ĐÁNH GIÁ CÁC PHƯƠNG PHÁP, QUY TRÌNH NHẬN DẠNG CHUỖI PLASMID HIỆN NAY VÀ ÁP DỤNG TRÊN DỮ LIỆU MÔ PHỎNG

Nguyễn Đình Quân
Học viện Công nghệ Bưu chính Viễn thông, Việt Nam

***Tóm tắt*—**Việc hiểu biết về di truyền của vi khuẩn và các cơ chế đằng sau khả năng kháng kháng sinh phụ thuộc quan trọng vào việc phát hiện plasmid. Trong bài viết này, chúng tôi cung cấp một phân tích toàn diện về các công cụ nhận diện plasmid khác nhau, bao gồm các phương pháp dựa trên căn chỉnh, dựa trên học máy, và phương pháp lai. Các nhà nghiên cứu và chuyên gia có thể thu được nhiều thông tin hữu ích từ việc xem xét ưu và nhược điểm của từng phương pháp. Chúng tôi cũng làm rõ các quy trình làm việc được sử dụng để chuẩn bị dữ liệu và đánh giá các công cụ này bằng cách sử dụng cả dữ liệu tổng hợp từ NCBI. Qua nghiên cứu kỹ lưỡng, chúng tôi chứng minh rằng các quy trình đề xuất có hiệu quả và độ tin cậy cao trong việc nhận diện chính xác plasmid trong cả môi trường mô phỏng và thực tế. Những phát hiện này góp phần vào khoa học nhận diện plasmid và mở ra hướng nghiên cứu tiếp theo về kháng kháng sinh và di truyền vi sinh vật.

***Từ khóa*—**Vi khuẩn kháng kháng sinh, Plasmid, tin sinh, học máy.

I. GIỚI THIỆU

Một phân tích chi tiết về các kỹ thuật nhận diện plasmid cho thấy rõ những ưu điểm và nhược điểm của các chiến lược khác nhau. Vì các phương pháp dựa trên căn chỉnh dựa vào sự tương đồng của trình tự, chúng có thể bỏ qua những khác biệt quan trọng khi cố gắng phân biệt giữa các plasmid có độ khác biệt lớn. Mặt khác, mặc dù các công cụ dựa trên học máy rất hứa hẹn, độ chính xác của chúng có thể bị giảm do phụ thuộc vào các mẫu đã được xác định trước đó. Các kỹ thuật căn chỉnh dành cho các plasmid có quan hệ gần gũi được kết hợp một cách khéo léo bởi các phương pháp lai, giúp việc nhận diện các trình tự tương tự trở nên dễ dàng hơn. Các công cụ dựa trên motif hoặc tần suất k-mer, chẳng hạn như PPR-Meta [8] và PlasFlow [2], hoạt động đặc biệt tốt trong các trường hợp mà các trình tự có các mẫu dễ nhận diện. Tuy nhiên, chúng có thể không dự đoán chính xác được các plasmid mới với các đặc tính bất thường hoặc không mong đợi. Thiếu sự căn chỉnh hoặc không có các motif có thể dẫn đến dự đoán không chính xác, buộc các công cụ phải dựa vào các tiêu chí cơ bản hơn như số lượng gen hoặc độ dài trình tự.

Các phương pháp lai như PLASMe [3] sử dụng các mô hình Transformer đặc thù theo thứ tự để dự đoán các liên kết dựa trên thứ tự trình tự thay vì căn chỉnh trực tiếp khi phải đối mặt với các plasmid khác biệt. Việc đưa các protein đặc trưng của nhiễm sắc thể vi khuẩn vào các cơ sở dữ liệu plasmid dường như là một bước đi khôn ngoan nhằm tăng độ chính xác của dự đoán. Vì hầu hết các protein thiết yếu cho sự sống của vi khuẩn đều được tìm thấy trong nhiễm sắc thể, việc bổ sung chúng có thể cải thiện hiệu suất của các mô hình học máy. Hơn nữa, các mô hình Transformer đã tiết lộ khả năng tồn tại các protein plasmid chưa được đặc trưng hóa, cho thấy vai trò chức năng tiềm năng vượt ra ngoài các chú thích hiện có. Có thể cung cấp thông tin mới về tầm quan trọng tiến hóa và sinh thái của các plasmid bằng cách dự đoán những vai trò này. Chúng tôi sẽ giải thích các phương pháp nhận diện plasmid hiện tại được sử dụng trong báo cáo này, bao gồm cả ưu điểm và nhược điểm của chúng. Tiếp theo, chúng tôi sẽ thảo luận về một quy trình chi tiết liên quan đến việc chuẩn bị dữ liệu để đánh giá phương pháp. Cuối cùng, chúng tôi sẽ đánh giá một số phương pháp bằng cách sử dụng dữ liệu thực tế và đề xuất các hướng nghiên cứu tiếp theo.

II. CÁC THUẬT TOÁN ĐỂ NHẬN DẠNG CHUỖI PLASMID

A. Phương pháp dựa trên căn chỉnh các chuỗi

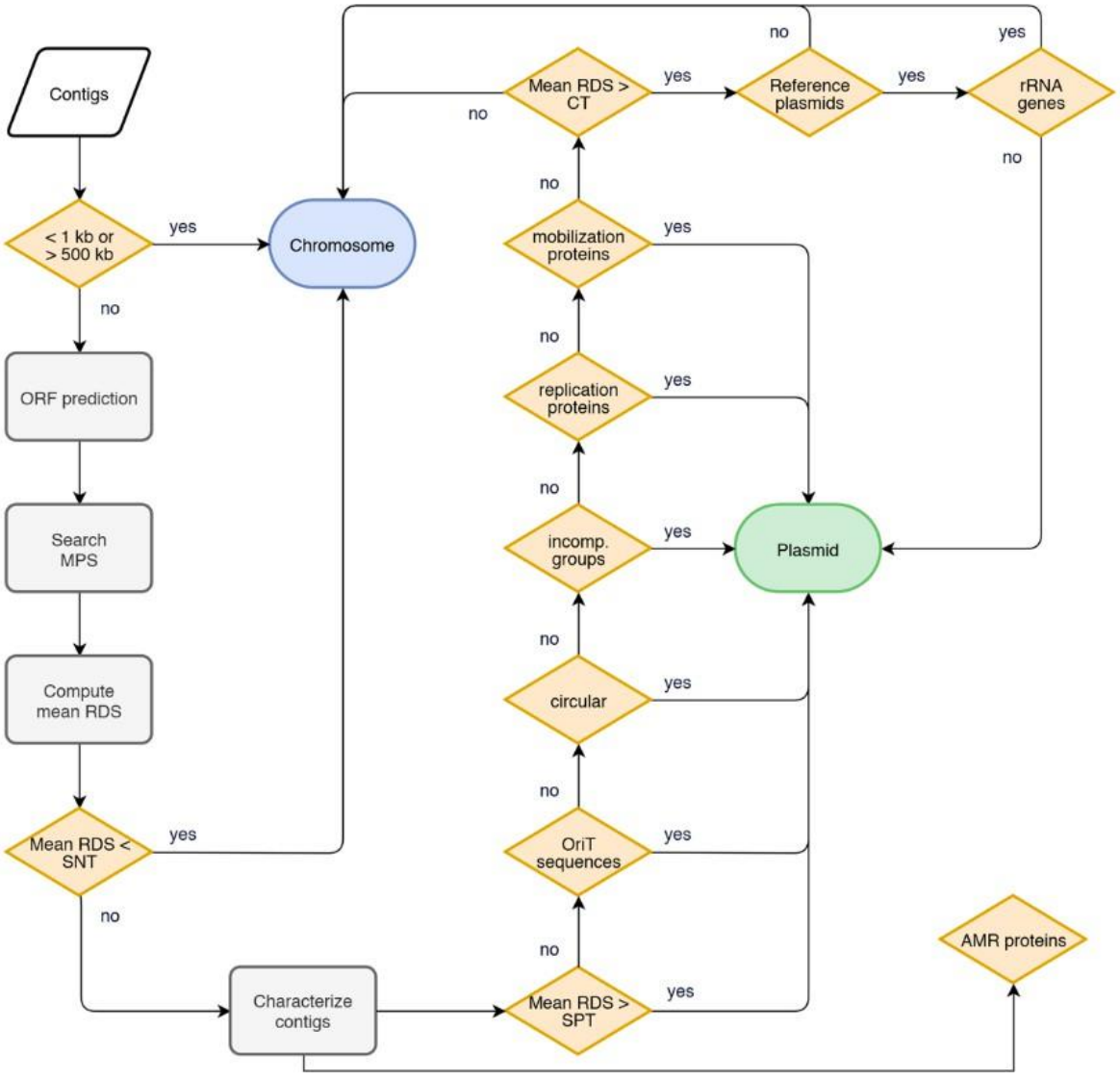
Một yếu tố ngày càng trở nên quan trọng là những hiểu biết về "dữ liệu lớn." Lượng dữ liệu trình tự ngày càng tăng đòi hỏi các phương pháp xử lý tự động với hiệu suất cao là cần thiết. Vì thiết kế tương tác hoặc triển khai dựa trên web, không phải tất cả các sản phẩm phần mềm tin sinh học hiện có trên thị trường đều đủ khả năng để phân tích với hiệu suất cao, chưa nói đến việc tích hợp mượt mà vào các quy trình nghiên cứu lớn hơn. Kiến trúc cơ sở dữ liệu chuyên biệt theo phân loại cũng đặt ra những thách thức bổ sung vì người dùng không có đủ hỗ trợ tin sinh học hoặc sức mạnh tính toán để tạo ra các cơ

Tác giả liên hệ: Nguyễn Đình Quân,
Email: quannd@ptit.edu.vn
Đến tòa soạn: 08/2024, chỉnh sửa: 09/2024, chấp nhận đăng: 10/2024.

sở dữ liệu đa phân loại lớn hoặc chuyên biệt. Hơn nữa, do dữ liệu này không phải lúc nào cũng dễ dàng truy cập, sự phụ thuộc vào dữ liệu thô hoặc dữ liệu trung gian—như các trình tự đọc và đồ thị lắp ráp—có thể cản trở quá trình phân tích. Xử lý với hiệu suất cao yêu cầu các công cụ phần mềm tin sinh học phải được thiết kế và triển khai theo cách đáp ứng các yêu cầu về khả năng mở rộng của dữ liệu lớn. Điều này bao gồm việc cung cấp các đầu ra phân loại có thể hành động, chẳng hạn như phân loại nhị phân chính xác, tạo ra giao diện dòng lệnh không tương tác và, nếu có thể, áp dụng thiết kế cơ sở dữ liệu độc lập với phân loại.

Để xác định nguồn gốc replicon của các contig, công cụ như Platon [4] kết hợp một loạt các đặc điểm cấp cao của contig với việc điều tra sự phân bố lệch của trình tự protein theo replicon. Platon đầu tiên phân loại các contig dài hơn 500 kbp hoặc ngắn hơn 1 kbp là nhiễm sắc thể. Heuristic này dựa trên quan sát rằng các trình tự chứa 1 k b p hiếm khi chứa mã CDS hoặc thông tin có thể khai thác được để cho phép phân loại chính xác. Tuy nhiên, theo kinh nghiệm của chúng tôi, các contig lớn hơn 500 kbp hiếm khi, nếu có, được lấy từ plasmid vì chúng thường bao gồm các đặc điểm di truyền khiến các trình tự lớn khó lắp ráp, như transposon và integron. Do đó, heuristic này cải thiện tốc độ xử lý mà không làm giảm đáng kể hiệu suất phân loại.

Trong giai đoạn tiếp theo, các trình tự mã hóa (CDS) được dự đoán bằng cách sử dụng Prodigal và được truy



Hình 1. Sơ đồ quy trình mô tả luồng công việc được triển khai trong Platon. ORF, các khung đọc mở; MPS, trình tự protein đánh dấu; RDS, điểm phân bố replicon; SNT, ngưỡng độ nhạy; SPT, ngưỡng đặc trưng; nhóm không tương thích, các nhóm không tương thích; CT, ngưỡng bảo thủ.

vấn với cơ sở dữ liệu các yếu tố di truyền di động (MGE) thông qua Diamond, áp dụng các ngưỡng phát hiện nghiêm ngặt phù hợp với thông số cụm của các cụm UniRef90 cơ bản. Các ngưỡng này yêu cầu độ bao phủ không dưới 80% và độ tương đồng của trình tự không dưới 90%. Sau đó, tỷ lệ dồi dào tương đối trung bình (RDS) của tất cả các MGE được xác định sẽ được tính toán cho mỗi contig. Các contig có RDS trung bình thấp hơn ngưỡng độ nhạy (SNT) được phân loại là các trình tự nhiễm sắc thể. Ngược lại, các contig vượt qua SNT sẽ được phân tích kỹ lưỡng theo quy trình đã được mô tả trong phần trước.

Việc phân loại tiếp theo các contig là plasmid phụ thuộc vào việc thỏa mãn một hoặc nhiều tiêu chí sau:

(i) đạt được RDS trung bình vượt ngưỡng đặc trưng (SPT); (ii) chứng minh khả năng circularizability;

(iii) chứa ít nhất một protein sao chép hoặc huy động; (iv) chứa một yếu tố không tương thích; (v) chứa một trình tự oriT; (vi) đạt được RDS trung bình vượt qua ngưỡng bảo thủ (CT) và có kết quả BLAST+ khớp với cơ sở dữ liệu plasmid RefSeq mà không mã hóa các gen ribosome.

B. Các phương pháp dựa trên học máy

Sử dụng các phương pháp tiếp cận học máy để tự động khám phá các trình tự plasmid trong bất kỳ metagenome nào là một cách tiếp cận thú vị khác để nhận diện plasmid từ dữ liệu metagenomic bắn phá (shotgun). Các phương pháp dựa trên dấu hiệu di truyền đã tìm ra mối quan hệ giữa thành phần %G+C của bộ gen và sự tương đồng giữa plasmid và vật chủ. Hơn nữa, các mã vạch của bộ gen plasmid thường thể hiện các đặc điểm tương tự, có thể là do áp lực chọn lọc tương tự gây ra bởi sự chuyển giao thường xuyên của chúng giữa các môi trường nuôi cấy tế bào.

Có một loạt các thuật toán được thiết kế cho vấn đề phân loại contig, với nhiều trong số đó mới được phát triển gần đây. Các phương pháp này chủ yếu tận dụng các tiếp cận học máy. Phương pháp sớm nhất cho việc phân loại contig, cBar [5], tiên phong sử dụng hồ sơ k-mer của một contig làm đặc tính chính trong mô hình phân loại học máy. Trong cBar, mô hình được huấn luyện bằng một tập dữ liệu lớn bao gồm các tổ hợp bộ gen vi khuẩn đã đóng. Nguyên tắc chung của việc sử dụng các thuộc tính k-mer làm đặc tính phân loại cũng đã được áp dụng bởi nhiều bộ phân loại học máy hiện đại, bao gồm PlasFlow, mlplasmids [6], và PlasClass [7]. PPR-Meta [8] đại diện cho một phương pháp học sâu dựa trên các chuỗi contig mã hóa một-hot.

C. Các phương pháp lai

PlasForest và Deeplasmid là hai phương pháp gần đây dựa trên các mô hình học máy, tận dụng các đặc tính đa dạng của một contig, chẳng hạn như thành phần GC (thường khác nhau giữa plasmid và nhiễm sắc thể) và sự hiện diện của các trình tự đặc trưng của plasmid, được xác định thông qua việc đối chiếu với cơ sở dữ liệu plasmid tham chiếu. RFPlasmid kết hợp cả hồ sơ k-mer và các đặc tính của trình tự plasmid.

Quy trình PLASMe dựa trên thiết kế trên được thể hiện trong Hình 2. Đầu tiên, chúng tôi lọc các contig có độ dài ngắn hơn 1k hoặc dài hơn 350k. Sau đó, chúng tôi sẽ đối chiếu chúng với cơ sở dữ liệu plasmid có tên DB, chứa 33 125 trình tự plasmid tham chiếu. Nếu các contig được đối chiếu với mức độ bao phủ truy vấn và độ tương đồng cao (Tcov = Tident = 90% theo mặc định), chúng sẽ được phân loại là plasmid. Ngược lại, chúng sẽ được gán vào các thứ tự dựa trên giá trị e của việc đối chiếu (nhỏ hơn 10). Sau đó, Transformer đặc trưng theo thứ tự sẽ được gọi để dự đoán plasmid. Nếu truy vấn không thể được gán vào bất kỳ thứ tự nào bằng BLAST dưới giá trị e mặc định, nó sẽ được phân loại là không phải plasmid. Dựa trên phân tích, sự tương đồng giữa các plasmid trong cùng một thứ tự có thể dẫn đến các đối chiếu với giá trị e đáng kể. Do đó, bước này có thể giúp lọc ra một số lượng lớn các trình tự không phải plasmid. Các phần tiếp theo sẽ chi tiết cách chúng tôi huấn luyện Transformer để nhận diện plasmid.

III. PHƯƠNG PHÁP

Để đánh giá hiệu quả của các mô hình một cách hiệu quả, điều cần thiết là chuẩn bị các tập dữ liệu phù hợp và khách quan. Các tập dữ liệu này cần phải đủ lớn và bao gồm một loạt các loài đa dạng. Hơn nữa, phần này cũng sẽ trình bày một quy trình toàn diện để tạo điều kiện truy cập vào dữ liệu cho những người có nền tảng hạn chế trong lĩnh vực tin sinh học.

A. Tải dữ liệu từ ngân hàng dữ liệu NCBI

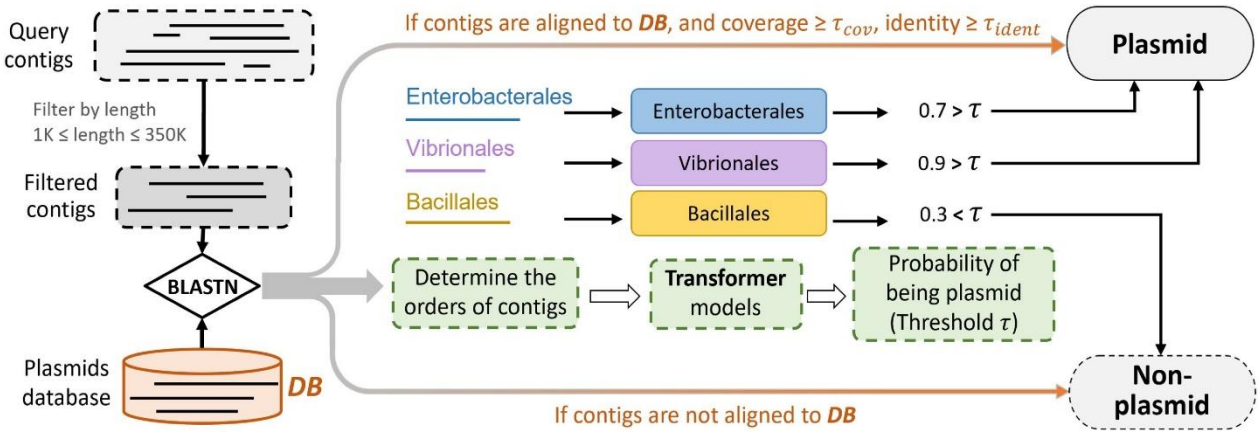
Tập dữ liệu ban đầu được tải xuống từ cơ sở dữ liệu NCBI, tập trung cụ thể vào các loài vi khuẩn được báo cáo là có khả năng kháng thuốc đa dạng, bao gồm Escherichia coli, Staphylococcus aureus và Klebsiella pneumoniae. Công cụ ete3 được sử dụng để hỗ trợ việc thu thập thông tin dựa trên các định danh phân loại. Tổng cộng có 8.759 mẫu vi khuẩn đã được xử lý sau khi lọc bỏ các mẫu có thông tin không đầy đủ.

B. Chuỗi đôi

Dựa trên dữ liệu đã thu thập, cho mỗi mẫu, chúng tôi đã xác định toàn bộ bộ gen của các plasmid và nhân đôi chiều dài trình tự plasmid bằng công cụ Biopython. Quy trình này là cần thiết vì plasmid thường có cấu trúc hình tròn, trong khi dữ liệu tải xuống lại trình bày dưới dạng các chuỗi tuyến tính. Thông tin tại các vị trí nối của plasmid, bao gồm các đoạn gen, có thể không sẵn có ngay. Do đó, việc nối chuỗi hai lần đảm bảo rằng các contig thu được bao quát đầy đủ thông tin về plasmid.

C. Sinh ra các đoạn đọc NGS tổng hợp

Sau đó, chúng tôi sử dụng Bộ mô phỏng đọc Assembly Read Simulator (ART) để tạo ra các contig từ các bản lắp ghép đọc ngắn. ART là một tập hợp các công cụ mô phỏng tạo ra các đọc tổng hợp



Hình 2. Quy trình của PLASMe gồm ba bước: lọc các contig theo độ dài, căn chỉnh với cơ sở dữ liệu plasmid có độ tương đồng cao, và xác định các contig là plasmid dựa trên độ đồng nhất và độ phủ. Sau đó, PLASMe sắp xếp thứ tự contig, đưa vào mô hình transformer, và dự đoán plasmid nếu xác suất vượt ngưỡng đã chọn.

của chuỗi thế hệ tiếp theo. Chức năng này rất quan trọng để kiểm tra và đánh giá các công cụ phân tích dữ liệu chuỗi thế hệ tiếp theo, bao gồm căn chỉnh đọc, lắp ghép de novo và phát hiện biến thể di truyền. ART tạo ra các đọc chuỗi giả bằng cách mô phỏng quy trình chuỗi với các mô hình lỗi đọc cụ thể theo công nghệ được tích hợp sẵn và các hồ sơ giá trị chất lượng cơ sở được tham số hóa thực nghiệm trong các tập dữ liệu chuỗi lớn. Hiện tại, chúng tôi hỗ trợ cả ba nền tảng chuỗi thế hệ tiếp theo thương mại chính: Roche's 454, Illumina's Solexa và Applied Biosystems' SOLiD. ART cũng cho phép linh hoạt trong việc sử dụng các tham số mô hình lỗi đọc tùy chỉnh và các hồ sơ chất lượng. Đối với mỗi bộ gen, các đọc chuỗi Illumina được tạo ra bằng cách sử dụng ART (v2.5.8) dưới các tham số cụ thể: chuỗi ghép đôi với chiều dài đọc là 100 cặp cơ sở (bp), đạt được độ phủ gấp 50 lần, và mô phỏng kích thước mảnh trung bình là 400 bp.

D. Xử lý và tạo ra các contig

Dữ liệu thu được sẽ được xử lý qua quy trình AMRomics. AMRomics là một bộ phần mềm được thiết kế để phân tích các bộ dữ liệu di truyền vi sinh vật. Nó bao gồm một quy trình tích hợp các phương pháp tiên tiến cho nhiều khía cạnh của phân tích kháng thuốc (AMR). Kết quả của phân tích quy trình có thể được truy cập và trực quan hóa thông qua một ứng dụng web, ứng dụng này cũng cung cấp khả năng quản lý dữ liệu hiệu quả.

E. Căn chỉnh contig với cơ sở dữ liệu tham chiếu

Cuối cùng, Minimap2 được sử dụng để căn chỉnh các contig với cơ sở dữ liệu tham chiếu. Minimap2 là một chương trình căn chỉnh đa mục đích để ánh xạ các chuỗi DNA hoặc mRNA dài chống lại một cơ sở dữ liệu tham chiếu lớn. Nó hoạt động với các đọc ngắn chính xác có chiều dài lớn hơn 100 bp, các đọc gen 1 kb với tỷ lệ lỗi 15%, các đọc RNA hoặc cDNA dài đầy đủ có độ ồn và các contig lắp ghép hoặc nhiễm sắc thể đầy đủ có liên quan chặt chẽ với chiều dài hàng trăm megabases. Minimap2 thực hiện căn chỉnh đọc tách rời, sử dụng chi phí khoảng trống lõm cho các chèn và xóa dài và giới thiệu các phương pháp mới để giảm căn chỉnh giả. Nó nhanh gấp 3–4 lần so với các công cụ ánh xạ đọc ngắn phổ biến ở độ chính xác tương đương và nhanh hơn 30 lần so với các công cụ ánh xạ gen hoặc cDNA dài ở độ chính xác cao hơn, vượt trội hơn hầu hết các công cụ căn chỉnh chuyên dụng cho một loại căn chỉnh.

Đối với các chuỗi đọc ồn 10 kb, Minimap2 nhanh gấp hàng chục lần so với các công cụ ánh xạ đọc dài phổ biến như BLASR, BWA- MEM, NGMLR và GMAP. Nó chính xác hơn trên các đọc dài giả và tạo ra căn chỉnh có ý nghĩa sinh học sẵn sàng cho các phân tích tiếp theo. Đối với các đọc ngắn Illumina lớn hơn 100 bp, Minimap2 nhanh gấp ba lần so với BWA-MEM và Bowtie2, và chính xác tương đương trên dữ liệu giả.

IV. THỰC NGHIỆM

Chúng tôi trước tiên so sánh hiệu suất của từng phương pháp trên bộ dữ liệu NCBI. Trên bộ dữ liệu này, chúng tôi cũng so sánh các công cụ nhận diện plasmid khác nhau, bao gồm các phương pháp dựa trên học máy (PlasClass, RFPlasmid, mlplasmid), các phương pháp dựa trên căn chỉnh (PlasmidFinder, Platon) và các phương pháp lai (plasmidVerify, PLASMe, Deeplasmid). Chúng tôi đánh giá kết quả bằng cách sử dụng các chỉ số phổ biến, chẳng hạn như độ chính xác, độ thu hồi (recall), độ chính xác (precision) và điểm F1. Tất cả các công cụ được đánh giá đều nhận các contig làm đầu vào và do đó có thể được đánh giá trên cùng một bộ thử nghiệm. Để có một so sánh công bằng, chúng tôi không bao gồm các công cụ lắp ghép plasmid dựa trên đồ thị, vì chúng yêu cầu các đầu vào khác và không được tối ưu hóa cho các contig.

A. Dữ liệu mô phỏng: Cơ sở dữ liệu trình tự tham chiếu NCBI

Cơ sở dữ liệu NCBI (Trung tâm Thông tin Sinh học Quốc gia) là một bộ sưu tập toàn diện về dữ liệu sinh học được công bố công khai bởi Viện Y tế Quốc gia (NIH) tại

Hoa Kỳ. Bộ dữ liệu này bao gồm một loạt thông tin sinh học, bao gồm các chuỗi gen, tài liệu y sinh, cấu trúc phân tử và nhiều hơn nữa. Các nhà nghiên cứu, khoa học gia và cộng đồng khoa học rộng lớn sử dụng bộ dữ liệu NCBI cho nhiều mục đích khác nhau như nghiên cứu genomics, phát hiện thuốc, điều tra bệnh tật và các nghiên cứu tiến hóa. Tính khả dụng và sự phong phú của bộ dữ liệu NCBI làm cho nó trở thành một tài nguyên quý giá để thúc đẩy nghiên cứu y sinh và hiểu biết về những phức tạp của sự sống ở cấp độ phân tử.

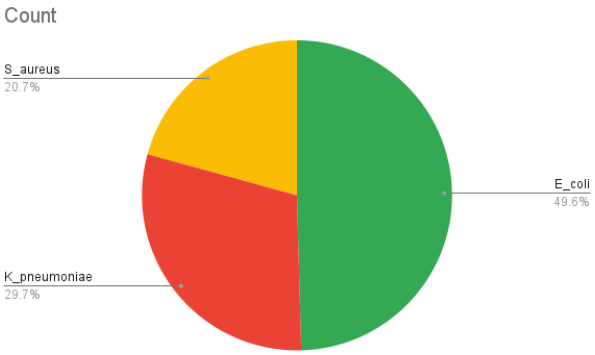
Bộ dữ liệu NCBI được mô phỏng từ 8759 bộ gen hoàn chỉnh của ba loài: Escherichia coli, Staphy- lococcus aureus và Klebsiella pneumoniae. Đây là ba loài vi khuẩn quan trọng có ảnh hưởng đa dạng đến sức khỏe con người. E. coli là một cư dân phổ biến trong ruột người, đóng vai trò cả có lợi và gây bệnh, gây ra các bệnh nhiễm trùng đường tiêu hóa và đường tiết niệu. S. aureus cư trú trên da và niêm mạc nhưng có thể trở thành gây bệnh, gây ra nhiễm trùng da, viêm phổi và nhiễm trùng máu, thường thể hiện kháng kháng sinh. Klebsiella pneumoniae, thường có trong ruột, có thể dẫn đến các bệnh nhiễm trùng hô hấp và đường tiết niệu, đặc biệt là trong các cơ sở y tế, nơi mà các chủng kháng đa thuốc gây khó khăn trong điều trị. Hiểu biết về các vi khuẩn này là rất quan trọng để quản lý nhiễm trùng và phát triển các chiến lược điều trị hiệu quả.

Chúng tôi thu thập dữ liệu và áp dụng các bước tiền xử lý đã nêu trong quy trình để có được kết quả cuối cùng bao gồm các contig được phân loại theo loài, kèm theo nhãn tương ứng. Tập dữ liệu mô phỏng mô phỏng rất gần với các quy trình xử lý dữ liệu trong thế giới thực và có thể được sử dụng làm tiêu chuẩn để đánh giá hiệu quả của các mô hình phát hiện plasmid hiện tại.

B. Các chỉ số đánh giá

Đối với mỗi nhiệm vụ phân loại, chúng tôi đánh giá một số chỉ số hiệu suất, bao gồm độ chính xác (precision), độ hồi tưởng (recall), điểm F1 (trung bình hài hòa của độ chính xác và độ hồi tưởng) và độ chính xác tổng thể (accuracy). Đối với các phương pháp gán điểm hoặc xác suất cho các contig, kết quả phân loại phụ thuộc cao vào ngưỡng điểm đã chọn. Lưu ý rằng, độ chính xác và độ hồi tưởng không xác định khi mẫu số bằng không, và điểm F1 không xác định nếu độ chính xác hoặc độ hồi tưởng không xác định. Chúng tôi đánh giá từng chỉ số này trên mỗi tập hợp được bao gồm trong bộ thử nghiệm (nơi có thể) và báo cáo giá trị trung vị. Tất cả các công cụ mà chúng tôi so sánh đều thể hiện sự biến thiên lớn về độ chính xác của các dự đoán giữa các mẫu riêng lẻ. Giá trị trung vị được chọn vì các chỉ số thu được ít bị ảnh hưởng bởi các giá trị ngoại lai.

Các nhãn thật và dự đoán tạo ra các số lượng true positives (TP), true negatives (TN), false positives (FP) và false negatives (FN) cho nhiệm vụ phân loại. Mỗi contig được tính là một đơn vị duy nhất, bất kể chiều dài của nó. Các contig không có nhãn thực không được đưa vào đánh giá. Lưu ý rằng, các contig có nhãn thực là "mơ hồ" được coi là dương tính cho cả nhiệm vụ phân loại plasmid và nhiễm sắc thể.



Hình 3. Phân bố thành phần các loài vi khuẩn trong bộ dữ liệu

V. KẾT QUẢ

Chúng tôi đã đánh giá ban đầu hiệu suất của các phương pháp này bằng cách sử dụng một tập dữ liệu thử nghiệm bao gồm 834 contig, được phân loại đều thành hai nhóm: plasmid và nhiễm sắc thể, đặc thù cho một chủng E. coli duy nhất. Kết quả cho thấy các phương pháp dựa trên căn chỉnh (alignment-based) luôn thể hiện giá trị độ chính xác (Precision) cao đáng kể. Quan sát này có thể được giải thích bởi sự phổ biến rộng rãi của E. coli và sự phong phú thông tin có sẵn trong các cơ sở dữ liệu liên quan đến loài này. Đặc biệt, phương pháp Platon đạt được điểm F1 cao nhất (0.91), cho thấy hiệu quả của nó trong việc xác định plasmid. Tuy nhiên, cần có thêm đánh giá trên nhiều loài khác nhau để xác định tính tổng quát của phương pháp này. Ngược lại, các phương pháp dựa trên học máy (learning-based) thường cho kết quả không tối ưu do phụ thuộc vào các đặc điểm đơn giản như chiều dài contig và phân phối k-mer, nhấn mạnh tầm quan trọng của việc tận dụng thông tin từ các cơ sở dữ liệu phong phú.

Hiện nay, có nhiều phương pháp tích hợp chiến lược lai (hybrid) kết hợp giữa tài nguyên cơ sở dữ liệu với các kỹ thuật học máy hoặc học sâu tiên tiến. PLASMe đại diện cho một phương pháp sáng tạo kết hợp kiến trúc Transformer với tài nguyên cơ sở dữ liệu, thể hiện hiệu quả đáng kể. Điều này được chứng minh bởi PLASMe đạt được điểm F1 cao thứ hai (0.854), chỉ sau Platon (0.91). Phương pháp này sử dụng các mã token cụm protein như một thay thế cho các phương pháp phân đoạn k-mer thông thường, mã hóa Byte-Pair (BPE) hoặc phân đoạn ở mức ký tự. PLASMe đạt được độ chính xác vượt trội trong khi giảm độ phức tạp của mô hình, từ đó nâng cao hiệu quả tính toán.

Để thực hiện một đánh giá toàn diện hơn về các phương pháp dựa trên căn chỉnh và các phương pháp học, chúng tôi đã mở rộng phân tích của mình sang một tập dữ liệu lớn hơn. Cụ thể, số lượng contig nhiễm sắc thể đã tăng từ 419 lên 26.630. Đối với đánh giá mở rộng này, chúng tôi đã chọn một mô hình dựa trên căn chỉnh, Platon, và hai mô hình dựa trên học, RFPlasmid và

Bảng I
KẾT QUẢ ĐÁNH GIÁ TRÊN DỮ LIỆU MÔ PHỎNG NCBI.

Method	Type	plasmid.fasta	chromosome.fasta	Precision	Recall	F1-score
Platon	Alignment-based	0.85	0.005	0.99	0.86	0.91
PlasmidFinder	Alignment-based	0.15	0	1	0.15	0.26
PlasClass	Learning based	0.77	0.026	0.967	0.77	0.838
mlplasmid	Learning based	0.58	0.01	0.976	0.575	0.724
RFPlasmid	Learning based	0.74	0.03	0.957	0.742	0.836
PLASme	Hybrid	1	0.34	0.746	1	0.854
Deeplasmid	Hybrid	0.277	0.024	0.92	0.277	0.426
plasmidVerify	Hybrid	0.828	0.124	0.867	0.828	0.848

mlplasmid (chúng tôi đã loại trừ các mô hình lai khỏi thử nghiệm này do hạn chế về thời gian, với kế hoạch đánh giá chúng trong các báo cáo tương lai với thông tin chi tiết hơn). Kết quả cho thấy với tập dữ liệu lớn hơn, các phương pháp dựa trên căn chỉnh vẫn tiếp tục thể hiện hiệu suất vượt trội so với các phương pháp chỉ dựa vào học máy hoặc các mô hình học sâu cơ bản.

VI. THẢO LUẬN

Dựa trên kết quả thực nghiệm của chúng tôi, các contig ngắn có xu hướng chứa nhiều kết quả dương giả hơn do sự chuyển gen ngang giữa các nhiễm sắc thể và plasmid. Việc xác định plasmid từ các trình tự ngắn có thể dẫn đến các kết quả dương giả. Do đó, việc xác định các trình tự plasmid bị phân mảnh vẫn là một nhiệm vụ thách thức.

Phương pháp tiếp cận lai như PLASMe sử dụng token cấp protein (PC-level token) làm đặc điểm đầu vào. So với k-mer, mã hóa byte-pair (BPE) và mã hóa cấp ký tự, token cấp protein tốt hơn trong việc học tầm quan trọng của các protein khác nhau và mối quan hệ giữa chúng. Đặc biệt đối với các trình tự dài hơn, việc sử dụng các protein làm token và nhúng chúng vào các vector tạo ra các vector ngắn hơn nhiều so với các vector được tạo ra bởi k-mer hoặc BPE. Việc sử dụng các vector đầu vào ngắn hơn có thể làm giảm độ phức tạp của mô hình (ít tham số hơn), điều này giúp quá trình huấn luyện mô hình trên GPU. Do đó, dưới tài nguyên tính toán hạn chế, việc sử dụng một bộ phân loại cấp protein có thể tích hợp thông tin từ các trình tự dài hơn tốt hơn. Tuy nhiên, việc mã hóa cấp protein cũng có một số hạn chế, vì nó dễ gặp phải vấn đề từ vựng ngoài (Out Of Vocalbulary). Do đó, đối với các plasmid mới hơn, việc sử dụng token cấp protein có thể dẫn đến các kết quả âm tính giả.

Bảng II
SO SÁNH HIỆU SUẤT GIỮA CÁC MÔ HÌNH DỰA TRÊN CĂN CHỈNH VÀ CÁC MÔ HÌNH HỌC MÁY

Model	Type 2	F1-score
Platon	Alignment-based	0.915
RFPlasmid	Learning based	0.765
mlplasmid	Learning based	0.714

VII. KẾT LUẬN

Trong bài này, chúng tôi đã đánh giá nhiều công cụ để xác định plasmid trong các lắp ráp đọc ngắn. Vẫn còn một số vấn đề cần giải quyết trong công việc tương lai. Đầu tiên, do plasmid có sự đa dạng lớn, plasmid mới có thể chứa các protein không thể căn chỉnh với cơ sở dữ liệu hiện tại, dẫn đến dự đoán không chính xác. Nếu một contig chỉ chứa các protein mới không thể căn chỉnh với cơ sở dữ liệu, điều này sẽ là một trường hợp khó khăn cho hầu hết các công cụ. Cụ thể, các công cụ lai và dựa trên căn chỉnh không thể xác định các contig này vì không có sự căn chỉnh. Nếu không có sự căn chỉnh, chúng chỉ có thể sử dụng các đặc điểm như chiều dài chuỗi và số lượng gen chứa trong đó, dẫn đến độ chính xác thấp. Các công cụ dựa trên học máy phụ thuộc vào động lực hoặc tần suất k-mer chỉ có thể dự đoán chính xác nếu các trình tự chứa các động lực cụ thể hoặc phân phối tần suất k-mer đã được nhìn thấy trước đó. Tuy nhiên, các công cụ căn chỉnh như Platon có thể phân loại các contig này là nhiễm sắc thể do các căn chỉnh kém. Để cải thiện độ nhạy của mô hình, bên cạnh việc cập nhật và mở rộng cơ sở dữ liệu kịp thời, chúng tôi sẽ đào sâu hơn vào mối quan hệ giữa các protein plasmid để xây dựng cầu nối giữa các token đã biết và chưa biết.

Thứ hai, các token hiện tại từ PLASMe chỉ chứa các protein của plasmid. Các nghiên cứu đã chỉ ra rằng các protein quan trọng đối với sự sống sót của vi khuẩn có nhiều khả năng xuất hiện trong nhiễm sắc thể. Do đó, việc thêm các protein đặc hiệu cho nhiễm sắc thể có thể giúp cải thiện độ chính xác của mô hình học hơn nữa.

Cuối cùng, việc giải thích mô hình Transformer đã xác định các protein tiềm năng chưa được chú thích có thể cũng đóng vai trò quan trọng trong đời sống của plasmid. Dự đoán chức năng của các protein plasmid chưa được chú thích này có thể giúp nghiên cứu ý nghĩa tiến hóa cũng như sinh thái của plasmid.

TÀI LIỆU THAM KHẢO

[1] Zhencheng Fang, Jie Tan, Shufang Wu, Mo Li, Congmin Xu, Zhongjie Xie, and Huaiqiu Zhu. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience*, 8(6):giz066, 2019. Oxford University Press.

[2] Pawel S Krawczyk, Leszek Lipinski, and Andrzej Dziem-

bowski. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic acids research*, 46(6):e35–e35, 2018. Oxford University Press.

[3] Xubo Tang, Jiayu Shang, Yongxin Ji, and Yanni Sun. PLASMe: a tool to identify PLASMid contigs from short- read assemblies using transformer. *Nucleic Acids Re- search*, 51(15):e83–e83, 2023. Oxford University Press.

[4] Oliver Schwengers, Patrick Barth, Linda Falgenhauer, Torsten Hain, Trinad Chakraborty, and Alexander Goesmann. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microbial Genomics*, 6(10):e000398, 2020. Microbiology Society.

[5] Fengfeng Zhou and Ying Xu. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, 26(16):2051–2052, 2010. Oxford University Press.

[6] Sergio Arredondo-Alonso, Malbert RC Rogers, Johanna C Braat, Tess D Verschuuren, Janetta Top, Jukka Corander, Rob JL Willems, and Anita C Schu¨rch. mlplasmids: A user-friendly tool to predict plasmid-and chromosome-derived sequences for single species. *Microbial Genomics*, 4(11):e000224, 2018. Microbiology Society.

[7] David Pellow, Itzik Mizrahi, and Ron Shamir. PlasClass improves plasmid sequence classification. PLoS Computational Biology, 16(4):e1007781, 2020. Public Library of Science San Francisco, CA USA.

[8] Zhencheng Fang, Jie Tan, Shufang Wu, Mo Li, Congmin Xu, Zhongjie Xie, and Huaiqiu Zhu. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. Gigascience, 8(6):giz066, 2019. Oxford University Press.

EVALUATION OF CURRENT METHODS AND PROCESSES FOR PLASMID SEQUENCE RECOGNITION AND APPLICATION ON SIMULATED DATA.

***Abstract*—**Understanding the genetics of bacteria and the mechanisms underlying antibiotic resistance depends critically on the detection of plasmids. Here, we provide a comprehensive analysis of various plas- mid identification tools, including alignment-based, learning-based, and hybrid approaches. Researchers and practitioners can gain significant insights from this examination of each method's strengths and limits. We also clarify workflows that are used to prepare data, and evaluate these tools using synthetic data from the NCBI. We show via thorough study how effective and reliable the suggested pipelines are at correctly recognizing plasmids in both simulated and real-world settings. These discoveries further the science of plasmid identification and set the stage for next investigations into antibiotic resistance and microbial genetics.

***Keywords*—**Antimicrobial resistance, Plasmid, Bioinformatics, machine learning.

Nguyễn Đình Quân Thạc sĩ, giảng viên khoa Trí tuệ nhân tạo - Học viện Công nghệ Bưu chính Viễn thông. Thạc sĩ Nguyễn Đình Quân tốt nghiệp bằng thạc sĩ ngành Khoa học máy tính ở Đại học Cleveland State ở thành phố Cleveland, Mỹ. Lĩnh vực nghiên cứu bao gồm học máy, ứng dụng hoc sâu trong xử lý dữ liệu tin sinh, dữ liệu y tế.