TIN SINH HỌC ĐẠI CƯƠNG
(Introduction to Bioinformatics)

PGS.TS. Trần Văn Lăng
Email: langtv@vast.vn
langtv@gmail.com

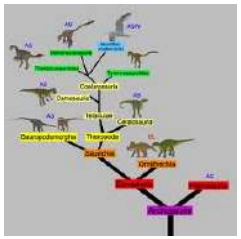
Chương 5:

CÂY PHÂN LOÀI

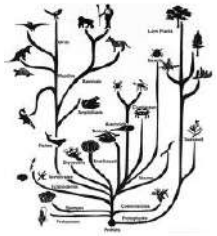
- Khái niệm cây phân loài
- Nguồn gốc cây phân loài
- Các phương pháp xây dựng cây phân loài

Khái niệm

- Cây phân loài (Phylogenetic tree) hay còn gọi là:
  - Cây phả hệ
  - Cây phát sinh loài

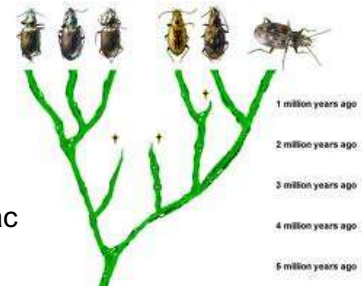


- Cây được dùng để mô hình hóa lịch sử thực tế của một nhóm các trình tự hay các sinh vật.

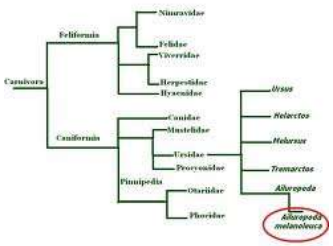


- Đối tượng nghiên cứu truyền thống của cây phân loài là biểu diễn mối quan hệ giữa các loài.

- Khi biểu diễn trong cây phân loài
  - *n* loài hiện tại được biểu diễn ở *n* lá của cây
  - Các nút bên trong (các nhánh) đại diện cho phần chung nay đã tuyệt chủng

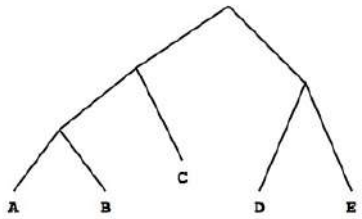


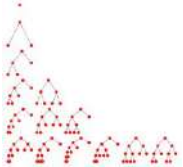
- Các nút bên trong đôi khi còn được coi:
  - Sự đại diện cho một nhóm các loài
  - Một sự kiện riêng biệt



- Cách biểu diễn: có 2 dạng
  - Cây có gốc (rooted tree)
  - Cây không gốc (unrooted tree)
- Gọi là biểu diễn Phylip hay NEWICK

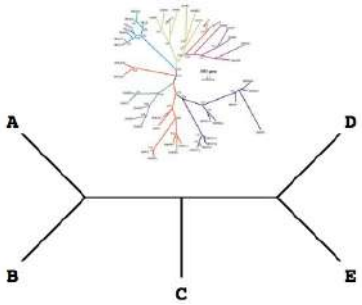
Biểu diễn cây có gốc





(((A,B),C),(D,E))

Các biểu diễn cây không gốc



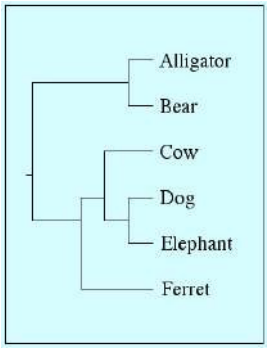
(((A,B),C),(D,E))
((A,B),(C,(D,E)))
(A,(B,(C,(D,E))))
((A,B),C,(D,E))
etc.

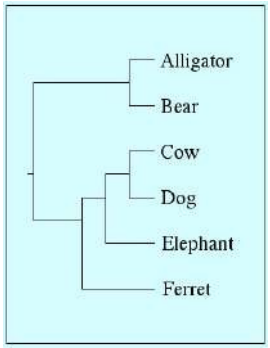
- Biểu diễn cây (A, (B, C)) và ((B, C), A) giống nhau hoàn toàn.
- Theo tự nhiên, cây có nút gốc được vẽ từ dưới lên.
- Tuy nhiên, khi biểu diễn cây có gốc thường từ đỉnh xuống hoặc từ trái sang phải.
- Cây không gốc được vẽ từ trung tâm đi ra.

Ví dụ: cá sấu, …, chồn

((Alligator,Bear),((Cow,(Dog,Elephant)),Ferret))

((Alligator,Bear),(((Cow,Dog),Elephant),Ferret))

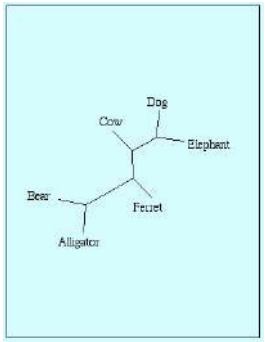


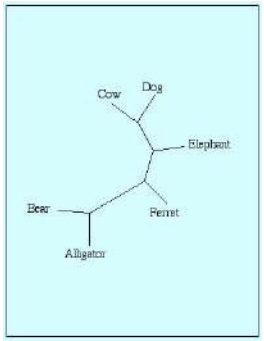


Trường hợp cây không gốc

((Alligator,Bear),((Cow,(Dog,Elephant)),Ferret))

((Alligator,Bear),(((Cow,Dog),Elephant),Ferret))

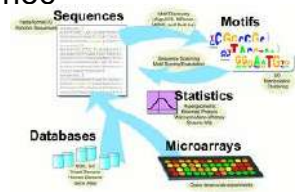




PHƯƠNG PHÁP KHOẢNG CÁCH ĐỂ TẠO CÂY PHÂN LOÀI

Phương pháp UPGMA

- UPGMA (Unweighted Pair Group Method using arithmetic Averages)
- Là phương pháp gom cụm không có trọng số dùng trung bình số hoc



Phương pháp

D-UPGMA

- Trên cơ sở khoảng cách giữa từng cặp trình tự, biểu diễn thành dạng ma trận khoảng cách
- Ma trận khoảng cách là ma trận đối xứng
- Trên cơ sở ma trận khoảng cách, tìm các cụm gần nhất một cách lần lượt

Khoảng cách trong cây phân loài

- Ma trận khoảng cách D = (d_{ij}) là ma trận trong đó mỗi phần tử d_{ij} là khoảng cách giữa 2 nút lá trong cây phân loài.
- Ngoài ra, trong cây phân loài, còn chỉ rõ khoảng cách giữa các nút lá và các nút bên trong cây.

- Khoảng cách trình tự d_{ij} thỏa mãn các điều kiện sau đây:
  - Tính đối xứng: $d_{ij} = d_{ji}$ với mọi i, j
  - Tính phân biệt: $d_{ij} \neq 0$ nếu và chỉ nếu $i \neq j$
  - Bất đẳng thức tam giác: $d_{ij} \leq d_{ik}+d_{kj}$ với mọi i, j, k

- Về mặt sinh học, độ dài cạnh d_{ij} tương ứng với thời gian trôi qua từ khi phân tách i và j khỏi nút chung.
- Điều đó có nghĩa chiều dài cạnh được đo bởi một “molecular clock” với tỉ lệ không đổi.

Minh họa

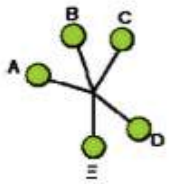
- Cho 5 trình tự A, B, C, D, E
- Từ đây, suy ra cần 10 khoảng cách giữa 5 trình tự này để tạo ma trận khoảng cách
  - 10 = n(n-1)/2, với n = 5

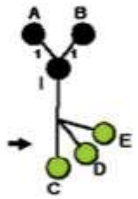
Ví dụ

- Giả sử 5 trình tự này có ma trận khoảng cách như bảng
- Lần lượt tính toán khoảng cách giữa các trình tự gom nhóm và không gom nhóm

	A	B	C	D	E
A					
B	2				
C	6	6			
D	4	4	6		
E	7	7	9	5	

- Trong ma trận này, khoảng cách giữa A và B là ngắn nhất, nên gom nhóm A và B lại.
- Như vậy, A và B có chung tổ tiên là I
- Chiều dài từ I đến A (hoặc B) là 1.



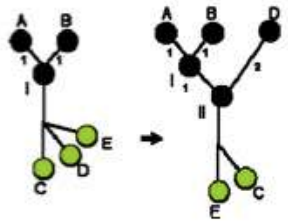


- Tính lại ma trận khoảng cách trong đó có khoảng cách giữa nhóm AB với các loài (trình tự) C, D, E còn lại
- Khoảng cách từ một loài đến nhóm là khoảng cách trung bình từ loài này đến các loài trong nhóm

- $d_{(AB)C} = (d_{AC}+d_{BC})/2$
- $d_{(AB)D} = (d_{AD}+d_{BD})/2$
- $d_{(AB)E} = (d_{AE}+d_{BE})/2$
  - Kết quả như bảng:

	AB	C	D	E
AB				
C	6			
D	4	6		
E	7	9	5	

- Sau khi có ma trận khoảng cách mới, tiếp tục gom cụm với tiêu chí khoảng cách nhỏ nhất được chọn
- 4 là khoảng cách nhỏ nhất, nên nhóm AB được gom cụm với trình tự D để có nhóm (AB)D
- Khoảng cách từ AB (hoặc D) đến II là 4/2 = 2

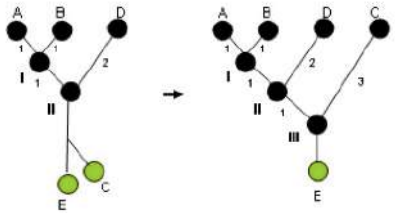


Có chung tổ tiên là II

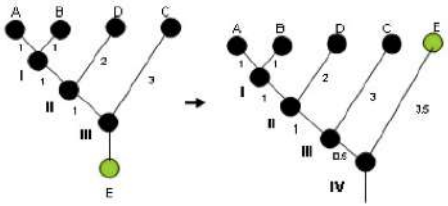
- Tính toán khoảng cách trung bình từ nhóm AB,D đến các trình tự còn lại theo quy tắc trên
- $d_{(AB,D)C} = (d_{AB,C}+d_{DC})/2$
- $d_{(AB,D)E} = (d_{AB,E}+d_{DE})/2$

	AB,D	C	E
AB,D			
C	6		
E	6	9	

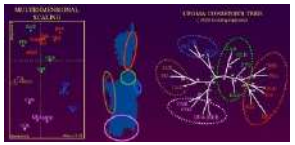
- Theo ma trận khoảng cách mới, giá trị nhỏ nhất là 6 nên có thể chọn cụm ((AB)D)C với nút trung tâm III
- Khoảng cách từ III đến (AB)D (hoặc đến C) là 6/2 = 3.



- Tương tự, khoảng cách giữa cụm ((AB)D)C với trình tự E là:
- $d_{((AB,D)C)E} = (d_{(AB,D)E} + d_{CE})/2 = (6+9)/2 = 7,5$
- Khoảng cách là 7,5/2 = 3,75



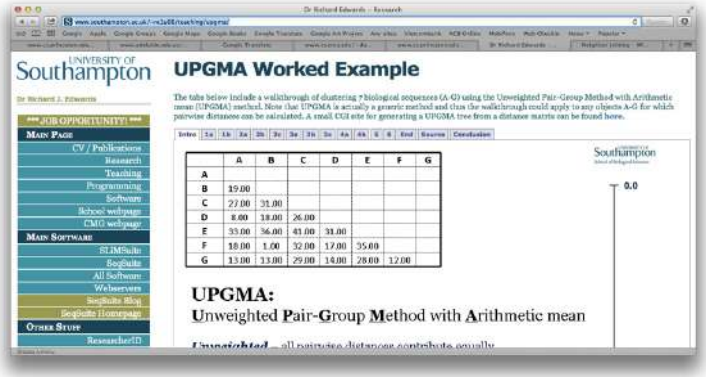
Bài tập



	A	B	C	D	E
A	0	200	300	600	600
B		0	300	600	600
C			0	600	600
D				0	200
E					0

- Hãy vẽ cây với ma trận khoảng cách như các bảng bên cạnh

	A	B	C	D	E
A					
B	17				
C	22	30			
D	32	34	29		
E	23	21	41	45	



- Minh họa trên web

Tổng quát về phương pháp gom cụm

- Có 4 phương pháp gom cụm
- Những phương pháp này khác nhau ở cách tính khoảng cách:
  - Liên kết đơn giản: $d_{ku} = \min(d_{ki}, d_{kj})$
  - Liên kết phức tạp: $d_{ku} = \max(d_{ki}, d_{kj})$
  - UPGMA: $d_{ku} = (d_{ki} + d_{kj})/2$
  - WPGMA: $d_{ku} = (n_i d_{ki} + n_j d_{kj})/(n_i + n_j)$

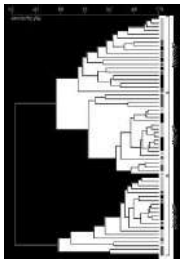
    Trong đó n_i là số phần tử của cụm i

Thuật toán

- Bao gồm 5 bước
  1. Tìm cặp cụm (i,j) có khoảng cách d_{ij} là bé nhất
  2. Tạo cụm u gồm cụm i và j
  3. Tính chiều cao của cụm u (khoảng cách đến lá) là $l_{ij} = d_{ij}/2$
  4. Tính khoảng cách d_{ku} với k không thuộc cụm u
  5. Loại cụm u (cụm i,j) từ ma trận khoảnh cách

- Sự khác nhau giữa các phương pháp
  - Liên kết đơn giản: $d_{ku} = \min(d_{ki},d_{kj})$
  - Liên kết phức tạp: $d_{ku} = \max(d_{ki},d_{kj})$
  - WPGMA: $d_{ku} = (n_i d_{ki} + n_j d_{kj})/(n_i+n_j)$
  - UPGMA: $d_{ku} = (d_{ki} + d_{kj})/2$

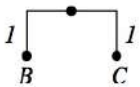
    Trong đó n_i là số phần tử của cụm i



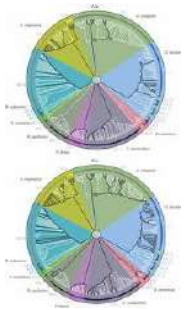
Ví dụ

- Cho các trình tự ký hiệu A B, C, D, E và ma trận khoảng cách như hình.
- Khoảng cách $d_{BC} = 2$ là nhỏ nhất
- Liên kết B và C thành cụm (BC) với độ cao là $d_{bc}/2 = 2/2 = 1$

	A	B	C	D	E
A	0	8	8	12	8
B		0	2	12	4
C			0	12	4
D				0	12
E					0

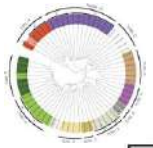


- Tính các khoảng cách theo WPGMA
  - $d_{A(BC)} = (1x8 + 1x8)/(1+1) = 8$
  - $d_{D(BC)} = (1x12 + 1x12)(1+1) = 12$
  - $d_{E(BC)} = (1x4 + 1x4)/(1+1) = 4$

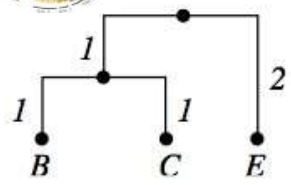


- Loại bỏ B, C để có ma trận khoảng cách mới

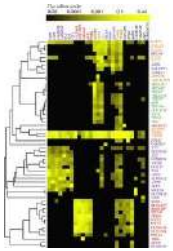
	A	(BC)	D	E
A	0	8	12	8
(BC)		0	12	4
D			0	12
E				0



- Theo ma trận khoảng cách: khoảng cách giữa cụm (BC) và E là bé nhất
- Nên tạo cụm (BC) với E để có cụm (BC)E với chiều cao là 4/2 = 2



- Tiếp tục tính khoảng cách từ cụm (BC)E đến các trình tự còn lại
  - $d_{A((BC)E))} = (2xd_{A(BC)} + 1xd_{AE})/(2+1)$
  - $= (2x8 + 1x8)/3 = 8$
  - $d_{D((BC)E))} = (2xd_{D(BC)} + 1xd_{DE})/(2+1)$
  - $= (2x12 + 1x12)/3 = 12$

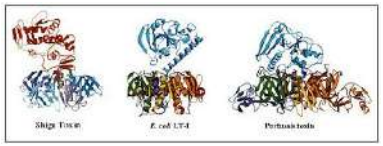


- Ma trận khoảng cách mới được viết lại

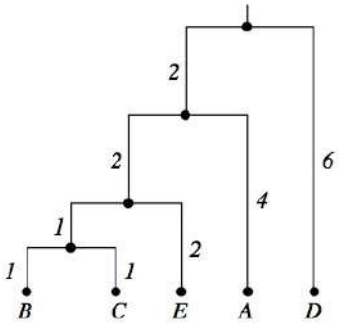
	A	$((BC)E)$	D
A	0	8	12
$((BC)E)$		0	12
D			0

- Do khoảng cách giữa A và cụm (BC)E là bé nhất, nên tạo cụm mới ((BC)E)A có chiều cao bằng 8/2 = 4

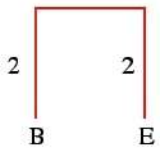
- Khoảng cách giữa D với cụm ((BC)E)A
  - $d_{D((BC)E)A} = (3xd_{D((BC)E)} + 1xd_{DA})/(3+1)$
  - $= (3x12 + 1x12)/4 = 12$
- Từ đây suy ra chiều cao của cây là 12/2 = 6



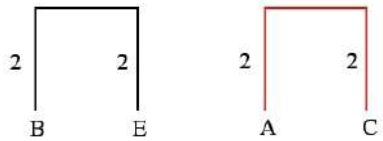
- Kết quả



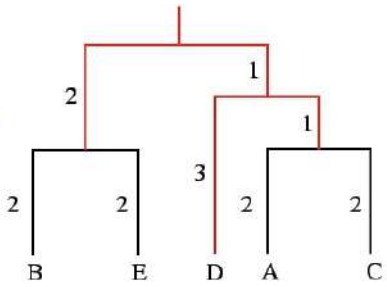
	A	B	C	D	E
A	0	8	4	6	8
B		0	8	8	4
C			0	6	8
D				0	8
E					0

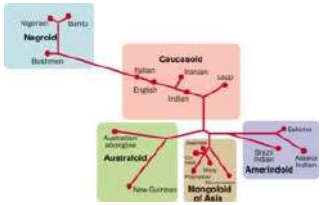


	A	(B E)	C	D
A	0	8	4	6
(B E)		0	8	8
C			0	6
D				0



	$\binom{A}{C}$	$\binom{B}{E}$	D
(A C)	0	8	6
(B E)		0	8
C			0





PHƯƠNG PHÁP NEIGHBOR - JOINING

- Do Naruya Saitou và Masatoshi Nei đưa ra vào năm 1987

- Với cách thức vẽ cây phân loài có gốc, chúng ta luôn nhận được khoảng cách (molecular clock) sau khi vẽ không còn. đúng với ban đầu.
- Chẳng hạn, với

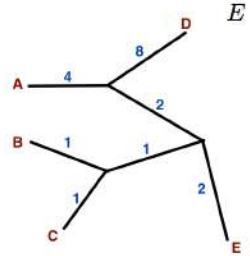
	A	B	C	D	E
A	0	8	8	12	8
B		0	2	12	4
C			0	12	4
D				0	12
E					0

- Kết quả sẽ là như bên
- Trong trường hợp này cần vẽ cây không có gốc

	A	B	C	D	E
A		8	8	12	12
B			2	12	4
C				12	4
D					12
E					

- Cây không gốc ứng với ma trận khoảng cách như bên sẽ là

	A	B	C	D	E
A	0	8	8	12	8
B		0	2	12	4
C			0	12	4
D				0	12
E					0

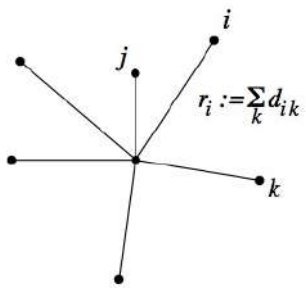


Neighbor - Joining

- Phương pháp Neighbor – Joining là phương pháp tương tự như phương pháp gom cụm.
- Tuy nhiên, khái niệm cụm hàng xóm có khác:
  - Hai trình tự được gọi là hàng xóm (lân cận) trong một cây nếu như giữa chúng chỉ có duy nhất một nút.

Phương pháp

- Cho ma trận khoảng cách chứa khoảng cách d_{ij} giữa các trình tự trong tập hợp n trình tự.
- Các trình tự ban đầu được biểu diễn như hình ngôi sao.



Các bước

- Bước 1: Ở mỗi nút i có thể tính tổng khoảng cách r_i:

$$r_i = \sum_{k=1}^{n} d_{ik}$$

- Bước 2: Mỗi cặp nút lá tính M_{ij}, lấy các giá trị nhỏ nhất.

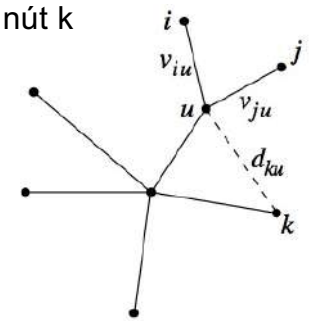
$$M_{ij} = d_{ij} - \frac{r_i + r_j}{n-2}$$

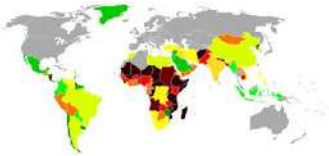
- Bước 3: Liên kết nút i và nút j thành một nút mới ký hiệu u. Khi đó chiều dài từ u đến i và j là:

$$v_{iu} = \frac{d_{ij}}{2} + \frac{r_i - r_j}{2n-4}, \text{ và } \quad v_{ju} = d_{ij} - v_{iu}$$

- Bước 4: Từ đây có thể tính khoảng cách từ u đến nút k khác là:

$$d_{ku} = \frac{d_{ik} + d_{jk} - d_{ij}}{2}$$



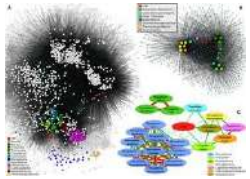


- Bước 5: Xóa nút i và j từ ma trận khoảng cách. Nếu còn lại nhiều hơn 2 cụm, quay trở lại bước 1

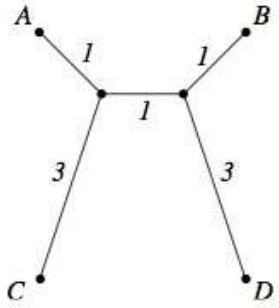
Ví dụ

- Cho ma trận khoảng cách với n = 4 trình tự ký hiệu A, B, C, D

	A	B	C	D
A	0	3	4	5
B		0	5	4
C			0	7
D				0



- Khoảng cách d_{AB} là nhỏ nhất, nhưng có thể A, B không phải là láng giềng; mà có thể là A, C như hình bên.
- Vì vậy, khoảng cách nhỏ nhất không cần thiết.



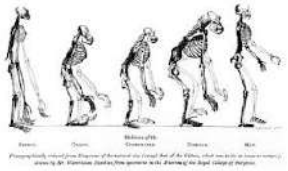
Bước 1

- $r_A = d_{AB} + d_{AC} + d_{AD} = 3 + 4 + 5 = 12$
- $r_B = d_{BA} + d_{BC} + d_{BD} = 3 + 5 + 4 = 12$
- $r_C = d_{CA} + d_{CB} + d_{CD} = 4 + 5 + 7 = 16$
- $r_D = d_{DA} + d_{DB} + d_{DC} = 5 + 4 + 7 = 16$

Bước 2

- $M_{AB} = d_{AB} – (r_A + r_B)/(4-2) = 3 – 24/2 = -9$
- $M_{AC} = d_{AC} – (r_A + r_C)/(4-2) = 4 – 28/2 = -10$
- $M_{AD} = d_{AD} – (r_A + r_D)/(4-2) = 5 – 28/2 = -9$
- $M_{BC} = d_{BC} – (r_B + r_C)/(4-2) = 5 – 28/2 = -9$
- $M_{BD} = d_{BD} – (r_B + r_D)/(4-2) = 4 – 28/2 = -10$
- $M_{CD} = d_{CD} – (r_C + r_D)/(4-2) = 7 – 32/2 = -9$

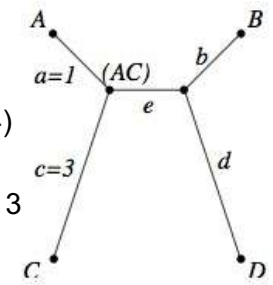
Giá trị nhỏ nhất là M_{AC} và M_{BD}

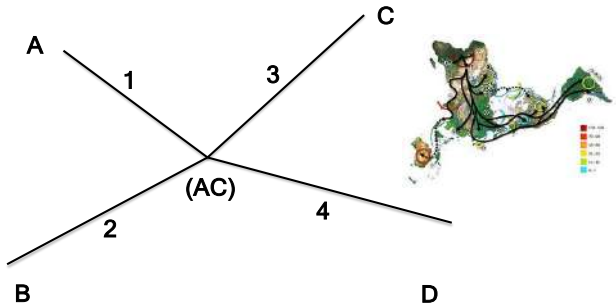


- Như vậy có 2 cụm là AC và BD
- Sử dụng cụm AC, tạo ra nút mới ký hiệu (AC) ở giữa 2 nút A, C này.

Bước 3

- Khi đó
  - $d_{A(AC)} = d_{AC}/2 + (r_A - r_C)/(2x4-4)$
  - $= 4/2+(12-16)/4 = 1$
  - $d_{C(AC)} = d_{AC} - d_{A(AC)} = 4 – 1 = 3$





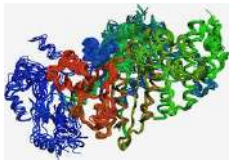
Bước 4

- Khoảng cách các nút còn lại (B và D) đến nút (AC) được tính như sau:
- $d_{B(AC)} = (d_{AB} + d_{CB} – d_{AC})/2$
- $= (3 + 5 – 4)/2 = 2$
- $d_{D(AC)} = (d_{AD} + d_{CD} – d_{AC})/2$
- $= (5 + 7 - 4)/2 = 4$

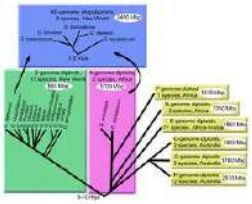
Bước 5

- Loại bỏ trình tự A và C, ma trận khoảng cách còn lại như bên cạnh

	(AC)	B	D
(AC)	0	2	4
B		0	4
D			0

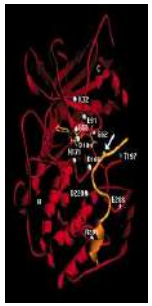


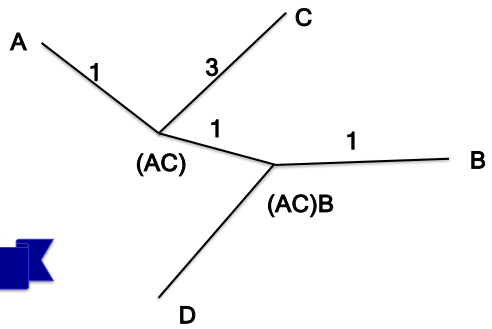
- Tiếp tục quay lại Bước 1 với n = 3
  - $r_{AC} = d_{(AC)B} + d_{(AC)D} = 2 + 4 = 6$
  - $r_B = d_{B(AC)} + d_{BD} = 2 + 4 = 6$
  - $r_D = d_{D(AC)} + d_{DB} = 4 + 4 = 8$



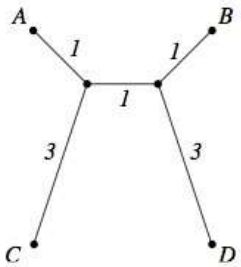
- Với Bước 2:
  - $M_{(AC)B} = d_{(AC)B} – (r_{AC} + r_B)/(3-2)=2-(6+6)/(3-2)= -10$
  - $M_{(AC)D} = d_{(AC)D} – (r_{AC} +r_D)/(3-2)=4-(6+8)/(3-2)= -10$
  - $M_{BD} = d_{BD} – (r_B +r_D)/(3-2)=4-(6+8)/(3-2)= -10$
- Cả 3 đều có giá trị -10, nên có thể gom thành cụm (AC)B

- Tính toán theo Bước 3:
  - $d_{AC((AC)B)} = d_{(AC)B}/2 + (r_{AC} - r_B)/(2x3-4)$
  - $= 2/2+(6-6)/2 = 1$
  - $d_{B((AC)B)} = d_{(AC)B} – d_{AC((AC)B)} = 2 – 1 = 1$



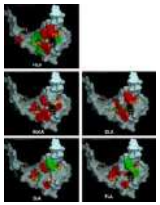


- Tính khoảng cách từ nút còn lại (Bước 4)
  - $d_{((AC)B)D} = (d_{(AC)D} + d_{BD} - d_{(AC)B})/2$
  - $= (4 + 4 - 2)/2 = 3$
- Khi đó có cây như hình

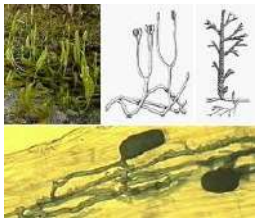


Bài tập

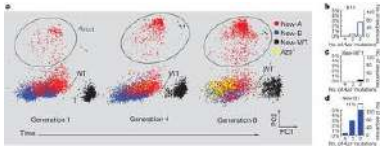
- Vẽ cây không gốc theo Neighbor – Joining với ma trận khoảng cách là:



	A	B	C	D
A	0	7	11	14
B	7	0	6	9
C	11	6	0	7
D	14	9	7	0



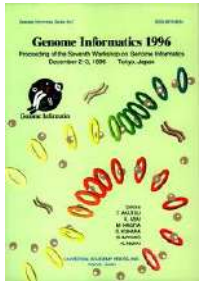
KHOẢNG CÁCH GIỮA CÁC TRÌNH TỰ



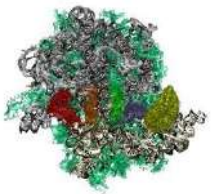
- Khoảng cách của 2 trình tự là tỷ số giữa các trính tự không bắt cặp (đột biến) và số cặp không kể gap.
- Thực chất đó là số nucleotide khác nhau giữa 2 trình tự

- Cho 4 trình tự A, B, C, D, mỗi trình tự có 20 nucleotide:

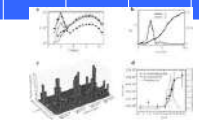
A. AGGCCATGAATTAAGAATAA
B. AGCCCATGGATAAAGAGTAA
C. AGGACATGAATTAAGAATAA
D. AAGCCAAGAATTACGAATAA



- Khoảng cách giữa
  - A và B là 4/20 (có 4 mismatch)
  - A và C là 1/20
  - A và D là 3/20
  - B và C là 5/20
  - B và D là 7/20
  - C và D là 4/20



- Ma trận khoảng cách có thể viết

	A	B	C	D
A		4	1	3
B			5	7
C				4
D				

	A	B	C	D
A		0,2	0,05	0,15
B			0,25	0,35
C				0,2
D