ĐOÁN NHẬN GEN BẰNG KỸ THUẬT PHÂN CỤM TRONG TIN HỌC

Phan Thị Thanh Thủy*

Tóm tắt

Bài toán thường gặp trong sinh học đó là phân chia tập các dữ liệu thí nghiệm thành các cụm sao cho các điểm dữ liệu trong cùng cụm có độ tương đồng cao, và nếu ở khác cụm thì chúng sẽ khác biệt nhau. Có nhiều cách phân cụm, và không có cách phân cụm nào được cho là tốt nhất mà nó tùy thuộc vào mục đích của việc phân cụm. Việc phân cụm các gen sẽ hy vọng các gen cùng cụm có liên quan với nhau cùng thực hiện một chức năng nào đó. Từ đó có thể tìm ra chức năng của một số gen mới dựa vào những gen đã biết trước đó. Các nhà sinh học sẽ quyết định chọn cách phân cụm nào là hợp lý nhất.

Từ khóa: *phân cụm gen, kỹ thuật, tin học*

Tin sinh học là một lĩnh vực nghiên cứu khá mới ở Việt Nam được ra đời bởi sự kết hợp giữa hai ngành khoa học chính là công nghệ thông tin và công nghệ sinh học. Tin sinh học hiện đang và sẽ tiếp tục đóng góp nhiều thành tựu trong khoa học sinh học, như tìm ra nguyên nhân các loại bệnh làm đẩy nhanh quá trình chẩn đoán bệnh và tìm ra các loại thuốc chữa bệnh mới, tìm ra các giống cây trồng và vật nuôi mới cho năng suất cao. Việc phân tích về sự giống nhau giữa các chuỗi DNA từ các sinh vật khác nhau cũng mở ra hướng mới trong việc nghiên cứu lí thuyết tiến hóa.

Với sự phát triển mạnh trong cả hai lĩnh vực là công nghệ sinh học và công nghệ thông tin, ngày nay một khối lượng khổng lồ dữ liệu sinh học phân tử được thu thập và phục vụ cho quá trình nghiên cứu. Một trong những ví dụ tiêu biểu nhất có lẽ là sự hoàn thành việc giải mã bản đồ gen của người vào năm 2003. Bộ gen của người bao gồm khoảng 3 tỷ nucleotide và được lưu trữ dưới dạng số hóa. Tuy nhiên, việc giải mã thành công bộ gen của người hay các sinh vật khác như chuột hay lúa mới chỉ là bước đầu tiên trong quá trình tìm hiểu về chúng. Và để hiểu được chức năng của tất cả các gen lại là một bài toán khác và còn lâu mới giải quyết xong, cũng như nhiều bài toán khác đang được quan tâm nghiên cứu.

Số lượng gen trong một loài là rất lớn, vì vậy ứng dụng các thuật toán vào việc biểu diễn gen sẽ giúp giảm bớt số lượng các thí nghiệm, rút ngắn thời gian nghiên cứu, giảm bớt công sức và chi phí đáng kể.

1. Các khái niệm cơ bản về sinh học

Mọi sinh vật được cấu tạo bởi các tế bào. Mỗi tế bào là một hệ thống phức tạp gồm nhiều khối tạo dựng khác nhau bọc bởi các màng. Trong cơ thể người có khoảng 6x1013 tế bào, với khoảng 320 kiểu khác nhau, như tế bào da, cơ bắp,

* ThS, Khoa KT-CN, Trường ĐH Phú Yên

não…Một đặc tính cơ bản của mọi tế bào sống là khả năng phát triển trong một môi trường thích hợp và trải qua sự phân chia tế bào. Mục tiêu hàng đầu của tin sinh học gắn liền với quá trình phân tích các thông tin sinh học đó.

1.1 DNA

DNA nằm trong nhân tế bào, được biết đến như là chất hóa học chứa các thông tin di truyền ở hầu hết các sinh vật sống. Về cấu tạo, bất kì chuỗi ADN nào cũng đều chứa 4 loại nucleotide là A, T, G và C. Trong xử lý dữ liệu tin học, trình tự DNA được xử lý như chuỗi các ký tự.

1.2 Gen

Gen cấu trúc là đoạn DNA mang thông tin cần thiết mã hóa một chuỗi polypeptide. Trong đó, các polypeptide là thành phần cấu trúc tạo nên các protein. Đây là nhóm phân tử đóng vai trò quan trọng trong việc quy định kiểu hình của sinh vật.

1.3 Sự biểu hiện của gen

Biểu hiện gen (gene expression), chỉ mọi quá trình liên quan đến việc chuyển đổi thông tin di truyền chứa trong gen để chuyển thành các axit amin (hay protein) (mỗi loại protein sẽ thể hiện một cấu trúc và chức năng riêng của tế bào).

Gen được biểu hiện thành protein thông qua con đường phiên mã và dịch mã. Biểu hiện gen là quá trình đa giai đoạn. Từ phân tử DNA thông tin được mã hoá sang mRNA rồi phân tử mRNA được vận chuyển ra ngoài nhân, tại đó thông tin được giải mã để sản xuất ra protein tương ứng. mRNA đóng vai trò là một loại phân tử truyền tải.

2. Phân cụm để giải bài toán tương đồng của gen

Việc xác định chức năng của một gen mới có ý nghĩa rất quan trọng trong các nghiên cứu sinh học và y học. Mỗi gen đảm nhận một chức năng nào đó và có mối liên hệ với các gen khác. Cho *n* gen, trong đó có một số gen đã biết chức năng, người ta muốn tìm ra chức năng của những gen mới trong số đó. Dựa vào kỹ thuật phân cụm, ta có thể xác định được những gen mới này thuộc cụm gen nào. Những gen được xếp cùng một cụm thì ta có thể kết luận rằng nó có liên quan với nhau về chức năng.

2.1 Phân tích biểu hiện của gen

Dựa vào việc phân tích mức độ biểu hiện gen từ dãy DNA trong quá trình điều hòa phiên mã - lượng mRNA được sinh ra trong tế bào trong nhiều thời điểm, điều kiện khác nhau. Không phải tất cả các gen đều có biểu hiện liên tục. Mức độ biểu hiện của gen khác nhau giữa các tế bào hoặc khác nhau theo giai đoạn trong chu trình tế bào. Tất cả các tế bào đều chứa cùng thông tin di truyền, những tế bào khác nhau chỉ ở những gen hoạt động. Trong nhiều trường hợp, hoạt tính của gen được điều hòa ở mức độ phiên mã, cả qua những tín hiệu bắt đầu bên trong tế bào và cả phản ứng với những điều kiện bên ngoài.

Kết quả của các thí nghiệm nghiên cứu này là một ma trận biểu hiện I(n×m), n là số các gen và m cột là số các thí nghiệm, tương ứng với các thời điểm và các điều kiện khác nhau. Phần tử $I_{i,j}$ của ma trận biểu hiện tượng trưng cho mức độ biểu hiện của gen i trong thí nghiệm j; toàn bộ dòng i gọi là mẫu biểu hiện của gen i. Nếu hai gen i_1 và i_2 có mẫu tương đồng nhau thì chúng ta có quyền hi vọng rằng 2 gen này có chức năng tương tự nhau hay chúng có liên quan với nhau trong cùng một quá trình sinh học. Do đó, nếu mẫu biểu hiện của một gen mới mà tương đồng với mẫu biểu hiện của một gen mà ta đã biết chức năng thì nhà sinh học có thể có lí do để nghi ngờ rằng những gen này cùng thực hiện chức năng tương tự hay có liên hệ với nhau. Tuy nhiên, phân tích biểu hiện gen sẽ không thực hiện được nếu dữ liệu sinh ra bị nhiễu với tỉ lệ lỗi cao.

Theo cách phân tích ở trên, mỗi gen được đặc trưng bởi một vectơ *m* chiều, hay là một điểm trong không gian R^m. Như vậy sự tương đồng giữa hai gen có thể được định lượng bằng khoảng cách giữa hai điểm tương ứng của chúng trong không gian R^m. Cuối cùng, ta có ma trận khoảng cách giữa các gen là $(D(i,j))_{n\times n}$, trong đó khoảng cách giữa 2 điểm tương ứng bởi 2 gen trong bài này là khoảng cách Euclide mà được định nghĩa như sau:

Cho *2* điểm X, Y trong không gian *m* chiều có tọa độ lần lượt là $X=(x_1,x_2,\ldots, x_m)$ và $Y=(y_1,y_2, \ldots, y_m)$, thì khoảng cách Euclid giữa chúng được xác định bởi:

$$D(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_m - y_m)^2}.$$

Chẳng hạn, ta có ma trận biểu hiện I ghi lại kết quả thí nghiệm của 10 gen trong 3 thời điểm khác nhau được cho ở Bảng 2.1 dưới đây:

Times	**1hr**	**2hr**	**3hr**
g1	10.0	8.0	10.0
g2	10.0	0.0	9.0
g3	4.0	8.5	3.0
g4	9.5	0.5	8.5
g5	4.5	8.5	2.5
g6	10.5	9.0	12.0
g7	5.0	8.5	11.0
g8	2.7	8.7	2.0
g9	9.7	2.0	9.0
g10	10.2	1.0	9.2

Bảng 2.1 Ma trận biểu hiện I(10x3) của 10 gen trong 3 thời điểm

Từ dữ liệu của ma trận này, ta có thể tính toán ma trận khoảng cách giữa các gen trong không gian 3 chiều theo khoảng cách Euclide được tính như dưới đây.

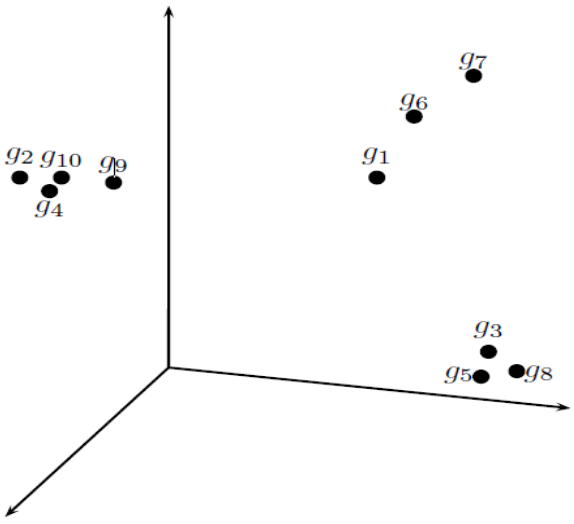
	g1	***g2***	***g3***	***g4***	***g5***	***g6***	***g7***	***g8***	***g9***	***g10***
g1	0.0	***8.1***	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g10	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

Bảng 2.2 Ma trận khoảng cách của 10 gen trong không gian 3 chiều

Trong đó, chẳng hạn khoảng cách giữa hai gen g1 và g2 được xác định bởi:

$$D(1,2) = \sqrt{(10-10)^2 + (0-8)^2 + (9-10)^2} = \sqrt{65} \approx 8.1.$$

Các mẫu biểu hiện là các điểm trong không gian 3 chiều:



Hình 2.1 Biểu diễn các điểm dữ liệu của ma trận I trong không gian 3 chiều.

Theo cách biểu diễn trực quan như trên thì ta thấy các gen có tọa độ gần nhau được phân thành các cụm. Các gen trong cùng cụm thì có khoảng cách gần nhau và

cách xa với các cụm còn lại. Để xử lý việc phân cụm các gen bằng chương trình máy tính thì đòi hỏi phải dùng đến các kỹ thuật phân cụm và các kỹ thuật này được giới thiệu ở phần tiếp theo.

2.2 Các thuật toán phân cụm trong tin sinh học

2.2.1 Giới thiệu kỹ thuật phân cụm

Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm, sao cho các đối tượng trong cùng một cụm tương tự nhau và các đối tượng khác cụm thì không tương tự nhau.

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu và gom các đối tượng dữ liệu này thành từng nhóm phụ thuộc vào sự đánh giá tương tự giữa các đối tượng. Các thuật toán phân cụm đều sinh ra các cụm. Tuy nhiên, để đánh giá hiệu quả của phân tích phân cụm thì không có tiêu chí nào được xem là tốt nhất mà điều này phụ thuộc vào mục đích của phân cụm như: rút gọn dữ liệu, đoán số nhóm tự nhiên trong tập dữ liệu, tìm ra những nhóm có ích và thích hợp, phát hiện ra các nhóm bất thường…

2.2.2 Thuật toán phân cụm các gen

Thuật toán phân cụm nhóm các gen vào trong các cụm và hi vọng rằng những cụm này tương ứng với các nhóm gen liên quan với nhau về chức năng. Để phân cụm, ma trận biểu hiện $I(n \times m)$ được chuyển sang ma trận khoảng cách $D(n \times n)$, trong đó $D_{i,j}$ phản ánh độ tương đồng giữa gen i và gen j. Mục tiêu của phân cụm là nhóm những gen vào các cụm mà thỏa mãn 2 điều kiện sau:

Tính thuần nhất: các gen trong cùng một cụm phải có độ tương đồng cao. Nói cách khác, khoảng cách D(i,j) phải nhỏ nếu gen i và gen j thuộc cùng một cụm.

Tính tách biệt: các gen ở các cụm khác nhau phải rất khác nhau. Nói cách khác, khoảng cách D(i,j) phải lớn nếu gen i và gen j thuộc về các cụm khác nhau.

2.2.3 Phân cụm phân cấp

Phân cụm phân cấp là một kỹ thuật tổ chức các phần tử vào trong một cây, thay vì phân chia dứt khoát các phần tử vào trong các cụm. Trong trường hợp này, những gen được đại diện cho các nút lá của cây. Các cạnh của cây được gán các độ dài và khoảng cách giữa các nút lá là độ dài đường đi nối hai nút lá chính là các giá trị cho trong ma trận khoảng cách.

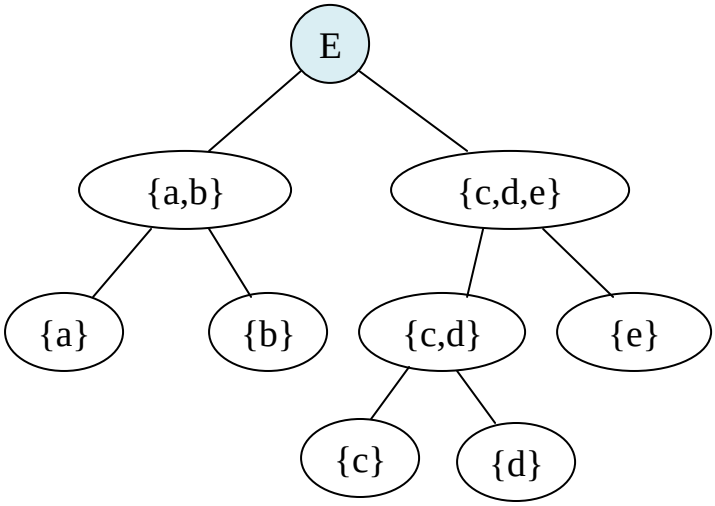
Cho H là một tập khác rỗng các tập con của E. H được gọi là một sự phân cấp trên E nếu các điều kiện sau đây thỏa mãn:

(i) $E \in H$,

(ii) $\forall x \in E, \{x\} \in H$ (lớp đơn),

(iii) $\forall h_i, h_j \in H: h_i \cap h_j \neq \varnothing \rightarrow h_i \subset h_j$ hoặc $h_j \subset h_i$.

Đồ thị biểu diễn một sự phân cấp là một cây, trong đó: gốc biểu diễn lớp lớn nhất là E, n lớp đơn biểu diễn lá của cây. Chẳng hạn, với tập E= {a, b, c, d, e}, ta có một cách phân cụm (không duy nhất) H={E, {a}, {b}, {c}, {d}, {e}, {a,b}, {c,d,e}, {c,d}}, mà được minh họa trong Hình 2.2.



Hình 2.2 Cây gốc E,biểu diễn một phân cấp của tập H.

Phát biểu bài toán phân cụm:

Input: Ma trận khoảng cách giữa các đối tượng $D_{n\times n}$.

Output: Tạo cây phân cấp T biểu diễn các đối tượng.

Những gen trong các cụm này có chức năng tương tự nhau, như thế việc thừa nhận kết quả phân cụm sẽ giúp ta xác định được chức năng của các gen mới, tăng thêm hiểu biết về sinh học.

2.2.4 Phân cụm phân hoạch

Vì *n* dòng trong ma trận biểu hiện I(n×m) chính là tập n điểm trong không gian *m* chiều nên bài toán đặt ra là tìm cách phân chia những điểm này vào k tập con, với giả thiết rằng k là số cụm đã được biết trước. Một trong những phương pháp phân cụm phân hoạch phổ biến nhất đối với các điểm trong không gian đa chiều là phân cụm k-means.

Cho một tập n điểm dữ liệu trong không gian m chiều và một số nguyên k. Vấn đề đặt ra là xác định k điểm hay k-tâm trong không gian m chiều sao cho bình phương khoảng cách từ các điểm đến tâm là nhỏ nhất. Cho một điểm dữ liệu *v* và tập k tâm $\chi = \{x_1, x_2, \ldots, x_k\}$, khoảng cách từ *v* đến các tâm χ được đo bởi khoảng cách từ *v* đến điểm gần nhất trong χ

$$d(v, \chi) = \min_{x_i \in \chi} d(v, x_i).$$

Khoảng cách trung bình bình phương từ tập n điểm $V = \{v_1, v_2, \ldots, v_n\}$ đối với tập tâm $\chi = \{x_1, x_2, \ldots, x_k\}$ được định nghĩa là khoảng cách trung bình bình phương từ các điểm dữ liệu đến tập tâm:

$$d(V, \chi) = \frac{\sum_{i=1}^{n} d(v_i, \chi)^2}{n}.$$

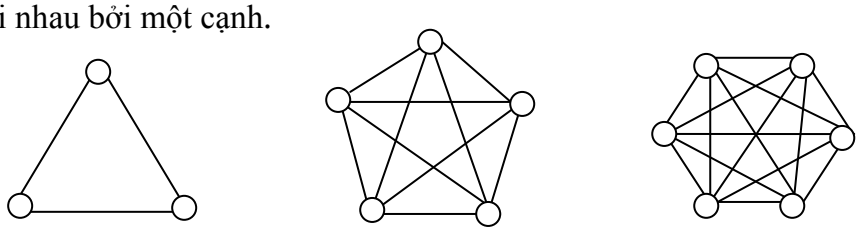
Phát biểu bài toán phân cụm *k*-means:

Input: Một tập V gồm *n* điểm và tham số *k*.

Output: Một tập χ gồm *k* điểm sao cho $d(V, \chi)$ là nhỏ nhất.

2.2.5 Phân cụm dựa trên đồ thị khối

Một đồ thị đầy đủ K_n là một đồ thị vô hướng, gồm *n* đỉnh mà với 2 đỉnh bất kì được nối với nhau bởi một cạnh.

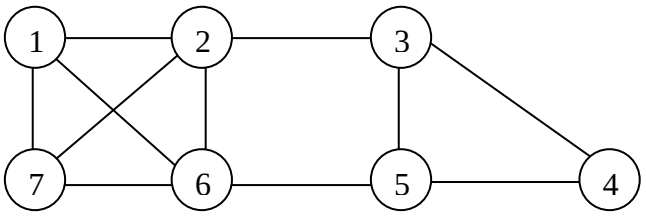


K_3 K_5 K_6

Hình 2.3 Đồ thị đầy đủ với k =3, k=5 và k =6

Một "khối" (clique) trong đồ thị là một đồ thị con đầy đủ lớn nhất mà không được chứa bên trong bất kỳ một đồ thị con đầy đủ bất kì nào khác.

Đồ thị khối (clique graph) là một đồ thị mà mỗi thành phần liên thông là một khối.



Hình 2.4 Đồ thị 7 đỉnh có 4 khối được tạo bởi các đỉnh {1,2,6,7}, {2,3}, {5,6}, {3,4,5}.

Ví dụ ở Hình 2.4 các đỉnh 1,6,7 tạo thành một đồ thị con đầy đủ nhưng nó không tạo thành một khối, còn các đỉnh 1,2,6,7 thì tạo thành một khối.

Cho n gen, ta có thể xây dựng một đồ thị với n đỉnh, trong đó hai gen có khoảng cách gần nhau dưới một ngưỡng cho phép được xem như có cạnh nối giữa chúng. Rõ ràng nếu đây là đồ thị khối thì việc chọn mỗi khối là một cụm là cách phân cụm hợp lý nhất. Tuy nhiên thường thì đây không phải là đồ thị khối. Vấn đề là

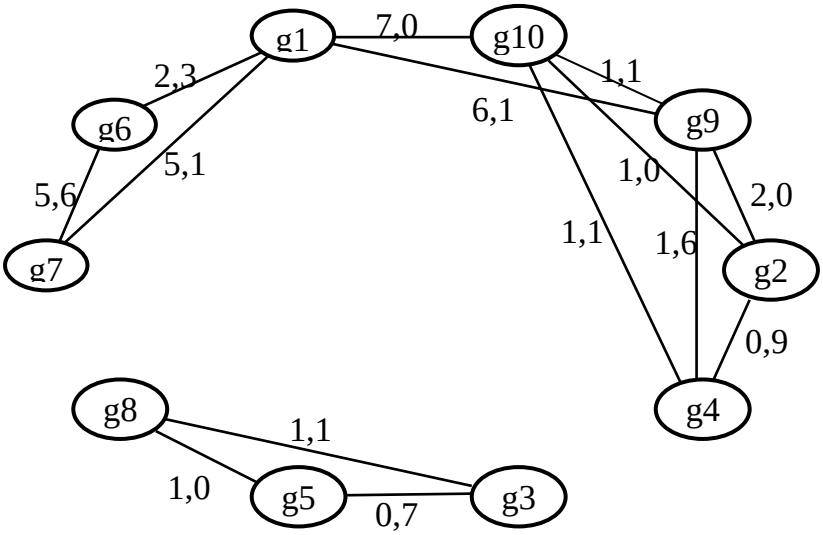
tìm một đồ thị khối xấp xỉ với đồ thị này. Vì vậy, mỗi cách chia n phần tử vào k cụm có thể được biểu diễn bởi một đồ thị khối gồm n đỉnh với k thành phần. Một tập con các đỉnh V'⊂V trong đồ thị G(V,E) có dạng một đồ thị con đầy đủ nếu đồ thị chỉ dựa trên các đỉnh này là đầy đủ. Tức là, hai đỉnh *v* và *w* trong V' được nối với nhau bởi một cạnh trong đồ thị.

Trong những nghiên cứu phân tích biểu hiện gen, ma trận khoảng cách $(D_{i,j})_{n\times n}$ thường được chuyển thành đồ thị khoảng cách G = G(θ), trong đó các đỉnh là các gen và có một cạnh nối giữa gen i và j nếu và chỉ nếu khoảng cách giữa chúng nhỏ hơn ngưỡng θ, tức là $D_{i,j} < \theta$. Các gen được phân cụm nếu thỏa mãn hai tính chất tính đồng nhất và tính tách biệt với một ngưỡng θ được chọn phù hợp sẽ tương ứng với một đồ thị khoảng cách và đồ thị này cũng là một đồ thị khối. Tuy nhiên, những sai lệch trong dữ liệu biểu hiện và ngưỡng θ không thích hợp thường cho kết quả trả về là một đồ thị khoảng cách mà không phải là đồ thị khối. Ta có thể gặp trường hợp một vài gen có khoảng cách nhỏ hơn ngưỡng θ nhưng những gen này lại không có liên quan nhau, điều này dẫn đến trên đồ thị sẽ xuất hiện thêm các cạnh được nối với nhau ở các cụm khác nhau. Trong khi đó, có những gen khác có khoảng cách trong ma trận khoảng cách lại vượt quá ngưỡng θ nhưng những gen này lại có liên quan với nhau, điều này sẽ dẫn đến trên đồ thị có các cạnh liên quan nhau lại bị xóa bỏ trong cụm. Như vậy những cạnh không đúng sẽ làm cho đồ thị khoảng cách không phải là đồ thị khối, người ta gọi là đồ thị khối rạn nứt (Corrupted Cliques). Bài toán đặt ra, làm thế nào để có thể chuyển từ đồ thị khoảng cách sang đồ thị khối với số cạnh thêm vào và xóa đi là ít nhất.

Ma trận khoảng cách D được cho ở Bảng 2.2 có các khoảng cách dưới ngưỡng θ=7.0 được in đậm.

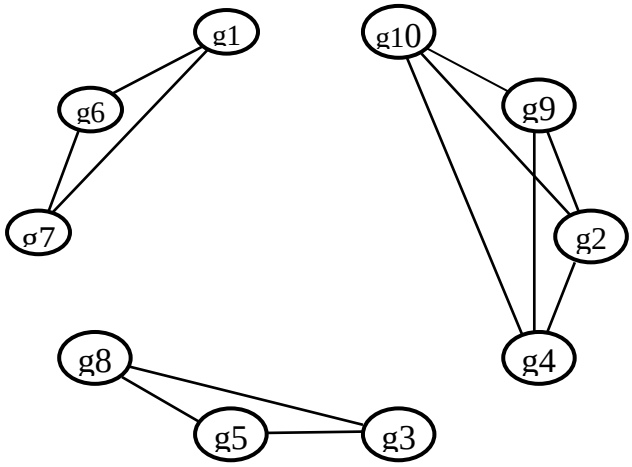
	g1	***g2***	***g3***	***g4***	***g5***	***g6***	***g7***	***g8***	***g9***	***g10***
g1	0	8.1	9.2	7.7	9.3	**2.3**	**5.1**	10.2	**6.1**	**7.0**
g2	8.1	0	12.0	**0.9**	12.0	9.5	10.1	12.8	**2.0**	**1.0**
g3	9.2	12.0	0	11.2	**0.7**	11.1	8.1	**1.1**	10.5	11.5
g4	7.7	**0.9**	11.2	0	11.2	9.2	9.5	12.0	**1.6**	**1.1**
g5	9.3	12.0	**0.7**	11.2	0	11.2	8.5	**1.0**	10.6	11.6
g6	**2.3**	9.5	11.1	9.2	11.2	0	**5.6**	12.1	7.7	8.5
g7	**5.1**	10.1	8.1	9.5	8.5	**5.6**	0	9.1	8.3	9.3
g8	10.2	12.8	**1.1**	12.0	**1.0**	12.1	9.1	0	11.4	12.4
g9	**6.1**	**2.0**	10.5	**1.6**	10.6	7.7	8.3	11.4	0	**1.1**
g10	**7.0**	**1.0**	11.5	**1.1**	11.6	8.5	9.3	12.4	**1.1**	0

Từ ma trận này, ta có đồ thị khoảng cách được thể hiện như sau :



Hình 2.5 Đồ thị khoảng cách

Đồ thị khoảng cách Hình 2.5 đã thể hiện trên đây không phải là đồ thị khối, sau khi loại bỏ 2 cạnh (g1, g10) và (g1, g9) thì đồ thị khoảng cách trên được chuyển thành đồ thị khối.



Hình 2.6 Đồ thị khối sau khi loại bỏ 2 cạnh (g1, g10), (g1, g9) từ đồ thị khoảng cách

Bài toán các khôi bị rạn nứt (Corrupted Cliques Problem).

Giả sử ta có một đồ thị G. Nếu G không phải là đồ thị khối thì bài toán đặt ra là: xác định số cạnh ít nhất cần thêm vào hay xóa đi để chuyển G thành một đồ thị khối.

Input: Đồ thị khoảng cách G

Output: Số cạnh cần thêm vào hay xóa đi ít nhất để chuyển G thành đồ thị khối.

Bài toán các khối bị rạn nứt được đưa ra thuộc lớp bài toán NP-khó, nên người ta đề xuất một vài giải thuật heuristic như PCC (Parallel Classification with

Cores), giải thuật này tốn nhiều thời gian, CAST (Cluster Affinity Search Technique) được kế thừa từ PCC và tỏ ra thực tế hơn.

Ý tưởng giải thuật PCC

Giả sử chúng ta cố gắng phân cụm một tập gen S và S' là tập con của S. Nếu ta đưa ra được một cách phân cụm đúng của S' là $\{C_1, \ldots, C_k\}$. Liệu có thể mở rộng sự phân cụm của S' ra toàn bộ của tập gen S? Đặt S\S' là tập các gen chưa được phân cụm, và $N(j,C_i)$ là số cạnh giữa gen $j \in S\backslash S'$ và những gen trong cụm C_i trong đồ thị khoảng cách. Chúng ta đánh giá sự tương đồng của gen j với cụm C_i là tỷ số sau đây:

$$\frac{N(j,C_i)}{|C_i|}$$

Gen j sẽ được phân vào cụm C_i nếu tỷ số trên có giá trị lớn nhất, ta nói rằng gen j tương đồng với cụm C_i nhất. Theo cách này, sự phân cụm của S' có thể mở rộng ra thành phân cụm toàn bộ tập gen S.

Ý tưởng giải thuật CAST

Định nghĩa khoảng cách giữa gen *i* và cụm *C* là khoảng cách trung bình giữa gen i và tất cả các gen trong cụm C: $d(i, C) = \frac{\sum_{j \in C} d(i, j)}{|C|}$. Cho một ngưỡng θ, gen *i* gần với cụm C nếu $d(i,C) < \theta$ và cách xa cụm C nếu $d(i,C) > \theta$.

Giải thuật CAST phân cụm S theo đồ thị khoảng cách G và ngưỡng θ. CAST tạo ra sự phân chia P của tập S bằng cách tìm cụm C mà không có gen $i \notin C$ mà gần với C và không có gen $i \in C$ cách xa C. P được khởi tạo là tập rỗng.

Kết luận: Trên đây là một số cách phân cụm cơ bản trong kỹ thuật khai phá dữ liệu, tuy nhiên để tìm một thuật toán được cho là tối ưu trên tập dữ liệu lớn đòi hỏi kỹ thuật xử lý có độ phức tạp rất lớn. Tin sinh học liên quan chặt với lĩnh vực data mining (khai phá dữ liệu) and machine learning (học máy) để giải quyết các bài toán sinh học. Bằng công nghệ đó người ta đã tìm ra chức năng của nhiều loại gen mới ở nhiều sinh vật khác nhau giúp cho ngành sinh học phân tử ngày càng phát triển đa dạng, phong phú❑

TÀI LIỆU THAM KHẢO

[1] Nguyễn Văn Cách (2009), *Tin sinh học*, Nxb Khoa học và kỹ thuật.
[2] Lê Minh Hoàng (2002), *Giáo trình giải thuật và lập trình*, Nxb ĐHSP Hà Nội.
[3] Nguyễn Viết Nhân (2007), *Giáo trình Di truyền y học*, Nxb Đại học Huế.
[4] Neil C. Join, Pavel A. Pevzner (2004), *An Introduction to Bioinformatics Algorithms*, A Bradford book The MIT Press Cambridge, Massachuetts London.

[5] Pang-Ning Tan, Michael Steibach, Vipin Kumar (2006), *Introduction to Data Mining*, Michigan Stage University, University of Minnesota and Army High Performance Computing Research Center.
[6] T.Chandrasekhar, K.Thangavel and E.Elayaraja (2011), *Performance Analysis of Enhanced Clustering Algorithm for Gene Expression Data*, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.
[7] Gregory A Wilkin1 and Xiuzhen Huang (2008), *A practical comparison of two K-Means clustering algorithms*, BMC Bioinformatics.

Abstract

Determining genes by clustering algorithms in information technology

A common problem in biology is to divide a set of experimental data into clusters (groups) in such a way that the data points in each cluster are highly similar, while the data points in different clusters are different. There are several algorithms that performs different types of clustering; each situation has its own best way of clustering and there is no common best choice in a general situation. Clustering algorithms group genes with similar expression patterns into clusters with the hope that the genes in each cluster has a common function. It, therefore, helps us to determine the new genes based on the information of already known genes. Biologists will determine the most reasonable choice of clustering.

Key words: *genes clustering, clustering algorithms, information technology*