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

TÓM TẮT LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

HÀ NỘI - 2016

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. 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… 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. Nên 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 và ứng dụng vào việc dự đoán khả năng ức chế của siRNA. Luận văn tập trung vào việc tổng hợp lại các quy tắc thiết kế siRNA hiệu quả và phương pháp dự đoán khả năng ức chế của siRNA. Đồng thời cũng tiến hành áp dụng thực nghiệm bằng một số phương pháp học máy và so sánh kết quả đạt được với các phương pháp học máy đã được tổng hợp. 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.

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 tìm hiểu của các nhà nghiên cứu về cách tìm ra 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.

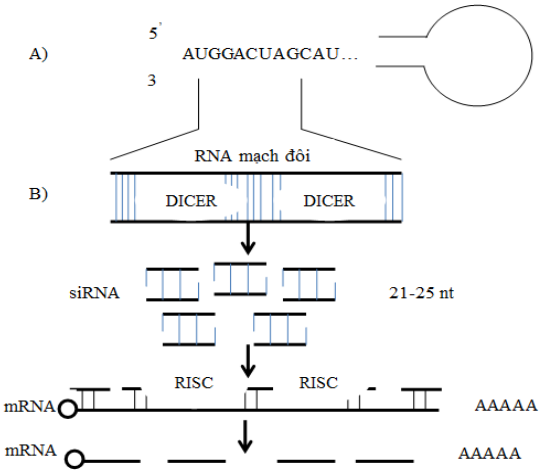
CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ ĐOẠN NGẮN RNA CÓ KHẢ NĂNG ỨC CHẾ (siRNA)

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.

1.1.1. Các cơ chế, thành phần chính của RNAi

RNAi chính là quá trình phân hủy mRNA (Hình 1.1), các dsRNA (Double stranded RNA) mạch kép hoặc sợi tóc bị cắt thành các đoạn ngắn RNA (siRNA) bởi các enzyme ribonuclease III Dicer. Các siRNA tháo xoắn thành hai sợi sense và antisense và họ protein được gọi là RNA- phức hệ gây sự im lặng (RISC) sẽ mang sợi antisense 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.



Hình 1.1: Sơ đồ hoạt động của RNAi và các siRNA

Có ba thành phần chính liên quan đến quá trình can thiệp RNA: siRNA, enzyme Dicer, và phức hệ (RISC).

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, điều hòa biểu hiện gen, điều khiển sự phát triển của tổ chức và duy trì hình dạng nhiễm sắc thể và tăng cường phiên mã…

1.1.3. Thành phần của RNAi

siRNA là các RNA ngắn có kích thước khoảng 21 đến 25 nucleotit, được hình thành từ các RNA sợi đôi, tham gia vào quá trình tổng hợp protein,

miRNA (micro RNA) là những đoạn RNA ngắn khoảng từ 19 đến 24 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 được phát hiện khi thực hiện biến đổi gen trên cây dạ yến thảo với dự kiến là có màu tím hơn, năm 1990. 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 (Hình 1.2),

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

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 và gen ngoại sinh phiên mã ra.

Can thiệp trong các tế bào động vật có vú.

Tuschl và đồng nghiệp phát hiện 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).

Ức chế sự biểu hiện của các gen mục tiêu thường kéo dài 5-7 ngày, Một siRNA chống những thành phần protein có chức năng vận chuyển lipid trong hệ thống tuần hoàn 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 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.

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 21 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. 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. Năm 1994, Cogoni và các cộng sự đã tiến hành một thí nghiệm nhằm phát triển màu cam của nấm. 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. 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. 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.

1.2.2. Chức năng của siRNA

- 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

1.2.3. Ứng dụng siRNA

Nghiên cứu các chức năng của gen

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 tiêu 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.

Ứng dụng điều trị

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.

Bệnh về mắt

Nghiên cứu can thiệp RNA lần đầu tiên được bắt đầu 2004 với một siRNA chống lại yếu tố tăng trưởng nội mạc (VEGF). Các siRNA được thử nghiệm dưới tên Bevasiranib. Phương pháp điều trị siRNA bắt đầu các nghiên cứu lâm sàng đầu tiên với biến đổi hóa học của một siRNA. 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

Kể từ khi các báo cáo đầu tiên về tác dụng kháng virus của siRNA chống virus hợp bào hô hấp (RSV), 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].

Ung thư

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 siRNA thâm nhập vào các tế bào khối u nội mô như lipoplexes và khối mạch

Các thử nghiệm lâm sàng khác

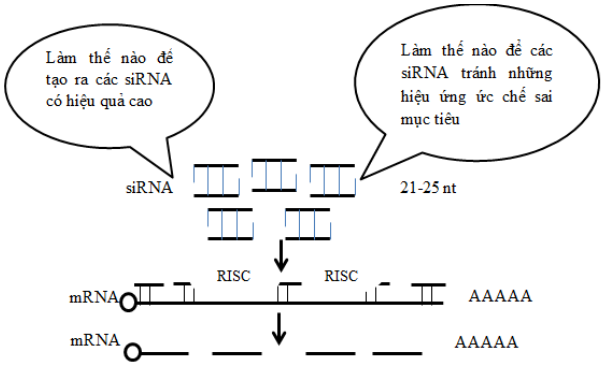
RNA đang được sử dụng như là một chiến lược điều trị chống suy thận cấp. Năm 2008, Transderm Inc đã bắt đầu một nghiên cứu lâm sàng để điều trị các nhiễm sắc thể di truyền bệnh dày móng bẩm sinh (Pachyonychia congenital). Các siRNA được tiêm vào và đặc biệt là ức chế sự biểu hiện của các keratin đột biến K6a [40].

1.2.4. Những thách thức trong nghiên cứu siRNA

siRNA là 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 từ dsRNA và bị cắt thành các siRNA nhờ Dicer, và siRNA mở xoắn tạo thành hai sợi sense và antisense, sợi antisense sẽ bám vào mRNA và nhờ phức hợp RISC nó cắt mRNA và phân hủy mRNA tương đồng với nó.

Để tạo ra các siRNA có khả năng ức chế cao ta phải giải quyết hai vấn đề quan trọng sau đây (Hình 1.3):

- Làm thế nào các siRNA tránh hiệu ứng ức chế sai mục tiêu
- Làm thế nào để tạo ra các siRNA có hiệu quả cao



Hình 1.3: Hai vấn đề quan trọng trong RNA

Tạo các siRNA hiệu quả cao

Vấn đề 1: Tìm quy tắc thiết kế siRNA hiệu quả.

Vấn đề 2: Xây dựng mô hình dự báo để dự đoán hiệu quả ức chế siRNA.

Vấn đề 3: Tạo siRNAs hiệu quả cao.

1.3. Kết luận

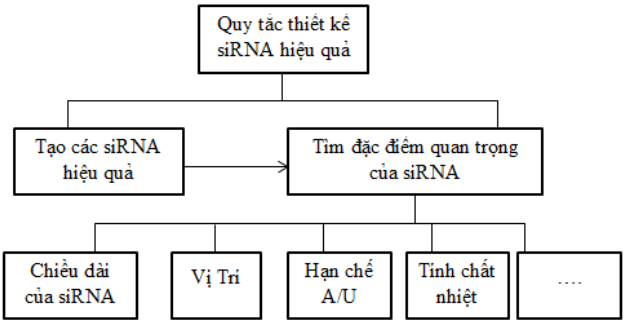
Các siRNA có thể được tổng hợp và đưa vào tế bào để làm ức chế gen đích dẫn việc tạo nhiều loại thuốc mới nhưng các siRNA làm ức chế các mRNA ở các cấp độ khác nhau nên việc tạo ra nhiều siRNA hiệu quả cao là một vấn đề rất quan trọng. Ngoài ra việc thực hiện các mô hình dự báo hiện tại rất ít trong 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

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

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. Có rất nhiều quy tắc được đưa ra (Hình 2.2). Sau đây là các quy tắc dự đoán quan trọng.

Quy tắc thiết kế Tuschl

Quy tắc thiết kế của Reynolds

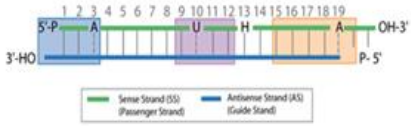
Quy tắc Amarzguioui

Quy tắc thiết kế Stockholm

Quy tắc thiết kế Ui-Tei

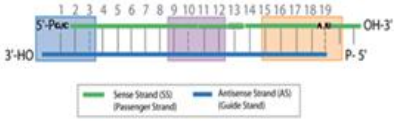
Quy tắc thiết kế Hseih

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

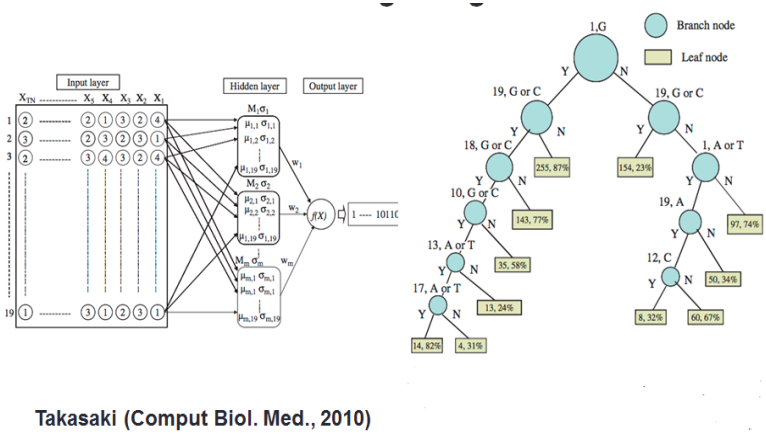
2.3. Các quy tắc thiết kế trong cách tiếp cận sinh học tính toán

Trong phương pháp sinh học, 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, 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 2.3)

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

Teramoto [46] và đồng nghiệp sử dụng Máy vector hỗ trợ (Support Vector Machine (SVM)) sử dụng để phân biệt các siRNA hiệu quả và không hiệu quả đã phát hiện được 20 đặc điểm của siRNA. Ladunga và đồng nghiệp [33] cũng sử dụng gói SVMLight với đa thức kernel để huấn luyện hơn 2200 siRNA, dựa trên các mạng nơron và cây quyết định (Hình 2.3) để lựa chọn siRNA hiệu quả từ nhiều mục tiêu có thể



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

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. Đó 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. 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

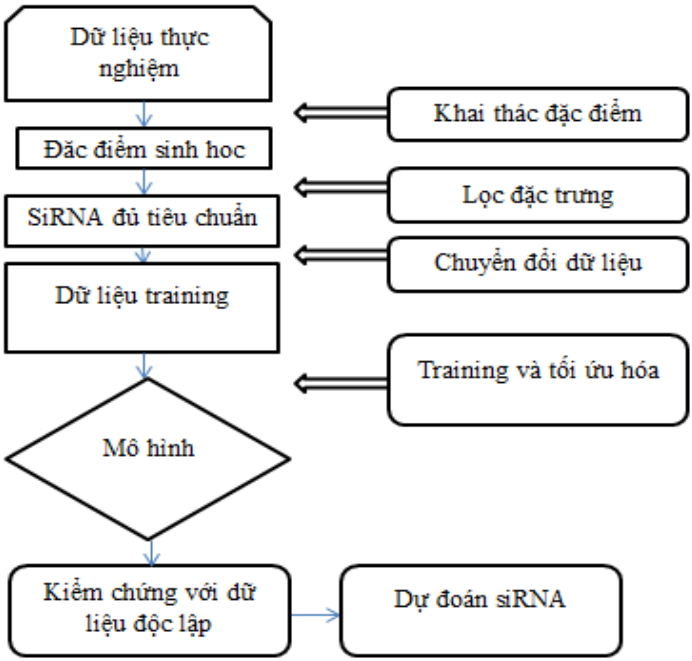
Như vậy là để tạo siRNA có hiệu quả cao trong cả hai cách tiếp cận sinh học và sinh học tính toán đã có nhiều quy tắc thiết kế siRNA đã được đưa ra. Tuy nhiên vẫn còn nhiều hạn chế, do đó để tạo ra quy tắc thiết kế siRNA hiệu quả cao ta vẫn phải tiếp tục nghiên cứu và thử nghiệm để tìm ra các quy tắc 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

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, áp dụng các phương pháp học máy để xây dựng mô hình dự báo đưa ra kết 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

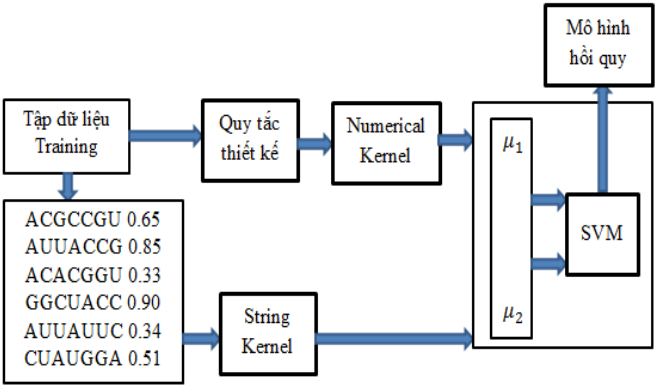
Giải pháp giải quyết việc xây dựng mô hình dự báo, nhiều kỹ thuật học máy đã được áp dụng để dự đoán hiệu quả ức chế siRNA như bảng 3.1

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 một siRNA được gia tăng là (0, 7). Huesken *et al* đã đề xuất các mô hình dự báo, trong đó biểu tượng nhận biết siRNA hiệu quả và không hoạt động đã được phát hiện bởi một mạng nơron nhân tạo (ANN) được huấn luyện trên 2.182 siRNA và thử nghiệm trên 249 siRNA với R= 0.66. Các chức năng của BIOPREDsi được phát triển bởi các số đặc trưng và sự nhạy cảm đối với ANN. 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 nhiều vector hỗ trợ hồi quy với đa nhân và cho dự đoán hiệu quả siRNA với R=0.62 với bộ dữ liệu Huesken gồm 2431siRNA. 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 (Hình 3.2).

Hầu hết các phương pháp đó bị một số nhược điểm. Mối tương quan giữa các giá trị dự báo và giá trị thực nghiệm của biến phụ thuộc khác nhau, từ 0, 60 tới 0, 68 đã được giảm đáng kể khi thử nghiệm trên bộ dữ liệu độc lập.Bởi vì thực tế rằng các số liệu Huesken vẫn còn quá nhỏ để có thể đại diện cho siRNA có khoảng 4^{19} thể siRNA



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

Ngoài ra việc thực hiện các phương pháp học máy phụ thuộc rất nhiều vào sự lựa chọn của biểu diễn dữ liệu (hoặc các tính năng) 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

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

Máy véc tơ hỗ trợ SVM

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 liên quan đến bài toán phân lớp hay dự báo như: Dự báo vị trí cắt-nối 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, và trong bài toán đoán nhận khả năng ức chế siRNA cũng đã được áp dụng (Hình 3.2)

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.

3.3. Phương pháp dự đoán rừng ngẫu nhiên (Random Forest)

Phân lớp dữ liệu có số chiều lớn có nhiễu như dữ liệu gen 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 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, ý tưởng chính của giải thuật random forest:

+ Từ tập học LS có N phần tử

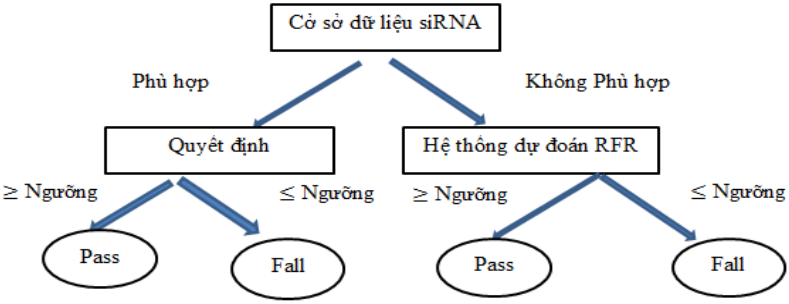
+ Xây dựng tập hợp T mô hình cơ sở độc lập nhau

+ 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

- Một bootstrap : Lấy mẫu N phần tử có hoàn lại từ tập LS
- Khi phân loại : Sử dụng majority vote(Đa số phiếu)

Ví dụ quy trình sử dụng Random forest



Hình 3.8 Quy trình dự báo của RFR

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

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. Trong phần thực nghiệm tôi sử dụng dụng phương pháp biểu diễn bằng cách chuyển đổi siRNA thành ma trận [3]. 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 và ý tưởng 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.

Ví dụ phương pháp biểu diễn chuyển chuỗi siRNA thành ma trận

Bảng 3.7: Chuyển đổi chuỗi siRNA thành ma trận

Chuỗi	Mã hóa ma trận X	Biến đổi thành ma trận T	Vecto dữ liệu chuyển đổi
AUGCU	1000	0.5 0.7 0.32 0.2 0.5	(0.5, 0.1, 0.08, 0.6, 0.1)
	0001	0.3 0.1 0.6 0.6 0.3	
	0010	0.1 0.1 0.08 0.1 0.1	
	0100	0.1 0.1 0 0.1 0.1	
	0001		

3.5. Kết luận

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 0.66 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.

CHƯƠNG 4 THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1. Dữ liệu thực nghiệm và cài đặt

Dữ liệu

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.

Dữ liệu được sử dụng 7 quy tắc thiết kế siRNA của Reynolds, Uitei, Amarzguioui, Jalag, Hsieh, Takasaki, and Huesken và bốn bộ dữ liệu gồm:

• 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 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) sử dụng phương pháp học biểu diễn của siRNA mới đó là chuyển dữ liệu sang ma trận.

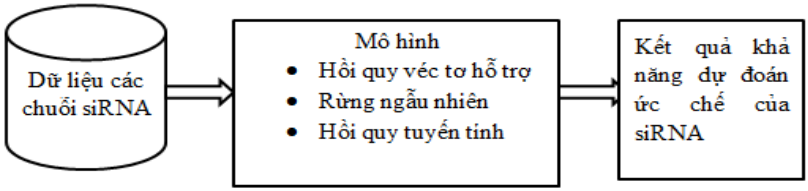
Môi trường thực nghiệm

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 dự đoán khả năng ức chế gen của siRNA theo các phương pháp hồi qui véc tơ hỗ trợ, rừng ngẫu nhiên, hồi quy tuyến tínhvới quy trình theo hình 4.1



Hình 4.1 Quy trình giải quyết bài toán

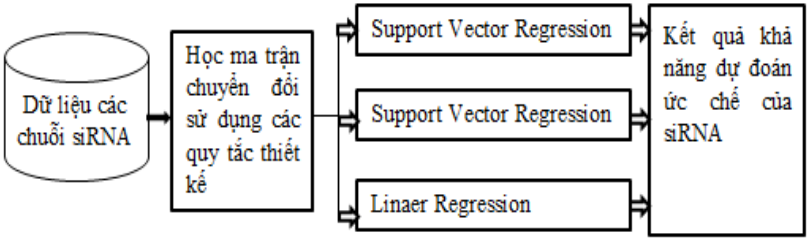
Sử dụng dữ liệu trong đó huấn luyện trên tập dataset Huesken và thử nghiệm trên ba tập dữ liệu độc lập của Reynolds, Vicker, Harborth với tập Huesken sử dụng phương thức kiểm tra chéo (k – fold cross validation)

Quá trình huấn luyện và dự đoán 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

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 hình 4.2;



Hình 4.2 Quá trình thực nghiệm các phương pháp đề xuất

Phương pháp Random forest

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

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

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

So sánh các mô hình đề xuất với phương pháp SVM nhiều nhân được đưa ra bởi Qui *et al*. [37] khi cùng thực nghiệm trên tập dữ liệu Huesken phương thức 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 phương pháp SVM nhiều nhân thì với dữ liệu được chuyển đổi sang ma trận khi thực nghiệm với cùng một phương pháp là SVR thì ta thấy kết quả thực nghiệm cao hơn.Với phương pháp Linear Regression thì có kết quả cùng với Qui là 0.62, 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 đã đạt cao hơn so với mô hình nhiều nhân của Qui.

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)

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). 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 rất nhiều một số phương pháp khác. 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. 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.

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**

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 và so sánh với phương pháp đã được báo cáo thì thấy rằng đã đạt được kết quả cao.

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 chưa đưa ra được phương pháp 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.

5.2. Công việc nghiên cứu trong tương lai

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 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.
- Thiết kế siRNA hiệu quả nên nghiên cứu với từng gen gây bệnh cụ thể.
- Xây dựng mô hình dự báo để giảm thiểu ức chế sai mục tiêu.