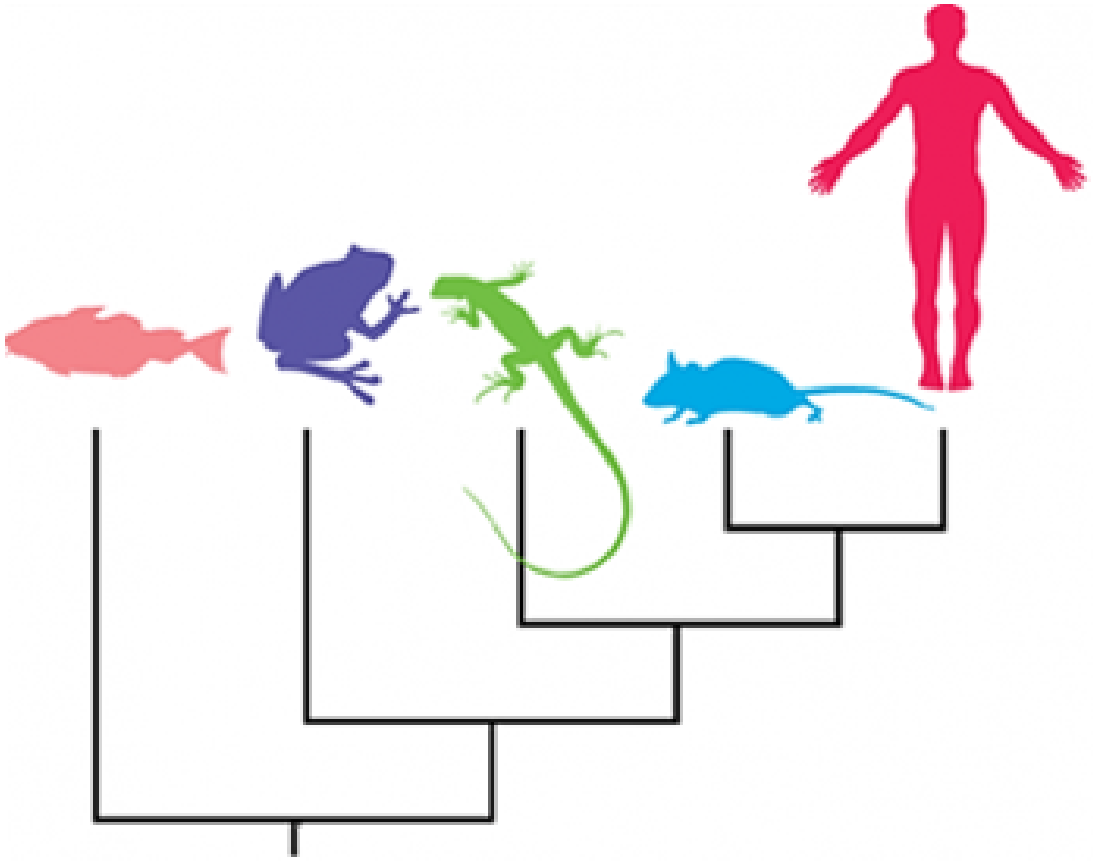
Filogenesi molecolare: le basi teoriche

Giuseppe Manna



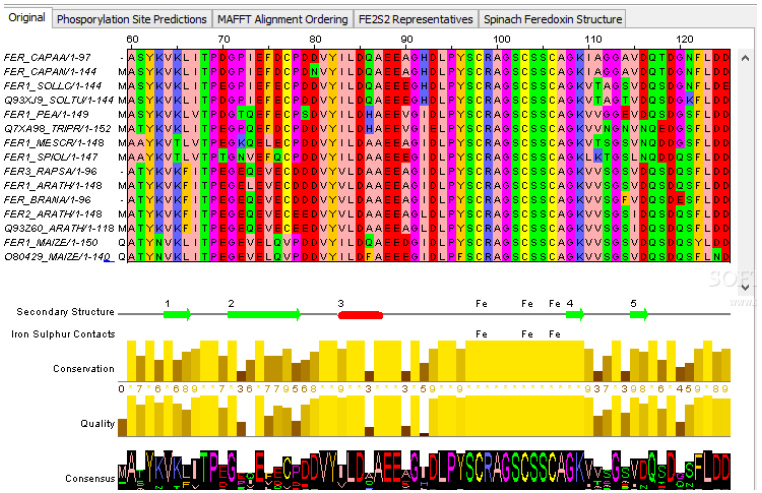
21 giugno 2016

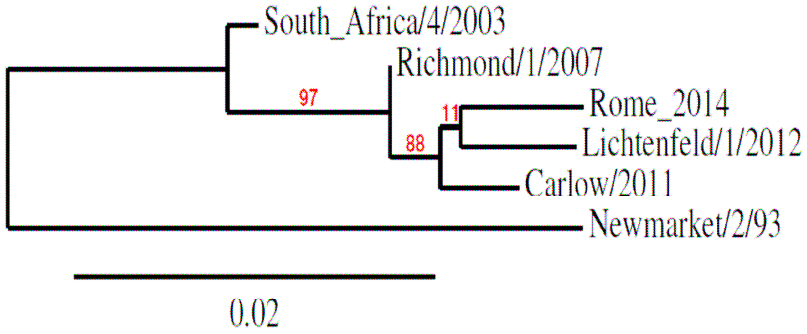
Direzione Operativa Diagnosi delle Malattie Virali

Introduzione

Obiettivi del Corso

Sapere “leggere” un albero filogenetico

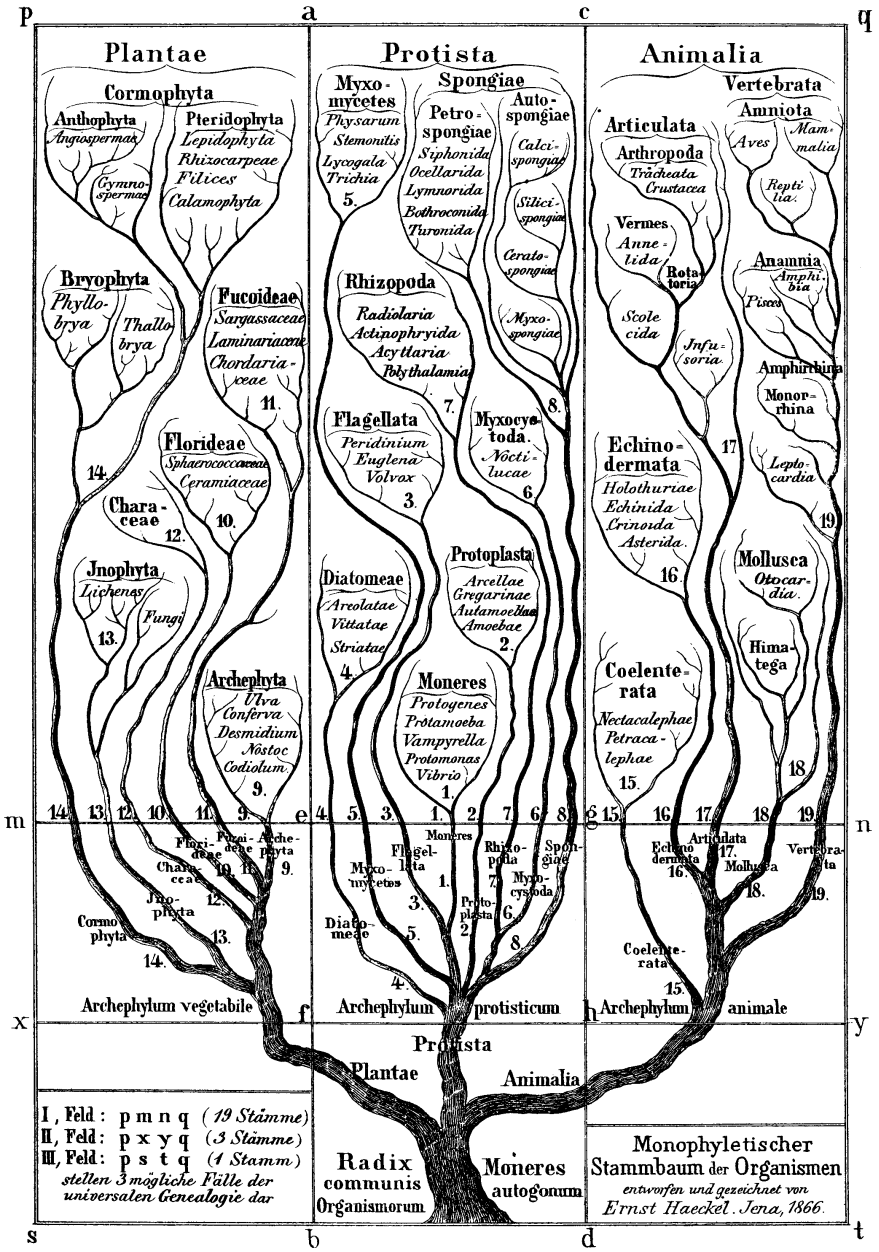




In questa presentazione

parleremo della filogenesi molecolare

come branca della filogenetica



Filogenesi

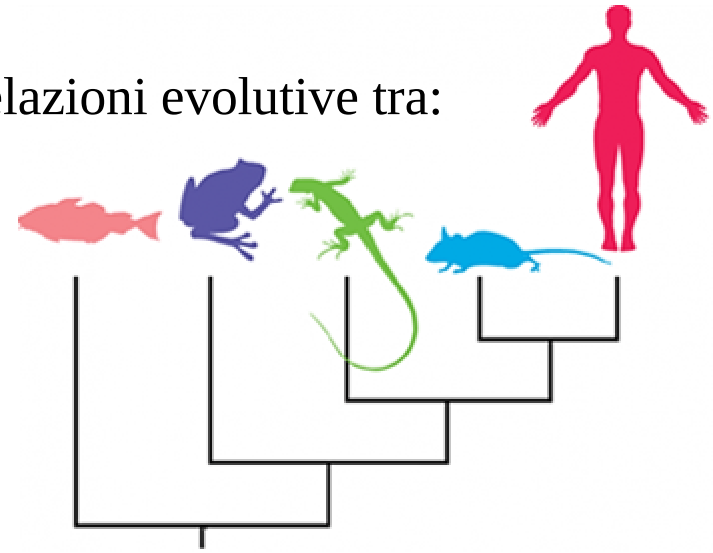
Filogenesi: storia evolutiva di un gruppo di organismi alla luce delle loro relazioni reciproche di discendenza e di affinità. Per lo studio della f. in biologia evoluzionistica ci si avvale di **dati morfologici, paleontologici** ecc.

Notevole è stato il contributo della **biochimica** allo **studio delle diverse strutture primarie di alcune proteine** (emoglobina, citocromo γ), che ha permesso l'individuazione dei cosiddetti **orologi molecolari**.

Nello studio della filogenesi è possibile riconoscere diversi livelli di indagine: dall'analisi dei dati si può passare agli scenari evolutivi, alberi filogenetici associati a dati concernenti gli adattamenti, l'ecologia, la biogeografia, la tettonica delle placche ecc.
(Treccani.it – Enciclopedie on line)

Riepilogando

La filogenetica è la scienza che analizza delle relazioni evolutive tra:



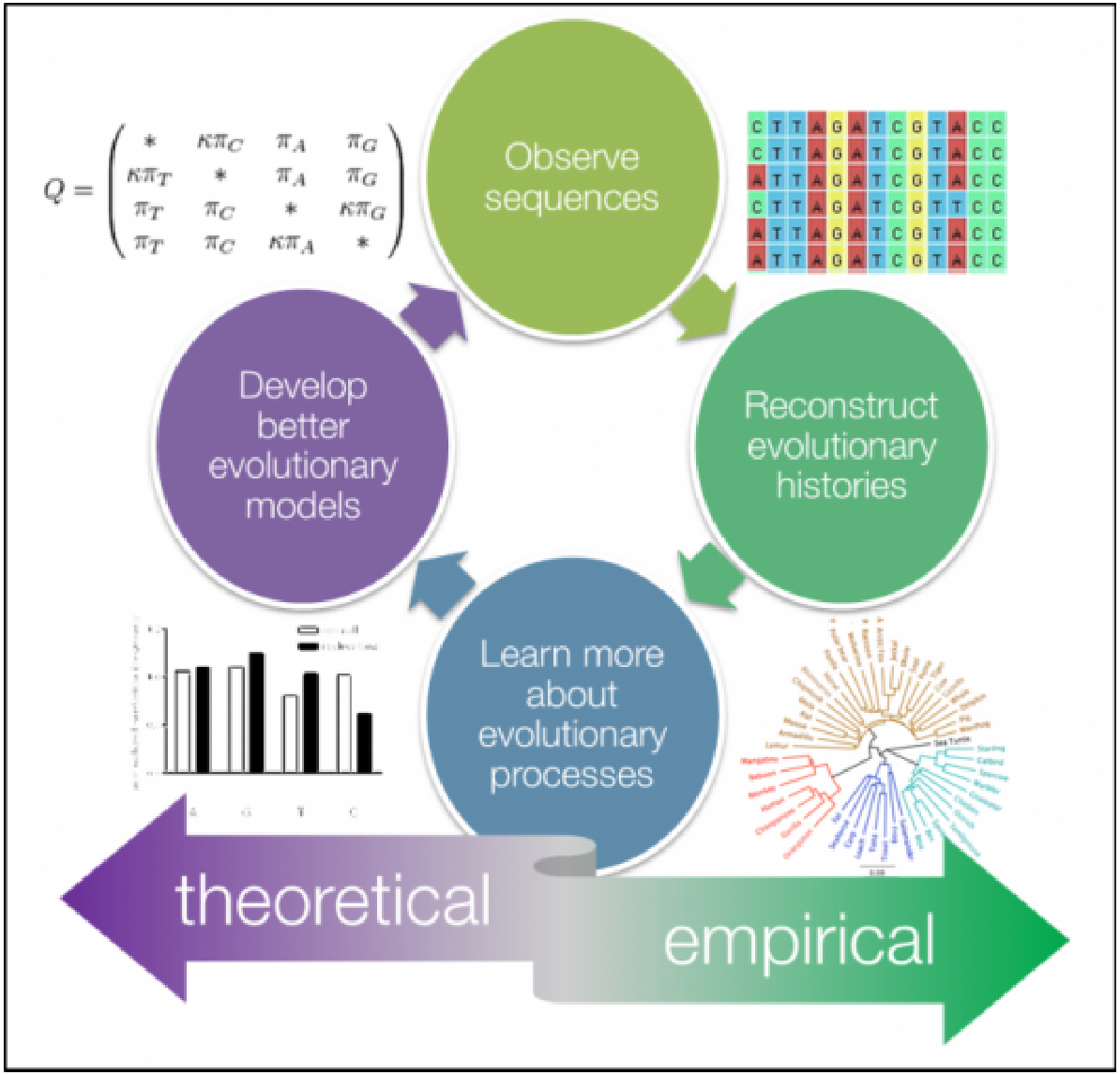
- specie
- individui
- geni
- entità biologiche

Ci chiediamo:

- qual'è la relazione evolutiva tra le varie entità
- in che modo si trasformano le sequenze durante il processo evolutivo
- se esistono dei modelli matematici affidabili che possono descrivere l'evoluzione attraverso l'analisi delle sequenze di nucleotidi/aminoacidi

Filogenesi molecolare

Filogenesi molecolare

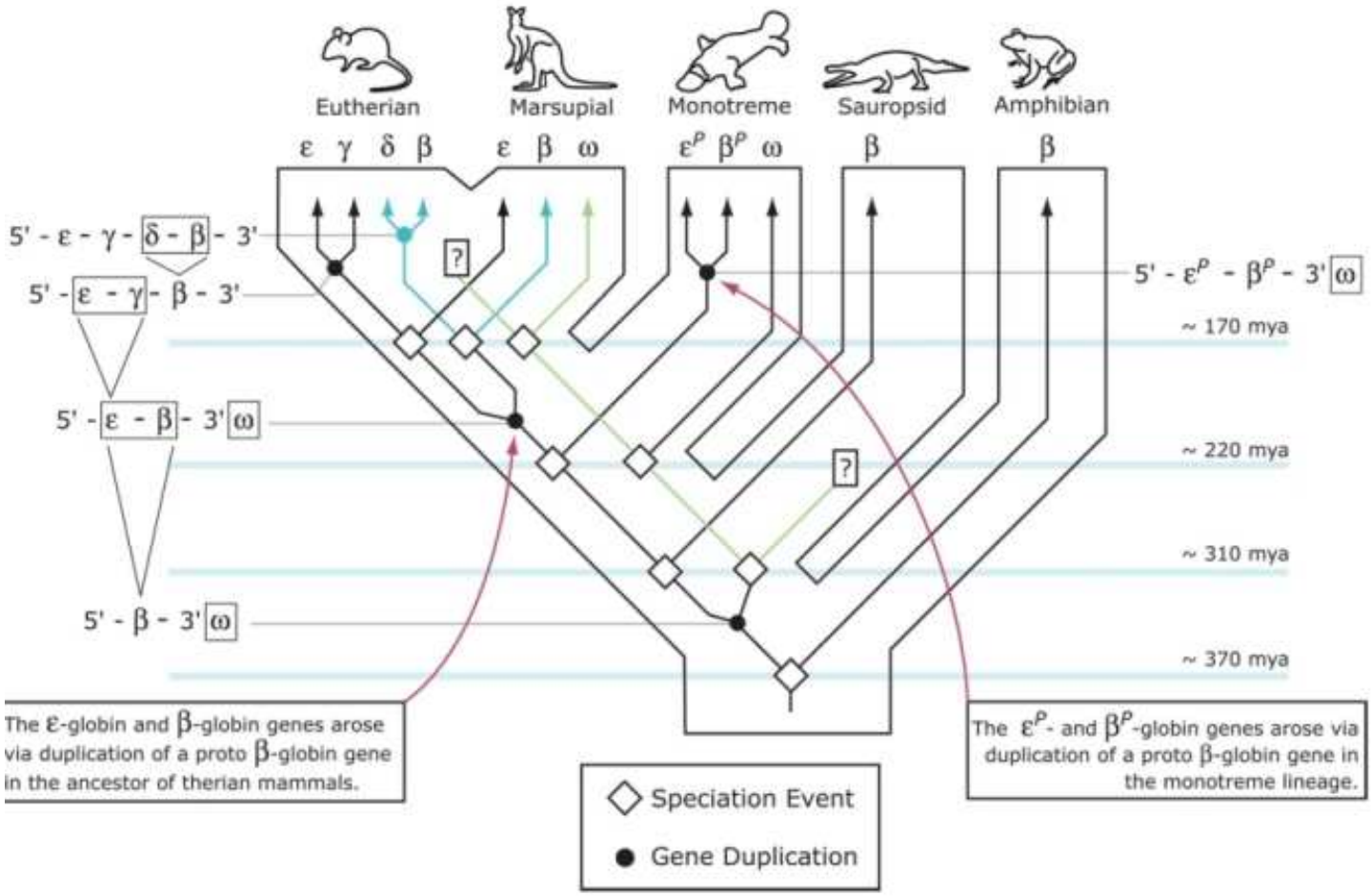


Possiamo costruire un albero filogenetico analizzando le trasformazioni evolutive in sequenze nucleotidiche o aminoacidiche.

Lo studio degli eventi del passato ci aiuta a creare un modello evolutivo di mutazione nel tempo, e nello spazio.

L'iterazione di questi metodi ci aiuta a migliorare i modelli matematici per lo studio dell'evoluzione.

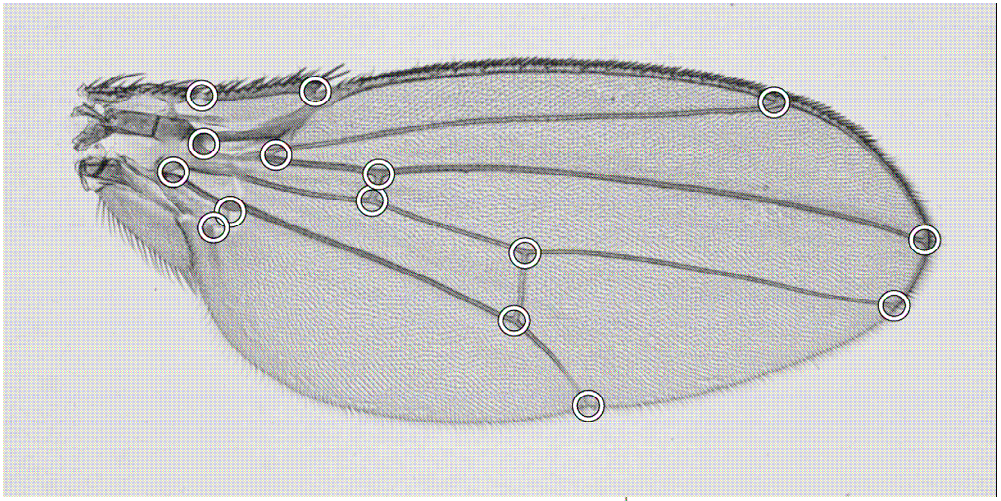
Perchè usare le sequenze (1)



Le sequenze sono molto adatte agli studi filogenetici perché a una maggiore distanza evolutiva corrisponde una maggiore differenza tra sequenze omologhe (orologio molecolare)

Fare sequenziamenti di materiale genetico è semplice, affidabile, poco costoso.

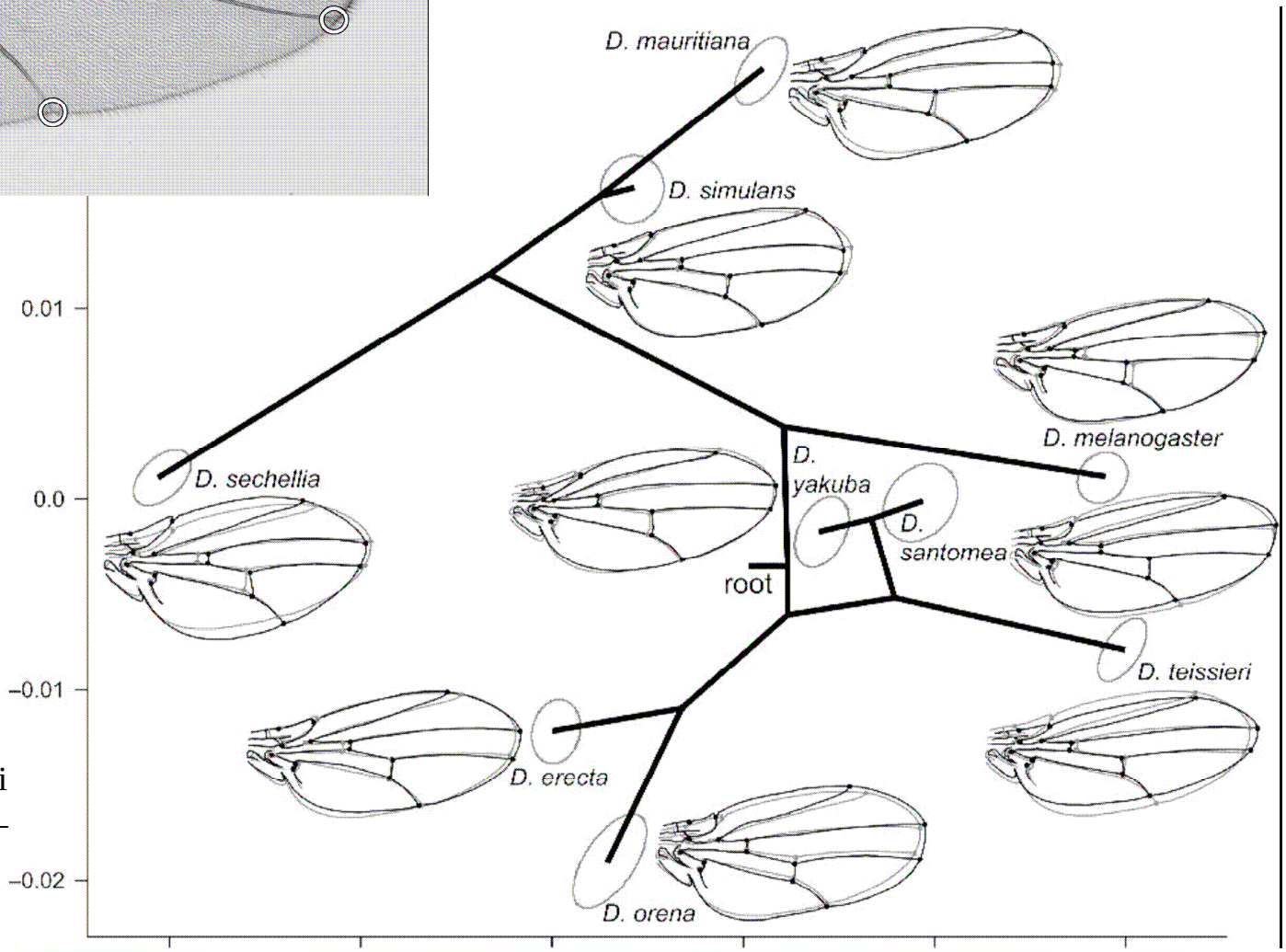
Le sequenze utilizzate sono molto specifiche e molto ricche di informazioni.



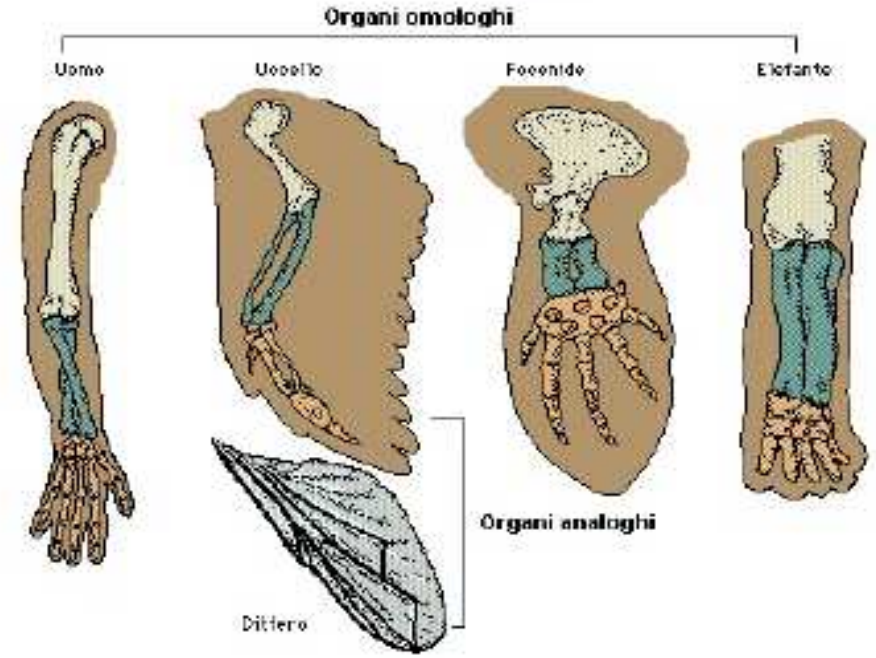
Perchè usare le sequenze (2)

Se non fosse possibile reperire materiale genetico per gli studi evoluzionistici (come nei di fossili) si possono utilizzare caratteristiche morfometriche.

Ma non ci assicura lo stesso livello di affidabilità delle sequenze perché spesso caratteristiche morfologiche simili provengono da linee evolutive indipendenti (es. ali negli uccelli e nei pipistrelli – caratteri analoghi)



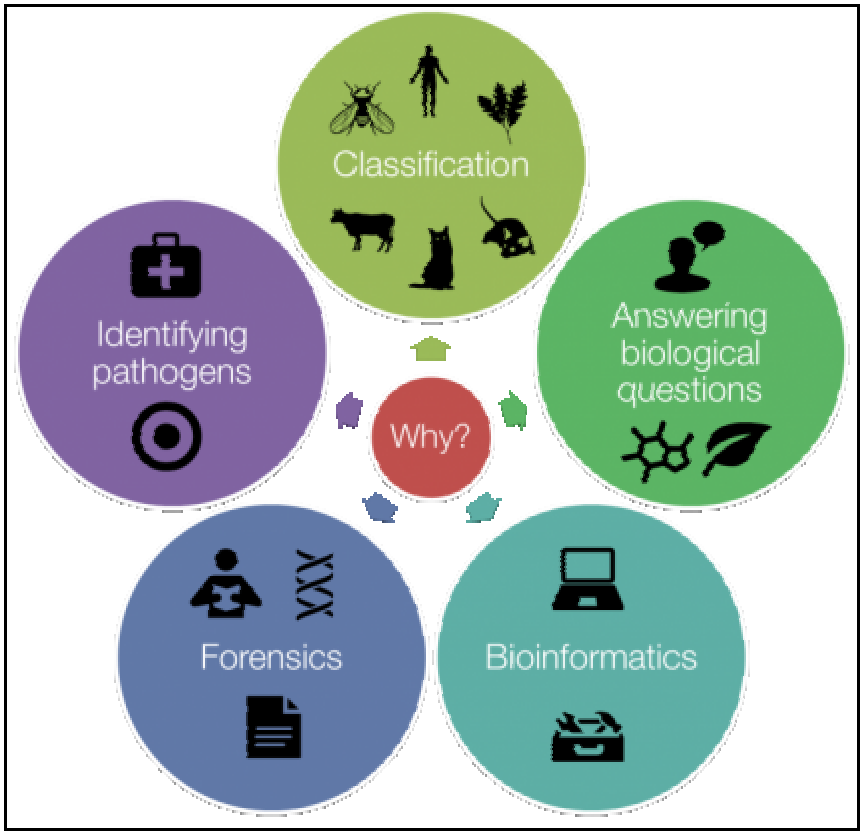
Strutture con un'origine comune, ma non necessariamente una funzione comune sono definite **Omologhe** (Es. l'ala di un Uccello e la zampa di un cavallo).



Strutture che svolgono la medesima funzione, che presentano una somiglianza superficiale, ma origini evolutive differenti sono definite **Analoghe** (Es. l'ala di un uccello e l'ala di un insetto).

Omologia e Analogia (Omoplasia)

Applicazioni



Classificazione degli organismi: La Filogenesi basata sulle sequenze genetiche ci fornisce modelli più accurati dei gradi di parentela.

Forense: la filogenesi viene utilizzata per valutare le tracce di DNA, per individuare l'autore di un crimine, per stabilire l'origine della contaminazione microbica di alimenti, stabilire la paternità di un figlio.

Identificare l'origine e le vie di diffusione di un patogeno: il sequenziamento molecolare e la filogenesi possono fornirci molte informazioni sulle origini di un focolaio di infezione. Possono aiutarci ad elaborare nuove linee guida sanitarie.

La filogenesi è spesso indispensabile per dare un significato biologico ai nostri studi

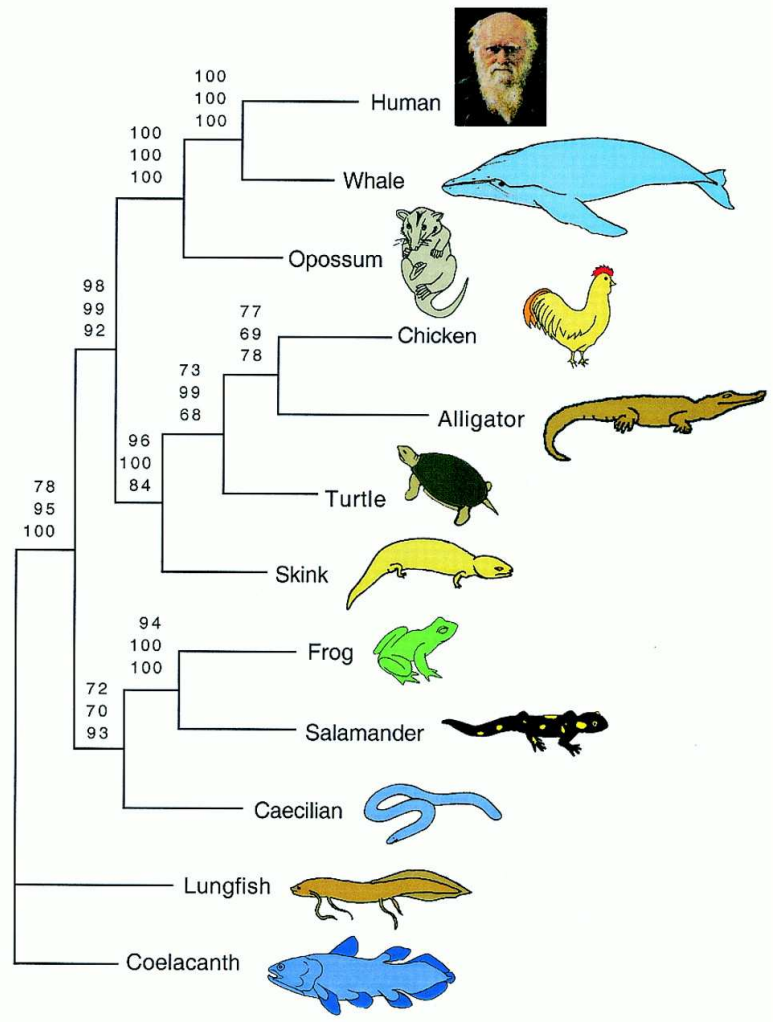
Molti degli algoritmi sviluppati per la filogenetica, sono stati poi utilizzati in altri settori della conoscenza

Le sequenze sono informative sulle relazioni evolutive tra le specie

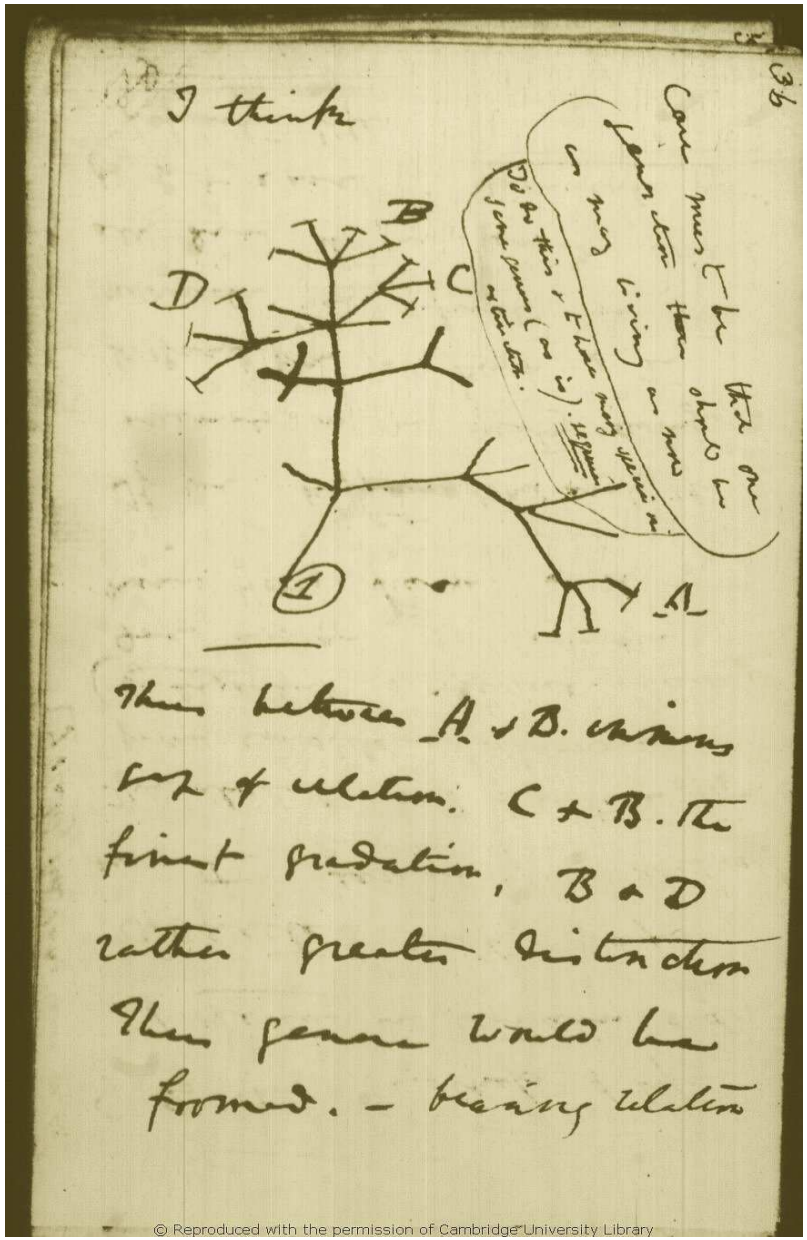
Data l'enormità di informazioni ottenute dal sequenziamento genomico su larga scala (NGS), c'è un grande interesse nel comparare le sequenze di molecole correlate tra specie diverse.

Per correlare differenze nella sequenza con differenze nella funzione (specialmente proteine).

Le sequenze delle macromolecole non contengono solo informazioni sulla loro funzione ma anche informazioni sulla loro storia evolutiva.



Rafael Zardoya, and Axel Meyer PNAS 2001;98:7380-7383



Schizzo di un albero filogenetico, disegnato da Charles Darwin,

Alberi filogenetici

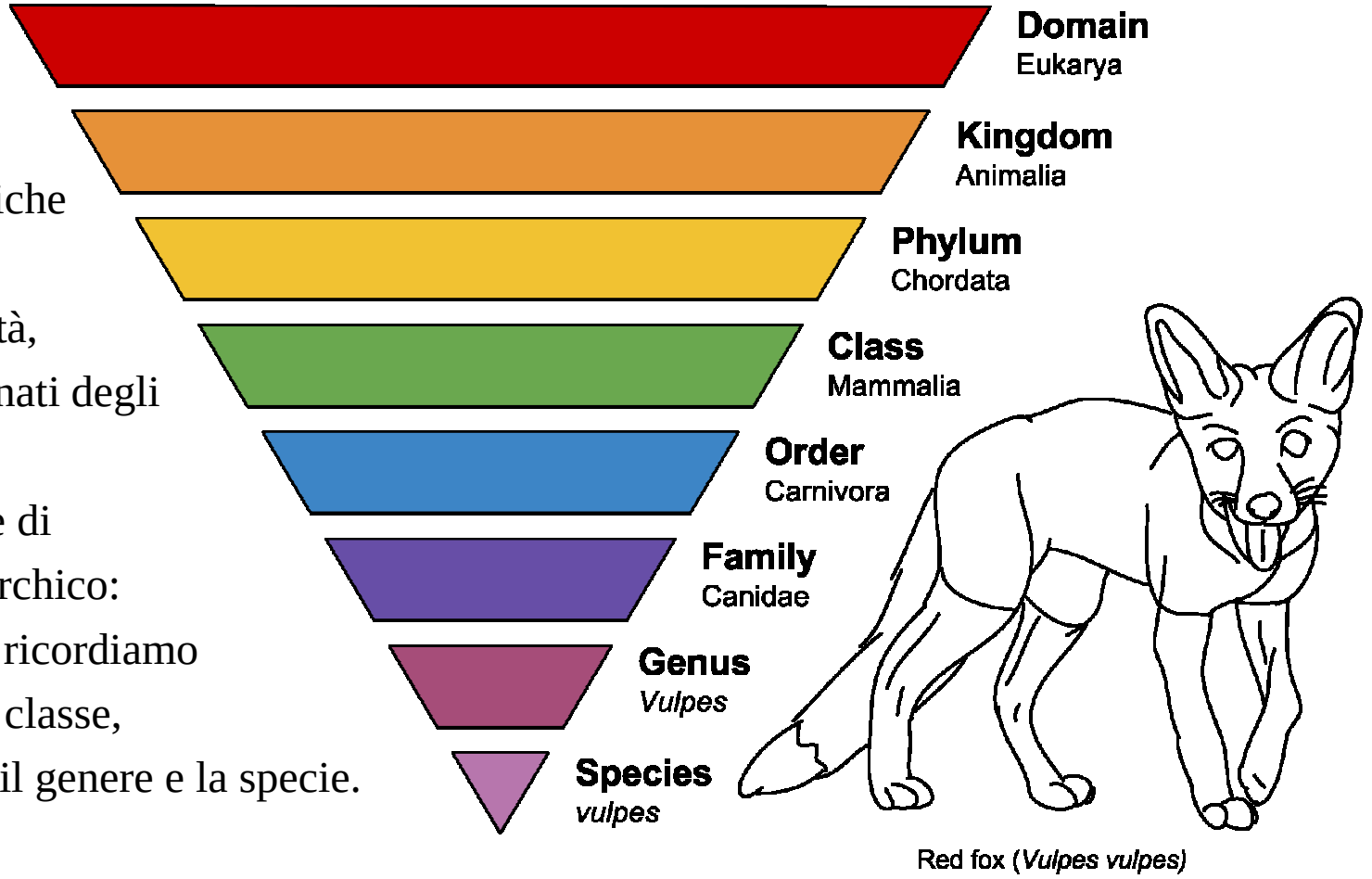
Le relazioni filogenetiche possono essere rappresentate con un diagramma, chiamato albero filogenetico, in modo immediato ci fornisce informazioni sul rapporto genealogico tra le specie in esame e sulla loro evoluzione.

Si ritiene che Charles Darwin abbia usato per primo la metafora dell'albero per rappresentare le relazioni evolutive (Albero, rami, foglie)

Il concetto di Taxa

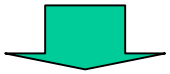
Taxa

In biologia,
le categorie sistematiche
(*taxon*, al singolare)
corrispondenti a entità,
raggruppamenti ordinati degli
esseri viventi.
I taxa possono essere di
qualsiasi livello gerarchico:
in senso decrescente ricordiamo
il *phylum* (o tipo), la classe,
l'ordine, la famiglia, il genere e la specie.

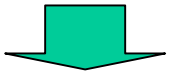


L’albero genetico è un diagramma della storia evolutiva di taxa

Poichè le sequenze cambiano attraverso il tempo



Le differenze fra sequenze si accumulano con il passare del tempo



Il numero delle differenze tra ogni coppia di sequenze può essere utilizzato come una misura del tempo trascorso da quando si sono separate

L’orologio molecolare

L’evoluzione è un processo inevitabilmente divergente e il numero di mutazioni che si accumulano nel tempo è direttamente proporzionale al tempo intercorso dalla divergenza delle sequenze in analisi.

Se ciò è vero, data una distanza genetica calcolata osservando le divergenze, è possibile ottenere il tempo trascorso dal momento in cui due sequenze hanno cominciato a divergere

Inoltre, se la velocità di accumulo delle mutazioni è costante, è possibile la datazione degli organismi in base a un solo dato verificato di distanza temporale

Anche se l’orologio molecolare è vero, non è universale, perché siti diversi hanno diversi tassi di mutazione

L’albero genealogico

Quello genealogico è un classico esempio di albero filogenetico che può aiutarci a capire i modelli di relazione che si applicano più in generale agli alberi filogenetici

Graficamente si vede che vi sono più stretti legami tra fratelli che tra cugini primi. Percorrendo a ritroso il ramo dell’albero vediamo che il più recente antenato comune (Most Recent Common *Ancestor*) tra due fratelli è la madre (o il padre).

Al contrario, il MRCA che si condivide con il primo cugino è la nonna (o il nonno) materna.

Questi principi valgono anche quando si stanno leggendo filogenesi più complesse.

Caratteristiche degli alberi

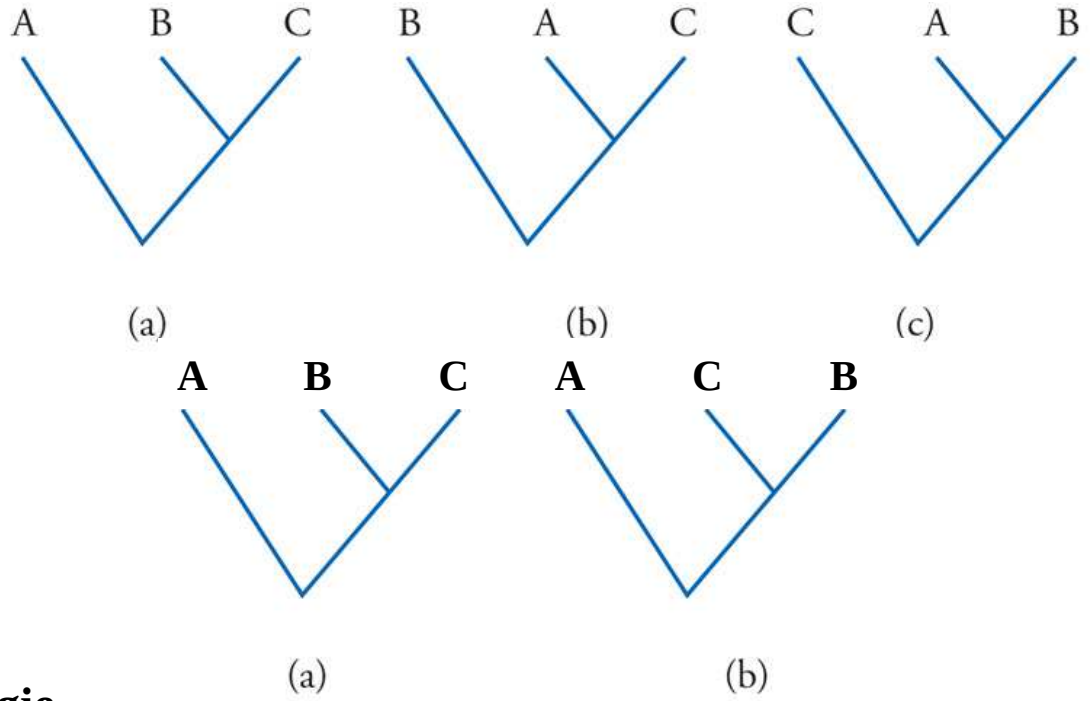
Gli alberi filogenetici hanno una tipica struttura:

Topologia

Rami

Nodi

Topologia



La topologia di un albero indica
la parentela fra i taxa.
Alberi con la stessa topologia e radice
hanno uguale significato filogenetico.

I tre alberi in alto hanno **diverse topologie**.

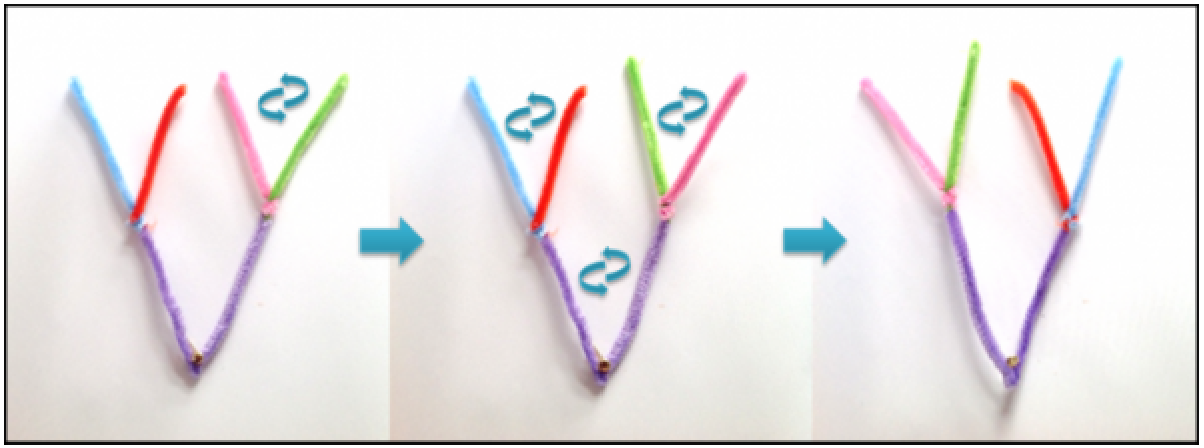
Nell'albero a sinistra, B e C sono più strettamente correlati di quanto lo siano con A.

In quello centrale, sono A e C sono più strettamente correlati.

A destra A e B sono più strettamente correlati.

I due alberi in basso della figura hanno la **stessa topologia**. La tesi "A e B sono più strettamente correlati tra di loro che con C", è vera per i due alberi.

Analizziamo un modello di albero



Se trovate difficile immaginare che alberi **“diversi” possano avere la stessa topologia**, quindi illustrano la stessa filogenesi, può essere utile fare un modello di un albero utilizzando degli scovolini da pipa colorati e un perno.

Gli scovolini rappresentano i rami e il perno indica la radice.

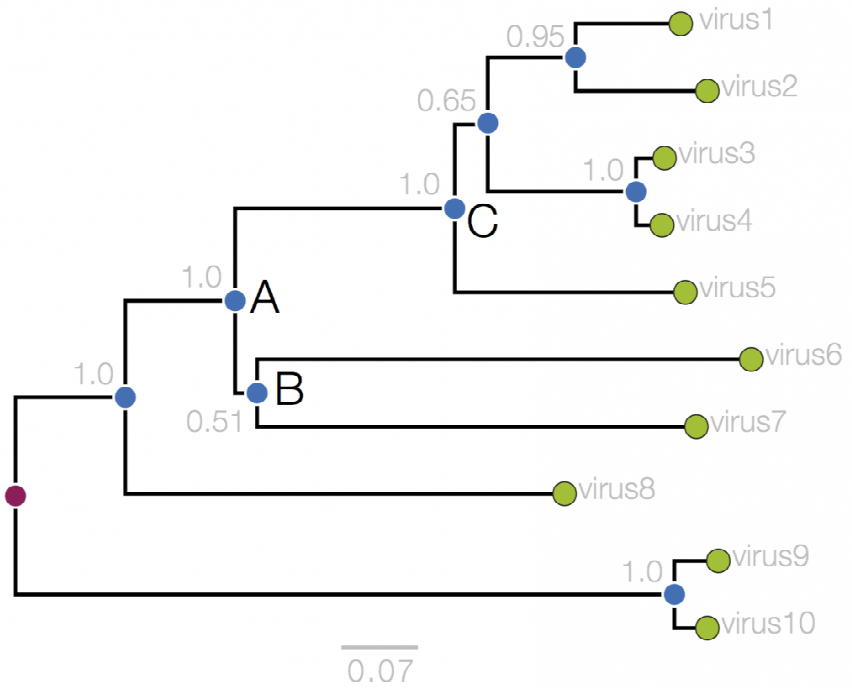
Le rotazioni dei rami intorno alla radice comporta **sempre la stessa topologia**, e quindi lo stesso modello di parentela.

Rami

Definiscono il percorso di trasmissione delle informazioni genetiche da una generazione a quella successiva. **Le lunghezze sono proporzionali al cambiamento genetico**, cioè più è lungo il ramo, maggiori cambiamenti genetici sono avvenuti.

In genere si misura la quantità di cambiamento genetico stimando il **numero** medio di **sostituzioni di nucleotidi o aminoacidi per sito.**

Alla base del diagramma di un albero vi è in genere un segmento che riporta in scala il numero di sostituzioni per sito. A volte le lunghezze sono accanto ad ogni ramo, ma è molto più comune vedere le **lunghezze rappresentate da una barra di scala** (a destra).



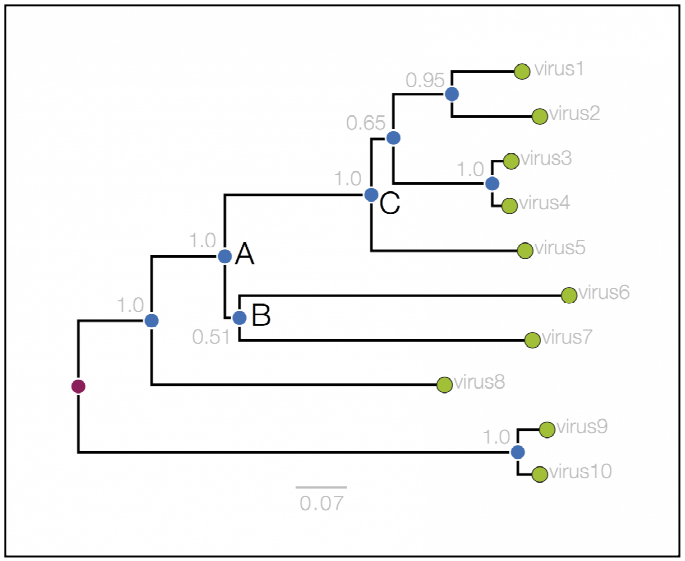
Rami (2) - stima dell'entità del cambiamento

Un metodo semplice è quello di **allineare le coppie di sequenze**, contare il numero di differenze e dividere per la lunghezza della sequenza.

In figura, possiamo vedere che c'è un sito diverso tra le due sequenze, possiamo dire che ci sono 1/10 = **0,1 sostituzioni per sito**. La nostra ipotesi è che ci sia stata solo quella singola sostituzione tra le due sequenze, e quindi non teniamo conto di eventuali sostituzioni multiple che si sono verificati in uno qualsiasi dei siti.

Abbiamo anche ipotizzato che ogni sostituzione (ad esempio, da T> C, o A> G) **ha la stessa probabilità**, anche se sappiamo che questo non è realistico.

Nodi



I nodi sono collocati alle estremità dei rami e rappresentano sequenze reali o ipotetiche a vari livelli della storia evolutiva.

Vi sono tre tipi di nodi:

Nodi esterni (tips):

Rappresentano le sequenze che abbiamo utilizzato per costruire l'albero

Nodi interni:

Sono collocati dove due o più rami si incontrano e rappresentano le (ipotetiche) **recenti sequenze ancestrali comuni**.

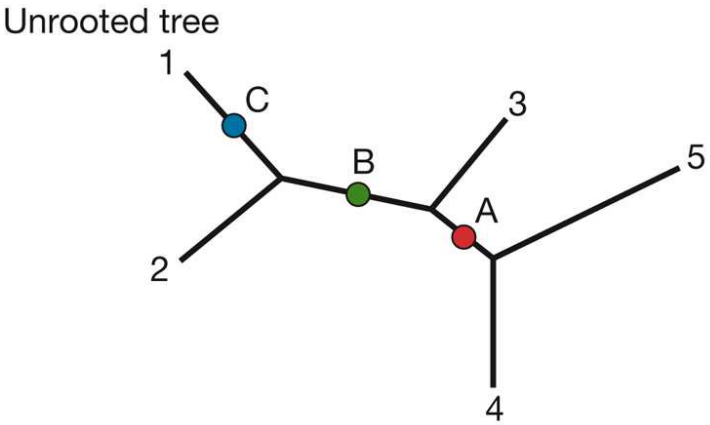
Radice:

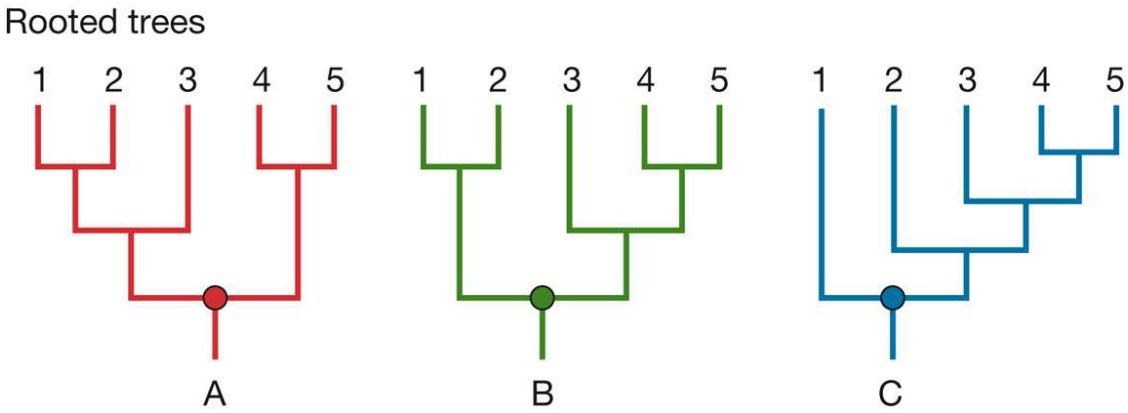
La radice è un importante nodo interno che rappresenta il più recente antenato comune (MRCA) di tutte le sequenze nell'albero filogenetico

Radice

La radice è il più recente antenato comune (ancestore) di tutti i taxa nella struttura. E' quindi **la sequenza più antica dell'albero** e ci indica la direzione dell'evoluzione, con il flusso di informazioni genetiche che partono dalla radice verso i nodi esterni ad ogni generazione successiva.

La maggior parte dei metodi di ricostruzione filogenetica non stimano la posizione della radice, anche perché questo **aumenta in modo esponenziale il numero di possibili alberi**, allungando moltissimo i tempi di calcolo.

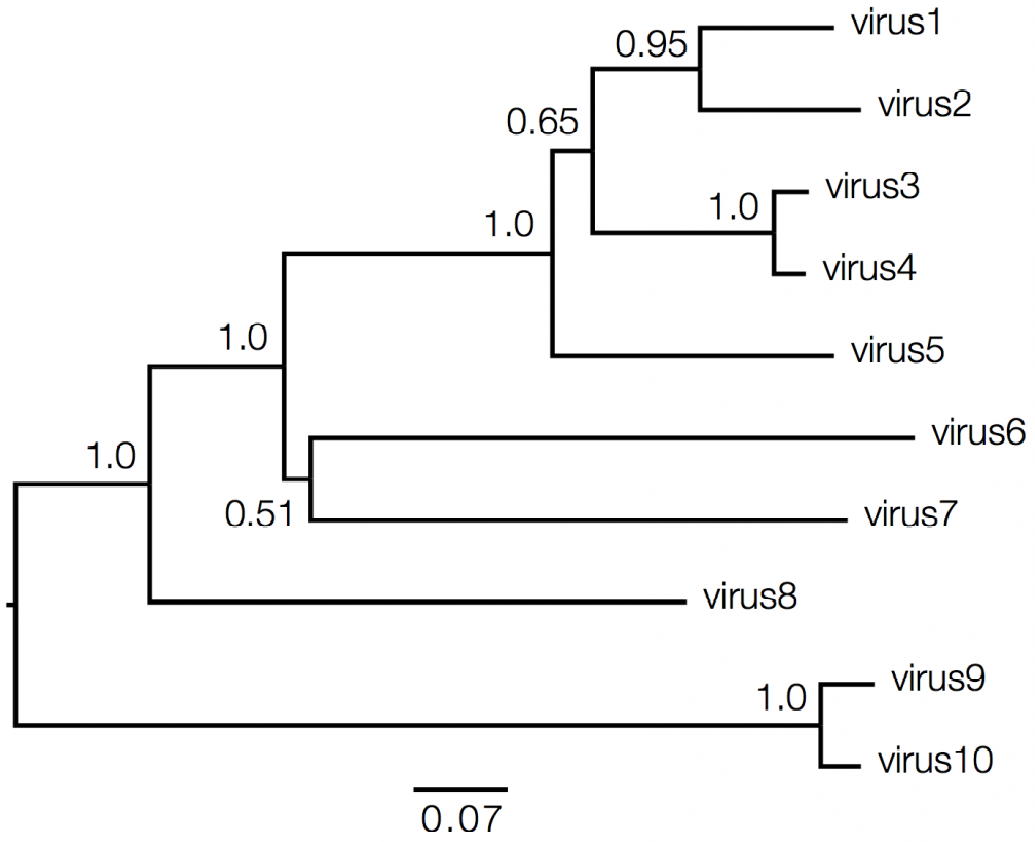




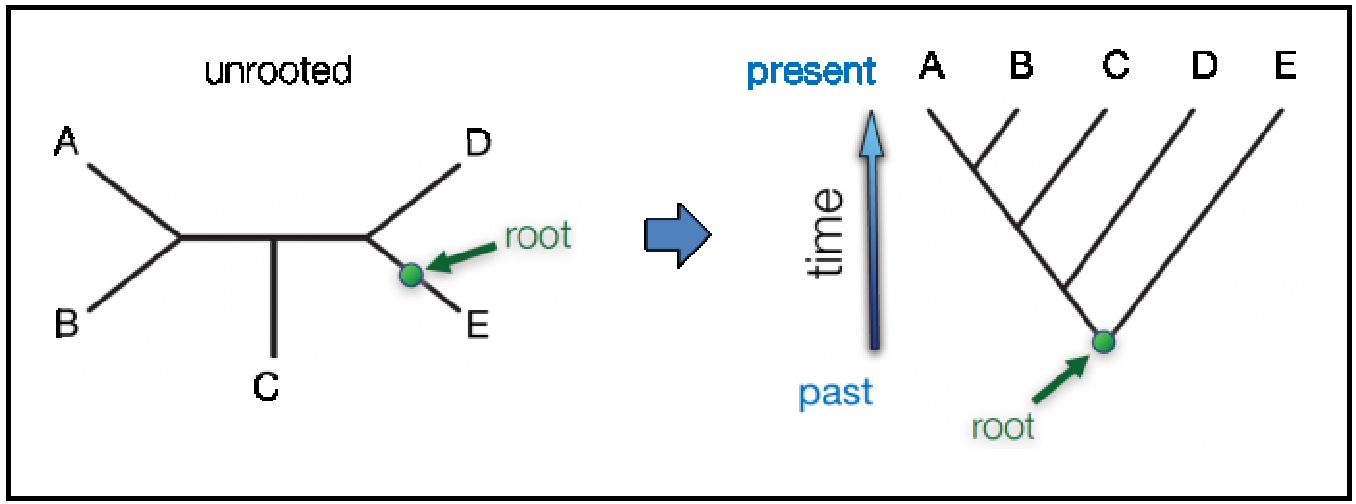
Esempio di albero che è stato successivamente radicato

Il Radicamento di un albero influisce sul suo significato

Dotare un albero filogenetico di una radice appropriata è fondamentale per l'interpretazione filogenetica perché questa ci indica la **direzione di evoluzione** e influisce sui modelli di parentela.



Come radicare un albero



Con Outgroup

E' l'approccio preferito: includere nell'albero una o più sequenze sicuramente correlate alle altre ma piuttosto distanti evolutivamente (outgroup).

La radice è semplicemente il punto in cui il nostro outgroup si unisce al resto dell'albero filogenetico.

I migliori outgroup sono quelli che pur essendo distanti evolutivamente sono più strettamente correlati alle nostre sequenze di interesse. Se gli outgroup sono troppo lontani possono essere inaffidabili, in quanto potrebbero essere difficili da allineare in modo affidabile.

Al Punto Medio

Richiede l'ipotesi che tutte le sequenze stiano evolvendo di pari passo.

In questo caso, la radice è posizionata a metà tra i due rami più lunghi.

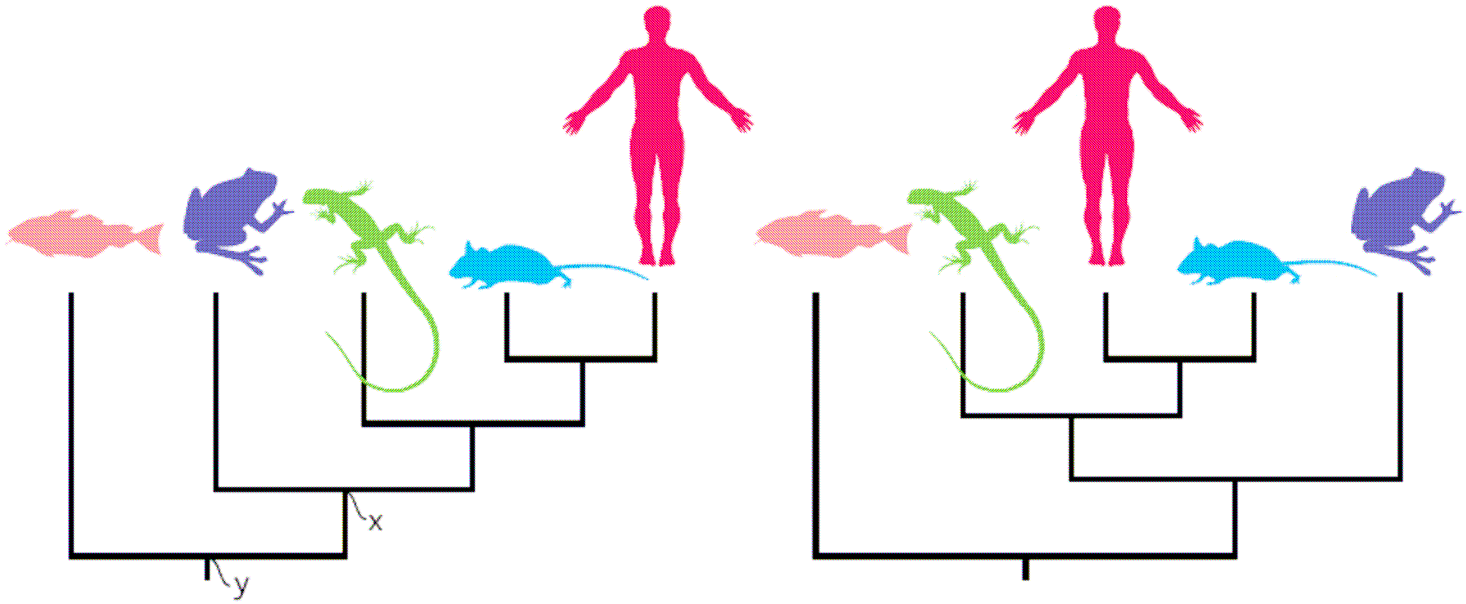
Verifica dell'affidabilità di un albero

L'Inferenza Filogenetica è un processo incerto perché **di solito non abbiamo altre informazioni se non le sequenze odierne dei taxa**, a cui applichiamo dei modelli evolutivi magari dedotti dall'analisi di altre specie.

Alcune sequenze sono più informativo rispetto ad altre, e così ci forniscono una migliore **stima di relazioni genealogiche**.

Stimare l'affidabilità degli alberi ottenuti è un problema complesso. In altri settori della bioinformatica possiamo ottenere parametri quantitativi come sensibilità e specificità, **perché disponiamo di riferimenti "reali"** anche se ottenuti empiricamente.

Questo **non è possibile in caso di filogenesi**, perché **non sono note le sequenze ancestrali**.



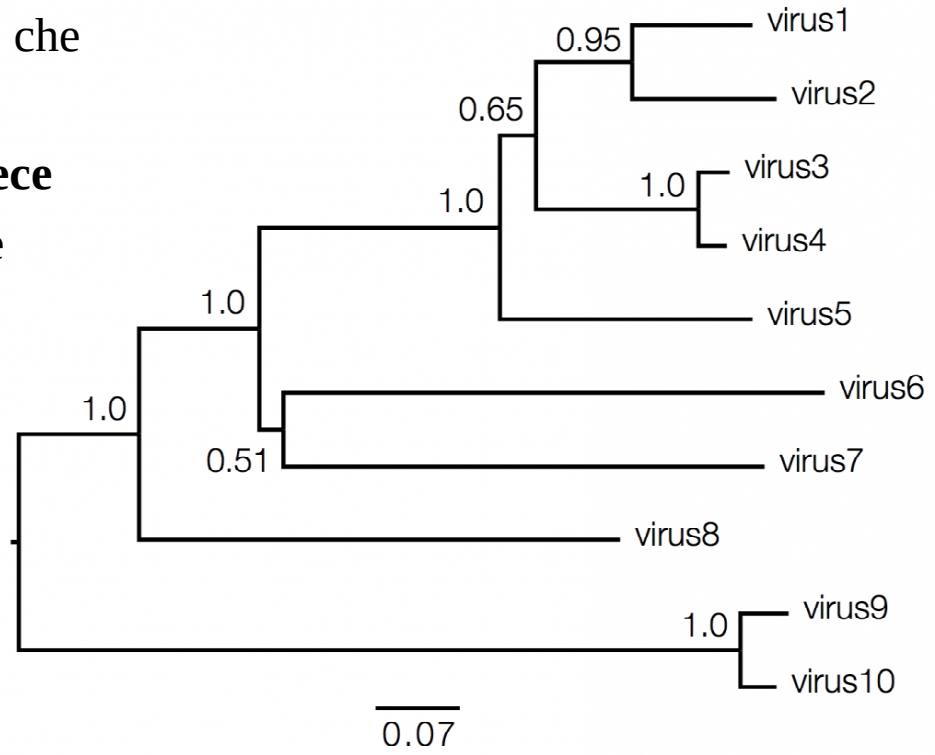
Qual è l'albero più affidabile? Dx o Sin?

Verifica dell'affidabilità di un albero: Bootstrap

E' il test più semplice e più diffuso:

è oramai abbastanza raro vedere un albero **senza i valori del bootstrap** sui rami.

Mediante questa tecnica verifichiamo che il nostro albero sia sostenuto da un intero set di dati e **non sia invece un vincitore marginale** fra molte alternative quasi uguali.



Verifica di un albero filogenetico

Il caso di un albero genealogico è un raro esempio in cui noi di solito conosciamo parte del vero albero biologico; abbiamo una spiegazione completa delle relazioni genealogiche tra i membri della famiglia. Questo è perché sappiamo:
- **tutti gli antenati e discendenti che sono** o sono stati recentemente **viventi**;
- l'**identità dei genitori biologici** di ogni bambino.

Purtroppo, quando vogliamo analizzare i rapporti filogenetici da oggi al passato di sequenze che non sono così strettamente e chiaramente collegate, raramente conosceremo il vero albero.

Ciò è dovuto al fatto che non sappiamo cosa si è realmente verificato nelle sequenze degli ancestori, e quali cambiamenti genetici si sono verificati per rendere le sequenze così come sono oggi.

Questo è il motivo per cui abbiamo bisogno di metodi di ricostruzione filogenetica per dedurre il vero albero.

Le nostre sono sempre ipotesi!

Grazie per l'attenzione