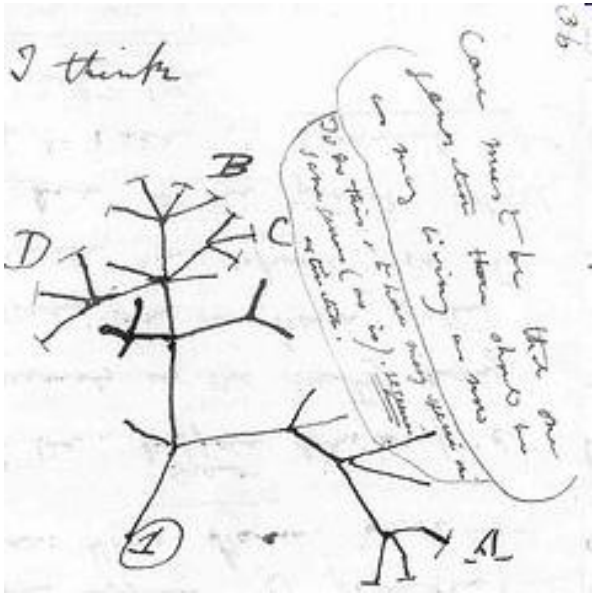
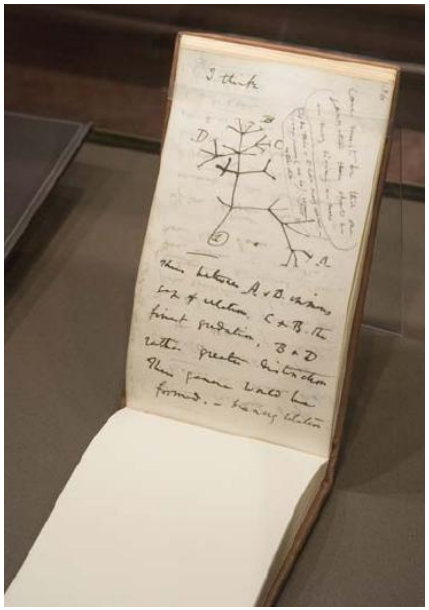
Filogenesi e alberi filogenetici





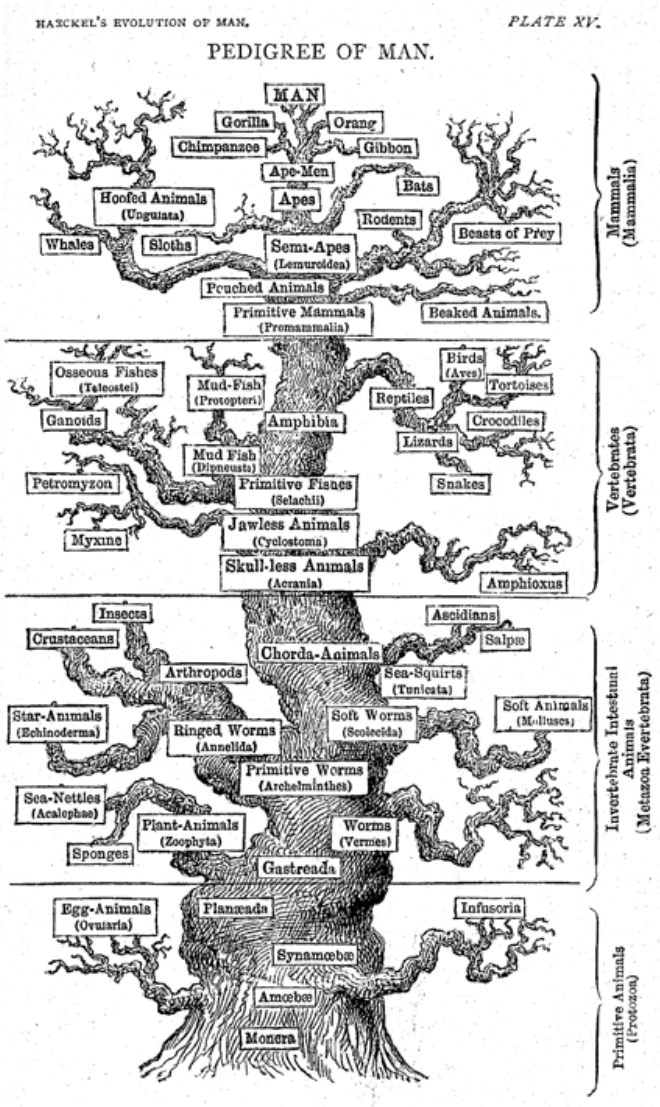
Darwin, 1837

"The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth... As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications." (from "The Origin")

Definizione di filogenesi

- La filogenesi è lo studio delle relazioni evolutive tra entità biologiche (non solo specie) che condividono antenati comuni
- La sua rappresentazione grafica è l'albero filogenetico
- L'albero filogenetico contiene i tempi e gli schemi temporali dei processi di divergenza.

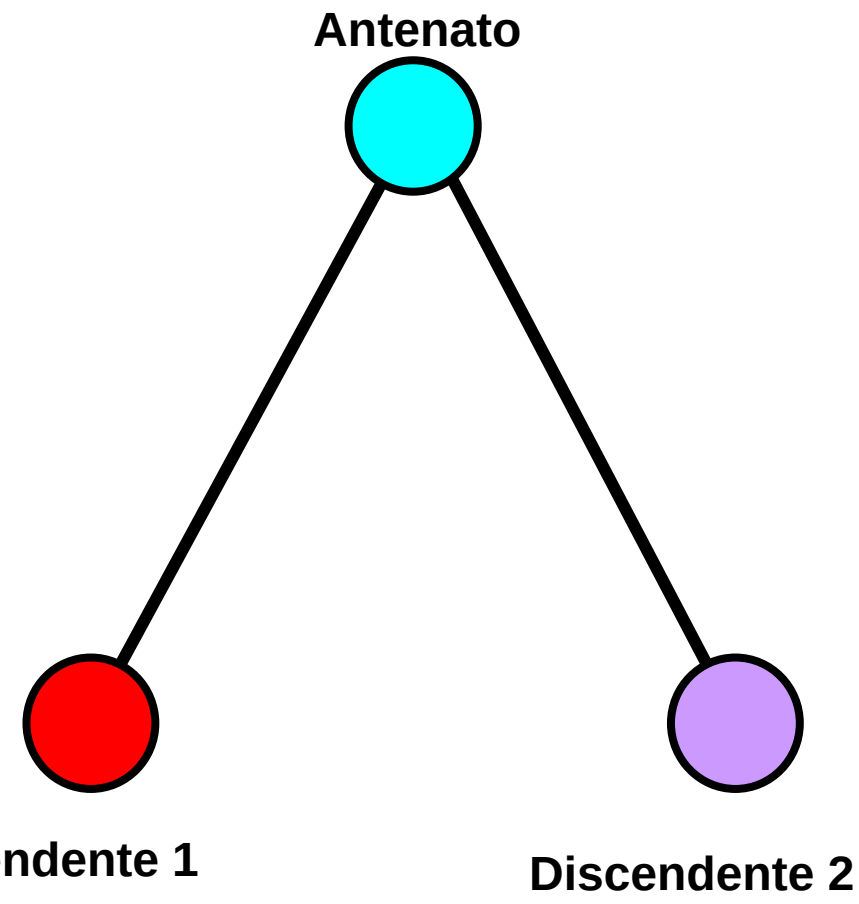
Un albero filogenetico (non di Darwin!) che ancora risente della *Scala Naturae*



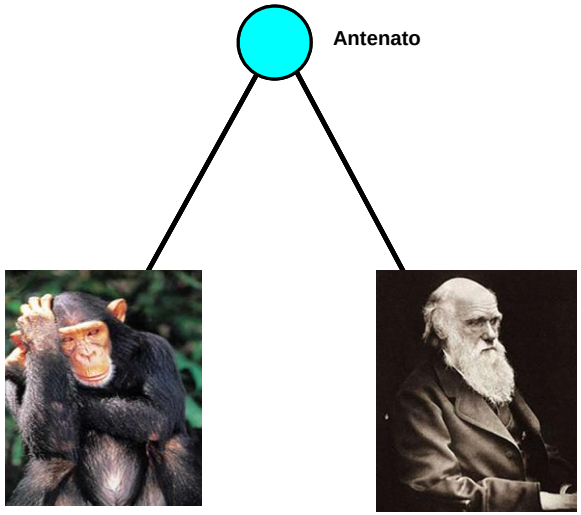
Ernst Haeckel (1834-1919)

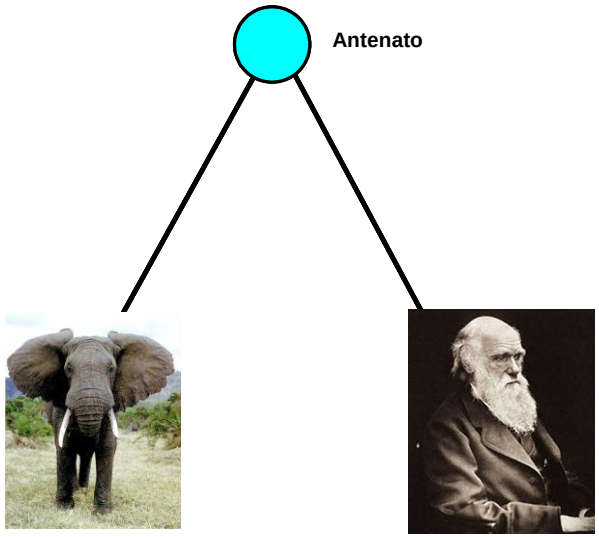
Logica alla base di un albero filogenetico

- Tutti gli organismi hanno un unico antenato comune nel passato
- Ogni coppia di organismi ha un antenato comune nel passato
- Eventi di speciazione si susseguono nel tempo creando nuove specie

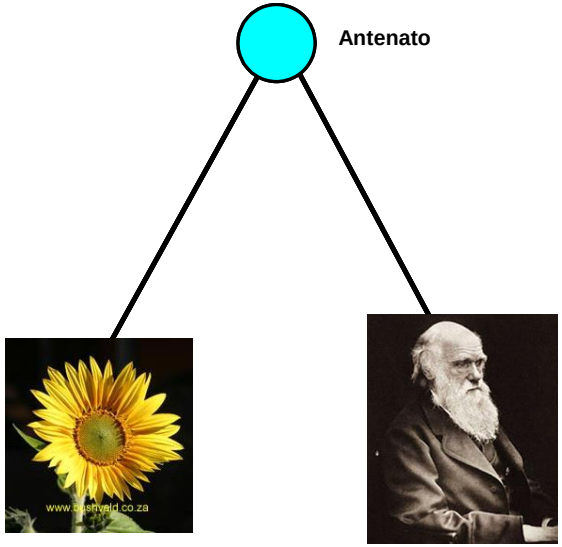


Ognuno di questi alberi è corretto

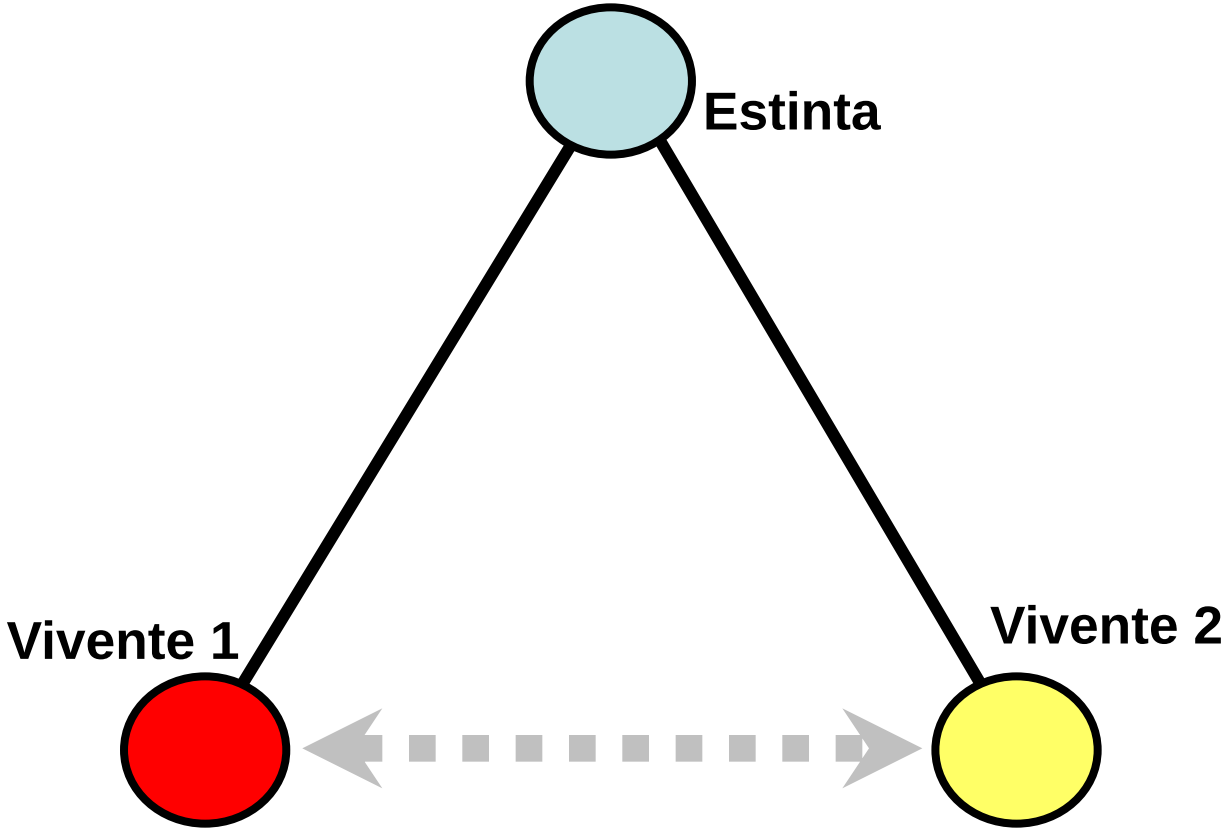




Ma qual'è la differenza?

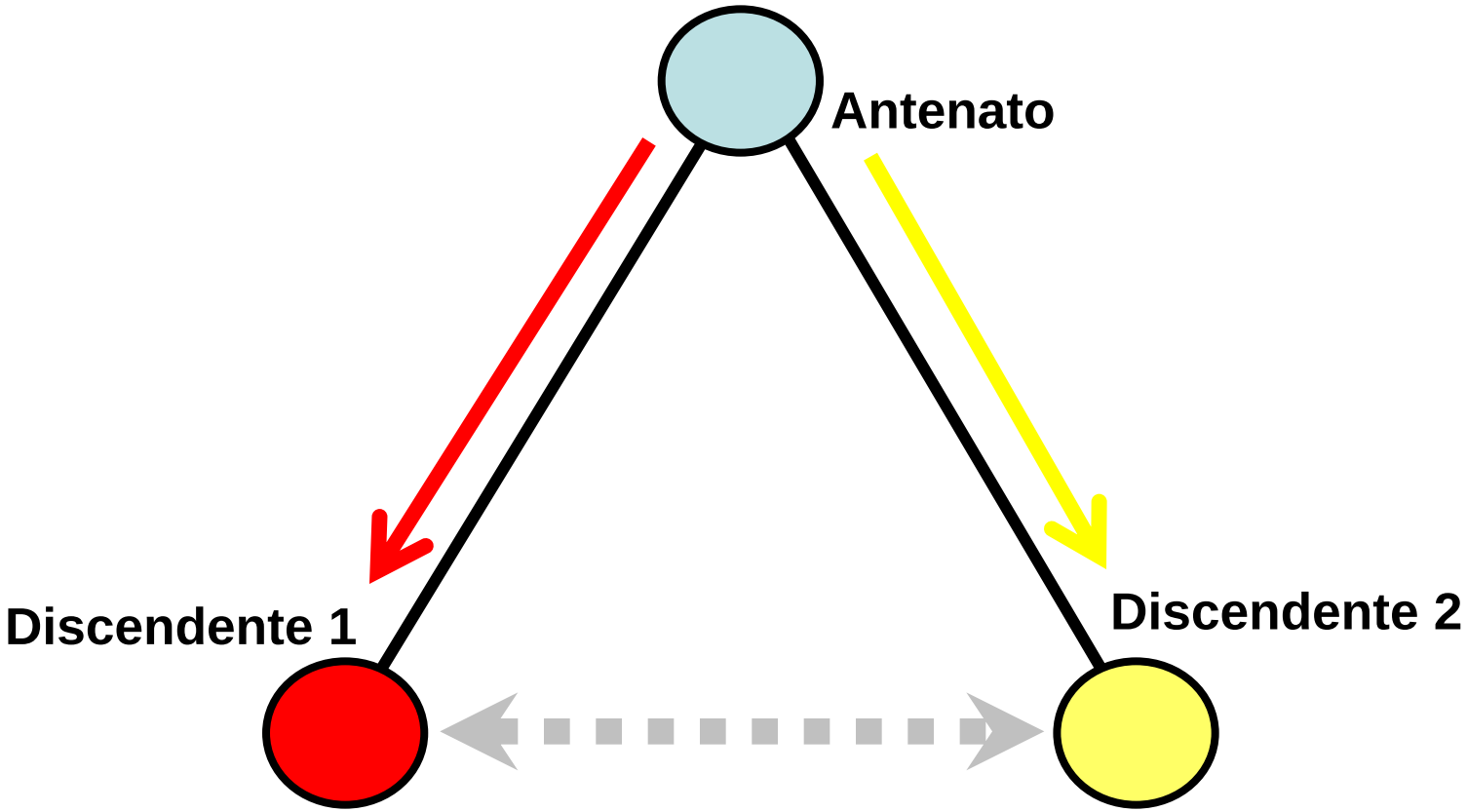


Le distanze tra ogni coppia di specie vivente è stimabile



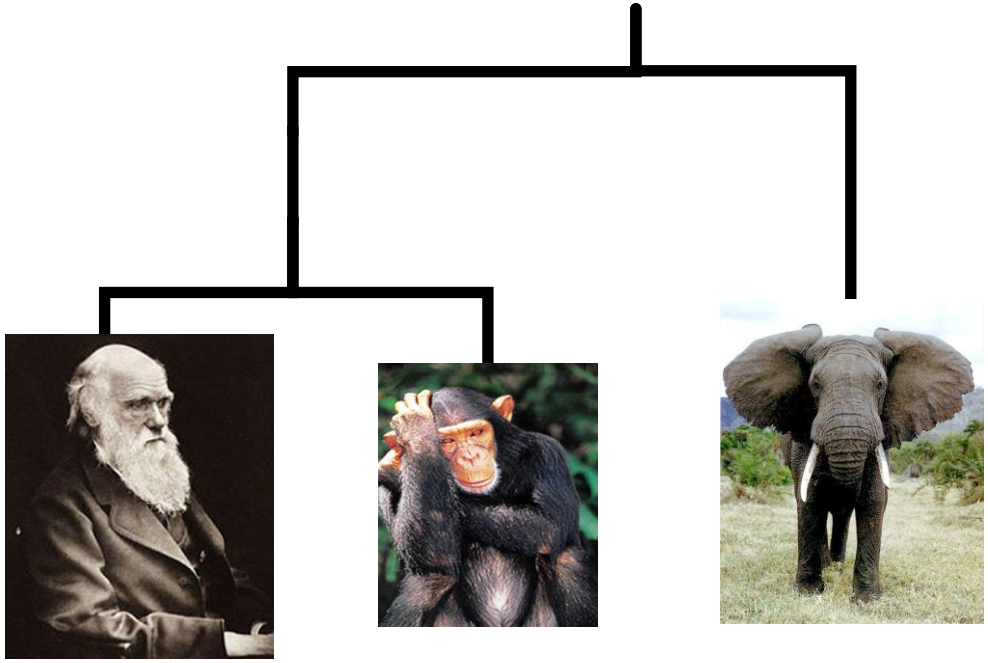
- Anche tra specie estinte, quando posso
- Con quali dati?

La distanza dipende dalla somma dei cambiamenti lungo le 2 linee



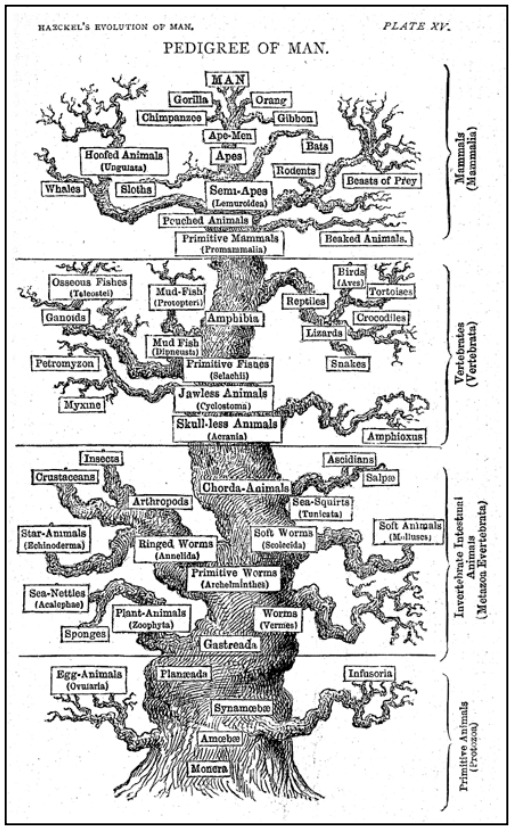
Un altro passaggio logico

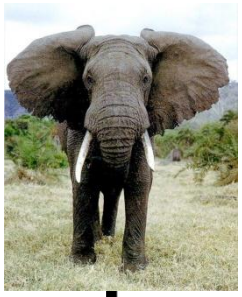
- Due linee evolutive si assomigliano di più tra di loro rispetto ad una terza linea evolutiva se condividono PRIMA (in tempi più recenti) un antenato comune

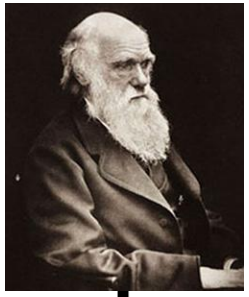


- Le ipotesi filogenetiche sono ipotesi che riguardano gli antenati comuni

Gli eventi di divergenza fanno parte di un unica *storia* (che non è solo nostra)





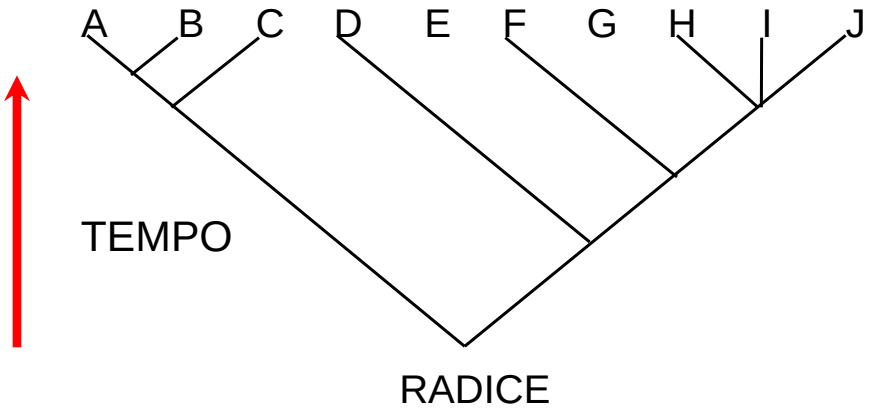


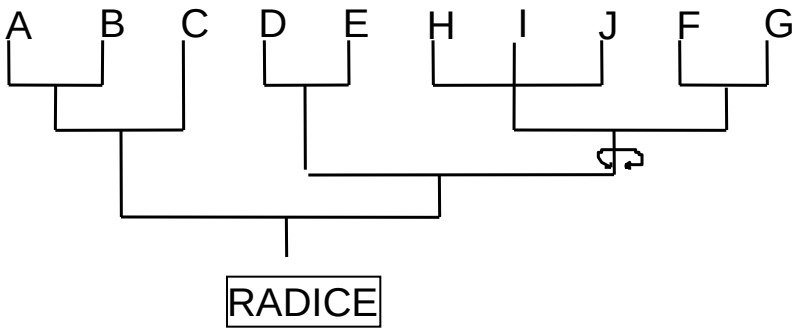
5 MYA

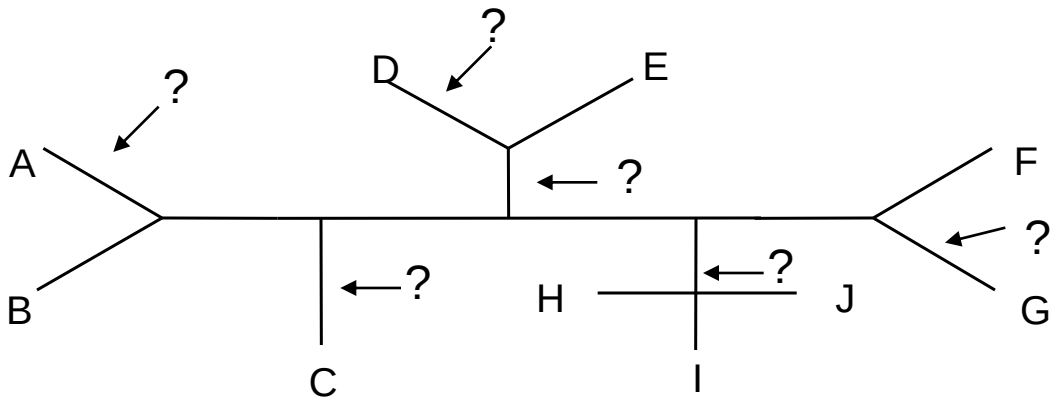
120 MYA

1,500 MYA

Alberi con e senza radice

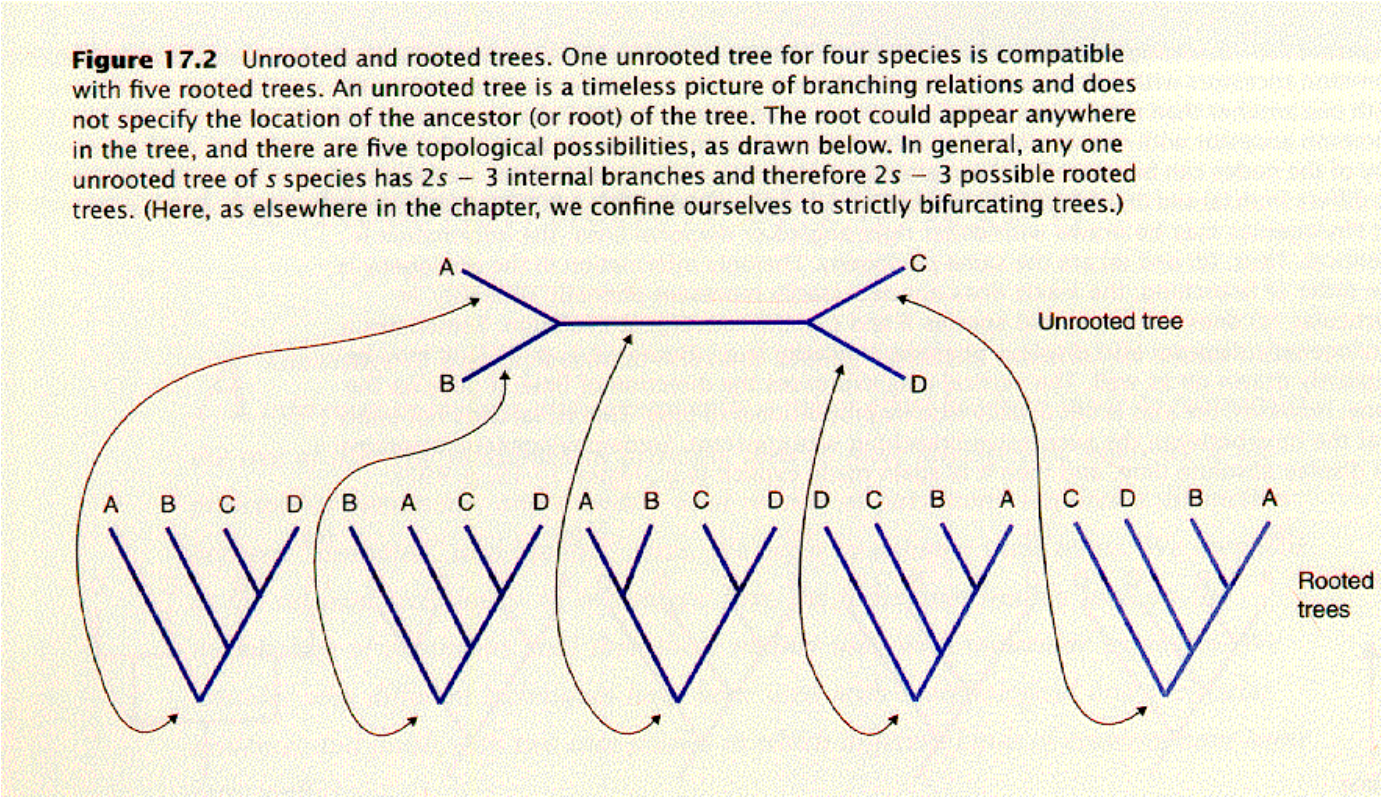






La figura in basso rappresenta un albero senza radice

Alberi con e senza radice



Per ogni albero senza radice, ci sono 2s-3 alberi con radice, dove s è il numero di unità tassonomiche. 2s-3 corrisponde ovviamente al numero di rami. Si considerano solo alberi dicotomici.

Alberi con e senza radice

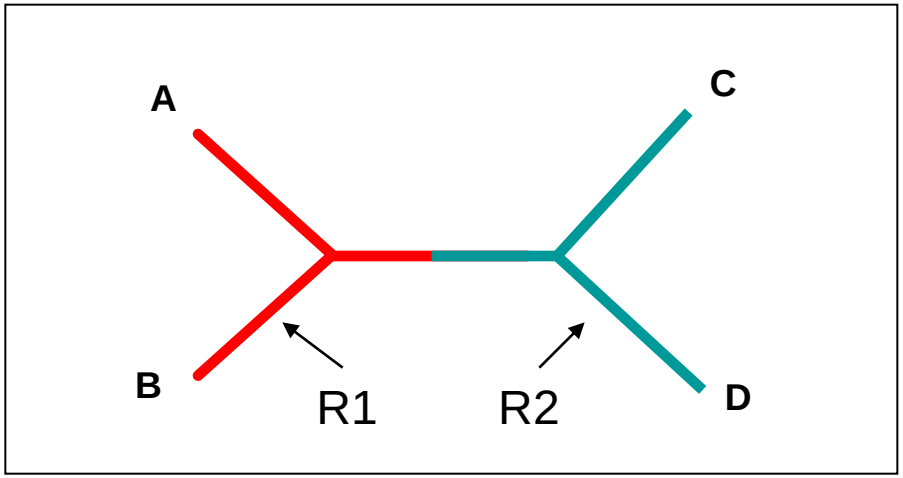
TABLE 5.1 Numbers of possible rooted and unrooted trees for up to 20 OTUs

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025
11	654,729,075	34,459,425
12	13,749,310,575	654,729,075
13	316,234,143,225	13,749,310,575
14	7,905,853,580,625	316,234,143,225
15	213,458,046,676,875	7,905,853,580,625
16	6,190,283,353,629,375	213,458,046,676,875
17	191,898,783,962,510,625	6,190,283,353,629,375
18	6,332,659,870,762,850,625	191,898,783,962,510,625
19	221,643,095,476,699,771,875	6,332,659,870,762,850,625
20	8,200,794,532,637,891,559,375	221,643,095,476,699,771,875

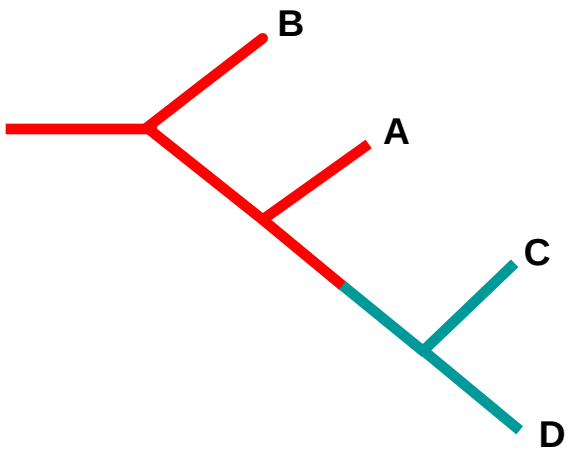
Data from Felsenstein (1978b).

Il numero di possibili alberi è molto alto. Quelli con radice sono di più, quindi il processo inferenziale è più semplice se ci si accontenta di un albero senza radice

Gli alberi con radice identificano direzioni di cambiamento

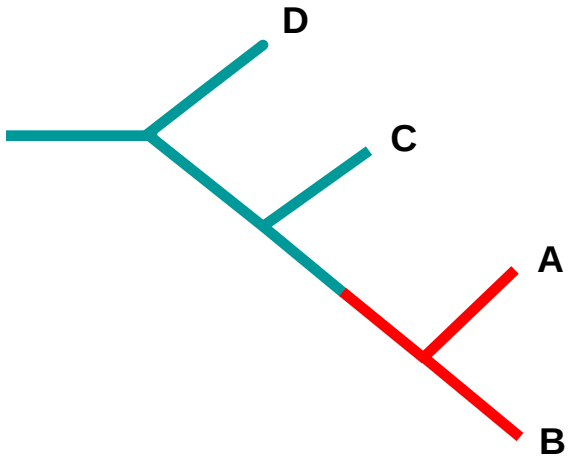


Se R1 è corretta



A+B+C non sono un gruppo monofiletico

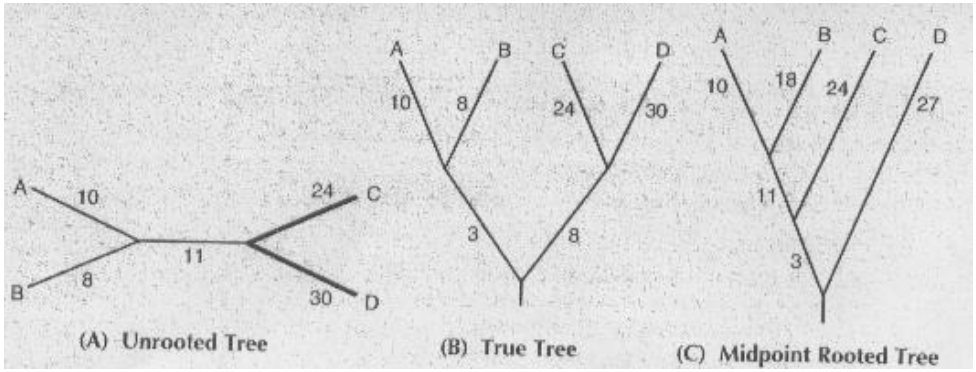
Se R2 è corretta



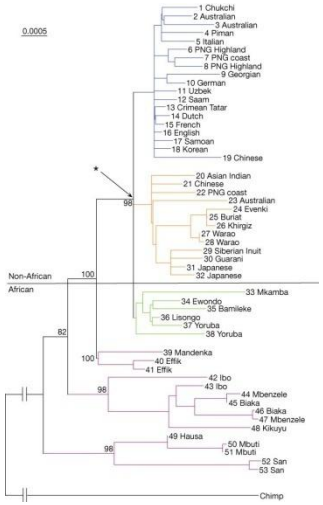
A+B+C sono un gruppo monofiletico

Rooting: trovare la radice di un albero filogenetico

- 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

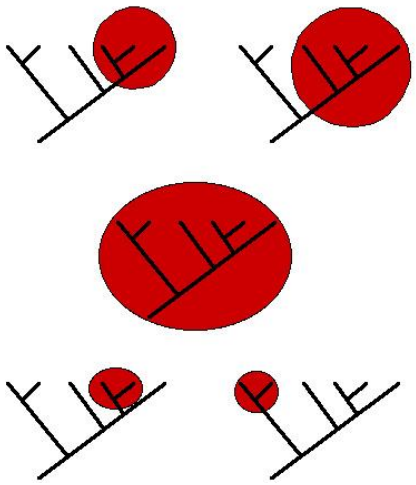


- Metodo dell'Outgroup: utilizzo un gruppo "esterno" alla filogenesi che sto analizzando
  - Assumo che l'outgroup si sia separato prima di tutti gli altri (devo fare un'ipotesi filogenetica esterna al gruppo che mi interessa)
  - La divergenza dell'outgroup non deve essere né troppo piccola né troppo grande

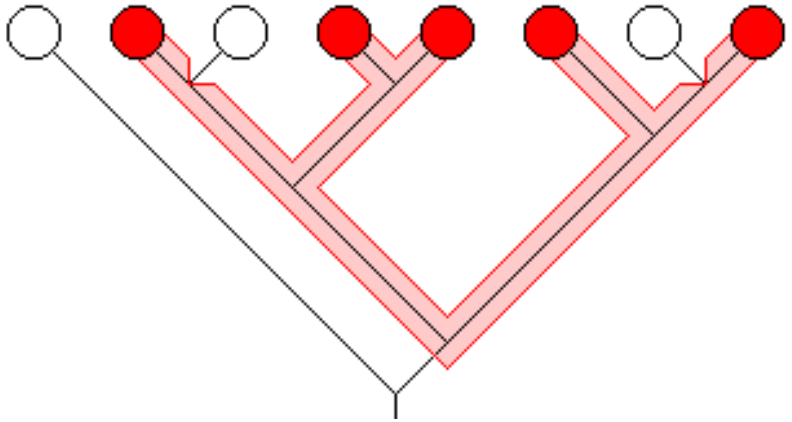


Classificazione e alberi: alcune definizioni per i taxa

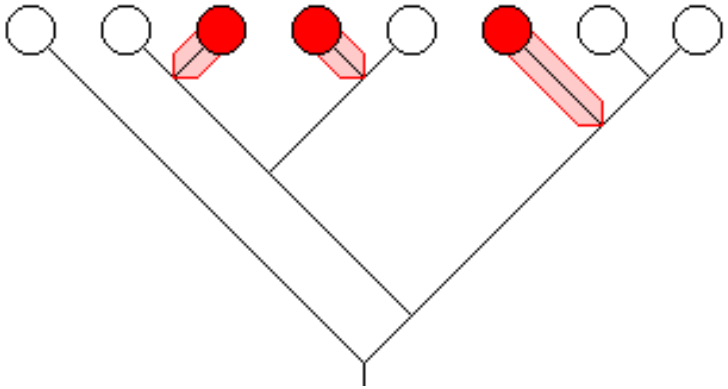
Monophyletic Groups



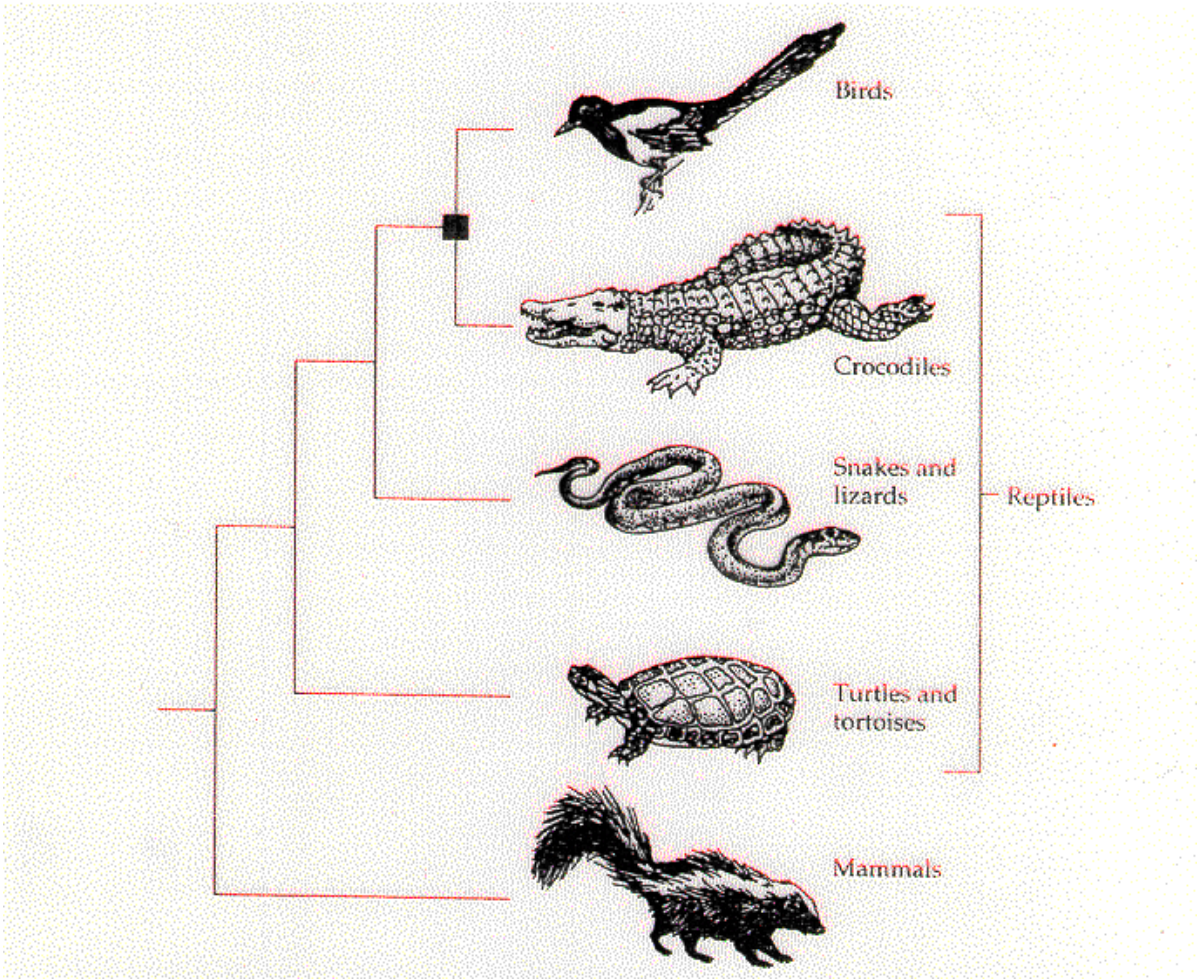
Paraphyletic taxon :



Polyphyletic taxon :

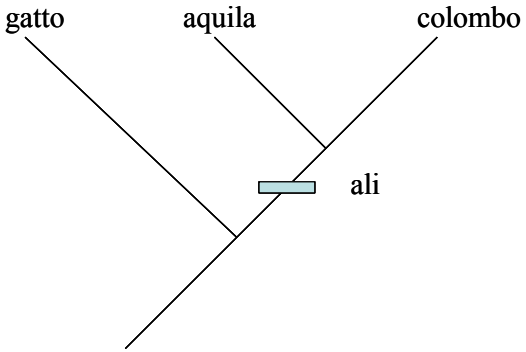


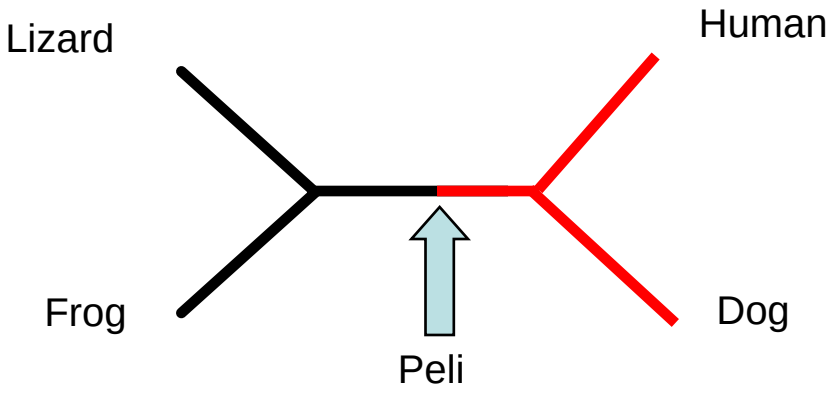
Classificazione e alberi: alcune definizioni per i taxa



Un classico esempio di parafilia

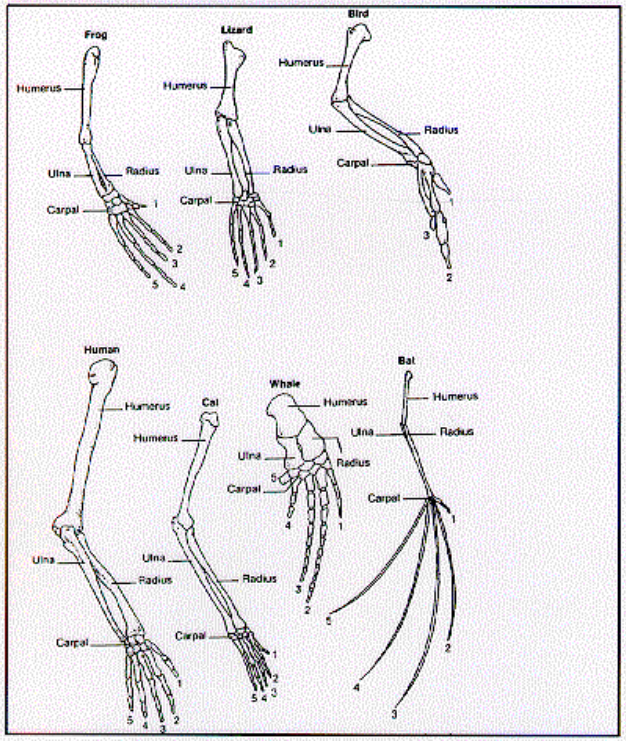
La “vera” somiglianza per costruire alberi: l’omologia





- 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 percè l’antenato comune era alato. Stesso ragionamento per i peli in cane e uomo

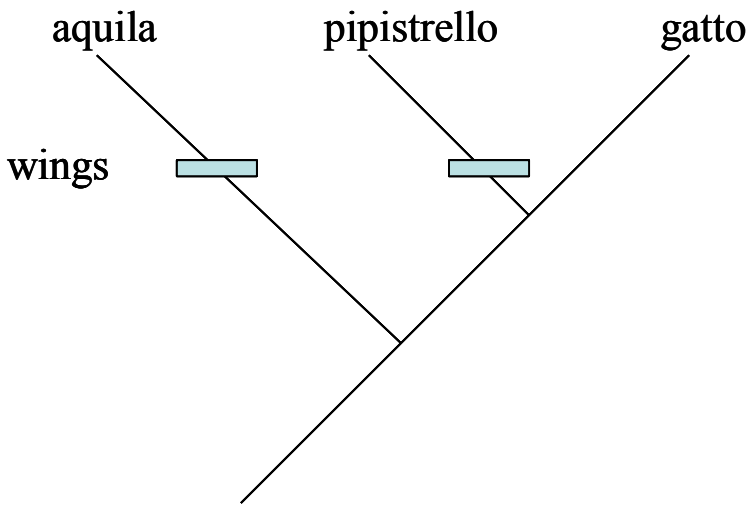
Omologia

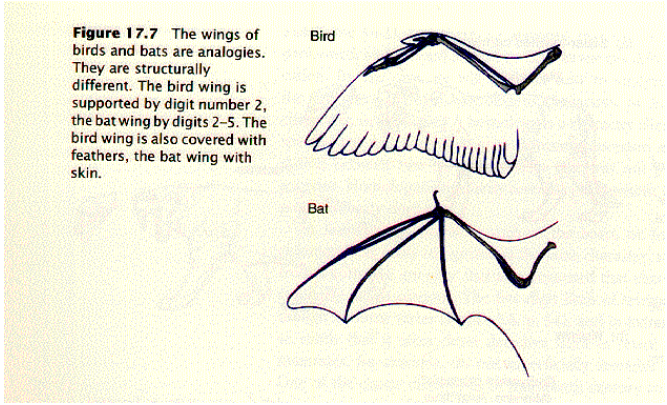


This is what we want, the best classification. The feature is shared because it derives directly from a common ancestor. For example the bony features of the forelimbs in vertebrates.

L’omologia nella struttura dell’arto anteriore dei vertebrati: questa struttura era presente nell’antenato comune. Caratteristiche tipiche dell’omologia: stessa struttura fondamentale, stesse relazioni con caratteri circostanti, forti somiglianze nello sviluppo embriologico

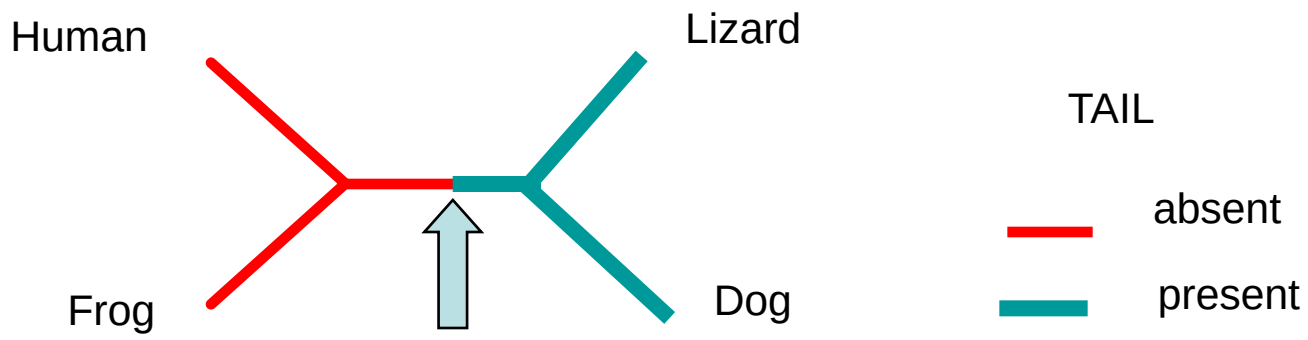
La somiglianza per convergenza può creare problemi: l'analogia



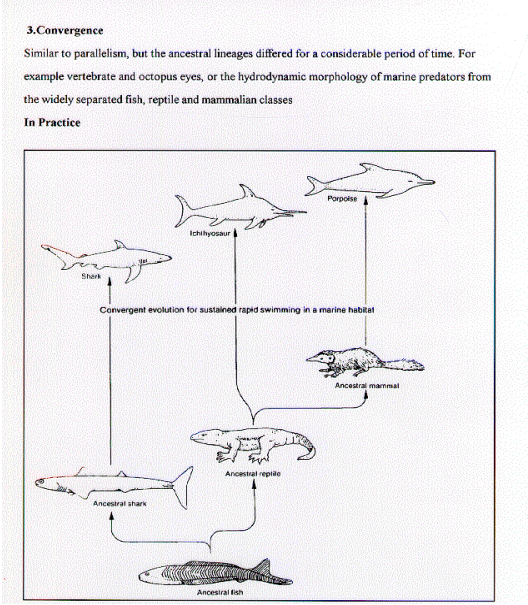


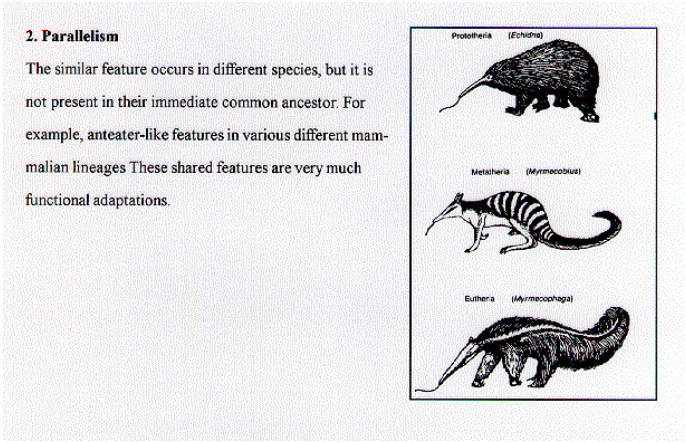
- Un carattere è simile (o presente) in due linee filogenetiche a causa di due eventi evolutivi indipendenti
- Questa somiglianza, non dovuta alla presenza del carattere nell'antenato comune delle due linee, è detta *omoplasia* o *analogia*
- Le ali nell'aquila e nel pipistrello sono un'omoplasia, perchè non erano presenti nell'antenato comune (un rettile tetrapode) non alato
- L'omoplasia non è sempre facile da riconoscere, e può produrre false filogenesi

False filogenesi considerando omoploasie



I diversi tipi di omoplasia





Convergenza nella morfologia: in genere, funzioni simili con strutture diverse

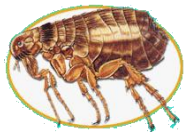
Convergenza in sequenze proteiche: in genere, a funzioni simili corrispondono sequenze molto diverse

Evoluzione parallela della morfologia: in genere, a funzioni simili corrispondono anche strutture simili

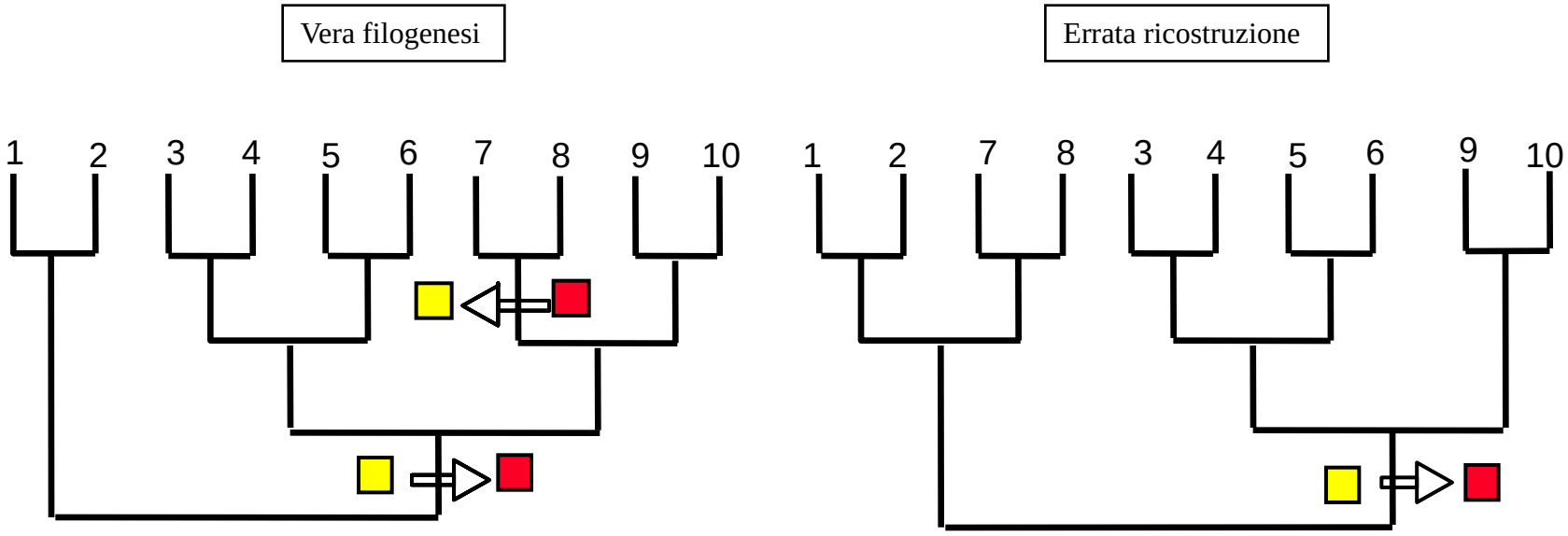
Evoluzione parallela in sequenze proteiche: in genere, a funzioni simili corrispondono anche sequenze simili

Reversione

Assenza di ali in tisanuri e pulci



Esempio di errata ricostruzione filogenetica in presenza di reversione



Distinguere se possibile omologie da omoplasie, altrimenti congruenza

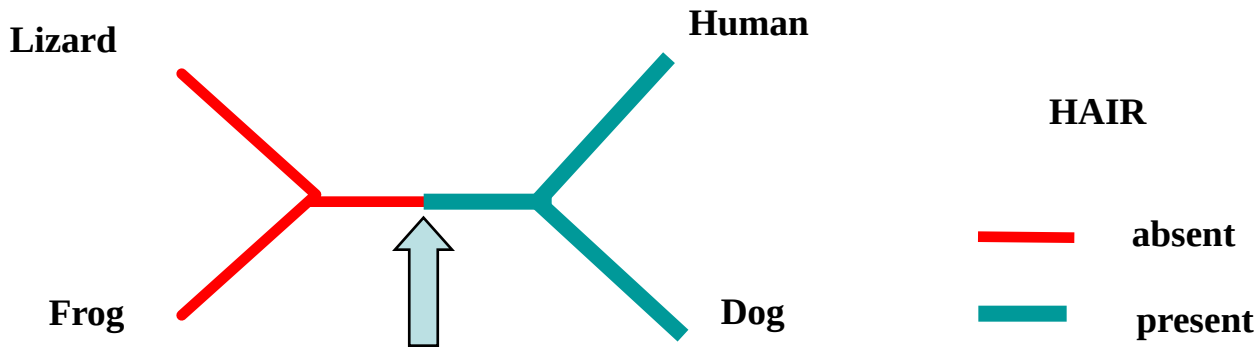
- Se c'e' omologia, la somiglianza fenotipica tende ad essere più profonda: nella struttura, nella posizione, nel modello di sviluppo embrionale.

- Darwin stesso però riconosce l'importanza del concetto di congruenza: la presenza di molti altri caratteri che suggeriscono se un una somiglianza è una omologia è la miglior prova. Fondamentale (e implicito) il concetto di congruenza nelle ricostruzioni basate su sequenze di DNA.

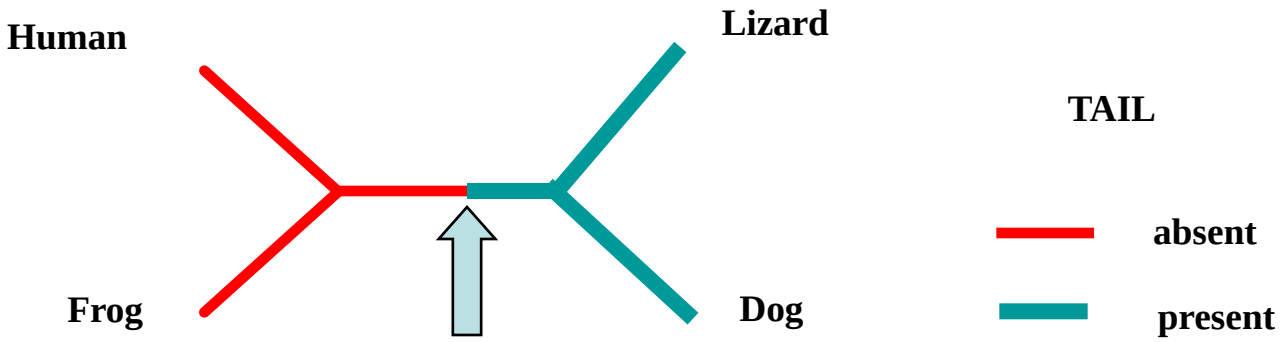
"The importance, for classification, of trifling characters, mainly depends on their being correlated with several other characters of more or less importance. The value indeed of an aggregate of characters is very evident ..... a classification founded on any single character, however important that may be, has always failed."

Charles Darwin, The Origin of Species

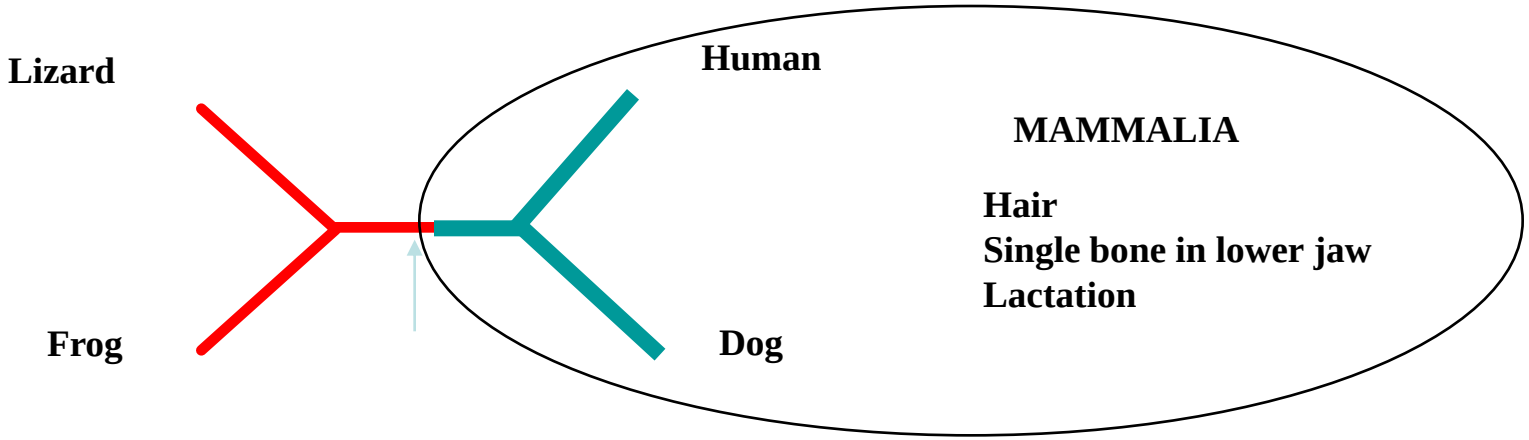
Le omoplasie producono incongruenze



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



Congruenza e alberi preferiti



Viene preferito l'albero supportato da molti caratteri conguenti

Ricostruzioni filogenetiche basate sul DNA: vantaggi

- Descrizione dei caratteri non ambigua
- Somiglianza dovuta a effetti ambientali non genetici non interferisce
- Evoluzione convergente implica spesso fenotipi simili ma genotipi differenti
- 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 non codificante
- Tutti gli individui hanno DNA!

Ricostruzioni filogenetiche basate sul DNA: svantaggi

- Omoplasia può essere frequente
  - Pochi stati del carattere (A,C,T,G)
  - Tasso di mutazione può essere elevato

- Mutazioni ricorrenti modificano la relazione tra distanza genetica e distanza temporale

- Duplicazioni e trasferimento orizzontale di geni possono essere identificati, ma 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 (ma almeno sono espliciti!)

- Alberi di geni e alberi di specie possono essere diversi

Accenni a tre tipologie di metodi per ricostruire filogenesi

1. Metodi basati sulla stima di distanze

2. Metodi basati sulla parsimonia

3. Metodi basati sulla verosimiglianza

- Nelle metodologie di tipo 2 e 3 bisogna valutare tante topologie (teoricamente tutte)
- Ci sono poi i metodi bayesiani!

Metodi basati sulle distanze

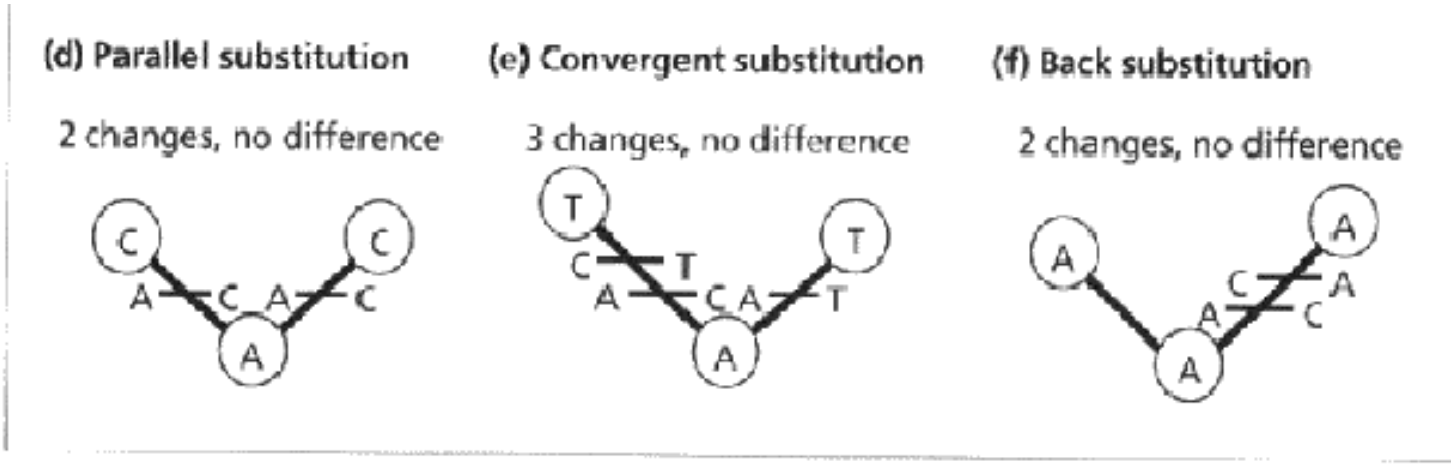
Vantaggi

- Veloce: va bene per analizzare grandi data sets
- Basta avere matrici di distanze (per alcuni dati (DNA hybridization, fingerprinting pattern, reazioni ad anticorpi, etc) sono gli unici dati disponibili

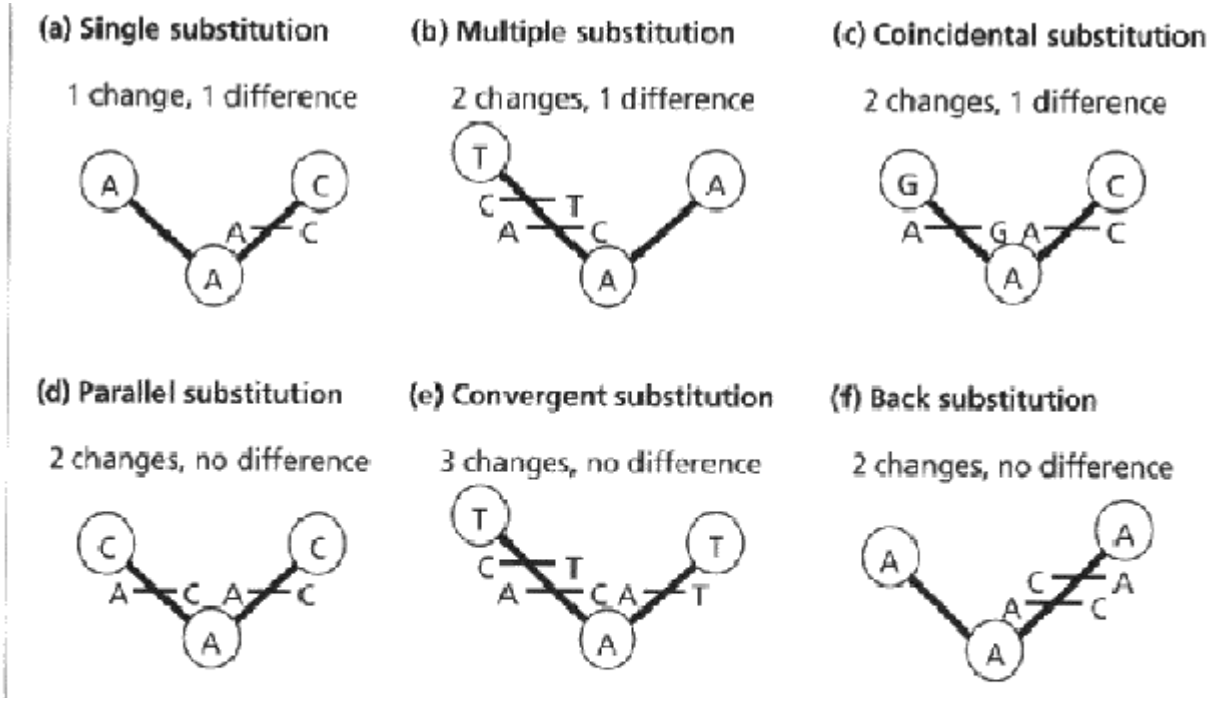
Svantaggi

- Netta perdita di informazione: dalle distanze non si torna indietro alle sequenze!
- Problemi con misure di distanza non lineari con il tempo

Omoplasie a livello molecolare



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 (casi b,c,d,e ed f)
- Servono correzioni alle misure di distanza
- In alcuni casi, l'eccessivo numero di mutazioni satura l'informazione

Mutazioni ricorrenti, sottostima della distanza e saturazione

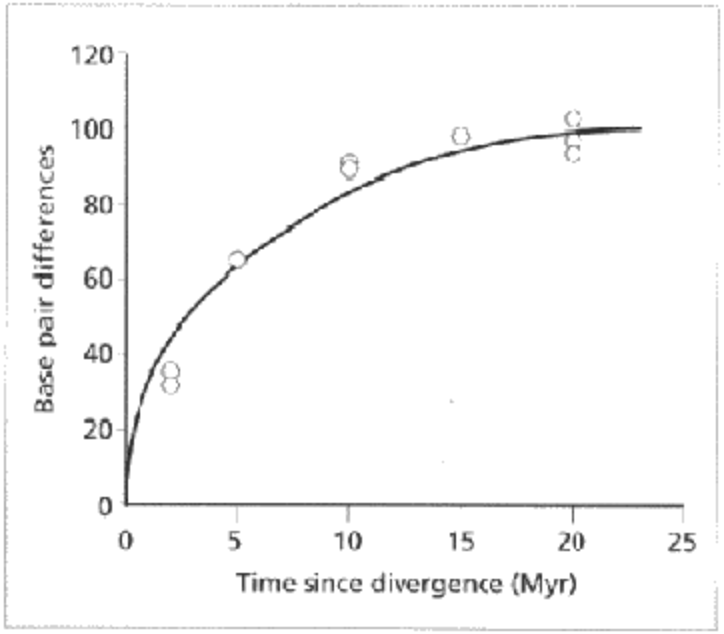
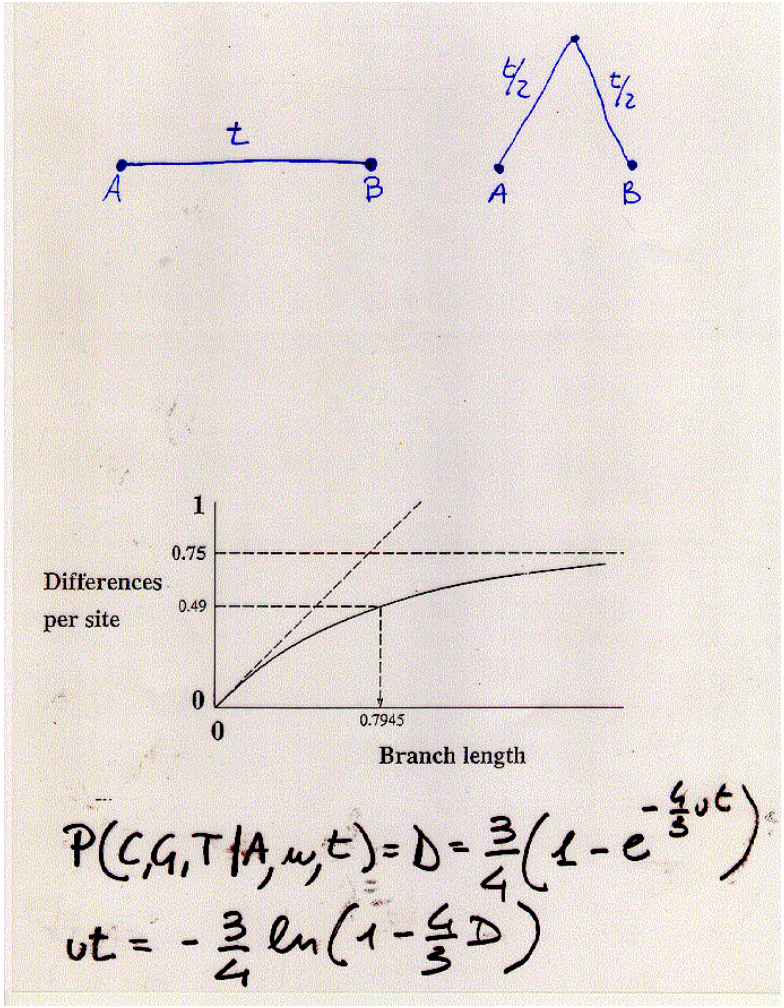


Fig. 5.11 Number of nucleotide substitutions between pairs of bovid mammal mitochondrial sequences (684 basepairs from the *COII* gene) against estimated time of divergence. Notice that the observed number of substitutions is not linear with time but curvilinear. Data from Janecek *et al.* (1996).

All'aumentare della distanza temporale, il numero di differenze non può accumularsi in maniera lineare. Più passa il tempo, maggiore è la frazione di mutazioni che avvengono a siti già mutati e quindi non aumentano (e a volte diminuiscono) la distanza.

Mutazioni ricorrenti: servono distanze corrette



- E' necessario un modello per convertire distanze molecolari in numero di mutazioni

- Il più semplice è il modello di Jukes e Cantor

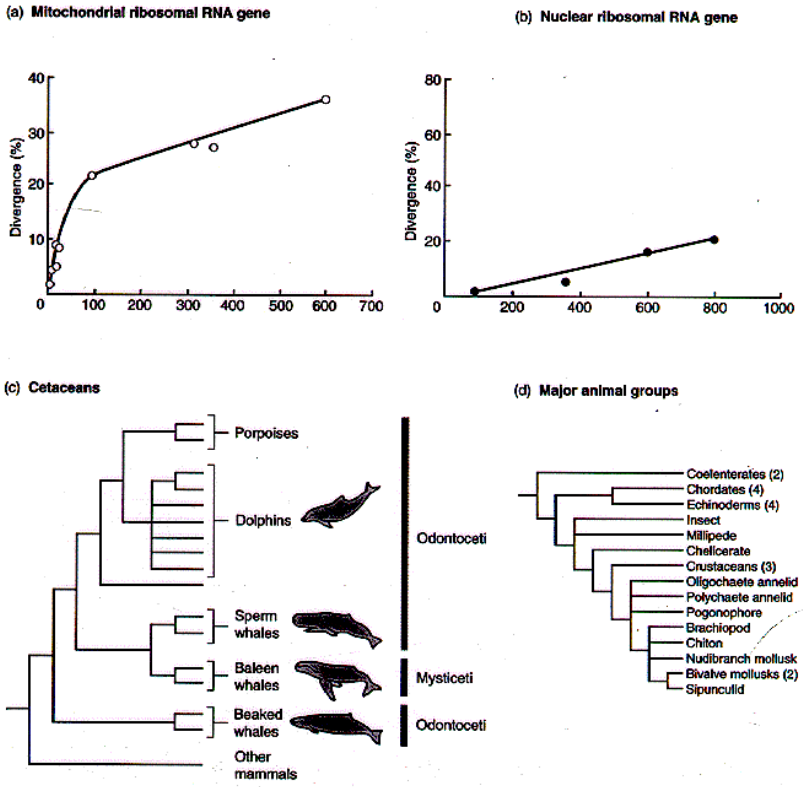
Esempio

- Se sequenzio 500bp, e A e B differiscono di 50bp, D=0.1, quindi ut = numero di mutazioni effettive per sito = tasso di mutazione per generazione x tempo di divergenza in generazioni = 0.1073, ossia mi aspetto che siano avvenute 0.1073*500 = 53.66 mutazioni in quella sequenza .
- Se D = 0.5, ut = 0.82

Per tempi molto grandi, si raggiunge la saturazione: troppe mutazioni, le differenze sono casuali e pari a 0.75 per sito (perchè)?

Mutazioni ricorrenti: scegliere bene i geni da studiare

Figure 17.14 Matching the molecule to the phylogenetic problem. The ribosomal RNA genes in (a) the mitochondria evolve more rapidly than those in (b) the nucleus. The different points indicate species pairs, for which the date of their common ancestor can be estimated from fossils. The graphs tail off (at about 33% divergence) because of multiple substitutions at a site. (c) Phylogeny of dolphins and whales, using mitochondrial rRNA genes; the deepest root is about 35 million years ago. (d) Relations of major animal groups, as revealed by nuclear rRNA genes; the deepest root is probably over 600 million years ago. Reprinted, by permission of the publisher, from Mindell and Honeycutt (1990), Milinkovitch *et al.* (1993), and Lake (1990).



A volte la buona scelta dei marcatori a seconda della scala temporale che si sta considerando rende le correzioni per mutazioni ricorrenti non indispensabili

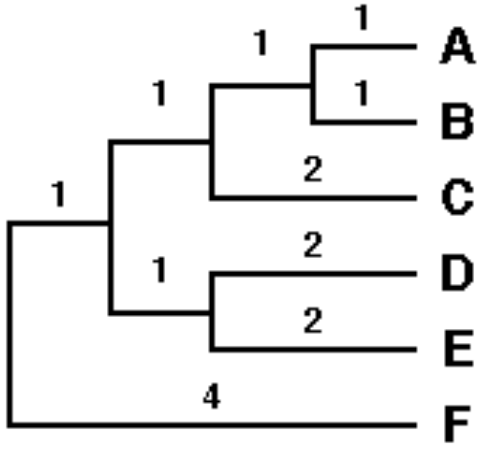
Trovare l'albero a partire dalla matrice delle distanze

UPGMA (Unweight Pair Group Method with Arithmetic mean)

• Funziona al meglio per alberi ultrametrici (tassi deterministicamente costanti)

• Posiziona automaticamente la radice

• Vediamo un esempio



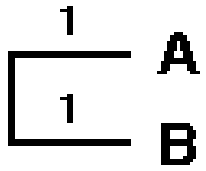
	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

Albero vero (ignoto) da ricostrure

Dati: matrice di distanze a coppie

UPGMA all'opera

• Unisco taxa con distanza minore, stimo le distanze dal nodo, e calcolo le distanze delle specie rimanenti dal gruppo appena fromato, e modifico la matrice

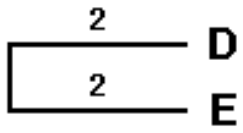


dist(A,B),C = (distAC + distBC) / 2 = 4
dist(A,B),D = (distAD + distBD) / 2 = 6
dist(A,B),E = (distAE + distBE) / 2 = 6
dist(A,B),F = (distAF + distBF) / 2 = 8

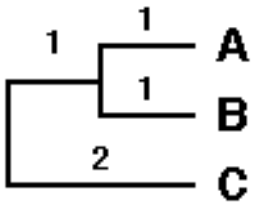
	A,B	**C**	**D**	**E**
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

• Procedo iterativamente nello stesso modo

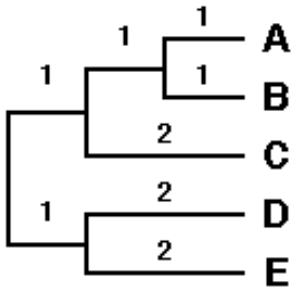
UPGMA all'opera



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

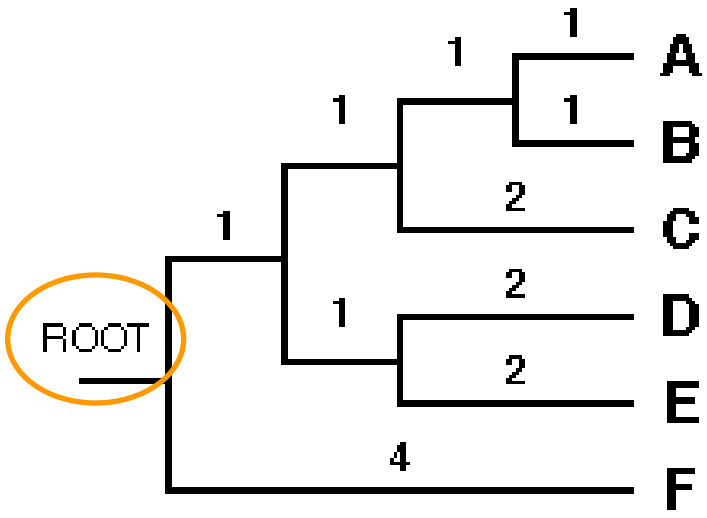


	AB,C	D,E
D,E	6	
F	8	8

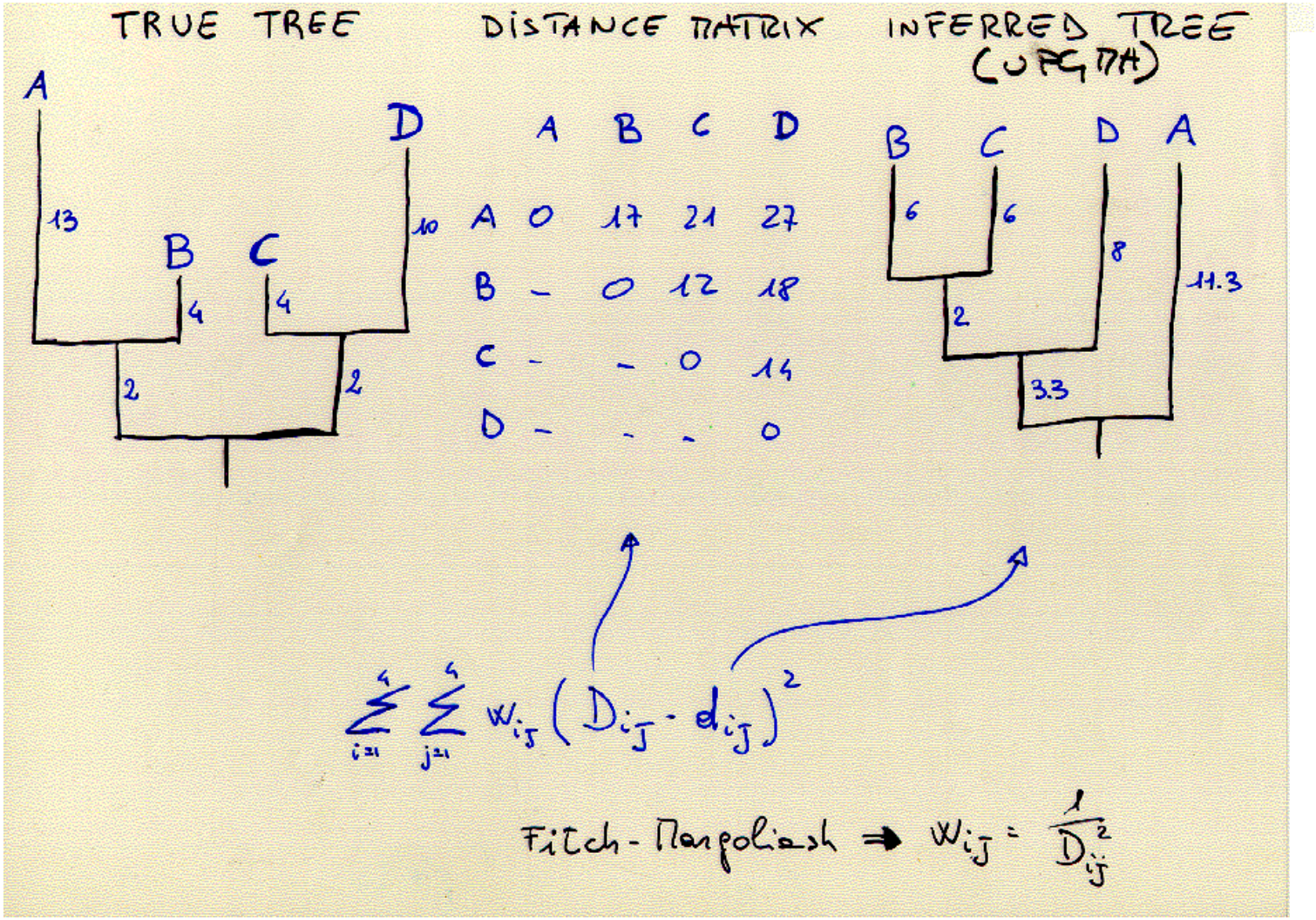


	ABC,DE
F	8

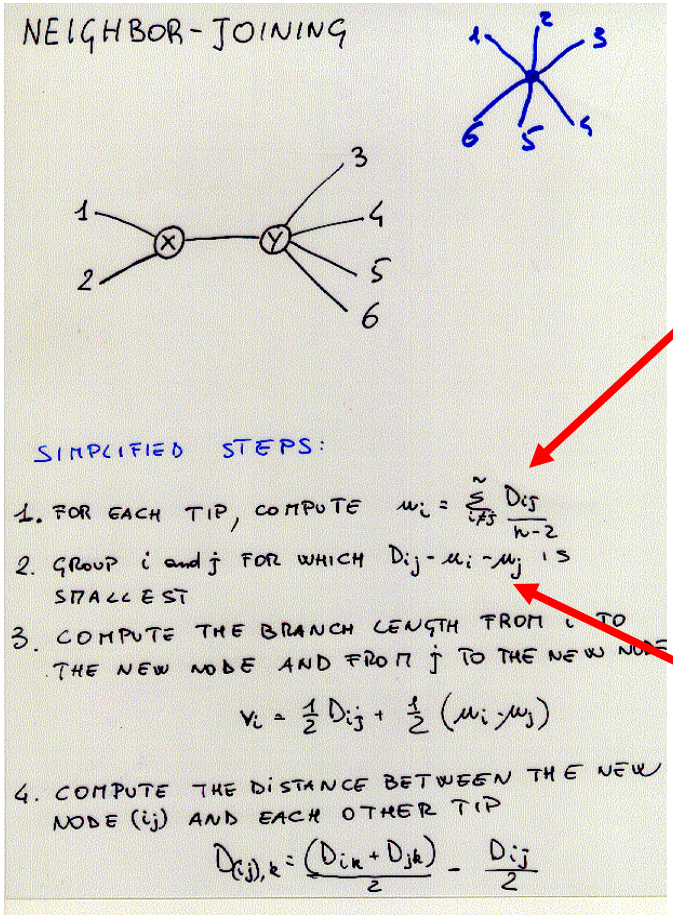
UPGMA all'opera



Gli errori di UPGMA con tassi non costanti



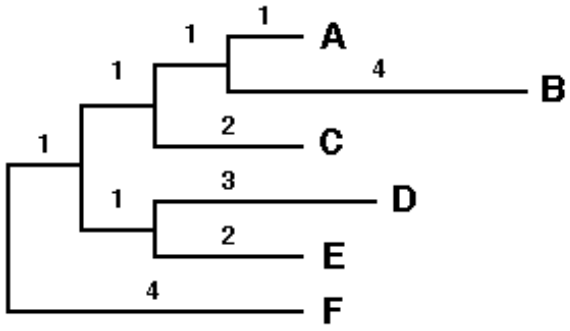
Un'alternativa: Neighbor-Joining



Considera la divergenza media di ciascun gruppo da tutti gli altri

In successione, si cerca l'albero con la minor lunghezza

Un' alternativa: Neighbor-Joining



ALBERO VERO

	A	B	C	D	E
B	5				
C	4	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

UPGMA

Neighbor-Joining

$m_A = (5+4+7+6+8)/4=7.5$
$m_B = (5+7+10+9+11)/4=10.5$
$m_C = (4+7+7+6+8)/4=8$

$D_{AB} = 5 - 7.5 - 10.5 = -13$
$D_{AC} = 4 - 7.5 - 8 = -11.5$
$D_{BC} = 7 - 10.5 – 8 = -11.5$

Metodo della massima parsimonia

• L’albero migliore è quello con il minor numero di cambiamenti (eventi evolutivi,mutazioni, …), quello cioè più parsimonioso

• Ci possono essere molte topologie che implicano lo stesso numero di cambiamenti. Sono tutte ugualmente valide.

• Poche assunzioni sul modello di evoluzione molecolare, trovo l’albero con il minor numero di omoplasie

• Importante il concetto di siti informativi

Metodo della massima parsimonia

Quali sono le basi logiche di questo criterio? Ovvero, perché l'albero con meno cambiamenti dovrebbe essere quello corretto?

• Principio filosofico e teologico dei 13esimo secolo (William of Ockham): tra diverse spiegazioni, la più semplice è da preferire. Inutile ricorrere a molte assunzioni se posso spiegare qualcosa con poche.

• Dio ha creato tutto, e Dio non avrebbe creato nulla di complesso se poteva fare la stessa cosa in maniera semplice

• La selezione naturale favorisce gli adattamenti rapidi, ossia in un numero minore di passi

• Statisticamente, i cambiamenti evolutivi sono rari, quindi è improbabile che avvengano molte volte.

Metodi basati sulla parsimonia

Vantaggi

• Metodo molto semplice. Le operazioni sono chiare

• 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 suii diversi rami

Svantaggi

- Se l'omoplasia è frequente e non distribuita in maniera omogenea nell'albero produce false filogenesi
- Sottostima la lunghezza dei rami
- Modello di evoluzione è implicito, difficile studiare l'esatto funzionamento in diverse condizioni
- Fa affidamento al semplice assunto che le cose più parsimoniose sono quelle che probabilmente sono avvenute

Metodi basati sulla parsimonia

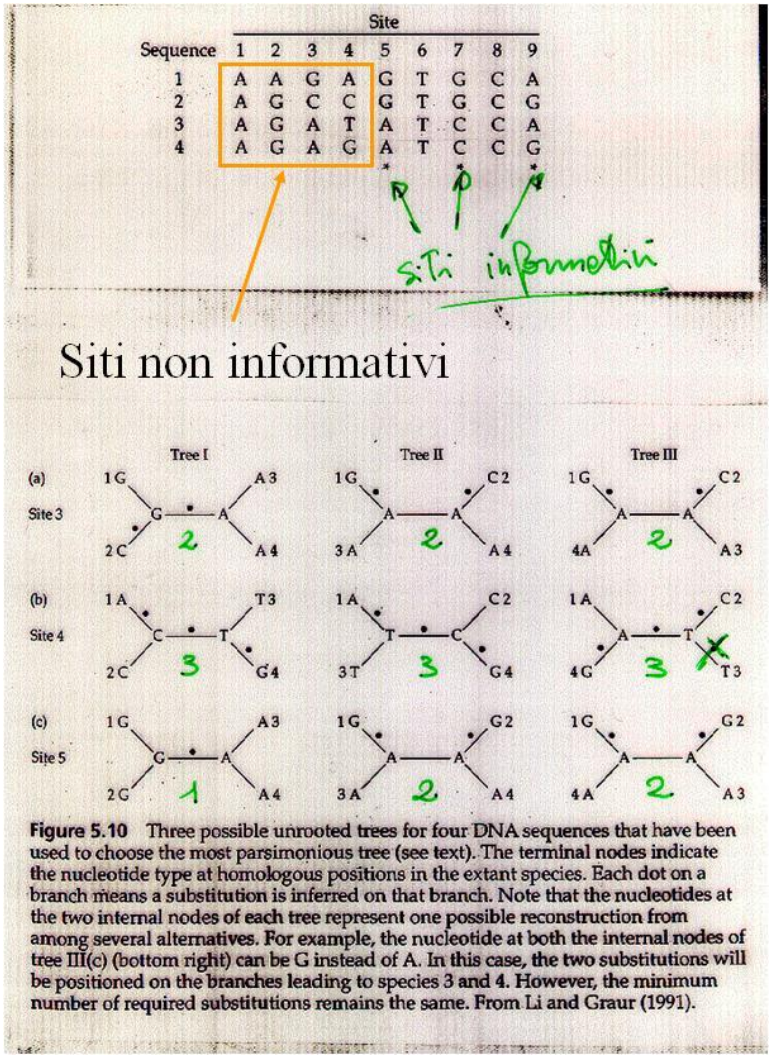
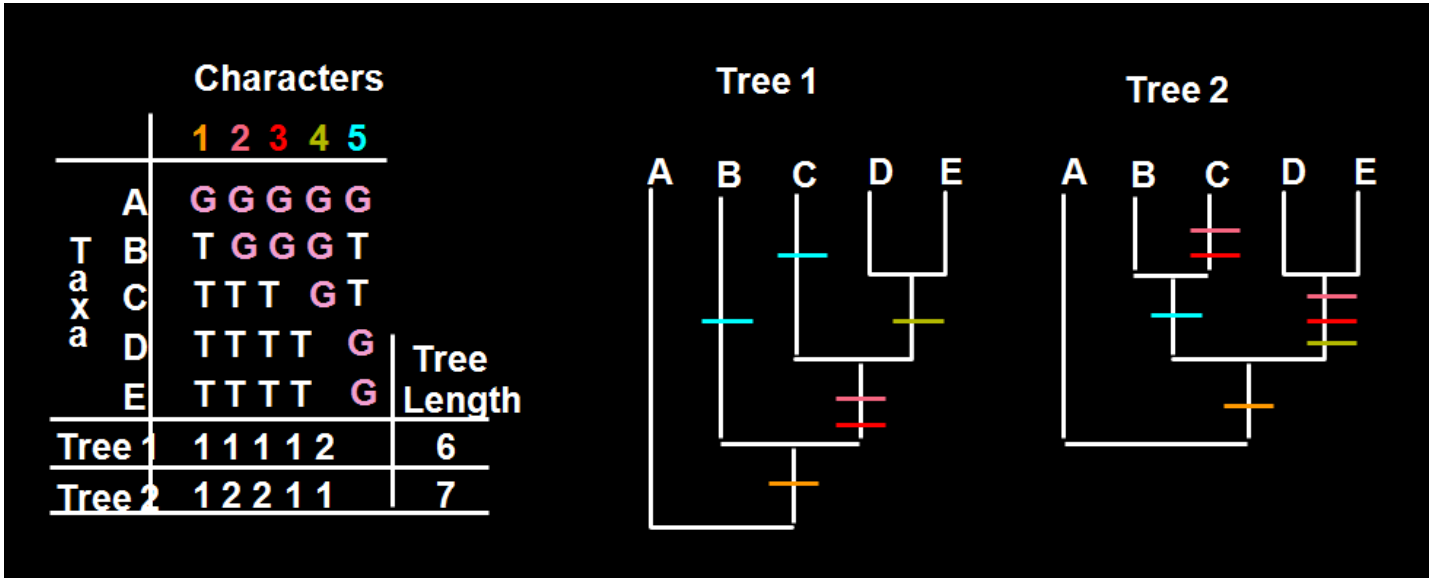


Figure 5.10 Three possible unrooted trees for four DNA sequences that have been used to choose the most parsimonious tree (see text). The terminal nodes indicate the nucleotide type at homologous positions in the extant species. Each dot on a branch means a substitution is inferred on that branch. Note that the nucleotides at the two internal nodes of each tree represent one possible reconstruction from among several alternatives. For example, the nucleotide at both the internal nodes of tree III(c) (bottom right) can be G instead of A. In this case, the two substitutions will be positioned on the branches leading to species 3 and 4. However, the minimum number of required substitutions remains the same. From Li and Graur (1991).

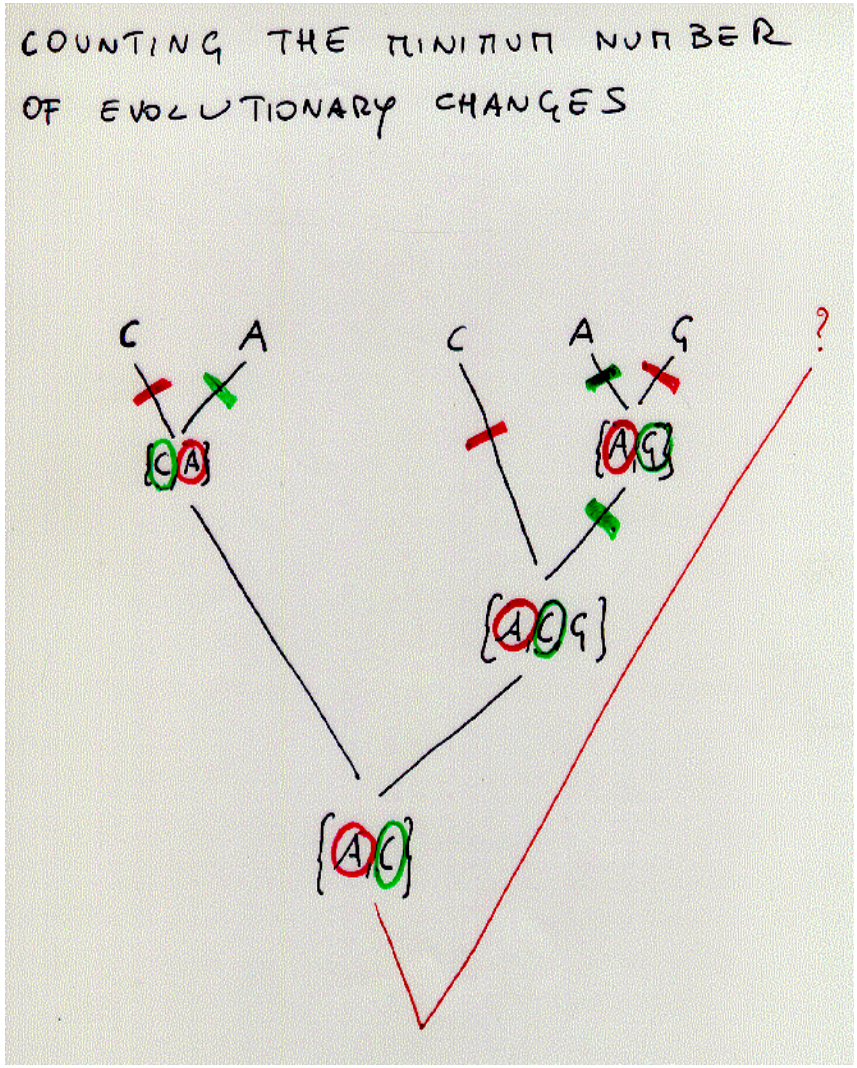
- Solo i siti informativi vengono utilizzati
- Sono quei siti che discriminano ipotesi topologiche

Parsimonia e DNA



L’albero 1 è più parsimonioso, ma tutti e due richiedono “extra steps” (omoplasia)

Parsimonia: l'algoritmo da applicare ad ogni topologia



- Ripeto il calcolo per ogni sito informativo e poi faccio la somma per trovare il numero di sostituzioni compatibili con una certa topologia.

- Devo analizzare tutte le topologie (o molte di loro)

Parsimonia: cosa si ottiene

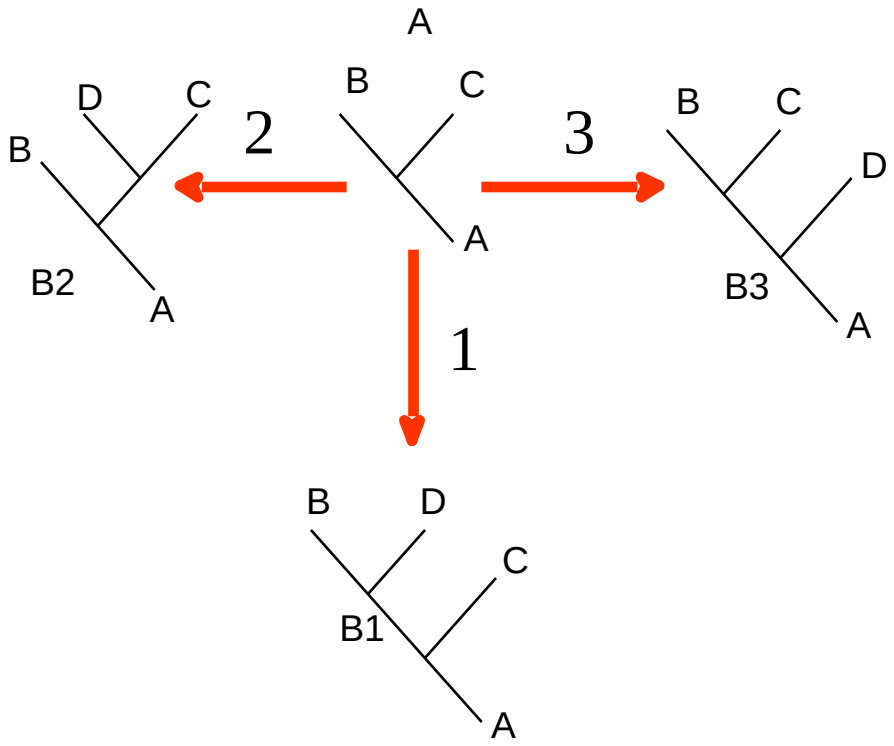
- Uno o più alberi maggiormente parsimoniosi
- Direttamente le ipotesi (dove e quando) riguardo i cambiamenti ad ogni carattere
- Direttamente la lunghezza dei rami (in numero di cambiamenti)
- Se si vuole, si possono indagare alberi subottimali (con più cambiamenti rispetto a quello più parsimonioso)

E’ necessario analizzare tutte le topologie?

- Le soluzioni esatte (ricerca esaustiva) non sono possibili con tanti taxa
- Devo per forza ricorrere a soluzioni euristiche: si esplora solo una parte della spazio
- Vediamo un esempio

Seleziono un “buon” albero di partenza

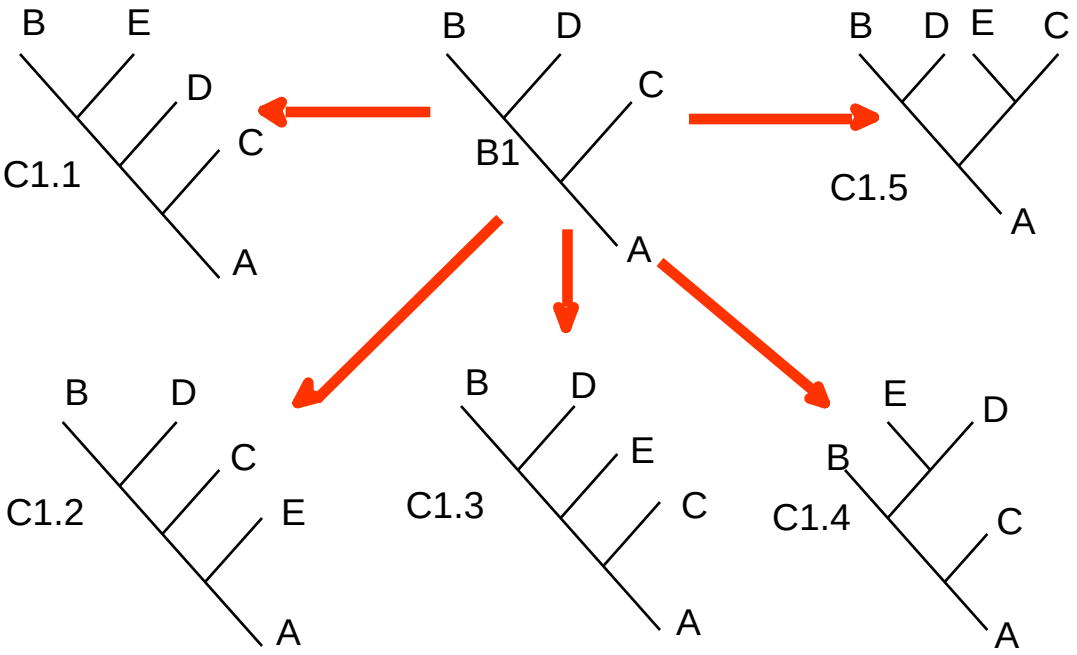
Per esempio, con il metodo di *sequential addition.* Inizio da una tripletta casuale.



Scelgo l’albero (B1, B2, o B3) migliore

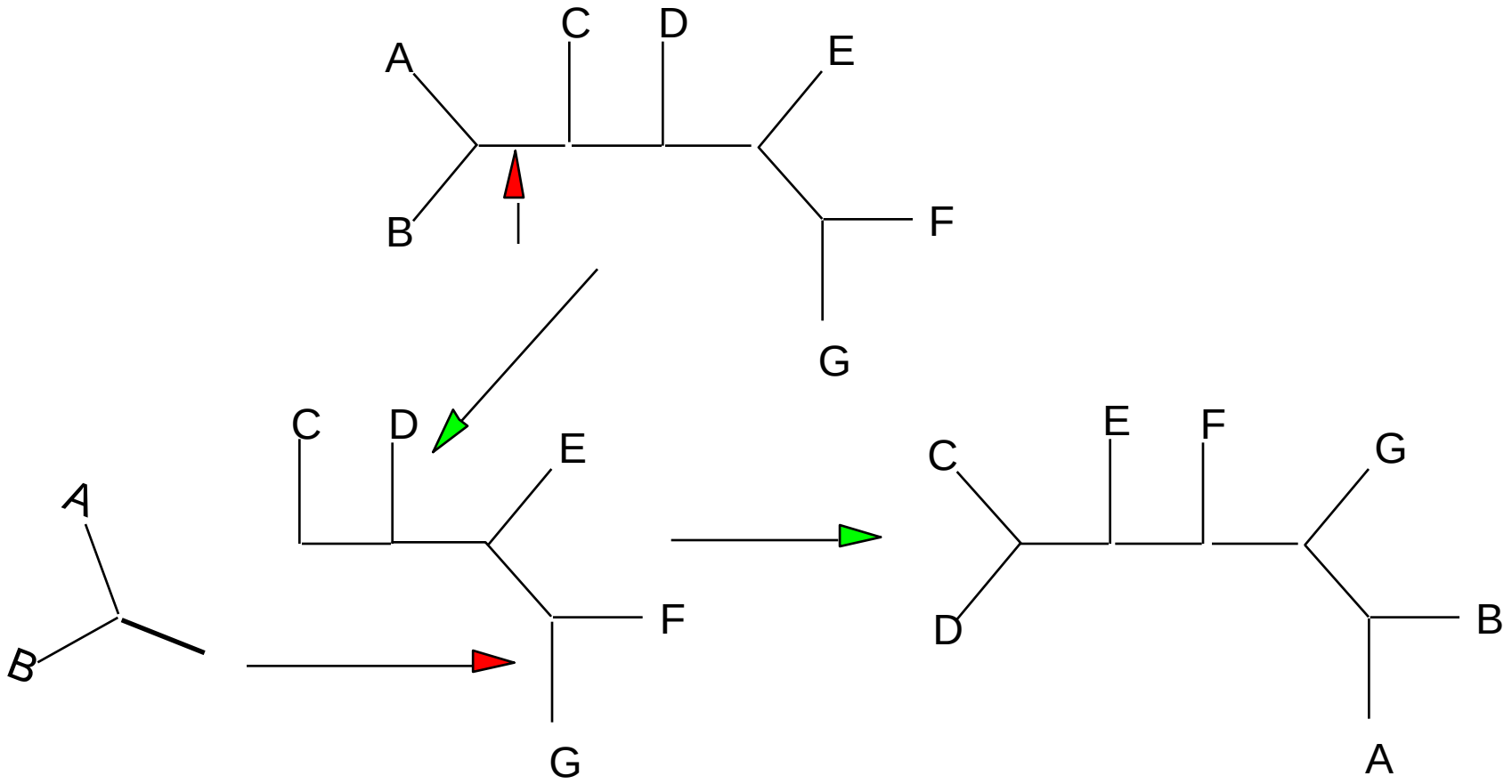
Seleziono un “buon” albero di partenza

Procedo iterativamente



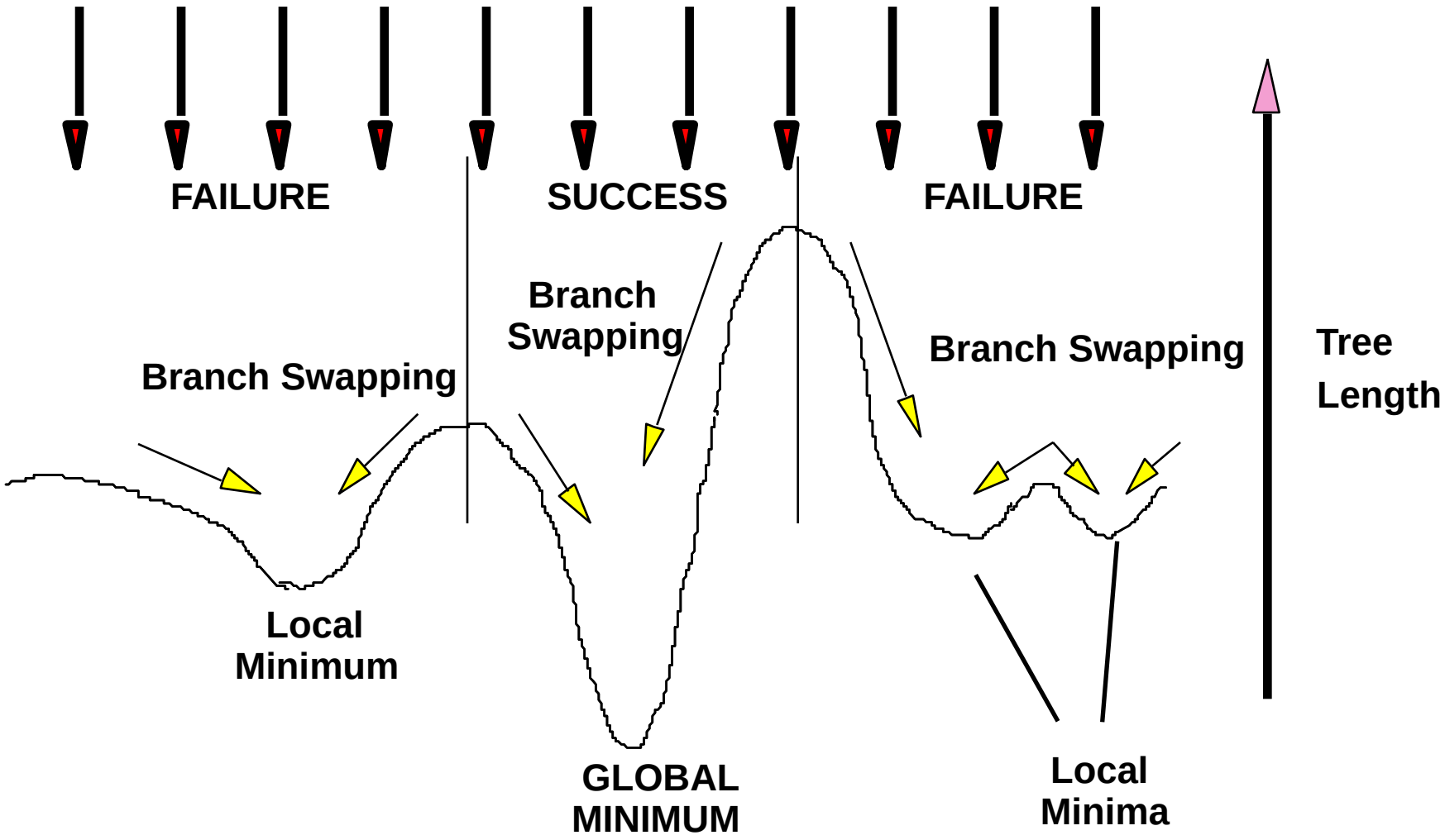
Scelgo il C1 migliore, e vado avanti fino all’albero di partenza (che dipenderà dall’ordine di addizione di taxa!)

Dall’albero di partenza inizio una serie di riarrangiamenti



Spesso si procede con strategie intermedie, ossia riarrangiamenti anche prima di giungere all’albero di partenza

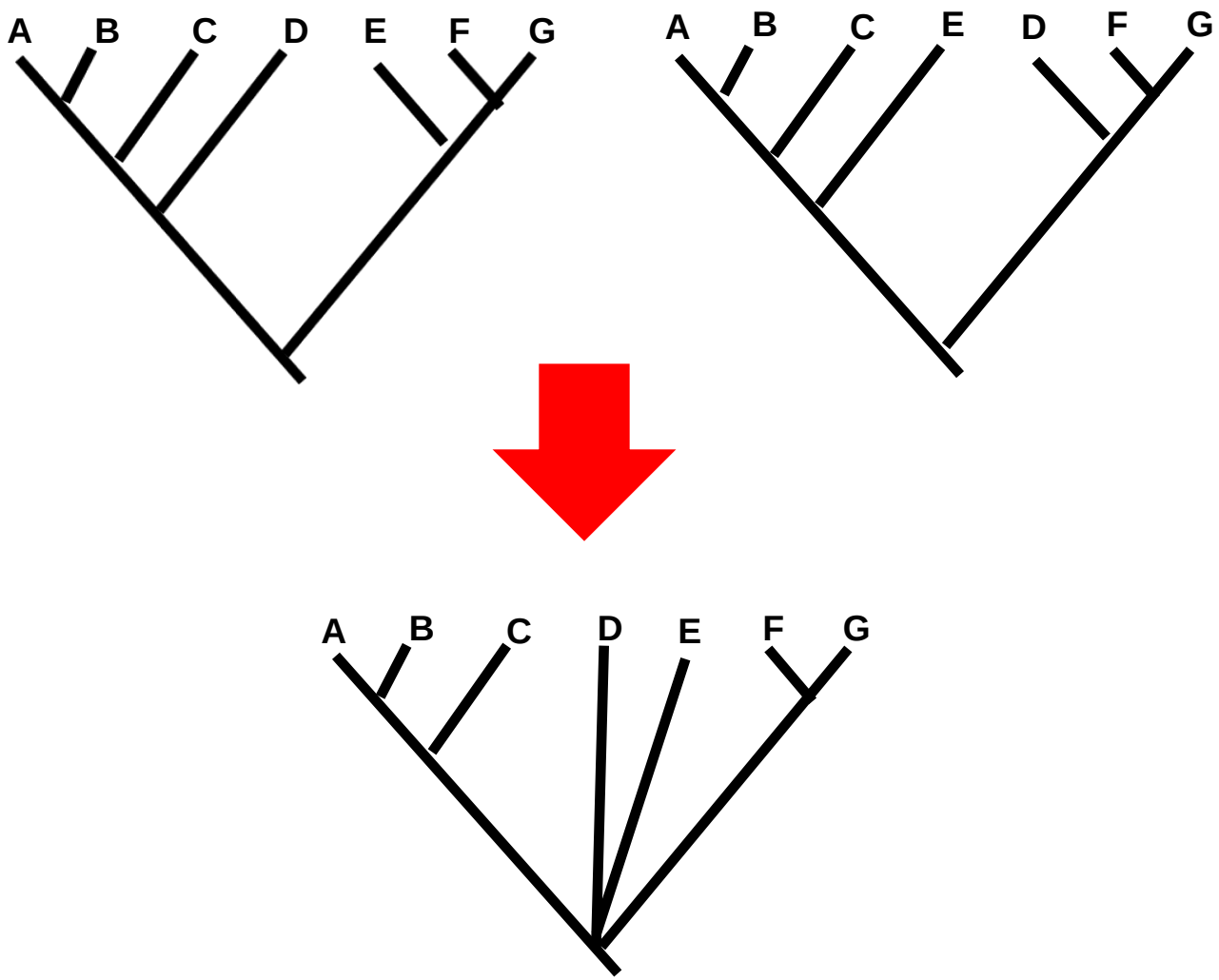
Lo spazio degli alberi è popolato da minimi locali!



E' necessario verificare la topologia finale esplorando lo spazio a partire da più punti

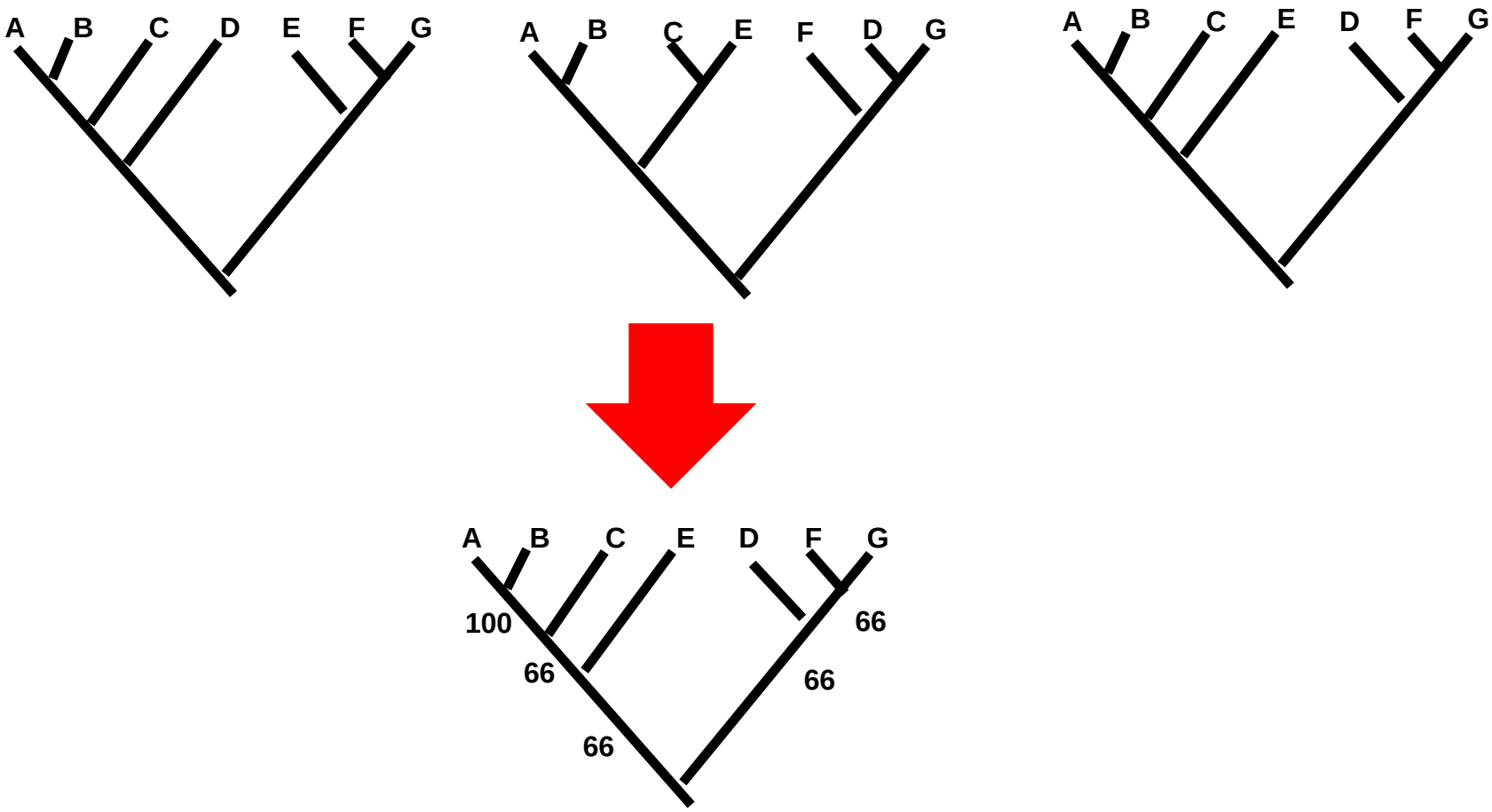
Come riassumere tanti alberi ugualmente parsimoniosi?

Il consenso stretto tra due alberi ugualmente parsimoniosi



Come riassumere tanti alberi ugualmente parsimoniosi

Il consenso con majority rule tra tre alberi ugualmente parsimoniosi



I numeri sull'albero indicano la frequenza dei clade negli alberi ugualmente parsimoniosi

Metodi di massima verosimiglianza

- Il modello di evoluzione deve essere esplicito (per esempio, rapporto tra trasversioni e transizioni, distribuzione dei tassi di mutazione lungo la sequenza, ecc.)

- Si basano sul concetto statistico di verosimiglianza (il cui massimo diventa il criterio per questi metodi)

Concetti di verosimiglianza

Verosimiglianza di un ipotesi = L (Likelihood) = Probabilità di osservare i dati D se è vera l'ipotesi H

$$L(H|D) = Prob(D|H) = P(D|H)$$

[Il modello specifica la funzione P(D|H)]

L'ipotesi che possiede la maggiore L, ossia il maggior valore di verosimiglianza, è quella da preferire.

Concetti di verosimiglianza

Supponiamo di lanciare una moneta e ottenere testa. La probabilità di osservare i dati (la moneta caduta con il lato testa verso l'alto) varia con le ipotesi che posso fare sulla moneta stessa.

Se la moneta ha una testa e una croce, ed è equilibrata

$$\rightarrow \quad P(D|H_1) = 0.5$$

Se la moneta ha due teste

$$\rightarrow \quad P(D|H_2) = 1$$

DIVERSE IPOTESI POSSONO DETERMINARE GRANDI VARIAZIONI SULLA PROBABILITA' DI OSSERVARE I DATI

La verosimiglianza applicata a RC

D = sequenze

H = topologia e lunghezza dei rami

Modello = modello evolutivo di sostituzione

Devo trovare la topologia (con lunghezza dei rami) che massimizza la probabilità di generare le sequenze che osservo,

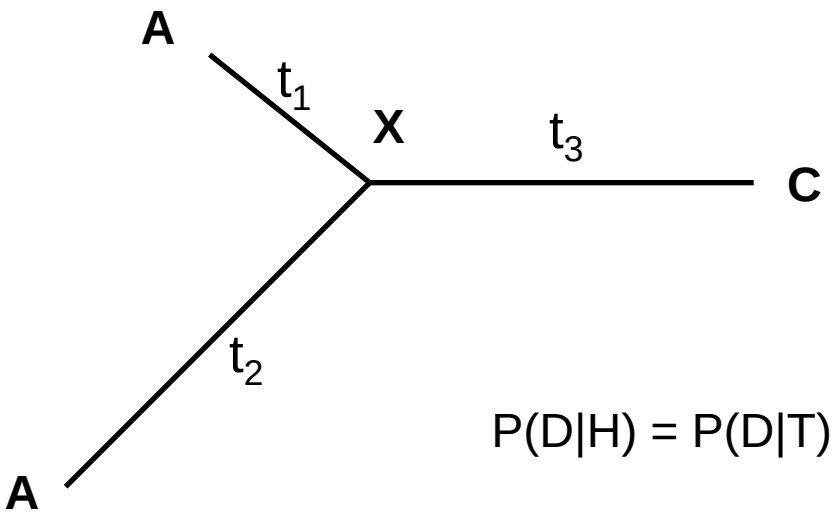
cioè l'albero con la massima verosimiglianza (likelihood)

Attenzione a non confondere likelihood con probabilità. La likelihood di un albero non è la probabilità che un albero sia quello vero [P(H|D)], ma la probabilità che quell'albero abbia originato i dati che osserviamo [P(D|H) = L(H|D)].

- *Le probabilità sommano sempre a 1, non le likelihood*

- *Il calcolo delle P(H|D) si fanno con metodi bayesiani, non di massima verosimiglianza*

La verosimiglianza applicata a RC



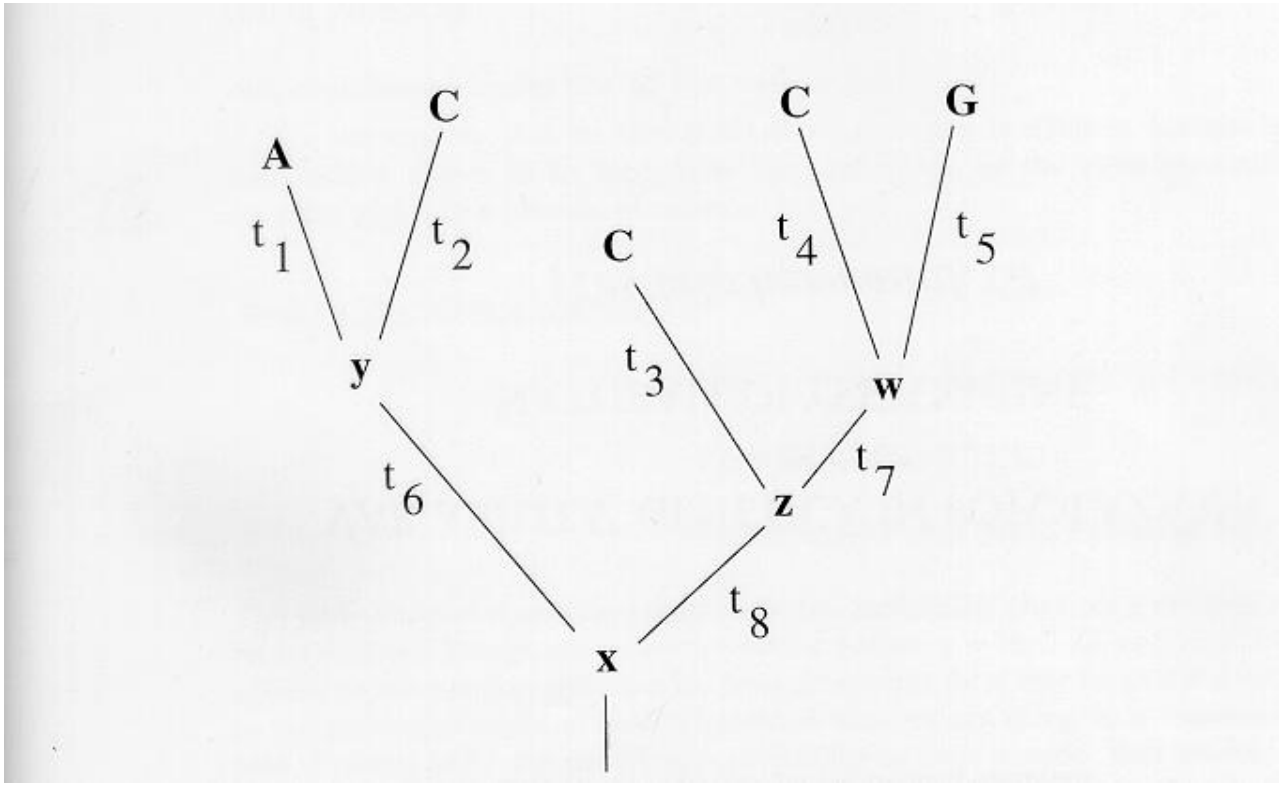
P(D|H) = P(D|T) =

$$\sum_{Tutti\ X} P(A, A, C, X \mid T)$$

se X = A $P(A|A,t_1)*P(A|A,t_2)*P(C|A,t_3)$
se X = C $P(A|C,t_1)*P(A|C,t_2)*P(C|C,t_3)$
se X = T $P(A|T,t_1)*P(A|T,t_2)*P(C|T,t_3)$
se X = G $P(A|G,t_1)*P(A|G,t_2)*P(C|G,t_3)$

Per i vari t (lunghezza dei rami) si parte di solito da valori plausibili per poi modificarli. Gradualmente, ottenendo per ogni topologia i valori dei vari t che massimizzano la verosimiglianza

La verosimiglianza applicata a RC



P(D|T) =

$$\sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w \mid T) =$$

$$\sum_x \sum_y \sum_z \sum_w P(x)P(y \mid x,t_6)P(A \mid y,t_1)P(C \mid y,t2).....$$

I metodi di massima verosimiglianza

Vantaggi

- Utilizzano tutti i dati (non solo i siti informativi!)
- Permettono di testare statisticamente diverse ipotesi
- Sono spesso superiori a gli altri metodi (se il modello evolutivo non è troppo diverso dalla realtà)
- Permettono di confrontare diversi modelli evolutivi sulla base di un albero prefissato (problemi di circolarità??)

Svantaggi

- Problemi se il modello è scorretto
- Lenti

Il bootstap per testare la robustezza di un albero (o parte di esso)

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

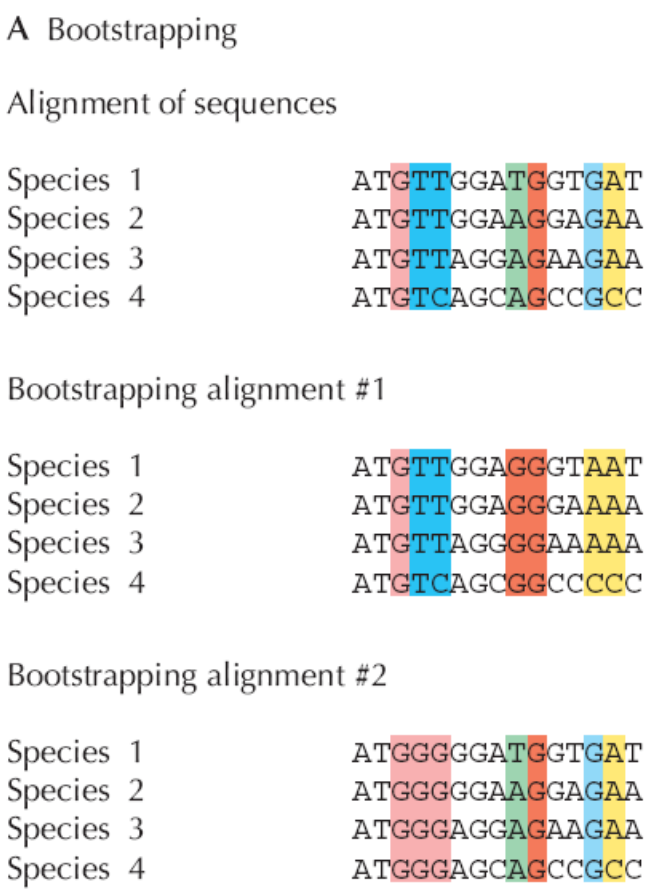
• I caratteri (colonne in un allineamento di seqeunze) sono estratte con rimpiazzo per generare molti (almeno 1000) pseudo data set

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

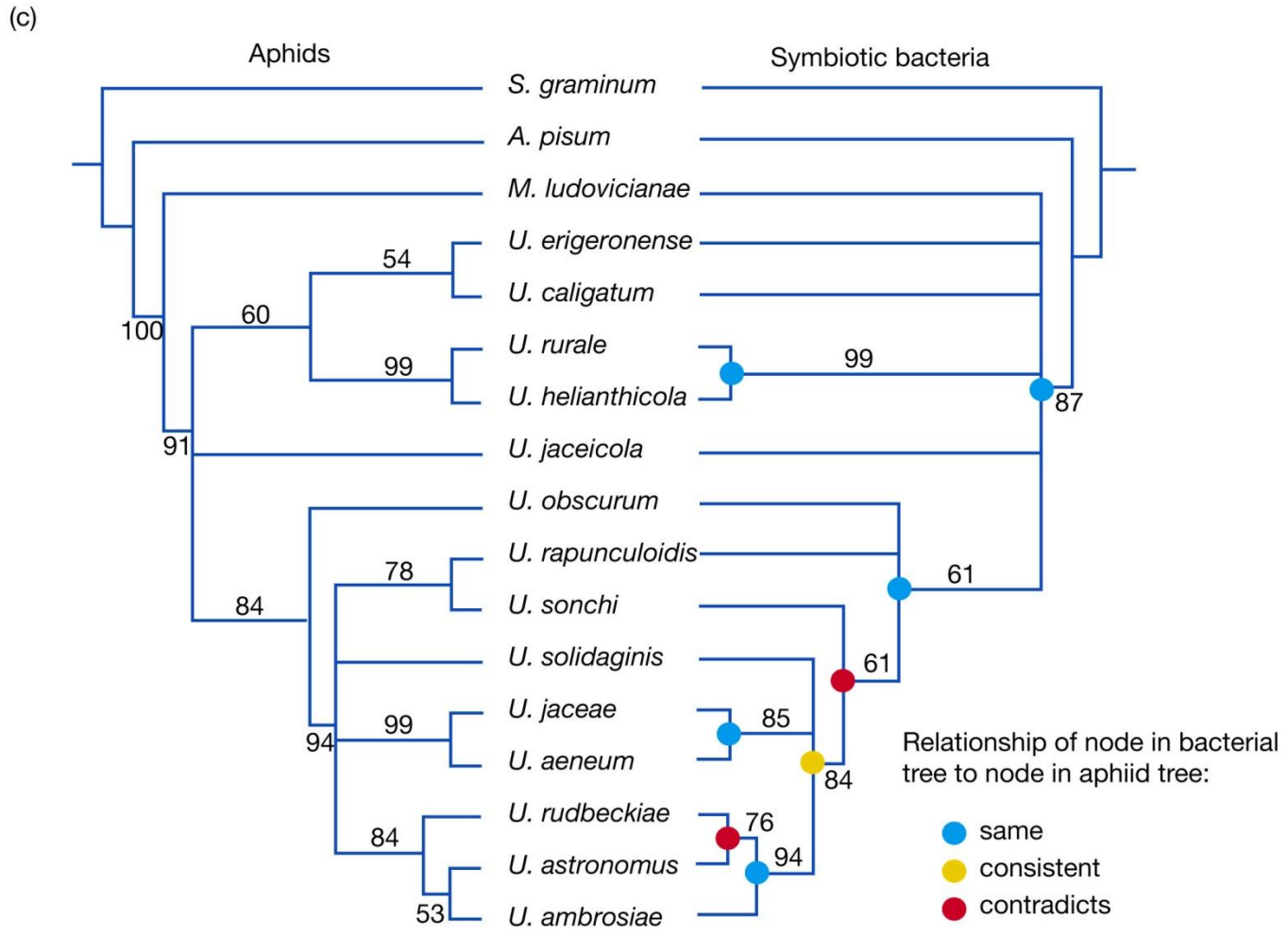
• L'albero che sintetizza i (per esempio 1000) data set viene costruito di solito con il metodo del majority rule consensus

• 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

Il bootstap per testare la robustezza di un albero (o parte di esso)



Il bootstap per testare gruppi e cospeciazione



Interpretare i valori di bootstrap

• Non semplice interpretare il valore di boostrap

• Si può dire che valori superiori all’80% indicano un supporto molto forte

• Anche valori superiori al 50% indicano comunque che un gruppo è presente frequentemente negli pseudo data set (altre combinazioni hanno valori sicuramente molto più bassi)

• Un supporto basso non indica che il clade è sbagliato, ma solo che il supporto statistico è basso

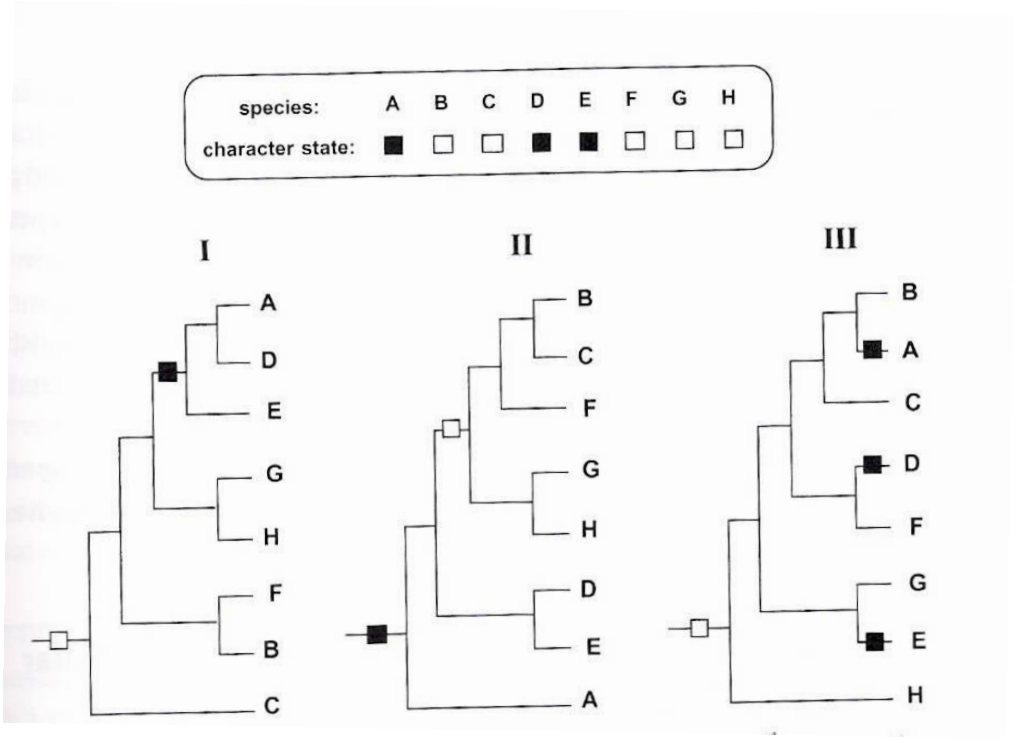
Quale metodo conviene utilizzare?

- I metodi devono essere confrontati sulla base di
  - Dati veri dei quali si conosce l'albero esatto (ma solo le topologie ovvie sono note, e per queste tutti i metodi funzionano bene)
  - Dati simulati (si conosce ovviamente l'albero vero, ma bisogna fare moltissime simulazioni e le analisi con alcuni metodi sono molto lente)

- In generale, NJ e MP sono simili, o meglio ognuno leggermente meglio dell'altro in situazioni diverse. ML tende ad essere quello che ritrova più frequentemente, nelle condizioni considerate da questi studi, la vera filogenesi.

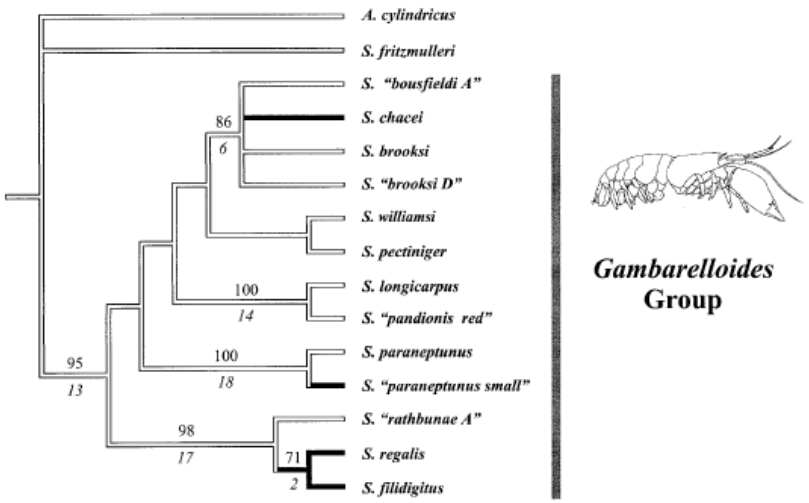
- Spesso si utilizzano tutti i metodi e si considera "robusto" un raggruppamento quando viene identificato da tutti i metodi

Se ho una filogenesi di specie, posso studiare *altri* caratteri



PCM = Phylogenetic Character Mapping

Esempio: evoluzione eusocialità in gamberetti



Gambarelloides Group

FIG. 2. Consensus of two most parsimonious trees resulting from the analysis of the combined morphological and molecular data. Symbols as in Figure 1.

Evolution, 54(2), 2000, pp. 503–516

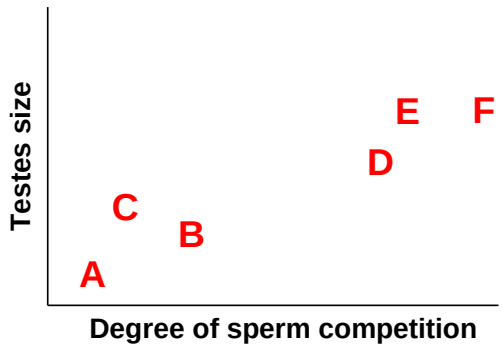
MULTIPLE ORIGINS OF EUSOCIALITY AMONG SPONGE-DWELLING SHRIMPS (*SYNALPHEUS*)

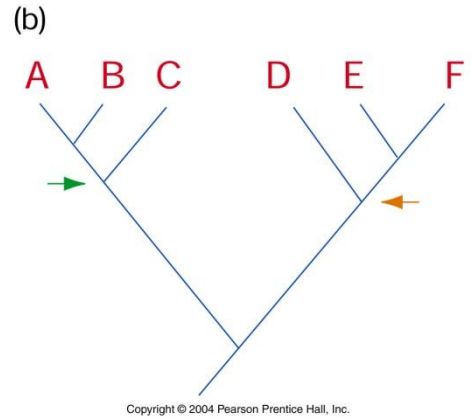
J. EMMETT DUFFY,[1] CHERYL L. MORRISON, AND RUBÉN RÍOS
College of William and Mary, School of Marine Science, Gloucester Point, Virginia 23062-1346
[1]*E-mail: jeduffy@vims.edu*

Abstract.—As the most extreme expression of apparent altruism in nature, eusociality has long posed a central paradox for behavioral and evolutionary ecology. Because eusociality has arisen rarely among animals, understanding the selective pressures important in early stages of its evolution remains elusive. Employing a historical approach to this problem, we used morphology and DNA sequences to reconstruct the phylogeny of 13 species of sponge-dwelling shrimps (*Synalpheus*) with colony organization ranging from asocial pair-bonding through eusociality. We then used phylogenetically independent contrasts to test whether sociality was associated with evidence of enhanced competitive ability, as suggested by hypotheses invoking an advantage of cooperation in crowded habitats. The molecular, morphological, and combined data each strongly supported three independent origins of monogynous, multigenerational (eusocial) colony organization within this genus. Phylogenetically independent contrasts confirmed that highly social taxa, with strong reproductive skew, have significantly higher relative abundance within the host sponge than do less social taxa, a result that was robust to uncertainty in tree topology and varying models of character change. A similar tendency for highly social species to share their sponge with fewer congener species was suggestive, but not significant. Because unoccupied habitat appears to be limiting for many sponge-dwelling shrimp species, these data are consistent with hypotheses that cooperative social groups enjoy a competitive advantage over less organized groups or individuals, where independent establishment is difficult, and that enemy pressure is of central importance in the evolution of animal sociality.

Eusocialità in questi gamberetti si è evoluta 3 volte, favorita dalla competizione

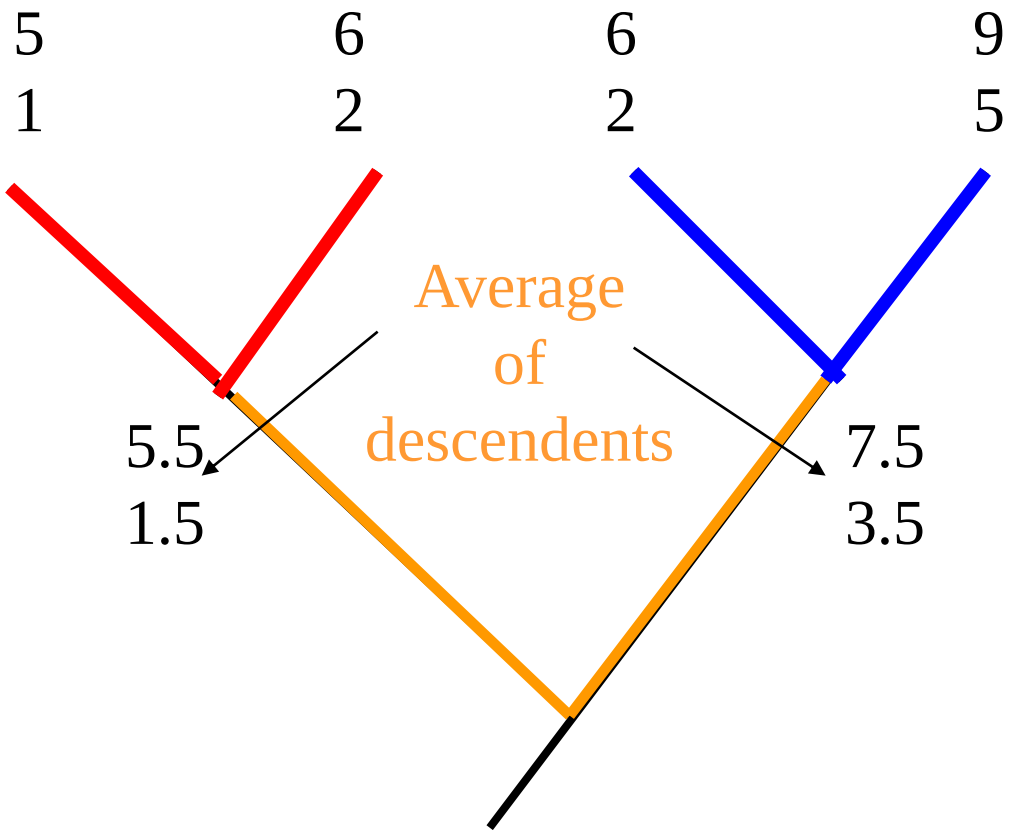
Contrasti indipendenti: un accenno



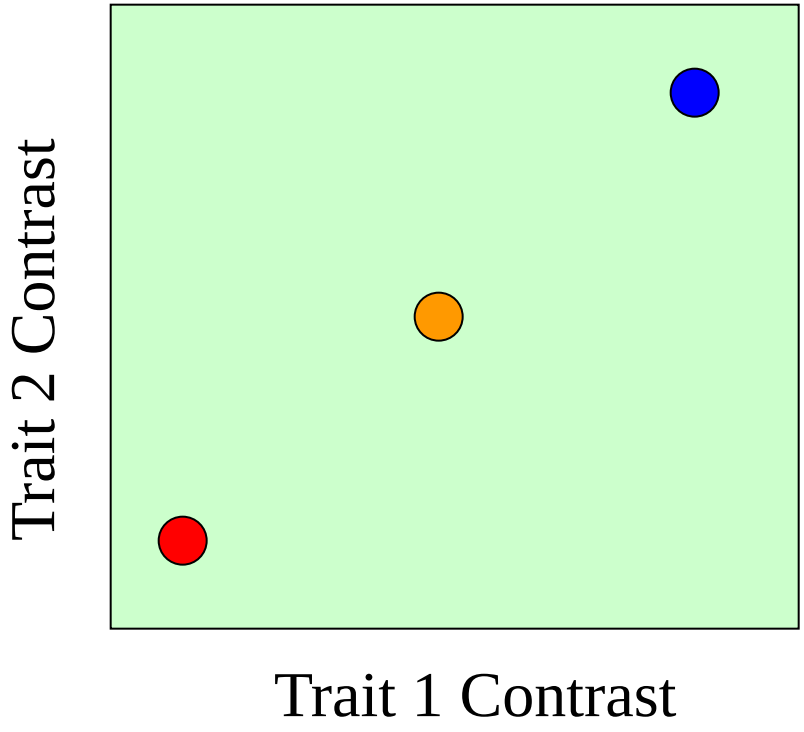


La correlazione semplice tra due tratti non significa molto se esiste forte correlazione filogenetica tra specie. D,E e F possono avere evoluto una sola volta i testicoli più grandi

Contrasti indipendenti: un accenno

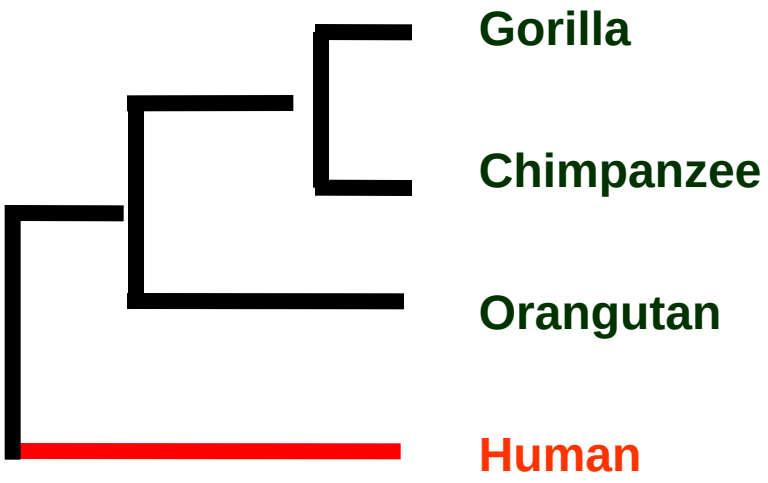


Trait 1: 7.5-5.5=2
Trait 2: 3.5-1.5=2
contrast: (2,2)

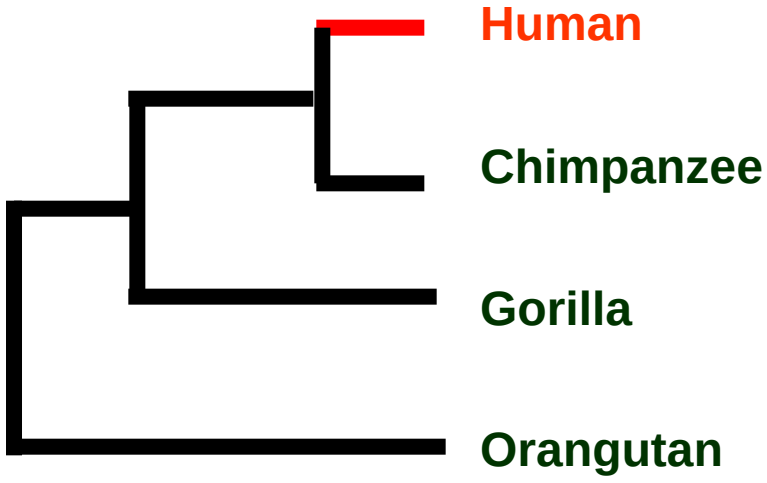


Album di famiglia: a chi siamo più vicini filogeneticamente?

Un albero filogenetico è un'ipotesi tra tante possibili

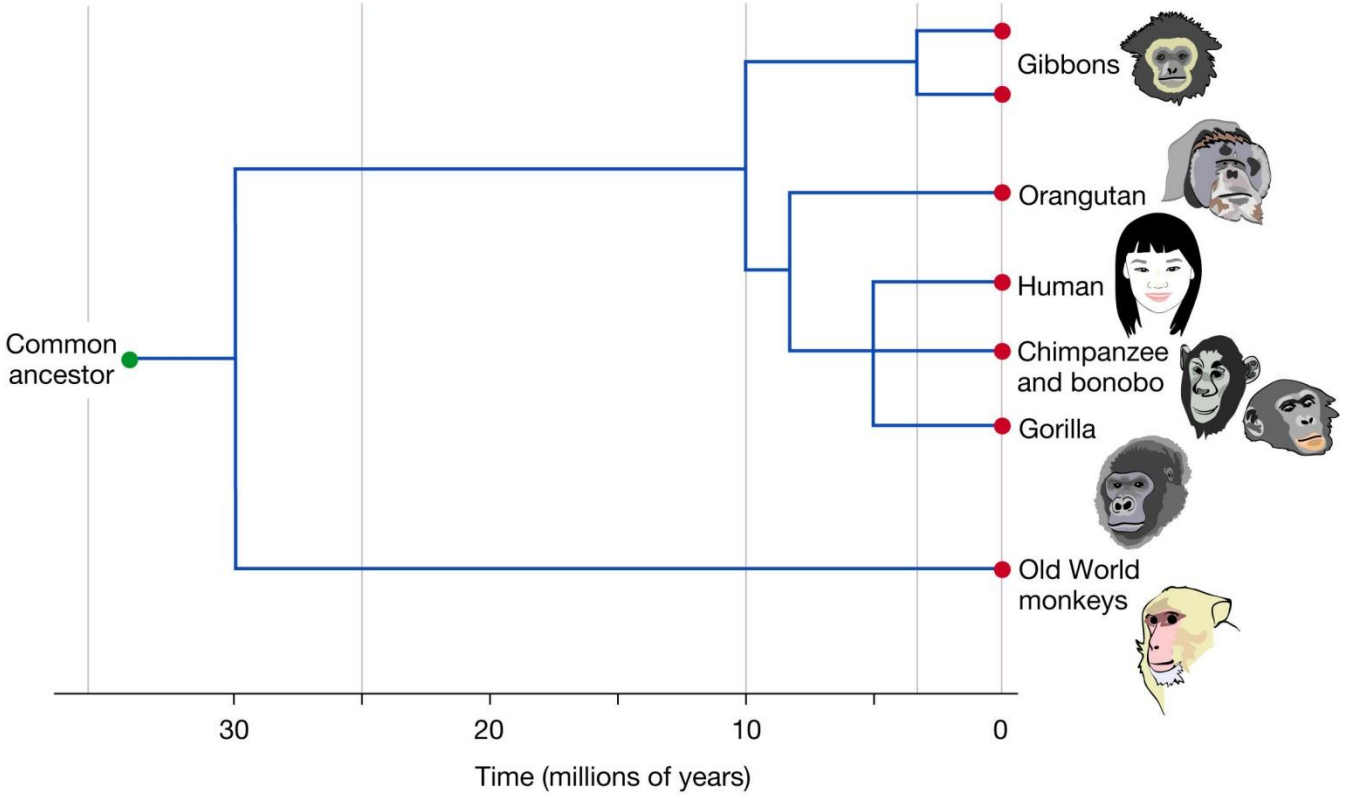


Analisi fossili (fino anni 60). Grande differenza tra uomo e altri primati e separazione antica (>15MY)



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

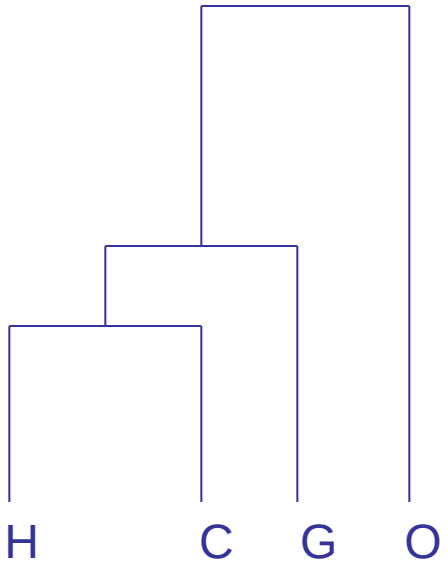
In realtà la tricotomia non è stata facile da risolvere



Sequenze di DNA

	H-C	H-G	C-G	H-O	C-O	G-O
Average divergence at non-coding sites (autosomal)	1.24%	1.62%	1.63%	3.08%	3.12%	3.09%

Suggests:



Am. J. Hum. Genet. 68:444–456, 2001

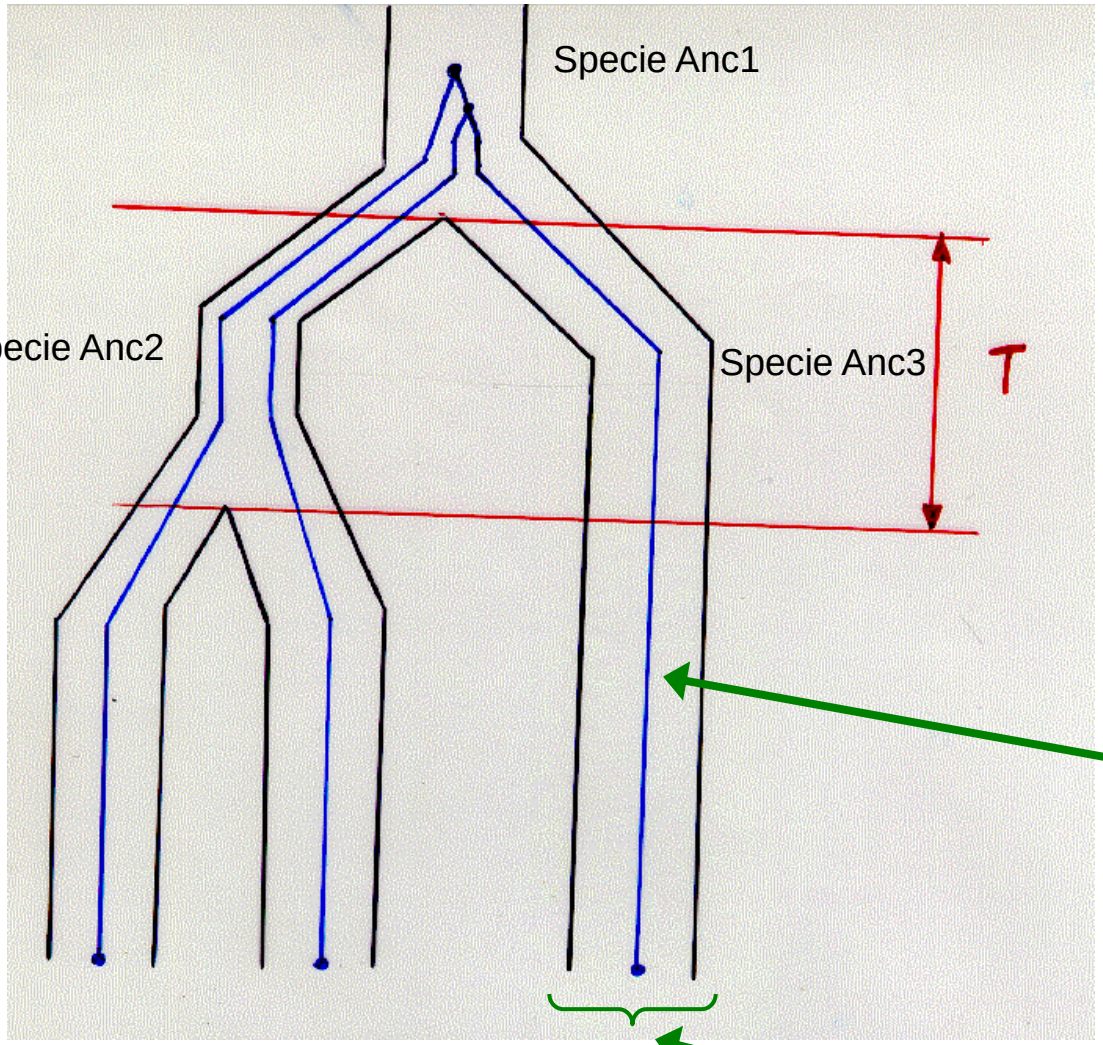
Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees

Feng-Chi Chen[1,*] and Wen-Hsiung Li[2]

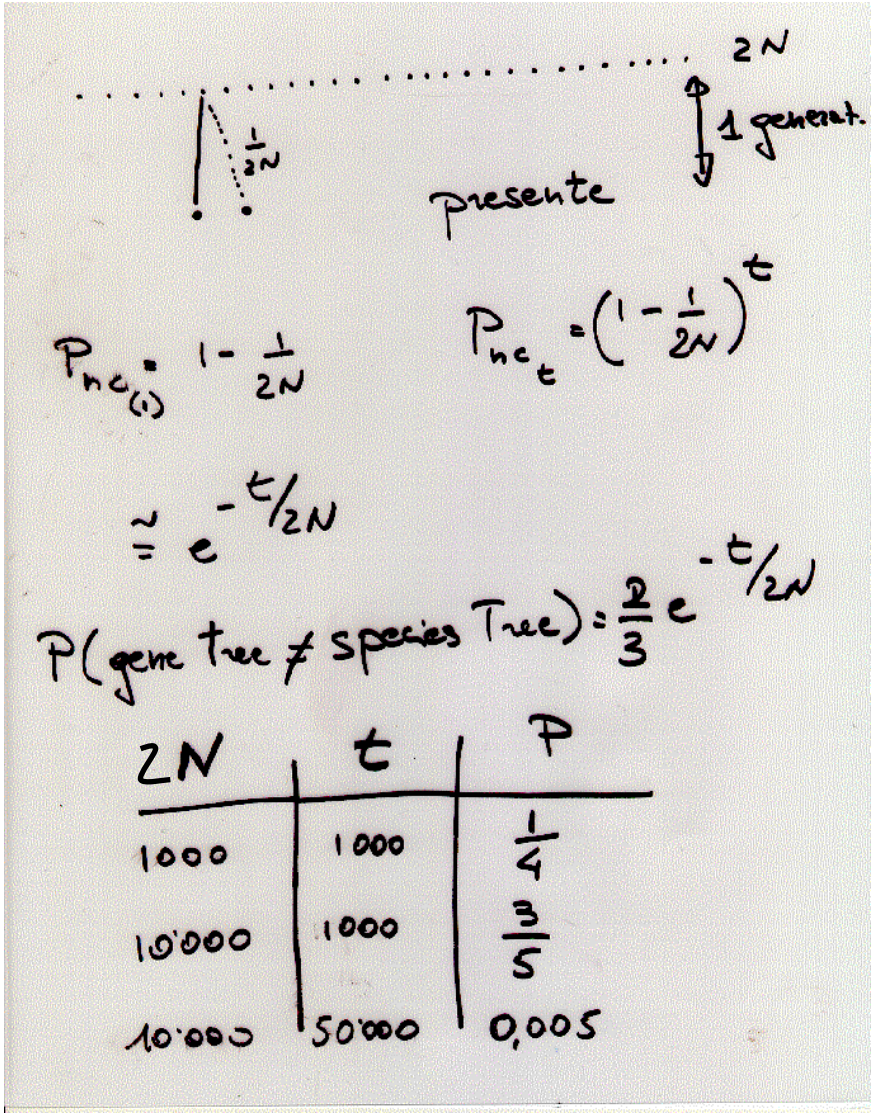
[1]Department of Life Science, National Tsing Hua University, Taiwan, and [2]Department of Ecology and Evolution, University of Chicago, Chicago

To study the genomic divergences among hominoids and to estimate the effective population size of the common ancestor of humans and chimpanzees, we selected 53 autosomal intergenic nonrepetitive DNA segments from the human genome and sequenced them in a human, a chimpanzee, a gorilla, and an orangutan. The average sequence divergence was only 1.24% ± 0.07% for the human-chimpanzee pair, 1.62% ± 0.08% for the human-gorilla pair, and 1.63% ± 0.08% for the chimpanzee-gorilla pair. These estimates, which were confirmed by additional data from GenBank, are substantially lower than previous ones, which included repetitive sequences and might have been based on less-accurate sequence data. The average sequence divergences between orangutans and humans, chimpanzees, and gorillas were 3.08% ± 0.11%, 3.12% ± 0.11%, and 3.09% ± 0.11%, respectively, which also are substantially lower than previous estimates. The sequence divergences in other regions between hominoids were estimated from extensive data in GenBank and the literature, and *Alu*s showed the highest divergence, followed in order by Y-linked noncoding regions, pseudogenes, autosomal intergenic regions, X-linked noncoding regions, synonymous sites, introns, and nonsynonymous sites. The neighbor-joining tree derived from the concatenated sequence of the 53 segments—24,234 bp in length—supports the *Homo-Pan* clade with a 100% bootstrap value. However, when each segment is analyzed separately, 22 of the 53 segments (~42%) give a tree that is incongruent with the species tree, suggesting a large effective population size (N_e) of the common ancestor of *Homo* and *Pan*. Indeed, a parsimony analysis of the 53 segments and 37 protein-coding genes leads to an estimate of N_e = 52,000 to 96,000. As this estimate is 5 to 9 times larger than the long-term effective population size of humans (~10,000) estimated from various genetic polymorphism data, the human lineage apparently had experienced a large reduction in effective population size after its separation from the chimpanzee lineage. Our analysis assumes a molecular clock, which is in fact supported by the sequence data used. Taking the orangutan speciation date as 12 to 16 million years ago, we obtain an estimate of 4.6 to 6.2 million years for the *Homo-Pan* divergence and an estimate of 6.2 to 8.4 million years for the gorilla speciation date, suggesting that the gorilla lineage branched off 1.6 to 2.2 million years earlier than did the human-chimpanzee divergence.

Alberi di specie e di geni non sempre corrispondono



Alberi di specie e di geni non sempre corrispondono



2N	t	P
1000	1000	1/4
10'000	1000	3/5
10'000	50'000	0,005

Popolazione con N individui e 2N copie di un gene

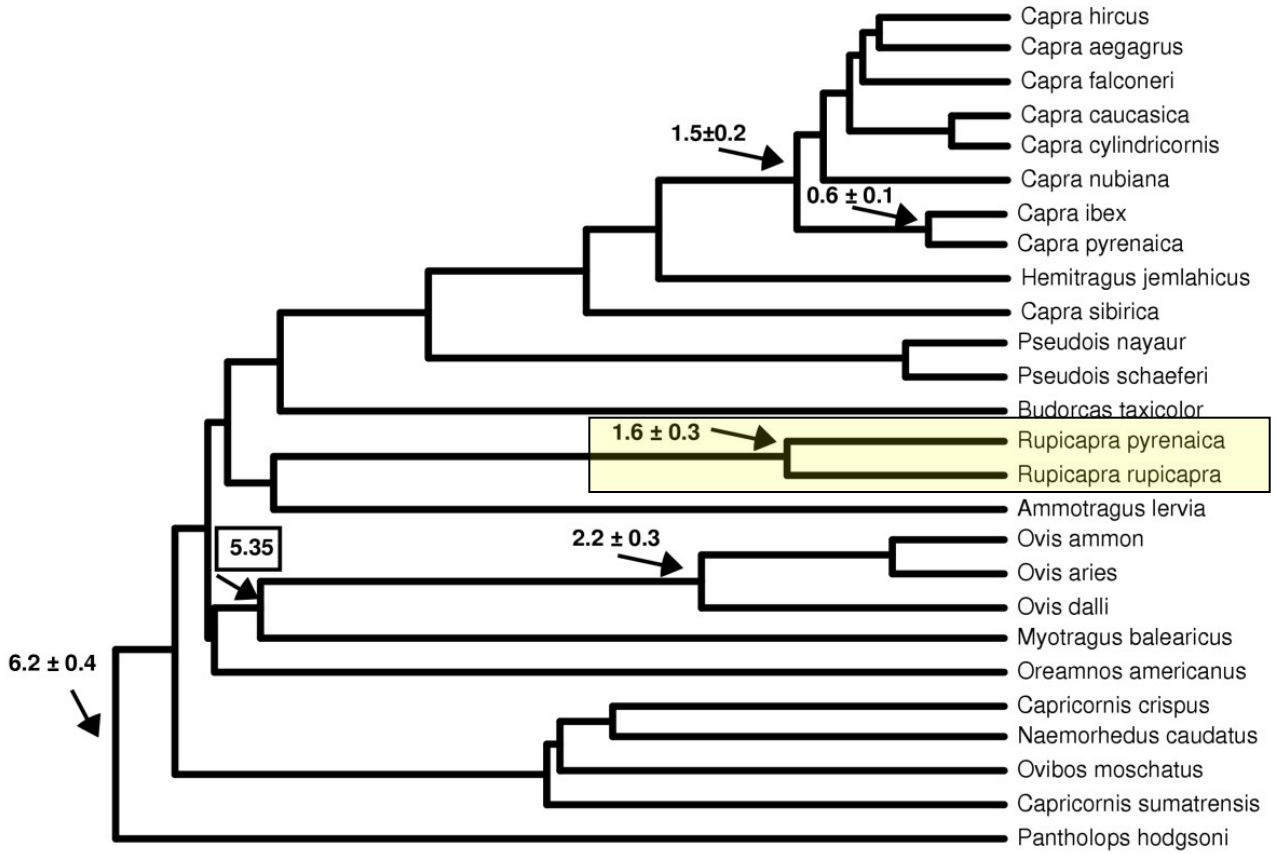
Ogni generazione, la probabilità che due linee (per esempio, due tratti omologhi di DNA provenienti da due individui scelti a caso) derivino da un antenato comune (padre o madre) è data da 1/(2N). E' la probabilità di coalescenza.

La probabilità che non ci sia coalescenza nelle due linee è facile da calcolare, in una o in t generazioni.

Se non ci sono coalescenze in t generazioni nella Specie Anc2, allora ci sono 2 probabilità su 3 che le due linee di questa specie abbiano per caso una coalescenza prima con la linea della Specie 3 che tra di loro. Quindi è possibile calcolare la probabilità che l'albero delle specie e quello del tratto di DNA siano diversi.

Questa probabilità aumenta all'aumentare di N e al diminuire di t, e può essere molto grande.

Datare gli eventi di separazione tra specie



Nuove specie e ibridazione tra specie

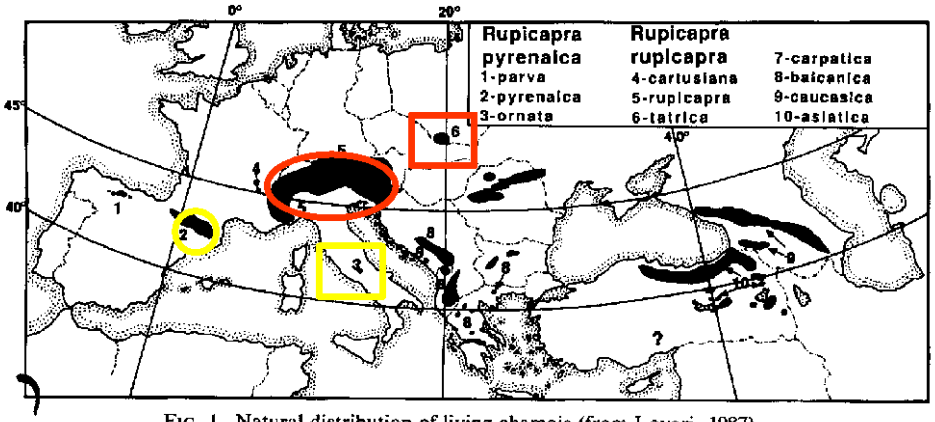
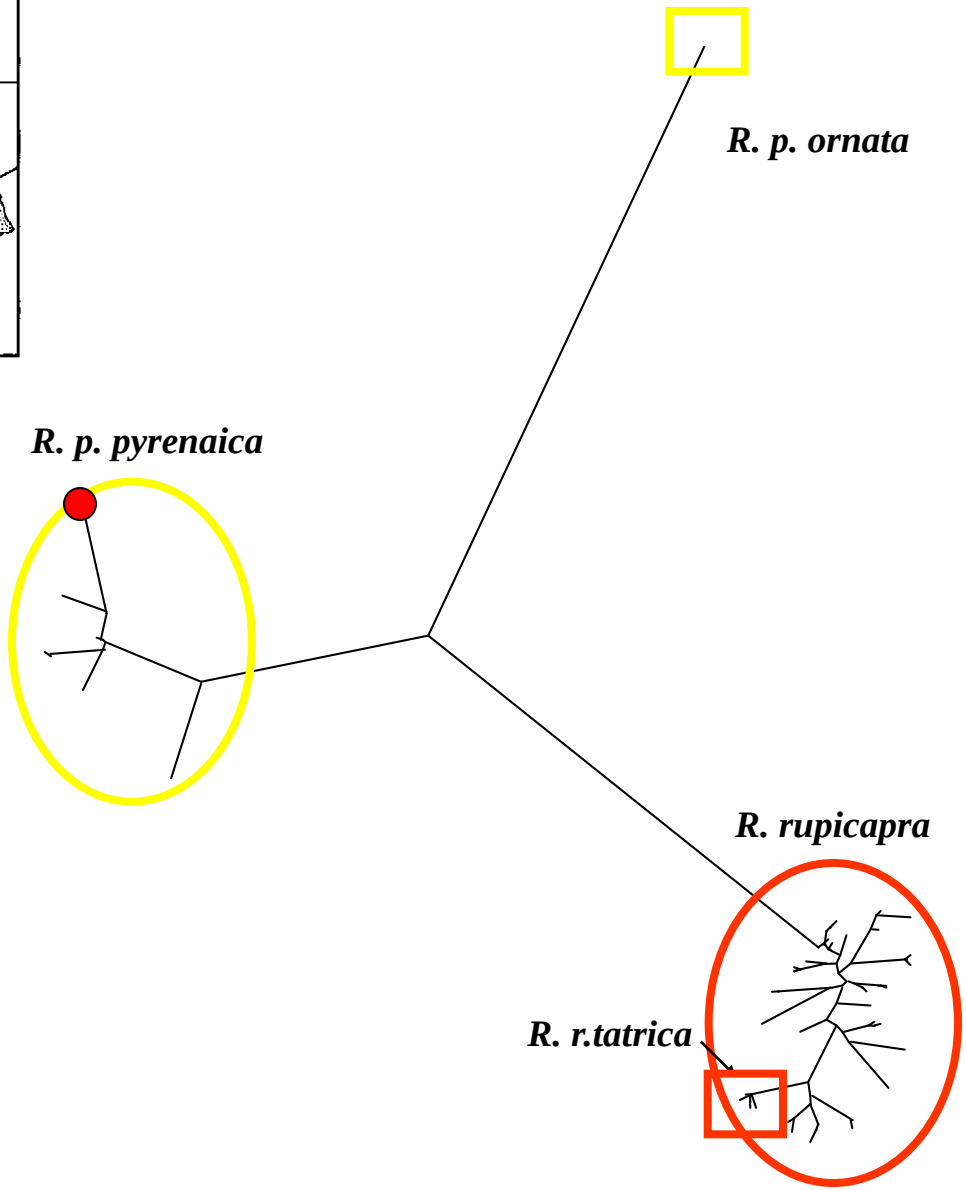


FIG. 1. Natural distribution of living chamois (from Lovari, 1987).



Intermezzo: ancora su filogenesi per testare ipotesi evolutive

Domanda: La deriva dei continenti spiega il pattern di speciazione nei camaleonti?

Chameleon radiation by oceanic dispersal

C. J. Raxworthy*, M. R. J. Forstner† & R. A. Nussbaum‡

* *American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA*
† *Department of Biology, Southwest Texas State University, San Marcos, Texas 78666, USA*
‡ *Museum of Zoology, University of Michigan, Ann Arbor, Michigan 48109-1079, USA*

Historical biogeography is dominated by vicariance methods that search for a congruent pattern of fragmentation of ancestral distributions produced by shared Earth history[1–3]. A focus of vicariant studies has been austral area relationships and the break-up of the supercontinent Gondwana[3–5]. Chameleons are one of the few extant terrestrial vertebrates thought to have biogeographic patterns that are congruent with the Gondwanan break-up of Madagascar and Africa[6,7]. Here we show, using molecular and morphological evidence for 52 chameleon taxa, support for a phylogeny and area cladogram that does not fit a simple vicariant history. Oceanic dispersal—not Gondwanan break-up—facilitated species radiation, and the most parsimonious biogeographic hypothesis supports a Madagascan origin for chameleons, with multiple 'out-of-Madagascar' dispersal events to Africa, the Seychelles, the Comoros archipelago, and possibly Reunion Island. Although dispersal is evident in other Indian Ocean terrestrial animal groups[8–16], our study finds substantial out-of-Madagascar species radiation, and further highlights the importance of oceanic dispersal as a potential precursor for speciation.

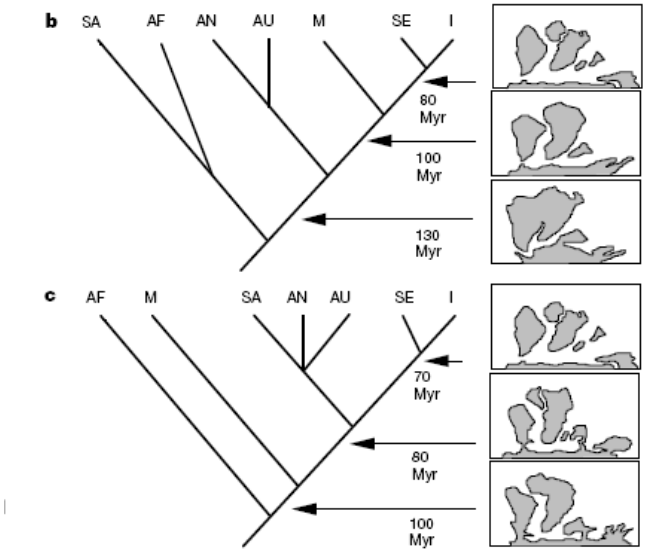
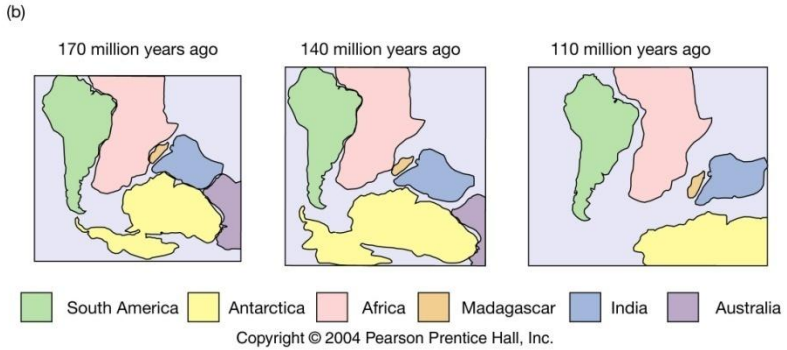


Figure 2 Continental area cladograms for the Indian Ocean region based on chameleon phylogeny and geological break-up models. AF, Africa; AN, Antarctica; AU, Australia; M, Madagascar; SA, South America; SE, the Seychelles; I, India and Sri Lanka. **a**, Chameleon area cladogram based on Fig. 1. **b**, Conventional hypothesis of Gondwanan break-up for the Indian Ocean region[17–19]. **c**, Alternative hypothesis of Gondwanan break-up for the Indian Ocean region proposed by ref. 20. The chameleon area cladogram is incongruent with both of the break-up hypotheses.

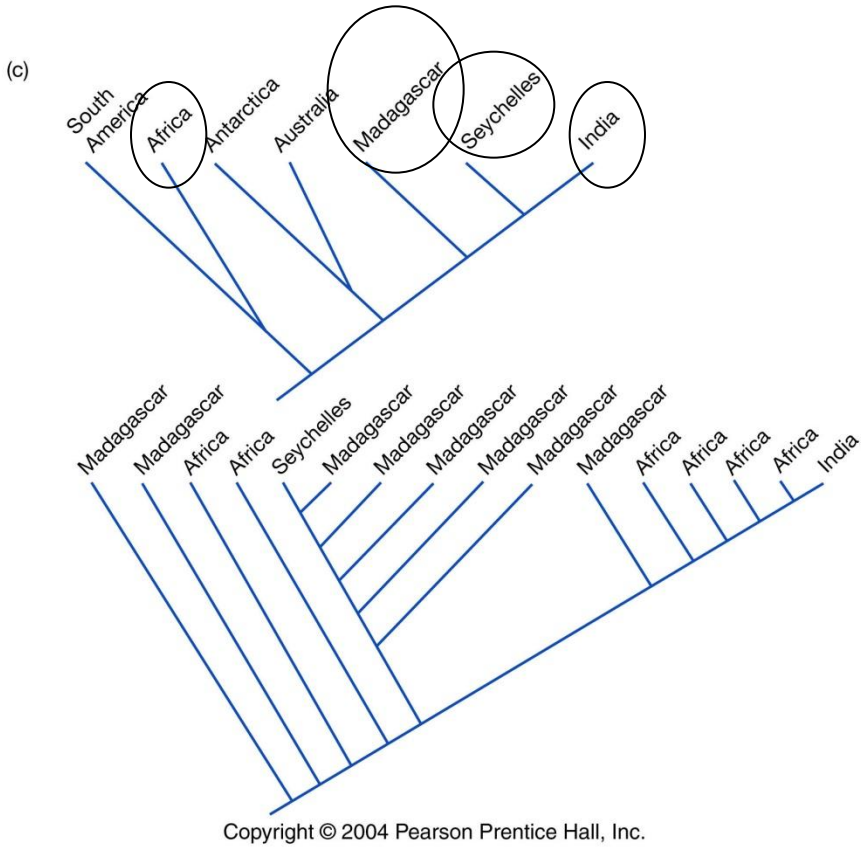
Nature, 2002

Intermezzo: ancora su filogenesi per testare ipotesi evolutive

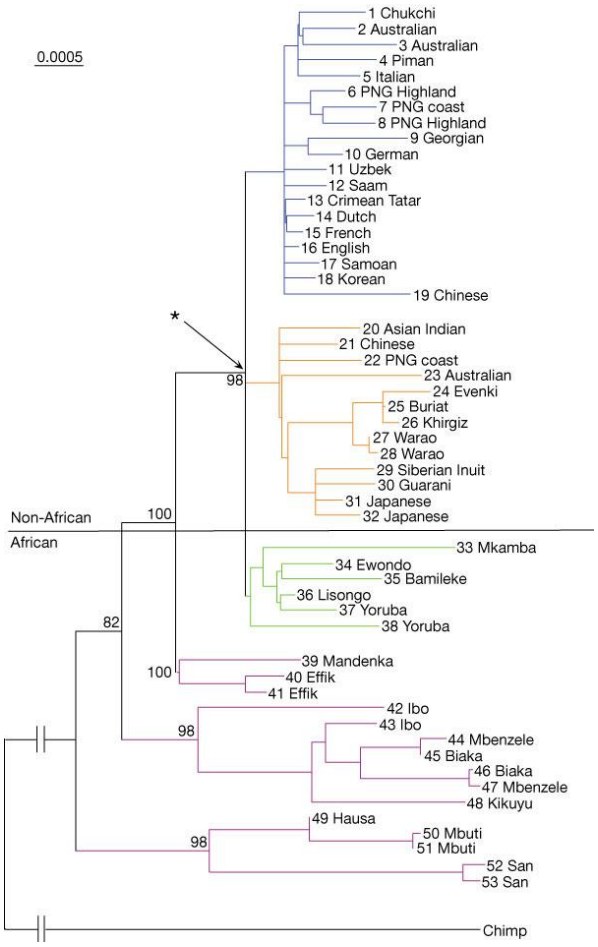
Domanda: La deriva dei continenti spiega il pattern di speciazione nei camaleonti?



- Separation of Gondwanaland
- Upper graph is phylogeny of chameleons based on sequence of separation of southern continents (vicariance hypothesis)
- Lower graph is phylogeny estimated from morphological, behavioral, and molecular data (Raxworthy et al. 2002). This tree implies that chameleons have *dispersed* from Madagascar to Africa on several occasions, from Madagascar to the Seychelles, and from Africa to India
- The dispersal hypothesis is supported by the presence of chameleons on Reunion and the Comoros Is., which are volcanic and have never been in contact with continental land masses.

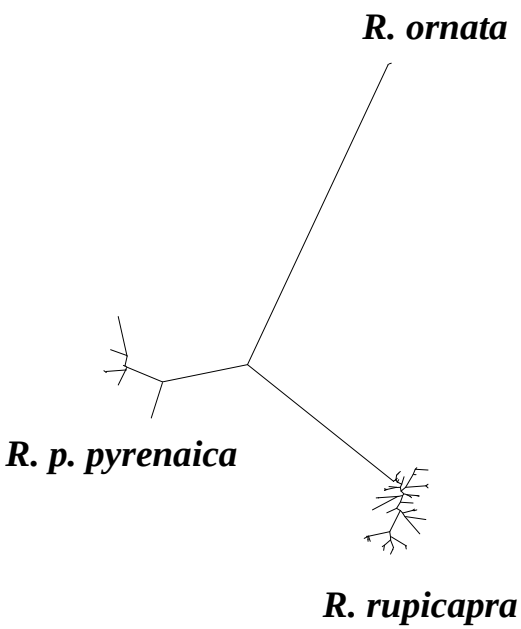


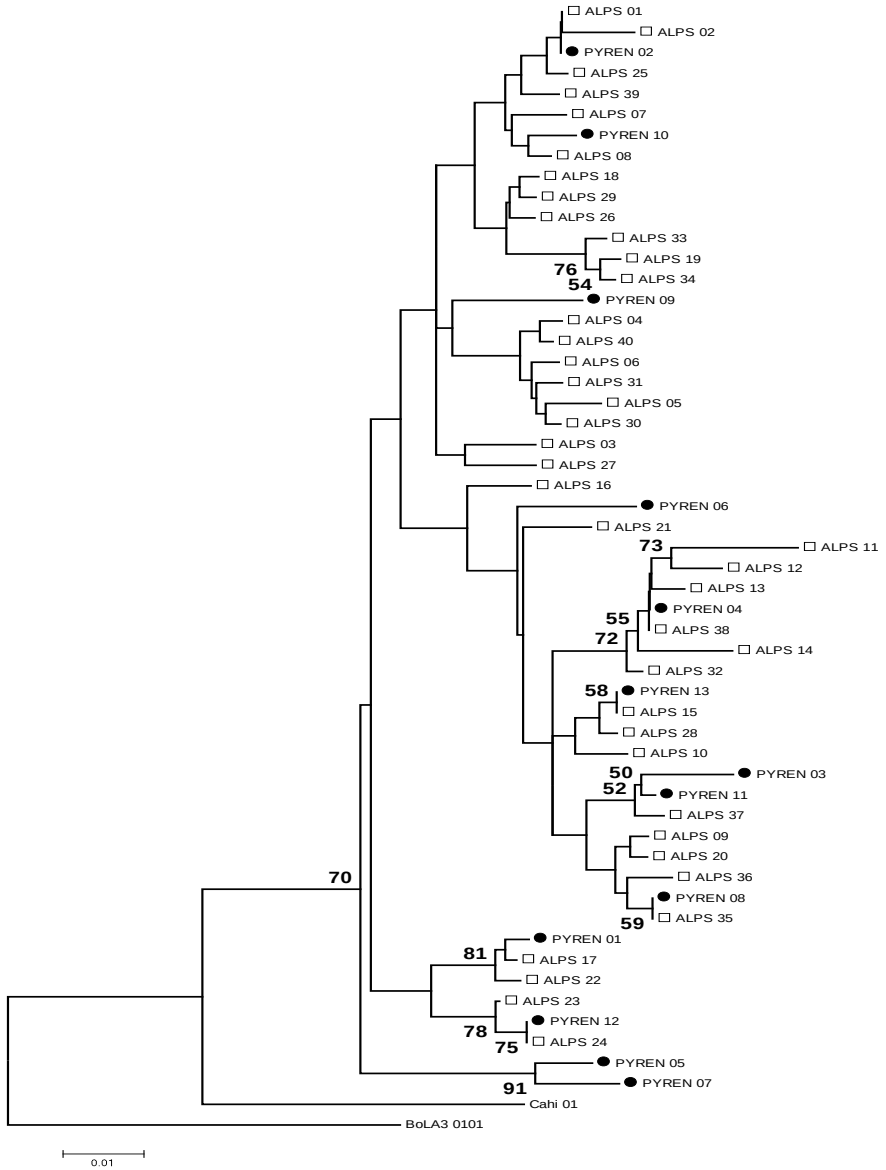
Alberi di popolazioni



Cosa ci dice quest'albero sui liveli di variabilità? E sulla direzione della dispersione? Attenzione alle conseguenze del fatto che le popolazioni possono mescolarsi, e a volte lo fanno

Confronto gli alberi in diversi tratti del genoma

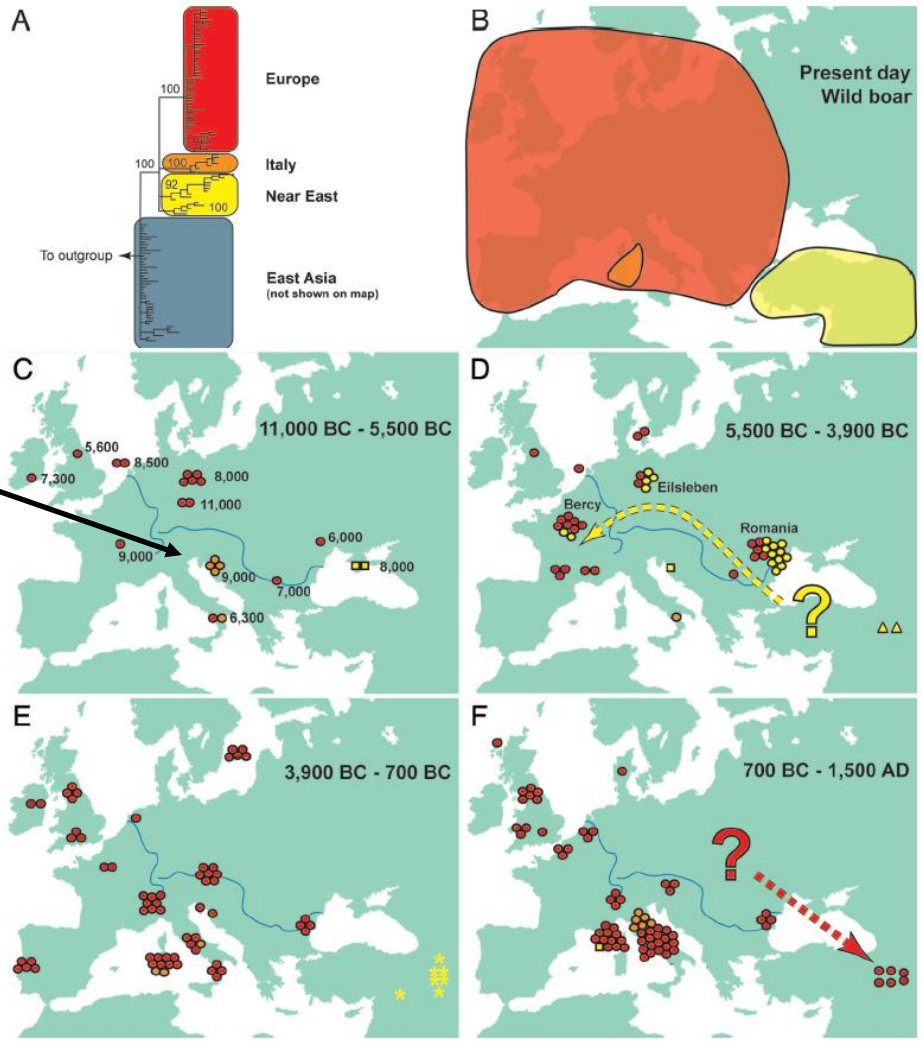




Alberi di aplotipi e DNA antico

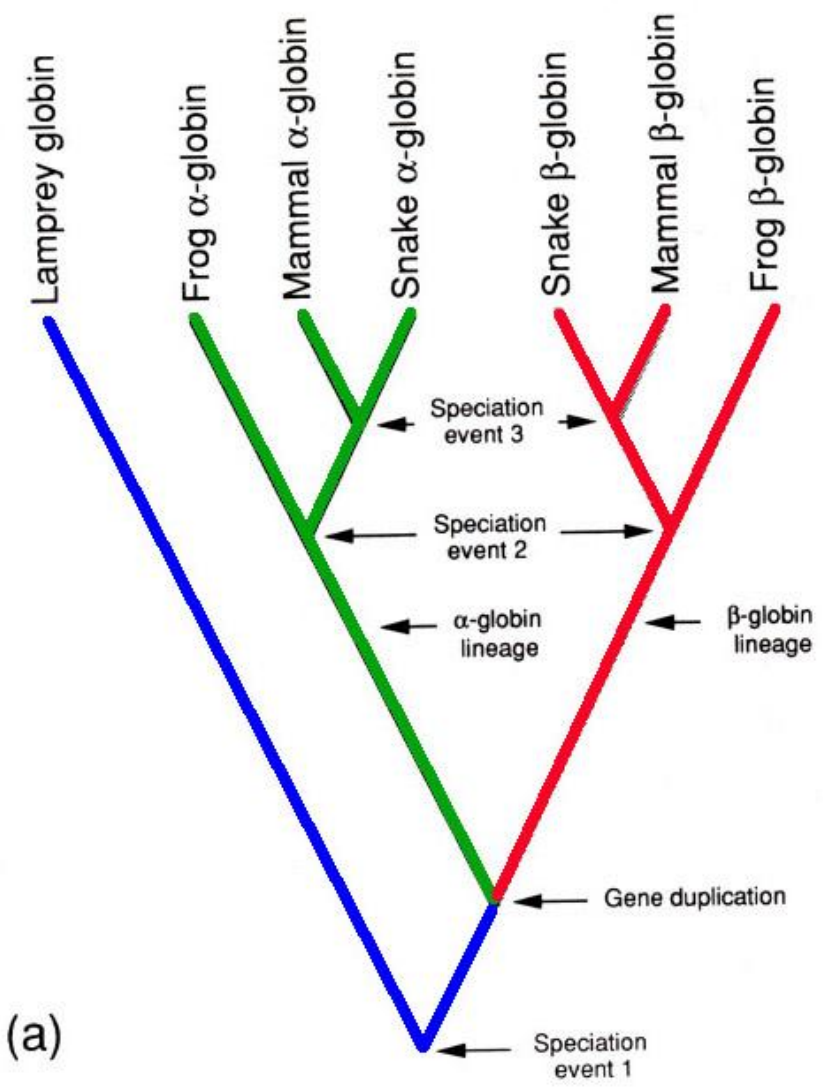
Sito Mesolitico
(Riparo Biarzo, UD)

➔ Due sequenze “gialle” e una “gialla” (se 6 campioni)



Mapping geografico delle sequenze

Gli alberi delle specie e gli alberi dei geni: le emoglobine



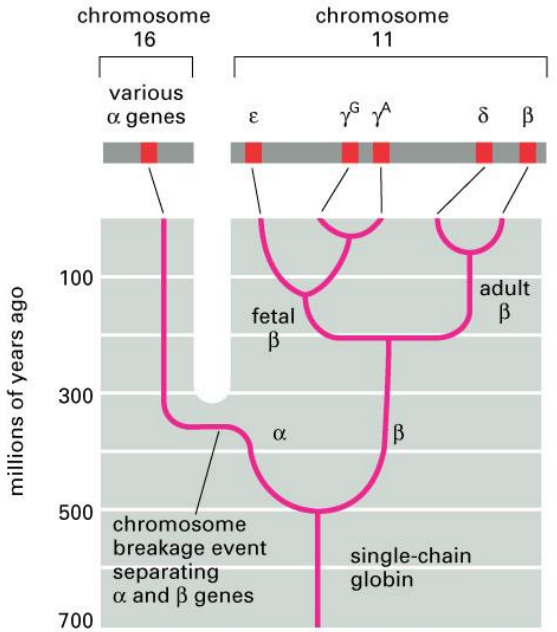
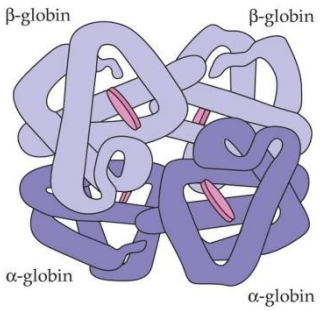
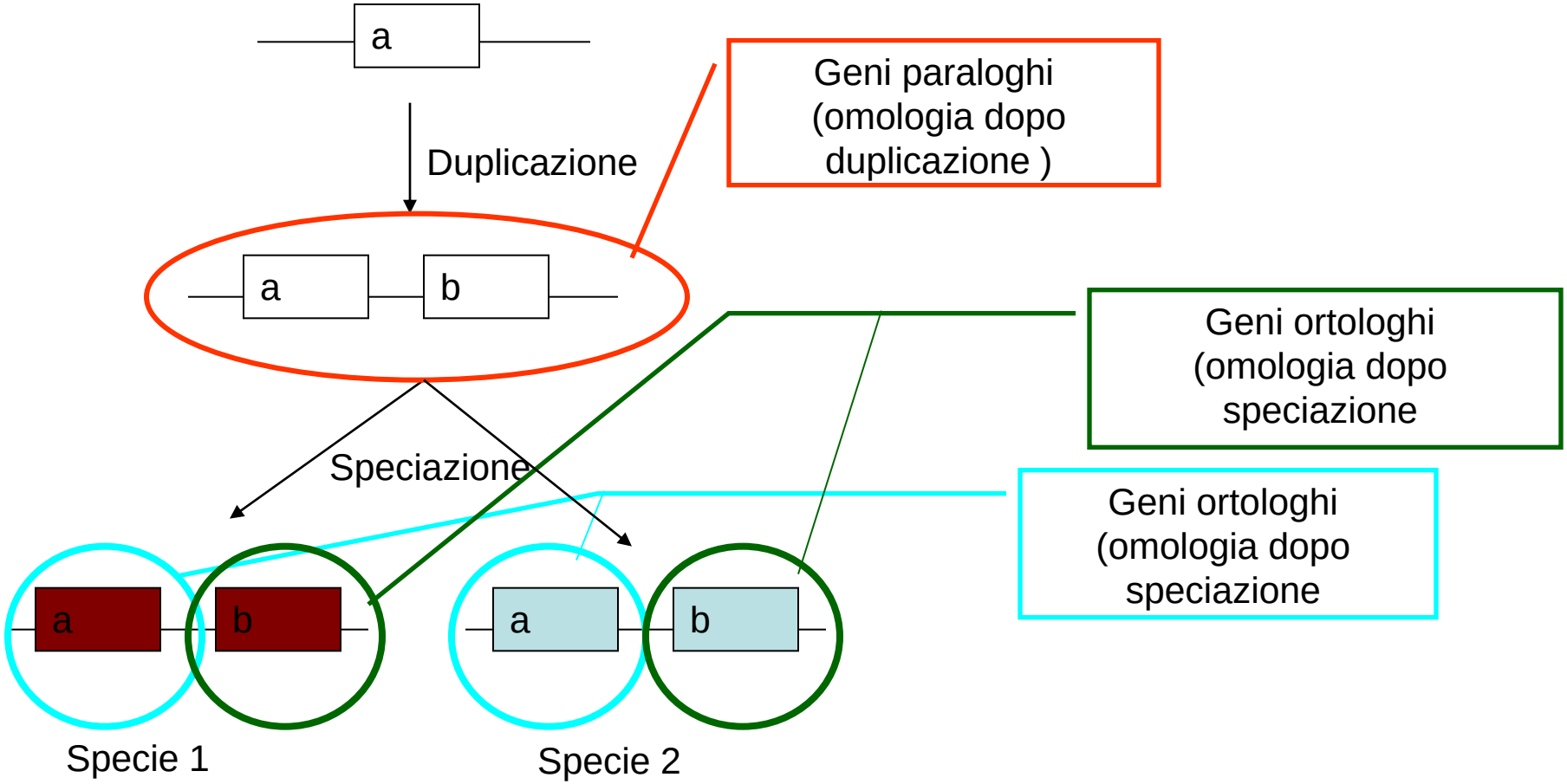


Figure 9-7 Essential Cell Biology, 2/e. (© 2004 Garland Science)

(c) Mammalian adult blood hemoglobin

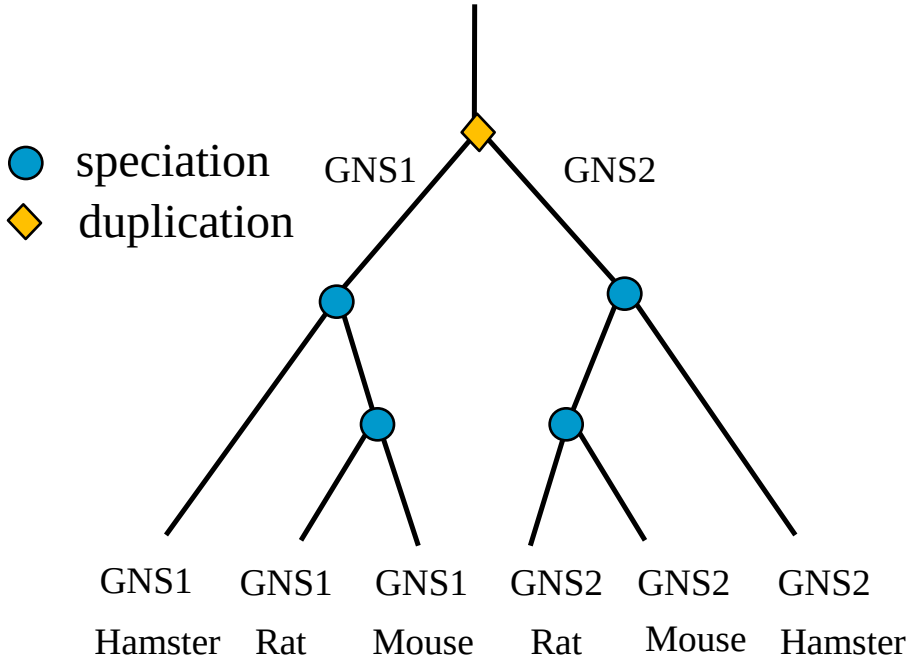


Quando la duplicazione può indurre false filogenesi

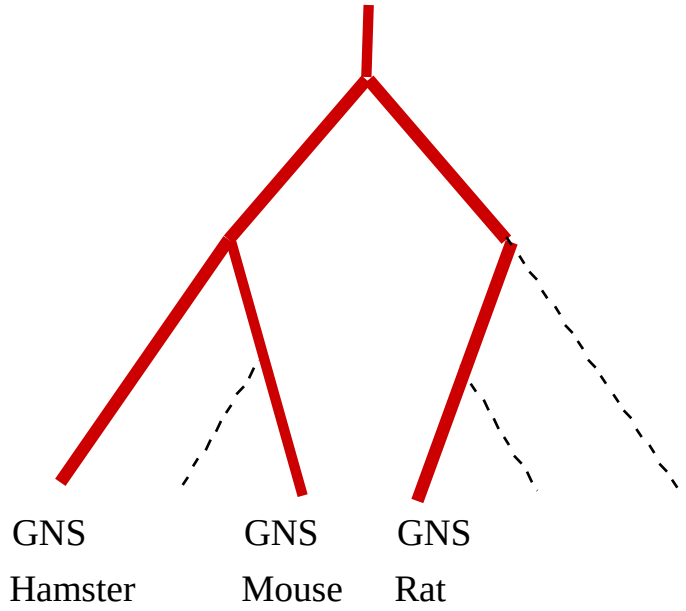


Se siamo in questa situazione, ma analizziamo 3 specie senza sapere che il gene studiato è duplicato (oppure alcuni geni, dopo duplicazione, sono rapidamente “degenerati”) …

Quando la duplicazione può indurre false filogenesi



Albero vero



Albero ricostruito con un campionamento parziale di geni omologhi

!! Gene loss can occur during evolution : even with complete genome sequences it may be difficult to detect paralogy !!