

Thiết kế ứng dụng phân tích sai khác trình tự nucleotide gen mã hóa protein P72 của virus dịch tả lợn châu Phi bằng công cụ Vibe-coding lập trình bằng ngôn ngữ tự nhiên - Websim

Souriya Viliddeth¹, Đặng Hữu Anh^{1*}, Mai Thị Ngân¹, Nguyễn Hữu Hân²

¹Học viện Nông nghiệp Việt Nam,

đường Ngô Xuân Quảng, xã Gia Lâm, Hà Nội, Việt Nam

²Trường Đại học Xây dựng Hà Nội,

55 Giải Phóng, phường Bạch Mai, Hà Nội, Việt Nam

Ngày nhận bài 8/7/2025; ngày chuyển phản biện 10/7/2025; ngày nhận phản biện 31/7/2025; ngày chấp nhận đăng 5/8/2025

Tóm tắt:

Bệnh dịch tả lợn châu Phi (DTLCP) đã xuất hiện tại Việt Nam và Lào từ năm 2019. Theo thời gian, virus gây bệnh có những biến đổi về kiểu gen và kiểu hình gây khó khăn cho việc phòng chống bệnh. Để phát hiện sự thay đổi về kiểu gen của virus, các phần mềm tin sinh học được sử dụng phân tích sự sai khác trình tự gen. Tuy nhiên, việc sử dụng phần mềm tin sinh học mất khá nhiều thao tác và yêu cầu người dùng phải có kiến thức về gen để hiểu những sự sai khác đó. Mục đích của nghiên cứu này nhằm thiết kế ứng dụng phân tích sai khác trình tự nucleotide với ví dụ là các trình tự gen mã hóa protein P72 của virus DTLCP ở Lào. Kết quả phân tích gen của ứng dụng được thể hiện bằng văn bản và file Excel bằng tiếng Việt để người dùng dễ hiểu, phục vụ mục đích phân tích sự khác biệt các gen ở dữ liệu đầu vào ứng dụng.

Từ khóa: dịch tả lợn châu Phi, gen, protein P72, ứng dụng phân tích, Vibe-coding, Websim.

Chỉ số phân loại: 1.2, 4.2

Design of application to analyse differences in the nucleotide sequence of the P72 protein-encoding gene of the African swine fever virus using the Vibe-coding tool programming in natural language - Websim

Souriya Viliddeth¹, Huu Anh Dang^{1*}, Thi Ngan Mai¹, Huu Han Nguyen²

¹Vietnam National University of Agriculture,

Ngo Xuan Quang Street, Gia Lam Commune, Hanoi, Vietnam

²Hanoi University of Civil Engineering,

55 Giai Phong Street, Bach Mai Ward, Hanoi, Vietnam

Received 8 July 2025; revised 31 July 2025; accepted 5 August 2025

*Tác giả liên hệ: Email: dhanh@vnua.edu.vn

Abstract:

African swine fever (ASF) has appeared in Vietnam and Laos since 2019. Over time, the virus causing the disease has changed in genotype and phenotype, making it difficult to prevent the disease. To detect changes in the genotype of the virus, bioinformatics software is used to analyse gene sequence differences. However, using bioinformatics software requires numerous operations, and the results require users to have a basic understanding of genes to comprehend those differences. The purpose of this study is to design an application to analyse nucleotide sequence differences with the example of the gene sequences encoding the P72 protein of the ASF virus in Laos. The application's gene analysis results are presented in both text and Excel files in Vietnamese, enabling users to easily understand them and analyse gene differences in the application's input data.

Keywords: African swine fever, analysed application, gene, protein P72, Vibe-coding, Websim.

Classification numbers: 1.2, 4.2

1. Đặt vấn đề

Bệnh DTLCP là bệnh truyền nhiễm nguy hiểm ở lợn, có khả năng gây chết lợn ở mọi lứa tuổi với tỷ lệ lên đến 100%. Bệnh DTLCP đã xuất hiện ở Việt Nam và Lào vào năm 2019 và gây thiệt hại nặng nề cho ngành chăn nuôi lợn ở hai nước [1]. Năm 2024, virus gây bệnh DTLCP có sự biến đổi lớn khi kết hợp giữa genotype 1 và genotype 2 khiến cho các vắc xin được phát triển chống lại bệnh không có hiệu quả [2]. Việc phân tích nhanh sự biến đổi gen của virus DTLCP là cần thiết để cập nhật thông tin virus gây bệnh và cảnh báo nhằm xây dựng các biện pháp ứng phó phù hợp. Để phân tích sự sai khác các nucleotide trong gen của virus DTLCP đã có nhiều phần mềm và ứng dụng tin sinh học trực tuyến hỗ trợ nhà khoa học. Tuy nhiên, các ứng dụng này mất khá nhiều thời gian để thao tác, giao diện bằng tiếng Anh và việc sử dụng các ứng dụng cần thời gian để thành thạo với nhà khoa học không có chuyên môn sâu về tin sinh học. Công nghệ trí tuệ nhân tạo (Artificial Intelligence) đang được ứng dụng rộng rãi và có thể giải quyết nhu cầu phân tích sai khác trình tự gen phục vụ nghiên cứu.

Vibe-coding là một phương pháp phát triển phần mềm phụ thuộc nhiều vào AI để tạo, tinh chỉnh và gỡ lỗi mã dựa trên lời nhắc ngôn ngữ tự nhiên. Các nhà nghiên cứu AI đã phổ biến ý tưởng là các nhà phát triển có thể tập trung vào “Vibe” (chức năng mong muốn của phần mềm), trong khi AI xử lý việc triển khai mã phức tạp. Phương pháp này nhằm mục đích làm cho việc tạo phần mềm dễ tiếp cận hơn và có khả năng nhanh hơn, đặc biệt là đối với các dự án đơn giản. Nghiên cứu này nhằm thiết kế một

ứng dụng phân tích sai khác trình tự nucleotide và đưa ra kết quả là văn bản tiếng Việt trình bày sự sai khác sau khi phân tích.

2. Vật liệu và phương pháp nghiên cứu

2.1. Vật liệu

Trình tự gen sử dụng để phân tích sai khác là trình tự gen P72 của virus gây bệnh DTLCP ở Lào năm 2019 và 2024. File dữ liệu được đăng tải và có thể lấy tại địa chỉ Google Drive [3].

Các phần mềm, công cụ tin học và các chương trình máy tính, website sử dụng cho trích xuất dữ liệu nucleotide và phân tích sai khác gen như công cụ Clustalw, ứng dụng Sequence and Similarity (SIAS) và ứng dụng online <https://trinhtugen.netlify.app/> [4-6].

2.2. Phương pháp nghiên cứu

2.2.1. Phương pháp phân tích sai khác trình tự nucleotide sử dụng các ứng dụng tin sinh học khác nhau

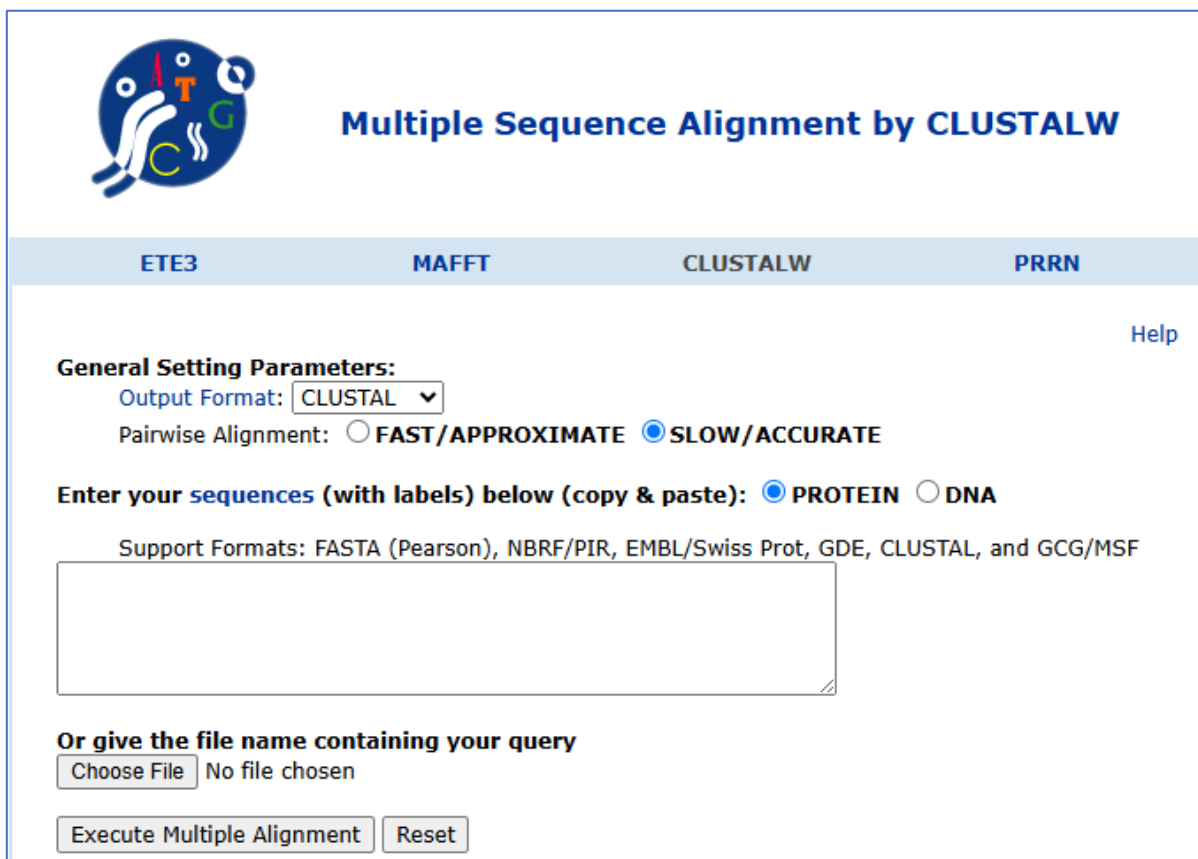
Để phân tích sai khác trình tự nucleotide giữa gen P72 của các chủng virus DTLCP ở Lào, nhiều chương trình online đã được kết hợp, chi tiết các bước thực hiện như sau.

* Bước 1: Lưu trữ trình tự nucleotide.

Gen mã hóa protein P72 của ASFV ở Lào (gồm 7 gen, có 12 variants được thu thập ở hai năm 2019 và 2024) được lưu trữ trình tự nucleotide dưới dạng file Fasta, có thể đọc được dưới dạng text bằng NotePad. File dữ liệu được đăng tải và có thể lấy tại địa chỉ Google Drive [3].

* Bước 2: Tạo các Alignment.

Sử dụng công cụ phân tích các Alignment bằng Clustalw tại website <https://www.genome.jp/tools-bin/clustalw> với lựa chọn Slow/Accurate và các tính năng mặc định của ứng dụng [4]. Chọn nút “Choose File” để upload file chứa trình tự gen dạng Fasta lên công cụ. Kế tiếp, chọn nút “Execute Multiple Alignment” để trích xuất file chứa các Alignment (hình 1). Ở trang kết quả sau khi trích xuất, lựa chọn liên kết “clustalw.aln” để tạo file định dạng đuôi “.aln” và sau đó lựa chọn lần 2 vào liên kết “clustalw.aln” sau khi kết quả đã xuất hiện để tự động tải file Alignment dạng “.aln” về máy tính.



The screenshot shows the 'Multiple Sequence Alignment by CLUSTALW' web interface. At the top, there is a logo on the left and the title 'Multiple Sequence Alignment by CLUSTALW' in blue. Below the title is a navigation bar with four tabs: 'ETE3', 'MAFFT', 'CLUSTALW' (which is selected), and 'PRRN'. On the right side of the interface is a 'Help' link. The main content area is titled 'General Setting Parameters:' and includes the following elements:

- 'Output Format:' with a dropdown menu set to 'CLUSTAL'.
- 'Pairwise Alignment:' with two radio buttons: 'FAST/APPROXIMATE' (unselected) and 'SLOW/ACCURATE' (selected).
- 'Enter your sequences (with labels) below (copy & paste):' with two radio buttons: 'PROTEIN' (selected) and 'DNA' (unselected).
- A text box for entering sequences, with a note above it stating 'Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF'.
- A section titled 'Or give the file name containing your query' with a 'Choose File' button and the text 'No file chosen'.
- At the bottom, there are two buttons: 'Execute Multiple Alignment' and 'Reset'.

Hình 1. Các chức năng Alignment của đa trình tự bằng Clustalw.

* Bước 3: Phân tích sai khác trình tự nucleotide.

Sử dụng ứng dụng SIAS tại website <http://imed.med.ucm.es/Tools/sias.html> [5]. Tại thẻ “Choose File”, lựa chọn file “Clustalw.aln” đã tải về ở bước trên và lựa chọn “Submit” (hình 2). Sau khi chờ một thời gian, kết quả sẽ trả lại một bảng dữ liệu so sánh sự sai khác theo cặp giữa các gen mã hóa protein P72 của DTLCP.

Tools >> Sias

SIAS

Sequence Identity And Similarity

Plea for Support
If you've found SIAS useful for your research, please the SECRETARIA GENERAL DE CIENCIA, TECNOLOGIA E INNOVACION OF SPAIN (sgci@mineco.es, dgi@mineco.es). Thank you.
Your support is crucial to keep the site up and running.

SIAS calculates pairwise sequence identity and similarity from multiple sequence alignments

MSA

INPUT

MSA Name (optional):

Paste your alignment (CLUSTAL, FASTA or GCG/PileUp format)

or Upload a file with the alignment in CLUSTAL, FASTA or GCG/PileUp format

clustalw (1).aln

OPTIONS

IDENTITY (ID) & SIMILARITY (SIM) PERCENTAGE

$$ID|SIM_{\%} = 100 \times \frac{\text{Identical | Similar Residues}}{\text{Sequence Length}}$$

SIMILARITY AMINO ACID GROUPING

Aromatic: ☒ F ☒ Y ☒ W ☐ H

Aliphatic: ☒ V ☒ I ☒ L

Charged positive: ☒ R ☒ K ☒ H

Charged negative: ☒ D ☒ E

Small: ☒ S ☒ T ☐ A ☐ G ☐ C

Polar (Not charged): ☒ N ☒ Q ☐ T ☐ S ☐ Y ☐ C

Hydrophobic: ☐ V ☐ I ☐ L ☐ M ☐ A ☐ F ☐ P ☐ W ☐ G

SEQUENCE LENGTH

☒ Length of smallest sequence ☐ Length of largest sequence

☐ Mean length of sequences ☐ Length of Multiple Sequence Alignment

OPTIONS

NORMALIZED SIMILARITY SCORE

$$S = (\sum M_{ij}) + oP_o + eP_e / \sum M_{ii}$$

Select a matrix:

☒ BLOSUM62 ☐ PAM250 ☐ GONNET

Select gap penalties:

Cost for creating the gap, P_o (0-100)

Cost for extending the gap, P_e (0-100)

Credits & Complaints: Pedro Reche parecheg@med.ucm.es

Hình 2. Phân tích sai khác trình tự nucleotide bằng SIAS.

2.2.2. Sử dụng các câu lệnh để thiết kế và vận hành ứng dụng phân tích sai khác trình tự nucleotide gen mã hóa protein P72 của virus dịch tả lợn châu Phi bằng công cụ Vibe-coding lập trình bằng ngôn ngữ tự nhiên - Websim

* Lệnh trong Websim: Tạo Tab “So sánh các trình tự trong file Fasta”, có chức năng cho phép người dùng upload file.fasta (File.fasta là một định dạng văn bản được sử dụng để biểu diễn trình tự nucleotide hoặc axit amin, thường được sử dụng trong lĩnh vực sinh học phân tử và tin sinh học. Mỗi trình tự bắt đầu bằng một dòng mô tả duy nhất (bắt đầu bằng ký tự “>”), theo sau là các dòng chứa dữ liệu trình tự). Sau khi người dùng upload file.fasta, ứng dụng trích xuất tất cả các trình tự nucleotide trong file.fasta vào nhiều textarea khác nhau. Mỗi trình tự nucleotide có 1 textbox chứa thông tin mô tả của nucleotide (bắt đầu bằng ký tự “>”) và 1 textarea chứa trình tự nucleotide tương ứng. Tạo bảng so sánh trình tự nucleotide giữa các chủng trong file.fasta, bảng này có thể xuất sang file Excel”.

Trong môi trường một ứng dụng Vibe-coding như Websim, lệnh bằng ngôn ngữ tự nhiên được chuyển đổi thành các thành phần phần mềm và logic hoạt động tương ứng. Dưới đây là cách hệ thống sẽ diễn giải và xây dựng chức năng này:

* Bước 1: Khởi tạo cấu trúc giao diện (Tab).

Diễn giải lệnh: Tạo Tab “So sánh các trình tự trong file Fasta”.

Hành động của Websim: Hệ thống Vibe-coding sẽ nhận diện “Tạo tab” là một chỉ thị để thêm một thành phần giao diện dạng Tab vào ứng dụng. Hệ thống sẽ đặt tên hiển thị cho Tab là “So sánh các trình tự trong file Fasta” và tạo ra một vùng làm việc riêng biệt cho Tab này. Về mặt kỹ thuật, hệ thống có thể tạo ra một phần tử HTML (ví dụ: <div>) được quản lý như một tab trong một bộ quản lý tab (Tab panel).

* Bước 2: Triển khai chức năng tải lên File.

Diễn giải lệnh: “...có chức năng cho phép người dùng upload file.fasta...”.

Hành động của Websim:

Thành phần Input: Hệ thống sẽ tự động thêm một thành phần nhập liệu file (ví dụ: <input type="file">) vào vùng Tab mới tạo.

Validation (Kiểm tra định dạng): Hệ thống cấu hình thuộc tính accept=".fasta" (hoặc logic kiểm tra mở rộng hơn nếu được cấu hình mặc định) cho thành phần input này để chỉ chấp nhận các file có đuôi .fasta.

Sự kiện (Event Listener): Một sự kiện “on change” (khi file được chọn) hoặc “on submit” (khi nút tải lên được nhấn) sẽ được gắn vào thành phần input này, kích hoạt bước xử lý tiếp theo.

* Bước 3: Xử lý và hiển thị dữ liệu Fasta.

Diễn giải lệnh: “Sau khi người dùng upload file Fasta, ứng dụng trích xuất tất cả các trình tự nucleotide trong file.fasta vào nhiều textarea khác nhau. Mỗi trình tự

nucleotide có 1 textbox chứa thông tin mô tả của nucleotide (bắt đầu bằng ký tự ">") và 1 textarea chứa trình tự nucleotide tương ứng".

Hành động của Websim: Đây là phần phức tạp nhất, đòi hỏi logic xử lý dữ liệu:

(1) Đọc file: Khi file.fasta được tải lên, hệ thống sẽ đọc toàn bộ nội dung file dưới dạng văn bản.

(2) Phân tích cú pháp Fasta (Fasta Parsing Logic): Hệ thống duyệt qua từng dòng của file; hệ thống sử dụng một cơ chế nhận dạng mẫu (Pattern recognition) hoặc trạng thái (State machine) để phân biệt giữa dòng mô tả (bắt đầu bằng ">") và các dòng trình tự.

Khi gặp >: Dòng này được lưu trữ là tiêu đề (mô tả) của một trình tự mới.

Khi không gặp >: Dòng này được coi là một phần của trình tự hiện tại và sẽ được nối vào trình tự đang được xây dựng.

Quy trình này lặp lại cho đến hết file, thu thập tất cả các cặp (mô tả, trình tự).

(3) Tạo giao diện động (Dynamic UI Generation).

Đối với mỗi cặp (mô tả, trình tự) đã được phân tích, hệ thống sẽ tự động tạo ra một cặp phần tử giao diện mới:

Một `<input type="text">` hoặc `<textarea>` nhỏ hơn (để đóng vai trò là textbox) và điền nội dung mô tả vào đó.

Một `<textarea>` lớn hơn và điền toàn bộ trình tự nucleotide tương ứng vào đó.

Các cặp phần tử này sẽ được sắp xếp một cách có hệ thống trên giao diện tab, có thể là theo chiều dọc hoặc trong một lưới, đảm bảo mỗi trình tự được hiển thị rõ ràng.

(4) Cập nhật Trạng thái: Sau khi xử lý và hiển thị, ứng dụng có thể cập nhật trạng thái (ví dụ: "Tìm thấy X trình tự", như trong ví dụ đã cung cấp).

* Bước 4: Triển khai chức năng so sánh trình tự và hiển thị bảng.

Diễn giải lệnh: "tạo bảng so sánh trình tự nucleotide giữa các chủng trong file fasta".

Hành động của Websim: Đây là một bước phức tạp, đòi hỏi xử lý dữ liệu và thuật toán:

(i) Thu thập dữ liệu nguồn: Hệ thống sẽ sử dụng tất cả các trình tự nucleotide đã được trích xuất và lưu trữ ở Bước 3. Giả sử N là tổng số trình tự được tìm thấy.

(ii) Lựa chọn và thực thi thuật toán so sánh trình tự: Với lệnh "so sánh các trình tự", Websim sẽ tự động kích hoạt một thuật toán so sánh trình tự cơ bản và phổ biến trong tin sinh học.

(5) So sánh theo cặp (Pairwise Comparison): Tính toán độ tương đồng (ví dụ: percentage identity) giữa mỗi cặp trình tự.

- Đối với các trình tự có độ dài bằng nhau: Có thể sử dụng khoảng cách Hamming (Hamming distance) hoặc tỷ lệ phần trăm không trùng khớp (percentage mismatch).

- Đối với các trình tự có độ dài khác nhau hoặc cần phát hiện các đoạn tương đồng cục bộ/toàn cục: Websim có thể sử dụng các thuật toán căn chỉnh trình tự đơn giản hóa như thuật toán Needleman-Wunsch (căn chỉnh toàn cục) hoặc Smith-Waterman (căn chỉnh cục bộ) để tìm ra căn chỉnh tối ưu và từ đó tính toán percentage identity.

(6) Đầu ra thuật toán: Kết quả sẽ là một ma trận đối xứng ($N \times N$) hiển thị mức độ tương đồng (hoặc khoảng cách) giữa mọi cặp trình tự. Ví dụ: [Trình tự A] vs [Trình tự B] = 95% tương đồng.

(7) Tạo cấu trúc bảng giao diện động (Dynamic UI Table Generation): Hệ thống sẽ tạo một phần tử bảng HTML (ví dụ: `<table class="comparison-table">`) trên giao diện người dùng, đặt trong vùng của tab "So sánh các trình tự trong file Fasta".

(8) Cấu trúc bảng:

- Dòng tiêu đề (header row) sẽ chứa tên/mô tả của tất cả các trình tự (ví dụ: "Chủng 1", "Chủng 2" ...).

- Cột đầu tiên của mỗi hàng sẽ chứa tên/mô tả của trình tự tương ứng.

- Các ô giao nhau sẽ hiển thị kết quả so sánh giữa trình tự hàng và trình tự cột (ví dụ: "95%", "88%" ...).

- Các ô trên đường chéo chính (ví dụ: A vs A) sẽ hiển thị 100% (hoặc giá trị tương đương).

- Tối ưu hiển thị: Nếu số lượng trình tự lớn, Websim sẽ tự động áp dụng các thuộc tính CSS hoặc JavaScript để làm cho bảng có thể cuộn được (scrollable) và có thể tìm kiếm/sắp xếp để cải thiện trải nghiệm người dùng.

* Bước 5: Triển khai chức năng xuất bảng sang Excel.

Diễn giải lệnh: "bảng này có thể xuất sang file excel".

Hành động của Websim: Để hỗ trợ việc trích xuất dữ liệu, Websim sẽ thực hiện các hành động sau:

(1) Thêm thành phần Xuất (Export Component): Websim sẽ tự động thêm một nút (ví dụ: `<button id="exportToExcel">Xuất ra Excel</button>`) vào giao diện người dùng, đặt gần bảng so sánh.

(2) Gắn sự kiện (Event Listener): Một sự kiện "click" sẽ được gắn vào nút "Xuất ra Excel".

(3) Chuyển đổi Dữ liệu và Định dạng (Data Transformation & Formatting): Khi nút được nhấn, hệ thống sẽ thu thập toàn bộ dữ liệu hiện có trong bảng so sánh (bao gồm tiêu đề và các giá trị so sánh).

(4) Chuyển đổi dữ liệu sang định dạng Excel: Websim sẽ chuyển đổi dữ liệu thành một định dạng tương thích với Microsoft Excel như CSV (Comma Separated Values): Đây là một định dạng văn bản đơn giản, được phân tách bằng dấu phẩy (hoặc dấu chấm phẩy) mà Excel có thể mở và diễn giải dễ dàng. Websim có thể sinh mã JavaScript để tạo chuỗi CSV và kích hoạt tải xuống.

(5) Kích hoạt Tải xuống File (File Download Initiation): Sau khi file CSV được tạo, hệ thống sẽ gửi dữ liệu đến trình duyệt của người dùng, kích hoạt một cửa sổ tải xuống để người dùng lưu file vào máy tính của họ. Tên file mặc định là: “bang_so_sanh_trinh_tu.csv”.

* Các yếu tố ngầm định:

Mặc dù lệnh rất ngắn gọn, hệ thống Vibe-coding thông minh sẽ tự động tích hợp các yếu tố sau để đảm bảo ứng dụng hoạt động hiệu quả và thân thiện:

- Xử lý lỗi (Error Handling): Nếu file không đúng định dạng (ví dụ: không phải Fasta thực sự), hệ thống sẽ hiển thị thông báo lỗi. Nếu file trống hoặc không có trình tự nào được tìm thấy hệ thống sẽ hiển thị thông báo lỗi.

- Phản hồi người dùng (User Feedback): Hiển thị thông báo khi file đã được tải lên thành công và hiển thị số lượng trình tự được tìm thấy.

- Hiệu suất (Performance): Đối với các file Fasta rất lớn, hệ thống sẽ cần tối ưu hóa việc đọc và phân tích để tránh treo ứng dụng.

- Styling và Layout: Vibe-coding sẽ áp dụng các kiểu mặc định để các textbox và textarea có kích thước và bố cục hợp lý, dễ đọc.

3. Kết quả và bàn luận

3.1. Kết quả thực hiện phương pháp phân tích sai khác trình tự nucleotide sử dụng các ứng dụng tin sinh học khác nhau

Từ sau bước 2, tại website kết quả trích xuất Alignment, những dấu “*” bị xóa bỏ biểu thị vị trí sai khác chi tiết về nucleotide của các gen (hình 3).

```
clustalw.aln

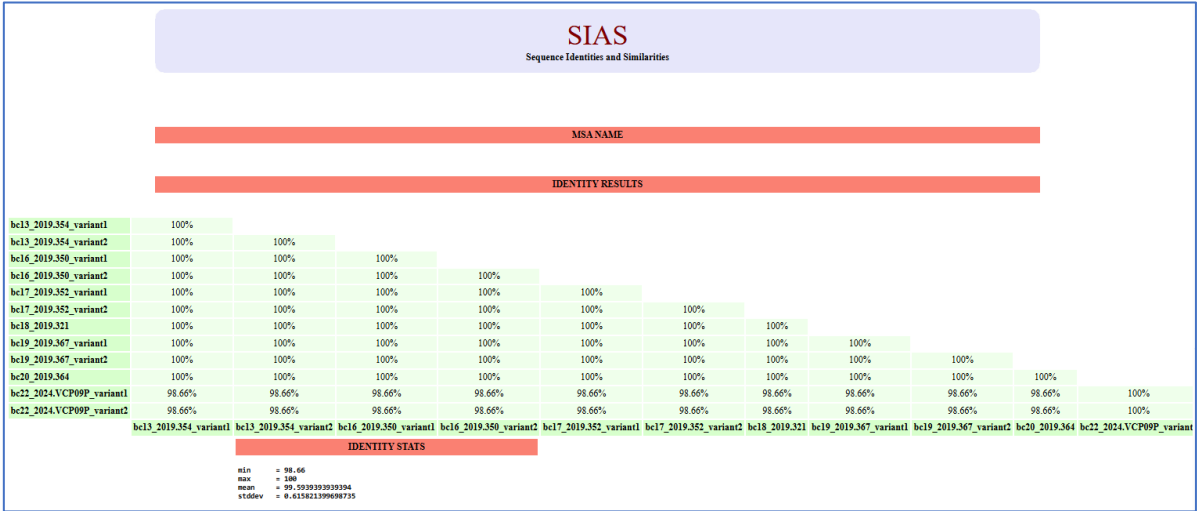
CLUSTAL 2.1 multiple sequence alignment

bc13_2019.354_variant1      ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc13_2019.354_variant2      ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc16_2019.350_variant1      ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc16_2019.350_variant2      ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc17_2019.352_variant1      ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc17_2019.352_variant2      ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc18_2019.321               ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc19_2019.367_variant1      ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc19_2019.367_variant2      ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc20_2019.364               ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc22_2024.VCP09P_variant1    ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
bc22_2024.VCP09P_variant2    ATGGCATCAGGAGGAGCTTTTGTCTTATTGCTAACGATGGGAAGGCCGA
*****

bc13_2019.354_variant1      CAAGATTATATTGGCCCAAGACTTGCTGAATAGCAGGATCTCTAACATTA
bc13_2019.354_variant2      CAAGATTATATTGGCCCAAGACTTGCTGAATAGCAGGATCTCTAACATTA
bc16_2019.350_variant1      CAAGATTATATTGGCCCAAGACTTGCTGAATAGCAGGATCTCTAACATTA
bc16_2019.350_variant2      CAAGATTATATTGGCCCAAGACTTGCTGAATAGCAGGATCTCTAACATTA
bc17_2019.352_variant1      CAAGATTATATTGGCCCAAGACTTGCTGAATAGCAGGATCTCTAACATTA
bc17_2019.352_variant2      CAAGATTATATTGGCCCAAGACTTGCTGAATAGCAGGATCTCTAACATTA
bc18_2019.321               CAAGATTATATTGGCCCAAGACTTGCTGAATAGCAGGATCTCTAACATTA
bc19_2019.367_variant1      CAAGATTATATTGGCCCAAGACTTGCTGAATAGCAGGATCTCTAACATTA
bc19_2019.367_variant2      CAAGATTATATTGGCCCAAGACTTGCTGAATAGCAGGATCTCTAACATTA
bc20_2019.364               CAAGATTATATTGGCCCAAGACTTGCTGAATAGCAGGATCTCTAACATTA
bc22_2024.VCP09P_variant1    CAAGATTATATTGGCCCAAGACTTGCTTAATAGCAGGATTTCTAACATTA
bc22_2024.VCP09P_variant2    CAAGATTATATTGGCCCAAGACTTGCTTAATAGCAGGATTTCTAACATTA
*****
```

Hình 3. Vị trí sai khác nucleotide được hiển thị trên website.

Kết quả “Submit” sau bước 3 ở ứng dụng SIAS trả lại một bảng dữ liệu so sánh sự sai khác theo cặp giữa các gen mã hóa protein P72 của DTLCP (hình 4).



Hình 4. Kết quả phân tích sai khác trình tự nucleotide.

Kết quả hình 4 cho thấy, gen mã hóa protein P72 của các chủng virus DTLCP ở Lào năm 2024 có sự sai khác với các chủng năm 2019. Sự tương đồng của các chủng ở hai giai đoạn khác nhau chỉ là 98,66%. Các chủng năm 2019 có sự tương đồng 100% khi thực hiện phân tích sai khác bằng các ứng dụng tin sinh học khác nhau.

Nhược điểm của việc sử dụng các phương pháp phân tích sai khác này bao gồm: (i) Người dùng phải thực hiện nhiều ứng dụng, website và thao tác; (ii) Kết quả thu được không phân tích chi tiết ý nghĩa của các vị trí sai khác và vị trí tương đồng một cách rõ ràng (chỉ thể hiện ở các dấu *), gây khó khăn cho người dùng ứng dụng.

3.2. Kết quả của việc vận hành ứng dụng phân tích sai khác trình tự nucleotide gen mã hóa protein P72 của virus dịch tả lợn châu Phi bằng công cụ Vibe-coding lập trình bằng ngôn ngữ tự nhiên - Websim

Sau khi thiết lập ứng dụng bằng công cụ Vibe-coding, ứng dụng được đưa lên trình duyệt ở dạng website tại địa chỉ: <https://trinhtugen.netlify.app/> [6]. Ứng dụng có nhiều chức năng trong đó có chức năng “So sánh các trình tự trong file.fasta” (hình 5). Khi lựa chọn chức năng này, ứng dụng sẽ tự động phân tích dữ liệu gen P72 có tất cả 12 trình tự và biểu hiện trên website (hình 6). Người dùng chỉ cần tải file dữ liệu trình tự gen lên và lựa chọn nút “so sánh tất cả các trình tự”.

Hình 5. Giao diện của website và chức năng phân tích trình tự DNA.

Hình 6. Trình tự các đoạn gen được trình bày ở giao diện website.

Sau khi lựa chọn “so sánh tất cả các trình tự”, ứng dụng cần một khoảng thời gian vài phút để phân tích và so sánh. Kết quả của việc phân tích và so sánh trình tự gen được thể hiện bằng văn bản và bằng bảng dữ liệu so sánh. Rất nhiều thông tin được ứng dụng phân tích chi tiết, giúp người dùng biết được kết quả so sánh tổng quan, so sánh từng cặp trình tự. Ứng dụng cũng đưa ra thông tin về các vùng bảo tồn chung, phân tích cây phát sinh loài cũng như nêu ra ý nghĩa sinh học tiềm năng của các mối quan hệ giữa các trình tự (hình 7).

Kết Quả Phân Tích Trình Tự FASTA

Dưới đây là phân tích và so sánh các trình tự DNA được cung cấp:

1. Thông tin tổng quan về các trình tự

Chúng ta có tổng cộng 12 trình tự DNA.

- Định dạng:** Tất cả các trình tự được cung cấp dưới định dạng FASTA, bắt đầu bằng một dòng tiêu đề (ký hiệu >) theo sau là chuỗi nucleotide.
- Độ dài:** Sau khi kiểm tra sơ bộ, tất cả 12 trình tự đều có cùng độ dài chính xác là **1656 base pair (bp)**.
- Ký tự:** Các chuỗi chỉ bao gồm các nucleotide chuẩn: A, T, G, C.
- Điểm khởi đầu và kết thúc tiềm năng:** Tất cả các trình tự đều bắt đầu bằng **ATG** (codon khởi đầu tiềm năng) và kết thúc bằng **TAA** (codon kết thúc tiềm năng). Điều này gợi ý rằng chúng có thể là các vùng mã hóa protein (coding sequences - CDS) hoặc các vùng liên quan đến gen.

2. So sánh từng cặp trình tự

Sau khi thực hiện so sánh chi tiết giữa tất cả các trình tự, chúng ta có thể chia chúng thành hai nhóm chính dựa trên sự giống nhau hoàn toàn:

- Nhóm 1 (Nhóm gốc/phổ biến):** Bao gồm 10 trình tự đầu tiên:
 - bc13_2019.354_variant1
 - bc13_2019.354_variant2
 - bc16_2019.350_variant1
 - bc16_2019.350_variant2
 - bc17_2019.352_variant1
 - bc17_2019.352_variant2
 - bc18_2019.321
 - bc19_2019.367_variant1
 - bc19_2019.367_variant2
 - bc20_2019.364Tất cả 10 trình tự trong Nhóm 1 này đều giống hệt nhau 100%.

Hình 7. Kết quả phân tích trình tự Fasta.

Tại bảng so sánh chi tiết trình tự, sự khác biệt được trình bày dưới dạng số. Ứng dụng cho phép xuất dữ liệu sang Excel (bảng 1). Kết quả cho thấy, khi so sánh giữa các gen P72 của virus DTLCP ở Lào năm 2019 có chỉ số 0,999-1, nghĩa là có sự tương đồng đến 100%. Đối với những gen P72 của virus DTLCP ở Lào năm 2024, có sự khác biệt (tương đồng chỉ đạt 99,5%).

Bảng 1. So sánh trình tự gen P72 của virus dịch tả lợn châu Phi ở Lào năm 2019 và 2024.

Seq	Seq1	Seq2	Seq3	Seq4	Seq5	Seq6	Seq7	Seq8	Seq9	Seq10	Seq11	Seq12
Seq1	1	1	0,999	1	1	1	1	1	1	1	0,995	0,995
Seq2	0,999	1	1	1	1	0,999	1	1	1	0,999	0,995	0,995
Seq3	1	1	1	1	1	1	1	1	1	0,999	0,995	0,995
Seq4	0,999	0,999	1	1	1	0,999	1	1	1	1	0,995	0,995
Seq5	1	1	1	1	1	1	0,999	1	1	1	0,995	0,995
Seq6	1	1	1	1	1	1	1	1	1	1	0,995	0,995
Seq7	0,999	0,999	1	1	1	1	1	0,999	1	1	0,995	0,995

Seq8	1	1	0,999	0,999	1	1	1	1	1	1	0,995	0,995
Seq9	0,999	1	1	1	1	0,999	1	0,999	1	1	0,995	0,995
Seq10	1	0,999	0,999	1	1	1	1	1	1	1	0,995	0,995
Seq11	0,995	0,995	0,995	0,995	0,995	0,995	0,995	0,995	0,995	0,995	1	1
Seq12	0,995	0,995	0,995	0,995	0,995	0,995	0,995	0,995	0,995	0,995	1	1

Kết quả ở bảng 1 có một số khác biệt với kết quả khi sử dụng các phần mềm tin sinh học khác nhau để phân tích trình tự gen. Nguyên nhân có thể do việc thiết lập thuật toán của các phần mềm này đã loại bỏ các đoạn trình tự “nhiều”. Mặc dù vậy, kết quả sử dụng ứng dụng có kết quả dò chính xác từng nucleotide trong trình tự gen đã nhập ở dữ liệu đầu vào và kết quả này là đáng tin cậy. Ưu điểm của ứng dụng là phân tích chi tiết bằng văn bản để người dùng có thể hiểu chi tiết về sự khác biệt giữa các gen được so sánh.

4. Kết luận

Việc sử dụng công cụ tạo ứng dụng bằng ngôn ngữ tự nhiên (Vibe-coding) như Websim thể hiện sức mạnh của công cụ lập trình với công nghệ AI. Điều này giúp đẩy nhanh tốc độ phát triển và giảm thiểu lỗi do lập trình thủ công. Ứng dụng phân tích sai khác trình tự gen bước đầu đã hoạt động tốt và có khả năng phân tích sai khác các nucleotide chính xác tương đương với việc kết hợp nhiều phần mềm, ứng dụng tin sinh học trước kia. Kết quả đầu ra của phần mềm là văn bản phân tích bằng tiếng Việt, thân thiện hơn với người dùng, phục vụ thiết thực cho công tác nghiên cứu sự biến đổi gen nói chung và sự biến đổi gen của virus DTLCP nói riêng.

TÀI LIỆU THAM KHẢO

- [1] B.T.T. Nga, L.V. Phan, B.T.A. Dao, et al. (2020), “Clinical and pathological features of African swine fever in the first outbreaks in Vietnam”, *Vietnam J. Agri. Sci.*, **18(7)**, pp.485-494 (in Vietnamese).
- [2] N.V. Diep, N.V. Duc, N.T. Ngọc, et al. (2024), “Genotype II live-attenuated asfv vaccine strains unable to completely protect pigs against the emerging recombinant ASFV genotype i/ii strain in Vietnam”, *Vaccines*, **12(10)**, DOI: 10.3390/vaccines12101114.
- [3] https://drive.google.com/file/d/1SPSjg_IL-q27ztwsquTQfvamseKSPR6s/view?usp=sharing, accessed 8 July 2025.
- [4] Multiple Sequence Alignment by CLUSTALW (2025), <https://www.genome.jp/tools-bin/clustalw>, accessed 8 July 2025.
- [5] Sequence Identity and Similarity (2025), <http://imed.med.ucm.es/Tools/sias.html>, accessed 8 July 2025.
- [6] DNA Sequence Analysis (2025), <https://trinhtugen.netlify.app/>, accessed 8 July 2025.