GIẢI TRÌNH TỰ HỆ GEN - ĐỘT PHÁ MANG TÍNH CÁCH MẠNG TRONG NHẬN THỨC VÀ CẢI TẠO SỰ SỐNG

GS.TS ĐỖ NĂNG VỊNH
Viện Di truyền nông nghiệp

Thông qua hợp tác với các nhà khoa học Vương quốc Anh, Việt Nam đã thành công trong việc giải trình tự hệ gen của một loài thực vật bậc cao rất quan trọng là cây lúa, mở ra hướng mới về ứng dụng bioinformatics để khai thác trình tự hệ gen phục vụ công tác nghiên cứu và chọn tạo giống lúa. Tuy nhiên, vấn đề đặt ra hiện nay là làm cách nào để phân tích và xử lý dữ liệu về hệ gen nhằm phát hiện chính xác vị trí, trình tự của từng gen trên từng nhiễm sắc thể (NST), chức năng của gen, các gen tham gia điều khiển hoạt hóa và các điều kiện để gen biểu hiện thành các tính trạng quan trọng như năng suất, chất lượng, khả năng chống chịu sâu bệnh và các biến đổi khắc nghiệt của khí hậu.

Khám phá cấu trúc, chức năng của hệ gen và khả năng cải tạo sinh giới

Gen và cấu trúc của hệ gen quy định mọi đặc tính của cơ thể sống và tính đa dạng phong phú của hệ sinh thái. Muốn cải tạo sinh giới phải hiểu biết cấu trúc, chức năng và tác động trực tiếp vào gen và hệ gen. Mặc dù sự sống trên trái đất đã trải qua quá trình tiến hóa khoảng 3,5 tỷ năm. Tuy nhiên, chỉ từ cuối thế kỷ XIX đến nay, nhân loại mới bắt đầu có những phát minh về gen. Cuối thế kỷ XIX, đầu thế kỷ XX, các phát minh của Mendel và Morgan đã khẳng định: tính trạng của cơ thể sống do gen quy định. Các gen nằm trên NST theo một trật tự nhất định. Toàn bộ các gen trên NST trong nhân tế bào và các sợi ADN trong tế bào quan (như ty thể, lục lạp ở cây trồng...) tạo ra hệ gen. Nhưng mãi đến năm 1944, người ta mới xác định được gen chính là các phân tử axit deroxyribonucleic (ADN). Năm 1953, Watson và Crick đã phát hiện cấu trúc hình sợi của ADN gồm hai chuỗi xoắn kép tương đồng. Mỗi sợi ADN là một polymer được tạo bởi các trình tự sắp xếp khác nhau của 4 loại nucleotid hay là 4 monomers của sự sống. Có 4 loại deroxyribonucleotid: Adenin, Guanin, Cytosin và Thimin. Cứ 3 nucleotid lại tạo ra một tổ hợp chập ba (triplet) hay còn gọi là đơn vị mã di truyền (codon), 4 nucleotid với tổ hợp chập ba sẽ tạo ra $4^3 = 64$ codons khác nhau. Chiều dài của gen thường rất khác nhau. Nếu tính một gen trung bình được tạo bởi khoảng 900 nucleotid, tương đương với 300 codons, theo lý thuyết với 64 codons sẽ tạo ra 64^{300} phân tử ADN hay các gen khác nhau.

Với 7 nốt nhạc, nhân loại đã tạo ra hàng triệu bài ca. Với 32 chữ cái, các nhà văn đã viết ra hàng triệu cuốn sách. Với 64 codons, có thể sáng tạo ra hằng hà sa số các cấu trúc đa dạng vô cùng vô tận của phân tử ADN và các gen khác nhau. Điều đó giải thích tại sao thế giới sống của chúng ta lại phong phú, đa dạng và tuyệt vời đến như vậy. Hàng triệu loài sinh vật trên trái đất, trên 7 tỷ người hầu như không ai giống hệt ai. Tất cả tính đa dạng phong phú, tất cả những vẻ đẹp thiên thần của sự sống đều do gen và cấu trúc của gen quy định. Trật tự của các nucleotid quy định cấu trúc đặc thù và sự khác biệt về chức năng của các gen. Số lượng đa dạng của gen quy định tính đa dạng của sinh giới.

Chức năng gen hay vai trò của gen được thể hiện như thế nào trong tế bào sống?

Cấu trúc đặc thù của gen lại quy định cấu trúc đặc thù của axit ribonucleic (ARN) cấu trúc và ARN thông tin. Đến lượt nó, ARN thông tin lại quy định cấu trúc đặc thù của các protein. Trật tự của 64 codons trong ADN sẽ quy định trật tự sắp xếp của các axit amin (gồm trên 20 axit amin khác nhau) trong phân tử protein. Trình tự axit amin đặc thù lại quy định cấu trúc và tính đặc thù của các enzym, trong đó mỗi phản ứng sinh hoá hoặc mỗi bước

trong quá trình dẫn truyền điện tử trong hệ thống màng tế bào đều được quy định bởi một enzym hoặc một protein cấu trúc. Hàng nghìn, hàng vạn chuỗi phản ứng sinh hóa học trong mỗi tế bào đều do các enzym hoặc hệ thống enzym quy định. Xét cho cùng, gen và các sản phẩm trực tiếp của gen (ARN và các protein) quy định mọi cấu trúc và quá trình xảy ra trong cơ thể sống. Gen và hệ gen; enzym và hệ enzym; protein và các tổ chức tế bào, mô và cơ thể sống; các quá trình sinh hóa và trao đổi chất; mọi tính trạng của mỗi cơ thể sống; tất cả nằm trong mối quan hệ biện chứng phức tạp giữa chúng với nhau và giữa chúng với môi trường sống, trong đó cấu trúc hay trình tự của gen là chìa khoá của mọi quá trình vận động sinh học. Cho đến nay, khoa học mới chỉ khám phá được cấu trúc và chức năng của một phần rất nhỏ các gen trong hệ gen của cây trồng, vật nuôi và con người. Nhiều gen quan trọng ở con người quy định sức khỏe, các bệnh tật khác nhau, trí tuệ, trí nhớ, khả năng sáng tạo của con người vẫn còn nằm trong bí ẩn. Điều đó cho thấy, khoa học Genomics (Genome học), bao gồm giải trình tự hệ gen (structural genomics) và các dự án nghiên cứu chức năng hệ gen (functional genomics) có ý nghĩa rất quan trọng. Đó là những dự án mang tầm lịch sử nhân loại. Thành công của các dự án này chắc chắn sẽ giúp cho nhân loại, một trong khoảng hai triệu loài sinh vật trên trái đất, trở thành chủ thể của những khả năng sáng tạo vô tận. Con người có thể nhận thức và cải tạo được hầu hết các quá trình của sự sống, có thể thiết kế các cơ thể sống mới, tạo ra hàng loạt các giống cây trồng, vật nuôi, chủng vi sinh vật mới, các phương pháp và phương tiện trị liệu các bệnh hiểm nghèo như ung thư, tim mạch, lão hóa... Đồng thời, nhân loại sẽ nắm trong tay khả năng cải thiện sức khỏe, sự dẻo dai, trí nhớ, tuổi thọ, khả năng nhận thức và sáng tạo của chính bản thân con người. Điều đó đang giải thích tại sao, các siêu cường đang đua tranh và cộng tác với nhau, đầu tư vào các dự án genomics mang tầm thiên niên kỷ trong lịch sử văn minh của nhân loại.

Sơ lược lịch sử giải mã trình tự hệ gen

Những bước đi đầu tiên

Năm 1976 được đánh dấu bằng một kỳ tích, được gọi là "anh hùng" trong lịch sử nhân loại khi mà hệ gen đầu tiên được giải trình tự. Đó là một trực khuẩn có cấu trúc đơn giản, có tên là bacteriophage MS2. Hệ gen của nó được cấu trúc chỉ từ một sợi ARN với chiều dài 3.569 nucleotides và chỉ mang 4 gen. Walter Fiers và nhóm của ông là những người đã thực hiện kỳ tích "anh hùng" đó trong lịch sử khoa học về sự sống. Những thành tựu giải mã gen ở giai đoạn tiếp theo thuộc về nhà bác học người Anh nổi tiếng, 2 lần giành Giải Nobel, Frederick Sanger. Năm 1975, Frederick Sanger đã phát minh ra phương pháp giải trình tự ADN bằng enzym, được gọi là phương pháp Dideoxy. Rất nhiều máy giải trình tự gen dựa trên nguyên tắc của phương pháp này. Năm 1977, hệ gen thứ 2 được giải trình tự là bacteriophage Φ X174 với phương pháp của Sanger. Đây là một thực thể đầu tiên trên thế giới có hệ gen từ ADN được giải trình tự. Hệ gen của bacteriophage Φ X174 có cấu trúc phức tạp hơn, gồm 1 ADN đơn sợi, mang 11 gen khác nhau (5.386 bp). Năm 1982, Sanger và các cộng sự công bố trình tự bộ gen tương đối lớn đầu tiên được giải mã, đó là thực khuẩn thể lambda (Phage λ). Phage λ là một loại virus ký sinh vi khuẩn Escherichia coli và được xem là một mô hình nổi tiếng nhất của sinh học phân tử cổ điển. Phage λ có hệ gen gồm 2 sợi kép ADN dài 48.502 bp.

Dựa trên các kết quả giải trình tự gen, năm 2007, Craig Venter và các cộng sự đã tổng hợp được hệ gen hoàn chỉnh của một vi khuẩn được gọi là "sinh vật tổng hợp" đầu tiên, do con người tạo ra, có khả năng sản xuất năng lượng sinh học. Nhân loại bước sang giai đoạn tự thân sáng tạo các dạng thức mới của sự sống vì lợi ích con người như sản xuất năng lượng, dược chất, khử ô nhiễm môi trường, vật liệu mới…

Cuộc cách mạng giải mã hệ gen trên toàn thế giới

Việc giải trình tự bộ gen của các sinh vật tiền nhân và nhân thực, bao gồm cả hệ gen người và một số cây trồng quan trọng sẽ không bao giờ trở thành hiện thực nếu không có các phương tiện tin học hiện đại. Cuộc cách mạng giải mã hệ gen trên toàn thế giới là thành tựu vĩ đại của khoa học và công nghệ (KH&CN) ở cuối thế kỷ XX, đầu thế kỷ XXI. Cuộc cách mạng đó sẽ không thể xảy ra nếu không có các tiến bộ nhanh chóng của các ngành khoa học như vật lý học, toán học, công nghệ vi mạch và các phần mềm xử lý, các khoa học sinh học như hóa sinh, di truyền học phân tử…

Việc xác định sự tương đồng và khác biệt về gen giữa các sinh vật với hệ gen gồm hàng triệu nucleotid sẽ không thể thực hiện được nếu không có sự tiến bộ nhanh chóng của công nghệ vi mạch và các bộ vi xử lý. Trình tự bộ gen của một cơ thể sống đầu tiên (các virus và thực khuẩn chưa được xem là cơ thể sống), vi khuẩn Haemophilus influenzae, sẽ không thể được giải mã nếu thiếu những phương pháp tính toán được phát triển tại Viện Nghiên cứu genome (TIGR - The Institute for Genomic Research) do J. Craig Venter lãnh

đạo. Lý do là vì vi khuẩn tuy là cơ thể đơn bào giản đơn, nhưng đã có genome tương đối lớn. Phần mềm, phát triển bởi TIGR, được gọi là Assembler TIGR đã thực hiện nhiệm vụ ghép khoảng 24.000 đoạn ADN riêng biệt vào hệ gen. Hệ gen của vi khuẩn Haemophilus influenzae đã được giải mã hoàn tất vào năm 1995, gồm 1.830.140 cặp base với 1.740 gen. Venter đã thành công trong giải trình tự bộ gen Haemophilus influenzae trong 13 tháng với chi phí 50 cent cho mỗi bp, bằng một nửa chi phí và nhanh hơn so với phương pháp trước đó. Bằng phương pháp mới này, Viện TIGR hoàn thành giải trình tự gen của vô số sinh vật khác, trong đó có vi khuẩn Genitalium mycoplasma, vi khuẩn có liên quan với nhiễm trùng đường sinh sản và nổi tiếng vì có bộ gen ngắn nhất trong tất cả các sinh vật sống tự do được giải mã trong thời gian 8 tháng (từ tháng 1 đến tháng 8.1995). TIGR sau đó công bố trình tự bộ gen đầu tiên của các sinh vật cổ *Archaea:* Methanococcus jannaschii, Archaeoglobus fulgidus, trình tự bộ gen của vi khuẩn gây bệnh loét dạ dày tá tràng, Helicobacter pylori…

Dự án giải mã hệ gen người là một dự án khổng lồ, khám phá bí mật cấu trúc gen của một sinh vật tiến hóa nhất trong sinh giới, một thực thể sống trí tuệ độc nhất trên hành tinh và có thể là duy nhất trong thiên hà của chúng ta. Dự án đã được bắt đầu vào năm 1990 với mục tiêu xác định trình tự sắp xếp của khoảng trên 3,3 tỷ cặp base (bp) trong hệ gen của con người. Dự án giải mã toàn bộ hệ gen của người đầu tiên đã được hoàn thành chỉ sau 13 năm, từ năm 1990 đến 2003, với chi phí khoảng 3,8 tỷ USD.

Theo thống kê của KEGG (Bách khoa toàn thư Kyoto về gen và genome - Kyoto encyclopedia of genes and genomes), cho đến nay, đã có 192 loài sinh vật nhân thực (Eukaryotes), 1.444 loài vi khuẩn (bacteria) và 160 loài sinh vật cổ sinh (archaea) đã được giải trình tự bộ gen.

Giải mã hệ gen cây lúa Oryza sativa

Hệ gen của cây lúa Oryza sativa có 12 cặp NST, một phân tử ADN ty thể và một ADN lục lạp tròn. Hình 1 mô tả hình dạng và kích thước tương đối của 12 NST lúa và 2 tế bào quan (ty thể và lục lạp). Bảng 1 cho thấy kích thước (đơn vị Mbp = triệu cặp base) và dự đoán số lượng gen cho mỗi NST ở lúa thuộc loài Oryza sativa.

Hình 1: hình dạng và kích thước tương đối của các NST lúa và 2 tế bào quan

Tên NST	Kích thước (Mb), 1 Mb = 1.000.000 bp	Dự tính số lượng gen/NST
Chr1	45,1	4.467
Chr2	36,8	3.011
Chr3	37,3	3.197
Chr4	35,9	2.679
Chr5	30,0	2.426
Chr6	32,1	2.342
Chr7	30,4	2.507
Chr8	28,5	2.286
Chr9	23,8	1.618
Chr10	23,7	1.724
Chr11	30,8	1.834
Chr12	27,8	1.870
Ty thể Mitochondria	0,491	96
Lục lạp Chloroplast	0,135	159
Tổng số	Khoảng 383 Mb	30.216

Bảng 1: kích thước và dự tính số lượng gen ở các NST lúa (nguồn: http://www.patentlens.net/daisy/RiceGenome/2697/2950.html)

Dự án quốc tế giải trình tự toàn bộ hệ gen của giống lúa Niponbare kéo dài 7 năm có sự tham của 10 quốc gia, với chi phí hơn 100 triệu USD. Toàn bộ hệ gen giống lúa này dài khoảng 383 Mb (383 triệu bp).

Viện Nghiên cứu lúa quốc tế (IRRI) đã đưa ra kế hoạch giải trình tự bộ gen của 10.000 giống lúa trong 2 năm nhằm thực thi chiến lược hỗ trợ chọn tạo giống lúa nhanh hơn, hiệu quả hơn. Ngày 15.11.2011, IRRI đã ký với Viện Hàn lâm khoa học nông nghiệp Trung Quốc (CAAS) thỏa thuận giải mã hệ gen của 3.000 giống. Viện BGI (Viện Genome Bắc Kinh) sẽ giải trình tự toàn bộ các hệ gen và xử lý dữ liệu cơ bản, các nhà khoa học thuộc CAAS, IRRI, BGI và có thể một số viện nghiên cứu khác sẽ làm việc cùng nhau để phân tích dữ liệu. Đây sẽ là một dự án kỷ nguyên liên kết giữa di truyền học và tin học trong nhận thức bản chất di truyền tính trạng ở cây lúa.

BGI và CAAS đã xây dựng một cơ sở dữ liệu về genome cây lúa, gọi là cơ sở dữ liệu về hệ gen lúa (Rice Genome Knowledgebase - RGKbase). Hiện tại, RGKbase lưu trữ dữ liệu hệ gen của 5 giống lúa và các loài dại: Nipponbare (japonica), 93-11 (indica), PA64s (indica), lúa châu Phi (Oryza glaberrima) và một loài lúa hoang (Oryza brachyantha). Các dữ liệu mới sẽ được cập nhật thường xuyên.

Dự án genome lúa Việt Nam

Hội đồng nghiên cứu công nghệ sinh học và các khoa học sinh học Vương quốc Anh (BBSRC) đã ký biên bản ghi nhớ với Bộ KH&CN nước ta về giải trình tự bộ gen của 30 giống lúa Việt Nam. Mục

tiêu đặt ra là tạo cơ sở dữ liệu về trình tự gen đầy đủ của các giống lúa có các đặc tính sinh học nông nghiệp quan trọng như năng suất, chất lượng cao, chịu ngập, mặn, hạn hán và chống sâu bệnh, làm nền tảng cho chọn tạo giống lúa. Trung tâm John Innes (John Innes Centre) và Trung tâm Phân tích hệ gen (Genome Analysis Centre) thuộc BBSRC sẽ xây dựng cơ sở dữ liệu di truyền để giúp Viện Di truyền nông nghiệp và các viện nghiên cứu khác khai thác, sử dụng cho cải thiện giống lúa ở Việt Nam.

BBSRC đã tài trợ 250.000 bảng Anh (khoảng 409.433 USD) và Bộ KH&CN nước ta đã tài trợ khoảng 100.000 bảng Anh cho dự án này. Việc Chính phủ Anh tài trợ cho dự án là một cử chỉ vinh danh cho những thành tựu rực rỡ của nông nghiệp nước ta cũng như vai trò quan trọng của cây lúa Việt Nam đối với nước ta và an ninh lương thực thế giới. Ông David Willets, Bộ trưởng Phụ trách các trường đại học và khoa học Anh, cho biết "Với sự đầu tư tương đối khiêm tốn này, chúng ta đã có tiềm năng để tạo ra một sự khác biệt rất lớn - nó có thể dẫn đến sự ra đời của các giống lúa mới có khả năng đối phó với biến đổi khí hậu và đáp ứng sự tăng trưởng dân số trong tương lai". Bà Janet Allen, Giám đốc nghiên cứu BBSRC, nói thêm rằng dân số toàn cầu dự kiến đạt 9 tỷ người vào năm 2050, và nếu một nửa dân số thế giới ăn gạo như một lương thực chủ yếu, mức tăng năng suất lúa sẽ phải ở mức đủ để nuôi thêm khoảng 1 tỷ người vào năm 2050.

Hội thảo Anh - Việt "Kết quả nghiên cứu giải mã genome một số giống lúa bản địa của Việt Nam" được tổ chức ngày 28.8.2013 đã tổng kết thành tựu giải mã hệ gen của 36 giống lúa với các đặc tính di truyền quan trọng. Phía Anh hứa sẽ tiếp tục giúp Việt Nam giải trình tự của 800 giống lúa khác nữa. Với việc khai thác và phân tích các dữ liệu về 36 hệ gen này, chúng ta hy vọng sẽ phát hiện được các gen quy định các tính trạng quan trọng như năng suất, chất lượng, chịu hạn, chịu mặn, chống chịu bệnh đạo ôn, bạc lá...

Vấn đề khai thác dữ liệu về hệ gen lúa để xác định chức năng gen

Với các tiến bộ hiện nay của KH&CN, việc giải mã hệ gen đã trở nên vô cùng nhanh chóng và với chi phí ngày càng thấp. Người ta tính rằng, việc giải mã toàn bộ hệ gen của cây lúa hoặc con người có thể hoàn thành trong 1-2 ngày với chi phí chỉ khoảng 1.000 USD. Vấn đề lớn nhất đặt ra hiện nay không còn là giải trình tự nữa mà là làm cách nào để phân tích và xử lý dữ liệu về hệ gen nhằm phát hiện chính xác vị trí, trình tự của từng gen trên NST, chức năng của gen, các gen tham gia điều khiển hoạt hóa và các điều kiện để gen biểu hiện thành các tính trạng quan trọng như năng suất, chất lượng, khả năng chống chịu sâu bệnh và các biến đổi khắc nghiệt của khí hậu.

Do vậy, mục tiêu tiếp theo của giải trình tự hệ gen là phải xác định được chức năng của từng gen, từng nhóm gen, phát hiện các gen mới ở lúa. Đây sẽ là một cơ hội vô cùng to lớn và hiếm hoi đối với khoa học nước nhà. Nó có thể tạo ra những đột phá trong tạo giống lúa và cây nông nghiệp ở Việt Nam, nhưng đồng thời cũng là những thách thức lớn. Con đường khám phá chức năng hệ gen còn đầy khó khăn ở phía trước. Khoa học về hệ gen đòi hỏi tổ chức hệ thống nghiên cứu và các cơ chế chính sách đặc thù.

Như trên chúng tôi đã trình bày, chúng ta rất cần đến các trí tuệ lớn và niềm đam mê để tiếp tục nghiên cứu và khai thác các dữ liệu về trình tự hệ gen. Đọc trình tự hệ gen của một giống lúa cũng gần giống như đọc khoảng 128.000 trang sách, mỗi trang khoảng 3.000 ký tự, trong khi người đọc còn chưa hiểu nghĩa của từng chữ, từng câu trong đó. Để đọc, so sánh, hiểu cấu trúc và phát hiện chức năng hệ gen của hàng chục, hàng trăm giống lúa, đòi hỏi sự đầu tư của các trí tuệ lớn về tin học và sinh học, với trái tim nhiệt huyết sẵn sàng cống hiến và sáng tạo... Đây là thách thức đặt ra không chỉ đối với các nhà khoa học, mà là vấn đề quyết sách quốc gia, vấn đề sách lược, chiến lược, quy hoạch phát triển các ngành khoa học ưu tiên của đất nước. Nếu không giải quyết được các vấn đề vĩ mô trên, thì các vấn đề cục bộ như các cơ hội phát minh của chúng ta về hệ gen sẽ rất hạn chế. Mặt khác, trong tương lai, các nhà khoa học về di truyền phải phối hợp với các viện toán học, tin học và các khoa học gia khác mới mong hoàn thành nhiệm vụ đặt ra. Genome học đang trở thành một ngành khoa học chính xác, một ngành quan trọng của toán sinh và tin sinh học. Một ngành khoa học mà tôi cho rằng Viện Nghiên cứu cao cấp về toán của các nhà toán học nổi tiếng như GS Ngô Bảo Châu nên quan tâm ■