SỬ DỤNG CÔNG CỤ TIN SINH TRONG NGHIÊN CỨU METAGENEOMICS – HƯỚNG NGHIÊN CỨU VÀ ỨNG DỤNG MỚI TRONG SINH HỌC

NGUYỄN MINH GIANG[*], ĐỖ THỊ HUYỀN[**], TRƯƠNG NAM HẢI[***]

TÓM TẮT

Metagenomics là ngành khoa học nghiên cứu về đa hệ gene – nguyên liệu được thu hồi trực tiếp từ các mẫu môi trường. Kĩ thuật này cho phép khai thác tối đa các gene của hệ thống vi sinh vật không nuôi cấy được trong hệ sinh thái. Số liệu của metagenome chỉ có thể khai thác hiệu quả khi có sự hỗ trợ của các công cụ tin sinh học. Đây thực sự là bước đột phá trong nghiên cứu và ứng dụng của công nghệ sinh học.

Từ khóa: kĩ thuật nghiên cứu đa hệ gen, đa hệ gen, tin sinh học.

ABSTRACT

Using bioinformatic technology in studying metagenomics – A new research approach and application in biology

Metagenomics is the study of metagenome, the genetic material recovered directly from environment samples. The technique allows maximum exploitation of the enormous genes of uncultured microorganism in biota. Metagenome statistics can only be effectively exploited with the aid of bioinformatic technology, which is really a breakthrough in researching and applying biological technology.

Từ khóa: Metagenomics, metagenome, bioinformatics.

1. Tổng quan về metagenomics

1.1. Khái niệm

Thuật ngữ "metagenomics" lần đầu tiên được sử dụng bởi Jo Handelsman, Jon Clardy, Robert M. Goodman cùng các tác giả khác và được xuất bản vào năm 1998. Metagenomics là ngành khoa học nghiên cứu về đa hệ gen (metagenome) – nguyên liệu di truyền được thu hồi trực tiếp từ các mẫu môi trường. Metagenome còn được biết đến như là "hệ gen cộng đồng" (community genomics) hay "hệ gen môi trường" (enviromental genomics). Metagenomics là kĩ thuật cho phép khai thác được tối đa các gen của vi sinh vật không nuôi cấy được trong các quần thể sinh vật . Tùy vào từng loại mẫu môi trường số lượng vi sinh vật không nuôi cấy được dao động từ 99,0 đến 99,7%. Nếu tất cả các gen của vi sinh vật trong mẫu môi trường được tập hợp lại sẽ là nguồn nguyên liệu vô cùng phong phú cho việc khai thác gen, cũng như tìm hiểu cơ chế tác động giữa các vi sinh vật đảm bảo sự ổn định, phát triển chung của hệ sinh thái.

[*] NCS, Trường Đại học Sư phạm TPHCM

[**] TS, Phòng Kĩ thuật Di truyền, Viện Công nghệ Sinh học – Viện Hàn lâm Khoa học quốc gia

[***] GS TS, Phòng Kĩ thuật Di truyền, Viện Công nghệ Sinh học – Viện Hàn lâm Khoa học quốc gia

1.2. Cách tiếp cận trong nghiên cứu metagenomics [1]

Metagenomics nghiên cứu metagenome quần xã sinh vật thông qua ba bước gồm: 1) tách chiết nucleic acid trong mẫu thu thập; 2) thiết lập thư viện metagenome hoặc giải trình tự DNA metagenome; 3) sàng lọc gen dựa vào ngân hàng gene hoặc phân lập gen dựa vào số liệu giải trình tự gene. Việc phân lập gen từ metagenome được thực hiện tương tự như các nghiên cứu phân lập gen trong một hệ gen (genome). [9]

Hiện nay sau khi tách chiết nucleic acid người ta ít tiến hành lập thư viện gen mà tiến hành giải trình tự. Sau đó dựa trên số liệu giải trình tự kết hợp với các công cụ tin sinh để tìm kiếm, khai thác gen hay vùng gen mã hóa cho các protein quan tâm trước khi đưa vào thực nghiệm.

1.3. Một số mục tiêu cụ thể của metagenomics

Mục đích của metagenomics là để tìm hiểu thành phần và hoạt động của tập đoàn vi sinh vật phức tạp trong các mẫu môi trường thông qua phân tích trình tự ADN của chúng [4]. Mặt khác, khi có số liệu về đa hệ gen, chúng ta có thể thực hiện hàng loạt dự án phân lập gen tùy theo mục đích nghiên cứu. Ví dụ người ta không chỉ phân lập được gen phân hủy sinh khối thực vật từ metagenome của hệ vi sinh vật trong các mẫu ủ phân hữu cơ mà còn có thể phân lập được cả những gen tham gia vào chuyển hóa các hợp chất béo, protein, vitamin… cũng từ chính hệ vi sinh vật này.

Kĩ thuật metagenomics tạo ra dữ liệu khổng lồ về DNA dẫn đến việc phân tích bằng các thao tác thủ công khó mang lại hiệu quả cao. Do đó, hàng loạt các công cụ tin sinh học ra đời giúp nhà nghiên cứu tiết kiệm được thời gian và mang lại hiệu quả cao khi xử lí số liệu metagenome. Tin sinh học khi xử lí dữ liệu metagenome bước đầu tập trung vào ba nhiệm vụ cơ bản là phân tích phân loại, phân tích chức năng và phân tích so sánh.

Một số mục tiêu của metagenomics là: Xác định tính đa dạng phân loài sử dụng 16S rRNA, các mẫu gene đa dạng và cây phân loài của vi sinh vật [7, 9]. Số liệu đó được sử dụng để theo dõi và dự đoán các biến đổi môi trường; xác định gen hay operon mã hóa cho các enzyme cần thiết, có đặc tính mới (như cellulases, chitinases, lipases, thuốc kháng sinh, các sản phẩm tự nhiên khác…). Những enzyme này có thể được ứng dụng trong công nghiệp hoặc dược phẩm [6, 8]; xác định biến thể hoặc đa dạng trong gen cho các enzyme quan trọng và thiết kế tối ưu các điều kiện xúc tác của enzyme; xác định các cơ chế điều hòa và truyền tín hiệu của các gen quan tâm; xác định vi khuẩn hoặc các trình tự plasmid, đánh giá ảnh hưởng của chúng đến cấu trúc và sự đa dạng của các cộng đồng vi sinh vật [5]. Xác định các sự kiện chuyển gen tiềm năng [3] hay các gene/operons cho việc thu nhận dinh dưỡng, trung tâm trao đổi chất trung gian… Từ đó, cung cấp những hiểu biết về tương tác giữa các sinh vật trong chuỗi và lưới thức ăn, hoặc khám phá nền tảng thành công của vi sinh vật trong môi trường của chúng; xác định con đường trao đổi chất để có thể thiết kế môi trường nuôi cấy tăng trưởng cho các loài vi sinh vật chưa thể nuôi cấy được (Aylward FO & CS, 2012).

1.4. Một số thành tựu của Metagenomics

1.4.1. Trên thế giới

Thành công của metagomics phụ thuộc rất lớn vào các phần mềm tin sinh học và nguồn dữ liệu thu thập được. Trong khu vực châu Á, các nước như Trung Quốc, Hàn Quốc, Nhật Bản đã có những đột phá trong lĩnh vực huy động nguồn nhân lực hoạt động trong các lĩnh vực như sinh học, toán học, vật lí, hóa học, tin học… để tham gia nghiên cứu các dự án lớn của tin sinh học. Nhật Bản đã công bố ngân hàng dữ liệu DNA khổng lồ DDBJ (DNA Data Bank of Japan: tại http://www.ddgj.nig.ac.jp). Ở các nước châu Âu và Mĩ đã cho ra đời ngân hàng dữ liệu nổi tiếng như: NCBI - Trung tâm Quốc gia về Thông tin Công nghệ Sinh học (National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov) của Mĩ; EMBL-Phòng Thí nghiệm Sinh học phân tử European Molecular Biology Laboratory: http://www.embl.org) của châu Âu hoặc một phần của nó là EBI - Viện Sinh tin học châu Âu đặt ở Anh (European Bioinformatics Institute: http://www.ebi.ac.uk/); Đồng thời với sự ra đời của các ngân hàng dữ liệu thì hàng loạt phần mềm giúp xử lí các trình tự sinh học DNA và protein cũng ra đời như: Align (so sánh từng cặp trình tự DNA hoặc protein); CENSOR (sàng lọc các trình tự lặp và các đoạn DNA tương đồng); ClustalW2, Kalign, T-coffee, MAFFT, MUSCLE (so sánh đồng thời nhiều đoạn trình tự DNA hoặc protein); BLAST (tìm trên cơ sở dữ liệu ngân hàng gen trình tự DNA/protein tương đồng với trình tự cần phân tích); CpG Plot/CpGreport (dò tìm đảo CpG); Dna Block Aligner Form (phân tích promoter); GeneWise (so sánh protein với DNA); PromoterWise (so sánh hai trình tự DNA (thường là promoter) có tính đến trường hợp đảo đoạn hay chuyển đoạn); Transeq, ChromasPro (dịch mã trình tự DNA sang protein); WebPRANK (so sánh nhiều trình tự DNA cùng với nghiên cứu mất đoạn, thêm đoạn để tìm thông tin về tiến hóa và phát sinh loài). Hầu hết các phần mềm tin sinh học được cung cấp miễn phí trên những trang web xuất xứ từ Bắc Mĩ và châu Âu. [1, 9]

Sự kết hợp giữa các công cụ tin sinh, các ngân hàng dữ liệu giúp metagenomics thành công trên thế giới hơn 20 năm qua và được ứng dụng trong rất nhiều lĩnh vực: khoa học Trái Đất, khoa học sự sống, khoa học y sinh, năng lượng, xử lí môi trường, công nghệ sinh học, nông nghiệp và bảo vệ sinh học… [4, 9]. Trong một số năm trở lại đây với khả năng giải trình tự ngày càng nhanh chóng và chi phí giảm dần thì kĩ thuật metagenomics đang làm bùng nổ cuộc cách mạng về số liệu di truyền dựa trên việc phân tích trình tự bộ gen. Dữ liệu và siêu dữ liệu metagenomics không chỉ dừng lại ở việc mô tả sự phát sinh loài hay một số đặc điểm của gen thông qua hệ thống di truyền 16S. Dựa trên số liệu về metagenome của cộng đồng vi sinh vật toàn bộ chức năng của gen, mối quan hệ giữa các gen trong một nhóm sinh vật và giữa các nhóm sinh vật đều được làm sáng tỏ một cách rõ ràng . Các thí nghiệm này tập trung vào việc xác định vai trò của các gen và các vi sinh vật trong việc thành lập cộng đồng vi sinh vật động [9]. Mặt khác, dữ liệu này còn được ứng dụng trong thực tế để nâng cao kiến thức trong nhiều lĩnh vực và giải quyết những thách thức trong y học, kĩ thuật, nông nghiệp, phát

triển bền vững và hệ sinh thái… [2, 12]. Ví dụ các nhà khoa học tại Bộ Nông nghiệp Mĩ đã sử dụng công cụ metagenomics để xác định nguyên nhân dẫn đến giảm trọng lượng, gây tử vong ở gà là do chúng nhiễm virus dẫn đến các hội chứng hoại tử đường ruột, suy nhược và còi cọc. Bên cạnh những virus thường gặp ở gia cầm như astrovirus, reovirus và rotavirus và virus RNA thuộc nhóm Picornaviridae, họ đã phát hiện ra những virus hoàn toàn mới mà trước đây chưa được biết đến như: Picobirnavirus - một loại virus liên quan đến bệnh đường ruột ở vật nuôi khác; calicivirus - một loại virus có liên quan tới các bệnh đường ruột của con người [11]. Bằng cách sử dụng kĩ thuật metagenomics, Laszlo Zsak - người chủ trì nghiên cứu tại đơn vị Nghiên cứu bệnh do virus đặc thù ở gia cầm tại Phòng Thí nghiệm nghiên cứu gia cầm khu vực Đông Nam (Athens), đã phát hiện ra một loại virus mới có khả năng ứng dụng trong sản xuất một loại kháng sinh trong tương lai. Zsak và nhà vi sinh vật học Michael Day đã tìm thấy một chuỗi ngắn DNA của các virus mới được phát hiện và đã xây dựng một kĩ thuật để lập trình tự toàn bộ hệ gen của nó. Virus này được gọi là "phiCA82" - là loại virus giết chết vi khuẩn một cách tự nhiên và nằm trong một nhóm "tiểu thực bào" hoặc thể thực khuẩn. Đây là một giải pháp mới thay thế việc sử dụng thuốc kháng sinh, đồng thời cũng là công cụ để chống lại tác nhân đa kháng thuốc.

Trong khoa học sự sống: Số liệu về metagenome cung cấp những hiểu biết về lịch sử tiến hóa cũng như các khả năng của cộng đồng vi sinh vật chuyên sống trong môi trường. Các câu hỏi "vi sinh vật nào ở đó?", "vi sinh vật đang làm gì?" và "vi sinh vật hoạt động như thế nào?" đều có thể được giải đáp [2, 9]. Trong 20 năm qua, các nhà khoa học đã nhiều lần khoan sâu dưới lớp nền, trầm tích dưới đáy đại dương và họ đã khám phá một thế giới vi sinh vật vô cùng phong phú. Những quần xã đa dạng của những loài tế bào nhân sơ (prokaryote) được phát hiện ở tận sâu hơn 1km dưới nền đất đá của đáy biển. Phần lớn những vi sinh vật này không thể nuôi cấy được hoặc có rất ít quan hệ với những thế giới sinh vật bên trên bề mặt. Người ta chỉ biết được sự có mặt của chúng thông qua những trình tự DNA đặc trưng bằng cách sử dụng kĩ thuật metagenomics. Wei Xie và CS (2014) cũng đã tìm hiểu được nguồn năng lượng nào đã duy trì cuộc sống ở những hệ sinh thái bị trôn vùi như vậy. Việc giải trình tự metagenome có ý nghĩa rất lớn trong nghiên cứu quần xã virus, do virus không có marker để phân loại (như 16S RNA đối với vi khuẩn và vi khuẩn cổ, 18S RNA cho các sinh vật nhân chuẩn) nên cách duy nhất để nghiên cứu đa dạng di truyền và tiến hóa của virus là thông qua metagenomics. [7, 9]

Trong y học: Cộng đồng vi khuẩn đóng một vai trò quan trọng trong việc bảo vệ sức khỏe con người. Tuy nhiên, thành phần và cơ chế hoạt động của chúng vẫn còn rất nhiều bí ẩn. Dự án của "Human Microbiom" bước đầu đã sử dụng trình tự metagenome của các cộng đồng vi khuẩn ở 15-18 vị trí khác nhau trên cơ thể của ít nhất 250 người để đánh giá sự thay đổi và mối quan hệ của chúng với sức khỏe của con người. Một nghiên cứu y tế khác của dự án MetaHit (Metagenomics of the Human Intestinal Tract)

tiến hành ở 124 cá nhân từ Đan Mạch và Tây Ban Nha mắc bệnh đường ruột, thừa cân và cáu kỉnh đã công bố thông tin về sự đa dạng phát sinh loài vi khuẩn tiêu hóa. Nghiên cứu đã chứng minh rằng hai ngành vi khuẩn Bacteroidetes và Firmicutes chiếm hơn 90% của các loài được biết đến đang thống trị vi khuẩn đường ruột. Sử dụng tần số gen liên quan được tìm thấy trong ruột đã xác định 1244 cụm gen của metagenome là cực kì quan trọng cho sức khỏe của đường ruột. Bệnh nhân bị hội chứng ruột kích thích chỉ có 75% các gen trên và tính đa dạng vi khuẩn thấp hơn so với cá nhân không bị hội chứng ruột kích thích. Nghiên cứu cũng đã chỉ ra sự thay đổi đa dạng của quần xã vi sinh vật của bệnh nhân có thể liên quan với bệnh đường ruột hoặc béo phì. Trên cơ sở các nghiên cứu về metagenome của hệ vi sinh vật hoạt động ở cơ thể người để phát triển các công cụ và công nghệ sinh học mới hỗ trợ các mục tiêu của y học. Một số nghiên cứu khác về metagenomics cho phép phát hiện ra virus - nguyên nhân gây ra một số bệnh ung thư ở người [Erika Cosset &cs, 2013].

Trong sản xuất nhiên liệu sinh học: Ở quy mô công nghiệp sản xuất nhiên liệu sinh học đòi hỏi các enzym mới có năng suất cao hơn và chi phí thấp hơn. Phương pháp tiếp cận metagenomics phân tích cộng đồng vi sinh vật tự nhiên phức tạp, cho phép sàng lọc các enzym có hiệu quả để đưa vào ứng dụng trong sản xuất nhiên liệu theo hình thức công nghiệp. Trong thực tế rất nhiều các kết quả đã công bố về phân tích và so sánh metagenome giữa các hệ thống vi sinh vật trong hệ thống lên men khí sinh học, trong đường tiêu hóa của các động vật ăn cỏ như côn trùng, nấm, thú ăn cỏ. Thế giới đã công bố khoảng 75 hệ gen có sẵn của các loại vi sinh vật giữ vai trò nhất định trong quá trình sản xuất năng lượng sinh học. Trong đó có 21 bộ gen của vi khuẩn cổ sản xuất methan, 24 bộ gen của vi khuẩn sản xuất hyđro hoặc điện năng và 30 bộ gen của cyanobacteria vốn là sinh vật sản xuất diesel sinh học tiềm năng. Ít nhất một nửa bộ gen vi khuẩn hoàn thiện có liên quan đến năng lượng sinh học tạo ra trong 2 năm qua, trên 80 bộ gen liên quan đến năng lượng sinh học hiện đang được thiết lập trình tự [11]. Quỹ thông tin về hệ gen càng ngày càng phát triển, sẽ cung cấp nhiều mục tiêu phân tử hỗ trợ nghiên cứu tiền di truyền và hậu di truyền, mang lại thông tin thiết yếu về các loại vi sinh vật có mặt trong cộng đồng, cũng như các phản ứng trao đổi chất mà chúng thực hiện. Hệ gen cùng với ngành khoa học sắp xếp trình tự ADN và nghiên cứu protein, sẽ làm tăng hiểu biết của chúng ta về các vi sinh vật sản xuất năng lượng sinh học.

Trong xử lí môi trường: Các số liệu về metagenome của cộng đồng vi sinh vật khi sử dụng kĩ thuật metagenomics có thể cải thiện các chiến lược để theo dõi tác động của các chất gây ô nhiễm hệ sinh thái và làm sạch môi trường bị ô nhiễm [4]. Tăng hiểu biết về cách mà các cộng đồng vi sinh vật đối phó với ô nhiễm, cải thiện đánh giá về tiềm năng phục hồi của các hệ thống bị nhiễm bẩn và làm tăng khả năng thử nghiệm và ứng dụng các kích thích hoặc ức chế sinh học.

Trong công nghệ sinh học: Cộng đồng vi khuẩn sản xuất một loạt các hóa chất có hoạt tính sinh học được sử dụng trong cạnh tranh và truyền thông. Ngày nay rất nhiều loại thuốc sử dụng ban đầu được phát hiện ở vi khuẩn. Thành tựu trong khai thác tài nguyên di truyền phong phú của vi khuẩn không thể nuôi cấy đã phát hiện ra gen, enzyme và các sản phẩm tự nhiên mới. Việc áp dụng metagenomics đã cho phép phát triển các sản phẩm và hóa chất nguyên chất, hóa chất nông nghiệp và dược phẩm.

Trong nông nghiệp: Các cộng đồng vi sinh vật sống trong đất rất phức tạp, cao gấp 10 lần so với các vùng biển mà khoa học vẫn chưa khám phá hết. Sự hiểu biết về cấu trúc, sự đa dạng, chức năng và sự ổn định của cộng đồng vi sinh vật là điều cần thiết khám phá sự tiến hóa, hình thành và phát triển bền vững của sự sống trên Trái Đất [5]. Tuy nhiên, việc thu thập thông tin này rất khó khăn, do 99% các vi sinh vật đó hiện đang không nuôi cấy được dưới điều kiện phòng thí nghiệm. Trong thực tế nhiều dự án phân tích các mẫu đất khác nhau đã thành công nhờ sử dụng metagenomics. Người ta đang thực hiện các dự án khám phá về bản chất các mối quan hệ giữa các yếu tố vật lí, hóa học và sinh học của các loại đất trên toàn cầu.

1.4.2. Ở Việt Nam

Việt Nam đã có một số nghiên cứu trong lĩnh vực phân tích gen, xác định trình tự DNA của một số loài quan trọng để đánh giá về mặt di truyền, biến dị, xác định hệ số di truyền tìm ra các họ hàng thân thích, đánh giá mức độ biến đổi tính di truyền, nghiên cứu về đa dạng sinh học, xây dựng ngân hàng gen (gen bank)... ở một số viện nghiên cứu, trường đại học lớn như Khoa Công nghệ Sinh học, Trường Đại học Khoa học Tự nhiên TP Hồ Chí Minh; Viện Công nghệ Sinh học, Viện Khoa học và Công nghệ Việt Nam; Trường Đại học Y Dược TP Hồ Chí Minh. Phân viện Công nghệ thông tin tại TP Hồ Chí Minh, trong những năm qua đã hợp tác với một số nhà nghiên cứu của Viện Công nghệ Sinh học; của NCBI/NLM/NIH và NIAID/NIH, xây dựng hướng nghiên cứu với hai mục tiêu chính: Xây dựng Website về ngân hàng dữ liệu cung cấp thông tin di truyền phục vụ công tác huấn luyện và nghiên cứu công nghệ sinh học và xây dựng phần mềm để xử lí và phân tích các trình tự sinh học, bước đầu tạo ra sản phẩm phần mềm mang thương hiệu Việt Nam trong lĩnh vực tin sinh học. [11]

Đáng chú ý nhất là sản phẩm phần mềm tin sinh học do Trần Văn Lăng (Phân viện Công nghệ thông tin tại TP Hồ Chí Minh) chủ trì đã tạo ra sản phẩm phần mềm HiBio riêng với một số tính năng cần thiết cho việc tìm hiểu về sinh học phân tử. Bên cạnh đó các phần mềm nguồn mở như ClustalX, RasTop, Blastn cũng được tích hợp vào hệ thống hoạt động. Ngoài ra, nhóm đã xây dựng trang Website IOIT-HCMC Bioinformatics tại địa chỉ: http://www.ioit-hcm.ac.vn/index.htm. Trang website này bao gồm các phần mềm do nhóm thực hiện xây dựng và những phần mềm khác do nhóm thu thập được trên Internet nhằm phổ biến kiến thức về sinh học phân tử.

Chúng ta có được những lợi thế về những nguồn thông tin to lớn, hữu ích, nhưng việc sử dụng vẫn chưa đủ để phát triển một ngành tin sinh học mạnh cho Việt Nam.

Nguồn dữ liệu miễn phí thường cho số liệu rất hạn chế, do đó cần tạo ra những ngân hàng dữ liệu đặc trưng cho riêng nước ta. Các ngân hàng đó có thể khai thác từ công nghệ sinh học sẵn có trong nước về nhiều lĩnh vực khác nhau như nông nghiệp, chăn nuôi, hải sản, phòng chống bệnh, vacxin, kit chẩn đoán, y dược phẩm... Việc đào tạo một đội ngũ chuyên gia về tin sinh học chính là điều quyết định cho sự thành công của sự phát triển tin sinh học. Đội ngũ này không những phải có trình độ tư duy toán học xuất sắc mà còn phải thông hiểu những vấn đề hiện nay của sinh học.

Với những hạn chế nhất định về sự phát triển của tin sinh học do đó đến thời điểm hiện nay chưa có nhiều công bố về phân tích metagenome của các cộng đồng vi sinh vật. Các nghiên cứu về metagenome ở Việt Nam chủ yếu sử dụng theo phương pháp lập ngân hàng gen để chọn lọc nên khả năng thành công thấp. Cơ sở đi đầu trong việc áp dụng kĩ thuật giải toàn bộ trình tự metagenome kết hợp với xử lí số liệu bằng công cụ tin sinh là Phòng Kĩ thuật Di truyền, Viện Công nghệ Sinh học, Viện Hàn lâm khoa học quốc gia. Tại đây đã có những công bố trong nước và quốc tế về metagenome của vi sinh vật cộng sinh trong ruột mối. [10]

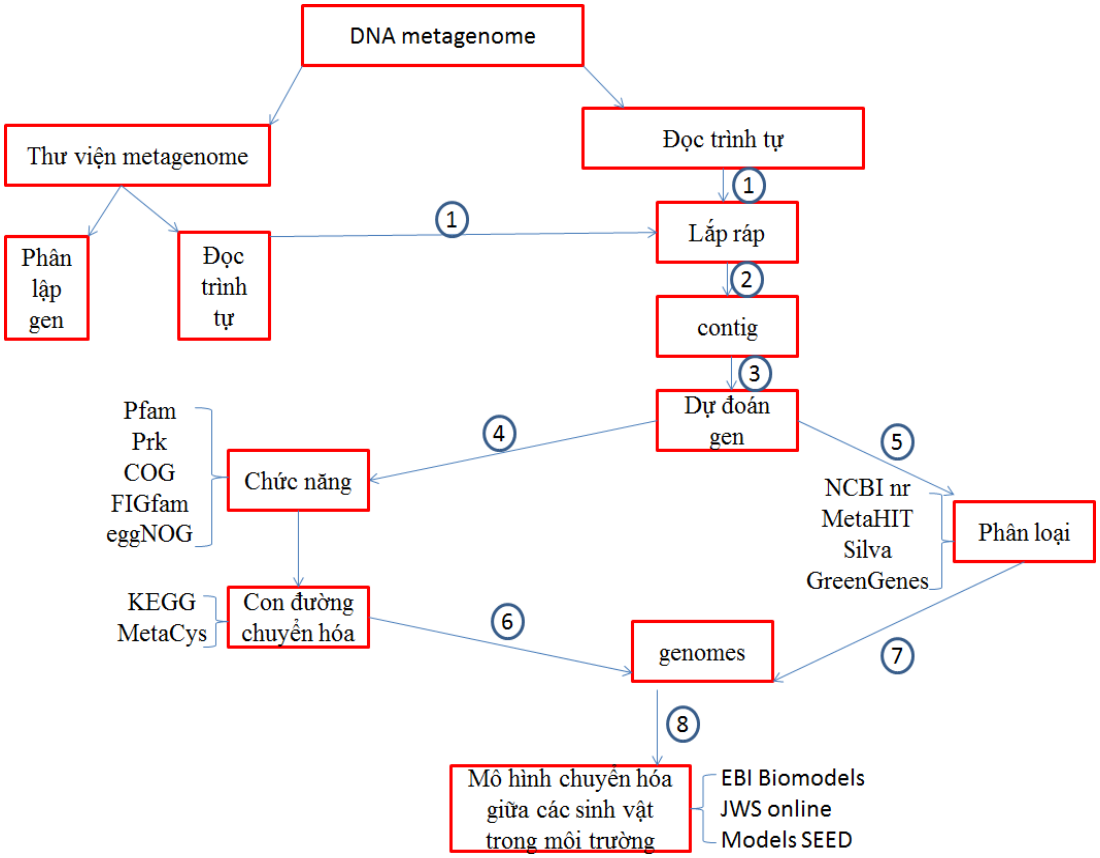
2. Khai thác metagenome

2.1. Phân lập gen dựa vào việc thiết lập thư viện metagenome

Tương tự như việc thiết lập thư viện genome để phân lập gen, toàn bộ DNA metagenome sẽ được phân cắt bằng enzym hạn chế thành các đoạn có kích thước nhất định, sao cho chúng chứa được trọn vẹn gen. Sau đó, các đoạn DNA này được gắn vào vector thích hợp và chuyển vào chủng vi sinh vật chủ. Với số lượng dòng đủ lớn, thư viện có thể chứa được toàn bộ các gen của metagenome. Các dòng biểu hiện protein ngoại lai sau đó sẽ được sàng lọc hoạt tính (ví dụ như sản xuất vitamin, tính kháng kháng sinh, enzyme...) trên môi trường có cơ chất đặc hiệu. Nhiều enzym, chất kháng sinh và các cơ chế đề kháng đã được phát hiện nhờ phương pháp này. Tuy nhiên, việc phân lập gen dựa trên việc sàng lọc thư viện metagenome trên môi trường có cơ chất thường tốn rất nhiều thời gian và công sức, do phải sàng lọc một số lượng quá lớn các dòng trong thư viện. Hơn nữa, cách tiếp cận này yêu cầu số lượng dòng thư viện phải rất lớn và chất lượng thư viện phải cao. Mặt khác một gen nguyên vẹn trong thư viện có thể biểu hiện ra được hoạt tính (được phát hiện) hay không cũng lại phụ thuộc rất nhiều vào sự tương thích và vị trí của nó với promoter của vector dùng để tạo thư viện. Để xác định chính xác trình tự DNA sau khi đã sang lọc hoạt tính có thể sử dụng thêm phương pháp đọc trình tự.

2.2. Sử dụng công cụ tin sinh khai thác metagenome

Việc chuẩn bị mẫu metageneome để đọc trình tự rất quan trọng, nếu mẫu không đủ sạch sẽ gây nhiễu khi đưa vào máy đọc tự động có thể gây ra sai lệch trong kết quả.



Quy trình chung xử lí số liệu metagenome

(Ghi chú: dấu ngoặc: các bộ dữ liệu; số 1- 8: là một số phần mềm tin sinh học)

1. Soap denovo, Soap aligner
2. MetaVelvet, Genovo, MetaORFA, MetaIDBA, IDBA-UD
3. MetaGeneAnnotator, FragGeneScan, Glimmer-MG, GeneMark
4. HMMer3, RPSBlast, BLAST, RAST, RAPsearch
5. MEGAN, CARMA, Sort-ITEMS, Sphinx
6. MinPath, Pathway Tools, KEGG mapper
7. RAST, MG-RAST, Model SEED
8. Pathway Tools, COBRA, Model SEED

Toàn bộ DNA tách chiết được từ mẫu môi trường đủ tiêu chuẩn sẽ được đưa vào máy đọc trình tự tự động. Sau khi đọc, máy sẽ xác lập được số lượng lớn các trình tự đọc ngắn (short – reads). Công việc tiếp theo là sắp dãy (assembly) các short - reads này để thu được bộ gien hoàn chỉnh. Tuy nhiên, trong quá trình xác lập trình tự ADN của các kĩ thuật có khả năng sinh lỗi cho từng nucleotide với tỉ lệ khoảng từ 1% đến 2% trên chiều dài của short - reads. Các nucleotide lỗi phải được sửa chữa để phục vụ

cho việc sắp dãy lại thành một bộ gien hoàn chỉnh. Ở bước này, một số phần mềm như SOAPdenovo được sử dụng để lắp ráp lại bộ gen từ các short - reads (hay "reads") thu được trong quá trình giải trình tự gen. Phần mềm này gồm có 6 module được dùng để 1) sửa chữa các lỗi đọc trình tự; sau đó 2) xây dựng đồ thị *de Bruijn* để 3) lắp ghép các contig, rồi 4) kiểm tra lại kết quả lắp ráp bằng cách so sánh các contig với các trình tự đọc được dùng để tạo ra nó; tiếp đến 5) tối ưu độ bao phủ và chiều dài các contig để 6) thu nhỏ các vùng gen không đọc được trình tự. Bằng công cụ như SOAPaligner các trình tự sau đó được đem so sánh lại (map) với các contig của chính nó để tìm ra bao nhiêu trình tự được sử dụng để tạo contig. PE (pair-end reads) là các trình tự mà cả hai đầu của nó đều tương đồng với contig và mối quan hệ hai đầu này là chính xác, cho độ tin cậy cao. Các trình tự mà chỉ có một đầu của nó tương đồng với contig hoặc mối quan hệ hai đầu không chính xác thì được gọi là SE (single-end reads).

Sau khi có các contig từ metageneome, các cặp mồi sẽ được thiết kế để phân lập các gen mong muốn. Phương pháp này đã được áp dụng để phân tích quần xã vi sinh vật trong rất nhiều môi trường như đại dương, đất, dải san hô, xác cá voi, suối nước nóng và các quần xã vi sinh vật liên kết với nhiều cơ thể sống khác nhau như người, mối, rệp, giun. [6]

Tùy theo mục đích nghiên cứu có thể lựa chọn các phần mềm phù hợp. Sau đó sử dụng các phần mềm dự đoán gen như: MetaGene Annotator (MGA), FragGeneScan, Glimmer-MG, GeneMark… được sử dụng để dự đoán tất cả các khung đọc mở (ORF – open reading frame) từ các contig. Dựa trên các ORF đã được xác định, sẽ tiếp tục dự đoán bằng cách so sánh ORF với hàng loạt các dữ liệu khác nhau như: Dữ liệu NCBI NR, MetaHIT, Silva, GreenGene để phân tích độ đa dạng loài; dữ liệu KEGG, MetaCys để phân loại gen vào các con đường chuyển hóa khác nhau; dữ liệu eggNOG, Pfam, Prk, COG, FIGfam để sắp xếp gen vào các nhóm chức năng.

Nếu nghiên cứu tập trung vào DNA và protein của metagenome thì công cụ BLASTall (Basic Local Alignment Search Tool: http://blast.ncbi.nlm.nih.gov/Blasti) được sử dụng rộng rãi nhất trong tin sinh học. BLAST sử dụng thuật toán tìm kiếm cục bộ heuristic và do đó có thể phát hiện ra mối liên hệ giữa các trình tự có những sự tương đồng riêng biệt. Có rất nhiều loại tìm kiếm khác nhau trên BLAST phục vụ cho những mục đích khác nhau: 1) BLASTp tìm kiếm tất cả các trình tự protein tương đồng với trình tự protein cần phân tích trong cơ sở dữ liệu protein; 2) BLASTn tìm kiếm tất cả các trình tự nucleotide tương đồng với trình tự DNA cần phân tích trong cơ sở dữ liệu DNA; 3) TBLASTn tìm trình tự protein tương đồng trong cơ sở dữ liệu DNA bằng cách dịch mỗi trình tự DNA ra tất cả 6 khung đọc mở; 4) BLASTx tìm trình tự nucleotide tương đồng trong cơ sở dữ liệu protein bằng cách dịch trình tự nucleotide cần phân tích sang tất cả 6 khung đọc mở. Sau khi có được các khung đọc mở cần quan tâm, sử dụng công cụ tìm kiếm trình tự amino acid tương đồng trong BLASTp; 5) Công cụ Blastpby được sử dụng trong so sánh các ORF với cơ sở dữ liệu NR để tiến hành phân loài.

Cấp độ phân loài của mỗi ORF được xác định bằng thuật toán dựa trên cơ sở LCA (Least Common Ancestors) được sử dụng trong phần mềm MEGAN (MEtaGenomic ANalyser). Thuật toán LCA sẽ xếp trình tự vào nhóm phân loại mà cấp độ phân loại của nhóm phân loại đó phản ánh được mức độ bảo thủ của trình tự gen. Căn cứ vào các ORF đã được đối chiếu với chức năng và các con đường chuyển hóa để lựa chọn các gen hay nhóm quan tâm. Trình tự các axit amin được dịch từ các ORF sẽ được sử dụng để dự đoán cụ thể về cấu trúc và các đặc tính của protein (trung tâm hoạt động, cơ chế xúc tác enzyme, khả năng chịu nhiệt, khả năng chịu kiềm…),… bằng phần mềm Phyre 2 (http://www.sbg.bio.ic.ac.uk/phyre2), Expasy (http://www.expasy.org)… Hoặc có thể xây dựng mô hình chuyển hóa các chất giữa các sinh vật trong môi trường bằng các công cụ Pathway Tools, COBRA, Model SEED…

3. Kết luận

Metagenomics tạo ra dữ liệu khổng lồ của metageneome đang mở ra rất nhiều hướng khai thác trong cả nghiên cứu cơ bản và nghiên cứu ứng dụng. Các bộ dữ liệu metageneome chỉ có thể được phân tích hiệu quả khi sử dụng các công cụ tin sinh học.

TÀI LIỆU THAM KHẢO

1. Carlotta De Filippo, Matteo Ramazzotti, Paolo Fontana and Duccio Cavalieri (2012), *Bioinformatic approaches for functional annotation and pathway inference in metagenomics data*. Briefings in bioiformatic, Vol 13. No 6. 696-710 doi:10.1093/bib/bbs070.
2. Edited by Diana Marco (2010), *Metagenomics: Theory, methods and applications*, Caister Academic press, Norfolk, UK. ISBN 978-1-904455-54-7.
3. Frans J. de Bruijn (2011), *Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats,* ISBN 978-0-47064719-6.
4. George I et al. (2010), *Application of Metagenomics to Bioremediation. Metagenomics: Theory, Methods and Applications*. Caister Academic Press, ISBN 978-1-904455-54-7.
5. Jones BV; Sun F; Marchesi JR. (2010), *Comparative metagenomic analysis of plasmid encoded functions in the human gut microbiome. BMC Genomics; 11: 46.*
6. Kennedy J1, O'Leary ND, Kiran GS, Morrissey JP, O'Gara F, Selvin J, Dobson AD (2011), *Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems*, J Appl Microbiol. 2011 Oct;111(4):787-99. doi: 10.1111/j.1365-2672.2011.05106
7. Shrikant Sharma1, Shashank Rana1, Raghvendar Singh (2012), *A SHORT NOTE-METAGENOMICS*. IJBR 3[04], pp.181-186.

(Xem tiếp trang 184)

SỬ DỤNG CÔNG CỤ TIN SINH…

(Tiếp theo trang 176)

8. The New Science of Metagenomics (2007), *Revealing the Secrets of Our Microbial Planet Committee on Metagenomics: Challenges and Functional Applications,* National Research Council, ISBN: 0-309-10677-X, 170 p, 6 x 9.
9. Thi Huyen Do,Thi Thao Nguyen, Thanh Ngoc Nguyen, Quynh Giang Le, Cuong Nguyen, Keitarou Kimura, and Nam Hai Truong (2014), *Mining biomass-degrading genes through Illumina-basedde novosequencing and metagenomic analysis of free-living bacteria in the gut of the lower termite Coptotermes gestroi harvested in Vietnam*, J Biosci Bioeng. 2014 Dec;118(6):665-71. doi: 10.1016/j.jbiosc.2014.05.010, Epub 2014 Jun 11.
10. Torsten, Thomas, Jack Gilbert and Folker Meyer (2012), *Metagenomics - a guide from sampling to data analysis. Microbial Informatics and Experimentation* 2012, **2**:3 doi:10.1186/2042-5783-2-3.
11. *http://vi.wikipedia.org/wiki/Metagenomics*
12. *http://tinsinhhoc.org/72-tong-quan-ve-tin-sinh-hoc*

(Ngày Tòa soạn nhận được bài: 26-12-2014; ngày phản biện đánh giá: 09-02-2015; ngày chấp nhận đăng: 12-02-2015)