ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẶNG CAO CƯỜNG

CÁC PHƯƠNG PHÁP XÂY DỰNG MA TRẬN BIẾN ĐỔI AXÍT AMIN

Chuyên ngành: Khoa học Máy tính
Mãsố: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

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

Người hướng dẫn khoa học:

1. TS. Lê Sỹ Vinh
2. TS. Lê Sĩ Quang

Phản biện 1: PGS.TSKH. Vũ Đình Hòa
Trường Đại học Sư phạm Hà Nội

Phản biện 2: PGS.TS. Lương Chi Mai
Viện Công nghệ thông tin, Viện Hàn lâm KH&CN VN

Phản biện 3: PGS.TS. Nguyễn Đức Nghĩa
Trường Đại học Bách khoa Hà Nội

Luận án sẽ được bảo vệ trước hội đồng cấp Đại học Quốc gia chấm luận án tiến sĩ họp tại Trường Đại học Công nghệ vào hồi 9 giờ 00 ngày 10 tháng 01 năm 2014.

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

MỞ ĐẦU

1. Tính cấp thiết của luận án

Ứng dụng công nghệ thông tin để nghiên cứu và giải quyết các bài toán trong sinh học phân tử đang rất được quan tâm. Tin sinh học là lĩnh vực nghiên cứu kết hợp cả hai ngành công nghệ thông tin và sinh học phân tử. Tin sinh học đang được đầu tư lớn do khả năng mang lại sự tiến bộ về khoa học và hiệu quả kinh tế thông qua việc thúc đẩy sự phát triển công nghệ sinh học và ứng dụng trong y tế, nông nghiệp và các lĩnh vực khác.

Các bài toán liên quan đến chuỗi prôtêin như sắp hàng đa chuỗi, tìm kiếm chuỗi tương đồng, xây dựng cây phân loài đều là các bài toán cơ bản và quan trọng của tin sinh học. Tất cả các bài toán này đều cần đến một thành phần rất quan trọng là mô hình (ma trận) biến đổi axít amin. Mô hình biến đổi axít amin có số lượng tham số lớn (khoảng 200 tham số) và thường khó có thể ước lượng trực tiếp trong quá trình phân tích dữ liệu. Chúng ta thường ước lượng trước một mô hình chung (general model) và mô hình này được sử dụng cho mọi bộ dữ liệu prôtêin. Mô hình tổng quát đầu tiên là PAM và gần đây nhất là LG.

Quá trình ước lượng mô hình biến đổi axít amin là một quá trình phức tạp và trải qua nhiều bước tính toán khác nhau, mỗi bước là một bài toán khó. Ba bước chính của quá trình ước lượng mô hình là:

1. Xây dựng cây phân loài từ tập các sắp hàng đa chuỗi. Các thuật toán xây dựng cây dùng trong quá trình ước lượng mô hình còn tốn rất nhiều thời gian. Ví dụ phải mất vài ngày để ước lượng được mô hình LG.
2. Xác định các ràng buộc liên quan đến mô hình. Độ chính xác của mô hình hiện tại vẫn còn hạn chế do việc mô hình hoá đã loại bỏ một số điều kiện ràng buộc trong sinh học phân tử.
3. Xây dựng các mô hình riêng biệt cho các loài sinh vật khác nhau. Đây là một bước rất quan trọng bởi vì trong nhiều trường hợp các mô hình chung không mô hình hoá được hết các đặc điểm biến đổi riêng biệt của các loài.

2. Các đóng góp của luận án

1. Đề xuất một số phương pháp mới để tăng tốc độ quá trình xây dựng cây, giảm bớt số bước tối ưu cấu trúc cây, từ đó giúp giảm thời gian ước lượng mô hình.

2. Sử dụng thêm các ràng buộc trong sinh học phân tử vào quá trình mô hình hoá. Việc này sẽ giúp nâng cao tính chính xác của mô hình biến đổi axít amin khi phân tích dữ liệu.
3. Xây dựng một hệ thống ước lượng tự động mô hình biến đổi axít amin từ dữ liệu của người dùng, qua đó giúp người dùng có thể ước lượng các mô hình riêng biệt cho các loài sinh vật khác nhau.
4. Bên cạnh đó, luận án cũng xây dựng thử nghiệm mô hình biến đổi axít amin cho riêng vi rút cúm và kiểm nghiệm tính hiệu quả của mô hình mới này.

Các kết quả của luận án đã được công bố trong 03 bài báo ở tạp chí SCI quốc tế và 02 báo cáo ở hội nghị quốc tế.

3. Bố cục của luận án

Ngoài phần kết luận, luận án được tổ chức như sau.z

Chương 1 giới thiệu khái quát về chuỗi ADN, chuỗi axít amin, các phép biến đổi, mô hình biến đổi và bài toán ước lượng mô hình biến đổi axít amin. Tiếp theo là phần trình bày về hai cách tiếp cận chính để ước lượng mô hình biến đổi axít amin là phương pháp đếm và phương pháp cực đại khả năng (maximum likelihood). Phần cuối của chương này giới thiệu về phương pháp xây dựng cây phân loài bằng phương pháp cực đại khả năng và các phương pháp so sánh hai mô hình biến đổi axít amin.

Chương 2 đề xuất phương pháp ước lượng nhanh mô hình biến đổi axít amin. Để làm được điều đó, chúng tôi đề xuất hai phương pháp chia tách nhỏ dữ liệu đầu vào. Hai phương pháp này giúp giảm thời gian xây dựng cây phân loài, một bước chiếm rất nhiều thời gian trong quá trình ước lượng mô hình biến đổi axít amin. Các thực nghiệm ở phần sau của chương đã chứng tỏ được hiệu quả của hai phương pháp này.

Chương 3 của luận án giới thiệu mô hình biến đổi axít amin sử dụng nhiều ma trận, một cải tiến mới so với các mô hình đơn ma trận hiện nay. Mô hình mới này sử dụng thêm các ràng buộc trong sinh học phân tử giúp tăng cường khả năng mô hình hoá các quá trình biến đổi của các chuỗi axít amin. Các thực nghiệm với hai bộ dữ liệu HSSP và TreeBase đã chứng tỏ mô hình biến đổi đa ma trận có độ chính xác cao hơn các mô hình hiện tại.

Chương 4 đề xuất một thuật toán ước lượng mô hình biến đổi axít amin cải tiến giúp giảm 50% thời gian ước lượng mô hình. Có được điều này chính là do thuật toán mới đã tìm cách giảm bớt số bước tối ưu cấu trúc cây phân loài – một bước chiếm nhiều thời gian trong quá trình ước lượng. Chương này

cũng giới thiệu hệ thống ước lượng mô hình tự động cài đặt thuật toán cải tiến trên.

Chương 5 trình bày mô hình biến đổi axít amin cho vi rút cúm, gọi là mô hình FLU. Phần sau của chương là các kết quả so sánh mô hình FLU với các mô hình khác. Qua các thực nghiệm, mô hình FLU đã chứng tỏ được hiệu quả cao hơn hẳn các mô hình hiện tại khi phân tích dữ liệu vi rút cúm.

Chương 1. BÀI TOÁN ƯỚC LƯỢNG SỰ BIẾN ĐỔI AXÍT AMIN

1.1. Giới thiệu chung

1.1.1. ADN và axít amin

Giới thiệu về cấu tạo của ADN và axít amin. Chuỗi axít amin là một thành phần vô cùng quan trọng cho sự sống. Prôtêin là thứ vật chất đã phát huy tác dụng quan trọng trong hoạt động của cơ thể, đồng thời còn đóng vai trò chất kích thích hệ miễn dịch, là thành phần cung cấp vitamin và năng lượng cho cơ thể

1.1.2. Các phép biến đổi trên chuỗi chuỗi axít amin

Hai chuỗi axít amin ở hai sinh vật khác nhau cùng tiến hoá từ một chuỗi axít amin tổ tiên thì gọi là hai chuỗi axít amin tương đồng. Hai chuỗi axít amin tương đồng có các khác biệt là do có các biến đổi (còn gọi là đột biến) trong quá trình tiến hoá. Các phép biến đổi thông thường được chia làm ba loại chính là:

- **Thay thế**: một axít amin này bị thay thế bằng một axít amin khác.
- **Xoá**: một hoặc một số axít amin bị xoá khỏi chuỗi.
- **Chèn**: một hoặc một số axít amin được chèn vào chuỗi.

1.1.3. Sắp hàng đa chuỗi axít amin

Quá trình biến đổi làm cho các chuỗi axít amin tương đồng khác nhau về nội dung cũng như độ dài. Sắp hàng đa chuỗi sẽ giúp làm rõ các phép biến đổi giữa các chuỗi axít amin. Sắp hàng đa chuỗi có thể được hiểu như một ma trận các axít amin, trong đó mỗi hàng chính là một chuỗi axít amin; còn mỗi cột (vị trí) chứa các axít amin tương đồng của các chuỗi. Chúng ta có thể sử dụng sắp hàng đa chuỗi để xây dựng cây phân loài giúp đánh giá nguồn gốc tiến hóa của các chuỗi.

1.1.4. Cây phân loài

Cây phân loài (cây tiến hóa) là một dạng sơ đồ phân nhánh thể hiện quá trình tiến hóa của các loài sinh vật và cho biết sự tương đồng và khác biệt về giữa chúng. Các sinh vật liên kết với nhau trong cây được cho là có cùng một tổ tiên

chung. Trong cây phân loài mỗi nút lá biểu diễn cho một loài sinh vật hiện tại, mỗi nút cha đại diện cho tổ tiên gần nhất của các nút con. Độ dài cạnh có thể được hiểu như là ước lượng khoảng cách về thời gian giữa các loài.

1.2. Mô hình hoá quá trình biến đổi axít amin

1.2.1. Sự khác biệt giữa hai chuỗi tương đồng

Có sự khác nhau giữa hai chuỗi axít amin tương đồng cùng tiến hóa từ một tổ tiên chung là do có các biến đổi giữa các axít amin trong quá trình tiến hóa. Hai loại khoảng cách thường dùng để đo sự khác biệt giữa hai chuỗi axít amin tương đồng **x** và **y** là khoảng cách quan sát và khoảng cách di truyền:

- **Khoảng cách quan sát** giữa hai chuỗi axít amin **x** và **y** là tỷ lệ giữa số vị trí trên hai chuỗi có các axít amin không giống nhau so với chiều dài chuỗi.
- **Khoảng cách di truyền** giữa hai chuỗi axít amin **x** và **y** là tỷ lệ giữa số lượng thực tế các biến đổi đã xảy ra giữa hai chuỗi trong quá trình tiến hoá so với chiều dài chuỗi.

Có ba hiện tượng xảy ra trong quá trình tiến hoá của các chuỗi axít amin làm cho khoảng cách quan sát nhỏ hơn rất nhiều khoảng cách di truyền là:

- **Đa biến đổi (multiple substitutions)**: Có nhiều phép biến đổi cùng xảy ra tại một vị trí trong quá trình tiến hoá nhưng chúng ta chỉ quan sát được nhiều nhất 1 phép biến đổi.
- **Biến đổi song song (parallel substitutions)**: Hai phép biến đổi giống hệt nhau cùng xảy ra tại một ví trí trên hai chuỗi con. Chúng ta không quan sát được phép biến đổi này vì trên hai chuỗi con không có sự khác.
- **Biến đổi ngược (back substitutions)**: Có nhiều phép biến đổi xảy ra nhưng axít amin ban đầu và cuối cùng lại giống nhau, chúng ta không quan sát được biến đổi nào giữa hai chuỗi con.

1.2.2. Mô hình Markov cho quá trình biến đổi axít amin

Xét quá trình biến đổi giữa các axít amin tại một vị trí trên chuỗi prôtêin. Quá trình biến đổi này là ngẫu nhiên và liên tục theo thời gian 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} chính là tập 20 axít amin. Quá trình biến đổi axít amin có thể được mô hình hóa bởi một quá trình Markov với các thuộc tính sau đây:

- Độc lập với quá khứ (memoryless): Tốc độ biến đổi từ axít amin *x* thành axít amin *y* không phụ thuộc vào quá trình biến đổi trước đó của axít amin *x*.

- Đồng nhất (homologous): Tốc độ biến đổi giữa các axít amin là đồng nhất trong toàn bộ quá trình biến đổi.
- Liên tục (continuous): Quá trình biến đổi giữa các axít amin có thể diễn ra bất cứ thời điểm nào trong suốt quá trình biến đổi.
- Ổn định (stationary): Tần số của các axít amin là không đổi trong suốt quá trình biến đổi. Gọi $\mathbf{\Pi} = \{\pi_i\}$ với $i = 1, \ldots 20$ là véc tơ tần số xuất hiện của 20 axít amin, khi đó $\sum_{i=1}^{20} \pi_i = 1$ và các π_i không đổi theo thời gian.

Gọi $\mathbf{P}(t) = \{p_{ij}(t),\ i \in \mathbf{S},\ j \in \mathbf{S}\}$ là ma trận xác suất chuyển giữa các axít amin sau một khoảng thời gian t; $p_{ij}(t)$ là xác suất chuyển từ axít amin i ($i \in \mathbf{S}$) sang axít amin j ($j \in \mathbf{S}$) sau một khoảng thời gian t. $\mathbf{P}$ có kích thước 20*20 và với mỗi axít amin i, ta có:

$$\sum_{j \in \mathbf{S}} p_{ij}(t) = 1 \tag{1.1}$$

và $p_{ij}(t) > 0$ với $\forall t > 0$.

$\mathbf{P}(t)$ cũng thỏa mãn công thức Chapman-Kolmogorov:

$$\mathbf{P}(t+s) = \mathbf{P}(t) + \mathbf{P}(s) \tag{1.2}$$

trong đó *t*, *s* là các giá trị thời gian, các điều kiện khởi tạo là:

$$p_{ii}(0) = 1,\ cho\ \forall i = j$$
$$p_{ij}(0) = 0,\ cho\ \forall i \neq j$$

Với giá trị Δt nhỏ, ma trận xác suất chuyển $\mathbf{P}(\Delta t)$ có thể được tính xấp xỉ tuyến tính theo khai triển Taylor như sau:

$$\mathbf{P}(\Delta t) \approx \mathbf{P}(0) + \Delta t * \mathbf{Q} \tag{1.3}$$

trong đó $\mathbf{Q} = \{q_{ij},\ i \in \mathbf{S},\ j \in \mathbf{S}\}$ là ma trận tốc độ biến đổi tức thì (instantaneous substitution rate matrix) giữa các axít amin; $\mathbf{Q}$ có kích thước 20*20 và q_{ij} là tốc độ biến đổi tức thì từ axít amin i sang axít amin j.

Xét một axít amin i, để đảm bảo điều kiện tổng xác suất chuyển từ i đến các trạng thái khác bằng 1 sau một khoảng thời gian t bất kì (Công thức 1.1) thì các giá trị của $\mathbf{Q}$ phải thỏa mãn điều kiện:

$$\sum_{j \in \mathbf{S}} q_{ij} = 0 \text{ hay } q_{ii} = -\sum_{j \in \mathbf{S}, j \neq i} q_{ij} \tag{1.4}$$

Chúng ta có thể coi q_{ij} là lượng biến đổi từ axít amin i sang axít amin j trong một đơn vị thời gian, còn q_{ii} là tổng lượng biến đổi rời khỏi axít amin *i*. Giá trị q_{ij} càng lớn thể hiện tốc độ biến đổi từ axít amin *i* sang axít amin *j* càng lớn.

Dựa vào công thức Chapman-Kolmogorov (Công thức 1.2), chúng ta có thể tính $\mathbf{P}(t)$ từ $\mathbf{Q}$ và t như sau:

$$\mathbf{P}(t) = e^{t\mathbf{Q}} \tag{1.5}$$

Chúng ta gọi

$$\mu = -\sum_{i \in \boldsymbol{S}} \pi_i q_{ii} \tag{1.6}$$

là tổng số lượng biến đổi axít amin trong một đơn vị thời gian. Ta có $d = \mu t$ là tổng số lượng biến đổi axít amin sau một khoảng thời gian t. Ma trận tốc độ biến đổi **Q** được chuẩn hóa sao cho tổng số lượng axít amin biến đổi trong một đơn vị thời gian bằng 1 ($\mu = 1$). Tức là, $p_{ij}(t)$ là xác xuất axít amin i biến đổi thành axít amin j nếu có d biến đổi giữa axít amin i và axít amin j.

Quá trình biến đổi axít amin thường được giả sử có tính thuận nghịch theo thời gian (time reversible), tức là số lượng biến đổi từ axít amin i sang axít amin j bằng với số lượng biến đổi từ axít amin j sang axít amin i (mặc dù tần số xuất hiện của hai axít amin i, j có thể khác nhau), điều này được thể hiện bằng công thức:

$$\pi_i q_{ij} = \pi_j q_{ji} \tag{1.7}$$

hay

$$\frac{q_{ij}}{\pi_j} = \frac{q_{ji}}{\pi_i}$$

Ta kí hiệu $r_{ij} = \frac{q_{ij}}{\pi_j}$; $r_{ji} = \frac{q_{ji}}{\pi_i}$ và gọi r_{ij}, r_{ji} $(r_{ij} = r_{ji})$ là hệ số hoán đổi (exchangeability coefficient) giữa hai axít amin i và j. Hệ số hoán đổi (hay tốc độ biến đổi tương đối) giữa hai axít amin i và j càng lớn thể hiện sự biến đổi giữa hai axít amin i và j xảy ra càng nhiều và ngược lại.

Ma trận tốc độ biến đổi tức thì **Q** có thể được biểu diễn bởi ma trận hoán đổi $\mathbf{R} = \{r_{ij}\}$ và vectơ tần số xuất hiện $\mathbf{\Pi} = \{\pi_i\}$ như sau:

$$q_{ij} = \begin{cases} \pi_j r_{ij} & nếu\ i \neq j \\ -\sum_{x \neq i} q_{ix} & nếu\ i = j \end{cases} \tag{1.8}$$

hoặc có thể viết gọn dưới dạng: $\mathbf{Q} = \mathbf{\Pi} * \mathbf{R}$. Chúng ta cũng thấy ma trận hệ số hoán đổi **R** có dạng đối xứng qua đường chéo chính. Như vậy chúng ta có thể ước lượng **Π** và **R** thay cho ước lượng **Q**. Do **R** có dạng đối xứng nên chúng ta chỉ cần lưu trữ một nửa ma trận nằm dưới đường chéo chính.

Số tham số cần ước lượng của **Π** là 19 do véc tơ **Π** có 20 thành phần nhưng tổng của 20 thành phần bằng 1. Số tham số cần ước lượng của **R** là 19 * 20/2 - 1 = 189, do **R** là ma trận đối xứng và được chuẩn hoá (công thức 1.6 và 1.8). Để ước lượng **Q** chúng ta cần phải ước lượng tổng cộng 208 tham số. Trong

nhiều nghiên cứu về mô hình biến đổi axít amin, ma trận biểu diễn tốc độ biến đổi tức thì **Q** còn được gọi là mô hình **Q**.

1.3. Bài toán ước lượng mô hình biến đổi axít amin

Quá trình biến đổi của axít amin có thể được mô hình hoá bởi mô hình **Q**. Các tham số của mô hình **Q** có thể được ước lượng từ các sắp hàng đa chuỗi axít amin. Bài toán xây dựng mô hình biến đổi axít amin được tóm tắt ngắn gọn như sau:

Dữ liệu vào: Dữ liệu đầu vào là một tập các sắp hàng đa chuỗi axít amin. Các sắp hàng thường có độ dài từ vài chục cho đến vài chục nghìn axít amin. Tập các sắp hàng thường được ký hiệu là $\mathbf{A} = \{D^1, \ldots D^N\}$. Trong đó N là số lượng sắp hàng còn D^a ($1 \le a \le N$) là ký hiệu sắp hàng thứ a trong tập **A**.

Bài toán: Ước lượng mô hình biến đổi axít amin để mô tả quá trình tiến hóa của các chuỗi prôtêin đầu vào.

Dữ liệu ra: Một mô hình biến đổi axít amin **Q** thể hiện quá trình tiến hoá của các chuỗi axít amin ở dữ liệu đầu vào **A**.

Ước lượng mô hình **Q** là một bài toán phức tạp bởi ta phải xác định một lượng lớn tham số. Các phương pháp có thể chia theo hai hướng tiếp cận chính: phương pháp đếm (counting approach) và phương pháp hợp lý nhất (maximum likelihood approach).

1.4. Các phương pháp ước lượng mô hình biến đổi axít amin

1.4.1. Phương pháp đếm

Trong phương pháp đếm, các tham số cần ước lượng của mô hình được tính toán một cách trực tiếp từ dữ liệu. Hai ma trận phổ biến được ước lượng bằng phương pháp đếm là PAM và BLOSUM.

1.4.1.1. Ma trận PAM (Point Accepted Mutation)

Tác giả của mô hình PAM là Dayhoff và các cộng sự đã sử dụng bộ dữ liệu gồm 71 nhóm prôtêin, trong đó mỗi nhóm bao gồm các chuỗi prôtêin có quan hệ gần nhau (giống nhau ít nhất 85%). Sự giống nhau cao giữa các chuỗi prôtêin giúp đảm bảo các biến đổi trực tiếp giữa các axít amin (ví dụ A → R) chiếm phần lớn, còn các biến đổi gián tiếp (ví dụ A→ X → R) chỉ chiếm phần nhỏ.

Ma trận PAM1 cho biết xác suất thay thế giữa các axít amin nếu có khoảng 1% tổng số axít amin bị biến đổi. Các giá trị của ma trận PAM1 cho biết xác suất biến đổi từ axít amin i thành axít amin j sau một đơn vị thời gian. Các phần từ không nằm trên đường chéo chính của ma trận được tính bởi công thức:

$$\text{PAM1}(i,j) = \frac{\lambda m_j b_{ij}}{\sum_{i \in S} b_{ij}} \quad (0.9)$$

trong đó m_j là độ đột biến của axít amin j, được tính tương đối so với các axít amin khác; b_{ij} là số lần biến đổi giữa hai axít amin i và j quan sát được từ dữ liệu và λ là hằng số được chọn sao cho tổng số biến đổi trên toàn bộ dữ liệu là 1%. Các phần tử nằm trên đường chéo chính của ma trận PAM được chọn sao cho tổng của bất kỳ cột nào cũng bằng một.

1.4.1.2. Ma trận BLOSUM (BLOcks SUbstitution Matrix)

Ma trận BLOSUM được giới thiệu lần đầu tiên bởi Henikoff và Henikoff vào năm 1992. Ma trận này được dùng chủ yếu cho bài toán sắp hàng đa chuỗi. Các tác giả đã sử dụng bộ dữ liệu BLOCKS, đây là bộ dữ liệu chứa các chuỗi prôtêin do chính nhóm tác giả xây dựng. Họ đã tìm các đoạn bảo tồn (conserved regions) để từ đó tính ra các tần số xuất hiện của các axít amin và xác suất biến đổi giữa các cặp các axít amin. Sau đó, các tác giả tính giá trị log-odds cho mỗi cặp biến đổi axít amin có thể có.

1.4.2. Phương pháp cực đại khả năng (maximum likelihood)

1.4.2.1. Giới thiệu chung

Một trong các nhược điểm chính của các phương pháp đếm là chỉ áp dụng được cho các tập dữ liệu có độ tương đồng cao. Để khắc phục hạn chế trên, phương pháp cực đại khả năng (maximum likelihood, viết tắt là ML) đã được đề xuất để xây dựng mô hình **Q**. Một số nghiên cứu đã chỉ ra rằng phương pháp cực đại khả năng có thể giúp tránh các lỗi có tính hệ thống và giúp tận dụng các thông tin trong các sắp hàng đa chuỗi prôtêin hiệu quả hơn so với phương pháp đếm. Năm 1996, nhóm tác giả Adachi và Haseqawa sử dụng phương pháp ML để phân tích các chuỗi prôtêin ti thể của 20 loài động vật có xương sống để xây dựng mô hình mtREV. Nhóm tác giả cho thấy mô hình mtREV tốt hơn các mô hình khác khi phân tích quá trình tiến hóa giữa các loài sinh vật dựa vào các chuỗi prôtêin ti thể.

Tuy nhiên, thời gian tính toán là một trong những cản trở lớn nhất trong việc áp dụng phương pháp ML trên những tập dữ liệu prôtêin lớn. Nhóm tác giả Whelan và Goldman đã đề xuất phương pháp ML xấp xỉ và áp dụng trên cơ sở dữ liệu gồm 3905 chuỗi prôtêin và xây dựng mô hình WAG vào năm 2002. Mô hình WAG cho kết quả tốt hơn các mô hình khác khi được dùng để phân tích quá trình tiến hóa giữa các sinh vật dựa vào các chuỗi prôtêin.

Gần đây nhất, vào năm 2008, nhóm tác giả Le và Gascuel đã cải tiến phương pháp của Whelan và Goldman bằng cách kết hợp thêm thông tin về tính không đồng nhất trong tốc độ biến đổi theo vị trí vào quá trình xây dựng mô hình **Q**.

1.4.2.2. Ước lượng mô hình bằng phương pháp cực đại khả năng

Giả sử $D = \{D_1, \dots D_l\}$ là một sắp hàng đa chuỗi có chiều dài l trong đó D_i ($1 \le i \le l$) là vị trí thứ i của sắp hàng. Gọi T là cây phân loài tương ứng với sắp hàng đa chuỗi D. Sử dụng mô hình **Q** như đã trình bày ở phần 1.2.1, giá trị likelihood của **Q** và T đối với D được tính theo công thức:

$$L(\mathbf{Q},T \mid D) = \prod_{i=1}^{l} L(\mathbf{Q},T \mid D_i) \tag{1.10}$$

trong đó $L(\mathbf{Q},T|D_i)$ là likelihood của **Q** và T đối với vị trí D_i , giá trị này có thể tính một cách hiệu quả bằng thuật toán cắt tỉa của Felsenstein.

Phương pháp cực đại khả năng để ước lượng mô hình biến đổi axít amin được giới thiệu lần đầu bởi Adachi và Haseqawa. Giả sử chúng ta có một bộ dữ liệu gồm N sắp hàng đa chuỗi prôtêin ký hiệu là $\mathbf{A} = \{D^1, \dots D^N\}$. Ký hiệu $\mathbf{T} = \{ T^1, T^2, \dots T^N \}$ là tập các cây, trong đó mỗi $T^a \in \mathbf{T}$ là cây tương ứng được xây dựng từ sắp hàng D^a với mô hình **Q**. Giá trị likelihood của mô hình **Q** và **T** được tính theo công thức:

$$L(\mathbf{Q}, \mathbf{T}) = \prod_{a=1}^{N} L(\mathbf{Q},T^a \mid D^a) \tag{1.11}$$

Mô hình **Q** khi đó được ước lượng bằng cách tìm cực đại của giá trị likelihood $L(\mathbf{Q}, \mathbf{T})$ theo công thức sau:

$$\mathbf{Q} = \arg\max_{\mathbf{Q}} \{L(\mathbf{Q}, \mathbf{T})\} \tag{1.12}$$

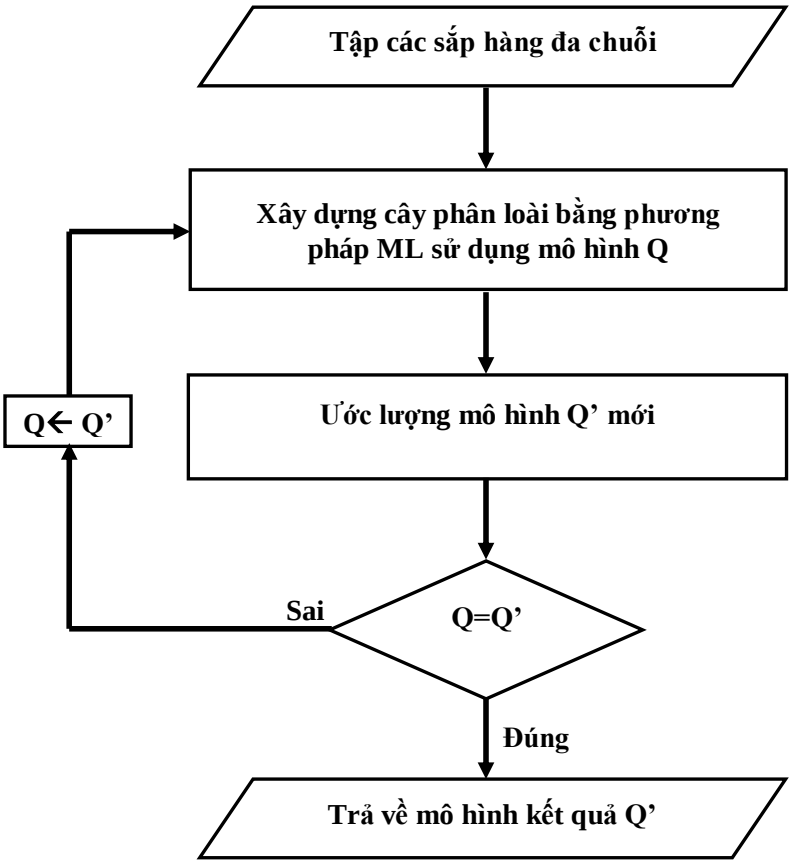
Quá trình tìm cực đại cho giá trị likelihood $L(\mathbf{Q}, \mathbf{T})$ theo công thức 1.11 là một bài toán rất khó vì chúng ta phải tối ưu cùng lúc các tham số của mô hình **Q** cùng tất cả các cây phân loài **T** (bao gồm cả cấu trúc và độ dài các cạnh). Các nghiên cứu đã chỉ ra rằng các hệ số của **Q** được ước lượng tương đối chính xác khi sử dụng cây phân loài gần tối ưu. Vì vậy, công thức 1.11 có thể được đơn giản hóa và xấp xỉ bởi:

$$L(\mathbf{Q}, \mathbf{T}) = \prod_{a=1}^{N} L(\mathbf{Q} \mid T^{*a}, D^a) \tag{1.13}$$

với T^{*a} là cây phân loài *gần tối ưu* của D^a. Do đó công thức để ước lượng mô hình **Q** có dạng:

$$\mathbf{Q} = \arg\max_{\mathbf{Q}} \left\{ \prod_{a=1}^{N} L(\mathbf{Q} \mid T^{*a}, D^a) \right\} \quad (1.14)$$

Lược đồ thuật toán ước lượng mô hình biến đổi axít amin bằng phương pháp cực đại khả năng được trình bày ở Hình 1.1 (xem chương 2 để biết thêm chi tiết về thuật toán).



Hình 1.1: Lược đồ quá trình ước lượng mô hình biến đổi axít amin bằng phương pháp ML.

1.5. Xây dựng cây phân loài bằng phương pháp ML

Trong phương pháp ML, cây "tốt nhất" được hiểu là cây có giá trị likelihood lớn nhất. Giá trị likelihood của một cây T đối với một mô hình biến đổi **Q** và dữ liệu D được tính như sau:

$$L(T \mid \mathbf{Q}, D) = \prod_{i=1}^{l} L(T \mid \mathbf{Q}, D_i) \quad (1.15)$$

Như vậy chúng ta sẽ cần tìm cây T (bao gồm cấu trúc cây và độ dài các cạnh) sao cho giá trị likelihood theo công thức 1.15 đạt cực đại.

Bài toán tối ưu cây T là một bài toán NP-khó do số lượng cây có cấu trúc khác nhau tương ứng với cùng một sắp hàng là (2n-5)!!. Số lượng này tăng

nhanh theo số lượng chuỗi. Một số phương pháp tìm kiếm gần đúng đã được đề xuất.

Chương 2. PHƯƠNG PHÁP ƯỚC LƯỢNG NHANH MÔ HÌNH BIẾN ĐỔI AXÍT AMIN BẰNG PHƯƠNG PHÁP CỰC ĐẠI KHẢ NĂNG

2.1. Giới thiệu

Phương pháp cực đại khả năng cho kết quả tốt tuy nhiên chúng yêu cầu một lượng tính toán lớn cho nên rất khó áp dụng cho các bộ dữ liệu lớn. Một trong những bước tốn nhiều thời gian nhất trong quá trình xây dựng mô hình **Q** là xây dựng cây phân loài từ các sắp hàng đa chuỗi. Luận án đề xuất một phương pháp mới để vượt qua trở ngại này bằng cách phân chia các sắp hàng lớn thành những sắp hàng nhỏ mà vẫn giữ được các thông tin của các ma trận cần ước lượng. Thực nghiệm với cả hai bộ dữ liệu Pfam và FLU cho thấy phương pháp cải tiến này nhanh hơn so với phương pháp tốt nhất hiện nay từ ba đến sáu lần trong khi các ma trận ước lượng vẫn gần như không khác biệt. Như vậy, phương pháp cải tiến này sẽ cho phép các nhà nghiên cứu ước lượng các ma trận từ những tập dữ liệu rất lớn.

2.2. Ước lượng mô hình bằng phương pháp cực đại khả năng

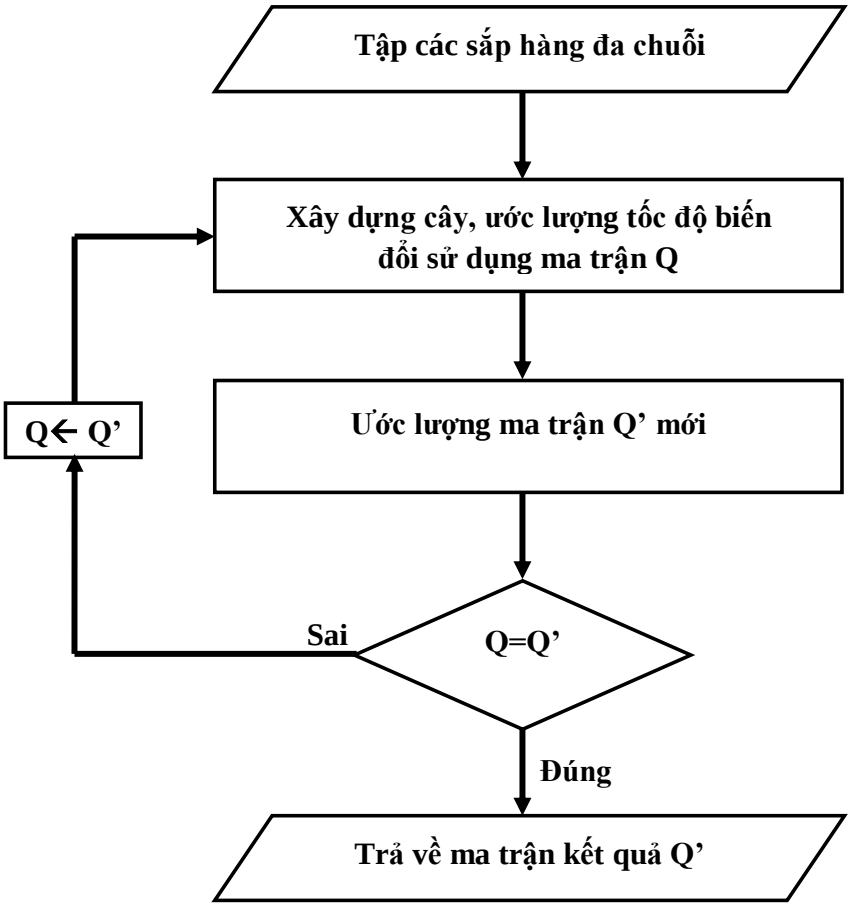
Cho một tập dữ liệu các sắp hàng đa chuỗi prôtêin *A*, nhiệm vụ của chúng ta là ước lượng ma trận **Q** sao cho **Q** thể hiện chính xác nhất tất cả các quá trình biến đổi trong các chuỗi prôtêin này.

Thông thường, tập dữ liệu *A* có thể bao gồm hàng trăm sắp hàng đa chuỗi prôtêin và chứa đến hàng trăm ngàn chuỗi prôtêi. Cụ thể ba bước của quá trình ước lượng ma trận **Q** bằng phương pháp ML là: (xem thêm Hình 2.1).

Xây dựng cây bằng ML: Xây dựng cây phân loài từ các sắp hàng sử dụng ma trận **Q** bằng phương pháp ML.

Ước lượng các tham số của mô hình: ước lượng ma trận **Q'** mới từ tất cả các sắp hàng và cây tương ứng ở bước *Xây dựng cây* bằng thuật toán cực đại kỳ vọng (expectation maximization).

So sánh mô hình: So sánh **Q** và **Q'**. Nếu **Q'** ~ **Q,** kết thúc và **Q'** là ma trận kết quả. Nếu không, thay **Q** bằng **Q'** và quay lại bước *Xây dựng cây.*



Hình 2.1: Lược đồ quá trình ước lượng mô hình biến đổi axít amin.

2.3. Các phương pháp chia tách dữ liệu

Trong mục này, dựa vào các phân tích của mục trước, luận án trình bày hai phương pháp để tăng tốc quá trình xây dựng cây phân loài. Ý tưởng ở đây là chia nhỏ các sắp hàng kích thước lớn thành nhiều sắp hàng kích thước nhỏ hơn. Với các sắp hàng kích thước nhỏ, quá trình xây dựng cây có thể được tăng tốc rất nhiều.

2.3.1. Phương pháp chia tách ngẫu nhiên

Đây là một ý tưởng đơn giản để giảm số lượng chuỗi trong mỗi sắp hàng. Xét một sắp hàng D^a gồm m chuỗi và một số nguyên dương k ($k \geq 4$) là ngưỡng chia tách. Các chuỗi của sắp hàng D^a được chia tách ngẫu nhiên thành các sắp hàng nhỏ có số lượng chuỗi nằm trong đoạn từ k đến $2k$. Các sắp hàng nhỏ này sẽ được sử dụng để ước lượng mô hình **Q**. Giả sử M là mô hình được ước lượng từ các sắp hàng không chia tách thì M_k^{R} sẽ là mô hình được ước lượng từ các sắp hàng được chia tách ngẫu nhiên với ngưỡng k. Ví dụ $\mathrm{LG}_8^{\mathrm{R}}$ là mô hình được ước lượng với cùng bộ dữ liệu như mô hình LG nhưng các sắp hàng có kích thước từ 8 đến 16 chuỗi. Các bước cụ thể của phương pháp chia tách sắp hàng ngẫu nhiên được trình bày ở **Thuật toán 2.1**.

procedure Thuật toán chia tách ngẫu nhiên;
input: Một sắp hàng D^a với m chuỗi axít amin và số nguyên dương $k \geq 4$;
output: Các sắp hàng con với kích thước từ k đến $2k$;
begin
while (số lượng chuỗi trong $D^a \geq k + 4$)

- Sinh ngẫu nhiên một số tự nhiên s thoả mãn $k \leq s \leq 2k$;
- Chọn ngẫu nhiên s chuỗi trong D^a để tạo thành một sắp hàng con;
- Loại bỏ các chuỗi đã chọn ra khỏi D^a;

endwhile;
Đưa ra tất cả các sắp hàng con;
end;

Thuật toán 2.1: Thuật toán chia tách sắp hàng ngẫu nhiên.

2.3.2. Phương pháp chia tách dựa theo cấu trúc cây

Phương pháp chia tách ngẫu nhiên có thể tạo ra các sắp hàng nhỏ chứa các chuỗi có quan hệ xa. Điều này có thể dẫn tới các cây phân loài tương ứng với các sắp hàng nhỏ này có độ chính xác không cao và làm giảm độ chính xác cuả mô hình **Q**. Để khắc phục vấn đề này, chúng tôi đề xuất một phương pháp tách dựa trên cấu trúc cây.

Phương pháp này dựa theo tư tưởng của thuật toán BIONJ. Thuật toán có độ phức tạp là $O(m^3)$ với m là số chuỗi. Trong phương pháp chia tách dựa theo cấu trúc cây, các chuỗi lần lượt được nhóm lại nếu như số lượng chuỗi trong nhóm mới nằm trong đoạn từ k đến $2k$.. Cụ thể phương pháp chia tách dựa theo cấu trúc cây gồm các bước như trong Thuật toán 2.2 sau đây:

procedure Thuật toán chia tách dựa theo cấu trúc cây;
input: Sắp hàng D^a với m chuỗi axít amin và số nguyên dương $k \geq 4$;
output: Các sắp hàng con với kích thước từ k đến $2k$;
begin
Mỗi chuỗi prôtêin của D^a được coi như một nhóm. Tính tất cả các khoảng cách giữa hai nhóm một dựa vào ma trận khoảng cách và thuật toán BIONJ;
repeat
Tìm hai nhóm có khoảng cách nhỏ nhất, giả sử là G_1 và G_2. Gọi m_1 và m_2 là số lượng chuỗi của G_1 và G_2 tương ứng;

```
        if  m1 + m2 ≤ 2k  then
            Kết hợp G1 và G2 thành một nhóm mới;
            Tính toán lại khoảng cách giữa nhóm mới này và các nhóm khác
            theo thuật toán BIONJ;
        else / / m1 > k hoặc m2 > k
            if m1 > k  then
                Xem G1 là một sắp hàng con;
            else / / s2 > k
                Xem G2 là một sắp hàng con;
            endif
        endif
    until (chỉ còn một nhóm G0);
    Giả sử m0 là số lượng chuỗi của G0.
    if  m0≥3  then
        Xem G0 là một sắp hàng con;
    else
        Kết hợp G0 vào một sắp hàng con trước đó
    Đưa ra tất cả các sắp hàng con;
end;
```

Thuật toán 2.2: Thuật toán chia tách sắp hàng dựa theo cấu trúc cây.

2.4. Kết quả

Các thực nghiệm với hai bộ dữ liệu Pfam và vi rút cúm cho thấy phương pháp chia tách dựa trên cấu trúc cây cho kết quả tốt. Với ngưỡng k = 8, phương pháp chia tách dựa trên cây tốt như phương pháp không chia tách trên cả hai bộ dữ liệu nhưng thời gian ước lượng mô hình nhanh hơn từ ba đến sáu lần. Như vậy, các phương pháp chia tách này cho phép các nhà nghiên cứu ước lượng mô hình từ bộ dữ liệu lớn với thời gian giảm đáng kể. Phương pháp tách dựa trên cây với ngưỡng k = 8 được chúng tôi khuyên dùng để có một kết quả tốt và hiệu quả. Các kết quả nghiên cứu của chương này đã được công bố tại hội nghị quốc tế KSE năm 2011 (công trình khoa học số 3).

Chương 3. XÂY DỰNG MÔ HÌNH BIẾN ĐỔI ĐA MA TRẬN

Phần lớn các mô hình biến đổi axít amin sử dụng một ma trận để mô hình hoá sự biến đổi giữa các axít amin. Tuy nhiên quá trình biến đổi ở các vị trí trên chuỗi axít amin là không giống nhau và phụ thuộc vào nhiều yếu tố. Trong hầu hết các trường hợp, một ma trận là không đủ để mô hình hóa sự phức tạp của quá trình biến đổi giữa các axít amin. Ở chương này, chúng tôi sẽ nghiên cứu việc sử dụng mô hình với nhiều ma trận cho các vị trí khác nhau trên chuỗi axít amin.

3.1. Tính không đồng nhất của tốc độ biến đổi theo vị trí

Nhiều nghiên cứu đã chỉ ra rằng tốc độ biến đổi có tính không đồng nhất, tức là tốc độ biến đổi giữa các vị trí khác nhau trong cùng một chuỗi có sự khác biệt đáng kể. Hiện tượng này thường được giải thích bởi sự hiện diện của các nhu cầu tiến hóa khác nhau ở các vị trí khác nhau. Để không bỏ qua hiện tượng quan trọng này, chúng ta cần sử dụng một mô hình phân phối biểu diễn tốc độ biến đổi axít amin tại các vị trí khác nhau trong chuỗi prôtêin .

Tính không đồng nhất của tốc độ biến đổi axít amin tại các vị trí khác nhau có thể được mô hình hoá bằng một phân phối gamma (Γ) với kỳ vọng là 1,0 và phương sai là 1/α (α>0) theo công thức sau:

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

3.2. Mô hình biến đổi đa ma trận

Với mô hình chuẩn ta cần ước lượng 208 tham số của mô hình **Q**. Ký hiệu D là một sắp hàng, T là cây cây phân loài tương ứng của D được xây dựng bằng phương pháp ML với mô hình **Q**. Khi đó likelihood của **Q** và T đối với D được tính theo công thức:

$$L(\mathbf{Q}, T \mid D) = \prod_{i=1}^{l} L(\mathbf{Q}, T \mid D_i) \qquad (3.1)$$

trong đó $D = \{D_1, \ldots D_l\}$ là một sắp hàng đa chuỗi có chiều dài l và D_i ($1 \le i \le l$) là vị trí thứ i của sắp hàng. Yang đã giới thiệu một mô hình hỗn hợp dựa trên một mô hình biến đổi axít amin duy nhất nhưng tốc độ của các vị trí biến thiên theo một phân phối gamma rời rạc với c phân loại tốc độ có trọng số bằng nhau. Likelihood được tính bằng công thức:

$$L(\mathbf{Q},T,\alpha \mid D)=\prod_{i=1}^{l}\left(\frac{1}{c}\sum_{k=1}^{c}L(\Gamma(\alpha,k)\mathbf{Q},T \mid D_i)\right) \quad (3.2)$$

với Γ(α, *k*) là tốc độ thứ *k* của một phân bố gamma rời rạc với tham số α. Các trọng số của các tốc độ đều bằng 1/*c*. Cả T và α được ước tính bằng phương pháp ML từ tập dữ liệu đầu vào.

Mô hình đa ma trận đã được đề xuất trong một số nghiên cứu. Với các mô hình đa ma trận này, likelihood được tính như sau:

$$L(\mathbf{Q}=\{Q_1,..,Q_M\},T,W=\{w_1,...,w_M\} \mid D)=\prod_{i=1}^{l}\left(\sum_{m=1}^{M}w_m L(Q_m,T \mid D_i)\right) \quad (3.3)$$

trong đó *M* là số lượng ma trận và w_m là trọng số của ma trận Q_m với điều kiện $\sum_{m=1}^{M}w_m=1$.

Các nghiên cứu gần đây đã kết hợp mô hình của Yang (công thức 3.2) với công thức 3.3 ở trên để tạo thành mô hình đa ma trận:

$$\begin{aligned}&L(\mathbf{Q}=\{Q_1,..,Q_M\},T,W=\{w_1,...,w_M\},\alpha \mid D)\\&=\prod_{i=1}^{l}\left(\sum_{m=1}^{M}\frac{w_m}{c}\sum_{k=1}^{c}L(\Gamma(\alpha,k)Q_m,T \mid D_i)\right)\end{aligned} \quad (3.4)$$

với điều kiện $\sum_{m=1}^{M}w_m=1$ vẫn được giữ nguyên.

Công thức 3.4 thể hiện hai cấp độ hỗn hợp, một cho các loại tốc độ phân phối gamma và một cho các ma trận thay thế. Các mô hình tương ứng là EX2 (bao gồm hai ma trận) và UL3 (bao gồm ba ma trận).

Trong luận án này, chúng tôi đơn giản hóa công thức 3.4. Mặc dù các mô hình EX2, UL3 là tốt nhưng chúng yêu cầu một lượng tính toán lớn và tốn nhiều bộ nhớ. Điều này chủ yếu là do số lượng lớn các phân loại vị trí, ví dụ như UL3 có tới 12 phân loại vị trí và 4 phân loại gamma. Để đơn giản hóa công thức 3.4, chúng tôi sử dụng bốn phân loại tốc độ và bốn ma trận tương ứng (*c* = 4, M = 4). Các trọng số của cả 4 phân loại đều được cho bằng ¼. Mô hình với bốn ma trận này được đặt tên là LG4M. Giả sử **Q** = (Q_1, Q_2, Q_3, Q_4) là tập bốn ma trận, khi đó likelihood của mô hình **Q**, cây phân loài *T* và tham số *α* được tính như sau:

$$L(\mathbf{Q},T,\alpha \mid D)=\prod_{i=1}^{l}\left(\frac{1}{4}\sum_{k=1}^{4}L(\Gamma(\alpha,k)Q_k,T \mid D_i)\right) \quad (3.5)$$

Công thức 3.5 này là một sự kết hợp giữa công thức 3.2 của Yang và công thức 3.4 của các mô hình hỗn hợp hai cấp. Thay vì dùng chung một ma trận như

trong mô hình của Yang, mỗi tốc độ có ma trận riêng và mỗi ma trận được áp dụng chỉ cho một loại tốc độ thay vì cho tất cả các tốc độ như trong mô hình hỗn hợp hai cấp. Như vậy, công thức 3.5 là tổng quát hơn so với mô hình của Yang, nhưng vẫn giữ các tham số tự do được ước tính từ các dữ liệu (α và T) như trong mô hình của Yang.

Mô hình LG4M trong công thức 3.5 sử dụng một phân phối gamma rời rạc để phân lớp các tốc độ biến đổi giữa các axít amin theo vị trí. Chúng tôi loại bỏ đi phân phối gamma để có một mô hình tổng quát hơn, gọi là mô hình LG4X. Likelihood khi đó được tính như sau:

$$L(\mathbf{Q}, T, P = \{\rho_1, \rho_2, \rho_3, \rho_4\}, W = \{w_1, w_2, w_3, w_4\} \mid D) = \prod_{i=1}^{l} \left(\sum_{k=1}^{4} w_k L(\rho_k Q_k, T \mid D_i) \right) \quad (3.6)$$

trong đó w_k và ρ_k là các trọng số và tốc độ của ma trận Q_k thoả mãn $\sum_{k=1}^{4} w_k = 1$ và $\sum_{k=1}^{4} w_k \rho_k = 1$. Như vậy LG4X chỉ còn có 3 trọng số w_k và 3 tốc độ ρ_k là các tham số cần ước lượng.

3.3. Thuật toán ước lượng mô hình

Dựa vào các lập luận trong mục 3.2, chúng ta có thuật toán ước lượng mô hình như trong Thuật toán 3.1 sau đây:

procedure Thuật toán ước lượng mô hình;
input: Tập N sắp hàng A = { D^1 , …, D^N }, mô hình khởi tạo ban đầu **S**;
output: Mô hình **Q** = {**Q**$_1$, **Q**$_2$, **Q**$_3$, **Q**$_4$};
begin
  Q = {**Q**$_1$ = **Q**$_2$ = **Q**$_3$ = **Q**$_4$ = **S**};
  repeat
    foreach sắp hàng D^a trong A
      - $T^a \leftarrow$ Cây phân loài của D^a xây dựng bằng ML với **Q**;
      - Ước lượng các tốc độ $\rho^a = \rho_1^a$, …, ρ_4^a và các trọng số $w^a = w_1^a$, …, w_4^a;
      - Phân lớp cho vị trí $D^a{}_i$ của D^a vào tập $L_{c_i}^a$ sao cho thỏa mãn $c_i = \arg\max_{k=1...4} w_k L(T^a, \rho_k^a Q_k \mid D_i^a)$;
      - Chia các sắp hàng D^a và cây T^a thành 4 sắp hàng và 4 cây con theo phân lớp ở trên, các cây con được nhân với các tốc độ ρ_1^a,

…, ρ_4^a tương ứng: $(L_1^a, T^a * \rho_1^a)$, $(L_2^a, T^a * \rho_2^a)$, $(L_3^a, T^a * \rho_3^a)$, $(L_4^a, T^a * \rho_4^a)$;

end foreach;

for (k = 1..4)

Ước lượng mô hình $\mathbf{Q}^*_k$ từ các sắp hàng và cây con thuộc phân lớp k ở trên $(L_k^a, T^a * \rho_k^a)$ bằng thuật toán cực đại kỳ vọng với $\mathbf{Q}_k$ là mô hình khởi tạo ban đầu của thuật toán cực đại kỳ vọng;

endfor;

until ($Q_k \approx Q^*_k$ với mọi k)**;**

Q ← Q';

end;

Thuật toán 3.1: Thuật toán ước lượng mô hình LG4M và LG4X

3.4. Kết quả

Các thực nghiệm với bộ dữ liệu HSSP và TreeBase cho thấy LG4M và LG4X cho các cây có likelihood cao hơn và cấu trúc khác so với các mô hình đơn ma trận. Như vậy cả hai mô hình mới của chúng tôi đều cho kết quả tốt hơn các mô hình đơn ma trận trong khi chỉ cần cùng một lượng bộ nhớ và thời gian thực hiện. Các kết quả nghiên cứu của chương này đã được công bố trên tạp chí quốc tế *Molecular Biology and Evolution* năm 2012 (công trình khoa học số 5).

Chương 4. HỆ THỐNG ƯỚC LƯỢNG MÔ HÌNH TỰ ĐỘNG

4.1. Giới thiệu

Nhiều mô hình biến đổi axít amin chung đã được đề xuất như JTT, WAG và LG. Ngoài ra, một số mô hình cho các tập dữ liệu riêng biệt đã được đề xuất như HIVw và HIVb cho vi rút HIV; FLU cho vi rút cúm, mtREV cho các prôtêin ty thể). Các mô hình riêng biệt này thường cho kết quả tốt hơn các mô hình chung khi áp dụng cho các nhóm prôtêin tương ứng. Do đó, việc ước lượng mô hình cho các tập dữ liệu riêng biệt là cần thiết.

Chúng tôi muốn xây dựng một hệ thống tự động để đáp ứng nhu cầu trên. Hệ thống cần phục vụ được cùng lúc nhiều người dùng và thời gian chờ của người dùng càng ngắn càng tốt. Do đó chúng tôi đã nghiên cứu và áp dụng một cải tiến khác để tăng tốc quá trình ước lượng mô hình.

Trong phương pháp ước lượng mô hình **Q,** bước tối ưu cấu trúc cây bằng ML được lặp lại nhiều lần. Các nghiên cứu đã chỉ ra rằng ước lượng mô hình với các cây gần tối ưu cũng cho các mô hình có chất lượng tốt. Từ đây chúng tôi đề xuất một phương pháp ước lượng nhanh với chỉ một lần tối ưu cấu trúc cây.

4.2. Phương pháp ước lượng nhanh

Chúng tôi thống kê với nhiều tập dữ liệu và bộ tham số khác nhau thì số lần lặp ước lượng lại ma trận Q trung bình là 3 và bước xây dựng cây bằng ML là tốn thời gian nhất. Từ những phân tích này, thuật toán được cải tiến như sau:

- Chỉ tối ưu cấu trúc cây một lần duy nhất ở lần lặp 2.
- Thay thế tần số axít amin trong mô hình khởi tạo ban đầu bằng tần số axít amin của dữ liệu.
- Sử dụng 4 phân loại tốc độ gamma.

Các bước cụ thể của thuật toán ước lượng nhanh mô hình biến đổi axít amin được trình bày trong Thuật toán 4.1 sau đây:

procedure Thuật toán ước lượng nhanh;
input: Tập N sắp hàng $A = \{D^1, \ldots D^N\}$ và mô hình khởi tạo ban đầu **S**;
output: Mô hình **Q**;
begin
    Thay thế tần số axít amin trong S bằng tần số tính từ dữ liệu;
    Q ← **S**;
    for (i = 1 .. 3)

foreach sắp hàng D^a trong A

if (i == 1) **then**

$T^a \leftarrow$ Cây phân loài của D^a xây dựng bằng thuật toán BioNJ;

endif;

if (i == 2) **then**

Tối ưu cấu trúc của T^a với **Q** bằng thuật toán SPR;

endif;

- Tối ưu độ dài các cạnh của T^a với **Q**;
- Tối ưu tham số của phân phối gamma với 4 phân lớp tốc độ biến đổi theo vị trí;
- Tách D^a thành 4 sắp hàng con D_1^a, D_2^a, D_3^a, D_4^a dựa theo xác suất của các phân phối tốc độ theo vị trí.
- Tạo ra 4 cây con T_1^a, T_2^a, T_3^a, T_4^a có cấu trúc giống T^a, các cạnh của 4 cây con được nhân tỷ lệ theo các tốc độ đã ước lượng của mỗi phân loại theo phân phối gamma;

end foreach;

Ước lượng ma trận **Q'** từ các sắp hàng và cây con ở trên bằng thuật toán EM với **Q** là ma trận khởi tạo ban đầu;

Q ← **Q'**;

endfor;

Đưa ra **Q**;

end;

Thuật toán 4.1: Thuật toán ước lượng nhanh mô hình biến đổi axít amin.

Trong thuật toán cải tiến, mỗi lần lặp chúng tôi chỉ tối ưu lại tham số gamma và chiều dài cạnh của cây ML đã xây dựng ở lần chạy trước với mô hình **Q** mới mà không tối ưu cấu trúc cây. Chúng tôi chỉ thực hiện tối ưu cấu trúc cây tại lần lặp thứ 2 (i=2). Cải tiến này giúp giảm thời gian đáng kể do thuật toán tối ưu cấu trúc của cây tốn rất nhiều thời gian.

4.3. Kết quả

Các thực nghiệm với hai bộ dữ liệu Pfam và FLU cho thấy trung bình tốc độ ước lượng bằng phương pháp mới giảm 50% so với phương pháp truyền thống. Mô hình ước lượng bằng phương pháp mới gần như giống hệt với mô hình ước lượng bằng phương pháp truyền thống (độ tương quan Pearson lớn hơn 0,999). Giá trị likelihood chênh lệch giữa hai mô hình là không đáng kể. Các cấu trúc

cây cũng không có nhiều khác biệt giữa các mô hình được ước lượng bằng hai phương pháp.

Chúng tôi đã ứng dụng phương pháp mới để xây dựng một hệ thống ước lượng tự động các ma trận biến đổi từ dữ liệu của người dùng. Các kết quả nghiên cứu của chương này đã được công bố trên tạp chí quốc tế *Bioinformatics* năm 2011 (công trình khoa học số 2).

Chương 5. MÔ HÌNH BIẾN ĐỔI AXIT AMIN CHO VIRÚT CÚM

5.1. Giới thiệu về vi rút cúm và sự cần thiết của các mô hình biến đổi axít amin riêng biệt cho từng loài

Các mô hình biến đổi axít amin chung bởi chúng như PAM, JTT, WAG, LG được xây dựng dựa vào một tập các chuỗi prôtêin từ các loài sinh vật khác nhau. Chúng được sử dụng để phân tích các chuỗi prôtêin của tất cả các loài. Tuy nhiên, những nghiên cứu mới nhất gần đây cho thấy các mô hình chung này không cho kết quả tốt nhất khi sử dụng để phân tích dữ liệu prôtêin của một số loài cụ thể riêng biệt, ví dụ như các loại vi rút HIV. Nguyên nhân là vì các mô hình chung này không thể phản ánh đầy đủ bản chất sinh học, hóa học cũng như quá trình tiến hóa của một số loài sinh vật riêng biệt.

Do đó, một hướng mới đang được các nhà nghiên cứu quan tâm và phát triển là xây dựng các mô hình biết đổi axít amin riêng biệt cho các đối tượng sinh vật khác nhau. Năm 2007, Nickle và đồng nghiệp áp dụng phương pháp hợp lý nhất được đề xuất bởi Whelan và Goldman để xây dựng mô hình biến đổi axít amin cho vi rút HIV. Nhóm tác giả xây dựng hai mô hình, HIVw để mô phỏng quá trình biến đổi của vi rút bên trong người bệnh, và HIVb để mô phỏng quá trình biến đổi của vi rút giữa các người bệnh. Các kết quả của nhóm tác giả cho thấy HIVb và HIVw tốt hơn các mô hình chung khác.

Trong những năm gần đây, dịch bệnh do vi rút cúm đang xảy ra trên toàn thế giới. Từ đó nổi lên vấn đề cần phải nghiên cứu toàn diện về loại vi rút nguy hiểm này, đặc biệt là các nghiên cứu về quá trình tiến hóa, lan truyền và lây nhiễm của chúng.

Vi rút cúm là một loại vi rút RNA và thuộc họ Orthomyxoviridae. Chúng được chia thành ba loại là: cúm A, cúm B và cúm C, trong đó có cúm A là phổ biến và nguy hiểm nhất. Trong những năm gần đây, vi rút cúm A đã gây ra nhiều vấn đề nghiêm trọng cho sức khỏe con người và kinh tế xã hội, nổi bật là dịch bệnh H5N1 (cúm gia cầm) và cúm H1N1.

Do đó trong chương này, luận án đề xuất mô hình FLU cho vi rút cúm để giúp tăng cường sự hiểu biết của chúng ta về sự tiến hóa của loại vi rút này. Mô hình FLU được xây dựng với phương pháp ước lượng nhanh đã đề xuất trong Chương 2. Các kết quả thực nghiệm đã chỉ ra rằng FLU tốt hơn hẳn các mô hình hiện tại khi phân tích prôtêin của vi rút cúm.

5.2. Ước lượng mô hình FLU

Chúng tôi sử dụng bộ dữ liệu chuẩn của vi rút cúm, kết hợp với phương pháp chia tách sắp hàng theo cấu trúc cây ở chương 2 để ước lượng mô hình FLU. Ngưỡng chia tách được chọn bằng 8 (k=8), có nghĩa là các sắp hàng sau khi được chia tách sẽ có kích thước từ 8 đến 16 chuỗi. Tổng số sắp hàng trước khi chia chia tách là 992, số lượng sắp hàng sau khi chia tách là 3970. Tiếp tục thực hiện các bước ước lượng mô hình như trong chương 2, chúng tôi có một mô hình biến đổi axít amin cho vi rút cúm gọi là FLU.

5.3. Kết quả

Chúng tôi đã ước lượng mô hình FLU cho dữ liệu vi rút cúm và thu được kết quả rất tốt. Các phân tích đã cho thấy sự khác biệt giữa FLU và các mô hình hiện tại ở cả véc tơ tần số axít amin và ma trận hệ số hoán đổi. Các thực nghiệm cho thấy FLU mô hình hoá các đặc điểm tiến hóa của vi rút cúm tốt hơn so với các mô hình chung. Cả hai thử nghiệm toàn cục và thử nghiệm chéo đều khẳng định rằng FLU tốt hơn so với các mô hình hiện tại trong việc xây dựng cây ML.

KẾT LUẬN

Các nghiên cứu về chuỗi axít amin đóng vai trò quan trọng trong sinh học phân tử và tin sinh học. Mô hình biến đổi axít amin là một thành phần có vai trò rất quan trọng trong nghiên cứu chuỗi axít amin. Phương pháp cực đại khả năng là một trong những phương pháp tốt nhất hiện nay để ước lượng mô hình biến đổi axít amin. Tuy nhiên các phương pháp hiện tại vẫn còn gặp nhiều hạn chế về thời gian thực hiện cũng như độ chính xác.

Luận án đã đề xuất hai cải tiến quan trọng để giảm thời gian của phương pháp ước lượng mô hình biến đổi axít amin hiện tại. Đề xuất đầu tiên là hai phương pháp chia tách nhỏ dữ liệu đầu vào giúp giảm đáng kể thời gian ước lượng mô hình. Đề xuất thứ hai là giảm bớt các bước tối ưu tham số khi xây dựng cây phân loài giúp giảm 50% thời gian ước lượng mô hình. Độ chính xác của các phương pháp cải tiến tương đương với phương pháp cũ.

Luận án cũng đưa ra một mô hình đa ma trận mới giúp mô hình hoá tốt hơn quá trình biến đổi của các chuỗi axít amin. Mô hình này cũng đã chứng tỏ được những ưu việt của nó so với các mô hình hiện tại khi độ chính xác được cải thiện đáng kể trong khi thời gian chạy vẫn tương đương với mô hình đơn ma trận.

Luận án đã xây dựng một hệ thống ước lượng mô hình tự động giúp ước lượng các ma trận biến đổi axít amin từ dữ liệu của người dùng. Hệ thống là kết quả nghiên cứu kết hợp cùng Viện nghiên cứu LIRMM, Cộng hoà Pháp. Hệ thống hoạt động được gần hai năm và đã có nhiều người sử dụng.

Chúng tôi cũng xây dựng mô hình FLU cho vi rút cúm. Mô hình FLU đã được tích hợp vào phần mềm xây dựng cây phân loài PhyML và đã chứng tỏ được hiệu quả khi phân tích các chuỗi axít amin của vi rút cúm. Mô hình này giúp tăng cường hiểu biết về vi rút cúm, giúp chúng ta có cách đối phó hữu hiệu hơn với loại vi rút rất nguy hiểm này.

Như vậy luận án đã tập trung phân tích và đề xuất những cải tiến cho các thành phần quan trọng nhất của phương pháp xây dựng mô hình biến đổi axít amin gồm: Dữ liệu đầu vào (Chương 2), Mô hình biến đổi (Chương 3) và Xây dựng cây phân loài bằng ML (Chương 4). Những cải tiến này đã giúp giảm đáng kể thời gian xây dựng và tăng độ chính xác của ma trận. Các kết quả của từng chương có thể gộp lại thành một kết quả thống nhất là những cải tiến cho phương pháp xây dựng ma trận biến đổi axít amin. Tuỳ vào điều kiện bài toán cụ thể mà chúng ta có thể lựa chọn áp dụng một hay nhiều cải tiến.

DANH MỤC CÁC CÔNG TRÌNH KHOA HỌC CỦA TÁC GIẢ LIÊN QUAN ĐẾN LUẬN ÁN

1. Cuong DC, Quang LS, Gascuel O, and Vinh LS (2010), "FLU, an amino acid substitution model for influenza proteins", *BMC Evolutionary Biology* vol. 10 (1), p. 99-110.
2. Cuong DC, Lefort V, Vinh LS, Quang LS and Gascuel O (2011), "ReplacementMatrix: a web server for maximum-likelihood estimation of amino acid replacement rate matrices", *Bioinformatics* vol. 27 (19), pp. 2758–2760.
3. Dat LV, Cuong DC, Quang LS and Vinh LS (2011), "A Fast and Efficient Method for Estimating Amino Acid Substitution Models", *Proc. of the 2011 Third International Conference on Knowledge and Systems Engineering*, pp. 85 –91.
4. Sau NV, Cuong DC, Quang LS and Vinh LS (2011), "Protein Type Specific Amino Acid Substitution Models for Influenza Viruses", *Proc. of the 2011 Third International Conference on Knowledge and Systems Engineering*, pp. 98 –103.
5. Quang LS, Cuong DC, and Gascuel O (2012), "Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates", *Mol Biol Evol* vol. 29 (10), pp. 2921–2936.