**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**Viện Công nghệ Thông tin và Truyền thông**

---------------------------------------------------------------



**Báo cáo bài tập lớn**

**Môn học: Tin sinh học**

**Đề tài: Dự đoán anti-CRISPR protein**

**Giảng viên hướng dẫn: TS. Nguyễn Hồng Quang**

**Nhóm sinh viên thực hiện:**

**Lê Hải Quân - 20173316**

**Nguyễn Hữu Dũng - 20173048**

**Tóm tắt đề tài**

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) là kĩ thuật chỉnh sửa DNA, được ứng dụng trong cây trồng, vật nuôi, nông nghiệp, y học,... Do đó, việc nghiên cứu ra protein chống CRISPR-Cas cũng được thúc đẩy mạnh để thuận tiện cho quá trình chỉnh sửa gen trở nên chính xác hơn. Một phương pháp dựa trên học máy để hỗ trợ xác định các protein anti-CRISPR chỉ sử dụng thông tin trình tự protein là SVM. Phương pháp học máy này được xây dựng bằng thuật SVM trên một tập dữ liệu đã biết về anti-CRISPRs. Việc bổ sung các phương pháp học máy vào bộ công cụ khám phá chống CRISPR cho phép các nhà nghiên cứu xếp hạng trực tiếp các gen được cho là chống CRISPR để tăng tốc độ thử nghiệm và xác nhận các gen chống CRISPR mới.

**Từ khóa:** CRISPR, DNA, anti-CRISPR, SVM, CRISPR-Cas, PaCRISPR

Mục lục

[1. Giới thiệu 2](#_Toc77544694)

[2. Mô tả bài toán 4](#_Toc77544695)

[2.1. Chi tiết bài toán 4](#_Toc77544696)

[2.2. Tập dữ liệu sử dụng 5](#_Toc77544697)

[2.3. Các độ đo đánh giá 5](#_Toc77544698)

[3. Các hướng nghiên cứu liên quan (Related Works) 6](#_Toc77544699)

[3.1. PaCRISPR - dự đoán và trực quan hóa các protein anti-CRISPR 6](#_Toc77544700)

[3.2. AcrHub - dự đoán và lập bản đồ các protein chống CRISPR 6](#_Toc77544701)

[3.3. AcRanker – xếp hạng các anti-CRISPR protein bằng học máy 8](#_Toc77544702)

[3.4. Phương pháp học máy mở rộng danh mục các họ anti-CRISPR protein 9](#_Toc77544703)

[3.5. AcrFinder: khai thác bộ gen các operon anti-CRISPR ở sinh vật nhân sơ và vi rút 9](#_Toc77544704)

[3.6. Phát hiện chính xác các Acr trong sinh vật nhân sơ chỉ sử dụng sáu tính năng 9](#_Toc77544705)

[4. Đề xuất mô hình 10](#_Toc77544706)

[5. Thực hiện hệ thống 12](#_Toc77544707)

[5.1. Xử lý dữ liệu đầu vào 12](#_Toc77544708)

[5.2. Thực hiện hệ thống 13](#_Toc77544709)

[5.3. Xử lý kết quả đầu ra 14](#_Toc77544710)

[6. Thử nghiệm và kết quả 14](#_Toc77544711)

[6.1. Môi trường thử nghiệm 14](#_Toc77544712)

[6.2. Các thử nghiệm 14](#_Toc77544713)

[7. Thảo luận 16](#_Toc77544714)

[8. Tổng kết và phương hướng phát triển (Conclusions and Perspectives) 16](#_Toc77544715)

[Tài liệu tham khảo 18](#_Toc77544716)

# Giới thiệu

CRISPR được viết tắt của những chữ cái đầu của cụm từ Clustered Regularly Interspaced Short Palindromic Repeats. CRISPR được mô tả là một họ các trình tự DNA ngắn xuôi ngược lặp lại (đọc theo hướng xuôi hay ngược thì đều có trình tự DNA giống nhau) và các trình tự này được ngăn cách với nhau bởi các vùng đệm (spacer). Kỹ thuật chỉnh sửa gen CRISPR-Cas9 dựa trên một cơ chế do vi khuẩn phát triển để tự bảo vệ chúng khỏi thực khuẩn thể (bacteriophage). CRISPR/Cas9 là hệ thống được mô tả lần đầu tiên và được dùng phổ biến. Hệ thống CRISPR/Cas9 bao gồm 2 thành phần: enzyme Cas9 nuclease và RNA dẫn đường. Nhờ đoạn trình tự bổ sung của RNA dẫn đường với trình tự đích mà phức hợp này có thể tìm thấy vị trí cần chỉnh sửa trên hệ gen. Để có thể hoạt động, hệ thống CRISPR còn yêu cầu một trình tự ngắn từ 2-5 nucleotides trên DNA đích được gọi là protospacer associated motif (PAM) ngay sau đoạn bổ sung của RNA dẫn đường (đối với CRISPR/Cas9 của vi khuẩn S.pyogenes thì trình tự này là 5’-NGG, trong đó N là bất kỳ nucleotide nào). Khi phức hợp Cas9 và guide RNA bám vào trình tự đích, enzyme Cas9 sẽ dùng 2 tiểu phần HNH và RuVC của mình để cắt đoạn DNA trên cả 2 mạch tại vị trí nucleotide thứ 3-4 phía trước PAM. CRISPR-Cas9 liên tục gây nhiễu hệ thống sửa sai DNA trong tế bào bằng cách: enzyme này cắt tại vị trí mục tiêu, sau khi tế bào sửa sai DNA, CRISPR-Cas9 lại cắt lần nữa. Chu trình gây lỗi này được lặp lại cho đến khi một đột biến được hình thành trên DNA ngăn chặn enzyme gắn vào, và khi đó phân tử CRISPR-Cas9 sẽ di chuyển để tìm một vị trí gắn khác

Nghiên cứu hiện nay chỉ ra rằng các thực khuẩn thể đối phó với vi khuẩn bằng cách tạo ra các protein ức chế được biết đến là các anti-CRISPR. Các protein này có thể được dùng để cải thiện kỹ thuật CRISPR-Cas9 như một công cụ của liệu pháp gen, nhằm làm giảm hiện tượng chỉnh sửa gene không phải gene mục tiêu, từ đó hạn chế những tác dụng phụ do hiện tượng này gây ra.

Các protein anti-CRISPR được khám phá gần đây có thể làm giảm đến bốn lần các tác động phụ của việc chỉnh gene không trúng đích và hoạt động như một công tắc khẩn để bất hoạt CRISPR-Cas9 sau khi nó hoàn thành việc chỉnh sửa gene. Nghiên cứu cho thấy một protein anti-CRISPR là AcrllA4 có khả năng làm giảm bốn lần tác động lên gene không phải gene mục tiêu trong quá trình phân tử CRISPR-Cas9 chỉnh sửa gene. Protein này sử dụng một RNA hướng dẫn để tìm, cắt và thay thế gene hemoglobin đột biến gây nên bệnh hồng cầu hình liềm. Nó thực hiện việc này mà không làm giảm hiệu quả chỉnh sửa gene mục tiêu.

Protein anti-CRISPR bắt chước DNA (*Hình 1*), làm cho CRISPR-­Cas9 nhầm lẫn, gắn chặt với nó không bị rời ra và bất hoạt khả năng cắt DNA của enzyme này. Protein anti CRISPR bắt vào một vị trí trên protein Cas9, là nơi rất quan trọng trong hoạt động chức năng của Cas9. Cas9 không thể vận hành việc cắt DNA khi anti-CRISPR gắn vào điểm này.

Map

Description automatically generated with medium confidence

Hình 1: Protein anti-CRISPR (màu đỏ bên phải) bắt chước DNA gắn vào vị trí tương tác của enzyme Cas9 với DNA (trái) trước khi enzyme tiến hành cắt DNA. Tuy nhiên, protein anti-CRISPR không rời khỏi vị trí tương tác khiến cho Cas9 không còn khả năng chỉnh sửa gen

Một số chiến lược đã được sử dụng để xác định anti-CRISPRs [1] [2], bao gồm các phân tích tin sinh học như ‘Guilt by association’ [3] hoặc phương pháp self-targeting [4], và các xét nghiệm hoặc sàng lọc chức năng. Mặc dù các cách tiếp cận này đã xác định thành công anti- CRISPR protein, nhưng các nghiên cứu này chỉ xác định được một số tập hợp con của anti- CRISPR protein và phụ thuộc nhiều vào kiến ​​thức trước đây về các đặc điểm chức năng của mối quan hệ thực khuẩn - vật chủ riêng lẻ. Ban đầu, các tìm kiếm dựa trên BLAST để lấy các điểm tương đồng của anti-CRISPRs từ các phage liên quan đã giúp xác định mức độ phổ biến của một số anti-CRISPRs [3]. Tuy nhiên, do một số anti- CRISPR protein được phát hiện gần đây không có trình tự tương tự như những chất hiện được biết đến, nên không thể chỉ dựa vào các phương pháp tương đồng để xác định các loại anti- CRISPR protein mới.

Để giải quyết vấn đề này, các phương pháp học máy đã được giới thiệu để đưa ra các dự đoán chống CRISPR chính xác hơn. Trong bài báo cáo, phương pháp học máy SVM (support vector machine) sẽ được trình bày để xác định hiệu quả và chính xác các anti-CRISPR dựa trên trình tự protein. SVM được sử dụng rộng rãi để giải các bài toán phân loại nhị phân trong lĩnh vực sinh học tính toán [5]. Đặc biệt SVM với nhân hàm cơ sở xuyên tâm (RBF) đã được sử dụng thành công để phân loại trình tự sinh học phi tuyến ( 29 , 30). Triển khai mô hình tổng thể như một máy chủ web thân thiện với người dùng, hệ thống cung cấp dự đoán anti-CRISPR protein dễ sử dụng. Bằng cách này, hệ thống được kỳ vọng sẽ tạo điều kiện thuận lợi cho việc phát hiện và mô tả đặc điểm của các anti- CRISPR protein mới từ các proteome của thể thực khuẩn và vi khuẩn, do đó thúc đẩy việc khám phá và hiểu các nguyên tắc đằng sau tương tác giữa thực khuẩn thể và sự đồng tiến hóa của thể thực khuẩn và vi khuẩn.

# Mô tả bài toán

## Chi tiết bài toán

Bài toán đặt ra trong đề tài là xây dựng một hệ thống để có thể dự đoán trước một protein có thể là anti-CRISPR protein bằng việc áp dụng các phương pháp học máy. Phương pháp học máy chỉ sử dụng đặc điểm trình tự protein để đưa ra dự đoán.

Đầu vào của hệ thống sẽ là chuỗi trình tự của protein cần dự đoán. Các trình tự của protein này sẽ được đưa vào hệ thống dưới dạng định dạng file *.fasta*. Nội dụng file như sau:

*>anti\_CRISPR0001*

*MKFIKYLSTAHLNYMNIAVYENGSKIKARVENVVNGKSVGARDFDSTEQLESWFYGLPGSGLGRIENAMNEISRRENP*

*>anti\_CRISPR0002*

*MMTISKTDIDCYLQTYVVIDPVSNGWQWGIDENGVGGALHHGRVEMVEGENGYFGLRGATHPTEKEAMAAALGYLWKCRQDLVAIARNDAIEAEKYRAKA*

*………………….*

*>anti\_CRISPR0005*

*MSRPTVVTVTETPRNPGSYEVNVERDGKMVVGRARAGSDPGAAAAKAMQMAMEWGSPNYVILGSNKVLAFIPEQLRVKM*

Đầu ra của hệ thống sẽ trả về một trong hai giá trị 1 hoặc -1 với 1 tức là protein đó khả năng là anti-CRISPR protein và với -1 thì là không. Kết quả đầu ra lưu dưới định dạng là file *.csv*  với hai trường là *id\_protein* và *nhãn dự đoán.* Ví dụ:

Bảng : Nội dung file .csv đầu ra sau mỗi lần dự đoán bao gồm hai cột là proteinID và label tương ứng là id của protein được dự đoán và kết quả dự đoán label= 1 thì protein đó là một anti-CRISPR và label = -1 thì không.

|  |  |
| --- | --- |
| protein ID | Label |
| WP\_028357638.1 | -1 |
| YP\_007392738.1 | -1 |
| AcrllA4 | 1 |
| AcrllA3 | 1 |

## Tập dữ liệu sử dụng

Những thông tin về anti-CRISPR cho các protein được thu nhập từ Anti-CRISPRdb [6] [*https://doi.org/10.1093/nar/gkaa857*](https://doi.org/10.1093/nar/gkaa857). Cơ sở dữ liệu chứa thông tin hơn 400 anti-CRISPR protein đã được kiểm chứng bằng thực nghiệm và từ các tài liệu khác.

Trong hệ thống được xây dựng, bộ dữ liệu tương ứng với nhãn positives sẽ có 488 anti-CRISPR protein đã được kiểm chứng bằng thực nghiệm. Ngoài ra, bộ dữ liệu tương ứng với nhãn negatives cũng có 902 anti-CRISPR protein với các đặc điểm sau:

* Không được biết hoặc tự đặt tên là anti-CRISPR,
* Phải được phân lập khỏi thực khuẩn thể hoặc từ MGE của vi khuẩn (có thể được biết đến hoặc giả định là MGE), trong đó các chi vi khuẩn nhất định được biết là chứa anti-CRISPR protein,
* Phải có ít hơn 40% trình tự tương đồng với nhau,
* Phải có độ dài nằm trong khoảng từ 50 đến 350.

Trong khi thực hiện đề tài, bộ dữ liệu này được chia làm tập huấn luyện và tập kiểm thử với tỷ lệ 7:3.

## Các độ đo đánh giá

Các phép đo hiệu suất bao gồm Sensitivity (SN), Specificity (SP), Accuracy (ACC), F-value và hệ số tương quan Matthews (MCC- Matthews correlation coefficient) [7], được xác định như sau:

SN =

SP =

ACC =

F- value = 2 ×

MCC =

Trong đó TP , TN , FP và FN lần lượt biểu thị số lượng True Positives, True Negatives, False Positives và False Negatives. Đối với một công cụ dự đoán, SN và SP đo lường khả năng xác định các mẫu positives và negatives của nó tương ứng. ACC, F -value và MCC đo khả năng toàn diện của nó trong việc xác định cả mẫu positives và negatives. Bên cạnh đó, đường cong đặc tính hoạt động của máy thu (ROC), với giá trị AUC (diện tích dưới đường cong) được tính toán, được sử dụng để hình dung hiệu suất dự đoán của một bộ dự báo.

# Các hướng nghiên cứu liên quan (Related Works)

## PaCRISPR - dự đoán và trực quan hóa các protein anti-CRISPR

Link bài báo: https://academic.oup.com/nar/article/48/W1/W348/5847774?searchresult=1

PaCRISPR [2] là một công cụ dự đoán dựa trên máy học để xác định hiệu quả và chính xác các anti-CRISPR dựa trên trình tự protein. PaCRISPR trích xuất bốn loại tính năng tiến hóa để khai thác các mẫu và đặc điểm từ tập dữ liệu đã được xác thực bằng thực nghiệm và đào tạo một tập hợp các mô hình cơ sở với từng tính năng. Sau đó, PaCRISPR tích hợp các mô hình cơ sở này để xây dựng một mô hình tổng thể cho các dự đoán cuối cùng.

Dữ liệu để thực hiện phương pháp PaCRISPR được lấy từ Anti-CRISPRdb [3] gồm 488 anti-CRISPR đã được xác thực bằng thí nghiệm.

Để giải quyết vấn đề mất cân bằng dữ liệu, đối với mỗi đặc trưng, 10 tập hợp con bằng cách kết hợp các mẫu anti-CRISPR được xây dựng và cùng số lượng mẫu không phải anti-CRISPR được chọn ngẫu nhiên từ bộ dữ liệu huấn luyện. Theo đó, 10 bộ phân loại đã được đào tạo bằng cách sử dụng máy vectơ hỗ trợ (SVM) và tích hợp chúng bằng cách lấy trung bình các kết quả dự đoán của chúng. SVM được sử dụng rộng rãi để giải các bài toán phân loại nhị phân trong lĩnh vực sinh học tính toán. Đặc biệt SVM với Radial Basis Function (RBF) kernel đã được sử dụng thành công để phân loại trình tự sinh học phi tuyến. Hai tham số ảnh hưởng đến hiệu suất của SVM dựa trên RBF kernel là Cost (chi phí phân loại sai của tập dữ liệu đào tạo) và Gamma (một tham số cụ thể của RBF kernel). Trong nghiên cứu này, đối với mỗi bộ phân loại dựa trên SVM, các tham số Cost và Gamma được tối ưu hóa bằng cách tìm kiếm trong không gian {2 ^−10 ,…, 2^10 }. Bằng cách này, một mô hình tổng thể như là mô hình cơ sở cho từng đặc trưng (được gọi là mô hình feature-based đơn lẻ). Mô hình tổng hợp cuối cùng được tạo thành bằng cách lấy trung bình điểm dự đoán của các mô hình dựa trên mô hình feature-based đơn lẻ của chúng.

PaCRISPR và AcRanker dự đoán chính xác 258 và 265 không phải anti-CRISPR trong số 266 protein của phage, và 596 và 595 không phải anti-CRISPR trong 597 protein của vi khuẩn. PaCRISPR đã xác định thành công 4 trong 5 nghiên cứu điển hình về anti-CRISPR

## AcrHub - dự đoán và lập bản đồ các protein chống CRISPR

Link bài báo: https://academic.oup.com/nar/article/49/D1/D630/5952202?searchresult=1

Các protein chống CRISPR (Acr) tự nhiên ức chế hệ thống miễn dịch thích ứng CRISPR-Cas trên các lĩnh vực của vi khuẩn. Lĩnh vực mới này hứa hẹn một số ứng dụng hữu ích từ chỉnh sửa gen đến liệu pháp thực khuẩn. Khi số lượng Acrs đã được xác minh và dự đoán mở rộng nhanh chóng, rất ít tài nguyên được phát triển để xử lý lượng thông tin phong phú này. Để khắc phục nhược điểm này, một cơ sở dữ liệu tích hợp (ArcHub) cung cấp giải pháp để điều tra, dự đoán và lập bản đồ các protein Acr. AcrHub [8] lập danh mục 339 Acr không dư thừa đã được xác thực bằng thực nghiệm và hơn 70 000 Acrs được trích xuất từ dữ liệu trình tự bộ gen từ một loạt các sinh vật nhân sơ và vi rút của chúng. Nó tích hợp các công cụ dự đoán hiện đại để dự đoán các Acrs tiềm năng và kết hợp ba mô-đun phân tích: phân tích tương tự, phân tích phát sinh loài và phân tích mạng tương đồng, để phân tích mối quan hệ của chúng với các Acrs đã biết. Bằng cách kết nối tất cả các mô-đun như một nền tảng, AcrHub trình bày phân tích chuyên sâu và phong phú về các Acr đã biết và tiềm năng, do đó cung cấp những hiểu biết mới và thú vị về tương lai của việc khám phá và xác thực Acr.

Vi khuẩn và vi khuẩn cổ đã phát triển nhiều loại hệ thống CRISPR-Cas để tự bảo vệ chúng khỏi các yếu tố di truyền di động có hại (MGE), chẳng hạn như phage và plasmid. Trái lại, MGE đã phát triển một loạt các chất ức chế mạnh, được gọi là anti CRISPR (Acrs), để chống lại hệ thống phòng thủ CRISPR-Cas của vật chủ . Ngoại trừ các chuỗi ngắn phổ biến của chúng, các Acr có rất ít điểm chung với nhau, bao gồm sự tương đồng về trình tự và cấu trúc rất thấp. Ít nhất 50 họ Acr riêng biệt đã được xác định trên cả lĩnh vực sống của vi khuẩn và vi khuẩn cổ, nơi mỗi họ sử dụng các cơ chế phân tử khác nhau để ức chế hệ thống CRISPR-Cas.

Để phân loại hoặc dự đoán các Acrs mới. Ba tài nguyên chính liên quan đến việc phân loại Acrs là Anti-CRISPRdb, đây là một bản tóm tắt các Acrs đã được kiểm chứng bằng thực nghiệm, bảng tính online hợp nhất các bản ghi Acr theo quy ước đặt tên thống nhất và AcrCatalog là danh sách các Acr dự đoán được tạo bằng mô hình dựa trên rừng ngẫu nhiên. Ngoài ra, để xác định các Acrs giả định từ bên trong các chuỗi do người dùng xác định, có sẵn ba bộ công cụ trực tuyến. Đầu tiên, AcRanker, xếp hạng các proteome phage cho các Acr tiềm năng bằng cách sử dụng mô hình dựa trên eXtreme Gradient Boosting. Thứ hai, AcrFinder, tích hợp các phương pháp tiếp cận dựa trên tương đồng, dựa trên lỗi liên kết và tự nhắm mục tiêu để xác định các Acr tiềm năng từ sinh vật nhân sơ và hệ gen của phage. Thứ ba, PaCRISPR , kết hợp thông tin tiến hóa trong một mô hình tổng hợp để dự đoán các protein Acr từ các trình tự gen và hệ gen. Tuy nhiên, các bộ công cụ này hướng đến dự đoán và có giới hạn khả năng hạ nguồn để khám phá thêm các Acrs dự đoán đó.

Ở đây, giới thiệu một nền tảng tích hợp, AcrHub, để cung cấp một giải pháp tất cả trong một, thân thiện với người dùng để dự đoán protein Acr, phân tích đặc tính và mối quan hệ AcrHub lập danh mục 339 Acrs không dư thừa, đã được xác thực bằng thực nghiệm và 71728 Acrs giả định, mỗi Acr được chú thích với mối quan hệ của chúng với Acrs đã biết. Ngoài ra, ba yếu tố dự đoán đã được tích hợp vào AcrHub để dự đoán các điểm tương đồng của Acr và Acrs mới: một dự đoán dựa trên Mô hình Markov ẩn (HMM), AcRanker và PaCRISPR . Để tạo điều kiện thuận lợi cho việc phân tích mối quan hệ giữa các Acrs đã biết và tiềm năng, chúng tôi đã phát triển ba mô-đun phân tích: phân tích tương đồng dựa trên BLAST, phân tích phát sinh loài dựa trên liên kết nhiều trình tự và phân tích mạng dựa trên tương đồng. Các công cụ này có thể hoạt động độc lập hoặc bên trong đường ống AcrHub để tạo điều kiện thuận lợi cho việc phân tích mối quan hệ hạ nguồn của một Acr mới được dự đoán và do đó rút ngắn khoảng cách giữa dự đoán, mô tả chức năng và xác nhận thực nghiệm cuối cùng.

## AcRanker – xếp hạng các anti-CRISPR protein bằng học máy

Link bài báo: https://academic.oup.com/nar/article/48/9/4698/5819938?login=true

AcRanker [9], một phương pháp dựa trên máy học để xác định trực tiếp các anti-CRISPR protein (Acrs). Chỉ sử dụng các đặc điểm về thành phần axit amin, AcRanker xếp hạng một tập hợp các protein ứng cử viên về khả năng trở thành anti-CRISPR protein. Một đánh giá chéo nghiêm ngặt của chương trình được đề xuất cho thấy các Acrs đã biết được xếp hạng cao trong số các proteome. Sau đó, AcRanker được sử dụng để dự đoán 10 Acrs ứng cử viên mới từ các proteome của vi khuẩn có mảng self-targeting CRISPR và xác nhận ba trong số chúng về mặt hóa sinh.

AcRanker đã thu thập thông tin chống CRISPR cho các protein từ Anti-CRISPRdb [6] gồm hơn 400 Acr và sử dụng CD-HIT [10] để xác định một tập hợp (ở ngưỡng tương tự trình tự 40%) làm tập huấn luyện cho công cụ dự đoán. Các protein này gồm 20 Acrs đã được xác minh bằng thực nghiệm , thuộc các lớp Acr khác nhau: 12 trong số các protein hoạt động chống lại hệ thống Cas CRISPR kiểu phụ IF, 4 loại chống lại IE và 4 loại chống lại II-A. Tất cả các protein còn lại trong Anti-CRISPRdb sẽ được dùng làm tập kiểm thử.

Mô hình học máy cơ bản cho AcRanker đã được xây dựng bằng cách sử dụng EXtreme Gradient Boosting (XGBoost) [11]. Trong học máy, boosting là một kỹ thuật trong đó nhiều bộ phân loại yếu được kết hợp để tạo ra một bộ phân loại mạnh. XGBoost là một phương pháp dựa trên cây sử dụng tăng cường theo kiểu end-to-end, tức là mọi cây tiếp theo đều cố gắng giảm thiểu lỗi do người tiền nhiệm của nó tạo ra. XGBoost đã được chứng minh là một thuật toán học nhanh và có thể mở rộng và đã được sử dụng rộng rãi trong nhiều ứng dụng học máy.

Công cụ AcRanker không chỉ ra protein nào là Acr mà chỉ đưa ra danh sách các protein với thứ hạng càng cao thì khả năng đó là Acr cũng cao. Từ danh sách dự đoán đó, các nhà nghiên cứu sẽ lấy các protein có điểm số cao theo một ngưỡng đã được xác định để chứng minh bằng thực nghiệm, và tìm ra các Acr mới. Các dự đoán của AcRanker đã giup các nhà nghiên cứu phát hiện ra hai anti-CRISPR chưa từng biết trước đây AcrllA20 (ML1) và AcrIIA21 (ML8).

## Phương pháp học máy mở rộng danh mục các họ anti-CRISPR protein

Link bài báo [12]: https://www.nature.com/articles/s41467-020-17652-0

Mô hình cho thấy khả năng dự đoán cao khi được thử nghiệm với một tập hợp thử nghiệm không nhìn thấy và được sử dụng để dự đoán 2.500 họ Acr ứng cử viên. Thử nghiệm xác thực các ứng cử viên hàng đầu cho thấy hai Acr không xác định (AcrIC9, IC10) và ba ứng cử viên hàng đầu khác tình cờ được xác định và phát hiện có hoạt động chống CRISPR. Những kết quả này mở rộng đáng kể danh mục các Acr được dự đoán và cung cấp một nguồn tài nguyên để khám phá Acr thử nghiệm.

## AcrFinder: khai thác bộ gen các operon anti-CRISPR ở sinh vật nhân sơ và vi rút

Link bài báo: https://academic.oup.com/nar/article/48/W1/W358/5836766

AcrFinder [13], một máy chủ web ( http://bcb.unl.edu/AcrFinder) kết hợp ba ý tưởng được chấp nhận tốt bởi các nghiên cứu thử nghiệm trước đây để sàng lọc trước dữ liệu bộ gen cho các ứng cử viên Acr. Những ý tưởng này bao gồm tìm kiếm tương đồng, kết hợp tội lỗi (GBA) và bộ đệm tự nhắm mục tiêu CRISPR-Cas. So với các công cụ tin sinh học hiện có, AcrFinder có các chức năng độc đáo sau: (i) đây là máy chủ trực tuyến đầu tiên khai thác cụ thể các bộ gen cho các operon Acr-Aca; (ii) nó cung cấp cơ sở dữ liệu Acr và Aca (cơ quan quản lý liên quan đến Acr) toàn diện nhất (được điền bởi bộ dữ liệu Acr và Aca dựa trên GBA); (iii) nó kết hợp các phương pháp tiếp cận dựa trên tương đồng, dựa trên GBA và tự nhắm mục tiêu trong một gói phần mềm; và (iv) nó cung cấp giao diện web thân thiện với người dùng để lấy cả tệp trình tự nucleotide và protein làm đầu vào và xuất ra trang kết quả với biểu diễn đồ họa về bối cảnh bộ gen của các operon Acr-Aca. Việc xác nhận chéo để lại một lần trên các operon Acr-Aca được đặc trưng bằng thực nghiệm cho thấy rằng AcrFinder có tỷ lệ thu hồi 100%. AcrFinder sẽ là một nguồn tài nguyên web có giá trị để giúp các nhà vi sinh vật thực nghiệm khám phá ra các Anti-CRISPR mới.

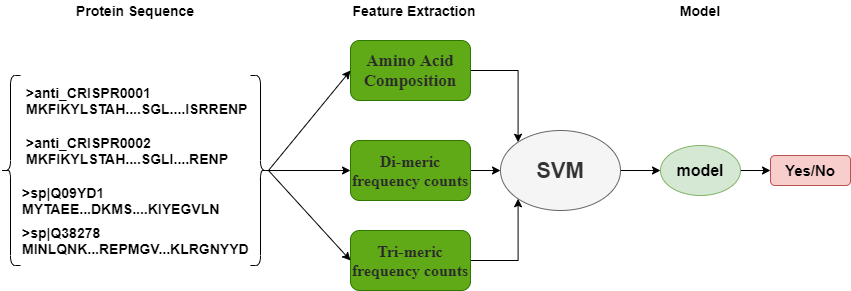
## Phát hiện chính xác các Acr trong sinh vật nhân sơ chỉ sử dụng sáu tính năng

Link bài báo: <https://www.biorxiv.org/content/10.1101/2020.05.23.112011v1>

Trong bài báo này, một công cụ đã được phát triển dựa trên cây rừng ngẫu nhiên, AcrDetector [14], để xác định các Acr trong toàn bộ tỷ lệ gen bằng cách sử dụng chỉ sáu tính năng. AcrDetector có thể đạt được độ chính xác trung bình là 99,65%, độ thu hồi trung bình là 75,84%, độ chính xác trung bình là 99,24% và điểm F1 trung bình là 85,97%; trong nhiều vòng, xác nhận chéo 5 lần (30 trạng thái ngẫu nhiên khác nhau). Để chứng minh rằng AcrDetector có thể xác định Acr thực một cách chính xác ở toàn bộ quy mô bộ gen, chúng tôi đã thực hiện xác nhận giữa các loài, kết quả là 71,43% Acr thực được xếp hạng trong top 10. AcrDetector được áp dụng để phát hiện Acr trong dữ liệu mới nhất. Nó có thể xác định chính xác 3 Acr mà trước đó đã được xác minh bằng thực nghiệm. Phiên bản độc lập của AcrDetector có sẵn tại https://github.com/RiversDong/AcrDetector . Ngoài ra, kết quả của chúng tôi cho thấy rằng hầu hết các Acr được chuyển vào bộ gen chủ của chúng trong giai đoạn gần đây chứ không phải là sớm.

# Đề xuất mô hình

Mô hình học máy cơ bản cho đề tài đã được xây dựng bằng cách sử dụng SVM (support vector machine) (*Hình 2*). Phương pháp học máy SVM (support vector machine) được sử dụng rộng rãi để giải các bài toán phân loại nhị phân trong lĩnh vực sinh học tính toán.



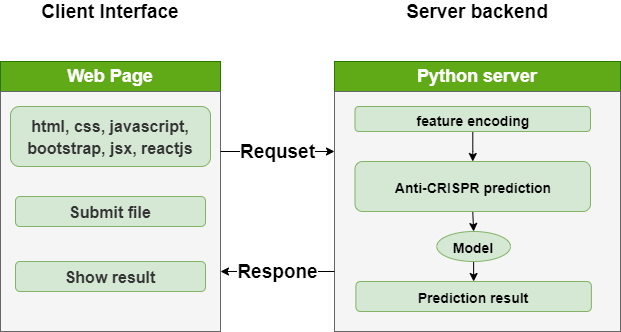
Hình 2: Mô hình học máy SVM tạo ra model: mã hóa các chuỗi trong tập dữ liệu theo ba thành phần amino acid compostion, di- and tri-meric fequency counts.

Để phù hợp với các kỹ thuật dự đoán chức năng protein dựa trên thuật toán học máy SVM, mỗi chuỗi protein sẽ được chia thành ba nhóm dựa trên đặc điểm trình tự [15]: thành phần axit amin và số lượng tần số dimer và trimer [16]. Vì mục đích này, các axit amin đầu tiên được nhóm thành bảy lớp (*Bảng 2)* dựa trên các đặc tính hóa lý của chúng [16] và số lượng tần số của tất cả các nhóm có thể được đánh dấu là dimer và trimers trong một trình tự protein nhất định được sử dụng cùng với thành phần axit amin. Tất cả ba loại feature (thành phần axit amin, số lượng tần số di- và tri-meric) được chuẩn hóa thành định mức đơn vị dẫn đến có 20 +72+ 73 = 412 chiều biểu diễn vectơ đặc trưng cho một trình tự protein nhất định [17] [18]. Các thông số của mô hình học máy được giữ mặc định của thư viện hỗ trợ thuật toán SVM.

|  |  |
| --- | --- |
| Lớp | Axit amin |
| 1 | A, G, V |
| 2 | I, L, F, P |
| 3 | Y, M, T, S |
| 4 | H, N, Q, W |
| 5 | R, K |
| 6 | D, E |
| 7 | C |

Bảng 2: Phân nhóm các axit amin dựa trên tính chất hóa lý. Các nhóm axit amin có chuỗi bên tương tự được nhóm lại với nhau để giảm số lượng các feature cần kiểm tra trong mô hình học máy.

Mô hình web-server trong đề tài bao gồm hai thành phần: giao diện web dành cho người dùng và server xử lý các yêu cầu (*Hình 3*).



Hình 3: Mô hình web-sever: giao diện người dùng được xây dựng bằng thư viện reactjs, và phần server được xậy dựng bằng thư viện flask trên python.

Giao diện web chịu trách nhiệm tương tác với người dùng thông qua hiển thị đầu vào và đầu ra, và xử lý logic dịch vụ bao gồm phát hiện ký tự bất hợp pháp, xác nhận trình tự và định dạng. Giao diện web được xây dựng dựa trên thư viện ReacJS (<https://reactjs.org/>). React là thư viện JavaScript phổ biến nhất để xây dựng giao diện người dùng (UI). Nó cho tốc độ phản hồi tuyệt vời khi người dùng nhập liệu bằng cách sử dụng phương pháp mới để render trang web. Một số ưu điểm: dễ sử dụng, viết component dễ dàng, hỗ trợ tái sử dụng lại các components có cùng chức năng, hiệu suất tốt với DOM.

Phần sever backend chịu trách nhiệm thực hiện toàn bộ quy trình dự đoán, bao gồm nhận dữ liệu đầu vào từ giao diện web người dùng, mã hóa các tính năng, đưa ra dự đoán và tạo dữ liệu sẵn sàng trực quan hóa. Chương trình dự đoán được viết bằng ngôn ngữ Python (https://www.python.org/) phiên bản 3.x, thư viện hỗ trợ xây dựng server trên Python là Flask(https://flask.palletsprojects.com/en/2.0.x/).

# 5. Thực hiện hệ thống

## 5.1. Xử lý dữ liệu đầu vào

Đầu vào của hệ thống là chuỗi protein được lưu lại dưới định dạng file *.fasta* có các thông tin cơ bản về protein đó như *protein\_id, sequence, …,*  sau đó sẽ được tiền xử lý để mã hóa về số để máy cả thể hiểu.

Ví dụ cho một chuỗi protein sau:

*>anti\_CRISPR0001*

*MKFIKYLSTAHLNYMNIAVYENGSKIKARVENVVNGKSVGARDFDSTEQLESWFYGLPGS*

*GLGRIENAMNEISRRENP*

Chuỗi protein anti\_CRSPR0001 có độ dài: L = 78.

Tính số lượng thành phần axit amin, nhóm dimer và nhóm trimer chỉ dựa vào đặc điểm thành phần có trong chuỗi:

* Thành phần axit amin có trong chuỗi (có 20 loại axit amin):

*'A': 5, 'C': 0, 'D': 2, 'E': 7, 'F': 3, 'G': 7, 'H': 1, 'I': 5, 'K': 5, 'L': 5,*

*'M': 3, 'N': 8, 'P': 2, 'Q': 1, 'R': 5, 'S': 7, 'T': 2, 'V': 5, 'W': 1, 'Y': 4*

* Thành phần nhóm dimer có trong chuỗi:

*11:4, 12:2, 13:4, 14:2, 15:4, 16:1, 17 0,*

*21:3, 22:2, 13:3, 24:1, 25:2, 26:3, 27:0,*

*31:4, 32:1, 33:3, 34:3, 35:3, 36:2, 37:0,*

*41:4, 42:5, 43:1, 44:0, 45:0, 46:1, 47:0,*

*51:2, 52:3, 53:2, 54:0, 55:1, 56:2, 57:0,*

*61:0, 62:2, 63:2, 64:5, 65:0, 66:0, 67:0,*

*71:0, 72:0, 73:0, 74:0, 75:0, 76:0, 77:0.*

Các axit amin đươc chia làm 7 lớp (*Bảng 2*), 7 lớp này được tạo nên 72 = 49 thành phần trong nhóm dimer. Tức là: thành phần 11 sẽ là số lượng xuất hiện nhóm hai axit amin được tạo nên từ lớp 1 trong (*Bảng 2*) (tổng AA, AG, AV, GA, GG, GV, VA, VG, VV có trong chuỗi protein). Trong chuỗi anti\_CRSPR0001 có 4 lần xuất hiện như vậy. Thành phần 17 sẽ là số lượng xuât hiện nhóm hai axit amin được tạo nên từ lớp 1 và lớp 7 (tổng của AC, GC, VC có trong chuỗi protein). Trong chuỗi anti\_CRSPR0001 có 0 lần xuất hiện như vậy. Các thành phần còn lại sẽ được tính toán như vậy.

* Thành phần nhóm trimer:

*111: 1, 112: 0, 113: 1, 114: 1, 115: 1, 116: 0, 117: 0,*

*121: 1, 122: 1, 123: 0, 124: 0, 125: 0, 126: 0, 127: 0,*

*………*

*771: 0, 772: 0, 773: 0, 774: 0, 775: 0, 776: 0, 777: 0*

Tương tự như nhóm dimer thì nhóm trimer được tạo nên từ 73 = 343 thành phần. Thành phần 111 trong nhóm sẽ là tổng số lần xuất hiện nhóm 3 axit amin được tạo nên từ lớp 1 trong *Bảng 2* (AGV, AVG, GAV, GVA, VAG, VGA). Trong chuỗi protein anti\_CRSPR0001 chỉ có VGA xuất hiện 1 lần nên thành phần 111 sẽ có giá trị là 1. Tương tự với các thành phần còn lại có trong nhóm trimer.

Sau khi tính các thành phần axit amin, nhóm dimer, nhóm trimer, các giá trị này sẽ chia với độ dài L của chuỗi để lấy phần trăm xuất hiện trong chuỗi, sau đó sẽ được chuẩn hóa dữ liệu thông qua hàm của thư viện sklearn.preprocessing: *sklearn.preprocessing.normalize(X, norm='l2', \*, axis=1, copy=True, return\_norm=False)*

Cuối cùng, mỗi thành phần đã được chuẩn hóa trên sẽ được hợp lại thành một vector có 20 + 72 + 73 = 412 chiều. Protein anti\_CRSPR0001 sau khi mã hóa:

*[0.2469324 0. 0.09877296 0.34570536 0.14815944 0.34570536*

*0.04938648 0.2469324 0.2469324 0.2469324 0.14815944 0.39509184*

*………………*

*0. 0. 0. 0. 0. 0.*

*0. 0. 0. 0. ]*

## 5.2. Thực hiện hệ thống

Sau khi mã hóa và gán nhãn các chuỗi protein có trong tập dữ liệu, kết quả thu được sẽ có 2 ma trận đầu vào cho mô hình SVM:

* Ma trận X: là vector sau khi mã hóa protein
* Ma trận y: là nhãn được gắn tương ứng với các vector có trong X (nhãn 1 là positive: protein là một anti\_CRISPR protein, -1 là nagative: protein không phải là một anti\_CRISPR protein).

Với đầu vào là 2 ma trận X, y như trên, ta cần chia ra thành tập huấn luyện và tập kiểm thử với tỷ lệ tương ứng là 7:3 như sau:

*from sklearn.model\_selection import train\_test\_split*

*X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=.3, random\_state=1)*

Model của bài toán sẽ được tạo ra và lưu lại dưới dạng file .*pickle* để dễ dàng sử dụng sau khi thực hiện:

*import sklearn*

*from sklearn import svm*

*from sklearn.svm import SVC*

*model = SVC()*

*model.fit(X\_train, y\_train)*

*pickle.dump(model, open('./model/svc.pickle', 'wb'))*

## 5.3. Xử lý kết quả đầu ra

Kết quả đầu ra sau mỗi lần dự đoán sẽ được lưu dưới dạng file *.csv* có nội dung như Bảng 1.

# 6. Thử nghiệm và kết quả

## 6.1. Môi trường thử nghiệm

Cấu hình máy tính CPU, RAM, GPU

* CPU: intel core i5
* RAM: 8GB
* GPU:

Các phần mềm, bộ công cụ toolkits đã sử dụng.

* Phần mền: Visual Studio Code, Jupyter Notebook
* Python: 3.x.x
* Packages are required: Biopython, sklearn, scipy, ….

## 6.2. Các thử nghiệm

Kết quả đo hiệu suất của thuật toán:

|  |  |
| --- | --- |
| Sensitivity | 0.84 |
| Specificity | 0.96 |
| Accuracy | 0.93 |
| F-value | 0.89 |

Độ chính xác Accuracy khá các lên đến 93%. Sau khi thử dự đoán 98 anti-CRISPR protein với đặc điểm là các protein này giống nhau đến trên 40% về trình tự thì cho phương pháp dự đoán ra 36 mẫu là negative trong khi cả 98 mẫu đều là positive. Có sự sai lệch này có thể do bộ dữ liệu để tạo ra model dự đoán chưa đồng đều, chưa bao quát hết các trường hợp.

Để kiểm tra thêm phương pháp được đề xuất, 26 trình từ anti-CRISPR protein mới được phát hiện, rất khác biệt sau đó đã được thu thập từ các tài liệu mới nổi được ghi lại trong tài nguyên chống CRISPR trực tuyến thống nhất [19]. Các mẫu anti-protein này bao gồm tập dữ liệu độc lập, trong đó chúng có độ tương đồng ít hơn 10% so với 98 anti-CRISPR trong tập dữ liệu huấn luyện, ngoại trừ hai mẫu có độ tương đồng là 21,38% và 56,12% . Sau đó, chúng tôi thu thập 260 không không phải là anti-CRISPR sử dụng các tiêu chí tương tự để lựa chọn các mẫu negative trong tập dữ liệu huấn luyện, các mẫu này có độ tương đồng trình tự <40% so với tập dữ liệu đào tạo và các mẫu dương tính trong tập dữ liệu độc lập. Tổng cộng, tập dữ liệu độc lập có 26 mẫu positive và 266 mẫu negative.

Đối với tập dữ liệu này thì phương pháp SVM được trình bài đã dự đoán chính xác 260 mẫu negative trong tổng số 266 mẫu negative ở trên, còn đối với 26 mẫu positive thì kết quả dự đoán chỉ cho ra 5 kết quả postive (điều này thực sự kém).

Sau đó, hai mẫu dữ liệu trên đã được bổ sung vào trong tập để tạo ra model dự đoán mới. Sau khi đánh giá kết quả của model thì thu được các thông số sau:

|  |  |
| --- | --- |
| Sensitivity | 0.81 |
| Specificity | 0.94 |
| Accuracy | 0.93 |
| F-value | 0.87 |

Các thông số đánh giá không khác biệt nhiều so với tập dữ liệu ban đầu nên việc dự đoán lại hai mẫu dữ liệu trên không có thay đổi đáng kể với 9 mẫu positive trong 26 mẫu positive.

# 7. Thảo luận

Khi dự đoán các protein dài không phải là anti-CRISPR từ thể thực khuẩn và MGE của vi khuẩn, cả PaCRISPR và AcRanker đều đạt được độ chính xác dự đoán âm tính thực sự cao. Cụ thể, PaCRISPR và AcRanker dự đoán chính xác 258 và 265 protein không phải là anti-CRISPR trong số 266 protein của phage, và 596 và 595 protein không phải là anti-CRISPR trong 597 protein của vi khuẩn. Điều này, cùng với sự cân bằng của các giá trị SN và SP cao trên cả xác nhận chéo 5 lần và các xét nghiệm độc lập, đã chứng minh hiệu suất vượt trội của PaCRISPR trong việc xác định các anti-CRISPR từ thể thực khuẩn và proteome vi khuẩn với tỷ lệ dương tính giả thấp.

Phương pháp SVM trong bài báo cáo này cũng đã được dự đoán với tập dữ liệu như hai phương pháp thì cho ra kết quả là 260 protein không phải là anti-CRISPR trong tổng số 266 protein của phage và 592 protein không phải là anti-CRISPR trong 597 protein của vi khuẩn. Từ các kiểm tra như trên cho thấy rằng phương pháp SVM trong bài báo cáo này có thể đạt độ chính xác như PaCRISPR và AcRanker.

Tuy nhiên, nhiều kết quả thu được từ phương pháp đã trình bày trong bài có được cho là chưa thực sự tốt so với các kết quả từ bộ công cụ AcRanker và PaCRISPR bởi:

* Mô hình chỉ dự vào đặc điểm trên trình tự, chưa bao quát hết các trường hợp như: sự đa dạng về trình tự trong protein (độ dài, thứ tự các axit amin,…)
* Thuật toán SVM với các tham số mặc định chưa phải đủ tốt để tạo ra model dự đoán. Để khắc phục cần có các thuật toán tốt hơn như XGBoost, … hay các phương pháp học sâu CNN, DNN, RNN,…
* Bộ dữ liệu trong đề tài này chưa đủ đa dạng để bao quát hết các trường hợp. Bộ dữ liệu chỉ có 488 protein là anti-CRISPR và 902 là các protein khác. Do đó, cần có bộ dữ liệu đủ lớn để các thể tạo ra các dự đoán tốt hơn

# 8. Tổng kết và phương hướng phát triển (Conclusions and Perspectives)

Việc khám phá ra anti-CRISPR protein mới mở ra một con đường để sử dụng máy móc CRISPR-Cas như một công cụ trong việc chỉnh sửa gen hoặc liệu pháp gen và cung cấp những hiểu biết mới về tương tác của thể thực khuẩn với vật chủ vi khuẩn của chúng. Trong bài báo cáo này, phương pháp học máy SVM, một công cụ dự đoán tổng hợp và phổ quát để xác định chính xác và hiệu quả các anti-CRISPR từ dữ liệu trình tự có nguồn gốc từ hệ gen và hệ số. Phương pháp nhằm mục đích khai thác và trích xuất các mẫu và đặc điểm khác biệt từ các anti-CRISPR đã biết. Phương pháp này tạo sự khác biệt với môi trường tương tác mở rộng, nơi người dùng có thể dự đoán về anti-CRISPR giả định dựa trên sự tương đồng về trình tự và phân tích phát sinh loài. Để đảm bảo bộ công cụ PaCRISPR duy trì tính cạnh tranh và cập nhật, nó sẽ được nâng cấp định kỳ khi các bộ chống CRISPR mới được xác định và xác thực bằng thực nghiệm. Trong đề tài này, một phương pháp học máy đã được áp dụng là SVM, như một bước đầu tiên hướng tới dự đoán trực tiếp của các gen Acr chỉ dựa vào các đặc điểm trình tự. Sử dụng một sự kết hợp của SVM và thực nghiệm, các nhà nghiên cứu có thể nhanh chóng làm giảm tới một ứng viên gen để tổng hợp trực tiếp và thử nghiệm các đặc tính của anti-CRISPR protein. Áp dụng các phương pháp học máy sẽ làm thêm một công cụ khác vào hộp công cụ của tìm kiếm anti-CRISPR protein bằng cách cung cấp một giải pháp thay thế cho BLAST để tìm các gia đình Acr mới. Tuy nhiên, nhiều trường hợp BLAST cho ra kết quả tốt hơn, cung cấp một cách tiếp cận bổ sung tiềm năng, mặc dù BLAST bị cho là phương pháp ít có khả năng dẫn đến các Acrs mới hơn. Khả năng xác định các Acr mới tiềm năng trực tiếp từ trình tự protein với học máy SVM mở ra cánh cửa cho việc thử nghiệm nhiều loại protein mới mà không cần nỗ lực sàng lọc tốn nhiều công sức. Tìm kiếm bên trong các bộ gen có chứa các mảng self-targeting-CRISPR hứa hẹn sẽ đặc biệt hiệu quả, vì các chất ức chế tiềm năng cho một hệ thống CRISPR cụ thể có thể nhanh chóng được dự đoán để tạo danh sách ngắn các ứng viên để kiểm tra. Hy vọng rằng các phương pháp dự đoán Acr trực tiếp như SVM sẽ tiếp tục tiết lộ thêm nhiều Acr phân bố trên nhiều loài vi khuẩn, tìm ra các Acr mới với các đặc tính độc đáo cho các ứng dụng công nghệ sinh học chưa lường trước được trong tương lai.

Phương pháp trong bài báo cáo này chưa thực sự tốt để có thể đưa vào thực tế do các khó khăn khi nhận dạng protein. Chỉ dựa vào đặc điểm trình tự là chưa đủ để dự đoán chính xác các protein mà cần xem xét đến thứ tự các axit amin có trong chuỗi trình tự. Do đó, hướng pháp triển trong tương lai có thể áp dụng các mô hình học sâu như DNN, RNN,… để có thể dự đoán tốt hơn so với các phương pháp học máy. Mã hóa tính năng protein cần kết hợp nhiều tính năng tiến hóa trong một khuôn khổ tổng hợp. So với các đặc điểm dựa trên trình tự, các đặc điểm tiến hóa được chứng minh là phù hợp hơn để dự đoán loại protein tiến hóa cao này. Các phương pháp học máy tạo sự khác biệt với môi trường tương tác mở rộng, nơi mà máy móc có thể dự đoán về anti-CRISPR giả định dựa trên sự tương đồng về trình tự và phân tích phát sinh loài. Máy móc có thể đóng vai trò như một bộ công cụ sàng lọc sơ bộ hữu ích để xác định các anti-CRISPR tiềm năng, và do đó xúc tiến việc phát hiện ra các anti-CRISPR mới để xác nhận thử nghiệm tiếp theo của chúng. Để đảm bảo bộ công cụ các sự đoán chính xác hơn, các dự đoán này cần được xác định và xác thực bằng thực nghiệm.

# Tài liệu tham khảo

|  |  |
| --- | --- |
| [1] | K. L. Maxwell, "The Anti-CRISPR Story: A Battle for Survival," *Molecular Cell,* vol. 68, no. 1, pp. 8-14, 2017. |
| [2] | K. L. M. Sabrina Y. Stanley, " Phage-encoded anti-CRISPR defenses," *Annual Review of Genetics,* vol. 52, no. 1, pp. 445-464, 2018. |
| [3] | R. H. S. C. T. B. N. W. S. S. P. C. F. K. L. M. &. A. R. D. April Pawluk, "Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species," *Nature Microbiology,* vol. 1, no. 1, p. 16085, 2016. |
| [4] | M. R. S. J. F. H. M. J. M. N. J. K. J. B.-D. Benjamin J. Rauch, "Inhibition of CRISPR-Cas9 with Bacteriophage Proteins," *Cell,* vol. 168, no. 1-2, pp. 150-158, 2017. |
| [5] | Z. R. Yang, "Biological applications of support vector machines," *Briefings in Bioinformatics,* vol. 5, no. 4, pp. 328-338, 2004. |
| [6] | G.-F. H. H.-L. H. S. L. A. A. L. G. C. J. H. N. R. F.-B. G. Chuan Dong, "Anti-CRISPRdb: a comprehensive online resource for anti-CRISPR proteins," *Nucleic Acids Research,* vol. 46, no. D1, p. D393–D398, 2018. |
| [7] | B.W.Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure,* vol. 405, no. 2, pp. 442-451, 1975. |
| [8] | W. D. J. L. Q. L. R. X. Y. Z. C. S. T. L. Jiawei Wang, "AcrHub: an integrative hub for investigating, predicting and mapping anti-CRISPR proteins," *Nucleic Acids Research,* vol. 49, no. D1, p. D630–D638, 2021. |
| [9] | A. A. W. K. I. A. K. G. D. J. M. F. Eitzinger S., "Machine learning predicts new anti-CRISPR proteins.," *Nucleic Acids Research,* vol. 48, no. 9, p. 4698–4708, 2020. |
| [10] | B. N. Y. G. L. F. W. L. Ying Huang, "CD-HIT Suite: a web server for clustering and comparing biological sequence," *Bioinformatics,* vol. 26, no. 5, pp. 680-682, 2010. |
| [11] | C. G. Tianqi Chen, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* vol. 1, no. 1, pp. 785-794, 2016. |
| [12] | A. E. P. A. L. B. S. A. S. K. S. M. Y. I. W. J. B.-D. &. E. V. K. Ayal B. Gussow, "Machine-learning approach expands the repertoire of anti-CRISPR protein families," *Nature Communications,* vol. 11, no. 1, p. 3784, 2020. |
| [13] | L. H. ,. B. Y. ,. J. G. ,. H. Z. ,. Y. Y. Haidong Yi, "AcrFinder: genome mining anti-CRISPR operons in prokaryotes and their viruses," *Nucleic Acids Research,* vol. 48, no. W1, pp. 358-365, 2020. |
| [14] | D.-K. P. C. M. X. W. Q.-F. W. Z. Z. V. O. P.-B. G. Chuan Dong, "Precise detection of Acrs in prokaryotes using only six features," *The preprint server for biology,* vol. 1, no. 1, p. 2, 2020. |
| [15] | M. M. &. E. M. N. Rabie Saidi, "Protein sequences classification by means of feature extraction with substitution matrices," *BMC Bioinformatics,* vol. 11, no. 1, p. 175, 2010. |
| [16] | J. Z. X. L. W. Z. K. Y. K. C. Y. L. H. J. Juwen Shen, "Predicting protein–protein interactions based only on sequences information," *PNAS,* vol. 104, no. 11, pp. 4337-4341, 2007. |
| [17] | E. E. A. W. S. N. Christina Leslie, "The spectrum kernel: A string kernel for SVM protein classification," *Biocomputing,* vol. 1, no. 1, pp. 564-575, 2002. |
| [18] | O. C. Eisenhaber, "Data Mining Techniques for the Life Sciences," *Methods in Molecular Biology,* vol. 609, no. 1, pp. 223-239, 2010. |
| [19] | A. R. D. J. A. D. P. C. F. K. L. M. S. M. X. P. E. J. S. B. W. Joseph Bondy-Denomy, "A Unified Resource for Tracking Anti-CRISPR Names," *The CRISPR Journal,* vol. 1, no. 5, pp. 302-305, 2018. |
| [20] | A. A. K. E. W. A. T. I. G. J. K. J. A. D. F. u. A. A. M. Simon Eitzinger, "Machine learning predicts new anti-CRISPR proteins," *Nucleic Acids Research,* vol. 48, no. 9, p. 4698–4708, 2020. |