

Docente: **Matteo Re**

UNIVERSITÀ DEGLI  
STUDI DI MILANO



C.d.I. Informatica

# Bioinformatica

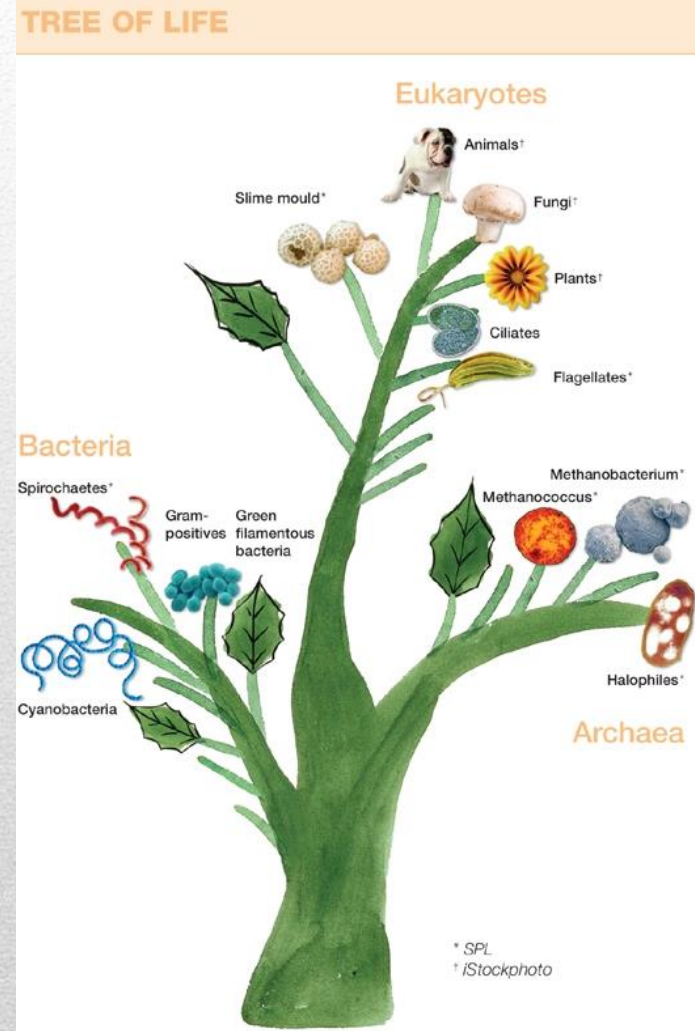
A.A. 2013-2014 semestre II

**4**

**Evoluzione e filogenesi**

---

- **Definizione**
  - Studio delle relazioni evolutive tra vari gruppi di organismi
- **La vita si è evoluta da un singolo organismo unicellulare**
  - Cenancestor
- **Tecniche tradizionali:**
  - Basate su differenze fenotipiche (caratteristiche osservabili, o “tratti”, degli organismi)

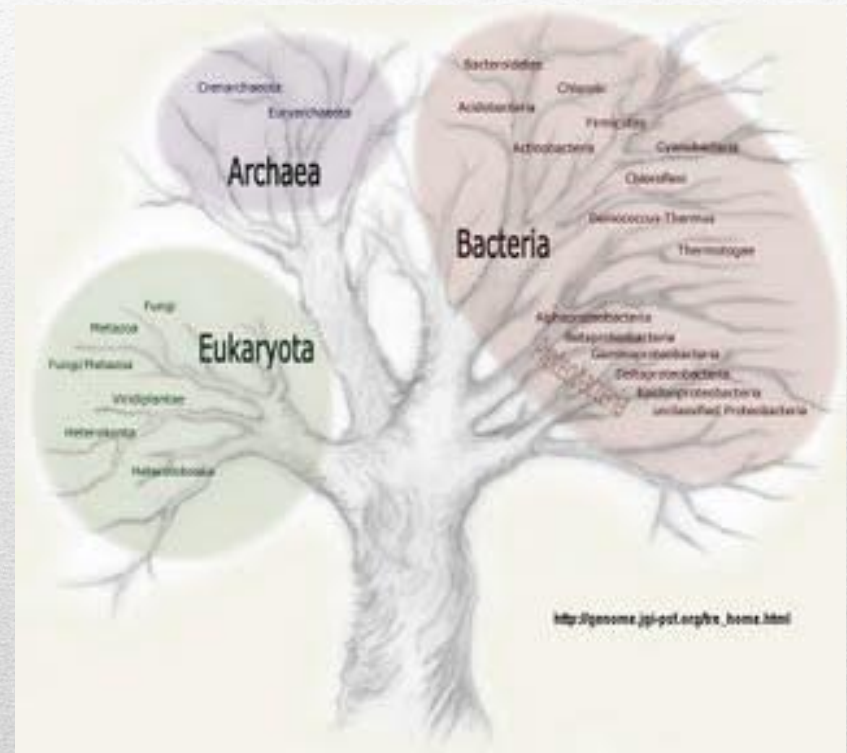


- **Comprendere l'origine dei viventi**
  - Chi siamo? Da dove veniamo? (in senso evolutivo)
- **Se riuscissimo a comprendere i sistemi biologici e la loro origine...**
  - Potremmo riuscire a predire
    - Reazioni a variazioni ambientali
    - Reazioni a farmaci (organismi "simili" probabilmente reagiranno in maniera simile)
    - E molto altro...
- **Cosa ci riserva il futuro**
  - Come evolveremo ( problema estremamente complesso)

# Perché è importante?

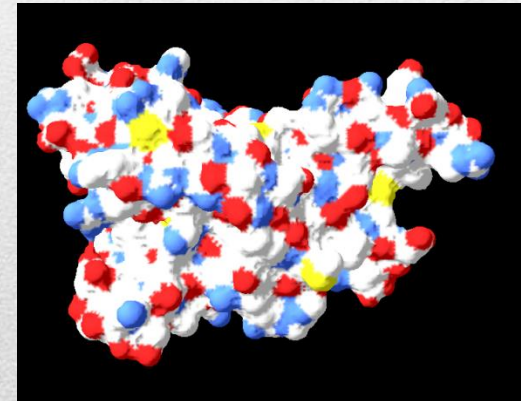
---

- DNA è “simile” in organismi **evolutive** correlati
- Come misuriamo la “similarità” del DNA?
  - Dobbiamo **allineare**
  - Dobbiamo utilizzare geni **omologhi\***
  - Conteggio delle posizioni in cui nt o aa sono differenti.



\* E' quindi richiesta **conoscenza a priori** durante la costruzione di una collezione di sequenze da analizzare.

- **Differenti** velocità evolutive (frequenza dei cambiamenti)
  - **Organismi:** fattori ambientali differenti
  - **Proteine:** pressioni selettive differenti
  - *Regioni* delle proteine:
    - Regioni interne, altamente compatte, idrofobiche
    - Loop esterni, meno importanti per l'integrità strutturale



Un allineamento è un'**ipotesi evolutiva**. Quando osserviamo un gap esso indica che, nel corso dell'evoluzione in una delle sequenze allineate si è verificata l'inserzione o la delezione di parte della sequenza.

**Dobbiamo** tener conto del fatto che:

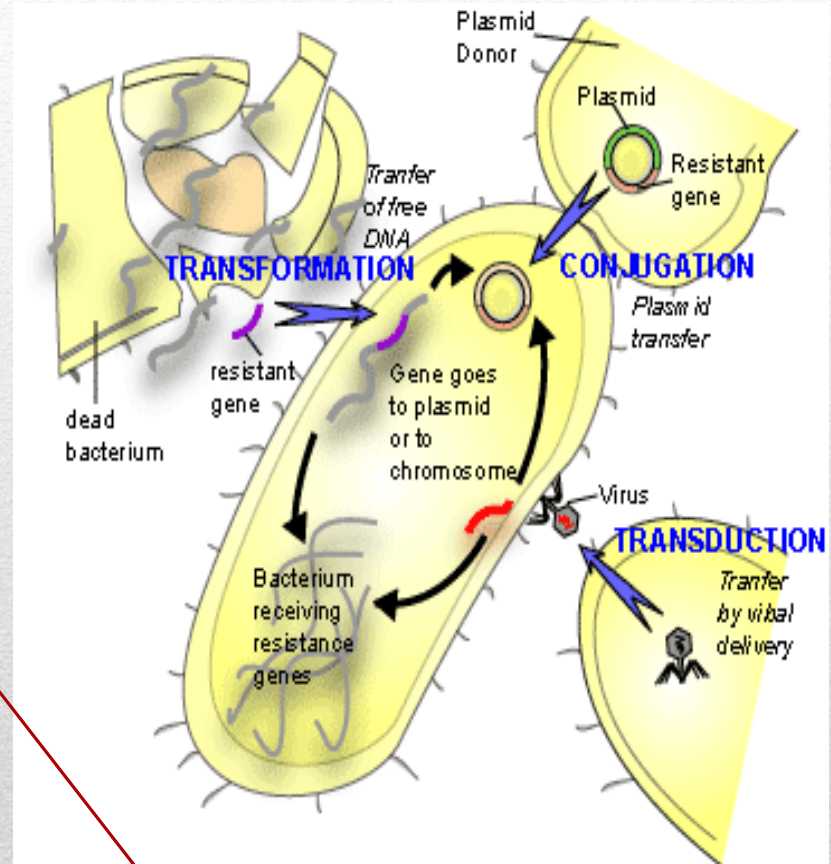
- Gap di ogni lunghezza possono avvenire in un **singolo** evento evolutivo
- Stiamo cercando di studiare l'evoluzione partendo da una serie di informazioni **PARZIALI** (non disponiamo delle sequenze di tutti gli organismi che si sono esistiti nel corso dell'evoluzione ma solo di alcune delle specie esistenti)

Buchnera/1-356	MENL-----DKKKALDRVIMEIEKAYGKGAIMKLG-EMA
Lactobacillus/1-363	MAKD-----EKKAALDAALKKIEKNFGKGAVMRMG-EKA
Geobacter/1-338	MTQ-----EREKAIELALSQIEKQFGKGAIMRLGADEA
Actinobacillus/1-376	MAADNKKAQKNTVTKQIDPEQKEKALAAALAQIEKQFGKGSIMRLG-DTQ
Salmonella/1-353	MAID-----ENKQKALAAALGQIEKQFGKGSIMRLG-EDR

# gaps

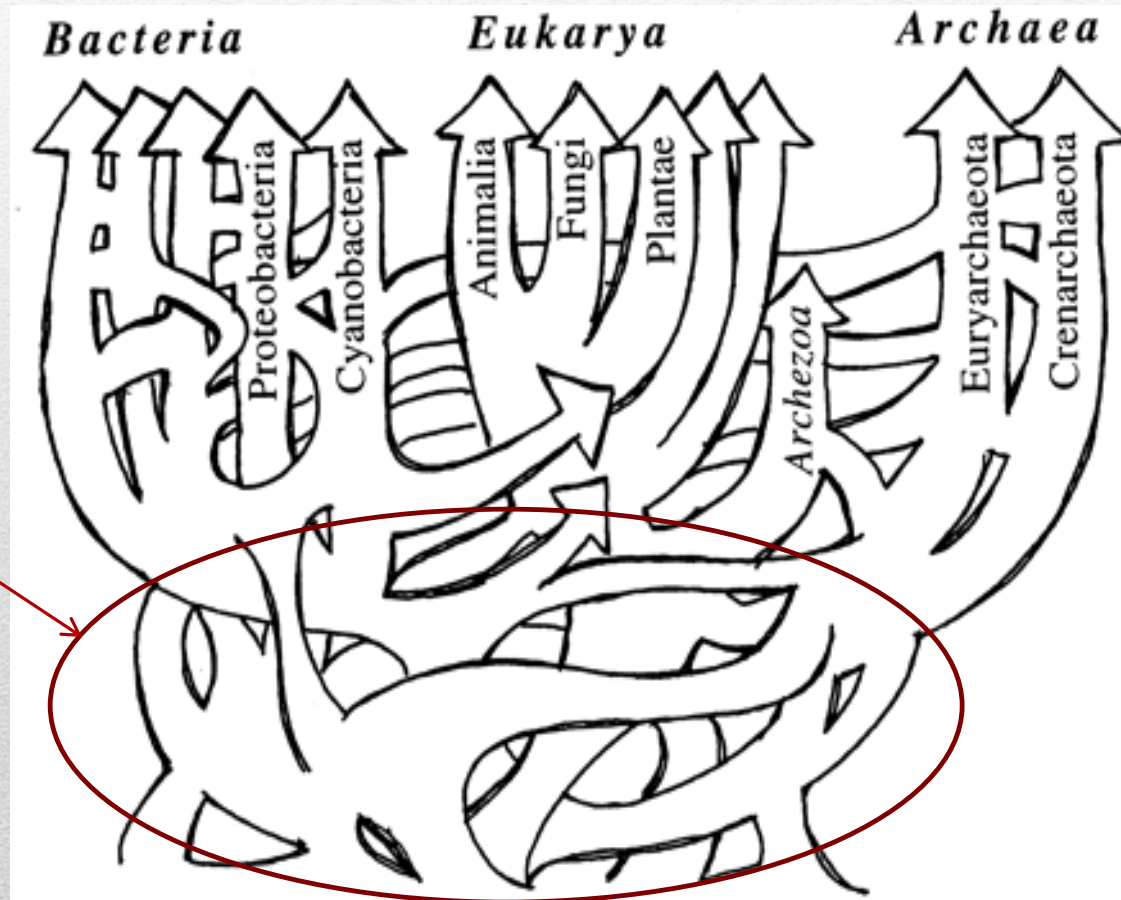
---

- DNA può **muoversi** da un organismo all'altro
- La riproduzione nei batteri è asessuale ma DNA può spostarsi per mezzo di :
  - plasmidi
  - virus
  - assunzione diretta
- Meccanismi "meno" sorprendenti...
  - Meiosi, mitosi, traslocazione



**trasferimento orizzontale**

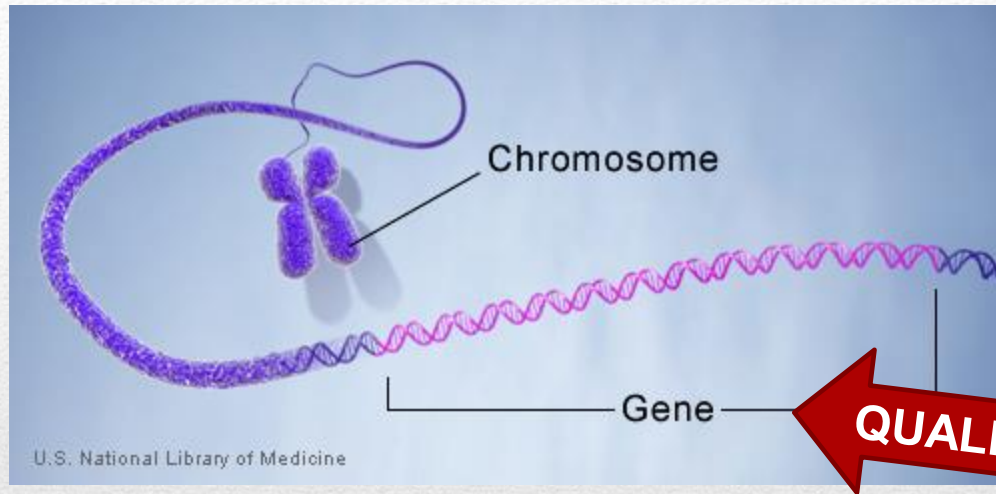
# Albero "reticolato"



Specialmente  
vicino alla  
radice



## SOLUZIONE : Scegliere il gene "GIUSTO"



- Serve un gene che si trovi in **tutti** gli organismi (ubiquitario)
  - Il gene dovrebbe essere **evolutiveamente** “**stabile**” (alta similarità in tutti gli organismi)
  - Dovremmo basare i confronti su regioni del gene che sono **altamente conservate**.
-

- DNA circolare localizzato in organelli (al di fuori del nucleo)
- Niente **crossing-over**: ereditato dalla cellula uovo
- **Copia esatta** ereditata dalla Madre
- I **mitocondri** sono le "centrali energetiche" della cellula
  - Elaborazione nutrienti, processamento e **rilascio energia**

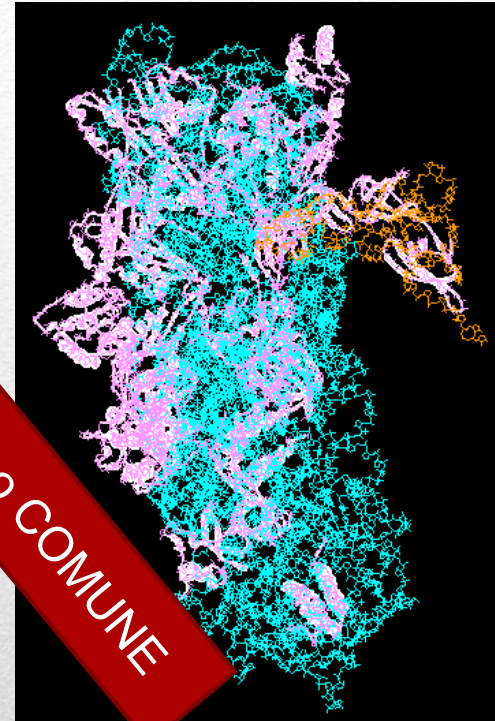


Processi COMUNI

# DNA mitocondriale

- Componente principale ribosomi procarioti (processo: **traduzione**)
- **Ubiquitario**, stesso ruolo in ogni organismo
- **Altamente** conservato

Processo COMUNE



# RNA ribosomale (16S)

---

Bio

CS

Ora abbiamo una COLLEZIONE di sequenze!  
COME POSSIAMO ALLINEARLE?

Strumenti per l'allineamento  
(lezioni precedenti)

- **Metodi di programmazione dinamica**
  - Needleman-Wunsch (allineamento globale)
  - Smith-Waterman (allineamento locale)
- **BLAST** (euristica) ←

**Veloce (lineare)**  
**...ma non molto sensibile!**  
**possibili soluzioni ...**

Alcune classi  
di complessità  
algoritmica

**Fissa:** la migliore  
**Lineare:** seconda migliore  
**Polinomiale** ( $n^2$ ): non male  
**Esponenziale** ( $3^n$ ): pessima

Confronto seq. proteiche  
Matrici di scoring specializzate  
...

Bio

CS

Ora abbiamo una COLLEZIONE di sequenze!  
COME POSSIAMO ALLINEARLE?

Strumenti per l'allineamento  
(lezioni precedenti)

- **BLAST** (euristica): veloce ma non molto sensibile... questo è un **grosso problema** dato che vogliamo confrontare sequenze che, evolutivamente, possono essere anche molto distanti!
  - L'ideale sarebbe utilizzare strumenti che garantiscono un **allineamento ottimo** (NW o SW), ma sono troppo costosi in termini di tempo!
-

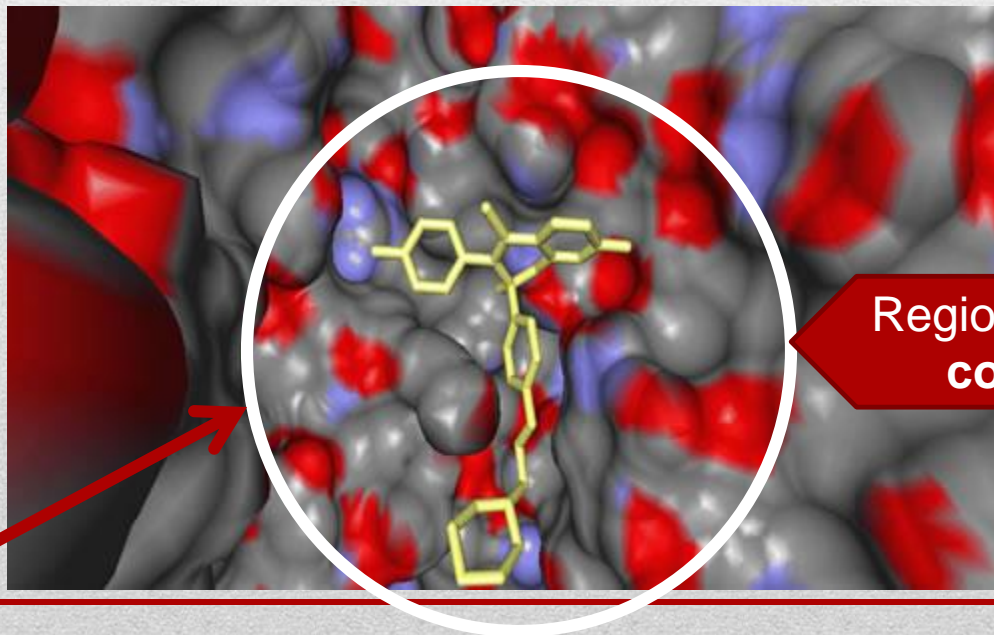
- Alcune parti delle proteine sono **estremamente importanti** per mantenere la funzione molecolare
- L'assunzione **biologica** è che queste parti debbano essere **simili** nelle sequenze provenienti da specie differenti
- **OBIETTIVO**: **evidenziare** queste regioni mediante un processo di allineamento.

↓ ↓ ↓ ↓ ↓ ↓ ↓

atg**cg**ca-actg**ccgcaggagatcaggactttc**atgaatat**catcatg**cg**tg**ggga-ttc**ag**  
acct**cg**atac**gtgccgcaggagatcaggactttc**acct--tgg**atcatg**cg**accg**tac**ctac**

---

- Spesso le regioni conservate sono vicine (o corrispondono a) **siti attivi** (qui “attivi” è utilizzato in maniera generica)
  - Riconoscimento di **ligandi, substrati** ecc.
  - Interfaccia di contatto **tra** proteine
  - Regioni importanti per la struttura **terziaria**



Molto utile per ipotizzare una funzione o per **riconoscere** proteine **funzionalmente correlate**

- La conservazione evolutiva emerge con più chiarezza durante il confronto di più sequenze.
- **Maggior confidenza** rispetto alla conservazione rilevata confrontando coppie di sequenze

atgcgca-actgccgcaggagatcaggactttcatgaatatcatcatgcttggga-ttcag  
 acctcgataacgtgccgcaggagatcaggactttcacct--tggatcatgcgaccgtacctac

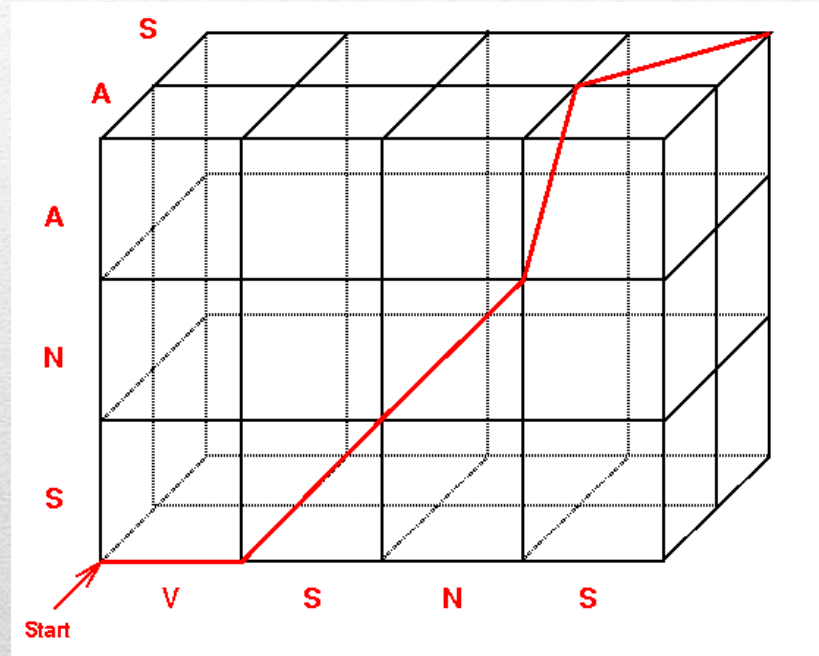
↓   ↓   ↓   ↓   ↓

atgccgca-actgccgcaggagatcaggactttcatgaatatcatcatgcttggga-ttcag  
 acctcatacgtgcccaggagatctggactttcacc---tggatcatgcgaccgtacctac  
 t-atgg-t-cgtgccgcaggagatcaggactttca-gt--g-aatcattgg-cgc--c-a  
 t--tcgt-ac-tgccccaggagatctggactttcaaa---ca-atcatgcgcc-g-tc-tat  
 aattcgtacgtgccgcaggagatcaggactttcag-t--a-tatcattgtc-ggc--tag

---



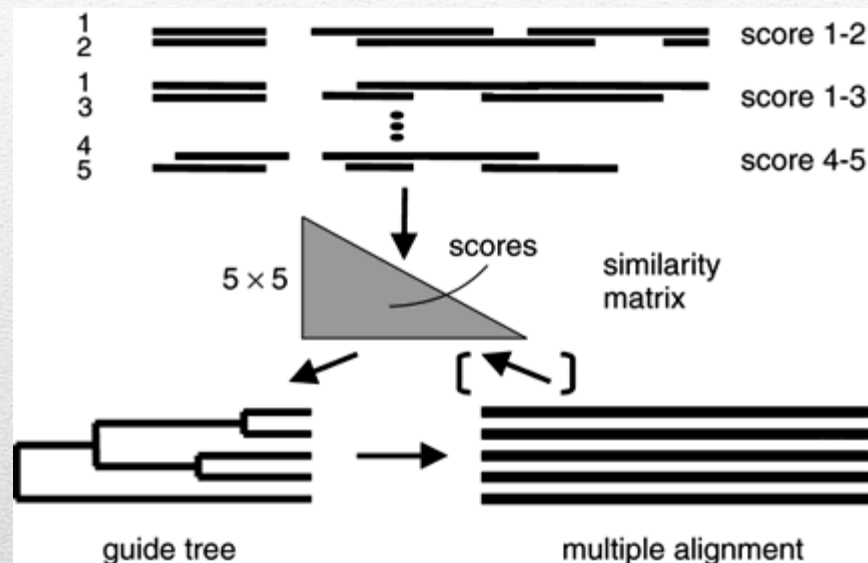
- Programmazione dinamica **iperdimensionale** (una dimensione per ogni sequenza)
- **Complessità : esponenziale** rispetto al numero di sequenze!!!
- $O(n^L)$  con  $L =$  numero di sequenze



**NON APPLICABILE!**

## ALLINEAMENTO PROGRESSIVO:

- Calcolo di tutte le **distanze** pairwise
  - Modo veloce: numero di match tra k-meri
  - Modo lento: allineamento globale
- Parto dalla coppia di sequenze + **simili**, e allineo
- Poi **allineo alla coppia** la sequenza più simile **tra le rimanenti**
- Continuo **fino a quando** non restano più sequenze da allineare



**ClustalW :**  
**cluster-alignment**

### Allineamento progressivo basato su PROFILI:

- **Profilo**: matrice (una riga per ogni simbolo, una colonna per ogni posizione nell'allineamento) di valori reali ognuno associato alla probabilità di un dato simbolo in ogni posizione dell'allineamento multiplo di sequenze
- Versione modificata dell'algoritmo Smith/Waterman
  - “Grado di match” tra aa di una sequenza e profilo è dato dalla probabilità dell' aa nel **profilo** del multiallineamento

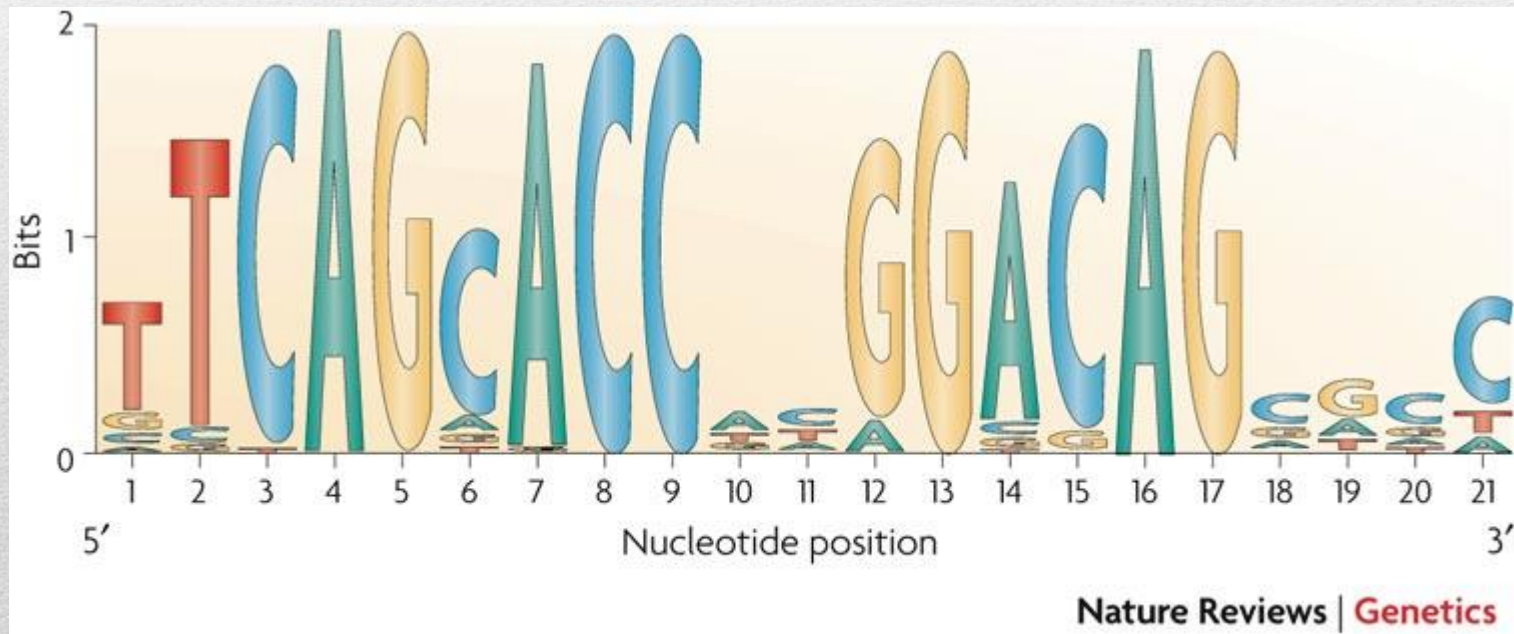
```

Consensus      1 M.ERS.HLPEG.PFAAALSGARFAAQSSGN.ASVL..DWNVLP.E 38
                | : : : || : ::::: : | : | ::| : : | :
OPSD_XENLA     1 MNG.GTE..EGPN.NFYVP.PMS...SN.NKTGVVRS.P..PFD 33
  
```

**PROBLEMA:** come possiamo allineare UNA sequenza ad un SET di sequenze precedentemente allineate?

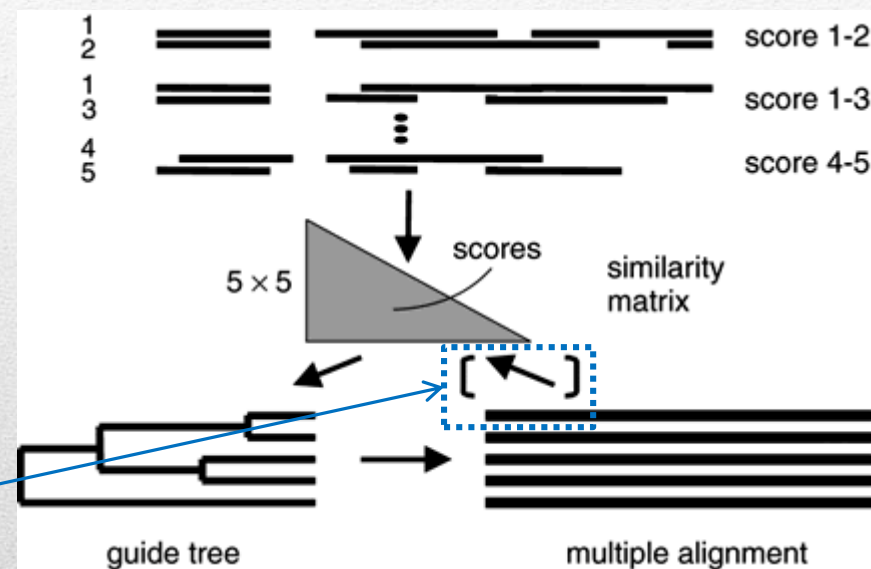
Visualizzazione di profili mediante LOGO:

- **LOGO:** l'altezza di una lettera è rappresentativa della **frequenza** del simbolo in una data posizione:

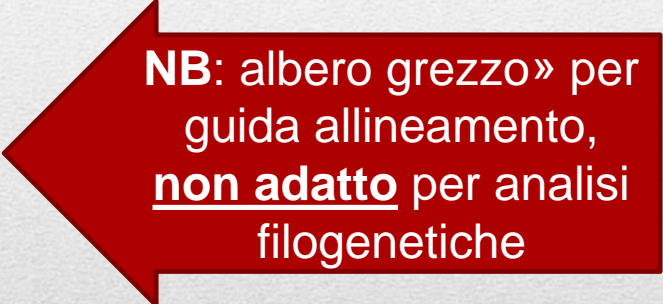


PROBLEMI DELL'ALLINEAMENTO PROGRESSIVO:

- Questo approccio è **PROGRESSIVO** ... **errori** di allineamento verificatisi nelle prime fasi vengono **propagati** in tutti i passi successivi del processo.
- Una volta che abbiamo allineato due sequenze queste **non** vengono più modificate (assenza raffinamento)
- Versioni più recenti del metodo allineano in modo "**iterativo**" (una volta ottenuto il profilo dell'intero allineamento ripartoo utilizzando questo profilo "più informativo")
- Versione più recente di ClustalW (version 2) include iterazione

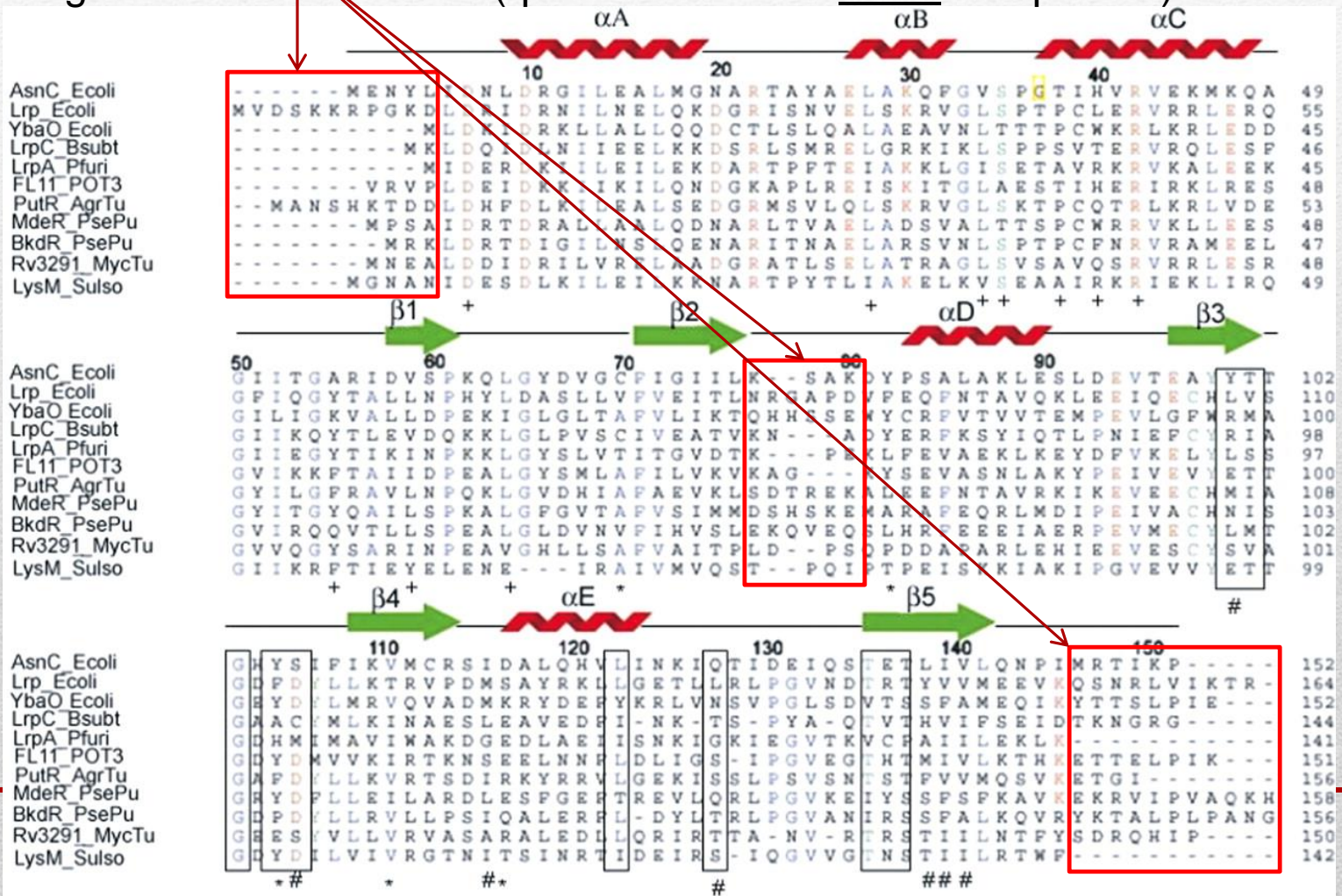


- Costruzione di una **matrice delle distanze** di tutte le  $N(N-1)/2$  coppie di sequenze utilizzando un metodo di allineamento basato su programmazione dinamica seguita da conversione (approssimata) degli score di similarità in distanze evolutive.
- Costruzione di un “**albero guida**”
- Allineare progressivamente partendo dai nodi più simili e procedendo verso il nodo a similarità minima. **NB:** un nodo può rappresentare allineamento tra, sequenza e sequenza, sequenza e profilo, profilo e profilo.



**NB:** albero grezzo» per guida allineamento, non adatto per analisi filogenetiche

Molto spesso un allineamento multiplo prodotto in modo **automatico** viene **rifinito manualmente** prima di procedere ad ulteriori analisi filogenetiche. (questo è un caso *molto* semplice...)



Caratteristiche peculiari di un allineamento multiplo:

- Conservazione **varia tra colonne** (position-specific scores)
- Le sequenze **non sono indipendenti** ( le relazioni tra di esse sono espresse da un albero filogenetico ... ma esso non è noto a priori).

### **Ipotesi di soluzione:**

Creare una rappresentazione probabilistica che modelli l'evoluzione. Il modello sarebbe in grado di descrivere ogni sequenza osservata in termini di variazioni tra sequenze ed ogni sequenza sarebbe generata tenendo conto delle velocità evolutive lungo i vari rami dell'albero.

**Soluzione NON PRATICABILE: non abbiamo dati a sufficienza per creare un modello probabilistico così complesso!**

Inoltre questo modello **richiede** la conoscenza del **vero** albero filogenetico ... mentre noi stiamo cercando di stimare una buona approssimazione dello stesso!



Per risolvere il problema dobbiamo fare alcune assunzioni. In particolare assumiamo che le colonne di un allineamento **siano indipendenti** (anche se non è vero) ed ignoriamo l'albero filogenetico!

$$S(m) = G + \sum_i S(m_i)$$

score  
multiallineamento  
(composto da i colonne)

gaps

score  
i-esima colonna

Somma di score tra tutte le coppie di simboli confrontati (Sum of Pairs o SP score) ... **causa problemi!**

$$S(m_i) = \sum_{k < l} sim(m_i^k, m_i^l)$$

score similarità ottenuti mediante matrici **PAM** o **BLOSUM**

## Ora abbiamo gli strumenti necessari

- Allineamenti multipli
- ClustalW (allin. progr. basato su profili). Risultato eventualmente rifinito **manualmente**.

E' facile identificare regioni **altamente conservate**



atg**ccg**ca-act**gccc**gaggagat**ca**ggacttt**ca**tgaatat**catcat**g**cg**tggga-ttcag  
acct**cc**atac**g**t**gccc**caggagat**ct**ggacttt**ca**acc---tggat**cat**g**cg**accgtacctac  
t-at**gg**-t-c**g**t**gccc**gaggagat**ca**ggacttt**ca**-gt--g-aat**cat**ct**g**g-c**g**c--c-a  
t--t**cg**t-ac-t**gccc**caggagat**ct**ggacttt**ca**aa---ca-at**cat**g**cg**cc-g-tc-tat  
aatt**cc**gtac**g**t**gccc**gaggagat**ca**ggacttt**ca**g-t--a-tat**cat**ct**g**tc-ggc--tag

---

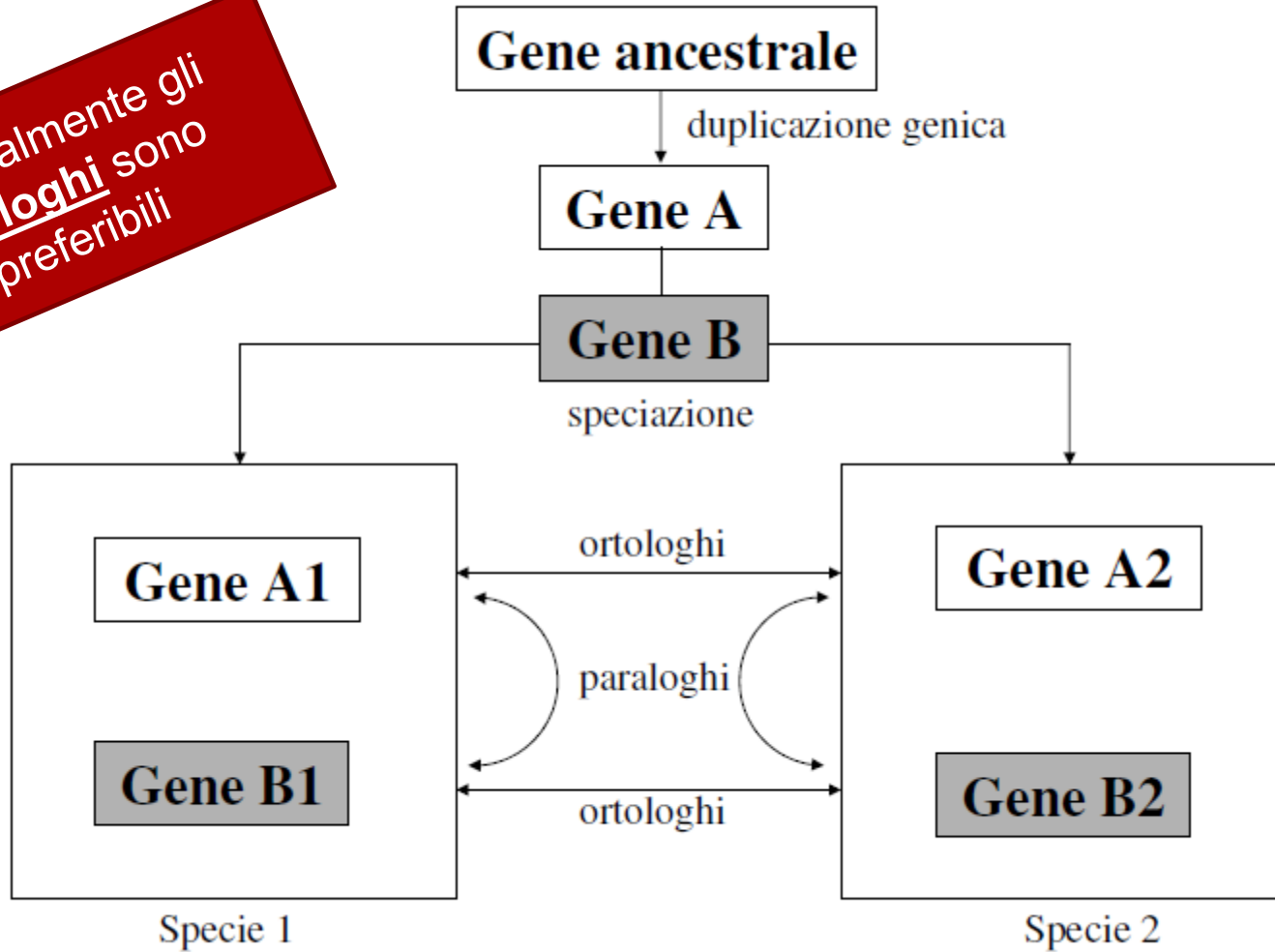
**Geni ortologhi**: geni simili riscontrabili in organismi correlati tra loro. Il fenomeno della speciazione porta alla divergenza dei geni e quindi delle proteine che essi codificano.

es. l'  $\alpha$ -globina di uomo e di topo hanno iniziato a divergere circa 80 milioni di anni fa, quando avvenne la divisione che dette vita ai primati e ai roditori. I due geni sono da considerarsi **ortologhi**.

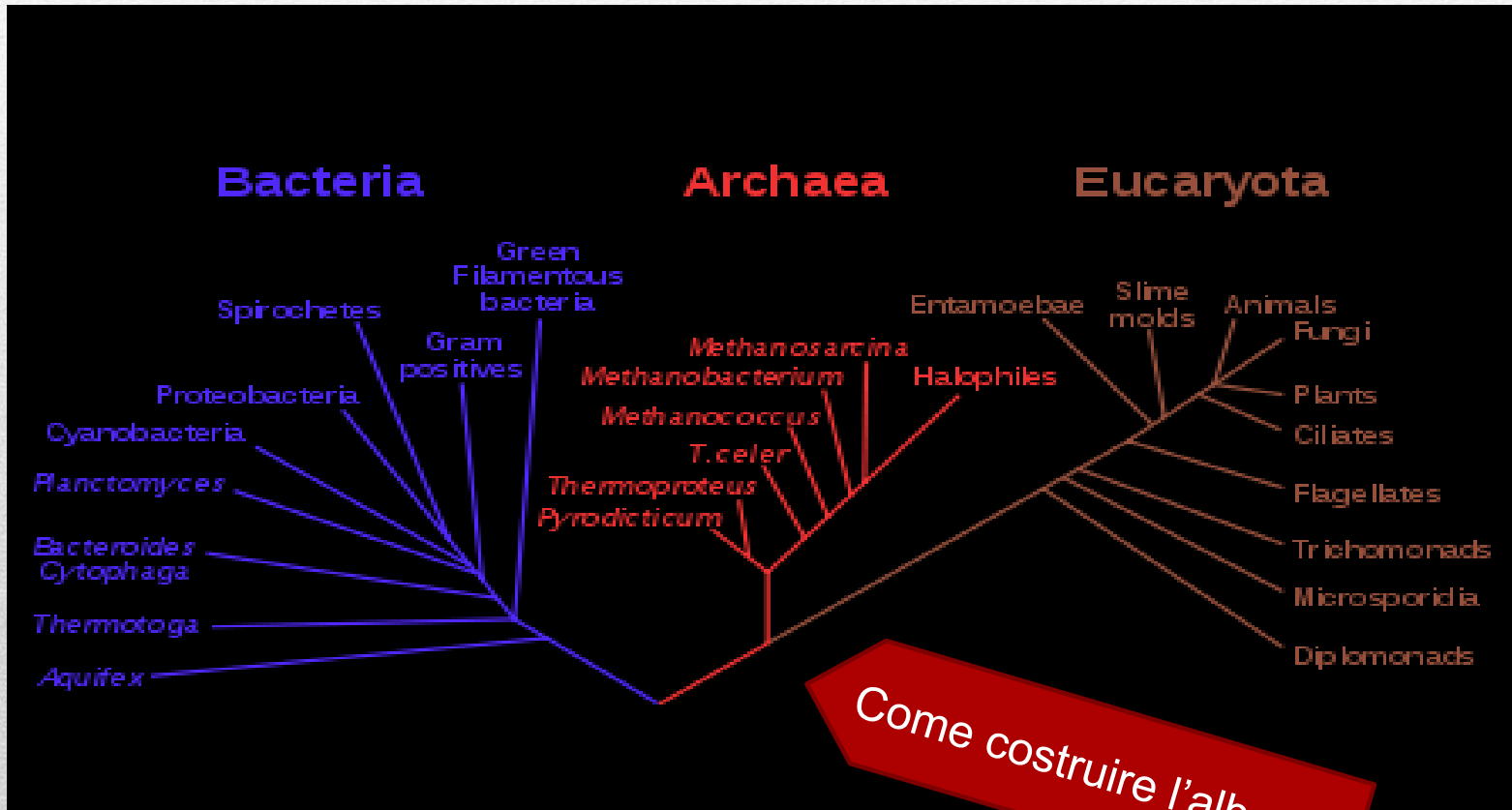
**Geni paraloghi**: geni originati dalla duplicazione di un unico gene nello stesso organismo. es.  $\alpha$ -globina e  $\beta$ -globina umana hanno iniziato a divergere in seguito alla duplicazione di un gene globinico ancestrale. I due geni sono da considerarsi **paraloghi**.

---

Generalmente gli  
ortologi sono  
preferibili

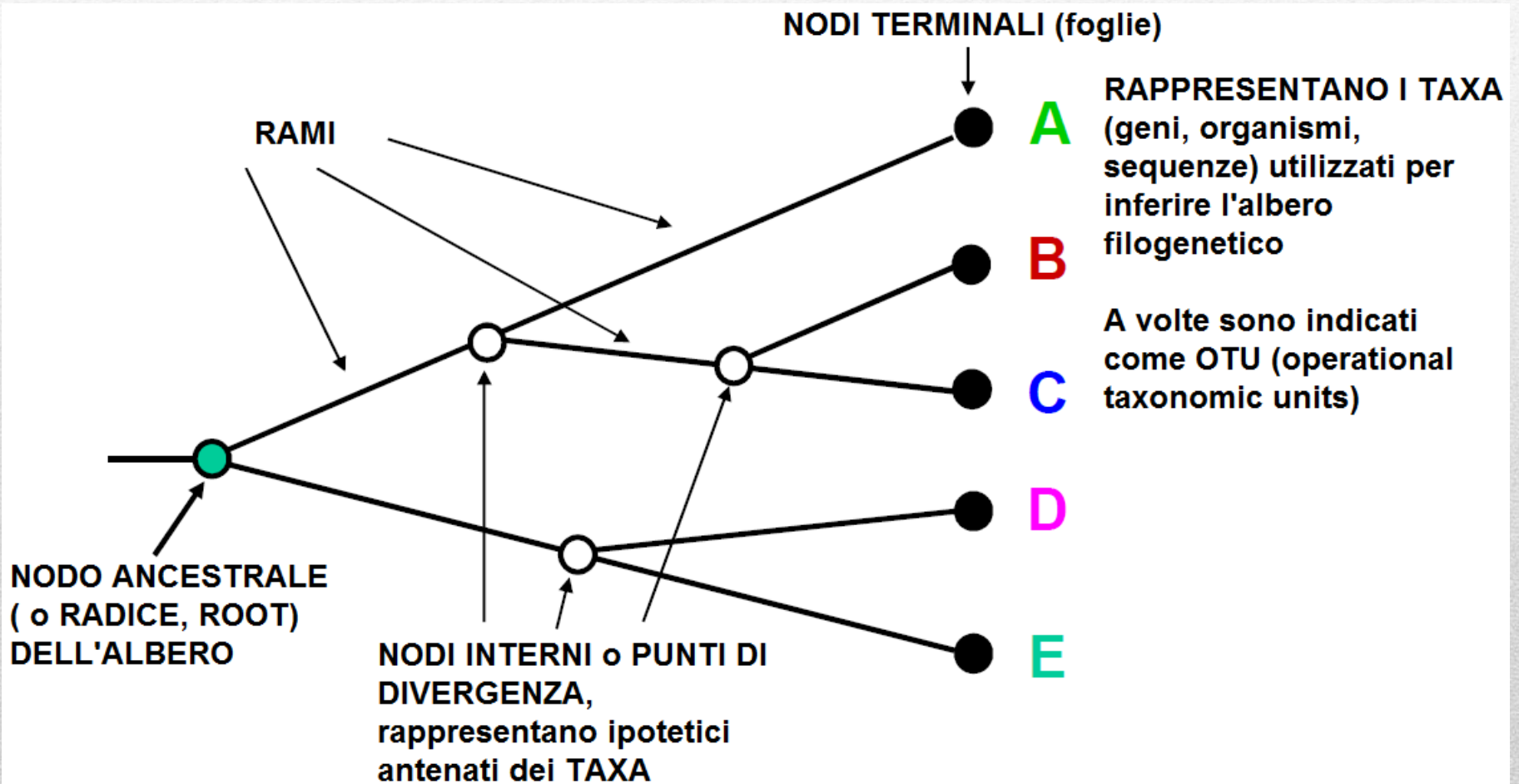


- Utilizzo di **16S rRNA** per indagini sull'albero della vita
- Identificati **tre** domini (non due)



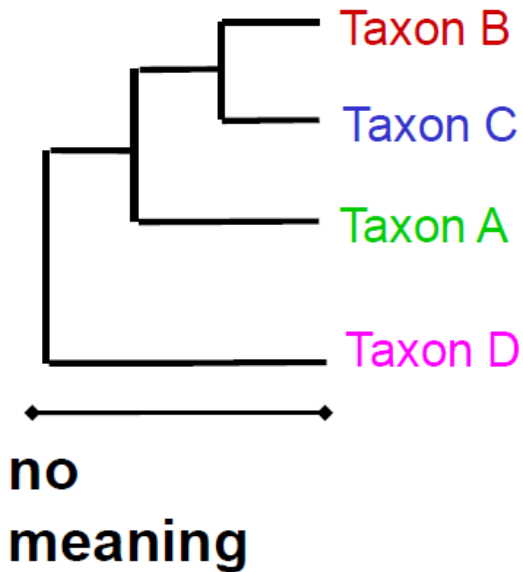
**Woese *et al.* 1987**

## Terminologia:

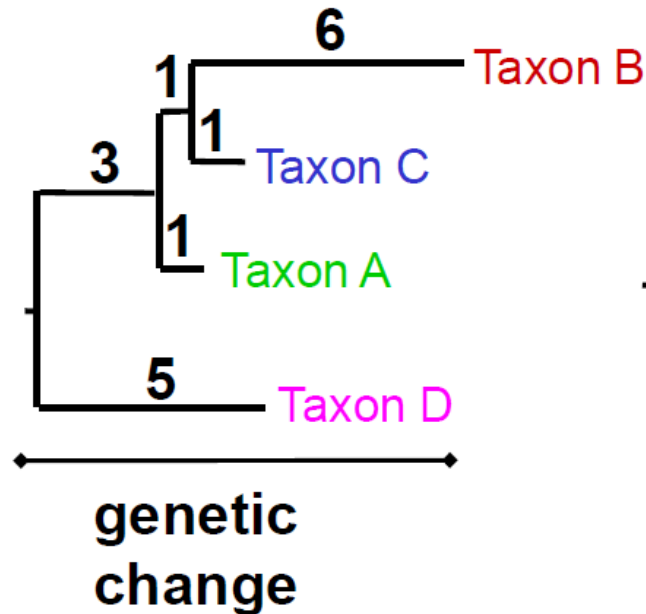


## Tipi di albero filogenetico (I):

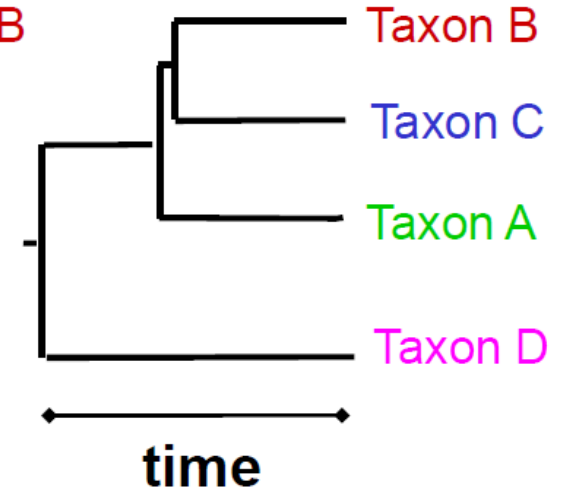
Cladogram  
(topology only matters)



Phylogram  
(topology and individual lengths)



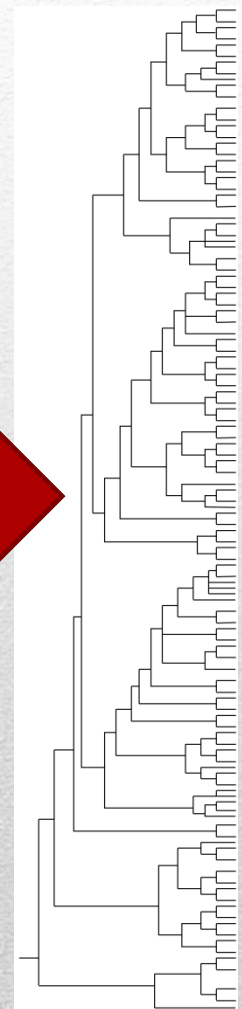
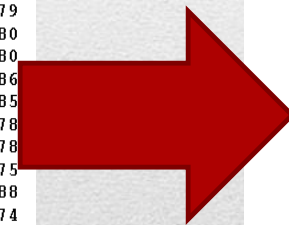
Chronogram (ultrametric)  
(topo & divergence times)



**NB:** tutti mostrano la stessa topologia

# Ruolo dei metodi filogenetici :

Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLGMKNTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLGMKNTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLGMKNTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLGMKNTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKOMQIIRMSLRGK-AVVLGMKNTMMRKAIRGHLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLGMKNTMMRKAIRGHLENN--PALE	76
Q7ZUG3_BRARE	-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLGMKNTMMRKAIRGHLENN--PALE	76
RLA0 ICTPU	-----MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKOMQIIRMSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_DROME	-----MVRENKAAWKAQYFIVKVELDFEFPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLGMKNTMMRKAIRGHLENN--PALE	76
RLA0_DICDI	-----MSGAG-SKRKKLFIEKATKLFYTDKMIIVAEADFGVSSOLQKIRKSIRGI-GAVLMGKNTMIRKIVIRDLADSK--PELD	75
Q54LP0_DICDI	-----MSGAG-SKRKNVFIEKATKLFYTDKMIIVAEADFGVSSOLQKIRKSIRGI-GAVLMGKNTMIRKIVIRDLADSK--PELD	75
RLA0_PLAF8	-----MAKLSKQKKQMYIEKISSLIQQYSKLIVHVDNVGSKOMQIIRMSLRGK-ATILMGKNTIRIRTALKKNLQAV--PQIE	76
RLA0_SULAC	-----MIGLAVTTTKIAKWKVDEVAELTEKLRKHTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFPNIALKNAG----YDTE	79
RLA0_SULTO	-----MRIMAVITQERKIAKWKIEEVEKLECKLREYHTIIIANIEGFPADKLHDIRKKMRGK-AEIKVTKNTLFGIAAKNAG----LDVS	80
RLA0_SULSO	-----MKRLALALQQRKVASWKEEVEKLETELKNSNTLIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFGIAAKNAG----IDIE	80
RLA0_AERPE	MSVVSIVGQMYKREKPIPEWKTMLMLRELEELFKSRHVVLFADLTGTPFVVRVQVKKLWKK-YPMVAKKRIILKAMKAAGLE--LDDN	86
RLA0_PYRAE	-----MMLAIGKRRYVRTQYPAKVKVIYSEATELLQYYPYVFLFDLHGLSRIILHEYRRLRRY-GVIKIIKPLFKIAFTKYVGG--IPAE	85
RLA0_METAC	-----MAEERHHTEHIPQWKDEIENIKELIQSHKVFQMGVIEGILATKIKKIRRDLDV-AVLKVSRTLTALRNQLG----ETIP	78
RLA0_METMA	-----MAEERHHTEHIPQWKDEIENIKELIQSHKVFQMGVIEGILATKIKKIRRDLDV-AVLKVSRTLTALRNQLG----ESIP	78
RLA0_ARCFU	-----MAAVRGS--PPEYKVRAVEEIKRMISKPVVAIVSFRNVPAGOMQIRREFRGK-AEIKVVKNTLLEALDALG----GDYL	75
RLA0_METKA	MAVKAAGQPPSGEYEPKVAEWRREVEKLEKLMDEYENGLVDLEGIPAPOLQEIRAKLRERTTIIRMSRNTLMRALDEEKLDER--PELE	88
RLA0_METTH	-----MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLLISLALEKAGREL--ENVYD	74
RLA0_METTL	-----MITAESEHKIAPWKIEEVENKLEKLLKNGQIIVALVDMMEVPAVQLQEIRDKIR-CTMTLKMRSNTLLEALKEVAEETGNPEFA	82
RLA0_METVA	-----MIDAKSEHKIAPWKIEEVENKLEKLLKSNVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMRSNTLLEALKEVAEETGNPEFA	82
RLA0_METJA	-----METKVAHVAPWKIEEVENKLEKLLKSPVVAIVDMMVDPAPOLQEIRDKIR-DKVKLRMSRNTLLEALKEVAEELNPNKLA	81
RLA0_PYRAE	-----MAHVAEWKKKEVEELANLIKSYVVALVDVSSMPAYPLSQMRRLLRENGGLLRVSRNTLLELAIKKAAGELGKPELE	77
RLA0_PYRHO	-----MAHVAEWKKKEVEELAKLIKSYVVALVDVSSMPAYPLSQMRRLLRENGGLLRVSRNTLLELAIKKAAGELGKPELE	77
RLA0_PYRFU	-----MAHVAEWKKKEVEELANLIKSYVVALVDVSSMPAYPLSQMRRLLRENGGLLRVSRNTLLELAIKKAAGELGKPELE	77
RLA0_PYRKO	-----MAHVAEWKKKEVEELANLIKSYVVALVDVAGVPAYPLSKMRDCLR-GKALLRVSRNTLLELAIKKAAAGELGQPELE	76
RLA0_HALMA	-----MSAESEKRTETIPWQKEEVDVAIVEIMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLEALDDVD--DGLE	79
RLA0_HALVO	-----MSESEYRQTEVIPQWKREVEDELVDVIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRRALDEVN--DGFE	79
RLA0_HALSA	-----MSAEQRTTEEVPEWKRQVEVAELVDLLETYDSVGVVNVGTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEAG----DGLD	79
RLA0_THEAC	-----MKEVSQQKKEIVNEITRRIKASRSVAIVDTAGIRIROTQDIRGKNRGK-INLKVTKNTLFLKALENLGD--EKLS	72
RLA0_THEVO	-----MRKINPKKKEIVSELAODITKSKAVAIVDIKGVRIROMQDIRAKNRDK-VKIKVYKKTLLFKALDSIND--EKLT	72
RLA0_PICTO	-----MTEPAQWKIDFVKNLENEINSRKKVAIVSIRGLRNNFQKIRNSIRDK-ARIKVSRRARLLRLAIENTGK--NNIV	72
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90	



«caratteri» (molecolari) → DISTANZE → ALBERO FILOGENETICO

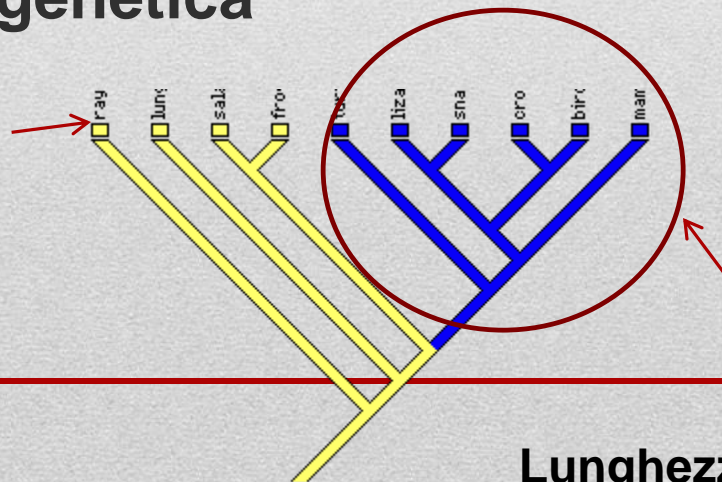


## Obiettivo:

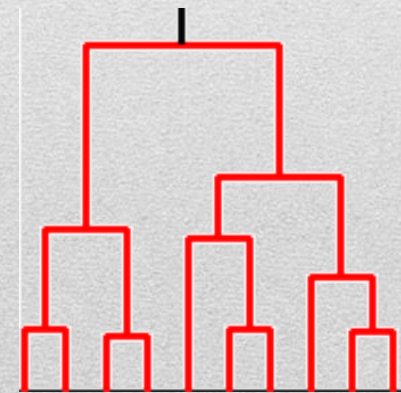
Costruzione di un cladogramma o di un filogramma

- **Lunghezza** di ogni ramo rappresenta il **numero** di cambiamenti osservati tra le sequenze (eccezione: in cladogramma lunghezza rami non ha significato)
- Vicinanza **topologica** rappresenta vicinanza **filogenetica**

Ogni sequenza è un TAXA



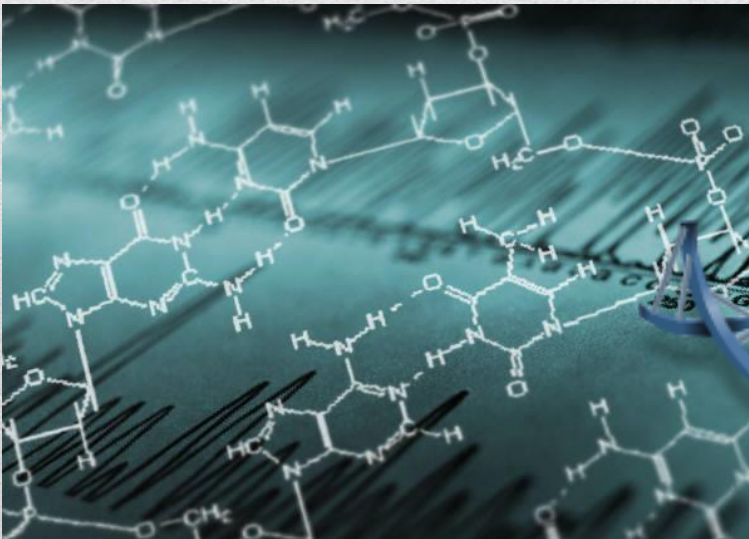
Ogni sottoalbero è un CLADE



Lunghezza albero = **SOMMA**(lunghezze rami)

## L'ipotesi dell'orologio molecolare

- Assunzione di velocità di mutazione uniforme per tutti i rami dell'albero
- E' ragionevole?
- Permette di testare in maniera semplice ipotesi che, altrimenti, richiederebbero test estremamente complessi



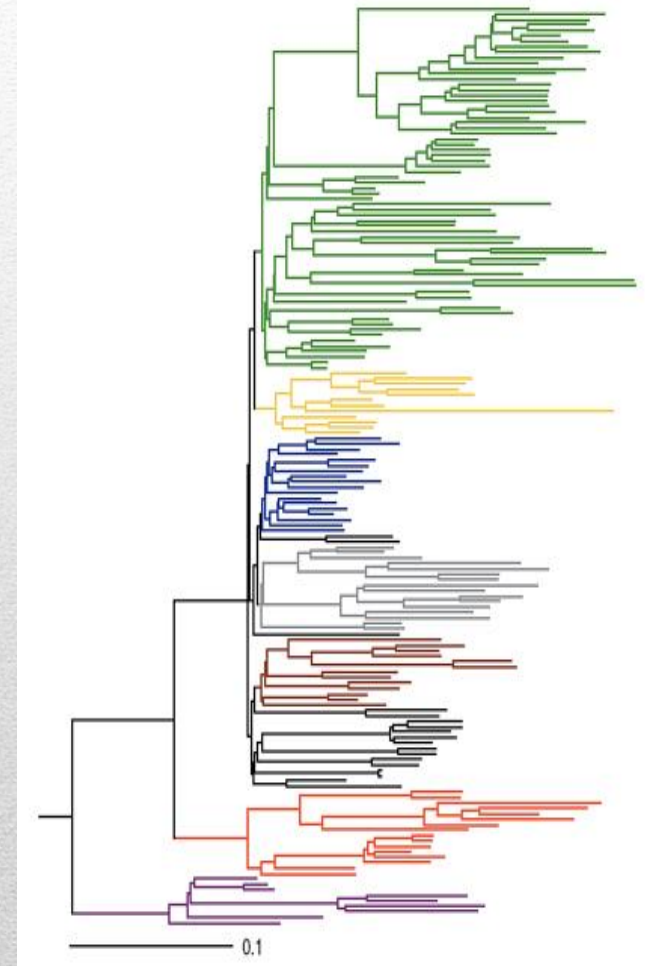
## Metodi basati su:

- **Distanza**
- **Massima parsimonia** (minima evoluzione)
- **Massima verosimiglianza**

Strumenti disponibili :

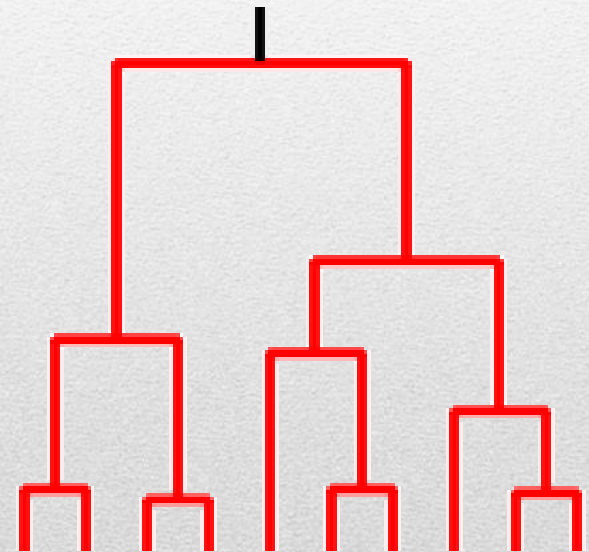
PAUP

PHYLIP



## Metodi basati su distanze

- Unweighted pair group method with arithmetic mean (**UPGMA**)
  - Uno dei primi (e più semplici) metodi basati su distanze
- Dal punto di vista informatico è un problema di **clustering gerarchico**



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

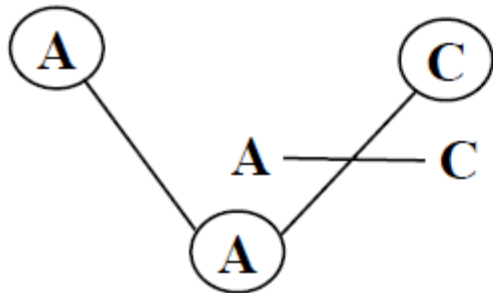
Quando confrontiamo siti omologhi in 2 sequenze di DNA osserviamo semplicemente se le sequenze sono le stesse o no.



Il numero **massimo** di differenze per sito che possiamo osservare è uno. Ciò significa che se più di una sostituzione è avvenuta ad un sito **perdiamo** l'informazione della precedente sostituzione

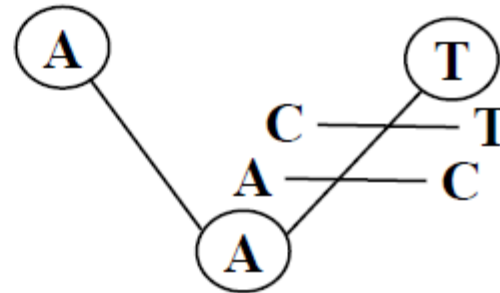
---

sost. singola



1 mutazione,  
1 differenza

sost. multipla



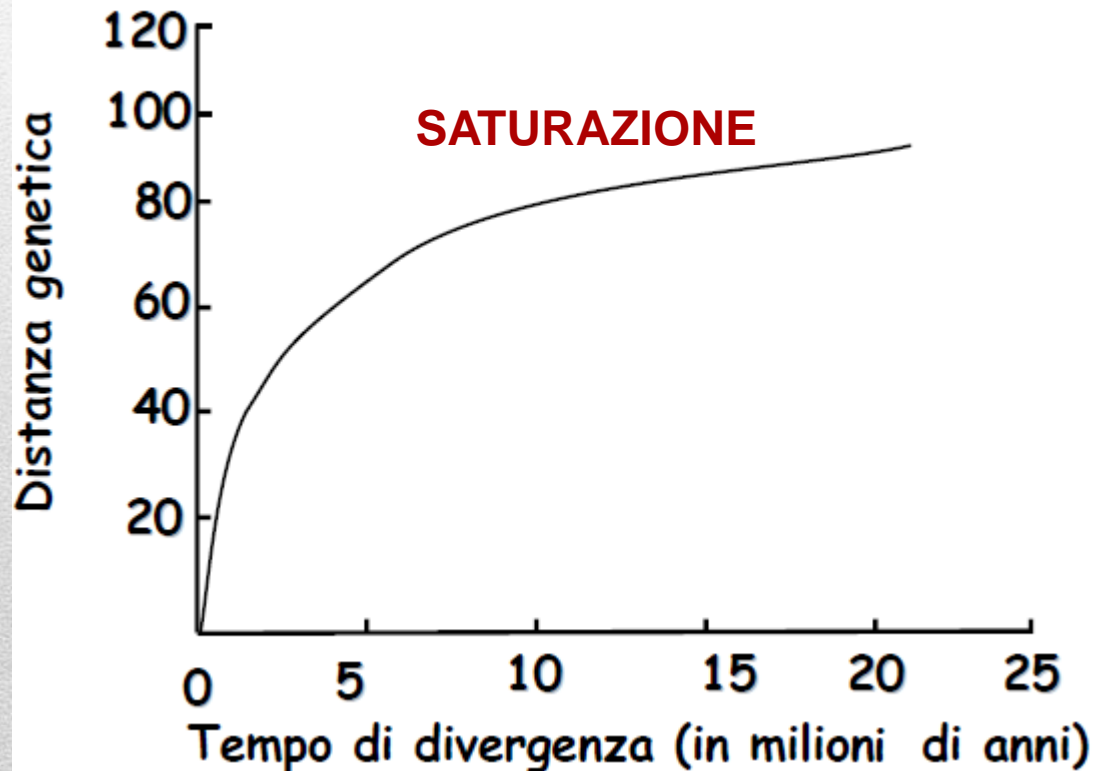
2 mutazioni,  
**1** differenza

Il semplice conteggio del numero di differenze tra sequenze ( $p$  distance =  $n.$ sostituzioni/ $n.$ 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**, aumenta la probabilità che più di un sito vada incontro a sostituzioni **multiple**



A causa delle sostituzioni multiple, le distanze osservate possono sottostimare il reale ammontare del cambiamento evolutivo. Sono stati, quindi, 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<sub>AC</sub>** è la probabilità che **A** muti in **C** nell'intervallo **t**

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

---

## Modello di Jukes-Cantor

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

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

$\alpha$  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 problemi dovuti a cambiamenti multipli nello stesso sito

Es.1:

$D = 0.05$  (identità = 95%)

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

sequenze **molto simili** : ci si aspettano pochi cambiamenti multipli nello stesso sito, poichè il tempo di divergenza **è breve**.

---

Il logaritmo naturale è usato per **correggere** i problemi dovuti a cambiamenti multipli nello stesso sito

Es.2:

$D = 0.5$  (identità = 50%)

$$d_{xy} = - (3/4) \ln (1 - 4/3 D) = - (3/4) \ln (1 - 4/3 0.5) = 0.824$$

sequenze **poco simili** : ci si aspettano molti cambiamenti multipli nello stesso sito, poichè il tempo di divergenza **è grande**. (Il rischio di sottostimare le distanze è maggiore)

---

## 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

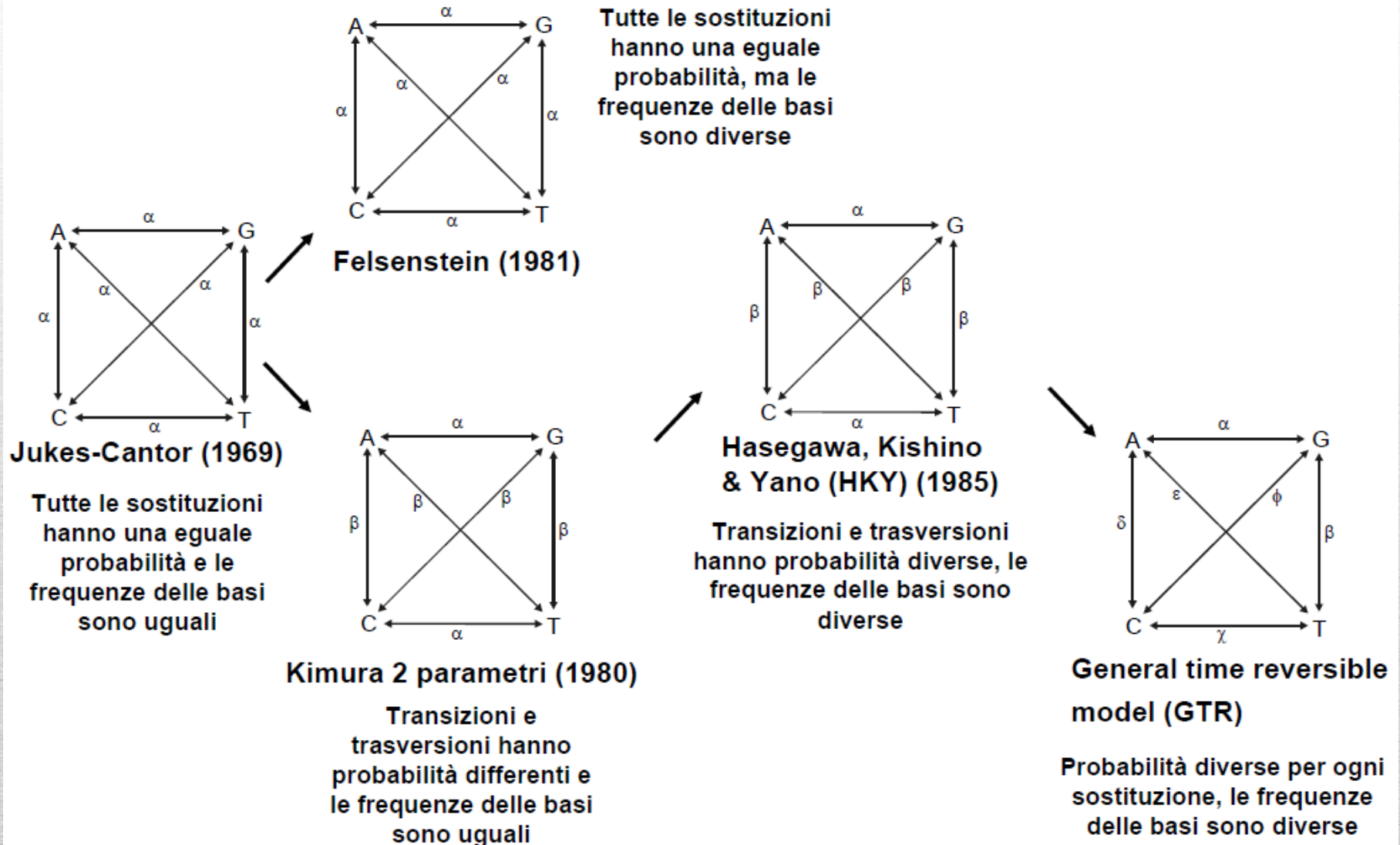
I parametri più comuni che vengono aggiunti sono:

- 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

---

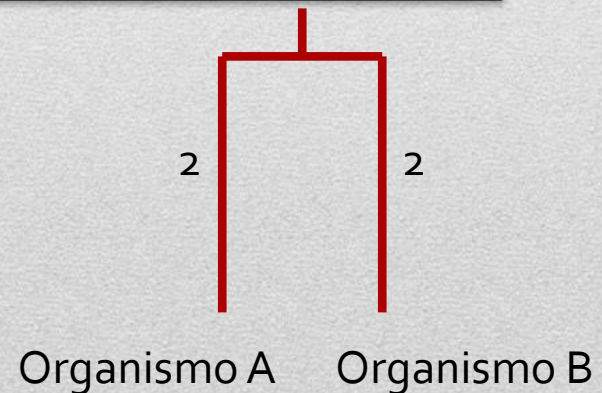
## «Evoluzione» dei modelli evolutivi :



- Servono le **distanze** tra tutte le coppie di sequenze
- Come misurare le distanze?
- Vogliamo misurare il **numero di mutazioni** verificatesi da quando le specie si sono separate

Contiamo il numero di colonne dell'allineamento pairwise in cui le sequenze sono differenti e dividiamo per la lunghezza delle sequenze: **probabilità di mutazione per sito** (NB: STIMA NON CORRETTA)

**Distanza tra  
organismo A e B è 4**

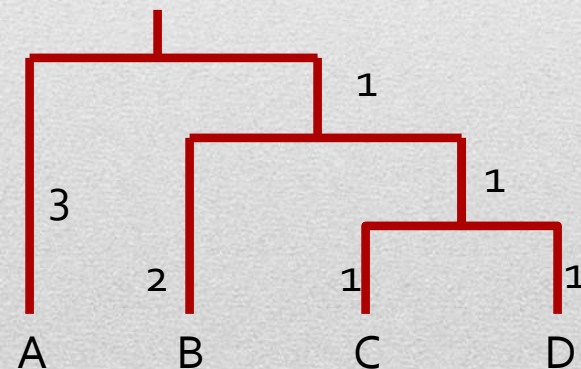




- Tutte le distanze pairwise

	A	B	C	D
A	0	6	6	6
B		0	4	4
C			0	2
D				0

- Quel che vogliamo ottenere ( albero )



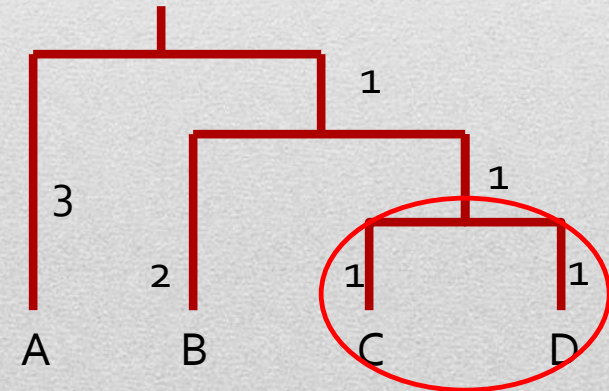
## Esempio ( 4 OTU ): Algoritmo UPGMA per costruire un albero

1. Troviamo le OTU più vicine
2. Mettiamole **vicine** nell'albero
3. Calcoliamo la distanza **MEDIA** dal resto delle OTU

	A	B	C	D
A	0	6	6	6
B		0	4	4
C			0	2
D				0



	A	B	CD
A	0	6	6
B		0	4
CD			0



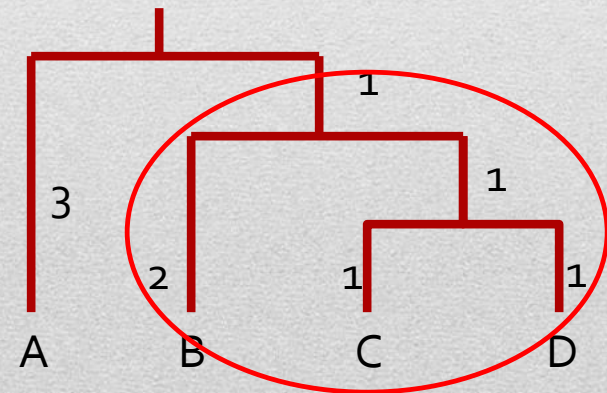
Distanza media:  $(4 + 4) / 2 = 4$

Distanza media:  $(6 + 6 + 6) / 3 = 6$

1. Troviamo la prossima OTU più vicina
2. Mettiamola vicina nell'albero
3. SE ESISTONO ALTRE OTU
  - I. Calcoliamo distanza media dal resto dell OTU
  - II. Ripartiamo da 1

	A	B	CD
A	0	6	6
B		0	4
CD			0

	A	BCD
A	0	6
BCD		0

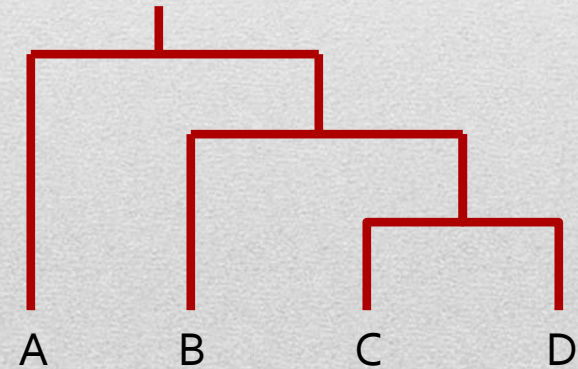


## Esempio 2 ( 4 OTU ) : Algoritmo UPGMA per costruire un albero

- Nuova matrice delle distanze

	A	B	C	D
A	0	6	6	7
B		0	4	5
C			0	3
D				0

- Quel che vogliamo ottenere ( albero )



## Esempio 2 ( 4 OTU ) : Algoritmo UPGMA per costruire un albero

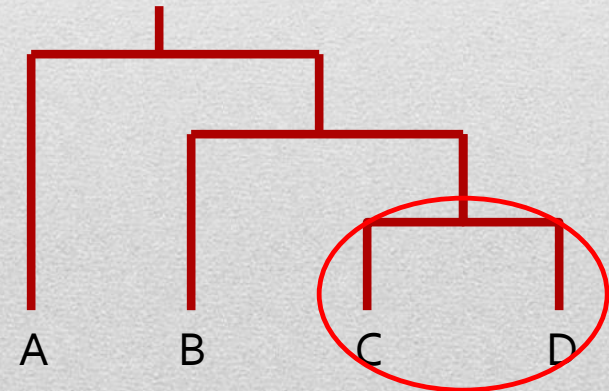
- C e D sono ancora le OTU più vicine
- Iniziamo a costruire l'albero usando C e D
- Calcolare la distanza MEDIA dal resto delle OTU

	A	B	C	D
A	0	6	6	7
B		0	4	5
C			0	3
D				0

	A	B	CD
A	0	6	6.5
B		0	4.5
CD			0

$$l(B, CD) = \frac{l(B, C) + l(B, D)}{2} = \frac{4 + 5}{2} = 4.5$$

$$l(A, CD) = \frac{l(A, C) + l(A, D)}{2} = \frac{6 + 7}{2} = 6.5$$



- Troviamo la OTU più vicina
- Posizioniamola vicino nell'albero (collassiamo B con CD)
- Calcolare la distanza MEDIA dal resto delle OTU

	A	B	CD
A	0	6	6.5
B		0	4.5
CD			0

$$l(A, BCD) = \frac{l(A, B) + l(A, CD)}{2} = \frac{6 + 6.5}{2} = 6.25$$

	A	BCD
A	0	6.25
BCD		0

