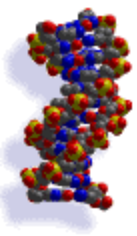


Đại học Quốc gia thành phố Hồ Chí Minh
Trường ĐH Khoa học Tự nhiên
Khoa Sinh học

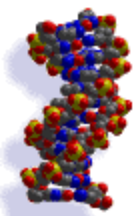
Bioinformatics

ThS. Cao Thị Ngọc Phượng



Nội dung

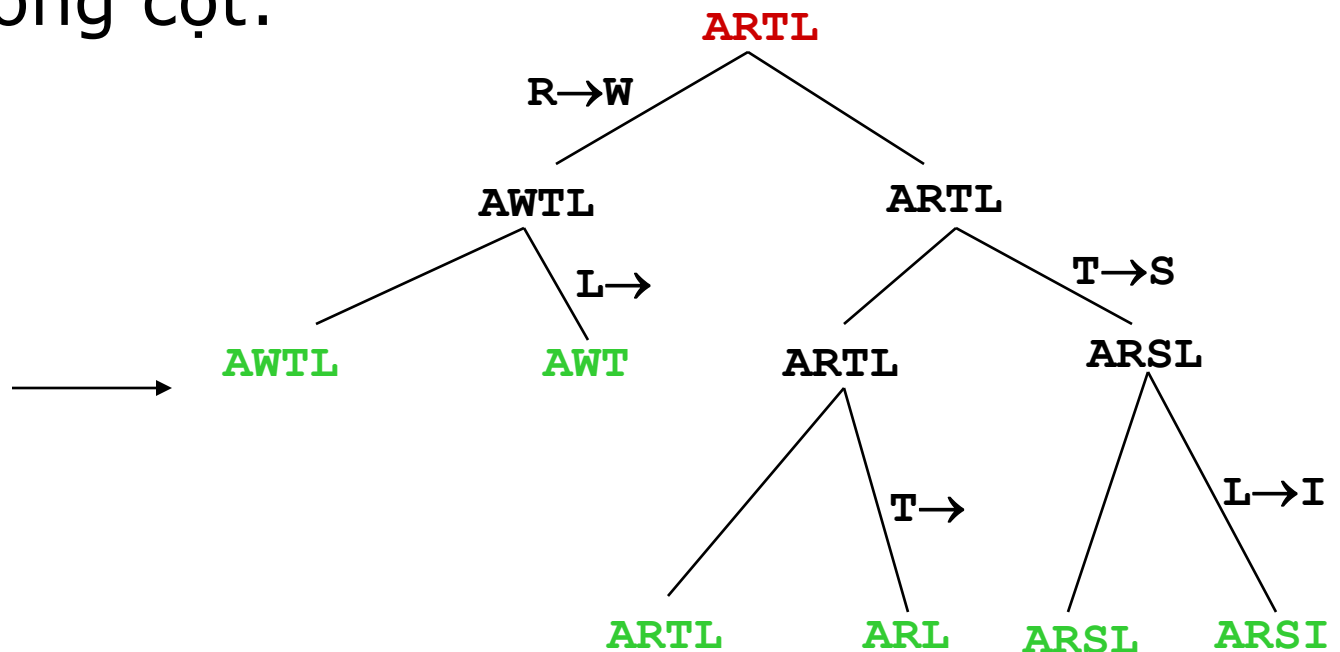
- ❑ Tìm kiếm trình tự tương đồng trong CSDL: Blast
 - ❑ Sắp giống cột nhiều trình tự sinh học (MSA)
 - ❑ **Tiến hóa phân tử và cây phát sinh loài**
 - ❑ Tin sinh học cấu trúc protein
-

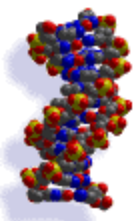


MSA và cây tiến hóa (Phylogenetic Tree - PT)

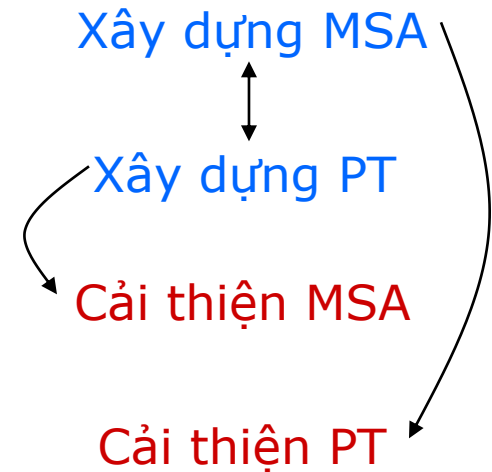
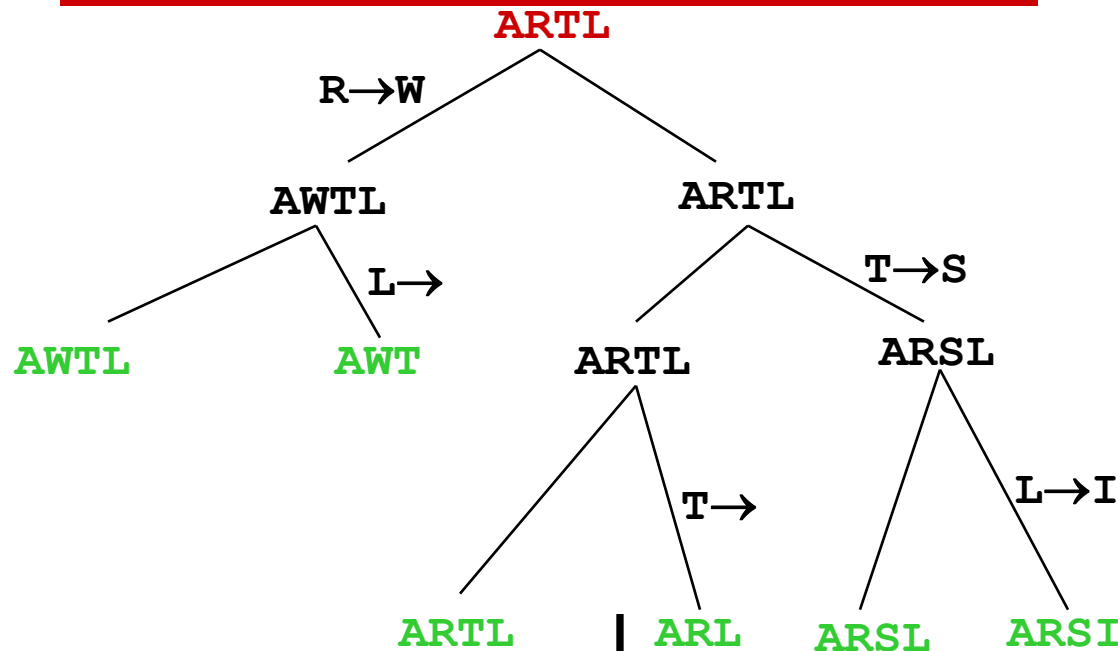
- ❑ Tập trình tự {**ARL**, **ARTL**, **ARSI**, **ARSL**, **AWTL**, **AWT**}
- ❑ Sắp giống cột:

AR-L
ARTL
ARSI
ARSL
AWTL
AWT-





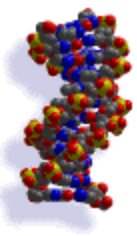
MSA và cây tiến hóa (PT)



AWTL	ARTL	ARSL
AWT-	AR-L	ARSI

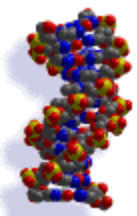
AWTL	ARTL
AWT-	AR-L
ARSL	ARSI

AWTL
AWT-
ARTL
AR-L
ARSL
ARSI

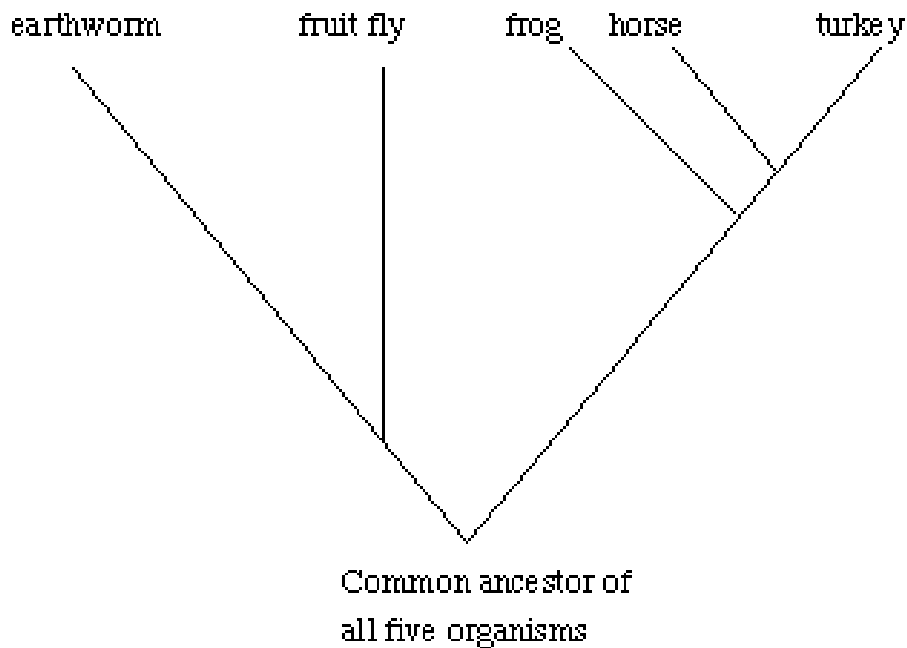


Mục tiêu xây dựng cây phát sinh loài

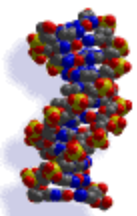
- ❑ Xây dựng phả hệ (mối quan hệ) giữa các sinh vật
- ❑ Ước tính thời gian tiến hóa phân kỳ của các sinh vật từ tổ tiên chung



Thuật ngữ

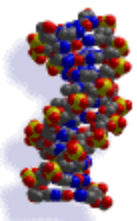


- Cây tiến hóa: hình ảnh thể hiện mối quan hệ tiến hóa sử dụng các node và các nhánh
- Node: một đơn vị phân loại (OTU)
- Nhánh: xác định mối quan hệ
- Chiều dài nhánh: số thay đổi xảy ra trong nhánh
- Node ngoại: OTU
- Node nội: tổ tiên chung



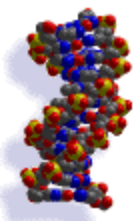
Cơ sở xây dựng cây phát sinh loài

- ❑ Dựa trên các đặc tính hình thái, đặc tính sinh hóa... của sinh vật
- ❑ Dựa trên các trình tự sinh học (cây phát sinh loài mức phân tử):
 - Sự phát triển bùng nổ của dữ liệu trình tự
 - Có công cụ so sánh các trình tự và tạo cây
 - Cây phát sinh loài dựa trên trình tự SSU rRNA là cây phát sinh loài mức phân tử đầu tiên
 - Xác định cây phát sinh loài của họ gen hay họ protein



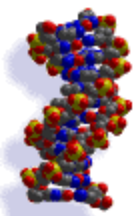
Cây gene (protein) và cây loài

- ❑ Cây loài: thể hiện con đường tiến hóa của một nhóm sinh vật
- ❑ Cây gene (protein): được xây dựng dựa trên trình tự một gene (protein) của mỗi loài
- ❑ Cây gene khác cây loài:
 - Sự phân kỳ của 2 gene có thể xảy ra trước sự phân kỳ của 2 loài
 - Mẫu (đặc trưng) nhánh của cây gene có thể khác với mẫu nhánh của cây loài



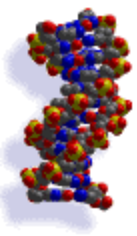
Cây tiến hóa cho m trình tự

- ❑ M node ngoại, mỗi node đại diện cho một trình tự
- ❑ Bao gồm 1 tập các node nội nối các node ngoại và các node nội khác
- ❑ Cây có gốc (node nội có 2 nốt con), cây không gốc (node nội có 3 nhánh)
- ❑ Cạnh: số thay đổi xảy ra giữa các node cạnh



Cây gene (protein)

- Khai khác dữ liệu trình tự bộ gen
 - Cây tiến hóa bộ gen
 - Cây tiến hóa sinh vật
 - Nghiên cứu sự tiến hóa của gen hay protein
 - So sánh các homolog
 - Các trình tự có cùng nguồn gốc
 - Homolog: Ortholog, Paralog, Xenolog
-



Homolog

□ Ortholog:

- Tạo ra do quá trình hình thành loài
- Tiến hóa phân kỳ từ một tổ tiên chung do sự tiến hóa phân kỳ của các loài
- Có chức năng chung

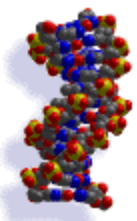
□ Paralog:

- Tạo ra do quá trình nhân đôi gen
- Tiến hóa từ 1 gen chung được nhân đôi trong 1 loài
- Có xu hướng khác chức năng

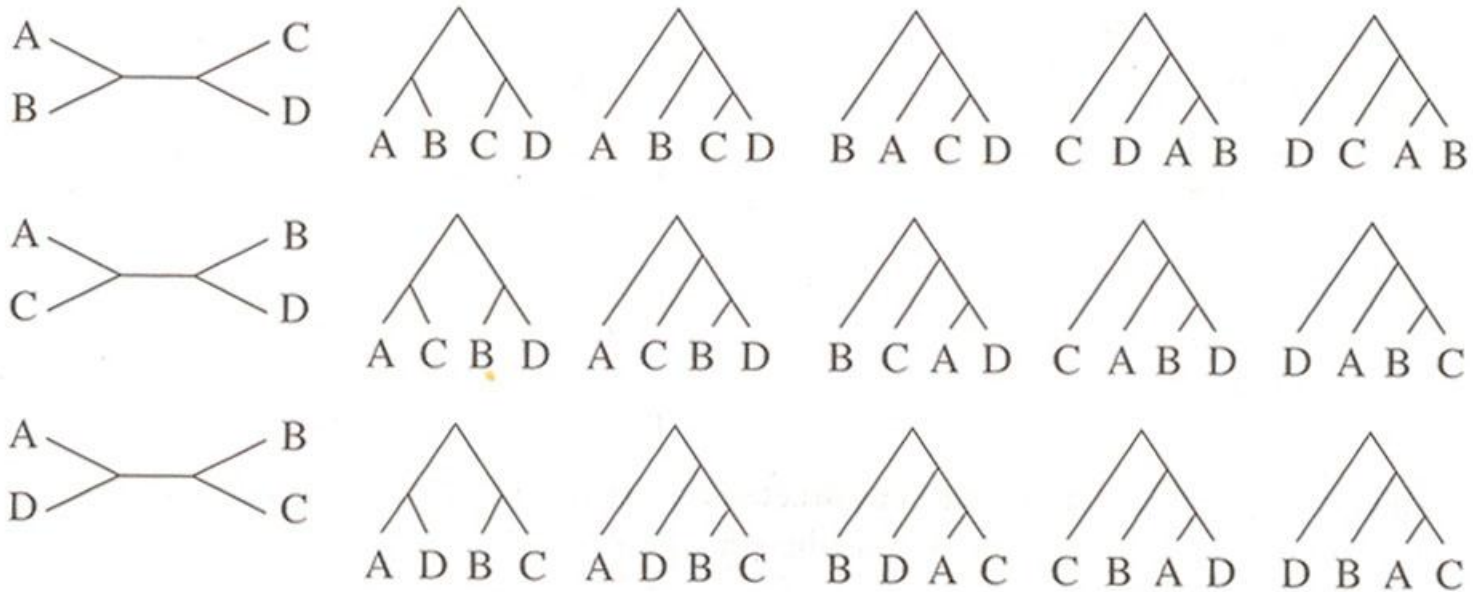
□ Xenolog:

- Tạo ra do quá trình chuyển gen
- Khó phân biệt
- Chức năng có thể thay đổi



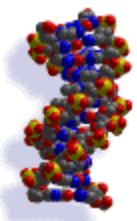


Số lượng hình học topo của cây

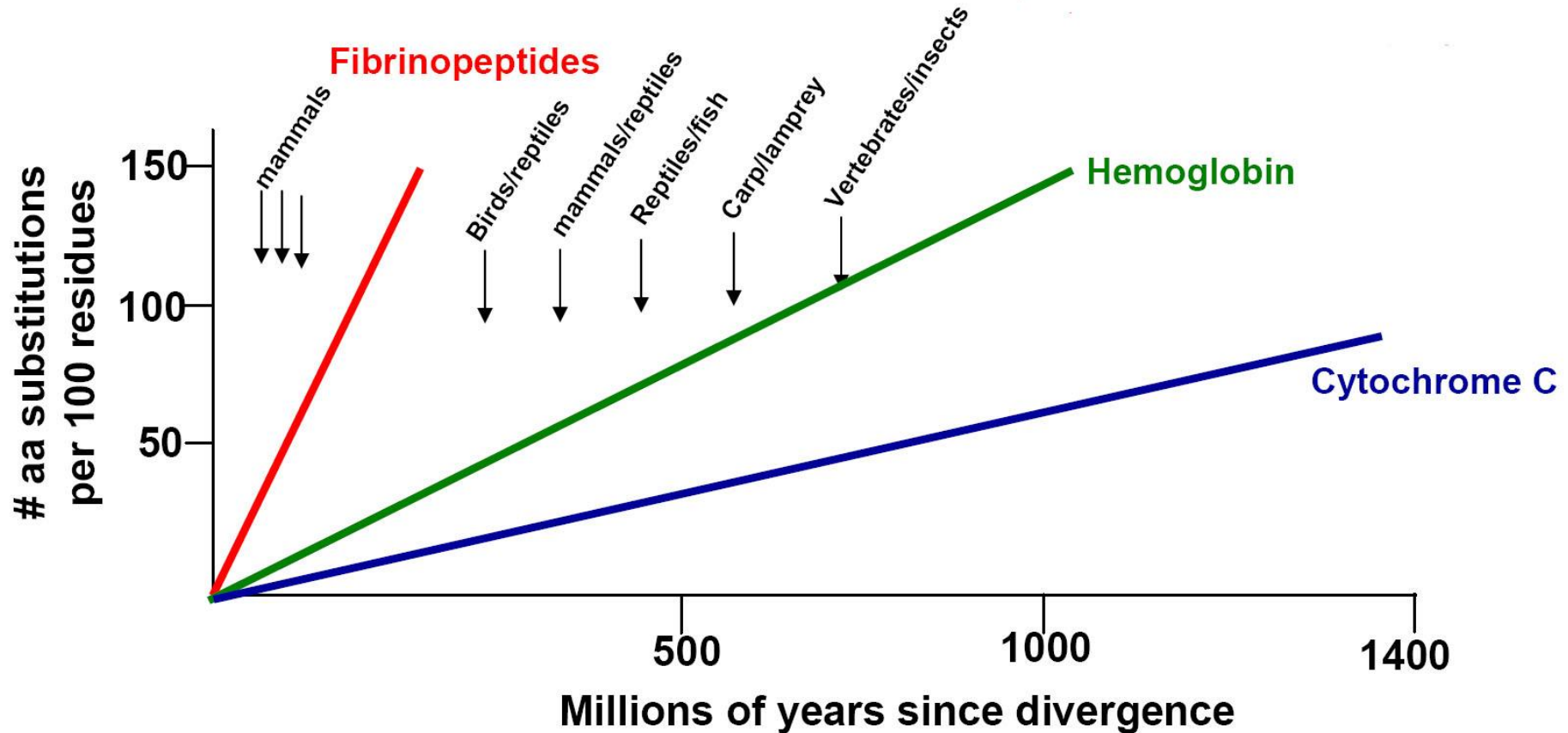


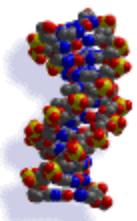
$$T_{unroot}(m) = \frac{(2m-5)!}{2^{m-3}(m-3)!}$$

$$T_{root}(m) = \frac{(2m-3)!}{2^{m-2}(m-2)!}$$



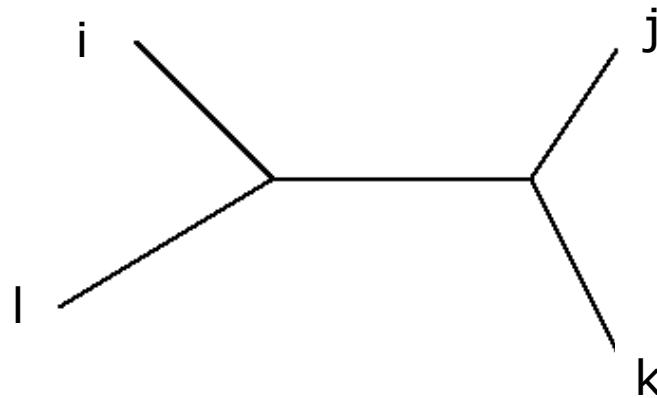
Đồng hồ phân tử

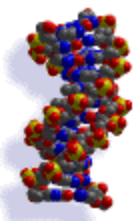




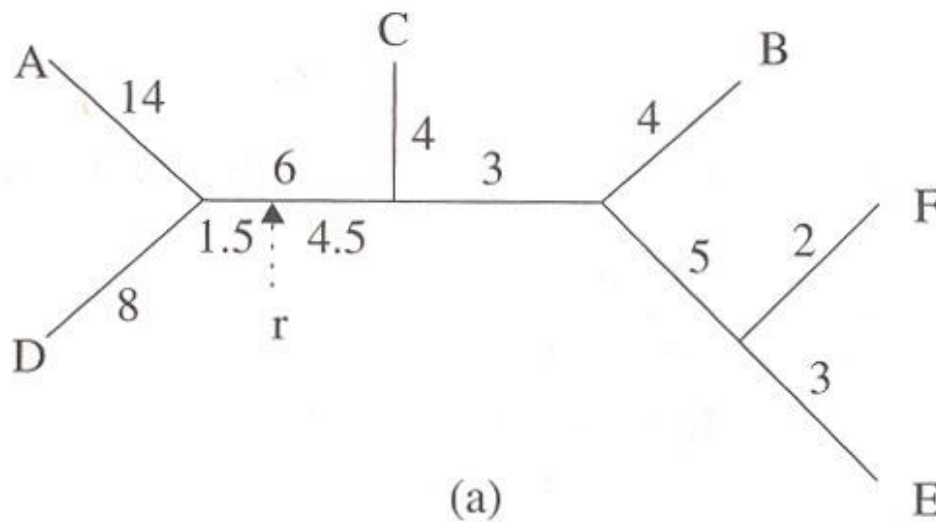
Cây additive

$$D_{i,j} + D_{k,l} = D_{i,k} + D_{j,l} \geq D_{i,l} + D_{j,k}$$



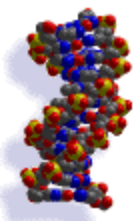


Cây additive



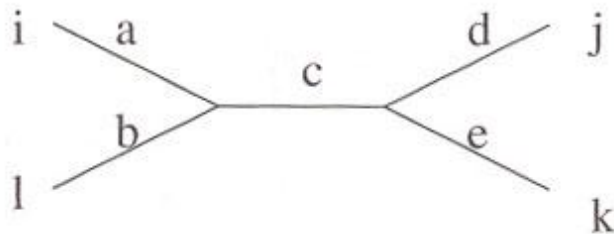
	B	C	D	E	F
A	27	24	22	31	30
B		11	21	12	11
C			18	15	14
D				25	24
E					5

(b)

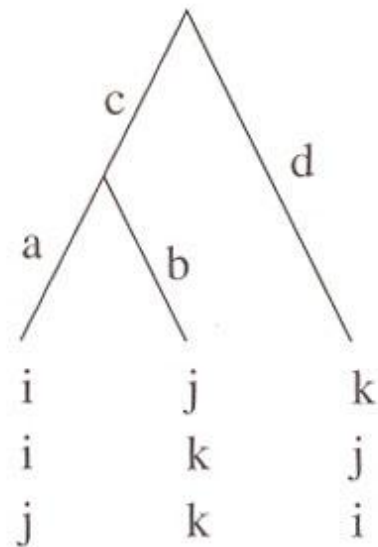


Cây ultrametric

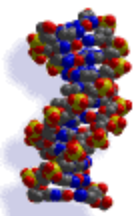
$$D_{i,j} \leq \max(D_{i,k}, D_{k,j})$$



(a)

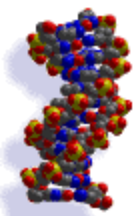


(b)



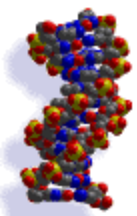
Các phương pháp xây dựng cây gene (protein)

- Maximum Parsimony
- Maximum Likelihood
- Dựa trên ma trận khoảng cách
 - Khoảng cách từng cặp trình tự (UPGMA)
 - Neighbor joining



Maximum Parsimony

- ❑ Nguồn gốc: phát triển tạo cây tiến hóa cho các trình tự protein
- ❑ Input: MSA
- ❑ Xác định hình học cây (Topology) có tổng số thay đổi tiến hóa nhỏ nhất (mất, thêm, thay thế)
- ❑ Nếu 2 hay nhiều cây được tìm thấy, các cây được gọi là các cây parsimony cân bằng



Các cột thông tin

□ S1: A **A** G A G T T C A

□ S2: A **G** C C G T T C T

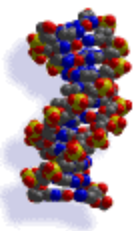
□ S3: A **G** A T A T C C A

□ S4: A **G** A G A T C C T

Cột không
thay đổi

Cột thông tin?

Cột thay đổi?



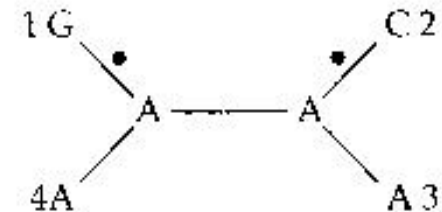
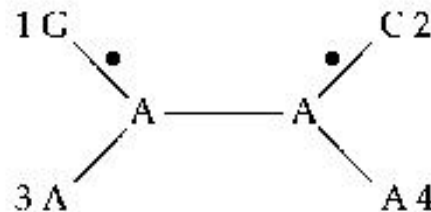
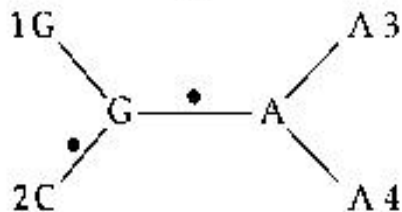
S1: A **A** G A G T T C A
 S2: A **G** C C G T T C T
 S3: A **G** A T A T C C A
 S4: A **G** A G A T C C T

Tree 1
 ((1,2),(3,4))

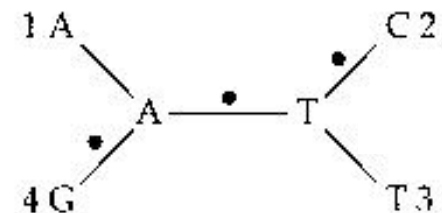
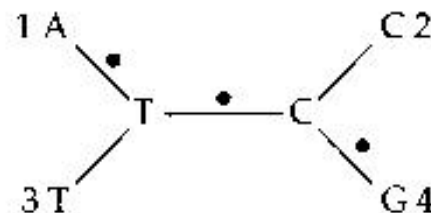
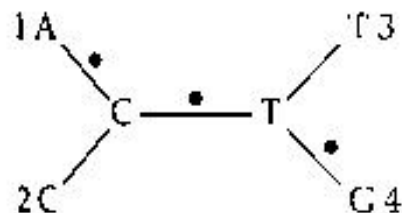
Tree 2
 ((1,3),(2,4))

Tree 3
 ((1,4),(2,3))

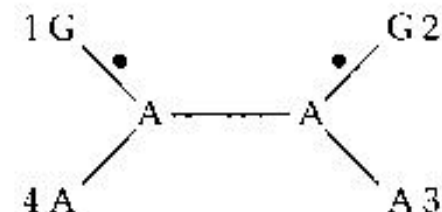
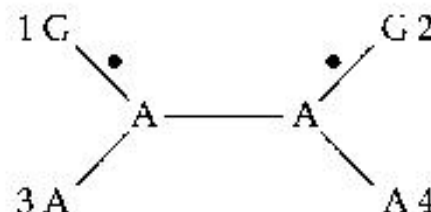
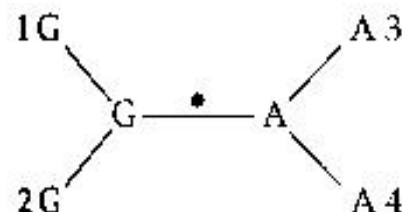
Vị
 trí 3

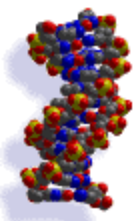


Vị
 trí 4



Vị
 trí 5



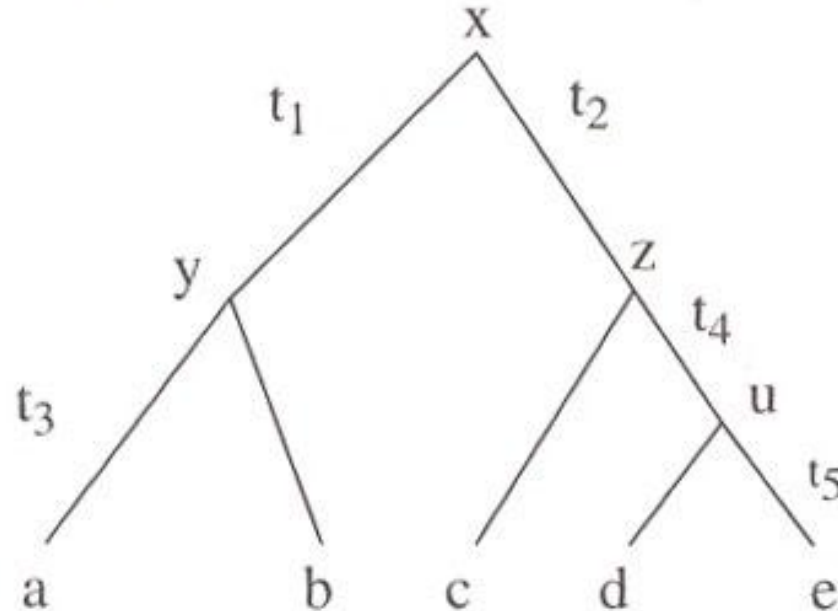


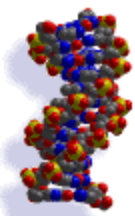
Maximum likelihood

Giả sử trình tự nucleotide là a ở thời điểm 0, $P_{ab}(t)$ là xác suất trình tự nucleotide là b ở thời điểm t .

Xác suất của các trình tự a, b, c, d, e ở node ngoại của cây là:

$P_{xy}(t_1)P_{ya}(t_3)P_{yb}(t_3)P_{xz}(t_2)P_{zc}(t_4+t_5)P_{zu}(t_4)P_{ud}(t_5)P_{ue}(t_5)$





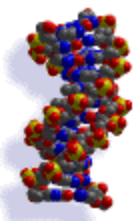
Xây dựng cây dựa trên khoảng cách cặp trình tự

☐ UPGMA

- Cây ultrametric
- Cây có gốc

☐ Neighbor joining

- Cây additive
- Cây không gốc



Khoảng cách 2 trình tự

□ SEQ1 ACGTACGTAA

□ SEQ2 ACGTTCGTAT

□ SEQ3 TCCATCGTAAA

	S1	S2	S3
S1	0		
S2	0.2	0	
S3	0.4	0.4	0

Độ giống nhau

Khoảng cách

(1-2) 80%

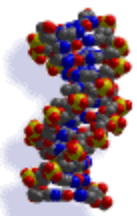
$1 - 0.8 = 0.2$

(1-3) 60%

$1 - 0.6 = 0.4$

(2-3) 60%

$1 - 0.6 = 0.4$

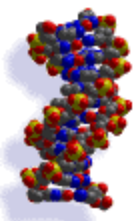


UPGMA

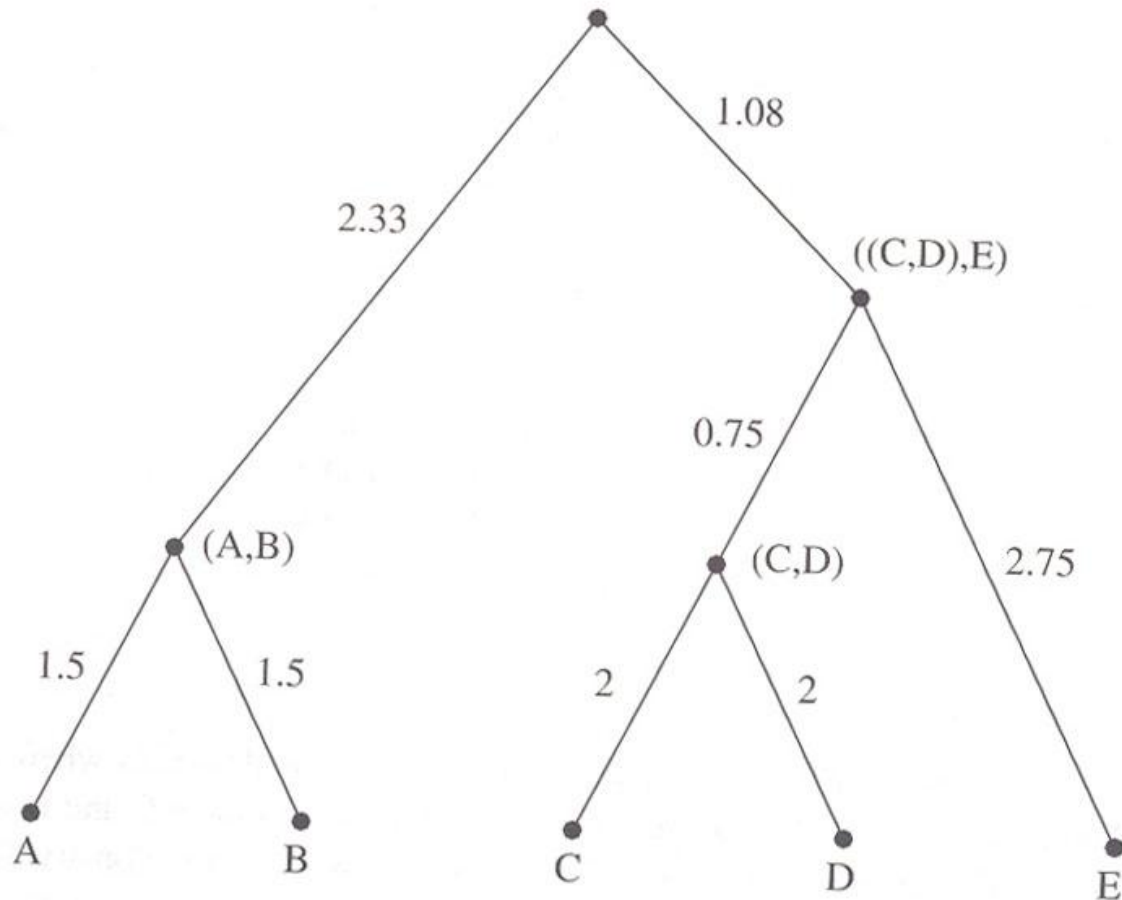
- Unweighed **Pair**-Group with Arithmetic Mean

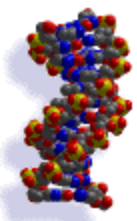
Species	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

- Bước 1: Nhóm 2 trình tự gần nhau nhất vào 1 nhóm
- Bước 2: Tạo ma trận khoảng cách mới giữa (AB) và C và D
- Bước 3: Sử dụng ma trận mới nhóm 2 trình tự gần nhau nhất vào trong một nhóm. Lập lại cho đến khi tất cả các trình tự đều được phân nhóm

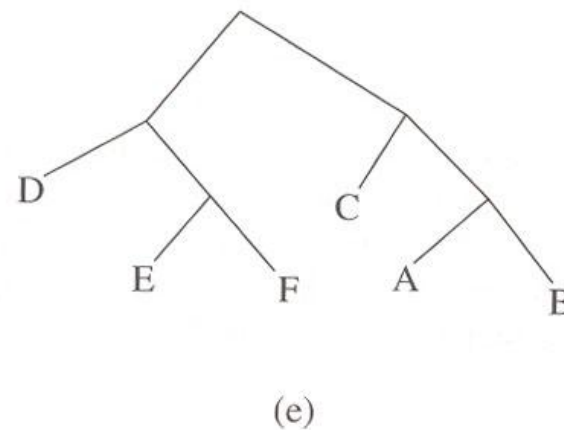
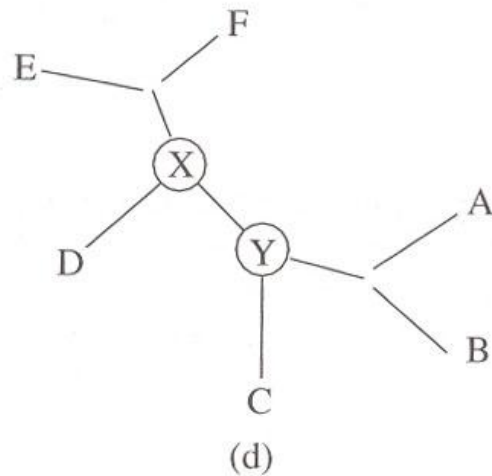
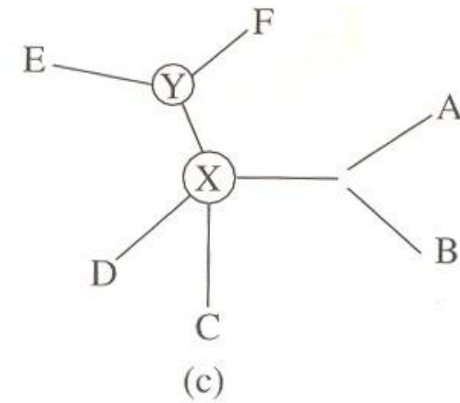
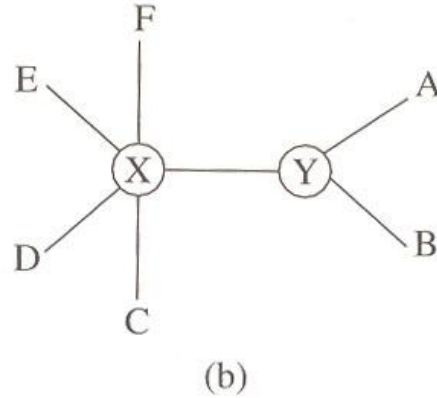
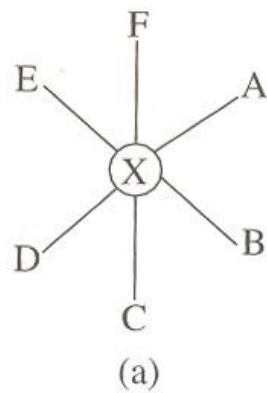


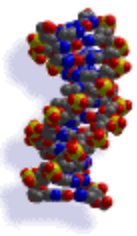
UPGMA





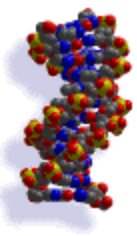
Neighbor joining





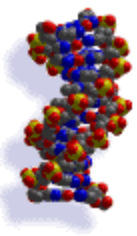
Bootstrap

- ❑ Kỹ thuật ước định độ chính xác của cây
- ❑ Phát minh năm 1979, áp dụng năm 1985



Các chương trình

- ☐ Phylip
- ☐ PAUL
- ☐ Clustal
- ☐ Puzzle
- ☐ Molphy



Bài tập

1. Vẽ tất cả các cây không gốc (binary) cho 5 trình tự?

2. Tìm vị trí thông tin cho sắp giống cột sau:

ATGTA

TAGTA

CGCTG

GCCTG

Tìm cây bằng phương pháp maximum parsimony cho sắp giống cột này?