# Bioinformatica

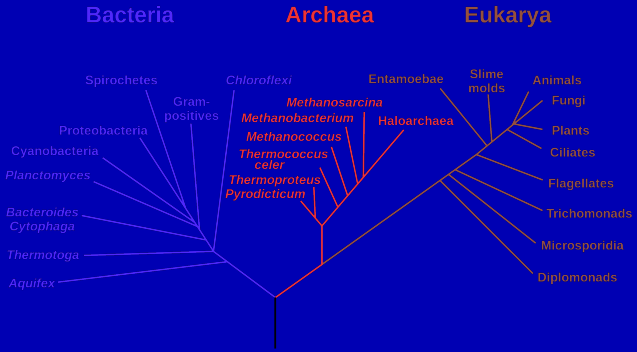
Gianluca Della Vedova

Univ. Milano–Bicocca
https://www.unimib.it/gianluca-della-vedova

15 marzo 2024

## Alberi evolutivi — Filogenesi

# Alberi evolutivi — Filogenesi

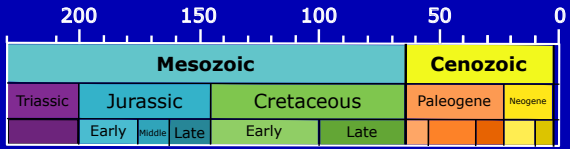


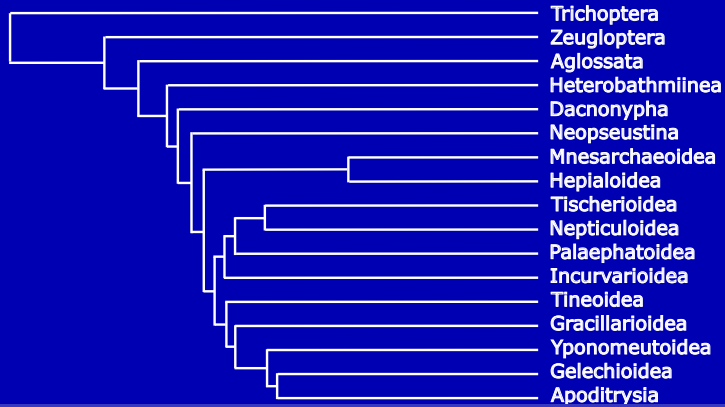

## Alberi etichettati

etichetta = distanza stimata

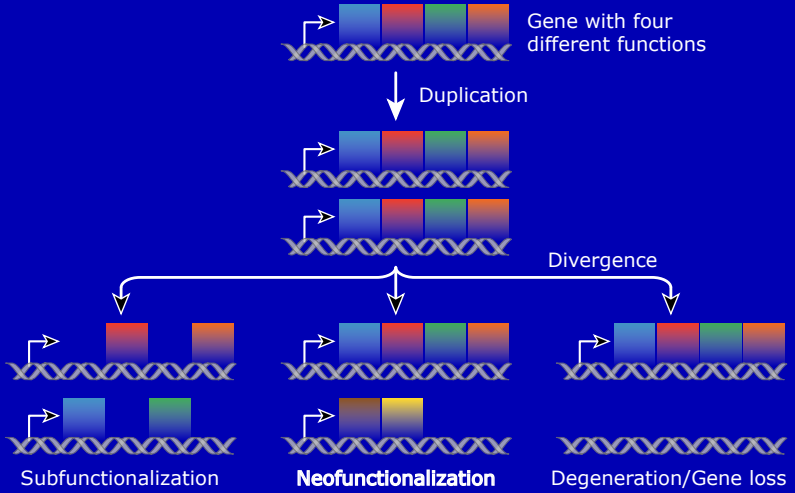
Public domain `https://en.wikipedia.org/wiki/File:Phylogenetic_tree.svg`

# Distanze


200
150
100
50
0
Mesozoic
Cenozoic
Triassic
Jurassic
Cretaceous
Paleogene
Neogene
Early
Middle
Late
Early
Late


Trichoptera
Zeugloptera
Aglossata
Heterobathmiinea
Dacnonypha
Neopseustina
Mnesarchaeoidea
Hepialoidea
Tischerioidea
Nepticuloidea
Palaephatoidea
Incurvarioidea
Tineoidea
Gracillarioidea
Yponomeutoidea
Gelechioidea
Apoditrysia

# Geni duplicati



# Approcci basati su distanze.

## Distanza

$d : S \times S \mapsto \mathbb{R}^+$ tale che:

1. $d(a, b) = 0 \Leftrightarrow a = b, \forall a, b \in S$
2. $d(a, b) = d(b, a), \forall a, b \in S$ (simmetria)
3. $d(a, b) \leqslant d(a, c) + d(c, b), \forall a, b, c \in S$ (disuguaglianza triangolare)

# Approcci basati su distanze.

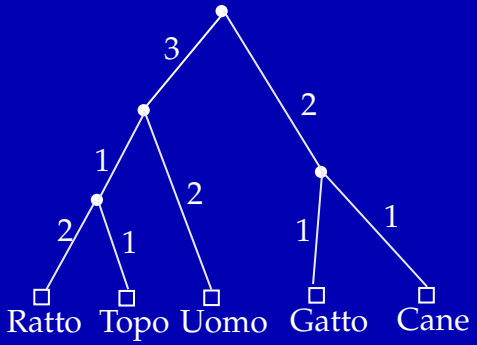
| | U | T | R | C | G |
|---|---|---|---|---|---|
| Uomo | 0 | 4 | 5 | 7 | 6 |
| Topo | - | 0 | 3 | 8 | 5 |
| Ratto | - | - | 0 | 9 | 7 |
| Cane | - | - | - | 0 | 2 |
| Gatto | - | - | - | - | 0 |

## Problema

- Input: matrice $d$ di distanze stimate
- Output: un albero con distanze $p = d$, se esiste

# Approcci basati su distanze.

| | U | T | R | C | G |
|---|---|---|---|---|---|
| Uomo | 0 | 4 | 5 | 9 | 9 |
| Topo | - | 0 | 3 | 8 | 8 |
| Ratto | - | - | 0 | 9 | 9 |
| Cane | - | - | - | 0 | 2 |
| Gatto | - | - | - | - | 0 |



# Approcci basati su distanze.

## Problema

- Input: matrice $d$ di distanze stimate
- Output: un albero con distanze $p$

# Approcci basati su distanze.

## Problema

- Input: matrice $d$ di distanze stimate
- Output: un albero con distanze $p$

## Funzioni obiettivo

1. $\max_{i,j} |d_{i,j} - p_{i,j}|$
2. $\sum_{i,j} |d_{i,j} - p_{i,j}|$
3. $\sum_{i,j} (d_{i,j} - p_{i,j})^2$

# Matrice di incidenza arco-percorso

## Variabili

- $y^e_{i,j} = 1$ sse il lato $e$ appartiene al cammino da $i$ a $j$
- $d_{i,j}$: distanza in input da $i$ a $j$
- $p_{i,j}$: distanza predetta da $i$ a $j$
- $w_e$: peso predetto di $e$

# Matrice di incidenza arco-percorso

## Variabili

- $y^e_{i,j}$ = 1 sse il lato $e$ appartiene al cammino da $i$ a $j$
- $d_{i,j}$: distanza in input da $i$ a $j$
- $p_{i,j}$: distanza predetta da $i$ a $j$
- $w_e$: peso predetto di $e$

$$\min \sum_{i \neq j} \left(d_{i,j} - p_{i,j}\right)^2$$

soggetto a

$$p_{i,j} = \sum_{e \in E(p_{i,j})} y^e_{i,j} w_e \quad \forall i, j$$

$$w_e = \sum_{i \neq j} \sum_{e \in E(p_{i,j})} y^{e_t}_{i,j} d_{i,j} \quad \forall e$$

# Bilanciamento

**archi incidenti su una foglia vs. archi centrali**

Diverso peso

$t_{i,j}$: distanza non pesata da $i$ a $j$ nell'albero predetto

# Bilanciamento

$$\min \sum_{i} \sum_{j, i \neq j} d_{i,j} 2^{-t_{i,j}}$$

$t_{i,j}$: distanza non pesata da $i$ a $j$ nell'albero predetto

**archi incidenti su una foglia vs. archi centrali**

Diverso peso

# Bilanciamento

$$\min \sum_{i} \sum_{j, i \neq j} d_{i,j} 2^{-t_{i,j}}$$

**archi incidenti su una foglia vs. archi centrali**

Diverso peso

$t_{i,j}$: distanza non pesata da $i$ a $j$ nell'albero predetto

**Legami on TSP**

# Bilanciamento

$$\min \sum_{i} \sum_{j, i \neq j} d_{i,j} 2^{-t_{i,j}}$$

$t_{i,j}$: distanza non pesata da $i$ a $j$ nell'albero predetto

## archi incidenti su una foglia vs. archi centrali

Diverso peso

## Legami on TSP

$$2 \sum_{e} w_e = \sum_{i=1}^{n} d_{\Pi(i),\Pi(i+1)}$$

# Caratterizzazione matrice arco-percorso

## Matrice lunghezza $l_{i,j}$

1. $l_{i,j} = 0$
2. $l_{i,j} = l_{j,i}$
3. $l_{i,j} + l_{j,k} - l_{i,k} \geqslant 2, \forall i \neq j \neq k, i \neq k$
4. $\sum_{i \neq j} 2^{-l_{i,j}} = \frac{1}{2}$ (uguaglianza di Kraft)
5. $\min \quad \sum_i \sum_{j, i \neq j} l_{i,j} 2^{-l_{i,j}} = 2n - 3$
6. Esattamente una delle seguenti
   - $l_{i,j} + l_{p,q} + 2 \leqslant l_{i,q} + l_{j,p} = l_{i,p} + l_{j,q}$
   - $l_{i,q} + l_{j,p} + 2 \leqslant l_{i,j} + l_{p,q} = l_{i,p} + l_{j,q}$
   - $l_{i,p} + l_{j,q} + 2 \leqslant l_{p,q} + l_{p,q} = l_{i,q} + l_{j,p}$

# UPGMA

- Unweighted Pair Group with Arithmetic Mean
- $D(C_1, C_2) \leftarrow \frac{1}{|C_1||C_2|} \sum_{i \in C_1} \sum_{j \in C_2} D(i, j)$
- All'inizio $h = 0$ per ogni cluster/specie
- Fondi i due cluster $C_1$, $C_2$ con minimo $D(\cdot, \cdot)$, ottenendo $C$
- Per ogni cluster $C^* \neq C$, $D(C, C^*) = \frac{1}{|C||C^*|} \sum_{i \in C} \sum_{j \in C^*} D(i, j)$
- $h(C) \leftarrow \frac{1}{2} D(C_1, C_2)$
- $h(C) - h(C_1)$ etichetta $(C, C_1)$; $h(C) - h(C_2)$ etichetta $(C, C_2)$
- UPGMA produce ultrametrica

# Neighbor Joining.

- $D(C_1, C_2) \leftarrow \frac{1}{|C_1||C_2|} \sum_{i \in C_1} \sum_{j \in C_2} D(i, j)$
- $u(C) \leftarrow \frac{1}{\text{num. cluster} - 2} \sum_{C_3} D(C, C_3)$
- Fondi i due cluster $C_1$, $C_2$ con minimo $D(C_1, C_2) - u(C_1) - u(C_2)$, ottenendo $C$
- Per ogni cluster $C^* \neq C$, $D(C, C^*) = \frac{1}{|C||C^*|} \sum_{i \in C} \sum_{j \in C^*} D(i, j)$
- $\frac{1}{2}(D(C_1, C_2) + u(C_1) - u(C_2))$ etichetta $(C, C_1)$
- $\frac{1}{2}(D(C_1, C_2) + u(C_2) - u(C_1))$ etichetta $(C, C_2)$

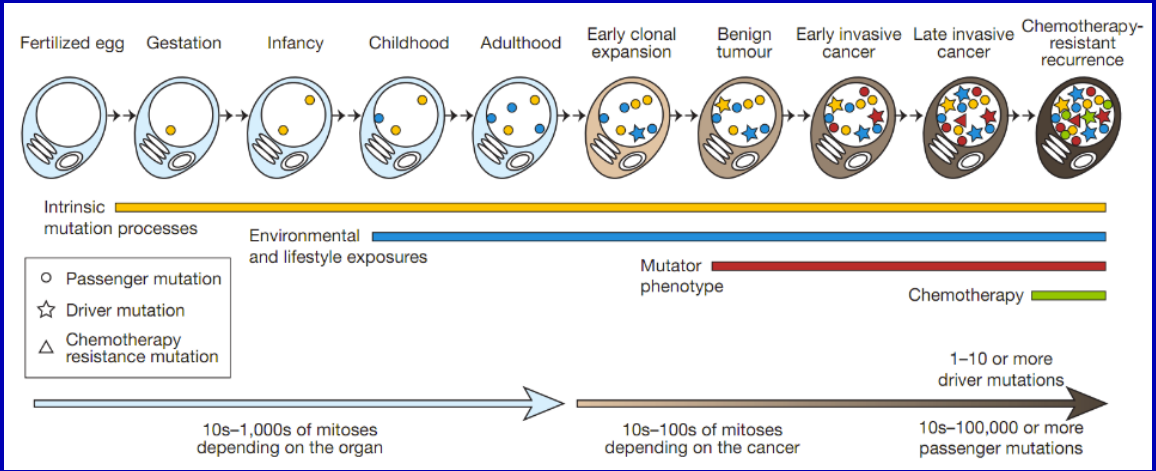
# Trovare $i, j$ che minimizza $Q(i, j) = D(C_i, C_j) - u(C_i) - u(C_j)$

## Stutture dati

- S: D con ogni riga in ordine crescente
- I: mappa da S a D

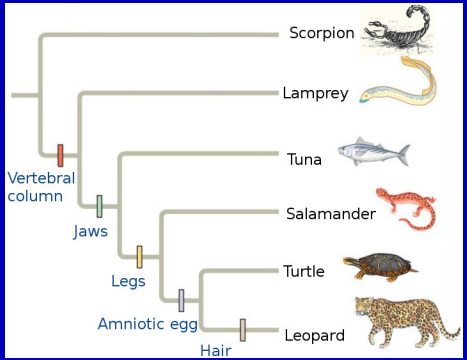
1. $q_{min} \leftarrow \infty, i \leftarrow -1, j \leftarrow -1$
2. $\forall r, c$
   1. se $S(r, c) - u(r) - u_{max} > q_{min}$ allora vai alla prossima riga
   2. se $Q(r, I(r, c)) < q_{min}$ allora $q_{min} \leftarrow Q(r, I(r, c))$, $i, j \leftarrow r, I(r, c)$

# Evoluzione in un individuo




- Cellule accumulano mutazioni durante la vita

# Evoluzione basata su caratteri

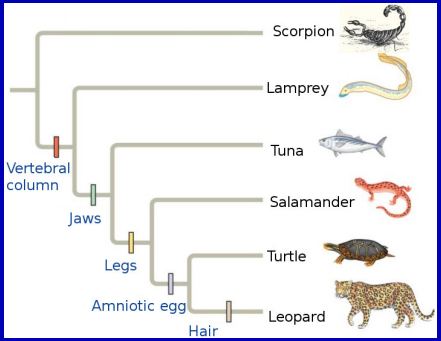



## Regola 1 (semplice)

Ogni carattere è acquisito esattamente una volta nell'albero.

# Filogenesi perfetta

| | A | J | H | L | V |
|---|---|---|---|---|---|
| Scorpione | 0 | 0 | 0 | 0 | 0 |
| Anguilla | 0 | 0 | 0 | 0 | 1 |
| Tonno | 0 | 1 | 0 | 0 | 1 |
| Salamandra | 0 | 1 | 0 | 1 | 1 |
| Tartaruga | 1 | 1 | 0 | 1 | 1 |
| Leopardo | 1 | 1 | 1 | 1 | 1 |




## Problema

- Input: matrice binaria $M$
- Output: un albero che spiega $M$, se esiste

## Algorithm di Gusfield — lineare

1. Radix Sort delle colonne, in ordine decrescente (anche del numero di 1)
2. Costruire l'albero, una specie alla volta

# Caratteri e stati

## Cambio di stato

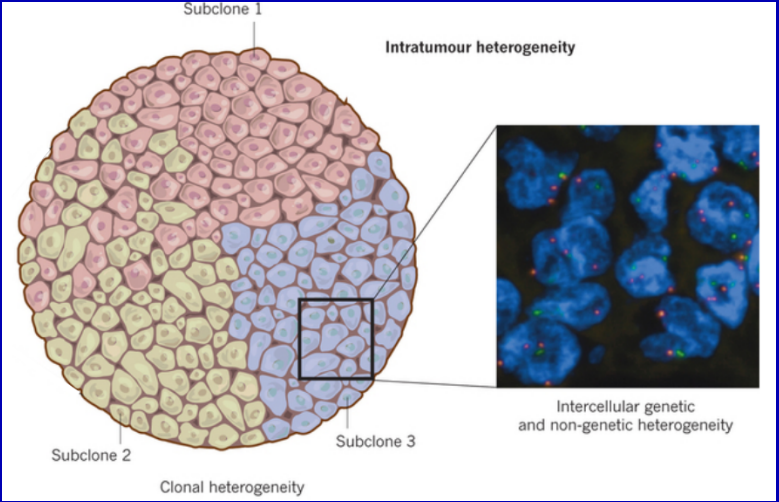
- Un carattere c è acquisito $\Rightarrow$ lo stato di c passa da 0 a 1 in un arco
- Un carattere c è perso $\Rightarrow$ lo stato di c passa da 1 a 0 in un arco (mutazione ricorrente)

## Modelli di evoluzione

Ogni carattere c è acquisito esattamente una volta nell'albero.

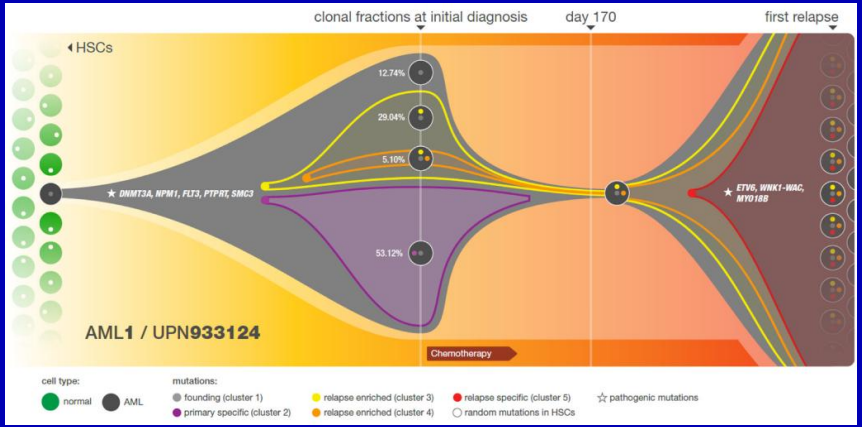
1. Filogenesi perfetta: nessuna mutazione ricorrente, nessuna perdita
2. Dollo: mutazioni ricorrenti senza limiti, ma senza perdite
3. Camin-Sokal: Perdite senza limiti, ma senza mutazioni ricorrenti

# Tumori



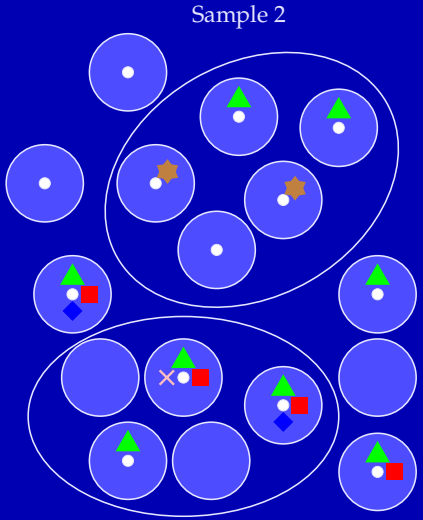

- Un tumore contiene sia cellule cancerose che sane
- Un tumore è un miscuglio di cloni (sottopopolazioni) diverse.

# Evoluzione tumorale




- I cloni compaiono con numerosità differente nel tumore

# Evoluzione tumorale



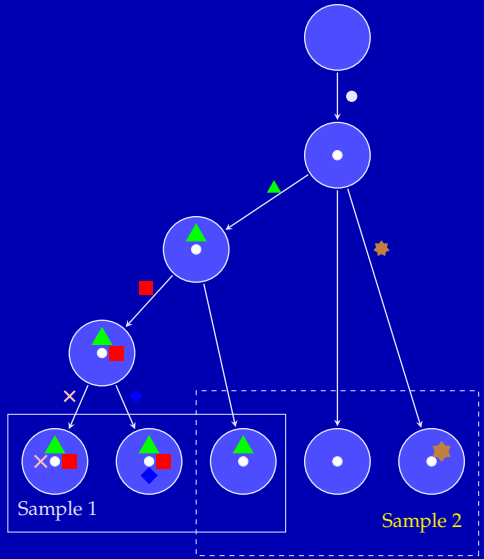

- Un campione contiene diversi cloni
- Per ogni campione, abbiamo la frequenza con cui ogni mutazione appare
- matrice di frequenze F

| | ◆ | ▲ | ● | ■ | ✕ | ✸ |
|---|---|---|---|---|---|---|
| $S_1$ | 0.2 | 0.6 | 0.6 | 0.4 | 0.2 | 0.0 |
| $S_2$ | 0.0 | 0.4 | 1.0 | 0.0 | 0.0 | 0.4 |

# Calcolare l'evoluzione tumorale

## Matrice B spiegata da T

| ◆ | ▲ | ● | ■ | ✕ | ✸ |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 1 | 1 | 0 |
| 0 | 1 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 0 | 0 |



# Modelli di evoluzione.

- Probabilità di transizione fra stati (A, C, G, T).
- dipende dal tempo trascorso fra i due eventi
- tasso istantaneo di mutazione
- probabilità di mutazione *in una generazione*: somma su ogni riga = 1

J. Felsenstein. Theoretical Evolutionary Genetics

# Modelli di evoluzione: Jukes-Cantor.

- ogni mutazione è equiprobabile
- $1 - \mu$: nessuna mutazione
- $\mu/3$: mutazione

# Modelli di evoluzione: Kimura 2 parametri

- Distinzione transizioni ($A \leftrightarrow G$, $C \leftrightarrow T$), transversioni
- $1 - \mu$: nessuna mutazione
- $\frac{R}{R+1}\mu$: probabilità transizione
- $\frac{1}{2(R+1)}\mu$: probabilità di trasversione $A \leftrightarrow C$ o $G \leftrightarrow T$
- $\frac{1}{2(R+1)}\mu$: probabilità di trasversione $A \leftrightarrow T$ o $C \leftrightarrow G$
- $R = \frac{R}{R+1}\mu / \left(2\frac{1}{2(R+1)}\mu\right)$: rapporto probabilità di transizioni / probabilità trasversioni

# Modelli di evoluzione: General time-reversible

- matrice simmetrica
- consequenza: alberi senza radice

# Massima verosimiglianza.

# Licenza d'uso