XÁC ĐỊNH VÙNG BẢO TỒN CHỨC NĂNG VÀ DỰ ĐOÁN EPITOPE TẾ BÀO T TRÊN CÁC PROTEIN VIRUS CÚM A

Văn Hải Vân, Lê Thị Thanh Thủy, Cao Thị Ngọc Phượng, Vũ Thị Bích
Trần Linh Thước
Trường Đại học Khoa học Tự nhiên, ĐHQG –HCM
(Bài nhận ngày 06 tháng 01 năm 2009, hoàn chỉnh sửa chữa ngày 26 tháng 01 năm 2009)

TÓM TẮT: *Virus cúm A hiện đang là mối quan tâm toàn cầu do sự biến đổi nhanh chóng không ngừng về cấu trúc di truyền. Dựa trên các cơ sở dữ liệu thực nghiệm virus cúm A, chúng tôi đã tiến hành phân tích các vùng bảo tồn chức năng trên các protein của virus nhằm hỗ trợ cho quá trình thiết kế vắcxin đa trị và dự đoán xu hướng biến đổi của các chủng virus. Nghiên cứu được thực hiện trên 11 protein chức năng là HA, NA, PA, NS1, NS2, M1, M2, NP, PB1, PB1_F2 và PB2. Từ các nhóm chức năng, các trình tự protein được phân nhóm theo các subtype, vật chủ, quốc gia và năm phân lập, sau đó được thực hiện sắp gióng cột nhiều trình tự bằng hai công cụ ClustalW và MAFFT. Các trình tự bảo tồn dài 9 amino acid được chọn để dự đoán epitope tế bào T bằng hệ thống dự đoán SEP (System for Epitope Prediction). Ngoài ra, chúng tôi cũng đã thực hiện việc dự đoán chức năng các vùng bảo tồn này dựa trên thông tin chức năng của protein virus cúm A từ cơ sở dữ liệu Swissprot.*

Từ khóa: *epitope bảo tồn, virus cúm A, vắcxin đa trị, vắcxin in silico, khai khoáng dữ liệu*

1.GIỚI THIỆU

Virus cúm A thuộc họ *Orthomyxoviridae* có hình thái rất đa dạng với đường kính 80 – 120nm và chiều dài có thể lên đến 2µm. Bộ gen virus cúm A bao gồm 8 mảnh RNA sợi đơn, mạch âm mã hóa cho các protein với chức năng chính được thể hiện trên ***bảng 1*** [1],[3].

Bảng 1. Các mảnh RNA và các protein được mã hóa của virus cúm A

Mảnh RNA	Tên protein	Chức năng
4	Hemagglutinin (HA)	Gắn thụ thể, dung hợp màng tế bào chủ và virus khởi đầu quá trình xâm nhiễm
6	Neuraminidase (NA)	Tránh sự kết tụ virus và hỗ trợ giải phóng phần tử virus mới
7	Matrix protein (MP) M1 M2	Tương tác với bộ gen và các nhân tố ngoại nhân, hỗ trợ đóng gói virus Kênh ion, kiểm soát pH trong Golgi của quá trình tổng hợp HA và cởi bỏ lớp vỏ virion
5	Nucleoprotein (NP)	Tổng hợp virus Phức hợp phiên mã
1	Basic polymerase protein 2 (PB2)	Tiểu đơn vị gắn mũ chụp, polymerase, quyết định độc lực
2	Basic polymerase protein 1 (PB1) và PB1_F2	Tiểu đơn vị xúc tác của RNA polymerase
3	Acidic polymerase protein (PA)	Tiểu đơn vị, RNA polymerase
8	Non-structural protein 1 (NS1) Non-structural protein 1 (NS2)	Kiểm soát RNA sau phiên mã, trung hòa interferon Hỗ trợ sự di chuyển ra ngoài nhân của RNA virus, đóng gói virus

Phần vỏ của virus cúm A có bản chất là lipid, trên bề mặt có các glycoprotein HA và NA. Đây không chỉ là các nhân tố quan trọng khởi đầu sự xâm nhiễm và giải phóng những phân tử virus mới mà còn là những kháng nguyên chính cho các đáp ứng miễn dịch của vật chủ. Dựa trên sự đa dạng kháng nguyên của HA và NA, virus cúm A được phân ra thành nhiều subtype. Trong số đó chỉ có các subtype H1N1, H2N3, H3N2, H5N1, H7N7 và H9N2 đã được phân lập trên người.

Từ khi xuất hiện đến nay, bệnh cúm do virus cúm A đã gây nên nhiều trận dịch bệnh [4,6]. Những trận dịch cúm A lớn trong thế kỷ qua đều do sự biến đổi thành phần bộ gen ban đầu (quá trình 'antigenic drift') hoặc do sự tái sắp xếp bộ gen của các subtype (quá trình antigenic shift). Năm 1918, sự biến đổi bộ gen của virus cúm A subtype H1N1 tạo khả năng lây nhiễm trên người gây nên dịch cúm Tây Ban Nha với ít nhất 50 triệu người thiệt mạng. Subtype H1N1 còn gây nên các trận dịch khác vào năm 1950, năm 1977 và cho đến nay vẫn còn lưu hành trong quần thể người. Năm 1957, H1N1 ở người tái sắp xếp 3 mảnh gen với H2N2 ở chim (PB1, HA và NA) tạo thành subtype H2N2 ở người gây nên trận đại dịch cúm Châu Á. Năm 1968, xảy ra sự tái sắp xếp giữa gen mã hóa PB1 và HA từ một virus cúm H3 ở chim và các mảnh còn lại của H2N2 ở người tạo thành subtype H3N2 và gây nên đại dịch cúm Hồng Kông năm 1997. Dịch cúm H5N1 bắt đầu ở Hồng Kông năm 1997, bùng phát trở lại ở Nga vào năm 2003, lan sang Đông Nam Á năm 2004, sau đó trải rộng sang Nga, Châu Âu, Châu Phi, lục địa Ấn Độ và Trung Đông trong suốt cuối năm 2005 đến nay. Do đó, để bảo vệ con người khỏi các trận đại dịch cúm tiếp theo, phát triển vắcxin phòng bệnh virus cúm A đang rất được quan tâm nghiên cứu trên thế giới. Hiện nay, các nghiên cứu tập trung chủ yếu trên các vắcxin bất hoạt và vắcxin nhược độc, nhưng các vắcxin này chỉ có khả năng phòng bệnh đối với các virus trong các trận dịch bệnh đã xảy ra và chỉ đặc hiệu cho một vài chủng virus. Ngoài ra, các hạn chế khác trong việc nghiên cứu vắcxin hiện nay là tiêu tốn nhiều thời gian, khó nuôi cấy *in vitro* một số chủng virus độc lực cao, vắcxin bất hoạt có thể hồi tính tạo độc lực… Vì vậy, yêu cầu một quy trình sản xuất vắcxin nhanh chóng, linh hoạt, đặc biệt với sự biến đổi liên tục của virus cúm A, vấn đề phát triển vắcxin phổ rộng là xu hướng tất yếu.

Số lượng trình tự protein của virus cúm A tăng bùng nổ trong các cơ sở dữ liệu sinh học công cộng. Bên cạnh đó, Tin sinh học ra đời từ cuối thế kỉ 20 với hạt nhân là so sánh các trình tự sinh học đã cung cấp nhiều phương pháp và công cụ giúp khai thác nguồn dữ liệu khổng lồ của sinh học thực nghiệm. Sắp gióng cột nhiều trình tự được biết đến như một trong những công cụ tiên quyết trong phân tích gen và protein cũng như cung cấp các thông tin cần thiết để nghiên cứu mối quan hệ chức năng và tiến hóa của các trình tự [2,7,8].

Từ các cơ sở trên, trong nghiên cứu này, các công cụ sắp gióng cột nhiều trình tự trong tin sinh học được sử dụng nhằm khai thác nguồn dữ liệu trình tự của virus cúm A. Bằng kết quả sắp gióng cột, chúng tôi đã xác định và phân tích được các vùng bảo tồn chức năng cho các nhóm protein của virus cúm A. Các vùng bảo tồn dài 9 amino acid đã được sử dụng để dự đoán epitope tế bào T bảo tồn góp phần vào việc phát triển vắcxin phòng bệnh cúm A có phổ rộng. Ngoài ra, đây là cơ sở khoa học định hướng cho các nghiên cứu tiếp theo về xu hướng tiến hóa của virus cúm A để từ đó có thể dự đoán được sự biến đổi của virus cúm A trong tương lai.

2.VẬT LIỆU VÀ PHƯƠNG PHÁP

2.1.Vật liệu

Dữ liệu trình tự protein virus cúm A được thu nhận trực tiếp từ cơ sở dữ liệu "Influenza Virus Resource" của NCBI từ địa chỉ http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html. Các trình tự protein mẫu và thông tin chức năng của các protein virus cúm A lần lượt được thu

nhận từ các cơ sở dữ liệu RefSeq và SwissProt bằng công cụ tìm kiếm của NCBI. Quy trình nghiên cứu được thực hiện trên nền Linux RedHat Enterprise 5.0 và được viết bằng ngôn ngữ lập trình Perl. Các protein được sắp gióng cột hai trình tự bằng công cụ Needle 5.0.0 tích hợp trong gói chương trình EMBOSS và được sắp gióng cột đa trình tự sử dụng chương trình Clustalw-mpi 0.13 và Mafft 6.240. Clustalw là chương trình được sử dụng phổ biến dựa trên thuật toán sắp gióng cột lũy tiến của Freg và Doolittle đưa ra năm 1987. Clustalw-mpi là một phiên bản của Clustalw hổ trợ chạy song song trên nhiều máy tính. Mafft [5] là chương trình sắp gióng cột nhiều trình tự tích hợp nhiều phương pháp nên khá linh động và phù hợp cho nhiều tập hợp sắp gióng cột khác nhau. Mafft có ba chiến lược sắp gióng cột chính: a) phương pháp lũy tiến sử dụng ma trận điểm (FFT-NS-2), b) phương pháp cải tiến có lặp sử dụng hàm tính điểm Weigh Sum of Pair-WSP (FFT_NS-i), c) phương pháp cải tiến có lặp sử dụng hàm tính điểm WSP và hàm điểm dựa trên độ nhất quán (L-INS-i). Phương pháp lũy tiến trong Mafft sử dụng hai kỹ thuật quan trọng nhằm giảm thời gian tính toán của bộ vi xử lý đó là thuật giải sắp gióng cột nhóm-nhóm FFT (Fast Fourier Transform) và phương pháp 6mer để so sánh các cặp trình tự.

Các vùng trình tự bảo tồn dài 9 amino acid được dự đoán epitope tế bào T bằng hệ thống dự đoán SEP (System for Epitope Prediction). SEP được xây dựng bởi Phòng thí nghiệm Tin-Sinh học, Trường Đại học Khoa học Tự nhiên tích hợp 3 mô hình dự đoán epitope tế bào T là HMMs (Hidden Markov Models), SVMs (Support Vector Machines) và ANNs (Artificial Neural Networks).

2.2.Phương pháp

Quy trình thực hiện xác định vùng bảo tồn chức năng và dự đoán epitope tế bào T của protein virus cúm A được tóm tắt chi tiết trong sơ đồ ở ***hình 1***.

Thu nhận dữ liệu trình tự protein

Tất cả trình tự protein virus cúm A được thu nhận trực tiếp từ cơ sở dữ liệu Influenza Virus Resource, trong đó loại bỏ các trình tự giống nhau. Các trình tự protein mẫu đại diện cho protein virus cúm A được thu nhận từ cơ sở dữ liệu protein RefSeq. Các trình tự protein mẫu này được dùng để loại bỏ các trình tự không có độ tin cậy cao trong cơ sở dữ liệu Influenza Virus Resource. Các thông tin về chức năng của protein được thu nhận từ các mẫu tin trình tự protein thuộc cơ sở dữ liệu SwissProt.

Tinh lọc và phân loại trình tự

Tất cả các trình tự protein thô được xử lý loại bỏ các trình tự con của các trình tự lớn và các trình tự chứa ký tự không thuộc bảng mã 20 amino acid. Sau đó, các trình tự này được thực hiện sắp gióng cột 2 trình tự bằng chương trình Needle với các trình tự mẫu để phân vào các nhóm protein chức năng tương ứng. Bước này nhằm đảm bảo loại bỏ các trình tự rác (ngắn), hay các trình tự giải mã chưa hoàn chỉnh. Từ các nhóm protein chức năng, dựa vào thông tin đặc tả, các trình tự tiếp tục được phân vào các nhóm subtype, vật chủ, quốc gia và năm phân lập.

Sắp gióng cột các nhóm trình tự

Các trình tự protein sau khi được phân thành các nhóm mục tiêu sẽ được sắp gióng cột nhằm xác định các vị trí bảo tồn đặc trưng cho từng nhóm. Clustalw là chương trình sắp gióng cột được phát triển từ lâu, đồng thời có được độ tin cậy cao của người sử dụng. Thông số cho chương trình Clustalw-mpi là mặc định. Tuy nhiên, chương trình Mafft với các thuật toán cải tiến được chứng minh là có độ chính xác cao và thời gian thực hiện nhanh hơn so với Clustalw khi thực hiện sắp gióng cột trên số lượng trình tự lớn. Thông số khảo sát trên Mafft là thông số

mặc định sử dụng chiến lược sắp gióng cột FFT-NS-2, và thông số tự động thay đổi theo số lượng trình tự.

Xác định vùng bảo tồn trên các nhóm trình tự

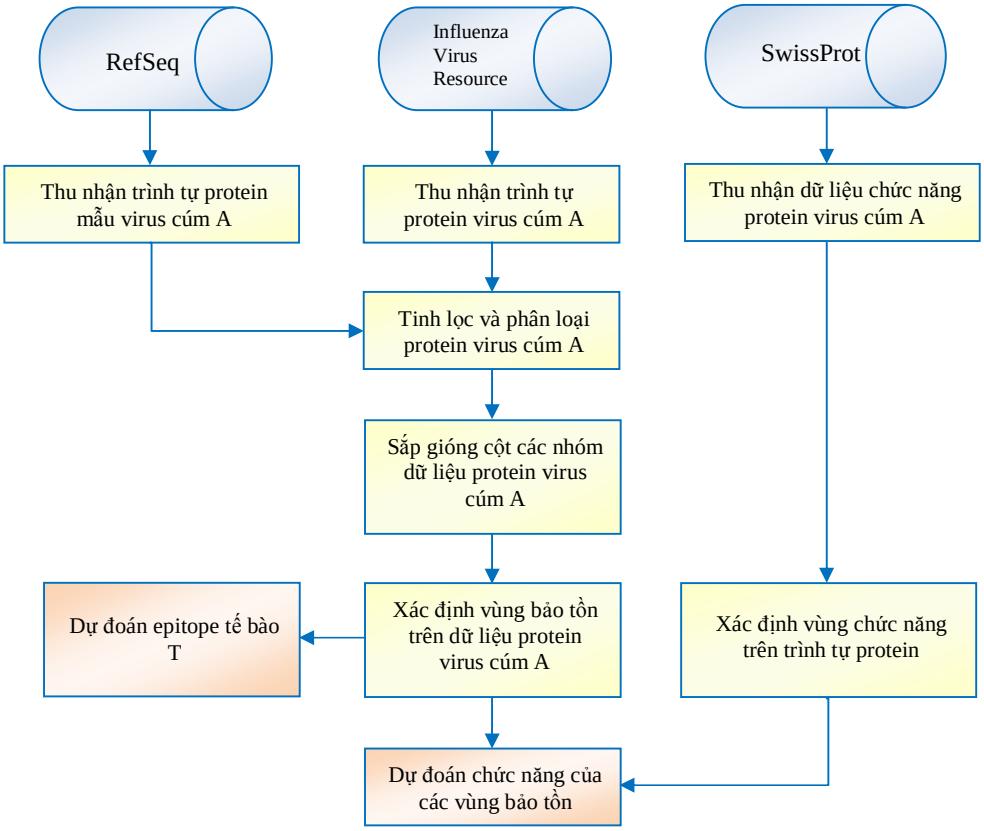
Các vùng trình tự bảo tồn có chiều dài tối thiểu một amino acid của sắp gióng cột được xác định dựa vào tỉ lệ bảo tồn trên 25% (bảo tồn trên 25% tổng số trình tự được sắp gióng cột trong từng nhóm).

Dự đoán chức năng của vùng bảo tồn

Dữ liệu chức năng protein của virus cúm A được thu nhận từ NCBI với giới hạn nguồn cơ sở dữ liệu gốc là Swissprot. Vị trí vùng bảo tồn trên sắp gióng cột và vị trí vùng chức năng trên trình tự thuộc sắp gióng cột được so sánh với nhau để dự đoán chức năng cho vùng bảo tồn (nếu có). Vị trí vùng bảo tồn và vị trí vùng chức năng đạt được sự phù hợp khi vùng bảo tồn nằm trong và gần nhất với vùng chức năng.

Dự đoán epitope tế bào T

Các trình tự bảo tồn dài 9 amino acid được thu nhận để thực hiện dự đoán epitope tế bào T bằng hệ thống SEP.

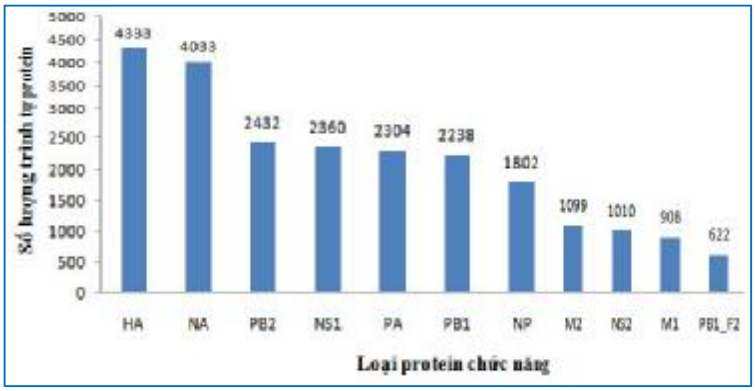


Hình 1. Quy trình xác định vùng bảo tồn, dự đoán chức năng và epitope tế bào T của protein virus cúm A

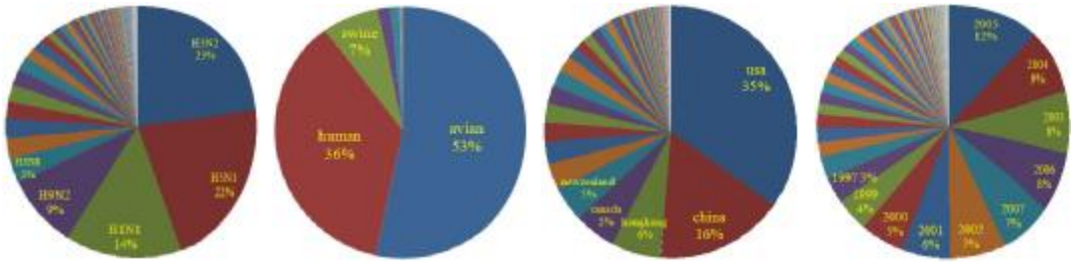
3.KẾT QUẢ VÀ THẢO LUẬN

3.1.Trình tự protein virus cúm A

Sau khi tinh lọc, chúng tôi thu nhận được 23.141 trình tự các protein virus cúm A. Trong đó, số lượng từng loại protein chức năng được thể hiện trên ***hình 2*** với 2 protein HA và NA có số lượng trình tự lớn nhất do được quan tâm nghiên cứu nhiều. Kết quả phân nhóm trình tự protein virus cúm A theo subtype, vật chủ, quốc gia và năm phân lập được mô tả trên biểu đồ ***hình 3***.



Hình 2. Biểu đồ số lượng trình tự của các protein virus cúm A



Hình 3. Phần trăm số lượng trình tự protein virus cúm A phân loại theo subtype, vật chủ, quốc gia và năm

Số liệu phân nhóm trình tự theo subtype cho thấy số lượng trình tự thuộc các subtype H3N2, H5N1, H1N1, H9N2 và H3N8 chiếm gần 75% tổng số trình tự. Đây cũng là các subtype được chứng minh là có độc lực cao, có khả năng gây nhiễm trên người và đã gây ra các trận đại dịch lớn. Vật chủ có số lượng trình tự nhiều nhất là chim (53%) tiếp đến là người (36%) và lợn (7%). Quốc gia phân lập virus cúm A nhiều nhất là Mỹ (35%), Trung Quốc (16%) và Hồng Kông (6%), đây là những quốc gia đã từng bùng phát đại dịch cúm A (Trung Quốc, Hồng Kông) hoặc là những quốc gia đi đầu trong kế hoạch phòng chống dịch cúm (Mỹ). Thống kê số lượng trình tự theo năm cho thấy virus cúm A được phân lập chủ yếu những năm gần đây cùng với thời điểm bùng phát của dịch cúm H5N1, bắt đầu từ 3% năm 1997 tăng dần qua các năm và cao điểm là năm 2005 với 12% tổng số trình tự.

3.2.Các vùng bảo tồn chức năng trên các nhóm protein

Số lượng vùng bảo tồn của các protein virus cúm A được thể hiện trên ***bảng 2***. Sắp gióng cột bằng chương trình Mafft với thông số tự động (Mafft_auto) và thông số mặc định

(Mafft_default) cho cùng số lượng vùng bảo tồn đối với tất cả các protein. So sánh giữa hai chương trình Mafft và Clustalw-mpi, số lượng vùng bảo tồn chỉ có sự khác biệt không đáng kể trên các protein HA, NA, PB2, PB1 và NP. Như vậy, với thời gian thực hiện sắp gióng cột của Clustalw_mpi là 48 giờ trên 3 máy tính, Mafft_auto là 20 giờ và Mafft_default là 12 giờ trên một máy đơn, chúng tôi đề nghị sử dụng chương trình Mafft_default với thời gian thực hiện nhanh nhất.

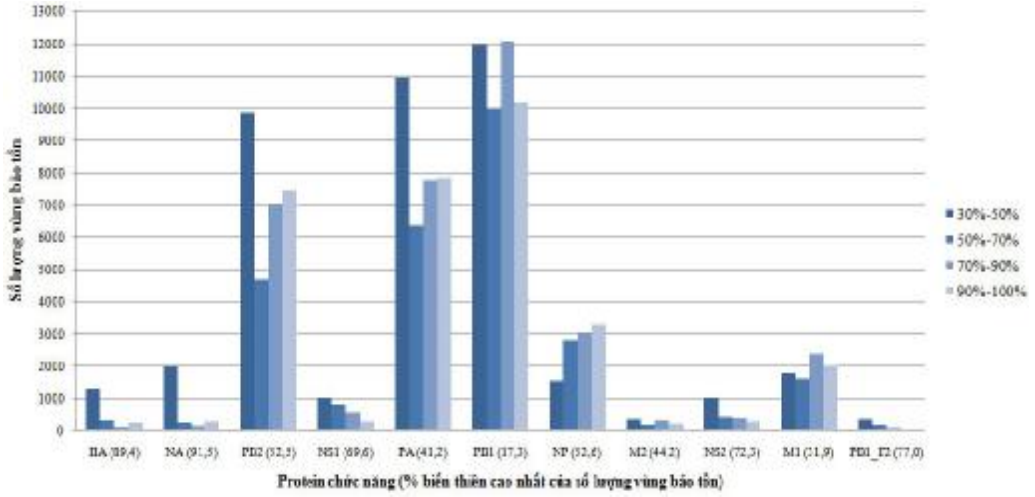
Bảng 2. Số lượng vùng bảo tồn trên các protein virus cúm A

	HA	NA	PB2	NS1	PA	PB1	NP	M2	NS2	M1	PB1_F2
Clustalw_mpi	2008	2722	29122	2761	32986	44277	10730	1110	2150	7903	725
Mafft_auto	2001	2695	29101	2761	32986	43919	10572	1110	2150	7903	725
Mafft_default	2001	2695	29101	2761	32986	43919	10572	1110	2150	7903	725

Số lượng vùng bảo tồn ở các tỉ lệ bảo tồn khác nhau của các protein virus cúm A trên ***hình 4*** cho thấy độ bảo tồn giảm dần theo thứ tự PB1, M1, PA, M2, PB2, NP, NS1, NS2, PB1_F2, HA và NA. Như vậy, HA và NA là hai protein có độ biến động cao nhất của virus cúm A với đa số các vùng bảo tồn ở tỉ lệ bảo tồn 30-50%. Các protein nền và protein trong phức hợp phiên mã có độ bảo tồn cao nhất. Tham khảo chức năng của các protein trên cơ sở dữ liệu SwissProt, vùng bảo tồn chức năng đặc trưng trên các protein virus cúm A được thể hiện trên ***bảng 3***. Protein HA có các vùng bảo tồn chức năng quan trọng là 'cleavage', 'lipid-binding', 'signal' và 'transmembrane region'. Tương tự, protein NA có các vùng 'sialidase', 'active', 'transmembrane region' và 'binding'. Đặc biệt, chức năng sialidase bảo tồn trong NA ở 2 vị trí khác nhau trên kết quả sắp gióng cột (vị trí 447 và vị trí 474, cách nhau khoảng 26 amino acid). Điều này cho thấy có thể chức năng này của NA được tiến hóa theo 2 hướng khác nhau.

Bảng 3. Vùng bảo tồn chức năng đặc trưng trên các protein virus cúm A

Tên protein	Tên vùng chức năng	Chiều dài (tối thiểu – tối đa)	Số trình tự (tối thiểu – tối đa)	Tên vùng chức năng	Chiều dài (tối thiểu – tối đa)	Số trình tự (tối thiểu – tối đa)
HA	Signal	1-2	1301-2868	Transmembrane region	1-6	1299-2890
	Cleavage	1-2	4159-4333	lipid-binding	1-1	3428-3649
NA	Sialidase	1-5	1503-1747	Transmembrane region	1-7	1229-3337
	Active	1-1	3979-2049	binding	1-1	4027-4030
M2	Flu_M2	4-26	329-944	Transmembrane region	1-20	335-1097



Hình 4. Số lượng vùng bảo tồn của các protein cúm A ở các tỉ lệ bảo tồn khác nhau

3.3.Các epitope tế bào T bảo tồn

Bảng 4. Các epitope tế bào T bảo tồn được dự đoán tốt nhất của HA

STT	Trình tự peptide	Số lượng trình tự chứa peptide	HMM-SVM	HMM-ANN	SVM-ANN	HMM-SVM-ANN
1	GRIQDLEKY	1322	hla_b_2705	hla_b_2705	hla_b_2705	hla_b_2705
2	KIDLWSYNA	1324	hla_a_0201, hla_a_0202, hla_a_0206	hla_a_0201	hla_a_0201, hla_a_0203	hla_a_0201
3	GLFGAIAGF	3187	hla_a_0201	hla_a_0201, hla_b_1501	hla_a_0201, hla_a_0202, hla_a_0203	hla_a_0201

Bảng 5. Các epitope tế bào T bảo tồn được dự đoán tốt nhất của NA

STT	Trình tự peptide	Số lượng trình tự chứa peptide	HMM-SVM	HMM-ANN	SVM-ANN	HMM-SVM-ANN
1	WSWPDGAEL	1248	hla_b_3501	hla_b_3501	hla_a_0206, hla_b_3501, hla_b_5101	hla_b_3501
2	DVFVIREPF	1356	hla_b_3501	hla_b_3501	hla_a_6801, hla_a_6802, hla_b_3501	hla_b_3501
3	YICSGVFGD	1420	hla_a_0201	hla_a_0201	hla_a_0201, hla_a_0202, hla_a_6801	hla_a_0201

4	HLECRTFFL	1461	hla_a_0201, hla_a_0202	hla_a_0201, hla_a_0202	hla_a_0201, hla_a_0202, hla_a_0206, hla_a_6801	hla_a_0201, hla_a_0202
5	APFSKDNSI	1857	hla_b_0702, hla_b_5401	hla_b_5401,	hla_b_5401	hla_b_5401
6	NPNQKIITI	2868	hla_b_5301	hla_b_5101, hla_b_5301, hla_b_5401	hla_b_5301	hla_b_5301

Các peptide bảo tồn dài 9 amino acid của 11 protein chức năng được thực hiện dự đoán khả năng gắn với 20 alen của HLA lớp I bằng các phương pháp HMM, ANN và SVM. Các epitope tế bào T bảo tồn được dự đoán tốt nhất bởi 2 phương pháp và 3 phương pháp trên cùng alen của HA và NA được trình bày trên ***bảng 4*** và ***bảng 5***. Trong đó, 2 peptide GLFGAIAGF và NPNQKIITI bảo tồn cao trong các trình tự HA và NA.

4.KẾT LUẬN

Chúng tôi đã xác định thành công các vùng bảo tồn chức năng cho các protein của virus cúm A. Các vùng bảo tồn này là cơ sở cho các nghiên cứu về sự biến đổi và tiến hóa của virus cúm A. Bên cạnh đó, từ các vùng bảo tồn dài 9 amino acid, các epitope tế bào T bảo tồn được dự đoán tốt nhất đã được đề xuất nhằm phục vụ cho mục tiêu thiết kế vắcxin cúm A có phổ rộng.

INDENTIFYING FUNCTIONALLY CONSERVED REGIONS AND PREDICTING T-CELL EPITOPES ON PROTEINS OF INFLUENZA A VIRUS

Van Hai Van, Le Thi Thanh Thuy, Cao Thi Ngoc Phuong, Vu Thi Bich
Tran Linh Thuoc
University of Science, VNU-HCM

ABSTRACT: *Influenza A viruses are of worldwide concerns because of their rapidly and endlessly genetic changes. Based on the experimental influenza A virus databases, we analyzed conserved regions on the protein sequences of influenza A virus to facilitate the design of universal vaccine and the prediction of changing tendency of influenza A viral strains. Our study was carried out on eleven viral functional proteins: HA, NA, PA, NS1, NS2, M1, M2, NP, PB1, PB1_F2 and PB2. From these groups, the clusters were formed on subtypes, hosts, countries and years of collection, followed by multiple sequence alignments by two tools ClustalW and MAFFT. Conserved sequences of 9 amino acid residues were selected and used for T-cell epitope prediction by SEP (System for Epitope Prediction). In addition, we also predicted the function of these conserved regions using information on function of influenza A viral proteins from the Swissprot database.*

Key words: *conserved epitope, influenza A virus, universal vaccine, vaccine in silico, data mining.*

TÀI LIỆU THAM KHẢO

[1].Cheung KWH, Poon LLM. *Biology of Influenza A Virus*, Ann NY Acad Sci 1102: 1-25 (2007).

[2].Edgar RC, *Batzoglou S.Multiple sequence alignment*, Curr Opin Struct Biol 16: 368 – 373 (2006).

[3].Engelhardt OG, Fodor E. *Functional association between viral and cellular transcription during influenza virus infection*, Rev Med Virol 16: 329 – 345 (2006).

[4].Horimoto T, Kawaoka Y. *Influenza: lessons from past pandemics, warnings from current incidents*, Nat Rev Microbiol 3: 591 – 600 (2005).

[5].Katoh K, Toh H. *Recent developments in Mafft multiple sequence aligment program*, Brief Bioinform 9: 286 – 298 (2008).

[6].Monto AS, Gravenstein SG, Elliot M, Colopy M, Schweinle. *Clinical Signs and Symptoms Predicting Influenza Infection*, Archives of Internal Medicine 160: 3243-47 J (2002).

[7].Nuin PA, Wang Z, Tillier ER. *The accuracy of several multiple sequence alignment programs for proteins*, BMC Bioinformatics 7: 471 (2006).

[8].Pirovan W, Heringa J. *Multiple Sequence Alignment*, *Methods Mol Biol* 452: 143 – 161 (2008).