Università degli Studi di Padova

DIPARTIMENTO DI MATEMATICA
Corso di Laurea in Matematica

Tesi di laurea

Inferenza Filogenetica con il metodo della Minima Evoluzione Bilanciata

Candidato:
Francesco Zaccaria
Matricola 609492

Relatore:
Ch.mo Prof. Marco Di Summa

10 Ottobre 2013

Indice

Capitolo 1

Sommario

La filogenesi molecolare è una branca della biologia che si occupa di determinare gerarchicamente le relazioni evolutive fra organismi, che nel seguito chiameremo anche taxa. Tale operazione trova applicazione in svariati ambiti quali ricerche tossicologiche, epidemiologia e dinamica delle popolazioni.

Il problema che si intende risolvere è denominato Inferenza Filogenetica. Le informazioni disponibili sono misure di diversità biologica fra gli organismi, solitamente percentuali di sequenze di DNA o RNA dissimili, estratte da esemplari di tali specie, in rapporto alla lunghezza complessiva dei codoni. Ciò che interessa determinare è il processo evolutivo che ha condotto alla diversificazione di tali taxa a partire da un, ipotetico, antenato comune.

Nel capitolo 2 vedremo in che modo si possa modellizzare tale processo. Successivamente, nel capitolo 3, entreremo nel merito del problema, elencandone le diverse istanze e motivando la preferenza per quella basata sul metodo della Minima Evoluzione. Al capitolo 4 introdurremo la variante del criterio della Minima Evoluzione detta Bilanciata, analizzando i suoi vantaggi dal punto di vista della coerenza biologica, da quello computazionale e soprattutto da quello statistico. Infine nel capitolo 5 presenteremo gli approcci risolutivi alla Minima Evoluzione Bilanciata, prima basandoci su algoritmi costruiti ad hoc e in seguito formulando l'intero problema in termini di Programmazione Lineare Intera.

Capitolo 2

Filogenie

2.1 Preliminari

Richiamiamo anzitutto alcune nozioni sui grafi non orientati.

Definizione 2.1. Un grafo (non orientato) G è una coppia ordinata $G = (V, E)$ di insiemi finiti V ed E, ove E è un insieme di coppie non ordinate di V. Gli elementi di V si chiamano *nodi* o *vertici*, mentre gli elementi di E si chiamano *archi*.

Denoteremo l'arco $e = \{u, v\}$ con uv e diremo che incide su o ha come *estremi* ciascuno dei due vertici. Si dice *grado* $d(v)$ di un vertice $v \in V$ il numero di archi incidenti su di esso.

Definizione 2.2. Due grafi $G = (V, E), G' = (V', E')$ si dicono *isomorfi* se esiste un isomorfismo di grafi fra di essi, ovvero una biiezione $\varphi : V \to V'$ tale che:

$$uv \in E \Leftrightarrow \varphi(u)\varphi(u) \in E' \qquad \forall u, v \in E$$

La relazione vista sopra è effettivamente di isomorfismo. Diremo che due grafi isomorfi sono lo stesso grafo *non etichettato*, mentre è possibile distinguere due grafi isomorfi ponendo i suoi vertici in corrispondenza biunivoca con $\{1, 2, \ldots, |V|\}$, ovvero *etichettandoli*.

Definizione 2.3. Si dice sottografo del grafo $G = (V, E)$ un grafo $G' = (V', E')$ tale che $V' \subseteq V, E' \subseteq E$. Si dice che un sottografo $G_{\setminus S}$ di G è ottenuto *rimuovendo* l'insieme di vertici $S \subseteq V$ o che è indotto dal suo complementare $V \setminus S$ se $G_{\setminus S} = G(V \setminus S, E_{V \setminus S})$, dove $E_{V \setminus S}$ è l'insieme degli archi in E con entrambe le estremità in $V \setminus S$. Si dice che un sottografo $G_{/S}$ di G è ottenuto *contraendo* l'insieme di vertici $\emptyset \neq S \subseteq V$ se $G_{/S} = G((V \setminus S) \cup \{S\}, E(V \setminus S) \cup E')$, dove $E(V \setminus S)$ è l'insieme degli archi in E con entrambe le estremità in $V \setminus S$ e E' è l'insieme degli archi di E aventi un estremo in $V \setminus S$ e l'altro in S.

Proposizione 2.1. *In ogni grafo $G = (V, E)$ vale*

$$\sum_{v \in V} d(v) = 2|E|. \tag{2.1}$$

Dimostrazione. Segue dal fatto che ogni arco incide su esattamente due nodi. □

Un *percorso* $P_{v_1 v_k}$ fra due nodi v_1 e v_k di un grafo è una sequenza alternata di nodi e archi $v_1, e_1, v_2, ..., v_{k-1}, e_{k-1}, v_k$ nella quale ogni arco ha come estremi il nodo che lo precede e quello che lo segue; i nodi v_1 e v_k sono detti *estremi* del percorso . Se nel percorso non vi sono ripetizioni di nodi esso si dice un *cammino*, di *lunghezza* uguale al numero dei suoi archi, se invece gli unici nodi uguali sono gli estremi si chiama *ciclo*. Due nodi distinti si dicono *connessi* se sono estremi di un cammino.

Definizione 2.4. Si dice *albero* un grafo che non contiene cicli e in cui ciascuna coppia di nodi è connessa. I vertici di grado uno di un albero sono detti foglie, gli altri sono detti vertici interni.

Gli archi di un albero sono anche chiamati *rami*. In un albero i rami incidenti su una foglia si dicono *esterni*, tutti gli altri sono i rami *interni*.

Osservazione 1. Dalla definizione segue che ogni coppia di nodi di un albero è collegata da uno, per la connessione, ed uno solo, per l'aciclicità, cammino.

Lemma 2.2. *Ogni albero T con almeno due vertici ha almeno due foglie.*

Proposizione 2.3. *Per ogni albero vale $|E| = |V| - 1$.*

Dimostrazione. Si dimostra per induzione sul numero dei vertici n. È banalmente vero per $n = 1$. Sia poi $n > 1$ e si consideri un albero T con n vertici; esso ha almeno una foglia v per il Lemma 2.2. Allora il suo sottografo $V_{\setminus v}$ ottenuto rimuovendo $\{v\}$, e di conseguenza un arco, ha $(n-1) - 1 = n - 2$ archi per ipotesi induttiva, da cui segue la tesi. □

2.2 Definizione e proprietà

I processi evolutivi che conducono alla diversificazione dei taxa sono modellizzati da un particolare tipo di albero, detto *filogenia*.

Definizione 2.5. Una *filogenia* (o *albero binario*) è un albero i cui vertici interni hanno tutti grado 3.

Dalla (2.1) combinata con la (2.3) segue subito che una filogenia con n taxa ha esattamente $n - 2$ vertici interni, in quanto:

$$2(n + a - 1) = 1n + 3a \Leftrightarrow a = n - 2$$

e dunque, sempre per la (2.3), $(n + (n-2)) - 1 = 2n - 3$ archi. Si noti che, essendo un albero con più di $n > 2$ foglie per garantire l'esistenza di vertici interni, una filogenia non può contenere rami fra le foglie, da cui l'insieme di tutti i suoi possibili archi ha cardinalità:

$$|E| = \frac{(n-2)(n-3)}{2} + n(n-2) = \frac{3}{2}(n-1)(n-2).$$

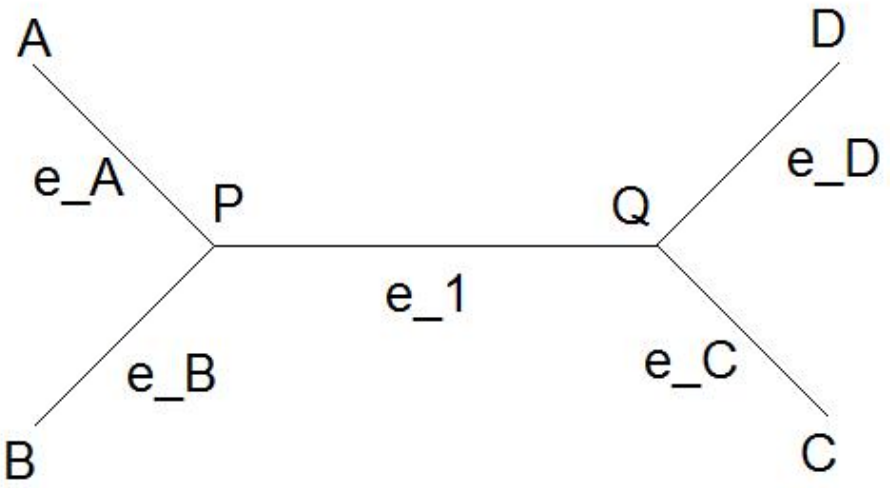

Figura 2.1: Un esempio di filogenia con quattro foglie.

Una variante degli alberi binari è la seguente

Definizione 2.6. Una filogenia si dice *radicata* o *con radice* se ha uno ed un solo vertice interno di grado 2, che viene chiamato appunto *radice*.

Ad ogni foglia di una filogenia viene assegnato uno specifico taxon, a rappresentare una specie esistente, mentre i vertici interni non vengono etichettati, poiché rappresentano gli antenati comuni, ricavabili solo a posteriori dal problema dell'Inferenza Filogenetica. L'insieme dei taxa di una filogenia verrà denotato con Γ. Inoltre nei modelli che vedremo tutte le filogenie saranno grafi *pesati*, ovvero ad ogni arco sarà assegnato un numero reale (e in particolare non negativo) che sia misura di diversità biologica fra gli organismi associati ai suoi estremi. Assumendo (in realtà forzatamente) che le mutazioni genetiche si verifichino con una certa regolarità, i pesi possono anche essere interpretati come periodi di tempo intercorsi fra la nascita di una specie e quella successiva. Un esempio di assegnazione di taxa ad una filogenia pesata è illustrato in Figura 2.2.

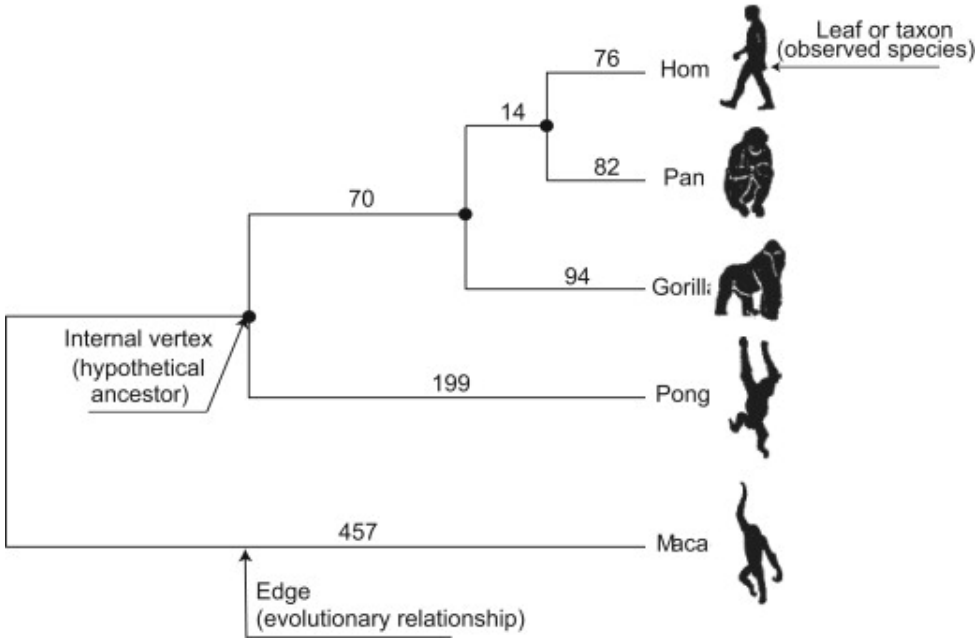

Figura 2.2: Una filogenia rappresentante l'evoluzione delle specie più vicine all'uomo.

Definizione 2.7. Un filogenia si dice albero *metrico* se i pesi sui suoi rami sono tutti non-negativi. Si dice poi *ultrametrico* se può essere scelto un vertice tale che le lunghezze dei cammini che lo congiungono alle foglie siano tutte uguali.

L'ipotesi sul grado dei vertici interni di una filogenia non ha significato biologico, è soltanto una condizione aggiuntiva, imposta per semplificare la formalizzazione del problema soprattutto nel caso del metodo della Minima Evoluzione.

Essa permette inoltre di definire su tali alberi un ordine con un vincolo sul numero di successori [2]. Difatti rimuovendo dalla filogenia avente n taxa un nodo interno, che assumiamo essere il progenitore comune dei taxa, si ottengono tre alberi binari con radice in un vertice adiacente a tale antenato. Trattandosi di alberi esiste un solo cammino dalla radice ad ogni altro vertice v, e la sua lunghezza è il *livello* di v, indicato con $level(v)$. In un tale albero ogni nodo diverso dalla radice è adiacente ad un unico vertice v' del livello precedente, altrimenti se per assurdo ve ne fosse un altro v'' si potrebbero formare cicli giustapponendo i cammini da v'' alla radice, dalla radice a v',$v'v$ e vv'' , in contraddizione con la definizione di albero. Dunque si può definire un ordinamento parziale su V come segue:

$$v \prec v' \Leftrightarrow level(v) < level(v'), \tag{2.2}$$

dove si può interpretare v come "genitore" e v' come "figlio". Grazie al vincolo sui gradi dei vertici interni, si ha che in una filogenia **ogni "genitore" ha sempre due "figli"**.

D'altro canto il vincolo sui gradi può essere imposto senza perdita di generalità a partire da un generico albero pesato, ottenendo una filogenia equivalente mediante l'aggiunta di vertici interni e di archi di peso 0 [1], come illustrato in Figura 2.3.

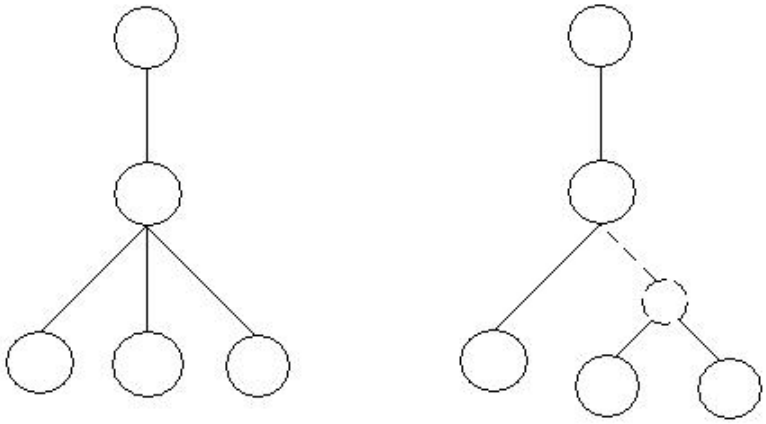

Figura 2.3: Un albero generico (sulla sinistra) viene trasformato in una filogenia aggiungendo un nodo ausiliario ed un arco di peso 0 (tratteggiati, sulla destra).

L'interpretazione biologica di questa operazione è la scomposizione delle multiforcazioni degli alberi genealogici in biforcazioni separate da un intervallo temporale nullo o comunque ridotto; ciò segue dall'assunzione che l'evoluzione procede per biforcazioni, che avvengono quando da una specie, che continua ad esistere, se ne sviluppa una nuova.

Nel seguito utilizzeremo la rappresentazione matriciale delle filogenie, necessaria per la formalizzazione della stima ai minimi quadrati dei pesi sui rami. Ad ogni filogenia T rimane associata una *matrice di incidenza archi-cammini*, cioè una matrice $\mathbf{X}$ avente una riga per ogni cammino P_{ij} fra due taxa, unico poiché una filogenia è un albero, ed una colonna per ogni arco e, aventi entrate:

$$x_{ij,e} := \begin{cases} 1 & \text{se } e \in P_{ij} \\ 0 & \text{altrimenti} \end{cases}$$

	e_A	e_B	e_C	e_D	e_1
AB	1	1	0	0	0
AC	1	0	1	0	1
AD	1	0	0	1	1
BC	0	1	1	0	1
BD	0	1	0	1	1
CD	0	0	1	1	0

Tabella 2.1: Matrice di incidenza archi-cammini, relativa alla filogenia in Figura 2.1.

Un esempio di matrice di incidenza archi-cammini, relativa alla filogenia in Figura 2.1, è quella in Tabella 2.1. Si noti che esiste una corrispondenza biunivoca fra tali matrici di incidenza e le filogenie. Infatti in un albero binario con radice esiste sempre un vertice interno g 'genitore' (nel senso specificato precedentemente) di due taxa v, v', altrimenti esisterebbe un vertice interno nel penultimo livello adiacente solo al suo "genitore" e ad una foglia, dunque di grado 2. Costruendo il cammino di lunghezza 2 vg, gv' e contraendo i tre vertici si ottiene un nuovo albero binario, e iterando il procedimento è possibile ottenere una filogenia ben definita.

Più avanti, al capitolo 5, saremo interessati alla formulazione in termini di Programmazione Lineare Intera del problema dell'Inferenza Filogenetica. A tale scopo risulta utile rappresentare con l'algebra lineare non solo gli alberi soluzione, ma anche i dati iniziali. Indicando allora con d_{ij} la *distanza evolutiva* fra due taxa i e j, ovvero la loro misura di diversità biologica, chiamiamo *matrice delle distanze* la matrice $\mathbf{D} = (d_{ij})_{1 \leq i,j \leq n}$, dove n è il numero totale di specie in esame.

Definiamo ora alcune proprietà delle matrici delle distanze [3], che utilizzeremo nel corso della trattazione.

Definizione 2.8. Una matrice delle distanze $\mathbf{D} = (d_{ij})_{1 \leq i,j \leq n}$ si dice:

matrice di dissimilarità se valgono:

$$d_{ij} > 0,\ d_{ij} = d_{j,i},\ d_{ii} = 0 \qquad \forall i \neq j \in \Gamma; \tag{2.3}$$

matrice metrica se è una matrice di dissimilarità e vale la *disuguaglianza triangolare*:

$$d_{ij} \leq d_{ik} + d_{kj} \qquad \forall i, j, k \in \Gamma; \tag{2.4}$$

matrice additiva se esiste un albero metrico tale che la somma dei pesi sugli archi del cammino che collega i taxa i e j sia uguale a d_{ij} per ogni $i, j \in \Gamma$;

matrice ultrametrica se soddisfa la condizione delle matrici additive, ma con un albero ultrametrico.

L'additività di una matrice delle distanze traduce l'interpretazione delle misure di diversità come somme dei pesi di una filogenia nei cammini fra le foglie, mentre la proprietà ultrametrica segue dall'assunzione che i taxa osservati siano tutte specie *esistenti*, e dunque equidistanti nel tempo dall'antenato comune nella radice.

Le proprietà additiva e ultrametrica possono essere sinteticamente caratterizzate in termini di disuguaglianze [3] dalle Proposizioni seguenti.

Proposizione 2.4. *Una matrice* $\mathbf{D} = (d_{ij})_{1 \leq i,j \leq n}$ *è additiva se e solo se soddisfa la* condizione dei quattro punti:

$$d_{zi} + d_{kj} \leq max\{d_{zj} + d_{ik}, d_{kz} + d_{ij}\} \qquad \forall i, j, k, z \in \Gamma \tag{2.5}$$

o equivalentemente

$$d_{zi} + d_{kj} \leq d_{zj} + d_{ik} = d_{kz} + d_{ij} \qquad \exists i, j, k, z \in \Gamma.$$

Dimostrazione. ($\Rightarrow$) Si consideri l'albero metrico associato alla matrice additiva, quattro suoi taxa qualsiasi i, j, k, z ed il sottoalbero indotto dai cammini fra tali foglie. Tale sottoalbero può essere rappresentato senza perdita di generalità come in Figura 2.1, dove i cammini pesati sono simboleggiati da singoli archi. Ordinando i taxa in modo tale che $A = i, B = z, C = j, D = k$ si ha allora:

$$d_{zi}+d_{kj} = e_A+e_B+e_C+e_D \leq e_A+e_B+e_C+e_D+2e_1 = d_{ik}+d_{zj} = d_{kz}+d_{ij}$$

dove la disuguaglianza segue dal fatto che $e_1 \geq 0$ poiché l'albero è metrico.

($\Leftarrow$) Si dimostra per induzione in modo costruttivo, vedi [4].

□

Proposizione 2.5. *Una matrice* $\mathbf{D} = (d_{ij})_{1 \leq i,j \leq n}$ *è ultrametrica se e solo se soddisfa la* condizione dei tre punti:

$$d_{ij} \leq max\{d_{ik}, d_{kj}\} \qquad \forall i, j, k \in \Gamma \tag{2.6}$$

o equivalentemente

$$d_{ij} \leq d_{jk} = d_{ik} \qquad \exists i, j, k \in \Gamma.$$

Dimostrazione. ($\Rightarrow$) Si consideri l'albero ultrametrico associato alla matrice additiva, tre suoi taxa qualsiasi i, j, k ed il sottoalbero indotto dai cammini fra tali foglie. Tale sottoalbero può essere rappresentato senza perdita di generalità come in Figura 2.4, dove i cammini pesati sono simboleggiati da singoli archi. Scegliendo un vertice

come radice si può poi costruire l'ordinamento visto sopra, e ordinando i taxa in modo che l'antenato comune a di i e j sia di livello inferiore rispetto a quello b di tutta la terna si ha:

$$d_{ij} = d_{ia} + d_{aj} \leq d_{ib} + d_{bk} = d_{ik} = d_{jk}$$

dove la disuguaglianza e l'ultima uguaglianza seguono dal fatto che l'albero è ultrametrico.

($\Leftarrow$) Come per la Proposizione precedente si dimostra per induzione in modo costruttivo, vedi [4].

□

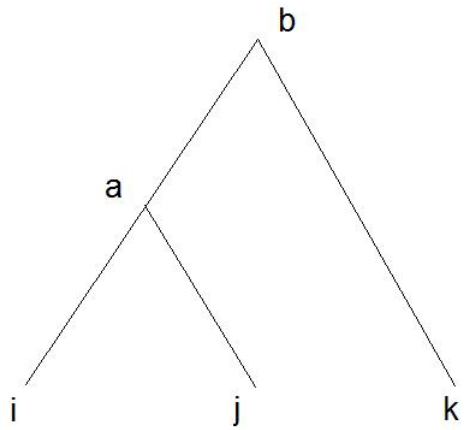

Figura 2.4: Sottoalbero generato dai cammini fra i taxa i, j, k di un albero ultrametrico.

Si noti che le equivalenze in (2.5),(2.6) si ottengono confrontando le disuguaglianze a sinistra per tutte le 3 possibili disposizioni una generica quadrupla, per (2.5), o terna, per (2.6), di indici.

Capitolo 3

Metodi di inferenza

La soluzione del problema di Inferenza Filogenetica è costituita da un processo evolutivo che si sviluppa nel corso di migliaia o anche milioni di anni, il quale non è dunque empiricamente osservabile. La filogenia che modellizza questo processo non è perciò determinabile sperimentalmente nella maggior parte dei casi.

Per tale ragione la selezione di un albero binario attendibile segue un metodo teorico, basato su diversi criteri forniti a priori [1]. Tali criteri possono solitamente essere espressi in termini di problemi di ottimizzazione su matrici di incidenza archi-cammini. Fissato un determinato metodo la soluzione del corrispondente problema di ottimizzazione è detta filogenia *ottima*.

I motivi che inducono a preferire un determinato metodo a discapito degli altri sono molteplici. Anzitutto è determinante la complessità computazionale dei problemi di ottimizzazione corrispondenti; importa inoltre che la soluzione fornita sia biologicamente sensata anche quando i dati forniti in input, ovvero le distanze evolutive fra le specie, siano perturbati da errori sperimentali. Più specificatamente, vedremo a quali vincoli devono sottostare le matrici di distanza nei vari metodi affinché i pesi della filogenia ottima siano non negativi, preferendo le condizioni meno stringenti.

Infine un'ultima proprietà che terremo in considerazione nell'analisi dei vari criteri sarà la *consistenza statistica*. Le misure di diversità che solitamente si hanno a disposizione sono infatti incomplete, relative cioè a sequenze di DNA o RNA troncate.

Definizione 3.1. Nel seguito chiameremo la filogenia ottima del problema completo filogenia *vera*.

Definizione 3.2. Si ha consistenza statistica in un fissato metodo quando, al tendere, nella topologia indotta da una delle norme matriciali equivalenti di $\mathbb{R}^{n\times n}$, delle matrici di distanza delle sequenze troncate a quella ottenuta nel caso completo le corrispondenti filogenie ottime tendono alla filogenia vera. Tale limite è inteso per la prima parte della

successione nella topologia $\{0,1\}^{\frac{3}{2}(n-1)(n-2)}$ di tutti gli alberi binari non pesati su n taxa (vedi al Capitolo 2), la quale essendo discreta implica che le filogenie ottime raggiungano definitivamente la struttura topologica di quella vera. Nella parte finale invece si adotta la topologia indotta da una delle norme equivalenti di $\mathbb{R}^{2n-3}$, ovvero lo spazio dei pesi sugli archi degli alberi binari con n taxa.

Osservazione 2. Si assuma che le stime dei pesi, **nota la configurazione non pesata del grafo**, siano lineari rispetto alla matrice delle distanze e si noti che gli spazi topologici introdotti sopra sono generati da spazi normati di dimensione finita: si ha allora che tali funzioni sono continue rispetto a $\mathbf{D} = (d_{ij})_{1 \leq i,j \leq n}$.

Se la configurazione degli archi della filogenia limite di quelle ottime soddisfa il problema **con le distanze complete in input** ciò è sufficiente a concludere la consistenza statistica. Infatti al tendere delle matrici di distanza delle sequenze troncate a quella ottenuta nel caso completo le corrispondenti filogenie ottime tendono alla filogenia vera per la continuità delle stime dei pesi. In sostanza se i pesi sono stimati linearmente la parte finale del limite è sempre verificata, mentre rimane da controllare la parte nella topologia discreta dei grafi.

Dunque per dimostrare la consistenza statistica è sufficiente dimostrare che la configurazione degli archi della filogenia limite di quelle ottime è ottima anche per il problema con le distanze complete in input. La condizione è poi anche necessaria per la definizione data di filogenia vera.

Esistono tre classi principali di criteri [1]:

criterio di parsimonia : questo criterio afferma che tra varie possibili spiegazioni di un fenomeno osservato è da preferirsi quella che richiede il minor numero di assunzioni. Di conseguenza viene selezionata come soluzione ottima quella la cui somma dei pesi sui cammini fra tutte le coppie di foglie è minima. Trovare la filogenia più parsimoniosa per un insieme di taxa è però un problema che è stato dimostrato essere $\mathcal{NP}$-difficile, per cui non sono noti algoritmi polinomiali, e un ulteriore inconveniente è rappresentato dal fatto che, in alcune circostanze, tale criterio si è rivelato statisticamente non consistente [1];

criteri di verosimiglianza : questi criteri affermano che tra varie possibili spiegazioni di un fenomeno osservato è da preferirsi quella con la più alta probabilità di verificarsi, che viene stimata in diversi modi [1]. Di conseguenza viene selezionata come soluzione ottima quella con la maggiore probabilità, definita nei vari modi, di giustificare le specie osservate. Anche trovare la filogenia più verosimile per un insieme di taxa è però un problema che è stato

dimostrato essere $\mathcal{NP}$-difficile, sebbene abbia il vantaggio, rispetto al criterio di parsimonia, di essere statisticamente consistente [1];

criteri basati sulla distanza : tali criteri mirano a trovare la filogenia che meglio si adatta alla matrice delle distanze fornita in input. A seconda della definizione di "adatta" si possono implementare diversi criteri basati sulla distanza.

Di seguito vengono esposti tre importanti criteri basati sulla distanza ed il loro comportamento al tendere dei dati iniziali ai valori veri. Il principale aspetto che induce a preferire il metodo della Minima Evoluzione è che si può dimostrare **in generale** la sua consistenza statistica. Inoltre, prendendo in considerazione un caso particolare e determinati parametri, i primi due criteri presentano problemi dei quali invece il terzo è dimostrato, con simulazioni al computer (vedi [5]), non soffrire. Gli altri motivi, ovvero la minore complessità computazionale e la coerenza biologica di tale metodo, verranno investigati più avanti, nella trattazione del caso Bilanciato al Capitolo 4.

3.1 Metodo dei Minimi Quadrati Ordinari

Il metodo dei Minimi Quadrati Ordinari seleziona come filogenia pesata ottima fra tutte quelle con n taxa quella A avente la minima *somma dei residui quadratici* SS_A, ovvero:

$$SS_A := \sum_{i<j} (d_{ij} - \mathbf{A}_A^{ij}\hat{\mathbf{b}})^2 = (\mathbf{d} - \mathbf{A}_A\hat{\mathbf{b}})^t(\mathbf{d} - \mathbf{A}_A\hat{\mathbf{b}}) \tag{3.1}$$

dove d_{ij} sono le entrate della matrice delle distanze al variare di i e di j (le quali possono essere disposte nel vettore $\mathbf{d}$ in ordine lessicografico), $\mathbf{A}_A^{ij}$ è la riga della matrice di incidenza archi-cammini relativa al cammino fra i taxa i e j nella filogenia A e $\hat{\mathbf{b}}$ è il vettore incognita dei pesi dell'albero.

Osservazione 3. Si noti che nell'equazione (3.1) **si assume nota la matrice di incidenza** della filogenia soluzione.

Dalla precedente osservazione segue che nella formulazione in esame del metodo dei Minimi Quadrati Ordinari la topologia della filogenia A si considera già nota. Dunque quello che si vuole risolvere è in realtà il sotto-problema di determinare i pesi $\hat{\mathbf{b}}$ dell'albero binario ottimo conoscendone già la struttura. Sceglieremo una particolare topologia del grafo nell'analisi di questo metodo e di quello dei Minimi Quadrati Generalizzati poiché siamo interessati principalmente ad evidenziare gli svantaggi di questi criteri rispetto a quello della Minima Evoluzione, per il quale invece verrà considerato anche il primo sotto-problema.

Le misure di diversità **nel caso completo** d_{ij} possono essere considerate come somme dei pesi incogniti sui cammini (noti perché lo è la

topologia) della filogenia **vera**:

$$d_{ij} = \sum_{e \in E} (A_A)_e^{ij} \hat{b}_e + e_{ij} \qquad \forall i, j \in \Gamma \tag{3.2}$$

dove gli e_{ij} sono gli errori di misurazione delle d_{ij}, componenti del vettore $\mathbf{e}$ in ordine lessicografico. Nel modello che trattiamo assumiamo che **gli** e_{ij} **sono stimatori corretti** delle misure complete, dunque li porremo come variabili aleatorie distribuite con media 0 (la varianza verrà discussa più avanti). In forma matriciale si scriverà allora:

$$\mathbf{d} = \mathbf{A}_A \hat{\mathbf{b}} + \mathbf{e} \tag{3.3}$$

Trascurando gli errori di misurazione, il metodo dei Minimi Quadrati Ordinari è allora equivalente a cercare il vettore $\hat{\mathbf{b}}$ che meglio risolve l'equazione (3.3) omogeneizzata. Non si vuole necessariamente una soluzione esatta poiché questa può anche non esistere se la matrice delle distanze $(d_{ij})_{i,j}$ non è additiva; quello che si cerca è invece il vettore mandato dall'applicazione lineare φ_A (quelle associata alla matrice A) nell'elemento dell'immagine $Im_{\varphi_A}(\mathbb{R}^{2n-3})$ più vicino a $\mathbf{d}$ nella norma euclidea. Questa è esattamente l' *approssimazione ai minimi quadrati* descritta nell'Appendice A, dalla quale otteniamo la stima:

$$\hat{\mathbf{b}} = (\mathbf{A}_A^t \mathbf{A}_A)^{-1} \mathbf{A}_A^t \mathbf{d}. \tag{3.4}$$

3.1.1 Consistenza statistica del metodo dei Minimi Quadrati Ordinari

Il calcolo dell'espressione (3.4) ha una complessità computazionale $O(n^4)$ nella taglia dei taxa [1], dunque pur essendo polinomiale richiede una eccessiva capacità di calcolo per grandi numeri di specie.

Si ha poi che se la matrice delle distanze in input è additiva allora per definizione esiste una soluzione esatta di (3.3) omogeneizzata e con $\hat{\mathbf{b}} \geq 0$, data da un albero pesato addirittura metrico, ma in assenza di tale condizione la soluzione ottima può avere entrate negative. Per imporre il vincolo di non-negatività sono stati sviluppati modelli più elaborati, tutti con complessità computazionale maggiore di questo criterio [1].

Consideriamo ora il comportamento del metodo al tendere dei dati iniziali ai valori veri, seguendo il ragionamento di [5]. Anzitutto supponiamo la matrice $\mathbf{D}$ additiva ed esprimiamo le distanze complete in termini dell'albero vero come fatto in (3.2), indicando con b_e le medie dei pesi sugli archi: in questo modo il valore di SS_A su una topologia A è funzione dei soli b_e e delle variabili aleatorie e_{ij}.

Si vuole verificare dunque che la somma dei residui quadratici SS_A per la topologia limite di quelle ottime A è minima nel caso completo.

Per farlo si calcola il valore atteso $E(SS_B - SS_A)$, dove B è una topologia qualsiasi del grafo diversa da A, per valori di d_{ij} che tendono a quelli delle distanze complete. Risulta utile a tal fine determinare le varianze $V(d_{ij})$ e le covarianze $Cov(d_{ik}, d_{kl})$: in accordo col modello di Jukes-Cantor [5]:

$$V(d_{ij}) = \frac{p_{ij}(1-p_{ij})}{m(1-p_{ij}/c)^2} \ , \ Cov(d_{ik}, d_{kl}) = \frac{p_{ij,kl} - p_{ij}p_{kl}}{m(1-p_{ij}/c)(1-p_{kl}/c)} \quad (3.5)$$

dove p_{ij} è la proporzione di nucleotidi differenti fra le sequenze genetiche dei taxa i e j, $p_{ij,kl}$ è la proporzione di nucleotidi differenti fra le coppie di sequenze i-j e k-l rispetto alla loro lunghezza totale, m è il numero dei nucleotidi esaminati (che supponiamo tendere a infinito nel caso delle sequenze complete) e $c = 3/4$. Tali valori sono ottenuti in sostanza stimando, in un processo di Poisson, i tassi di sostituzione dei nucleotidi nel tempo come uguali nei percorsi fra i diversi taxa, ovvero assumendo che le mutazioni genetiche si verifichino con la stessa frequenza indipendentemente da fattori connessi alle specie, quali adattamenti all'habitat o abitudini riproduttive [6]. Sviluppando in serie di Taylor e utilizzando i dati di alcuni ordini di mammiferi, si ricavano allora le uguaglianze (3.5) e il valore di c [5].

Eseguiamo i calcoli nel caso particolare in cui le topologie A e B sono quelle in Figura 3.1. In tal caso la matrice di incidenza archi-cammini per A è:

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

dunque le distanze evolutive sono:

$$\begin{aligned} d_{12} &= b_1 + b_2 + b_3 + b_4 + b_5 + e_{12}, \\ d_{13} &= b_1 + b_3 + b_5 + e_{13}, \\ d_{14} &= b_1 + b_4 + b_5 + e_{14}, \\ d_{23} &= b_2 + b_3 + b_5 + e_{23}, \\ d_{24} &= b_2 + b_4 + b_5 + e_{24}, \\ d_{34} &= b_3 + b_4 + e_{34}. \end{aligned}$$

Utilizzando le stime dei minimi quadrati (3.4), si ricava:

$$SS_A = (d_{13} + d_{24} + d_{14} + d_{23})^2/4 = (e_{13} + e_{24} + e_{14} + e_{23})^2/4.$$

Con lo stesso procedimento abbiamo, per B:

$$SS_B = (-d_{12} - d_{34} + d_{14} + d_{23})^2/4 = (2b_5 - e_{12} - e_{34} + e_{14} + e_{23})^2/4.$$

Otteniamo infine il valore atteso della differenza:

$$\begin{aligned}E(SS_B - SS_A) = \frac{1}{4}E(&4b_5^2 + e_{12}^2 + e_{14}^2 + e_{23}^2 + e_{34}^2 - 4b_5e_{12} + 4b_5e_{14} + 4b_5e_{23} \\ &- 4b_5e_{34} - 2e_{12}e_{14} - 2e_{12}e_{23} + 2e_{12}e_{34} + 2e_{14}e_{23} - 2e_{14}e_{34} \\ &- 2e_{23}e_{34} - e_{13}^2 - e_{14}^2 - e_{23}^2 - e_{24}^2 + 2e_{13}e_{14} + 2e_{13}e_{23} - 2e_{13}e_{24} \\ &- 2e_{14}e_{23} + 2e_{14}e_{24} + 2e_{23}e_{24}) = \\ &= b_5^2 + (v_{12} + v_{34} - v_{13} - v_{24})/4 + (-c_{12,14} - c_{12,23} + c_{12,34} \\ &- c_{14,34} - c_{23,34} + c_{13,14} + c_{13,23} - c_{13,24} + c_{14,24} + c_{23,24})/2\end{aligned} \tag{3.6}$$

con $v_{ij} := V(d_{ij})$ e $c_{ij,kl} := Cov(d_{ij}, d_{kl})$. Se assumiamo che i nucleotidi presi in esame siano sufficientemente pochi in proporzione alla lunghezza delle sequenze, allora è possibile usare la matrice varianza-covarianza seguente:

$$\mathbf{V} \approx m^{-1} \begin{pmatrix} b_1 + b_2 & b_1 & b_1 & b_2 & b_2 & 0 \\ b_1 & b_1 + b_3 + b_5 & b_1 + b_3 & b_3 + b_5 & b_5 & b_3 \\ b_1 & b_1 + b_5 & b_1 + b_4 + b_5 & b_5 & b_4 + b_5 & b_4 \\ b_2 & b_3 + b_5 & b_5 & b_2 + b_3 + b_5 & b_2 + b_5 & b_3 \\ b_2 & b_5 & b_4 + b_5 & b_2 + b_5 & b_2 + b_4 + b_5 & b_4 \\ 0 & b_3 & b_4 & b_3 & b_4 & b_3 + b_4 \end{pmatrix} \tag{3.7}$$

approssimando opportunamente le stime (3.5), riscritte utilizzando i valori attesi delle $p_{ij}, p_{ij,kl}$ [5]. Si ottiene così:

$$E(SS_B - SS_A) \approx b_5^2 + b_5/m > 0. \tag{3.8}$$

Ripetendo il procedimento anche per altre topologie si ottiene lo stesso valore positivo [5], il che suggerisce che la filogenia con la topologia A abbia la minima somma degli scarti quadratici con le distanze complete e che il metodo sia statisticamente consistente per l'Osservazione 2.

Se invece l'ipotesi sul numero di nucleotidi analizzati non vale allora non si può utilizzare la matrice approssimata (3.7), dunque la differenza dei residui rimane la (3.6), che riscritta nuovamente utilizzando i valori attesi delle $p_{ij}, p_{ij,kl}$ [5] fornisce:

$$\begin{aligned}E(SS_B - SS_A) = 0.001 + [&4.51908 - 12e^{\frac{4}{3}x} - 6e^{\frac{8}{3}x} + 18e^{\frac{4}{3}(0.01+x)} + 9e^{\frac{8}{3}(0.01+x)} \\ &- 6e^{\frac{4}{3}(0.21+x)} - 3e^{\frac{8}{3}(0.21+x)}]/(64m)\end{aligned} \tag{3.9}$$

che è positivo per $m = 100$ se $x < 0.158$, ma altrimenti è negativo, come si nota in Figura 3.1; perciò in questo caso il metodo restituisce un albero che non è il limite di quelli ottimi. Naturalmente al crescere di m le varianze e covarianze (3.5) tendono a 0, dunque l'albero fornito tende a essere quello vero.

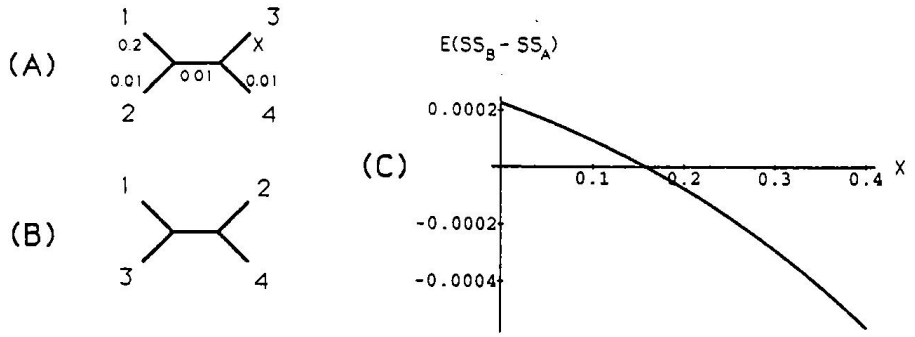

Figura 3.1: (A) Albero vero con relativi pesi sugli archi. (B) Un albero errato. (C) Grafico di (3.9) in funzione del peso x.

Riassumendo, **per questa particolare configurazione topologica** si è provato che il metodo dei Minimi Quadrati Ordinari è statisticamente consistente, nel senso che all'aumentare del numero m dei nucleotidi esaminati tende a scegliere l'albero vero, ma se le distanze evolutive sono grandi e i nucleotidi presi in esame sono pochi in proporzione alla lunghezza delle sequenze allora spesso si ottiene un albero errato e **il metodo non è corretto**.

3.2 Metodo dei Minimi Quadrati Generalizzati

Il metodo dei Minimi Quadrati Generalizzati seleziona come filogenia pesata ottima fra tutte quelle con n taxa quella A avente il valore minimo di:

$$SS_{GA} := (\mathbf{d} - \mathbf{A}_A\hat{\mathbf{b}})^t\mathbf{V}^{-1}(\mathbf{d} - \mathbf{A}_A\hat{\mathbf{b}}) \tag{3.10}$$

dove $\mathbf{V}$ è la matrice varianza-covarianza ottenuta dalle stime viste sopra (3.5) e gli altri fattori sono gli stessi di (3.1).

Esattamente come nel caso dei Minimi Quadrati Ordinari assumiamo già nota la topologia della filogenia limite A e consideriamo le misure di diversità d_{ij} come somme dei pesi incogniti sui cammini (noti perché lo è la topologia) di A, con **gli** e_{ij} **corretti** delle misure complete. Rimangono dunque valide le espressioni (3.2) e (3.3).

Conviene a questo punto riscrivere il valore (3.10) nel seguente modo:

$$SS_{GA} := (\mathbf{d} - \mathbf{A}_A\hat{\mathbf{b}})^t\mathbf{L}^t\mathbf{L}(\mathbf{d} - \mathbf{A}_A\hat{\mathbf{b}}) \tag{3.11}$$

dove $\mathbf{L}$ è una matrice tale che $\mathbf{L}^t\mathbf{L} = \mathbf{V}^{-1}$, che esiste poiché la matrice varianza-covarianza $\mathbf{V}^{-1}$ è simmetrica [7]. Trascurando poi gli errori di misurazione e definendo $\mathbf{d}' := \mathbf{Ld}, \mathbf{A}'_A := \mathbf{LA}_A$, la (3.11) si può ancora

riscrivere come:

$$SS_{GA} = (\mathbf{d} - \mathbf{A}_A\hat{\mathbf{b}})^t\mathbf{L}^t\mathbf{L}(\mathbf{d} - \mathbf{A}_A\hat{\mathbf{b}}) = (\mathbf{d}' - \mathbf{A'}_A\hat{\mathbf{b}})^t(\mathbf{d}' - \mathbf{A'}_A\hat{\mathbf{b}})$$

Ci si è allora ricondotti nuovamente all' *approssimazione ai minimi quadrati* descritta nell'appendice A, dalla quale otteniamo lo stimatore dei Minimi Quadrati Generalizzati:

$$\hat{\mathbf{b}} = (\mathbf{A'}^t_A\mathbf{A'}_A)^{-1}\mathbf{A'}^t_A\mathbf{d}' = (\mathbf{A}^t_A\mathbf{L}^t\mathbf{L}\mathbf{A}_A)^{-1}\mathbf{A}^t_A\mathbf{L}^t\mathbf{L}\mathbf{d} = (\mathbf{A}^t_A\mathbf{V}^{-1}\mathbf{A}_A)^{-1}\mathbf{A}^t_A\mathbf{V}^{-1}\mathbf{d}. \tag{3.12}$$

Il metodo appena esposto può essere modificato ponendo in (3.10) una matrice diagonale $\mathbf{\Omega} = Diag((\omega_{ij})_{1\leq i,j\leq n})$ al posto di $\mathbf{V}^{-1}$, per ottenere lo stimatore dei minimi quadrati pesati:

$$\hat{\mathbf{b}} = (\mathbf{A}^t_A\mathbf{\Omega}^{-1}\mathbf{A}_A)^{-1}\mathbf{A}^t_A\mathbf{\Omega}^{-1}\mathbf{d}. \tag{3.13}$$

Ad esempio $\omega_{ij} = \frac{1}{d^2_{ij}}$ nella variante di Fitch e Margoliash oppure $\omega_{ij} = \frac{1}{d_{ij}}$ in quella di Beyer et al. [1]. Usando simulazioni al computer si è mostrato che questi modelli sono statisticamente consistenti [5], cosa che non è in generale vero per quantità ω_{ij} qualsiasi [1].

Le modifiche al modello dei Minimi Quadrati Ordinari che portano a quelli Pesati o Generalizzati sono dovute al fatto che, a causa della storia evolutiva comune dei taxa in esame e della maggiore varianza delle distanze grandi rispetto alle altre, l'assunzione che le distanze d_{ij} siano variabili aleatorie i.i.d.(i.e. indipendenti e identicamente distribuite) non è vera in generale, e dunque è necessario introdurre le quantità ω_{ij}, per rappresentare le varianze delle d_{ij} nel primo caso, e le entrate non diagonali di $\mathbf{V}^{-1}$, per rappresentare le covarianze nel modello generalizzato [1],[12].

Nel seguito non ci occuperemo dei Minimi Quadrati Pesati, ma osserveremo al Capitolo 4 che il Metodo della Minima Evoluzione Bilanciata ne è una forma speciale [12], dunque ciò suggerisce che per le quantità ω_{ij} di tale criterio esso possa essere statisticamente consistente, come verrà in effetti dimostrato [12] allo stesso Capitolo.

3.2.1 Consistenza statistica del metodo dei Minimi Quadrati Generalizzati

Il calcolo dell'espressione (3.12) ha una complessità computazionale $O(n^6)$ nella taglia dei taxa [1], dunque pur essendo polinomiale richiede una eccessiva capacità di calcolo per grandi numeri di specie, analogamente al caso dei Minimi Quadrati Ordinari.

Sempre come per (3.3) si ha poi che se la matrice delle distanze in input è additiva allora per definizione esiste una soluzione esatta di (3.3),

che quindi minimizza (3.10), e con $\hat{\mathbf{b}} \geq 0$, data da un albero pesato addirittura metrico, ma in assenza di tale condizione la soluzione ottima può avere entrate negative. Per imporre il vincolo di non-negatività sono stati sviluppati modelli più elaborati, tutti con complessità computazionale maggiore di questo criterio [1].

Consideriamo ora il comportamento del metodo al tendere dei dati iniziali ai valori veri, seguendo di nuovo il ragionamento di [5]. Supponiamo sempre la matrice **D** additiva e gli stimatori e_{ij} **corretti**, esprimendo le distanze complete in termini dell'albero vero come fatto in (3.2).

Si vuole verificare dunque che la somma dei residui quadratici SS_A per la topologia limite di quelle ottime A è minima nel caso completo.

Calcoliamo ancora il valore atteso $E(SS_B - SS_A)$ con le stesse topologie A e B in Figura 3.1; la matrice di incidenza archi-cammini per A è la stessa vista sopra e così pure le distanze evolutive; con le nuove stime (3.12), si ottiene così [5]:

$$\begin{aligned}
&\mathrm{E}(\mathrm{SS}_G A) = 1,\\
&E(SS_G B) = \frac{4b_5^2}{v_{12}+v_{34}+v_{14}+v_{23}+2c_{12,34}-2c_{12,14}-2c_{12,23}-2c_{14,34}-2c_{23,34}+2c_{14,23}} + 1,\\
&E(SS_G B - SS_G A) = \frac{4b_5^2}{v_{12}+v_{34}+v_{14}+v_{23}+2c_{12,34}-2c_{12,14}-2c_{12,23}-2c_{14,34}-2c_{23,34}+2c_{14,23}} \geq 0.
\end{aligned}$$

Ripetendo il procedimento anche per altre topologie si ottengono ancora valori positivi [5], il che suggerisce che la filogenia con la topologia A abbia la minima somma degli scarti quadratici generalizzati con le distanze complete e che il metodo sia statisticamente consistente per l'Osservazione 2; inoltre al crescere di m le varianze e covarianze (3.5) tendono a 0, dunque l'albero fornito tende a essere quello vero.

Si è provato in sostanza che, **per questa particolare configurazione topologica**, il metodo dei Minimi Quadrati Generalizzati è statisticamente consistente. Si ha però che quando i d_{ij} tendono a 0 la matrice **V**, ottenuta utilizzando le varianze $v_{ij} = V(d_{ij})$ e le covarianze $c_{ij,kl} = Cov(d_{ik}, d_{kl})$ del modello di Jukes-Cantor (3.5), **tende a diventare singolare**: in (3.7), valida quando i nucleotidi presi in esame siano sufficientemente pochi in proporzione alla lunghezza delle sequenze, si può ottenere la quarta riga sommando la seconda con la quinta e togliendo la terza. Per valori di d_{ij} maggiori i nucleotidi analizzati sono solitamente di più, dunque (3.7) non vale e il problema non sussiste, ma in caso contrario questo inconveniente rende impossibile calcolare $\mathbf{V}^{-1}$ (producendo ad esempio errori di overflow) e quindi inapplicabile l'intero metodo.

3.3 Metodo della Minima Evoluzione

Il metodo della Minima Evoluzione seleziona come filogenia pesata ottima fra tutte quelle con n taxa quella A avente il valore minimo di:

$$S_A := \sum_{e \in E} \hat{b}_e \tag{3.14}$$

dove la notazione è la stessa di (3.1), con $\hat{\mathbf{b}}$ vettore incognita dei pesi dell'albero stimati con il criterio dei Minimi Quadrati Ordinari (quelli Generalizzati possono portare talvolta alla scelta di un albero errato [5]). Consideriamo le misure di diversità d_{ij} come somme dei pesi incogniti sui cammini (noti perché lo è la topologia) di A, con **gli** e_{ij} **corretti** delle misure complete. Rimangono dunque valide le espressioni (3.2) e (3.3).

Nell'implementazione del metodo fornita da Rzhetsky e Nei si applica il metodo dell'Unione degli Intorni (NJ, Neighbor Joining) di Saitou e Nei per determinare la topologia della filogenia ottima A [9]. La procedura è la seguente:

1. costruire un albero secondo la procedura NJ;

2. elencare tutte le filogenie la cui distanza topologica d_T dall'albero NJ è 2 o 4. Tale distanza è definita come:

   $$d_T = 2[min(q_1, q_2) - p] + |q_1 - q_2|$$

   dove q_1 e q_2 sono i numeri totali di partizioni creati dai rami interni dei due alberi e p è il numero di partizioni comuni ai due alberi. Per alberi generici q_1 e q_2 non sono in generale uguali, ma per le filogenie ciò segue dalla definizione, e dunque in tal caso d_T è uguale al doppio del numero delle partizioni differenti generate dai rami interni degli alberi; tale valore è in effetti intuitivamente una misura di distanza sulla topologia discreta dei grafi;

3. stimare la differenza $D = S - S_{NJ}$ tra i valori di S per le filogenie di cui al punto precedente e per l'albero NJ;

4. se $D >> 0$ per ogni topologia testata scegliere l'albero NJ, altrimenti considerare anche le filogenie per cui tale condizione non vale, in particolare quelle per le quali $D < 0$;

Comunque per i nostri fini assumeremo nel seguito già noto tale dato, e ci occuperemo solo del sotto-problema di determinare i pesi $\hat{\mathbf{b}}$ dell'albero binario ottimo conoscendone già la struttura.

3.3.1 Consistenza statistica del metodo della Minima Evoluzione

Un primo vantaggio del metodo della Minima Evoluzione è che si è dimostrato, con simulazioni al computer (vedi [5]), non soffrire dei problemi riscontrati sopra con i criteri dei Minimi Quadrati, prendendo in considerazione una topologia particolare e determinati parametri.

Consideriamo poi il comportamento del metodo al tendere dei dati iniziali ai valori veri. Per dimostrare la consistenza statistica (e in questo caso lo si farà **in generale**) è ancora una volta sufficiente e necessario dimostrare che la somma dei pesi per la topologia limite di quelle ottime T è minima nel caso completo, seguendo il ragionamento di [10].

Supponiamo sempre la matrice **D** additiva e gli stimatori e_{ij} **corretti**, esprimendo le distanze in termini dell'albero vero come fatto in (3.2). Si calcola dunque il valore atteso $E(S_W - S_T)$, dove W è una topologia qualsiasi del grafo diversa da T costruita su n sequenze molecolari relative ai diversi taxa, per valori di d_{ij} uguali a quelli delle distanze complete.

Utilizzando le stime dei minimi quadrati (3.4), si può dimostrare che la media della somma dei pesi dell'albero vero è esattamente uguale a:

$$E(S_T) := \sum_{e \in E} b_e \tag{3.15}$$

Ciò segue direttamente dalle ipotesi, ricordando che abbiamo qui assunto in (3.4) la matrice **D** additiva (da cui la soluzione esatta a (3.3)), gli stimatori e_{ij} corretti e i valori di d_{ij} uguali a quelli delle distanze complete in termini dei pesi dell'albero vero b_e. Una dimostrazione basata sul calcolo diretto viene comunque fornita nel capitolo successivo, nell'analisi della verità del modello Bilanciato nel caso additivo.

Per determinare ora il valore di $E(S_W)$ conviene riscrivere le stime (3.4) senza fare ricorso all'algebra matriciale: ciò ci servirà per facilitare l'esposizione, ma è utile anche per ridurre enormemente il carico di memoria che è richiesto ad un computer per eseguire i calcoli quando il numero di sequenze è grande (si ricordi che l'espressione (3.4) ha una complessità computazionale $O(n^4)$ nella taglia dei taxa [1]). Si consideri l'albero (A) in Figura 3.3 come esempio. Se scegliamo un particolare ramo interno di questa filogenia, tale albero può essere disegnato nella forma (B) della stessa figura, dove A,B,C e D rappresentano ciascuna un gruppo di sequenze. Ad esempio, per il braccio interno b in Figura 3.3(A) A,B,C e D rappresentano i gruppi (3),(1,2),(4) e (5,6,7,8) rispettivamente. In questo caso il peso di b nell'albero (B) può essere stimato con la seguente espressione:

$$\begin{aligned}\hat{b} = \frac{1}{2}\{&\gamma[d_{AC}/(n_A n_C) + d_{BD}/(n_B n_D)] \\ &+ (1-\gamma)[d_{BC}/(n_B n_C) + d_{AD}/(n_A n_D)] - d_{AB}/(n_A n_B) - d_{CD}/(n_C n_D)\}\end{aligned} \tag{3.16}$$

dove

$$\gamma := (n_B n_C + n_A n_D)/[(n_A + n_B)(n_C + n_D)]$$

n_A, n_B, n_C, n_D sono i numeri delle sequenze nei gruppi A,B,C e D rispettivamente e d_{AC} è la somma delle distanze fra i gruppi, dove una sequenza appartiene ad A e l'altra a C, con $d_{BD}, d_{BC}, d_{AD}, d_{AB}, d_{CD}$ definiti in modo analogo. L'espressione (3.16) è una riscrittura della (3.4): lo si dimostra in [10] imponendo anzitutto la condizione che $\hat{b}$ sia uno stimatore lineare nelle distanze date in input, ovvero:

$$\hat{b} = \sum_{i=1}^{K} \alpha_i d^i_{AC} + \sum_{i=1}^{L} \beta_i d^i_{BD} + \sum_{i=1}^{M} \gamma_i d^i_{AD} + \sum_{i=1}^{N} \delta_i d^i_{BC} + \sum_{i=1}^{P} \epsilon_i d^i_{AB} + \sum_{i=1}^{Q} \zeta_i d^i_{CD}$$

dove $\alpha_i, \beta_i, \gamma_i, \delta_i, \epsilon_i, \zeta_i$ sono coefficienti ignoti, d^i_{AC} è l'i-esima distanza fra due sequenze di cui una appartenente al gruppo A e l'altra al gruppo C, $d^i_{BD}, d^i_{AD}, d^i_{BC}, d^i_{AB}, d^i_{CD}$ sono definite analogamente e infine $K = n_A n_C, L = n_B n_D, M = n_A n_D, N = n_B n_C, P = n_A n_B, Q = n_C n_D$. Il motivo dei segni meno risulterà chiaro nel capitolo successivo, nell'analisi della verità del modello Bilanciato nel caso additivo. Si aggiunge poi la condizione che gli stimatori dei minimi quadrati sono quelli con varianza minima nella classe degli stimatori lineari, assumendo che tutte le varianze delle stime delle distanze evolutive siano uguali e le corrispondenti covarianze nulle, ovvero che la matrice varianza-covarianza sia:

$$\mathbf{V} = v\mathbb{I}_{1 \leq i,j \leq \frac{n(n-1)}{2}}.$$

Eseguendo i calcoli ([10]) si ottengono così i valori (3.16). La (3.16) può essere applicata anche ad un braccio esterno dell'albero (C) in Figura 3.3, per il quale si ha:

$$\hat{b} = \frac{1}{2}[d_{CA}/n_A + d_{CB}/n_B - d_{AB}/(n_A n_B)] \tag{3.17}$$

ponendo $n_C = n_D = 1, d_{AC} = d_{AD}, d_{BC} = d_{BD}, d_{CD} = 0$, i.e. considerando il caso speciale dove i gruppi C e D diventano una singola foglia C.

A questo punto possiamo esplicitare il valore di $E(S_W - S_T)$ in funzione dei pesi dell'albero vero b_i.

Anzitutto calcoliamo le medie delle stime $\hat{a}_i$ dei pesi su W:

$$E(\hat{a}_i) = \alpha_{i,1} b_1 + \alpha_{i,2} b_2 + \cdots + \alpha_{i,2n-3} b_{2n-3} \tag{3.18}$$

dove gli $\alpha_{i,j}$ sono i coefficienti dei b_j (questi ultimi sono in totale $2n-3$ perché W è una filogenia). Tali coefficienti sono univocamente definiti da (3.4):

$$\begin{aligned} \mathbf{a} &= (\mathbf{A}_W^t \mathbf{A}_W)^{-1} \mathbf{A}_W^t \mathbf{d} \\ &= (\mathbf{A}_W^t \mathbf{A}_W)^{-1} \mathbf{A}_W^t \mathbf{A}_T \mathbf{b} \\ &= (\alpha_{i,j})_{1 \leq i,j \leq 2n-3} \mathbf{b} \end{aligned} \tag{3.19}$$

dove $\mathbf{A}_W, \mathbf{A}_T$ sono le matrici di incidenza delle topologie errata e vera rispettivamente, $\mathbf{a} = (E(\hat{a}_i))_i$ e $\mathbf{b} = b_i$. Si noti che si sono potute scrivere le medie (3.18) come funzioni lineari dei pesi $\mathbf{b} = b_i$ poiché le $E(\hat{a}_i)$ sono funzioni lineari delle $E(d_{ij})$ (3.4), che a loro volta sono funzioni lineari dei b_i poiché **gli stimatori** e_{ij} **sono corretti**, i.e. hanno media 0, in (3.2).

Esprimiamo infine il valore di $E(S_W - S_T)$ in funzione dei pesi dell'albero vero b_e:

$$E(S_W - S_T) = \beta_1 b_1 + \beta_2 b_2 + \cdots + \beta_{2n-3} b_{2n-3} \tag{3.20}$$

dove

$$\beta_j = \alpha_{1,j} + \alpha_{2,j} + \cdots + \alpha_{2n-3,j} - 1. \tag{3.21}$$

Vogliamo ora dimostrare che $E(S_W - S_T) > 0$ per qualsiasi scelta di W. Questo può essere fatto mostrando che almeno un termine $\beta_k b_k$ nell'equazione (3.20) è positivo e che gli altri sono non-negativi. Anzitutto notiamo che i β_i associati ai rami esterni sono sempre 0 indipendentemente dalla topologia: infatti se i pesi di tutti gli archi interni fossero 0, allora tutti gli alberi avrebbero lo stesso valore di (3.14), in quanto le condizioni sulle distanze fra i taxa determinerebbero univocamente i pesi dei rami esterni, da cui $E(S_W - S_T) = 0$. Consideriamo allora un β_i associato ad un arco interno dell'albero vero e studiamone il valore per un albero sbagliato. Per semplificare la trattazione, in (3.3) denotiamo tutte le sequenze da una parte del ramo i nell'albero vero in bianco e le altre in grigio (si noti che rimuovendo un arco interno da un albero se ne ottengono altri due); coloriamo inoltre le sequenze dell'albero errato con lo stesso colore usato nell'albero vero. Così facendo si ottengono due diverse configurazioni di sequenze colorate nell'albero errato:

1. una configurazione congruente rispetto al ramo i se esiste un particolare arco k dell'albero errato che separa la sequenze bianche e grigie;

2. una configurazione incongruente rispetto al ramo i se ciò non avviene, e di conseguenza si formano almeno 4 gruppi monocromatici.

A questo punto notiamo che i valori β_i possono essere ricavati dagli $\alpha_{j,i}$ tramite (3.21)(dove l'indice j è legato al ramo dell'albero errato e i a quello dell'albero vero dalla forma di (3.19)), e questi ultimi possono a loro volta essere ottenuti da (3.18) ponendo $b_i = 1, b_m = 0 \forall m \neq i$. Invece di utilizzare la (3.18), che abbiamo visto derivare da (3.4), ci serviamo però delle equivalenti (3.16) dimostrate sopra, traducendo la condizione precedente con la sostituzione di 1 per ogni distanza che include nel suo cammino l'i-esimo ramo dell'albero vero(cioè fra una sequenze bianca e una grigia) e 0 per le altre (cioè distanze fra taxa dello stesso colore). Si può fare questo considerando quattro gruppi di sequenze A,B,C,D

rispetto al j-esimo ramo interno dell'albero errato e contando il numero di taxa di ogni colore in ognuna di esse, da cui si ottiene:

$$\begin{aligned}\alpha_{j,i} = \gamma &\left(\frac{W_A G_C + G_A W_C}{2n_A n_C} + \frac{W_B G_D + G_B W_D}{2n_B n_D}\right) \\ + (1-\gamma) &\left(\frac{W_B G_C + G_B W_C}{2n_B n_C} + \frac{W_A G_D + G_A W_D}{2n_A n_D}\right) \\ - &\frac{G_A W_B + G_B W_A}{2n_A n_B} - \frac{G_C W_D + G_D W_C}{2n_C n_D}\end{aligned} \tag{3.22}$$

dove $W_A, G_A; W_B, G_B; W_C, G_C; W_D, G_D$ sono i numeri di sequenze bianche e grigie nei gruppi A,B,C,D rispettivamente e il resto della notazione è lo stesso di (3.16). Analogamente se j è un ramo esterno nell'albero errato da (3.17) si ottiene:

$$\alpha_{j,i} = \begin{cases} \frac{G_A G_B}{n_A n_B} & \text{se la sequenza C è bianca,} \\ \frac{W_A W_B}{n_A n_B} & \text{se la sequenza C è grigia.} \end{cases} \tag{3.23}$$

Calcolando ora β_i per le diverse configurazioni si ha:

1. per una configurazione congruente con k ramo che separa nell'albero errato le sequenze di colore diverso, si ha:

$$\alpha_{j,i} = \begin{cases} 1 & \text{se } j = k, \\ 0 & \text{se } j \neq k. \end{cases}$$

   da cui $\beta_i = 0$;

2. per una configurazione incongruente ci si può ricondurre ad una congruente mediante le due operazioni in Figura 3.2, che consistono nello scambiare due gruppi di sequenze separate da uno o due rami interni. Si può dimostrare [10] che tali operazioni diminuiscono β_i, da cui $\beta_i > 0$.

Osservando dunque che l'albero corretto ha per definizione una configurazione congruente di sequenze per ogni ramo interno e che un albero errato ne ha almeno una incongruente rispetto a quello vero, si ha $\beta_i = 0 \ \ \forall i \in E$ in T e $\exists i \in E$ t.c. $\beta_i > 0$ in W, da cui, se i pesi dell'albero vero b_i sono tutti positivi, si ha $E(S_W - S_T) > 0$ per qualsiasi scelta di W, come volevasi dimostrare.

Quindi abbiamo dimostrato che $E(S_W - S_T) > 0$ per qualsiasi scelta di W, e che dunque il valore di $E(S)$ per la topologia limite di quelle ottime è, nel caso completo, il minimo tra quelli di tutte le filogenie. Equivalentemente si può scrivere in forma matriciale:

$$E(S_W - S_T) = \mathbf{u}[(\mathbf{A}_W^t \mathbf{A}_W)^{-1} \mathbf{A}_W^t \mathbf{A}_T - \mathbf{I}]\mathbf{b} > 0$$

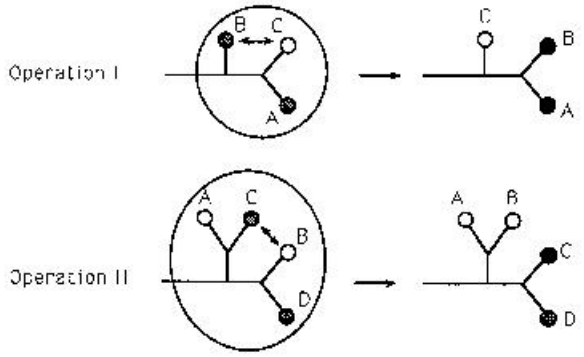

Figura 3.2: Operazioni che permettono di trasformare una configurazione incongruente in una congruente nel caso di sequenze separate da uno (I) o due (II) rami interni.

con $\mathbf{u} = (1 \dots 1)^t$ e $\mathbf{I}$ matrice identità. Ciò permette di concludere che vale **sempre** la consistenza statistica per l'Osservazione 2, anche se consideriamo le medie delle misure evolutive perturbate da errori sperimentali.

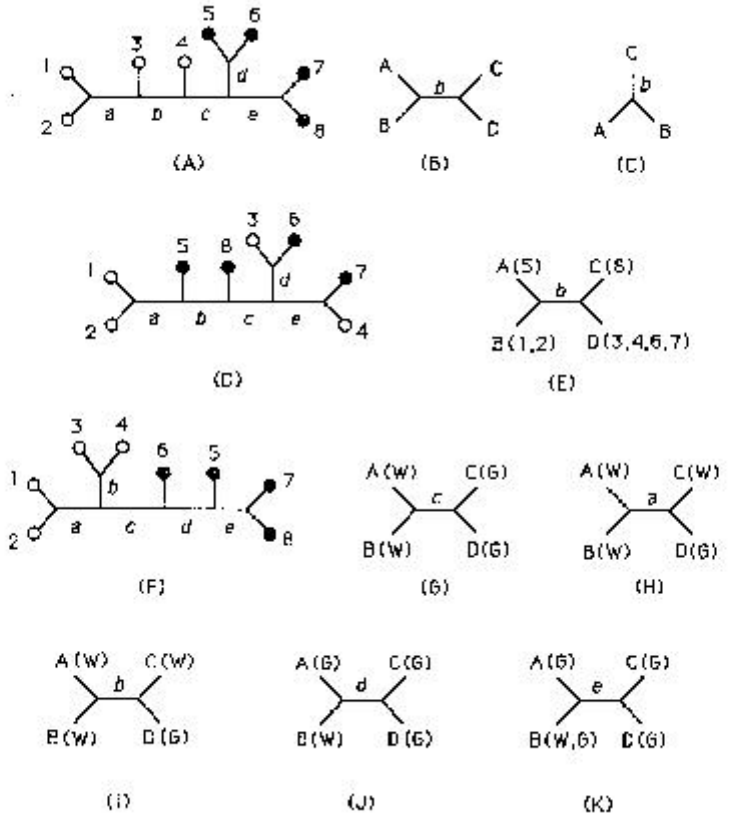

Figura 3.3: (A) Filogenia vera per otto taxa. Il ramo c di quest'albero separa le sequenze nei due gruppi colorati in bianco e in grigio. (B) Quattro gruppi di sequenze A,B,C,D generati considerando il ramo interno b. (C) Gruppi di una e più sequenze generati considerando il ramo esterno b. (D) Albero errato sugli stessi taxa di (A) con una configurazione incongruente. (E) Quattro gruppi di sequenze generati considerando il ramo b nell'albero (D). Un altro albero errato sugli stessi taxa di (A) con una configurazione congruente. (G)-(K) Quattro gruppi generati considerando diversi rami interni dell'albero (F); (W) e (G) stanno per gruppi di sequenze bianche e grigie rispettivamente, (W,G) per misto di sequenze bianche e grigie.

Capitolo 4

Metodo della Minima Evoluzione Bilanciata

Il metodo della Minima Evoluzione esposto nel capitolo precedente può essere così formulato:

Problema 1 (Problema della Minima Evoluzione)**.**

$$\begin{aligned} \min_{(\mathbf{X},\mathbf{w})} \quad & L(\mathbf{X},\mathbf{w}) \\ \text{soggetto a} \quad & f(\mathbf{D},\mathbf{X},\mathbf{w}) = 0 \\ & \mathbf{X} \in \mathcal{X} \\ & \mathbf{w} \in \mathcal{R}_{0+}^{2n-3} \end{aligned}$$

dove $L(\mathbf{X},\mathbf{w})$ indica la lunghezza di un albero $\mathbf{X}$ (modellizzata da una matrice di incidenza archi-cammini nell'insieme delle filogenie $\mathcal{X}$) con pesi sugli archi $\mathbf{w}$, e $f(\mathbf{D},\mathbf{X},\mathbf{w})$ è una funzione che correla la matrice delle distanze $\mathbf{D}$ in input con $\mathbf{X}$ e $\mathbf{w}$. Dunque ogni versione del problema è completamente caratterizzata dalla scelta di $L(\mathbf{X},\mathbf{w})$ e $f(\mathbf{D},\mathbf{X},\mathbf{w})$. Nella formulazione al capitolo precedente, in particolare, si era imposto:

$$L(\mathbf{X},\mathbf{w}) = \sum_{e=1}^{2n-3} w_e$$
$$f(\mathbf{D},\mathbf{X},\mathbf{w}) = \mathbf{w} - (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{d}$$

ovvero che la funzione obiettivo fosse la somma dei pesi della filogenia e che tali valori fossero basati sugli stimatori dei minimi quadrati ordinari (3.4).

Nel seguito introdurremo invece la versione della Minima Evoluzione Bilanciata.

In essa si pone anzitutto $L(\mathbf{X},\mathbf{w}) = \sum_{e=1}^{2n-3} w_e$, ovvero la somma dei pesi della filogenia. Questa funzione obiettivo, la stessa della formulazione data nel capitolo precedente, è stata osservata avere un fondamento

biologico [1]. Infatti sebbene sia inverosimile che l'evoluzione proceda segua sempre il processo più breve possibile, ovvero che non ci siano mutazioni genetiche parallele cioé che nel tempo si annullino l'un l'altra, questo generalmente avviene a livello locale nel caso specifico dei dati molecolari ben-conservati. Questi sono quelle parti delle sequenze genetiche legate a funzioni biochimiche basilari piuttosto che a quelle esteriori, che nel corso dell'evoluzione subiscono cambiamenti minimi. Tale minimo è comunque soltanto locale e in generale non globale nel lungo periodo, principalmente perché gli alleli hanno nei codoni un insieme di vicini limitato e i fattori evolutivi non sono costanti nel tempo. Dunque i pesi di una filogenia di lunghezza minima forniscono solo un limite inferiore sulle distanze evolutive reali fra le specie.

Nel modello Bilanciato si pone poi come $f(\mathbf{D}, \mathbf{X}, \mathbf{w})$ un'espressione dei pesi degli archi derivante da un modifica delle stime del modello dei Minimi Quadrati Ordinari introdotta da Pauplin [11]. Nel seguito illustreremo i vantaggi legati a tale stima, ovvero la consistenza nel caso additivo, la semplicità computazionale, la consistenza statistica e il rispetto del vincolo di non-negatività sotto la sola condizione (2.4).

4.1 Consistenza nel caso additivo

Il metodo della Minima Evoluzione Bilanciata ha come funzione obiettivo:

$$S_{DC} := \sum_i \sum_j D_{ij}/2^{B_{ij}} \tag{4.1}$$

dove e sono gli archi della filogenia, w_e i loro pesi, D_{ij} le misure di diversità fra le specie contenute nella matrice delle distanze e infine B_{ij} le distanze topologiche fra i taxa, ovvero il numero di rami che separano le corrispondenti foglie nella topologia del grafo. L'espressione (4.1), dovuta a Pauplin [11], è motivata dal fatto che nel caso in cui le distanze in input siano additive rispetto alla topologia associata alle B_{ij}, allora S_{DC} è esattamente la somma dei pesi sugli archi di tale albero, dunque in tal caso si può scrivere:

$$S_{DC} = \sum_{e \in E} w_e.$$

Per dimostrarlo consideriamo separatamente le espressioni del peso di un ramo esterno e di uno interno in una filogenia le cui somme dei pesi sui cammini fra i taxa sono esattamente le D_{ij}.

1. **Calcolo del peso di un ramo esterno.** Si consideri l'esempio più semplice di albero binario avente almeno un vertice interno, illustrato in Figura 4.1. In tale albero il peso del ramo esterno $1Z$ è dato da:
$$L_{1Z} = \frac{D_{12} + D_{13} - D_{23}}{2}. \tag{4.2}$$

Si noti che tale espressione rimane valida anche nel caso in cui i taxa 2 e 3 non siano adiacenti a Z, infatti si ha:

$$P_{23} = (P_{12} \cup P_{13}) \setminus \{1Z\}$$

poiché, come visto al Capitolo 2, in un albero esiste uno ed un solo cammino fra due vertici.

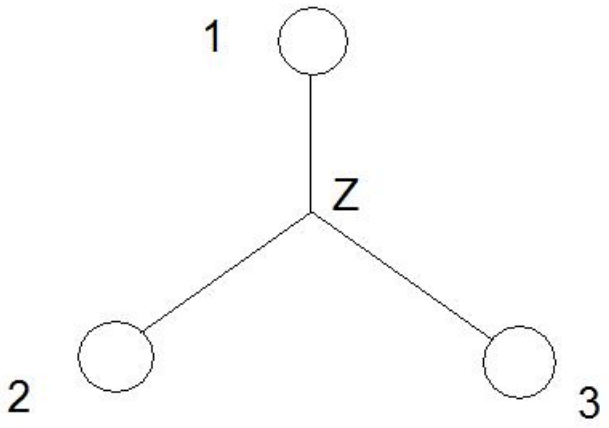

Figura 4.1: Un albero binario con un vertice interno.

Nel caso di una filogenia generica si ha che un ramo esterno partiziona i taxa in tre gruppi, contenenti le foglie dei tre alberi con radice generati rimuovendo il vertice interno del ramo (vedi Capitolo 2). Ad esempio in Figura 4.2 l'arco $1Z$, di vertice interno Z, divide il singoletto $\{1\}$, I e J. Considerando tale partizione si verifica la stima del peso di $1Z$ (3.17):

$$L_{1Z} := \left\{ (1/n_I) \sum_{i \in I} D_{1i} + (1/n_J) \sum_{j \in J} D_{1j} - (1/n_I n_J) \sum_{i \in I} \sum_{j \in J} D_{ij} \right\} /2$$

dove $n_I = |I|, n_J = |J|$. Infatti si ha:

$$\begin{aligned} L_{1Z} &= \frac{1}{2n_I n_J} \{ n_J \sum_{i \in I} D_{1i} + n_I \sum_{j \in J} D_{1j} - \sum_{i \in I} \sum_{j \in J} D_{ij} \} \\ &= \frac{1}{2n_I n_J} \sum_{i \in I} \sum_{j \in J} (D_{1i} + D_{1j} - D_{ij}) \\ &= \frac{1}{2n_I n_J} (2n_I n_J L_{1Z}) \end{aligned}$$

con l'ultima uguaglianza che segue dall'equazione (4.2). Si noti che nella stima (3.17) compaiono le medie non pesate delle distanze evolutive. Pauplin invece calcola la media in modo tale che ogni ramo conti ugualmente dopo una biforcazione (sia cioè pesato o "bilanciato") [11], dunque poiché il numero di biforcazioni sul cammino fra due taxa è pari a $B_{ij} - 2$ si ha:

$$\begin{aligned} L_{1Z} :&= \left[\sum_{i\in I}(1/2^{B_{1i}-2})D_{1i} + \sum_{j\in J}(1/2^{B_{1j}-2})D_{1j} - \sum_{i\in I}\sum_{j\in J}(1/2^{B_{ij}-2})D_{ij}\right]/2 \\ &= \sum_{i\in I}(1/2^{B_{1i}-1})D_{1i} + \sum_{j\in J}(1/2^{B_{1j}-1})D_{1j} - \sum_{i\in I}\sum_{j\in J}(1/2^{B_{ij}-1})D_{ij} \end{aligned} \tag{4.3}$$

Osservazione 4. Si è visto che rimuovendo il vertice interno Z si ottengono due alberi binari con radice (tre se si considera quello degenere $\{1\}$) aventi come insieme delle foglie rispettivamente I e J. Dunque, imponendo l'ordinamento introdotto al Capitolo 2, è possibile pesare ogni taxa di tali alberi dimezzando i coefficienti ad ogni livello, e poiché ogni nodo ha sempre due "figli" si ha:

$$\sum_{m\in M}\frac{1}{2^{B_{1m}-2}} = 1 \text{ con } M = I, J \tag{4.4}$$

Dall'osservazione precedente, e dal fatto che

$$\frac{1}{2^{B_{ij}-2}} = \frac{1}{2^{B_{1i}-2}2^{B_{1j}-2}} \qquad \forall (i,j) \in I \times J$$

si ottiene l'esattezza della stima (4.3) nel caso additivo:

$$\begin{aligned} L_{1Z} &= \frac{1}{2}\left[\left(\sum_{j\in J}\frac{1}{2^{B_{1j}-2}}\right)\sum_{i\in I}\frac{D_{1i}}{2^{B_{1i}-2}} + \left(\sum_{i\in I}\frac{1}{2^{B_{1i}-2}}\right)\sum_{j\in J}\frac{D_{1j}}{2^{B_{1j}-2}} - \sum_{i\in I}\sum_{j\in J}\frac{D_{ij}}{2^{B_{1i}-2}2^{B_{1j}-2}}\right] \\ &= \frac{1}{2}\sum_{i\in I}\sum_{j\in J}\frac{D_{1i}+D_{1j}-D_{ij}}{2^{B_{1i}-2}2^{B_{1j}-2}} \\ &= \frac{1}{2}2L_{1Z}\left(\sum_{i\in I}\frac{1}{2^{B_{1i}-2}}\right)\left(\sum_{j\in J}\frac{1}{2^{B_{1j}-2}}\right) \end{aligned}$$

con l'ultima uguaglianza che segue dall'equazione (4.2).

2. **Calcolo del peso di un ramo interno.** Si consideri l'esempio più semplice di albero binario avente almeno un arco interno, illustrato in Figura 4.3. In tale albero il peso del ramo interno YZ è dato da:

$$L_{YZ} = \frac{D_{13} + D_{14} + D_{23} + D_{24} - 2D_{12} - 2D_{34}}{4}. \tag{4.5}$$

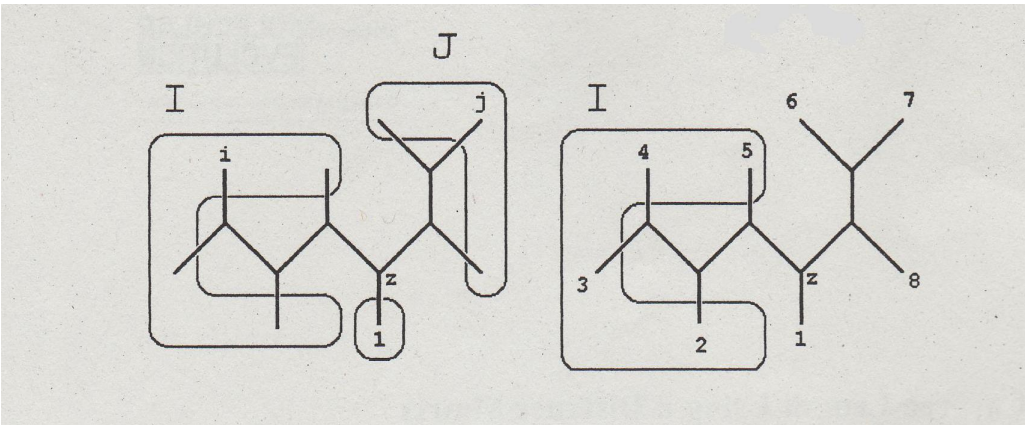

Figura 4.2: Esempio di filogenia in cui il ramo esterno $1Z$, di vertice interno Z, partiziona i taxa negli insiemi $\{1\}$, I e J.

Si noti che tale espressione rimane valida anche nel caso in cui i taxa 1,2 e 3,4 non siano adiacenti rispettivamente a Y e Z, infatti si ha:

$$P_{12} \cup P_{34} = (P_{13} \cup P_{24}) \setminus \{YZ\} = (P_{14} \cup P_{23}) \setminus \{YZ\}$$

poiché, come visto al Capitolo 2, in un albero esiste uno ed un solo cammino fra due vertici.

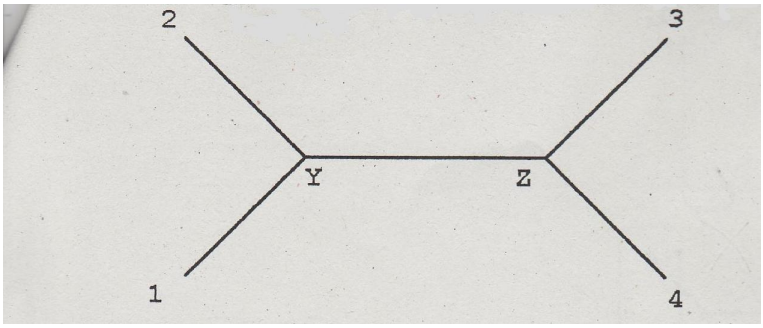

Figura 4.3: Un albero binario con un arco interno.

Nel caso di una filogenia generica si ha che un ramo interno partiziona i taxa in quattro gruppi, contenenti le foglie dei quattro alberi con radice generati rimuovendo tale arco (vedi Capitolo 2). Ad esempio in Figura 4.4 l'arco YZ divide I, J, K, L. Considerando tale partizione si verifica la stima del peso di YZ (3.16):

$$\begin{aligned} L_{YZ} := \frac{1}{2}\{&\gamma[d_{IK}/(n_I n_K) + d_{JL}/(n_J n_L)] \\ &+ (1-\gamma)[d_{JK}/(n_J n_K) + d_{IL}/(n_I n_L)] - d_{IJ}/(n_I n_J) - d_{KL}/(n_K n_L)\} \end{aligned}$$

dove $n_I = |I|, n_J = |J|, n_K = |K|, n_L = |L|$, γ e $d_{IK}, d_{JL}, d_{JK}, d_{IL}, d_{IJ}, d_{KL}$ sono definiti come nell'espressione (3.16). Infatti si ha:

$$\begin{aligned} L_{YZ} &= \frac{1}{2n_I n_J n_K n_L}\{\gamma[n_J n_L d_{IK} + n_I n_K d_{JL}] + (1-\gamma)[n_I n_L d_{JK} + n_J n_K d_{IL}] \\ &\quad - n_K n_L d_{IJ} - n_I n_J d_{KL}\} \\ &= \frac{1}{2n_I n_J n_K n_L}\sum_{i\in I}\sum_{j\in J}\sum_{k\in K}\sum_{l\in L}(d_{ik} + d_{jl} - d_{ij} - d_{kl}) \\ &= \frac{1}{2n_I n_J n_K n_L}(2n_I n_J n_K n_L L_{YZ}) \end{aligned}$$

con l'ultima uguaglianza che segue dall'equazione (4.5) e la penultima che segue dal fatto che $d_{ik} + d_{jl} = d_{il} + d_{jk}$ $\forall (i,j,k,l) \in I \times J \times K \times L$. Si noti che nella stima (3.16) compaiono le medie non pesate delle distanze evolutive. Pauplin invece calcola nuovamente la media in modo tale che ogni ramo conti ugualmente dopo una biforcazione (sia cioè pesato o "bilanciato") [11], dunque poiché il numero di biforcazioni sul cammino fra due taxa è pari a $B_{ij} - 2$ si ha:

$$\begin{aligned} L_{YZ} :&= \Bigg[\sum_{i\in I}\sum_{k\in K}(1/2^{B_{ik}-3})D_{ik} + \sum_{i\in I}\sum_{l\in L}(1/2^{B_{il}-3})D_{il} + \sum_{j\in J}\sum_{k\in K}(1/2^{B_{jk}-3})D_{jk} \\ &\quad + \sum_{j\in J}\sum_{l\in L}(1/2^{B_{jl}-3})D_{jl} - 2\sum_{i\in I}\sum_{j\in J}(1/2^{B_{ij}-2})D_{ij} - 2\sum_{k\in K}\sum_{l\in L}(1/2^{B_{kl}-2})D_{kl}\Bigg]/4 \\ &= \sum_{i\in I}\sum_{k\in K}(1/2^{B_{ik}-1})D_{ik} + \sum_{i\in I}\sum_{l\in L}(1/2^{B_{il}-1})D_{il} + \sum_{j\in J}\sum_{k\in K}(1/2^{B_{jk}-1})D_{jk} \\ &\quad + \sum_{j\in J}\sum_{l\in L}(1/2^{B_{jl}-1})D_{jl} - \sum_{i\in I}\sum_{j\in J}(1/2^{B_{ij}-1})D_{ij} - \sum_{k\in K}\sum_{l\in L}(1/2^{B_{kl}-1})D_{kl} \end{aligned} \tag{4.6}$$

Dalla stessa osservazione del caso del ramo esterno si può scrivere:

$$\sum_{m\in M}\frac{1}{2^{B_{Zm}-2}} = 1 \text{ con } M = I, J \ , \ \sum_{n\in N}\frac{1}{2^{B_{Yn}-2}} = 1 \text{ con } N = K, L$$

e

$$\begin{aligned} &\frac{1}{2^{B_{ik}-3}} = \frac{1}{2^{B_{Zi}-2}2^{B_{Yk}-2}} && \forall (i,k) \in I \times K \text{ e analogamente per le } (i,l),(j,k),(j,l) \\ &\frac{1}{2^{B_{ij}-2}} = \frac{1}{2^{B_{Zi}-2}2^{B_{Zj}-2}} && \forall (i,j) \in I \times J \\ &\frac{1}{2^{B_{kl}-2}} = \frac{1}{2^{B_{Yk}-2}2^{B_{Yl}-2}} && \forall (k,l) \in K \times L \end{aligned}$$

ottenendo l'esattezza della stima (4.6) nel caso additivo:

$$
\begin{aligned}
L_{YZ} = \frac{1}{4}\Bigg[& \left(\sum_{j\in J}\frac{1}{2^{B_{Zj}-2}}\right)\left(\sum_{l\in L}\frac{1}{2^{B_{Yl}-2}}\right)\sum_{i\in I}\sum_{k\in K}\frac{D_{ik}}{2^{B_{Zi}-2}2^{B_{Yk}-2}} \\
& + \left(\sum_{j\in J}\frac{1}{2^{B_{Zj}-2}}\right)\left(\sum_{k\in K}\frac{1}{2^{B_{Yk}-2}}\right)\sum_{i\in I}\sum_{l\in L}\frac{D_{il}}{2^{B_{Zi}-2}2^{B_{Yl}-2}} \\
& + \left(\sum_{i\in I}\frac{1}{2^{B_{Zi}-2}}\right)\left(\sum_{l\in L}\frac{1}{2^{B_{Yl}-2}}\right)\sum_{j\in J}\sum_{k\in K}\frac{D_{jk}}{2^{B_{Zj}-2}2^{B_{Yk}-2}} \\
& + \left(\sum_{i\in I}\frac{1}{2^{B_{Zi}-2}}\right)\left(\sum_{k\in K}\frac{1}{2^{B_{Yk}-2}}\right)\sum_{j\in J}\sum_{l\in L}\frac{D_{jl}}{2^{B_{Zj}-2}2^{B_{Yl}-2}} \\
& - 2\left(\sum_{k\in K}\frac{1}{2^{B_{Yk}-2}}\right)\left(\sum_{l\in L}\frac{1}{2^{B_{Yl}-2}}\right)\sum_{i\in I}\sum_{j\in J}\frac{D_{ij}}{2^{B_{Zi}-2}2^{B_{Zj}-2}} \\
& - 2\left(\sum_{i\in I}\frac{1}{2^{B_{Zi}-2}}\right)\left(\sum_{j\in J}\frac{1}{2^{B_{Zj}-2}}\right)\sum_{k\in K}\sum_{l\in L}\frac{D_{kl}}{2^{B_{Yk}-2}2^{B_{Yl}-2}}\Bigg] \\
= & \frac{1}{4}\sum_{i\in I}\sum_{j\in J}\sum_{k\in K}\sum_{l\in L}\frac{D_{ik}+D_{il}+D_{jk}+D_{jl}-2D_{ij}-2D_{kl}}{2^{B_{Zi}-2}2^{B_{Zj}-2}2^{B_{Yk}-2}2^{B_{Yl}-2}} \\
= & \frac{1}{4}4L_{YZ}\left(\sum_{i\in I}\frac{1}{2^{B_{Zi}-2}}\right)\left(\sum_{j\in J}\frac{1}{2^{B_{Zj}-2}}\right)\left(\sum_{k\in K}\frac{1}{2^{B_{Yk}-2}}\right)\left(\sum_{l\in L}\frac{1}{2^{B_{Yl}-2}}\right)
\end{aligned}
$$

con l'ultima uguaglianza che segue dall'equazione (4.5).

Nelle stime (4.3),(4.6) sono presenti molte frazioni che possono condurre a errori di arrotondamento nei calcolatori. A tal proposito se si chiama B_{max} il massimo numero di archi che separano due taxa nella filogenia si ha:

$$1/2^{B_{ij}-1} = 2^{B_{max}-B_{ij}}/2^{B_{max}-1} \qquad \forall i,j \in \Gamma$$

con $B_{max} - B_{ij}, B_{max} - 1 \geq 0\ \forall i,j \in \Gamma$ dalla definizione di B_{max}. Quindi le stime (4.3),(4.6) si possono ridurre a comune denominatore:

$$L_{1Z} = \left(\sum_{i\in I} 2^{B_{max}-B_{1i}}D_{1i} + \sum_{j\in J} 2^{B_{max}-B_{1j}}D_{1j} - \sum_{i\in I}\sum_{j\in J} 2^{B_{max}-B_{ij}-1}D_{ij}\right)/2^{B_{max-1}} \tag{4.7}$$

$$
\begin{aligned}
L_{YZ} = \Bigg(& \sum_{i\in I}\sum_{k\in K} 2^{B_{max}-B_{ik}}D_{ik} + \sum_{i\in I}\sum_{l\in L} 2^{B_{max}-B_{il}}D_{il} + \sum_{j\in J}\sum_{k\in K} 2^{B_{max}-B_{jk}}D_{jk} \\
& + \sum_{j\in J}\sum_{l\in L} 2^{B_{max}-B_{jl}}D_{jl} - \sum_{i\in I}\sum_{j\in J} 2^{B_{max}-B_{ij}}D_{ij} - \sum_{k\in K}\sum_{l\in L} 2^{B_{max}-B_{kl}}D_{kl}\Bigg)/2^{B_{max-1}}
\end{aligned} \tag{4.8}
$$

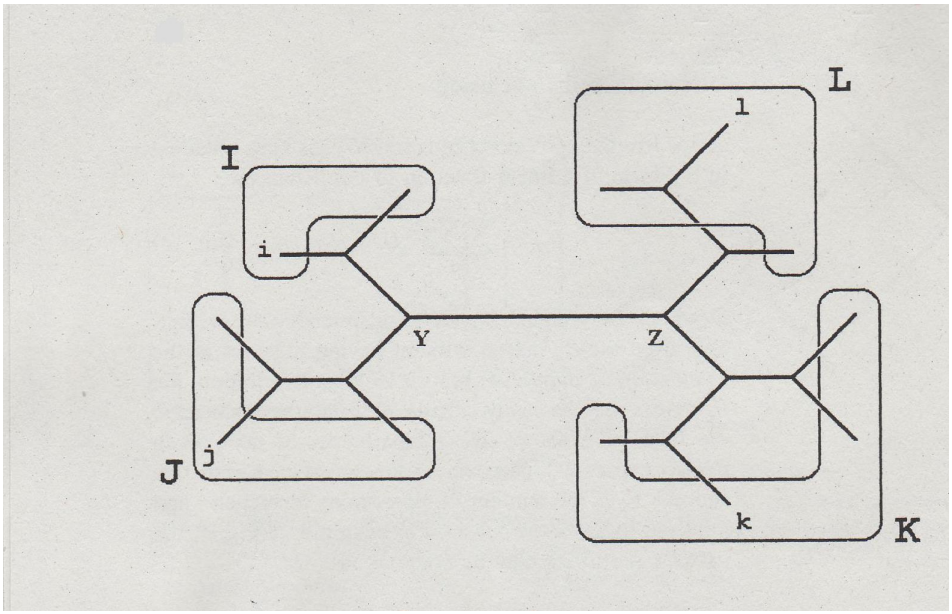

Figura 4.4: Esempio di filogenia in cui il ramo interno YZ partiziona i taxa negli insiemi I, J, K, L.

Osservazione 5. Nelle stime (4.7),(4.8) ogni somma dei pesi D_{ij} su di un cammino $P_{ij},\ i, j \in \Gamma$ viene contata:

- positivamente se il ramo di cui si calcola il peso appartiene a P_{ij};
- negativamente se il ramo è adiacente a P_{ij};
- come nulla se il ramo non ha vertici in comune a P_{ij}.

Dall'osservazione precedente e dal fatto che ogni cammino fra due foglie $i, j \in \Gamma$ ha $B_{ij} - 1$ archi adiacenti in un albero binario, segue che nella somma dei pesi stimati dei rami i termini D_{ij} hanno sempre, al numeratore, coefficiente:

$$B_{ij}2^{B_{max}-B_{ij}} - (B_{ij} - 1)2^{B_{max}-B_{ij}} = 2^{B_{max}-B_{ij}}.$$

Si ha allora che la lunghezza della filogenia è data da:

$$S_{DC} := \left(\sum_{i<j} 2^{B_{max}-B_{ij}} D_{ij} \right) / 2^{B_{max}-1}$$

e poiché $D_{ii} = 0$ per definizione, si può riscrivere:

$$S_{DC} := \sum_i \sum_j D_{ij}/2^{B_{ij}}$$

che è esattamente l'espressione (4.1). Si ha allora consistenza nel caso additivo, come volevasi dimostrare.

Si noti che nell'equazione (4.1) **non compaiono i pesi dei singoli archi ma solo le misure di diversità e le distanze topologiche**, dunque il calcolo può essere eseguito **direttamente** avendo a disposizione la matrice delle distanze e la topologia dell'albero. Un altro vantaggio di tale stima è il fatto che la sua complessità computazionale è $O(n^2)$, dunque sono necessari molti meno calcoli rispetto alle espressioni (3.16) e (3.17), di complessità $O(n^3)$, e (3.4), di complessità $O(n^4)$ [11].

4.2 Consistenza statistica

L'espressione della somma dei pesi (4.1) può essere ricavata anche sommando le stime dei pesi ottenute nel modello dei Minimi Quadrati Pesati (3.13), dunque con covarianze delle misure di differenza nulle, ponendo

$$\omega_{ij} = Var(D_{ij}) = k2^{B_{ij}} \tag{4.9}$$

ovvero assumendo che le varianze dei D_{ij} crescano esponenzialmente con le distanze topologiche, con una costante di proporzionalità k che solitamente incorpora l'inverso della lunghezza delle sequenze analizzate [12]. Vale infatti il seguente

Teorema 4.1. *Assumendo le varianze delle D_{ij} date da* (4.9)*, l'espressione* (4.1)*:*

i) è lo stimatore della somma dei pesi di una filogenia con varianza minima;

ii) è uguale alla somma delle stime dei pesi date da (3.13)*.*

Dimostrazione. i) Ogni stimatore della somma dei pesi che sia lineare nelle distanze può essere scritto come $\mathbf{f}^t\mathbf{d}$, con $\mathbf{f} = (f_{ij})$ vettore dei coefficienti. Affinché poi sia consistente si deve avere, nel caso additivo:

$$\mathbf{f}^t\mathbf{d} = \mathbf{f}^t\mathbf{A}_A\hat{\mathbf{b}} = \mathbf{u}^t\hat{\mathbf{b}} \qquad \forall\hat{\mathbf{b}} \Leftrightarrow \mathbf{A}_A^t\mathbf{f} = \mathbf{u}$$

con $\mathbf{u} = (1 \dots 1)^t$ e le stesse notazioni di (3.13). Si vuole allora trovare uno stimatore che risolva il problema:

$$\begin{aligned} &\text{minimizza} && \sum_{i,j} Var(D_{ij}) f_{ij}^2 \\ &\text{soggetto a} && \mathbf{A}_A^t\mathbf{f} = \mathbf{u} \end{aligned}$$

Poiché i vincoli sono lineari e la funzione obiettivo quadratica nei f_{ij}, il problema ammette un'unica soluzione, e si devono trovare moltiplicatori di Lagrange μ_e tali che:

$$2Var(D_{ij}) f_{ij} = \sum_{e \in P_{ij}} \mu_e \qquad \forall i, j$$

Se tale soluzione è la (4.1), allora sostituendo i coefficienti $f_{ij} = 2^{1-B_{ij}}$ nel problema si devono trovare i moltiplicatori. Poiché è stato dimostrato alla Sezione precedente la consistenza di tale stimatore i vincoli sono rispettati, e utilizzando l'ipotesi (4.9) si ha che i μ_e devono soddisfare:

$$\sum_{e \in P_{ij}} \mu_e = 4k \qquad \forall i, j$$

Tale equazione corrisponde a pesare gli archi in modo che la distanza fra ogni taxa sia $4k$, dunque la sola soluzione è:

$$\mu_e = \begin{cases} 2k & \text{se } e \text{ è un ramo esterno,} \\ 0 & \text{altrimenti.} \end{cases}$$

Si è dunque verificato che (4.1) è uno stimatore che risolve il problema di varianza minima.

ii) Per dimostrare che (4.1) è uguale alla somma delle stime dei pesi date da (3.13) con le varianze date in ipotesi si può verificare che pure quest'ultimo stimatore risolve il problema di varianza minima, ed essendo la soluzione unica segue la tesi. Anzitutto si noti che lo stimatore è dato da:

$$\mathbf{f}^t = \mathbf{u}^t (\mathbf{A}_A^t \mathbf{\Omega}^{-1} \mathbf{A}_A)^{-1} \mathbf{A}_A^t \mathbf{\Omega}^{-1}$$

da cui risultano verificati i vincoli imposti:

$$\mathbf{f}^t \mathbf{A}_A = \mathbf{u}^t (\mathbf{A}_A^t \mathbf{\Omega}^{-1} \mathbf{A}_A)^{-1} \mathbf{A}_A^t \mathbf{\Omega}^{-1} \mathbf{A}_A = \mathbf{u}^t$$

Inoltre riscrivendo le condizioni sui moltiplicatori di Lagrange in forma matriciale si ha:

$$\mathbf{V}\mathbf{f} = \mathbf{V}(\mathbf{u}^t (\mathbf{A}_A^t \mathbf{\Omega}^{-1} \mathbf{A}_A)^{-1} \mathbf{A}_A^t \mathbf{\Omega}^{-1})^t = \mathbf{A}_A (\mathbf{A}_A^t \mathbf{\Omega}^{-1} \mathbf{A}_A)^{-1} \mathbf{u}$$

trovando dunque come soluzione i valori

$$\mathbf{m} := (\mu_e)_{e \in E}^t = (\mathbf{A}_A^t \mathbf{\Omega}^{-1} \mathbf{A}_A)^{-1} \mathbf{u}.$$

□

Si noti che il risultato precedente non è immediato: gli stimatori non sono in generale indipendenti, dunque non è a priori garantito che lo stimatore della somma dei pesi di varianza minima sia uguale alla somma degli stimatori dei pesi di varianza minima [12].

Poiché usando simulazioni al computer si è mostrato che alcuni modelli di Minimi Quadrati Pesati sono statisticamente consistenti [5], si è

portati a credere che il criterio Bilanciato, che come appena visto ne è un caso particolare, sia statisticamente consistente.

Per dimostrarlo basta, sempre per l'Osservazione 2, verificare che la somma dei pesi per la topologia limite di quelle ottime è minima nel caso completo, ovvero dimostrare il seguente

Teorema 4.2. *Sia T la topologia limite di quelle ottime e W una qualsiasi altra topologia errata, allora si ha $S_{DC}(W) > S_{DC}(T)$, con lo stimatore* (4.1) *nel caso le distanze D_{ij} siano quelle complete.*

Dimostrazione. Denotiamo anzitutto con $X|Y$ ogni bipartizione sull'insieme dei taxa $\Gamma \supset X, Y$ generata da un ramo della filogenia e con $S(T)$ l'insieme di tali bipartizioni indotte dagli archi della topologia T. Analogamente alla dimostrazione del metodo della Minima Evoluzione con gli stimatori OLS al capitolo precedente, definiamo le distanze relative ad una bipartizione $X|Y$ come $\mathbf{D}^{X|Y} = (D_{ij}^{X|Y})$, con:

$$D_{ij}^{X|Y} := \begin{cases} 1 & \text{se } i \neq j \text{ e } |\{i,j\} \cap X| = 1, \\ 0 & \text{altrimenti} \end{cases}$$

ovvero il vettore delle distanze, in ordine lessicografico, nulle se non includono nel loro cammino l'arco che genera $X|Y$. Sia poi $w(X|Y)$ il peso del ramo che induce $X|Y$. Si noti che

$$\mathbf{d} = \sum_{X|Y \in S(T)} w(X|Y)\mathbf{D}^{X|Y}$$

da cui, per la linearità dell'espressione nelle (4.1)

$$S_{DC}(W, \mathbf{d}) = \sum_{X|Y \in S(T)} w(X|Y) S_{DC}(W, \mathbf{D}^{X|Y})$$

dove $S_{DC}(W, \mathbf{d})$ è la stima (4.1) per la topologia W nel caso delle distanze complete. Dunque, analogamente al caso con gli stimatori OLS al capitolo precedente, si può determinare la positività di

$$S_{DC}(W, \mathbf{d}) - S_{DC}(T, \mathbf{d}) = \sum_{X|Y \in S(T)} w(X|Y)(S_{DC}(W, \mathbf{D}^{X|Y}) - S_{DC}(T, \mathbf{D}^{X|Y}))$$

mostrando che almeno un termine $S_{DC}(W, \mathbf{D}^{X|Y}) - S_{DC}(T, \mathbf{D}^{X|Y})$ è positivo e che gli altri sono non-negativi, se si assume che $w(X|Y) > 0$ almeno per gli archi generano bipartizioni diverse, vero poiché altrimenti W e T sarebbero in sostanza la stessa filogenia seppure con topologie diverse. Ancora seguendo il ragionamento del caso OLS si ha:

1. se una bipartizione $X|Y$ è presente sia in $S(T)$ che in $S(W)$ allora $S_{DC}(W, \mathbf{D}^{X|Y}) = S_{DC}(T, \mathbf{D}^{X|Y}) = 1$;

2. se una bipartizione $X|Y$ è presente in $S(T)$ ma non in $S(W)$ allora $S_{DC}(W, \mathbf{D}^{X|Y}) > S_{DC}(T, \mathbf{D}^{X|Y}) = 1$. Infatti si può ricondurre la configurazione W ad una con la bipartizione $X|Y$ mediante le stesse due operazioni del caso OLS. Si può ancora dimostrare [12] che tali operazioni diminuiscono $S_{DC}(W, \mathbf{D}^{X|Y})$.

Osservando dunque che esiste almeno una bipartizione $X|Y$ presente in $S(T)$ ma non in $S(W)$ visto che le due topologie sono diverse si conclude. □

4.3 Vincolo di non-negatività

Nel metodo della Minima Evoluzione con gli stimatori OLS si è dimostrato [1] che talvolta le stime dei pesi possono essere negative. Nel caso invece della Minima Evoluzione Bilanciata vale il vincolo di non-negatività sotto la sola condizione (2.4).

Si definisca anzitutto [12] la trasformazione NNI (Nearest Neighbor Interchange) della topologia T in una T' ottenuta scambiando due gruppi di taxa partizionati e divisi da un ramo interno, come J e K in Figura 4.4. Si ha in tal caso:

$$S_{DC}(T) - S_{DC}(T') = \frac{1}{4}\left[\left(\sum_{i\in I}\sum_{j\in J}(1/2^{B_{ij}-3})D_{ij} + \sum_{k\in K}\sum_{l\in L}(1/2^{B_{kl}-3})D_{kl}\right) - \left(\sum_{i\in I}\sum_{k\in K}(1/2^{B_{ik}-3})D_{ik} + \sum_{j\in J}\sum_{l\in L}(1/2^{B_{jl}-3})D_{jl}\right)\right] \tag{4.10}$$

Teorema 4.3. *Sia T una filogenia ottenuta applicando la trasformazione NNI ad un albero finché la somma dei suoi pesi secondo* (4.1) *con matrice delle distanze* ***metrica*** *non diminuisce ulteriormente, cioè sia un minimo locale per tale algoritmo. Allora le stime dei pesi sui suoi rami date da* (4.3) *e* (4.6) *sono tutte positive.*

Dimostrazione. Consideriamo anzitutto le stime per i rami interni (4.6). Poiché T è un minimo locale, applicare ad esso la trasformazione NNI rispetto ai gruppi J e K in Figura 4.4 ne aumenta il valore S_{DC} (almeno in un caso), da cui:

$$\sum_{i\in I}\sum_{j\in J}(1/2^{B_{ij}-3})D_{ij}+\sum_{k\in K}\sum_{l\in L}(1/2^{B_{kl}-3})D_{kl} < \sum_{i\in I}\sum_{k\in K}(1/2^{B_{ik}-3})D_{ik}+\sum_{j\in J}\sum_{l\in L}(1/2^{B_{jl}-3})D_{jl}$$

Analogamente per J e L:

$$\sum_{i\in I}\sum_{j\in J}(1/2^{B_{ij}-3})D_{ij}+\sum_{k\in K}\sum_{l\in L}(1/2^{B_{kl}-3})D_{kl} < \sum_{i\in I}\sum_{l\in L}(1/2^{B_{il}-3})D_{il}+\sum_{j\in J}\sum_{k\in K}(1/2^{B_{jk}-3})D_{jk}$$

Dalle precedenti disuguaglianze, da (4.6) e dal fatto che una matrice metrica è anche una matrice di dissimilarità (vedi Capitolo 2) segue la tesi. Infine dalla disuguaglianza triangolare (2.4), da (4.3) e ancora dal fatto che una matrice metrica è anche una matrice di dissimilarità (vedi Capitolo 2) si ottiene la tesi anche nel caso esterno. □

Capitolo 5

Approcci risolutivi al problema

Si è visto al capitolo precedente che la formulazione del problema della Minima Evoluzione Bilanciata ha, tra i vari vantaggi, quello di poter essere implementabile avendo a disposizione solamente la matrice delle distanze e la topologia dell'albero. Dunque, poiché le distanze evolutive sono fornite in input nell'Inferenza Filogenetica, il problema da risolvere si riduce a trovare la configurazione di archi che minimizza l'espressione (4.1), senza specificarne necessariamente i pesi. Si ricordi che si assume che le filogenie, non pesate in questo caso, siano etichettate solo sui taxa per il motivo visto al Capitolo 2.

Nel seguito vedremo prima un algoritmo esatto per la risoluzione di questa istanza e poi una formulazione interamente lineare.

5.1 Un algoritmo esatto

Per la risoluzione del problema della Minima Evoluzione Bilanciata sono stati proposti diversi algoritmi, tutti particolarmente complessi o approssimati [13],[1]. D'altro canto l'algoritmo esatto più banale, ovvero quello di calcolare, data la matrice delle distanze, il valore S_{DC} secondo la stima (4.1) per ogni possibile filogenia su n taxa è inapplicabile dal punto di vista pratico a causa del numero di filogenie semi-etichettate su n taxa, pari a $(2n-5)!!$ come dimostrato in Appendice B, che diventa enorme per valori anche modesti di n, come si nota in Tabella 5.1.

Un algoritmo esatto applicabile è stato invece introdotto in [13], e si basa sulla suddivisione del problema in due fasi:

1. elencare tutte le possibili classi di isomorfismo (vedi Capitolo 2) delle filogenie su n taxa;

2. risolvere il problema per una fissata classe di isomorfismo, ottenendo così una topologia ottima in tale insieme;

Taxa	Filogenie semi-etichettate	Filogenie non etichettate
4	3	1
5	15	1
6	105	2
7	945	2
8	10395	3
9	135135	4
10	2027025	11
15	$\sim 7.9 \times 10^{12}$	265
20	$\sim 2.2 \times 10^{20}$	11020
25	$\sim 2.5 \times 10^{28}$	565734

Tabella 5.1: Numero di filogenie semi-etichettate e non etichettate per un dato numero di taxa.

3. selezionare fra le filogenie ottime trovate al punto precedente quella che minimizza (4.1).

dove la prima serve a ridurre notevolmente, per la discussione precedente, il numero di casi sui quali applicare la seconda. In particolare, per quanto riguarda la seconda istanza, essa consiste nel trovare la filogenia appartenente ad una determinata classe di isomorfismo che minimizzi il valore di (4.1). Poiché si etichettano le filogenie solo sui taxa per il motivo visto al Capitolo 2, tale richiesta è equivalente a

Problema 2 (Problema di assegnamento delle foglie)**.**

$$\begin{aligned} \min_{\phi} \quad & \sum_{i,j \in \Gamma} D_{ij} / 2^{B_{\phi(i)\phi(j)}} \\ \text{soggetto a} \quad & \phi : \Gamma \mapsto L \text{ permutazione} \end{aligned}$$

dove L è l'insieme delle foglie della filogenia non etichettata in esame e le notazioni sono le stesse di (4.1).

La prima operazione può essere eseguita con diversi algoritmi di enumerazione, come ad esempio quello proposto in [13]. Ci interessiamo nel seguito del secondo passaggio, dimostrando che si tratta di un caso particolare del Problema di Assegnamento Quadratico e che si può linearizzarlo [14],[15].

5.1.1 Formulazione ILP del Problema di Assegnamento Quadratico

Definizione 5.1 (QAP, versione di Koopmans-Beckmann)**.** Date due matrici $n \times n$ $\mathbf{A} = (a_{ij}), \mathbf{B} = (b_{ij})$, si dice *Problema di Assegnamento*

Quadratico (Quadratic Assignment Problem, QAP) il seguente:

$$\min_{\phi \in \mathcal{S}_n} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} a_{\phi(i)\phi(j)} b_{ij} \tag{5.1}$$

dove $\mathcal{S}_n$ è l'insieme delle permutazioni su n elementi.

Si ha allora che il problema di assegnamento delle foglie è un'istanza del QAP, con $n = |\Gamma|$ e matrici $\mathbf{A} := (2^{B_{kl}}),\ k, l \in L,\ \mathbf{B} := \mathbf{D} = (D_{ij})$.

Siamo ora interessati a riscrivere il QAP (5.1) come un problema di Programmazione Lineare Intera (Integer Linear Programming, ILP), in modo da potervi applicare tutte le relative tecniche di risoluzione.

Per farlo diamo anzitutto una formulazione del QAP (5.1) come un problema di Programmazione Quadratica Intera [14],[15].

Definizione 5.2. Una matrice $X = (x_{ij})$ di dimensioni $n \times n$ si dice *matrice di permutazione* se è caratterizzata dai seguenti *vincoli di assegnamento*:

$$\begin{aligned} \sum_{i=1}^{n} x_{ij} &= 1, \qquad j = 1, 2, \ldots, n \\ \sum_{j=1}^{n} x_{ij} &= 1, \qquad i = 1, 2, \ldots, n \\ x_{ij} &\in \{0, 1\}, \qquad i, j = 1, 2, \ldots, n \end{aligned}$$

Osservazione 6. Ogni permutazione $\phi \in \mathcal{S}_n$ può essere rappresentata da una matrice di permutazione $X = (x_{ij})$ di dimensioni $n \times n$ tale che:

$$x_{ij} := \begin{cases} 1 & \text{se } \phi(i) = j, \\ 0 & \text{altrimenti.} \end{cases}$$

Usando le matrici di permutazione in luogo delle permutazioni si ha allora:

Definizione 5.3 (QAP, formulazione di Koopmans-Beckmann)**.** Il QAP (5.1) ha la seguente formulazione equivalente come problema di Programmazione Quadratica Intera:

$$\begin{aligned} \min_{X=(x_{ij})} \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} b_{kl} x_{ki} a_{ij} x_{lj} \\ \text{soggetto a} \quad & \sum_{i=1}^{n} x_{ij} = 1, \qquad j = 1, 2, \ldots, n \\ & \sum_{j=1}^{n} x_{ij} = 1, \qquad i = 1, 2, \ldots, n \\ & x_{ij} \in \{0, 1\}, \qquad i, j = 1, 2, \ldots, n \end{aligned} \tag{5.2}$$

o equivalentemente, in maniera più compatta:

$$\begin{aligned} \min_{\mathbf{X}=(x_{ij})} \quad & \langle \mathbf{B}, \mathbf{X}\mathbf{A}\mathbf{X}^t \rangle \\ \text{soggetto a} \quad & \mathbf{X} \text{ matrice di permutazione} \end{aligned}$$

dove

$$\langle \mathbf{A}, \mathbf{B} \rangle := \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} b_{ij}$$

è il *prodotto interno* fra matrici [14].

L'equivalenza di tale formulazione segue dal fatto che se X è la matrice di permutazione associata alla permutazione ϕ allora per definizione:

$$\mathbf{X}\mathbf{A}\mathbf{X}^t = (a_{\phi(i)\phi(j)}).$$

Infine possiamo trasformare il problema di Programmazione Quadratica Intera (5.2) in un problema di Programmazione Lineare Intera. Per farlo introduciamo nella funzione obiettivo di (5.2) nuove variabili binarie imponendo:

$$y_{ijkl} := x_{ij} x_{kl} \tag{5.3}$$

Per esprimere i vincoli di queste nuove variabili coerentemente agli altri si può osservare che è equivalente richiedere che:

$$y_{ijkl} := \begin{cases} 1 & \text{se } x_{ij} = 1 = x_{kl}, \\ 0 & \text{altrimenti.} \end{cases}$$

ovvero, utilizzando i vincoli logici suggeriti in [16]:

$$\begin{aligned} y_{ijkl} &\leq x_{ij} \qquad \forall i,j,k,l = 1,2,\ldots,n \\ y_{ijkl} &\leq x_{kl} \qquad \forall i,j,k,l = 1,2,\ldots,n \\ y_{ijkl} &\geq x_{ij} + x_{kl} - 1 \qquad \forall i,j,k,l = 1,2,\ldots,n \\ y_{ijkl} &\in \{0,1\} \qquad \forall i,j,k,l = 1,2,\ldots,n \end{aligned}$$

o considerando i particolari vincoli del problema (5.2) si possono usare anche [14]:

$$\begin{aligned} 2y_{ijkl} &\leq x_{ij} + x_{kl} \qquad \forall i,j,k,l = 1,2,\ldots,n \\ \sum_{i,j,l,k=1}^{n} y_{ijkl} &= n^2 \\ y_{ijkl} &\in \{0,1\} \qquad \forall i,j,k,l = 1,2,\ldots,n \end{aligned}$$

Inserendo in (5.2) uno dei gruppi di vincoli lineari visti ed eseguendo la sostituzione delle variabili (5.3) nella funzione obiettivo si ottiene così una riscrittura ILP del problema QAP 5.1, come si voleva.

5.2 Formulazione ILP del problema completo

Per generalizzare il risultato della sezione precedente si potrebbe volere una riscrittura dell'**intero** problema della Minima Evoluzione Bilanciata in termini di Programmazione Lineare Intera. La formulazione che seguiamo è quella fornita in [17].

Anzitutto si introducono alcune proprietà delle distanze topologiche su n taxa B_{ij} $i, j \in \Gamma$, nelle stesse notazioni usate per (4.1).

Proposizione 5.1 (Uguaglianza simmetrica)**.** *Le distanze topologiche di una filogenia soddisfano:*

$$B_{ij} = B_{ji} \qquad \forall i, j \in \Gamma \tag{5.4}$$

Dimostrazione. Segue dal fatto che una filogenia è un grafo non orientato. □

Proposizione 5.2 (Uguaglianza di Kraft, Parker e Ram)**.** *Una sequenza di interi $(B_{ij})_{j\in\Gamma\setminus\{i\}}$ rappresenta in una filogenia l'insieme delle distanze topologiche di un fissato taxa $i \in \Gamma$ dagli altri se e solo se:*

$$\sum_{j\in\Gamma\setminus\{i\}} \frac{1}{2^{B_{ij}}} = \frac{1}{2} \tag{5.5}$$

Dimostrazione. ($\Rightarrow$) Segue direttamente da (4.4) nell'Osservazione vista al Capitolo 4, considerando i due alberi binari con radice generati rimuovendo la foglia i (vedi Capitolo 2). L'unione degli insiemi delle foglie di tali alberi è infatti proprio $\Gamma \setminus \{i\}$.

($\Leftarrow$) Segue dalla dimostrazione in [18].

□

Proposizione 5.3 (Uguaglianza 3)**.** *Le distanze topologiche di una filogenia soddisfano:*

$$\sum_{i\in\Gamma} \sum_{j\in\Gamma\setminus\{i\}} \frac{B_{ij}}{2^{B_{ij}}} = 2n - 3 \tag{5.6}$$

Dimostrazione. Segue direttamente dalla consistenza nel caso additivo dell'espressione (4.1). Infatti inserendo in S_{DC} le distanze evolutive relative alla filogenia con struttura topologica data e pesi unitari su tutti gli archi, ovvero ponendo $D_{ij} = B_{ij}$, si ha:

$$\sum_{i\in\Gamma} \sum_{j\in\Gamma\setminus\{i\}} \frac{B_{ij}}{2^{B_{ij}}} = S_{DC} = \sum_{e\in E} w_e = (2n-3)1 = 2n - 3$$

□

Proposizione 5.4 (Condizione dei quattro punti estesa)**.** *Una filogenia non contiene triangoli e le distanze topologiche fra* ***tutti i suoi vertici*** *soddisfano l'equazione* (2.5)*, dunque con* $i, j, k, z \in \Gamma \cup V$ *anziché* $i, j, k, z \in \Gamma$*.*

La proposizione precedente è derivata dalla restrizione di una proprietà più generale relativa alle matrici additive dimostrata da Buneman in [19], dove si dimostra valere anche l'implicazione inversa.

Dunque si è visto che la proprietà (5.5) caratterizza completamente le sequenze di distanze topologiche di un fissato taxa $i \in \Gamma$ dagli altri in una filogenia. Di conseguenza la proprietà (5.5) unitamente alla condizione dei quattro punti estesa determinano tutte le distanze topologiche in una filogenia. Infatti una sequenza di interi B_{ij} che soddisfi la condizione dei quattro punti estesa corrisponde alle somme dei pesi lungo i vari cammini fra **tutti i vertici** di una filogenia pesata. Se poi vale anche l'uguaglianza di Kraft, Parker e Ram per i soli cammini fra le foglie allora tali valori della sequenza sono pure distanze topologiche fra i taxa di un'unica topologia, **quella della filogenia introdotta per l'intera sequenza**; inoltre ciò implica che i pesi siano tutti unitari, da cui **tutti** i B_{ij} devono essere distanze topologiche.

Le proprietà delle distanze topologiche introdotte permettono di dare una riscrittura ILP del problema della Minima Evoluzione Bilanciata, come voluto. Infatti, introducendo le variabili binarie in modo che:

$$x_{ij}^k = \begin{cases} 1 & \text{se } B_{ij} = k, \\ 0 & \text{altrimenti.} \end{cases} \qquad \forall i, j \in \Gamma \cup V, i \neq j, 1 \leq k \leq n-1$$

e quelle per linearizzare la condizione dei quattro punti estesa:

$$y_{ijqt} = \begin{cases} 1 & \text{se } B_{it} + B_{jq} \geq B_{iq} + B_{jt}, \\ 0 & \text{altrimenti.} \end{cases} \qquad \forall i, j, q, t \in \Gamma \cup V, i \neq j \neq q \neq t$$

si ottiene la formulazione fornita in [17]:

Definizione 5.4 (BMEP, formulazione ILP)**.**

$$\min_{X=(x_{ij})} \quad \sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} D_{ij}\left(\sum_{k=2}^{n-1} 2^{-k}x_{ij}^k\right) \tag{5.7}$$

soggetto a

$$\sum_{k=1}^{n-1} x_{ij}^k = 1, \qquad \forall i\neq j\in\Gamma\cup V \tag{5.8}$$

$$x_{ij} = x_{ji}, \qquad \forall i<j\in\Gamma\cup V, 1\leq k\leq n-1 \tag{5.9}$$

$$\sum_{j\in\Gamma\setminus\{i\}}\sum_{k=2}^{n-1} 2^{-k}x_{ij}^k = \frac{1}{2}, \qquad \forall i\in\Gamma \tag{5.10}$$

$$\sum_{k=2}^{n-1} k2^{-k}\sum_{i\in\Gamma}\sum_{j\in\Gamma\setminus\{i\}} x_{ij}^k = 2n-3 \tag{5.11}$$

$$\sum_{k=1}^{n-1} k(x_{ij}^k + x_{qt}^k) \leq \sum_{k=1}^{n-1} k(x_{iq}^k + x_{jt}^k) + (2n-2)y_{ijqt}, \tag{5.12}$$

$$\forall i\neq j\neq q\neq t\in\Gamma\cup V$$

$$\sum_{k=1}^{n-1} k(x_{ij}^k + x_{qt}^k) \leq \sum_{k=1}^{n-1} k(x_{it}^k + x_{jq}^k) + (2n-2)(1-y_{ijqt}), \tag{5.13}$$

$$\forall i\neq j\neq q\neq t\in\Gamma\cup V$$

$$x_{ij}^1 = 0, \qquad \forall i\neq j\in\Gamma \tag{5.14}$$

$$\sum_{i,j\in\Gamma\cup V, i\neq j} x_{ij}^1 = 2n-3 \tag{5.15}$$

$$\sum_{j\in V} x_{ij}^1 = 1, \qquad \forall i\in\Gamma \tag{5.16}$$

$$\sum_{j\in\Gamma\cup V, i\neq j} x_{ij}^1 = 3, \qquad \forall i\in V \tag{5.17}$$

$$x_{ij}^1 + x_{il}^1 + x_{lj}^1 \leq 2, \qquad \forall i\neq j\neq l\in V \tag{5.18}$$

$$x_{ij}^k + 1 \geq x_{il}^{k-1} + x_{lj}^1, \qquad \forall i\neq j\in\Gamma, l\in V, 2\leq k\leq n-2 \tag{5.19}$$

$$x_{ij}^k + x_{ij}^{k-2} + 1 \geq x_{il}^{k-1} + x_{lj}^1, \qquad \forall i\neq j\neq l\in\Gamma\cup V, 3\leq k\leq n-2 \tag{5.20}$$

$$x_{ij}^k \in \{0,1\}, \qquad \forall i,j\in\Gamma\cup V, 1\leq k\leq n-1 \tag{5.21}$$

$$y_{ijqt} \in \{0,1\}, \qquad \forall i\neq j\neq q\neq t\in\Gamma\cup V \tag{5.22}$$

In tale formulazione i vincoli traducono le condizioni del problema:

- (5.8) impone che le B_{ij} assumano esattamente un valore compreso fra 1 e $n-1$;

- (5.9) impone l'uguaglianza simmetrica (5.4);
- (5.10) impone l'uguaglianza di Kraft, Parker e Ram (5.5);
- (5.11) impone l'uguaglianza 3 (5.6);
- (5.12) e (5.13) impongono la condizione dei quattro punti estesa;
- (5.14)–(5.20) descrivono la struttura di una filogenia, specificatamente:
  - (5.14) impone che non ci siano archi fra i taxa;
  - (5.15) impone che il grafo abbia $2n - 3$ archi;
  - (5.16) e (5.17) impongono il vincolo sui gradi;
  - (5.18) impone che non ci siano triangoli;
  - (5.19) e (5.20) collegano le variabili relative agli archi $x_{ij}^k, k = 1$ a quelle relative ai cammini $x_{ij}^k, k \geq 2$.

Appendice A

Approssimazione ai minimi quadrati

Osservazione 7. Nel seguito considereremo il caso generale di matrici a entrate complesse, i risultati si possono particolarizzare al caso reale sostituendo all'operazione di trasposizione e coniugio $\mathbf{M}^H$ di una matrice $\mathbf{M}$ quella di sola trasposizione $\mathbf{M}^t$. Le dimostrazioni sono tratte da [7].

Se un sistema lineare $\mathbf{Ax} = \mathbf{b}$ ha la matrice dei coefficienti $\mathbf{A}$ (quadrata) invertibile, allora il vettore $\mathbf{A}^{-1}\mathbf{b}$ è l'unica soluzione del sistema. Si può generalizzare la nozione di matrice inversa nel caso generico di una matrice $m \times n$ $\mathbf{A}$ con la seguente.

Definizione A.1 (Pseudo-inversa di Moore-Penrose)**.** Data una qualunque matrice $m \times n$ $\mathbf{A}$, si chiama *pseudo-inversa* di $\mathbf{A}$ una matrice $\mathbf{A}^+$ he soddisfa alle quattro condizioni:

$$\mathbf{AA}^+\mathbf{A} = \mathbf{A},\ \mathbf{A}^+\mathbf{AA}^+ = \mathbf{A}^+,\ \mathbf{AA}^+ = (\mathbf{AA}^+)^H,\ \mathbf{A}^+\mathbf{A} = (\mathbf{A}^+\mathbf{A})^H \tag{A.1}$$

da cui la matrice $\mathbf{A}^+$ ha necessariamente dimensioni $n \times m$

Tale matrice esiste sempre ed è unica [7]. Vedremo di seguito come il motivo di tale generalizzazione sia che il vettore $\mathbf{A}^+\mathbf{b}$ ha un significato particolare per il sistema lineare visto sopra con $\mathbf{A}$ matrice $m \times n$ qualunque.

Si ricordi che il sistema $\mathbf{Ax} = \mathbf{b}$ ha soluzione se e solo se $\mathbf{b} \in C(\mathbf{A})$, spazio delle colonne di $\mathbf{A}$ e che esiste un unico vettore in $C(\mathbf{A})$ che ha distanza euclidea minima da $\mathbf{b}$; tale vettore coincide con la proiezione ortogonale $\mathbf{b}_0$ del vettore $\mathbf{b}$ su $C(\mathbf{A})$, e $\mathbf{b} = \mathbf{b}_0$ se e solo se $\mathbf{b} \in C(\mathbf{A})$. Il teorema seguente caratterizza le soluzioni del sistema compatibile $\mathbf{Ax} = \mathbf{b}_0$.

Teorema A.1. *Dato il sistema lineare* $\mathbf{Ax} = \mathbf{b}$*, con* $\mathbf{A}$ *matrice* $m \times n$*, e detto* $\mathbf{b}_0$ *il vettore proiezione di* $\mathbf{b}$ *su* $C(\mathbf{A})$*, le seguenti condizioni per un vettore* $\mathbf{v} \in \mathbb{C}^n$ *sono equivalenti:*

(a) $\mathbf{Av} = \mathbf{b}_0$*;*

(b) $||\mathbf{Av}-\mathbf{b}||_2 < ||\mathbf{Au}-\mathbf{b}||_2$ *per ogni vettore* $\mathbf{u} \in \mathbb{C}^n$ *tale che* $\mathbf{Av} \neq \mathbf{Au}$*;*

(c) $\mathbf{A}^H\mathbf{Av} = \mathbf{A}^H\mathbf{b}$*.*

Dimostrazione.

(a)⇔(b) Discende dalle proprietà della proiezione ortogonale (vedi [7]) e dal fatto che un vettore appartiene a $C(\mathbf{A})$ se e solo se è del tipo $\mathbf{Au}$ per un opportuno vettore $\mathbf{u} \in \mathbb{C}^n$.

(c)⇔(a) Ricordando che $C(\mathbf{A})^\perp = N(\mathbf{A}^H)$ per le proprietà dei sottospazi fondamentali di una matrice (vedi [7]), si ha:

$$\mathbf{A}^H\mathbf{Av} = \mathbf{A}^H\mathbf{b} \Leftrightarrow \mathbf{Av} - \mathbf{b} \in N(\mathbf{A}^H) = C(\mathbf{A})^\perp.$$

Inoltre, poiché $\mathbf{b} = \mathbf{Av} + (\mathbf{b} - \mathbf{Av})$ e $\mathbf{Av} \in C(\mathbf{A})$, ancora dalle proprietà della proiezione ortogonale (vedi [7]) segue che:

$$\mathbf{Av} - \mathbf{b} \in C(\mathbf{A})^\perp \Leftrightarrow \mathbf{Av} = \mathbf{b}_0.$$

In conclusione si ottiene che $\mathbf{A}^H\mathbf{Av} = \mathbf{A}^H\mathbf{b} \Leftrightarrow \mathbf{Av} = \mathbf{b}_0$.

□

Il sistema $\mathbf{A}^H\mathbf{Ax} = \mathbf{A}^H\mathbf{b}$ si chiama il *sistema delle equazioni normali* associato al sistema $\mathbf{Ax} = \mathbf{b}$. Il teorema precedente dice che il sistema delle equazioni normali è risolubile e ha esattamente le soluzioni del sistema $\mathbf{Ax} = \mathbf{b}_0$.

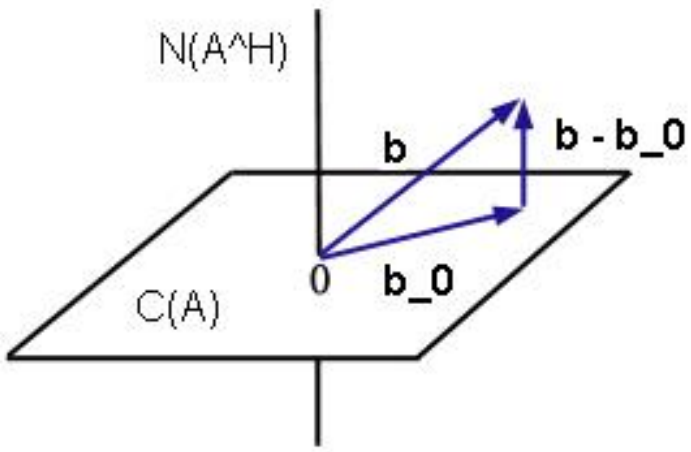

Figura A.1: Proiezione ortogonale $\mathbf{b}_0$ del vettore $\mathbf{b}$ su $C(\mathbf{A})$.

Otteniamo allora il significato del vettore $\mathbf{A}^+\mathbf{b}$, enunciato di seguito.

Teorema A.2. *Dato il sistema lineare* $\mathbf{Ax} = \mathbf{b}$ *e detto* $\mathbf{b}_0$ *il vettore proiezione di* $\mathbf{b}$ *su* $C(\mathbf{A})$*, il vettore* $\mathbf{A}^+\mathbf{b}$ *è l'unica soluzione del sistema* $\mathbf{Ax} = \mathbf{b}_0$ *di norma euclidea minima.*

Dimostrazione. Proviamo innanzitutto che il vettore $\mathbf{A}^+\mathbf{b}$ è soluzione del sistema $\mathbf{Ax} = \mathbf{b}_0$.

Per il teorema precedente ciò accade se e solo se $\mathbf{A}^H\mathbf{A}\mathbf{A}^+\mathbf{b} = \mathbf{A}^H\mathbf{b}$. Ma $\mathbf{A}^H\mathbf{A}\mathbf{A}^+ = \mathbf{A}^H$ è una delle proprietà della matrice pseudo-inversa che segue dalla definizione (vedi [7]), da cui l'asserto.

Proviamo ora che, se $\mathbf{u} \in \mathbb{C}^n$ è un'altra soluzione di $\mathbf{Ax} = \mathbf{b}_0$ diversa da $\mathbf{A}^+\mathbf{b}$, allora $||\mathbf{u}||_2 > ||\mathbf{A}^+\mathbf{b}||_2$.

Osserviamo preliminarmente che $\mathbf{A}^+\mathbf{b} \in N(\mathbf{A})^\perp$, Infatti, se $\mathbf{w} \in N(\mathbf{A})$ risulta, per le proprietà della matrice pseudo-inversa che seguono dalla definizione (vedi [7])

$$\begin{aligned}\mathbf{w}^H\mathbf{A}^+\mathbf{b} &= \mathbf{w}^H\mathbf{A}^+\mathbf{A}\mathbf{A}^+\mathbf{b} = \mathbf{w}^H(\mathbf{A}^+\mathbf{A})^H\mathbf{A}^+\mathbf{b} = \mathbf{w}^H\mathbf{A}^H(\mathbf{A}^+)^H\mathbf{A}^+\mathbf{b} \\ &= (\mathbf{A}\mathbf{w})^H(\mathbf{A}^+)^H\mathbf{A}^+\mathbf{b} = 0.\end{aligned}$$

Osserviamo poi che, essendo $\mathbf{A}^+\mathbf{b}$ una soluzione del sistema $\mathbf{Ax} = \mathbf{b}_0$, ogni altra soluzione $\mathbf{u}$ è del tipo $\mathbf{u} = \mathbf{A}^+\mathbf{b} + \mathbf{y}$, dove $\mathbf{y} \in N(\mathbf{A})$. Ora, se $\mathbf{u} \neq \mathbf{A}^+\mathbf{b}$, risulta $\mathbf{y} \neq 0$. Poiché $\mathbf{y} \in N(\mathbf{A})$ e $\mathbf{A}^+\mathbf{b} \in N(\mathbf{A})^\perp$, si ha

$$||\mathbf{u}||_2^2 = ||\mathbf{A}^+\mathbf{b}||_2^2 + ||\mathbf{y}||_2^2 > ||\mathbf{A}^+\mathbf{b}||_2^2$$

da cui l'asserto. □

Dal teorema precedente e dalla definizione di norma euclidea, il vettore $\mathbf{A}^+\mathbf{b}$ si chiama *soluzione ai minimi quadrati* del sistema $\mathbf{Ax} = \mathbf{b}$.

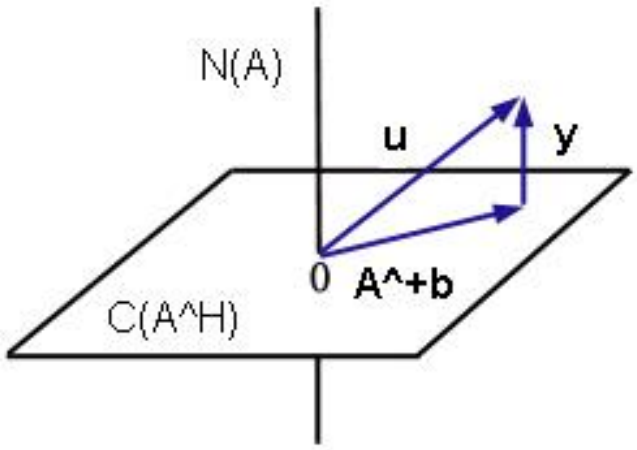

Figura A.2: Soluzione $\mathbf{u} = \mathbf{A}^{+}\mathbf{b} + \mathbf{y}$, dove $\mathbf{y} \in N(\mathbf{A})$, del sistema $\mathbf{Ax} = \mathbf{b}_0$.

Appendice B

Numero di Filogenie semi-etichettate

Si è visto al Capitolo 2 che è biologicamente sensato etichettare solamente le foglie di una filogenia; siamo allora interessati a contare il numero di filogenie in tal modo semi-etichettate, seguendo il ragionamento in [2].

Innanzitutto calcoliamo quanti sono gli alberi completamente etichettati con una fissata sequenza dei gradi. Osserviamo anzitutto che, per (2.1) e (2.3), la sequenza dei gradi $(g_i)_{1\leq i\leq n}$ interi positivi di un albero con n vertici deve soddisfare:

$$\sum_{i=1}^{n} g_i = 2n - 2 \tag{B.1}$$

Vale poi anche il viceversa.

Lemma B.1. *Sia $(g_i)_{1\leq i\leq n}$ una sequenza di interi positivi che soddisfi* (B.1)*. Allora esiste un albero con n vertici per il quale tale sequenza è la sequenza dei gradi.*

Dimostrazione. Si dimostra per induzione su n.

($n = 1$) Si tratta dell'albero costituito da un solo vertice senza archi.

($n \Rightarrow n+1$) Sia $g_1, \ldots, g_{n+1}$ una sequenza con

$$\sum_{i=1}^{n+1} g_i = 2(n+1) - 2 = 2n.$$

Non tutti i g_i possono essere uguali a 1, poiché altrimenti

$$\sum_{i=1}^{n+1} g_i = \sum_{i=1}^{n+1} 1 = n + 1 < 2n \qquad \text{se} n > 1.$$

Inoltre non tutti i g_i possono essere maggiori di 1, poiché altrimenti

$$2n = \sum_{i=1}^{n+1} g_i \geq \sum_{i=1}^{n+1} 2 = 2(n+1).$$

Dunque, senza perdita di generalità, possiamo assumere che $g_{n+1} = 1$ e $g_n > 1$. Definiamo poi $(g'_i)_{1 \leq i \leq n}$ così:

$$g'_i := \begin{cases} g_i & \text{per } 1 \leq i \leq n-1, \\ g_n - 1 & \text{per } i = n. \end{cases}$$

Per tale sequenza vale

$$\sum_{i=1}^{n} g'_i = \sum_{i=1}^{n-1} g_i + (g_n - 1) + (g_{n+1} - 1) = \sum_{i=1}^{n+1} g_i - 2 = 2n - 2.$$

Per l'ipotesi induttiva esiste un albero $T' = (V' = \{v_1, \ldots, v_n\}, E')$ tale che $d(v_i) = g'_i$. Allora l'albero $T = (V' \cup \{v_{n+1}\}, E' \cup \{v_n v_{n+1}\})$ soddisfa l'asserto.

□

Possiamo allora ottenere il valore voluto, enunciato di seguito.

Teorema B.2. *Sia $n \geq 2$ e $(g_i)_{1 \leq i \leq n}$ una sequenza di interi positivi e si denoti con $t(n, g_1, \ldots, g_n)$ il numero di differenti alberi etichettati $T = (\{v_1, \ldots, v_n\}, E)$ di n vertici con la sequenza dei gradi*

$$d_T(v_i) = g_i \ \forall 1 \leq i \leq n.$$

Allora

$$t(n, g_1, \ldots, g_n) = \frac{(n-2)!}{\prod_{i=1}^{n}(g_i - 1)!} \tag{B.2}$$

se (B.1) *vale, e*

$$t(n, g_1, \ldots, g_n) = 0$$

altrimenti.

Dimostrazione. Per il teorema precedente $t(n, g_1, \ldots, g_n) > 0$ vale se e solo se (B.1) è soddisfatta.

Senza perdita di generalità, possiamo assumere che

$$g_1 \geq g_2 \geq \cdots \geq g_n.$$

Allora v_n deve essere una foglia. Sia C_i l'insieme di tutti gli alberi T con vertici $v_1, \ldots, v_n$ e gradi $g_j = d_T(v_j)$, tali che la foglia v_n sia adiacente a v_i. Assumendo $g_i \geq 2$ abbiamo

$$|C_i| = t(n-1, g_1, \ldots, g_{i-1}, g_i - 1, g_{i+1}, \ldots, g_{n-1}).$$

Poiché l'insieme di tutti gli alberi è l'unione degli insiemi C_i per $g_i \geq 2$ otteniamo

$$t(n, g_1, \ldots, g_n) = \sum_{g_i \geq 2} t(n-1, g_1, \ldots, g_{i-1}, g_i - 1, g_{i+1}, \ldots, g_{n-1}).$$

Usiamo ora l'induzione. Il teorema è vero per $n = 2$. Si assuma che $n \geq 3$ e che il teorema sia vero per $n - 1$. Allora

$$\begin{aligned}
t(n, g_1, \ldots, g_n) &= \sum_{g_i \geq 2} t(n-1, g_1, \ldots, g_{i-1}, g_i - 1, g_{i+1}, \ldots, g_{n-1}) \\
&= \sum_{g_i \geq 2} \frac{(n-3)!}{(g_1-1)! \cdots (g_{i-1}-1)!(g_i-2)!(g_{i+1}-1)! \cdots (g_{n-1}-1)!} \\
&= \frac{(n-3)!}{(g_{a+1}-1)! \cdots (g_{n-1}-1)!} \cdot \\
&\quad \left(\sum_{i=1}^{a} \frac{1}{(g_1-1)! \cdots (g_{i-1}-1)!(g_i-2)!(g_{i+1}-1)! \cdots (g_a-1)!} \right) \\
&= \frac{(n-3)!}{(g_{a+1}-1)! \cdots (g_{n-1}-1)!} \cdot \frac{\sum_{i=1}^{a} g_i - a}{(g_1-1)! \cdots (g_a-1)!} \\
&= \frac{((2n-2-(n-a))-a)(n-3)!}{(g_1-1)! \cdots (g_{n-1}-1)!} \\
&= \frac{(n-2)!}{(g_1-1)! \cdots (g_n-1)!}
\end{aligned}$$

dove la prima uguaglianza è stata dimostrata sopra, la seconda è l'ipotesi induttiva, nella terza si assume che $v_1, \ldots, v_a$ siano i vertici interni con $g_i \geq 2$, la quinta segue da (B.1) e la sesta dal fatto che v_n deve essere una foglia. □

Utilizzando il risultato precedente si può allora contare il numero di filogenie semi-etichettate, enunciato di seguito.

Teorema B.3. *a) Il numero di alberi binari con n foglie etichettate e $n-2$ vertici interni etichettati è*

$$\frac{(2n-4)!}{2^{n-2}}$$

b) (Cavalli-Sforza, Edwards) Il numero di alberi binari con n foglie etichettate e $n-2$ vertici interni non etichettati (cioè il numero di filogenie su n taxa) è

$$(2n-5)!! := 1 \cdot 3 \cdot 5 \cdots (2n-5) \tag{B.3}$$

Dimostrazione. a) Si ha, per (B.2)

$$\binom{(2n-2)-2}{\underbrace{(1-1)\dots(1-1)}_{n\text{volte}}\underbrace{(3-1)\dots(3-1)}_{n-2\text{volte}}} = \frac{(2n-4)!}{2^{n-2}}$$

dove a primo membro si ha il cosiddetto coefficiente multinomiale, così definito

$$\binom{n}{k_1 \dots k_m} := \frac{n!}{k_1! \cdots k_m!}.$$

b) Se i vertici interni non sono etichettati, allora il numero ottenuto in a) deve essere diviso per $(n-2)!$, da cui

$$\begin{aligned}\frac{(2n-4)!}{2^{n-2}(n-2)!} &= \frac{(2n-4)(2n-5)(2n-6)(2n-7)\cdots 4\cdot 3\cdot 2\cdot 1}{2(n-2)2(n-3)\cdots 2\cdot 2\cdot 2\cdot 1} \\ &= (2n-5)(2n-7)(2n-9)\cdots 5\cdot 3\cdot 1 \\ &= (2n-5)!!\end{aligned}$$

□

Bibliografia

[1] D. Catanzaro, *The minimum evolution problem: overview and classification*, Networks 53 (2009), 112–25.

[2] D. Cieslik, *Counting phylogenetic trees*, University of Greifswald.

[3] M. Farach, S. Kannan, and T. Warnow, *A Robust Model for Finding Optimal Evolutionary Trees*, Algorithmica (1995) 13:155–179.

[4] Amitabh Sinha, *Metric Embeddings and Methods, R. Ravi, Lecture 12*, Note, (2003)

[5] A. Rzhetsky and M. Nei, *Statistical properties of the ordinary least-squares generalized least-squares and minimum evolution methods of phylogenetic inference*, J Mol Evol 35 (1992), 367-375.

[6] M. Bulmer, *Estimating the variability of substitution rates*, Genetics 123:615–619.

[7] E. Gregorio, L. Salce, *Algebra Lineare*, Edizioni Libreria Progetto Padova (2010).

[8] R. Desper and O. Gascuel, *Theoretical foundations of the balanced minimum evolution method of phylogenetic inference and its relationship to the weighted least-squares tree fitting*, Mol Biol Evol 21 (2004), 587-598.

[9] A. Rzhetsky and M. Nei, *A Simple Method for Estimating and Testing Minimum-Evolution Trees*, Mol Biol Evol 9:945-967.

[10] A. Rzhetsky and M. Nei, *Theoretical foundations of the minimum evolution method of phylogenetic inference*, Mol Biol Evol 10 (1993), 1073-1095.

[11] Y. Pauplin, *Direct calculation of a tree length using a distance matrix*, J Mol Evol 51 (2000), 41-47.

[12] R. Desper and O. Gascuel, *Theoretical foundations of the balanced minimum evolution method of phylogenetic inference and its relationship to the weighted least-squares tree fitting*, Mol Biol Evol 21 (2004), 587-598.

[13] R. Aringhieri, D. Catanzaro, M. Di Summa, *Optimal solutions for the balanced minimum evolution problem*, Computers & Operations Research 38 (2011), 1845–1854.

[14] R. E. Burkard, E. Çela, P. M. Pardalos, L. S. Pitsoulis, *The Quadratic Assignment Problem*, Handbook of Combinatorial Optimization (1997).

[15] E. Çela, *The Quadratic Assignment Problem- Theory and Algorithms*, Springer-Science+Business Media, B.V. (1998).

[16] M. Di Summa, *Programmazione Lineare Intera*, Dispense, Università di Padova.

[17] D. Catanzaro, M. Labbé, R. Pesenti, J. J. Salazar-González, *The Balanced Minimum Evolution Problem*, INFORMS Journal on Computing (2011), 1–19.

[18] Parker, D. S., P. Ram, *The construction of Huffman codes is a submodular (convex) optimization problem over a lattice of binary trees*, SIAM J. Comput. (1996), 28(5) 1875-1905.

[19] P. Buneman, *A note on the metric properties of trees*, J. Combin. Theory (1974), Ser. B 17 48-50.

Ringraziamenti

Ringrazio il prof. Di Summa per l'assistenza nella redazione e i chiarimenti. Un ringraziamento inoltre alla mia famiglia per il supporto.