

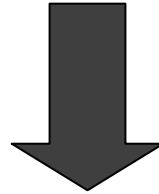
Metodi di Distanza



G.Allegrucci riproduzione vietata

La misura più semplice della distanza tra due sequenze nucleotidiche è contare il numero di siti nucleotidici che differiscono tra le due sequenze

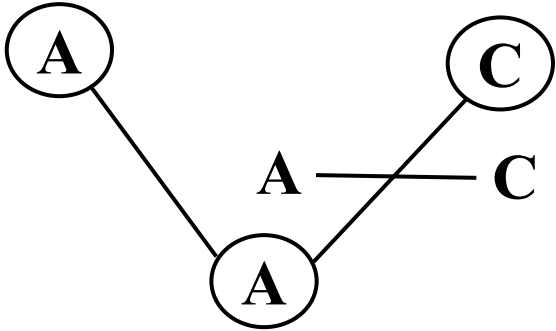
Quando confrontiamo siti omologhi in 2 sequenze di DNA, osserviamo semplicemente se le sequenze sono le stesse o no



Il numero massimo di differenze per sito che noi possiamo osservare è uno.

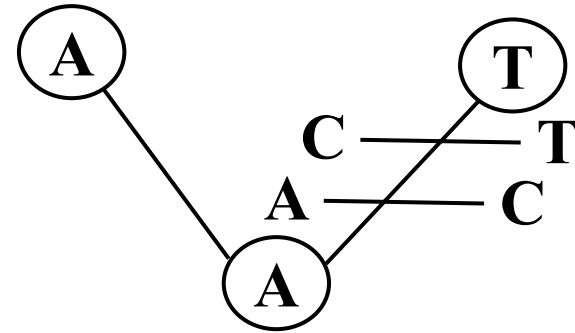
Ciò significa che se più di una sostituzione è avvenuta ad un sito noi perdiamo l'informazione della precedente sostituzione

Sostituzione singola

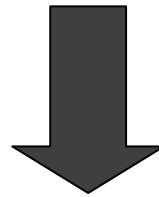


1 mutazione, 1 differenza

Sostituzione multipla

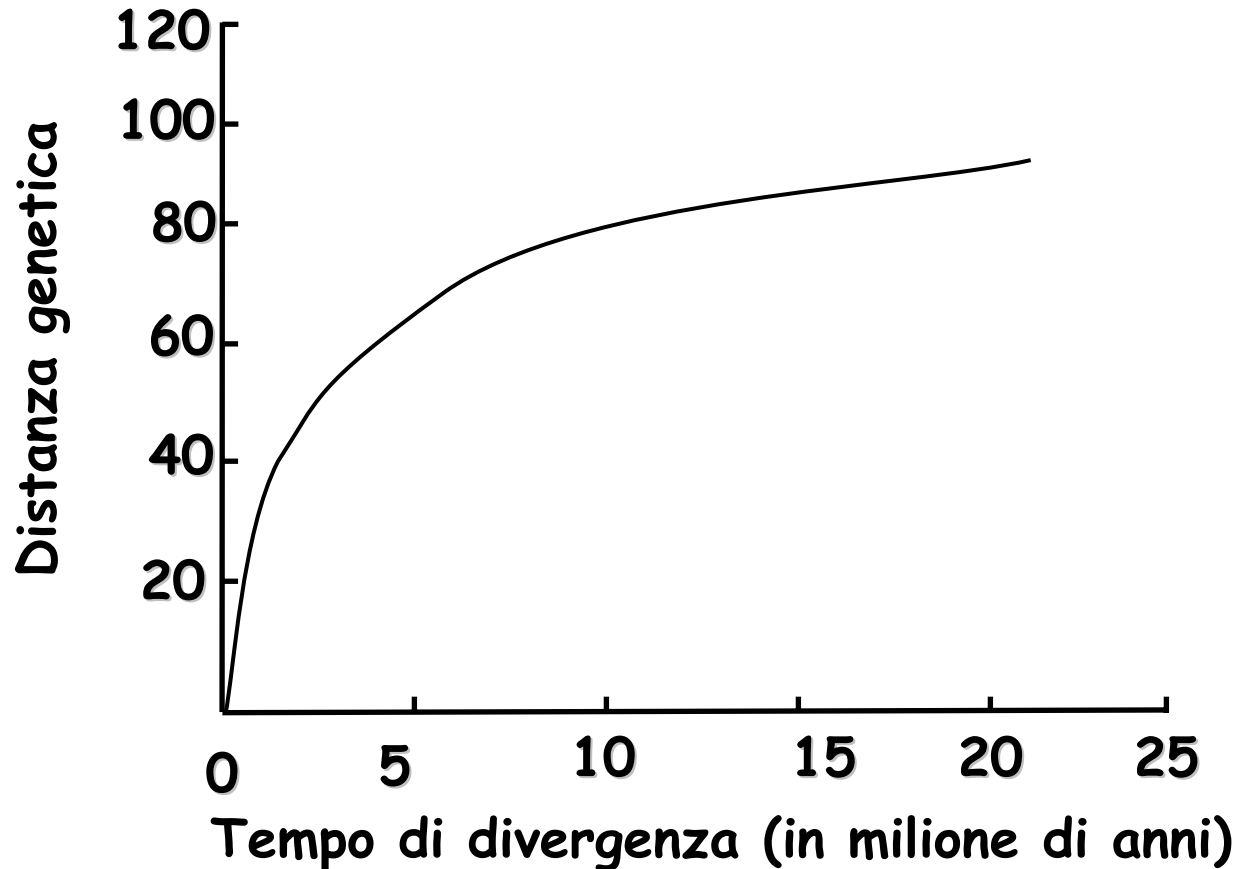


2 mutazioni, 1 differenza



La semplice conta del numero di differenze tra sequenze (p distance= $n.\text{sostituzioni}/n.\text{totale di basi considerate}$) può sottostimare la quantità di cambiamento, specialmente se queste sono poco simili, a causa dei molteplici cambiamenti.

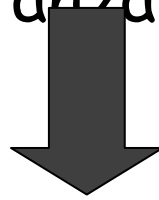
La relazione tra la distanza genetica e il tempo di divergenza non è lineare perchè lo stesso sito può aver subito più sostituzioni con il passare del tempo



Quando si accumulano più sostituzioni tra le due sequenze esse diventano progressivamente saturate, più di un sito è sottoposto a sostituzioni multiple

A causa delle sostituzioni multiple, le distanze osservate possono sottostimare il reale ammontare del cambiamento evolutivo.

Sono stati, perciò, sviluppati diversi metodi che convertono le distanze osservate nella "reale" misura della distanza evolutiva



***MODELLI EVOLUTIVI
(METODI DI CORREZIONE DELLA DISTANZA)***

**"Correggono" la distanza osservata
valutando l'ammontare del
cambiamento evolutivo**

Considerando che la probabilità di sostituzione di un dato nucleotide è costante nel tempo e che la composizione in basi della sequenza è in equilibrio otteniamo

MATRICE PROBABILITA' DI SOSTITUZIONE

$$\mathbf{P}_t = \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}$$

p_{AC} è la probabilità che A muti in C nell'intervallo t

In molti modelli la matrice è simmetrica ossia $p_{AC} = p_{CA}$

Modelli di evoluzione di sequenze comunemente usati nell'analisi di distanza:

- Modello di Jukes-Cantor

Le 4 basi hanno uguale frequenza e tutte le sostituzioni
sono ugualmente probabili

$$P_t = \begin{pmatrix} - & \alpha & \alpha & \alpha \\ \alpha & - & \alpha & \alpha \\ \alpha & \alpha & - & \alpha \\ \alpha & \alpha & \alpha & - \end{pmatrix}$$

α è la probabilità di una sostituzione

Il modello di Jukes-Cantor è il più semplice:

$$d_{xy} = -(3/4) \ln (1 - 4/3 D)$$

- ❑ d_{xy} = distanza fra la sequenza x e la sequenza y, espressa come numero di cambiamenti per sito
- ❑ D = proporzione osservata di nucleotidi che differiscono fra due sequenze (dissimilarità frazionaria)
- ❑ $\ln$ = log naturale usato per correggere le sostituzioni ripetute
- ❑ I termini 3/4 e 4/3 indicano che ci sono quattro tipi di nucleotidi e tre modi in cui un secondo nucleotide può o meno essere uguale al precedente - con tutti i tipi di cambiamento ugualmente probabili (cioè, sequenze non affini dovrebbero essere identiche per il 25% solo per effetto del caso).

Il logaritmo naturale è usato per correggere i cambiamenti multipli nello stesso sito

- Se due sequenze sono identiche per il 95%, sono diverse per il 5% o hanno una D pari a 0.05, perciò:

$$\text{➤ } d_{xy} = -3/4 \ln (1-4/3 \cdot 0.05) = 0.0517$$

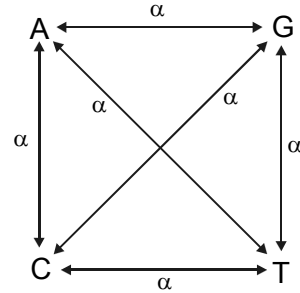
- Da notare che la dissimilarità osservata di 0.05 aumenta soltanto leggermente, essendo la distanza stimata pari a 0.0517 - ciò ha un senso perchè in due sequenze molto simili ci si aspettano pochi cambiamenti multipli nello stesso sito, poichè il tempo di divergenza è breve.
- Comunque, se due sequenze sono identiche soltanto per il 50%, sono diverse per il 50% o hanno una D pari a 0.50, perciò:
$$\text{➤ } d_{xy} = -3/4 \ln (1-4/3 \cdot 0.5) = 0.824$$
- Per sequenze dissimili, che possono essersi diversificate da lungo tempo, l'uso del logaritmo naturale mette in risalto il fatto che ci possono essere stati diversi cambiamenti multipli nello stesso sito.

Per aumentare il realismo dei modelli di distanza si possono considerare ulteriori parametri

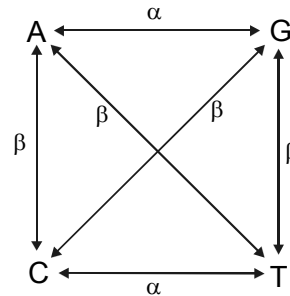
- E' meglio usare un modello che sia conforme ai dati piuttosto che imporre, alla cieca, un modello sui dati
- I parametri più comuni che vengono aggiunti sono:
 - ✓ Una correzione per la proporzione di siti invarianti
 - ✓ Una correzione per i tassi di variazione per i siti variabili
 - ✓ Una correzione che permetta tassi di sostituzione differente per ogni tipo di cambiamento nucleotidico
- PAUP è il programma in grado di stimare tutti questi parametri

Modelli di evoluzione delle sequenze

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

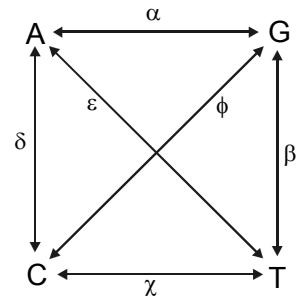


Felsenstein (1981)



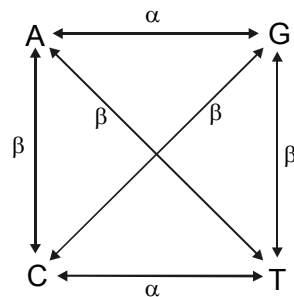
Hasegawa, Kishino & Yano (HKY) (1985)

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



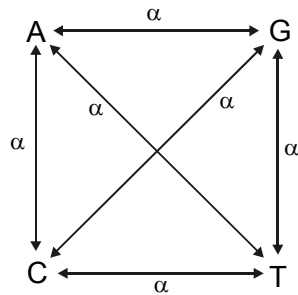
General time reversible model (GTR)

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

Vantaggi dell'analisi di distanza

- ❖ Veloce - adatta per analizzare set di dati grandi
- ❖ Sono disponibili molti modelli che migliorano le stime di distanza

Svantaggi dell'analisi di distanza

- Si perde informazione - date soltanto le distanze è impossibile risalire alle sequenze originali
- In alcuni casi i branch lengths sono validi da un punto di vista matematico ma non da uno biologico

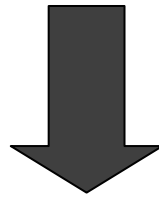
ES:

2.5% tasso di sostituzione per milione di anni quindi in una seq. di 100 bp ci aspettiamo una media di 2.5 sostituzioni

Biologicamente non possiamo avere mezza sostituzione!!!!

Ottenere un albero utilizzando le misure di distanza

Abbiamo dati (distanze osservate) che vogliamo adattare ad un albero e trovare il valore ottimale per determinati parametri (pattern di ramificazione e lunghezza dei rami)



METODI DI DISTANZA

- Neighbour Joining
- UPGMA

Distanze Osservate-Tree Distance

- Distanze Osservate- ottenute direttamente dalle sequenze
- Tree Distance- ottenute dall'albero

In realtà le distanze osservate e le tree distance non coincidono esattamente perchè le prime non possono essere completamente rappresentate dalle seconde, la loro discrepanza può essere utilizzata per misurare quanto l'albero migliore è adatto a rappresentare le distanze osservate



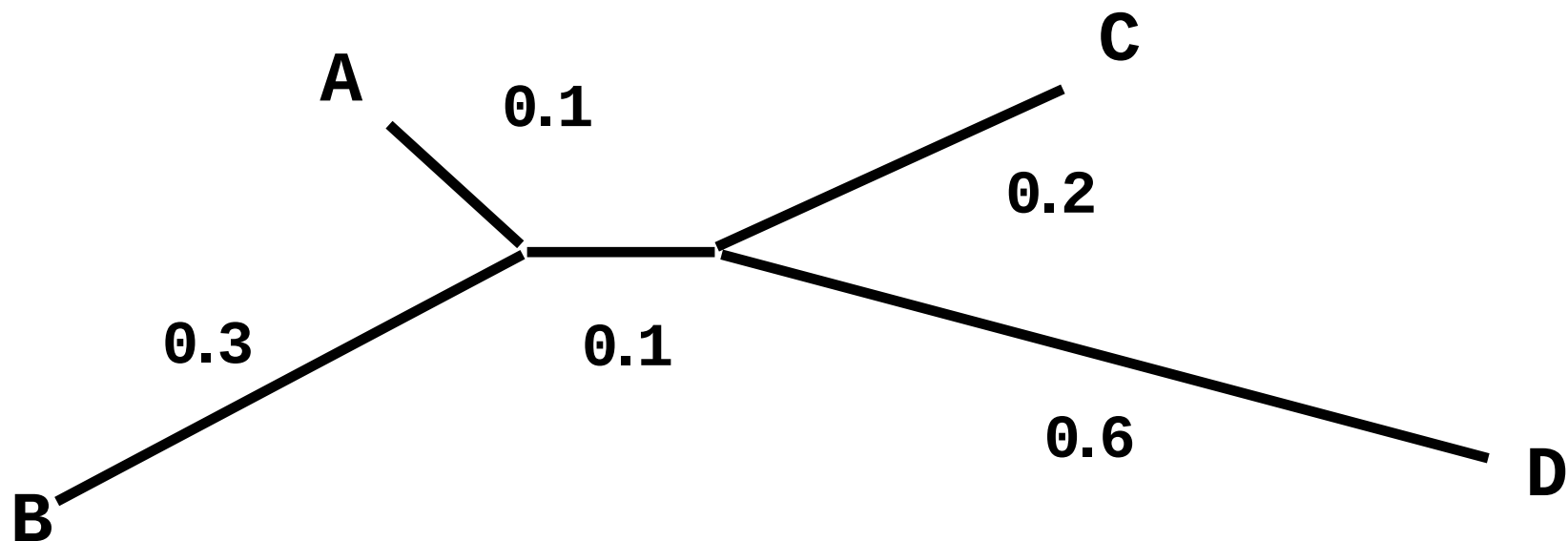
- Goodness of Fit Measures
- Minimum evolution (Metodo dell'evoluzione minima)
- Least squares (Metodo dei minimi quadrati)

Ottenere un albero utilizzando le misure di distanza

Distanze Additive :

Se si potesse determinare esattamente la vera distanza evolutiva corrispondente ad un certo cambiamento osservato fra le sequenze di ogni coppia di taxa sotto studio, avremmo che la lunghezza delle braccia dell'albero saranno proporzionali al cambiamento evolutivo (albero additivo)

Un albero additivo perfetto



	A	B	C	D
A	-	0.4	0.4	0.8
B	0.4	-	0.6	1.0
C	0.4	0.6	-	0.8
D	0.8	1.0	0.8	-

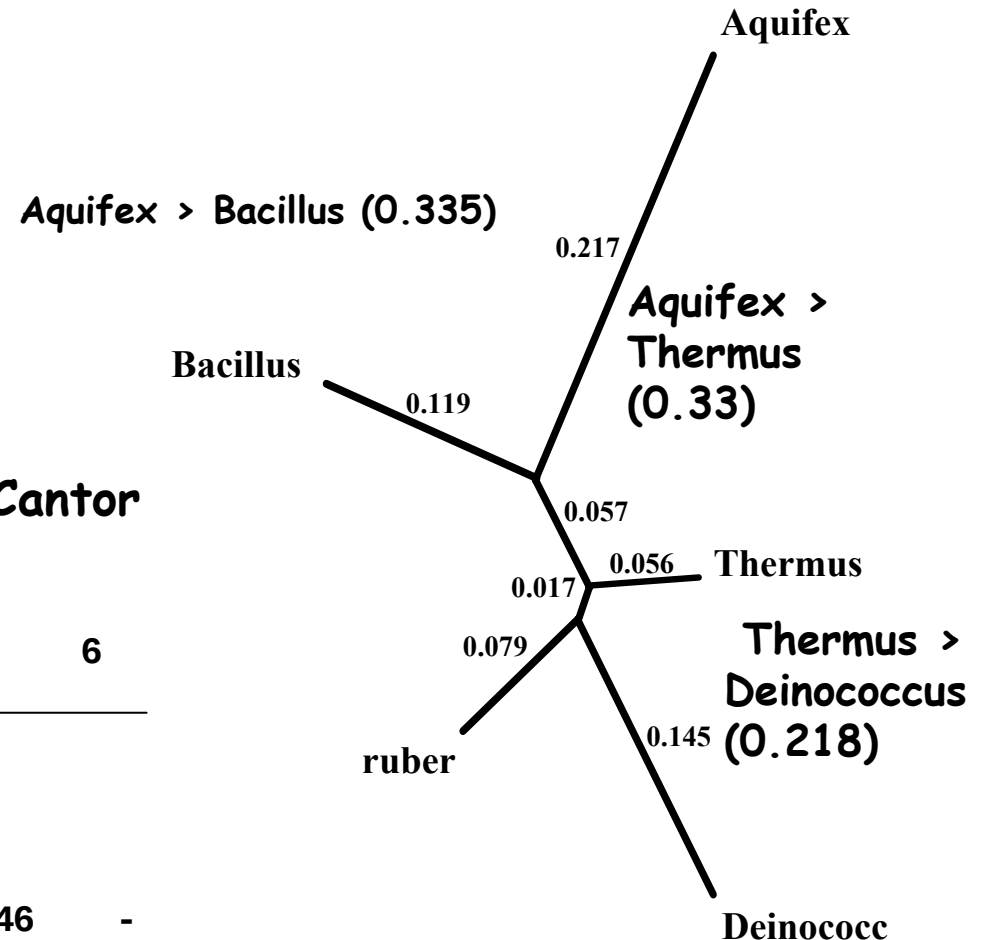
Le lunghezze dei rami nella matrice e le lunghezze dei sentieri sull'albero corrispondono esattamente: c'è un solo e unico albero additivo

Le stime di distanza possono formare un albero *non* additivo

Alcuni percorsi sono più lunghi ed altri più corti, rispetto alla matrice di distanze.

Matrice di distanza secondo Jukes-Cantor

	1	2	4	5	6
1 ruber	-				
2 Aquifex	0.38745	-			
4 Deinococc	0.22455	0.47540	-		
5 Thermus	0.13415	0.27313	0.23615	-	
6 Bacillus	0.27111	0.33595	0.28017	0.28846	-



Ottenere un albero utilizzando le misure di distanza

- In alcuni casi le distanze dell'albero (tree distance) sono più lunghe di quelle osservate a causa di omoplasia (es. sostituzioni multiple)
- Errori casuali possono causare deviazioni dalle distanze stimate da un albero additivo perfetto anche quando l'evoluzione è in accordo con il modello scelto
- Stime ottenute utilizzando un modello inappropriato complicano il problema