Phân tích gen và sản phẩm của gen

Bởi:

PGS. TS. Phạm Thành Hổ

Các kỹ thuật phân tích axit nucleic

Điện di phân tích ADN và ARN

Có nhiều phương pháp khác nhau có thể được sử dụng để phân tích ADN và ARN, nhưng cho đến nay **điện di trên gel** là phương pháp được sử dụng phổ biến hơn cả nhờ ưu điểm nhanh và tương đối đơn giản. Nguyên tắc của phương pháp là do: dưới tác động của điện trường, các phân tử ADN (thường tích điện âm) khác nhau về kích thước, điện tích, mức độ cuộn xoắn và dạng phân tử (mạch thẳng hay mạch vòng) sẽ di chuyển qua hệ mạng của gel từ cực âm (cathode) sang cực dương (anode) với tốc độ di chuyển khác nhau. Vì vậy, chúng dần dần tách nhau ra trên trường điện di, qua đó người ta có thể thu thập và phân tích được từng phân đoạn ADN hoặc gen riêng rẽ. Trên điện trường các phân đoạn ADN có kích thước càng nhỏ càng có tốc độ di chuyển trên điện trường nhanh hơn. Sau khi điện di kết thúc, các phân tử ADN có thể quan sát thấy nhờ sử dụng các thuốc nhuộm phát huỳnh quang, chẳng hạn như ethidium (chất này gắn kết với ADN bằng cách cài vào khe ở giữa các nucleotit). Mỗi một băng điện di thường phản ánh một tập hợp các phân tử ADN có cùng kích thước.

Có hai loại vật liệu làm gel được sử dụng phổ biến trong điện di, đó là **agarose** và **polyacrylamid**. Trong đó, gel polyacrylamid có khả năng phân tách cao, nhưng khoảng kích thước ADN có thể phân tích hẹp. Vì vậy, điện di trên gel polyacrylamid có thể phân tách được các phân đoạn ADN khác nhau thậm trí chỉ một cặp nucleotit (1 bp) duy nhất, nhưng thường chỉ để phân tích các đoạn ADN kích thước vài trăm bp. Còn gel agarose có khả năng phân tách thấp hơn đối với các phân đoạn ADN kích thước nhỏ, nhưng rất hiệu quả khi phân tách các phân đoạn ADN kích thước lớn tới hàng chục hoặc hàng trăm kb (1 kb = 1000 bp).

Các phân đoạn ADN kích thước lớn không thể "lọt" qua các lỗ có kích thước nhỏ trên các bản gel, kể cả gel agarose. Thay vào đó, chúng sẽ "trườn" qua mạng lưới của gel bằng việc đầu này của phân tử đi trước, còn đầu kia theo sau. Kết quả là các phân đoạn ADN kích thước lớn (từ 30 đến 50 kb) có tốc độ di chuyển trên điện trường gần như

tương đương và khó phân tách được bằng phương pháp điện di thông thường. Đối với các phân đoạn ADN kích thước lớn như vậy, người ta có thể phân tách bằng việc sử dụng phương pháp **điện di xung trường** (pulsed-field gel electrophoresis). Trong phương pháp này, người ta sử dụng 2 cặp điện cực nằm chéo góc trên bản điện di (hình 9). Việc "bật" và "tắt" luân phiên 2 cặp điện cực sẽ làm cho các phân đoạn ADN lớn thay đổi chiều di chuyển như mình họa trên hình vẽ. Các phân đoạn ADN có kích thước càng lớn càng chậm hơn trong quá trình thay đổi chiều di chuyển. Nhờ vậy, các phân đoạn có kích thước khác nhau sẽ phân tách ra khỏi nhau trong quá trình di chuyển. Kỹ thuật điện di xung trường trong thực tế có thể sử dụng để xác định được kích thước đầy đủ của nhiễm sắc thể vi khuẩn, hoặc nhiễm sắc thể các loài sinh vật nhân chuẩn bậc thấp, như nấm men. Những loài này có kích thước hệ gen khoảng vài Mb.

Điện di không những có thể phân tách các phân đoạn ADN khác nhau về kích thước mà cả về hình dạng và cấu hình không gian của chúng. Các phân tử ADN ở dạng mạch vòng giãn xoắn hoặc bị "đứt gãy" ở một số nucleotit di chuyển chậm hơn trên trường điện di so với các phân tử ADN ở dạng mạch thẳng có cùng khối lượng. Tương tự như vậy, các phân tử ADN ở dạng siêu xoắn, kích thước và thể tích thu nhỏ thường di chuyển nhanh hơn trên trường điện di so với các phân tử ADN dạng mạch vòng giãn xoắn hoặc có mức độ cuộn xoắn thấp hơn có cùng khối lượng.

Kỹ thuật điện di cũng được sử dụng để phân tách các phân tử ARN. Các phân đoạn ADN sợi kép mạch thẳng có cấu trúc bậc hai đồng nhất nên tốc độ di chuyển trên trường điện di tương quan tỉ lệ thuận với khối lượng phân tử của chúng. Cũng giống như ADN, các phân tử ARN cũng thường tích điện âm, nhưng ngoài cấu trúc cơ bản ở dạng mạch đơn, các cấu trúc bậc 2 và 3 của phân tử ARN có ảnh hưởng đến tốc độ di chuyển của chúng trên trường điện di. Để hạn chế điều này, thông thường người ta phải xử lý ARN với một số hóa chất ngăn cản sự hình thành các liên kết cục bộ trong phân tử ARN, chẳng hạn như glyoxal. Hợp chất này liên kết với nhóm $-NH_2$ trong các bazơ nitơ và ngăn cản sự kết cặp của các nucleotit. Các phân tử ARN được xử lý glyoxal không hình thành được các cấu trúc bậc 2 và 3 vì vậy có tốc độ di chuyển trên trường điện di dường như đơn thuần phụ thuộc vào khối lượng phân tử của chúng.

Sử dụng enzym giới hạn trong phân tích ADN

Hầu hết các phân tử ADN trong tự nhiên đều lớn hơn nhiều so với kích thước có thể thao tác và phân tích một cách thuận lợi trong phòng thí nghiệm. Trong các tế bào, phần lớn các nhiễm sắc thể thường là một phân tử ADN dài chứa hàng trăm thậm trí hàng nghìn gen khác nhau. Vì vậy, để có thể phân lập và phân tích từng gen, người ta phải cắt các phân tử ADN kích thước lớn thành các phân đoạn nhỏ. Công việc này được thực hiện bởi một nhóm các enzym đặc biệt gọi là **enzym giới hạn**.

Tất cả các enzym giới hạn đều có hai đặc tính: 1) *nhận biết một trình tự đặc hiệu trên phân tử ADN* (gọi là trình tự giới hạn); và 2) *cắt bên trong phân tử ADN tại vị trí đặc*

hiệu (hoặc ngay tại vị trí giới hạn như đối với nhóm enzym giới hạn loại II; hoặc cách vị trí giới hạn một số nucleotit nhất định như đối với các nhóm enzym giới hạn thuộc các nhóm I và III). Trong các nhóm enzym giới hạn, nhóm thường được dùng trong các nghiên cứu di truyền phân tử và kỹ nghệ gen là nhóm II nhờ vị trí và trình tự cắt của chúng được xác định rõ. Trong phạm vi giáo trình này, vì vậy chúng ta chỉ đề cập đến việc ứng dụng của nhóm enzym giới hạn này. Các trình tự giới hạn của enzym nhóm II thường gồm 4 - 8 bp, thông thường có tính đối xứng và vị trí cắt thường nằm trong trình tự giới hạn này. Ví dụ như enzym giới hạn *EcoRI* được tìm thấy ở vi khuẩn *E. coli* có trình tự giới hạn là 5'-GAATTC-3' với vị trí cắt ở giữa G và A. Tên enzym gồm 3 ký tự đầu chỉ tên loài vi khuẩn mà từ đó enzym được tìm thấy (*Eco = Escherichiacoli*), các ký tự sau chỉ tên của chủng vi khuẩn và số thứ tự của enzym được tìm thấy ở loài vi khuẩn đó (*EcoRI* là enzym giới hạn đầu tiên được tìm thấy ở *E. coli*).

Một enzym giới hạn có trình tự giới hạn gồm 6 bp giống *EcoRI* thông thường được trông đợi sẽ có trung bình một vị trí cắt trong một đoạn trình tự có kích thước khoảng 4 kb (bởi theo nguyên tắc xác suất tại một vị trí nhất định xác suất để có một loại nucleotit nhất định là 1/4, vì vậy xác suất để có một trình tự nhất định gồm 6 bp sẽ là $1/4^6 = 1/4096$). Giả sử có một phân tử ADN mạch thẳng có 6 vị trí cắt của enzym *EcoRI*. Việc cắt phân tử ADN này bằng *EcoRI* sẽ cho ra 7 phân đoạn ADN khác nhau. Do đó, khi điện di trên gel sản phẩm cắt, 7 phân đoạn ADN sẽ phân tách nhau ra do chúng khác nhau về khối lượng (vì chúng khác nhau về thành phần và trình tự các nucleotit). Như vậy, một phân đoạn ADN sẽ tương ứng với một vùng của phân tử ADN ban đầu.

Việc sử dụng một enzym giới hạn khác, chẳng hạn *HindIII* cũng có trình tự giới hạn gồm 6 bp, nhưng có trình tự giới hạn thay đổi (5'-AAGCTT-3') sẽ cho ra các sản phẩm cắt khác với khi sử dụng *EcoRI* (với cùng phân tử ADN ban đầu). Như vậy, việc sử dụng đồng thời nhiều enzym giới hạn sẽ tạo ra một kiểu hình phổ điện di các phân đoạn cắt giới hạn đặc thù đối với từng gen phân tích.

Đối với một số enzym giới hạn khác, chẳng hạn như *Sau*3A1 (tìm thấy ở vi khuẩn *Staphylococcusaureus*) có trình tự giới hạn ngắn hơn (5'-GATC-3'), nên tần số cắt của chúng thường cao hơn các enzym có trình tự giới hạn dài. Theo xác suất, *Sau*3A1 có trung bình 1 vị trí cắt trong một đoạn trình tự khoảng 250 bp ($1/4^4 = 1/256$). Ngược lại, enzym *NotI* có trình tự giới hạn dài (5'-GCGGCCGC-3') trung bình cứ một đoạn trình tự dài khoảng 65 kb, mới có 1 vị trí cắt ($1/4^8 = 1/65536$).

Các enzym giới hạn không chỉ khác nhau về trình tự giới hạn và độ dài đoạn trình tự giới hạn đặc trưng của chúng, mà chúng còn khác nhau về cách "cắt" phân tử ADN. Chẳng hạn như enzym *HpaI* tạo ra các phân tử ADN dạng đầu bằng (đầu tù), còn các enzym *EcoRI*, *HindIII* và *PstI* cắt phân tử ADN tạo ra các phân đoạn có đầu dính. Sở dĩ gọi là "đầu dính" bởi phần các trình tự ở hai đầu sau khi được enzym cắt ra bổ trợ với nhau theo nguyên tắc Chargaff và vì vậy chúng có xu hướng "dính" trở lại với nhau, hoặc với

các phân tử ADN được cắt bởi cùng một loại enzym giới hạn. Tính chất này được ứng dụng rộng rãi trong công nghệ ADN tái tổ hợp và các kỹ thuật tách dòng phân tử.

Các phương pháp lai phân tử và mẫu dò

Các phân tử ADN sợi kép có một tính chất đặc biệt là khả năng biến tính (phân tách thành hai mạch đơn) và hồi tính (hai mạch đơn có trình tự bổ trợ có xu hướng liên kết trở lại khi vắng mặt các tác nhân gây biến tính). Khả năng liên kết bổ trợ giữa các bazơ nitơ cũng cho phép hai mạch ADN có nguồn gốc khác nhau nhưng có trình tự bổ trợ liên kết với nhau trong điều kiện phù hợp (về nhiệt độ, độ pH, ion hóa …) để tạo nên một phân tử ADN mới. Hiện tượng liên kết như vậy cũng có thể xảy ra giữa hai mạch ADN với nhau, hoặc giữa hai mạch ARN hoặc giữa ADN và ARN. Phân tử axit nucleic sợi kép mới hình thành được gọi là phân tử lai và quá trình kết cặp giữa các bazơ thuộc hai mạch đơn axit nucleic có nguồn gốc khác nhau theo nguyên tắc bổ trợ như vậy được gọi là **quá trình lai phân tử**.

Nhiều kỹ thuật trong nghiên cứu di truyền phân tử được dựa trên nguyên tắc lai phân tử. Chẳng hạn, bằng nguyên tắc này người ta có thể dùng một trình tự ADN biết trước để xác định một trình tự bổ trợ tương ứng có trong hệ gen của mẫu phân tích. Phân đoạn ADN có trình tự biết trước dùng trong phản ứng lai như vậy được gọi là **mẫu dò**. Các mẫu dò có thể có nguồn gốc từ các phân đoạn ADN trong tự nhiên hoặc được tổng hợp theo nguyên tắc hóa học, và để nhận biết được chúng, các mẫu dò thường được đánh dấu với các chất phóng xạ hoặc phát huỳnh quang (ở đây, gọi tắt là *chất phát quang*).

Có hai phương pháp đánh dấu mẫu dò ADN. Phương pháp thứ nhất dùng nguyên tắc tổng hợp hóa học phân tử ADN mới với tiền chất là các phân tử đánh dấu. Phương pháp thứ hai là gắn một phân tử đánh dấu vào đuôi của một trình tự ADN có sẵn. Chẳng hạn, bằng việc sử dụng enzym polynucleotide kinase, người ta có thể bổ sung nhóm phosphat ? của ATP vào nhóm 5’-OH của một phân tử ADN định đánh dấu. Nếu nhóm phosphat này được đánh dấu bằng đồng vị phóng xạ ^{32}P thì phân tử ADN sẽ được dánh dấu phóng xạ.

Phương pháp đánh dấu ADN thứ hai (sử dụng các tiền chất đánh dấu) thường được thực hiện dựa trên phản ứng PCR, hoặc đôi khi chỉ cần sử dụng các đoạn mồi ngắn rồi cho enzym ADN polymerase thực hiện phản ứng kéo dài chuỗi. Các tiền chất đánh dấu được sử dụng thường là 1 trong 4 loại nucleotit được cải biến thành dạng được đánh dấu bằng cách gắn với các nhóm chất phát quang hoặc nguyên tử phóng xạ. Với phương pháp này, khoảng 25% nucleotit trong phân tử ADN được đánh dấu và điều đó đủ đáp ứng hầu hết các nhu cầu nghiên cứu khác nhau.

Các phân tử ADN đánh dấu với các tiền chất phát quang được phát hiện bằng cách chiếu xạ mẫu ADN với ánh sáng UV có bước sóng phù hợp và đo ở bước sóng phát xạ tương ứng. Các phân tử ADN được đánh dấu phóng xạ thường được phát hiện bằng cách chụp

mẫu ADN với phim tia X, hoặc đo bằng máy khuếch đại tín hiệu hạt β từ các nguyên tố phóng xạ ^{32}P và ^{35}S (đây là hai nguyên tố phóng xạ được sử dụng phổ biến để dánh dấu ADN).

Trong các nghiên cứu di truyền học phân tử hiện nay, có nhiều cách để xác định các phân đoạn ADN và ARN đặc hiệu dựa trên các phương pháp lai. ở đây, chúng ta chỉ đề cập đến hai phương pháp được dùng phổ biến:

Xác định các phân đoạn ADN và ARN bằng điện di và mẫu dò

Phương pháp sử dụng mẫu dò kết hợp với điện di là một phương pháp cơ bản có thể giúp xác định mức độ phổ biến hoặc kích thước của một đoạn trình tự ADN hoặc ARN được quan tâm nghiên cứu. Chẳng hạn như bằng kỹ thuật này, người ta có thể xác định và so sánh được mức biểu hiện của một gen ở các loại tế bào và mô khác nhau thông qua định lượng bản phiên mã mARN tương ứng của gen đó tại các tế bào và mô tương ứng; hay như để xác định kích thước của một đoạn ADN cắt giới hạn mang trình tự gen được quan tâm nghiên cứu.

Giả sử chúng ta cắt hệ gen của nấm men bằng enzym giới hạn *EcoRI* và cần xác định được kích thước của phân đoạn ADN cắt giới hạn mang trình gen A. Sản phẩm ADN tổng số sau khi được cắt bằng *EcoRI* sẽ tạo ra một số lượng lớn các phân đoạn có kích thước xấp xỉ 4 kb (vì $4^6 = 4096$ bp). Vì vậy, nếu đem sản phẩm cắt giới hạn nhuộm với EtBr, thì sản phẩm điện di sẽ là một dải các phân đoạn liên tục có kích thước xấp xỉ 4 kb, và không thể xác định được chính xác phân đoạn nào mang gen A. Trong trường hợp đó, kỹ thuật **thẩm tách Southern** (còn gọi là **lai Southern**) có thể được dùng để xác định phân đoạn mang gen đó. Lúc này, người ta sẽ đem sản phẩm cắt giới hạn ngâm vào một dung dịch có tính kiềm để làm biến tính các phân đoạn ADN sợi kép. Các phân đoạn này sau đó được chuyển sang một màng tích điện dương gọi là màng "thẩm tách" theo hình thức "đóng dấu". Nghĩa là các phân đoạn ADN định vị trên màng thẩm tách sẽ tương ứng với các phân đoạn định vị trên gel điện di. Các phân đoạn ADN gắn trên màng thẩm tách sau đó được ủ với mẫu dò là phân đoạn ADN chứa một đoạn trình tự đặc trưng bổ trợ với trình tự của gen A. Quá trình ủ được tiến hành trong điều kiện về nhiệt độ và nồng độ muối tương ứng với sự biến tính và hồi tính của axit nucleic. Trong điều kiện như đó, các mẫu dò sẽ chỉ lai đặc hiệu với phân đoạn ADN mang gen A. Do các phân đoạn gen có kích thước thường lớn hơn nhiều so với mẫu dò, nên khả năng hồi tính khó xảy ra hơn khả năng lai với mẫu dò. Sau đó, nhờ phương pháp phóng xạ tự chụp (mẫu dò được đánh dấu phóng xạ), các phân đoạn ADN bắt cặp với các mẫu dò có thể được xác định và phân lập rõ ràng. Các phân đoạn này chính là các phân đoạn mang trình tự gen A cần phân tích.

Một phương pháp tương tự như vậy cũng có thể được áp dụng trực tiếp để phân tích các sản phẩm phiên mã của gen là mARN, và được gọi là phương pháp **thẩm tách Northern** (**lai Northern**). Tuy vậy, vì so với ADN các phân tử mARN thường có kích

thước ngắn hơn (khoảng 5 kb), nên trong thẩm tách Northern, các phân tử mARN không cần cắt bằng enzym giới hạn (nếu có cắt thì số vị trí cắt giới hạn của một enzym trên phân tử mARN thường thấp). Các phân tử mARN sau khi phân tách bằng điện di được chuyển lên màng tích điện dương và lai với các mẫu dò ADN có trình tự bổ trợ tương ứng thường được đánh dấu phóng xạ (trong trường hợp này, sản phẩm lai là do sự kết cặp của các bazơ nitơ thuộc hai mạch ARN và ADN có trình tự bổ trợ).

Trong thực tiễn nghiên cứu, phương pháp thẩm tách Northern thường được sử dụng để định lượng một loại phân tử mARN nhất định nào đó có trong một mẫu phân tích, hơn là để xác định kích thước của nó. Lượng mARN được xác định được xem như thông số cơ bản phản ánh mức độ biểu hiện của gen mã hóa tương ứng. Chẳng hạn như, bằng phương pháp thẩm tách Northern, các nhà nghiên cứu có thể đánh giá được ảnh hưởng của một tác nhân phiên mã nào đó đến sự biểu hiện của một gen nhất định khi tiến hành so sánh lượng mARN do gen đó mã hóa có trong các tế bào được xử lý và không được xử lý với tác nhân phiên mã. Tương tự như vậy, kỹ thuật này cho phép xác định và so sánh mức độ biểu hiện của các gen khác nhau ở các loại tế bào, mô và cơ quan khác nhau của cùng một cơ thể trong cùng một giai đoạn hoặc ở các giai đoạn khác nhau của quá trình phát triển cơ thể.

Trong kỹ thuật thẩm tách Northern, mẫu dò thường được đưa vào phản ứng lai với một lượng dư vừa đủ để đảm bảo lượng phân tử lai tạo thành tương ứng với lượng mARN có mặt trong các mẫu nghiên cứu. Hay nói cách khác, qua lượng sản phẩm lai, có thể định lượng được lượng mARN.

Nguyên lý của các phương pháp lai Northern và Southern cũng chính là cơ sở của các kỹ thuật **phân tích gen bằng vi dãy phản ứng** (**microarray**) được phát triển và ngày các được sử dụng rộng rãi trong các nghiên cứu di truyền học phân tử gần đây. Trong các phương pháp microarray, các mẫu dò thường là các đoạn cADN được tạo ra từ việc phiên mã ngược các phân tử mARN tương ứng được tách chiết từ các mô hoặc tế bào. Các mẫu dò này sau đó được tiến hành lai với một dãy các phân tử ADN, trong đó mỗi dãy thường liên quan đến một hoặc một số gen khác nhau trong cơ thể sinh vật nghiên cứu. Cường độ biểu hiệu của sản phẩm lai (nhờ sự có mặt của các phân tử đánh dấu) ở các dãy khác nhau sẽ góp phần phản ánh mức độ biểu hiện khác nhau của các gen phân tích.

Tách dòng phân tử và xây dựng thư viện hệ gen

Khái niệm về tách dòng phân tử và thư viện hệ gen

Tách dòng phân tử (molecular cloning) là khái niệm chỉ một nhóm các phương pháp được sử dụng để: i) *phân lập một một trình tự gen* (ADN) đặc hiệu từ hỗn hợp các phân tử ADN ban đầu được tách chiết từ các mẫu sinh học vốn có cấu trúc phức tạp, kích

thước lớn; và ii) *khuếch đại (sao chép) trình tự* lên một số lượng lớn đủ để có thể tiến hành phân tích về cấu trúc và chức năng của gen tương ứng.

Khả năng tinh sạch các đoạn gen (ADN) đặc hiệu có số lượng đủ lớn là cần thiết để có thể "thao tác" các đoạn gen đó vì các mục tiêu nghiên cứu khác nhau. Chẳng hạn, từ các phân đoạn ADN được tách dòng, người ta có thể tạo ra các phân tử ADN tái tổ hợp mang các phân đoạn ADN mang nguồn gốc khác nhau. Các phân tử ADN tái tổ hợp mới có thể làm thay đổi mức độ biểu hiện bình thường của một gen (chẳng hạn bằng việc dung hợp giữa một trình tự mã hóa của một loài này với trình tự promoter của một loài khác) hoặc thậm chí mã hóa tổng hợp một loại protein "dung hợp" mới (protein lai) mang các trình tự axit amin từ các protein có nguồn gốc khác nhau. Hiện nay, các kỹ thuật tách dòng phân tử (bao gồm cả PCR) đã trở thành các công cụ thiết yếu trong nghiên cứu về sự điều hòa và biểu hiện của các gen và hệ gen ở các loài sinh vật khác nhau.

Quá trình tách dòng ADN và tạo nên các phân tử ADN tái tổ hợp điển hình thường liên quan đến việc sử dụng các **véctơ** là trình tự mang thông tin điều khiển hoạt động nhân lên (khuếch đại) và/hoặc biểu hiện trong tế bào của phân tử ADN tái tổ hợp mang **đoạn ADN cài** (đoạn trình tự được phân lập) bao gồm trình tự gen được quan tâm nghiên cứu. Các "công cụ" chính để tạo nên các phân tử ADN tái tổ hợp là các enzym giới hạn giúp cắt các phân tử ADN tại các vị trí xác định và các enzym nối cho phép ghép nối các phân đoạn ADN có nguồn gốc khác nhau với nhau. Bằng việc tạo nên các phân tử ADN tái tổ hợp có thể tự nhận lên trong tế bào chủ, một đoạn ADN cài xác định nào đó có thể được phân lập, tinh sạch và nhân lên thành một số lượng lớn các bản sao.

Tiếp theo đây, chúng ta sẽ mô tả bằng cách nào các phân tử ADN được cắt, tái tổ hợp và nhân lên, đồng thời đề cập đến việc xây dựng **thư viện hệ gen** gồm tập hợp các phân tử ADN lai chứa các đoạn cài xuất phát từ một hệ gen cần được thiết lập để phục vụ cho việc nghiên cứu, phân tích một hệ gen. Thông thường một thư viện hệ gen được thiết lập bằng việc sử dụng chung cùng một loại véctơ mang các phân đoạn ADN cài khác nhau. Chúng ta sẽ thấy bằng cách nào các phân đoạn ADN đặc hiệu có thể được xác định và phân lập từ các thư viện hệ gen.

Tách dòng ADN trong các véctơ plasmit

Sau khi một phân đoạn ADN được cắt khỏi một phân tử ADN có kích thước lớn hơn bằng enzym giới hạn, phân đoạn ADN đó cần được "cài" vào một véctơ để có thể nhân lên. Hay nói cách khác là một phân đoạn ADN cần được cài vào một phân tử ADN thứ hai (véctơ) để có thể nhân lên được trong tế bào chủ như đã nói ở trên. Cho đến nay, tế bào chủ được sử dụng rộng rãi nhất để nhân lên các đoạn ADN trong công nghệ ADN tái tổ hợp là vi khuẩn *E. coli*.

Các véctơ ADN điển hình thường có 3 đặc tính:

1. Chúng phải chứa một trình tự khởi đầu sao chép (tái bản), cho phép chúng tự sao chép độc lập với nhiễm sắc thể của tế bào chủ.
2. Chúng phải mang một dấu chuẩn chọn lọc cho phép dễ dàng xác định và phân lập được các tế bào mang véctơ tái tổ hợp (mang đoạn ADN cài) với các tế bào không mang véctơ tái tổ hợp.
3. Có vị trí cắt của một hoặc nhiều enzym giới hạn khác nhau. Đây chính là vị trí cài của phân đoạn ADN cần tách dòng vào véctơ.

Véctơ tách dòng phổ biến nhất là các phân tử ADN sợi kép, mạch vòng kích thước nhỏ (khoảng 3 kb) được gọi là các plasmit. Các véctơ này phần lớn có nguồn gốc từ các loài vi khuẩn và một số từ các sinh vật nhân chuẩn đơn bào (nấm men). Trong nhiều trường hợp, các phân tử ADN này trong tự nhiên đã mang sẵn các gen mã hóa tính kháng chất kháng sinh. Như vậy, các plasmid trong tự nhiên đã có sẵn hai thuộc tính là: khả năng tự sao chép trong tế bào chủ và có trình tự dấu chuẩn chọn lọc. Ngoài ra, các véctơ plasmit còn một ưu điểm nữa là chúng có thể đồng thời tồn tại nhiều bản sao trong tế bào. Điều này có thể giúp khuếch đại và phân lập được một số lượng lớn một phân đoạn ADN nào đó từ một quần thể tế bào nhỏ.

Trước đây, một số véctơ plasmit chỉ có một vị trí cắt của enzym giới hạn duy nhất. Cùng với thời gian, cấu trúc của các véctơ plasmit được cải tiến theo hướng cắt bỏ bớt các trình tự không cần thiết và gắn thêm vào đoạn trình tự có thể được cắt bằng nhiều loại enzym giới hạn khác nhau. Vị trí trên véctơ mang đoạn trình tự như vậy được gọi là **vị trí đa tách dòng** (polycloning site). Có những vị trí đa tách dòng hiện nay có kích thước ngắn nhưng có thể được cắt bởi trên 20 loại enzym giới hạn khác nhau. Nhờ đặc tính này, một véctơ có thể được dùng để tách dòng nhiều phân đoạn ADN có nguồn gốc khác nhau. Trên cơ sở các nguyên tắc tương tự, ngoài véctơ plasmit, hiện nay người ta đã phát triển được nhiều loại véctơ khác nhau có nguồn gốc phagơ, hoặc lai giữa vi khuẩn - phagơ (như phagemid, cosmid …), hoặc có nguồn gốc từ nấm men (ví dụ: YAC).

Việc cài một phân đoạn ADN vào một véctơ thường là một công việc tương đối đơn giản. Người ta thường sử dụng cùng một loại enzym giới hạn để cắt véctơ và đoạn ADN cài. Chẳng hạn như việc sử dụng enzym *EcoRI* sẽ cắt và chuyển véctơ từ dạng "vòng" sang dạng "mạch thẳng" có hai đầu dính. Do được cắt bởi cùng enzym giới hạn, nên đoạn ADN cài cũng có hai đầu dính với trình tự bổ trợ với trình tự đầu dính của véctơ. Các đầu dính này sẽ liên kết véctơ và đoạn ADN cài lại với nhau hình thành nên một phân tử ADN mạch vòng mới (phân tử ADN tái tổ hợp) chỉ còn thiếu hai liên kết phosphodieste duy nhất còn lại ở mỗi mạch. Hai liên kết này sẽ được "hàn" kín nhờ sử dụng enzym **ADN ligase** trong sự có mặt của ATP. Để hạn chế khả năng gắn kết lại của hai đầu dính của chính véctơ (vì hai đầu dính này có trình tự bổ trợ) sau khi được cắt bởi enzym giới hạn, người ta thường cho lượng ADN cài dư thừa so với véctơ để phần lớn các véctơ sau khi gắn lại là các véctơ tái tổ hợp mang các đoạn ADN cài.

Một số loại véctơ không những cho phép phân lập và tinh sạch được một phân đoạn gen nào đó, mà còn có thể điều hòa sự biểu hiện của gen nằm trong phân đoạn ADN cài. Những véctơ như vậy được gọi là các **véctơ biểu hiện**. Các véctơ này thường phải chứa trình tự promoter nằm ngược dòng sát với vị trí cài gen. Nếu vùng mã hóa của gen (không chứa promoter) được gắn vào véctơ theo đúng chiều khung đọc của gen, thì gen cài sẽ được phiên mã thành mARN và dịch mã thành protein trong tế bào chủ. Các véctơ biểu hiện thường được sử dụng để biểu hiện các gen đột biến hoặc các gen lai để tiến hành phân tích chức năng của chúng. Chúng cũng có thể được dùng để sản xuất một lượng lớn một loại protein nào đó vốn không thu được hiệu suất tương tự bằng các con đường tự nhiên. Ngoài ra, các promoter trong các véctơ biểu hiện có thể được lựa chọn sao cho sự biểu hiện của gen cài có thể được điều hòa bằng việc bổ sung một hợp chất đơn giản vào môi trường nuôi cấy (như một loại đường hoặc axit amin chẳng hạn). Việc có thể điều khiển chủ động sự biểu hiện của một gen nào đó có ý nghĩa quan trọng, đặc biệt đối với các gen gây độc.

Biến nạp các véctơ ADN vào tế bào chủ

Biến nạp là quá trình ở đó một cơ thể chủ có thể tiếp nhận một phân tử ADN ngoại lai từ môi trường bên ngoài. Một số vi khuẩn (trong đó không có *E. coli*) có khả năng biến nạp tự nhiên được gọi là các **vi khuẩn khả biến di truyền**. Vi khuẩn *E. coli* có thể trở nên khả biến khi được xử lý với ion Ca^{2+}. Mặc dù cơ chế khả biến chưa được biết đầy đủ, nhưng dường như ion Ca^{2+} bao bọc lại các điện tích âm trên phân tử ADN và tăng cường khả năng của chúng xuyên qua màng tế bào. Các tế bào được xử lý với Ca^{2+} vì vậy được gọi là các tế bào khả biến. Người ta có thể sử dụng một chất kháng sinh mà véctơ plasmid mang gen dấu chuẩn mã hóa tính kháng chất kháng sinh đó để chọn lọc được các thể mang plasmid tái tổ hợp. Các tế bào mang véctơ tái tổ hợp có thể sinh trưởng trong môi trường chứa chất kháng sinh, trong khi các tế bào khác thì không.

Thông thường quá trình biến nạp có hiệu suất không cao. Chỉ có một tỉ lệ nhỏ các tế bào được xử lý biến nạp có thể tiếp nhận được plasmid tái tổ hợp. Nhưng, chính hiệu quả biến nạp thấp giúp hầu hết các tế bào mang véctơ tái tổ hợp thường chỉ tiếp nhận một plasmid duy nhất. Thuộc tính này giúp cho các tế bào biến nạp và dòng tế bào do chúng sinh ra (do phân chia trực phân) chỉ mang một véctơ ADN tái tổ hợp duy nhất và cho phép các nhà nghiên cứu thể phân lập và tinh sạch được các gen hoặc hoặc sản phẩm của các gen riêng rẽ từ hỗn hợp biến nạp mang cả các phân tử ADN khác.

Thiết lập thư viện hệ gen

Đối với các hệ gen đơn giản, kỹ thuật tách dòng thường khá đơn giản. Chẳng hạn như với hệ gen một số virut có kích thước khoảng 10 kb, người ta có thể trực tiếp tách chiết ADN, cắt chúng bằng enzym giới hạn rồi tiến hành phân tích điện di. Các phân đoạn ADN tách biệt trên gel điện di được cắt khỏi gel, tinh sạch rồi cài vào các véctơ.

Trong khi đó, đối với các hệ gen phức tạp, kích thước lớn (như hệ gen người), việc cắt ADN tổng số bằng enzym giới hạn rồi phân tích điện di thường dẫn đến sự hình thành một dải băng điện di liên tục do có sự phân bố liên tục của các phân đoạn ADN được cắt giới hạn chỉ khác nhau một hoặc một vài nucleotit. Vì vậy, để đơn giản hóa quy trình tách dòng ở những hệ gen này, người ta thường tiến hành biến nạp toàn bộ các phân đoạn ADN (thu được sau khi cắt bằng enzym giới hạn) vào véctơ tách dòng rồi biến nạp tất cả chúng vào tế bào chủ, sau đó mới phân lập các dòng tế bào mang véctơ tái tổ hợp có các đoạn cài ADN khác nhau. Tập hợp các dòng tế bào như vậy được gọi là **thư viện hệ gen**. Như vậy, thư viện hệ gen là tập hợp các dòng tế bào mang các véctơ tái tổ hợp chứa các đoạn ADN cài khác nhau có cùng nguồn gốc (cùng hệ gen).

Để thiết lập một thư viện gen, hệ gen của tế bào đích (chẳng hạn ADN hệ gen người) được cắt bằng enzym giới hạn để tạo ra các phân đoạn ADN có kích thước trung bình mong muốn. Kích thước các phân đoạn (đoạn cài) có thể dao động từ 100 bp đến trên 1 Mb (đối với các phân đoạn ADN kích thước rất lớn, phân tử ADN thường được cắt không hoàn toàn bằng một enzym giới hạn). Các phân đoạn ADN cắt giới hạn sau đó được "trộn" với một loại véctơ phù hợp (trước đó được cắt bởi cùng loại enzym giới hạn) và ADN ligase. Kết quả của bước này sẽ tạo ra một tập hợp các véctơ mang các đoạn ADN cài khác nhau.

Người ta có thể tạo ra nhiều thư viện gen khác nhau bắt nguồn từ các nguồn vật liệu khác nhau. Thư viện hệ gen đơn giản nhất bắt nguồn từ ADN hệ gen tổng số được cắt bằng một enzym giới hạn duy nhất, gọi là **thư viện hệ gen**. Loại thư viện này có ý nghĩa ứng dụng rõ rệt nhất nhằm giải mã trình tự các hệ gen. Ngược lại, đối với mục đích tìm và phân lập ra một phân đoạn mang gen mong muốn, thì thư viện hệ gen chỉ tỏ ra hiệu quả đối với hệ gen chỉ chứa một phần tương đối nhỏ các vùng không mã hóa. Đối với các hệ gen phức tạp hơn, thư viện hệ gen không phù hợp cho việc tìm ra phân đoạn mang gen mong muốn bởi vì phần lớn các dòng trong thư viện mang các đoạn cài thuộc các vùng không mã hóa.

Để có được thư viện gen mang hầu hết các đoạn cài là các vùng mã hóa, người ta sử dụng thư viện cADN. Các bước thiết lập thư viện cADN được minh họa trên hình 10. Theo đó, trong bước đầu tiên thay vì bắt đầu từ ADN, người ta phiên mã ngược trình tự mARN thành trình tự ADN phiên bản (cADN). Quá trình này được gọi là quá trình phiên mã ngược và được thực hiện nhờ một enzym ADN polymerase đặc biệt là reverse transcriptase. Enzym này có khả năng tổng hợp ADN bắt nguồn từ phân tử ARN mạch đơn làm khuôn ban đầu. Khi có mặt enzym reverse transcriptase, các trình tự mARN được phiên mã ngược thành các phân tử ADN sợi kép, và những phân tử này có thể được gắn vào các véctơ.

Để có thể phân lập được các đoạn cài riêng rẽ từ thư viện hệ gen, các tế bào *E. coli* được tiến hành biến nạp với tất cả các phân đoạn trong thư viện. Mỗi một tế bào biến nạp thường chỉ mang duy nhất một véctơ mang đoạn cài ADN. Vì vậy, khi các tế bào nhân

lên sau đó chúng sẽ sẽ tạo ra nhiều dòng tế bào, mỗi dòng chứa nhiều bản sao của một phân đoạn trong thư viện cADN. Khuẩn lạc được tạo ra từ các tế bào mang các trình tự ADN mong muốn có thể được phân lập và từ đó thu lại ADN. Có một số cách để xác định các dòng tế bào đã mang gen biến nạp. Chẳng hạn phương pháp được nêu dưới đây sử dụng các mẫu dò ARN và/hoặc ADN để xác định các quần thể tế bào mang một đoạn ADN nhất định nào đó.

Sử dụng lai mẫu dò để xác định các dòng tế bào trong thư viện gen

Trong phương pháp sử dụng mẫu dò, người ta sử dụng các đoạn trình tự ADN/ARN có trình tự bổ trợ được đánh dấu và lai với các dòng tế bào mang các đoạn gen cài khác nhau trong thư viện hệ gen. Kỹ thuật này được gọi là **phương pháp lai khuẩn lạc**. Một thư viện hệ gen điển hình thường chứa hàng ngàn đoạn cài khác nhau được mang bởi cùng một loại véctơ tách dòng. Sau khi véctơ được biến nạp vào một chủng vi khuẩn phù hợp, các tế bào vi khuẩn được cấy trên bề mặt đĩa petri chứa môi trường bắn rắn chứa agar. Mỗi tế bào sau đó sẽ phát triển lên thành một khuẩn lạc riêng rẽ, và các tế bào trong cùng một khuẩn lạc đều mang cùng loại véctơ và đoạn gen cài giống nhau.

Trong kỹ thuật lai khuẩn lạc, người ta cũng có thể sử dụng loại màng lai được dùng trong các phương pháp thẩm tách Southern và Northern để thu hồi được một lượng "vết" ADN đủ cho sự kết cặp với mẫu dò. ở đây, người ta sẽ dùng màng lai ép lên bề mặt đĩa nuôi cấy chứa khuẩn lạc và in hình chúng lên màng lai (cùng với ADN của chúng) sao cho vị trí của các dòng tế bào trên màng lai tương ứng với các vị trí khuẩn lạc của chúng trên đĩa petri (theo nguyên tắc "đóng dấu"). Điều này đảm bảo cho việc khi chúng ta xác định được một vị trí trên màng lai có kết quả "dương tính" và việc bắt cặp với mẫu dò thì chúng ta sẽ xác định được tương ứng khuẩn lạc mang dòng tế bào chứa véctơ tái tổ hợp mang đoạn gen cài mong muốn.

Màng lai được đem lai với mẫu dò như sau: người ta tiến hành xử lý màng lai sao cho màng tế bào vỡ ra và các phân tử ADN thoát ra ngoài gắn lên màng lai tại chính vị trí tế bào của chúng. Các màng lai sau đó được ủ với các mẫu dò được đánh dấu từ trước trong các điều kiện giống như khi tiến hành các kỹ thuật thẩm tách Northern hay Southern.

Trong công nghệ ADN tái tổ hợp, ngoài các véctơ plasmit có nguồn gốc vi khuẩn, người ta còn có thể sử dụng véctơ có nguồn gốc virut (bacteriophage), hoặc từ các sinh vật bậc cao hơn như nhiễm sắc thể nhân tạo nấm men (YAC), hay nhiễm sắc thể nhân tạo có nguồn gốc vi khuẩn (BAC), hoặc một số véctơ lai giữa chúng (vd: cosmid, …). Trong các véctơ có nguồn gốc bacteriophage, phage λ được sử dụng phổ biến. ADN của virut này được cải biến và sử dụng như véctơ tách dòng. Véctơ này được sử dụng để tách dòng các thư viện hệ gen về nguyên tắc giống hệt như khi sử dụng các véctơ plasmit. Chỉ có một điểm khác lai khi tiến hành lai để xác định các dòng gen biến nạp thì vị trí các mẫu dò được xác định trên màng lai tương ứng với vị trí các vết tan thay cho vị trí các khuẩn lạc như khi sử dụng véctơ tách dòng plasmit.

Tổng hợp hóa học và sử dụng các đoạn oligonucleotit

Các đoạn ADN có trình tự xác định kích thước ngắn được sử dụng nhiều trong các nghiên cứu khác nhau của di truyền học phân tử, chẳng hạn như sử dụng làm mồi trong các phản ứng PCR, hay dùng làm mẫu dò trong các phương pháp lai phân tử. Các đoạn trình tự này thường được tổng hợp theo phương pháp hóa học và được gọi là các **đoạn oligonucleotit**. Phương pháp hóa học được sử dụng phổ biến nhất được tiến hành trên bề mặt cứng sử dụng máy tự động. Tiền chất để tạo nên các nucleotit lần lượt gắn với đoạn oligonucleotit theo trật tự nhất định được gọi là các hợp chất **phosphoamidine** (xem minh họa hình 11). Phản ứng kéo dài chuỗi oligonucleotit diễn ra bằng việc gắn thêm tiền chất nucleotit mới vào phía đầu 5', như vậy ngược chiều với phản ứng kéo dài chuỗi ADN sử dụng enzym ADN polymerase.

Phương pháp tổng hợp hóa học có hiệu quả và độ chính xác cao khi tiến hành tổng hợp các phân tử ADN mạch đơn có độ dài tới 30 nucleotit. Với các thiết bị tổng hợp oligonucleotit hiện nay, một người nghiên cứu có thể dễ dàng tổng hợp bất cứ một trình tự ADN ngắn nào đơn giản bằng cách gõ và nhập trình tự nucleotit mong muốn vào phần mềm máy tính điều khiển máy tổng hợp tự động. Tuy vậy, khi phân tử ADN được tổng hợp có kích thước lớn thì độ chính xác của trình tự và độ đồng đều của sản phẩm tạo thành giảm đi do các lỗi cố hữu mang tính kỹ thuật. Thực tế, các phân tử ADN có trình tự dài hơn 100 nucleotit khó có thể được tổng hợp bằng các phương pháp hóa học mà đảm bảo được số lượng và chất lượng mong muốn.

Các đoạn oligonucleotit kích thước ngắn ngoài các ứng dụng làm mồi phản ứng PCR hay làm mẫu dò như được nêu ở trên, còn có thể được sử dụng cho nhiều mục đích khác trong nghiên cứu di truyền học phân tử. Chẳng hạn như các đoạn oligonucleotit kết cặp sai với một đoạn ADN được tách dòng có thể được dùng để tạo ra một đột biến định vị trí. Phương pháp này, gọi là **phương pháp gây đột biến định vị trí**, được thực hiện như sau: một trình tự nucleotit nhất định được tiến hành lai với một phân đoạn ADN, ở đây đoạn oligonucleotit được sử dụng làm mồi còn phân đoạn ADN đích được dùng làm khuôn. Trong đoạn ADN sợi kép hình thành sẽ có một vị trí kết cặp sai, và theo nguyên tắc PCR, các phân đoạn ADN được tạo ra trong các phản ứng về sau đều mang đột biến tại vị trí kết cặp sai.

Các đoạn oligonucleotit cũng có thể được sử dụng theo cách tương tự như vậy để tạo nên một điểm cắt giới hạn mới trong một phân tử ADN, để rồi sau đó điểm cắt giới hạn này được sử dụng để cài đoạn ADN đích vào giữa một trình tự mã hóa và một trình tự không mã hóa, hoặc sau một trình tự promoter hay vị trí gắn của ribosom, v.v…

Như vậy, rõ ràng trong các nghiên cứu di truyền học phân tử, việc các đoạn oligonucleotit có thể được tổng hợp theo một trình tự bất kỳ có nhiều ý nghĩa ứng dụng quan trọng trong việc xác định và khuếch đại các đoạn ADN đặc hiệu, để giải mã trình

tự ADN, cũng như trong các nghiên cứu tạo đột biến điểm xác định vị trí, hay sử dụng cho công nghệ ADN tái tổ hợp.

Phản ứng PCR

Ngoài các kỹ thuật tách dòng và khuếch đại gen sử dụng các loại tế bào chủ, một phương pháp khuếch đại gen có hiệu quả cao giờ đây đã trở thành một kỹ thuật phổ biến trong hầu hết các nghiên cứu di truyền học phân tử là kỹ thuật **phản ứng chuỗi trùng hợp PCR** (polymerase chain reaction). Đây là một kỹ thuật hóa sinh invitro thuần túy. Kỹ thuật PCR sử dụng enzym ADN polymerase để tổng hợp nên các phân tử ADN mới từ trình tự của phân tử ADN làm khuôn với tiền chất là các deoxyribonucleotit. Như đã nêu ở Chương 2, các enzym ADN polymerase tổng hợp ADN theo chiều 5'→ 3' và có thể xúc tác gắn nucleotit tiếp theo vào phía đầu 3' của một trình tự oligonucleotit có sẵn. Nghĩa là, nếu ta có một đoạn trình tự oligonucleotit đã gắn vào một mạch ADN làm khuôn có trình tự bổ sung với trình tự của đoạn oligonucleotit, thì enzym ADN polymerase có thể sử dụng đoạn oligonucleotit đó làm đoạn mồi và tiếp tục kéo dài chuỗi ADN về phía đầu 3' dựa trên trình tự của mạch ADN làm khuôn.

Vậy, bằng cách nào người ta có thể sử dụng phản ứng PCR và enzym ADN polymerase để khuếch đại một đoạn trình tự ADN đặc hiệu? Người ta sẽ tổng hợp và sử dụng hai đoạn oligonuleotit. Đoạn thứ nhất có trình tự bổ sung với đầu 5' của một mạch phân đoạn ADN cần khuếch đại (đoạn này gọi là **mồi xuôi**), còn đoạn trình tự thứ hai bổ sung với đầu 5' của mạch đối diện (**mồi ngược**). Phân tử ADN làm khuôn sẽ được làm biến tính (bởi nhiệt) và các mồi xuôi và môi ngược sẽ gắn vào trình tự bổ sung ở hai đầu đoạn ADN cần khuếch đại. Với sự có mặt của các cơ chất là các deoxyribonucleotit, enzym ADN polymerase sẽ tổng hợp và kéo dài một mạch phân tử ADN bắt đầu từ hai đoạn mồi.

Trong chu kỳ tiếp theo, phân tử ADN lại được gây biến tính và quá trình tổng hợp ADN được lặp lại với cùng cặp mồi. Kết quả của chu kỳ thứ hai tạo ra 4 bản sao của phân đoạn gen mong muốn. Theo cơ chế đó, chu kỳ biến tính và tổng hợp ADN diễn ra lặp đi lặp lại sẽ làm khuếch đại phân đoạn ADN nằm giữa 2 trình tự mồi theo cấp số nhân (số bản sao tương ứng qua các chu kỳ sẽ lần lượt là 2, 4, 8, 16, 32, 64, v.v…). Như vậy, chỉ cần từ một bản sao ADN duy nhất có mặt trong một lượng mẫu nhỏ, chúng ta có thể thu được một lượng lớn bản sao xuất phát từ bản sao đầu tiên.

Xét về một khía cạnh nào đó, kỹ thuật tách dòng ADN và PCR đều được dùng để khuếch đại một phân đoạn ADN đặc thù lên một số lượng lớn. Nhưng điểm khác biệt là ở chỗ, trong kỹ thuật tách dòng, chúng ta thường sử dụng một hóa chất chọn lọc hay một thiết bị nào đó để định vị được trình tự ADN đã được khuếch đại trong một thư viện ADN đã có sẵn gồm nhiều dòng tế bào khác nhau, trong khi ở kỹ thuật PCR, hóa chất chọn lọc chính là cặp mồi sẽ giúp hạn chế phản ứng khuếch đại chỉ tập trung vào đoạn ADN được quan tâm ngay từ đầu.

Giải mã trình tự ADN

Trong phần này chúng ta sẽ xem xét bằng cách nào có thể xác định được trình tự nucleotit của các phân đoạn hoặc toàn bộ phân tử ADN mong muốn. Về một khía cạnh nào đó, có thể coi giải mã trình tự các nucleotit là việc đánh dấu mẫu dò triệt để nhất của một hệ gen với tính chọn lọc cao. Chúng ta sẽ xác định toàn bộ trình tự hệ gen của các cơ thể sinh vật có mức độ cấu tạo phức tạp khác nhau từ vi khuẩn cho đến loài người, và điều này cho phép chúng ta tìm thấy mọi trình tự đặc hiệu một cách nhanh và chính xác thông qua việc sử dụng các phần mềm máy tính với các thuật toán phù hợp. Hay nói cách khác, "các chất chọn lọc" của chúng ta ở đây là các chuỗi bazơ nitơ được chúng ta nhập vào phần mềm máy tính. Do cơ sở dữ liệu về các hệ gen ngày càng trở nên phong phú, nên ngày càng trở nên dễ dàng hơn để có thể tìm thấy các bản sao của trình tự các hệ gen hoặc của các trình tự có liên quan trong cùng một loài hoặc của các loài khác. Rõ ràng, việc giải mã trình tự các nucleotit đã tạo ra một cơ sở dữ liệu khổng lồ phục vụ cho các nghiên cứu giải mã trình tự và so sánh giữa các hệ gen được đề cập dưới đây.

Nguyên tắc giải mã trình tự ADN về cơ bản dựa trên việc phân tách các phân đoạn ADN có kích thước khác nhau được giới hạn bởi hai đầu. Các phân tử ADN đều giống nhau ở phần đầu 5', nhưng kết thúc ở phía đầu 3' có các nucleotit khác nhau. Các thành viên của một nhóm sẽ có nucleotit ở phía đầu 3' giống nhau. Như vậy, trong một nhóm sẽ bao gồm tất cả các phân tử ADN tận cùng đầu 3' bằng G, nhóm khác tương ứng là A, C và T. Trong mỗi nhóm các phân tử sẽ có kích thước khác nhau phụ thuộc vào vị trí của nucleotit tương ứng (ví dụ như G) nằm trên phân tử ADN. Các phân đoạn khác biệt về chiều dài như vậy có thể phân tách được nhờ sử dụng kỹ thuật điện di trên gel polyacrylamid. Chẳng hạn khi chạy hỗn hợp các phân tử ADN tận cùng đầu G ta sẽ thu được thang các băng điện di tương ứng với các phân đoạn, trong đó mỗi băng tương ứng với một phân đoạn có chiều dài phản ánh vị trí của nucleotit G trên phân tử ADN….

Giải mã trình tự hệ gen vi khuẩn bằng kỹ thuật shotgun ("giải mã từng đoạn ngẫu nhiên")

Vi khuẩn gây bệnh kiết lị ở người *Hemophilus influenza* là loài sinh vật đầu tiên được giải mã toàn bộ hệ gen. Sở dĩ hệ gen của loài này được hoàn thành việc giải mã đầu tiên là nhờ hệ gen của nó nhỏ, chỉ chứa một phân tử ADN duy nhất kích thước 1,8 Mb. Hệ gen của vi khuẩn này được "cắt" thành các phân đoạn nhỏ có kích thước trung bình khoảng 1 kb. Các đoạn ADN hệ gen này sau đó được tách dòng bằng các véctơ ADN plasmit tái tổ hợp. ADN từ các dòng vi khuẩn chứa các phân đoạn ADN tái tổ hợp riêng rẽ rồi được giải mã trình tự riêng rẽ trên các máy giải mã trình tự tự động sử dụng phương pháp ddNTP như mô tả ở phần trên. Phương pháp này được gọi là phương pháp giải mã trình tự kiểu "shotgun" (bắn ngẫu nhiên). Các khuẩn lạc mang các véctơ tơ tái tổ hợp mang đoạn ADN cài ngẫu nhiên được phân lập, xử lý và giải mã trình tự. Để chắc chắn rằng mọi nucleotit trong hệ gen vi khuẩn đều có mặt trong các dòng vi khuẩn của thư viện hệ gen, tổng cộng có khoảng 30.000 - 40.000 dòng tái tổ hợp khác nhau được

sử dụng và giải mã trình tự. Từ đó, tạo ra khoảng 20 Mb dữ liệu thô về hệ gen (các phản ứng tạo ra trình tự có kích thước trung bình 600 bp, và 20 Mb = 600 bp x 33.000 dòng vi khuẩn). Dữ liệu này được gọi là **vùng trình tự 10x**. Bởi vì, mỗi nucleotit trong hệ gen được đọc lặp lại khoảng 10 lần.

Phương pháp này dường như là tốn nhiều công sức, nhưng chi phí rẻ hơn và nhanh hơn so với các phương pháp truyền thống khác. Một phương pháp giải mã trình tự trước đây dựa trên nguyên tắc giải mã từng phân đoạn ADN cắt giới hạn trên bản đồ vật lý của nhiễm sắc thể vi khuẩn. Một hạn chế của kỹ thuật này là hầu hết các phân đoạn cắt giới hạn có kích thước lớn hơn kích thước có thể giải mã trình tự hoàn toàn trong mỗi phản ứng được thực hiện. Do vậy, để giải mã toàn bộ hệ gen, người ta phải tiến hành cắt giới hạn, lập bản đồ và giải mã trình tự nhiều lần. Các bước này nếu lặp đi lặp lại nhiều lần sẽ tốn nhiều thời gian hơn khi sử dụng phương pháp giải mã trình tự tự động của các phân đoạn ADN ngẫu nhiên. Hay nói cách khác, nhờ sử dụng phần mềm máy tính việc sắp xếp lại các phân đoạn ADN ngẫu nhiên vẫn nhanh hơn nhiều việc lập bản đồ các phân đoạn cắt giới hạn trên NST vi khuẩn.

Khoảng 30.000 đoạn trình tự ADN được giải mã trình tự ngẫu nhiên được trực tiếp nhập vào phần mềm máy tính. Nhiều phần mềm máy tính chuyên dụng hiện nay có thể xếp các đoạn trình tự theo đúng thứ tự dựa trên các trình tự gối lên nhau của chúng. Sự "lắp ráp" thành trình tự của các phân đoạn ADN ngắn cuối cùng sẽ có một trình tự liên tục duy nhất, còn được gọi là một **contig**.

Kỹ thuật giải mã trình tự kiểu shotgun cho phép "ráp nối" từng phần của hệ gen lớn

Như đã trình bày ở trên việc giải mã các đoạn trình tự ADN kích thước khoảng 600 bp hiện nay có thể thực hiện một cách tương đối đơn giản và nhanh chóng. ở đây, chúng ta sẽ xem bằng cách nào kỹ thuật "shortgun" được áp dụng để giải mã trình tự các hệ gen lớn.

Chẳng hạn, nhiễm sắc thể người có kích thước trung bình khoảng 150Mb. Do vậy, mỗi đoạn trình tự ~600 bp được giải mã chỉ chiếm 0,0004% của mỗi NST. Kết quả là để có thể xác định được trình tự đầy đủ của một NST, người ta cần tạo ra một số lượng lớn các dữ liệu trình tự từ nhiều phân đoạn ADN ngắn (hình 12). Các phân đoạn ADN nhỏ được tạo ra từ 23 NST của hệ gen người, rồi sau đó được cắt ngắn thành một thư viện các đoạn ADN nhỏ bằng một kỹ thuật "kim áp lực". Thông thường, có 2 hoặc 3 thư viện hệ gen chứa các đoạn trình tự có kích thước khác nhau (tăng dần) được tạo ra, chẳng hạn tương ứng với các đoạn trình tự có kích thước 1, 5 và 100 kb. Các phân đoạn này sau đó được tách dòng ngẫu nhiên vào các plasmit của vi khuẩn theo phương pháp được mô tả ở trên.

Các phân tử ADN tái tổ hợp mang các phân đoạn ngẫu nhiên của NST người sau đó được phân lập từ các plasmit vi khuẩn rồi giải mã bằng máy giải mã trình tự tự động. Để

đảm bảo mọi nucleotit trong hệ gen đều được giải mã, người ta phải tiến hành giải mã riêng rẽ khoảng 2 triệu phân đoạn ADN khác nhau. Với kích thước của mỗi phân đoạn có thể giải mã chính xác khoảng 600 bp, quy trình này tạo ra dữ liệu khoảng 1 tỉ bp, hay nói các khác là gấp ~10 lần kích thước trung bình của một NST. Như đã trình bày ở trên với kỹ thuật giải mã trình tự ở vi khuẩn, việc phân tích các mẫu với lượng trình tự gấp khoảng 10 lần lượng ADN thực cần giải mã trình tự sẽ đảm bảo mọi phần của NST đều được phân tích.

Quá trình tạo ra các thư viện tái tổ hợp mang các trình tự ngẫu nhiên và một lượng lớn ADN cần phải giải mã trình tự ngẫu nhiên dường như là một việc làm rất lãng phí. Tuy vậy, với việc sử dụng hệ thống một trăm máy giải mã trình tự tự động gồm 384 cột sẽ cho phép phân tích 10 lần một nhiễm sắc thể người chi tiết trong vòng 3 tuần. Phương pháp này vì vậy vẫn nhanh hơn nhiều phương pháp phân lập từng phần đã biết trong NST, rồi sau đó giải mã trình tự một tập hợp đã biết của các đoạn ADN được đặt so le. Vì vậy, bản chất của công nghệ cốt lõi được sử dụng để thúc đẩy việc giải mã hệ gen người dựa trên **kĩ thuật giải mã trình tự ngẫu nhiên tự động**, rồi sau đó sử dụng phần mềm máy tính để sắp xếp lại các đoạn ADN khác nhau giống như trò chơi "ghép hình" vậy. Việc kết hợp sử dụng máy giải mã trình tự tự động với phần mềm máy tính đã giúp dự án giải mã toàn bộ hệ gen người kết thúc sớm hơn nhiều năm so với kế hoạch ban đầu.

Các chương trình máy tính phức tạp được sử dụng để tập hợp các đoạn ADN ngắn được giải mã trình tự ngẫu nhiên thành những đoạn trình tự dài kích thước lớn kế tiếp nhau được gọi là những contig. Các đoạn trình tự nằm gối lên nhau sẽ được phần mềm xử lý rồi nối lại với nhau thành các trình tự lớn hơn. Kích thước của các đoạn contig phụ thuộc vào lượng trình tự đã được giải mã. Nếu lượng trình tự giải mã càng nhiều, thì các đoạn contig càng có kích thước lớn và khoảng cách trống chưa được giải mã càng nhỏ.

Thông thường các đoạn contig riêng rẽ thường có kích thước 50.000 - 200.000 bp. Nghĩa là ngắn hơn nhiều so với kích thước NST ở người. Tuy vậy, các đoạn contig rất hiệu quả khi phân tích các hệ gen nhỏ. Chẳng hạn, hệ gen của ruồi dấm (*Drosophila*) trung bình có mật độ 1 gen / 10 kb. Vì vậy, một contig điển hình thường chứa vài gen liên kết với nhau. Rất tiếc là các hệ gen lớn lại thường chứa mật độ gen thấp. Hệ gen người có mật độ trung bình là 1 gen / 100 kb, vì vậy một contig điển hình thường không chứa được trình tự trọn vẹn của một gen, chứ chưa nói đến là một dãy gen liên kết. Bây giờ, chúng ta sẽ nói đến bằng cách nào các đoạn contig tương đối ngắn có thể được lắp ráp lại thành các **đoạn khung** có kích thước 1-2Mb.

Phương pháp giải mã trình tự đầu cuối cho phép lắp ráp các contig thành các đoạn khung ở các hệ gen kích thước lớn

Một khó khăn lớn gặp phải khi thiết lập các đoạn contig là sự xuất hiện của các đoạn ADN lặp lại. Các đoạn trình tự này làm việc ráp nối trở nên khó khăn và phức tạp do

các đoạn ADN không liên kết (từ các NST khác nhau) nhưng có thể bị xếp thành các đoạn trình tự nằm gối lên nhau do chúng có cùng trình tự lặp lại. Một phương pháp được sử dụng để khắc phục trở ngại này là kĩ thuật **giải mã phần nối trình tự đầu cuối**. Kỹ thuật này tương đối đơn giản nhưng hiệu quả mà nó mang lại cao.

Ngoài việc ADN hệ gen được dùng để tạo nên một thư viện các đoạn ADN ngắn nhằm giải mã trình tự ngẫu nhiên, thì chính ADN hệ gen đó đồng thời được dùng để tạo nên các đoạn ADN tái tổ hợp mang các đoạn có kích thước lớn, thường có kích thước 3 - 100 kb. Giả sử chúng ta có một mẫu ADN từ một NST người. Một phần của mẫu này được dùng để tạo nên các phân đoạn có kích thước 1 kb, trong khi một phần khác được dùng để tạo nên các phân đoạn có kích thước 5 kb. Kết quả của quá trình đó là người ta thu được 2 thư viện hệ gen khác nhau, một mang các đoạn cài kích thước ngắn, còn thư viện kia là các đoạn cài kích thước lớn (hình 12).

Tiếp theo, người ta sử dụng các đoạn mồi "đa năng" (có tính chọn lọc thấp) có thể gắn vào phần đoạn nối giữa plasmit và hai vùng biên của đoạn ADN cài kích thước lớn. Mỗi một phản ứng giải mã trình tự cho phép tạo ra thông tin về trình tự của một đoạn kích thước khoảng 600 bp ở hai đầu của một đoạn cài bất kỳ. Một bản ghi nhớ sẽ ghi chép lại các trình tự ở hai đầu của cùng một phân đoạn kích thước lớn. Việc dùng phần mềm sau đó cho thấy một trình tự được tìm thấy ở contig A, còn trình tự kia được tìm thấy ở contig B. Nếu contig A và B cùng có các trình tự có mặt trong một phân đoạn kích thước khoảng 5 kb thì có thể giả thiết chúng cùng xuất xứ từ một vùng của một NST. Trong khi đó hầu hết các phân đoạn ADN lặp lại thường có kích thước nhỏ hơn 2-3 kb. Vì vậy, các đoạn trình tự ADN đầu cuối xuất xứ từ các đoạn cài ~5kb là đủ để nối các contig bị ngắt quãng bởi các đoạn ADN có trình tự lặp lại.

Các nghiên cứu ban đầu thường chỉ tạo ra các đoạn contig có kích thước nhỏ hơn 500 kb. Để thu được dữ liệu từ các đoạn có trình tự dài, có kích thước vài Mb hoặc dài hơn, người ta cần dữ liệu từ các trình tự đầu cuối từ các phân đoạn ADN lớn có kích thước ít nhất là 100 kb. Các đoạn ADN này có thể thu được từ bằng một véctơ tách dòng đặc biệt gọi là **nhiễm sắc thể nhân tạo vi khuẩn - BAC** (*bacterialartificialchromosome*). Nguyên tắc các đoạn này được dùng để tạo nên thông tin của các trình tự dài là giống như trường hợp sử dụng các đoạn 5 kb được mô tả ở trên. Các đoạn mồi được dùng để xác định trình tự ~600kb ở hai đầu của đoạn cài BAC. Việc sử dụng BAC cho phép sắp xếp nhiều đoạn contig khác nhau vào cùng một đoạn khung duy nhất có kích thước lớn tới vài Mb (hình 13).

Chất lượng của việc ráp nối hệ gen là một phép đo kích thước đoạn khung trung bình. Những đoạn khung nào có kích thước từ 1 Mb trở lên được tìm thấy được xem là có kết quả ráp nối tốt. Ví dụ như, ở loài cá bể dẹt (Tetraodontidae) có kích thước hệ gen 800 Mb, và trình tự ráp nối của toàn hệ gen này gồm 500 đoạn khung khác nhau, như vậy mỗi đoạn khung có kích thước trung bình 1,6 Mb. Một "hiệu quả" ráp nối cao như vậy cũng tạo thuận lợi cho nhiều phân tích di truyền khác, chẳng hạn như có thể dễ dàng xác

định được tất cả các vùng mã hóa của hệ gen. Đến năm 2000, kích thước trung bình của các đoạn khung được xây dựng cho hệ gen người có kích thước là 2 Mb. Điều này là đủ để có thể tin cậy về số gen ước lượng có trong hệ gen (xấp xỉ 30.000 gen).

Phân tích mở rộng hệ gen

Đối với các hệ gen nhỏ như của vi khuẩn hay các loài sinh vật nhân chuẩn đơn giản, việc xác định các trình tự mã hóa protein thường có thể ngoại suy trực tiếp từ kết quả giải mã trình tự, mà thực chất là thông qua việc xác định các ORF. Mặc dù không phải tất cả các ORF (đặc biệt là các ORF ngắn) đều thực sự là các gen mã hóa protein, thì việc xác định như vậy thường cũng rất hiệu quả, việc khó khăn hơn thường là việc xác định được chức năng của các gen đó hoặc sản phẩm (protein) của nó.

Việc xác định được vùng mã hóa protein ở hệ gen các loài động vật vốn phổ biến chứa cấu trúc exon - intron thực tế phức tạp hơn nhiều. Trong trường hợp này, người ta phải sử dụng "một loạt" các công cụ **tin sinh học** để xác định được các gen và thành phần di truyền của các hệ gen phức tạp. Các chương trình máy tính đã được lập trình để có thể xác định được các vùng có tiềm năng mã hóa protein dựa trên một số tiêu chí nhất định, bao gồm sự xuất hiện của các ORF được chặn bởi các vị trí cắt ở hai đầu và gần kề một trình tự khởi đầu phiên mã (promoter). Tuy vậy, các chương trình phân tích gen này đến nay vẫn chưa hoàn thiện để có thể khẳng định sự chính xác là 100%. Một tỉ lệ khoảng 3/4 số gen có thể được xác định bằng phương pháp này, nhưng cũng có rất nhiều gen bị bỏ sót; và thậm chí chi tiết hơn trong một gen, một số trình tự exon cũng có thể bị bỏ sót.

Một hạn chế đáng kể nữa của **các chương trình tìm gen** hiện nay là đôi khi không xác định được đầy đủ các promoter. Ví dụ như một promoter lõi điển hình ở động vật đa bào có kích thước khoảng 60 bp, chứa các trình tự định dạng (motif), như TATA, INR và DPE, là những motif cần thiết cho sự gắn vào của phức hệ khởi động TFIID và phức hệ phiên mã của enzym ARN Polymerase II. Đáng tiếc là trình tự của yếu tố khởi đầu phiên mã lõi này có mức độ biến đổi rất lớn. Mặc dù trong khi phức hệ khởi đầu phiên mã của tế bào đủ "thông minh" để xác định được những trình tự này, thì đến nay con người chưa viết được các chương trình máy tính cho phép xác định được đầy đủ các promoter lõi dạng này. Tất nhiên, hiện nay các nhà sinh tin học đang tiếp tục hoàn thiện các chương trình phần mềm để đến một ngày nào đó chúng ta có thể xác định, phân tích được tất cả các thuộc tính của gen đã nêu ở trên, bao gồm các yếu tố promoter lõi, các ORF, các điểm cắt và tái tổ hợp gen, v.v… để xác định được đúng và đầy đủ các gen mã hóa protein.

Phương pháp quan trọng nhất để kiểm chứng các gen mã hóa protein suy đoán và xác định các gen bị bỏ sót bởi các phần mềm máy tính là sử dụng dữ liệu cADN. cADN được tạo ra theo nguyên tắc phiên mã ngược từ các phân tử mARN hoàn thiện, vì vậy nó phản ánh đúng các trình tự exon thực sự. Các phân tử cADN được dùng để tạo ra cơ sở

dữ liệu **EST**, hay còn gọi là **nhãn xác định trình tự biểu hiện** (*expressedsequencetag*), thực chất là các đoạn trình tự ngắn được trích ra từ một trình tự cADN đã biết. Các trình tự cADN ngẫu nhiên (có thể là trình tự đầy đủ hay các trình tự một phần EST) được xác định bằng sử dụng phương pháp giải mã trình tự ngẫu nhiên rồi được đối chiếu với các đoạn khung của hệ gen. Các vùng tương ứng với các EST được xác định là các exon, còn các vùng nằm giữa các exon tương ứng với các intron (mặc dù, nguyên tắc cắt intron khác nhau có thể sử dụng một exon không có mặt trong cADN hay EST được giải mã trình tự). Các thông tin giải mã trình tự cADN và EST cũng giúp tìm được sự liên kết giữa các contig, giữa các đoạn khung và giữa chúng với nhau. Chẳng hạn như giả sử có một phân tử cADN được phiên mã từ một gen kích thước rất lớn có chiều dài intron là 100 kb hoặc hơn. Có hai đoạn khung cùng chứa các trình tự khác nhau của phân tử cADN chung này, thì nhiều khả năng chúng là các vùng liên kết của hệ gen và biểu hiện là các đoạn của cùng một gen.

Phân tích so sánh các hệ gen

Việc so sánh hệ gen giữa các loài động vật khác nhau là cơ sở trực tiếp để đánh giá sự biến đổi về cấu trúc gen và trình tự của chúng xuất hiện trong quá trình tiến hóa. Việc so sánh các hệ gen như vậy đồng thời cùng giúp khẳng định chắc chắn hơn về các vùng gen mã hóa protein trong một hệ gen của loài nào đó. Ví dụ như các exon của các gen đồng tiến hóa có mức độ bảo thủ cao hơn nhiều so với các intron. Việc so sánh hệ gen người và chuột đã tìm thấy nhiều exon có tính bảo thủ cao. Việc so sánh giữa các hệ gen cũng đồng thời giúp xác định các trình tự exon ngắn (hay tìm thấy ở phần đầu 5’ của gen và vùng promoter lõi) vốn thường bị sót khi xác định bằng phần mềm máy tính.

Một trong những khám phá nổi bật của phép phân tích so sánh các hệ gen là việc tìm ra sự phổ biến của **tính bảo thủ liên kết giữa các gen trên cùng NST**. ở người và chuột, sự bảo thủ của tính liên kết giữa các gen trên cùng NST là rất phổ biến. Trong nhiều trường hợp, tính bảo thủ này được tìm thấy ở cả các loài rất xa nhau trong quá trình tiến hóa, ví dụ như ở loài cá bể dẹt có tổ tiên chung với các loài động vật có vú từ 400 triệu năm trước đây. Hiện tượng phổ biến của sự bảo thủ trong tính liên kết của nhiều gen cho thấy có nhiều khả năng các gen “láng giềng” cùng dùng chung các trình tự điều hòa gen. Một điều tra dùng phần mềm máy tính gần đây tìm thấy trong một đoạn NST có kích thước 100 - 200 kb ở ruồi dấm *Drosophila* có 10 - 20 gen liên kết có hình thức điều hòa sự biểu hiện giống hệt nhau. ở ruồi dấm có khoảng 500 - 1000 đoạn NST duy trì sự liên kết bảo thủ này có thể là do các gen liên kết cùng phụ thuộc vào các trình tự điều hòa chung ở vùng NST đó.

Các trình tự mã hóa protein không chỉ là các vùng của hệ gen được giới hạn về chức năng. Các trình tự điều hòa (vị trí gắn của các yếu tố phiên mã và các yếu tố điều hòa hoạt động gen, như các yếu tố tăng cường enhancer) thường có tính bảo thủ cao. Các trình tự này thường được xác định là các trình tự không mã hóa protein ngắn và bảo thủ. Ví dụ một chương trình máy tính gọi là VISTA (không phải hệ điều hành mới đây của

Microsoft) khi phân tích hệ gen ở nhiều loài khác nhau tìm thấy sự bảo thủ ở ở tỉ lệ 70% trong một đoạn trình tự phân tích 50 - 75 bp đối với một số trình tự ADN có vai trò điều hòa. Hai loài cá bể dẹt và chuột cùng có khoảng 10.000 các đoạn trình tự không mã hóa ngắn giống nhau, rất có thể chúng là các trình tự tăng cường đặc trưng mô. Tuy vậy, cả hai loài này, đặc biệt ở chuột, dường như có nhiều trình tự điều hòa bị bỏ sót khi sử dụng phần mềm máy tính để phân tích trình tự gen. Người ta đã xác định được ở loài động vật bậc thấp *Ciona intestialis* có chứa khoảng 20.000 các trình tự enhancer, và vì vậy không có gì là ngạc nhiên nếu người và chuột sẽ có khoảng 50.000 - 100.000 các trình tự enhancer trong hệ gen.

Các phương pháp được sử dụng để xác định các trình tự tăng cường dựa trên việc xác định các vị trí liên kết của các yếu tố hoạt hóa hoặc ức chế phiên mã. Việc xác định được các trình tự điều hòa trong phân tử ADN còn là thách thức lớn hơn so với việc xác định được các trình tự mã hóa protein bởi các trình tự điều hòa không bị hạn chế bởi các nguyên lý của mã di truyền. Vì vậy, dường như việc phải phối hợp nhiều phương pháp sinh tin học và chương trình máy tính là cần thiết để có thể xác định được các trình tự ADN điều hòa trong toàn bộ hệ gen.

Công cụ phần mềm phân tích hệ gen được sử dụng rộng rãi nhất hiện nay là BLAST (*basiclocalalignmenttool*). Có một số cải biến khác nhau trong các chương trình BLAST, tuy vậy tất cả các chương trình này đều có các đặc điểm chung là tìm được những vùng giống nhau giữa các gen mã hoa protein khác nhau. Có nhiều cách để tìm dữ liệu từ BLAST. Một trong những cách đó là sử dụng công cụ tìm kiếm hệ gen hoặc các hệ gen đối với tất cả các trình tự protein được dự đoán trước gọi là "**querry sequence**". Chẳng hạn như ví dụ sau: gen *eve* mã hóa trong một protein điều hòa phiên mã thiết yếu cho sự phân hóa tế bào ở phôi *Drosophila*. Protein *Eve* có 376 axit amin. Vùng chức năng của protein này nằm giữa các axit amin 71 - 130. Khi sử dụng trình tự của 60 axit amin này để tìm kiếm, kết quả cho thấy hệ gen Drosophila có 75 gen mã hóa chứa trình tự này. Như vậy, chương trình BLAST đã nhanh chóng xác định được một loạt các gen có chức năng tương tự.

Một cách khác để khai thác cơ sở dữ liệu của BLAST là tra cứu theo trình tự nucleotit. Chẳng hạn như trong thí dụ trên, người ta có thể sử dụng tương ứng trình tự 180 bp mã hóa cho hộp định loại gen (homeobox).

Tóm lại, việc trình tự các hệ gen đầy đủ của các loài khác nhau ngày càng tăng lên đã cung cấp một cơ sở dữ liệu ngày càng phong phú và đầy đủ cho các nghiên cứu hệ gen học so sánh. Ngày càng có nhiều các chương trình máy tính được phát triển và hoàn thiện để khai thác vốn thông tin di truyền đang ngày càng được tạo ra đầy đủ hơn qua các chương trình giải mã ADN tự động.

Các kỹ thuật phân tích protein

Chuẩn bị dịch chiết tế bào để tinh sạch protein

Việc phân lập và tinh sạch được các loại protein riêng rẽ có ý nghĩa quyết định đến khả năng tìm hiểu được chức năng của chúng. Mặc dù trong một số trường hợp, chúng ta có thể nghiên cứu chức năng của protein ở dạng hỗn hợp phức tạp, nhưng phần lớn những nghiên cứu này thường dẫn đến những kết luận "mù mờ". Chẳng hạn như khi chúng ta nghiên cứu về hoạt tính của một enzym ADN polymerase trong một hỗn hợp protein thô (chẳng hạn từ dịch phân giải tế bào), các enzym ADN polymerase và protein thành phần khác cũng có thể ảnh hưởng đến hiệu suất tổng hợp ADN quan sát được trong thực nghiệm. Vì vậy, việc tinh sạch các protein là một bước quan trong trong quá trình tìm hiểu về chức năng của chúng.

Mỗi một protein thường có một số đặc tính riêng làm việc tinh sạch chúng thường có tính đặc thù. Điều này thì trái ngược với ADN, vốn cơ bản giống nhau về cấu trúc và thành phần, chỉ khác nhau về trình tự của các nucleotit. Các bước tinh sạch từng loại protein thường dựa trên các đặc tính đặc thù của nó về kích thước, hình dạng, điện tích và nhiều khi là chức năng của chúng.

Vật liệu khởi đầu cho hầu hết các quá trình tinh sạch protein từ sinh vật là các dịch chiết tế bào. Không giống ADN vốn có tính phục hồi cao trong các điều kiện nhiệt độ sống khác nhau, thì protein rất dễ bị biến tính và phá hủy sau khi bị giải phóng ra khỏi tế bào. Vì lý do này, hầu hết quá trình chuẩn bị các dịch chiết và tinh sạch protein được tiến hành ở nhiệt độ lạnh (4^{oC}). Có một số cách chuẩn bị dịch chiết tế bào. Các tế bào có thể phân giải bằng sử dụng chất tẩy, các lực làm vỡ thành tế bào, xử lý với dung dịch nhược trương (làm tế bào trương lên do nước đi vào và vỡ ra), hoặc thay đổi đột ngột áp suất. Điểm chung của tất cả các phương pháp là làm thành tế bào vỡ ra và các protein được giải phóng. Trong một số trường hợp, các tế bào được chuyển về trạng thái đông lạnh trước khi được nghiền bằng những máy nghiền mẫu trong phòng thí nghiệm.

Sử dụng sắc ký cột trong tinh chế protein

Phương pháp chiết xuất và tinh sạch protein phổ biến nhất là **sắc ký cột**. Trong trường hợp này, các phân đoạn protein được cho chạy qua cột nhồi bằng các hạt agarose hoặc polyacrylamit nhỏ được cải biến cho phù hợp. Có một số phương án khác nhau trong việc sử dụng cột tách chiết và tinh sạch protein. Các quy trình khác nhau được thiết lập có thể khác nhau do đặc tính khác nhau của các loại protein. ở đây mô tả ba phương pháp. Trong hai phương pháp đầu, protein được phân lập dựa vào kích thước và tính tích điện của chúng. Tóm tắt về các phương pháp này được nêu trên hình 14.

Sắc ký trao đổi ion: Trong phương pháp này, các phân tử protein được phân lập dựa trên điện tích ion hóa bề mặt của chúng bằng việc sử dụng các vật liệu làm cột là các hạt

mang các nhóm chức tích điện âm hoặc dương (đây còn được gọi là **pha tĩnh**). Các phân tử protein tương tác yếu với các hạt (chẳng hạn như một phân tử protein tích điện dương được cho chạy qua cột mang các hạt tích điện âm) sẽ được hồi lưu khi sử dụng một dung dịch muỗi loãng chảy qua cột sau đó (dung dịch chạy mẫu được gọi là **pha động**). Các phân tử tương tác với pha động càng mạnh, càng cần dung dịch hàm lượng muối cao để hồi lưu mẫu (bởi muối làm "trung hòa" các vùng mang điện tích và vì vậy cho phép các phân tử protein được giải phóng khỏi cột. Bằng việc tăng dần nồng độ muối trong các dung dịch đệm thu hồi mẫu, các phân tử protein khác nhau, kể cả các phân tử có đặc tính tích điện gần giống nhau cũng được phân tách thành các phân đoạn khác nhau khi chúng được hồi lưu từ cột.

Sắc ký lọc gel: Kỹ thuật này cho phép phân tách các loại protein trên cở sở đặc điểm khác nhau của các loại protein về hình dạng và kích thước. Khác với trong kỹ thuật sắc ký trao đổi ion, các hạt được sử dụng để nhồi cột trong kỹ thuật này không mang các nhóm tích điện mà thay vào đó là mang các lỗ có kích thước khác nhau. Các phân tử protein càng nhỏ càng có nhiều khả năng thâm nhập vào tất cả các lỗ; vì vậy, thời gian chạy qua cột dài hơn và thời gian hồi lưu muộn hơn. Ngược lại, các phân tử protein kích thước càng lớn càng có thời gian hồi lưu (chạy qua toàn bộ cột) sớm hơn.

Đối với mỗi loại cột, các phân đoạn sắc ký được thu ở các nồng độ muối khác nhau hoặc ở các thời gian hồi lưu khác nhau để thu được từng loại protein được quan tâm nghiên cứu. Các phân đoạn có hoạt tính protein được quan tâm cao nhất sẽ được tích lũy và tiến hành tinh sạch bổ sung.

Độ tinh sạch của sản phẩm protein sẽ tăng lên khi các phân đoạn protein được chạy qua nhiều cột sắc ký khác nhau. Thông thường một cột sắc ký đơn lẻ không đủ để tinh sạch được một loại phân tử protein mong muốn nào đó dù quá trình sắc ký được lặp đi lặp lại nhiều lần, thay vào đó người ta thường phải áp dụng một chuỗi các bước kỹ thuật để có thể thu được một phân đoạn chứa một lượng lớn loại protein cần quan tâm nghiên cứu. Chẳng hạn như, dù có rất nhiều phân tử protein được hồi lưu trong dung dịch muối đậm đặc từ cột tích điện dương (đối với protein tích được âm) hoặc được hồi lưu trong sắc ký lọc gel (đối với các protein kích thước tương đối nhỏ), các kỹ thuật này đơn lẻ thường không đủ để thu được một sản phẩm protein được tinh sạch hoàn toàn.

Sắc ký ái lực hỗ trợ quá trình tinh chế protein

Các đặc tính đặc trưng của từng loại protein có thể được tận dụng để giúp tinh sạch loại protein tương ứng được thuận tiện và hiệu quả hơn. Giả sử, nếu chúng ta biết một loại protein khi hoạt động liên kết đặc hiệu với ATP, thì có thể dùng cột sắc ký mang vật liệu gắn kết ATP để phân tách protein đó. Chỉ có protein liên kết với ATP mới được cột giữ lại, và cho phép hầu hết các loại protein không liên kết với ATP được chảy trôi qua cột. Kỹ thuật tinh sạch này được gọi là **sắc ký ái lực**. Có nhiều hợp chất khác nhau có thể được sử dụng để gắn kết với cột và giúp quá trình tinh sạch protein dễ dàng và hiệu quả

hơn. Các hợp chất này bao gồm cả các trình tự ADN (để tinh sạch protein liên kết ADN) hoặc thậm trí là một loại protein để tinh sạch một loại protein khác được biết hoặc được mong đợi có tương tác phân tử với loại protein trên. Như vậy, để bắt đầu tinh sạch một loại protein nào đó, cần phải có những hiểu biết cơ bản về loại protein đó và tìm cách khai thác, ứng dụng các đặc tính có nó cho quá trình tách chiết và tinh sạch.

Một dạng rất phổ biến của sắc ký ái lực là **sắc ký ái lực miễn dịch**. Trong phương pháp này, người ta gắn kháng thể đặc hiệu với protein đích lên vật liệu làm cột sắc ký. Trong trường hợp lý tưởng, loại kháng thể này chỉ liên kết đúng với loại protein cần quan tâm còn cho phép tất cả các loại protein khác chảy trôi qua cột. Loại protein liên kết sau đó sẽ được thu hồi (hồi lưu) bằng cách sử dụng dung dịch muối hoặc đôi khi là các dung dịch chất tẩy nhẹ chảy qua cột. Khó khăn cơ bản gặp phải đối với phương pháp này là đôi khi liên kết giữa kháng thể và protein khá bền vững đến mức phải gây biến tính protein mới thu hồi được sản phẩm. Trong khi khác với ADN, protein sau khi biến tính thường không có khả năng hồi tính, vì vậy protein thu được theo cách này nhiều khi ở dạng không hoạt động chức năng và mất đi giá trị sử dụng trong nghiên cứu.

Để tăng hiệu quả tinh sạch, các protein cũng có thể được cải biến. Những cải biến này có thể là sự bổ sung các trình tự axit amin ngắn hoặc là vào đầu C hoặc vào đầu N của phân tử protein cần phân tích. Những bổ sung này, hay còn gọi là "trình tự đánh dấu" có thể tạo ra được bằng công nghệ ADN tái tổ hợp. Các trình tự peptit đánh dấu giúp thay đổi thuộc tính của một phân tử protein mong muốn và giúp tinh chế protein này dễ dàng hơn. Ví dụ như, đối với một số protein người ta tiến hành bổ sung một chuỗi gồm 6 histidin giúp những protein này liên kết với Ni^{2+} gắn trên cột chặt hơn và dễ phân tách hơn, trong khi thuộc tính này thường không có ở phần lớn các loại protein khác. Ngoài ra việc sử dụng các **epitop** đặc hiệu (thường là một trình tự peptit có 5 - 7 axit amin đặc hiệu xác định kháng nguyên) cũng có thể được gắn vào phân tử protein cần tinh chế. Các cải biến này cho phép tinh sạch các loại protein trên cơ sở nguyên tắc sắc ký ái lực miễn dịch và sử dụng dị kháng nguyên mang các epitop được bổ sung. Điều đặc biệt là, các kháng thể và epitop này có thể thay đổi tính liên kết kháng thể tùy thuộc vào điều kiện khác nhau của môi trường (chẳng hạn, ái lực tăng khi không có Ca^{2+}, và ái lực giảm khi có Ca^{2+}). Điều này cũng giúp làm giảm ảnh hưởng của các yếu tố gây biến tính khác.

Nguyên tắc sắc ký ái lực miễn dịch cũng có thể được dùng để làm kết tủa nhanh một loại protein đặc hiệu nào đó (và mọi loại protein liên kết chặt với nó) từ một dịch chiết thô. Trong trường hợp này, phản ứng kết tủa thu được do kháng thể được gắn vào cùng loại hạt được sử dụng trong sắc ký cột. Do các hạt này có kích thước lớn, nên chúng lắng rất nhanh xuống đáy ống nghiệm mang theo các kháng thể và protein liên kết với chúng. Kỹthuật này được gọi là kỹ thuật **kết tủa miễn dịch**, cũng là một kỹ thuật ngày càng sử dụng phổ biến để tinh sạch nhanh protein hoặc phức hệ protein từ các dịch chiết thô. Mặc dù nếu chỉ sử dụng phương pháp này riêng rẽ, hiếm khi thu được sản phẩm protein được tinh sạch hoàn toàn, song phương pháp này thường rất hiệu quả để xác định các

loại phân tử protein và các hợp chất khác (ví dụ như ADN) có tương tác với một loại protein đích nào đó.

Phân tích protein trên gel polyacrylamid

Các loại protein thường không mang điện tích âm đồng đều hay có cấu trúc bậc hai đồng nhất. Thay vào đó, chúng được tạo ra từ 20 loại axit amin khác nhau, một số chúng không mang điện tích, một số mang điện tích âm, còn một số mang điện tích dương. Ngoài cấu trúc bậc hai, protein hoạt động còn có các cấu trúc bậc ba, bậc bốn điển hình. Tuy vậy, nếu các protein được xử lý với các chất tẩy ion hóa mạnh như **SDS** (sodium dodecyl sulphat) và một hợp chất khử, ví dụ như mercapthoethanol, thì các cấu trúc bậc 2, 3 và 4 của phân tử protein sẽ bị phá vỡ. Nghĩa là, sau khi xử lý với SDS, protein trở thành dạng phân tử polymer không có cấu trúc. Đồng thời, SDS tạo thành lớp vỏ ion và làm phân tử protein trở nên tích điện âm đồng đều hơn. Mecarptoethanol thì làm giảm các liên kết disulphit hình thành giữa các tiểu phần cystein. Kết quả là, nếu các phân tử ADN và ARN được gây biến tính bởi SDS và mercaptoethanol, thì sự phân tách của các loại phân tử protein trong điện di chủ yếu là do sự khác biệt về trọng lượng và kích thước phân tử. Sau khi điện di, các phân tử protein được nhuộm với các thuốc nhuộm liên kết protein như **Coomassie brilliant blue** và quan sát. Khi không có SDS, điện di vẫn có thể phân tách được các loại protein, nhưng lúc này còn có các yếu tố khác nhau của các loại protein là trọng lượng phân tử, tổng điện tích và điểm đẳng điện (xem dưới đây).

Định tính protein dựa trên phương pháp thẩm tách miễn dịch (immunoblotting)

Mặc dù có bản chất khác ADN và ARN, việc xác định sự có mặt của một loại protein trong số các protein có trong mẫu sinh học theo nguyên tắc gần giống với các phương pháp thẩm tách (còn gọi là lai) Southern và Northern (tương ứng với trường hợp của ADN và ARN). Tuy vậy, đối với protein, phương này định tính này dựa trên dựa trên nguyên tắc miễn dịch đặc thù của protein nên còn được gọi là phương pháp thẩm tách miễn dịch (immunoblotting) hay phương pháp thẩm tách (lai) Western. Trong phương pháp thẩm tách miễn dịch, các phân tử protein sau khi được phân tách trên điện di được chuyển và gắn lên màng. Màng này sau đó được ủ trong một dung dịch chứa kháng thể đặc hiệu với loại protein tinh sạch được quan tâm nghiên cứu. Kháng thể sẽ tìm thấy loại protein tương ứng ở trên màng lọc và gắn vào. Cuối cùng, bằng việc sử dụng phản ứng enzym hiện màu, người ta có thể quan sát được vị trí kháng thể liên kết trên màng. Như vậy, tất cả các phương pháp lai Southern, Northern và Western đều có điểm chung là sử dụng các hợp chất chọn lọc để quan sát sự có mặt của một hoặc một số loại phân tử đặc thù trong các hỗn hợp phức tạp.

Giải trình tự trực tiếp protein

Mặc dù có cấu tạo phức tạp hơn so với ADN và ARN, các phân tử protein cũng có thể giải trình tự trực tiếp, tức là việc xác định thành phần và thứ tự của các axit amin trên chuỗi polypeptit. Có hai phương pháp được sử dụng phổ biến hơn cả là: 1) phương pháp biến tính Edman và 2) phương pháp khối phổ kế tiếp. Khả năng xác định được trình tự protein có ý nghĩa quan trong trong trong nhiều nghiên cứu về hệ protein học (proteomics) và hệ gen học (genomics). Bởi vì với việc xác định được dù chỉ là một trình tự ngắn của các axit amin của một phân tử protein nào đó, chúng ta có thể xác định được gen mã hóa cho protein đó trên cơ sở đối chiếu với khung đọc mở có trong hệ gen của nhiều loài vốn đã được giải mã hoàn toàn hoặc gần như hoàn toàn nhưng nhưng chưa biết đầy đủ về chức năng của nhiều gen.

Phương pháp biến tính Edman là một phản ứng hóa học trong đó các tiểu phần axit amin được giải phóng lần lượt từ đầu N của chuỗi polypeptit. Điểm mấu chốt của phương pháp này là axit amin ở tận cùng đầu N của một chuỗi polypeptit có thể được cải biến khi xử lý với **phenylisothyocyanate** (**PITC**) dẫn đến sự thay đổi ở gốc α-amino. Axit amin được cải biến này sau đó được cắt rời khỏi chuỗi polypeptit khi xử lý với axit trong điều kiện không làm phá hủy phần còn lại của chuỗi polypeptit. Việc xác định trình tự các axit amin được tiến hành dựa trên thứ tự hồi lưu của từng axit amin được cắt khỏi chuỗi bằng kỹ thuật sắc ký lỏng hiệu năng cao HPLC (high performance liquid chromatography), nhờ đặc điểm từng loại axit amin có thời gian hồi lưu đặc trưng. Mỗi một vòng gây biến đổi axit amin và cắt khỏi chuỗi polypeptit như vậy tạo ra một chuỗi polypeptit và một nhóm α-amino. Với việc lặp đi, lặp lại phản ứng cắt axit amin này sẽ cho phép xác định được trình tự ở đầu N của một chuỗi polypeptit. Trong thực tế, chu kỳ gây biến tính như vậy được lặp đi lặp lại từ 8 - 15 lần để xác định protein. Số chu kỳ như vậy thông thường là đủ để xác định được một loại protein đặc thù nào đó.

Phương pháp giải mã trình tự tự động dựa trên nguyên lý biến tính Edman là một phương pháp hiệu quả và cũng đã được sử dụng rộng rãi. Tuy vậy kỹ thuật này gặp trở ngại khi đầu N tận cùng của phân tử protein bị cải biến về mặt hóa học (ví dụ như bởi các nhóm acetyl hoặc formyl), mà hiện tượng này thì vốn có thể xảy ra tự nhiên trong điều kiện in vivo, hoặc trong quá trình tách chiết và tinh sạch protein. Để khắc phục hiện tượng này, người ta có thể sử dụng protease để cắt chuỗi polypeptit và phân tích trình tự bên trong chuỗi.

Phương pháp khối phổ kế tiếp (MS/MS) có thể được dùng để xác định các vùng trình tự protein khác nhau. Khối phổ là phương pháp xác định chính xác khối lượng của các các phân tử nhỏ. Một cách vắn tắt phương pháp này được mô tả như sau: phân tử cần phân tích được cho bay qua một thiết bị (trong điều kiện chân không) trong điều kiện tốc độ di chuyển của nó tương quan với tỉ số khối lượng / điện tích. Trên cơ sở này người ta có thể đo được thời gian bay của phân tử và xác định được khối lượng của nó. Đối với

các phân tử sinh học có kích thước nhỏ như các chuỗi peptit và các phân tử protein nhỏ, thì trọng lượng phân tử của nó có thể xác định chính xác đến từng Dalton.

Để xác định khối lượng phân tử protein bằng phương pháp MS/MS, trước tiên phân tử protein thường được cắt thành các đoạn peptit có trình tự ngắn (thường dưới 20 axit amin) nhờ sử dụng enzym đặc hiệu, chẳng hạn như trypsin. Hỗn hợp các đoạn peptit này sau đó được đưa vào phân tích khối phổ và chúng phân tách nhau ra dựa trên tỉ số khối lượng / điện tích. Các đoạn peptit riêng rẽ sau đó sẽ bị bắt giữ và phân đoạn thành các chuỗi peptit thành phần. Khối lượng của mỗi peptit thành phần được xác định bằng phổ khối như minh họa trên hình 15. Sự kết hợp các dữ liệu về khối lượng của các đoạn peptit thành phần sẽ cho biết trình tự rõ ràng của phân tử protein ban đầu. Cũng giống như trong phương pháp biến tính Edman, việc xác định được trình tự của của khoảng 15 axit amin là đủ để so sánh với trình tự protein được luận ra từ các trình tự ADN đã được giải mã.

Kỹ thuật MS/MS thực sự là một kỹ thuật có tính cách mạng trong việc xác định và giải trình tự protein. Trong phương pháp này, thông thường chỉ cần một lượng mẫu nhỏ và hơn hết là có thể phân tích hỗn hợp nhiều loại protein đồng thời.

Hệ protein học (proteomics)

Việc phát triển của các kỹ thuật giải mã trình tự ADN, kết hợp với các phương pháp tách chiết, tinh sạch và phân tích trình tự protein đã mở đường cho sự hình thành một chuyên ngành mới gọi là hệ protein học. Hệ protein học (proteomics) là chuyên ngành nghiên cứu về toàn bộ tập hợp protein được một mô, tế bào hoặc cơ thể tạo ra trong các điều kiện đặc thù, phân tích mức độ phổ biến của từng protein và sự tương tác của chúng với nhau và với các phân tử khác (ví dụ như ADN) trong quá trình hoạt động của tế bào. Nếu như các kỹ thuật phân tích vi dãy phản ứng (microarray) có thể giúp nhận biết được sự biểu hiện của các gen trên cở sở phân tích hệ gen, thì các kỹ thuật proteomics cho phép xác định được hình ảnh tổng thể về toàn bộ vốn protein của một tế bào, mô hoặc cơ thể.

Hệ protein học được dựa trên ba phương pháp cơ bản: điện di hai chiều trên gel polyacrylamid để tiến hành phân tách các protein, khối phổ để xác định khối lượng phân tử và định tính protein (hoặc các đoạn peptit từ phân tử protein đó), và sinh tin học để đối chiếu các phân tử protein và các đoạn peptit với các trình tự mã hóa của chúng trong hệ gen. Một tế bào riêng lẻ thường ít nhất tạo ra hàng nghìn loại protein khác nhau, trong khi việc sử dụng đơn lẻ phương pháp điện di truyền thống trên gel polyacrylamid thường không đủ để phân lập và tách biệt các protein này. Vì vậy, như tên gọi của nó, phương pháp điện di hai chiều cho phép phân tách các phân tử protein trên gel điện di theo hai chiều được thực hiện kế tiếp nhau.

Trong bước thứ nhất, các phân tử protein được phân đoạn dựa trên điểm đẳng điện của chúng (theo nguyên tắc hội tụ đẳng điện). Theo nguyên tắc này, khi một gradient pH được tạo ra từ đầu này đến đầu kia của bản điện di, thì tại vị trí pH tương ứng làm trung hòa điện tích của một phân tử protein sẽ làm các phân tử protein tập trung (hội tụ) tại vị trí đó. Trong bước thứ hai, các phân tử protein tại điểm hội tụ được tiếp tục phân tách trên cơ sở khối lượng và kích thước của chúng khi di chuyển trên trường điện di polyacrylamid đã được gây biến tính bởi SDS như mô tả ở trên. Do các phân tử protein được đồng thời phân tách dựa trên hai thuộc tính (điểm đẳng điện và trọng lượng phân tử), nên hàng nghìn loại protein khác nhau có thể được phân tách khỏi nhau trong một thí nghiệm duy nhất. Sau khi được phân đoạn theo phương pháp điện di hai chiều, mỗi loại protein riêng biệt sẽ được đưa vào phân tích khối phổ để xác định chính xác khối lượng phân tử. Như đã nói ở trên, để tăng hiệu quả phân tích, thông thường protein được cắt thành các đoạn peptit nhỏ nhờ sử dụng protease thay cho việc phân tích phân tử protein ở dạng nguyên vẹn có phân tử lượng lớn và cấu trúc phức tạp. Kỹ thuật phân tích MS/MS cho phép giải trình tự chi tiết của các đoạn peptit và xác định phân tử protein.

Cuối cùng các dữ liệu về trình tự hệ gen hoàn chỉnh của một cơ thể và các trình tự peptit từ các loại protein được quan tâm nghiên cứu sẽ được đưa vào phân tích bằng các phần mềm sinh tin học để có thể xác định được một trình tự mã hóa đặc thù trong hệ gen tương ứng với loại protein có chức năng được quan tâm nghiên cứu. Như vậy, các nghiên cứu hệ gen học (genomics) và hệ protein học (proteomics) có mối quan hệ chặt chẽ trong các nghiên cứu di truyền học phân tử.