

ĐẠI HỌC QUỐC GIA HÀ NỘI  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ



**TỔNG QUAN VỀ PHƯƠNG PHÁP PHÂN LOẠI THỰC  
KHUẨN DỰA TRÊN TÍNH TOÁN**

**BÁO CÁO MÔN HỌC**  
**INT 7021 - Tin sinh học cho dữ liệu lớn**

**HÀ NỘI - 2025**

**ĐẠI HỌC QUỐC GIA HÀ NỘI  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**



**TỔNG QUAN VỀ PHƯƠNG PHÁP PHÂN LOẠI THỰC  
KHUẨN DỰA TRÊN TÍNH TOÁN**

**BÁO CÁO MÔN HỌC**

**INT 7021 - Tin sinh học cho dữ liệu lớn**

**Cán bộ hướng dẫn: TS. Hoàng Thị Diệp  
TS. Đặng Cao Cường**

**Nhóm học viên: 23025097 - Trịnh Bá Tú  
23025086 - Vũ Quang Sơn  
23025079 - Lê Thế Nam**

**HÀ NỘI - 2025**

# Mục lục

Mục lục

Danh sách hình vẽ

Danh sách bảng

Danh mục các từ viết tắt

Tóm tắt

<b>Chương 1 Mở đầu</b>	<b>1</b>
1.1 Giới thiệu . . . . .	1
1.1.1 Vi khuẩn, vi-rút, thực khuẩn và vùng protein bảo tồn . . . . .	1
1.1.2 Động lực thúc đẩy . . . . .	2
1.2 Phát biểu bài toán . . . . .	2
1.2.1 Định nghĩa bài toán . . . . .	3
1.2.2 Đặc điểm của bài toán . . . . .	3
1.2.3 Thách thức trong quá trình thực hiện . . . . .	3
1.3 Nguồn dữ liệu và định dạng dữ liệu . . . . .	5
1.3.1 Nguồn dữ liệu . . . . .	5
1.3.2 Định dạng dữ liệu . . . . .	5
<b>Chương 2 Các nghiên cứu liên quan</b>	<b>8</b>
2.1 Phân loại theo phương pháp học máy . . . . .	8
2.2 Sắp xếp theo thứ tự thời gian công bố . . . . .	8
2.3 Phương pháp trình bày trong các phần tiếp theo . . . . .	9
2.4 PHACTS . . . . .	10
2.4.1 Dữ liệu huấn luyện . . . . .	10
2.4.2 Phương pháp . . . . .	10
2.4.3 Kết quả và đánh giá . . . . .	12

2.5	PhageAI . . . . .	13
2.5.1	Dữ liệu huấn luyện . . . . .	13
2.5.2	Phương pháp . . . . .	14
2.5.3	Kết quả và đánh giá . . . . .	16
2.6	BACPHLIP . . . . .	17
2.7	DeePhage . . . . .	19
2.8	PhaTYP . . . . .	22
2.9	DeepPL . . . . .	24
<b>Chương 3</b>	<b>Thực nghiệm</b>	<b>28</b>
3.1	Xây dựng bộ dữ liệu . . . . .	28
3.1.1	Xử lý nhãn . . . . .	28
3.2	Kịch bản thực nghiệm . . . . .	29
3.3	Các chỉ số đánh giá . . . . .	29
3.4	Kết quả . . . . .	29
3.4.1	Thực nghiệm với mô hình PHACTS . . . . .	29
3.4.2	Kết quả thực nghiệm với PhageAI . . . . .	30
3.4.3	So sánh hiệu suất phân loại giữa DeePhage và XGBoost trên bộ dữ liệu xây dựng . . . . .	31
<b>Chương 4</b>	<b>Kết luận</b>	<b>33</b>
4.1	Dữ liệu . . . . .	33
4.2	Phương pháp . . . . .	34
4.3	Kết quả thực nghiệm . . . . .	35
4.4	Hướng nghiên cứu tiếp theo . . . . .	36
	<b>Tài liệu tham khảo</b>	<b>37</b>

# Danh sách hình vẽ

1.1	Sơ đồ mô tả bài toán phân loại thực khuẩn dựa trên tính toán . . . .	3
2.1	Quy trình giải pháp đề xuất của PhageAI . . . . .	14
2.2	Cấu trúc mạng nơ-ron học sâu và hình ảnh hóa 5 lớp bằng cách giảm kích thước của DeePhage . . . . .	21
2.3	Kiến trúc mô hình PhaTYP sử dụng BERT . . . . .	22
2.4	Kiến trúc mô hình DeepPL sử dụng NDABERT . . . . .	26
3.1	Kết quả hiệu suất phân loại của mô hình XGBoost trên tập dữ liệu xây dựng. . . . .	31

# Danh sách bảng

2.1	Bacteriophages life cycle prediction benchmark for 11 supervised ML tuned classifiers with 10-fold cross-validation . . . . .	16
2.2	So sánh kết quả của BACPHLIP với các Mavrich và PHACTS. . . .	18
2.3	So sánh hiệu suất của PhaTYP với các công cụ khác. . . . .	24
2.4	So sánh hiệu suất giữa DeepPL và các công cụ khác. . . . .	27
3.1	Kết quả thực nghiệm với mô hình PHACTS . . . . .	29
3.2	Kết quả so sánh các thuật toán học máy theo Accuracy, Sensitivity và Specificity trên bộ dữ liệu kết hợp. . . . .	30
4.1	Bảng tổng hợp thông tin các bộ dữ liệu . . . . .	34
4.2	Bảng tổng hợp phương pháp và mô hình các bài báo sử dụng . . . .	35

# Danh mục các từ viết tắt

STT	Từ viết tắt	Cụm từ đầy đủ (tiếng Anh)	Giải nghĩa tiếng Việt
1	Phage	Bacteriophage	Thực khuẩn - vi-rút ký sinh trên vi khuẩn
2	Virulent phage	Virulent phage	Thực khuẩn thể độc lực
3	Temperate phage	Temperate phage	Thực khuẩn thể ôn hòa
4	Lytic cycle	Lytic cycle	Chu kỳ tan
5	Lysogenic cycle	Lysogenic cycle	Chu kỳ tiền tan
6	Genome	Genome	Bộ gen
7	Contig	Contig	Đoạn trình tự liên tiếp
8	NGS	Next-Generation Sequencing	Giải trình tự thế hệ tiếp theo
9	k-mer	k-mer	Đoạn con độ dài k
10	Sliding window	Sliding window	Cửa sổ trượt – kỹ thuật cắt chuỗi
11	AUC	Area Under Curve	Diện tích dưới đường cong ROC
12	MetaSim	MetaSim	Trình mô phỏng dữ liệu giải trình tự

## Tóm tắt

Thực khuẩn là các vi-rút có khả năng xâm nhiễm vào vi khuẩn. Dựa trên chu kỳ sống, thực khuẩn được phân thành hai nhóm chính. Nhóm thứ nhất là thực khuẩn thể độc lực, thực hiện chu kỳ tan, trong đó thực khuẩn nhân lên nhanh chóng và phá hủy tế bào vi khuẩn sau khi xâm nhiễm. Nhóm thứ hai là thực khuẩn thể ôn hòa, có khả năng tích hợp bộ gen của mình vào nhiễm sắc thể của vi khuẩn và sao chép cùng tế bào chủ thông qua chu kỳ tiềm tan. Từ đó, việc xác định chính xác chu kỳ sống của thực khuẩn là một bước quan trọng trong việc phát triển các ứng dụng phù hợp, đặc biệt trong trị liệu bằng phage, một lựa chọn tiềm năng cho các bệnh nhân dị ứng hoặc kháng thuốc kháng sinh.

Trước đây, các phương pháp phân loại thực khuẩn truyền thống chủ yếu dựa trên nuôi cấy trong phòng thí nghiệm, vốn đòi hỏi nhiều thời gian, chi phí cao và không hiệu quả khi xử lý khối lượng lớn dữ liệu chưa được gán nhãn. Hiện nay, dữ liệu di truyền ngày càng phong phú, được thu thập từ các nguồn như NCBI, PhageScope hay PhageDB cho phép thực hiện các phương pháp phân loại phage dựa trên tính toán. Với sự phát triển của công nghệ, các kỹ thuật học máy và học sâu đã được áp dụng để thực hiện phân loại thực khuẩn đạt được các kết quả tốt và rút ngắn thời gian phân loại so với các phương pháp truyền thống.

Mục tiêu của báo cáo là cung cấp một cái nhìn tổng quan và hệ thống về bài toán phân loại thực khuẩn dựa trên phương pháp tính toán. Trong báo cáo này, nhóm sinh viên chia các phương pháp thành hai nhóm chính: (1) các phương pháp học máy truyền thống giải quyết bài toán với dữ liệu bộ gen đầy đủ (PHACTS, PhageAI, BACPHLIP), và (2) các phương pháp học sâu có khả năng xử lý bài toán với dữ liệu không hoàn chỉnh (DeePhage, PhaTYP, DeepPL).

Báo cáo gồm có 4 chương:

1. **Giới thiệu.** Nội dung của chương này là giới thiệu về các khái niệm liên quan, phát biểu bài toán và nguồn dữ liệu.
2. **Các nghiên cứu liên quan.** Chương này tập trung trình bày về các phương pháp phân loại thực khuẩn dựa trên tính toán tiêu biểu.
3. **Thực nghiệm.** Chương này trình bày về dữ liệu, kịch bản, chỉ số, kết quả liên quan tới các thực nghiệm mà nhóm sinh viên đạt được.
4. **Kết luận.** Trong chương này, nhóm sinh viên tập trung thảo luận về kết quả đạt được và hướng nghiên cứu tiếp theo.



# Chương 1

## Mở đầu

### 1.1 Giới thiệu

#### 1.1.1 Vi khuẩn, vi-rút, thực khuẩn và vùng protein bảo tồn

Vi khuẩn là sinh vật đơn bào có kích thước nhỏ (0,2 - 5  $\mu\text{m}$ ), có cấu tạo đơn giản bao gồm tế bào chất, màng tế bào và vách tế bào. Vi khuẩn sinh sản chủ yếu bằng phương pháp phân đôi. Ngược lại, vi-rút không phải là sinh vật sống hoàn chỉnh do không có cấu trúc tế bào, mà chỉ bao gồm vật liệu di truyền (ADN hoặc ARN) được bao bọc bởi lớp vỏ protein. Vi-rút chỉ có thể sao chép khi xâm nhập vào tế bào vật chủ và khai thác bộ máy sinh học của tế bào đó.

Trong nhóm vi-rút, **thực khuẩn** là loại xâm nhiễm vào vi khuẩn. Mỗi loại vi khuẩn thường bị nhiễm bởi một hoặc một vài loại thực khuẩn đặc trưng. Thực khuẩn đóng vai trò quan trọng trong điều hòa quần thể vi khuẩn tự nhiên và đang được nghiên cứu ứng dụng trong y học. Dựa trên chu kỳ sống bên trong vi khuẩn, thực khuẩn được chia thành hai nhóm chính: **Thực khuẩn thể độc lực** và **thực khuẩn thể ôn hoà**. Với thực khuẩn thể độc lực, sau khi xâm nhập vào tế bào vi khuẩn, thực khuẩn chiếm quyền kiểm soát, nhân bản, và phá hủy tế bào để giải phóng các bản sao mới, quá trình này được gọi là **chu kỳ tan**. Với thực khuẩn thể ôn hoà, chúng tích hợp vật liệu di truyền của mình vào hệ gen của vi khuẩn, nhân bản cùng vật chủ mà không phá hủy tế bào, quá trình này gọi là **chu kỳ tiền tan** và có thể chuyển sang chu kỳ tan khi gặp điều kiện kích hoạt.

Vùng protein bảo tồn trong bộ gen thực khuẩn là các vùng chức năng trong chuỗi axit amin của protein, được duy trì qua quá trình tiến hóa và thường xuất hiện ở nhiều loài thực khuẩn khác nhau. Những protein này thường liên quan đến các chức năng thiết yếu như lắp ráp cấu trúc vi-rút, xâm nhập tế bào chủ, sao chép dữ liệu di truyền và ly giải tế bào vi khuẩn.

### **1.1.2 Động lực thúc đẩy**

Việc phân loại thực khuẩn là có độc lực hay ôn hòa mang đến những lợi ích to lớn cho việc nghiên cứu hệ vi sinh vật và ứng dụng của chúng. Đầu tiên, phân loại thực khuẩn là chìa khóa để khám phá mối quan hệ phức tạp giữa thực khuẩn và vật chủ, từ đó làm sáng tỏ vai trò của thực khuẩn trong cộng đồng vi sinh vật. Tiếp theo, việc nhận diện chính xác thực khuẩn có độc lực là bước đệm quan trọng cho việc phát triển các liệu pháp điều trị bằng thực khuẩn, nhằm tiêu diệt các vi khuẩn gây bệnh, mở ra hướng đi mới cho việc điều trị.

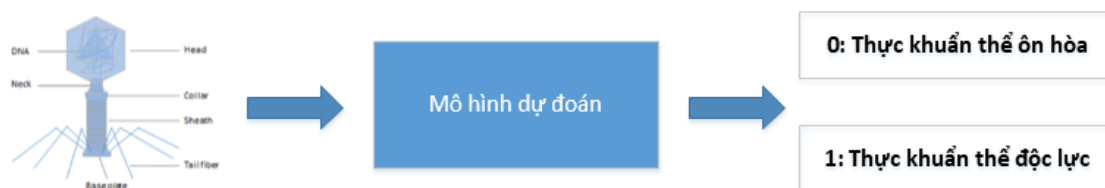
Để phân loại một thực khuẩn mới có thuộc thể độc lực hay ôn hòa, phương pháp truyền thống là nuôi cấy trong phòng thí nghiệm và theo dõi. Các kỹ thuật nuôi cấy này không chỉ tốn thời gian và chi phí mà còn không có khả năng áp dụng được đối với các trình tự thực khuẩn mới được tổng hợp từ môi trường. Do đó, một phương pháp phân loại thực khuẩn dựa trên tính toán thông qua dữ liệu trình tự gen của thực khuẩn là cần thiết.

Với mục tiêu cải tiến phương pháp cũ hay phát triển phương pháp mới, báo cáo này cung cấp một góc nhìn tổng quan về các phương pháp tiêu biểu đã được công bố, giúp tiếp cận bài toán phân loại thực khuẩn dựa trên tính toán một cách khái quát và nhanh chóng nhất.

## **1.2 Phát biểu bài toán**

Trong bối cảnh số lượng lớn thực khuẩn được phát hiện thông qua các công nghệ giải trình tự thế hệ mới, việc phân loại thực khuẩn bằng các phương pháp truyền thống là rất tốn kém, về cả nguồn lực và tài chính. Cùng với sự phát triển của công nghệ tính toán, các phương pháp học máy và học sâu được sử dụng để thực hiện phân loại thực khuẩn.

### 1.2.1 Định nghĩa bài toán



Hình 1.1: Sơ đồ mô tả bài toán phân loại thực khuẩn dựa trên tính toán

Hình 1.1 mô tả bài toán phân loại thực khuẩn dựa trên phương pháp tính toán, trong đó:

- **Dữ liệu đầu vào:** Chuỗi DNA đầy đủ hoặc tập hợp các đoạn DNA ngắn được lấy từ bộ gen của thực khuẩn.
- **Mô hình dự đoán:** Thuật toán học máy, học sâu được huấn luyện.
- **Dữ liệu đầu ra:** Dự đoán thực khuẩn là thể có độc lực hay thể ôn hòa.

### 1.2.2 Đặc điểm của bài toán

- Xét ở khía cạnh học máy, bài toán phân loại thực khuẩn là bài toán phân loại nhị phân.
- Dữ liệu đầu vào có độ dài và chất lượng không đồng đều, bao gồm cả các trình tự ngắn hoặc có chứa nhiễu sinh học.
- Các mô hình cần đảm bảo tính chính xác cao, khả năng khái quát tốt với dữ liệu chưa thấy trong quá trình huấn luyện.
- Phải xử lý được dữ liệu có nguồn gốc từ các họ thực khuẩn khác nhau và có tính đa dạng về mặt di truyền học.

### 1.2.3 Thách thức trong quá trình thực hiện

Với cách tiếp cận dựa trên tính toán, việc phân loại thực khuẩn hiện tại đang gặp phải 4 khó khăn chính:

1. **Thiếu dữ liệu huấn luyện chất lượng cao:** Mặc dù số lượng thực khuẩn trong tự nhiên được ước tính lên đến khoảng  $10^{31}$  [5], nhưng hiện nay chỉ có một lượng rất nhỏ bộ gen thực khuẩn đã được giải trình tự hoàn chỉnh và đưa vào cơ sở dữ liệu. Phần lớn các thực khuẩn này vẫn chưa được xác định đặc điểm sinh học cụ thể, bao gồm cả thông tin về vòng đời (chu kỳ sinh tan hay tiềm tan). Điều này dẫn đến sự thiếu hụt dữ liệu được gán nhãn rõ ràng và chính xác, gây khó khăn cho quá trình huấn luyện các mô hình học máy trong bài toán phân loại vòng đời thực khuẩn.
2. **Không tồn tại chỉ dấu sinh học phổ quát:** Không giống như vi khuẩn có gen 16S rRNA được bảo tồn cao, đóng vai trò như một chỉ dấu phân tử phổ quát trong phân loại và phân tích phát sinh loài, thực khuẩn không sở hữu bất kỳ gen nào có mặt đồng nhất và bảo tồn trên toàn bộ các nhóm thực khuẩn [1]. Do đó, việc phân loại thực khuẩn thường dựa trên nội dung bộ gen hoặc các gen đặc trưng cho từng nhóm nhỏ.
3. **Tính tương đồng di truyền với vật chủ:** Nhiều đoạn gen của phage có thể có trình tự tương tự hoặc thậm chí đồng nhất với gen vi khuẩn, do quá trình tiến hóa đồng hành hoặc sự trao đổi gen thông qua cơ chế di truyền ngang.
4. **Xử lý dữ liệu không đầy đủ:** Trong những nghiên cứu gần đây, dữ liệu được thu thập thông qua metagenomics – phương pháp giải trình tự toàn bộ DNA có trong mẫu môi trường (như đất, nước, ruột sinh vật) mà không cần phải nuôi cấy vi sinh vật trong phòng thí nghiệm. Tuy nhiên, một đặc điểm hạn chế của phương pháp này là chỉ thu được các đoạn DNA ngắn, rời rạc và không hoàn chỉnh, thay vì toàn bộ bộ gen đầy đủ của một loại thực khuẩn. Điều này gây khó khăn vì các đoạn trình tự thu được có thể không chứa đủ thông tin đặc trưng, dễ bị nhiễu hoặc nhầm lẫn với vật liệu di truyền từ các sinh vật khác trong mẫu. Do đó, việc xử lý dữ liệu từ metagenomics đòi hỏi các phương pháp phân tích chuyên biệt để bù đắp cho tính không đầy đủ và phân mảnh của dữ liệu..

Với các đặc điểm nêu trên, bài toán phân loại thực khuẩn đòi hỏi sự kết hợp giữa kiến thức sinh học phân tử và các kỹ thuật trong lĩnh vực học máy và học sâu. Đây chính là trọng tâm của các phương pháp được khảo sát trong báo cáo này.

## 1.3 Nguồn dữ liệu và định dạng dữ liệu

### 1.3.1 Nguồn dữ liệu

Các phương pháp phân loại được đề cập đến trong báo cáo này lấy dữ liệu từ các nguồn sau:

- **NCBI (National Center for Biotechnology Information - <https://www.ncbi.nlm.nih.gov/>):** Là một trong những kho dữ liệu sinh học lớn và toàn diện nhất, chứa thông tin về trình tự DNA, protein, gen, và các chú thích liên quan. NCBI cung cấp cả dữ liệu thô và đã được gán nhãn, là nguồn phổ biến cho việc huấn luyện và đánh giá mô hình phân loại.
- **PhageScope - <https://phagescope.deepomics.org/database>:** Là cơ sở dữ liệu chuyên biệt dành cho thực khuẩn, hỗ trợ phân tích chức năng, xác định vòng đời và các đặc điểm phân tử. PhageScope tích hợp các công cụ tính toán, hỗ trợ xử lý và khám phá dữ liệu thực khuẩn ở mức độ chi tiết.
- **PhageDB - <https://phagesdb.org>:** Là kho dữ liệu tập trung vào thu thập và lưu trữ thông tin thực nghiệm về các chủng thực khuẩn đã được phát hiện. PhageDB thường được sử dụng để đánh giá mô hình phân loại trên các bộ dữ liệu thực tế.

### 1.3.2 Định dạng dữ liệu

Khi xử lý dữ liệu thực khuẩn, hai định dạng tập tin phổ biến là **FASTA** và **GenBank**. Mỗi định dạng cung cấp một cách biểu diễn riêng cho trình tự nucleotide và thông tin liên quan.

#### Định dạng FASTA

Định dạng **FASTA** là định dạng văn bản đơn giản dùng để lưu trữ trình tự nucleotide hoặc amino acid. Tập tin định dạng **FASTA** có phần mở rộng là \*.fasta. Cấu trúc của tập tin **FASTA** bao gồm hai phần:

- **Dòng tiêu đề (header):** Bắt đầu bằng ký tự >, tiếp theo là thông tin mô tả, thường bao gồm ID chuỗi, tên sinh vật, tên gen hoặc nguồn gốc.

- **Trình tự** (sequence): Nằm ở các dòng tiếp theo, chứa các ký tự đại diện cho nucleotide (A, T, G, C) hoặc acid amin.

Ví dụ: file *NC\_000866\_Lytic.fasta* có nội dung như sau:

```
>NC_000866.4 Enterobacteria phage T4, complete genome
AATTTTCCTTATTAGGCCGCAAGGGCCTTCATAGTTTTAGCGATTGGGAACTTCATCA
TCACTTAAAGAGTTGCGATAACCGATGAAGTCGGAAACAATACGGAATTTCTTGGTAAAC
TCAGCAACCATTTTATCACTGTTTTTTGAAGCATTATTTGATAATACATCAAAAAGATTA
GTTACTGTCCAAATGTCATGACCGATGGTATCTTTTCCACCATTAAAATATACACCCTGT
.....
```

## Định dạng GenBank

Định dạng GenBank là định dạng tiêu chuẩn của NCBI, chứa thông tin chi tiết hơn so với FASTA. Tập tin định dạng FASTA có phần mở rộng là \*.gb. Một tập tin GenBank bao gồm ba phần chính:

- **Phần tiêu đề** (Header): Mô tả chung về bộ gen, bao gồm số hiệu truy cập, nguồn sinh vật, độ dài chuỗi và các đặc điểm chung.
- **Phần đặc trưng** (Features): Liệt kê các đặc trưng sinh học của chuỗi, bao gồm vị trí gen, loại protein mã hóa, chú thích chức năng.
- **Phần trình tự** (Sequence): Trình bày toàn bộ chuỗi nucleotide của bộ gen.

Ví dụ dưới đây là một phần của file *AF503408\_Lysogenic.gb*

```
LOCUS      AF503408      101660 bp      DNA      linear      PHG 28-APR-2006
DEFINITION Enterobacteria phage P7, complete genome.
ACCESSION  AF503408
.....
FEATURES             Location/Qualifiers
     source            1..101660
                        /organism="Enterobacteria phage P7"
                        /mol_type="genomic DNA"
```

/db\_xref="taxon:10682"

....

#### ORIGIN

```
1 acattatacg aagttatatt aagggttatt gaacatgac aatttacctg taaatccata
61 cagttcaata ccttatcagg tcaaatagtg atcacttgat catttgatca agtttgcgct
121 acgtaaaatc tgtgaaaagt tggcagtggt agtgctccag atttcgcgta gcgcacttag
181 caccaccaat caatcagagg tgaaaaatgg gatattcagc tgctaaagtg tccactcatc
241 ttgagcttga gaaaaaccgt gggtactggc gggcaaaagg gtttgatcgt gatagttgcc
301 aactgtcatt atcgcgcggt gaagagaaaa tagtacgcac gcgcggtcgc tggcgtttct
```

Định dạng GenBank hỗ trợ các phần mềm sinh học phân tử và hệ thống phân tích chú thích gen trong quá trình tiền xử lý dữ liệu.

# Chương 2

## Các nghiên cứu liên quan

Chương này trình bày tổng quan và phân tích các phương pháp phân loại thực khuẩn dựa trên dữ liệu di truyền. Hai tiêu chí chính được sử dụng để tổ chức và đánh giá các phương pháp bao gồm: (1) phân loại theo loại thuật toán học máy được áp dụng, và (2) sắp xếp theo trình tự thời gian công bố nhằm phản ánh tiến trình phát triển của lĩnh vực.

### 2.1 Phân loại theo phương pháp học máy

Dựa trên kỹ thuật xử lý dữ liệu và loại mô hình học máy sử dụng, các phương pháp phân loại thực khuẩn có thể chia thành hai nhóm chính:

- **Nhóm 1:** Các phương pháp học máy truyền thống, sử dụng các thuật toán như Random Forest (Rừng ngẫu nhiên), SVM (Support Vector Machine - Máy vector hỗ trợ), v.v. với đầu vào là bộ gen thực khuẩn đầy đủ. Báo cáo này thực hiện trên 3 phương pháp của **PHACTS**, **PhageAI**, và **BACPHLIP**.
- **Nhóm 2:** Các phương pháp học sâu, có khả năng xử lý dữ liệu không hoàn chỉnh nên có khả năng tận dụng được lượng dữ liệu lớn từ nguồn dữ liệu metagenomics. Báo cáo này thực hiện trên 3 phương pháp của **DeePhage**, **PhaTYP**, và **DeepPL**.

### 2.2 Sắp xếp theo thứ tự thời gian công bố

Bên cạnh phân loại theo kỹ thuật, các phương pháp còn được trình bày theo thứ tự thời gian công bố để thể hiện xu hướng phát triển qua các giai đoạn và so sánh kết quả của phương pháp mới với phương pháp trước đó. Thời gian công bố của từng phương pháp như sau:

1. **PHACTS** – tháng 01 năm 2012



2. **PhageAI** – tháng 07 năm 2020
3. **BACPHLIP** – tháng 05 năm 2021
4. **DeePhage** – tháng 09 năm 2021
5. **PhaTYP** – tháng 01 năm 2023
6. **DeepPL** – tháng 10 năm 2024

## 2.3 Phương pháp trình bày trong các phần tiếp theo

Để đảm bảo tính nhất quán và thuận tiện cho việc đánh giá, mỗi phương pháp được trình bày một số điểm chính bao gồm:

1. **Mô tả chung** về phương pháp, động lực, phương pháp thực hiện và mục tiêu.
2. **Tập dữ liệu:** Mô tả nguồn dữ liệu, số lượng mẫu, tính chất dữ liệu (hoàn chỉnh hay contig), và phương pháp gán nhãn.
3. **Phương pháp thực hiện:** Thuật toán hoặc mô hình được áp dụng, kiến trúc mạng (nếu có), kỹ thuật xử lý đặc trưng và quy trình huấn luyện.
4. **Kết quả thu được:** Các chỉ số đánh giá như độ chính xác, độ nhạy, độ đặc hiệu, F1-score, v.v..
5. **So sánh với các phương pháp trước:** Ưu điểm nổi bật, cải tiến kỹ thuật, sự khác biệt về khả năng dự đoán.

## 2.4 PHACTS

PHACTS[4] là một trong những công cụ tiên phong trong việc phân loại thực khuẩn thể theo vòng đời dựa trên bộ gen đầy đủ. Công cụ này sử dụng thuật toán rừng ngẫu nhiên để xây dựng mô hình học máy dựa trên mức độ tương đồng protein giữa các thực khuẩn.

### 2.4.1 Dữ liệu huấn luyện

Dữ liệu huấn luyện được sử dụng trong bài báo là dữ liệu thu thập từ cơ sở dữ liệu PHANTOME, gồm 654 bộ gen thực khuẩn. Trong đó, dữ liệu được sử dụng cho huấn luyện gồm 227 thực khuẩn có vòng đời đã được xác nhận bằng phương pháp thủ công từ nhiều nguồn tài liệu khác nhau, cụ thể tập huấn luyện chứa 148 thực khuẩn thể ôn hoà và 79 thực khuẩn thể độc lực. Đặc điểm của tập dữ liệu này là tỷ lệ lớp không cân bằng (2:1), phản ánh đúng về sự phân bố dữ liệu của thực khuẩn thể ôn hoà trong dữ liệu.

Để đảm bảo tính khách quan trong đánh giá mô hình, các thực khuẩn có độ tương đồng cao về protein ( $>90\%$  protein giống nhau với  $>90\%$  độ tương đồng) với thực khuẩn đang được kiểm tra sẽ bị loại khỏi tập huấn luyện. Điều này giúp tránh tình trạng mô hình "học tử" và đánh giá chính xác khả năng tổng quát hóa.

### 2.4.2 Phương pháp

**Tạo tập protein chuẩn** (*query proteins*): Một tập hợp các chuỗi protein truy vấn  $Q = \{P_1, P_2, \dots, P_M\}$  được tạo ra bằng cách chọn ngẫu nhiên  $M$  protein, trong đó  $M$  là số lượng protein do người dùng chỉ định để sử dụng cho việc tạo tập huấn luyện. Từ mỗi lớp,  $M/C$  protein được chọn ngẫu nhiên, thuộc về các thể thực khuẩn của lớp đó, với  $C$  là số lượng lớp trong tập huấn luyện. Trong các thí nghiệm của nhóm tác giả, việc chọn  $M = 600$  cho kết quả tốt nhất được lựa chọn theo kinh nghiệm. Khi giảm giá trị  $M$ , độ chính xác giảm xuống; ngược lại, khi tăng giá trị  $M$ , thời gian chạy tăng lên mà không mang lại sự cải thiện tương ứng về độ chính xác.

**Tạo tập dữ liệu huấn luyện:** Để tạo tập dữ liệu huấn luyện cho thuật toán rừng ngẫu nhiên, một tập hợp gồm  $N$  vectơ tương đồng được xây dựng, trong đó  $N$  là số lượng thể thực khuẩn được sử dụng làm các trường hợp huấn luyện. Từ mỗi lớp,

$N/C$  bộ gen của thực khuẩn được chọn ngẫu nhiên mà không hoàn lại, với  $C$  là số lượng lớp. Từ  $N$  thực khuẩn này, danh sách  $L = \{G_1, G_2, \dots, G_N\}$  được tạo ra. Lớp có số lượng mẫu đại diện ít nhất sẽ giới hạn số lượng trường hợp huấn luyện có thể sử dụng. Từ kinh nghiệm thực nghiệm của nhóm tác giả, giá trị  $N = 100$  cho kết quả tốt nhất. Và với 50 thực khuẩn mỗi lớp là đủ để thu được kết quả chính xác đồng thời đảm bảo tính đa dạng trong việc lấy mẫu ngẫu nhiên.

Đối với mỗi bộ gen trong số  $N$  bộ gen này, một vector tương đồng  $\mathbf{X}$  được xây dựng. Các protein của một thực khuẩn được so sánh với tất cả các protein trong tập  $Q$  bằng chương trình FASTA. Từ đó, vector tương đồng  $X = [S_1, S_2, \dots, S_M]$  được xây dựng, trong đó  $S_i$  là phần trăm độ tương đồng cao nhất giữa bất kỳ protein nào của phage đầu vào và protein chuẩn  $P_i$  trong tập  $Q$ , như mô tả bên dưới:

$$\begin{aligned}\mathbf{X}_1 &= [S_{1,1}, S_{1,2}, \dots, S_{1,M}] \\ \mathbf{X}_2 &= [S_{2,1}, S_{2,2}, \dots, S_{2,M}] \\ &\vdots \\ \mathbf{X}_N &= [S_{N,1}, S_{N,2}, \dots, S_{N,M}]\end{aligned}$$

**Huấn luyện thuật toán rừng ngẫu nhiên:** Để phân loại tập kiểm tra, PHACTS sử dụng thuật toán rừng ngẫu nhiên. Trong bộ phân loại rừng ngẫu nhiên, một tập hợp các cây quyết định được tạo ra. Đối với mỗi cây, kỹ thuật bootstrapping được áp dụng bằng cách chọn  $N$  mẫu có hoàn lại từ tập huấn luyện gồm  $N$  mẫu. Mỗi cây được phát triển bằng cách chọn ngẫu nhiên  $m$  biến tại mỗi nút, trong đó  $m$  bằng căn bậc hai của tổng số biến. Phép tách tốt nhất tại nút đó được xác định từ tập  $m$  biến này, và cây được phát triển tối đa mà không bị cắt tỉa.

Mỗi cây đưa ra một dự đoán về chu trình (độc lực hay ôn hòa), và dự đoán cuối cùng được xác định bằng quy tắc bỏ phiếu đa số từ tất cả các cây trong rừng. Rừng ngẫu nhiên cũng cung cấp thông tin về tỷ lệ bỏ phiếu, được tính bằng phần trăm số lượng cây dự đoán một chu trình cụ thể chia cho tổng số cây. Vì thuật toán rừng ngẫu nhiên không dẫn đến hiện tượng quá khớp, nên có thể tạo ra số lượng lớn cây. Trong các thí nghiệm của nhóm tác giả, 1001 cây được tạo ra để đảm bảo đủ mức bao phủ cho tập huấn luyện.

**Cách đưa ra dự đoán chính xác, ổn định:** Dự đoán thu được từ thuật toán rừng ngẫu nhiên với  $N$  thực khuẩn đã biết, được chọn ngẫu nhiên làm các trường hợp huấn luyện, và  $M$  protein, được chọn ngẫu nhiên để tạo các vector tương đồng. Do việc lựa chọn dữ liệu huấn luyện là ngẫu nhiên, một thực khuẩn chưa biết có thể bị dự đoán là thuộc các chu trình sống khác nhau trong các lần phân loại rừng ngẫu nhiên khác nhau.

Để xử lý tốt hơn sự biến thiên này trong các dự đoán, nhóm tác giả thực hiện 10 lần lặp với các tập hợp thực khuẩn huấn luyện khác nhau và tập hợp protein truy vấn khác nhau. Mười lần lặp được chọn nhằm cân bằng giữa thời gian thực thi và độ chính xác. Các dự đoán dựa trên 5 lần lặp cho độ chính xác thấp hơn, trong khi dự đoán với 20 lần lặp làm tăng đáng kể thời gian thực thi mà không mang lại sự cải thiện tương ứng về độ chính xác.

Điểm xác suất cuối cùng được xem là “đáng tin cậy” nếu có sự đồng thuận từ 10 lần lặp cho một chu trình sống cụ thể. Để xác định liệu một dự đoán có đáng tin cậy hay không, giá trị trung bình và độ lệch chuẩn của 10 lần lặp được tính toán. Dự đoán được xem là đáng tin cậy nếu điểm xác suất trung bình của chu trình sống được dự đoán cao hơn 2 độ lệch chuẩn so với điểm xác suất trung bình của chu trình còn lại.

**Lọc đặc trưng quan trọng:** Để tăng độ chính xác và hiệu quả tính toán, PHACTS sử dụng phương pháp Gini Importance để đánh giá mức độ quan trọng của từng protein trong việc phân loại. Gini Importance đo lường mức độ đóng góp của một protein vào khả năng phân biệt giữa các lớp (ôn hòa và độc lực). Giá trị Gini Importance cao cho thấy protein đó có vai trò quan trọng trong việc phân loại và giá trị Gini Importance thấp cho thấy protein đó ít đóng góp vào việc phân loại. Chỉ các protein có Gini Importance vượt qua một ngưỡng nhất định (trong nghiên cứu là gấp đôi giá trị trung bình) mới được giữ lại để xây dựng vector tương đồng. Việc loại bỏ các protein ít quan trọng giúp giảm nhiễu, tăng tốc độ xử lý và cải thiện độ chính xác của mô hình.

### 2.4.3 Kết quả và đánh giá

Kết quả thực nghiệm của PHACTS:

- PHACTS đạt độ chính xác cao trong việc phân loại thực khuẩn.

- Độ chính xác: đạt 99% (197/199 phage được phân loại chắc chắn).
- Độ nhạy: đạt 88%.
- Đối với bộ gen không đầy đủ (chỉ sử dụng một phần protein), PHACTS vẫn duy trì được độ chính xác tương đối cao: xấp xỉ 90% khi sử dụng khoảng 20 proteins.

Từ các kết quả thực nghiệm cho thấy rằng, ưu điểm của PHACTS là có thể đưa ra dự đoán với độ chính xác chấp nhận được ngay cả khi chỉ có một phần bộ gen của thực khuẩn. Việc sử dụng mô hình Rừng ngẫu nhiên cũng giúp cho PHACTS dễ hiểu và có thể được giải thích một cách rõ ràng. Bên cạnh những ưu điểm, PHACTS cũng tồn tại hạn chế đáng kể. Mô hình học máy của PHACTS được huấn luyện trên một số lượng nhất định các thực khuẩn đã biết. Nếu gặp các thực khuẩn rất khác biệt, khả năng phân loại chính xác của PHACTS sẽ giảm.

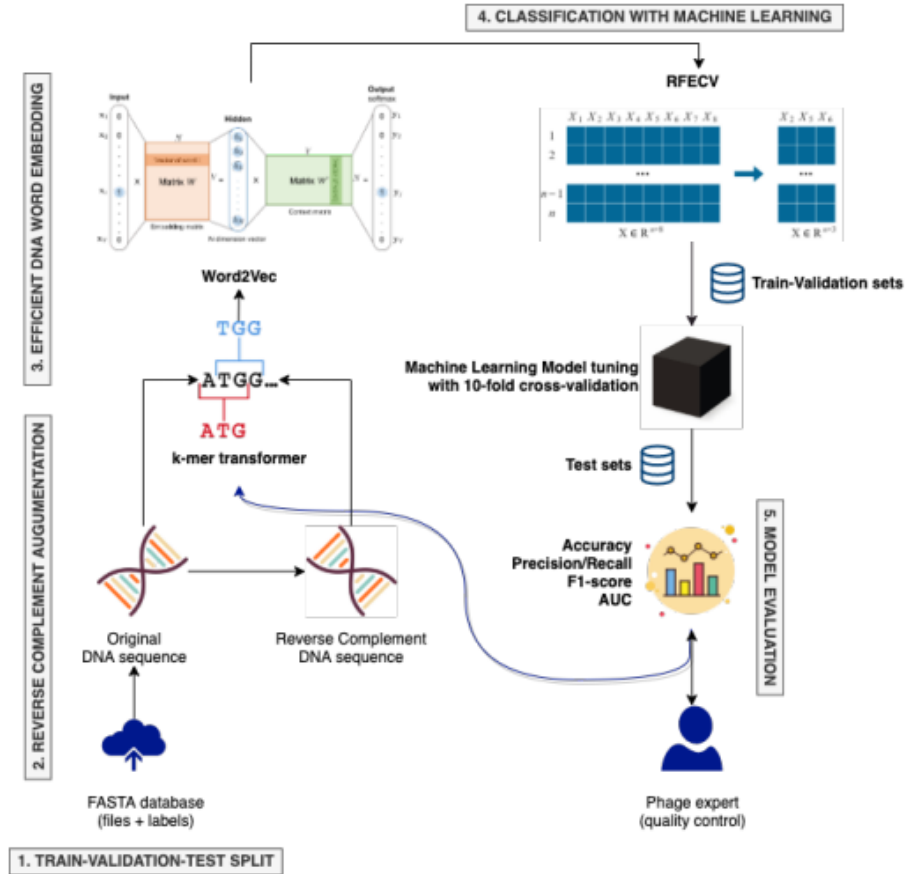
## 2.5 PhageAI

PhageAI[7] sử dụng phương pháp tiếp cận dựa trên học máy và xử lý ngôn ngữ tự nhiên để phân loại thực khuẩn dựa trên trình tự nucleotide. Quy trình giải pháp của PhageAI được mô tả trong Hình 2.1

### 2.5.1 Dữ liệu huấn luyện

Trong nghiên cứu này, dữ liệu được thu thập từ hai cơ sở dữ liệu chuyên biệt về thực khuẩn thể là ACLAME và PhagesDB. Tổng cộng có hơn 600 bộ gen thực khuẩn. Trong đó mỗi mẫu đều được gán nhãn về chu kỳ sống — bao gồm chu kỳ tan tương ứng với thực khuẩn thể độc lực và chu kỳ tiềm tan tương ứng với thực khuẩn thể ôn hoà.

Tập dữ liệu được phân chia thành hai phần chính: (1) Tập huấn luyện Gồm 278 mẫu thực khuẩn thể độc lực và 174 mẫu thực khuẩn thể ôn hoà. (2) Tập kiểm tra gồm 54 mẫu thực khuẩn thể độc lực và 30 mẫu thực khuẩn thể ôn hoà, được lựa chọn từ các họ và loài thực khuẩn thể khác với các mẫu trong tập huấn luyện, nhằm đảm bảo khả năng tổng quát hóa của mô hình.



Hình 2.1: Quy trình giải pháp đề xuất của PhageAI

### 2.5.2 Phương pháp

**Phân chia dữ liệu huấn luyện:** Để kiểm soát và theo dõi quá trình học của mô hình phân loại, PhageAI sử dụng chiến lược phân chia dữ liệu như sau:

- **Tập huấn luyện:** Dữ liệu được chia ngẫu nhiên thành 10 phần bằng nhau. Trong mỗi vòng lặp, 80% dữ liệu được sử dụng để huấn luyện, và 20% còn lại dùng để kiểm tra trong quá trình học. Việc phân tầng được thực hiện dựa trên vòng đời và họ của thực khuẩn, nhằm đảm bảo tính đại diện của các nhóm trong từng phần dữ liệu.
- **Tập thử nghiệm:** Một tập gồm 84 mẫu chưa từng được sử dụng trong quá trình huấn luyện được giữ lại để làm dữ liệu kiểm tra độc lập. Tập này được dùng để đánh giá khách quan hiệu suất của mô hình sau khi quá trình học kết thúc.

- Bộ dữ liệu kiểm tra bên ngoài: Một tập dữ liệu thứ hai gồm 61 mẫu, do công ty Proteon Pharmaceuticals S.A. cung cấp, cũng không được sử dụng trong giai đoạn huấn luyện. Tập này được dùng để ước lượng các chỉ số hiệu suất cuối cùng của mô hình khi áp dụng lên dữ liệu thực tế hoàn toàn mới.

**Tăng cường dữ liệu bằng phương pháp bổ sung chuỗi đảo ngược:** Sau khi phân chia dữ liệu, PhageAI áp dụng kỹ thuật tăng cường dữ liệu bằng cách sử dụng chuỗi bổ sung đảo ngược của các trình tự thực khuẩn như các mẫu bổ sung. Điều này cho phép mô hình học máy tự động học các mối quan hệ phức tạp giữa các trình tự DNA sợi đôi. Kỹ thuật này còn giúp tăng gấp đôi kích thước tập dữ liệu, từ đó cải thiện hiệu suất của mô hình.

**Phương pháp vector hóa:** Trình tự gen của thực khuẩn ở tệp tin dạng FASTA thường là các chuỗi tương đối dài (từ 5.000 đến 300.000 bp) bao gồm các nucleotide {A, C, G, T}. PhageAI đã áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên phổ biến để xây dựng không gian vector đại diện cho các trình tự thực khuẩn và giảm đáng kể yêu cầu bộ nhớ nhằm tăng tốc quá trình phân loại. Cụ thể, nhóm nghiên cứu đã sử dụng phương pháp biểu diễn phân tán của các thành phần k-mer chồng lấn và nhúng từ.

Để có được các vector đặc trưng có kích thước cố định đại diện cho các gen, nhóm nghiên cứu đã áp dụng phương pháp nhúng từ dựa trên Word2Vec với mô hình Skip-gram. Cuối cùng, DNA của thực khuẩn được biểu diễn bằng trung bình của các vector nhúng k-mer của các từ cấu thành trình tự, có nghĩa là mỗi gen được mô tả bằng các giá trị số trung bình trong không gian vector.

**Phân loại với Học máy:** Trong nghiên cứu này, các đặc trưng không đồng nhất được trích xuất từ trung bình của các vector nhúng k-mer có thể phản ánh thông tin mẫu tốt hơn. Vì mục đích này, nhóm nghiên cứu đã áp dụng RFECV (Recursive feature elimination with cross-validation), một phương pháp lựa chọn tính năng hiệu quả để loại bỏ các thuộc tính không liên quan và tăng khả năng tổng quát hóa của mô hình ở bước tiếp theo. Thông qua quá trình này, 150 đặc trưng quan trọng đã được chọn từ tổng số 300 đặc trưng để sử dụng trong quá trình phân loại. Trong phạm vi bài báo này, nhóm tác giả đã thực hiện huấn luyện và so sánh kết quả từ 11 thuật toán học máy có giám sát:

- Mô hình Bayesian: MultinomialNB - Multinomial Naive Bayes

- Máy vector hỗ trợ: SVC - Support Vector Classification, SGDClassifier - Stochastic Gradient Descent Classifier
- Mô hình tuyến tính: Logistic Regression
- Mạng nơ-ron: MLPClassifier - Multilayer Perceptron Classifier
- Cây quyết định: Random Forest Classifier
- Thuật toán dựa trên tương đồng: K-Neighbors Classifier
- Gradient boosting: Gradient Boosting Classifier, XGBoost - Extreme Gradient Boosting, CatBoostClassifier - Categorical Boosting Classifier, LightGBM - Light Gradient Boosting Machine

Để điều chỉnh siêu tham số của mô hình, thay vì sử dụng các kỹ thuật như Grid Search và Randomized Search - vốn tìm kiếm qua toàn bộ không gian các kết hợp tham số có sẵn theo cách biệt lập mà không cải thiện dựa trên các kết quả trước đó - nhóm nghiên cứu đã áp dụng phương pháp tối ưu hóa Bayesian, giúp giảm thiểu thời gian cần thiết để có được một tập hợp tham số mô hình tối ưu.

### 2.5.3 Kết quả và đánh giá

Bảng 2.1: Bacteriophages life cycle prediction benchmark for 11 supervised ML tuned classifiers with 10-fold cross-validation

Classifier	Accuracy (%)	AUC (%)	Precision	Recall	F1-score
MultinomialNB	83.53	80.79	0.85	0.84	0.82
SGDClassifier	87.06	89.01	0.88	0.87	0.87
MLPClassifier	92.35	94.40	0.93	0.92	0.92
LogisticRegression	92.94	94.53	0.93	0.93	0.93
RandomForestClassifier	92.35	97.42	0.92	0.92	0.92
Support Vector Classification	<b>98.90</b>	<b>99.84</b>	<b>0.99</b>	<b>0.99</b>	<b>0.99</b>
KNeighborsClassifier	95.44	96.18	0.94	0.95	0.95
GradientBoostingClassifier	87.06	88.36	0.87	0.87	0.87
XGBoost	97.80	98.90	0.98	0.98	0.98
CatBoostClassifier	91.18	96.03	0.91	0.91	0.91
LightGBM	96.90	97.68	0.97	0.97	0.97



Kết quả tốt nhất đạt được với bộ phân loại Support Vector Machine với kernel tuyến tính, cho độ chính xác trung bình là 98,90% trên các tập đánh giá. Bên cạnh đó, thuật toán còn đạt độ chính xác trên tập kiểm thử là 97.18%. Và để kiểm nghiệm khả năng thực sự của mô hình, nhóm tác giả thực nghiệm trên bộ dữ liệu riêng của Proteon Pharmaceuticals, và cho kết quả dự đoán chính xác toàn bộ 61 thực khuẩn trong tập dữ liệu.

Từ kết quả thực nghiệm có thể nhận thấy, ưu điểm của PhageAI đó là sử dụng các mô hình học máy hiện đại (như SVM, XGBoost) với kỹ thuật vector hóa trình tự DNA bằng xử lý ngôn ngữ tự nhiên (word2vec, k-mer embeddings) cho độ chính xác cao và có thể dự đoán trực tiếp từ trình tự DNA của thực khuẩn. Nhưng, PhageAI cũng tồn tại những nhược điểm như: khó giải thích sinh học điều này gây hạn chế việc hiểu rõ cơ chế tiềm ẩn phía sau; các mô hình học máy trên yêu cầu gán nhãn dữ liệu chính xác và đa dạng để đạt được độ chính xác cao, nếu dữ liệu bị bias sẽ ảnh hưởng xấu tới khả năng của mô hình.

## 2.6 BACPHLIP

BACPHLIP [2] là công cụ phân loại thực khuẩn được phát triển vào năm 2021, sử dụng mô hình học máy **Rừng ngẫu nhiên (Random Forest)** và tập trung vào việc khai thác các **protein domain bảo tồn** trong bộ gen thực khuẩn.

Protein domain bảo tồn trong bộ gen thực khuẩn là các vùng chức năng trong chuỗi axit amin của protein, được duy trì qua quá trình tiến hóa và thường xuất hiện ở nhiều loài thực khuẩn khác nhau. Những domain này thường liên quan đến các chức năng thiết yếu như lắp ráp cấu trúc virus, xâm nhập tế bào chủ, sao chép DNA và ly giải tế bào vi khuẩn. Những domain này cũng được đánh giá là có ảnh hưởng lớn đến vòng đời của thực khuẩn thể, vì chúng có thể quyết định cách thức mà thực khuẩn thể tương tác với tế bào chủ và môi trường xung quanh.

### Dữ liệu sử dụng

BACPHLIP sử dụng bộ dữ liệu bao gồm 1.057 bộ gen thực khuẩn được Mavrich và Hatfull thu thập năm 2017. Bộ dữ liệu được chia thành 2 tập nhỏ với tỉ lệ 60:40 để làm tập dữ liệu huấn luyện và tập dữ liệu kiểm thử.

- **Tập huấn luyện:** 634 bộ gen thực khuẩn thể đã được gán nhãn.
- **Tập kiểm thử độc lập:** 423 bộ gen thực khuẩn thể khác, không trùng lặp với tập huấn luyện. Trong đó có: 240 thực khuẩn thể ôn hoà và 183 thực khuẩn thể độc lực

## Phương pháp và kỹ thuật chính

BACPHLIP xây dựng mô hình theo các bước chính sau:

1. **Xác định domain protein:** Sử dụng công cụ **HMMER3** để phát hiện các *protein domain* có mặt trong bộ gen thực khuẩn, dựa trên các mô hình Markov ẩn (Hidden Markov Models – HMMs).
2. **Tạo vector đặc trưng nhị phân:** Với mỗi domain, nếu nó xuất hiện trong bộ gen thực khuẩn thì gán giá trị 1, nếu không thì gán 0. Kết quả là một vector nhị phân đại diện cho mỗi bộ gen thực khuẩn.
3. **Huấn luyện mô hình:** Sử dụng thuật toán **Random Forest Classifier** để phân loại vòng đời dựa trên vector đặc trưng nhị phân.

## Kết quả thực nghiệm

Bảng 2.2: So sánh kết quả của BACPHLIP với các Mavrich và PHACTS.

	<b>BACPHLIP</b>	<b>Mavrich</b>	<b>PHACTS</b>
Accuracy	<b>0.983</b>	0.955	0.790
Balanced accuracy	<b>0.970</b>	0.917	0.528
MCC	<b>0.967</b>	0.911	0.586
F1-score	<b>0.985</b>	0.939	0.837

Mô hình **BACPHLIP** cho thấy hiệu năng vượt trội so với **PHACTS**, cả về độ chính xác và các chỉ số đánh giá khác. Một trong những nguyên nhân chính là BACPHLIP được huấn luyện trên tập dữ liệu lớn hơn đáng kể, với tổng cộng **1057** bộ gen thực khuẩn thể có nhãn rõ ràng về vòng đời, so với chỉ **277** mẫu được sử dụng trong PHACTS. Việc mở rộng quy mô dữ liệu huấn luyện giúp mô hình học được đặc trưng đa dạng và tổng quát hơn của các loại thực khuẩn thể.

Kết quả đánh giá trên tập kiểm thử cho thấy BACPHLIP đạt độ chính xác lên tới **98.3%**, cao hơn nhiều so với **79.0%** của PHACTS. Các chỉ số khác cũng phản ánh sự vượt trội của BACPHLIP: *balanced accuracy* đạt **97.0%** so với **52.8%**, hệ số tương quan Matthews (*MCC*) là **96.7%** so với **58.6%**, và *F1-score* đạt **98.5%** trong khi PHACTS chỉ đạt **83.7%**.

## 2.7 DeePhage

Các phương pháp và công cụ trước đây trong bài toán phân loại thực khuẩn thường sử dụng các mô hình học máy truyền thống và yêu cầu dữ liệu DNA đầy đủ của thực khuẩn. Tuy nhiên, trong thực tế, nguồn dữ liệu đầy đủ như vậy rất hạn chế, do nhiều loài thực khuẩn chưa được nuôi cấy và giải trình tự toàn bộ hệ gen.

Trong bối cảnh đó, phương pháp metagenomics đã mở ra hướng tiếp cận mới bằng cách thu thập trực tiếp các đoạn DNA từ mẫu môi trường tự nhiên mà không cần nuôi cấy. Nhờ áp dụng các kỹ thuật giải trình tự thế hệ mới, metagenomics có thể tạo ra khối lượng dữ liệu lớn với thời gian ngắn. Dữ liệu metagenimics bao gồm hàng triệu đến hàng tỷ đoạn trình tự ngắn, cho phép khai thác thông tin phong phú để phục vụ bài toán phân loại thực khuẩn. Tuy nhiên, loại dữ liệu này chứa các đoạn DNA của nhiều loài khác nhau trong cùng quần thể. Do đó, dữ liệu metagenomics gây khó khăn trong việc lắp ráp hệ gen hoàn chỉnh và tỷ lệ lớn các đoạn gen không thể gán chức năng do thiếu thông tin đối chiếu trong cơ sở dữ liệu hiện có.

Nhằm tận dụng hiệu quả dữ liệu metagenomics mà không phải phụ thuộc vào bộ dữ liệu DNA đầy đủ, công cụ DeePhage [8] đã được phát triển dựa trên kiến trúc mạng nơ-ron tích chập (Convolutional Neural Network – CNN), với thiết kế cho phép xử lý các đoạn DNA ngắn thu được từ dữ liệu metagenomics. Đây là một cách tiếp cận mới trong việc mở rộng khả năng phân loại thực khuẩn từ các nguồn dữ liệu chưa hoàn chỉnh.

### Dữ liệu sử dụng

Dữ liệu được sử dụng trong nghiên cứu bao gồm hai tập chính.

- **Tập dữ liệu được đề cập đến trong công cụ PHACTS:** 77 thực khuẩn

thể độc lực và 148 thực khuẩn thể ôn hoà.

- **Tập dữ liệu NCBI:** 1211 thực khuẩn thể độc lực và 429 thực khuẩn thể ôn hoà.

Các dữ liệu này được mô phỏng lại bằng công cụ **MetaSim** để tạo thành các đoạn contig có độ dài khác nhau:

- **Nhóm A:** 100–400 bp
- **Nhóm B:** 400–800 bp
- **Nhóm C:** 800–1200 bp
- **Nhóm D:** 1200–1800 bp

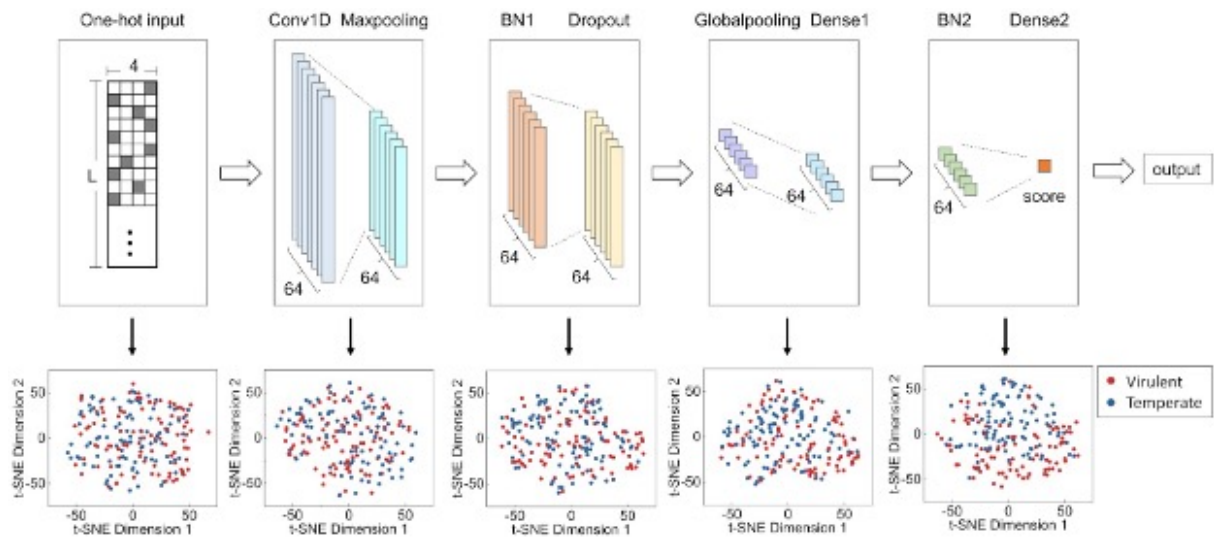
## Phương pháp và kỹ thuật chính

1. **Mã hóa one-hot:** Trình tự DNA được mã hóa thành vector nhị phân:

$$\bullet A = [0, 0, 0, 1], C = [0, 0, 1, 0], G = [0, 1, 0, 0], T = [1, 0, 0, 0]$$

2. **Mạng CNN:** Kiến trúc mạng nơ-ron tích chập (CNN) được thiết kế nhằm trích xuất các đặc trưng cục bộ từ chuỗi mã hoá one-hot. Mạng bao gồm các lớp tích chập, lớp gộp (pooling) và lớp kết nối đầy đủ (fully connected).

3. **Huấn luyện và đánh giá:** Mạng được huấn luyện với mục tiêu phân loại nhị phân (độc lực hoặc ôn hoà), sử dụng các chỉ số đánh giá tiêu chuẩn như *accuracy*, *precision*, và *recall*.



Hình 2.2: Cấu trúc mạng nơ-ron học sâu và hình ảnh hóa 5 lớp bằng cách giảm kích thước của DeePhage

## Kết quả thực nghiệm

Dựa trên bảng kết quả kiểm thử chéo 5 lần, DeePhage vượt trội so với PHACTS trên cả ba tiêu chí đánh giá Độ nhạy - Sensitivity, Độ đặc hiệu - Specificity và Độ chính xác - Accuracy ở tất cả các nhóm độ dài contig (từ 100–400 bp đến 1,200–1,800 bp).

Cụ thể:

- Độ nhạy: DeePhage đạt từ 77.3% đến 87.5%, trong khi PHACTS chỉ đạt từ 64.7% đến 73.7%.
- Độ đặc hiệu: DeePhage đạt từ 74.6% đến 89.5%, trong khi PHACTS chỉ đạt từ 26.3% đến 42.3%.
- Độ chính xác: DeePhage đạt độ chính xác cao từ 76.2% đến 88.9%, trong khi PHACTS dao động trong khoảng 48.6% đến 54.8%

Kết quả này cũng khẳng định ưu thế vượt trội của mô hình học sâu so với các mô hình học máy truyền thống trong việc giải quyết cùng nhiệm vụ phân loại thực khuẩn.

Ngoài ra, khi so sánh về hiệu năng, DeePhage xử lý 100 trình tự DNA chỉ mất 10s, nhanh hơn PHACTS 810 lần (135 phút)

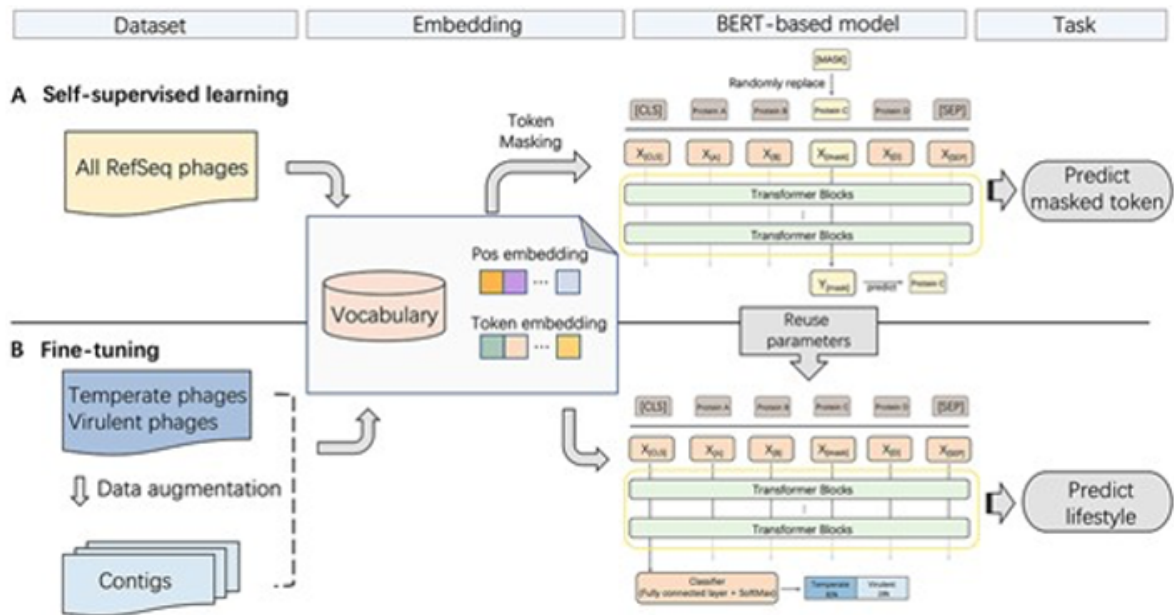
## 2.8 PhaTYP

PhaTYP[6] là một mô hình học sâu thực hiện nhiệm vụ phân loại thực khuẩn thể trên dữ liệu metagenomics, tương tự như DeePhage. Tuy nhiên, thay vì sử dụng mạng nơ-ron tích chập (CNN) như DeePhage, PhaTYP áp dụng kiến trúc **BERT** (*Bidirectional Encoder Representations from Transformers*).

### Chiến lược huấn luyện

PhaTYP được huấn luyện qua hai nhiệm vụ: Học tự giám sát và Tinh chỉnh.

- **Nhiệm vụ học tự giám sát (Self-supervised learning)**: Mô hình học biểu diễn chuỗi DNA bằng cách dự đoán các đoạn bị che khuất tương tự như BERT trong xử lý ngôn ngữ tự nhiên.
- **Nhiệm vụ tinh chỉnh (Fine-tuning)**: Sau khi học biểu diễn DNA, mô hình được tinh chỉnh để phân loại thực khuẩn.



Hình 2.3: Kiến trúc mô hình PhaTYP sử dụng BERT

## Tập dữ liệu sử dụng

**Giai đoạn học tự giám sát:** dữ liệu được lấy từ cơ sở dữ liệu **NCBI RefSeq 2022**, bao gồm tổng cộng 3474 bộ gen của thực khuẩn. Mỗi bộ gen được cắt thành các đoạn có độ dài khác nhau: 5, 10, 15 và 20 kilobase pairs (kbp). Đối với mỗi bộ gen, 10 đoạn contig ngẫu nhiên được tạo ra, dẫn đến tổng cộng **142,434** đoạn contig được thu thập để phục vụ quá trình huấn luyện mô hình.

**Giai đoạn phân loại:** dữ liệu được lấy từ cùng nguồn **NCBI RefSeq 2022** như ở giai đoạn học tự giám sát. Tập dữ liệu bao gồm 1290 thực khuẩn thể độc lực (*virulent*) và 577 thực khuẩn thể ôn hoà (*temperate*). Từ mỗi loại, 10.000 contig được tạo ngẫu nhiên với độ dài nằm trong khoảng từ 100 base pairs (bp) đến 20 kilobase pairs (kbp). Tổng cộng, tập dữ liệu huấn luyện cho giai đoạn này gồm **160,000 contig**, được xây dựng sao cho cân bằng giữa hai lớp.

## Phương pháp và kỹ thuật chính

PhaTYP sử dụng biểu diễn k-mer và kiến trúc BERT để học đặc trưng ngữ nghĩa của DNA, sau đó tinh chỉnh cho nhiệm vụ phân loại vòng đời thực khuẩn thể.

Cụ thể:

- Sử dụng biểu diễn DNA dưới dạng chuỗi k-mer làm đầu vào (tương tự token trong NLP).
- Áp dụng kiến trúc Transformer của BERT để học biểu diễn ngữ nghĩa của DNA.
- Tinh chỉnh lớp đầu ra cho nhiệm vụ phân loại thực khuẩn.

## Kết quả thực nghiệm

Dựa trên kết quả so sánh hiệu suất trên tập dữ liệu kiểm tra có mức độ tương đồng thấp, có thể nhận thấy rằng **PhaTYP vượt trội rõ rệt so với cả DeePhage và PHACTS** ở cả ba tiêu chí: độ nhạy (Sensitivity), độ đặc hiệu (Specificity) và độ chính xác (Accuracy).

Bảng 2.3: So sánh hiệu suất của PhaTYP với các công cụ khác.

Công cụ	Sensitivity	Specificity	Accuracy
PhaTYP	<b>0.99</b>	<b>0.89</b>	<b>0.94</b>
PhaTYP (without SSL)	0.98	0.86	0.92
DeePhage	0.96	0.86	0.91
BACPHLIP	0.98	0.84	0.90
PHACTS	0.90	0.69	0.74
PhagePred	0.57	0.83	0.67

So với DeePhage, PhaTYP đạt độ nhạy cao hơn (0.99 so với 0.96), độ đặc hiệu cao hơn (0.89 so với 0.86), và tổng độ chính xác cũng cao hơn (0.94 so với 0.91). Mặc dù mức chênh lệch không quá lớn, kết quả này cho thấy việc áp dụng học tự giám sát (*Self-Supervised Learning – SSL*) và kiến trúc BERT đã giúp PhaTYP học được các đặc trưng ngữ nghĩa hiệu quả hơn từ dữ liệu DNA, đặc biệt trong các mẫu khó và ít tương đồng.

So với PHACTS là công cụ sử dụng mô hình học máy truyền thống, PhaTYP tạo ra kết quả khác biệt lớn. PHACTS chỉ đạt 0.90 về độ nhạy, 0.69 về độ đặc hiệu và 0.74 về độ chính xác – thấp hơn nhiều so với PhaTYP ở mọi chỉ số.

Những điều này cho thấy khả năng vượt trội của các mô hình học sâu khi thực hiện phân loại trên dữ liệu metagenomics. Từ đó, PhaTYP khẳng định khả năng ứng dụng hiệu quả trong thực tế với nhiệm vụ phân loại thực khuẩn.

## 2.9 DeepPL

Tương tự PhaTYP, DeepPL[9] là một mô hình phân loại vòng đời thực khuẩn thể được phát triển dựa trên kiến trúc **DNABERT** – phiên bản thích ứng của BERT cho dữ liệu chuỗi DNA. Điểm khác biệt của DeepPL là tập trung khai thác thông tin từ các đoạn gen liên quan đến chu kỳ tiềm tan, vốn chỉ có ở phage ôn hoà, nhằm cải thiện độ chính xác trong phân loại.

### Dữ liệu sử dụng

### Tập huấn luyện:



- 1262 bộ gen thực khuẩn thể độc lực
- 557 bộ gen thực khuẩn thể ôn hoà

### **Tập kiểm thử:**

- 245 bộ gen thực khuẩn thể độc lực
- 129 bộ gen thực khuẩn thể ôn hoà

**Tiền xử lí dữ liệu:** Do chỉ có các thực khuẩn thể ôn hoà mới có các gen kích hoạt/duy trình chu kỳ tiềm tan, nhóm tác giả thực hiện trích xuất những đoạn gen có chức năng này. Tiếp theo, thực hiện thao tác chuẩn hoá dữ liệu:

- Loại bỏ các đoạn gen nếu có nhiều hơn 10 ký tự không thuộc loại A, T, G, C.
- Với các đoạn có không quá 10 ký tự không hợp lệ, thay thế ngẫu nhiên thành A, T, G hoặc C.

### **Chiến lược tạo mẫu**

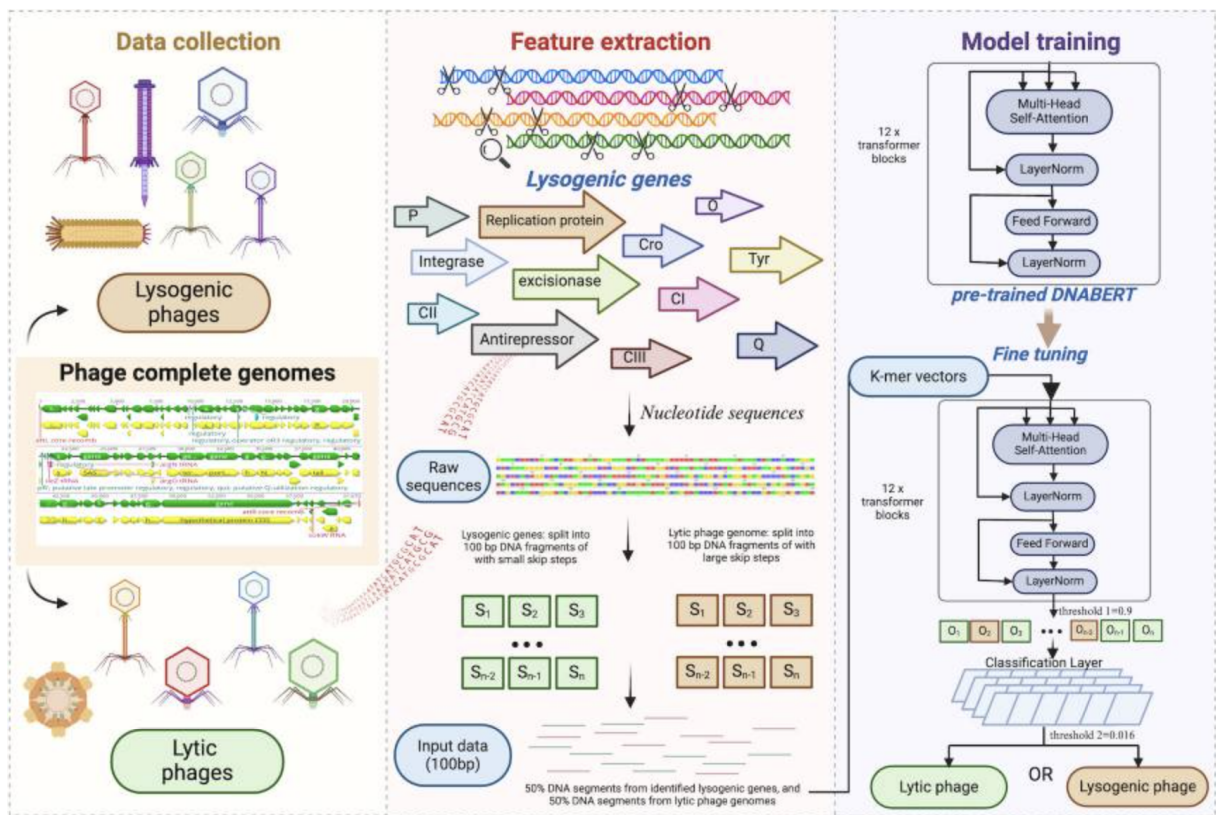
Mục tiêu của giai đoạn này, nhóm tác giả muốn tạo mẫu cân bằng giữa hai nhãn. Trong khi đặc điểm của các mẫu gen của thực khuẩn thể ôn hoà thường rất ngắn, còn mẫu gen của thực khuẩn thể độc lực thường dài hơn. Do đó, nhóm tác giả dùng kĩ thuật của sỏ trượt với kích thước cửa sổ là 100bp nhưng với các bước trượt khác nhau:

- Với mẫu gen của thực khuẩn thể ôn hoà: bước trượt chỉ 1 bp để tạo ra nhiều mẫu hơn.
- Với mẫu gen của thực khuẩn thể ôn hoà: bước trượt 91 bp để tạo ra số mẫu tương đương.

Kết quả thu được 547,810 chuỗi từ gen thực khuẩn thể độc lực và 500,765 chuỗi từ gen thực khuẩn thể ôn hoà. Gần đạt được tỉ lệ cân bằng (50:50) đối với hai nhãn.

## Phương pháp và kiến trúc mô hình

- **Tiền xử lý:** Chuỗi DNA được chia thành các đoạn *6-mer*, tạo thành chuỗi tương tự văn bản.
- **Biểu diễn:** Sử dụng mô hình huấn luyện trước DNABERT để ánh xạ các 6-mer sang không gian vector.
- **Tinh chỉnh:** Tinh chỉnh DNABERT trên tập dữ liệu đã gán nhãn để thực hiện phân loại thực khuẩn.



Hình 2.4: Kiến trúc mô hình DeepPL sử dụng NDABERT

## Kết quả thực nghiệm

DeepPL cho thấy hiệu suất cao và ổn định trong bài toán phân loại thực khuẩn, với độ nhạy 92.24%, độ đặc hiệu 95.91%, độ chính xác 94.65%, F-score đạt 0.92 và hệ số tương quan Matthews (MCC) là 0.53.

So với PhaTYP, DeepPL không tạo được sự khác biệt lớn. DeepPL đạt độ nhạy cao hơn (92.24% so với 90.44%). Nhưng so sánh độ chính xác tổng thể, DeepPL

Bảng 2.4: So sánh hiệu suất giữa DeepPL và các công cụ khác.

Công cụ	Sensitivity	Specificity	Accuracy	F1-score	MCC
DeepPL	<b>92.24</b>	95.91	94.65	0.92	0.53
PhaTYP	90.44	97.47	<b>94.91</b>	0.92	0.53
DeePhage	78.61	<b>98.13</b>	89.83	0.86	0.49
PHACTS	38.94	79.77	48.66	0.53	0.14
PhageAI	83.33	96.08	91.17	0.87	0.50

(94.65%) thấp hơn so với PhaTYP (94.91%).

# Chương 3

## Thực nghiệm

Để có cái nhìn tổng quan về các phương pháp đã đề cập đến ở chương trước, nhóm báo cáo đã tiến hành tổng hợp bộ dữ liệu được sử dụng của mỗi phương pháp và tổng hợp kết quả so sánh được đề cập trong mỗi phương pháp. Chương này trình bày bảng tổng hợp bộ dữ liệu, xây dựng bộ dữ liệu chung, sử dụng bộ dữ liệu chung để tiến hành các thực nghiệm trên 3 phương pháp là PHACTS, Phage AI và DeePhage. Cuối cùng, nhóm báo cáo sẽ trình bày kết quả thực nghiệm và phân tích kết quả.

### 3.1 Xây dựng bộ dữ liệu

Do các phương pháp sử dụng nhiều bộ dữ liệu khác nhau, việc so sánh trực tiếp hiệu suất phân loại giữa chúng là bất khả thi. Vì vậy, nhóm sinh viên thực hiện xây dựng một bộ dữ liệu chung để sử dụng cho các thực nghiệm của mình.

#### 3.1.1 Xử lý nhãn

Bộ dữ liệu mà nhóm sinh viên xây dựng là sự kết hợp của 2 bộ được sử dụng trong bài báo DeepPL và DeePhage. Nhãn  $y$  của bản ghi  $X$  được nhóm báo cáo xử lý như sau:

1. Nếu  $X \in DeePhage \Rightarrow y = y_{DeePhage}$
2. Nếu  $X \in DeepPL \Rightarrow y = y_{DeepPL}$
3. Nếu  $X \in DeePhage \cap DeepPL \Rightarrow y = y_{DeePhage}$

Sau khi kết hợp, bộ dữ liệu thu được bao gồm: 2241 bộ dữ liệu di truyền thực khuẩn, trong đó có 1534 bản ghi được gán nhãn là 1 (thực khuẩn thể độc lực) và 707 bản ghi được gán nhãn 0 (thực khuẩn thể ôn hòa). Sau đó, tập được chia nhỏ thành 2 tập: huấn luyện và kiểm thử. Trong đó, tập huấn luyện gồm 1733 bản ghi

với số lượng nhãn 1 và 0 lần lượt là 1184 và 549. Tập kiểm thử có số lượng nhãn 1 và 0 lần lượt là 296 và 138.

## 3.2 Kịch bản thực nghiệm

Nhóm sinh viên thực hiện mô phỏng lại 3 phương pháp: PHACTS, PhageAI và DeePhage trên bộ dữ liệu chung. Ngoài ra, nhóm có thử nghiệm phân loại bộ dữ liệu với mô hình XGBoost và so sánh mô hình này với phương pháp DeePhage.

## 3.3 Các chỉ số đánh giá

Để đánh giá 1 cách toàn diện hiệu suất phân loại của mô hình chứ không chỉ tập trung vào nhãn 1, nhóm báo cáo sử dụng các chỉ số sau:

- Accuracy: sử dụng để đo lường hiệu suất phân loại chung của mô hình trên 2 nhãn.
- Sensitivity: sử dụng để đo lường độ phủ của mô hình trên nhãn 1.
- Specificity: sử dụng để đo lường độ phủ của mô hình trên nhãn 0.

## 3.4 Kết quả

### 3.4.1 Thực nghiệm với mô hình PHACTS

Dựa trên mã nguồn mà nhóm tác giả công bố, nhóm đã thực hiện đánh giá phương pháp PHACTS trên bộ dữ liệu của mình. Mã nguồn được công bố được thiết kế với đầu vào là dữ liệu amino axit nên nhóm cần thực hiện chuyển đổi dữ liệu di truyền sang dạng này. Để có kết quả khách quan nhất, kết quả thực nghiệm là trung bình cộng của 10 lần lặp lại.

Chỉ số	Giá trị	Độ lệch chuẩn
Accuracy	71.24%	1.67%
Sensitivity	82.77%	2.25%
Specificity	46.52%	3.22%

Bảng 3.1: Kết quả thực nghiệm với mô hình PHACTS

Nhóm báo cáo nhận thấy rằng, với bộ dữ liệu dùng chung ở trên, mô hình PHACTS cho kết quả Accuracy là 71.24%, Sensitivity là 82.77% và Specificity là 46.52%. Kết quả này thấp hơn nhiều so với kết quả công bố của nhóm tác giả PHACTS (đã đề cập ở phần trước).

### 3.4.2 Kết quả thực nghiệm với PhageAI

Trong thí nghiệm này, chúng tôi thực hiện thử nghiệm các thuật toán học máy trong bài báo PhageAI với bộ dữ liệu mới. Các mô hình đều sử dụng BayesianSearch để tối ưu lựa chọn các bộ tham số tốt nhất. Riêng với thuật toán SVM, chúng tôi sử dụng bộ tham số đã được tác giả công bố trong bài báo.

Model	Accuracy	Sensitivity (Class 1)	Specificity (Class 0)
GaussianNB	0.649	0.780	0.365
SGDClassifier	0.864	0.861	0.869
MLPClassifier	0.901	0.946	0.803
RandomForestClassifier	0.928	<b>0.963</b>	0.854
SVC	0.901	0.882	<b>0.942</b>
KNeighborsClassifier	0.910	0.909	0.912
GradientBoostingClassifier	0.930	0.951	0.882
XGBoost	<b>0.936</b>	0.959	0.883
LightGBM	0.933	0.951	0.878
CatBoost	0.931	0.953	0.870

Bảng 3.2: Kết quả so sánh các thuật toán học máy theo Accuracy, Sensitivity và Specificity trên bộ dữ liệu kết hợp.

Dựa trên kết quả Bảng 3.2, có thể thấy rằng các mô hình thuộc nhóm ensemble learning, đặc biệt là các thuật toán Boosting như XGBoost, LightGBM, CatBoost và GradientBoostingClassifier, đều cho hiệu suất rất cao trên cả ba chỉ số: Accuracy, Sensitivity và Specificity. Trong đó, XGBoost đạt Accuracy cao nhất (0.936), đồng thời duy trì mức Sensitivity (0.959) và Specificity (0.883) rất cân bằng, cho thấy khả năng phân biệt tốt giữa hai lớp.

RandomForestClassifier cũng thể hiện hiệu quả cao với Accuracy 0.928, tuy nhiên có phần thiên lệch hơn về Sensitivity (0.963) so với Specificity (0.854), nghĩa là mô hình có xu hướng nhạy hơn với lớp dương (Class 1).

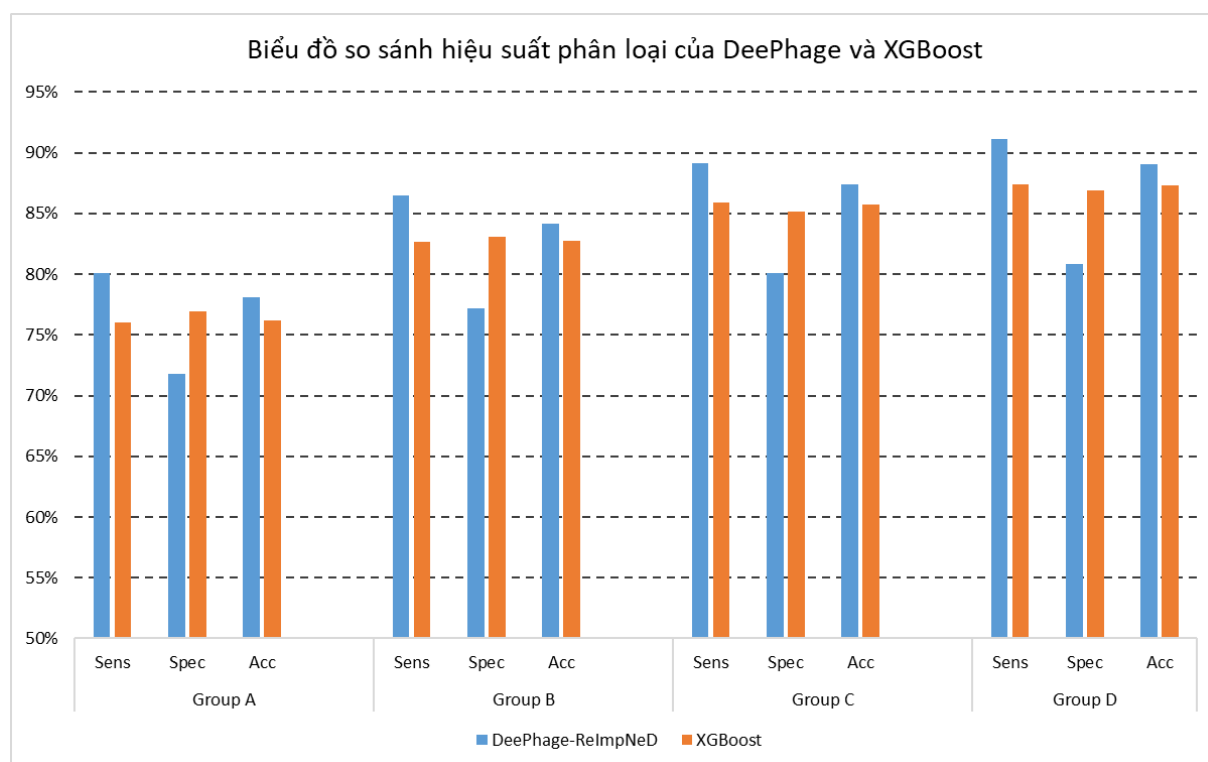
Ở nhóm mô hình cơ bản hơn như GaussianNB và SGDClassifier, kết quả không

thực sự khả quan. Đặc biệt, GaussianNB cho Specificity chỉ ở mức 0.365, cho thấy mô hình này khó phân biệt được lớp âm (Class 0), mặc dù có Sensitivity tương đối tốt (0.780). Điều này thường xảy ra khi dữ liệu không phù hợp với giả định phân phối chuẩn của Naive Bayes.

MLPClassifier và SVC đều có Accuracy trên 0.9, nhưng mỗi mô hình có thiên hướng riêng: MLPClassifier nghiêng về phát hiện Class 1 (Sensitivity cao 0.946), trong khi SVC lại rất mạnh trong việc nhận diện Class 0 (Specificity 0.942).

Tổng thể, như vậy khác với tập dữ liệu của bài báo với SVC cho kết quả tốt nhất thì trong thực nghiệm này, các mô hình Boosting như XGBoost và GradientBoostingClassifier tỏ ra vượt trội về độ chính xác và cân bằng giữa các chỉ số, phù hợp với các bài toán phân loại nhị phân có yêu cầu cao về độ tin cậy.

### 3.4.3 So sánh hiệu suất phân loại giữa DeePhage và XGBoost trên bộ dữ liệu xây dựng



Hình 3.1: Kết quả hiệu suất phân loại của mô hình XGBoost trên tập dữ liệu xây dựng.

Hình 3.1 là biểu đồ so sánh hiệu suất phân loại của DeePhage và XGBoost trên tập dữ liệu mà nhóm báo cáo xây dựng. Có thể thấy, DeePhage cho kết quả tốt hơn

1 chút, khoảng từ 2% - 5% khi hơn XGBoost ở 2 chỉ số Sensitivity và Accuracy. Nghĩa là DeePhage cho khả năng nhận diện nhãn 1 và độ chính xác tổng thể cao hơn. Với chỉ số Specificity, XGBoost cho kết quả tốt hơn khoảng 5%, nghĩa là khả năng nhận diện nhãn 0 của XGBoost tốt hơn DeePhage.



# Chương 4

## Kết luận

Trong chương này, nhóm sinh viên tóm tắt lại các kết quả đã đạt được, thảo luận thêm về các phương pháp đã được trình bày trong Chương 2 và các đề xuất cho hướng nghiên cứu tiếp theo.

### 4.1 Dữ liệu

Các bộ dữ liệu được đề cập đến trong các phương pháp bao gồm:

- PHACTS: Nhóm tác giả PHACTS đã sử dụng bộ dữ liệu PHANTOME. Trong bộ dữ liệu PHANTOME có 654 bộ gen thực khuẩn. Tuy nhiên, nhóm tác giả PHACTS đã lựa chọn thủ công và chỉ sử dụng 227 dữ liệu. Trong đó bao gồm 148 bộ gen thực khuẩn thể ôn hoà và 79 bộ gen thực khuẩn thể độc lực. Tỷ lệ nhân trong bộ dữ liệu này là 2:1. Bộ dữ liệu này cũng được đề cập và sử dụng lại ở các phương pháp khác là DeePhage, PhaTYP và PhageBERT.
- ACLAME và PhagesDB: Nhóm tác giả PhageAI đã sử dụng hai bộ dữ liệu này để huấn luyện mô hình phân loại thực khuẩn. Trong bài báo, nhóm tác giả không mô tả chi tiết về số lượng dữ liệu cũng như cách chọn dữ liệu từ hai tập dữ liệu này. Nhóm tác giả chỉnh sửa thủ công và chọn ra bộ dữ liệu huấn luyện gồm: 278 bộ gen thực khuẩn thể độc lực và 174 bộ gen thực khuẩn thể ôn hoà. Và bộ dữ liệu thử nghiệm gồm 54 bộ gen thực khuẩn thể độc lực và 30 bộ gen thực khuẩn thể ôn hoà. Tổng cả có thể xem bộ dữ liệu này gồm 536 bộ gen thực khuẩn.
- Mavrich & Hatfull (2017): Bộ dữ liệu này được giới thiệu ở bài báo của Mavrich & Hatfull (2017). Bộ dữ liệu này gồm 1057 bộ gen thực khuẩn.
- NCBI-tháng 3 năm 2021: Bộ dữ liệu này gồm 1640 bộ gen thực khuẩn. Trong đó, 1211 bộ gen thực khuẩn thể độc lực và 429 bộ gen thực khuẩn thể ôn hoà.
- NCBI RefSeq 2022: Bộ dữ liệu này gồm 3474 bộ gen thực khuẩn được công bố trước năm 2022. Bộ dữ liệu này được PhaTYP sử dụng cho nhiệm vụ học

tự giám sát và nhiệm vụ tinh chỉnh. Trong đó, nhiệm vụ học tự giám sát sử dụng 3474 bộ gen thực khuẩn. Nhiệm vụ tinh chỉnh sử dụng 1290 thực khuẩn thể độc lực và 577 thực khuẩn thể ôn hoà.

- NCBI- tháng 10 năm 2024: Nhóm tác giả của DeepPL sử dụng cơ sở dữ liệu NCBI với các cập nhật mới hơn bản tháng 3 năm 2021. Nhóm tác giả DeepPL sử dụng bộ dữ liệu huấn luyện gồm 1262 bộ gen thực khuẩn thể độc lực và 557 bộ gen thực khuẩn thể ôn hoà. Tập dữ liệu kiểm thử gồm 245 bộ gen thực khuẩn thể độc lực và 129 bộ gen thực khuẩn thể ôn hoà.

Các bộ dữ liệu trên được sử dụng ở các phương pháp như sau:

STT	Phương pháp	Nguồn dữ liệu	Kích thước bộ dữ liệu	Nhân 1	Nhân 0
1	PHACTS	PHANTOME	227	148	79
2	PhageAI	ACLAME và PhagesDB	597	381	216
3	BACHPHLIP <sup>1</sup>	PHACTS, Mavrich & Hatfull	1057		
4	DeePhage <sup>2</sup>	PHACTS, NCBI	1865	1288	577
5	PhaTYP <sup>3</sup>	DeePhage, NCBI, RefSeq	5339		
6	DeepPL <sup>4</sup>	DeePhage, NCBI	1819	1262	557

Bảng 4.1: Bảng tổng hợp thông tin các bộ dữ liệu

Từ bảng 4.1, ta có thể thấy rằng các phương pháp được phát triển sau có xu hướng sử dụng bộ dữ liệu lớn hơn so với các phương pháp trước đó. Điều này cho thấy rằng, các phương pháp sau đã có sự cải tiến trong việc thu thập và xử lý dữ liệu đầu vào.

## 4.2 Phương pháp

Trong báo cáo này, nhóm sinh viên đã trình bày một cách tổng quan về các phương pháp phân loại thực khuẩn dựa trên tính toán. Bảng 4.2 tổng hợp lại các phương pháp mã hóa dữ liệu và mô hình mà các tác giả sử dụng. Các phương pháp này bao gồm: PHACTS, PhageAI, BACPHLIP, DeePhage, PhaTYP, DeepPL. Tuy cùng giải quyết một bài toán là phân loại thực khuẩn, hai nhóm lại có cách tiếp

<sup>1</sup>Nguồn tải dữ liệu: <https://zenodo.org/records/4058664>

<sup>2</sup>Nguồn tải dữ liệu: <https://gigadb.org/dataset/100918>

<sup>3</sup>Nguồn tải dữ liệu: <https://github.com/KennthShang/PhaTYP>

<sup>4</sup>Nguồn tải dữ liệu: Phần Supporting information trong bài báo

STT	Phương pháp	Phương pháp mã hóa	Mô hình sử dụng
1	PHACTS	Tạo véc-tơ đặc trưng tương đồng thông qua tập protein được xây dựng dựa trên bộ dữ liệu	Random Forest
2	PhageAI	Sử dụng kmer để đưa dữ liệu di truyền về dạng văn bản, sau đó sử dụng Word2Vec để mã hóa dữ liệu	Các mô hình học máy truyền thống
3	BACPHLIP	Tạo véc-tơ đặc trưng nhị phân biểu diễn sự xuất hiện của vùng protein bảo tồn.	Random Forest
4	DeePhage	Tạo các véc-tơ one-hot biểu diễn cho các nucleotide	CNN
5	PhaTYP	Sử dụng Bert để mã hóa dữ liệu với token là k-mer	Bert
6	DeepPL	Sử dụng Bert để mã hóa dữ liệu với token là protein.	Bert

Bảng 4.2: Bảng tổng hợp phương pháp và mô hình các bài báo sử dụng

cận khác nhau. Nhóm 1 gồm có PHACTS, PhageAI, BACPHLIP là các phương pháp sử dụng các thuật toán học máy truyền thống với đầu vào là bộ dữ liệu di truyền đầy đủ của thực khuẩn. Ta có thể dễ dàng nhận thấy, bài toán mà nhóm phương pháp 1 giải quyết có điều kiện lý tưởng khi mà dữ liệu đầu vào của các thuật toán là bộ dữ liệu di truyền đầy đủ của thực khuẩn, điều khó có thể thu thập được trong thực tế. Sau khi đã đạt các kết quả gần như tuyệt đối với bài toán này, các nhà khoa học dần chuyển sang giải quyết bài toán với điều kiện sát với thực tế hơn, khi mà dữ liệu đầu vào là các đoạn di truyền không đầy đủ, điều thường xuất hiện khi dữ liệu được thu thập. Các phương pháp nhóm 2 gồm có: DeePhage, PhaTYP, DeepPL dần được phát triển để thực hiện phân loại đoạn dữ liệu di truyền không đầy đủ của thực khuẩn.

### 4.3 Kết quả thực nghiệm

Trong báo cáo, nhóm sinh viên đã thực hiện mô phỏng lại DeePhage và so sánh với mô hình XGBoost. Với thực nghiệm 2, nhóm sinh viên đã thực hiện tạo một bộ dữ liệu mới dựa trên hai bộ dữ liệu của hai bài báo: DeePhage[8] và DeepPL[9]. Kết quả của mô hình DeePhage trên tập này tốt hơn XGBoost ở 2 chỉ số Sensitivity và Accuracy với chênh lệch khoảng từ 2% đến 5%. Điều này cho thấy DeePhage nhận diện nhãn 1 - thực khuẩn thể độc lực tốt hơn XGBoost và dự đoán chính xác hơn khi có Accuracy cao hơn. XGBoost tốt hơn DeePhage ở chỉ số Specificity cho thấy mô hình này nhận diện nhãn 0 - thực khuẩn thể ôn hòa tốt hơn DeePhage.

## 4.4 Hướng nghiên cứu tiếp theo

Trong tương lai, nhóm sinh viên dự định sẽ tiếp tục tìm hiểu, áp dụng và cải tiến các phương pháp phân loại thực khuẩn dựa trên tính toán khác. Hiện tại, các phương pháp được áp dụng chủ yếu là tinh chỉnh một mô hình đã được huấn luyện trước sau đó dùng chúng để thực hiện phân loại. Nhóm tác giả của mô hình DNABERT[3] đã công bố 2 phiên bản nâng cấp là DNABERT-2[10] và DNABERT-S[11]. Nhóm sinh viên sẽ tiếp tục tìm hiểu và tìm cách áp dụng để cải thiện hiệu suất phân loại trên bộ dữ liệu của mình. Ngoài ra, các phương pháp về mạng nơ-ron đồ thị cũng là một giải pháp tiềm năng khi có thể áp dụng cho dữ liệu di truyền.

# Tài liệu tham khảo

- [1] Ho Bin Jang, Benjamin Bolduc, Olivier Zablocki, Jens H Kuhn, Simon Roux, Evelien M Adriaenssens, J Rodney Brister, Andrew M Kropinski, Mart Krupovic, Rob Lavigne, Dann Turner, and Matthew B Sullivan. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature biotechnology*, 37(6):632—639, June 2019.
- [2] Adam J Hockenberry and Claus O Wilke. Bacphlip: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ*, 9:e11396, 2021.
- [3] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.
- [4] Katelyn McNair, Barbara A Bailey, and Robert A Edwards. Phacts, a computational approach to classifying the lifestyle of phages. *Bioinformatics*, 28(5):614–618, 2012.
- [5] A. R. Mushegian. Are there  $10^{31}$  virus particles on earth, or more, or fewer? *Journal of Bacteriology*, 202(9):10.1128/jb.00052–20, 2020.
- [6] Jiayu Shang, Xubo Tang, and Yanni Sun. Phatyp: predicting the lifestyle for bacteriophages using bert. *Briefings in Bioinformatics*, 24(1):bbac487, 2023.
- [7] Piotr Tynecki, Arkadiusz Guziński, Joanna Kazimierczak, Michał Jadczuk, Jarosław Dastyh, and Agnieszka Onisko. Phageai-bacteriophage life cycle recognition with machine learning and natural language processing. *BioRxiv*, pages 2020–07, 2020.
- [8] Shufang Wu, Zhencheng Fang, Jie Tan, Mo Li, Chunhui Wang, Qian Guo, Congmin Xu, Xiaoqing Jiang, and Huaiqiu Zhu. Deephage: distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach. *Gigascience*, 10(9):giab056, 2021.
- [9] Yujie Zhang, Mark Mao, Robert Zhang, Yen-Te Liao, and Vivian CH Wu. Deeppl: A deep-learning-based tool for the prediction of bacteriophage lifecycle. *PLOS Computational Biology*, 20(10):e1012525, 2024.

- [10] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.
- [11] Zhihan Zhou, Weimin Wu, Harrison Ho, Jiayi Wang, Lizhen Shi, Ramana V Davuluri, Zhong Wang, and Han Liu. Dnabert-s: Learning species-aware dna embedding with genome foundation models. *arXiv preprint arXiv:2402.08777*, 2, 2024.