Applicazioni della geometria algebrica alla biologia

Ciro Ciliberto, Enrico Rogora.

Introduzione

In questo articolo vogliamo dare un'idea di *nuove* interazioni tra matematica e biologia. In verità non sono mancate nel passato applicazioni importanti della matematica alla biologia. Si pensi al modello di Lotke-Volterra per la competizione tra le specie [47, 29], ai modelli differenziali per la dinamica delle popolazioni [22, 25, 50], ecc. Si tratta di modelli analitici deterministici ispirati a quelli della meccanica classica. Come tali non sono *propri* della biologia e mostrano limiti anche gravi quando vengono applicati a situazioni reali. Probabilmente questa è una delle ragioni che ha spinto Gian Carlo Rota ad affermare [38], p. 2:

> *La mancanza di contatto reale tra la matematica e la biologia è una tragedia o uno scandalo o una sfida: è difficile da decidere.*
>
> G. C. Rota

Negli ultimi decenni l'affermazione della biologia molecolare ha sollecitato l'uso di strumenti matematici del tutto diversi. Innanzitutto si è imposta una analisi probabilistica di molti fenomeni biologico molecolari. Queste tecniche probabilistiche, per la natura stessa dei problemi trattati, hanno carattere discreto e un contenuto combinatorico e algebrico. Questo ha richiesto l'uso e in alcuni casi la creazione di raffinate tecniche algebrico-geometriche. Si tratta dunque di un quadro completamente nuovo ed estremamente stimolante che giustifica il titolo di un recente articolo di J.E. Cohen [14]:

> *La Matematica è il prossimo microscopio della biologia, ed è migliore. La biologia è la prossima fisica per la matematica, ed è migliore.*
>
> J. E. Cohen

Come è ben noto le esigenze della fisica teorica contemporanea hanno reso necessario lo sviluppo di tecniche matematiche tanto innovative e raffinate da far sì che i loro creatori, in origine fisici, venissero insigniti della medaglia Fields, Possiamo chiederci, in modo forse un po' provocatorio con Sturmfels, [45],

> Sarà mai possibile che un biologo possa vincere la medaglia Fields?

1 Genetica e biologia molecolare

L'ambito in cui, a nostro avviso, si sono sviluppate le tecniche cui abbiamo accennato precedentemente è quello della *genetica* e della *biologia molecolare*. Ci sembra opportuno richiamare alcuni concetti basilari di queste discipline.

La *genetica* nasce intorno al 1865 con i lavori di *Gregor Mendel* che postulò l'esistenza di unità discrete di informazione, i *geni*, che governano la trasmissione delle caratteristiche individuali in un organismo.

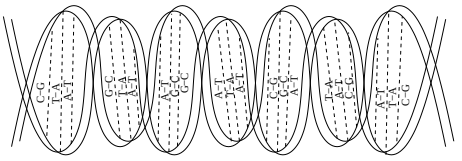
Gregor Mendel (1822-1884)

Nella prima metà del ventesimo secolo si capì che i geni sono contenuti nelle macromolecole complesse di *acido deossiribonucleico* (DNA) situate in strutture chiamate cromosomi presenti in ogni cellula vivente.

Nel 1953 *J. Watson e F. Crick* determinarono la struttura del DNA.

James Watson (n. 1928) e Francis Crick (1916-2004), vincitori del premio Nobel nel 1962.

Secondo questo modello il DNA è composto da due catene avvolte a spirale a formare una doppia elica.



Rappresentazione schematica della doppia elica del DNA.

L'informazione genetica contenuta nel DNA è codificata in una successione di *basi azotate.* Le basi azotate sono

Adenina, *Citosina*, *Guanina*, *Timina*.

Le basi azotate di un elica determinano quelle dell'altra. Infatti ogni base su un catena è legata ad una *base complementare* sull'altra. Le coppie complementari sono (C, G) e (T, A).

Dal punto di vista genetico ogni *specie* viene descritta dal suo DNA, la cui struttura primaria si modella come una *stringa sull'alfabeto*

$$\boxed{\Omega = \{A, C, G, T\}}$$

ll contenuto totale delle molecole di DNA entro i cromosomi costituisce il *genoma* di un organismo. Il genoma umano, formato da circa 3 miliardi di coppie di basi complementari, corrisponde a circa 700 megabytes di informazione, *quanta ne può essere memorizzata in un CD Rom.* Il genoma di due individui della stessa specie non è identico: le differenze per gli esseri umani sono di circa *una base ogni mille*, sufficienti a spiegare la variabilità tra i diversi individui.

Alcuni tratti del genoma codificano degli elementi fondamentali per la vita cioè le *proteine*, le quali sono i *mattoni* di ogni edificio biologico. Le loro funzioni principali consistono nel catalizzare le reazioni chimiche, nel regolare le attività cellulari, nel mediare le comunicazioni tra le cellule.

Dal punto di vista biochimico una *proteina* è una catena di *amminoacidi*, codificate nel DNA in tratti detti *geni*.

Ogni cellula (escluse le cellule uovo e gli spermatozoi) contiene copia dell'intero genoma ed è quindi teoricamente in grado di produrre ogni proteina, tuttavia essa in genere produce solo le proteine connesse con le sue funzioni.

Vi sono VENTI amminoacidi *principali*, cioè quelli presenti in maggioranza in tutte le specie e sono indicati nella Tabella seguente

A	*ALA*	Alanina	*V*	*VAL*	Valina	*L*	*LEU*	Leucine
I	*ILE*	Isoleucina	*F*	*PHE*	Fenilanina	*P*	*PRO*	Prolina
M	*MET*	Metionina	*D*	*ASP*	Acido Aspartico	*E*	*GLU*	Acido Glutammico
K	*LYS*	Lisina	*R*	*ARG*	Arginina	*S*	*SER*	Serina
T	*THR*	Treonina	*C*	*CYS*	Cisteina	*N*	*ASN*	Asparagina
Q	*GLU*	Glutamina	*H*	*HIS*	Histidina	*Y*	*TYR*	Tirosina
W	*TRP*	Triptopane	*G*	*GLY*	Glicina.			

Tabella degli amminoacidi

Ogni amminoacido viene codificato da una *tripletta di basi azotate adiacenti* detta *codone*. Il numero delle possibili triplette *(64)* è più grande del

numero degli amminoacidi *(20)*. Quindi diversi codoni codificano lo stesso amminoacido.

Due codoni che codificano la stesso amminoacido si dicono *sinonimi*. Due codoni sinonimi differiscono generalmente solo nella terza base della tripletta.

Tra i codoni ne esistono tre, TAG, TAA, TGA, che non codificano alcun amminoacido ma *marcano la fine di una proteina*, e uno, ATG che codifica la *Metionina*, che *marca l'inizio di ogni proteina*.

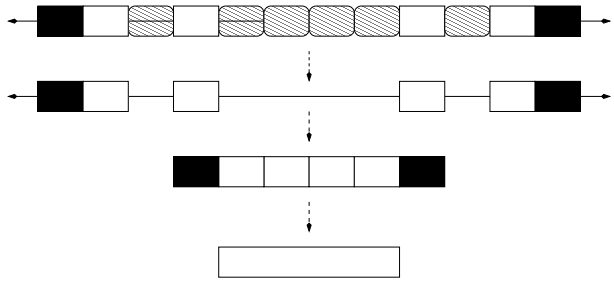
Un cambiamento di una base in un gene che induce una modifica non sinonima del corrispondente codone altera la proteina e può portare ad una malattia genetica.

	T	C	A	G
T	TTT ↦ Phe	TCT ↦ Ser	TAT ↦ Tyr	TGT ↦ Cys
	TTC ↦ Phe	TCC ↦ Ser	TAC ↦ Tyr	TGC ↦ Cys
	TTA ↦ Leu	TCA ↦ Ser	TAA ↦ *stop*	TGA ↦ *stop*
	TTG ↦ Leu	TCG ↦ Ser	TAG ↦ *stop*	TGG ↦ Trp
C	CTT ↦ Leu	CCT ↦ Pro	CAT ↦ His	CGT ↦ Arg
	CTC ↦ Leu	CCC ↦ Pro	CAC ↦ His	CGC ↦ Arg
	CTA ↦ Leu	CCA ↦ Pro	CAA ↦ Gln	CGA ↦ Arg
	CTG ↦ Leu	CCG ↦ Pro	CAG ↦ Gln	CGG ↦ Arg
A	ATT ↦ Ile	ACT ↦ Thr	AAT ↦ Asn	AGT ↦ Ser
	ATC ↦ Ile	ACC ↦ Thr	AAC ↦ Asn	AGC ↦ Ser
	ATA ↦ Ile	ACA ↦ Thr	AAA ↦ Lys	AGA ↦ Arg
	ATG ↦ Met	ACG ↦ Thr	AAG ↦ Lys	AGG ↦ Arg
G	GTT ↦ Val	GCT ↦ Ala	GAT ↦ Asp	GGT ↦ Gly
	GTC ↦ Val	GCC ↦ Ala	GAC ↦ Asp	GGC ↦ Gly
	GTA ↦ Val	GCA ↦ Ala	GAA ↦ Glu	GGA ↦ Gly
	GTG ↦ Val	GCG ↦ Ala	GAG ↦ Glu	GGG ↦ Gly

Tabella di conversione codoni - amminoacidi.

I *geni* sono sottoinsiemi di segmenti del DNA che contengono l'informazione necessaria alla costruzione delle *proteine*. Nel genoma umano ci sono circa *30.000* geni.

Nel processo di formazione di una proteina negli organismi *eucarioti*, cioè quelli in cui le cellule sono dotate di nucleo, alcuni codoni all'interno del gene, vengono eliminati (cfr. la figura seguente). Essi sono detti *introni*. I rimanenti codoni, detti *esoni*, codificano gli amminoacidi che costituiscono la proteina.



Rimozione degli introni

Solo una piccola parte del genoma umano, circa l' 1.2%, sembra codificare proteine. Non sono ancora chiare le funzioni della parte restante. Tuttavia è interessante osservare che anche in questa parte devono essere registrate delle informazioni importanti. Infatti si osserva la presenza di tratti non codificanti si ritrovano esattamente uguali in specie diverse ma aventi delle affinità. Per esempio, sono noti circa 500 segmenti più lunghi di 200bp detti *sequenze ultraconservate*, comuni ai genomi dell'uomo, del topo e del ratto; Inoltre è noto *almeno* un segmento, detto MEANING OF LIFE SEQUENCE, di lunghezza 42, comune a 10 specie di vertebrati tra cui l'uomo [34]:

$$\boxed{TTTAATTGAAAGAAGTTAATTGAATGAAAATGATCAACTAAG}$$

La probabilità della meaning of life sequence, calcolata in base ai modelli matematici di evoluzione usualmente impiegati e che descriveremo nel seguito, non supera 10^{-50} [34].

Come osserva E. Borel, in [8]

> Un fenomeno la cui probabilità è 10^{-50} NON ACCADRÀ MAI,
> o, quanto meno NON SARÀ MAI OSSERVATO.

Quindi i modelli matematici per l'evoluzione delle sequenze biologiche devono essere ancora raffinati per descrivere adeguatamente questi fenomeni.

2 Modelli deterministici e modelli probabilistici.

Nel corso del novecento si sono affermati, a complemento o in alternativa ai *modelli deterministici* classici, *modelli probabilistici* per la descrizione di fenomeni fisici (*meccanica statistica* e *meccanica quantistica*). Anche in biologia, come in fisica, ci sono modelli matematici classici di tipo deterministico per l'interpretazione di alcuni fenomeni, si pensi al famoso modello di Volterra-Lotka. Per quanto raffinati, questi modelli si sono tuttavia rivelati di scarsa capacità sia descrittiva che predittiva (cfr. [26]) Per contro, l'uso di modelli probabilistici risulta di grande utilità pratica anche in biologia in forza delle loro capacità descrittive e predittive. Infatti essi trovano ogni giorno impiego nei laboratori di ricerca. Di alcuni di questi modelli, impiegati nella descrizione dell'evoluzione delle sequenze biologiche, vogliamo parlare.

Supponiamo di osservare sequenze di dati dicotomici. Per esempio, il sesso dei nati in un certo periodo, o la successione di introni/esoni in un gene. Vogliamo costruire un modello probabilistico per la generazione di questi dati. Il più semplice è quello del *lancio di una moneta*.

Modello del lancio della moneta Supponiamo di avere una moneta con testa (T) e croce (C) e supponiamo che l'esito di ogni lancio sia indipendenti dagli altri. Sia p la probabilità di osservare testa in un lancio ($0 \leq p \leq 1$). La probabilità di osservare croce è quindi $1 - p$. Se $p \neq \frac{1}{2}$ la moneta è truccata. Sulla base di questo modello siamo in grado di valutare la probabilità di ogni evento. Per esempio la probabiità di osservare k teste e $n - k$ croci è $\binom{n}{k} p^k (1-p)^{n-k}$.

Sulla base di questo modello ci possiamo proporre di *stimare* p, e quindi capire se la moneta è truccata. Per far ciò si può applicare il principio di *massima verosimiglianza*.

Lanciamo ripetutamente la moneta ottenendo ad esempio

$$\boxed{TCCTCTTCTCCTTT}$$

La probabilità di osservare 8 teste e 6 croci in 14 lanci, come nella sequenza proposta, è

$$\boxed{L(p) = p^8(1-p)^6.}$$

La *stima di massima verosimiglianza* di p è il valore che rende massimo $L(p)$, nell'intervallo $[0, 1]$. Nell'esempio vale $\frac{4}{7}$.

Nei modelli probabilistici per la descrizione di sequenze biologiche il dato osservato è *una successione di lettere dell'alfabeto*

$$\Omega = \{A, C, G, T\}.$$

Il modello probabilistico più semplice per la produzione di queste sequenze è il *modello dell'urna*.

Modello dell'urna Un'urna contiene n_A palline marcate con A, e analogamente per n_C n_G e n_T. Per generare una sequenza si pensi di: *estrarre una pallina*, *annotare la marca*, *rimettere la pallina nell'urna*, *agitare bene e ripetere*.

Questo modello è caratterizzato dai seguenti parametri: la probabilità

$$\boxed{p_A = \frac{n_A}{n_A + n_C + n_G + n_T}}$$

di estrarre una pallina contrassegnata con A, e analogamente per le altre p_C, p_G e p_T. Si osservi che $\boxed{p_A + p_C + p_G + p_T = 1}$, quindi il modello dell'urna dipende da *tre parametri essenziali*.

Data una sequenza biologica ci poniamo due problemi.

1. *Stimare i parametri* p_A, p_C, ecc. nell'ipotesi che la generazione della sequenza sia descrivibile con il meccanismo dell'urna.

2. *Valutare l'adeguatezza* del modello.

La stima dei parametri si effettua utilizzando il principio di massima verosimiglianza, che abbiamo discusso in precedenza.

Per valutare l'adeguatezza del modello esistono diversi metodi statistici che non tratteremo [2].

Il modello dell'urna è troppo semplicistico in quanto prevede l'indipendenza di ogni carattere dagli altri e questo non è biologicamente plausibile: per esempio, come abbiamo visto parlando dei codoni, le basi azotate che appaiono nella terza posizione di ogni codone hanno una probabilità di cambiare significativamente maggiore delle altre in quanto producono modificazioni sinonime.

Catene di Markov Un modello più complesso funziona così. Si considerino *quattro urne*, la prima marcata con A, la seconda con G, ecc. e una *quinta urna* marcata con I (per inizializzazione).

Nell'urna A, il numero della palline marcate A è $n_{A,A}$, quelle marcate con C è $n_{A,C}$ ecc. L' urna I, come nell'esempio precedente, contiene n_A palline marcate con A, ecc.

Il *processo di generazione* di una sequenza consiste nei passi seguenti:

- *Inizializzazione:* estrarre un pallina dall'urna I;
- *Iterazione: estrarre una pallina dall'urna contrassegnata dalla marca pescata al passo precedente*, *annotare la marca*, *rimettere la pallina nell'urna e mescolare*, *ripetere il passo di iterazione.*

Questo nuovo modello introduce un *meccanismo di dipendenza* nella generazione della sequenza. L'estrazione di una marca infatti *dipende da quella estratta al passo precedente e solo da quella.* Si tratta di una *catena di Markov.*

I parametri da cui dipende il modello sono le probabilità iniziali, p_A, p_C, p_G e p_T e le probabilità

$$\boxed{p_{X,Y} = \frac{n_{X,Y}}{n_{X,A} + n_{X,C} + n_{X,G} + n_{X,T}}}$$

di estrarre una pallina contrassegnata con Y pescando dall'urna contrassegnata con X.

Questo modello dipende dunque da 15 parametri (3 per ogni urna). La stima dei parametri e la valutazione di adeguatezza si affrontano come per il modello precedente.

Modelli a stati nascosti Nella pratica c'è spesso l'esigenza di introdurre, in un modello probabilistico, *stati nascosti*, cioè stati *non osservabili*, da cui dipendono le osservazioni. Ad esempio l'esito *osservato* nelle prove d'esame dipende dall' umore *non osservabile* dell'esaminatore.

Questi modelli vengono introdotti con lo scopo principale di *stimare gli stati nascosti* a fronte degli stati osservati.

Un semplice modello a stati nascosti considera *due urne*, $\mathcal{T}$ e $\mathcal{C}$, ciascuna contenente palline marcate A, C, G, T e *una moneta* con due facce marcate $\mathcal{T}$ e $\mathcal{C}$. La probabilità $p(X, \mathcal{Y})$ di estrarre la marca X dall'urna $\mathcal{Y}$ è funzione dei numeri $n_{X,\mathcal{Y}}$ di palline marcate X nell'urna $\mathcal{Y}$.

I parametri da cui dipende il modello sono le probabilità di estrazione dalle urne e la probabilità che esca $\mathcal{T}$ o $\mathcal{C}$ nel lancio della moneta: di questi 10 parametri solo 7 sono indipendenti.

In base a questo modello una sequenza si genera così: *si lancia la moneta, si pesca una pallina dalla corrispondente urna, si annota la marca, si rimette la pallina nell'urna e si rimescola, si ripete.*

Questo modello descrive un *processo visibile*, quello che produce le marche, basato su un *processo non osservabile*, il lancio della moneta. Esso non è adeguato alla descrizione di sequenze biologiche in quanto gli eventi del processo nascosto sono indipendenti.

Catene di Markov nascoste Per introdurre una forma di dipendenza tra gli stati nascosti possiamo modellarli a loro volta come una *catena di Markov*.

Per esempio, consideriamo due piatti contrassegnati $\mathcal{T}$ e $\mathcal{C}$. Su ogni piatto c'è un'urna contrassegnata con lo stesso simbolo del piatto, analoga a quella del precedente modello a stati nascosti e una moneta con i simboli $\mathcal{T}$ e $\mathcal{C}$ sulle facce. Le monete sui due piatti sono truccate in maniera diversa.

C'è infine una terza moneta con gli stessi simboli sulle facce per inizializzare il processo.

In base a questo modello una sequenza si genera così: *Inizializzazione:* Si lancia la terza moneta per scegliere il piatto da cui cominciare. *Iterazione:* Si pesca una pallina dal piatto *corrente*, *si annota la marca*, *si lancia la moneta sul piatto corrente per scegliere il nuovo piatto*, *si ripete.*

Questo processo, che dipende da 9 parametri indipendenti, è una *catena di Markov nascosta* e costituisce, come vedremo nell'esempio discusso a p. 13, un utile modello per descrivere alcuni aspetti delle sequenze biologiche. Un riferimento standard per le applicazioni dei modelli probabilistici all'analisi delle sequenze biologiche è [15].

In generale, una catena di Markov nascosta è descritta dai dati seguenti:

1. L'alfabeto $N = \{n_1, \ldots, n_h\}$ degli *stati nascosti.*

2. L'alfabeto $V = \{v_1, \ldots, v_k\}$ dei *simboli visibili.*

3. Il vettore $p = (p_1, \ldots, p_h)$ delle *probabilità iniziali*: p_i è la probabilità che lo stato iniziale sia n_i.

4. La matrice $T = (t_{ij})$ di *transizione* tra gli stati nascosti: t_{ij} è la probabilità di passare dallo stato n_i allo stato n_j.

5. La matrice $E = (e_{is})$ di *emissione*: e_{is} è la probabilità che lo stato n_i emetta il simbolo v_s.

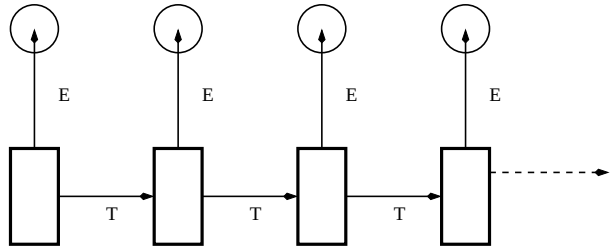
Il meccanismo di generazione di una sequenza di simboli visibili è il seguente.

1. Viene prodotto uno *stato nascosto iniziale* x_1 mediante il lancio di una "moneta" con h facce, con probabilità descritte dal vettore p.

2. Il *primo simbolo visibile* y_1 viene prodotto a partire da x_1 pescando da un'urna opportuna con probabilità di estrazione data dalla riga di E corrispondente a x_1.

3. Il *nuovo stato nascosto* x_2 viene prodotto a partire da x_1 lanciando una "moneta" con h facce con probabilità descritta dalla riga di T corrispondente a x_1, e così via.

Questo meccanismo può essere raffigurato mediante il grafo nella Figura seguente.



Grafo associato ad una catena di Markov nascosta.

In base a questo modello è possibile calcolare la probabilità di ogni evento in funzione dei parametri del modello.

Data una successione degli *stati nascosti*

$$\sigma = (\sigma_1, \ldots, \sigma_n) \qquad \sigma_i \in N$$

ed una di *stati visibili*

$$\tau = (\tau_1, \ldots, \tau_n) \qquad \tau_j \in V$$

la *probabilità* di osservare τ in corrispondenza di σ è

$$p_{\sigma\tau} = p_{\sigma_1} e_{\sigma_1\tau_1} t_{\sigma_1\sigma_2} e_{\sigma_2\tau_2} t_{\sigma_2\sigma_3} e_{\sigma_3\tau_3} \ldots t_{\sigma_{n-1}\sigma_n} e_{\sigma_n\tau_n}$$

che è un *monomio* nei parametri del modello.

Quindi la *probabilità* di osservare τ *qualunque siano gli stati nascosti* è il *polinomio*

$$\boxed{p_\tau = \sum_{\sigma \in N^n} p_{\sigma\tau}.}$$

Un'applicazione: riconoscimento vocale Le catene di Markov nascoste costituiscono una classe di processi stocastici che hanno numerose applicazioni pratiche. Storicamente la prima è quella relativa al *riconoscimento vocale* (cfr,[37]). In questo caso gli insiemi N e V coincidono con l'insieme dei *fonemi di una lingua*. Il dato osservabile è la successione $y_1, \dots y_n \in V$ dei *fonemi registrati* da un riconoscitore vocale. Il dato nascosto è la successione $x_1, \dots, x_n \in N$ dei *fonemi emessi* da riconoscere.

La successione $y_1, \dots, y_n$ *non coincide* in generale con $x_1, \dots, x_n$ a causa della scarsa affidabilità del riconoscimento dei fonemi. Il problema basilare è quello di ricostruire la successione degli stati nascosti.

Per fare ciò questa modellizzazione è molto efficace. La matrice di transizione T è determinata dalle caratteristiche fonetiche della lingua e la matrice di emissione E dalle caratteristiche tecniche del riconoscitore vocale.

Data la successione $y_1, \dots, y_n$ dei fonemi riconosciuti, per determinare la successione $x_1, \dots, x_n$ del discorso da riconoscere si applica il principio di massima verosimiglianza, ovvero si *massimizza* la probabilità $p(x_1, y_1, \dots, x_n, y_n)$ calcolata precedentemente.

Per determinare la successione $x_1, \dots, x_n$ nei tempi necessari alle applicazioni pratiche è necessario sviluppare algoritmi *iterativi* efficienti che evitino di passare in rassegna tutte le possibili successioni $x_1, \dots, x_n$. Esistono algoritmi di questo tipo, ad esempio l' *Algoritmo di Viterbi* [37].

Un'altra applicazione: il riconoscimento dei geni Lo stesso modello probabilistico si può applicare al riconoscimento dei geni. Qui l'insieme N è costituito da due elementi $\{I, E\}$ dove I sta per *Introne* ed E per *Esone*. L'insieme V coincide con l'insieme dei 64 codoni. La successione $y_1, \dots y_n \in V$ è quella dei codoni osservati in un tratto di DNA. La successione $x_1, \dots, x_n \in N$ è l'*annotazione* della sequenza in introni ed esoni.

Anche qui vogliamo ricostruire la successione $x_1, \dots, x_n$ a partire da $y_1, \dots, y_n$. Questo è fondamentale per riconoscere la proteina codificata

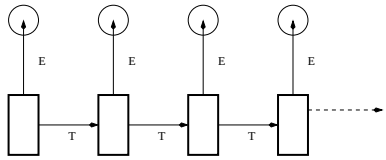
da un gene. Questo problema si affronta con gli stessi metodi discussi nell'esempio precedente.

3 Modelli grafici e statistica algebrica

La formula

$$p_\tau = \sum_{\sigma \in N^n} p_{\sigma\tau}$$

e il grafo



insieme alle matrici T, E e al vettore delle probabilità iniziali, sintetizzano in maniera equivalente la struttura del modello di Markov a stati nascosti.

Più in generale, esiste una ampia classe di modelli probabilistici discreti, i cosiddetti *modelli grafici*, in cui aspetti *algebrici* e *combinatorici* interagiscono in maniera analoga prestandosi anche ad utili e suggestive interpretazioni geometriche. Discutiamo alcuni esempi.

Il modello di indipendenza Consideriamo il grafo

A ciascuno dei due vertici assegnamo un alfabeto di due simboli $\{E, I\}$ e le probabilità $p_E^{(i)}, p_I^{(i)}$ di osservare E oppure I nel vertice i.

Questo modello assegna alle quattro possibili osservazioni le probabilità, date da *monomi*, indicate nella seguente tabella

EE	EI	IE	II
$p_E^{(1)} p_E^{(2)}$	$p_E^{(1)} p_I^{(2)}$	$p_I^{(1)} p_E^{(2)}$	$p_I^{(1)} p_I^{(2)}$

Lo spazio delle distribuzioni di probabilità sulle possibili osservazioni $\{EE, EI, IE, II\}$ è l'insieme Δ delle quaterne (x_0, x_1, x_2, x_3) di numeri reali tali che

$$0 \leq x_i \leq 1 \qquad i = 0, \ldots, 3; \qquad x_0 + x_1 + x_2 + x_3 = 1.$$

Il modello di indipendenza seleziona in Δ il sottoinsieme dato dalla *equazione algebrica*

$$\boxed{x_0 x_3 - x_1 x_2 = 0.}$$

Nella sua versione più generale il modello di indipendenza è associato al grafo con m vertici

Al vertice i è associato un alfabeto di $n_i + 1$ simboli $a_j^{(i)}$, $j = 0, \ldots, n_i$, e le probabilità $p_j^{(i)}$ di osservare $a_j^{(i)}$ in tale vertice.

Il modello di indipendenza assegna all'osservazione $a_{i_1}^{(1)}, \ldots, a_{i_m}^{(m)}$ la probabilità data dal monomio

$$\boxed{p(i_1, \ldots, i_m) = p_{i_1}^{(1)} \cdot \cdots \cdot p_{i_m}^{(m)}.}$$

Lo spazio delle distribuzioni di probabilità sulle possibili osservazioni è l'insieme Δ delle $N+1$-ple

$$(x_{i_1 \ldots i_m}), \qquad i_j = 0, \ldots, n_j, \quad j = 1, \ldots, m$$

dove

$$N + 1 = (n_1 + 1) \cdot \cdots \cdot (n_m + 1)$$

verificanti le condizioni

$$0 \leq x_{i_1 \ldots i_m} \leq 1, \qquad \sum_{i_1 \ldots i_m} x_{i_1 \ldots i_m} = 1. \tag{1}$$

Il modello di indipendenza seleziona in Δ il sottoinsieme dei punti

$$x_{i_1 \dots i_m} = p_{i_1}^{(1)} \cdot \dots \cdot p_{i_m}^{(m)} \tag{2}$$

al variare dei parametri $p_j^{(i)}$.

Tali punti sono le soluzioni di un *sistema di equazioni algebriche* omogeneeche definiscono *varietà algebriche* note col nome di *prodotti Segre*. Le loro parametrizzazioni (2) sono noto come *mappe di Segre*.

È opportuno e naturale cominciare lo studio di questi oggetti sul campo complesso.

Le mappe di Segre Le mappe di Segre hanno la loro naturale collocazione nell'ambito della *geometria proiettiva*. Lo *spazio proiettivo numerico* $\mathbb{P}^n$ su $\mathbb{C}$ è definito come il quoziente di $\mathbb{C}^{n+1} \setminus \{0\}$ rispetto alla relazione di proporzionalità tra vettori numerici. Quindi un punto di $\mathbb{P}^n$ è determinato da un vettore non nullo $(x_0, \dots, x_n)$ e quindi da tutti e soli i vettori del tipo $(\lambda x_0, \dots, \lambda x_n)$ con λ diverso da zero. Ogni tale $n+1$-pla si dice una $n+1$-pla di *coordinate omogenee* del punto. Ciò si denota scrivendo che il punto ha coordinate omogenee $[x_0, \dots, x_n]$.

Lo spazio proiettivo $\mathbb{P}^n$ è ottenuto aggiungendo allo *spazio affine numerico* $\mathbb{C}^n$ i *punti all'infinito*. Più precisamente, $\mathbb{C}^n$ si identifica con il sottoinsieme dei punti $[1, x_1, \dots, x_n]$ di $\mathbb{P}^n$. l'insieme complementare è costituito dai punti del tipo $[0, x_1, \dots, x_n]$. Un tale punto si identifica con il punto *all'infinito* delle rette parallele al vettore $x_1, \dots, x_n$.

Mentre il prodotto di due spazi affini è uno spazio affine, il prodotto di due spazi proiettivi non è uno spazio proiettivo. Tuttavia esso si immerge in modo naturale in un opportuno spazio proiettivo *più grande* mediante la *mappa di Segre*. Siano infatti $\mathbb{P}^n$ e $\mathbb{P}^m$ due spazi proiettivi. La mappa di Segre

$$s_{m,n} : \mathbb{P}^m \times \mathbb{P}^n \to \mathbb{P}^{mn+m+n}$$

definita ponendo

$$s_{m,n}([x_0, \dots, x_m], [y_0, \dots, y_n]) = [\dots, x_i y_j, \dots].$$

È facile rendersi conto che questa applicazione è iniettiva. Tramite $s(m,n)$ si identifica $\mathbb{P}^m \times \mathbb{P}^n$ con l'immagine $S_{m,n}$ di $s_{m,n}$. Questa immagine si chiama *varietà di Segre*.

In generale un sottoinsieme $V \subseteq \mathbb{P}^N$ si dice *varietà algebrica* se esistono polinomi omogenei $F_i(x_0, \ldots, x_N)$ per $i = 1, \ldots, h$ tali che V coincide con l'insieme dei punti $[x_0, \ldots, x_N]$ di $\mathbb{P}^N$ le cui coordinate omogenee annullano tutti i polinomi F_i. Si noti che se un $n+1$-pla $(x_0, \ldots, x_N)$ annulla un polinomio omogeneo, lo stesso accade anche per tutte le $n+1$-ple ad essa proporzionali. I polinomi F_i che intervengono nella definizione di una varietà algebrica si dicono *equazioni* della varietà. Un insieme di equazioni per la varietà di Segre $S(m,n)$ si ottiene nella maniera seguente. Si consideri una matrice $(m+1) \times (n+1)$ di variabili z_{ij}. I minori di ordine due sono polinomi omogenei nelle variabili z_{ij}. È un divertente esercizio di algebra lineare verificare che la varietà algebrica che ha per equazioni tali polinomi è esattamente $S(m,n)$.

La mappa di Segre si può estendere al prodotto di più spazi proiettivi in maniera analoga

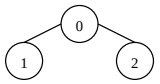
$$s(n_1, \ldots, n_h) : \mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_k} \to \mathbb{P}^{(n_1+1)\cdots\cdot(n_h+1)-1}.$$

Questa mappa è ancora iniettiva e la sua immagine $S(n_1, \ldots, n_h)$ è ancora una varietà algebrica.

Naturalmente la definizione delle mappe di Segre ha perfettamente senso anche sostituendo $\mathbb{C}$ con $\mathbb{R}$. Le parametrizzazioni dei modelli di indipendenza sono mappe di Segre.

Ad esempio l'equazione $x_0x_3 - x_1x_2 = 0$ che abbiamo incontrato discutendo il più semplice dei modelli di indipendenza definisce una *quadrica* nello spazio proiettivo a tre dimensioni, che è il prodotto di due rette proiettive.

Un primo modello di dipendenza Consideriamo il grafo



A ciascuno dei tre vertici è associato un alfabeto:

0	$a_0, \ldots, a_h$
1	$b_0, \ldots, b_n$
2	$c_0, \ldots, c_m$

I *parametri* del modello sono i seguenti:

- Le probabilità $p_0, \ldots, p_h$ di osservare $a_0, \ldots, a_h$ in 0.
- La matrice T di tipo $n \times h$, dove t_{ij} è la probabilità di osservare b_i in 1 se è stato osservato a_j in 0.
- La matrice S di tipo $m \times h$, dove s_{ij} è la probabilità di osservare c_i in 2 se è stato osservato a_j in 0.

In questo modello, la probabilità dell'osservazione (b_i, c_j) in $(1, 2)$, data l'osservazione a_α in 0 è

$$p_{\alpha,i,j} := p_\alpha t_{i\alpha} s_{j\alpha}.$$

Nelle applicazioni di questo modello, 0 è uno stato nascosto. La probabilità dell'osservazione (b_i, c_j) qualunque sia lo stato in 0 è

$$p_{ij} = \sum_{\alpha=0}^{h} p_\alpha t_{i\alpha} s_{j\alpha}. \tag{3}$$

Lo spazio delle distribuzioni di probabilità sulle possibili osservazioni di questo modello è l'insieme Δ delle $(n+1)(m+1)$-ple

$$(x_{ij}), \qquad i = 0, \ldots, n, \quad j = 0, \ldots, m$$

che verificano usuali restrizioni analoghe a (1), di cui da ora in poi non ci occuperemo.

Le matrici p_{ij} ottenute mediante la parametrizzazione (3) sono tutte e sole quelle che si ottengono per combinazione lineare di $h+1$ matrici di rango 1. È un esercizio di algebra lineare verificare che in tal modo si ottengono tutte e sole le matrici di rango al più $h+1$. Pertanto l'insieme

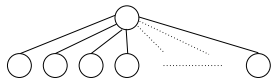
descritto parametricamente da (3) coincide con l'insieme delle matrici che verificano il sistema di equazioni omogenee di grado $h+2$ dato da

$$\mathrm{rk}(x_{ij}) \leq h+1.$$

Esplicitamente tale sistema si ottiene annullando tutti i minori di ordine $h+2$ della matrice $\{x_{ij}\}$.

Anche questi sistemi di equazioni definiscono varietà algebriche notevoli legate alle varietà di Segre. Infatti una siffatta varietà è la chiusura dell'insieme descritto da tutti i sottospazi di dimensione h generati da $h+1$ punti indipendenti della varietà di Segre $S(n,m)$. Questa varietà di dice *varietà degli spazi* $h+1$*-secanti la varietà di Segre.*

Questo modello si può generalizzare considerando il grafo



Dal punto di vista algebrico - geometrico questo corrisponde a considerare *varietà di spazi secanti varietà di Segre con più fattori*[28].

Varietà algebriche associate alle catene di Markov nascoste Anche in una catena di Markov nascosta, le espressioni algebriche per le probabilità *parametrizzano* una varietà algebrica.

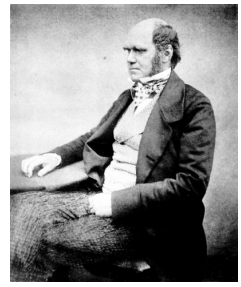
Gran parte di queste varietà non sono state precedentemente studiate e offrono *problemi interessanti alla geometria algebrica.* Ad esempio, un problema particolarmente rilevante è la determinazione di un *sistema di equazioni.*[9]

Viceversa, recenti *tecniche combinatoriche e computazionali* in geometria algebrica (*basi di Gröbner* [12, 13, 33], *geometria torica* [23], *geometria tropicale* [24, 32]) suggeriscono algoritmi per risolvere i problemi di *stima dei parametri* e verifica di *adeguatezza del modello.* Torneremo più avanti su alcuni di questi temi.

4 Filogenetica

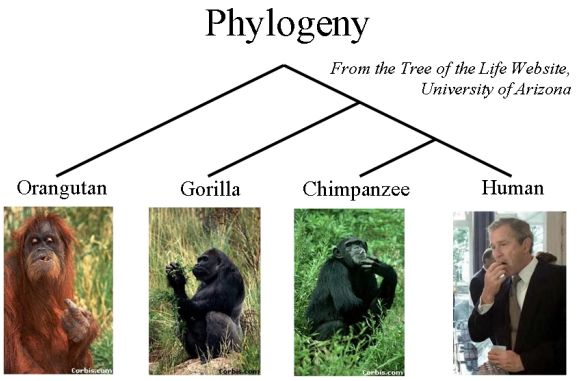
L'approccio algebrico, combinatorico e geometrico introdotto per l'analisi delle catene di Markov torna utile per altre applicazioni alla biologia, in particolare alla *Filogenetica*.

Evoluzionismo Il creatore dell'evoluzionismo è stato Charles Robert Darwin (1809 - 1882), naturalista inglese le cui teorie scientifiche costituiscono uno dei fondamenti della biologia moderna. Queste teorie sono volte a spiegare la diversità biologica tra le specie viventi partendo dall'ipotesi che esse si siano evolute a partire da antenati comuni. Non è il caso di dare qui una definizione formale del concetto di *specie*. Ad ogni modo il concetto di specie è legato a quello di riproduzione. Ad esempio, per i cosiddetti animali superiori, una specie è determinata da quegli individui che incrociandosi tra loro possono generare una prole illimitatamente feconda

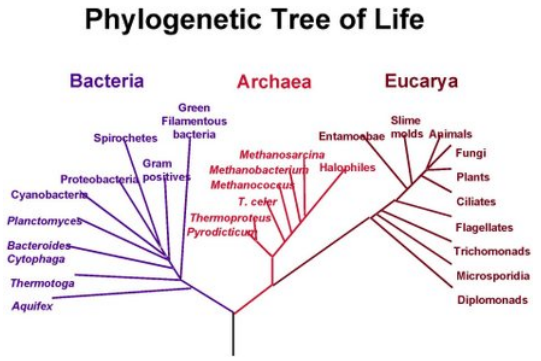


Darwin studiò medicina ad Edimburgo e teologia a Cambridge. Il suo viaggio intorno al mondo, durato cinque anni sulla nave Beagle fornì un ricco materiale di osservazioni su cui fondò le teorie esposte nel libro *On the Origin of Species (1859)*. Esse purtroppo sono ancora oggi oggetto di violente critiche antiscientifiche.

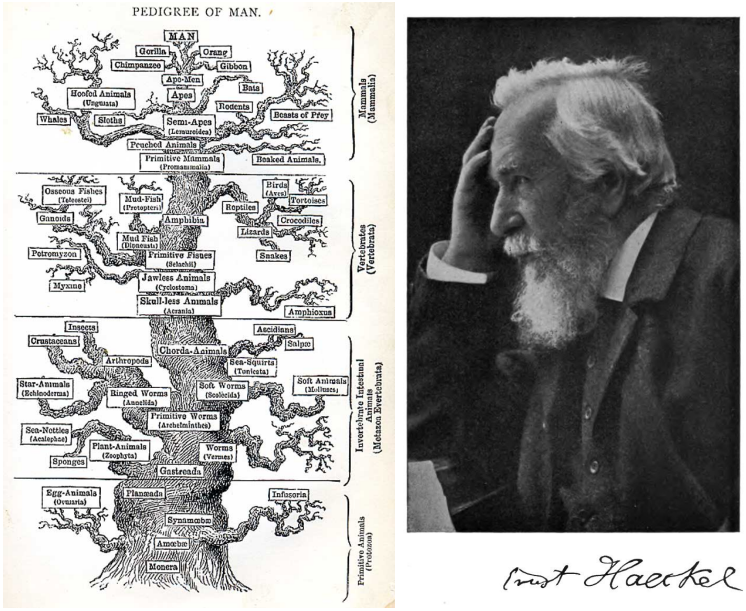
Alberi filogenetici L'evoluzionismo prevede l'esistenza di *alberi filogenetici* alla cui *radice* vi è l'antenato comune delle specie che si trovano alle *foglie*.



Gli *alberi filogenetici* o *filogenie* mostrano le *relazioni evolutive* tra diverse specie o altre entità biologiche che si suppone abbiano un antenato comune.



I termini filogenia, come anche il termine *ecologia*, furono proposti da Ernst Haeckel (1834 - 1919), biologo, naturalista, filosofo, medico ed artista tedesco. Fu un grande promotore delle idee di Darwin in Germania e diede nome a migliaia di specie viventi. Propose un albero filogenetico per tutte le forme di vita.



Ernst Heckel e il suo *Albero della vita*.

La costruzione e lo studio degli alberi filogenetici ha anche numerose e importanti applicazioni pratiche, ad esempio:

1. allo scopo di capire l'evoluzione di differenti ceppi virali per determinarne la pericolosità e valutare la possibilità di trovare vaccini efficaci;

2. valutare la *distanza evolutiva* tra diverse specie al fine di estendere l'efficacia di interventi terapeutici.

La costruzione degli alberi filogenetici è *in generale* un problema insolubile per la sua *enorme complessità*.

In pratica è possibile determinare alberi filogenetici che descrivono solo alcuni aspetti evolutivi di un ristretto insieme di specie, sfruttando un numero limitato di caratteri, che possono essere morfologici oppure biochimici.

5 Approccio matematico alla filogenetica

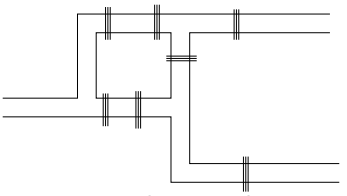
La nozione di albero filogenetico è suscettibile di una rigorosa definizione matematica di cui ora ci occupiamo e la cui utilità apparirà chiara in seguito.

Entra in scena la matematica: i grafi La nozione di *grafo* è utilizzata in matematica, in informatica e in numerosi altri contesti scientifici per modellare relazioni tra coppie di elementi di un insieme. Le applicazioni sono vastissime.

Il primo a riconoscere l'importanza del concetto di grafo è stato *Leonhard Euler*.

Leonhard Euler (1707-1783)

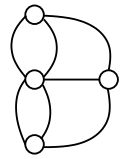
Il problema di cui si occupò Eulero questo proposito è il cosiddetto *problema dei ponti di Könisberg*. Nella città di Könisberg (ex Prussia, ora Kalinigrad, in Russia) ci sono due grosse isole nel fiume Pregel collegate tra loro e con il resto della città da 7 ponti, come schematizzato nella figura.



I ponti di Könisberg

Il problema consiste nel trovare, se possibile, un percorso che attraversi tutti ponti esattamente una volta e riporti al punto di partenza.

Eulero rappresentò ognuna della quattro zone in cui il fiume divide la città con un *nodo* e ogni ponte che congiunge due zone con un *arco* tra i corrispondenti nodi, cioè attraverso il *multigrafo* rappresentato in figura



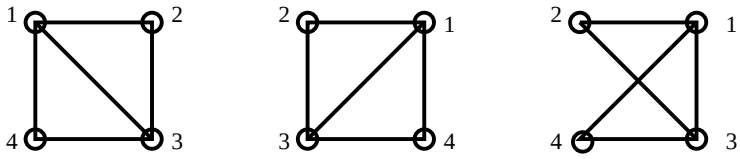
Eulero capì che la risolubilità del problema dipende dal numero degli archi che escono da ciascun nodo (detto *grado* del nodo). Eulero dimostrò che un percorso della forma desiderata esiste se e solo se non ci sono nodi di grado dispari. Un tale percorso è detto *circuito euleriano*.

Diamo la definizione di grafo. Un *grafo* (V, E) è una coppia di insiemi finiti e non vuoti V ed E:

- gli elementi di V si dicono *vertici* o *nodi*;
- gli elementi di E si dicono *archi* e sono coppie non ordinate di vertici distinti di V. In altre parole, E è contenuto nell'insieme di tutti i sottoinsiemi di due elementi di V.

Se $\{a, b\}$ è un arco, si dice che esso *congiunge* i vertici a, b. Due vertici collegati da un arco si dicono *adiacenti*.

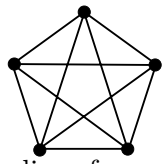
È possibile rappresentare ogni grafo con un diagramma, disegnando per ogni vertice un cerchietto e ogni arco con un arco di curva congiungente i cerchietti che rappresentano i vertici.



Tre diagrammi che rappresentano il grafo in cui $V = \{1, 2, 3, 4\}$ e $E = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{3, 4\}, \{4, .1\}\}$.

Attenzione!

- Il diagramma che rapprenta un grafo non è univocamente determinato.
- Non tutti i grafi si possono rappresentare nel piano in modo tale che gli archi si intersechino solo nei vertici. Se esiste una tale rappresentazione, il grafo si dice *planare*.

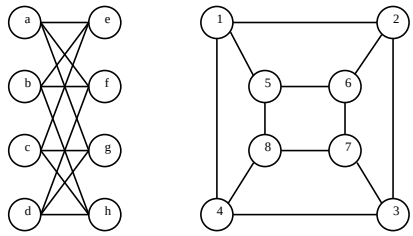


Esempio di grafo non planare

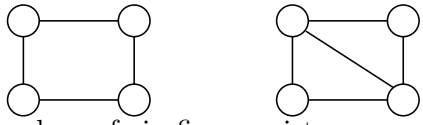
I grafi in cui ogni coppia di vertici è unita da un arco (grafi completi, come quello in figura) non sono planari se i vertici sono almeno quattro.

Due grafi possono essere diversi solo per il modo di chiamare i vertici. In tal caso si dicono *isomorfi* e hanno le stesse proprietà. Più precisamente

- Un *morfismo* tra due grafi (V_1, E_1) e (V_2, E_2) è una funzione $f : V_1 \to V_2$ che *preserva le adiacenze*, ovvero tale che
$$\{a, b\} \in E_1 \Longrightarrow \{f(a), f(b)\} \in E_2 \text{ per ogni } a, b \in V_1.$$
- Se f è un morfismo biunivoco e anche f^{-1} è un morfismo, si dice che f è un *isomorfismo*.



Un esempio di grafi isomorfi con isomorfismo ϕ dato da $\phi(a) = 1$, $\phi(b) = 6$, $\phi(c) = 8$, $\phi(d) = 3$, $\phi(e) = 5$, $\phi(f) = 2$, $\phi(g) = 4$, $\phi(h) = 7$.



Tra il primo e il secondo grafo in figura esiste un morfismo biunivoco che non è un isomorfismo.

Verificare se due grafi sono isomorfi può esser molto complicato.

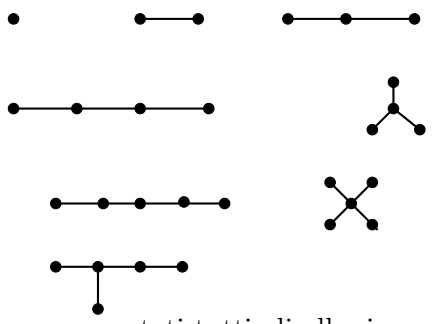
Pensando al problema di Eulero introduciamo la nozione di cammino in un grafo.

- Un *cammino* in un grafo G è una successione di vertici adiacenti $\{v_1, v_2, \ldots, v_n\}$.
- Un cammino si dice *chiuso* se $v_1 = v_n$.
- Un cammino chiuso, con almeno tre vertici e con vertici distinti salvo il primo e l'ultimo, si dice *ciclo*.
- Un grafo senza cicli si dice *aciclico*.
- Se ogni coppia di vertici di un grafo è congiunta da un cammino, il grafo si dice *connesso*.

Alberi Possiamo finalmente introdurre il concetto di albero. Gli alberi sono utilizzati nella modellizzazione di varie situazioni, per esempio in matematica (struttura delle formule), biologia (alberi filogenetici), informatica (struttura del file system di un calcolatore), araldica (alberi genealogici), ecc.

Dal punto di vista matematico un *albero* è un grafo connesso e aciclico.

Un albero con n vertici ha $n - 1$ archi. Questa proprietà caratterizza gli alberi tra i grafi connessi.

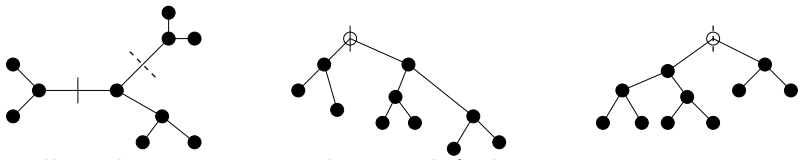


Nella figura sono rappresentati tutti gli alberi con 1,2,3,4,5 vertici.

Inoltre:

- una *foglia* di un albero è un vertice di grado 1.
- Un albero si dice *binario* se ha al più un vertice di grado 2, che si dice *radice*, e ogni vertice che non sia una foglia ha grado 3.

Un albero binario con radice con k foglie ha $2k - 1$ vertici.

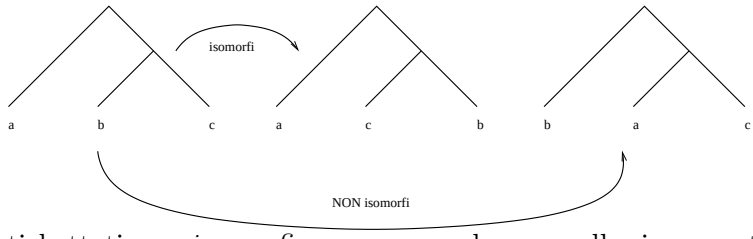


A ogni albero binario senza radici con k foglie si possono associare $2k - 2$ alberi con radice, uno per ogni arco. In figura sono indicati due modi per aggiungere una radice ad un albero con 6 foglie. In base a questa osservazione, la teoria degli alberi binari con o senza radice è essenzialmente la stessa.

Alberi filogenetici Per descrivere l'evoluzione delle specie biologiche si utilizza la struttura di *albero filogenetico* o *filogenia*.

Un *albero filogenetico* è un albero *binario* le cui foglie sono *numerate* o, come pure si dice *etichettate*: il numero assegnato a ciascuna foglia ne è l'etichetta.

La nozione di isomorfismo per alberi etichettati è più restrittiva di quella generale. Si richiede infatti che l'isomorfismo conservi le etichette.



Alberi etichettati non *isomorfi* possono esserlo come alberi senza etichette.

Il problema fondamentale della filogenetica è il seguente:

> Date delle specie e delle osservazioni ad esse relative, si vuole determinare l'albero filogenetico che *è in migliore accordo con le osservazioni* sulla base di una serie di *ipotesi di lavoro*.

Metodo di massima parsimonia Supponiamo di avere cinque specie, per ognuna delle quali si osservano sei caratteri, codificati nelle sequenze binarie riportate nella Tabella.

S1	1	0	0	1	1	0
S2	0	0	1	0	0	0
S3	1	1	0	0	0	0
S4	1	1	0	1	1	1
S5	0	0	1	1	1	0

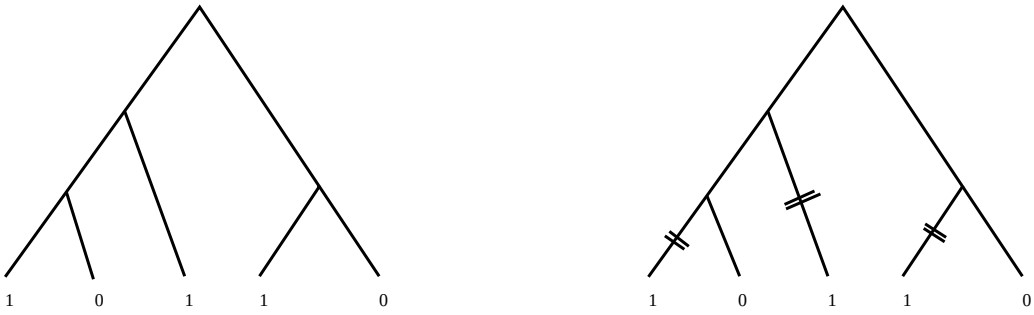
Quale dei possibili alberi filogenetici con cinque foglie è in *migliore accordo* con i caratteri osservati?

Esistono vari metodi per determinarlo. Essi portano in generale a risultati diversi. Il più semplice ed anche il primo ad essere stato impiegato è il metodo della *massima parsimonia*. Ritorneremo più avanti su metodi più raffinati.

Secondo il principio di massima parsimonia si giudicano *migliori* gli alberi che spiegano i caratteri con il *minimo numero di cambiamenti*.

Nel nostro esempio, per ogni albero filogenetico con cinque foglie $(S_1, \ldots, S_5)$ calcoliamo il minimo numero di cambiamenti necessari per spiegare i caratteri osservati nella Tabella (ogni carattere corrisponde a una colonna):

Consideriamo per esempio l'albero filogenetico a sinistra in Figura.



Se lo stato del primo carattere è uguale a 0 nella radice dell'albero, si ha una transizione allo stato 1 nei rami indicati nella figura a destra.

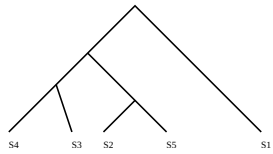
La seguente tabella fornisce il conteggio dei cambiamenti.

Carattere	cambiamenti da 0	cambiamenti da 1
1	3	**2**
2	*2*	*2*
3	**1**	2
4	*2*	*2*
5	*2*	*2*
6	**1**	2

Dunque per spiegare i sei caratteri utilizzando l'albero filogenetico proposto sono necessari almeno 10 cambiamenti.

Per applicare il principio di massima verosimiglianza occorre ripetere lo stesso conteggio per ogni albero filogenetico con cinque foglie e scegliere quello (o uno di quelli) che minimizzano i valori calcolati.

L'albero di massima parsimonia per l'esempio considerato è



La determinazione di un albero filogenetico secondo il principio di massima parsimonia prevede due passi. Il primo consiste di calcolare in relazione ad un dato albero, il numero minimo di cambiamenti necessari ad ottenere i dati osservati. Per questo esistono vari algoritmi di tipo *combinatorico*. Tra i più usati quelli di *Fitch* e di *Sankoff* [20, 40]. Il secondo passo consiste nell'iterare questo calcolo per *tutti* gli alberi filogenetici aventi un dato numero di etichette e confrontare i risultati. Questo è concettualmente semplice ma *di grande costo computazionale*. Infatti il numero degli alberi etichettati cresce *enormemente* al crescere del numero delle etichette.

Vale il seguente teorema: Il numero degli alberi binari con radice con k foglie etichettate, detto *numero di Schroeder*, è

$$(2k-3)!! = (2k-3)(2k-5)(2k-7)\cdot\cdots\cdot 5\cdot 3\cdot 1$$

Il numero di Schroeder cresce molto rapidamente al crescere di k e dunque non c'è speranza di determinare *esattamente* le filogenie quando il numero di specie supera la decina.

etichette	alberi filogenetici
6	945
10	~ 35.000
12	$\sim 13\cdot 10^9$
30	$\sim 10^{38}$
52	$\sim 10^{81}$

In questa tabella si riportano alcuni numeri di Schroeder. Si noti che il numero stimato degli atomi di idrogeno in tutte le stelle dell'universo, 4×10^{79} è minore del numero di Schroeder relativo a $k = 52$.

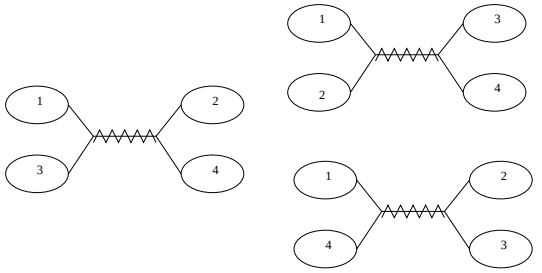
Esistono algoritmi basati su principi diversi per la ricerca di *buone approssimazioni* della soluzione ottimale. Essi si basano su una *struttura ma-*

tematica più raffinata che riguarda l'intero insieme degli alberi filogenetici con un dato numero di etichette di cui parliamo ora.

Grafo degli alberi filogenetici L'insieme degli alberi filogenetici con un numero fissato di foglie ha a sua volta una struttura di grafo determinata specificando quali coppie di alberi sono adiacenti.

La nozione di adiacenza comunemente usata si basa sull'idea dei *Nearest-neighbor interchanges*.

Essa funziona nel seguente modo. Dato un albero G e consideratone un arco si possono costruire due nuovi alberi che diremo *adiacenti* a G. La costruzione è illustrata nella seguente figura



Il numero degli alberi filogenetici adiacenti ad uno con k etichette è dunque uguale al doppio del numero dei suoi archi, ovvero $2k - 4$.

In generale, per ricercare un massimo locale di una funzione definita sui vertici di un grafo l'idea è di partire da un vertice qualsiasi e di *visitare tutti quelli adiacenti*. Tra questi si sceglie il vertice in cui la funzione è massima, e si ricomincia. Si percorre in questa maniera un cammino sul grafo che porta ad un *massimo locale*, che non può essere migliorato da quelli adiacenti. Questo non permette in generale di trovare massimi globali. Tuttavia esistono raffinamenti probabilistici di questo algoritmo, detti di *simulated annealing*[31, 27, 30], volti ad evitare di rimanere intrappolati in un massimo locale che non sia sufficientemente robusto.

Questo principio si può applicare al grafo di tutti gli alberi filogenetici per effettuare la ricerca di una soluzione del problema fondamentale della filogenetica con il metodo della massima parsimonia.

6 Modelli grafici su alberi.

I semplici modelli probabilistici che abbiamo discusso nella sezione 2 utilizzano strutture ad albero e proprio per questo risultano utili in filogenetica. Si possono considerare modelli probabilistici più generali che utilizzano grafi, anche orientati. Questi modelli prendono il nome di *modelli grafici* e *reti Bayesiane* nel caso orientato. Hanno numerose applicazioni tra cui riconoscimento di immagini, teoria delle decisioni, intelligenza artificiale, bioinformatica, ecc, [5]. Qui ci limitiamo a considerare solo modelli grafici su alberi.

Un modello grafico specifica come calcolare la probabilità di effettuare una serie di osservazioni alle foglie di un albero. Il modello che si usa è una naturale generalizzazione delle catene di Markov a stati nascosti in cui i nodi interni sono trattati in maniera analoga agli stati nascosti della catena.

Dato un albero con radice, il modello che noi consideriamo è assegnato definendo:

1. un *alfabeto* $\Omega = \{\omega_1, \ldots, \omega_n\}$;
2. il *vettore delle probabilità iniziali* $p = (p(\omega_1), \ldots, p(\omega_n))$ dove $p(\omega)$ è la probabilità di osservare ω nella radice;
3. la *matrice di transizione* $T = (t_{\omega,\omega'})$ dove $t_{\omega,\omega'}$ è la probabilità di passare da ω ad ω' lungo un qualsiasi arco dell'albero.

Le entrate della matrice T e del vettore p si dicono *parametri del modello.*

Il nostro modello assegna uguale probabilità di transizione tra i caratteri ω e ω' lungo un qualsiasi arco. Non sempre questo è realistico. Si possono considerare modelli più complicati in cui le matrici di transizione dipendano dagli archi e gli alfabeti dai vertici.

Formule polinomiali per la probabilità Come abbiamo detto, il modello ci occorre per calcolare la probabilità di effettuare una data osservazione alle foglie. Sia $p_W(\omega)$ la probabilità di osservare il carattere ω nella foglia W. Per calcolare $p_W(\omega)$ si procede così:

- si determina l'unico *cammino minimale* $V_1, \ldots, V_k = W$, cioè un cammino senza ripetizione di vertici, tale che V_1 sia la radice;
- per una qualsiasi assegnazione di caratteri $\omega^{(1)}, \ldots, \omega^{(k-1)}, \omega^{(k)} = \omega$ ai vertici $V_1, \ldots, V_{k-1}, V_k$ la probabilità di questa assegnazione è

$$\boxed{\pi(\omega^{(1)}, \ldots, \omega^{(k-1)}, \omega^{(k)}) := p(\omega^{(1)}) \cdot T_{\omega^{(1)}\omega^{(2)}} \cdot T_{\omega^{(2)}\omega^{(3)}} \cdot \cdots \cdot T_{\omega^{(k-1)}\omega^{(k)}}}$$

  che è un *monomio* nei parametri del modello;
- si calcola la probabilità $p_W(\omega)$ mediante la formula

$$\boxed{p_W(\omega) = \sum_{\omega^{(1)}, \ldots, \omega^{(k-1)}} \pi(\omega^{(1)}, \ldots, \omega^{(k-1)}, \omega)}$$

  che è un *polinomio* nei parametri del modello.

Infine, la *probabilità di osservare* $\omega^{[1]}, \ldots, \omega^{[h]}$ nelle foglie $W_1, \ldots, W_h$ è

$$\boxed{p(\omega^{[1]}, \ldots, \omega^{[h]}) := p_{W_1}(\omega^{[1]}) \cdot \cdots \cdot p_{W_h}(\omega^{[h]})}$$

Questo è ancora un polinomio nei parametri del modello. Come abbiamo già osservato questo apre la strada alla possibilità di usare metodi *algebrico geometrici* per l'analisi di questi modelli. Diamo qui di seguito un esempio dell'utilità di questo approccio.

Varietà algebriche e invarianti filogenetici Dato un modello probabilistico del tipo appena considerato, relativo ad un albero con m foglie e ad un alfabeto con n caratteri, si ottengono

$$\boxed{N = m^n}$$

polinomi $p(\omega^{[1]}, \ldots, \omega^{[m]})$ che calcolano le probabilità delle osservazioni alle foglie in funzione dei parametri del modello. Tra questi quelli *indipendenti* sono

$$\boxed{r = n^2 - 1}$$

Infatti i parametri, che sono $n + n^2$ verificano le seguenti $n + 1$ relazioni indipendenti

$$\sum_{\omega \in \Omega} p(\omega) = 1 \qquad \sum_{\omega' \in \Omega} t(\omega, \omega') = 1.$$

Considerando i polinomi $p(\omega^{[1]}, \ldots, \omega^{[m]})$ sul campo complesso, risulta definita un' applicazione

$$\boxed{\phi : \mathbb{C}^r \to \mathbb{C}^N}$$

Applicazioni di questo tipo sono stati già considerati nella sezione 3, ad esempio le mappe di Segre.

La *chiusura* dell'immagine di una mappa $\phi : \mathbb{C}^r \to \mathbb{C}^N$ definita da polinomi è una *varietà algebrica*, cioè il luogo Z degli zeri di un *insieme di polinomi*. Un famoso teorema di Hilbert (cfr. [16]) assicura che, data una varietà algebrica Z esiste un insieme finito di polinomi $f_1, \ldots, f_h$ tale che Z è l'insieme delle soluzioni del sistema di equazioni $f_1 = \cdots = f_h = 0$ e inoltre ogni polinomio g che si annulla su tutti i punti di Z si esprime come $g = a_1 f_1 + \cdots + a_h f_h$ per opportuni polinomi $a_1, \ldots, a_h$. Tutti i polinomi g siffatti costituiscono l'*ideale generato da* $f_1, \ldots, f_h$ che prende il nome di *ideale della varietà* Z. Nell'esempio delle mappe di Segre i polinomi che definiscono la ϕ sono omogenei dello stesso grado ed è inoltre possibile scegliere i polinomi $f_1, \ldots, f_h$ in maniera che siano omogenei. Questo ci ha permesso di riguardare la chiusura dell'immagine delle mappe di Segre come varietà in uno spazio proiettivo. In generale i polinomi che definiscono la ϕ non sono omogenei dello stesso grado e le equazioni per la chiusura dell'immagine pure non sono omogenee. Quindi definiscono solo varietà algebriche *affini*. Molte delle varietà che nascono da modelli grafici non sono mai state studiate in precedenza e risultano interessanti sia per le applicazioni che per le loro proprietà algebrico-geometriche [10, 17, 18, 34, 46]. I polinomi delle varietà associate ai nostri modelli grafici sono detti *invarianti filogenetici*.

L'approccio algebrico-geometrico allo studio dei modelli grafici può essere molto utile. Ad esempio, se il modello che stiamo considerandoè adeguato alla descrizione di un insieme di dati sperimentali, ogni suo invariante filogenetico, valutato sulle frequenze empiriche stimate dai dati, deve assumere

valori *prossimi a zero*. Quindi ogni invariante filogenetico offre un *test* per *validare il modello* o per *verificare la bontà dei dati*.

Il problema generale è quello di determinare i polinomi che si annullano sui punti di una varietà definita *parametricamente* da un insieme di equazioni polinomiali

$$y_i = p_i(x_1, \ldots, x_r) \quad i = 1, \ldots, N.$$

Ad esempio, le seguenti equazioni parametriche

$$y_i = t^i \quad i = 1, \ldots, N$$

definiscono la *curva razionale normale* di $\mathbb{C}^N$. Equazioni per la varietà assegnata parametricamente si ottiene *eliminando* i parametri dalle equazioni parametriche. Per esempio, eliminando il parametro t nell'esempio della curva razionale normale, si ottengono le equazioni

$$y_i = y_1^i \qquad i = 1, \ldots, N.$$

In questo caso i polinomi $y_i - y_1^i$ sono addirittura generatori dell'ideale della curva. In generale innanzitutto è difficile eliminare le variabili. Inoltre non è detto che si ottenga come risultato un sistema di generatori. Eliminare le variabili ed ottenere le equazioni della varietà è un problema classico della geometria algebrica [41, 48]. Oltre ai metodi classici basati sui *risultanti* negli ultimi quarant'anni sono stati sviluppati algoritmi che usano la nozione di *base di Gröbner*. Una base di Gröbner per un ideale è una scelta particolarmente comoda per i calcoli di un sistema di generatori dell'ideale e si ottiene con un algoritmo, dovuto a Buchberger, che generalizza l'algoritmo di eliminazione di Gauss per la risoluzione di un sistema di equazioni lineari[33, 12, 13, 16]. Questi algoritmi hanno una complessità in generale doppiamente esponenziale ma in casi non patologici la complessità è polinomiale. Essi rendono dunque teoricamente abbordabile il calcolo di invarianti filogenetici.

7 Matrici di dissimilarità per la semplificazione dei dati

I dati relativi all'osservazione di m *caratteri* su n *specie*, si possono raccogliere in una *matrice* $n \times m$: l'elemento sulla riga di posto i e sulla colonna di posto j è il carattere j-esimo relativo alla i-esima specie. Tipicamente si considera un numero *basso* di specie (in genere nell'ordine della decina) e un numero molto più alto di caratteri, ad esempio sequenze di qualche migliaio di basi. Quindi $m >> n$.

Si può ridurre la complessità dei dati mediante il seguente espediente. Usando i caratteri a disposizione si attribuisce ad ogni coppia di specie una *misura di dissimilarità* che è un numero che misura la differenza tra i caratteri osservati. Vedremo tra breve in base a quali criteri convenga scegliere tale misura. Con queste misure di *dissimilarità filogenetica* si forma una matrice quadrata e simmetrica con n righe ed n colonne ed entrate nulle sulla diagonale principale, detta *matrice di dissimilarità*.

Dare una matrice di dissimilarità equivale ad assegnare una *funzione di dissimilarità*

$$\delta : \{1, \dots, n\} \times \{1, \dots, n\} \to \mathbb{R}_{\geq 0}$$

tale che

1. $\delta(i, i) = 0 \qquad \forall i = 1, \dots, n;$
2. $\delta(i, j) = \delta(j, i) \qquad \forall i, j = 1, \dots, n.$

Se una funzione di dissimalarità d verifica la disuguaglianza triangolare

$$d(i, j) + d(j, k) \geq d(i, k) \qquad \forall i, j, k = 1, \dots, n$$

allora è una *distanza*.

Il processo che associa la matrice di dissimilarità a quella dei caratteri è una drastica semplificazione dell'informazione a nostra disposizione. Nonostante tale semplificazione possa apparire troppo drastica, nelle applicazioni essa risulta sorprendentemente efficace per opportune scelte della nozione di dissimilarità. In altri termini la matrice di dissimilarità contiene ancora una quantità sufficiente di informazione per costruire filogenie piuttosto accurate.

Quale misura di dissimilarità scegliere per le sequenze di DNA? La misura più semplice consiste nel numero delle differenze tra i caratteri corrispondenti di due sequenze. Questa è una distanza ma *è poco significativa dal punto di vista biologico* e va corretta usando un modello probabilistico di evoluzione, per esempio il *modello di Jukes e Cantor*. Infatti in biologia importa non tanto considerare le differenze osservate tra tratti allineati di DNA ma piuttosto il *numero totale di sostituzioni* di caratteri occorse nel processo di differenziazione delle due sequenze. Quindi il numero delle differenze tra i caratteri omologhi delle sequenze osservate è *solo* una sottostima di tale misura. Dal punto di vista computazionale il calcolo di queste dissimilarità è *molto leggero*.

Il modello di Jukes e Cantor Descriviamo il più semplice dei modelli probabilistici utilizzati dai biologi per descrivere l'evoluzione di sequenze biologiche (cfr. [19]). Sulla base di questo modello è possibile, come vedremo, determinare una dissimilarità biologicamente significativa. Tale dissimilarità non è di solito una distanza.

In questo modello si suppone di avere una filogenia T e si parte dall'ipotesi semplificativa che ogni carattere evolva lungo gli archi di T indipendentemente dagli altri. Sia X un carattere che assume valori in Ω (tipicamente $\Omega = \{A, C, G, T\}$ o $\Omega = \{0, 1\}$). Si consideri un arco l di estremi x e y a cui attribuiamo una *lunghezza* o *peso* che è un numero positivo che conviene scrivere nella forma $\frac{3}{4}\alpha_l t_l$. La quantità t_l si può pensare in prima approssimazione come il tempo trascorso nell'evoluzione da x a y. La quantità α_l rappresenterà il *tasso di sostituzioni* nei caratteri che hanno luogo lungo l'arco l.

Possiamo pensare alla seguente realizzazione concreta del processo evolutivo. Si pone un contatore geiger davanti ad una sostanza radioattiva avente tempo di dimezzamento uguale ad α. Ogni volta che il contatore misura l'emissione di una particella radioattiva si pesca da un'urna contenente una pallina di colore diverso per ogni elemento di Ω. La pallina estratta (che viene poi reimbussolata) fornisce lo stato attuale del carattere.

Più formalmente, questo processo si compone di due fasi. La prima si

modella con un *processo di Poisson* di parametro α_l e consiste nell' innesco della seconda fase. In altre parole, la probabilità che la fase due venga innescata k volte nell'intervallo di tempo $[0, t]$ è

$$p(k,t) = \frac{(\alpha t)^k}{k!} e^{-\alpha t}.$$

Ogni volta che la seconda fase viene innescata, la probabiltà di effettuare l'osservazione di un elemento di Ω è uniforme, cioè $1/|\Omega|$.

Possiamo a questo punto calcolare facilmente la probabilità di transizione da un carattere ω in x a un carattere ω' in y. La probabilità che la fase 2 non venga mai innescata nell'intervallo $[0, t_l]$ è $p(0, t_l) = e^{(-\alpha_l t_l)}$ e quindi quella che la fase 2 venga innescata almeno una volta è il complementare, $1 - e^{(-\alpha_l t_l)}$. Nell'ipotesi di uniformità nella fase due e nell'ipotesi che ci sia stato almeno un innesco la probabilità di osservare un qualsiasi elemento di Ω vale $\frac{(1-e^{(-\alpha_l t)})}{|\Omega|}$.

In questo modello siamo interessati a calcolare due probabilità. Quella che il carattere X in y abbia lo stesso valore che in x, che indicheremo con θ_l, e quella che il carattere X in y abbia un valore diverso che in x, che indicheremo con π_l. Da quello che abbiamo detto segue che

$$\pi_l = \frac{(n-1)}{n}(1 - e^{(-\alpha_l t_l)}).$$

Invece, per calcolare θ_l osserviamo che un carattere può essere conservato in due casi: o se non cè stato nessun innesco ovvero se almeno un innesco c'è stato e l'ultima estrazione ha riprodotto le stesso carattere. Pertanto

$$\theta_l = \frac{(1 - e^{(-\alpha_l t_l)})}{n} + e^{-\alpha_l t_l} = \frac{1 + (n-1)e^{-\alpha_l t_l}}{n}.$$

Riassumendo, questo modello nel caso dell'evoluzione di una sequenza biologica di DNA calcola le seguenti probabilità

$$P_l(X(y) = \omega' | X(x) = \omega, t) = \begin{cases} \pi_l = \frac{1-e^{(-\alpha_l t_l)}}{4} & \text{If } \omega' \neq \omega \\ \theta_l = \frac{1+3e^{(-\alpha_l t_l)}}{4} & \text{If } \omega' = \omega \end{cases} \tag{4}$$

Questi dati si possono compendiare nella *matrice di transizione*

$$P_l = \begin{pmatrix} \theta_l & \pi_l & \pi_l & \pi_l \\ \pi_l & \theta_l & \pi_l & \pi_l \\ \pi_l & \pi_l & \theta_l & \pi_l \\ \pi_l & \pi_l & \pi_l & \theta_l \end{pmatrix} \tag{5}$$

Si notino le relazioni $\theta_l \geq 0$, $\pi_l \geq 0$ e $\theta_l + 3\pi_l = 1$.

Nel caso binario la matrice di transizione assume la forma

$$P_l = \begin{pmatrix} \theta_l & \pi_l \\ \pi_l & \theta_l \end{pmatrix}. \tag{6}$$

Per dedurre da questo modello una mappa di dissimilarità si procede nel seguente modo.

Ricordiamo che stiamo confrontando m caratteri su n specie. Fissiamo due specie. Sia k il numero di caratteri che assumono valori diversi su queste due specie. Supponiamo che l'evoluzione delle due specie dal più prossimo comune antenato sia descritta da un semplice albero con due foglie corrispondenti alle specie, un unica radice, corrispondente all'antenato comune e due archi con tassi di sostituzione dei caratteri dati da α_1 e α_2. È facile verificare che la probabilità p che un carattere abbia lo stesso valore alle due foglie è

$$p = \theta_1\theta_2 + 3\pi_1\pi_2$$

e quindi, richiedendo che il valore aspettato pm dei caratteri conservati sia $m - k$ si ha

$$12m\pi_1\pi_2 - 3m\pi_1 - 3m\pi_2 + k = 0.$$

Questa equazione si può riscrivere nella forma

$$(1 - 4\pi_1)(1 - 4\pi_2) = 1 - \frac{4k}{3m}.$$

Ricordando che

$$1 - 4\pi_i = e^{-\alpha_i t_i}$$

e che le lunghezze degli archi valgono $3/4\alpha_i t_i$, si ottiene che la somma δ di tali lunghezze vale

$$\delta = -3/4 \log \left(1 - \frac{4k}{3m}\right).$$

Questa è la formula di Jukes e Cantor per calcolare la dissimilarità di due sequenze.

8 Filogenie pesate.

Anche alla luce delle considerazioni che abbiamo appena svolto è naturale considerare un arricchimento di un albero filogenetico ottenuto assegnando ad ogni arco un numero positivo, detto *peso* o *lunghezza*, che *rappresenta il numero di sostituzioni* avvenute nella sequenza considerata nell'evoluzione da una specie all'altra lungo l'arco.

Un albero *filogenetico con pesi* T determina una ovvia distanza d_T, detta *distanza arborea* tra le foglie e quindi una *matrice delle distanze*.

> **Problema** Assegnata una matrice di dissimilarità filogenetiche determinare una filogenia pesata la cui matrice delle distanze arboree *meglio approssimi* la matrice data.

L'interesse di questo problema è chiaro. A partire dai dati sperimentali vogliamo costruire un albero filogenetico che li spieghi.

È bene osservare che non c'è modo di rispondere in modo univoco al problema di cui sopra. Basta pensare al problema di costruire una filogenia tra due foglie x e y, assegnata la distanza d tra queste. Dal punto di vista combinatorico esiste una sola filogenia su due foglie che prevede un unico progenitore z per x e y.

FIGURA

Questa filogenia prende il nome di *ciliegia* interpretando x e y come ciliegie collegate al picciuolo z. Quella che resta determinata è la somma delle distanze di z tra x e y mentre non c'è modo di determinare ciascuna delle due distanze separatamente. La scelta più ingenua è quella di assumere

queste due distanze uguali a $d/2$. È bene però porre l'attenzione su due aspetti delicati:

- se si parte da una data filogenia pesata sulle due foglie x e y, questa scelta in generale non la ricostruisce;

- dal punto di vista biologico questa scelta è giustificata soltanto se il tasso di sostituzione dei nucleotidi lungo entrambi i rami dell'evoluzione può ritenersi uguale ovvero che tale tasso sia proporzionale al tempo di evoluzione con la stessa costante di proporzionalità su entrambi i rami. Queste ipotesi prende il nome di *orologio molecolare*.

Esiste un semplice algoritmo che fornisce buone ricostruzioni di filogenie pesate nell'ipotesi dell'orologio molecolare. Si tratta del cosiddetto UPGMA (Unweighted Pair Group Method with Arithmetic mean) [42].

Siano date n specie che indichiamo con i numeri da 1 a n di cui abbiamo stimato le dissimilarità: $\delta(i,j)$ indica la dissimilarità tra la specie i e la specie j.

UPGMA

1. Formiamo la matrice quadrata $Q_\delta = (\delta(i,j))_{1\leq i,j\leq n}$.
2. Identifichiamo una entrata $\delta(i,j)$ in Q_δ fuori della diagonale principale di grandezza minima. Colleghiamo i e j con una ciliegia il cui picciuolo indichiamo con $z(i,j)$.
3. Sostituiamo i e j con $z(i,j)$. La dissimilarità di $z(i,j)$ da una specie k diversa da i e j viene definita come la media aritmetica delle dissimilarità di k da i e j rispettivamente meno la metà della dissimilarità tra i e j. Le rimanenti dissimilarità restano invariate.
4. Iteriamo il procedimento avendo cura di effettuare le medie necessarie non più calcolando le medie aritmetiche bensì medie pesate: il peso di un vertice è uguale al numero di vertici che ha assorbito fino a quel momento.

È bene ribadire che questo algoritmo produce una filogenia pesata ma che se si parte da una data metrica arborea esso non riproduce la filogenia pesata di partenza. La condizione necessaria e sufficiente perché ciò avvenga è quella che venga verificata l'ipotesi dell'orologio molecolare ovvero che tutte le foglie abbiano la stessa distanza dalla radice.

9 Alberi semietichettati senza radice

In generale la matrice di dissimilarità filogenetica non è una distanza e, se pure lo fosse, non è detto sia arborea. Se però ciò accade allora esiste un'unica filogenia pesata la cui matrice delle distanza coincide con la data matrice di dissimilarità, come risulta dall'algoritmo di Neighbor Joining che ci accingiamo a spiegare.

Sia data una funzione di dissimilarità δ sull'insieme $X_n = \{0, \ldots, n+1\}$.

Si verifica che, se esiste una filogenia pesata T tale che $\delta = d_T$, allora una coppia (x, y) di punti distinti di X_n che minimizza

$$q_\delta(x, y) := (n-1) \cdot \delta(x, y) - \sum_{k \neq x} \delta(x, k) - \sum_{h \neq y} \delta(y, h),$$

è una *ciliegia*, ossia esiste un vertice interno z di T, il *picciuolo*, congiunto da un arco sia a x che a y.

Neighbor Joining

Algoritmo di Neighbor Joining

1. Formiamo la matrice quadrata $Q_\delta = (q_\delta(i, j))_{0 \leq i,j \leq n}$.
2. Identifichiamo una entrata fuori della diagonale principale che minimizza $q_\delta(x, y)$ di Q_δ .
3. Se esistesse un albero T tale che $\delta = d_T$, avremmo un picciuolo z per x e y e, per ogni $t \in X_n$ avremmo

$$\delta(z, t) = \frac{1}{2}(\delta(x, t) + \delta(y, t) - \delta(x, y)).$$

4. Rimuoviamo x e y sostituendoli con z, la cui distanza dai rimanenti punti di X_n sia data dalla formula precedente.
5. Iteriamo il procedimento.

Se c'è una filogenia pesata T tale che $\delta = d_T$, è chiaro che quest'albero è unico e l'algoritmo di neighbor joining ricostruisce l'albero T.

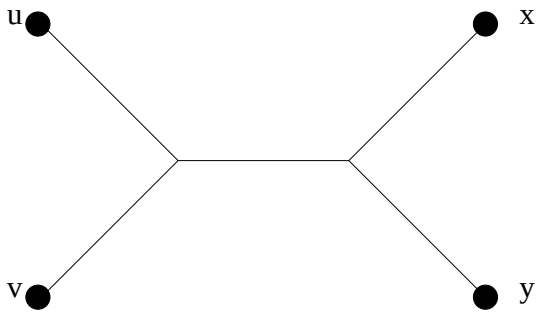
Condizione dei quattro punti Le distanze arboree si possono caratterizzare con la seguente condizione algebrica:

per ogni quaterna di punti x, y, u, v il massimo delle quantità

$$\boxed{d(u,v) + d(x,y), \qquad d(u,x) + d(v,y), \qquad d(u,y) + d(v,x)}$$

viene raggiunto almeno due volte.

La necessità di questa condizione è evidente. Basta osservare che la più piccola delle tre quantità corrisponde ad aggregare le coppie di vertici che che sono parenti più prossimi. Per esempio, nella figura seguente, la quantità minore è $d(u,v) + d(x,y)$, corrispondente all'aggregazione di x con y e di u con v. Da questa si ottengono le altre aggiungendo il doppio della distanza tra l'antenato comune più prossimo a x e y dall'antenato comune più prossimo a u e v.



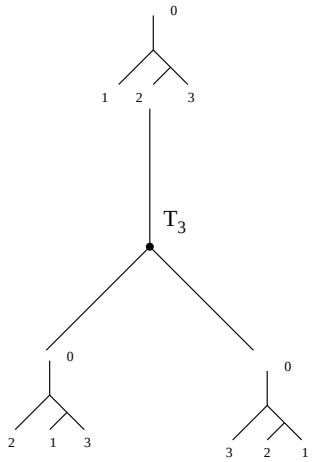
Per la sufficienza si rimanda a [11].

Lo spazio degli alberi di Billera, Holmes e Vogtmann Billera, Holmes e Vogtmann hanno costruito uno *spazio metrico* $\mathcal{T}_n$ i cui punti corrispondono agli alberi binari, con radice, con n foglie etichettate e con *archi interni* di lunghezza positiva [4]. Per cominciare, supponiamo che agli archi che contengono una foglia non venga attribuita lunghezza.

$\mathcal{T}_n$ è costituito da $(2n-3)!!$ *ortanti* di dimensione $n-2$, i cui bordi sono *opportunamente incollati*. Ricordiamo che un ortante di dimensione h è il prodotto h volte di $[0,+\infty)$. Ciascun ortante corrisponde ad una possibile struttura combinatorica di un albero. Ci si muove all'interno di ciascun ortante mantenendo la struttura combinatorica, ma variando la lunghezza degli archi interni: quando uno di questi diviene di lunghezza nulla si raggiunge il bordo dell'ortante. A un punto del bordo corrispondente ad un albero con un solo lato interno di lunghezza nulla si può giungere anche da altri due ortanti: ciò corrisponde ai due alberi adiacenti a quello dato nel grafo degli alberi filogenetici (cfr. p. 31). Gli incollamenti degli ortanti in punti del bordo corrispondenti ad alberi con più lati interni di lunghezza nulla sono più complicati. Ci limiteremo tra breve ad esaminare un paio di esempi. Osserviamo che in ogni caso le origini di tutti gli ortanti vanno identificate in punto che chiameremo *origine*. Inoltre è ben definito il prodotto di un numero non negativo per un elemento di $\mathcal{T}_n$ e questa moltiplicazione trasforma ogni ortante in sè. Dato un elemento T di $\mathcal{T}_n$, l'insieme $\lambda \cdot T$ con $\lambda \in [0,+\infty)$ è una *semiretta* pasante per l'origine. Quind $\mathcal{T}_n$ si può pensare come un *cono* di vertice l'origine. Se si considera in ciascun ortante il sottoinsieme $\mathcal{P}(i)$ definito dall'equazione $x_1 + \cdots + x_{n-2} = 1$, l'unione $\mathcal{P}_n = \cup_i \mathcal{P}(i)$ è una *base* del cono $\mathcal{T}_n$ che contiene tutte le informazioni necessarie a ricostruire $\mathcal{T}_n$.

La metrica di $\mathcal{T}_n$ è quella euclidea su ogni ortante. La distanza tra due punti in ortanti diversi è la minima lunghezza di un cammino che congiunge i due punti. Volendo attribuire anche agli archi che contengono una foglia una lunghezza si ottiene semplicemente il prodotto di $\mathcal{T}_n$ per l'ortante $[0,+\infty)^n$. Per questo motivo basta capire la struttura di $\mathcal{T}_n$.

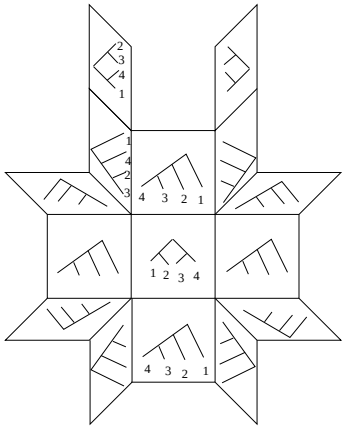
Lo spazio $\mathcal{T}_3$



Esistono tre tipi combinatorici distinti di alberi binari con radice e con tre foglie, ognuno con un solo arco interno, pesato con un numero positivo. Corrispondentemente lo spazio $\mathcal{T}_3$ consiste di tre ortanti uno dimensionali. Azzerando la lunghezza dell'arco interno degeneriamo uno qualsiasi dei tre tipi combinatorici ad un albero con un solo vertice interno da cui si dipartono tre archi che congiungono il vertice interno alle foglie.

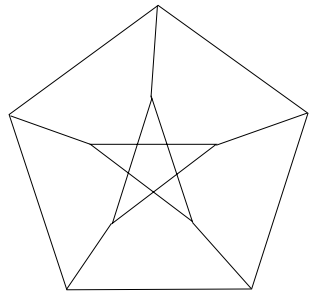
Lo spazio $\mathcal{P}_3$ consiste di tre punti distinti.

Lo spazio $\mathcal{T}_4$



Esistono 15 tipi combinatorici distinti di alberi binari con radice e con quattro foglie, ognuno con due archi interni, ciascuno pesato con un numero positivo. Corrispondentemente lo spazio $\mathcal{T}_4$ consiste di *15* ortanti bidimensionali, aventi in comune l'origine. Ogni ortante unodimensionale *corrisponde ad un albero con solo un arco interno*, che si può deformare ad un albero binario in tre modi distinti. Dunque ogni ortante unodimensionale appartente a *tre ortanti bidimensionali*. Dalla figura si ottiene $\mathcal{T}_4$ con opportune identificazioni degli ortanti unodimensionali. In particolare *tutti i vertici vanno identificati*.

Lo spazio $\mathcal{P}_4$ è un grafo *trivalente*, i cui vertici corrispondono ai 10 ortanti unidimensionali, e gli archi ai 15 ortanti bidimensionali. Si tratta del ben noto *grafo di Peterson*:



Utilità in Biologia degli spazi $\mathcal{T}_n$. Le proprietà di $\mathcal{T}_n$, in particolare la metrica che abbiamo introdotto, sono assai utili in biologia.

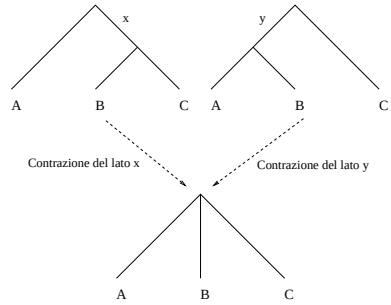
Per esempio, supponiamo di essere pervenuti a un certo numero $T_1, \ldots, T_h$ di filogenie che hanno probabilità $p_1, \ldots, p_n$ di spiegare i nostri dati e di voler determinare una filogenia T su cui si esprima il *consenso* dei dati raccolti. Si tratta di *interpolare* tra le filogenie trovate. A tale scopo si può prendere, ad esempio, il *baricentro*, il *circocentro* o il *centroide*. Queste nozioni coincidono nello spazio euclideo, ma danno luogo a concetti diversi, e diversamente utili, in $\mathcal{T}_n$.

Il *baricento* di un insieme di punti pesati $(T_1, p_1), \ldots, (T_h, p_h)$ è definito come il punto T che minimizza il numero $p_1^2 d(T, T_1)^2 + \ldots + p_h^2 d(T, T_h)^2$. Si dimostra che questo punto esiste ed è unico.

Se la distribuzione $p_1, \ldots, p_n$ è uniforme, cioè se $p_1 = \ldots = p_h$, si può considerare anche il *circocentro*, cioè il centro della più piccola sfera nella metrica d che contiene tutti i punti $T_1, \ldots, T_h$. Anche qui si dimostra che tale punto esiste ed è unico. Se $h = 2$ questo non è altro che il punto medio della geodetica che congiunge i due punti.

Un'altro concetto analogo è quello di *centroide* di $T_1, \ldots, T_h$. Se $h = 2$ questo coincide col circocentro. Se $h > 2$, esso si ottiene nel seguente modo induttivo: si prende il centroide su ogni $(h-1)$-upla di punti di $\{T_1, \ldots, T_h\}$, ottenendo coì un nuovo insieme di punti $\{T_{1,1}, \ldots, T_{h,1}\}$; si itera questa costruzione ottenendo per ogni intero positivo m un insieme di punti $\{T_{1,m}, \ldots, T_{h,m}\}$; si dimostra che per $m \to \infty$ e per ogni $i = 1, \ldots, h$, ciascuno dei punti $T_{i,m}$ tende allo stesso punto T, che è appunto il centroide.

Per esempio se per le specie A, B, C gli esperimenti forniscono i due alberi mostrati nella parte superiore della seguente figura, con i lati interni della stessa misura, scegliendo il centroide si conclude che le specie hanno avuto un unico progenitore comune.



10 Algebra tropicale

Come abbiamo già visto la geometria algebrica fornisce un contesto naturale nel quale considerare i modelli probabilistici utili per estrarre informazioni dalle sequenze biologiche.

Lo spazio $\mathcal{T}_n$ che appare in maniera naturale in ambito filogenetico non sembra a prima vista munito di una struttura di *varietà algebrica*. È sorprendente però scoprire che invece anche questo spazio *è* una varietà algebrica su un'appropriata struttura algebrica. Tale struttura, nata all'incirca vent'anni fa [36, 43] è quella dell'*algebra tropicale*. Il nome deriva dal fatto che fu sviluppata in collaborazione da alcuni matematici brasiliani e francesi e questi ultimi ne proposero la denominazioni in onore del collega brasiliano Imre Simon.

L'algebra tropicale si può riguardare come ottenuta, a partire dall'algebra classica, con un opportuno procedimento di *passaggio al limite* di cui daremo qualche cenno. Per maggiori dettagli, cfr. [49]. Questo passaggio al limite semplifica gli oggetti algebrico geometrici producendo degli oggetti combinatorici che riflettono molte delle proprietà di quelli di partenza. Applicazioni molto interessanti della geometria algebrica tropicale a pro-

blemi classici di geometria enumerativa sono state recentemente fornite da G. Mikhalkin [32] (cfr. anche [24]).

Il semianello tropicale. L'oggetto fondamentale dell' algebra tropicale è il *semianello* $(\mathbb{R} \cup -\infty, \oplus, \odot)$ dove le operazioni sono definite nel modo seguente:

- la *somma tropicale* $\oplus$ è il *massimo* rispetto all'ordinamento naturale;
- il *prodotto tropicale* $\odot$ è la *somma* ordinaria.

L'*elemento neutro* per $\oplus$ è $-\infty$ e quello per $\odot$ è 0. Si tratta di un semianello e non di un anello perchè la somma non ammette opposto.

Nel semianello tropicale si possono definire i polinomi, che diremo *tropicali*. Essi si interpretano come *funzioni lineari a tratti*. Ad esempio, al polinomio in due variabili

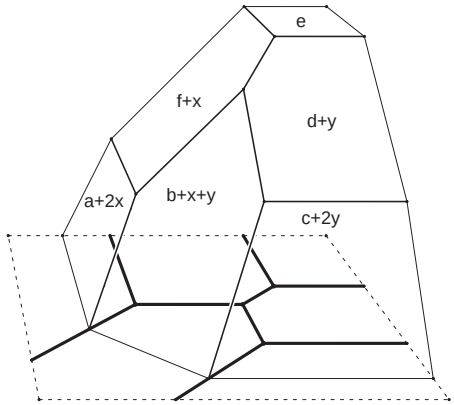
$$a \odot x^{\odot 2} \oplus b \odot x \odot y \oplus c \odot y^{\odot 2} \oplus d \odot x \oplus e \odot y \oplus f \tag{7}$$

corrisponde la funzione lineare a tratti

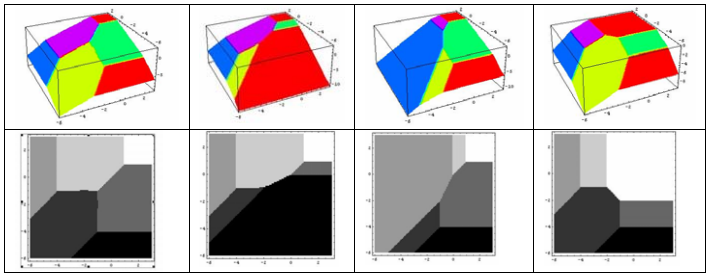
$$\max\{2x + a, x + y + b, 2y + c, x + d, y + e, f\}.$$

Ad un polinomio tropicale è associato il suo *luogo singolare* o *corner locus*. Si tratta dell'insieme dei punti dove la corrispondente funzione lineare a tratti non è differenziabile, ovvero il luogo di punti in cui le espressioni lineari che assumono il massimo sono almeno due. Il luogo singolare di un polinomio tropicale di grado d in n variabili si dice *ipersuperficie tropicale di grado d*. Ad esempio il luogo singolare del polinomio tropicale (7) è una *conica tropicale*.

Le coniche tropicali Nella figura seguente si riporta il grafico di una funzione lineare a tratti il cui *luogo singolare* è la *conica tropicale* disegnata in neretto.



Nella prossima figura sono riportati gli esempi di coniche tropicali provenienti da funzioni lineari a tratti il cui grafico consiste di 6 facce planari. Esse si dicono coniche tropicali *proprie*.



Amebe e curve tropicali Le curve algebriche tropicali, e più in generale le ipersuperfici tropicali, si ottengono a partire dai rispettivi oggetti classici con un procedimento di *limite del logaritmo*.

Il logaritmo *linearizza* e quindi *tropicalizza* i monomi. Inoltre *innesca* la tropicalizzazione dei polinomi in un senso che ci accingiamo a descrivere nel caso delle curve piane (il caso delle ipersuperfici è del tutto analogo).

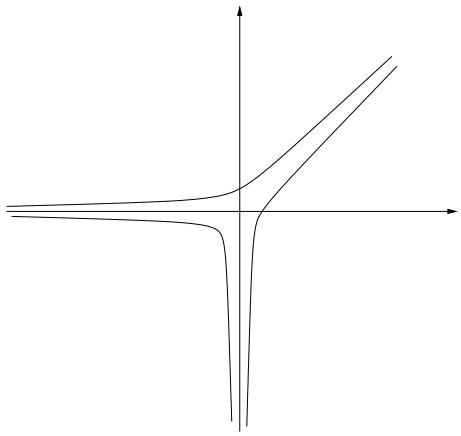
Considerata una curva piana di equazione

$$f(x, y) = 0$$

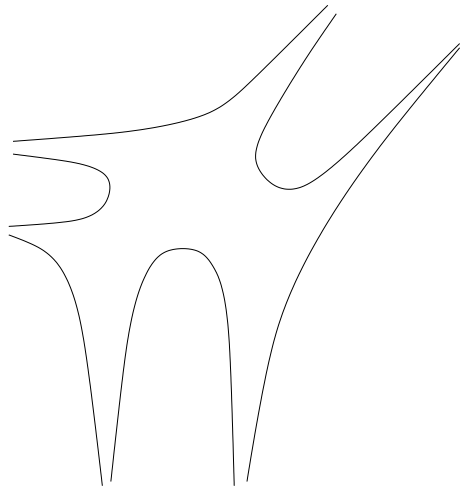
la si trasformi mediante la mappa

$$\begin{array}{rcl} \mathrm{Log}_t : (\mathbb{C}^*)^2 & \to & \mathbb{R}^2 \\ (z_1, z_2) & \mapsto & (\log_t |z_1|, \log_t |z_2|) \end{array}$$

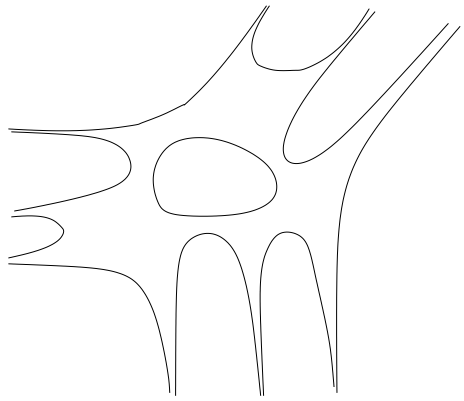
L'immagine si dice *ameba* della curva. Essa cattura basilari proprietà topologiche e geometriche della curva di partenza, come illustrato dai disegni che seguono.



Ameba della retta $z_1 + z_2 = 1$.



Ameba di una conica.



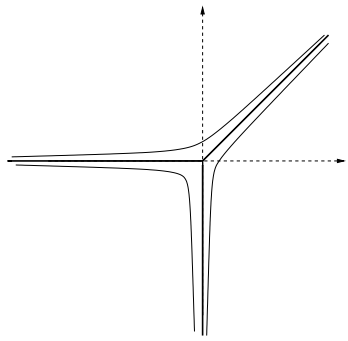
Ameba di una cubica.

Osserviamo che il *numero dei tentacoli* dell'ameba è il triplo del grado della curva: i tentacoli orizzontali corrispondono alle intersezioni della curva con l'asse delle x, quelli verticali alle intersezioni con l'asse delle y e

quelli nella direzione della bisettrice del primo quadrante ai punti all'infinito della curva. Inoltre il numero dei buchi dell'ameba corrisponde al *genere* della curva. Ricordiamo che ogni curva algebrica proiettiva complessa non singolare è, come spazio topologico, una superficie orientabile compatta. La sua caratteristica di Eulero-Poincaré vale $2-2g$ dove g è il genere di C. Per una curva piana non singolare di grado d il genere vale $\binom{d-1}{2}$.

Per una stessa curva $\mathcal{C}$ si ha una famiglia di amebe dipendente dalla base t del logaritmo.

Nel caso della retta di equazione $z_1 + z_2 = 1$, al tendere di t a zero, le amebe *collassano* su tre semirette come mostrato nella seguente figura.



Questo processo di limite, *opportunamente eseguito* sulle amebe di una curva, conduce ad un grafo composto di segmenti e semirette, queste ultime sono limite dei *tentacoli* dell'ameba. Questa è la cosiddetta *tropicalizzazione* della curva $\mathcal{C}$.

Curve definite sul campo delle serie di Puiseux e loro tropicalizzazione. Un modo elegante per evitare le complicazioni legate al processo di limite accennato nel paragrafo precedente consiste nel considerare una famiglia di curve dipendenti da un parametro come una singola curva definita sul campo $\mathbb{K}$ delle serie di *Puiseux* in una variabile complessa (il quale è la chiusura algebrica del campo delle serie formali di Laurent in t). Un elemento

$f = f(t) \in \mathbb{K}$ è una serie formale di potenze

$$f = \sum_{q \in \mathbb{Q}} a_q t^q.$$

tale che il sottoinsieme $A_f = \{a_q \neq 0\}$ sia limitato inferiormente e abbia un insieme finito di denominatori, perciò A_f è dotato di minimo. Su $\mathbb{K}$ è quindi possibile definire una *valutazione*

$$\mathrm{val} : \mathbb{K}^* \to \mathbb{Q}$$

ponendo $\mathrm{val}(f) = \min A_f$.

La valutazione è legata al logaritmo. Infatti,

$$\log_t(f) = \log_t \left(a_{\mathrm{val}(f)} t^{\mathrm{val}(f)} (1 + g(t)) \right)$$

dove $g(0) = 0$. Quindi, quando ha senso considerare f come funzione analitica (polidroma), $\log_t(f)$ tende, al tendere di t a zero, a

$$\log_t \left(a_{\mathrm{val}(f)} t^{\mathrm{val}(f)} \right) = \mathrm{val}(f) + \log_t \left(a_{\mathrm{val}(f)} \right)$$

e, in ultima analisi a $\mathrm{val}(f)$.

La teoria delle curve algebriche piane a coefficienti in $\mathbb{K}$ è formalmente identica a quella delle curve piane complesse. Un polinomio in due variabili $F \in \mathbb{K}[x, y]$ determina una curva algebrica piana

$$V(F) = \{(f, g) \in \mathbb{K}^2 \mid F(f, g) = 0\} \subseteq \mathbb{K}^2.$$

Possiamo restringere a $V_0(F) = V(F) \cap (\mathbb{K}^*)^2$ la mappa

$$\begin{array}{rccc} \mathrm{Val} : & (\mathbb{K}^*)^2 & \to & \mathbb{Q}^2 \\ & (f, g) & \mapsto & (-\mathrm{val}(f), -\mathrm{val}(g)) \end{array}$$

Come abbiamo visto, se è possibile riguardare f e g come funzioni analitiche, si ha $\lim_{t \to 0} \mathrm{Log}_t(f, g) = \mathrm{Val}(f, g)$ ed è dunque naturale sostituire a limiti di amebe di famiglie di curve algebriche piane complesse oggetti del tipo

$\mathrm{Trop}(\mathcal{C})$ che sono chiusura dell'immagine di $\mathrm{Val}(V_0(F))$ in $\mathbb{Q}^2$, $F \in \mathbb{K}[x,y]$. Chiamiamo *curva piana tropicale* ogni sottoinsieme del tipo $\mathrm{Trop}(\mathcal{C}_F)$, dove $\mathcal{C}_F$ è la curva algebrica di equazione $F = 0$ con $F \in \mathbb{K}[x,y]$. Questa definizione è puramente algebrica e non fa intervenire nessuna nozione di limite.

Consideriamo per esempio la curva $\mathcal{C}$ di equazione $F(x,y) := t^{-3}x + t^{-2}y - 1 = 0$. Per ogni $f \in \mathbb{K}$ esiste una g che $(f,g) \in \mathcal{C}$. Se la valutazione di f è maggiore di 3, allora quella di g deve valere 2. Questo dimostra che la semiretta costituita dai punti con coordinate del tipo $(-3-\lambda, -2)$, con $\lambda > 0$ è contenuta in $\mathrm{Trop}(\mathcal{C})$.

Se (f,g) è soluzione e la valutazione di g è maggiore di 2 allora la valutazione di f deve valere 3 e si ha quindi la semiretta $(-3,-2) - \lambda(0,1)$, $\lambda > 0$.

Se (f,g) è soluzione, la valutazione di g è minore o uguale a 2 e la valutazione di f è minore o uguale a 3 allora deve essere $\mathrm{val}(f) = \mathrm{val}(g)+1$ e si ha quindi la semiretta $(-3,-2) + \lambda(1,1)$, $\lambda > 0$.

In definitiva $\mathrm{Trop}(\mathcal{C})$ è unione delle tre semirette $(-3-\lambda, -2)$, $(-3, -2-\lambda)$ e $(-3+\lambda, -2+\lambda)$, con $\lambda > 0$.

Notiamo che $\mathrm{Trop}(\mathcal{C})$ non è altro che il corner locus della funzione lineare a tratti corrispondente al polinomio tropicale

$$3 \odot x \oplus 2 \odot y \oplus 0.$$

Questo è un fatto generale. Dato un polinomio

$$F = \sum a_{ij}(t) x^i y^j \qquad \in \mathbb{K}[x,y]$$

se ne definisce la *tropicalizzazione*

$$\mathrm{Trop}(F) = \oplus -\mathrm{val}(a_{ij}(t)) \odot x^{\odot i} \odot y^{\odot j}.$$

Allora $\mathrm{Trop}(\mathcal{C})$ è il corner locus della funzione lineare a tratti corrispondente a $\mathrm{Trop}(F)$. Questa si dice *tropicalizzazione* della curva $\mathcal{C}$ di equazione $F = 0$.

Tropicalizzazione di varietà. La *tropicalizzazione* si può effettuare per ogni varietà algebrica. Si tratta di un processo di *linearizzazione* e di *limite*, analogo a quello descritto precedentemente per le curve.

Sia X una varietà algebrica definita sul campo $\mathbb{K}$ delle serie di Puiseaux. Per ogni polinomio F che si annulla su X, consideriamo il corrispondente *polinomio tropicale* $\mathrm{Trop}(F)$. L'intersezione del luogo singolare delle funzioni lineari a tratti corrispondenti a $\mathrm{Trop}(F)$ al variare di F nell'ideale di X è un *complesso poliedrale* detto *tropicalizzazione* di X.

Si noti che l'ideale di X è finitamente generato. Tuttavia non è detto che se $F_1, \ldots, F_h$ sono un insieme di generatori allora la intersezione delle ipersuperfici tropicali associate a $\mathrm{Trop}(F_1), \ldots, \mathrm{Trop}(F_h)$ coincida con la tropicalizzazione di X. Tuttavia esiste un algoritmo per determinare un insieme finito di generatori per cui invece ciò accada [7].

Come abbiamo accennato dalla tropicalizzazione di una varietà si possono ottenere diverse informazioni concernenti la varietà di partenza. Ad esempio, oltre alle proprietà già indicate per le curve, si ha che la *dimensione topologica* della tropicalizzazione di una varietà coincide con la *dimensione* della varietà stessa [3].

Grassmanniana tropicale e spazi $\mathcal{T}_n$. Ricordiamo che una metrica d è *arborea* se e solo se vale la condizione dei quattro punti, ossia *per ogni scelta* di indici $1 \leq i_0 < i_1 < j_0 < j_1 \leq n$ il *massimo* delle tre quantità $d_{i_0,i_1} + d_{j_0,j_1}$, $d_{i_0,j_0} + d_{i_1,j_1}$, $d_{i_0,j_1} + d_{j_0,i_1}$ *viene raggiunto almeno due volte* ($d_{i,j}$ è la lunghezza dell'arco tra i e j).

La *riformulazione tropicale* della condizione dei quattro punti è la seguente: *i numeri $d_{i,j}$ appartengono al luogo singolare di ognuno dei polinomi tropicali*

$$X_{i_0,i_1} \otimes X_{j_0,j_1} \oplus X_{j_0,i_1} \otimes X_{i_0,j_1} \oplus X_{j_1,i_1} \otimes X_{i_0,j_0}. \tag{8}$$

Questi polinomi sono la *tropicalizzazione* dei polinomi

$$p_{i_0,i_1}p_{j_0,j_1} + p_{i_1,j_0}p_{i_0,j_1} + p_{j_1,i_1}p_{i_0,j_0}$$

che sono i *generatori dell'ideale* di una ben nota varietà algebrica, la *Grassmanniana* $\mathbb{G}(1, n-1)$ *delle rette dello spazio proiettivo di dimensione* $n-1$.

Questa è a sua volta una varietà proiettiva di dimensione $2(n-2)$ i cui punti corrispondono alle rette di $\mathbb{P}^{n-1}$.

La grassmanniana tropicale coincide con l'insieme delle matrici di dissimilarità arboree. Tra queste ci sono le matrici $T(a_1, \ldots, a_n)$ con $T_{ij} = a_i + a_j$, che sono l'immagine di $\mathbb{R}^n$ con la mappa

$$\phi : \mathbb{R}^n \to \mathbb{R}^N$$

dove $N = \binom{n}{2}$ e $\phi(a_1, \ldots, a_n)_{ij} = a_i + a_j$. Gli alberi pesati associati a queste matrici hanno un solo vertice interno. Si noti che matrici di questo tipo non hanno necessariamente entrate non negative.

Ogni matrice di dissimilarità D si può scrivere come $D' + T(a_1, \ldots, a_n)$, dove D' è una matrice di dissimilarità con la proprietà che la dissimilarità delle ciliegie è nulla. Le matrici D' corrispondono agli alberi pesati per cui gli archi che contengono le foglie hanno peso zero. Posiamo quindi identificare l'immagine della grassmanniana tropicale in $\mathbb{R}^N/\phi(\mathbb{R}^n)$ con lo spazio $\mathcal{T}_{n-1}$ di Billera, Holmes e Vogtmann.

È anche possibile introdurre la nozione di *retta tropicale* e dimostrare che la grassmanniana tropicale parametrizza l'insieme delle rette tropicali.

Mappa di m-dissimilarità. Una semplificazione meno drastica dei dati si può ottenere considerando, al posto della matrice delle distanze tra le n specie, *mappe di m-dissimilarità*. Si tratta di applicazioni

$$D : \{1, \ldots, n\}^m \to \mathbb{R}$$

tali che $D(i_1, \ldots, i_m) = D(i_{\sigma(1)}, \ldots, i_{\sigma(m)})$ per ogni permutazione $\sigma \in S_m$ e valgono zero sulle m-ple di indici non tutti distinti.

Ad ogni albero pesato con n foglie è associata una mappa di m dissimilarità definendo $D(i_1, \ldots, i_m)$ come il *peso del minimo sottoalbero* contenente le foglie etichettate con $i_1, \ldots, i_m$. L'albero è *univocamente determinato* da questa mappa se $n \geq 2m - 1$ [34].

Ad una serie di dati si può associare una mappa di m dissimilarità a partire dalle distanze di Jukes-Cantor tra le foglie. Come nel caso $m = 2$ il problema è quello di capire se questa mappa di dissimilarità *è arborea*, ovvero di determinare una sua approssimazione arborea.

Neighbour joining generalizzato e grassmanniane tropicali generali Esiste una generalizzazione dell'algoritmo di Neighbour Joining che fornisce l'approssimazione arborea di una mappa di dissimilarità. Sembra che adoperando mappe di m dissimilarità con $m > 2$ si ottengano migliori approssimazioni dell'albero di massima verosimiglianza [34].

Una mappa di m dissimilarità arborea è completamente determinata dalla matrice delle distanze, ossia dalla mappa di 2 dissimilarità. Tuttavia, a differenza del caso $m = 2$, non è nota, in generale, una completa *caratterizzazione tropicale* delle mappe di m dissimilarità.

A riguardo c'è una recente ed interessante ricerca di *Bocci e Cools* [6]:

- Sia $\mathcal{T}(m, n)$ la *varietà grassmanniana tropicale a tre termini* le cui equazioni tropicali si ottengono tropicalizzando solo le cosiddette *equazioni a tre termini* della grassmanniana $\mathbb{G}(m, n)$ (studiate classicamente da B. Segre, L. Bassotti et al., [1].).

- Le mappe di m dissimilarità sono contenute in $\mathcal{T}(m, n)$.

- Per $m = 3$ sono tutte e sole gli elementi dell'intersezione di $\mathcal{T}(3, n)$ con una ben determinata sottovarietà tropicale.

Queste considerazioni suggeriscono dunque delle estensioni che coinvolgono *grassmanniane tropicali più generali*. Esse non hanno soltanto un interesse speculativo ([44]) ma possono anche essere utili per le applicazioni biologiche.

Come suggerisce il risultato di Bocci e Cools, per $m > 1$, la tropicalizzazione della grassmanniana $\mathbb{G}(m, n)$, o meglio, della *grassmanniana a tre termini*, si lega alle mappe di dissimilarità.

A differenza del caso della grassmanniana delle rette questo approccio è ancora un terreno di ricerca scarsamente esplorato sia dal punto di vista matematico che da quello della applicazioni alla biologia, e merita dunque ulteriori ricerche.

Riferimenti bibliografici

[1] Bassotti L., Sulle relazioni a tre termini fra le coordinate di Grassmann. *R. Acc. Linc.*, v. XX (1956), Nota I, 200-204, Nota II, 318-325,

[2] Bickel P. J., Doksum K. A., *Mathematical statistics: Basic ideas and Selected Topics*, Prentics Hall, (2000).

[3] Bieri R., Groves J. R. J., The geometry of the set of characters induced by valutations. *J. Reine Angew. Math.* 347 (1984), 168-195.

[4] L. J. Billera, S. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. Advances in Applied Mathematics 27 (2001) 733-767.

[5] Bayesian Networks: A Practical Guide to Applications, Olivier Pourret (Editor), Patrick Naïm (Co-Editor), Bruce Marcot (Co-Editor), Chichester, UK: Wiley. (2008).

[6] Bocci C., Cools C., A tropical interpretation of m-dissimilarity maps, arXiv:0803.2184.

[7] Bogart T., Jensen A., Speyer D., Sturmfels B., Thomas R., Computing Tropical Varieties, *Journal of Symbolic Computation* 42, 1-2, (2007), 54-73.

[8] Borel E. *Les probabilités et la vie.* Paris, Presses Universitaires de France, (1943).

[9] Bray N., Morton J. Equations Defining Hidden Markov Models, in [35], pp. 237-249.

[10] Buczynska W., Wisniewski J., On phylogenetic trees - a geometer?s view, math.AG/0601357.

[11] Buneman P., A note on the metric properties of trees, *Journal of Combinatorial Theory (B), 17 (1974) 48-50.*

[12] Cox D., Little J. O'Shea D., *Ideals, Varieties and Algorithms*, second edition, Springer-Verlag, New York, 1996.

[13] Cox D., *Introduction to Gröbner basis* in Applications of Computational Algebraic Geometry, Proceedings of Symposia in Applied Mathematics, vol. 53, ed. Cox D. and Sturmfels B., American Mathematical Society

[14] Cohen J. E., Mathematics is Biology's next microscope, only better; biology is mathematics' next physics, only better, PLOS Biology 2 (2004), N. 12.

[15] Durbin R., Eddy S., Krogh A., Mitchison G., *Biological sequence analysis. Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University press, (1998).

[16] Eisenbud D., Commutative Algebra: with a View Toward Algebraic Geometry, New York, Springer, (1995).

[17] Eriksson *Simulation studies of algebraic invariants for phylogeny* preprint.

[18] Eriksson N., Ranestad K., Sturmfels B., Sullivant S., math/0407033 *Phylogenetic algebraic geometry* in *Projective Varieties with Unexpected properties* Ciliberto C., Geramita A. V., Harbourne B., Mirò-Roig R. M., Ranestad K. (eds.), Berlin, Walter de Gruyter (2005), 237-256.

[19] Felsentein J. *Inferring phylogenies*, Sinauer associates, Sunderland, Massachusetts, (2004).

[20] Fitch W. M., Toward defining the course of evolution: Minimum change for a specified tree topology, *Systematic Zoology*, 20 (1971), 406-416.

[21] Fitch W. M., Margoliash E. Construction of phylogenetic trees, *Science* 155, (1967), 279-284.

[22] Fischer R., *The Genetical Theory of Natural Selection*, Clarendon Press, Oxford (1930).

[23] Fulton, W. Introduction to Toric Varieties. Princeton, NJ: Princeton University Press, (1993).

[24] Gathmann A., *Tropical algebraic geometry* arXiv:math.AG/0601322

[25] Hardy, G. H. Mendelian proportions in a mixed population. *Science* 28, (1908), 49 ? 50.

[26] Israel G., *La visione matematica della realtà¿ Introduzione ai temi e alla storia della modellistica matematica*, Bari, Laterza, (1996).

[27] Kirkpatrick S., Gelatt C. D., Vecchi M. P. Optimization by Simulated Annealing, Science, Vol 220, Number 4598, (1983), pp 671-680.

[28] Landsberg J. M., Manivel L., On the ideals of secant varieties of Segre varieties, Found. Comput. Math. 4 (2004) 397-422.

[29] Lotka A. J., *Elements of Physical Biology*, Baltimore, Wiliams and Wilkins, (1925).

[30] Lundy M., Applications of the annealing algorithm to combinatorial problems in statistics. *Biometrika*, 72 (1985) 191-198.

[31] Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E.. Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics, 21(6) (1953), 1087-1092.

[32] Mikhalkin G., *Tropical Geometry*,

`http:\\www.math.toronto.edu/mikha/book.pdf`

[33] Bayer D., Mumford D. "What can be computed in algebraic geometry" in *Computational Algebraic Geometry and Commutative Algebra* Eisenbud E. and Robbiano L. editors, Cambridge U. Press, Cambridge, 1993, 1-48.

[34] Pachter L., Sturmfels B. The Mathematics of Phylogenomics, SIAM Review, 49, (2007), pp. 3?31.

[35] Pachter L., Sturmfels B. (eds.) *Algebraic Statistics for Computational Biology* Cambridge, Cambridge University Press, (2005).

[36] Pin J.-E., Tropical semirings, *Idenpotency* (Bristol 1994), 50-69, Publ. Newton Inst. Cambridge University Press, Cambridge, 1998.

[37] Rabiner L. R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, 77 (2), (1989), p. 257?286.

[38] Kac M., Rota G. C., Schwartz J. T. *Discrete Thoughts*, Birkäuser, Boston, 1986.

[39] Saitou N., Nei M., The neighbor-joining method: A new method for constructing phylogenetic trees, *Molecular Biology and evolution*, 4, (1987), 406-425

[40] Sankoff D. Minimal mutation trees of sequences. SIAM Journal of Applied Mathematics, 28, (1975), 35-42.

[41] Segre B., *Prodromi di Geometria Algebrica*, Cremonese, Roma, (1972).

[42] Sokal R. R., Michener C. D. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, (1958), 1409-1438.

[43] Simon I., Recognizable sets with multiplicities in the tropical semiring. Mathematical foundations of computer science (Carlsbad, 1988), 107-120, Lecture Notes in Comp. Sci., 324, Springer, Berlin, 1988.

[44] Speyer D., Sturmfels B., *The tropical grassmannian* arXiv:math.AG/0304218

[45] Sturmfels B., Can biology lead to new Theorems? In Annual Report 2005 of the Clay Mathematics Institute.

[46] Sturmfels B., Sullivant S., Toric ideals of Phylogenetic invariants, *Journal of computational biology* 12 (2005) 204-228.

[47] Volterra V., Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Memoria della R. Acc. dei Lincei*, S. 6, vol. 2, p. 31-113, (1926).

[48] Van der Waerden B. L., *Moderne Algebra*. 2d edition. Part 1. Berlin, Springer, 1937

[49] Viro O., Dequantization of Real Algebraic Geometry on a Logarithmic Paper, Proceedings of the 3rd European Congress of Mathematicians, Birkhäuser, Progress in Math, 201, (2001), 135–146.

[50] Wright, S. *Evolution and the Genetics of Populations: Genetics and Biometric Foundations*, 4 voll. University of Chicago Press (1968, 1969, 1977, 1978).