21 Các hệ gene và sự tiến hoá của chúng

▲ **Hình 21.1 Thông tin nào trong hệ gene đã tạo nên con người hoặc tinh tinh?**

CÁC KHÁI NIỆM THEN CHỐT

21.1 Các phương pháp mới đã giúp gia tăng tốc độ giải trình tự các hệ gene

21.2 Các nhà khoa học ứng dụng tin sinh học để phân tích các hệ gene và chức năng của chúng

21.3 Các hệ gene khác nhau về kích cỡ, số gene và mật độ gene

21.4 Sinh vật nhân thực đa bào có nhiều DNA không mã hoá và nhiều họ đa gene

21.5 Lặp đoạn, tái cấu trúc và đột biến trong trình tự DNA đóng góp vào quá trình tiến hoá

21.6 So sánh các trình tự hệ gene cung cấp bằng chứng về các quá trình tiến hoá và phát triển

TỔNG QUAN

Đọc các lá trên cây sự sống

Người phụ nữ trên **Hình 21.1** và con tinh tinh bên cạnh cô đang cười đùa với nhau – có thật vậy không? Họ có hiểu những "câu đùa giỡn" và đáp lại bằng vẻ mặt cùng với các tiếng phát âm của nhau không? Nhờ những kỹ thuật được phát triển gần đây trong việc giải trình tự nhanh toàn bộ các hệ gene, giờ đây chúng ta có thể tuyên bố về cơ sở di truyền liên quan đến các câu hỏi hấp dẫn như vừa được nêu.

Tinh tinh (*Pan troglodytes*) là loài hiện còn tồn tại có họ hàng gần chúng ta nhất trên cây tiến hoá của sự sống. Hệ gene của nó được giải trình tự hoàn toàn vào năm 2005, nghĩa là khoảng 2 năm sau khi việc giải trình tự hệ gene người hoàn thành phần lớn. Giờ đây chúng ta đã có thể so sánh hệ gene của chúng ta với hệ gene của tinh tinh và đối chiếu từng base nitrogen nhằm làm sáng tỏ những thông tin di truyền khác nhau nào đã dẫn đến các đặc điểm khác biệt giữa hai loài linh trưởng này.

Ngoài việc đã xác định được trình tự hệ gene đầy đủ của người và tinh tinh, các nhà nghiên cứu cũng đã thu được trình tự hệ gene đầy đủ của vi khuẩn *E. coli* và nhiều loài sinh vật nhân sơ khác, cũng như của một số loài sinh vật nhân thực, bao gồm *Saccharomyces cerevisiae* (nấm men bia), *Caenorhabitis elegans* (một loài giun tròn), *Drosophila melanogaster* (ruồi quả), *Mus musculus* (chuột nhà) và *Macaca mulatta* (khỉ rhezut). Thậm chí các đoạn DNA từ các loài đã bị tuyệt chủng, như gấu hang (*Ursus spelaneus*) hay voi mamut rậm lông (*Mammuthus primigenius*) cũng đã được giải trình tự. Các trình tự hệ gene đầy đủ hoặc từng phần, bản thân chúng là đối tượng được quan tâm nghiên cứu, đồng thời chúng cung cấp những thông tin sâu hơn về tiến hoá và nhiều quá trình sinh học khác. Bằng việc mở rộng so sánh hệ gene người và tinh tinh với các loài linh trưởng khác cũng như với các loài động vật có quan hệ di truyền xa hơn, chúng ta có thể tìm thấy tập hợp các gene quy định sự khác biệt rõ rệt của mỗi nhóm sinh vật. Xa hơn một chút, sự so sánh với các hệ gene vi khuẩn, vi sinh vật cổ (archaea), nguyên sinh động vật và các loài thực vật sẽ giúp chúng ta làm sáng tỏ lịch sử tiến hoá lâu dài liên quan đến các gene được các loài cùng nhau chia sẻ cùng với các sản phẩm của chúng.

Với việc hệ gene của nhiều loài đã được giải trình tự đầy đủ, các nhà khoa học có thể nghiên cứu các tập hợp gene hoàn chỉnh và sự tương tác của chúng theo một hướng nghiên cứu được gọi là **hệ gene học** (genomics). Các nỗ lực giải trình tự theo hướng nghiên cứu này đã và đang tiếp tục tạo ra những khối dữ liệu khổng lồ. Nhu cầu cần xử lý một lượng thông tin tràn ngập đang tăng lên nhanh chóng đã dẫn đến sự hình thành của lĩnh vực **tin sinh học** (bioinformatics), lĩnh vực ứng dụng các phương pháp khoa học máy tính vào việc lưu giữ và phân tích các số liệu sinh học.

Chúng ta sẽ bắt đầu chương này bằng việc thảo luận về hai hướng nghiên cứu, gồm các kỹ thuật giải trình tự hệ gene và một số tiến bộ trong việc ứng dụng tin sinh học. Sau đó chúng ta sẽ tóm lược những hiểu biết thu nhận được từ việc giải trình tự các hệ gene đã được tiến hành đến nay. Tiếp đến, chúng ta sẽ mô tả thành phần hệ gene người như một hệ gene đại diện cho các sinh vật nhân thực đa bào. Cuối cùng, chúng ta sẽ cùng tìm hiểu những quan điểm về quá trình tiến hoá và các cơ chế phát triển vốn là cơ sở tạo nên sự đa dạng vĩ đại của sự sống hiện có trên Trái Đất.

KHÁI NIỆM **21.1**

Các phương pháp mới đã giúp gia tăng tốc độ giải trình tự các hệ gene

Việc giải trình tự hệ gene người, một dự án tham vọng với tên gọi **Dự án Hệ gene Người** (HGP) được bắt đầu vào năm 1990. Được tổ chức thành một Tổ hợp gồm nhiều nhà khoa học quốc tế được cộng đồng tài trợ, dự án đã được triển khai ở 20 trung tâm giải trình tự lớn thuộc 6 quốc gia, cùng nhiều phòng thí nghiệm nhỏ thực hiện các nhánh của dự án.

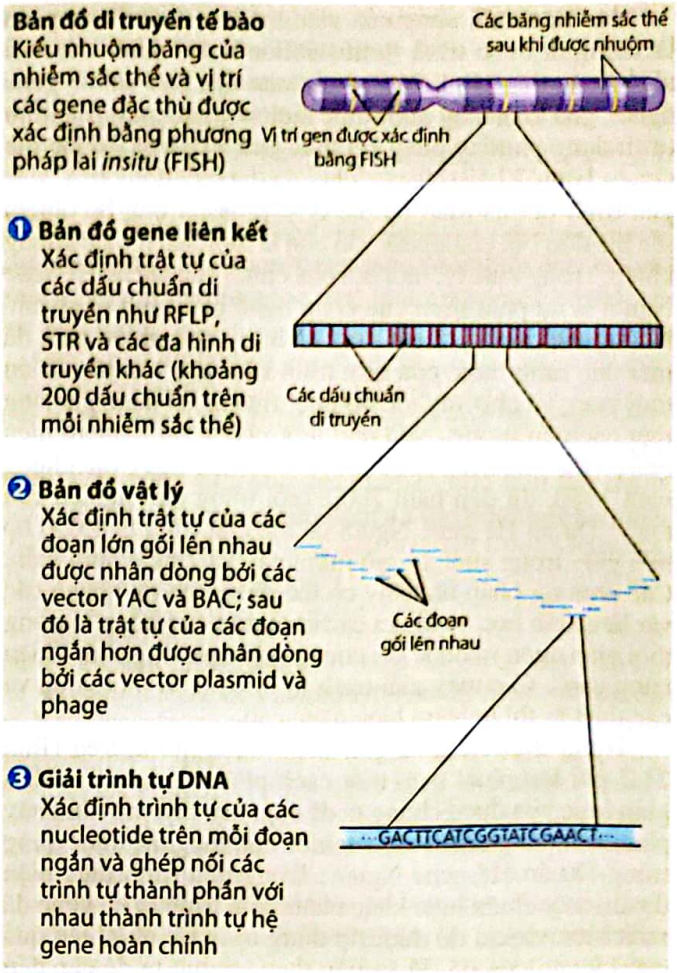
Sau khi việc giải trình tự hệ gene người được hoàn thành phần lớn vào năm 2003, trình tự của mỗi nhiễm sắc thể đã được phân tích kỹ lưỡng và được mô tả trong hàng loạt các bài báo khoa học, trong đó bài báo cuối cùng liên quan đến trình tự của nhiễm sắc thể số 1 được công bố vào năm 2006. Với kết quả này, các nhà nghiên cứu coi việc giải trình tự hệ gene người đã "chính thức hoàn thành". Để đạt được những cột mốc đó, dự án đã được triển khai qua ba giai đoạn với các phát hiện ngày càng chi tiết hơn về hệ gene người; ba giai đoạn đó gồm: lập bản đồ liên kết, lập bản đồ vật lý và giải trình tự DNA.

Giải trình tự hệ gene qua ba giai đoạn

Trước khi Dự án Hệ gene Người bắt đầu, các nghiên cứu trước đó đã phác thảo được một bức tranh sơ bộ về tổ chức hệ gene của nhiều cơ thể sinh vật khác nhau. Ví dụ như, việc phân tích kiểu nhân (karyotype) của nhiều loài đã cho biết số lượng nhiễm sắc thể và kiểu nhuộm băng của chúng (xem Hình 13.3). Và đối với một số gene, vị trí của chúng trên nhiễm sắc thể đã được xác định bởi phương pháp lai huỳnh quang tại chỗ (FISH), phương pháp mà trong đó người ta đem lai các mẫu dò phát huỳnh quang với các nhiễm sắc thể nguyên vẹn được cố định (xem Hình 15.1). Bản đồ di truyền tế bào được xây dựng theo cách này đã cung cấp những thông tin khởi đầu cho việc lập bản đồ chi tiết hơn sau này.

Khi đã có trong tay bản đồ di truyền tế bào của các nhiễm sắc thể, giai đoạn đầu tiên của tiến trình giải trình tự hệ gene người là xây dựng một bản đồ gene liên kết (một loại bản đồ di truyền; xem Chương 15) của khoảng vài nghìn dấu chuẩn di truyền được phân bố khắp các nhiễm sắc thể (**Hình 21.2** giai đoạn ❶). Trật tự vị trí của các dấu chuẩn và khoảng cách giữa chúng trên bản đồ được xác định trên cơ sở tần số tái tổ hợp (xem Hình 15.11). Các dấu chuẩn di truyền có thể là các gene hoặc là các đoạn trình tự DNA khác có thể xác định được, chẳng hạn như các RFLP hay các trình tự lặp lại kế tiếp ngắn (STR) đã được đề cập ở Chương 20. Tính đến năm 1992, các nhà nghiên cứu đã tập hợp được một bản đồ gene liên kết ở người gồm khoảng 5.000 dấu chuẩn khác nhau. Một bản đồ như vậy đã giúp họ xác định được vị trí của các dấu chuẩn khác, bao gồm cả các gene, bằng việc kiểm tra sự liên kết di truyền của chúng với các dấu chuẩn đã biết trước đó. Ngoài ra, nó còn có giá trị là phần cốt lõi của việc lập bản đồ chi tiết hơn tại những vùng nhất định trong hệ gene.

Giai đoạn tiếp theo là việc lập bản đồ vật lý hệ gene người. Trong **bản đồ vật lý**, khoảng cách giữa các dấu chuẩn được biểu diễn bởi đơn vị vật lý, thường là số cặp base nitrogen (bp) dọc theo phân tử DNA. Để lập một



▲ **Hình 21.2 Phương pháp giải trình tự toàn hệ gene qua ba giai đoạn.** Bắt đầu từ một bản đồ di truyền tế bào của mỗi nhiễm sắc thể, các nhà nghiên cứu liên quan đến Dự án Hệ gene Người đã tiến hành các nghiên cứu qua ba giai đoạn để đạt được mục tiêu cuối cùng, đó là giải trình tự toàn bộ từng nucleotide trên mỗi nhiễm sắc thể.

bản đồ hệ gene hoàn chỉnh, một bản đồ vật lý được thiết lập bằng cách cắt phân tử DNA của một nhiễm sắc thể thành một số các đoạn giới hạn rồi xác định trình tự của các đoạn trên phân tử DNA của nhiễm sắc thể gốc. Chìa khoá để thực hiện điều này là cần tạo ra các đoạn DNA gối lên nhau, rồi sử dụng các mẫu dò hoặc phương pháp giải trình tự tự động các trình tự đầu cuối của những đoạn này để tìm ra các trình tự gối lên nhau đó (Hình 21.2, giai đoạn ❷). Bằng cách đó, có thể đặt các đoạn vào đúng trật tự tương ứng của chúng trên nhiễm sắc thể.

Nguồn cung cấp các đoạn DNA dùng cho việc lập bản đồ vật lý dựa trên việc nhân dòng DNA. Để giải trình tự các hệ gene lớn, các nhà khoa học phải thực hiện lặp lại nhiều lần các công việc cắt DNA, nhân dòng và lập bản đồ vật lý. Các vector nhân dòng đầu tiên thường được sử dụng là nhiễm sắc thể nhân tạo nấm men (YAC) cho phép mang những đoạn DNA cài dài đến hàng triệu bp, hoặc nhiễm sắc thể nhân tạo vi khuẩn (BAC) vốn điển hình có thể mang các đoạn cài dài từ 100.000 đến 300.000 bp. Sau khi những đoạn DNA dài như vậy đã được xác định trình tự trên nhiễm sắc thể chính xác, chúng sẽ được cắt thành những phân đoạn nhỏ hơn, rồi được nhân dòng vào các vector plasmid hoặc phage, trước khi những đoạn nhỏ này được dùng để giải trình tự chi tiết.

Mục tiêu cuối cùng của việc lập bản đồ một hệ gene là xác định được trình tự nucleotide hoàn chỉnh của mỗi nhiễm sắc thể (Hình 21.2, giai đoạn ❸). Đối với hệ gene người, giai đoạn này được thực hiện nhờ các máy giải trình tự sử dụng phương pháp kết thúc chuỗi dideoxy được mô tả trên Hình 20.12. Ngay cả khi đã được tự động hoá, việc giải trình tự của toàn bộ 3,2 tỷ cặp base trong bộ nhiễm sắc thể đơn bội của người vẫn còn là một thách thức khủng khiếp. Trong thực tế, một đột phá chính của Dự án Hệ gene Người là sự phát triển của công nghệ giải trình tự nhanh. Những cải tiến kỹ thuật được tích luỹ qua nhiều năm đã mài dũa từng bước của quy trình kỹ thuật vốn tốn nhiều thời gian, và nhờ vậy tốc độ giải trình tự đã được gia tăng một cách ấn tượng. Nếu như một phòng thí nghiệm hiệu quả có thể giải trình tự được 1.000 bp mỗi ngày vào những năm 1980, thì đến năm 2000, mỗi trung tâm nghiên cứu thuộc Dự án Hệ gene Người có thể giải trình tự 1.000 bp *mỗi giây* trong suốt 24 giờ mỗi ngày và 7 ngày mỗi tuần. Các phương pháp như vậy có thể phân tích rất nhanh các vật liệu sinh học và tạo ra các khối dữ liệu khổng lồ trong thời gian ngắn và được gọi chung là các phương pháp "hiệu năng cao". Các máy giải trình tự tự động là một ví dụ về các thiết bị thí nghiệm hiệu năng cao.

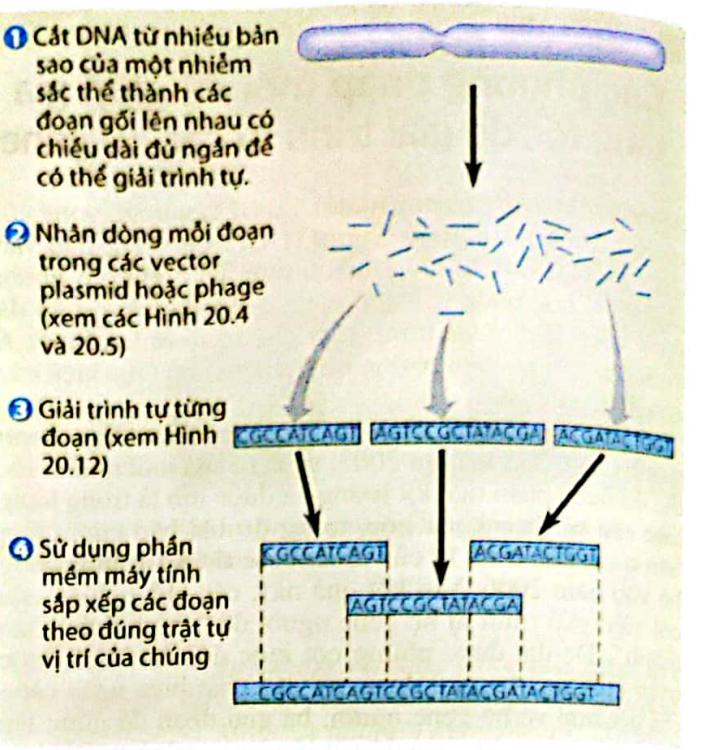
Trong thực tiễn, ba giai đoạn được mô tả trên Hình 21.2 gối lên nhau theo một cách phức tạp hơn mô hình giản lược vừa được chúng ta đề cập; tuy vậy, mô hình này phản ánh đúng chiến lược nghiên cứu tổng thể được dùng trong Dự án Hệ gene Người. Trong quá trình thực hiện dự án, một chiến lược khác nhằm giải trình tự hệ gene đã xuất hiện và sau đó được áp dụng rộng rãi nhờ hiệu quả cực kỳ cao của nó. Phần tiếp theo, chúng ta đề cập đến chiến lược giải trình tự này.

Giải trình tự ngẫu nhiên toàn hệ gene

Năm 1992, mạnh dạn dựa vào các thành tựu mới của kỹ thuật giải trình tự và công nghệ máy tính, J. Craig Venter - một nhà sinh học phân tử - đã phát minh ra một phương pháp giải trình tự toàn hệ gene mới. Được đặt tên là *phương pháp giải trình tự ngẫu nhiên toàn hệ gene* (hay *phương pháp shotgun*), thực chất phương pháp này đã bỏ qua các giai đoạn lập bản đồ gene liên kết và bản đồ vật lý; thay vào đó, nó bắt đầu ngay bằng việc giải trình tự các đoạn DNA ngẫu nhiên của toàn hệ gene. Sau đó, các chương trình máy tính mạnh sẽ tiến hành sắp xếp một số lượng lớn các đoạn DNA đã được giải trình tự, dựa trên các đoạn trình tự ngắn nằm gối lên nhau của chúng, thành một trình tự liên tục duy nhất **(Hình 21.3)**.

Mặc dù ban đầu bị hoài nghi bởi nhiều nhà khoa học, giá trị của phương pháp Venter trở nên rõ ràng vào năm 1995 khi ông và cộng sự công bố hệ gene của một loài sinh vật được giải trình tự hoàn chỉnh đầu tiên, đó là vi khuẩn gây bệnh tiêu chảy *Haemophilus influenzae*. Năm 1998, Venter thành lập một công ty có tên là Celera Genomics và tuyên bố dự định giải trình tự toàn bộ hệ gene người của mình. Năm năm sau, Cerela Genomics và Tổ hợp public consortium đồng thời thông báo việc giải trình tự hệ gene người đã hoàn thành phần lớn, nghĩa là sớm hơn hai năm so với tiến độ dự kiến ban đầu của Dự án Hệ gene Người.

Các đại diện của Tổ hợp public consortium chỉ ra rằng việc hoàn thành giải trình tự hệ gene người của Celera phải dựa nhiều vào các bản đồ di truyền và số liệu trình tự của họ, cũng như các trang thiết bị mà họ thiết lập cho



▲ **Hình 21.3 Phương pháp giải trình tự ngẫu nhiên toàn hệ gene.** Theo phương pháp này, được phát triển bởi Craig Venter và các đồng nghiệp tại Công ty Celera Genomics do chính ông sáng lập, các đoạn DNA ngẫu nhiên được giải trình tự, rồi sau đó chúng được sắp xếp theo đúng trật tự vị trí tương đối với nhau. Hãy so sánh phương pháp này với phương pháp giải trình tự toàn hệ gene qua ba giai đoạn được mô tả trên Hình 21.2.

? *Các đoạn ở giai đoạn 2 trên hình này được vẽ nằm rải rác, trong khi những đoạn ở giai đoạn 2 trên Hình 21.2 được vẽ nằm theo trật tự vị trí. Sự khác biệt trong cách vẽ như vậy phản ánh sự khác biệt giữa hai phương pháp như thế nào?*

dự án đã hỗ trợ nhiều cho các nỗ lực của Celera. Ngược lại, Venter cũng đã dùng lý lẽ để biện hộ cho hiệu quả và giá thành hạ trong phương pháp giải trình tự của Celera, đồng thời chỉ ra rằng Tổ hợp public consortium cũng đã sử dụng các số liệu của họ. Rõ ràng cả hai phương pháp đều có giá trị và cùng đóng góp vào việc nhanh chóng hoàn thành việc giải trình tự hệ gene của một số loài.

Hiện nay phương pháp giải trình tự ngẫu nhiên toàn hệ gene đang được dùng rộng rãi. Theo một cách điển hình, các đoạn DNA được nhân dòng bằng ba loại vector khác nhau, mỗi loại được cài một đoạn DNA có chiều dài xác định. Khoảng cách đã biết giữa các đầu của đoạn DNA cài là một thông tin bổ sung giúp máy tính có thể sắp xếp đúng các trình tự. Một nghiên cứu gần đây so sánh hai chiến lược giải trình tự đã chỉ ra rằng phương pháp shotgun có thể mắc lỗi bỏ qua một số trình tự lặp lại, vì vậy có thể phản ánh không chính xác kích thước thực của hệ gene và có thể bỏ qua một số gene trong những vùng như vậy trên nhiễm sắc thể. Các phương pháp phối hợp cuối cùng đã được áp dụng cho hệ gene người; trong đó phương pháp shotgun có tốc độ nhanh được hỗ trợ bởi bản đồ di truyền của các dòng gene có lẽ là cách hữu hiệu nhất cho những ứng dụng lâu dài.

Đến năm 2007, vẫn còn một phần nhỏ của hệ gene người chưa được giải trình tự. Do có các trình tự DNA lặp lại và bởi một số nguyên nhân chưa biết khác, một số phần nhất định trên nhiễm sắc thể của các cơ thể đa bào rất khó giải trình tự chi tiết bởi các phương pháp thông thường.

Thoạt nhìn thì dường như trình tự hệ gene của người và các sinh vật khác đơn giản chỉ là những trình tự "khô khốc" của các nucleotide, nghĩa là hàng triệu các "chữ cái" A, T, G và C sắp xếp kế tiếp nhau một cách buồn tẻ. Điều cốt yếu để lượng dữ liệu khổng lồ này trở nên có nghĩa là các phương pháp phân tích mà chúng ta sẽ đề cập đến ở tiểu mục tiếp theo.

KIỂM TRA KHÁI NIỆM 21.1

1. Bản đồ gene liên kết và bản đồ vật lý của một nhiễm sắc thể khác nhau cơ bản ở đặc điểm gì?
2. Xét tổng thể, phương pháp lập bản đồ hệ gene được dùng trong Dự án Hệ gene Người và phương pháp giải trình tự ngẫu nhiên toàn hệ gene khác nhau như thế nào?
3. **ĐIỀU GÌ NẾU?** Giả sử bạn quyết định tiến hành giải trình tự hệ gene của một loài chuột đồng, vốn là một loài có quan hệ gần gũi với loài chuột thí nghiệm có trình tự hệ gene đã được xác định hoàn toàn. Tại sao trình tự hệ gene chuột thí nghiệm đã biết đưa bạn đến quyết định chọn phương pháp giải trình tự ngẫu nhiên toàn hệ gene thay cho phương pháp ba giai đoạn?

Câu trả lời có trong Phụ lục A.

KHÁI NIỆM 21.2

Các nhà khoa học ứng dụng tin sinh học để phân tích các hệ gene và chức năng của chúng

Mỗi một trung tâm trong số khoảng 20 trung tâm giải trình tự tham gia dự án Hệ gene Người ngày này qua ngày khác đã tạo ra một lượng khổng lồ các trình tự DNA. Khi số liệu ngày càng được tích luỹ, thì nhu cầu nảy sinh là phải có cách quản lý và theo dõi tất cả các trình tự đã được phát hiện. Nhờ đã chuẩn bị từ trước, các nhà khoa học và các cơ quan quản lý tham gia Dự án Hệ gene Người đã đặt ra một mục tiêu ngay từ đầu là thiết lập các ngân hàng dữ liệu, hay còn gọi là cơ sở dữ liệu, và ngày càng hoàn thiện các phần mềm phân tích dữ liệu. Những cơ sở dữ liệu và những phần mềm này sau đó được tập hợp lại và có thể dễ dàng truy cập và sử dụng trên môi trường internet. Việc hoàn thành mục tiêu này của dự án đã góp phần thúc đẩy quá trình phân tích các trình tự DNA do các nhà khoa học toàn thế giới có thể tiếp cận các tài nguyên tin sinh học, cũng như thúc đẩy việc truyền bá và trao đổi các thông tin có liên quan.

Tập hợp dữ liệu để phân tích các hệ gene

Các cơ quan được chính phủ tài trợ thực hiện vai trò thiết lập các cơ sở dữ liệu và cung cấp các phần mềm nhờ đó các nhà khoa học có thể phân tích các dữ liệu trình tự hệ gene. Chẳng hạn, ở Mỹ, một chương trình hợp tác giữa Thư viện Y học Quốc gia và Viện Y học Quốc gia (NIH) đã thiết lập nên Trung tâm Quốc gia về Thông tin Công nghệ Sinh học (NCBI) đồng thời duy trì một trang Web (www.ncbi.nlm.nih.gov) lưu giữ các tài nguyên tin sinh học hết sức phong phú. Tại trang web này, các đường "link" dẫn đến các cơ sở dữ liệu, các phần mềm và các kho chứa các thông tin về các hệ gene và các chủ đề có liên quan khác. Các trang web tương tự cũng đã được thiết lập bởi Phòng thí nghiệm Sinh học phân tử châu Âu và Ngân hàng Dữ liệu DNA Nhật Bản; đây cũng chính là hai trung tâm nghiên cứu hệ gene cùng hợp tác với NCBI. Những trang web lớn và toàn diện này còn được bổ sung thêm bởi những trang web khác được duy trì bởi các phòng thí nghiệm nhỏ hơn hoặc bởi các cá nhân. Các trang web nhỏ hơn thường cung cấp các cơ sở dữ liệu và các phần mềm được thiết kế cho các mục đích nghiên cứu hẹp hơn, chẳng hạn như để tìm hiểu về những thay đổi di truyền hoặc trong hệ gene liên quan đến một bệnh ung thư nhất định.

Các cơ sở dữ liệu về các trình tự của NCBI được gọi chung là Ngân hàng gene (Genbank). Tính tới tháng 8 năm 2007, Genbank đã chứa trình tự của 76 triệu phân đoạn DNA hệ gene khác nhau, gồm tổng cộng 80 tỷ cặp base ! Các trình tự trong ngân hàng gene liên tục được cập nhật, và ước tính lượng dữ liệu của nó cứ sau khoảng 18 tháng lại tăng lên gấp đôi. Mọi trình tự trong Genbank có thể được truy xuất và phân tích bằng các phần mềm ở trang web của NCBI hoặc từ các trang web khác.

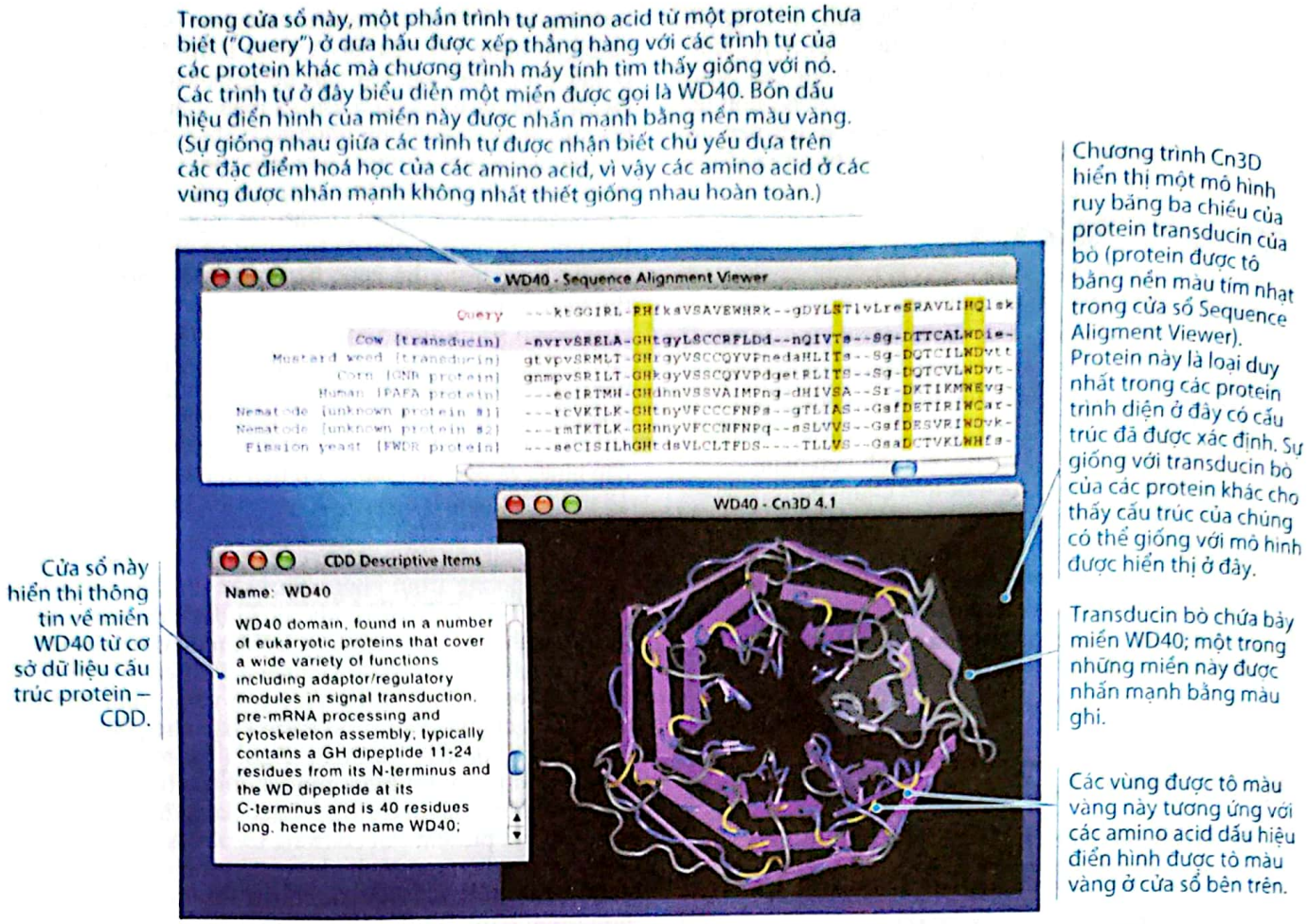
Một chương trình phần mềm sẵn có trên trang Web của NCBI, gọi là BLAST, cho phép bất cứ ai truy cập có thể so sánh được một trình tự DNA nhất định với bất cứ trình tự nào sẵn có trong Genbank trên cơ sở đối chiếu từng cặp base, qua đó tìm thấy các vùng trình tự giống nhau giữa chúng. Một phần mềm khác cho phép so sánh các trình tự protein dự đoán. Ngoài ra, một phần mềm thứ ba cho phép tìm kiếm một chuỗi amino acid (miền) có chức năng sinh học đã biết hoặc đang được dự đoán từ mọi trình tự protein sẵn có trong Genbank; đồng thời, nó có thể biểu diễn mô hình không gian ba chiều của miền chức năng đó cùng với các thông tin có liên quan phù hợp (xem **Hình 21.4** ở trang sau). Thậm chí còn có một chương trình phần mềm có thể so sánh một tập hợp các trình tự, hoặc là các trình tự acid nucleic hoặc là các trình tự polypeptide, và biểu diễn chúng ở dạng cây tiến hoá trên cơ sở mối quan hệ giữa các trình tự. (Chúng ta sẽ đề cập kỹ hơn về những sơ đồ này ở Chương 26).

Trang web của NCBI cũng còn duy trì một cơ sở dữ liệu bao gồm tất cả các cấu trúc ba chiều của protein đã được xác định (để hình dung các cấu trúc protein được phân tích ra sao hãy xem Hình 5.25). Bằng phần mềm máy tính, người xem có thể quay những cấu trúc này để có thể quan sát protein từ mọi phía. Giả sử một nhà nghiên cứu có một trình tự amino acid là trình tự đầy đủ hoặc một phần của một protein chưa biết nào đó, mà nó lại có trình tự giống với một trình tự amino acid có cấu trúc không gian đã biết. Trong trường hợp này, nhà nghiên cứu có thể dự đoán cấu trúc của protein chưa biết bằng một phần mềm, và sử dụng một phần mềm khác để so sánh nó với tất cả các cấu trúc protein đã biết. Những thông tin này có thể giúp nhà nghiên cứu xác định được chức năng của protein chưa biết.

Hiện nay, trên toàn thế giới có rất nhiều nguồn tài nguyên sẵn có cho các nhà nghiên cứu sử dụng. Bây giờ chúng ta sẽ nói đến các chủ đề mà những nguồn tài nguyên này đề cập đến.

Xác định các gene mã hoá protein trong các trình tự DNA

Bằng việc sử dụng các trình tự DNA sẵn có, các nhà di truyền học có thể nghiên cứu trực tiếp các gene mà không nhất thiết phải phỏng đoán về kiểu gene trên cơ sở phân tích kiểu hình như trong các nghiên cứu di truyền học kinh điển trước đây. Tuy vậy, cách tiếp cận này lại có một trở ngại khác: đó là việc xác định kiểu hình trên cơ



▲ Hình 21.4 Các công cụ tin sinh học sẵn có trên internet. Một trang web được Trung tâm Quốc gia Thông tin về Công nghệ Sinh học (Mỹ) duy trì cho phép các nhà khoa học và cộng đồng tiếp cận các trình tự protein và DNA. Trang web này gồm cả kết nối tới một cơ sở dữ liệu cấu trúc protein - CDD (Conserved Domain Database) giúp tìm và mô tả những miền giống nhau ở các protein có quan hệ với nhau, cũng như các phần mềm quan sát ba chiều - Cn3D - cho phép quan sát mô hình ba chiều của các miền cấu trúc đã được xác định. Hình ảnh được minh hoạ ở trên là kết quả tìm kiếm các vùng protein giống với một trình tự amino acid tìm thấy ở một protein của dưa hấu.

sở kiểu gene đã biết. Trên cơ sở một trình tự DNA dài có trên cơ sở dữ liệu như Genbank, bằng cách nào chúng ta có thể nhận ra các gene mã hoá protein vốn chưa từng được biết tới và xác định chức năng của chúng?

Cách thông thường là sử dụng một phần mềm để tìm kiếm trong những trình tự này sự có hay không của các tín hiệu khởi đầu và kết thúc phiên mã hoặc dịch mã, hoặc là các vị trí cắt - nối RNA hay các tín hiệu khác, thường có ở các gene mã hoá protein. Phần mềm này đồng thời cũng tìm kiếm các đoạn trình tự ngắn tương ứng với các trình tự thường có trên các phân tử mRNA đã biết. Hàng nghìn các trình tự như vậy, được gọi là các *đoạn đánh dấu trình tự biểu hiện* hay *EST* được thu thập từ các trình tự cDNA và được máy tính tập hợp lại thành các cơ sở dữ liệu. Kiểu phân tích này cho phép xác định được các trình tự tương ứng với các gene mã hoá protein mà trước đó chưa từng được biết tới.

Khoảng một nửa số gene ở người đã được biết từ trước khi dự án hệ gene người bắt đầu. Vậy đối với những gene còn lại, việc phân tích các trình tự DNA bằng cách nào cho biết chúng là các gene chưa được biết trước đó? Manh mối để xác định những gene này xuất phát từ việc so sánh trình tự của các "gene ứng cử viên" (các trình tự được dự đoán là gen) với trình tự của các gene đã biết có nguồn gốc từ các sinh vật khác bằng việc sử dụng các phần mềm đã được nhắc đến ở trên. Do tính thoái hoá của mã di truyền, bản thân trình tự DNA có thể có mức độ biến đổi lớn hơn so với các trình tự protein tương ứng. Vì vậy, với các nhà khoa học quan tâm đến protein, họ thường tiến hành so sánh giữa trình tự amino acid của protein phỏng đoán với các trình tự của các protein đã biết.

Đôi khi một trình tự vừa mới được xác định khớp hoàn toàn hay một phần với trình tự của một gene hoặc một protein mà chức năng đã biết rõ. Ví dụ như, một phần của một gene mới có thể khớp với một gene đã biết mã hoá cho một protein kinase, một protein quan trọng tham gia vào một con đường truyền tín hiệu (xem Chương 11), chỉ ra nhiều khả năng gene mới này có thể có chức năng tương tự. Theo một cách khác, trình tự của một gene mới lại giống với một trình tự đã từng được biết từ trước nhưng chưa rõ chức năng. Một khả năng khác là trình tự mới được xác định không giống với bất cứ một trình tự nào đã từng được biết đến. Điều này là đúng đối với ít nhất một phần ba các gene của *E. coli* khi hệ gene của vi khuẩn này được giải trình tự. Trong trường hợp cuối cùng, chức năng của protein thường được suy diễn bằng việc kết hợp giữa các nghiên cứu về chức năng phân tử và hoá sinh học. Các nghiên cứu về hoá sinh nhằm xác định cấu trúc không gian ba chiều cũng như các thuộc tính hoá lý của protein, chẳng hạn như các vị trí liên kết của protein với các phân tử khác. Trong khi đó, các nghiên cứu về chức năng phân tử thường tiến hành làm bất hoạt

hoặc làm giảm mức độ biểu hiện của các gene mới xác định rồi theo dõi sự thay đổi của kiểu hình, qua đó xác định chức năng của gene. RNAi được mô tả ở Chương 20, là một ví dụ về kỹ thuật phòng thí nghiệm được dùng để bất hoạt chức năng của gene.

Tìm hiểu các gene và các sản phẩm của gene ở cấp độ sinh học hệ thống

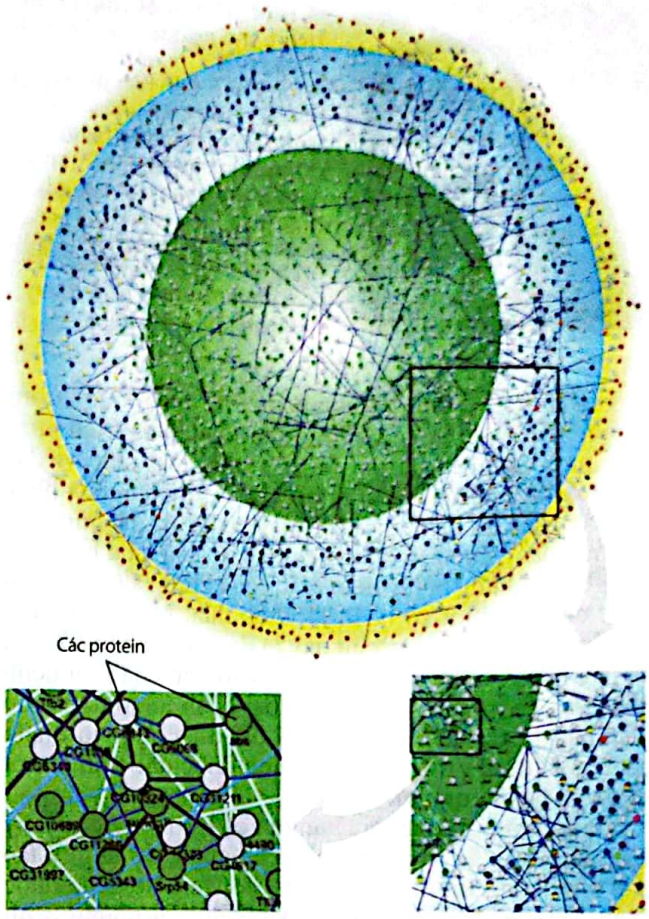
Sức mạnh đầy ấn tượng của các công cụ tin sinh học và máy tính cho phép các nhà khoa học giờ đây có thể nghiên cứu toàn bộ các gene thuộc các bộ nhiễm sắc thể và sự tương tác của chúng với nhau, cũng như có thể so sánh hệ gene từ các loài khác nhau. Hệ gene học là một tài nguyên thông tin phong phú và chuyên sâu có thể trả lời các câu hỏi cơ bản về cách tổ chức của các hệ gene, về sự điều hoà sự biểu hiện các gene, về các quá trình sinh trưởng và phát triển, kể cả tiến hoá.

Những thành công trong lĩnh vực giải trình tự các hệ gene và khả năng nghiên cứu toàn bộ các gene thuộc các bộ nhiễm sắc thể khác nhau đã thúc đẩy các nhà khoa học nỗ lực nghiên cứu một hệ thống tương tự các hệ protein (*proteomes*) được mã hoá tương ứng bởi các hệ gene, từ đó hình thành nên một lĩnh vực nghiên cứu mới gọi là **hệ protein học** (proteomics). Các protein, chứ không phải các gene mã hoá chúng, trong thực tế thực hiện phần lớn các hoạt động sống của tế bào. Vì vậy, để tìm hiểu sự biểu hiện chức năng sinh học của các tế bào và cơ thể, chúng ta phải tìm hiểu các protein được tạo ra khi nào và ở đâu trong mỗi cơ thể, cũng như việc chúng tương tác với nhau thế nào trong các mạng lưới tương tác phân tử.

Các hệ thống được nghiên cứu ra sao: Một ví dụ

Các lĩnh vực hệ gene học và hệ protein học cho phép các nhà sinh học tiến hành các nghiên cứu về sự sống ở quy mô ngày càng rộng lớn và theo xu hướng toàn cầu. Bằng việc sử dụng các công cụ mà chúng ta đã mô tả, các nhà sinh học đã bắt đầu tập hợp các dữ liệu về các gene và các protein, tức là liệt kê tất cả các cấu phần tham gia vào việc điều hành các hoạt động của tế bào, mô và cơ thể. Với tập hợp các dữ liệu như vậy, các nhà nghiên cứu có thể chuyển mối quan tâm của họ từ mỗi cấu phần đơn lẻ sang sự biểu hiện chức năng ở dạng tổ hợp gồm nhiều cấu phần ở các cấp độ của hệ thống sinh học. Nhớ lại ở Chương 1, chúng ta đã đề cập đến sinh học hệ thống là lĩnh vực mô hình hoá các biểu hiện hoạt động năng động của các hệ thống sinh học toàn bộ.

Một ứng dụng cơ bản của hướng nghiên cứu sinh học hệ thống là xác định được các mạng lưới gene và các mạng lưới tương tác của các protein. Chẳng hạn như, để xây dựng được sơ đồ mạng lưới tương tác giữa các protein ở ruồi *Drosophila* như được nêu ở Chương 1, các nhà nghiên cứu đã bắt đầu từ trên 10.000 RNA dự đoán. Sau đó, bằng các phương pháp phân tử, họ đã kiểm tra sự tương tác giữa toàn bộ hoặc một phần các sản phẩm protein do các phân tử này mã hoá. Bằng việc sử dụng các phép phân tích thống kê để chọn ra các mối tương tác có số liệu thuyết phục nhất, họ đã tìm ra khoảng 4.700 loại protein dường như tham gia vào trên 4.000 mối tương tác khác nhau. Một phần trong những mối tương tác này được minh hoạ ở dạng sơ đồ trên **Hình 21.5**; chi tiết có thể được nhìn dễ hơn ở hai hình phóng to bên dưới. Để có thể xử lý một số lớn các dữ liệu thu được về các mối tương tác protein - protein phức tạp thu được từ các thí nghiệm này, đồng thời có thể tổ hợp chúng với nhau dưới dạng các sơ đồ mô hình, chúng ta cần đến các hệ thống



▲ **Hình 21.5 Sinh học hệ thống tiếp cận các tương tác protein.** Bản đồ tương tác protein tổng thể này hiển thị một tập hợp con của các tương tác nhiều khả năng nhất (đường kẻ nối) từ 2.300 protein (vòng tròn nhỏ) ở ruồi *Drosophila*. Ba màu nền khác nhau trên bản đồ tương ứng với vị trí chung của mỗi protein: màu xanh lục là nhân, xanh lam là tế bào chất và vàng là màng sinh chất. Các protein được mã hoá bằng màu tương ứng với vị trí định vị trong tế bào đặc thù của chúng; ví dụ, các vòng tròn màu xanh lục là các protein trong nhân.

máy tính hiệu năng cao, các công cụ toán học và các phần mềm được phát triển mới. Như vậy, có thể nói sinh học hệ thống trong thực tế đã trở thành hiện thực nhờ các tiến bộ của tin sinh học.

Ứng dụng sinh học hệ thống trong Y học

Dự án Atlat Hệ gene Ung thư là một ví dụ khác về sinh học hệ thống mà ở đó người ta đồng thời tiến hành phân tích một số lớn các gene và sản phẩm của gene tương tác với nhau. Dự án này đặt dưới sự chỉ đạo phối hợp của Viện Ung thư Quốc gia (Mỹ)và NIH nhằm tìm hiểu những thay đổi trong các hệ thống sinh học dẫn đến sự phát sinh ung thư. Trong giai đoạn 3 năm thử nghiệm dự án (từ 2007 đến 2010), các nhà nghiên cứu tập trung phân tích ba loại ung thư - ung thư phổi, ung thư buồng trứng và u nguyên bào đệm (glioblastoma) của não- thông qua việc tìm hiểu sự khác nhau trong trình tự của các gene và sự biểu hiện của chúng ở các tế bào ung thư so với các tế bào bình thường. Một tập hợp gồm khoảng 2.000 gene ở các tế bào ung thư sẽ được giải trình tự vào các thời điểm khác nhau trong quá trình tiến triển của bệnh nhằm tìm ra những thay đổi hoặc gây ra do đột biến hoặc gây ra bởi các cơ chế tái cấu trúc nhiễm sắc thể khác. Nếu những nghiên cứu này thành công, chúng sẽ được mở rộng áp dụng để nghiên cứu các loại bệnh ung thư khác.

◀ Hình 21.6 Một chip phân tích gene người. Các điểm nhỏ chứa DNA được xếp thành các đường kẻ ô trên bản silicon này đại diện cho hầu hết các gene trong hệ gene người. Nhờ sử dụng chip này, các nhà nghiên cứu có thể phân tích cùng lúc mức biểu hiện của tất cả các gene, qua đó giúp giảm lượng hoá chất cần dùng tối đa đồng thời đảm bảo điều kiện đồng đều cho tất cả các gene.

Sinh học hệ thống có tiềm năng ứng dụng to lớn trong y học, mặc dù hiện nay nó mới bắt đầu được triển khai. Đến nay, người ta đã tạo ra được các loại chip vi dãy (microarray) làm bằng thuỷ tinh hoặc silicon chứa phần lớn các gene đã biết của người **(Hình 21.6)**. Những chip như vậy đang được sử dụng để phân tích sự biểu hiện của các gene ở những bệnh nhân mắc các chứng bệnh ung thư khác nhau và một số bệnh lý khác nữa. Mục đích cuối cùng của những nghiên cứu này là đề ra các phác đồ điều trị phù hợp đặc thù với bản chất di truyền của mỗi bệnh nhân và đặc trưng đối với mỗi loại bệnh ung thư mà họ mắc phải. Cách tiếp cận này đã đạt được một số thành công nhất định trong việc xác định được đặc tính ở một số nhóm bệnh ung thư.

Cuối cùng, mỗi người chúng ta có thể có một hồ sơ y học cùng với các trình tự DNA của mình; đó là một tập hợp nhỏ thông tin di truyền với các vùng hệ gene được đánh dấu cho biết xu hướng mẫn cảm với những bệnh nhất định. Lúc này, tiềm năng ứng dụng trong phòng tránh và điều trị bệnh đối với mỗi người sẽ thành hiện thực.

Sinh học hệ thống là một cách tiếp cận hiệu quả cao để nghiên cứu các đặc tính nổi trội ở cấp độ phân tử. Từ Chương 1 chúng ta nhớ lại rằng, theo khái niệm đặc tính nổi trội, các đặc tính mới xuất hiện ở mỗi bậc tổ chức phức tạp hơn thường bắt nguồn từ sự sắp xếp các bộ phận cấu thành của cấp độ tổ chức thấp hơn. Khi chúng ta hiểu biết ngày càng đầy đủ hơn về cách sắp xếp và tổ hợp của các cấu phần thuộc các hệ thống di truyền, chúng ta càng hiểu biết sâu hơn về hoạt động của các cơ thể sống. Phần còn lại của chương này chúng ta sẽ tổng quát lại những kiến thức mà chúng ta đã học được từ các nghiên cứu hệ gene học cho đến giờ.

KIỂM TRA KHÁI NIỆM 21.2

1. Internet có vai trò như thế nào trong các nghiên cứu hiện nay về các hệ gene học và protein học?
2. Hãy giải thích ưu thế của các nghiên cứu theo hướng sinh học hệ thống khi tìm hiểu về ung thư so với phương pháp nghiên cứu độc lập từng gene vào mỗi thời điểm.
3. **ĐIỀU GÌ NẾU?** Giả sử bạn đang dùng một phương pháp nghiên cứu di truyền kinh điển để tìm hiểu một tính trạng di truyền ở ruồi *Drosophila*. Cụ thể, bạn đã gây đột biến ở ruồi và chọn lọc ra được các cá thể có kiểu hình mà bạn quan tâm. Giả thiết bạn cũng có thể sử dụng các công cụ sinh học phân tử để thu được vùng DNA mang đột biến. Bạn sẽ tiếp tục phân tích đột biến đó như thế nào để có thể xác định được cách mà nó liên quan đến kiểu hình được quan tâm?

Câu trả lời có trong Phụ lục A.

KHÁI NIỆM 21.3

Các hệ gene khác nhau về kích cỡ, số gene và mật độ gene

Tính đến đầu năm 2008, việc giải trình tự của trên 700 hệ gene đã hoàn thành và khoảng trên 2.700 hệ gene khác đang tiếp tục được giải trình tự. Trong nhóm các hệ gene đã được giải trình tự hoàn toàn, có khoảng 600 hệ gene vi khuẩn và 50 hệ gene của các loài Archaea (vi sinh vật cổ). Trong số 65 loài sinh vật nhân thực thuộc nhóm này có các loài động vật có xương sống, các loài động vật không xương sống, các nguyên sinh động vật, nấm và thực vật. Các trình tự hệ gene đã được tích luỹ chứa đựng một tài nguyên thông tin phong phú mà hiện nay chúng ta mới bắt đầu khai thác. Cho đến nay chúng ta đã học được gì từ việc so sánh các hệ gene đã được giải trình tự? Trong mục này, chúng ta sẽ xem xét các đặc tính về kích cỡ hệ gene, số gene và mật độ gene của chúng. Do xét về chi tiết, các đặc tính này rất đa dạng, nên chúng ta chỉ nhấn mạnh vào các xu hướng chung; tuy vậy, bên cạnh các xu hướng chung thì thường xuất hiện các ngoại lệ.

Kích cỡ hệ gene

Khi so sánh hệ gene giữa ba siêu giới (vi khuẩn, vi sinh vật cổ và sinh vật nhân thực), chúng ta nhận thấy một xu hướng khác biệt chung về kích cỡ hệ gene giữa các sinh vật nhân sơ (vi khuẩn và sinh vật cổ) với sinh vật nhân thực **(Bảng 21.1)**. Ngoài một số ngoại lệ, phần lớn hệ gene vi khuẩn có kích cỡ từ 1 đến 6 triệu cặp base (bp); chẳng hạn như hệ gene của *E. coli* là 4,6 triệu bp. Hệ gene của các vi sinh vật cổ trong phần lớn trường hợp có kích cỡ giống với hệ gene vi khuẩn. (Tuy vậy, cần phải nhớ rằng mới chỉ có một số ít hệ gene sinh vật cổ đã được giải trình tự hoàn toàn, vì vậy "bức tranh toàn cảnh" này cũng có thể sẽ thay đổi.) Các hệ gene sinh vật nhân thực có xu hướng lớn hơn. Hệ gene của nấm men đơn bào *Saccharomyces cerevisiae* là khoảng 13 Mb (triệu cặp base); trong khi đó, phần lớn các loài động vật và thực vật, tức là các sinh vật đa bào, có kích cỡ hệ gene ít nhất là 100 Mb. Hệ gene ruồi quả có kích cỡ là 180 Mb, còn hệ gene người là 3.200 Mb, nghĩa là lớn hơn từ 500 đến 3.000 lần so với một hệ gene vi khuẩn điển hình.

Bên cạnh sự khác biệt chung giữa hệ gene của các sinh vật nhân sơ và sinh vật nhân thực, thì việc so sánh kích cỡ hệ gene trong phạm vi các loài sinh vật nhân thực lại không phản ánh mối tương quan có hệ thống giữa kích cỡ hệ gene với kiểu hình của các loài sinh vật. Chẳng hạn như, hệ gene của loài *Fritillaria assyriaca*, một loài hoa thuộc họ loa kèn (huệ tây), có kích cỡ là 120 tỷ cặp base (120.000 Mb), tức là lớn hơn khoảng 40 lần so với hệ gene người. Nhưng còn kinh ngạc hơn là hệ gene một loài amip đơn bào, *Amoeba dubia*, có kích cỡ khoảng 670.000 Mb. (Hệ gene loài này chưa được giải trình tự.) Trong phạm vi hẹp hơn, việc so sánh hệ gene giữa hai loài côn trùng cho thấy hệ gene của dế (*Anabrus simplex*) lớn hơn 11 lần so với hệ gene của ruồi quả (*Drosophila melanogaster*). Kích cỡ hệ gene cũng biến động rộng trong phạm vi mỗi nhóm loài nguyên sinh động vật, côn trùng, lưỡng cư và thực vật; nhưng ít biến động hơn trong phạm vi các loài thú và bò sát.

Số lượng gene

Một xu hướng khác biệt tương tự cũng đúng khi xét về số gen: nghĩa là, nhìn chung các vi khuẩn và sinh vật cổ có ít gene hơn so với sinh vật nhân thực. Các vi khuẩn và sinh vật cổ sống tự do có từ 1.500 đến 7.500 gene, trong khi số gene ở các sinh vật nhân thực dao động từ khoảng 5.000 gene ở các nấm đơn bào cho đến ít nhất 40.000 gene ở một số loài sinh vật nhân thực đa bào (xem Bảng 21.1).

Trong phạm vi các loài sinh vật nhân thực, số gene ở mỗi loài thường thấp hơn số gene được dự đoán đơn thuần trên cơ sở kích cỡ hệ gene của chúng. Nhìn vào Bảng 21.1, bạn có thể thấy hệ gene giun tròn *C. elegans* có kích cỡ là 100 Mb và chứa khoảng 20.000 gene. Trong khi đó, hệ gene *Drosophila* có kích cơ gần gấp đôi (180 Mb), song chỉ có số gene bằng khoảng hai phần ba - tức là, chỉ có 13.700 gene.

Hãy xem một ví dụ khác gần gũi hơn, chúng ta để ý thấy hệ gene người chứa 3.200 Mb, tức là lớn hơn khoảng trên 10 lần so với các hệ gene *Drosophila* và *C. elegans*. Khi Dự án Hệ gene Người khởi động, trên cơ sở số protein đã biết, các nhà sinh học mong đợi sẽ có khoảng từ 50.000 đến 100.000 gene sẽ được xác định sau khi hoàn thành việc giải trình tự hệ gene. Dọc theo tiến trình triển khai dự án, số gene ước lượng có trong hệ gene người được sửa đổi nhiều lần theo xu hướng giảm dần; và đến năm 2007, số gene ước lượng được tin cậy hơn cả dừng ở con số 20.488 gene. Số lượng gene tương đối thấp này, chỉ gần giống số gene có ở loài giun tròn *C. elegans*, đã gây sửng sốt nhiều nhà sinh học vốn đã luôn mong đợi hệ gene người có nhiều gene hơn.

Thuộc tính di truyền nào đã cho phép loài người (và nhiều loài động vật có xương sống khác) tiến hoá mà không cần nhiều gene hơn so với giun tròn? Một yếu tố quan trọng đó là các trình tự mã hoá trong các hệ gene động vật có xương sống có đặc điểm "một vốn bốn lời" do chúng có nhiều cách cắt - nối các bản phiên mã khác nhau. Chúng ta nhớ lại rằng quá trình này có thể tạo ra nhiều hơn một loại protein biểu hiện chức năng xuất phát từ một gene duy nhất (xem Hình 18.11). Ví dụ như, hầu hết các gene ở người đều chứa nhiều exon, và ước lượng có khoảng 75% số gene gồm nhiều exon này được cắt - nối ít nhất bằng hai cách khác nhau. Nếu chúng ta giả thiết mỗi gene khi được cắt nối theo các cách khác nhau, trung bình xác định 3 chuỗi polypeptide khác nhau, thì tổng số chuỗi polypeptide khác nhau ở người sẽ đạt con số khoảng 75.000. Sự đa dạng của các chuỗi polypeptide thực tế còn hơn bởi các biến đổi sau dịch mã, chẳng hạn bởi sự cắt tỉa các amino acid hay gắn thêm các gốc cacbohydrate diễn ra khác nhau ở các tế bào khác nhau hoặc ở các giai đoạn khác nhau của quá trình phát triển.

Bảng 21.1 Kích cỡ hệ gene và số gene ước tính*

Loài	Kích cỡ hệ gene đơn bội (Mb)	Số gene	Số gene/ Mb
Vi khuẩn			
Haemophilus influenzae	1,8	1.700	940
Escherichia coli	4,6	4.400	950
Sinh vật cổ			
Archaeoglobus fulgidus	2,2	2.500	1,130
Methanosarcina barkeri	4,8	3.600	750
Sinh vật nhân thực			
Saccharomyces cerevisiae (nấm men)	13	6.200	480
Caenorhabditis elegans (giun tròn)	100	20.000	200
Arabidopsis thaliana (cây thuộc họ Mù tạt)	118	25.500	215
Drosophila melanogaster (ruồi quả)	180	13.700	76
Oryza sativa (lúa gạo)	390	40.000	140
Danio rerio (cá ngựa)	1.700	23.000	13
Mus musculus (chuột nhà)	2.600	22.000	11
Homo sapiens (người)	3.200	20.500	7
Fritillaria assyriaca (cây thuộc họ Loa kèn)	120.000	ND	ND

* Một số số liệu trên đây có thể sẽ được chỉnh lý sau này do các phân tích hệ gene vẫn đang tiếp tục tiến hành. Mb = 1 triệu cặp base (bp). ND = chưa xác định.

Mật độ gene và các trình tự DNA không mã hoá

Bên cạnh kích cỡ hệ gene và số gene, chúng ta cũng có thể so sánh mật độ gene ở những loài khác nhau, nghĩa là có bao nhiêu gene trên một đơn vị chiều dài của DNA. Khi chúng ta so sánh hệ gene giữa các loài vi khuẩn, sinh vật cổ và sinh vật nhân thực, chúng ta thấy sinh vật nhân thực thường có hệ gene lớn hơn nhưng lại có số gene ít hơn trên cùng một số nhất định các cặp base. Người có kích cỡ hệ gene lớn hơn hàng trăm thậm chí hàng nghìn lần so với hệ gene của phần lớn các loài vi khuẩn, nhưng như chúng ta đã nói, người chỉ có số gene gấp từ 5 đến 15 lần so với những loài này; như vậy, mật độ gene ở người là thấp hơn (xem Bảng 21.1). Ngay cả các loài sinh vật nhân thực đơn bào, như nấm men, cũng có ít gene hơn trong mỗi một triệu cặp base so với các loài vi khuẩn và sinh vật cổ. Trong số các hệ gene đã được giải trình tự hoàn toàn đến nay, người và các loài thú có mật độ gene thấp nhất.

Trong tất cả các hệ gene vi khuẩn đã được nghiên cứu đến nay, phần lớn DNA chứa các gene mã hoá cho protein, tRNA hoặc rRNA; một lượng nhỏ của các trình tự DNA còn lại gồm chủ yếu là các trình tự điều hoà không được phiên mã, chẳng hạn như các trình tự khởi động (promoter). Trình tự các nucleotide dọc theo một gene mã hoá protein ở vi khuẩn thường không bị ngắt quãng từ vị trí bắt đầu cho đến vị trí kết thúc bởi các trình tự không mã hoá (intron). Ngược lại, ở các hệ gene sinh vật nhân thực, phần lớn DNA hoặc không được dùng để mã hoá cho protein hoặc không được phiên mã thành các phân tử RNA biểu hiện chức năng (như tRNA chẳng hạn), đồng thời DNA chứa nhiều trình tự điều hoà phức tạp. Trong thực tế, hệ gene người chứa DNA không mã hoá nhiều hơn khoảng 10.000 lần so với hệ gene vi khuẩn. Một số trình tự DNA không mã hoá này ở sinh vật nhân thực đa bào xuất hiện trong các intron của các gene. Thực tiễn cho thấy các intron là nhân tố chính dẫn đến phần lớn các khác biệt về chiều dài trung bình giữa các gene của người (27.000 bp) so với các gene của vi khuẩn (1.000 bp).

Bên cạnh các intron, các sinh vật nhân thực đa bào chứa một lượng lớn DNA không mã hoá ở giữa các gene.

Trong mục tiếp theo, chúng ta sẽ mô tả thành phần và cách sắp xếp các chuỗi trình tự lớn của DNA như vậy trong hệ gene người.

KIỂM TRA KHÁI NIỆM 21.3

1. Theo các số liệu ước tính hiện nay, hệ gene người chứa khoảng 20.500 gene. Tuy vậy, có bằng chứng cho thấy các tế bào người có thể sản sinh nhiều hơn 20.500 loại chuỗi polypeptide khác nhau. Những quá trình nào có thể giúp giải thích cho sự không nhất quán này?
2. Số hệ gene được giải trình tự đang tiếp tục tăng lên đều đặn. Hãy sử dụng trang web www.genomesonline.org để tìm số hệ gene hiện tại thuộc các siêu giới khác nhau đã được giải trình tự hoàn toàn, cũng như số hệ gene đang tiếp tục được giải trình tự (gợi ý: Hãy dùng chuột nháy kép vào khẩu lệnh "GOLD tables" rồi sau đó nháy kép vào "Published Complete Genomes" để có thêm thông tin.)
3. **ĐIỀU GÌ NẾU?** Các quá trình tiến hoá nào có thể giải thích cho việc các sinh vật nhân sơ có hệ gene nhỏ hơn so với các sinh vật nhân thực?

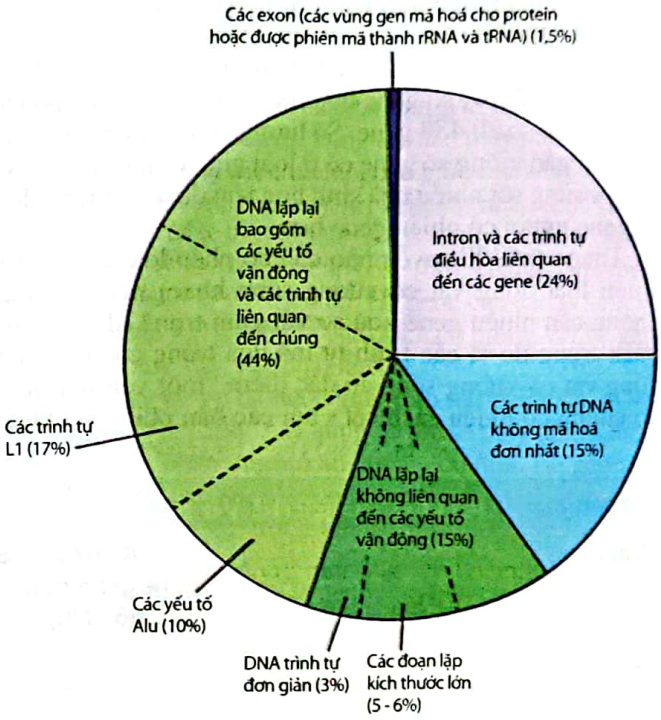
Câu trả lời có trong Phụ lục A.

KHÁI NIỆM 21.4

Sinh vật nhân thực đa bào có nhiều DNA không mã hoá và nhiều họ đa gene

Đến đây, có thể nói chúng ta đã dùng phần lớn dung lượng của chương này, mà thực tế là của cả Phần III, để tập trung nói về các gene mã hoá protein. Nhưng trong thực tế, các vùng mã hoá của những gene này và các gene mã hoá cho các sản phẩm RNA như rRNA, tRNA và tiểu-RNA (miRNA hay microRNA) chỉ chiếm một tỷ lệ nhỏ trong hệ gene của phần lớn các sinh vật nhân thực đa bào. Một phần lớn hệ gene của hầu hết sinh vật nhân thực là các trình tự DNA hoặc không mã hoá cho protein hoặc không được phiên mã để tạo nên các loại RNA có chức năng đã biết; những trình tự DNA không mã hoá này trước kia thường được mô tả như các "DNA dư thừa". Tuy vậy, ngày càng có nhiều bằng chứng cho thấy những trình tự DNA này giữ vai trò quan trọng trong hoạt động sống của tế bào; ý tưởng này đồng thời được củng cố bởi sự tồn tại một cách bền vững qua hàng trăm thế hệ của những trình tự này ở nhiều hệ gene khác nhau. Chẳng hạn, khi so sánh hệ gene giữa người với chuột đồng và chuột nhà, các nhà nghiên cứu tìm thấy có đến 500 vùng DNA không mã hoá trong hệ gene giống hệt nhau ở cả ba loài. Ở những loài này, mức độ bảo thủ của những trình tự này thậm chí còn cao hơn so với các vùng mã hoá protein; điều này ủng hộ mạnh mẽ cho giả thiết các vùng không mã hoá có những chức năng quan trọng. Trong mục này chúng ta sẽ tìm hiểu các gene và các trình tự DNA không mã hoá được tổ chức như thế nào trong hệ gene của các sinh vật nhân thực, với ví dụ chủ yếu chính là hệ gene người của chúng ta. Cách tổ chức của hệ gene cho chúng ta biết con đường mà các hệ gene đã và đang tiếp tục tiến hoá; đây cũng là nội dung được đề cập tiếp theo.

Khi hệ gene người đã được giải trình tự hoàn toàn, một sự thật được bộc lộ rõ ràng là chỉ có 1,5% trình tự nucleotide trong hệ gene được dùng hoặc để mã hoá cho các protein hoặc được phiên mã thành các phân tử rRNA và tRNA. **Hình 21.7** cho thấy những gì chúng ta đã biết cấu trúc nên 98,5% trình tự còn lại của hệ gene người. Các trình tự điều hoà liên quan đến các gene và các trình tự intron chiếm 24% hệ gene người; phần còn lại, nằm giữa các gene biểu hiện chức năng, gồm các trình tự không mã hoá đơn nhất, chẳng hạn như các đoạn của gene và các **gene giả**, tức là các gene cũ vốn từng tồn tại nhưng sau đó do tích luỹ các đột biến đã trở nên mất chức năng. Tuy vậy, phần lớn các trình tự DNA giữa các gene là những trình tự **DNA lặp lại**, tức là các trình tự có mặt với nhiều bản sao trong hệ gene. Điều đáng ngạc nhiên là ba phần tư của các trình tự DNA lặp lại này (tương ứng với 44% của toàn bộ hệ gene người) tạo nên các đơn vị được gọi là các yếu tố di truyền vận động hoặc các trình tự có liên quan đến chúng.



▲ Hình 21.7 Các loại trình tự DNA trong hệ gene người. Các trình tự gene mã hoá cho protein hoặc được phiên mã thành các phân tử rRNA hay tRNA chỉ chiếm khoảng 1,5% hệ gene người (màu tía sẫm trên biểu đồ toả tròn), trong khi các trình tự điều hoà và các intron liên quan đến các gene (màu tía nhạt) chiếm khoảng 1/4 hệ gene. Phần lớn hơn cả của hệ gene người là những trình tự không mã hoá cho protein và cũng không được dùng để tạo ra các loại RNA đã biết, mà phần nhiều trong những trình tự này là các DNA lặp lại (màu xanh lục sẫm và nhạt). Do DNA lặp lại là những trình tự khó phân tích và khó giải trình tự hơn cả, nên sự phân loại của một phần những trình tự này ở trên chỉ có tính ước đoán, và các tỷ lệ phần trăm được nêu có thể sẽ thay đổi đôi chút khi các nghiên cứu phân tích hệ gene vẫn đang tiếp diễn. Những gene mã hoá các miRNA mới được tìm thấy gần đây thuộc các vùng DNA không mã hoá đơn nhất (tức là không lặp lại) và trong các intron; nghĩa là, chúng thuộc hai vùng của đồ thị toả tròn trên đây.

Các yếu tố di truyền di động và các trình tự có liên quan

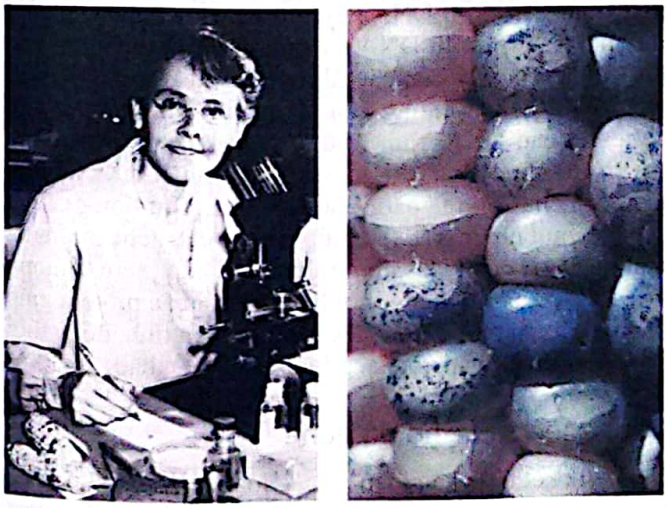
Cả sinh vật nhân sơ cũng như sinh vật nhân thực đều có trong hệ gene những đoạn trình tự DNA có thể di chuyển từ vị trí này sang vị trí khác trong hệ gene. Những đoạn trình tự DNA như vậy được gọi là *các yếu tố di truyền di động*, hay được gọi tắt là các yếu tố di động. Trong quá trình được gọi là *chuyển vị*, một yếu tố di động sẽ di chuyển từ một vị trí trên DNA trong tế bào tới một vị trí đích khác nhờ một quá trình tái tổ hợp. Đôi khi các yếu tố di động được gọi là các "gene nhảy", nhưng thuật ngữ này thực tế dễ gây hiểu nhầm bởi trong thực tế những đoạn trình tự DNA vận động không bao giờ rời khỏi DNA của tế bào. (Các vị trí gốc và vị trí đích mới của các yếu tố di động được đưa đến gần nhau bởi cơ chế "bẻ cong" DNA.)

Bằng chứng đầu tiên về các đoạn DNA có thể di chuyển được phát hiện từ các thí nghiệm lai giống ở cây ngô được nhà nữ di truyền học người Mỹ là Barbara McClintock tiến hành vào những năm 1940 và 1950 **(Hình 21.8)**. Khi theo dõi các cây ngô qua nhiều thế hệ, McClintock xác định được sự thay đổi màu nội nhũ của các hạt ngô chỉ có thể giải thích được nếu như có sự tồn tại của các yếu tố di truyền có thể di chuyển từ những vị trí khác trong hệ gene vào trong các gene quy định tính trạng màu nội nhũ, làm phá vỡ những gene này và dẫn đến hiện tượng màu nội nhũ thay đổi. Phát hiện của McClintock ban đầu được đón nhận bằng nhiều hoài nghi và thậm chí bị phản đối. Phải mất nhiều năm sau đó, công trình nghiên cứu kỳ công cùng những ý tưởng sâu sắc của McClintock về các yếu tố di động mới được xác nhận bởi các nhà di truyền học vi khuẩn và vi sinh vật khi họ tìm ra cơ sở phân tử của quá trình chuyển vị của những yếu tố này.

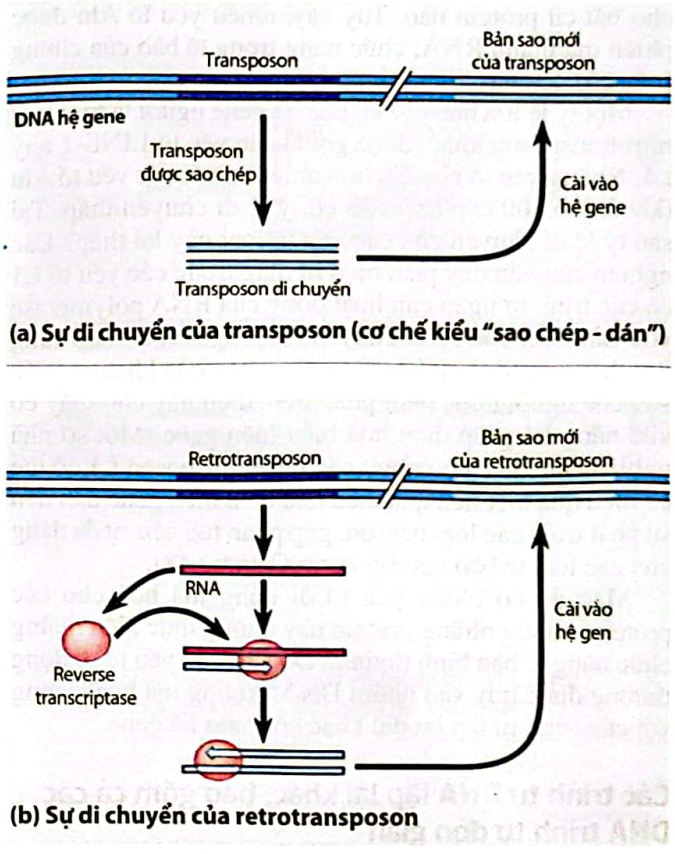
Sự di chuyển của các transposon và retrotransposon

Các sinh vật nhân thực có hai loại yếu tố di động. Loại thứ nhất được gọi là các transposon; loại yếu tố này di chuyển trong hệ gene thông qua một DNA trung gian. Các transposon có thể di chuyển hoặc bởi cơ chế "cắt - dán" và chúng được chuyển dời khỏi vị trí gốc, hoặc bởi cơ chế "sao chép - dán" và chúng để lại một bản sao tại vị trí gốc **(Hình 21.9a)**.

Phần lớn các yếu tố di động trong hệ gene sinh vật nhân thực thuộc loại thứ hai, được gọi là các retrotransposon; loại yếu tố này di chuyển trong hệ gene thông qua một RNA trung gian; đây là bản phiên mã của chính DNA retrostransposon. Các retrotransposon luôn để lại một bản sao tại vị trí đích trong quá trình chuyển vị, do chúng được phiên mã thành RNA trung gian **(Hình 21.9b)**. Trước khi cài vào vị trí đích, phân tử RNA trung gian được phiên mã ngược trở lại thành DNA bởi enzyme phiên mã ngược - reverse transcriptase - do chính retrotransposon mã hoá. Điều này có nghĩa là enzyme phiên mã ngược có thể có trong các tế bào mà chúng không nhất thiết phải bị lây nhiễm bởi retrovirus. (Trong thực tế, các retrovirus, như đã được đề cập ở Chương 19, có thể đã tiến hoá bắt nguồn từ chính các retrotransposon.) Hoạt động cài trình tự DNA được phiên mã ngược vào vị trí mới được xúc tác bởi enzyme trong tế bào.



▲ **Hình 21.8 Ảnh hưởng của các yếu tố vận động đến màu hạt ngô.** Barbara McClintock là người đầu tiên đưa ra ý tưởng về những yếu tố di truyền có khả năng di chuyển khi quan sát hiện tượng có nhiều đốm màu khác nhau trong nhân của các hạt ngô. Tuy ban đầu ý tưởng của bà vào những năm 1940 được đón nhận bởi những mối hoài nghi, nhưng sau này đã được kiểm chứng là hoàn toàn xác thực. Bà được nhận giải Nobel năm 1983 khi ở tuổi 81 nhờ công trình mang tính tiên phong của mình.



▲ **Hình 21.9 Sự di chuyển của các yếu tố di động ở sinh vật nhân thực. (a)** Sự di chuyển của các transposon hoặc theo cơ chế "cắt - dán" hoặc theo cơ chế "sao chép - dán" (được minh hoạ ở đây) liên quan đến một phân tử DNA sợi kép trung gian sau đó được cài vào hệ gene. **(b)** Sự di chuyển của các retrotransposon bắt đầu bằng sự hình thành một phân tử RNA mạch đơn trung gian. Các bước còn lại về bản chất giống với một phần chu kỳ sinh sản của retrovirus (xem Hình 19.8). Trong kiểu di chuyển của các transposon theo kiểu "sao chép - dán" và kiểu di chuyển của retrotransposon, trình tự DNA vừa được duy trì ở vị trí gốc vừa xuất hiện ở vị trí mới.

? *Phần (a) ở trên sẽ khác như thế nào nếu cơ chế được minh hoạ ở đây là cơ chế di chuyển kiểu "cắt - dán"?*

Các trình tự liên quan đến các yếu tố di động

Ở sinh vật nhân thực, nhiều bản sao của các yếu tố di động và các trình tự liên quan đến chúng nằm rải rác khắp hệ gene. Mỗi đơn vị riêng lẻ của yếu tố di động thường dài từ vài trăm đến vài nghìn cặp base, và các "bản sao" nằm phân tán thường giống nhau, nhưng không giống hệt nhau. Một số yếu tố di động như vậy có khả năng di chuyển; các enzyme cần thiết cho sự di chuyển của nó có thể được mã hoá bởi một yếu tố di động bất kỳ, bao gồm cả chính yếu tố di động đang hoạt động. Những trình tự khác là những trình tự có liên quan nhưng đã mất hoàn toàn khả năng di chuyển. Các yếu tố di động và các trình tự có liên quan chiếm khoảng 25% - 50% hệ gene ở phần lớn động vật có vú (xem Hình 21.7); tỷ lệ này thậm chí còn cao hơn ở các loài lưỡng cư và nhiều loài thực vật.

Ở người và nhiều loài linh trưởng khác, một tỷ lệ lớn các trình tự DNA liên quan đến các yếu tố di động bao gồm một họ các trình tự giống nhau được gọi là các *yếu tố Alu*. Riêng những trình tự này đã chiếm khoảng 10% hệ gene người. Các yếu tố *Alu* có chiều dài khoảng 300 nucleotide, tức là ngắn hơn nhiều so với phần lớn các yếu tố di động còn hoạt động khác, và chúng không mã hoá cho bất cứ protein nào. Tuy vậy, nhiều yếu tố Alu được phiên mã thành RNA; chức năng trong tế bào của chúng (nếu có) đến nay chưa rõ.

Một tỷ lệ lớn hơn (17%) của hệ gene người là một loại retrotransposon khác, được gọi là các yếu tố LINE-1 hay L1. Những yếu tố này dài hơn nhiều so với các yếu tố Alu (khoảng 6.500 cặp base) và có tỷ lệ di chuyển thấp. Tại sao tỷ lệ di chuyển của các yếu tố loại này lại thấp? Các nghiên cứu gần đây phát hiện ra rằng trong các yếu tố L1 có các trình tự ngăn cản hoạt động của RNA polymerase vốn cần thiết cho sự di chuyển. Một nghiên cứu bổ sung tìm thấy các trình tự L1 có trong intron của khoảng 80% số gene người được đem phân tích, điều này cho thấy có khả năng L1 giúp điều hoà biểu hiện gene. Một số nhà nghiên cứu khác cho rằng: các retrotransposon L1 có thể có hiệu quả biệt hoá qua điều hoà biểu hiện gene dẫn đến sự phát triển các loại neuron, góp phần tạo nên sự đa dạng của các loại tế bào neuron (xem Chương 48).

Mặc dù có nhiều yếu tố di động mã hoá cho các protein, nhưng những protein này không thực hiện những chức năng tế bào bình thường. Do vậy, các yếu tố di động thường được quy vào nhóm DNA "không mã hoá", cùng với các trình tự lặp lại dài khác có trong hệ gene.

Các trình tự DNA lặp lại khác, bao gồm cả các DNA trình tự đơn giản

Các trình tự DNA lặp lại vốn không liên quan đến các yếu tố di động có vẻ xuất hiện do các sai sót trong các quá trình sao chép hoặc tái tổ hợp của DNA. Những trình tự DNA như vậy chiếm khoảng 15% hệ gene người (xem Hình 21.7). Khoảng một phần ba trong số này (tức là khoảng 5 - 6% hệ gene người) là những đoạn DNA dài lặp lại hai lần với mỗi đơn vị lặp lại dài từ 10.000 đến 300.000 cặp base. Các đoạn DNA dài như vậy dường như đã được sao chép từ vị trí này sang vị trí khác thuộc cùng một nhiễm sắc thể hoặc thuộc hai nhiễm sắc thể khác nhau.

Không giống như các bản sao của các trình tự DNA dài phân tán khắp hệ gene, các DNA trình tự đơn giản thường gồm nhiều bản sao của các đoạn trình tự ngắn lặp lại liên tiếp như ví dụ được minh hoạ dưới đây (ở đây, chỉ minh hoạ một mạch):

...GTTACGTTACGTTACGTTACGTTACGTTAC...

Trong trường hợp này, đơn vị lặp lại (GTTAC) gồm 5 nucleotide. Trong thực tế, các đơn vị lặp lại như vậy có thể dài đến 500 nucleotide, nhưng thường thì ngắn hơn 15 nucleotide như ví dụ trên đây. Khi đơn vị lặp lại chỉ chứa từ 2 đến 5 nucleotide, thì đoạn trình tự lặp lại liên tiếp như vậy được gọi là **trình tự ngắn lặp lại liên tiếp**, hay còn gọi là **STR** (short tandem repeat). Chúng ta đã nói về việc sử dụng chỉ thị STR trong xây dựng tàng thư di truyền ở Chương 20. Số bản sao của cùng một đơn vị lặp lại có thể khác nhau ở những vị trí khác nhau trong hệ gene. Chẳng hạn như, đơn vị lặp lại GTTAC có thể xuất hiện liên tiếp hàng trăm nghìn lần tại một vị trí trong hệ gen; nhưng ở một vị trí khác, số lần lặp lại của đơn vị này chỉ bằng một nửa. Số lần lặp lại cũng rất khác nhau giữa người này với người khác, tạo nên sự khác biệt trong tàng thư di truyền của mỗi cá nhân trên cơ sở phân tích các trình tự STR. Tính tổng cộng, các DNA trình tự đơn giản chiếm khoảng 3% hệ gene người.

Thành phần nucleotide của các đoạn DNA trình tự đơn giản khác biệt với thành phần của các đoạn trình tự DNA khác trong hệ gene đến mức chúng tạo nên sự khác biệt về tỷ trọng. Nếu DNA hệ gene được cắt thành các đoạn nhỏ, rồi được ly tâm ở tốc độ cao, thì các đoạn DNA có tỷ trọng khác nhau sẽ "định vị" ở những vị trí khác nhau trong ống ly tâm. Các đoạn DNA lặp lại vốn ban đầu được phân lập theo cách này được gọi là các trình tự *DNA vệ tinh* bởi vì các băng ly tâm của chúng tách biệt khỏi phần băng ly tâm chung gồm các trình tự DNA còn lại của hệ gene giống như một "vệ tinh". Thuật ngữ "DNA vệ tinh" và DNA trình tự đơn giản hiện nay thường được dùng thay thế cho nhau.

Một lượng lớn DNA trình tự đơn giản của hệ gene tập trung ở các đầu mút và tâm động của nhiễm sắc thể, cho thấy những trình tự DNA này giữ vai trò cấu trúc nhiễm sắc thể. Các trình tự DNA tại tâm động là thiết yếu cho hoạt động phân ly của các nhiễm sắc tử trong quá trình phân bào (xem Chương 12). Trình tự DNA tâm động, cùng với các DNA trình tự đơn giản khác, có thể đóng vai trò tổ chức chất nhiễm sắc trong nhân tại kỳ trung gian của chu trình tế bào. Các DNA trình tự đơn giản tại các đầu mút nhiễm sắc thể giúp bảo vệ các gene không bị mất do DNA ngắn lại sau mỗi lần sao chép (xem Chương 16). DNA đầu mút đồng thời liên kết với các protein giúp bảo vệ đầu mút nhiễm sắc thể khỏi bị biến tính, đồng thời không bị dính chập với các nhiễm sắc thể khác.

Các gene và các họ đa gene

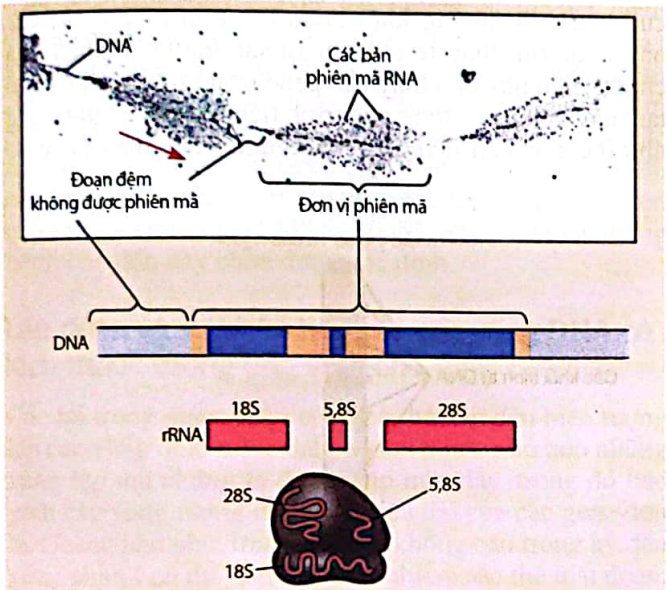
Chúng ta kết thúc bàn luận về các loại trình tự DNA khác nhau trong các hệ gene sinh vật nhân thực bằng việc xem xét các gene một cách chi tiết hơn. Chúng ta nhớ lại rằng tổng cộng các trình tự DNA mã hoá hoặc cho các protein hoặc cho các loại tRNA và rRNA chỉ chiếm có 1,5% hệ gene người (xem Hình 21.7). Nếu chúng ta tính cả các trình tự intron và các trình tự điều hoà liên quan đến gene, thì tổng cộng tất cả các trình tự DNA có liên quan đến gene (bao gồm cả những đoạn mã hoá và không mã hoá)

chiếm khoảng 25% hệ gene người. Nói cách khác, trung bình chỉ có khoảng 6% (tức là 1,5% của 25%) trình tự đầy đủ của một gene có mặt trong sản phẩm cuối cùng của gene.

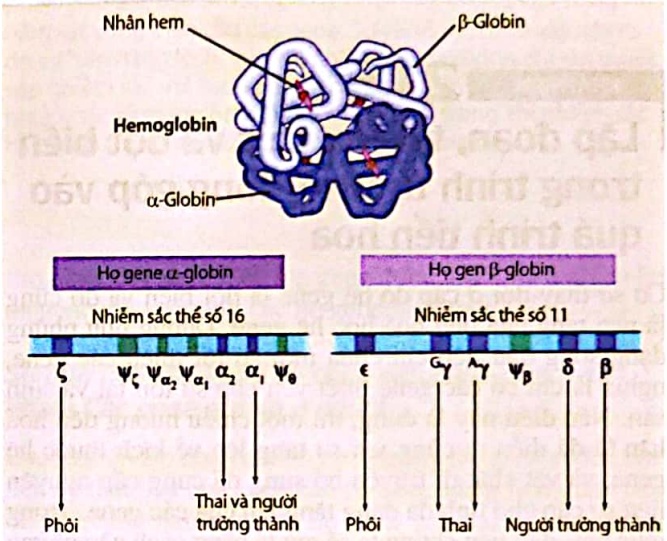
Giống với các gene của vi khuẩn, nhiều gene ở sinh vật nhân thực là những trình tự đơn nhất và chỉ có một bản sao duy nhất trong mỗi bộ nhiễm sắc thể đơn bội. Tuy vậy, trong hệ gene người và hệ gene của nhiều động vật và thực vật khác, những gene "đơn độc" như vậy chiếm ít hơn một nửa tổng số trình tự DNA được phiên mã. Các gene còn lại xuất hiện thành các **họ đa gene**, tức là tập hợp của hai hay nhiều gene giống hệt hoặc rất giống nhau.

Trong các họ đa gene gồm các trình tự DNA *giống hệt nhau*, các trình tự DNA lặp lại liền kề nhau, và ngoại trừ các gene mã hoá protein histone, chúng mã hoá cho sản phẩm cuối cùng là RNA. Một ví dụ về họ các trình tự DNA giống hệt nhau là cụm các gene mã hoá cho ba loại phân tử rRNA lớn nhất **(Hình 21.10a)**. Những phân tử rRNA này được phiên mã thành các bản phiên mã duy nhất gồm hàng trăm thậm chí hàng nghìn lần lặp lại kế tiếp nhau và tập hợp thành một hoặc một số cụm trong hệ gene sinh vật nhân thực. Với nhiều bản sao cùng có mặt trong một đơn vị phiên mã như vậy, tế bào có thể nhanh chóng tạo ra hàng triệu ribosome cần cho quá trình tổng hợp protein. Bản phiên mã sơ cấp của các gene rRNA sau đó được cắt xén để hình thành nên ba loại phân tử rRNA. Những phân tử rRNA này sau đó được kết hợp với các protein và một loại rRNA khác (rRNA 5S) để tạo nên các tiểu phần ribosome.

Các ví dụ kinh điển về các họ đa gene có trình tự *không giống hệt nhau* gồm hai họ gene có quan hệ với nhau mã hoá cho globin; đây là một nhóm các protein gồm các tiểu phần (chuỗi polypeptide) α và β của hemoglobin. Có một họ gene nằm trên nhiễm sắc thể số 16 ở người mã hoá cho các dạng khác nhau của α-globin; một họ gene còn lại nằm trên nhiễm sắc thể số 11 mã hoá cho các dạng khác nhau của β-globin **(Hình 21.10b)**. Các dạng khác nhau của mỗi tiểu phần globin được biểu hiện vào các thời điểm khác nhau của quá trình phát triển, qua đó giúp hemoglobin biểu hiện chức năng hiệu quả trong các điều kiện môi trường thay đổi trong quá trình phát triển ở động vật. Chẳng hạn như, ở người, các dạng hemoglobin có trong phôi và thai có ái lực với oxygen cao hơn so với dạng hemoglobin ở người trưởng thành; điều này giúp đảm bảo hiệu quả vận chuyển oxygen từ mẹ sang thai nhi. Trong các cụm họ gene mã hoá globin, người ta còn tìm thấy một số gene giả.



(a) Một phần họ gene mã hoá RNA ribosome. Ba trong số hàng trăm bản sao của các đơn vị phiên mã rRNA trong hệ gene của loài kỳ giông được minh hoạ ở phần trên (ảnh TEM). Mỗi một "chiếc lông" tương ứng với một đơn vị phiên mã với khoảng 100 phân tử đang được tổng hợp bởi RNA polymerase (điểm màu sẫm dọc theo sợi DNA) dịch chuyển từ trái qua phải. Các bản phiên mã RNA đang được "mở rộng" từ DNA. Sơ đồ bên dưới ảnh TEM mô tả một đơn vị phiên mã. Nó bao gồm các gene (màu xanh lam) mã hoá ba loại rRNA xen giữa các vùng được phiên mã nhưng sau đó được cắt bỏ (màu vàng). Ban đầu chỉ một bản phiên mã RNA duy nhất được tạo ra, nhưng sau đó nó được cắt xén để tạo nên ba phân tử rRNA khác nhau (mỗi loại một phân tử); chúng là các thành phần thiết yếu của ribosome. Một loại rRNA thứ tư (5S rRNA) cũng là thành phần của ribosome, nhưng gene mã hoá nó không thuộc cùng đơn vị phiên mã này.



(b) Các họ gene α-globin và β-globin ở người. Hemoglobin được cấu tạo từ hai tiểu phần (chuỗi) polypeptide loại α-globin và hai tiểu phần loại β-globin. Các gene (màu xanh lam) mã hoá cho α-globin và β-globin được tìm thấy trong hai họ gene có cấu trúc tổ chức như minh hoạ trên hình. Các trình tự DNA không mã hoá xen giữa các gene chức năng trong mỗi họ gene gồm các gene giả (màu xanh lục) và các dạng biến đổi không biểu hiện chức năng của các gene chức năng bình thường. Tên gọi các gene và các gene giả được ký hiệu và đọc theo tiếng Hy Lạp.

▲ Hình 21.10 Các họ gene.

? *Trong phần (a) của hình trên, bằng cách nào bạn có thể xác định được chiều phiên mã, nếu như không có mũi tên màu đỏ?*

Sự sắp xếp các gene thành các họ gene đã giúp các nhà sinh học có những hiểu biết sâu hơn về quá trình tiến hoá của các hệ gene. Trong mục tiếp theo, chúng ta sẽ đề cập đến một số quá trình dẫn đến sự định hình các hệ gene của các loài khác nhau qua quá trình tiến hoá.

KIỂM TRA KHÁI NIỆM 21.4

1. Hãy nêu các đặc điểm của hệ gene động vật có vú làm chúng trở nên lớn hơn so với các hệ gene sinh vật nhân sơ?
2. Các intron, các yếu tố vận động và các trình tự DNA lặp lại đơn giản phân bố trong hệ gene khác nhau như thế nào?
3. Nêu sự khác nhau trong cấu trúc của các họ gene mã hoá rRNA và mã hoá các protein globin ở người. Với mỗi họ gene, hãy giải thích lợi thế của sự tồn tại cấu trúc kiểu họ gene đối với sinh vật.
4. **ĐIỀU GÌ NẾU?** Giả sử bạn tìm thấy một trình tự DNA giống với trình tự của một gene đã biết, nhưng chúng lại khác nhau rõ rệt ở một vài nucleotide nhất định. Bằng cách nào bạn có thể xác định trình tự mới tìm thấy có phải là một "gen" biểu hiện chức năng hay không?

Câu trả lời có trong Phụ lục A.

KHÁI NIỆM 21.5

Lặp đoạn, tái cấu trúc và đột biến trong trình tự DNA đóng góp vào quá trình tiến hoá

Cơ sở thay đổi ở cấp độ hệ gene là đột biến và đó cũng là nền tảng của tiến hoá học hệ gene. Dường như những dạng sống đầu tiên chỉ chứa một số tối thiểu các gene, nghĩa là chỉ có các gene thiết yếu cho sự tồn tại và sinh sản. Nếu điều này là đúng, thì một chiều hướng tiến hoá hẳn là đã diễn ra cùng với sự tăng lên về kích thước hệ gene, và vật chất di truyền bổ sung đã cung cấp nguyên liệu sơ cấp cho tính đa dạng tăng lên của các gene. Trong mục này, đầu tiên chúng ta sẽ mô tả bằng cách nào những bản sao của toàn bộ hay một phần của hệ gene có thể xuất hiện, rồi sau đó đề cập những quá trình xảy ra tiếp theo dẫn sự tiến hoá của các protein (hoặc các sản phẩm RNA) có chức năng hoàn toàn mới hoặc thay đổi chút ít.

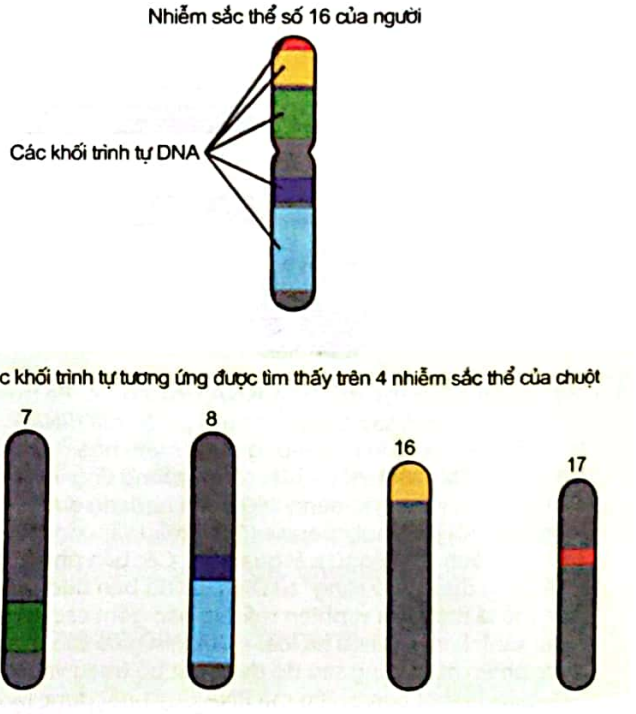
Sự lặp lại cả bộ nhiễm sắc thể

Các sự kiện ngẫu nhiên trong giảm phân có thể dẫn đến tế bào có thêm một hoặc nhiều bộ nhiễm sắc thể; hiện tượng này được gọi là đa bội thể. Mặc dù, trong phần lớn trường hợp những sự kiện ngẫu nhiên đó thường gây chết, nhưng trong một số hiếm trường hợp, chúng lại thúc đẩy sự tiến hoá của các gene. Ở một cơ thể đa bội, một bộ các gene có thể cung cấp đủ các chức năng thiết yếu cho cơ thể đó. Những gene ở những bộ nhiễm sắc thể bổ sung có thể phân hoá trong quá trình tích luỹ các đột biến; những biến dị này có thể được duy trì nếu như cơ thể mang các đột biến sống sót và sinh sản được. Bằng cách đó, các gene có thể tiến hoá với những chức năng mới. Cùng với việc một bản sao của gene thiết yếu được biểu hiện, sự phân hoá của một bản sao khác có thể dẫn đến một loại protein vẫn do gene đó mã hoá song hoạt động theo một cách mới, qua đó làm thay đổi kiểu hình của sinh vật. Kết quả của sự tích luỹ các đột biến này có thể dẫn đến sự phân nhánh tiến hoá của một loài mới, giống như biểu hiện thường thấy ở thực vật (xem Chương 24). Các động vật đa bội cũng tồn tại, song rất hiếm.

Sự thay đổi cấu trúc nhiễm sắc thể

Từ lâu các nhà khoa học đã biết rằng vào một thời điểm nào đó trong vòng 6 triệu năm trước khi các dạng tổ tiên của người hiện đại và tinh tinh phân ly khỏi nhau và hình thành nên các loài riêng biệt, một sự dung hợp hai nhiễm sắc thể khác nhau vốn có ở dạng tổ tiên đã dẫn đến loài người có số nhiễm sắc thể đơn bội (n = 23) khác với của tinh tinh (n = 24). Với sự bùng nổ thông tin về trình tự các hệ gene, giờ đây chúng ta có thể so sánh cấu trúc và tổ chức nhiễm sắc thể giữa nhiều loài ở cấp độ chi tiết hơn. Những thông tin này giúp chúng ta có thể tìm hiểu sâu hơn về các quá trình tiến hoá đã dẫn đến sự hình thành các nhiễm sắc thể cũng như sự phát sinh các loài.

Ví dụ, trong một nghiên cứu, các nhà khoa học đã tiến hành so sánh trình tự DNA giữa mỗi nhiễm sắc thể của người với trình tự toàn bộ hệ gene của chuột. **Hình 21.11** cho thấy kết quả so sánh với nhiễm sắc thể số 16 của người là: những khối gene lớn trên nhiễm sắc thể này được tìm thấy trên 4 nhiễm sắc thể khác nhau của chuột; điều này cho thấy các gene trong mỗi khối đã tồn tại cùng với nhau trong quá trình tiến hoá của chuột cũng như ở các nhánh tiến hoá của người. Thực hiện phép so



▲ Hình 21.11 Các khối trình tự giống nhau trên các nhiễm sắc thể của người và chuột. Các trình tự DNA rất giống nhau được tìm thấy trong một khối trình tự lớn thuộc nhiễm sắc thể số 16 của người được tìm thấy trên các nhiễm sắc thể số 7, 8, 16 và 17 của chuột. Điều này cho thấy các trình tự DNA trong mỗi khối đã luôn tồn tại cùng nhau ở các dòng tiến hoá dẫn đến sự hình thành người và chuột kể từ thời điểm chúng phân ly khỏi nhau từ tổ tiên chung.

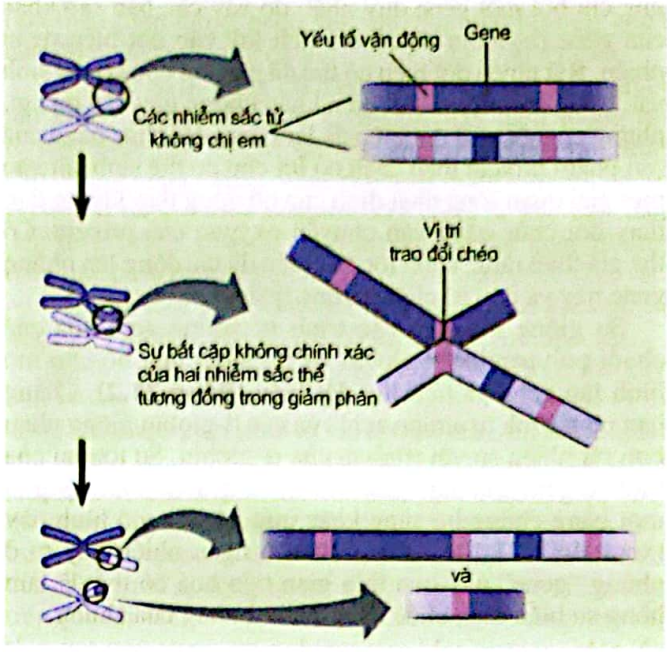
sánh tương tự giữa nhiễm sắc thể của người với sáu loài động vật có vú khác cũng đã giúp các nhà nghiên cứu tái thiết được lịch sử tiến hoá tổ chức nhiễm sắc thể ở tám loài động vật có vú này. Qua đó, các nhà nghiên cứu đã tìm ra nhiều lặp đoạn và đảo đoạn trên các đoạn lớn của NST là kết quả của các lỗi tái tổ hợp xảy ra trong giảm phân dẫn đến sự đứt gãy và nối lại không chính xác của DNA. Tần số suất hiện những sự kiện này dường như đã tăng nhanh trong khoảng 100 triệu năm trước, tức là khoảng thời gian những loài khủng long kích thước lớn trở nên tuyệt chủng và số loài động vật có vú tăng lên nhanh chóng. Sự trùng lặp ngẫu nhiên này rõ ràng là rất thú vị bởi vì sự tái cấu trúc nhiễm sắc thể được cho là đã đóng góp vào sự hình thành các loài mới. Mặc dù hai cá thể mang các nhiễm sắc thể được sắp xếp khác nhau vẫn có thể giao phối với nhau và sinh sản, nhưng các cá thể con sinh ra sẽ có hai bộ nhiễm sắc thể không tương đồng. Vì vậy, sự tái cấu trúc các nhiễm sắc thể có thể dẫn đến sự hình thành hai quần thể không còn có khả năng giao phối với nhau nữa, và nó trở thành một bước trong con đường dẫn đến sự hình thành hai loài tách biệt (chúng ta sẽ đề cập kỹ hơn về vấn đề này ở Chương 24).

Điều gây ngạc nhiên một chút là những nghiên cứu tương tự đã phát hiện ra những mối liên quan đến y học. Việc phân tích các điểm đứt gãy nhiễm sắc thể liên quan đến sự tái sắp xếp của chúng cho thấy những điểm này không hề phân bố ngẫu nhiên, mà chúng là những điểm đặc thù được dùng đi dùng lại nhiều lần. Nhiều "điểm nóng" tái tổ hợp như vậy tương ứng với vị trí tái cấu trúc nhiễm sắc thể trong hệ gene người vốn có liên quan đến các bệnh bẩm sinh. Tất nhiên, các nhà nghiên cứu còn quan tâm cả những vị trí khác có thể liên quan đến những bệnh cho đến nay chưa được xác định.

Lặp đoạn và sự phân hoá của các vùng DNA có kích thước tương ứng với gene

Các lỗi trong giảm phân cũng có thể dẫn đến hiện tượng lặp các vùng nhiễm sắc thể có kích thước nhỏ hơn những vùng lặp mà chúng ta đã đề cập trên đây, trong đó bao gồm các vùng tương ứng với chiều dài của các gene đơn lẻ. Chẳng hạn như, trao đổi chéo không cân trong kỳ đầu giảm phân I có thể dẫn đến một nhiễm sắc thể mất đoạn, trong khi một nhiễm sắc thể khác lặp đoạn. Như minh hoạ trên **Hình 21.12**, các yếu tố di động trong hệ gene là những vị trí mà các nhiễm sắc tử không chị em có thể trao đổi chéo với nhau, thậm chí ngay cả khi chúng không có những trình tự tương đồng xếp thẳng hàng chính xác với nhau.

Ngoài ra, hiện tượng "trượt" có thể xảy ra trong sao chép DNA, chẳng hạn như mạch làm khuôn xê dịch so với mạch tương đồng mới được tổng hợp, hoặc một phần của mạch làm khuôn bị bộ máy sao chép bỏ qua hay trong trường hợp khác nó được dùng làm khuôn hai lần. Kết quả là một đoạn DNA bị mất đi hoặc lặp lại. Có thể dễ dàng tưởng tượng ra cách mà những lỗi như vậy có thể xuất hiện trong các vùng trình tự lặp lại giống như các trình tự DNA lặp lại đơn giản đã được mô tả ở trên. Các trình tự DNA lặp lại đơn giản với số lượng biến động tại một vị trí nhất định, vốn được dùng cho phân tích STR, có thể là do những lỗi giống như vậy. Các bằng chứng về trao đổi chéo không cân và hiện tượng "trượt" của mạch khuôn trong sao chép DNA dẫn đến lặp gene được tìm thấy ở nhiều họ đa gene tồn tại trong các hệ gene hiện nay.



▲ **Hình 21.12 Lặp gene do trao đổi chéo không cân.** Một cơ chế mà qua đó một gene (hoặc một đoạn DNA khác) có thể bị lặp lại (nhân đôi) là sự tái tổ hợp xảy ra trong quá trình giảm phân giữa các bản sao khác nhau của một yếu tố vận động nằm sát vùng biên của các gene. Sự tái tổ hợp như vậy xảy ra do sự "sắp hàng lệch" của hai nhiễm sắc tử không chị em thuộc cặp nhiễm sắc thể tương đồng dẫn đến sự hình thành một nhiễm sắc tử mang hai bản sao của gene, trong khi nhiễm sắc tử còn lại thì không có bản sao nào của gene đó.

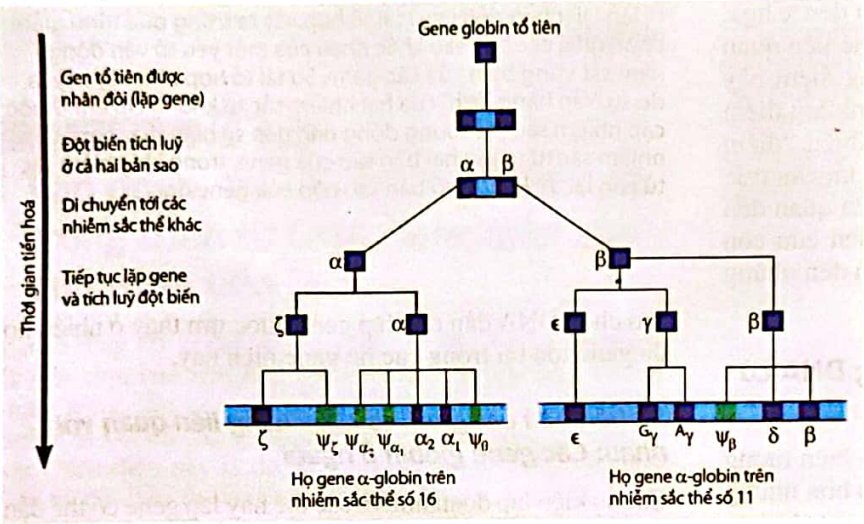
Sự tiến hoá các gene có chức năng liên quan với nhau: Các gene globin ở người

Các sự kiện lặp đoạn nhiễm sắc thể hay lặp gene có thể dẫn đến sự tiến hoá của các gene có chức năng liên quan đến nhau, chẳng hạn như các họ gene mã hoá cho α-globin và β-globin (xem Hình 21.10b). Việc so sánh các trình tự gene trong một họ đa gene có thể chỉ ra thứ tự các gene xuất hiện. Cách tiếp cận để tái tạo lại lịch sử tiến hoá của các gene mã hoá globin đã chỉ ra rằng tất cả những gene này đều có nguồn gốc từ một gene globin tổ tiên chung; gene tổ tiên này đã trải qua hiện tượng lặp gene rồi phân hoá thành các gene α-globin và β-globin tổ tiên khoảng 450 - 500 triệu năm trước (**Hình 21.13**, ở trang sau). Mỗi gene tổ tiên này sau đó tiếp tục được lặp lại một vài lần, rồi những bản sao của chúng phân hoá về trình tự, dẫn đến hình thành các gene thành viên thuộc họ gene như hiện nay. Trong thực tế, gene globin tổ tiên chung cũng có thể là nguồn gốc của gene mã hoá protein cơ liên kết oxygen có tên gọi là myoglobin và protein ở thực vật là leghemoglobin. Hai loại protein này hoạt động ở dạng đơn phân, và các gene của chúng thuộc "siêu họ globin".

Tiếp theo sau các sự kiện lặp gene, sự khác biệt giữa các gene trong các họ globin rõ ràng xuất phát từ các đột biến được tích luỹ trong các bản sao của gene qua nhiều thế hệ. Ví dụ, một mô hình hiện nay cho rằng chức năng thiết yếu của protein α-globin trước đây có thể được đáp

ứng chỉ bởi một gene duy nhất, do vậy các bản sao khác của gene α-globin đã có thể tích luỹ các đột biến ngẫu nhiên. Rất nhiều đột biến có thể đã gây hại cho cơ thể sinh vật, trong khi một số đột biến khác không gây hậu quả gì, nhưng có một số ít đột biến đã làm thay đổi chức năng của sản phẩm protein theo cách có lợi cho cơ thể sinh vật vào một giai đoạn sống nhất định của nó đồng thời không làm thay đổi chức năng vận chuyển oxygen của protein. Có thể giả thiết rằng: chọn lọc tự nhiên đã tác động lên những gene này và duy trì chúng trong quần thể.

Sự giống nhau về các trình tự amino acid của các chuỗi polypeptide α-globin và β-globin ủng hộ cho mô hình lặp gene và tích luỹ đột biến **(Bảng 21.2)**. Chẳng hạn như, trình tự amino acid của các β-globin giống nhau hơn rất nhiều so với trình tự của α-globin. Sự tồn tại của một số gene giả nằm giữa các gene globin hoạt động là một bằng chứng bổ sung khác ủng hộ cho mô hình này (xem Hình 21.10b): Các đột biến ngẫu nhiên xảy ra ở những "gene" này qua thời gian tiến hoá có thể đã làm hỏng sự biểu hiện chức năng bình thường của chúng.



▲ Hình 21.13 Một mô hình tiến hoá của các họ gene α-globin và β-globin từ gene globin "tổ tiên" duy nhất.

? *Các yếu tố trình tự màu xanh lục là các gene giả. Hãy giải thích bằng cách nào chúng có thể xuất hiện sau khi đã xảy ra các sự kiện lặp gene.*

Bảng 21.2 Tỷ lệ giống nhau trong trình tự amino acid giữa các protein globin ở người

		Các loại α-globin		Các loại β-globin		
		α	ζ	β	γ	α
Các loại α-globin	α	–	58	42	39	37
	ζ	58	–	34	38	37
Các loại β-globin	β	42	34	–	73	75
	γ	39	38	73	–	80
	ε	37	37	75	80	–

Sự tiến hoá của các gene có chức năng mới

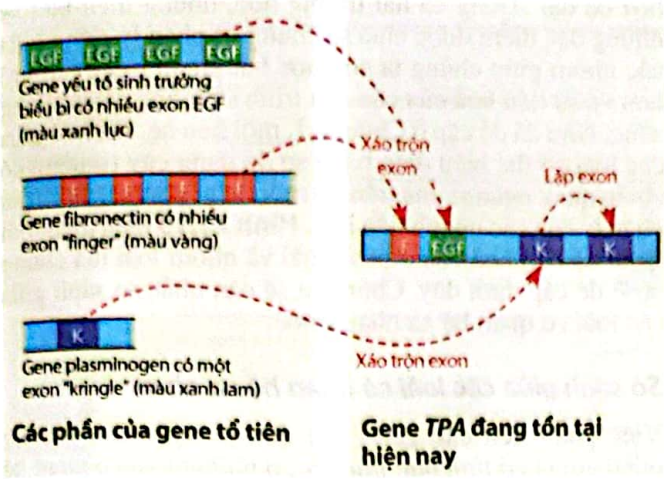
Trong quá trình tiến hoá của các họ gene globin, hiện tượng lặp gene và phân ly sau đó đã tạo nên các gene thành viên mà sản phẩm của chúng đều thực hiện chức năng giống nhau (vận chuyển oxygen). Theo một cách khác, một bản sao của gene lặp có thể trải qua những biến đổi dẫn đến sự xuất hiện một chức năng hoàn toàn mới của sản phẩm protein. Các gene mã hoá lysozyme và α-lactalbumin là một ví dụ như vậy.

Lysozyme là một enzyme giúp bảo vệ cơ thể động vật khỏi sự lây nhiễm của vi khuẩn bằng việc xúc tác thuỷ phân thành tế bào vi khuẩn; α-lactalbumin là một protein không có chức năng enzyme, thay vào đó nó giữ vai trò trong quá trình sản xuất sữa ở động vật có vú. Hai protein này rất giống nhau về trình tự amino acid và cấu trúc không gian ba chiều. Cả hai gene đều được tìm thấy ở các loài động vật có vú, nhưng ở chim chỉ tìm thấy gene mã hoá lysozyme. Điều này chỉ ra rằng vào một thời điểm nhất định nào đó trong quá khứ, sau khi các nhánh dẫn đến hình thành các loài động vật có vú và chim phân ly khỏi nhau, gene lysozyme đã trải qua một sự kiện lặp gene trong nhánh tiến hoá hình thành các động vật có vú, nhưng không xảy ra trong nhánh tiến hoá của chim. Cuối cùng, một bản sao của gene lysozyme đã dẫn đến sự tiến hoá hình thành gene mã hoá α-lactalbumin vốn là một protein có chức năng khác biệt hoàn toàn.

Sự sắp xếp lại các phần của gene: lặp và xáo trộn exon

Sự sắp xếp lại các trình tự DNA sẵn có trong các gene cũng đã góp phần vào sự tiến hoá hệ gene. Sự có mặt của intron trong phần lớn các gene ở sinh vật nhân thực đa bào có thể đã thúc đẩy sự tiến hoá của các protein có tiềm năng hữu dụng mới bằng việc gia tăng hiện tượng lặp exon hay sắp xếp lại vị trí của các exon trong hệ gene. Chúng ta nhớ lại từ Chương 17 rằng mỗi exon thường mã hoá cho một miền có cấu trúc và chức năng đặc thù của protein.

Chúng ta cũng đã biết trao đổi chéo không cân trong quá trình giảm phân có thể dẫn đến hiện tượng lặp gene trên một nhiễm sắc thể đồng thời làm mất gene trên nhiễm sắc thể tương đồng với nó (xem Hình 21.12). Bằng một quá trình tương tự, một exon nhất định trong gene có thể lặp lại trên một nhiễm sắc thể, song lại bị mất đi trên nhiễm sắc thể kia. Các gene mang các exon lặp lại có thể mã hoá cho một loại protein chứa hai bản sao của một miền protein. Sự thay đổi này trong cấu trúc có thể làm tăng cường sự biểu hiện chức năng của protein nếu protein đó lúc này trở nên ổn định hơn, và tăng khả năng liên kết với một chất gắn nhất định hoặc làm thay đổi một số thuộc tính khác. Khá nhiều gene mã hoá protein có nhiều bản sao của các exon có quan hệ với nhau



▲ Hình 21.14 Sự tiến hoá của một gene mới bằng cơ chế xáo trộn exon. Sự xáo trộn exon có thể gồm sự di chuyển exon từ các dạng tiền thân của gene mã hoá yếu tố sinh trưởng biểu bì, của fibronectin và của plasminogen (bên trái) vào gene mã hoá yếu tố hoạt hoá plasminogen mô - *TPA* (bên phải). Thứ tự xảy ra các sự kiện là chưa rõ. Sự nhân đôi của exon "kringle" từ gene plasminogen khi nó di chuyển giải thích cho sự xuất hiện hai bản sao của exon này trong gene *TPA*. Mỗi loại exon mã hoá cho một miền đặc thù của protein TPA.

? *Bằng cách nào sự có mặt của các yếu tố vận động có trong các intron lại có thể thúc đẩy sự xáo trộn exon diễn ra như được mô tả trên đây ?*

mà có thể giả thiết chúng hình thành sau một quá trình lặp exon và phân hoá exon. Một ví dụ điển hình về điều này là gene mã hoá protein mạng ngoại bào collagen. Collagen là một protein cấu trúc có trình tự amino acid với mức độ lặp lại cao phản ánh sự lặp lại của các exon trong gene collagen.

Theo một cách khác, chúng ta cũng có thể tưởng tượng sự kết cặp và đôi khi phối trộn giữa các exon khác nhau của cùng một gene hoặc giữa hai gene không allele với nhau do các lỗi tái tổ hợp xảy ra trong quá trình giảm phân. Quá trình này, được gọi là sự *xáo trộn exon*, có thể dẫn đến sự hình thành những protein mới với những tổ hợp chức năng mới. Hãy xem ví dụ về gene mã hoá chất hoạt hoá plasminogen mô (TPA, *tissue plasminogen activator*). Protein TPA là một loại protein ngoại bào giúp điều khiển sự hình thành huyết khối (trong quá trình đông máu). Protein này gồm có 4 miền chức năng thuộc 3 loại khác nhau; mỗi miền được mã hoá bởi một exon, trong đó có một exon xuất hiện với hai bản sao. Do mỗi loại exon này cũng được tìm thấy ở những protein khác nữa, nên người ta cho rằng gene mã hoá TPA đã hình thành sau một số sự kiện lặp và xáo trộn exon **(Hình 21.14)**.

Các yếu tố di động góp phần vào sự tiến hoá của hệ gene như thế nào?

Sự có mặt ổn định của các yếu tố di động vốn chiếm một phần lớn hệ gene ở một số sinh vật nhân thực phù hợp với ý tưởng cho rằng chúng giữ một vai trò quan trọng trong quá trình tiến hoá hệ gene của những sinh vật này. Những yếu tố này có thể góp phần vào sự tiến hoá của hệ gene theo một số cách. Chúng có thể thúc đẩy các hiện tượng tái tổ hợp, làm đứt gãy các gene hoặc các trình tự điều hoà biểu hiện gene, hoặc vận chuyển toàn bộ một gene nào đó hoặc các vùng exon riêng lẻ tới các vị trí mới.

Các yếu tố di động có trình tự giống nhau nằm phân tán khắp hệ gene là điều kiện thúc đẩy hiện tượng tái tổ hợp giữa các nhiễm sắc thể khác nhau bởi nó cung cấp những vùng tương đồng cho hoạt động trao đổi chéo. Phần lớn những thay đổi như vậy có lẽ là gây hại, dẫn đến hiện tượng chuyển đoạn nhiễm sắc thể hoặc những thay đổi khác trong hệ gene vốn có thể gây chết sinh vật. Nhưng qua thời gian tiến hoá lâu dài, một sự kiện tái tổ hợp ngẫu nhiên cũng có thể có lợi cho cơ thể sinh vật.

Sự di chuyển của các yếu tố di động cũng có thể gây nên những hậu quả trực tiếp. Ví dụ, nếu một yếu tố di động nhảy vào giữa trình tự mã hoá protein, thì nó sẽ ngăn cản tế bào sản xuất bản phiên mã bình thường của gene. Nếu một yếu tố di động cài vào giữa một trình tự điều hoà, thì sự di chuyển đó có thể dẫn đến việc sinh tổng hợp một hoặc một số protein tăng lên hoặc giảm đi. Sự di chuyển của các yếu tố di động có thể gây nên cả hai kiểu hiệu ứng trên đối với các gene mã hoá cho các enzyme tổng hợp sắc tố ở hạt ngô trong thí nghiệm của McClintock. Một lần nữa, phần lớn những thay đổi như vậy thường có hại, song trong một thời gian tiến hoá dài thì một số thay đổi đó lại tạo nên ưu thế về khả năng sống sót.

Trong quá trình di chuyển, các yếu tố di động có thể mang theo một gene hoặc một nhóm gene tới một vị trí mới trong hệ gene. Cơ chế này có thể giải thích cho việc các họ gene α-globin và β-globin ở người nằm trên các nhiễm sắc thể khác nhau, cũng như hiện tượng các gene thành viên của một số họ gene khác nằm phân tán khắp hệ gene. Bởi một quá trình tương tự diễn ra lâu dài, một exon từ một gene có thể được cài vào một gene khác bởi cơ chế giống với hiện tượng tráo exon trong tái tổ hợp. Ví dụ như, một exon có thể được cài vào trong một intron của một gene mã hoá protein bởi hoạt động của một yếu tố di động. Nếu exon được cài vào đó được duy trì ở bản phiên mã RNA trong quá trình hoàn thiện RNA, thì protein được tổng hợp ra sẽ có thêm một miền mới; điều này có thể dẫn đến một chức năng mới của protein.

Một nghiên cứu gần đây còn chỉ ra một cách khác mà các yếu tố di động có thể tạo nên các trình tự mã hoá mới. Nghiên cứu này cho thấy một yếu tố *Alu* có thể "nhảy" vào trong một intron theo cách tạo nên một vị trí cắt intron mới hoạt động yếu trên bản phiên mã RNA. Trong quá trình hoàn thiện bản phiên mã, các vị trí cắt intron bình thường được dùng thường xuyên hơn, nhưng đôi khi intron lại được cắt ở vị trí mới, dẫn đến hình thành một số bản phiên mã mRNA hoàn thiện chứa cả yếu tố *Alu*; kết quả là yếu tố này mã hoá cho một phần mới của protein. Bằng cách này, một kiểu tổ hợp di truyền mới có thể được "thử nghiệm" trong khi chức năng của sản phẩm gene gốc vẫn tiếp tục được duy trì.

Rõ ràng, tất cả các quá trình được thảo luận trong mục này phổ biến hơn cả là gây hại, thậm chí có thể gây chết, hoặc đơn giản là không gây nên bất cứ hậu quả gì. Tuy vậy, trong một số ít trường hợp, những thay đổi có lợi có thể xuất hiện. Qua nhiều thế hệ, sự đa dạng di truyền thu được sẽ là nguồn nguyên liệu có giá trị cho chọn lọc tự nhiên. Sự đa dạng hoá các gene và sản phẩm của chúng là một nhân tố quan trọng trong quá trình tiến hoá của một loài mới. Vì vậy, sự tích luỹ những thay đổi trong hệ gene của mỗi loài cũng chính là bản ghi chép về lịch sử tiến hoá của nó. Để đọc được bản ghi chép này, chúng ta phải xác định được

những thay đổi diễn ra trong hệ gene. So sánh hệ gene của các loài khác nhau giúp chúng ta thực hiện được điều đó, đồng thời giúp chúng ta hiểu rõ hơn các hệ gene tiến hoá như thế nào. Chúng ta sẽ đề cập những chủ đề này trong mục cuối cùng dưới đây thuộc chương này.

KIỂM TRA KHÁI NIỆM 21.5

1. Hãy nêu ba ví dụ về các lỗi xảy ra trong các quá trình của tế bào có thể dẫn đến hiện tượng lặp đoạn DNA?
2. Giải thích bằng cách nào nhiều exon có thể xuất hiện trong các gene EGF tiền thân và fibronectin được vẽ trên Hình 21.14 (phần bên trái)?
3. Ba cách mà các yếu tố vận động được cho là đã góp phần vào sự tiến hoá của các hệ gene là gì?
4. **ĐIỀU GÌ NẾU?** Năm 2005, các nhà khoa học Iceland công bố tìm thấy một đảo đoạn lớn trên nhiễm sắc thể ở 20% số người Bắc Âu, và họ nhấn mạnh rằng những phụ nữ Iceland mang đảo đoạn này có nhiều con hơn đáng kể so với những người phụ nữ không mang đảo đoạn này. Tần số của đảo đoạn này trong quần thể người Iceland ở các thế hệ tương lai được mong đợi sẽ như thế nào?

Câu trả lời có trong Phụ lục A.

KHÁI NIỆM 21.6

So sánh các trình tự hệ gene cung cấp bằng chứng về các quá trình tiến hoá và phát triển

Một nhà nghiên cứu đã ví giai đoạn phát triển hiện nay của sinh học như Kỷ nguyên Khám phá vào thế kỷ thứ XV sau khi lĩnh vực hàng hải và đóng tàu vận tải nhanh có được hàng loạt các tiến bộ kỹ thuật. Trong vòng 20 năm qua, chúng ta đã chứng kiến nhiều tiến bộ nhanh chóng trong giải trình tự các hệ gene và tập hợp các dữ liệu, cũng như sự phát triển của những kỹ thuật mới cho phép đánh giá hoạt động của các gene trong khắp hệ gene, và các phương pháp tinh vi cho phép tìm hiểu bằng cách nào các gene và sản phẩm của chúng cùng phối hợp hoạt động trong các hệ thống phức tạp. Chúng ta mới ở đầu ngưỡng cửa của một thế giới mới.

Việc so sánh trình tự hệ gene từ các loài khác nhau đã cung cấp nhiều thông tin về lịch sử tiến hoá của sự sống từ giai đoạn cổ đại cho đến gần đây. Tương tự như vậy, các nghiên cứu so sánh về chương trình di truyền đã điều khiển quá trình phát triển phôi ở các loài khác nhau đang bắt đầu làm sáng tỏ các cơ chế tạo nên sự phong phú và đa dạng của các dạng sống hiện nay. Trong mục này, chúng ta sẽ bàn luận về việc chúng ta đã học được gì từ những hướng nghiên cứu này.

So sánh hệ gene

Khi các gene và hệ gene của hai loài càng giống nhau về trình tự, thì chúng càng có quan hệ gần gũi trong lịch sử tiến hoá. Việc so sánh hệ gene của các loài có quan hệ gần gũi giúp làm sáng tỏ nhiều sự kiện tiến hoá trong thời gian gần đây; trong khi đó, việc so sánh hệ gene của các loài có khoảng cách xa hơn giúp chúng ta hiểu về lịch sử tiến hoá cổ đại. Trong cả hai trường hợp, những hiểu biết về những đặc điểm được chia sẻ chung và phân ly riêng giữa các nhóm giúp chúng ta có được bức tranh ngày càng rõ hơn về sự tiến hoá của các quá trình sinh học và các dạng sống. Như đã đề cập ở Chương 1, mối liên hệ tiến hoá giữa các loài có thể biểu diễn bằng sơ đồ dạng cây (thường có chiều quay ngang), mà trên đó mỗi điểm phân cành chỉ sự phân ly của các nhánh tiến hoá. **Hình 21.15** biểu diễn mối quan hệ tiến hoá của một số loài và nhóm loài mà chúng ta sẽ đề cập dưới đây. Chúng ta sẽ cân nhắc so sánh giữa các loài có quan hệ xa nhau trước.

So sánh giữa các loài có quan hệ xa nhau

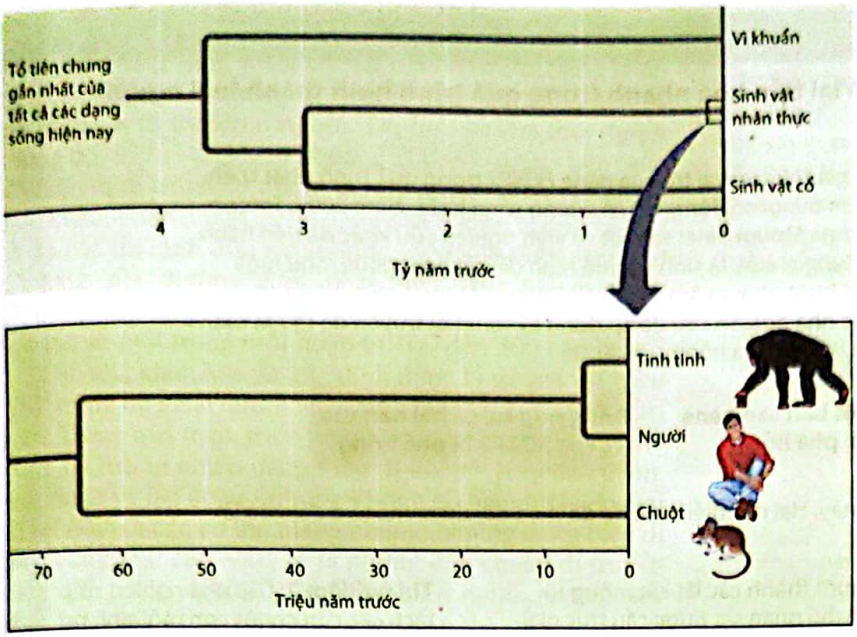
Việc phân tích các gene vẫn còn giống nhau, thường quen gọi là *có tính bảo thủ cao*, ở những loài có quan hệ xa nhau giúp làm sáng tỏ mối quan hệ tiến hoá giữa các loài vốn phân ly khỏi nhau từ một thời điểm rất xa trong quá khứ. Trong thực tế, việc so sánh trình tự hệ gene đầy đủ của vi khuẩn, sinh vật cổ và sinh vật nhân thực đã chỉ ra rằng ba nhóm loài này đã phân ly khỏi nhau khoảng từ 2 tỷ đến 4 tỷ năm trước, đồng thời ủng hộ mạnh mẽ giả thuyết chúng là những siêu giới cơ bản của các sinh vật (xem Hình 21.15).

Ngoài giá trị sử dụng trong nghiên cứu tiến hoá, các nghiên cứu hệ gene học so sánh còn giúp khẳng định sự phù hợp trong việc lựa chọn nghiên cứu ở các sinh vật mô hình từ đó giúp chúng ta hiểu biết ngày càng đầy đủ hơn về sinh học nói chung và về sinh học người nói riêng. Nhiều gene đã tiến hoá qua một thời gian dài, song có thể vẫn giống nhau một cách ngạc nhiên ở các loài khác hẳn nhau. Một ví dụ về điều này là một số gene ở nấm men giống với một số gene gây bệnh nhất định ở người đến mức những nhà nghiên cứu có thể suy luận ra chức năng của những gene gây bệnh này thông qua nghiên cứu các gene tương ứng ở nấm men. Sự giống nhau đáng ngạc nhiên này cho thấy nguồn gốc chung của hai loài có quan hệ xa nhau này.

So sánh giữa các loài có quan hệ gần gũi

Hệ gene của hai loài có quan hệ gần gũi nhiều khả năng có tổ chức giống nhau bởi vì chúng mới chỉ phân ly khỏi nhau trong thời gian gần đây. Như chúng ta đã đề cập ở trên, điều này cho phép hệ gene của một loài đã được giải trình tự hoàn toàn có thể được dùng làm khung lắp ráp các trình tự hệ gene của một loài có quan hệ gần gũi với nó, qua đó làm tăng tốc độ lập bản đồ hệ gene của loài thứ hai. Ví dụ như, bằng việc sử dụng hệ gene người làm bản hướng dẫn, các nhà nghiên cứu có thể nhanh chóng giải trình tự hệ gene của chuột.

Sự phân ly gần đây của hai loài có quan hệ gần cũng là cơ sở của hiện tượng chỉ có một số ít sự khác biệt về gene được tìm thấy khi so sánh hệ gene của chúng với nhau. Những khác biệt di truyền nhất định nhờ vậy có thể dễ dàng đối chiếu với những khác biệt hình thái giữa hai loài. Một ứng dụng lý thú của kiểu phân tích này được phát hiện khi các nhà nghiên cứu so sánh hệ gene người với các hệ gene của tinh tinh, chuột nhắt, chuột đồng và các động vật có vú khác. Việc xác định được các gene đồng thời có mặt trong hệ gene của tất cả những loài này nhưng không có trong hệ gene của các loài khác vốn không phải động vật có vú sẽ cung cấp manh mối về quá trình tiến hoá và phát sinh của lớp động vật này; cùng lúc



▲ **Hình 21.15 Mối quan hệ tiến hoá của ba siêu giới sinh vật.** Sơ đồ hình cây này cho thấy sự phân ly từ cổ xưa của ba siêu giới vi khuẩn, sinh vật cổ và sinh vật nhân thực. Một phần của nhánh tiến hoá của sinh vật nhân thực được tách riêng cho thấy sự phân ly của ba loài sinh vật nhân thực được đề cập ở chương này.

đó, những gene được "chia sẻ" chung giữa người và tinh tinh nhưng không có ở chuột đồng có thể cung cấp bằng chứng về quá trình tiến hoá của các loài linh trưởng. Và, tất nhiên, việc so sánh giữa hệ gene người với hệ gene tinh tinh có thể giúp chúng ta trả lời câu hỏi đầy thách thức đã được nêu ngay ở đầu chương này, đó là: thông tin nào trong hệ gene đã tạo nên con người và tinh tinh?

Một phân tích tổng thể các thành phần của hệ gene người và tinh tinh vốn được cho là phân ly khỏi nhau chỉ khoảng 6 triệu năm trước (xem Hình 21.15) cho thấy một số khác biệt cơ bản. Khi cân nhắc các thay thế đơn nucleotide, hai hệ gene người và tinh tinh chỉ khác nhau khoảng 1,2%. Tuy vậy, khi các nhà nghiên cứu phân tích các đoạn DNA dài hơn, họ đã rất ngạc nhiên khi tìm thấy thêm 2,7% khác biệt do việc thêm hay mất những vùng lớn hơn trong hệ gene của loài này hoặc của loài kia; nhiều trình tự cài thêm là những trình tự được nhân đôi hoặc là những đoạn trình tự DNA lặp lại khác. Trong thực tế, một phần ba các đoạn trình tự nhân đôi ở người không có trong hệ gene của tinh tinh, và một số trong những trình tự nhân đôi nay chứa các vùng có liên quan đến các bệnh ở người. Yếu tố *Alu* có trong hệ gene người nhiều hơn so với trong hệ gene tinh tinh, trong khi đó hệ gene tinh tinh chứa nhiều bản sao trình tự tiền virus của các retrovirus vốn không có trong hệ gene người. Tất cả những phát hiện này đã cung cấp manh mối về các áp lực đã làm phân tách hai hệ gene theo hai con đường khác nhau; nói vậy, nhưng chúng ta vẫn chưa có bức tranh đầy đủ về nó. Ngoài ra, chúng ta còn chưa rõ bằng cách nào những khác biệt này dẫn đến những đặc điểm đặc trưng ở mỗi loài.

Để phát hiện ra cơ sở dẫn đến sự khác biệt hình thái giữa hai loài, các nhà sinh học đã nghiên cứu các gene đặc thù và các loại gene khác nhau giữa người và tinh tinh và so sánh chúng với những gene tương ứng ở các loài động vật có vú khác. Hướng nghiên cứu này đã chỉ ra một số gene rõ ràng đã biến đổi (tiến hoá) nhanh hơn ở người so với tinh tinh cũng như so với chuột. Trong số những gene này có các gene liên quan đến các cơ chế bảo vệ cơ thể chống lại các bệnh sốt rét lao và ít nhất liên quan đến một gene điều hoà kích thước não. Khi xét về chức năng, thì các gene dường như tiến hoá nhanh nhất là các gene mã hoá cho các yếu tố phiên mã. Đây là một thông tin hấp dẫn bởi vì các yếu tố phiên mã điều hoà sự biểu hiện của gene và do đó giữ vai trò chính trong điều phối các chương trình di truyền chung.

Một gene mã hoá cho yếu tố phiên mã có biểu hiện biến đổi nhanh trong nhánh tiến hoá ở người được gọi là FOXP2. Một số bằng chứng chỉ ra rằng gene *FOXP2* có chức năng phát triển khả năng phát âm ở động vật có xương sống. Trước hết, các đột biến xảy ra ở gene này gây nên những sai hỏng nghiêm trọng về khả năng phát triển ngôn ngữ và lời nói ở người. Ngoài ra, gene *FOXP2* cũng được biểu hiện trong não của các loài chim sẻ và các hoàng yến trong giai đoạn các loài chim này đến độ tuổi tập hót. Nhưng có lẽ những bằng chứng thuyết phục nhất bắt nguồn từ những thí nghiệm "knock-out (bất hoạt) gen" mà Joseph Buxhaum và cộng sự đã tiến hành nhằm làm hỏng gene *FOXP2* ở chuột rồi tiến hành phân tích kiểu hình thu được (**Hình 21.16**, xem trang bên). Các chuột đột biến đồng hợp tử có não phát triển bất thường và mất khả năng phát ra âm thanh siêu âm bình thường, đồng thời các cá thể chuột mang một bản sao gene này bị hỏng cũng gặp vấn đề rõ rệt trong phát triển âm thanh. Những kết quả này ủng hộ cho ý tưởng cho rằng gene *FOXP2* đã tiến hành bật các gene liên quan đến khả năng phát âm.

Mở rộng từ khái niệm này, các nhà nghiên cứu đang khám phá liệu sự khác nhau giữa protein FOXP2 ở người và tinh tinh có phải là nguyên nhân dẫn đến khả năng phát triển ngôn ngữ giao tiếp ở người vốn không có được ở tinh tinh hay không. Protein FOXP2 ở người và tinh tinh chỉ khác nhau 2 amino acid duy nhất, và ảnh hưởng của sự khác biệt này đến chức năng của protein ở người như thế nào đến nay vẫn là một câu hỏi bí ẩn chưa có câu trả lời.

Câu chuyện về gene *FOXP2* là một ví dụ điển hình về việc bằng cách nào các cách tiếp cận khác nhau có thể bổ sung cho nhau trong việc giúp khám phá các hiện tượng sinh học có ý nghĩa quan trọng. Trong thí nghiệm được minh hoạ trên Hình 21.16, chuột được dùng làm mô hình thay thế cho con người, bởi vì trong những thí nghiệm như vậy, việc thực hiện các nghiên cứu trên người là không phù hợp về đạo đức (cũng như là không thực tế). Chuột và người phân ly khỏi nhau cách đây khoảng 65,5 triệu năm (xem Hình 21.15) và 85% các gene giữa hai loài là giống nhau. Sự giống nhau về vật chất di truyền như vậy có thể được khai thác trong các nghiên cứu về các rối loạn di truyền khác ở người. Nếu các nhà nghiên cứu đã biết các mô và cơ quan bị ảnh hưởng bởi một rối loạn di truyền nhất định, họ có thể tìm ra các gene được biểu hiện ở những vị

▼ Hình 21.16 Tìm hiểu

Chức năng của gene (*FOXP2*) là gì mà nó lại tiến hoá nhanh trong quá trình hình thành loài người?

THÍ NGHIỆM Một số bằng chứng đã ủng hộ cho giả thiết về vai trò của gene *FOXP2* trong quá trình phát triển lời nói và ngôn ngữ ở người và khả năng phát âm ở một số động vật có xương sống khác. Năm 2005, Joseph Buxbaum và các cộng sự tại Trường Đại học Y khoa Mount Sinai và một số viện nghiên cứu khác đã tiến hành tìm hiểu chức năng của gene *FOXP2*. Họ đã sử dụng chuột, là sinh vật mô hình để bất hoạt gene, như một động vật có xương sống có khả năng phát âm. Chuột phát ra âm thanh siêu âm có âm vực cao, giống như tiếng rít, mỗi khi diễn đạt trạng thái "stress". Các nhà nghiên cứu đã áp dụng kỹ thuật di truyền để tạo ra các con chuột có một hoặc hai bản sao của gene *FOXP2* bị phá hỏng.

Kiểu dại : có hai bản sao gene *FOXP2* bình thường | **Dị hợp tử: một bản sao gene *FOXP2* bị phá hỏng** | **Đồng hợp tử: cả hai bản sao gene *FOXP2* bị phá hỏng**

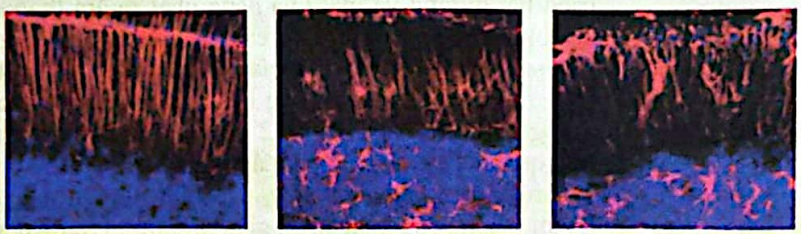
Sau đó họ so sánh kiểu hình của các con chuột này. Hai đặc điểm mà họ đã theo dõi được mô tả ở đây, đó là: giải phẫu não và khả năng phát tiếng.

Thí nghiệm 1: Các nhà nghiên cứu cắt não chuột thành các lát cắt mỏng rồi nhuộm chúng với các hoá chất phù hợp để có thể quan sát được cấu trúc giải phẫu của não dưới kính hiển vi huỳnh quang nguồn sáng UV.

Thí nghiệm 2: Các nhà nghiên cứu tách các con chuột con mới sinh ra khỏi mẹ của chúng và ghi âm số tiếng rít siêu âm do chuột con phát ra.

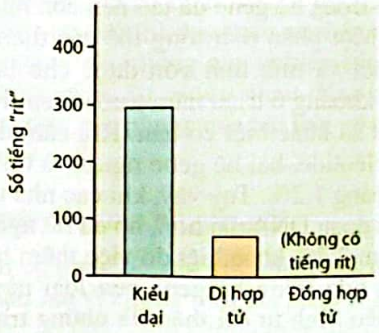
KẾT QUẢ

Thí nghiệm 1: Sự phá hỏng cả hai bản sao gene *FOXP2* dẫn đến sự bất thường trong cấu trúc não, biểu hiện ở sự hỗn độn của các tế bào. Ảnh hưởng kiểu hình đối với cá thể dị hợp tử ở mức độ ít nghiêm trọng hơn.



Kiểu dại | Dị hợp tử | Đồng hợp tử

Thí nghiệm 2: Sự phá hỏng cả hai bản sao gene *FOXP2* dẫn đến việc mất khả năng phát tiếng khi đáp ứng lại với "stress". Ảnh hưởng đối với khả năng phát tiếng của dị hợp tử là đáng kể.



KẾT LUẬN Gene *FOXP2* giữ vai trò quan trọng trong sự phát triển hệ thống liên lạc bằng âm thanh ở chuột. Kết quả nghiên cứu này củng cố thêm bằng chứng cho các nghiên cứu ở chim và người cho thấy gene FOX2P có thể hoạt động chức năng giống nhau ở nhiều loài động vật khác nhau.

NGUỒN W. Shu et al., Altered ultrasonic vocalization in mice with a disruption in the *Foxp2* gene, *Proceedings of the National Academy of Sciences* 102:9643-9648 (2005).

ĐIỀU GÌ NẾU? Do kết quả nghiên cứu này ủng hộ giả thiết về vai trò của gene *FOXP2* trong khả năng phát âm ở chuột, bạn có thể băn khoăn liệu protein FOXP2 có phải là protein có vai trò điều hoà chính trong khả năng phát triển lời nói ở người hay không. Nếu biết trình tự amino acid của các protein FOXP2 bình thường và đột biến ở người, cũng như của protein FOXP2 ở tinh tinh. Bằng cách nào bạn kiểm chứng được câu hỏi trên? Những thông tin bổ sung nào khác có thể tìm thấy khi so sánh những trình tự này với trình tự amino acid của protein FOXP2 ở chuột?

trí đó trong các thí nghiệm được tiến hành trên chuột. Hướng nghiên cứu này đã giúp làm sáng tỏ một số gene đáng quan tâm ở người, bao gồm cả gene góp phần gây nên hội chứng Down.

Các nỗ lực khác đang tiếp tục được triển khai nhằm mở rộng các nghiên cứu hệ gene ở các loài vi sinh vật, các loài linh trưởng khác, kể cả các loài đã từng bị lãng quên thuộc các nhánh khác nhau của cây sự sống. Những nghiên cứu này giúp nâng cao hiểu biết của chúng ta về tất cả các khía cạnh sinh học khác nhau, bao gồm sức khỏe và sinh thái cũng như tiến hoá.

So sánh hệ gene trong phạm vi một loài

Một triển vọng sáng sủa khác bắt nguồn từ khả năng phân tích các hệ gene là chúng ta sẽ ngày càng hiểu biết hơn về phổ biến dị di truyền ở người. Do lịch sử của loài người tương đối ngắn - có lẽ chỉ khoảng 200.000 năm - nên mức độ biến dị di truyền ở người là nhỏ khi so sánh với nhiều loài khác. Phần nhiều sự đa dạng của chúng ta dường như là do các đa hình đơn nucleotide (SNP, đã được mô tả ở Chương 20), thường được phát hiện bằng giải trình tự DNA. Trong hệ gene người, các SNP xuất hiện trung bình với tần số một trong mỗi đoạn từ 100 đến 300 cặp base.

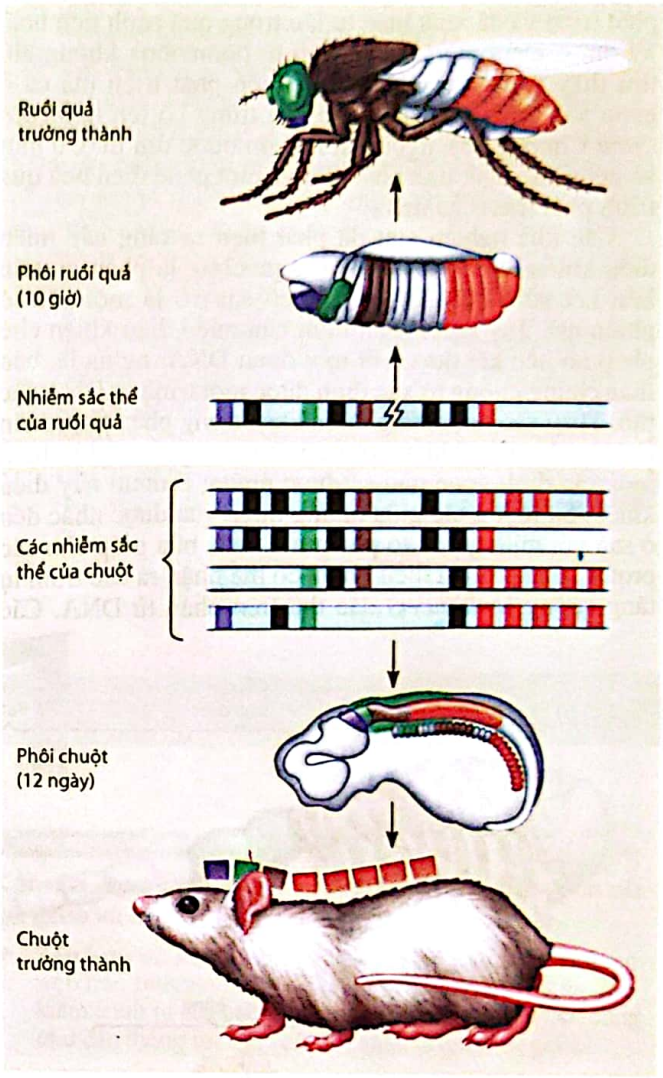
Các nhà khoa học đã xác định được vị trí của vài triệu SNP trong hệ gene người và sẽ tiếp tục tìm thêm các vị trí mới. Trong quá trình triển khai hướng nghiên cứu này, họ cũng đã tìm ra nhiều dạng biến dị khác - gồm đảo đoạn, mất đoạn và lặp đoạn - nhưng không có biểu hiện gây hại rõ rệt đối với các cơ thể mang chúng. Những dạng biến dị này, cũng như các SNP, sẽ là những dấu chuẩn di truyền hiệu quả trong nghiên cứu tiến hoá ở người, trong việc phát hiện các khác biệt giữa các quần thể người, và tìm ra con đường di cư của các quần thể người qua lịch sử. Sự đa dạng di truyền như vậy trong DNA của người cũng sẽ là những dấu chuẩn có giá trị để xác định được các gene gây bệnh cũng như các gene có những ảnh hưởng đến sức khỏe của chúng ta một cách ít rõ ràng hơn. Ngoài việc cung cấp cho chúng ta những thông tin về quá trình tiến hoá, việc phân tích những đặc điểm khác biệt trong hệ gene của các cá thể có thể sẽ làm thay đổi các liệu pháp y học sau này trong thế kỷ XXI.

So sánh quá trình phát triển

Các nhà sinh học thuộc lĩnh vực sinh học tiến hoá - phát triển hay còn thường được gọi tắt là **evo-devo** thường tiến hành so sánh các quá trình phát triển của các cơ thể đa bào khác nhau. Mục tiêu của họ là tìm hiểu những quá trình này đã tiến hoá như thế nào và bằng cách nào những thay đổi trong quá trình phát triển có thể làm biến đổi những đặc điểm của cơ thể hoặc thậm chí dẫn đến sự hình thành các đặc điểm mới. Nhờ các tiến bộ trong kỹ thuật phân tử và làn sóng thông tin về các hệ gene gần đây, chúng ta bắt đầu nhận ra rằng ở các loài có quan hệ họ hàng, dù cho chúng có hình dạng khác nhau rõ rệt, song sự khác biệt trong trình tự của các gene cũng như sự điều hoà biểu hiện của chúng thường rất nhỏ. Việc phát hiện ra cơ sở phân tử dẫn đến những khác biệt này đồng thời giúp chúng ta có được những hiểu biết về nguồn gốc của vô số các dạng sống đa dạng đang cùng chung sống trên hành tinh này, qua đó cung cấp thông tin cho các nghiên cứu tiến hoá của chúng ta.

Tính bảo thủ phổ biến của các gene điều khiển phát triển ở các loài động vật

Ở Chương 18, chúng ta đã đề cập các gene điều khiển phát triển (homeotic genes) ở ruồi *Drosophila* có vai trò của chúng trong việc xác định sự phân đốt cơ thể (xem Hình 18.18). Việc phân tích phân tử các gene điều khiển phát triển ở *Drosophila* cho thấy tất cả các cá thể đều có một trình tự dài 180 nucleotide được gọi là **hộp điều khiển** (homeobox) mã hoá cho một *miền điều khiển* (*homeo domain*) gồm 60 amino acid trong phân tử protein. Một trình tự giống hệt hoặc rất giống với homeobox của ruồi quả đến nay đã được tìm thấy trong các gene điều khiển ở nhiều loài động vật có xương sống và không xương sống. Những trình tự này giữa người và ruồi quả giống nhau đến nỗi, trong thực tế, một nhà nghiên cứu đã ví von "ruồi là những con người nhỏ mang cánh". Sự giống nhau của những gene này còn biểu hiện ở cách tổ chức của chúng: Các gene ở động vật có xương sống tương đồng với các gene điều khiển phát triển ở ruồi quả đều có cùng cách sắp xếp trên nhiễm sắc thể **(Hình 21.17)**. Các trình tự chứa homeobox cũng được tìm thấy ở các gene điều hoà ở nhiều sinh vật nhân thực có quan hệ họ hàng rất xa nhau, chẳng hạn như giữa thực vật và nấm men. Từ những đặc điểm giống nhau này, chúng ta có thể suy ra rằng trình tự



▲ Hình 21.17 Sự bảo thủ của gene điều khiển phát triển (homeotic genes) ở ruồi quả và chuột. Các gene điều khiển phát triển có vai trò điều phối sự hình thành các cấu trúc đầu - đuôi của cơ thể xuất hiện trên nhiễm sắc thể theo các trật tự rất giống nhau giữa ruồi *Drosophila* và chuột. Mỗi băng được tô màu trên nhiễm sắc thể ở đây biểu diễn cho một gene "homeotic". Ở ruồi quả, tất cả các gene điều khiển phát triển được tìm thấy trên cùng một nhiễm sắc thể. Chuột và các loài động vật có vú khác có các bộ gene giống nhau hoặc giống hệt nhau phân bố trên bốn nhiễm sắc thể khác nhau. Các khối màu được vẽ trên hình biểu diễn các phần của phôi mà ở đó những gene có màu này được biểu hiện mà cuối cùng dẫn đến sự hình thành các phần tương ứng ở cơ thể trưởng thành. Những gene này giống hệt nhau khi so sánh giữa ruồi quả và chuột, ngoại trừ các gene được tô màu đen, chúng giống nhau ít hơn so với những gene kia.

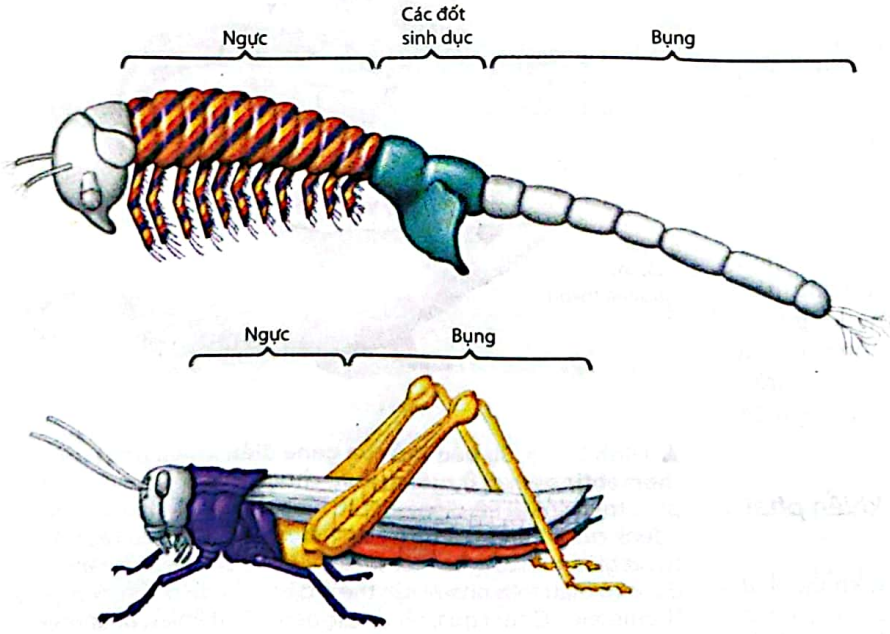
DNA của homeobox đã hình thành từ rất sớm trong lịch sử tiến hoá của sự sống và vai trò của chúng đối với các cơ thể quan trọng đến mức dường như chúng không biến đổi qua hàng trăm triệu năm ở cả động vật và thực vật.

Các gene điều khiển phát triển ở động vật được gọi đặt tên là các gene *Hox* (viết tắt của các gene mang trình tự *h*omeob*ox*), bởi vì những gene điều khiển phát triển đầu tiên được tìm thấy chứa trình tự homeobox. Có những gene chứa homeobox sau này được tìm thấy không hoạt động như các gene điều khiển phát triển; nghĩa là chúng không trực tiếp xác định các phần và vị trí các phần của cơ thể. Tuy vậy, phần lớn những gene này, ít nhất ở động vật, đều liên quan đến quá trình phát triển, qua đó cho thấy chúng có vai trò quan trọng cơ bản trong quá trình phát triển và đã xuất hiện từ lâu trong quá trình tiến hoá. Ví dụ, ở *Drosophila*, các trình tự homeobox không chỉ tìm thấy trong các gene điều khiển phát triển mà cả ở gene xác định tính phân cực của trứng có tên là *bicoid* (xem Chương 18), ngoài ra, nó còn được tìm thấy ở một số gene xác định tính phân đốt và một gene điều hoà quá trình phát triển của mắt.

Các nhà nghiên cứu đã phát hiện ra rằng các miền điều khiển được mã hoá bởi homeobox là phần protein liên kết với DNA khi protein có vai trò là một yếu tố phiên mã. Tuy nhiên, cấu hình của miền điều khiển cho phép nó liên kết được với mọi đoạn DNA; nghĩa là, bản thân chúng không tự xác định được một trình tự DNA đặc thù. Thay vào đó, những miền khác trong phân tử protein chứa miền điều khiển, vốn có mức độ biến dị lớn hơn, mới xác định gene nào sẽ được những protein này điều khiển. Sự tương tác giữa những miền vừa được nhắc đến ở sau với những yếu tố phiên mã khác nữa giúp cho các protein mang miền điều khiển có thể nhận ra các trình tự tăng cường (*enhancer*) đặc thù trên phân tử DNA. Các protein có homeodomain có thể đã điều hoà quá trình phát triển bằng việc điều phối hoạt động phiên mã của những bộ gene sinh trưởng khác nhau, làm chúng "bật" hay "tắt". Ở phôi của *Drosophila* và nhiều loài động vật khác, những tổ hợp khác nhau của các gene homeobox được hoạt hoá ở những phần khác nhau. Sự biểu hiện chọn lọc của những gene điều hoà như vậy, vốn khác nhau về thời gian và vị trí trong quá trình phát triển của phôi, là trung tâm của sự hình thành mẫu hình cơ thể.

Bên cạnh các gene điều khiển phát triển, các nhà sinh học phát triển còn tìm thấy nhiều gene khác liên quan đến quá trình phát triển và có tính bảo thủ rất cao khi so sánh giữa các loài. Những gene này bao gồm nhiều gene mã hoá cho các thành phần của các con đường truyền tín hiệu. Sự giống nhau một cách đặc biệt giữa các gene sinh trưởng nhất định ở các loài động vật khác nhau làm nảy sinh một câu hỏi: Bằng cách nào những gene giống nhau có thể cùng tham gia vào quá trình phát triển ở những loài động vật mà hình dạng của chúng rất khác nhau?

Những nghiên cứu gần đây đã góp phần gợi ý câu trả lời đối với câu hỏi này. Trong một số trường hợp, những thay đổi nhỏ trong trình tự điều khiển của những gene nhất định có thể làm thay đổi kiểu biểu hiện của gene dẫn đến những thay đổi lớn về hình dạng cơ thể. Hãy xem ví dụ sau: các kiểu biểu hiện khác nhau của các gene *Hox* dọc theo trục cơ thể ở côn trùng và giáp xác có thể giải thích cho các dạng biến dị về số các đốt thân có chân ở các loài động vật phân đốt **(Hình 21.18)**. Tương tự như vậy, những nghiên cứu gần đây chỉ ra rằng cùng một sản phẩm của gene *Hox* có thể gây nên những hiệu ứng khác nhau đôi chút ở những loài khác nhau, giúp "bật" những gene mới hoặc "bật" những gene giống nhau song ở mức độ biểu hiện tăng lên hoặc giảm đi. Trong những trường hợp khác, có những gene giống nhau nhưng điều khiển các quá trình phát triển khác nhau khi xét ở các loài khác nhau, dẫn đến sự đa dạng về hình dạng cơ thể. Chẳng hạn như, một số gene *Hox* được biểu hiện trong giai đoạn phôi và ấu trùng ở loài nhím biển vốn là một loài động vật không thuộc nhóm phân đốt và có sơ đồ cơ thể khác hoàn toàn so với côn trùng và chuột. Các con nhím biển trưởng thành có hình dạng cơ thể như những "chiếc gối cắm kim". Chúng thuộc nhóm loài từ lâu được dùng cho nghiên cứu phôi học kinh điển (xem Chương 47).



▲ Hình 21.18 Ảnh hưởng của sự điều hoà biểu hiện gene *Hox* khác nhau trong quá trình phát triển ở giáp xác và côn trùng. Những thay đổi về kiểu biểu hiện của bốn gene *Hox* đã diễn ra qua thời gian tiến hoá. Những thay đổi này phần nào giải thích cho sự khác biệt về sơ đồ cơ thể giữa loài tôm biển *Artemia* (một loài giáp xác, hình trên) với loài châu chấu (một loài côn trùng, hình dưới). Được minh hoạ ở đây là các phần cơ thể được tô màu nhằm phản ánh sự biểu hiện của các gene *Hox* khác nhau có vai trò xác định các phần cơ thể trong quá trình phát triển phôi. Mỗi màu đại diện cho một gene *Hox* đặc thù.

So sánh quá trình phát triển giữa động vật và thực vật

Tổ tiên chung gần nhất của động vật và thực vật có lẽ là một sinh vật nhân thực đơn bào sống cách đây hàng trăm triệu năm; do vậy, các quá trình phát triển hẳn là đã tiến hoá độc lập với nhau tạo nên hai nhánh sinh vật đa bào. Thực vật tiến hoá cùng với sự xuất hiện thành tế bào cứng, gây ngăn cản sự vận động của các mô và tế bào vốn là đặc điểm quan trọng ở động vật. Thay vào đó, sự phát sinh hình thái ở thực vật chủ yếu phụ thuộc vào sự hình thành các mặt phẳng phân bào khác nhau và tính giãn nở chọn lọc của tế bào. (Chúng ta sẽ đề cập về những quá trình

này ở Chương 35.) Tuy có nhiều khác biệt giữa động vật và thực vật, nhưng các cơ chế phân tử của quá trình phát triển ở động vật và thực vật có nhiều điểm giống nhau; có lẽ, đó là di sản chung mà chúng thừa hưởng từ tổ tiên.

Ở cả động vật và thực vật, quá trình phát triển phụ thuộc vào một chuỗi (gồm nhiều mắt xích) các yếu tố điều hoà phiên mã có vai trò bật hoặc tắt các gene khác nhau theo các thứ tự được điều khiển tinh vi. Ví dụ như, một số nghiên cứu ở một loài thực vật có hoa nhỏ thuộc họ mù tạt là *Arabidopsis thaliana* cho thấy sự sắp xếp toả tròn của các phần thuộc hoa cũng giống như sự hình thành trục đầu - đuôi ở *Drosophila* đều liên quan đến một chuỗi các yếu tố phiên mã. Tuy vậy, những gene điều khiển những quá trình này lại khác nhau đáng kể giữa động vật và thực vật. Nếu như khá nhiều các công tắc điều hoà ở *Drosophila* là các gene *Hox* mang các homeobox, thì các gene ở *Arabidopsis* thuộc một họ gene hoàn toàn khác, được gọi là các gene *Mads-box*. Mặc dù các gene chứa các homebox cũng được tìm thấy ở thực vật, cũng như các gene *Mads-box* cũng được tìm thấy ở động vật, song trong cả hai trường hợp chúng không thực hiện những vai trò chính giống nhau trong quá trình phát triển giống như ở nhóm kia. Như vậy, các bằng chứng phân tử ủng hộ cho giả thiết rằng: chương trình phát triển đã tiến hoá độc lập với nhau ở động vật và thực vật.

Qua chương cuối cùng này thuộc khối kiến thức di truyền học, chúng ta đã biết bằng cách nào các nghiên cứu về thành phần các hệ gene và so sánh hệ gene giữa các loài có thể làm sáng tỏ quá trình tiến hoá của các hệ gene. Ngoài ra, bằng việc so sánh chương trình phát triển, chúng ta có thể thấy sự thống nhất của sinh giới được biểu hiện qua sự giống nhau về các cơ chế phân tử và tế bào được dùng để thiết lập nên các mẫu hình cơ thể, mặc dù các gene điều khiển quá trình phát triển có thể khác nhau giữa các sinh vật khác nhau. Sự giống nhau giữa các hệ gene đồng thời phản ánh tổ tiên chung của tất cả các dạng sống trên Trái Đất. Nhưng sự khác nhau cũng là đáng kể, từ đó chúng đã tạo nên sự đa dạng phong phú của các loài qua tiến hoá. Trong phần còn lại của cuốn sách này, chúng ta sẽ mở rộng tầm quan sát ra khỏi các cấp độ phân tử, tế bào và gene để khám phá sự đa dạng của sinh vật ở cấp độ cơ thể.

KIỂM TRA KHÁI NIỆM 21.6

1. Theo bạn, hệ gene của khỉ giống hệ gene của chuột hơn hay giống hệ gene của người hơn? Tại sao?
2. Các trình tự DNA được gọi là các homeobox, giúp các gene điều khiển phát triển ở động vật có thể điều phối được quá trình phát triển, rất giống nhau giữa ruồi và chuột. Hãy giải thích tại sao mặc dù có sự giống nhau này, nhưng hình thái của các động vật này là rất khác nhau.
3. **ĐIỀU GÌ NẾU?** Các yếu tố *Alu* trong hệ gene người nhiều gấp 3 lần so với hệ gene tinh tinh. Theo bạn, bằng cách nào hệ gene người có thêm những bản sao của các yếu tố *Alu*? Hãy nêu một vai trò có thể có của yếu tố *Alu* trong quá trình tiến hoá phân ly giữa hai loài.

Câu trả lời có trong Phụ lục A.

Ôn tập chương 21

TÓM TẮT CÁC KHÁI NIỆM THEN CHỐT

KHÁI NIỆM 21.1

Các phương pháp mới đã giúp gia tăng tốc độ giải trình tự các hệ gene (tr. 427 – 429)

- **Giải trình tự hệ gene qua ba giai đoạn** Trong giai đoạn lập bản đồ gene liên kết, trật tự của các gene và các dấu chuẩn di truyền khác trong hệ gene và khoảng cách tương đối giữa chúng có thể được xác định thông qua các tần số tái tổ hợp. Trong giai đoạn tiếp theo, bản đồ vật lý dùng các trình tự gối lên nhau của các phân đoạn DNA để sắp xếp các đoạn này vào đúng trật tự của chúng trong hệ gene, đồng thời xác định khoảng cách giữa các dấu chuẩn tính theo đơn vị cặp base. Cuối cùng, ở giai đoạn thứ ba, các đoạn DNA đã được giải trình tự được xếp theo đúng trật tự, từ đó chúng ta thu được trình tự hệ gene đầy đủ.
- **Giải trình tự ngẫu nhiên toàn hệ gene** Toàn hệ gene được phân cắt thành nhiều đoạn nhỏ, gối lên nhau. Những đoạn này sau đó được giải trình tự, rồi được ráp nối lại thành các trình tự hoàn chỉnh nhờ phần mềm máy tính. Nếu có sẵn các thông tin về bản đồ di truyền, thì việc ráp nối sẽ thuận lợi và chính xác.

KHÁI NIỆM 21.2

Các nhà khoa học ứng dụng tin sinh học để phân tích các hệ gene và chức năng của chúng (tr. 429 – 432)

- **Tập hợp dữ liệu để phân tích các hệ gene** Nhiều trang web trên Internet cung cấp tài nguyên truy cập và tìm kiếm trình tự của các hệ gene, các công cụ phân tích cũng như các thông tin khác có liên quan đến các hệ gene.
- **Xác định các gene mã hoá protein trong các trình tự DNA** Việc phân tích các trình tự hệ gene nhờ máy tính giúp các nhà nghiên cứu xác định được các trình tự nhiều khả năng mã hoá cho các protein. Việc so sánh các trình tự của các gene "mới" với các gene đã biết ở những loài khác có thể giúp xác định chức năng của các gene mới. Đối với gene chưa rõ chức năng, việc gây bất hoạt gene thực nghiệm và quan sát hiệu quả kiểu hình thu được có thể cung cấp đầu mối thông tin về chức năng của chúng.
- **Tìm hiểu các gene và các sản phẩm của gene ở cấp độ sinh học hệ thống** Bằng việc sử dụng máy tính và các công cụ sinh tin học, các nhà khoa học có thể so sánh các hệ gene và nghiên cứu tập hợp các gene và protein như các hệ thống hoàn chỉnh (hệ gene học và hệ protein học). Những nghiên cứu này bao gồm các phân tích về tương tác protein ở quy mô lớn.

KHÁI NIỆM 21.3

Các hệ gene khác nhau về kích cỡ, số gene và mật độ gene (tr. 432 - 434)

	Vi khuẩn	Vi sinh vật cổ	Sinh vật nhân thực
Kích cỡ hệ gene	Phần lớn trong khoảng 1 - 6 Mb		Phần lớn trong khoảng 10 - 4.000 Mb, nhưng một số lớn hơn
Số lượng gene	1.500 - 7.500		5.000 - 40.000
Mật độ gene	Cao hơn so với sinh vật nhân thực		Thấp hơn so với sinh vật nhân sơ (Trong các sinh vật nhân thực, mật độ gene thấp hơn ở các hệ gene lớn hơn)
Các intron	Không có ở các gene mã hoá protein	Có ở một số gene	Ở sinh vật nhân thực đơn bào: có, song chỉ phổ biến ở một số loài Ở sinh vật nhân thực đa bào: có ở phần lớn gene
DNA không mã hoá khác	Chỉ có rất ít trong hệ gene		Chiếm lượng lớn trong hệ gen; nhìn chung có nhiều trình tự lặp lại không mã hoá ở các sinh vật nhân thực đa bào

KHÁI NIỆM 21.4

Sinh vật nhân thực đa bào có nhiều DNA không mã hoá và nhiều họ đa gene (tr. 434 – 438)

- Chỉ có 1,5% hệ gene người mã hoá cho protein hoặc cho rRNA và tRNA; phần còn lại là các trình tự DNA không mã hoá, bao gồm nhiều trình tự DNA lặp lại.
- **Các yếu tố di truyền di động và các trình tự có liên quan** Loại trình tự DNA lặp lại phổ biến nhất trong hệ gene sinh vật nhân thực đa bào là các yếu tố di truyền di động và các trình tự có liên quan. Có hai loại yếu tố di động trong hệ gene sinh vật nhân thực: *transposon* di chuyển thông qua một phân tử DNA trung gian, và *retrotransposon* có mức độ phổ biến cao hơn và di chuyển thông qua một phân tử RNA trung gian.
- **Các trình tự DNA lặp lại khác, bao gồm cả các DNA trình tự đơn giản** Nhiều trình tự ngắn không mã hoá lặp lại liên tục hàng nghìn lần (gọi là các DNA trình tự đơn giản, bao gồm cả các STR) có đặc biệt phổ biến ở vùng tâm động và các đầu mút nhiễm sắc thể; chúng có thể có vai trò quan trọng trong cấu trúc của nhiễm sắc thể.
- **Các gene và các họ đa gene** Mặc dù nhiều gene ở sinh vật nhân thực chỉ có một bản sao duy nhất trong bộ nhiễm sắc thể đơn bội của chúng, các gene còn lại (phần lớn ở một số loài) là thành viên của cùng họ các gene có quan hệ với nhau. Đơn vị phiên mã tương ứng với ba loại rRNA lớn nhất lặp lại liên tiếp hàng trăm nghìn lần trên một hoặc một số vị trí khác nhau của các nhiễm sắc thể; cấu trúc này giúp tế bào có thể nhanh chóng tổng hợp rRNA cần cho hàng triệu ribosome. Các gene trong họ gene globin khác nhau đôi chút. Các gene này mã hoá cho chuỗi polypeptide được sử dụng trong các giai đoạn phát triển khác nhau của con vật.

KHÁI NIỆM 21.5

Lặp đoạn, tái cấu trúc và đột biến trong trình tự DNA đóng góp vào quá trình tiến hoá (tr. 438 – 442)

- **Sự lặp lại cả bộ nhiễm sắc thể.** Các sự kiện ngẫu nhiên trong quá trình phân bào có thể dẫn đến các tế bào có thêm những bản sao của tất cả hay một phần hệ gene. Những hệ gene này sau đó có thể phân ly nếu như một bộ nhiễm sắc thể tích luỹ các biến đổi trong trình tự DNA.
- **Sự thay đổi cấu trúc nhiễm sắc thể** Cấu trúc nhiễm sắc thể các hệ gene có thể được đem so sánh giữa các loài, qua đó cung cấp thông tin về mối quan hệ tiến hoá. Trong phạm vi một loài nhất định, sự sắp xếp lại các nhiễm sắc thể được cho là một cơ chế đóng góp vào sự phát sinh loài mới.
- **Lặp đoạn và sự phân hoá của các vùng DNA có kích thước tương ứng với gene** Các gene mã hoá cho các loại protein globin khác nhau được tiến hoá bắt nguồn từ một gene globin tổ tiên chung; gene tổ tiên này đã được lặp gene và phân hoá thành các gene tổ tiên của α-globin và β-globin. Hiện tượng lặp gene sau đó kết hợp với các đột biến ngẫu nhiên đã dẫn đến sự hình thành các gene hiện nay; tất cả những gene này đều mã hoá cho các protein liên kết oxygen. Các bản sao khác nhau của gene lặp đã phân hoá đến mức mà chức năng của các protein hiện nay do chúng mã hoá đã trở nên khác nhau một cách cơ bản.
- **Sự sắp xếp lại các phần của gen: lặp và xáo trộn exon** Sự sắp xếp lại các exon trong phạm vi một gene và giữa các gene trong quá trình tiến hoá đã dẫn đến các gene chứa nhiều bản sao của các exon giống nhau và/hoặc của một số exon khác nhau bắt nguồn từ các gene khác.
- **Các yếu tố di động góp phần vào sự tiến hoá của hệ gene như thế nào?** Sự di chuyển của các yếu tố di động hay tái tổ hợp giữa các bản sao của cùng một yếu tố di động đôi khi tạo ra những tổ hợp trình tự mới có lợi cho cơ thể sinh vật. Những cơ chế như vậy có thể làm thay đổi các chức năng của các gene hoặc kiểu biểu hiện hay điều hoà biểu hiện của chúng.

KHÁI NIỆM 21.6

So sánh các trình tự hệ gene cung cấp bằng chứng về các quá trình tiến hoá và phát triển (tr. 442-447)

- **So sánh các hệ gene** Các nghiên cứu so sánh hệ gene giữa các loài có quan hệ gần gũi cũng như giữa các loài có mức độ phân ly xa hơn cung cấp nhiều thông tin giá trị tương ứng về lịch sử tiến hoá cận đại và cổ xưa. Các trình tự hệ gene người và tinh tinh khác nhau khoảng 4%, chủ yếu do thêm đoạn, mất đoạn, và lặp đoạn trong một nhánh tiến hoá. Cùng với các biến đổi về các nucleotide trong những gene đặc thù (ví dụ như gene *FOXP2*, một gene ảnh hưởng đến khả năng phát âm), những thay đổi này có thể giải thích cho các đặc điểm khác biệt giữa hai loài. Các đa hình đơn nucleotide giữa các cá thể trong phạm vi một loài cũng có thể cung cấp thông tin về lịch sử tiến hoá của loài đó.
- **So sánh quá trình phát triển** Các gene điều khiển phát triển và một số gene khác có liên quan đến quá trình phát triển ở động vật chứa một vùng homeobox; đó là trình tự giống hệt nhau hoặc rất giống nhau ở nhiều loài. Nhiều trình tự có quan hệ với nhau được tìm thấy đồng thời ở các hệ gene thực vật và nấm men. Các gene điều hoà phát triển khác cũng có tính bảo thủ cao ở các loài động vật, nhưng chúng có thể có vai trò khác nhau trong quá trình phát triển của các loài khác nhau. Trong quá trình phát triển phôi ở động vật cũng như thực vật có một chuỗi các yếu tố phiên mã giúp bật hoặc tắt các gene theo một trật tự nghiêm ngặt. Tuy vậy, các gene điều khiển quá trình phát triển tương tự lại có trình tự khác nhau đáng kể khi so sánh giữa động vật và thực vật; có lẽ do tổ tiên của chúng đã phân ly từ lâu trong quá trình tiến hoá.

KIỂM TRA KIẾN THỨC CỦA BẠN

TỰ KIỂM TRA

1. Tin sinh học bao gồm tất cả các nội dung sau, *trừ*
   a. sử dụng các chương trình máy tính để so sánh các trình tự DNA.
   b. phân tích tương tác giữa các protein trong một loài.
   c. sử dụng sinh học phân tử để kết hợp DNA từ các nguồn khác nhau trong điều kiện *invitro*.
   d. phát triển các công cụ máy tính để phân tích hệ gene.
   e. sử dụng các công cụ toán học để tìm hiểu các hệ thống sinh học.

2. Loài nào dưới đây có hệ gene lớn nhất và mật độ gene thấp nhất tính theo số cặp base có trong hệ gene?
   a. *Haemophilus influenzae* (vi khuẩn)
   b. *Saccharomyces cerevisiae* (nấm men)
   c. *Arabidopsis thaliana* (thực vật)
   d. *Drosophila melanogaster* (ruồi quả)
   e. *Homo sapiens* (người hiện đại)

3. Một đặc điểm của các retrotransposon là
   a. chúng mã hoá cho một enzyme tổng hợp DNA sử dụng RNA làm mạch khuôn.
   b. chúng chỉ được tìm thấy ở các tế bào động vật.
   c. chúng thường vận động bởi cơ chế cắt - dán.
   d. chúng đóng góp một phần đáng kể vào các biến dị di truyền quan sát thấy trong một quần thể các giao tử.
   e. chúng được nhân bản phụ thuộc vào một retrovirus.

4. Các họ đa gene là
   a. các nhóm trình tự tăng cường điều khiển phiên mã.
   b. các gene thường kết thành cụm ở các đầu mút.
   c. các cấu trúc tương ứng với các operon của sinh vật nhân sơ.
   d. các nhóm gene được điều khiển đồng thời.
   e. các bộ gene giống hệt nhau hoặc rất giống nhau tiến hoá qua quá trình lặp gene.

5. Hai protein ở sinh vật nhân thực chứa một miền chức năng giống nhau, song phần còn lại của chúng thì rất khác nhau. Quá trình nào dưới đây nhiều khả năng góp phần gây ra sự giống nhau này của chúng?
   a. Lặp gene
   b. Xén RNA
   c. Trao đổi exon
   d. Biến đổi histone
   e. Các đột biến điểm ngẫu nhiên

6. Các gene điều khiển phát triển (homeotic genes)
   a. mã hoá cho các yếu tố phiên mã điều khiển sự biểu hiện của các gene có vai trò xác định cấu trúc giải phẫu đặc thù.
   b. chỉ được tìm thấy ở *Drosophila* và các loài thân đốt.
   c. là những gene duy nhất mang miền homeobox.
   d. mã hoá cho các protein hình thành nên cấu trúc giải phẫu của ruồi quả.
   e. có vai trò xác định mẫu hình phát triển của thực vật.

7. **HÃY VẼ** Ở phía trên của cột bên là các trình tự amino acid (dùng kiểu viết tắt một chữ cái; xem Hình 5.17) thuộc bốn phân đoạn ngắn của protein FOXP2 được tìm thấy ở 6 loài khác nhau, gồm: tinh tinh, đười ươi, khỉ gorila, khỉ rhesus chuột và người. Đây chính là những phân đoạn chứa tất cả các amino acid khác nhau trong protein FOXP2 khi so sánh giữa những loài này.

ATETI ... PKSSD ... TSSTT ... NARRD
ATETI ... PKSSE ... TSSTT ... NARRD
ATETI ... PKSSD ... TSSTT ... NARRD
ATETI ... PKSSD ... TSSNT ... SARRD
ATETI ... PKSSD ... TSSTT ... NARRD
VTETI ... PKSSD ... TSSTT ... NARRD

Hãy dùng bút đánh dấu bôi màu vào các amino acid khác biệt giữa các loài. (phủ màu lên amino acid đó ở tất cả các loài.) Sau đó, trả lời các câu hỏi dưới đây:
   a. Các trình tự của tinh tinh (T), khỉ gorila (G) và khỉ rhesus (R) giống hệt nhau. Những dòng nào tương ứng với những loài này.
   b. Trình tự ở người khác với trình tự của các loài T, G và R ở hai amino acid. Dòng nào tương ứng với trình tự của người? Gạch chân hai amino acid khác biệt.
   c. Trình tự của đười ươi khác với trình tự của các loài T, G và R ở một amino acid (thay thế alanine bằng valine) và khác với trình tự của người ở ba amino acid. Dòng nào tương ứng với trình tự của đười ươi?
   d. Có bao nhiêu amino acid khác biệt giữa trình tự của chuột so với trình tự của các loài T, G và R? Khoanh tròn vào các amino acid khác biệt ở chuột. Có bao nhiêu amino acid khác biệt giữa chuột và người? Vẽ hình vuông bao quanh amino acid khác biệt ở chuột.
   e. Các loài linh trưởng và bộ Gặm nhấm phân ly khỏi nhau cách đây khoảng từ 60 đến 100 triệu năm, còn người và tinh tinh phân ly khỏi nhau cách đây khoảng 6 triệu năm. Từ cơ sở đó, bạn có thể kết luận điều gì khi so sánh sự khác biệt về trình tự amino acid giữa chuột với các loài T, G và R đồng thời đối chiếu với sự khác biệt giữa người với các loài T, G và R?

Đáp án cho câu hỏi trắc nghiệm có trong Phụ lục A.

LIÊN HỆ VỚI TIẾN HOÁ

8. Các gene giữ vai trò quan trọng trong phát triển phôi ở động vật, chẳng hạn như các gene mang miền homeobox, có tính bảo thủ tương đối cao trong quá trình tiến hoá; nghĩa là, khi so sánh giữa các loài, chúng giống nhau nhiều hơn so với nhiều gene khác. Tại sao lại như vậy?

KHOA HỌC, CÔNG NGHỆ VÀ XÃ HỘI

9. Các nhà khoa học khi lập bản đồ các SNP trong hệ gene người đã để ý thấy hiện tượng các nhóm SNP có xu hướng di truyền với nhau thành từng khối và được gọi là các đơn dạng (hoặc đơn nhánh; haplotype) có chiều dài từ 5.000 đến 200.000 cặp base. Mỗi haplotype chỉ có khoảng từ 4 đến 5 kiểu tổ hợp của các SNP xuất hiện phổ biến. Hãy nêu giả thiết giải thích cho hiện tượng này trên cơ sở kết hợp các thông tin thu nhận được từ chương này nói riêng và khối kiến thức di truyền học nói chung.