ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

PHẠM THỊ MAI HOA

CÁC PHƯƠNG PHÁP DỰ ĐOÁN KHẢ NĂNG ỨC CHẾ BỆNH DỰA TRÊN CÁC BIỂU DIỄN KHÁC NHAU CỦA RNA VÀ ỨNG DỤNG

Ngành: Công nghệ thông tin

Chuyên ngành: Hệ thống thông tin

Mã số: 14025126

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. Bùi Ngọc Thăng

HÀ NỘI – 2017

MỤC LỤC

DANH MỤC KÝ HIỆU VÀ CÁC CHỮ VIẾT TẮT

Từ viết tắt	Từ chuẩn	Diễn giải

ANN	Artificial Neural Network	Mạng nơ ron nhân tạo
CHS	Chalcone synthase	Gen quy định màu tím
DNA	Axit deoxyribonucleic	Axít deoxyribonucleic
dsRNA	Double-strand RNA	RNA xoắn kép
EIIP	Electron-ion interaction exon prediction	Dự đoán exon tương tác điện tử-ion
Endonuclease		enzyme phân cắt liên kết bên trong một mạch nucleic acid; chúng có thể mang tính đặc hiệu đối với một phân tử RNA, một phân tử DNA mạch đơn hay mạch kép
vivo		Cơ thể sống
vitro		Trong ống nghiệm
Interferon		Loại prôtêin do tế bào cơ thể sinh ra khi bị vírut tấn công, nhằm ngăn không cho virut phát triển
Lentivirus		Một phân họ của Retrovirus, đặc trưng của chúng là hướng tới các tế bào bạch cầu đơn nhân và đại thực bào
Ligase		Enzyme nối quan trọng trong tế bào
MiRNA	Micro RNA	Micro RNA
mRNA	Messenger RNA	RNA thông tin
Nuclease		enzyme thủy phân liên kết của phân tử nucleic acid (phân tử DNA và RNA)
PTGS	Post transcriptional gene silencing	Im lặng gen sau phiên mã
Retrovirus		Cách gọi các loại virus mà vật chất di truyền của chúng là phân tử RNA
RF	Random forest	Rừng ngẫu nhiên
RISC	RNA – incluced silencing complex	Phức hệ gây sự im lặng
RNA	Axit ribonucleic	Axit ribonucleic
ROC	Receiver operating characteristic	Đường cong đặc trưng hoạt động của bộ thu nhận

shRNA	Short hairpin RNA	
SiRNA	Short interfering RNA	RNA ngắn can thiệp
SVM	Support vector machine	Máy vecto hỗ trợ

DANH MỤC HÌNH VẼ VÀ ĐỒ THỊ

DANH MỤC BẢNG

MỞ ĐẦU

Như chúng ta đã biết, trong tế báo có nhiều loại RNA khác nhau, mỗi loại đảm nhận một chức năng sinh học riêng biệt. Kể từ khi khám phá ra RNAi thì việc nghiên cứu cơ chế và ứng dụng của nó ngày càng trở thành một vấn đề lý thú thu hút sự quan tâm của các nhà sinh học góp phần tạo nên cơn sốt "Thế giới RNA-RNA world".

Andrew Fire và Craig Mello đã tiến hành nghiên cứu về cơ chế điều khiển biểu hiện gene ở giun tròn Caenorhabditis elegans (C.elegans). Hai ông đã thực hiện hàng loạt các thí nghiệm ngoạn mục nhằm kiểm tra kiểu hình ảnh hưởng của việc tiêm RNA vào bộ phận sinh dục của C.elegans. Kết quả của quá trình nghiên cứu đã đưa ra được suy luận RNA chuỗi đôi có thể làm các gene ngừng hoạt động (bất hoạt gene). Cơ chế can thiệp RNA này mang tính đặc trưng đối với gene mang mã di truyền giống với mã di truyền của phân tử RNA được tiêm vào. Ngoài ra, cơ chế can thiệp RNA có thể lan giữa các tế bào và thậm chí được di truyền sang đời sau. Chỉ cần tiêm một lượng nhỏ phân tử RNAi cũng có thể đạt được kết quả mong muốn.

RNAi được sử dụng trong khoa học cơ bản nghiên cứu chức năng của gene. Ngoài ra, cơ chế này có ý nghĩa rất quan trọng đối với việc điều khiển các biểu hiện gene, tham gia bảo vệ cơ thể chống nhiễm virus và kiểm soát gene thay đổi đột ngột. Với nghiên cứu mới này, giới khoa học cũng đang tìm ra các ứng dụng của RNAi trong những nghiên cứu y học chữa bệnh bằng liệu pháp gene, các ứng dụng trên cây trồng, vật nuôi trong nông nghiệp nhằm tạo ra các sản phẩm với chất lượng tốt hơn; trong điều trị các bệnh nhiễm khuẩn, các bệnh do virut, bệnh tim, ung thư, rối loạn nội tiết và nhiều chứng bệnh khác. Bộ máy can thiệp RNAi bao gồm 2 thành phần siRNA và miRNA, trong đó cơ chế tắt gene bởi siRNA có hiệu quả rất cao, chỉ cần một lượng nhỏ siRNA được đưa vào tế bào cố thể đủ để làm tắt hoàn toàn sự biểu hiện của một gene nào đó (vốn có rất nhiều bản sao trong cơ thể đa bào).

Trong ngữ cảnh đó, đã có rất nhiều nghiên cứu ứng dụng học máy vào việc dự đoán khả năng ức chế bệnh của siRNA. Các nghiên cứu tập trung vào việc tìm kiếm cách thiết kế siRNA có khả năng ức chế cao, đồng thời xây dựng các mô hình dự đoán khả năng ức chế bệnh của siRNA. Các mô hình đã xây dựng bằng nhiều phương pháp tiếp cận những hầu hết còn bị hạn chế do hệ số tương quan của mô hình còn thấp. Một trong những ảnh hưởng lớn tới kết quả này là sự biểu

diễn dữ liệu siRNA, do vậy một hướng tiếp cận trong việc xây dựng mô hình dự đoán này là tìm biểu diễn siRNA nhằm đại diện được những đặc tính quan trọng nhất của siRNA mà vẫn đạt hiệu năng tính toán tốt.

Với hướng tiếp cận biểu diễn dữ liệu siRNA, nghiên cứu này khảo sát một số phương pháp xây dựng mô hình dự đoán khả năng ức chế bệnh của siRNA và tập trung vào việc biểu diễn dữ liệu siRNA theo nhiều cách khác nhau và đánh giá mô hình dự đoán được xây dựng bằng một số phương pháp như Hồi quy tuyến tính, Luật kết hợp. Kết quả thực nghiệm cho đánh giá và kết luận được phương pháp biểu diễn dữ liệu siRNA cho hiệu quả tốt nhất đã được nghiên cứu và mở ra hướng nghiên cứu tiếp là tìm cách tối ưu phương pháp học máy đã áp dụng trên biểu diễn đó để thu được hệ số tương quan tốt hơn.

Luận văn được trình bày trong 5 chương:

Chương 1: Giới thiệu về khả năng ức chế bệnh của RNA. Chương này giới thiệu tổng quan về RNA, RNAi và đi sâu vào siRNA, ý nghĩa của chúng trong nghiên cứu và thực tiễn.

Chương 2: Các hướng nghiên cứu khả năng ức chế của RNA. Chương này sẽ trình bày một số nghiên cứu tiếp cận theo hướng sinh học và tin sinh học.

Chương 3: Các cách thức biểu diễn RNA. Trình bày các cách thức biểu diễn chuỗi RNA

Chương 4: Đánh giá thực nghiệm các mô hình dự đoán khả năng ức chế của siRNA theo các biểu diễn dữ liệu khác nhau. Chương này trình bày các áp dụng cụ thể một số phương pháp dự đoán như Hồi quy tuyến tính và Luật kết hợp trên các biểu diễn khác nhau của chuỗi siRNA và đánh giá kết quả

Chương 5: Kết luận. Tổng kết lại nội dung đã nghiên cứu, đưa ra khả năng áp dụng thực tế và hướng đi tiếp theo.

Phần còn lại là các nội dung bổ sung cho luận văn và các tài liệu tham khảo đã được sử dụng cho nghiên cứu.

CHƯƠNG 1: GIỚI THIỆU VỀ KHẢ NĂNG ỨC CHẾ BỆNH CỦA RNA

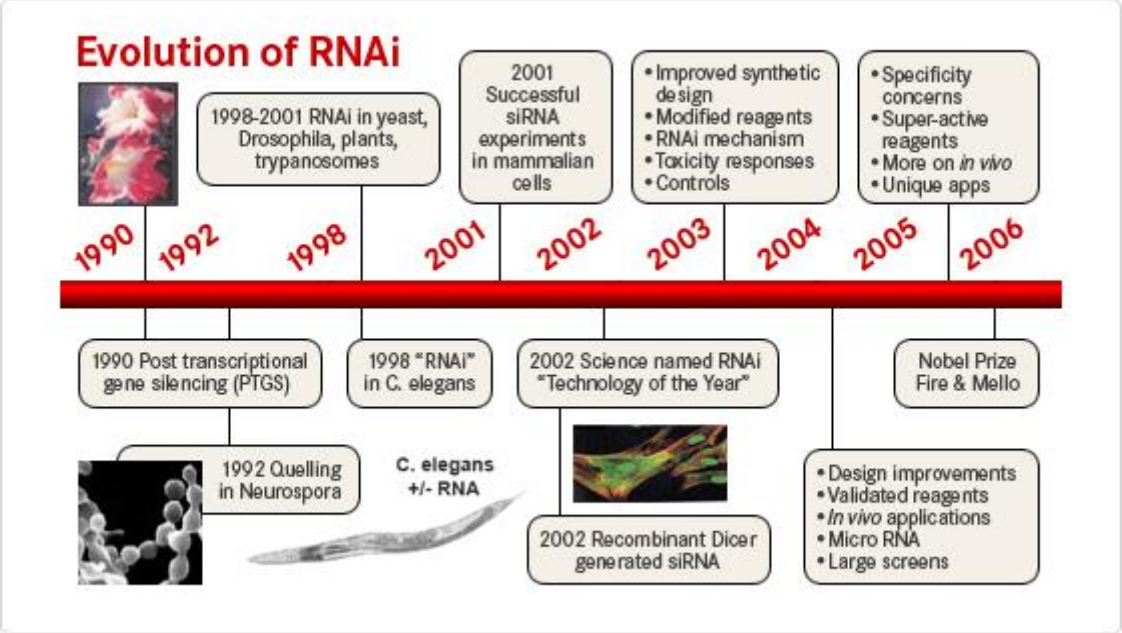
1. Tổng quan RNA can thiệp (RNAi)

1.1. Khái niệm RNAi

RNA can thiệp (RNA interference, RNAi) là một hệ thống bên trong các tế bào sống, giúp kiểm soát được các gene đang hoạt động. RNAi là một cơ chế để bất hoạt gene gây nên bởi RNA mạch kép (dsRNA). Đó là trình tự đặc biệt và liên quan đến sự suy thoái của cả hai loại phân tử RNA: RNA sợi kép (dsRNA) và RNA sợi đơn thường mRNA là những sợi tương đồng trong trình tự dsRNA làm kích hoạt phản ứng trả lời.

Các phân tử RNAi này có thể gây nên các hiệu ứng: Ức chế dịch mã đơn vị mRNA, ức chế sự phiên mã của gene ở trong nhân, phân giải mRNA.

1.2. Lịch sử nghiên cứu RNAi



Hình 1: Lịch sử nghiên cứu RNAi [1]

Trong lịch sử, sự can thiệp RNA được biết đến với những tên gọi khác như: RNA silening, quelling, cosuppresion, RNA inteference

- Năm 1984, Pesthea và các cộng sự đã nghiên cứu kỹ thuật Antiense-RNA trên vi khuẩn Escherichia Coli được đăng trên tạp chí PNAS số 81. Tuy nhiên ở giai đoạn này vẫn chưa hình dung được cơ chế gây ra sự ức chế gen.
- Đến những năm đầu thập niên 1990, một số kết quả nghiên cứu được công bố trên các tạp chí quốc tế (Napoli và cộng sự, Vander Krol và cộng sự đều vào năm 1990) dựa trên quan sát hiện tượng của hoa dạ yến thảo (pentunia) khi cố gắng tạo cánh hoa màu tím bằng cách chuyển gen quy định màu tím Chalcone synthase (CHS) dưới tác động của promoter 35S. Tuy nhiên cánh hoa lại bị đốm màu, có chỗ còn màu trắng, hiện tượng này được gọi là "đồng ức chế"
- Năm 1992, phát hiện "quelling" ở Neurospora (Neurospora crassa - vi khuẩn mốc bánh mì màu đỏ (red bread mold)). Năm 1994, Cogoni và cộng sự đã tiến hành thí nghiệm tăng màu cam của nấm Neurospora crassa, và kết quả hầu như nấm không thể hiện và hiện tượng này được gọi là "quelling".
- Năm 1995, trên tạp chí Cell số 81, nhóm nghiên cứu của Guo và Kemphues đã đưa ra bằng chứng đầu tiên trên tuyến trùng Caenorhabditis elegans rằng: Phân tử RNA chiều thuận (sense RNA) cũng gây ra sự ức chế gene tương đương với với phân tử RNA chiều ngược. Điều này gây ra sự lúng túng do kết quả khác với điều các nhà khoa học mong đợi.
- Phải đến ba năm sau 1998, nhóm nghiên cứu Fire đã giải thích được điều nghịch lý này bằng những thí nghiệm trên tuyến trùng C. elegans. Mục đích của các thí nghiệm này là nhằm kiểm tra sự hỗ trợ lẫn nhau giữa các phân tử RNA theo cả hai chiều trong quá trình ức chế sự biểu hiện của gen.
- Năm 2000, trên tạp chí Nature cũng công bố việc phát hiện hiện tượng RNAi trên loài ruồi giấm ProSophila do nhóm nghiên cứu của Richard Cathew tiến hành.
- Năm 2001, lần đầu tiên RNAi được mô tả trong các tế bào động vật có vú (Tuschl và cộng sự).
- 2002, Tạo ra tái tổ hợp dicer để tạo siRNA, công nghệ iRNA trở thành công nghệ của năm
- 2003-2005, khoảng thời gian cải tiến và tìm hiểu rõ hơn về công nghệ iRNA.
- Năm 2006, giải thưởng Nobel sinh lý và y học cho phát hiện cơ chế RNAi của hai nhà bác học Mỹ là Andrew Fire (ĐH Stanford) và Craig C. Mello (ĐH Massachusetts)

Đóng góp quan trọng nhất là việc phát hiện cơ chế RNAi từ việc nghiên cứu và thí nghiệm của Andrew Fire và C. Mello.

Ý nghĩa khoa học của công trình nghiên cứu:

- Cung cấp lời giải thích cho các hiện tượng nghiên cứu ở thực vật: Phiên mã bổ nhiệm gen im lặng (PTGS – post transcriptional gene silencing) từ đó làm sáng tỏ nhiều quan sát thí nghiệm mâu thuẫn và khó hiểu trong nhiều năm trước đây.
- Đồng thời tiết lộ một cơ chế tự nhiên để kiểm soát dòng thông tin di truyền trong tế bào
- Với nghiên cứu mới này, giới khoa học cũng đang tìm ra các ứng dụng của RNAi trong nghiên cứu y học chữa bệnh bằng liệu pháp gen, các ứng dụng trên cây trồng, vật nuôi trong nông nghiệp nhằm tạo ra các sản phẩm với chất lượng tốt hơn.
- Từ kết quả của nghiên cứu này đã mở ra nhiều hướng nghiên cứu và được tạp chí Science bình chọn là "Break Through in 1998" tức "Bước đột phá của năm 1998" dựa theo số lượng ra tăng cấp số nhân các bài báo khoa học đăng trên các tạp chí khoa học quốc tế hàng đầu.

1.3. Ý nghĩa của việc phát hiện ra RNAi

- Can thiệp RNA chống lại sự nhiễm virus
- Can thiệp RNA bảo đảm ổn định hệ gen
- Can thiệp RNA như cơ chế kiểm soát quá trình tổng hợp protein và điều khiển sự phát triển
- Can thiệp RNA như cơ chế bảo vệ nhiễm sắc tử cô đặc và tăng cường phiên mã
- Can thiệp RNA cống hiến một phương pháp mới để kiềm chế gen chuyên biệt
- Can thiệp RNA đã đề xuất một giải pháp hiệu quả trong điều trij bệnh di truyền trong tương lai

2. Cơ chế can thiệp RNAi

2.1. Các loại RNAi

Có 3 loại RNAi bao gồm: shRNA, siRNA và miRNA.

shRNA có thể dược đưa vào bởi DNA plasmid, mẫu tuyến tính hoặc vector virus hoặc vi khuẩn.

Trung tâm của quá trình can thiệp RNAi gồm 2 thành phần siRNA và miRNA và những ARN này có thể liên kết với các mRNA khác, tăng hoặc giảm hoạt động của chúng hoặc là ngăn không cho mRNA tổng hợp protein. Con đường RNAi xuất hiện ở nhiều sinh vật nhân chuẩn, bắt nguồn từ enzyme Dicer, chúng cắt các sợi dài dsRNA thành các đoạn ngắn khoảng 20 nucleotide (siRNA). Mỗi siRNA được tách thành 2 sợi đơn ssRNA, sợi hành khách và sợi hướng dẫn. Sợi hành khách bị suy thoái còn sợi hướng dẫn sẽ kết hợp vào RNA gây ra sự im lặng phức tạp (RISC). Kết quả nghiên cứu tốt nhất là sự im lặng gen sau khi phiên mã, xảy ra khi sợi hướng dẫn ghép cặp theo trình tự bổ sung với mRNA và gây ra sự phân cắt bởi Argonaute 2 (Ago2), thành phần xúc tác của phức hợp RISC.

siRNA (small interfeing RNA, short interfering RNA) là các RNA ngắn có kích thước khoảng 19 đến 25 nucleotit, được hình thành từ các RNA sợi đôi, tham gia vào quá trình tổng hợp protein, siRNA có khả năng điều khiển protein họ Argomaute tới đích điều hòa.

miRNA (micro RNA) là những đoạn RNA ngắn khoảng từ 19 đến 25 nucleotit, không tham gia vào quá trình tổng hợp protein. Tiền thân miRNA (Pre-miRNA) có cấu trúc dạng thân vòng (steen-loop) hay dạng kẹp tóc (hairpin).

2.2. Cơ chế can thiệp RNA

Khi các phần khác nhau của cơ chế RNAi đang được phát hiện, cơ chế RNAi đang trở nên ngày càng rõ ràng hơn. Trong vài năm gần đây, các nhà khoa học đã thu được những hiểu biết quan trọng trong việc làm sáng tỏ cơ chế RNAi. Sự kết hợp của các kết quả thu được từ một số thí nghiệm trên cơ thể sống (vivo) và trong ống nghiệm (vitro) đã tạo thành mô hình cơ học hai bước cho RNAi/PTGS. Bước đầu tiên, được gọi là bước khởi đầu RNAi, liên quan đến việc gắn các phân tử RNA vào một sợi kép dsRNA lớn và sự phân tách của nó thành các đoạn RNA rời rạc có kích thước xấp xỉ 21 đến 25 nucleotide (siRNA). Trong bước thứ hai, các siRNA này tham gia một phức hợp đa nuclease (enzyme thủy phân), làm giảm các mRNA đơn mạch tương đồng. Khi các phân tử mRNA này biến mất thì gen tương ứng bị bất hoạt, không có protein nào do gen đó mã hóa được tạo thành. Cơ chế can thiệp gồm 3 bước: (1) Quá trình dsRNA trở thành siRNA, (2) Khuếch đại siRNA, (3) Sự thoái hóa mRNA

2.3. Ứng dụng RNAi và thách thức

Việc phát hiện ra RNAi và cơ chế làm im lăng gen khiến các nhà khoa học không ngừng nghiên cứu và tìm cách ứng dụng RNAi vào nhiều lĩnh vực đặc biệt là khám chữa bệnh [5]

- Ứng dụng RNAi trong các bệnh liên quan đến đường uống trên cá thể sống
  - Ung thư biểu mô vòm họng
  - Ung thư đầu và cổ
  - Ung thư tế bào vảy miệng
  - Phát triển rang
- Ứng dụng RNAi trong ống nghiệm các bệnh liên quan đến đường uống trong ống nghiệm
- Ứng dụng trên cá thể sống RNAi trong các biến thể quy luật ghép
- Ứng dụng RNAi trên cá thể sống trong các bệnh hoặc chứng rối loạn thần kinh trung ương
- Ứng dụng RNAi trên cá thể sống trong bệnh viêm mãn tính và cấp tính

2.3.1. Ứng dụng của siRNA

- Sử dụng trong nghiên cứu và thử nghiệm lâm sàng
- Sử dụng để điều trị ung thư và các bệnh liên quan đến virus, các bệnh về mắt

2.3.2. Thách thức tránh các hiệu ứng không mong muốn

- Miễn dịch cơ thể: quá nhiều siRNA có thể dẫn đến các sự kiện không mong muốn do kích hoạt phản ứng miễn dịch bẩm sinh.
- Ức chế sai mục tiêu: sai mục tiêu là một thách thức nữa đối với việc sử dụng siRNAs như một công cụ bất hoạt gen

  - Đáp ứng miễn dịch thích nghi: Các chuỗi RNA có thể là các gen miễn dịch kém, nhưng kháng thể có thể dễ dàng được tạo ra đối với các phức hợp RNA-protein. Nhiều bệnh tự miễn dịch xem các loại kháng thể này.

CHƯƠNG 2: CÁC HƯỚNG NGHIÊN CỨU KHẢ NĂNG ỨC CHẾ CỦA RNA

Việc phát hiện ra RNA can thiệp đã tạo ra một trào lưu rộng lớn trong việc nghiên cứu, thử nghiệm và ứng dụng RNAi không chỉ để tạo sự hiểu biết sâu hơn mà còn mở ra những bước tiến trong việc điều trị bệnh và ngành nuôi trồng. Việc nghiên cứu RNA còn gặp nhiều thách thức, và một trong số đó là tìm ra những RNAi có khả năng ức chế cao mà không gây ra những phản ứng phụ như ức chế sai mục tiêu hay miễn dịch. Các nhà khoa học trên thế giới vẫn không ngừng nghiên cứu về khả năng ức chế của RNA, chủ yếu đi theo hai hướng tiếp cân: (1) Hướng tiếp cận sinh học và (2) Hướng tiếp cận tin sinh học. Cũng có những khoa học có thể nghiên cứu theo cả hai hướng tiếp cận này đã đưa ra được những kết quả vô cùng gia trị cho ngành nghiên cứu này.

1. Hướng nghiên cứu sinh học

Nghiên cứu của Angela Reynolds và công sự nhằm đưa ra một thiết kế hợp lý để lựa chọn được các siRNA tiềm năng [6]. Để xác định tính năng của siRNA đặc hiệu, nhóm nghiên cứu đã thực hiện một phân tích có hệ thống của 180 siRNA nhằm mục tiêu mRNA của hai gen. Tám đặc điểm liên quan đến chức năng siRNA được xác định: hàm lượng G/C thấp, sự thiên vị với nội bộ bên trong sự bền vững ở sợi ý nghĩa đầu 3’, thiếu các lặp đảo ngược.

Một số nhà nghiên cứu sinh học khác cũng thực hiện băng phương pháp thí nghiệm và quan sát như Tuschl, Amarzguioui, Stockholm, Ui-Tei, Hseih mục đích nhằm tìm ra những mẫu siRNA có hiệu quả ức chế cao nhất và tránh được các tác dụng không mong muốn như ức chế sai mục tiêu.

2. Hướng nghiên cứu sinh học kết hợp tin sinh học

Huesken nghiên cứu theo hướng lai sinh học và tin học, ông sử dụng phương pháp mạng neuron để xây dựng mô hình dự báo từ dữ liệu thực tế, và sinh ra được dữ liệu nhân tạo và sử dụng dữ liệu sinh học để kiểm thử. Bộ dữ liệu Huesken của ông bào gồm 2431 chuỗi siRNA hiện đang sử dụng rộng rãi trong các bài toán xây dựng mô hình dự đoán.

3. Hướng nghiên cứu tin sinh học

Sử dụng các phương pháp học máy để xây dựng mô hình dự đoán, đa số sử dụng bộ dataset là tập dữ liệu Huesken đã công bố. Một số nhà nghiên cứu trong danh sách này như: Shibalina với phương pháp hồi quy tuyến tính, Vert và cộng sự với phương pháp hồi quy Laso, Ichihara và cộng sự với phương pháp MKSVR, Qui và công jswj sử dụng phương pháp Assembel learning, Sciablola và cộng sử sử dụng SVR và Bùi Ngọc Thăng sử dụng dụng Tensor regression.

CHƯƠNG 3: CÁC CÁCH THỨC BIỂU DIỄN RNA

Như đã trình bày ở chương trước, việc biểu diễn dữ liệu ảnh hưởng lớn tới kết quả xây dựng mô hình. RNA là một chuỗi các nucleotide gồm 4 loại: Adenin (A), Guanin (G), Uraxin (U), Cytozin (C). Các cách thức biểu diễn RNA được trình bày trong chương này xuất phát từ trình tự của các nucleotide A, C, G, U (nguyên tắc bổ sung A-U, G-C).

1. Biểu diễn theo tần số xuất hiện của các bộ 1-merge, 2-merge, 3-merge

- Các định nghĩa:
  - 1-merge: bộ gồm duy nhất 1 nucleotide
  - 2-merge: bộ gồm 2 nucleotide đứng cạnh nhau có phân biệt thứ thự
  - 3-merge: bộ gồm 3 nucleotide đứng cạnh nhau có phân biệt thứ tự
- Như vậy theo định nghĩa trên với 4 loại nucleotide ta sẽ có:
  - 4 bộ 1-merge phân biệt với nhau
  - 16 (tương đương với 4^2) bộ 2-merge phân biệt với nhau
  - 64 (tương đương với 4^3) bộ 3-merge phân biệt với nhau
- Bộ dữ liệu ban đầu để xây dựng biểu diễn gồm một tập các RNA có độ dài bằng nhau (n nucleotide) được chia thành 4 tập con:
  - Low: tập các chuỗi siRNA có khả năng ức chế thấp ký hiệu là S_1
  - Medium: tập các chuỗi siRNA có khả năng ức chế trung bình ký hiệu là S_2
  - High: tập các chuỗi siRNA có khả năng ức chế cao ký hiệu là S_3
  - tập các chuỗi siRNA có khả năng ức chế rất cao ký hiệu là S_4

Việc biểu diễn dữ liệu RNA được thực hiện như sau:

- Thống kê số lần xuất hiện của từng bộ 1-merge, 2-merge, 3-merge:
  - Thống kê số lần xuất hiện của mỗi bộ 1-merge trong mỗi tập S_1, S_2, S_3, S_4 lần lượt là x, y, z, t

- Thống kê số lần xuất hiện của mỗi bộ 2-merge trong mỗi tập S_1, S_2, S_3, S_4 lần lượt là x', y', z', t'
- Thống kê số lần xuất hiện của mỗi bộ 3-merge trong mỗi tập S_1, S_2, S_3, S_4 lần lượt là x'', y'', z'', t''

- Với mỗi chuỗi RNA, ta biểu diễn tần số của từng bộ 1-merge, 2-merge, 3-merge có mặt trong chuỗi RNA như sau:
  - Với chuỗi RNA có chiều dài n, sẽ có n bộ 1-merge xuất hiện ở các vị trí từ 1 cho tới n (có thể có giá trị trùng nhau). Tại mỗi vị trí của chuỗi RNA sẽ có 1 bộ 1-merge có số lần xuất hiện trong các tập S_1, S_2, S_3, S_4 lần lượt là x, y, z, t. Khi đó tại mỗi vị trí, biểu diễn dữ liệu sẽ là 4 giá trị tần số xuất hiện của bộ 1-merge đó trong các tập S_1, S_2, S_3, S_4 tức $\frac{x}{x+y+z+t}, \frac{y}{x+y+x+t}, \frac{z}{x+y+z+t}, \frac{t}{x+y+z+t}$
    Như vậy n vị trí sẽ biểu diễn thành 4n giá trị tần số của các bộ 1-merge.
  - Với chuỗi RNA có chiều dài n, sẽ có n-1 bộ 2-merge xuất hiện ở các vị trí từ 1 cho tới n-1. Tương tự như cách biểu diễn bộ 1-merge, tại mỗi vị trí trong chuỗi RNA (trừ vị trí cuối cùng) sẽ tồn tại 1 bộ 2-merge có số lần xuất hiện trong các tập S_1, S_2, S_3, S_4 lần lượt là x', y', z', t'. Tại mỗi vị trí sẽ biểu diễn dữ liệu bằng 4 giá trị tần số $\frac{x'}{x'+y'+z'+t'}, \frac{y'}{x'+y'+x'+t'}, \frac{z'}{x'+y'+z'+t'}, \frac{t'}{x'+y'+z'+t'}$
    Như vậy n vị trí sẽ biểu diễn được 4(n-1) giá trị tần số của các bộ 2-merge
  - Với chuỗi RNA có chiều dài n, sẽ có n-2 bộ 3-merge xuất hiện ở các vị trí từ 1 cho tới n-2. Tương tự tại mỗi vị trí trong chuỗi RNA (trừ vị trí cuối cùng) sẽ tồn tại 1 bộ 3-merge có số lần xuất hiện trong các tập S_1, S_2, S_3, S_4 lần lượt là x'', y'', z'', t''. Tại mỗi vị trí sẽ biểu diễn dữ liệu bằng 4 giá trị tần số $\frac{x''}{x''+y''+z''+t''}, \frac{y''}{x''+y''+x''+t''}, \frac{z''}{x''+y''+z''+t''}, \frac{t''}{x''+y''+z''+t''}$
    Như vậy n vị trí sẽ biểu diễn được 4(n-2) giá trị tần số của các bộ 3-merge
- Tổng kết, chuỗi RNA có chiều dài n sẽ được biểu diễn thành 1 vecto có số chiều 4n + 4(n-1) + 4(n-2). Trong đó 4n chiều đầu tiên biểu diễn tần số của các bộ 1-merge, 4(n-1) chiều tiếp theo biểu diễn tần số của các bộ 2-merge, 4(n-2) chiều cuối cùng biểu diễn tần số của các bộ 3-merge

2. Biểu diễn theo tần số của một bộ các nucleotide có tính thứ tự

- Cách biểu diễn này giống với các biểu diễn tần số đã trình bày ở mục trước Biểu diễn theo tần số xuất hiện của các bộ 1-merge, 2-merge, 3-merge
- Nếu bộ nucleotide và thứ tự không xuất hiện trong chuỗi siRNA thì nó sẽ biểu diễn bằng giá trị (0,0,0,0)
- Điểm khác, biểu diễn này không giới hạn chỉ bộ 1-merge, 2-merge, 3-merge mà có thể là một bộ gồm k nucleotide được chọn ra và có phân biệt thứ tự.
- Số lượng bộ k-nucleotide tùy thuộc vào thuật toán lựa chọn.

3. Biểu diễn thành số tương ứng với loại nucleotide và vị trí

- Quy đổi các loại nucleotide thành các giá trị: A = 0, C = 1, G = 2, U = 3
- Với một chuỗi RNA có độ dài n sẽ được biểu diễn bằng vector n chiều tương ứng với mỗi vị trí nucleotide trong chuỗi RNA. Tại vị trí i (1, 2, …, n) trong vector n chiều:
  - Nếu A xuất hiện tại vị trí i trong chuỗi RNA thì giá trị tại chiều thứ i là 4i.
  - Nếu C xuất hiện tại vị trí i trong chuỗi RNA thì giá trị tại chiều thứ i là (4i+1)
  - Nếu G xuất hiện tại vị trí i trong chuỗi RNA thì giá trị tại chiều thứ i là (4i+2)
  - Nếu U xuất hiện tại vị trí i trong chuỗi RNA thì giá trị tại chiều thứ i là (4i+3)

4. Phương pháp biểu diễn chuỗi DNA không suy thoái

Phương pháp biểu diễn này, nếu có một chuỗi DNA có độ dài n. Tại các vị trí i trên chuỗi (i=1, 2, .., n) ta dễ dàng tính được a, g, c, t. Mỗi vị trí của chuỗi DNA sẽ được ánh xạ thành một điểm tương ứng với một cặp giá trị (x, y) trong đó theo công thức:

$$a\left(\frac{1}{2},-\frac{\sqrt{3}}{2}\right)+g\left(\frac{\sqrt{3}}{2},-\frac{1}{2}\right)+c\left(\frac{\sqrt{3}}{2},\frac{1}{2}\right)+t\left(\frac{1}{2},\frac{\sqrt{3}}{2}\right)=(x,y)$$

Như vậy một chuỗi DNA có độ dài n sẽ được biểu diễn bằng n điểm với 2 tọa độ (x, y) và không tạo thành mạch (biểu diễn đồ họa). Ta biểu thị số cho đồ họa đó bằng một vector 2n chiều chứa biểu diễn liên tiếp 2 tọa độ (x,y) của n điểm trong chuỗi DNA để thu được biểu diễn số học cuối cùng.

Ngoài các cách biểu diễn trên, một loạt các biểu diễn số học chuỗi DNA đã được tổng kết lại trong tài liệu [7] trong các phần tiếp theo bao gồm 11 cách biểu diễn: VOSS, TETRAHEDRON, INTEGER, REAL, COMPLEX, QUATERNION, EIIP, ATOMIC NUMBER, PAIRED NUMERIC, DNA WALK, Z-CURVE. Cách biểu diễn này khi áp dụng RNA thì sẽ thay thế uraxin (U) cho Thymine (T). Các cách biểu diễn này được chia thành hai nhóm. Nhóm 1 Fixed mapping (Ánh xạ cố định) các ribonucletide trong dữ liệu DNA được chuyển đổi thành một loạt các chuỗi số tùy ý. Ánh xạ cố định bao gồm các phương pháp VOSS, TETRAHEDRON, INTEGER, REAL, COMPLEX. Nhóm 2 Physico Chemical Property Based Mapping (Ánh xạ dựa trên cơ sở các thuộc tính vật lý hóa học), trong đó các thuộc tính sinh lý và sinh hóa của các phân tử sinh học DNA được sử dụng cho việc ánh xạ chuỗi DNA, khá mạnh và thường được sử dụng để tìm kiếm các nguyên lý sinh học và các cấu trúc trong phân tử sinh học. Các phương pháp ánh xạ thuộc nhóm 2 bao gồm các phương pháp biểu diễn EIIP, ATOMIC NUMBER, PAIRED NUMERIC, DNA WALK, Z-CURVE.

Phương pháp	Biểu diễn	S(n) = [CGAT]	Số chuỗi chỉ thị
VOSS	$X_n = 1$ với S(n) = X $X_n = 1$ với S(n) ≠ X X_n áp dụng cho mỗi C_n, G_n, A_n, T_n	C_n = [1,0,0,0] G_n = [0,1,0,0] A_n = [0,0,1,0] T_n = [0,0,0,1]	4
TETRA HEDRO N	$X_r(n) = \frac{\sqrt{2}}{3}[2T_n - C_n - G_n]$ $X_g(n) = \frac{\sqrt{6}}{3}[C_n - G_n]$ $X_b(n) = \frac{1}{3}[3A_n - T_n - C_n - G_n]$	$X_r(n) = \frac{\sqrt{2}}{3}[-1,-1,0,2]$ $X_g(n) = \frac{\sqrt{6}}{3}[1,-1,0,0]$ $X_b(n) = \frac{1}{3}[-1,-1,3,-1]$	3
INTEGE R	A = 2, C = 1, G = 3, T = 0	[1, 3, 2, 0]	1
REAL	A = -1.5, C = 0.5, G = -0.5, T= 1.5	[0.5, -0.5, -1.5, 1.5]	1

COMPLEX	A = 1+j, C = -1+j, G = -1-j, T = 1-j	[-1+j, -1-j, 1+j, 1-j]	1,4
QUATERNION	A = i+j+k, C = i-j-k, G = -i-j+k, T = -i+j-k	[i-j-k, -i-j+k, i+j+k, -i+j-k]	1,4
EIIP	A = 0.1260, C = 0.1340, G = 0.0806, T = 0.1335	[0.1340, 0.0806, 0.1260, 0.1335]	1,4
ATOMIC NUMBER	A = 70, C = 58, G = 78, T = 66	[58, 78, 70, 66]	1,4
PAIRED NUMERIC	A hoặc T = 1, C hoặc G = -1	P_{1n} = [-1, -1, 1, 1]	1
		P_{2n} = [-1, -1, 0, 0] & [0, 0, 1, 1]	2
DNA WALK	C hoặc T = 1, A hoặc G = -1	[1, 0, -1, 0]	1
Z-CURVE	$x_n = (A_n + G_n) - (C_n + T_n) \equiv R_n - Y_n$ $y_n = (A_n + C_n) - (G_n + T_n) \equiv M_n - K_n$ $z_n = (A_n + T_n) - (C_n + G_n) \equiv W_n - S_n$	x = [-1, 0, 1, 0] y = [1, 0, 1, 0] z = [-1,-2,-1, 0]	3

Bảng 1: Tóm tắt các phương pháp biểu diễn số học cho chuỗi DNA

CHƯƠNG 4: ĐÁNH GIÁ THỰC NGHIỆM CÁC MÔ HÌNH DỰ ĐOÁN KHẢ NĂNG ỨC CHẾ CỦA SIRNA THEO CÁC BIỂU DIỄN DỮ LIỆU KHÁC NHAU

Sau khi đã khảo sát một số phương pháp xây dựng mô hình dự đoán khả năng ức chế của RNA và các phương pháp biểu diễn chuỗi DNA và RNA. Chương này báo cáo lại quá trình thực nghiệm và đánh giá một số mô hình dự đoán khả năng ức chế của siRNA theo một số cách biểu diễn dữ liệu đã trình bày ở chương 3. Các phương pháp xây dựng mô hình dự đoán bao gồm: Quy hồi tuyến tính, Phân lớp (Naïve Bayes) và Kết hợp (thuật toán Apriori).

Phần thực nghiệm sử dụng dữ liệu dataset bao gồm 2 loại: Scored Dataset và Label Dataset. Scored Dataset bao gồm: Huesken19_train (2182 siRNA), Huesken19_test (249 siRNA), Vicker (76 siRNA), Isis (67 siRNA), Uitei (81 siRNA), Sloan (601 siRNA), Reynolds (244 siRNA), Ncbi (653 siRNA). Labeled Dataset gồm file dữ liệu siRecords (1261 siRNA nhãn "Low", 1253 siRNA nhãn "Medium", 2459 siRNA nhãn "High", 2470 siRNA nhãn "Very High" trong tổng 7443 siRNA được gán nhãn về khả năng ức chế bệnh).

Để xây dựng mô hình dự đoán, Weka 3.8 được sử dụng để thực hiện các giải thuật học máy cần thiết khi nạp dữ liệu đầu vào là biểu diễn dữ liệu đã được tính toán và thể hiện lại trong file arff. Các file arff là kết quả thực hiện chạy các thuật toán biểu diễn dữ liệu đã trình bày ở chương 3 và ghi lại ra file theo định dạng arff – là định dạng phần mềm Weka hỗ trợ.

Phương pháp đánh giá mô hình: sử dụng Cross-Validation 10-Folds.

Môi trường thử nghiệm: Máy tính cá nhân Dell 64 bit, 8G Ram, Core i5-6200U, tốc độ 2.3 GHz.

1. Thực nghiệm thuật toán kết hợp Apriori

Trong phần thực nghiệm này, dữ liệu để xây dựng mô hình được lấy từ bộ dữ liệu Labeled Datasets bao gồm các chuỗi siRNA có độ dài 19 nucleotide được gán nhãn Low và Very High về khả năng ức chế bệnh.

Các chuỗi siRNA từ tập dữ liệu là trình tự sắp xếp của 19 nucleotide (A, C, G, U). Nguyên tắc bổ sung của RNA là A-U và G-C.

Sử dụng phương pháp biểu diễn dữ liệu số 3 (Biểu diễn thành số tương ứng với loại nucleotide và vị trí). Khi đó mỗi chuỗi siRNA sẽ được biểu diễn thành vector 20 chiều. Chiều thứ nhất là thuộc tính nhãn lấy từ file siRecords của chuỗi siRNA là một trong bốn giá trị trị {“Low”, “Medium”, “High”, “Very High”}. 19 chiều tiếp theo được biểu diễn bởi một số nguyên không âm chính là vector biểu diễn RNA theo phương pháp số 3.

Thực hiện phương pháp biểu diễn dữ liệu trên với 4 tập riêng biệt {“Low”, “Medium”, “High”, “Very High”} để thu được 4 file arff cho mỗi tập và chạy thuật toán Apriori (Kết hợp) bằng weka 3.8 với cấu hình Apriori -N 20 -T 0 -C 0.9 -D 0.05 -U 0.01 -M 0.01 -S -1.0 -c -1.

Kết quả trên mỗi tập “Low”, “High”, “Medium”, “Very High” ta thu được 20 luật kết hợp, và tổng ta có 80 luật kết hợp trên cả 4 tập. Chi tiết 80 rules kết hợp xin tham chiếu phần Danh mục bổ sung, mỗi luật thể hiện luật kết hợp giữa vài nucleotide và vị trí xuất hiện của nó tại vị trí nào đó với khả năng ức chế bệnh.

Ngoài ra, để nâng cao độ tin cậy, thực hiện lọc những luật có tần số lớn hơn 30%, tức là những luật đã được tìm thấy ở một tập ví dụ “Low” thì nó phải có tần số xuất hiện >= 30% tổng số lần xuất hiện luật đó trên cả bốn tập “Low”, “Medium”, “High”, “Very High”. Sau khi thực hiện lọc với tỉ lệ 30%, số lượng luật kết hợp đã giảm từ 80 xuống còn 30 luật kết hợp. Chi tiết xem Danh mục bổ sung.

Đánh giá chung: Sau khi lọc với tỉ lệ 30% thì số luật giảm đáng kể, thể hiện độ chính xác của thuật toán chưa cao. Cách biểu diễn số 3 chưa thể hiện được mức độ liên kết giữa các nucleotide với khả năng ức chế bệnh của chuỗi siRNA.

2. Thực nghiệm thuật toán Phân lớp Naïve Bayes

Trong phần thực nghiệm này, dữ liệu để xây dựng mô hình được lấy từ bộ dữ liệu Labeled Datasets bao gồm các chuỗi siRNA có độ dài 19 nucleotide được gán nhãn Low và Very High về khả năng ức chế bệnh.

Biểu diễn VOSS

Thực hiện biểu diễn dữ liệu theo phương pháp VOSS kết hợp với thuộc tính nhãn. Khi đó mỗi chuỗi siRNA sẽ được biểu diễn bởi một vector có số chiều là 77. Chiều thứ nhất là nhãn của siRNA (“Low”, “Very High”). 76 thuộc tính tiếp

theo là biểu diễn dạng binary là các số 0,1 theo biểu diễn VOSS. Dữ liệu đã sinh ra được ghi vào một file arff để chạy thuật toán.

Chạy thuật toán Phân lớp Naïve Bayes của Weka 3.8 với tập dữ liệu đã biểu diễn để xây dựng mô hình phân lớp với thuộc tính nhãn (thuộc tính thứ nhất) là mục tiêu cho kết quả như sau: tỉ lệ phân lớp đúng đạt 65.4784% và tỉ lệ phân lớp sai là 34.5214%.

Biểu diễn DNA không suy thoái

Thực hiện biểu diễn dữ liệu theo phương pháp biểu diễn DNA không suy thoái kết hợp với thuộc tính nhãn. Khi đó mỗi chuỗi siRNA sẽ được biểu diễn bởi một vector có số chiều là 39. Chiều thứ nhất là nhãn của siRNA ("Low", "Very High"). 38 thuộc tính tiếp theo là biểu diễn dạng tọa độ (x,y) tương ứng với các vị trí từ 1 đến vị trí 19 trên chuỗi RNA. Dữ liệu đã sinh ra được ghi vào một file arff để chạy thuật toán.

Chạy thuật toán Phân lớp Naïve Bayes của Weka 3.8 với tập dữ liệu đã biểu diễn để xây dựng mô hình phân lớp với thuộc tính nhãn (thuộc tính thứ nhất) là mục tiêu cho kết quả như sau: tỉ lệ phân lớp đúng đạt 56.2252 % và tỉ lệ phân lớp sai là 43.7748 %.

3. Thực nghiệm thuật toán Phân lớp Hồi quy tuyến tính

Trong quá trình thực nghiệm cũng kết hợp một số phương pháp biểu diễn với nhau và so sánh kết quả hệ số tương quan được thể hiện tổng hợp trong bảng đầy đủ sau:

	Data				
	Huesken19_train	**Huesken19_test**	**Reynolds**	**Utei**	**Vicker**
1-merge	0.5991	N/A	N/A	N/A	N/A
2-merge	0.4767	N/A	N/A	N/A	N/A
3-merge	0.3191	N/A	N/A	N/A	N/A

rules80	0.2482	0.214	0.0695	0.2548	0.1529
rules38	0.1626	0.115	0.1043	0.1219	0.1103
1-merge + 2-merge	0.5985	N/A	N/A	N/A	N/A
1-merge + 3-merge	0.5903	N/A	N/A	N/A	N/A
1-merge + rules80	0.5872	N/A	N/A	N/A	N/A
1-merge + rules38	0.5928	N/A	N/A	N/A	N/A
2-merge + 3-merge	0.4684	N/A	N/A	N/A	N/A
1-merge + 2-merge + 3-merge	0.588	0.6137	0.5225	0.6641	0.5147
1-merge + 2-merge + 3-merge + rules38	0.5772	0.6097	0.5262	0.6455	0.4843
1-merge + 2-merge + 3-merge + rules80	0.5792	0.5986	0.5091	0.6603	0.4573
2-merge + 3-merge + rules38	0.4583	0.4876	0.3694	0.5052	0.3665
2-merge + 3-merge + rules80	0.4645	0.5133	0.3252	0.5208	0.329
VOSS + 1-merge + 2-merge + 3-merge	0.5874	0.6145	0.5329	0.666	0.5063
VOSS + 1-merge	0.6032	0.6238	0.5397	0.6428	0.5757
VOSS + 2-merge	0.5968	0.6244	0.5224	0.665	0.547

VOSS + 3-merge	0.5935	0.6069	0.5337	0.6433	0.5807
VOSS + 2-merge + 3-merge	0.5838	0.6168	0.5486	0.6772	0.515
Biểu diễn số học - VOSS	0.6024	0.6187	0.5394	0.6326	0.5668
Biểu diễn không suy thoái Yau	0.6031	N/A	0.5377	0.6205	0.588
Biểu diễn số học - TetraHedron	0.6047	0.6218	0.5471	0.6355	0.5681
Biểu diễn số học - Integer	0.3663	0.451	0.2993	0.2101	0.381
Biểu diễn số học - Real	0.218	0.2514	0.2036	0.0219	0.0846
Biểu diễn số học - EIIP	0.3277	0.405	0.2414	0.2569	0.2958
Biểu diễn số học - Atomic	0.1427	0.1125	0.127	0.1659	0.1081
Biểu diễn số học - DNA Walker	0.341	0.3003	0.3448	0.4688	0.2594

Bảng 2: Tổng hợp kết quả thực nghiệm phương pháp Hồi quy tuyến tính với các cách biểu diễn siRNA khác nhau

4. Đánh giá kết quả thực nghiệm

Nhìn chung kết quả của các mô hình còn thấp với hệ số tương quan < 0.65. Kết quả như vậy vì so với các mô hình hiện tại, chưa có sự cải tiến về mặt phương pháp xây dựng mô hình, mà chú trọng việc biểu diễn dữ liệu. Tuy nhiên đối với những biểu diễn dữ liệu dạng số học với số chiều khá thấp (39 chiều hoặc 77 chiều) nên chưa thể hiện được sự tương quan của chuỗi siRNA với score mục tiêu gây ra kết quả rất thấp. Một số biểu diễn có kết quả gần như ngang bằng với những các mô hình hiện tại như biểu diễn theo tần số của các bộ 1-merge, 2-merge, 3-merge hoặc biểu diễn VOSS, TETRAHEDRON, biểu diễn DNA không suy thoái do có số chiều tương đối lớn và cách biểu diễn có đề cập đến vị trí tương quan của các nucleotide trong chuỗi. Tuy nhiên sự tương quan được biểu diễn chưa đủ tốt để đạt kết quả xây dựng mô hình như mong đợi do chưa tối ưu được mô hình dự đoán.

KẾT LUẬN

Kết quả từ quá trình thực nghiệm cho thấy việc kết hợp các phương pháp xây dựng mô hình dự đoán và các phương pháp biểu diễn hiện có chưa đem lại kết quả mong đợi. Có nhiều nguyên nhân để dẫn tới kết quả đó như dữ liệu để thực nghiệm chưa đủ lớn để đem lại kết quả chính xác. Dữ liệu để thực nghiệm được lấy từ kết quả của công trình nghiên cứu của một số nhà khoa học hiện có một số ý kiến trái chiều với nhau nên kết quả test với mô hình đã xây dựng từ dữ liệu training không thực sự cao. Ngoài ra kết quả thực nghiệm chỉ ngang bằng với hiện tại do chưa có sự tối ưu mô hình dự đoán trong quá trình thực nghiệm. Và nguyên nhân chính là do các phương pháp biểu diễn đã được trình bày và thực nghiệm còn bộc lộ nhiều thiếu xót như số chiều chưa đủ lớn, thiếu các cấu trúc dữ liệu bậc 1, 2, 3 và chưa đủ tính đai diễn cho số lượng siRNA vô cùng lớn 4^{19}.

Từ những vấn đề còn tồn tại trong quá trình làm luận văn, và kết quả thực nghiệm, nghiên cứu này có thể tiếp tục để giải quyết một khía canh đã gặp phải đó là tối ưu mô hình dự đoán. Phương pháp được đề xuất để tối ưu mô hình dự đoán đó là phải tối ưu ma trận F (ma trận chuyển đổi) bằng phương pháp Lagrange sao cho sai số bình phương tối thiểu đạt mức nhỏ nhất. Việc tối ưu ma trận F được trông đợi sẽ đem lại mô hình dự đoán có độ tương quan đủ tốt đối với khả năng ức chế bệnh của siRNA.