

Maximum Likelihood

- Potrebbe essere chiamato massima probabilità.
- E' il metodo più nuovo
- E' stato reso popolare da Joseph Felsenstein, Seattle, Washington.
- Il suo lento apparire nella comunità scientifica è dovuto sia alla difficoltà della teoria sia all'assenza, quantomeno iniziale, di software di buona qualità che permettessero la scelta di modelli e una facile interazione con i dati.
- Inoltre, all'inizio, era praticamente impossibile trattare grandi matrici di dati, considerando che alla metà degli anni 80, un PC tipico aveva un processore con una velocità inferiore a 30 MHz.
- Negli ultimi anni sono stati superati tutti i problemi tecnici e ML è diventato un metodo correntemente usato.

ML: confronto con gli altri metodi

- ✓ ML è simile a molti altri metodi in molti aspetti.
- ✓ In molti aspetti è fondamentalmente diverso.
- ✓ ML assume un modello di evoluzione, come la Massima Parsimonia e come i metodi di distanza.
- ✓ ML tenta di rispondere alla domanda: Qual è la probabilità di osservare questi dati, dato un particolare modello di evoluzione?
- ✓ ML usa un "modello". Ciò è possibile poichè può essere dimostrato che i dati di sequenza variano secondo un processo stocastico.

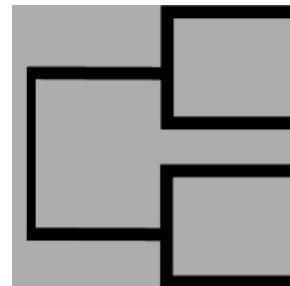
Maximum Likelihood - scopo

- Stimare la probabilità di osservare un particolare set di dati, a partire da un albero filogenetico e da qualche notizia su come il processo evolutivo ha lavorato nel tempo.

Probabilità di

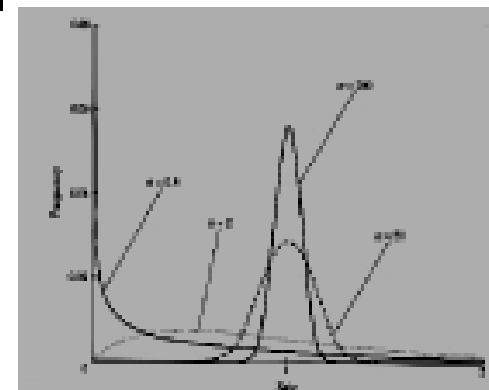
```
Seq1 aacg  
seq2 accg  
seq3 aaca  
seq4 aatg
```

dato


$$\begin{Bmatrix} a & b & c & d \\ b & a & e & f \\ c & e & a & g \\ d & c & f & a \end{Bmatrix}$$

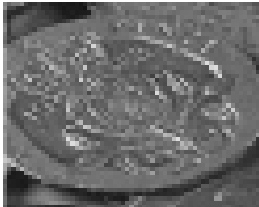
$$\pi = [a, c, g, t]$$

Giuliana Allegrucci - Riproduzione vietata



Qual è la probabilità di osservare un dato?

- Se lanciamo una moneta e otteniamo testa, e sappiamo che la moneta ha due facce (testa e croce) e che il tiro è bilanciato, la probabilità di osservare testa è uguale a 0.5.
- Se invece sappiamo che il tiro della moneta è sbilanciato e ci aspettiamo di avere testa l'80% delle volte, allora la probabilità di osservare questo dato è uguale a 0.8.
- Perciò: La probabilità di fare certe osservazioni è completamente dipendente dal modello sottostante le nostre assunzioni.

$$p \text{  = ?$$

Osservazione: Il dato non è cambiato, il modello sì.
Sotto il nuovo modello la probabilità di osservare il dato è cambiata.

Qual è la probabilità di osservare un nucleotide 'G'

- **Domanda:** Se abbiamo una sequenza di DNA lunga un solo nucleotide e l'identità di questo nucleotide è 'G', qual è la probabilità di osservare 'G'?
- **Risposta:** Come nel caso del lancio della moneta, la probabilità di osservare 'G' dipende dal modello di evoluzione ammesso.
- **Cioè:**
 - Modello 1: frequenza of $G = 0.4 \Rightarrow$ probabilità (G) = 0.4
 - Modello 2: frequenza of $G = 0.1 \Rightarrow$ probabilità (G) = 0.1
 - Modello 3: frequenza of $G = 0.25 \Rightarrow$ probabilità (G) = 0.25

La somma delle probabilità di tutte le possibilità è sempre uguale ad 1.

Es. Per il DNA $p(a)+p(c)+p(g)+p(t)=1$

Che cosa succede con sequenze più lunghe?

- Consideriamo un gene di lunghezza 2:
Gene 1: ga
- La probabilità di osservare questo gene è uguale al *prodotto* delle probabilità di osservare ogni carattere.
- Es.
 - $p(g) = 0.4$; $p(a) = 0.15$ (per esempio)
 - probabilità (ga) = $0.4 \times 0.15 = 0.06$

...o sequenze ancora più lunghe?

- **Gene 1:** gactagctagacagatacgaattac
- **Modello** (simple base frequency model):
 - $p(a)=0.15$; $p(c)=0.2$; $p(g)=0.4$; $p(t)=0.25$;
 - (la somma di tutte le probabilità deve essere uguale a 1)

Probabilità (*Gene* 1) =
0.0000000000000000000018452813

Osservazioni sui modelli

- Il modello usato è quello ottimale, per i dati osservati?

$$(p(a)=0.15; p(c)=0.2; p(g)=0.4; p(t)=0.25)$$

Proviamo ad usare quest'altro modello:

$$p(a)=0.4; p(c)=0.2; p(g)=0.2; p(t)=0.2;$$

La probabilità di osservare questo gene è:

Probabilità (gene 1) =

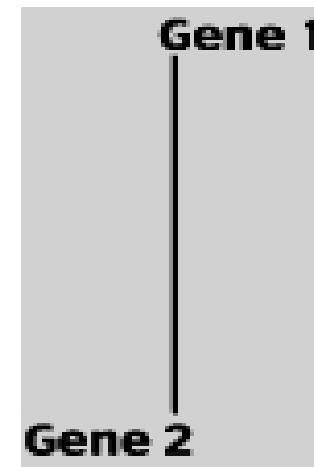
0.0000000000000335544320000

(un valore che è quasi 10.000 volte più grande)

Osservazione: Il dato non è cambiato, il modello sì.
Sotto il nuovo modello la probabilità di osservare il dato è cambiata.

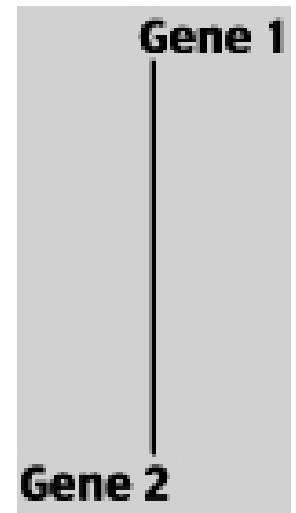
Come tutto ciò si collega agli alberi filogenetici?

- Consideriamo un allineamento di due sequenze:
 - Gene 1: gaac
 - Gene 2: gacc
- Assumiamo che questi geni siano collegati da un semplice albero filogenetico con lunghezze dei rami differenziate.



Utilizzare un modello più sofisticato

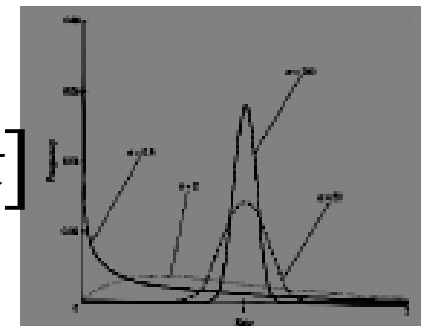
- Oltre a pensare alla composizione in basi, dobbiamo includere nel nostro modello anche il cambiamento in sequenza e la stasi.
- Ci sono due parti in questo modello - l'albero e il processo (quest'ultimo è erroneamente pensato come modello, sebbene il modello sia costituito anche dall'albero).



Il modello

- Il modello è costituito da due parti: l'albero e il processo evolutivo sottostante.
- Il processo prende in considerazione la composizione in basi e il processo di sostituzione - il tasso di cambiamento da uno stato del carattere ad un altro.

$$\text{Modello } p = \text{[Diagramma dell'albero]} + \begin{Bmatrix} a & b & c & d \\ b & a & e & f \\ c & e & a & g \\ d & c & f & a \end{Bmatrix} \pi = [a, c, g, t]$$



Simple “time-reversible” model

- Un modello semplice prevede che il tasso di cambiamento da a a c o viceversa è 0.4, la composizione di a è 0.25 e quella di c è 0.25 (è questa una versione semplificata del modello di Jukes and Cantor, 1969)

$$P = \begin{Bmatrix} . & 0.4 & . & . \\ 0.4 & . & . & . \\ . & . & . & . \\ . & . & . & . \end{Bmatrix} \quad \pi = [0.25 \quad 0.25 \quad . \quad .]$$

Probabilità di trovare la terza posizione del nucleotide nel nostro allineamento

- $p(a) = 0.25$; $p(c) = 0.25$; $p_{a \rightarrow c} = 0.4$

Partendo da a, la probabilità di trovare il nucleotide a è 0.25 e la probabilità di sostituzione (ramo) è 0.4. Quindi la probabilità di osservare questi dati é:

* $\text{Likelihood}(D|M) = 0.25 \times 0.4 = 0.01$

Nota: Dal momento che il modello è reversibile, avremmo ottenuto lo stesso risultato se fossimo partiti da c

* likelihood dei dati, dato il modello

Matrice di sostituzione

- Per le sequenze nucleotidiche, ci sono 16 possibili modi di descrivere le sostituzioni, ovvero una matrice 4x4.

$$P = \begin{Bmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \\ m & n & o & p \end{Bmatrix}$$

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

Nota: per gli aminoacidi, la matrice è 20x20 e per i codoni la matrice è 61x61

Un esempio di matrice di sostituzione

$$P = \begin{Bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ 0.002 & 0.983 & 0.005 & 0.01 \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ 0.003 & 0.01 & 0.979 & 0.007 \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ 0.002 & 0.013 & 0.005 & 0.979 \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{Bmatrix}$$

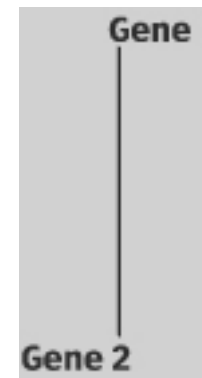
- In questa matrice la probabilità che A cambi in C è 0.01 e la probabilità che C non muti è 0.983, etc...

Nota: Le righe di questa matrice hanno somma =1. Ciò significa che per ogni nucleotide abbiamo considerato tutti i possibili cambiamenti. La somma delle colonne non ha significato.

Calcolare la likelihood di un dataset completo, a partire da una matrice di sostituzione, una certa composizione in basi e una lunghezza dei rami ottenuta da un particolare indice di distanza evolutiva

$$P = \begin{Bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{Bmatrix}$$

Likelihood di Gene 1: ccat
Gene 2: ccgt dato



$$\pi = [0.1, 0.4, 0.2, 0.3]$$

Likelihood di un allineamento di due sequenze.

- ccat

- ccgt

$$\pi_c P_{c \rightarrow c} \pi_c P_{c \rightarrow c} \pi_a P_{a \rightarrow g} \pi_t P_{t \rightarrow t}$$

$$= 0.4 \times 0.983 \times 0.4 \times 0.983 \times 0.1 \times 0.007 \times 0.3 \times 0.979$$
$$= 0.0000300$$

La likelihood di passare dalla prima alla seconda sequenza è 0.0000300

Lunghezze dei rami differenti

- Quando le lunghezze dei rami sono molto corte, la probabilità che un carattere rimanga nello stesso stato è alta, mentre è bassa la probabilità che cambi.
- Per lunghezze dei rami più lunghe, si alza la probabilità di cambiamento e si abbassa quella di non cambiamento.
- I calcoli precedenti si basano sull'assunzione che la lunghezza dei rami descriva una particolare distanza evolutiva.
- Se vogliamo considerare una lunghezza dei rami doppia (ovvero distanza evolutiva doppia), dobbiamo moltiplicare la matrice di sostituzione per se stessa (matrice²).

Modello distanza evolutiva doppia

$$P = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix} \times P = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}$$
$$= \begin{bmatrix} 0.953 & 0.02 & 0.013 & 0.015 \\ 0.005 & 0.966 & 0.015 & 0.029 \\ 0.01 & 0.029 & 0.939 & 0.022 \\ 0.007 & 0.038 & 0.015 & 0.94 \end{bmatrix}$$

La likelihood risultante è pari a 0.0000559

Likelihood precedente = 0.0000300

distanza evolutiva tripla

$$P^3 = \begin{bmatrix} 0.93 & 0.029 & 0.019 & 0.022 \\ 0.007 & 0.949 & 0.015 & 0.029 \\ 0.01 & 0.029 & 0.939 & 0.022 \\ 0.007 & 0.038 & 0.015 & 0.94 \end{bmatrix}$$

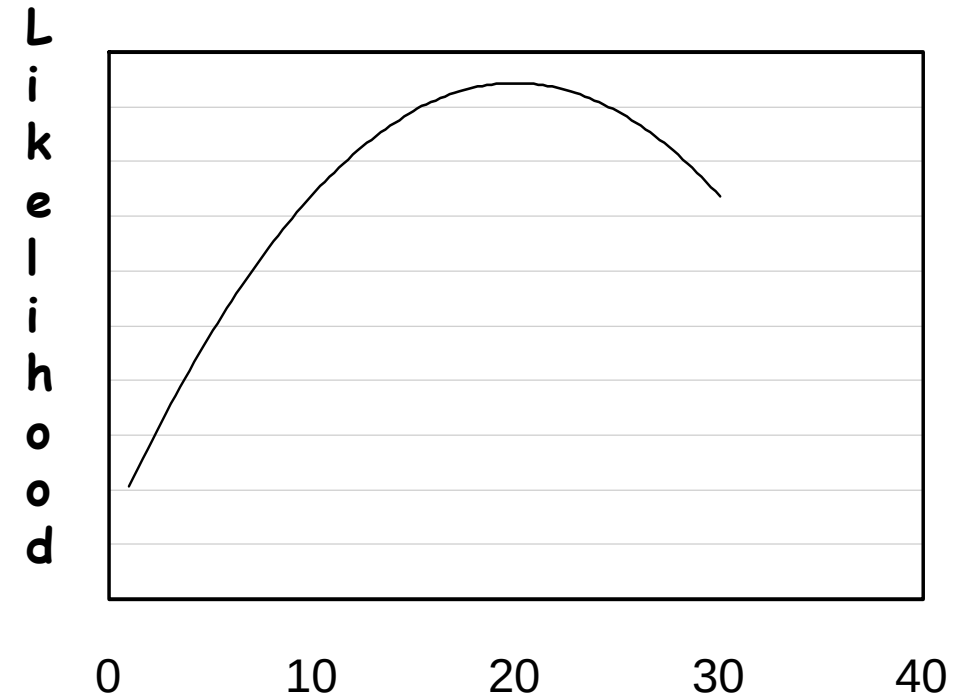
La likelihood risultante è uguale a 0.0000782

Bisogna notare che all'aumentare della lunghezza dei rami, diminuiscono i valori sulla diagonale e aumentano gli altri.

Per valori più alti di distanza evolutiva

ccat
ccgt

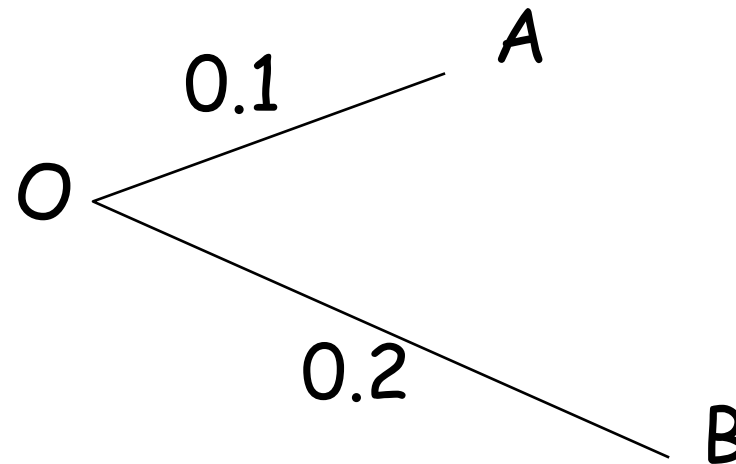
1	0.0000300
2	0.0000559
3	0.0000782
10	0.0001620
15	0.0001770
20	0.0001750
30	0.0001520



Lunghezza dei rami

Il valore massimo della likelihood è 0.0001777, quando
la lunghezza dei rami è 0.330614

Likelihood di un albero con due rami

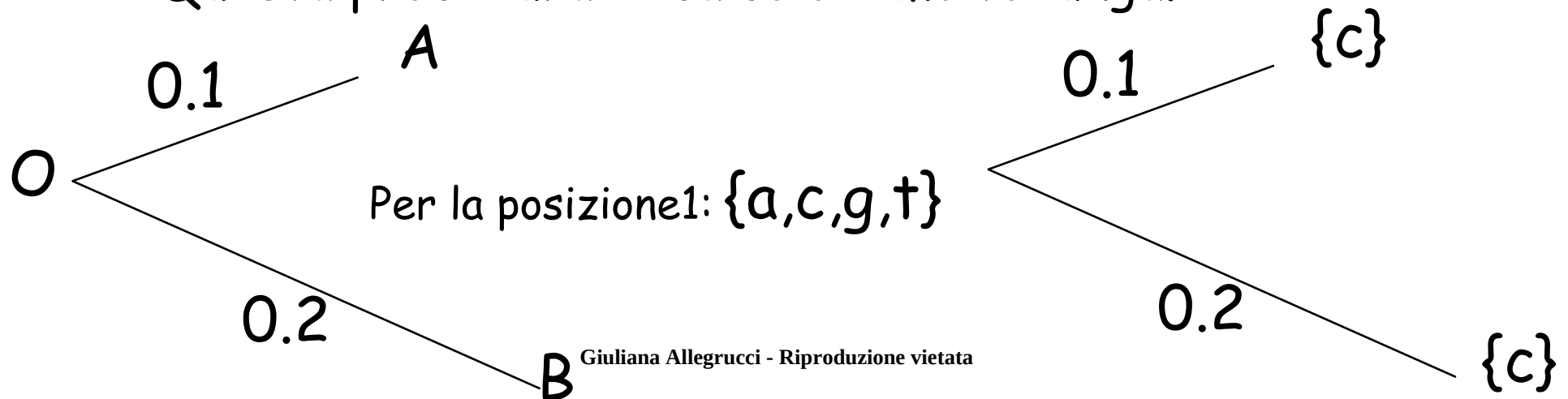


O è l'origine o radice, i numeri rappresentano le lunghezze dei rami. La likelihood può essere calcolata in tre modi:

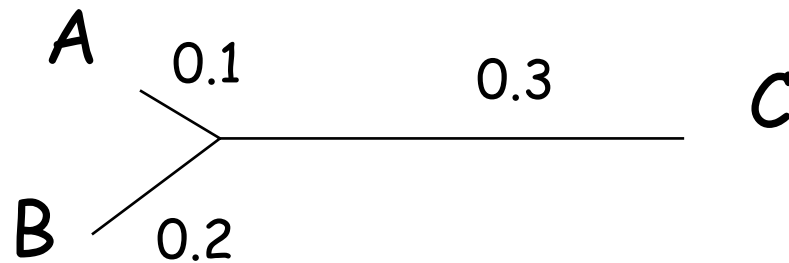
- da A a B in un passaggio
- da A a B in due passaggi (attraverso O)
- in due parti a partire da O .

Osservazioni su O

- O è una sequenza sconosciuta.
- Possiamo solo ipotizzare che ogni posizione nell'allineamento sia uguale a quello che osserveremmo se conoscessimo la sequenza O .
- Ciò che sappiamo è che la somma di tutte le possibilità è uguale ad 1.
- Perciò dobbiamo sommare le probabilità di tutte le possibilità di O .
- Questa procedura di calcolo è molto lunga.



Un albero a tre rami



L'albero può essere radicato ovunque e le sostituzioni calcolate di conseguenza.

Aumentare la complessità dei modelli

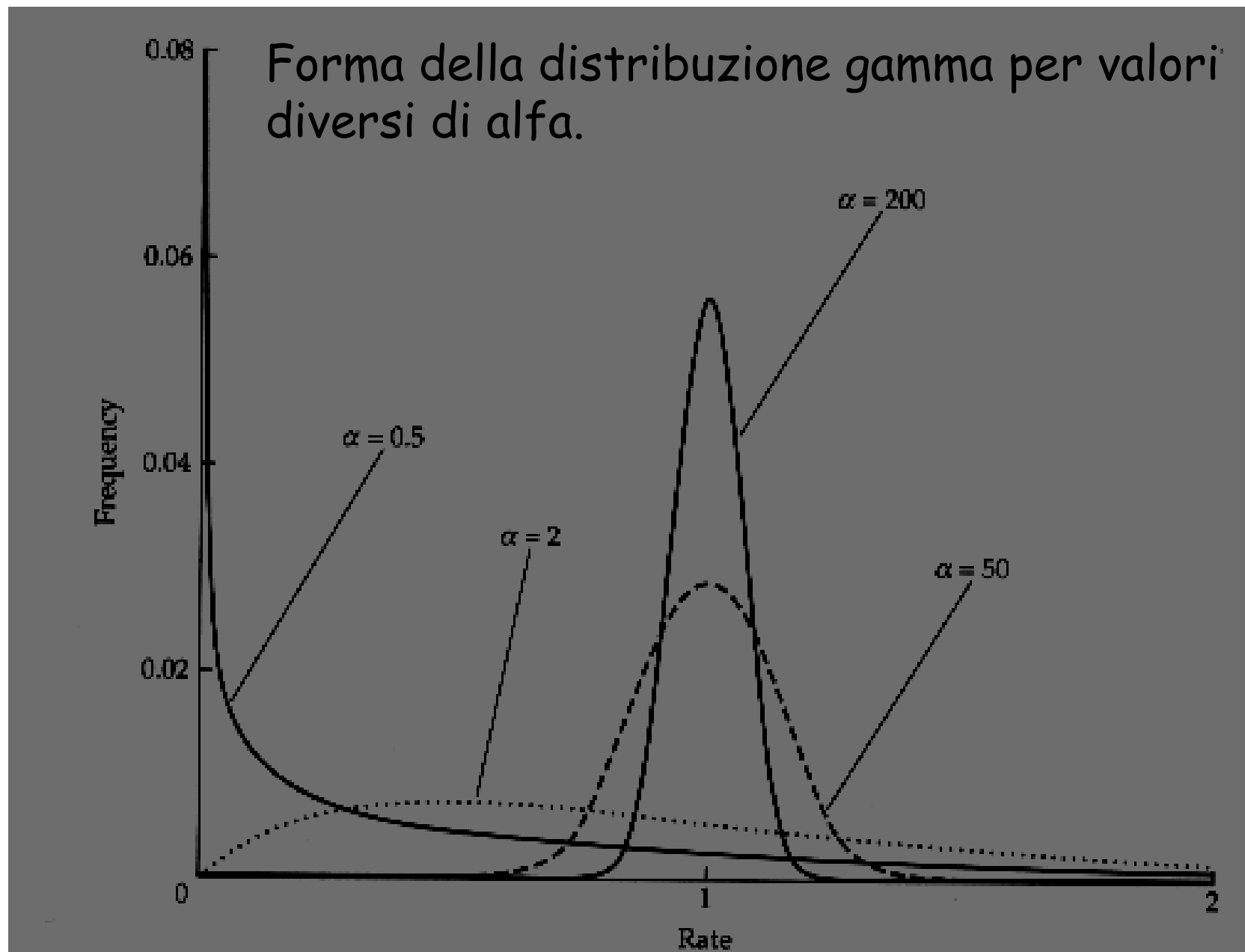
- Finora abbiamo assunto che i cambiamenti sono tutti egualmente probabili in tutte le posizioni e che il tasso di cambiamento è costante durante la filogenesi.
- Ciò non rappresenta un modello realistico per tutte le sequenze (è un modello neutrale con un orologio molecolare costante)

Siti non variabili

- Per un particolare set di dati possiamo assumere che una certa proporzione di siti non sono liberi di variare - la selezione purificante (correlata alla funzione) fa in modo che questi siti non possano cambiare.
- Si possono perciò osservare posizioni non variabili che sono tali o per effetto del caso, o perchè sotto pressioni selettive, o a causa di fenomeni quali l'omoplasia o le retromutazioni.
- La probabilità che un sito non sia variabile può essere calcolata considerando questa possibilità nel nostro modello e calcolando per ogni sito la probabilità che esso sia un sito non variabile.
- La likelihood del nostro set di dati può essere migliorata rimuovendo una certa percentuale di siti non variabili.
- Utilizzando PAUP è possibile stimare la proporzione di siti non varianti usando Maximum Likelihood.

Siti variabili

- Altri siti nel set di dati sono liberi di variare.
- L'intensità della selezione su questi siti non è uniforme, così è meglio utilizzare un modello che consideri un tasso di variazione eterogeneo.
- Ciò si può fare in due modi:
 - Considerare il tasso in dipendenza del sito (per esempio, posizione nel codone, o la struttura secondaria)
 - Usare una approssimazione discreta ad una distribuzione continua (distribuzione gamma)
- Queste variabili vengono modellate su tutte le possibilità di cambiamento nella sequenza, su tutte le possibilità delle lunghezze dei rami e su tutte le possibili topologie dell'albero.



Cambiando il modello il risultato cambia?

Ci sono diversi modelli

Jukes and Cantor (JC69):

Composizione in basi eguale (ognuna ha frequenza 0.25), tasso di cambiamento costante ed uguale

Kimura 2-Parametri (K2P):

Composizione in basi eguale (ognuna ha frequenza 0.25), tasso di sostituzione diverso per le transizioni e le trasversioni

Hasegawa-Kishino-Yano (HKY):

Come K2P, ma con composizione in basi variabile

General Time Reversible (GTR):

Composizione in basi variabile, tasso di sostituzione diverso per ogni tipo.

Tutti questi modelli possono considerare anche la proporzione di siti non variabili e il tasso di variazione sito per sito.

Vantaggi di ML

- Non c'è bisogno di alcuna correzione perchè i modelli tengono conto anche delle sostituzioni sovrainposte.
- Stima accuratamente le lunghezze dei rami.
- Ogni sito ha una likelihood.
- Se il modello è corretto anche l'albero è corretto *.
- Si può usare un modello che si adatti ai dati.
- ML usa tutti i caratteri, tutti sono informativi.
- ML ci indica la filogenesi delle sequenze e il processo evolutivo sottostante.

*Se le sequenze sono abbastanza lunghe e se viene utilizzato un modello abbastanza sofisticato.

Modelli

- Si possono usare modelli che:
 - Considerano rapporti T_i/T_v differenti.
 - Considerano una composizione in base ineguale.
 - Considerano tassi di sostituzione eterogenei.
 - Considerano l'eterogeneità del processo di sostituzione (tassi diversi lungo le linee e in differenti parti dell'albero).
- Più i parametri sono liberi, meglio il modello si adatterà ai dati. (vantaggio)
- Più i parametri sono liberi , maggiore sarà la varianza della stima (svantaggio).
- Usa un modello che si adatta ai dati.

Debolezze di ML

- Può dare risultati inconsistenti se il modello seguito non è accurato.
- Il modello potrebbe essere non abbastanza complesso.
- Il procedimento di calcolo è molto complicato. Potrebbe non essere possibile esaminare tutti i modelli.

Raccomandazioni

- Interagire con i dati.
- Se ci sono abbastanza dati, potrebbe essere possibile avere un buon quadro del modello sottostante l'evoluzione delle sequenze.
- Usare un test di modelli alternativi.
- Non scegliere un modello a caso, usare un modello che si adatti ai dati.

“How often have I said to you that when you have eliminated the impossible, whatever remains, *however improbable*, must be the truth.”

Sherlock Holmes to Dr. Watson in
The Sign of Four, by A. Conan Doyle.