**Xây dựng mô hình dự đoán chuỗi peptide kháng khuẩn bằng học máy**

**Tóm tắt**

Antimicrobial Peptides (AMPs) là các phân tử tự nhiên có khả năng kháng khuẩn, virus và các tác nhân gây bệnh, trở thành một giải pháp triển vọng trong y học. Việc ứng dụng học máy cho phép phân tích, khai thác các tập dữ liệu sinh học lớn để dự đoán và tối ưu hóa hoạt tính kháng khuẩn của AMPs. Trong nghiên cứu này, hai mô hình học máy Support Vector Machine (SVM) và Long Short-Term Memory (LSTM) được áp dụng để dự đoán hoạt tính sinh học của AMPs. Kết quả thực nghiệm cho thấy mô hình LSTM vượt trội hơn mô hình SVM về khả năng dự đoán. Cụ thể, mô hình LSTM đạt được độ chính xác (Accuracy) lên tới 90%, trong khi SVM chỉ đạt khoảng 86%. Các chỉ số khác như F1-Score, Precision và Recall đều cho thấy LSTM có hiệu quả vượt trội, đặc biệt trong việc xử lý chuỗi dữ liệu tuần tự và tự động trích xuất đặc trưng mà không cần qua bước xử lý thủ công. LSTM không chỉ chính xác trong dự đoán mà còn giảm thiểu sự phụ thuộc vào các quy trình xử lý thủ công, tối ưu hóa việc khai thác thông tin tuần tự trong chuỗi peptide. Điều này đánh dấu bước tiến quan trọng trong phát triển thuốc kháng sinh mới, giúp giải quyết thách thức kháng thuốc trong y học hiện đại.

Từ khóa: Chuỗi peptides, AMPs, DNA, Học sâu, LSTM.

**1. Giới thiệu**

Kháng lại thuốc kháng sinh (AMR) hiện đang là một trong những thách thức nghiêm trọng nhất đối với y học hiện đại. Các chủng vi khuẩn như Staphylococcus aureus kháng methicillin (MRSA), Enterococcus kháng vancomycin (VRE) và các vi khuẩn Enterobacteriaceae kháng carbapenem (CRE) đang trở thành mối đe dọa lớn đối với sức khỏe cộng đồng. Sự phát triển của các siêu vi khuẩn này không chỉ làm phức tạp việc điều trị mà còn tạo ra gánh nặng lớn đối với hệ thống y tế. Kháng thuốc kháng sinh xảy ra khi vi khuẩn phát triển khả năng đề kháng với các loại thuốc vốn được sử dụng để tiêu diệt chúng. Các vi khuẩn như MRSA và VRE đặc biệt nguy hiểm đối với những bệnh nhân có hệ miễn dịch yếu, hoặc những người cần chăm sóc y tế dài ngày. Sự kháng thuốc của các loại vi khuẩn làm giảm hiệu quả của các kháng sinh phổ biến như methicillin và vancomycin, trong đó có nhóm vi khuẩn sản xuất beta-lactamase phổ rộng (ESBL) ngày càng gia tăng, làm vô hiệu hóa nhiều loại kháng sinh thông dụng. Đặc biệt, chủng Enterobacteriaceae kháng carbapenem (CRE), có khả năng kháng lại carbapenem – một nhóm kháng sinh mạnh nhất hiện nay, gây ra rất nhiều khó khăn trong điều trị.

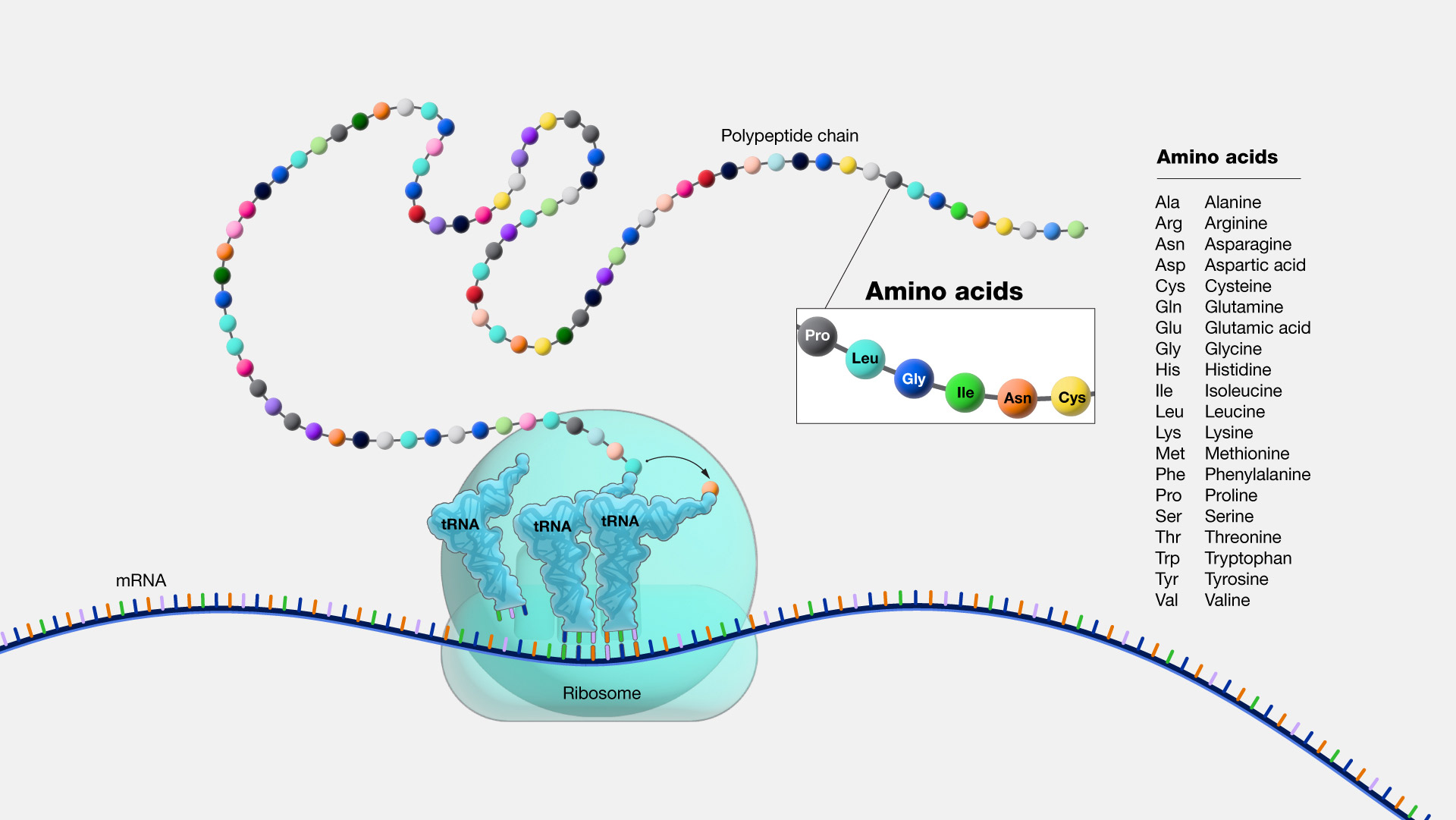
Chuỗi peptide kháng khuẩn (AMPs) được cấu tạo từ các amino acid, là đơn vị cơ bản quyết định tính chất hóa học và hoạt tính kháng khuẩn. Chúng thuộc nhóm peptide, gồm các chuỗi ngắn từ 10–50 amino acid liên kết với nhau qua liên kết peptide. AMPs mang điện tích dương và có tính phân cực, cho phép chúng tương tác mạnh với màng tế bào vi sinh vật, vốn mang điện tích âm. Các peptide này thường có cấu trúc phân cực, với một đầu kỵ nước và một đầu ưa nước, giúp chúng dễ dàng xuyên qua lớp màng lipid kép của vi khuẩn, như defensin và cathelicidin, chúng tiêu diệt vi khuẩn bằng cách phá vỡ màng tế bào, làm giảm khả năng vi khuẩn phát triển kháng thuốc. AMPs có thể hình thành từ sự cắt nhỏ của các protein lớn hơn.

A diagram of different colored circles

Description automatically generated

**Hình1.**

AMPs được tìm thấy trong nhiều sinh vật khác nhau, bao gồm động vật, thực vật và vi sinh vật, chúng đóng vai trò quan trọng trong hệ miễn dịch tự nhiên. AMPs là các chuỗi peptide ngắn, được mã hóa bởi các gen trong DNA và trải qua quá trình phiên mã (mRNA) và dịch mã (ribosome) để hình thành. Trong quá trình dịch mã, các amino acid được sắp xếp theo trình tự quy định bởi mRNA, tạo nên chuỗi polypeptide.



**Hình 2**

Quá trình tổng hợp peptide kháng khuẩn (AMPs) bắt đầu từ DNA trong nhân tế bào, nơi thông tin di truyền được sao chép thành mRNA. mRNA này chứa vùng khung đọc mở (ORF), nơi mã hóa chuỗi peptide. Khung đọc mở bắt đầu từ codon khởi đầu (AUG) và kết thúc tại một trong các codon kết thúc (UAA, UAG hoặc UGA). Sau khi quá trình dịch mã hoàn tất, mRNA được chuyển thành chuỗi peptide, trong đó các amino acid được sắp xếp theo một trình tự cụ thể, tạo thành AMPs hoàn chỉnh.

Mô hình Open Reading Frame (ORF)

A diagram of a dna

Description automatically generated with medium confidence

**Hình 3**

Quá trình tổng hợp peptide bắt đầu từ DNA, nơi lưu trữ thông tin di truyền của sinh vật. Đầu tiên, DNA được sử dụng làm khuôn mẫu để phiên mã thành mRNA. mRNA này chứa bản sao mã hóa thông tin di truyền và đóng vai trò như một bản hướng dẫn, truyền tải thông tin cần thiết cho việc tổng hợp peptide. Sau đó, mRNA di chuyển ra ngoài nhân tế bào, nơi nó được dịch mã thành chuỗi peptide qua sự tham gia của ribosome và các tRNA mang amino acid tương ứng.

Trong bối cảnh hiện tại, việc nghiên cứu và phát triển thuốc mới nhanh chóng là vô cùng cấp thiết. Việc ứng dụng công nghệ học máy trong việc dự đoán và thiết kế peptide kháng khuẩn (AMPs) trước khi bước vào giai đoạn sản xuất, đóng vai trò vô cùng quan trọng trong việc giảm thiểu rủi ro, tiết kiệm chi phí. Phương pháp này giúp quá trình phát triển các loại thuốc kháng sinh mới được thực hiện nhanh chóng, hiệu quả, giảm gánh nặng về chi phí nhân vật lực trong quá trình nghiên cứu và sản xuất. Học máy với khả năng xử lý khối lượng dữ liệu lớn, giúp nhận diện mối liên hệ giữa các đặc tính của peptide và tính kháng khuẩn của chúng. Việc sử dụng các thuật toán học máy, có thể dự đoán hiệu quả của peptide trước khi tiến hành thí nghiệm thực tế, từ đó tiết kiệm thời gian và chi phí. Đồng thời, cũng hỗ trợ tối ưu hóa thiết kế peptide, cải thiện tính đặc hiệu và tăng cường hiệu lực của chúng đối với các vi khuẩn kháng thuốc. Sự kết hợp giữa mô hình học máy và các phương pháp phân tích dữ liệu tiên tiến đã mở ra cơ hội khám phá những peptide nhanh chóng và chính xác. Học máy có thể đánh giá hàng triệu chuỗi peptide, chọn lọc và tối ưu hóa các chuỗi có hiệu quả kháng khuẩn cao nhất, góp phần đẩy nhanh quá trình phát triển thuốc mới, rút ngắn thời gian nghiên cứu, thử nghiệm, giúp ngăn chặn kịp thời tình trạng kháng thuốc khánh sinh hiện nay.

**2.** **Phương pháp nghiên cứu peptide kháng khuẩn (AMPs)**

**2.1.** **Phương pháp truyền thống trong nghiên cứu peptide kháng khuẩn (AMPs)**

Phần lớn các peptide kháng khuẩn (AMPs) tự nhiên được phát hiện thông qua phương pháp phân lập từ nguồn sinh học và đặc trưng hóa bằng các kỹ thuật truyền thống, tuy nhiên, các phương pháp này thường tốn nhiều thời gian và công sức. Để khắc phục những hạn chế này, phương pháp tính toán dựa trên dữ liệu peptide và bộ gen của các sinh vật đã được áp dụng. Phương pháp này có thể chia thành năm nhóm chính: Nhóm một, dựa trên peptide trưởng thành (Mature peptide), tập trung vào các đặc điểm cấu trúc như tính axit, phân cực và khả năng tương tác với màng tế bào để xác định các peptide kháng khuẩn. Nhóm hai, dựa trên propeptide, phân tích các peptide nguyên mẫu chưa qua xử lý, dự đoán các peptide có khả năng trở thành AMPs sau khi trải qua quá trình cắt và biến đổi dịch mã. Nhóm ba, kết hợp peptide trưởng thành và propeptide, giúp tăng độ chính xác trong dự đoán AMPs bằng cách tận dụng đặc điểm của cả hai giai đoạn trong quá trình sinh tổng hợp peptide. Nhóm bốn, dựa trên enzyme xử lý (Processing Enzyme), phân tích các enzyme chịu trách nhiệm cắt và xử lý propeptide thành AMPs, hỗ trợ dự đoán các peptide kháng khuẩn tiềm năng. Nhóm 5, dựa trên ngữ cảnh gen (Genomic Context), khai thác thông tin bộ gene của sinh vật để xác định các gen mã hóa AMPs và làm rõ các yếu tố môi trường và sinh lý có thể kích hoạt quá trình sinh tổng hợp AMPs.

**2.2. Phương pháp Global Protein Sequence Descriptor (GPSD)**

Global Protein Sequence Descriptor (GPSD) là phương pháp mã hóa trình tự protein dưới dạng số, dựa trên các đặc tính lý hóa của các amino acid trong chuỗi protein. Mục tiêu của GPSD là rút trích thông tin từ chuỗi AMPs thành các giá trị số, giúp các mô hình học máy phân tích và dự đoán các tính chất quan trọng của chuỗi AMPs, như quá trình gấp cuộn protein (protein folding), chức năng sinh học và khả năng kháng khuẩn. Nhờ việc mã hóa thông tin dựa trên các thuộc tính lý hóa của amino acid, GPSD cung cấp một công cụ hiệu quả trong nghiên cứu và ứng dụng sinh học phân tử.

**2.2.1. Đặc điểm chính của GPSD**

Các đặc điểm chính của Global Protein Sequence Descriptor, sử dụng các thuộc tính lý hóa của các amino acid như độ kỵ nước (hydrophobicity), thể tích van der Waals chuẩn hóa (normalized van der Waals volume), độ phân cực (polarity), khả năng phân cực (polarizability), điện tích (charge), cấu trúc thứ cấp (secondary structure) và khả năng tiếp xúc với dung môi (solvent accessibility). Những đặc tính này ảnh hưởng trực tiếp đến cấu trúc và chức năng của protein []. Nhờ vào các kỹ thuật học máy, việc dự đoán AMP đã đạt được độ chính xác và hiệu quả cao hơn, đóng góp vào sự phát triển các liệu pháp điều trị kháng sinh, đặc biệt trong bối cảnh kháng thuốc đang trở thành thách thức lớn trong y học hiện đại.

**2.2.2. Phân nhóm acid amin**

Việc phân nhóm các amino acid, để dễ dàng phân tích và xử lý, các amino acid trong chuỗi protein được phân loại thành các nhóm dựa trên các thuộc tính lý hóa. Nếu xét theo thuộc tính độ kỵ nước (hydrophobicity), các amino acid có thể được chia thành ba nhóm chính:

**Polar (Phân cực**): R, K, E, D, Q, N

**Neutral (Trung tính)**: G, A, S, T, P, H, Y

**Hydrophobic (Kỵ nước)**: C, L, V, I, M, F, W

**2.2.3. Distribution Descriptor**

Distribution Descriptors (Mô tả phân bố vị trí), là phương pháp tính toán quan trọng của GPSD, giúp mô tả phân bố của các nhóm amino acid trong chuỗi protein. Phương pháp này xác định tỷ lệ vị trí của các amino acid thuộc các nhóm khác nhau (polar, neutral, hydrophobic) ở các phần tử đầu tiên, 25%, 50%, 75%, và 100% trong chuỗi. Distribution Descriptors cung cấp cái nhìn chi tiết về phân bố của từng nhóm amino acid, từ đó giúp nhận diện các mẫu phân bố đặc trưng trong protein.

Công thức tính giá trị của Distribution Descriptor

(1)

**Trong đó**:

P: là vị trí của amino acid trong chuỗi.

L: là tổng số amino acid trong chuỗi.

**2.2.4 Các thành phần của mô tả chuỗi protein**

Để trích xuất thông tin từ chuỗi protein, các đặc điểm được tính dựa trên ba thành phần chính:

**Composition (C):** Mô tả thành phần tổng thể của chuỗi protein, tức là tỷ lệ phần trăm các nhóm axit amin trong chuỗi.

Công thức tính:

(2)

Trong đó:

: Số lượng amino acid thuộc nhóm i.

: Tổng số amino acid trong chuỗi.

**Transition (T):** Tần suất thay đổi tính chất (như sự chuyển đổi từ nhóm polar sang hydrophobic) dọc theo toàn bộ chiều dài chuỗi protein.

**Distribution (D)**: Mô tả mẫu phân bố các tính chất (polar, neutral, hydrophobic) trên toàn bộ trình tự protein.

Distribution Descriptor là chỉ số quan trọng trong GPSD, giúp chuyển đổi trình tự protein thành các giá trị số phản ánh đặc tính lý-hóa, như tính axit, tính phân cực, và khả năng tương tác với môi trường. Các giá trị này cung cấp thông tin định lượng, hỗ trợ các mô hình học máy phân tích và dự đoán các đặc tính sinh học quan trọng của chuỗi protein.

**2.3. Phương pháp học máy trong dự đoán các peptide kháng khuẩn (AMPs**)

Việc ứng dụng học máy trong dự đoán các peptide kháng khuẩn (AMP) đã trở thành một lĩnh vực nghiên cứu quan trọng, đặc biệt khi các phương pháp truyền thống như thử nghiệm thực nghiệm và các mô hình tính toán gặp nhiều hạn chế về thời gian, chi phí và khả năng xử lý dữ liệu phức tạp. Các phương pháp học máy được áp dụng phổ biến trong dự đoán AMP bao gồm:

Máy học giám sát (Supervised Learning): Các mô hình giám sát, chẳng hạn như cây quyết định, máy vector hỗ trợ (SVM - Support Vector Machine) và mạng nơ-ron nhân tạo (ANN - Artificial Neural Networks). Các mô hình này được huấn luyện trên tập dữ liệu gồm các peptide có tính kháng khuẩn và không kháng khuẩn. Sau khi huấn luyện, mô hình có thể dự đoán khả năng kháng khuẩn của các peptide. Những mô hình học máy truyền thống cần phải qua bước rút trích thông tin, cụ thể là phương pháp Global Protein Sequence Descriptor (GPSD), điều này có thể tiêu tốn nhiều thời gian và chi phí thí nghiệm.

Học sâu (Deep Learning): Các mô hình học sâu, đặc biệt là các mô hình xử lý chuỗi như RNN (Recurrent Neural Networks) và LSTM (Long Short-Term Memory), đã chứng minh hiệu quả vượt trội trong dự đoán AMP. Những mô hình này có khả năng xử lý chuỗi amino acid phức tạp và tự động trích xuất các đặc trưng sâu hơn mà các phương pháp truyền thống có thể bỏ qua. Ngoài ra, học sâu cho phép bỏ qua giai đoạn trích xuất đặc trưng thủ công, đồng thời bảo toàn thông tin về thứ tự của các amino acid trong chuỗi protein. Điều này đặc biệt hữu ích khi có một lượng lớn dữ liệu để huấn luyện, góp phần nâng cao độ chính xác trong dự đoán.

**2.4. Mô hình học máy đề xuất**

Nghiên cứu này xây dựng và đánh giá thực nghiệm trên hai mô hình học máy để dự đoán khả năng kháng khuẩn của chuỗi peptide trước khi đưa vào thử nghiệm và sản xuất công nghiệp.

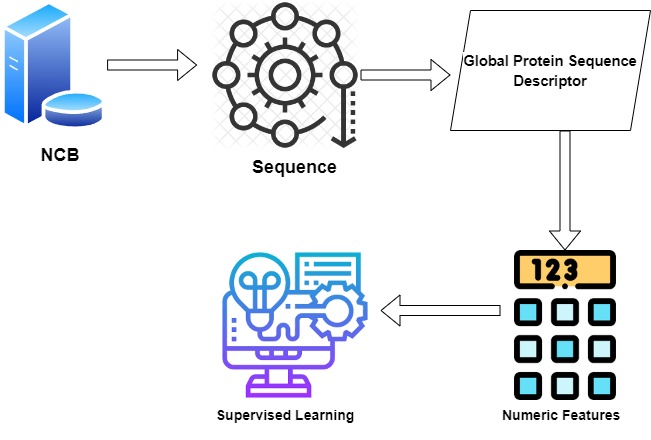
Mô hình 1: Học máy truyền thống

Mô hình này áp dụng các phương pháp học máy truyền thống, kết hợp các thuật toán phân tích, để học từ dữ liệu về cấu trúc, tính chất lý hóa và hành vi sinh học của chuỗi peptide kháng khuẩn. Dữ liệu đầu vào được thu thập từ National Center for Biotechnology Information (NCBI), một cơ sở dữ liệu sinh học lớn, cung cấp thông tin về chuỗi DNA, RNA và protein từ nhiều sinh vật khác nhau. Mô hình này sẽ tận dụng các đặc trưng của dữ liệu để phân tích và dự đoán các tính chất quan trọng của peptide kháng khuẩn, hỗ trợ nghiên cứu và đánh giá trước khi đưa vào sản xuất công nghiệp.

Quy trình xử lý dữ liệu và dự đoán của mô hình bao gồm:

* Thu thập dữ liệu từ NCBI.
* Chuẩn bị chuỗi peptide (Sequence).
* Trích xuất đặc trưng bằng phương pháp Global Protein Sequence Descriptor (GPSD).
* Tạo tập đặc trưng (Features).
* Áp dụng thuật toán học máy, Support Vector Machine (SVM), để dự đoán khả năng kháng khuẩn của các chuỗi peptide.

Mô hình học máy truyền thống trong nghiên cứu này, được thể hiện như trong Hình 4.



**Hình 4.**

Phương pháp học máy truyền thống kết hợp thuật toán Support Vector Machine (SVM) với bước rút trích đặc trưng từ chuỗi peptide kháng khuẩn. Tuy nhiên, phương pháp này có một số hạn chế:

Phụ thuộc vào bước trích xuất đặc trưng thủ công, điều này làm tăng chi phí và thời gian xử lý.

Không bảo toàn được tính liên kết tuần tự của các amino acid trong chuỗi peptide, ảnh hưởng đến khả năng phản ánh đầy đủ các đặc trưng sinh học.

Mặc dù có những hạn chế trên, mô hình học máy truyền thống đóng vai trò quan trọng làm cơ sở so sánh để đánh giá hiệu quả của mô hình được đề xuất trong nghiên cứu này.

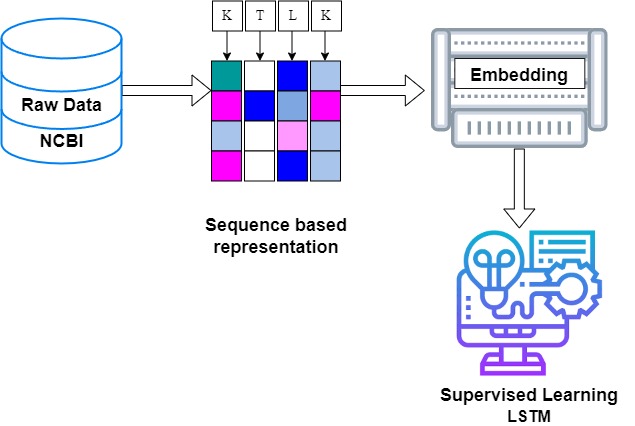
Mô hình 2: Học sâu (Deep Learning)

Dựa trên những hạn chế của mô hình học máy truyền thống, nghiên cứu đề xuất một mô hình học sâu sử dụng Long Short-Term Memory (LSTM) để trực tiếp xử lý chuỗi protein, bỏ qua bước trích xuất đặc trưng thủ công. Mô hình này tận dụng khả năng giữ nguyên thứ tự của các amino acid trong chuỗi, giúp bảo toàn thông tin tuần tự quan trọng.

Quy trình xử lý của mô hình học sâu:

* Thu thập dữ liệu: Dữ liệu peptide được lấy từ NCBI.
* Chuẩn bị chuỗi peptide: Biểu diễn chuỗi peptide (Sequence).
* Embedding: Biểu diễn chuỗi peptide dưới dạng vector số hóa.
* Dự đoán: Sử dụng mô hình LSTM để dự đoán khả năng kháng khuẩn.

Việc áp dụng mô hình học sâu này, không chỉ cải thiện độ chính xác trong dự đoán mà còn giảm thiểu phụ thuộc vào các bước xử lý thủ công, đồng thời tận dụng toàn bộ dữ liệu tuần tự của peptide. Điều này giúp giải quyết các bài toán phức tạp trong nghiên cứu và phát triển peptide kháng khuẩn. Mô hình đề xuất được thể hiện như trong Hình 5.



**Hình 5.**

**2.4.1. Tập dữ liệu**

Nghiên cứu này sử dụng bộ dữ liệu từ National Center for Biotechnology Information (NCBI), bao gồm hai nhóm peptide chính: peptide kháng khuẩn (AMPs) và peptide không kháng khuẩn. Bộ dữ liệu tổng hợp 6000 peptide, trong đó có 3000 peptide kháng khuẩn và 3000 peptide không kháng khuẩn. Các peptide vi sinh vật được lựa chọn là đối tượng nghiên cứu chính vì dễ khai thác, dễ thực hiện thí nghiệm, và dữ liệu về chúng đã có sẵn trên NCBI. Bộ dữ liệu này cung cấp một cái nhìn toàn diện về các loại peptide khác nhau, hỗ trợ việc nghiên cứu, dự đoán peptide kháng khuẩn và phát triển các loại thuốc kháng sinh mới. Điều này có ý nghĩa quan trọng trong cải thiện hiệu quả điều trị và giải quyết vấn đề kháng thuốc – một thách thức lớn trong y học hiện đại.

**2.4.2. Tiền xử lý dữ liệu**

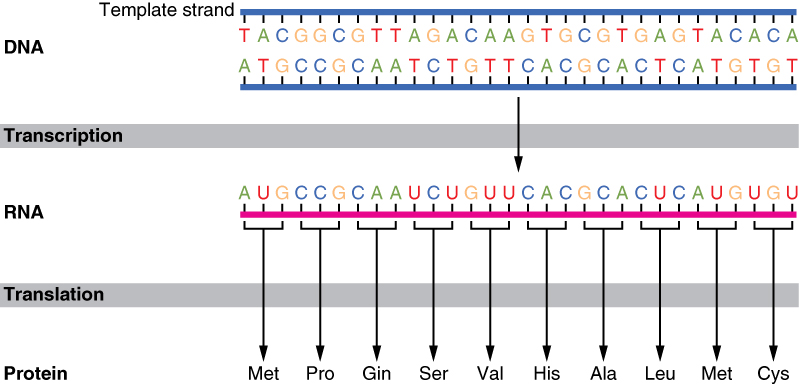
Quá trình tiền xử lý dữ liệu được tiến hành theo các bước sau:

Bước 1: Sử dụng thư viện BioPython để tải dữ liệu trực tiếp từ cơ sở dữ liệu NCBI. Bộ dữ liệu đã được gán nhãn, bao gồm 60000 mẫu, với 30 000 peptide kháng khuẩn và 30 000 peptide không kháng khuẩn.

Bước 2: Tiến hành trích xuất đặc trưng bằng phương pháp Global Protein Sequence Descriptor.

Quá trình này bắt đầu từ chuỗi DNA, chuyển hóa thành RNA thông qua phiên mã, sau đó dịch thành chuỗi AMPs thông qua khung đọc mở (ORF).

Việc tổng hợp protein từ DNA được chia thành hai giai đoạn chính: phiên mã (transcription) và dịch mã (translation), đảm bảo việc chuyển đổi thông tin di truyền từ gene thành chuỗi protein chức năng.



*Bước 2.1 Phiên mã (Transcription):*

Trong quá trình phiên mã, chuỗi DNA khuôn mẫu (template strand) được sử dụng để tổng hợp RNA thông tin (mRNA) thông qua sự hoạt động của enzyme RNA polymerase. Quá trình này tuân theo nguyên tắc bổ sung giữa các nucleotide, trong đó Adenine (A) trên DNA bắt cặp với Uracil (U) trên RNA, Thymine (T) trên DNA bắt cặp với Adenine (A) trên RNA, Cytosine (C) trên DNA bắt cặp với Guanine (G) trên RNA, và Guanine (G) trên DNA bắt cặp với Cytosine (C) trên RNA. Kết quả là một chuỗi mRNA được tổng hợp, mang trình tự tương ứng với gene mã hóa trên DNA.

*Bước 2.2. Dịch mã (Translation):*

Giai đoạn dịch mã diễn ra tại ribosome, nơi mRNA được đọc theo từng bộ ba nucleotide liên tiếp gọi là codon. Mỗi codon trên mRNA mã hóa một axit amin cụ thể, dựa trên bảng mã di truyền. Ví dụ, codon AUG mã hóa Methionine (Met) và đánh dấu điểm khởi đầu, trong khi các codon như CCU, CAA, UCC lần lượt mã hóa Proline (Pro), Glutamine (Gln), và Serine (Ser). Ribosome kết hợp các axit amin lại với nhau thành chuỗi polypeptide, theo trình tự xác định bởi mRNA. Quá trình này kết thúc khi ribosome gặp một codon kết thúc (stop codon), chẳng hạn như UAA, UAG hoặc UGA, tín hiệu để giải phóng chuỗi polypeptide hoàn chỉnh.

Cả hai giai đoạn này cùng phối hợp để chuyển đổi thông tin di truyền từ DNA thành một chuỗi protein có chức năng, phục vụ cho các hoạt động sinh học của tế bào.

*Bước 2.3. Chuyển từ RNA sang Khung Đọc Mở (Open reading frame - ORF):*

Để dịch một chuỗi RNA thành Khung đọc mở (ORF), thực hiện một số bước cơ bản. Đầu tiên, chuỗi RNA phải bắt đầu với một codon khởi đầu, thường là AUG, mã hóa cho axit amin Methionine. Tiếp theo, RNA được đọc theo từng bộ ba nucleotide (codon), mỗi codon mã hóa cho một axit amin hoặc một tín hiệu dừng. Quá trình dịch mã kết thúc khi gặp một codon kết thúc, như: UAA, UAG, hoặc UGA, báo hiệu kết thúc quá trình dịch mã.

Quá trình trích xuất ORF từ chuỗi RNA bao gồm các bước như: Tìm codon khởi đầu AUG; Đọc các codon tiếp theo cho đến khi gặp codon kết thúc; Trích xuất chuỗi axit amin tương ứng với phần RNA.

Trong quá trình sinh học thực tế, RNA sẽ được dịch mã thành protein bằng cách sử dụng bảng mã di truyền, trong đó mỗi codon tương ứng với một axit amin cụ thể.

A diagram of a sequence of dna

Description automatically generated

*Bước 2.4. Dịch mã từ ORF thành peptide:*

Để chuyển đổi một Khung đọc mở (ORF) thành peptide, quá trình dịch mã RNA diễn ra theo các bước sau: Đầu tiên, các codon trong ORF được xác định. Mỗi codon là một bộ ba nucleotide RNA, mã hóa một axit amin cụ thể. Quá trình dịch mã bắt đầu từ codon khởi đầu AUG, mã hóa cho axit amin Methionine (Met), đánh dấu điểm bắt đầu của chuỗi polypeptide. Tiếp theo, các codon sau đó sẽ được dịch thành các axit amin tương ứng theo bảng mã di truyền. Mỗi codon trong chuỗi RNA chỉ định một axit amin duy nhất, quá trình dịch mã diễn ra cho đến khi gặp một codon kết thúc (UAA, UAG hoặc UGA). Khi gặp codon kết thúc, chuỗi axit amin được hoàn thành và sẽ hình thành một peptide hoàn chỉnh.

Quá trình dịch mã này giúp tạo ra protein từ thông tin di truyền trong RNA, và đây là cơ sở để thực hiện các chức năng sinh học trong tế bào. Việc dịch mã dựa vào bảng ánh xạ codon thành axit amin, trong đó mỗi bộ ba nucleotide tương ứng với một axit amin cụ thể.

Dưới đây là bảng ánh xạ codon sang axit amin dưới dạng bảng:

Bảng 1: **Bảng ánh xạ Codon sang Axit amin**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Codon | Amino Acid | Codon | Amino Acid | Codon | Amino Acid | Codon | Amino Acid |
| AUG | M | UUU | F | UUC | F | UUA | L |
| UUG | L | UCU | S | UCC | S | UCA | S |
| UCG | S | UAU | Y | UAC | Y | UGU | C |
| UGC | C | UGG | W | UAA | \* | UAG | \* |
| UGA | \* | CUU | L | CUC | L | CUA | L |
| CUG | L | CCU | P | CCC | P | CCA | P |
| CCG | P | CAU | H | CAC | H | CAA | Q |
| CAG | Q | CGU | R | CGC | R | CGA | R |
| CGG | R | AUU | I | AUC | I | AUA | I |
| ACU | T | ACC | T | ACA | T | ACG | T |
| AAU | N | AAC | N | AAA | K | AAG | K |
| AGU | S | AGC | S | AGA | R | AGG | R |
| GUU | V | GUC | V | GUA | V | GUG | V |
| GCU | A | GCC | A | GCA | A | GCG | A |
| GAU | D | GAC | D | GAA | E | GAG | E |
| GGU | G | GGC | G | GGA | G | GGG | G |

Giả sử muốn chuyển ORF Sequence: **AUG**CGUAUGCG**UAA** Peptide Sequence, thực hiện theo các bước sau:

Tìm mã di truyền (codon) trong ORF:

* AUG → Methionine (M)
* CGU → Arginine (R)
* AUG → Methionine (M)
* CGU → Arginine (R)
* UAA → Stop codon (ngừng dịch mã)

Vậy ta thu được peptide sequence từ ORF sequence “**AUG**CGUAUGCG**UAA***”* là “*MRMR”.*

***Bướ 3. Phân tích các đặc trưng của chuỗi peptide:***

Sử dụng Global Protein Sequence Descriptor (GPSD) để rút trích đặc trưng từ chuỗi peptide dựa trên bảy thuộc tính chính, giúp mô tả tính chất sinh học và hóa học của peptide. Các đặc trưng này bao gồm:

* Độ kỵ nước (Hydrophobicity): Phân loại amino acid thành ba nhóm:
* Polar (P): D, K, E , Q, N.
* Neutral (N): G, A, S, T, P, H, Y.
* Hydrophobic (H): C, L, V, I, M, F, W.
* **Thể tích van der Waals chuẩn hóa (Normalized van der Waals volume):** Đo thể tích không gian mà các nguyên tử chiếm giữ trong peptide.
* **Độ phân cực (Polarity):** Đo sự phân cực của amino acid, ảnh hưởng đến khả năng tương tác với môi trường.
* **Khả năng phân cực (Polarizability):** Đo khả năng thay đổi phân bố điện tích của peptide dưới tác động của trường điện từ.
* **Điện tích (Charge):** Đo lường điện tích của peptide, ảnh hưởng đến sự tương tác phân tử.
* **Cấu trúc thứ cấp (Secondary structure):** Mô tả các cấu trúc không gian của peptide như alpha helix, beta sheet.
* **Khả năng tiếp xúc với dung môi (Solvent accessibility):** Đo lường mức độ tiếp xúc của amino acid với dung môi xung quanh.

Các đặc trưng này hỗ trợ trong việc mô tả và dự đoán tính chất sinh học, hóa học của peptide, từ đó giúp đánh giá khả năng kháng khuẩn.

Giả sử chúng ta có chuỗi peptide cần phân tích “**GLFDIIKKIAESI**” với độ dài 13. Quy trình rút trích đặc trưng cho chuỗi peptide này được thực hiện qua các bước sau:

#### Bước 1: Phân loại amino acid theo độ kỵ nước (hydrophobicity)

* **Polar (P):** D, K, E
* **Neutral (N):** G, A, S
* **Hydrophobic (H):** L, F, I, I, I

Kết quả phân loại chuỗi peptide "GLFDIIKKIAESI" sẽ được biểu diễn như sau: “NHHPHHPPHNPNH”.

*Bước 2: Tính toán Distribution Descriptor cho từng nhóm (Polar)*

Để tính toán các giá trị phân bố cho lớp **Polar (P)**, ta sẽ tính các giá trị Descriptor tại các điểm phân bố trong chuỗi. Áp dụng công thức (1). Giá trị Descriptor cụ thể cho lớp Polar trong chuỗi peptide “GLFDIIKKIAESI” được thể hiện trong bảng sau:

*Bảng các giá trị Descriptor cho lớp* ***Polar (P)*** *của chuỗi peptide "GLFDIIKKIAESI".*

|  |  |  |
| --- | --- | --- |
| **Residue** | **Vị trí trong chuỗi peptide** | **Descriptor Value (%)** |
| First residue (P) | Vị trí 4 (D - Aspartic acid) | 30.77% |
| 25% residue (P) | Vị trí 4 (D - Aspartic acid) | 30.77% |
| 50% residue (P) | Vị trí 8 (K - Lysine) | 61.54% |
| 75% residue (P) | Vị trí 9 (E - Glutamic acid) | 69.23% |
| 100% residue (P) | Vị trí 11 (K - Lysine) | 84.62% |

Quá trình này cung cấp cái nhìn chi tiết về và sự phân bố của các amino acid trong chuỗi peptide,

từ đó hỗ trợ trong việc đánh giá các tính chất sinh học và hóa học của các peptide.

**Bước** **4:** **Phân tích đặc trưng Composition, Transition và Distribution**

Tiếp theo sử dụng distribution descriptors để tính đặc trưng phân bố của chuỗi peptide theo các thuộc tính Composition (C), Transition (T), và Distribution (D), ta cần phân tích chuỗi peptide và cách các axit amin phân bố trong chuỗi.

*Bước 4.1. Composition (C):*

Composition mô tả tỷ lệ của từng loại axit amin trong chuỗi peptide, phản ánh thành phần tổng thể của chuỗi protein. Dưới đây là cách phân loại axit amin trong chuỗi peptide “GLFDIIKKIAESI” theo độ kỵ nước:

* Polar (P):
* D (Aspartic acid)
* K (Lysine)
* K (Lysine)
* E (Glutamic acid)
* S (Serine)

🡪 Có 5 phần tử

* Neutral (N):
* G (Glycine)
* A (Alanine)

🡪 Có 2 phần tử

* Hydrophobic (H):
* L (Leucine)
* F (Phenylalanine)
* I (Isoleucine)
* I (Isoleucine)
* K (Lysine)
* A (Alanine)
* I (Isoleucine)

🡪 Có 7 phần tử

Tỷ lệ Composition cho từng nhóm được tính theo công thức (2) như sau:

* Polar (P):
* Neutral (N):
* Hydrophobic (H):

*Bước 4.2****. Transition (T):***

Transition đo lường tần suất thay đổi tính chất (hoặc nhóm axit amin) trong chuỗi. Ta sẽ tính số lần chuyển từ nhóm này sang nhóm khác (ví dụ, từ Polar sang Hydrophobic, hoặc từ Neutral sang Polar). Mỗi lần thay đổi sẽ được tính là một sự chuyển tiếp.

Dựa trên chuỗi peptide **"NHHPHHPPHNPNH"**, có thể tính số lần chuyển tiếp giữa các nhóm:

* **Từ N -> H**: 2 lần
* **Từ H -> P**: 3 lần
* **Từ P -> N**: 2 lần
* **Từ N -> P**: 1 lần

Tổng số chuyển tiếp trong chuỗi là 8 lần.

*Bước 4.3. Distribution (D):*

Distribution mô tả sự phân bố của các tính chất trong chuỗi peptide. Dựa trên các giá trị đã tính toán từ bước trước, chúng ta xác định sự phân bố của các axit amin trong chuỗi peptide ở các vị trí khác nhau:

|  |  |
| --- | --- |
| **Residue** | **Tỷ lệ phân bố (%)** |
| First residue | 30.77% |
| 25% residue | 30.77% |
| 50% residue | 61.54% |
| 75% residue | 69.23% |
| 100% residue | 84.62% |

Các giá trị trên cho thấy sự phân bố của các axit amin thuộc lớp Polar (P) trong chuỗi peptide. Việc tính toán các đặc trưng này giúp phân tích cấu trúc và tính chất của peptide, từ đó hỗ trợ trong việc nghiên cứu hoạt tính sinh học và tính chất hóa học của chúng.

Sau khi hoàn thành bốn bước xử lý này, kết quả thu được là một tập hợp các đặc trưng bao gồm 126 biến, đại diện cho 60.000 mẫu dữ liệu. Tập dữ liệu này sẽ được sử dụng làm đầu vào cho các mô hình máy học được đề xuất trong nghiên cứu này.

*2.4.3. Tiến hành thực nghiệm trên mô hình 1.*

Trong nghiên cứu này, mô hình Support Vector Machine (SVM) được sử dụng để dự đoán khả năng kháng khuẩn của peptide dựa trên bộ dữ liệu được thu thập từ cơ sở dữ liệu National Center for Biotechnology Information (NCBI). Nhóm peptide vi sinh vật được chọn làm đối tượng nghiên cứu chính nhờ tính dễ khai thác, khả năng thực hiện thí nghiệm thuận lợi và sự sẵn có của dữ liệu trên NCBI. Bộ dữ liệu bao gồm 60.000 peptide, được phân chia thành hai nhóm chính với số lượng cân bằng: 30.000 peptide kháng khuẩn và 30.000 peptide không kháng khuẩn. Đây là nguồn dữ liệu phong phú, cung cấp thông tin quan trọng không chỉ để nghiên cứu đặc trưng peptide mà còn để phát triển các phương pháp dự đoán tiên tiến, góp phần vào việc giải quyết vấn đề kháng thuốc đang ngày càng gia tăng.

Quá trình thực nghiệm trên mô hình SVM bắt đầu với bước tiền xử lý dữ liệu nhằm làm sạch và chuẩn bị bộ dữ liệu, đảm bảo tính nhất quán và chính xác. Các đặc trưng được trích xuất từ chuỗi peptide, tạo thành tập dữ liệu đầu vào cho mô hình. Tiếp theo, bộ dữ liệu được phân chia thành hai phần theo tỷ lệ 80:20, trong đó 48.000 mẫu được sử dụng cho tập huấn luyện (training set) và 12.000 mẫu còn lại được sử dụng cho tập kiểm tra (test set). Mô hình SVM được huấn luyện trên tập huấn luyện để học cách phân biệt peptide kháng khuẩn và không kháng khuẩn dựa trên các đặc trưng đã trích xuất. Sau khi huấn luyện, mô hình được đánh giá trên tập kiểm tra thông qua các chỉ số hiệu suất quan trọng như độ chính xác (Accuracy), độ tin cậy (Kappa), điểm F1 (F1-Score), độ nhạy (Sensitivity), độ đặc hiệu (Specificity) và chỉ số phân loại tổng thể (C-measure.

Kết quả từ mô hình SVM cho thấy khả năng dự đoán chính xác và hiệu quả các peptide kháng khuẩn, khẳng định vai trò của phương pháp học máy truyền thống trong việc hỗ trợ phát triển các công cụ dự đoán peptide kháng khuẩn. Những kết quả này không chỉ tạo nền tảng cho các nghiên cứu tiếp theo mà còn mở ra hướng ứng dụng trong thiết kế và phát triển peptide kháng khuẩn mới, góp phần quan trọng vào việc giải quyết các thách thức liên quan đến kháng thuốc trong y học hiện đại.

**Đánh giá kết quả của mô hình 1.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Accuracy | Kappa | F1\_Score | Sensitivity | Specificity | C-measure |
| 0.90 | 0.87 | 0.90 | 0.92 | 0.94 | 0.93 |

**A screenshot of a computer

Description automatically generated**

**2.4.5 Thực hiện trên mô hình LSTM**

Trong nghiên cứu này, mô hình Long Short-Term Memory (LSTM) được áp dụng nhằm khai thác các mối quan hệ tuần tự trong dữ liệu peptide mà không yêu cầu bước rút trích đặc trưng thủ công. Điều này giúp tự động hóa và tối ưu hóa quá trình xử lý dữ liệu, phù hợp với bản chất chuỗi tuần tự của peptide. Các bước thực hiện được mô tả chi tiết như sau:

Bước 1: Embedding

Chuỗi peptide đầu vào được biểu diễn dưới dạng ký tự, mỗi axit amin được ánh xạ thành một số nguyên dựa trên bảng từ vựng chuẩn: “ACDEFGHIKLMNPQRSTVWXYZ”, trong đó mỗi ký tự được gán giá trị từ 1 đến 23. Để xử lý các chuỗi có độ dài khác nhau, kỹ thuật padding được áp dụng, bổ sung giá trị 0 để chuẩn hóa chiều dài của tất cả các chuỗi peptide.

Bước 2: Xây dựng mô hình LSTM

Mô hình được thiết kế với các thành phần chính:

* Embedding Layer: Biến đổi các chỉ số của axit amin thành không gian đặc trưng có thể học được.
* LSTM Layer: Xử lý chuỗi peptide tuần tự, học các mối quan hệ dài hạn và phụ thuộc thời gian.
* Dense Layers: Các đặc trưng từ LSTM được đưa qua hai lớp Dense, trong đó lớp cuối cùng sử dụng hàm kích hoạt sigmoid để dự đoán peptide kháng khuẩn (AMPs) hay không.

Bước 3: Huấn luyện và đánh giá mô hình

Mô hình LSTM được huấn luyện với bộ dữ liệu 60.000 mẫu, sử dụng hàm tối ưu Adam và độ đo chính là Accuracy. Trong quá trình huấn luyện, mô hình đạt độ chính xác 90% sau 50 epochs, vượt trội so với mô hình SVM với độ chính xác 85%. Ngoài ra, mô hình LSTM cũng cải thiện các thông số khác, bao gồm:

* F1-Score: 0.91 (so với 0.87 của SVM)
* Sensitivity (Recall): 0.92 (so với 0.85 của SVM)
* Specificity: 0.89 (so với 0.83 của SVM)

Kết luận

Mô hình LSTM không chỉ cải thiện hiệu suất dự đoán mà còn cho thấy ưu thế vượt trội trong việc phát hiện peptide kháng khuẩn. Bằng cách khai thác hiệu quả tính chất tuần tự của chuỗi peptide, mô hình này mở ra tiềm năng lớn cho việc nghiên cứu và phát triển các phương pháp điều trị kháng khuẩn mới.

Từ kết quả của mô hình SVM, mô hình đề xuất sử dụng Deep Learning, cụ thể là mô hình LSTM, để cải thiện độ chính xác và khả năng dự đoán trong các nghiên cứu peptide kháng khuẩn.

**3. Kết quả mô hình**

**4. Kết luận**

Antimicrobial Peptides (AMPs) là các phân tử tự nhiên với tiềm năng kháng khuẩn, virus và các tác nhân gây bệnh, đóng vai trò quan trọng trong việc phát triển giải pháp đối phó với kháng thuốc kháng sinh (AMR). Trong nghiên cứu này, chúng tôi đã ứng dụng hai mô hình học máy, Support Vector Machine (SVM) và Long Short-Term Memory (LSTM), để dự đoán và tối ưu hóa hoạt tính kháng khuẩn của AMPs trên tập dữ liệu peptide sinh học lớn. Kết quả thực nghiệm cho thấy mô hình LSTM đạt độ chính xác (Accuracy) 90%, vượt trội so với SVM (86%). Ngoài ra, các chỉ số F1-Score, Precision, và Recall cũng khẳng định khả năng của LSTM trong việc xử lý chuỗi dữ liệu tuần tự và tự động trích xuất đặc trưng mà không cần bước xử lý thủ công. Điều này chứng minh tiềm năng của học sâu trong việc nâng cao hiệu quả và độ chính xác của dự đoán peptide kháng khuẩn. Nghiên cứu này không chỉ đặt nền móng cho các phương pháp học máy trong lĩnh vực phát triển thuốc mà còn nhấn mạnh tính ứng dụng thực tiễn của LSTM trong việc rút ngắn thời gian và giảm chi phí nghiên cứu peptide mới. Cụ thể, sự tích hợp học sâu giúp giảm thiểu các quy trình thủ công, tối ưu hóa khả năng khai thác thông tin tuần tự trong peptide, và tăng tốc độ phát hiện các peptide kháng khuẩn tiềm năng. Trong bối cảnh các siêu vi khuẩn kháng thuốc như MRSA, VRE, và CRE ngày càng phổ biến, kết quả nghiên cứu này đóng vai trò quan trọng trong phát triển thuốc kháng sinh thế hệ mới. Việc kết hợp học máy, sinh học phân tử và các phương pháp phân tích hiện đại hứa hẹn mang lại giải pháp bền vững, góp phần đảm bảo sức khỏe cộng đồng và giảm thiểu nguy cơ từ kháng thuốc kháng sinh trong tương lai.

.