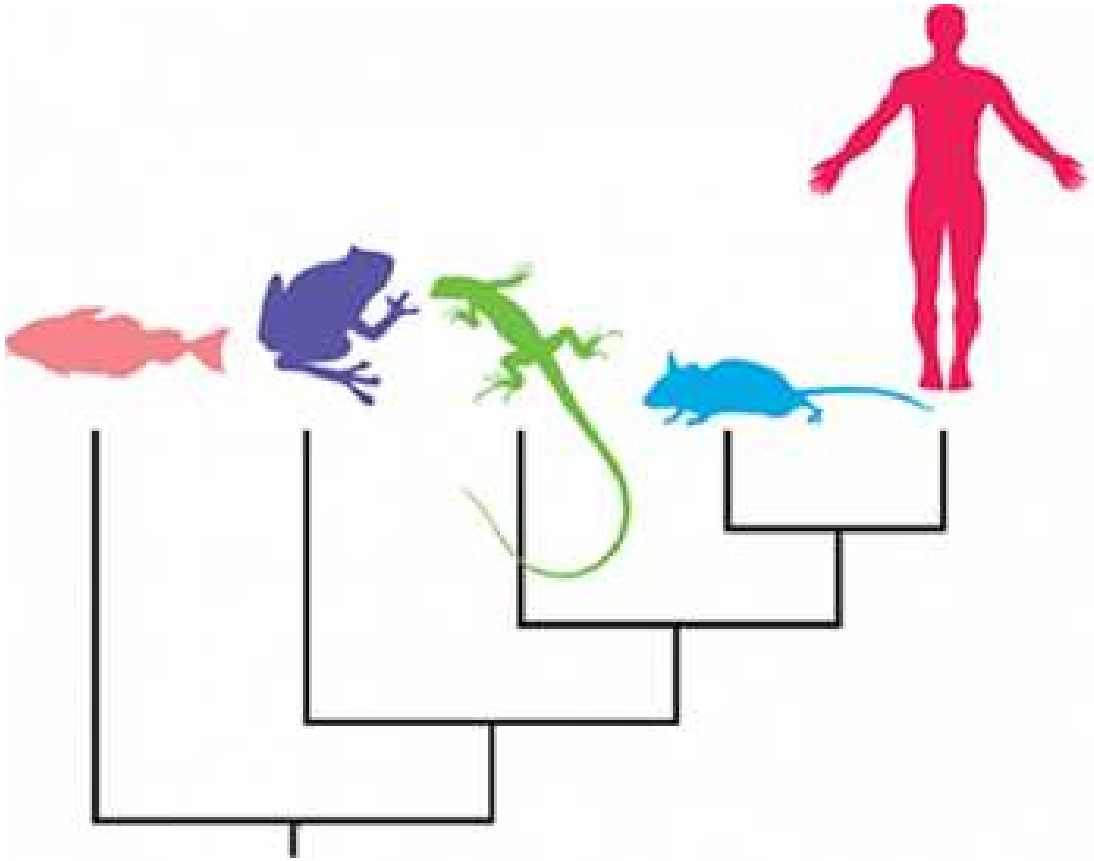
Costruzione di Alberi filogenetici – Verifiche di affidabilità

Giuseppe Manna



28 giugno 2016

Direzione Operativa Diagnosi delle Malattie Virali

Ricostruzioni filogenetiche basate sul DNA: vantaggi

Descrizione dei caratteri non ambigua

Nessun problema di omoplasie fenotipiche

Posso analizzare tanti caratteri ==> tanta variabilità e maggiore possibilità che i siti congruenti prevalgano su quelli incongruenti

Maggiore facilità di stimare tempi di divergenza (cioè la lunghezza dei rami)

Modelli statistici rigorosi

Posso analizzare DNA/RNA non codificante

Tutti gli individui hanno DNA!

Ricostruzioni filogenetiche basate sul DNA: svantaggi

Pochi stati del carattere (A,C,T,G)

Tasso di mutazione a volte molto elevato

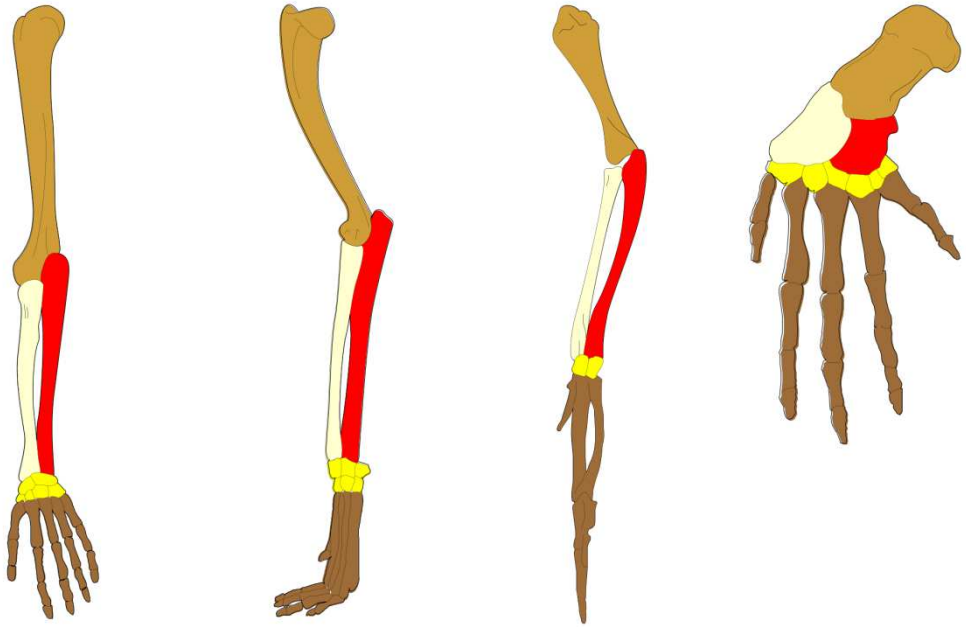
Mutazioni ricorrenti modificano la relazione tra distanza genetica e distanza temporale

Duplicazioni e trasferimento orizzontale di geni possono creare problemi nella ricostruzione filogenetica

Omologia e omoplasia non possono essere distinte attraverso una analisi dettagliata come per caratteri fenotipici

I modelli di evoluzione del DNA possono essere molto complessi

Alberi di **geni** e alberi di **specie** possono essere diversi



La “vera” somiglianza per costruire alberi: l’omologia

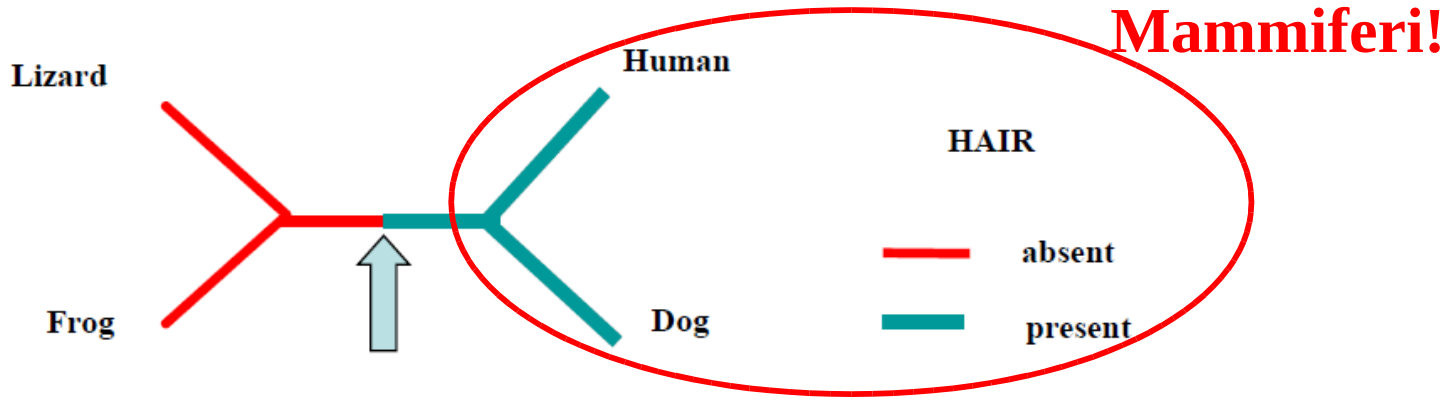
Un carattere che si è evoluto una volta e non ha subito reversioni ha un valore filogenetico

La somiglianza in due linee filogenetiche per un carattere di questo tipo è detta *omologia*

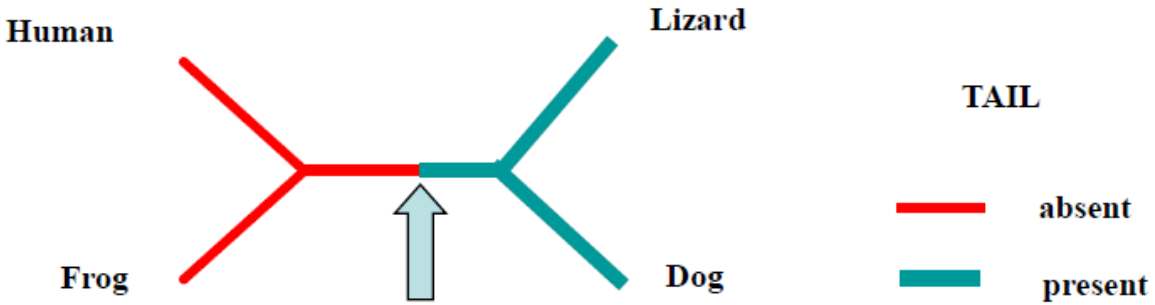
In altre parole, un carattere di questo tipo è simile (o presente) in due specie perchè era così nel loro antenato comune

Le ali sono un carattere omologo in aquila e colombo perché l’antenato comune era alato. Stesso ragionamento per i peli in cane e uomo

Le omoplasie producono incongruenze



I due alberi sono diversi, ma esiste solo un albero "vero" ==> i due caratteri sono incongruenti, almeno uno deve essere omoplasico



Costruzione di un albero filogenetico

1. **Costruire il dataset**

   Ottenere le sequenze geniche o proteiche degli organismi in esame

2. **Allineamento multiplo di sequenze**

   Registrare le sequenze tra di loro (inserire gap)

3. **Verifica statistiche sulle sequenze**
   Scelta del **modello di sostituzione** adeguato
   Verifiche sull'informatività filogenetica dell'allineamento

4. **Clustering e costruzione dell’albero**

   Derivare l’albero filogenetico

5. **Validazione**

   Determinare la robustezza statistica dell’albero

Proteine o acidi nucleici?

Sequenze proteiche:

- necessitano di matrici si sostituzione 20x20, molto complesse da trattare
- sono espressione di sole regioni codificanti
- aminoacidi identici possono essere espressione di più codoni
- informazione grezza ma pulita: **adatte a trovare relazioni generali**.

Sequenze nucleotidiche:

- sono descrivibili con matrici 4x4
- possono essere estratte da sequenze genomiche non codificanti
- non hanno degenerazione né ridondanza
- informazione dettagliata ma “rumorosa”: adatte a trovare relazioni tra organismi “evolutivamente” vicini.
- Si possono anche utilizzare le parti non codificanti del genoma!

In analisi più complesse si possono considerare entrambi

Allineamento

Allineamento tra due sequenze: trovare un sistema di riferimento comune, inserendo gap (-)

Problema: occorre allineare simultaneamente molte sequenze

Soluzione: “Progressive sequence alignment”

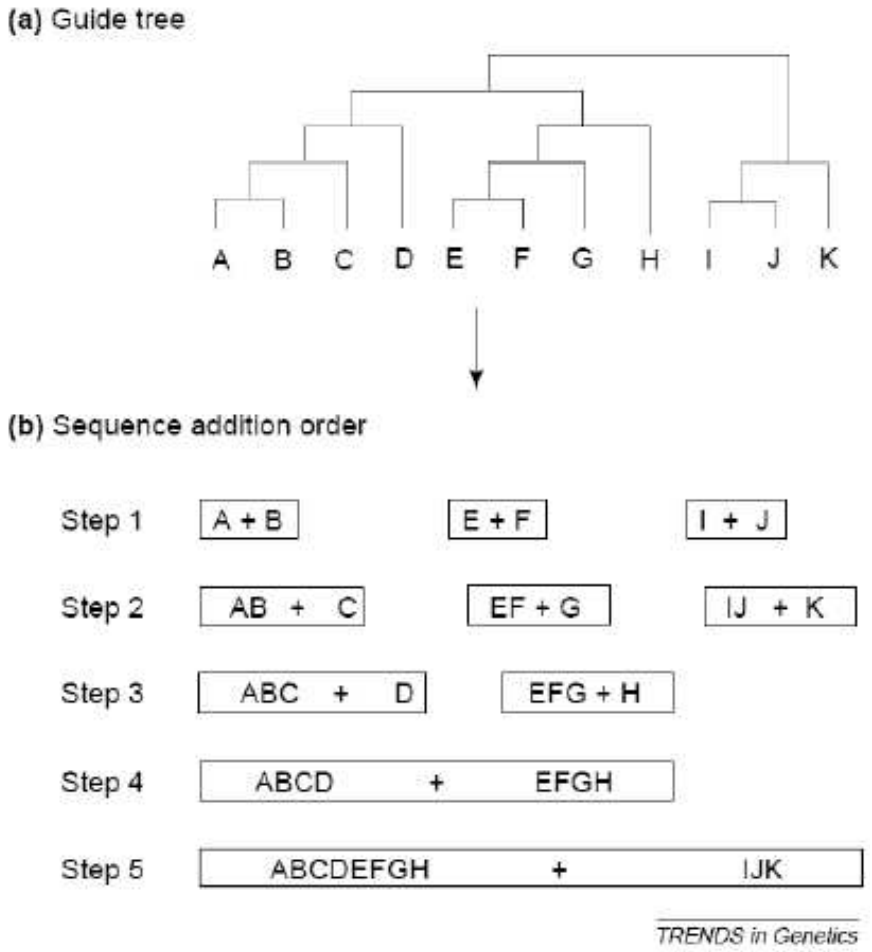
```
CAGATCTTGACGATCCCAAGTGGTTCATTGGCTTTAGATGAAG
TACCGATCTTGACGATCCCAAGTCATTGGCTTTAGATGAAG
CAGATCTTGACGATCCCAAGTGGTTCATTGGCTTTAGATGAAG
CAGATCTTCACGATCCCAAGTGGTTCATTGGCTTTAGAT

CA--GATCTTGACGATCCCAAGTGGTTCATTGGCTTTAGATGAAG
TACCGATCTTGACGATCCCAAG----TCATTGGCTTTAGATGAAG
CA--GATCTTGACGATCCCAAGTGGTTCATTGGCTTTAGATGAAG
CA--GATCTTCACGATCCCAAGTGGTTCATTGGCTTTAGAT----
```

Allineamento

Progressive Sequence Alignment: Allineare le sequenze passo dopo passo, aggiungendo ogni volta una sola sequenza

La sequenza considerata ad ogni iterazione è la più simile alle sequenze già considerate

Necessità di un albero “grezzo” che faccia da guida (per determinare l’ordine in cui le sequenze vengono aggiunte all’allineamento)



Algoritmi per la costruzione di un albero

Metodi distance-based

Viene calcolata una matrice delle distanze “evolutiva” tra tutte le coppie di sequenze.

La distanza si chiama evolutiva perché tiene conto di un modello evolutivo

L’albero filogenetico viene costruito a partire da questa matrice di distanza

Metodi di clustering generici (UPGMA)

Metodi nati ad hoc per la filogenesi (Neighbor Joining)

Algoritmi per la costruzione di un albero

Metodi tree-searching, due approcci principali

Maximum Parsimony

trova l'albero che minimizza il numero di eventi evolutivi (mutazioni) dall'organismo ancestrale

PRO: può gestire facilmente inserzioni e cancellazioni in alcune condizioni è molto efficiente

CONTRO: se l'albero vero ha un particolare tipo di struttura, la tecnica MP può fallire

Maximum Likelihood

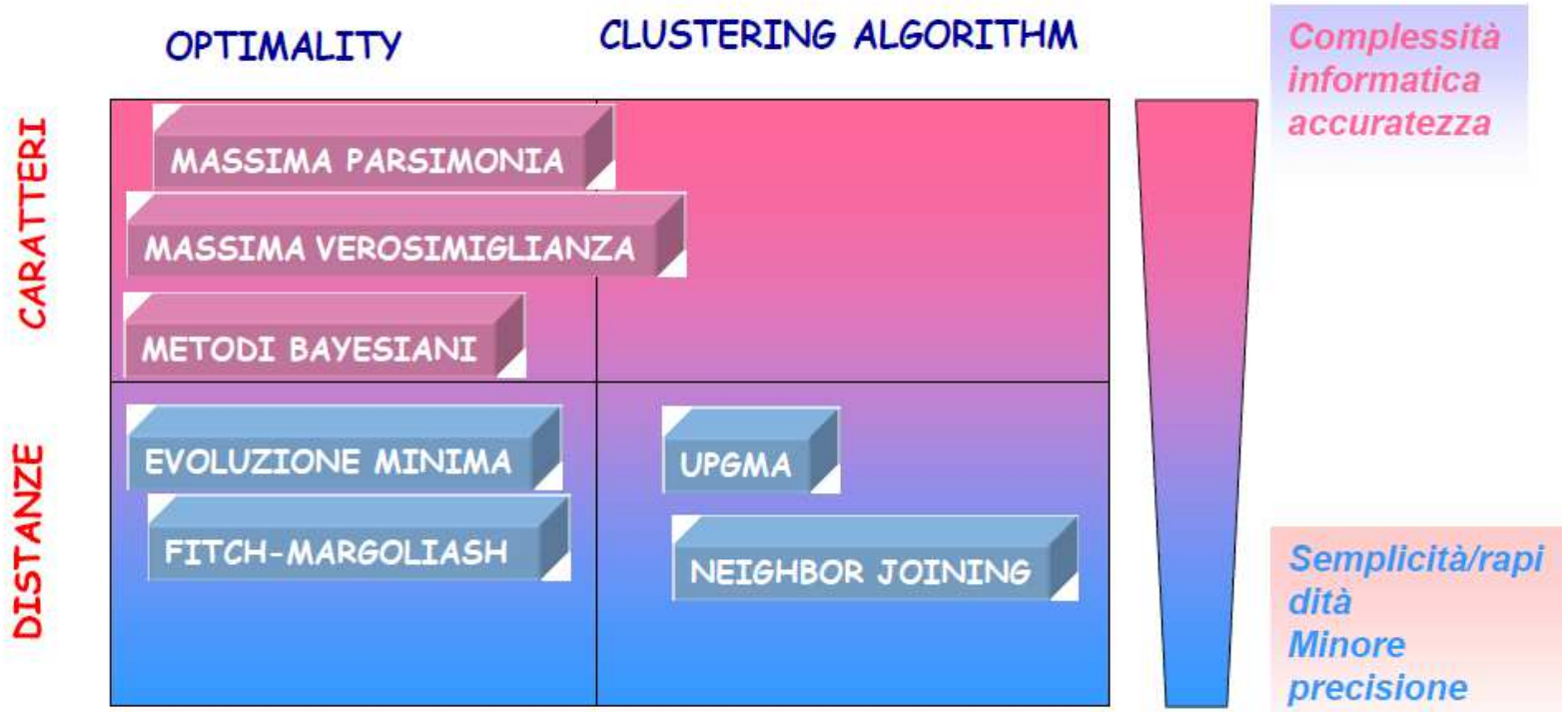
Dato un modello evolutivo, questo metodo seleziona l'ipotesi (l'albero) che meglio spiega i dati osservati

PRO: produce risultati molto accurati

CONTRO: lento (ricerca nel tree-space)

Possibili estensioni con modelli Bayesiani (MCMC)

METODI PER LA COSTRUZIONE DEGLI ALBERI



Modelli di sostituzione

Calcolo delle distanze e modelli evolutivi di sostituzione

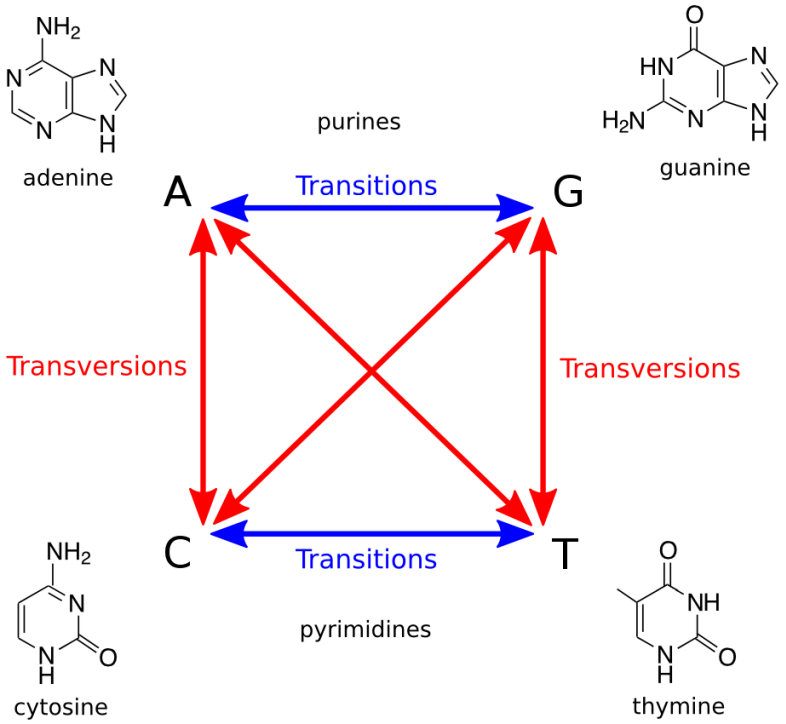
Numero di sostituzioni che ci sono tra le due sequenze (numero di caratteri diversi)

Considerare il numero totale di siti analizzati

Pesare diversamente transizioni e transversioni (mutazioni tra strutture chimiche diverse: purine (A,G) e pirimidine (C,T))

Valutare la frequenza dei nucleotidi

Considerare o meno la frequenza con cui abbiamo transversioni rispetto a transizioni



Distanze evolutive
modelli semplici

Distanze semplici:

Number of differences:

numero di siti dove le due sequenze differiscono

P-distance

Percentuale di siti nucleotidici dove le due sequenze sono differenti

Nessuna assunzione, solo normalizza sulla lunghezza

Modelli più complessi

Distanze più complesse: assumono un modello di sostituzione
(cioè un modello che mi dice quanto pesare una sostituzione)

Jukes-Cantor
1 parametro, quanto peso dare ad una sostituzione

Modelli più complessi

Distanza Tajima-Nei

Pesa in modo diverso le sostituzioni tenendo conto della frequenza che i nucleotidi hanno all'interno delle sequenze

	A	T	C	G
A	-	αg_T	αg_C	αg_G
T	αg_A	-	αg_C	αg_G
C	αg_A	αg_T	-	αg_G
G	αg_A	αg_T	αg_C	-

Distanza di Kimura

Pesa in modo diverso le transversioni dalle transizioni

	A	T	C	G
A	-	β	β	α
T	β	-	α	β
C	β	α	-	β
G	α	β	β	-

Modelli più complessi

Anche per le sequenze proteiche vi sono modelli di sostituzione per calcolare le distanze simili a quelli usati per il confronto di due sequenze aminoacidiche

NOTA

Cambiando il metodo il modello di sostituzione l'albero risultante potrebbe essere piuttosto diverso!

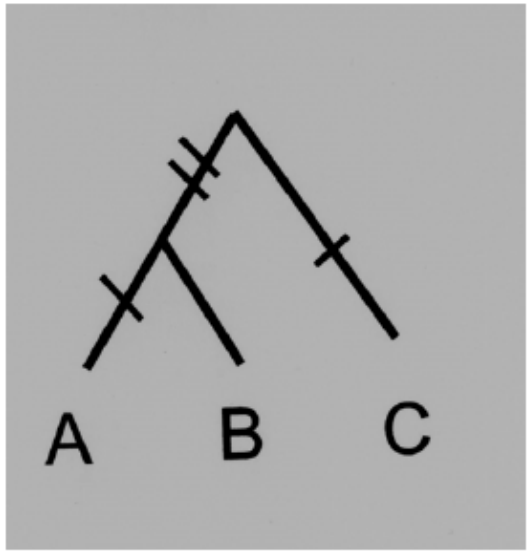
La scelta deve avvenire considerando anche le informazioni a priori

Ci sono molte altri modelli di distanze, introdotti negli ultimi anni, che tengono conto di altri fattori

ESEMPIO: Contenuto in GC

Matrici di distanza

Ogni albero filogenetico rispecchia una matrice di distanze fra coppie di sequenze



Albero

	A	B	C
A	0		
B	1	0	
C	4	3	0

Matrice di distanze

Metodi basati sulle distanze

Vantaggi

Veloce: va bene per analizzare grandi data set

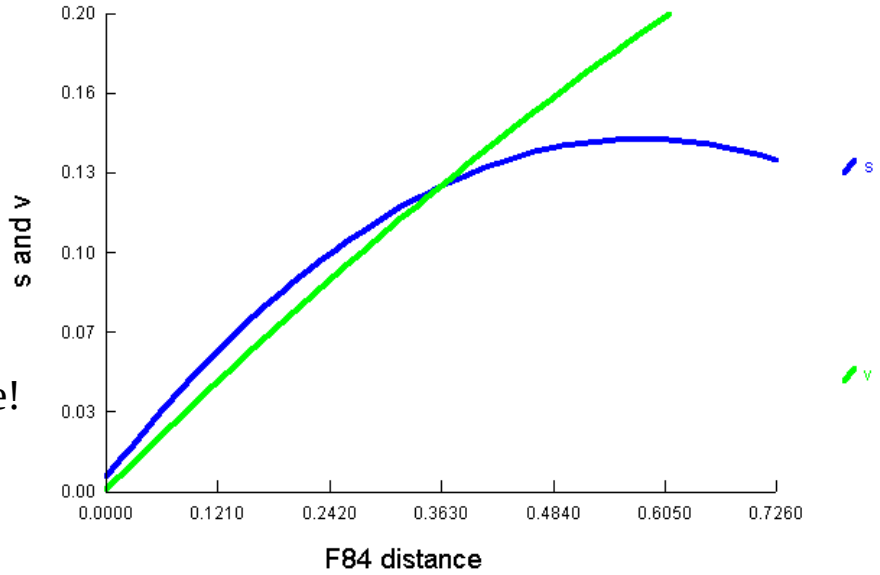
Basta avere le matrici di distanza

Svantaggi

Netta perdita di informazione:

dalle distanze non si torna indietro alle sequenze!

Problemi con misure di distanza non lineari con il tempo



Mutazioni ricorrenti, sottostima della distanza e saturazione

Anche assumendo che l'accumulo di mutazioni sia proporzionale al tempo che passa, non posso osservare direttamente questo numero ma il numero di differenze tra sequenze

Il **numero di differenze**, a causa delle mutazioni ricorrenti (mutazioni che si verificano più volte allo stesso sito nucleotidico) è **spesso inferiore al numero di mutazioni**

Servono modelli di correzioni alle misure di distanza (es. Jukes-Cantor)

In alcuni casi, l'eccessivo numero di mutazioni **satura l'informazione**

Clustering analysis (UPGMA)

Classico algoritmo agglomerativo gerarchico

La distanza tra cluster è definita come la media delle distanze di tutte le possibili coppie formate

Questo metodo produce alberi radicati e ultametrici (cioè, due di tre coppie di distanze fra tre taxa sono uguali e più piccole della terza, ciò implica l'assunzione che i tassi di evoluzione non possono variare tra i taxa e lungo le linee).

Da un punto di vista biologico è piuttosto povero: assume un tasso di evoluzione (sostituzione) costante

Clustering analysis (UPGMA)

Si costruisce la tabella delle distanze.

Si sceglie la **coppia di taxa più simili** (cioè i due taxa che sono separati dalla distanza più piccola).

Questi vengono uniti a formare una nuova entità.

Vengono ricalcolate le distanze genetiche con gli altri taxa, come media delle distanze fra le coppie originali.

Di nuovo viene unita insieme la coppia di taxa più simili.

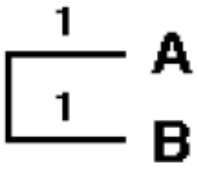
Si ricalcolano le distanze genetiche.

Questo processo viene ripetuto fino a che non sono stati aggiunti all'albero tutti i taxa.

Trovare l'albero a partire dalla matrice delle distanze (UPGMA)

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

Unisco i taxa con distanza minore
stimo le distanze dal nodo
calcolo le distanze delle specie rimanenti dal gruppo appena formato
modifico la matrice

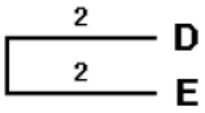


dist(A,B),C = (distAC + distBC) / 2 = 4
dist(A,B),D = (distAD + distBD) / 2 = 6
dist(A,B),E = (distAE + distBE) / 2 = 6
dist(A,B),F = (distAF + distBF) / 2 = 8

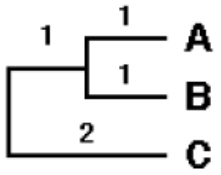
	A,B	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

Procedo iterativamente nello stesso modo

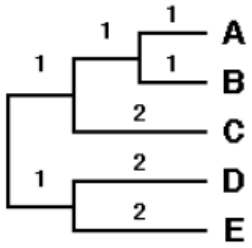
Trovare l’albero a partire dalla matrice delle distanze (UPGMA)



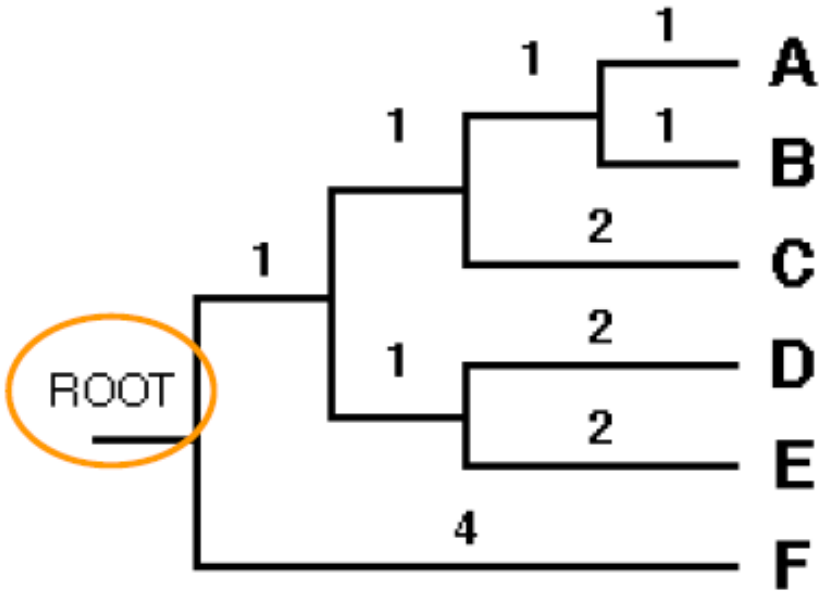
	A,B	C	D,E
C	4		
D,E	6	6	
F	8	8	8



	AB,C	D,E
D,E	6	
F	8	8



	ABC,DE
F	8



Clustering analysis (Neighbour joining)

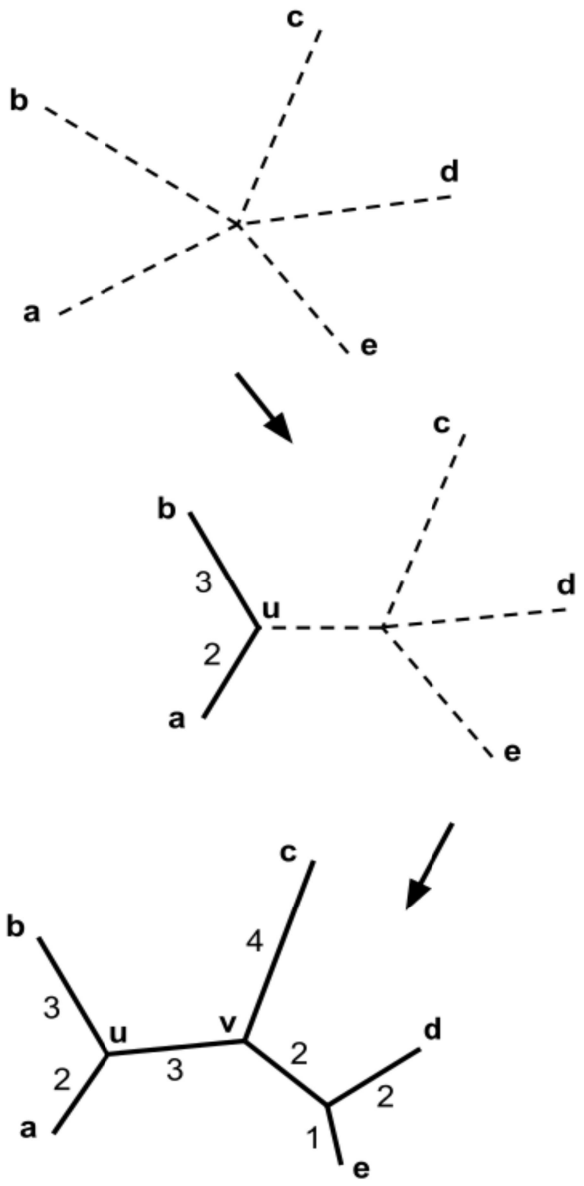
È un metodo che produce alberi non radicati, seguendo la proprietà additiva

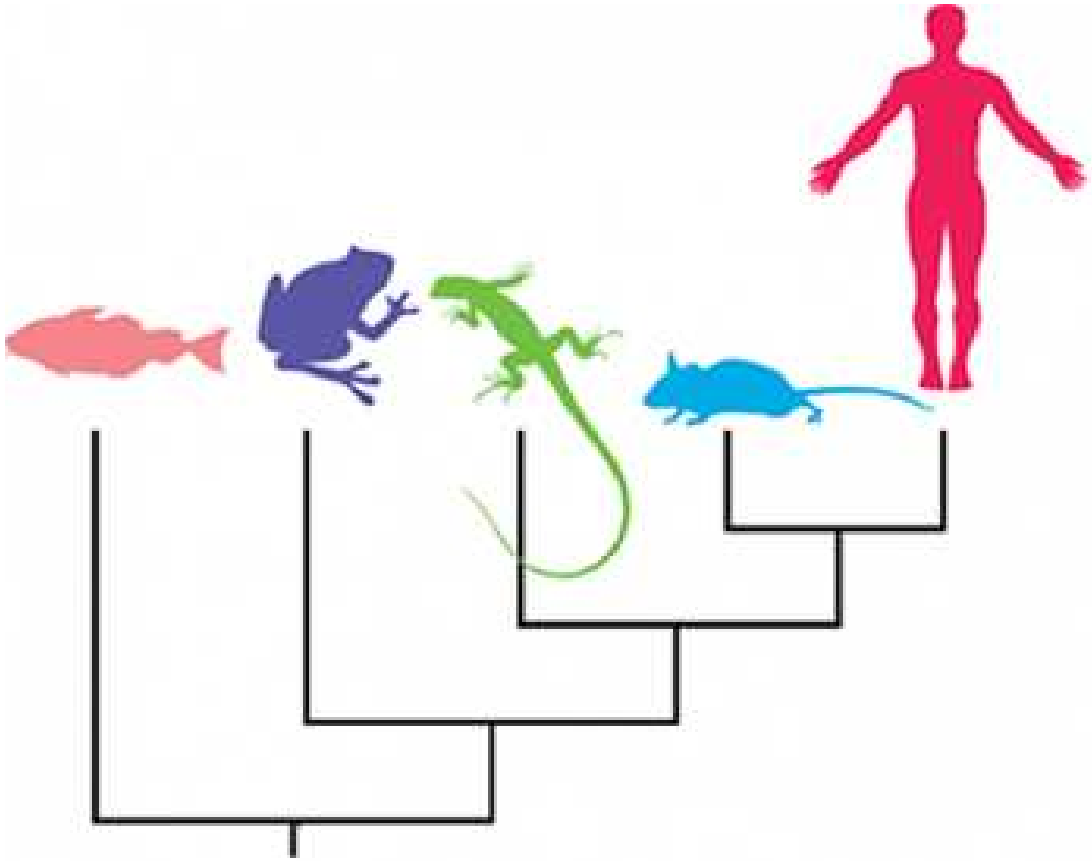
Due taxa si dicono vicini (neighbors) se sono connessi attraverso un solo nodo interno.

Il metodo NJ inizia con un set di nodi terminali non connessi che rappresentano le sequenze da analizzare. Sulla base delle distanze genetiche note, il metodo sceglie due nodi vicini **a** e **b** e li connette attraverso un nuovo nodo interno **u**.

I nodi terminali originali a e b vengono eliminati, in quanto già connessi.

Il processo viene ripetuto fino a che tutti i nodi terminali non vengono connessi in un singolo albero.





Metodo della Massima Parsimonia

Vantaggi

Metodo molto semplice.

Fornisce insieme l'albero e le ipotesi di evoluzione dei caratteri

Sembra che funzioni abbastanza bene se l'omoplasia è rara o comunque se è distribuita casualmente sui diversi rami

Svantaggi

Se l'omoplasia è frequente e non distribuita in maniera omogenea nell'albero produce false filogenesi

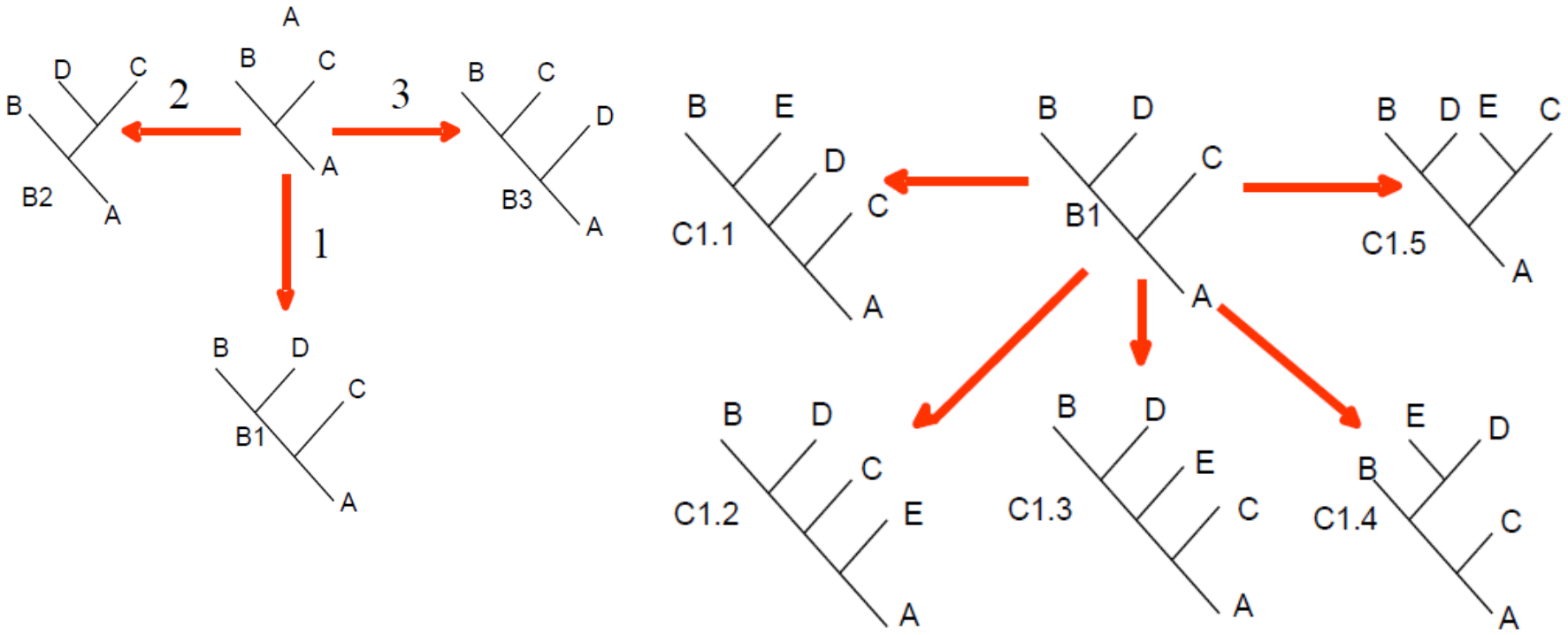
Sottostima la lunghezza dei rami

Il modello di evoluzione è implicito, difficile studiare l'esatto funzionamento in diverse condizioni

Fa affidamento sull'assunto che le **mutazioni più parsimoniose** sono quelle più probabili

Algoritmo di Massima Parsimonia

Si parte da 3 taxa, si fanno tutti i possibili alberi e si calcola mediante una formula la parsimonia. L'albero con punteggio migliore viene utilizzato per costruire gli alberi con 4 taxa… e così di seguito.



Massima Parsimonia

Quando dobbiamo costruire alberi con molti taxa diventa impossibile utilizzare la ricerca esaustiva.

Devo per forza ricorrere a programmi che utilizzano algoritmi euristici: ad ogni fase si analizza solo un campione del set di dati ottenuti.

Alla fine avremo una buona soluzione ma, non avendo esaminato tutte le soluzioni possibili, non sapremo mai se poteva esserci una soluzione migliore!

Number of trees increases with number of taxa

- 4 taxa → 3 unrooted trees
- 8 taxa → 10,395 unrooted trees
- 10 taxa → 2,027,025 unrooted trees
- 22 taxa → 3×10^{23} (almost a mole)
- 50 taxa → 3×10^{74} (More trees than the number of atoms in the universe)
- "Exhaustive searches" of tree topologies are nearly impossible with modern data sets
- **Algorithms attempt to find best tree anyway**

Barry Hall, Phylogenetic Trees Made Easy

Maximum likelihood

Ricerca la topologia dell'albero e la relativa lunghezza dei rami che massimizzano la verosimiglianza dei dati osservati con il modello considerato

Il modello di evoluzione deve essere accuratamente stabilito prima di iniziare a costruire l'albero (per esempio, rapporto tra transizioni e trasversioni, distribuzione dei tassi di mutazione lungo la sequenza, ecc.)

Il programma costruisce tutti i possibili alberi ottenibili con i taxa a disposizione

Per ognuno calcola un punteggio di massima verosimiglianza, secondo un complicato modello statistico e secondo i parametri evolutivi che noi abbiamo scelto.

L'albero migliore è quello che ha il più alto punteggio, corrispondente alla più alta probabilità di essere generato sulla base del modello e dei dati.

Le operazioni di scelta del modello evolutivo e di altri parametri sono eseguite mediante altri software che sono presenti nel pacchetto per l'analisi.

Maximum Likelihood

Utilizza modelli molto realistici:

I siti evolvono indipendentemente l'uno dall'altro

I siti possono seguire processi di sostituzione differenti (siti sinonimi vs siti non sinonimi; transizioni verso trasversioni)

Le probabilità di sostituzione possono variare tra i rami

Dal punto di vista teorico è il miglior metodo

Esperimenti di simulazione su sequenze hanno dimostrato che questo è il metodo che, nella maggior parte dei casi, lavora meglio

Uno degli svantaggi è rappresentato dal fatto che è molto complesso da eseguire, richiedendo molto tempo di calcolo al computer

E' quasi sempre impossibile valutare tutti i possibili alberi. Viene in genere fatta una esplorazione parziale dello spazio dei possibili alberi.

Di conseguenza non vi è la certezza matematica di ottenere l'albero più affidabile.

Svantaggi

Risultati discutibili se il modello evolutivo scelto è scorretto

Molto complesso e lento dal punto di vista computazionale

Metodi Bayesiani

Strettamente connessi ai metodi di Maximum Likelihood

Si basano su un modello probabilistico che spiega come i dati osservati sono stati prodotti

Ogni parametro del modello ha un valore di probabilità

Confrontano la *probabilità a priori* del modello (stabilito prima di analizzare i dati) con la *probabilità a posteriori* (ovvero la probabilità che il valore di un parametro sia uguale alla probabilità dell'osservazione, dato quel valore di parametro)

La probabilità bayesiana

Probabilità inversa

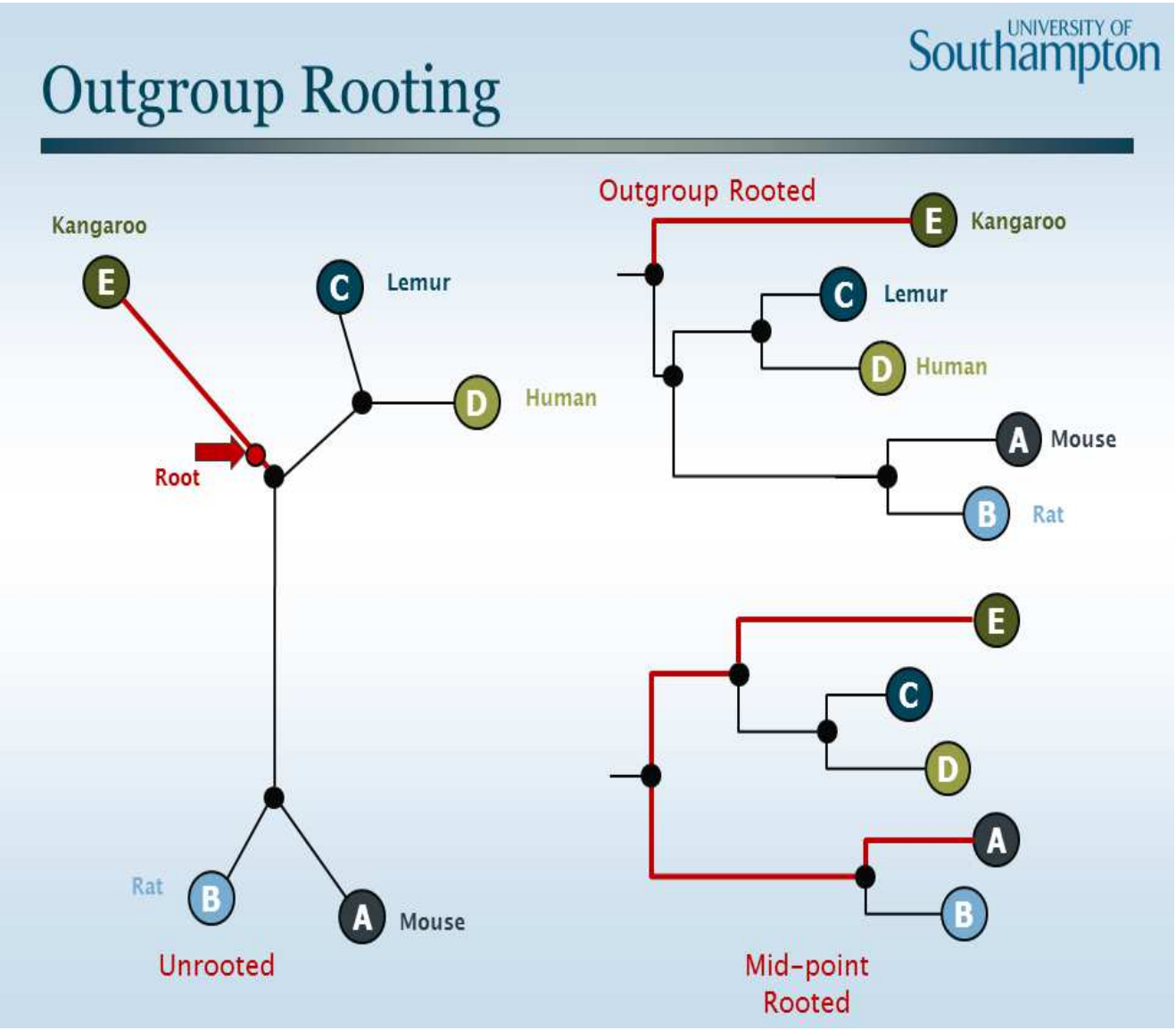
consiste nel risalire dalle frequenze osservate alla probabilità.

Nell'approccio bayesiano si utilizzano considerazioni “personali" per assegnare la probabilità ad un dato evento prima di fare l'esperimento.

La probabilità è quindi legata al grado di credibilità dell'evento stabilito da “persone informate".

Il teorema di Bayes consente in seguito, alla luce delle frequenze osservate, di “aggiustare" la probabilità a priori (ossia precedentemente assegnata in base all'interpretazione soggettiva) per arrivare alla probabilità a posteriori.

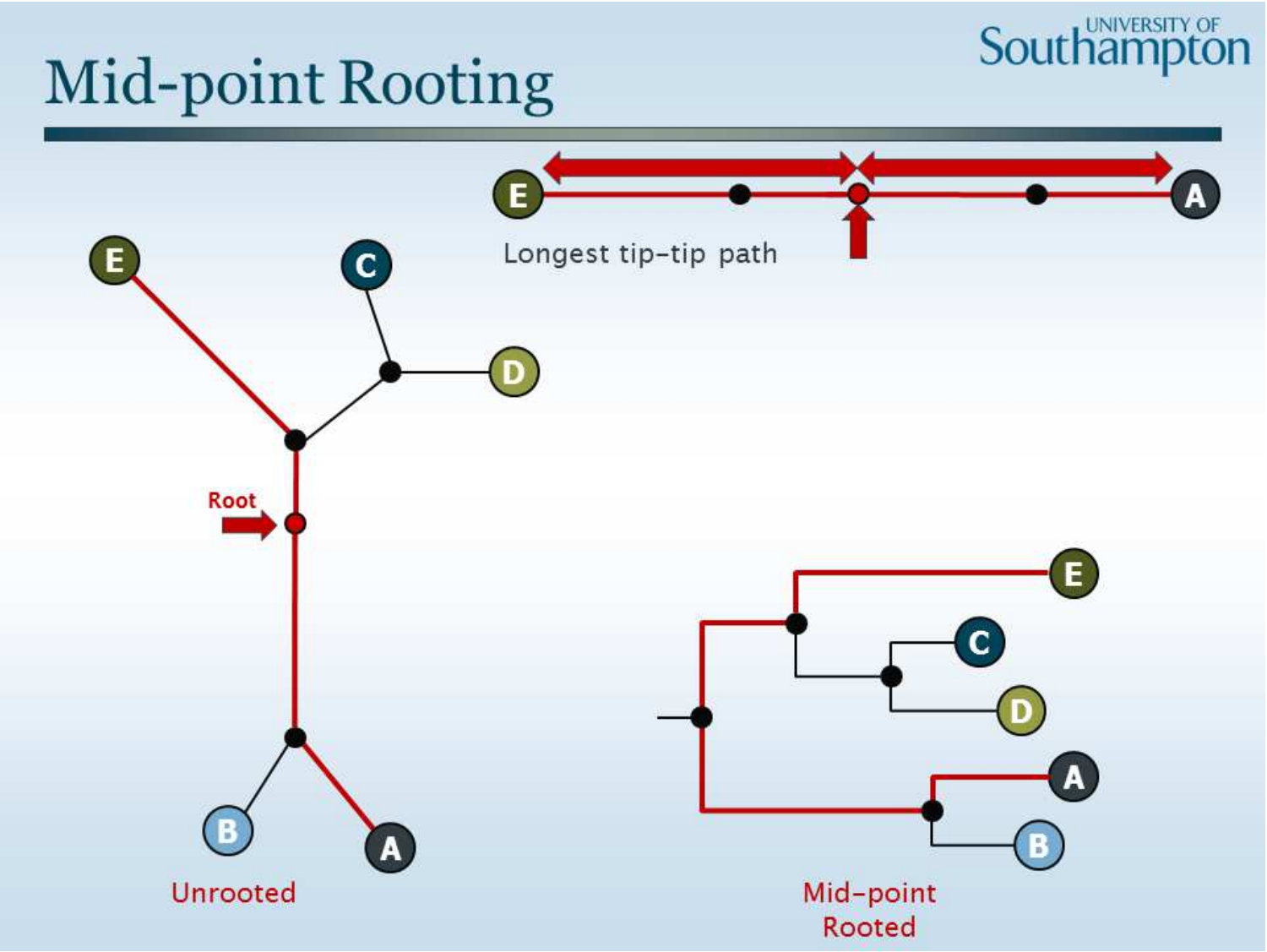
Quindi, tramite tale approccio, si usa una stima del grado di credibilità di una data ipotesi prima dell'osservazione dei dati, al fine di associare un valore numerico al grado di credibilità di quella stessa ipotesi successivamente all'osservazione dei dati.



Metodo dell'Outgroup: utilizzo un gruppo "esterno" alla filogenesi che sto analizzando

Assumo che l'outgroup si sia separato prima di tutti gli altri (devo fare un'ipotesi filogenetica esterna al gruppo che mi interessa)

La divergenza dell'outgroup non deve essere né troppo piccola né troppo grande



Metodo del Mid-point: a metà del ramo più lungo che unisce due OTU

Assume tassi di evoluzione approssimativamente costanti

Se l'assunzione non è verificata, commette errori

Il bootstap per testare la robustezza di un albero

Tecnica di randomizzazione: la confidenza si calcola ricampionando i dati disponibili

I caratteri (colonne in un allineamento di sequenze) sono estratte con rimpiazzo per generare molti (almeno 1000) pseudo data set

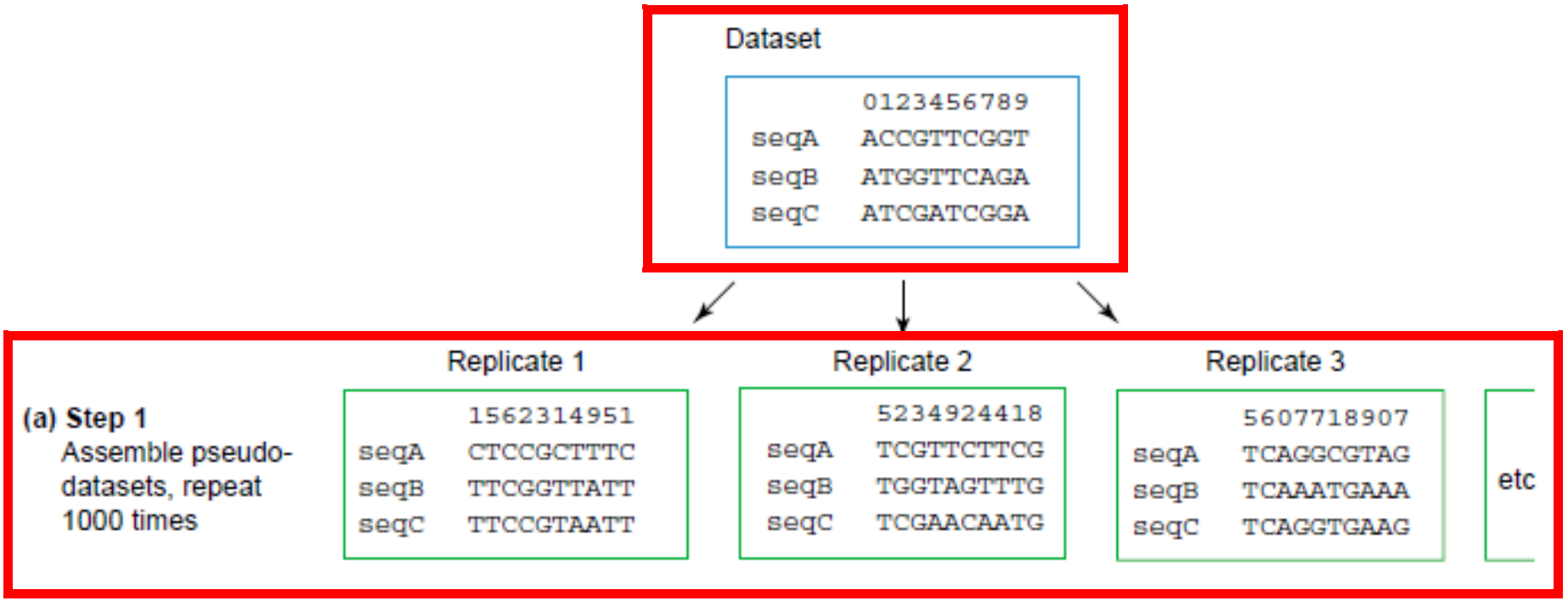
Ogni pseudo data set viene analizzato per ricostruire una filogenesi (con uno dei metodi visti)

La frequenza con cui i diversi gruppi si ritrovano nell'albero di consenso così costruito (le bootstrap proportions) sono una misura del supporto statistico per quel gruppo

Bootstrap

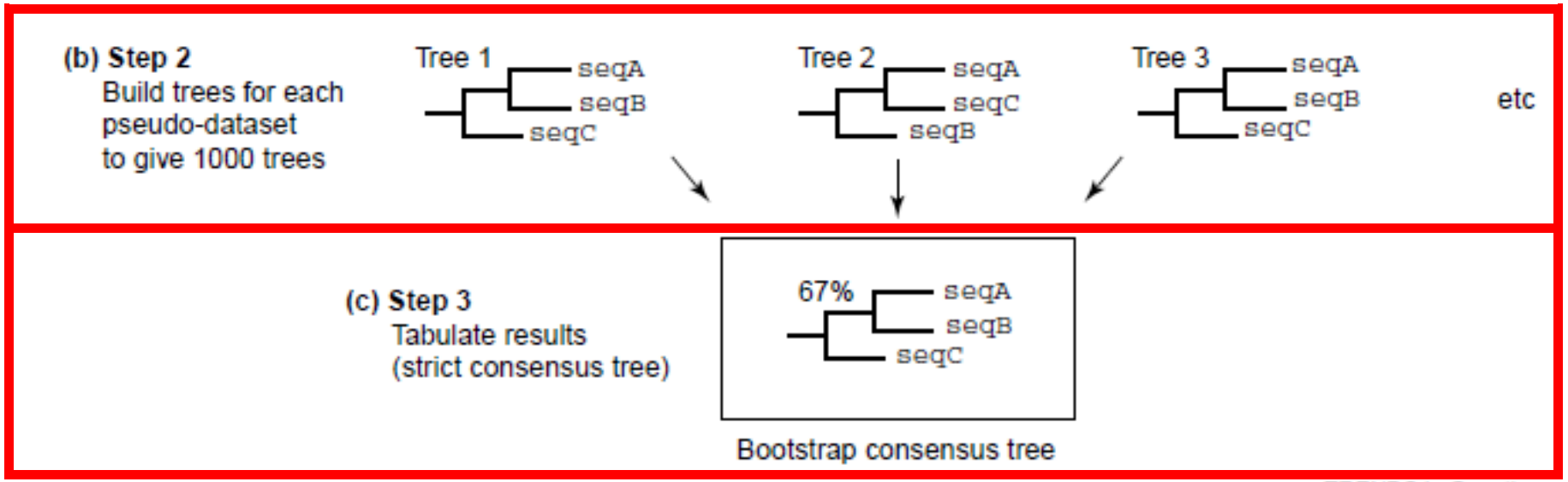
Si fanno dei **sottocampioni casuali** del set di dati

Ciascuno dei sottocampioni è della **stessa dimensione dell'originale**.



Bootstrap

Si costruiscono gli alberi e si **valuta la frequenza** con cui i rami dell'albero in esame sono presenti in ciascuno di questi sottocampioni casuali.

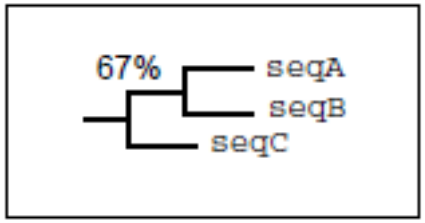


Bootstrap

Se il ramo X è presente in ogni albero sottocampione, il suo valore di bootstrap sarà del **100 %**

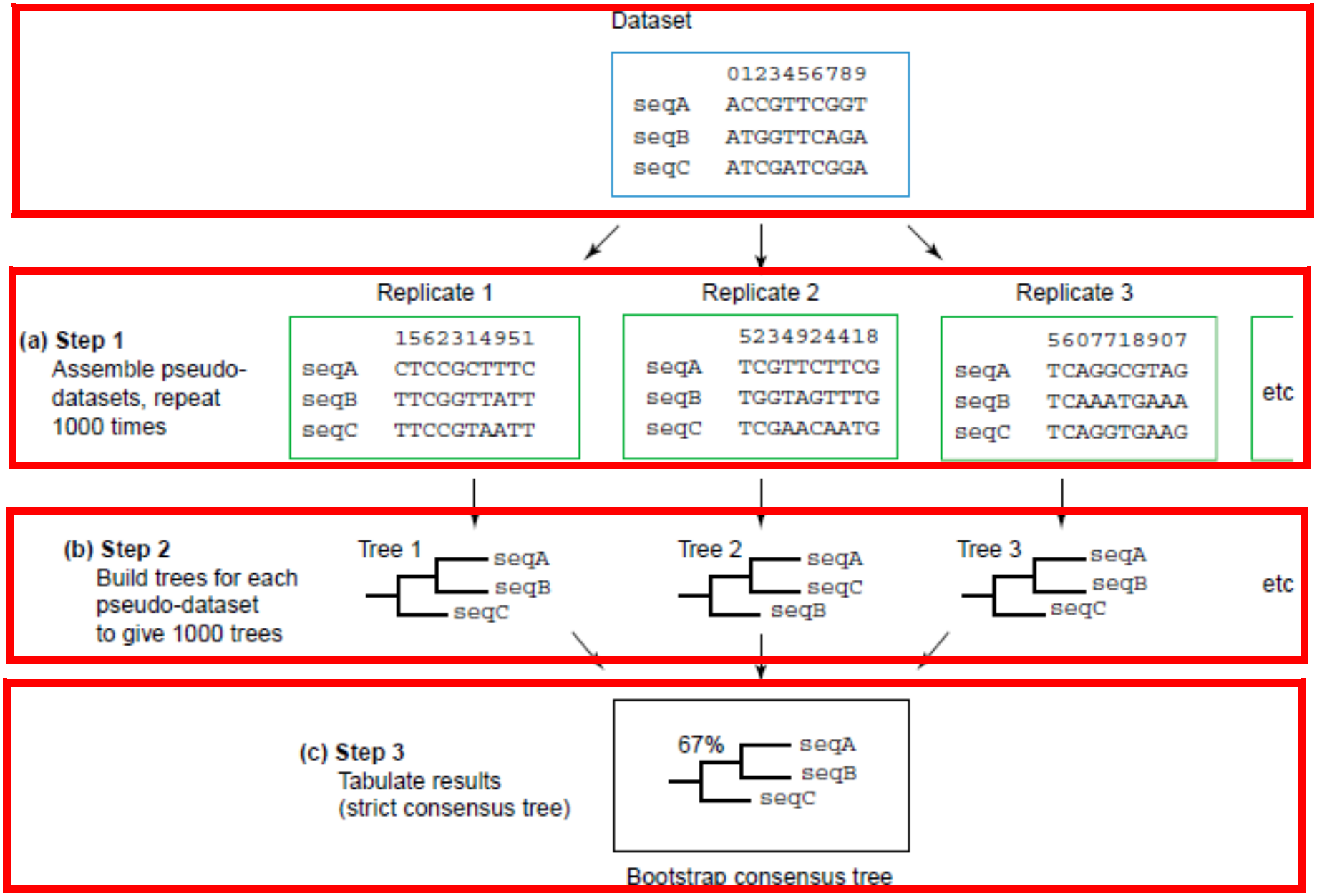
Se è presente solo nei due terzi degli alberi sottocampione, il suo bootstrap sarà del **67%.**

(c) Step 3
Tabulate results
(strict consensus tree)



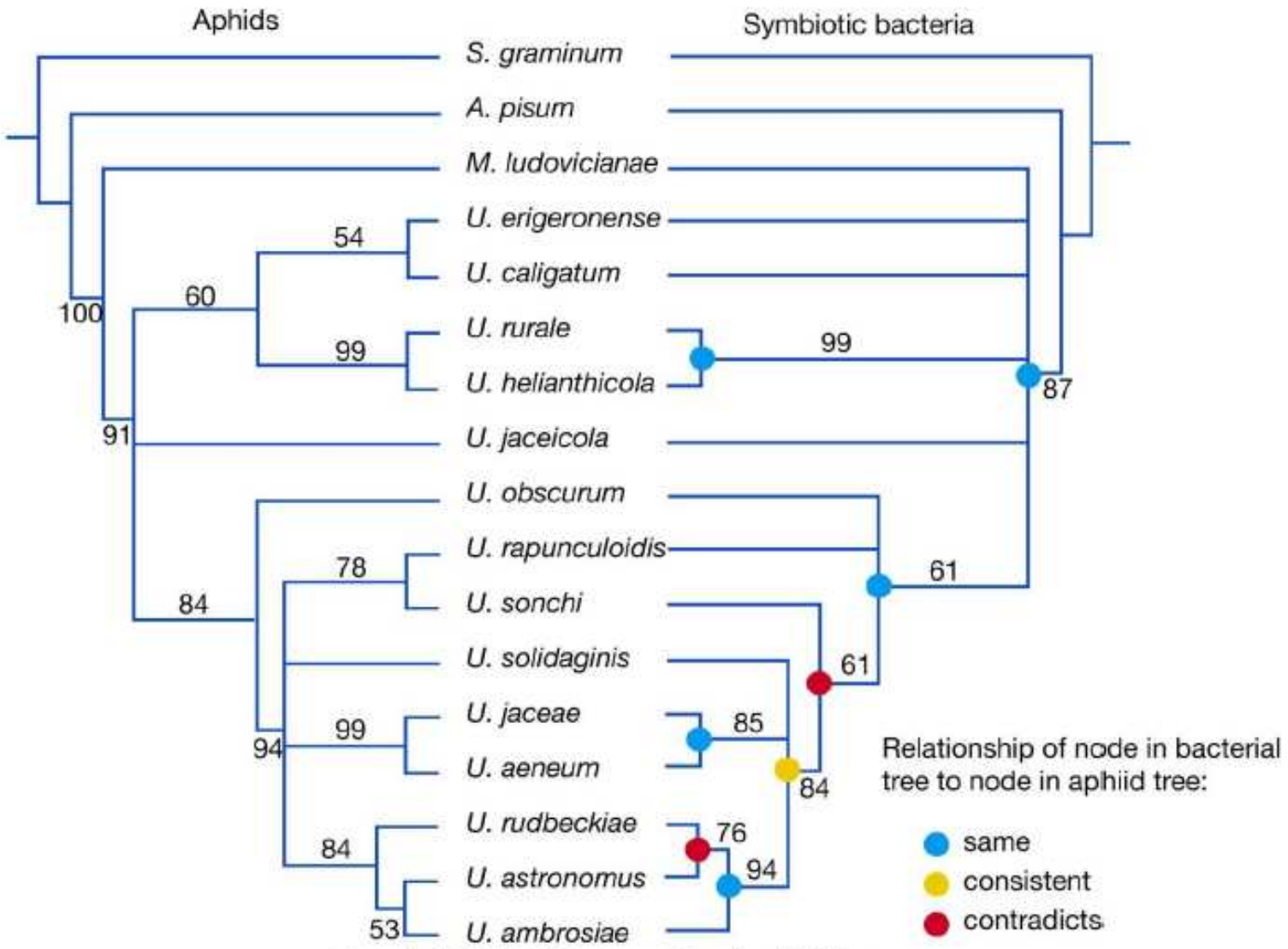
Bootstrap consensus tree

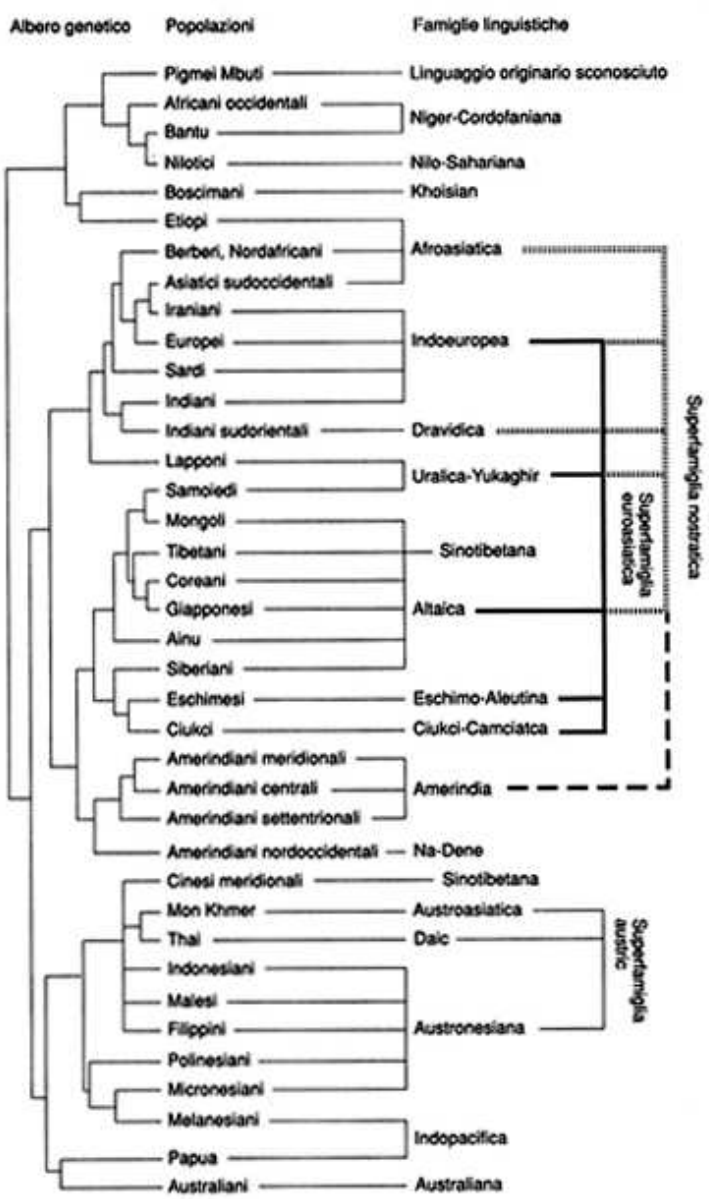
TRENDS in Genetics



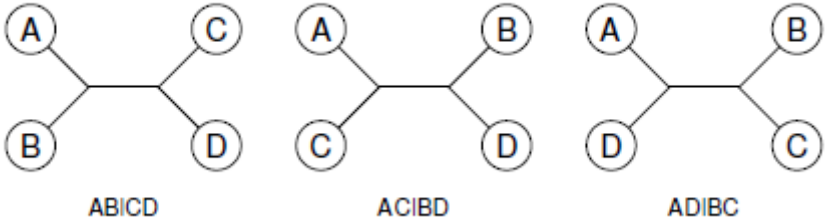
TRENDS in Genetics

Il bootstap per testare gruppi e cospeciazione




Albero genetico
Popolazioni
Famiglie linguistiche
Pigmei Mbuti
Linguaggio originario sconosciuto
Africani occidentali
Bantu
Niger-Cordofaniana
Nilotici
Nilo-Sahariana
Boscimani
Khoisian
Etiopi
Berberi, Nordafricani
Asiatici sudoccidentali
Afroasiatica
Iraniani
Europei
Sardi
Indiani
Indoeuropea
Indiani sudorientali
Dravidica
Lapponi
Samoiedi
Uralica-Yukaghir
Mongoli
Tibetani
Coreani
Giapponesi
Ainu
Siberiani
Sinotibetana
Altaica
Eschimesi
Ciukci
Eschimo-Aleutina
Ciukci-Camciatca
Amerindiani meridionali
Amerindiani centrali
Amerindiani settentrionali
Amerindia
Amerindiani nordoccidentali
Na-Dene
Cinesi meridionali
Sinotibetana
Mon Khmer
Thai
Austroasiatica
Daic
Indonesiani
Malesi
Filippini
Austronesiana
Polinesiani
Micronesiani
Melanesiani
Papua
Indopacifica
Australiani
Australiana
Superfamiglia eurasiatica
Superfamiglia nostratica
Superfamiglia austric

Likelihood mapping



Metodo grafico che si basa sui valori di Maximum Likelihood

Il metodo prende quattro sequenze a caso del nostro dataset

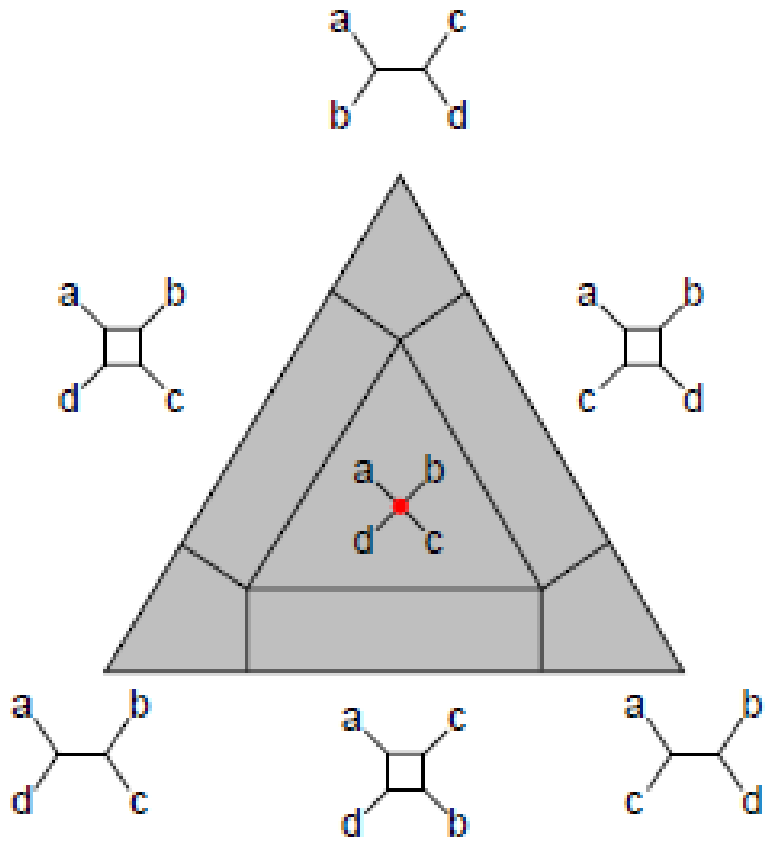
Con il metodo Maximum Likelihood trova tre alberi filogenetici diversi e ne calcola il punteggio

Se **uno** dei tre alberi ha un punteggio molto migliore degli altri due mette un puntino ad uno dei vertici del triangolo

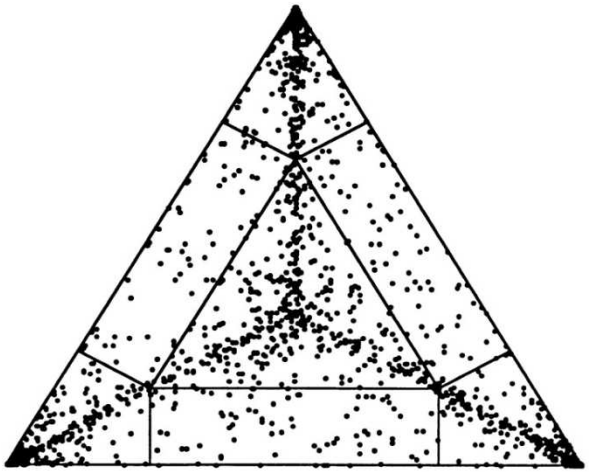
Se vi sono **due alberi con punteggio superiore** mette un puntino nelle aree laterali

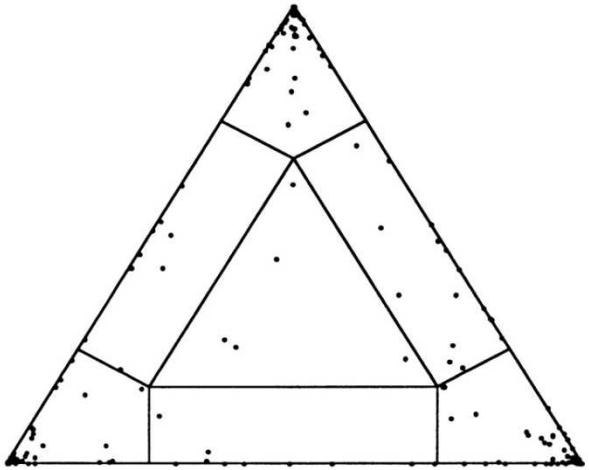
Se tutti e tre hanno un punteggio simile mette un puntino nell'aera centrale

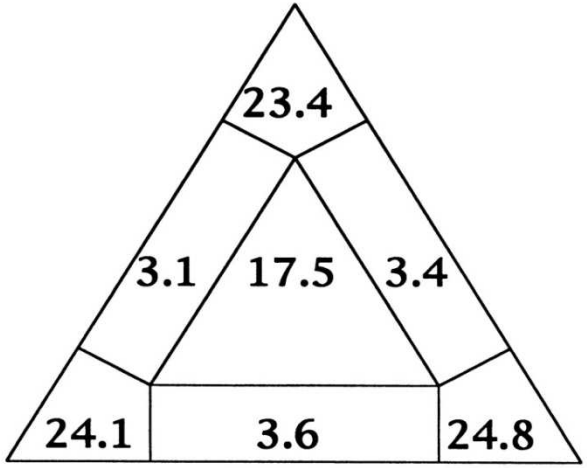
Il programma ripete l'operazione per migliaia di volte

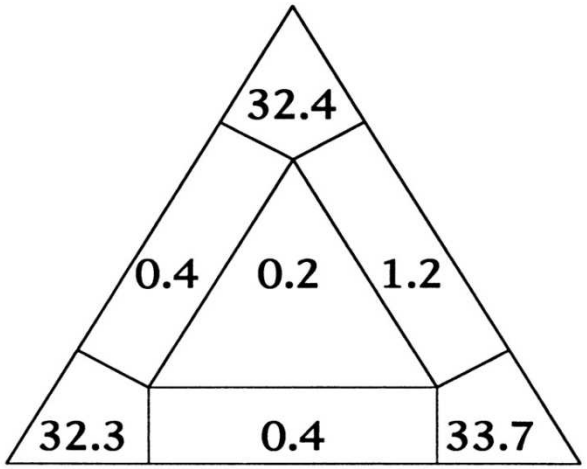


Likelihood mapping





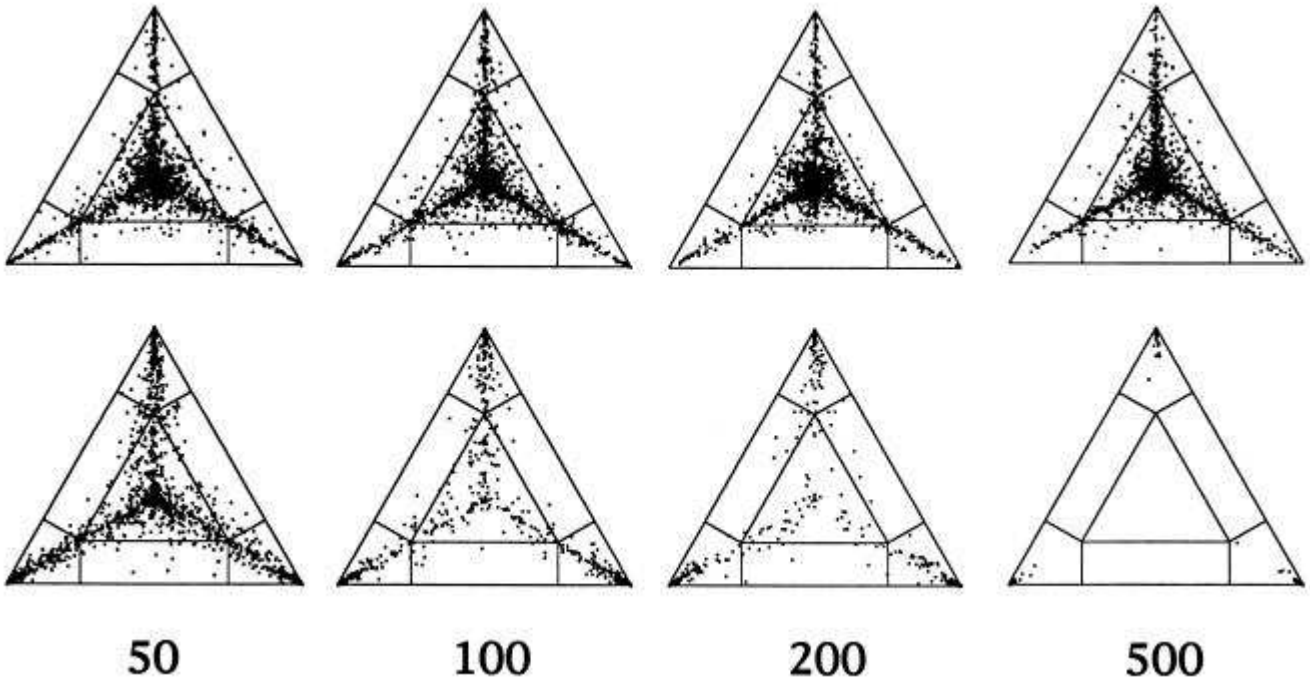




A

B

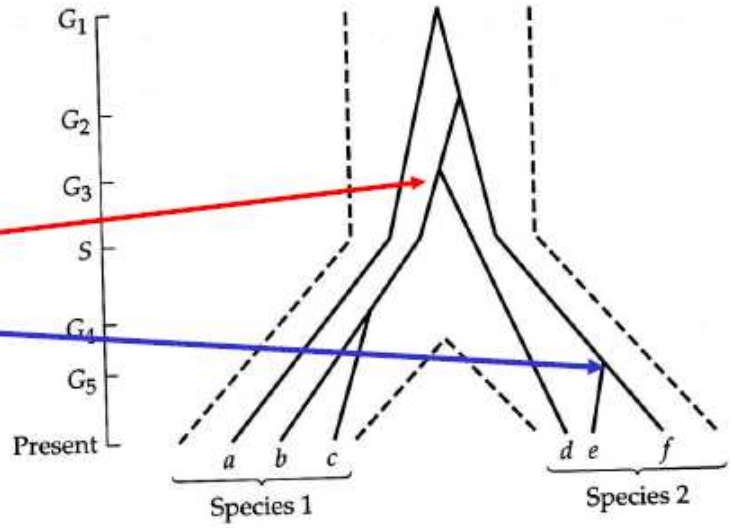
Lunghezza delle sequenze

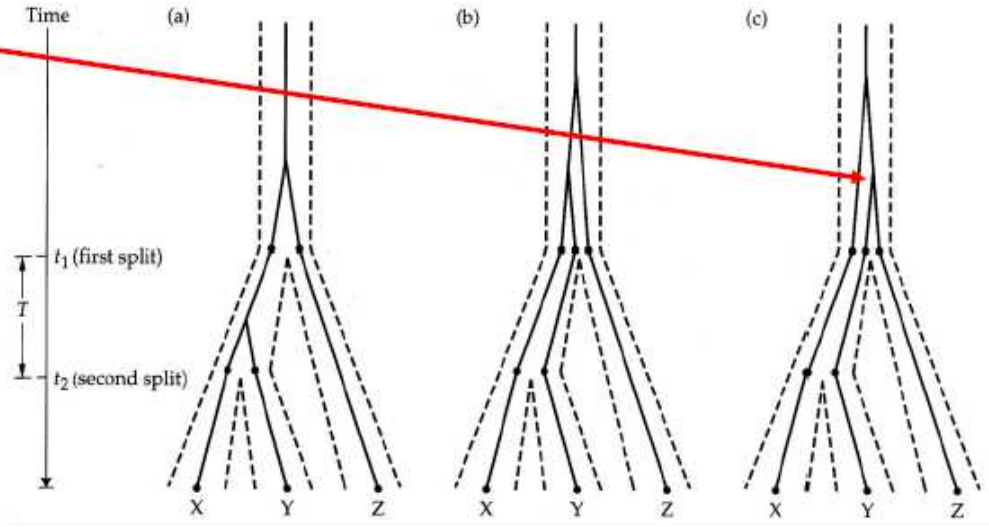


L’albero filogenetico ottenuto da un buon set di sequenze allineate migliora moltissimo con l’aumento della lunghezza!

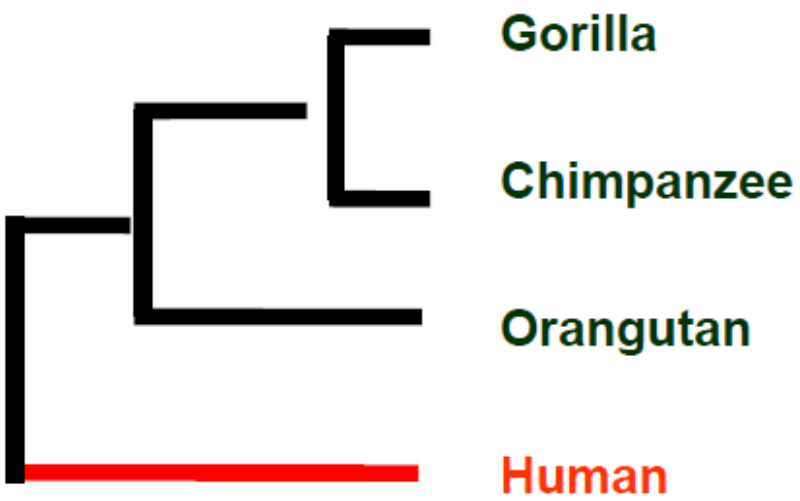
GENI vs SPECIE

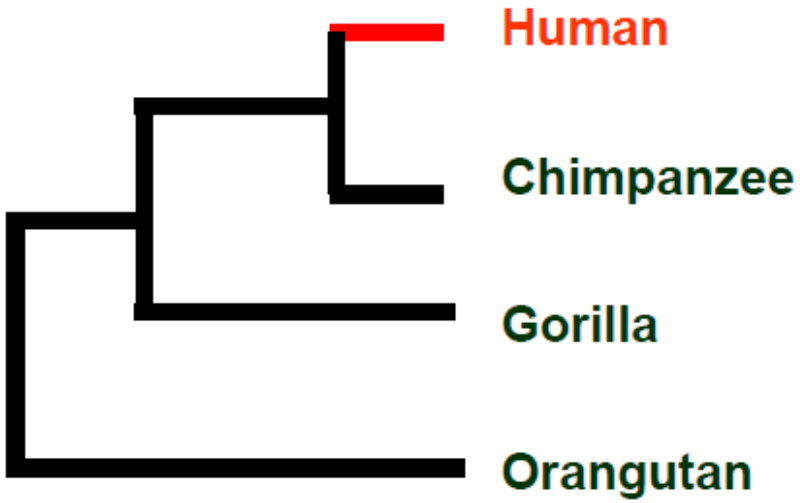
- Gli alberi costruiti con i geni non sempre coincidono con gli alberi delle specie
- La divergenza di due alleli può precedere o seguire la speciazione (cladogenesi-la separazione di due specie)
- Un allele può mostrare una ramificazione che è diversa da quella della specie
- Duplicazione genica
- Bisogna studiare più geni per ottenere un albero più affidabile





Un albero filogenetico è un'ipotesi tra tante possibili





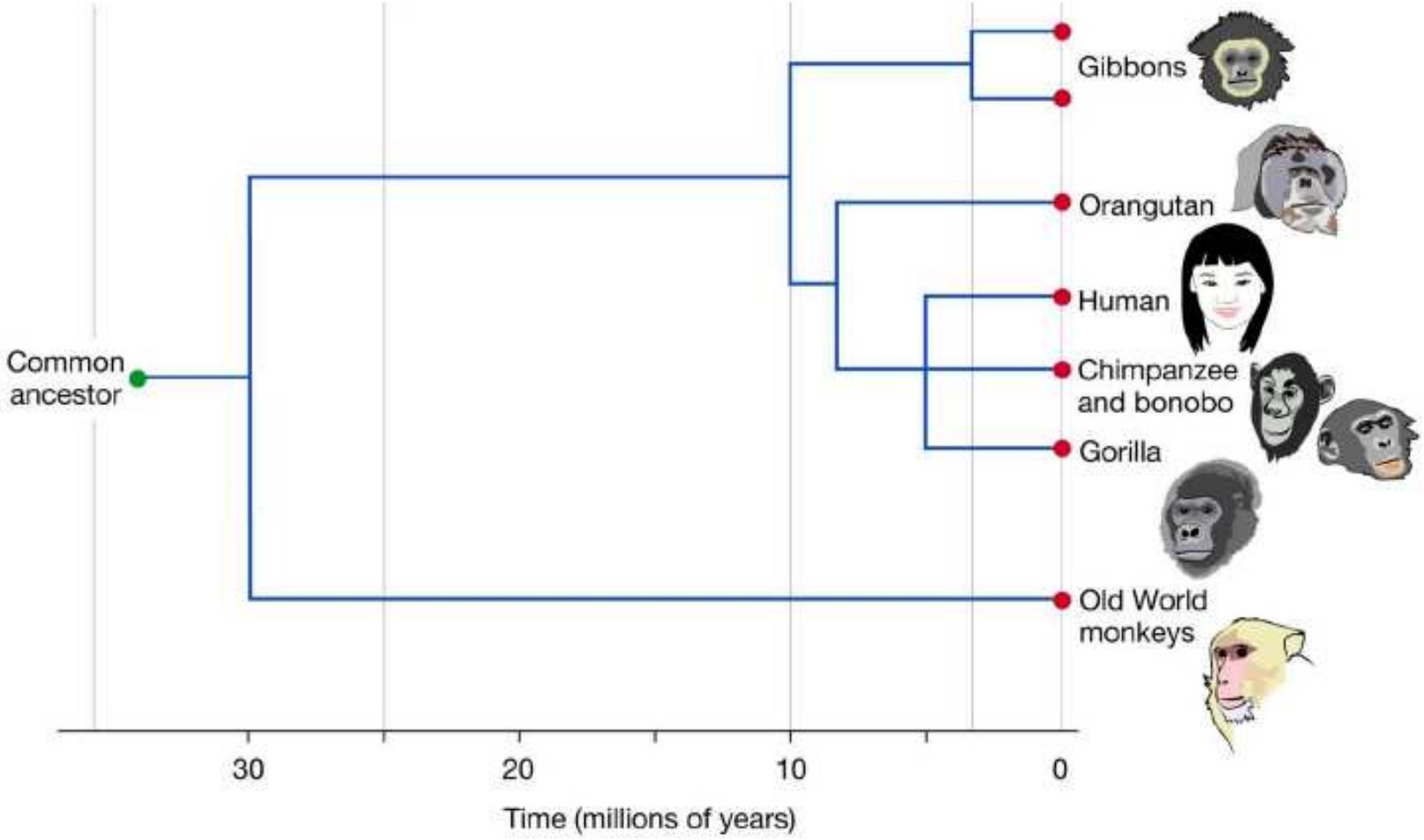
Analisi fossili (fino anni 60).

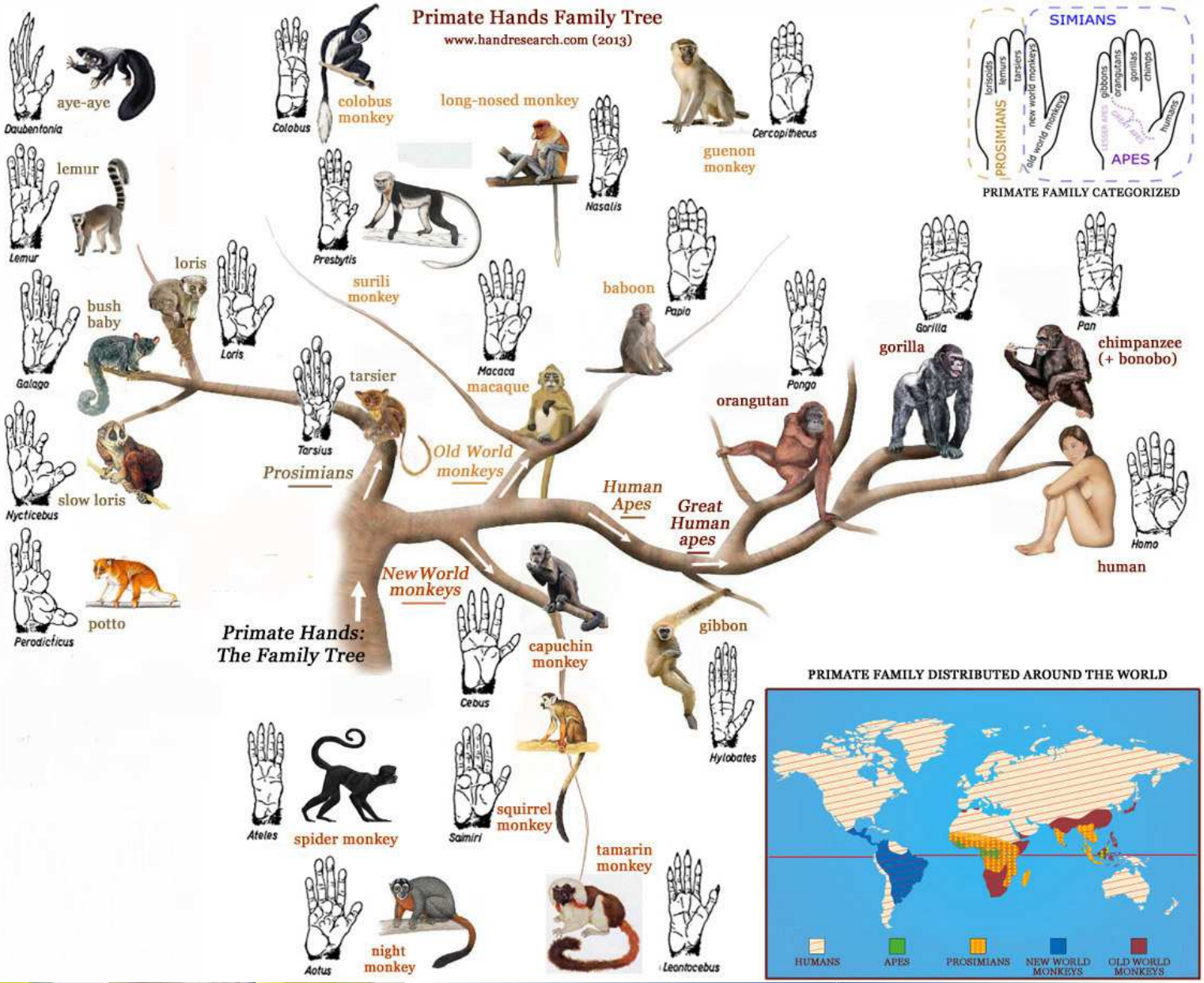
Grande differenza tra uomo e altri primati e separazione antica (>15MY)

Analisi molecolari.

Lo scimpanzè è più vicino all'uomo che non al gorilla (split a circa 5MY)

In realtà la tricotomia non è stata facile da risolvere




Primate Hands Family Tree
www.handresearch.com (2013)
aye-aye
Daubentonia
colobus monkey
Colobus
long-nosed monkey
Nasalis
guenon monkey
Cercopithecus
lemur
Lemur
Presbytis
surili monkey
loris
Loris
bush baby
Galago
tarsier
Tarsius
Macaca
macaque
baboon
Papio
Pongo
orangutan
Gorilla
gorilla
Pan
chimpanzee (+ bonobo)
Homo
human
slow loris
Nycticebus
potto
Perodicticus
Prosimians
Old World monkeys
Human Apes
Great Human apes
New World monkeys
Primate Hands: The Family Tree
capuchin monkey
Cebus
gibbon
Hylobates
Ateles
spider monkey
Saimiri
squirrel monkey
tamarin monkey
Aotus
night monkey
Leontocebus
SIMIANS
PROSIMIANS
APES
lorisoids
lemurs
tarsiers
new world monkeys
old world monkeys
gibbons
orangutans
gorillas
chimps
humans
PRIMATE FAMILY CATEGORIZED
PRIMATE FAMILY DISTRIBUTED AROUND THE WORLD
HUMANS
APES
PROSIMIANS
NEW WORLD MONKEYS
OLD WORLD MONKEYS