BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

VŨ DUY SƠN

XÂY DỰNG HỆ THỐNG PHÁT HIỆN VIRUS TRÊN MÁY TÍNH

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công nghệ Thông tin

Mã ngành: 60480201

CÁN BỘ HƯỚNG DẪN KHOA HỌC: PGS. TS. VŨ THANH NGUYÊN

TP. HỒ CHÍ MINH, tháng 05 năm 2016

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP.HCM

Cán bộ hướng dẫn khoa học: PGS. TS. Vũ Thanh Nguyên

(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày 31 tháng 05 năm 2016

Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:

(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ Luận văn Thạc sĩ)

TT	Họ và tên	Chức danh Hội đồng
1	PGS.TS. Võ Đình Bảy	Chủ tịch
2	TS. Đặng Trường Sơn	Phản biện 1
3	TS. Cao Tùng Anh	Phản biện 2
4	TS. Lư Nhật Vinh	Ủy viên
5	TS. Nguyễn Thị Thúy Loan	Ủy viên, Thư ký

Xác nhận của Chủ tịch Hội đồng đánh giá Luận sau khi Luận văn đã được sửa chữa (nếu có).

Chủ tịch Hội đồng đánh giá LV

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM
PHÒNG QLKH – ĐTSĐH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

TP. HCM, ngày 19 tháng 01 năm 2015

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên:	Vũ Duy Sơn	Giới tính:	Nam
Ngày, tháng, năm sinh:	28/05/1990	Nơi sinh:	Hậu Giang
Chuyên ngành:	Công nghệ thông tin	MSHV:	1241860017

I- Tên đề tài:

XÂY DỰNG HỆ THỐNG PHÁT HIỆN VIRUS TRÊN MÁY TÍNH

II- Nhiệm vụ và nội dung:

Nghiên cứu xây dựng hệ thống phát hiện virus dựa trên hệ miễn dịch nhân tạo và các thuật toán . Thực hiện thuật toán và phát hiện nhận dạng virus một cách chính xác, có khả năng nhận dạng được biến thể của virus để từ đó ngăn chặn kịp thời và chủ động phòng tránh các tình huống lây nhiễm virus.

III- Ngày giao nhiệm vụ **:** 20/01/2016

IV- Ngày hoàn thành nhiệm vụ **:** 14/05/2016

V- Cán bộ hướng dẫn **: PGS. TS. VŨ THANH NGUYÊN**

CÁN BỘ HƯỚNG DẪN
(Họ tên và chữ ký)

KHOA QUẢN LÝ CHUYÊN NGÀNH
(Họ tên và chữ ký)

PGS. TS. VŨ THANH NGUYÊN

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Luận văn này đã được cảm ơn và các thông tin trích dẫn trong Luận văn đã được chỉ rõ nguồn gốc.

Học viên thực hiện Luận văn

(Ký và ghi rõ họ tên)

Vũ Duy Sơn

LỜI CẢM ƠN

Trước tiên tôi xin chân thành cảm ơn thầy giáo PGS.TS. Vũ Thanh Nguyên đã tận tình hướng dẫn, chỉ bảo tôi trong thời gian qua.

Tôi xin bày tỏ lòng biết ơn tới các thầy cô giáo trong khoa Công nghệ Thông tin nói riêng và trường Đại học Công nghệ Tp.HCM nói chung đã dạy bảo, cung cấp những kiến thức quý báu cho tôi trong suốt quá trình học tập và nghiên cứu tại trường.

Tôi cũng xin gửi lời cảm ơn tới gia đình, bạn bè, những người luôn cổ vũ, quan tâm và giúp đỡ tôi trong suốt thời gian học tập cũng như làm luận văn.

Do thời gian và kiến thức có hạn nên luận văn chắc không tránh khỏi những thiếu sót nhất định. Tôi rất mong nhận được những sự góp ý quý báu của thầy cô và các bạn.

Hồ Chí Minh, 01-2015

Vũ Duy Sơn

TÓM TẮT

Hiện nay, với sự phát triển nhanh chóng của CNTT, dẫn đến vấn đề an ninh máy tính là một vấn đề hết sức cần thiết. Trong đề tài này, tiến hành nghiên cứu một số dạng virus trên máy tính, tìm hiểu về một số khái niệm mạng miễn dịch sinh học, miễn dịch nhân tạo, và một số thuật toán trong hệ miễn dịch nhân tạo.

Tổng quan về hệ miễn dịch nhân tạo và một số thuật toán xử lý trong hệ miễn dịch và Nghiên cứu áp dụng một số thuật toán máy học vào hệ thống phát hiện Virus bằng cách lựa chọn các thuật toán phân lớp như thuật toán mạng RBF, thuật toán phân lớp SVM…, mô hình lai giữa mạng nơ-ron tiến hóa và thuật toán miễn dịch ứng dụng trong phát hiện virus bằng phương pháp sử dụng mạng nơ-ron nhân tạo kết hợp với thuật giải di truyền nhằm xây dựng một hệ thống phát hiện virus.

Hệ thống miễn dịch nhân tạo (AIS) là một chi nhánh của lĩnh vực tình báo tính toán lấy cảm hứng từ hệ thống miễn dịch sinh học, và đã đạt được nhiều sự quan tâm của các nhà nghiên cứu trong việc phát triển các mô hình và kỹ thuật miễn dịch dựa trên tính toán để giải quyết các vấn đề phức tạp hoặc kỹ thuật đa dạng.

Trọng tâm chính của luận văn này là xây dựng một hệ thống phát hiện virus dựa trên hệ thống miễn dịch nhân tạo bởi sự kết hợp của AIS và một số thuật toán phân lớp như KNN, SVM, và RBF, .. nhằm xử lý bài toán phát hiện virus.

ABSTRACT

Nowaday, the development of infomation of technology rapidly. Therefore, security issues are really necessary problems. The thesis research some kinds of virus on computer, learning some concept about natural immune system and artificial immune system and some althgorithm in AIS.

Overview of artificial immune system and a processing algorithm in the immune system and study and apply some machine learning algorithms into the virus detected system by selecting classification algorithms such as RBF network algorithm, SVM classification algorithm…, a hybrid system by using artificial neural network combined with the genetic algorithm to build a virus detection system.

Artificial Immune System (AIS) is a branch of computational intelligence field inspired by the biological immune system, and has gained increasing interest among researchers in the development of immune-based models and techniques to solve diverse complex computational or engineering problems.

The main focus of this research is devoted to building a virus detection system based on the artificial immune system by combination of AIS and some algorithms of classification such as KNN, SVM, and RBF,.. which aims to handle virus detection problem.

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

CNTT	Công Nghệ Thông Tin
KN	Kháng Nguyên
KT	Kháng Thể
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
RBF	Radial Basis Function
NSA	Negative Selection Algorithms
PSA	Positive Selection algorithms
CLONALG	Clonal Selection Algorithm
CSDL	Cơ Sở Dữ Liệu
MHC	Major Histocompatibility Complex
NK	Natural Killer
APC	Antigen Presenting Cell
TCR	T-cell receptor
AIS	Artificial Immune Systems
BIS	Biology Immune System
VDS	Virus Detector System

DANH MỤC BẢNG

DANH MỤC HÌNH

CHƯƠNG 1: TỔNG QUAN

1.1 Giới thiệu

Trong thực tế hiện nay bảo mật thông tin đang đóng một vai trò thiết yếu chứ không còn là "thứ yếu" trong mọi hoạt động liên quan đến việc ứng dụng công nghệ thông tin. Tôi muốn nói đến vai trò to lớn của việc ứng dụng CNTT đã và đang diễn ra sôi động, không chỉ thuần túy là những công cụ (Hardware, software), mà thực sự đã được xem như là giải pháp cho nhiều vấn đề. Khởi động từ những năm đầu thập niên 90, với một số ít chuyên gia về CNTT, những hiểu biết còn hạn chế và đưa CNTT ứng dụng trong các hoạt động sản xuất, giao dịch, quản lý còn khá khiêm tốn và chỉ dừng lại ở mức công cụ, và đôi khi tôi còn nhận thấy những công cụ "đắt tiền" này còn gây một số cản trở, không đem lại những hiệu quả thiết thực cho những Tổ chức sử dụng nó.

Internet cho phép chúng ta truy cập tới mọi nơi trên thế giới thông qua một số dịch vụ. Ngồi trước máy tính của mình bạn có thể biết được thông tin trên toàn cầu, nhưng cũng chính vì thế mà hệ thống máy tính của bạn có thể bị xâm nhập vào bất kỳ lúc nào mà bạn không hề được biết trước, kéo theo việc bảo mật và an toàn thông tin của bạn bị đe dọa..

Trong bối cảnh đó, đề tài " Xây dựng hệ thống phát hiện virus trên máy tính " được tiến hành nhằm góp phần giải quyết vấn đề bị virus xâm nhập cũng như việc bảo mật thông tin và an toàn máy tính cho người sử dụng.

1.2 Mục tiêu nghiên cứu

Nghiên cứu xây dựng hệ thống phát hiện virus trên máy tính, tìm hiểu về hệ miễn dịch nhân tạo và các thuật toán, nghiên cứu các khả năng bị xâm phạm an toàn thông tin và phương thức xâm nhập máy tính dựa trên các tiêu chí:

Nhận dạng virus nhanh và phát hiện một cách chính xác các trường hợp lây nhiễm virus.

Có khả năng dự báo được biến thể của virus để từ đó ngăn chặn kịp thời và chủ động phòng tránh các tình huống lây nhiễm virus..

1.3 Đối tượng nghiên cứu

Đối tượng nghiên cứu là các virus về máy tính, về hệ miễn dịch sinh học, hệ miễn dịch nhân tạo và các thuật toán trong hệ miễn dịch nhân tạo để từ đó xây dựng hệ thống có khả năng nhận dạng được virus và chủ động phòng tránh các trường hợp lây nhiễm virus.

1.4 Phạm vi nghiên cứu

Luận văn sẽ tìm hiểu về hệ miễn dịch nhân tạo, các thuật toán trong hệ miễn dịch nhân tạo. Từ đó tạo nền tảng để xây dựng hệ thống có thể nhận dạng và phát hiện virus một cách chính xác. Thực hiện thuật toán và xác định tính hiệu quả của phương pháp này bằng việc xây dựng bộ dữ liệu huấn luyện và kiểm thử. Kết quả thực nghiệm cho thấy, tỉ lệ phát hiện là khá tốt, với hướng tiếp cận đưa ra sẽ là nền tảng khá tốt cho việc nghiên cứu và các hướng phát triển trong tương lai.

1.5 Bố cục luận văn

Luận văn có cấu trúc như sau:

Chương 1: tổng quan về luận văn gồm các mục: giới thiệu, mục tiêu nghiên cứu, đối tượng và phạm vi nghiên cứu.

Chương 2: khái quát về virus máy tính và các nghiên cứu liên quan trong và ngoài nước.

Chương 3: khái niệm về hệ miễn dịch sinh học và các thành phần chức năng của hệ miễn dịch sinh học.

Chương 4: hệ miễn nhân tạo và các thuật toán đề từ đó đưa ra định hướng, tiến hành thực nghiệm và đánh giá kết quả.

Chương 5: kết luận.

CHƯƠNG 2: KHÁI QUÁT VỀ VIRUS MÁY TÍNH VÀ CÁC NGHIÊN CỨU LIÊN QUAN

2.1 Virus máy tính

Virus máy tính là một chương trình máy tính, nó có thể tự lây lan bằng cách gắn vào các chương trình khác và tự sao chép chính nó để lây nhiễm các máy khác trong cùng hệ thống. Khi virus phát tác, chúng gây ra nhiều hậu quả : từ thông báo bậy bạ cho đến những tác động làm lệch lạc khả năng thực hiện của phần mềm hệ thống, hoặc xóa sạch mọi thông tin trên đĩa cứng.

Khi nghiên cứu virus máy tính, có 3 vấn đề cần cân nhắc là :

- Môi trường: hệ điều hành, kiến trúc máy tính.
- Phương tiện: nơi chứa tin, cơ chế lây lan.
- Cơ hội: cộng đồng sử dụng, tần suất kích hoạt,…

Hiện nay thì do tính phổ biến của hệ điều hành Windows nên virus máy tính trên hệ điều hành này cũng nhiều hơn. Và để đáp ứng nhu cầu thực tiễn cấp bách, đề tài tập trung nghiên cứu các loại virus máy tính hoạt động trên các hệ điều hành Windows dành cho máy tính IBM-PC (máy vi tính xách tay hoặc máy tính để bàn).

Mặc dù vậy, đề tài cũng được định hướng nghiên cứu để có thể mở rộng kết quả nghiên cứu cho các hệ anti-virus sử dụng các hệ điều hành khác Windows.

Worm cũng là một chương trình có khả năng tự nhân bản và tự lây nhiễm trong hệ thống tuy nhiên nó có khả năng “tự đóng gói”, điều đó có nghĩa là worm không cần phải có “file chủ” để mang nó khi nhiễm vào hệ thống. Như vậy, có thể thấy rằng chỉ dùng các chương trình quét file sẽ không diệt được worm trong hệ thống vì worm không “bám” vào một file hoặc một vùng nào đó trên đĩa cứng. Mục tiêu của worm bao gồm cả làm lãng phí nguồn lực băng thông của mạng và phá hoại hệ thống như xoá file, tạo backdoor, thả keylogger,... Tấn công của worm có đặc trưng là lan rộng cực kỳ nhanh chóng do không cần tác động của con người (như khởi động máy, copy file hay đóng/mở file). Worm có thể chia làm 2 loại:

- Network Service Worm: lan truyền bằng cách lợi dụng các lỗ hổng bảo mật của mạng, của hệ điều hành hoặc của ứng dụng. Sasser là ví dụ cho loại sâu này.
- Mass Mailing Worm: là một dạng tấn công qua dịch vụ mail, tuy nhiên nó tự đóng gói để tấn công và lây nhiễm chứ không bám vào vật chủ là email. Khi sâu này lây nhiễm vào hệ thống, nó thường cố gắng tìm kiếm sổ địa chỉ và tự gửi bản thân nó đến các địa chỉ thu nhặt được. Việc gửi đồng thời cho toàn bộ các địa chỉ thường gây quá tải cho mạng hoặc cho máy chủ mail. Netsky, Mydoom là ví dụ cho thể loại này.

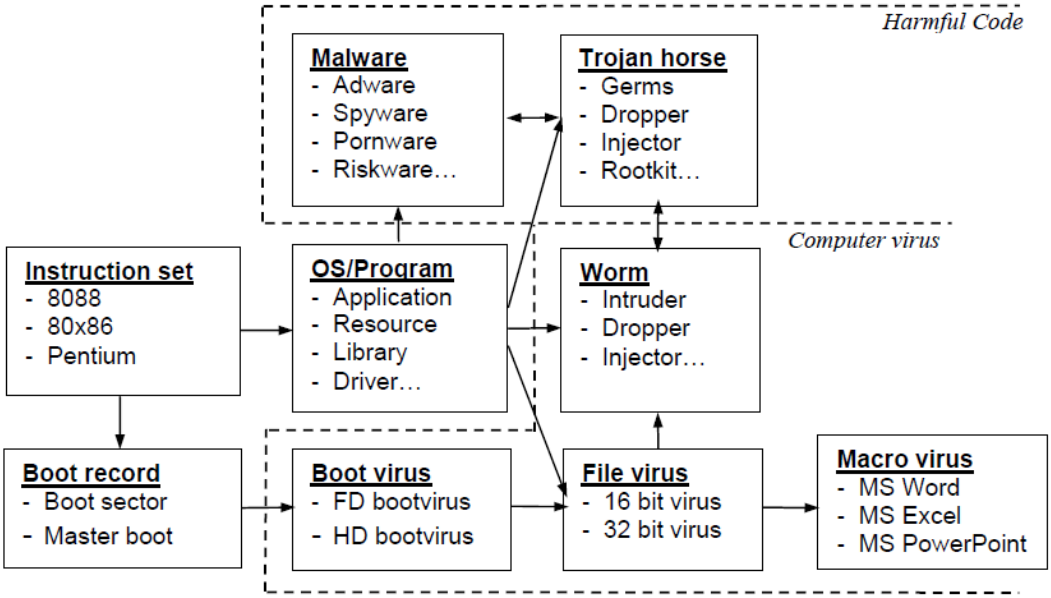
Trojan Horse: là loại mã độc hại được đặt theo sự tích "Ngựa thành Troa". Trojan horse không tự nhân bản tuy nhiên nó lây vào hệ thống với biểu hiện rất ôn hoà nhưng thực chất bên trong có ẩn chữa các đoạn mã với mục đích gây hại. Trojan có thể lựa chọn một trong 3 phương thức để gây hại:

- Tiếp tục thực thi các chức năng của chương trình mà nó bám vào, bên cạnh đó thực thi các hoạt động gây hại một cách riêng biệt (ví dụ như gửi một trò chơi dụ cho người dùng sử dụng, bên cạnh đó là một chương trình đánh cắp password)
- Tiếp tục thực thi các chức năng của chương trình mà nó bám vào, nhưng sửa đổi một số chức năng để gây tổn hại (ví dụ như một trojan giả lập một cửa sổ login để lấy password) hoặc che dấu các hành động phá hoại khac (ví dụ như trojan che dấu cho các tiến trình độc hại khác bằng cách tắt các hiển thị của hệ thống).
- Thực thi luôn một chương trình gây hại bằng cách núp dưới danh một chương trình không có hại (ví dụ như một trojan được giới thiệu như là một chò chơi hoặc một tool trên mạng, người dùng chỉ cần kích hoạt file này là lập tức dữ liệu trên PC sẽ bị xoá hết).

Malware là tên gọi chung các loại phần mềm độc hại như:

- Adware: loại phần mềm tự động bật (popup) các cửa sổ quảng cáo, thay đổi các thiết lập hệ thống…, gây phiền phức cho người sử dụng.

- Spyware: loại phần mềm gián điệp, săn lùng thông tin thương mại, đánh cắp thông tin cá nhân như địa chỉ e-mail, độ tuổi, giới tính, thói quen mua sắm…
- Pornware: loại phần mềm đồi trụy phát tán từ các trang web khiêu dâm, tự động bật lên các hình ảnh gợi dục, phim sex... Pornware rất nguy hiểm cho trẻ em và thanh thiếu niên, làm suy đồi đạo đức xã hội.
- Riskware: phần mềm trôi nổi, không được kiểm định chất lượng. Riskware tiềm ẩn nhiều lỗi nghiêm trọng, làm suy giảm chất lượng phục vụ của hệ thống, có nguy cơ ảnh hưởng dữ liệu của người dùng.



Hình 2.1 Nguồn gốc mã độc và phân loại mã độc

Trapdoor (hay backdoor) rất được giới hacker ưa chuộng. Khi lây vào hệ thống, backdoor có nhiệm vụ mở cổng (port, điểm truy cập ứng dụng), làm nội gián chờ đáp ứng thao tác quét cổng của hacker. Khi nhận được tín hiệu, backdoor khống chế hàng rào bảo vệ hệ thống, dọn đường đón các đợt thâm nhập từ bên ngoài.

Gần đây xuất hiện loại trojan đặc biệt nguy hiểm gọi là rootkit. Ban đầu, rootkit là tên gọi các bộ công cụ (kit) giúp người quản trị nắm quyền điều khiển hệ thống ở mức cao nhất (root). Trong tay hacker, rootkit trở thành công cụ đắc lực để đánh cắp mật khẩu truy nhập, thu thập thông tin trên máy nạn nhân hoặc che đậy

các hoạt động thâm nhập bất hợp pháp. Hacktool là một dạng rootkit sơ cấp. Cao cấp hơn có các loại rootkit thám báo như keylogger (theo dõi hoạt động bàn phím), sniffer (theo dõi gói tin qua mạng), filehooker (theo dõi truy nhập tập tin)…

2.2 Các nghiên cứu liên quan

2.2.1 Tình hình nghiên cứu trong nước

Một số công trình nghiên trong nước được công bố :

Năm 1997, luận văn Cao học về nhận dạng virus tự động đầu tiên của Việt Nam được bảo vệ ở Viện Tin học Pháp ngữ. Mục tiêu của đề tài nhằm xây dựng một hệ suy diễn nhận dạng virus máy tính thông qua các hành vi cơ sở [6]. Sử dụng thuật giải tìm kiếm hành vi với tri thức bổ sung tại mỗi nút trên cây, đề tài cho kết quả chẩn đoán boot virus khá tốt. Để tăng cường độ tin cậy và an toàn hệ thống, tác giả đề xuất một không gian chẩn đoán đặc biệt gọi là máy ảo chẩn đoán.

Năm 1998, Trần Quốc Việt (Khoa CNTT, ĐH Cần Thơ) tiếp tục hướng nghiên cứu máy ảo mở rộng bài toán cho file virus. Đề tài rút ra kết luận: máy ảo chỉ thích hợp để chẩn đoán boot virus. Khi chẩn đoán file, máy ảo cần bộ xử lý lệnh tương thích với tập lệnh của HĐH nên phức tạp, cồng kềnh và kém hiệu quả [4].

Năm 2005, Hồ Ngọc Thơ (Khoa CNTT, ĐH Cần Thơ) thực hiện đề tài nhận dạng biến thể virus hướng text mining. Tác giả đưa ra giả thuyết virus máy tính di truyền mã lệnh của nó cho các thế hệ con cháu, vì vậy có thể nhận dạng các biến thể virus thông qua tập gien (chuỗi mã lệnh phổ biến) của các thành viên trong họ. Giải pháp của đề tài là phân tích tập virus mẫu hướng text mining để xây dựng cây phả hệ, sau đó áp dụng thuật giải nhận dạng tập gien xuất hiện trên cây. Mặc dù phần demo còn hạn chế (cài đặt phức tạp, chạy chậm, tiêu tốn nhiều tài nguyên…), tuy nhiên tác giả đã phần nào chứng minh được giả thuyết của đề tài [3].

2.2.2 Tình hình nghiên cứu thế giới

Một số công trình nghiên cứu trên thế giới được công bố :

Databases That Learn: dự án của Symantec Research Labs ở Santa Monica (USA) luyện học thói quen truy nhập vào các cơ sở dữ liệu để rút quy luật tấn công của tin tặc, bảo vệ hệ thống tránh bị xâm nhập [7].

MLX Proofpoint Zero-Hour Anti-virus: sản phẩm của ProofPoint Inc., bảo vệ hệ thống trong thời gian thực. Bằng các kỹ thuật máy học, ProofPoint phân tích các email có tập tin đính kèm và phát cảnh báo khi có file nghi ngờ mã độc [10]. Công cụ này được cài đặt cho hệ F-Secure Message Security Gateway [8] để lọc thư rác.

DDI (Distributed Detection and Inference): dự án của Intel triển khai ở đại học Berkeley (2005) suy luận trên hệ thống mạng nhằm phát hiện các cuộc tấn công lan trản [9]. Nghiên cứu này đặt ra giả thuyết nếu một nút mạng bị tấn công thì các nút mạng còn lại trong hệ thống cũng có thể bị tân công. Giải pháp của đề tài là cài đặt một thuật toán học cho mỗi nút mạng để phát hiện các cuộc truy nhập cục bộ. Mỗi nút có liên lạc với các nút kế cận để suy luận và cảnh báo các tình huống hệ thống bị tấn công lan tràn [11].

Malicious Software Detection for Resource Constrained Devices: ý tưởng cơ bản của dự án là sản sinh tập mã độc chưa biết từ số ít dấu hiệu nhận dạng các loại mã độc đã biết [12].

CHƯƠNG 3: HỆ MIỄN DỊCH SINH HỌC

3.1 Khái niệm về hệ miễn dịch sinh học

Hiện nay, đấu tranh sinh tồn là một trong các quy luật tự nhiên, cho nên mọi sinh vật đều có ít nhiều khả năng tự bảo vệ để chống lại sự xâm nhập đối với các yếu tố gây hại đến chúng. Cùng với sự tiến hóa của sinh vật, các biện pháp tự bảo vệ ngày càng phong phú và hoàn thiện hơn, trong đó đáp ứng miễn dịch là một biện pháp quan trọng và phức tạp nhất.

Theo những quan niệm ban đầu "miễn dịch là trong khi cơ thể này không mắc bệnh truyền nhiễm còn những cơ thể khác lại mắc bệnh truyền nhiễm tuy ở trong cùng điều kiện".

Hiện nay miễn dịch được định nghĩa: "Miễn dịch là khả năng phòng vệ của toàn bộ cơ thể đối với các yếu tố mang thông lạ".

Theo (R.V.Petrov, 1978). "Miễn dịch là khả năng đề kháng của sinh vật chống lại một sinh vật khác và các chất mang trên bản thân chúng những dấu hiệu thông tin di truyền ngoại lai. Tính sinh miễn dịch được hình thành trong quá trình tiến hóa của sinh vật".

Miễn dịch học là một chuyên ngành rộng trong y sinh học, nghiên cứu mọi phương diện của hệ miễn dịch của tất cả các sinh vật. Đối tượng nghiên cứu của miễn dịch học bao gồm: hoạt động sinh lý của hệ miễn dịch ở cơ thể khỏe mạnh và cả khi bệnh (các đặc điểm lý, hóa của các thành phần thuộc hệ miễn dịch); các rối loạn của hệ miễn dịch (các bệnh tự miễn, các phản ứng quá mẫn, sự suy giảm miễn dịch); và hiện tượng thải ghép. Miễn dịch học được ứng dụng trong nhiều ngành khoa học khác, bản thân nó cũng phân thành các ngành chuyên sâu hơn.

3.2 Các thành phần của hệ miễn dịch sinh học

Các thành phần chính của hệ miễn dịch là tế bào lymphô, tế bào trình diện kháng nguyên, và tế bào hiệu quả.

- *Tế bào lymphô*: là những tế bào có khả năng nhận diện một cách đặc hiệu kháng nguyên lạ và tạo phản ứng chống lại chúng. Do vậy, lymphô bào là tế bào trung gian của cả miễn dịch dịch thể và miễn dịch tế bào. Có nhiều

tiểu quần thể tế bào lymphô khác nhau về cả cách nhận diện kháng nguyên lẫn chức năng của chúng. Tế bào lymphô B là tế bào duy nhất có thể sản xuất kháng thể. Chúng nhận diện kháng nguyên ngoại bào (kể cả kháng nguyên trên bề mặt tế bào) và biệt hoá thành tế bào tiết kháng thể, do đó chúng tác dụng như tế bào trung gian của miễn dịch dịch thể. Tế bào lymphô T nhận diện kháng nguyên của vi sinh vật nội bào và có chức năng tiêu diệt những vi sinh vật này hoặc những tế bào bị nhiễm trùng. Thụ thể kháng nguyên của chúng là những phân tử màng khác với kháng thể nhưng có cấu trúc liên quan. Tế bào T có tính đặc hiệu rất chặt chẽ đối với kháng nguyên. Chúng chỉ nhận diện những phân tử pepxid gắn với một protein bản thân được mã hoá bởi những gen trong phức hệ hòa hợp mô chủ yếu (MHC) và được thể hiện trên bề mặt của những tế bào khác. Như vậy, tế bào T nhận diện và phản ứng với kháng nguyên gắn trên bề mặt tế bào chứ không phải kháng nguyên hoà tan. Tế bào T có nhiều nhóm mang chức năng khác nhau. Được biết nhiều nhất là tế bào T giúp đỡ, T gây độc. Khi đáp ứng với kháng nguyên, tế bào T giúp đỡ tiết ra những protein gọi là cytokin có chức năng kích thích sự tăng sinh và biệt hoá của tế bào T và một số tế bào trong đó có tế bào B, đại thực bào và các bạch cầu khác. Tế bào T gây độc giết các tế bào sản xuất ra kháng nguyên lạ như các tế bào bị nhiễm virus hay những vi khuẩn nội bào khác. Một số tế bào T được gọi là T điều hoà có chức năng ức chế đáp ứng miễn dịch. Bản chất và vai trò sinh lý của tế bào T điều hoà chưa được biết đầy đủ. Có một nhóm tế bào lymphô thứ ba là tế bào giết (NK), đây là những tế bào tham gia vào hệ thống miễn dịch bẩm sinh chống lại nhiễm trùng virus và các vi sinh vật nội bào khác.

- Sự khởi động và phát triển đáp ứng miễn dịch thu được bao giờ cũng đòi hỏi kháng nguyên phải được bắt giữ và trình diện cho tế bào lymphô. Tế bào chịu trách nhiệm làm việc này được gọi là *tế bào trình diện kháng nguyên* (APC). Tế bào trình diện kháng nguyên

được chuyên môn hoá cao nhất là tế bào hình sao (dendritic), chúng bắt giữ những vi sinh vật từ bên ngoài xâm nhập vào, vận chuyển những kháng nguyên này đến các cơ quan lymphô và trình diện kháng nguyên cho những tế bào T để khởi động đáp ứng miễn dịch.

- Sự hoạt hoá tế bào lymphô bởi kháng nguyên dẫn đến sự hình thành nhiều cơ chế loại bỏ kháng nguyên. Sự loại bỏ kháng nguyên đòi hỏi sự tham gia của những tế bào gọi là *tế bào hiệu quả*. Tế bào lymphô hoạt hoá, thực bào đơn nhân, và một số bạch cầu khác có thể làm chức năng tế bào hiệu quả trong những đáp ứng miễn dịch khác nhau.

Hệ thống miễn dịch có thể chia làm 2 loại là miễn dịch tự nhiên và miễn dịch thích nghi. Cả 2 loại miễn dịch này có liên quan chặt chẽ với nhau.

3.2.1 Miễn dịch tự nhiên

Miễn dịch tự nhiên (không đặc hiệu) là khả năng bảo vệ cơ thể sẵn có và mang tính di truyền trong các cá thể cùng loài. Nói cách khác, đó là khả năng tự bảo vệ của một cá thể có từ lúc mới sinh, không đòi hỏi phải có sự tiếp xúc trước của cơ thể với các kháng nguyên của sinh vật lạ từ môi trường bên ngoài.

Hệ thống miễn dịch tự nhiên bao gồm các biểu mô tạo nên lớp rào chắn chống lại sự xâm nhập của vi sinh vật, các tế bào trong hệ tuần hoàn và trong các mô, và một số protein huyết tương. Các thành phần này có những vai trò khác nhau nhưng bổ trợ cho nhau để ngăn chặn không cho vi sinh vật xâm nhập vào các mô của cơ thể, và một khi vi sinh vật đã vào mô rồi thì loại bỏ chúng.

Tính đặc hiệu của miễn dịch tự nhiên có một số điểm khác biệt so với tính đặc hiệu của các tế bào lympho là thành phần mấu chốt đóng vai trò nhận diện kháng nguyên và tạo nên tính đặc hiệu của đáp ứng miễn dịch thích ứng.

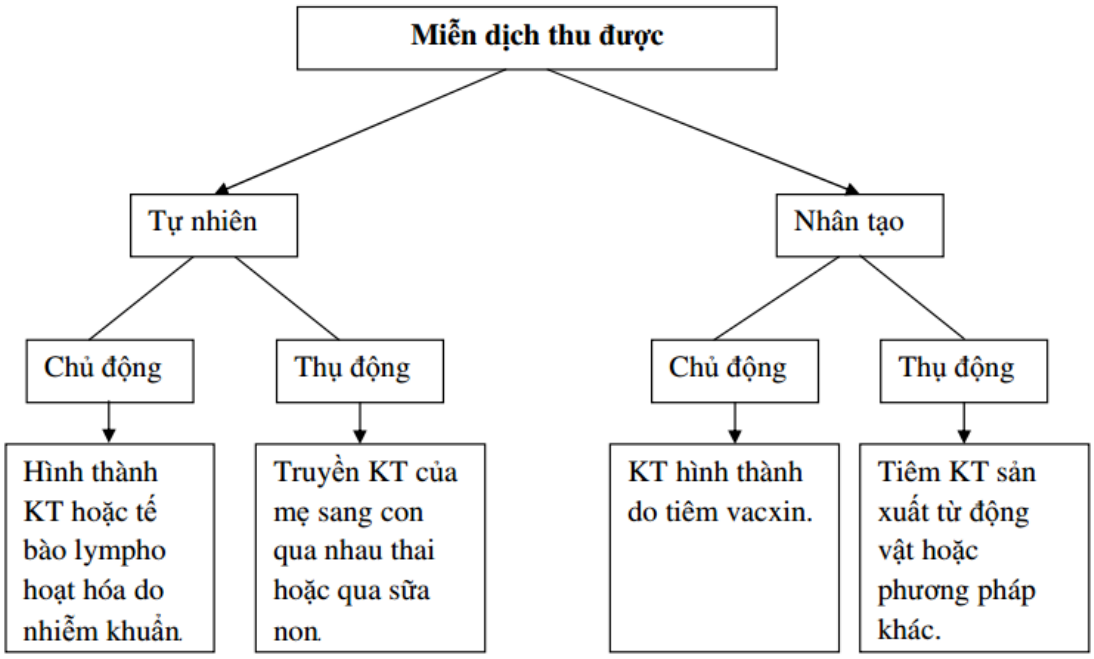
3.2.2 Miễn dịch thích nghi

Miễn dịch thích nghi (miễn dịch đặc hiệu) là trạng thái miễn dịch khi cơ thể đáp ứng lại một cách đặc hiệu với kháng nguyên. Đáp ứng miễn dịch là kết quả của sự hợp tác rất chặt chẽ, phức tạp và hài hòa giữa các tế bào và các phân tử của hệ

thống miễn dịch. Hệ thống miễn dịch đặc hiệu ở động vật có xương sống có 3 chức năng chính:

- Nhận diện bất kỳ "kẻ" nào được coi là lạ đối với cơ thể.
- Đáp ứng lại (đánh trả) "kẻ xâm nhập" ấy.
- Ghi nhớ "kẻ xâm phạm".

Sự nhận diện mang tính đặc hiệu cao. Hệ thống miễn dịch của cơ thể phân biệt được các vật lạ, các tế bào ung thư, tế bào nhiễm virut, nhận diện được các protein và tế bào của mình (self), các protein và tế bào không phải của mình (nonself).



Hình 3.1 Các loại miễn dịch thu được

Sau khi nhận diện là giai đoạn đáp ứng. Hệ thống miễn dịch tuyển mộ các tế bào và các phân tử phù hợp để tấn công "kẻ xâm phạm". Hiện tượng này gọi là đáp ứng hiệu ứng (effector respond). Đáp ứng này nhằm loại trừ vật lạ hoặc biến chúng thành vô hại đối với vật chủ, do đó ngăn chặn được bệnh.

Nếu tác nhân gây bệnh xâm phạm lần sau thì hệ thống miễn dịch sẽ nhớ để đáp ứng lại một cách nhanh và mạnh hơn. Miễn dịch đặc hiệu được chia làm hai

loại là miễn dịch thể dịch (còn gọi là miễn dịch qua trung gian kháng thể) và miễn dịch tế bào (hay miễn dịch qua trung gian tế bào)

Miễn dịch thể dịch dựa trên sự hoạt động của kháng thể (protein hòa tan trong thể dịch của cơ thể và có trên màng tế bào B). Kháng thể lưu động gắn đặc hiệu với vi sinh vật, độc tố do chúng sinh ra và virut ngoại bào để trung hòa hoặc làm tan chúng theo một cơ chế riêng.

Miễn dịch tế bào dựa trên sự hoạt động của các loại tế bào T đặc hiệu tấn công trực tiếp tế bào nhiễm virut, tế bào ung thư, các tế bào của mô ghép…Tế bào T có thể làm tan các tế bào này hoặc tiết ra các chất hóa học gọi là cytokin để tăng cường đáp ứng miễn dịch. Miễn dịch thu được chia ra thành miễn dịch thu được tự nhiên và miễn dịch thu được nhân tạo, có thể là chủ động hay thụ động.

Miễn dịch thu được tự nhiên chủ động được hình thành khi có sự xâm nhập của kháng nguyên, ví dụ khi bị nhiễm khuẩn. Hệ thống miễn dịch đáp lại bằng cách sản ra kháng thể và hoạt hóa các tế bào lympho để làm bất hoạt hoặc phá hủy kháng nguyên. Miễn dịch có thể tồn tại suốt đời (ví dụ đậu mùa) hoặc chỉ vài năm (ví dụ uốn ván).

Miễn dịch thu được tự nhiên thụ động được hình thành khi truyền kháng thể từ cá thể này cho cá thể khác. Ví dụ kháng thể miễn dịch của mẹ truyền sang thai nhi. Một số kháng thể của mẹ cũng có thể truyền cho con ở thời kỳ đầu qua dòng sữa non. Điều này rất cần thiết vì hệ thống miễn dịch của bé chưa hoàn thiện để có thể tự lập. Tiếc rằng loại miễn dịch này chỉ tồn tại trong một thời gian ngắn, trong vài tuần hoặc vài tháng.

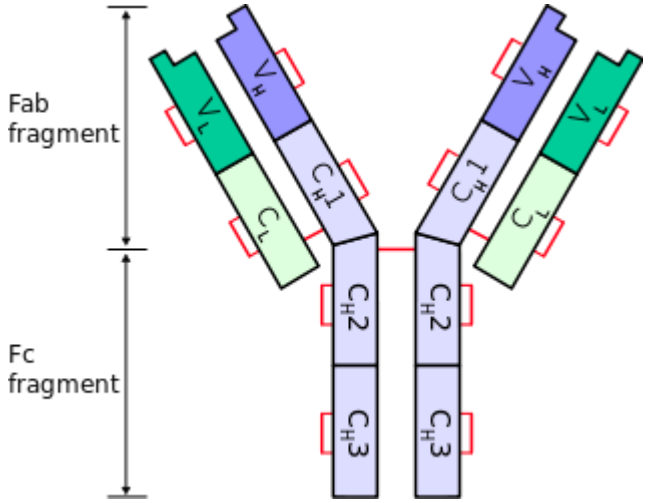
Miễn dịch thu được nhân tạo chủ động được hình thành khi đưa vacxin vào cơ thể để tạo đáp ứng miễn dịch hoặc đưa các tế bào lympho đã hoạt hóa.

Miễn dịch thu được nhân tạo thụ động được hình thành khi tiêm vào cơ thể kháng huyết thanh (huyết thanh chứa kháng thể). Đây là hành động chi viện có tác dụng ngay, nhưng thời gian sống của kháng thể rất ngắn, chỉ mấy tuần hoặc mấy tháng.

3.3 Kháng Thể

Kháng thể (antibody): là các gama globulin (Ig) có trong huyết thanh của động vật có khả năng liên kết đặc hiệu với kháng nguyên đã kích thích sinh ra nó.

Cấu trúc của kháng thể: phân tử kháng thể cấu tạo từ 4 chuỗi polypeptide, gồm hai chuỗi nặng (H, heavy, tiếng Anh, màu tím trong hình 3) giống hệt nhau và hai chuỗi nhẹ (L, light, tiếng Anh, màu xanh lá trong hình 3) cũng giống hệt nhau. Có hai loại chuỗi nhẹ κ (kappa) và λ (lambda), do đó hai chuỗi nhẹ của mỗi phân tử immunoglobulin chỉ có thể cùng là κ hoặc cùng là λ. Các chuỗi của immunoglobulin liên kết với nhau bởi các cầu nối disulfide và có độ đàn hồi nhất định. Một phần cấu trúc của các chuỗi thì cố định nhưng phần đầu của hai "cánh tay" chữ Y thì rất biến thiên giữa các kháng thể khác nhau, để tạo nên các vị trí kết hợp có khả năng phản ứng đặc hiệu với các kháng nguyên tương ứng, điều này tương tự như một enzyme tiếp xúc với cơ chất của nó. Có thể tạm so sánh sự đặc hiệu của phản ứng kháng thể-kháng nguyên với ổ khóa và chìa khóa.



Hình 3.2 Sơ đồ các chuỗi của một kháng thể

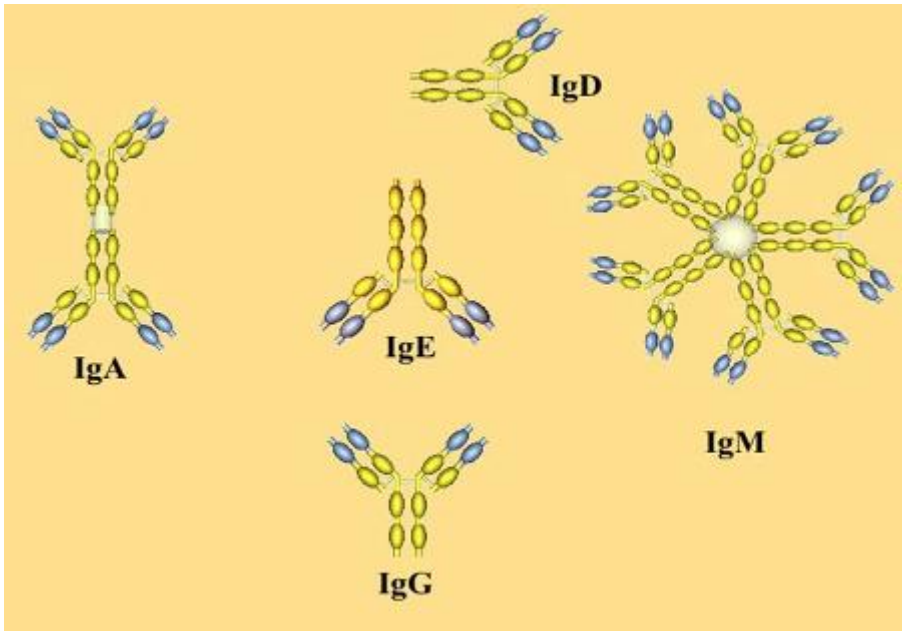
Các lớp kháng thể: Có 5 lớp kháng thể mang tên chuỗi nặng là IgG, IgM, IgA, IgD và IgE.

- IgG chiếm 80% tổng Ig trong huyết thanh người bình thường. Cấu tạo gồm chuỗi nhẹ Kappa hoặc lamda và hai chuỗi nặng gama. Ở người Việt nồng độ IgG trong máu là 1400 mg/100ml. IgG là kháng thể duy

nhất được truyền từ mẹ sang thai nhi qua nhau thai. IgG giữ vai trò chính bảo vệ cơ thể chống tác nhân gây bệnh, đảm nhiệm các chức năng opsonin hóa (giúp đại thực bào bắt giữ kháng nguyên), hoạt hóa bổ thể, gây độc qua trung gian tế bào phụ thuộc kháng thể (giúp tế bào K diệt tế bào đích), trung hòa ngoại độc tố (độc tố uốn ván, độc tố do Clostridium botulinum gây ngộ độc thức ăn, nọc rắn, nọc côn trùng…), gây ngưng kết vi khuẩn và trung hòa virus.

- IgM chiếm 5-10% trong huyết thanh ở người bình thường. IgM có cấu tạo gồm hai chuỗi nhẹ kappa hoặc lamda và hai chuỗi muy. Năm globulin chụm lại với nhau thành phân tử lớn hình sao năm cánh nhờ cầu nối disulfua và chuỗi protein J, do vậy có tới 10 paratop. IgM xuất hiện sớm, thực hiện các chức năng như hoạt hóa bổ thể, ngưng kết hồng cầu cùng loài trong trường hợp nhóm máu ABO, ngưng kết vi khuẩn.
- IgA có cấu tạo gồm 2 chuỗi nhẹ kappa hoặc lamda với hai chuỗi nặng alpha. Có hai loại: IgA trong huyết thanh chủ yếu ở dạng mônome và IgA tiết (sIgA) luôn có dạng dime, có trong dịch tiết của cơ thể như sữa, nước bọt, nước mắt, trong dịch nhầy đường tiêu hóa, sinh dục, hô hấp. IgA monome làm nhiệm vụ hoạt hóa bổ thể theo con đường nhánh. IgA tiết chống vi khuẩn trên bề mặt niêm mạc gây nhiễm trùng đường hô hấp, tiêu hóa đồng thời chống kháng nguyên nhóm máu ABO.
- IgD có nồng độ trong máu rất thấp và cấu tạo gồm hai chuỗi nhẹ kappa hoặc lamda và hai chuỗi nặng delta. Hoạt tính sinh học của IgD còn chưa rõ nhưng nó có mặt trên tế bào B làm thụ thể cho kháng nguyên.
- IgE có nồng độ trong huyết thanh rất thấp và có cấu tạo gồm hai chuỗi nhẹ kappa hoặc lamda và hai chuỗi nặng epsilon. IgE tham gia vào quá mẫn tức thì (hay dị ứng) bằng cách gắn phần Fc với tế bào mast

(dưỡng bào) và phần Fab với kháng nguyên. Tế bào mast được hoạt hóa sẽ giải phóng các chất hoạt mạch như histamin, leukotrien, serotonin gây giãn mạch và tăng tính thấm thành mạch.



Hình 3.3 Các lớp kháng thể

Vai trò của kháng thể: Trong một đáp ứng miễn dịch, kháng thể có 3 chức năng chính: gắn với kháng nguyên, kích hoạt hệ thống bổ thể và huy động các tế bào miễn dịch.

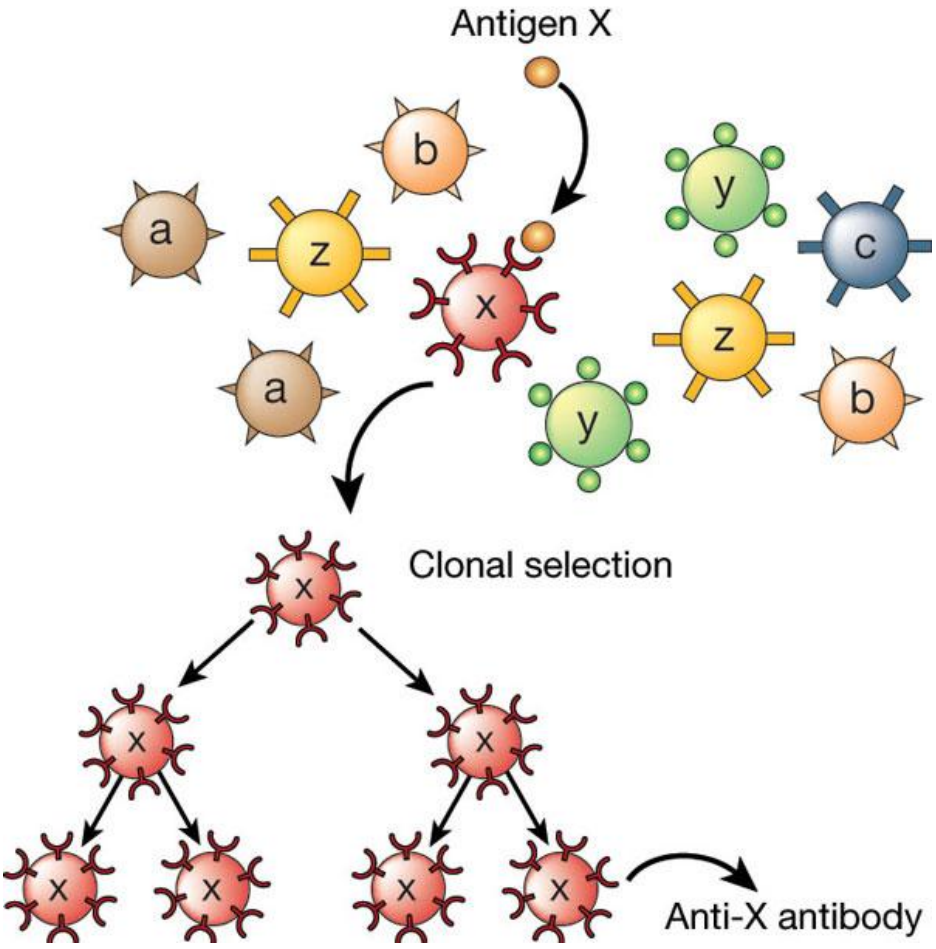
- Liên kết với kháng nguyên: Các immunoglobulin có khả năng nhận diện và gắn một cách đặc hiệu với 1 kháng nguyên tương ứng. Một thí dụ để miêu tả lợi ích của kháng thể là trong phản ứng chống độc tố vi khuẩn. Kháng thể gắn với và qua đó trung hòa độc tố, ngăn ngừa sự bám dính của các độc tố trên lên các thụ thể tế bào. Như vậy, các tế bào cơ thể tránh được các rối loạn do các độc tố đó gây ra. Tương tự như vậy, nhiều virus và vi khuẩn chỉ gây bệnh khi bám được vào các tế bào cơ thể. Vi khuẩn sử dụng các phân tử bám dính là adhesine, còn virus sở hữu các protein cố định trên lớp vỏ ngoài. Các kháng thể kháng-adhesine và kháng-proteine capside virus sẽ ngăn chặn các vi sinh vật này gắn vào các tế bào đích của chúng.

- Hoạt hóa bổ thể: Một trong những cơ chế bảo vệ cơ thể của kháng thể là việc hoạt hóa dòng thác bổ thể. Bổ thể là tập hợp các protein huyết tương khi được hoạt hóa sẽ tiêu diệt các vi khuẩn xâm hại bằng cách: đục các lỗ thủng trên vi khuẩn, tạo điều kiện cho hiện tượng thực bào, thanh lọc các phức hợp miễn dịch và (4) phóng thích các phân tử hóa hướng động.
- Hoạt hóa các tế bào miễn dịch: Sau khi gắn vào kháng nguyên ở đầu biến thiên (Fab), kháng thể có thể liên kết với các tế bào miễn dịch ở đầu hằng định (Fc). Những tương tác này có tầm quan trọng đặc biệt trong đáp ứng miễn dịch. Như vậy, các kháng thể gắn với một vi khuẩn có thể liên kết với một đại thực bào và khởi động hiện tượng thực bào. Các tế bào lympho NK (Natural Killer) có thể thực hiện chức năng độc tế bào và ly giải các vi khuẩn bị opsonine hóa bởi các kháng thể.

3.4 Thụ Thể Tế Bào T Và Quá Trình Chọn Lọc Nhân Bản

Tế bào T có khả năng nhận diện kháng nguyên thông qua thụ thể bề mặt, viết tắt là TCR (T-cell receptor). Sự nhận diện này mang tính đặc hiệu cao. Chẳng hạn tế bào Tc có thể phân biệt được mỗi loại virut khi chúng xâm nhập vào cơ thể. TCR có cấu tạo gần giống KT, gồm hai chuỗi peptit: α và β, gắn với nhau bởi cầu nối disulfua. TCR cũng có hai vùng: vùng biến đổi nằm ở phía đầu amin của mỗi chuỗi tạo nên vị trí kết hợp KN. Vùng cố định nằm phía đầu cacboxyl và cắm sâu vào màng sinh chất của tế bào T. Các gen của thụ thể tế bào T: Các gen mã hóa cho các chuỗi α và β của TCR rất giống với các gen mã hóa KT. Vùng biến đổi của TCR được mã hóa bởi các gen V và MHC-I đối với chuỗi α và các gen V, D, MHC-I đối với chuỗi β. Hầu hết khả năng biến đổi được tập trung tại các điển nối giữa V-J và V-D-J, tạo thành những vùng chứa vị trí liên kết với KN lúc KN này đang nằm trên rãnh của MHC. Do vậy sự đa dạng của TCR cũng được thực hiện theo cùng một cơ chế như cơ chế tạo ra sự đa dạng của thụ thể tế bào B và KT. Tuy nhiên có một số

điểm khác là vùng cố định của TCR không có các biến dị idiotyp, không tồn tại ở dạng tiết và không có vùng xuyên màng.



Hình 3.4 Minh họa quá trình chọn lọc nhân bản

Chọn lọc nhân bản hay sinh sản có chọn lọc là một quá trình mà con người chọn các loài động vật khác và thực vật theo một vài tính trạng đặc biệt. Quá trình này nhằm đào thải những biến dị bất lợi cho con người và tích lũy những biến dị có lợi. Chọn lọc nhân bản gồm các quá trình loại bỏ các bản sao tự gây phản ứng miễn dịch, sự tăng trưởng và biệt hóa của các tế bào lympho trưởng thành, duy trì mô hình của tế bào miễn dịch tốt bằng cách nhân bản vô tính, tạo ra nhiều biến đổi mới trong các tế bào miễn dịch để tăng sự đa dạng của các tế bào đặc biệt trong nhận dạng kháng nguyên.

CHƯƠNG 4: KẾT HỢP THUẬT TOÁN PHÂN LỚP VÀ HỆ MIỄN DỊCH NHÂN TẠO

4.1 Giới Thiệu Hệ Miễn Dịch Nhân Tạo

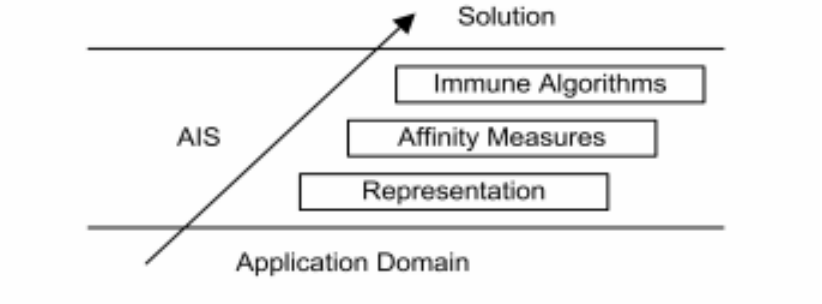
Hiện nay, hệ miễn dịch nhân tạo (AIS) đang thu hút được sự quan tâm rất lớn của các nhà khoa học trên thế giới trong việc cung cấp các giải pháp thông minh trong nhiều lĩnh vực khác nhau và được lấy cảm hứng từ hệ miễn dịch sinh học (Biology Immune System - BIS). Như đã được đề cập ở chương 3.

Do ứng dụng các nguyên lý cơ bản của BIS, đặc biệt là các nguyên lý phòng vệ của hệ này nên AIS cũng có những khả năng thông minh vượt trội như nhận dạng, ghi nhớ, và tự tổ chức. AIS có thể ứng dụng trong các lĩnh vực như nhận dạng máy tính, chế tạo robot, kiểm soát hệ thống, phát hiện virus, phát hiện xâm nhập trái phép, nhận dạng mẫu dữ liệu, tối ưu hóa, và xử lý ảnh,..

Về mặt định nghĩa: "hệ miễn dịch nhân tạo là một hệ thống thích nghi lấy ý tưởng của miễn dịch học thuyết và những chức năng, nguyên tắc, mô hình miễn dịch quan sát được, áp dụng giải các bài toán thực tế" (Castro & Timmis – 2002).

4.2 Cấu Trúc Của Hệ Miễn Dịch Nhân Tạo

Trong hệ miễn dịch nhân tạo có ba yếu tố cơ bản để thiết kế là: biểu diễn mô hình trừu tượng cho các thành phần của hệ miễn dịch gồm tế bào, phân tử và các phần tử miễn dịch; một tập các hàm xác định độ thích hợp để định lượng sự tương tác của các phần tử; một tập các thuật toán để điều khiển tính động của hệ thống.



Hình 4.1 Cấu trúc phân tầng của hệ miễn dịch nhân tạo

Application Domain: đây là tầng lĩnh vực ưng dụng, đối với lĩnh vực ứng dụng khác nhau sẽ quyết định những thành phần và cách thức biểu diễn khác nhau và dẫn tới các thao tác trên các thành phần cũng khác nhau.

Representation: đây là tầng biểu diễn, định nghĩa các thành phần của hệ miễn dịch như kháng nguyên, kháng thể,…

Affinity Measures: tầng thứ ba là các phương pháp đánh giá độ thích hợp, để đánh giá độ thích hợp có thể sử dụng nhiều phương pháp khác nhau như khoảng cách Hamming, khoảng cách Euclid, hoặc khoảng cách Mahattan.

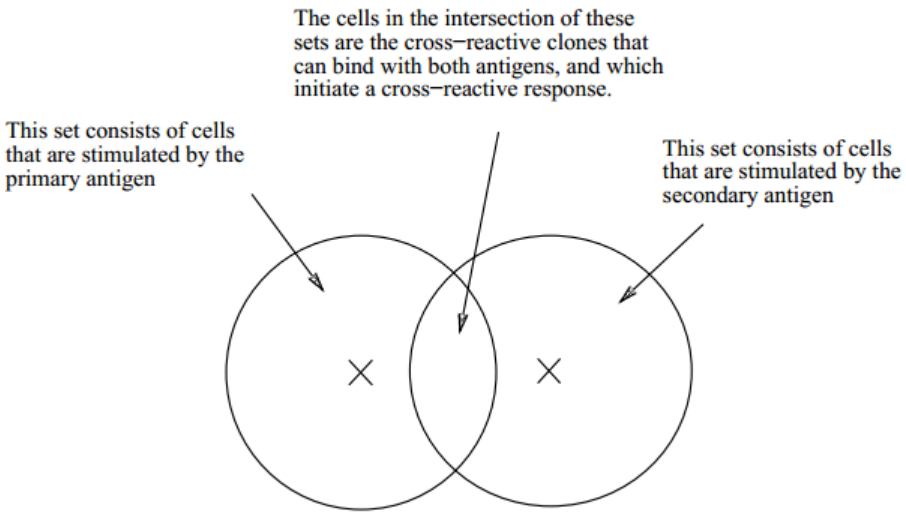
Immnue Algortihms: tầng thứ tư là sử dụng các thuật toán miễn dịch, có thể dùng các thuật toán miễn dịch như thuật toán chọn lọc tích cực, thuật toán chọn lọc tiêu cực, thuật toán chọn lọc clone, thuật toán aiNet,..để điều chỉnh tính động của hệ AIS.

Solution: tầng thứ năm là đưa ra lời giải cho bài toán, lời giải cho bài toán sẽ được cập nhật lại sau khi một quần thể mới được tạo và đưa ra kết quả cuối cùng khi đạt đến điều kiện kết thúc nào đó ví dụ như sau một số bước lặp nhất định.

4.2.1 Không gian hình (Shape-space)

Trong không gian hình sự liên kết giữa các thành phần miễn dịch như kháng nguyên và kháng thể được biểu diễn. Các thành phần như kháng nguyên và kháng thể sẽ được đặc trưng biểu một các thuộc tính nào đó. Và không gian hình đã trở thành một khuôn mẫu cho để biểu diễn các thành phần miễn dịch trong các ứng dụng hệ miễn dịch nhân tạo.

Thông thường trong không gian hình các khoảng cách giữa các thành phần thường được đo bằng khoảng cách giữa các điểm biểu cho chúng theo một luật so khớp nào đó.



Hình 4.2 Hình mô phỏng quá trình tương tác giữa 2 kháng nguyên

Trong không gian hình, các thành phần của hệ miễn dịch như kháng nguyên, kháng thể, tế bào T, tế bào B có thể được biểu diễn theo các cách như: chuỗi nhị phân, chuỗi hữu hạn các ký tự alphabet, các vector giá trị thực. Ngoài ra, còn có thể biểu diễn bằng cách phối hợp các kiểu dữ liệu khác nhau.

4.2.2 Các Thành Phần Sinh Học Của Hệ Miễn Dịch

Tế bào thực bào: là các tế bào có khả năng nuốt, tiêu các vi sinh vật.

Tiểu thực bào: là các bạch cầu hạt trung tính ở máu làm nhiệm vụ ăn các đối tượng có cỡ khổ bé.

Đại thực bào: là tế bào có nguồn gốc từ tuỷ xương phân hoá thành tế bào Monocyte ở máu rồi di chuyển đến các mô để trở thành các tế bào của hệ thống võng nội mô có những tên khác nhau: tế bào kupffer (gan), langerhans (ở da)... Đối với một số chất sau khi năng nuốt và xử lý thì đại thực bào sẽ trình diện kháng nguyên để hoạt hoá tế bào lympho T.

Thực bào: là hiện tượng các bạch cầu hình thành chân giả bắt và nuốt các vi khuẩn vào tế bào rồi tiêu hoá chúng.

+ Có 2 loại bạch cầu tham gia vào thực bào: Bạch cầu trung tính và bạch cầu mono (đại thực bào).

+ Limpho B tiết ra kháng thể vô hiệu hoá kháng nguyên.

+ Limpho T phá huỷ các tế bào cơ thể bị nhiễm vi khuẩn, virus bằng cách tiết ra các prôtêin đặc hiệu (kháng thể) làm tan màng tế bào bị nhiễm để vô hiệu hoá kháng nguyên.

4.3 Một Số Luật So Khớp Chuỗi

Luật so khớp chuỗi là nền tảng của các giải thuật phát hiện, phân lớp và nhận dạng. Và việc chọn một luật so khớp phù hợp nhằm mục đích đạt hiệu quả trong ứng dụng. Phần này chủ yếu tập trung vào 3 luật so khớp là: Hamming, Edit và R-Contiguous.

4.3.1 Luật So Khớp Hamming

Luật so khớp Hamming (hay khoảng cách Hamming) giữa hai dãy ký tự (strings) có chiều dài bằng nhau là số các ký hiệu ở vị trí tương đương có giá trị khác nhau. Nói một cách khác, khoảng cách Hamming đo số lượng thay thế cần phải có để đổi giá trị của một dãy ký tự sang một dãy ký tự khác, hay số lượng lỗi xảy ra biến đổi một dãy ký tự sang một dãy ký tự khác.

Đối với hai dãy ký tự nhị phân (binary strings) a và b, phép toán này tương đương với phép toán a XOR b. Khoảng cách Hamming của các dãy ký tự nhị phân còn tương đương với khoảng cách Manhattan (Manhattan distance) giữa hai giao điểm của một hình giả phương cấp n (n-dimensional hypercube), trong đó n là chiều dài của các từ.

Ví dụ: Khoảng cách Hamming giữa chuỗi A= 10**11**01 và B= 1001001 là 2.

Khoảng cách Hamming h(a,b) giữa 2 chuỗi a và b được định nghĩa dưới dạng công thức như sau:

$$h(a,b) = \sum_{i=1}^{N} (\overline{A_i \oplus B_i})$$

Trong đó:

N là độ dài của chuỗi.

A_i là ký hiệu cho bit thứ i trong chuỗi a.

B_i là ký hiệu cho bit thứ i trong chuỗi b.

$\oplus$ là phép toán logic XOR trong hệ nhị phân.

4.3.2 Luật So Khớp Edit

Luật so khớp Edit giữa 2 chuỗi x1 và x2 là số lượng phép biển đổi chuỗi tối thiểu để chuyển từ chuỗi x1 thành x2. Để thực hiện phép biến đổi chuỗi ta có thể thêm 1 ký tự, xóa hoặc thay đổi 1 ký tự. Đó là sự tổng quá hóa của luật so khớp Hamming.

4.3.3 Luật So Khớp R-Contiguous

Luật so khớp R-Contiguous được phát biểu như sau: Nếu a và b là những chuỗi có độ dài bằng nhau được định nghĩa dựa trên bảng chữ cái Alphabet hữu hạn thì match(a,b) là true khi a và b giống ít nhất ở r vị trí liên tiếp.

Ví dụ:

Chuỗi A = 011010**1010**

và B = 0101**011010**

Trong đó A và B là 2 chuỗi được định nghĩa trên Alphabet {0,1} và match(A,B) ở 011010.

Trong trường hợp nếu các phần tử là dạng chuỗi nhị phân thì luật so khớp R-Contiguous Bits sẽ phát huy được tác dụng trong đó detector d là một chuỗi nhị phân c và có ngưỡng r. Detector d khớp với một chuỗi x nếu rcb của c và x đạt hoặc vượt ngưỡng r. Việc chọn rcb sẽ đơn giản hóa các phân tích tính toán học và là một mô hình tốt trong việc so khớp gần đúng của tế bào T. Tham số r xác định bậc của detector của một trường hợp đặc biệt; nếu giá trị r càng nhỏ thì detector càng tổng quát.

4.4 Một Số Thuật Toán Trong Hệ Miễn Dịch Nhân Tạo

Trong phần này sẽ trình bày một số thuật toán trong hệ miễn dịch nhân tạo và nó đóng vai trò quan trọng trong việc có đưa ra một giải pháp hiệu quả hay không trong ứng dụng sử dụng hệ miễn dịch nhân tạo để giải quyết bài toán.

4.4.1 Thuật Toán Chọn Lọc Clone (Clonal Selection Algorithm: CLONALG)

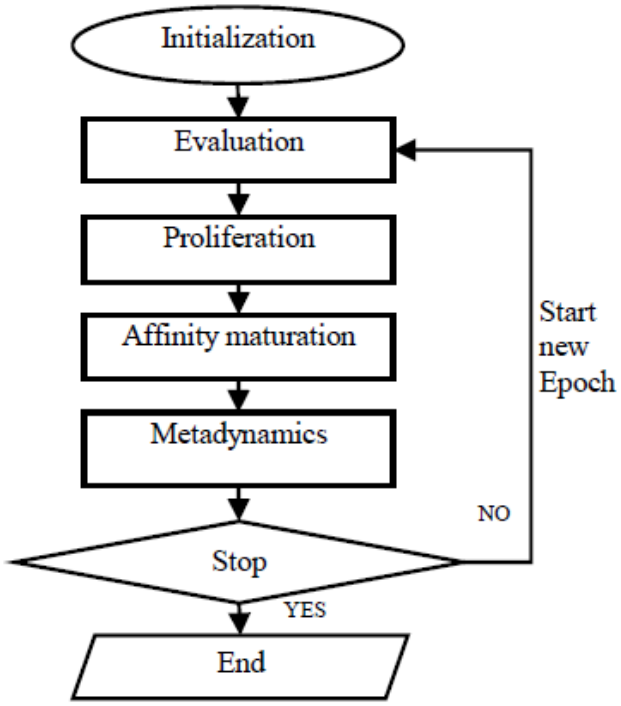
Thuật toán chọn lọc nhân bản được sử dụng làm cơ sở để cải tiến hệ miễn dịch nhân tạo trong các tác vụ về tối ưu hóa tính toán và nhận dạng mẫu. Trong đó,

mô hình mà ta sử dụng sẽ mô phỏng quá trình "trưởng thành ái lực đặc trưng cho từng kháng nguyên" của các tế bào loại B cùng với cơ chế siêu đột biến liên quan.

Các hệ miễn dịch nhân tạo cũng thường sử dụng ý tưởng về các tế bào nhớ để lưu lại các giải pháp tốt cho những vấn đề đã được giải quyết. Trong tài liệu [15] "Artificial immune systems: a new computational intelligence approach" của de Castro và Timmis, các tác giả nêu bật hai đặc tính quan trọng của quá trình trưởng thành ái lực của các tế bào loại B, mà theo họ, hoàn toàn có thể được khai thác khi xét từ góc độ tính toán.

Đặc tính đầu tiên đề cập rằng mức độ sinh sôi các tế bào loại B tỉ lệ thuận với ái lực đối với kháng nguyên mà nó tương ứng, do đó, ái lực càng cao, càng nhiều dòng vô tính được tạo ra. Đặc tính thứ hai, sự đột biến của các kháng thể tế bào loại B tỉ lệ nghịch với ái lực của kháng nguyên mà nó tương ứng. Tận dụng hai đặc tính này, de Castro và Von Zuben [13] đã phát triển CLONALG - một trong những phép chọn lọc nhân bản được sử dụng rộng rãi nhất hiện nay. Phép chọn lọc này được sử dụng để triển khai các tác vụ tối ưu hóa so khớp mẫu và hàm đa phương thức.

Khi ứng dụng vào so khớp mẫu, một tập mẫu S được xem như kháng nguyên để so khớp. Nhiệm vụ của CLONALG là sản sinh ra một tập các kháng thể nhớ M, các kháng thể nhớ này sẽ tương ứng với từng thành phần trong S. Thuật toán CLONALG được trình bày như sau:



Hình 4.3 Thuật toán chọn lọc nhân bản

input : S = tập các mẫu cần nhận dạng.

n = số lượng các phần tử xấu nhất sẽ bị loại bỏ.

output : M = tập các bộ dò ghi nhớ có khả năng phân loại các mẫu chưa biết.

begin

Tạo ra tập ngẫu nhiên các kháng thể ban đầu A.

forall patterns in S do

Xác định ái lực với từng kháng thể trong A.

- Sản sinh các dòng vô tính của một tập con các kháng thể trong A mà có độ ái lực lớn nhất.
- Số lượng các dòng vô tính của một kháng thể tỉ lệ thuận với ái lực của nó.
- Biến đổi thuộc tính các dòng vô tính này, đặt chúng vào tập A, và đặt một phiên bản của kháng thể có ái lực cao nhất trong A vào tập nhớ M.

- Thay thế n kháng thể có ái lực thấp nhất trong A với các kháng thể mới được sản sinh tự động.

end

end.

4.4.2 Thuật Toán Chọn Lọc Âm Tính (Negative Selection Algorithms: NSA)

Thuật toán lấy ý tưởng từ việc quan sát hệ thống miễn dịch bảo vệ cơ thể vật chủ từ xâm nhập mầm bệnh. Trong quá trình chọn lọc của thuật toán, chỉ giữ lại những tế bào T không nhận dạng được liên kết giữa Major Histocompatibility Complex (MHC) và self-peptide. Còn đối với các tế bào T nhận dạng được sẽ bị loại bỏ. Theo [15] đã đề xuất một mô hình tính toán phân biệt đối với self và nonself mô phỏng quá trình chọn lọc âm tính – quá trình tăng trưởng của tế bào T ở tuyến ức, nó được gọi là giải thuật chọn lọc âm tính (Negative Selection Algorithm – NSA)

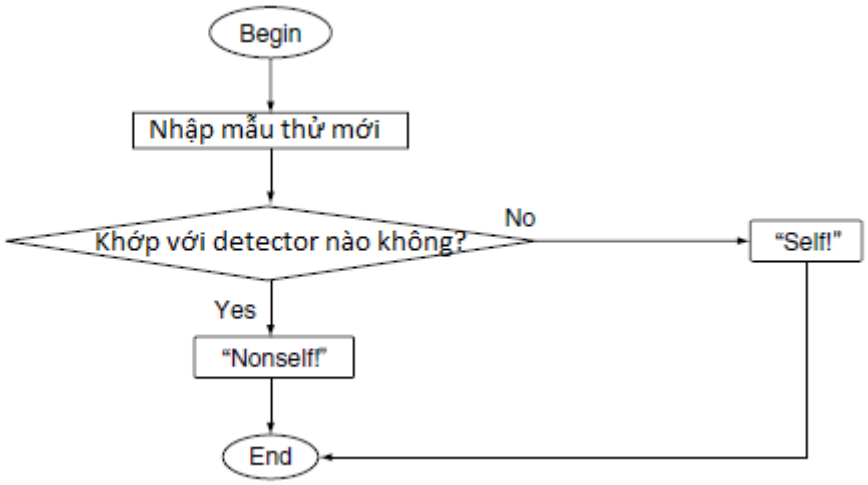
Thuật toán được mô tả như sau [15]:

```
input: S = set of self strings characterising benign, normal data.
output: A = Stream of nonself strings detected.

begin
    Create empty set of detector strings D          ▷ Generation of detector strings
    Generate random strings R.
    for all random strings r ∈ R do
        for all self strings s ∈ S do
            if r matches s then
                Discard r
            else
                Place r in D
            end if
        end for
    end for

    while There exist protected strings p to check do          ▷ Detection stage
        Retrieve protected string p
        for all detector strings d ∈ D do
            if p matches d then
                Place p in A and output.          ▷ Nonself string detected
            end if
        end for
    end while
end
```

Hình 4.4 Thuật toán chọn lọc âm tính



Hình 4.5 Mô hình thuận toán NSA

Begin

Giai đoạn 1. Tạo tập detectors không nhận dạng tập self đã chuẩn bị trước

Giai đoạn 2: Kiểm tra tập mẫu chưa xác định bằng tập detector và mẫu bị detector nhận dạng thì bị đánh dấu là nonself.

End

4.4.3 Thuật Toán Chọn Lọc Dương Tính (Positive Selection algorithms: PSA)

Ngược lại với quá trình chọn lọc âm tính, trong chọn lọc dương tính, các tế bào T được kiểm tra bằng cách nhận dạng các phần tử giữa Major Histocompatibility Complex (MHC) được biểu thị ở các tế bào biểu mô của vỏ não. Nếu một tế bào T thất bại trong việc nhập dạng một trong các phân tử MHC, nó sẽ bị loại, ngược lại thì giữ lại.

Tương tự như giải thuật chọn lọc âm tính, trong chọn lọc dương tính ta cũng tạo ra bộ detector nhưng là bộ detector để nhận dạng self.

Ở bước giai đoạn 1: tạo detector, các detector được cho nhận dạng tập mẫu self và nếu detector không nhận dạng được tất cả các mẫu self thì chúng sẽ bị loại bỏ.

Ở giai đoạn 2: thì các detector ở giai đoạn 1 được sử dụng để nhận dạng mẫu đưa vào kiểm là self nếu chúng khớp với nhau và ngược lại là nonself.

4.5 Các Thuật Toán Phân Lớp

Phân lớp dữ liệu là một trong những hướng nghiên cứu chính của ngành khai phá dữ liệu. Có nhiều thuật toán máy học cho bài toán phân lớp có thể được dùng để kết hợp với một hệ AIS. Trong luận văn này sẽ tiến hành nghiên cứu 3 thuật toán phân lớp dùng trong việc kết hợp với hệ miễn dịch nhân tạo đó là: K-Nearest Neighbors (KNN), Support Vector Machine (SVM) và Radial Basis Function (RBF).

4.5.1 Thuật toán K – Láng giềng gần nhất (K-Nearest Neighbors: KNN)

Theo [16] KNN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp với tất cả các đối tượng trong Training Data.

Một đối tượng được phân lớp dựa vào k láng giềng của nó. K là số nguyên dương được xác định trước khi thực hiện thuật toán. Người ta thường dùng khoảng cách Euclidean để tính khoảng cách giữa các đối tượng.

Thuật toán được thực hiện theo các bước:

- Xác định giá trị tham số K (số láng giềng gần nhất).
- Tính khoảng cách giữa đối tượng cần phân lớp với tất cả các đối tượng trong training data (thường sử dụng khoảng cách Euclidean, Cosine…).
- Sắp xếp khoảng cách theo thứ tự tăng dần và xác định k láng giềng gần nhất với đối tượng cần phần lớp.
- Lấy tất cả các lớp của k láng giềng gần nhất đã xác định.
- Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho đối tượng.

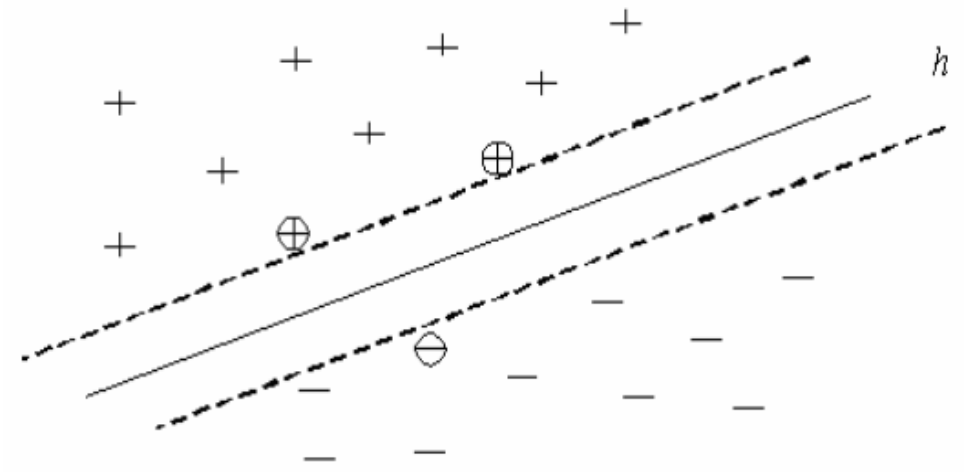
4.5.2 Thuật Toán Phân Loại SVM

Theo [2] (SVM) là một phương pháp phân lớp dựa trên lý thuyết học thống kê, được đề xuất bởi Vapnik (1995).

Ý tưởng chính của thuật toán này là cho trước một tập huấn luyện được biểu diễn trong không gian vector trong đó mỗi tài liệu là một điểm, phương pháp này tìm ra một mặt phẳng h quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng lớp + và lớp -. Chất lượng của siêu mặt phẳng

này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này.

Khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt đồng thời việc phân loại càng chính xác. Mục đích thuật toán SVM tìm ra được khoảng cách biên lớn nhất để tạo kết quả phân lớp tốt.



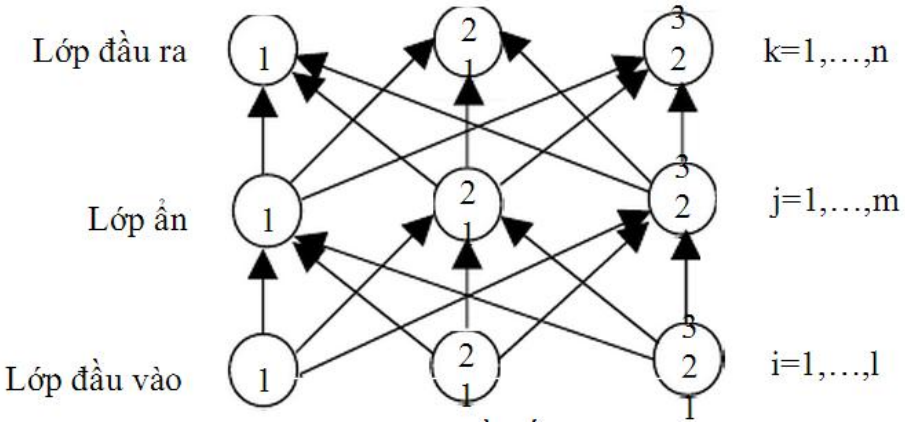
Hình 4.6: Minh họa thuật toán SVM

Hình 4.6 miêu tả Siêu phẳng h phân chia dữ liệu huấn luyện thành 2 lớp + và - với khoảng cách biên lớn nhất. Các điểm gần h nhất là các vector hỗ trợ (Support Vector - được khoanh tròn).

SVM sử dụng thuật toán học nhằm xây dựng một siêu phẳng làm cực tiểu hoá độ phân lớp sai của một đối tượng dữ liệu mới. Độ phân lớp sai của một siêu phẳng được đặc trưng bởi khoảng cách bé nhất tới siêu phẳng đấy. SVM có khả năng rất lớn cho các ứng dụng được thành công trong bài toán phân lớp văn bản.

4.5.3 Thuật Toán Phân Loại Mạng RBF

Thuật Toán RBF đã có từ lâu trong lý thuyết xấp xỉ và được sử dụng để xấp xỉ hàm chưa biết dựa trên cơ sở các cặp điểm vào – ra biểu diễn hàm chưa biết đó.



Hình 4.7 Sơ đồ cấu trúc mạng RBF

Trên hình 4.7 biểu diễn mạng RBF nhiều đầu vào nhiều đầu ra. Lớp đầu vào phân bố mỗi thành phần của véc tơ đầu vào cho tất cả các nút ẩn. Mỗi nút ẩn trong lớp ẩn chứa một trongnhững tâm của RBF và áp hàm cơ sở W cho khoảng cách euclidean giữa véc tơ đầu vào và tâm. Do đó mỗi nút trong lớp ẩn đưa ra một giá trị vô hướng dựa trên tâm mà nút đó có.

Các đầu ra của lớp ẩn được truyền đến lớp đầu ra với các liên kết trọng số. Nút ở lớp đầu ra cộng các đầu vào của nó để tạo ra các đầu ra của mạng.

CHƯƠNG 5: THỬ NGHIỆM, ĐÁNH GIÁ VÀ HƯỚNG PHÁT TRIỂN

Chương này đưa ra kết quả cài đặt thử nghiệm giữa sự kết hợp của thuật toán phân lớp và hệ miễn dịch nhân tạo qua đó tiến hành so sánh đánh giá kết quả thực nghiệm.

5.1 Chuẩn Bị Dữ Liệu

Bảng 5.1 Bộ dữ liệu thử nghiệm

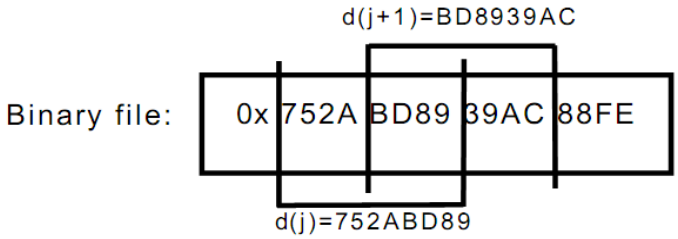
	Bộ huấn luyện		Bộ kiểm thử	
Data sets (Bộ dữ liệu)	Benign Files (File sạch)	Virus Files (File virus)	Benign Files (File sạch)	Virus Files (File virus)
Dataset 1	71	885	213	2662
Dataset 2	142	1773	142	1773
Dataset 3	213	2662	71	885

Bộ dữ liệu chuẩn bị bao gồm tổng cộng 284 file sạch với dung lượng ~78MB, và 3547 file virus với dung lượng ~7.8MB. Bộ data set khác chứa 208 file sạch sử dụng cho chọn lọc âm tính với tổng dung lượng là 189MB. Tất cả những file sạch đều là file phổ biến và có đuôi mở rộng là *.exe

Như trên bảng 5.1 bộ dữ liệu được chia một cách ngẫu nhiên giữa hai bộ huấn luyện và bộ kiểm thử và không bị chồng lấp lên nhau.

5.2 Xây Dựng Bộ Detector (Virus Detector System: VDS)

Theo [1], [5], [14], các bộ phát hiện sẽ có chiều dài l = 32 và là các chuỗi bit nhị phân, tức là m = 2. Các chuỗi nay sẽ được rút trích trực tiếp từ các file virus theo nguyên tắc :



Hình 5.1 Nguyên tắc rút trích đoạn bit nhị phân

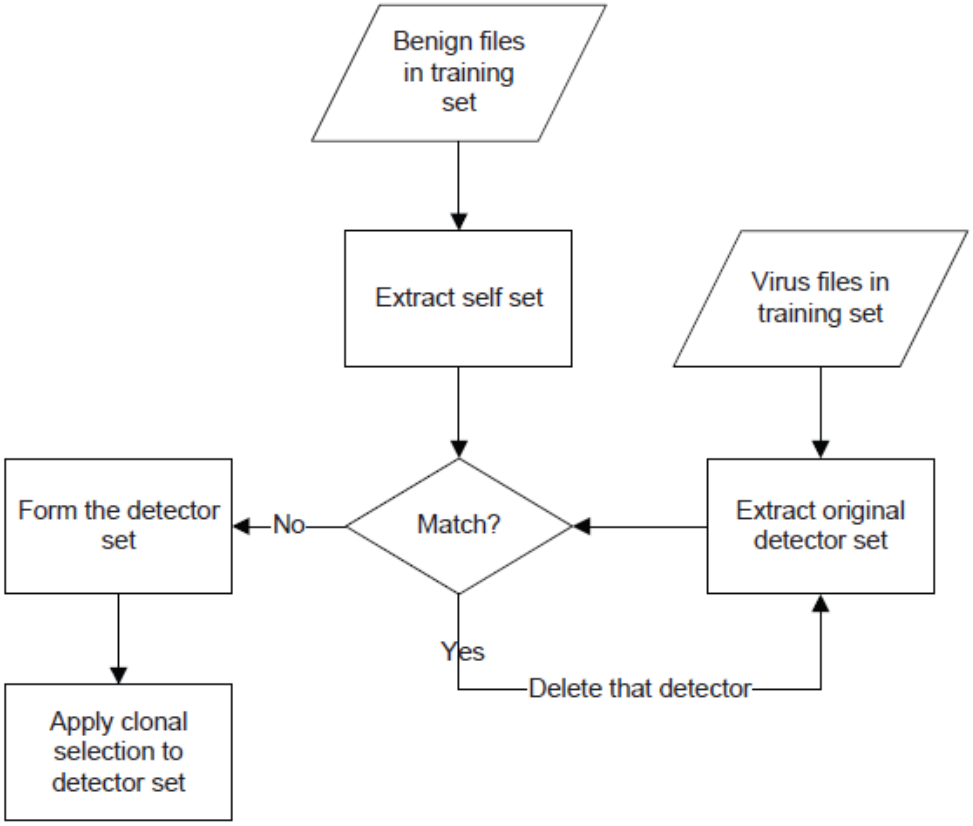
Các đoạn chuỗi nhị phân dài l = 32 bit sẽ được rút trích từ file nhị phân một cách liên tục, mỗi chuỗi nhị phân liên tiếp nhau sẽ trùng lên nhau một đoạn l/2. Việc rút trích trực tiếp từ file virus với mật độ cao như vậy nhằm làm tăng tính đa dạng của các bộ phát hiện và tránh bỏ sót các dấu hiệu có thể dùng để nhận dạng virus.

Sự biểu diễn của chuỗi bit nhị phân l = 32 cho phép hệ thống VDS có thể nhận ra các phần tử khác nhau thông qua việc so khớp chuỗi. Nhưng như đã phân tích về hệ miễn dịch, một đặc tính cực kỳ thú vị đó là sự khái quát hóa và suy rộng của việc so khớp chuỗi có thể được thi công trong hệ VDS sử dụng thuật toán so khớp gần đúng.

5.3 Tiến Hành Xử Lý Dữ Liệu Bằng Chọn Lọc Âm Tính

Sau khi ta có hai tập gen sạch và gen virus đó là về mặt lý thuyết thì các tập tin virus cũng chính là các chương trình và có thể tập tin bị nhiễm virus. Trong tập gen virus lúc này hoàn toàn có thể chứa các gen hay chi chỉ thông thường ngoài các gen đặc biệt có ở một chương trình virus. Chính vì thế, một yêu cầu đặt ra là làm sạch tập gen virus hiện tại.

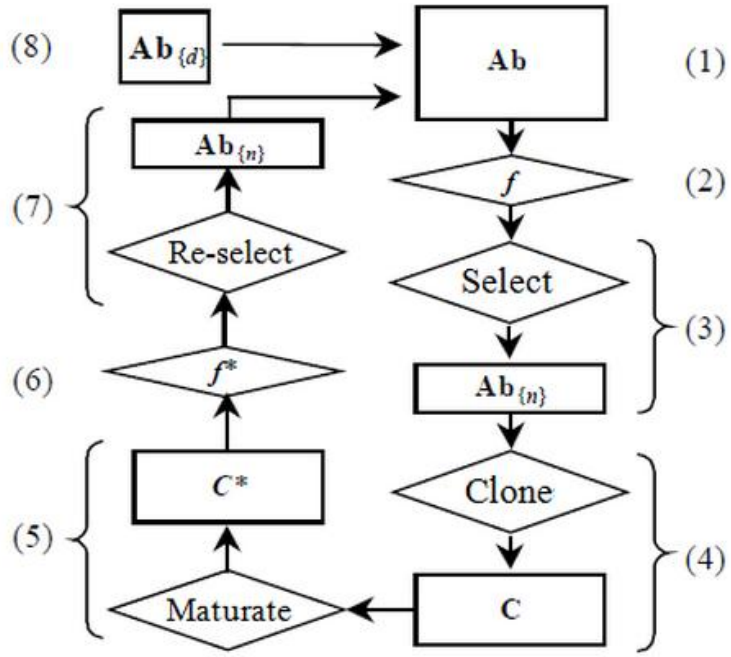
Để làm sạch tập gen virus này ta áp dụng giải thuật NSA. Quá trình làm sạch tập gen virus có thể mô tả như sau:



Hình 5.2 Quá trình xử lý NSA

5.4 Tiến Hành Xử Lý Dữ Liệu Bằng Chọn Lọc Nhân Bản

Sau khi thực hiện tiến trình ở bước 5.3 qua bước này ta tiếp tục thực hiện thuật toán CLONALG để sản sinh ra tập các kháng thể nhớ m. Sau quá trình huấn luyện , tập m bao gồm các kháng thể ghi nhớ sẽ được chọn làm kết quả của thuật toán để nhận diện các mẫu kháng nguyên mà hệ thống cần phát hiện.



Hình 5.3 Mô hình thuật toán CLONALG

5.5 Tiến Hành Đo Khoảng Cách

Sau khi tiến hành phép chọn lọc âm tính và chọn lọc nhân bản ta tiến hành đo khoảng cách dựa trên khoảng cách Hamming và khoảng cách r-Continuous để xác định đặc trưng cho thuật toán phân lớp.

5.6 Affinity Vector (Đo Độ Vector thích hợp)

Theo [14] đã giới thiệu một công thức tính độ nguy hiểm để từ đó đánh giá lựa chọn các yếu tố phù hợp cho hệ thống.

Công thức tính độ nguy hiểm DL (Danger Level) của một chuỗi bit x được mô tả như sau:

$$DL(x) = \frac{\sum_{i=1}^{|S_{detector}|} \left\langle HA(x, S_{\det ector}), RCBA(x, S_{\det ector}, 12), RCBA(x, S_{\det ector}, 24) \right\rangle}{|S_{\det ector}|}$$

Trong đó: $S_{detector}$ là tập detector, x là một chuỗi bit được trích xuất từ tập tin L, HA(x, $S_{detector}$) là khoảng cách hamming trung bình của giữa x với các detector trong $S_{detector}$, RCBA(x, $S_{detector}$, m) là giá trị trung bình của các kết quả của luật so khớp R-Contiguous bit giữa x và các detector trong tập $S_{detector}$ với ngưỡng là m.

Mỗi phép so khớp R- Contiguous sẽ trả về 1 nếu hai chuỗi khớp với nhau và 0 nếu hai chuỗi không khớp dựa trên ngưỡng m.

5.7 Tiến Hành Xây Dựng Phân Lớp

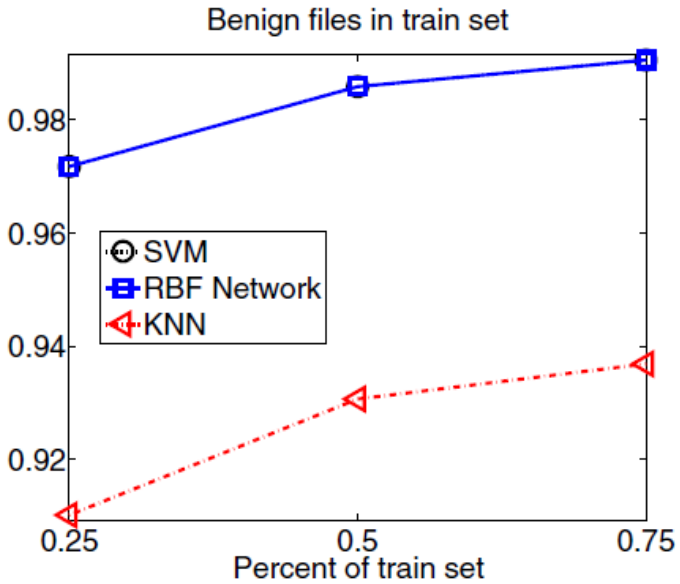
Sau khi ta đo được độ vector thích hợp ở bước này ta sẽ tiến hành xây dựng phân lớp dựa vào thuật toán phân lớp KNN, SVM, và RBF để phát hiện virus và so sánh hiệu suất giữa 3 thuật toán phân lớp này .

5.8 Kết Quả Thực Nghiệm Và Đánh Giá

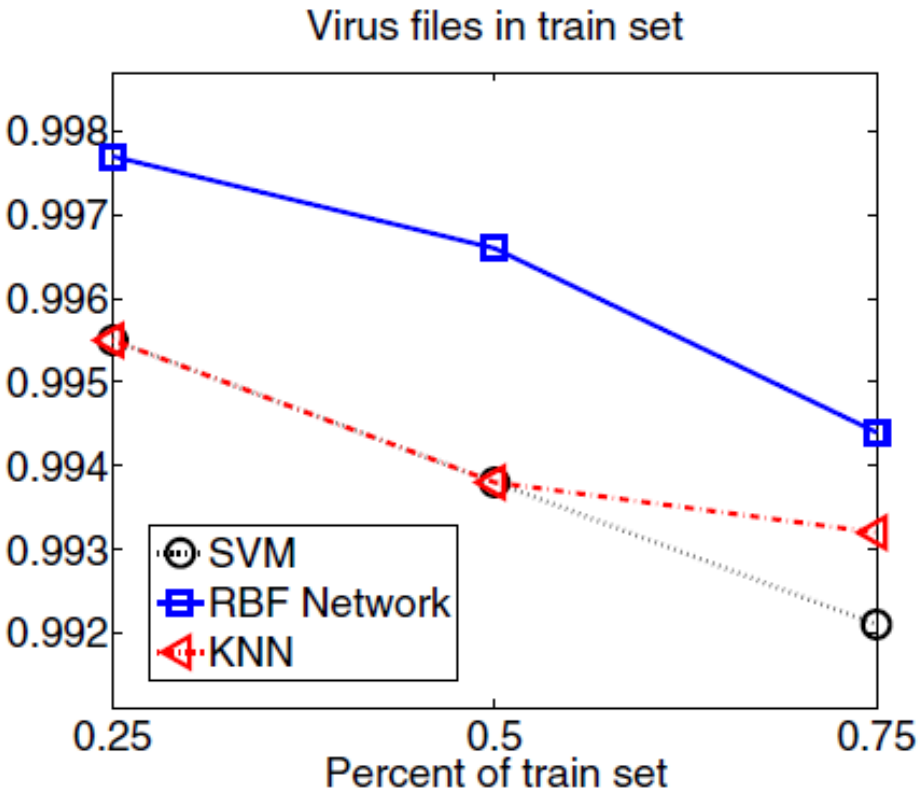
Qua quá trình thực nghiệm thu được tỉ lệ phát hiện trung bình của SVM khi L= 32 và L= 64, với thư mục được chọn một cách ngẫu nhiên từ bộ dữ liệu.

Bảng 5.2 Tỉ lệ phát hiện trung bình của SVM

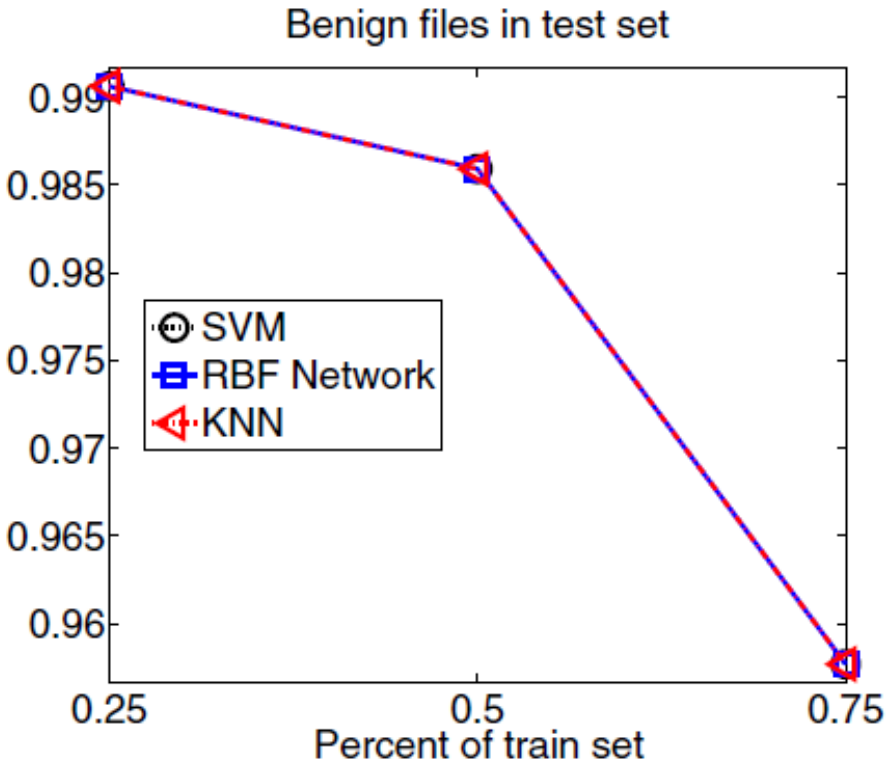
Tỉ lệ phát hiện		L =32		L =64	
Database		Virus	File sạch	Virus	File sạch
Dataset 1	Traning Set	99.55%	97.18%	100%	97.18%
	Testing Set	91.28%	99.06%	84.44%	99.53%
Dataset 2	Traning Set	99.38%	98.59%	100%	97.18%
	Testing Set	92.45%	98.59%	89.06%	97.89%
Dataset 3	Traning Set	99.21%	99.06%	100%	99.53%
	Testing Set	93.46%	95.77%	89.06%	97.18%



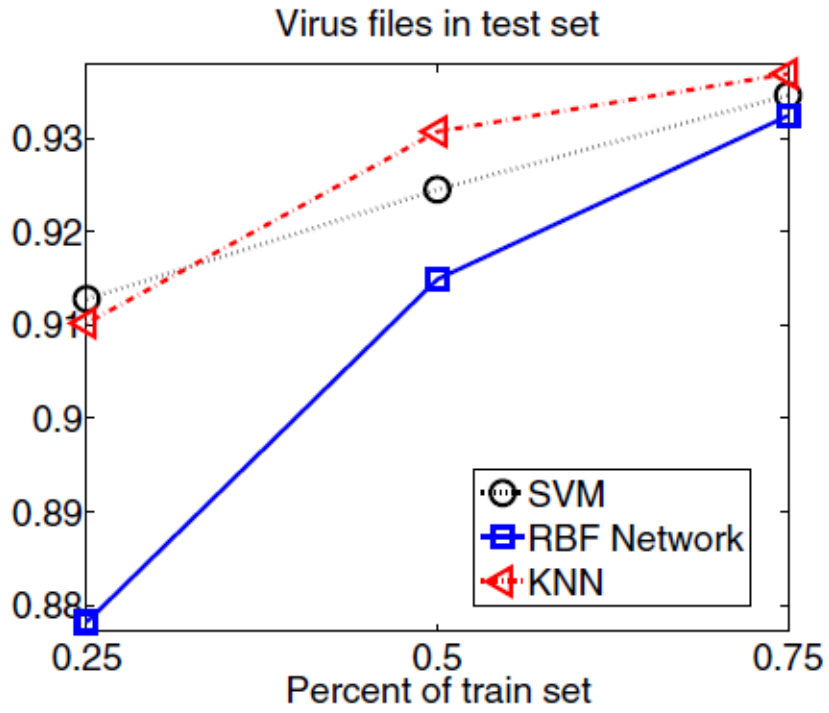
Hình 5.4 Kết quả tỉ lệ phát hiện trung bình của SVM, KNN và RBF với L=32



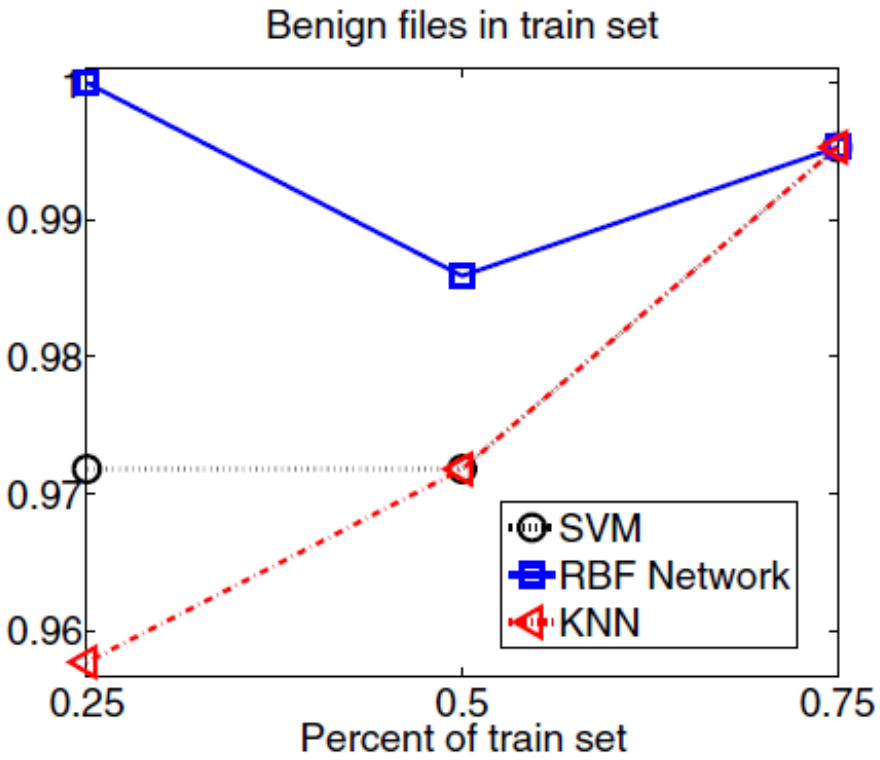
Hình 5.5 Kết quả tỉ lệ phát hiện trung bình của SVM, KNN và RBF với L=32



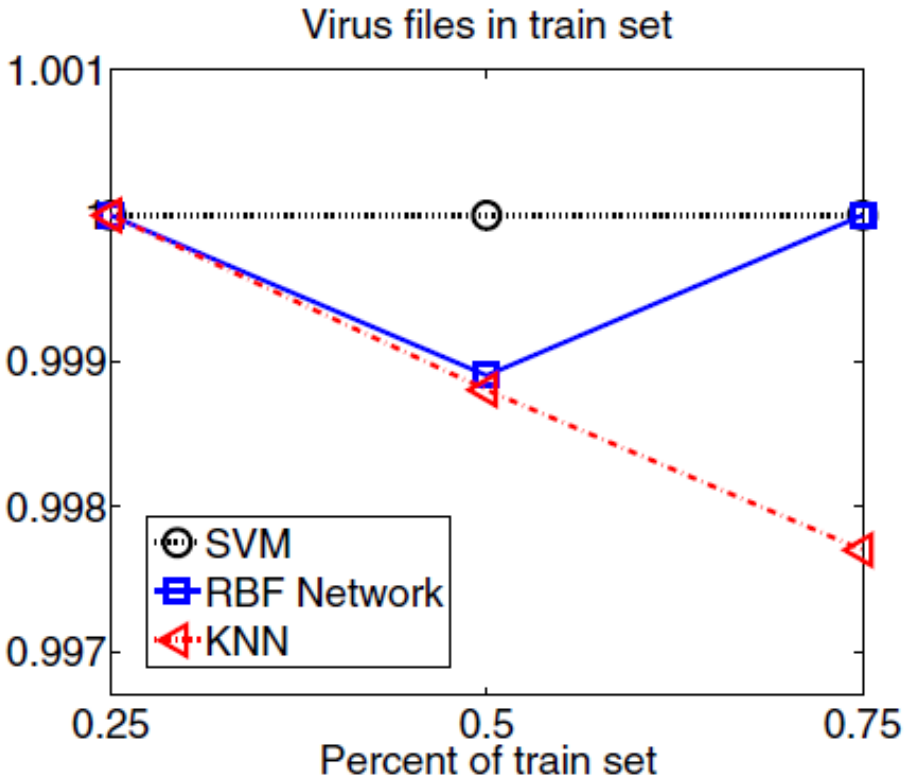
Hình 5.6 Kết quả tỉ lệ phát hiện trung bình của SVM, KNN và RBF với L=32



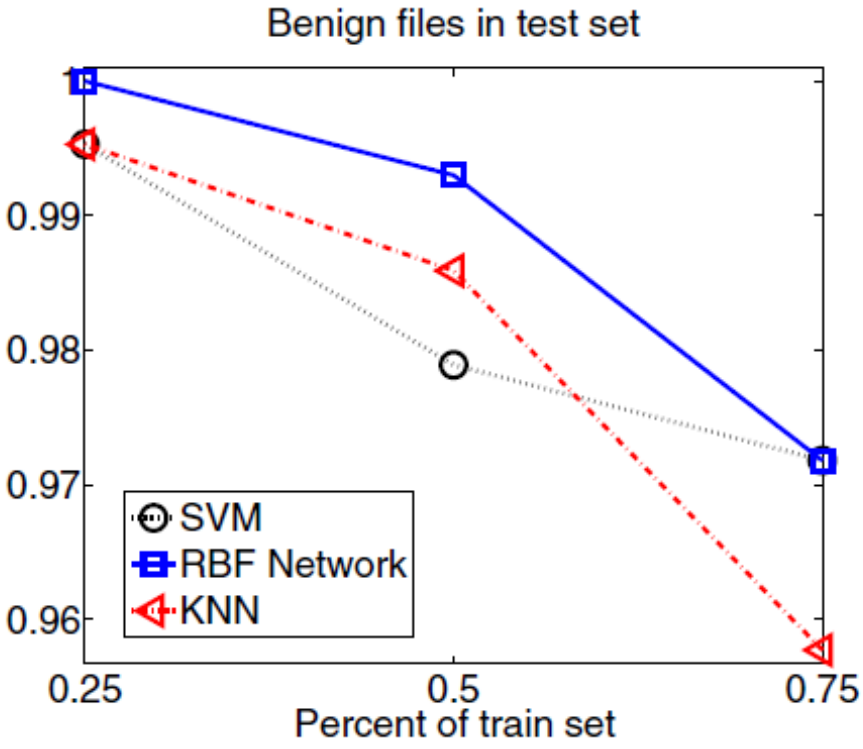
Hình 5.7 Kết quả tỉ lệ phát hiện trung bình của SVM, KNN và RBF với L=32



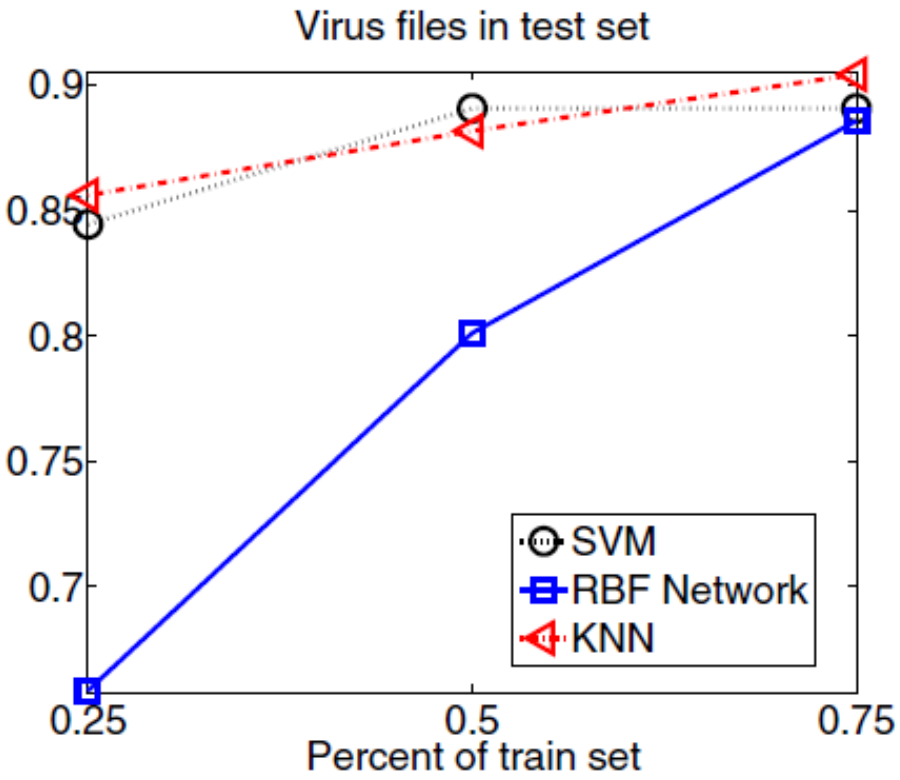
Hình 5.8 Kết quả tỉ lệ phát hiện trung bình của SVM, KNN và RBF với L=64



Hình 5.9 Kết quả tỉ lệ phát hiện trung bình của SVM, KNN và RBF với L=64



Hình 5.10 Kết quả tỉ lệ phát hiện trung bình của SVM, KNN và RBF với L=64



Hình 5.11 Kết quả tỉ lệ phát hiện trung bình của SVM, KNN và RBF với L=64

Nhận xét:

Qua bảng 5.2 ta nhận thấy tỉ lệ phát hiện sử dụng phân lớp SVM với chiều dài L khác nhau thì có kết quả khác nhau. Bộ dữ liệu testing có độ chính xác tốt hơn trong bộ detector 32-bits, Vì trong bộ phân mảnh dữ liệu 64 bits do chứa quá nhiều mã benign nên cần phải giảm bớt một số thông tin không rõ về chúng .

Ở kết quả các biểu đồ chỉ ra kết quả hiệu suất của quá trình so sánh giữa các thuật toán phân lớp SVM, RBF và KNN. Và trong 3 phân lớp đó thì thuật toán RBF là có hiệu suất tốt hơn so với 2 thuật toán còn lại ngoại trừ file virus trong bộ testing set.

RBF có hiệu suất tốt nhưng nó lại có nhược điểm là khả năng tạo ra các detector yếu hơn khi dữ liệu training nhỏ.

KNN và SVM có tỉ lệ phát hiện tương đối gần nhau với L =32 bit. Trong trường hợp L = 64 bit thì SVM có hiệu suất tốt hơn so với phân lớp KNN. Khi dữ liệu training nhiều hơn thì việc thực thi của KNN có tầm quan trọng hơn nhưng về tỉ lệ phát hiện thì SVM vẫn cho kết quả ổn định hơn.

Hệ thống phát hiện virus (VDS) đạt được độ chính xác cao trong việc phát hiện đã nhận biết và chưa biết đặc biệt là trong bộ dataset 1. Nhưng trong kết quả các biểu đồ thì tỉ lệ phát hiện của file sạch trong bộ dữ liệu testing giảm khi số lượng file virus trong bộ training tăng.

Kết quả cũng chỉ ra rằng tỉ lệ phát hiện virus của file virus trong bộ dữ liệu testing là tăng lên khi số lượng của file virus tăng thêm trong bộ dữ liệu training.

Trong dataset, kích thước của file dữ liệu sạch lớn hơn rất nhiều so với kích thước của file virus. Nó cũng là vấn đề nan giải chung cho tình hình phát triển phần mềm máy tính . Một khi kích thước của dataset ngày càng lớn hơn thì cần phải giảm một phần và tăng tính trung hòa lẫn nhau để tỉ lệ phát hiện đạt được mức ổn định. Và đó cũng là hướng để chúng ta nghiên cứu phát triển trong tương lai.

CHƯƠNG 6: KẾT LUẬN

Qua quá trình nghiên cứu, bản thân cũng đã trang bị được nhiều kiến thức quan trọng và hữu ích để có thể phát triển tiếp trong tương lai như:

- Hiểu rõ được một số kiến thức về virus máy tính
- Các mô hình lý thuyết về hệ miễn dịch sinh học
- Hệ miễn dịch nhân tạo và một số thuật toán trong hệ miễn dịch nhân tạo.
- Tìm hiểu và có kiến thức về xây dựng hệ thống phát hiện virus máy tính dựa trên việc kết hợp giữa các thuật toán với nhau.

Những kết quả mà luận văn thực hiện:

- Về lý thuyết: luận văn tập trung vào việc nghiên cứu về hệ miễn dịch cũng như các thuật toán để có thể ứng dụng trong việc phát hiện virus trên máy tính.
- Về thực tiễn, luận văn đã đánh giá và đưa ra tỉ lệ phát hiện virus cũng như so sánh hiệu suất giữa các thuật toán phân lớp kết hợp với hệ miễn dịch nhân tạo để giải quyết bài toán.

Một số ưu điểm và nhược điểm trong luận văn đã thực hiện :

6.1 Ưu điểm

Luận văn đã nghiên cứu được những kiến thức nền tảng về hệ miễn dịch tạo tiền đề cho hướng phát triển tương lai.

Tìm hiểu và tiến hành thực nghiệm, đánh giá và đưa ra kết quả cho việc phát hiện virus dựa trên sự kết hợp giữa thuật toán phân lớp và hệ miễn dịch nhân tạo.

6.2 Nhược Điểm

Hiện tại, hệ thống phát hiện virus cần phải có thời gian để thực hiện việc huấn luyện dữ liệu và kích thước của dataset ngày càng lớn dẫn đến việc chiếm tài nguyên máy tính khi sử dụng.

Những kiến thức về hệ miễn dịch sinh học cũng như hệ miễn dịch nhân tạo chỉ ở mức cơ sở chưa thực sự chuyên sâu và cần thiết phải bổ sung.

6.3 Hướng Phát Triển

Tiến hành kết hợp nhiều thuật toán hơn nữa trong hệ miễn dịch nhân tạo để thực thi xây dựng hệ thống phát hiện virus đạt được tỉ lệ phát hiện chính xác và thời gian thực hiện nhanh chóng, bên cạnh đó hướng tới việc sử dụng giảm thiểu tối đa tài nguyên của máy tính.

TÀI LIỆU THAM KHẢO

[1]. *Tiếp cận máy học và hệ chuyên gia để nhận dạng, phát hiện virus máy tính*. Trương Minh Nhật Quang. Luận án tiến sĩ toán học, Đại học Khoa Học Tự Nhiên, ĐHQG Tp.HCM, 2009.

[2] *Nghiên cứu một số thuật toán máy học và hệ miễn dịch nhân tạo trong phát hiện virus máy tính*. Mai Trọng Khang, Nguyễn Hoàng Ngân. Khóa luận tốt nghiệp đại học, Đại học Công Nghệ Thông Tin, ĐHQG Tp.HCM, 2013.

[3] *Tiếp cận sinh học để nhận dạng biến thể virus tin học.* Hồ Ngọc Thơ. Khoa CNTT Đại học Cần Thơ, 2005.

[4] *Immunological Computation: Theory and Applications.* Dipankar Dasgupta, Luis Fernando Niño. CRC Press, Taylor & Francis Group, 2009 .

[5] *Système Intelligent Diagnostiquer et Detruire*. Truong Minh Nhat Quang, 2008.

[6] *Các giải pháp cho phần mềm chống virus thông minh.* Nguyễn Thanh Thủy, Trương Minh Nhật Quang. Tạp chí Tin học và Điều khiển, T.13, S.3

[7] *Massachusetts Institute of Technology.* Technology Review (US-2006). http://www.technologyreview.com/read_article.aspx?id=17608&ch=infotech

[8] *F-Secure Corporation*. (Finland-2008). http://www.f-secure.com

[9] *Intel Corporation. Distributed Detection and Inference.* (US-2005). http://www.intel.com/research/distributed_detection.htm

[10] *Proofpoint.Inc*.(US-2008).

[11] *Intel Corporation. Distributed Detection and Inference*. (2005). http://www.intel.com/research/distributed_detection.htm

[12] *National Institute of Standards & Tech.* (USA-2008). http://www.nist.gov.

[13] *Artificial Immune Systems. Part I: Basic Theory and Applications*. L. N. de Castro, F. J. Von Zuben. Technical Report TR-DCA 01/99, FEEC/UNICAMP, Brazil, 1999.

[14] *A Virus Detection System Based on AIS.* Rui Chao, Ying Tan: In: Proceedings of the 2009 International Conference on Computational Intelligence & Security, vol. 1, pp. 6-10 (2009) .

[15] *Artificial immune systems: a new computational intelligence approach*, L.N. de Castro, J. Timmis , Springer, 2002.

[16] *Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification*. Su, M.Y., 2011.Journal of Network and Computer Applications, 34(2):722-730.