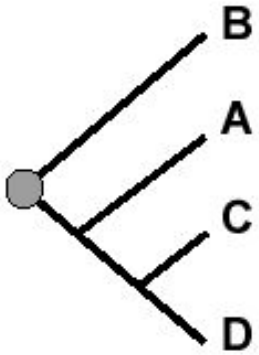
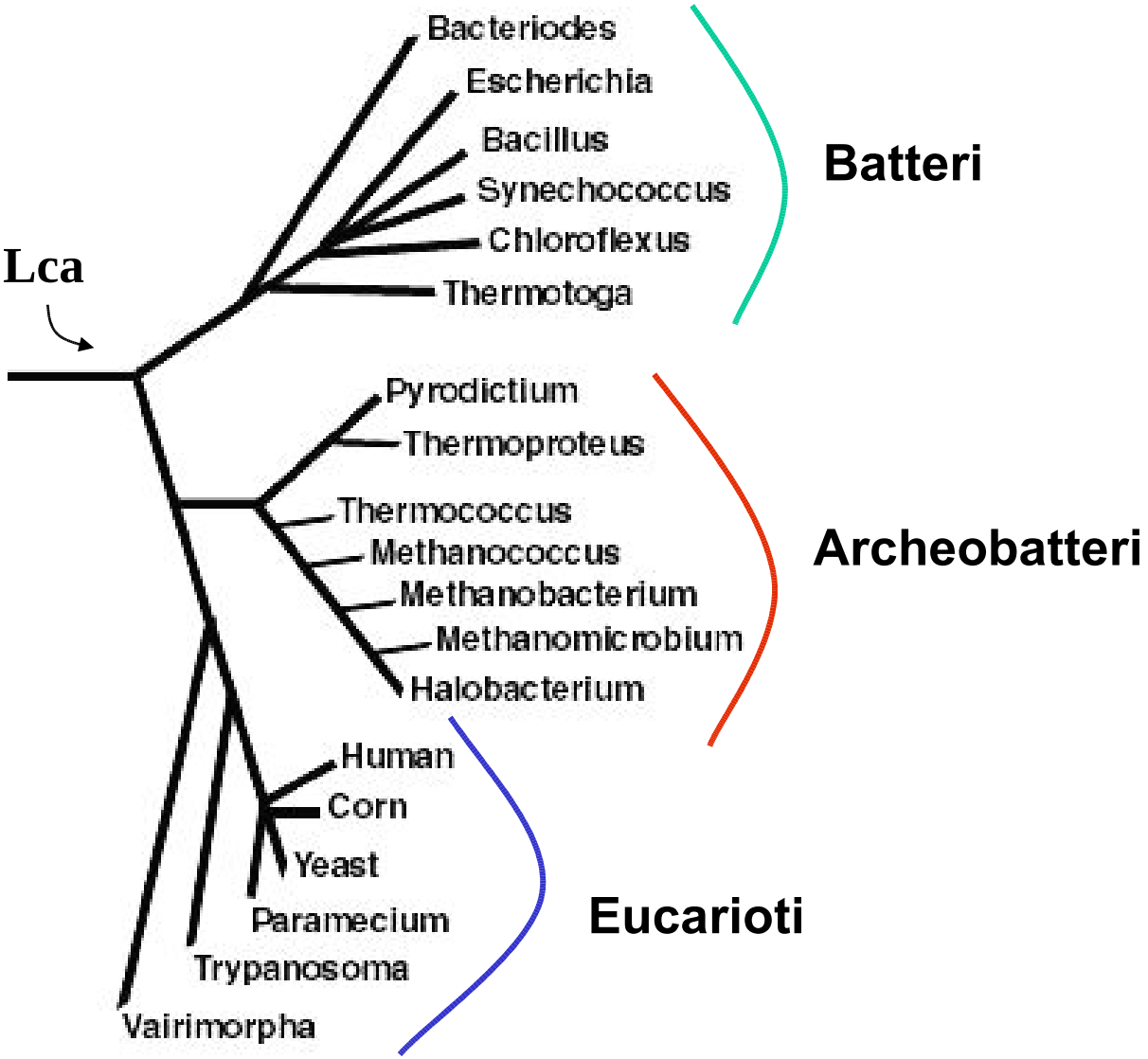
Alberi filogenetici

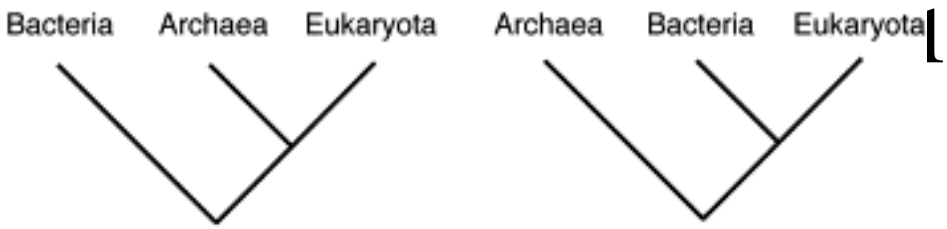


The tree of life

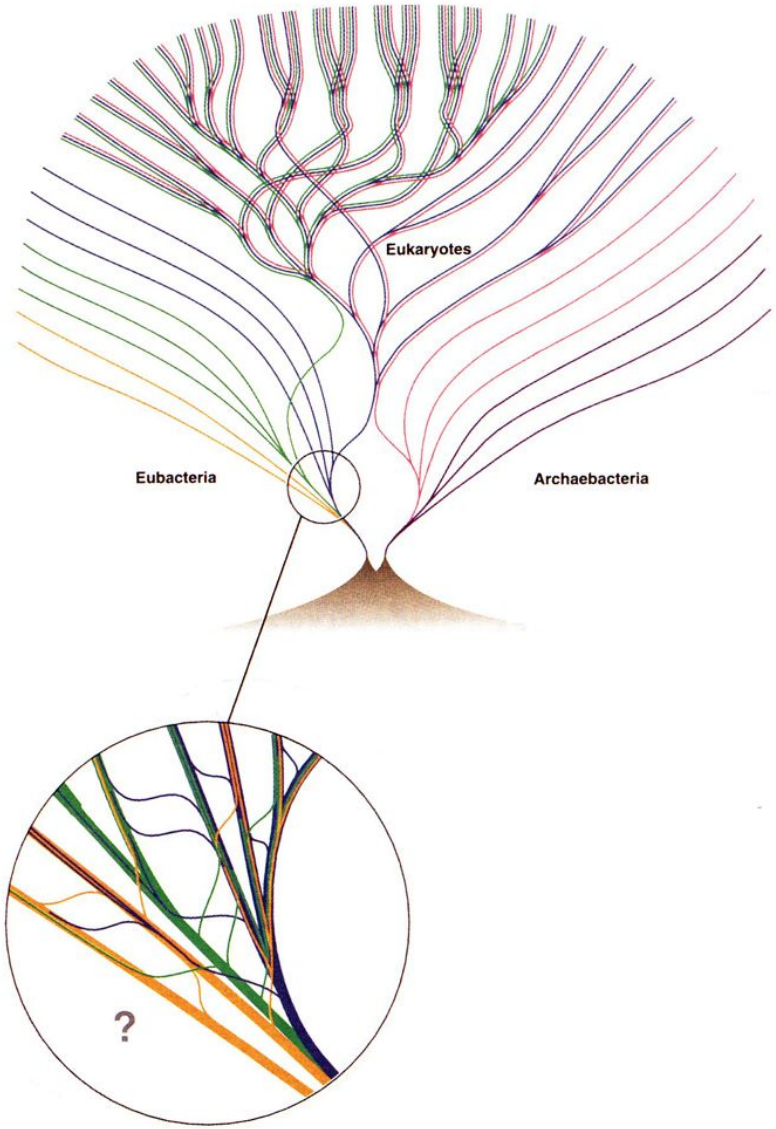
Albero filogenetico costruito con le sequenze della subunità piccola dell'RNA ribosomale. Tutte le forme viventi condividono un comune ancestore (LCA, "last common ancestor") e sono raggruppabili in tre gruppi principali. Batteri, Archeobatteri e Eucarioti. L'ordine di diramazione tra i tre gruppi (vale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



Complessità dell'evoluzione dei i



argininosuccinate synthase	adenylosuccinate synthase
aspartate transcarbamoylase	argininosuccinate lyase
aspartyl-tRNA synthetase	aspartate aminotransferase
ATP synthase alpha chain	deoxyribodipyrimidine photolyase
ATP synthase beta chain	dihydroorotate oxidase
DNA-dependent RNA polymerase	ferredoxin
DNA polymerase II	glutamate dehydrogenase
elongation factor Tu	glutamine synthetase
elongation factor 2	glycine hydroxymethyltransferase
histidyl-tRNA synthetase	lactate dehydrogenase
hydroxymethylglutaryl-CoA reductase	pyrroline-5-carboxylate reductase
indole-3-glycerol phosphate	pyruvate kinase
isocitrate dehydrogenase	ribose-phosphate pyrophosphokinase
isoleucyl-tRNA synthetase	threonyl-tRNA synthetase
ribosomal proteins (various)	triose-phosphate isomerase
RNA polymerase subunit A	UDP-glucose epimerase
RNA polymerase subunit B	valyl-tRNA synthetase
tryptophanyl-tRNA synthetase	
tyrosyl-tRA synthetase	



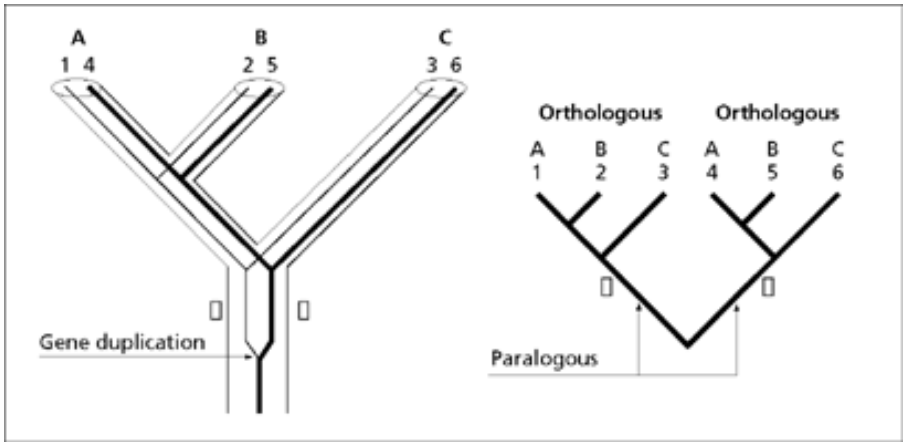
L'analisi delle relazioni filogenetiche esistenti per numerose famiglie di geni rivela per i genomi una storia più complessa e non descrivibile con un semplice albero filogenetico. L'albero della vita ha probabilmente una struttura reticolare che testimonia di numerosi eventi di trasmissione di materiale genetico successivamente alla divergenza tra le varie linee evolutive.

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

Xenologia: I geni, successivamente alla separazione (per speciazione o duplicazione) vengono trasmessi orizzontalmente da un organismo all'altro. La filogenesi riflette solo in parte la storia dei geni



I geni possono essere mosaici complessi:
i diversi domini di una proteina modulare possono avere una diversa storia evolutiva

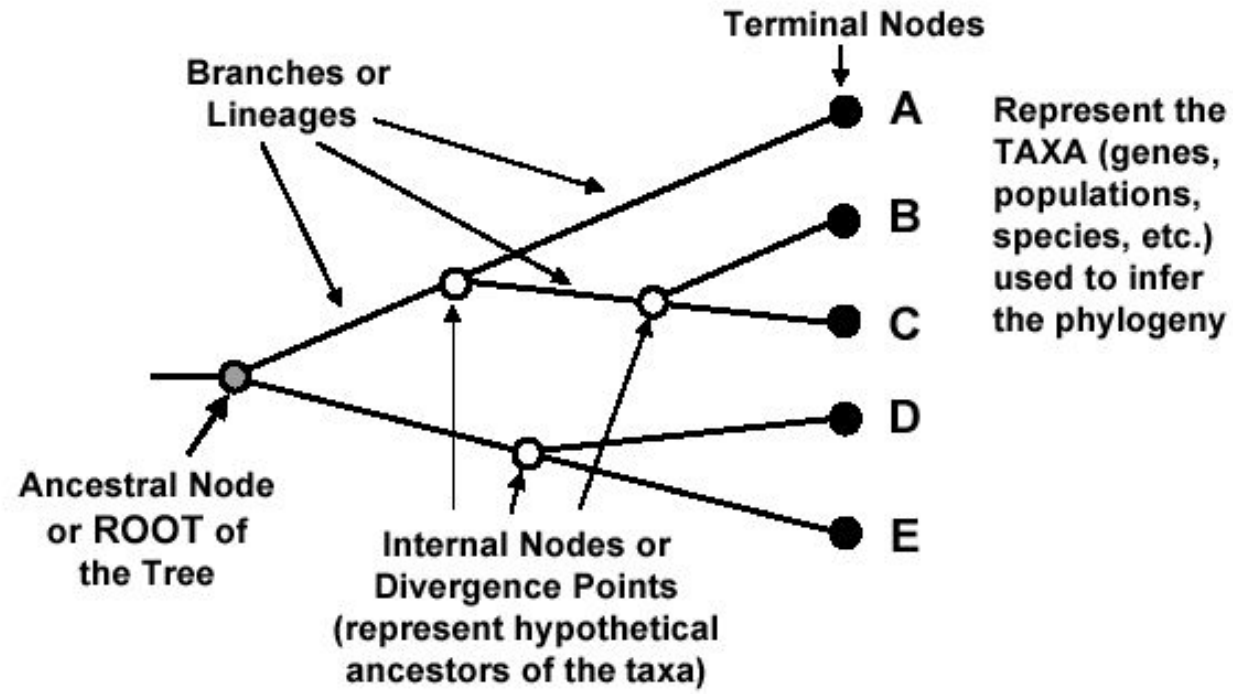
Alberi evolutivi

Albero evolutivo	
:=	Un grafo bidimensionale che mostra le relazioni evolutive esistenti tra diverse entità (geni,sequenze,organismi)

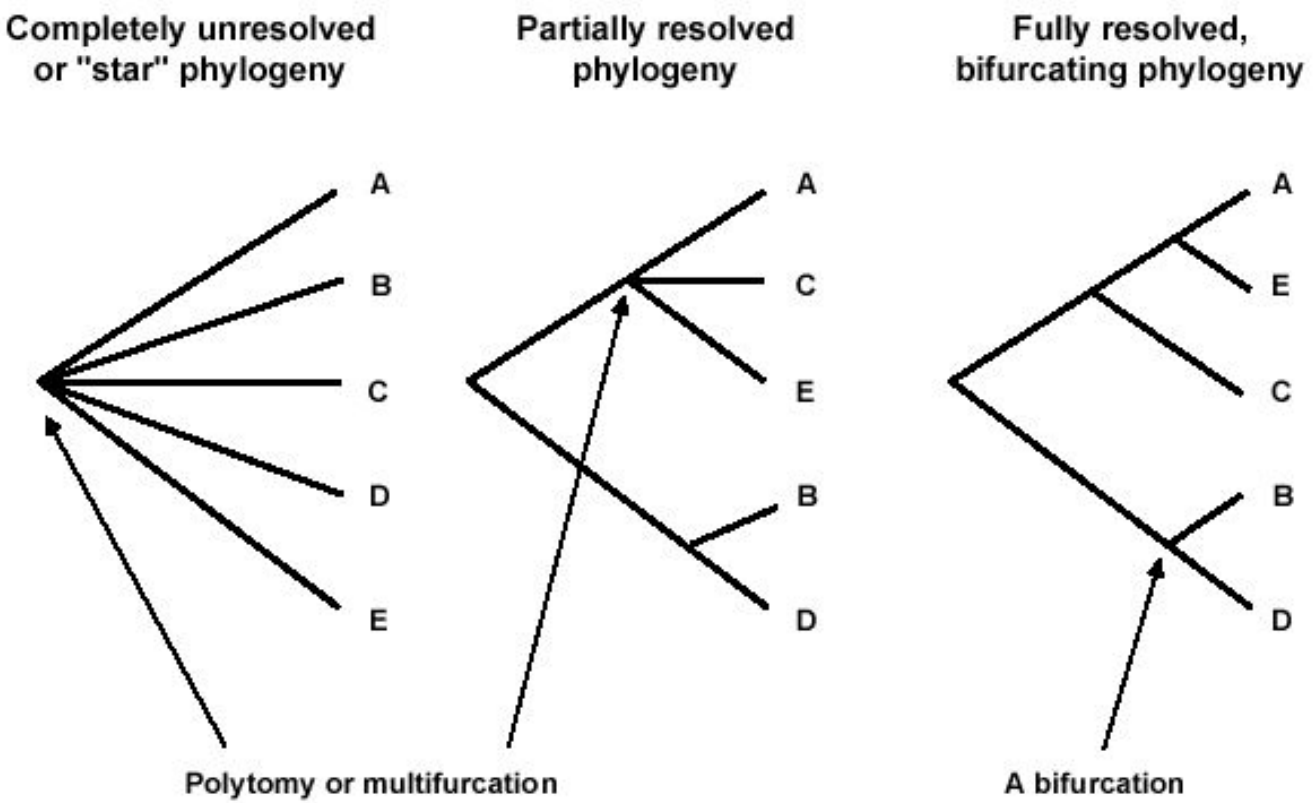
Terminologia degli alberi evolutivi

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

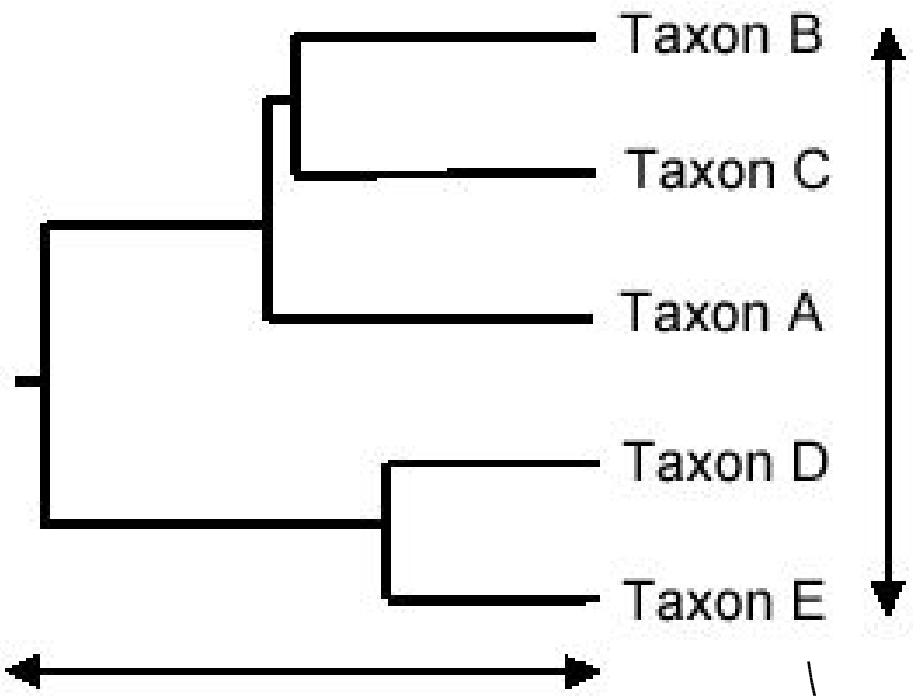


Risoluzione dell'ordine di diramazione dei taxa



Lo scopo dell'inferenza filogenetica è la risoluzione dell'ordine di diramazione dei taxa. Questo corrisponde ad un albero che presenta solamente biforcazioni (dicotomie). Una multiforcazione (politomia) può essere "soft", ovvero risolvibile con l'aggiunta di altri dati filogenetici, o "hard", ovvero causata dal fatto che si è avuta separazione circa contemporanea tra i taxa.

Rappresentazione delle relazioni evolutive



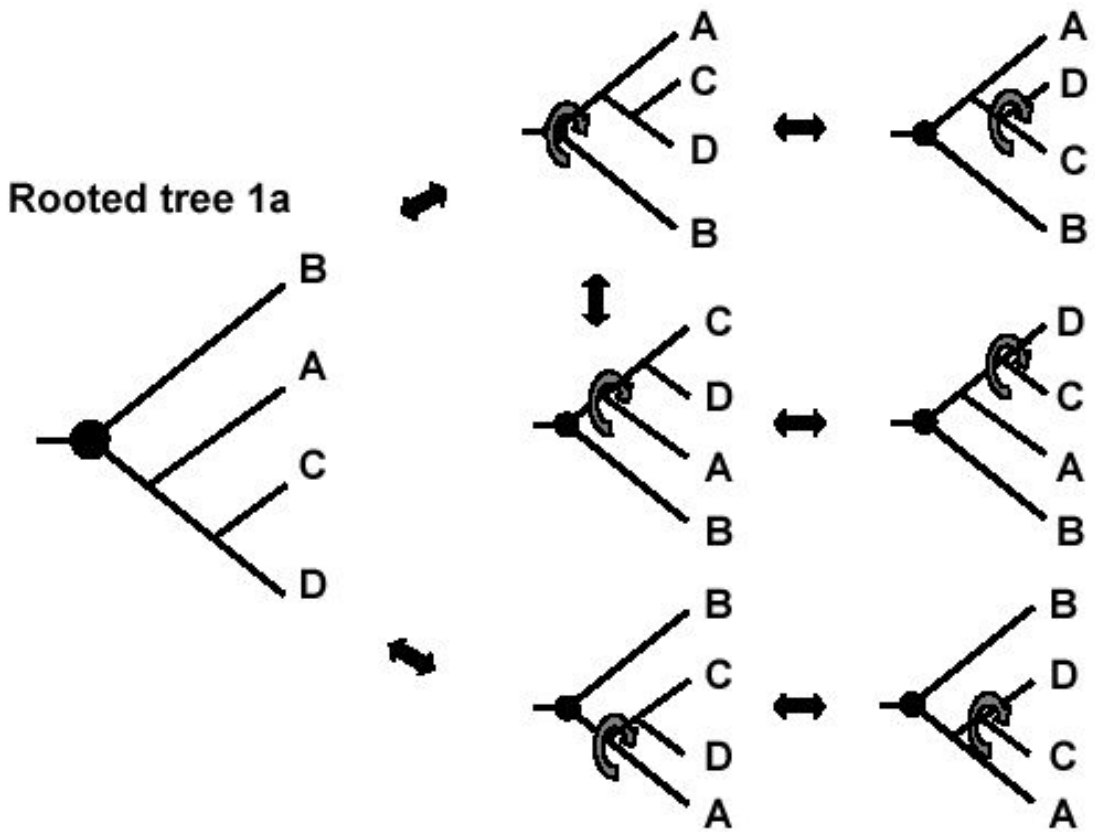
Quest'asse non ha nessun significato

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

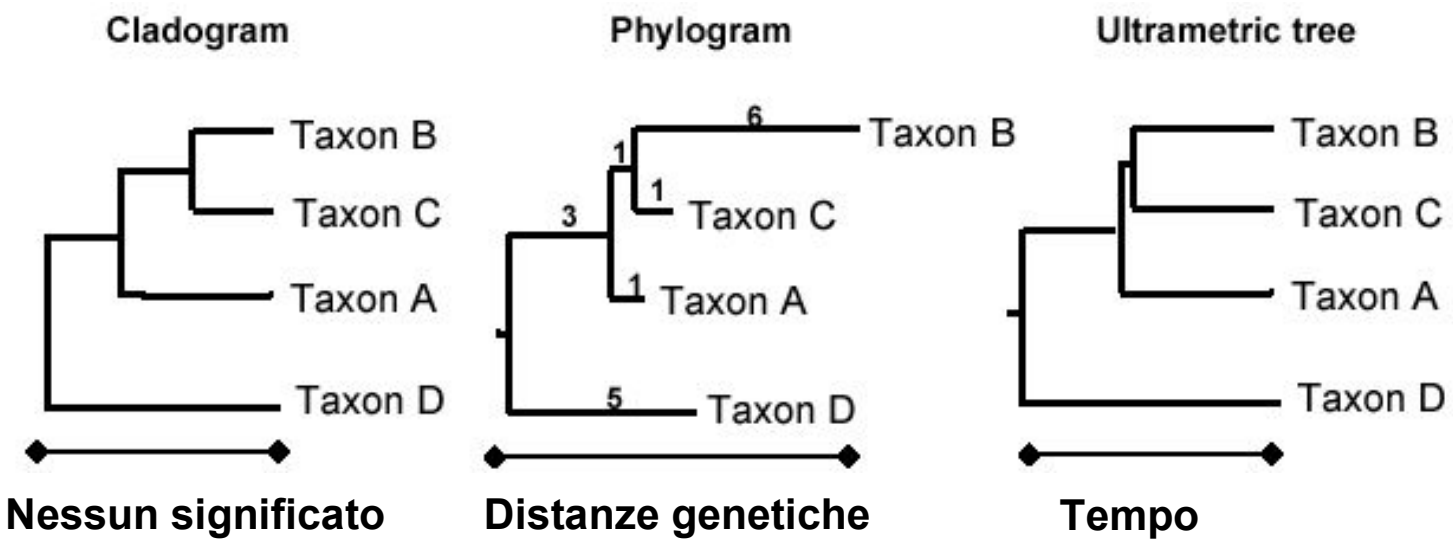
L'albero e le parentesi rappresentano le stesse relazioni evolutive. Ad esempio che **B** e **C** sono più vicini tra di loro di quanto non lo sia **A** a ciascuno dei due, e che **A,B,C** formano un "**clade**" che è il "**sister group**" del clade composto da D e E. In un albero con una scala temporale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



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

Due tipi principali di analisi filogenetica

Inferenza filogenetica o “costruzione di alberi”:
L’inferenza sull’ordine di diramazione delle linee evolutive, sulle relazioni evolutive esistenti tra le entità presenti o “taxa” (specie, geni, popolazioni).

Analisi dei caratteri o dei tassi evolutivi:
Utilizza l’analisi filogenetica come supporto analitico per la comprensione rigorosa dell’evoluzione di alcuni caratteri o condizioni di interesse.

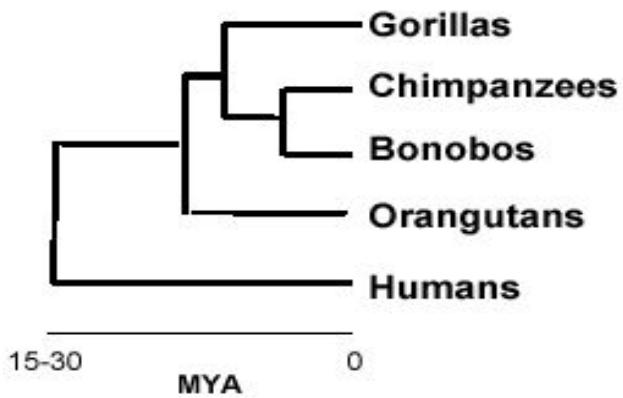
Inferenza filogenetica

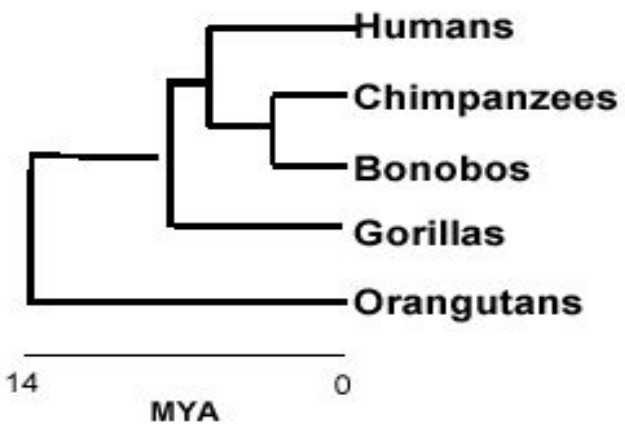
Ricostruzione delle relazioni esistenti tra geni o organismi. In un albero viene dedotto dell'ordine di diramazione dei taxa

Esempi domande a cui si può rispondere grazie alle inferenze filogenetiche

- *Quali sono i rapporti di parentela tra l'uomo e gli altri primati?*
- *Ho davvero sequenziato il DNA di un dinosauro?*
- *Il dentista della Florida ha infettato i propri pazienti con il virus dell'HIV?*
- *Da quali tRNA derivano le SINEs?*
- *E molte altre…*

Qual è la specie più vicina all'uomo moderno?





L'opinione prevalente nell'epoca pre-molecolare era che l'uomo fosse in una linea evolutiva separata rispetto alle grandi scimmie, le quali erano unite in un raggruppamento ("clade") separatosi dall'uomo 15-30 My di anni fa.

Le ricostruzioni filogenetiche effettuate con il DNA e proteine nucleari e mitocondriali mostrano che lo scimpanzé e il bonobo sono più vicini all'uomo che non entrambi al gorilla. Il tempo di separazione è < di 10 My.

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.

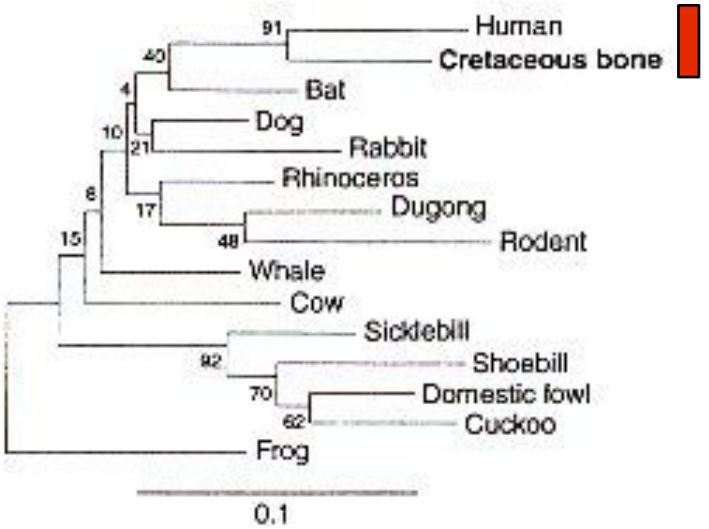
```
          15.627
Consensus CC CTT CTA TTA TCC ATT GTC ATT CTA TTC GTT ATT CCT GTA CTC CAC ACA TCC (C) AAA
2-37                      ..A ... ... ... ..C ... C.. ..C ... T.. ... ... .A. ...  /  ...
3-37      .. ... G.. ... ... ... T.. ... ... ... ... ... ... ... ... ..T ... ..G ..T  /  ...
4-37      .. ... .G. ... ... ... ... ... ... ... ... G.. ... .CC ... ... ... ... ..G  .  ...
31-44                          ... ... ... ..T ... ... ... ... ... ... ... ... ...  /  .G.
2-61      .. ... ... C.. .T. ..C ... ... ... ... .CA ... ... A.. ... ... .T. ...  /  ...
2-18                     ... ... ... ... ... T.. ... ... ... ... .G. ... ...  /  ...
20-61                        ..T ... ... ... ... ... ... ... ... ... .T. ...  /  ...
5-37      .. ... ... C.. T.. .A. T.. ..C ... ... ..A ... ..C A.. ..T ... ... .GT  .  ...
6-37                .. ..T ... ..T ..C ..G ... ..A T.. ... .C. ..T ... ... .GT  /  ...

Consensus CAA CAA AGC ATA ATA TTC CAC CCA TTC AGT CAA TTC CTA TCC TGA TTC TTA GTC CCC GAA
2-37      ... .G. .C. ... .C. ... ... ... ... T.. ...
3-37      ... ... .C. ... ... ... ..A ... ... ... ..G
4-37      ... ... ... ..C ... ... .G. ... ... ... ...
31-44     ..G ... ... ... ... ... ... ... ... C.. ...
2-61      ... ... ... ... ... ..T ... ... ... ... ...
2-18      ... ..T ... ... ... ... ... ... ... ..C ...
20-61     T.. ... ... ..T ... ..T ... ... ... ... ...
5-37      ... ... G.. ... ... ... GGT ... C.. G.C ...
6-37      ... ... ... ... ... ... CG. ... C.A G.. ...

Consensus CCT TTT ACA CTC ACA TG
2-37
3-37      .TA ... ... ... ... ..
4-37      .G.
31-44     ... ... .
2-61      ... ... ... ..A ..C
2-18      .T. ... ... ... ...
20-61     ... ..
```

(c)2000 Benoit Leblanc

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

Analisi dei caratteri e dei tassi evolutivi

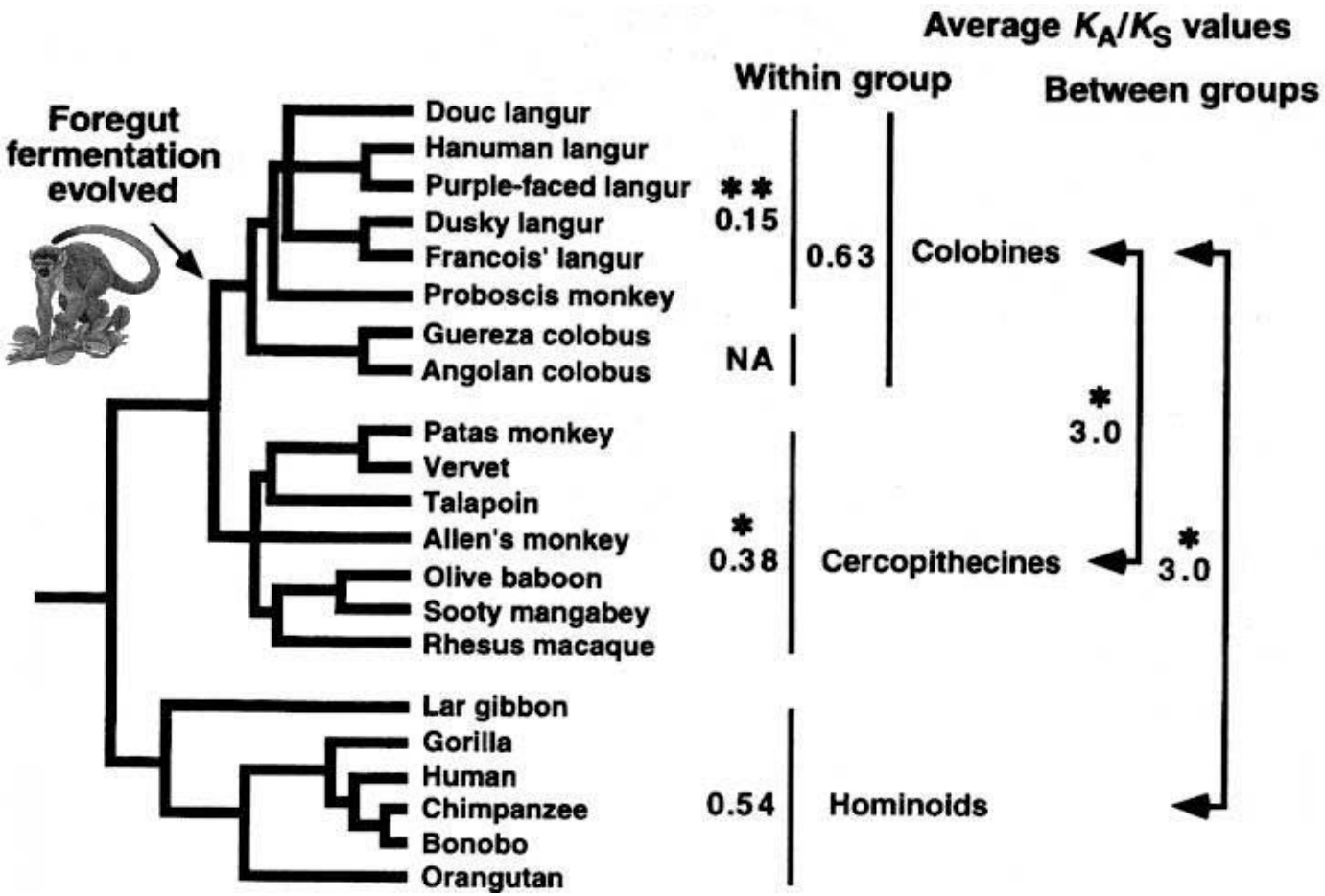
L'analisi filogenetica è usata come supporto per l'analisi dell'evoluzione di alcuni caratteri

Esempi di domande a cui si può rispondere grazie all'analisi dei caratteri e dei tassi evolutivi

- *Ci sono stati episodi di selezione darwiniana positiva nel corso dell'evoluzione?*
- *I tartufi derivano da funghi epigei che hanno perso la capacità di formare corpi fruttiferi emersi? O viceversa?*
- *Sono comparse prima nel corso della evoluzioni le tRNA sintetasi o i tRNA?*
- *E molte altre…*

Selezione darwiniana positiva nelle proteine ?

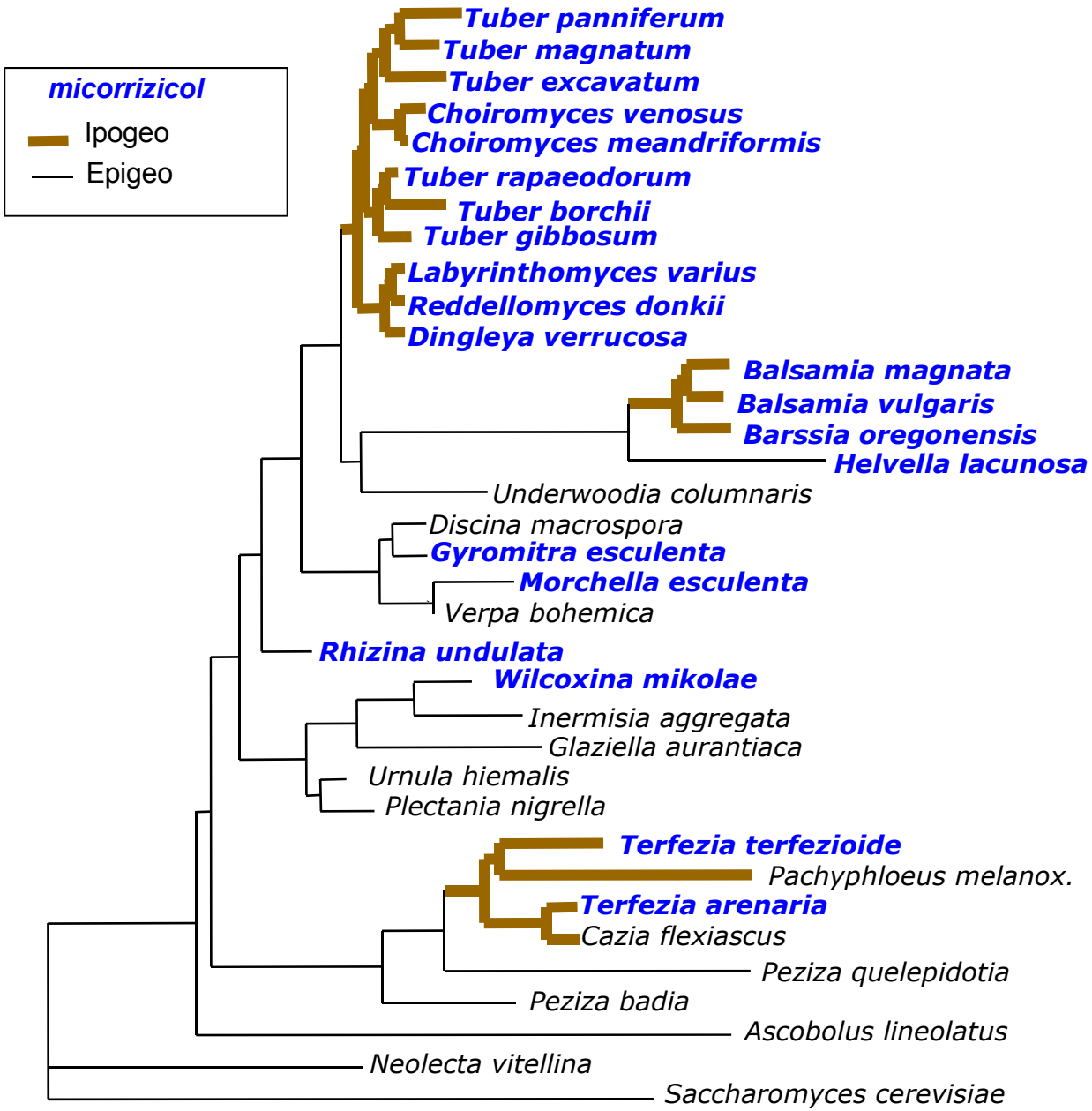
Messier et al. Nature, 1997

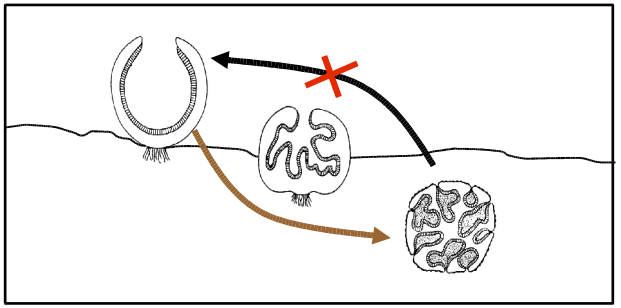


Le scimmie colobine hanno evoluto la ruminazione in modo indipendente da altri mammiferi ruminanti. Il lisozima, proteina normalmente presente nella saliva e nelle lacrime si è adattato all'utilizzazione nel rumine. L'analisi del rapporto **Ka/Ks** (mutazioni non sinonime/mutazioni sinonime) mostra selezione purificatrice (**Ka/Ks<1**) nei confronti all'interno del gruppo delle colobine e selezione darwiniana positiva (**Ka/Ks>1**) nei confronti tra le colobine e gli altri gruppi di scimmie.

Evoluzione dei funghi: emersione o affossamento?

Percudani et al., Mol Biol Evol, 1998





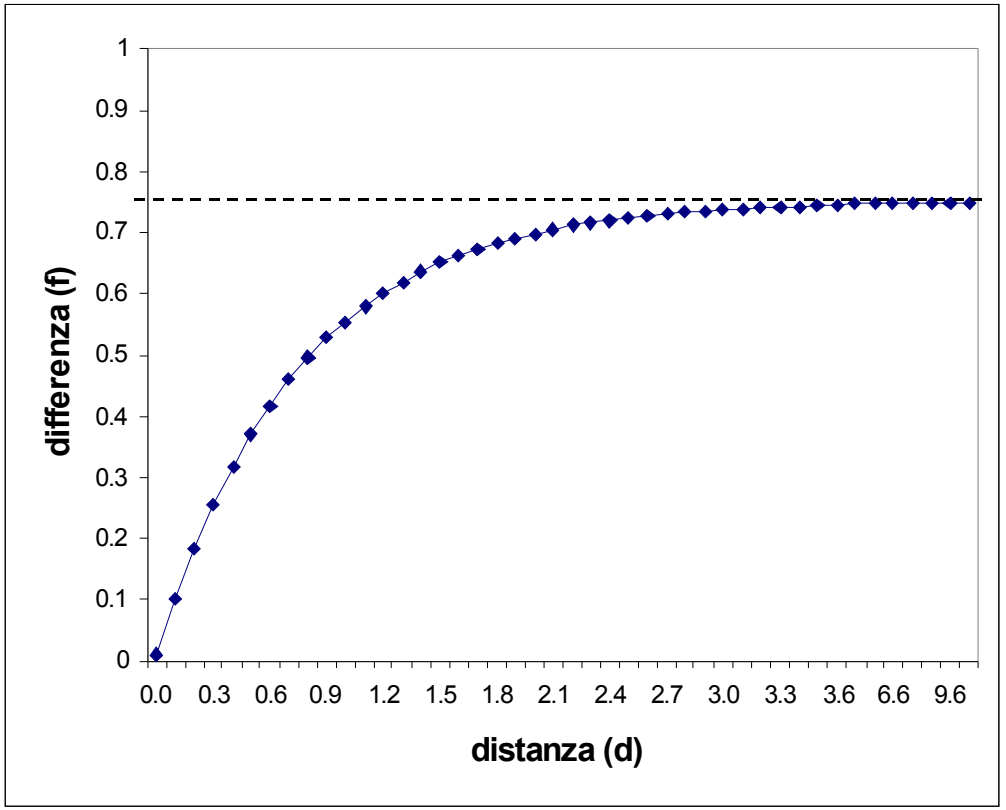
La mappatura “parsimoniosa” (quella che assume il numero minimo di transizioni) dei caratteri ipogeo (corpo fruttifero sotterraneo) e epigeo (corpo fruttifero emergente) nell’albero filogenetico delle Pezizales mostra che la transizione è avvenuta più volte ed in modo indipendente nell’evoluzione dei funghi. La transizione è sempre nella direzione epigeo -> ipogeo, e mai in direzione contraria. Il carattere “micorrizico” è invece meno raggruppato e senza una direzione precisa.

Stima della distanze tra le sequenze

Le distanze genetiche sono calcolate in base alle differenze osservate

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

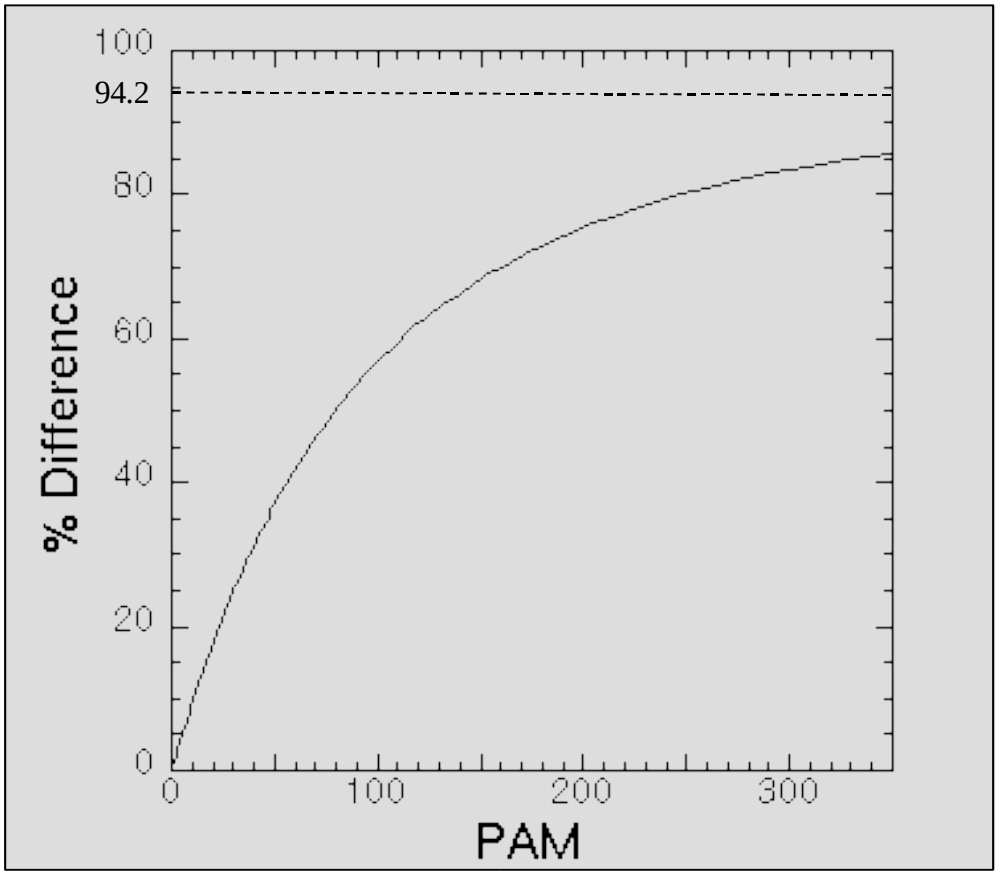
Calcolo delle distanze per sequenze nucleotidiche



$$d = -3/4 \ln(1 - 4/3 f)$$

Formula di Jukes-Cantor

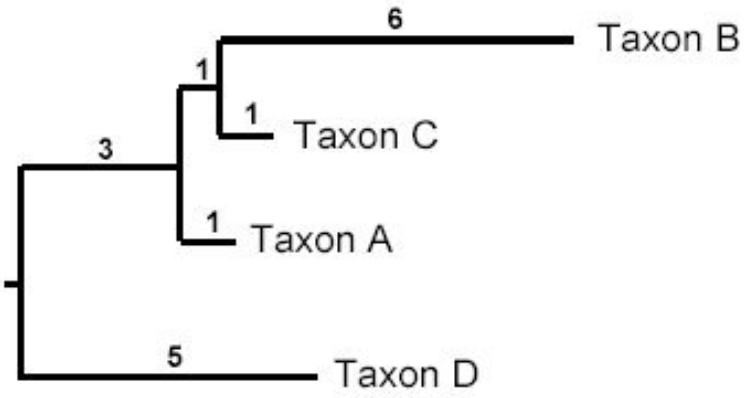
Calcolo delle distanze per sequenze proteiche



%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

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)

Procedimento per l'analisi filogenetica di sequenze

- Selezione delle sequenze (omologhe)
- Allineamento multiplo delle sequenze (selezione di porzione adatta dell'allineamento)
- Applicazione di un metodo di ricostruzione filogenetica
- Visualizzazione dell'albero (eventuale polarizzazione dell'albero con radice)
- Validazione della ricostruzione filogenetica

Due tipi di dati per le ricostruzioni filogenetiche

A) Sequenze allineate:

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

B) 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 vero numero di mutazioni)

Due tipi principali di metodi di ricostruzione filogenetica

Sistemi basati sui caratteri
Si basano su un criterio di ottimizzazione

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

Sistemi basati sulle distanze
Calcolo delle distanze + Metodo di clustering

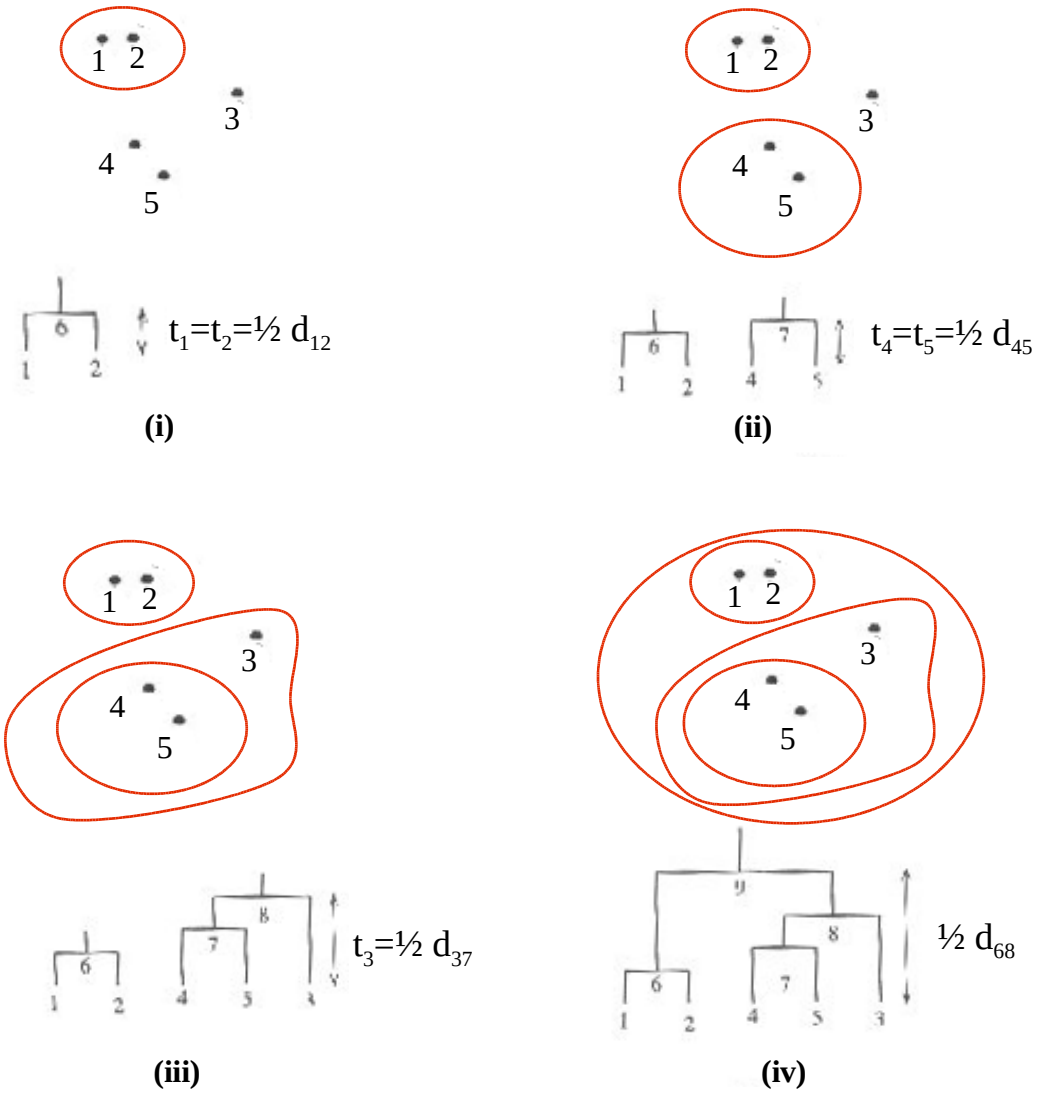
- **Neighbor-joining**
- **UPGMA**

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.

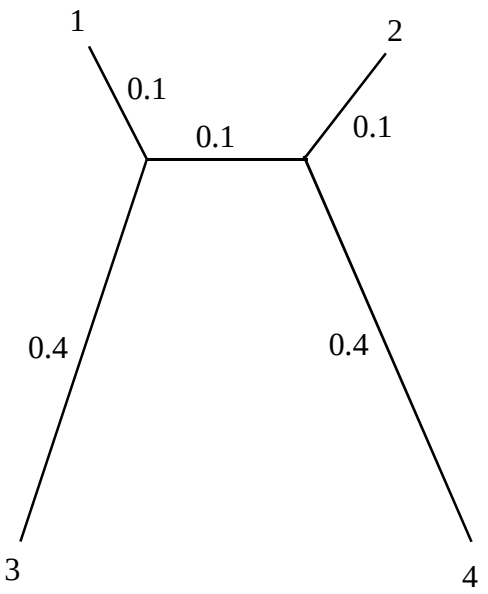


Neighbour-joining

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



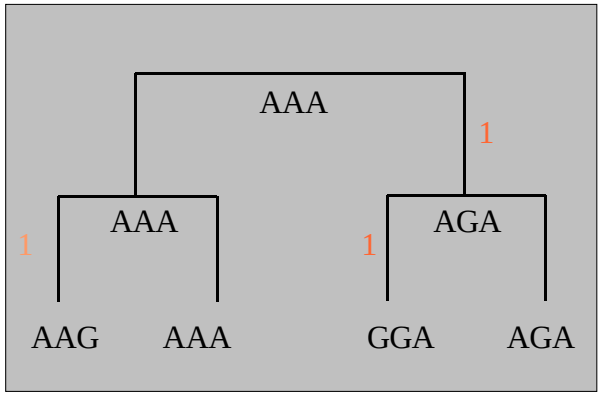
Massima parsimonia

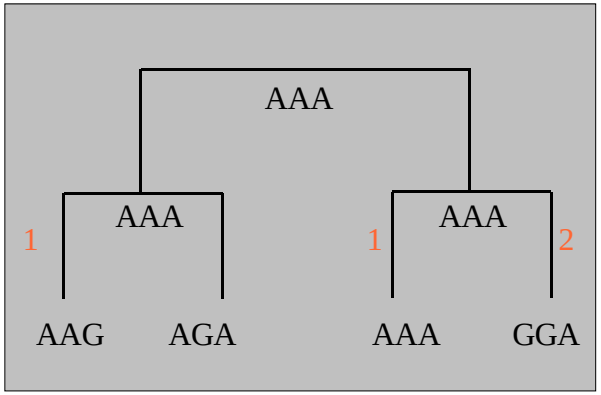
Trova l'albero (unrooted) che spiega le sequenze osservate con il numero minimo di sostituzioni. L'algoritmo a due componenti:

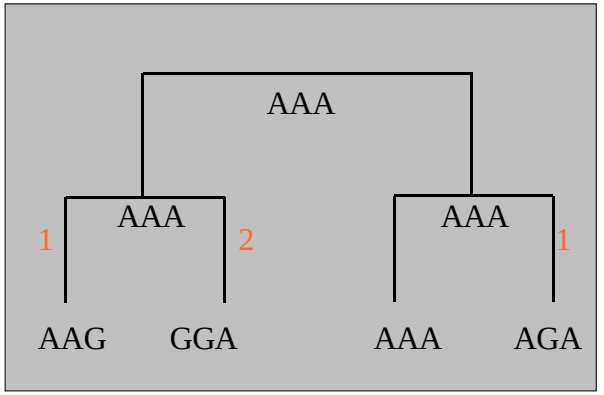
1) Valutazione del **costo di un albero** in termini di mutazioni

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)

Maximum likelihood

Il **likelihood** è la **probabilità di osservare i dati nel caso l'ipotesi sia corretta**

Il sistema statistico di **maximum likelihood** sceglie tra le varie ipotesi quella che massimizza il likelihood, ovvero **l'ipotesi che rende i dati più plausibili**

$$L = Pr(D|H)$$

Maximum Likelihood

Il sistema di **maximum likeliohood (ML)** valuta **la probabilità che i dati di sequenza siano stati prodotti un dato modello evolutivo e da un dato percorso filogenetico**. L'albero (unrooted) che ha il valore più alto di probabilità è detto **albero di ML** e viene considerato l'albero ottimale.

L'algoritmo ha due componenti:

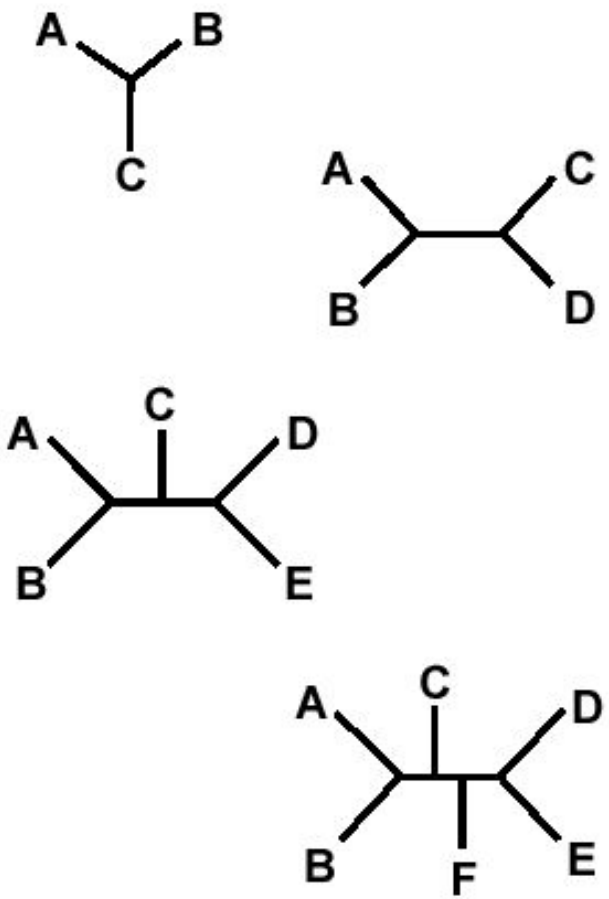
1) Valutazione della **probabilità di ciascuna colonna dell'allineamento di derivare dalle relazioni rappresentate da un dato albero** (likelihood).

2) **Ricerca tra tutti gli alberi possibili** per trovare l'albero con la maggiore verosimiglianza

Vantaggi: solida base probabilistica.

Svantaggi: richiedere un tempo di elaborazione elevato, specialmente in caso di numerosi taxa.

Numero possibile di alberi unrooted



# 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

Ricerca dell'albero "ottimale"

I sistemi di ricostruzione filogenetica basati sui caratteri sono indirizzati a trovare l'albero ottimale secondo un dato criterio oggettivo (numero minore di eventi mutazionali, massima probabilità). Questi sistemi devono necessariamente esaminare tutti gli alberi possibili per le sequenze date e selezionare l'albero ottimale o più alberi ugualmente ottimali. Esistono due sistemi di ricerca:

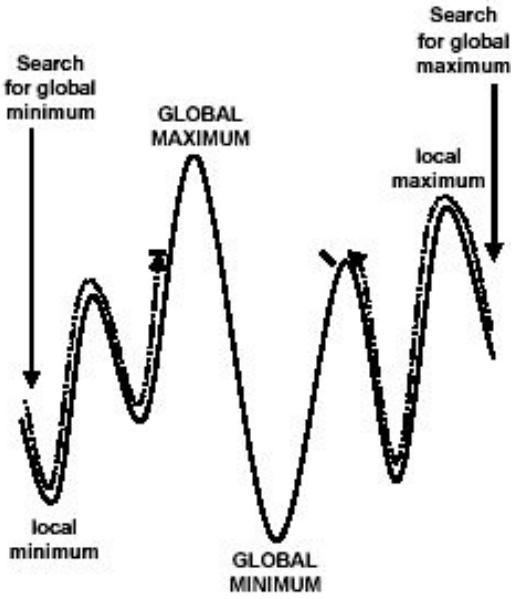
Sistemi esatti:

•*Ricerca esaustiva:* valuta tutti gli alberi possibili.

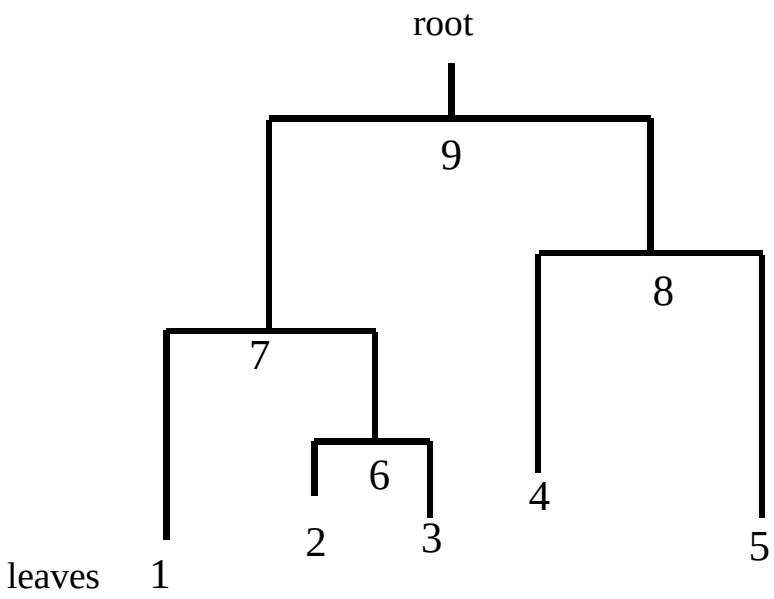
•*Branch and Bound:* inizia a costruire l'albero con un numero crescente di foglie ma si interrompe se l'albero ha un costo che supera quello ottenuto fino a quel punto dall'albero migliore.

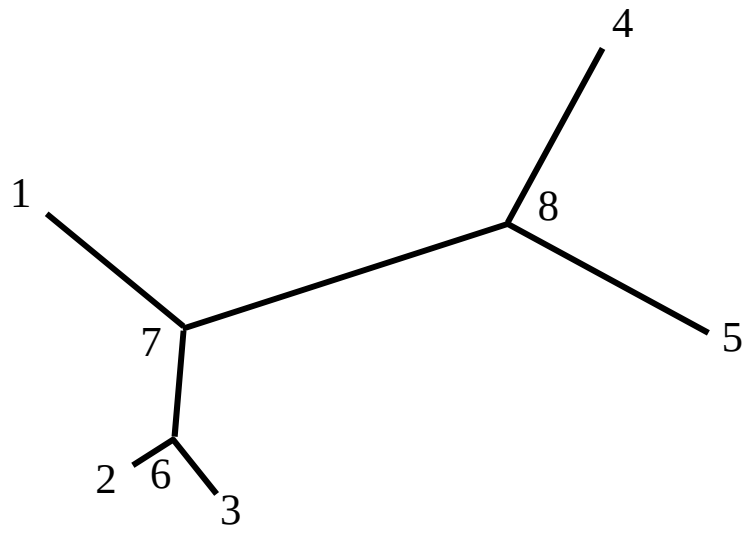
Sistemi euristici

Metodi che cercano di trovare l'albero ottimale senza esplorare tutte le soluzioni. Non garantiscono l'individuazione dell'albero ottimale. La soluzione dipende spesso dall'ordine con il quale vengono aggiunti i vari taxa nella costruzione dell'albero, poiché a seconda dei percorsi si possono incontrare dei minimi locali che impediscono l'individuazione del minimo globale



Alberi rooted e unrooted

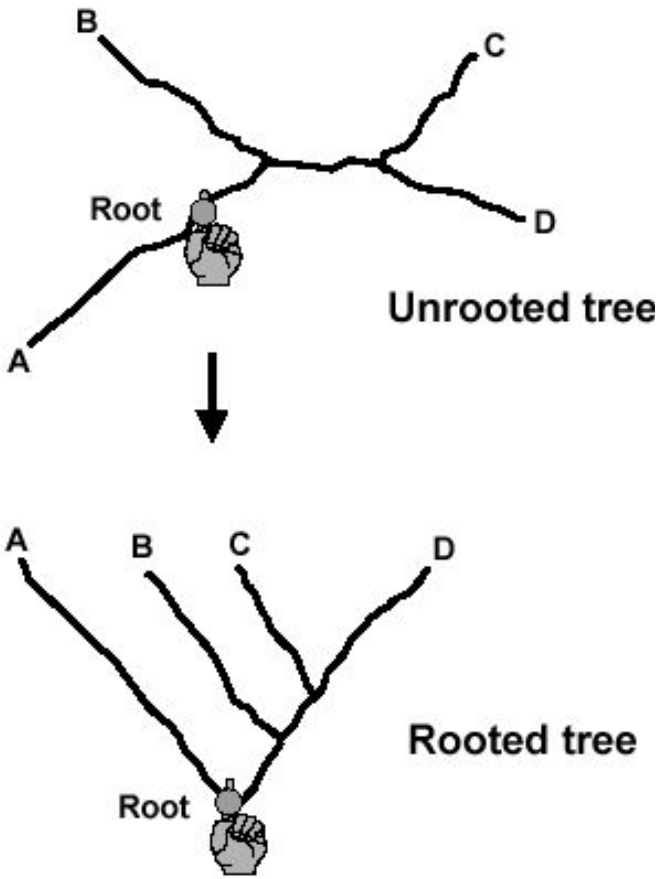




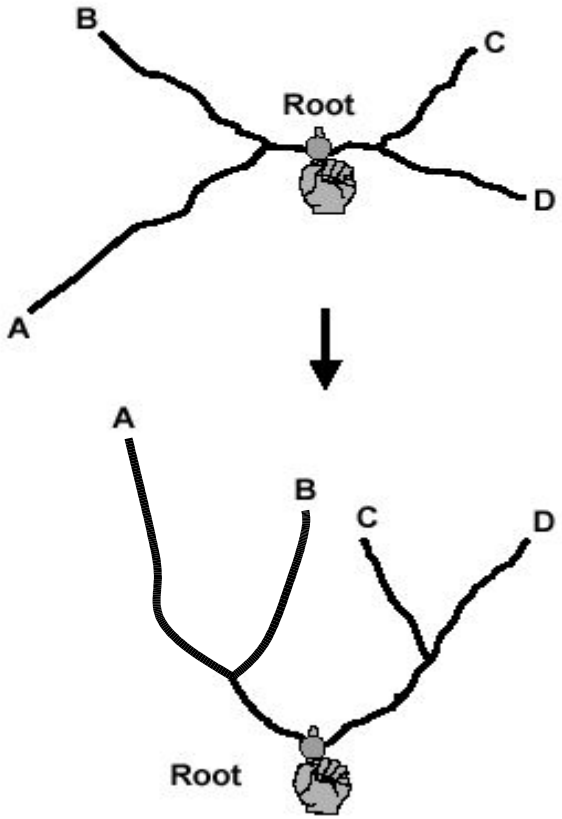
Gli alberi vengono numerati a partire dalle foglie e quindi dai nodi interni via via più lontani. Un albero rooted con n foglie o taxa ha 2n-1 nodi e 2n-2 bracci, un albero unrooted ha 2n-2 nodi e 2n-3 bracci. I bracci sono indicati relativamente al nodo col numero inferiore. Ad esempio, il braccio che porta al braccio 1 è indicato con d1.

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

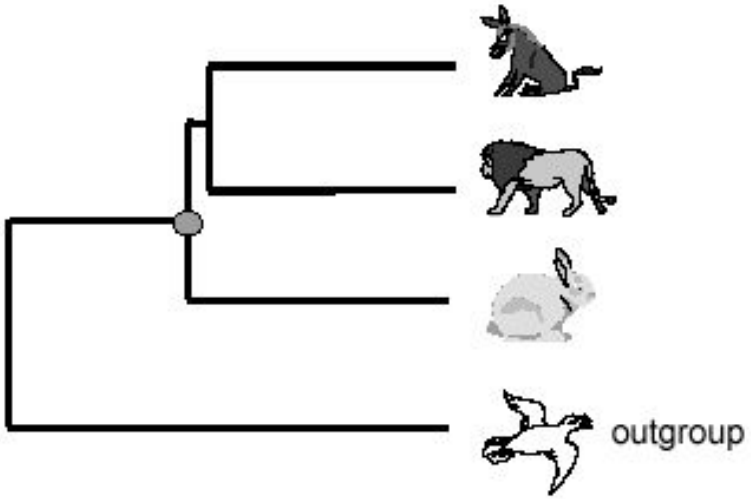


Con questa radice A è più vicino a B di quanto non lo siano C e D

Due modi di posizionare la radice

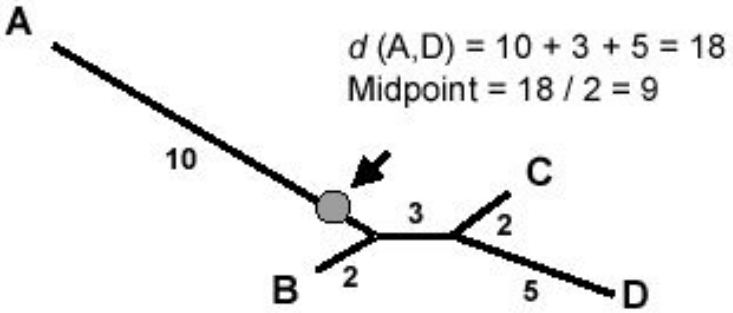
Outgroup

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



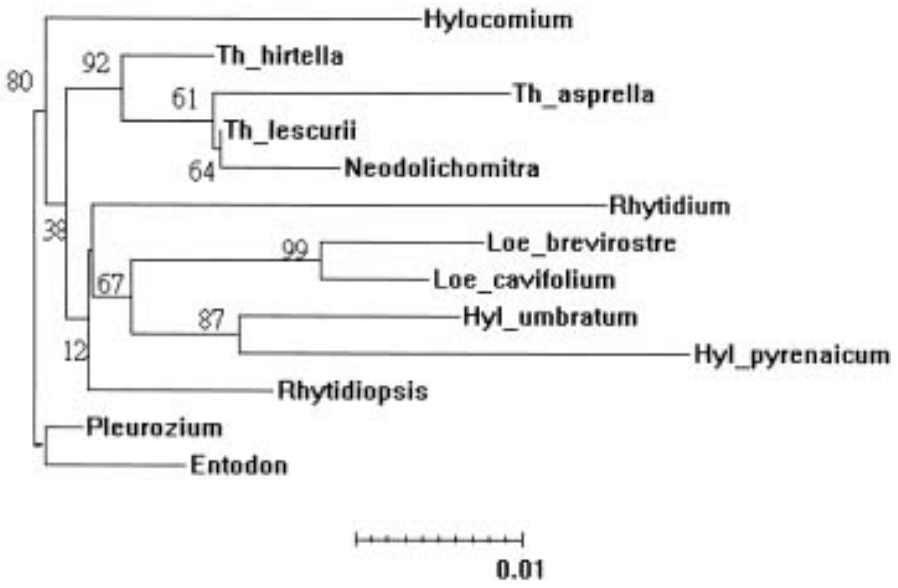
Affidabilità della predizione filogenetica

Ci sono due sistemi per stimare il grado di fiducia che va accordato ad una data ricostruzione filogenetica. E' consigliato l'uso di entrambi

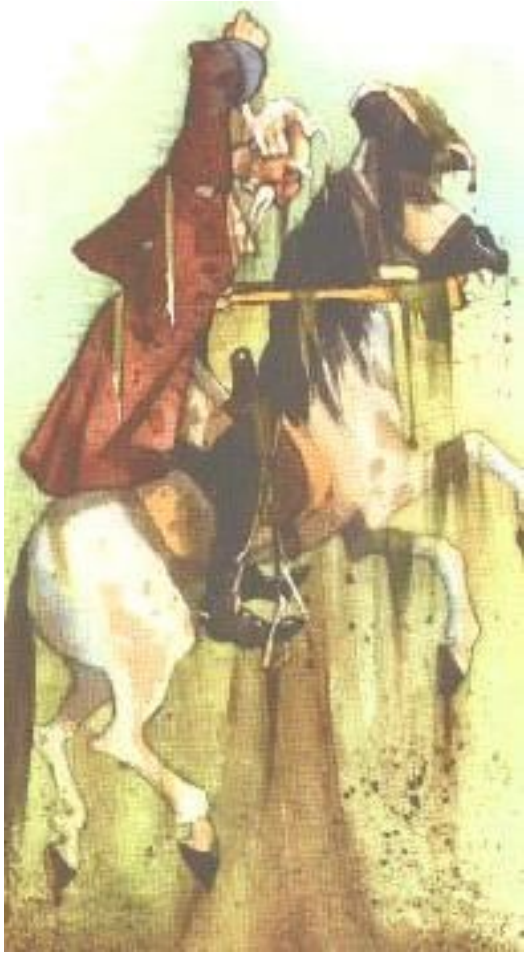
1) Confronto delle topologie ottenute con metodi diversi di costruzione di alberi, preferibilmente uno basato sulla distanza e uno sui caratteri

2) Stima statistica della affidabilità dei vari raggruppamenti attraverso sottocampionamenti casuali dei dati (**bootstrap**)

Albero filogenetico con indici di bootstrap. I valori in corrispondenza dei nodi rappresentano il numero di volte che un dato raggruppamento è stato ottenuto nei set campionati a partire dai dati originari. Si usano di solito campionature da 100 a 1000 set.

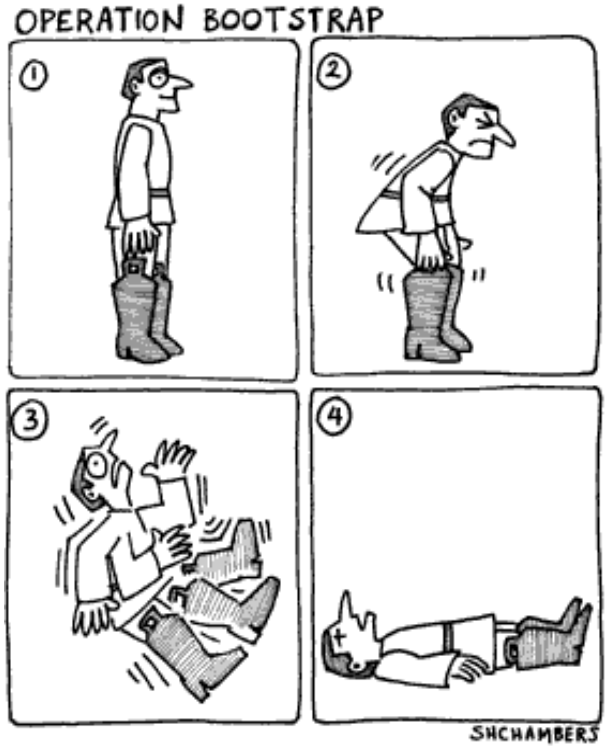


Bootstrap



:=

Metodo statistico per stimare distribuzioni difficili da calcolare attraverso la creazione e l'analisi di dataset artificiali.
Bootstraps sono le linguette per calzare gli stivali. Il nome deriva da una espressione inglese che evoca la "miracolosa" capacità del metodo di generare distribuzioni statistiche dal nulla.



"Pull itself up by its bootstrap"

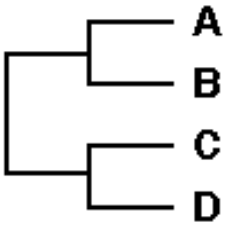
Bootstrap trees

Campione1 0 1 2 0 3 0 1 2 0 1 (<- volte in cui il sito è campionato)

	0	1	2	0	3	0	1	2	0	1
a	A	G	G	C	U	C	C	A	A	A
b	A	G	G	U	U	C	G	A	A	A
c	A	G	C	C	C	C	G	A	A	A
d	A	U	U	U	C	C	G	A	A	C

replica1

a	G	G	G	U	U	U	C	A	A	A
b	G	G	G	U	U	U	G	A	A	A
c	G	C	C	C	C	C	G	A	A	A
d	U	U	U	C	C	C	G	A	A	C

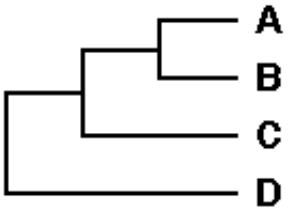


Campione2 1 0 0 0 2 2 2 0 0 3

	1	0	0	0	2	2	2	0	0	3
a	A	G	G	C	U	C	C	A	A	A
b	A	G	G	U	U	C	G	A	A	A
c	A	G	C	C	C	C	G	A	A	A
d	A	U	U	U	C	C	G	A	A	C

replica2

a	A	U	U	C	C	C	C	A	A	A
b	A	U	U	C	C	G	G	A	A	A
c	A	C	C	C	C	G	G	A	A	A
d	A	C	C	C	C	G	G	C	C	C

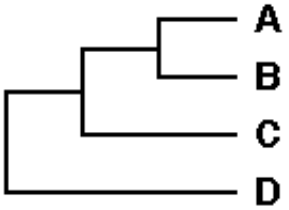


Campione3 1 0 0 0 2 2 2 0 0 3

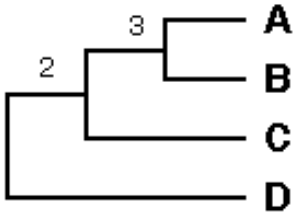
	1	0	0	0	2	2	2	0	0	3
a	A	G	G	C	U	C	C	A	A	A
b	A	G	G	U	U	C	G	A	A	A
c	A	G	C	C	C	C	G	A	A	A
d	A	U	U	U	C	C	G	A	A	C

replica3

a	A	U	U	C	C	C	C	A	A	A
b	A	U	U	C	C	G	G	A	A	A
c	A	C	C	C	C	G	G	A	A	A
d	A	C	C	C	C	G	G	C	C	C



Le repliche hanno lo stesso numero di siti dei dati originali. Ovvero le colonne vengono campionate casualmente nei set di replica fino ad avere un numero di colonne pari al set originale.



Albero di consenso con indici di bootstrap