Chương 2

Cấu trúc genome

Genome (hệ gen, bộ gen) là thuật ngữ được dùng với các nghĩa khác nhau như sau:

- Nguyên liệu di truyền của một cơ thể: 1) nhiễm sắc thể trong tế bào vi khuẩn (hoặc một trong mỗi loại nhiễm sắc thể nếu hơn một loại có mặt, ví dụ: các nhiễm sắc thể lớn hoặc bé của *Vibrio cholerae*), 2) DNA hoặc RNA trong một virion, 3) nhiễm sắc thể cùng với mọi plasmid được kết hợp (ví dụ: nhiễm sắc thể và hai plasmid nhỏ trong vi khuẩn *Buchnera*).

- Tất cả các gen (khác nhau) trong tế bào hoặc virion.

- Bộ nhiễm sắc thể đơn bội hoặc genome đơn bội trong tế bào.

Chuỗi genome hoàn chỉnh (nghĩa là trình tự hoàn chỉnh của các nucleotide trong genome) đã được công bố cho một số loài vi khuẩn. Các trình tự khác cũng đã được công bố, ví dụ genome của cây cúc dại (*Arabidopsis thaliana*) và genome người.

Genome chứa toàn bộ thông tin di truyền và các chương trình cần thiết cho cơ thể hoạt động. Ở các sinh vật nhân thật (eukaryote), 99% genome nằm trong nhân tế bào và phần còn lại nằm trong một số cơ quan tử như ty thể và lạp thể. Đa số genome vi khuẩn và phần genome chứa trong các cơ quan tử thường có kích thước nhỏ và ở dạng vòng khép kín. Ngược lại, phần genome trong nhân thường rất lớn và phân bố trên các nhiễm sắc thể dạng thẳng.

Dự án genome là dự án xác định cấu trúc di truyền chính xác của một genome cơ thể sống, nghĩa là trình tự DNA của tất cả các gen của nó. Dự án genome của một số sinh vật mô hình (model organisms) đã được hoàn thành như sau:

- Các genome vi khuẩn. Các trình tự hoàn chỉnh của genome *Escherichia coli* đã được xác định theo phương thức tổ hợp/tập hợp (consortium) của các phòng thí nghiệm. Năm 1995, hai trình tự genome hoàn chỉnh của vi khuẩn *Haemophilus influenzae* và *Mycoplasma genitalium* cũng được hoàn thành. Loài *M. genitalium* có một genome đơn

giản (khoảng 580.067 base), do nó dựa vào vật chủ để vận hành nhiều bộ máy trao đổi chất của mình. Loài *H. influenzae* là một vi khuẩn đặc trưng hơn, và có genome khoảng 1.830.121 base với 1.749 gen.

- Chuỗi genome hoàn chỉnh của nấm men *Saccharomyces cerevisiae* đã được hoàn chỉnh trong năm 1996, nhờ một consortium của các phòng thí nghiệm. Genome của chúng dài 12.146.000 base.

- Các dự án genome ở động vật như: chuột, cừu, lợn, giun tròn (*Caenorhabditis elegans*), ruồi giấm (*Drosophila melanogaster*)…, hoặc ở thực vật như: lúa nước, lúa mì, ngô, táo, cúc dại…, mà nổi bật nhất trong số đó là dự án genome người cũng đã được thực hiện.

Ngày 12. 2. 2001 genome người đã được công bố với khoảng 30.000 gen, ít hơn nhiều so với dự kiến trước đây (hàng trăm ngàn gen), và chỉ gấp hai lần giun tròn hoặc ruồi giấm. Người ta đã xác định hệ gen người giống 98% so với tinh tinh và có đến 99% là giống nhau giữa các dân tộc, các cá thể. Do đó, vấn đề hình thành và phát triển nhân cách, chỉ số thông minh... phải chủ yếu trên cơ sở xã hội và sự rèn luyện của từng người để phát triển tiềm năng sinh học của bản thân.

Trình tự genome của những sinh vật mô hình rất có ý nghĩa trong những nghiên cứu của một chuyên ngành khoa học mới đó là genome học (genomics). Dựa vào đây, các nhà sinh học phân tử có thể phân tích cấu trúc, hoạt động và chức năng của các gen, làm sáng tỏ được vai trò của DNA lặp lại, DNA không chứa mã di truyền, DNA nằm giữa các gen... Điều đặc biệt có ý nghĩa là khi so sánh các genome với nhau, có thể hiểu được hoạt động của genome trong các cơ thể sống, mối quan hệ giữa chúng, sự đa dạng sinh học và mức độ tiến hóa.

Kết quả bước đầu so sánh genome giữa các loài sinh vật với nhau đã cho thấy có ba đặc điểm nổi bật: 1) các gen phân bố trong genome không theo qui luật, 2) kích thước của genome thay đổi không tỷ lệ thuận (tương quan) với tính phức tạp của loài, 3) số lượng nhiễm sắc thể cũng rất khác nhau ngay giữa những loài rất gần nhau.

I. Thành phần và đặc điểm của genome

Genome chứa mọi thông tin di truyền đặc trưng cho từng loài, thậm chí cho từng cá thể trong loài. Genome có thể bao gồm các phân tử DNA

hoặc RNA. Đối với sinh vật bậc cao, kích thước genome thay đổi từ 10^9 bp (động vật có vú) đến 10^{11} bp (thực vật). Khác với tế bào tiền nhân (prokaryote), các gen trong genome của eukaryote thường tồn tại nhiều bản sao và thường bị gián đoạn bởi các đoạn mã mù không mang thông tin di truyền (các intron). Vì vậy, một trong những vấn đề đang được quan tâm là cần phải biết số lượng các gen khác nhau có mặt trong genome cũng như số lượng các gen hoạt động trong từng loại mô, từng giai đoạn phát triển và tỷ lệ các gen so với kích thước genome...

1. Genome của cơ quan tử

Hầu hết genome của cơ quan tử, nhưng không phải luôn luôn, có dạng phân tử DNA mạch vòng đơn của một chuỗi duy nhất.

Genome của cơ quan tử mã hóa cho một số, không phải tất cả, các protein được tìm thấy trong cơ quan tử. Do có nhiều cơ quan tử trong một tế bào, cho nên có nhiều genome của cơ quan tử trên một tế bào. Mặc dù bản thân genome của cơ quan tử là duy nhất. Nhưng nó cấu tạo gồm một chuỗi lặp lại[1] liên quan với mọi chuỗi không lặp lại[2] của nhân. Về nguyên tắc, các gen cơ quan tử được phiên mã và dịch mã bởi các cơ quan tử.

1.1. Genome của ty thể

DNA ty thể (mitochondrial DNA-mtDNA) là một genome độc lập, thường là mạch vòng, được định vị trong ty thể.

- DNA ty thể của tế bào động vật mã hóa đặc trưng cho 13 protein, 2 rRNA và 22 tRNA.

- DNA ty thể của nấm men *S. cerevisiae* dài hơn mtDNA của tế bào động vật năm lần do sự có mặt của các đoạn intron dài.

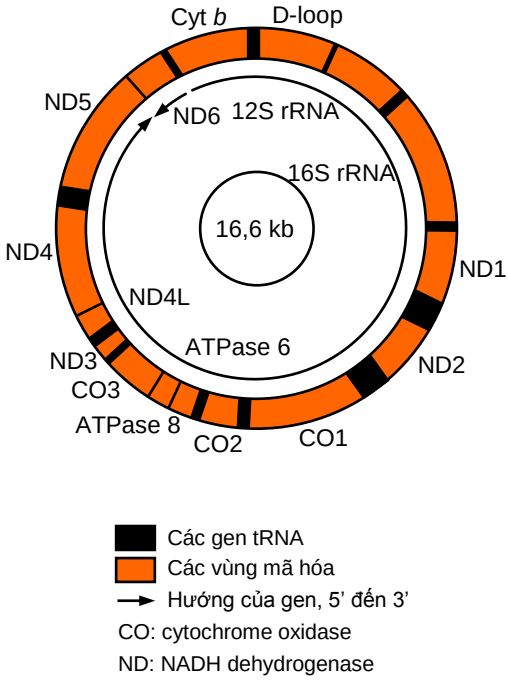
Các genome ty thể có kích thước tổng số rất khác nhau, các tế bào động vật có kích thước genome nhỏ (khoảng 16,5 kb ở động vật có vú) (Hình 2.1). Có khoảng một vài trăm ty thể trên một tế bào. Mỗi ty thể có

[1] DNA lặp lại mô tả các chuỗi hiện diện hơn một bản sao trong mỗi genome.
[2] DNA không lặp lại chứa các chuỗi duy nhất: chỉ có một bản sao trong genome đơn bội.

nhiều bản sao DNA. Số lượng tổng số của DNA ty thể so với DNA nhân là rất nhỏ (<1%).

Trong nấm men *S. cerevisiae*, genome ty thể có kích thước khá lớn (khoảng 80 kb) và khác nhau tùy thuộc vào từng chủng. Có khoảng 22 ty thể trên một tế bào, tương ứng khoảng 4 genome trên một cơ quan tử. Ở những tế bào sinh trưởng, tỷ lệ mtDNA có thể cao hơn (khoảng 18%).

Kích thước của genome ty thể ở các loài thực vật là rất khác nhau, tối thiểu khoảng 100 kb. Kích thước lớn của genome đã gây khó khăn cho việc phân lập nguyên vẹn DNA, nhưng bản đồ cắt hạn chế (restriction map) trong một vài loài thực vật đã cho thấy genome ty thể thường là một chuỗi đơn, được cấu tạo như một mạch vòng. Trong mạch vòng này có những chuỗi tương đồng ngắn và sự tái tổ hợp giữa chúng đã sinh ra các phân tử tiểu genome (subgenome) mạch vòng nhỏ hơn, cùng tồn tại với genome "chủ" (master genome) hoàn chỉnh, đã giải thích cho sự phức tạp của các DNA ty thể ở thực vật.



Hình 2.1. DNA ty thể của người. Bao gồm 22 gen tRNA, 2 gen rRNA, và 13 vùng mã hóa protein.

Bảng 2.1 tóm tắt sự phân công của các gen trong một số genome ty thể. Tổng số gen mã hóa protein là khá ít, và không tương quan với kích thước của genome. Ty thể động vật có vú sử dụng các genome 16 kb của chúng để mã hóa cho 13 protein, trong khi đó ty thể nấm men *S. cerevisiae* dùng các genome từ 60-80 kb mã hóa cho khoảng 8 protein. Thực vật với genome ty thể lớn hơn nhiều mã hóa cho nhiều protein hơn. Các intron được tìm thấy trong hầu hết các genome của ty thể, nhưng lại không có trong các genome rất nhỏ của động vật có vú.

Hai rRNA chính luôn được mã hóa bởi genome ty thể. Số lượng các tRNA được mã hóa bởi genome ty thể dao động từ không cho đến đầy đủ (25-26 trong ty thể). Nhiều protein ribosome được mã hóa trong genome ty thể của thực vật và sinh vật nguyên sinh, nhưng chỉ có một ít hoặc không có trong genome của nấm và động vật.

Bảng 2.1. Các genome ty thể có các gen mã hóa cho các protein, rRNA và tRNA

Ty thể mã hóa cho các RNA và protein			
Loài	**Kích thước (kb)**	**Các gen mã hóa protein**	**Các gen mã hóa RNA**
Nấm	19-100	8-14	10-28
Sinh vật nguyên sinh	6-100	3-62	2-29
Thực vật	186-366	27-34	21-30
Động vật	16-17	13	4-24

1.2. Genome của lạp thể

DNA lạp thể (chloroplast DNA-ctDNA) cũng là một DNA genome độc lập, thường là mạch vòng, được tìm thấy trong lạp thể của thực vật.

- Genome của lạp thể rất khác nhau về kích thước, nhưng đủ lớn để mã hóa cho khoảng 50-100 protein cũng như rRNA và tRNA.

- DNA lạp thể dài từ 120-190 kb. Các genome của lạp thể đã được phân tích trình tự cho thấy có khoảng 87-183 gen. Bảng 2.2 mô tả các chức năng được mã hóa bởi genome lạp thể ở cây trồng.

Bảng 2.2. Genome của lạp thể ở các cây trồng mã hóa cho 4 rRNA, 30 tRNA và khoảng 60 protein

Các lạp thể có hơn 100 gen
Các gen
- Mã hóa RNA
16S rRNA
23S rRNA
4,5S rRNA
5S rRNA
tRNA
- Biểu hiện gen
Các r-protein
RNA polymerase
Khác
- Các chức năng của lạp thể
Rubisco và thylakoids
NADH dehydrogenase

Nói chung, các đặc điểm của genome lạp thể tương tự ở ty thể, ngoại trừ lạp thể mang nhiều gen hơn. Genome lạp thể mã hóa cho tất cả các loại rRNA và tRNA cần thiết trong tổng hợp protein, và cho khoảng 50 protein, bao gồm cả RNA polymerase và các protein ribosome.

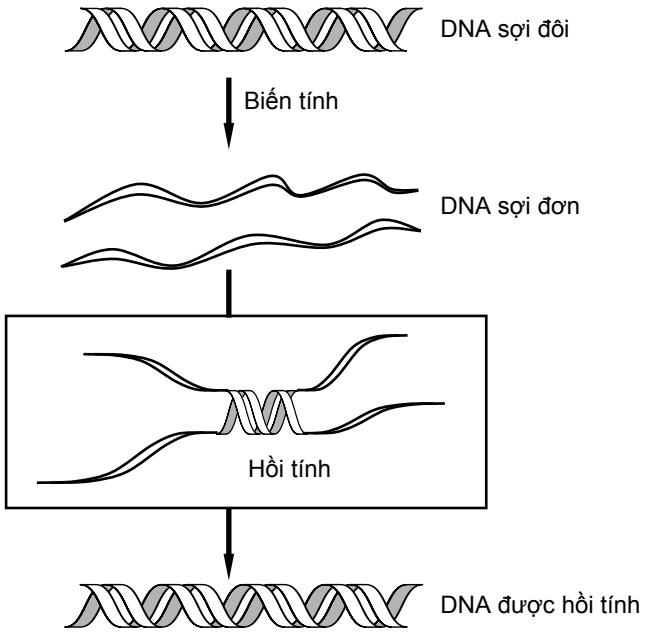
Các intron trong lạp thể được chia thành hai nhóm: 1) những intron ở trên các gen tRNA thường (mặc dù không chắc chắn) được định vị trong vòng anticodon, giống như các intron được tìm thấy trong các gen tRNA

của nhân nấm men *S. cerevisiae*; 2) những intron trong các gen mã hóa protein tương tự với các intron của các gen ty thể.

Vai trò của lạp thể là thực hiện quá trình quang hợp. Do đó, nhiều gen của nó mã hóa cho các protein của các phức hợp định vị trong các màng thylakoid. Một vài phức hợp protein của lạp thể giống các phức hợp protein của ty thể: có một số tiểu đơn vị được mã hóa bởi genome của cơ quan tử và một số khác được mã hóa bởi genome của nhân. Nhưng các phức hợp còn lại được mã hóa hoàn toàn bởi genome lạp thể.

2. Động học của phản ứng lai DNA

Bản chất chung của eukaryotic genome được phản ánh qua động học của sự tái liên kết các DNA (DNA reassociation kinetics) bị biến tính. Sự tái liên kết giữa các chuỗi DNA bổ sung xảy ra nhờ bắt cặp base, ngược lại với quá trình biến tính (denaturation) mà nhờ đó chúng được tách rời (Hình 2.2) để thực hiện sự tái bản hoặc phiên mã. Động học của phản ứng tái liên kết phản ánh sự khác nhau của các chuỗi hiện diện, vì thế phản ứng này có thể được dùng để định lượng các gen và các sản phẩm RNA của chúng.



Hình 2.2. DNA có thể biến tính và hồi tính

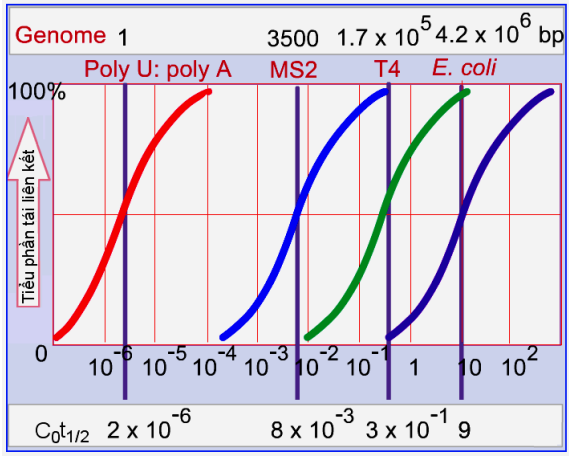
Bảng 2.3 mô tả phản ứng tái liên kết. Sự hồi tính của DNA (renaturation) phụ thuộc vào sự va chạm ngẫu nhiên của các chuỗi bổ sung. Phản ứng của các DNA riêng biệt có thể được mô tả bằng các điều kiện cần thiết cho sự hoàn thành một nửa (half-completion). Đây là tích số của $C_0 \times t_{1/2}$ và được gọi là $C_0t_{1/2}$. Giá trị này tỷ lệ nghịch với hằng số tốc độ. Do $C_0t_{1/2}$ là tích số của nồng độ và thời gian yêu cầu cho một nửa đường, nên một giá trị $C_0t_{1/2}$ lớn hơn dẫn đến một phản ứng chậm hơn.

Bảng 2.3. Một phản ứng tái liên kết của DNA được mô tả bởi $C_0t_{1/2}$

Phản ứng lai phụ thuộc vào C_0t
Tốc độ phản ứng
Phản ứng theo phương trình bậc hai $\frac{dC}{dt} = -kC^2$ *C* là nồng độ của DNA sợi đơn ở thời điểm t. *K* là hằng số tốc độ tái liên kết.
Tiến độ phản ứng
Lấy tích phân phương trình tốc độ giữa các giới hạn: nồng độ ban đầu của DNA = C_0 ở thời điểm t = 0; nồng độ duy trì sợi đơn = C sau thời gian t $\frac{C}{C_0} = \frac{1}{1 + k.C_0t}$
Thông số tới hạn là $C_0t_{1/2}$
Khi phản ứng hoàn thành một nửa ở thời điểm t = ½ $\frac{C}{C_0} = \frac{1}{2} = \frac{1}{1 + k.C_0t_{1/2}}$ Vì thế $C_0t_{1/2} = C_0t_{1/2} = \frac{1}{k}$

Sự hồi tính của DNA thường có dạng đường cong C_0t, đường cong biểu diễn đồ thị phân số của DNA được tái liên kết ($1-C/C_0$) theo log của C_0t. Hình 2.3 trình bày đường cong C_0t của một số genome đơn giản. Các đường cong có dạng tương tự nhau, nhưng giá trị $C_0t_{1/2}$ của mỗi đường là khác nhau.

Các genome trong hình 2.3 đại diện cho các nguồn DNA khác nhau (PolyU:PolyA, thực khuẩn thể MS2, thực khuẩn thể T4 và vi khuẩn *E. coli*). $C_0t_{1/2}$ liên quan trực tiếp với lượng DNA trong genome. Điều này phản ánh tình trạng khi genome trở nên phức tạp hơn, thì sẽ có thêm một số bản sao của một vài chuỗi đặc biệt trong một lượng DNA có trước. Ví dụ: nếu C_0 của DNA là 12 pg, thì nó sẽ chứa khoảng 3.000 bản sao của mỗi trình tự trong genome vi khuẩn.



Hình 2.3. $C_0t_{1/2}$ phụ thuộc vào độ phức tạp của genome. PolyU:PolyA, thực khuẩn thể MS2, thực khuẩn thể T4 và vi khuẩn *E. coli*.

3. Kích thước của genome

Không phải tất cả các đoạn DNA trong genome đều tương ứng với các gen (mã hóa cho protein hoặc một sản phẩm cần thiết cho hoạt động sống của tế bào). Từ những năm 1970, bằng các thí nghiệm gây bão hòa đột biến người ta đã có thể xác định được số gen nằm trên một đoạn nhiễm sắc thể. Ngày nay, nhờ các kỹ thuật phân tích DNA và RNA hiện đại (Southern blot, Northern blot, microarray...) các nhà khoa học có thể xác định số gen hoạt

động trong một tế bào. Ví dụ: ở tế bào nấm men *S. cerevisiae* (sinh vật eukaryote bậc thấp) có khoảng 4.000 gen hoạt động, còn tế bào động vật có vú khoảng 10.000-15.000 gen. Như vậy, nếu độ dài trung bình của một gen khoảng 10 kb thì tổng số chiều dài các gen hoạt động trong một tế bào cũng chỉ chiếm 1-2% genome. Hay nói cách khác, chỉ một phần rất nhỏ genome mang thông tin di truyền cần thiết cho hoạt động sống của tế bào. Vậy phần genome còn lại có vai trò gì, và tính phức tạp của loài có liên quan gì với kích thước genome hay không?

Để làm sáng tỏ vấn đề trên, chúng ta cần xem xét kích thước genome của một số loài gần nhau trong bậc thang tiến hóa (có độ phức tạp loài tương tự nhau) cũng như genome của những loài xa nhau (có tính phức tạp khác nhau). Chẳng hạn:

- Genome của người có kích thước khoảng $3{,}3\times10^9$ bp, trong khi đó genome của những loài lưỡng cư dài khoảng $3{,}1\times10^9$ bp hoặc thực vật có thể lên đến 10^{11} bp. Như vậy, có phải là các loài lưỡng cư có tính phức tạp tương tự cơ thể chúng ta?

- Hay là ngay trong cùng một loại, chúng ta cũng nhận thấy có sự mâu thuẫn về kích thước genome? Ví dụ: ruồi nhà (*Musca domestica*) có genome khoảng $8{,}6\times10^8$ bp, lớn gấp sáu lần kích thước genome của ruồi giấm khoảng $1{,}4\times10^8$bp. Ngoài ra, trong các loài lưỡng cư kích thước genome của chúng cũng thay đổi khá lớn từ 10^9-10^{11} bp. Vì sao ngay trong cùng một loại mà kích thước genome lại biến thiên nhiều như vậy, có phải ruồi nhà có cấu tạo phức tạp hơn nhiều so với ruồi giấm?

Từ những dữ liệu trên, chúng ta có thể nhận định rằng tính phức tạp của loài không liên quan đến kích thước của genome. Tuy nhiên, vai trò của phần genome còn lại (phần không mã hóa) đến nay vẫn chưa được biết nhiều.

4. Tổng số gen được biết ở một số loài eukaryote

Có 6.000 gen ở nấm men *S. cerevisiae*, 18.500 gen ở giun tròn, 13.600 gen ở ruồi giấm, 25.000 gen ở *Arabidopsis*, và có khả năng 30.000 gen ở chuột và < 30.000 gen ở người.

Như chúng ta đã biết, mối quan hệ giữa kích thước genome và số lượng gen đã không còn nữa. Genome của các sinh vật eukaryote đơn bào