# DỰ ĐOÁN EPITOPE TẾ BÀO B KHÔNG LIÊN TỤC TRÊN PROTEIN MATRIX CỦA VIRUS H5N1

**Trần Ngọc Vinh, Võ Cẩm Quy, Trần Linh Thước**
Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM
*(Bài nhận ngày 08 tháng 01 năm 2009, hoàn chỉnh sửa chữa ngày 06 tháng 08 năm 2009)*

***TÓM TẮT:*** *Mặc dù epitope không liên tục chiếm đến 90% trong tổng số các epitope tế bào B nhưng do các khó khăn trong việc phát triển các phương pháp dự đoán epitope không liên tục nên hiện nay các phương pháp dự đoán epitope tế bào B đa số đều tập trung vào việc dự đoán epitope liên tục. Để phục vụ cho việc phát triển văcxin phòng virus H5N1, chúng tôi đang tiến hành nghiên cứu dự đoán in silico các epitope liên tục liên quan đến tế bào T và tế bào B cũng như epitope không liên tục liên quan đến tế bào B của các kháng nguyên virus H5N1. Trong nghiên cứu này, bằng phương pháp mô hình hóa tương đồng, chúng tôi đã tạo được các cấu trúc của protein matrix của virus H5N1 từ đó dự đoán epitopes tế bào B không liên tục. 60 trong 72 amino acid đã dự đoán được là trùng hợp với kết quả dự đoán bởi phương pháp dự đoán epitope cấu hình của CEP (Conformational Epitope Prediction). Tất cả các amino acid dự đoán đều thuộc nhóm thích nước, phân cực, tích điện và nằm trên bề mặt cấu trúc của kháng nguyên.*

***Từ khóa:*** *epitope tế bào B không liên tục, protein matrix, H5N1, dự đoán epitope*

## 1. GIỚI THIỆU

Epitope tế bào B không liên tục là một đối tượng khó thực hiện trong việc dự đoán *in silico* do tính chất liên quan đến cấu trúc cũng như lượng dữ liệu thực nghiệm còn ít. Tuy vậy, nó có tầm quan trọng lớn đối với đáp ứng miễn dịch tế bào B khi có đến 90% epitope tế bào B thuộc dạng không liên tục. Chúng tôi đã xây dựng quy trình dự đoán epitope không liên tục dựa trên phương pháp mô hình hóa tương đồng và ứng dụng bước đầu vào các trình tự protein matrix của virus H5N1 ở Việt Nam.

## 2. PHƯƠNG PHÁP

Chúng tôi thực hiện việc xây dựng quy trình dự đoán epitope như miêu tả trong **Hình** với 4 bước chính: (1) Thu nhận và xử lý dữ liệu, (2) Xây dựng và đánh giá cấu trúc, (3) Ứng dụng, (4) Dự đoán epitope.

***Bước 1: Thu nhận và xử lý dữ liệu***

Từ CSDL CED và IEDB: thu dữ liệu về cấu trúc epitope tế bào B không liên tục.Từ CSDL NCBI, UniProKB, CED và IEDB: thu tập trình tự mục tiêu dùng xây dựng mô hình tạo cấu trúc. Đồng thời thu trình tự protein MP dùng để dự đoán epitope. Dữ liệu sau khi thu nhận từ các CSDL trên được xử lý bằng các đoạn chương trình Perl để lấy ra các thông tin cần thiết cũng như định dạng dữ liệu phù hợp với các chương trình sử dụng.

***Bước 2: Xây dựng và đánh giá cấu trúc***

Phương pháp mô hình hóa tương đồng được sử dụng để tạo ra cấu trúc cho các trình tự protein. Quy trình gồm 4 bước như sau:

1) Tìm trình tự mẫu trong CSDL tương đồng với trình tự mục tiêu bằng câu lệnh blastall của chương trình BLAST.

2) Thực hiện sắp gióng cột giữa trình tự mục tiêu và các trình tự mẫu bằng các chương trình sắp gióng cột Clustal, T-Coffee và một công cụ sắp gióng cột có sẵn của Modeller.

3) Sử dụng chương trình Modeller để tạo cấu trúc cho trình tự mục tiêu từ kết quả sắp gióng cột đa trình tự ở bước 2.

4) Kiểm tra, đánh giá độ chính xác của cấu trúc đã tạo ra bằng chương trình Prosa và Pymol. Các chỉ tiêu đánh giá bao gồm: chênh lệch giá trị z-score thực nghiệm và lý thuyết của cấu trúc không quá 3,5; giá trị năng lượng rk-comb của cấu trúc dưới 10; giá trị RMSD giữa cấu trúc thực nghiệm và cấu trúc được tạo ra bằng phương pháp mô hình hóa không quá 1,5Å.

Chúng tôi xây dựng 5 phương án tạo cấu trúc dựa vào sự kết hợp giữa các chương trình sắp gióng cột nêu ở trên cùng với cách thức tạo cấu trúc tự động hay bán tự động của chương trình Modeller.

**Bảng 1.** Các phương án tạo mô hình tương đồng

| STT | Tên phương án | Công cụ sắp gióng cột | Cách thức tạo cấu trúc |
|---|---|---|---|
| 1 | *auto* | MODELLER | Tự động |
| 2 | *autoclustal* | Clustal | Tự động |
| 3 | *autotcoffee* | T-Coffee | Tự động |
| 4 | *clustal* | Clustal | Bán tự động |
| 5 | *tcoffee* | T-Coffee | Bán tự động |

Để chọn phương án tạo cấu trúc tối ưu nhất sử dụng cho toàn bộ quy trình dự đoán, chúng tôi thực hiện 4 lần kiểm tra như sau:

**Bảng 2.** 4 lần kiểm tra để chọn phương pháp tối ưu

| | *Lần 1* | *Lần 2* | *Lần 3* | *Lần 4* |
|---|---|---|---|---|
| *Dữ liệu* | 1034 trình tự từ CED và IEDB | 168 trình tự lấy ngẫu nhiên từ CED và IEDB, chia làm 2 bộ | 283 trình tự tạo ra ngẫu nhiên, chia làm 3 bộ | 220 trình tự lấy ngẫu nhiên từ CED và IEDB |
| *Phương pháp tạo cấu trúc* | 5 phương pháp | 5 phương pháp | 5 phương pháp | Phương pháp cho kết quả tốt nhất từ 3 lần kiểm tra trước |
| *Mục tiêu* | Đánh giá khái quát khả năng tạo cấu trúc của 5 phương pháp | Kiểm tra khả năng tối ưu hóa cấu trúc của chương trình Modeller | Đánh giá khả năng tạo cấu trúc với trình tự có độ tương đồng thấp | Kiểm tra giá trị RMSD giữa cấu trúc tạo ra và cấu trúc thực nghiệm |

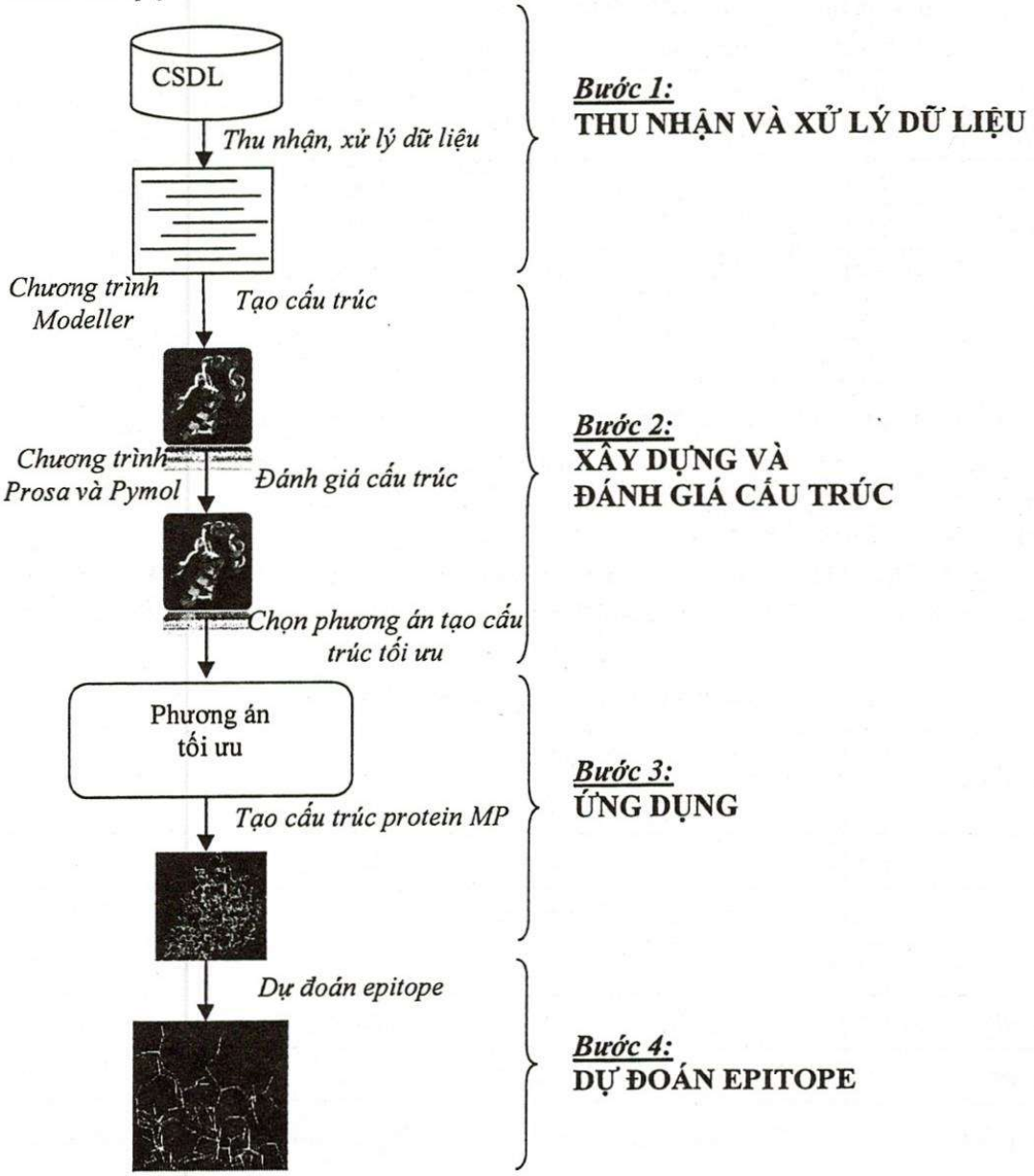
***Bước 3: Ứng dụng***

Sau bước 2, chúng tôi xác định được phương án tạo cấu trúc tối ưu là *autoclustal* và phương án này được sử dụng để tạo cấu trúc cho 53 trình tự protein MP của virus H5N1 ở Việt Nam.

***Bước 4: Dự đoán epitope***

Cấu trúc của các protein MP đã tạo ra được sắp gióng cột cấu trúc với các cấu trúc epitope mẫu bằng chương trình CE. nhằm dự đoán epitope trên trình tự mục tiêu. Kết quả dự đoán epitope bằng quy trình do chúng tôi xây dựng được so sánh với kết quả dự đoán bằng server

CEP (một server dự đoán epitope không liên tục với độ chính xác 75%) để đánh giá khả năng dự đoán của quy trình.




**Hình 1**. Sơ đồ quy trình thực hiện

## 3. KẾT QUẢ VÀ BIỆN LUẬN

### 3.1. Kết quả thu nhận dữ liệu

Kết quả thu nhận dữ liệu từ các CSDL CED, IEDB, NCBI và UniProtKB được tóm tắt trong bảng 3.

**Bảng 3**. Kết quả thu nhận và xử lý dữ liệu

| CSDL | Số mẫu tin ban đầu | Số lượng trình tự thu được | | | Chức năng |
|---|---|---|---|---|---|
| | | Dạng pdb | | Dạng fasta | |
| CED | 293 | 1406 | 1412 | **1034** | Trình tự mục tiêu dùng **xây dựng mô hình** |
| IEDB | 73 | 51 | | | |
| NCBI | 87 | | | **53** | Trình tự protein MP dùng **dự đoán epitope** |
| UniProtKB | 84 | | | | |
| PDB | 43.994 | **43.994** | | **43.895** | **Trình tự cấu trúc** mẫu |
| CED | 293 | 1174 | **1180** | | Trình tự **epitope có cấu trúc** |
| IEDB | 73 | 34 | | | |

### 3.2. Kết quả các lần kiểm tra chọn phương án tạo cấu trúc tối ưu

➢ Lần 1: Kết quả đánh giá tỉ lệ tạo cấu trúc tốt của 5 phương án *auto, autoclustal, autotcoffee, clustal, tcoffee lần lượt là* 82,84%, 84,23%, 82,94%, 68,33% và 66,63%. Như vậy, *autoclustal* cho tỷ lệ số cấu trúc tốt là cao nhất (84,23%).

➢ Lần 2: Thực hiện ưu hóa với thuật giải CG với 2 bộ dữ liệu dùng để kiểm tra. Bộ dữ liệu 1 gồm có 86 trình tự, bộ dữ liệu 2 có 82 trình tự. Kết quả của bộ dữ liệu 1 được thể hiện ở **Bảng 4.**

**Bảng 4.** Tỷ lệ số cấu trúc tốt của bộ dữ liệu 1

| Phương pháp | Số bước tối ưu hóa | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 00 | 50 | 00 |
| ***clustal*** | 5,05% | 4,52% | 2,47% | 2,19% | 1,42% | 0,51% |
| ***tcoffee*** | 2,30% | 1,75% | 9,81% | 9,23% | 7,17% | 7,10% |
| ***auto*** | 2,64% | 3,20% | 1,82% | 2,13% | 2,05% | 1,98% |
| ***autoclustal*** | 2,04% | 1,77% | 0,86% | 2,08% | 2,77% | 1,98% |
| ***autotcoffee*** | 0,94% | 1,22% | 1,48% | 1,67% | 1,34% | 0,35% |

Bảng 5 mô tả kết quả tối ưu hóa của bộ dữ liệu thứ 2 theo 5 phương pháp tạo cấu trúc.

**Bảng 5.** Tỷ lệ số cấu trúc tốt của bộ dữ liệu 2

| Phương pháp | Số bước tối ưu hóa | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 |
| ***clustal*** | 64,96% | 63,96% | 63,43% | 62,41% | 60,38% | 60,40% | 59,47% | 58,22% | 56,68% |
| ***tcoffee*** | 69,55% | 66,87% | 65,80% | 64,16% | 63,46% | 62.03% | 61.78% | 61.11% | 59.96% |
| ***auto*** | 81,74% | 81,50% | 80,15% | 80,18% | 79,90% | 79,77% | 80,26% | 79,56% | 78,74% |
| ***autoclustal*** | 78,86% | 78,88% | 77,34% | 76,37% | 76,73% | 78,89% | 79,17% | 78,43% | 76,09% |
| ***autotcoffee*** | 78,92% | 78,61% | 77,51% | 76,71% | 76,85% | 78,10% | 78,02% | 78,92% | 77,85% |

Các lần thực hiện ở cả 2 bộ dữ liệu với các bước tối ưu hóa khác nhau cho thấy số bước tối ưu hóa không tương quan tuyến tính đến số cấu trúc tốt. Vì vậy chúng tôi không thực hiện bước tối ưu hóa trong quy trình.

➢Lần 3: Thực hiện kiểm tra với 3 bộ dữ liệu (bộ 1 có 95 trình tự, bộ 2 và 3 có 94 trình tự), kết quả ở Bảng 6 cho thấy phương pháp autoclustal cho tỷ lệ số cấu trúc tốt cao nhất ở 2 trong 3 bộ dữ liệu.

**Bảng 6.** Tỷ lệ số cấu trúc tốt của 3 bộ dữ liệu

| Bộ dữ liệu | Phương pháp | | | | |
|---|---|---|---|---|---|
| | *auto* | *autoclustal* | *autotcoffee* | *clustal* | *tcoffee* |
| **1** | 45% | 54% | 51% | 24% | 27% |
| **2** | 49% | 52% | 48% | 32% | 31% |
| **3** | 51% | 54% | 57% | 32% | 31% |

➢Lần 4: Sau khi xác định phương pháp cho kết quả tốt nhất ở các lần kiểm tra trên là phương pháp autoclustal, chúng tôi sử dụng phương pháp này để tạo cấu trúc và kiểm tra cấu trúc thông qua giá trị RMSD. Kết quả cho thấy có 90% số cấu trúc kiểm tra tương đồng với cấu trúc thực nghiệm (tương ứng giá trị RMSD trong khoảng 0-1.5). Điều này nói lên độ chính xác cao của cấu trúc tạo ra bởi phương pháp autoclustal.

### 3.3. Kết quả dự đoán epitope tế bào B không liên tục virus H5N1 ở Việt Nam

Phương pháp autoclustal sẽ được sử dụng trong quy trình dự đoán epitope. Chúng tôi tạo được 22 cấu trúc trong tổng số 53 trình tự protein MP mục tiêu. Sau đó đem dự đoán epitope tế bào B không liên tục trên 22 cấu trúc này. Kết quả dự đoán sẽ so sánh với kết quả tính toán bằng server CEP. **Bảng 7** mô tả kết quả dự đoán epitope trên 22 cấu trúc protein MP:

**Bảng 7.** Kết quả dự đoán epitope

| *STT* | *Trình tự mục tiêu* | *Chiều dài (amino acid)* | *Vị trí epitope* |
|---|---|---|---|
| 1 | Q1KJN4_9INFA | 252 | E23, E29*, K98*, K104* |
| 2 | Q1KJM8_9INFA | 252 | E29*, K104 |
| 3 | Q1KJK8_9INFA | 252 | E23*, E29*, K98, K104* |
| 4 | 72398515 | 252 | E23*, E29*, K98, K104* |
| 5 | 72398530 | 252 | E23*, E29*, K98*, K104* |
| 6 | 72398545 | 252 | E23*, E29*, K98*, K104* |
| 7 | 72398548 | 252 | E23*, E29*, K98, K104* |
| 8 | 72398557 | 252 | E23*, K98 |
| 9 | 86753786 | 252 | E23, E29*, K98*, K104* |
| 10 | 93212892 | 252 | E15*, E21*, K90*, K96* |
| 11 | 93212895 | 252 | E23*, E29*, K98*, K104* |
| 12 | 93212952 | 252 | E15*, E21*, K90, K96* |
| 13 | 93212994 | 252 | E15*, E21*, K90*, K96* |
| 14 | 93213036 | 252 | E7*, E13*, K82*, K88* |

| | | | |
|---|---|---|---|
| 15 | 95116892 | 252 | E23*, K98 |
| 16 | 115951813 | 252 | E23*, E29*, K98*, K104* |
| 17 | 145284445 | 252 | E23, E29*, K98*, K104* |
| 18 | 172052831 | 252 | E22, E28* , K97*, K103* |
| 19 | 172052926 | 252 | E23*, K98 |
| 20 | 172053268 | 252 | E18*, E24*, K93 *, K99* |

*(Ghi chú: các vị trí đánh dấu sao "*" là các amino acid được dự đoán trùng với kết quả của server CEP.)*

20/22 cấu trúc được dự đoán có chứa các amino acid có khả năng nằm trong epitope, trong đó có 60/72 vị trí giống với kết quả dự đoán bằng server CEP. Đồng thời qua khảo sát, tất cả các amino acid dự đoán đều nằm trên bề mặt của cấu trúc, hai amino acid Glutamate (E) và Lysine (K) đều thuộc nhóm thích nước, phân cực và tích điện. Điều này cho thấy các amino acid E và K trên có khả năng là thành phần của epitope không liên tục trong các trình tự protein MP đem dự đoán.

## 4. KẾT LUẬN

Phương pháp mô hình hóa tương đồng là một phương pháp dự đoán cấu trúc cho độ chính xác cao, đặc biệt là khi thu được các trình tự mẫu phù hợp. Các cấu trúc của phương pháp này tạo ra có đủ khả năng áp dụng vào những nghiên cứu đòi hỏi cấu trúc có độ phân giải cao như thiết kế thuốc, dự đoán epitope... Quy trình dự đoán epitope dựa trên phương pháp mô hình hóa tương đồng đã được xây dựng và ứng dụng trên đối tượng là protein matrix của virus H5N1 đã đem lại kết quả tương đối tốt. Các amino acid được dự đoán có nhiều khả năng thuộc về các epitope không liên tục của trình tự mục tiêu.

## PREDICTION OF B-CELL DISCONTINUOUS EPITOPES ON MATRIX PROTEIN OF H5N1 VIRUS

**Tran Ngoc Vinh, Vo Cam Quy, Tran Linh Thuoc**
University of Science, VNU-HCM

***ABSTRACT:*** *Although discontinuous epitopes make up 90% of total number of B-cell epitopes, however, because of difficulties in the development of method for their prediction, most of the B-cell epitope prediction methods today focus on continuous epitopes. To serve for the development of vaccine against H5N1 virus, we have been studying on in silico prediction of T- and B-cell continuous as well as B-cell discontinuous epitopes on H5N1 viral antigens. In this study, using the homology modeling method, we have generated structures of matrix protein of the H5N1 virus and predicted B-cell discontinuous epitopes. 60 out of 72 predicted residues were similar with those reported by the CEP method (Conformational Epitope Prediction). All predicted aminoacid residues were hydrophilic, polar, electrically charged and located on the surface of the antigen structures.*

***Key words:*** *discontinuous B-cell epitope, matrix protein, H5N1, epitope prediction*

## TÀI LIỆU THAM KHẢO

[1].Boris Steipe, *Homology Modeling*, University of Toronto.

[2].Marc A. Marti-Renom, Andras Fiser, M.S. Madhusudhan, Bino John, Ashley Stuart, Narayanan Eswar, Ursula Pieper, Min-yi Shen & Andrej Sali, *Modeling Protein Structure from Its Sequence,* Current Protocols in Bioinformatics 5.1.1-5.1.32, (2003).

[3].Robert Sánchez & Andrej Sali, *Comparative Protein Structure Modeling*, Protein Structure Prediction: Methods and Protocols (143), pp. 97-129, (2000).