LẬP CHỈ MỤC THEO NHÓM ĐỂ NÂNG CAO HIỆU QUẢ KHAI THÁC CƠ SỞ DỮ LIỆU VIRUS CÚM

Trương Thị Đức, Trương Thị Quỳnh Hương, Nguyễn Thụy Mai Trâm
Võ Hồng Bảo Châu, Tạ Thúc Nhu
Khoa Công nghệ thông tin, Trường Đại học Lạc Hồng
10 Huỳnh Văn Nghệ, Biên Hòa, Đồng Nai
{duc,huong,maitram,chau,nhu}@lhu.edu.vn

TÓM TẮT

Virus cúm (influenza) là một loại RNA virus, chính là nguyên nhân gây ra bệnh cúm ở người và động vật. Với khả năng biến đổi và lan truyền nhanh từ động vật sang động vật, từ động vật sang người, và đặc biệt là từ người sang người; virus cúm là một trong những loài virus nguy hiểm nhất cho nền kinh tế cũng như sức khỏe con người trên toàn thế giới từ trước đến nay. Chính vì vậy, sự hiểu biết về cấu trúc phân tử của nó là một nhu cầu lớn trong các nghiên cứu về dịch bệnh. Hiện nay, các tổ chức y tế, cũng như các ngân hàng dữ liệu trên thế giới đã lưu trữ nhiều trình tự sinh học liên quan đến virus cúm. Tuy nhiên, các ngân hàng dữ liệu sinh học này không chứa thông tin chi tiết đến các tỉnh thành của một quốc gia. Vì vậy, chúng ta không có đầy đủ thông tin để biểu diễn quá trình lây nhiễm, cũng như phân tích virus cúm ở Việt Nam một cách đầy đủ, đặc biệt có đủ thông tin để phục vụ cộng đồng.

Bài viết này trình bày giải pháp xây dựng cơ sở dữ liệu để có thể bổ sung dữ liệu virus cúm ở Việt Nam cho đến mức độ tỉnh thành; đồng thời đưa ra thuật toán lập chỉ mục theo nhóm qua đó có thể giúp cho việc khai thác thông tin theo tiêu chí người dùng về virus cúm nhanh chóng và hiệu quả. Thuật toán cho phép chọn lựa những trình tự sinh học với mức độ tương đồng khác nhau để truy vấn; sau đó nhóm những kết quả dựa trên quan hệ họ hàng của chúng với nhau. Bên cạnh đó, bài viết cũng trình bày giải pháp cho phép cập nhật dữ liệu một cách tự động từ các ngân hàng dữ liệu về virus cúm trên thế giới, đặc biệt là ngân hàng dữ liệu của NCBI (National Center for Biotechnology Information)

1. Đặt vấn đề

Sự phát triển mạnh mẽ của công nghệ sinh học đã giúp chúng ta giải mã bộ gen của virus cúm trong một thời gian ngắn với chi phí vừa phải. Dự án giải mã toàn bộ hệ gen của virus cúm đã được triển khai tại nhiều nơi như Viện nghiên cứu quốc gia về các bệnh truyền nhiễm, Hoa Kỳ (NIAID) từ những năm 2004 [1]

Một lượng lớn dữ liệu sinh học phân tử (các trình tự DNA/protein) của virus cúm đã được giải mã và lưu trữ ở các cơ sở dữ liệu dùng chung của thế giới như Trung tâm Thông tin về công nghệ sinh học Hoa Kỳ - NCBI (National Center for Biotechnology Information). NCBI hiện đang lưu giữ hơn 100.000 trình tự DNA/protein của virus cúm được thu thập và giải mã từ nhiều quốc gia trên thế giới trong suốt thời gian qua.

Với một lượng dữ liệu khổng lồ đã được thu thập, việc xây dựng các hệ thống thông tin, xây dựng các công cụ tìm kiếm và phân tích dữ liệu đang được phát

triển mạnh mẽ trên thế giới. Qua đó giúp chúng ta hiểu được cơ chế lây nhiêm, tạo ra vắc-xin mới, theo dõi và kiểm soát dịch bệnh.

Nổi bật trong các hệ thống đó là hệ thống thông tin virus cúm của NCBI (http://www.ncbi.nlm.nih.gov/genomes/FLU/) được phát triển bởi Bao và các đồng nghiệp năm 2008 [2]**Error! Reference source not found.**. Hệ thống hiện lưu giữ hơn 100.000 trình tự DNA/protein của các loài virus cúm khác nhau. Một số chức năng chính của hệ thống là:

- Cung cấp thông tin về virus cúm theo nhiều tiêu chí khác nhau như: loại virus cúm (cúm A, cúm B, cúm C), động vật chủ (người, gia cầm,..), quốc gia, loại protein.

Cung cấp một số công cụ tìm kiếm và phân tích dữ liệu như: tìm kiếm BLAST [1]**Error! Reference source not found.**, sắp hàng đa trình tự **Error! Reference source not found.**, xây dựng cây phát sinh loài [8], v.v…

Tuy nhiên, các thông tin do hệ thống NCBI cung cấp chỉ chi tiết đến mức độ quốc gia. Tức là không chi tiết đến mức độ các tỉnh thành trong một quốc gia. Hệ thống cũng không cung cấp công cụ cho phép hiện thị và theo dõi quá trình lây nhiễm của virus cúm.

Một số nghiên cứu về virus cúm tiêu biểu:

- Ngoài nước:
  - Trung tâm Thông tin về công nghệ sinh học Hoa Kỳ - NCBI (National Center for Biotechnology Information) http://www.ncbi.nlm.nih.gov/genomes/FLU/
  - Viện nghiên cứu genome Bắc Kinh, Trung Quốc, xây dựng cơ sở dữ liệu virus cúm IVDB (http://influenza.psych.ac.cn/).
  - Phòng thí nghiệm Quốc gia Los Alamos (http://flu.lanl.gov/)
  - Trường đại học Hàn Quốc và Viện Sức khỏe quốc gia xây dựng "Cơ sở dữ liệu genome cúm và quyết định kháng nguyên" ISED (http://influenza.korea.ac.kr)
- Trong nước:
  - Viện Công nghệ sinh học (Institute of Biotechnology - IBT) đã tiến hành nghiên cứu và giải mã nhiều trình tự virus cúm H5N1
  - Cục thú y trung ương đã tiến hành giải mã toàn bộ hệ gen của 33 virus cúm ở nhiều tỉnh thành khác nhau từ 10/2005 đến 5/2007: Đồng tháp, Sóc Trăng, An Giang, Hà Tây, Vĩnh Long, Hà Nội,v.v…
  - Nhóm nghiên cứu của TS. Lê Sỹ Vinh ở Trường Đại học Công nghệ, thuộc Đại học Quốc gia Hà Nội tiến hành phát

triển các phương pháp và công cụ tin sinh học để phân tích dữ liệu virus cúm thu được

- Nhóm nghiên cứu của PGS. Trần Văn Lăng ở Phân viện Công nghệ thông tin tại TPHCM trước đây, nay là Viện Cơ học và Tin học ứng dụng (Institute of Mechanics and Informatics – IAMI) thuộc Viện Khoa học và Công nghệ Việt Nam đã nhiều năm nghiên cứu, xây dựng các công cụ tin sinh phục vụ cho việc nghiên cứu các trình tự DNA/protein làm nền tảng cho việc nghiên cứu vi khuẩn và virus.

Mặc dù nhiều nghiên cứu về virus cúm đã được tiến hành ở Việt Nam, các nghiên cứu chủ yếu tập trung vào việc giải mã các trình tự DNA và protein, qua đó tiến hành một số phân tích để tìm hiểu mối quan hệ giữa chúng.

Tuy nhiên, hiện nay chúng ta còn thiếu một hệ thống tin giúp các nhà quản lý (bộ, ngành y tế); các nhà chuyên môn; và người dân có được thông tin, dữ liệu, cũng như những công cụ phân tích (thống kê, mô hình) về virus cúm trên thế giới, đặc biệt chi tiết hóa cho virus cúm ở Việt Nam.

Nghiên cứu này tập trung xây dựng công cụ cung cấp thông tin về virus cúm bao gồm các chức năng:

- Thiết kế một cơ sở dữ liệu chứa thông tin về virus cúm trên thế giới và chi tiết hóa dữ liệu virus cúm ở Việt Nam cho đến mức độ tỉnh thành
- Tự động cập nhật dữ liệu từ ngân hàng dữ liệu NCBI.
- Lập chỉ mục theo nhóm
- Xây dựng công cụ cung cấp thông tin virus cúm

2. Phương pháp nghiên cứu

Thiết kế một cơ sở dữ liệu chứa thông tin về virus cúm trên thế giới và chi tiết hóa dữ liệu virus cúm ở Việt Nam cho đến mức độ tỉnh thành

Bắt đầu từ nguồn dữ liệu mà NCBI lưu trữ **ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/**

Và thông tin từng file dữ liệu của Nucleotic, Protein, Gene

Các file này chứa đầy đủ thông tin của 1 gene, 1 protein hoặc 1 nucleotic.. Yêu cầu cần thiết phải thiết kế một cơ sở dữ liệu có thể lưu trữ các thông tin này nhưng phải thêm phần chi tiết đến tỉnh thành ở Việt Nam, đồng thời phải dễ dàng cho việc cập nhật tự động, truy xuất và hiển thị thông tin.

Xem hình về file thông tin của 1 nucleotic

Influenza A virus (A/chicken/Egypt/1052S-NLQP/2010(H5N1)) segment 4 hemagglutinin (HA) gene, partial cds

```
LOCUS       GU811748               1584 bp    cRNA    linear   VRL 21-APR-2010
DEFINITION  Influenza A virus (A/chicken/Egypt/1052S-NLQP/2010(H5N1)) segment 4
            hemagglutinin (HA) gene, partial cds.
ACCESSION   GU811748
VERSION     GU811748.1  GI:289900038
KEYWORDS    .
SOURCE      Influenza A virus (A/chicken/Egypt/1052S-NLQP/2010(H5N1))
  ORGANISM  Influenza A virus (A/chicken/Egypt/1052S-NLQP/2010(H5N1))
            Viruses; ssRNA negative-strand viruses; Orthomyxoviridae;
            Influenzavirus A.
REFERENCE   1  (bases 1 to 1584)
  AUTHORS   Arafa,A.A., Hagag,N.M., Abdullah,M.H., Yehia,N.M.,
            Abdel-Halim,A.M., Kilany,W.H., Ahmed,M.S., Zanaty,A.M.,
            Abdel-Aziz,O.M., Hassan,M.K. and Aly,M.M.
  TITLE     Genetic analysis of recent Egyptian H5N1 viruses
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1584)
  AUTHORS   Arafa,A.A., Hagag,N.M., Abdullah,M.H., Yehia,N.M.,
            Abdel-Halim,A.M., Kilany,W.H., Ahmed,M.S., Zanaty,A.M.,
            Abdel-Aziz,O.M., Hassan,M.K. and Aly,M.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-FEB-2010) National Laboratory for Veterinary Quality
            Control on Poultry Production, Nadi-Elsaid Street, Dokki, Giza
            12618, Egypt
FEATURES             Location/Qualifiers
     source          1..1584
                     /organism="Influenza A virus
                     (A/chicken/Egypt/1052S-NLQP/2010(H5N1))"
                     /mol_type="viral cRNA"
                     /strain="A/chicken/Egypt/1052S-NLQP/2010"
                     /serotype="H5N1"
                     /isolation_source="farm"
                     /host="chicken"
                     /db_xref="taxon:720653"
                     /segment="4"
                     /country="Egypt: Qaliobia"
                     /collection_date="Feb-2010"
     gene            <1..>1584
                     /gene="HA"
     CDS             <1..>1584
                     /gene="HA"
                     /codon_start=3
                     /product="hemagglutinin"
                     /protein_id="ADD21384.1"
                     /db_xref="GI:289900039"
                     /translation="ANNSTEQVDTIMEKNVTVTHAQDILEKTHNGKLCDLDGVKPLIL
                     RDCSVAGWLLGNPMCDEFPNVSEWSYIVEKTNPANDLCYPGNFNNYEELKHLLSRINR
                     FEKIKIIPKSSWPDHEASLGVSSACPYQGGPSFYRNVVWLIKKNNTYPTIKESYHNTN
                     QEDLLVLWGIHHPNDEEEQTRIYKNPTTYISVGTSTLNQRLVPKIATRSKVNGQSGRV
                     EFFWTILKSNDTINFESNGNFIAPENAYKIVKKGDSTIMKSELEYGNCSTKCQTPVGA
                     INSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNSPQGEGRRKKRGLFGAIAGFIEG
                     GWQGMVDGWYGYHHSNEQGSGYAADRESTQKAIDGVTNKVNSIIDKMNTQFEAVGREF
                     NNLEKRIENLNKKMEDGFLDVWTYNAELLVLMENERTLDFHDSNVKNLYDKVRLQLRD
                     NAKELGNGCFEFYHRCDNECMESVRNGTYDYPQYSEEARLKREEISGVKLESIGTYQI
                     LSIYSTVASSLALAIIVAG"
```

Hình 1: Thông tin đầy đủ của nucleotic

Từ các thông tin trên, mô hình quan niệm dữ liệu được thiết kế.

Tự động download dữ liệu từ ngân hàng dữ liệu NCBI

Ngân hàng dữ liệu NCBI cho phép download dữ liệu về nhưng phải sử dụng thủ công. Số lượng các file virus cúm rất lớn, hơn 100.000, việc download từng file là không thực hiện được. Module tự động download dữ liệu sẽ tự động lấy dữ liệu và lưu trữ vào thư mục được chỉ định. Yêu cầu của module này là phải được kết nối với Internet. Tốc độ thực hiện tùy thuộc vào tốc độ đường truyền Internet.

Tự động cập nhật dữ liệu vào cơ sở dữ liệu

Các file virus được download về là từng file riêng lẻ. Thông tin của các virus này cần phải được trích ra và lưu vào cơ sở dữ liệu để có thể truy xuất sau này. Việc trích lọc các thông tin từ các file phải được thực hiện tự động và yêu cầu chính xác, nhanh chóng. Module cập nhật tự động có đầy đủ các khả năng này.

Lập chỉ mục theo nhóm

Dữ liệu virus cúm sau khi được thu thập sẽ được lập chỉ mục theo các nhóm ưu tiên cho việc tìm kiếm và khai thác thông tin virus cúm. Việc lập chỉ mục được dựa trên các công cụ được cung cấp sẵn như Blast và dựa vào tính tương đồng của các chuỗi trình tự. Sau khi các chuỗi trình tự được lập chỉ mục sẽ giúp cho việc tìm kiếm, thống kê và biểu diễn trở nên hiệu quả hơn

Công cụ cung cấp thông tin virus cúm

Công cụ cung cấp thông tin virus cúm thực chất là một website cho phép người dùng tìm kiếm, thống kê các thông tin về virus cúm. Hệ thống website có giao diện thân thiện, dễ sử dụng và cho truy xuất, hiển thị thông tin

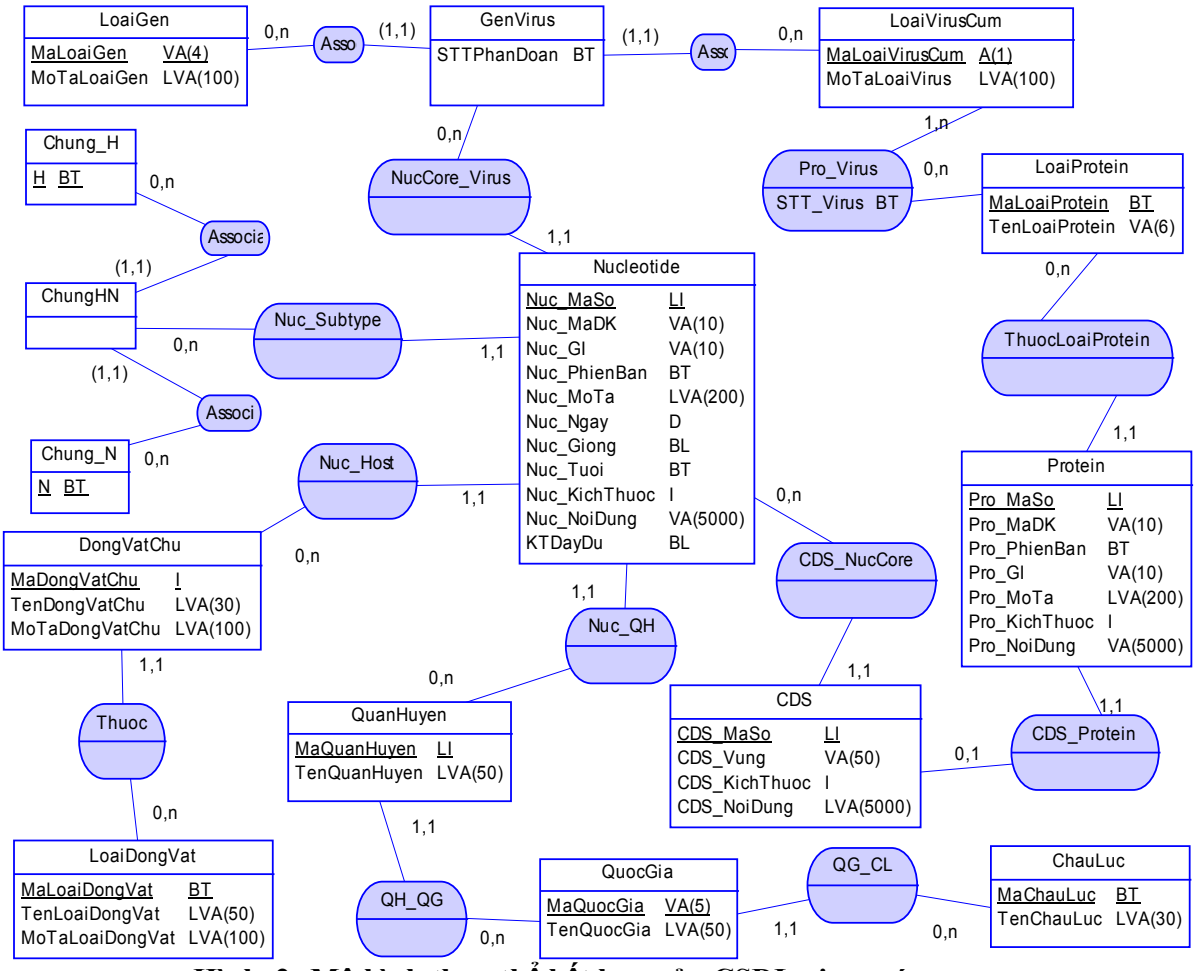
Việc cung cấp các công cụ thống kê về dữ liệu và sự lây lan của virus cúm là hết sức cần thiết. Công cụ gồm các chức năng:

- Cho phép người dùng lựa chọn thống kê về virus cúm theo nhiều tiêu chí khác nhau
- Thống kê và biểu diễn kết quả về virus cúm theo vị trí địa lý (quốc gia, tỉnh thành ở Việt Nam)
- Thống kê và biểu diễn kết quả về virus cúm theo thời gian
- Thống kê và biểu diễn kết quả sự phát triển của virus cúm theo loại và chủng virus

3. Kết quả thực hiện

Nghiên cứu đã đạt được các kết quả như sau:

- Cơ sở dữ liệu Virus cúm chi tiết đến từng tỉnh thành



Hình 2: Mô hình thực thể kết hợp của CSDL virus cúm

- Module tự động download dữ liệu từ NCBI
- Module tự động cập nhật dữ liệu virus cúm, chi tiết hóa đến từng tỉnh thành

Hình 3: Giao diện module tự động download và cập nhật dữ liệu virus cúm

- Cơ sở dữ liệu virus cúm được lập chỉ mục.
- Hệ thống website cung cấp các thông tin về virus cúm

Hình 4: Giao diện website cung cấp thông tin virus cúm

4. Kết luận

Các nghiên cứu ở Việt Nam thường được thực hiện riêng rẽ, chưa có sự gắn kết. Hệ thống sẽ giúp lưu trữ dữ liệu một cách tập trung qua đó giúp cho việc tìm kiếm, hiển thị và nghiên cứu về virus cúm ở Việt Nam một cách đầy đủ và tổng thể, làm phong phú thêm ngân hàng dữ liệu về virus cúm. Nhờ dữ liệu được lập chỉ mục, việc khai thác các thông tin virus cúm trở nên nhanh và dễ dàng hơn.

Hệ thống website được đưa lên mạng Internet có thể giúp cho người dân có những hiểu biết nhất định về sự phân bổ virus cúm trên toàn lãnh thổ, đồng thời cũng có thể cung cấp dữ liệu có các tổ chức y tế có nhu cầu

Tài liệu tham khảo

[1] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). *Basic local alignment search tool*. *J Mol Biol* **215** (3): 403–410.

[2] Bao Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, D. Lipman (2008) *The Influenza Virus Resource at the National Center for Biotechnology Information*. J. Virol. 2008 Jan; 82(2):596-601.

[3] Chang, S., Zhang, J., Liao, X., Zhu, X., Wang, D., Zhu, J., Feng, T., Zhu, B., Gao, G.F., Wang, J. et al. (2007) *Influenza Virus Database (IVDB): an integrated information resource DNA analysis platform for influenza virus research*. Nucleic Acids Res, 35, D376-380

[4] Dang Cao Cuong, Le Si Quang, Le Sy Vinh (2009). *Influenza-specific amino acid substitution model*, The first international conference on knowledge DNA systems engineering, Hanoi.

[5] Edgar RC (2004) *MUSCLE: multiple sequence alignment with high accuracy DNA high throughput*. Nucl. Acids Res. 2004, 32:1792–1797.

[6] Fauci A: Race against time. Nature 2009, 435:423–42

[7] Nguyen TD, et al (2008) *Multiple Sublineages of Influenza A Virus* (H5N1), Vietnam, 2005-2007. Emerging Infectious Diseases 2008, 14:632–636.

[8] Saitou N, Nei M (1987). *The Neighbor-Joining method: a new method for reconstructing phylogenetic trees*. Mol Biol Evol **4** (4): 406-425

[9] Trần Văn Lăng và cộng sự. *Nghiên cứu để xây dựng công cụ tin học xử lý thông tin về Gene và Protein*. Đề tài cấp bộ, Viện Khoa học và Công nghệ Việt Nam quản lý, 2003-2004

[10] Trần Văn Lăng và cộng sự. *Tính toán hiệu năng cao và tính toán lưới trong một số bài toán sinh học*. Đề tài thuộc chương trình Nghiên cứu cơ bản, 2006-2007

[11] Trần Văn Lăng. *Ứng dụng Tin học trong việc giải một số bài toán thuộc Sinh học phân tử*, Nxb. Giáo dục, 2008