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

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

HÀ NỘI - 2017

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: 8480104

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. Bùi Ngọc Thăng

HÀ NỘI - 2017

LỜI CAM ĐOAN

Tôi là Phạm Thị Mai Hoa, học viên khóa K21, ngành Công nghệ thông tin, chuyên ngành Hệ Thống Thông Tin. Tôi xin cam đoan luận văn "Các phương pháp dự đoán khả năng ức chế bệnh dựa trên các biểu diễn khác nhau của RNA và ứng dụng" là do tôi nghiên cứu, tìm hiểu và phát triển dưới sự hướng dẫn của TS. Bùi Ngọc Thăng. Luận văn không phải sự sao chép từ các tài liệu, công trình nghiên cứu của người khác mà không ghi rõ trong tài liệu tham khảo. Tôi xin chịu trách nhiệm về lời cam đoan này.

Hà Nội, ngày tháng năm 2017

LỜI CẢM ƠN

Đầu tiên tôi xin gửi lời cảm ơn tới các thầy cô Trường Đại học Công nghệ, Đại học Quốc Gia Hà Nội đã tận tình giảng dạy và truyền đạt kiến thức trong suốt thời gian tôi học tập và nghiên cứu tại trường. Tôi cũng xin được gửi lời cảm ơn đến các thầy cô trong Bộ môn Hệ thống thông tin cũng như Khoa công nghệ thông tin đã mang lại cho tôi những kiến thức vô cùng quý giá và bổ ích trong quá trình học tập tại trường.

Đặc biệt xin chân thành cảm ơn thầy giáo, TS. Bùi Ngọc Thăng, người đã định hướng, giúp đỡ, trực tiếp hướng dẫn và tận tình chỉ bảo tôi trong suốt quá trình nghiên cứu, xây dựng và hoàn thiện luận văn này.

Tôi cũng xin được cảm ơn tới gia đình, những người thân, các đồng nghiệp và bạn bè đã thường xuyên quan tâm, động viên, chia sẻ kinh nghiệm, cung cấp các tài liệu hữu ích trong thời gian học tập, nghiên cứu cũng như trong suốt quá trình thực hiện luận văn tốt nghiệp.

Hà Nội, ngày tháng năm 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
Antisense ODNs	Antisense oligonucleotides	
ATP	Adenosine triphosphate	Phân tử năng lượng
CHS	Chalcone synthase	Gen quy định màu tím
Codon		Bộ ba các ribo-nucleotide có gốc là nucleobase đối ứng với các nucleobase của nucleotide trong triplet đối ứng gốc
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
Helicase		Enzyme helicase (còn có tên là enzyme deroulase) có nhiệm vụ giúp chuỗi DNA từ dạng siêu xoắn sang dạng dãn thành hai sợi đơn
Heuristic		Các kỹ thuật dựa trên kinh nghiệm để giải quyết vấn đề, học hỏi hay khám phá nhằm đưa ra một giải pháp mà không được đảm bảo là tối ưu
Interferon		Loại prôtêin do tế bào cơ thể sinh ra khi bị virut 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

Luciferase		Enzyme phát sáng 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)
Ovo		In ovo có nghĩa trong trứng
PCR	Polymerase Chain Reaction	Phản ứng chuỗi polymerase, cũng có sách gọi là "phản ứng khuếch đại gen"
PTGS	Post transcriptional gene silencing	Im lặng gen sau phiên mã
Renilla luc	Renilla luciferase	Protein ở cây ngải biển (Renilla reniformis)
Reporter gene		Gen chỉ thị
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 can thiệp ngắn
SVM	Support vector machine	Máy vecto hỗ trợ
Triplet		Các bộ ba nucleotide trong mỗi mạch đơn của chuỗi xoắn kép ADN khi giản phân, là một tổ hợp của 3 trong bốn loại nucleotide này
UTR	Untranslated region	Vùng không dịch mã
vivo		Cơ thể sống
vitro		Trong ống nghiệm

DANH MỤC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

Bộ máy di truyền ở cá thể sống là một cơ chế kỳ diệu mà con người luôn mong muốn khám phá, tìm ra cơ chế hoạt động mà tự nhiên đã ban tặng cho mỗi loài. Việc nghiên cứu liên quan tới thông tin di truyền không chỉ mang lại hiểu biết cho con người mà còn để ứng dụng vào nhiều lĩnh vực quan trọng, đặc biệt là lĩnh vực y học, sinh học. Mã di truyền trên DNA quy định protein được hình thành. Thông tin di truyền lưu trữ trong DNA được sao chép sang RNA và sau đó được dùng để tổng hợp protein. Dòng thông tin được truyền từ DNA qua mRNA đến protein được gọi là "Học thuyết trung tâm" của lĩnh vực sinh học phân tử. Cơ chế kiểm soát của bộ máy sao chép DNA sang mRNA trong quá trình phiên mã quyết định gen nào được biểu hiện. Quá trình phiên mã cũng bị điều khiển bởi nhiều nhân tố khác và được con người nghiên cứu, tìm hiểu ngày càng rõ.

Như chúng ta đã biết, trong tế bào có nhiều loại RNA khác nhau, mỗi loại đảm nhận một chức năng sinh học riêng biệt. Một số chức năng quan trọng của RNA: 1. Chức năng vận chuyển thông tin (mRNA); 2. Chức năng tham gia tổng hợp và vận chuyển protein (tRNA và rRNA); 3. Chức năng hoàn thiện các phân tử RNA. Hơn nữa, bằng những quan sát của sự ức chế phiên mã nhờ biểu hiện RNA đối khuôn trong thực vật chuyển gen được thực hiện bởi các nhà thực vật học ở Mỹ và Hà Lan trong những năm đầu của thập kỷ 1990, con người đã phát chức năng điều hòa biểu hiện gen của RNA hay còn gọi là can thiệp RNA (RNAi).

Andrew Fire và Craig Mello đã tiến hành nghiên cứu về cơ chế điều khiển biểu hiện gen 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 gen ngừng hoạt động (bất hoạt gen). Cơ chế can thiệp RNA này mang tính đặc trưng đối với gen 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 gen. Ngoài ra, cơ chế này có ý nghĩa rất quan trọng đối với việc điều khiển các biểu hiện gen, tham gia bảo vệ cơ thể chống nhiễm virus và kiểm soát gen 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 gen, 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 gen 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 gen 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 bệnh 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, các cách biểu diễn dữ liệu siRNA theo nhiều cách khác nhau và phần thực nghiệm tập trung vào việc biểu diễn siRNA khác nhau bằng các chương trình Java và ghi lại biểu diễn ra file, và đánh giá các phương pháp biểu diễn siRNA trong một số mô hình dự đoán bằng phương pháp như Hồi quy tuyến tính, Luật kết hợp bằng phần mềm Weka 3.8. Kết quả thực nghiệm mang lại đánh giá và so sánh giữa các phương pháp biểu diễn dữ liệu siRNA khác nhau cho hiệu quả khác nhau, từ đó 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 bệnh 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 bệnh 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

Phần Kết luận sẽ 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)

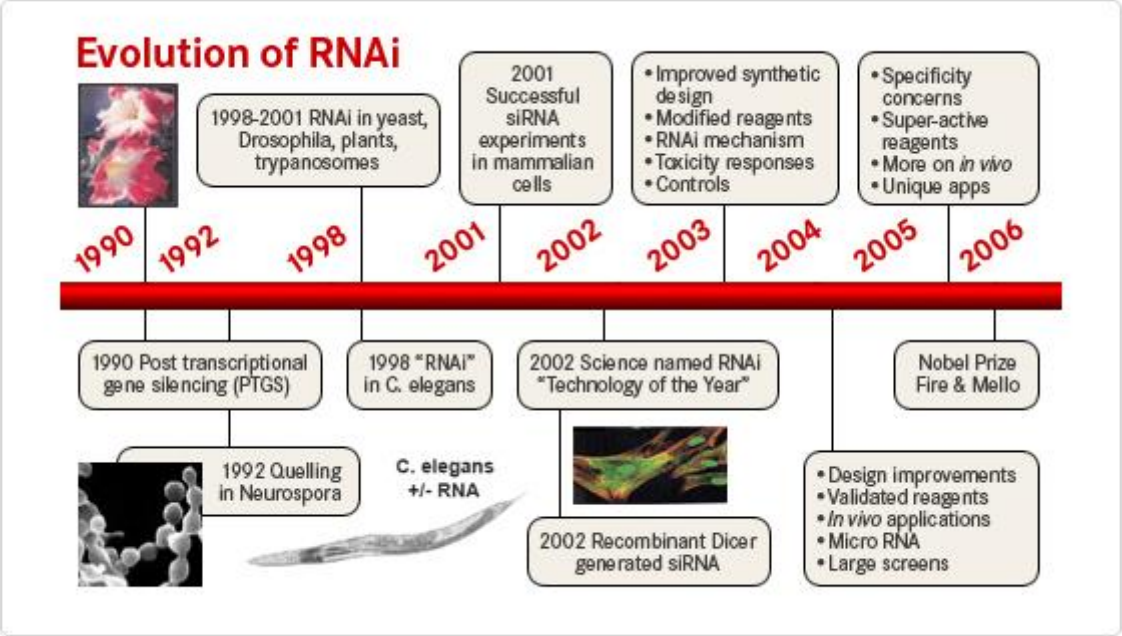
1.1. Tổng quan RNAi

RNA (Hoặc ARN) là axit ribonucleic – một trong hai loại axit nucleic (ADN, ARN) và là cơ sở di truyền cấp độ phân tử. Ở những loài không có ADN ví dụ một số loại virut thì ARN đóng vai trò là vật chất di truyền [1]. RNA tham gia vào quá trình phiên mã và dịch mã thông tin di truyền với nhiều vai trò khác nhau được đảm nhận bởi ba loại RNA (mRNA – RNA thông tin, tRNA – RNA vận chuyển, rRNA – RNA riboxom). Ngoài ra RNA còn có các chức năng điều hòa biểu hiện gen hoặc có chức năng tham gia các quá trình phát triển, biệt hoá tế bào như RNAi (interfering RNA).

RNA can thiệp (RNA interference, RNAi) là một cơ chế điều hòa biểu hiện gen được hướng dẫn (guiding) bởi RNA mà bằng cách này RNA mạch kép ức chế biểu hiện của các gen bằng các trình tự nucleotide bổ sung. 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 [1].

Khả năng ức chế của RNAi có thể gây nên các hiệu ứng: Ức chế dịch mã đơn vị mRNA, ức chế sự phiên mã của gen ở trong nhân, phân giải mRNA. Các hiệu ứng này gây nên sự ức chế biểu hiện của gen (ức chế gen), cụ thể sự tổng hợp protein sẽ bị giảm (knockdown) hoặc ngừng hoàn toàn (knock out) dẫn đến các tính trạng được quy định bởi gen đó bị suy giảm hoặc không xuất hiện.

1.2. Lịch sử nghiên cứu RNAi



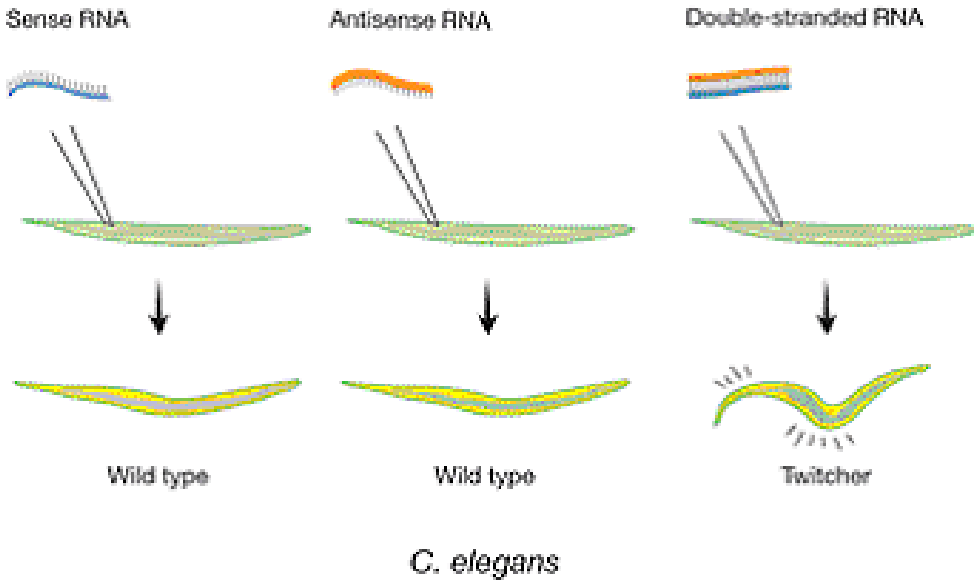
Hình 1: Lịch sử nghiên cứu RNAi [2]

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

- Năm 1984, Pesthea và các cộng sự đã nghiên cứu kỹ thuật Antisense-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 gen 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.

- 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. Năm 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. Kết quả là dsRNA ức chế sự biểu hiện của gen gấp 10 lần so với việc dùng phân tử RNA đơn lẻ theo chiều thuận hay chiều nghịch khi dùng phân tử RNA đơn lẻ còn dần dần mất tác dụng ức chế gen. Như vậy nhóm nghiên cứu của giáo sư Fire đã xác định được ngyên nhân chủ yếu của hiện tượng RNA silencing chính là do phân tử dsRNA gây nên. Hiện tượng này được các nhà khoa học đặt cho một thuật ngữ là RNA interference (RNAi). Việc tiêm mRNA mã hóa protein cơ không gây ra sự thay đổi nào ở giun. Mã di truyền của mRNA được mô tả như là một trình tự sense. Việc tiêm RNA antisense, một trình tự bổ sung với mRNA, cũng không mang lại tác động nào. Nhưng khi Fire và Mello tiêm RNA sence và antisense cùng với nhau thì họ quan sát thấy giun có những biểu hiện co giật đặc trưng. Những biểu hiện tương tự cũng được ghi nhận ở các giun bị khuyết hoàn toàn gen chức năng mã hóa protein cơ.



Hình 2: Biểu hiện của giun khi tiêm RNA liên quan đến mã hóa protein cơ [3]

- 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 RNAi trở thành công nghệ của năm.
- 2003-2005, khoảng thời gian cải tiến và tìm hiểu rõ hơn về công nghệ RNAi.
- 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)

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 trị bệnh di truyền trong tương lai

2. Cơ chế can thiệp RNAi

2.1. Các loại RNAi

Trung tâm của quá trình can thiệp RNAi gồm 2 thành phần siRNA và miRNA và những RNA 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.

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. siRNA tổng hợp hóa học là dạng đơn giản nhất của RNAi. Một trong những rào cản lớn nhất để đạt được hiệu quả RNAi với siRNA là nhiều tế bào khó để chuyển nạp. Thử nghiệm RNAi thường được coi là thành

công khi biểu hiện của gen mục tiêu giảm đến hơn 70%, khó có thể đạt được ở nhiều loại tế bào do hiệu quả của việc truyền thấp. Một nhược điểm nữa của việc sử dụng siRNA tổng hợp là thời gian hạn chế của các hiệu ứng sau khi truyền, điển hình là các hoạt động im lặng gen trong 24 giờ và giảm trong 48 giờ. Tổng hợp hóa học của siRNA tốn kém trong việc chuyển nạp cơ sở liên quan tới các thuốc thử nghiệm dựa trên vector DNA.

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).

Ngoài ra, một loại RNAi khác là shRNA có thể được đưa vào bởi DNA plasmid, mẫu tuyến tính hoặc vector virus hoặc vi khuẩn. Chính vì vậy loại RNAi này gây ra mối quan ngại về sự an toàn khi sử dụng.

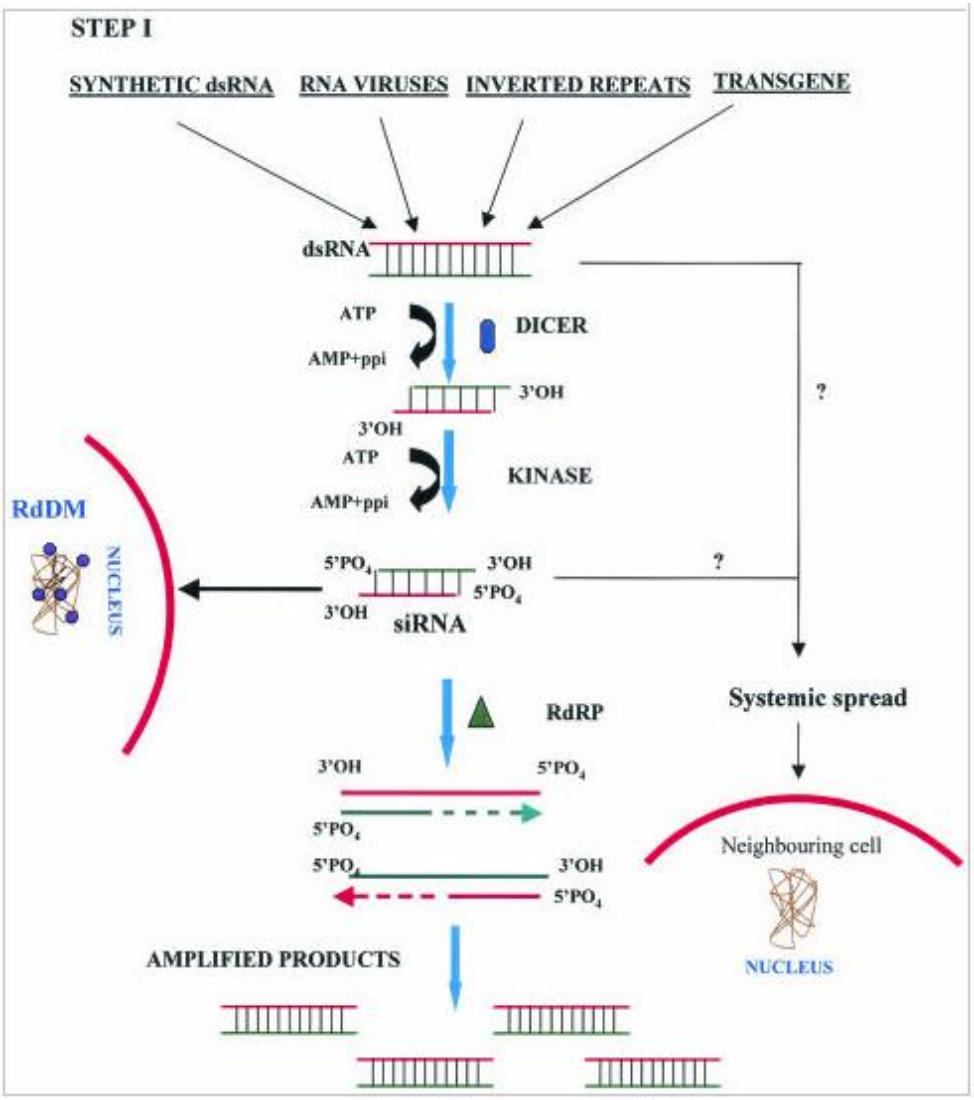
2.2. Cơ chế can thiệp RNA

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 (mô hình 2 bước được mô tả trong hình bên dưới). 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, mỗi siRNA kép được tách thành 2 sợi đơn siRNA, sợi passenger và sợi guider. Sợi passenger bị suy thoái còn sợi guider sẽ kết hợp vào RNA gây ra sự im lặng phức tạp (RISC). 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.

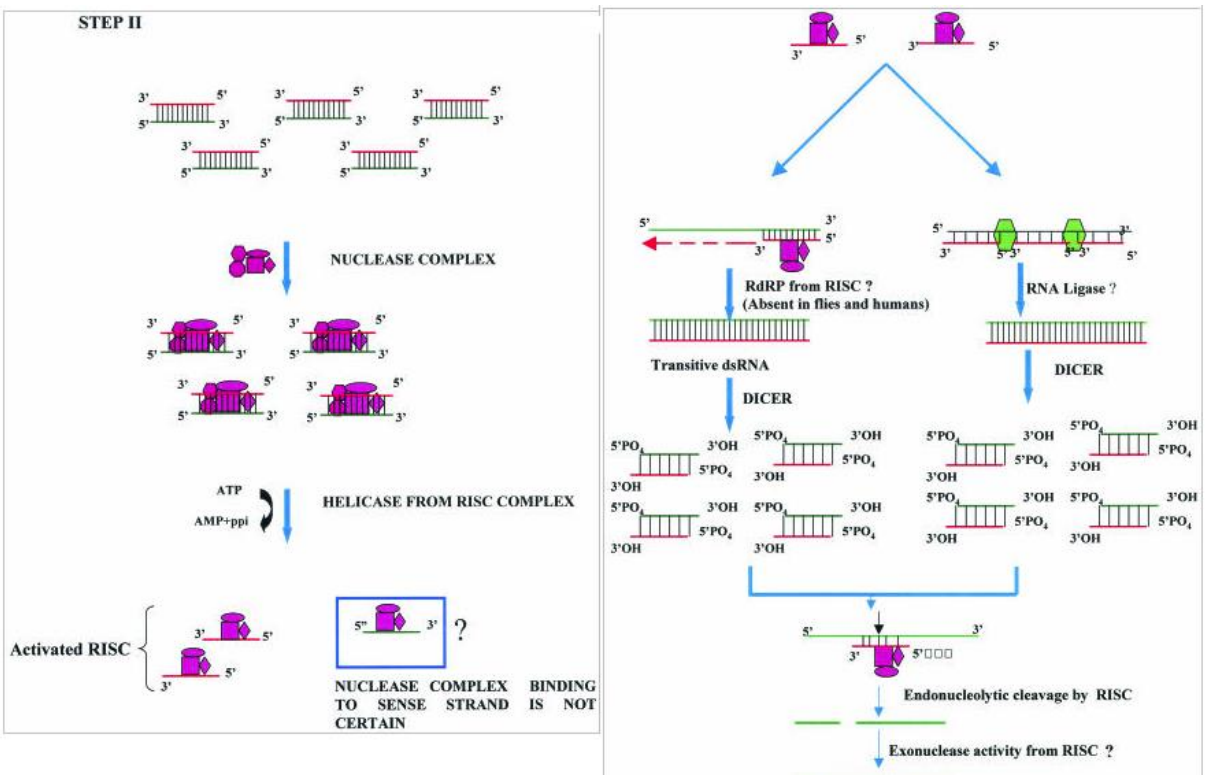
Ngoài ra trong bước 1 có thể xảy ra sự khuếch đại siRNA. Vì các đột biến gen mã hóa polymeraza RNA phụ thuộc RNA (RNA-dependent RNA polymerase - RdRP) ảnh hưởng đến RNAi nên loại polymerase này được đề xuất là có thể sao chép siRNA như các tác nhân biểu sinh, cho phép chúng lan truyền khắp cây trồng và giữa các thế hệ trong C. elegans. Các nghiên cứu của Lipardi và cộng sự và Sijen và cộng sự cung cấp các bằng chứng sinh học và di truyền thuyết phục rằng

RdRP thực sự đóng một vai trò quan trọng trong việc khuếch đại các hiệu ứng RNAi.

Cơ chế tắt gen 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 gen nào đó (vốn có rất nhiều bản sao trong cơ thể đa bào).



Hình 3: Bước 1, dsRNA bị cắt bởi enzyme Dicer để tạo ra các siRNA [4]



Hình 4: Bước 2, kết quả phân tách endonucleolytic của mRNA [4]

2.3. Ứng dụng RNAi và thách thức

Việc phát hiện ra RNAi và cơ chế làm im lặng 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. Mục tiêu của các nghiên cứu này là tìm ra những RNAi có khả năng ức chế cao đối với một số bệnh gây ra bởi gen (ví dụ ung thư) và ứng dụng nó vào cá thể để chữa bệnh.

- 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 răng
- 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

Vì RNAi giao nhau với một số con đường khác, không có gì đáng ngạc nhiên khi các hiệu ứng không mong muốn được kích hoạt bởi việc đưa một siRNA ra thử nghiệm.

- 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. Một phương pháp đầy hứa hẹn để giảm các hiệu ứng không mong muốn là chuyển đổi siRNA thành một microRNA. MicroRNAs xảy ra tự nhiên, và bằng cách khai thác con đường nội sinh này, nên có thể đạt được sự loại bỏ gen tương tự ở các nồng độ siRNA tương đối thấp. Điều này sẽ giảm thiểu các hiệu ứng không mong muốn.
- 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. Ở đây, các gen có bổ sung không hoàn chỉnh được vô tình giảm xuống bởi siRNA (có hiệu lực, siRNA hoạt động như một miRNA), dẫn đến các vấn đề trong việc giải đoán dữ liệu và độc tính tiềm ẩn. Tuy nhiên, điều này có thể được giải quyết bằng cách thiết kế các thí nghiệm kiểm soát thích hợp, và các thuật toán thiết kế siRNA hiện đang được phát triển để tạo ra các siRNAs miễn phí. Phân tích biểu hiện gen toàn bộ, ví dụ, bằng công nghệ vi mô, sau đó có thể được sử dụng để xác minh điều này và tinh chỉnh thêm các thuật toán.
- 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 đã bắt gặp các loại kháng thể này. Chưa có báo cáo về kháng thể chống lại siRNA gắn với protein.

3. Phát biểu bài toán

Những tri thức đã trình bày ở các phần trước đã chỉ ra những hiệu quả và lợi ích tiềm năng của RNAi trong việc chữa các bệnh gây ra bởi gen. Việc chữa bệnh lợi dụng vào khả năng ức chế của RNAi, cụ thể là tìm ra những RNAi có khả năng ức chế cao đối với bệnh, tức là suy giảm hoặc ngừng hoàn toàn biểu hiện của gen gây bệnh. Tuy nhiên việc ứng dụng RNAi vào thực thế còn gặp rất

nhiều thách thức như miễn dịch, sai mục tiêu. Những thách thức này đòi hỏi giải quyết các vấn đề liên quan: (i) Phải tìm ra những RNAi có khả năng ức chế hiệu quả và tránh được ức chế sai mục tiêu, (ii) Sau đó là giảm chi phí sản xuất RNAi và đưa nó vào cơ thể một cách an toàn. Để giải quyết vấn đề thứ nhất, đã có rất nhiều nghiên cứu được thực hiện và công bố từ năm 2001 cho tới nay nhằm thiết kế ra những siRNA hiệu quả có khả năng ức chế cao, hoặc dự đoán được khả năng ức chế của siRNA.

Và xét ở khía cạnh ứng dụng công nghệ thông tin, nghiên cứu khả năng ức chế bệnh của RNAi xoay quanh việc dự đoán khả năng ức chế bệnh của siRNA, và cũng là mối quan tâm của tôi khi thực hiện đề tài này. Dựa vào dữ liệu thực nghiệm từ nghiên cứu thiết kế siRNA hiệu quả của các nhà nghiên cứu sinh học và phương pháp xây dựng mô hình dự đoán của các nhà nghiên cứu tin học, tôi đã thực nghiệm các phương pháp biểu diễn RNA để biểu diễn siRNA và xây dựng mô hình dự đoán bằng thuật toán Hồi quy tuyến tính.

Trong công việc này, tôi đã thống kê tần số của các siRNA gắn nhãn có độ dài 19nt trong bộ dữ liệu siRecord, sau đó biểu diễn các siRNA trong các tập scored dataset bao gồm Huesken train, Huesken test, Vicker, Reynolds, Ui-tei theo phương pháp biểu diễn tần số k-merges và ghi lại biểu diễn này vào các file arff. Biểu diễn các siRNA trong Huesken train được sử dụng làm dữ liệu huấn luyện mô hình dự đoán sử dụng thuật toán Hồi quy tuyến tính, và các biểu diễn của các tập scored dataset còn lại được sử dụng làm dữ liệu kiểm chứng. Mô hình dự đoán được tạo ra được đánh giá bằng phương pháp Cross Validation 10 folds. Việc xây dựng, kiểm chứng và đánh giá mô hình dự đoán được thực hiện bằng phần mềm Weka 3.8.

Chương tiếp theo của luận văn sẽ trình bày tóm tắt các nghiên cứu liên quan tới bài toán ức chế bệnh của RNA từ những năm 2001 cho tới nay. Trong các phần tiếp theo, thuật ngữ "ức chế bệnh" được viết ngắn gọn là "ức chế".

CHƯƠNG 2. CÁC HƯỚNG NGHIÊN CỨU KHẢ NĂNG ỨC CHẾ BỆNH 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 cận: (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

Thời gian từ 2001 đến 2005, có rất nhiều nghiên cứu theo tiếp cận sinh học xác định các quy tắc thiết kế cơ bản của siRNA. Một số quy tắc thiết kế hợp lý cho siRNA đã được đề xuất bởi các nhóm nghiên cứu khác nhau như Tuschl [5], Reynolds [6], Chalk [7], Amarzguioui [8], Ui-tei [9], Hsieh [10], Jagla [11] sẽ được trình bày dưới đây. Các quy tắc thiết kế này chủ yếu dựa trên thông tin về hàm lượng G/C, ưu tiên hoặc tránh các nucleotide cụ thể ở vị trí nào đó và các motif chuỗi siRNA.

Vào năm 2001, M. Elbashir [5] và các cộng sự đã sử dụng Drosopila trong hệ thống ống nghiệm để chứng minh rằng các đoạn RNA 21 và 22 nucleotide là các trung gian với trình tự xác định của RNAi. Nhóm nghiên cứu đã đưa ra được quy tắc thiết kế Tuschl: (i) Lựa chọn miền mục tiêu ưu tiên 50-100nt hạ lưu của codon bắt đầu, (ii) Tìm chuỗi 5’-AA (N19) UU trong sợi antisense với N là bất kì nucleotide nào, (iii) Tìm các chuỗi 5’-(N’19) TT trên sợi sense với N là bất kì nucleotide nào, (iv) Hàm lượng G/C từ 32-79%.

Hai năm sau, Scherer và cộng sự [12] đã thăm dò cơ chế cơ bản của hoạt động ở chỉ các tác nhân ức chế antisense bao gồm Antisense ODNs, Ribozymes, DNAzymes, RNAi và so sánh ưu điểm, nhược điểm giữa chúng nhằm cung cấp nền tảng để đánh giá tác nhân nào phù hợp nhất với mục đích của thử nghiệm hoặc ứng dụng điều trị. RNAi có ưu điểm là (1) Hiệu quả ngay cả ở nồng độ thấp, (2) Bỏ qua interferon pathway, (3) Có thể phân phối theo nhiều cách, (4) Có thể

thể hiện mô cụ thể nhưng lại có những nhược điểm như (1) Không thể nhắm mục tiêu RNAs của nhân, (2) Không có lựa chọn để cải thiện nếu kháng mục tiêu, (3) Một vài báo cáo về việc ức chế sai mục tiêu. Ngoài ra nghiên cứu cũng báo cáo rằng các tính chất nhiệt động học nhắm mục tiêu mRNA là đặc trưng quan trọng.

Sau những nghiên cứu này, nhiều nguyên tắc thiết kế siRNAs hiệu quả đã được đề xuất. Năm 2003, Schwars và các cộng sự [13] đã chỉ ra rằng chỉ một thay đổi nhỏ trong chuỗi siRNA có những ảnh hưởng sâu sắc và có thể dự đoán được mức độ mà các sợi đơn trong một siRNA kép tham gia vào con đường RNAi, hiện tượng này được gọi là tính bất đối xứng của siRNA. Một số kết luận về tính bất đối xứng có được từ nghiên cứu: (i) Hai sợi siRNA có hiệu quả như nhau khi là các sợi đơn nhưng thể hiện hoạt động khác nhau đáng kể khi ghép cặp lại, điều này thể hiện tính bất đối xứng trong hoạt động của chúng được thiết lập tại một bước trong con đường RNAi trước khi gặp RISC được lập trình với mục tiêu RNA của nó, (ii) Hai sợi của siRNA kép được nạp khác nhau vào RISC và sợi đơn siRNA không được lắp ráp vào RISC sẽ bị phá hủy, (iii) Sự lắp ráp RISC thiên về các sợi siRNA có đầu 5' có xu hướng xung đột, (iv) Sự khác biệt về một liên kết Hidro đơn có ảnh hưởng đo được đối với sự đối xứng của sự lắp ráp RISC.

Ngay trong năm 2003, Khvorova A, Reynolds A, Jayasena SD [14] đã phân tích thống kê về sự ổn định nội tại của các chuỗi miRNA được tạo ra từ kẹp tóc tiền thân miRNA đã cho thấy sự linh hoạt tăng cường của các tiền thân miRNA, đặc biệt ở cặp bazo cuối đầu 5’ sợi antisense. Xu hướng tương tự đã được quan sát thấy ở siRNA, với các sợi kép hoạt động có độ ổn định bên trong thấp hơn (Δ 0.5 kcal / mol) ở đầu 5'-antisense (AS) so với các sợi kép không hoạt động. Sự ổn định nội tại trung bình của các siRNA thu được từ tế bào thực vật sau khi đưa ra các chuỗi RNA dài cũng cho thấy dấu hiệu nhiệt động học đặc trưng này. Một số kết luận từ quá trình nghiên cứu: (i) Tính ổn định bất đối xứng nội tại của sợi là một đặc trưng của tiền thân kẹp tóc miRNA, (ii) Các siRNA hoạt động có đầu 5’ antisense không ổn định, (iii) Tính chất nhiệt động học đặc trưng của siRNA tương quan mạnh với tính hoạt động. Các kết luận này cho thấy các tính chất nhiệt động học của các siRNA đóng một vai trò trung tâm trong việc xác định chức năng bằng cách tạo điều kiện cho một vài bước liên quan đến RISC trong con đường RNAi, cụ thể là sự trải ra của sợi đôi, lựa chọn sợi và sự chuyển đổi mRNA.

Năm 2004, Angela Reynolds và cộng sự [6] đã giới thiệu quy tắc thiết kế (quy tắc Reynolds) chuỗi siRNA sense độ dài 19 nt (nucleotide) được khái quát

lại theo 8 tiêu chuẩn: (i) Hàm lượng G/C từ 30 đến 52%, (ii) Có ít nhất 3 bazo A/U ở vị trí 15-19, (iii) Không lặp lại bên trong ($T_m < 20^o$), (iv) Một bazo A ở vị trí 19, (v) Một bazo A ở vị trí 3, (vi) Một bazo U ở vị trí 10, (vii) Một bazo khác G hoặc C ở vị trí 19, (viii) Một bazo khác G ở vị trí 13. Các phân tích trong nghiên cứu đã chỉ ra việc áp dụng một thuật toán kết hợp cả 8 tiêu chuẩn trên cải thiện đáng kể việc lựa chọn siRNA tiềm năng.

Cùng năm 2004, Amarzguioui M và cộng sự [8] đã thực hiện phân tích thống kê 46 siRNA, xác định các đặc điểm khác nhau của 19 cặp bazo có tương quan đáng kể với tính hoạt động ở mức độ knockdown 70% và đã xác minh các kết quả này dựa trên một bộ dữ liệu độc lập với 34 siRNA. Kết quả của nghiên cứu khuyến cáo nên sử dụng siRNA độ dài 19 nt có thiết kế theo tiêu chuẩn sau: (i) Có một đầu kép 3-nt dương chênh lệch A/U (nên là +2 hoặc +3), (ii) Nên kết hợp với nhiều nhân tố tích cực (S1, A6, W19), (iii) Tránh các nhân tố tiêu cực (U1 và G19). Trong đó S1 là G hoặc C ở vị trí 1 (S=G, C), W19 là A hoặc U ở vị trí 19 (W=A, U), U1 là U ở vị trí 1, A6 là A ở vị trí 6, G19 là G ở vị trí 19. Ngoài ra nghiên cứu cũng khuyên nên thiết kế siRNA với hàm lượng GC 32-53% và nhắm các mục tiêu có hàm lượng GC từ thấp đến trung bình. Những biện pháp phòng ngừa bổ sung này có thể hỗ trợ ngăn hiệu quả siRNA bị giới hạn bởi sự kết hợp mRNA mục tiêu hoặc cấu trúc phụ mRNA rộng.

Với cùng mục tiêu nghiên cứu, cũng năm 2004, nhóm nghiên cứu của Ui-Tei [9] đã phân tích mối quan hệ giữa chuỗi siRNA và hiệu quả RNAi sử dụng 63 mục tiêu của 4 gen ngoại sinh và 2 gen nội sinh và 3 tế bào của động vật có vú và ruồi giấm Drosophila. Dựa trên một số thành quả nghiên cứu của Schwars [13] điều kiện về chuỗi. Các nguyên tắc thiết kế được đề xuất bởi nghiên cứu phù hợp để thiết kế các siRNA hiệu quả cao cần thiết cho hệ thống gen của động vật có vú. Việc đáp ứng đồng thời cả 4 điều kiện chuỗi sau đây có khả năng gây ra hiệu quả im lặng gen rất cao trong tế bào động vật có vú: (i) A hoặc U ở đầu 5’ của sợi antisense, (ii) G hoặc C ở đầu 5’ của sợi sense, (iii) Có ít nhất 5 bazo A/U ở đầu 5’ một phần ba của sợi antisense, (iv) Sự vắng mặt của bất kì đoạn GC nào có chiều dài hơn 9 nucleotide. siRNAs đối nghịch với 3 điều kiện đầu tiên đưa ra tăng lên rất ít hoặc không gây ra sự im lặng gen ở tế bào động vật có vú. Về cơ bản các quy tắc tương tự cho chuỗi siRNA ưu tiên được tìm thấy có thể áp dụng cho DNA-based RNAi trong tế bào động vật có vú và trong RNA trong trứng (in

ovo) sử dụng phôi gà. Trái ngược với động vật có vú và gà, sự lựa chọn chuỗi siRNA có thể ít được phát hiện trong RNAi ở cá thể ruồi giấm Drosophila.

Ngoài ra, quy tắc thiết kế Stockholm của nhóm nghiên cứu Chalk [7], quy tắc thiết kế Hseih do Hsieh và cộng sự [10] cũng đưa ra vào năm 2004. Quy tắc Stockholm được tóm tắt như sau: (i) Tổng năng lượng kẹp tóc < 1, (ii) Đầu 5' antisense có năng lượng ràng buộc < 9, (iii) Đầu 5' sense có năng lượng ràng buộc trong khoảng 5-9 riêng biệt, (iv) Hàm lượng G/C từ 36-53%, (v) Năng lượng liên kết đoạn giữa (7-12) < 13, (vi) Sai khác năng lượng < 0, (vii) Sai khác năng lượng nằm trong khoảng -1 và 0. Quy tắc thiết kế Hseih: (i) Tránh mục tiêu giữa chuỗi mã hóa gen mục tiêu, (ii) Hợp nhất 4 hoặc 5 siRNA duplex cho mỗi gen, (iii) A ở vị trí 19 của sợi sense, (iv) G hoặc C ở vị trí 13 của sợi sense.

Tiếp theo năm 2005, nhóm nghiên cứu của Jagla [11] đã phân tích tổng hợp 601 siRNA kép có độ dài 21 nt với hai trong đó có 30 nucleotide nhô ra. Họ đã sử dụng thuật toán cây quyết định kết hợp với phân tích thông tin, các phân tích cho thấy bốn bộ quy tắc thiết kế chuỗi siRNA sense độ dài 19nt với hiệu quả knockdown trung bình từ 60% đến 73%. Bộ quy tắc thứ nhất: (i) A hoặc U ở vị trí 19, (ii) Có hơn 3 bazo A hoặc U ở vị trí từ 13-19, (iii) G hoặc C ở vị trí 1, (iv) A hoặc U ở vị trí 10. Bộ quy tắc thứ hai: (i) A hoặc U ở vị trí 19, (ii) Có hơn 3 bazo A hoặc U ở vị trí từ 13-19, (iii) G hoặc C ở vị trí 1, (iv) G hoặc C ở vị trí 10. Bộ quy tắc thứ ba: (i) G hoặc C ở vị trí 19, (ii) Có hơn 6 bazo A hoặc U ở vị trí từ 5-19, (iii) G hoặc C ở vị trí 1, (iv) G hoặc C ở vị trí 11. Bộ quy tắc thứ tư: (i) A hoặc U ở vị trí 19, (ii) Có hơn 3 bazo A hoặc U ở vị trí từ 13-19, (iii) A hoặc U ở vị trí 1. Quy tắc thứ nhất là quy tắc tốt nhất cho cơ hội 99,9% thiết kế một siRNA hiệu quả trong một bộ ba với hiệu quả knockdown hơn 50% trong chỉ thị sinh học. Bộ quy tắc tốt nhất đã áp dụng đối với tất cả các gen của con người (ENSEMBL 19) và cho thấy rằng 99,2% bộ gen có ít nhất ba điểm mục tiêu của siRNA, với hiệu quả trung bình dự đoán là 73%.

Nhìn chung các thiết kế đã được giới thiệu ở trên được phát triển vào thời gian 2001-2005 đều gặp hạn chế là dữ liệu về hiệu quả siRNA rất hạn chế nên không có cách đơn giản để các nhà phát triển nhận được phản hồi về các phương pháp của họ hoặc kiểm chứng hiệu quả của chúng. Hầu hết các phương pháp này đều thiếu cách đánh giá hữu hiệu ý nghĩa thống kê của các siRNAs được dự đoán.

Tới năm 2006, nhóm nghiên cứu Ren Y, Gong W, Xu Q, Zheng X, Lin D, Wang Y và cộng sự [15] đã xây dựng siRecords, một cơ sở dữ liệu của siRNAs

thực nghiệm đã được kiểm tra bởi nhiều nhà nghiên cứu với tỉ lệ hiệu quả nhất quán. Hơn 4100 chuỗi siRNA được ghi chú cẩn thận thu được từ hơn 1200 nghiên cứu siRNA đã công bố đã nằm trong siRecords và cơ sở dữ liệu này sẽ tiếp tục mở rộng khi có nhiều hơn siRNA được kiểm tra thử nghiệm được công bố. Trang chính của siRecords có thể được truy cập tại http://siRecords.umn.edu/siRecords/. Cơ sở dữ liệu này không chỉ giúp các nhà nghiên cứu RNA phát triển các quy tắc thiết kế siRNA đáng tin cậy hơn mà còn cung cấp thông tin về các siRNA đã được kiểm tra thực nghiệm và mức độ hiệu quả của nó khi nhắm mục tiêu các gen mà họ quan tâm.

Năm 2006, nhóm nghiên cứu Gong W [16] đã sử dụng một tập hợp lớn các dữ liệu về hiệu quả của siRNA được tập hợp từ nhiều nguồn gốc khác nhau (dữ liệu siRecords, có chứa 3277 siRNA thử nhắm mục tiêu 1518 gen, lấy từ 1417 nghiên cứu độc lập), tiến hành phân tích sâu rộng về tất cả các đặc điểm đã biết liên quan tăng hiệu quả RNAi. Một số đặc trưng có tác động dương tính lên hiệu quả siRNA đã được xác định. Bằng cách phân tích định lượng về hiệu quả hợp tác giữa các đặc trưng này, sau đó áp dụng một thuật toán hợp nhất (DRM), họ đã phát triển một nhóm quy tắc thiết kế siRNA với mức độ chính xác được kiểm soát (stringency level) và đã hạn chế được vấn đề dương tính giả. Họ đã so sánh với 15 công cụ thiết kế trực tuyến siRNA tại thời điểm đó và cho thấy một số quy tắc đã vượt qua tất cả các công cụ thiết kế thường được sử dụng trong thực tiễn thiết kế siRNA trong các giá trị tiên đoán dương (PPVs). Bộ quy tắc DRM RS 0.951 được chỉ thị cho mức α (stringency level) cao nhất (α = 0.951, được biểu thị là RS 0.951) chứa bảy quy tắc. Ngoài ra, với mức α thấp hơn thì số lượng lớn hơn các quy tắc được đưa vào bộ quy tắc (xem Bảng 6 tại Additionial file 1 được đính kèm vào bài báo [16]). Và bộ quy tắc DRM RS 0.951 với độ chính xác cao nhất với 7 quy tắc và 17 đặc điểm có tác động dương tính lên hiệu quả siRNA được mô tả trong hình bên dưới:

Non-redundant DRM rule set for the highest α level:$RS_{0.951}$.

Feature	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}	F_{13}	F_{14}	F_{15}	F_{16}	F_{17}
Rule 1	√	√			√										√		
Rule 2		√			√			√						√	√		
Rule 3	√	√				√									√	√	√
Rule 4		√			√		√	√							√	√	
Rule 5	√				√								√	√	√	√	
Rule 6		√			√	√	√	√							√		
Rule 7		√			√		√	√					√		√		

Bảng 1: Bộ quy tắc DRM RS 0.951 [16]

List of features:	
Feature Index	**Feature Names**
F_1	2nd nucleotide = A
F_2	4th nucleotide = C
F_3	6th nucleotide ≠ C
F_4	7th nucleotide ≠ U
F_5	9th nucleotide = C
F_6	17th nucleotide = A
F_7	18th nucleotide ≠ C
F_8	19th nucleotide = (A/U)
F_9	At least three (A/U)s in the seven nucleotides at the 3' end
F_{10}	No occurrences of four or more identical nucleotides in a row
F_{11}	No occurrences of G/C stretches of length 7 or longer
F_{12}	G/C content is between 35 and 60%
F_{13}	T_m is between 20 and 60°C
F_{14}	Binding energy of N16–N19 > -9 KCal/Mol
F_{15}	Binding energy of N16–N19 – binding energy of N1–N4 is between 0 and 1 KCal/Mol
F_{16}	Local folding potential (mean) ≥ -22.72 KCal/Mol
F_{17}	Target site is on CDS

Bảng 2: Các đặc điểm có tác động dương tính lên hiệu quả siRNA [16]

Tuy có nhiều hướng tiếp cận sinh học đã đươc công bố nhưng hiệu suất của chúng khi kiểm tra thực nghiệm không cao (65% siRNA được tạo ra bởi các quy

tắc thiết kế nói trên đã thất bại khi kiểm tra thực nghiệm, và gần 20% trong siRecords không hoạt động [17]). Các phân tích thực nghiệm này chỉ dựa trên các bộ dữ liệu nhỏ và tập trung vào siRNAs cho các gen cụ thể nên chưa thể mang tính đại diện cho toàn bộ siRNA với số lượng lớn hơn rất nhiều nên các quy tắc thiết kế này không đủ sự tin cậy để thiết kế được siRNA hiệu quả cao. Ngoài phương pháp tiếp cận bằng thực nghiệm và sử dụng công cụ để thiết kế siRNA, có một số nhà nghiên cứu đi theo hướng xây dựng các mô hình tiên đoán bằng cách sử dụng các kỹ thuật học máy đã được học qua các bộ dữ liệu lớn hơn.

2. Hướng nghiên cứu tin sinh học

Sau những nghiên cứu mở đầu với nhiều hạn chế theo hướng tiếp cận sinh học, thời gian tiếp theo là sự tiếp nối nghiên cứu ngày càng tăng về thiết kế siRNA, tuy nhiên khác với gian đoạn đầu các nghiên cứu này áp dụng phương pháp học máy thống kê tiên tiến để phân tích hiệu quả siRNA, dẫn đến sự phát triển của các quy tắc thiết kế siRNA tin cậy và mạnh mẽ hơn. Các phương pháp này dựa trên các kỹ thuật như SVM, self-organizing map, mạng nơ ron nhân tạo, cây quyết định và nhiều phương thức hạt nhân.

Đi tiên phong là Huesken và các đồng nghiệp [18], nghiên cứu theo hướng lai sinh học và tin học. Năm 2005, nhóm nghiên cứu đã sử dụng mô phỏng mạng nơ ron Stuttgart để huấn luyện các thuật toán trên tập dữ liệu gồm 2182 siRNA được chọn ngẫu nhiên nhắm mục tiêu 34 loại mRNA, được khảo sát thông qua một hệ gen chỉ thị huỳnh quang thông lượng cao. Thuật toán (BIOPREDsi) có thể dự đoán tin cậy hoạt động của 249 siRNAs của một bộ kiểm tra độc lập và siRNA nhắm mục tiêu gen nội sinh tại mRNA và protein với hệ số tương quan Pearson R = 0.66. Mạng nơ ron được huấn luyện trên một chuỗi guider có độ dài 21 nt có tính chất bổ sung là ưu việt hơn hẳn so với những phương pháp khác được huấn luyện trên chuỗi có độ dài 19 nt. Nhóm nghiên cứu đã đưa ra 5 quy tắc thiết kế siRNA hiệu quả: (i) Một bazo A hoặc U ở vị trí 1 của sợi antisense, (ii) Một bazo U ở vị trí 2, 7, 11 của sợi antisense, (iii) Một bazo A ở vị trí 10 của sợi antisense, (iv) Một bazo C ở vị trí 19 của sợi antisense, (v) Một bazo G ở vị trí 21 trong đầu 3’ nhô ra. Ngoài ra, nghiên cứu cung cấp bộ dữ liệu bao gồm 2431 scored siRNA là những siRNA mà hiệu quả knockdown được ghi điểm số đã được thực nghiệm quan sát, bộ dữ liệu này hiện được sử dụng rộng rãi cho việc huấn luyện và kiểm tra ở nhiều mô hình dự đoán khác.

Năm 2006, Shabalina SA, Spiridonov AN, Ogurtsov AY [19] đã thực hiện phân tích tính chất nhiệt động học và sự tương quan trên một bộ gồm 653 siRNA không đồng nhất được thu thập từ tài liệu. Họ đã sử dụng tập huấn luyện này để lựa chọn đặc điểm và tối ưu mô hình tính toán. Và nhóm cũng xác định được 18 tham số có độ tương quan đáng kể với hiệu quả im lặng, những tham số này hoặc đặc trưng cho chuỗi siRNA hoặc liên quan tới toàn bộ mRNA. Họ đã sử dụng multiple linear regression là kỹ thuật xây dựng mô hình cơ sở, dự đoán hiệu quả của tập dữ liệu kiểm chứng Novartis với 2431 siRNAs [18] sử dụng mô hình 3 tham số được tạo ra từ tập training. Nhóm nghiên cứu không mong đợi mô hình tuyến tính có thể làm việc việc tốt trên dữ liệu của họ nên họ đã áp dụng mạng nơ ron được cho là phù hợp với bất kì đặc trưng nào. Họ đã sử dụng tiêu chuẩn 7-fold cross validation chạy trên cả hai mô hình hồi quy tuyến tính và mạng nơ ron để đảm bảo không có siRNA nào có độ tương tự cao trên tập kiểm chứng và tập huấn luyện. Họ cũng tối ưu mô hình mạng nơ ron trên tập huấn luyện sử dụng ba tham số đặc trưng cho chuỗi RNAi và hệ số tương quan giữa hiệu quả dự đoán và hiệu quả quan sát được là 0.75. Sự khác biệt cũng là điểm mới trong mạng nơ ron của họ và mạng nơ ron nhân tạo trước đó là hướng tiếp cận của họ dựa trên cả nhiệt động học và sự tổng hợp các đặc trưng. Phương pháp ThermoComposition-21 của họ cũng có lợi thế so với BIOPREDsi là số lượng tham số nhỏ, họ chỉ sử dụng 3 tham số thay vì 84 nên đòi hỏi tập training nhỏ hơn để tạo ra kết quả đồng nhất. Mô hình của họ có thể sử dụng với bộ dữ liệu thực nghiệm nhỏ hơn dưới các điều kiện thực nghiệm khác nhau và hiệu quả để dự đoán hiệu quả siRNA ở cả nồng độ cao và nồng độ thấp. Nhóm nghiên cứu cũng đề xuất quy tắc thiết kế Shabalina bao gồm: (i) Một bazo U ở vị trí 1 của sợi antisense, (ii) Hàm lượng A/U cao hơn ở vị trí 1-3 của sợi antisense, (iii) Một bazo U ở vị trí 13-14 của sợi antisense, (iv) Một bazo khác A ở vị trí 17-19 của sợi antisense, (v) Một bazo G/C ở vị trí 19 của sợi antisense, (vi) Chỉ số hàm lượng dinucleotide bị tránh, (vii) Có ít hơn bản sao mục tiêu tiềm năng trong mRNAs, (viii) Khác biệt ΔG đáng kể giữa các vị trí 1 và 18.

Cùng năm 2006, Vert và cộng sự [20] đã huấn luyện mô hình tuyến tính trên tập dữ liệu Huesken dataset với phương thức hồi quy LASSO dẫn đến mô hình tuyến tính thưa thớt, tự động loại bỏ các đặc trưng không liên quan, cho phép một số lượng lớn các đặc trưng và tập trung trên những đặc trưng nhiều thông tin nhất. Nhóm đã hạn chế trên mô hình tuyến tính với hai tập đặc trưng đơn giản của

chuỗi: sự xuất hiện nucleotide ở mỗi vị trí trong chuỗi siRNA và hàm lượng tổng của chuỗi siRNA trong motif ngắn. Họ đã cho thấy các biểu diễn đều liên quan và bổ sung một phần thông tin để dự đoán hiệu quả siRNA, sự kết hợp của các biểu diễn dẫn đến một mô hình tuyến tính đơn giản (DSIR) và chính xác như mạng nơ ron BIOPREDsi. Sử dụng 5-fold cross validation trên tập huấn luyện để xác định độ chính xác của mô hình. Hơn nữa, qua quan sát họ cho thấy rằng kết hợp thêm các biểu diễn quang phổ (19 hoặc 21) vào các biểu diễn 21-sparse cho hiệu suất tốt hơn mô hình BIOPREDsi (0.67 so với 0.66), do đó xác nhận rằng cách tiếp cận đơn giản này có hiệu suất cao nhất trên bộ dữ liệu này. Nghiên cứu đã phát hiện và định lượng một xu hướng mạnh mẽ của siRNAs tiềm năng chứa các motif ngắn bất đối xứng trong chuỗi, và những motif này ít nhất có nhiều thông tin liên quan để dự báo tiềm năng ví dụ sự ưu tiên nucleotide cho các vị trí cụ thể. Quy tắc thiết kế Vert đã đề xuất như sau: (i) Có nhiều A/U hướng về phía đầu 5’ của sợi antisense, (ii) Có nhiều G/C hướng về đầu 5’ của sợi antisense, (iii) Một bazo khác C ở vị trí 7, 21 của sợi sense, (iv) Một bazo khác G ở vị trí 14 của sợi sense, (v) Motif AAC, UC, AAG, AGC có trong sợi antisense, (vi) Tránh motif CUU, CUA, GUU, GU, GAU trong sợi antisense.

Năm 2007, Ichihara và các cộng sự [21] đã phát triển một thuật toán đơn giản, i-Score (inhibitory-Score), để dự đoán siRNA hoạt động. Họ đã áp dụng mô hình hồi quy tuyến tính cho bộ dữ liệu A bao gồm 2431 siRNAs để xây dựng thuật toán i-Score dự đoán điểm số khả năng ức chế, và xác nhận nó với tập dữ liệu B bao gồm 419 siRNAs. Thuật toán i-Score dự đoán dựa vào trung bình 65 siRNA hoạt động trên mỗi mRNA trong phân tích hệ gen cũng như các thuật toán khác, sử dụng độ ưu tiên nucleotide (nt) ở mỗi vị trí được chuẩn hóa từ 0 đến 100 để tính điểm $i\text{-}Score = \sum_{m=1}^{19} P_{mn}(n: A, C, G, U)$. Độ chính xác dự đoán của i-Score đã được so sánh là tốt như của s-Biopredsi, ThermoComposition21 và DSIR, sử dụng một mô hình mạng thần kinh hoặc nhiều tham số trong mô hình hồi quy tuyến tính. Reynolds và Katoh cũng dự đoán các siRNA hoạt động hiệu quả, nhưng số lượng các siRNA dự đoán để được hoạt động ít hơn một phần tám của i-Score. Nhóm nghiên cứu đã phát hiện thêm rằng việc loại trừ các siRNA chịu nhiệt có toàn bộ năng lượng xếp chồng (ΔG) ít hơn 234.6 kcal / mol, cải thiện độ chính xác dự đoán trong i-Score, s-Biopredsi, ThermoComposition21 và DSIR. Vector mục tiêu phổ biến pSELL, được họ phát triển có thể xác định hoạt động siRNA của bất kỳ chuỗi sense và antisense. Kết quả khảo sát 86 siRNA trong tế

bào HEK293 sử dụng pSELL, đã xác nhận tính khả dụng của i-Score và giá trị ΔG trong việc thiết kế siRNAs.

Nhóm nghiên cứu Matveeva [22] đã sử dụng 4 cơ sở dữ liệu độc lập tổng cộng 3336 thí nghiệm xác minh siRNAs để so sánh một số phương pháp dự đoán hiệu quả tách siRNA. Theo đặc điểm hoạt động của máy thu (ROC) và phân tích tương quan, các chương trình tốt nhất là BioPredsi, ThermoComposition và DSIR tại thời điểm đó. Các tham số được sử dụng trong các phương pháp lúc đó được chia thành 2 nhóm, liên quan và không liên quan đến sự ổn định của đầu cuối của sợi siRNA duplex. Nhóm 1 bao gồm tính ổn định của đầu duplex (được tính bằng ΔG_{37}^{0}) hoặc sự có mặt hoặc vắng mặt của một nucleotide xác định tại một vị trí đầu cuối duplex nào đó. Nhóm 2 bao gồm các tham số như tỷ lệ % nucleotide trên sợi siRNA sense hoặc antisense, sự hiện diện hoặc vắng mặt của một nucleotide cụ thể tại một vị trí bên trong nào đó, sự ổn định của cấu trúc thứ cấp của mRNA mục tiêu hoặc sự ổn định của siRNA antisense. Nhóm đã sử dụng những bộ dữ liệu lớn nhất từ các phương pháp này để phát triển một phương pháp mới với các thông số tối ưu nhóm 1 và nhóm 2, hiện đã trở thành một công cụ thiết kế siRNA qua web với tên gọi "siRNA scale". Phương pháp này sử dụng hồi quy tuyến tính phù hợp với sự ổn định duplex bên trong, mức ưu tiên nucleotide tại các vị trí, hàm lượng G/C của duplexes siRNA là các tham số đầu vào. Khả năng phân biệt siRNA hiệu quả và không hiệu quả có thể so sánh được với các phương pháp tốt nhất đã được đưa ra nhưng các tham số của nó liên quan hơn đến cơ chế hoạt động siRNA so với BioPredsi. Phương pháp mới của họ cũng dự đoán hiệu quả siRNA nhanh hơn của ThermoComposition vì nó không tốn nhiều thời gian tính toán cấu trúc RNA thứ cấp và có ít thông số hơn DSIR.

Đáng chú ý là công trình nghiên cứu của năm 2009, Qiu S và Lane T [23] đã phát triển môt mô hình học máy hồi quy vector hỗ trợ đa nhân (MKSVR) hồi quy với RNA chuỗi hạt nhân để dự đoán hiệu quả siRNA. Trước đó, các quy tắc thiết kế siRNA mô tả các đặc tính cấu trúc và nhiệt động lực học đề xuất dưới dạng mô tả số học nhiều chiều dẫn đến khoảng trống đầu vào cho các mô hình học máy. Huesken và đồng nghiệp [18] đã xây dựng một mạng nơ ron nhân tạo và báo cáo tương quan dự đoán hệ số, nhưng tỷ lệ lỗi, chẳng hạn như bình phương bình phương sai (MSE), đã không được hiển thị. Hơn nữa, mạng nơ-ron là được đào tạo bằng tìm kiếm độ dốc, phụ thuộc vào giá trị ban đầu và không đảm bảo sự tối ưu toàn cầu. Vert và cộng sự sử dụng hồi quy tuyến tính để dự đoán hiệu

quả và lựa chọn các bước quan trọng nhưng chỉ sử dụng các mô hình tuyến tính và không khai thác tính phi tuyến. Không gian vector cũng có thể được xây dựng sử dụng các motif hạt nhân và các mã hóa thưa thớt thường được sử dụng để biểu diễn các chuỗi trong không gian vector, cộng với sử dụng ngưỡng hiệu quả để phân loại siRNA thành hai lớp hoạt động nếu bằng hoặc vượt ngưỡng và không hoạt động cho các trường hợp còn lại (dưới ngưỡng) sẽ tạo điều kiện để áp dụng các thuật toán phân loại, chẳng hạn như máy vector hỗ trợ (SVM), cây quyết định, và mạng thần kinh. Để dự đoán chính xác hiệu quả hơn phân loại, phương pháp hồi quy đã được sử dụng. Vì một máy vector hỗ trợ sử dụng giảm thiểu hóa rủi ro cấu trúc và chương trình lồi dẫn đến tối ưu hóa toàn cục, nó có khả năng tổng quát tốt hơn các mô hình học tập khác. Năm 2007, Qui và Lan cũng đã đạt được một số thành tựu nghiên cứu liên quan trực tiếp đến mô hình MKSVR, thứ nhất xây dựng không gian véc tơ đa chiều từ mô tả các quy tắc thiết kế siRNA để sử dụng hạt nhân số hồi quy vector hỗ trợ (SVR) và đạt được những tính chính xác đáng kể trong hiệu quả dự, thứ hai phát triển và áp dụng chuỗi hạt nhân để dự đoán và đạt được độ chính xác cao hơn hơn các hạt nhân số. Nhóm nghiên cứu đã đề xuất một nền tảng hồi quy hạt nhân nhiều để thống nhất thông tin trong không gian đặc trưng hạt nhân - tổ hợp tuyến tính của chuỗi và hạt nhân số để cải thiện mô hình học tập. Họ đã xây dựng học đa nhân thành một bài toán quy hoạch toàn phương bậc 2 (QCQP). Kết quả thực nghiệm đã chứng minh rằng hồi quy đa nhân đã cải thiện hiệu suất dự đoán và đơn giản sự phức tạp của mô hình bằng cách giảm số lượng các vector hỗ trợ. Mặc dù công thức QCQP tạo ra giải pháp tối ưu toàn cục, nhưng không hiệu quả về mặt tính toán và yêu cầu giải quyết thương mại. Do đó, họ tiếp tục đề xuất heuristic cho học đa nhân để tăng tốc tính toán và đơn giản hóa việc sử dụng. Thử nghiệm trên bốn bộ dữ liệu sinh học chứng minh rằng các heuristics cho hồi quy hạt nhân nâng cao độ chính xác dự đoán và tăng tốc độ tính toán hiệu suất. Hơn nữa, nó cung cấp cái nhìn sâu sắc vào hạt nhân, mang lại lợi ích bổ sung cho việc so sánh ý nghĩa tương đối của các quy tắc thiết kế.

Trong năm đó, Klingelhoefer và cộng sự [24] đã xây dựng một tập dữ liệu meta lớn nhất thời điểm đó bao gồm 6483 siRNAs đã công khai (nhắm mục tiêu đến mRNA động vật có vú), sau đó áp dụng một phân tích Bayesian phù hợp với những đặc trưng không chắc chắn. Thuật toán dựa trên hồi quy logistic ngẫu nhiên là được thiết kế để khám phá một không gian mô hình rộng lớn của 497 đặc trưng compositional, các đặc tính cấu trúc và nhiệt động lực học, xác định các liên kết

với siRNA hiệu lực. Tập dữ liệu meta của họ được thu thập dữ liệu thực nghiệm của một loạt các phương pháp thống kê bao gồm các phương pháp hồi quy tuyến tính đơn giản (bao gồm Matveeva và cộng sự năm 2007, Shabalina và cộng sự năm 2006, Ui-Tei và cộng sự năm 2004, Vert và cộng sự năm 2007), các phương pháp phức tạp hơn như mạng thần kinh (nghiên cứu của Huesken và đồng nghiệp năm 2005, Shabalina và cộng sự năm 2006), đồ thị Euler (nghiên cứu của Pancoska và cộng sư năm 2004), máy vecto hỗ trợ (nghiên cứu của Ladunga năm 2007, nghiên cứu của Peek năm 2007, nghiên cứu của Saetrom năm 2004, nghiên cứu Teramoto và cộng sự năm 2005), lập trình di truyền (nghiên cứu của Saetrom năm 2004) và hợp nhất các luật rời rạc (nghiên cứu của Gong và cộng sự năm 2008). Họ lựa chọn phương thức dựa trên hồi quy tuyến tính vì tính đơn giản nhưng hiệu quả lại so sánh được với phần lớn các phương pháp phức tạp khác. Họ đã kiểm tra thuật toán lựa chọn đặc trưng Bayesian Markov chain Monte Carlo (Bayesian MCMC), sử dụng Bayesian Information Criteria (BIC) để ước lượng yếu tố Bayes cho một mô hình hồi quy logitic trên một siêu dữ liệu lớn của 6483 siRNA đã xây dựng. Thuật toán đạt hiệu quả thành công trong việc xác định những đặc trưng mới ảnh hưởng đáng kể đến hiệu quả của siRNA và hiệu năng có thể so sánh được với các phương pháp dự đoán thành công thời điểm đó. Bằng thuật toán của mình, họ đã đề xuất 10 quy tắc thiết kế siRNA hiệu quả đối với mRNA của loài người như sau: (i) Một bazo A ở vị trí 1 hoặc 10 của sợi antisense, (ii) Một bazo U ở vị trí 1 của sợi antisense, (iii) Một bazo khác A ở vị trí 19 của sợi antisense, (iv) Một bazo khác G ở vị trí 14 hoặc 18 của sợi antisense, (v) Hàm lượng G/C từ 35-73%, (vi) Motif UCU, UCCG có mặt trong sợi antisense, (vii) Tránh các motif ACGA, GCC, GUGG trong sợi antisense, (viii) ΔG cao tại các vị trí từ 1-4, 5-8 và 13-14 của sợi antisense, (ix) ΔG thấp tại các vị trí từ 18-19 của sợi antisense, (x) Tránh sự gấp lại trong siRNA.

Năm 2012, Sciabola và cộng sự [25] chứng minh sử dụng các mô tả 3 chiều đã cải thiện sự phân biệt giữa siRNA hoạt động và không hoạt động trong các mô hình thống kê. Đã có 5 loại mô tả được sử dụng: (i) Vị trí nucleotide dọc theo chuỗi siRNA, (ii) Thành phần nucleotide liên quan tới sự có mặt/vắng mặt của một tổ hợp di hoặc tri-nucleotides cụ thể, (iii) Tương tác nucleotide bởi ý nghĩa của một hàm tự hiệp và hiệp phương sai chéo đã có sửa đổi, (iv) Sự ổn định của nhiệt động học nucleotide thu được biểu diễn mô hình hàng xóm gần nhất, (v) Sự linh hoạt trong cấu trúc axit nucleic. Mô tả tính linh hoạt của duplex có nguồn gốc

từ mô phỏng động lực học phân tử mở rộng có thể mô tả các đặc tính đàn hồi phụ thuộc trình tự của RNA duplexes, ngay cả đối với các oligonucleotides không tiêu chuẩn. Ma trận của các mô tả được phân tích bằng cách sử dụng ba gói thống kê (bình phương nhỏ nhất từng phần, rừng ngẫu nhiên, và máy vector hỗ trợ). Việc thực hiện các mô tả RNA mới cùng với các thuật toán thống kê thích hợp đã cải thiện hiệu suất mô hình cho việc lựa chọn siRNA khi so sánh với công cụ dự đoán siRNA công khai và các bộ thử nghiệm được công bố trước đó. Việc sử dụng các mô tả 1D và 3D dựa trên trình tự đã được kiểm tra trong các mô hình hiệu quả siRNA. Sử dụng những mô tả này và bộ dữ liệu Huesken, các mô hình hiệu quả siRNA được tạo ra bằng cách sử dụng ba kỹ thuật hồi quy: (i) Bình phương nhỏ nhất từng phần (PLS), (ii) Rừng ngẫu nhiên (RF), (iii) Máy vector hỗ trợ (SVM). Việc xác nhận kết quả đã được thực hiện thông qua cross-validation và các dự đoán bên ngoài dựa bộ kiểm tra độc lập cho phép xác định các kết hợp tốt nhất các mô tả và thuật toán hồi quy.

Trong năm 2012, nhóm nghiên cứu Qi Liu và cộng sự [26] đã khảo sát chi tiết về thiết kế siRNAs tiên tiến, tập trung vào một số vấn đề chính với thực tại trong các nghiên cứu RNAi silic, bao gồm: (i) Sự không nhất quán giữa các hướng đề xuất cho việc thiết kế siRNAs và danh sách các đặc trưng của siRNAs không đầy đủ, (ii) Tích hợp không chính xác các dữ liệu siRNAs nền tảng không đồng nhất, (iii) Xem xét không đầy đủ về độ ràng buộc cụ thể của mRNA mục tiêu và (iv) Giảm hiệu ứng sai mục tiêu trong thiết kế siRNAs. Họ tin rằng giải quyết các vấn đề như vậy trong nghiên cứu siRNA sẽ cung cấp các đầu mối mới cho thiết kế cải tiến thiết kế siRNA hiệu quả hơn và đặc trưng hơn trong RNAi. Nhóm nghiên cứu của Mysara và cộng sự [27] đã sử dụng mạng nơ ron để huấn luyện một mô hình dự đoán hiệu quả/điểm số của siRNA mới, được phát triển dựa trên việc kết hợp hai thuật toán ghi điểm (ThermoComposition21 và i-Score), cùng với năng lượng xếp chồng (ΔG), trong một mạng nơ ron nhân tạo đa lớp. Mô hình MysiRNA của họ đã được huấn luyện trên 2431 siRNA và được thử nghiệm bằng cách sử dụng ba bộ dữ liệu khác A (Novartis), B (bao gồm dữ liệu thực nghiệm Reynold, Vickers, haborth, Ui-tei, Khovorova), C (được trích từ B). MysiRNA đã thu được kết quả AUCs là 0.855, 0.808 và 0.834, và hệ số tương quan Pearson là 0.687, 0.600 và 0.699 trên các tập A, B và C cao hơn so với các công cụ sẵn có. Mô hình MysiRNA là một phần của gói thiết kế MysiRNA được mong đợi sẽ đóng một vai trò quan trọng trong việc lựa chọn và đánh giá siRNA. Nhóm nghiên

cứu Chang và cộng sự [28] đã tạo ra một công cụ thiết kế siRNA "DEsi" nhanh chóng chọn siRNAs có hoạt động RNAi cao đối với mRNA mong muốn. DEsi kết hợp các bộ lọc tính năng truyền thống, mô hình học máy và BLAST để tối ưu hóa thiết kế siRNA. Trong đó, SVMs được huấn luyện trên hai tập dữ liệu siRNA 19-nt và 21-nt của siRecord với các tham số tối ưu. Các mô hình dự đoán trong DEsi có năng lực dự đoán đoán đáng kể, đã được xác nhận bởi phân tích thống kê. So với các công cụ thiết kế siRNA khác, DEsi có thể nhanh chóng và chính xác thiết kế siRNAs chống lại các mRNA mong muốn. DEsi thể hiện hiệu quả rất cao khi dự đoán siRNA trong siRecord, kết quả đều cao hơn khi so sánh với DSIR, Invitrogen sTarget Finder, và Abions Target Finder.

Năm 2015, Bùi Ngọc Thăng và cộng sự [17] đã phát triển một nền tảng chung để tăng cường dự đoán hiệu quả knockdown của siRNA. Ý tưởng chính trước hết là làm giàu chuỗi siRNA bằng kết hợp chúng với các quy tắc đã được tìm thấy trong thiết kế các siRNA hiệu quả và biểu diễn chúng dưới dạng ma trận được làm giàu, sau đó sử dụng hồi quy tensor song tuyến tính để dự đoán hiệu quả knockdown của các ma trận này. Để thực hiện ý tưởng, nhóm nghiên cứu đã thực hiện 4 công việc sau đây. Thứ nhất, xây dựng một biểu diễn thích hợp của siRNAs, biểu diễn ma trận làm giàu, bằng cách kết hợp các quy tắc thiết kế siRNA sẵn có bao gồm bảy quy tắc Reynolds, quy tắc Uitei, quy tắc Amarzguioui, quy tắc Jalag, quy tắc Hsieh, quy tắc Takasaki và quy tắc Huesken và sử dụng cả hai siRNAs gắn nhãn và ghi điểm số. Thứ hai, xây dựng một phương pháp tiên đoán cao hơn và ổn định để dự đoán hiệu quả của siRNA bằng cách xây dựng mô hình hồi quy tensor song tuyến tính. Các quá trình học của ma trận chuyển đổi và các thông số của mô hình được kết hợp với nhau để tạo ra biểu diễn siRNA chính xác hơn. Các siRNA được gắn nhãn được sử dụng để giám sát quá trình học của các tham số. Thứ ba, xác định định lượng các vị trí trên trên siRNAs mà các nucleotide có thể ảnh hưởng mạnh đến khả năng ức chế siRNAs. Thứ tư, cung cấp hướng dẫn dựa trên các đặc trưng về vị trí tạo siRNA hiệu quả cao. Mô hình hồi quy tensor song tuyến tính do nhóm nghiên cứu phát triển được gọi là BiLTR được thử nghiệm so với các mô hình được công bố trên tập dữ liệu Huesken và ba bộ dữ liệu độc lập thường được sử dụng bởi cộng đồng nghiên cứu Reynolds, Vicker và Harborth. Kết quả cho thấy hiệu suất của các dự đoán BiLTR hầu hết là ổn định hơn và cao hơn các mô hình khác với hệ số tương quan Pearson đạt 0.64 với Huesken dataset, 0.67 với HU test, 0.57 với Reynold, 0.58 với Vicker, 0.57 với Harborth.

Và gần nhất năm 2017, hai nhóm nghiên cứu của Fei He [29], và nhóm của Ye Han [30] đã công bố nghiên cứu mới nhất đã cho kết quả dự đoán tốt hơn các mô hình đã công bố trước đó. Nhóm nghiên cứu của Ye Han [30] đã giới thiệu 2-3NTs là các đặc trưng mới. Một bộ đặc trưng hỗn hợp 230 chiều được tạo ra bằng cách kết hợp 191 đặc trưng truyền thống và 39 đặc trưng do nhóm đề xuất. Vì có nhiều đặc trưng tiềm năng, nên thuật toán lựa chọn đặc trưng tìm kiếm nhị phân (Binary Search Feature Selection - BSFS) dựa trên tầm quan trọng của RF-variable được đề xuất để chọn bộ đặc trưng tối ưu. Và 9 đặc trưng do nhóm đề xuất được đưa vào trong tổng cộng 57 đặc trưng đã được chọn làm vectơ đầu vào của mô hình RF để dự đoán hoạt động của siRNA. Đáng chú ý là motif trinucleotide ở vị trí 19 đã được đưa vào bộ đặc trưng lựa chọn, đây là vị trí gắn kết của protein Argonaute, họ cũng thấy rằng "CUG" xảy ra thường xuyên nhất ở vị trí 19 siRNAs tiềm năng. Nhóm đã phát triển và đánh giá một công cụ có tên "siRNApred" với một bộ đặc trưng hỗn hợp nói trên để dự đoán hoạt động siRNA. Việc đánh giá so sánh thử nghiệm về các tập dữ liệu được sử dụng phổ biến cho thấy siRNApred tạo ra kết quả tốt hơn so với phương pháp thiết kế siRNA thế hệ thứ nhất và thế hệ thứ hai. Kết quả khi huấn luyện bằng Huesken train và kiểm chứng với tập Huesken test cho kết quả PCC = 0.722, cao hơn lần lượt 9.39%, 10.39%, 9.56%, và 7.76% so với các phương pháp Biopredsi, i-score, ThermoComposition-21 và DSIR. Do đó siRNApred được xem một công cụ xứng đáng để thiết kế siRNA hiệu quả cho một mRNA đầu vào sử dụng bộ đặc trưng tối ưu.

Trong lúc đó, Fei He và cộng sự [29] cố gắng để mô tả siRNA từ cả hai khía cạnh định lượng và định tính. Đối với các phân tích định lượng, họ tạo thành bốn nhóm các đặc trưng hiệu quả, bao gồm tần số nucleotide, hồ sơ ổn định nhiệt động lực học, nhiệt động lực học của sự tương tác siRNA-mRNA, và đặc trưng liên quan đến mRNA, là một biểu diễn hỗn hợp mới, trong đó tương tác nhiệt động lực học của siRNA mRNA lần đầu tiên được đưa vào để tiên đoán hiệu quả siRNA. Sau đó họ sẽ chọn một đặc trưng dựa trên điểm F để kiểm tra sự đóng góp của mỗi đặc trưng và loại bỏ các đặc trưng có liên quan yếu. Trong khi đó, họ mã hóa chuỗi siRNA và các quy tắc thiết kế theo thực nghiệm đã có thành một biểu diễn siRNA định lượng. Hai loại biểu diễn siRNA được kết hợp để dự đoán hiệu quả siRNA bằng hồi quy vecto hỗ trợ (SVR) ở mức điểm số. Kết quả thực nghiệm của họ cho thấy phương pháp họ đề xuất có thể chọn các đặc trưng có khả năng

phân biệt mạnh mẽ và làm cho hai loại biểu diễn siRNA thể hiện đầy đủ khả năng. Kết quả dự báo với PCC (hệ số tương quan Pearson) là 0.73, cao hơn 10.61% sơ với Biopredsi, 11.62% so với i-Score, 10.77% so với ThermoComposition-21 và 8.96% so với DSIR cũng chứng minh rằng phương pháp của họ có thể vượt trội các thuật toán tiên đoán hiệu quả siRNA khác.

Trong luận văn này, tôi đã trình bày tóm tắt đa dạng các phương pháp từ các phương pháp tiếp cận sinh học để đề xuất các quy tắc thiết kế siRNA hiệu quả thế hệ đầu tiên chủ yếu sử công cụ và làm thực nghiệm, cho đến thế hệ thứ hai các phương pháp học máy thống kê từ năm 2005 cho tới hiện này. Kết quả tìm hiểu đã cho thấy sự tiến bộ dần của các phương pháp nghiên cứu, độ tin cậy của mô hình dự đoán và các phương pháp thiết kế đã được cải thiện. Trong đó đáng chú ý nhất phải nói đến mô hình học máy sử dụng SVR đã được đề xuất của Fei He và cộng sự [29] đã cho hệ số tương quan Pearson lên tới 0.73 cao nhất cho tới nay. Tuy nhiên chúng ta vẫn mong đợi một mô hình dự đoán với hệ số tương quan cao hơn để có thể dự đoán hiệu quả siRNA một cách tốt nhất. Kết quả của phương pháp thế hệ thứ nhất rất thấp do hạn chế của việc nghiên cứu như bộ dữ liệu quả nhỏ nên không bao phủ được hết các đặc trưng. Trong khi đó, hiệu suất của các thuật toán thế hệ thứ hai phụ thuộc rất nhiều vào việc lựa chọn các đặc trưng được đưa vào. Bởi vì chuỗi siRNA là yếu tố quan trọng quyết định hoạt động của RNAi nên nhiều tính năng tiềm ẩn được nhúng vào chuỗi siRNA cần được khai thác để tăng độ chính xác dự đoán. Tuy vậy, các phương pháp ở thế hệ thứ hai vẫn còn thiếu xót như: (1) một số phương pháp tập trung vào đặc tính trình tự và profile của siRNAs theo trình tự nhưng bỏ qua việc áp dụng các quy tắc thực nghiệm, (2) một số phương pháp còn đặt đặc trưng tương tác nhiệt động học của siRNA-mRNA và các tính năng liên quan tới mRNA để xem xét, tuy nhiên người ta đã chứng minh rằng các mRNA liên quan tính năng có thể giúp dự đoán hiệu quả siRNA, (3) Mặc dù công cụ siPred đã cố gắng kết hợp các đặc trưng với nhau với các quy tắc là đầu vào, nhưng nó đã không chú trọng việc đối phó sự không đồng nhất dữ liệu giữa các dữ liệu liên tục và nhị phân, mà có thể ảnh hưởng đến tính chính xác của mô hình hóa một hệ thống hồi quy tuyến tính và (4) Các cách biểu diễn siRNA có ảnh hưởng lớn đối với hiệu quả dự đoán của các mô hình.

Cụ thể, một số cách biểu diễn siRNA đã được đề xuất ở các nghiên cứu được trình bày phía trên. Vert và cộng sự [20] đã biểu diễn 1 cho sự xuất hiện, 0 cho sự vắng mặt của 84 motif 1, 2, 3 nucleotide trong chuỗi siRNA. Trong phương

pháp dự đoán i-Score, nhóm nghiên cứu của Ichihara [21] đã biểu diễn siRNA theo độ ưu tiên của các nucleotide tại các vị trí, độ ưu tiên được chuẩn hóa từ 0 cho đến 100. Qui và Lane [23] kết hợp các quy tắc thiết kế đã có là Ui-tei, Reynolds, Jagla, Huesken trong biểu diễn của họ sử dụng 2 thuộc tính nhị phân, cụ thể (1, 0) thể hiện một quy tắc nào đó xuất hiện trong chuỗi siRNA và (0, 1) nếu ngược lại. Sciabola và cộng sự [25] biểu diễn siRNA dưới dạng số học, biến đổi tương ứng mỗi nucleotide thành một bộ ba với A(-1, -1, +1), C(+1, -1, -1), G(-1, +1, -1) và U/T(+1, +1, +1). Bùi Ngọc Thăng và cộng sự [17] đã sử dụng phương pháp ma trận chuyển đổi, ban đầu các chuỗi siRNA được mã hóa thành ma trận nx4 trong đó n là số nucleotide của chuỗi siRNA, thực chất là các nucleotide sẽ tương ứng với một vetor nhị phân với A=(1, 0, 0, 0), B(0, 1, 0, 0), C(0, 0, 1, 0), D(0, 0, 0, 1), sau đó ma trận này sẽ được nhân với ma trận chuyển đổi để tạo ra một vector n chiều là biểu diễn chính xác hơn cho siRNA. Và mới nhất là Fei He và cộng sự [29] đã biểu diễn định lượng cho siRNA trên cả hai khía cạnh định lượng và định tính. Nhóm đặc trưng định tính bao gồm, bao gồm tần số nucleotide (các motif 1, 2 và 3 nucleotide), hồ sơ ổn định nhiệt động lực học, nhiệt động lực học của sự tương tác siRNA-mRNA, và đặc trưng liên quan đến mRNA, là một biểu diễn hỗn hợp mới. Nhóm đặc trưng định lượng bao gồm các quy tắc thiết kế đã được thực nghiệm liên quan đến sự xuất hiện của một loại nucleotide tại vị trí xác định trong chuỗi siRNA được biểu diễn bởi tập giá trị nguyên (1, 0, -1), nếu quy tắc đó xuất hiện trong chuỗi siRNA và tương thích với siRNA hiệu quả cao (tác động dương tính) sẽ được mã hóa là 1, nếu quy tắc đó xuất hiện tương thích siRNA không hiệu quả cao (có tác động âm tính) sẽ được mã hóa là -1, nếu quy tắc đó không xuất hiện sẽ được mã hóa là 0. Như vậy cách biểu diễn siRNA góp phần gây ra hiệu quả dự đoán khác nhau ở các mô hình. Do đó, chương tiếp theo sẽ trình bày về một số cách biểu diễn RNA khác và phần thực nghiệm ở chương 4 sẽ đánh giá hiệu quả của các cách biểu diễn này khi kết hợp với một số phương pháp học máy cũng là đóng góp chính của nghiên cứu được trình bày trong luận văn này.

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
  - Very high: 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

Biểu diễn đồ họa của chuỗi DNA cung cấp một cách đơn giản để xem, phân loại và so sánh các cấu trúc gen khác nhau. Một phương thức biểu diễn đồ họa mới chiều hai chiều sử dụng một hệ 2 trục tọa độ Cartesian đã thu được từ việc biểu thị toán học của chuỗi DNA [31]. Biểu diễn đồ họa hai chiều giải quyết vấn đề suy thoái của chuỗi và được chứng minh là loại bỏ sự hình thành mạch. Cho trước x-chiếu và y-chiếu của bất kỳ điểm nào trên biểu diễn đồ họa, thì số lượng các nucleotide A, G, C, T từ đầu chuỗi cho tới điểm đó có thể tìm ra được.

Phương pháp này sử dụng các vector đơn vị để biểu diễn cho bốn nucleotide A, G, C, T như sau:

$$\left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right) \to A$$

$$\left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right) \to G$$

$$\left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right) \to C$$

$$t \to T$$

Để chứng minh không có mạch hoặc suy thoái trong biểu diễn đồ họa hai chiều, chúng ta giả thiết rằng:

(1) Số lượng nucleotide cấu thành mạch là n

(2) Số lượng các loại A, G, C, T trong một mạch tương ứng là a, g, c, t. Do đó a + g + c + t = n

Bởi vì aA, gG, cC và tT tạo thành một mạch nên sẽ thu được phương trình:

$$a\left(\frac{1}{2},-\frac{\sqrt{3}}{2}\right)+g\left(\frac{\sqrt{3}}{2},-\frac{1}{2}\right)+c\left(\frac{\sqrt{3}}{2},\frac{1}{2}\right)+t\left(\frac{1}{2},\frac{\sqrt{3}}{2}\right)=0$$

Tức là: $\begin{cases} a+\sqrt{3}g+\sqrt{3}c+t=0 \\ -\sqrt{3}a-g+c+\sqrt{3}t=0 \end{cases}$

Dễ nhận thấy điều kiện trên thoải mãn khi vào chỉ khi a = g = c = t = 0. Do đó n = 0, điều này có nghĩa là không tồn tại mạch trong biểu diễn đồ họa này.

Hơn nữa, nếu cho trước x chiếu và y chiếu của bất kỳ điểm p = (x, y) trong chuỗi ta có:

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

Tức là: $\begin{cases} a+\sqrt{3}g+\sqrt{3}c+t=2x \\ -\sqrt{3}a-g+c+\sqrt{3}t=2y \end{cases}$

Vì số lượng nucleotide là số nguyên nên ta dễ nhận thấy 2x, 2y là mẫu của các số có dạng $m+n\sqrt{3}$ trong đó m và n là các số nguyên. Sau khi xác định được duy nhất các số m_x, n, m_y, n_y từ 2x và 2y. Ta có thể dễ dàng tìm được các số a, g, c, t bằng cách giải hệ phương trình tuyến tính sau:

$$\begin{cases} a+t=m_x \\ g+c=n_x \\ -g+c=m_y \\ -a+t=n_y \end{cases}$$

Như vậy với tọa độ x, y của các điểm trên chuỗi ta có thể khôi phục được chuỗi DNA ban đầu duy nhất từ đồ họa DNA.

Tổng kết lại lại 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. Cách biểu diễn này khi áp dụng RNA thì sẽ thay thế uraxin (U) cho Thymine (T).

Ngoài các cách biểu diễn trên, một loạt các biểu diễn số học chuỗi DNA đã được tổng kết lại trong tài liệu [32] 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	$X_r(n)=\frac{\sqrt{2}}{3}[2T_n-C_n-G_n]$	$X_r(n)=\frac{\sqrt{2}}{3}[-1,-1,0,2]$	3

	$X_g(n) = \frac{\sqrt{6}}{3}[C_n - G_n]$ $X_b(n) = \frac{1}{3}[3A_n - T_n - C_n - G_n]$	$X_g(n) = \frac{\sqrt{6}}{3}[1, -1, 0, 0]$ $X_b(n) = \frac{1}{3}[-1, -1, 3, -1]$	
INTEGER	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 3: Tóm tắt các phương pháp biểu diễn số học cho chuỗi DNA

5. VOSS

Các biểu diễn VOSS ánh xạ các nucleotide A, C, G, T thành 4 chuỗi chỉ thị binary A_n, C_n, G_n, T_n thể hiện sự xuất hiện là 1 và không xuất hiện là 0 cho mỗi nucleotide tương ứng.

Như vậy, nếu ta ký hiệu S_n là chuỗi DNA bao gồm n nucleotide, thì chuỗi này sẽ được biểu diễn bằng 4 chuỗi chỉ thị binary tương ứng được ký hiệu lần lượt là A_n, C_n, G_n, T_n. Trong đó tại mỗi vị trí i của mỗi chuỗi:

- $A_i = 1$ nếu $S_i \equiv A$, ngược lại $A_i = 0$
- $C_i = 1$ nếu $S_i \equiv C$, ngược lại $C_i = 0$
- $G_i = 1$ nếu $S_i \equiv G$, ngược lại $G_i = 0$
- $T_i = 1$ nếu $S_i \equiv T$, ngược lại $T_i = 0$

6. TETRAHEDRON

Trong cách biểu diễn này, bốn chuỗi [A_n, C_n, G_n, T_n] trong biểu diễn VOSS được ánh xạ thành bốn đỉnh của một tứ diện thường sẽ làm giảm số chuỗi chỉ thị từ bốn còn ba nhưng theo cách đối xứng cho cả bốn chuỗi.

Ba chuỗi chỉ thị được ký hiệu lần lượt là x_r, x_g, x_b

$$X_r(n) = \frac{\sqrt{2}}{3}[2T_n - C_n - G_n]$$

$$X_g(n) = \frac{\sqrt{6}}{3}[C_n - G_n]$$

$$X_b(n) = \frac{1}{3}[3A_n - T_n - C_n - G_n]$$

7. INTEGER

- Cách biểu diễn này là một ánh xạ một chiều bằng cách ánh xạ các số {0, 1, 2, 3} đến các loại nucleotide là: A = 2, C = 1, G = 3, T = 0
- Với một chuỗi DNA có độ dài n sẽ được biểu diễn bằng vector n chiều tương ứng với mỗi vị trí nucleotide trong chuỗi DNA. Tại vị trí i (1, 2, …, n) trong vector n chiều:
  - Nếu A xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i là 2.
  - Nếu C xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i là 1
  - Nếu G xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i là 3

  - Nếu T xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i là 0
- Các ánh xạ này dẫn đến một cấu trúc các nucleotide như purine (A,G) > pyrimidine (C, T)
- Purines và Pyrimidines là những bazơ nitơ làm nên 2 loại bazo nucleotide khác nhau đó là DNA và RNA. Các bazo nito carbon 2 vòng Adenine và Guanine là Purine, trong khi đó các base một vòng Nito-Carbon như Thymine và Cytosine là Pyrimidines.

8. REAL

Cách biểu diễn số thực (REAL) cũng tương tự cách biểu diễn số nguyên (INTEGER) sẽ ánh xạ các số thực đến các nucleotide cụ thể: A = -1.5, T = 1.5, C = 0.5, G = -0.5

- Với một chuỗi DNA có độ dài n sẽ được biểu diễn bằng vector n chiều tương ứng với mỗi vị trí nucleotide trong chuỗi DNA. Tại vị trí i (1, 2, …, n) trong vector n chiều:
  - Nếu A xuất hiện tại vị trí i trong chuỗ iDNA thì giá trị tại chiều thứ i là -1.5.
  - Nếu C xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i là 0.5
  - Nếu G xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i là -0.5
  - Nếu T xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i là 1.5
- Các biểu diễn này thể hiện được tính chất bổ sung và hiệu quả trong việc tìm kiếm sợi bổ sung của chuỗi DNA
- Tuy nhiên việc gán một số thực tới một trong bốn bazo không nhất thiết phản ảnh cấu trúc hiện diện trong chuỗi DNA

9. COMPLEX

Phương pháp biểu diễn phức hợp thể hiện được tính chất bổ sung của các cặp A-T, C-G khi ánh xạ các nucleotide A = 1+j, C = -1+j, G = -1-j, T = 1-j

- Với một chuỗi DNA có độ dài n sẽ được biểu diễn bằng vector n chiều tương ứng với mỗi vị trí nucleotide trong chuỗi DNA. Tại vị trí i (1, 2, …, n) trong vector n chiều:
  - Nếu A xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là (1+j)

- Nếu C xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn (-1+j)
- Nếu G xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là (-1-j)
- Nếu T xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là (1-j)

10. QUATERNION

Phương pháp biểu diễn quaternion của các bazo trong DNA, quaternion thuần túy được gán cho các bazo với: A= i+j+k, C= i-j-k, G= -i-j+k, và T= -i+j-k

- Với một chuỗi DNA có độ dài n sẽ được biểu diễn bằng vector n chiều tương ứng với mỗi vị trí nucleotide trong chuỗi DNA. Tại vị trí i (1, 2, …, n) trong vector n chiều:
  - Nếu A xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là (i+j+k)
  - Nếu C xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn (i-j-k)
  - Nếu G xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là (-i-j+k)
  - Nếu T xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là (-i+j-k)

11. EIIP

Cơ sở của phương pháp biểu diễn này dựa trên năng lượng của các electron tự do dọc theo chuỗi DNA. Một chuỗi chỉ thị EIIP đơn được tạo thành từ việc thay thế EIIP của các nucleotide A = 0.1260, C = 0.1340, G = 0.0860 và T = 0.1335 trong chuỗi DNA.

- Với một chuỗi DNA có độ dài n sẽ được biểu diễn bằng vector n chiều tương ứng với mỗi vị trí nucleotide trong chuỗi DNA. Tại vị trí i (1, 2, …, n) trong vector n chiều:
  - Nếu A xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là 0.1260
  - Nếu C xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn 0.1340

- Nếu G xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là 0.0860
- Nếu T xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là 0.1335

12. ATOMIC NUMBER

Cơ sở của phương pháp biểu diễn này lại dựa trên số lượng nguyên tử của mỗi loại nucleotide. Theo cách này, một chuỗi chỉ thị atomic number đơn sẽ được tạo thành bằng cách gán số lượng nguyên tử của mỗi loại nucleotide là: A = 70, C = 58, G = 78, T = 66 trong chuỗi DNA.

- Với một chuỗi DNA có độ dài n sẽ được biểu diễn bằng vector n chiều tương ứng với mỗi vị trí nucleotide trong chuỗi DNA. Tại vị trí i (1, 2, …, n) trong vector n chiều:
  - Nếu A xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là 70
  - Nếu C xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn 58
  - Nếu G xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là 78
  - Nếu T xuất hiện tại vị trí i trong chuỗi DNA thì giá trị tại chiều thứ i sẽ được biểu diễn là 66

13. PAIRED NUMERIC

Trong biểu diễn paired numeric (số ghép cặp), các nucleotide (A-T, C-G) sẽ được ghép cặp theo cách bổ sung và các giá trị +1 và -1 sẽ được sử dụng tương ứng để biểu thị các cặp nucleotide A-T và C-G. Nó có thể được biểu diễn dưới dạng một hoặc hai chuỗi chỉ thị. Phương pháp biểu diễn này Biểu hiện này kết hợp chặt chẽ với thuộc tính cấu trúc của DNA với độ phức tạp giảm.

14. DNA WALK

Mô hình DNA-Walk cho thấy một đồ thị của một chuỗi DNA trong đó một bước được đưa lên trên (+1) nếu nucleotide là pyrimidin (C hoặc T) hoặc xuống dưới (-1) nếu nó là purine (A hoặc G). Đồ thị tiếp tục di chuyển lên và xuống dưới khi trình tự tiến hành với một cách thức một tích lũy, với số bazo của nó được thể

hiện dọc theo trục x. DNA Walk có thể được sử dụng như một công cụ để hình dung sự thay đổi trong sự tổ hợp các nucleotide, mô hình cặp base, và tiến hóa dọc theo trình tự ADN.

15. Z-CURVE

Đường cong Z-curve là một đường cong 3-D cung cấp cách biểu diễn duy nhất để hình dung và phân tích chuỗi DNA. Ba thành phần của đường cong Z-curve, $\{x_n, y_n, z_n\}$, biểu diễn ba phân bố nucleotide độc lập, mô tả đầy đủ một chuỗi DNA. Các thành phần x_n, y_n, z_n hiển thị tương ứng sự phân bố của purine so với pyrimidin (R so với Y), amino so với keto (M so với K), và liên kết Hiđro mạnh so với bazo liên kết Hiđro yếu (S so với W) dọc theo chuỗi.

CHƯƠNG 4. ĐÁNH GIÁ THỰC NGHIỆM CÁC MÔ HÌNH DỰ ĐOÁN KHẢ NĂNG ỨC CHẾ BỆNH 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: Hồi quy tuyến tính, Phân lớp (Naïve Bayes) và Kết hợp (thuật toán Apriori).

Trong đó. phương pháp hồi quy tuyến tính là phương pháp đơn giản và hiệu quả so sánh được với phần lớn các phương pháp khác. Đây là phương pháp được sử dụng nhiều nhất để xây dựng mô hình dự đoán trong các nghiên cứu đã được trình bày ở chương 2, nên tôi đã sử dụng phương pháp này để xây dựng mô hình dự đoán trong phần thực nghiệm chính của mình. Ngoài ra tôi cũng thực nghiệm phương pháp Naïve Bayes do phù hợp với lựa chọn đặc trưng không chắc chắn, và sử dụng thuật toán Apriori nhằm mong muốn tìm ra được được đặc trưng liên quan tới sự kết hợp của loại nucleotide và vị trí xuất hiện trong chuỗi siRNA. Tuy nhiên kết quả của thuật toán Apriori trong phần thực nghiệm chưa tìm được đặc trưng khả quan hơn, cũng như hiệu quả phân lớp Naïve Bayes còn thấp.

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 chương trình viết bằng Java thực thi 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.

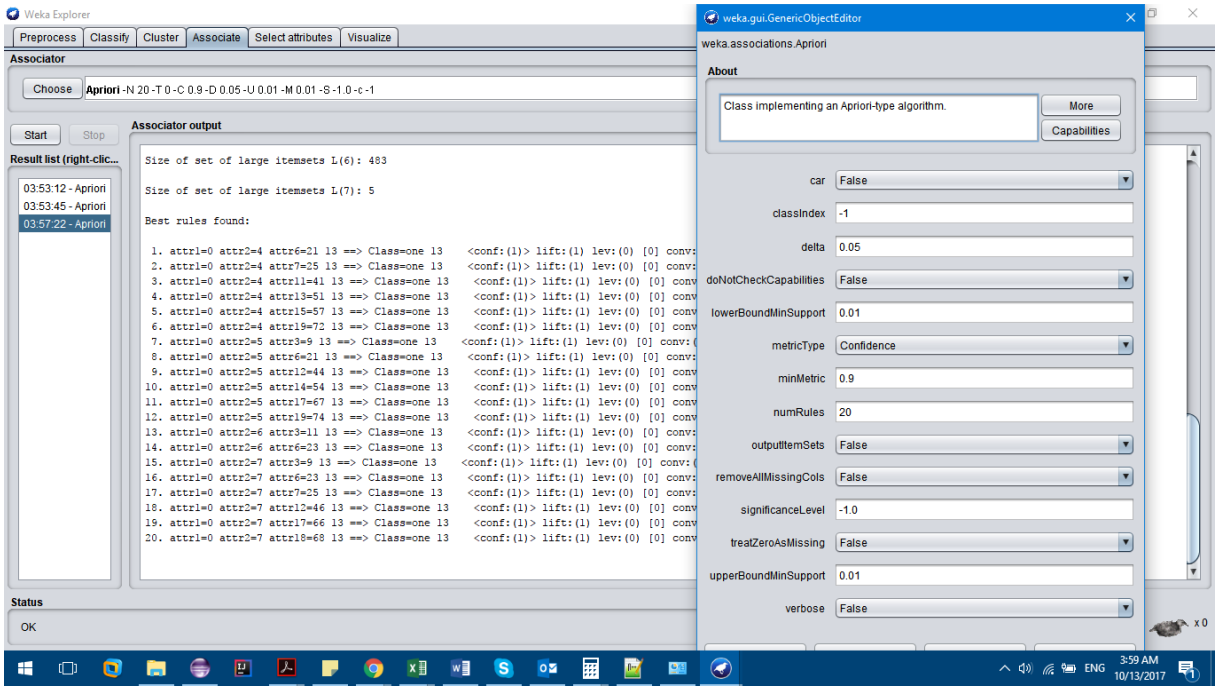
1. Thực nghiệm thuật toán kết hợp Apriori

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.



Hình 5: Chạy thuật toán Apriori (Association) trên weka 8.0

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 Phụ Lục, 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.

Ví dụ Trong tập “Low” có luật (A,0) (A,7) (A,8) có ý nghĩa là: những siRNA có A xuất hiện ở vị trí 0, A xuất hiện ở vị trí 7 và A xuất hiện ở vị trí 8 sẽ có khả năng ức chế bệnh thấp.

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.

2.1. 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:

```
=== Summary ===
```

```
Correctly Classified Instances        2443               65.4784 %
Incorrectly Classified Instances      1288               34.5216 %
Kappa statistic                          0.1457
Mean absolute error                      0.4146
Root mean squared error                  0.4687
Relative absolute error                 92.6332 %
Root relative squared error             99.0947 %
Total Number of Instances             3731

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area
PRC Area  Class
                 0.291     0.160     0.482       0.291     0.363       0.154     0.623
0.462     low
                 0.840     0.709     0.699       0.840     0.763       0.154     0.623
0.746     very_high
Weighted Avg.    0.655     0.523     0.626       0.655     0.628       0.154     0.623
0.650

=== Confusion Matrix ===

   a    b   <-- classified as
 367  894 |    a = low
 394 2076 |    b = very_high
```

2.2. 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:

```
=== Summary ===

Correctly Classified Instances        1418               56.2252 %
Incorrectly Classified Instances      1104               43.7748 %
Kappa statistic                          0.1245
Mean absolute error                      0.4486
Root mean squared error                  0.579
Relative absolute error                 89.7135 %
Root relative squared error            115.8078 %
Total Number of Instances             2522

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area
PRC Area  Class
                 0.514     0.389     0.569       0.514     0.540       0.125     0.582
0.577     low
                 0.611     0.486     0.557       0.611     0.582       0.125     0.582
0.560     very_high
Weighted Avg.    0.562     0.438     0.563       0.562     0.561       0.125     0.582
0.569

=== Confusion Matrix ===

   a   b   <-- classified as
 648 613 |   a = low
 491 770 |   b = very_high
```

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

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

- Sử dụng bộ dữ liệu siRecords lấy ra các siRNA có độ dài 19 nucleotide và chia thành 4 tập S-one, S-two, S-three, S-four tương ứng với khả năng ức chế lần lượt là “Low”, “Medium”, “High”, “Very High” của các siRNA.
- Thực hiện thống kê số lần xuất hiện của các bộ 1-merge, 2-merge, 3-merge trên 4 tập S-one, S-two, S-three, S-four và tính toán tần số xuất hiện của từng bộ trên mỗi tập. Với mỗi bộ, tổng các tần số trên cả 4 tập phải là 1.

- Tính toán biểu diễn dữ liệu cho các chuỗi siRNA cho các tập dữ liệu scored Dataset: Huesken_train, Huesken_test, Vicker, Reynolds, Uitei. Mỗi chuỗi siRNA có độ dài 19 nucleotide sẽ được biểu diễn bởi một vector 217 chiều (217 = 1 + 4 (19 + 18 + 17)). Chiều thứ nhất là score thể hiện khả năng ức chế bệnh của chuỗi siRNA, 216 chiều tiếp theo biểu diễn dữ liệu theo phương pháp thống kê tần số của các bộ 1-merge, 2-merge, 3-merge.
- Biểu diễn dữ liệu trên file arff đưa vào phần mềm Weka 3.8 để chạy thuật toán xây dựng và đánh giá mô hình.
- Sử dụng dữ liệu training là Huesken_train để training mô hình với thuộc tính score (thuộc tính thứ nhất) là mục tiêu.
- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.588
Mean absolute error                      0.1285
Root mean squared error                  0.1622
Relative absolute error                 79.2692 %
Root relative squared error             81.1968 %
Total Number of Instances             2182
```

Kết quả supplied test trên các tập dữ liệu còn lại, chỉ thống kê Correlation coefficient (hệ số tương quan)

Huesken19_train	**Huesken19_test**	**Reynolds**	**Utei**	**Vicker**
0.588	0.6137	0.5225	0.6641	0.5147

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

- Sử dụng bộ 80 rule và 38 rule thu được từ thực nghiệm phương pháp luật kết hợp sử dụng thuật toán Apriori để biểu diễn dữ liệu siRNA
- Mỗi bộ dữ liệu có 2 cho tới 3 nucleotide đi kèm với vị trí xuất hiện của nó trong chuỗi siRNA.
- Với bộ 80 rules, mỗi chuỗi siRNA có độ dài 19 nucleotide sẽ được biểu diễn bởi một vector 321 chiều. Với chiều thứ nhất là score của chuỗi siRNA, 320

chiều còn lại biểu diễn rule xuất hiện trong chuỗi. Với những rule không xuất hiện sẽ được điền giá trị 0.

- Tính toán biểu diễn dữ liệu cho các chuỗi siRNA cho các tập dữ liệu scored Dataset: Huesken_train, Huesken_test, Vicker, Reynolds, Uitei.
- Biểu diễn dữ liệu trên file arff đưa vào phần mềm Weka 3.8 để chạy thuật toán xây dựng và đánh giá mô hình.
- Sử dụng dữ liệu training là Huesken_train để training mô hình với thuộc tính score (thuộc tính thứ nhất) là mục tiêu.
- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train cho bộ 80 rules:

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.2482
Mean absolute error                      0.156
Root mean squared error                  0.1939
Relative absolute error                 96.2278 %
Root relative squared error             97.104  %
Total Number of Instances             2182
```

- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train cho bộ 38 rules

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.1626
Mean absolute error                      0.1595
Root mean squared error                  0.1975
Relative absolute error                 98.3752 %
Root relative squared error             98.8776 %
Total Number of Instances             2182
```

Kết quả supplied test trên các tập dữ liệu còn lại, chỉ thống kê Correlation coefficient (hệ số tương quan)

	Huesken19_train	Huesken19_test	Reynolds	Utei	Vicker
Bộ 80 rules	0.2482	0.214	0.0695	0.2548	0.1529
Bộ 38 rules	0.1626	0.115	0.1043	0.1219	0.1103

3.3. Phương pháp biểu diễn DNA không suy thoái

- Mỗi chuỗi siRNA độ dài 19 nucleotide sẽ được biểu diễn bởi một vector 39 chiều. Chiều thứ nhất là giá trị score của chuỗi siRNA đó, 38 chiều còn lại là biểu diễn DNA không suy thoái.
- Tính toán biểu diễn dữ liệu cho các chuỗi siRNA cho các tập dữ liệu scored Dataset: Huesken_train, Huesken_test, Vicker, Reynolds, Uitei.
- Biểu diễn dữ liệu trên file arff đưa vào phần mềm Weka 3.8 để chạy thuật toán xây dựng và đánh giá mô hình.
- Sử dụng dữ liệu training là Huesken_train để training mô hình với thuộc tính score (thuộc tính thứ nhất) là mục tiêu.
- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.6031
Mean absolute error                      0.1268
Root mean squared error                  0.1593
Relative absolute error                 78.2349 %
Root relative squared error             79.7662 %
Total Number of Instances             2182
```

Kết quả supplied test trên các tập dữ liệu còn lại, chỉ thống kê Correlation coefficient (hệ số tương quan)

Huesken19_train	Huesken19_test	Reynolds	Utei	Vicker
0.6031	N/A	0.5377	0.6205	0.588

3.4. VOSS

- Mỗi chuỗi siRNA độ dài 19 nucleotide sẽ được biểu diễn bởi một vector 77 chiều. Chiều thứ nhất là giá trị score của chuỗi siRNA đó, 76 chiều còn lại là biểu diễn VOSS.
- Tính toán biểu diễn dữ liệu cho các chuỗi siRNA cho các tập dữ liệu scored Dataset: Huesken_train, Huesken_test, Vicker, Reynolds, Uitei và ghi các biểu diễn ra file arff.
- Sử dụng dữ liệu training là Huesken_train để training mô hình với thuộc tính score (thuộc tính thứ nhất) là mục tiêu.
- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.6024
Mean absolute error                      0.1271
Root mean squared error                  0.1595
Relative absolute error                 78.4031 %
Root relative squared error             79.8555 %
Total Number of Instances             2182
```

Kết quả supplied test trên các tập dữ liệu còn lại, chỉ thống kê Correlation coefficient (hệ số tương quan)

Huesken19_train	Huesken19_test	Reynolds	Utei	Vicker
0.6024	0.6187	0.5394	0.6326	0.5668

3.5. TETRAHEDRON

- Mỗi chuỗi siRNA độ dài 19 nucleotide sẽ được biểu diễn bởi một vector 77 chiều. Chiều thứ nhất là giá trị score của chuỗi siRNA đó, 76 chiều còn lại là biểu diễn TETRAHEDRON.
- Tính toán biểu diễn dữ liệu cho các chuỗi siRNA cho các tập dữ liệu scored Dataset: Huesken_train, Huesken_test, Vicker, Reynolds, Uitei và ghi các biểu diễn ra file arff.
- Sử dụng dữ liệu training là Huesken_train để training mô hình với thuộc tính score (thuộc tính thứ nhất) là mục tiêu.
- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.6047
Mean absolute error                      0.1267
Root mean squared error                  0.1591
Relative absolute error                 78.1187 %
Root relative squared error             79.6736 %
Total Number of Instances             2182
```

Kết quả supplied test trên các tập dữ liệu còn lại, chỉ thống kê Correlation coefficient (hệ số tương quan)

Huesken19_train	Huesken19_test	Reynolds	Utei	Vicker
0.6047	0.6218	0.5471	0.6355	0.5681

3.6. INTEGER

- Mỗi chuỗi siRNA độ dài 19 nucleotide sẽ được biểu diễn bởi một vector 77 chiều. Chiều thứ nhất là giá trị score của chuỗi siRNA đó, 76 chiều còn lại là biểu diễn INTEGER.

- Tính toán biểu diễn dữ liệu cho các chuỗi siRNA cho các tập dữ liệu scored Dataset: Huesken_train, Huesken_test, Vicker, Reynolds, Uitei và ghi các biểu diễn ra file arff.
- Sử dụng dữ liệu training là Huesken_train để training mô hình với thuộc tính score (thuộc tính thứ nhất) là mục tiêu.
- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3663
Mean absolute error                      0.1477
Root mean squared error                  0.1858
Relative absolute error                 91.1151 %
Root relative squared error             93.0365 %
Total Number of Instances             2182
```

Kết quả supplied test trên các tập dữ liệu còn lại, chỉ thống kê Correlation coefficient (hệ số tương quan)

Huesken19_train	Huesken19_test	Reynolds	Utei	Vicker
0.3663	0.451	0.2993	0.2101	0.381

3.7. REAL

- Mỗi chuỗi siRNA độ dài 19 nucleotide sẽ được biểu diễn bởi một vector 77 chiều. Chiều thứ nhất là giá trị score của chuỗi siRNA đó, 76 chiều còn lại là biểu diễn REAL.
- Tính toán biểu diễn dữ liệu cho các chuỗi siRNA cho các tập dữ liệu scored Dataset: Huesken_train, Huesken_test, Vicker, Reynolds, Uitei và ghi các biểu diễn ra file arff.
- Sử dụng dữ liệu training là Huesken_train để training mô hình với thuộc tính score (thuộc tính thứ nhất) là mục tiêu.

- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.218
Mean absolute error                      0.1559
Root mean squared error                  0.195
Relative absolute error                 96.1335 %
Root relative squared error             97.6288 %
Total Number of Instances             2182
```

Kết quả supplied test trên các tập dữ liệu còn lại, chỉ thống kê Correlation coefficient (hệ số tương quan)

Huesken19_train	Huesken19_test	Reynolds	Utei	Vicker
0.218	0.2514	0.2036	0.0219	0.0846

3.8. EIIP

- Mỗi chuỗi siRNA độ dài 19 nucleotide sẽ được biểu diễn bởi một vector 77 chiều. Chiều thứ nhất là giá trị score của chuỗi siRNA đó, 76 chiều còn lại là biểu diễn EIIP.
- Tính toán biểu diễn dữ liệu cho các chuỗi siRNA cho các tập dữ liệu scored Dataset: Huesken_train, Huesken_test, Vicker, Reynolds, Uitei và ghi các biểu diễn ra file arff.
- Sử dụng dữ liệu training là Huesken_train để training mô hình với thuộc tính score (thuộc tính thứ nhất) là mục tiêu.
- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.3277
Mean absolute error                      0.1504
Root mean squared error                  0.1887
```

```
Relative absolute error                 92.7591 %
Root relative squared error             94.4762 %
Total Number of Instances              2182
```

Kết quả supplied test trên các tập dữ liệu còn lại, chỉ thống kê Correlation coefficient (hệ số tương quan)

Huesken19_train	Huesken19_test	Reynolds	Utei	Vicker
0.3277	0.405	0.2414	0.2569	0.2958

3.9. ATOMIC

- Mỗi chuỗi siRNA độ dài 19 nucleotide sẽ được biểu diễn bởi một vector 77 chiều. Chiều thứ nhất là giá trị score của chuỗi siRNA đó, 76 chiều còn lại là biểu diễn ATOMIC.
- Tính toán biểu diễn dữ liệu cho các chuỗi siRNA cho các tập dữ liệu scored Dataset: Huesken_train, Huesken_test, Vicker, Reynolds, Uitei và ghi các biểu diễn ra file arff.
- Sử dụng dữ liệu training là Huesken_train để training mô hình với thuộc tính score (thuộc tính thứ nhất) là mục tiêu.
- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.1427
Mean absolute error                      0.1592
Root mean squared error                  0.1978
Relative absolute error                 98.1929 %
Root relative squared error             99.0446 %
Total Number of Instances              2182
```

Kết quả supplied test trên các tập dữ liệu còn lại, chỉ thống kê Correlation coefficient (hệ số tương quan)

Huesken19_train	Huesken19_test	Reynolds	Utei	Vicker
0.1427	0.1125	0.127	0.1659	0.1081

3.10. DNA WALKER

- Mỗi chuỗi siRNA độ dài 19 nucleotide sẽ được biểu diễn bởi một vector 77 chiều. Chiều thứ nhất là giá trị score của chuỗi siRNA đó, 76 chiều còn lại là biểu diễn DNA WALKER.
- Tính toán biểu diễn dữ liệu cho các chuỗi siRNA cho các tập dữ liệu scored Dataset: Huesken_train, Huesken_test, Vicker, Reynolds, Uitei và ghi các biểu diễn ra file arff.
- Sử dụng dữ liệu training là Huesken_train để training mô hình với thuộc tính score (thuộc tính thứ nhất) là mục tiêu.
- Kết quả xây dựng mô hình khi chạy bằng Weka 3.8 trên tập Huesken_train

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.341
Mean absolute error                      0.1525
Root mean squared error                  0.1878
Relative absolute error                 94.065  %
Root relative squared error             94.0161 %
Total Number of Instances              218
```

Kết quả supplied test trên các tập dữ liệu còn lại, chỉ thống kê Correlation coefficient (hệ số tương quan)

Huesken19_train	Huesken19_test	Reynolds	Utei	Vicker
0.341	0.3003	0.3448	0.4688	0.2594

3.11. Kết hợp các phương pháp biểu diễn khác nhau

Ngoài thực nghiệm từng phương pháp biểu diễn, 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 4: 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

4.1. Tóm tắt kết quả thực nghiệm

- Các biểu diễn có hệ số tương quan cao nhất:
  - TetraHedron(R=0.6047/Huesken_train)
  - VOSS+2-merge (R=0.6244/Huesken_test)
  - VOSS+2-merge+3-merge (R=0.5486/Reynolds, R=0.6772/Ui-tei)

  - Biểu diễn Yau (R=0.588/Vicker).
- Các phương pháp biểu diễn số học với số chiều biểu diễn thấp 20 cho kết quả kém (Integer, Real, EIIP, Atomic, DNA Walker). Nguyên nhân do cách biểu diễn quá đơn giản chỉ phụ thuộc vào loại nucleotide và không xét đến đặc tính trình tự chuỗi siRNA và quan hệ ràng buộc giữa các nucleotide hoặc vị trí của nucleotide trong chuỗi.
- Việc áp dụng luật kết hợp để tìm ra những bộ nucleotide có khả năng đại diện cho các tập con của labled dataset (low, medium, high, very high) chưa đạt kết quả mong muốn nên xuất hiện nhiều siRNA trong dataset không khớp với rule nào dẫn tới kết quả thấp.

4.2. Đánh giá

- Dựa trên kết quả thực nghiệm, mô hình biểu diễn kết hợp "VOSS+2-merge+3-merge" với 217 chiều được coi là phương pháp biểu diễn tốt nhất trong số các phương pháp biểu diễn đã được giới thiệu với hệ số tương quan lần lượt 0.5838 trên tập Huesken train , 0.6168 trên tập Huesken test, 0.5486 trên tập Reynolds, 0.6772 trên Ui-tei, 0.515 trên tập Vicker.
- Nhìn chung các kết quả thực nghiệm được chỉ tương đương với các mô hình dự đoán đã có, thậm chí thấp hơn rõ rệt đối với một số mô hình dự đoán đề xuất gần đây như BiLTR (BN Thăng, 2015), siRNApred(Ye Han et al, 2017), Fei He's method (Fei He et al, 2017). 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.
  - Hơn nữa 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.
  - Đặc tính liên quan tới tính chất nhiệt động học của siRNA, tương tác nhiệt động học siRNA-mRNA và đặc điểm liên quan tới mRNA chưa được biểu diễn.

KẾT LUẬN

Các công việc đã thực hiện trong luận văn của tôi có đóng góp quan trọng nhất là kiểm chứng được hiệu quả của các phương pháp biểu diễn RNA đối với việc dự đoán khả năng ức chế bệnh của siRNA và cung cấp một số thông tin khác liên quan đến khả năng ức chế bệnh của RNA. Thứ nhất, bài luận đã cung cấp được những kiến thức cơ bản về khả năng ức chế bệnh của RNA. Thứ hai, tổng hợp được một số các phương pháp nghiên cứu theo hai hướng tiếp cận sinh học và tin sinh học để giải quyết bài toán đã đặt ra. Thứ ba, trình bày các phương pháp biểu diễn đã được giới thiệu bởi các nhà nghiên cứu khác và ba phương pháp biểu diễn mới. Thứ tư, thực nghiệm mô hình dự đoán khả năng ức chế bệnh của siRNA theo các phương pháp biểu diễn khác nhau.

Trong công việc này, giảng viên hướng dẫn của tôi đã đề xuất phương pháp biểu diễn dựa vào thống kê tần số căn cứ vào các đặc tính về trình tự và số lần xuất hiện của các bộ thứ tự nucleotide trong chuỗi siRNA. Kết quả từ quá trình thực nghiệm của phương pháp biểu diễn này cũng như các phương pháp biểu diễn khác khi kết hợp với các phương pháp xây dựng mô hình dự đoán 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 các thử nghiệm trước đó và thấp hơn so với công bố năm 2017 của nhóm nghiên cứu Fei He và Ye Han một phần 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 diện 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 cạnh 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 tốt hơn đối với việc dự đoán khả năng ức chế bệnh của siRNA.

TÀI LIỆU THAM KHẢO

1 Montgomery, Mary K: "RNA Interference - RNA Interference, Editing, and Modification: Methods and Protocols", *Methods in Molecular Biology*,3-21, 2010.

2 slideshare.net, https://www.slideshare.net/mariyazaman58/role-of-antisense-and-rnaibased-gene-silencing-in-crop-improvement

3 Nobelprize.org, "The Nobel Prize in Physiology or Medicine 2006"

4 Neema Agrawal, P. V. N. Dasaradhi, Asif Mohmmed, Pawan Malhotra, Raj K. Bhatnagar, and Sunil K. Mukherjee*: "RNA Interference: Biology, Mechanism, and Applications", *Microbiol Mol Biol Rev*, 67(4):657-85, 2003.

5 Sayda M. Elbashir, Winfried Lendeckel and Thomas Tuschl: "RNA interference is mediated by 21- and 22-nucleotide RNAs", *Genes Dev*, 15:188–200, 2001.

6 Angela Reynolds, Devin Leake, Queta Boese, Stephen Scaringe, William S Marshall, Anastasia Khvorova: "Rational siRNA design for RNA interference", *Nat Biotechnol*, 22:326–30, 2004.

7 Chalk AM, Wahlestedt C, Sonnhammer EL: "Improved and automated prediction of effective siRNA", *Biochem Biophys Res Commun*, 319(1):264–74, 2004.

8 Amarzguioui M, Prydz H: "An algorithm for selection of functional siRNA sequences", *Biochem Biophys Res Commun*, 316:1050–8, 2004.

9 Ui-Tei K, Naito Y, Takahashi F, Haraguchi T, Ohki–Hamazaki H, Juni A, et al: "Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference", *Nucleic Acids Res*, 32:936–48, 2004.

10 Hsieh AC, Bo R, Manola J, et al: "A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens", *Nucleic Acids Res,* 32:893–901, 2004.

11 Jagla B, Aulner N, Kelly PD, Song D, Volchuk A, Zatorski A, et al: "Sequence characteristics of functional siRNAs", *RNA,* 11:864–72, 2005.

12 Lisa J Scherer, John J Rossi: "Approaches for the sequence-specific knockdown of mRNA", *Nat Biotechnol,* 21:1457–65, 2003.

13 Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD: "Asymmetry in the Assembly of the RNAi Enzyme Complex", *Cell,* 115(2):199–208, 2003.

14 Khvorova A, Reynolds A, Jayasena SD: "Functional siRNAs and miRNAs Exhibit Strand Bias", *Cell,* 115:209–16, 2003.

15 Ren Y, Gong W, Xu Q, Zheng X, Lin D, Wang Y, et al: "siRecords: an extensive database of mammalian siRNAs with efficacy ratings", *Bioinformatics,* 22:1027–8, 2006.

16 Gong W, Ren Y, Xu Q, Wang Y, Lin D, Zhou H, et al: "Integrated siRNA design based on surveying of features associated with high RNAi effectiveness", *BMC Bioinf,* 7:516, 2006.

17 Bui Ngoc Thang, Tu Bao Ho and Tatsuo Kanda: "A semi–supervised tensor regression model for siRNA efficacy prediction", *BMC Bioinformatics,* 2015.

18 Huesken D, Lange J, Mickanin C, Weiler J, Asselbergs F, Warner J, et al: "Design of a genome-wide siRNA library using an artificial neural network", *Nat Biotechnol,* 23:955–1001, 2005.

19 Shabalina SA, Spiridonov AN, Ogurtsov AY: "Computational models with thermodynamic and composition features improve siRNA design", *BMC Bioinf,* 7:65, 2006.

20 Vert JP, Foveau N, Lajaunie C, Vandenbrouck Y: "An accurate and interpretable model for siRNA efficacy prediction", *BMC Bioinf,* 7:520, 2006.

21 Ichihara M, Murakumo Y, Masuda A, Matsuura T, Asai N, Jijiwa M, et al: "Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities", *Nucleic Acids Res,* e123:35, 2007.

22 Matveeva O, Nechipurenko Y, Rossi L, Moore B, Ogurtsov AY, Atkins JF, et al: "Comparison of approaches for rational siRNA design leading to a new efficient and transparent method", *Access,* 35:1–10, 2007.

23 Qiu S, Lane T: "A Framework for Multiple Kernel Support Vector Regression and Its Applications to siRNA Efficacy Prediction", *IEEE/ACM Trans Comput Biol Bioinform,* 6:190–9, 2009.

24 Klingelhoefer JW, Moutsianas L, Holmes CC: "Approximate Bayesian feature selection on a large meta-dataset offers novel insights on factors that effect siRNA potency", *Bioinformatics,* 25:1594–601, 2009.

25 Sciabola S, Cao Q, Orozco M, Faustino I, Stanton RV: "Improved nucleic acid descriptors for siRNA efficacy prediction", *Nucl Acids Res,* 41:1383–94, 2012.

26 Qi L, Han Z, Ruixin Z, Ying X, Zhiwei C: "Reconsideration of in silico siRNA design from a perspective of heterogeneous data integration: problems and solutions", *Brief Bioinform,* 15:292–305, 2012.

27 Mysara M, Elhefnawi M, Garibaldi JM: "MysiRNA: Improving siRNA efficacy prediction using a machine-learning model combining multi-tools and whole stacking energy", *J Biomed Inform,* 45:528–34, 2012.

28 Chang PC, Pan WJ, Chen CW, Chen YT, Chu YW: "A design engine of siRNA that integrates SVMs prediction and feature filters", *Biocatal Agric Biotechnol*, 1:128–34, 2012.

29 Fei He, Ye Han, Jianting Gong, Jiazhi Song, Han Wang and Yanwen Li: "Predicting siRNA efficacy based on multiple selective siRNA representations and their combination at score level", *Scientific Reports 7, Article number 44836*, 2017.

30 Ye Han, Yuanning Liu, Hao Zhang, Fei He, et al: "Utilizing Selected Di- and Trinucleotides of siRNA to Predict RNAi Activity", *Computational and Mathematical Methods in Medicine, Volume 2017 (2017), Article ID 5043984*, 2017.

31 Stephen S. -T. Yau*, Jiasong Wang1, Amir Niknejad, Chaoxiao Lu, Ning Jin1: "DNA sequence representation without degeneracy", *Nucleic Acids Research*, 31:3078–3080, 2003.

32 Hon Keung Kwan, Swarna Bai Arniker: "Numerical Representation of DNA Sequences", *IEEE International Conference on Electro/Information Technology*, 307-310, 2009.

PHỤ LỤC

1. 80 luật kết hợp đầy đủ

STT	Rule	S-one	S-two	S-three	S-four
1	(A,0) (A,2) (C,4)	25	18	31	41
2	(A,0) (A,2) (A,5)	25	20	45	42
3	(A,0) (A,7) (A,8)	25	16	22	20
4	(A,0) (A,7) (C,14)	25	13	9	20
5	(A,0) (G,9) (C,18)	25	12	23	10
6	(A,0) (A,12) (C,18)	25	6	18	27
7	(A,0) (C,13) (U,14)	25	12	31	36
8	(A,0) (C,13) (U,15)	25	14	27	30
9	(A,0) (C,15) (C,18	25	14	18	20
10	(C,0) (A,1) (A,2)	25	18	52	47
11	(A,0) (A,1) (C,5)	13	13	25	20
12	(A,0) (A,1) (C,6)	13	17	19	32
13	(A,0) (A,1) (C,10)	13	11	22	29
14	(A,0) (A,1) (U,12)	13	14	18	14
15	(A,0) (A,1) (C,14)	13	9	18	15
16	(A,0) (A,1) (A,18)	13	15	36	36
17	(A,0) (C,1) (C,2)	13	12	18	19
18	(A,0) (C,1) (C,5)	13	14	15	6
19	(A,0) (C,1) (A,11)	13	13	34	25
20	(A,0) (C,1) (G,13)	13	15	26	20
21	(U,0) (U,8)	54	25	55	64
22	(A,0) (G,1) (U,12)	21	25	42	32
23	(A,0) (A,2) (U,9)	20	25	40	41
24	(A,0) (A,2) (U,18)	18	25	29	29
25	(A,0) (A,3) (A,7)	21	25	16	21
26	(A,0) (C,7) (U,9)	21	25	31	18
27	(A,0) (U,9) (U,12)	17	25	28	24
28	(A,0) (U,12) (U,18)	14	25	22	21
29	(A,0) (A,17) (U,18)	17	25	28	24
30	(C,0) (A,1) (A,5)	26	25	54	47
31	(A,0) (A,1) (G,3)	19	13	26	30
32	(A,0) (A,1) (C,5)	13	13	25	20
33	(A,0) (A,1) (A,9)	18	13	12	30
34	(A,0) (A,1) (C,9)	12	13	15	22
35	(A,0) (A,1) (U,9)	12	13	23	22
36	(A,0) (A,1) (G,10)	17	13	17	19
37	(A,0) (A,1) (A,11)	18	13	15	29
38	(A,0) (A,1) (U,11)	20	13	21	28
39	(A,0) (A,1) (G,12)	10	13	16	26
40	(A,0) (A,1) (A,13)	23	13	15	30

41	(A,0) (G,1) (A,18)	17	17	49	56
42	(A,0) (A,11) (A,18)	11	17	49	34
43	(A,0) (A,12) (A,17)	18	12	49	26
44	(A,0) (A,14) (A,18)	13	19	49	37
45	(A,0) (A,17) (A,18)	14	17	49	48
46	(C,0) (A,1) (U,9)	17	28	49	57
47	(C,0) (A,1) (G,13)	23	16	49	45
48	(C,0) (C,1) (G,5)	17	19	49	29
49	(C,0) (C,1) (A,6)	25	28	49	46
50	(C,0) (C,1) (C,7)	30	25	49	32
51	(A,0) (A,1) (C,5)	13	13	25	20
52	(A,0) (A,1) (G,6)	24	12	25	26
53	(A,0) (A,1) (A,12)	24	14	25	27
54	(A,0) (A,1) (G,13)	12	15	25	27
55	(A,0) (C,1) (U,9)	14	20	25	34
56	(A,0) (C,1) (A,17)	12	14	25	16
57	(A,0) (A,2) (A,4)	17	14	25	22
58	(A,0) (A,2) (U,7)	15	10	25	24
59	(A,0) (A,2) (G,9)	23	12	25	22
60	(A,0) (A,2) (C,11)	10	18	25	24
61	(C,0) (A,1) (G,5)	17	16	31	49
62	(C,0) (C,1) (A,11)	28	31	38	49
63	(C,0) (U,1) (A,18)	21	28	43	49
64	(C,0) (A,2) (U,14)	20	19	44	49
65	(C,0) (C,3) (U,4)	21	24	38	49
66	(C,0) (C,3) (A,7)	14	26	57	49
67	(C,0) (C,3) (U,9)	17	27	45	49
68	(C,0) (A,4) (G,5)	15	19	48	49
69	(C,0) (A,5) (C,6)	21	27	47	49
70	(C,0) (A,5) (C,8)	19	12	30	49
71	(A,0) (A,1) (A,6)	15	12	19	25
72	(A,0) (A,1) (G,9)	22	10	27	25
73	(A,0) (C,1) (A,11)	13	13	34	25
74	(A,0) (G,1) (A,4)	17	16	26	25
75	(A,0) (G,1) (C,13)	22	17	36	25
76	(A,0) (G,1) (G,14)	18	14	39	25
77	(A,0) (A,2) (G,15)	18	26	20	25
78	(A,0) (A,2) (G,17)	12	11	15	25
79	(A,0) (C,2) (G,9)	13	14	13	25
80	(A,0) (C,2) (A,13)	9	7	19	25

2. 38 luật kết hợp sau khi filter với tần số lớn hơn hoặc bằng 30%

STT	Rule	S-one	S-two	S-three	S-four
1	(A,0) (A,7) (A,8)	25	16	22	20
2	(A,0) (A,7) (C,14)	25	13	9	20
3	(A,0) (G,9) (C,18)	25	12	23	10
4	(A,0) (A,12) (C,18)	25	6	18	27
5	(A,0) (C,15) (C,18)	25	14	18	20
6	(A,0) (A,3) (A,7)	21	25	16	21
7	(A,0) (U,12) (U,18)	14	25	22	21
8	(A,0) (G,1) (A,18)	17	17	49	56
9	(A,0) (A,11) (A,18)	11	17	49	34
10	(A,0) (A,12) (A,17)	18	12	49	26
11	(A,0) (A,14) (A,18)	13	19	49	37
12	(A,0) (A,17) (A,18)	14	17	49	48
13	(C,0) (A,1) (U,9)	17	28	49	57
14	(C,0) (A,1) (G,13)	23	16	49	45
15	(C,0) (C,1) (G,5)	17	19	49	29
16	(C,0) (C,1) (A,6)	25	28	49	46
17	(C,0) (C,1) (C,7)	30	25	49	32
18	(A,0) (A,1) (C,5)	13	13	25	20
19	(A,0) (A,1) (G,13)	12	15	25	27
20	(A,0) (C,1) (A,17)	12	14	25	16
21	(A,0) (A,2) (A,4)	17	14	25	22
22	(A,0) (A,2) (U,7)	15	10	25	24
23	(A,0) (A,2) (G,9)	23	12	25	22
24	(A,0) (A,2) (C,11)	10	18	25	24
25	(C,0) (A,1) (G,5)	17	16	31	49
26	(C,0) (C,1) (A,11)	28	31	38	49
27	(C,0) (U,1) (A,18)	21	28	43	49
28	(C,0) (A,2) (U,14)	20	19	44	49
29	(C,0) (C,3) (U,4)	21	24	38	49
30	(C,0) (C,3) (A,7)	14	26	57	49
31	(C,0) (C,3) (U,9)	17	27	45	49
32	(C,0) (A,4) (G,5)	15	19	48	49
33	(C,0) (A,5) (C,6)	21	27	47	49
34	(C,0) (A,5) (C,8)	19	12	30	49
35	(A,0) (A,1) (A,6)	15	12	19	25
36	(A,0) (A,2) (G,17)	12	11	15	25
37	(A,0) (C,2) (G,9)	13	14	13	25
38	(A,0) (C,2) (A,13)	9	7	19	25

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ
------🕮------

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

Hà Nội, ngày 02 tháng 12 năm 2017

QUYẾT NGHỊ
CỦA HỘI ĐỒNG CHẤM LUẬN VĂN THẠC SĨ

Căn cứ Quyết định số **1162/QĐ-ĐT, ngày 23 tháng 11 năm 2017** của Hiệu trưởng trường Đại học Công nghệ về việc thành lập Hội đồng chấm luận văn thạc sĩ của học viên **Phạm Thị Mai Hoa,** Hội đồng chấm luận văn Thạc sĩ đã họp vào **11h, thứ 7, ngày 02 tháng 12 năm 2017, Phòng 212, Nhà E3,** Trường Đại học Công nghệ - ĐHQGHN.

Tên đề tài luận văn: **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:

Sau khi nghe học viên trình bày tóm tắt luận văn Thạc sĩ, các phản biện đọc nhận xét, học viên trả lời các câu hỏi, Hội đồng đã họp, trao đổi ý kiến và thống nhất kết luận:

1. Về tính cấp thiết, tính thời sự, ý nghĩa lý luận và thực tiễn của đề tài luận văn:

Luận văn có tính thực tiễn ứng dụng cao

2. Về bố cục, phương pháp nghiên cứu, tài liệu tham khảo, ... của luận văn:

Phương pháp nghiên cứu đáng tin cậy, có tham khảo các bài báo khoa học

3. Về kết quả nghiên cứu:

Tìm hiểu khả năng ức chế bệnh của RNA
Tìm hiểu các hướng nghiên cứu khả năng ức chế của RNA
Tìm hiểu cách thức biểu diễn RNA
Thực nghiệm mô hình dự đoán khả năng ức chế của siRNA theo các biểu diễn dữ liệu khác nhau

4. **Hạn chế của luận văn** *(nếu có)*:

Nội dung chương 1, 2 quá sơ sài so với bài toán được giải quyết trong luận văn
Chưa mô tả rõ bài toán được giải quyết trong luận văn
Chưa phân tích lý do lựa chọn kỹ thuật trong luận văn
Cách đánh chỉ số mục, chương chưa đúng
Chú ý sử dụng chính xác thuật ngữ "ức chế bệnh"
Chưa có phần tiết, đánh giá kết quả thu được

5. **Đánh giá chung và kết luận**:

Luận văn đạt yêu cầu của luận văn cao học chuyên ngành HTTT

Luận văn đạt 8,3 / 10 điểm. Quyết nghị này được 05 / 05 thành viên của Hội đồng nhất trí thông qua.

THƯ KÝ HỘI ĐỒNG

TS. Nguyễn Thu Hậu

CHỦ TỊCH HỘI ĐỒNG

PGS. TS. Nguyễn Trí Thành

XÁC NHẬN CỦA CƠ SỞ ĐÀO TẠO

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

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

NHẬN XÉT PHẢN BIỆN LUẬN VĂN THẠC SỸ

Họ tên học viên: Phạm Thị Mai Hoa
Đề tài luận văn: ***"Các phương pháp dự đoán khả năng ức chế bệnh dựa trên các biểu diễn khác nhau của RNA và ứng dụng"***
Chuyên ngành: Hệ thống thông tin Mã số: **8480104** (2017)
Họ tên người nhận xét: Hà Quang Thụy
Học hàm, học vị: PGS. TS. Chuyên ngành: Hệ thống thông tin
Cơ quan công tác: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Luận văn dài 77 trang với bốn chương nội dung là Chương 1 *"Giới thiệu về khả năng ức chế bệnh của RNA"* (trang 12-22), Chương 2 *"Các hướng nghiên cứu khả năng ức chế của RNA"* (trang 23-40), Chương 3 *"Các cách thức biểu diễn RNA"* (trang 41-51), Chương 4 *"Đánh giá thực nghiệm các mô hình dự đoán khả năng ức chế của siRNA theo các biểu diễn dữ liệu khách nhau"* (trang 52-68). Luận văn còn một phu lục gồm hai danh sách 80 luật kết hợp đầy đủ (trang 75-76), 38 luật kết hợp sau khi lọc với tần số không nhỏ thua 30% (trang 77).

NHẬN XÉT

1. Về đề tài luận văn

- Đề tài luận văn *"Các phương pháp dự đoán khả năng ức chế bệnh dựa trên các biểu diễn khác nhau của RNA và ứng dụng"* đề cập tới chủ đề xây dựng các biểu diễn RNA trong phân lớp xâu RNA vào một trong bốn lớp năng lực ức chế bệnh là có ý nghĩa khoa học và thực tiễn.
- Đề tài luận văn phù hợp với chuyên ngành đào tạo Hệ thống thông tin (mã số **8480104**), trong đó, nội dung luận văn đề cập tới các kỹ thuật ứng dụng khai phá dữ liệu trong sinh học,
- Về cơ bản, nội dung luận văn phù hợp với tên đề tài luận văn.

2. Về độ tin cậy

- Nội dung hai chương 3, 4 và các tài liệu tham khảo (TLTK) liên quan hai chương này có điểm khác biệt so với các luận văn Thạc sỹ trong nước và thể hiện một độ tin cậy nhất định. Hai chương 1,2 đề cập tới vấn đề quá rộng so với nội dung nghiên cứu thực chất của luận văn, do đó, chúng chứa các yếu tố chưa tin cậy.
- Tài liệu tham khảo được mô tả tương đối phù hợp, tuy nhiên, không cần sử dụng quá nhiều TLTK đề cập rất ít tới các nội dung chính của luận văn. Tham chiếu TLTK tương đối phù hợp.

3. Về kết quả và hạn chế

3.1. Kết quả

- Trình bày được bốn phương pháp biểu diễn RNA liên quan tới đoán nhận khả năng ức chế bệnh của RNA, đồng thời, giới thiệu 11 phương pháp biểu diễn RNA ít phổ biến hơn (đặc điểm của 11 phương này được trình bày tại Bảng 3),
- Trình bày được giải pháp biểu diễn dữ liệu RNA phù hợp với các phương pháp biểu diễn RNA tưng ứng và sử dụng ba thành phần trong công cụ WEKA tiến hành thực nghiệm trên bộ dữ liệu Labeled Datasets như sau:

- Sử dụng thành phần Apriori với ràng buộc 20 luật kết hợp cho mỗi mức ức chế, thu được 80 luật kết hợp cho toàn bộ 4 mức ức chế. Bổ sung mức lọc độ hỗ trợ 30%, luận văn thu được 38 luật kết hợp.
- Sử dụng thành phần phân lớp Naïve Bayes trên tập mẫu có nhãn Low hoặc Very High theo hai biểu diến và hiển thị kết quả đánh giá phân lớp,
- Sử dụng thành phần phân lớp Hồi quy tuyến tính theo 11 biểu diễn và hiện thi kết quả đánh giá phân lớp.

3.2. Hạn chế

- Phát biểu chưa tường minh (đầu vào, đầu ra và hướng giải pháp) bài toán cần giải quyết trong luận văn là khảo sát các giải pháp biểu diễn dữ liệu RNA và hiệu năng của các giải pháp dữ liệu này này trong bài toán phân lớp khả năng ức chế bệnh của RNA. Đây là nguyên nhân chính làm cho chương 1,2 đề cập tới các nội dung vượt quá tầm của một luận văn Thạc sỹ,
- Đã nắm bắt được các giải pháp biểu diễn dữ liệu RNA và biết sử dụng công cụ WEKA, tuy nhiên, mức độ nắm bắt của học viên mới ở mức triển khai kỹ thuật mà chưa đạt mức độ giải thích được lý do sử dụng các giải pháp biểu diễn đó cũng như phân tích được các kết quả phân lớp.
- Luận văn còn các lỗi trình bày, chẳng hạn, chỉ số mục không theo quy định hoặc một số tiêu đề mục có "*ức chế*" mà không là "*ức chế bệnh*", v.v.

4. Câu hỏi cho học viên

- Phát biểu chính xác bài toán được giải quyết trong luận văn.
- Hai bộ luật kết hợp kết quả thực nghiệm dùng để làm gì?

5. KẾT LUẬN

- Tuy còn cần phải chỉnh sửa về bố cục và loại bỏ lỗi, luận văn "*Các phương pháp dự đoán khả năng ức chế bệnh dựa trên các biểu diễn khác nhau của RNA và ứng dụng*" của Học viên Phạm Thị Mai Hoa đáp ứng yêu cầu của một luận văn Thạc sỹ chuyên ngành HTTT mã số **8480104**.
- Luận văn đủ điều kiện được đưa ra bảo vệ tại Hội đồng chấm luận văn Thạc sỹ chuyên ngành HTTT.

Hà nội, ngày 01 tháng 12 năm 2017

Người nhận xét

PGS.TS. Hà Quang Thụy

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

BẢN NHẬN XÉT PHẢN BIỆN LUẬN VĂN THẠC SĨ

Họ và tên cán bộ phản biện: Bùi Thu Lâm

Học hàm, học vị: PGS TS

Chuyên ngành: CNTT

Cơ quan công tác: Học viện KTQS

Họ và tên học viên cao học: Phạm Thị Mai Hoa

Tên đề tài luận văn: Các phương pháp dự đoán khả năng ức chế bệnh dựa trên các biểu diễn khác nhau của RNA và ứng dụng

Chuyên ngành: HTTT Mã số: 60480104

Ý KIẾN NHẬN XÉT

- **Tính cấp thiết, tính thời sự, ý nghĩa khoa học và thực tiễn của đề tài luận văn:**

  Chúng ta đang sống trong giai đoạn bùng nổ thông tin. Công nghệ dữ liệu ngày càng có những tiến bộ đáng kể. Dựa trên các kho dữ liệu, các nhà nghiên cứu đã xây dựng nhiều công cụ để phân tích khám phá tri thức. Lĩnh vực tin sinh học cũng là nơi có nhiều dữ liệu và rất cần các công cụ phân tích và khai phá dữ liệu, đồng thời rất cần thiết các mô hình toán học để mô tả các mối quan hệ giữa các đối tượng sinh học, bài toán dự đoán khả năng ức chế bệnh dựa trên các biểu diễn khác nhau của RNA là một trong những ví dụ điển hình. Chính vì vậy, đề tài có tính cấp thiết và nhiều ý nghĩa khoa học.

- **Sự không trùng lặp của đề tài nghiên cứu so với các công trình khoa học, luận văn đã công bố ở trong và ngoài nước; tính trung thực, rõ ràng và đầy đủ trong trích dẫn tài liệu tham khảo.**

  Đóng góp của tác giả phần lớn là tìm hiểu công nghệ, cài đặt và triển khai thí nghiệm. Tác giả đã bỏ nhiều công sức trong thu thập và tổng hợp thông tin, triển khai thí nghiệm có tính hệ thống, kết nối các chương. Chính vì vậy, đề tài cơ bản là không trùng lặp với các công trình khác. Việc trích dẫn tài liệu cơ bản là phù hợp.

- **Sự phù hợp giữa tên đề tài với nội dung nghiên cứu cũng như với chuyên ngành và mã số đào tạo**

Phù hợp.

- **Độ tin cậy và tính hiện đại của phương pháp nghiên cứu đã sử dụng để hoàn thành luận văn**

  Đáp ứng theo yêu cầu.

- **Kết quả nghiên cứu mới của tác giả, đóng góp mới cho sự phát triển chuyên ngành, đóng góp mới phục vụ sản xuất, kinh tế, xã hội, an ninh, quốc phòng và đời sống. Giá trị và độ tin cậy của những kết quả nghiên cứu**

  Nội dung luận văn có tính mới không cao. Tác giả cố gắng tìm hiểu các công cụ và phương pháp dự đoán khả năng ức chế bệnh dựa trên các biểu diễn khác nhau cửa RNA. Các kết quả thu được mạc dù đã có nêu nhưng chưa thực sự được kiểm chứng một cách rõ ràng.

- **Nhận xét về nội dung, bố cục và hình thức của luận văn**

  Việc chia các chương như hiện tại cơ bản là phù hợp.

- **Các ý kiến nhận xét khác (về khả năng viết báo, phát triển sản phẩm, hoặc định hướng nghiên cứu tiếp theo,...)**

  - Không rõ mô hình bài toán dự đoán như thế nào. Tác giả lệ thuộc quá nhiều vào Weka.
  - Quá nhiều nội dung tổng quan.
  - Ứng dụng ở đây là gì? Cần làm rõ hơn.

- **Kết luận chung (khẳng định mức độ đáp ứng các yêu cầu đối với một luận văn Thạc sĩ; bản tóm tắt luận văn phản ánh trung thực nội dung cơ bản của luận văn; luận văn có thể đưa ra bảo vệ để nhận học vị Thạc sĩ được hay không?)**

Cơ bản đáp ứng yêu cầu. Đồng ý cho học viên được bảo vệ để nhận học vị Thạc sỹ.

Hà Nội, ngày 9 tháng 8 năm 2017

XÁC NHẬN CỦA CƠ QUAN CÔNG TÁC **CÁN BỘ PHẢN BIỆN**

Bùi Thu Lâm

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Số: 1162 /QĐ-ĐT

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

Hà Nội, ngày 23 tháng 11 năm 2017

QUYẾT ĐỊNH
Về việc thành lập Hội đồng chấm luận văn thạc sĩ

HIỆU TRƯỞNG
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Căn cứ Quy định về Tổ chức và hoạt động của các đơn vị thành viên và đơn vị trực thuộc Đại học Quốc gia Hà Nội ban hành theo quyết định số 3568/QĐ-ĐHQGHN ngày 08/10/2014 của Giám đốc Đại học Quốc gia Hà Nội;

Căn cứ Quy định về Tổ chức và hoạt động của Trường ĐH Công nghệ ban hành kèm theo Quyết định số 520/QĐ-ĐHCN ngày 19/7/2016 của Hiệu trưởng Trường ĐH Công nghệ;

Căn cứ Quy chế đào tạo sau đại học ở Đại học Quốc gia Hà Nội, ban hành kèm theo Quyết định số 1555/QĐ-ĐHQGHN ngày 25/5/2011 và Quyết định sửa đổi, bổ sung số 3050/QĐ-ĐHQGHN ngày 17/9/2012 của Giám đốc Đại học Quốc gia Hà Nội;

Căn cứ Quy chế Đào tạo thạc sĩ tại Đại học Quốc gia Hà Nội, ban hành theo Quyết định số 4668/QĐ-ĐHQGHN ngày 10/12/2014 của Giám đốc Đại học Quốc gia Hà Nội ;

Căn cứ Quyết định công nhận học viên cao học số 1010/QĐ-CTSV ngày 12/12/2014 của Hiệu trưởng Trường Đại học Công nghệ;

Căn cứ Công văn số 153/CNTT-ĐT ngày 31/10/2017 và Công văn số 169/CNTT-ĐT ngày 21/11/2017 của Chủ nhiệm Khoa Công nghệ thông tin về việc đề xuất hội đồng chấm luận văn;

Xét đề nghị của Trưởng phòng Đào tạo,

QUYẾT ĐỊNH:

Điều 1. Thành lập Hội đồng chấm luận văn thạc sĩ của học viên Phạm Thị Mai Hoa, sinh ngày 29/10/1989 tại Hà Nội là học viên cao học khóa 21,

Ngành: Hệ thống thông tin

Chuyên ngành: Hệ thống thông tin Mã số: 60480104

Tên đề tài luận văn: Các phương pháp dự đoán khả năng ức chế bệnh dựa trên các biểu diễn khác nhau của RNA và ứng dụng.

Cán bộ hướng dẫn: TS. Bùi Ngọc Thăng

Danh sách các thành viên Hội đồng kèm theo quyết định này.

Điều 2. Chủ nhiệm Khoa Công nghệ thông tin có nhiệm vụ tổ chức để học viên bảo vệ luận văn thạc sĩ trước Hội đồng theo đúng Quy chế Đào tạo thạc sĩ ở Đại học Quốc gia Hà Nội và các quy định hiện hành khác. Hội đồng tự giải thể sau khi hoàn thành nhiệm vụ.

Điều 3. Trưởng phòng Hành chính – Quản trị, Trưởng phòng Đào tạo, Chủ nhiệm Khoa Công nghệ thông tin, các Thủ trưởng đơn vị có liên quan, các thành viên Hội đồng và học viên Phạm Thị Mai Hoa chịu trách nhiệm thi hành quyết định này./.

KT. HIỆU TRƯỞNG
PHÓ HIỆU TRƯỞNG

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Chử Đức Trình

Nơi nhận:
- Như Điều 3;
- Lưu: VT, ĐT, CH11.

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

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

DANH SÁCH HỘI ĐỒNG CHẤM LUẬN VĂN THẠC SĨ

(theo Quyết định số: 1162/QĐ-ĐT ngày 23 tháng 11 năm 2017
của Hiệu trưởng Trường Đại học Công nghệ)

STT	Họ và tên	Cơ quan công tác	Trách nhiệm trong Hội đồng
1	PGS.TS. Nguyễn Trí Thành	Trường ĐH Công nghệ, ĐHQGHN	Chủ tịch
2	TS. Nguyễn Thị Hậu	Trường ĐH Công nghệ, ĐHQGHN	Thư ký
3	PGS.TS. Hà Quang Thuỵ	Trường ĐH Công nghệ, ĐHQGHN	Phản biện 1
4	PGS.TS. Bùi Thu Lâm	Học viện Kỹ thuật quân sự	Phản biện 2
5	PGS.TS. Nguyễn Kim Anh	Trường ĐH Bách khoa Hà Nội	Ủy viên

Hội đồng gồm có 05 thành viên./.