受基因理论启发的计算机病毒进化模型

张 瑜[1]，刘庆中[2]，石元泉[3]，曹均阔[1]

(1. 海南师范大学信息科学技术学院 海口 571158；2. 美国萨姆休斯顿州立大学计算机科学系 休斯顿 77340；
3. 怀化学院计算机科学与工程学院 湖南 怀化 418000)

【摘要】作为一种重要的网络空间安全威胁，计算机病毒及其生存与繁衍一直是网络空间安全领域研究热点。该文借鉴生物基因理论和人工生命思想，提出了一种基于基因理论的计算机病毒进化模型：给出了计算机病毒形式化定义，构建了计算机病毒在基因、DNA、染色体等3个层次的进化数学模型，以模拟计算机病毒在自然选择中的进化模式。仿真实验表明，即使遭遇严酷的外部环境，具有算法特性与生命特性的计算机病毒仍具极强的进化能力。

关 键 词 人工生命；计算机病毒；进化；基因理论
中图分类号 TP309 **文献标志码** A doi:10.3969/j.issn.1001-0548.2018.06.014

Gene-Inspired Model for Computer Viruses Evolution

ZHANG Yu[1], LIU Qing-zhong[2], SHI Yuan-quan[3], and CAO Jun-kuo[1]

(1. School of Information Science & Technology, Hainan Normal University Haikou 571158;
2. Department of Computer Science, Sam Houston State University Houston 77340;
3. School of Computer Science and Engineering, Huaihua University Huaihua Hunan 418000)

Abstract As an important sort of cyberspace security threats, computer viruses have been a hotspot in cyberspace security research. Inspired by the gene theory and the artificial life theory, we present a gene-based model of computer viruses evolution. The form definition of the computer virus is given and the mathematical models of computer viruses evolution in the three levels of gene, DNA, and chromosome are constructed. And then the computer viruses evolution process in natural selection is simulated. Experimental results show that, even in face of harsh external environment, computer viruses which have characteristics of both algorithm and life still have strong ability of adaptation and evolution.

Key words artificial life; computer virus; evolution; gene theory

网络信息技术的普适性、软硬件漏洞的客观存在性、数字复制技术的便捷性以及网络犯罪的高收益性，为计算机病毒进化提供了坚实的物质基础与内驱动力。当前，计算机病毒仍是网络空间所面临的主要安全威胁之一。据最新研究报告[1]称：2016年中国计算机病毒感染率为57.88%。其安全威胁[2-4]主要表现为：1) 窃取敏感数据，威胁用户数据隐私安全；2) 篡改系统配置、潜伏于目标系统中，威胁工业基础设施安全；3) 传播网络谣言、勒索用户，威胁现实社会公共安全。

作为一种重要的网络空间安全威胁，计算机病毒的生存、适应、延续、发展等进化理论研究意义重大，且可促进如下技术发展：1) 网络武器技术。作为一种高效的网络战武器，计算机病毒可为网络武器开发与利用提供技术与策略支撑；2) 反病毒技术。从攻防博弈的角度，计算机病毒与反病毒技术之间的魔道之争，促使反病毒技术预测病毒未来进化趋势与发展方向，从而未雨绸缪、防患未然；3) 漏洞修复技术。作为最佳的漏洞利用范例，计算机病毒可促使软件漏洞修复技术的应用与发展；4) 软件演化技术。计算机病毒是一种能自我复制的程序代码，其进化模型能为软件代码进化提供理论与实践支撑；5) 人工生命技术。作为一种绝佳的人工生命体，计算机病毒可为人工生命技术研究提供实验素材。

迄今，计算机病毒及其进化发展一直为信息安全界所关注，研究者从不同视角研究了计算机病毒及其进化规律[5]，概括起来可分为4类：1) 病毒自我复制性，如文献[6]从图灵机的可自我复制性的角度

收稿日期：2016－10－8；修回日期：2017－5－22
基金项目：国家自然科学基金(61462025,61862022, 61262077)；海南省重点研发计划项目(ZDYF2016013)；海南省重大科技计划(ZDKJ2017012)
作者简介：张瑜(1975－)，男，博士，教授，主要从事网络安全、恶意代码分析与取证、智能计算等方面的研究.

讨论了计算机病毒进行自我复制、自我演化的特性；2) 人工生命体的自我复制性，如文献[7-9]从人工生命体的角度讨论了计算机病毒的演化特性；3) 遗传算法，如文献[10-13]从遗传算法的角度研究了计算机病毒演化规律；4) 协同进化论，如文献[14-15]从协同进化的角度研究了计算机病毒进化特性。

然而，上述研究未能有效的从病毒基因视角探索计算机病毒进化发展规律。本质而言，计算机病毒具有双重特性：1) 算法特性。作为一种能自我复制的计算机程序，计算机病毒具有普通程序所具备的算法特性：按照编程者意图，通过所设计的算法逻辑完成相关操作；2) 生命特性。计算机病毒还是一种人工生命体，通过编程者的“上帝之手”，具备自我繁衍、传播感染、适应环境等生命特性。由此可见，在计算机病毒代码世界，病毒基因蕴含着独特的进化意义，有助于洞悉病毒进化内涵，并借此揭开计算机病毒进化之谜：病毒基因的变化能使其更好地适应不断改变的外部环境，为病毒进化提供了独一无二的证据，也为病毒检测提供了明显的特征码支持。因此，如能从病毒基因视角，借助于病毒基因的独特进化内涵，通过算法和生命的双重特性去研究计算机病毒的进化模型，将有助于全面有效地理解计算机病毒进化逻辑。

鉴于此，本文提出一种基于基因理论的计算机病毒进化模型。首先，从生命特性的角度，通过借鉴生物学病毒相关机理与概念定义了计算机病毒及其相关概念；其次，从算法特性的视角，设计了计算机病毒的相关进化算子；最后，从病毒基因的角度，构建了计算机病毒进化模型。仿真实验表明，即使遭遇严酷的外部环境，具有算法和生命双重特性的计算机病毒仍具极强的进化能力。

1　模型理论

1.1　基因理论

基因理论[16]是研究生物体的遗传和变异的科学，是生物学的一个重要分支。基因理论认为：1) 从载体的角度，基因是位于染色体的DNA上的遗传物质；2) 从定义的角度，基因是指带有遗传信息的DNA片段；DNA是由4类不同的核苷酸组成的链状分子，DNA上的核苷酸序列就是生物体的遗传信息；染色体是细胞核内由核蛋白组成、能用碱性染料染色、有结构的线状体，其本质是脱氧核甘酸；3) 从功能的角度，基因通过指导蛋白质的合成来表达自己所携带的遗传信息，从而控制生物体的性状表现；DNA用于长期性的遗传信息储存，以引导生物发育与生命机能运作；染色体是细胞核中载有遗传信息(基因)的物质，主要由DNA和蛋白质组成，是遗传物质的载体。

可知，生物病毒体是由细胞组成，细胞包括细胞核，细胞核内有染色体，染色体是DNA的载体，DNA包含多种基因。生物病毒的遗传结构可用集合代数描述为：基因⊂DNA⊂染色体⊂细胞核⊂细胞⊂生物病毒，其包含关系如图1所示。

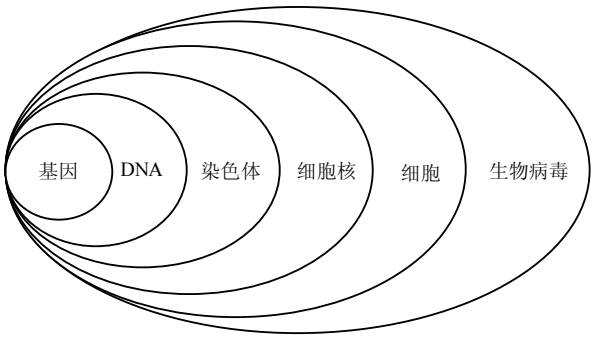

图1　生物病毒基因及其载体

与生物病毒体类似，计算机病毒是一种能自我复制的人工生命体(程序代码)。从程序代码的角度，计算机病毒是编制者借助于计算机语言编写的能在计算机系统中运行的计算机程序代码，其最终表现为大量数字基因的有机组合。从程序结构的角度，计算机病毒包含：进程、过程(函数)、模块、指令等结构。

从抽象逻辑的角度，生物基因理论与计算机程序理论的映射关系[11-12]如表1所示。两者的对应关系为：“基因”对应“特殊指令序列”，“DNA”对应“模块”，“染色体”对应“过程(函数)”，“细胞核”对应“内核进程”，“细胞”对应“应用进程”，“生物体”对应“程序”。

表1　生物基因理论与计算机程序理论的映射关系

生物学基因理论	计算机程序理论
基因	特殊指令序列
DNA	模块
染色体	过程(函数)
细胞核	内核进程
细胞	应用进程
生物体	程序

1.2　相关定义

计算机病毒是指一组能自我复制且具有表现(破坏)作用的计算机指令或程序代码。为完成相关功能，计算机病毒需要相关程序结构支撑，即计算机病毒功能决定其结构。因此，计算机病毒结构是其

充分利用系统资源以完成相关功能的最佳体现。在功能结构上，计算机病毒通常由5个部分组成：初始化模块、感染标记、触发模块、表现模块和感染模块，如图2所示。

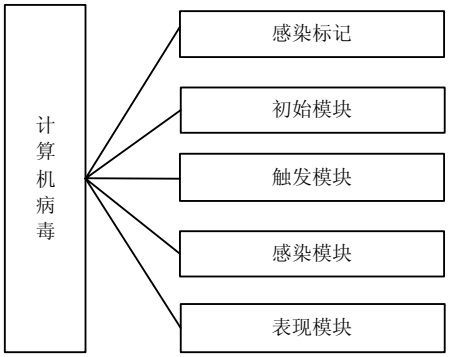

图2 计算机病毒的逻辑结构

由此，可得到计算机病毒的形式化定义如下。

定义 1 计算机病毒是由5个功能模块组成的可自我复制的计算机指令或程序代码集合，表示为：$V=(V_1, V_2, V_3, V_4, V_5)$，其中$V_1$为感染标记集，$V_2$为初始化模块集，$V_3$为触发模块集，$V_4$为感染模块集，$V_5$为表现模块集。

借鉴生物学病毒基因理论，可将计算机病毒结构表示为长度为5的染色体模式，即$V=(V_{c1}, V_{c2}, V_{c3}, V_{c4}, V_{c5})$。由此可得计算机病毒在染色体、DNA、基因等3个层次上的形式化定义如下：

定义 2 计算机病毒染色体是指计算机病毒中具有某些复杂功能的程序模块，表示为：$V_{Ci}=(V_{c1i}, V_{c2i}, \cdots, V_{c5i})$，$i=1,2,3,4,5$。

定义 3 计算机病毒DNA是指染色体中具有某一特定功能的程序小模块，表示为：$V_{DNA}=(V_{c1iDNA}, V_{c2iDNA}, \cdots, V_{cniDNA})$，$i=1,2,3,4,5$。

定义 4 计算机病毒基因是指病毒DNA中为实现某些微操作设计的计算机特殊指令序列，表示为：$V_g=(V_{DNA1}, V_{DNA2}, \cdots, V_{DNAi})$，$i=1,2,3,4,5$。

至此，计算机病毒可表示为$n\times5$阶矩阵，即：

$$\boldsymbol{V}=\begin{pmatrix} \boldsymbol{v}_{11} & \boldsymbol{v}_{12} & \boldsymbol{v}_{13} & \boldsymbol{v}_{14} & \boldsymbol{v}_{15} \\ \boldsymbol{v}_{21} & \boldsymbol{v}_{22} & \boldsymbol{v}_{23} & \boldsymbol{v}_{24} & \boldsymbol{v}_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{v}_{n1} & \boldsymbol{v}_{n2} & \boldsymbol{v}_{n3} & \boldsymbol{v}_{n4} & \boldsymbol{v}_{n5} \end{pmatrix}。$$

1.3 基于基因理论的计算机病毒进化模型

依据基因理论，生物的不断进化得益于遗传与变异。生物的遗传与变异主要包括2类：1) 基因突变与重组；2) 染色体变异。基因突变是指某个基因上的某些碱基对的增添、缺失或替换，从而导致基因结构的改变，产生新的基因；基因重组是生物变异中最普遍的现象，是在有性生殖过程中才会发生的，主要是在减数分裂中染色体的重新组合，像积木推倒重建一样，没有产生新的基因，只是原有基因的重新组合；染色体变异是指染色体片段的缺失，引起染色体上基因的数目或排列顺序发生改变，从而导致性状的改变。上述两类遗传变异，都会导致生物性状与功能的改变，适者生存，不适者淘汰。

计算机病毒的算法特性与生命特性决定了病毒的进化模式：即由低级到高级、从已知到未知、由简单到复杂的螺旋式上升。具体而言，计算机病毒编制者在编写新病毒时，通常采用如下步骤：首先，对已知病毒进行编码分析，并提取病毒各种模块(基因)；其次，应用不同算法对已知病毒的模块(基因)进行修改利用或创新而得到新病毒；最后，新病毒将通过外部环境(运行环境与反病毒软件)考验，适者生存，不适者淘汰。

借鉴基因理论并结合计算机病毒的算法与生命特性，构建基于基因理论的计算机病毒进化模型(如图3所示)。初始病毒群体在病毒进化算子的作用下，生成新生病毒群体。新生病毒群体在自然选择(反病毒软件)作用下，如成功存活则继续进化繁衍，否则淘汰。

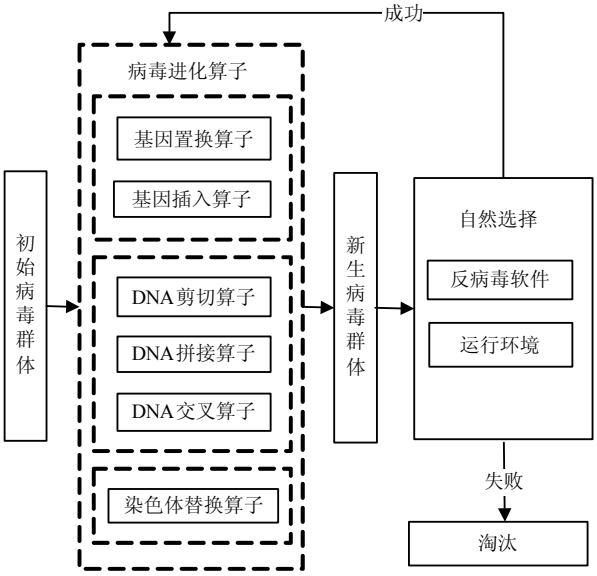

图3 基于基因理论的计算机病毒进化模型

鉴于计算机病毒的代码特殊性，本模型中的病毒进化算子将分别作用于基因、DNA、染色体等3个层次。因此，病毒进化算子可抽象为3类：1) 基因算子；2) DNA算子；3) 染色体算子。其中，基因算子用于计算机病毒基因层面的变异操作，主要包括：基因置换算子和基因插入算子；DNA算子则较基因算子更抽象，作用于计算机病毒DNA的变异，

主要包括：DNA剪切算子、DNA拼接算子、DNA交叉算子；染色体算子只包含染色体替换算子，用于在计算机病毒染色体层面进行变异操作。

1.3.1 基因算子

计算机病毒是一种能自我复制的程序代码，不论由何种计算机语言编写，最终生成的能在计算机系统上运行的只能是二进制形式的机器码。在二进制中，只有两个表示基元：0和1。由二进制基元构成的计算机病毒，将产生多种类型的病毒基因。因此，从数字基因的角度，计算机病毒是由诸多数字基因构建的、有特殊功能的、能自我复制、自动运行的人工生命体。

基因算子[17]将在计算机病毒基因层面进行操作，通过改变病毒某些关键基因而发生基因突变，从而改变其某些性状(特征码)，以逃避反病毒软件查杀，更好适应外部运行环境、最大程度生存与繁衍自身。譬如，通过更换计算机病毒的部分寄存器(EAX、EBX、ECX、EDX等)，或插入部分花指令，可导致其特征码(基因)的改变。基因算子主要包括两类：基因置换算子和基因插入算子，其数学模型分别如下。

1) 基因置换算子：

$$V_{\mathrm{Re}}(t)=\begin{cases}V_{\mathrm{initial}} & t=0\\ V_{\mathrm{Re}}(t-1)-V_{\mathrm{del}}(t)+V_{\mathrm{new}}(t) & t\geqslant 1\end{cases} \tag{1}$$

$$V_{\mathrm{del}}(t)=\{v|v\in V_{\mathrm{Re}}(t\text{-}1)\wedge f_{\mathrm{adaption}}(v)=0\} \tag{2}$$

$$V_{\mathrm{new}}(t)=\{v\,|\,v\in V_{\mathrm{Re}}(t-1)\wedge f_{\mathrm{adaption}}(v)=1\} \tag{3}$$

$$f_{\mathrm{adaption}}(v)=\begin{cases}0 & \text{detected}\\ 1 & \text{otherwise}\end{cases} \tag{4}$$

$$\mathrm{R}(x)=\{x\,|\,\forall v\in V_g, x.v_i\rightarrow x.v_j \quad 1\leqslant i,j\leqslant 5\} \tag{5}$$

式(1)刻画了病毒群体通过基因置换算子的进化过程，其中V_{initial}为初始病毒集，V_{del}为被查杀而删除的病毒集，V_{new}是病毒进行基因置换操作后生成的新病毒集。f_{adaption}为病毒适应度，反映病毒的生存能力，如被查杀则其适应度为0，否则为1。式(5)模拟了借助基因置换算子生成新病毒的过程，即通过在病毒基因库里选择相关的等位基因，置换病毒相应基因座上的基因，从而模拟基因突变以产生新病毒。

2) 基因插入算子：

$$V_{\mathrm{Ins}}(t)=\begin{cases}V_{\mathrm{initial}} & t=0\\ V_{\mathrm{Ins}}(t-1)-V_{\mathrm{del}}(t)+V_{\mathrm{new}}(t) & t\geqslant 1\end{cases} \tag{6}$$

$$V_{\mathrm{del}}(t)=\{v\,|\,v\in V_{\mathrm{Ins}}(t-1)\wedge f_{\mathrm{adaption}}(v)=0\} \tag{7}$$

$$V_{\mathrm{new}}(t)=\{v\,|\,v\in V_{\mathrm{Ins}}(t-1)\wedge f_{\mathrm{adaption}}(v)=1\} \tag{8}$$

$$\mathrm{I}(x)=\{x\,|\,\forall v\in V_g, x.v_i\rightarrow x.v_iv_j \quad 1\leqslant i,j\leqslant 5\} \tag{9}$$

式(6)刻画了病毒群体通过基因插入算子的进化过程。式(9)模拟了借助基因插入算子生成新病毒的过程，即通过在病毒基因库里选择相关的等位基因，插入至病毒相应基因座上，模拟基因突变以产生新病毒。

1.3.2 DNA算子

由基因理论可知，基因重组是一种广泛存在的生物遗传机制，高等生物体、细菌、病毒、原核生物都存在基因重组。基因重组是指一个基因的DNA序列是由两个或两个以上的亲本DNA组合而成，是对基因的重新排列组合。从广义上讲，任何造成基因型变化的基因交流过程，都可视为基因重组。

借鉴生物基因理论，本文将在计算机病毒的DNA层面进行基因进化操作，以促进病毒基因交流与重组，形成具有不同特征码(基因)的计算机病毒，从而导致计算机病毒的多样性。譬如，通过同类计算机病毒之间的模块交换，或在模块中插入部分花指令，可生成功能相似而特征码(基因)不同的新型计算机病毒。DNA进化算子主要包括：DNA剪切算子、DNA拼接算子、DNA交叉算子，其数学模型分别如下。

1) DNA剪切算子：

$$V_{\mathrm{Cut}}(t)=\begin{cases}V_{\mathrm{initial}} & t=0\\ V_{\mathrm{Cut}}(t-1)-V_{\mathrm{del}}(t)+V_{\mathrm{new}}(t) & t\geqslant 1\end{cases} \tag{10}$$

$$V_{\mathrm{del}}(t)=\{v\,|\,v\in V_{\mathrm{Cut}}(t-1)\wedge f_{\mathrm{adaption}}(v)=0\} \tag{11}$$

$$V_{\mathrm{new}}(t)=\{v\,|\,v\in V_{\mathrm{Cut}}(t-1)\wedge f_{\mathrm{adaption}}(v)=1\} \tag{12}$$

$$\mathrm{C}(x)=\{x\,|\,\forall v\in V_{\mathrm{DNA}}, x.v_iv_j\rightarrow x.v_i \quad 1\leqslant i,j\leqslant 5\} \tag{13}$$

式(10)刻画了病毒群体通过DNA剪切算子的进化过程。式(13)模拟了借助DNA剪切算子生成新病毒的过程，即通过剪切掉病毒基因座上的部分等位基因，实现基因重组以产生新病毒。

2) DNA拼接算子：

$$V_{\mathrm{Joint}}(t)=\begin{cases}V_{\mathrm{initial}} & t=0\\ V_{\mathrm{Joint}}(t-1)-V_{\mathrm{del}}(t)+V_{\mathrm{new}}(t) & t\geqslant 1\end{cases} \tag{14}$$

$$V_{\mathrm{del}}(t)=\{v\,|\,v\in V_{\mathrm{Joint}}(t-1)\wedge f_{\mathrm{adaption}}(v)=0\} \tag{15}$$

$$V_{\mathrm{new}}(t)=\{v\,|\,v\in V_{\mathrm{Joint}}(t-1)\wedge f_{\mathrm{adaption}}(v)=1\} \tag{16}$$

$$\mathrm{J}(x)=\{x\,|\,\forall v\in V_{\mathrm{DNA}}, x.v_i\rightarrow x.v_jv_i\vee x.v_i\rightarrow x.v_iv_j \quad 1\leqslant i,j\leqslant 5\} \tag{17}$$

式(14)刻画了病毒群体通过DNA拼接算子的进化过程。式(17)模拟了借助DNA拼接算子生成新病毒的过程，即通过随机选择病毒DNA库中的DNA序列并将其拼接于病毒等位座上，实现基因重组以产生新病毒。

3) DNA交叉算子：

$$V_{\text{Intersection}}(t)=\begin{cases}V_{\text{initial}} & t=0\\ V_{\text{Intersection}}(t-1)-V_{\text{del}}(t)+V_{\text{new}}(t) & t\geqslant 1\end{cases} \tag{18}$$

$$V_{\text{del}}(t)=\{v\mid v\in V_{\text{Intersection}}(t-1)\wedge f_{\text{adaption}}(v)=0\} \tag{19}$$

$$V_{\text{new}}(t)=\{v\mid v\in V_{\text{Intersection}}(t-1)\wedge f_{\text{adaption}}(v)=1\} \tag{20}$$

$$\begin{aligned}\text{Int}(x,y)=\{(x,y)\mid v_i,v_j\in V_{\text{DNA}},x.v_i\leftrightarrow y.v_j\\ 1\leqslant i,j\leqslant 5\}\end{aligned} \tag{21}$$

式(18)刻画了病毒群体通过DNA交叉算子的进化过程。式(21)模拟了借助DNA交叉算子生成新病毒的过程，即通过将一对病毒的等位DNA进行交叉，实现基因重组以产生新病毒。

1.3.3 染色体算子

由基因理论可知，染色体变异是指因染色体片段的缺失、重复、倒位、易位，引起染色体上基因的数目或排列顺序发生改变，遗传信息随之改变，从而导致生物体后代性状的改变。其中，易位是指一条染色体的某一片段移接到另一条非同源染色体上，可改变基因连锁群，造成染色体融合而改变染色体数目，从而引起变异的现象。

借鉴染色体变异理论，利用计算机语言(如Java、C++等)的函数多态性，借助编译时的重载机制或运行时的虚函数机制，通过修改函数参数或函数内部实现来达到更换函数功能与特征码(基因)的目的。本文将在计算机病毒染色体层面进行染色体易位操作，其数学模型如下：

$$V_{\text{Substitution}}(t)=\begin{cases}V_{\text{initial}} & t=0\\ V_{\text{Substitution}}(t-1)-V_{\text{del}}(t)+V_{\text{new}}(t) & t\geqslant 1\end{cases} \tag{22}$$

$$V_{\text{del}}(t)=\{v\mid v\in V_{\text{Substitution}}(t-1)\wedge f_{\text{adaption}}(v)=0\} \tag{23}$$

$$V_{\text{new}}(t)=\{v\mid v\in V_{\text{Substitution}}(t-1)\wedge f_{\text{adaption}}(v)=1\} \tag{24}$$

$$\text{S}(x)=\{x\mid \forall v\in V_C,x.v_i\rightarrow x.v_j\quad 1\leqslant i,j\leqslant 5\} \tag{25}$$

式(22)刻画了病毒群体通过染色体易位算子的进化过程。式(25)模拟了借助染色体易位算子生成新病毒的过程，即通过从病毒染色体库中随机选取基因片段替换病毒的部分染色体，从而实现染色体变异以产生新病毒。

2 实验及分析

2.1 实验环境

本文涉及到实验平台、反病毒软件、病毒样本数、病毒基因库、病毒生产机等实验环境。实验平台模拟计算机病毒所处的外部运行环境，选择由VMWare Workstation V10作支撑平台，通过在其中安装Windows 8操作系统，以模拟真实主机环境。反病毒软件模拟计算机病毒进化过程中所遭遇的天敌，通过查杀(识别与杀灭)来模拟自然选择机制，遵循“物竞天择，适者生存”的进化论逻辑，选择为360杀毒V5.0。

病毒样本采用Windows脚本病毒，样本清单源于国际权威的WildList，并从国际著名的病毒样本组织VXHeavens网站中下载病毒样本，主要包括HappyTime、ILoveYou、Melissa、Redlof等1000个典型的脚本病毒样本，基本涵盖了当前脚本病毒所使用的各类技术。

病毒基因库的基因源于对1 000个脚本病毒的感染模块或表现模块的提取。

对比实验选择3种典型的脚本病毒生产机：1) VWG(VBS Worms Generator)；2) DVVM(Dr. VBS Virus Maker)；3) WSHWC(Windows Scripting Host Worm Constructor)，以验证本模型的有效性。

2.2 实验过程

为验证本模型的正确性与有效性，本文分别设计了3种实验予以论证与解释，包括：1) 病毒进化实验；2) 模型对比实验；3)病毒存活实验。

病毒进化实验主要目的在于：分析已知病毒并提取其相关基因，构建初始病毒基因库，并应用本模型提出的病毒进化算子生成新的病毒群体，为后续的病毒存活实验提供进化病毒群体支持。该实验将测试两个内容：1) 病毒基因库规模与病毒存活率的关系；2) 病毒进化算子对病毒进化的影响。

模型对比实验主要目的在于：通过本模型与3种典型病毒生产机的病毒群体存活率对比，评估本模型的有效性。

病毒存活实验主要目的在于：利用反病毒软件，模拟自然选择对本模型新生成的病毒群体进行查杀，采用存活率来评估病毒进化效率以验证模型的正确性。

实验具体步骤如下：

1) 病毒基因库建立。在受控的虚拟环境中，对入选的病毒样本进行分析并分别提取病毒的感染模块和表现模块，经冗余筛选处理后加入病毒基因库，为后续的病毒群体进化生成提供基因支撑。

2) 生成病毒进化群体。这里分为两个并行阶段来生成病毒群体：1) 对初始病毒群体应用本模型提出的病毒进化算子生成新的病毒群体；2) 使用典型

的病毒生产机生成新的病毒群体，为后续的识别杀灭提供病毒群支持。

3) 病毒识别与杀灭。借助反病毒软件，模拟自然选择机制对新生成的两类病毒群体进行查杀，以最后的病毒存活率来评估本模型的有效性。

2.3　实验结果及分析

1) 病毒进化算子实验

本实验利用病毒进化算子生成新的病毒群体，通过反病毒软件查杀而得到存活率，旨在检验本模型病毒进化算子的有效性。提取1 000个病毒基因，分别对本模型中的3类病毒进化算子(基因算子、DNA算子、染色体算子)进行了测试，结果如图4所示。

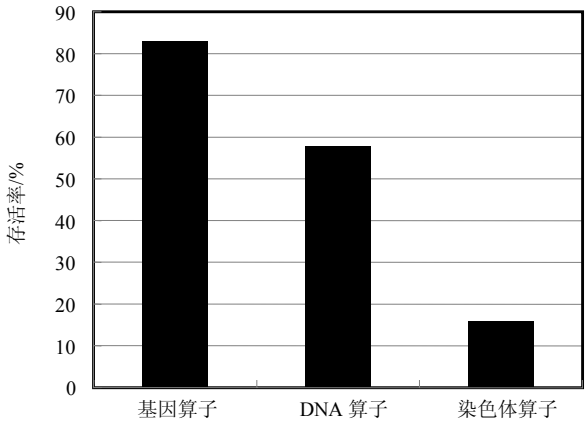

图4　病毒进化算子实验结果

实验结果表明，3类病毒进化算子对病毒进化(存活率)都有直接影响，按影响大小顺序排列依次为：基因算子影响最大、DNA算子影响次之、染色体算子影响最小。由模型理论可知，基因算子作用在病毒指令序列层面，而指令序列的改变将导致病毒特征码的改变，从而使基于特征码的反病毒软件难以识别与查杀，使病毒成功存活；DNA算子作用于病毒模块层面，模块中包含很多类似的指令序列，模块的改变可能改变了模块的名称、作用过程，其作用机理并未改变，因而使广谱反病毒软件容易识别而杀灭之；染色体算子作用于病毒过程(函数)层面，改变的只是病毒程序中的过程(函数)参数名称，其内部指令代码未变或少许改变，因此更易被反病毒软件查杀。

2) 病毒基因库实验

通过提取已知病毒的基因(特征码)，并经冗余处理后加入病毒基因库，为病毒进化提供基因支持。本实验旨在检验病毒基因库规模与病毒进化存活率的关系，结果如图5所示。

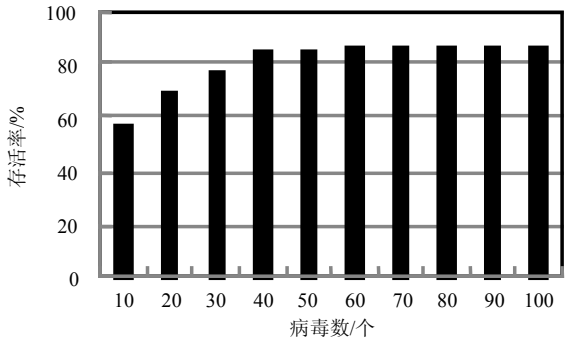

图5　病毒基因库与病毒进化存活率的关系

实验结果表明：1) 病毒进化存活率与病毒基因库规模成近似正比关系，即病毒基因库越大，基于此基因库的病毒进化群体的存活率越高。由模型理论知，随着病毒基因库的扩增，病毒进化所能使用的基因素材将显著增加，病毒特征码空间将明显扩大，导致反病毒软件更加难以识别，从而提高了病毒存活率；2) 在病毒基因库规模达到设定规模的半数后，病毒存活率变化不大。由模型理论及实验设置可知，由于病毒基因来自于同类病毒样本，当提取的病毒基因达到一定数量时将不可避免有近似基因存在，而拥有近似基因的病毒将为反病毒软件的广谱启发式所识别与杀灭。

3) 对比实验

为检验本模型的有效性，选择3种与本模型相似的典型脚本病毒生产机VWG、DVVM、WSHWC进行了对比实验，均分别产生病毒数为20个、40个、60个、80个、100个，通过反病毒软件进行查杀来验证其存活率。实验结果如图6所示。

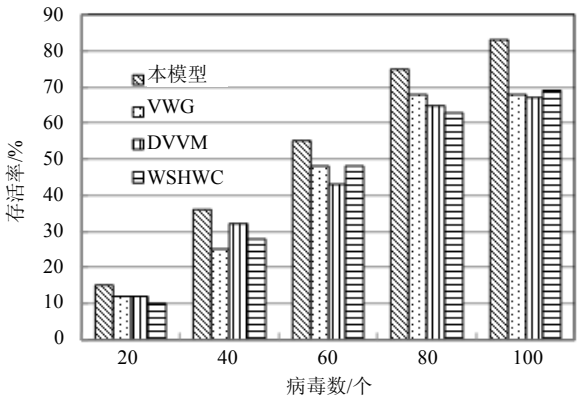

图6　本模型与其他病毒生产机的对比实验结果

实验结果表明，在病毒存活率方面，本模型优于其他3种病毒生成机。由模型理论知，由于常见的病毒生成机算法仅对有限的病毒模块进行变量替换或修改，而本模型在病毒基因、DNA和染色体3个层次对病毒进行了进化操作，既扩增了病毒特征码空

间，又使病毒在该空间中随机组合，从而提高了其存活率。

3 结 束 语

作为一种重要的网络空间安全威胁，计算机病毒的生存、适应、延续、发展等进化理论研究可促进网络武器技术、反病毒技术、漏洞修复技术、软件演化技术以及人工生命技术的发展。本文提出了一种基于基因理论的计算机病毒进化模型：从生命特性的角度，通过借鉴生物学病毒相关机理与概念定义了计算机病毒及其相关概念；从算法特性的视角，设计了计算机病毒的相关进化算子；从病毒基因的角度，模拟了计算机病毒的生存、繁衍、适应等进化过程。仿真实验表明，即使遭遇严酷的外部环境，具有算法和生命双重特性的计算机病毒仍具极强的进化能力。

参 考 文 献

[1] 国家计算机病毒应急处理中心, 第十七次计算机病毒和移动终端病毒疫情调查报告 [EB/OL].[2017-05-11] http://www.cverc.org.cn/ head/diaocha2017/report2017.pdf.
National Computer Virus Emergency Response Center. The 17th survey on computer virus and mobile terminal virus epidemic[EB/OL].[2017-05-11]. http://www.cverc.org.cn/ head/ diaocha2017/report2017.pdf.

[2] BALTHROP J, FORREST S, NEWMAN M E, et al. Technological networks and the spread of computer viruses[J]. Science, 2004, 304:527-529.

[3] SZOR P. The Art of computer virus research and defense[M]. Maryland: Symantec Press, 2005.

[4] 张瑜. 计算机病毒进化论[M]. 北京： 国防工业出版社, 2015.
ZHANG Yu. Theory of computer virus evolution[M]. Beijing: National Defense Industry Press, 2015.

[5] AGAPOW P M. Computer viruses: the inevitability of evolution[J]. Complex systems: from biology to computation, 1993: 46-54.

[6] NEUMANN J V. Theory and organization of complicated automata [C]// In Theory of Self-Reproducing Automata. Urbana: University of Illinois Press, 1966: 29-87.

[7] SPAFFORD E H. Computer viruses--a form of artificial life?[R]. Indiana： Purdue University, 1990.

[8] LUDWIG M A. Computer viruses, artificial Life and evolution[M]. Tucson, Arizona: American Eagle Publications, Inc. 1993.

[9] WILKE C O., WANG J L, OFRIA C, et al. Evolution of digital organisms at high mutation rates leads to survival of the flattest [J]. Nature, 2001, 412(6844): 331-333.

[10] PARSONS R J, FORREST S, BURKS C. Genetic algorithms, operators, and DNA fragment assembly[J]. Machine Learning, 1995, 21(1-2): 11-33.

[11] VALLEZ. Genetic programming in virus[EB/OL]. [2017-05-11]. https://download.adamas.ai/dlbase/Stuff/VX Heavens Library/vva00.html

[12] BLUEOWL. Implementing genetic algorithms in viruses[EB/OL].[2017-05-11] https://download.adamas.ai/ dlbase/Stuff/VX Heavens Library/vbo00.html

[13] 张瑜, 李涛, 吴丽华, 等. 计算机病毒演化模型及分析[J]. 电子科技大学学报, 2009, 38(3): 419- 422.
ZHANG Yu, LI Tao, WU Li-hua, et al. Computer virus evolution model and its analysis[J]. Journal of University of Electronic Science and Technology of China, 2009, 38(3): 419-422.

[14] NACHENBERG C. Computer Virus-antivirus Coevolution [J]. Communications of the ACM, 1997, 40(1): 46-51.

[15] ILIOPOULOS D, ADAMI C, SZOR P. Darwin inside the machines: Malware evolution and the consequences for computer security[C]// In Proceedings of Virus Bulletin Conference 2008，Ottawa：GROOTEN Martijn, 2008：187-194.

[16] MORGAN T H. The Theory of the Gene[J]. American Naturalist, 1917, 51(609): 513-544.

[17] 张瑜, 刘庆中, 宋丽萍, 等. 基于免疫和代码重定位的计算机病毒特征码提取与检测方法[J]. 北京理工大学学报, 2017, 37(10): 1036-1042.
ZHANG Yu, LIU Qing-zhong, SONG Li-ping, et al. Signature extraction and detection method of computer viruses based on immunity and code relocation[J]. Transactions of Beijing Institute of Technology, 2017, 37(10): 1036-1042.

编 辑 刘飞阳