ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN BÁ QUÂN

CÁC PHƯƠNG PHÁP DỰ ĐOÁN VÀ ỨNG DỤNG VÀO BÀI TOÁN ĐOÁN NHẬN KHẢ NĂNG ỨC CHẾ GEN CỦA siRNA

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

HÀ NỘI – 2016

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN BÁ QUÂN

CÁC PHƯƠNG PHÁP DỰ ĐOÁN VÀ ỨNG DỤNG VÀO BÀI TOÁN ĐOÁN NHẬN KHẢ NĂNG ỨC CHẾ GEN CỦA siRNA

Ngành: Hệ thống thông tin

Chuyên ngành: Hệ thống thông tin

Mã số: 60 48 01 04

LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. BÙI NGỌC THĂNG

HÀ NỘI - 2016

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của cán bộ hướng dẫn khoa học, thầy giáo, TS. Bùi Ngọc Thăng, các kết quả đạt được trong luận văn này là quá trình tìm hiểu, nghiên cứu của riêng tôi. Trong toàn bộ nội dung của luận văn, những điều được trình bày là của cá nhân tôi hoặc là được tổng hợp từ nhiều nguồn tài liệu khác. Các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Hà Nội, ngày …… tháng ..… năm 2016

Học viên thực hiện luận văn

Nguyễn Bá Quân

LỜI CẢM ƠN

Đầu tiên, tôi muốn gửi lời cảm ơn sâu sắc nhất đến cán bộ hướng dẫn khoa học, thầy giáo, TS. Bùi Ngọc Thăng, người đã đưa tôi đến lĩnh vực nghiên cứu này và đã trực tiếp giảng dạy trong suốt quá trình tôi học tập, nghiên cứu tại trường Đại học Công Nghệ - Đại học Quốc Gia Hà Nội, thầy luôn truyền cho tôi nguồn cảm hứng, nhiệt huyết nghiên cứu khoa học và hết sức tận tình hướng dẫn tôi, cho tôi những lời khuyên quý báu. Mặc dù thầy rất bận với công việc giảng dạy và nghiên cứu nhưng thầy đã dành cho tôi nhiều thời gian thảo luận các ý tưởng nghiên cứu, chỉ dẫn cách nghiên cứu, giải đáp thắc mắc và động viên tôi vượt qua những vấn đề khó khăn cũng như hướng tôi tới nhiều vấn đề có giá trị khác khiến tôi muốn tìm hiểu và nghiên cứu trong tương lai.

Tôi xin bày tỏ lòng biết ơn chân thành tới Thầy, Cô giáo các anh chị và các bạn trong bộ môn Hệ thống thông tin, Khoa Công nghệ thông tin, những người đã nhiệt tình giúp tôi mở rộng kiến thức về Công nghệ thông tin nói chung và Hệ thống thông tin nói riêng, đó là những kiến thức quý báu và sẽ rất có ích với tôi trong giai đoạn hiện tại và tương lai.

Tôi xin gửi lời cảm ơn chân thành tới Ban Giám hiệu Nhà trường, Phòng Đào tạo sau đại học, Đại học Công nghệ - Đại học Quốc gia Hà Nội đã tạo điều kiện tốt nhất giúp tôi trong suốt quá trình học tập.

Qua tất cả tôi gửi đến gia đình thân yêu mọi tình cảm của mình, cảm ơn bố mẹ đã luôn luôn tin tưởng, luôn luôn là chỗ dựa vững chắc, cảm ơn các anh chị em đã dành mọi điều kiện để giúp tôi tập trung vào nghiên cứu.

Học viên thực hiện luận văn

Nguyễn Bá Quân

MỤC LỤC

LỜI CAM ĐOAN **1**
LỜI CẢM ƠN **2**
MỤC LỤC **3**
DANH SÁCH HÌNH VẼ **5**
DANH SÁCH BẢNG BIỂU **6**
DANH MỤC CHỮ VIẾT TẮT **7**
MỞ ĐẦU **8**
CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ ĐOẠN NGẮN RNA CÓ KHẢ NĂNG ỨC CHẾ (siRNA) **10**
1.1. Can thiệp RNA 10
1.1.1. Các cơ chế, thành phần chính của RNAi 10
1.1.2. Vai trò của RNAi 11
1.1.3. Thành phần của RNAi 12
1.1.4. Nghiên cứu can thiệp RNA 12
1.2. Nghiên cứu siRNA 15
1.2.1. Lịch sử nghiên cứu siRNA 15
1.2.2. Chức năng của siRNA 16
1.2.3. Ứng dụng siRNA 16
1.2.4. Những thách thức trong nghiên cứu siRNA 18
1.3. Kết luận 22
CHƯƠNG 2. CÁC QUY TẮC THIẾT KẾ siRNA HIỆU QUẢ **23**
2.1 Quy tắc thiết kế siRNA 23
2.2. Quy tắc thiết kế siRNA hiệu quả trong phương pháp sinh học 23
2.3. Các quy tắc thiết kế trong cách tiếp cận sinh học tính toán 27
2.4. Kết luận 29
CHƯƠNG 3. PHƯƠNG PHÁP DỰ ĐOÁN KHẢ NĂNG ỨC CHẾ CỦA siRNA **30**
3.1. Tổng quan một số phương pháp xây dựng mô hình dự đoán ức chế của siRNA 30
3.2. Phương pháp máy vecto hỗ trợ (SVM- Support vector machine) 32
3.3. Phương pháp rừng ngẫu nhiên (Random Forest) 39
3.4. Sử dụng phương pháp học biểu diễn để nâng cao độ chính xác của các mô hình dự đoán 46

3.5. Kết luận 49
CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ 50
4.1. Dữ liệu thực nghiệm và cài đặt 50
4.2. Thực nghiệm các phương pháp học máy dự đoán khả năng ức chế của siRNA 52
4.3. Đánh giá thực nghiệm 55
4.4. Kết luận 57
CHƯƠNG 5. KẾT LUẬN 58
5.1. Những vấn đề được giải quyết trong luận văn. 58
5.2. Công việc nghiên cứu trong tương lai 59
TÀI LIỆU THAM KHẢO 60

DANH SÁCH HÌNH VẼ

Hình 1.1: Sơ đồ hoạt động của RNAi và siRNA 11
Hình 1.2: Đồng ức chế của cây dạ yến thảo, cây bên trái là cây dại, bên phải là cây chứa biến đổi gen 12
Hình 1.3: Hai vấn đề quan trọng trong RNAi 19
Hình 1.4: Tìm siRNA hiệu quả cao 21
Hình 2.1: Quy tắc thiết kế siRNA hiệu quả 23
Hình 2.2: Ví dụ về phát hiện ra quy tắc thiết kế siRNA hiệu quả trong cách tiếp cận sinh học 24
Hình 2.3: Tìm quy tắc thiết kế dựa trên mạng nơron và cây quyết định 29
Hình 3.1: Quy trình xây dựng mô hình dự đoán khả năng ức chế của siRNA 30
Hình 3.2: Ví dụ sử dụng mô hình SVR dự đoán khả năng ức chế của siRNA 31
Hình 3.3: Siêu phẳng với lề cực đại trong không gian R^2 34
Hình 3.4: Ví dụ của GSK 36
Hình 3.5: Phân loại các dữ liệu thử nghiệm bởi thuật toán GSK / SVM 38
Hình 3.6: Mối quan hệ giữa tự luciferase siRNA và điểm GSK / SVM 38
Hình 3.7: Giải thuật rừng ngẫu nhiên cho phân lớp dữ liệu 41
Hình 3.8: Quy trình dự báo của RFR 44
Hình 4.1: Quy trình giải quyết bài toán 51
Hình 4.2: Quá trình thực nghiệm các phương pháp đề xuất 52
Hình 4.3: Các tham số huấn luyện mô hình Random forest 53
Hình 4.4: Các tham số huấn luyện mô hình SVR 54
Hình 4.5: Các tham số huấn luyện mô hình Linear Regression 54

DANH SÁCH BẢNG BIỂU

Bảng 1.1: Các quy tắc thiết kế siRNA được xây dựng trong thực nghiệm sinh học 21
Bảng 2.1: Các mô hình tìm quy tắc thiết kế siRNA bằng phương pháp sinh học tính toán 28
Bảng 3.1: Các phương pháp học máy sử dụng xây dựng mô hình dự báo 31
Bảng 3.2: So sánh hiệu suất phân biệt giữa 1-, 2-, 3- và (1, 2, 3) - GSK/SVM 36
Bảng 3.3: Danh sách 20 của vectơ trọng lượng SVM với (1,2,3)-GSK..... 37
Bảng 3.4: Các tính năng được sử dụng trong các mô hình dự báo RFR..... 43
Bảng 3.5: Thực hiện mô hình RFR và mô hình SVM trong siRNA..... 45
Bảng 3.6: Hiệu suất trên bảng dữ liệu độc lập..... 45
Bảng 3.7: Chuyển đổi chuỗi siRNA thành ma trận..... 46
Bảng 3.8: Ví dụ về quy tắc thiết kế..... 48
Bảng 4.1: Kết quả huấn luyện của mô hình Random forest 53
Bảng 4.2: Kết quả huấn luyện của mô hình SVR..... 54
Bảng 4.3: Kết quả huấn luyện của mô hình Linear Regression 55
Bảng 4.4: Các giá trị của R áp dụng trên bộ dữ liệu Huesken..... 55
Bảng 4.5: So sánh phương pháp thực nghiệm với 18 phương pháp 56

DANH MỤC CHỮ VIẾT TẮT

Ký hiệu	Từ tiếng Anh	Tiếng Việt
RNA	Axit ribonucleic	Axít ribônuclêic
siRNA	Short interfering RNA	RNA ngăn can thiệp
RISC	RNA – incluced silencing complex	Phức hệ gây sự im lặng
PTGS	Post transcriptional gene silencing	Im lặng gen sau phiên mã
dsRNA	Double-strand RNA	RNA xoăn kép
DNA	Axit deoxyribonucleic	Axít đêôxiribônuclêic
mRNA	Messenger RNA	RNA thông tin
CHS	Chalcone synthase	Gen quy định màu tím
SVM	Support vector machine	Máy vecto hỗ trợ
RF	Random forest	Rừng ngẫu nhiên
ANN	Artificial Neural Network	Mạng noron nhân tạo
ROC	Receiver operating characteristic	Đường cong đặc trưng hoạt động của bộ thu nhận

MỞ ĐẦU

Andrew Fire và Craig Mello đã tiến hành nghiên cứu về cơ chế điều khiển biểu hiện gen ở giun tròn (C. Elegans), hai ông đã thực hiện hàng loạt các thí nghiệm của việc tiêm RNA vào bộ phận sinh dục của giun tròn và phát hiện ra cơ chế gọi là can thiệp RNA. Năm 2006 Fire và Mello đã nhận được giải thưởng Nobel cho những đóng góp của mình trong nghiên cứu về sự can thiệp RNA (RNAi). Quá trình nghiên cứu của họ và của người khác về việc phát hiện RNAi đã có một tác động to lớn về nghiên cứu y sinh học và rất có thể sẽ được áp dụng trong y tế để tạo ra các loại thuốc mới để điều trị nhiều loại bệnh như virus cúm A, HIV, virus viêm gan B, ung thư. RNAi là quá trình sinh học trong đó đoạn RNA ngắn (siRNA) làm ức chế của gen mục tiêu (mRNA). Trong RNAi, các siRNA có thể được tổng hợp và tiêm vào tế bào để ức chế các mRNA, nhằm mục đích kiểm soát bệnh do đó tổng hợp các siRNA có hiệu quả cao để thiết kế các loại thuốc mới là một trong những vấn đề quan trọng nhất về nghiên cứu can thiệp RNA.

Nghiên cứu trên siRNA được liên tục thử nghiệm để tìm ra các phương pháp hiệu quả trong đó nghiên cứu đầu tiên tập trung vào các vấn đề của việc tìm kiếm quy tắc thiết kế siRNA. Mỗi quy tắc thiết kế siRNA được tìm ra bởi các đặc tính quan trọng của nó tác động đến hiệu quả ức chế, nhiều quy tắc thiết kế để tìm các siRNA có khả năng ức chế cao đã được phát hiện ra bởi các quá trình thực nghiệm sinh học và sinh học tính toán. Hướng nghiên cứu tiếp theo đó là tập trung vào các vấn đề xây dựng mô hình dự báo để dự đoán hiệu quả ức chế của các siRNA, các kỹ thuật học máy chủ yếu được sử dụng để giải quyết theo hướng nghiên cứu này. Tuy nhiên vẫn còn một số các hạn chế đó là hầu hết các quy tắc thiết kế siRNA có hiệu suất thấp và nhiều siRNA tạo ra không hoạt động hoặc không khả năng ức chế không cao hoặc hiệu suất của các mô hình dự báo được đề xuất cũng vẫn còn thấp và giảm khi thử nghiệm trên bộ dữ liệu độc lập. Vì vậy việc tìm kiếm các giải pháp cho hai vấn đề nêu trên để tạo ra các siRNA có khả năng ức chế hiệu quả cao vẫn là một thách thức lớn. Do những hạn chế trên nên quá trình nghiên cứu tiếp theo để tìm ra các phương pháp để tạo ra các siRNA hiệu quả cao đã hầu như không xuất hiện.

Với hướng đi tìm hiểu và nghiên cứu "*Các phương pháp dự đoán và ứng dụng vào bài toán đoán nhận khả năng ức chế của siRNA*". Luận văn tập trung vào việc tổng hợp các giải pháp nhằm giải quyết bài toán siRNA bao gồm các quy tắc thiết kế siRNA hiệu quả và phương pháp dự đoán khả năng ức chế của siRNA. Đồng thời cũng tiến hành đề xuất áp dụng thực nghiệm bằng một số phương pháp học máy và so sánh kết quả đạt được với kết quả thực nghiệm trên các phương pháp học máy đã được công bố. Kết quả đạt được giúp chúng ta có cách nhìn tổng quan và áp dụng một cách phù hợp vào giải quyết bài toán nhằm xây dựng một số mô hình dự đoán khả thi để đoán nhận khả năng ức chế của siRNA hỗ trợ cho việc điều chế thuốc. Bài toán đoán nhận khả năng ức chế gen của siRNA là một trong những thách thức hiện nay trong cộng

đồng nghiên cứu nhằm tìm ra cách điều chế thuốc để điều trị nhiều loại bệnh như bệnh viêm gan, bệnh ung thư, bệnh cúm...

Luận văn được chia làm năm chương chính:

Chương 1: *Giới thiệu tổng quan về đoạn ngắn RNA có khả năng ức chế (siRNA)*. Ở chương đầu tiên mở đầu sẽ trình bày một số kiến thức nền tảng của RNAi và trình bày tổng quát về siRNA bao gồm chức năng, hoạt động, ứng dụng, hạn chế và các phương pháp giải quyết bài toán siRNA.

Chương 2: ***Các quy tắc thiết kế siRNA hiệu quả:*** Trình bày khái quát các phương pháp đã được các nhà khoa học thực nghiệm để giải quyết vấn đề của bài toán. Đó là tìm các quy tắc thiết kế siRNA hiệu quả trong cả hai cách tiếp cận sinh học và sinh học tính toán.

Chương 3: ***Phương pháp dự đoán khả năng ức chế gen của siRNA*.** Chương này sẽ tập trung vào giới thiệu tổng quan về nghiên cứu xây dựng các mô hình dự báo và cách áp dụng các phương pháp học SVM và RF để dự đoán khả năng ức chế gen của siRNA. Đồng thời trình bày phương pháp học biểu diễn dữ liệu áp dụng cho phần thực nghiệm.

Chương 4: ***Thực nghiệm đánh giá***. Đây là phần nêu lên kết quả đạt được trong suốt quá trình thực hiện, ngoài ra còn đề cập đến những khó khăn vấn đề vướng mắc phát sinh, sau đó là đánh giá những kết quả đạt được chi tiết ở từng bước thực hiện

Chương 5: ***Kết luận***. Tổng kết lại những nội dung chính của luận văn, đưa ra hướng đi và hướng áp dụng thực tế.

CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ ĐOẠN NGẮN RNA CÓ KHẢ NĂNG ỨC CHẾ (siRNA)

Phần đầu của chương này trình bày tổng quan về sự can thiệp RNA, phần thứ hai là thảo luận chi tiết về siRNA gồm lịch sử ra đời, cơ chế hoạt động, chức năng, ứng dụng của siRNA cũng như giải pháp giải quyết bài toán siRNA.

1.1. Can thiệp RNA

Can thiệp RNA (RNAi) là một hệ thống bên trong các tế bào sống, giúp kiểm soát các gen đang hoạt động đó là các đoạn ngắn RNA giúp tế bào ức chế sự biểu hiện của các gen có trình tự tương đồng với nó. Đây là hệ thống tự vệ của tế bào nhằm chống lại sự xâm nhập của siêu vi khuẩn, các phần tử di truyền ngoại lai khác, những yếu tố sử dụng chuỗi RNA xoắn kép trong chu kỳ sống của tế bào.

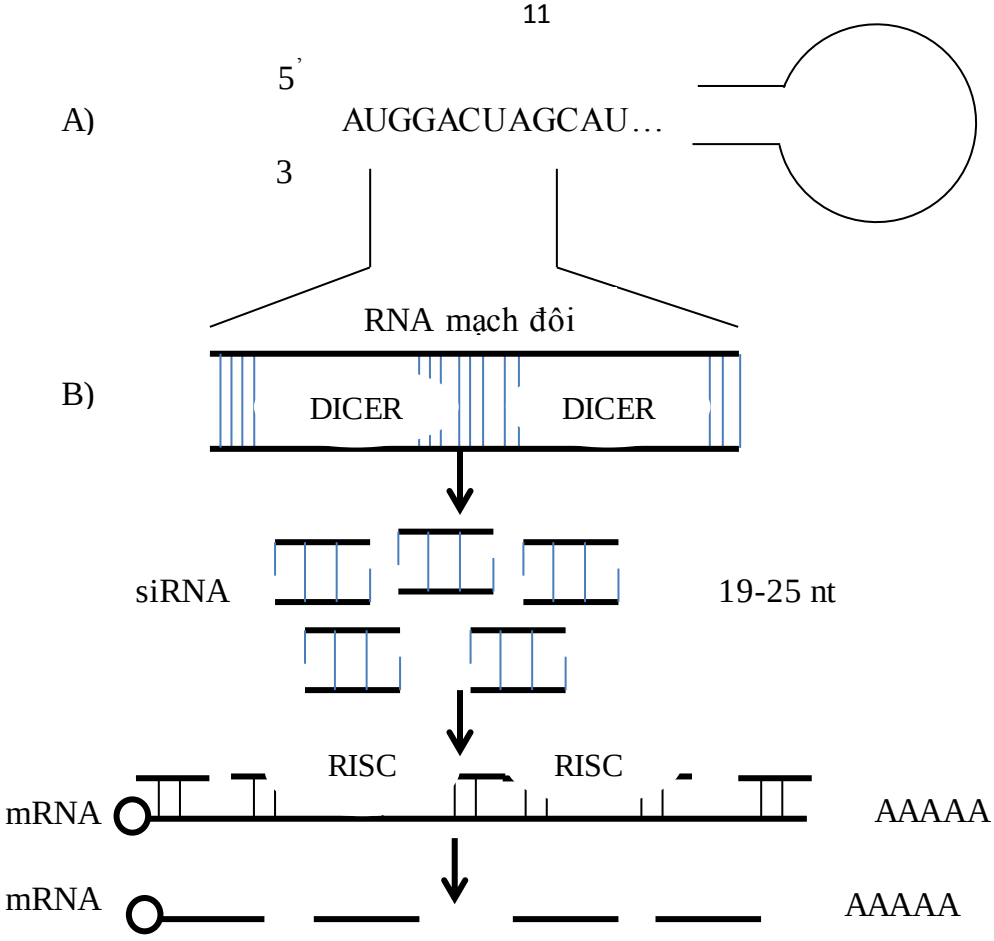
1.1.1. Các cơ chế, thành phần chính của RNAi

RNAi chính là quá trình phân hủy mRNA. Các dsRNA (Double stranded RNA) mạch kép hoặc dạng kẹp tóc bị cắt thành các đoạn ngắn RNA (siRNA) bởi các enzyme ribonuclease III Dicer, họ protein được gọi là RNA- phức hệ gây sự im lặng (RISC) sẽ mang các siRNA bám vào mRNA đích có trình tự tương đồng với nó và phân hủy mRNA. Nên quá trình chuyển hóa mRNA thành protein hay lây nhiễm virut RNA sẽ bị ngăn chặn. RNAi xảy ra trong quá trình im lặng gen sau phiên mã (PTGS), quá trình RNAi được bảo toàn mạnh mẽ ở các sinh vật nhân chuẩn đóng vai trò bảo vệ chống lại virus và sự bất ổn về di truyền phát sinh từ yếu tố di truyền động (Transposons).

Công trình nghiên cứu về RNAi của hai nhà khoa học Z.Fire và C.Mello đã được công bố trên tạp chí Nature vào ngày 19/2/1998, kết quả của nghiên cứu này vô cùng quan trọng bởi chúng cung cấp lời giải thích cho các hiện tượng nghiên cứu ở thực vật được các nhà nghiên cứu trước đó gọi là "Đồng ức chế". Khám phá của họ đã làm sáng tỏ nhiều quan sát thí nghiệm mâu thuẫn và khó hiểu trong nhiều năm trước đây, đồng thời tiết lộ một cơ chế tự nhiên để kiểm soát dòng thông tin di truyền trong tế bào, báo hiệu sự khởi đầu cho một lĩnh vực nghiên cứu mới.

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.

Quá trình RNAi bao gồm các bước sau (Hình1.1), bước đầu tiên của RNAi đó là RNA sợi kép (dsRNA) bị cắt thành những đoạn ngắn (siRNA) bởi một endonuclease gọi là dicer sẽ tách dsRNA thành các siRNA



Hình 1.1: Sơ đồ hoạt động của RNAi và siRNA

Bước thứ 2, siRNA được mở ra thành hai sợi đơn ngắn đó là hai sợi sence và antisence. Sợi antisense ngắn (siRNA) được nạp vào phức hợp RISC và sợi antisense RNA trong phức hợp RISC bắt cặp với mRNA bằng liên kết tương đồng giữa các bazơ. Tiếp theo RISC sẽ phân hủy mRNA tương đồng với nó. Các thành phần xúc tác mà tách hai sợi của siRNA được xác định là các protein thuộc họ Argonaut 2 (Ago2). Bằng cách phân tích cấu trúc tinh thể của Ago2, cho thấy Ago2 tương tự như RNase H, đó sẽ tách thành phần RNA của một DNA / RNA kép [34]

Có ba thành phần chính liên quan đến quá trình can thiệp RNA: siRNA, enzyme Dicer, và phức hệ (RISC). Trong đó siRNA là một đoạn ngắn của dsRNA (RNA mạch kép) có kích thước khoảng 19 đến 25 nucleotit với gốc phosphoryl là đầu 5 ' đến 2 phân tử nucleotide ở đầu hydroxy 3' (Hình 1.1A). Dicer là một endoonuclease giống như RNase III sẽ cắt RNA sợi đôi thành các đoạn ngắn RNA (siRNA) và RISC là một phức hợp đa protein (muti-protein) có chứa enzyme helicase và một số protein, trong đó quan trọng nhất là protein thuộc họ Agronaut hoạt động như một endonuclease và có vai trò cắt mRNA.

1.1.2. Vai trò của RNAi

RNAi có nhiều chức năng quan trọng trong tế bào như: Bảo vệ tế bào chống lại gen ký sinh trùng, virus và các yếu tố di truyền vận động (Transposon). Điều hòa biểu hiện gen. Duy trì hình dạng nhiễm sắc thể và tăng cường phiên mã…

1.1.3. Thành phần của RNAi

RNAi gồm 2 thành phần siRNA và miRNA

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

miRNA (micro RNA) là những đoạn RNA ngắn khoảng từ 19 đến 25 nucleotit, không tham gia vào quá trình tổng hợp protein.

1.1.4. Nghiên cứu can thiệp RNA

Can thiệp trong thực vật

Ở thực vật sự ức chế của RNA (RNA silencing) được phát hiện khi thực hiện biến đổi gen trên cây dạ yến thảo với dự kiến là có màu tím hơn. Năm 1990 phòng thí nghiệm R. Jorgensen muốn tăng cường hoạt động của gien tổng hợp chalcone synthase (chsA) một loại enzyme tham gia vào việc sản xuất sắc tố anthocyanin (Hình 1.2) họ đã thí nghiệm bằng cách chuyển gen quy định màu tím chalcone synthase dưới sự điều khiển của một promoter mạnh (promoter 35S), gen CHS là gen có liên quan đến chu trình hình thành chất anthocyanin trong hoa dạ yến thảo, tuy nhiên thay vì hình thành màu tím của cánh hoa như mong đợi thì chúng lại thể hiện các đốm màu khác nhau và thậm chí là màu trắng. Hiện tượng này các nhà khoa học đặt thuật ngữ là "cosuppresion" nghĩa là "đồng ức chế" bởi vì sự biểu hiện của gen ngoại sinh và gen nội sinh trong hoa dạ yến thảo đều bị ức chế như nhau. Thuật ngữ "đồng ức chế" là quá trình mô tả sự mất đi của các mRNA do gen nội sinh (gen có sẵn của tế bào) và gen ngoại sinh (gen được chuyển vào trong tế bào) phiên mã ra.

Hình 1.2: Đồng ức chế của cây dạ yến thảo, cây bên trái là cây dại, bên phải là cây chứa biến đổi gen

Trong khoảng thời gian này các phòng thí nghiệm khác [23] cũng cho thấy rằng việc chuyển gen ở dạng sense vào tế bào có thể làm giảm sự biểu hiện của gen nội sinh tương ứng. Sau đó nhiều trường hợp đồng ức chế tương tự đã được báo cáo trong các tài liệu khoa học. Tất cả các trường hợp đồng ức chế đều dẫn đến sự thoái hoá của các phân tử RNA của gen nội sinh và gien ngoại sinh sau khi quá trình phiên mã ở nhân

xảy ra [27].Vì sự thoái hoá của RNA sau phiên mã được quan sát thấy ở một loạt các gien ở thực vật, vi khuẩn hoặc virut, nó được đặt tên lại là sự bất hoạt gien sau phiên mã [PTGS], PTGS không chỉ dưới tác động của gen được chuyển dạng sense mà còn cả antisense và những bằng chứng về mặt hoá học cho thấy các cơ chế tương tự nhau có thể đã xảy ra trong cả hai trường hợp [17]. Nó chỉ ra rằng mặc dù hiện tượng đồng ức chế ban đầu được quan sát ở thực vật nhưng nó không chỉ giới hạn ở thực vật mà còn xuất hiện ở động vật đa bào và động vật có vú.

Cùng thời gian đó những biến đổi quan sát được ở các kiểu hình (motif) có liên quan tới PTGS được cho là do sự kết hợp ở nhiều vị trí, hình thành nên phân tử RNA dị thường, các cấu trúc được lặp đi lặp lại của gen được chuyển vào trong tế bào. Sau đó, người ta mới biết rõ rằng sự biểu hiện của gen được chuyển vào trong tế bào đã dẫn đến sự hình thành nên dsRNA và bắt đầu cho PTGS.

Các báo cáo từ một số phòng thí nghiệm trong vài năm qua đã cho thấy rằng sự mất đi khả năng tích tụ các mRNA đích là gần như hoàn toàn nếu gen được chuyển vào phiên mã ra các bản sao ở dạng mạch kép. Bằng chứng này cho thấy việc tạo ra dsRNA là cần thiết để khởi đầu PTGS ở thực vật, dựa vào điều này, thực vật mang gen chuyển có hoạt động phiên mã mạnh mẽ theo cả hai hướng tạo ra dạng sense và dạng antisense đã cho thấy những đặc tính PTGS mạnh mẽ. Những thực vật chuyển gen này có thể gây bất hoạt gien nội sinh, RNA virut có thể xâm nhập hoặc những gen lạ không mong muốn xuất hiện và đặc tính này có thể di truyền được.

Nói chung, thành phần sense và antisense của các gen được chuyển vào tế bào nói trên chỉ khác nhau ở một DNA bên trong một gen nhưng không tham gia vào việc mã hoá protein (intron) để tăng cường tính hiệu quả của PTGS [8], [42]. Ví dụ, 2 loại cà chua Arabidopsis thaliana và Lycopersicon esculentum được biến nạp bởi một gien được thiết kế với mục đích tạo ra các bản sao iaaM và ipt có khả năng tự bổ sung. (iaaM và ipt là các gien gây ung thư của vi khuẩn Agrobacterium), chịu trách nhiệm tạo thành những khối u ở thực vật bị nhiễm. Những dòng thực vật chuyển gen này vẫn giữ được tính mẫn cảm với sự biến nạp của Agrobacterium nhưng có thêm tính chống chịu cao đối với sự tạo thành khối u, mang lại sức đề kháng đối với bệnh bằng cách làm thoái các phân tử RNA phiên mã từ hai gen iaaM và ipt [15].

Can thiệp trong các tế bào động vật có vú.

Kỹ thuật gây ức chế gen trước hết có thể được áp dụng cho thực vật, giun tròn (C. Elegans) hoặc duồi giấm (D. melanogaster), nhưng chưa được áp dụng cho động vật có vú bởi vì dsRNA kích hoạt một loại kháng virut (INF) được lý giải như một tác nhân gây bệnh và protein kinase R kích hoạt chấm dứt sự tổng hợp protein trong tế bào bị ảnh hưởng [9]. Tuschl và đồng nghiệp là những người mở đường cho việc thí nghiệm RNAi trong các tế bào động vật có vú tạo ra các cơ hội mới cho phương pháp

điều trị nghiên cứu và điều trị. Các siRNA trước tiên tổng hợp phosphoryl ở 5' bởi kinase CLP1 sau khi đưa vào các tế bào [51] được mô tả RNAi (Hình 1B).

RNAi mở rộng nghiên cứu trên các phân tử DNA, RNA mạch đơn ngắn (Oligonucleotides) đã được sử dụng hơn 30 năm qua để gây ức chế sự biểu hiện của gen ở mRNA.

Antisense và RNAi có nhiều điểm chung chẳng hạn như sự cần thiết để xác định chuỗi liên kết phù hợp trên RNA đích, sự ổn định của các phân tử DNA, RNA mạch đơn ngắn (oligonucleotide) bởi biến đổi hóa học, hoặc vận chuyển của polymer điện tích âm qua màng tế bào. Với những kết quả đó antisense rất nhanh chóng được thực hiện với các chiến lược can thiệp RNA mới [10]. Tuy nhiên có sự khác biệt quan trọng giữa hai công nghệ: Antisense oligonucleotide là các phân tử sợi đơn ngắn (biến đổi) DNA chủ yếu là tách RNA đích trong nhân tế bào bằng cách kích hoạt các RNase H. Ngược lại can thiệp RNA được kích hoạt bởi RNA sợi đôi (dsRNA) có chức năng chủ yếu trong tế bào chất, trong đó Ago2 là thành phần quan trọng nhất của RISC [39] làm cho RNAi sẽ xuất hiện trong cấu trúc rời rạc của tế bào chất. Nó có thể tăng lên hiệu quả 1000 lần như antisense oligonucleotide truyền thống đối với các phân tử cùng một mục tiêu và vùng hạt giống [19] (vị trí 2-8 của sợi antisense, Hình 1 A) là rất quan trọng cho siRNA.

Sự suy thoái của RNA đích thường bắt đầu ngay lập tức sau khi siRNA vào tế bào. Tuy nhiên việc giảm số lượng protein phụ thuộc vào chu kỳ nửa phân rã của protein đích, thông thường hiệu quả ức chế có thể quan sát thấy trong vòng 48 giờ khi chuyển vào một siRNA trong tế bào, tuy nhiên, có những protein có sự luân chuyển với tốc độ rất chậm, có thể được quan sát thấy lâu hơn. Trong hầu hết các trường hợp các gen đích không hoàn toàn ngừng, đó là lý do can thiệp RNA được gọi là một công nghệ ức chế (ức chế trong trường hợp động vật biến đổi gen được tạo ra bởi sự tái tổ hợp tương đồng).

Ức chế sự biểu hiện của các gen mục tiêu thường kéo dài 5-7 ngày, hai thử nghiệm trong ống nghiệm [52] và ngoài ống nghiệm [11] thấy rằng một siRNA có thể làm việc với các thời gian khác nhau ở các loài khác nhau. Một siRNA chống những thành phần protein có chức năng vận chuyển lipid trong hệ thống tuần hoàn (apolipoprotein B) cho thấy có hoạt động ở chuột chỉ một vài ngày và sau chín ngày đã trở lại đến 70% của mức khởi điểm ban đầu trong khi sử dụng ức chế (knockdown) với các loài linh trưởng không phải con người là 11 ngày [49]. Thời gian tác dụng của một siRNA có thể phụ thuộc vào nhiều yếu tố, chẳng hạn như các cơ quan đích, gen đích và các loài. Trong tế bào shRNA có thể được sử dụng thay cho siRNA tổng hợp nhằm mở rộng gen im lặng. RNAi chính là một quy trình PTGS biểu hiện gen bị ức chế bởi một mRNA và RNAi có thể làm thay đổi cấu trúc nhiễm sắc thể trong nhân và do đó ảnh hưởng đến phiên mã. Điều này đã được quan sát đặc biệt đối với ruồi giấm, thực

vật. Tuy nhiên, tầm quan trọng của RNAi đối với ức chế gen ở động vật có vú đã không được chứng minh rõ ràng.

1.2. Nghiên cứu siRNA

Các đoạn ngắn RNA có khả năng ức chế (siRNA) là các phân tử RNA sợi kép nhỏ, kích thước khoảng 19 đến 25 nucleotit, được tạo bởi Dicer, một RNA endonuclease nhóm III, là thành phần trong phức hợp RISC có chức năng phân hủy mRNA đồng dạng của nó.

1.2.1. Lịch sử nghiên cứu siRNA

Nguồn gốc hình thành siRNA chính là từ kỹ thuật antisense-RNA, khi phân tử antisense RNA được hình thành thì việc tổng hợp protein beta-galactosidase bị ức chế gần như hoàn toàn (98%). Tuy nhiên, đến năm 1990 các nhà khoa học mới phát hiện ra cơ chế gây ra sự ức chế trên là do gen. Đó là nghiên cứu trên loài hoa dạ yến thảo (petunia), các nhà khoa học đã cố gắng tạo màu tím trên cánh hoa petunia bằng cách chuyển gen quy định màu tím Chalcone synthase (CHS) dưới sự điều khiển của promoter 35S. Gen CHS là gen có liên quan đến chu trình hình thành chất anthocyanin trong hoa petunia. Kết quả cánh hoa lại thể hiện các đốm màu khác nhau và màu trắng chứ không phải là màu tím. Năm 1994, Cogoni và các cộng sự đã tiến hành một thí nghiệm nhằm phát triển màu cam của nấm Neurospora crassa thông qua việc chuyển một gen có chức năng tạo ra carotenoid (một dạng sắc tố hữu cơ). Tuy nhiên nấm lại không có màu cam. Năm 1995, Guo và Kemphues đã đưa ra bằng chứng đầu tiên trên tuyến trùng Caenorhabditis elegans, đó là hiện tượng RNA sợi sense và antisense có hiệu quả ức chế biểu hiện gen như nhau.

Hiện tượng RNAi được khám phá đầu tiên trên giun tròn Caenorhabditis elegans do việc ức chế biểu hiện gen bởi RNA sợi đôi. Timmons L và Fire A đã dùng antisense RNA để ức chế biểu hiện gen. Hiệu quả tác động của antisense RNA hơn nhất 10 lần so với chỉ là dùng sợi sense.

Cho đến nay đa số các siRNA được công bố có nguồn gốc ngoại sinh. Tức là có nguồn gốc từ bên ngoài đưa vào tế bào và cơ thể sống bằng các con đường khác nhau (bằng tiêm hoặc có nguồn gốc từ các gen RNAi chuyển từ bên ngoài vào cơ thể). siRNA nội sinh lần đầu tiên được Baulcome và Hamilton vào năm 1999. Các tác giả đã chuyển gen *aco, gus* vào cây cà chua và thuốc lá. Trên các cây phát hiện hiện tượng PTGS, các tác giả đã phát hiện được các phân tử RNA nhỏ, đặc hiệu nhưng ngược chiều với gen chuyển (chứng tỏ không phải sản phẩm phân hủy mRNA của các gen trên). Sau đó nghiên cứu của Tuschl đã công bố phát hiện siRNA gây bất hoạt gen ở động vật.

Trong các tế bào người sự kích hoạt gen được tìm thấy đầu tiên do các siRNA kích hoạt các promoter của E-cadherin và p21, làm tăng mức độ biểu hiện của mRNA

và protein. Trong cơ thể sống (in vivo), siRNA đóng vai trò quan trọng trong việc hạn chế lây nhiễm virus vì nó làm bất hoạt RNA được tạo ra trong chu kỳ sống của virus.

Quá trình hình thành siRNA diễn ra ở tế bào chất (cytoplasma). RNAi được kích hoạt bởi dsRNA và được cắt thành những mảnh có độ dài khoảng 21 đến 25 bởi một enzyme dicer ở ngoài tế bào chất. Những đoạn dsRNA bị cắt được gọi tắt là siRNA. Hiệu quả ức chế của gen phụ thuộc vào mức độ tương đồng giữa siRNA và mRNA đích. Nếu sự tương đồng là hoàn toàn thì phân tử mRNA có xu hướng bị cắt và phân giải, do vậy không có mRNA sao mã cho protein đó.

Khả năng gây ức chế của siRNA có hiệu quả rất cao, chỉ cần một lượng nhỏ siRNA được đưa vào tế bào có thể đủ làm tắt hoàn toàn sự biểu hiện của một gen nào đó (vốn có rất nhiều bản sao trong cơ thể đa bào).

1.2.2. Chức năng của siRNA

Chức năng của siRNA đó là

- Bảo vệ tế bào chống lại gen ký sinh trùng, virut và các yếu tố di truyền vận động
- Giữ gìn nhiễm sắc thể và tăng cường phiên mã

Ngoài ra còn rất nhiều chức năng khác mà con người chưa khám phá ra và sẽ được khám phá dần trong tương lai

1.2.3. Ứng dụng siRNA

Nghiên cứu các chức năng của gen

Nghiên cứu các trình tự hệ gen người cũng như các sinh vật nhân chuẩn là một trong những phát triển quan trọng nhất của trong vài thập kỷ gần đây trong khoa học đời sống. Trong nhiều trường hợp chỉ có các trình tự hệ gen được biết đến nhưng các chức năng của protein được mã hóa vẫn chưa biết. Xác định chức năng của gen đã trở thành một trong những nhiệm vụ nghiên cứu quan trọng nhất hiện nay. Trong một vài năm gần đây việc áp dụng RNAi là một phương pháp chuẩn của nghiên cứu sinh học phân tử được các phòng thí nghiệm hóa sinh sử dụng với số lượng rất lớn. Kể từ khi ức chế gen được thực hiện với sự ghép đôi giữa mRNA và siRNA, chức năng của gen có thể được kiểm tra nhanh hơn nhiều. Ngoài ra các nhóm protein do một gen tạo ra (isoforms) có thể được chọn lọc tắt bằng cách lựa chọn phù hợp các trình tự đích để điều tra các chức năng cụ thể trong khi các chất dược lý không thể làm được, can thiệp RNA cung cấp một phương pháp nhanh chóng để tìm ra các mục tiêu.

Ứng dụng điều trị

Sự phát triển lâm sàng của các chuỗi ngắn của DNA (oligonucleotide antisense) [12] và ribozymes [43] đã được sử dụng trong các ứng dụng điều trị của các siRNA. Vì vậy các phương pháp điều trị can thiệp RNA đầu tiên được thử nghiệm bắt đầu trên

con người chỉ ba năm rưỡi sau khi siRNA lần đầu tiên được sử dụng trong các tế bào động vật có vú. Trong khi đó oligonucleotide antisense và siRNA khác nhau bởi kích thước của chúng, với số lượng lớn các oligomer gây ra những khó khăn và chi phí cao. Hơn nữa hai sợi của siRNA phải được tổng hợp riêng biệt và sau đó lai ghép tiếp, quá trình này phải đảm bảo sự hình thành của một loại thuốc thống nhất.

Bệnh về mắt

Chỉ có hai oligonucleotit chỉ đã được phê duyệt với cục quản lý thực phẩm và dược phẩm Hoa Kỳ là để điều trị các bệnh về mắt. Các nghiên cứu lâm sàng can thiệp RNA lần đầu tiên được bắt đầu vào cuối năm 2004 với một siRNA chống lại yếu tố tăng trưởng nội mạc (VEGF). Các siRNA được thử nghiệm dưới tên Bevasiranib trong một thử nghiệm giai đoạn III của công ty Opko Health. Phương pháp điều trị siRNA bắt đầu các nghiên cứu lâm sàng đầu tiên với biến đổi hóa học của một siRNA. Các siRNA được cố định bởi deoxythymidine lẻ với một liên kết phosphorothioate và hoán đổi một dư lượng đường cơ bản trên đầu sợi antisense và sense. Trong một nghiên cứu y học mới, các siRNA RTP801i-14 chống lại các rtp801 gen thiếu oxy gây ra đã được sử dụng để điều trị bệnh thoái hóa điểm vàng do tuổi theo dược phẩm Quark. Cách này có thể an toàn hơn và hiệu quả hơn so với các chất NTI-VEGF.

Nhiễm Virut

Nhiễm virus là một vấn đề lớn của y học hiện nay. Số lượng nhiễm virut liên quan đến HIV-1, cũng như viêm gan B (HBV) và viêm gan C (HCV), đang gia tăng liên tục, hơn nữa có những biến thể mới của virus như cúm virus H5N1 hoặc virus mới như SARS mà nổi lên như là mối đe dọa. Thực tế là do con người và động vật sống gần gũi với nhau trong một số khu vực trên thế giới có nghĩa là có nhiều mối nguy hiểm mới từ virus phải dự kiến được. Mặc dù, có rất nhiều các thuốc kháng virus phát hiện, chỉ có một số ít loại thuốc đã được phê duyệt để điều trị các bệnh do virus. Điều này chứng tỏ sự cần thiết cho sự phát triển của chiến lược chống virus mới.

RNAi được dựa trên các cặp bazơ bổ sung của một RNA đích và hướng các sợi siRNA cho phép thích ứng nhanh chóng với bất kỳ biến thể nhất định của một virus hoặc các loại virus mới. Đây là một trong những lợi thế lớn của RNAi so với các phương pháp khác. Kể từ khi các báo cáo đầu tiên về tác dụng kháng virus của siRNA chống virus hợp bào hô hấp (RSV), ứng dụng kỹ thuật RNAi thành công với hầu hết các virus có liên quan y tế, bao gồm cả HIV-1, HBV, HCV, SARS, virus cúm, virus bại liệt, đã được công bố [28].

Một vai trò quan trọng trong phương pháp tiếp cận can thiệp RNA chống lại virus đó là sự lựa chọn các trình tự mục tiêu phù hợp. RNA virus thường chứa các cấu trúc không quan trọng, có thể cản trở hiệu quả của sự ức chế siRNA.

Một trong những vấn đề lớn nhất đối với việc sử dụng RNAi lâu dài để chống lại virus là virus trốn thoát (escape). Đối với cả hai virus bại liệt [18] và HIV [5], đã

được mô tả trong đó bản sao virus có thể lúc đầu bị chặn hiệu quả, nhưng sau một thời gian tăng trở lại vì có các đột biến mà có thể vượt qua sự ức chế.

Ung thư

Sự khám phá ra cơ chế RNA can thiệp chính là công cụ cần thiết để dò tìm các cơ chế phân tử bị thay đổi trong tế bào ung thư. Sự biểu hiện của gen dẫn đến sự hình thành mạch trong khối u để tạo ra các mạch máu mới để cung cấp các khối u cũng có thể bị chặn. Mục tiêu nghiên cứu là di căn, vì trong nhiều trường hợp khối u chính có thể được phẫu thuật. Quan trọng nhất trong đó các tế bào khối u trở nên đề kháng với hóa trị liệu thông qua sự biểu hiện của các gen kháng đa thuốc (MDR). Do tính đặc hiệu của quá trình can thiệp RNA nên có thể dễ dàng tiến hành thực nghiệm trên hàng ngàn gen hoặc toàn bộ hệ gen trong mỗi thí nghiệm. Từ đó, khía cạnh ung thư sẽ được giải mã và sẽ tìm ra thuốc điều trị ung thư đặc hiệu.

Có nhiều nghiên cứu được công bố trong đó cho thấy rằng sự tăng trưởng của khối u sẽ bị chậm lại ở động vật bằng kỹ thuật RNAi. Ví dụ siRNA chống CD31 ức chế sự tăng trưởng của các khối u ở mô hình chuột mô ghép (xenograft) khác nhau [38]. Các siRNAs thâm nhập vào các tế bào khối u nội mô như lipoplexes và khối mạch

Các thử nghiệm lâm sàng khác

Trong một nghiên cứu lâm sàng khác, RNA đang được sử dụng như là một chiến lược điều trị chống suy thận cấp. Nó đã được chứng minh rằng sự ức chế tạm thời của p53 ức chế khối u có thể ngăn ngừa tổn thương tế bào [30] và các siRNA AKli-5 sẽ ức chế sự biểu hiện của p53 trong một thời gian hạn chế. Sự an toàn của AKli-5 là để được kiểm tra thử nghiệm giai đoạn I ở bệnh nhân mà có nguy cơ cao bị suy thận tồn tại vì hoạt động tim mạch.

Vào tháng Giêng năm 2008, Transderm Inc đã bắt đầu một nghiên cứu lâm sàng để điều trị các nhiễm sắc thể di truyền bệnh dày móng bẩm sinh (Pachyonychia congenital). Các siRNA được tiêm vào và đặc biệt là ức chế sự biểu hiện của các keratin đột biến K6a [40].

1.2.4. Những thách thức trong nghiên cứu siRNA

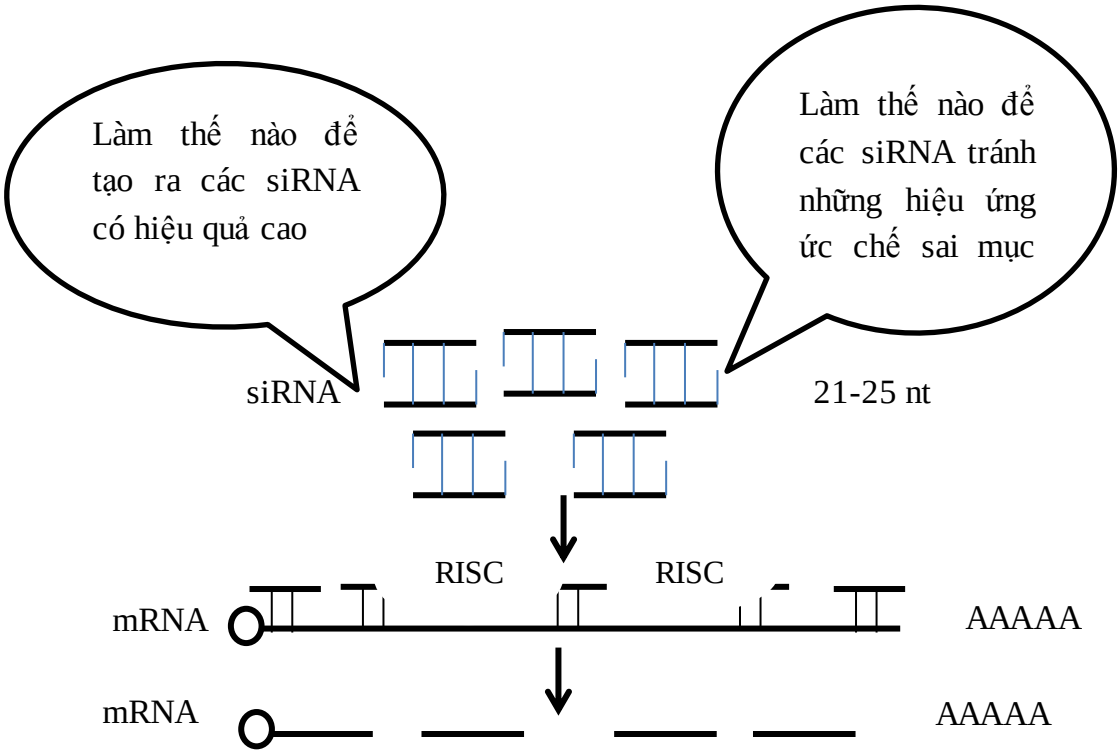
RNA sử dụng các RNA ngắn can thiệp (siRNA) có một cấu trúc được xác định rõ ràng, bao gồm một RNA mạch kép ngắn có khoảng 21-25 nucleotit với đầu 5’- P và 3’-OH có hai nucleotit nhô ra (Hình 1.1A). Chúng có thể được đưa trực tiếp bằng cách chuyển vào hoặc tạo ra trong tế bào và bị cắt thành các siRNA nhờ Dicer. Sau đó các siRNA được tháo xoắn dưới tác dụng của enzyme helicas thành hai sợi esense và antisense. Sợi antisense được nạp vào phức hợp RISC và antisense RNA bắt cặp với mRNA. Tiếp theo RISC phân hủy mRNA tương đồng với nó. Các siRNA có thể được tổng hợp để ức chế các gen đích. Trong thực tế các thí nghiệm có kết quả tốt nói lên

rằng các siRNA có cùng một mục tiêu nên được sử dụng độc lập để đảm bảo rằng các tác dụng sinh học là do sự ức chế của gen đích. Bằng các phân tích thực nghiệm, các nhà sinh học đã báo cáo rằng hiệu quả của các siRNA khác nhau có cùng RNA đích có thay đổi đáng kể [22] các siRNA cũng có thể hướng mục tiêu là không liên quan tới mRNA [25] được gọi là hiệu ứng ức chế sai mục tiêu của các siRNA. Trong nghiên cứu RNAi, tổng hợp các siRNA có hiệu quả cao và hiệu ứng ức chế sai mục tiêu, là những vấn đề rất quan trọng để thiết kế các loại thuốc mới. Do đó hai vấn đề quan trọng sau đây (Hình 1.3) có thể được coi là đáng kể:

(i) Làm thế nào các siRNA tránh hiệu ứng ức chế sai mục tiêu

(ii) Làm thế nào để tạo ra các siRNA có hiệu quả cao.

Hai vấn đề này được thảo luận chi tiết như sau



Hình 1.3: Hai vấn đề quan trọng trong RNAi

Tránh tác động hiệu ứng ức chế sai mục tiêu của siRNA

Hiệu ứng ức chế sai mục tiêu (Off–target effects) xảy ra khi một siRNA được xử lý bởi RISC và các mRNA. Các đặc trưng chi tiết của hiệu ứng ức chế sai mục tiêu được đưa ra đầu tiên vào năm 2003 [25]. Phương pháp phát hiện là sử dụng mẫu dò trên giá thể rắn (microarray profiling), các tác giả xác định thay đổi rất ít sau khi tăng 1,5 - 3 lần, các biểu hiện thay đổi của hàng chục đến hàng trăm gen sau đưa acid nucleic vào bên trong tế bào động vật của từng siRNA riêng biệt. Mức bổ sung giữa các sợi sence hoặc antisense của siRNA và các gen ứng hiệu ứng ức chế sai mục tiêu thay đổi đáng kể và nó là duy nhất cho mỗi siRNA.

Ban đầu những thay đổi khiêm tốn trong biểu hiện gen của hiệu ứng ức chế sai mục tiêu khiến nhiều người nghĩ rằng các kết quả thu được là không quan trọng và bỏ qua chúng. Nhưng gần đây đã được xua tan bởi các báo cáo rằng hiệu ứng ức chế sai mục tiêu có thể quan sát và đo lường được. Các nghiên cứu gần đây đã chỉ ra rằng nó không phải là tính chất nói chung của một mRNA với siRNA mà là sự tương ứng hoàn hảo giữa các phần của 3'-UTR (UTR vùng không được dịch mã) và các khu vực chính (vị trí 2-7 hoặc 2-8) của sợi antisense của siRNA để quyết định sự biểu hiện gen bị ảnh hưởng (Hình 1.1). Du *et al*. [14] chỉ ra rằng không những vị trí của cặp bazơ sai mà tính chất của các nucleotit hình thành không phù hợp đã ảnh hưởng đến tác dụng của hiệu ứng ức chế sai mục tiêu của siRNA.

Việc kết hợp hoàn hảo của cặp bazơ ở khu vực chính đã được tìm thấy là rất quan trọng trong hoạt động ức chế và siRNA rất nhạy cảm với sai lệch trong khu vực này. Đột biến nucleotit nhất định tại các vị trí 5, 7, 8 và 11 đã được tìm thấy và đột biến này được công nhận là khá tốt và biểu hiện của các gen tổng hợp bị ức chế, các cách tiếp cận mới đối với của hiệu ứng ức chế sai mục tiêu có liên quan đến thiết kế siRNA. Nghiên cứu của Birmingham [6], Lim [35] và Jackson [26] tiết lộ rằng gen của hiệu ứng ức chế sai mục tiêu thường xuyên chứa kết quả phù hợp giữa các khu chính của siRNA (vị trí 2 -7) và chuỗi trong 3 'UTR của gen của hiệu ứng ức chế sai mục tiêu. Điều này có nghĩa rằng của hiệu ứng ức chế sai mục tiêu có thể được giảm bằng cách sử dụng siRNA.Vì vậy cách tiếp cận nghiên cứu đầy hứa hẹn của RNA là nghiên cứu cả hai hướng (hiệu quả cao siRNA) và (hiệu ứng ức chế sai mục tiêu) về ức chế sự biểu hiện của một gen (knockdown gen). Hơn nữa các đặc trưng của các siRNA có thể được giảm xuống bằng sự kết hợp của các nucleotit sửa đổi, nó dễ dàng để hoàn toàn làm bất hoạt các sợi sense bằng sự giảm bớt các nguy cơ hiệu ứng ức chế sai mục tiêu đến mức tối thiểu. Mặt khác thay đổi sợi antisense tạo ra khó khăn cho hiệu quả ức chế của siRNA để nhằm mục tiêu làm cho các các mRNA không bị ảnh hưởng.

Liên quan đến vấn đề này theo phương pháp sinh học tính toán một vài nhóm nghiên cứu đã đề xuất các hàm đánh giá hoặc các mô hình để dự đoán ra hiệu ứng ức chế sai mục tiêu của các siRNA. Các hàm đánh giá đầu tiên được đề xuất bởi Alistair và các đồng nghiệp của mình [1] và họ đã phát triển một phương pháp đánh giá mới dựa trên kết quả của Du *et al*. [14]. Trong đó sử dụng thực nghiệm quan sát hiệu ứng ức chế sai mục tiêu ở mỗi vị trí siRNA. Trong năm 2010, Karol và đồng nghiệp [29] đề xuất phương pháp dựa trên hàm nhân (kernel function) để phân tích hiệu ứng ức chế sai mục tiêu. Họ đã phát triển một phương pháp dùng để đo trình tự giống nhau dựa trên sự xuất hiện chung của chiều dài các chuỗi con, tính với sai số. Mặc dù việc xây dựng một hàm chức năng cho các hiệu ứng ức chế sai mục tiêu của các siRNA được coi là một vấn đề quan trọng, tuy nhiên các tính năng thu được dựa trên vị trí và khu vực không phù hợp giữa siRNA và mRNA có thể không đủ thông tin để xây dựng một dự báo tốt. Do đó, nó trở thành một vấn đề thách thức.

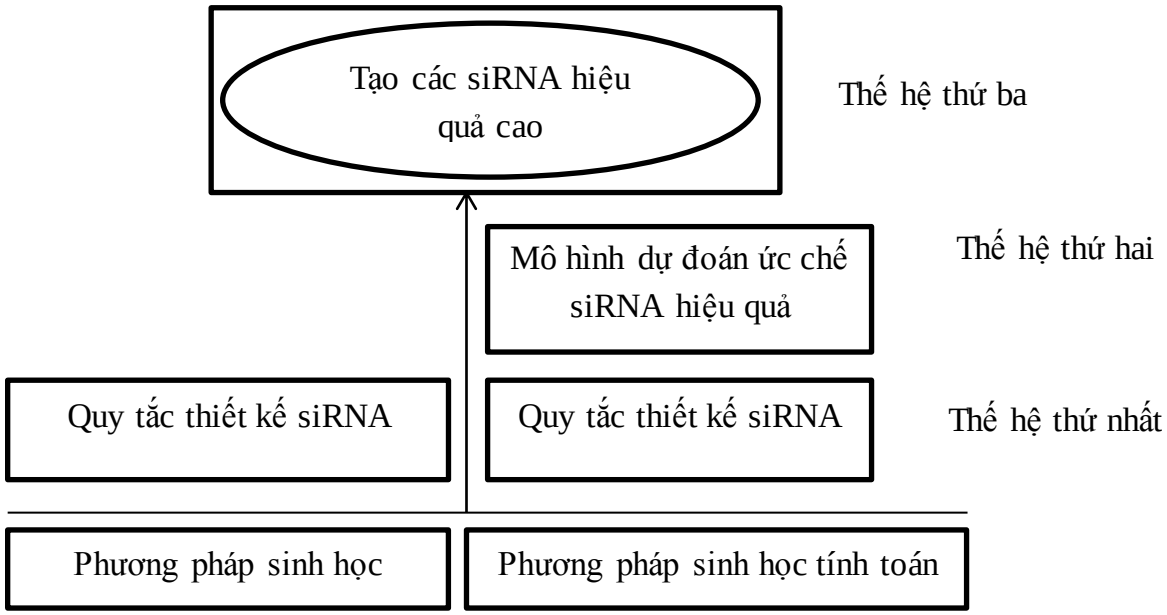
Tạo các siRNA hiệu quả cao

Như đã đề cập ở trên, các siRNA có thể được tổng hợp và đưa vào tế bào để làm ức chế gen đích, nó dẫn đến việc tạo nhiều loại thuốc mới dựa trên các siRNA để điều trị nhiều loại bệnh. Tuy nhiên các siRNA có thể làm ức chế các mRNA tương đồng ở các cấp độ khác nhau, do đó tạo ra nhiều siRNA hiệu quả cao là một vấn đề rất quan trọng, đã có rất nhiều các nghiên cứu để tìm ra siRNA có hiệu quả cao trong cả hai cách tiếp cận là sinh học và sinh học tính toán. Các vấn đề để giải quyết bài toán siRNA để tạo ra các siRNA đạt hiệu quả cao như hình 1.4.

Vấn đề 1: Tìm quy tắc thiết kế siRNA hiệu quả (thế hệ đầu tiên).

Vấn đề 2: Xây dựng mô hình dự báo để dự đoán hiệu quả ức chế siRNA (Thế hệ thứ hai).

Vấn đề 3: Tạo siRNAs hiệu quả cao (thế hệ thứ ba).



Hình 1.4: Tìm siRNA hiệu quả cao

Trong quá trình nghiên cứu để giải quyết các vấn đề của bài toán siRNA việc tìm quy tắc thiết kế siRNA hiệu quả, các nhà khoa học sử dụng cả hai cách tiếp cận sinh học và sinh học tính toán để tìm ra đặc điểm quan trọng của siRNA có ảnh hưởng đến hiệu quả của ức chế. Kết quả là đã có một số các quy luật thiết kế siRNA quan trọng được báo cáo (Bảng 1.1).

Bảng 1.1: Các quy tắc thiết kế siRNA được xây dựng trong thực nghiệm sinh học

Năm	Nhóm nghiên cứu	Số gen	Số siRNA	Công nghệ
2004	Reynolds *et al.*	2	197	Sequence features
2004	Ui - Tei *et al.*	6	72	Sequence features
2004	Amarzguioui *et al.*	4	46	Sequence features
2004	Hsieh *et al.*	22	138	Sequence features
2005	Jalag *et al.*	4	601	Sequence features

Bên cạnh đó, các phương pháp học máy cũng áp dụng để xây dựng mô hình để dự đoán hiệu quả ức chế của các siRNA, những kỹ thuật để xây dựng mô hình dự báo đã được coi là thế hệ thứ hai, khi thế hệ đầu tiên dựa trên tập dữ liệu nhỏ với một bảng đánh giá lượng [24], [41].

Mặc dù nhiều quy tắc thiết kế siRNA đã được báo cáo (Bảng 1.1). Kết quả có các quy tắc thiết kế có hiệu suất thấp và có cái hiệu quả, ngoài ra việc thử nghiệm các mô hình dự báo hiện tại rất ít trong khi dữ liệu của các siRNA là rất lớn, vì vậy để tạo ra nhiều siRNA hiệu quả cao vẫn là một thách thức.Các kỹ thuật tiên tiến nên được đề xuất để giải quyết vấn đề này và coi các kỹ thuật này là thế hệ thứ ba để tạo ra các siRNA hiệu quả cao.

Để tạo ra các siRNA hiệu quả cao, các nghiên cứu được tập trung việc giải quyết hai vấn đề đầu tiên. Trong luận văn này sẽ trình bày về việc tìm hiểu cách giải quyết hai vấn đề này

1.3. Kết luận

Các siRNA có thể được tổng hợp và đưa vào tế bào để làm ức chế gen đích dẫn việc tạo nhiều loại thuốc mới nhưng các siRNA làm ức chế các mRNA ở các cấp độ khác nhau nên việc tạo ra nhiều siRNA hiệu quả cao là một vấn đề rất quan trọng. Để tạo siRNA có hiệu quả cao trong cách tiếp cận sinh học và sinh học tính toán đã có nhiều quy tắc thiết kế siRNA đã được báo cáo có các quy tắc thiết kế có hiệu suất thấp và có cái hiệu quả. Ngoài ra việc thực hiện các mô hình dự báo hiện tại rất ít trong khi dữ liệu của các siRNA là rất lớn, vì vậy để tạo ra nhiều siRNA hiệu quả cao vẫn là một thách thức rất nhiều kỹ thuật tiên tiến nên được đề xuất để giải quyết vấn đề này. Trong luận văn này tập trung vào việc tìm hiểu những nghiên cứu của các nhà khoa học nhằm giải quyết vấn đề một và hai để tìm siRNA hiệu quả cao.

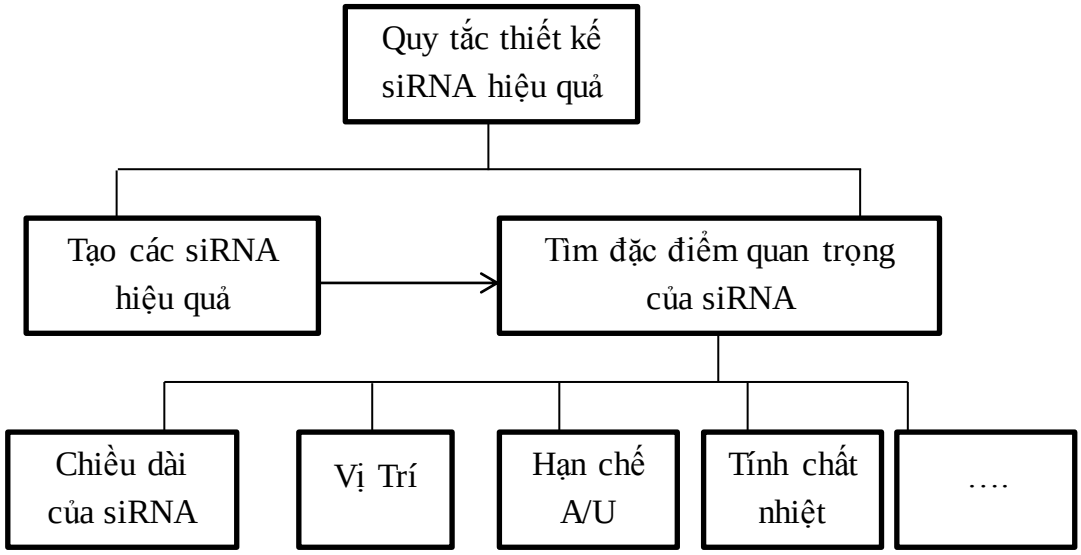
CHƯƠNG 2. CÁC QUY TẮC THIẾT KẾ siRNA HIỆU QUẢ

Trình bày khái quát các phương pháp đã được các nhà khoa học thực nghiệm để giải quyết vấn đề đề một của bài toán là tìm các quy tắc thiết kế siRNA hiệu quả trong cả hai cách tiếp cận sinh học và sinh học tính toán

2.1 Quy tắc thiết kế siRNA

Bài toán: Đầu vào là các chuỗi siRNA, sử dụng các phương pháp tiếp cận sinh học và sinh học tính toán để đưa ra các quy tắc thiết kế các siRNA hiệu quả.

Quy tắc thiết kế siRNA được tìm ra bởi đặc điểm ảnh hưởng đến hiệu quả của ức chế các siRNA, như chiều dài, vị trí, hạn chế tại A/U, tính chất nhiệt …Hình 2.1



Hình 2.1: Quy tắc thiết kế siRNA hiệu quả

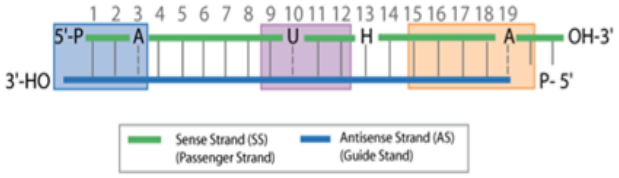
2.2. Quy tắc thiết kế siRNA hiệu quả trong phương pháp sinh học

Năm 1998 Fire và Mello đã khám phá ra vai trò quan trọng của dsRNA trong RNAi, dsRNA có thể được tổng hợp và tiêm vào tế bào để các sợi antisense ràng buộc với các mRNA. Sợi antisense với chiều dài đầy đủ không được phát hiện điều này dẫn đến tìm kiếm trên các sợi antisense ngắn (siRNA) có nguồn gốc từ các dsRNA. Năm 2001 Elbashir *et al*. [16] thấy rằng các siRNA có độ dài 19 đến 21 nucleotit với 2 nucleotit nhô ra ở hai đầu 3 'có ức chế mRNA hiệu quả khi họ đưa siRNA có độ dài 19 đến 21 nucleotide vào tế bào của chuột và người. Scherer *et a.*, đã báo cáo rằng các tính chất nhiệt động học ảnh hưởng quan trọng đối với mRNA. Ngay sau khi các công trình đầu tiên được công bố đã có một số quy tắc thiết kế được đưa ra (Hình 2.2). Sau đó nhiều quy tắc thiết kế hợp lý tạo nên các siRNA hiệu quả đã được báo cáo (Bảng 1.5). Đặc điểm của các quy tắc liên quan đến tính chất nhiệt, vị trí nucleotit, chiều dài, vị trí của các bazơ và chuỗi cụ thể…

Trong đó mặc dù các đặc điểm về vị trí được coi là yếu tố quan trọng nhất để xác định các quy tắc thiết kế siRNA một cách hiệu quả. Tuy nhiên có một số siRNA

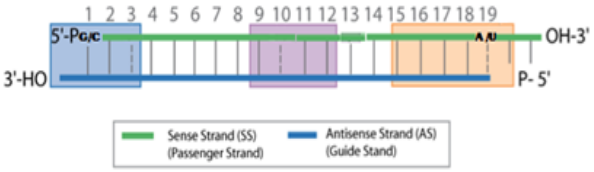
có kết quả rất tốt nhưng lại không giống với các tiêu chuẩn đề xuất. Trong khi rất nhiều siRNA thiết kế cẩn thận khác lại không hoạt động, gần đây ngay cả những giả thuyết cho rằng sự ổn định tương đối của hai đầu có ảnh hưởng đến hiệu quả của chúng. Việc tìm kiếm khả năng ức chế của siRNA không phải chỉ dựa vào các khảo sát thực nghiệm khác nhau của siRNA cũng không phải dựa vào quá trình phân tích toàn diện về siRNA được công bố hoặc các siRNA được đưa lên ngân hàng dữ liệu sẽ, các đặc điểm khác của siRNA cũng có thể đóng một vai trò quan trọng. Ngoài ra phân tích thực nghiệm trước đó chỉ dựa trên tập dữ liệu nhỏ và tập trung vào những gen cụ thể. Do đó những quy tắc này có thể không đủ thông tin để thiết kế các siRNA hiệu quả.

1	G / C hàm lượng 30- 52%
2	Ít nhất 3 'A hoặc' U là tại các vị trí 15- 19
3	Không lặp đi lặp lại bên trong
4	Một 'A' ở vị trí 19
5	Một 'U' ở vị trí 3
6	Một 'A' ở vị trí 10
7	Một bazơ khác so với 'G' hoặc 'C' ở vị trí 19
8	Một bazơ khác so với 'G' ở vị trí 13



Reynolds (*Nature Biotechno., 2004*)
197 siRNAs targeting 2 genes

1	A/U ở tại vị trí 19
2	G/C ở tại ví trí 1
3	Ít nhất 5 A/U tại vị trí 13 to 19
4	Không có mặt GC tại dải 9



Uitei (*Nucleic Acids Res., 2004*)
72 siRNAs targeting to 6 genes

Hình 2.2: Ví dụ về phát hiện ra quy tắc thiết kế siRNA hiệu quả trong cách tiếp cận sinh học

Các nghiên cứu với antisense của các phân tử DNA, RNA mạch đơn ngắn đã chỉ ra rằng khả năng tiếp cận của các khu vực bắt buộc đối với các RNA đích của chúng là quan trọng trong quá trình tạo ra ức chế có hiệu quả, một sự tương ứng giữa khả năng tiếp cận của chúng và siRNA đã được chứng minh. Trong một phân tích toàn diện hơn các RNA mục tiêu đã được thử nghiệm lặp đi lặp kết quả cho thấy các siRNA tại khu vực dễ tiếp cận dự đoán là hiệu quả hơn và sự ổn định nhiệt động học tương đối của hai đầu của siRNA đã được chứng minh.

Bên cạnh đó bản thân siRNA, RNA đích cũng có thể đóng một vai trò quan trọng trong sự ức chế, điều này có thể giúp giải thích tại sao một số có thể dễ dàng bị ức chế, một số khác khó khăn hơn. Trong một nghiên cứu với hàng ngàn siRNA với các gen khác nhau theo thuật toán BIOPREDsi, 70% gen kinase (một loại enzyme có vai trò chuyển hóa các gốc phosphate) khảo sát dễ dàng tạo ra im lặng, trong khi 6%

của các gen này có thể không được giảm biểu hiện (down-regulated) lên đến 10 siRNA khác nhau.

Các tính năng như định vị, nhiệt động học, cấu trúc bậc hai của siRNA được xem như là một yếu tố quan trọng để tìm quy tắc thiết kế siRNA. Sau đây là các quy tắc dự đoán quan trọng được tóm tắt trong các kết quả nghiên cứu sau

Quy tắc thiết kế Tuschl

Kỹ thuật này được sử dụng rộng rãi cho các thiết kế hiệu quả siRNA. Theo quy tắc này tổng hợp chuỗi siRNA có độ dài 19 nucleotit đến 21 nucleotit trên cơ sở ghép nối với 2 nucleotit 3' nhô ra ở cả hai đầu qua trung gian mRNA. Các kết quả nghiên cứu được tóm tắt dưới đây.

- Chọn khu vực có mục tiêu từ một chuỗi mRNA bắt đầu từ 50-100 nucletotit cùng hướng (downstream) bắt đầu từ codon (mã di truyền).
- Tìm kiếm 23 nucleotit với chuỗi là AA (N19) TT.
- Tìm kiếm 23 nucleotit với chuỗi là NA (N21).
- Đầu 3 'của siRNA có ý nghĩa với TT.
- Cuối cùng tìm kiếm NAR (N17) YNN, trong đó R = A, G và Y = C, T.
- Chuỗi mục tiêu cần phải có GC khoảng 50 tỷ.

Quy tắc thiết kế của Reynolds

Reynolds *et al*. [48] đã phân tích một tập hợp của 180 siRNA và đã chia các siRNA trong các nhóm khác nhau dựa trên chức năng của nó để tìm thuộc tính có mối tương quan cao với chức năng.

- < F50 - ức chế ít hơn 50%.
- > F50 - ức chế 50% hoặc nhiều hơn.
- > F80 - ức chế 80% hoặc nhiều hơn.
- > F95 - ức chế 95% hoặc nhiều hơn.

Đã mô tả tám nguyên tắc chủ yếu các chuỗi siRNA được đánh giá cao trong việc xác định mức độ ức chế mRNA được liệt kê dưới đây.

- G / C hàm lượng 30- 52%
- Ít nhất 3 'A hoặc' U là tại các vị trí 15-19
- Không lặp đi lặp lại bên trong
- Một 'A' ở vị trí 19
- Một 'U' ở vị trí 3
- Một 'A' ở vị trí 10
- Một bazơ khác so với 'G' hoặc 'C' ở vị trí 19
- Một bazơ khác so với 'G' ở vị trí 13

Thuật toán này chỉ định một số điểm dựa trên số lượng các quy tắc phù hợp và các siRNA thỏa mãn sáu hoặc nhiều các quy tắc được dự báo.

Quy tắc Amarzguioui

Một nghiên cứu khác của Amarzguioui *et al.*, [2]. Đó là đưa ra một phương pháp đánh giá tương tự nhưng xác định được một bộ các quy tắc khác nhau, họ đã nghiên cứu 46 siRNA và xác định các tính năng sau của 19 nucleotit siRNA tương quan với ức chế là hơn 70.

- Sự khác biệt về số lượng của A và U.
- Sự hiện diện của G hoặc C ở vị trí 1.
- Sự hiện diện của A tại vị trí 6.
- Sự vắng mặt của U ở vị trí 1.
- Sự vắng mặt của G ở vị trí 19.
- Sự hiện diện của A / U ở vị trí 19.

Mỗi quy tắc hoặc thêm hoặc bỏ đi một điểm thỏa mãn, những siRNA với số điểm là 3 hoặc nhiều hơn được coi là hiệu quả. Trong nghiên cứu này, các chức năng được chỉ định bởi một ức chế là 70.

Quy tắc thiết kế Stockholm

Đây là quy tắc được đưa ra bởi Chalk *et al*. [7] kết hợp các tính chất nhiệt động học của siRNA. Các quy định được gọi là quy tắc Stockholm được tóm tắt dưới đây.

- Tổng năng lượng kẹp tóc (hairpin) <1.
- Antisense 5 'năng lượng liên kết <9.
- Sense 5 ' năng lượng liên kết trong phạm vi 5 - 9.
- GC từ 36% đến 53%.
- Giữa (7-12) năng lượng liên kết <13.
- Chênh lệch năng lượng <0.
- Chênh lệch năng lượng trong phạm vị -1 và 0.

Quy tắc thiết kế Ui-Tei

Ui-Tei *et al*. [50] đã phân tích 72 siRNA trong các tế bào động vật có vú và các tế bào ruồi giấm và đã đưa ra với bốn tính năng mà cùng một lúc các siRNA phải đáp ứng để gây sự im lặng có hiệu quả. Những tính năng mà siRNA hiệu quả cần phải có là.

- A / U ở đầu 5 'của sợi antisense.
- G / C ở đầu 5 'của các sợi sense.
- Ít nhất là năm bazơ A / U từ các vị trí 13-19.
- Sự vắng mặt một vài GC căng ra của hơn 9 nucleotit

Những quy định này đã được tìm thấy đối với tế bào động vật có vú nhưng không áp dụng cho các tế bào ruồi giấm.

Quy tắc thiết kế Hseih

Hsieh *et al.* [20] thực hiện một thử nghiệm với 138 siRNAs và 22 gen có các đặc điểm sau:

- Nucleotide 'C' là tiêu cực ở vị trí 6
- Nucleoside 'C' hoặc G là tích cực và A hoặc U là tiêu cực ở vị trí 11
- Nucleotide 'A' là tích cực ở vị trí 13
- Nucleotide 'G' là tích cực ở vị trí 16
- Nucleotide 'U' là tích cực và nucleotide G là tiêu cực ở vị trí 19

Ngoài ra còn rất nhiều các quy tắc thiết kế dựa trên phương pháp tiếp cận sinh học đã được đưa ra

Mặc dù có rất nhiều các quy tắc thiết kế siRNA hiệu quả được đưa ra nhưng các quy tắc thiết kế được đề xuất lại không hoàn toàn giống nhau, có một số báo cáo phát hiện ở vị trí này, một số khác lại ở vị trí khác. Như là Reynolds đã không xem xét ở vị trí 1 nhưng các nhà khoa học khác được đề nghị rằng nên đặt G / C ở vị trí này và Huesken khuyến cáo rằng nó có kết quả tốt nếu siRNA có nucleotit trừ C ở vị trí 1. Reynolds và Huesken cũng xung đột với nhau khi quyết định thiết kế nucleotide ở vị trí 3 của siRNA. Mặt khác, khi kiểm tra các quy tắc với cơ sở dữ liệu siRecord dẫn đến việc tạo ra rất nhiều các quy tắc dẫn đến việc khó khăn cho quá trình tổng hợp siRNA hiệu quả, hơn nữa phân tích thực nghiệm ở trên chỉ dựa trên dữ liệu nhỏ và tập trung vào gen cụ thể, vì thế không đủ thông tin để thiết kế các siRNA hiệu quả.

Trong phương pháp sinh học, để thực nghiệm phải mất rất nhiều thời gian và tài chính vì vậy rất khó để xử lý trên tập dữ liệu lớn. Do đó nhiều nhóm nghiên cứu sử dụng kỹ thuật học máy trong nghiên cứu sinh học tính toán đó là áp dụng phương pháp học máy xây dựng mô hình cho việc tìm kiếm quy tắc thiết kế siRNA và dự đoán hiệu quả ức chế của siRNA. Sau đây là một số phương pháp học máy được áp dụng cho việc tìm kiếm quy tắc thiết kế siRNA.

2.3. Các quy tắc thiết kế trong cách tiếp cận sinh học tính toán

Trong phương pháp sinh học, các nhóm nghiên cứu phải mất rất nhiều thời gian và tài chính cho mỗi lần thực nghiệm. Do đó họ cũng có thể không xử lý trên tập dữ liệu lớn, nên đây có thể là một lý do các phương pháp được trong nghiên cứu trong cách tiếp cận sinh học là không đủ để thiết kế các siRNA hiệu quả.

Dựa trên các mục tiêu để tạo ra các quy quắc thiết kế siRNA hiệu quả mà trong phương pháp tiếp cận sinh học các nhà khoa học gặp một số các hạn chế nhất định. Các nhóm nghiên cứu chuyển sang hướng là tìm các các quy tắc thiết kế siRNA bằng

phương pháp sinh học tính toán. Bằng việc sử dụng kỹ thuật học máy xây dựng mô hình cho việc tìm kiếm quy tắc và dự đoán hiệu quả ức chế của siRNA (Bảng 2.1),

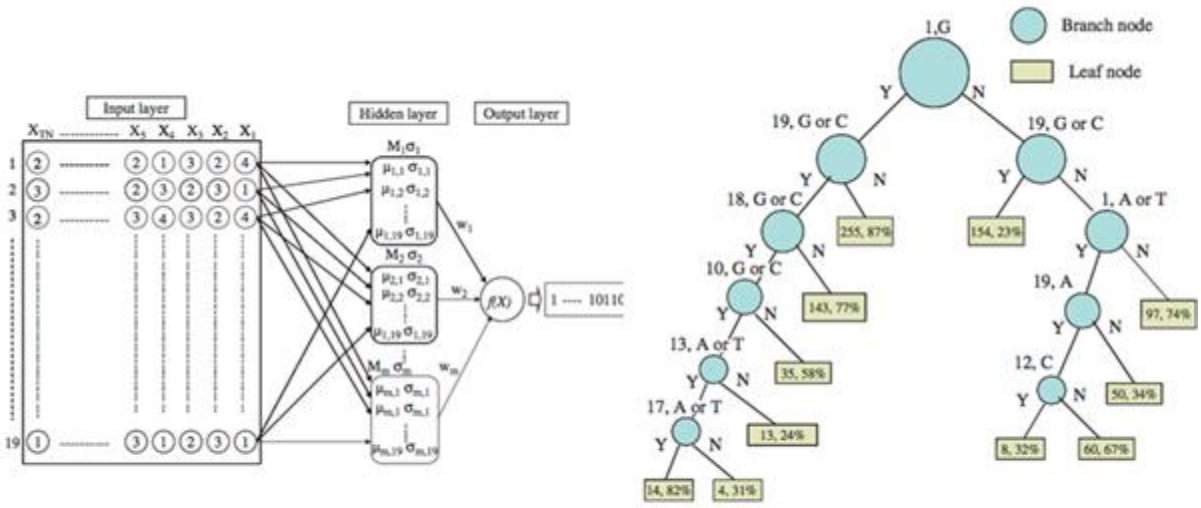
Bảng 2.1: Các mô hình tìm quy tắc thiết kế siRNA bằng phương pháp sinh học tính toán

Năm	Nhóm nghiên cứu	Số gen	Số siRNA	Công nghệ
2005	Teramoto *et al.*	2	94	SVM
2005	Huesken *et al.*	34	2182	Neural Networks
2007	Ludunga *et al.*	34	2252	SVM
2010	Takasaki *et al.*	490	833	Neural Networks Decision Tree

Liên quan đến việc tìm kiếm vấn đề quy tắc thiết kế siRNA, Teramoto [46] và đồng nghiệp sử dụng máy vector hỗ trợ (Support Vector Machine (SVM)) dựa trên nhân chuỗi tổng quát để chọn nhiều siRNA hiệu quả. Họ đã phát triển một thuật toán để dự đoán chức năng siRNA bằng cách sử dụng chuỗi kernel (GSK) kết hợp với các chương trình Libsvm để trích xuất tính năng trình tự và phân loại siRNA vào các lớp có hiệu quả và không hiệu quả bởi đại diện cho mỗi siRNA như chuỗi con k-mer, dựa trên các hệ số vector của mô hình. Họ cũng phát hiện 20 dấu hiệu đầu mà có thể được sử dụng để phân biệt các siRNA hiệu quả và không hiệu quả nhưng họ không thể suy ra một quy tắc thiết kế siRNA. Ladunga và đồng nghiệp [33] cũng sử dụng gói SVMLight với đa thức kernel để huấn luyện hơn 2200 siRNA, họ đã sử dụng 572 tính năng đại diện cho các siRNA quan đến đặc điểm trình tự, nhiệt động lực và khả năng tiếp cận. Higeru Takasaki và các đồng nghiệp của ông đề xuất phương pháp dự báo dựa trên các mạng nơron và cây quyết định (Hình 2.3). Để lựa chọn siRNA hiệu quả từ nhiều mục tiêu có thể [44, 45]. Đầu tiên, tác giả sử dụng thuật toán K-men để tính toán trong một giây, một cây quyết định được chia ra các nhánh, các dữ liệu thử nghiệm được sử dụng để kiểm tra các lỗi trong nhánh của cây, hơn nữa ông kết hợp hai phương pháp để tăng hiệu suất của các yếu tố dự báo. Mạng noron có một số hạn chế đó là các mối quan hệ của đặc điểm này là không rõ ràng, có thể quan sát được được và tạo ra kết quả khác nhau khi đào tạo lại với cùng một dữ liệu, ý nghĩa của cụm không được đề cập và khoảng cách Euclide cũng là không tốt để đánh giá sự tương tự của mỗi cặp siRNA. Như vậy thuật toán K-means trong trường hợp này có thể hiệu quả thấp. Hơn nữa, phương pháp cây quyết định không thể khái quát các dữ liệu vì hàm học được quá thích nghi với tập huấn luyện và kết quả cũng không ổn định vì sự thay đổi nhỏ trong dữ liệu có thể dẫn kết quả là đến cây khác nhau hoặc quy tắc thiết kế khác nhau.

Tóm lại các nhà nghiên cứu đã dùng cả hai cách tiếp cận với rất nhiều các quy tắc được tìm thấy để tìm kiếm siRNA hiệu quả cao nhưng đều có một hạn chế chung là không thống nhất giữa các quy tắc thiết kế siRNA. Hiệu năng đạt được rất thấp 20% siRNA tạo ra bởi các quy tắc không hoạt động, 65% siRNA tạo ra bởi quy tắc này hoạt

động không hiệu quả. Do vậy để tìm kiếm siRNA hiệu quả cao mục tiêu phải tiếp tục tìm ra các quy tắc thiết kế siRNA tốt hơn, đồng thời tìm ra các đặc điểm quan trọng của siRNA ảnh hưởng đến hiệu quả ức chế.



Takasaki (Comput Biol. Med., 2010)

Hình 2.3: Tìm quy tắc thiết kế dựa trên mạng nơron và cây quyết định

Trong quá trình nghiên cứu tìm kiếm quy tắc siRNA hiệu quả cao thì các nhà khoa học cũng đồng thời sử dụng các phương pháp học máy để xây dựng các mô hình dự đoán khả năng ức chế gen của siRNA.

2.4. Kết luận

Như vậy là để tạo siRNA có hiệu quả cao trong cả hai cách tiếp cận sinh học và sinh học tính toán đã có nhiều quy tắc thiết kế siRNA đã được đưa. Tuy nhiên vẫn còn nhiều hạn chế. Do đó để tạo ra quy tắc thiết kế siRNA hiệu quả cao ta vẫn phải tiếp tục nghiên cứu và thử nghiệm để tìm ra các quy tắc tốt hơn cũng như tìm ra các đặc điểm quan trọng của siRNA để phát hiện ra các quy tắc thiết kế hiệu quả.

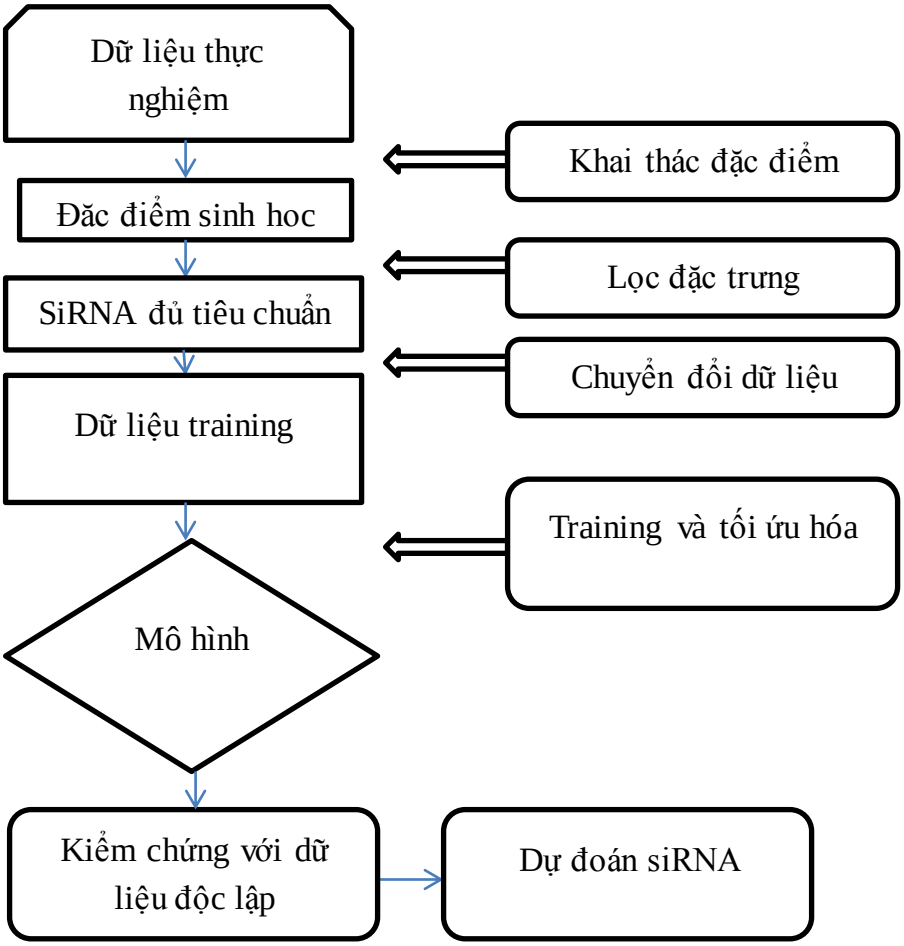
CHƯƠNG 3. PHƯƠNG PHÁP DỰ ĐOÁN KHẢ NĂNG ỨC CHẾ CỦA siRNA

Như đã trình bày ở các chương trước việc xây dựng mô hình dự báo dự đoán khả năng ức chế gen của siRNA là một trong hai vấn đề tạo siRNA hiệu quả cao. Trong chương này sẽ tập trung vào giới thiệu tổng quan về nghiên cứu xây dựng các mô hình dự báo và cách áp dụng các phương pháp học SVM và RF để dự đoán khả năng ức chế gen của siRNA. Đồng thời trình bày phương pháp học biểu diễn dùng để tiến hành thực nghiệm trong chương 4.

3.1. Tổng quan một số phương pháp xây dựng mô hình dự đoán ức chế của siRNA

Bài toán: Đưa vào tập dữ liệu siRNA được gán nhãn và một tập hợp các quy tắc thiết kế siRNA, sử dụng các phương pháp học máy để xây dựng mô hình dự báo đưa ra kết quả dự báo khả năng ức chế của siRNA.

Quy trình xây dựng các mô hình dự báo để đưa ra kết quả dự đoán khả năng ức chế của siRNA như Hình 3.1.



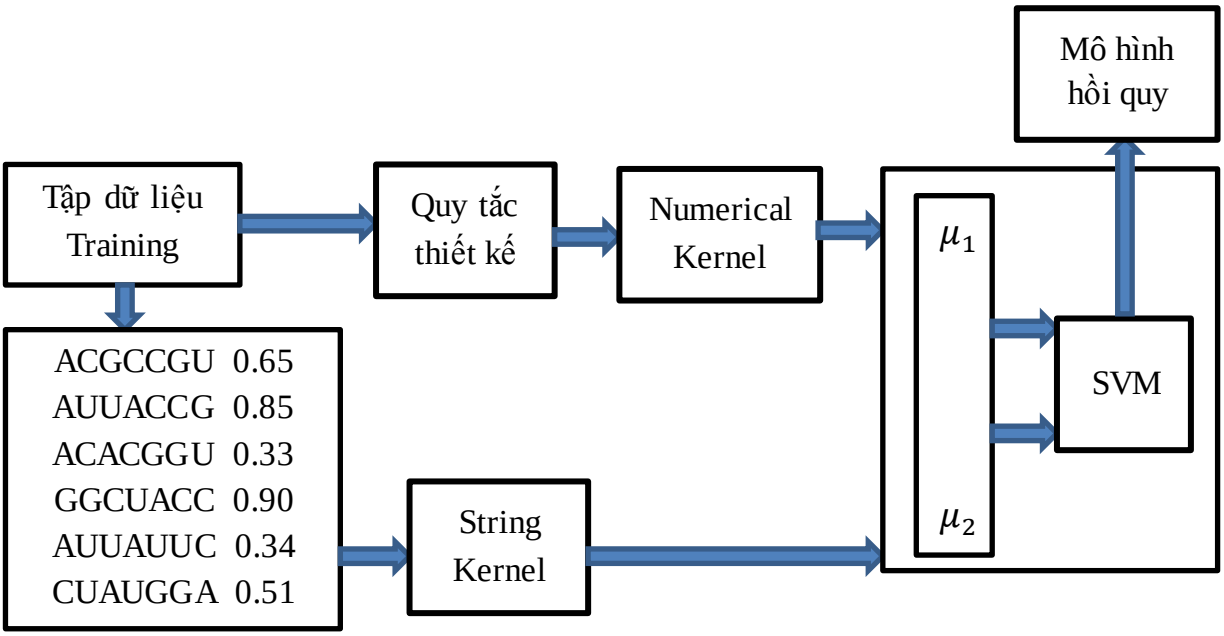
Hình 3.1: Quy trình xây dựng mô hình dự đoán khả năng ức chế của siRNA

Trong quá trình nghiên cứu về việc xây dựng mô hình dự báo hiệu quả ức chế của siRNA. Nhiều kỹ thuật học máy đã được áp dụng để dự đoán hiệu quả ức chế siRNA (Bảng 3.1)

Bảng 3.1: Các phương pháp học máy sử dụng xây dựng mô hình dự báo

Năm	Nhóm nghiên cứu	Dữ liệu	Công nghệ
2004	Chalk *et al.*	94	Regression tree
2005	Huesken *et al.*	2182	Neural Networks
2006	Shibalina *et al.*	Huesken Dataset	Linear regression
2006	Vert *et al.*	Huesken Dataset	Laso regression
2007	Ichihara *et al.*	Huesken Dataset	Linear regression
2009	Qui *et al.*	Huesken Dataset	MKSVR
2012	Mysara *et al.*	Huesken Dataset	Assemble learning
2013	Sciablola *et al.*	Huesken Dataset	SVR
2015	Bui Thang *et al.*	Huesken Dataset	Tensor regression

Chalk *et al* sử dụng tính chất nhiệt động học bằng cách sử dụng cây hồi quy trong phần mềm BioJava. Theo họ hệ số đánh giá của siRNA được gia tăng là (0, 7). Huesken *et al* đã đề xuất mô hình dự báo để nhận biết siRNA hiệu quả và không hiệu quả đã được phát hiện bởi một mạng nơron nhân tạo (ANN), được huấn luyện trên 2.182 siRNA và thử nghiệm với 249 siRNA đã đạt kết quả với R= 0.66. Bộ dữ liệu của họ đã được sử dụng rộng rãi và được thử nghiệm trong các mô hình hồi quy khác. Qui và các đồng nghiệp sử dụng mô hình vector hỗ trợ hồi quy đa nhân và cho dự đoán hiệu quả siRNA với R=0.62 với bộ dữ liệu Huesken gồm 2431siRNA(Hình 3.2). Đáng chú ý nhất Sciabola *et al* [41] sử dụng phương pháp học máy véc tơ hỗ trợ hồi quy và sử dụng cấu trúc ba chiều của siRNA để tăng khả năng dự báo của mô hình hồi quy đạt kết quả với R=0.8.



Hình 3.2: Ví dụ sử dụng mô hình SVR dự đoán khả năng ức chế của siRNA

Ngoài ra một số nhóm nghiên cứu [31, 13] sử dụng phương pháp phân lớp (classification methods) trên các siRNA đã được gán nhãn để thực nghiệm về khả năng ức chế có hiệu quả. Với tập dữ liệu siRNA được lấy từ cơ sở dữ liệu siRecord [47] bao

gồm siRNA phân thành bốn lớp với nhãn 'rất cao', 'cao', 'trung bình' và 'thấp', các phương pháp phân loại xây dựng phân lớp từ siRNA với bộ dữ liệu có gán nhãn để dự đoán nhãn lớp của siRNA chưa biết.

Peng Jiang và Liangjiang Wang [32] đã giới thiệu phương pháp rừng ngẫu nhiên (random forest (RF)) để dự đoán khả năng ức chế của siRNA. Với kết quả thu được cao hơn một số thuật toán học máy khác như SVR và các thuật khác trước đó. Mô hình RF đã đạt hiệu suất tốt nhất với một máy chủ (web-server) là RFRCDB-siRNA đã được phát triển, RFRCDB-siRNA bao gồm hai phần: Một cơ sở dữ liệu siRNA-trung tâm và một hệ thống dự báo RF, RFRCDB-siRNA hoạt động như sau: (1) Thay vì trực tiếp dự đoán các hoạt động ức chế gen của các siRNA, máy chủ có những siRNA để truy vấn và tìm kiếm cơ sở dữ liệu các chuỗi phù hợp được lưu giữ. (2) Các chuỗi không phù hợp sau đó được chuyển tới các hệ thống dự báo RFR để phân tích thêm.

Đã có rất nhiều phương pháp học máy được sử dụng để thử nghiệm nhưng hầu hết các phương pháp đó bị một số nhược điểm, R chỉ từ từ 0, 60 tới 0, 68 và bị giảm đáng kể khi thử nghiệm trên bộ dữ liệu độc lập. Bởi thực tế rằng các số liệu Huesken vẫn còn quá nhỏ để có thể đại diện cho siRNA có khoảng 4^{19} siRNA. Ngoài ra việc thực hiện các phương pháp học máy phụ thuộc rất nhiều vào sự lựa chọn phương pháp biểu diễn dữ liệu đang áp dụng. Đó là một lý do tại sao nhiều nỗ lực thực tế trong việc triển khai các thuật toán học máy đi vào việc tìm các phương pháp biểu diễn có thể hỗ trợ các phương pháp học máy hiệu quả.

Trong các mô hình trước đó siRNA được mã hóa bởi hệ nhị phân, quang phổ, tứ diện, và chuỗi, tuy nhiên nó không thể đại diện cho siRNA để xây dựng một mô hình tốt để dự đoán hiệu quả ức chế chính xác của siRNA.

Với bài toán dự đoán khả năng ức chế của siRNA thì hai phương pháp học máy véc tơ hỗ trợ của Vapnik (SVM). Rừng ngẫu nhiên của Breiman (RF), được biết đến như là những giải thuật phân lớp hiệu quả các tập dữ liệu có số chiều lớn như dữ liệu gien. Khi xử lý dữ liệu có số chiều lớn và số phần tử ít như dữ liệu gen thì rừng ngẫu nhiên và SVM là hai giải thuật học nhanh, chịu đựng nhiễu tốt và không bị tình trạng học vẹt, điều này ngược lại với AdaBoost, ArcX4 rất dễ bị học vẹt và ảnh hưởng lớn với nhiễu. Sau đây là chi tiết hai phương pháp SVM và RF được áp dụng vào bài toán dự đoán khả năng ức chế gen siRNA.

3.2. Phương pháp máy vecto hỗ trợ (SVM- Support vector machine)

Trong những thập kỷ gần đây, các nghiên cứu về gen và di truyền phát triển và đã có những thành công nhất định, đồng thời cũng tạo ra một khối lượng lớn các dữ liệu đa dạng về gen sinh học. Tuy nhiên, để có thể khám phá và khai thác những thông tin quý giá trong các dữ liệu này và để hiểu về các hệ thống sinh học, thì ta phải cần đến các phương pháp tính toán phức tạp với các giải thuật tính toán chính xác và hiệu

quả. Rất nhiều vấn đề quan trọng trong sinh học tính toán (Computational Biology) liên quan đến bài toán phân lớp (classification) hay dự báo (prediction), như: Dự báo vị trí cắt-nối (splice site prediction) để tìm kiếm gen, dự báo cấu trúc gen, chức năng của gen, sự tương tác, và vai trò của gen trong một số loại bệnh tật v.v. Một trong những kỹ thuật tính toán nổi tiếng cho bài toán phân lớp/dự báo cho độ chính xác cao và được sử dụng rộng rãi trong cộng đồng nghiên cứu tin sinh học trong những năm gần đây là kỹ thuật phân lớp sử dụng máy vec-tơ hỗ trợ SVM (support vector machine).

Trong phần này giới thiệu những vấn đề cơ bản của lý thuyết học máy (machine learning) cho bài toán đoán nhận khả năng ức chế của siRNA sử dụng SVM (Hình 3.2). Đó là một phương pháp nổi tiếng dựa trên việc cực đại hóa dải biên phân lớp (max margin classification) và việc lựa chọn các hàm nhân (kernel) phù hợp. Phương pháp này được sử dụng rộng rãi để giải quyết nhiều bài toán của tin sinh học do tính hiệu quả, độ chính xác cao, và khả năng xử lý đối với các bộ dữ liệu lớn.

Máy véc tơ hỗ trợ SVM

Máy vector hỗ trợ là một tập hợp các phương pháp học có giám sát bao gồm phân tích dữ liệu và phát hiện mẫu, được sử dụng cho phân lớp và phân tích hồi quy, là một giải thuật máy học dựa trên lý thuyết học thống kê do Vapnik and Chervonenkis (1974), Vapnik (1999) xây dựng. Sau đó, Corinna Cortes cùng với Vladimir Vapnik đề xuất hình thức chuẩn hiện nay.

Bài toán cơ bản của SVM là bài toán phân loại hai lớp: Cho trước r điểm trong không gian n chiều (mỗi điểm thuộc vào một lớp kí hiệu là +1 hoặc –1), mục đích của giải thuật SVM là tìm một siêu phẳng (hyperplane) phân hoạch tối ưu cho phép chia các điểm này thành hai phần sao cho các điểm cùng một lớp nằm về một phía với siêu phẳng này (Hình 3.3).

Xét tập r mẫu huấn luyện $\{(x_1, y_1), (x_2, y_2), \ldots, (x_r, y_r)\}$. Trong đó x_i là một vector đầu vào được biểu diễn trong không gian $X \subseteq R^n$, y_i là một nhãn lớp; $y_i \in \{1, -1\}$. Siêu phẳng tối ưu phân tập dữ liệu này thành hai lớp là siêu phẳng có thể tách rời dữ liệu thành hai lớp riêng biệt với lề (margin) lớn nhất, tức là cần tìm siêu phẳng H_0: $y = w.x + b = 0$ và hai siêu phẳng H_+, H_- hỗ trợ song song với H_0 và có cùng khoảng cách đến H_0. Với điều kiện không có phần tử nào của tập mẫu nằm giữa H_+ và H_-, khi đó:

H_+: $w.x + b >= +1$ với y = +1

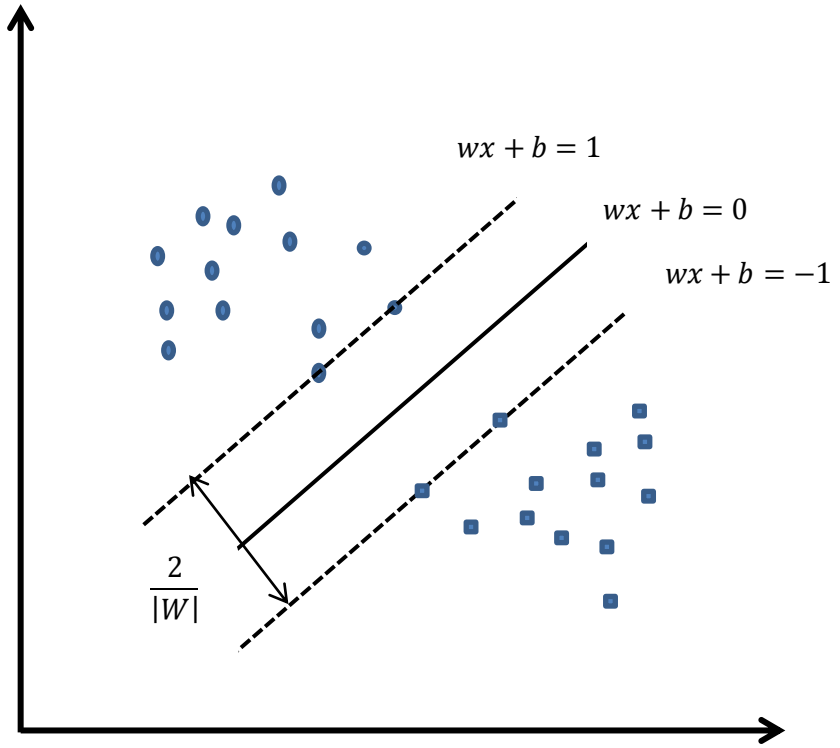
H-: $w.x + b >= -1$ với y = -1

Kết hợp hai điều kiện trên, có $y\ (w.x + b) >= 1$. Khoảng cách của siêu phẳng H_+ và H_- đến H_0 là $||w||$ Ta cần tìm siêu phẳng H với lề lớn nhất, tức là giải bài toán tối ưu tìm $\min_{w,b} \|w\|$ với ràng buộc $y\ (w.x + b) >= 1$.

Người ta có thể chuyển bài toán sang bài toán tương đương nhưng dễ giải hơn là $\min_{w,\, b} \frac{1}{2}\|w\|^2$ với ràng buộc *y (w.x + b) >= 1*. Lời giải cho bài toán tối ưu này là cực tiểu hóa hàm Lagrange:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\|w.x_i\| + b) - 1]$$

Trong đó α là các hệ số Lagrange với α≥0



Hình 3.3: Siêu phẳng với lề cực đại trong không gian R^2

Lời giải tìm siêu phẳng tối ưu trên có thể mở rộng trong trường hợp dữ liệu không thể tách rời tuyến tính bằng cách ánh xạ dữ liệu vào một không gian có số chiều lớn hơn bằng cách sử dụng một hàm nhân K (kernel). Một số hàm nhân thường dùng

- Hàm tuyến tính có dạng $K(x, y) = x.y$
- Hàm đa thức có dạng $K(x.y) = (x.y + 1)^d$
- Hàm RBF (Radial Basis Function) có dạng $K(x, y) = e^{\frac{-|x-y|^2}{2a^2}}$

Với khả năng vượt trội của SVM về tính hiệu quả, độ chính xác, khả năng xử lý các bộ dữ liệu một cách linh hoạt. Việc sử dụng máy vectơ hỗ trợ SVM đã và đang là sự lựa chọn tối ưu nhất trong việc giải quyết các bài toán phân loại, dự báo trong một số các ngành khoa học và trong nghiên cứu siRNA. Đã có rất nhiều nhóm nghiên cứu dựa trên SVM để áp dụng tìm ra các quy tắc thiết kế dự đoán siRNA. Sau đây là ví dụ về quá trình thực nghiệm việc sử dụng mô hình SVM sử dụng chuỗi tổng quát kernel [37].

Dự đoán các chức năng siRNA sử dụng chuỗi tổng quát kernel và máy vector hỗ trợ

Từ tập dữ liệu Khvorova chứa các siRNA, với 94 siRNA gồm các luciferase và gen cyclophilin B của người phân thành hai lớp chức năng, hiệu quả và không hiệu quả. Trong đó lớp hiệu quả chứa 53 siRNA với 90% khả năng ức chế gen và lớp không hiệu quả chứa 41 các siRNA có ít hơn 50% khả năng ức chế gen.

Quy tắc thiết kế siRNA

Chuỗi tổng quát kernel (GSK) dựa trên chuỗi kernel bất đối xứng (MSK-mismatch kernels) và phổ kernel, được ký hiệu là ((k, m)-MSK). Trong đó k chính là độ dài những đoạn có độ dài k có thể tạo được từ tất cả các ký tự thuộc tập Σ (Σ = {A, C, T, G} đối với các chuỗi ADN) từ các chuỗi con khác nhau bởi ít nhất một m bất đối xứng trong đó m quy định số lượng ký tự tối đa khi đếm số lần xuất hiện của một k-mer trên một chuỗi. Một vectơ đặc trưng cho một chuỗi x có độ dài cố định, xác định

$$K_{(k, m)}(x, y) = \langle \emptyset_{(k,m)}(x), \emptyset_{(k,m)}(y) \rangle.$$

Trong trường hợp m=0, $K_{(k, 0)}(x, y)$ là k-spectrum kernel với m#0 sử dụng $K_{(k, m)}(x, y)$ như sau:

$$K_{(k, m)}(x, y) \leftarrow \frac{K_{(k,m)}(x,y)}{\sqrt{K_{(k,m)}(x,x)}\sqrt{K_{(k,m)}(y,y)}}$$

GSK là tổng của tất cả các (k_i, m_i) –MSK

Các (k_1,m_1,...,k_s,m_s) - GSK $K_{(k1,m1,...,ks,ms)}(x,y)$ được định nghĩa như sau:

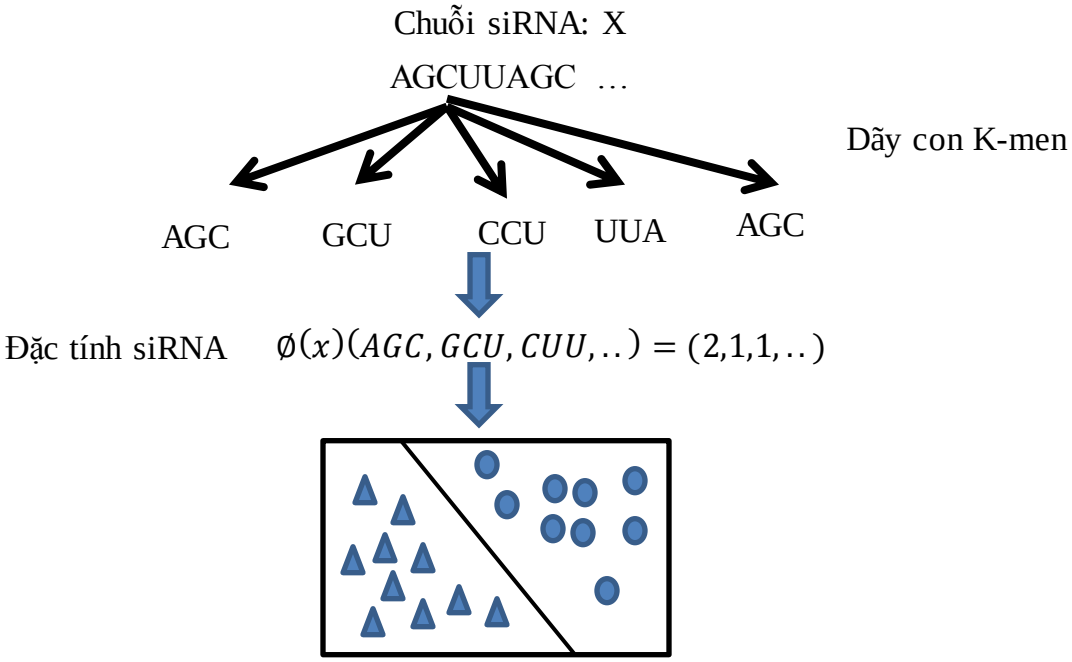
$$K_{(k1, m ,..., ks, ms)}(x,y) = \sum_i \langle \emptyset_{(k_i,m_i)}(y), \emptyset_{(k_i,m_i)}(y) \rangle = \sum_i K_{(k_i,m_i)}(x, y)$$

Thực nghiệm

Các đặc trưng được rút ra từ trình tự siRNA

Cơ sở của phương pháp tiếp cận đó là để mô tả trình tự siRNA, vectơ trong không gian đặc trưng đa chiều phản ánh các số 1, 2 và 3 chuỗi con trong mỗi siRNA và đưa các vector biểu diễn được học vào một thuật toán máy học có giám sát là SVM. Để đưa ra các tính năng từ trình tự siRNA phương pháp đã sử dụng GSK với một tập dữ liệu thử nghiệm của siRNA được công bố bởi Khvorova *et a*., đại diện cho 53 siRNA hiệu quả và 41 siRNA không hiệu quả (Hình 3.4).

Với GSK của k-mer, là 1-mer (1-GSK), 2-mer (2-GSK), hoặc 3-mer chỉ (3-GSK), hoặc GSK của tất cả các 1- 3-mer chuỗi con ((1,2,3) -GSK), chúng ta có thể phân loại các bộ dữ liệu thử nghiệm với độ chính xác 55,3%, 80,9%, 87,2% và 86,2%, tương ứng (Bảng 3.2). Những kết quả này chỉ ra rằng hiệu suất phân biệt cao với 3-GSK, và (1, 2, 3) -GSK hơn với 1-GSK hoặc 2-GSK. Bảng 3.3 cho thấy một danh sách cao nhất của 20 vector trọng lượng SVM cho (1, 2, 3) –GSK



Hình 3.4: Ví dụ của GSK

Bảng 3.2: So sánh hiệu suất phân biệt giữa 1-, 2-, 3- và (1, 2, 3) - GSK/SVM

Nhãn	TP	TN	FP	FN	Chính xác
1 - GSK	37	15	26	16	55.3% (52/94)
2 - GSK	44	32	9	9	80.9% (76/94)
3 - GSK	49	33	8	4	87.2% (82/94)
1, 2, 3 - GSK	48	33	8	5	86.2% (81/94)

Giá trị tuyệt đối của vector trọng lượng SVM cho mỗi dãy thể hiện sự quan trọng trong việc phân lớp mặc dù từ 17 đến 20 tốp đầu SVM bắt nguồn từ chuỗi con 3-mer và (10, 15, và 17) là từ chuỗi con 2-mer. Các vector trọng lượng bắt nguồn chuỗi con 1-mer tương ứng C, A, G và U là 0,087, 0,055, 0,030, và 0,027, những kết quả này chỉ ra rằng các tính năng bắt nguồn từ hoặc 1-mer, hoặc 2-mer vẫn có những đóng góp đáng kể vào việc thực hiện phân biệt, vì vậy nó được sử dụng (1, 2, 3) -GSK để phân tích thêm.

Hình 3.5A cho thấy phân bố các điểm GSK / SVM cho 94 siRNA, với 90, 6% hiệu quả và 80,5% của siRNA không hiệu quả có điểm tích cực và tiêu cực, tương ứng. Trong hình 3.5B, đồ thị cho thấy các tần xuất cộng dồn của siRNA hiệu quả được sắp xếp theo thứ tự điểm số GSK / SVM với các siRNA không hiệu quả, trong đó cả 36 siRNA đầu tiên và 24 siRNA cuối cùng đã được phân loại tương ứng là có hiệu quả và không hiệu quả

Bảng 3.3: Danh sách 20 của vectơ trọng lượng SVM với (1,2,3)-GSK

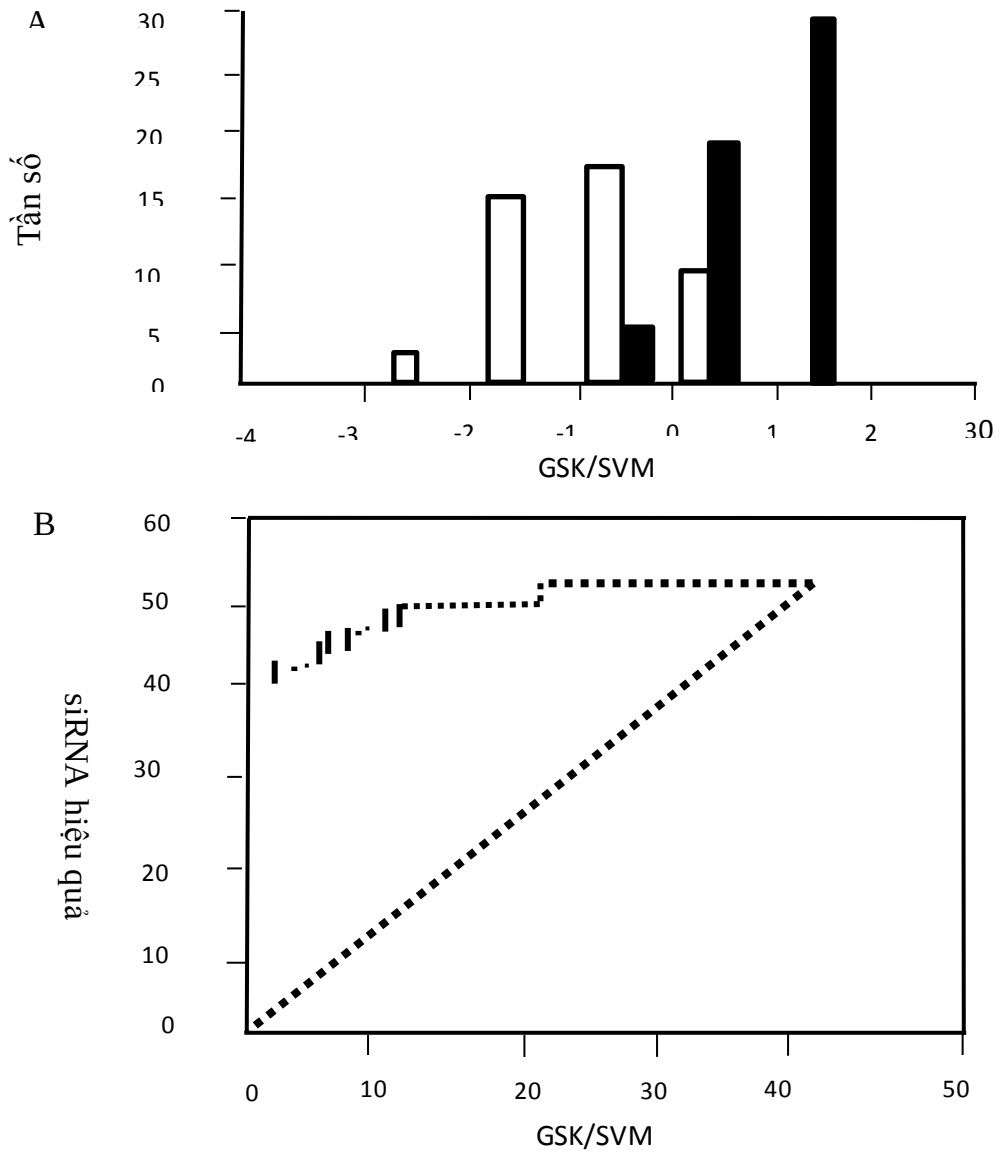
Thứ tự	Chuỗi con	Trọng Lượng
1	GAC	0.599
2	GGA	0.374
3	AU	0.368
4	UGC	0.338
5	CAA	0.334
6	AGC	0.317
7	CAU	0.3001
8	GGC	0.300
9	UGA	0.283
10	UG	0.276
11	AAG	0.274
12	CUG	0.268
13	CUC	0.265
14	GAG	0.253
15	GA	0.240
16	GCA	0.231
17	GU	0.230
18	UUC	0.228
19	CCA	0.224
20	CUU	0.198

Thanh màu đen và trắng thanh hiển thị phân bố các điểm GSK / SVM cho các siRNA hiệu quả và không hiệu quả

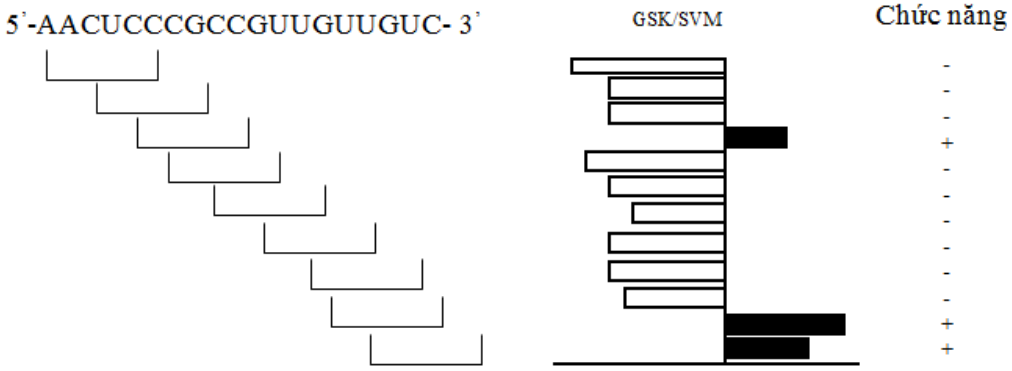
Hình 3.6 cho thấy ví dụ về trình tự siRNA cùng với GSK / SVM như đã trình bày chỉ trừ một trường hợp là GSK / SVM có thể phân biệt các siRNA hiệu quả từ siRNA không hiệu quả, những kết quả này gợi ý rằng các tính năng của siRNA chiết xuất bởi GSK đúng cách có thể đại diện cho siRNA.

Dấu ngoặc để xác định vị trí của siRNAvà điểm GSK / SVM cho mỗi siRNA được biểu diễn dưới dạng đồ thị thanh trong đó thanh đen và thanh trắng chỉ ra điểm tích cực và tiêu cực tương ứng với chức năng; Hiệu quả siRNA (+), siRNA không hiệu quả (-).

Để kiểm tra GSK/SVM cho dự đoán siRNA sử dụng LOOCV (Leave-one-out cross-validation) với 94 siRNAs có kết quả 75, 5% hiệu quả và 68, 3% không hiệu quả có điểm tích cực và tiêu cực, tương ứng. Độ chính xác tổng thể là 72, 3% (= 68/94), trong đó 40 là tích cực thực sự, 28 tiêu cực, 13 tích cực giả, và 13 tiêu cực tính giả.



Hình 3.5: Phân loại các dữ liệu thử nghiệm bởi thuật toán GSK / SVM



Hình 3.6: Mối quan hệ giữa tự luciferase siRNA và điểm GSK / SVM

Với kết quả đạt được độ chính xác là 72,3% là kết quả dự đoán khá cao trong các mô hình dự báo về khả năng ức chế của siRNA và đã phát hiện ra 20 dấu hiệu có thể được sử dụng để phân biệt các siRNA hiệu quả và không hiệu quả. Do đó có thể áp dụng phương pháp này để thử nghiệm trên bộ dữ liệu chuẩn khác, liên quan đến việc xây dựng mô hình dự báo có rất nhiều nhóm nghiên cứu đã tiến hành thực nghiệm với dữ liệu và phương pháp biểu diễn khác nhau như, Ladunga và đồng nghiệp cũng sử dụng gói SVMLight với đa thức kernel để huấn luyện hơn 2200 siRNA, họ đã sử dụng 572 tính năng đại diện cho các siRNA liên quan đến đặc điểm trình tự, nhiệt động lực và khả năng tiếp cận. Peilin Jia sử dụng thuật toán SVM bằng cách biểu diễn chuỗi hệ thống nhị phân có chiều dài cố định sử dụng bộ dữ liệu của Dieter và Reena Murali sử dụng SVM dựa vào các báo cáo hoạt động ức chế gen đã chia các siRNA thành hai loại, các siRNA với tỷ lệ hơn 60% hoạt động làm ức chế gen được coi là hiệu quả và siRNA ít hơn 30% được coi là không hiệu quả, ngoài ra còn rất nhiều các nhóm nghiên cứu khác cũng đã áp dụng phương pháp SVM áp dụng vào bài toán siRNA. Nhưng hiệu quả vẫn còn thấp, ngoài phương pháp SVM thì phương pháp rừng nhiên cũng được thử nghiệm và có kết quả tốt.

3.3. Phương pháp rừng ngẫu nhiên (Random Forest)

Phân lớp dữ liệu có số chiều lớn có nhiễu như dữ liệu gen (mỗi chiều cung cấp rất ít thông tin cho tách lớp) được biết là một trong 10 vấn đề khó của cộng đồng khai phá dữ liệu. Mô hình học phân lớp thường cho kết quả tốt trong khi huấn luyện lại cho kết quả rất thấp khi dự báo, vấn đề khó khăn thường gặp chính là số chiều quá lớn lên đến hàng nghìn chiều thậm chí đến cả triệu và dữ liệu thường tách rời nhau trong không gian có số chiều lớn việc tìm mô hình phân lớp tốt có khả năng làm việc với dữ liệu có số chiều lớn là khó khăn do có quá nhiều khả năng lựa chọn mô hình. Việc tìm một mô hình phân lớp hiệu quả (phân lớp dữ liệu tốt trong tập thử) trong không gian giả thiết lớn là vấn đề khó. Phương pháp rừng ngẫu nhiên cho độ chính xác cao khi so sánh với các thuật toán học có giám sát hiện nay, bao gồm cả AdaBoost, ArcX4, và SVM.

Thuật toán rừng ngẫu nhiên

Thuật toán tạo một rừng ngẫu nhiên được phát triển bởi Leo Breiman và Adele Cutler, thuật ngữ Random Forest được lấy làm tên phổ biến cho thuật toán này, thuật ngữ RF được xuất lần đầu tiên năm 1995. Sau đó kết hợp với phương pháp "bagging" trong lựa chọn các thuộc tính ngẫu nhiên của Leo Breiman năm 1996 để xây dựng phương pháp chọn các cây quyết theo các thay đổi có thể kiểm soát được. Năm 2001 Breiman xây dựng thuật toán RF có bổ sung thêm một lớp ngẫu nhiên để phân lớp, ngoài việc xây dựng mỗi cây sử dụng các mẫu dữ liệu khác nhau, các rừng ngẫu nhiên được thay đổi để xây dựng các cây phân loại và hồi quy khác nhau, các gói thư viện cài đặt thuật toán RF được xây dựng bằng ngôn ngữ Fortran bởi Leo Breiman và Cutler

Thuật toán RF - Random Forest là một thuật toán đặc biệt dựa trên kỹ thuật lắp ghép. Về bản chất thuật toán RF được xây dựng dựa trên nền tảng thuật toán phân lớp cây phân loại và hồi quy, sử dụng kỹ thuật có tên gọi là "bagging", thuật toán này cho phép lựa chọn một nhóm nhỏ các thuộc tính tại mỗi nút của cây để phân chia cho mức tiếp theo của cây phân lớp, bằng cách chia nhỏ không gian tìm kiếm thành các cây nhỏ hơn như vậy cho phép thuật toán có thể phân loại một cách rất nhanh chóng cho dù không gian thuộc tính rất lớn. Các tham số đầu vào của thuật toán khá đơn giản bao gồm các thuộc tính được chọn trong mỗi lần phân chia. Giá trị mặc định của tham số này là căn bậc hai của p với p là số lượng các thuộc tính, số lượng cây được tạo ra là không hạn chế và cũng không sử dụng bất kỳ kỹ thuật nào để hạn chế mở rộng cây, phải lựa chọn tham số cho biết số lượng cây sẽ được sinh ra sao cho đảm bảo rằng sẽ mỗi một thuộc tính sẽ được kiểm tra một vài lần. Thuật toán sử dụng kỹ thuật "out of bag" để xây dựng tập huấn luyện và phương pháp kiểm tra trên nó

Random Forest (rừng ngẫu nhiên) là một phương pháp phân lớp và hồi quy dựa trên việc kết hợp kết quả dự đoán của một số lượng lớn các cây quyết định. Có thể liên tưởng tới việc bầu cử theo nguyên tắc phổ thông đầu phiếu, nếu sử dụng một cây quyết định chẳng khác nào việc bầu cử mà chỉ có một người bỏ phiếu, việc sinh các cây quyết định từ một mẫu dữ liệu nhằm đa dạng hoá các "phiếu bầu" (giống như việc mọi thành phần, tầng lớp, giai cấp đều được đi bỏ phiếu) cho kết luận. Việc áp dụng các kỹ thuật sinh ra các mẫu dữ liệu hay việc lựa chọn rẽ nhánh ngẫu nhiên sẽ tạo ra các cây "dị tật" trong rừng (giống việc cho phép công dân không cần phân biệt trình độ học vấn, sức khỏe... đi bầu cử), càng nhiều loại hình, càng nhiều phiếu bầu sẽ cung cấp cho chúng ta cái nhìn đa chiều, chi tiết hơn và do đó kết luận sẽ có tính chính xác, gần với thực tế hơn. Trong thực tế RF đã trở thành một công cụ tin cậy cho phân tích dữ liệu đặc biệt là dữ liệu tin sinh học.

Trong random forest, sự phát triển của một tập hợp các cây đã làm cải thiện một cách đáng kể độ chính xác phân lớp, mỗi cây trong tập hợp sẽ "bỏ phiếu" cho lớp phổ biến nhất, để phát triển các tập hợp cây này thông thường các véc tơ ngẫu nhiên được tạo ra, các véc tơ này sẽ chi phối sự phát triển của mỗi cây trong các tập nói trên. Đối với cây thứ k trong tập các cây, một véc tơ ngẫu nhiên Θk được tạo ra, véc tơ này độc lập với các véc tơ được tạo ra trước đó Θ1, Θ2, …, Θk-1 nhưng sự phân bố của các véc tơ này là tương tự nhau, một cây được phát triển dựa vào tập huấn luyện và véc tơ Θk kết quả là được một phân lớp h(x, Θk) trong đó x là véc tơ đầu vào sau khi một số lượng lớn các cây được tạo ra các cây này "bỏ phiếu" cho lớp phổ biến nhất.

Random forest được định nghĩa như sau .Một random forest là một phân lớp bao gồm một tập các phân lớp có cấu trúc cây {h(x, Θk), k=1,... trong đó {Θk} là những véc tơ độc lập, tương tự nhau được phân bố một cách ngẫu nhiên và mỗi cây sẽ bỏ một phiếu bầu cho lớp phổ biến nhất ở véc tơ đầu vào x.

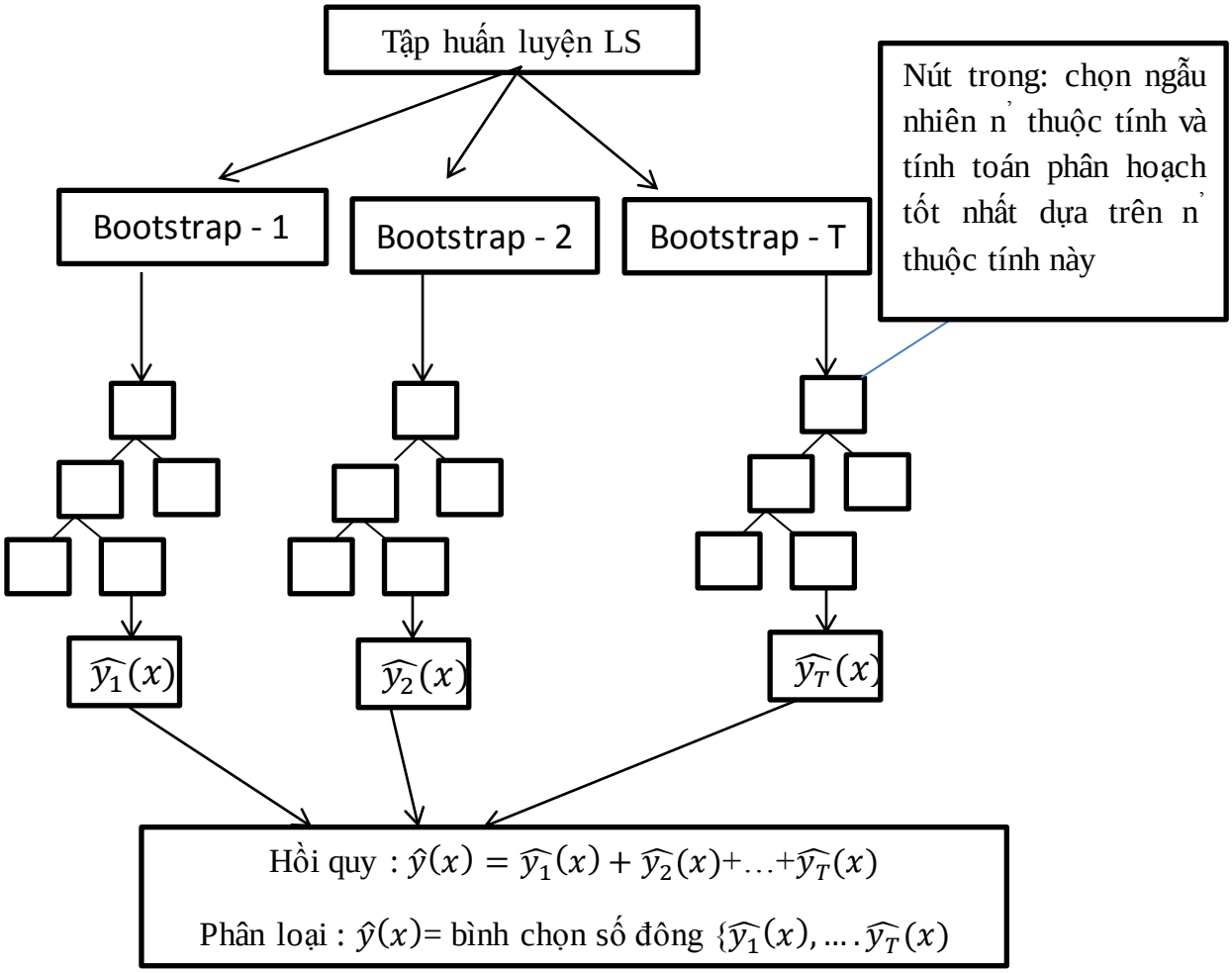
Ý tưởng chính của giải thuật random forest (Hình 3.7):

+ Từ tập học LS có N phần tử
+ Xây dựng tập hợp T mô hình cơ sở độc lập nhau
+ Mô hình thứ i được xây dựng trên tập mẫu bootstrap,

- Tại nút trong, chọn ngẫu nhiên n' thuộc tính (n'<<n) và tính toán phân hoạch tốt nhất dựa trên n' thuộc tính này

- Cây được xây dựng đến độ sâu tối đa không cắt nhánh

- Một bootstrap : Lấy mẫu N phần tử có hoàn lại từ tập LS
- Khi phân loại : Sử dụng majority vote(Đa số phiếu)
- Hồi quy : Tính giá trị trung bình của dự đoán của các mô hình



Hình 3.7: Giải thuật rừng ngẫu nhiên cho phân lớp dữ liệu

Việc áp dụng phương pháp rừng ngẫu nhiên cho siRNA cũng được rất nhiều nhóm nghiên cứu áp dụng như là. Peng Jiang đã nâng cao thiết kế của siRNA bằng mô hình hồi quy rừng ngẫu nhiên kết hợp với tìm kiếm cơ sở dữ liệu, Liangjiang Wang đã sử dụng SVM và RF để dự đoán hiệu năng siRNA. Simone sử dụng cả ba phương pháp học máy PLS (Phương pháp bình phương tối thiểu từng phần), SVM, RF để cải thiện mô tả cho siRNA và một số báo cáo của các nhà nghiên cứu khác. Kết quả cho thấy RF luôn có độ chính xác cao hơn so với các phương pháp học máy khác. Sau đây

là một ví dụ cải tiến thiết kế của siRNA bằng mô hình hồi quy rừng ngẫu nhiên kết hợp với cơ sở dữ liệu tìm kiếm [36]

Dữ liệu

Dựa trên 3589 siRNA từ 9 bộ dữ liệu đã được công bố của [21] bao gồm 2.431 siRNA đã được sử dụng để xây dựng và tối ưu hóa các mô hình hồi quy rừng ngẫu nhiên. Trong đó bao gồm 573 siRNA đã được sử dụng như một bộ dữ liệu độc lập để đánh giá mô hình RFR.

Phương pháp hồi quy rừng ngẫu nhiên

Rừng ngẫu nhiên lần đầu tiên được đề xuất bởi Breiman là bộ phân loại với cây B $\{T_1(X), ..., T_B(X)\}$. Trong đó $X = \{x_1, ..., x_p\}$ là vectơ p chiều của siRNA. Đầu ra là $B(\{\hat{y}_1 = T_1(X), ..., \hat{y}_B = = T_B(X)\}$ trong đó $\hat{y}_B$, b = 1,..., B, là giá trị dự đoán cho một chuỗi siRNA của cây, đầu ra của tất cả các cây được tổng hợp để đưa ra dự đoán cuối cùng, $trong$ đó $\hat{Y}$ là giá trị trung bình của các dự đoán cây riêng biệt.

Với những dữ liệu gồm một tập hợp của các chuỗi n siRNA cho huấn luyện, $D = \{(X_1, Y_1), ..., (X_n, Y_n)\}$, trong đó X_i, i = 1,. . ., N, là một vector của các tính năng và Y_i là thử nghiệm giá trị hiệu quả đánh giá, quy trình huấn luyện như sau.

- Từ các dữ liệu huấn luyện gồm chuỗi n siRNA, tạo bootstrap (tức là, lấy mẫu ngẫu nhiên, thay thế, n chuỗi siRNA).
- Đối với mỗi mẫu bootstrap, tại mỗi nút, chọn giá trị tốt nhất trong một tập hợp con lựa chọn ngẫu nhiên các chức năng và cây có kích thước tối đa (tức là cho đến khi không chia tách hơn nữa) và không tỉa lại.
- Lặp lại các bước trên cho đến khi (một số đủ lớn) cây B đó được tạo.

Việc thực hiện dự đoán của RFR được đánh giá bởi một phương thức kiểm tra chéo và cách sử dụng mẫu Out-Of-Bag (OOB), cụ thể trong quá trình huấn luyện, mỗi cây được xây dựng bằng cách sử dụng một mẫu bootstrap đặc biệt, Số lượng cây được thiết lập là 1000 và sai số bình phương trung bình (RMSE) để xác định giá trị m thử tốt nhất.

Lựa chọn đặc trưng

Trên cơ sở các nghiên cứu trước đây về quy tắc thiết kế siRNA lựa chọn 15 thuộc tính có liên quan chặt chẽ với hiệu quả siRNA, các tính năng này được thể hiện trong Bảng 3.4. Mỗi cặp bazơ lân cận trong chuỗi sense- antisense siRNA đã được tính toán theo phương pháp láng giềng gần nhất được mô tả bởi Xia *et al*. Tính năng được ước tính dựa trên mỗi giá trị khác biệt của RMSE đưa ra được các quy tắc sau.

- Nucleotit tại vị trí ưa thích: A1, U1, U2, U3, U5, A7, U7, A10, U13, A14, U14, C17, C18, C19; nucleotit tại vị trí khác: C1, G1, G2, A5, G6, C7, G7, A11, G13, C14, G14, A17, A19.

- Ổn định cho mỗi hai cặp bazơ lân cận của siRNA sense - antisense.
- dG (1, 18): Tiêu chuẩn chênh lệch năng lượng tự do giữa các vị trí 1 và 18

Đánh giá hiệu quả mô hình

Các tham số để đánh giá mô hình hồi quy được xác định là:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|efficacy_i - efficacy_i^*\right|\right)^2}$$

$$q = \sqrt{1 - \frac{\text{MSE}}{Var(efficacy_i{}^*)}}$$

$$R = \frac{\sum_{i=1}^{n}(efficacy - \overline{efficacy})(efficacy^* - \overline{efficacy^*})}{\sqrt{\sum_{i=1}^{n}(efficacy - \overline{efficacy})^2}\sqrt{\sum_{i=1}^{n}(efficacy^* - \overline{efficacy^*})^2}}$$

Với n là số các chuỗi siRNA trong tập dữ liệu trong đó $efficacy_i$ và $efficacy_i$* là giá trị thực và giá trị dự đoán được xác nhận bằng thực nghiệm, tương ứng

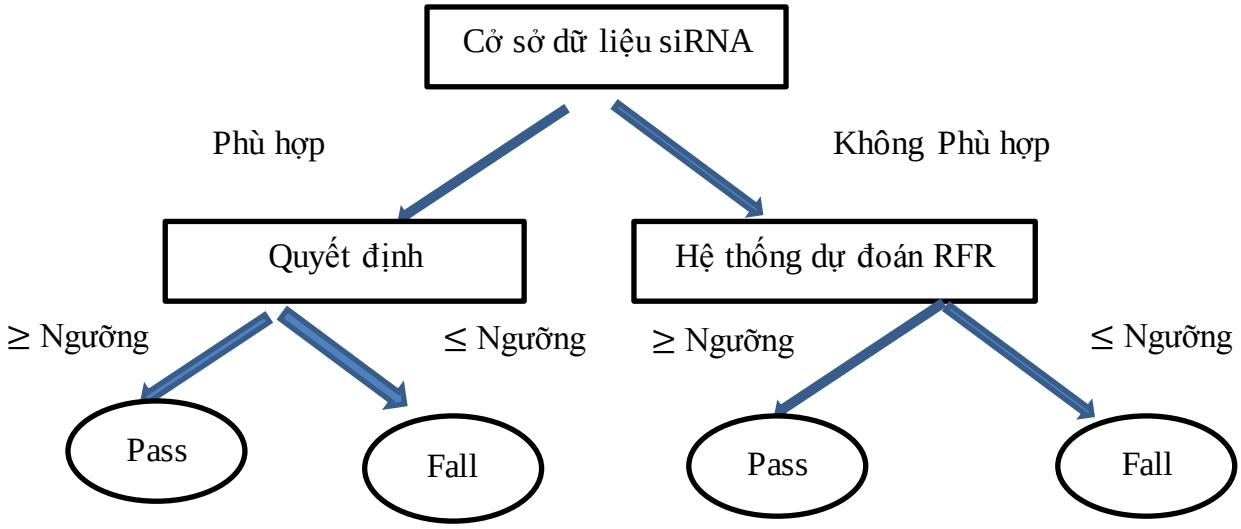
Bảng 3.4: Các tính năng được sử dụng trong các mô hình dự báo RFR

Đặc tính	Tầm quan trọng của tính năng	Độ xếp hạng quan trọng
Thành phần		
U (%)	3.87	5
G (%)	2.52	7
UU (%)	1.75	11
UC (%)	0.93	14
CA (%)	0.82	15
GC (%)	1.84	9
GG (%)	1.50	13
Nhiệt động học		
Position 1	11.90	1
Position 2	3.26	6
Position 6	1.79	10
Position 13	2.06	8
Position 14	1.54	12
Position 18	3.96	4
dG (1,18)[d]	8.13	2

Để so sánh hiệu suất của mô hình RFR với các thuật toán khác trong việc xác định siRNA có hiệu quả cao, thiết lập một ngưỡng giá trị chức năng 75% (siRNA với khả năng gen im lặng lớn hơn 75% được xác định là chuỗi hiệu quả ngược lại không hiệu quả)

Thực nghiệm

Sử dụng máy chủ tên RFR CDB-siRNA, đã được phát triển, RFR CDB-siRNA bao gồm hai thành phần độc lập: Một cơ sở dữ liệu siRNA-trung tâm và một hệ thống dự báo RFR, quá trình làm việc Hình 3.8:



Hình 3.8: Quy trình dự báo của RFR

Xây dựng các mô hình hồi quy rừng ngẫu nhiên

Đối với huấn luyện trong một thuật toán máy học, một tập dữ liệu đồng nhất và đầy đủ nhất là rất quan trọng. Tuy nhiên chức năng của một siRNA thay đổi khác nhau như điều kiện sinh học và thực nghiệm khác nhau, nhưng bộ dữ liệu Huesken được công bố có thể kết hợp trực tiếp các trình tự siRNA từ các nguồn khác nhau, sử dụng dữ liệu Huesken như các tập dữ liệu đào tạo trong mô hình dự báo RFR.

Nghiên cứu trước đây chỉ ra rằng nhiều thuộc tính, chẳng hạn như các tính năng trình tự, năng lượng của RNA ảnh hưởng đến chức năng của các siRNA và Shabalina *et al*. cải tiến những tính năng gồm 15 thuộc tính như thể hiện trong bảng 3.4 và những tính năng này đã được lựa chọn trong phương pháp RFR. Nó chỉ ra rằng sự nhất trí ở vị trí phụ thuộc năng lượng của hai cặp bazơ lân cận trong sense- antisense siRNA ở vị trí 1 và sự khác biệt năng lượng tự do giữa các vị trí 1 và 18 là các tính năng có liên quan chặt chẽ với chức năng siRNA. Để đánh giá hiệu quả của mô hình sử dụng phương thức kiểm tra chéo là3-fold cross-validation với tham số RFR (ntree = 1000 và mtry = 10) đã thu được một RMSE và R là 8,924 và 0,851, và q đã lên đến 0,851, cho thấy mô hình hồi quy tốt với các dữ liệu thực nghiệm, để cải thiện và tối ưu hóa các mô hình RFR thực hiện một loại trừ từng bước số liệu bất thường (outlier). Nếu ít nhất một chuỗi trong tập dữ liệu tạo một giá trị ước tính ≥ 15 đơn vị trong mô hình sau đó tiếp tục với giá trị còn lại, và một mô hình thay thế được xây dựng sử dụng các trình tự còn lại. Thủ tục này được lặp đi lặp lại cho đến khi tất cả các chuỗi trong tập dữ liệu có giá trị còn lại < 15 đơn vị. Sau khi loại trừ số liệu bất thường, mô hình RFR xác định và loại bỏ 277 giá trị ngoại lai R và q được tăng lên 0,917 và 0,918, tương ứng.

So sánh với các mô hình hồi quy máy vector hỗ trợ

Các nghiên cứu thuật toán hồi quy học máy trước đó chỉ ra rằng mô hình SVR tốt hơn nhiều các mô hình hồi quy khác trên nhiều khía cạnh. Do đó sử dụng mô hình SVR là một thuật toán thay thế, để so sánh với các mô hình dự báo của RFR.

Sử dụng phương thức kiểm tra chéo 3-fold cross-validation của SVR trên các số liệu có cùng một kết quả là RMSE bằng 9,414. Kết quả đạt được cao hơn so với các mô hình RFR (trước khi trừ outlier), cho thấy một sự sai lệch nhiều về kết quả dự đoán từ các dữ liệu quan sát, bên cạnh đó, cả R và các giá trị q của mô hình SVR thấp hơn so với các mô hình RFR, để cải thiện và tối ưu hóa các mô hình SVR. Một thủ tục loại trừ số liệu bất thường theo từng bước và làm tương tự với phần còn lại thì thấy rằng phương pháp RFR chính xác hơn so với phương pháp SVR [20], như thể hiện trong bảng 3.5, mô hình SVR bị giảm so với mô hình RFR sau khi số liệu bất thường đã được gỡ bỏ.

Bảng 3.5: Thực hiện mô hình RFR và mô hình SVM trong siRNA

Kết quả	RFR		SVM	
	With outlier	Without outlier	With outlier	Without outlier
RMSE	0.8924	6.904	9.414	7.403
R	0.851	0.917	0.832	0.907
q	0.851	0.918	0.832	0.907

So sánh với các phương pháp học máy khác

Huesken *et al*. Lựa chọn ngẫu nhiên 2.431 siRNA trong 34 loài mRNA, cùng hệ thống dự báo BIOPREDsi dựa trên mô hình mạng nơron. Vert *et al*, đề xuất mô hình hồi quy Lasso với cùng một bộ dữ liệu trên, ba bộ tính năng (thưa: sự hiện diện hay vắng mặt của mỗi nucleotide ở mỗi vị trí, phổ, số lần xuất hiện của mỗi mô típ nucleotide dài 1-3).

Để so sánh các mô hình RFR với hai tập dữ liệu độc lập (Reynolds gồm 240 siRNA và các số liệu Vickers gồm 76 siRNA) đã được sử dụng là các bộ dữ liệu thử nghiệm. Tất cả các phương pháp học máy được tập huấn luyện cùng tập dữ liệu (dataset Huesken). Kết quả, như thể hiện trong Bảng 3.6, chỉ ra rằng các hệ số tương quan Pearson giữa dự đoán và dự đoán RFR là cao hơn so với hai phương pháp khác trên cả hai bộ dữ liệu.

Bảng 3.6: Hiệu suất trên bảng dữ liệu độc lập

Dữ liệu	Mô hình LASSO			BIOPREDsi	RFR
	Sparse	Spectral	Composite		
Reynolds	0.54	0.49	0.55	0.55	0.58
Vicker	0.58	0.54	0.49	0.57	0.59

Như vậy là với kết quả đạt được ở trên ta thấy việc sử dụng mô hình RFR đạt được kết quả tốt hơn các phương pháp khác

Trên đây vừa trình bày hai phương pháp học máy SVM và RF trong quá trình thực nghiệm hai phương pháp thấy rằng kết quả của SVM và RF đều có độ chính xác cao so với các phương pháp học máy khác và tùy từng phương pháp biểu diễn có các kết quả khác nhau. Trong phần thực nghiệm tôi sử dụng SVR, RF để thực nghiệm so sánh kết quả đạt được với các phương pháp đã được đưa ra trong các báo cáo gần đây.

3.4. Sử dụng phương pháp học biểu diễn để nâng cao độ chính xác của các mô hình dự đoán

Như trên đã đề cập, việc tạo ra siRNA hiệu quả cao là một trong hai vấn đề quan trọng trong quá trình nghiên cứu siRNA để tạo các loại thuốc mới để điều trị nhiều loại bệnh. Trong cách tiếp cận sinh học, các nhà sinh học dựa trên thí nghiệm của mình để phát hiện quy tắc thiết kế siRNA đã tìm ra các đặc điểm quan trọng ảnh hưởng đến hiệu quả của việc ức chế siRNA. Trong cách tiếp cận tính toán, kỹ thuật học máy đã áp dụng không chỉ tìm thấy quy tắc thiết kế siRNA mà còn xây dựng mô hình dự báo để dự đoán hiệu quả ức chế của siRNA tuy nhiên, chúng có một số hạn chế như sau:

(i) Quy tắc thiết kế là không đủ để chọn siRNAs hiệu quả

(ii) Các mô hình phát triển có hiệu suất thấp và đạt được kết quả không tốt khi thử nghiệm trên bộ dữ liệu độc lập.

Ngoài ra, các quy tắc thiết kế có thể tạo ra hàng ngàn siRNA dự tuyển và nhiều siRNAs tạo ra là không hoạt động hoặc không hiệu quả. Mặt khác, quần thể của siRNAs là khoảng 4^{19}, vì vậy nó rất khó khăn để tạo ra một mô hình có thể dự đoán hiệu quả ức chế cho tất cả các siRNA. Vì vậy, để tạo ra siRNA hiệu quả cao, là tìm ra các quy tắc thiết kế và xây dựng mô hình dự báo tốt hơn, mục đích của cách này là sử dụng các quy tắc thiết kế để thu hẹp không gian tìm kiếm, dựa trên không gian tìm kiếm này, các mô hình dự báo có thể dự đoán siRNA với hiệu quả cao. Việc thực hiện các phương pháp học máy phụ thuộc rất nhiều vào sự lựa chọn của biểu diễn dữ liệu, dựa trên ý tưởng này và để khắc phục nhược điểm trên đã có rất nhiều nhóm nghiên cứu đưa ra các phương pháp biểu diễn khác nhau như nhị phân, quang phổ, tứ diện, chuỗi đại diện, Bui Thang sử dụng phương pháp biểu diễn bằng cách chuyển đổi siRNA thành ma trận (Bảng 3.7) và dùng một số quy tắc thiết kế siRNA đã được công bố để học ma trận chuyển đổi [3],

Bảng 3.7: Chuyển đổi chuỗi siRNA thành ma trận

Chuỗi	Mã hóa ma trận X	Biến đổi thành ma trận T	Vectơ dữ liệu chuyển đổi
AUGCU	1 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 1	0.5 0.7 0.32 0.2 0.5 0.3 0.1 0.6 0.6 0.3 0.1 0.1 0.08 0.1 0.1 0.1 0.1 0 0.1 0.1	(0.5, 0.1, 0.08, 0.6, 0.1)

Trong đó quy tắc thiết kế siRNA được tích hợp để làm giàu đại diện siRNA và phân cụm thứ tự nhãn siRNA cũng được bảo tồn. Trong phần thực nghiệm của luận

văn này dùng phương pháp biểu diễn là chuyển đổi dữ liệu sang ma trận và thực nghiệm với một số phương pháp học máy, phương pháp được trình bày sơ lược như sau.

Đưa vào: Hai bộ siRNA được gán nhãn có độ dài n, và một tập hợp các quy tắc thiết kế K siRNA.

Tìm: Ma trận chuyển đổi có thể chuyển đổi chuỗi siRNA để làm giàu ma trận.

Phương pháp này bao gồm ba bước. Bước đầu tiên là mã hóa siRNA, thứ hai là thiết kế và học ma trận chuyển đổi, cuối cùng là sử dụng ma trận chuyển đổi để làm giàu siRNA, các bước của phương pháp này được tóm tắt như sau:

- Để mã hóa mỗi chuỗi siRNA giống như mã hóa một ma trận X đại diện cho các nucleotide A, C, G và U ở vị trí n trong chuỗi, như vậy chuỗi siRNA được biểu diễn như ma trận mã hóa n x 4.
- Để học biến đổi ma trận T_k, k = 1, ..., K, mỗi đặc trưng cho khả năng ức chế của nucleotit A, C, G và U ở vị trí n trong chuỗi siRNA. Về nguyên tắc thiết kế thứ k, mỗi T_k là học được từ các bộ siRNA dán nhãn và các quy tắc thiết kế thứ k. thành lập từng quy tắc thiết kế với siRNA để giải quyết một vấn đề tối ưu hóa mới được hình thành.
- Chuyển đổi siRNA (ma trận mã hóa) để làm giàu ma trận bằng cách chuyển đổi K ma trận.

Bước 1 của phương pháp này có thể dễ dàng thực hiện, mỗi dãy siRNA với n nucleotit được mã hóa như một ma trận mã hóa nhị phân kích thước n x 4. Trong thực tế, bốn nucleotit A, C, G, hoặc U được mã hóa bằng cách mã hóa các vector (1,0,0,0), (0,1,0,0), (0,0,1,0) và (0,0,0,1), tương ứng, nếu một nucleotit từ A, C, G và U xuất hiện ở vị trí thứ n trong một chuỗi siRNA, j = 1, ..., n, vector mã hóa của nó sẽ được sử dụng để mã hóa hàng thứ j của ma trận mã hóa.

Bước 2 là học ma trận chuyển đổi T_k liên quan đến các quy tắc thiết kế thứ k, k =1, ..., K. T_k có kích thước của 4 x n, nơi các hàng tương ứng với nucleotit A, C, G và U và các cột tương ứng với vị trí n trên chuỗi. T_k được học từng cái một từ tập các siRNA và các quy tắc thiết kế thứ k, do đó để đơn giản sử dụng T thay vì T_k. Mỗi tế bào T [i, j], i = 1, ..., 4, j = 1, ..., n, đại diện cho khả năng ức chế nucleotit i ở vị trí j liên quan đến các quy tắc thiết kế thứ k. Mỗi tế bào T [i, j] để được học phải đáp ứng một số hạn chế, thứ nhất là những hạn chế cơ bản của T.

$$T[i, j] \geq 0 \; i = = 1, ..., 4; j = 1, 2, \ldots, n$$

$$\sum_{i=1}^{4} T[i, j] = 1, \; j = 1, \ldots, n$$

Thứ hai là những hạn chế liên quan đến thiết kế quy định, mỗi quy tắc thiết kế mô tả sự xuất hiện hay vắng mặt của các nucleotit ở các vị trí khác nhau của chuỗi siRNA hiệu quả. Do đó, nếu một quy tắc thiết kế cho thấy sự xuất hiện (vắng mặt) của một số

nucleotit vào vị trí thứ j, sau đó giá trị tương ứng của nó trong ma trận T sẽ lớn hơn (nhỏ hơn) so với các giá trị khác tại cột j.

Ví dụ, các quy tắc thiết kế trong bảng bên phải trong bảng 3.8 minh họa rằng ở vị trí 19, nucleotit A / U là hiệu quả và nucleotit C là không hiệu quả, nó có nghĩa là khả năng ức nucleotit A / U có kích thước lớn hơn so với các nucleotide G / C và khả năng ức chế của nucleotit C nhỏ hơn của các nucleotit khác, như vậy, giá trị T [1, 19], T [2, 19], T [3, 19] và T [4, 19] cho thấy khả năng ức chế của nucleotit A, C, G và U ở vị trí 19, tương ứng, vì vậy, năm hạn chế tại cột 19 của T được hình thành.

Bảng 3.8: Ví dụ về quy tắc thiết kế

Vị Trí	Khả năng ức chế	Nucleoti	Tạo T	Hạn chế trên T
19	Hiệu quả	A,U	T[1,19]	T[3,19]- T[1,19] < 0
			T[4,19]	T[3,19]- T[4,19] < 0
	Không hiệu quả	C	T[2,19]	T[2,19]- T[1,19] < 0
				T[2,19]- T[3,19] < 0
				T[2,19]- T[4,19] < 0

Các quy tắc thiết kế được xem xét bởi.

$$\{\mathrm{gr(T)} < 0\}_{r=1}^{R}$$

Thứ ba là những hạn chế liên quan đến bảo tồn các lớp siRNA sau khi được chuyển đổi bằng cách sử dụng các ma trận chuyển đổi T_k, nó có nghĩa là siRNA thuộc cùng lớp nên được thêm với nhau hơn siRNA thuộc các lớp khác.

Cho vectơ x_l có cỡ là 1 x n ký hiệu vectơ chuyển đổi của chuỗi siRNA thứ l sử dụng các ma trận chuyển đổi T, các phần tử thứ j của x_l là các phần tử của T ở cột j trong chuỗi siRNA, để tính x_l các cột bên trong tính như sau

$$X_l = T * X_l = (\langle Xl[1,.], T[.,1]\rangle, \langle Xl[2,.], T[.,2]\rangle, \ldots, \langle Xl[n,.], T[.,n]\rangle)$$

nơi X_l [j,.] và T [., j] là vector hàng thứ j và cột thứ j của ma trận X_l và T. Tương ứng, và $\langle x, y\rangle$ biểu thị kết quả bên trong của vectơ x và y.

Các giá trị trong bảng 3.7 cho thấy một ví dụ về mã hóa ma trận X, chuyển đổi ma trận T và chuyển đổi vector x của chuỗi AUGCU nhất định, các hàng của X đại diện cho các vector mã hóa của các nucleotit trong chuỗi, với ma trận biến đổi T kích thước 4 x 5, trình tự AUGCU được đại diện bởi các vector

$$x = (T[1, 1], T[4, 1], T[3, 3], T[2, 4], T[4, 5]) = (0.5, 0.1, 0.08, 0.6, 0.1)$$

Do đó, dữ liệu chuyển đổi có thể được tính bằng x = T * X.

Việc xây dựng các mô hình dự báo khả năng ức chế của siRNA đã có rất nhiều mô hình được tiến hành thực nghiệm và có nhiều kết quả được đưa ra. Tuy nhiên tất cả

các mô hình đều có một số hạn chế đó là hiệu năng nói chung là thấp R từ 0.62 đến 0.68, giảm dần khi sử dụng trên bộ dữ liệu độc lập. Các bộ siRNA để test có thể không đại diện cho toàn bộ siRNA, biểu diễn siRNA có thể không phù hợp.

Để giải quyết vấn đề này các nhà nghiên cứu phải tiếp tục tìm ra các phương pháp biểu diễn thích hợp, làm giàu siRNA đại diện bằng cách kết hợp những kiến thức từ những quy tắc thiết kế siRNA hiệu quả và xây dựng một mô hình dự báo tốt hơn để đánh giá chính xác khả năng ức chế của siRNA hiệu quả.

3.5. Kết luận

Như vậy để xây dựng các mô hình dự báo khả năng ức chế của siRNA đã có rất nhiều các phương pháp học máy liên tục được các nhóm nghiên cứu thử nghiệm để giải quyết bài toán nhằm xây dựng một số mô hình dự đoán khả thi. Nhưng hầu hết kết quả dự đoán đều không cao nên việc tìm kiếm các giải pháp để tạo ra các siRNA có khả năng ức chế hiệu quả cao vẫn là một thách thức lớn. Trong luận văn này ở trong chương thực nghiệm và đánh giá tôi áp dụng một số mô hình học máy với phương pháp học biểu diễn với dữ liệu là các siRNA chuyển sang dạng ma trận và so sánh với kết quả với các phương pháp học máy khác để có cái nhìn tổng quan về bài toán.

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Chương này sẽ trình bày quá trình thực nghiệm sử dụng một số mô hình học máy để dự đoán khả năng ức chế gen. Bằng cách sử dụng dữ liệu của phương pháp biểu diễn dữ liệu đã nêu với mục đích đưa ra so sánh kết quả giữa các mô hình thực nghiệm với kết quả nghiên cứu đã được báo cáo để lựa chọn ra mô hình tốt phù hợp cho bài toán, có thể làm minh chứng khi đưa ra áp dụng thực tế.

4.1. Dữ liệu thực nghiệm và cài đặt

Dữ liệu

Trong phần thực nghiệm này sử dụng các phương pháp học máy để dự đoán khả năng ức chế của siRNA với các điều kiện tương tự của các thực nghiệm khác. Trong đó sử dụng các kết quả báo cáo đã được đưa ra trong những năm gần đây để so sánh với kết quả đã đạt được.

Sử dụng bộ dữ liệu được biểu diễn chuyển đổi bằng cách sử dụng phương pháp biểu diễn chuyển đổi dữ liệu sang ma trận và dùng một số các quy tắc đã được báo cáo để học ma trận [3]. Đã trình bày trong phần 3.3 bao gồm.

Bảy quy tắc thiết kế siRNA của Reynolds, Uitei, Amarzguioui, Jalag, Hsieh, Takasaki, and Huesken và bốn bộ dữ liệu sau:

• Bộ dữ liệu Huesken với 2431 siRNA của 34 gen gồm người và động vật gặm nhấm [21].

• Bộ dữ liệu Reynolds với 244 siRNA [48].

• Bộ dữ liệu Vicker với 76 siRNA của hai gen [55].

• Bộ dữ liệu Harborth với 44 siRNA của một gen [44].

Mô hình đề xuất thực nghiệm

Quá trình thực nghiệm trong luận văn đề xuất dùng một số mô hình học máy đó là hồi qui véc tơ hỗ trợ (Support Vector Regression –SVR). Rừng ngẫu nhiên (Random Forest-RF). Hồi quy tuyến tính (Linear Regression) và sử dụng phương pháp học biểu diễn của siRNA đó là chuyển dữ liệu sang ma trận và sử dụng các quy tắc thiết kế được báo cáo để làm giàu ma trận.

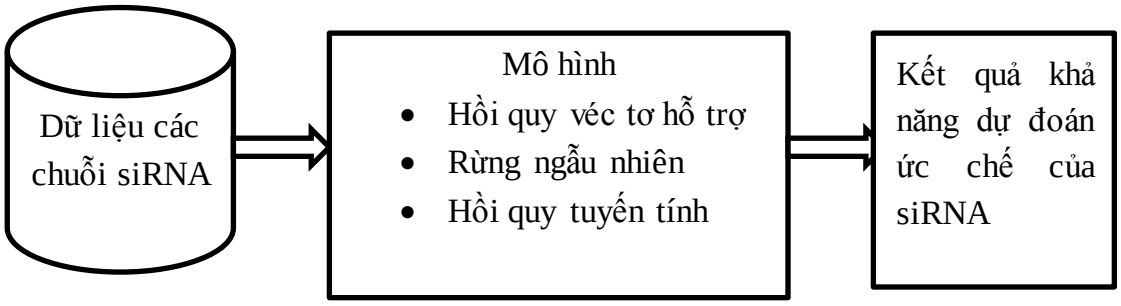
Môi trường thực nghiệm

Phần cứng máy tính Acer 4732z core i3 2.13 GHz, RAM 4GB

Phần mềm: Sử dụng bộ công cụ Weka phiên bản 3.7 được phát triển bởi nhóm nghiên cứu trường đại học Waikato Hamilton, New Zealan

Quy trình thực nghiệm

Áp dụng các kiến thức nghiên cứu ở trên để tiến hành thực nghiệm dự đoán khả năng ức chế gen của siRNA dựa trên các phương pháp hồi qui véc tơ hỗ trợ, rừng ngẫu nhiên, hồi quy tuyến tính được thực hiện với quy trình như hình 4.1.



Hình 4.1: Quy trình giải quyết bài toán

Sử dụng dữ liệu trong đó huấn luyện trên tập dataset Huesken và thử nghiệm trên ba tập dữ liệu độc lập của Reynolds, Vicker, Harborth với tập Huesken sử dụng phương thức kiểm tra chéo (k – fold cross validation) trên mỗi tập dữ liệu. Sử dụng phương thức k – fold, chia tập dữ liệu thành 10 - fold, sau đó tiến hành huấn luyện với 10 lần lặp, mỗi lần sử dụng 9 – fold dữ liệu làm tập huấn luyện mô hình, fold còn lại làm tập test.

Tính RMSE (sai số bình phương), R (hệ số tương quan), MAE sai số tuyệt đối trung bình. Phương pháp dự báo tốt là phương pháp nhận được sai số R lớn còn sai số RMSE và MAE nhỏ (càng gần về không càng tốt)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2}$$

$$\text{R} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(f_i-\bar{f}_i)}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(f_i-\bar{f})^2}}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|(y_i - f_i)|$$

Với y_i, f_i,$\bar{y}$ chỉ giá trị thực, giá trị dự đoán và giá trị trung bình của mẫu thứ i tương ứng

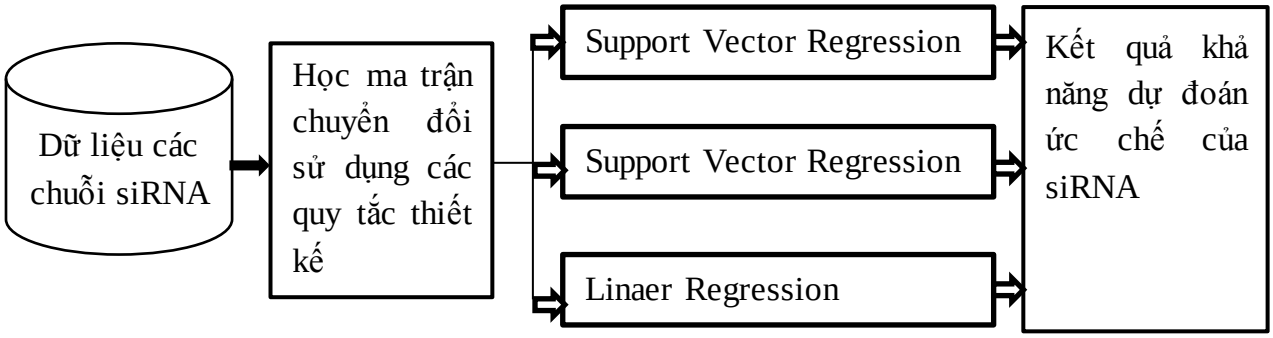
Trong nghiên cứu này sẽ thực hiện thực nghiệm sử dụng các phương pháp học máy sau: Random forest, SVR và Linear Regression. Quá trình huấn luyện và dự đoán các phương pháp được thực hiện như sau:

- Bước 1: Chọn dữ liệu
- Bước 2: Chọn phương pháp
- Bước 3: Chỉnh các tham số của phương pháp
- Bước 4: Chạy huấn luyện
- Bước 5: Lấy ra được mô hình

Kết quả cho ra mô hình huấn luyện với các tham số tối ưu, việc tìm ra được các tham số tối ưu là rất quan trọng, nó ảnh hưởng lớn đến độ chính xác của mô hình để đưa ra kết quả chính xác cao hay không. Quá trình huấn luyện mô hình được thực hiện với bốn tập dữ liệu.

4.2. Thực nghiệm các phương pháp học máy dự đoán khả năng ức chế của siRNA

Quá trình thực nghiệm được mô tả trong hình 4.2:

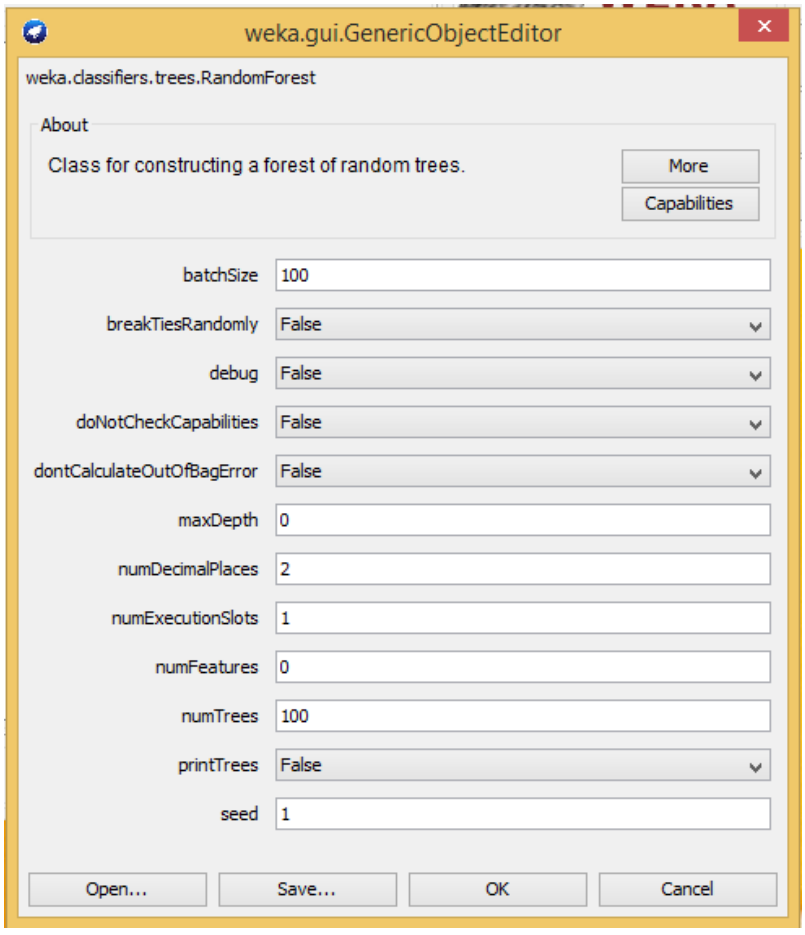


Hình 4.2: Quá trình thực nghiệm các phương pháp đề xuất

Sử dụng dữ liệu trong đó huấn luyện trên tập dataset Huesken và thử nghiệm trên ba tập dữ liệu độc lập của Reynolds, Vicker, Harborth với tập Huesken sử dụng phương thức kiểm tra chéo 10 – fold cross validation.

Phương pháp Random forest

Các tham số chính Random forest khi huấn luyện bằng Weka hình 4.3,



Hình 4.3: Các tham số huấn luyện mô hình Random forest

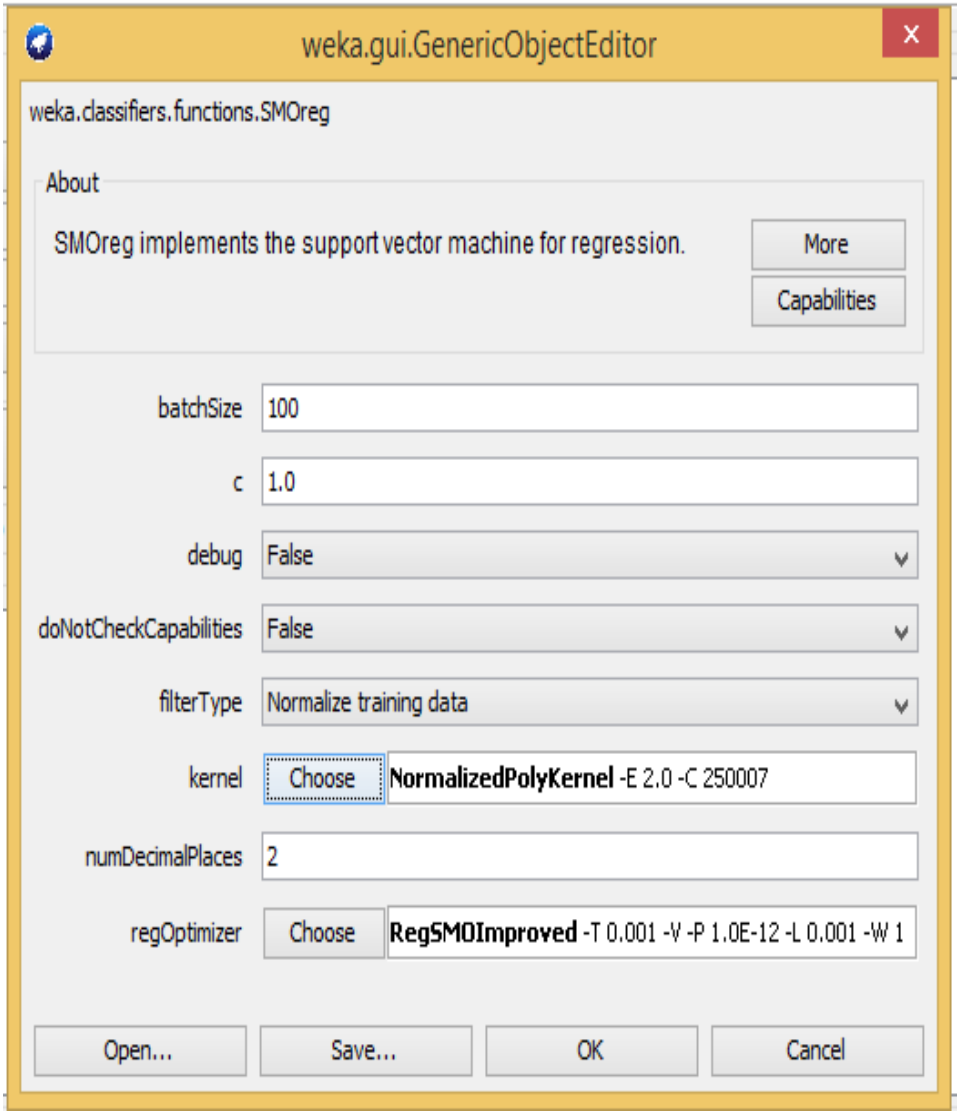
Sử dụng 4 tập dữ liệu Huesken Reynolds, Vicker, Harborth tiến hành thực nghiệm với phương pháp RF kết quả thu được trong Bảng 4.1.

Bảng 4.1: Kết quả huấn luyện của mô hình Random forest

Tập dữ liệu	RMSE	MAE	R	siRNA
Harborth	20.3246	18.7826	0.4502	44
Reynolds	28.1583	20.2544	0.5004	244
Huesken	15.4773	12.4966	0.60	2431
Vicker_	41.6252	36.5266	0.5258	76

Phương pháp SVR

Các tham số chính SVR khi huấn luyện bằng Weka Hình 4.4



Hình 4.4: Các tham số huấn luyện mô hình SVR

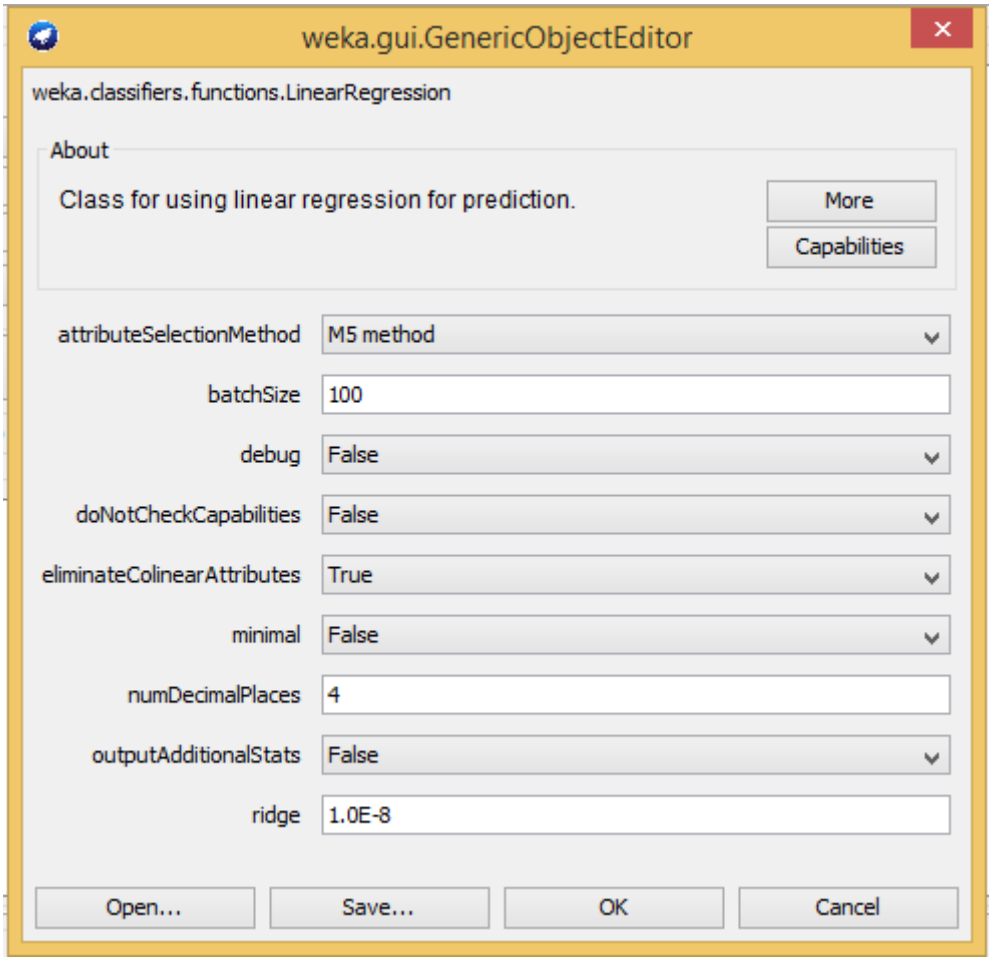
Sử dụng 4 tập dữ liệu Huesken Reynolds, Vicker, Harborth tiến hành thực nghiệm với phương pháp SVR kết quả thu được trong Bảng 4.2.

Bảng 4.2: Kết quả huấn luyện của mô hình SVR

Tập dữ liệu	RMSE	MAE	R	siRNA
Harborth	37.8097	32.5779	0.5412	44
Reynolds	37.195	33.252	0.54	244
Huesken	15.0423	12.0436	0.63	2431
Vicker_	19.2521	15.7425	0.5644	76

Phương pháp Linear Regression

Các tham số chính Linear Regression khi huấn luyện bằng Weka Hình 4.5



Hình 4.5: Các tham số huấn luyện mô hình Linear Regression

Sử dụng 4 tập dữ liệu Huesken Reynolds, Vicker, Harborth tiến hành thực nghiệm với phương pháp Linear Regression kết quả thu được trong Bảng 4.3

Bảng 4.3: Kết quả huấn luyện của mô hình Linear Regression

Tập dữ liệu	RMSE	MAE	R	siRNA
Harborth	24.2303	22.6723	0.4708	44
Huesken	15.0861	12.0568	0.62	2431
Reynolds	26.3556	19.3782	0.55	244
Vicker	39.7976	32.7644	0.5508	76

4.3. Đánh giá thực nghiệm

Các đánh giá thực nghiệm được thực hiện như sau:

So sánh các mô hình đề xuất với phương pháp SVM nhiều nhân được đưa ra bởi Qui *et al*. [37]. Kết quả là R = 0, 62 thu được dựa trên áp dụng phương thức k – fold đó là chia tập dữ liệu thành 10 - fold, sau đó tiến hành huấn luyện với 10 lần lặp, mỗi lần sử dụng 9 – fold dữ liệu làm tập huấn luyện mô hình, fold còn lại làm tập test trên tập dữ liệu Huesken.

Thực hiện thực nghiệm các phương gồm SVR, Linear Regression, Random forest với tập dữ liệu Huesken phương thức kiểm tra chéo là 10-fold cross validation ta được kết quả (Bảng 4.4).

Bảng 4.4: Các giá trị của R áp dụng trên bộ dữ liệu Huesken

Phương pháp	Dữ liệu	R	Đánh giá
Qui's method	Huesken	0.62	10 lần với phương thức10-folds cross validation
SVR	Huesken	0.63	
Linear Regression		0.62	
Random Forest		0.60	

Với kết quả thu được ta thấy R của SVR, Linear Regression, Random Forest lần lượt có giá trị 0.63, 0.62, 0.60. So sánh với R của Qui sử dụng phuong pháp SVM nhiều nhân thì với dữ liệu sử dụng phương pháp biểu diễn chuyển đổi sang ma trận và làm giầu siRNA với các quy tắc thiết kế đã được công bố thì khi thực nghiệm với cùng một phương pháp là SVR thì ta thấy kết quả thực nghiệm cao hơn. Phương pháp Linear Regression thì có kết quả cùng với Qui là 0.62. Tuy RF cho độ chính xác phân lớp cao khi so sánh với các thuật toán học có giám sát hiện nay bao gồm Boosting, Baging, các láng giềng gần nhất (Nearest neighbors), SVM, Neural Network, C45,... Tuy nhiên, tiếp cận cài đặt RF ban đầu chỉ cho kết quả tốt trên các dữ liệu có số chiều vừa phải và giảm đáng kể hiệu năng khi xử lý bài toán có số chiều rất cao, nhiều nhiễu, dung lượng mẫu ít và bài toán phân tích dữ siRNA là một trường hợp cụ thể. Nguyên nhân chính là trong quá trình xây dựng cây quyết định, tại mỗi nút, RF dùng phương pháp chọn ngẫu nhiên một tập con thuộc tính từ tập thuộc tính ban đầu để tìm thuộc tính phân hoạch tốt nhất phân tách nút và luật quyết định ở nút lá của các cây trong rừng ngẫu nhiên dựa vào luật bình chọn số đông. Điều này dẫn đến độ chính xác của giải thuật rừng ngẫu nhiên bị giảm khi phân lớp dữ liệu, nên khi xử lý với các dữ liệu nhiều nhiễu như siRNA, RF có thể lựa chọn ngẫu nhiên nhiều siRNA nhiễu vào không gian con thuộc tính dùng cho việc tách nút khi dựng cây, nên khả năng dự đoán của RF

giảm sút. Nên trong cả bốn phương pháp thì kết quả dự đoán trên mô hình RF có độ chính xác thấp hơn cả, như vậy với phương pháp biểu diễn khác thì ta thấy kết quả SVR với dữ liệu được sử dụng bởi [3] đã đạt cao hơn so với mô hình nhiều nhân của Qui và các con số này cho thấy hầu hết các siRNA trong bộ dữ liệu Huesken có khả năng ức chế cao.

So sánh ba phương pháp thực nghiệm trên với 18 phương pháp bao gồm BIOPREDsi, DSIR, Thermocomposition21,SVM … Khi huấn luyện trên tập dataset Huesken và thử nghiệm trên ba tập dữ liệu độc lập của Reynolds, Vicker và Harborth trong các báo cáo gần đây [41] (Bảng 4.5)

Bảng 4.5: So sánh phương pháp thực nghiệm với 18 phương pháp

Phương pháp	Năm	$R^{Reynolds}$	R^{Vicker}	$R^{Harborth}$
GPboot	2004	0.55	0.35	0.43
Uitei	2004	0.47	0.58	0.31
Amarzguioui	2004	0.45	0.47	0.34
Hsieh	2004	0.03	0.15	0.17
Takasaki	2010	0.03	0.25	0.01
Reynolds 1	2004	0.35	0.47	0.23
Reynolds 2	2004	0.37	0.44	0.23
Schawarz	2003	0.29	0.35	0.01
Khvorova	2003	0.15	0.19	0.11
Stockholm 1	2004	0.05	0.18	0.28
Stockholm 2	2004	0.00	0.15	0.41
Tree	2004	0.11	0.43	0.06
Luo	2004	0.33	0.27	0.40
i-score	2007	0.54	0.58	0.43
BIOPREDsi	2006	0.53	0.57	0.51
DSIR	2006	0.54	0.49	0.51
Katoh	2007	0.40	0.43	0.44
SVM	2013	0.54	0.52	0.54
SVR		**0.54**	**0.5644**	**0.5412**
Linear Regression		**0.55**	**0.5508**	**0.4708**
RF		**0.5004**	**0.5258**	**0.4502**

So sánh kết quả thực nghiệm khi thực nghiệm trên ba bộ dữ liệu độc lập với kết quả của 18 phương pháp đã được báo cáo (Bảng 4.5). Với kết quả đạt được ta thấy các phương pháp thử nghiệm với bộ dữ liệu được biểu diễn bằng cách chuyển sang ma trận với các phương pháp SVR, Linear Regression, Random Forest có kết quả cao hơn hầu hết các phương pháp khác để dự đoán khả năng ức chế gen của siRNA. Chẳng hạn như phương pháp SVR ta thầy kết quả ổn định trên cả ba bộ dữ liệu độc lập có thể so sánh với SVM được đưa ra 2013 sử dụng cấu trúc 3 chiều ta thấy kết quả cao hơn. Lý do là phương pháp biểu diễn dữ liệu đã kết hợp các quy tắc được tìm thấy từ các tập dữ liệu khác nhau trong các thực nghiệm. Tuy nhiên nó cũng có nhược điểm là ma trận biến đổi học dựa trên tính năng vị trí của các quy tắc thiết kế sẵn có. Do đó nó cũng thiếu một số đặc điểm ảnh hưởng hiệu quả ức chế của siRNA như là tính chất nhiệt, tương quan cặp bazơ, chiều dài … Điều đó chứng tỏ kết quả phụ thuộc nhiều vào việc lựa chọn các phương pháp dự đoán và các phương pháp biểu diễn, với mỗi phương pháp biểu diễn dữ liệu của cùng tập dữ liệu chúng ta thấy có các kết quả khác nhau.

4.4. Kết luận

Trong chương này đã tiến hành thực nghiệm các phương pháp SVR, RF, Linear Regression để đánh giá sự phù hợp của mô hình đối với bài toán dự đoán khả năng ức chế gen của siRNA. Đồng thời so sánh với phương pháp đã được báo cáo thì thấy rằng các phương pháp đề xuất thực nghiệm đã đạt được kết quả cao.

Tuy kết quả trong quá trình thực nghiệm không phải là tối ưu nhưng nó cũng có thể đóng góp thêm một cách tìm hiểu việc chọn lựa mô hình dự đoán cũng như phương pháp học biểu diễn cho các nhà khoa học khi nghiên cứu khi nghiên cứu về việc xây dựng mô hình dự đoán khả năng ức chế của siRNA.

CHƯƠNG 5. KẾT LUẬN

5.1. Những vấn đề được giải quyết trong luận văn.

Trong quá trình tìm hiểu để đưa ra cách giải quyết cho bài toán siRNA luận văn đã trình bày nghiên cứu một vấn đề sinh học đó là làm thế nào để tổng hợp siRNA hiệu quả để thiết kế các loại thuốc mới để điều trị nhiều loại bệnh như HIV, ung thư, virus cúm A, virus viêm gan B. Để giải quyết vấn đề này, các nhà sinh học đã được thực hiện và phân tích các quá trình thực nghiệm và họ phát hiện ra những đặc điểm quan trọng ảnh hưởng hiệu quả ức chế của siRNA, kết quả là, họ báo cáo quy tắc thiết kế cho siRNA hiệu quả. Trong nghiên cứu sinh học tính toán, các nhóm nghiên cứu đã được áp dụng kỹ thuật máy học thay thế để phát hiện quy tắc thiết kế siRNA và dự đoán hiệu quả ức của siRNA. Luận văn tổng hợp nghiên cứu về bài toán siRNA để giúp chúng ta có cách nhìn tổng quan và áp dụng một cách phù hợp vào giải quyết bài toán nhằm xây dựng một số mô hình dự đoán khả thi, để đoán nhận khả năng ức chế của siRNA hỗ trợ cho việc điều chế thuốc.

Liên quan đến việc phát hiện các quy tắc thiết kế cho vấn đề siRNA hiệu quả, có rất nhiều các phương pháp trong cả hai hướng tiếp cận sinh học và sinh học tính toán được đưa ra. Một số đặc điểm mới của siRNA ảnh hưởng đến hiệu quả của ức chế siRNA đã được phát hiện, những phương pháp này đã được trình bày trong chương 2.

Việc giải quyết bài toán siRNA không chỉ nhằm tìm kiếm các quy tắc thiết kế tạo ra các siRNA hiệu quả các nhà khoa học còn tập trung vào việc xây dựng các mô hình học máy để dự đoán khả năng ức chế của siRNA. Đã có rất nhiều các phương pháp học máy được đưa ra, với nhiều kết quả thử nghiệm khác nhau đã được trình bày trong chương 3. Trong chương này cũng trình bày một phương pháp biểu diễn để áp dụng cho phần thực nghiệm.

Kết quả chạy thực nghiệm đã chứng minh được rằng lựa chọn các phương pháp thực nghiệm và phương pháp biểu diễn dữ liệu đề xuất đã có hiệu quả hơn một số phương pháp khác. Tuy rằng luận văn mới dừng lại ở bước thực hiện thực nghiệm trên các phương pháp đưa ra, nhưng kết quả mang lại cũng có những ý nghĩa nhất định giúp các nhóm nghiên cứu khác có nhìn tổng quan về việc sử dụng các mô hình học máy để đoán nhận khả năng ức chế siRNA.

Trong quá trình thực hiện luận văn này tôi đã cố gắng tập trung nghiên cứu bài toán dự đoán khả năng ức chế của siRNA và tham khảo nhiều tài liệu liên quan. Luận văn chủ yếu tập trung vào việc tổng hợp nghiên cứu của các nhà khoa học để giải quyết bài toán. Tuy đưa chưa đạt được tối ưu, nhưng luận văn của tôi cũng có một số thực nghiệm đạt kết quả tốt để các nhà nghiên cứu tham khảo thêm trong quá trình thực nghiệm về siRNA. Tuy nhiên do thời gian và trình độ có hạn nên không tránh

khỏi những hạn chế và thiếu sót nhất định, do vậy tôi thật sự mong muốn nhận được những góp ý cả về kiến thức chuyên môn lẫn cách trình bày.

5.2. Công việc nghiên cứu trong tương lai

Như trình bày ở trên nghiên cứu của luận văn tập trung vào một vấn đề thú vị và đầy thử thách của sinh học, các kết quả đã đạt được trong thử nghiệm của luận văn cũng như các nghiên cứu trước đó vẫn còn một số hạn chế. Trong vấn đề phát hiện quy tắc thiết kế siRNA, các quy tắc thiết kế siRNA hợp lý và đặc điểm mới đã được tìm thấy bằng cách áp dụng một phương pháp mô tả, tuy nhiên, những quy tắc thiết kế hợp lý và đặc điểm mới cần phải được đánh giá bởi quá trình thực nghiệm cũng như các chuyên gia trong nghiên cứu sinh học. Vì vậy, nghiên cứu chung giữa các nhà sinh học và tin sinh học sẽ là một sự hợp tác mạnh mẽ để giải quyết các vấn đề sinh học và mang lại kết quả nghiên cứu để ứng dụng thực tế. Trong dự đoán ức chế của siRNA, tôi đề nghị quá trình học và dự báo các phương pháp đại diện siRNA bằng cách kết hợp những kiến thức nền tảng của quy tắc thiết kế siRNA, tại thời điểm này mô hình dự báo không đạt được hiệu suất cao, dựa trên những hạn chế và nghiên cứu hiện tại trong cả hai cách tiếp cận sinh học và sinh học tính toán, mục đích của chúng tôi là nghiên cứu những vấn đề sau đây trong tương lai.

- Tìm siRNA hiệu quả cao dựa trên các quy tắc thiết kế siRNA và mô hình dự báo: Trong các báo cáo trước đó, các mô hình hồi quy có thể dự đoán hiệu quả ức chế của siRNA và phát hiện quy tắc thiết kế có thể tạo ra siRNA hiệu quả, nhưng quy tắc thiết kế siRNA không thể tạo ra hiệu quả với số lượng 4^{19} siRNAs. Do đó, chúng ta nên có một chiến lược để tìm siRNA có hiệu quả cao, có thể được tổng hợp để làm thuốc. Trong luận văn này, tất cả các đặc điểm quan trọng được phát hiện bởi các nghiên cứu trước đây cần được xem xét để thực hiện quy tắc thiết kế siRNA và các mô hình tiên đoán hiệu suất chính xác hơn. Để có kết quả tốt cần sự hợp tác giữa các nhóm và các nhà sinh học và kết quả của các công trình nghiên cứu nên được đánh giá bởi các quá trình thực nghiệm.
- Thiết kế siRNA hiệu quả nên nghiên cứu với từng gen gây bệnh cụ thể. Có mô tả cụ thể đặc điểm của như là nhiễm trùng, biến dị di truyền, cấu trúc protein … Do đó, siRNA dựa cho việc điều trị và ngăn ngừa từng bệnh là vấn đề rất quan trọng.
- Xây dựng mô hình dự báo để giảm thiểu ức chế sai mục tiêu, ảnh hưởng ức chế sai mục tiêu của siRNA được định nghĩa là hiện tượng mà siRNA mục tiêu mRNA ngoài ý muốn và chúng ức chế những mRNA. Nó dẫn đến các tác dụng phụ của thuốc dựa siRNA, vấn đề này hiện đang xem xét một trong những vấn đề thách thức trong thiết kế của siRNA hiệu quả. Do đó, tôi dự định xây dựng mô hình có thể dự đoán khả năng ức chế sai mục tiêu của siRNA. Mô hình giúp đỡ để tìm ra siRNA không chỉ có hiệu quả ức chế cao nhưng cũng có giảm khả năng ức chế sai mục tiêu.

TÀI LIỆU THAM KHẢO

1. Alistair M. C., Erik L. L. (2008), "Sonnhammer: siRNA specificity searching incorporating mismatch tolerance data". *Bioinformatics*, 24(10), pp.1316–1317
2. Amarzguioui, M., Prydz, H. (2004), "An algorithm for selection of functional siRNA sequences", *Biochem Biophys Res Commun*, 316(4), pp.1050–8.
3. Bui Thang. (2014), "A Novel Framework to Improve siRNA Efficacy Prediction", *PAKDD (2)*, pp.400-412.
4. Bitko, V., Barik, S. (2001), "Phenotypic silencing of cytoplasmic genes us ing sequence-specific double-stranded short interfering RNA and its application in the reverse genetics of wild type negative-strand RNA viruses", *BMC Microbiol*, (1), pp.34.
5. Boden, D., Pusch, O., Lee, F., Tucker, L., Ramratnam, B. (2003), "Human Immunodeficiency Virus Type 1 Escape from RNA Interference", *J. Virol.*, 77, pp.11531– 11535.
6. Birmingham A., Anderson E.M., Reynolds A. (2006). *et al.*, ""3' UTR seed matches, but not overall identity, are associated with RNAi off–targets", *Nat. Methods*, (3), pp.199–204
7. Chalk, A.M., Wahlestedt, C., Sonnhammer, E.L.L. (2004), "Improved and automated prediction of effective siRNA", *Biochem Biophys Res Commun*, (319), pp.264–274.
8. Chuang, C. F., Meyerowitz, E. M. (2000): "Specific and heritable genetic interference by double-stranded RNA in Arabidopsis thaliana", *Proc. Natl. Acad. Sci*, (97), pp.4985–4990
9. Clemens, M.J, Elia, A. (1997), "The mRNA of the translationally controlled tumor protein P23/TCTP is a highly structured RNA, which activates the dsRNA-dependent protein, kinase", *PKR. J. Interferon Cytokine Res.*, 17, pp.503–524.
10. Corey, D. R (2007), "RNAi learns from antisense", *Nat. Chem. Bio.*, (3), pp.8–11).
11. Christoph, T., Grunweller, A., Mika, J., Schafer, M. K., Wade,E. J., Weihe, E., Erdmann, V. A., Frank, R., Gillen, C., Kurreck, J (2006), "Silencing of vanilloid receptor TRPV1 by RNAi reduces neuropathic and visceral pain in vivo", *Biochem, Biophys. Res. Commun.*, (350), pp.238–243
12. Crooke, S. T. (2004), "Progress in Antisense Technology", *Annu. Rev. Med.*, (55), pp.61–95.

13. Chang, P.C., Pan, W.J., Chen, C.W., Chen, Y.T., Chu DEsi, Y.W. (2012), "A design engine of siRNA that integrates SVMs prediction and feature filters", *Biocatalysis and Agricultural Biotechnology*, (1), pp.129–134.

14. Du Q, Thonberg H, Wang J, Wahlestedt C, Liang Z. (2005), "A systematic analysis of the silencing effects of an active siRNA at all single–nucleotide mismatched target sites*", Nucleic Acids Res*, 33(5), pp.1671-7.

15. Escobar, M. A., Civerolo, E. L., "Summerfelt, K. R., Dandekar, A. M. (2005), RNAi-mediated oncogene silencing confers resistance to crown gall tumorigenesis*", Proc. Natl. Acad. Sci*, (98), pp.13437–13442

16. Elbashir, S.M., Lendeckel, W., Tuschl, T. (2001), "RNA interference is mediated by 21– and 22–nucleotide RNAs*", Genes Dev*., (15), pp.188–200

17. Francesco, D. S., Hanspeter, S., Alejandro, L., Cornia, T., Estelle, B. (2001), *Frederick, M*.:"Sense and antisense mediated gene silencing in tobacco is inhibited by the same viral suppressors and is associated with accumulation of small RNAs", *Proc. Natl. Acad. Sci.*, 96, pp.6506–6510.

18. Gitlin, L., Stone, J. K., Andino, R. (2005), "Poliovirus Escape from RNA Interference: Short Interfering RNA-Target Recognition and Implications for Therapeutic Approaches", *J. Virol.*, 79, pp.1027–1035.

19. Grunweller, A.,Wyszko, E., Bieber, B., Jahnel, R., Erdmann, V.A. , Kurreck, J(2000), "Comparison of different antisense strategies in mammalian cells using locked nucleic acids, 2'–O–methyl RNA, phosphorothioates and small interfering RNA", *Nucleic Acids Res*., 31, pp.3185–3193.

20. Hsieh, A.C., Bo, R., Manola, J., Vazquez, F., Bare, O., Khvorova, A.,Scaringe, S., Sellers, W.R. (2004), "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(3), pp.893–901

21. Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Mellon, B., Engel, S., Rosenberg, A., Cohen, D., Labow, M., Reinhardt,M., Natt, F., Hall, J. (2005), "Design of a Genome–Wide siRNA Library Using an Artificial Neural Network". *Nature,, Biotechnology*, 23(8), pp. 955–1001.

22. Holen, T., Amarzguioui, M., Wiiger, M.T., Babaie, E., Prydz, H. (2002), "Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor", *Nucleic Acids Res.*, (30), pp.1757–1766.

23. Ingelbrecht, I., Van Houdt, H., Van Montagu, M., Depicker, A (1994), "Post-transcriptional silencing of reporter transgenes in tobacco correlates with DNA methylation". *Proc. Natl. Acad. Sci* , (91), pp.10502–10506.

24. Ichihara, M., Murakumo, Y., Masuda, A., Matsuura, T., Asai, N., Jijiwa, M., Ishida, M., Shinmi, J., Yatsuya, H., Qiao, S. *et al*. (2007), "Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities", *Nucleic Acids Res*., 35, e123.

25. Jackson A.L., Bartz S.R., Schelter J., *et al*. (2003), "Expression pro- filing reveals off-target gene regulation by RNAi", *Nature Biotechnol*, 21, pp.635–637.

26. Jackson A.L., Burchard J., Leake D., *et al*. (2006), *"*Position–specific chemical modification of siRNAs reduces"off–target" transcript silencing*", RNA*, 12, pp. 1197- 1205.

27. Kooter, J. M., Matzke, M. A., Meyer, P. (1999), "Listening to silent gene: transgene silencing, gene regulation and pathogen control", *Trends Plant Sci.*, (4), pp.340– 347.

28. Kurreck, J. (2009), "RNA interference: from basic research to therapeutic applications*", Angew, Chem.*, (121), pp.1404– 1426.

29. Karol K., Gabor C. (2010), "Kernel Based Off–Target Analysis of Rnai Experiments Global, Journal of Medical Research*"*, Vol. 1, Issue 1, Ver 1.0,

30. Komarov, P. G., Komarova, E. A., Kondratov, R. V., Christov– Tselkov, K., Coon, J. S., Chernov, M. V., Gudkov, A. V. (1999), "A Chemical Inhibitor of p53 That Protects Mice from the Side Effects of Cancer Therapy", *Science*, 285, pp.1733–1737

31. Klingelhoefer, J.W., Moutsianas, L., and Holmes, C.C. (2009), *"*Approximate Bayesian feature selection on a large meta-dataset offers novel insights on factors that effect siRNA potency*", Bioinformatics*, (25), pp.1594–1601

32. Liangjiang Wang, Caiyan Huang, and Jack Y Yang (2009), *"*Predicting siRNA potency with random forests".

33. Ladunga, I. (2007), "More complete gene silencing by fewer siRNAs: Transparent optimized design and biophysical signature*", Nucleic Acids Res*, (35), pp.433 – 440.

34. Liu J., Carmell, M.A.,Rivas F.V., Marsden, C.G.,Thomson, J.Ms., Song, J.J., Hammond, S.M., Joshua–Tor, L., Hannon, G.J 2004, "Argonaute2 is the catalytic engine of mammalian RNAi", *Science*, (305), pp.1437–1441.

35. Lim L., Lau N., Garrett–Engele P. *et al*. (2005), "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs", *Nature*, (433), pp.769–773

36. Peng Jiang, Haonan Wu, Yao Da, Fei Sang, Jiawei Wei, Xiao Sun, Zuhong Lu(2007), "RFRCDB-siRNAImproved design of siRNAs by random forest regression model coupled with database searching".

37. Qiu, S. and Lane, T. (2009), "A Framework for Multiple Kernel Support Vector Regression and Its Applications to siRNA Efficacy Prediction", *IEEE/ACM Trans*. Comput., *Biology Bioinform*. (6), pp.190–199

38. Santel, A., Aleku, M., Keil, O., Endruschat, J., Esche, V., Durieux, B., Fechtner, M., Rohl, T., Fisch, G., Dames, S., Arnold, W., Giese, K., Klippel, A., Kaufmann, J, "RNA interference in the mouse vascular endothelium by systemic administration of siRNA-lipoplexes for cancer therapy".

38. *Sen, G. L., Blau, H. M.* (2006), "Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies", *Nat. Cell Biol., 7, 633–636 (2005). Ther*, 13, pp.1360–1370

40. Smith, F. J., Hickerson, R. P., Sayers, J. M., Reeves, R. E., Contag, C. H., Leake, D., Kaspar, R. L., McLean, W. H. (2008), "Development of Therapeutic siRNAs for Pachyonychia", *Congenita*. *J. Invest. Dermatol*, 128, pp. 0–58

41. Sciabola, S., Cao, Q., Orozco, M., Faustino, I. and Stanton, R.V (2013), "Improved nucleic acid descriptors for siRNA efficacy prediction", *Nucl.Acids Res.*, (41), pp.1383–1394.

42. Smith, F. J., Hickerson, R. P., Sayers, J. M., Reeves, R. E., Contag, C. H., Leake, D., Kaspar, R. L., McLean, W. H. (2008), "Development of Therapeutic siRNAs for Pachyonychia Congenita", J. Invest. Dermatol, (128), pp.50–58

43. Schubert *et al*., 2004 Schubert, S., Kurreck, J (2004), "Human Gene Therapy", *Curr. Drug Target*, (5), pp.667–681

44. Takasaki, S. (2010), "Efficient prediction methods for selecting effective siRNA equences", *Comput Biol Med.*, (40), pp. 149–158

45. Takasaki, S(2013), "Methods for Selecting Effective siRNA Target Sequences Using a Variety of Statistical and Analytical Techniques", *Methods Mol Biol.*, (942), pp. 17–55.

46. Teramoto, R., Aoki, M., Kimura, T., Kanaoka, M. (2005), "Prediction of siRNA functionality using generalized string kernel and support vector machine", *FEBS Lett*., 579, pp.2878–2882.

47. Ren, Y., Gong, W., Xu, Q., Zheng, X., Lin, D. and *et al*. (2006), "siRecords: an extensive database of mammalian siRNAs with efficacy ratings", *Bioinformatics*, (22), pp.1027–1028.

48. Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S.,Khvorova, A.: "Rational siRNA design for RNA interference", *Nat Biotechnol*., 22(3), 326–330.

49. Zimmermann, T. S., Lee, A. C., *et al.* (2006), "RNAi-mediated gene silencing in non-human primates", *Nature*, (441), pp.111–114.

50. Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., Saigo, K (2004), "Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RN"A interference", *Nucleic Acids Res*., (32), pp.936–948.

51. Weitzer S1, Martinez J. (2007), "The human RNA kinase hClp1 is active on 3' transfer RNA exons and short interfering RNAs", *Nature*, (447), pp. 222 – 226

52. Watanabe, A., Arai, M., Yamazaki, M., Koitabashi, N., Wuytack, F., Kurabayashi, M. (2004): "Phospholamban ablation by RNA interference increases Ca2+uptake into rat cardiac myocyte sarcoplasmic reticulum", *J. Mol. Cell. Cardiol*., (37), pp. 691–698.

53. Warnecke, C., Zaborowska, Z., Kurreck, J., Erdmann, V. A., Frei, U., Wiesener, M., Eckardt, K. U(2004), "Differentiating the functional role of hypoxia inducible factor (HIF)-1alpha and HIF-2alpha (EPAS-1) by the use of RNA interfer ence: erythropoietin is a HIF-2alpha target gene in Hep3B and Kelly cells", *FASEB J*., (18), pp.1462–1464

54. Wu *et al*., 2003 Wu, H., Hait, W. N., Yang, J. M. (2003), "Small interfering RNA-induced sup pression of MDR1 (P-glycoprotein) restores sensitivity to multidrug–resistant cancer cells", *Cancer Res*., (63), pp. 1515–1519.

55. Vickers, T.A., Koo, S., Bennett, C.F., Crooke, S.T., Dean, N.M.and Baker, B.F. (2003), "Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents, A comparative analysis*", J. Biol. Chem*., (278), pp. 7108–7118