同时,成熟检测器库中也将有两个达到生命周期的检测器死亡。

```
temp:b802005be923012e8936760a
singal_max_percent : 0.333333
tottle_max_percent : 0.333333
current:48
radom mutation
self_max:0.375
self_max:0.375
92e2edb4e8bf310600da4991
self_max:0.458333
self_max:0.458333
3521212ed0ca92e2b41ab42a
add:
92e2edb4e8bf310600da4991
3521212ed0ca92e2b41ab42a
death:
a84900da75fa3dfe8c86d0ca
cbb4edb4e8bfa84900da4991
ab:92e2edb4e8bf310600da4991
temp:2600eb03e8b00172085ae805
singal_max_percent : 0.291667
singal_max_percent : 0.291667
ab:3521212ed0ca92e2b41ab42a
temp:ffcbb42fdab41aedb41abab4
singal_max_percent : 0.333333
tottle_max_percent : 0.333333
current:49
radom mutation
```

图 3 系统对 Baby 病毒的初次应答图

3.3 算法对免疫应答效率的影响

由于可控变异的引入,抗体变异的效率更高,与流行病毒抗原亲和力更高的抗体保留在记忆检测器库中,当流行病毒感染时,二次应答的速度更快。如图 4 所示,改进后的系统对 Tiny 病毒的二次应答比改进前的效率更高。随机变异改进了克隆选择算法,提高系统免疫学习的效率。如图 5 所示,改进后系统对 Jerusalem 病毒家族(系列变形病毒)的初次应答比改进前效率更高。在图 5 中,改进前后系统都随变异代数增加,亲和力增大,显然是克隆选择原理起作用。但改进算法的系统(如图 5 中菱形曲线)有个明显的阶跃点 a。a 点的产生,主要是由于随机变异,产生了与抗原亲和力更高的子代抗体,该子代抗体后经克隆选择遗传变异,能够明显提高成熟检测器的亲和力,明显提高曲线的纵坐标。b 点表示系统识别了攻击。

3.4　疑似病毒库对肯定错误率 FP 的影响

在病毒检测中,最大限度减少肯定错误,是 AIS 研究中的重要课题。通过引入疑似病毒,把那些貌似病毒的文件保留起来,不直接下结论,而是经抗体变异后再检测,提高系统免疫识别效率。如图 6 所示,改进后的系统能显著降低 FP。

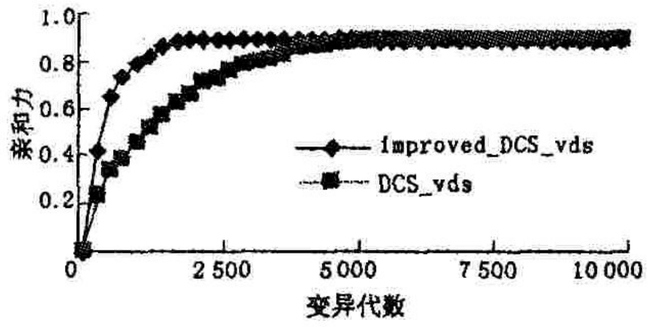

图4　系统二次应答效率提高图

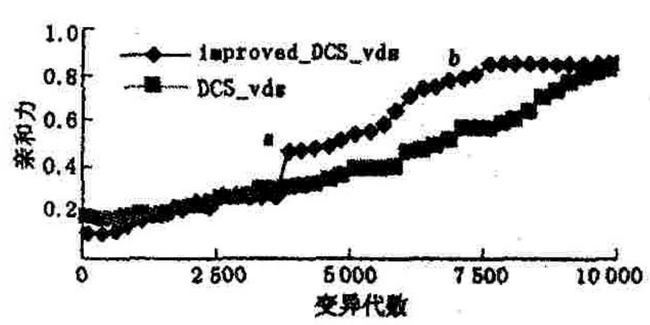

图5　系统初次应答效率提高图

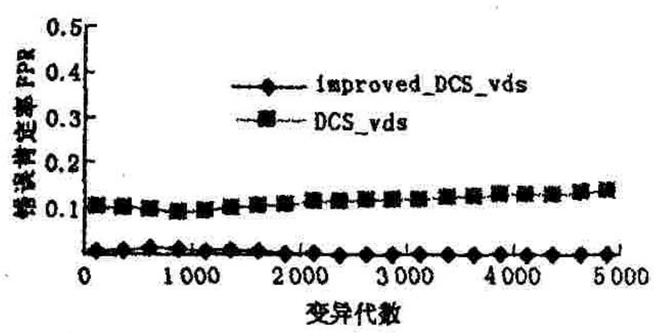

图6　FP 值有效降低图

4　结束语

基于人工免疫的病毒检测模型具有自学习性和自组织性,能够有效识别新的病毒。可控变异与随机变异方法改进的免疫算法能提高病毒检测的效率,疑似病毒库的建立能显著减少协同刺激,降低系统错误率。

参考文献:

[1]　林　闯,彭雪海. 可信网络研究[J]. 计算机学报,2005,28(5):38-42.

[2]　李　涛. 基于免疫的网络安全风险检测[J]. 中国科学,E 辑:信息科学,2005,35(8):789-816.

[3]　Kim J. Bentley P J. Immune Memory and Gene Library Evolution in the Dynamical Clonal Selection Algorithm[J]. Journal of Genetic Programming and Evolvable Machines,2004,5(4): 361-391.

[4]　李　涛. 计算机免疫学[M]. 北京:电子工业出版社,2004.

[5]　Forrest S, Perelson A S, Allen L. Self——Nonself Discrimination in a Computer[A]. Proceedings of IEEE Symposium on Research in Security and Privacy[C]. Oakland, CA, 1994, 202-212.

[6]　许　春,李　涛,刘孙涛,等. 一种改进的动态克隆选择免疫算法在入侵检测中的应用[J]. 空军工程大学学报(自然科学版),2006,7(3):50-54.

(编辑:门向生)

The Research for the Application of an Improved Dynamic Clonal Selection Algorithm to Virus Detection

XU Chun, LI Tao, LIANG Gang, ZHAO Hui, ZHAO Kui, HU Xiao-qin

(Sichuan University, Chengdu 610064, China)

Abstract: Methods of the controllable - aberrance, random - aberrance and an idiographie dynamic clonal selection algorithm are put forward. Sets of suspicions computer virus are established, which contribute to the reduction in the mistakes (FPR) of the immune system. A virus detection model based on the idiographic artificial immune system is put forward and realized. Emulation experiment shows that the model can identify not only old virus, but also new transmutation virus and is a self - adapted, and diverse model for virus detection.

Key words: dynamic clonal selection algorithm; artificial immune; computer virus detection; transmutation virus