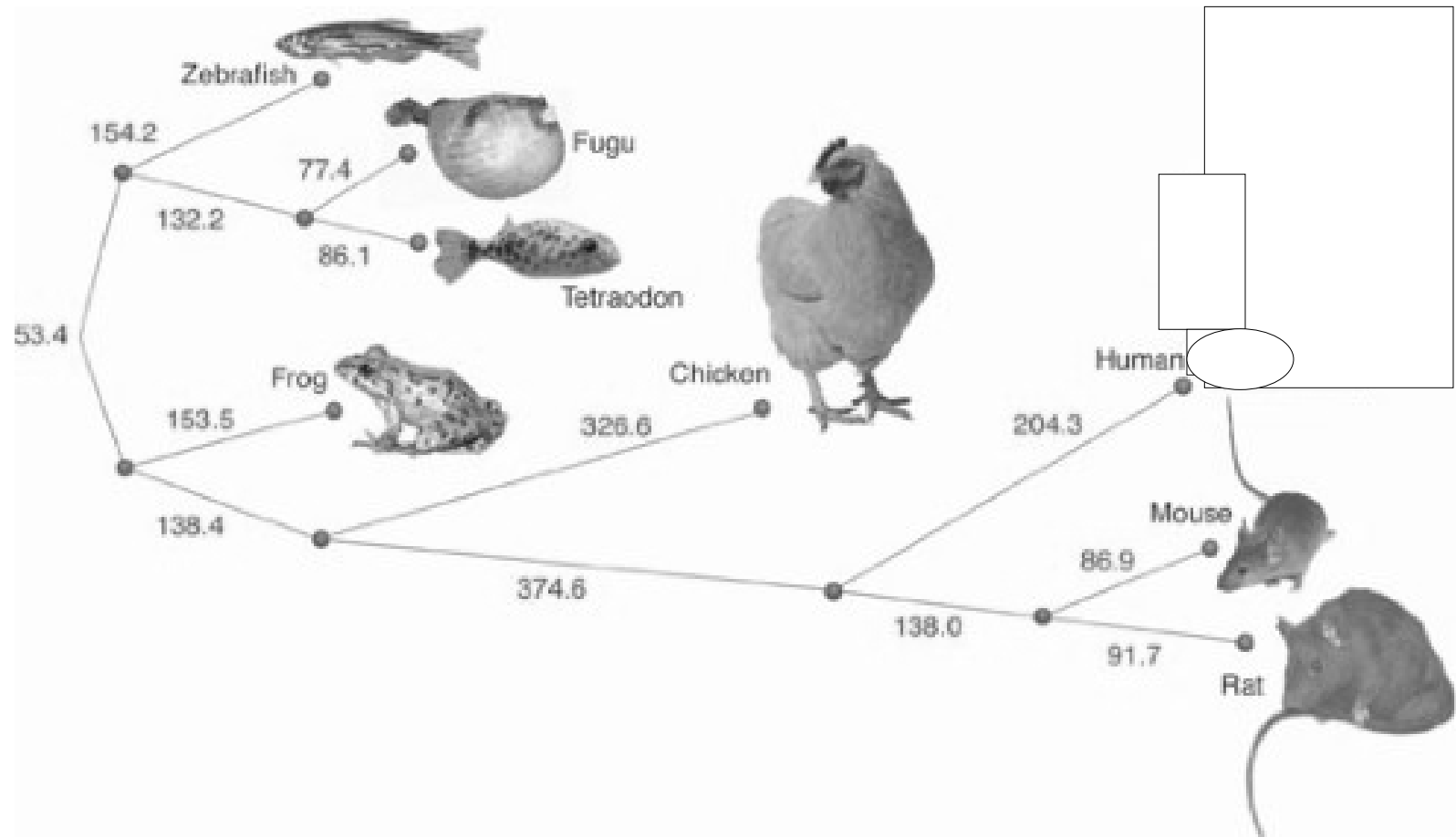
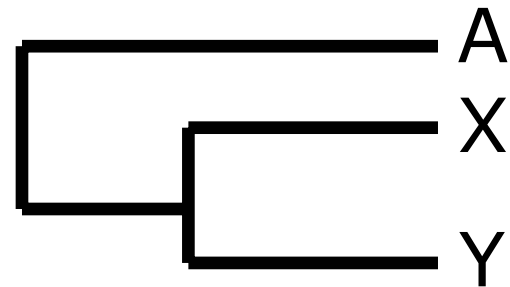


Metodi per la Ricostruzione Filogenetica



Ricostruire una filogenesi significa trovare la migliore stima delle relazioni evolutive storiche fra entità tassonomiche usando i dati disponibili.



I dati di base per riconoscere i gruppi (classificazione), per delineare le loro relazioni (filogenesi, sistematica) vengono detti caratteri tassonomici

Carattere tassonomico

Caratteristica di un taxon che è divisibile in almeno due condizioni contrastanti (stati, o diverse espressioni di uno stesso carattere).

I caratteri tassonomici non comprendono le differenze fra i due sessi o quelle indotte dall'ambiente.

Rango	Sottospecie	Specie	Genere	Taxa superiori
Tasso	molto veloce	veloce	moderato	lento

Criteri Filetici

Caratteri omologhi – Stati di carattere che derivano da un carattere condiviso dall'antenato comune del gruppo

Criteri per l'omologia:

- Similarità di posizione
- Somiglianza dei particolari
- Corrispondenza con altri caratteri
- Modalità di sviluppo

Caratteri Analoghi – Stati di carattere che non riflettono una origine evolutiva comune



Convergenza – sviluppo di caratteristiche simili indipendenti dall'antenato comune



Parallelismo – evoluzione indipendente di caratteristiche simili a partire da una condizione ancestrale simile



Criteri Logici

- Un carattere non deve essere una combinazione di altri
Es. lunghezza del capo + lunghezza del torace +
lunghezza dell'addome = lunghezza del corpo
- Un carattere complesso dovrebbe essere ridotto a
caratteri semplici dal momento che le interazioni fra molti
caratteri unitari può produrre valori simili anche se i
caratteri unitari sono molto diversi
- La presenza di un carattere può essere più informativa
della sua assenza
- Le strutture complesse hanno maggior valore delle
strutture semplici

Criteri Biologici

Comprendere la base biologica della variazione del carattere:

La variazione può essere indotta dall'ambiente. Ciò può essere molto importante per la sopravvivenza dell'organismo, ma non è utile come carattere tassonomico.

I cambiamenti possono essere legati allo stadio di sviluppo

Es: colori stagionali (ermellino, etc)
eteromorfia (differenze fra caste di insetti sociali)
età
differenze fra sessi
caratteri giovanili vs caratteri allo stadio adulto

Caratteri comunemente usati

Morfologici

Esterni

Interni

Embriologici

Cariologici

Molecolari

Allozimi

Sequenze nucleotidiche ed aminoacidiche

Ibridazione DNA-DNA

RFLP

RAPD

Tre scuole di tassonomia

Evolutiva (Simpson, Mayr)

La classificazione biologica deve rispecchiare la massima informazione evolutiva

Fenetica (Sokal, Sneath)

La classificazione deve incorporare il massimo numero di osservazioni oggettive e non essere guidata da una particolare teoria scientifica, come per esempio quella sull'evoluzione organica. Le genealogie sono ricostruite sulla base di indici di distanza

Cladista (Hennig)

"The task of systematics is the creation of a general reference system and the investigation of the relations that extend from it to all other possible and necessary systems in biology" (Hennig, 1966). Le genealogie sono ricostruite sulla base della condivisione di caratteri.

Hennig ha introdotto una speciale terminologia per indicare i diversi stati di un carattere:

stati primitivi del carattere (o plesiomorfi)

stati derivati (o apomorfi)

Caratteri sin-plesiomorfi: stati ancestrali del carattere condivisi da più specie

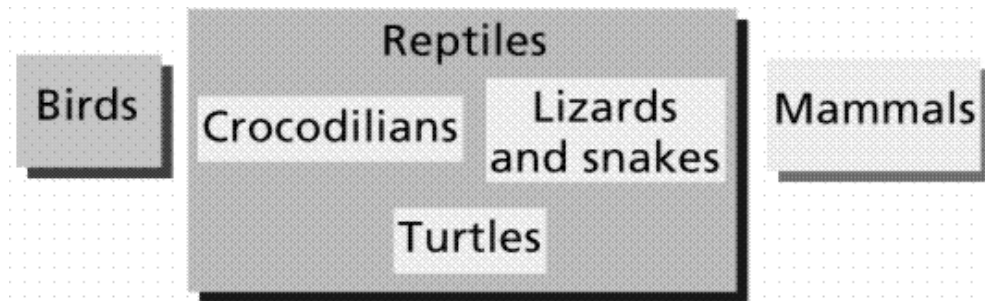
Caratteri sin-apomorfi: stati derivati del carattere condivisi da più specie

Caratteri auto-apomorfi: stati derivati del carattere posseduti da un solo taxon

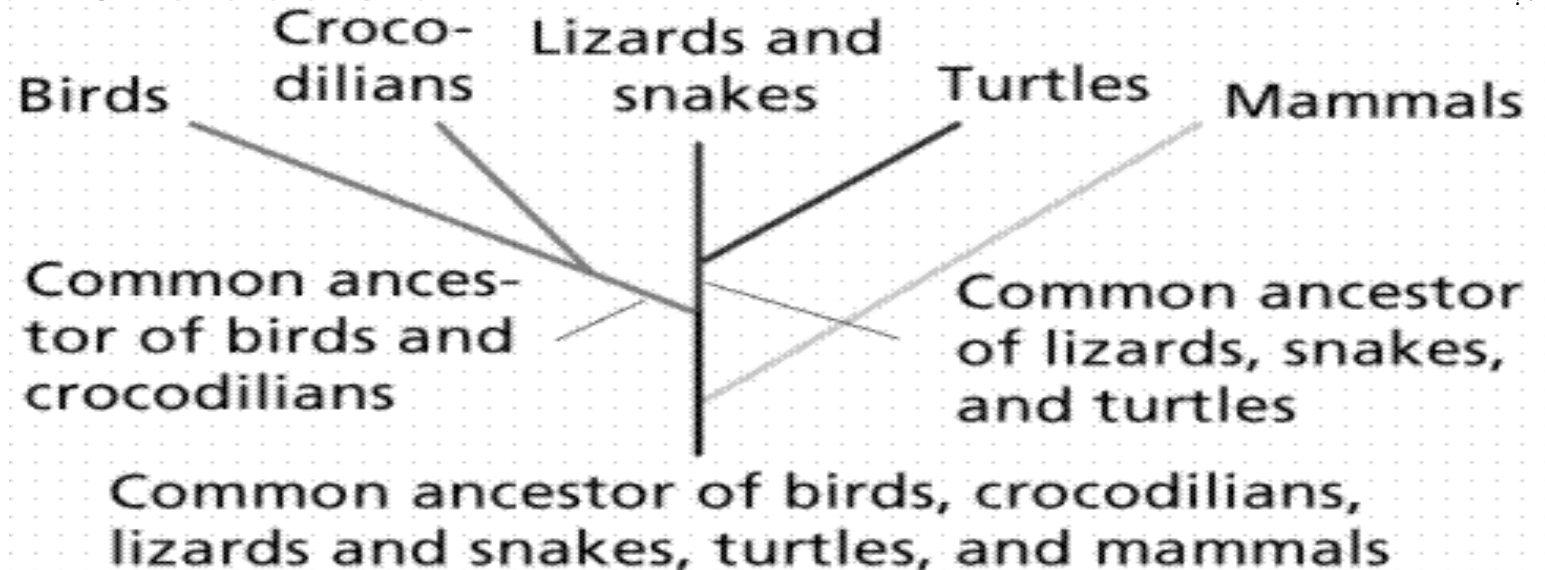
Classificazione evolutiva



Classificazione fenetica



Classificazione cladista



Come si ricostruisce una filogenesi?

Esistono due approcci principali per stimare una filogenesi:

1/ attraverso algoritmi – metodi fenetici che riassumono le differenze in indici di distanza/similarità

(UPGMA (unweighted pair group method using arithmetic averages) e Neighbour Joining (NJ)).

2/ attraverso i criteri di ottimizzazione, ovvero applicazione di una definizione relativa agli alberi migliori

Vi sono due step logici nella ricostruzione filogenetica:

- (a) una definizione di ciò che si intende per albero migliore,
- (b) un confronto fra i diversi alberi, sulla base del criterio di ottimizzazione scelto, per trovare l'albero migliore.

(cioè: massima parsimonia e massima verisimiglianza (o maximum likelihood))

Gli approcci basati sul criterio di ottimizzazione sono in genere i migliori, ma anche i più lenti. Se i taxa da analizzare sono superiori a 10, non è possibile fare una ricerca esatta, ovvero analizzare tutti i possibili alberi.

Il principale vantaggio di questo approccio è che gli alberi possono essere ordinati.

Ciò permette di stimare la potenza (robustezza) dei dati. Un equivalente non è disponibile per gli approcci basati sugli algoritmi, dal momento che è prodotto sempre un solo albero.

Gli algoritmi sono utilizzati in entrambi gli approcci, ma con funzioni differenti.

Negli approcci basati sugli algoritmi, l'algoritmo è centrale rispetto al metodo e definisce quale albero sarà prodotto.

Negli approcci basati sui criteri di ottimizzazione, gli algoritmi sono semplicemente dei mezzi per la ricerca dell'albero migliore che più si adatta al criterio di ottimizzazione prescelto.

Metodi per la ricerca degli alberi ottimali



♦ Algoritmi Esatti

"Garantiscono" l'albero ottimale o migliore

- Nella costruzione dell'albero possono essere usati due sistemi diversi:

Ricerca esaustiva: vengono valutati tutti i possibili alberi non radicati

Ricerca branch-and-bound: Elimina la parte dell'albero che presenta soluzioni sub-ottimali

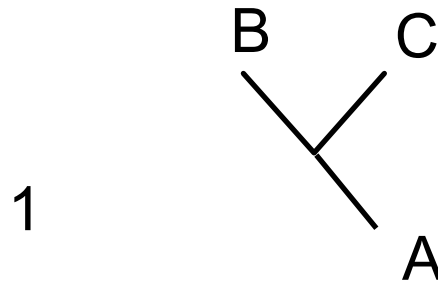


♦ Algoritmi euristici

"Non garantiscono" l'albero ottimale o migliore

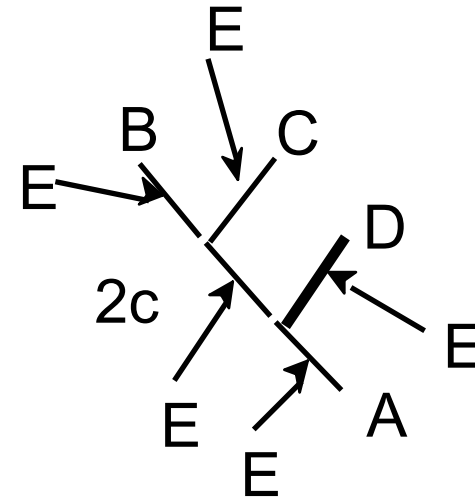
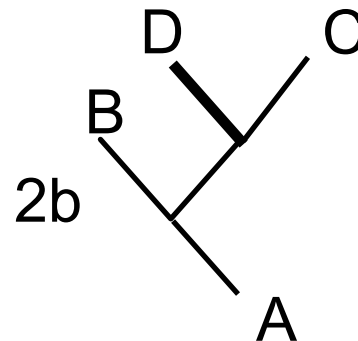
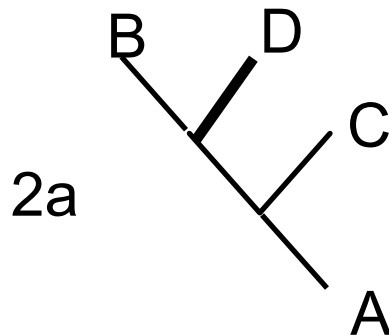
- Spesso operano con metodi "hill-climbing"

Ricerca degli alberi ottimali – ricerca esaustiva



Albero iniziale, 3 taxa qualsiasi

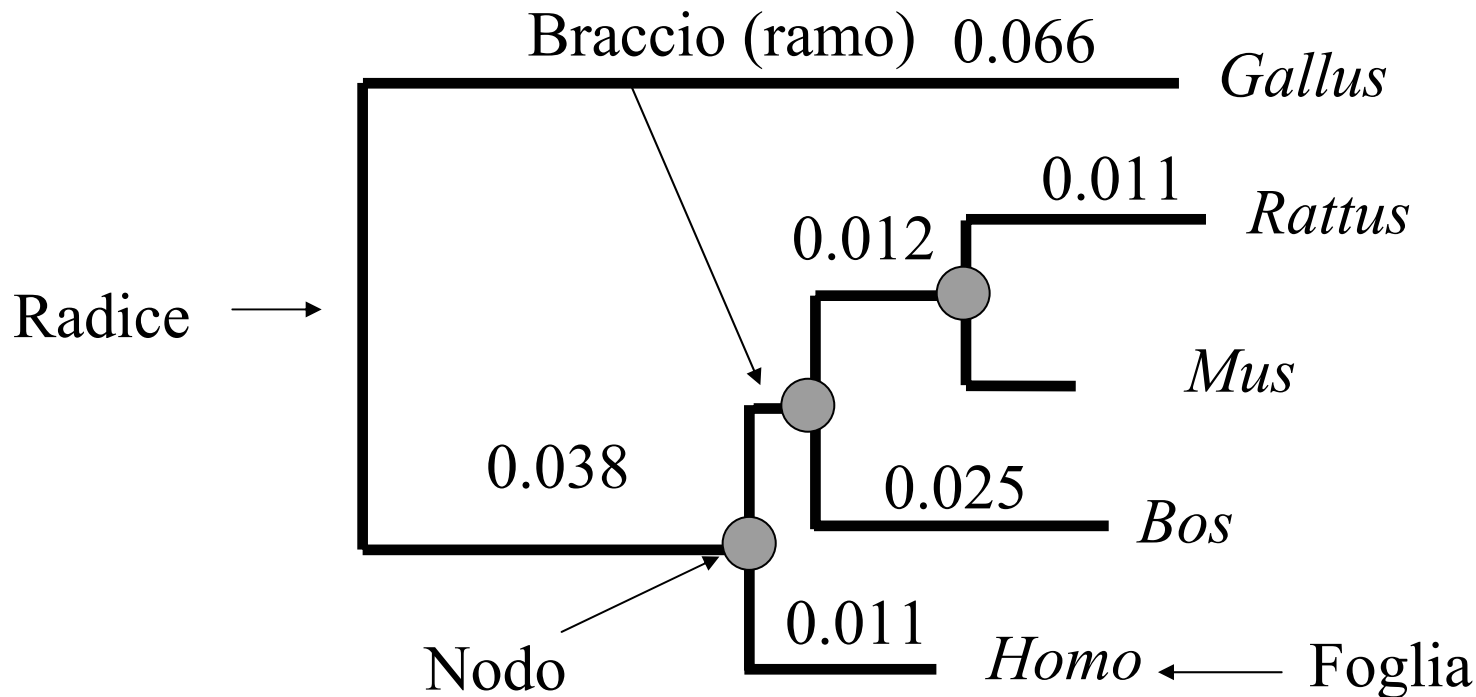
Aggiunge il quarto taxon (D) in ognuna delle tre possibili posizioni
-> tre alberi



Aggiunge il quinto taxon (E) in ognuna delle cinque possibili posizioni
su ognuno dei tre alberi -> 15 alberi, e così via

Albero Filogenetico

- Braccio (ramo) interno: fra due nodi
- Braccio (ramo) esterno: fra un nodo e una foglia
- La lunghezza orizzontale del ramo è proporzionale alle distanze evolutive fra le sequenze e i loro antenati (unità = sostituzione per sito)
- Topologia dell'albero= forma dell'albero= ordine dei rami fra i nodi



Metodi per la ricostruzione filogenetica

Metodi di Distanza

Principio generale:

✓ Registrazione dei dati



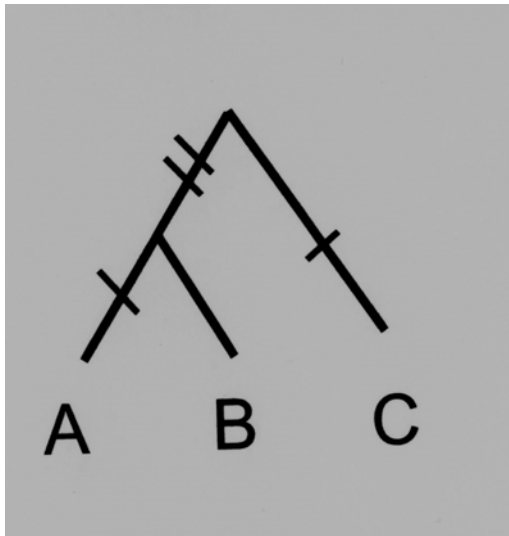
✓ Matrice di distanze fra coppie di osservazioni



✓ Costruzione dell'albero

Corrispondenza fra alberi e matrici di distanza

- ✓ Ogni albero filogenetico rispecchia una matrice di distanze fra coppie di osservazioni
- ✓ Matrici di distanza “perfette” corrispondono a un singolo albero filogenetico



Albero



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

Matrice di distanze

Clustering analysis (UPGMA)

Viene trovata 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, quindi, 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.

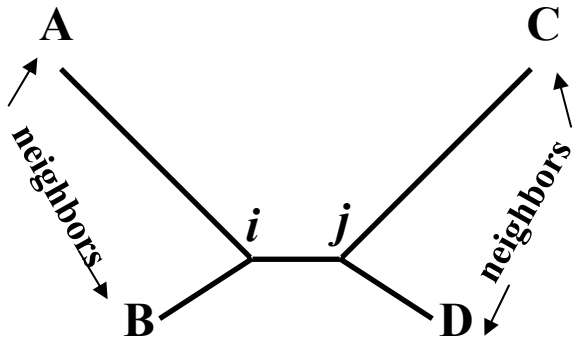
Questo processo viene ripetuto fino a che non sono stati aggiunti tutti i taxa all'albero.

Questo metodo produce alberi radicati e ultrametrici (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), ma è quasi sempre violata. UPGMA è ancora comunemente usato, nonostante confrontato ad altri ora esistenti, sia un metodo poco potente.

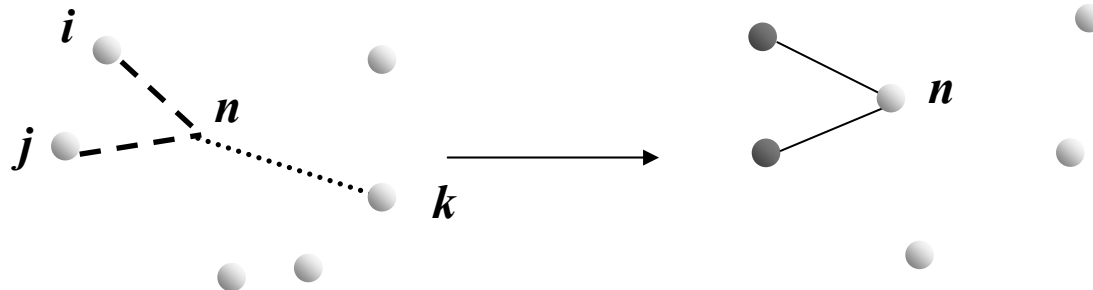
Clustering analysis (Neighbour joining):

È un metodo che produce alberi *non radicati*, seguendo la *proprietà additiva* (un albero è detto additivo se le distanze fra i taxa sono uguali alla somma delle lunghezze dei rami che li connettono).

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 (cerchi bianchi in figura). Sulla base delle distanze genetiche note, il metodo sceglie due nodi vicini i e j e li connette attraverso un nuovo nodo interno n . I nodi terminali originali i e j vengono eliminati, in quanto già connessi (cerchi grigi in figura). I cerchi bianchi sono ovviamente diminuiti e il processo viene ripetuto fino a che tutti i nodi terminali non vengono connessi in un singolo albero.



Metodi per la ricostruzione filogenetica

Metodi basati sul criterio di ottimizzazione

- ✓ **Massima Parsimonia**
- ✓ **Maximum Likelihood**
- ✓ **Metodi Bayesiani**

Metodi basati sui criteri di ottimizzazione

Gli approcci che utilizzano il criterio di ***Parsimonia*** lavorano direttamente con i dati carattere e assumono che l'albero che richiede il più piccolo numero di cambiamenti è quello che meglio spiega i dati.

Un semplice esempio:

sequenza A **GTTgTAATGT**

sequenza B **GTTaTAATGT**

E' più probabile che queste due sequenze abbiano subito un solo cambiamento $G \rightarrow A$, o che invece vi siano stati degli stati di carattere intermedi e che, quindi, il corso reale dell'evoluzione abbia coinvolto molti eventi?

sequence A **GTTgTAATGT**

sequence A' **GTTtTAGTGT**

sequence A'' **GTTtTAATCT**

sequence B **GTTaTAATGT**

In assenza di ulteriore informazione dovremmo preferire l'ipotesi di un singolo cambiamento, ovvero la più parsimoniosa.

L'analisi della *Massima Parsimonia* procede nel modo seguente...

- ❖ Inizia identificando i siti *informativi**, su cui verrà eseguita l'analisi
- ❖ Utilizzando i metodi di ricerca esatti o quelli euristici trova un set di possibili alberi. Tra questi sceglie quello(i) più corto.
- ❖ La lunghezza è funzione del numero dei rami, e dei caratteri

* Un sito è informativo se sono presenti almeno due stati di carattere in più di un taxon

Proprietà della Parsimonia

- Alberi differenti possono essere egualmente parsimoniosi (stessa lunghezza, la più corta di tutte le possibili lunghezze)
- La posizione dei cambiamenti su ogni ramo non è definita in modo univoco

La parsimonia non permette di definire un modo univoco

- Il numero di alberi da valutare cresce rapidamente con il numero di taxa da analizzare:

La parsimonia può essere un calcolo molto costoso

La ricerca dell'albero più corto deve essere fatta su un sottocampione di tutti i possibili alberi

Non vi è certezza matematica nel trovare l'albero più parsimonioso

Taxa (n)	rooted $(2n-3)!/(2n-2(n-2)!)$
2	1
3	3
4	15
5	105
6	954
7	10,395
8	135,135
9	2,027,025
10	34,459,425

Metodi per la ricostruzione filogenetica

Metodi basati sul criterio di ottimizzazione

✓ **Massima Parsimonia**

✓ **Maximum Likelihood**

✓ **Metodi Bayesiani**

FILOGENESI MOLECOLARE

Allineamento e Gaps

- La qualità dell'allineamento è essenziale: ogni colonna dell'allineamento (sito) si suppone contenga residui omologhi (nucleotidi, aminoacidi) che derivano da un antenato comune.
- La maggior parte dei metodi considera soltanto le sostituzioni, le inserzioni e/o le delezioni non vengono considerate

CLUSTAL X: multiple sequence alignment

<i>C.gracilipes</i>	AAACATGTCTTTTTGTAAATAATTTAAAGTCTGGCCTGCCCACTGA----TATAAG----
<i>T.cavicola</i>	AAACATGTCTTTTTGAGAATAATTTAAAGTCTGACCTGCCCACTGA----AATAAAAGAT
<i>H.cumberlandicus</i>	AAACATGTCTTTTTGATTATAATTTGAAGTCTGACCTGCCCACTGACGTTTATAAG----
<i>D.euxina</i>	AAACATGTCTTTTTGATAATAATTTAAAGTCTGACCTGCCCACTGA----AATAA-----
<i>D.ligustica</i>	AAACATGTCTTTTTGATGATAATTTAAAGTCTGACCTGCCCACTGA----TATAAA----
<i>D.aegilion</i>	AAACATGTCTTTTTGATAATAATTTAAAGTCTGACCTGCCCACTGA----TTTAAA----
<i>D.baccetti</i>	AAACATGTCTTTTTGATAATAATTTAAAGTCTGACCTGCCCACTGA----TATAAA----
<i>D.schiavazzii</i>	AAACATGTCTTTTTGATGATAATTTAAAGTCTGACCTGCCCACTGA----T-TATA----
<i>D.laetitiae</i>	AAACATGTCTTTTTGATGATAATTTAAAGTCTGACCTGCCCACTGA----TATAAA----
<i>D.geniculata</i>	AAACATGTCTTTTTGATGATAATTTAAAGTCTGACCTGCCCACTGA----TTGTAA----

Modelli di evoluzione delle sequenze

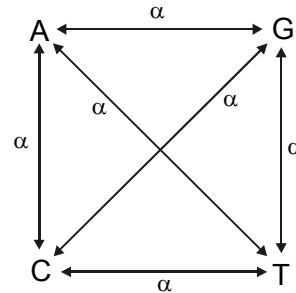
	260	*	280	*	300	*	320		
0841r :	CC	CAATTTTAT	AGC	GCC	ACTCC	TTGGGCAC	AGAGTTTTAGGAG	AAATAAGTATGTG	: 272
0992r :	CC	CAATTTTAT	AGC	GCC	ACTCC	TTGGGCAC	AGAGTTTTAGGAG	AAATAAGTATGTG	: 213
3803r :	CC	CAATTTTAT	AGC	GCC	ACTCC	TTGGGCAC	AGAGTTTTAGGAG	AAATAAGTATGTG	: 305
4062r :	CC	CAATTTTAT	AGC	GCC	ACTCC	TTGGGCAC	AGAGTTTTAGGAG	AAATAAGTATGTG	: 319
3802r :	CC	CAATTTTAT	AGC	GCC	ACTCC	TTGGGCAC	AGAGTTTTAGGAG	AAATAAGTATGTG	: 282
ph2f :	CC	CAATTTTAT	AGC	GCC	ACTCC	TTGGGCAC	AGAGTTTTAGGAG	AAATAAGTATGTG	: 306
	CCTcCAATTTTATTag ttgcctactcctttggg acAGAGTTTTAGGAGAAATAAGTATGTG								

Il problema

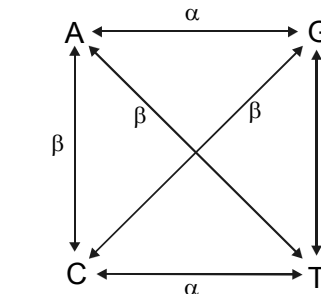
- Un processo basilare nell'evoluzione di una sequenza è il cambiamento di quella sequenza nel tempo
- Esistono diversi modelli matematici che ne descrivono il cambiamento
- E' importante avere un modello per capire quali sono stati i meccanismi di cambiamento e per stimare sia il tasso di evoluzione che la storia evolutiva delle sequenze

Modelli di evoluzione delle sequenze

Tutte le sostituzioni
hanno una eguale
probabilità, ma le
frequenze delle basi
sono diverse

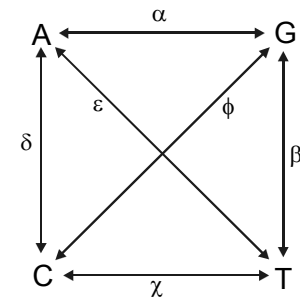


Felsenstein (1981)



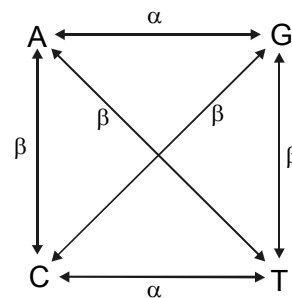
**Hasegawa, Kishino
& Yano (HKY) (1985)**

Transizioni e trasversioni
hanno probabilità diverse, le
frequenze delle basi
sono diverse



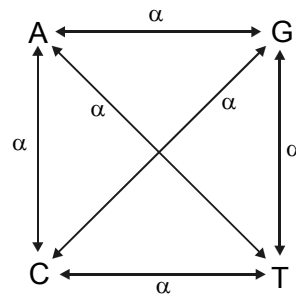
**General time reversible
model (GTR)**

Probabilità diverse per ogni
sostituzione, le frequenze
delle basi sono diverse



Kimura 2 parametri (1980)

Transizioni e
trasversioni hanno
probabilità differenti e
le frequenze delle basi
sono uguali



Jukes-Cantor (1969)

Tutte le sostituzioni
hanno una eguale
probabilità e le
frequenze delle basi
sono uguali

Metodi per la ricostruzione filogenetica

Metodi basati sul criterio di ottimizzazione

✓ **Massima Parsimonia**

✓ **Maximum Likelihood**

✓ **Metodi Bayesiani**

Maximum likelihood (massima verisimiglianza)

La Parsimonia può incorporare all'interno del suo criterio di ottimizzazione gli aspetti del cambiamento e, quindi non sono necessari modelli di evoluzione espliciti.

Un approccio alternativo consiste nello specificare un modello di evoluzione e stimare su questo l'albero migliore.

L'albero migliore è quello che ha la più alta probabilità di essere generato sulla base del modello e dei dati.

Metodi di maximum likelihood

✓ Ipotesi

- Il processo di sostituzione segue un modello probabilistico la cui espressione matematica, ma *non i parametri*, è nota *a priori*

✓ Principio

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

✓ Vantaggi

- Considera tutti i dati contemporaneamente, a differenza dei metodi di distanza e di parsimonia

Scelta di un modello realistico:

- 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

Proprietà del Maximum likelihood

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 probabile.

METODI BAYESIANI

Strettamente connessi ai metodi di Maximum Likelihood

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

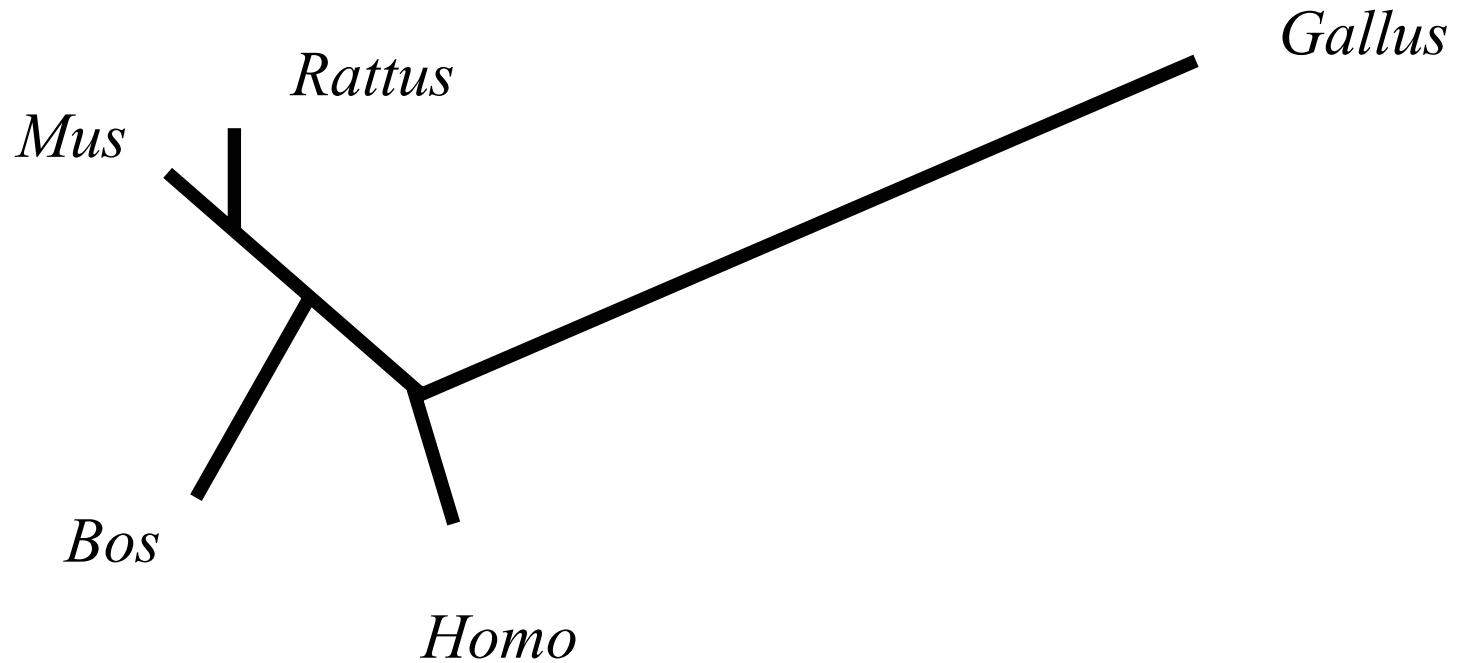
Ogni parametro del modello ha un valore di probabilità

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)

Orientamento degli Alberi

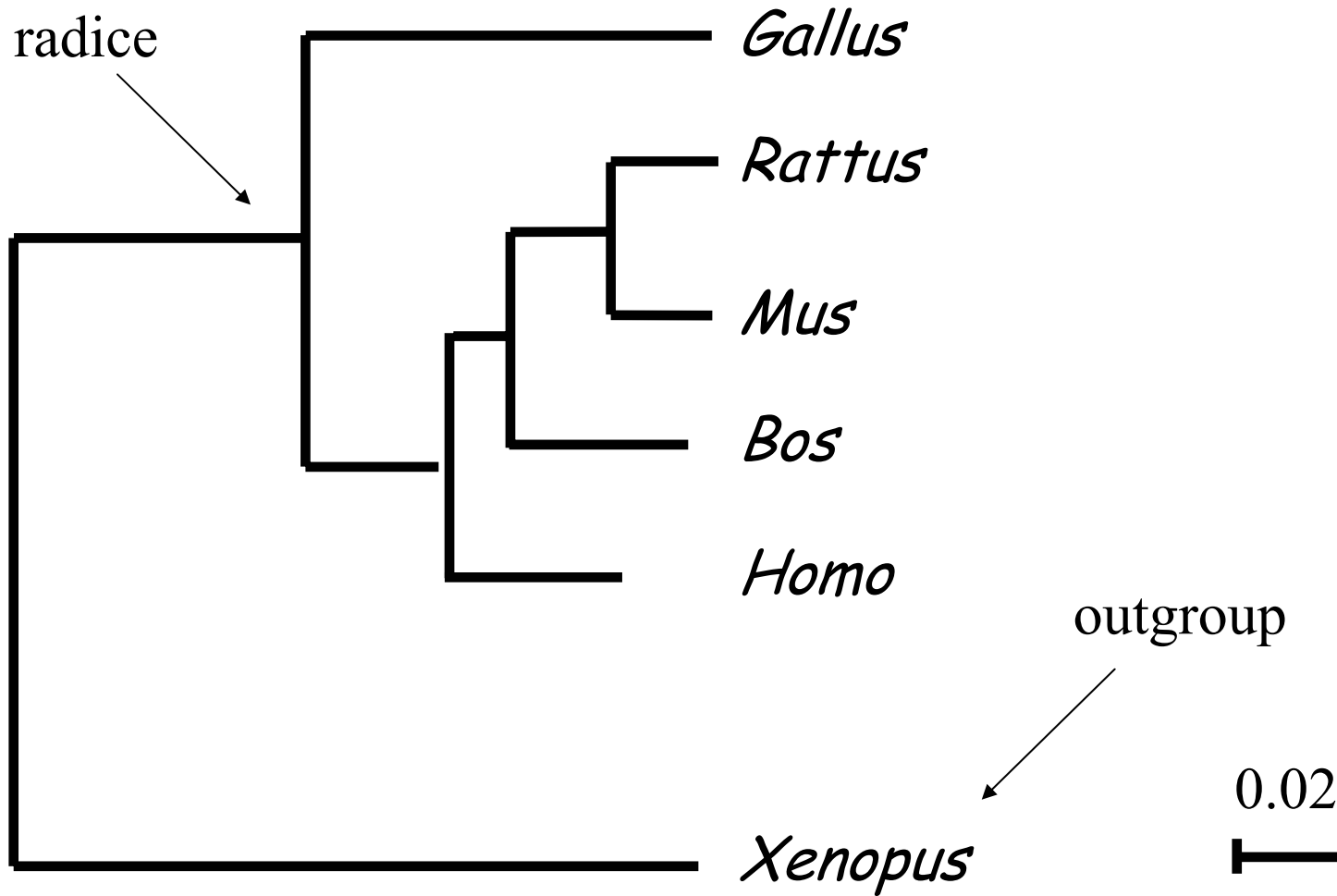
- La maggior parte dei metodi filogenetici produce alberi senza radice. Ciò avviene perché essi rivelano le differenze fra i taxa, ma non hanno modo di orientare i cambiamenti residui relativamente al tempo.
- Esistono due metodi per orientare gli alberi:
 - Il metodo dell'**outgroup**: comprende nell'analisi un gruppo di sequenze che devono essere esterne al gruppo sotto studio: la radice viene posta sul ramo che congiunge l'outgroup alle altre sequenze
 - Ipotesi dell'**orologio molecolare**: si suppone che tutte le linee si siano evolute con la stessa velocità dal momento della divergenza dall'antenato comune. La radice viene posta nel punto equidistante da tutte le foglie.

Albero non radicato



0.02
└───┘

Albero Radicato



Applicazioni della Filogenesi Molecolare

- Tassonomia (classificazione delle specie)
- Identificazione dei campioni (diagnostica, controllo qualità)
- Rilevamento di trasferimenti orizzontali
- Studio dell'evoluzione dei geni (duplicazioni, acquisizione di nuove funzioni)
- Epidemiologia (es.: origine di HIV)

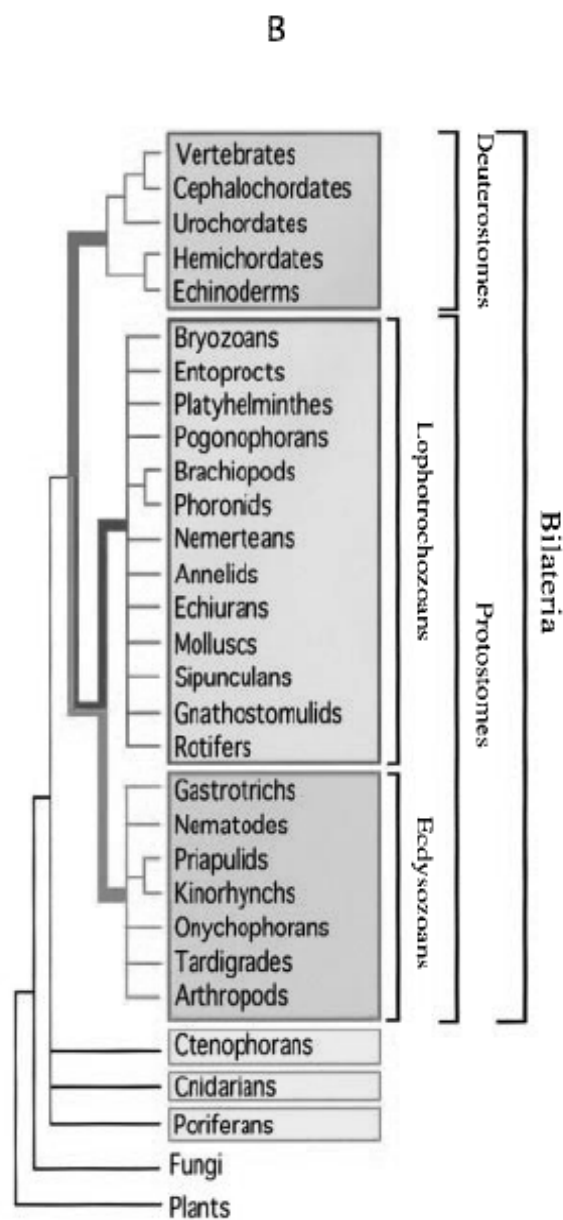
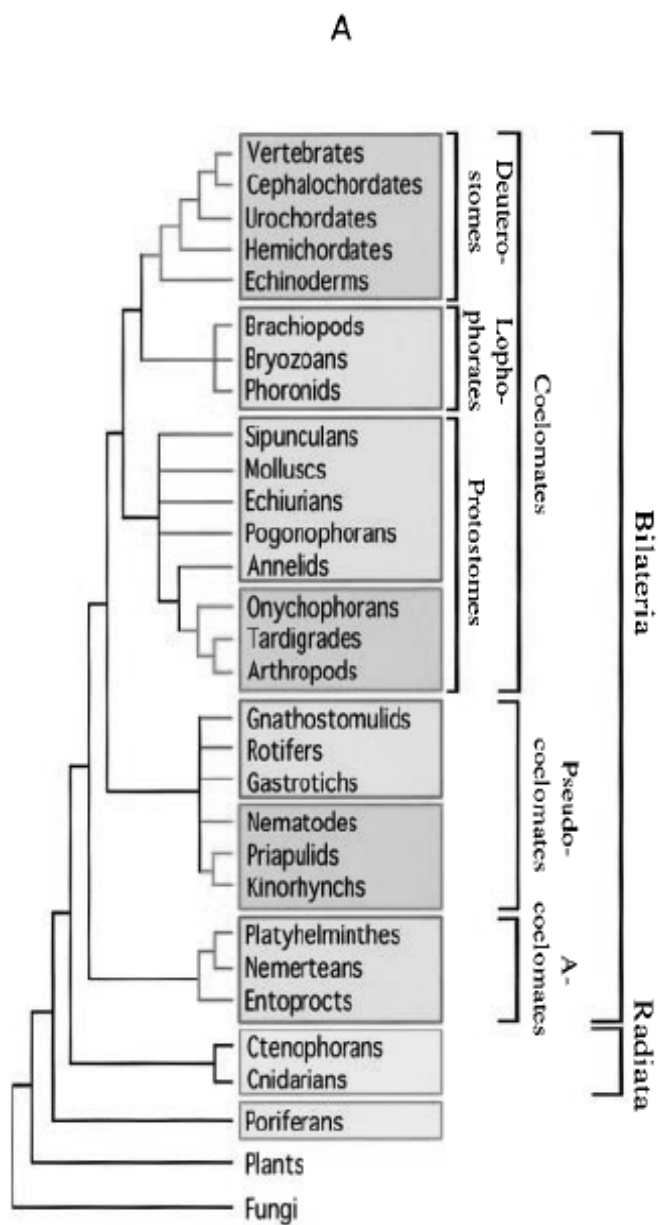


Fig. 1. Metazoan phylogenies. (A) The traditional phylogeny based on morphology and embryology, adapted from Hyman (11). (B) The new molecule-based phylogeny. A conservative approach was taken in B: i.e., some datasets provide resolution within some of the unresolved multifurcations displayed, but we have limited the extent of resolution displayed to that solidly provided by rRNA only.

