CƠ SỞ DỮ LIỆU GENOME: CÔNG CỤ ĐỂ PHÂN TÍCH BỘ GEN NGƯỜI VIỆT NAM

Nguyễn Đăng Tôn[1], Phan Thanh Hải[2], Trần Đức Nghĩa[2], Nông Văn Hải[1]

[1]*Viện Công nghệ sinh học*
[2]*Viện Công nghệ thông tin*

TÓM TẮT

Dữ liệu thông tin về bộ gen đầu tiên của dự án giải mã genome người (HGP) được đưa ra cho toàn thế giới cùng sử dụng gọi là trình tự chuẩn. Sau thành công của dự án này, hàng loạt dự án quốc tế về giải mã genome người đã và đang được tiến hành ở nhiều viện nghiên cứu, trường đại học và các tổ chức khác trên phạm vi toàn thế giới. Góp phần vào sự thành công của các dự án giải mã genome là các cơ sở dữ liệu genome và các công cụ tin sinh học được sử dụng để phân tích, so sánh và chú giải các trình tự mới đọc được. Nhằm mục đích thu thập các số liệu trình tự genome, chúng tôi đã tiến hành xây dựng cơ sở dữ liệu bước đầu về genome người tại Việt Nam, gọi tắt là "VGENOME" và trong tương lai sẽ được bổ sung thường xuyên các dữ liệu mới. Cơ sở dữ liệu có nhiệm vụ thu thập lưu trữ toàn bộ dữ liệu về trình tự genome người đã được công bố trên các ngân hàng gen quốc tế. Hệ thống cơ sở dữ liệu được xây dựng trên mô hình kiến trúc Web-base 3 lớp, thực hiện xử lý dữ liệu tập trung. Hiện tại, đã thiết lập được cơ sở dữ liệu với 5820 trình tự nucleotide, 10077 trình tự protein và 191059 trình tự gen đã biết chức năng trong cơ sở dữ liệu. Cơ sở dữ liệu cũng tích hợp công cụ BLAST nhằm tìm kiếm trình tự tương đồng trong phạm vi genome người được lưu trữ.

Từ khóa: BLAST, công cụ tin sinh học, bộ gen người, cơ sở dữ liệu, cơ sở dữ liệu genome người Việt Nam

ĐẶT VẤN ĐỀ

Bộ gen người có một cấu trúc hết sức tinh vi và phức tạp, gồm 2 thành phần: i) Bộ gen nhân: kích thước khoảng 3,2 tỷ bp; và ii) Bộ gen ty thể có kích thước chỉ hơn 16 kb. Mọi biểu hiện của sự sống, bao gồm các yếu tố quyết định sức khỏe mỗi người đều liên quan đến chức năng gen. Vì vậy, việc nghiên cứu cấu trúc và chức năng toàn bộ các gen của cơ thể là một vấn đề khoa học cơ bản có định hướng ứng dụng hết sức quan trọng.

Cho đến đầu thiên niên kỷ mới, trước khi các số liệu về bộ gen người của dự án Genome người (Human Genome Project - HGP) được công bố năm 2003, chúng ta còn biết rất ít về các gen trong bộ gen và tầm quan trọng của chúng đối với cuộc sống. Tuy nhiên, gần đây, hàng loạt dự án quốc tế về giải mã bộ gen người đã và đang được tiến hành ở nhiều viện nghiên cứu, trường đại học và các tổ chức trên phạm vi toàn thế giới. Bên cạnh các dự án bộ gen lớn mang tầm quốc tế như dự án Genome người, dự án Lập bản đồ kiểu gen đơn bội ở người (Haplotype Map of Human Genome), dự án 1000 bộ gen (1000 Genomes Project), dự án di truyền ngoại sinh ở người (Human Epigenome Project) được thực hiện dưới sự hợp tác của nhiều cơ quan, tổ chức đi đầu trong lĩnh vực genomics, nhiều dự án nhỏ hơn ở tầm quốc gia như các dự án giải mã genome người Trung Quốc, dự án xác định các SNP ở người Nhật Bản và nhiều dự án giải mã toàn bộ trình tự gen người bản địa được tiến hành ở nhiều quốc gia đang phát triển hiện nay cũng đang trong quá trình vận hành, chính là một nguồn dữ liệu vô cùng quan trọng, góp phần làm sáng tỏ nguồn gốc tiến hóa của loài người cũng như cơ chế, cách phòng ngừa và điều trị các bệnh xảy ra ở người.

Từ năm 2003, Viện Công nghệ sinh học thuộc Viện Khoa học và Công nghệ Việt Nam bắt đầu tiến hành nghiên cứu đặc điểm cấu trúc bộ gen người Việt Nam thông qua việc giải mã toàn bộ genome ty thể ở một số cá thể người Việt Nam (Huỳnh Thị Thu Huệ *et al.*, 2005; Nguyễn Thị Tú Linh *et al.*, 2005; Trần Thị Minh Nguyệt *et al.*, 2008; Vũ Hoài Thu *et al.*, 2005; Nguyễn Đăng Tôn *et al.*, 2008). Các kết quả này bước đầu đã bổ sung những số liệu quan trọng phục vụ cho các nghiên cứu về đặc điểm genome, các bệnh liên quan và phục vụ cho công tác giám định cá thể trong khoa học hình sự và pháp y.

Cho đến nay, chưa có cơ sở dữ liệu nào được xây dựng để tra cứu và phân tích trình tự gen của Việt Nam nói chung và phân tích genome nói riêng. Nhằm mục đích cho các nghiên cứu về giải mã và phân tích bộ gen người Việt Nam, chúng tôi tiến hành xây dựng cơ sở dữ liệu để lưu trữ các trình tự nucleotide, trình tự protein của người thu thập được

từ các ngân hàng gen quốc tế phục vụ cho việc phân tích nội bộ của dự án. Ngoài ra, cơ sở dữ liệu còn bước đầu tích hợp công cụ phân tích BLAST. Trong tương lai sẽ bổ sung, cập nhật nhiều chức năng mới.

PHƯƠNG PHÁP NGHIÊN CỨU

Thu thập dữ liệu gen người

Dữ liệu gen người được thu thập từ ngân hàng gen quốc tế Genbank dưới dạng chuẩn XML hoặc chuẩn fasta.

Môi trường và nền tảng phát triển cơ sở dữ liệu

Sử dụng giải pháp phát triển Web DotNetNuke Community 05.01.01 của Microsoft trên nền tảng công nghệ ASP.NET trong bộ Microsoft.Net. Cơ sở dữ liệu được thiết kế trên nền tảng Microsoft SQL Server 2005 Express Edition với webserver là Internet Information Services 6.0. DotNetNuke (DNN) là một giải pháp phát triển Website truy cập cơ sở dữ liệu, hay các Portal, dựa trên công nghệ mã nguồn mở của Microsoft.

KẾT QUẢ VÀ THẢO LUẬN

Các dự án giải mã genome người trên thế giới và công cụ để phân tích dữ liệu

Dự án Genome người (HGP), được thực hiện từ năm 1989 đến 2003, do Nhóm các cơ quan khoa học nhà nước do Mỹ đứng đầu với khoảng 20 nước và vùng lãnh thổ tham gia, đã giải mã hoàn chỉnh bộ gen người (~3,2 tỷ bp), với DNA lấy từ 5 cá thể đại diện 5 chủng tộc người trên thế giới. Đồng thời, việc giải mã bộ gen người cũng đã được Công ty tư nhân Celera Genomics của Mỹ tiến hành. Kết quả là mỗi nhóm giải mã hoàn chỉnh 1 bộ gen người, đồng thời công bố "bản nháp" trên 2 tạp chí khoa học danh tiếng nhất là Nature, Anh (McPherson *et al.*, 2001), và Science, Mỹ (Venter *et al.*, 2001). Trình tự của Nhóm được tài trợ từ ngân sách của các chính phủ đã được công khai, dữ liệu thông tin về bộ gen cho toàn thể giới cùng sử dụng, được gọi là "trình tự chuẩn" hay "trình tự tham chiếu" (reference sequence).

Dự án Lập bản đồ kiểu gen đơn bội quốc tế (hay còn gọi tắt là dự án HapMap) đã được khởi động từ năm 2002 với mục tiêu phát triển một bản đồ kiểu gen đơn bội của genome người, hay còn gọi là bản đồ HapMap, mô tả những kiểu đa hình phổ biến trong trình tự DNA của người. HapMap được kỳ vọng sẽ trở thành một công cụ quan trọng được sử dụng để phát hiện các gen có liên quan chặt chẽ tới sức khỏe, bệnh tật của con người và mở ra hướng mới trong nghiên cứu trị liệu. (http://hapmap.ncbi.nlm.nih.gov/).

Dự án 1000 Bộ gen, được khởi sự từ tháng 1 năm 2008, là một nghiên cứu quốc tế với nỗ lực nhằm thiết lập một bản đồ chi tiết các đa hình di truyền ở người. Năm 2010, dự án đã kết thúc giai đoạn thử nghiệm. Từ cuối năm 2010, dự án sẽ bước vào giai đoạn sản xuất với mục tiêu giải trình tự của 2000 cá thể (http://www.1000genomes.org/).

Được tổ chức và thực hiện bởi Tổ chức Di truyền ngoại sinh ở người (Human Epigenome Consortium - HEC), dự án Di truyền ngoại sinh ở người (Human Epigenome Project) là một nỗ lực chung của sự hợp tác quốc tế nhằm xác định và giải thích các mô hình methyl hóa DNA trên quy mô genome ở tất cả các gen ở người trong tất cả các mô chính trong cơ thể, qua đó giải thích các cơ chế của sự phát triển, tính nhạy cảm với bệnh tật cũng như tính bền vững của genome (Beck *et al.*, 1999, http://www. epigenome.org/).

Dự án genome người Neandertal là dự án hợp tác giữa các nhà khoa học thuộc viện Max Planck về Nhân chủng học tiến hóa tại Đức và công ty tư nhân 454 Life Science tại Hoa Kỳ để giải trình tự genome người Neandertal. Nghiên cứu này đã chỉ ra rằng có một sự pha trộn trong các gen xảy ra giữa người Neandertal và người hiện đại và đã đưa ra những bằng chứng cho rằng các yếu tố trong genome của người Neandertal vẫn còn sót lại trong genome của người hiện đại không có nguồn gốc từ châu Phi (Green *et al.*, 2010).

Trung Quốc là quốc gia đầu tiên, tiên phong trong việc giải trình tự toàn bộ genome trên quy mô lớn của nhiều cá thể người thuộc các dân tộc trong nước. Tháng 10 năm 2007, viện nghiên cứu Genome Bắc Kinh tại Thẩm Quyến (Beijing Genome Institute - BGI) đã thông báo hoàn thành việc giải trình tự genome lưỡng bội của một người Hán Trung Quốc, đại diện cho quần thể châu Á. Bộ gen này là bộ gen khởi đầu cho một dự án giải mã 100 cá thể người Trung Quốc trong ba năm - dự án Viêm Hoàng (Yanhuang Project) (theo tên của Hoàng Đế và Viêm Đế là hai vị hoàng đế - thủy tổ của người Trung Quốc) (Li *et al.*, 2009; Wang *et al.*, 2008).

Dự án được bắt đầu thực hiện vào năm 2000 và được phát triển thông qua dự án Thiên niên kỷ của Thủ tướng Nhật Bản. Mục tiêu của dự án là xác định và so sánh tới 150.000 SNP từ quần thể người Nhật

Bản, nằm trên các gen hoặc các vùng liên gen có thể ảnh hưởng tới trình tự mã hóa của gen. Dự án đã được tiến hành dưới sự hợp tác của Trung tâm Genome người (Human Genome Center - HGC) thuộc Viện Khoa học Y học (Institute of Medical Science - IMS) tại Đại học Tokyo và Tập đoàn Khoa học và Kỹ thuật Nhật Bản (Japan Science and Technology Corporation - JST). Mục tiêu ban đầu của dự án là nhằm xây dựng một cơ sở dữ liệu cơ bản để xác định các mối quan hệ giữa các đa hình và các bệnh thường gặp ở người hoặc phản ứng của cơ thể với các loại thuốc (Hirakawa *et al.*, 2002).

Tại một số nước khác, các dự án giải mã genome người cũng đã và đang được tiến hành như: dự án giải mã bộ gen người Nga (http://www.inauka.ru/news/article97736.html), dự án "100 bộ gen người châu Á" do Đại học Seoul và các nhóm nghiên cứu khác nhau tại Hàn Quốc và các nước khác tham gia (Kim *et al.*, 2009), dự án về cơ sở dữ liệu đa hình/đột biến gen người Thái của Thái Lan (http://www4a.biotec.or.th/GI)...

Với sự phát triển của công nghệ giải trình tự thế hệ mới, người ta có thể giải mã một cá thể trong thời gian rất ngắn. Tuy nhiên, công nghệ này thường tạo ra các đoạn đọc có trình tự rất ngắn, từ 30 đến 100 bp. Chính vì vậy, việc phát triển các công cụ tin sinh học cũng như các hệ thống tính toán hiệu năng cao và lưu trữ lớn là rất quan trọng, góp phần vào sự thành công của dự án.

Bước đầu xây dựng cơ sở dữ liệu genome người Việt Nam

Cơ sở dữ liệu lưu trữ các trình tự gen người trên thế giới

Ngoài trình tự toàn bộ genome của người làm trình tự chuẩn, các trình tự đơn lẻ về nucleotide, trình tự protein của người được lưu trữ trên ngân hàng genbank hiện tại lần lượt là 9670317 và 576742 (theo thống kê của ngân hàng Genbank). Ngoài ra, trên Genbank còn lưu trữ 42036 trình tự các gen đã biết chức năng. Nhằm mục đích lưu trữ toàn bộ số trình tự trên chúng tôi tiến hành xây dựng cơ sở dữ liệu để phân tích và so sánh các trình tự gen trong quá trình nghiên cứu.

Hệ thống cơ sở dữ liệu (CSDL) được xây dựng trên mô hình kiến trúc Web-base 3 lớp, thực hiện xử lý dữ liệu tập trung. Kiến trúc Web-base 3 lớp của hệ thống bao gồm các thành phần sau: 1) Tầng thứ 1: Giao diện cho người dùng, tầng này có nhiệm vụ cung cấp giao diện tương tác thân thiện với người dùng, cho phép người dùng có thể dễ dàng thực hiện các thao tác nghiệp vụ với hệ thống bằng các trình duyệt internet phổ biến như Firefox, IE; 2) Tầng thứ 2: Là phần mềm lớp giữa bao gồm lớp web và lớp các chức năng phần mềm ứng dụng tương tác với CSDL. Phần mềm ứng dụng bao gồm các chức năng chuyển đổi, cập nhật dữ liệu và khai thác thông tin trên CSDL; và 3) Tầng thứ 3: Các phần mềm hệ thống nền tảng và CSDL bên dưới. Với phần mềm hệ Quản trị CSDL trong hệ thống này để xuất sử dụng hệ quản trị CSDL SQL Server của Microsoft. Mô hình kiến trúc tổng thể của hệ thống CSDL được thể hiện ở Hình 1. Thiết kế một CSDL chung nhất cho cả Database và Private Database (như mô tả ở trên) lưu trữ các thông tin về gen của người (*Homo sapiens sapiens*); bao gồm các thông tin về trình tự Nucleotide, Protein và các gen đã biết chức năng. CSDL lưu trữ đầy đủ các thông tin phục vụ cho việc tra cứu, hiển thị theo các tiêu chuẩn chung hiển thị thông tin theo chuẩn Fasta và Genbank và phục vụ cho việc kết xuất báo cáo, thống kê theo yêu cầu.

CSDL lưu trữ ba nhóm thông tin chính là HomoLocus, HomoReference, HomoFeatures và SequenceType (Hình 1B) đáp ứng đầy đủ thông tin cho việc hiển thị theo chuẩn Fasta và Genbank. Chúng được kết nối với nhau qua giá trị "accession" là một giá trị không bị trùng lặp giữa các chuỗi khác nhau.

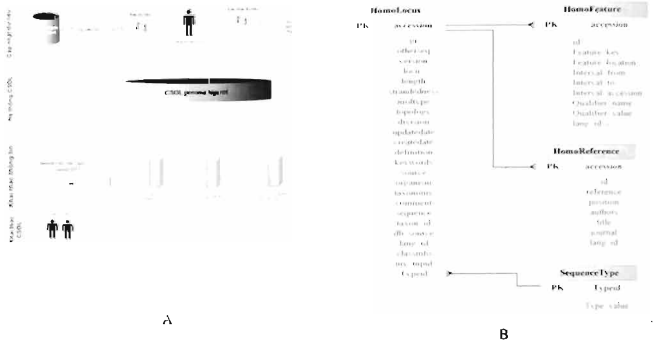
Khả năng tích hợp với các CSDL bên ngoài

Hệ thống được xây dựng có có khả năng tích hợp với các CSDL bên ngoài về gen của genome người, đặc biệt là với các hệ CSDL trên nền Hệ quản trị cơ sở dữ liệu quan hệ (RDBMS) như Microsoft SQL server, Oracle, PostgreSQL, MySQL... Các bộ kết nối CSDL (Database Adapter) có sẵn cho như ZMySQLDA, SyBaseDA, ZOracleDA... cho phép thực hiện các lệnh SQL trong các script của hệ thống CSDL, nhờ đó các ứng dụng khác có thể dễ dàng khai thác, tìm kiếm và trình bày những dữ liệu lấy từ hệ thống CSDL được xây dựng. Bộ kết nối ODBC cho phép tích hợp hệ thống CSDL với bất kỳ hệ thống nào hỗ trợ ODBC, bao gồm cả các phần mềm như Microsoft Exchange và IBM Lotus Notes.

Mô hình chức năng

Mô hình chức năng được thiết kế tương ứng với các lớp người sử dụng khác nhau với các quyền truy cập khác nhau. Mô hình chức năng cho lớp người sử dụng bình thường: Đây là giao diện Web đầu tiên khi người sử dụng bất kỳ kết nối Internet và truy cập địa chỉ Website CSDL genome sẽ được khai thác thông

tin trên CSDL chung (Database). Mô hình chức năng cho người sử dụng tác nghiệp và người quản trị: Lớp người dùng tác nghiệp và người quản trị, sau khi đăng nhập hệ thống qua tài khoản, có quyền cập nhật và khai thác thông tin trên CSDL chung (Database), đồng thời có quyền cập nhật và khai thác trên CSDL cục bộ (Private Database) được cập nhật từ nguồn số liệu.

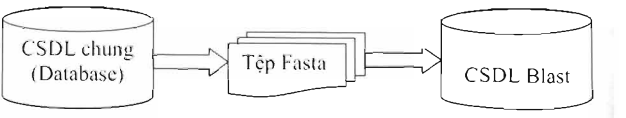


Hình 1. Sơ đồ cơ sở dữ liệu. A: Mô hình kiến trúc tổng thể của hệ thống CSDL genome; B. Mô hình dữ liệu. Dữ liệu gồm 4 bảng "HomoLocus" chứa các trường thông tin liên quan đến trình tự như "moltype" - kiểu trình tự, "gi" mã trình tự, "lenght": kích thước trình tự ; "HomoFeature" bao gồm các trường liên quan đến đặc tính của trình tự, "HomoReference" bao gồm các thông tin liên quan đến sự trích dẫn; và bảng "SequenceType" chưa thông tin về loại trình tự

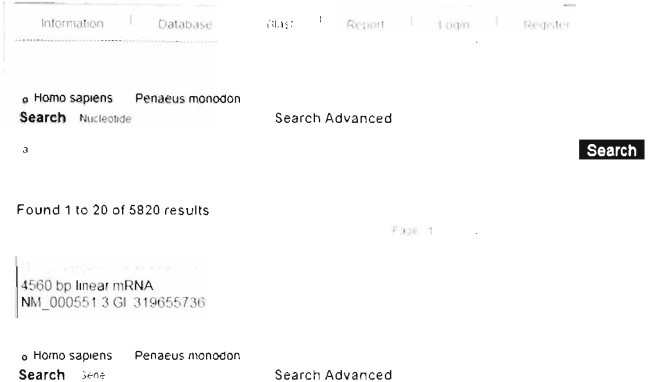
Chức năng BLAST

Ngoài yêu cầu về lưu trữ và khai thác dữ liệu trên CSDL, hệ thống có khả năng tích hợp các công cụ, tiện ích phân tích, thống kê, đặc biệt là công cụ BLAST. Công cụ BLAST sẽ cho phép người sử dụng tìm kiếm trình tự tương đồng trong phạm vi cơ sở dữ liệu genome người đã được xây dựng. Để có thể tích hợp công cụ này vào CSDL, trước hết phải tạo CSDL cho riêng chương trình BLAST (Hình 2). Với trình tự nhập vào (Nucleotide hoặc Protein) được, chương trình sẽ thực hiện BLAST (Blast N hoặc Blast P) trong phạm vi cơ sở dữ liệu.

Từ các dữ liệu trình tự genome người trên Genbank, chúng tôi đã thiết lập được cơ sở dữ liệu với 5820 trình tự nucleotide, 10077 trình tự protein và 191059 trình tự gen đã biết chức năng trong cơ sở dữ liệu (Hình 3). Các trình tự này được đưa vào cơ sở dữ liệu dưới định dạng Genbank, Fasta hoặc XML. Các dữ liệu sau khi đưa vào cơ sở dữ liệu sẽ được cập nhật thường xuyên và được sử dụng cho các phân tích tiếp theo của dự án giải mã bộ gen người Việt Nam.



Hình 2 Mô hình tạo CSDL BLAST



Hình 3. Cơ sở dữ liệu các trình tự nucleotide, trình tự protein và trình tự gen đã biết chức năng được cập nhật

KẾT LUẬN

Chúng tôi đã xây dựng thành công cơ sở dữ liệu genome phục vụ nghiên cứu giải mã genome người, bao gồm các dữ liệu hiện có trong ngân hàng GenBank (5820 trình tự nucleotide, 10077 trình tự protein và 191059 trình tự gen đã biết chức năng). Cơ sở dữ liệu của chúng tôi không chỉ giúp người sử dụng tra cứu các thông tin về trình tự Nucleotide, Protein và trình tự các gen đã biết chức năng của người, mà còn tích hợp công cụ BLAST nhằm tìm kiếm trình tự tương đồng trong phạm vi genome người. Dự kiến, cơ sở dữ liệu này sẽ được đưa lên trang web tại địa chỉ http://www.genome.ac.vn.

Lời cảm ơn: *Công trình này được là một nhiệm vụ của đề tài "Xây dựng cơ sở khoa học cho dự án giải mã genome người Việt Nam" do Bộ Khoa học và Công nghệ cấp kinh phí thực hiện.*

TÀI LIỆU THAM KHẢO

Beck S, Olek A, Walter J (1999) From genomics to epigenomics: a loftier view of life. *Nat Biotechnol* 17(12) 1144.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, *et al.* (2010) A draft sequence of the Neandertal genome. *Science* 328(5979): 710-722.

Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y (2002) JSNP: a database of common gene variations in the Japanese population. *Nucl Acids Res* 30(1): 158-162.

Huỳnh Thị Thu Huệ, Hoàng Thị Thu Yến, Nguyễn Đăng Tôn, Lê Thị Thu Hiền, Nguyễn Đình Cường, Phan Văn Chi, Nông Văn Hải (2005), Phân tích trình tự vùng điều khiển (D-loop) trên genome ty thể của 5 cá thể người Việt Nam, *Tạp chí Công nghệ Sinh học* 3(1). 15-22.

Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, et al. (2009) A highly annotated whole-genome sequence of a Korean

individual. *Nature* 460(7258): 1011-1015.

Li G, Ma L, Song C, Yang Z, Wang X, Huang H, *et al.* (2009) The YH database: the first Asian diploid genome database. *Nucl Acids Res* 37(Database issue): D1025-1028.

Nguyễn Thị Tú Linh, Nguyễn Đình Cường, Nguyễn Đăng Tôn, Lê Thị Thu Hiền, Huỳnh Thị Thu Huệ, Phan Văn Chi, Nông Văn Hải (2005), Phân tích trình tự gen ND5 ty thể của một số cá thể người Việt Nam, *Tạp chí Công nghệ Sinh học* 3(3): 279-286.

McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, *et al.* (2001) A physical map of the human genome. *Nature* 409(6822): 934-941

Trần Thị Minh Nguyệt, Lê Thị Bích Thảo, Bùi Thị Huyền, Phạm Đình Minh, Trần Thế Thành, Nguyễn Thị Tý, Nguyễn Bích Nhi, Đặng Diễm Hồng, Lê Quang Huấn, Quyền Đình Thi, Nguyễn Đăng Tôn, Nông Văn Hải, Phan Văn Chi (2008) Trình tự toàn bộ genome ty thể từ 9 cá thể người Việt Nam. *Tạp chí Công nghệ Sinh học* 6(4A): 569-578.

Nguyễn Đăng Tôn, Nguyễn Thị Tú Linh, Vũ Hải Chi, Trần Thị Ngọc Diệp, Địch Thị Kim Hương, Bùi Thị Tuyết, Nguyễn Hải Hà, Huỳnh Thị Thu Huệ, Lê Thị Thu Hiền, Trần Thị Phương Liên, Phan Văn Chi, Nông Văn Hải (2008) Đa hình kiểu đơn bội DNA ty thể của các cá thể người Việt Nam. *Tạp chí Công nghệ Sinh học*, 6(4A), pp. 579-590.

Vũ Hoài Thu, Nguyễn Đình Cường, Huỳnh Thị Thu Huệ, Nguyễn Đăng Tôn, Lê Thị Thu Hiền, Trần Thị Phương Liên, Phan Văn Chi, Nông Văn Hải (2005), Xác định trình tự gen ND6 ty thể của một số cá thể người Việt Nam, *Tạp chí Công nghệ Sinh học* 3(4): 415-421

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al.* (2001) The sequence of the human genome *Science* 291(5507) 1304-1351

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, *et al.* (2008) The diploid genome sequence of an Asian individual *Nature* 456(7218) 60-65.

GENOME DATABASE: A TOOL TO ANALYSE VIETNAMESE HUMAN GENOME

Nguyen Dang Ton[1], Phan Thanh Hai[2], Tran Duc Nghia[2], Nong Van Hai[1,*]

[1]*Institute of Biotechnology*
[2]*Institute of Information Technology*

SUMMARY

With the completion of the Human Genome Project (HGP), the reference sequence of human genome was published freely in 2003. After successful of the HGP, many other genome projects are being carried out by many institutions, universities and other organizations in the world. Bioinformatics software and genomes databases were useful tools to compare, analyse, and annotate structure and functions of query sequences. The aim of this study was to establish the genome database with the bioinformatics tools for analysis and annotation of the Vietnamese human genome DNA sequences. The database system was designed with 3 layer web base. Most of the sequences of human genome, including nucleotides, proteins, and known gene sequences, published in Genbank, we collected and inserted in to our database. There are 5280 nucleotide sequences, 10077 protein sequences, and 191059 sequences of known genes of human in our preliminary database. The BLAST services, with the most BLAST programs, and other tools are to be included into the database for the purpose of aligning query sequence against sequences in the database.

Keywords: *BLAST, bioinformatics tool, database, human genome, Vietnamese genome database*

* *Author for correspondence:* Tel/Fax: *84-4-8363222;* E-mail *vhnong@ibt.ac.vn*