Metodi di Geometria Algebrica per la ricostruzione statistica degli alberi filogenetici

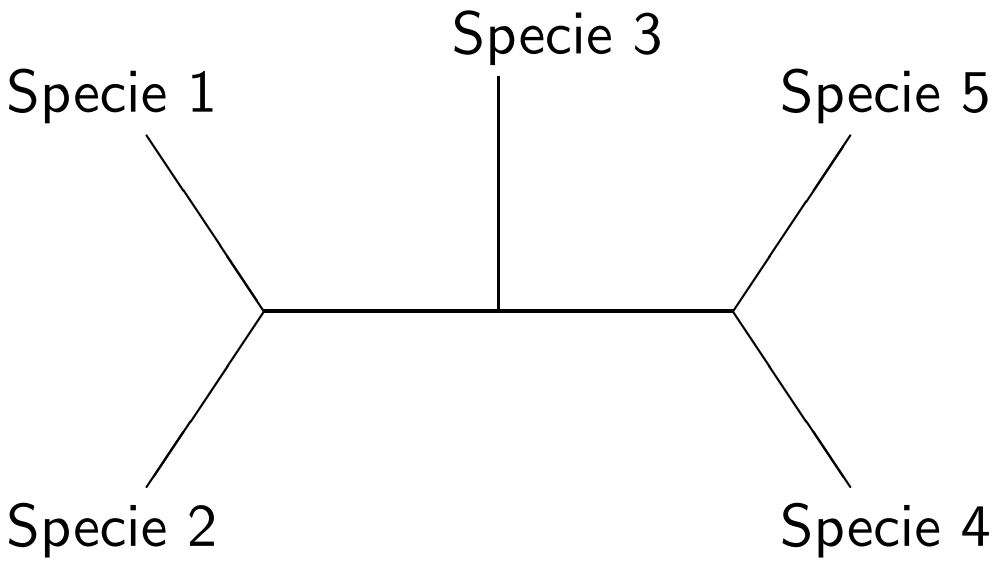
Luigi Biondi

20 Luglio 2011

Specie 1: ACGTACTACTGCAGTCCTAGCTGATCGT ...
Specie 2: ACTGTCGATCATGCTAATCGATGCATCG ...
Specie 3: GTCATCTACGACTACGACGCGATCGTAC ...
Specie 4: AGATCTGCTATCAGTCATCGACGTATAA ...
Specie 5: ACTGCACTCGGGACTCTCGCTACGACT ...

Allineamento di DNA

Variabili aleatorie a valori in $\{A, C, G, T\}$.



Abero Filogenetico

Un modello statistico discreto $\mathcal{M}$ è un sottoinsieme del *simplesso di probabilità* :

$$\mathcal{M} \subseteq \Delta_n = \{x \in \mathbb{R}^n : \sum_i x_i = 1, x_i \geq 0\}.$$

Un modello statistico discreto $\mathcal{M}$ è un sottoinsieme del *simplesso di probabilità* :

$$\mathcal{M} \subseteq \Delta_n = \{x \in \mathbb{R}^n : \sum_i x_i = 1, x_i \geq 0\}.$$

Ad esempio, se $X_1, \ldots, X_n$ sono variabili aleatorie a valori rispettivamente in $[k_1], \ldots, [k_n]$, possiamo definire un tensore $p \in \mathbb{R}^{k_1} \otimes \cdots \otimes \mathbb{R}^{k_n}$ ponendo

$$p_{i_1 \ldots i_n} = \mathbf{P}[X_1 = i_1, \ldots, X_n = i_n],$$

detto *tensore di probabilità*.

Un modello statistico discreto $\mathcal{M}$ è un sottoinsieme del *simplesso di probabilità* :

$$\mathcal{M} \subseteq \Delta_n = \{x \in \mathbb{R}^n : \sum_i x_i = 1, x_i \geq 0\}.$$

Ad esempio, se $X_1, \ldots, X_n$ sono variabili aleatorie a valori rispettivamente in $[k_1], \ldots, [k_n]$, possiamo definire un tensore $p \in \mathbb{R}^{k_1} \otimes \cdots \otimes \mathbb{R}^{k_n}$ ponendo

$$p_{i_1 \ldots i_n} = \mathbf{P}[X_1 = i_1, \ldots, X_n = i_n],$$

detto *tensore di probabilità*. Dato che

$$\sum_{i_1, \ldots, i_n} p_{i_1 \ldots i_n} = 1 \qquad \text{e} \qquad p_{i_1 \ldots i_n} \geq 0$$

$p \in \Delta_K$ con $K = \prod k_i$: ogni sottoinsieme di Δ_K può rappresentare una collezione di distribuzioni per $X_1, \ldots, X_n$.

$X_1, \ldots, X_n$ variabili aleatorie si dicono *mutuamente indipendenti* se

$$\mathbf{P}[X_1 = i_1, \ldots, X_n = i_n] = \prod_{k=1}^{n} \mathbf{P}[X_k = i_k] \qquad \forall i_1, \ldots i_n.$$

$X_1, \ldots, X_n$ variabili aleatorie si dicono *mutuamente indipendenti* se

$$\mathbf{P}[X_1 = i_1, \ldots, X_n = i_n] = \prod_{k=1}^{n} \mathbf{P}[X_k = i_k] \qquad \forall i_1, \ldots i_n.$$

Se p è il tensore di probabilità associato a queste variabili, questa condizione si può scrivere come

$$p = p^1 \otimes \cdots \otimes p^n,$$

dove p^k è il vettore delle *probabilità marginali* di X_k: $p^k_j = \mathbf{P}[X_k = j]$.

$\varphi \in V_1 \otimes \cdots \otimes V_m$ si dice *decomponibile* se esistono $v_i \in V_i$ tali che

$$\varphi = v_1 \otimes \cdots \otimes v_n.$$

$\varphi \in V_1 \otimes \cdots \otimes V_m$ si dice *decomponibile* se esistono $v_i \in V_i$ tali che

$$\varphi = v_1 \otimes \cdots \otimes v_n.$$

Definizione

$\mathcal{M}_{\perp\!\!\!\perp} = \{p \in \Delta_K : p \text{ decomponibile}\}$ è detto *modello di indipendenza*.

$\varphi \in V_1 \otimes \cdots \otimes V_m$ si dice *decomponibile* se esistono $v_i \in V_i$ tali che

$$\varphi = v_1 \otimes \cdots \otimes v_n.$$

Definizione

$\mathcal{M}_{\perp\!\!\!\perp} = \{p \in \Delta_K : p \text{ decomponibile}\}$ è detto *modello di indipendenza*.

I tensori 2-dimensionali decomponibili sono le matrici di rango 1.

$\varphi \in V_1 \otimes \cdots \otimes V_m$ si dice *decomponibile* se esistono $v_i \in V_i$ tali che

$$\varphi = v_1 \otimes \cdots \otimes v_n.$$

Definizione

$\mathcal{M}_{\perp\!\!\!\perp} = \{p \in \Delta_K : p \text{ decomponibile}\}$ è detto *modello di indipendenza*.

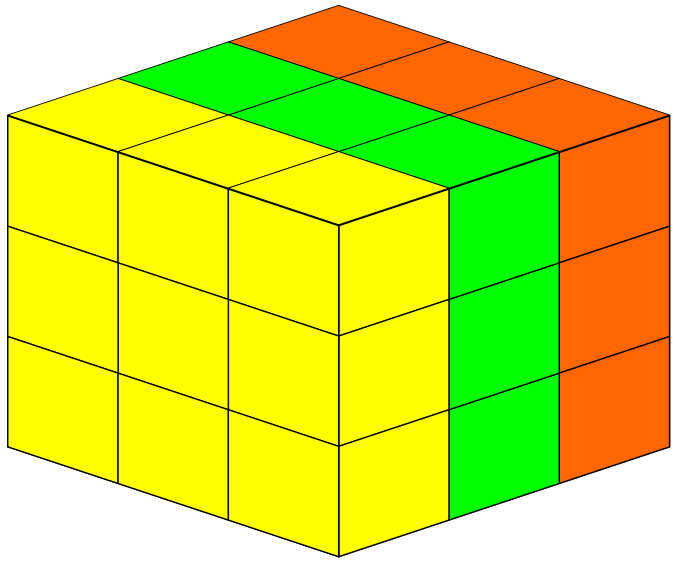
I tensori 2-dimensionali decomponibili sono le matrici di rango 1.

Definizione

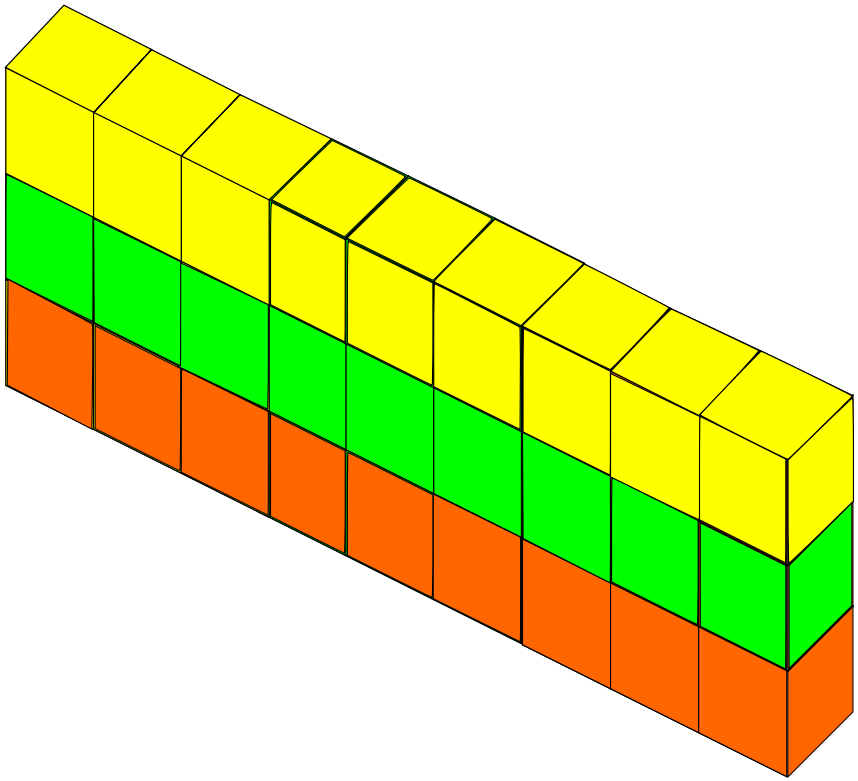
Sia $\varphi \in V_1 \otimes \cdots \otimes V_n$ e sia $\{A, B\}$ una partizione di $[n] = \{1, \ldots, n\}$. Il *flattening* di φ rispetto ad $\{A, B\}$ è l'immagine di φ (denotata con $\text{Flat}_A(\varphi)$) tramite l'isomorfismo naturale

$$\psi : V_1 \otimes \cdots \otimes V_n \longrightarrow \text{Hom}\left(\bigotimes_{a \in A} V_a^*, \bigotimes_{b \in B} V_b\right).$$

In pratica, un flattening è una matrice ottenuta grazie ad un "appiattimento" del tensore.
Esempio:

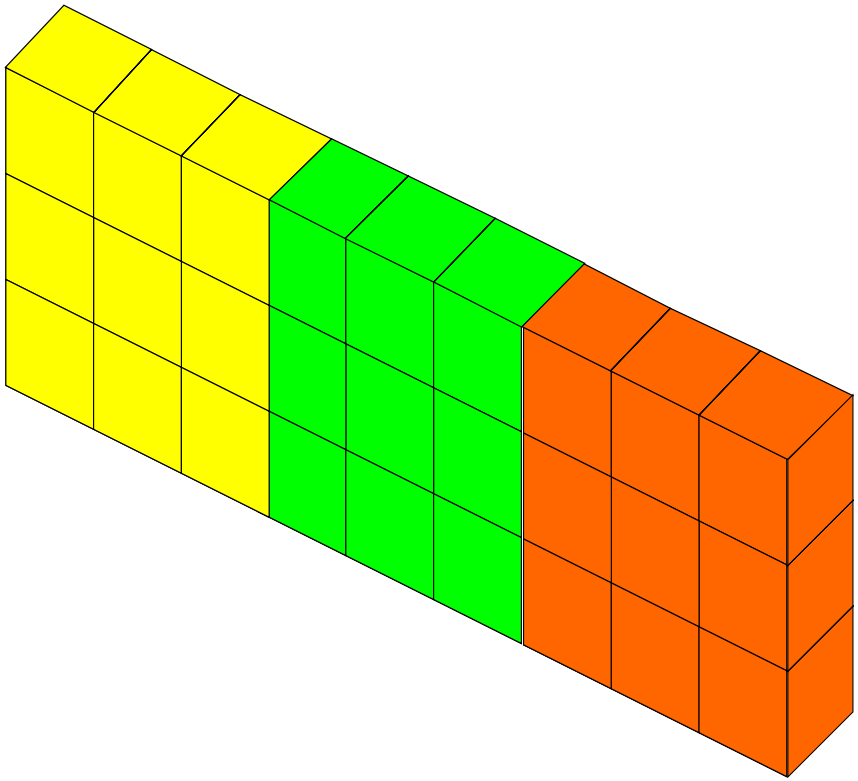


φ



$Flat_1(\varphi)$

Corrisponde alla partizione $\{1\}, \{2, 3\}$.



$Flat_3(\varphi)$

Corrisponde alla partizione $\{3\}, \{1, 2\}$.

Teorema

$\varphi \in V_1 \otimes \cdots \otimes V_n$ *è decomponibile se e soltanto se* $\operatorname{rk} \mathrm{Flat}_i(\varphi) \leq 1$ *per ogni* $i = 1 \ldots n$.

Teorema

$\varphi \in V_1 \otimes \cdots \otimes V_n$ *è decomponibile se e soltanto se* $\operatorname{rk} \operatorname{Flat}_i(\varphi) \leq 1$ *per ogni* $i = 1 \ldots n$.

Seguono le equazioni per il modello di indipendenza nel caso generale:

$$\mathcal{M}_{\perp\!\!\!\perp} = \Delta_K \cap V(F),$$

dove $F = \{\text{minori } 2 \times 2 \text{ dei flattening di } p \text{ rispetto alle coordinate}\}$.

Teorema

$\varphi \in V_1 \otimes \cdots \otimes V_n$ *è decomponibile se e soltanto se* $\operatorname{rk} \operatorname{Flat}_i(\varphi) \leq 1$ *per ogni* $i = 1 \ldots n$.

Seguono le equazioni per il modello di indipendenza nel caso generale:

$$\mathcal{M}_{\perp\!\!\!\perp} = \Delta_K \cap V(F),$$

dove $F = \{\text{minori } 2 \times 2 \text{ dei flattening di } p \text{ rispetto alle coordinate}\}$.
Il modello è descritto parametricamente dalla seguente applicazione:

$$\begin{aligned} \phi : \Delta_{k_1} \times \cdots \times \Delta_{k_n} &\longrightarrow \Delta_K \\ (p^1, \ldots, p^n) &\longmapsto p^1 \otimes \cdots \otimes p^n. \end{aligned}$$

Teorema

$\varphi \in V_1 \otimes \cdots \otimes V_n$ *è decomponibile se e soltanto se* $\operatorname{rk} \operatorname{Flat}_i(\varphi) \leq 1$ *per ogni* $i = 1 \ldots n$.

Seguono le equazioni per il modello di indipendenza nel caso generale:

$$\mathcal{M}_{\perp\!\!\!\perp} = \Delta_K \cap V(F),$$

dove $F = \{\text{minori } 2 \times 2 \text{ dei flattening di } p \text{ rispetto alle coordinate}\}$.
Il modello è descritto parametricamente dalla seguente applicazione:

$$\begin{aligned} \phi : \Delta_{k_1} \times \cdots \times \Delta_{k_n} &\longrightarrow \Delta_K \\ (p^1, \ldots, p^n) &\longmapsto p^1 \otimes \cdots \otimes p^n. \end{aligned}$$

Il rango di un tensore φ è il minimo numero di tensori decomponibili φ^i per i quali $\varphi = \sum_i \varphi^i$: $\mathcal{M}_{\perp\!\!\!\perp} = \{p \in \Delta_K : \operatorname{rk} p = 1\}$.

Definizione

Sia $\mathcal{M} \subseteq \Delta_n$ un modello statistico. L's-esimo modello mistura di $\mathcal{M}$ è

$$\text{Mixt}^s(\mathcal{M}) = \left\{ \sum_{i=1}^{s} \pi_i p^i : \ \pi \in \Delta_s, p \in \mathcal{M} \right\}.$$

Definizione

Sia $\mathcal{M} \subseteq \Delta_n$ un modello statistico. L's-esimo modello mistura di $\mathcal{M}$ è

$$\text{Mixt}^s(\mathcal{M}) = \left\{ \sum_{i=1}^{s} \pi_i p^i : \ \pi \in \Delta_s, p \in \mathcal{M} \right\}.$$

Se $\mathcal{M} = \mathcal{M}_{\perp\!\!\!\perp}$ il modello prende il nome di *modello a classi latenti*: corrisponde alla situazione in cui esistono delle variabili indipendenti se condizionate ad un'unica variabile nascosta.

Definizione

Sia $\mathcal{M} \subseteq \Delta_n$ un modello statistico. L's-esimo modello mistura di $\mathcal{M}$ è

$$\text{Mixt}^s(\mathcal{M}) = \left\{ \sum_{i=1}^{s} \pi_i p^i : \ \pi \in \Delta_s, p \in \mathcal{M} \right\}.$$

Se $\mathcal{M} = \mathcal{M}_{\perp\!\!\!\perp}$ il modello prende il nome di *modello a classi latenti*: corrisponde alla situazione in cui esistono delle variabili indipendenti se condizionate ad un'unica variabile nascosta.

Proposizione

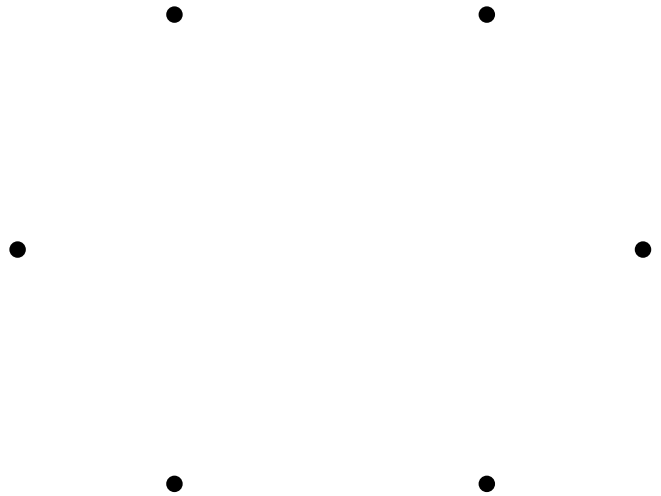
$p \in \text{Mixt}^s(\mathcal{M}_{\perp\!\!\!\perp})$ *allora* $\text{rk}(p) \leq s$.

Per definire modelli statistici più complessi useremo i grafi: sia $G = (V, E)$ un grafo orientato aciclico (DAG) e per ogni $v \in V$ consideriamo una variabile aleatoria X_v.

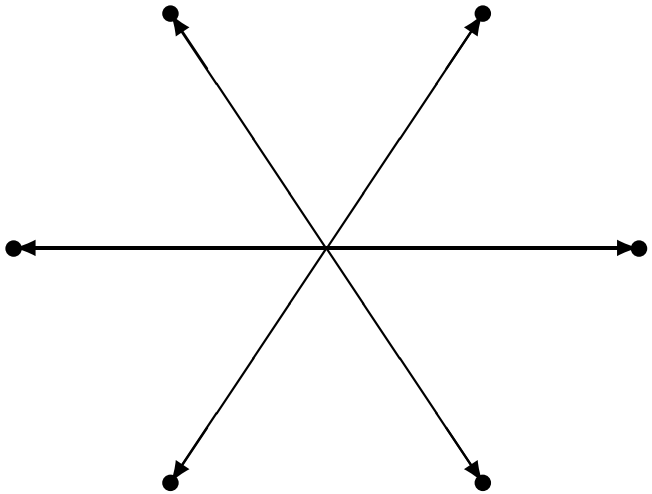
Definizione

Una distribuzione di probabilità per $\{X_v\}_{v \in V}$ soddisfa la *proprietà locale di Markov* per G se

$$X_v \perp\!\!\!\perp X_{\mathrm{nd}(v)} | X_{\mathrm{pa}(v)} \qquad \forall v \in V.$$



Grafo per il modello di indipendenza.



Grafo a stella.

Per poter sfruttare i risultati di geometria algebrica noti, cerchiamo la *chiusura di Zariski* dei modelli statistici.

Per poter sfruttare i risultati di geometria algebrica noti, cerchiamo la *chiusura di Zariski* dei modelli statistici.

- $\overline{\mathcal{M}_{\perp\!\!\!\perp}} = \mathbb{P}^{k_1-1} \times \cdots \times \mathbb{P}^{k_n-1} \subseteq \mathbb{P}^{K-1}$. (Varietà di Segre.)

Per poter sfruttare i risultati di geometria algebrica noti, cerchiamo la *chiusura di Zariski* dei modelli statistici.

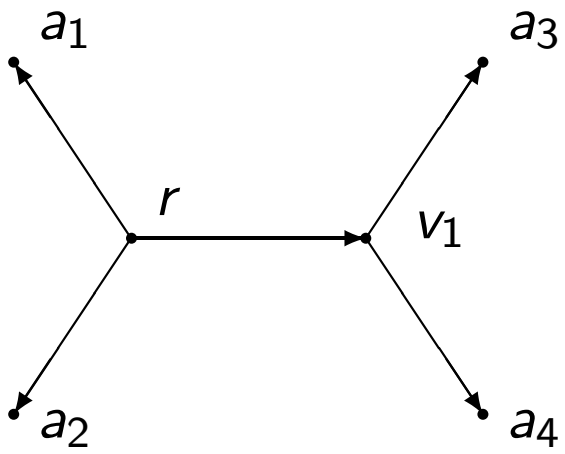
- $\overline{\mathcal{M}_{\perp\!\!\!\perp}} = \mathbb{P}^{k_1-1} \times \cdots \times \mathbb{P}^{k_n-1} \subseteq \mathbb{P}^{K-1}$. (Varietà di Segre.)
- $\overline{\text{Mixt}^s(\mathcal{M}_{\perp\!\!\!\perp})} = \sigma_s(\mathbb{P}^{k_1-1} \times \cdots \times \mathbb{P}^{k_n-1}) \subseteq \mathbb{P}^{K-1}$.

Per poter sfruttare i risultati di geometria algebrica noti, cerchiamo la *chiusura di Zariski* dei modelli statistici.

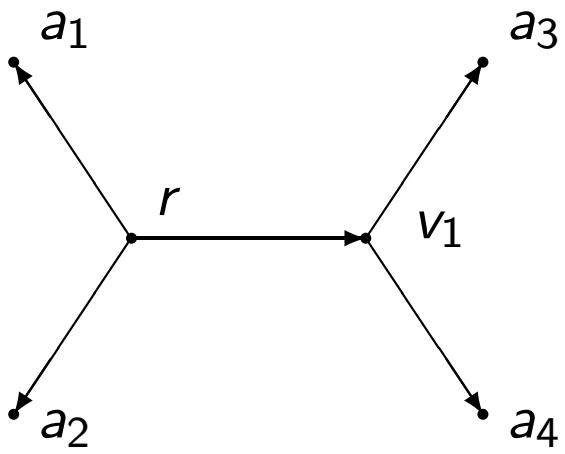
- $\overline{\mathcal{M}_{\perp\!\!\!\perp}} = \mathbb{P}^{k_1-1} \times \cdots \times \mathbb{P}^{k_n-1} \subseteq \mathbb{P}^{K-1}$. (Varietà di Segre.)
- $\overline{\text{Mixt}^s(\mathcal{M}_{\perp\!\!\!\perp})} = \sigma_s(\mathbb{P}^{k_1-1} \times \cdots \times \mathbb{P}^{k_n-1}) \subseteq \mathbb{P}^{K-1}$.

In generale, $\overline{\text{Mixt}^s(\mathcal{M})} = \sigma_s(\overline{\mathcal{M}})$.

Possiamo dare un'orientazione ad un albero fissando un suo vertice r (radice) e direzionando gli archi in senso uscente.

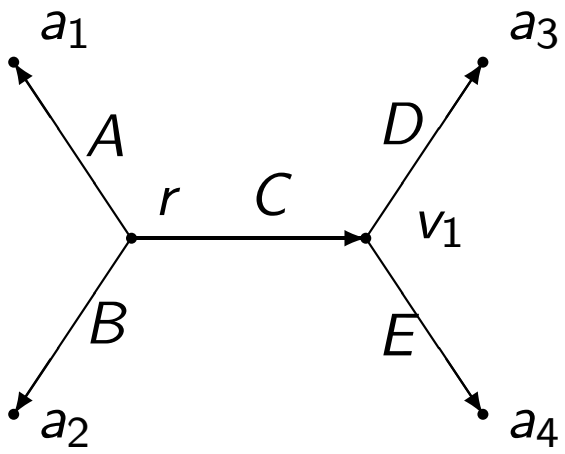


Possiamo dare un'orientazione ad un albero fissando un suo vertice r (radice) e direzionando gli archi in senso uscente.



Un modello grafico definito da un albero così orientato è determinato interamente da una distribuzione sulla radice e dalle probabilità di passaggio fra gli stati delle variabili associate ai vertici di uno stesso arco.

Possiamo dare un'orientazione ad un albero fissando un suo vertice r (radice) e direzionando gli archi in senso uscente.



Un modello grafico definito da un albero così orientato è determinato interamente da una distribuzione sulla radice e dalle probabilità di passaggio fra gli stati delle variabili associate ai vertici di uno stesso arco.

Vogliamo studiare i modelli $\mathcal{M}_T$ definiti da un albero T sotto le seguenti ipotesi:

Vogliamo studiare i modelli $\mathcal{M}_T$ definiti da un albero T sotto le seguenti ipotesi:

- T albero binario (vertici interni con valenza 3) con n foglie;

Vogliamo studiare i modelli $\mathcal{M}_T$ definiti da un albero T sotto le seguenti ipotesi:

- T albero binario (vertici interni con valenza 3) con n foglie;
- Variabili relative ai vertici a valori sempre nello stesso insieme $[q] = \{1 \dots q\}$;

Vogliamo studiare i modelli $\mathcal{M}_T$ definiti da un albero T sotto le seguenti ipotesi:

- T albero binario (vertici interni con valenza 3) con n foglie;
- Variabili relative ai vertici a valori sempre nello stesso insieme $[q] = \{1 \dots q\}$;
- Variabili relative ai vertici interni *latenti*: $\mathcal{M}_T \subseteq \Delta_{q^n}$.

Vogliamo studiare i modelli $\mathcal{M}_T$ definiti da un albero T sotto le seguenti ipotesi:

- T albero binario (vertici interni con valenza 3) con n foglie;
- Variabili relative ai vertici a valori sempre nello stesso insieme $[q] = \{1 \dots q\}$;
- Variabili relative ai vertici interni *latenti*: $\mathcal{M}_T \subseteq \Delta_{q^n}$.

$\mathcal{M}_T$ è chiamato in questo caso *modello filogenetico* e T *albero filogenetico*.
$V(T) = \overline{\mathcal{M}_T}$ è detta *varietà filogenetica*.

I modelli filogenetici sono modelli algebrici parametrici: è possibile definire una mappa polinomiale omogenea

$$\phi : \Delta_q \times \mathcal{A}^{|E|} \longrightarrow \Delta_{q^n},$$

dove $\mathcal{A}$ è l'insieme delle *matrici di transizione* associate agli archi, per cui $\mathcal{M}_T = \mathrm{Im}(\phi)$. Inoltre, l'immagine non dipende dalla posizione della radice.

I modelli filogenetici sono modelli algebrici parametrici: è possibile definire una mappa polinomiale omogenea

$$\phi : \Delta_q \times \mathcal{A}^{|E|} \longrightarrow \Delta_{q^n},$$

dove $\mathcal{A}$ è l'insieme delle *matrici di transizione* associate agli archi, per cui $\mathcal{M}_T = \mathrm{Im}(\phi)$. Inoltre, l'immagine non dipende dalla posizione della radice. Ci siamo occupati del *Modello Generale di Markov*, in cui

$$\mathcal{A} = \left\{ M \in \mathbb{R}^{q \times q} : \sum_j M_{ij} = 1, M_{ij} \geq 0 \right\}$$

è l'insieme delle matrici *stocastiche*.

Per poter usare concretamente i modelli filogenetici, dobbiamo trovare equazioni che lo generino.

Definizione

Un polinomio $f \in \mathbb{C}[x_1, \dots x_{q^n}]$ è detto *invariante filogenetico* per T se $f(p) = 0, \forall p \in V(T)$.

Indichiamo con $\mathcal{F}(T)$ l'insieme degli invarianti filogenetici per T.

Supponiamo di avere a disposizione una serie di realizzazioni indipendenti di una n-upla di variabili aleatorie $X_1, \ldots, X_n$, a valori in $[q]$ descritte da un albero filogenetico.
Teoricamente, se conoscessimo il tensore di probabilità p per queste variabili, troveremmo un albero i cui invarianti si annullano tutti in p.

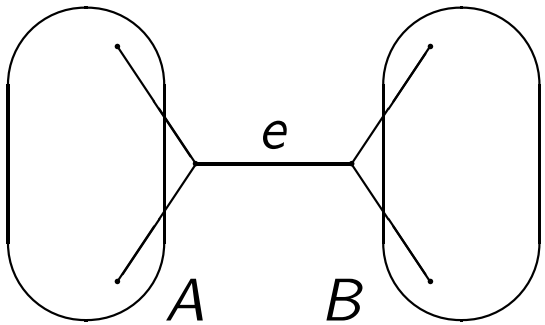
Supponiamo di avere a disposizione una serie di realizzazioni indipendenti di una n-upla di variabili aleatorie $X_1, \ldots, X_n$, a valori in $[q]$ descritte da un albero filogenetico.
Teoricamente, se conoscessimo il tensore di probabilità p per queste variabili, troveremmo un albero i cui invarianti si annullano tutti in p.

In pratica, dai dati possiamo ricavare soltanto il tensore delle frequenze $\hat{p}$:

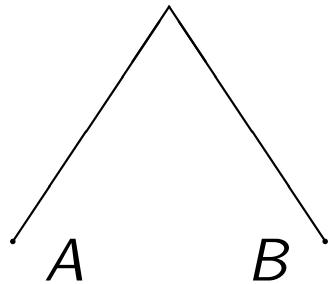
$$\hat{p}_{i_1,\ldots,i_n} = \frac{\sharp \text{ occorrenze di } (i_1, \ldots, i_n)}{\sharp \text{ realizzazioni totali}}.$$

Sceglieremo quindi l'albero corrispondente agli invarianti più "piccoli".

Per ogni arco di T consideriamo la bi-partizione delle foglie indotta dalla sua cancellazione (*split*).

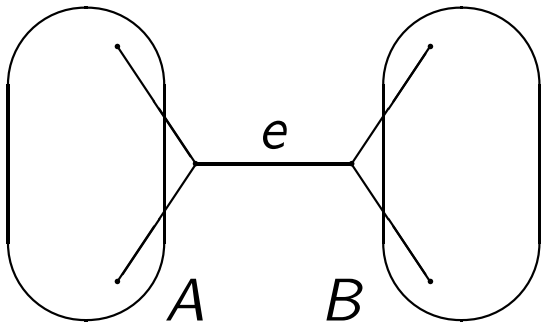


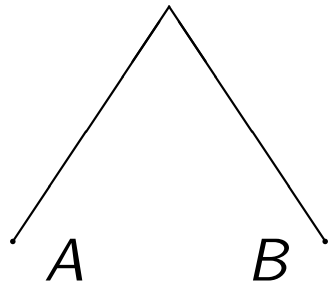
p



$\mathrm{Flat}_e(p)$

Per ogni arco di T consideriamo la bi-partizione delle foglie indotta dalla sua cancellazione (*split*).





p $\qquad\qquad$ $\mathrm{Flat}_e(p)$

Dunque $\mathrm{rk}(\mathrm{Flat}_e(p)) \leq q$: un insieme di invarianti filogenetici è

$$\mathcal{F}_{edge}(T) = \bigcup_e \{\text{minori}(q+1) \times (q+1) \text{ di } \mathrm{Flat}_e(p)\},$$

detti *invarianti degli archi*.

Nel caso di variabili binarie questi costituiscono l'intero insieme degli invarianti:

Teorema

$q = 2 \Rightarrow \mathcal{F}(T) = \mathcal{F}_{edge}(T)$.

In generale gli invarianti degli archi non definiscono interamente la varietà filogenetica.

Nel caso di variabili binarie questi costituiscono l'intero insieme degli invarianti:

Teorema

$q = 2 \Rightarrow \mathcal{F}(T) = \mathcal{F}_{edge}(T)$.

In generale gli invarianti degli archi non definiscono interamente la varietà filogenetica.

Teorema (Casanellas, Fernandez-Sanchez 2009)

Per il modello generale di Markov gli invarianti degli archi sono sufficienti alla ricostruzione filogenetica.

Algoritmo (Metodo degli Invarianti)

Input: *Un allineamento di dati di n variabili da un alfabeto Σ con q stati.*

Output: *Un albero binario con n foglie.*

Algoritmo (Metodo degli Invarianti)

Input: *Un allineamento di dati di n variabili da un alfabeto Σ con q stati.*

Output: *Un albero binario con n foglie.*

Passo 1: *Calcolare le probabilità empiriche $\hat{p}_{i_1 \ldots i_n}$ e scriverle in un tensore $\hat{p}$.*

Algoritmo (Metodo degli Invarianti)

Input: *Un allineamento di dati di n variabili da un alfabeto Σ con q stati.*

Output: *Un albero binario con n foglie.*

Passo 1: *Calcolare le probabilità empiriche $\hat{p}_{i_1 \dots i_n}$ e scriverle in un tensore $\hat{p}$.*

Passo 2: *Per ogni T_i topologia di alberi binari con n foglie*

- *determinare $\mathcal{F}(T_i)$ insieme di invarianti filogenetici.*
- *Calcolare*

$$t_i = \sum_{f \in \mathcal{F}(T_i)} |f(\hat{p})|.$$

Algoritmo (Metodo degli Invarianti)

Input: *Un allineamento di dati di n variabili da un alfabeto Σ con q stati.*

Output: *Un albero binario con n foglie.*

Passo 1: *Calcolare le probabilità empiriche $\hat{p}_{i_1 \dots i_n}$ e scriverle in un tensore $\hat{p}$.*

Passo 2: *Per ogni T_i topologia di alberi binari con n foglie*

- *determinare $\mathcal{F}(T_i)$ insieme di invarianti filogenetici.*
- *Calcolare*

$$t_i = \sum_{f \in \mathcal{F}(T_i)} |f(\hat{p})|.$$

Passo 3: *Scegliere T_i corrispondente al minimo t_i.*

Problemi del metodo:

Problemi del metodo:

- ♯ di alberi: gli alberi binari con n foglie da confrontare sono $(2n-5)!!$

Problemi del metodo:

- $\sharp$ di alberi: gli alberi binari con n foglie da confrontare sono $(2n-5)!!$
  $\longrightarrow$ per n grande il metodo è inutilizzabile;

Problemi del metodo:

- $\sharp$ di alberi: gli alberi binari con n foglie da confrontare sono $(2n-5)!!$ $\longrightarrow$ per n grande il metodo è inutilizzabile;
- comportamento degli invarianti: quali sono necessari per la ricostruzione? Considerarli tutti può disturbare l'identificazione?

Un approccio alternativo è stato proposto da N. Eriksson: invece di confrontare tutti gli alberi, cerchiamo di ricavare l'albero corretto rintracciando le coppie di specie più simili.

Un approccio alternativo è stato proposto da N. Eriksson: invece di confrontare tutti gli alberi, cerchiamo di ricavare l'albero corretto rintracciando le coppie di specie più simili.

Proposizione

- *Se* (A, B) *è uno split,* $\mathrm{rk}(\mathrm{Flat}_A(p)) \leq q$;
- *Se* (A, B) *non è uno split,* $\mathrm{rk}(\mathrm{Flat}_A(p)) \geq q^2$.

Se troviamo che una partizione del tipo $\{i,j\},[n]\setminus\{i,j\}$ è uno split, possiamo dire che le foglie i e j sono unite allo stesso vertice interno nell'albero.

Se troviamo che una partizione del tipo $\{i,j\},[n]\setminus\{i,j\}$ è uno split, possiamo dire che le foglie i e j sono unite allo stesso vertice interno nell'albero.

Idea: consideriamo i flattening di questo tipo, e uniamo le coppie (i,j) per le quali $\mathrm{rk}(Flat_{\{i,j\}}(\hat{p})) \leq q$.

Se troviamo che una partizione del tipo $\{i,j\},[n]\setminus\{i,j\}$ è uno split, possiamo dire che le foglie i e j sono unite allo stesso vertice interno nell'albero.

Idea: consideriamo i flattening di questo tipo, e uniamo le coppie (i,j) per le quali $\mathrm{rk}(Flat_{\{i,j\}}(\hat{p})) \leq q$.

Problema: su dati reali, il rango è quasi sempre massimo.

Se troviamo che una partizione del tipo $\{i,j\},[n]\setminus\{i,j\}$ è uno split, possiamo dire che le foglie i e j sono unite allo stesso vertice interno nell'albero.

Idea: consideriamo i flattening di questo tipo, e uniamo le coppie (i,j) per le quali $\mathrm{rk}(Flat_{\{i,j\}}(\hat{p})) \leq q$.

Problema: su dati reali, il rango è quasi sempre massimo.

Soluzione: consideriamo le coppie (i,j) per le quali $(Flat_{\{i,j\}}(\hat{p}))$ è più vicino all'insieme delle matrici di rango $\leq q$.

Se troviamo che una partizione del tipo $\{i,j\},[n]\setminus\{i,j\}$ è uno split, possiamo dire che le foglie i e j sono unite allo stesso vertice interno nell'albero.

Idea: consideriamo i flattening di questo tipo, e uniamo le coppie (i,j) per le quali $\text{rk}(Flat_{\{i,j\}}(\hat{p})) \leq q$.

Problema: su dati reali, il rango è quasi sempre massimo.

Soluzione: consideriamo le coppie (i,j) per le quali $(Flat_{\{i,j\}}(\hat{p}))$ è più vicino all'insieme delle matrici di rango $\leq q$.

Teorema

$$\min_{\text{rk}(B)\leq q} ||A-B||_F = \sqrt{\sum_{i\geq q+1} \sigma_i^2},$$

dove σ_i sono i valori singolari di A ottenuti con la fattorizzazione SVD.

Algoritmo (Costruzione degli alberi tramite SVD, N.Eriksson)

Input: *Un allineamento di dati genomici da n specie da un alfabeto Σ con q stati.*

Output: *Un albero binario con n foglie.*

Algoritmo (Costruzione degli alberi tramite SVD, N.Eriksson)

Input: *Un allineamento di dati genomici da n specie da un alfabeto Σ con q stati.*

Output: *Un albero binario con n foglie.*

Passo 1: *Calcolare le probabilità empiriche $\hat{p}_{i_1 \dots i_n}$ e scriverle in un tensore $\hat{p}$.*

Algoritmo (Costruzione degli alberi tramite SVD, N.Eriksson)

Input: *Un allineamento di dati genomici da n specie da un alfabeto Σ con q stati.*

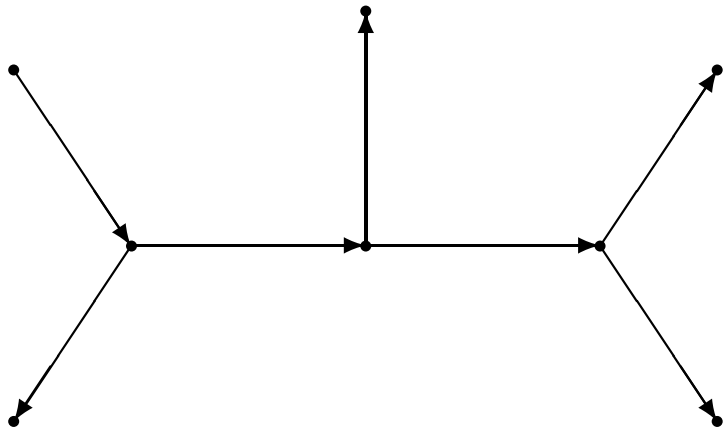
Output: *Un albero binario con n foglie.*

Passo 1: *Calcolare le probabilità empiriche $\hat{p}_{i_1 \dots i_n}$ e scriverle in un tensore $\hat{p}$.*

Passo 2: *Per $k = n \searrow 4$ effettuare le seguenti operazioni:*

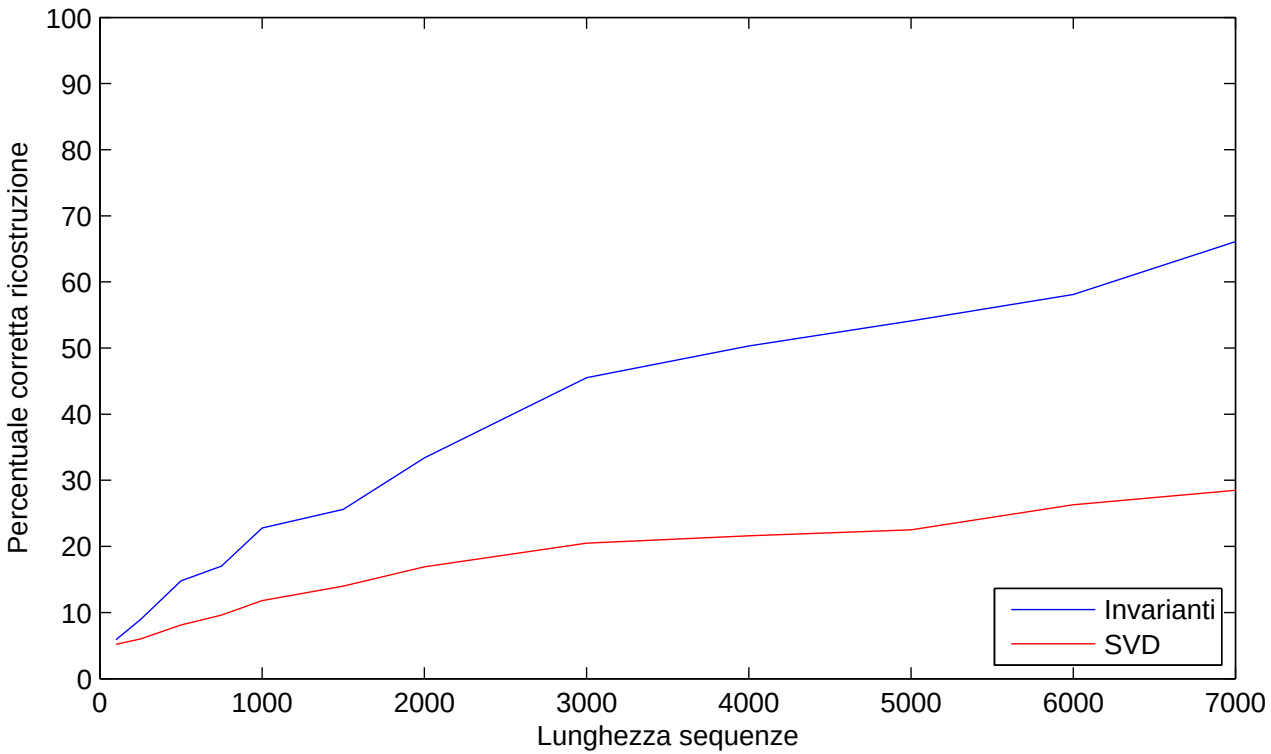
- *Per ognuna delle $\binom{k}{2}$ coppie di specie (i, j) scrivere $Flat_{\{i,j\}}(\hat{p})$ e trovarne la fattorizzazione SVD.*
- *Scegliere la coppia rispetto alla quale $\sqrt{\sum_{i \geq q+1} \sigma_i^2}$ sia minimo e unirla ad un unico vertice dell'albero. Considerare nei successivi passi queste due variabili come un'unica a valori in $[m_i] \times [m_j]$ con m_i e m_j stati rispettivamente di X_i e X_j.*

Per confrontare i due metodi abbiamo simulato con MATLAB dati provenienti dal modello definito da questo albero con variabili binarie.



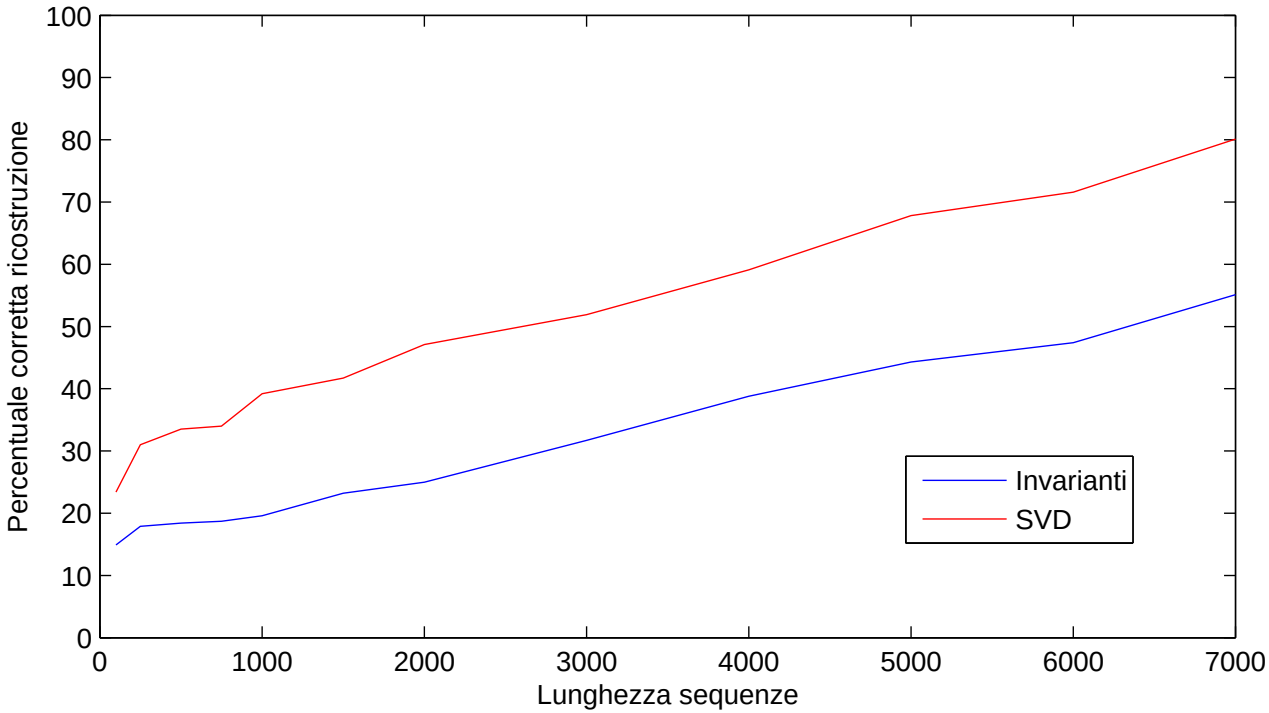
Albero filogenetico usato nella simulazione.

Abbiamo effettuato 1000 simulazioni per 3 volte, ognuna con diversi parametri generati casualmente.



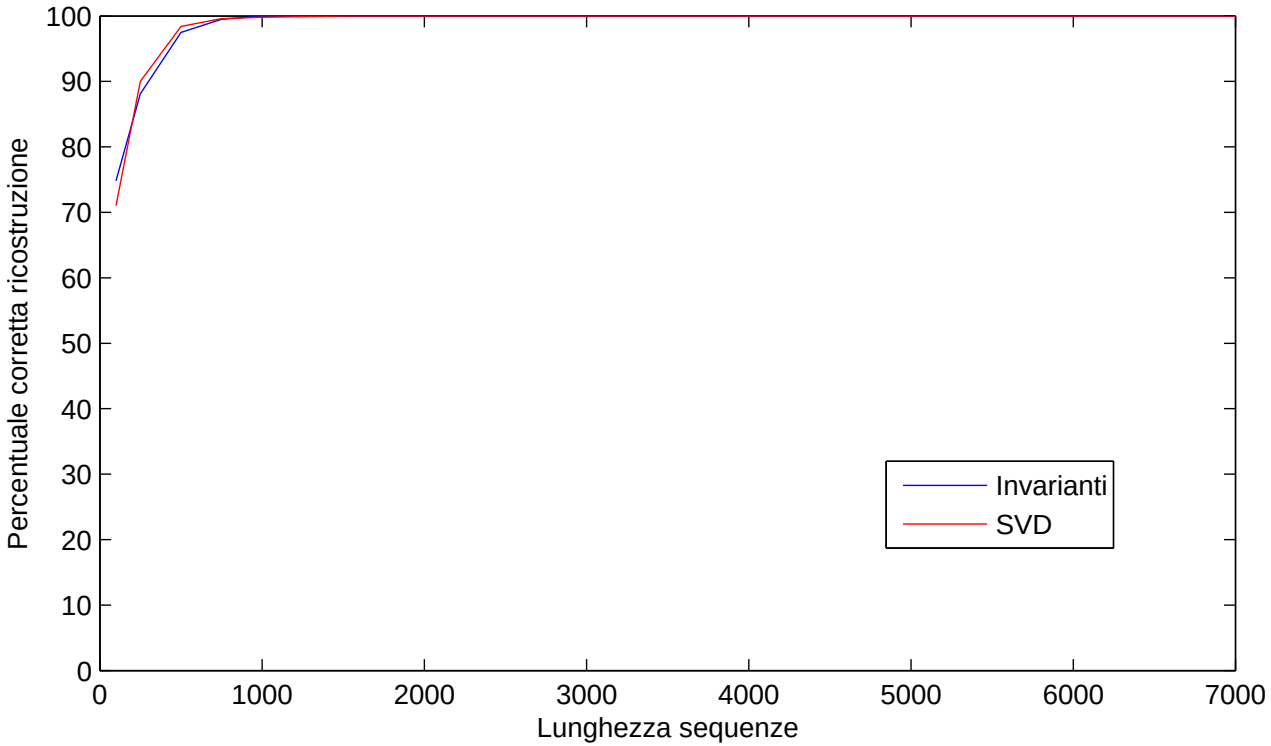
Confronto fra i due metodi con il primo set di parametri.

In questo caso le matrici di transizione sono state generate in maniera completamente casuale.



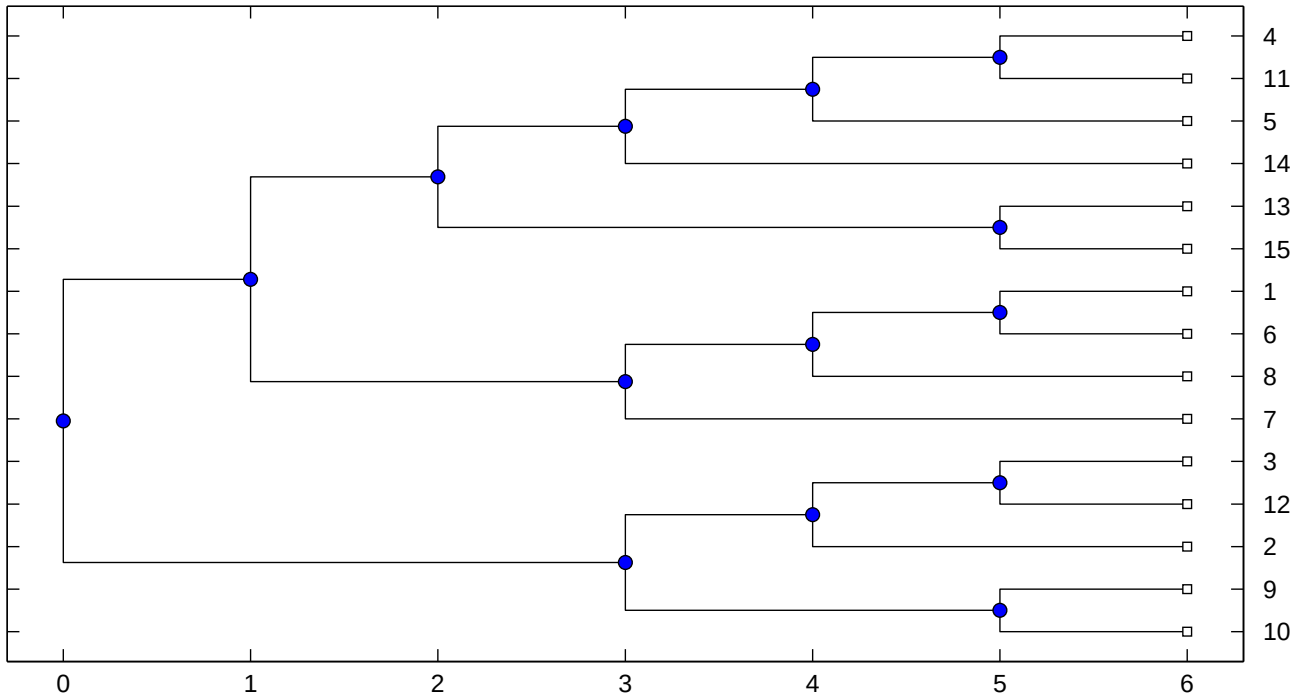
Confronto fra i due metodi con il secondo set di parametri.

In questo caso le matrici di transizione sono state generate nella forma $\frac{1}{2}M + \frac{1}{2}Id(2)$, con M matrice stocastica casuale.



Confronto fra i due metodi con il terzo set di parametri.

In questo caso le matrici di transizione sono state generate nella forma $\frac{1}{10}M + \frac{9}{10}Id(2)$, con M matrice stocastica casuale. È il caso più significativo dal punto di vista biologico.



Albero ricostruito con il metodo SVD.