Quá trình tiến hóa sinh kháng thuốc của vi-rút HIV với cây đột biến di truyền theo mô hình Markov

Nguyễn Văn Thế[1], Tạ Văn Nhân[2], Nguyễn Thị Kim Duyên[1], Trịnh Mai Phương[1], Nguyễn Thị Hồng Minh[1]
[1] Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội, Hà Nội
[2] Công ty LOBI Việt Nam, Hà Nội

Tác giả liên hệ: Tạ Văn Nhân, tavannhan@gmail.com
Ngày nhận bài: 23/09/2021, ngày sửa chữa: 29/10/2021, ngày duyệt đăng: 15/11/2021
Định danh DOI: 10.32913/mic-ict-research-vn.v2021.n2.1014

Tóm tắt: Trong bài báo này, chúng tôi dự đoán quá trình tiến hóa của vi-rút HIV qua 14 đột biến kháng thuốc trong phác đồ điều trị sử dụng thuốc Efavirenz bằng mô hình Markov ẩn và cây đột biến di truyền. Với dữ liệu mới gồm 396 bệnh nhân trên cơ sở dữ liệu kháng thuốc HIV của trường đại học Stanford, chúng tôi tiến hành kiểm định giả thiết và nhận thấy dữ liệu phù hợp để đưa vào mô hình tính toán. Phần thực nghiệm cho thấy thuật toán EM dùng để ước lượng và tối ưu tham số khi áp dụng vào mô hình có tốc độ hội tụ nhanh. Hơn nữa, dựa vào các tham số sau khi tối ưu, chúng tôi cũng xác định được thứ tự xuất hiện của các đột biến theo thời gian trong đó đột biến K103N xuất hiện sớm nhất.

Từ khóa: *Mô hình Markov ẩn, thuật toán EM, cây đột biến di truyền, thuốc Efavirenz.*

Title: **Evolution of Drug Resistance of HIV Virus with Mutagenetic Tree according to Markov Model**

Abstract: In this paper, we predict the evolution of 14 mutations associated with the HIV resistance to Efavirenz using mutagenetic tree hidden Markov model. With new data of 396 patients from the HIV drug resistance database by Stanford University, we test statistical assumptions and found this data set significant for further modeling. Model results show that applying EM algorithm to estimate and optimize parameters has fast convergence. Furthermore, based on optimized parameters, we determine the occurrence order of mutations over time in which the K103N mutation appeared earliest.

Keywords: *Hidden markov model, EM algorithm, mutagenetic tree, efavirenz.*

I. MỞ ĐẦU

Luận thuyết trung tâm của Crick đã cho thấy một sơ đồ tổng hợp nên Protein bắt đầu bằng quá trình phiên mã từ DNA thành RNA [1]. Không lâu sau đó, các nhà khoa học đã khám phá ra một quá trình mà ban đầu tưởng như mâu thuẫn với giáo điều của luận thuyết trung tâm đó là quá trình phiên mã ngược (Retrotranscription) [2]. Đối với quá trình này, các RNA thông tin (mRNA) được dùng làm khuôn để tạo ra sợi đơn DNA bổ sung (cDNA) giống như bản khuôn đã phiên mã ra nó. Một số loại vi-rút có khả năng phiên mã ngược (Retroviruses) do sở hữu men phiên mã ngược (Reverse Transcriptase, viết tắt là RT). Chẳng hạn như vi-rút T-lymphotropic gây bệnh bạch cầu ở người (HTLV) [3], hay vi-rút gây suy giảm miễn dịch ở người (Human Immunodeficiency Virus, viết tắt là HIV) [4]. Hướng tiếp cận chính để điều trị bệnh gây ra bởi các Retroviruses là phát triển các loại thuốc gây ức chế men phiên mã ngược của chúng. Trong bài báo này, chúng tôi tập trung nghiên cứu đối với vi-rút HIV, một loại vi-rút nguy hiểm phá hủy tế bào lympho CD4+, làm giảm khả năng miễn dịch qua trung gian tế bào, tăng nguy cơ nhiễm trùng và ung thư ở người mang vi-rút. Hiện tại, có rất nhiều loại thuốc khác nhau được sử dụng kết hợp trong các phác đồ điều trị HIV. Trong đó có hai nhóm chính là: (i) nhóm ức chế phiên mã ngược nucleoside (Nucleoside Reverse Transcriptase Inhibitors, viết tắt là NRTIs); và (ii) nhóm ức chế phiên mã ngược không Nucleoside (Non-Nucleoside Reverse Transcriptase Inhibitors, viết tắt là NNRTIs)[1]. Mặc dù đã trải qua nhiều năm với vô vàn nỗ lực, nhưng cho đến nay loài người vẫn chưa tìm ra phương pháp đặc trị vi-rút HIV. Nguyên nhân gây khó khăn được xác định là trong quá trình điều trị, vi-rút xuất hiện những đột biến theo thời

[1] https://www.msdmanuals.com/professional/infectious-diseases/human-immunodeficiency-virus-hiv/drug-treatment-of-hiv-infection

gian có khả năng kháng lại các thuốc ức chế Protein.

Để quá trình điều trị HIV hiệu quả hơn, các nhà khoa học đã tiến hành các nghiên cứu nhằm dự đoán các thời điểm xảy ra các đột biến kháng thuốc trong các phác đồ sử dụng thuốc khác nhau trên các dữ liệu thu được từ quá trình điều trị. Nghiên cứu của Niko Beerenwinkel và các cộng sự vào năm 2007 đã chỉ ra quá trình tiến hóa của vi-rút HIV qua 7 đột biến kháng thuốc điển hình [5]. Phương pháp chính dựa vào cây đột biến di truyền và mô hình Markov ẩn. Ngoài ra, việc xây dựng cây đột biến cũng là một quá trình quan trọng để xác định quá trình tiến hóa của vi-rút. Dữ liệu cho phân tích khi đó gồm 163 bệnh nhân với 3350 nhân bản trình tự (Clones) được lấy từ ba nghiên cứu về liệu pháp kết hợp sử dụng thuốc Efavirenz (DMP 266-003, -004, -005) [6]. Tiếp tục những kết quả nghiên cứu của Niko Beerenwinkel, với mục tiêu thực nghiệm trên các bộ dữ liệu mới làm cơ sở cho phát triển phương pháp, trong nghiên cứu này, chúng tôi tiến hành dự đoán quá trình tiến hóa của vi-rút HIV qua 14 đột biến kháng thuốc với dữ liệu gồm 396 bệnh nhân được thu thập trên cơ sở dữ liệu kháng thuốc HIV của trường Đại học Standford[2]. Với dữ liệu từ năm 2005 đến nay, nhiều đột biến kháng thuốc mới xuất hiện cũng như nhiều loại thuốc mới ra đời. Để thực hiện mô hình, chúng tôi đã tiến hành kiểm định lại ba giả thiết của Niko Beerenwinkel và nhận thấy chúng vẫn đúng với bộ dữ liệu lớn hơn và mới hơn. Mặc dù, quá trình tiến hóa của vi-rút HIV vẫn được phân tích dựa trên việc sử dụng thuốc Efavirenz của nhóm NNRTIs, tuy nhiên với bộ dữ liệu mới này, cây đột biến được xây dựng lại với các nhánh lớn hơn, số các phụ thuộc giữa các nút đại diện cho các đột biến cũng nhiều hơn. Với giả thiết các đột biến xuất hiện theo thời gian tuân theo quá trình Poisson, chúng tôi cũng đã đưa ra dự đoán các thời điểm xuất hiện các đột biến kháng thuốc trong các phác đồ điều trị có sử dụng Efavirenz. Trong các phần tiếp theo của bài báo, chúng tôi sẽ trình bày về bộ dữ liệu HIV cho nghiên cứu trong mục II, về mô hình Markov trong mục III. Các nội dung tính toán thực nghiệm được trình bày trong mục IV, các kết quả được bàn luận trong mục V, cuối cùng mục VI là kết luận.

II. DỮ LIỆU

1. Dữ liệu HIV

Dữ liệu về tập hợp trình tự các nhân bản của HIV theo thời gian thu được từ 25 nghiên cứu lâm sàng với các phác đồ điều trị kết hợp trong đó có sử dụng thuốc Efavirenz. Các nghiên cứu này được thực hiện trong thời gian từ năm 1998 tới 2018 với 416 bệnh nhân, thông tin của mỗi bệnh nhân gồm các nhân bản trình tự Protein (Clones) tại các thời điểm khác nhau, ở mỗi Clones đã xác định vị trí của các đột biến. Bộ dữ liệu này được công bố trên cơ sở dữ liệu về kháng thuốc với HIV của đại học Stanford. Sau quá trình loại bỏ các quan sát không đủ thông tin, chúng tôi giữ lại thông tin của 396 bệnh nhân. Những mô tả chi tiết về mẫu, khuếch tán RNA, nhân bản hay giải trình tự có thể xem tại [7]. Tất cả các kết quả nghiên cứu đưa vào phân tích đều của các bệnh nhân được điều trị bằng Efavirenz.

Nghiên cứu của Bacheler và cộng sự vào năm 2001 đã xác định được dãy biến đổi của các axit amin trong quá trình phiên mã ngược của HIV có ảnh hưởng tới việc kháng thuốc Efavirenz. Họ đã chỉ ra được 2 đường chuyển hóa thay thế để kháng lại Efavirenz, một là chuyển hóa bắt đầu từ đột biến K103N (đường 103) và còn lại là đường chuyển hóa từ đột biến Y188L (đường 188). Chúng tôi thực hiện đánh giá cả 2 đường này và sử dụng để xây dựng cây đột biến. Trong số các đột biến kháng thuốc Efavirenz, để đảm bảo thời gian thực hiện tính toán, trước mắt chúng tôi giới hạn chỉ lấy các đột biến có tần suất xuất hiện lớn hơn 2%.

Sau quá trình tiền xử lí dữ liệu, kết quả thu được là dãy 14 đột biến với tần suất xuất hiện như sau: K103N (52,4%), L100I (9,1%), N348I (8,4%), Y181C (7,8%), G190A (7,7%), H221Y (6,0%), G190S (5,5%), P225H (4,6%), Y188L (4,6%), V108I (4,3%), A98G (3,8%), K101E (3,6%), V179D (3,1%), V106I (2,4%).

Ở đây chúng tôi giải thích thêm về mẫu đột biến có trong bộ dữ liệu HIV. Lấy ví dụ dữ liệu của bệnh nhân mã LB4-P142, tại thời điểm trước khi điều trị (thời điểm 1996-12), có ba Clone mang đột biến K103N. Đến mốc thời điểm tiếp theo (1997-01), đột biến V106I được phát hiện thêm trong quần thể vi-rút của bệnh nhân. Những dữ liệu này là căn cứ cho việc kiểm chứng các giả thiết theo đề xuất của Niko Beerenwinke, và sẽ trình bày ở những phần tiếp theo của bài báo.

2. Kiểm định thống kê

Để mô hình hóa quá trình tiến hóa tích lũy, Niko Beerenwinke đã đưa ra 3 giả thiết [5]:

(A1) Những đột biến thay thế không xảy ra độc lập. Con đường tiến hóa của vi-rút kháng thuốc được xem xét với các đột biến có tính duy trì.

(A2) Sự tồn tại của các đột biến trong quần thể là vĩnh viễn, tức là những biến đổi đã diễn ra sẽ luôn được duy trì, không thể đảo ngược hoặc mất đi.

(A3) Tại mỗi thời điểm, quần thể vi-rút bị chi phối bởi một chủng duy nhất và các Clones là độc lập với nhau.

Trong 3 giả thiết trên, có thể kiểm chứng bằng thống kê đối với giả thiết (A2) và (A3). Chúng tôi đã thực hiện kiểm chứng hai giả thiết này với bộ dữ liệu đã được tiền xử lí bằng phương pháp kiểm tra ngẫu nhiên hóa (Randomization Test) [8]. Cụ thể như sau:

[2]https://hivdb.stanford.edu/cgi-bin/RTIPairs.cgi

a) Giả thiết (A2) nêu lên tính chất không thể đảo ngược của các thay thế, tức khi đột biến đã xuất hiện trong Clone k của bệnh nhân i tại 1 thời điểm thì đột biến đó vẫn xuất hiện tại tất cả các Clones sau đó, chứ không mất đi. Giả thiết được kiểm tra bằng cách theo dõi sự thay đổi tần suất alen đột biến cho mỗi bệnh nhân theo thời gian.

Kí hiệu $[N] = \{1, 2, \ldots, N\}$ là tập các bệnh nhân, $[M] = \{1, 2, \ldots, M\}$ là tập các đột biến được xem xét, $[K_{ij}]$ là tập K_{ij} Clones quan sát được của bệnh nhân i $(i = 1, \ldots, N)$ tại thời điểm j $(j = 1, \ldots, J_i)$. Hàm chỉ $y_{ijkm} \in \{0, 1\}$ thể hiện sự hiện diện hay không của đột biến $m \in M$ trong Clone $k \in [K_{ij}]$ của bệnh nhân i tại thời điểm j. Khi đó, tần suất alen đột biến $m \in [M]$ cho mỗi bệnh nhân $i = 1, \ldots, N$ theo thời gian $j = 1, \ldots, J_i$ xác định bởi công thức

$$f_{ijm} = \frac{1}{K_{ij}} \sum_{k \in [K_{ij}]} y_{ijkm} \tag{1}$$

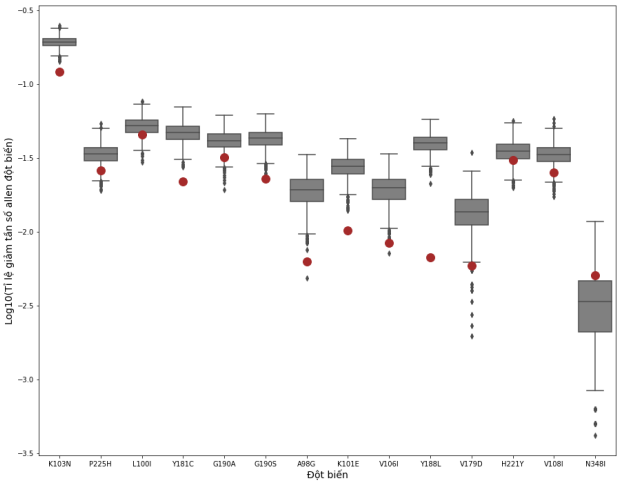
Với mỗi đột biến m, giá trị kiểm định thống kê A_m xác định mức độ tần suất alen của nó giảm từ một thời điểm này cho tới thời điểm khác. Với I là hàm chỉ, thì

$$A_m = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{J_i - 1} \sum_{j=1}^{J_i - 1} I\{f_{i,j,m} > f_{i,j+1,m}\} \tag{2}$$

Sử dụng phương pháp kiểm định ngẫu nhiên (Randomization Test) với giả thuyết không (Null Hypothesis) là xác suất của việc tăng và giảm tần số alen là bằng nhau. Phương pháp kiểm định ngẫu nhiên thực hiện bằng cách xáo trộn trình tự (Resampling) trong quần thể Clones. Tại mỗi lần Resampling dữ liệu, ta tính được một giá trị kiểm định thống kê. Với giả thiết sự thay thế không thể đảo ngược (A2), tỉ lệ % của việc giảm tần suất đột biến chiếm số lượng rất nhỏ trong quần thể. Để giả thiết này đúng, giá trị kiểm định thống kê của dữ liệu ban đầu cần phải nhỏ hơn các giá trị kiểm định thống kê sau khi Resampling dữ liệu.

Trong tính toán của chúng tôi, sử dụng Randomization Test với 1000 lần. Giá trị p_{value} được tính là tỉ lệ trên 1000 lần của số lần giá trị kiểm định thống kê với dữ liệu ngẫu nhiên mà nhỏ hơn giá trị kiểm định thống kê với dữ liệu ban đầu (Hình 1). Với hầu hết các đột biến, giá trị này nhỏ đáng kể ($p_{K101E} < 0,001$, $p_{Y181C} < 0,001$, $p_{K103N} < 0,001$, $p_{G190S} < 0,001$, $p_{Y188L} < 0,001$, $p_{A98G} = 0,001$, $p_{V106I} = 0,004$, $p_{V179D} = 0,017$, $p_{V108I} = 0,073$, $p_{G190A} = 0,078$, $p_{225H} = 0,084$), bên cạnh một số ngoại lệ ($p_{L100I} = 0,191$, $p_{H221Y} = 0,216$, $p_{N348I} = 0,841$). Chúng tôi đánh giá rằng giả thiết (A2) về sự thay thế không thể đảo ngược là hợp lý đối với phần lớn dữ liệu được sử dụng trong nghiên cứu này.

b) Giả thiết (A3) sẽ được kiểm định bằng cách xét sự đa dạng di truyền học của các Clones tại mỗi thời điểm $j \in [T_i]$ của bệnh nhân i, đa dạng càng nhỏ sẽ càng khẳng



Hình 1. Sự đột biến đã xuất hiện được duy trì theo thời gian. Với mỗi đột biến, chấm tròn biểu thị số lần giảm tần suất trong quần thể theo thời gian. Biểu đồ hộp biểu diễn phân phối của giá trị thống kê tạo bởi kiểm định ngẫu nhiên 1000 lần.

định được các Clones không sinh ra chủng loài mới mà bảo tồn gene từ một chủng chính.

Để đo sự đa dạng di truyền học giữa hai Clone c và c' có hay không đột biến m kí hiệu tương ứng là c_m và c'_m, khoảng cách Hamming được sử dụng theo công thức

$$D_H(c, c') = \sum_{m \in [M]} I\{c_m \neq c'_m\} \tag{3}$$

Độ đa dạng của bộ mẫu gồm $c_1, ..., c_K$ gồm K Clones là giá trị kỳ vọng của khoảng cách Hamming giữa 2 Clones bất kỳ trong mẫu đó,

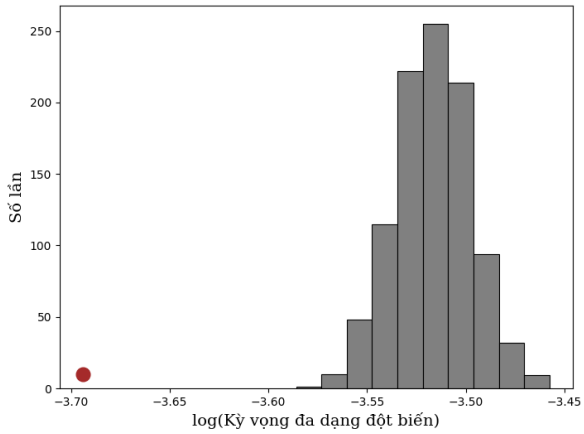
$$D_H(c_1, ..., c_K) = \frac{2}{K(K-1)} \sum_{k<k'} D_H(c_k, c_{k'}) \tag{4}$$

Đại lượng thống kê để đưa vào kiểm định là kỳ vọng độ đa dạng của toàn bộ mẫu, tính theo

$$B = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{J_i} \sum_{j=1}^{J_i} D_H(y_{ij1}, ..., y_{ijK_{ij}}) \tag{5}$$

Đặt giả thuyết bác bỏ là tại từng thời điểm, đều có thể quan sát được đa dạng di truyền học đầy đủ tính trên tất cả Clones của bệnh nhân. Giá trị của nó được ước lượng bằng cách tính kỳ vọng đa dạng di truyền học sau 1000 lần hoán đổi ngẫu nhiên các Clones vào thời điểm khác nhau. Một quan sát với B nhỏ hơn giá trị kỳ vọng này sẽ có mức đa dạng nhỏ hơn hơn đa dạng di truyền học đầy đủ và bác bỏ được giả thuyết trên nếu tỉ lệ nhỏ hơn là đáng kể ($p_{value} < 5\%$).

Kết quả tính toán cho thấy trên bộ dữ liệu hiện tại, đa dạng di truyền học tính theo từng thời điểm nhỏ hơn rất



Hình 2. Kết quả giá trị logarit đa dạng đột biến của dữ liệu mẫu (hình tròn) và phân phối tần suất sau khi thực hiện hoán đổi 1000 lần các Clones của một bệnh nhân vào thời điểm ngẫu nhiên (biểu đồ tần suất).

nhiều so với đa dạng của toàn bộ mẫu. Chúng tôi cũng quan sát được từ tính toán, các Clones tại mỗi thời điểm chỉ khác biệt bởi 1 trên 4900 đột biến được xét. Trong khi đó, kết quả trung bình sau khi ngẫu nhiên hóa là 1 trên 3300 đột biến với kết quả kiểm định p_{value} rất nhỏ. Như vậy, có thể kết luận rằng việc phát sinh chủng mới trong các Clones rất nhỏ và xảy ra giữa các thời điểm chứ không phải tại mỗi thời điểm. Từ đó khẳng định được rằng giả thiết (A3) là có ý nghĩa trên tập các nhân bản HIV được sử dụng trong nghiên cứu này.

III. MÔ HÌNH MARKOV

1. Cây đột biến

Với tập đột biến được xem xét $[M] = \{1, 2, \dots, M\}$, xét một cây có hướng T trên tập hợp đỉnh $V = \{0\} \cup [M]$ với gốc tại 0. Ứng với mỗi đỉnh $m \in V$ của T, xét một biến nhị phân ngẫu nhiên X_m trong đó mối tương quan của các biến ngẫu nhiên này phụ thuộc vào hình dạng của cây T và $\Pr(X_0 = 1) = 1$. Theo cách định nghĩa này, cây T cảm sinh một mô hình đồ thị có hướng cho phân bố đồng thời của véc-tơ ngẫu nhiên $X = (X_1, \dots, X_M)$. Cụ thể, phân bố của X cảm sinh từ mô hình này xác định bởi

$$\Pr(X = x) = \prod_{m \in [M]} \Pr\left(X_m = x_m \mid X_{\mathrm{pa}(m)} = x_{\mathrm{pa}(m)}\right) \quad (6)$$

trong đó $x = (x_1, \dots, x_M) \in \mathcal{I}$ với $\mathcal{I} = \{0, 1\}^M$ là không gian trạng thái của véc-tơ ngẫu nhiên X, $x_0 = 1$ và $\mathrm{pa}(m)$ là đỉnh cha của m trong cây T. Ví dụ

- Nếu T là đồ thị đường từ $0 \to 1 \to 2 \to \cdots \to M$, thì

$$\Pr(X = x) = \prod_{m=1}^{M} \Pr(X_m = x_m \mid X_{m-1} = x_{m-1}).$$

- Nếu T là một đồ thị sao với gốc tại 0 thì

$$\Pr(X = x) = \prod_{m \in [M]} \Pr(X_m = x_m \mid X_0 = 1).$$

Từ công thức (6), phân phối của X được xác định bởi các ma trận chuyển trạng thái có dạng

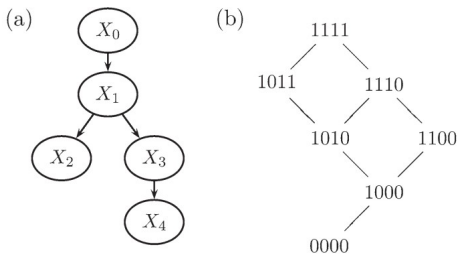
$$\vartheta^m = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 - \vartheta^m_{11} & \vartheta^m_{11} \end{pmatrix} \end{array}$$

trong đó $\vartheta^m_{ab} = \Pr(X_m = b | X_{\mathrm{pa}(m)} = a) \in [0, 1]$.

Phần còn lại của mục này mô tả việc áp dụng mô hình trên cho các sự kiện đột biến, gọi tắt là đột biến. Cụ thể, xét tập M đột biến $\{1, 2, \dots, M\}$, với mỗi đột biến $m \in [M]$, biến nhị phân ngẫu nhiên X_m với hai giá trị 1 và 0 lần lượt biểu thị sự xuất hiện và không xuất hiện của đột biến m. Một dãy đột biến $x \in \mathcal{I}$ được gọi là *tương thích* với cây T nếu trạng thái x có thể xuất hiện với xác suất khác 0 trong mô hình phân bố đồng thời cảm sinh bởi T. Do đó, x tương thích với T nếu tồn tại tham số $\vartheta = (\vartheta^1_{11}, \dots, \vartheta^M_{11}) \in [0, 1]^M$ sao cho

$$\Pr(X = x) = \prod_{m \in [M]} \vartheta^m_{\mathrm{pa}(m), x_m} > 0.$$

Kí hiệu $C(T)$ là tập hợp các trạng thái tương thích với cây T, tập hợp này tạo nên một mạng tinh thể $(C(T), \vee, \wedge)$ trong đó $\vee$ và $\wedge$ lần lượt là các toán tử max và min theo thành phần (xem [9, Lemma 14.3]). Biểu đồ Hasse trong hình vẽ dưới đây mô tả các con đường (dãy đột biến) được nhắc tới ở trên bằng việc tương ứng mỗi con đường tiến hóa với một đường đi từ trạng thái tự nhiên $(0, \dots, 0) \in C(T)$ đến trạng thái $(1, \dots, 1) \in C(T)$ khi tất cả các đột biến đều đã xuất hiện.



Hình 3. (a) Cây đột biến và (b) mạng tinh thể cảm sinh của các trạng thái tương thích.

Trong mô hình cây đột biến theo thời gian, các đột biến xuất hiện theo quá trình Poisson độc lập. Nếu λ_m là tỷ lệ

của quá trình này trên cạnh $\text{pa}(m) \to m$, xác suất để đột biến m xuất hiện trong khoảng thời gian Δt là

$$\vartheta_{11}^{m} = \Pr(X_m = 1 \mid X_{\text{pa}(m)} = 1) = 1 - e^{-\lambda_m \Delta t}.$$

Cây đột biến có thể được tái thiết lập từ dữ liệu bằng việc giải bài toán nhánh cực đại trong đồ thị đầy đủ trên tập đỉnh V bằng một tổ hợp thuật toán hiệu quả, xem thêm [9, 10] để biết chi tiết về thuật toán. Trong bài báo này, chúng tôi sử dụng thuật toán nêu trên để thiết lập hình dạng của T.

2. Mô hình Markov ẩn

Giả sử ta có thể quan sát được chuỗi đột biến vi-rút của một bệnh nhân nhiều hơn một thời điểm. Gọi X_{jm} là biến ngẫu nhiên chỉ sự xuất hiện của đột biến m tại thời điểm t_j với $j = 1, \ldots, J$ trong quần thể vi-rút của bệnh nhân. Ta giả thiết rằng quá trình tiến hóa bắt đầu tại thời điểm 0 ở trạng thái tự nhiên - không có đột biến nào. Do đó, $X_{1m} = 0$ với mọi $m \in [M]$ với $t_1 = 0$.

Sự phát triển của đột biến m tại thời điểm t_j với $j \geq 2$, mã hóa bởi biến ngẫu nhiên X_{jm}, lúc này phụ thuộc vào trạng thái thời điểm trước đó $X_{j-1,m}$, cũng như trạng thái hiện tại của đột biến cha $X_{j,\text{pa}(m)}$. Sự phụ thuộc này nảy sinh từ tính chất: sự hiện diện của đột biến m tại thời điểm t_j là kết quả của quá trình phát triển qua cạnh $\text{pa}(m) \to m$ tại thời điểm t_j hoặc từ tính không thể đảo ngược của nó trong quần thể vi-rút và do đó phụ thuộc vào sự hiện diện của nó tại thời điểm t_{j-1}. Cấu trúc phụ thuộc giữa các đột biến $(X_{jm} \mid j = 1, \ldots, J, m \in [M])$ có thể được biểu thị dưới một đồ thị có hướng không có chu trình (Xem Hình 4).

Ma trận chuyển của mô hình xích Markov này có dạng

$$\theta_j(\lambda_m) = \begin{array}{c} \\ 00 \\ 01 \\ 10 \\ 11 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1 & 0 \\ e^{-\lambda_m(t_j - t_{j-1})} & 1 - e^{-\lambda_m(t_j - t_{j-1})} \\ * & * \\ 0 & 1 \end{pmatrix} \end{array} \tag{7}$$

trong đó các hàng được đánh chỉ số bởi cặp $(m, \text{pa}(m)) \in \{0, 1\}^2$. Vị trí đánh dấu $*$ chỉ các vị trí không cần phải xem xét vì không có đột biến m nào xuất hiện trước đột biến mẹ $\text{pa}(m)$ của nó. Với ma trận này, ta định nghĩa mô hình cây Markov đột biến như là họ các phân bố đồng thời có dạng

$$\begin{aligned} &\Pr(X_{jm} = x_{jm}, j = 1, \ldots, J, m \in [M]) \\ &= \prod_{j=1}^{J} \prod_{m \in [M]} \theta_j(\lambda_m)_{(x_{j-1,m}, x_{j\text{pa}(m)}), x_{jm}}, \end{aligned}$$

trong đó $x_{j0} = 1$ và $t_0 = 0$. Điều này dẫn tới

$$\begin{aligned} &\Pr(X_{jm} = x_{jm} \mid X_{j-1,m} = x_{j-1,m}, X_{j\text{pa}(m)} = x_{j\text{pa}(m)}) \\ &\qquad = \theta_j(\lambda_m)_{(x_{j-1,m}, x_{j\text{pa}(m)}), x_{jm}}. \end{aligned}$$

3. Mô hình Markov ẩn cho cây đột biến

Chuyển qua trường hợp các mẫu Clones quan sát được tại các thời điểm khác nhau, mô hình hóa dữ liệu này bằng việc giả sử các Clones là các sao chép lỗi của một chuỗi đột biến ẩn mà chuỗi đột biến này tiến hóa dựa theo một cây đột biến. Như vậy, X_{jm} lúc này là một biến nhị phân ẩn. Dữ liệu quan sát được là ví dụ của các biến nhị phân ngẫu nhiên $Y_{jkm}, k \in [K_j]$ chỉ đột biến m có hiện diện trong đột Clone thứ k lấy mẫu từ quần thể vi-rút tại thời điểm t_j hay không. Các Clones là độc lập có điều kiện với điều kiện $(X_j), j = 1, \ldots, J$ cho trước. Mô hình đồ thị thu được xem như là một mô hình Markov ẩn cho cây đột biến (Mtree-HMM), một ví dụ xem ở Hình 4.

Sử dụng các kí hiệu $\varepsilon^+ = (\varepsilon_1^+, \ldots, \varepsilon_M^+) \in [0, 1]^M$ và $\varepsilon^- = (\varepsilon_1^-, \ldots, \varepsilon_M^-) \in [0, 1]^M$ là các véc-tơ tham số chứa xác suất quan sát phải dương tính giả và âm tính giả của từng đột biến. Tỷ lệ dương tính giả và âm tính giả cho biết sự khác biệt so với trạng thái quần thể có thể phát sinh từ các đột biến trong phản ứng PCR. Do đó, các tham số này định lượng kỳ vọng đa dạng di truyền của một quần thể vi-rút. Với điều kiện trạng thái X_{jm}, xác suất quan sát được đột biến m trong Clone k tại thời điểm t_j là

$$\theta'(\varepsilon_m^+, \varepsilon_m^-) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1 - \varepsilon_m^+ & \varepsilon_m^+ \\ \varepsilon_m^- & 1 - \varepsilon_m^- \end{pmatrix} \end{array}$$

Các yếu tố của ma trận này là xác suất có điều kiện

$$\theta'(\varepsilon_m^+, \varepsilon_m^-) = \Pr(Y_{jkm} = y_{jkm} \mid X_{jm} = x_{jm}).$$

Do đó, các Clones khác nhau $Y_{jk} = (Y_{jk1}, \ldots, Y_{jkM})$, với $k \in [K_j]$ được mô hình thành các biến ngẫu nhiên độc lập cùng phân bố. Đặt

$$Y = (Y_{jkm} \mid j = 1, \ldots, J, k \in [K_j], m \in [M])$$

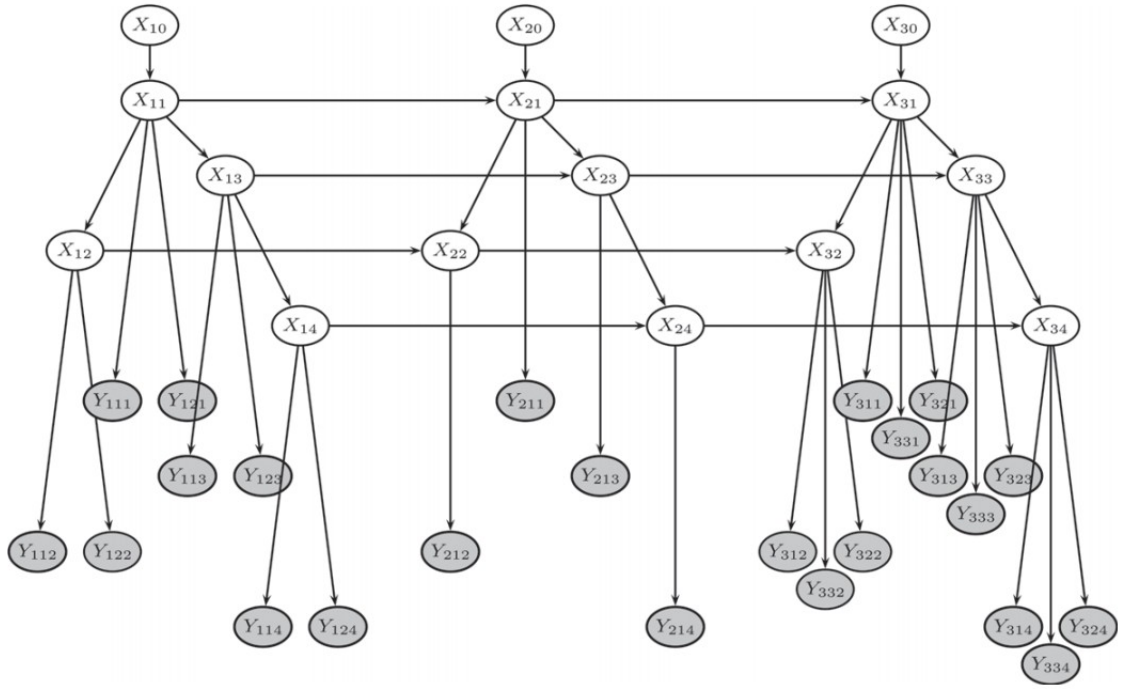
là vec tơ biểu thị tất cả dãy quan sát Clones. Mô hình Markov ẩn cho cây đột biến là họ các phân bố đồng thời của Y được cho bởi

$$\begin{aligned} \Pr(Y = y) = &\sum_{x_1 \in C(T)} \cdots \sum_{x_J \in C(T)} \prod_{m \in [M]} \prod_{j=1}^{J} \\ &\left(\theta_j(\lambda_m)_{(x_{j-1,m}, x_{j\text{pa}(m)}), x_{jm}} \prod_{k \in [K_j]} \theta'(\varepsilon_m^+, \vartheta_m^-)_{x_{jm}, y_{jkm}} \right), \end{aligned}$$

trong đó phép lấy tổng được thực hiện với tất cả các trạng thái ẩn của mô hình. Cấu trúc đồ thị của mô hình Markov ẩn cho cây đột biến được biểu diễn ở Hình 4.

4. Tính toán tham số

Với mỗi bệnh nhân $i \in [N] = 1, 2, \ldots, N$, ta có các quan sát tại các thời điểm $t_{i1}, t_{i2}, \ldots, t_{iJ_i}$. Gọi X_{ijm} là biến ngẫu



Hình 4. Đồ thị không có chu trình mô tả mô hình Markov ẩn cho cây đột biến theo thời gian với 3 thời điểm. Các đỉnh trắng ứng với biến ngẫu nhiên ẩn và các đỉnh xám là các biến quan sát được [5].

nhiên chỉ sự xuất hiện của đột biến m trong quần thể vi-rút của bệnh nhân i tại thời điểm t_{ij}. Biến ngẫu nhiên Y_{ijkm} chỉ sự xuất hiện của đột biến m trong Clone $k \in [K_{ij}]$ của bệnh nhân i tại thời điểm t_{ij}. Kí hiệu ma trận chuyển (7) tương ứng với bệnh nhân i bởi $\theta_{ij}(\lambda_m)$, chẳng hạn

$$\theta_{ij}(\lambda_m)_{01,0} = e^{-\lambda_m(t_{ij}-t_{i,j-1})}.$$

Giả sử các bệnh nhân độc lập với nhau và dữ liệu của mỗi bệnh nhân dựa theo một mô hình cây Markov ẩn trên một cây T cố định. Khi đó, kết quả của mô hình cho quan sát

$$Y = \left(Y_{ijkm} \,|\, i \in [N], j = 1, \ldots, J_i, k \in [K_{ij}], m \in [M]\right)$$

thực hiện tại các thời điểm $\{t_{ij} \,|\, i \in [N], j = 1, \ldots, J_i\}$ là các tham số $\lambda = (\lambda_1, \ldots, \lambda_n)$, $\varepsilon^+ = \left(\varepsilon_1^+, \ldots, \varepsilon_m^+\right)$, và $\varepsilon^- = \left(\varepsilon_1^-, \ldots, \varepsilon_m^-\right)$ làm cực đại hàm Likelihood

$$L_{\text{obs}}(\lambda, \varepsilon^+, \varepsilon^-) = \sum_{x_{11} \in C(T)} \cdots \sum_{x_{NJ_N} \in C(T)} \prod_{i \in [N]} \prod_{m \in [M]} \prod_{j=1}^{J_i} \left(\theta_{ij}(\lambda_m)_{x_{i,j-1,m}, x_{ij\text{pa}(m))}, x_{ijm}} \prod_{k \in [K_{ij}]} \theta'(\varepsilon^+, \varepsilon^-)_{x_{ijm}, y_{ijkm}} \right)$$

trong đó $x_{i0m} = 1$ và $t_{i0} = 0$ với mọi $i \in [N], m \in [M]$. Vì việc tính toán trên hàm Likelihood khó khăn hơn, nên hàm Log-Likelihood sẽ được sử dụng thay thế.

Gọi $\{x_{ijm}\}$ là giá trị của các biến ẩn $\{X_{ijm}\}$ tương thích với mô hình cây đột biến ẩn. Tính chất sau được suy ra trực tiếp từ mô hình cây Markov ẩn.

$$x_{ijm} = \begin{cases} 1 & \text{nếu } x_{i,j-1,m} = 1, \\ 0 & \text{nếu } x_{ij\text{pa}(m)} = 0. \end{cases}$$

Gọi I là hàm đặc trưng (hàm chỉ) nhận giá trị 0 và 1, đặt

$$\chi_{ijm}(a) = I\{x_{i,j-1,m} = 0, x_{ij\text{pa}(m)} = 1, x_{ijm} = a\},$$
$$\chi'_{ijkm}(a, b) = I\{x_{ijm} = a, y_{ijkm} = b\},$$

với $a, b = 0, 1$. Khi đó, hàm Log-Likelihood $\ell_{\text{hid}}(\lambda, \varepsilon^+, \varepsilon^-)$ của mô hình ẩn xác định bởi

$$\begin{aligned} \ell_{\text{hid}}(\lambda, \varepsilon^+, \varepsilon^-) = \sum_{i \in [N]} \sum_{m \in [M]} \sum_{j=1}^{J_i} \Big\{ & -\chi_{ijm}(0)\lambda_m \left(t_{ij} - t_{i,j-1}\right) \\ & + \chi_{ijm}(1) \log\left(1 - e^{\lambda_m(t_{ij}-t_{i,j-1})}\right) \\ & + \sum_{k \in [K_{ij}]} \Big[\chi'_{ijkm}(0,0) \log\left(1 - \varepsilon_m^+\right) + \chi'_{ijkm}(0,1) \log \varepsilon_m^+ \\ & + \chi'_{ijkm}(1,0) \log \varepsilon_m^- + \chi'_{ijkm}(1,1) \log\left(1 - \varepsilon_m^-\right) \Big] \Big\}. \end{aligned}$$

Để giải bài toán tối ưu này và tìm được ước lượng hợp lý cực đại, chúng tôi sử dụng thuật toán EM hay Forward-Backward của mô hình Markov ẩn. Chi tiết của thuật toán này có thể tham khảo ở nhiều tài liệu, chẳng hạn xem [11].

IV. TÍNH TOÁN THỰC NGHIỆM

Nhắc lại, chúng tôi thực hiện tính toán thực nghiệm trên bộ dữ liệu đã làm sạch và kiểm tra các giả thiết bao gồm 14 đột biến đối với vi-rút HIV. Theo cách dựng cây đột biến đã trình bày trong mục III.A với, nút gốc là K103N và Y188C được giả thiết là xuất hiện ngay từ đầu (chỉ có trạng thái 1) trong các tập đột biến có thể có (compatible set). Mặt khác, các nút này không có nút cha mẹ, do đó việc tính toán tham số λ_m, dương tính giả (ε_m^+), và âm tính giả (ε_m^-) được thực hiện dựa trên thống kê từ bộ dữ liệu ban đầu. Tham số của các nút còn lại được tối ưu nhờ quá trình học của thuật toán EM gồm hai bước Expectation (E-step) và Maximization (M-step) [12]. Trong bước E-step ta cần tính được xác suất chuyển trạng thái $u_{ijm,a}$ thông qua hàm **GST**, xác suất xuất hiện quan sát $u_{m,ab}$ thông qua hàm **GSE**, trong đó a, b nhận các giá trị 0, 1̃.

$$\textbf{GST}: \quad u_{ijm,a} = \textbf{Pr}(x_{i,j-1,m} = 0, x_{ijpa(m)} = 1, x_{ijm} = a|Y)$$

$$\textbf{GSE}: \quad u_{m,ab} = \sum_{i\in[N]} \sum_{j=1}^{J_i} \sum_{k\in[K_{ij}]} \textbf{Pr}(x_{ijm} = a, y_{ijkm} = b|Y)$$

Bước M-step được thực hiện nhờ vòng lặp để khép sai số các tham số λ_m, ε_m^+, ε_m^-, thuật toán dừng lại khi sai số giữa hai bước của λ_m nhỏ hơn 5.10^{-4} (Xem Thuật toán 1).

Để đánh giá hiệu quả của mô hình, chúng tôi dựa trên một số tiêu chí gồm độ chính xác, độ hội tụ, và thời gian tính toán. Độ chính xác của mô hình được đánh giá bằng xác suất dương tính giả, xác suất âm tính giả giữa các đột biến dự đoán và các đột biến quan sát trong dữ liệu. Độ hội tụ của mô hình được xem xét thông qua số vòng lặp của thuật toán tương ứng với mỗi đột biến. Cuối cùng, thời gian tính toán được đo cho mỗi đột biến khi thực nghiệm trên hệ thống với cấu hình máy tính core i7, 4 CPUs, 8GB RAM, ngôn ngữ Python.

Thuật toán 1:
Hidden Markov Model Maximization
HMMM($cSet$, $mSet$, $rTree$)

Dữ liệu vào:
  $cSet$: Compatible Set,
  $mSet$: Mutation Set,
  $rTree$: Reconstruction Tree.
Dữ liệu ra:
  pr: Parameters.
Khởi tạo:
  $\lambda_m = [0.1] * |mSet|$
  $\varepsilon_m^+ = 0.01$ *//Dương tính giả*
  $\varepsilon_m^- = 0.01$ *//Âm tính giả*
Chuẩn bị trước:
  GEP: Function to Get Emission Probability
  GST: Function to Get Sum of Transition Probability

begin
  for $m = mSet[1]$ to $mSet[-1]$ **do**
    $pr = empty$;
    while $error\ of\ \lambda_m \geq 5e-04$ **do**
      $u_{m,00} = \textbf{GEP}(cSet, rTree, \lambda_m, 0, 0, \varepsilon_m^+, \varepsilon_m^-)$;
      $u_{m,01} = \textbf{GEP}(cSet, rTree, \lambda_m, 0, 1, \varepsilon_m^+, \varepsilon_m^-)$;
      $u_{m,11} = \textbf{GEP}(cSet, rTree, \lambda_m, 1, 1, \varepsilon_m^+, \varepsilon_m^-)$;
      $u_{m,10} = \textbf{GEP}(cSet, rTree, \lambda_m, 1, 1, \varepsilon_m^+, \varepsilon_m^-)$;
      $\varepsilon_m^- = \frac{u_{m,01}}{u_{m,00}+u_{m,01}}$;
      $\varepsilon_m^+ = \frac{u_{m,10}}{u_{m,10}+u_{m,11}}$;
      $u_{ijm,0} = \textbf{GST}(cSet, rTree, \lambda_m, 0, \varepsilon_m^+, \varepsilon_m^-)$;
      $u_{ijm,1} = \textbf{GST}(cSet, rTree, \lambda_m, 1, \varepsilon_m^+, \varepsilon_m^-)$;
      $\lambda_m = \sum_{i\in[N]} \sum_{j=1}^{J_i} \frac{u_{ijm,1}}{\Delta_{i,j}(u_{ijm,0}+u_{ijm,1})}$
    end
    append $(\lambda_m, \varepsilon_m^+, \varepsilon_m^-)$ to pr
  end
  return pr
end

V. KẾT QUẢ

1. Cây đột biến

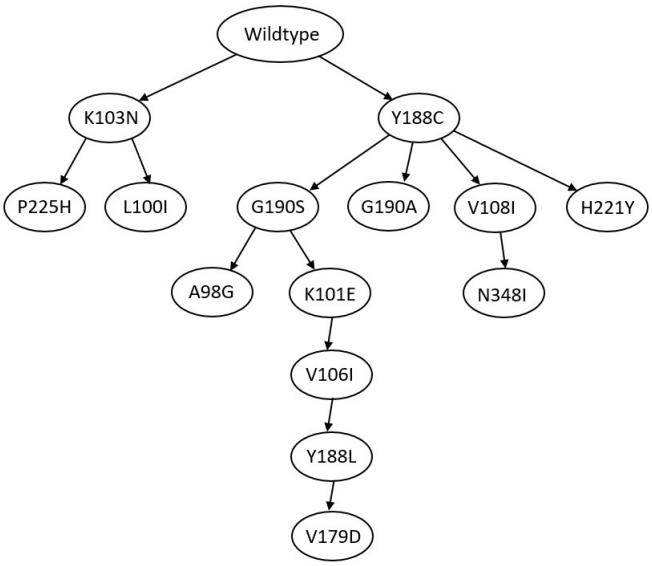
Sử dụng thuật toán tái xây dựng cây đột biến từ dữ liệu cắt ngang chúng tôi thu được cây đột biến với 15 đỉnh (Xem Hình 5). Nhánh một gồm 3 đột biến là K103N, L100I, P225H; nhánh hai gồm 11 đột biến là Y188C, G190S, G190A, V108I, H221Y, A98G, K101E, N348I, V106I, Y188L, và V179D.

2. Quá trình tiến hóa của vi-rút HIV

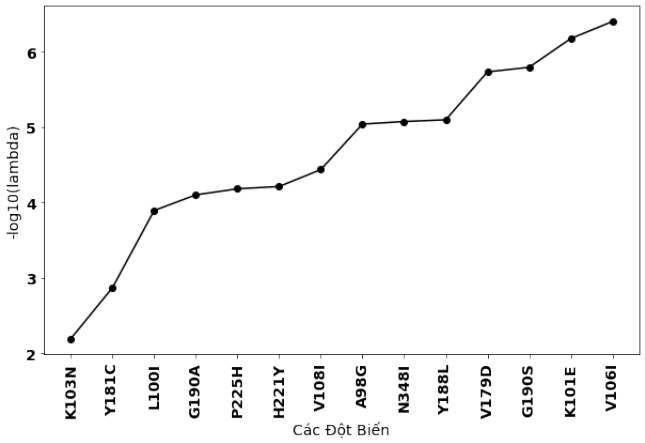
Sau khi tối ưu tham số với hai cây có các nút gốc là K103N và Y188C ta thu được các giá trị λ_m tương ứng với từng đột biến. Vì giá trị λ_m mang ý nghĩa là số lần xuất hiện đột biến trong khoảng thời gian một tuần nên giá trị này càng lớn thì đột biến xuất hiện càng sớm. Kết quả cho thấy đột biến K103N xuất hiện sớm nhất ở tuần $\lambda_{K103N}^{-1} = 157$, tiếp theo là các đột biến Y181C, L100I, G190A, P225H, H221Y, V108I, A98G, N348I, Y188L, V179D, G190S, K101E, V106I (Xem Hình 6).

3. Hiệu quả của mô hình

Dương tính giả (ε_m^+) xuất hiện khi một đột biến được dự đoán là tồn tại trong trong quần thể virus, nhưng thực tế lại không tìm thấy ở những nhân bản trình tự Protein (Clones). Ngược lại, âm tính giả (ε_m^-) xuất hiện khi một đột biến được dự đoán là không tồn tại trong quần thể nhưng lại tìm thấy trong các Clones quan sát. Theo kết quả tái xây dựng cây đột biến (xem phần V.1) ta thu được hai nút gốc là K103N và Y188C cuả hai nhánh. Với hai đột biến đóng vai trò hai nút gốc này, các tham số được tính dựa trên thống kê dữ liệu ban đầu nên dương tính giả khá cao, xấp xỉ 0,5. Trong các nghiên cứu tiếp theo, chúng tôi có thể cải thiện kết quả này bằng cách sử dụng phương pháp mặt cắt ngang để ước lượng các tham số khởi tạo cho hai nút gốc. Các



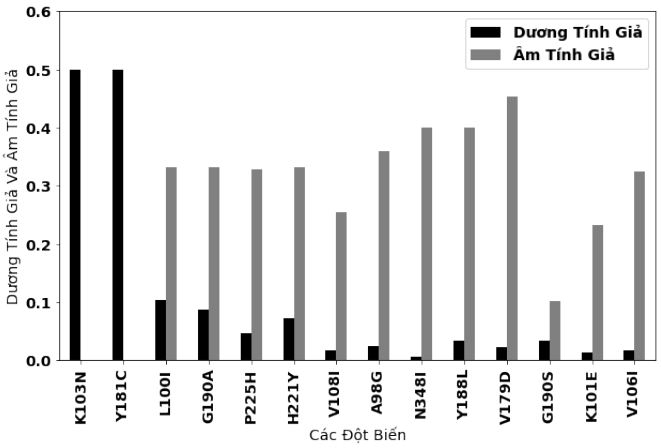
Hình 5. Mô hình cây đột biến của sự xuất hiện kháng thuốc Efavirenz trong HIV. Các đỉnh được đặt tên theo các thay thế axit amin.



Hình 6. Quá trình tiến hóa của vi-rút HIV qua 14 đột biến kháng thuốc. Trục hoành biểu diễn tên của các đột biến. Trục tung biểu diễn $-\log_{10}(\lambda)$. Chỉ số trên trục tung càng lớn thì đột biến xuất hiện càng muộn trong quá trình điều trị.

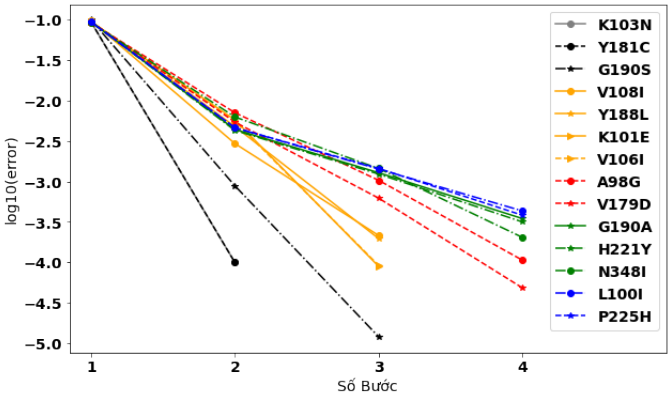
đột biến còn lại của hai nhánh từ L100I đến V106I có xác suất dương tính giả khá nhỏ (từ 0,5% tới 10,4%) so với xác suất âm tính giả (từ 10,2% đến 45,3%) (Xem Hình 7). Dương tính giả nhỏ chứng tỏ giả thiết A2 về sự không thể đảo ngược của đột biến có độ chính xác cao.

Trong bài báo này, phương pháp ước lượng và tối ưu tham số EM vào mô hình Markov ẩn của cây đột biến gene cho tốc độ hội tụ khá cao. Ngoài hai nút gốc là K103N và Y188C được tính toán không dựa vào vòng lặp thì các tham số λ_m của các đột biến còn lại hội tụ chỉ từ 3 đến 4 bước với ngưỡng sai số là 5.10^{-4} (Xem Hình 8). Với hai cây con được tạo ra từ việc xây dựng lại cây đột biến, nhánh một



Hình 7. Tỷ lệ dương tính giả và âm tính giả của dự đoán ứng với các đột biến. Hai đột biến đóng vai trò gốc ở hai nhánh của cây đột biến là K103N và Y181C có tỷ lệ dương tính giả lớn nhất, xấp xỉ 0.5. Ngược lại, các đột biến khác đều có tỷ lệ dương tính giả rất nhỏ so với tỷ lệ âm tính giả.

gồm 3 nút với thời gian tính toán chỉ từ 1,15 đến 2,89 giây với từng đột biến, nhánh hai gồm 11 nút có thời gian tính toán từ 168,1 đến 551,82 giây cho mỗi đột biến (xem bảng I).



Hình 8. Độ hội tụ của thuật toán khi thực hiện vòng lặp cho các đột biến khác nhau với ngưỡng sai số giữa hai bước nhỏ hơn 5.10^{-4}. Trục hoành biểu diễn số bước lặp, trục tung biểu diễn $\log_{10}(error)$.

VI. KẾT LUẬN

Với mô hình Markov ẩn và cây đột biến di truyền áp dụng cho dữ liệu kháng thuốc HIV được cập nhật mới nhất, chúng tôi đã tìm được quá trình tiến hóa của vi-rút HIV qua 14 đột biến kháng thuốc Efavirenz. Thời điểm xuất hiện đột biến được dự đoán thông qua tham số của quá trình poisson. Kết quả cho thấy đột biến K103N xuất hiện sớm nhất và V106I xuất hiện muộn nhất. Mặc dù vậy, do sự kết hợp của thuốc ngày càng phức tạp, nhiều thuốc

Bảng I
THỜI GIAN TÍNH TOÁN CHO MỖI ĐỘT BIẾN VỚI SỐ VÒNG LẶP VÀ SAI SỐ TƯƠNG ỨNG. NHÁNH 1 CHỈ GỒM BA ĐỘT BIẾN K103N, L100I, VÀ P225H NÊN THỜI GIAN TÍNH TOÁN NHANH HƠN CÁC ĐỘT BIẾN CỦA NHÁNH 2.

Đột biến	λ	Số bước	Sai số	Thời gian (s)
K103N	0,0063689	2	0	2,891
Y181C	0,0013390	2	0	168,1
L100I	0,0001269	4	0,00043	1,167
G190A	0,0000788	4	0,00035	551,783
P225H	0,0000654	4	0,00038	1,151
H221Y	0,0000610	4	0,00032	551,818
V108I	0,0000364	3	0,00022	360,737
A98G	0,0000091	4	0,00011	550,339
N348I	0,0000084	4	0,0002	528,277
Y188L	0,0000080	3	0,00019	363,145
V179D	0,0000018	4	0,00005	463,144
G190S	0,0000016	3	0,00001	278,808
K101E	0,0000007	3	0,00009	330,593
V106I	0,0000004	3	0,00009	358,929

được sử dụng trước khi sử dụng Efavirenz trong các phác đồ điều trị nên thời điểm xuất hiện của các đột biến được dự đoán là khá muộn. Để kết quả đạt độ chính xác cao hơn chúng tôi cần tiếp tục tinh sạch dữ liệu làm cho dữ liệu phù hợp nhất với mô hình tính toán.

Thời gian tính toán vẫn đảm bảo không quá lớn với cây đột biến di truyền được xây dựng lại gồm nhiều nút và số lượng ảnh hưởng giữa các nút tăng lên đáng kể. Tuy nhiên, các tham số được tính toán với các nút gốc K103N và Y181C dựa trên thống kê dữ liệu là lý do dẫn đến các đột biến này hầu như chưa tham gia vào quá trình tối ưu tham số. Do đó, chúng cần được tính toán kĩ lưỡng hơn dựa trên việc xây dựng lại cây đột biến di truyền.

Quá trình ước lượng và tối ưu tham số nhờ thuật toán EM tỏ ra khá hiệu quả với mô hình và dữ liệu trong bài báo. Cụ thể, với sai khác các giá trị tham số giữa hai vòng lặp liên tiếp là 5.10^{-4}, thuật toán hội tụ chỉ từ 3 đến 4 bước với mỗi đột biến. Đồng thời, xác suất dương tính giả của các đột biến trong quá trình tiến hóa cũng khá thấp, điều này có ý nghĩa lớn trong lâm sàng khi ta không cần lựa chọn thuốc thay thế cho các đột biến giả trong quá trình điều trị.

Cho rằng các nghiên cứu liên quan tới quá trình tiến hóa của các chủng vi-rút, đặc biệt với các vi-rút chứa men phiên mã ngược là rất có ý nghĩa vì sự cấp thiết trong điều trị một khi vi rút tấn công tế bào cơ thể người. Do đó, chúng tôi sẽ tiếp tục phát triển nghiên cứu này với những bộ dữ liệu mới hơn, lớn hơn. Bên cạnh mục tiêu hoàn thiện mô hình cây đột biến, nghiên cứu sự tiến hóa của vi-rút là việc áp dụng phương pháp với các loại vi-rút mới, biến đổi nhanh.

LỜI CẢM ƠN

Nghiên cứu được hoàn thành bởi nhóm nghiên cứu gồm các học viên và giảng viên chương trình đào tạo Thạc sĩ Khoa học Dữ liệu của Khoa Toán-Cơ-Tin học, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội với hai tác giả chính là Nguyễn Văn Thế và Tạ Văn Nhân. Nghiên cứu được hỗ trợ bởi VinIF trong khuôn khổ chương trình hợp tác đào tạo thạc sĩ Khoa học dữ liệu với khoa Toán-Cơ-Tin học, trường Đại học Khoa học Tự nhiên, Đại học quốc gia Hà Nội. Nguyễn Văn Thế được tài trợ bởi Tập đoàn Vingroup – Công ty CP và hỗ trợ bởi chương trình học bổng đào tạo thạc sĩ, tiến sĩ trong nước của Quỹ Đổi mới sáng tạo Vingroup (VINIF), Viện Nghiên cứu Dữ liệu lớn (VinBigdata), mã số VINIF.2020.ThS.KHTN.05.

TÀI LIỆU THAM KHẢO

[1] F. Crick, "Central Dogma of Molecular Biology," *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970.

[2] H. M. Temin and D. Baltimore, "RNA-Directed DNA Synthesis and RNA Tumor Viruses," in *Advances in Virus Research*, K. M. Smith, M. A. Lauffer, and F. B. Bang, Eds. Academic Press, Jan. 1972, vol. 17, pp. 129–186.

[3] R. C. Gallo, "Summary of Recent Observations on the Molecular Biology of RNA Tumor Viruses and Attempts at Application to Human Leukemia," *American Journal of Clinical Pathology*, vol. 60, no. 1, pp. 80–87, Jul. 1973.

[4] B. D. Preston, B. J. Poiesz, and L. A. Loeb, "Fidelity of HIV-1 reverse transcriptase," *Science (New York, N.Y.)*, vol. 242, no. 4882, pp. 1168–1171, Nov. 1988.

[5] N. Beerenwinkel and M. Drton, "A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data," *Biostatistics (Oxford, England)*, vol. 8, no. 1, pp. 53–71, Jan. 2007.

[6] S.-Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer, "Human immunodeficiency virus reverse transcriptase and protease sequence database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 298–303, Jan. 2003.

[7] L. Bacheler, S. Jeffrey, G. Hanna, R. D'Aquila, L. Wallace, K. Logue, B. Cordova, K. Hertogs, B. Larder, R. Buckery, D. Baker, K. Gallagher, H. Scarnati, R. Tritch, and C. Rizzo, "Genotypic correlates of phenotypic resistance to efavirenz in virus isolates from patients failing nonnucleoside reverse transcriptase inhibitor therapy," *Journal of Virology*, vol. 75, no. 11, pp. 4999–5008, Jun. 2001.

[8] C. E. Lunneborg, "Random assignment of available cases: Bootstrap standard errors and confidence intervals," *Psychological Methods*, vol. 6, no. 4, pp. 402–412, 2001.

[9] N. Beerenwinkel and M. Drton, *Mutagenetic Tree Models*. Cambridge University Press, 2005.

[10] R. Desper, F. Jiang, O.-P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer, "Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data," *Journal of Computational Biology*, vol. 6, no. 1, pp. 37–51, Jan. 1999, publisher: Mary Ann Liebert, Inc., publishers.

[11] O. Cappé, "Online EM Algorithm for Hidden Markov Models," *Journal of Computational and Graphical Statistics*, vol. 20, no. 3, pp. 728–749, 2011.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

SƠ LƯỢC VỀ TÁC GIẢ

Nguyễn Văn Thế

Nhận bằng cử nhân khoa học tài năng ngành Toán học năm 2020.
Hiện là học viên chương trình đào tạo thạc sĩ Khoa học dữ liệu, trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.
Các lĩnh vực nghiên cứu: Tổ hợp, lý thuyết đồ thị, Giải tích ma trận, thuật toán, tin sinh học.
Email: nguyenvanthe@hus.edu.vn

Tạ Văn Nhân

Nhận bằng thạc sĩ ngành Khoa học dữ liệu năm 2021, trường Đại học Khoa học tự nhiên, Đại học Quốc gia Hà Nội.
Hiện là chuyên viên tin sinh học tại công ty LOBI Việt Nam.
Lĩnh vực nghiên cứu: Giải trình tự hệ gen người, dược học hệ gen, mô hình Markov ẩn và mạng Bayes trong tin sinh học.
Email: tavannhan@gmail.com

Nguyễn Thị Kim Duyên

Nhận bằng cử nhân khoa học ngành Toán-tin năm 2010.
Hiện là học viên chương trình đào tạo thạc sĩ Khoa học dữ liệu, trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.
Lĩnh vực nghiên cứu: Điểm nguy cơ đa di truyền, áp dụng mô hình học máy trong tin sinh học.
Email: nguyenthikimduyen_ch2020@hus.edu.vn

Trịnh Mai Phương

Hiện là học viên chương trình thạc sĩ Khoa học dữ liệu, trường Đại học Khoa học tự nhiên, Đại học Quốc gia Hà Nội.
Lĩnh vực nghiên cứu: Dự báo trong kinh tế và kinh doanh, khai phá dữ liệu lớn, học máy có giám sát trong tin sinh học.
Email: trinhmaiphuong_ch2020@hus.edu.vn

Nguyễn Thị Hồng Minh

Nhận học vị tiến sĩ Toán - Tin năm 2001, học hàm Phó giáo sư ngành Công nghệ thông tin năm 2018.
Hiện là giảng viên cao cấp, trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội.
Lĩnh vực nghiên cứu: Độ phức tạp thuật toán, tính toán song song, khai phá dữ liệu, tính toán mềm, tin sinh học.
Email: minhnth@hus.edu.vn