ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

---oo0oo---

NGUYỄN HUY TÌNH

PHÁT TRIỂN VÀ ĐÁNH GIÁ CÁC PHƯƠNG PHÁP ƯỚC LƯỢNG MÔ HÌNH THAY THẾ AXIT AMIN CHO CÁC TẬP DỮ LIỆU CÓ KÍCH THƯỚC LỚN

Ngành : Khoa học Máy tính

Mã số : 9480101

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

HÀ NỘI - 2025

Công trình được hoàn thành tại: Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội

Người hướng dẫn khoa học: 1. GS.TS. Lê Sỹ Vinh

2. TS. Đặng Cao Cường

Phản biện: PGS.TS. Huỳnh Thị Thanh Bình

Phản biện: PGS.TS. Nguyễn Long Giang

Phản biện: PGS.TS. Vũ Thị Thơm

Luận án được bảo vệ trước Hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại

……..

vào hồi … giờ … ngày … tháng … năm 2025.

NGHIÊN CỨU SINH **CÁN BỘ HƯỚNG DẪN**

XÁC NHẬN CỦA ĐƠN VỊ ĐÀO TẠO

Có thể tìm hiểu luận án tại:

- Thư viện Quốc Gia Việt Nam
- Trung tâm Thông tin - Thư viện, Đại học Quốc gia Hà Nội

Lời mở đầu

1. Đặt vấn đề

Tin sinh học là một lĩnh vực nghiên cứu đa ngành với sự kết hợp của công nghệ thông tin, sinh học phân tử và toán thống kê. Dữ liệu chính và phổ biến trong tin sinh học là các trình tự DNA và trình tự protein. Cùng với sự phát triển của các máy giải trình tự thế hệ mới, số lượng trình tự này đang tăng liên tục cả về số lượng lẫn độ chính xác. Qua đó, một lượng lớn dữ liệu được cung cấp một cách thường xuyên tạo cơ sở giúp cho các nhà khoa học nghiên cứu đưa ra các kết quả mới để giải quyết những bài toán khó hiện tại đang gặp phải.

Các bài toán cơ bản trong tin sinh học tiến hóa liên quan đến trình tự protein bao gồm: sắp hàng đa trình tự, tìm kiếm trình tự tương đồng hay xây dựng cây phân loài. Trong đó bài toán xây dựng cây phân loài được coi là bài toán trung tâm trong tin sinh học tiến hóa. Để xây dựng cây phân loài ta cần mô hình hóa quá trình tiến hóa bởi một mô hình toán học dạng ma trận để giải thích một cách gần đúng nhất (có xác suất cao nhất) quá trình thay thế giữa các axit amin trong trình tự protein, các mô hình này được gọi là mô hình thay thế axit amin, sau đây gọi tắt là mô hình. Tối ưu các mô hình thay thế cũng như ước lượng các mô hình thay thế mới sẽ giúp quá trình nghiên cứu sát với thực tế tiến hóa và mang lại các kết quả tốt hơn.

2. Mục đích nghiên cứu

Các mô hình thay thế axit amin tùy vào thuộc tính về mặt thời gian mà chúng ta cần phải ước lượng 208 tham số tự do (thời gian thuận nghịch: tốc độ biến đổi giống nhau theo cả hai hướng tiến và lùi) hay 379 tham số tự do (thời gian không thuận nghịch: tốc độ biến đổi khác nhau theo hai hướng tiến và lùi). Các mô hình có thể tồn tại ở dạng đơn ma trận hay kết hợp của nhiều ma trận khác nhau. Xây dựng mô hình thay thế axit amin là bài toán khó, thường phải trải qua nhiều bước phức tạp và tốn kém cả về mặt thời gian lẫn chi phí tính toán. Các phương pháp truyền thống để giải quyết bài toán này bao gồm:

1. Phương pháp đếm: đây là phương pháp cổ điển được dùng để ước lượng các mô hình như PAM, Dayhoff.
2. Phương pháp cực đại hợp lý (maximum likelihood): đây là phương pháp phổ biến và mang lại kết quả tốt hơn. Với phương pháp này, trước hết chúng ta cần xây dựng cây phân loài dựa trên các mô hình chung như LG, sau đó dựa trên cây phân loài đã có, mô hình mới sẽ được ước lượng nhằm cực đại hóa giá trị hợp lý của cây.

Việc ước lượng dựa trên phương pháp cực đại hợp lý có thể dựa trên các đặc điểm sinh học khác nhau và tạo ra các mô hình đa ma trận khác nhau. Rất nhiều công trình nghiên cứu đã đề xuất các phương pháp tối ưu quá trình ước lượng dựa trên phương pháp cực đại hợp lý như XRate, RAxML. Gần đây, nhóm tác giả Minh Bùi và cộng sự đề xuất QMaker, Cường Đặng và cộng sự đã đề xuất nQMaker, đây là 2 phương pháp mới được dùng để tự động ước lượng mô hình biến đổi của trình tự protein. Với các phương

pháp này, nhiều mô hình đã được ước lượng từ các bộ dữ liệu sắp hàng khác nhau. Từ đó, một số câu hỏi nghiên cứu được đặt ra như:

- Hiệu quả của mô hình ước lượng thay đổi ra sao tương ứng theo dữ liệu hay số lượng dữ liệu cần thiết để ước lượng mô hình đạt yêu cầu về hiệu năng?
- Ngoài ra, cách tiếp cận dựa trên tính chất thời gian thuận nghịch cũng như việc dùng đơn ma trận để mô hình hóa quá trình tiến hóa có phù hợp với thực tế?

Tiếp theo, việc lựa chọn (hay ước lượng) nhanh mô hình thay thế axit amin cho một sắp hàng bất kì cho trước là vấn đề quan trọng. Do số lượng mô hình ngày càng nhiều nên khi nghiên cứu một dữ liệu sắp hàng bất kỳ, chúng ta cần tìm ra mô hình phù hợp nhất với dữ liệu đó. Thông thường, để chọn được mô hình hợp lý nhất chúng ta có thể sử dụng phương pháp dựa trên tiêu chuẩn cực đại hợp lý, tiêu biểu là phương pháp ModelFinder hay ModelTest-NG. Ngoài ra, để giải quyết bài toán này, nhiều nhà khoa học cũng sử dụng các mạng học sâu để tối ưu hơn về mặt thời gian. Đây là một hướng đi mới trong tin sinh học tiến hóa. Gần đây, với phương pháp này các nhà khoa học đã đề xuất các mạng học sâu để dự đoán mô hình cho dữ liệu DNA.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của luận án là các tập dữ liệu sắp hàng axit amin có kích thước lớn, với dữ liệu thật có thể lên đến hàng chục nghìn sắp hàng, với dữ liệu mô phỏng lên đến hàng triệu sắp hàng.

Luận án tập trung vào việc mô hình quá hóa trình tiến hóa của sắp hàng axit amin dựa trên các đặc tính sinh học như sự không đồng nhất về tốc độ tiến hóa hay tính không thuận nghịch về mặt thời gian trong quá trình tiến hóa của các vị trí trên sắp hàng. Luận án cũng đi sâu vào cách ước lượng mô hình thay thế sử dụng nhiều ma trận tốc độ biến đổi khác nhau. Ngoài ra, luận án nghiên cứu các phương pháp dựa trên học sâu nhằm lựa chọn nhanh các mô hình phù hợp cho sắp hàng với thời gian tối ưu.

4. Phương pháp nghiên cứu

Phương pháp được luận án sử dụng là sự kết hợp giữa nghiên cứu lý thuyết và làm các thực nghiệm, gồm các bước cơ bản như sau:

- Nghiên cứu cơ sở lý thuyết nhằm đạt được những nền tảng căn bản của tin sinh học tiến hóa.
- Khảo sát, nghiên cứu các công trình khoa học có liên quan đến nội dung nghiên cứu đồng thời thu thập các bộ dữ liệu tiêu chuẩn.
- Đề xuất các phương pháp hoặc mô hình mới theo nội dung của luận án.
- Tiến hành các thực nghiệm đánh giá phương pháp và mô hình mới trên các bộ dữ liệu đã thu thập, phân tích kết quả nhằm cải tiến phương pháp và mô hình đã được đề xuất.

5. Những kết quả và đóng góp chính của luận án

Các kết quả của luận án đã được công bố trong 05 bài báo ở tạp chí SCI quốc tế (3 bài Q1, 2 bài Q2) và 01 bài báo ở hội nghị quốc tế uy tín. Cụ thể như sau:

1. Đề xuất một quy trình đánh giá các phương pháp ước lượng mô hình dựa trên tiêu chuẩn cực đại hợp lý. Các kết quả được công bố trong công trình [CT1] và [CT2].
2. Đề xuất các mô hình thay thế axit amin mới gồm nhiều ma trận trong đó có sử dụng các đặc tính về mặt sinh học là tốc độ biến đổi tại các vị trí trên trình tự và thuộc tính thời gian không thuận nghịch. Các kết quả được công bố trong công trình [CT3], [CT4] và [CT6].
3. Xây dựng một mạng học sâu để lựa chọn nhanh mô hình thay thế axit amin của một sắp hàng đa trình tự bất kì. Các kết quả được công bố trong công trình [CT5].

6. Bố cục của luận án

Ngoài phần kết luận, luận án được tổ chức như sau:

Chương 1 giới thiệu tổng quan về trình tự DNA, trình tự axit amin và các phép biến đổi trên trình tự axit amin. Sau đó luận án giới thiệu về bài toán mô hình hoá quá trình biến đổi axít amin và bài toán ước lượng mô hình thay thế axít amin.

Chương 2 đề xuất một quy trình đánh giá các phương pháp ước lượng mô hình thay thế axit amin. Luận án đưa ra các tiêu chí đánh giá mô hình cũng như đề xuất các tiêu chuẩn về mặt dữ liệu nhằm giúp các nhà nghiên cứu vừa tiết kiệm thời gian cũng như đạt được hiệu quả mong muốn.

Chương 3 của luận án giới thiệu phương pháp ước lượng mô hình thay thế axít amin sử dụng nhiều ma trận. Phương pháp này được thể hiện dưới dạng một phần mềm và có thể tự động ước lượng một số lượng ma trận tùy ý. Đồng thời luận án cũng ước lượng và đề xuất các mô hình thay thế axit amin sử dụng đa ma trận mới được ước lượng dựa trên các bộ dữ liệu HSSP và bộ dữ liệu các loài thực vật.

Chương 4 trình bày một mạng học sâu và phương pháp để trích xuất đặc trưng dữ liệu nhằm giải quyết bài toán lựa chọn mô hình thay thế axit amin phù hợp nhất cho một sắp hàng đa trình tự bất kì cho trước.

Chương 1. Giới thiệu chung

1.1. Một số khái niệm cơ bản

1.1.1. Trình tự DNA và axit amin

Deoxyribo Nucleic Axit (viết tắt DNA) là phân tử sinh học mang thông tin di truyền được cấu tạo từ các phân tử nhỏ hơn gọi là nucleotit. Phân tử DNA được cấu tạo bởi hai trình tự nucleotit và có cấu trúc dạng xoắn kép, mỗi trình tự được tạo thành từ bốn loại nucleotit là: Adenine (A), Thymine (T), Cytosine (C), và Guanine (G). Protein hay còn gọi là chất đạm, là những phân tử sinh học chứa một hay nhiều mạch axit amin liên kết với nhau bởi liên kết peptit. Có 20 loại axit amin khác nhau được mã hóa bởi bộ gene.

1.1.2. Sự biến đổi trên trình tự axit amin

Có ba loại biến đổi chính trên trình tự protein là:

- **Xoá**: một hoặc một số axit amin bị xoá khỏi trình tự, độ dài của trình tự giảm đi.
- **Chèn**: một hoặc một số axit amin được chèn vào trình tự, độ dài trình tự tăng lên.
- **Thay thế**: một axit amin này bị thay thế bằng một axít amin khác, độ dài của trình tự không thay đổi.

1.1.3. Sắp hàng các trình tự axit amin tương đồng

Hai trình tự axit amin gọi là tương đồng nếu chúng cùng tiến hóa từ một trình tự axit amin tổ tiên. Sự khác biệt của các trình tự tương đồng chủ yếu do các biến đổi trong quá trình tiến hóa và do vậy làm chúng khác nhau cả về nội dung lẫn độ dài.

1.1.4 Cây phân loài

Trong nghiên cứu tin sinh học, cây phân loài (cây tiến hóa, cây phả hệ) là một loại biểu đồ dạng cây nhị phân, dùng để mô hình hóa mối quan hệ phát sinh loài hay lịch sử tiến hóa của một nhóm các sinh vật.

Trong luận án này, phương pháp cực đại hợp lý được sử dụng để giải quyết các bài toán về xây dựng mô hình thay thế axit amin và cây phân loài.

1.2 Bài toán mô hình hóa quá trình thay thế axit amin

1.2.1 Mô hình Markov

Quá trình thay thế axit amin tại một vị trí bất kì trên trình tự protein được xem là ngẫu nhiên và liên tục theo thời gian. Quá trình này có thể được mô hình hóa bằng chuỗi Markov [15, 16], với các thuộc tính căn bản là:

- Tính liên tục: sự biến đổi diễn ra liên tục tại bất kì thời điểm nào.

- Tính độc lập: sự biến đổi chỉ phụ thuộc vào trạng thái hiện tại, không phụ thuộc vào trạng thái trong quá khứ.

Do trước đây năng lực tính toán của các hệ thống phần cứng còn nhiều hạn chế, nên nhằm đơn giản hóa tối đa việc ước lượng thì các nhà khoa học thường sử dụng thêm các thuộc tính khác như sau:

- Tính đồng nhất: tốc độ biến đổi giữa các axit amin không đổi trong suốt quá trình biến đổi.
- Tính ổn định: tần số của các axit amin là không đổi trong suốt quá trình biến đổi.
- Tính thuận nghịch: tốc độ biến đổi là đồng nhất theo các hướng tiến và lùi.

1.2.2 Mô hình thay thế nucleotit

Do dữ liệu DNA chỉ bao gồm 4 loại ký tự là A, T, G và C nên chúng ta xem xét quá trình thay thế nucleotit trong chuỗi DNA với 4 trạng thái. Để mô tả quá trình thay thế này, chúng ta có thể sử dụng ma trận tốc độ biến đổi tức thì Q.

Do tính chất ổn định của mô hình nên tổng số biến đổi từ trạng thái *i* sang *j* sau khoảng thời gian t là bằng 0. Nghĩa là chúng ta có:

$$Q_{ii} = -\sum_{\dots} Q_{ij} \qquad (1.1)$$

1.2.3 Mô hình thay thế axit amin

Tương tự như trên, xét quá trình Markov với tập trạng thái **S** = {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}, đây là tập gồm 20 loại axit amin. Gọi $\Pi = \{\pi_i\}$ với i = 1, … 20 là véc tơ tần số xuất hiện của 20 axit amin, như vậy $\sum_{i=1}^{20} \pi_i = 1$ và các π_i không đổi theo thời gian. Ta cũng kí hiệu $R = \{r_{ij}\}$ là ma trận hệ số hoán đổi trong đó r_{ij} là hệ số hoán đổi giữa hai axit amin i và j.

Chúng ta ký hiệu $Q = \{q_{ij},\ i \in \boldsymbol{S},\ j \in \boldsymbol{S}\}$ là ma trận tốc độ biến đổi tức thì có kích thước $20 * 20$, trong đó q_{ij} là tốc độ biến đổi tức thì từ axit amin i sang axit amin j. Ma trận Q có thể được biểu diễn bởi ma trận $R = \{r_{ij}\}$ và vectơ $\Pi = \{\pi_i\}$ như sau:

$$q_{ij} = \begin{cases} \pi_j r_{ij} & nếu\ i \neq j \\ -\sum_{x \neq i} q_{ix} & nếu\ i = j \end{cases}$$

hoặc có thể được viết gọn dưới dạng: $Q = \Pi * R$ trong đó, Π là vector tần suất còn R được gọi là ma trận hệ số hoán đổi. Do tính chất thuận nghịch ta có $r_{ij} = r_{ji}$ hay ma trận hệ số hoán đổi R đối xứng qua đường chéo chính. Với lượng tham số lớn nên chúng ta cần phải dùng các tập dữ liệu có kích thước lớn để ước lượng mô hình đạt hiệu quả tốt nhất.

1.2.4 Bài toán ước lượng mô hình thay thế axit amin

1.2.4.1 Phát biểu bài toán

Dữ liệu đầu vào: Cho một tập *N* các sắp hàng $\mathbf{D} = \{D_1, \dots, D_N\}$. Mỗi sắp hàng bao gồm nhiều trình tự tương đồng.

Bài toán: Đề xuất phương pháp ước lượng mô hình thay thế axit amin cho tập *N* sắp hàng trên sao cho đảm bảo về độ chính xác của mô hình cũng như tối ưu về mặt thời gian thực thi.

Kết quả đầu ra: Mô hình thay thế axit amin Q tương ứng giải thích một cách tốt nhất sự biến đổi của axit amin trong các sắp hàng đa trình tự của tập **D**.

Ước lượng Q là một bài toán khó và phức tạp do có nhiều tham số cần phải được tối ưu cùng lúc trong khi tài nguyên tính toán và thời gian là có giới hạn.

1.2.4.2 Phương pháp cực đại hợp lý ước lượng mô hình thay thế axit amin

Cho tập dữ liệu đầu vào gồm N sắp hàng $\mathbf{D} = (D_1, \dots, D_N)$ và $\mathbf{T} = (T_1, \dots T_N)$ là tập các cây phân loài tương ứng với các sắp hàng trong **D**. Theo đó, giá trị hợp lý của mô hình Q và cây phân loài **T** đối với sắp hàng **D** được tính theo công thức như sau:

$$L(Q, \mathbf{T}|\mathbf{D}) = \prod_{i=1}^{N} L(Q, T_i|D_i) \tag{1.2}$$

Ở đây, $L(Q, T_i|D_i)$ là giá trị hợp lý của Q và T_i đối với sắp hàng D_i. Mô hình Q được ước lượng bằng cách tìm cực đại của giá trị hợp lý $L(Q, \mathbf{T}|\mathbf{D})$ theo công thức:

$$\begin{aligned} Q &= argmax_Q(L(Q, \mathbf{T}|\mathbf{D}) \\ &= argmax_Q\{\textstyle\prod_{i=1}^{N} L(Q, T_i|D_i)\} \end{aligned} \tag{1.3}$$

Gần đây, các nhà khoa học đã giới thiệu hai phương pháp mới để ước lượng nhanh mô hình *Q* gọi là QMaker [5] giành cho mô hình thời gian thuận nghịch và nQMaker [6] cho mô hình thời gian không thuận nghịch. Hai phương pháp này dùng tiêu chuẩn cực đại hợp lý để ước lượng mô hình thay thế axit amin theo lược đồ tóm tắt như sau:

Bước 1: Khởi tạo mô hình tốt nhất hiện tại theo tập mô hình có sẵn LG, WAG, JTT, giả sử chọn $Q = \text{LG}$.

Bước 2: Với mỗi sắp hàng thực hiện tìm cây phân loài tối ưu nhất theo mô hình Q tốt nhất hiện tại.

Bước 3: Với cây phân loài đã tìm được, thực hiện ước lượng mô hình mới $Q*$ theo công thức:

$$Q^* = argmax_Q(L(Q, \mathbf{T}|\mathbf{D})$$

Bước 4: Kiểm tra $Q \approx Q^*$ thì dừng. Ngược lại, gán $Q = Q^*$ và quay về Bước 2 tiếp tục vòng lặp đến khi đạt điều kiện dừng.

1.2.4.3 Mô hình tốc độ biến đổi tại các vị trí trên trình tự

Việc giả thiết rằng tốc độ biến đổi tại tất cả các vị trí đều như nhau là không phù hợp với thực tế, nhiều nghiên cứu đã cho thấy tốc độ biến đổi là không đồng nhất. Do đó, việc sử dụng một mô hình tốc độ biến đổi là cần thiết để giải thích hiện tượng này. Thông thường, chúng ta có thể mô hình hóa bằng phân phối gamma (Γ) với giá trị kỳ vọng 1 và phương sai $1/\alpha$ ($\alpha > 0$). Trong đó, công thức của hàm mật độ như sau:

$$Pdf(r) = \frac{\alpha^{\alpha} r^{\alpha-1}}{e^{\alpha r}\Gamma(\alpha)}$$

ở đây, $\Gamma(\alpha) = \int_0^{\infty} e^{-t} t^{\alpha-1} dt$.

Ước lượng các mô hình thay thế axit amin Q đi kèm với mô hình tốc độ tại các vị trí là nội dung sẽ được trình bày trong Chương 3 của luận án.

1.3 Các phương pháp so sánh hai mô hình thay thế axit amin

1.3.1 So sánh dựa trên giá trị các hệ số của hai mô hình

1.3.2 So sánh dựa trên giá trị hợp lý

Do giá trị hợp lý rất nhỏ nên để thuận tiện trong tính toán và tối ưu ta thường sử dụng giá trị logarit của giá trị hợp lý (LogLikelihood – LH), giá trị này là một số âm.

Bên cạnh giá trị hợp lý, hai tiêu chuẩn khác cũng thường được sử dụng là tiêu chuẩn thông tin Bayesian (Bayesian information criterion - BIC) và tiêu chuẩn thông tin Akaike (Akaike information criterion - AIC).

1.3.3 So sánh dựa trên cấu trúc cây phân loài

Trong luận án này, ta sử dụng khoảng cách Robinson-Foulds (RF).

1.3.4 So sánh sử dụng phân tích thành phần chính

Phân tích thành phần chính (principal component analysis - PCA) là phương pháp giảm số chiều của một bộ dữ liệu nhằm mục đích trích xuất ra mô hình và xu hướng quan trọng của dữ liệu. Mỗi mô hình thay thế axit amin bao gồm 400 phần tử, chúng ta có thể coi nó như một vector gồm có 400 chiều.

1.4 Các phương pháp lựa chọn mô hình phù hợp nhất với dữ liệu

1.4.1 Phương pháp dựa trên cực đại hợp lý

Đây là phương pháp truyền thống, có một số phương pháp tiêu biểu như PhyML, RAxML, ModelTest-NG và ModelFinder. Với mỗi mô hình thay thế axit amin cần lựa chọn, các phương pháp này sẽ xây dựng cây và tính giá trị hợp lý của cây xây dựng được. Mô hình cho kết quả tốt nhất là mô hình được lựa chọn.

1.4.2 Phương pháp dựa trên học máy

Đây là cách tiếp cận mới, có một số công trình đã được đề xuất như ModelTeller, ModelRevelator cho dữ liệu DNA. Các phương pháp này sẽ học các đặc trưng của dữ liệu rồi dự đoán mô hình phù hợp cho một sắp hàng bất kì.

1.5 Các bộ dữ liệu

Luận án sử dụng các bộ dữ liệu dùng chung như: HSSP, TreeBASE, Pfam và các bộ dữ liệu theo loài gồm: plant, bird, yeast, mammal và insect.

1.6 Kết luận chương

Quá trình tiến hóa luôn xảy ra những biến đổi trong trình tự axit amin, với đặc điểm môi trường sống ngày nay thì nó diễn biến lại càng phức tạp hơn khi mà hóa chất, thuốc men đang lan tràn và gây ảnh hưởng mạnh mẽ đến cơ thể sống của mọi sinh vật.

Chuỗi Markov thường được sử dụng để mô hình hóa quá trình biến đổi này, nó được thể hiện dạng ma trận hay mô hình Q. Ước lượng Q là bài toán khó và phức tạp, đòi hỏi kết hợp nhiều phương pháp khác nhau, trong đó phương pháp truyền thống và phổ biến nhất hiện nay là phương pháp cực đại hợp lý (maximum likelihood). Với phương pháp này, các nhà khoa học đã ước lượng được một số mô hình đơn ma trận dùng chung như LG, WAG hay JTT. Ngoài ra những mô hình biến đổi riêng cho các loài cũng đã được đề xuất nhằm thuận tiện hơn trong việc nghiên cứu chuyên sâu về các loài đó. Những mô hình này (chung hay riêng) là nền tảng để tiếp tục ước lượng các mô hình mới nhằm tối ưu hóa cây phân loài.

Chương 2. Quy trình đánh giá các phương pháp ước lượng mô hình thay thế axit amin

2.1 Giới thiệu chung

Luận án đề xuất quy trình đánh giá phương pháp ước lượng dựa trên dữ liệu mô phỏng gồm các bước như sau:

- Bước 1: Thu thập các dữ liệu sắp hàng thật, ước lượng mô hình thay thế Q, cây phân loài T và các tham số tốc độ biến đổi tại các vị trí.

- Bước 2: Sinh dữ liệu mô phỏng dựa trên mô hình Q, cây phân loài T và các tham số liên quan như tốc độ biến đổi tại các vị trí, tham số hình dạng α của phân phối gamma tìm được ở Bước 1.
- Bước 3: Ước lượng mô hình Q' mới dựa trên phương pháp ước lượng mô hình dựa trên dữ liệu từ Bước 2.
- Bước 4: Đánh giá mô hình Q' so với Q sử dụng các tiêu chuẩn như: giá trị hệ số biến đổi của mô hình, tiêu chuẩn cực đại hợp lý và khoảng cách RF.

Mục tiêu phần này của luận án là dựa trên việc sử dụng dữ liệu mô phỏng để ước lượng mô hình, rồi đánh giá xem độ tốt của mô hình được ước lượng thay đổi như thế nào đối với từng kích thước dữ liệu đào tạo.

2.2 Phương pháp

2.2.1 Phương pháp sử dụng dữ liệu đa sắp hàng mô phỏng

Với mỗi sắp hàng thật R_i quá trình để sinh một sắp hàng $D_i \in \mathbf{D}$ bao gồm hai bước chính cụ thể như sau:

- Ước lượng tham số: với mỗi sắp hàng thật R_i gồm m_i trình tự với độ dài l. Bằng cách chạy IQ-TREE [34] cho mỗi sắp hàng, ta sẽ xác định được các tham số như mô hình tốc độ biến đổi trên mỗi vị trí ρ_i, giá trị tham số hình dạng α.
- Sinh dữ liệu: mô hình $Q^{defined}$, cây $T^{defined}$ cho trước, kết hợp mô hình tốc độ ρ_i và α tìm được ở trên ta thực hiện sinh dữ liệu đa sắp hàng D_i với m_i sắp hàng có độ dài l_i kí tự sử dụng chương trình sinh dữ liệu mô phỏng.

2.2.2 Phương pháp ước lượng và đánh giá các mô hình từ dữ liệu mô phỏng

2.2.3 Đánh giá các mô hình được ước lượng từ dữ liệu mô phỏng

Ở đây ta sẽ đánh giá dựa trên các tiêu chuẩn như: giá trị các hệ số mô hình, giá trị hợp lý và khoảng cách RF giữa các cây phân loài.

2.3 Kết quả

2.3.1 Kết quả sử dụng dữ liệu mô phỏng

Luận án sử dụng phương pháp AliSim để sinh dữ liệu dựa trên các tham số từ dữ liệu sắp hàng thật (real alignments) và thực nghiệm trên hai bộ dữ liệu Plants và Birds.

2.3.2 Kết quả ước lượng và đánh giá mô hình từ dữ liệu mô phỏng

2.3.2.1 Kết quả ước lượng mô hình

Với dữ liệu sau khi sinh ra, các phương pháp QMaker và nQMaker được dùng để ước lượng mô hình thay thế axit amin theo thuộc tính thuận nghịch và không thuận nghịch thời gian.

2.3.2.2 Kết quả phân tích mô hình

Với mỗi kịch bản để tạo dữ liệu mô phỏng, giá trị trung bình của độ đo tương quan giữa mô hình đúng và năm mô hình mô phỏng được tính toán và lấy trung bình. Theo đó, xu hướng của độ tương quan tăng dần theo kích thước của số lượng gene dùng để ước lượng mô hình, tuy nhiên tất cả đều có độ tương quan mạnh với nhau (>0.95).

2.3.2.3 Đánh giá khả năng xây dựng cây cực đại hợp lý

Với mỗi sắp hàng, mô hình A tốt hơn mô hình B nếu cây phân loài được xây dựng bởi mô hình A có giá trị hợp lý cao hơn cây phân loài được xây dựng bởi mô hình B. Ở đây, tôi đã so sánh các mô hình mô phỏng với mô hình đúng để xem độ tốt của mô hình mô phỏng thay đổi ra sao theo các kích thước gene, và tôi cũng đánh giá hiệu suất của các mô hình mô phỏng với nhau.

Kết quả cho ta thấy rằng hầu hết cây phân loài xây dựng với mô hình đúng cho giá trị hợp lý cao hơn cây được xây dựng với mô hình mô phỏng.

2.3.2.4 Đánh giá ảnh hưởng của mô hình đến cấu trúc cây

Đến đây, sau khi đã so sánh các mô hình tạo ra từ hai phương pháp QMaker và nQMaker thông qua các hệ số biến đổi tốc độ cũng như khả năng xây dựng cây cực đại hợp lý, tôi tiếp tục đánh giá chi tiết hơn về hình dạng của cây được xây dựng bởi các mô hình này. Tôi tiến hành tính khoảng cách Robinson-Foulds (RF) giữa các cây tạo bởi mô hình đúng, mô hình mô phỏng và cây thực (real tree).

2.4 Kết luận chương

Tổng kết lại, trong chương này, luận án đã trình bày một luồng tiến trình từ sinh dữ liệu, ước lượng và đánh giá mô hình với dữ liệu mô phỏng.

Các thực nghiệm đã cho thấy, dữ liệu càng nhiều thì mô hình càng tốt trên cả khía cạnh xây dựng cây cực đại hợp lý lẫn cấu trúc của cây phân loài.

Chương 3. Ước lượng mô hình thay thế axit amin sử dụng đa ma trận

3.1 Giới thiệu chung

Cho đến hiện tại, nhiều mô hình đa ma trận đã được giới thiệu như EX2, CF4, UL4 hay gần đây là LG4X, LG4M. Các mô hình đa ma trận như LG4X, LG4M được sử dụng rộng rãi do đạt hiệu suất tốt. Trong phần này, luận án sẽ trình bày phương pháp thuận tiện hơn để ước lượng các mô hình đa ma trận, được gọi là QMix.

3.2 Phương pháp

3.2.1 Mô hình thay thế axit amin sử dụng đa ma trận

Như luận án đã giới thiệu ở Chương 1, tốc độ biến đổi tại các vị trí là không đồng nhất và thường được mô hình hóa bởi một phân phối, thông thường là phân phối gamma rời rạc với *C* lớp tốc độ khác nhau. Theo Chương 1, công thức tính giá trị hợp lý của mô hình *Q* và cây *T* đối với đa sắp hàng *D* có *l* vị trí như sau:

$$L(Q,T|\mathrm{D}) = \prod_{i}^{\iota} L(Q,T|D_i) \tag{3.1}$$

Chúng ta có thể áp dụng phân phối gamma với C nhóm tốc độ có trọng số tương ứng $\mathbf{R} = \{r_1, r_2, r_3, .., r_M\}$ vào công thức 3.1 để thu được công thức mới chứa các thành phần tốc độ như sau:

$$L(\mathrm{Q},T,\alpha|D) = \prod_{i=1}^{l}\left(\sum_{k=1}^{C} r_k L(\Gamma(\alpha,k)\mathrm{Q},T|D_i)\right) \tag{3.2}$$

trong đó Γ(α, *k*) là tốc độ thứ *k* của phân phối gamma với tham số hình dạng α, và các trọng số của các tốc độ là r_k. Cả *T* và α được ước lượng từ dữ liệu đầu vào dựa trên tiêu chuẩn cực đại hợp lý. Một số nghiên cứu đã đề xuất mô hình đa ma trận với *M* ma trận $\mathbf{Q} = \{Q_1, Q_2, \dots, Q_M\}$**,** và các trọng số tương ứng ký hiệu là $\boldsymbol{W} = \{w_1, w_2, \dots, w_M\ \}$ với điều kiện cho trước $\sum_m w_m = 1$.

$$L(\mathbf{Q} = \{Q_1, Q_2, \dots, Q_M\}, T, \boldsymbol{W} = \{w_1, w_2, \dots, w_M\}|D) = \prod_{i=1}^{l}\left\{\sum_{m=1}^{M} w_m L(Q_m, T|D_i)\right\} \tag{3.3}$$

Kết hợp công thức 3.2 và 3.3, ta có mô hình đa ma trận kết hợp với mô hình tốc độ biến đổi tổng quát như sau:

$$L\left(\begin{matrix}\mathbf{Q} = \{Q_1, Q_2, .., Q_M\}, \mathbf{W} = \{w_1, w_2, .., w_M\}, \\ \mathbf{P} = \{\rho_1, \rho_2, \dots, \rho_C\}, \mathbf{R} = \{r_1, r_2, .., r_C\}T\end{matrix}\middle| D\right) = \prod_{i=1}^{N}\left\{\sum_{m=1}^{M} w_m \sum_{k=1}^{C} r_k L(\rho_k Q_m, T|D_i)\right\} \tag{3.4}$$

ở đây, **Q** là tập các ma trận, **W** là các trọng số của các ma trận, **P** là tập các phân lớp tốc độ còn **R** là trọng số tương ứng của các phân lớp tốc độ. Từ công thức tổng quát này, áp dụng cho các mô hình tốc độ khác nhau các nhà khoa học đã ước lượng một số mô hình thay thế axit amin. Cụ thể, khi sử dụng phân phối gamma gồm C phân lớp tốc độ có trọng số bằng nhau, các tác giả Lê Sĩ Quang và Gascuel đã ước lượng một số mô hình đa ma trận như EX2 (gồm hai ma trận), UL3 (gồm ba ma trận) dựa trên công thức như sau:

$$L(\mathbf{Q} = \{Q_1, Q_2, .., Q_M\}, \mathbf{W} = \{w_1, w_2, .., w_M\}, T, \alpha|D) = \prod_{i=1}^{N}\left\{\sum_{m=1}^{M}\frac{w_m}{C}\sum_{k=1}^{C}L(\Gamma(\alpha,k)Q_m, T|D_i)\right\} \quad (3.5)$$

Với $C = M = 4$ đồng thời mỗi phân lớp tốc độ chỉ áp dụng cho một ma trận, các nhà khoa học đã ước lượng mô hình LG4M bao gồm 4 ma trận tương ứng với 4 phân lớp tốc độ có trọng số bằng nhau [28] dựa trên công thức nhau sau:

$$L(\mathbf{Q} = \{Q_1, Q_2, Q_3, Q_4\}, T, \alpha|D) = \prod_{i=1}^{N}\left\{\frac{1}{4}\sum_{k=1}^{4}L(\Gamma(\alpha,k)Q_k, T|D_i)\right\} \quad (3.6)$$

ở đây, các phân lớp tốc độ của phân phối gamma đều có trọng số bằng nhau (=1/4).

Áp dụng công thức tổng quát 3.4 với $C = M = 4$, và cũng như trên, mỗi phân lớp tốc độ được áp dụng cho một ma trận, chúng ta thu được công thức cho mô hình đa ma trận với tốc độ tuân theo phân phối tự do như sau:

$$L(\mathbf{Q}, T, \mathbf{P} = \{\rho_1, \rho_2, \rho_3, \rho_4\}, \mathbf{W} = \{w_1, w_2, w_3, w_4\}|D) = \prod_{i=1}^{N}\left\{w_k\sum_{k=1}^{4}L(\rho_k Q_k, T|D_i)\right\} \quad (3.7)$$

Ở đây, mỗi ma trận được gán trọng số w_k và tốc độ ρ_k; các tham số w_k, ρ_k được ước lượng từ dữ liệu và thỏa mãn hai điều kiện $\sum_{k=1}^{M} w_k = 1$ và $\sum_{k=1}^{M} w_k\rho_k = 1$. Dựa trên công thức 3.7, các nhà khoa học đã ước lượng mô hình LG4X [28] gồm 4 ma trận. Trong luận án này, dựa trên các công thức 3.6 và 3.7, tôi cũng ước lượng các mô hình thay thế axit amin sử dụng 4 ma trận với các phân lớp tốc độ tương ứng là "rất chậm", "chậm", "bình thường" và nhanh. Phương pháp ước lượng được trình bày trong phần tiếp theo ngày sau đây.

3.2.2 Phương pháp ước lượng mô hình thay thế axit amin sử dụng đa ma trận

Giả sử chúng ta có một tập *N* các sắp hàng kí hiệu là $\mathbf{D}= \{D^1, D^2, \ldots, D^N\}$ và chúng ta cần ước lượng mô hình gồm *M* ma trận $\mathbf{Q}^* = \{Q_1^*, Q_2^*, .., Q_M^*\}$ sao cho giá trị hợp lý đạt cực đại:

$$\mathbf{Q}^* = \underset{\mathbf{Q}=\{Q_1,Q_2,..,Q_M\},T,P,W}{\operatorname{argmax}}\left\{\prod_{n=1}^{N}L(\mathbf{Q}, T^n, \rho^n, w^n|D^n)\right\} \quad (3.8)$$

Ở đây, D^n là sắp hàng thứ *n* trong tập dữ liệu **D**, còn T^n, ρ^n, w^n tương ứng là cây phân loài, mô hình tốc độ biến đổi tại các vị trí và trọng số của sắp hàng D^n, còn $L(\mathbf{Q}, T^n, \rho^n, w^n | D^n)$ là giá trị hợp lý của mô hình **Q**, cây T^n, mô hình tốc độ ρ^n, trọng số w^n đối với sắp hàng D^n.

Để ước lượng được ma trận $\mathbf{Q}^*$ chúng ta phải tối ưu cùng lúc **Q**, **T**, **P**, và **W**, đây là một thách thức tính toán rất lớn. Tuy nhiên, theo các nghiên cứu đã công bố, chúng ta có thể tối ưu từng phần một cách tuần tự để thu được mô hình $\mathbf{Q}^*$. Cụ thể, quá trình tối ưu sẽ gồm hai bước chính như sau:

- Bước 1: Cố định các ma trận Q_k, tối ưu T^n, ρ^n, w^n bằng phương pháp cực đại hợp lý theo công thức 3.5 cho tốc độ theo phân phối gamma hoặc 3.6 cho mô hình tốc độ tự do. Giả sử ta thu được T^*, ρ^*, w^*.
- Bước 2: Sau khi có T^*, ρ^*, w^*, chúng ta tối ưu **Q** theo công thức 3.7 để thu được $\mathbf{Q}^*$.

Lặp lại hai bước này cho tới khi $\mathbf{Q}^*$ không tối ưu thêm được nữa thì dừng.

Do tính độc lập của các thành phần tốc độ, trọng số và cây phân loài nên chúng ta hoàn toàn có thể tối ưu T^n, ρ^n, w^n cho từng sắp hàng D^n riêng biệt rồi tổng hợp để có được kết quả cuối cùng là mô hình $\mathbf{Q}^*$. Ước lượng **Q*** gồm 4 ma trận cần tối ưu một lượng lớn tham số (4*208 cho thuộc tính thuận nghịch theo thời gian hoặc 4*379 cho thuộc tính không thuận nghịch theo thời gian). Nhóm tác giả Lê Sĩ Quang và cộng sự [28] đã đề xuất phương pháp xấp xỉ để tối ưu hóa quá trình ước lượng $\mathbf{Q}^*$ bằng cách sử dụng phân lớp tốc độ có xác suất lớn nhất. Như vậy $\mathbf{Q}^*$ sẽ được ước lượng theo công thức:

$$\mathbf{Q}^* = (Q_1^*, Q_2^*, \dots, Q_M^*) = \underset{\mathbf{Q}=\{Q_1,Q_2,\dots,Q_M\},T,P,W}{\operatorname{argmax}} \left\{ \prod_{n=1}^{N} \prod_{i=1}^{l} L(T^n, \rho_{c_i}^n Q_{c_i} | D_i^n) \right\} \tag{3.9}$$

Ở đây, D_i^n là vị trí thứ *i* của sắp hàng D^n, c_i là phân lớp tốc độ có xác suất cao nhất cho D_i^n còn ρ_{c_i} là tốc độ của c_i tương ứng với Q_{c_i}. Mỗi ma trận có thể được ước lượng riêng biệt thông qua công thức:

$$\forall k = 1 \dots M, Q_k^* = \underset{Q_k}{\operatorname{argmax}} \left\{ \prod_{n=1}^{N} \prod_{i=1, c_i=k}^{l} L(T^n, \rho_k^n Q_k | D_i^n) \right\} \tag{3.10}$$

Dựa vào công thức 3.9, chúng ta có thể ước lượng hai mô hình mới dựa trên thuộc tính thuận nghịch hoặc không thuận nghịch về thời gian đồng thời tốc độ biến đổi trên các vị trí tuân theo phân bố gamma hoặc phân bố tự do.

3.3 Kết quả

3.3.1 Thực nghiệm đánh giá độ ổn định phương pháp Qmix

Đầu tiên, tôi thực hiện ước lượng lại hai mô hình tương tự với LG4X, LG4M từ 1471 sắp hàng thuộc tập dữ liệu HSSP. Mục đích của việc này là đánh giá hiệu suất của mô hình xem có tương đương với hai mô hình gốc hay không.

3.3.2 Ước lượng hai mô hình chung sử dụng đa ma trận mới

Mục tiêu phần này của luận án là sử dụng QMix kết hợp với nQMaker ước lượng hai mô hình chung mới, nT4X và nT4M, tương ứng tuân theo mô hình tốc độ tự do và theo phân phối gamma.

3.3.2.1 Phân tích các hệ số của mô hình

3.3.2.2 Đánh giá khả năng xây dựng cây cực đại hợp lý

Luận án thực nghiệm trên 300 sắp hàng HSSP và 84 sắp hàng TreeBASE để đánh giá khả năng của mô hình trong việc xây dựng cây phân loài.

Với mô hình M và sắp hàng D, nhắc lại công thức tính AIC như sau:

$$AIC(M|D) = 2 * k - 2 * ln(L(Q,T|D))$$

Ở đây, $ln(L(Q,T|D))$ là giá trị hợp lý được lấy logarit, k là số lượng tham số tự do của mô hình M. Cho hai mô hình M_1 và M_2, nếu $\Delta AIC = AIC(M_1, D) - AIC(M_2, D) < 0$ thì M_1 được xem như tốt hơn M_2 và ngược lại.

Để đánh giá sâu hơn khi so sánh độ tốt giữa các mô hình, luận án sử dụng kiểm định KH-Test với p-value là 0.01. Về cách thực hiện, ta sử dụng kĩ thuật lấy mẫu có hoàn lại RELL bootstrap với 10,000 lần lặp để đánh giá phân phối của giá trị ΔAIC. Nếu giá trị p-value < 0.01 thì M_1 thực sự tốt hơn M_2 còn ngược lại chúng ta không thể khẳng định M_1 tốt hơn M_2.

Bảng 0.1. So sánh dựa trên AIC giữa nT4X, nT4M với 13 mô hình khác dựa trên các bộ dữ liệu HSSP và TreeBASE. Chú thích: #M_1>M_2: số lượng sắp hàng mà AIC của M_1 tốt hơn của M_2. #M_1>M_2(p<0.01): số lượng sắp hàng mà AIC của M_1 tốt hơn thực sự AIC của M_2. Tương tự cho #M_1>M_2 và #M_1>M_2(p<0.01).

M_1	M_2	300 sắp hàng HSSP			84 sắp hàng TreeBASE		
		#M_1>M_2	#M_1>M_2 (p<0.01)	#M_2>M_1 (p<0.01)	#M_1>M_2	#M_1>M_2 (p<0.01)	#M_2>M_1 (p<0.01)
nT4M	LG4M	217	68	18	72	16	6
nT4X	EXEHO	137	23	37	38	8	4
nT4X	LG4X	170	27	26	49	11	5
nT4M	EXEHO	103	12	37	11	3	8

nT4X	UL3	208	56	37	55	6	3
nT4M	UL3	183	33	42	23	5	14
nT4M	CF4	297	111	2	83	18	0
nT4X	CF4	297	111	2	83	25	0
nT4X	NQ.pfam	294	66	0	81	13	1
nT4M	C20	260	53	13	52	10	3
nT4X	C20	271	57	7	65	12	1
nT4X	NQ.insect	284	112	2	80	8	1
nT4M	C60	247	55	15	33	13	7
nT4X	NQ.yeast	284	101	1	79	17	0
nT4X	C60	258	61	15	48	18	3
nT4M	NQ.pfam	277	60	5	65	10	6
nT4M	NQ.insect	273	97	4	56	8	2
nT4M	NQ.yeast	276	82	3	66	10	3
nT4M	NQ.plant	272	105	8	71	17	4
nT4X	NQ.plant	277	125	2	76	21	1
nT4M	NQ.mammal	281	99	11	76	11	3
nT4X	NQ.mammal	283	110	10	77	16	1
nT4M	NQ.bird	283	102	7	77	14	3
nT4X	NQ.bird	286	123	7	77	24	2

Kết quả cho thấy nT4X và nT4M tốt hơn các mô hình đơn ma trận ở hầu hết các sắp hàng của 2 bộ HSSP và TreeBASE.

3.3.2.3 Đánh giá ảnh hưởng của mô hình đến cấu trúc cây

Ở phần này, luận án so sánh cấu trúc cây được xây dựng bởi nT4X, nT4M với các mô hình khác sử dụng khoảng cách nRF.

Với bộ kiểm tra thuộc TreeBASE, nT4M tốt hơn LG4M ở 72 sắp hàng, trong đó có 33 sắp hàng cho cây khác nhau và có 8 sắp hàng cho cây khác nhau mà nT4M thực sự tốt hơn LG4M.

3.3.2.4 Đánh giá khả năng xây dựng cây có gốc

Để kiểm nghiệm khả năng trong việc xây dựng cây có gốc của các mô hình không thuận nghịch về mặt thời gian, luận án sử dụng phương pháp rootstrap với 1000 lần lặp. Kết quả cho thấy mô hình nT4X có thể xây dựng cây có độ hỗ trợ rootstrap tại vị trí gốc trên 50% tại 134 trên 300 sắp hàng HSSP và 36 trên 84 sắp hàng TreeBASE.

3.3.3 Ước lượng hai mô hình riêng sử dụng đa ma trận cho các loài thực vật

Trong phần này, luận án tiếp tục sử dụng phương pháp QMix để ước lượng mô hình đa ma trận cho bộ dữ liệu thực vật gọi là QPlant.mix và nQPlant.mix, cả hai mô hình đều có 4 ma trận, trong đó QPlant.mix tuân theo thuộc tính thuận nghịch thời gian còn nQPlant.mix theo thuộc tính không thuận nghịch thời gian.

Để đánh giá hiệu suất của hai mô hình mới, luận án thực hiện các thí nghiệm trên 308 sắp hàng kiểm tra của bộ dữ liệu Plant.

3.3.3.1 Phân tích các hệ số của mô hình

3.3.3.2 Đánh giá khả năng xây dựng cây cực đại hợp lý

Các mô hình mới vượt trội hơn hẳn các mô hình đang tồn tại trên tập dữ liệu thực vật.

3.3.3.3 Đánh giá ảnh hưởng của mô hình đến cấu trúc cây

3.4 Kết luận chương

Mô hình biến đổi đa ma trận có nhiều ưu điểm phù hợp với các tính chất về mặt sinh học của trình tự axit amin. Độ phức tạp và yêu cầu cao về hiệu năng phần cứng khiến cho việc ước lượng mô hình đa ma trận trở nên khó khăn. Trong chương này, luận án đã trình bày một phương pháp mới nhằm đơn giản hóa quá trình ước lượng mô hình đa ma trận, đó là phần mềm QMix. Các nhà nghiên cứu có thể thực thi quá trình ước lượng bằng một câu lệnh duy nhất và rất nhanh chóng thu nhận được mô hình đa ma trận mới.

Chương 4. Phương pháp lựa chọn nhanh mô hình thay thế axit amin

4.1 Giới thiệu chung

Việc tìm mô hình theo tiêu chuẩn cực đại hợp lý yêu cầu lượng tính toán rất lớn do phải phân tích một loạt các mô hình khác nhau. Do vậy, chúng ta cần phải có một phương pháp khác nhanh chóng đưa ra được mô hình phù hợp nhất với sắp hàng cho trước mà tiêu tốn ít tài nguyên nhất.

Cách tiếp cận dựa trên học máy đã được áp dụng trong một số nghiên cứu tin sinh học và cho kết quả khả quan. Việc sử dụng học sâu để dự đoán mô hình biến đổi cho trình tự axit amin là cần thiết và có nhiều lợi ích lớn như tiết kiệm thời gian và chi phí tính toán. Trong chương này, luận án giới thiệu một kiến trúc học sâu rất tinh giản và phương pháp trích xuất thông tin từ sắp hàng để có thể dự đoán hiệu quả mô hình thay thế axit amin, phương pháp này được gọi là ModelDetector.

4.2 Phương pháp

4.2.1 Sinh dữ liệu mô phỏng

4.2.2 Phương pháp trích xuất thông tin theo cặp trình tự

Thiết lập các tổng hợp thống kê (summary statistics) từ các sắp hàng protein là một trong các bước quan trọng nhất của quá trình đào tạo các mạng học sâu để nhận biết các mô hình thay thế axit amin. Do có tổng cộng 20 axit amin nên mô hình Q bao gồm 400 hệ số tốc độ biến đổi giữa các cặp axit amin. Để trích xuất thông tin thống kê trong sắp hàng, tôi lựa chọn ngẫu nhiên $N = 1000$ cặp trình tự trong sắp hàng. Với mỗi cặp trình tự, tôi sẽ thực hiện đếm số lần thay thế axit amin trong đó và tổng hợp toàn bộ thông tin của N lần lặp. Cuối cùng, tôi thu được 400 giá trị đã được chuẩn hóa thể hiện tốc độ biến đổi của các axit amin trong sắp hàng.

4.2.3 Kiến trúc mạng

ModelDetector sử dụng hàm kích hoạt ReLU cho các lớp bên trong, hàm kích hoạt Softmax cho lớp cuối cùng và sử dụng hàm tối ưu Adam. Nhãn (label) của mỗi sắp hàng chính là mô hình được sử dụng để sinh ra sắp hàng đó. Luận án thực hiện gán nhãn cho từng sắp hàng với 1 trong 13 nhãn: Q.plant, Q.bird, Q.yeast, Q.mammal, Q.insect, Q.pfam, LG, WAG, JTT, VT, PMB, Blosum62, và Dayhoff.

4.3 Kết quả

4.3.1 Phân tích các giá trị tổng hợp thống kê

4.3.2 Đánh giá phương pháp trên bộ dữ liệu mô phỏng

Sau khi xây dựng kiến trúc mạng và thực hiện thu thập các tổng hợp thống kê từ dữ liệu đào tạo. Chúng ta tiến hành quá trình huấn luyện mạng theo các thông số như đã trình bày ở các phần trước. Độ chính xác trên tập xác thực (validation accuracy) đạt 0.99 và giá trị tổn hao trên tập xác thực (validation loss) giảm xuống 0.01 sau một vài epoch. Quá trình huấn luyện mạng kết thúc sau 70 epoch với độ chính xác trên tập xác thực là 0.998. Độ chính xác (accuracy) và độ tổn hao (loss) tương ứng giữa tập huấn luyện và tập xác thực chênh lệch rất nhỏ cho thấy mạng không bị hiện tượng quá khớp (overfitting).

Tiếp theo, tôi thực hiện đánh giá hiệu suất của mạng ModelDetector và so sánh với phương pháp rất thông dụng dựa trên tiêu chuẩn cực đại hợp lý là ModelFinder. Kết quả chung, hai phương pháp đều cho độ chính xác cao 99.71% với ModelDetector và 99.32% với ModelFinder.

4.3.3 Đánh giá phương pháp trên bộ dữ liệu độc lập PFAM

Nhằm tăng tính khách quan và đánh giá toàn diện hơn phương pháp ModelDetector, tôi sử dụng thêm một bộ dữ liệu độc lập là bộ PFAM để kiểm nghiệm khả năng của phương pháp trên dữ liệu thật lẫn dữ liệu mô phỏng dựa trên bộ PFAM. ModelFinder thực hiện tốt hơn một chút so với ModelDetector.

Tôi tiếp tục lựa ra 886 sắp hàng thật có tối thiểu 100 vị trí có sự biến đổi (variant sites) thuộc bộ PFAM để đánh giá khả năng chạy trên dữ liệu thật của phương pháp ModelDetector. Do không biết mô hình đúng của các sắp hàng, tôi thực hiện so sánh kết quả của ModelDetector với kết quả của ModelFinder để xem có bao nhiều sắp hàng mà hai phương pháp dự đoán trùng nhau, kết quả chi tiết xem **Bảng 0.1**. Theo đó, số lượng sắp hàng mà hai phương pháp dự đoán cùng mô hình tăng lên theo kích thước sắp hàng. Ví dụ, hai phương pháp dự đoán đúng ở mức 60.98% sắp hàng có độ dài 100-300 vị trí và 66.67% sắp hàng có trên 500 vị trí. Xem xét giá trị BIC của mô hình tốt nhất và mô hình tốt thứ hai trong kết quả của ModelFinder, ta thấy rằng các giá trị này có độ chênh lệch rất nhỏ. Kết quả này dường như do hiện tượng không đồng nhất về tốc độ biến đổi tại các vị trí (RHAS) nên các vị trí có thể tuân theo các mô hình biến đổi khác nhau. Đôi khi mô hình đúng không phải là có BIC tốt nhất, tôi tiếp tục thống kê thêm sự trùng lặp ở mô hình tốt thứ hai. Với cách thống kê này, hai phương pháp có kết quả trùng nhau ở 82.09% những sắp hàng từ 100 đến 300 vị trí và tăng tới 83.33% trên các sắp hàng ít nhất 500 vị trí.

Bảng 0.1. Số lượng sắp hàng thật thuộc bộ PFAM mà cả hai phương pháp ModelDetector và ModelFinder dự đoán cùng một mô hình.

Số vị trí trên sắp hàng	Số sắp hàng mà ModelDetector trùng với mô hình tốt nhất của ModelFinder	Số sắp hàng mà ModelDetector trùng với mô hình tốt thứ hai của ModelFinder
100-300	60.98%	82.09%
300-500	62.51%	82.16%
>500	66.67%	83.33%

Với các sắp hàng được dự đoán khác nhau, luận án đánh giá khoảng cách nRF [44] giữa các cây được xây dựng với mô hình dự đoán bởi ModelDetector và mô hình tốt nhất của ModelFinder. Giá trị nRF trung bình là 0.121. Với mỗi sắp hàng được dự đoán trái ngược này, luận án cũng thử nghiệm chạy IQ-TREE 3 lần để xây dựng 3 cây phân loài khác nhau với kịch bản là chạy với cùng một mô hình nhưng bằng các số nhân ngẫu nhiên (random seed) khác nhau. Mục đích của việc này là đánh giá kết quả xây dựng cây của IQ-TREE xem có đồng nhất hay không. Kết quả là các cây này có nRF trung bình là 0.143. Từ các kết quả trên, chúng ta có thể thấy độ tương đồng cao trong kết quả dự đoán của ModelDetector và ModelFinder.

4.3.4 Phân tích thời gian dự đoán

Yếu tố thời gian chạy là một trong những ưu điểm của phương pháp dựa trên học sâu. ModelDetector được huấn luyện rất nhanh trên các máy tính có 8 lõi mà không cần bất kì phần cứng GPU nào. Việc loại bỏ lớp tích chập đã làm giảm đáng kể thời gian huấn luyện khi giờ đây, mỗi epoch chỉ còn tiêu tốn 168 giây. Tổng cộng, để thu được các tham số mạng cho 2,246,400 sắp hàng, luận án chỉ cần chạy huấn luyện trong khoảng 3.3 giờ sau 70 epoch.

Với các dữ liệu mô phỏng, thời gian chạy của ModelDetector nhanh hơn ModelFinder ở toàn bộ số trường hợp. Thời gian chạy của ModelFinder tăng rất nhanh theo kích thước sắp hàng, ví dụ, ModelFinder mất lần lượt 14.3, 58.2 và 156.6 giây để tìm ra mô hình tốt nhất của sắp hàng có 100, 500 và 1100 vị trí. Trong khi đó, ModelDetector gần như không thay đổi nhiều thời gian chạy với những kích thước này. Cụ thể, ModelDetector cần 2.4 giây để dự đoán một sắp hàng có 100 vị trí và 3.6 giây cho một sắp hàng có 1100 vị trí.

Luận án cũng kiểm tra xem với những sắp hàng có hàng trăm nghìn đến hàng triệu vị trí thì thời gian chạy hai phương pháp này như thế nào. Luận án tạo ra các sắp hàng có 50 trình tự và độ dài tăng dần từ 1,000 đến 1,000,000 vị trí. Kết quả là, ModelDetector chỉ cần khoảng 693 giây để đưa ra kết quả dự đoán cho sắp hàng có 1,000,000 vị trí, nhanh gấp 27 lần so với ModelFinder.

4.4 Kết luận chương

Trong chương này, luận án đã trình bày một kiến trúc mạng mới dựa trên học sâu cùng với phương pháp trích chọn thông tin tổng hợp thống kê từ sắp hàng axit amin để dự đoán mô hình biến đổi của sắp hàng đó. Trong khi phương pháp ModelRevelator sử dụng tới 260,000 thuộc tính và cần dùng phần cứng GPU chuyên dụng để huấn luyện mô hình thì phương pháp của tôi, gọi là ModelDetector, chỉ cần dùng 400 giá trị tổng hợp thống kê để huấn luyện mạng và dự đoán mô hình của sắp hàng. Phương pháp của tôi có thời gian huấn luyện và dự đoán vượt trội so với ModelRevelator và độ chính xác có thể so sánh được với ModelFinder. Tuy nhiên, một giới hạn trong phương pháp dựa trên học sâu là chỉ có thể dự đoán được mô hình mà đã được huấn luyện. Như ModelDetector thì chỉ dự đoán dc 13 mô hình, muốn mở rộng thì bắt buộc chúng ta phải tạo thêm dữ liệu đào tạo và huấn luyện lại mạng. May mắn là mạng ModelDetector có thể được huấn luyện rất nhanh nên việc đó là hoàn toàn khả thi.

Một vấn đề nữa là mạng ModelDetector được dự đoán chỉ dựa vào các sắp hàng sinh ra với mô hình tốc độ tại các vị trí là +I+G, do vậy nó chưa thể dự đoán tốt được những sắp hàng không sử dụng mô hình tốc độ hoặc sử dụng mô hình tần số +F.

Dù cho vẫn còn nhiều điểm cần phải tối ưu và xem xét, nhưng tôi cho rằng phương pháp này rất hứa hẹn trong việc dự đoán mô hình thay thế axit amin và hoàn toàn có thể trở thành xu hướng tốt trong nghiên cứu tiến hóa.

KẾT LUẬN

Xây dựng và đánh giá mô hình thay thế axit amin luôn là bài toán quan trọng trong nghiên cứu sinh học tiến hóa. Ước lượng mô hình thay thế axit amin thường phức tạp và tốn kém cả về thời gian lẫn tài nguyên tính toán. Nhiều phương pháp đã được đề xuất để tối ưu quá trình ước lượng này. Trong luận án này, tôi thực hiện các phân tích lý thuyết và thực nghiệm để giải quyết ba bài toán cơ bản gồm: đánh giá

phương pháp ước lượng mô hình thay thế axit amin, đề xuất phương pháp ước lượng nhanh mô hình thay thế axit amin sử dụng đa ma trận và cuối cùng là đề xuất phương pháp lựa chọn nhanh mô hình thay thế axit amin cho một sắp hàng bất kì.

Trước tiên, luận án đưa ra một quy trình để khám phá hiệu suất các mô hình ước lượng từ dữ liệu mô phỏng. Qua các thực nghiệm trên hai bộ dữ liệu thử nghiệm, luận án đã xác nhận mối liên hệ giữa kích thước của dữ liệu đào tạo với hiệu suất mô hình đầu ra. Luận án đã đưa ra một số đề xuất để người dùng có thể sử dụng dữ liệu mô phỏng một cách hợp lý và chính xác trong việc ước lượng mô hình thay thế axit amin theo tiêu chuẩn cực đại hợp lý.

Tiếp đó, luận án trình bày QMix, là một phương pháp mới dùng để ước lượng mô hình thay thế axit amin sử dụng đa ma trận. Phương pháp mới này được thể hiện dưới dạng một công cụ phần mềm, với giao diện đơn giản, dễ sử dụng cũng như nhanh chóng đưa ra kết quả cho người dùng. Luận án cũng đã thử nghiệm QMix trên các bộ dữ liệu và giới thiệu một số mô hình thay thế axit amin mới, cụ thể, luận án giới thiệu hai mô hình chung dựa trên thuộc tính không thuận nghịch về thời gian nT4X và nT4M tương ứng có mô hình tốc độ tại vị trí tuân theo phân phối tự do và phân phối gamma. Luận án cũng đề xuất hai mô hình biến đổi đa ma trận cho bộ dữ liệu thực vật, QPlant.mix và nQPlant.mix. Các mô hình mới này đã thể hiện được những ưu điểm trong việc nghiên cứu tiến hóa chuỗi protein.

Cuối cùng, luận án trình bày phương pháp lựa chọn mô hình thay thế axit amin mới sử dụng mạng học sâu, gọi là ModelDetector. Với phương pháp này, người dùng có thể nhanh chóng lựa chọn mô hình tối ưu nhất cho một sắp hàng cho trước với độ chính xác tương đương ModelFinder. Một trong những đặc trưng của ModelDetector là không sử dụng lớp tích chập trong thời gian huấn luyện và dự đoán nhanh. Quá trình thực nghiệm và so sánh kết quả với phương pháp ModelFinder dựa trên tiêu chuẩn cực đại hợp lý cho thấy đây là một hướng đi mới đầy khả quan và triển vọng trong nghiên cứu tin sinh học tiến hóa.

Tuy nhiên, chúng tôi nhận thấy còn một số điểm hạn chế đòi hỏi những nghiên cứu và phân tích sâu hơn. Đầu tiên với quy trình đánh giá phương pháp ước lượng, luận án đang chỉ tập trung vào phương pháp ước lượng dựa trên tiêu chuẩn cực đại hợp lý, do đây đang là phương pháp thông dụng và hiệu quả, vì vậy không thể áp dụng quy trình này sang các phương pháp khác. Với phương pháp ước lượng mô hình đa ma trận, câu hỏi là liệu có thể tự động xác định số lượng ma trận phù hợp nhất với dữ liệu hay không? Hiện tại chúng ta cần xác định sẵn số lượng ma trận trước khi dùng phương pháp QMix để ước lượng mô hình. Với phương pháp lựa chọn mô hình dựa trên học sâu, vấn đề quan trọng là sinh dữ liệu mô phỏng sát với dữ liệu thật nhất, luận án sinh dữ liệu dựa trên các tham số được ước lượng từ dữ liệu thật nhưng quá trình đánh giá ModelDetector cho thấy kết quả dự đoán trên dữ liệu thật vẫn còn hạn chế. Ngoài ra là làm sao để có thể sử dụng dữ liệu thật trong quá trình đào tạo mạng? Đây là một câu hỏi rất thách thức đặt ra cho những nhà nghiên cứu về tin sinh học tiến hóa.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC

[CT1] N. H. Tinh, C. C. Dang, and L. S. Vinh, "Rooting Phylogenetic Trees from Protein Alignments," in *Proceedings - International Conference on Knowledge and Systems Engineering, KSE*, Oct. 2023, pp. 1–5. doi: 10.1109/KSE59128.2023.10299425.

[CT2] N. H. Tinh, C. C. Dang, and L. S. Vinh, "Estimating amino acid substitution models from genome datasets: A simulation study on the performance of estimated models," *J. Evol. Biol.*, vol. 37, no. 2, pp. 256–265, 2023, doi: 10.1093/jeb/voad017. (thuộc danh mục SCIE, Q1 theo SCImago)

[CT3] N. H. Tinh, C. C. Dang, and L. S. Vinh, "QMix: An Efficient Program to Automatically Estimate Multi-Matrix Mixture Models for Amino Acid Substitution Process," *J. Comput. Biol.*, vol. 31, no. 8, pp. 703–707, 2024, doi: 10.1089/cmb.2023.0403. (thuộc danh mục SCIE, Q2 theo SCImago)

[CT4] N. H. Tinh and L. S. Vinh, "Improving the study of plant evolution with multi-matrix mixture models," *Plant Syst. Evol.*, vol. 310, 2024, doi: 10.1007/s00606-024-01896-0. (thuộc danh mục SCIE, Q2 theo SCImago)

[CT5] N. H. Tinh and L. S. Vinh, "An efficient deep learning method for amino acid substitution model selection," *J. Evol. Biol.*, vol. 38, no. 1, pp. 129–139, 2024, doi: https://doi.org/10.1093/jeb/voae141. (thuộc danh mục SCIE, Q1 theo SCImago)

[CT6] N. H. Tinh, C. C. Dang, and L. S. Vinh, "nT4X and nT4M: Novel Time Non-reversible Mixture Amino Acid Substitution Models," *J. Mol. Evol.*, 2025, doi: https://doi.org/10.1007/s00239-024-10230-8. (thuộc danh mục SCIE, Q1 theo SCImago)