PGS,TS. NGUYỄN VĂN CÁCH

Tin Sinh học

NHÀ XUẤT BẢN KHOA HỌC VÀ KỸ THUẬT

PGS. TS. NGUYỄN VĂN CÁCH

TIN – SINH HỌC

(Tái bản lần thứ nhất, có sửa chữa và bổ sung)

NHÀ XUẤT BẢN KHOA HỌC VÀ KỸ THUẬT
HÀ NỘI – 2009

Lời nói đầu

Trong nửa cuối thế kỷ XX, nền khoa học công nghệ thế giới đã tạo ra bước phát triển mang tính đột phá ngoạn mục trên rất nhiều lĩnh vực khác nhau, trong đó đặc biệt nhất là ba lĩnh vực tin học, công nghệ thông tin trên nền tảng internet và công nghệ sinh học. Thành công trong lĩnh vực công nghệ sinh học phải kể đến bước phát triển đột phá của công nghệ lên men hiện đại, của sinh học phân tử và kỹ thuật gen, của công nghệ enzym và động học phản ứng... Chính trong thời khắc lịch sử ấy, một lĩnh vực khoa học mới đã ra đời là tin - sinh học.

Tin-sinh học chính là sự hội tụ, hợp tác hữu cơ và đặc biệt hiệu quả của cả ba lĩnh vực công nghệ hàng đầu: tin học- công nghệ truyền thông- công nghệ sinh học, cùng cộng tác với nhau khám phá thế giới sống. Thực tế đã cho thấy, ngay từ khi ra đời tin-sinh học đã thực sự trở thành công cụ nghiên cứu mới, trợ giúp đắc lực và hiệu quả để đẩy mạnh tốc độ nghiên cứu và ứng dụng công nghệ sinh học; chắp cánh cho công nghệ sinh học nói riêng và sinh học nói chung, bay lên tầm cao mới.

Cuốn "Tin - Sinh học" này nhằm cung cấp cho cán bộ và sinh viên ngành công nghệ sinh học, cũng như các đối tượng khác có liên quan, những kiến thức cơ bản về tin - sinh học và một vài thí dụ khai thác ứng dụng môn học. Trong lần tái bản, đã có vài điều chỉnh nhỏ về hình thức và cách trình bày để cập nhật thêm các thông tin mới trong lĩnh vực khoa học trẻ và phát triển năng động này.

Tác giả rất mong nhận được sự đóng góp của độc giả để hiệu chỉnh cho lần in sau được hoàn chỉnh hơn. Xin chân thành cảm ơn sự khích lệ và ủng hộ của đồng nghiệp và của bạn đọc.

Xin chân thành cảm ơn bạn đọc.

PGS. TS. Nguyễn Văn Cách

Hà Nội, 2009

MỤC LỤC

1. MỞ ĐẦU

Sự phát triển như vũ bão của khoa học và công nghệ trong thế kỷ XX đã tạo ra cơ sở lý luận, vật chất và sự liên kết hỗ trợ lẫn nhau, tác động thúc đẩy sự phát triển của mọi lĩnh vực hoạt động của đời sống xã hội. Trong lĩnh vực công nghệ sinh học, nhờ những thành tựu vô cùng to lớn của sinh học và sinh học ứng dụng (đặc biệt là trong các lĩnh vực: di truyền học, sinh học phân tử, kỹ thuật gen, công nghệ lên men hiện đại...), cùng với việc hoàn thiện và hiện đại hoá các trang thiết bị phục vụ nghiên cứu khoa học đã cho phép con người trong khoảng thời gian ngắn thu được khối lượng dữ liệu khoa học khổng lồ về công nghệ sinh học, nói riêng và về khoa học sự sống nói chung. Đồng thời, sự phát triển vô cùng mạnh mẽ của sinh học phân tử và kỹ thuật gen trong nửa cuối thế kỷ XX đã cho phép con người khám phá bản chất sinh học, ở cấp độ phân tử, các đơn vị cơ sở nhỏ nhất cấu thành nên từng bộ phận cơ thể và các quá trình vận động biến đổi xảy ra trong các cơ thể sống. Chính các yếu tố trên đã cấu thành nên cơ sở vật chất ban đầu cho các ngân hàng dữ liệu công nghệ sinh học.

Nguồn dữ liệu cơ sở này, thực tế là các dữ liệu kết quả nghiên cứu thu được của từng cá nhân hay của các cơ sở nghiên cứu rải rác khắp nơi trên thế giới. Với đặc thù là ngành khoa học thực nghiệm, đây chính là sản phẩm kết tinh của khối lượng rất lớn lao động trí tuệ, hao phí vật chất, tiền bạc và tiêu tốn thời gian, công sức. Việc bảo quản tại chỗ kết quả nghiên cứu này là không hiệu quả và không thể tránh khỏi mất mát hay thất lạc, do nhiều nguyên nhân khác nhau, thí dụ: do cơ sở hạ tầng vật chất kỹ thuật lạc hậu, năng lực tài chính hạn chế, điều kiện địa lý, khí hậu không thuận lợi hay các

yếu tố chính trị liên quan... Trong khi đó, việc sử dụng các trang thiết bị phân tích hiện đại đã cho phép thu được khối lượng thông tin rất lớn, cho mỗi nghiên cứu riêng biệt. Kết quả là trong hầu hết các trường hợp, bằng các phương tiện thông tin truyền thống (tạp chí, sách, hội nghị, hội thảo khoa học...) nhìn chung không đủ dung lượng và môi trường để truyền tải hết ý tưởng và dữ liệu kết quả nghiên cứu của các tác giả. Đây cũng là một nguyên nhân dẫn tới khả năng thất thoát tài nguyên trực tiếp hay gián tiếp, do lạc hậu về thông tin nên có thể ở nơi này vẫn đang tiêu tốn tiền bạc vào các mục tiêu nghiên cứu đã được giải quyết thành công ở nơi khác. Trong khi đòi hỏi thực tiễn đặt ra cho sự phát triển toàn diện và sâu rộng công nghệ sinh học ngày càng trở nên cấp bách. Như một hệ quả tất yếu để giải quyết các vấn đề trên, các trung tâm dữ liệu công nghệ sinh học đã ra đời và phát triển hết sức nhanh chóng, trên cả hai mặt quy mô và số lượng các đơn vị thành viên.

Về mặt bản chất, sinh học hiện đại đã chỉ rõ rằng: đặc tính riêng biệt của mỗi loài trong sự đa dạng của thế giới sinh học được quyết định chính trong kích thước và cấu trúc gen của từng cá thể, với đơn vị cấu trúc cơ sở là bốn loại nucleotide: Adenine, Guanine, Cytosine và Thymine (Uracil thay thế Thymine trong RNA). Đồng thời, protein (thành phần quan trọng nhất của mọi cơ thể sống) được tạo thành trên cơ sở kết nối của 20 amino axit khác nhau. Logic chính xác trong quy luật của thế giới sống trong môi trường tin học đã cho phép con người "số hoá và ký tự hoá " trong việc mô tả bản chất và sự vận động của thế giới sinh học. Kết hợp với khả năng kết nối trao đổi thông tin "vô hạn " của công nghệ thông tin và internet đã mở ra điều kiện lý tưởng cho các nhà sinh học để cất giữ, liên kết, xử lý và trao đổi kho tàng dữ liệu giữa các thành viên với nhau. Nhờ sự hợp tác và liên kết rộng rãi này, một mặt mở ra khả năng tư vấn, trao đổi và hỗ trợ cho nhà nghiên cứu hay các tổ chức thành viên tham gia. Nhưng mặt khác, chính sự liên kết này đã tạo ra công cụ mới để nghiên cứu sự biến đổi trong các cơ thể sống hay

các hiện tượng sống, trên cơ sở phân tích phát hiện tính quy luật từ vô số các dữ liệu thực nghiệm trong kho tàng dữ liệu khổng lồ này... Nghĩa là, thông qua xử lý hàng loạt mảng dữ liệu thực nghiệm rời rạc, người ta thu được các mảng dữ liệu thứ cấp, để từ đó có thể khái quát hoá thành quy luật vận động và biến đổi của nó; hoặc trên cơ sở xử lý cơ sở dữ liệu đã có để định hướng, hoạch định kế hoạch và tổ chức thực nghiệm khoa học của mình sao cho hiệu quả hơn, hay trên cơ sở nắm bắt được quy luật vận động của tự nhiên để "thiết kế " ra các sản phẩm hoàn toàn mới, thậm chí có thể chưa xuất hiện trong thiên nhiên... Chính từ các cơ sở lý luận và thực tiễn nêu trên, một lĩnh vực khoa học mới đã ra đời, đó chính là tin-sinh học.

Tin-sinh học (*Bioinformatic*) có thể hiểu là một ngành khoa học sinh học phân tích và dự đoán đặc tính của đối tượng sinh học, trên cơ sở tích hợp năng lực hoạt động hữu cơ của ba lĩnh vực khoa học công nghệ là khoa học sinh học, với tri thức về quy luật vận động của thế giới sống; năng lực quản trị và xử lý dữ liệu của computer với năng lực kết nối của công nghệ thông tin (qua mạng internet và hệ thống viễn thông hiện đại) để tổ chức quản lý và khai thác nguồn dữ liệu thông tin sinh học khổng lồ quy mô toàn cầu. Tin – Sinh học đảm nhiệm nhiệm vụ to lớn hỗ trợ cho việc hoạch định các thực nghiệm sinh học; hỗ trợ hiệu quả cho việc phân tích, dự đoán đặc tính của các vật liệu sinh học, cũng như nghiên cứu khám phá bản chất sinh học của giới tự nhiên, hay đảm nhiệm vai trò quan trọng trong việc "thiết kế" và sản xuất ra các sản phẩm sinh học mong muốn khác nhau phục vụ đời sống con người...

Sự ra đời của tin-sinh học không chỉ mở ra khả năng quản lý và khai thác hiệu quả hơn cơ sở dữ liệu sinh học thực nghiệm thu được; mà trong thực tế chính tin-sinh học đã thực sự trở thành công cụ nghiên cứu mới, trợ giúp đắc lực và hiệu quả để đẩy nhanh tốc độ nghiên cứu và ứng dụng công nghệ sinh học; chắp cánh cho công nghệ sinh học nói riêng và sinh học nói

chung, bay lên tầm cao mới. Cơ sở dữ liệu công nghệ sinh học không chỉ dừng lại ở tập hợp các kết quả nghiên cứu thực nghiệm đơn thuần, mà nó còn bao gồm khả năng khái quát hoá, mô phỏng hoá thành những "đối tượng số " của thế giới sinh học sống động. Thí dụ, với công cụ tin-sinh học đã cho phép con người tìm hiểu và khám phá các quá trình vận động nội tại trong bản thân mình, nhờ nghiên cứu dữ liệu thực nghiệm trên các đối tượng sinh vật khác, hay cho phép con người chế tạo ra cả những sinh giới mới vượt ra khỏi quy luật tiến hoá và chọn lọc tự nhiên...

Tin-sinh học có thể khái quát hoá thành ba nhiệm vụ cơ bản là:

- Xây dựng, bổ sung, tổ chức quản lý và khai thác cơ sở dữ liệu đa dạng và toàn diện trên quy mô toàn cầu liên quan đến sinh học và các ngành hay lĩnh vực khoa học liên quan. Vấn đề này đã và sẽ phát huy được lợi thế khổng lồ của nó khi huy động được sự tham gia thực sự của đông đảo các thành viên sở hữu thông tin sinh học trên toàn thế giới.

- Xây dựng và phát triển các chương trình xử lý dữ liệu ứng dụng, dưới dạng các chương trình xử lý dữ liệu độc lập hay được tích hợp ngay trong các thiết bị phân tích hiện đại, nhằm cung cấp cho các nhà sinh học phương tiện xây dựng phương án nghiên cứu hay phân tích xử lý kết quả thu được với sự "tư vấn và trao đổi của các chuyên gia " trên toàn thế giới.

- Đào tạo và cập nhật thường xuyên cho các nhà sinh học kỹ năng tư duy và năng lực khai thác hai nội dung trên vào hoạt động khoa học và công nghệ nhằm tạo ra bước chuyển biến đột phá trong phương cách tiếp cận và nghiên cứu khám phá thế giới sống, tạo ra cuộc cách mạng thực sự trong hoạt động sáng tạo của con người vì phồn vinh và hạnh phúc nhân loại.

2. ĐẠI CƯƠNG VỀ INTERNET

2.1. Khái niệm về internet và địa chỉ trên mạng

Internet là hệ thống gồm rất nhiều mạng máy tính cục bộ hay khu vực được kết nối lại với nhau thành mạng chung trên phạm vi toàn cầu (*Networks of the Networks*). Như vậy, internet kết nối nhiều triệu máy tính riêng lẻ đã hoà mạng vào hệ thống chung, trong đó giữa các máy đã nối mạng đều bình đẳng và có thể liên hệ trao đổi thông tin qua lại với nhau. Trên internet, người truy cập vào mạng từ khắp nơi trên hành tinh, nếu được phép của chủ sở hữu, có thể tìm kiếm và khai thác tất cả mọi thông tin và dữ liệu trong từng máy con với tốc độ "ánh sáng" vượt qua mọi trở ngại về không gian và lãnh thổ.

Điểm khởi đầu của internet là dự án nối mạng các máy tính của bốn đơn vị thành viên là Viện Nghiên cứu Stanford, Trường Đại học Tổng hợp California, Trường Đại học Tổng hợp UC-Santa Barbara và Trường Đại học Tổng hợp Utah do cơ quan quản lý dự án nghiên cứu phát triển của bộ quốc phòng Mỹ (*U.S. Defense Advance Research Projects Agency* – DARPA) tài trợ (tháng 7/1968). Việc kết nối thành công các máy tính tham gia của bốn thành viên trên (năm 1969) đã đánh dấu sự ra đời của mạng máy tính khu vực – viết tắt là ARPANET. Lịch sử phát triển của internet là quá trình phát triển và hoàn thiện không ngừng từ ARPANET, qua MILNET và NSFNET (*National Science Foundation Network*), đến

internet với khả năng khổng lồ và quy mô toàn cầu hiện nay (internet với đầy đủ ý nghĩa và thực sự bùng nổ mạnh mẽ chỉ từ 1995, sau thời điểm chính phủ Mỹ cho phép công khai và thương mại hoá công nghệ này trên phạm vi toàn cầu).

Internet là sự kết nối đa chiều các mạng diện rộng (*Wide Area Network* – WAN) của các quốc gia hay khu vực. Mỗi mạng WAN được hình thành do sự kết nối của nhiều mạng khu vực hẹp hơn (*Local Area Network* – LAN); trong đó, mỗi mạng LAN lại là mạng kết nối các máy tính riêng lẻ (hay mạng của cụm các máy tính riêng lẻ) lại với nhau. Việc kết nối giữa các mạng trên được thực hiện nhờ các cổng chuyển thông tin - thường là các cầu nối (*Bridges*) hoặc các bộ định tuyến (*Router*).

Từng máy tính con thường được kết nối vào internet qua một máy chủ (*Host*). Để các máy tính nối mạng có thể nhận biết và thông tin qua lại với nhau, mỗi máy chủ đều được nhận một miền gồm một số địa chỉ IP (*Identification Protocol*) nhất định và không trùng nhau với các máy chủ khác. Trung tâm thông tin điều phối internet quốc tế (*Network Information Center* – NIC) chủ trì phân phối các địa chỉ mạng (*Net* ID) cho mỗi quốc gia. Tiếp theo, tổ chức quản lý internet từng quốc gia sẽ phân phối miền địa chỉ cho các máy chủ trên mạng đó (*Host* ID). Theo hệ địa chỉ đang được sử dụng hiện tại *IPv4* mỗi địa chỉ mạng gồm bốn cụm số phân cách nhau bằng dấu chấm dạng A.B.C.D, với A, B, C, và D là một số nguyên có giá trị trong dải (0 – 255), thí dụ: `192.168.127.16`; `172.16.1.3` (mạng WAN một vài nước đã sử dụng hệ địa chỉ *IPv6*). Để thuận tiện cho người sử dụng trong giao tiếp, các địa chỉ IP kiểu số trên thường được máy chủ (do các nhà cung cấp dịch vụ internet quản lý) phiên mã thành dạng địa chỉ các cụm từ, thí dụ: http://www.vnn.vn; http://www.hut.edu.vn; http://www.atcc.org; http://merlin.bcm.tmc.edu...

Để truy cập vào mạng, người sử dụng internet (thường được gọi chung là khách hàng) phải đăng ký với các nhà cung cấp dịch vụ và sẽ được cấp một tên truy cập (*Account*) và với mật khẩu riêng tương ứng (*Password*). Với tên và mật khẩu đã đăng ký, thường khách hàng có thể truy cập vào mạng internet từ bất kỳ máy tính nào trong mạng LAN của nhà cung cấp dịch vụ đó hay thông qua kết nối trực tiếp một máy tính ngoài mạng với máy chủ bằng đường điện thoại (sử dụng Modem thường hay Modem ADSL). Việc kết nối giữa một máy tính con với máy chủ còn phụ thuộc vào chế độ kết nối. Có nhiều kiểu kết nối khác nhau, phụ thuộc vào kiểu dữ liệu sử dụng, phần mềm cài đặt trên máy chủ, phần mềm của khách hàng. Các kiểu kết nối này thường mang đặc trưng riêng với từng trường hợp cụ thể (*"service by service"*, *"user by user"*) và thường được xác định qua cổng kết nối (*Port*) đi kèm như một địa chỉ phụ, thí dụ

"`192.168.127.16:8080`" (port 8080); hay

"merlin.bcm.tmc.edu:23" (port 23)...

2.2. Thông tin trên internet

Internet chứa khối lượng thông tin khổng lồ, bao gồm dữ liệu của hầu như tất cả mọi lĩnh vực khác nhau trong đời sống xã hội hiện đại, từ khoa học, kinh tế, văn hoá, chính trị, xã hội đến cả vô số các thông tin quảng cáo sản phẩm hay các thông tin về dịch vụ thương mại điện tử... Các dữ liệu thông tin này được lưu giữ trong các máy chủ của hàng trăm ngàn mạng con (LAN và WAN) và trong các máy tính đang hoà mạng trên khắp thế giới. Khả năng khai thác các dữ liệu thông tin này, đương nhiên còn phụ thuộc vào việc cung cấp của chủ sở hữu và giới hạn khai thác của khách hàng được chủ sở hữu dữ liệu cấp phép. Ở góc độ khai thác, có thể chia cơ sở dữ liệu khổng lồ trên thành hai nhóm lớn là:

* Loại các thông tin công cộng: Bao gồm tất cả các loại dữ liệu thông tin mà bất kỳ khách hàng nào, từ mọi nơi trên khắp thế giới, khi đã vào internet đều có thể tự do truy cập và khai thác phục vụ cho mục đích riêng, điển hình cho kiểu dịch vụ thông tin công cộng là WWW (World Wide Web), thí dụ: http://www.vnn.vn ; http://www.sony.com...

* Loại các thông tin giới hạn truy cập: Bao gồm tất cả các dữ liệu hay các hệ thống dữ liệu trên mạng, nhưng việc truy cập và khai thác chỉ có thể được thực hiện nếu được phép của chủ sở hữu chúng. Thí dụ các thông tin phải trả tiền khi sử dụng, các thông tin chỉ dành cho các đối tượng đã được cấp quyền truy cập, các thông tin chỉ sử dụng nội bộ.... Thông thường, nguồn dữ liệu này được lưu giữ trên mạng nhưng với độ bảo mật rất cao; chỉ có những người đã được cấp phép (với tên và mật khẩu truy cập đã đăng ký) mới có thể truy cập và khai thác.

2.3. Một số dịch vụ trên internet

Các dịch vụ trên mạng rất đa dạng và được cải tiến, hoàn thiện và mở rộng không ngừng. Một số dịch vụ phổ dụng hiện nay của internet là:

- Truy cập khai thác thông tin từ xa (*Telnet*): Được xem là dịch vụ cơ sở và đầu tiên của việc kết nối mạng. Dịch vụ này cho phép từ một máy tính ở bất kỳ vị trí nào trên thế giới có thể truy cập vào một máy tính xác định khác trong mạng thông qua giao thức TCP/IP (*Transfer Control Protocol/Internet Protocol*). Khi dịch vụ đã được thiết lập, người sử dụng dịch vụ có thể thực hiện các thao tác đầy đủ trên máy tính kia cũng như trên máy đang sử dụng, thí dụ: gọi các chương trình hiện có, ghi hay xoá các tệp tin... Trong thực tế, việc khai thác dịch vụ

truy cập từ xa được thực hiện với sự trợ giúp của các chương trình hỗ trợ và giám sát mà các nhà quản lý hệ thống máy chủ phía sở hữu dữ liệu sử dụng. Nghĩa là người muốn truy cập vẫn phải được "cấp phép" dưới dạng được cấp tên đăng ký và mật khẩu riêng (*public login name and password*).

- Dịch vụ trao đổi các tệp dữ liệu (*files transfer* - ***ftp***): Dịch vụ *ftp* cũng là dịch vụ cơ sở đầu tiên của việc kết nối mạng, nhưng được xây dựng dành riêng cho những người sử dụng chỉ trao đổi một hay một số tệp dữ liệu nhất định, song không mong muốn truy cập (hay không được thẩm quyền truy cập) vào toàn bộ ngân hàng dữ liệu của máy chủ đó. Thao tác để sử dụng dịch vụ *ftp* nguyên thuỷ cũng hoạt động trên cơ sở tương tự như sử dụng dịch vụ *telnet*. Khi sử dụng dịch vụ *ftp*, thông thường khách hàng phải thực hiện hàng loạt dòng lệnh khác nhau mới có thể gửi (*put files*) hoặc nhận (*get files*) và phải phân biệt hai dạng dữ liệu là kiểu ký tự (*text mode*) và kiểu nhị phân (*binary mode*). Dịch vụ *ftp* với kiểu ký tự đã lưu ý đến sự khác biệt giữa các hệ điều hành (môi trường Unix sử dụng hệ ASCII 10, môi trường Macintosh sử dụng hệ ASCII 13 và môi trường MSDOS được thiết kế cho sử dụng một trong hai hệ trên, trong đó với kiểu nhị phân sẽ được trao đổi đúng nguyên bản gốc).

Nhằm giảm bớt trục trặc và để thuận tiện hơn cho khách hàng, người cung cấp tin có thể chuẩn bị sẵn các tệp dữ liệu hay một một số thư mục tệp dữ liệu liên quan thành các nhóm riêng, sao cho khi khách hàng cần trao đổi có thể thực hiện được dễ dàng mà không cần phải sử dụng đến mật khẩu. Khi xây dựng các trang WWW (*World Wide Web*) người ta sử dụng phổ biến kỹ thuật này giúp khách hàng đang ở trong trang Web vẫn có thể trao đổi thuận tiện các tệp dữ liệu mong muốn. qua truy cập các đường dẫn siêu liên kết dưới dạng dòng lệnh

"*Download*", "*Download now*" hay đường dẫn "*ftp://...*" (thông thường các tệp dữ liệu dạng này không có sẵn trong các trang WWW), thí dụ:

```
"The file is available by anonymous ftp.
        ftp to ftp.bcm.tmc.edu
        and retrieve mbcr/pub/file.txt"
```

Để trao đổi tệp trên có thể thực hiện nhờ sử dụng lệnh:

ftp://ftp.bcm.tcm.edu/bmcr/pub/file.txt

- Dịch vụ thư điện tử (*E-Mail*): Dịch vụ thư điện tử là dịch vụ đơn giản nhất nhưng lại rất hiệu quả và được nhiều người sử dụng nhất. Dịch vụ này dành cho cả những người không đăng ký quyền truy cập mạng hay thường xuyên được chọn với các khách hàng chỉ đăng ký sử dụng hạn chế các dịch vụ trên internet. Người gửi thư chỉ cần "gọi ra" một khung mẫu thư từ một máy chủ nhất định (các *mailserver*), sau đó sử dụng bàn phím để viết thư, điền địa chỉ điện tử của người nhận và nhấn lệnh gửi đi. Khi đó thư sẽ được chuyển ngay đến máy chủ rồi chuyển tiếp sang máy chủ của người nhận đăng ký địa chỉ và được lưu giữ ở đó. Người nhận thư, vào lúc thời gian thuận tiện, có thể truy cập vào "thùng thư" của mình trên máy chủ để xem các thư gửi đến. Ngày nay, kết hợp với các dịch vụ đi kèm khác, người gửi thư có thể gửi đồng thời một bức thư đến nhiều người nhận khác nhau (dịch vụ C.c. qua *listserver*), có thể chuyển cả "thư" dưới dạng âm thanh, hình ảnh hay tiếng nói đến người nhận và thường kết hợp kèm thêm dịch vụ chuyển tệp đơn giản để mở rộng năng lực phục vụ khách hàng (chế độ *attachment*). Nhìn chung, việc sử dụng dịch vụ thư điện tử rất đơn giản về thao tác, thuận tiện về thời gian và hết sức nhanh chóng. Vì vậy, để thu hút khách hàng truy cập, rất nhiều công ty kinh doanh trên internet thường có thêm *mailserver* phục vụ miễn phí cho mọi đối tượng được tự do đăng ký "thùng thư" cá nhân.

- Dịch vụ thông tin theo nhóm (*usenet*): Dịch vụ này cho phép người sử dụng mạng có thể tham gia "sinh hoạt" theo các nhóm thông tin (*Newsgroup*), trong đó họ có thể gửi hay nhận các thông tin cho các thành viên khác cùng tham gia trong chủ đề này. Các nhóm thông tin được trình bày theo chủ đề, không phân biệt thời gian cập nhật, tách biệt độc lập giữa các nhóm với nhau và độc lập với dịch vụ thư điện tư. Đồng thời, việc đăng ký tham gia vào nhóm tin, xoá tên đã đăng ký, gửi và nhận tin thao tác rất đơn giản và thuận tiện. Do dịch vụ này rất thuận lợi nên từ thời kỳ đầu internet chỉ có 7 nhóm tin (*sci*- khoa học, *soc*-xã hội, *comp*-computer...), song đến nay có thể tới hàng chục ngàn nhóm tin khác nhau trên mạng. Tuy nhiên, do những lý do nhất định, nhiều nhóm tin không tham gia vào hệ thống dịch vụ "*usenet*" chung, mà chúng tồn tại theo nhóm độc lập riêng hay các nhóm chỉ "trao đổi nội bộ" trong diện đối tượng hẹp trên mạng.

- Dịch vụ tìm kiếm thông tin *gopher*, WAIS (*Wide Area Information Server*) và dịch vụ truyền siêu văn bản HTTP (*Hyper Text Transport Protocol*) hoặc WWW *(World Wide Web)*: Với mục đích phối hợp với dịch vụ trao đổi tệp dữ liệu, *gopher* cho phép người sử dụng mạng có thể tìm kiếm và hiển thị thuận tiện các tệp dữ liệu có trên mạng, thường với các tên theo từ khoá và các đường dẫn từ trang *gopher* đến các trang khác. Cũng hoạt động tương tự, dịch vụ WAIS (*Wide Area Information Server*) tìm kiếm theo các cụm dữ liệu dưới dạng ký tự (*free-text databases*). Nhờ vậy, dịch vụ này có công năng rất mạnh để tìm kiếm, thu thập và cung ứng thông tin. Song song với hai dạng trên, phương án liên kết các tệp dữ liệu trong từng máy chủ để tạo ra dạng cung cấp thông tin hiệu quả hơn đã xuất hiện dịch vụ truyền thông tin siêu văn bản HTTP (*Hyper Text Transport Protocol*) và Web (*www*, *W3* hoặc *Web*). Với dịch vụ thông tin mới này, khả năng trình bày, nội

dung hiển thị, đường dẫn đến các cơ sở dữ liệu hay các dạng dịch vụ khác rất đa dạng. Nhờ vậy, đã tạo ra phương án cung cấp thông tin nhanh chóng và hiệu quả, môi trường giao tiếp thân thiện và hết sức thuận lợi cho khách hàng. Với ưu thế to lớn của mình, ngày nay hầu như dịch vụ WWW đã thế chỗ hoàn toàn cho dạng dịch vụ *gopher* và WAIS (các *Web server* đều có khả năng giao tiếp kết nối với các *gopher server* và *ftp server*). Để giao tiếp với các Web server khách hàng thường sử dụng các chương trình trình duyệt Web, trong đó ba chương trình trình duyệt mạnh nhất hiện nay là: *Microsoft Internet Explorer* (của *Microsoft Corp.*), *Netscape Explorer* (của *Netscape Communication Corp.*) và *AOL Browser* (của *American On Line Corp.*).

2.4. Truy cập tìm kiếm dữ liệu thông tin qua internet

Cũng như các lĩnh vực khoa học khác, người ta hầu như không thể hy vọng liệt kê ra được phần lớn các cơ sở dữ liệu liên quan đến công nghệ sinh học, thậm chí sẽ không có một giải pháp tối ưu nhất để tìm kiếm thông tin dù chỉ trong một lĩnh vực hẹp. Giải pháp tương đối đơn giản và thường áp dụng với những người khởi đầu tham gia khai thác thông tin qua internet là:

- Sử dụng các trang công cụ tìm kiếm phổ dụng trên internet như: www.yahoo.com; www.google.com; www.altavista.com; www.webferret.com...

* Vào một cơ sở dữ liệu lớn đã biết gần gũi với chuyên mục cần tìm kiếm. Sau đó sử dụng các đường dẫn siêu liên kết mặc định (các đường "*links*", "*hyperlink*", lệnh "*go*"...) để mở rộng khả năng tìm kiếm sang các cơ sở dữ liệu khác.

Cần chú ý rằng, với mỗi cơ sở dữ liệu đều chứa đựng khối lượng thông tin rất lớn, nguồn tin được cập nhật bổ sung và hoàn thiện liên tục, có thể có những thông tin lại được trình bày dưới các dạng chủ đề khác nhau và có thể tồn tại một vài khác biệt nhất định trong các chương trình xử lý dữ liệu thực nghiệm giữa các tổ chức sở hữu.

Bên cạnh việc tìm kiếm trên, một trong số các giải pháp cập nhật thông tin nhanh và hiệu quả là đăng ký tham gia dịch vụ trao đổi tin theo nhóm theo những chuyên đề hẹp quan tâm (dịch vụ *usenet* hoặc dạng tương tự). Ngoài ra, mỗi cá nhân có thể "sở hữu" kiểu tìm kiếm thông tin hữu hiệu hơn và việc tiếp thu thông tin bạn bè giới thiệu lại... trong nhiều trường hợp lại là cách tiếp cận nhanh chóng và hiệu quả đến nguồn dữ liệu mong muốn.

Bảng 2.1. Địa chỉ một số nhóm tin liên quan đến công nghệ sinh học

(http://www.bioremediationgroup.org/BioLinks/links/news.htm)

Agriculture	news:sci.agriculture
Agroforestry Research	news:bionet.agroforestry
Biology Announcements	news:bionet.announce
Audubon Society	news:alt.org.audubon
Biology (Journals and Publications)	news:bionet.journals.contents
Biology of Grasses	news:bionet.biology.grasses
Biotechnology	news:sci.bio.technology
Botany	news:sci.bio.botany
Chemistry	news:sci.chem
Chemical Engineering	news:sci.engr.chem
Civil Engineering	news:sci.engr.civil
Ecological Research	news:sci.bio.ecology
Energy, Science, & Technology	news:sci.energy
Entomology	news:sci.bio.entomology.misc
Environment and Ecology	news:sci.environment
Fisheries Science	news:sci.bio.fisheries

General Biology & Science	news:bionet.general
General Engineering	news:sci.engr
Hydrology	news:sci.geo.hydrology
Microbiology	news:sci.bio.microbiology
Microbiology (Bionet Newsgroup)	news:bionet.microbiology
Microscopy Techniques	news:sci.techniques.microscopy
Petroleum Geology	news:sci.geo.petroleum
Population Biology	news:bionet.population-bio
Scientific Research	news:sci.research
Toxicology	news:bionet.toxicology
Tropical Biology	news:bionet.biology.tropical
Energy and Renewable Resources	news:alt.energy.renewable
Environmentalist Causes	news:alt.save.the.earth
Technology Topics	news:alt.technology.misc
Symbiosis Discussion and Research	news:bionet.biology.symbiosis
Biosphere and Ecology	news:bit.listserv.biosph-l
Conservation	news:sci.bio.conservation
Meteorology	news:sci.geo.meteorology
Chaotic and other Nonlinear Systems	news:sci.nonlinear
Computational Fluid Dynamics	news:sci.physics.computational.fluid-dynamics
Polymer Science	news:sci.polymers
Systems Science	news:sci.systems
Magnetic Resonance Imaging and Spectroscopy	news:sci.techniques.mag-resonance
Mass Spectrum Techniques	news:sci.techniques.mass-spec
Spectrum Analysis	news:sci.techniques.spectroscopy
Miscellaneous Research Testing Techniques	news:sci.techniques.testing.misc
Nondestructive Testing Techniques	news:sci.techniques.testing.nondestructive
Crystallography	news:sci.techniques.xtallography
Environment	news:talk.environment
Waste Management	news:sci.environment.waste
Plant Science	news:bionet.plants

3. CƠ SỞ DỮ LIỆU CÔNG NGHỆ SINH HỌC

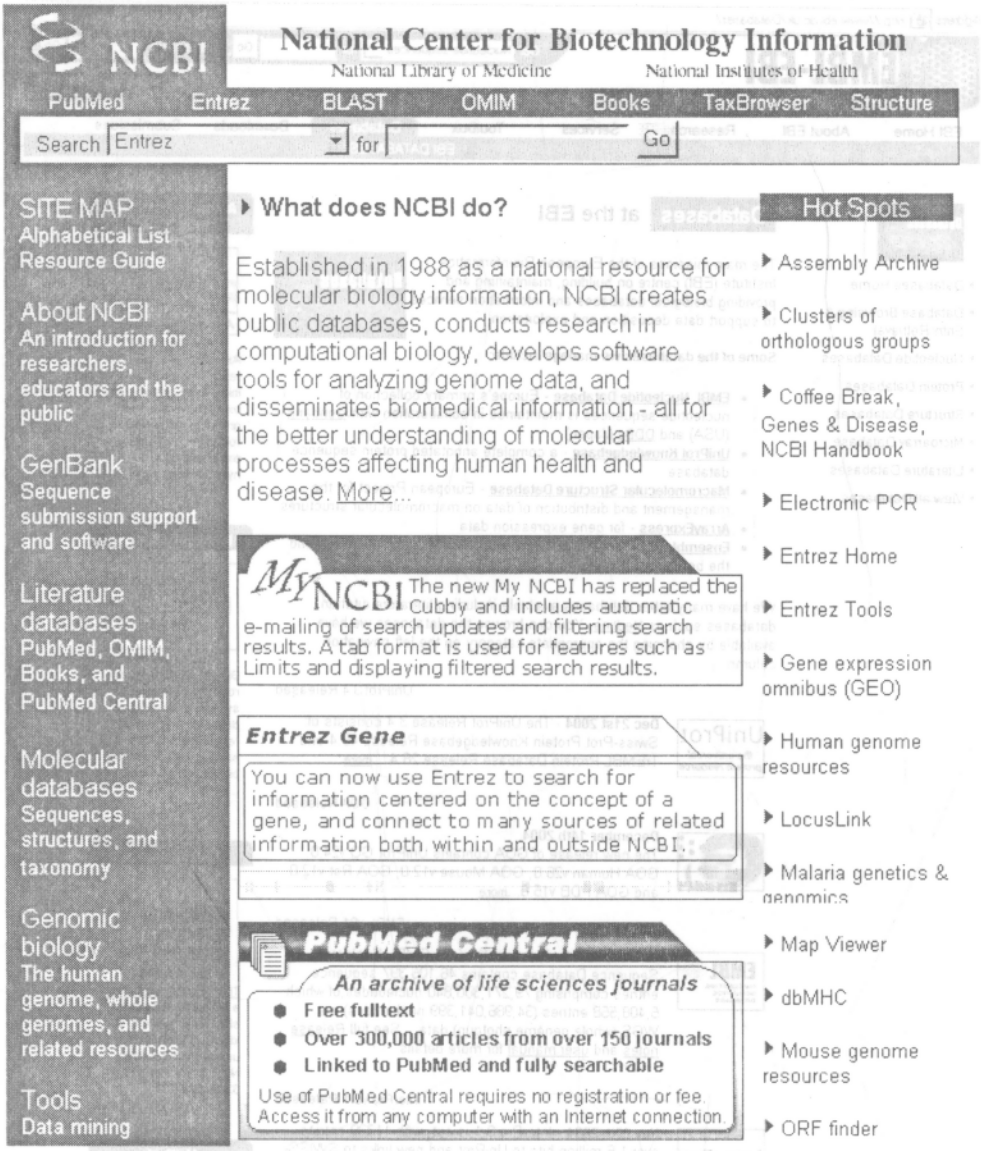
3.1. Đại cương

Công nghệ sinh học là một lĩnh vực khoa học trẻ, đa ngành, phát triển rất năng động và hết sức mạnh mẽ trong nửa cuối thế kỷ XX. Nếu như công nghệ thông tin và internet được xem là công nghệ của thế kỷ XX, thì rất nhiều ý kiến dự báo đều cho rằng công nghệ sinh học sẽ trở thành công nghệ phát triển mạnh mẽ và năng động nhất của thế kỷ XXI. Rất nhiều quốc gia trên thế giới đã xác định công nghệ sinh học là một lĩnh vực khoa học công nghệ trọng điểm trong chiến lược phát triển đất nước. Nhờ vậy, trong thời gian qua công nghệ sinh học đã nhận được sự đầu tư đáng kể của các chính phủ, đã huy động được tiềm lực khoa học và công nghệ không chỉ các cơ quan chuyên sâu, hoạt động trực tiếp trong lĩnh vực của mình, mà còn mở rộng sang cả nhiều công ty vốn không có truyền thống hoạt động về công nghệ sinh học.

Về tiềm lực khoa học và công nghệ sinh học, các cường quốc công nghiệp hàng đầu, do ưu tiên tập trung đầu tư từ rất sớm nên công nghệ sinh học của các quốc gia này phát triển hết sức mạnh mẽ, vượt trội toàn diện, triệt để và bỏ rất xa các quốc gia đang phát triển. Như một hệ quả tất yếu, năng lực lưu trữ, xử lý và khai thác cơ sở dữ liệu nói chung, và dữ liệu về công nghệ sinh học nói riêng, cũng tập trung cao độ trong các ngân hàng dữ liệu thuộc ba trung tâm khoa học và công nghệ hàng đầu thế giới là: Mỹ,

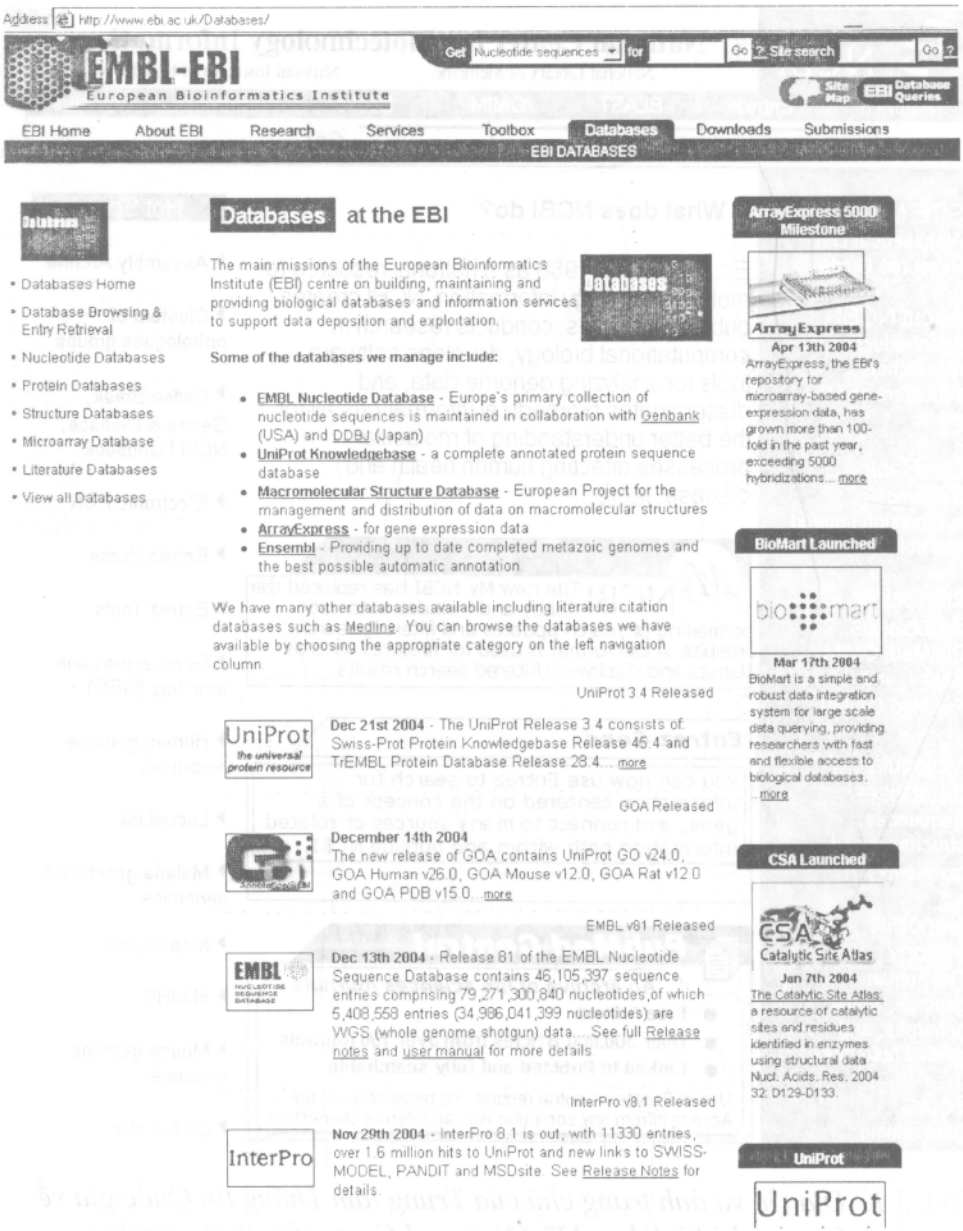
Cộng đồng Châu Âu và Nhật Bản. Một số quốc gia đang phát triển, nhờ chiến lược đầu tư trọng điểm nên cũng đã thu được một số thành công nhất định trong từng lĩnh vực (thí dụ thành tựu về lúa lai của Trung Quốc hay thành tựu về công nghệ sinh học trong sản xuất thuốc điều trị của Cuba...).

Tuy nhiên, trong kỷ nguyên công nghệ và hội nhập quốc tế hiện nay, để đẩy nhanh tốc độ phát triển công nghệ sinh học thì mỗi quốc gia, dù ở bất cứ trình độ công nghệ nào cũng phải xem hợp tác quốc tế là một thực tế tất yếu của thời đại. Hơn nữa, ưu thế về đa dạng sinh học lại tập trung cao ở vành đai xanh nhiệt đới, chứ không phải thuộc các nước công nghiệp phát triển. Nghĩa là, trong lĩnh vực công nghệ sinh học, mọi quốc gia trên thế giới đều rất cần sự "cộng tác và hỗ trợ" từ các quốc gia khác. Cũng nhờ đặc điểm này nên ngay các ngân hàng dữ liệu lớn của các quốc gia công nghiệp hàng đầu cũng rất "hào phóng" trong việc tiếp nhận thông tin mới và cung cấp những "trợ giúp cần thiết" cho các nhà khoa học sinh học trên toàn thế giới, thông qua dịch vụ internet. Thực tế này, đã tạo ra cơ hội thuận lợi cho các nhà khoa học và công nghệ ở nước đang phát triển trong việc tiếp thu thành tựu khoa học và công nghệ mới phục vụ cho mục tiêu nghiên cứu của mình. Trên nền tảng công nghệ thông tin và internet, cơ sở dữ liệu công nghệ sinh học và hợp tác trao đổi thông tin đã thực sự liên thông và liên kết quy mô toàn cầu. Từ hầu hết các cơ sở dữ liệu đều có thể tìm thấy các đường dẫn siêu liên kết đến các cơ sở dữ liệu khác. Đồng thời, các trung tâm dữ liệu lớn như NCBI, EBI, WFCC-MIRCEN và ExPASy thực hiện chế độ trao đổi dữ liệu và cập nhật thông tin trong ngày. Sau đây, cuốn sách cung cấp cho bạn đọc một vài địa chỉ của các ngân hàng dữ liệu lớn trên thế giới để tham khảo.



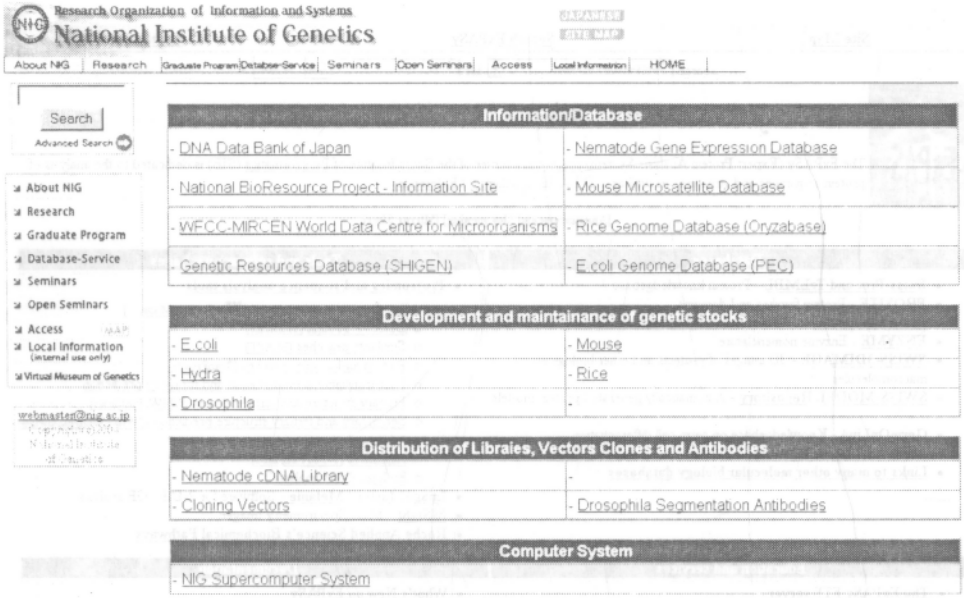
Hình 3.1. Địa chỉ và ảnh trang chủ của Trung tâm Thông tin Quốc gia về Công nghệ Sinh học Mỹ (National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, USA) – (http://www.ncbi.nlm.nih.gov/)

(Chú ý: giao diện trên có thể thay đổi, phụ thuộc vào thời điểm truy cập)



Hình 3.2. Địa chỉ và ảnh trang chủ của cơ sở dữ liệu thuộc Viện Tin-Sinh học Châu Âu (European Bioinformatics Institute, England) - (www.ebi.ac.uk/databases)

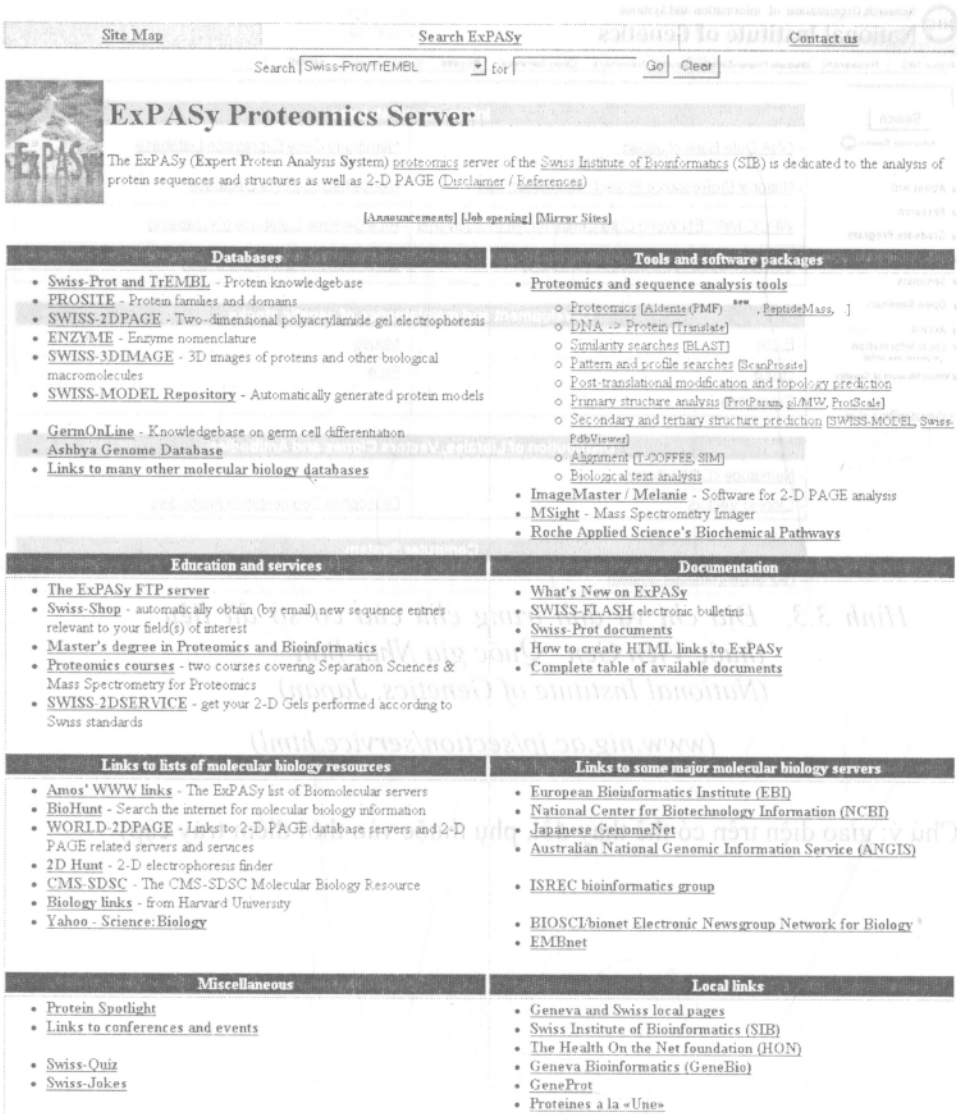
(Chú ý: giao diện trên có thể thay đổi, phụ thuộc vào thời điểm truy cập)



Hình 3.3. Địa chỉ và ảnh trang chủ của cơ sở dữ liệu thuộc Viện Gen Quốc gia Nhật Bản (National Institute of Genetics, Japan)

(www.nig.ac.jp/section/service.html)

(Chú ý: giao diện trên có thể thay đổi, phụ thuộc vào thời điểm truy cập)



Site Map | Search ExPASy | Contact us

Search Swiss-Prot/TrEMBL for Go Clear

ExPASy Proteomics Server

The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References)

[Announcements] [Job opening] [Mirror Sites]

Databases

- Swiss-Prot and TrEMBL - Protein knowledgebase
- PROSITE - Protein families and domains
- SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis
- ENZYME - Enzyme nomenclature
- SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules
- SWISS-MODEL Repository - Automatically generated protein models
- GermOnLine - Knowledgebase on germ cell differentiation
- Ashbya Genome Database
- Links to many other molecular biology databases

Tools and software packages

- Proteomics and sequence analysis tools
  - Proteomics [Aldente (PMF) ***, PeptideMass,]
  - DNA -> Protein [Translate]
  - Similarity searches [BLAST]
  - Pattern and profile searches [ScanProsite]
  - Post-translational modification and topology prediction
  - Primary structure analysis [ProtParam, pI/MW, ProtScale]
  - Secondary and tertiary structure prediction [SWISS-MODEL, Swiss-PdbViewer]
  - Alignment [T-COFFEE, SIM]
  - Biological text analysis
- ImageMaster / Melanie - Software for 2-D PAGE analysis
- MSight - Mass Spectrometry Imager
- Roche Applied Science's Biochemical Pathways

Education and services

- The ExPASy FTP server
- Swiss-Shop - automatically obtain (by email) new sequence entries relevant to your field(s) of interest
- Master's degree in Proteomics and Bioinformatics
- Proteomics courses - two courses covering Separation Sciences & Mass Spectrometry for Proteomics
- SWISS-2DSERVICE - get your 2-D Gels performed according to Swiss standards

Documentation

- What's New on ExPASy
- SWISS-FLASH electronic bulletins
- Swiss-Prot documents
- How to create HTML links to ExPASy
- Complete table of available documents

Links to lists of molecular biology resources

- Amos' WWW links - The ExPASy list of Biomolecular servers
- BioHunt - Search the internet for molecular biology information
- WORLD-2DPAGE - Links to 2-D PAGE database servers and 2-D PAGE related servers and services
- 2D Hunt - 2-D electrophoresis finder
- CMS-SDSC - The CMS-SDSC Molecular Biology Resource
- Biology links - from Harvard University
- Yahoo - Science:Biology

Links to some major molecular biology servers

- European Bioinformatics Institute (EBI)
- National Center for Biotechnology Information (NCBI)
- Japanese GenomeNet
- Australian National Genomic Information Service (ANGIS)
- ISREC bioinformatics group
- BIOSCI/bionet Electronic Newsgroup Network for Biology
- EMBnet

Miscellaneous

- Protein Spotlight
- Links to conferences and events
- Swiss-Quiz
- Swiss-Jokes

Local links

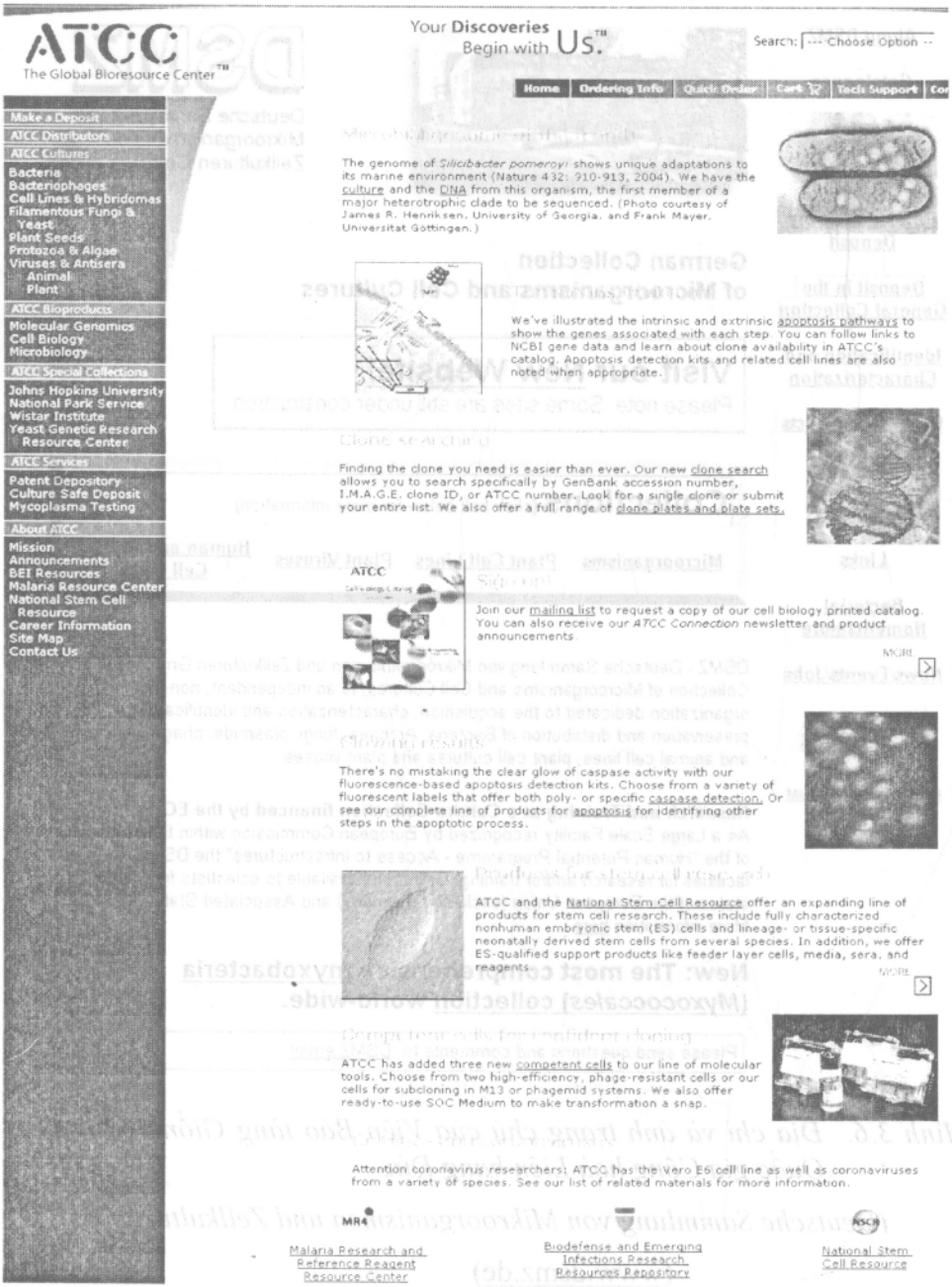
- Geneva and Swiss local pages
- Swiss Institute of Bioinformatics (SIB)
- The Health On the Net foundation (HON)
- Geneva Bioinformatics (GeneBio)
- GeneProt
- Proteines a la «Une»

Hình 3.4. Địa chỉ và ảnh trang chủ của cơ sở dữ liệu về hệ thống nghiên cứu phân tích cấu trúc protein của Thuỵ Sĩ

(ExPASy Proteomics Server, Swiss Institute of Bioinformatics)

(www.expasy.org)

(Chú ý: giao diện trên có thể thay đổi, phụ thuộc vào thời điểm truy cập)



Hình 3.5. Địa chỉ và ảnh trang chủ của Viện Bảo tàng Giống Quốc gia Mỹ (American Type Culture Collection - www.atcc.org)

(Chú ý: giao diện trên có thể thay đổi, phụ thuộc vào thời điểm truy cập)

Hình 3.6. Địa chỉ và ảnh trang chủ của Viện Bảo tàng Giống Quốc gia Cộng hoà Liên bang Đức

(Deutsche Sammlung von Mikroorganismen und Zellkulturen)

(www.dsmz.de)

(Chú ý: giao diện trên có thể thay đổi, phụ thuộc vào thời điểm truy cập)

3.2. Đặc điểm của dữ liệu công nghệ sinh học

Nguồn cơ sở dữ liệu liên quan đến sinh học được truyền tải trên mạng vô cùng đa dạng, phong phú về chủng loại và đồ sộ về khối lượng, với tốc độ gia tăng mạnh mẽ theo thời gian. Về nội dung, cơ sở dữ liệu trải rộng trên tất cả các mặt khác nhau, từ các thông tin chung về tiềm lực khoa học và công nghệ của các cơ quan, đến các thông tin về các công trình khoa học đã công bố, các tạp chí chuyên ngành... Trong đó chiếm khối lượng lớn và đa dạng nhất là các kết quả nghiên cứu trên đối tượng sinh học. Đặc điểm chung nhất của các dữ liệu này là được biểu diễn dưới dạng số hay ký tự trong các tệp dữ liệu đơn lẻ hay dưới dạng các chương trình thuật toán hoàn chỉnh rất thuận tiện để cất giữ hay trao đổi. Về đặc điểm cấu trúc, nguồn thông tin này có thể phân chia sơ bộ thành hai mảng lớn là mảng dữ liệu sơ cấp và mảng dữ liệu thứ cấp:

- Mảng dữ liệu sơ cấp bao gồm tất cả các dữ liệu thu được qua phân tích trực tiếp, bằng các trang thiết bị tương ứng, thí dụ cơ sở dữ liệu thực nghiệm phân tích cấu trúc protein, cấu trúc chuỗi amino axit, cấu trúc và đặc tính enzym, về các hợp chất hữu cơ khác (hydratcarbon, vitamin, lipid...) hay các đặc tính phân loại sinh học, thông tin về đa dạng sinh học, về các đường hướng trao đổi chất trong cơ thể sống...

- Mảng dữ liệu thứ cấp bao gồm các dữ liệu và thông tin thu được trên cơ sở phân tích, khái quát hoá, hệ thống hoá hay thông tin mô phỏng cho từng đối tượng hay nhóm đối tượng sinh học trong thế giới tự nhiên. Mảng dữ liệu này được hình thành thông qua việc xử lý hàng loạt mảng dữ liệu thực nghiệm rời rạc, để từ đó có thể khái quát hoá thành quy luật biến đổi của nó hay mảng dữ liệu hình thành khi xử lý các kết quả nghiên cứu cụ thể, trên cơ sở các quy luật đã phát hiện được qua khai thác cơ sở dữ liệu công nghệ sinh học. Mảng dữ liệu này bao gồm cả

mảng thông tin mà qua đó nhà sinh học có thể khai thác phục vụ cho việc định hướng, hoạch định kế hoạch và tổ chức thực nghiệm khoa học tiếp theo sao cho hiệu quả hơn. Hoặc trên cơ sở phát hiện nắm bắt được quy luật vận động của tự nhiên kết hợp với nền tảng logic chính xác của thế giới sống, nhà sinh học có thể xây dựng ý tưởng, mô phỏng "thiết kế" ra các sản phẩm hoàn toàn mới, thậm chí có thể chưa xuất hiện trong thiên nhiên... Để xử lý phân tích cơ sở dữ liệu trên, đương nhiên không thể xem nhẹ vai trò của các chương trình hay các thuật toán xử lý dữ liệu sinh học ứng dụng. Các chương trình này được thiết kế độc lập hay, từng phần hoặc toàn bộ, dưới dạng tích hợp ngay trong các thiết bị phân tích hiện đại. Chính các yếu tố này cũng là mảng dữ liệu hết sức quan trọng, góp phần tạo ra ưu thế ứng dụng to lớn của tin-sinh học.

3.3. Một số cơ sở dữ liệu sinh học lớn trên thế giới

Cơ sở dữ liệu sinh học là cả một kho tàng dữ liệu khổng lồ, được lưu giữ trong hệ thống rộng lớn các cơ sở dữ liệu thành viên (hay độc lập), dưới nhiều hình thức và định dạng khác nhau, trong đó chiếm khối lượng lớn và nội dung phong phú nhất là mảng dữ liệu sinh học phân tử và công nghệ sinh học. Quy mô và cấu trúc của từng cơ sở dữ liệu có những đặc điểm riêng, song nhìn chung có thể phân chia theo nội dung thành một số mảng dữ liệu chính lớn sau:

- **Dữ liệu về thông tin thông thường** (sách, tạp chí, tài liệu thông tin... dạng số hoá), thí dụ: cơ sở dữ liệu về các công trình khoa học đã công bố PUBMED (http://www.ncbi.nlm.nih.gov/PubMed/), cơ sở dữ liệu tập trung về mảng y - dược (http://www.embase.com), cơ sở dữ liệu về mảng nông nghiệp (http://www.nalusda.gov/general_info/agricola/agricola.html), cơ sở dữ liệu tập trung về mảng thông tin về cổ sinh học

và động vật hoang dã (http://www.biosis/org), cơ sở dữ liệu tập trung về mảng bệnh học trong nông nghiệp (http://www.cabi.org) ...

- **Dữ liệu về phân loại học**, thí dụ: cơ sở dữ liệu về phân loại sinh học của NCBI (http://www.ncbi.nlm.nih.gov/taxonomy/), cơ sở dữ liệu về hệ thống thông tin phân loại các giới (http://www.itis.usda.gov/itis/), cơ sở dữ liệu của tổ chức quốc tế về các thông tin chung về thực vật (http://www.iopi.csu.edu.au/iopi/) ... (mảng dữ liệu này rất phong phú về chủng loại, song trong chừng mực nhất định vẫn bị ràng buộc do sự khác biệt tương đối còn tồn tại giữa một vài hệ thống phân loại).

- **Dữ liệu về cấu trúc và đặc tính của nucleotide và genom**: Đây là một trong hai mảng lớn nhất, đa dạng và phong phú nhất trong kho tàng dữ liệu công nghệ sinh học. Về dữ liệu cấu trúc chuỗi nucleotide, trước hết phải kể đến cơ sở dữ liệu hợp tác liên kết chung giữa EBI, NCBI và DDBJ (khi cần khai thác có thể truy cập vào một trong ba địa chỉ: http://www.ncbi.nlm.nih.gov/Genbank/index.html, http://www.ebi.ac.uk/embl/databases/, hay http://www.ddbj.nig.ac.jp.

  Về dữ liệu genom có thể thí dụ một vài cơ sở dữ liệu lớn như: cơ sở dữ liệu về gen người (OMIM: http://www3.ncbi.nlm.nih.gov/Omim/ và GDB: http://www.gdb.org), cơ sở dữ liệu về vi khuẩn *E. coli* (http://cgsc.biology.yale.edu/top.html và http://www.susi.bio.uni-giessen.de/ecdc/ecdc.html), cơ sở dữ liệu về nấm men (http://www.mips.biochem.mpg.de/proj/yeast/ và http://genome-www.stanford.edu/Saccharomyces/)

- Dữ liệu về cấu trúc và đặc tính chuỗi amino axit và protein được xem là một trong hai mảng dữ liệu lớn nhất về công nghệ sinh học. Trong nhóm này phải kể đến các cơ sở dữ liệu lớn như: Protein Information Resources PIR (http://www.nbrf.georgetown.edu), SWISS-PROT (http://www.expasy.ch hay http://www.ebi.ac.uk/swissprot/) TrEMBL (http://www.ebi.ac.uk/trEMBL/),

PROSITE (http://www.expasy.ch/prosite/) , PRINTS (http://www.bioinf.man.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html).. cơ sở dữ liệu proteomic trong (http://www.genom.ad.jp/kegg/, http://wit.mcs.anl.gov/WIT2/, http://www.ncbi.nlm.nih.gov/COG) ...

Dữ liệu về enzyme và các đường hướng trao đổi chất, thí dụ ENZYME Databases (http://www.expasy.ch/enzyme/), về đặc tính enzyme BRENDA (http://www.brenda.uni-koeln.de/brenda/), về enzyme và phản ứng enzyme (http://www.genome.ad.jp/dbget/ligand.html) ...

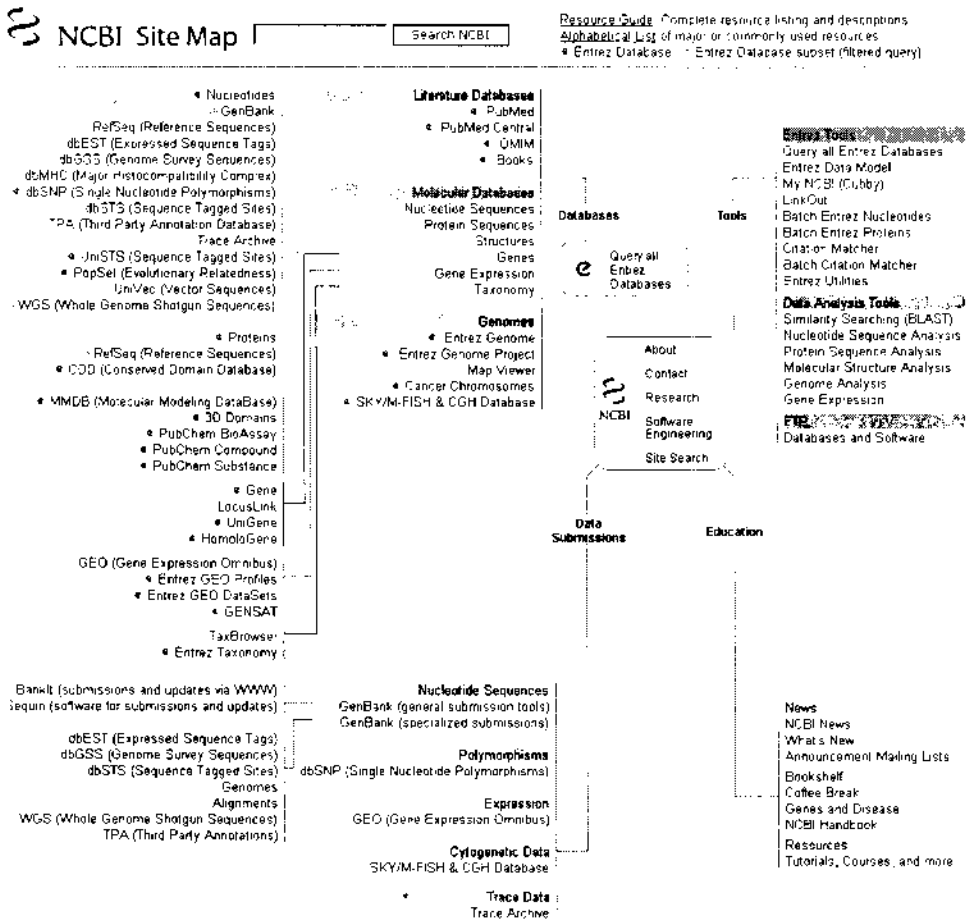
Mỗi cơ sở dữ liệu có thể định hướng tập trung vào những mảng thông tin riêng. Song tất cả mọi cơ sở dữ liệu đều được xây dựng với tiêu chí đảm bảo dễ dàng truy cập, quản lý, và khai thác cho người khai thác dữ liệu, nhằm hỗ trợ giúp họ dễ dàng tìm kiếm được thông tin mong muốn. Để thoả mãn yêu cầu trên, nhìn chung tất cả các cơ sở dữ liệu đều cung cấp cho khách hàng các chương trình tìm kiếm và kết nối liên thông dữ liệu rất hiệu quả, thí dụ Entrez trong NCBI, SRS trong EBI hay SRS trong DDBJ...

3.3.1. Cơ sở dữ liệu Trung tâm Thông tin Quốc gia về Công nghệ Sinh học Mỹ

Cơ sở dữ liệu Trung tâm Thông tin Quốc gia về Công nghệ Sinh học Mỹ (*National Centre for Biotechnology Informatic* - NCBI) được thành lập năm 1988. Đây là một trong số các cơ sở dữ liệu sinh học lớn nhất thế giới hiện nay. Cơ sở NCBI quản lý nguồn thông tin sinh học khổng lồ, với khoảng 25.10^6 nhóm dữ liệu khác nhau, bao gồm từ thông tin về các công trình đã công bố, đến dữ liệu về cấu trúc chuỗi DNA, cấu trúc chuỗi amino axit, cấu trúc gen các loài, cấu trúc không gian ba chiều của các cơ chất khác nhau... Nguồn thông tin dữ liệu trong ngân hàng được tổ chức và quản lý theo từng nhóm tin, với sự liên thông kết nối chặt chẽ giữa các nhóm với

nhau (hình 3.7). Khi truy cập vào ngân hàng, sử dụng công cụ tìm kiếm dữ liệu *Entrez*, người khai thác tin có thể dễ dàng truy cập khai thác các nhóm tin trong cơ sở dữ liệu của NCBI với các đường dẫn siêu liên kết để kết nối liên thông rất thuận tiện và hiệu quả. Sau đây là một số mảng dữ liệu lớn của trung tâm dữ liệu này:

- **PubMed**: NCBI là một trong số ít các địa chỉ tin cậy cho các nhà khoa học công bố kết quả nghiên cứu của mình. Mỗi công trình công bố này được định dạng phân loại bằng một giá trị số (*MEDLINE Unique Identifier* - MUID). NCBI sử dụng mã số này làm mã hiệu cơ sở để cung cấp hàng loạt dịch vụ thông tin kèm theo, thí dụ: thông tin về tác giả, điểm tóm tắt toàn bộ công trình, tóm tắt nội dung chính, đường dẫn đến các công bố khác có liên quan... Do nhu cầu công bố kết quả nghiên cứu nói chung, và khối lượng công trình công bố trong MEDLINE nói riêng, ngày càng tăng nên NCBI đã cung cấp loại hình dịch vụ mới *PubMed*. Dịch vụ PubMed sẽ cung cấp cho người khai thác tất cả các công trình khoa học đã công bố trong MEDLINE và các công trình liên quan của cùng tác giả hay các công trình của tác giả khác có cùng chủ đề. Thời gian gần đây, NCBI còn đưa ra dịch vụ *PubMed Central*, để cung cấp thêm cho người truy cập cả những công trình khoa học đã nằm trong kế hoạch sắp phát hành (do các nhà xuất bản cung cấp để giới thiệu trước, dưới dạng thông tin tóm tắt gửi cho PubMed).
- **GenBank**: Là mảng cơ sở dữ liệu về cấu trúc chuỗi DNA và chuỗi amino axit, với đơn vị cơ sở là các tệp dữ liệu của từng mạch đơn, kèm theo thông tin mô tả về đặc tính của chúng. Các tệp dữ liệu này được tổ chức theo nhóm (*Division*), các nhóm được tổ chức theo cấu trúc phân



Hình 3.7. Sơ đồ cấu trúc cơ sở dữ liệu NCBI

loại loài. Tất cả các thông tin liên quan đến chuỗi đều do chính tác giả cung cấp. Cơ sở dữ liệu GenBank đồng thời là sản phẩm hợp tác quốc tế giữa ba trung tâm dữ liệu gen lớn nhất thế giới là: GenBank of NCBI (*USA*), DNA Data Bank of Japan (*DDBJ, Mishima, Japan*) và European Molecular Biology Laboratory nucleotide database (*EMBL, at EBI, Hinxton, England*). Ba cơ sở này thực hiện chế độ kết nối trực tiếp và trao đổi cập nhật thông tin hàng ngày, nên thực chất cả ba cơ sở đều sở hữu tất cả khối lượng thông tin của hai cơ sở kia, và ngược lại, để trở

thành bộ ba cơ sở dữ liệu gen tập trung và lớn nhất thế giới. Về bản chất cấu trúc, cơ sở dữ liệu này gồm hai mảng lớn riêng biệt là: mảng dữ liệu về protein và mảng dữ liệu về nucleotide, trong đó cơ sở dữ liệu về nucleotide được sử dụng làm đường dẫn để truy cập sang cả dữ liệu tương ứng về protein (chú ý rằng việc thay đổi, sửa chữa hay bổ sung thêm thông tin vào từng tệp chỉ có thể thực hiện được tại cơ sở dữ liệu đăng ký đầu tiên).

- **Entrez System**: Thông thường, mỗi tệp dữ liệu đều truyền tải hàng loạt thông tin khác nhau, trên cơ sở tổ chức theo nhóm, từng thông tin này được sắp xếp tại các thư mục thích hợp trong kho tàng cơ sở dữ liệu của NCBI. Dịch vụ Entrez ra đời nhằm kết nối liên thông giữa các mảng dữ liệu này, giúp cho người truy cập tiếp cận nhanh và đầy đủ các thông tin tìm kiếm. Như vậy, tự Entrez không phải là một cơ sở dữ liệu, mà là một dịch vụ và khi sử dụng dịch vụ này người khai thác có thể dễ dàng tiếp cận các thông tin liên quan từ nhiều mảng dữ liệu khác nhau, thí dụ: dữ liệu truyền thống từ PubMed, cấu trúc và các thông tin liên quan của chuỗi xoắn kép DNA và chuỗi nucleotide, cấu trúc không gian ba chiều của chuỗi protein... Dịch vụ Entrez bao gồm nhiều mảng dịch vụ nhỏ như: Neighboring (tìm kiếm thông tin có nội dung gần gũi nhau), BLAST (Basic Local Alignment Search Tool), VAST (Vector Alignment Search Tool), Hard Links...

3.3.2. Cơ sở dữ liệu EMBL

Phòng thí nghiệm Sinh học Phân tử Châu Âu *(European Molecular Biology Laboratory* - EMBL, 1974) là hệ thống liên kết các phòng thí nghiệm sinh học của 17 nước Tây Âu và Israel, trong đó tập trung vào năm trung tâm nghiên cứu lớn ở Heidelberg và Hamburg (CHLB Đức),

Grenoble (Pháp), Hinxton (Anh) và Monterotondo (Italia). Với mục tiêu xây dựng, lưu giữ, xử lý cơ sở dữ liệu và cung cấp các dịch vụ thông tin liên quan đến sinh học phân tử và tin-sinh học, Viện Tin-Sinh học Châu Âu (*Europian Bioinformatics Institute*, trực thuộc EMBL) được thành lập chính thức vào năm 1994. Qua quá trình xây dựng và phát triển cơ sở dữ liệu của EBI (EBI Databases) hiện đã trở thành một trong ba ngân hàng dữ liệu sinh học lớn nhất trên thế giới.

Cơ sở dữ liệu này được tổ chức và quản lý theo khoảng tám mươi mảng khác nhau, trong đó lớn nhất tập trung vào các mảng: EMBL Nucleotide Sequence Databases, TrEMBL and SWISS-PROT protein sequence databases, Macromolecular Structure Database (EBI-MSD) of 3D co-ordinates of biological macromolecules và RHdb database of radiation hybrid maps. Đồng thời, EBI còn cung cấp hầu hết các chương trình phân tích và xử lý thông tin sinh học như: FASTA (Smith và Waterman, 1981), BLAST (Altschul và đồng nghiệp, 1990), CLUSTALW (Thompson và đồng nghiệp, 1994) and Smith & Waterman (Smith và Waterman, 1981), DALI (Holm và Sander, 1997) ... Việc quản lý, tìm kiếm và khai thác cơ sở dữ liệu khổng lồ này được thực hiện dễ dàng qua chương trình SRS (Sequence Retrieval System). Sau đây điểm một vài thông tin chính về ba cơ sở dữ liệu lớn của EBI:

- **Mảng dữ liệu cấu trúc DNA** (EMBL Nucleotide Sequence Database, gọi tắt là EMBL - thành lập năm 1998) hiện đang lưu giữ thông tin về cấu trúc và đặc tính liên quan của khoảng trên hai triệu đoạn chuỗi DNA (với khoảng 2,3 tỉ cặp nucleotide). Đồng thời, như phần trên đã trình bày, EMBL kết nối liên thông chặt chẽ với hai trung tâm dữ liệu DNA lớn khác trên thế giới là GenBank (Mỹ) và DDBJ (Nhật Bản)...

- **Mảng dữ liệu cấu trúc Protein** (SWISS-PROT và TrEMBL protein sequence database): SWISS-PROT ra đời năm 1986 tại Trường Đại học Tổng hợp Giơnevơ (Thuỵ Sĩ) là một thành viên hợp tác thường xuyên với EBI (từ 1987). Đây là một sơ sở dữ liệu lớn về cấu trúc chuỗi protein và các đặc tính của chúng, cùng với các chương trình xử lý, mô phỏng cấu trúc và đặc tính phân tử protein. Do nhu cầu cung cấp và xử lý thông tin liên quan đến mảng này rất lớn nên, sau đó, EBI đã thiết lập thêm cơ sở dữ liệu TrEMBL, cùng tồn tại song song và kết nối chặt chẽ với SWISS-PROT. TrEMBL cho phép tự động hoàn toàn các dịch vụ lưu giữ, bảo quản và phân tích xử lý thông tin, đảm bảo cung cấp dịch vụ khai thác trực tuyến 24/24 giờ cho người truy cập.

- **Mảng dữ liệu cấu trúc các chất phân tử lượng lớn** (Macromolecular Structure Database - EBI-MSD), là cơ sở dữ liệu liên quan đến các hợp chất sinh học có phân tử lượng lớn. EBI-MSD chính là sản phẩm của dự án "Macromolecular Structure Database Project" của EBI nhằm hợp tác cùng khai thác thông tin chung với US-RCSB *(Research Collaboratory for Structural Bioinformatics*, USA, nơi quản lý cơ sở dữ liệu lớn về protein - Protein Data Bank -PDB).

3.3.3. Cơ sở dữ liệu CIB - DDBJ

Cơ sở dữ liệu CIB - DDBJ (*Center for Information Biology and DNA Data Bank of Japan*) là cơ sở dữ liệu đặt dưới sự quản lý của Trung tâm Thông tin Sinh học, Viện Di truyền Quốc gia Nhật Bản (*Japan National Institute of Genetics*). CIB-DDBJ là cơ sở dữ liệu công nghệ sinh học quan trọng và là cơ sở dữ liệu DNA duy nhất ở Nhật Bản. Cơ sở dữ liệu này được xây dựng trước hết nhằm phục vụ cho hoạt động khoa học của các nhà sinh học Nhật Bản. Tuy nhiên, do hợp tác và liên kết thông tin với hai

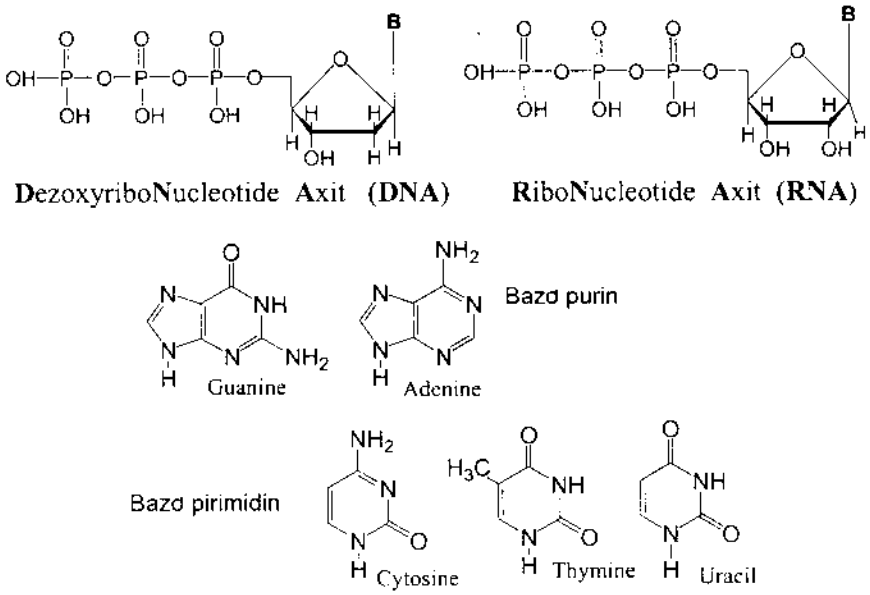
trung tâm dữ liệu hàng đầu thế giới NCBI và EBI, nên CIB-DDBJ đã trở thành là một trong ba trung tâm dữ liệu lớn nhất thế giới hiện nay. Cơ sở dữ liệu này cung cấp trực tuyến cho người sử dụng rất nhiều nhóm thông tin khác nhau, bao gồm cả thông tin thường hay truy cập và khai thác hay các chương trình xử lý thông tin, thí dụ: SRS, gententry, FASTA BLAST, S&W, Search SQmatch XML, TXSearch GIB, ClustalW, GTOP LIBRA...

Bên cạnh CIB-DDBJ, Viện Di truyền Quốc gia Nhật Bản còn quản lý nhiều mảng dữ liệu khác như: WFCC-MIRCEN (*World Data Centre for Microorganisms*, www.wdcm.nig.ac.jp), Genetic Resources Databaces SHIGEN (SHared Inform. of GENetic resources, www.shigen.nig.ac.jp) ...

4. NGHIÊN CỨU CẤU TRÚC CHUỖI DNA VÀ AMINO AXIT

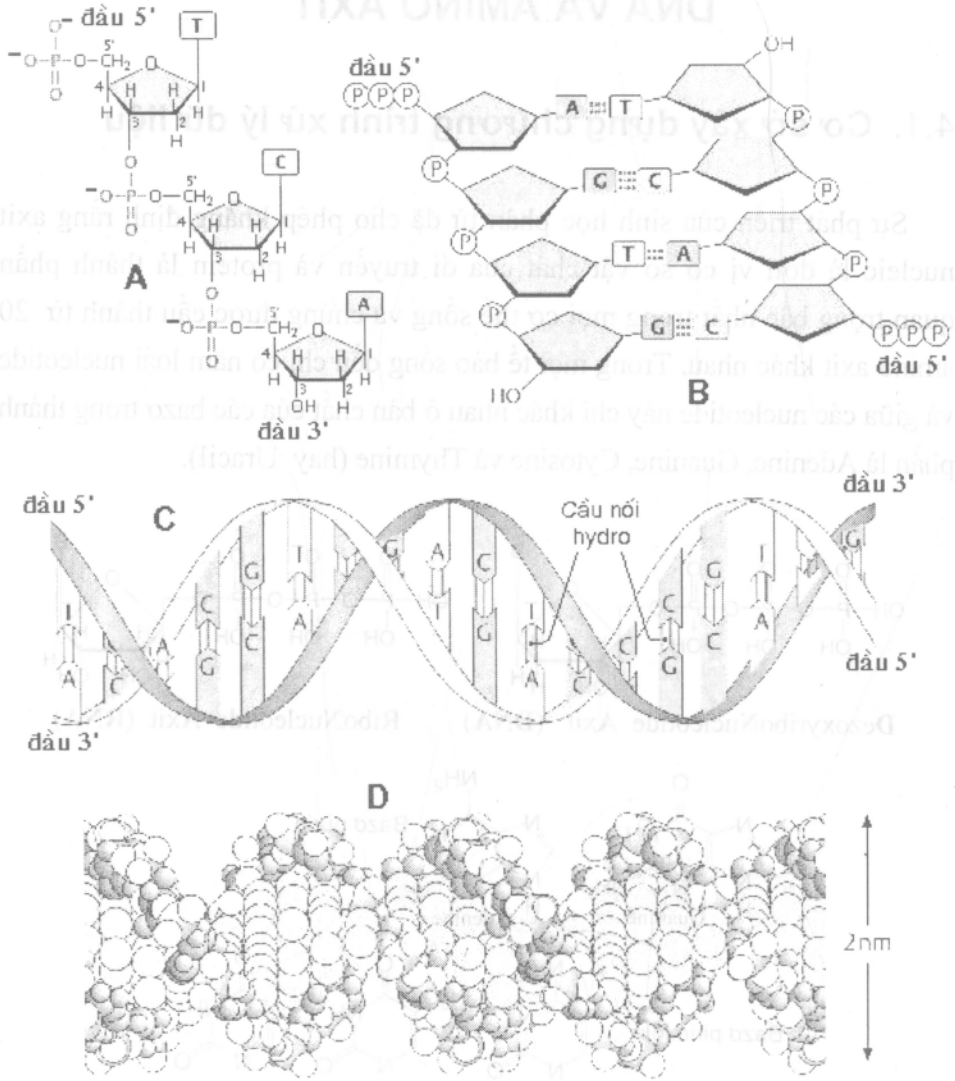
4.1. Cơ sở xây dựng chương trình xử lý dữ liệu

Sự phát triển của sinh học phân tử đã cho phép khẳng định rằng axit nucleic là đơn vị cơ sở vật chất của di truyền và protein là thành phần quan trọng bậc nhất trong mọi cơ thể sống và chúng được cấu thành từ 20 amino axit khác nhau. Trong mọi tế bào sống đều chỉ có năm loại nucleotide và giữa các nucleotide này chỉ khác nhau ở bản chất của các bazơ trong thành phần là Adenine, Guanine, Cytosine và Thymine (hay Uracil).



Hình 4.1. Đơn vị cơ sở của mã thông tin di truyền

Các nucleotide này liên kết và xắp xếp theo trật tự nhất định để hình thành các đoạn đơn vị DNA mang thông tin di truyền, được gọi là các gen. Sơ đồ nguyên lý cấu trúc DNA được mô tả trong hình 4.2.



Hình 4.2. Nguyên lý cấu trúc xoắn kép DNA

A: Sơ đồ cấu trúc liên kết các nucleotide

B: Liên kết cặp bazơ tương đồng đặc hiệu trên chuỗi

C và D: Mô hình cấu trúc xoắn kép DNA

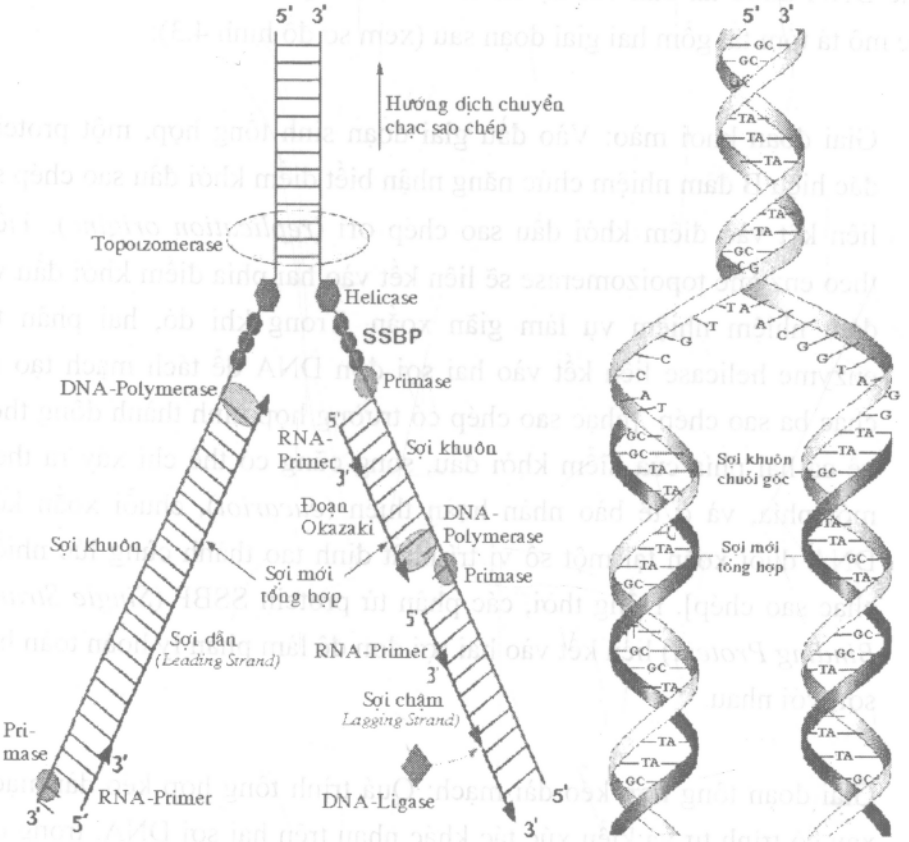
Thành phần mang thông tin di truyền của mọi sinh giới đều có bản chất DNA (chỉ có một số loài virut là RNA). Sự khác biệt giữa các loài chính là do đặc trưng DNA của chúng, ở cấu trúc gen, ở số lượng, hoạt tính và sự tương tác giữa các gen trong quá trình sống. Cấu trúc DNA của sinh giới mang tính ổn định rất cao, do hình thành cấu trúc xoắn kép đặc trưng. Liên kết này là kết quả của sự kết cặp giữa hai bazơ nitơ tương ứng trên hai sợi luôn tuân thủ quy luật của hai cặp bazơ purin – pirimidin là A-T và G-C (hình 4.2).

Nhờ cấu trúc xoắn kép trên nên trong quá trình sinh sản, trật tự cấu trúc DNA được tái bản với độ chính xác cao. Quá trình tái bản DNA có thể mô tả tóm tắt gồm hai giai đoạn sau (xem sơ đồ hình 4.3):

- Giai đoạn khơi mào: Vào đầu giai đoạn sinh tổng hợp, một protein đặc hiệu B đảm nhiệm chức năng nhận biết điểm khởi đầu sao chép sẽ liên kết vào điểm khởi đầu sao chép ori (*replication origine*). Tiếp theo enzyme topoizomerase sẽ liên kết vào hai phía điểm khởi đầu và đảm nhiệm nhiệm vụ làm giãn xoắn. Trong khi đó, hai phân tử enzyme helicase liên kết vào hai sợi đơn DNA để tách mạch tạo ra chạc ba sao chép [chạc sao chép có trường hợp hình thành đồng thời về cả hai phía của điểm khởi đầu, song cũng có thể chỉ xảy ra theo một phía, và ở tế bào nhân hoàn thiện (*eucariot*), chuỗi xoắn kép DNA duỗi xoắn tại một số vị trí nhất định tạo thành cùng lúc nhiều chạc sao chép]. Đồng thời, các phân tử protein SSBP (*Single Strand Binding Protein*) liên kết vào hai sợi đơn để làm phân ly hoàn toàn hai sợi với nhau.

- Giai đoạn tổng hợp kéo dài mạch: Quá trình tổng hợp kéo dài mạch xảy có trình tự và kiểu xúc tác khác nhau trên hai sợi DNA, trong đó một sợi được tổng hợp kéo dài liên tục (sợi dẫn – *Leading Strand*), còn sợi kia (sợi chậm – *Lagging Strand*) được tổng hợp theo từng

đoạn Okazaki rồi mới nối lại với nhau. Quá trình kéo dài này được xúc tác bởi hệ enzyme DNA-polymerase. Trên sợi dẫn, đầu tiên enzyme primase sẽ gắn vào sợi có đầu tự do 3' một đoạn mồi RNA. Tiếp theo, phức hợp enzyme DNA-polymerase III sẽ đọc trình tự mạch khuôn và kéo gắn tiếp các nucleotide tiếp lại thành mạch và vào đúng vị trí tương ứng với trình tự chuỗi khuôn theo hướng 5'–3' (nếu lắp ghép sai, hoạt tính exonuclease sẽ cắt lùi nucleotide này và lắp ghép nucleotide khác tương ứng đúng trở lại). Các nucleotide trước khi tham gia phản ứng được phosphoryl hoá thành dạng hoạt động mang năng lượng.

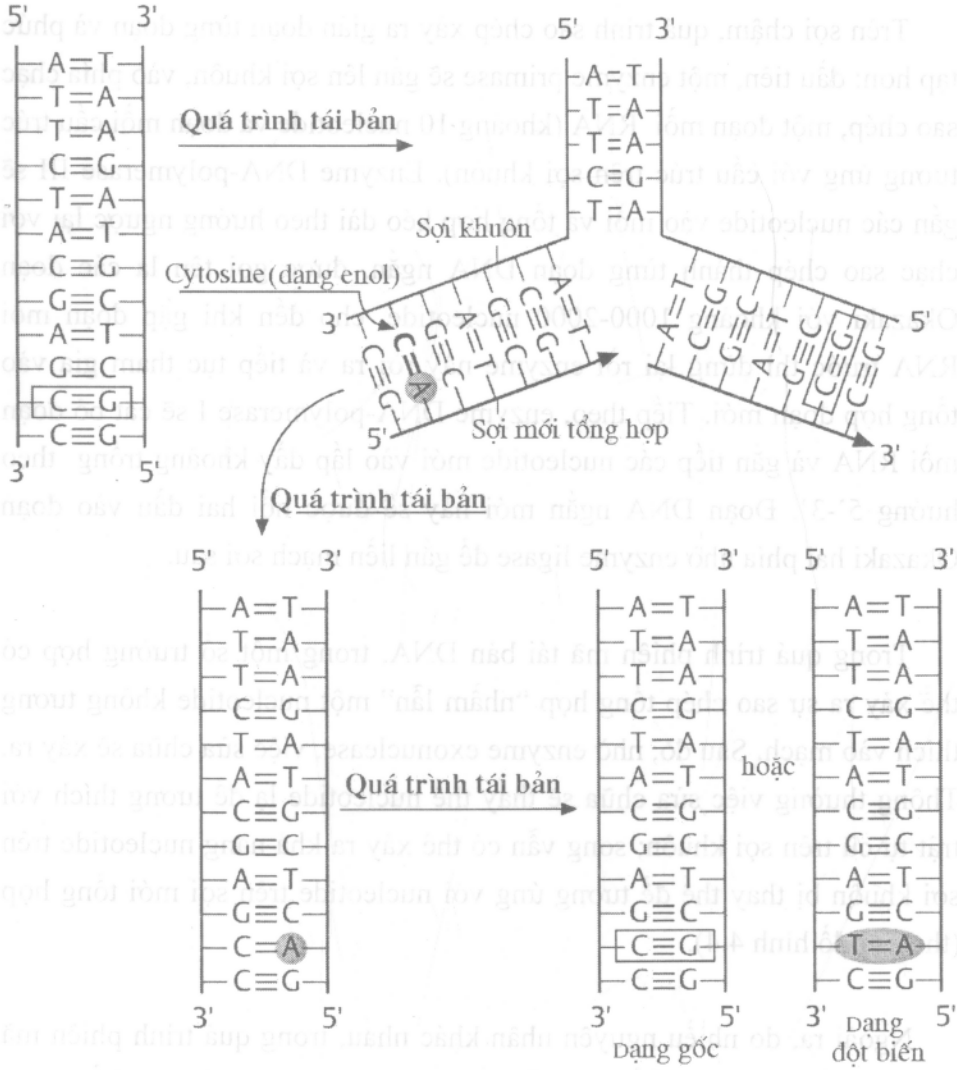


Hình 4.3. Sơ đồ nguyên lý tái bản bảo toàn DNA

Trên sợi chậm, quá trình sao chép xảy ra gián đoạn từng đoạn và phức tạp hơn: đầu tiên, một enzyme primase sẽ gắn lên sợi khuôn, vào phía chạc sao chép, một đoạn mồi RNA (khoảng 10 nucleotide và đoạn mồi cấu trúc tương ứng với cấu trúc trên sợi khuôn). Enzyme DNA-polymerase III sẽ gắn các nucleotide vào mồi và tổng hợp kéo dài theo hướng ngược lại với chạc sao chép thành từng đoạn DNA ngắn, được gọi tên là các đoạn Okazaki với khoảng 1000-2000 nucleotide, cho đến khi gặp đoạn mồi RNA trước thì dừng lại rồi enzyme này rời ra và tiếp tục tham gia vào tổng hợp đoạn mới. Tiếp theo, enzyme DNA-polymerase I sẽ cắt bỏ đoạn mồi RNA và gắn tiếp các nucleotide mới vào lấp đầy khoảng trống theo hướng 5'-3'. Đoạn DNA ngắn mới này sẽ được nối hai đầu vào đoạn Okazaki hai phía nhờ enzyme ligase để gắn liền mạch sợi sau.

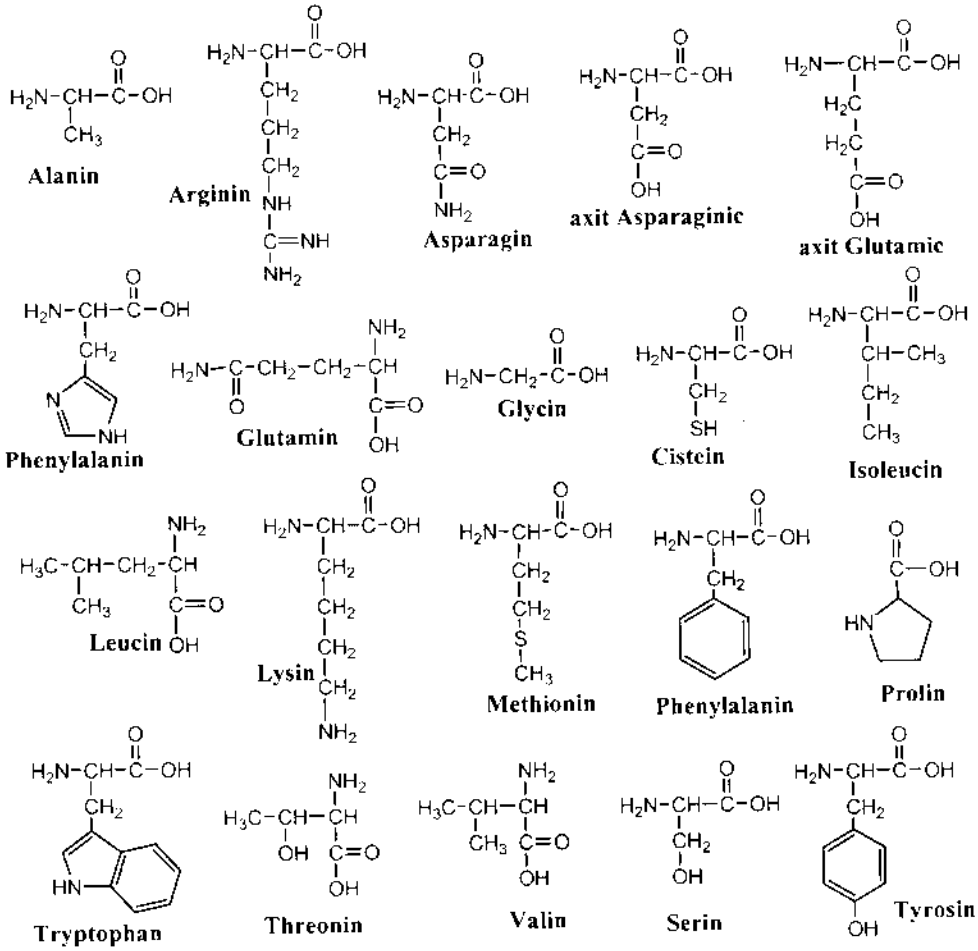
Trong quá trình phiên mã tái bản DNA, trong một số trường hợp có thể xảy ra sự sao chép tổng hợp "nhầm lẫn" một nucleotide không tương thích vào mạch. Sau đó, nhờ enzyme exonuclease, việc sửa chữa sẽ xảy ra. Thông thường việc sửa chữa sẽ thay thế nucleotide lạ để tương thích với trật tự cũ trên sợi khuôn; song vẫn có thể xảy ra khả năng nucleotide trên sợi khuôn bị thay thế để tương ứng với nucleotide trên sợi mới tổng hợp (theo sơ đồ hình 4.4).

Ngoài ra, do nhiều nguyên nhân khác nhau, trong quá trình phiên mã hoặc ngay cả vào thời điểm không xảy ra quá trình sao chép, có thể xảy ra việc đứt đoạn mất một số nucleotide, hay bị chèn thêm vào chuỗi một đoạn nucleotide khác, hoặc xảy ra hiện tượng nối đảo đoạn DNA bị đứt gãy. Tất cả các trường hợp này đều làm thay đổi bản chất trình tự chuỗi xoắn kép DNA ban đầu, nghĩa là gây ra đột biến cấu trúc DNA. Sự biến đổi này, phụ thuộc vào bản chất và vị trí thay thế, có thể không làm thay đổi tính trạng của chúng (đột biến lặn) hoặc làm thay đổi tính trạng ban đầu, hoặc làm xuất hiện tính trạng hoàn toàn mới (đột biến trội).



Hình 4.4. Sơ đồ nguyên lý xuất hiện đột biến trong quá trình phiên mã

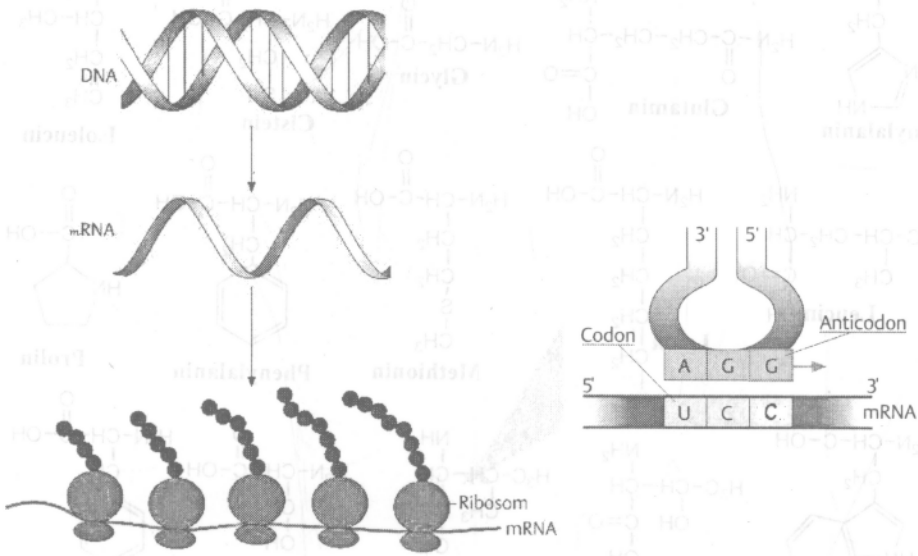
Protein là thành phần chính và quan trọng bậc nhất của mọi cơ thể sống. Các phân tử protein có cấu trúc phức tạp hơn nhiều so với axit nucleic. Ở dạng mang hoạt tính sinh học trong điều kiện tự nhiên chúng tồn tại dưới dạng cấu trúc không gian ba chiều phức tạp. Về bản chất, phân tử protein là một polymer cấu thành từ 20 amino axit khác nhau như trong hình 4.5.



Hình 4.5. Cấu trúc hoá học của các amino axit

Cơ chế quá trình sinh tổng hợp protein có thể mô tả tóm tắt như sau: Thông tin di truyền mã hoá cho phân tử protein lưu giữ trong cấu trúc chuỗi DNA, đầu tiên trải qua quá trình phiên mã để tổng hợp ra phân tử RNA thông tin mRNA (ở các sinh vật nhân hoàn thiện, quá trình phiên mã không xảy ra liên tục mà đứt quãng do bỏ qua các đoạn không mang mã sinh tổng hợp intron nằm trên sợi DNA). Tiếp theo, phân tử mRNA này sẽ trở thành sợi khuôn cho quá trình dịch mã trên ribosom để tổng hợp nên phân tử protein tương ứng. Song song với quá trình trên, các amino axit

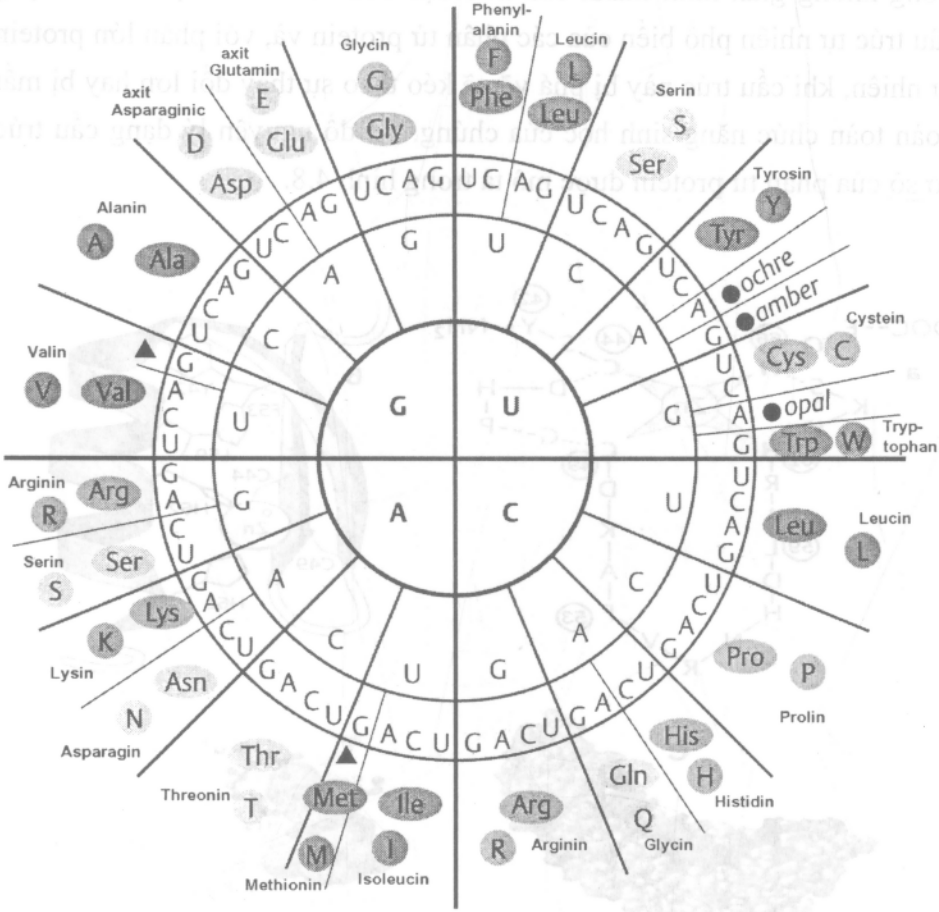
tham gia vào quá trình được hoạt hoá và sẽ liên kết với các phân tử RNA vận tải (tRNA) tương ứng. Tiếp theo, các phân tử tRNA sẽ vận chuyển chúng đến ribosom. Với sự nhận biết tương thích của cặp liên kết codon-anticodon (hình 4.6), phân tử tRNA sẽ vận chuyển amino axit này vào đúng vị trí liên kết, được quy định trên trình tự cấu trúc chuỗi mRNA.



Hình 4.6. Sơ đồ nguyên lý quá trình phiên mã và dịch mã tổng hợp protein

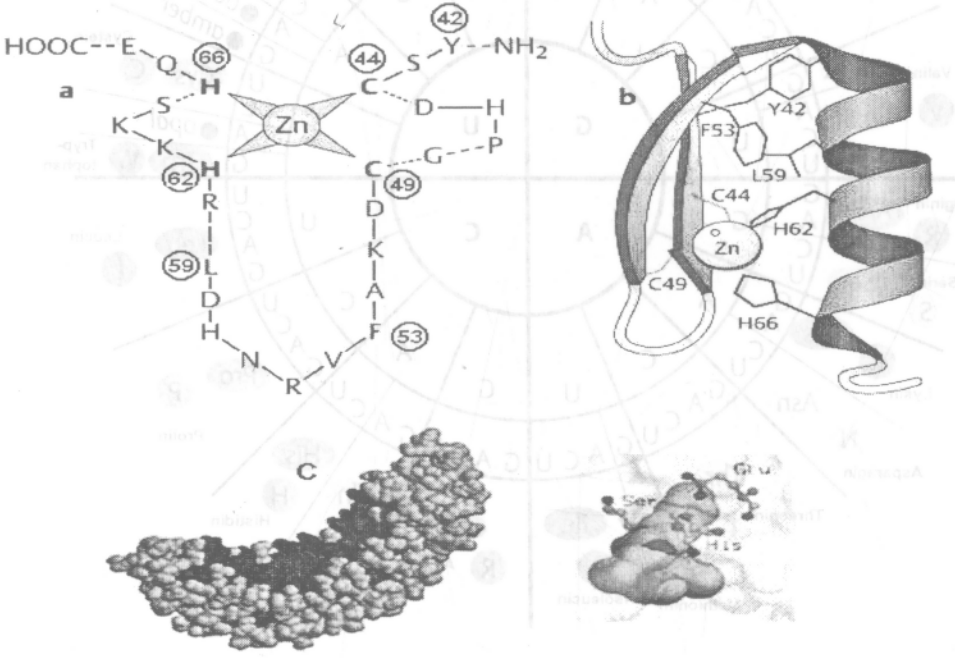
Theo cơ chế trên, trình tự cấu trúc chuỗi amino axit được tổng hợp ra tương đồng hoàn toàn với cấu trúc chuỗi khuôn mRNA. Điều đó cho phép rút ra kết luận là vị trí tương đối của các amino axit trên chuỗi được mã hoá trên chính cấu trúc chuỗi DNA đã được sử dụng làm khuôn trong quá trình phiên mã. Đồng thời, như đã trình bày ở phần trên, gen là một đơn vị chức năng cơ sở của bộ máy di truyền và xác định một tính trạng nhất định. Như vậy, có thể nói thông tin di truyền được mã hoá trong chính cấu trúc của các gen tương ứng. Sinh học hiện đại đã xác định được mỗi bộ ba nucleotide là một đơn vị mã thông tin di truyền. Mối liên hệ giữa các đơn

vị mã di truyền với các amino axit hay thông tin di truyền tương ứng được trình bày trong biểu đồ hình 4.7. Trong biểu đồ này, thứ tự bộ ba nucleotide mã hoá cho amino axit tương ứng được đọc từ vòng tròn trong ra vòng tròn bên ngoài và cụm ký tự UAA, UAG, UGA đảm nhiệm vai trò tín hiệu kết thúc chuỗi.



Hình 4.7. Biểu đồ xác định mã di truyền

Trình tự chuỗi polypeptide trên được gọi là cấu trúc bậc một của phân tử protein. Ở trạng thái tự nhiên, chuỗi polymer này liên kết với nhau theo kiểu nhất định để hình thành cấu trúc xoắn đặc trưng α và cấu trúc β (*α-helix* và *β-sheet* – cấu trúc bậc hai). Các chuỗi α và β này cuộn xoắn lại trong không gian theo kiểu trật tự riêng của mỗi dạng protein tạo ra cấu trúc không gian bậc ba và các phân tử protein bậc ba tiếp tục cuộn lại trong không gian hình thành cấu trúc bậc bốn. Cấu trúc bậc cao là dạng cấu trúc tự nhiên phổ biến của các phân tử protein và, với phần lớn protein tự nhiên, khi cấu trúc này bị phá vỡ sẽ kéo theo sự thay đổi lớn hay bị mất hoàn toàn chức năng sinh học của chúng. Sơ đồ nguyên lý dạng cấu trúc cơ sở của phân tử protein được mô tả trong hình 4.8.

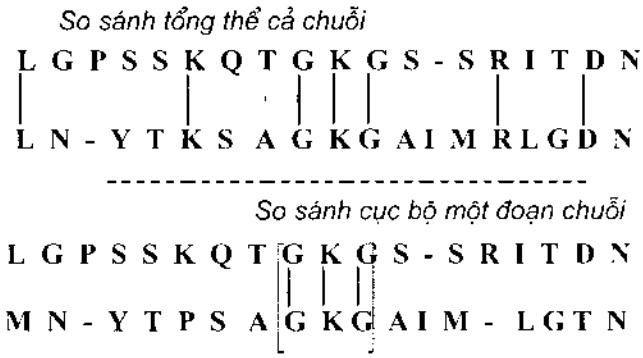


Hình 4.8. Mô hình cấu trúc xoắn protein

Như đã trình bày ở phần trên, cấu trúc chuỗi xoắn kép DNA, cơ quan mang thông tin di truyền, cấu trúc phân tử protein, quá trình tái bản DNA và quá trình sinh tổng hợp protein xảy ra theo cơ chế đặc thù và logic. Sử dụng các ký tự viết tắt tên các nucleotide và tên các amino axit người ta dễ dàng áp dụng công nghệ thông tin để mô phỏng và mô hình hoá các quá trình ấy, nghĩa là có thể dễ dàng số hoá để mô tả đặc tính tự nhiên của các vật liệu sinh học. Trên cơ sở này, việc nghiên cứu, so sánh, xử lý dữ liệu và thiết kế mô phỏng chương trình nghiên cứu thực nghiệm có thể tiến hành một cách thuận lợi và hiệu quả hơn so với cách thức đã thực hiện theo công nghệ cổ điển.

4.2. Nghiên cứu so sánh cấu trúc chuỗi

So sánh cấu trúc chuỗi là kỹ thuật hay thuật toán để so sánh cấu trúc hai chuỗi (pair-wise alignment) hay so sánh đồng thời nhiều chuỗi với nhau (multiple sequence alignment), bằng cách tìm kiếm xác định các đặc điểm hoặc các thuộc tính riêng giống nhau giữa các chuỗi. Việc so sánh có thể tiến hành theo từng vùng (local alignment) hay thực hiện trên toàn bộ chuỗi (global alignment). Mô hình so sánh đơn giản nhất có thể mô tả qua sơ đồ hình 4.9.



Hình 4.9. Mô hình hai dạng so sánh chuỗi giản đơn

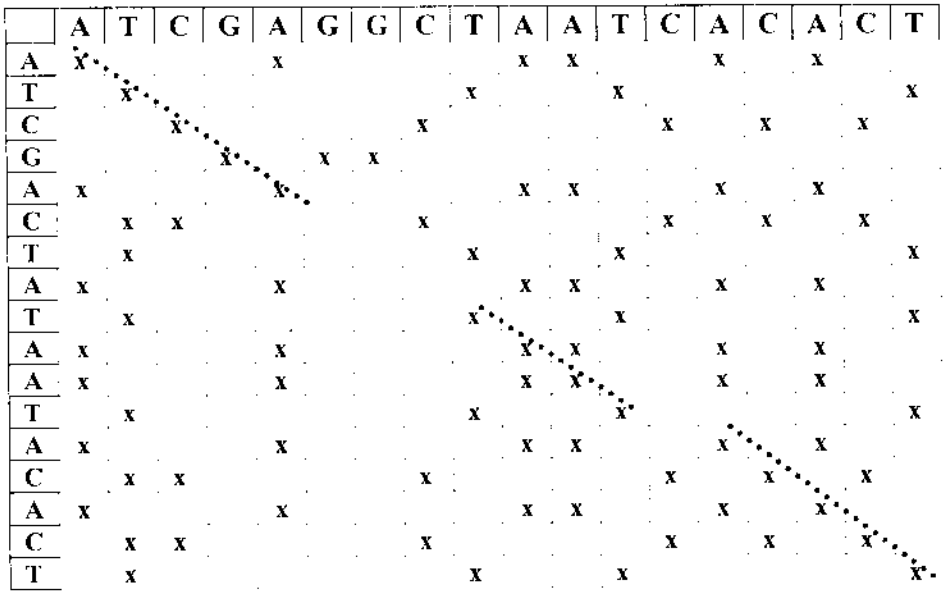
Kỹ thuật so sánh cấu trúc chuỗi được ứng dụng để khám phá thông tin về chức năng, cấu trúc chuỗi và mối quan hệ tiến hoá thể hiện trong sự biến đổi cấu trúc giữa các chuỗi với nhau. Thí dụ: hai chuỗi ADN tương đồng cao với nhau về cấu trúc rất có khả năng cùng có nguồn gốc từ cùng một chuỗi, nghĩa là có quan hệ họ hàng gần gũi về mặt tiến hoá với nhau, và rất nhiều khả năng chúng sẽ có những chức năng tương đồng với nhau, hay hai chuỗi protein đồng nhất cao với nhau, nhiều khả năng sẽ có đặc tính hoá sinh và có cấu trúc không gian tương ứng giống nhau. Từ kết quả so sánh này, cho phép nhà nghiên cứu có thể căn cứ vào đặc tính đã biết của chuỗi nọ để dự đoán đặc tính của chuỗi kia. Nhờ vậy, cho phép rút ngắn rất lớn khối lượng thực nghiệm kiểm tra các đặc tính trên và làm cơ sở để xây dựng các phương án tổ chức nghiên cứu tiếp theo...

Để tìm hiểu cơ sở thuật toán so sánh chúng ta hãy xem phương pháp phân tích so sánh sử dụng ma trận điểm đơn giản sau đây: Giả sử người phân tích cần so sánh độ tương đồng (hay phân ly) của hai chuỗi với nhau. Đầu tiên người ta thiết lập bảng ô vuông và chép trình tự một chuỗi theo hàng và một chuỗi theo cột dọc vuông góc với nhau. Sau đó, đánh dấu vào tất cả các ô vuông tương ứng cùng với một nucleotide, dùng thước kẻ nối các ô được đánh dấu liền kề nhau theo chiều đường chéo phía góc trên bên trái kẻ xuống để xác định đoạn chuỗi tương đồng theo sơ đồ hình 4.10.

Trong thí dụ so sánh này, có thể thấy dường như tồn tại một khả năng là hai chuỗi có cùng nguồn gốc, với sự sao chép "nhầm lẫn" giữa chúng ở đoạn GGC và một đột biến đứt đoạn tại C theo sơ đồ như sau:

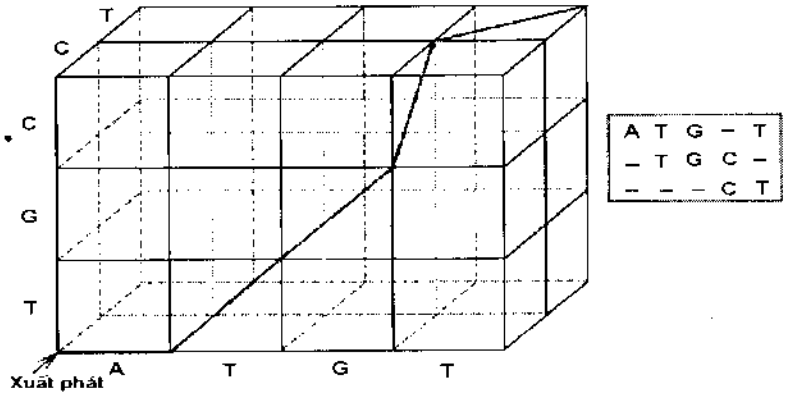
A T C G A G G C T A A T C A C A C T

A T C G A C T A T A A T – A C A C T



Hình 4.10. Sơ đồ ma trận điểm so sánh xác định cấu trúc chuỗi

Với phương án so sánh đồng thời ba chuỗi: ATGT, TGC và CT thì sơ đồ nguyên lý giản đơn nhất có thể mô tả qua hình 4.11.



Hình 4.11. Sơ đồ nguyên lý so sánh cấu trúc ba chuỗi

Tuy nhiên, để tìm hiểu và khám phá quy luật về sự tương đồng và/hay phân ly của các sinh giới trong tự nhiên, đòi hỏi phải nghiên cứu trên lượng rất lớn các chuỗi có đặc tính gần gũi nhau. Nghĩa là phải tiến hành phép so sánh đồng thời từng cặp với nhau và tất cả các cặp đối tượng nghiên cứu. Để thực hiện được mục tiêu trên, nhiều nhóm tác giả đã hoàn thiện các chương trình xử lý dữ liệu đa chuỗi *(Dinamic Programing for Multiple Sequence Alignment* - trên cơ sở ứng dụng nhiều thuật toán khác nhau, thí dụ các thuật toán ma trận PAM, ma trận BLOSUM, thuật toán mô hình hộp đen Markov...). Sau đây là một số địa chỉ (hay đường dẫn siêu liên kết) và đặc điểm chính của một số chương trình phân tích cấu trúc hiện nay:

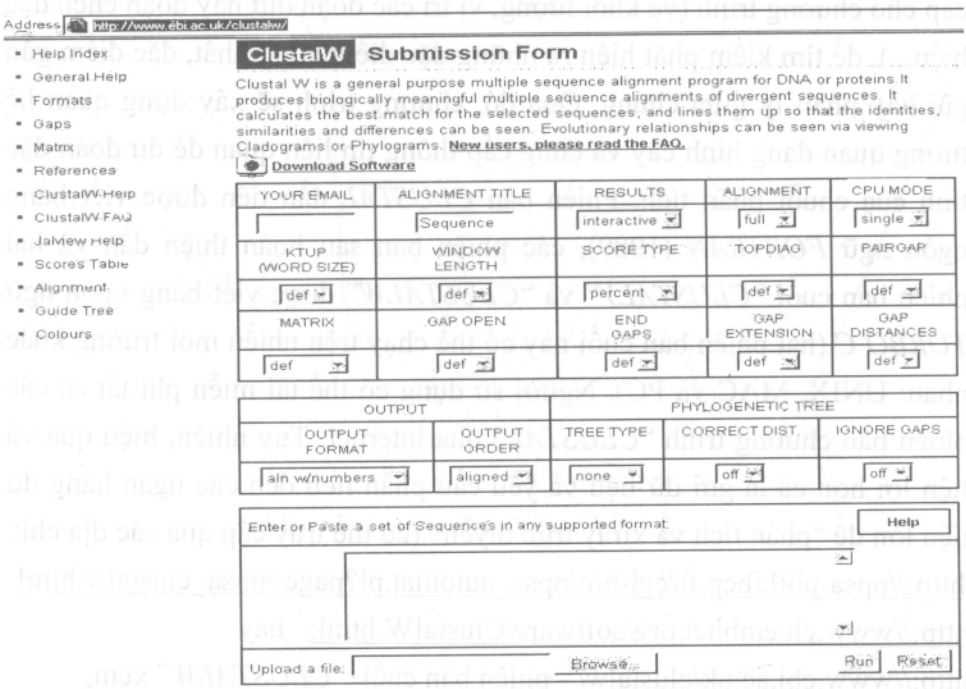
- ***Diaglin 2.2.1*** : http://bibiserv.techfak.uni-bielefeld.de/dialign/
- ***AlignACE3.0*** : http://atlas.med.harvard.edu/cgi-bin/alignace.pl
- ***Genome Vista*** : http://pipeline.lbl.gov/cgi-bin/GenomeVista
- ***MAVID multiple alignment***: http://baboon.math.berkeley.edu/mavid/
- ***Partial Order Alignment*** : http://www.bioinformatics.ucla.edu/poa/
- multiple alignment of genomic sequences using CHAOS and DIALIGN : http://dialign.gobics.de/chaos-dialign-submission
- Wavis Alignment visualization tools : http://wavis.img.cas.cz/
- ***The Gibbs Motif Sampler (for DNA)*** : http://bayesweb.wadsworth.org/cgi-bin/gibbs.9.pl?data_type=DNA
- ***Meta-MEME*** : http://metameme.sdsc.edu/submit-verify.html
- ***GP Sequence Homology Search*** : http://spock.genes.nig.ac.jp/~genome/adseqsch.html
- ***MaliP / Multiple alignment for protein sequences*** : http://www.softberry.com/berry.phtml?topic=mali&group=programs &subgroup=mali

5. CHƯƠNG TRÌNH PHÂN TÍCH CẤU TRÚC CHUỖI CLUSTALW

5.1. Đại cương về chương trình CLUSTAL

Chương trình "*CLUSTAL*" là dãy các phiên bản phần mềm phân tích kết quả thí nghiệm về cấu trúc chuỗi DNA hay protein, bằng phương pháp so sánh đồng thời giữa tất cả các chuỗi mà người yêu cầu đã lựa chọn cung cấp cho chương trình (về khối lượng, vị trí các đoạn đứt hay đoạn chèn đặc hiệu...), để tìm kiếm phát hiện ra những đặc điểm đồng nhất, đặc điểm gần gũi hay phân ly giữa chúng. Qua đó, chương trình sẽ xây dựng quan hệ tương quan dạng hình cây và cung cấp thông tin liên quan để dự đoán đặc tính của chuỗi phân tích. Phiên bản *CLUSTAL* đầu tiên được viết bằng ngôn ngữ *FORTRAN* (1989), các phiên bản sau hoàn thiện dần và hai phiên bản cuối "*CLUSTALV*" và "*CLUSTALW*" được viết bằng ngôn ngữ *TURBO C* (hai phiên bản cuối này có thể chạy trên nhiều môi trường khác nhau: UNIX, MAC và PC). Người sử dụng có thể tải miễn phí tất cả các phiên bản chương trình "*CLUSTAL*" qua internet. Tuy nhiên, hiệu quả và tiện lợi hơn cả là gửi dữ liệu và yêu cầu phân tích đến các ngân hàng dữ liệu lớn để "phân tích và xử lý trực tuyến" (có thể truy cập qua các địa chỉ: http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html http://www.ch.embnet.org/software/ClustalW.html; hay http://www.ebi.ac.uk/clustalw/ - phiên bản cuối "*CLUSTALW*" xem *Thompson, J. D.; Higgins, D.G. and Gibson, T.J. (1994) - Nucleic acids research, 22(22):4673-4680).*

Ngày nay, việc xử lý phân tích cấu trúc chuỗi đã trở thành công cụ hết sức quan trọng trong công nghệ sinh học. Thí dụ kết quả so sánh cấu trúc chuỗi DNA cho phép chỉ ra các vùng bảo toàn và vùng phân ly của chuỗi kiểm tra. Trên cơ sở đó, nhà sinh học có thể dự đoán được đặc tính chuỗi, hoạch định các thực nghiệm để kiểm tra lại các đặc tính của chuỗi, hoặc tìm kiếm phương án tác động nhằm làm biến đổi cấu trúc chuỗi, hay từ đó dự đoán được cấu trúc và đặc tính của các gen (và protein) mới (bao gồm cả các sản phẩm nhân tạo được tạo ra mang các đặc tính mong muốn). Việc so sánh có thể tiến hành theo phương án toàn bộ (*global alignment*) hay từng đoạn (*local alignment*), so sánh tổng hợp với tất cả các chuỗi đã lựa chọn hay so sánh từng cặp chuỗi riêng rẽ với nhau... Giao diện trực tuyến chương trình ClustalW có dạng như trong hình 5.1.



Hình 5.1. Giao diện chương trình ClustalW trực tuyến (http://www.ebi.ac.uk/clustalw/)

5.2. Sử dụng chương trình

Dữ liệu của chuỗi cần phân tích cấu trúc phải được lưu toàn bộ trong một tệp dưới một trong các định dạng ngôn ngữ sau: *FASTA (Pearson), NBRF/PIR, EMBL/UniProt/Swiss Prot, GDE, ALN/CLUSTALW, GCG/MSF* và *GCG9/RSF* (Thường các chương trình tích hợp sẵn trong thiết bị phân tích hiện đại hay dữ liệu lưu trữ trong các ngân hàng dữ liệu trực tuyến đều đã chuyển kết quả định dạng thành một trong các ngôn ngữ trên). Chương trình xử lý sẽ tự động nhận dạng các dạng tệp trên và phân biệt chuỗi DNA/RNA hay chuỗi amino axit.

Giả sử người phân tích muốn nghiên cứu xác định quy luật vận động tương đối tồn tại trong trong một nhóm chuỗi có đặc tính sinh học gần gũi nhau (nhóm chuỗi cùng nguồn). Các chuỗi này được tìm kiếm và lựa chọn từ trong các ngân hàng dữ liệu (xem kết quả tìm kiếm, chương 9), sau khi đã phân tích kỹ lưỡng đặc tính tương đồng về mặt sinh học, thí dụ từ kết quả phân tích đặc tính sinh học đã lựa chọn được 13 chuỗi sau:

BF056441	**BE848719**	**BF022813**	**BF452255**	**BG089808**
BG147728	**AF186109**	**AF310722**	**AF362887**	**BI817778**
	AF186110	**AF362886**	**AF087679**	

Để phát hiện quy luật vận động tương đối của nhóm chuỗi này, người sử dụng tải về tất cả 13 tệp trên, từ ngân hàng dữ liệu, rồi chép tuần tự vào thành một tệp chung. Yêu cầu bắt buộc là các chuỗi là phải cùng viết theo một kiểu định dạng thống nhất, được chương trình ClustalW chấp nhận, mà không phân biệt thứ tự các chuỗi được chép. Thí dụ, theo ngôn ngữ FASTA mỗi chuỗi gồm hai phần: dòng thông tin đầu (gồm 4 thông số, phân cách bằng dấu ";" là: ký hiệu khởi đầu ">", ký mã hiệu chuỗi, tên chuỗi và mô tả tóm tắt đặc tính chuỗi; còn dòng thứ hai là trình tự cấu trúc của chuỗi. Tệp dữ liệu chung có dạng như sau:

>embl:BF056441 BF056441; 7k05a04.x1 NCI_CGAP_GC6 Homo sapiens cDNA clone IMAGE:3443238 3' similar to SW:TPM4_HUMAN P07226 TROPOMYOSIN, FIBROBLAST NON-MUSCLE TYPE ;, mRNA sequence.

acagttgcaagaatctaaagtgtggattttattccattgcacaatttgctagtgtatttc
ctgggtagtgtggtgctgaataaataggaataaatgctacttaaggaaaaaataagagag
ctgaaaaagctggtgccatttgaaaaaaaaaagggaaggaatgagatttaactggtgctc
aaagcttctccgatacaaaatatttggtcatgtattcataatttgcttgacatttccagc
aaagcgaagatggcaataacaaaaggaacttcttacaagagaagagaaagacccacggag
ctccagagtttctgttggaacaagactcttctgttttgcttatatacagttaagttcgtt
tagtgtctgatccagtgtctgatgtaagcccacgttctcttctttggcctgggcaagttt
ctcttccaggtcatcaattgtcttttccagttttgcaaccgttctctctgcaaattcagc
acgggtctcagcctctttcagtttgtcagacagaagtttaatttcttcttcatatttgtc
ctccttttcagaatacttttcagatgcagcctccagagatttcagattgtagttacaatt
ctgagttcttctccaggtcaccacattttattcagacacctccgcacgcttctctgccct
ctcagctaacccttc

>embl:BE848719 BE848719; uw40c07.y1 Soares_thymus_2NbMT Mus musculus cDNA clone IMAGE:3419148 5' similar to SW:TPM4_HUMAN P07226 TROPOMYOSIN, FIBROBLAST NON-MUSCLE TYPE ;, mRNA sequence.

tgttaccaatctgcttggcatttcctgcaaggtggaaacctggtaataagcggaacttct
tacaaaagaggaagacagggcacactctctggagtggagttggtgttaaaacagtactct
tctggtttagtttatatacagttaagttcgtttagtgtctggtccagtgtctgatgtaag
cccacattctcttctttggcctgggcaagtttttcttccaggtcatcgattgtcttctcc
agtttagaaactgtcctttctgcaaactcagctcgggtctcagcctccttcagcttgtca
gacagaagcttgatttcttcttcatatttatcctccttttcagaatacttttcagaagca
gcctccagtgatttcagattgttagttacattcttgagctcttcttccaggtcaccacac
tttagttcagatacctccgccctctcctctgctctcttcagctcaccctccaggatgacc
aacttacgagcaacctcctcatacttgcggtcggctacgtcagtgatgtgcttggctgct
ttgagctgcatctccaggatctacatcttttgctcatctttcatggcttcgttctctatt
accttcatgcctctctcactctcatcagcagccctctctgcctctttcagattctgcaag
gctgtggccagctgcttctgaccctgtccaattccttc

>embl:BF022813 BF022813; uw40c07.x1 Soares_thymus_2NbMT Mus musculus cDNA clone IMAGE:3419148 3' similar to SW:TPM4_RAT P09495 TROPOMYOSIN 4, EMBRYONIC FIBROBLAST ISOFORM ;, mRNA sequence.

ccggatcccagcagaacgattgagctatggccggcctcaactcactggaggcagtgaagc
gcaagatccaggccctgcagcagcaggcagacgacgcagaggatcgcgcgcaaggcctgc
agcgcgaactggatggcgagcgcgagcggcgcgagaaagctgaaggagatgcagccgctc
tcaaccgccgcatccaactgctggaggaggaactggaccgggctcaggagcagctggcca
cagccctgcagaatctggaagaggcagagaaggctgctgatgagagtgagagaggcatga
aggtaatagagaaccgagccatgaaagatgaggaaaagatggagatcctggagatgcagc
tcaaagaagccaagcacatcactgacgaagccgaccgcaagtttgaggaggttgctcgt

>embl:BF452255 BF452255; uz86d11.y1 NCI_CGAP_Lu29 Mus musculus cDNA clone IMAGE:3675957 5' similar to SW:TPM4_RAT P09495 TROPOMYOSIN 4, EMBRYONIC FIBROBLAST ISOFORM ;, mRNA sequence.

gagcccagcagaacgattgagctatggccggcctcaactcactggaggcagtgaagcgca
agatccaggccctgcagcagcaggcagacgacgcagaggatcgcgcgcaaggcctgcagc
gcgaactggatggcgagcgcgagcggcgcgagaaagctgaaggagatgcagccgctctca

accgccgcatccaactgctggaggaggaactggaccgggctcaggagcagctggccacag
ccctgcagaatctggaagaggcagagaaggctgctgatgagagtgagagaggcatgaagg
taatagagaaccgagccatgaaagatgaggaaaagatggagatcctggagatgcagctca
aagaagccaagcacatcactgacgaagccgaccgcaagtantgagaggttgctcgtaagt
tggtcatcctggagggtgagctgaagagagcagaggagagggcgaggtatctgaactaaa
gttggtgacctggagaagagctcaagaatgtaactaac

>embl:BG089808 BG089808; mab82b11.x1 NCI_CGAP_BC3 Mus musculus cDNA clone IMAGE:3976676 3' similar to SW:TPM4_HUMAN P07226 TROPOMYOSIN, FIBROBLAST NON-MUSCLE TYPE ;, mRNA sequence.

agctgtcgccggagcccagcagaacgagtgagctatggccggcctcaactcactggaggc
agtgaagcgcaagatccaggccctgcagcagcaggcagacgacgcagaggatcgcgcgca
aggcctgcagcgcgaactggatggcgagcgcgagcggcgcgagaaagctgaaggagatgc
agccgctctcaaccgccgcatccaactgctggaggaggaactggaccgggctcaggagca
gctggccacagccctgcagaatctggaagaggcagagaaggctgctgatgagagtgagag
acgcatgaaggtaatagagaaccgagccatgaaagatgaggaaaagatggagatcctgga
gatgcagctcagagaagccaagcacatcactgatgaagccgaccgcaagtatgatgaggt
tgctcgtaagttggtcatcctggagggtgagctgaagagagcagatgagcgggcggaggt
atctgaactaaagtgtggtgacctgtaataagagctcatgaatgtaactaacaatctgaa
atgactggaggctgtatttgagaagtattctgaataggaggattagtatgaagaagaaga
taagcttatgcctgataagctgaaggtaggtggaaaccagtctggatttgcagacaga

>embl:BG147728 BG147728; mab53f06.x1 Soares_NMEBA_branchial_arch Mus musculus cDNA clone IMAGE:3974147 3' similar to SW:TPM4_HUMAN P07226 TROPOMYOSIN, FIBROBLAST NON-MUSCLE TYPE ;, mRNA sequence.

caactcactggaggcagtgaagcgcaagatccaggccctgcagcagcaggcagacgacgc
anaggatcgcgcgcaaggcctgcagcgcgaactggatggcgagctctagcggcgcgagaa
agctgagggagagggggcgctctcaaccgccgcatccaactgctggaggaggaactgga
ccgggctcatgagcagctggccacagccctgcagaatctggaagaggcagagaaggctgc
tgatgagagtgagagaggcatgaaggtaatagagaaccgagccatgaaagatgaggaaaa
gatggagatcctggagatgcagctcaaagaagccaagcacatcactgatgaagccgaccg
caagtatgaggaggttgctcgtaagttggtcatcctggagggtgagctgaagagagcata
ggagcgggcggaggtatctgaactaaagtgtggtgaccctgaagaagagctcaagaatgt
aactaacaatctgaaatcactggaggctgcttctgaaaagtattctgaaaaggag

>embl:BI817778 BI817778; G3-F20 Axolotl Lambda Zap Library Ambystoma mexicanum cDNA similar to Homo sapiens gb|AAG17014.1|AF186109_1 (2.0e-40), TPM4-ALK fusion oncoprotein type 2, mRNA sequence.

ccggggtaccctaagccttctcggatccgagactcttcttcccgttgaggcccccccccc
gccccccagcagggaagatgtcgggtggcagttccatcgatgcggtgaagaagaagatcc
agagccttcagcaggtggcggacgaggccgaagagcgggccgagatcctgcagagggagg
tggacgccgagaggcagtgcagggagcgggccgaggcagacgtgggatcgctcaaccgcc
gcatccagctggtagaggaggagctggaccgtgcccaggagcgccttgccactgccctgc
tgaagttggaggaggcggagaaagctgcagacgagagtgaacgaggcatgaaggtcattg
aaaaccgagccaccaaggacgaggagaagatggagatccaggagatgcagttgaaagagg

ccaaacacatagcagaggaggccgaccgcaaa

>embl:AF186109 AF186109; Homo sapiens TPM4-ALK fusion oncoprotein type 2 (TPM4-ALK fusion) mRNA, partial cds.

gccatggccggcctcaactccctggaggcggtgaaacgcaagatccaggccctgcagcag
caggcggacgaggcggaagaccgcgcgcagggcctgcagcgggagctggacggcgagcgc
gagcggcgcgagaaagctgaaggtgatgtggccgccctcaaccgacgcatccagctcgtt
gaggaggagttggacagggctcaggaacgactggccacggccctgcagaagctggaggag
gcagaaaaagctgcagatgagagtgagagaggaatgaaggtgatagaaaaccgggccatg
aaggatgaggagaagatggagattcaggagatgcagctcaaagaggccaagcacattgcg
gaagaggctgaccgcaaatacgaggaggtagctcgtaagctggtcatcctggagggtgag
ctggagagggcagaggagcgtgcggaggtgtctgaactaaaatgtggtgacctggaagaa
gaactcaagaatgttactaacaatctgaaatctctggaggctgcatctgaaaagtattct
gaaaaggaggacaaatatgaagaagaaattaaacttctgtctgacaaactgaaagaggct
gagacccgtgctgaatttgcagagagaacggttgcaaaactggaaaagacaattgatgac
ctggaagtgtaccgccggaagcaccaggagctgcaagccatgcagatggagctgca

>embl:AF186110 AF186110; Homo sapiens TPM4-ALK fusion oncoprotein type 1 (TPM4-ALK fusion) mRNA, partial cds.

ctctcagccaggcggattgaaggatggaattccaacgaggctcccccgcctcgtcccacc
ttggctgaaggtgatgtggccgccctcaaccgacgcatccagctcgttgaggaggagttg
gacagggctcaggaacgactggccacggccctgcagaagctggaggaggcagaaaaagct
gcagatgagagtgagagaggaatgaaggtgatagaaaaccgggccatgaaggatgaggag
aagatggagattcaggagatgcagctcaaagaggccaagcacattgcggaagaggctgac
cgcaaatacgaggaggtagctcgtaagctggtcatcctggagggtgagctggagagggca
gaggagcgtgcggaggtgtctgaactaaaatgtggtgacctggaagaagaactcaagaat
gttactaacaatctgaaatctctggaggctgcatctgaaaagtattctgaaaaggaggac
aaatatgaagaagaaattaaacttctgtctgacaaactgaaagaggctgagacccgtgct
gaatttgcagagagaacggttgcaaaactggaaaagacaattgatgacctggaagtgtac
cgccggaagcaccaggagctgcaagccatgcagatggagctgcagagccctgagtacaag
ctgagcaagctccgcacctcgaccatcatgaccgactacaaccccaactactgctttgct
ggcaagacctcctccatcagtgacctgaaggaggtgccgcggaaaaacatcaccctcatt
cggggtctgggccatggcgcctttggggaggtgtatgaaggccaggtgtccggaatgccc
aacgacccaagcccccctgcaagtggctgtgaagacgctgcctg

>embl:AF310722 AF310722; Homo sapiens tropomyosin 4-anaplastic lymphoma kinase fusion protein (TPM4-ALK) mRNA, partial cds.

cgcgccatggccggcctcaactccctggaggcggtgaaacgcaagatccaggccctgcag
cagcaggcggacgaggcggaagaccgcgcgcagggcctgcagcgggagctggacggcgag
cgcgagcggcgcgagaaagctgaaggtgatgtggccgccctcaaccgacgcatccagctc
gttgaggaggagttggacagggctcaggaacgactggccacggccctgcagaagctggag
gaggcagaaaaagctgcagatgagagtgagagaggaatgaaggtgatagaaaaccgggcc
atgaaggatgaggagaagatggagattcaggagatgcagctcaaagaggccaagcacatt
gcggaagaggctgaccgcaaatacgaggaggtagctcgtaagctggtcatcctggagggt
gagctggagagggcagaggagcgtgcggaggtgtctgaactaaaatgtggtgacctggaa

gaagaactcaagaatgttactaacaatctgaaatctctggaggctgcatctgaaaagtat
tctgaaaaggaggacaaatatgaagaagaaattaaacttctgtctgacaaactgaaagag
gctgagacccgtgctgaatttgcagagagaacggttgcaaaactggaaaagacaattgat
gacctggaagtgtaccgccggaagcaccaggagctgcaagccatgcagatggagctgcag
agccctgagtacaagctgagcaagctccgcacctcgaccatcatgaccgactacaacccc
aactactgctttgctggcaagacctcctccatcagtgacctgaaggaggtgccgcggaaa
aacatcaccctcattcggggtctgggccatggcgcctttggggaggtgtatgaaggccag
gtgtccggaatgcccaacgacccaagccccctgcaagtggctgtgaagacgctgcctgaa
gtgtgc

>embl:AF362886 AF362886; Homo sapiens tropomyosin 4-anaplastic lymphoma kinase fusion protein major isoform mRNA, partial cds.

ctggcagagtcccgttgccgagagatggatgagcagattagactgatggaccagaacctg
aagtgtctgagtgctgctgaagaaaagtactctcaaaaagaagataaatatgaggaagaa
atcaagattcttactgataaactcaaggaggcagagacccgtgctgaatttgcagagaga
acggttgcaaaactggaaaagacaattgatgacctggaagtgtaccgccggaagcaccag
gagctgcaagccatgcagatggagctgcagagccctgagtacaagctgagcaagctccgc
acctcgac

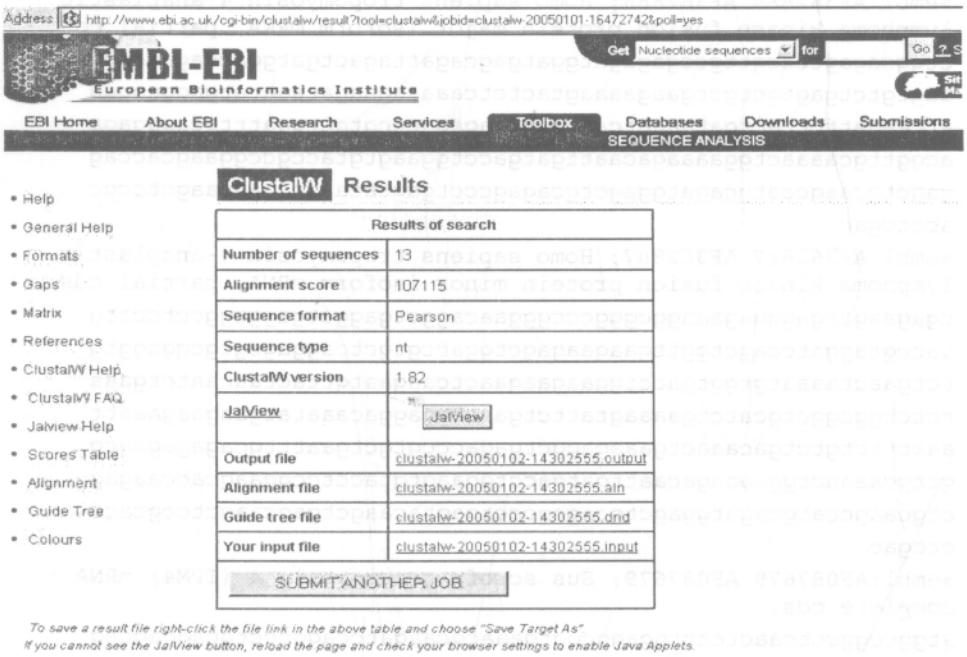
>embl:AF362887 AF362887; Homo sapiens tropomyosin 4-anaplastic lymphoma kinase fusion protein minor isoform mRNA, partial cds.

cgagaagttgagggagaaaggcgggcccgggaacaggctgaggctgaggtggcctccttg
aaccgtaggatccagctggttgaagaagagctggaccgtgctcaggagcgtgcggaggtg
tctgaactaaaatgtggtgacctggaagaagaactcaagaatgttactaacaatctgaaa
tctctggaggctgcatctgaaaagtattctgaaaaggaggacaaatatgaagaagaaatt
aaacttctgtctgacaaactgaaagaggctgagacccgtgctgaatttgcagagagaacg
gttgcaaaactggaaaagacaattgatgacctggaagtgtacctccggaagcaccaagag
ctgcaagccatgcagatggagctgcagagccctgagtacaagctgagcaagctccgcacc
ctcgac

>embl:AF087679 AF087679; Sus scrofa tropomyosin 4 (TPM4) mRNA, complete cds.

atggccggcctcaactccctggaggcggtgaaacgcaagatccaggccctgcagcagcag
gcggacgaggcagaggatcgcgcgcagggcctgcagcgggagctggacggcgagcgcgag
cggcgggagaaagccgaaggggatgtagcagctctcaatcggcgcatccaactcgttgag
gaggagttggacagggctcaggaacgactggccacagccctgcagaagcttgaggaggca
gaaaaggctgcagatgagagcgagagaggaatgaaggtgatagaaaaccgggccatgaaa
gatgaggagaagatggagattcaggagatgcagctcaaagaggccaagcacattgccgag
gaggccgaccgcaaatacgaggaggtagctcgtaagttggtcatcctggagggcgagctg
gagagggcagaggagcgtgccgaggtgtctgaactaaaatgtggtgacctggaagaagaa
ctcaagaatgtcaccaacaacctgaagtcgctagaggctgcatctgaaaagtattctgaa
aaggaggataaatatgaagaagagattaaacttctgtctgacaaactgaaagaggctgag
acccgtgctgaatttgcagagagaacagttgcaaaactggaaaagaccatcgatgacctg
gaagaaaaacttgcccaggccaaagaagagaacgtgggcttacatcagacactggatcag
acactaaacgaactaaactgtatataaccaaaacagaagagtctcgttccatcagaaact
ccagagctacgtgtttttctcttctcttgtaagaagtttcttttgttattgcctctttgc
tttgctggaaatg

Sau khi chép xong, copy và chèn toàn bộ tệp dữ liệu chung trên vào ô nhập dữ liệu (*Enter or Paste a set of sequences in any supported format*) hay chỉ đường dẫn đến tệp dữ liệu (*Uploade a file*) rồi nhấn lệnh chuyển dữ liệu đi xử lý trực tuyến (*run*). Sau khoảng thời gian nhất định, chương trình xử lý dữ liệu trực tuyến sẽ phản hồi lại kết quả xử lý với dạng giao diện như hình 5.2.



Hình 5.2. Giao diện thông báo kết quả phân tích của ClustalW

Trong giao diện kết quả hiển thị, cần chú ý đến bốn tệp dữ liệu: *.input; *.output; *. aln và tệp *.dnd. Các tệp tin kết quả này chỉ được lưu trong đệm máy chủ sau một khoảng thời gian nhất định rồi sẽ bị xoá đi, phụ thuộc vào khả năng cung cấp của ngân hàng dữ liệu đó. Vì vậy, khi nhận được thông báo kết quả xử lý, thường người ta phải tải về máy mình để lưu giữ. Trong ba tệp dữ liệu kết quả trên, tệp kết quả dữ liệu so sánh kiểu ký tự được biểu diễn dưới dạng "*.aln" và có cấu trúc như trong các trang sau.

CLUSTAL W (1.82) multiple sequence alignment

```
embl_BF022813      ------------------------------------------------------------
embl_BF452255      ------------------------------------------------------------
embl_BG089808      ------------------------------------------------------------
embl_BG147728      ------------------------------------------------------------
embl_AF087679      ------------------------------------------------------------
embl_AF362886      ------------------------------------------------------------
embl_AF362887      ------------------------------------------------------------
embl_AF186110      ------------------------------------------------------------
embl_AF310722      CGCGCCATGGCCGGCCTCAACTCCCTGGAGGCGGTGAAACGCAAGATCCAGGCCCTGCAG 60
embl_AF186109      ---GCCATGGCCGGCCTCAACTCCCTGGAGGCGGTGAAACGCAAGATCCAGGCCCTGCAG 57
embl_BI817778      ------------------------------------------------------------
embl_BF056441      ------------------------------------------------------------
embl_BE848719      ------------------------------------------------------------

embl_BF022813      ------------------------------------------------------------
embl_BF452255      ------------------------------------------------------------
embl_BG089808      ------------------------------------------------------------
embl_BG147728      ------------------------------------------------------------
embl_AF087679      ------------------------------------------------------------
embl_AF362886      ------------------------------------------------------------
embl_AF362887      ------------------------------------------------------------
embl_AF186110      ---------------CTCTCAGCCAGGCGGATTGAAGGATGGAATTCCAACGAGGCTCCC 45
embl_AF310722      CAGCAGGCGGACGAGGCGGAAGACCGCGCGCAGGGCCTGCAGCGGGAGCTGGACGGCGAG 120
embl_AF186109      CAGCAGGCGGACGAGGCGGAAGACCGCGCGCAGGGCCTGCAGCGGGAGCTGGACGGCGAG 117
embl_BI817778      ------------------------------------------------------------
embl_BF056441      ----------------------------ACAGTTGCAAGAATCTAAAGTGTGGATTTTA 31
embl_BE848719      ------------------------------------------------------------
```

```
embl_BF022813     ------------------------------------------------------------
embl_BF452255     ------------------------------------------------------------
embl_BG089808     ------------------------------------------------------------
embl_BG147728     ------------------------------------------------------------
embl_AF087679     ------------------------------------------------------------
embl_AF362886     ------------------------------------------------------------
embl_AF362887     ------------------------------------------------------------
embl_AF186110     CCGCCTCGTCCCACCTTGGCTGAAGGTGATGTGGCCGCCCTCAACCGACGCATCCAGCTC 105
embl_AF310722     CGCGAGCGGCGCGAGAAAGCTGAAGGTGATGTGGCCGCCCTCAACCGACGCATCCAGCTC 180
embl_AF186109     CGCGAGCGGCGCGAGAAAGCTGAAGGTGATGTGGCCGCCCTCAACCGACGCATCCAGCTC 177
embl_BI817778     ------------------------------------------------------------
embl_BF056441     TTCCATTGCACAATTTGCTAGTGTATTTCCTGGGTAGTGTGGTGCTGAATAAATAGGAAT 91
embl_BE848719     ------------------------------------------------------------

embl_BF022813     ------------------------------------------------------------
embl_BF452255     ------------------------------------------------------------
embl_BG089808     ------------------------------------------------------------
embl_BG147728     ------------------------------------------------------------
embl_AF087679     ------------------------------------------------------------
embl_AF362886     ------------------------------------------------------------
embl_AF362887     ------------------------------------------------------------
embl_AF186110     GTTGAGGAGGAGTTGGACAGGGCTCAGGAACGACTGGCCACGGCCCTGCAGAAGCTGGAG 165
embl_AF310722     GTTGAGGAGGAGTTGGACAGGGCTCAGGAACGACTGGCCACGGCCCTGCAGAAGCTGGAG 240
embl_AF186109     GTTGAGGAGGAGTTGGACAGGGCTCAGGAACGACTGGCCACGGCCCTGCAGAAGCTGGAG 237
embl_BI817778     ---------------------------------------------CCGGGGTACCCTAAGCC 17
embl_BF056441     AAATGCTACTTAAGGAAAAAATAAGAGAGCTGAAAAAGCTGGTGCCATTTGAAAAAAAAA 151
embl_BE848719     ------------------------------------------------------------
```

```
embl_BF022813     ---------------------------------CCGGATCCCAGCAGAACGATTGAGCT 26
embl_BF452255     -------------------------------------GAGCCCAGCAGAACGATTGAGCT 23
embl_BG089808     -------------------------AGCTGTCGCCGGAGCCCAGCAGAACGAGTGAGCT 34
embl_BG147728     ------------------------------------------------------------
embl_AF087679     ------------------------------------------------------------
embl_AF362886     ------------------------------------------------------------
embl_AF362887     ------------------------------------------------------------
embl_AF186110     GAGGCAGAAAAAGCTGCAGATGAGAGTGAGAGAGGAATGAAGGTGATAGAAAACCGGGCC 225
embl_AF310722     GAGGCAGAAAAAGCTGCAGATGAGAGTGAGAGAGGAATGAAGGTGATAGAAAACCGGGCC 300
embl_AF186109     GAGGCAGAAAAAGCTGCAGATGAGAGTGAGAGAGGAATGAAGGTGATAGAAAACCGGGCC 297
embl_BI817778     TTCTCGGATCCGAGACTCTTCTTCCCGTTGAGGCCCCCCCCCCGCCCCCCAGCAGGGAAG 77
embl_BF056441     AGGGAAGGAATGAGATTTAACTGGTGCTCAAAGCTTCTCCGATACAAAATATTTGGTCAT 211
embl_BE848719     ------------------------------------------------------------

embl_BF022813     ATGGCCGGCCTCAACTCACTGGAGGCAGTGAAGCGCAAGATCCAGGCCCTGCAGCAGCAG 86
embl_BF452255     ATGGCCGGCCTCAACTCACTGGAGGCAGTGAAGCGCAAGATCCAGGCCCTGCAGCAGCAG 83
embl_BG089808     ATGGCCGGCCTCAACTCACTGGAGGCAGTGAAGCGCAAGATCCAGGCCCTGCAGCAGCAG 94
embl_BG147728     -----------CAACTCACTGGAGGCAGTGAAGCGCAAGATCCAGGCCCTGCAGCAGCAG 49
embl_AF087679     ATGGCCGGCCTCAACTCCCTGGAGGCGGTGAAACGCAAGATCCAGGCCCTGCAGCAGCAG 60
embl_AF362886     ------------------------------------------------------------
embl_AF362887     ---------------------------------CGAGAAGTTGAGGGAGAAAGGCGGGCC 27
embl_AF186110     ATGAAGGATGAGGAGAAGATGGAGATTCAGGAGATGCAGCTCAAAGAGGCCAAGCACATT 285
embl_AF310722     ATGAAGGATGAGGAGAAGATGGAGATTCAGGAGATGCAGCTCAAAGAGGCCAAGCACATT 360
embl_AF186109     ATGAAGGATGAGGAGAAGATGGAGATTCAGGAGATGCAGCTCAAAGAGGCCAAGCACATT 357
embl_BI817778     ATGTCGGGTGGCAGTTCCATCGATGCGGTGAAGAAGAAGATCCAGAGCCTTCAGCAGGTG 137
embl_BF056441     GTATTCATAATTTGCTTGACATTTCCAGCAAAGCGAAGATGGCAATAACAAAAGGAACTT 271
embl_BE848719     -TGTTACCAATCTGCTTGGCATTTCCTGCAAGGTGGAAACC-TGGTAATAAGCGGAACTT 58
```

```
embl_BF022813     GCAGACGACGCAGAGGATCGCGCGCAAGGCCTGCAGCGCGAACTGGATGGCGAGCGCGAG 146
embl_BF452255     GCAGACGACGCAGAGGATCGCGCGCAAGGCCTGCAGCGCGAACTGGATGGCGAGCGCGAG 143
embl_BG089808     GCAGACGACGCAGAGGATCGCGCGCAAGGCCTGCAGCGCGAACTGGATGGCGAGCGCGAG 154
embl_BG147728     GCAGACGACGCANAGGATCGCGCGCAAGGCCTGCAGCGCGAACTGGATGGCGAGCTCTAG 109
embl_AF087679     GCGGACGAGGCAGAGGATCGCGCGCAGGGCCTGCAGCGGGAGCTGGACGGCGAGCGCGAG 120
embl_AF362886     ------------------------------------------------------------
embl_AF362887     CGGGAACAGGCTGAGGCTGAGGTGGCCTCCTTGAACCGTAGGATCCAGCTGGTTGAAGAA 87
embl_AF186110     GCGGAAGAGGCTGACCGCAAATACGAGGAGGTAGCTCGTAAGCTGGTCATCCTGGAGGGT 345
embl_AF310722     GCGGAAGAGGCTGACCGCAAATACGAGGAGGTAGCTCGTAAGCTGGTCATCCTGGAGGGT 420
embl_AF186109     GCGGAAGAGGCTGACCGCAAATACGAGGAGGTAGCTCGTAAGCTGGTCATCCTGGAGGGT 417
embl_BI817778     GCGGACGAGGCCGAAGAGCGGGCCGAGATCCTGCAGAGGGAGGTGGACGCCGAGAGGCAG 197
embl_BF056441     CTTACAAGAGAAGAGAAAGACCCACGGAGCTC----CA-GAGTTTCTGTTGGAACAAGAC 326
embl_BE848719     CTTACAAAAGAGGAAGACAGGGCACACTCTCTGGAGTG-GAGTTGGTGTTAAAACAGTAC 117

embl_BF022813     CGGCGCGAGAAAGCTGAAGGAGATGCAGCCGCTCTCAACCGCCGCATCCAACTGCTGGAG 206
embl_BF452255     CGGCGCGAGAAAGCTGAAGGAGATGCAGCCGCTCTCAACCGCCGCATCCAACTGCTGGAG 203
embl_BG089808     CGGCGCGAGAAAGCTGAAGGAGATGCAGCCGCTCTCAACCGCCGCATCCAACTGCTGGAG 214
embl_BG147728     CGGCGCGAGAAAGCTGAGGGAGAGGGGGGCGCTCTCAACCGCCGCATCCAACTGCTGGAG 169
embl_AF087679     CGGCGGGAGAAAGCCGAAGGGGATGTAGCAGCTCTCAATCGGCGCATCCAACTCGTTGAG 180
embl_AF362886     ------------------------------CTGGCAGAGTCCCGTTGCCGAGAGATGGAT 30
embl_AF362887     GAGCTGGACCGTGCTCAGGAGCGTGCGGAGGTGTCTGAACTAAAATGTGGTGACCTGGAA 147
embl_AF186110     GAGCTGGAGAGGGCAGAGGAGCGTGCGGAGGTGTCTGAACTAAAATGTGGTGACCTGGAA 405
embl_AF310722     GAGCTGGAGAGGGCAGAGGAGCGTGCGGAGGTGTCTGAACTAAAATGTGGTGACCTGGAA 480
embl_AF186109     GAGCTGGAGAGGGCAGAGGAGCGTGCGGAGGTGTCTGAACTAAAATGTGGTGACCTGGAA 477
embl_BI817778     TGCAGGGAGCGGGCCGAGGCAGACGTGGGATCGCTCAACCGCCGCATCCAGCTGGTAGAG 257
embl_BF056441     TCTTCTGTTTT-GCTTATATACAGTTAAGTTCGTTTAGTGTCTG-ATCCAGTGTCTGATG 384
embl_BE848719     TCTTCTGGTTTAGTTTATATACAGTTAAGTTCGTTTAGTGTCTG-GTCCAGTGTCTGATG 176
                                                                            *
```

```
embl_BF022813     GAGGAACTGGACCGGGCTCAGGAGCAGCTGGCCACAGCCCTGCAGAATCTGGAAGAGGCA 266
embl_BF452255     GAGGAACTGGACCGGGCTCAGGAGCAGCTGGCCACAGCCCTGCAGAATCTGGAAGAGGCA 263
embl_BG089808     GAGGAACTGGACCGGGCTCAGGAGCAGCTGGCCACAGCCCTGCAGAATCTGGAAGAGGCA 274
embl_BG147728     GAGGAACTGGACCGGGCTCATGAGCAGCTGGCCACAGCCCTGCAGAATCTGGAAGAGGCA 229
embl_AF087679     GAGGAGTTGGACAGGGCTCAGGAACGACTGGCCACAGCCCTGCAGAAGCTTGAGGAGGCA 240
embl_AF362886     GAGCAGATTAG-----------ACTGATGGACCAGAACC-TGAAGTGTCTGAGTGCTGCT 78
embl_AF362887     GAAGAACTCAA-----------GAATGTTACTAACAATC-TGAAATCTCTGGAGGCTGCA 195
embl_AF186110     GAAGAACTCAA-----------GAATGTTACTAACAATC-TGAAATCTCTGGAGGCTGCA 453
embl_AF310722     GAAGAACTCAA-----------GAATGTTACTAACAATC-TGAAATCTCTGGAGGCTGCA 528
embl_AF186109     GAAGAACTCAA-----------GAATGTTACTAACAATC-TGAAATCTCTGGAGGCTGCA 525
embl_BI817778     GAGGAGCTGGACCGTGCCCAGGAGCGCCTTGCCACTGCCCTGCTGAAGTTGGAGGAGGCG 317
embl_BF056441     TAAGCCCACGTTCTCTTCTTTGGCCTGGGCAAGTTTCTCTTCCAGGTCATCAATTGTCTT 444
embl_BE848719     TAAGCCCACATTCTCTTCTTTGGCCTGGGCAAGTTTTTCTTCCAGGTCATCGATTGTCTT 236
                   *                                     * *        *

embl_BF022813     GA-GAAGGCTGCTGATGAGAGTGAGAGAGGCATGAAGGTAATAGAGAACCGAGCCATGAA 325
embl_BF452255     GA-GAAGGCTGCTGATGAGAGTGAGAGAGGCATGAAGGTAATAGAGAACCGAGCCATGAA 322
embl_BG089808     GA-GAAGGCTGCTGATGAGAGTGAGAGACGCATGAAGGTAATAGAGAACCGAGCCATGAA 333
embl_BG147728     GA-GAAGGCTGCTGATGAGAGTGAGAGAGGCATGAAGGTAATAGAGAACCGAGCCATGAA 288
embl_AF087679     GA-AAAGGCTGCAGATGAGAGCGAGAGAGGAATGAAGGTGATAGAAAACCGGGCCATGAA 299
embl_AF362886     GAAGAAAAGTACTCTCAAAAAGAAGATAAATATGAGGAAGAAATCAAGATTCTTACTGAT 138
embl_AF362887     TCTGAAAAGTATTCTGAAAAGGAGGACAAATATGAAGAAGAAATTAAACTTCTGTCTGAC 255
embl_AF186110     TCTGAAAAGTATTCTGAAAAGGAGGACAAATATGAAGAAGAAATTAAACTTCTGTCTGAC 513
embl_AF310722     TCTGAAAAGTATTCTGAAAAGGAGGACAAATATGAAGAAGAAATTAAACTTCTGTCTGAC 588
embl_AF186109     TCTGAAAAGTATTCTGAAAAGGAGGACAAATATGAAGAAGAAATTAAACTTCTGTCTGAC 585
embl_BI817778     GA-GAAAGCTGCAGACGAGAGTGAACGAGGCATGAAGGTCATTGAAAACCGAGCCACCAA 376
embl_BF056441     TTCCAGTTTTGCAACCGTTCTCTCTGCAAAT-TCAGCACGGGTCTCAGCCTCTTTCAGTT 503
embl_BE848719     CTCCAGTTTAGAAACTGTCCTTTCTGCAAAC-TCAGCTCGGGTCTCAGCCTCCTTCAGCT 295
                     *                     *    * *           *
```

```
embl_BF022813     AGATGAGGAAAAGATGGAGATCCTGGAGATGCAGCT-CAAAGAAGCCAAGCACATCACTG 384
embl_BF452255     AGATGAGGAAAAGATGGAGATCCTGGAGATGCAGCT-CAAAGAAGCCAAGCACATCACTG 381
embl_BG089808     AGATGAGGAAAAGATGGAGATCCTGGAGATGCAGCT-CAGAGAAGCCAAGCACATCACTG 392
embl_BG147728     AGATGAGGAAAAGATGGAGATCCTGGAGATGCAGCT CAAAGAAGCCAAGCACATCACTG 347
embl_AF087679     AGATGAGGAGAAGATGGAGATTCAGGAGATGCAGCT-CAAAGAGGCCAAGCACATTGCCG 358
embl_AF362886     AAACT-CAAGGAGGCAGAGACCC--GTGCTGAATTTGCAGAGAGAACGGTTGCAAAACTG 195
embl_AF362887     AAACT-GAAAGAGGCTGAGACCC--GTGCTGAATTTGCAGAGAGAACGGTTGCAAAACTG 312
embl_AF186110     AAACT-GAAAGAGGCTGAGACCC--GTGCTGAATTTGCAGAGAGAACGGTTGCAAAACTG 570
embl_AF310722     AAACT-GAAAGAGGCTGAGACCC--GTGCTGAATTTGCAGAGAGAACGGTTGCAAAACTG 645
embl_AF186109     AAACT-GAAAGAGGCTGAGACCC--GTGCTGAATTTGCAGAGAGAACGGTTGCAAAACTG 642
embl_BI817778     GGACGAGGAGAAGATGGAGATCCAGGAGATGCAGTT-GAAAGAGGCCAAACACATAGCAG 435
embl_BF056441     TGTCAGACAGAAGTTTAATTTCTTCTTCATATTTGTCCTCCTTTTCAGAATACTTTTCAG 563
embl_BE848719     TGTCAGACAGAAGCTTGATTTCTTCTTCATATTTATCCTCCTTTTCAGAATACTTTTCAG 355
                          *  **    *            *     *                 *    * *

embl_BF022813     ACGAAGCCGACCGCAAGTTTGAGGAGGTTGCTCGT------------------------- 419
embl_BF452255     ACGAAGCCGACCGCAAGTANTGAGAGGTTGCTCGTAAGTTGGTCATCCTGGAGGGTGAGC 441
embl_BG089808     ATGAAGCCGACCGCAAGTATGATGAGGTTGCTCGTAAGTTGGTCATCCTGGAGGGTGAGC 452
embl_BG147728     ATGAAGCCGACCGCAAGTATGAGGAGGTTGCTCGTAAGTTGGTCATCCTGGAGGGTGAGC 407
embl_AF087679     AGGAGGCCGACCGCAAATACGAGGAGGTAGCTCGTAAGTTGGTCATCCTGGAGGGCGAGC 418
embl_AF362886     GAAAAGACAATTGATGACCTGGAAGTGTACCGCCGGAAGCACCAGGAGCTGCAAGCCATG 255
embl_AF362887     GAAAAGACAATTGATGACCTGGAAGTGTACCTCCGGAAGCACCAAGAGCTGCAAGCCATG 372
embl_AF186110     GAAAAGACAATTGATGACCTGGAAGTGTACCGCCGGAAGCACCAGGAGCTGCAAGCCATG 630
embl_AF310722     GAAAAGACAATTGATGACCTGGAAGTGTACCGCCGGAAGCACCAGGAGCTGCAAGCCATG 705
embl_AF186109     GAAAAGACAATTGATGACCTGGAAGTGTACCGCCGGAAGCACCAGGAGCTGCAAGCCATG 702
embl_BI817778     AGGAGGCCGACCGCAAA------------------------------------------- 452
embl_BF056441     ATGCAGCCTCCAGAGATTTCAGATTGT-AGTTACAATTCTGAGTTCTTCT-CCAGGTCAC 621
embl_BE848719     AAGCAGCCTCCAGTGATTTCAGATTGTTAGTTACATTCTTGAGCTCTTCTTCCAGGTCAC 415
                       * *    *
```

```
embl_BF022813     ------------------------------------------------------------
embl_BF452255     TGAAGAGAGCAGAGGAGAGGGCG-AGGTATCTGAACTAAAGT-TGGTGACCTGGAG-AAG 498
embl_BG089808     TGAAGAGAGCAGATGAGCGGGCGGAGGTATCTGAACTAAAGTGTGGTGACCTGTAATAAG 512
embl_BG147728     TGAAGAGAGCATAGGAGCGGGCGGAGGTATCTGAACTAAAGTGTGGTGACCCTGAAGAAG 467
embl_AF087679     TGGAGAGGGCAGAGGAGCGTGCCGAGGTGTCTGAACTAAAATGTGGTGACCTGGAAGAAG 478
embl_AF362886     CAGATGGAGCTGCAGAGCCCTGAGTACAAGCTGAGC--AAGCTCCGCACC-TCGAC---- 308
embl_AF362887     CAGATGGAGCTGCAGAGCCCTGAGTACAAGCTGAGC--AAGCTCCGCACCCTCGAC---- 426
embl_AF186110     CAGATGGAGCTGCAGAGCCCTGAGTACAAGCTGAGC--AAGCTCCGCACC-TCGACCATC 687
embl_AF310722     CAGATGGAGCTGCAGAGCCCTGAGTACAAGCTGAGC--AAGCTCCGCACC-TCGACCATC 762
embl_AF186109     CAGATGGAGCTGCA---------------------------------------------- 716
embl_BI817778     ------------------------------------------------------------
embl_BF056441     CACATTTTA-TTCAGACACCTCCGCACGCTTCTCTGCCCTCT-CAGCTAACCCTTC---- 675
embl_BE848719     CACACTTTAGTTCAGATACCTCCGCCCTCTCCTCTGCTCTCTTCAGCTCACCCTCCAGGA 475

embl_BF022813     ------------------------------------------------------------
embl_BF452255     AGCTCAAGAATGTAACTAAC---------------------------------------- 518
embl_BG089808     AGCTCATGAATGTAACTAACAATCTGAAATGACTGGAGGCTGTATTTGAGAAGTATTCTG 572
embl_BG147728     AGCTCAAGAATGTAACTAACAATCTGAAATCACTGGAGGCTGCTTCTGAAAAGTATTCTG 527
embl_AF087679     AACTCAAGAATGTCACCAACAACCTGAAGTCGCTAGAGGCTGCATCTGAAAAGTATTCTG 538
embl_AF362886     ------------------------------------------------------------
embl_AF362887     ------------------------------------------------------------
embl_AF186110     ATGACCGACTACAACCCCAACTACTGCTTTGCTGGCAAGACCTCCTCCATCAGTGACCTG 747
embl_AF310722     ATGACCGACTACAACCCCAACTACTGCTTTGCTGGCAAGACCTCCTCCATCAGTGACCTG 822
embl_AF186109     ------------------------------------------------------------
embl_BI817778     ------------------------------------------------------------
embl_BF056441     ------------------------------------------------------------
embl_BE848719     TGACCAACTTACGAGCAACCTCCTCATACTTGCGGTCGGCTACGTCAGTGATGTGCTTGG 535
```

```
embl_BF022813     ------------------------------------------------------------
embl_BF452255     ------------------------------------------------------------
embl_BG089808     AATAGGAGGATTAGTATGAAGAAGAAGATAAGCTTATGCCTGATAAGCTGAAGGTAGGTG 632
embl_BG147728     AAAAGGAG-------------------------------------------------- 535
embl_AF087679     AAAAGGAGGATAAATATGAAGAAGAGATTAAACTTCTGTCTGACAAACTGAAAGAGGCTG 598
embl_AF362886     ------------------------------------------------------------
embl_AF362887     ------------------------------------------------------------
embl_AF186110     AAGGAGGTGCCGCGGAAAAACATCACCCTCATTCGGGGTCTGGGCCATGGCGCCTTTGGG 807
embl_AF310722     AAGGAGGTGCCGCGGAAAAACATCACCCTCATTCGGGGTCTGGGCCATGGCGCCTTTGGG 882
embl_AF186109     ------------------------------------------------------------
embl_BI817778     ------------------------------------------------------------
embl_BF056441     ------------------------------------------------------------
embl_BE848719     CTGCTTTGAGCTGCATCTCCAGGATCTACATCTTTTGCTCATCTTTCATGGCTTCGTTCT 595

embl_BF022813     ------------------------------------------------------------
embl_BF452255     ------------------------------------------------------------
embl_BG089808     GAAACCAGTCTGGATTTGCAGACAGA--------------------------------- 658
embl_BG147728     ------------------------------------------------------------
embl_AF087679     AGACCCGTGCTGAATTTGCAGAGAGAACAGTTGCAAAACTGGAAAAGACCATCGATGACC 658
embl_AF362886     ------------------------------------------------------------
embl_AF362887     ------------------------------------------------------------
embl_AF186110     GAGGTGTATGAAGGCCAGGTGTCCGGAATGCCCAACGACCCAAGCCCCCTGCAAGTGGCT 867
embl_AF310722     GAGGTGTATGAAGGCCAGGTGTCCGGAATGCCCAACGACCCAAGCCCCCTGCAAGTGGCT 942
embl_AF186109     ------------------------------------------------------------
embl_BI817778     ------------------------------------------------------------
embl_BF056441     ------------------------------------------------------------
embl_BE848719     CTATTACCTTCATGCCTCTCTCACTCTCATCAGCAGCCCTCTCTGCCTCTTTCAGATTCT 655
```

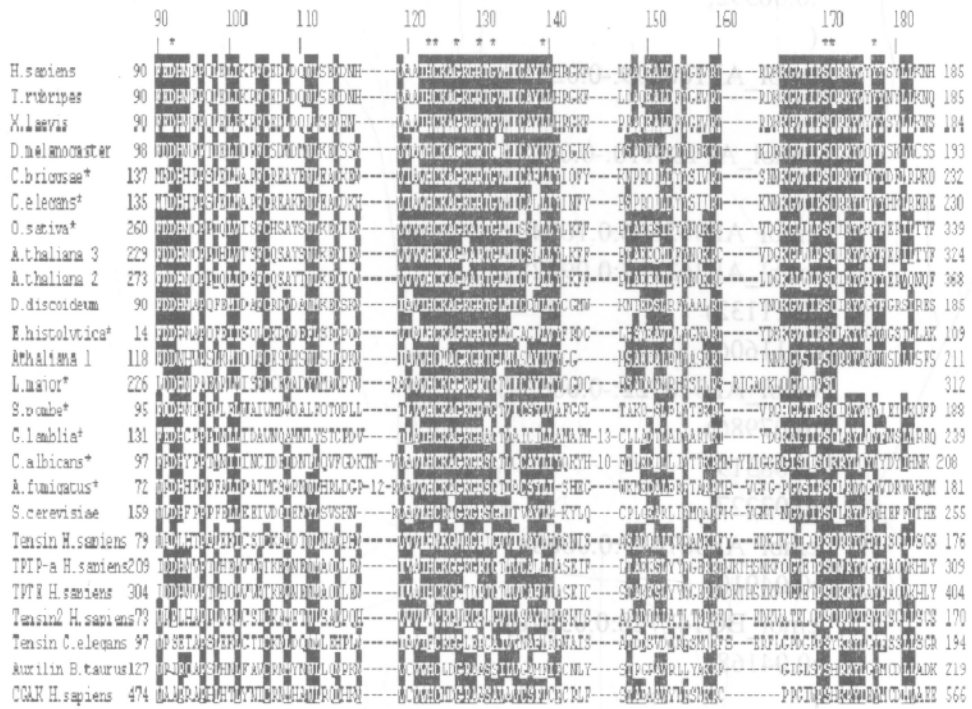
```
embl_BF022813     ------------------------------------------------------------
embl_BF452255     ------------------------------------------------------------
embl_BG089808     ------------------------------------------------------------
embl_BG147728     ------------------------------------------------------------
embl_AF087679     TGGAAGAAAAACTTGCCCAGGCCAAAGAAGAGAACGTGGGCTTACATCAGACACTGGATC 718
embl_AF362886     ------------------------------------------------------------
embl_AF362887     ------------------------------------------------------------
embl_AF186110     GTGAAGACGCTGCCTG------------------------------------------- 883
embl_AF310722     GTGAAGACGCTGCCTGAAGTGTGC----------------------------------- 966
embl_AF186109     ------------------------------------------------------------
embl_BI817778     ------------------------------------------------------------
embl_BF056441     ------------------------------------------------------------
embl_BE848719     GCAAGGCTGTGGCCAGCTGCTTCTGACCCTGTCCAATTCCTTC---------------- 698

embl_BF022813     ------------------------------------------------------------
embl_BF452255     ------------------------------------------------------------
embl_BG089808     ------------------------------------------------------------
embl_BG147728     ------------------------------------------------------------
embl_AF087679     AGACACTAAACGAACTAAACTGTATATAACCAAAACAGAAGAGTCTCGTTCCATCAGAAA 778
embl_AF362886     ------------------------------------------------------------
embl_AF362887     ------------------------------------------------------------
embl_AF186110     ------------------------------------------------------------
embl_AF310722     ------------------------------------------------------------
embl_AF186109     ------------------------------------------------------------
embl_BI817778     ------------------------------------------------------------
embl_BF056441     ------------------------------------------------------------
embl_BE848719     ------------------------------------------------------------
```

```
embl_BF022813      ------------------------------------------------------------ ----
embl_BF452255      ------------------------------------------------------------
embl_BG089808      ------------------------------------------------------------
embl_BG147728      ------------------------------------------------------------
embl_AF087679      CTCCAGAGCTACGTGTTTTTCTCTTCTCTTGTAAGAAGTTTCTTTTGTTATTGCCTCTTT 838
embl_AF362886      ------------------------------------------------------------
embl_AF362887      ------------------------------------------------------------
embl_AF186110      ------------------------------------------------------------
embl_AF310722      ------------------------------------------------------------
embl_AF186109      ------------------------------------------------------------
embl_BI817778      ------------------------------------------------------------
embl_BF056441      ------------------------------------------------------------
embl_BE848719      ------------------------------------------------------------

embl_BF022813      ---------------
embl_BF452255      ---------------
embl_BG089808      ---------------
embl_BG147728      ---------------
embl_AF087679      GCTTTGCTGGAAATG 853
embl_AF362886      ---------------
embl_AF362887      ---------------
embl_AF186110      ---------------
embl_AF310722      ---------------
embl_AF186109      ---------------
embl_BI817778      ---------------
embl_BF056441      ---------------
embl_BE848719      ---------------
```

Trong tệp kết quả so sánh dạng ký tự trên sẽ xuất hiện thêm dòng kết quả dưới cùng, với các ký tự: "*" biểu diễn vị trí đồng nhất hoàn toàn giữa tất cả các chuỗi, ký hiệu ":" biểu diễn vị trí có sự sai lệch nhất định và ký hiệu "." biểu diễn vị trí có sự sai lệch tương đối lớn hơn giữa các chuỗi với nhau. Khi số lượng chuỗi so sánh tăng cao thì xác suất xuất hiện các vị trí tương đồng "*", ":" hay "." sẽ giảm đi, gây khó khăn cho việc quan sát và đánh giá trực quan độ tương đồng (hay phân ly) giữa chúng. Để dễ dàng nhận diện hơn, chương trình còn sử dụng phương pháp đánh dấu bằng màu sắc khác nhau hay thay đổi cường độ màu để biểu diễn quan hệ tương đồng giữa các chuỗi (xem chế độ JAVAWIEW - hình 5.3).

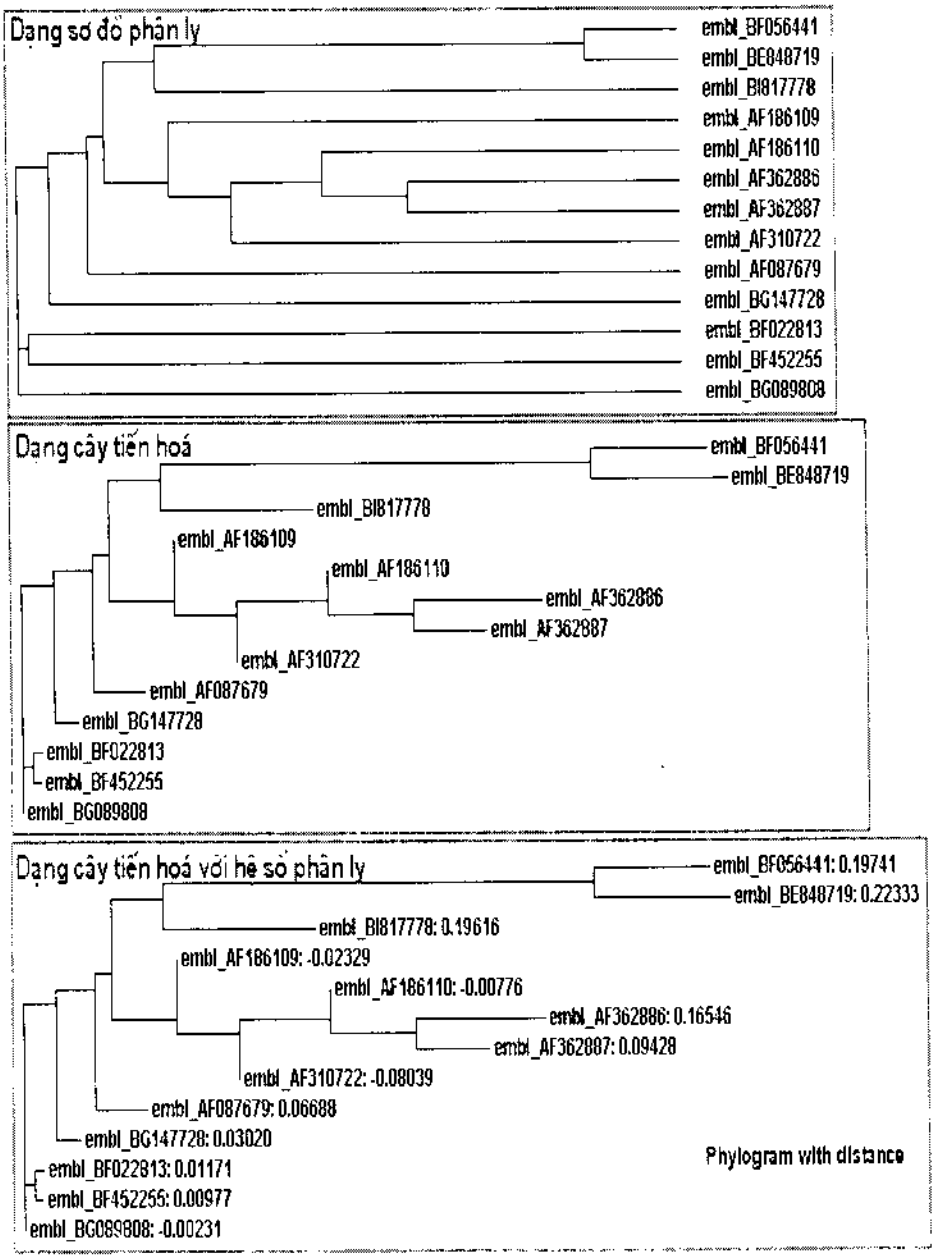


Hình 5.3. Giao diện kết quả xử lý cấu trúc chuỗi trên CLUSTALW

Kết quả so sánh hiển thị dưới dạng cây tương quan (dạng *.dnd – *Guide tree file*) có thể dễ dàng hiển thị các kiểu giao diện khác nhau: dạng

ký tự văn bản, dạng sơ đồ phân ly (*Cladogram*), dạng tương quan hình cây (*Phylogram*) đơn giản hay kèm theo hệ số phân ly (xem hình 5.4). Kết quả biểu diễn sơ đồ phân ly dạng ký tự có giao diện như sau:

```
(
(
(
(
(
(
 embl_BF056441:0.19741,
embl_BE848719:0.22333)
:0.55475,
 embl_BI817778:0.19616)
:0.06592,
(
embl_AF186109:-0.02329,
(
 embl_AF186110:-0.00776,
(
embl_AF362886:0.16546,
embl_AF362887:0.09428)
:0.11132)
:0.11606,
embl_AF310722:-0.08039)
:0.07986)
:0.08319)
:0.02227,
embl_AF087679:0.06688)
:0.04910,
embl_BG147728:0.03020)
:0.04168,
(
embl_BF022813:0.01171,
embl_BF452255:0.00977)
:0.01247,
embl_BG089808:-0.00231);
```



*Hình 5.4. Các dạng giao diện hiển thị kiểu tệp kết quả dạng *.dnd*

Trước khi gửi tệp tin đi xử lý, phục vụ cho việc trả lời kết quả, người sử dụng cần xác định thêm các yêu cầu cho chế độ xử lý sau:

- Đặt chế độ cho kiểu tệp dữ liệu lấy ra, với phương án lựa chọn dưới một trong các định dạng sau: ALN, GCG, PIR, PHYLIP và GDE. Chế độ này do người yêu cầu tự lựa chọn trên cửa sổ "*Output format*" trước khi gửi thông tin đi xử lý.
- Yêu cầu về trật tự sắp xếp các chuỗi trong tệp kết quả, với phương án theo trình tự gửi dữ liệu đi hay theo quan hệ tương quan về khoảng cách phân ly giữa các chuỗi khi xử lý (*output order*).
- Đặt thêm các thông số phụ, khi lựa chọn chế độ so sánh nhanh từng cặp (Alignment fast pairwise) như:
  - KTUP để lựa chọn số ký tự khi xử lý so sánh.
  - WINDOWS để lựa chọn kích thước mảng xử lý.
  - SCORE để đặt chế độ tỉ lệ khi tính kết quả.
  - TOPDIAG đặt chế độ so sánh chéo.
  - PAIRGAP đặt chế độ khoảng đứt (hay chèn) giới hạn...

Lựa chọn kiểu thuật toán xử lý:

- BLOSUM là kiểu ma trận thích hợp nhất để xác định độ tương đồng của chuỗi. Kích thước ma trận sử dụng trong chương trình là: Blosum80, 62, 40 và 30.
- PAM được sử dụng rộng rãi từ cuối thập kỷ bảy mươi của thế kỷ XX. Kích thước ma trận được sử dụng là: 120, 160, 250 và 350.
- GONNET tương tự như PAM, nhưng cập nhật nâng cấp thường xuyên hơn nên độ nhạy cao hơn. Kích thước mảng được sử dụng là: 40, 80, 120, 160, 250 và 350.

- Lựa chọn khoảng trống ký tự giới hạn GAP cho thuật toán, với các tham số: Gapopen, Endgap, Gapext, Gapdist...

*** Trong trường hợp không sử dụng dịch vụ xử lý trực tuyến, người ta có thể tải chương trình ClustalW miễn phí về máy cá nhân từ nhiều ngân hàng dữ liệu khác nhau, thí dụ EBI hay DDBJ.

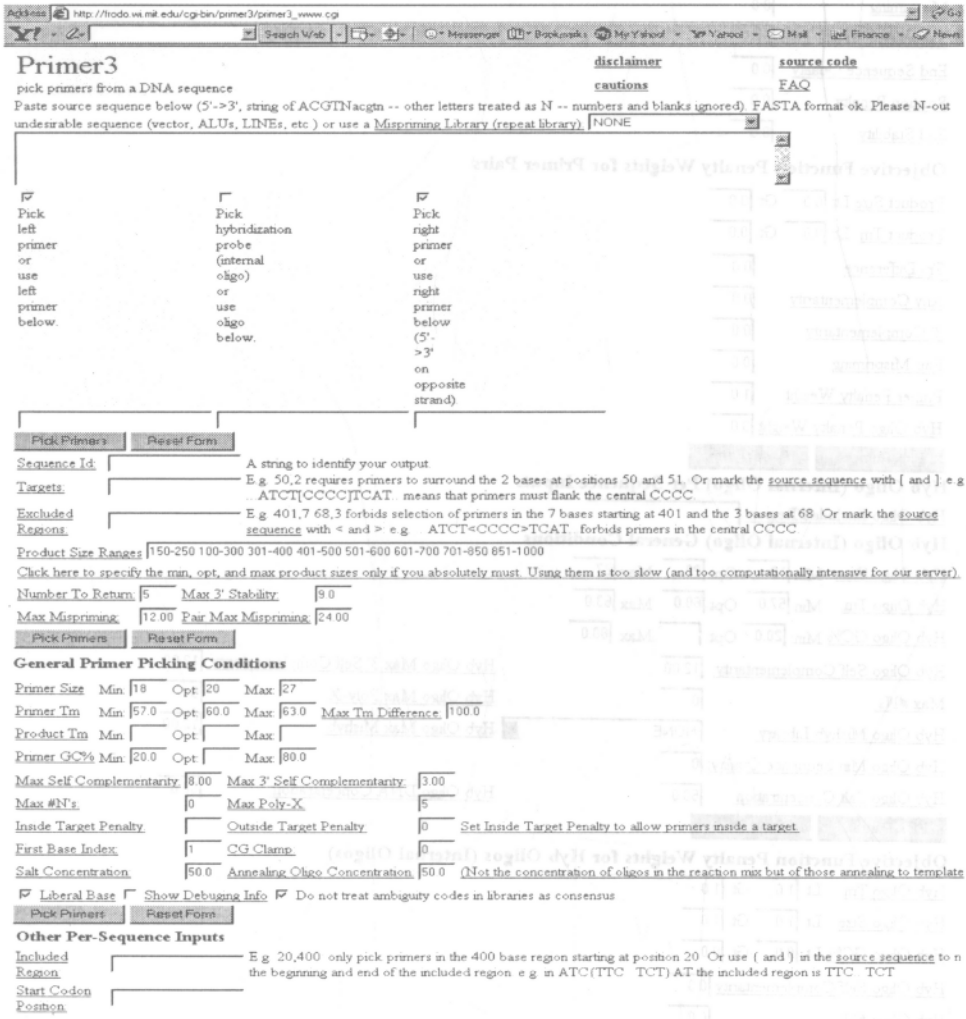
6. CHƯƠNG TRÌNH THIẾT KẾ VÀ LỰA CHỌN ĐOẠN MỒI PRIMER3

6.1. Đại cương

Chương trình thiết kế và lựa chọn đoạn mồi (*Primer Design*) là chương trình tìm kiếm và lựa chọn xác định đoạn nucleotide tương đồng với cấu trúc chuỗi phân tích, phục vụ cho kỹ thuật nhân gen PCR hay sử dụng cho nhiều kỹ thuật lai ứng dụng khác nhau. Để giải quyết nhiệm vụ trên, nhiều phần mềm đã được xây dựng và cung cấp cho người sử dụng (bao gồm cả phần mềm miễn phí và loại phải trả tiền), thí dụ: OLIGO Primer Analysis Software (http://www.oligo.net/ - *Molecular Biology Insights, Inc.*), OLIGO® (http://www.medprobe.com/no/oligo.html - *Molecular Biology Insights, Inc.*), Oligo Perfect™ Designer (http://www.invitrogen.com - *Invitrogen Corp.*), Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi - *Whitehead Institute for Biomedical Research*)...

Primer3 là một trong số các chương trình ứng dụng để thiết kế đoạn mồi (định hướng vào việc thiết kế mồi phục vụ cho kỹ thuật nhân gen PCR và LA-PCR) của Rozen và Skeletsky, viết cho môi trường Sun, OS, Unix và Linux (chương trình này không chạy trên môi trường Windows). *Whitehead Institute for Biomedical Research* đã cung cấp miễn phí chương trình xử lý trực tuyến trên cho người sử dụng, qua địa chỉ truy cập: http://fokker.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi. Cơ sở tính toán lập trình cho chương trình này dựa vào các thông số chính là: nhiệt độ gắn mồi, kích thước mồi, hàm lượng GC trong thành phần, khả năng bắt cặp

dimer, trị số nhiệt động và cấu trúc không gian bậc hai của đoạn mồi, kích thước sản phẩm PCR..., trên cơ sở dữ liệu phân tích của các đoạn mồi tương ứng đã biết trong các ngân hàng dữ liệu. Giao diện trực tuyến của chương trình Primer3 có dạng như sau:



Hình 6.1. Giao diện trực tuyến của chương trình Primer3 (còn tiếp)

(Chú ý: giao diện trên có thể thay đổi, phụ thuộc vào thời điểm truy cập)

Objective Function Penalty Weights for Primers

Tm Lt 1.0 Gt 1.0
Size Lt 1.0 Gt 1.0
GC% Lt 0.0 Gt 0.0
Self Complementarity 0.0
3' Self Complementarity 0.0
#N's 0.0
Mispriming 0.0
Sequence Quality 0.0
End Sequence Quality 0.0
Position Penalty 0.0
End Stability 0.0

Objective Function Penalty Weights for Primer Pairs

Product Size Lt 0.0 Gt 0.0
Product Tm Lt 0.0 Gt 0.0
Tm Difference 0.0
Any Complementarity 0.0
3' Complementarity 0.0
Pair Mispriming 0.0
Primer Penalty Weight 1.0
Hyb Oligo Penalty Weight 0.0

Pick Primers | Reset Form

Hyb Oligo (Internal Oligo) Per-Sequence Inputs

Hyb Oligo Excluded Region:

Hyb Oligo (Internal Oligo) General Conditions

Hyb Oligo Size: Min 18 Opt 20 Max 27
Hyb Oligo Tm: Min 57.0 Opt 60.0 Max 63.0
Hyb Oligo GC% Min: 20.0 Opt: Max 80.0
Hyb Oligo Self Complementarity 12.00
Hyb Oligo Max 3' Self Complementarity 12.00
Max #Ns 0
Hyb Oligo Max Poly-X 5
Hyb Oligo Mishyb Library NONE
Hyb Oligo Max Mishyb 12.00
Hyb Oligo Min Sequence Quality 0
Hyb Oligo Salt Concentration 50.0
Hyb Oligo DNA Concentration 50.0

Pick Primers | Reset Form

Objective Function Penalty Weights for Hyb Oligos (Internal Oligos)

Hyb Oligo Tm Lt 1.0 Gt 1.0
Hyb Oligo Size Lt 1.0 Gt 1.0
Hyb Oligo GC% Lt 0.0 Gt 0.0
Hyb Oligo Self Complementarity 0.0
Hyb Oligo #N's 0.0
Hyb Oligo Mishybing 0.0
Hyb Oligo Sequence Quality 0.0

Pick Primers | Reset Form

Copyright Notice and Disclaimer

(Phần nửa dưới của giao diện chương trình Primer3 - Tiếp hình 6.2)

6.2. Thao tác sử dụng chương trình

Việc thao tác sử dụng chương trình trên có thể tóm tắt qua các bước chính sau: kết nối mạng internet, hiển thị giao diện trang chủ Primer3, nhập dữ liệu, đặt chế độ xử lý (được xác định qua việc lựa chọn giá trị khi đặt các chế độ xử lý tương ứng), sau đó nhấn cửa sổ "**Pick primers**" để gửi dữ liệu đi xử lý trực tuyến. Sau khoảng thời gian chờ, phụ thuộc vào tốc độ đường truyền của mạng kết nối, người xử lý sẽ nhận lại được kết quả xử lý của chương trình Primer3 (xem phần thí dụ phía dưới).

- Trong tệp dữ liệu kết quả, có thể xảy ra hai khả năng: chương trình không lựa chọn được đoạn mồi thoả mãn với các các thông số đã chọn. Trong trường hợp này, người xử lý quay ngược trở lại giao diện nhập dữ liệu để thay đổi các thông số đầu vào rồi gửi đi xử lý tiếp, các bước lặp lại như quy trình ban đầu, cho đến khi xác định được các đoạn mồi mong muốn.

- Chương trình Primer3 lựa chọn được đoạn mồi phù hợp nhất cho yêu cầu người gửi tin (thường là sau một số lần gửi và chỉnh sửa lại thông tin đầu vào. Đương nhiên, người ta vẫn có thể nhận được kết quả mong muốn ngay sau lần yêu cầu đầu tiên).

Các thao tác chính và thông số lựa chọn ban đầu bao gồm:

A/ **Nhập dữ liệu**: là thao tác chèn chuỗi dữ liệu được chọn làm khuôn để thiết kế mồi vào trong ô nhập dữ liệu ở đầu giao diện. Chương trình xử lý chỉ chấp nhận chuỗi ký tự viết theo định dạng FASTA hay lấy trực tiếp chuỗi ký tự tên các cặp bazơ nitơ thu được khi giải trình tự cấu trúc chuỗi, dưới dạng sau "...ACTGNacgtn..." (Với chuỗi kết quả giải trình

tự, trước khi chèn vào cửa sổ nhập cần kiểm tra để thay ký tự N bằng một ký tự nucleotit bất kỳ, hay chọn chế độ mispriming library).

B/ **Đặt chế độ xử lý**: Trong mục này, người phân tích phải lựa chọn xác định hàng loạt thông số khác nhau bao gồm:

+ **Sequence Id**: Đặt tên chuỗi nhận dạng đầu ra để lựa chọn đoạn mồi hay đoạn lai ghép.

+ **Targets**: Là thông số xác định vị trí đoạn mồi, được viết dưới dạng hai cụm số hay khung ký tự. Thí dụ: "Targets: 50,2" có nghĩa đoạn mồi phải nằm sát vị trí 50 hay 51; hoặc đánh dấu ngoặc vuông trên chuỗi "ATAC[CCCC]TACT..." nghĩa là đoạn mồi phải nằm sát một vị trí trong đoạn được đánh dấu khung. Vị trí đích của đoạn mồi thường là trong các vùng bảo thủ cấu trúc của chuỗi (theo kết quả trên chương trình phân tích quy luật vận động ccủa nhóm chuỗi cùng nguồn, thí dụ CLUSTALW).

+ **Excluded Regions**: Là vùng không được lựa chọn đoạn mồi, đánh dấu bằng hai giá trị: khởi đầu và độ dài. Thí dụ "Excluded Regions: 120,42" nghĩa là đoạn mồi lựa chọn không được chứa các cặp nucleotide tương ứng với 42 ký tự, tính từ vị trí thứ 120. Vùng loại trừ không thiết kế mồi thường là các vùng phân ly cấu trúc của chuỗi (theo kết quả trên chương trình phân tích quy luật vận động của nhóm chuỗi cùng nguồn, thí dụ CLUSTALW).

+ **Product Size Range**: Trong ô cửa sổ này người thiết kế có thể điền vào một hay nhiều khoảng số khác nhau, thí dụ người thiết kế đặt chế độ:

"Product Size Range: 100-300 301-400 401-500 501-600", Khi đó, đầu tiên Primer3 chỉ lựa chọn mồi trong đoạn từ vị trí 100 đến 300; Nếu trong khoảng này không tìm được mồi thì Primer3 tiếp tục lặp lại quy trình chọn mồi trên đoạn 301-400, ... và cứ thế tiếp tục cho đến khi

Primer3 chọn được mồi, hoặc lặp lại việc tìm kiếm cho đến hết khoảng đặt cuối cùng.

Trong trường hợp người thiết kế lựa chọn đặt thêm các thông số "Minimum, Optimum and Maximum lengths" thì chương trình Primer3 sẽ không chọn các đoạn mồi tương ứng với sản phẩm PCR ngắn hơn Minimum hay dài hơn Maximum, mà sẽ ưu tiên lựa chọn các đoạn mồi tương ứng với sản phẩm kích thước lân cận giá trị Optimum.

+ **Number to Return**: Xác định số cặp mồi lựa chọn và sắp xếp theo thứ tự "chất lượng" từ thấp đến cao. Thí dụ: đặt chế độ "Number to Return: 5" thì chương trình Primer3 sẽ lựa chọn và sắp xếp 5 đoạn mồi theo mức chất lượng từ dòng 1 đến dòng 5.

+ **Max 3' Stability**: Chỉ số lựa chọn độ ổn định của chuỗi mồi, được tính theo ΔG của octamers, với giá trị lựa chọn cao nhất là 9.00.

+ **Max Mispriming**: Là đặt số lượng (theo trị số hiệu quả) phương án gắn mồi có độ ổn định cao nhất với chuỗi bất kỳ trong "Mispriming Library"; giá trị mặc định của chương trình là 12.00.

+ **Pair Max Mispriming**: Là trị cực đại của tổng số cặp mồi tương đồng so với một chuỗi bất kỳ trong "Mispriming Library"; giá trị mặc định của chương trình là 24.00.

+ **Primer Size**: Là kích thước giới hạn của đoạn mồi được chọn: Min, Max và Opt; với Min ≥ 1, Max ≤ 36, với Min $\leq$ Max và $1 <$ Opt<36. Khi đó, Primer3 sẽ chỉ chọn các đoạn mồi với Min $\leq$ kích thước mồi $\leq$ Max và ưu tiên lựa chọn các đoạn kích thước gần giá trị Opt.

+ **Primer Tm**: Nhiệt độ phân ly cặp mồi (hay còn gọi là nhiệt độ tan mồi), tính theo đơn vị ^{o}C; với ba mức Min, Opt và Max. Đây là điều kiện biên thông báo cho chương trình xử lý chỉ tìm kiếm các đoạn mồi có nhiệt độ phân ly mồi trong khoảng Min - Max và ưu tiên các đoạn có nhiệt độ phân ly mồi lân cận giá trị Opt.

+ **Maximum Tm Difference**: Là chênh lệch nhiệt độ phân ly mồi cao nhất có thể chấp nhận được giữa đoạn mồi bên phải và đoạn mồi về phía trái.

+ **Product Tm**: Với ba giá trị là Minimum, Optimum và Maximum. Khi đặt các chế độ này, chương trình Primer3 chỉ lựa chọn các đoạn mồi tương ứng với sản phẩm có Tm trong khoảng: $Tm_{min} \leq$ nhiệt độ phân ly chuỗi sản phẩm $\leq Tm_{max}$ và với ưu tiên chọn các đoạn có Tm lân cận giá trị Tm_{opt}.

+ **Primer GC%**: Là điều kiện đặt trước về tỉ lệ % tổng số của hai bazơ Guanine và Cytosine, với ba giá trị là Minimum, Optimum và Maximum; Primer3 sẽ ưu tiên lựa chọn tương tự như với thông số Tm đã đặt trước (thường tỉ lệ cặp G-C càng lớn thì Tm càng cao).

+ **Max Self Complementaty**: Là tổng trị số lớn nhất đánh giá khả năng tự bắt cặp của đoạn mồi với đoạn mồi khác. Chương trình Primer3 có bốn mức là:

1.00 mức tương hợp	-0.25 mức nhầm lẫn thay thế bằng N
-1.00 mức sai lệch	-2.00 mức đứt trống GAP

```
5' A T C G N A 3'
   | |   |   |
3' T A - C G T 5'
```

```
5' A T C C G N A 3'
   | |     |   |
3' T A - - C G T 5'
```

* Tổng trị số trên trong sơ đồ là 1.75 (trái) và 0.00 (phải: -0.25).

** Trường hợp tổng trị số bằng 0.0 có nghĩa là đoạn mồi không tự kết cặp với đoạn mồi khác được.

+ **Max 3' Self Complementary**: Là tổng trị số đánh giá khả năng tự kết cặp của đoạn mồi trái với đoạn mồi bên phải, phía đầu 3' theo sơ đồ:

```
5' A T C G C A T G C G C A T C G C A T G C G C A 3'
                          | | | | | | | |   | | |
                      3' T A G C G T A C T C G T C G T A C T C G T 5'
```

+ **Mispriming Library:** để loại bỏ các đoạn mồi không hiệu quả, dựa theo dữ liệu thống kê trong cơ sở dữ liệu.

+ **Max #N's**: Là chỉ số các vị trí bazơ không xác định N cực đại cho phép trong đoạn mồi; Giá trị mặc định của chương trình là 0.

+ **Max Poly-X**: Là chỉ số lặp liên tiếp của một loại nucleotide cực đại cho phép.

+ **Inside Target Penalty**: Trị số xác định số lần đoạn mồi trùm lên vị trí đích; Trị số này không cần xác định nếu trong chuỗi chỉ đặt một điểm đích.

+ **Outside Target Penalty**: Trị số xác định khoảng cách đoạn mồi đến vị trí đích, trong trường hợp mồi tương ứng với đoạn nucleotide bên cạnh đích, nhưng không trùm lên đích này.

+ **First Base Index**: Chỉ số của bazơ đầu tiên trong chuỗi nhập vào. Trên giao diện xử lý trực tuyến trị số này mặc định là 1.

+ **CG Clamp**: Trị số xác định số bazơ G và C liên tục nhau từ phía đầu 3' của cả hai đoạn mồi.

+ **Salt Concentration**: Nồng độ muối trong phản ứng PCR, tính theo milimol (thường dùng là muối KCl).

+ **Annealing Oligo Concentration**: Nồng độ mồi trong phản ứng PCR, tính bằng nanomol. Primer3 sử dụng nồng độ này để tính nhiệt độ phân ly mồi. Giá trị mặc định của chương trình Primer3 là 50 nM.

+ **Liberal Base**: Phương án lựa chọn chế độ mã IUB/IUPAC cho các bazơ không xác định. Nếu không chấp nhận tồn tại dạng này trong đoạn mồi, phải đặt chế độ "Max Ns Accepted: 0".

+ **Show Debuging Info**: Phương án lựa chọn chế độ thông báo sửa lỗi đầu vào trong kết quả ra.

+ **Included Region**: Đặt khoảng giới hạn chọn mồi, dạng số "x, y" hoặc dạng cụm ký tự : ...{TGA and ATT}.... Khi đó chương trình chỉ lựa chọn mồi phù hợp với đoạn chuỗi trong khoảng giới hạn trên.

+ **Start Codong Position**: Vị trí xác định trong thực nghiệm.

+ **Objective Function Penalty Weights for Primers**: Trong mục này yêu cầu người xử lý chọn các chế độ đặt tương ứng cho các giá trị:

- Tm
- Size
- GC%
- Self Complementary
- 3' Self Complementary
- #N's
- Mispriming
- Sequence Quality
- End Sequence Quality
- Position Penalty và
- End Stability

Trong đó, tại ba tham số đầu đều có hai ô cửa để đặt chế độ, với "Lt" để đặt cận trên (Less than) và "Gt" để đặt cận dưới (Greater than). Mục đích thay đổi các tham số này cho phép người sử dụng công cụ để xử lý chọn ra đoạn mồi tốt nhất.

Các thông số yêu cầu trong mục "**Objective Function Penalty Weights for Primer Pair**" cũng tương tự như phần trên. Việc lựa chọn đặt chế độ cho các phần còn lại khác có thể xem trực tiếp trong phần trợ giúp của chương trình, trong http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www_help.cgi#generic_penalty_weights

Thí dụ, khi sử dụng chương trình trực tuyến Primer3 để lựa chọn mồi cho chuỗi sau:

ctcagctgtgtcaaagtttcacagatcctcgtcttctattccggctacactcagtctcctccagcttaga
tctttgtccttctcctgggtactctccgactccttcttccagctaatgtccggtcattagaaaagttttaaa
gtttgaattgtcnntccctgtcaaagtttccagacctcgtcgtccttctcttctccgtcagctctcagtctt
cattggaacagatctgtctttattccgcctgctacactcagtctcctccttcagtctcttaaaagtttgttc
agtcttagatgaatttctctgggtactttgtcctccgactccgtccagctaatcggtcttgtcgtcattag
atttccttcttctagatgattcatgtctacctattgtcnntcgtcttcccgtgtnnnccaggtccgtttcgtc
cgcctgtcgtctattctatctcggtccttacacaaagttgtccttaaagtttttttgtgtccctagtccaag
gtccaatttttccatctgtttcgtcctgtctttttgnggtcgcgtccgtttcccgttctctatgcctccctcct
cttatc

Với một số thông số đặt trước, bao gồm:

- Sequen ID : H_A_N
- Targets: 300,250
- Excluded Region: 30,15
- Number To Return: 5 Max 3' Stability: 9.0
- Max Mispriming: 12.00 Pair Max Mispriming: 24.00
- Primer Size: Min: 15 Opt: 20 Max: 25
- Primer Tm: Min: 55 Opt: 60 Max: 65
- Product Tm Min: Opt: 50 Max:
- Các tham số khác giữ nguyên giá trị mặc định của chương trình.

Sau khi nhấn "Pick Primer" để gửi thông tin đi xử lý, người phân tích sẽ nhận lại được kết quả lựa chọn đoạn mồi, với giao diện tương ứng như sau:

Primer3 Output

```
PRIMER PICKING RESULTS FOR H_A_N

No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len   tm     gc%   any   3'   seq
LEFT PRIMER        173   20   59.99  60.00  4.00  0.00 cagacctcgtcgtccttctc
RIGHT PRIMER       572   20   59.27  50.00  2.00  0.00 gaggcatagagaacgggaaa
SEQUENCE SIZE: 584
INCLUDED REGION SIZE: 584

PRODUCT SIZE: 400, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00
TARGETS (start, len)*: 300,250
EXCLUDED REGIONS (start, len)*: 30,15

    1 ctcagctgtgtcaaagtttcacagatcctcgtcttctattccggctacactcagtctcct
                                   XXXXXXXXXXXXXXX

   61 ccagcttagatctttgtccttctcctgggtactctccgactccttcttccagctaatgtc

  121 cggtcattagaaaagttttaaagtttgaattgtcnntccctgtcaaagtttccagacctc
                                                         >>>>>>>>

  181 gtcgtccttctcttctccgtcagctctcagtcttcattggaacagatctgtctttattcc
      >>>>>>>>>>>>

  241 gcctgctacactcagtctcctccttcagtctcttaaaagtttgttcagtcttagatgaat
                                                                *

  301 ttctctgggtactttgtcctccgactccgtccagctaatcggtcttgtcgtcattagatt
```

```
      ************************************************************

  361 tccttcttctagatgattcatgtctacctattgtcnntcgtcttcccgtgtnnnccaggt
      ************************************************************

  421 ccgtttcgtccgcctgtcgtctattctatctcggtccttacacaaagttgtccttaaagt
      ************************************************************

  481 ttttttgtgtccctagtccaaggtccaattttttccatctgtttcgtcctgtctttttg
      ************************************************************

  541 nggtcgcgtccgtttcccgttctctatgcctccctcctcttatc
      *********   <<<<<<<<<<<<<<<<<<<<
```

KEYS (in order of precedence):
XXXXXX excluded region
****** target
>>>>>> left primer
<<<<<< right primer

ADDITIONAL OLIGOS	start	len	tm	gc%	any	3'	seq
1 LEFT PRIMER	185	20	59.68	55.00	2.00	2.00	tccttctcttctccgtcagc
RIGHT PRIMER	572	20	59.27	50.00	2.00	0.00	gaggcatagagaacgggaaa
PRODUCT SIZE: 388, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 2.00							
2 LEFT PRIMER	236	20	60.42	55.00	3.00	3.00	attccgcctgctacactcag
RIGHT PRIMER	572	20	59.27	50.00	2.00	0.00	gaggcatagagaacgggaaa
PRODUCT SIZE: 337, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00							
3 LEFT PRIMER	183	20	60.52	55.00	2.00	1.00	cgtccttctcttctccgtca
RIGHT PRIMER	572	20	59.27	50.00	2.00	0.00	gaggcatagagaacgggaaa

```
   PRODUCT SIZE: 390, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 2.00

 4 LEFT PRIMER        228   20   59.48   50.00  2.00  0.00 ctgtctttattccgcctgct
   RIGHT PRIMER       572   20   59.27   50.00  2.00  0.00 gaggcatagagaacgggaaa
   PRODUCT SIZE: 345, PAIR ANY COMPL: 6.00, PAIR 3' COMPL: 3.00

Statistics
         con   too    in    in          no    tm    tm  high  high        high
         sid  many   tar  excl   bad    GC   too   too   any    3'  poly   end
        ered    Ns   get   reg   GC% clamp   low  high compl compl     X  stab    ok
Left    2384    26     0    38    88     0  1274   109     1     0     0    25   823
Right    141     0     0     0     0     0    71    12     0     0     0     7    51
Pair Stats:
considered 84012, unacceptable product size 84005, high end compl 2, ok 5
primer3 release 1.0

(primer3_www_results.cgi v 0.4)
```

Trong tệp kết quả này, chương trình Primer3 đã lựa chọn được năm cặp mồi. Cặp mồi phù hợp nhất với các yêu cầu đầu vào, được hiển thị đầu tiên trong tệp kết quả, là:

- Mồi xuôi: **cagacctcgtcgtccttctc**
- Mồi ngược: **gaggcatagagaacgggaaa**

Sau đó là bốn cặp mồi khác kém hơn, xếp theo thứ tự chất lượng (được đánh số theo thứ tự chất lượng...) như sau:

1	Mồi xuôi	tccttctcttctccgtcagc
	Mồi ngược	gaggcatagagaacgggaaa
2	Mồi xuôi	attccgcctgctacactcag
	Mồi ngược	gaggcatagagaacgggaaa
3	Mồi xuôi	cgtccttctcttctccgtca
	Mồi ngược	gaggcatagagaacgggaaa
4	Mồi xuôi	ctgtctttattccgcctgct
	Mồi ngược	gaggcatagagaacgggaaa

Thông tin thu được trên được sử dụng để tổng hợp đoạn mồi, hay đặt mua các đoạn mồi, phục vụ cho mục tiêu thực hiện phản ứng nhân khuyếch đại PCR đoạn gen chờ đợi (có cấu trúc tương đồng với sợi khuôn được lựa chọn để thiết kế mồi) có trong mẫu DNA được cấp vào hỗn hợp phản ứng.

7. CHƯƠNG TRÌNH PHÂN TÍCH CẤU TRÚC TƯƠNG ĐỒNG BLAST

7.1. Đại cương

WU-BLAST2 (*Washington University Basic Local Alignment Tools version 2 - Warren Gish*) là chương trình so sánh cấu trúc của chuỗi DNA, chuỗi amino axit phân tích với các chuỗi tương ứng lưu giữ trong ngân hàng dữ liệu, nhằm tìm kiếm các chuỗi có độ tương đồng cao nhất với chuỗi kiểm tra. Sau đó, người phân tích sẽ khai thác các thông tin về đặc điểm hay đặc tính đã biết của các chuỗi trong ngân hàng để dự đoán xác định cấu trúc và đặc tính của chuỗi kiểm tra này. Trọng tâm của kỹ thuật phân tích là tìm kiếm xác định các chuỗi (và các vùng trên chuỗi) lưu trữ trong ngân hàng dữ liệu có cấu trúc tương đồng cao với chuỗi cần phân tích. Trên nguyên tắc sự tồn tại về độ tương đồng cấu trúc cho phép chờ đợi sự gần gũi nhau về đặc tính; hay nói cách khác, từ sự tương đồng này cho phép dự đoán được tính chất của chuỗi cần phân tích, dựa theo đặc tính của các chuỗi tương đồng đã nêu phần trên (đã mô tả trong cơ sở dữ liệu). Về phương diện kỹ thuật, chương trình BLAST cho phép phát hiện sự tương đồng cấu trúc ở hai mức độ là mang tính cục bộ ở một vùng hay mang tính tổng thể giữa hai chuỗi với nhau. Đơn vị cơ bản của đầu ra theo thuật toán BLAST là chỉ số *HSP (High-Scoring Segment Pair)*. HSP là hai đoạn chuỗi, ở vị trí bất kỳ, có độ dài bằng nhau có chỉ số so sánh bằng hoặc vượt ngưỡng hay bằng hoặc vượt hệ số so sánh ***cutoff***. Mỗi bộ HSP được xác định qua ba phần là trình tự hai đoạn chuỗi (một của chuỗi phân tích và một của chuỗi lấy trong ngân hàng dữ liệu), chỉ số hệ thống và chỉ số *cutoff*.

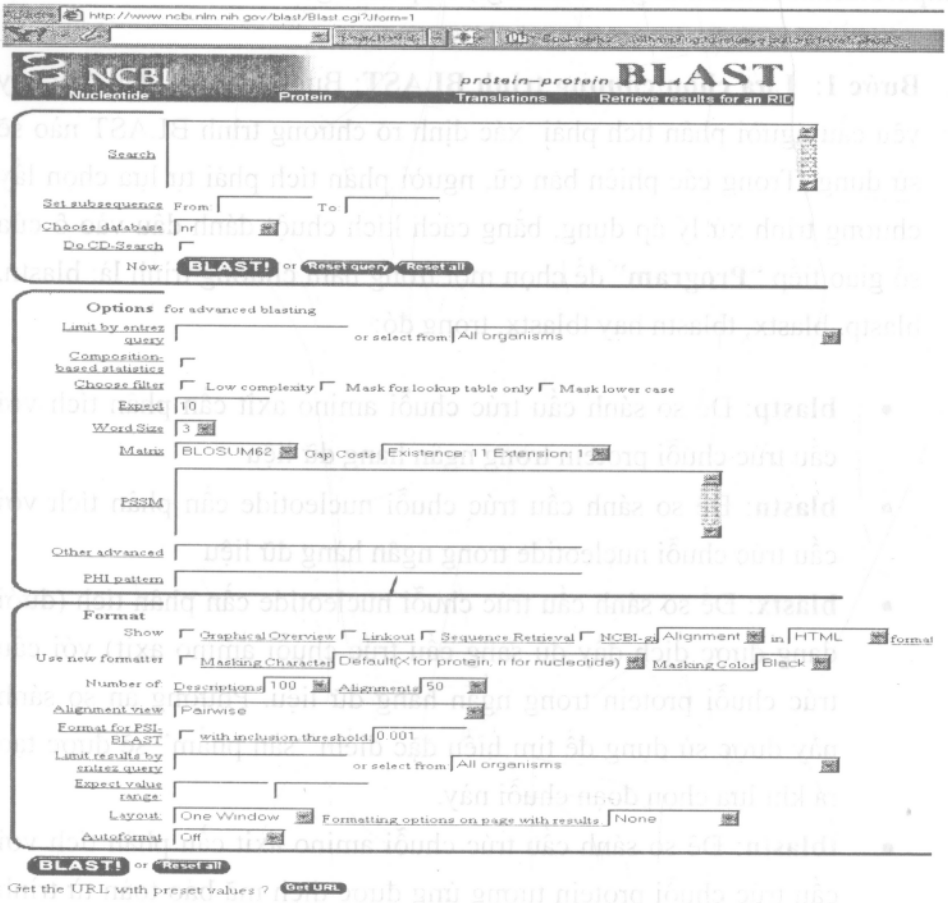
7.2. Sử dụng chương trình BLAST trực tuyến

Thao tác cơ bản khi sử dụng chương trình phân tích cấu trúc chuỗi BLAST trực tuyến thường bắt đầu bằng kết nối internet, rồi hiển thị giao diện chung của chương trình theo đường dẫn:
http://blast.ncbi.nlm.nih.gov/Blast.cgi; Tiếp theo gồm các bước chính sau:

1. **Bước 1: Lựa chọn chương trình BLAST**: Bước thao tác đầu tiên này yêu cầu người phân tích phải xác định rõ chương trình BLAST nào sẽ sử dụng. Trong các phiên bản cũ, người phân tích phải tự lựa chọn lấy chương trình xử lý áp dụng, bằng cách kích chuột đánh dấu vào ô cửa sổ giao tiếp "**Program**" để chọn một trong năm chương trình là: blastn, blastp, blastx, tblastn hay tblastx, trong đó:

   - **blastp**: Để so sánh cấu trúc chuỗi amino axit cần phân tích với cấu trúc chuỗi protein trong ngân hàng dữ liệu
   - **blastn**: Để so sánh cấu trúc chuỗi nucleotide cần phân tích với cấu trúc chuỗi nucleotide trong ngân hàng dữ liệu
   - **blastx**: Để so sánh cấu trúc chuỗi nucleotide cần phân tích (dưới dạng được dịch đầy đủ sang cấu trúc chuỗi amino axit) với cấu trúc chuỗi protein trong ngân hàng dữ liệu. Phương án so sánh này được sử dụng để tìm hiểu đặc điểm "sản phẩm" sẽ được tạo ra khi lựa chọn đoạn chuỗi này.
   - **tblastn**: Để so sánh cấu trúc chuỗi amino axit cần phân tích với cấu trúc chuỗi protein tương ứng được dịch mã bảo toàn từ trình tự chuỗi nucleotide trong ngân hàng dữ liệu.
   - **tblastx**: Là phương án so sánh cấu trúc chuỗi amino axit cần phân tích với cấu trúc chuỗi protein trong ngân hàng dữ liệu, theo từng đoạn khung gồm các cụm sáu ký tự một.

Phiên bản xử lý trực tuyến BLAST mới nhất của ngân hàng dữ liệu NCBI hiển thị các chương trình này độc lập với nhau, dưới dạng "nucleotide - nucleotide BLAST", "protein - protein BLAST" và "translating BLAST", nên ngay từ đầu người phân tích đã phải truy cập trực tiếp vào các trang chương trình riêng này. Hình 7.1. biểu thị giao tiếp hiển thị "protein - Protein BLAST".



Hình 7.1. Giao diện chương trình protein-protein BLAST

2/ Bước 2: **Nhập dữ liệu**: Chương trình xử lý trực tuyến BLAST cho phép nhập dữ liệu chuỗi phân tích trực tiếp dạng ký tự qua bàn phím hay nhập

dữ liệu đã được viết theo một trong ba ngôn ngữ là "**FASTA** sequence format", "**Identifiers**" (NCBI Accessions numbers, Gis) và "**Bare Sequence**".

- Theo ngôn ngữ FASTA, chuỗi dữ liệu được viết thành hai phần: Dòng đầu bắt đầu bằng ký tự ">" hoặc ">gi[...] (chỉ kích thước chuỗi lớn hơn...) và tiếp theo là thông tin chung về chuỗi; các dòng sau là trình tự cấu trúc chuỗi (viết liên tục, không để cách dòng trống ở giữa), thí dụ:

>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNK
MKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKVL
MALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQF
RADHP FLFLIKHNPTNTIVYFGRYWSP

- Ngôn ngữ Bare Sequence cũng viết tương tự như ngôn ngữ FASTA, song không có dòng thông tin chung ban đầu, mà chỉ có dòng trình tự cấu trúc chuỗi, với dạng như sau:

QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPV
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVK
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDI
KHSPESEQFRADHP FLFLIKHNPTNTIVYFGRYWSP

hoặc:

1 qikdllvsss tdldttlvlv naiyfkgmwk tafnaedtre mpfhvtkqes kpvqmmcmnn
61 sfnvatlpae kmkilelpfa sgdlsmlvll pdevsdleri ektinfeklt ewtnpntmek
121 rrvkvylpqm kieekynlts vlmalgmtdl fipsanltgi ssaeslkisq avhgafmels
181 edgiemagst gviedikhsp eseqfradhp flflikhnpt ntivyfgryw sp

- Ngôn ngữ Identifiers được viết có dạng như sau:

ACCESSION P01013 AAA68881. 1 gi| 129295

3/ **Bước 3**: **Đặt vùng phân tích "Set Subsequence"**: Trong mục này, người phân tích phải cung cấp thông tin vị trí trên đoạn chuỗi cần phân tích bằng hai giá trị số chỉ vị trí giới hạn đầu-cuối đoạn chuỗi ấy. Trong

trường hợp cần phân tích toàn chuỗi, dữ liệu nhập sẽ có dạng *from 1, to length.*

4/ **Bước 4: Lựa chọn ngân hàng dữ liệu "choose databases"**: Trong bước lựa chọn này, người phân tích phải xác định nhóm dữ liệu cụ thể của ngân hàng dữ liệu được chỉ định làm đối tượng so sánh. Thao tác lựa chọn này được thực hiện bằng cách dùng chuột đánh dấu vào một trong các mảng cấu trúc chuỗi, trong cửa sổ giao tiếp "**Choose databases**", tương ứng với đối tượng chuỗi cần phân tích. Để phục vụ cho mục đích trên, chương trình xử lý BLAST đã phân chia và sử dụng các ký hiệu viết tắt để chỉ các nhóm đối tượng cơ sở dữ liệu tương ứng như sau:

A/ Cơ sở dữ liệu protein bao gồm:

- **Nr**: Cho các chuỗi được dịch đầy đủ từ các cơ sở dữ liệu GenBank CDS + PDB+ SwissProt+ PIR+ PRF.
- **Month**: Cho các chuỗi được dịch đầy đủ từ các cơ sở dữ liệu GenBank CDS + PDB+ SwissProt+ PIR+ PRF, chỉ xét đến các chuỗi mới đăng ký bổ sung vào ngân hàng dữ liệu trong 30 ngày gần nhất.
- **Swissprot**: Dành cho phương án lựa chọn so sánh với phiên bản dữ liệu mới nhất mà NCBI nhận được từ cơ sở dữ liệu "SWISS-PROT protein sequence database của EMBL".
- **Patents**: Khi lựa chọn so sánh với chuỗi Protein đã đăng ký bảo hộ sáng chế trong ngân hàng "Patent division of GenBank".
- **Yeast**: Là phương án lựa chọn cơ sở dữ liệu protein tương ứng, được biên dịch đầy đủ theo cấu trúc genome hoàn chỉnh của nấm men *Saccharomyces cerevisiae.*

- **E. coli** : Là cơ sở dữ liệu protein tương ứng, được biên dịch đầy đủ theo cấu trúc genome hoàn chỉnh của vi khuẩn *Escherichia coli*.
- **Pdb**: Là các chuỗi tương ứng với các chuỗi protein trong ngân hàng dữ liệu "Brookhaven Protein Data Bank".
- **kabat [kabatpro]:** Là các chuỗi có liên quan đến hoạt tính miễn dịch trong ngân hàng dữ liệu "Kabat's database" (chi tiết hơn xem trong trang Web: http://immuno.bme.nwu.edu/).
- **alu**: Chuỗi dịch từ ngân hàng dữ liệu "REPBASE" (đặc tính chi tiết hơn vào trang ftp://ncbi.nlm.nih.gov/pub/jmc/alu, xem nội dung trong đường dẫn Claverie and Makalowski, Nature vol. 371, page 752 (1994).

B/ Cơ sở dữ liệu nucleotide bao gồm:

- **nr**: Các chuỗi hoàn chỉnh của các ngân hàng dữ liệu: GenBank+ EMBL+ DDBJ+ PDB (song không bao gồm chuỗi thuộc các mảng EST, STS, GSS, hoặc HTGS).
- **Month**: Các chuỗi mới cập nhật vào các ngân hàng dữ liệu GenBank+ EMBL+ DDBJ+ PDB trong vòng 30 ngày gần nhất.
- **Dbest**: Các chuỗi hoàn chỉnh của các ngân hàng dữ liệu: GenBank+EMBL+DDBJ EST.
- **mouse_ests**: Các chuỗi gen chuột các ngân hàng dữ liệu: GenBank+EMBL+DDBJ EST.
- **human_ests**: Các chuỗi gen người các ngân hàng dữ liệu: GenBank+EMBL+DDBJ EST.

- **other_ests**: Các chuỗi gen của các sinh giới khác trong các ngân hàng dữ liệu: GenBank+EMBL+DDBJ EST, không xét đến mảng gen người và gen chuột.
- **yeast**: Cấu trúc các đoạn chuỗi gen hoàn chỉnh, lấy từ mảng genome nấm men *Saccharomyces cerevisiae.*
- **E. coli**: Cấu trúc các chuỗi gen hoàn chỉnh, lấy từ mảng genome của vi khuẩn *E. coli.*
- **Pdb**: cấu trúc chuỗi gen hoàn chỉnh tương ứng với cấu trúc không gian ba chiều của protein trong ngân hàng dữ liệu PDB.
- **kabat [kabatnuc]**: Là các chuỗi có liên quan đến hoạt tính miễn dịch trong ngân hàng dữ liệu "Kabat's database" (chi tiết hơn xem trong trang Web: http://immuno.bme.nwu.edu/).
- **patents**: Cấu trúc chuỗi nucleotide đã đăng ký bảo hộ sáng chế trong mảng dữ liệu Patent division of GenBank.
- **Vector**: Cấu trúc các đoạn Vector trong ngân hàng GenBank (R), NCBI, (xem trong ftp://ncbi.nlm.nih.gov/pub/blast/db/)
- **Mito**: Dữ liệu về chuỗi của ty thể.
- **Alu**: Chuỗi dịch từ ngân hàng dữ liệu REPBASE (xem trong trang ftp://ncbi.nlm.nih.gov/pub/jmc/alu, như đã nêu trong phần protein trên).
- **Gss**: Dữ liệu về bộ gen hoàn chỉnh (Genome Survey Sequence) bao gồm cả các đoạn sợi đơn, các chuỗi có exon và các chuỗi Alu PCR.
- **Htgs**: Dữ liệu về các chuỗi gen có độ hiệu dụng cao (High Throughput Genomic Sequences).

Thao tác tiếp theo, người phân tích phải xác định thêm một số thông số yêu cầu trong mục "**Options**" và "**Format**". Các thông tin trong mục **Option** bao gồm:

- Hạn chế chuỗi lựa chọn (**Limit by entrez Query... or select from...**) để giảm số lượng chuỗi cần phân tích. Chương trình BLAST cho phép sử dụng mọi mã hay cụm ký tự được chương trình tìm kiếm Entrez chấp nhận, thí dụ: **Protease NOT hiv 1 [Organism]** là giới hạn chỉ tìm các chuỗi protease và bỏ qua cả các chuỗi dạng này trong HIV 1.
- Lựa chọn phin lọc (**Choose filter**): Với ba phương án là: **Low complexity** (loại không xét đến các thông tin riêng biệt của chuỗi), **Mask for lookup table only** (tìm kiếm theo chế độ low complexity, sau đó mới xem xét đến toàn bộ thông tin riêng biệt trong các chuỗi đã tìm được) và **Mask lower case** (cho phép sử dụng thông tin, viết theo ngôn ngữ FASTA) và các thông số khác...

Trong mục **Format**, người phân tích cần lựa chọn đặt trước các chế độ sau:

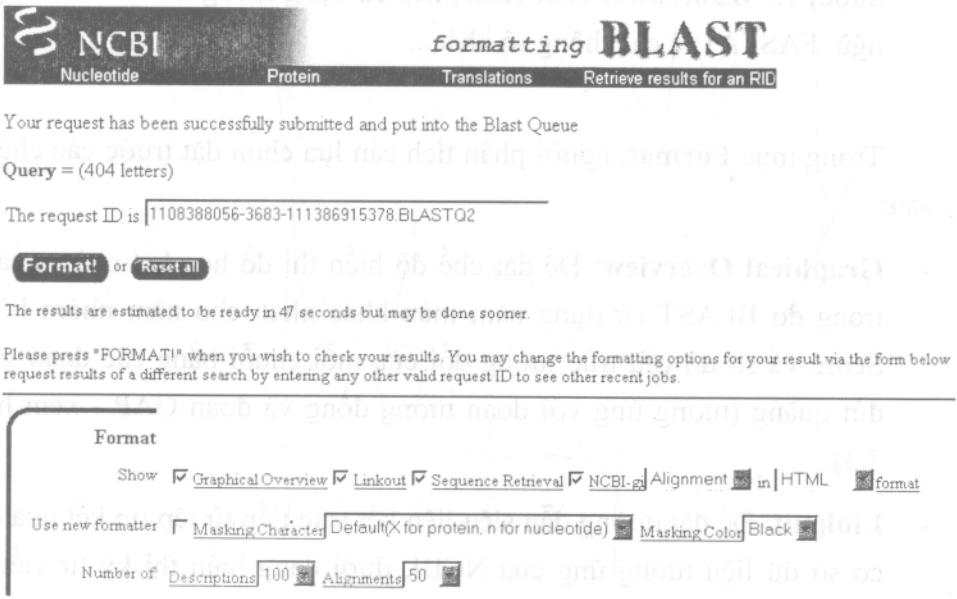
- **Graphical Overview**: Để đặt chế độ hiển thị đồ hoạ kết quả so sánh, trong đó BLAST sử dụng năm màu khác nhau cho năm nhóm hệ số Score và sơ đồ cấu trúc tương đối của mỗi chuỗi bằng các đoạn gạch đứt quãng (tương ứng với đoạn tương đồng và đoạn GAP - xem hình 7.3).
- **Linkout**: Để đặt đường dẫn siêu liên kết trực tiếp từ tệp tin kết quả đến cơ sở dữ liệu tương ứng của NCBI, dưới dạng hiển thị ký tự viết tắt trong ô nền màu (hình 7.3). Thí dụ hai ký tự (**L U**) là vị trí đường dẫn siêu liên kết trên giao diện hiển thị kết quả đến tệp dữ liệu tương ứng trong **LocusLink** và **UniGene.**

- **NCBI-gi ... in ...** : Để đặt chế độ hiển thị kết quả theo một trong ba phương án (Alignment, PSSM hay Bioseq), dưới một trong bốn dạng (HTML, Plain Text, ASN.1 hay XML).

- Ngoài ra nếu cần thiết phải đặt tiếp chế độ cho một số tham số khác theo yêu cầu phân tích.

Trong trường hợp không đặt lựa chọn các thông số trong hai mục trên, chương trình sẽ xử lý theo chế độ mặc định của ngân hàng dữ liệu đã chọn.

5/ **Bước 5 - Gửi yêu cầu xử lý**: Sau khi khai báo xong, người phân tích nhấn lệnh "BLAST" để gửi tin đi. Sau khoảng thời gian chờ đợi ngắn, chương trình BLAST sẽ phản hồi yêu cầu với dạng giao diện như trong hình 7.2.



Hình 7.2. Giao diện trung gian của chương trình xử lý trực tuyến BLAST

Sau khi cung cấp các thông tin bổ sung cần thiết, người phân tích lại tiếp tục nhấn lệnh "FORMAT" để gửi tin. Sau mỗi lần gửi tin bằng lệnh FORMAT này, người phân tích sẽ nhận được một tệp dữ liệu kết quả, với các mức từ thấp đến cao. Nghĩa là khi tìm được, trong thông tin phản hồi sẽ hiển thị các chuỗi theo độ tương đồng từ mức cao xuống mức thấp hơn. Trong trường hợp chưa tìm được chuỗi mong muốn, người phân tích vẫn có thể thay đổi lại lần nữa các thông số đi và gửi đi tiếp, cho đến khi thu được kết quả mong muốn hay dừng lại. Chi tiết hơn về các chế độ này có thể xem hướng dẫn trực tuyến tại địa chỉ:

http://www.ncbi.nlm.nih.gov/
last/html/blastcgihelp.html#get_subsequence

Để hiểu rõ hơn thao tác xử lý trên hãy làm thí dụ sau: Giả sử cần tiến hành khi phân tích đặc tính chuỗi nucleotide (giả định) với cấu trúc sau:

gggttaccaatctgcttggcatattgagattcctgcaaggtggaaacctggtaataagcgg
aacttcttacaaaagaggaagacagggcacactctctggagtggagttggtgttaaaacagt
actcttctggttgtagtaattatatacagttaagttcgtagtgagtgtctggtccagtgtctgatgtaa
gcccacattctcttctagtgggcctgggcaagttaaaaatagtgcttccaggtcatcgattgtcttc
tccagtagtgccgagaaactgtcctagtgctgcaaactcagctcgggtctcagcctccttcagc
ttgtcagacagaagcttgatagtgcttcttcatatagtgatcctcctattgacagaatacttggccg
cttcagaagcagcc

Một trong các giải pháp là sử dụng chế độ phân tích trực tuyến qua ngân hàng dữ liệu NCBI. Khi đó, thao tác qua các bước chính sau:

- Để lựa chọn chương trình cần thao tác theo trình tự sau: http://www.ncbi.nlm.nih.gov > Tools > BLAST > Nucleotide -

Nucleotide BLAST (blastn). Kết thúc các dòng lệnh trên, giao diện "Nucleotide - Nucleotide BLAST" sẽ xuất hiện.

- Nhập dữ liệu và đặt chế độ yêu cầu phân tích, bao gồm các thao tác là: chèn tệp cấu trúc chuỗi vào ô cửa sổ "**search**"; Với giả sử chọn các chế độ là: đặt khoảng tìm kiếm "**set Subsequences**" (From 1 to Length); chọn cơ sở dữ liệu so sánh "**Choose Databaces**" (est_other); các thông số khác theo chế độ mặc định của chương trình. Sau đó, nhấn cửa sổ "**BLAST**" để gửi thông tin đi.

- Sau khoảng thời gian ngắn, chương trình xử lý trực tuyến sẽ phản hồi lại thông tin với dạng giao diện như hình 7.2. Sau khi lựa chọn cung cấp các thông tin bổ sung cần thiết, người phân tích lại tiếp tục nhấn lệnh " FORMAT"... Trong trường hợp tìm được kết quả mong muốn, chương trình BLAST sẽ phản hồi lại tệp tin kết quả với giao diện như trong hình 7.3.

Về cấu trúc, tệp tin kết quả gồm bốn phần là:

- Phần đầu hiển thị kết quả sơ bộ dạng đồ hoạ hình ảnh màu của các chuỗi có độ tương đồng cao nhất, như trong hình 7.3.

NCBI **results of BLAST**

BLASTN 2.2.10 [Oct-19-2004]

Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

RID: 1108388056-3683-111386915378.BLASTQ2

Query= (404 letters)

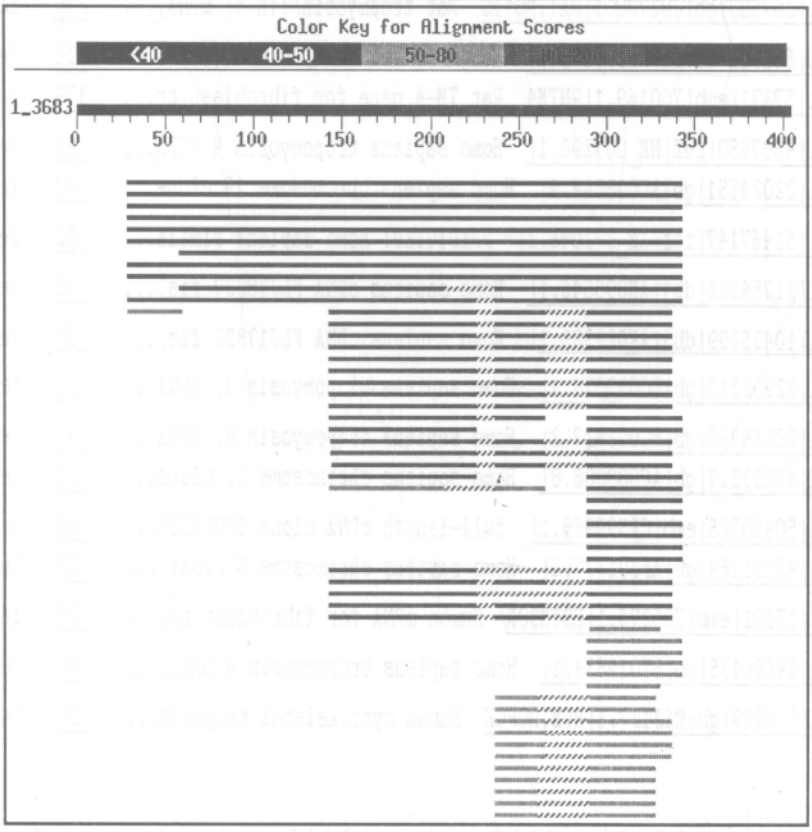
Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,environmental samples or phase 0, 1 or 2 HTGS sequences)
2,891,993 sequences; 13,289,160,675 total letters

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**

Taxonomy reports

Distribution of 153 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments



Hình 7.3. Giao diện kết quả chương trình BLAST

- Phần tiếp theo hiển thị kết quả dạng ký tự tóm tắt kết quả, dạng như sau:

```
                                                          Score    E
Sequences producing significant alignments:               (bits) Value
gi|47894397|ref|NM_001001491.1|  Mus musculus tropomyosin 4 ...   373   e-100  GU
gi|23958443|gb|BC023701.1|  Mus musculus tropomyosin 4, mRNA...   373   e-100  GUE
gi|21618969|gb|BC032174.1|  Mus musculus cDNA clone IMAGE:53...   373   e-100  GU
gi|21327808|gb|BC032175.1|  Mus musculus cDNA clone IMAGE:53...   373   e-100  GU
gi|56031571|dbj|AK207394.1|  Mus musculus cDNA, clone:Y2G010...   315   5e-83  U
gi|6981671|ref|NM_012678.1|  Rattus norvegicus tropomyosin 4...   230   2e-57  GUE
gi|207503|gb|J02780.1|RATTRO4IS  Rat tropmyosin (TM-4) mRNA,...   230   2e-57  GUE
gi|56030266|dbj|AK206089.1|  Mus musculus cDNA, clone:Y2G010...   216   3e-53  U
gi|57371|emb|Y00169.1|RNTM4  Rat TM-4 gene for fibroblast tr...   174   1e-40
gi|4507650|ref|NM_003290.1|  Homo sapiens tropomyosin 4 (TPM...    92   1e-15  GUE
gi|22024551|gb|AC008894.9|  Homo sapiens chromosome 19 clone...    92   1e-15
gi|51467147|ref|XM_372046.2|  PREDICTED: Homo sapiens simila...    92   1e-15  GU
gi|21754822|dbj|AK095546.1|  Homo sapiens cDNA FLJ38227 fis,...    92   1e-15  U
gi|10435299|dbj|AK023385.1|  Homo sapiens cDNA FLJ13323 fis,...    92   1e-15  GUE
gi|22902217|gb|BC037576.1|  Homo sapiens tropomyosin 4, mRNA...    92   1e-15  GU
gi|38114798|gb|BC002827.2|  Homo sapiens tropomyosin 4, mRNA...    92   1e-15  GUE
gi|17223217|gb|AC009808.8|  Homo sapiens chromosome 8, clone...    92   1e-15
gi|50480765|emb|CR599958.1|  full-length cDNA clone CS0DK009...    92   1e-15  U
gi|46391164|gb|AF201337.4|  Homo sapiens chromosome 8 clone ...    92   1e-15
gi|37201|emb|X05276.1|HSTM30R  Human mRNA for fibroblast tro...    92   1e-15  GUE
gi|54696135|gb|BT019634.1|  Homo sapiens tropomyosin 4 mRNA,...    90   5e-15  GU
gi|339959|gb|M12127.1|HUMTROPCK  Human cytoskeletal tropomyo...    90   5e-15  GU
```

- Phần thứ ba hiển thị kết quả cụ thể khi so sánh từng cặp chuỗi, với giao diện có dạng:

```
>gi|47894397|ref|NM_001001491.1|  GU  Mus musculus tropomyosin 4 (Tpm4), mRNA
          Length = 2082
 Score =  373 bits (188), Expect = e-100
 Identities = 286/314 (91%), Gaps = 21/314 (6%)
 Strand = Plus / Minus
Query: 30   ttcctgcaaggtggaaacctggtaataagcggaacttcttacaaaagaggaagacagggc 89
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 887 ttcctgcaaggtggaaacctggtaataagcggaacttcttacaaaagaggaagacagggc 828

Query: 90   acactctctggagtggagttggtgttaaaacagtactcttctggttgtagtaattatata 149
            ||||||||||||||||||||||||||||||||||||||||||||| ||||  |||||||
Sbjct: 827 acactctctggagtggagttggtgttaaaacagtactcttctggtt-tagt--ttatata 771

Query: 150 cagttaagttcgtagtgagtgtctggtccagtgtctgatgtaagcccacattctcttcta 209
           |||||||||||||  | ||||||||||||||||||||||||||||||||||||||||||
Sbjct: 770 cagttaagttcgt--ttagtgtctggtccagtgtctgatgtaagcccacattctcttct- 714

Query: 210 gtgggcctgggcaagttaaaaatagtgcttccaggtcatcgattgtcttctccagtagtg 269
           | |||||||||||||||  | |||||||||||||||||||||||||||||||||||||
Sbjct: 713 -ttggcctgggcaagtt-------tttcttccaggtcatcgattgtcttctccagt---- 666

Query: 270 ccgagaaactgtcctagtgctgcaaactcagctcgggtctcagcctccttcagcttgtca 329
             |||||||||||||  | ||||||||||||||||||||||||||||||||||||||||
Sbjct: 665 -ttagaaactgtcct--ttctgcaaactcagctcgggtctcagcctccttcagcttgtca 609

Query: 330 gacagaagcttgat 343
           ||||||||||||||
Sbjct: 608 gacagaagcttgat 595

>gi|23958443|gb|BC023701.1|  GUE  Mus musculus tropomyosin 4, mRNA (cDNA clone MGC:38384
          IMAGE:5345587), complete cds
         Length = 2118
```

- Phần cuối cùng tóm tắt thông tin về chế độ chạy yêu cầu cho BLAST.

Phần đầu của kết quả cung cấp cho người phân tích bức tranh tổng thể về quan hệ tương đồng về cấu trúc bậc 1 của các chuỗi có trong cơ sở dữ liệu được chọn lựa so sánh với chuỗi dữ liệu được gửi đi phân tích, trong đó độ tương đồng được sắp xếp từ trên xuống dưới theo mức độ từ cao đến thấp (trong bảng ô vuông, các chuỗi được biểu thị dưới dạng các đoạn thẳng với màu sắc tương ứng với mức độ tương đồng trên các vùng của chuỗi).

Phần thứ 2, các chuỗi cũng được sắp xếp theo mức độ tương đồng giảm dần từ trên xuống dưới; song trong phần này chương trình hiển thị cả tên chuỗi, hệ số tương đồng và các dạng dữ liệu về cấu trúc của chuỗi có trong cơ sở dữ liệu (bằng các ô màu bên góc phải của chuỗi).

Phần thứ ba giao diện hiển thị chi tiết hơn về trình tự cấu trúc giữa chuỗi gửi đi phân tích (**Query**) với chuỗi có cấu trúc tương đồng cao nhất được tìm thấy trong cơ sở dữ liệu lựa chọn (Subject - **Subject**), với chỉ số tương đồng (Identities) và các đoạn trống giữa hai cấu trúc (Gap).

Kết quả so sánh về độ tương đồng này cho phép người phân tích có thể dự đoán được, phụ thuộc vào mức độ tương đồng, đặc tính của chuỗi sản phẩm gửi đi phân tích, dựa theo các đặc tính của chuỗi có cấu trúc tương đồng đã được các xác định và mô tả trong cơ sở dữ liệu. Các đặc tính này dễ dàng nhận được, nếu kích chuột vào vị trí tên của chuỗi hiển thị trên giao diện kết quả. Đương nhiên, đặc tính thực của chuỗi sản phẩm, nói riêng và bản chất khoa học sinh học nói chung, chỉ có thể được xác định bằng con đường thực nghiệm; Song kết quả phép phân tích này có tác dụng quan trọng để hoạch định hướng kiểm tra và giải pháp kỹ thuật sẽ áp dụng để kiểm tra; nghĩa là qua đó đã cho phép giảm rất nhiều khối lượng các thử nghiệm cần triển khai để xác định thuộc tính của chuỗi này.

8. CHƯƠNG TRÌNH HIỂN THỊ PHÂN TÍCH CẤU TRÚC KHÔNG GIAN CN3D

8.1. Đại cương

Cấu trúc không gian của tất cả các chất là một thuộc tính rất quan trọng quy định tính chất và đặc tính của chúng, đặc biệt là các vật liệu hữu cơ. Vì vậy, việc hiển thị, nghiên cứu, so sánh đặc điểm cấu trúc không gian này là yêu cầu và cũng là giải pháp giúp nhà khoa học phân tích và dự đoán được đặc tính của đối tượng nghiên cứu. Hướng vào mục tiêu trên, nhiều tác giả đã hoàn thiện và cung cấp cho người sử dụng các phần mềm ứng dụng khác nhau, thí dụ: chương trình hiển thị phân tích cấu trúc Cn3D www.ncbi.nlm.nih.gov/ Structure/CN3D/cn3dtut.shtml#cn3d,
Rasmol (http://www.bernstein-plus-sons.com/software/rasmol/ChangeLog.html),
Protein Explorer (http://www.umass.edu/microbio/chime/explorer),
PDB Lite (http://www.umass.edu/microbio/rasmol/pdblite.htm),
DRuMS Standard Color Scheme for Macromolecules
(http://www.umass.edu/molvis/drums)...

Cn3D là chương đồ hoạ hiển thị cấu trúc không gian của các phân tử sinh học, cấu trúc không gian của chuỗi amino axit và các công cụ để phân tích cấu trúc của chúng, được NCBI cung cấp miễn phí cho người sử dụng. Người phân tích có thể sử dụng chương trình này để vẽ ảnh hay hiển thị cấu trúc không gian của phân tử protein tương ứng với chuỗi phân tích, để hiển thị so sánh cấu trúc không gian giữa các phân tử, hay để phân tích, dự đoán tính trạng của chúng; thí dụ tìm kiếm vùng cấu trúc bị đột biến hay vùng bảo toàn cấu trúc giữa các chuỗi gần gũi nhau. NCBI cung cấp cho người

sử dụng đồng thời cả hai phương án khai thác là: dịch vụ Cn3D trực tuyến hay tải toàn bộ Cn3D về máy cá nhân phục vụ mục đích phân tích tại chỗ.

Để xác định cấu trúc phân tử , người ta thường sử dụng phương pháp phân tích khối phổ cộng hưởng từ hạt nhân (*Nuclear Magnet Responce Spectroscopy*) hay phương pháp phân tích nhiễu xạ Rơn-ghen (*X-Ray Crystallography*). NCBI đã sử dụng các dữ liệu kết quả phân tích thực nghiệm này làm cơ sở vật chất để xây dựng mảng dữ liệu cấu trúc MMDB (*Molecular Modeling DataBase*), nhằm góp phần làm phong phú thêm lượng thông tin truyền tải về chức năng sinh học, về cơ chế hoạt động của các phân tử và phục vụ cho mục tiêu nghiên cứu quan hệ giữa các phân tử có đặc điểm cấu trúc không gian gần gũi nhau. Như vậy, MMDB chỉ là mảng dữ liệu về cấu trúc không gian ba chiều trong kho tàng dữ liệu chung về protein PDB (MMDB được viết bằng ngôn ngữ ASN.1 (*Abstract Syntax Notation One*) và chương trình Cn3D được thiết kế trong môi trường này. Nghĩa là, chương trình Cn3D không đọc trực tiếp được dữ liệu chung từ PDB, mà trước hết dữ liệu này phải được dịch sang dạng ngôn ngữ giao tiếp MMDB). Về giao diện, chương trình được thiết kế nhằm cung cấp cho người sử dụng ảnh không gian ba chiều của đối tượng ở mọi kích thước, mọi góc độ theo yêu cầu.

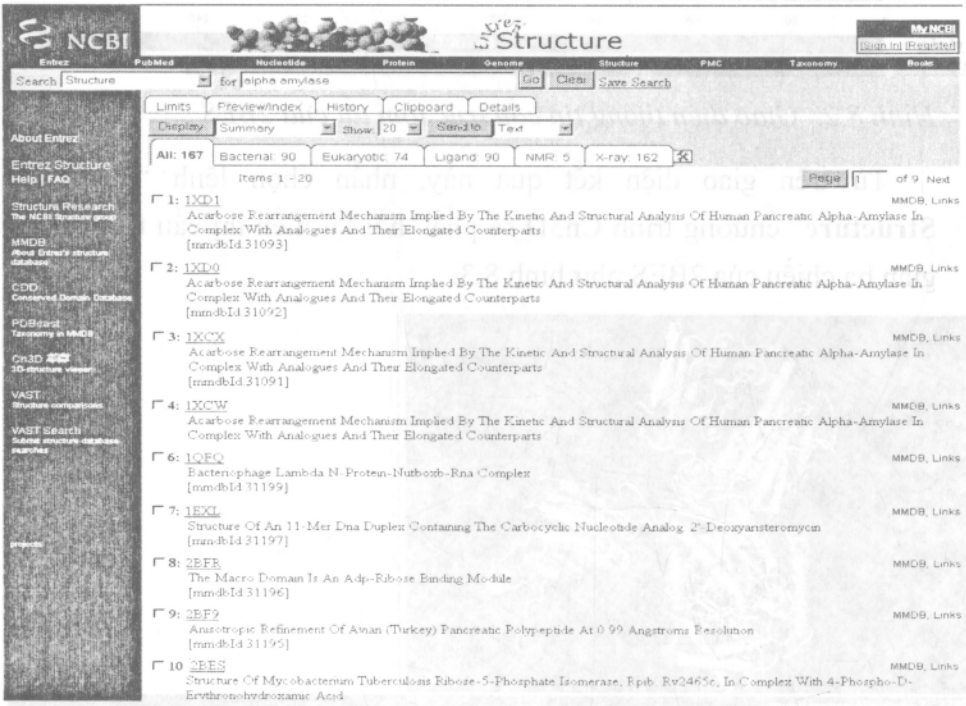
8.2. Sử dụng chương trình

Để hiển thị cấu trúc không gian từ dữ liệu MMDB, đầu tiên, người ta phải tải chương trình Cn3D về và cài đặt vào máy của mình. Sau đó, có thể sử dụng nhiều con đường khác nhau để hiển thị hình ảnh cấu trúc chuỗi bằng chương trình Cn3D. Khi vào trong chương trình này người phân tích

có thể sử dụng các lệnh tương ứng để thay đổi chế độ hiển thị, theo mục tiêu phân tích. Sau đây là bốn giải pháp thường áp dụng trong NCBI.

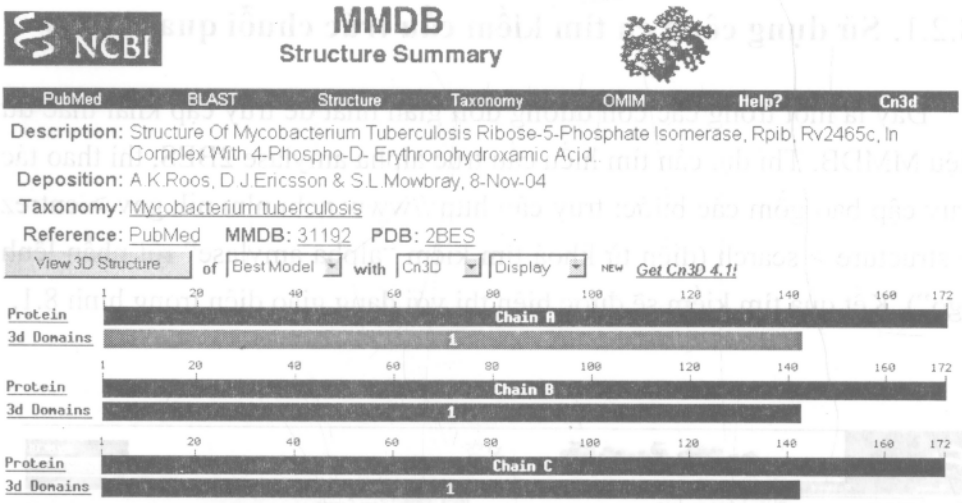
8.2.1. **Sử dụng công cụ tìm kiếm cấu trúc chuỗi qua Entrez**

Đây là một trong các con đường đơn giản nhất để truy cập khai thác dữ liệu MMDB. Thí dụ, cần tìm hiểu cấu trúc alpha amylase 2BES, thì thao tác truy cập bao gồm các bước: truy cập http://www.ncbi.nlm.nih.gov > entrez > structure > search (điền từ khoá tìm kiếm "alpha amylase" rồi nhấn lệnh "go"). Kết quả tìm kiếm sẽ được hiển thị với dạng giao diện trong hình 8.1.



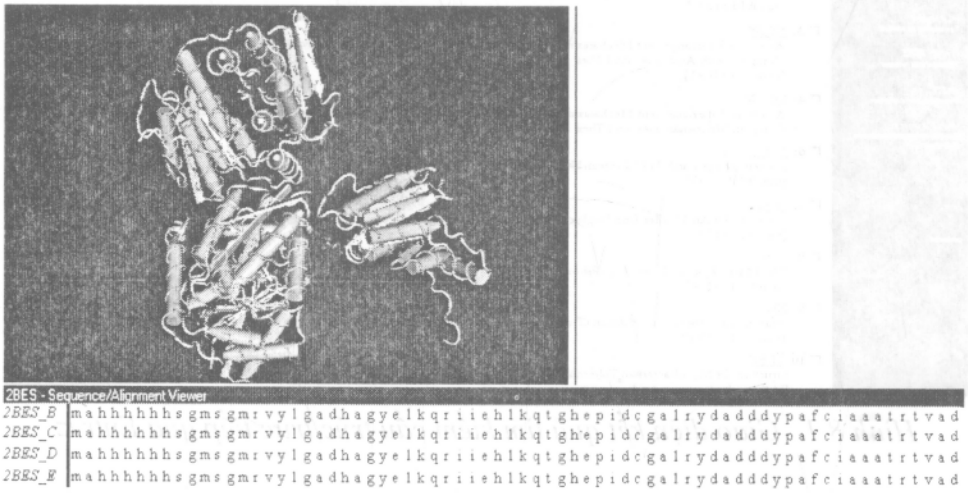
Hình 8.1. Giao diện kết quả tìm kiếm cấu trúc trực tiếp qua Entrez

Tiếp theo, nhấn chọn vào một trong hai đường dẫn siêu liên kết **2BES** hay **MMDB** (phía bên phải dòng tin). Sau đó, chương trình tìm kiếm cấu trúc **Entrez Structure** sẽ phản hồi lại kết quả với dạng giao diện như sau:



Hình 8.2. Giao diện thông tin cấu trúc tóm tắt của 2BES

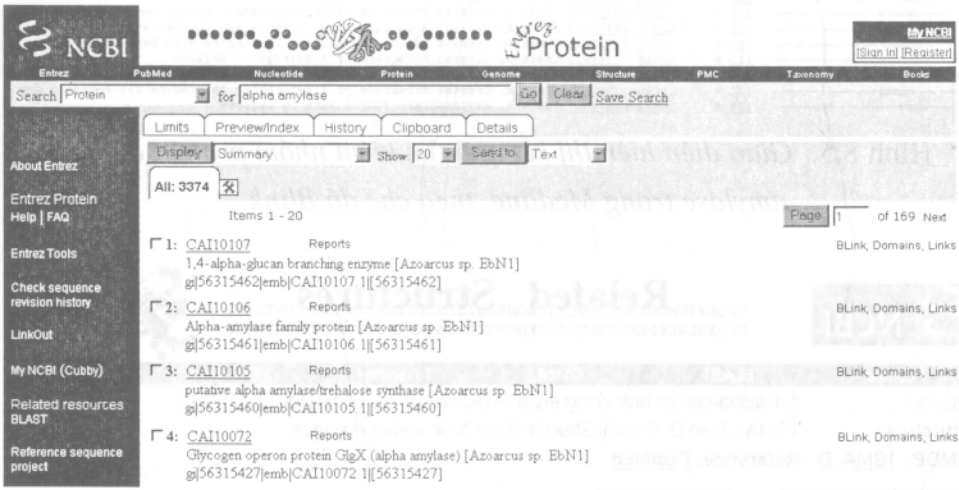
Từ trên giao diện kết quả này, nhấn chọn lệnh "**View 3D Structure**" chương trình Cn3D sẽ phản hồi lại hiển thị cấu trúc không gian ba chiều của 2BES như hình 8.3.



Hình 8.3. Cấu trúc không gian ba chiều của alpha amylase 2BES

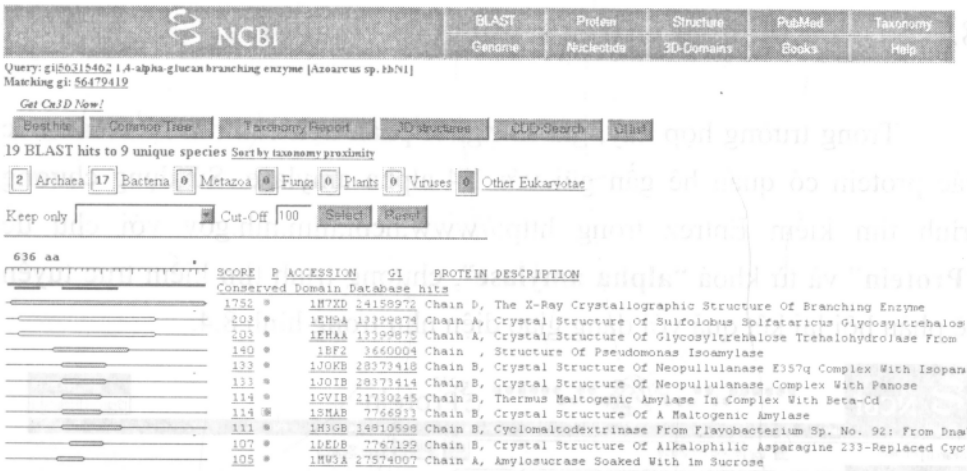
8.2.2. Từ dịch vụ entrez sequence neighbor

Trong trường hợp này, giả sử người phân tích cần tìm hiểu cấu trúc các protein có quan hệ gần gũi với 1,4-alpha amylase. Sử dụng chương trình tìm kiếm Entrez trong http://www.ncbi.nlm.nih.gov với chủ đề "**Protein**" và từ khoá "**alpha amylase**", chương trình tìm kiếm trực tuyến sẽ phản hồi lại kết quả với dạng giao diện như trong hình 8.4.

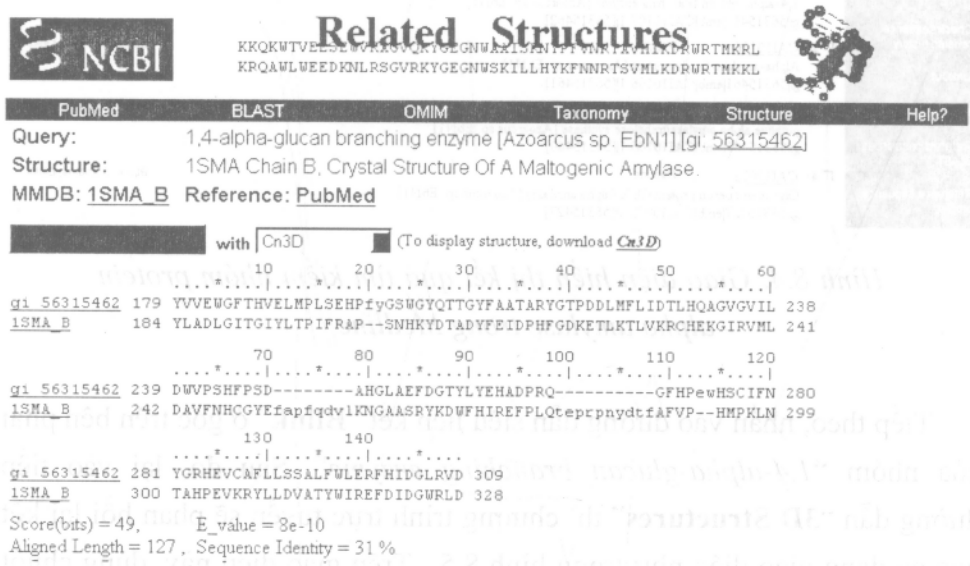


Hình 8.4. Giao diện hiển thị kết quả tìm kiếm nhóm protein alpha amylase trong Medline

Tiếp theo, nhấn vào đường dẫn siêu liên kết "**Blink**" ở góc trên bên phải của nhóm "*1,4-alpha-glucan branching enzyme*". Sau đó, lại vào tiếp đường dẫn "**3D Structures**" thì chương trình trực tuyến sẽ phản hồi lại kết quả có dạng giao diện như trong hình 8.5. Trên giao diện này, dùng chuột kích hoạt vào đường dẫn siêu liên kết tại vị trí có điểm tròn nhỏ màu nhạt tương ứng với cấu trúc chuỗi có đặc tính cần lựa chọn (thí dụ chuỗi cấu trúc tinh thể của một maltogenic amylase, với mã hiệu chuỗi là 1SMA_B). Sau thao tác đó, người phân tích trực tuyến sẽ nhận được kết quả phản hồi lại với giao diện như trong hình 8.6.



Hình 8.5. *Giao diện hiển thị kết quả tìm kiếm nhóm protein alpha amylase trong Medline, theo chế độ Blink*



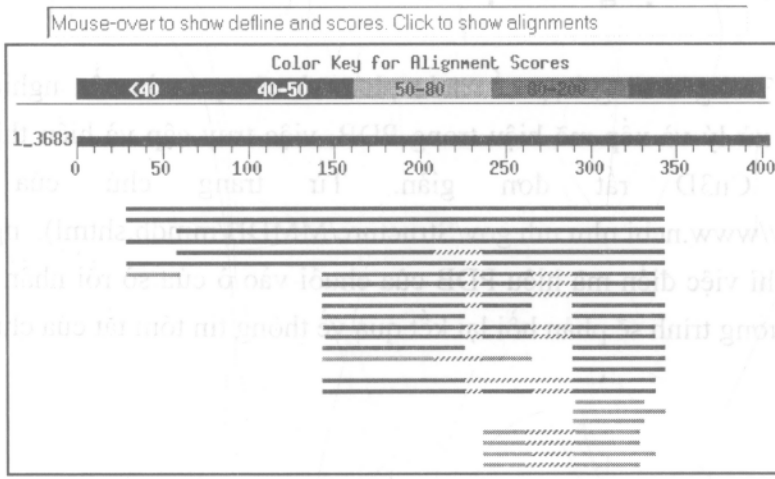
Hình 8.6. Nhóm chuỗi tương đồng cấu trúc với 1SMA_B

Trong giao diện này chỉ rõ cấu trúc chuỗi 1SMA_B và một chuỗi có cấu trúc gần gũi với chúng là 1,4-alpha-glucan branching enzyme (từ vi khuẩn *Azoarcus sp.* EbN1, với mã hiệu là gi: ***56315462***). Để hiển thị cấu trúc chuỗi của nhóm này chỉ việc nhấn chuột vào cửa sổ "**View 3D Structure**", rồi sử

dụng các công cụ trên giao diện chương trình Cn3D để thay đổi chế độ hiển thị.

8.2.3. Từ dịch vụ phân tích cấu trúc chuỗi BLAST

Chương trình phân tích cấu trúc BLAST cung cấp cho người sử dụng cả dịch vụ kết nối trực tiếp với chương trình hiển thị cấu trúc Cn3D ngay trong quá trình phân tích cấu trúc protein. Giả sử, người phân tích đang sử dụng chương trình "**Protein-Protein BLAST**" với chuỗi phân tích mang mã hiệu là **gi 54696134** và nhận được kết quả phản hồi với giao diện như trong hình 8.7.



Sequences producing significant alignments:		Score (bits)	E Value	
gi|47894397|ref|NM_001001491.1|	Mus musculus tropomyosin 4 ...	373	e-100	G U
gi|23958443|gb|BC023701.1|	Mus musculus tropomyosin 4, mRNA...	373	e-100	G U E
gi|21618969|gb|BC032174.1|	Mus musculus cDNA clone IMAGE:53...	373	e-100	G U
gi|21327808|gb|BC032175.1|	Mus musculus cDNA clone IMAGE:53...	373	e-100	G U
gi|56031571|dbj|AK207394.1|	Mus musculus cDNA, clone:Y2G010...	315	5e-83	U
gi|6981671|ref|NM_012678.1|	Rattus norvegicus tropomyosin 4...	230	2e-57	G U E
gi|207503|gb|J02780.1|RATTRO4IS	Rat tropmyosin (TM-4) mRNA,...	230	2e-57	G U E
gi|56030266|dbj|AK206089.1|	Mus musculus cDNA, clone:Y2G010...	216	3e-53	U

Hình 8.7. Giao diện hiển thị kết quả Protein-Protein BLAST

Kết quả này cho biết, trên vị trí đầu tiên là nhóm chuỗi có chỉ số **Score** và **E-value** tương đồng cao nhất với chuỗi kiểm tra. Tiếp theo, vào đường dẫn siêu liên kết của chuỗi để hiển thị thông tin tóm tắt về nhóm này. Giả sử nhấn chuột lựa chọn nhóm "**Mus Musculus Tropomyosin 4**"; tiếp theo vào "**Blink**" trong giao diện kết quả; rồi chọn tiếp đường dẫn "**3D Structure**" chương trình sẽ hiển thị các chuỗi protein trong ngân hàng dữ liệu MMDB có cấu trúc gần gũi với chuỗi kiểm tra. Sau đó kích chuột vào đường dẫn siêu liên kết tại vị trí vòng tròn màu nhạt rồi thao tác tiếp tương tự như mục 8.2.2 ở trên.

8.2.4. Sử dụng mã hiệu chuỗi PDB Identifier

Trong trường hợp cấu trúc phân tử của protein cần nghiên cứu đã được xử lý và xếp mã hiệu trong PDB, việc truy cập và hiển thị cấu trúc nhờ Cn3D rất đơn giản. Từ trang chủ của MMDB (http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml), người phân tích chỉ việc điền mã hiệu PDB của chuỗi vào ô cửa sổ rồi nhấn lệnh "**go**" là chương trình sẽ phản hồi lại kết quả về thông tin tóm tắt của chuỗi đó.

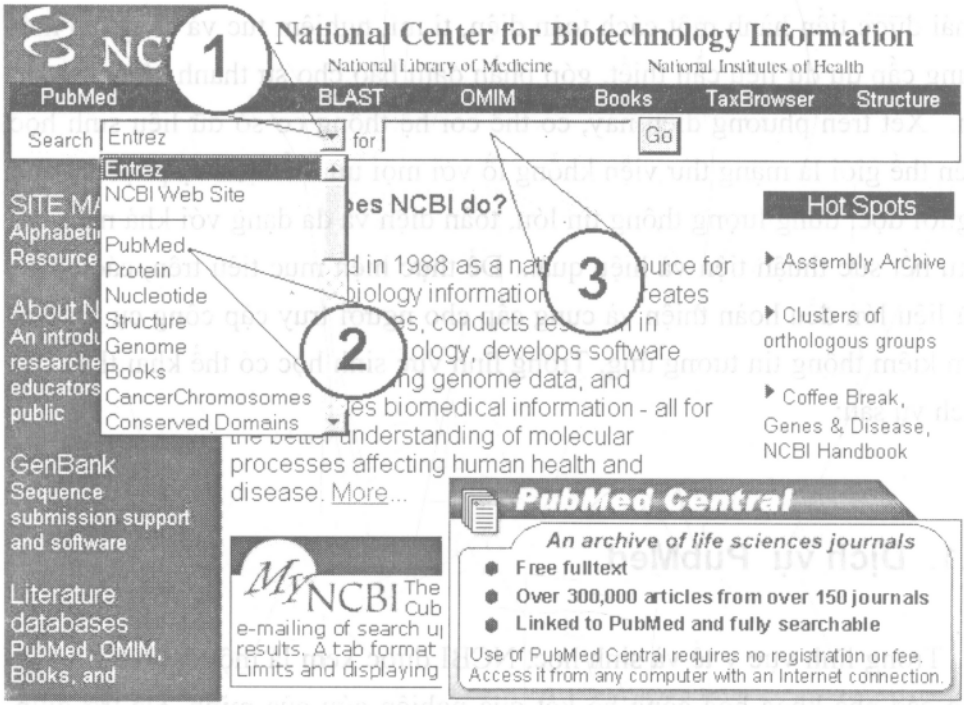
9. TRA CỨU TÀI LIỆU QUA INTERNET

Mọi dự án hay kế hoạch triển khai hoạt động nghiên cứu khoa học đều được bắt đầu bằng giai đoạn thu thập và nghiên cứu tài liệu. Công tác này phải được tiến hành một cách toàn diện, tỉ mỉ, nghiêm túc và sáng tạo mới cung cấp đủ dữ liệu cần thiết, góp phần đảm bảo cho sự thành công của dự án. Xét trên phương diện này, có thể coi hệ thống cơ sở dữ liệu sinh học trên thế giới là mạng thư viện khổng lồ với mọi ưu thế dịch vụ phục vụ cho người đọc: dung lượng thông tin lớn, toàn diện và đa dạng với khả năng tra cứu hết sức thuận tiện và hiệu quả... Để thực hiện mục tiêu trên, các cơ sở dữ liệu lớn đều hoàn thiện và cung cấp cho người truy cập công cụ tra cứu tìm kiếm thông tin tương ứng. Trong lĩnh vực sinh học có thể khai thác các dịch vụ sau:

9.1. Dịch vụ PubMed

Trong lĩnh vực y tế và sinh học, NCBI được xem là một địa chỉ tin cậy cho các nhà khoa học công bố kết quả nghiên cứu của mình. Để trợ giúp khách hàng khai thác nhóm dữ liệu này, NCBI đã hoàn thiện và cung cấp cho khách hàng công cụ dịch vụ tìm kiếm thông tin **PubMed** và **PubMed Central**. Dịch vụ PubMed cung cấp cho người khai thác thông tin (đầy đủ hoặc tóm lược) của tất cả các công trình khoa học đã công bố trong

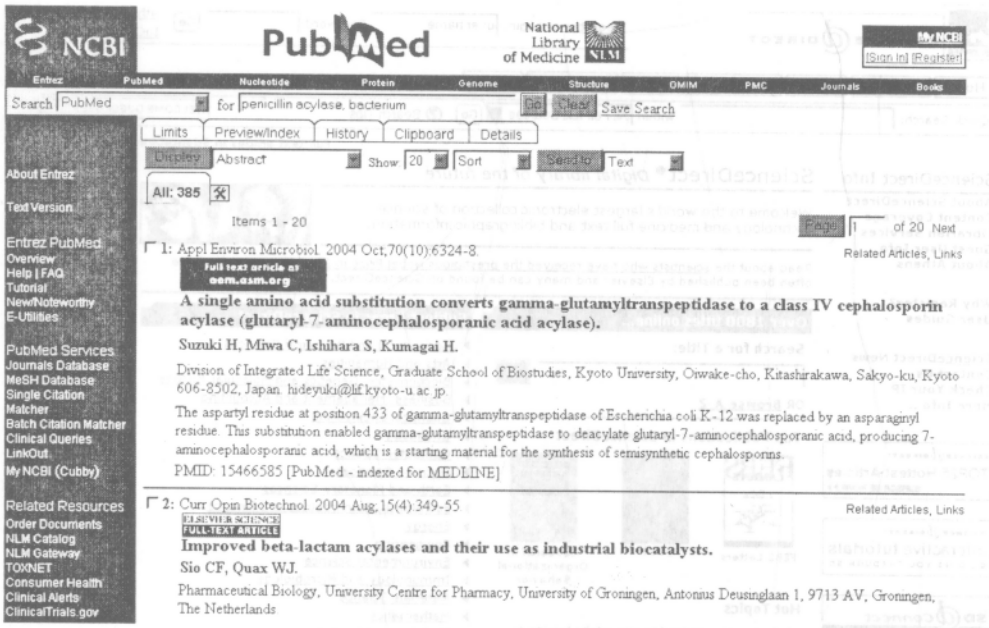
MEDLINE và các công trình liên quan của cùng tác giả hay các công trình của tác giả khác có cùng chủ đề tìm kiếm. Với dịch vụ *PubMed Central*, NCBI còn cung cấp thêm cho người truy cập cả thông tin của các công trình khoa học sắp phát hành (do một số nhà xuất bản cung cấp để giới thiệu trước, dưới dạng thông tin tóm tắt gửi cho PubMed). Với ưu thế to lớn và đa dạng về cơ sở dữ liệu, PubMed hiện được xem là một trong các công cụ tìm kiếm phổ dụng nhất trong lĩnh vực công nghệ sinh học. Để sử dụng dịch vụ này, cần phải truy cập trang chủ của NCBI rồi thao tác qua các bước là lựa chọn PubMed (kích chuột vào vị trí 1, sau đó kích chuột vào vị trí 2 để chọn PubMed) và cung cấp thông tin tìm kiếm (bước 3 - xem hình 9.1).



Hình 9.1. Giao diện tra cứu tài liệu PubMed trong NCBI

Sau khi cung cấp dữ liệu tìm kiếm, dạng số hay ký tự, người tìm tin chỉ việc nhấn lệnh "Go" để gửi yêu cầu đi. Chương trình PubMed sẽ tìm kiếm và gửi kết quả phản hồi lại cho người tìm tin. Người tìm tin có thể thay đổi

chế độ hiển thị khác nhau theo nhu cầu (lựa chọn tại cửa sổ **Display**). Giao diện kết quả tìm kiếm thông tin dạng tóm lược như sau:



Hình 9.2. Giao diện kết quả tìm tin qua PubMed

Trên giao diện kết quả này, người tìm tin chỉ cần nhấn chuột vào vị trí các đường dẫn siêu liên kết là có thể tải về được các tệp tin mong muốn. Trong nhiều trường hợp, người tìm tin có thể được quyền tải miễn phí toàn bộ nội dung công trình công bố hoàn chỉnh (*full text article*).

9.2. Dịch vụ thư viện qua mạng ScienceDirect®

ScienceDirect® là thương hiệu dịch vụ thư viện qua mạng internet của *Elsevier Copr.* ScienceDirect® được xem là một trong số rất ít địa chỉ cung cấp dịch vụ thông tin lớn nhất thế giới, với khoảng 60 triệu tin tóm lược các công trình khoa học kèm theo đường dẫn siêu liên kết tới công bố trong

khoảng 1800 tạp chí khoa học của trên 170 nhà xuất bản khác nhau trên thế giới, trong đó rất nhiều tài liệu có thể tải về dưới dạng nội dung công bố hoàn chỉnh. Giao diện trang chủ này có dạng như sau:

Hình 9.3. Giao diện trang chủ của ScienceDirect®

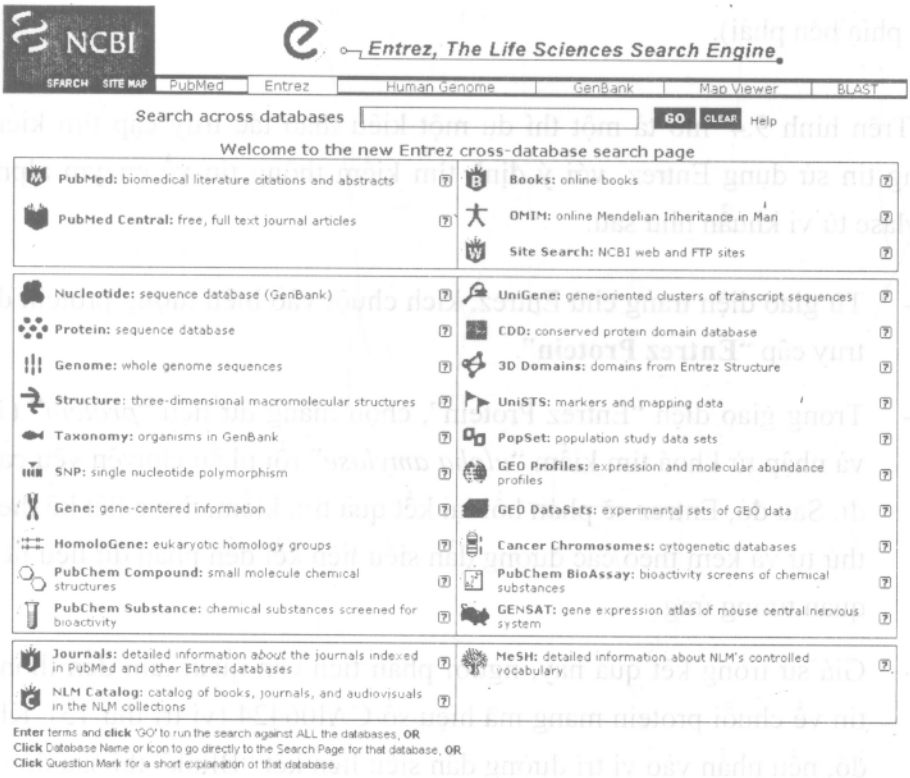
Thông tin của cơ sở dữ liệu này trải rộng trên hầu hết mọi lĩnh vực khoa học và công nghệ và dịch vụ cung cấp tin này miễn phí truy cập cho cá nhân và trả phần trăm cho các cơ sở có nhu cầu in lại nên rất hữu dụng cho các nước đang phát triển. Từ 02/01/2005, Elsevier đã liên kết hợp tác với hai tổ chức lớn trên thế giới là hiệp hội sách các trường đại học Mỹ (*US Sabre Foundation* - www.sabre.org) và tổ chức hỗ trợ sách cho các nước nghèo của Cộng đồng Châu Âu (*Europian Book Aid International* - www.bookaid.org)

Để sử dụng dịch vụ này, khách hàng chỉ cần đăng ký trực tuyến với ScienceDirect® để được cấp quyền truy cập (đường dẫn siêu liên kết

"register" nằm ở phía trên bên phải của giao diện). Sau khi đăng ký, ScienceDirect® sẽ tự động cấp cho khách hàng tên và mật khẩu (do khách hàng đăng ký) để sử dụng truy cập sau này. Khi đã ở trong ScienceDirect®, việc tra cứu và khai thác thông tin cũng tương tự như trong phần 9.1.

Ngoài ra, nhiều cơ sở dữ liệu khác cũng cung cấp cho khách hàng có thể khai thác khả năng cung cấp thông tin tư liệu đủ mạnh tương tự, thí dụ: HighWire (www://highWire.stanford.edu) hay
Scirus (http://www.scirus.com)

9.3. Dịch vụ Entrez của NCBI và SRS của EBI



Hình 9.4. Giao diện dịch vụ Entrez của NCBI

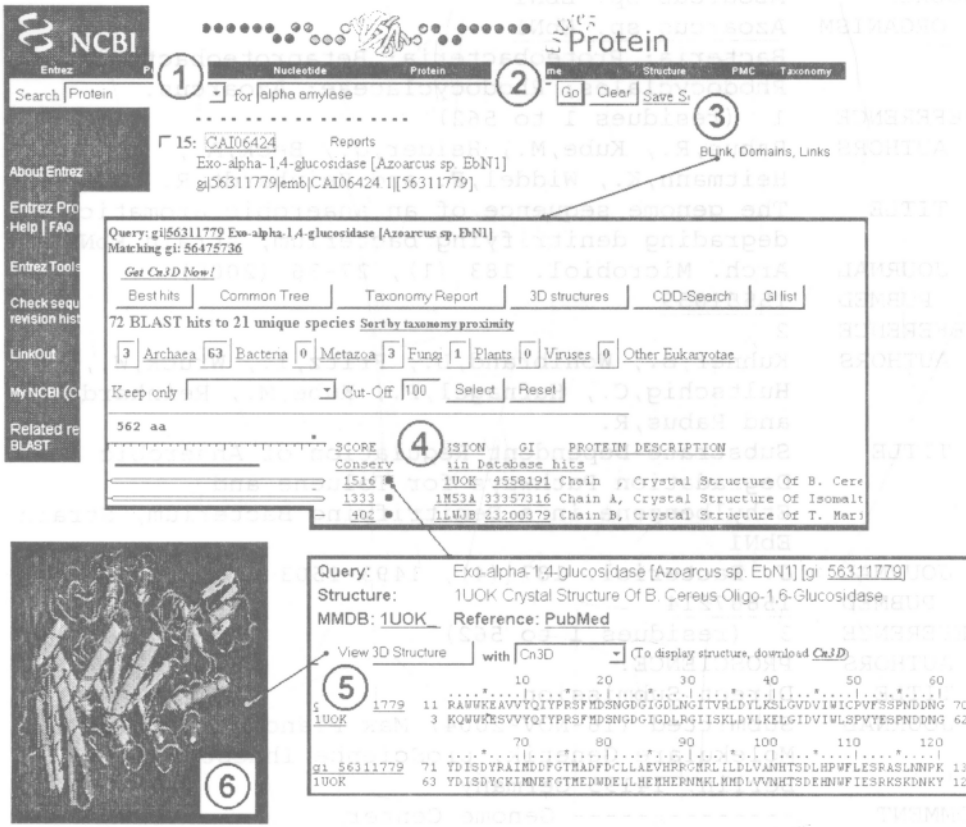
Entrez là dịch vụ quản lý và liên kết thông tin của cơ sở dữ liệu NCBI. Giao diện trang chủ của dịch vụ này có dạng như trong hình 9.4. Với cơ cấu tổ chức quản lý thông tin theo từng mảng riêng biệt của NCBI, dịch vụ này đảm nhiệm vai trò kết nối liên thông giữa các mảng thành một tổng thể hữu cơ, giúp cho người truy cập tiếp cận nhanh và đầy đủ các thông tin tìm kiếm. Thí dụ, đang từ cơ sở dữ liệu tư liệu PubMed khách hàng dễ dàng truy cập sang mảng dữ liệu về cấu trúc chuỗi xoắn kép DNA hay chuỗi nucleotide trong PDB, hoặc chuyển qua mảng dữ liệu về cấu trúc không gian protein MMDB, nhờ đường dẫn siêu liên kết do Entrez thiết lập. Người tìm tin có thể truy cập trực tiếp cơ sở dữ liệu bằng tìm kiếm theo từ khoá (điền từ khoá vào ô cửa sổ: "search across databases") hay qua các đường dẫn siêu liên kết vào các mảng dữ liệu rồi mới tìm kiếm thông tin sau (kích chuột vào vị trí biểu tượng của mảng dữ liệu hay dấu hỏi tương ứng phía bên phải).

Trên hình 9.4 mô tả một thí dụ một kiểu thao tác truy cập tìm kiếm thông tin sử dụng Entrez, với ý định tìm kiếm thông tin về enzym alpha amylase từ vi khuẩn như sau:

- Từ giao diện trang chủ Entrez, kích chuột vào biểu tượng protein để truy cập "**Entrez Protein**".
- Trong giao diện "Entrez Protein", chọn mảng dữ liệu "*protein*" (1) và nhập từ khoá tìm kiếm "*alpha amylase*" rồi nhấn chuyển yêu cầu đi. Sau đó, Entrez sẽ phản hồi lại kết quả tìm kiếm, được liệt kê theo thứ tự và kèm theo các đường dẫn siêu liên kết đến phần dữ liệu liên quan tương ứng.
- Giả sử trong kết quả này, người phân tích cần quan tâm đến thông tin về chuỗi protein mang mã hiệu số CAI06424 (vị trí thứ 15). Khi đó, nếu nhấn vào vị trí đường dẫn siêu liên kết "Blink" thì sau đó sẽ hiển thị kết quả về so sánh chuỗi BLAST.

- Trong giao diện kết quả BLAST này, nếu kích hoạt vị trí vòng tròn nhỏ nhạt màu của đường dẫn đến cơ sở cấu trúc (4) thì Entrez sẽ trả lời thông tin về mức độ tương đồng của chuỗi CAI06424 với các chuỗi lưu giữ trong cơ sở dữ liệu của NCBI.

- Để hiển thị nghiên cứu hình ảnh, chỉ cần kích hoạt đường dẫn "**View 3D Structure**" (5) sẽ hiển thị được ảnh để so sánh cấu trúc không gian của chúng (6 - sau đó, sử dụng thanh công cụ trong giao diện Cn3D để thay đổi chế độ hiển thị khi phân tích đánh giá).



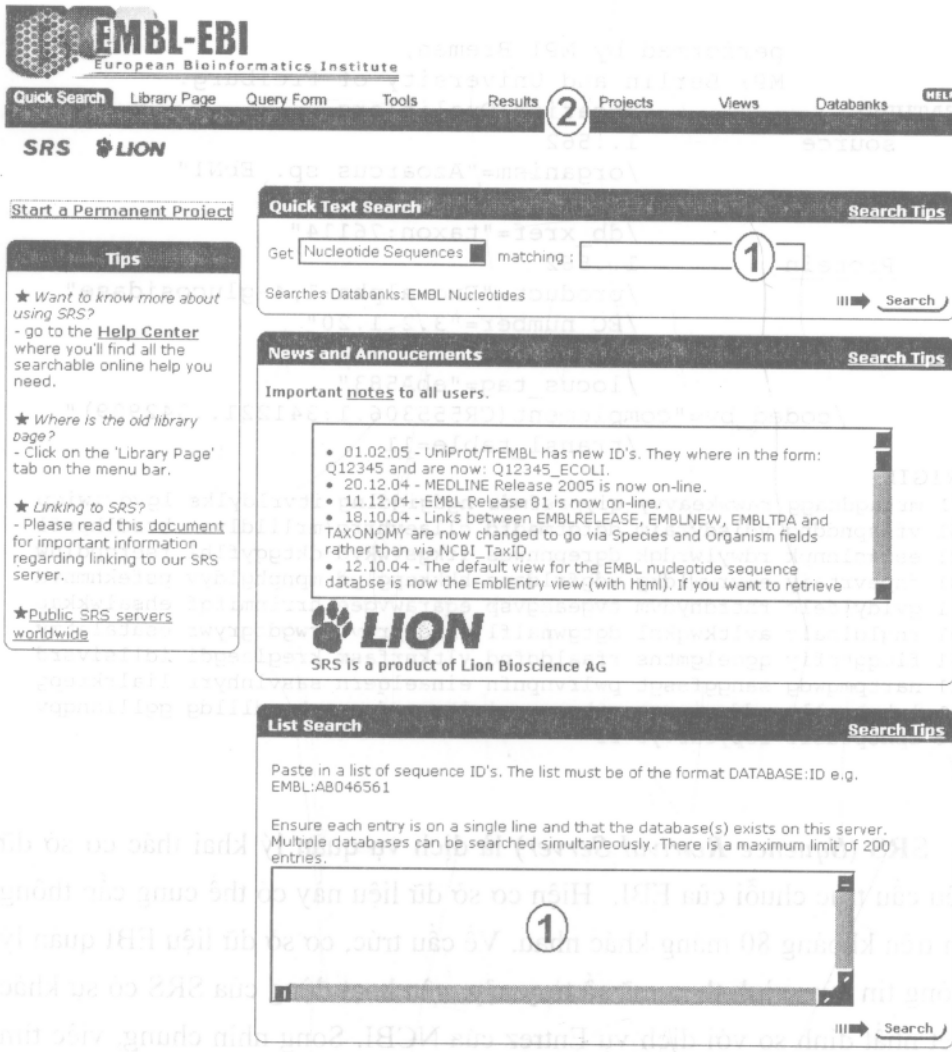
Hình 9.5. Sơ đồ một dạng khai thác dữ liệu qua dịch vụ Entrez (6 bước)

- Trường hợp kích hoạt vào đường dẫn siêu liên kết mã hiệu chuỗi (tại vị trí ghi mã số CAI06424), người truy cập có thể hiển thị thông tin về cấu trúc và đặc tính của chuỗi này, chế độ ký tự có dạng như sau:

```
CAI06424. Reports Exo-alpha-1,4-glu...[gi:5631779] Blink Domain Link
LOCUS       CAI06424      562 aa    linear   BCT 09-FEB-2005
DEFINITION  Exo-alpha-1,4-glucosidase [Azoarcus sp. EbN1].
ACCESSION   CAI06424
VERSION     CAI06424.1  GI:56311779
DBSOURCE    embl accession CR555306.1
KEYWORDS    .
SOURCE      Azoarcus sp. EbN1
  ORGANISM  Azoarcus sp. EbN1
            Bacteria; Proteobacteria; Betaproteobacteria;
            Rhodocyclales; Rhodocyclaceae; Azoarcus.
REFERENCE   1  (residues 1 to 562)
  AUTHORS   Rabus,R., Kube,M., Heider,J., Beck,A.,
            Heitmann,K., Widdel,F. and Reinhardt,R.
  TITLE     The genome sequence of an anaerobic aromatic-
            degrading denitrifying bacterium, strain EbN1
  JOURNAL   Arch. Microbiol. 183 (1), 27-36 (2005)
   PUBMED   15551059
REFERENCE   2
  AUTHORS   Kuhner,S., Wohlbrand,L., Fritz,I., Wruck,W.,
            Hultschig,C., Hufnagel,P., Kube,M., Reinhardt,R.
            and Rabus,R.
  TITLE     Substrate-Dependent Regulation of Anaerobic
            Degradation Pathways for Toluene and
            Ethylbenzene in a Denitrifying Bacterium, Strain
            EbN1
  JOURNAL   J. Bacteriol. 187 (4), 1493-1503 (2005)
   PUBMED   15687214
REFERENCE   3  (residues 1 to 562)
  AUTHORS   PROSCIENCE.
  TITLE     Direct Submission
  JOURNAL   Submitted (16-NOV-2004) Max Planck Institut Fuer
            Molekulare Genetik, proScience Ihnestrasse 73,
            Berlin, 14195 Germany
COMMENT     -------------- Genome Center
            Center: Max-Planck-Institute for Molecular
             Genetics Center code:     MPIMG
            see http://www.micro-genomes.mpg.de/ebn1/ for
            detailed gene
            annotation annotation of the genome was
```

```
            performed by MPI Bremen,
            MPI Berlin and University of Freiburg.
FEATURES             Location/Qualifiers
     source          1..562
                     /organism="Azoarcus sp. EbN1"
                     /strain="EbN1"
                     /db_xref="taxon:76114"
     Protein         1..562
                     /product="Exo-alpha-1,4-glucosidase"
                     /EC_number="3.2.1.20"
     CDS             1..562
                     /locus_tag="ebA583"
        /coded_by="complement(CR555306.1:341221..342909)"
                     /transl_table=11
ORIGIN
  1 mrtagdnagg rawwkeavvy qiyprsfmds ngdgigdlng itvrldylks lgvdviwicp
 61 vfsspnddng ydisdyraim ddfgtmadfd cllaevhrrg mrlildlvan htsdlhpwfl
121 esraslnnpk rdwylwrdgk dgrepnnwes ifkgsvwkyd dktgqyflhl fserqpdlnw
181 dnpevrtaiy emvrwwldkg vdgfrldavs hkkkepglpd mpnphgldyv psfekhmnvd
241 gvldyldelc rhtfdhydvm tvgeangvsp eqarawvgee hrrlnmifqf ehsalwkkap
301 rngldlpalr avltkwqksl dgtgwnalfl enhdlprvvs rwgdtgrywr esatalatmy
361 flmqgtpfiy qgqelgmtns rfaalddfnd vltknrfasm kreglaegdi idllsivsrd
421 nartpmqwdg sanggfssgt pwlrvnpnfn einaelgern sasvlnhyrr lialrkrepg
481 lvhghcellm eddaqiyays rthgsqryvi itnmtrdear yhqpdllldg gglllanqpv
541 ephtptdsll lepyearvyr fr
//
```

SRS (***Squence Retrival Server***) là dịch vụ quản lý khai thác cơ sở dữ liệu cấu trúc chuỗi của EBI. Hiện cơ sở dữ liệu này có thể cung cấp thông tin trên khoảng 80 mảng khác nhau. Về cấu trúc, cơ sở dữ liệu EBI quản lý thông tin của mình theo mã số truy cập, nên hoạt động của SRS có sự khác biệt nhất định so với dịch vụ Entrez của NCBI. Song nhìn chung, việc tìm kiếm thông tin cũng thuận tiện và đơn giản tương tự như sử dụng công cụ Entrez trình bày trong phần trên. Hình 9-6 mô tả giao diện trang chủ của dịch vụ SRS. Từ giao diện này, khách hàng có thể dùng từ khoá hay số mã hiệu thông tin để truy cập thẳng đến cơ sở dữ liệu tương ứng (1), hoặc kích hoạt biểu tượng nhóm dịch vụ tương ứng trên thanh công cụ (2) để truy cập thẳng vào nhóm này.



Hình 9.6. Giao diện trang chủ dịch vụ SES của EBI

10. KHAI THÁC THÔNG TIN CƠ SỞ DỮ LIỆU CẤU TRÚC ĐỂ THIẾT KẾ GEN

10.1. Cơ sở dữ liệu RFLP (*Restriction Fragment Length Polymorphism*) và cơ sở dữ liệu ESTs (*Expressed Sequence Tags*)

Trong quá trình sinh tổng hợp protein, thông tin di truyền từ nhân được phiên mã tổng hợp phân tử mRNA và, tiếp theo, quá trình dịch mã trên ribosom đã tổng hợp nên phân tử protein theo khuôn mẫu được xác định bởi chính trình tự cấu trúc gen tương ứng của chúng. Với nguyên lý một gen - một protein cho phép mở ra khả năng chỉ cần xác định cấu trúc của một trong ba phân tử gen-mRNA-protein là có thể suy ra cấu trúc của hai phân tử kia và ngược lại.

10.1.1. Cơ sở dữ liệu RFLP (*Restriction Fragment Length Polymorphism*)

Việc xác định trực tiếp cấu trúc DNA trong toàn bộ genom có thể được giải quyết nhờ kỹ thuật thiết lập bản đồ gen hạn chế cho từng chuỗi DNA của chúng, bằng cách sử dụng enzyme giới hạn cắt nhỏ sợi DNA thành các đoạn ngắn để xác định trình tự (theo phương pháp tổng hợp từng nucleotide một *Maxam-Gilbert* hay phương pháp tổng hợp giới hạn sử dụng dideoxy-

nucleotide *Sanger*) rồi tổ hợp các đoạn này theo phương án đúng với trình tự cắt thành chuỗi hoàn chỉnh. Do kích thước chuỗi xoắn kép DNA rất lớn (trên mỗi chuỗi có thể chứa hàng chục ngàn gen và mỗi gen có thể dài tới hàng chục ngàn nucleotide), nên việc thiết lập bản đồ gen hạn chế cho từng sinh vật nhất định cũng là cả khối lượng công việc đồ sộ, tốn kém tiền của và hết sức phức tạp; trong khi thế giới sống vốn đã hết sức đa dạng về các chủng loài và đặc tính sinh học của chúng. Nghĩa là trong tương lai gần người ta không thể thiết lập xong bản đồ gen hạn chế cho tất cả mọi sinh giới. Chính trong bối cảnh này, tin-sinh học đã trở thành phương tiện nghiên cứu vô cùng hiệu quả để giải quyết bài toán trên.

Trên phương diện lý thuyết, nếu lựa chọn được các phức hệ enzyme giới hạn thích hợp để cắt hệ genom của các sinh vật có quan hệ gần gũi nhau về di truyền thì người ta sẽ thu được ba nhóm phức hệ các phân đoạn nucleotide là:

- Nhóm các phân đoạn nucleotide giống nhau hoàn toàn, tương ứng với các đoạn cắt bảo toàn về di truyền,

- Nhóm các phân đoạn nucleotide có sự sai khác về cấu trúc (về số lượng hay về trình tự sắp xếp các nucleotide, tương ứng với các đoạn nucleotide có sự biến đổi nhất định về di truyền, có thể chỉ ở một vị trí hay ở nhiều vị trí khác nhau) và

- Nhóm các phân đoạn khác nhau hoàn toàn (do sự biến đổi di truyền có thể gây ra hiện tượng đứt gãy hay chèn vào toàn bộ các phân đoạn hoàn chỉnh này).

Mở rộng quan hệ trên khi phân tích các sinh vật có quan hệ xa hơn về mặt di truyền (và cuối cùng là mở rộng ra cho toàn bộ sinh giới) thì vẫn tồn tại cả ba nhóm phân đoạn cấu trúc dạng như vậy (với xu hướng nếu quan hệ

họ hàng về di truyền càng xa nhau thì số lượng các phân đoạn thuộc nhóm đầu có thể giảm đi, sự sai khác trong nhóm giữa càng đa dạng hơn và số lượng các phân đoạn thuộc nhóm cuối có xu hướng tăng lên). Điều đó có nghĩa là về lý thuyết, nếu lựa chọn được bộ phức hệ các enzyme giới hạn phù hợp thì khi cắt hệ enzyme của mỗi sinh giới sẽ tạo ra một tập hợp các phân đoạn nucleotide tương ứng nhất định đặc trưng cho bản chất di truyền của loài. Để từ đó, căn cứ vào đặc tính cấu trúc của các đoạn chuỗi và mối quan hệ tương đồng (hay phân ly) giữa các phân đoạn chuỗi tương ứng với nhau trong các tập hợp này để nhận biết xác định hay để phân loại sinh giới. Trong thực tế, do kích thước của các phân tử DNA rất lớn nên ngay khi sử dụng một số hữu hạn các enzyme giới hạn để cắt hệ genom các sinh vật khác nhau (chưa cần với toàn bộ sinh giới) đã tạo ra được ba bộ tập hợp rất lớn các nhóm phân đoạn trên, đủ để xây dựng một ngân hàng dữ liệu khổng lồ về cấu trúc chuỗi. Khi khối lượng dữ liệu từng nhóm chuỗi đủ lớn, thì trong chừng mực nhất định, người ta có thể dựa vào cơ sở dữ liệu ngân hàng này tiến hành phân tích cấu trúc gen của đối tượng thử nghiệm. Các luận điểm phân tích trên chính là cơ sở lý luận cho sự ra đời và phát triển hết sức mạnh mẽ của mảng cơ sở dữ liệu đa dạng cấu trúc phân đoạn cắt RFLP (*Restriction Fragment Length Polymorphism*). Thư viện dữ liệu RFLP được hình thành trên cơ sở phân cắt ngẫu nhiên DNA của các sinh giới bằng các enzyme giới hạn thành các phân đoạn nhỏ rồi dùng kỹ thuật lai *Southern* sử dụng chỉ thị đánh dấu (dùng đồng vị phóng xạ nhân tạo ^{32}P hay chỉ thị phát xạ huỳnh quang *Enhanced Chemical Luminescence*), để nhận biết các phân đoạn này. Theo thời gian, số lượng các enzyme giới hạn được sử dụng tăng lên và việc phân tích không ngừng mở rộng ra các đối tượng sinh giới khác nhau nên dung lượng thông tin nhóm cơ sở dữ liệu này tăng hết sức nhanh chóng. Nhờ vậy, kỹ thuật phân tích cấu trúc chuỗi dựa trên cơ sở dữ liệu RFLP đã trở thành công cụ nghiên cứu mạnh của sinh học hiện đại. Đồng thời cùng với xu hướng trên, nhờ hoàn thiện kỹ thuật phân cắt hạn chế cấu trúc DNA kết hợp với sự phát triển và hoàn thiện kỹ

thuật lai đánh dấu đã làm xuất hiện thêm các cơ sở dữ liệu mới dạng này như: VNTR (*Variable Number of Tadem Repeat Units*), STRPs (*Short Tandem Repeat polymorphisms*), SSLPs (*Short Sequence Length Polymorphisms*)... Cấu trúc tệp thông tin trang 129 là tệp dữ liệu chuỗi mã hiệu số AY366356, dạng dữ liệu RFLP của NCBI.

10.1.2. Cơ sở dữ liệu ESTs (*Expresed Sequence Tags*)

Như phần đầu đã trình bày, cấu trúc của chuỗi phân tử gen-mRNA-protein còn có thể được xác định theo con đường thứ hai là xác định cấu trúc phân tử mRNA. Với sự ra đời của kỹ thuật nhân bản ngược chuỗi polymerase R-PCR (R*eversed PCR*), người ta dễ dàng tách chiết, tinh chế và xác định được cấu trúc chuỗi mRNA (dưới dạng cấu trúc của cDNA - *complementary* DNA). Sự tồn tại của các phân tử mRNA này trong tế bào gắn liền với quá trình tổng hợp ra các phân tử protein tương ứng. Do đó, việc xây dựng cơ sở dữ liệu cDNA đủ mạnh sẽ mở ra con đường dự đoán xác định nhanh chóng và hiệu quả các gen mang mã sinh tổng hợp protein.

Xét trên quan hệ giữa gen-mRNA, trên các đối tượng sinh vật nhân chưa hoàn thiện cấu trúc cDNA cũng chính là trình tự cấu trúc gen trên chuỗi DNA. Song đối với sinh vật nhân hoàn thiện, dữ liệu cấu trúc cDNA không thể phản ánh được cấu trúc đầy đủ của gen, do quá trình phiên mã đã bỏ qua các đoạn DNA-intron. Điều này rất có thể dẫn tới khả năng bỏ sót lượng thông tin quý giá mã hoá trong cấu trúc thực của DNA (trên đoạn intron hay các đoạn mang tín hiệu cảm ứng, kiềm toả nằm trong nhân...). Mặc dù có những hạn chế nhất định (do độ bền vững của cấu trúc đơn mRNA kém hơn hay nguy cơ nhiễm tạp vật liệu di truyền khác trong quá trình thao tác có thể dẫn đến làm sai lệch cấu trúc mRNA-cDNA), song về

tổng thể ưu thế về dung lượng thông tin truyền tải trong cấu trúc của các chuỗi cDNA là cơ bản và nổi trội (vượt xa rất nhiều so với những hạn chế của nó), trong khi kỹ thuật tổng hợp chúng lại tương đối đơn giản. Bên cạnh đó, khi dung lượng cơ sở dữ liệu càng lớn thì độ tin cậy của thông tin truyền tải càng cao (theo luật số lớn) và chất lượng quá trình phiên mã ngược sẽ cao hơn khi kích thước chuỗi không quá lớn. Nhờ vậy, chỉ trong thời gian rất ngắn, dung lượng các cơ sở dữ liệu chuỗi cDNA đã phát triển hết sức nhanh chóng và dần trở thành mảng dữ liệu cấu trúc chuỗi trọng yếu trong cơ sở dữ liệu GenBank của NCBI hiện nay. Cơ sở dữ liệu cDNA này được gọi là cơ sở dữ liệu ESTs (do để xác định cấu trúc các chuỗi cDNA phải qua bước biến nạp biểu hiện vào tế bào chủ). Quy trình thiết lập dữ liệu cấu trúc cDNA có thể tóm tắt qua các bước chính sau:

- Xử lý tách và tinh chế thu mRNA tinh sạch, từ nguồn nguyên liệu gen (tế bào hay mô mong muốn),

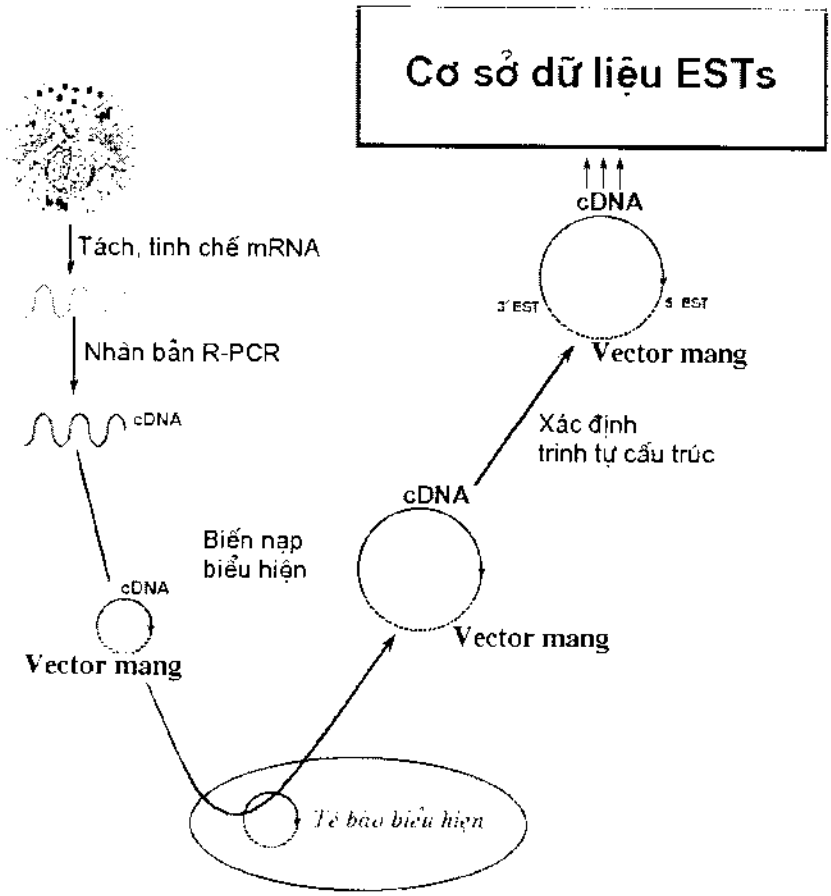
- Tổng hợp và tinh chế thu cDNA, sử dụng kỹ thuật thuật nhân bản ngược chuỗi polymerase R-PCR (phối hợp với kỹ thuật RACE - ***Rapid Amplication of cDNA Ends***) và điện di tách thu cDNA,

- Tạo dòng phân tử EST (gắn cDNA vào vectơ mang, biểu hiện vào tế bào chủ để thu dòng vector mang cDNA)

- Thiết lập dữ liệu dbEST (xác định cấu trúc dòng EST và đưa vào cơ sở dữ liệu ESTs).

Do đặc điểm lựa chọn và tạo dòng như trên nên chỉ có các chuỗi ngắn hoặc trung bình mới có thể biến nạp và xác định cấu trúc được, còn các chuỗi cDNA kích thước lớn không thể xác định được cấu trúc trực tiếp theo phương pháp mô tả trên). Chính vì lý do này nên các tệp dữ liệu EST phổ biến là thuộc các nhóm chuỗi có độ dài trung bình (khoảng 400 nucleotide, thường bằng hoặc ngắn hơn kích thước phân tử mRNA tương ứng, với đầu

3' thường là đoạn ologo(dT), còn đầu kia tương ứng với vùng 5' không phiên mã trên chuỗi mRNA; khi phối hợp kỹ thuật RACE sẽ tạo ra cấu trúc poly-A cho đầu kia). Trong trang 132 trình bày cấu trúc tệp dữ liệu chuỗi mRNA với số mã hiệu CX198004, dạng dữ liệu EST của NCBI.



Hình 10.1 Sơ đồ quy trình thiết lập dữ liệu ESTs

Cấu trúc dữ liệu chuỗi DNA mã số AY366356S1 - dạng dữ liệu cấu trúc RFLP:

```
LOCUS       AY366356S1              962 bp    DNA     linear   INV 01-SEP-2004
DEFINITION  Litopenaeus vannamei alpha-amylase gene, exons 1 through 4 and
            partial cds.
ACCESSION   AY366356
VERSION     AY366356.1  GI:38373486
KEYWORDS    .
SEGMENT     1 of 2
SOURCE      Litopenaeus vannamei (Pacific white shrimp)
  ORGANISM  Litopenaeus vannamei
            Eukaryota; Metazoa; Arthropoda; Crustacea; Malacostraca;
            Eumalacostraca; Eucarida; Decapoda; Dendrobranchiata; Penaeoidea;
            Penaeidae; Litopenaeus.
REFERENCE   1  (bases 1 to 962)
  AUTHORS   Glenn,K.L., Suwanasopee,T., Sornthep,T., Rothschild,M.F. and
            Harris,D.L.
  TITLE     Polymorphisms in the alpha-amylase and cathepsin-L genes in
            Litopenaeus vannamei, detectable by PCR-RFLP analysis
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 962)
  AUTHORS   Glenn,K.L., Suwanasopee,T., Sornthep,T., Rothschild,M.F. and
            Harris,D.L.
  TITLE     Direct Submission
  JOURNAL   Submitted (12-AUG-2003) Animal Science, Iowa State University,
            2255J Kildee Hall, Ames, IA 50011, USA
FEATURES             Location/Qualifiers
```

```
     source          1..962
                     /organism="Litopenaeus vannamei"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:6689"
     mRNA            join(<1..32,239..328,515..672,794..>962)
                     /product="alpha-amylase"
     CDS             join(<1..32,239..328,515..672,794..>962)
                     /codon_start=3
                     /product="alpha-amylase"
                     /protein_id="AAR19061.1"
                     /db_xref="GI:38373488"
                     /translation="QWDPNSSNGQVIVHLFEWKWSDIAAECENFLGPRGFAGVQVSPP
                     NEYVEVYQGDVKRPWWERYQPVSYKLVTRSGDENAFKDMVTRCNNVGVRIYVDAVINH
                     MSGGWPMGTGASGGSSFDSGAESYPGVPYSAFDFNDGNCHTGSGNIE"
     exon            <1..32
                     /number=1
     exon            239..328
                     /number=2
     exon            515..672
                     /number=3
     exon            794..962
                     /number=4
ORIGIN
        1 cgcagtggga tcccaactct agcaatggac aggtgagtca cttccgatcc aggaagttat
       61 ttcttccagt cccacgtgag acatgggcta atcaaatgtt aaactgtagt ggatgttttc
      121 actgaattct tatacgtttc atgacagaac gagttcacta ttgttcgatt gaagtgtcgt
      181 actctacgtt gttcttagga gagaaaacct aaatacaacc aaactaaaac aattttaggt
      241 tatcgtccac ttgtttgagt ggaagtggtc ggacatcgcc gccgaatgcg agaacttctt
```

```
301 gggtcctcga ggattcgccg gcgttcaggt aaacactgac tcacgccata gtacttaaga
361 atagatatag catgaggctt acgtagataa taatataaaa ccaagactct gcagtaaata
421 taatcaagta gctacagaaa aatcaaaagg cttaacaaaa taactatata tttcaatcac
481 acacaaaaga gaaaaaccag ccctctctcg gcaggtatca ccgcctaacg aatacgtgga
541 ggtgtaccag ggagacgtga agcggccgtg gtgggagagg taccagcccg tctcctataa
601 actcgtcact cgctccggtg acgaaaatgc tttcaaagac atggtcacac gctgcaacaa
661 cgtgggagtc aggtgaggaa ggaactttag ggcatattct aattatgaag agcttcatct
721 acgtatataa tggcagattt atcgatccaa aatgaattca atcgtcacca aagtgataca
781 accacgcttc caggatctac gtcgacgctg tgataaacca catgtcaggg ggatggccga
841 tgggcacagg agcctccggg gggtcctcct tcgactcggg cgcggagtcc taccccgggg
901 ttccttactc cgctttcgac ttcaacgacg gcaactgcca caccgggtcc gggaacattg
961 aa
```

Cấu trúc dữ liệu chuỗi mRNA mã số* CX198004 *- dạng dữ liệu cấu trúc EST:

```
LOCUS       CX198004                 720 bp    mRNA    linear   EST 28-DEC-2004
DEFINITION  yde90c01.y1 Sea urchin EST Lib1 Strongylocentrotus purpuratus cDNA
            clone yde90c01 5' similar to TR:Q9U0F9 Q9U0F9 ALPHA-AMYLASE ;, mRNA
            sequence.
ACCESSION   CX198004
VERSION     CX198004.1  GI:56845428
KEYWORDS    EST.
SOURCE      Strongylocentrotus purpuratus
  ORGANISM  Strongylocentrotus purpuratus
            Eukaryota; Metazoa; Echinodermata; Eleutherozoa; Echinozoa;
            Echinoidea; Euechinoidea; Echinacea; Echinoida;
            Strongylocentrotidae; Strongylocentrotus.
REFERENCE   1  (bases 1 to 720)
  AUTHORS   Coffman,J.A., Robertson,A.J., Clifton,S., Pape,D., Hillier,L.,
            Martin,J., Wylie,T., Dante,M., Meyer,R., Theising,B., Bowers,Y.,
            Gibbons,M., Ronko,I., Tsagareishvili,R., Ritter,E., Kennedy,S. and
            Wilson,R.
  TITLE     WashU Sea Urchin EST Project
  JOURNAL   Unpublished (2004)
COMMENT     Contact: Dr. James A. Coffman
            WashU Sea urchin EST Project
            Washington University School of Medicine
            4444 Forest Park Parkway, Box 8501, St. Louis, MO 63108
            Tel: 314 286 1800
            Fax: 314 286 1810
            Email: est@watson.wustl.edu
            DNA sequencing by: Washington University Genome Sequencing Center
            Seq primer: -28RPpOT
```

```
            High quality sequence stop: 649.
FEATURES             Location/Qualifiers
     source          1..720
                     /organism="Strongylocentrotus purpuratus"
                     /mol_type="mRNA"
                     /db_xref="taxon:7668"
                     /clone="yde90c01"
                     /lab_host="DH10B"
                     /clone_lib="Sea urchin EST Lib1"
                     /note="Vector: pCMVSport6.1 (Invitrogen); Site_1: Not1;
                     Site_2: Sma1; Arrayed normalized library of full-length
                     cDNAs representing blastula stage transcriptome of the sea
                     urchin Strongylocentrotus purpuratus, cloned into the
                     vector pCMVSport6.1 (Invitrogen)"

ORIGIN   1 aactgtaaaa gagctggtgg ccaaataaaa gattacaaca atgcagatga agtcaggtac
        61 tgtaacctga atggcctgga cgatcttgat ttttcgaggg aagacacaac aaatcatgtg
       121 gtcaattacc tcaacaagct cttgagtttt ggtgttgcgg gcttcaggat cgacgctgcg
       181 aagcacatgc acccgttgga gattcaaaaa atcgtcaaca gtttgcttga ttgcactttt
       241 ggtgatcgtc catacatcta ccaggaagtc atagcggtga gtcaggatga acctatctcc
       301 tcatctgaat acacaggcaa cggtgacgtg actgagttca aattctgcct caagatggcc
       361 gagctggcca acggaaagac tgcactgaag tacttccaaa actttggaga accctggggt
       421 ttgcttggtt cctcacaagc tttggtcttc gttgacaacc atgataacca gagaggacac
       481 ggtggtgggg gcagccttgt gcacttcaac caaggagcag cttacaagaa agccacagca
       541 tttgctctag cctggaacta cggaactgcg aggatcatga gctcctacaa attcagtaac
       601 cctgatcagg gacctccatc tggtgactgg ggcaacacgt tgtctccgca attcaaccac
       661 gactcctcct gcatgggtga ctgggtctgc aaacatccct gggaaacgat tcccaacatg
```

10.2. Khai thác thông tin cơ sở dữ liệu chuỗi trong thiết kế và tách dòng gen

Thiết kế và tách dòng gen là nhiệm vụ trung tâm của kỹ thuật gen, một trong những ngành khoa học trẻ và phát triển năng động nhất của công nghệ sinh học hiện đại. Ngành công nghệ này cho phép nghiên cứu tách chọn lọc những gen có giá trị trong sinh giới làm nguyên liệu cơ sở để tái tổ hợp (hay điều chỉnh) hệ genom và hoạt động của chúng trong cơ thể chủ để tạo ra các biến chủng có giá trị cao hơn nhiều so với chủng ban đầu, phục vụ cho mục tiêu ứng dụng khác nhau. Thí dụ nhờ kỹ thuật này, người ta có thể chèn hay cắt đoạn nhằm làm thay đổi hay vô hiệu hoá hoạt lực gen (để điều chỉnh đường hướng trao đổi chất ở cấp độ phân tử trên chủng gốc theo chiều mong muốn) hay tách thu các gen có giá trị từ chủng gốc (chủng có năng lực sinh tổng hợp sản phẩm thấp, hay chủng đòi hỏi khắt khe về điều kiện lên men hoặc khó tách thu sản phẩm...) để đưa vào tái cấu trúc hệ genom trên các chủng công nghiệp (với ưu thế không đòi hỏi nghiêm ngặt về điều kiện lên men, hay chủng không tích tụ chất độc, hoặc dễ tách sản phẩm hơn...). Để có thể xây dựng và phát triển được ngành công nghệ này, tin-sinh học đã và đang đảm nhiệm vai trò đồng hành không thể thay thế, kể từ giai đoạn hình thành ý tưởng công nghệ ban đầu cho đến giai đoạn kiểm tra giám định chất lượng sản phẩm mới tạo ra.

Về nguyên lý, kỹ thuật thiết kế tách dòng gen có thể áp dụng trên các chủng đã biết cấu trúc di truyền (đã xác định được trong hệ genom có gen cần tách dòng), các chủng lạ mang hoạt tính gen cần tách hay tìm kiếm tách dòng gen từ các chủng mới (chủng chưa kiểm tra đánh giá hoạt tính biểu hiện của gen hay hoạt tính biểu hiện gen quá thấp không đủ để nhận biết tính trạng của chúng). Vai trò tin-sinh học trong các đường hướng này, ở chừng mực nhất định, có thể khác nhau; song nhìn chung trong cả ba đều hết sức quan trọng.

10.2.1. Tách dòng gen trên các loài đã biết cấu trúc di truyền

Việc xây dựng bản đồ hoàn chỉnh hệ genom của sinh giới được đánh giá là một trong các thành công to lớn của sinh học hiện đại (*Haemophilus influenza* 1995, *Escherichia coli* và *Bacillus subtilis* 1997...). Thông tin cấu trúc di truyền hoàn chỉnh của loài này có thể ứng dụng để nghiên cứu tách dòng gen. Để tách dòng gen quý từ các loài đã biết trước cấu trúc di truyền có thể áp dụng nhiều giải pháp kỹ thuật khác nhau, trong đó giải pháp khai thác thế mạnh điển hình của tin-sinh học có thể mô tả tóm tắt qua các bước sau:

- **Phân tích thông tin cấu trúc gen**: Bằng việc truy cập cơ sở dữ liệu để thu thông tin về hệ genom loài cần tách. Trên cơ sở dữ liệu genom này, sử dụng công cụ tìm kiếm để truy cập dữ liệu chuỗi gen cần tách. Thông tin về cấu trúc chuỗi gen này (tạm gọi là chuỗi khuôn) được sử dụng làm cơ sở để nghiên cứu thiết kế (hay lựa chọn) đoạn mồi tương ứng.

- **Nhân bản khuếch đại thu nhận gen**: Quá trình này bao gồm việc nuôi hoạt hoá chủng mang gen để thu sinh khối. Sau đó, áp dụng các giải pháp xử lý thích hợp để tách thu DNA của tế bào. Tiếp theo, sử dụng đoạn mồi đã thiết kế ở trên để tiến hành phản ứng nhân khuếch đại gen PCR và khâu cuối cùng của công đoạn này là điện di tách và tinh chế thu đoạn sản phẩm DNA đã khuếch đại.

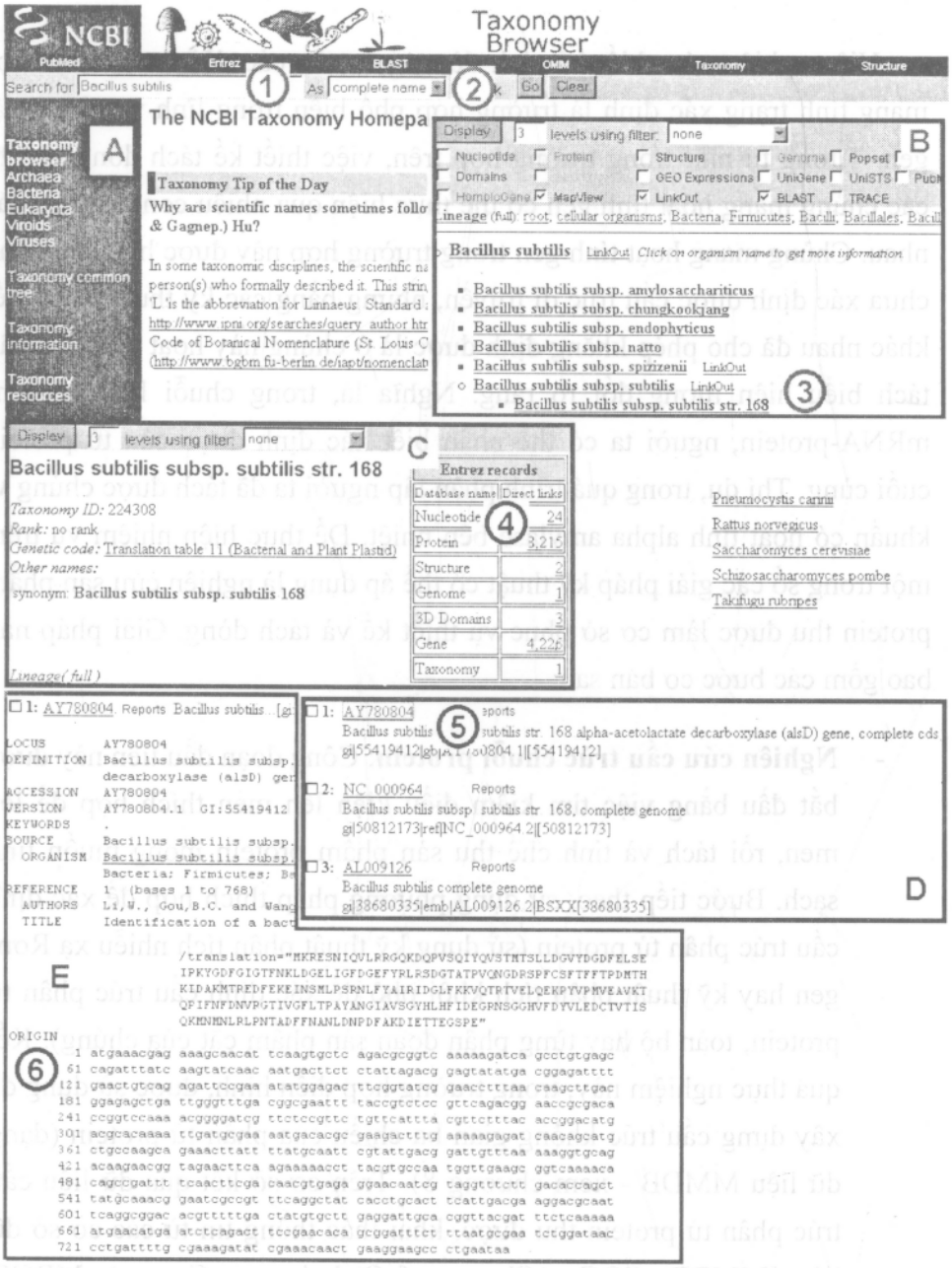
- **Xác định đặc tính gen tách dòng**: Việc xác định hoạt tính gen của đoạn DNA trên, trong trường hợp điển hình, được thực hiện qua việc gắn đoạn DNA vừa nhân khuếch đại vào các vector mang thích hợp để chèn chúng vào trong tế bào chủ lựa chọn. Tiếp theo, người ta tiến hành lên men rồi tách thu vector mang đoạn DNA trên. Bước cuối cùng người ta áp dụng các kỹ thuật thích hợp để cắt và giải trình

tự để tìm lại cấu trúc đoạn DNA chèn ban đầu [trong một vài trường hợp có thể tiến hành trực tiếp trên sản phẩm gen tinh sạch thu được, bằng cách thiết lập bản đồ gen hạn chế cho chuỗi (xem mục 10.1.1) và đối chứng là chuỗi khuôn]. Sản phẩm DNA thu được là một đoạn gen nếu thỏa mãn đồng thời hai yếu tố là vận động độc lập trong quá trình di truyền và mã hóa cho một tính trạng nhất định...

Như vậy, trong quá trình triển khai thực hiện tất cả các bước trên, đều không thể tách rời thế mạnh của tin sinh-học (khai thác cơ sở dữ liệu hay sử dụng công cụ xử lý dữ liệu). Để minh hoạ, giả sử người ta cần tách dòng gen *alpha-acetolactate decarboxylase* từ loài vi khuẩn *Bacillus subtilis*. Bản đồ hệ genom của loài vi khuẩn này đã được xác định hoàn toàn và có thể khai thác thông tin liên quan đến chúng trong các ngân hàng dữ liệu. Để thuận tiện cho việc minh hoạ, giả sử lựa chọn cơ sở dữ liệu NCBI cho mục đích trên. Hình 10.2 mô tả sơ đồ truy cập tìm kiếm dữ liệu chuỗi làm khuôn như sau: sau khi truy cập vào cơ sở dữ liệu NCBI (khung nền A), sử dụng đường dẫn siêu liên kết để truy cập sang cơ sở dữ liệu phân loại học "*Taxonomy Browser*". Tiếp theo, điền từ khoá "*Bacillus subtilis* as *complete name*" vào cửa sổ tìm kiếm (1 và 2 - A) rồi nhấn lệnh "Go" để gửi yêu cầu đi.

Sau đó, chương trình Entrez sẽ phản hồi lại thông tin với một phần giao diện có dạng như trong ô cửa sổ nhỏ B. Trong giao diện này, kích hoạt đường dẫn lựa chọn (giả sử loài gần gũi nhất để chọn là *Bacillus subtilis subsp. subtilis str. 168* -3) chương trình sẽ phản hồi lại dạng như trong cửa sổ nhỏ C. Kích hoạt đường dẫn về chuỗi nucleotide (4) để nhận lại thông tin dạng trong cửa sổ D. Trong giao diện này, chúng ta đã tìm thấy chuỗi mã hoá cho đặc tính enzyme cần tìm mang mã hiệu "AY780804". Truy cập vào đường dẫn siêu liên kết của chuỗi này (5-D) để thu thông tin chi tiết trong đó có cả cấu trúc chuỗi cần

tìm (tại 6-E). Cuối cùng, sử dụng dữ liệu cấu trúc chuỗi này để thiết kế đoạn mồi *primer* để phục vụ cho nhiệm vụ tách dòng gen tiếp theo.



Hình 10.2. Thao tác truy cập thu dữ liệu chuỗi trong NCBI

10.2.2. Thiết kế tách dòng gen từ chủng mang hoạt tính gen

Việc nghiên cứu thiết kế tách dòng gen trên các đối tượng sinh học mang tính trạng xác định là trường hợp phổ biến trong lĩnh vực kỹ thuật gen. Tương tự như trong trường hợp trên, việc thiết kế tách dòng gen từ các chủng mang hoạt tính gen có thể thực hiện qua nhiều con đường khác nhau. Chủng mang hoạt tính gen trong trường hợp này được hiểu là chủng chưa xác định được cấu trúc di truyền, nhưng bằng các kỹ thuật nhận biết khác nhau đã cho phép khẳng định được là ở chủng này hoạt tính gen cần tách biểu hiện tương đối rõ ràng. Nghĩa là, trong chuỗi liên hệ gen-mRNA-protein, người ta có thể nhận biết xác định được cấu tử protein cuối cùng. Thí dụ, trong quá trình phân lập người ta đã tách được chủng vi khuẩn có hoạt tính alpha amylase bền nhiệt. Để thực hiện nhiệm vụ trên, một trong số các giải pháp kỹ thuật có thể áp dụng là nghiên cứu sản phẩm protein thu được làm cơ sở phục vụ thiết kế và tách dòng. Giải pháp này bao gồm các bước cơ bản sau:

- **Nghiên cứu cấu trúc chuỗi protein**: Công đoạn đầu tiên này được bắt đầu bằng việc tìm kiếm điều kiện lên men thích hợp để lên men, rồi tách và tinh chế thu sản phẩm protein mong muốn tinh sạch. Bước tiếp theo, sử dụng phương pháp thích hợp để xác định cấu trúc phân tử protein (sử dụng kỹ thuật phân tích nhiễu xạ Rơn-gen hay kỹ thuật phân tích khối phổ để xác định cấu trúc phân tử protein, toàn bộ hay từng phân đoạn sản phẩm cắt của chúng). Kết quả thực nghiệm này, trong trường hợp điển hình, được sử dụng để xây dựng cấu trúc không gian ba chiều của phân tử protein (dạng dữ liệu MMDB - xem chương 8). Trên cơ sở kết quả dữ liệu cấu trúc phân tử protein thu được, khai thác thông tin từ các cơ sở dữ liệu [MMDB, COGs (*Cluster of Orthologous Groups*), MBGD (*Microbial Genome Database*), WIT (*What Is There?*)...] và sử

dụng các chương trình phân tích cấu trúc phù hợp (BLAST, FASTA hay thuật toán Smith Waterman...) để phân tích dự đoán xác định cấu trúc gen tương ứng (khai thác cơ sở dữ liệu để dự đoán xác định cấu trúc gen mã hoá - do tính thoái hoá của mã di truyền). Bước cuối cùng của công đoạn đầu tiên này là thiết kế đoạn mồi primer để nhân bản khuếch đại gen.

- Hai công đoạn tiếp theo **Nhân bản khuếch đại thu nhận gen** và **Xác định cấu trúc gen đã tách dòng**, về nguyên tắc cũng có thể tiến hành tương tự như đã mô tả trong mục 10.2.1 và sử dụng chính cấu trúc gen đã xác định được ở trên làm mẫu khuôn trong các khâu xử lý phân tích kết quả.

(Việc nghiên cứu gen dựa trên cơ sở phân tích protein, và xa hơn là việc thiết kế tạo ra các protein có đặc tính mong muốn là một trong các ứng dụng của Proteomic, hướng nghiên cứu đang được tập trung nỗ lực nghiên cứu hiện nay và để triển khai được chắc chắn không thể xem nhẹ vai trò tích cực, xuyên suốt và hiệu quả của tin-sinh học)

10.2.3. **Thiết kế tách dòng gen từ các chủng mới**

Việc thiết kế tách dòng gen từ chủng mới cũng có thể thực hiện được, dựa trên cơ sở phân tích cơ sở dữ liệu công nghệ sinh học. Kỹ thuật này thường được áp dụng nhằm nghiên cứu mở rộng khả năng khai thác tiềm năng chủng hiện có hay phục vụ cho việc nghiên cứu toàn diện đặc tính của các chủng mới. Các chủng mới trong trường hợp này thường là các chủng đã biết có các tính trạng nhất định (thường là các tính trạng quý và có lợi), song người ta lại chưa có trong tay thông tin di truyền về đối tượng gen cần tách hay hoạt tính biểu hiện gen quá thấp không đủ để nhận

biết tính trạng của chúng. Để thực hiện mục đích trên, một trong các giải pháp có thể áp dụng là khai thác cơ sở dữ liệu ESTs (RFLP, UniGene...). Kỹ thuật này được xây dựng trên cơ sở của nguyên lý tiến hoá và nguyên lý một gen-một protein. Nghĩa là nếu chủng nghiên cứu có mang gen cần tách dòng thì gen ấy sẽ có điểm tương đồng nhất định với gen tương ứng trên các loài gần gũi khác. Giải pháp kỹ thuật này có thể tóm tắt qua ba bước cơ bản là khai thác thông tin cơ sở dữ liệu để thiết kế mồi, nhân bản khuếch đại thu nhận gen cần tách dòng và xác định cấu trúc gen đã tách được, tương tự như hai trường hợp trên. Trong ba bước này, sự khác biệt lớn nhất tập trung ở khâu đầu tiên và được tóm tắt như sau.

- **Khai thác thông tin cơ sở dữ liệu**: Bước phân tích khai thác thông tin cơ sở dữ liệu này được bắt đầu bằng việc tìm kiếm thông tin về các gen mang tính trạng cần tách dòng trong ngân hàng dữ liệu ESTs (ngân hàng dữ liệu RFLP, SSLP...). Trên cơ sở kết quả thu được, tiến hành phân tích quy luật cấu trúc tương đồng (vùng ổn định về di truyền) và phân ly giữa chúng. Tiếp theo, lựa chọn trong các chuỗi kết quả trên lấy chuỗi có quan hệ tiến hóa gần gũi nhất với đối tượng cần tách dòng làm khuôn để thiết kế đoạn mồi phục vụ cho việc nhân bản tách dòng gen. Chuỗi khuôn lựa chọn thường là các chuỗi có nguồn gốc sinh học gần gũi hơn cả với chủng nghiên cứu và trong qua trình thiết kế đoạn mồi người ta quan tâm khai thác thông tin các vùng cấu trúc ổn định về di truyền nhiều hơn các vùng khác.

- Hai bước **Nhân bản khuếch đại thu nhận gen** và **xác định cấu trúc gen đã tách dòng** cũng được tiến hành tương tự như đã mô tả trong phần trên.

Những thao tác tóm tắt trình bày phần trên nhằm xác định vai trò quan trọng của tin-sinh học trong kỹ thuật tách dòng gen. Đương nhiên, trong thực tiễn có thể khai thác cơ sở dữ liệu công nghệ sinh học theo nhiều phương cách khác nhau. Song bao trùm lên tất cả là từ xây dựng, quy hoạch phương án thí nghiệm cho đến khâu kiểm tra phân tích chất lượng sản phẩm gen đã tách dòng người cán bộ sinh học đều phải lĩnh hội, sử dụng và khai thác hiệu quả tin-sinh học mới có thể thành công được.

Đồng thời, việc tách dòng gen thành công, trong thực tiễn mới đi được một phần con đường để tạo ra sản phẩm công nghệ. Vấn đề là ở chỗ làm thế nào để có thể ứng dụng được kết quả nghiên cứu thu được vào mục tiêu ứng dụng cụ thể. Nghĩa là phải biến nạp, hoạt hoá hay kiềm toả được hoạt động của gen ấy. Hay nói cách khác là phải xác định được cấu trúc và đặc tính sinh học của sản phẩm protein tạo ra cuối cùng. Để thực hiện mục tiêu trên đòi hỏi phải áp dụng nhiều giải pháp kỹ thuật và công nghệ khác nhau, song chắc chắn trong các giải pháp ấy bao giờ cũng có vai trò tích cực của tin-sinh học.

TÀI LIỆU THAM KHẢO

1. Baxevanis, A.D. and Francis Ouellette, B.F.
   Bioinformatics a Practical Guide to the Analysis of Genes and Protein
   John Wiley & Sons, Inc., Publication, 2001.

2. Bryan, B.
   Bioinformatics Computing
   Pretice Hall Pub., 2001.

3. David W. Mount
   Bioinformatics: Sequence and Genome Analysis
   Cold Spring Harbor Press, New York, 2001.

4. Dov Stekel
   Microarray Bioinformatics
   Cambridge Uni. Press., 2003.

5. Michael R. Barnes and Ian C. Gray
   Bioinformatics for Geneticists
   John Wiley & Sons Ltd. Publication, 2003.

6. Phipip E. B. and Helge W.
   Structural Bioinformatics
   Wiley-Liss Inc. 2003.

7. Stephen A.K. and David D.W.
   Introdution to Bioinformatics: A theorotical and prctical
   Human Press,2003.

8. Teresa, K. and Attwood
In troduction to Bioinformatics.Prentice Hall Pub., 1999.

9. 2can Bioinformatics Educational Resource Nucleotide Analysis at EBI Data Bank (Tài liệu đào tạo trực tuyến của cơ sở dữ liệu EBI: www.ebI.ac.uk/2can/home.html

10. Education: Teaching resources and on-line tutorials at Medline (www.ncbi.nlm.nih.gov/education/)

PGS. TS. NGUYỄN VĂN CÁCH

TIN – SINH HỌC

Chịu trách nhiệm xuất bản: TS. PHẠM VĂN DIỄN
Biên tập: NGUYỄN KIM ANH

NHÀ XUẤT BẢN KHOA HỌC VÀ KỸ THUẬT
70 Trần Hưng Đạo, Hà Nội

In 700 cuốn, khổ 16 x 24cm, tại Xưởng in NXB Văn hoá Dân tộc.
Số đăng ký KHXB: 209-2009/CXB/346-10/KHKT ngày 18/3/2009
Quyết định xuất bản số: 216/QĐXB-NXBKHKT ngày 8/7/2009
In xong và nộp lưu chiểu Quý III năm 2009.

PGS.TS. NGUYỄN VĂN CÁCH

Giá: 30.000đ