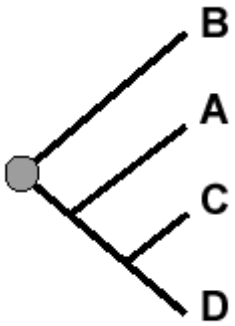
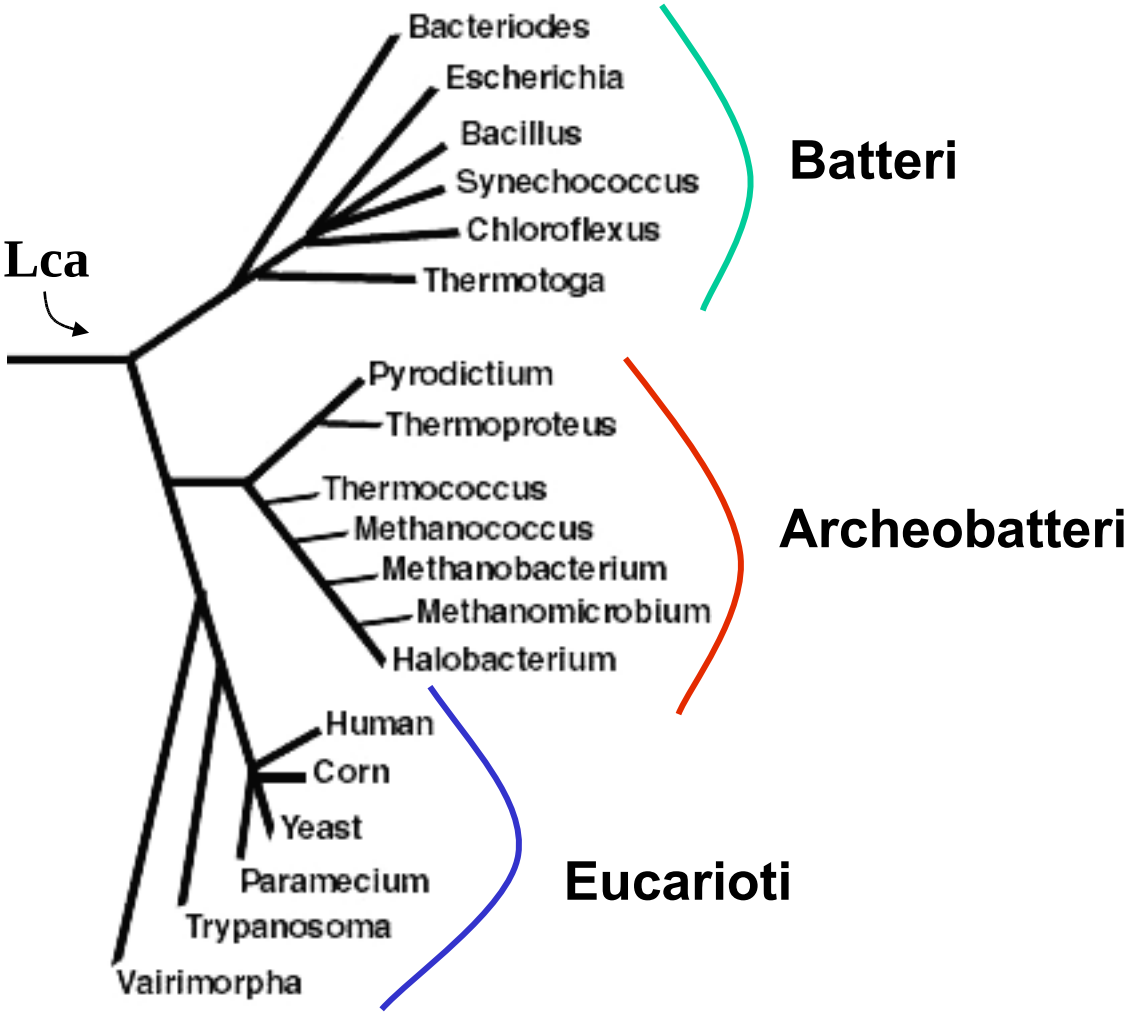
Alberi filogenetici

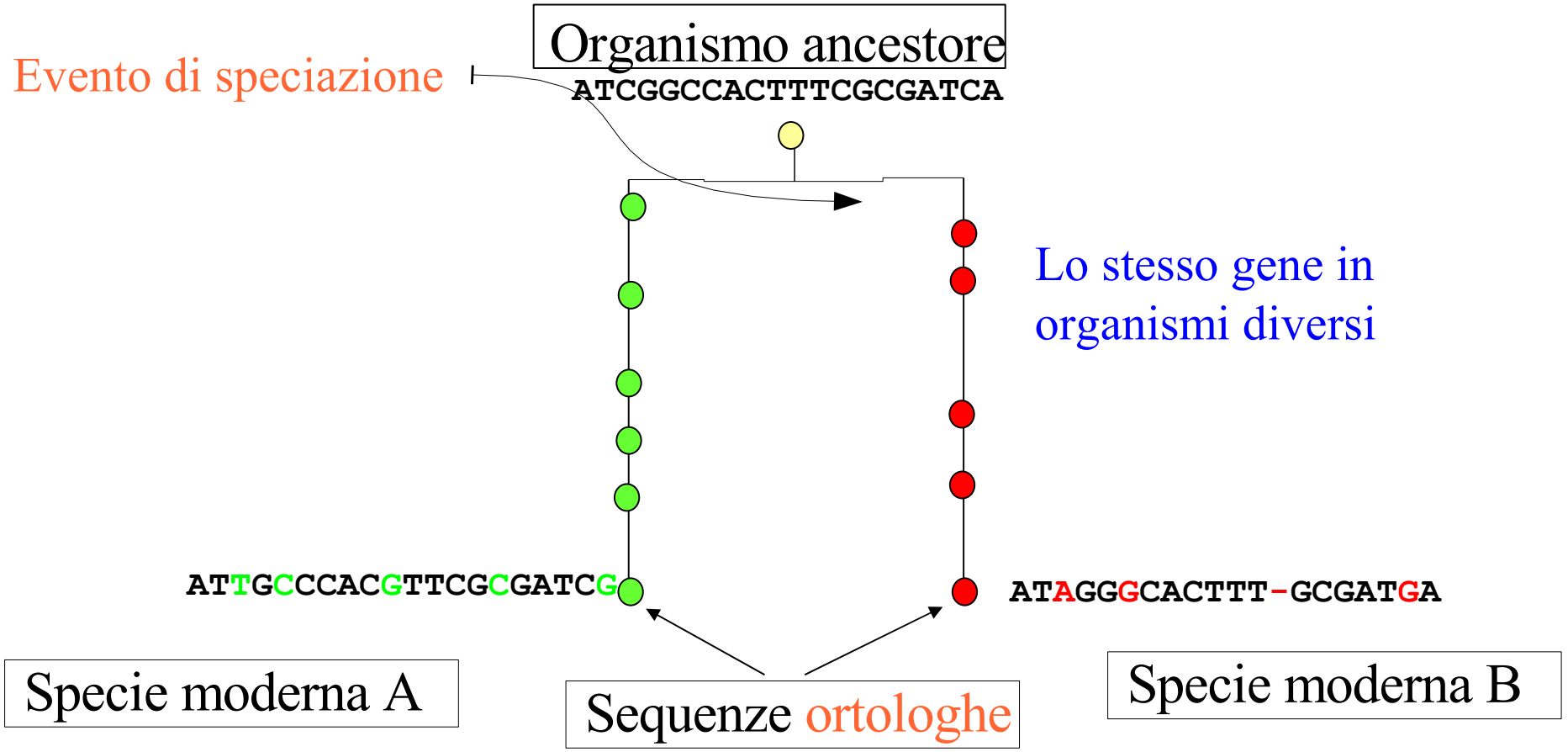


The tree of life

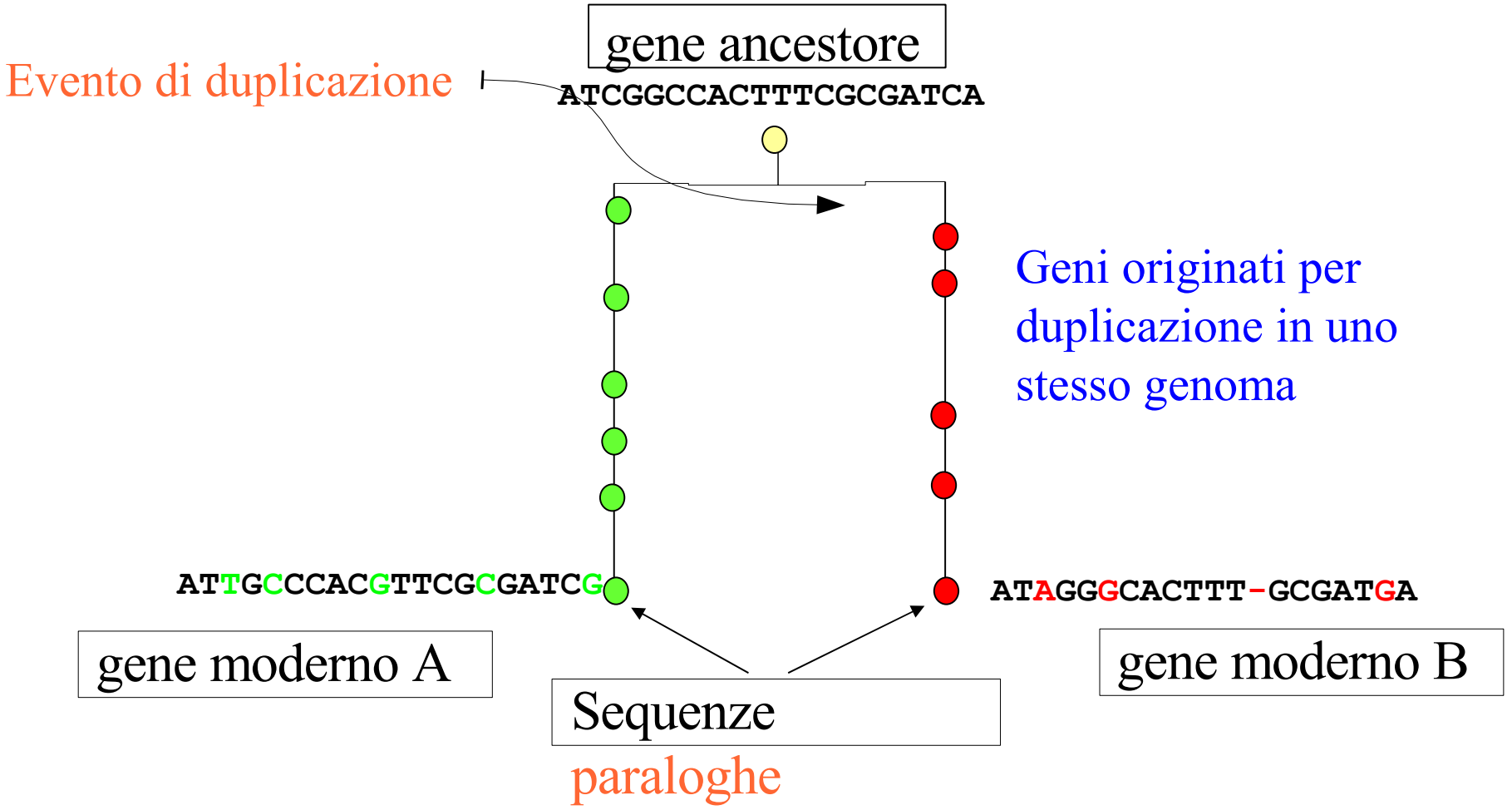
Albero filogenetico costruito con le sequenze della subunità piccola dell'RNA ribosomale. Tutte le forme viventi condividono un comune ancestore (LCA, "last common ancestor") e sono raggruppabili in tre gruppi principali. Batteri, Archeobatteri e Eucarioti. L'ordine di diramazione tra i tre gruppi (vale a dire, ad esempio, se gli Archeobatteri siano più vicini agli Eucarioti o ai Batteri) non è tuttora chiaro.



Separazione per speciazione



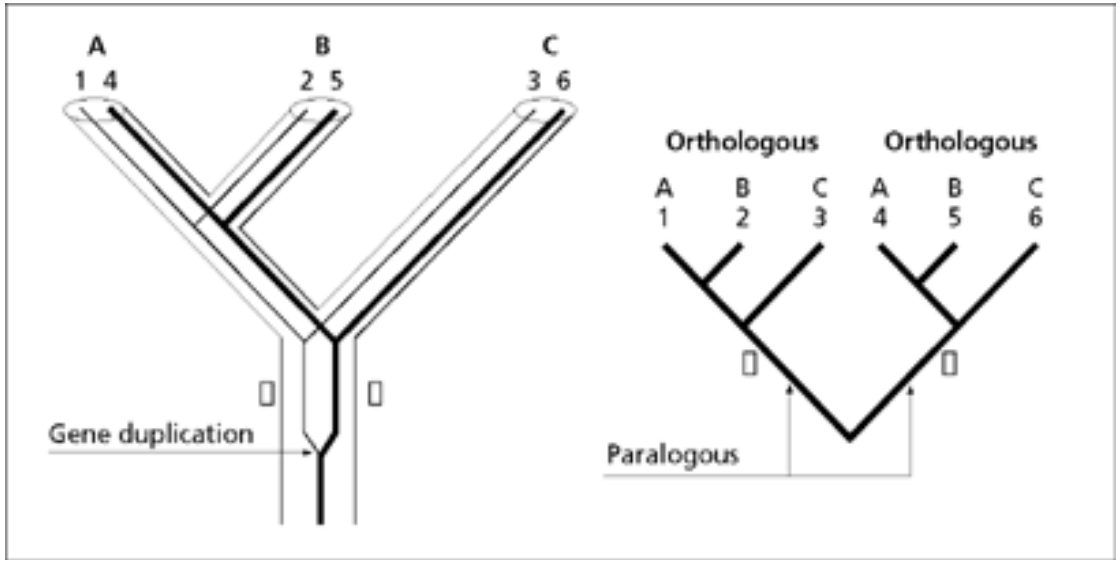
Separazione per duplicazione genica



Rapporti evolutivi tra geni

Ortologia: I geni si separano per speciazione. La filogenesi dei geni riflette la storia degli organismi

Paralogia: I geni si separano per duplicazione all'interno di uno stesso organismo. La filogenesi riflette la storia dei geni



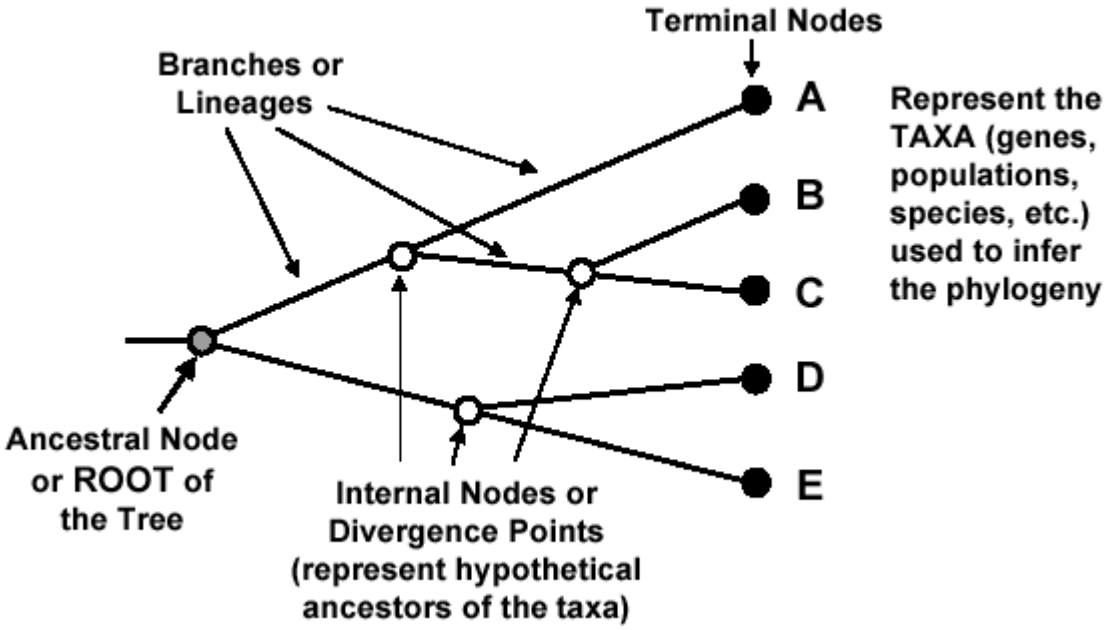
Alberi evolutivi

Albero evolutivo	
:==	Un grafo bidimensionale che mostra le relazioni evolutive esistenti tra organismi o tra geni/sequenze

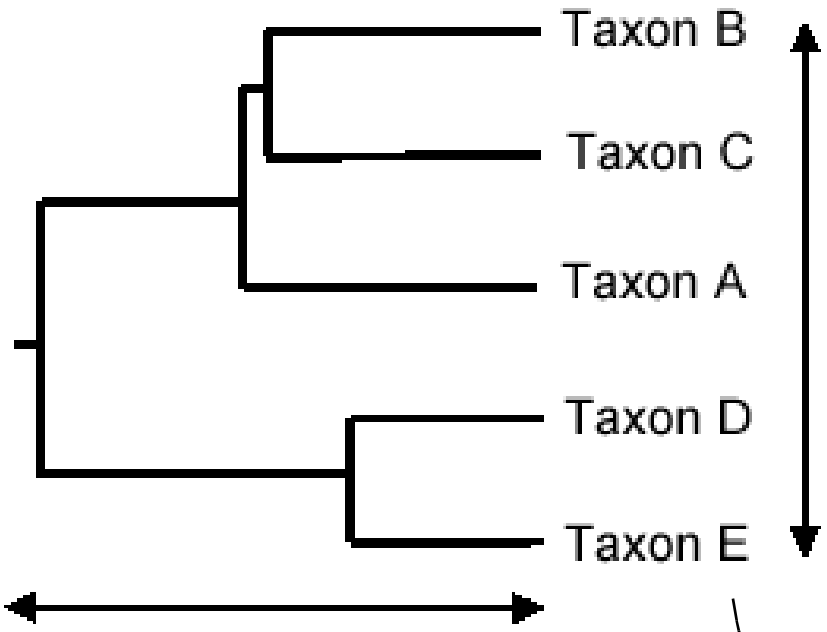
Terminologia degli alberi evolutivi

Un albero si compone di:

- **nodi terminali** o foglie o taxa che rappresentano oggetti esistenti
- **nodi interni** o punti di divergenza o biforcazione che rappresentano ipotetici ancestori dei taxa
- un nodo iniziale (solo nel caso degli alberi “rooted”) , o **radice** che rappresenta l’ancestore di tutti i taxa
- **bracci** o linee che congiungono i vari nodi



Rappresentazione delle relazioni evolutive



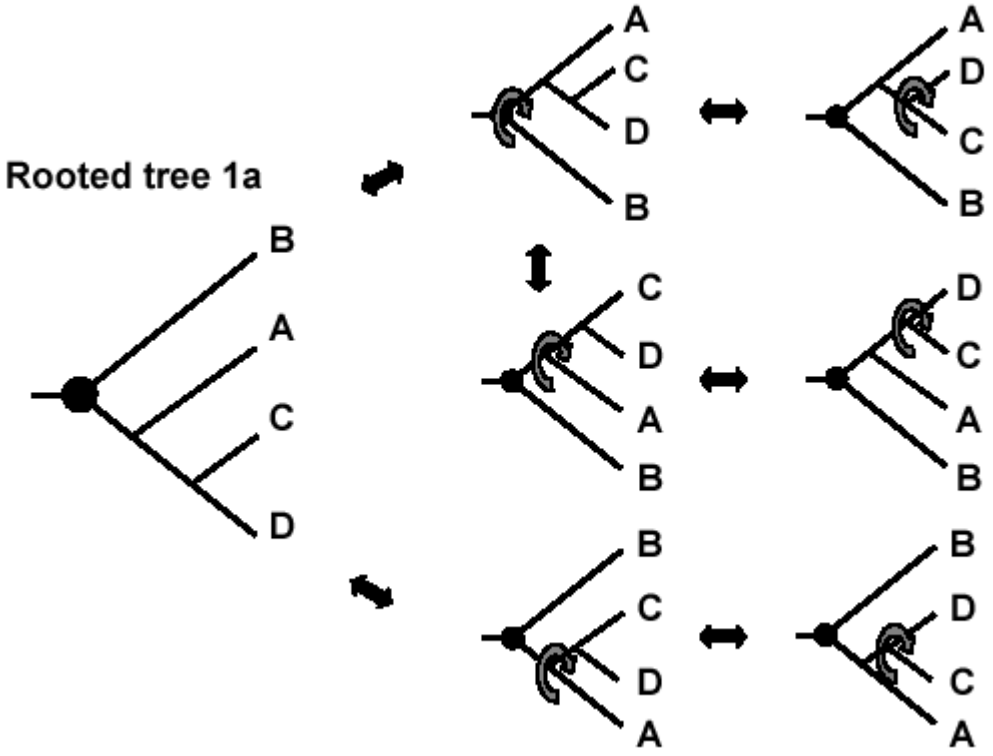
Quest'asse non ha nessun significato

Quest'asse può non avere nessuna scala (cladogramma), oppure essere proporzionale alla distanza genetica (filogrammi, o alberi additivi) o essere proporzionale al tempo (alberi ultrametrici)

((A,(B,C)),(D,E)) = le stesse relazioni filogenetiche descritte come parentesi

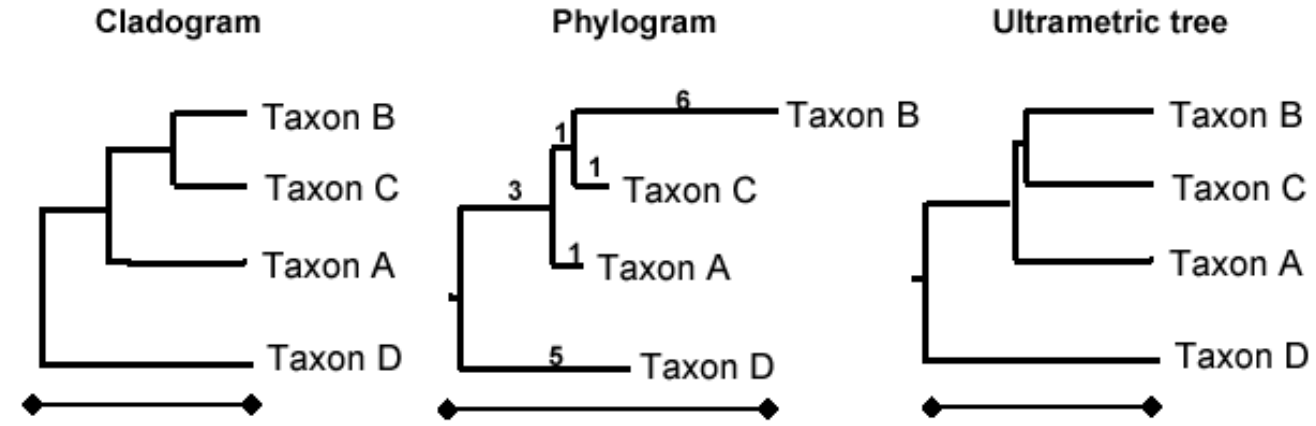
L'albero e le parentesi rappresentano le stesse relazioni evolutive. Ad esempio che **B** e **C** sono più vicini tra di loro di quanto non lo sia **A** a ciascuno dei due, e che **A,B,C** formano un "**clade**" che è il "**sister group**" del clade composto da D e E. In un albero con una scala temporale D e E sono anche i più vicini in assoluto.

La rotazione di un nodo non modifica l'albero



Tutte le operazioni di rotazione attorno ad un nodo forniscono alberi con topologia equivalente.

Tre tipi di alberi

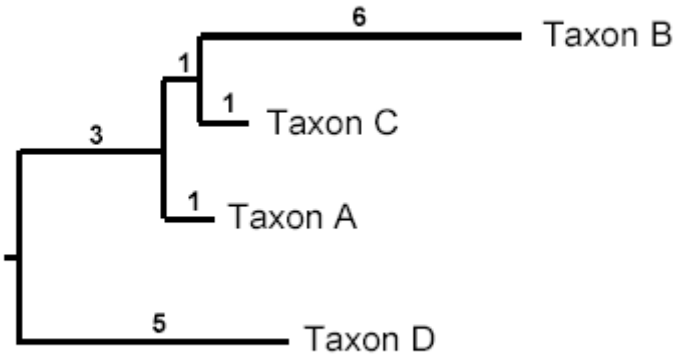


Nessun significato **Distanze genetiche** **Tempo**

Questi alberi hanno la **stessa topologia**, ovvero rappresentano le stesse relazioni evolutive tra i taxa. Il significato della lunghezza dei bracci è diverso nei tre casi

Somiglianza ⊨ relazione evolutiva

I taxa B e C sono evolutivamente più vicini tra loro (vale a dire hanno un ancestore comune più recente) rispetto al taxon A benché i taxa C e A siano più simili in sequenza (la distanza tra A e C è uguale a 3 [1+1+1], mentre la distanza tra B e C è uguale a 7 [6+1])



Somiglianza di sequenza:= proprietà additiva del confronto (un osservazione)

Relazione evolutiva:= connessione genetica nel tempo (un fatto storico)

Metodi di ricostruzione filogenetica

:= metodi matematici o statistici per inferire l'ordine di divergenza tra i taxa e la lunghezza dei bracci che connettono i vari nodi.

Due tipi principali di metodi di ricostruzione filogenetica

Sistemi basati sulle distanze

Calcolo delle distanze + Metodo di clustering

- **Neighbor-joining**
- **UPGMA**

Sistemi basati sui caratteri

Si basano su un criterio di ottimizzazione

- **Parsimonia**
- **Maximum Likelihood**

Dati usati per le ricostruzioni filogenetiche

Sequenze allineate:

Taxa	Characters
Species A	ATGGCTATTCTTATAGTACG
Species B	ATCGCTAGTCTTATATTACA
Species C	TTCACTAGACCTGTGGTCCA
Species D	TTGACCAGACCTGTGGTCCG
Species E	TTGACCAGTTCTCTAGTTCG

Distanze genetiche:

I dati di sequenza vengono trasformati in matrici di distanze utilizzando un modello evolutivo.

	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

Matrice 'non corretta' di distanza (differenze osservate)

Correzione (stima del numero di mutazioni per sito)

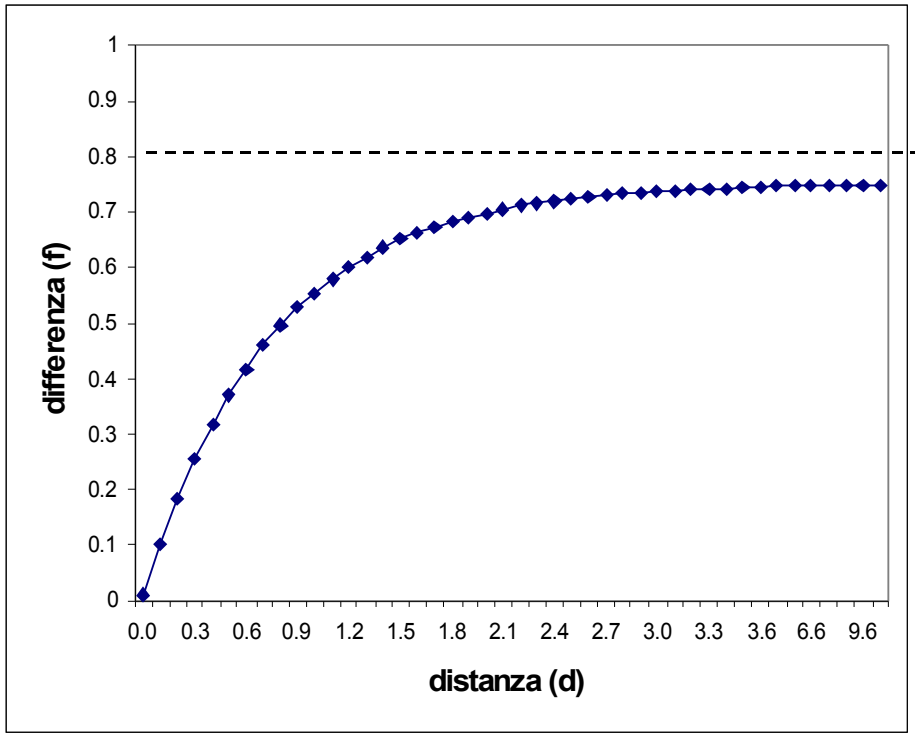
Stima della distanze tra le sequenze

Unità di misura della distanza genetica: Mutazioni / siti

Le distanze genetiche sono calcolate in base alle differenze osservate

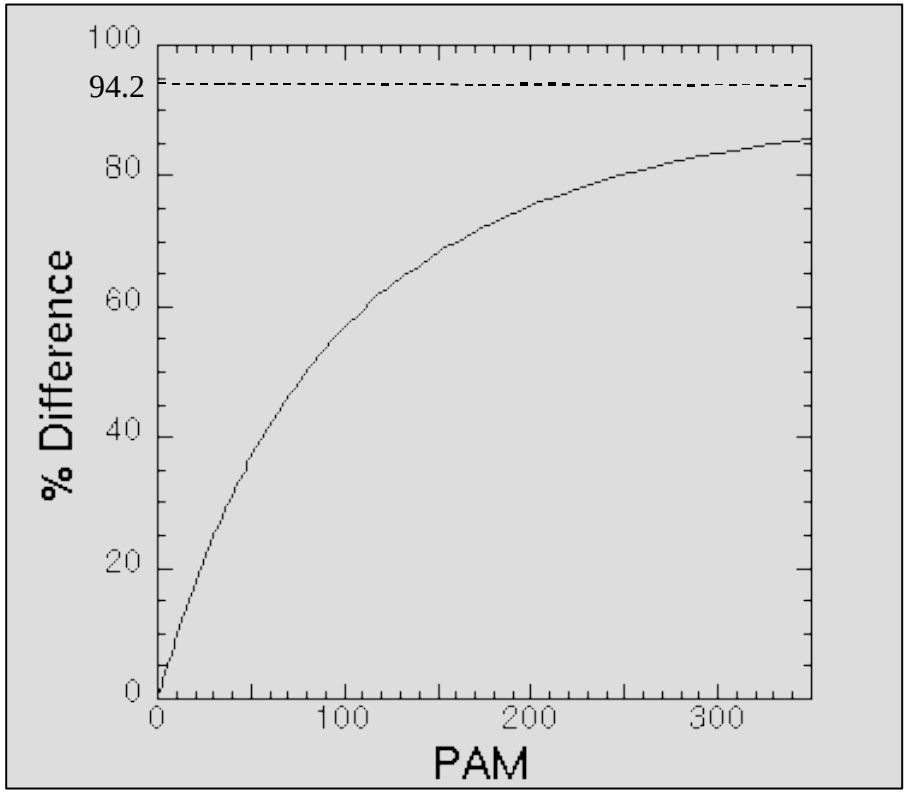
Il calcolo deve tener conto che non tutte le mutazioni sono osservabili

Calcolo delle distanze per sequenze nucleotidiche



$$d = -3/4 \ln(1 - 4/3\, f)$$ Formula di Jukes-Cantor

Calcolo delle distanze per sequenze proteiche



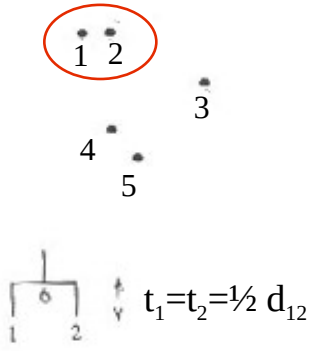
%Difference	PAM
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

UPGMA

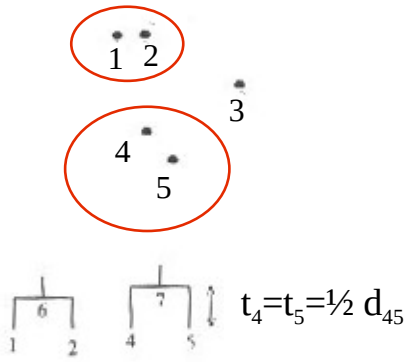
Sokal & Michner 1958

UPGMA è un sistema di clustering basato su “Unweighted Pair Group Method using aritmetic Average”. Raggruppa successivamente le sequenze a partire dalle più simili ed aggiungendo via via un nodo all’albero. Le distanze tra due taxa, tra un nodo e un taxon, o tra due nodi (ovvero le lunghezze dei bracci) sono dati dalla media aritmetica delle distanze. L’albero può essere immaginato essere costruito dal basso verso l’alto con ciascun nodo aggiunto sopra i successivi. L’ultimo nodo aggiunto è la radice.

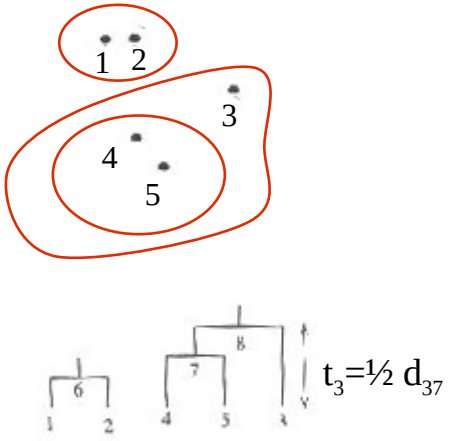
UPGMA produce alberi **rooted** ed **ultrametrici**. Può dare alberi con corretta topologia solo se le sequenze rispettano l’orologio molecolare.



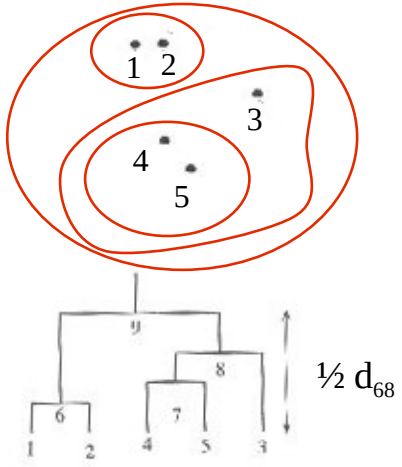
(i)



(ii)



(iii)



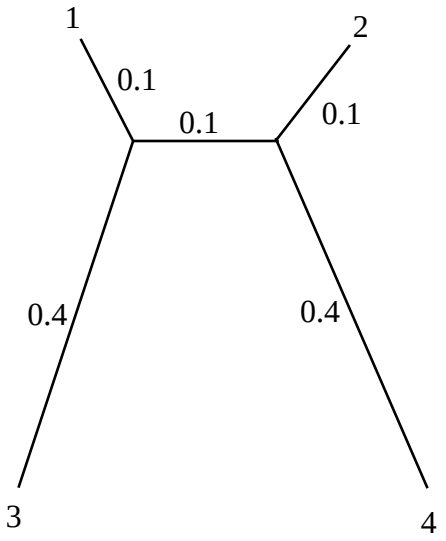
(iv)

Neighbour-joining

Saitou & Nei, 1987

Il sistema usato da neighbour-joining per trovare i neighbour si basa sulla valutazione della distanza tra due foglie sottraendo la distanza media di ciascuna di queste rispetto a tutte le altre foglie. In altre parole, neighbor-joining **non considera semplicemente la distanza tra le coppie per costruire l'albero, ma valuta la distanza rispetto a tutti gli altri punti**.

Gli alberi costruiti con neighbor-joining sono **additivi** e **unrooted**. Possono ricostruire in modo esatto la topologia di sequenze che non seguono l'orologio molecolare



Due tipi principali di metodi di ricostruzione filogenetica

Sistemi basati sulle distanze

Calcolo delle distanze + Metodo di clustering

- **Neighbor-joining**
- **UPGMA**

→ **Sistemi basati sui caratteri**

Si basano su un criterio di ottimizzazione

- **Parsimonia**
- **Maximum Likelihood**

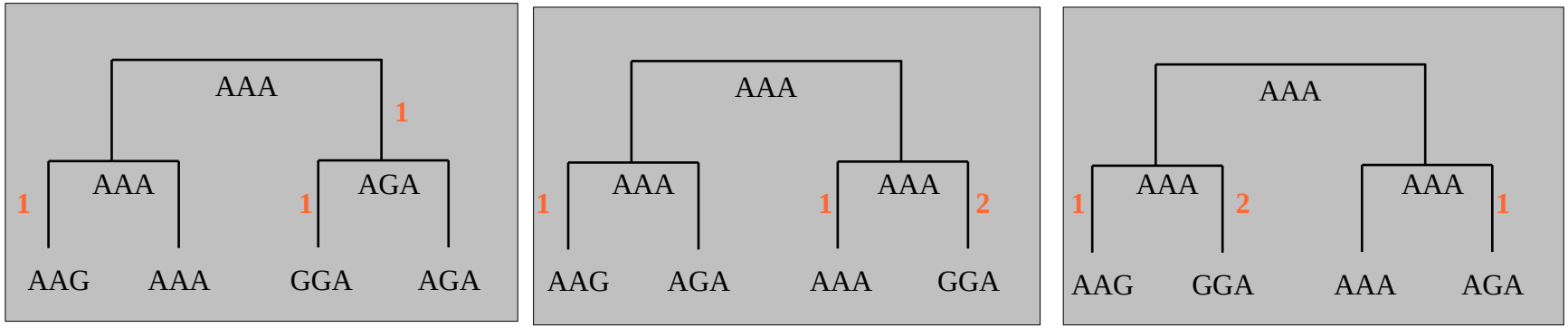
Massima parsimonia

Trova l'albero (unrooted) che spiega le sequenze osservate con il numero minimo di sostituzioni. L'algoritmo a due componenti:

1) Valutazione del **costo di un albero** in termini di mutazioni

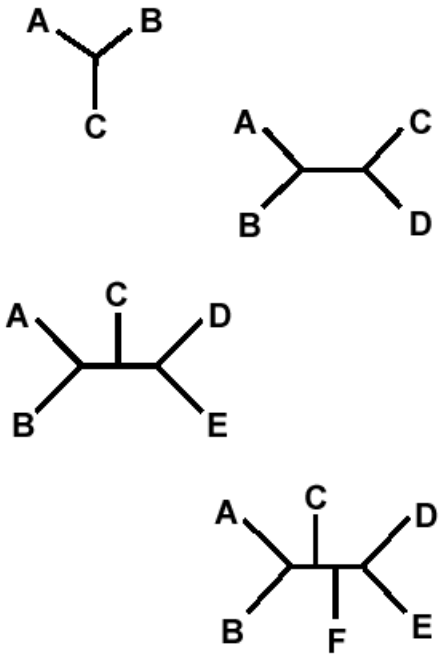
2) **Ricerca tra tutti gli alberi possibili** per trovare l'albero con il costo inferiore

Seq1 AAG
Seq2 AAA
Seq3 GGA
Seq4 AGA



Dei tre alberi rappresentati viene selezionato quello a sinistra perché ha un costo inferiore (tre) rispetto agli altri (che necessitano di quattro mutazioni)

Numero possibile di alberi



# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
.	.
.	.
.	.
.	.
30	$\approx 3.58 \times 10^{36}$

(2N - 5)!! = # unrooted trees for N taxa

Il numero degli alberi possibili cresce in modo più che esponenziale con l'aumentare dei taxa vi sono (2n-5)!! [ovvero 3*5*…*(2n-5)] alberi unrooted con n taxa

Come conseguenza Gli algoritmi basati sui caratteri devono ricorrere a metodi euristici per trovare l'albero ottimale nel caso di molte sequenze.

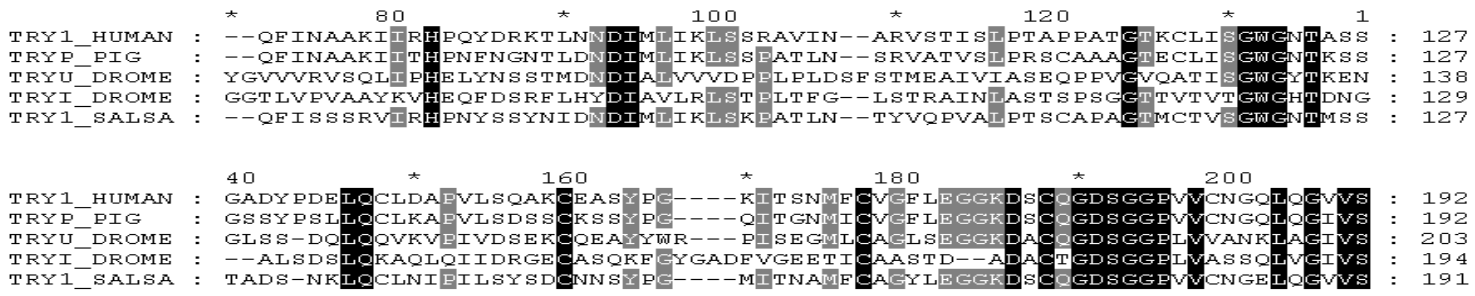
Come si realizza un albero filogenetico 1

1) Scelta delle sequenze da confrontare

NB Tutte le sequenze del confronto devono essere omologhe!

Come si realizza un albero filogenetico 2

2) Allineamento multiplo.



Come si realizza un albero filogenetico 3

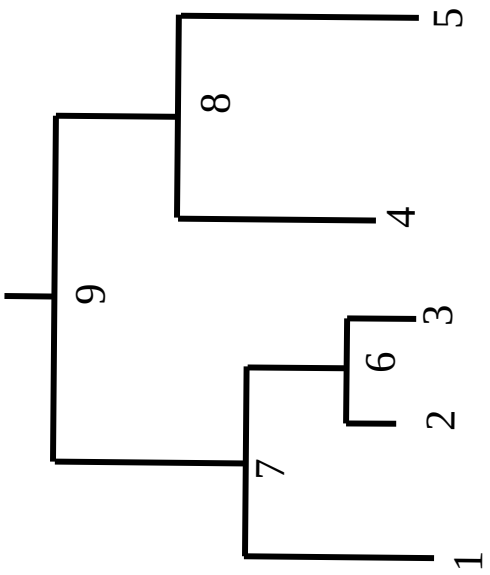
3) Applicazione di un metodo di ricostruzione filogenetica

Metodi basati sui caratteri:
Usano direttamente le sequenze allineate (acidi nucleici o amino acidi) per la costruzione dell'albero

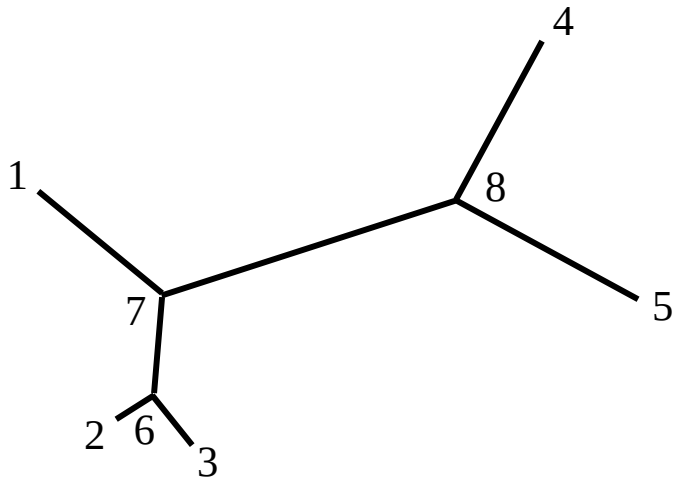
Metodi basati sulle distanze:
I dati di sequenza vengono trasformati in distanze utilizzando un modello evolutivo.

Come si realizza un albero filogenetico 4

4) Rappresentazione dell'albero



Con radice



In forma radiale
(senza radice)

In pratica...

- Scelta delle sequenze; allineamento con ClustalX

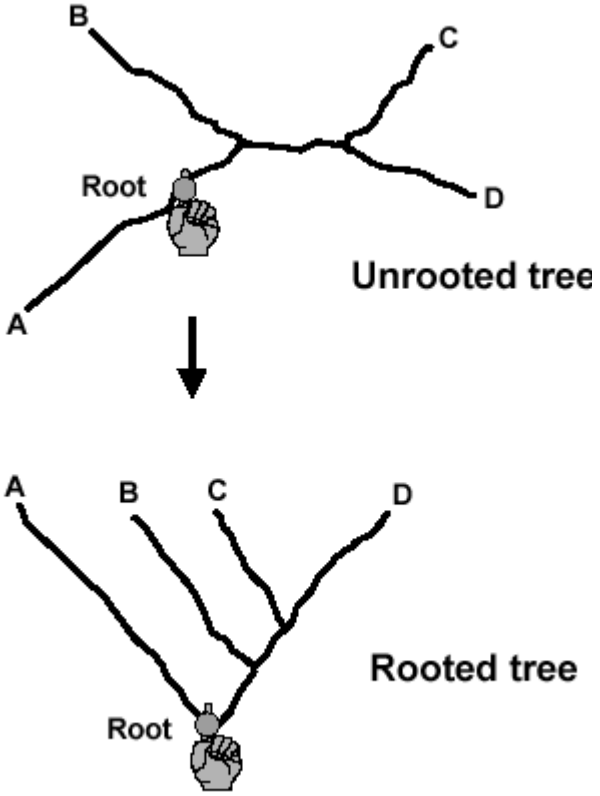
- Ricostruzione filogenetica di NJ con ClustaX ("Draw tree")

- Visualizzazione delle sequenze con Treeview

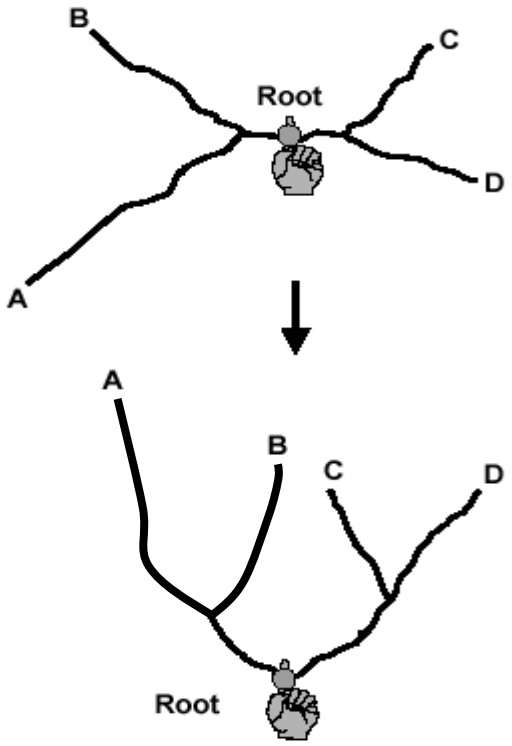
- Posizionamento della radice (o visualizzazione radiale)

La radice determina l'ordine di diramazione

Per inserire la radice in un albero si può immaginare di avere un albero composto di lacci, di afferrare la radice e di tirarla fino a portare tutte le foglie all'estremità opposta alla radice.



Con questa radice A non è più vicino a B di quanto non lo siano C e D

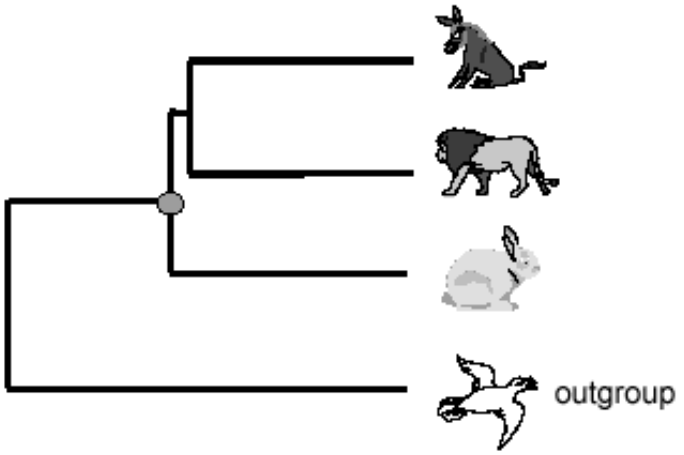


Con questa radice A è più vicino a B di quanto non lo siano C e D

Due modi di posizionare la radice

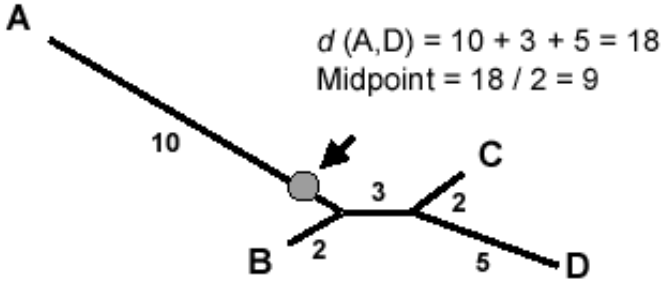
Outgroup

Si usa un taxa (“Outgroup”) che precede nell’ordine di diramazione il gruppo di interesse (“ingroup”). Richiede una conoscenza preliminare delle relazioni tra i vari taxa



Midpoint distance

Pone la radice dell’albero a metà tra i due taxa più distanti dell’albero, come si deduce dalla lunghezza dei bracci. Richiede che non ci siano deviazioni importanti dall’orologio molecolare



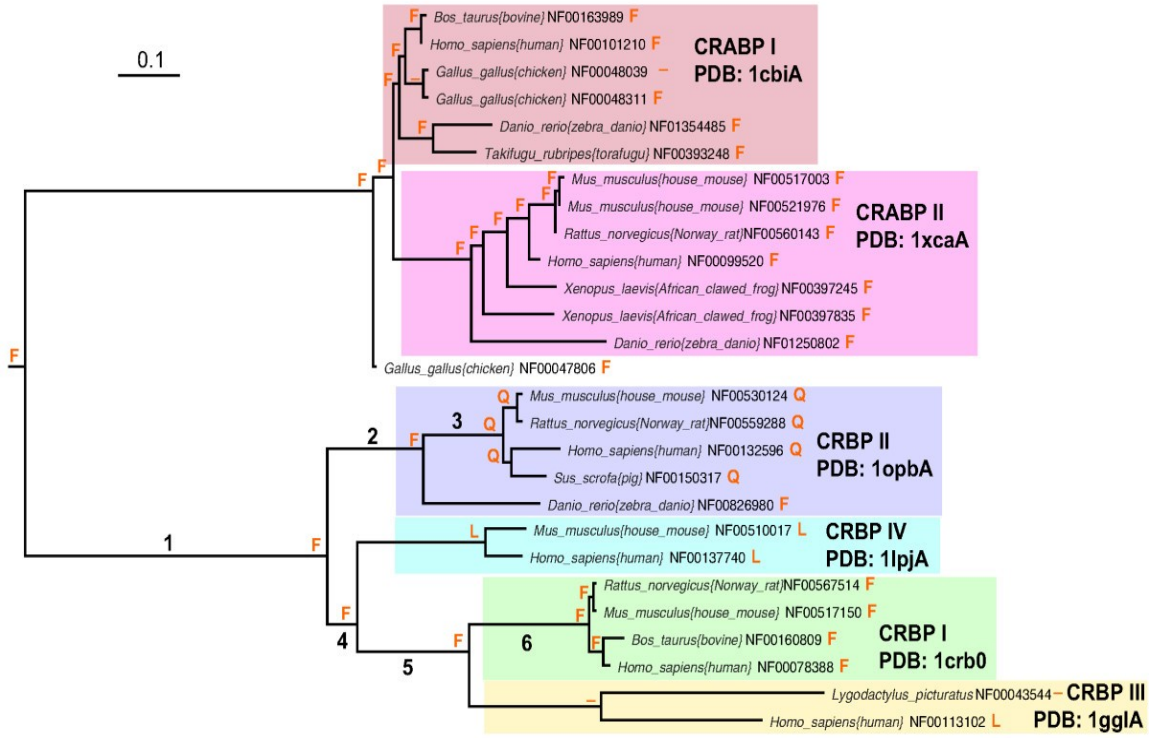
Scopi dell'inferenza filogenetica

Ricostruzione delle relazioni esistenti tra geni o organismi. In un albero ciò si deduce dell'ordine di diramazione dei taxa

Esempi domande a cui si può rispondere grazie alle inferenze filogenetiche

- *Quali sono i rapporti di parentela tra l'uomo e gli altri primati?*
- *Quali sono i rapporti di parentela tra proteine omologhe?*
- *Ho veramente sequenziato il DNA di un dinosauro?*

Classificazione evolutiva delle proteine



Trasportatori di acido retinoico

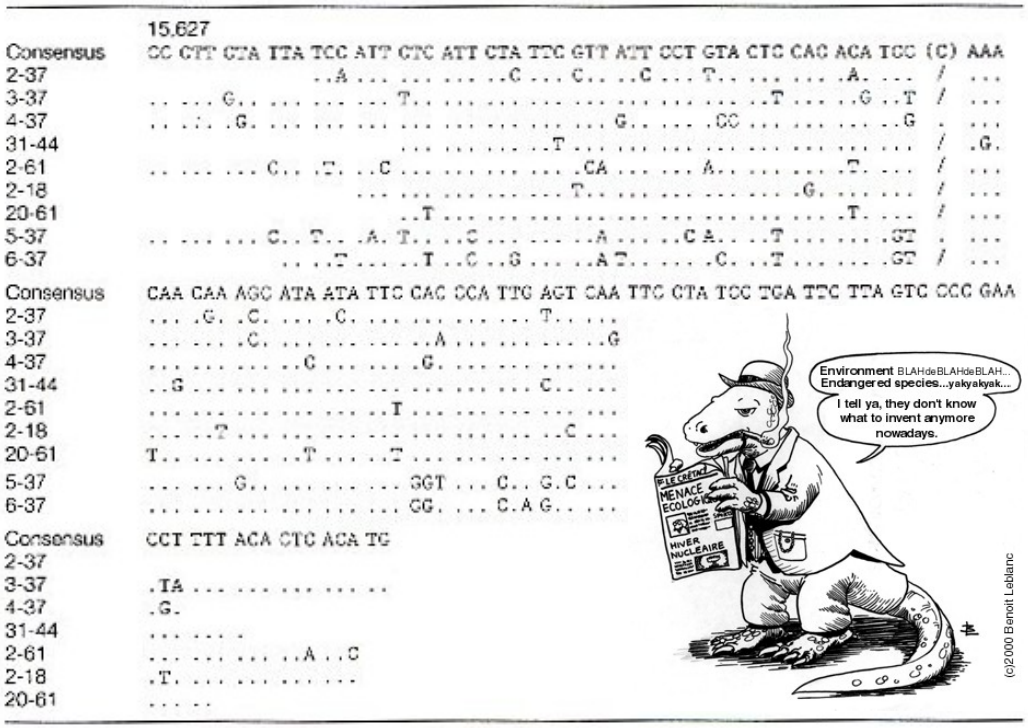
Trasportatori di retinolo

DNA da un dinosauro?

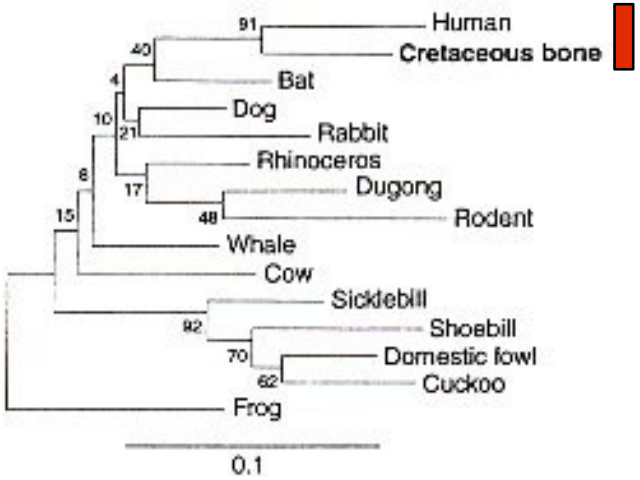
DNA Sequence from Cretaceous Period Bone Fragments

Scott R. Woodward,* Nathan J. Weyand, Mark Bunnell

DNA was extracted from 80-million-year-old bone fragments found in strata of the Upper Cretaceous Blackhawk Formation in the roof of an underground coal mine in eastern Utah. This DNA was used as the template in a polymerase chain reaction that amplified and sequenced a portion of the gene encoding mitochondrial cytochrome b. These sequences differ from all other cytochrome b sequences investigated, including those in the GenBank and European Molecular Biology Laboratory databases. DNA isolated from these bone fragments and the resulting gene sequences demonstrate that small fragments of DNA may survive in bone for millions of years.



Nel 1994 un articolo su SCIENCE affermò l'amplificazione con successo tramite PCR di un frammento del citocromo B mitocondriale da un osso di dinosauro. Le sequenze ottenute vennero giudicate valide poiché "erano diverse da tutte le altre in banca dati". Una serie di articoli sulla stessa rivista dimostrò con analisi filogenetica che il DNA ottenuto era in realtà una contaminazione di una rara variante umana.



Replica su SCIENCE, 26 maggio 1995