Algebra, geometria e inattese applicazioni

Ciro Ciliberto

TFA – ROMA TOR VERGATA, 14 Febbraio, 2013

Vorrei parlare ...

... di matematica solitamente ritenuta astratta: cioè ALGEBRA e GEOMETRIA, o, più precisamente, di GEOMETRIA ALGEBRICA. Anche se ...

الجبر

al–ğabr in arabo significa unione, connessione, completamento, ma anche aggiustare, deriva dal nome del libro del matematico persiano arabo Muhammad ibn Musa al–Kwarizmi (780–850), intitolato Al-Kitab al-Jabr wa-l-Muqabala, cioè Compendio sul Calcolo per Completamento e Bilanciamento, che tratta la risoluzione delle equazioni di primo e di secondo grado in vista di applicazioni a problemi molto concreti.

L'algebra ...

... infatti fu importata in occidente nel secolo XIII per motivi assai pratici, cioè per far di conto negli affari, principalmente da Leonardo Fibonacci (1170–1250), autore del Liber Abaci e Practica Geometriae.

Il primo a usare il termine algebra nel mondo occidentale fu il maestro d'abaco fiorentino Raffaello di Giovanni Canacci, autore dei Ragionamenti di Algebra (1490).

Geometria ...

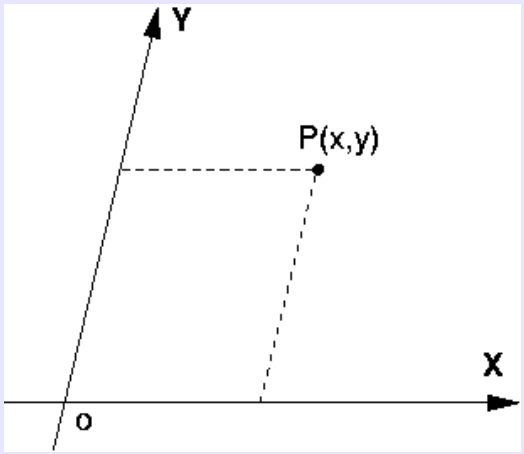
... viene dal greco

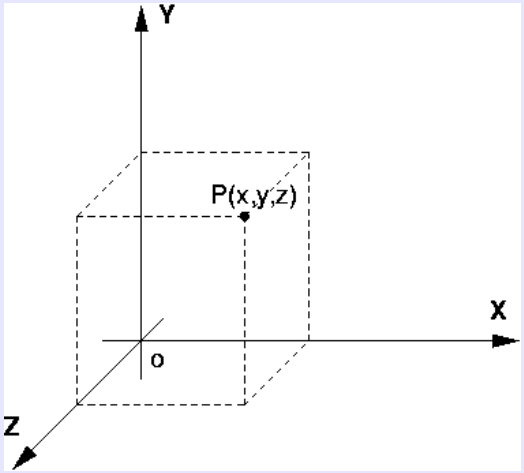
$$\gamma\epsilon\omega\mu\epsilon\tau\rho\iota\alpha$$

che significa misura della terra. Dunque le sue origini molto concrete sono fuori di ogni dubbio.

La Geometria Algebrica ...

... può pensarsi come la naturale prosecuzione del cammino intrapreso da R. Descartes (1596–1650) con l'introduzione delle coordinate cartesiane.





Ora prima un po' di storia, poi qualche nozione volta ad introdurre il concetto di razionalità, che con Cartesio ci sta proprio bene! Poi mi concentrerò su alcune questioni volte a mostrare come questo concetto, apparentemente astratto, si ritrovi nell'affrontare problemi molto concreti.

Un po' di storia (I)

La Geometria Algebrica nasce come studio di curve del piano e superficie dello spazio definite da una equazione polinomiale e successivamente di varietà nello spazio definite da sistemi di equazioni polinomiali.

Alla base dell'interesse in queste questioni c'erano problemi derivanti dalle applicazioni alla Fisica, all'Ingegneria, all'Architettura e all'Arte.

L'idea di spazio proiettivo, ambiente naturale di lavoro in Geometria Algebrica, trae origine dagli studi sulla prospettiva dei pittori del rinascimento e diviene strumento consolidato per le applicazioni nell'insegnamento di G. Monge (1746–1818) presso l' École Polytechnique di Parigi.

Un po' di storia (II)

Intorno alla metà dell'ottocento la Geometria Algebrica diviene disciplina autonoma.

Con J. Plücker (1801–1868) e H. G. Grassmann (1809–1877) si iniziano a studiare varietà algebriche in spazi proiettivi di dimensione qualunque, ad esempio le varietà i cui punti corrispondono ai sottospazi lineari di data dimensione di uno spazio proiettivo fissato, dette varietà di Grassmann, o grassmanniane.

Un po' di storia (III)

Questo indirizzo di ricerca trova successivamente ampio sviluppo presso la Scuola Italiana di Geometria Algebrica fondata da L. Cremona (1830–1903).

Contributi fondamentali allo studio delle varietà algebriche si devono a G. Veronese (1854–1917) e C. Segre (1863–1924), da cui prendono il nome due importanti classi di varietà.

La Scuola Italiana si sviluppa poi con G. Castelnuovo (1865–1952), F. Enriques (1871–1946) e F. Severi (1879–1961), i cui contributi alla classificazione delle curve e superficie algebriche gettano le basi per lo sviluppo della Geometria Algebrica contemporanea.

Lo Spazio Proiettivo Complesso $\mathbb{P}^n$ ...

... è ottenuto aggiungendo i punti all'infinito, cioè le direzioni delle rette, allo Spazio Affine Numerico Complesso

$$\mathbb{A}^n = \mathbb{C}^n$$

i cui punti sono le n–ple ordinate $(x_1, \ldots, x_n)$ di numeri complessi.

Analiticamente, i punti di $\mathbb{P}^n$ sono $(n+1)$-ple $[x_0, \ldots, x_n]$ non nulle, date a meno di un fattore di proporzionalità, che si dicono coordinate omogenee del punto.

Allora $\mathbb{A}^n$ si identifica con il sottoinsieme dei punti $[1, x_1, \ldots, x_n]$ di $\mathbb{P}^n$.

L'insieme complementare è costituito dai punti del tipo $[0, x_1, \ldots, x_n]$. Un tale punto si identifica con il punto all'infinito, cioè la direzione, delle rette parallele al vettore non nullo $(x_1, \ldots, x_n)$.

Uno dei pregi dello spazio proiettivo è di essere compatto.

Prodotti di spazi proiettivi

Il prodotto di due spazi affini è uno spazio affine; il prodotto di due spazi proiettivi invece non è uno spazio proiettivo.

Tuttavia esso si immerge in modo naturale in un opportuno spazio proiettivo più grande.

Applicazioni di Segre e varietà di Segre

L'applicazione di Segre si definisce nel seguente modo

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

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

Essa è iniettiva. La sua immagine

$$S_{m,n}$$

è una varietà algebrica, che si chiama varietà di Segre.

Analogamente per il prodotto di più di due spazi proiettivi.

Varietà algebriche

Un sottoinsieme $V \subseteq \mathbb{P}^N$ è varietà algebrica se esistono polinomi omogenei

$$F_i(x_0, \ldots, x_N), \quad i = 1, \ldots, h$$

tali che V coincida con l'insieme dei punti $[x_0, \ldots, x_N]$ di $\mathbb{P}^N$ le cui coordinate omogenee annullano tutti i polinomi F_i.

Nota bene: 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.

Le equazioni

$$\boxed{F_i(x_0, \ldots, x_N) = 0, \quad i = 1, \ldots, h}$$

si dicono equazioni della varietà V.

Equazioni delle varietà di Segre

I determinanti di ordine due subordinati a una matirce

$$Z = (z_{ij})_{i=0,\dots,m;j=0,\dots,n}$$

di variabili sono polinomi omogenei di grado 2 nelle z_{ij}.

La varietà di Segre $S_{m,n}$ ha per equazioni tali polinomi. Essa coincide con l'insieme delle classi di proporzionalità delle matrici di rango 1.

Ad esempio l'equazione

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

definisce una quadrica nello spazio proiettivo a tre dimensioni, che è il prodotto di Segre $S_{1,1}$ di due rette proiettive $\mathbb{P}^1$.

Dimensione

La dimensione di una varietà $V \subseteq \mathbb{P}^N$ è il numero di parametri da cui dipendono i suoi punti:

$$\dim(V) \geq N - \text{numero di equazioni che definiscono } V$$

Ad esempio:

- $\mathbb{P}^n$ ha dimensione n;
- $S_{m,n} = \mathbb{P}^n \times \mathbb{P}^m$ ha dimensione $n + m$;
- una curva piana definita da una sola equazione in $\mathbb{P}^2$ ha dimensione 1;
- una superficie definita da una sola equazione in $\mathbb{P}^3$ ha dimensione 2.

Razionalità

Una varietà $V \subseteq \mathbb{P}^N$ di dimensione n è unirazionale se esiste un'applicazione suriettiva

$$f : \mathbb{P}^n \to V$$

definita da polinomi omogenei dello stesso grado.

Ossia V è unirazionale se è possibile parametrizzare i punti di V nel modo seguente

$$x_i = p_i(y_0, \ldots, y_n), \quad i = 0, \ldots, N$$

$$\text{con} \quad p_i \quad \text{polinomi} \quad \text{omogenei} \quad \text{dello} \quad \text{stesso} \quad \text{grado}$$

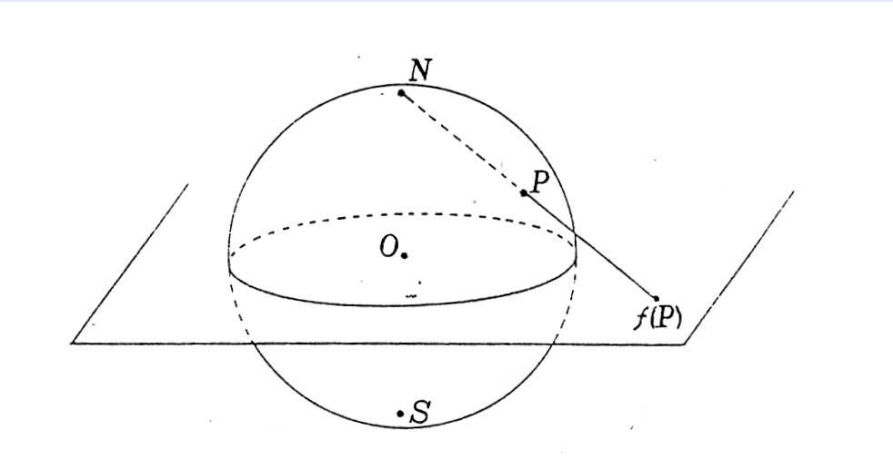
o, passando a coordinate affini

$$X_i = q_i(Y_1, \ldots, Y_n), \quad i = 1, \ldots, N, \quad \text{con} \quad q_i \quad \text{funzioni} \quad \text{razionali.}$$

V è razionale se la parametrizzazione f è generalmente invertibile ossia invertibile su un aperto non vuoto.

Esempi

- gli spazi proiettivi e i loro sottospazi sono razionali;
- le varietà di Segre sono razionali;
- le coniche o in generale le quadriche irriducibili, definite cioè da un polinomio irriducibile di grado 2, sono razionali: si possono parametrizzare razionalmente mediante la proiezione stereografica.

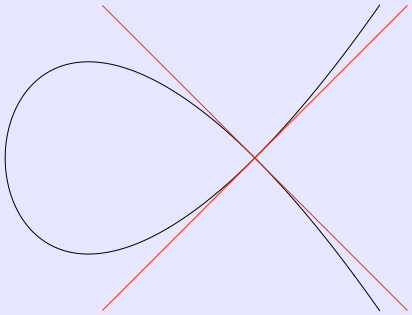


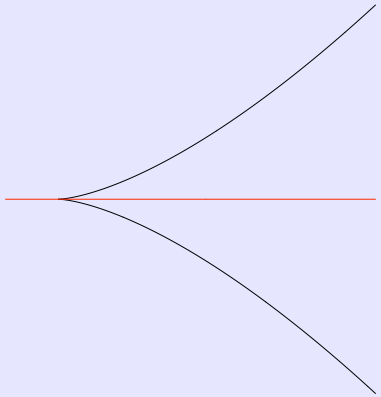
Il nome proiezione stereografica (dal greco $\sigma\tau\epsilon\rho\epsilon o\nu$ =corpo solido e $\gamma\rho\alpha\phi\eta$ =disegno) fu introdotto dal gesuita F. d'Aquilon (1567–1617), autore del geniale trattato di ottica Opticorum libri sex, illustrato con acqueforti di P. P. Rubens (1577–1640), pensato anche per architetti, astronomi, naviganti, ingegneri militari, pittori.

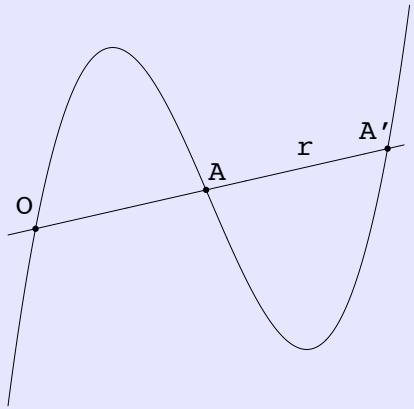
... e le curve di grado superiore?

Il grado di una curva è il minimo grado di una sua equazione e coincide col numero di punti a comune tra la curva e una retta generale del piano. Questo è il Teorema di Bezout.

- Le cubiche singolari sono razionali;
- Le cubiche non singolari NON sono razionali!







Come si fa a riconoscere la razionalità?

Per le curve si introduce un invariante, il genere g.

Se la curva è piana di grado d con δ punti doppi, si ha la formula di Clebsch

$$\boxed{g = \frac{(d-1)(d-2)}{2} - \delta}$$

Teorema di Clebsch: una curva è razionale se e solo se $g = 0$.

Per le superficie si ha un (ben più complicato) teorema analogo, il criterio di Castelnuovo (1893).

Curve e superficie sono razionali se e solo se sono unirazionali.

Per le varietà di dimensione superiore ciò non è sempre vero, nè si hanno criteri generali di unirazionalità o razionalità. Il problema di riconoscere se una varietà è unirazionale o razionale è tuttora aperto.

Le varietà razionali, che apparentemente sembrano le più semplici, perchè più vicine di altre agli spazi proiettivi, in realtà non sono affatto tali, vista la difficoltà perfino di riconoscerle.

Ora mi concentrerò su ...

... due temi, il primo classico, il secondo recentissimo:

- integrazione di funzioni algebriche e l'importanza di essere RAZIONALE;
- algebra, geometria e ... biologia.

Il legame tra loro è fornito proprio dal concetto di razionalità e dalle tecniche della geometria algebrica, che, a dispetto della loro astrazione (o forse proprio a causa di questa!), hanno trovato in passato, e trovano ancora notevoli e talvolta inattese applicazioni in campi molto differenti.

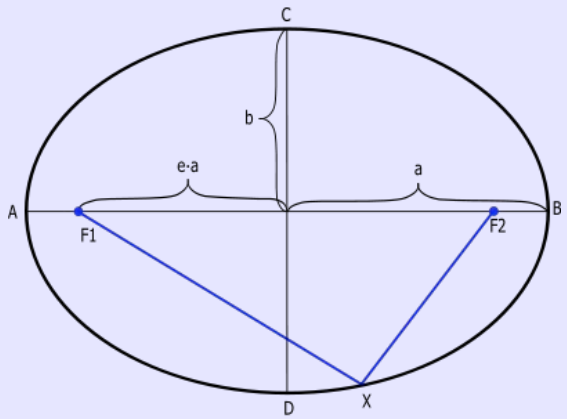
Gli integrali ellittici ...

... si presentano nel calcolo della lunghezza di un arco di ellisse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \quad \text{con} \quad a > b > 0$$

che si può parametricamente rappresentare mediante le equazioni

$$x = a \sin \phi, \quad y = b \cos \phi, \quad \phi \in [0, 2\pi).$$



L'ellisse è il luogo dei punti X del piano la cui distanza da due punti fissi F_1, F_2 detti fuochi, è costante, uguale a $2a$. Se si indica con $2c$ la distanza tra i fuochi e si pone $b = \sqrt{a^2 - c^2}$, allora $2a$ e $2b$ sono le lunghezze degli assi.

Il calcolo della lunghezza dell'arco di ellissi ...

... si incontra in vari problemi di natura applicativa. Ad esempio:

- in astronomia: le orbite dei pianeti sono ellissi;
- in architettura: calcolo di aree di volte a botte di forma ellittica;
- in statica: flessione di aste caricate di punta;
- in navigazione: il calcolo della lunghezza delle geodetiche su un ellissoide di rotazione

Alcuni di questi problemi, ed altri ancora, si posero ai matematici fin dagli albori del calcolo differenziale e integrale.

Considerazioni sugli integrali ellittici si trovano in J. Wallis (1616–1703), che tra il 1643 e il 1689 fu capo crittografo del Parlamento del Regno Unito e successivamente della corte reale.

Altri precursori della teoria degli integrali ellittici furono Giacomo (1654–1705) e Giovanni Bernoulli (1667–1748).

Il calcolo della lunghezza dell'arco di ellisse...

... conduce all'integrale

$$\int \sqrt{\frac{1-k^2x^2}{1-x^2}}\,dx \quad \text{dove} \quad 0 < k^2 = \frac{a^2-b^2}{a^2} < 1$$

Questo, a prima vista non è tanto dissimile dall'innocente

$$\arcsin x = \int \frac{1}{\sqrt{1-x^2}}\,dx$$

che, si integra con la semplice sostituzione $x = \sin t$, ed è una funzione elementare.

Integrali dello stesso tipo di quest'ultimo sono

$$\log x = \int \frac{1}{x}\,dx, \quad \operatorname{arctg} x = \int \frac{1}{1+x^2}\,dx.$$

Diversamente da questi integrali ...

... non esistono sostituzioni che coinvolgano funzioni elementari come seno, coseno, tangente, esponenziale e loro inverse, che semplifichino l'integrale ellittico

$$\int \sqrt{\frac{1-k^2x^2}{1-x^2}}\,dx$$

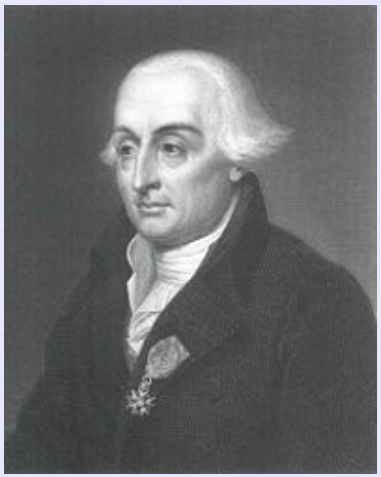
Proviamo a renderci conto della ragione profonda di questo fatto.

J. L. Lagrange (1736–1813) ...

... fu il primo ad osservare che gli oggetti di cui ci stiamo occupando rientrano nella classe molto generale di integrali della forma

$$\int R(x, \sqrt{f(x)})dx$$

dove $R(x, y)$ è una funzione razionale, e $f(x)$ è un polinomio.



Qui entrano in gioco geometria e razionalità!

Consideriamo la curva piana C di equazione

$$y^2 = f(x).$$

Se C è razionale, allora può rappresentarsi parametricamente mediante equazioni del tipo

$$x = p(t), \quad y = q(t), \quad \text{con} \quad p(t), q(t) \quad \text{funzioni} \quad \text{razionali}$$

Sostituendo queste espressioni in $\int R(x, \sqrt{f(x)})dx$, esso diventa del tipo

$$\int R(p(t), q(t))p'(t)dt$$

cioè un integrale di funzione razionale, che si può esprimere in termini di funzioni elementari.

Questo è il caso dell'integrale

$$\arcsin x = \int \frac{1}{\sqrt{1-x^2}}dx.$$

in cui C è una conica.

Invece, per l'integrale ellittico ...

$$\int \sqrt{\frac{1-k^2x^2}{1-x^2}}dx = \int \frac{\sqrt{(1-k^2x^2)(1-x^2)}}{1-x^2}dx$$

la curva cubica C di equazione

$$y^2 = (1-k^2x^2)(1-x^2)$$

non è razionale, e questa è la ragione per cui l'integrale non può esprimersi in termini di funzioni elementari.

Gli integrali che abbiamo considerato rientrano nel tipo ancora più generale

$$\int R(x,y)dx$$

con $R(x,y)$ funzione razionale e x e y sono legati da una relazione del tipo

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

con g polinomio irriducibile, che definisce una curva piana Γ.

Questi integrali si dicono abeliani.

Contributi alla teoria degli integrali abeliani

- Giulio de' Toschi, conte di Fagnano (1682–1766);
- L. Euler (1707–1783);
- A. M. Legendre (1752–1833);
- C. F. Gauss (1777–1855).

e ...

- E. Galois (1811–1832);
- C. G. Jacobi (1804–1851);
- N. Abel (1802–1829).

Il contributo decisivo di B. Riemann (1826–1866)

Riemann sposta definitivamente l'attenzione dagli oggetti analitici, cioè gli integrali abeliani, all'oggetto geometrico costituito dalla curva algebrica Γ sul campo complesso di equazione

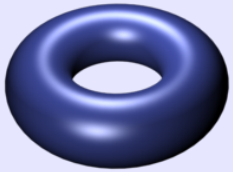
$$g(x, y) = 0$$

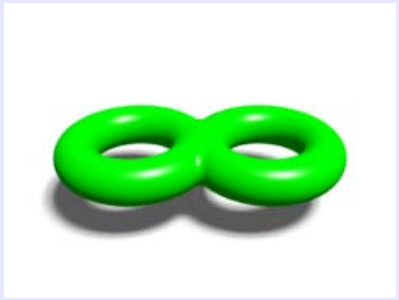
cui essi sono legati.

Riemann ...

... presuppone l'esistenza (provata solo più tardi) di un modello liscio X per Γ, che è una curva senza singolarità in qualche spazio proiettivo $\mathbb{P}^N$.

Il modello liscio X, dal punto di vista topologico, è una superficie orientabile compatta, omeomorfa ad una sfera con un certo numero g di manici attaccati.





Il numero g è denominato da Riemann numero di classe (Klassenzahl) della curva, mentre un po' più tardi verrà detto genere (Geschlecht) da A. Clebsch (1833–1872), ed è proprio l'invariante di cui abbiamo già parlato.

Superficie di Riemann

La curva algebrica non singolare X ha una struttura di varietà complessa, oggi detta superficie di Riemann. Su X si può effettuare il calcolo differenziale e quello integrale come su $\mathbb{C}$: gli integrali abeliani sono integrali di forme differenziali razionali su X.

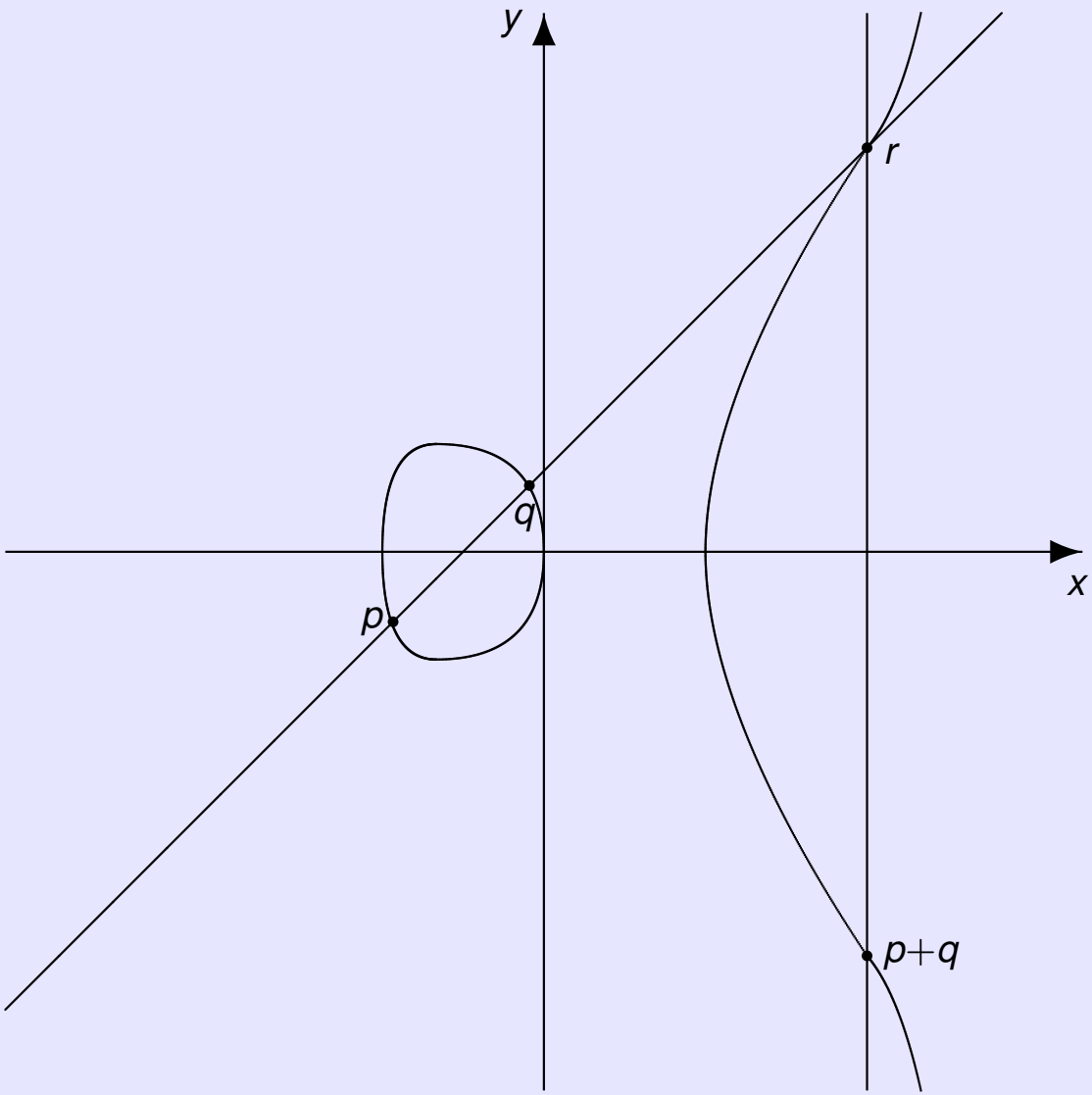
Se $g = 0$, siamo nel caso razionale: gli integrali abeliani sono integrali di funzioni razionali sulla sfera di Riemann $\mathbb{P}^1$, e si esprimono in termini di funzioni elementari. Il caso $g = 1$ dà luogo agli integrali ellittici.

Se $g > 0$, gli integrali abeliani sono funzioni plurivoche su X. Il loro studio richiede l'introduzione di una varietà algebrica $J(X)$ di dimensione g, la varietà jacobiana di X, che ha anche una struttura di gruppo abeliano, cioè è un toro complesso.

Nel caso ellittico $g = 1$ si ha

$$\boxed{X = J(X)}$$

Legge di gruppo sulle curve ellittiche



Dopo Riemann ...

... la teoria venne ulteriormente geometrizzata da A. Clebsch, il quale tradusse in termini algebrico–geometrici i concetti basilari introdotti da Riemann.

Le applicazioni di questa teoria ad altre parti della matematica sono molteplici, ad esempio alla teoria dei numeri. Un ruolo primario giocano le curve ellittiche (cioè quelle di genere $g = 1$) nella famosa dimostrazione di R. Taylor e A. Wiles dell'Ultimo Teorema di Fermat.

Di recente lo studio delle curve ellittiche in teoria dei numeri ha trovato notevolissime applicazioni alla crittografia a chiave pubblica.

Forse Wallis si sarebbe interessato nell'apprendere che la matematica di cui si era occupato nel secolo XVII avrebbe avuto, in un lontano futuro, importanti applicazioni al suo lavoro di crittografo.

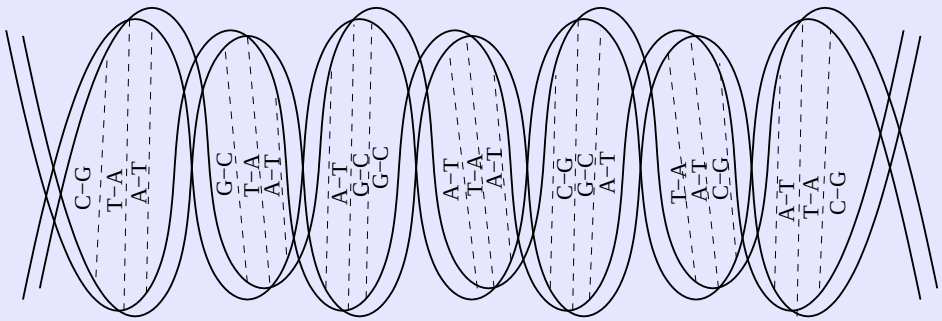
Ora cambiamo scenario e passiamo alla biologia

Il genoma umano è costituito da acido desossiribonucleico (DNA) ...

... che ha la struttura di una doppia elica formata da circa 3 miliardi di coppie (circa 700 megabytes di informazione, quanta ne può essere memorizzata in un CD Rom) di basi complementari:

(Adenina, Timina), (Citosina, Guanina).

La struttura primaria del genoma si può modellare come una sequenza di lettere tratta dall'alfabeto $\Omega\{A, C, G, T\}$.



Sequenze biologiche

Alcuni tratti del genoma codificano degli elementi fondamentali per la vita cioè le proteine, catene di amminoacidi, contenute nel DNA in tratti detti geni.

La successione delle basi in un segmento di DNA, quella degli amminoacidi in una proteina, ecc. sono esempi di sequenze biologiche.

Importanti problemi sono il riconoscimento e l'estrazione dell'informazione codificata in una sequenza biologica.

Esempi

1. Distinguere la parte di un gene che codifica una proteina dalla parte non codificante.
2. Segmentare una porzione di DNA in frazioni con diverse funzioni.
3. Allineare porzioni di DNA appartenenti a specie diverse.
4. Costruire un albero filogenetico.
5. Riconoscere sequenze ultraconservate.

Sequenze ultraconservate

1. Solo l' 1.2% del genoma umano sembra codificare proteine;
2. sono noti circa 500 segmenti più lunghi di 200bp (sequenze ultraconservate), assolutamente conservati in tratti *non codificanti* dei genomi dell'uomo, del topo e del ratto;
3. È noto *almeno* un segmento, detto MEANING OF LIFE SEQUENCE, di lunghezza 42, comune a 10 specie di vertebrati tra cui l'uomo:

   $$TTTAATTGAAAGAAGTTAATTGAATGAAAATGATCAACTAAG$$

4. La probabilità della meaning of life sequence, sotto le ipotesi di un modello semplice di evoluzione con sostituzioni indipendenti ad ogni sito non supera 10^{-50} (Pachter L., Sturmfels B. *The mathematics of Phylogenomics*, 2006).

E. Borel, *Le probabilités et la vie.*

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

Modelli probabilistici e modelli deterministici

Nel corso del novecento si sono affermati, in alternativa ai modelli deterministici classici, modelli probabilistici per la descrizione di fenomeni fisici (meccanica quantistica e meccanica statistica).

Prescindendo da considerazioni di natura filosofica, l'uso dei modelli probabilistici, anche nella biologia, risulta di grande utilità pratica.

Questi modelli, per la natura stessa dei problemi trattati, hanno carattere discreto e un contenuto combinatorico e algebrico che ha richiesto l'uso, e in alcuni casi la creazione, di raffinate tecniche algebrico-geometriche. Si tratta dunque di un quadro completamente nuovo rispetto al passato.

Per discutere i modelli impiegati nella descrizione dell'evoluzione delle sequenze biologiche è necessario far uso di alcuni concetti di calcolo delle probabilità.

Lancio di una moneta

Quando si comincia a parlare di probabilità la prima cosa che viene in mente è il risultato del lancio di una moneta.

Sia p la probabilità di osservare testa ($0 \leq p \leq 1$). La probabilità di osservare croce è quindi $1 - p$. Se $p \neq \frac{1}{2}$ la moneta *è truccata*.

Come si fa a stimare p, e quindi a capire se la moneta è truccata? Si può applicare il principio di massima verosimiglianza, basato sulla seguente:

Assunzione fondamentale

Gli esiti dei lanci sono indipendenti, ovvero l'esito di ogni lancio non dipende da quello degli altri.

Stima 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 osservata, è

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

Stima di massima verosimiglianza

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

Modelli probabilistici per il DNA

In questi modelli il dato osservato è una successione di lettere dell'alfabeto

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

Il modello più semplice è il 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.

A questo modello è associata 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.

Poiché

$$\boxed{p_A + p_C + p_G + p_T = 1}$$

il modello dell'urna dipende da tre parametri essenziali.

Stima dei parametri e adeguatezza del modello

Data una sequenza biologica ci poniamo due problemi.

- Stimare i parametri p_A, p_C, ecc. ad esempio nell'ipotesi che la generazione della sequenza sia descrivibile con il meccanismo dell'urna.
- Valutare l'adeguatezza del modello.

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

Per valutare l'adeguatezza del modello esistono diversi metodi statistici che qui non tratterò, e qualche metodo algebrico, cui accennerò più avanti.

Modelli di A. A. Markov (1856–1922) (I)

Il modello dell'urna è poco utile per descrivere le sequenze di DNA. Un po' più appropriato è un modello in cui l'urna da cui si pesca viene scelta sulla base di un processo di Markov.

Ad esempio possiamo considerare quattro urne, la prima marcata con A, la seconda con G, ecc. e un'ulteriore urna marcata con I (per inizializzazione).

Nell'urna contrassegnata con A, il numero della palline marcate A sarà $n_{A,A}$, quelle marcate con C è $n_{A,C}$ ecc. L' urna I contiene n_A palline marcate con A, etc.

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

- Inizializzazione: estrarre un pallina dall'urna I;
- Iterazione: pescare 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.

Catene di Markov (II)

Questo processo introduce un MECCANISMO DI DIPENDENZA nel processo di scelta. Infatti l'estrazione di una marca dipende da quella estratta al passo precedente. Questo è un esempio di catena di Markov.

Il numero

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

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

Questo modello dipende da 15 parametri (3 per ogni urna).

Gli stessi problemi di stima dei parametri e valutazione di adeguatezza del modello si pongono anche in questo caso e si affrontano come per il modello precedente.

Modelli di Markov a stati nascosti

Nella pratica c'è spesso l'esigenza di considerare, 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 delle osservazioni.

Esempio

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 marcata $\mathcal{T}$ 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.

Stati nascosti indipendenti

Nell'esempio precedente, 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.

Attenzione

Questo semplice modello a stati nascosto NON è adeguato alla descrizione di sequenze biologiche in quanto gli eventi del processo nascosto sono indipendenti.

Catene di Markov nascoste - Esempio

Per introdurre una forma di dipendenza tra gli stati nascosti possiamo modellarli come una catena di Markov.

Abbiamo 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 dell'esempio precedente, 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.

Una sequenza si genera così:

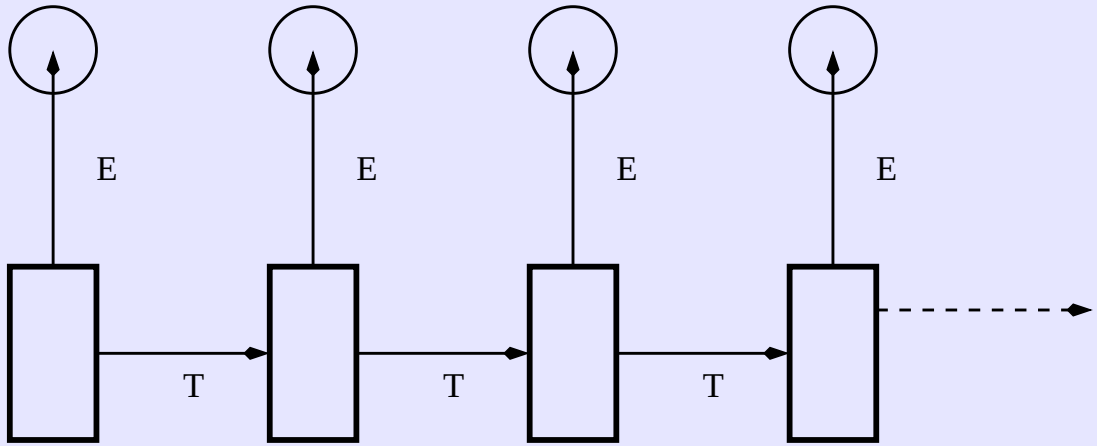
- Inizializzazione: Si lancia la terza moneta per scegliere il piatto da cui cominciare.
- Iterazione: Si pesca una pallina dal piatto *corrente*, si annota la marca, si lancia la moneta sul piatto corrente per scegliere il nuovo piatto, si ripete.

Questo processo, che dipende da 9 parametri indipendenti, è una *catena di Markov nascosta* e costituisce un valido modello per descrivere alcuni aspetti delle sequenze biologiche.

Le catene di Markov nascoste ...

... sono in generale 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.



Catene di Markov nascoste (II)

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.

Applicazioni delle catene di Markov nascoste

Le catene di Markov nascoste costituiscono una classe di processi stocastici che hanno numerose applicazioni pratiche.

Storicamente la prima applicazione importante è quella relativa al riconoscimento vocale:

- 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.

Problema

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.

Riconoscimento vocale

Per ricostruire gli stati nascosti 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 $y_1, \ldots, y_n$ dei fonemi riconosciuti per determinare la successione $x_1, \ldots, x_n$ del discorso da riconoscere si applica il principio di massima verosimiglianza, ovvero si massimizza la probabilità $p(x_1, y_1, \ldots, x_n, y_n)$ che $x_1, \ldots, x_n$ produca $y_1, \ldots, y_n$, che ora calcoleremo.

Probabilità in una catena di Markov nascosta

Successione degli stati nascosti

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

Successione degli stati visibili

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

La probabilità di osservare τ in corrispondenza di σ è il monomio

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

La probabilità di osservare τ qualunque siano gli stati nascosti è il polinomio

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

Modelli grafici

La formula polinomiale

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

per la probabilità nel modello di Markov a stati nascosti fa entrare in gioco l'algebra e, di conseguenza, la geometria algebrica!

In generale, esiste un'ampia classe di modelli probabilistici discreti, i cosiddetti modelli grafici, in cui aspetti algebrici e combinatorici interagiscono in maniera analoga prestandosi ad utili e suggestive interpretazioni geometriche.

Discutiamo alcuni esempi di questi modelli.

Il modello di indipendenza

Consideriamo il grafo

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

Il modello di indipendenza 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

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

Il modello di indipendenza (II)

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

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

Aspetti algebrico–geometrici

Lo spazio delle distribuzioni di probabilità sulle possibili osservazioni è contenuto nell'insieme Δ delle $(n_1 + 1) \dots (n_m + 1)$-ple

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

Il modello di indipendenza seleziona in Δ un sottoinsieme dato da un sistema di equazioni algebriche omogenee e di secondo grado.

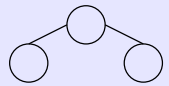
Ad esempio, se $m = 2$ l'insieme di equazioni si compendia in

$$\mathrm{rk}(x_{ij}) = 1$$

Questi sistemi di equazioni definiscono le Varietà di Segre.

Un semplice modello di dipendenza (I)

Nel 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.

Questo modello, in cui 0 è stato nascosto, assegna all'osservazione (b_i, c_j) la probabilità data dal polinomio

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

Un semplice modello di dipendenza (II)

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

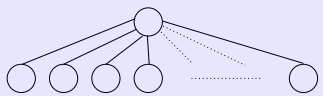
$$(x_{ij}), \qquad i = 0, \ldots, n, \quad j = 0, \ldots, m.$$

Il modello seleziona in Δ un sottoinsieme dato ancora da un sistema di equazioni algebriche omogenee compendiate in

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

Anche questi sistemi di equazioni definiscono varietà algebriche notevoli, ossia Varietà di spazi secanti le varietà di Segre $\mathbb{P}^n \times \mathbb{P}^m$.

Questo modello si può generalizzare considerando il grafo



Dal punto di vista algebrico–geometrico questo corrisponde a considerare varietà di spazi secanti a prodotti di Segre con più fattori.

Varietà algebriche associate a catene di Markov

Nelle catene di Markov nascoste, le espressioni algebriche per le probabilità parametrizzano varietà algebriche razionali.

Molte 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.

Viceversa, recenti tecniche combinatoriche e computazionali in geometria algebrica (basi di Gröbner, geometria torica, geometria tropicale) suggeriscono algoritmi per risolvere i problemi di stima dei parametri e verifica di adeguatezza del modello.

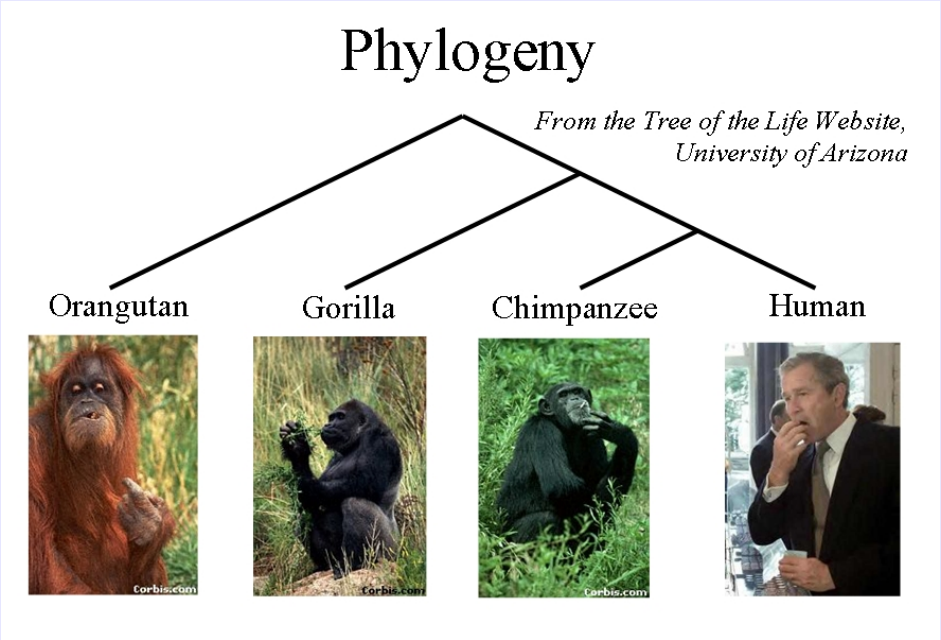
Filogenetica

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

Evoluzionismo

La teoria di DARWIN presuppone che le specie si evolvano da antenati comuni.

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



Darwin

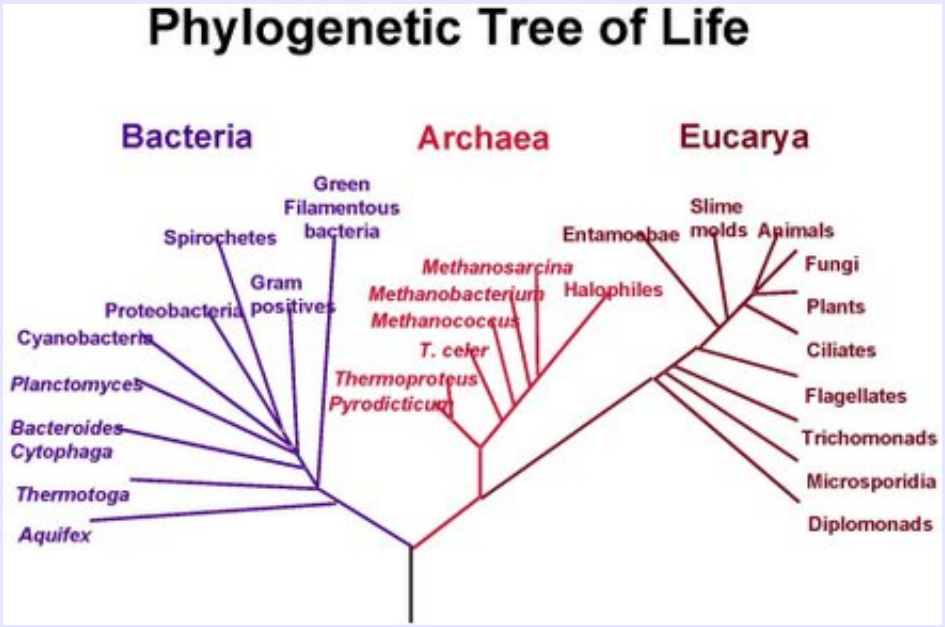
Charles Robert Darwin (1809 - 1882)

Naturalista inglese le cui scoperte scientifiche costituiscono il fondamento della biologia moderna: esse forniscono una spiegazione logica unificata per la diversità delle specie viventi.

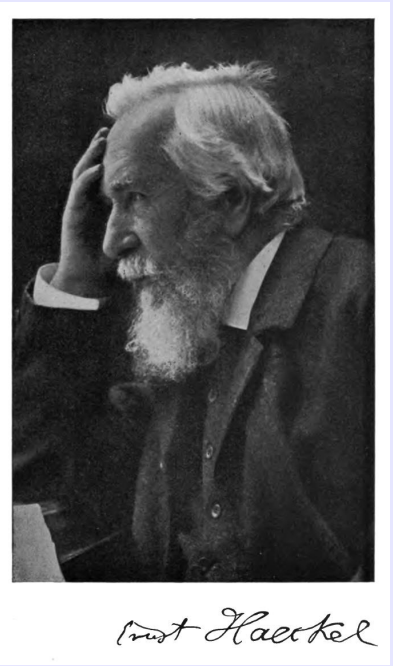
Studiò medicina ad Edimburgo e teologia a Cambridge. Il suo viaggio intorno al mondo, durato cinque anni sulla nave Beagle fornì un ricco materiale di osservazioni su cui fondò le teorie esposte nel libro On the Origin of Species (1859). Esse purtroppo sono ancora oggi oggetto di violente critiche antiscientifiche.

Alberi filogenetici

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



Haeckel

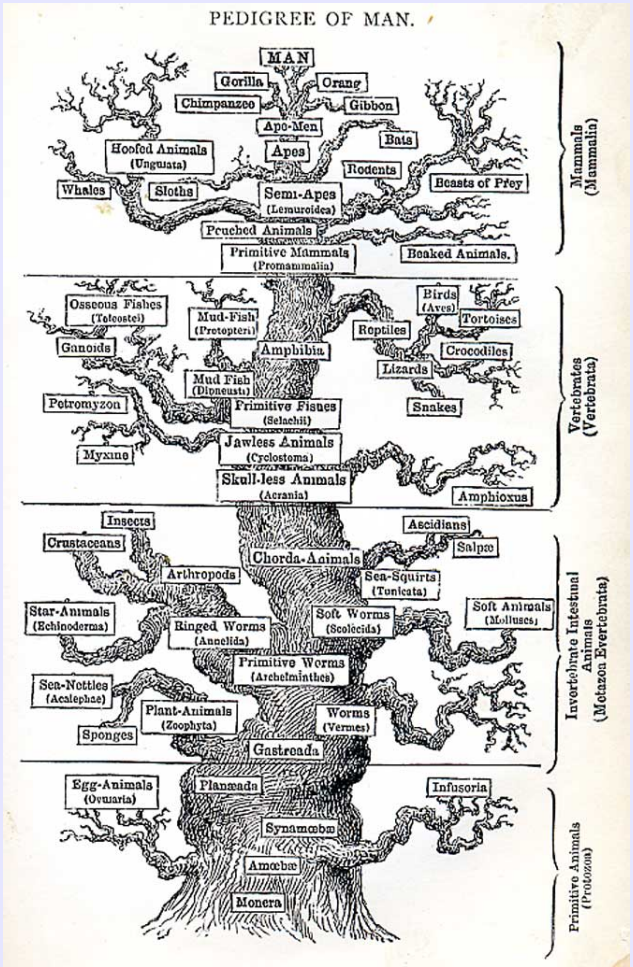


Ernst Haeckel (1834 - 1919)

Biologo, naturalista, filosofo, medico ed artista tedesco. Diede nome a migliaia di specie. Propose un albero filogenetico per tutte le forme di vita.

I termini filogenia ed ecologia furono proposti da lui. Fu un grande promotore delle idee di Darwin in Germania.

l'albero della vita di Haeckel



Applicazioni e complessità della Filogenetica

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.

Alcune applicazioni pratiche

1. capire l'evoluzione di differenti ceppi virali allo scopo di 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.

Complessità del calcolo

Non mi occuperò qui del problema della determinazione di un albero filogenetico plausibile che risponda a una serie di osservazioni su un dato numero di specie. Tale determinazione che si effettua di solito secondo il cosiddetto principio di massima parsimonia è concettualmente semplice ma di grande costo computazionale.

È necessario considerare tutti gli alberi filogenetici aventi un dato numero di foglie etichettate.

Questo numero cresce enormemente al crescere del numero delle etichette.

Teorema

Il numero degli alberi binari con radice con k foglie etichettate, detto numero di Schroeder è

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

Considerazioni sul numero degli alberi filogenetici

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

Il numero stimato degli atomi di idrogeno in tutte le stelle dell'universo è 4×10^{79}.

Non c'è speranza di determinare esattamente le filogenie quando il numero di specie supera la decina.

Esistono invece algoritmi basati su principi diversi per la ricerca di *buone approssimazioni* della soluzione ottimale.

Essi si basano su una struttura matematica più raffinata che riguarda l'intero insieme degli alberi filogenetici con un dato numero di etichette.

Modelli grafici su alberi

Dato un albero filogenetico vogliamo calcolare la probabilità di effettuare una serie di osservazioni alle foglie.

Questo calcolo deve essere effettuato nell'ambito di un modello probabilistico.

Il modello che si usa è una naturale generalizzazione dei modelli di Markov a stati nascosti. Esso è descritto per ogni albero con radice da:

1. $\Omega = \{\omega_1, \ldots, \omega_n\}$, un alfabeto. Un dato è l'assegnazione di un elemento dell'alfabeto ad ogni vertice.
2. $p = (p_1, \ldots, p_n)$, il vettore delle probabilità iniziali: p_i è la probabilità di osservare ω_i nella radice.
3. $T = (t_{ij})$, la matrice di transizione: t_{ij} è la probabilità di passare da ω_i ad ω_j lungo un qualsiasi arco dell'albero.

Questo modello assegna uguale probabilità di transizione tra due stati lungo ogni 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 per la probabilità

La probabilità $p(\omega_W)$ di osservare ω_W nella foglia W si calcola così.

- C'è un unica sequenza $V_1, \ldots, V_k = W$ tale che il vertice V_j è discendente diretto di V_{j-1} per $i = 2, \ldots, k$ e V_1 è la radice.
- La probabilità di osservare ω_{V_i} in V_i per $i = 1, \ldots, k$ è il monomio

$$p(\omega_{V_1}, \ldots, \omega_{V_k}) := p(\omega_{V_1}) \cdot T_{\omega_{V_1}\omega_{V_2}} \cdot T_{\omega_{V_2}\omega_{V_3}} \cdot \cdots \cdot T_{\omega_{V_{k-1}}\omega_{V_k}}$$

- Quindi la probabilità $p(\omega_W)$ è il polinomio

$$p(\omega_W) = \sum_{\omega_{V_1}, \ldots, \omega_{V_k}} p(\omega_{V_1}, \ldots, \omega_{V_{k-1}}, \omega_W)$$

La probabilità di osservare $\omega_{W_1}, \ldots, \omega_{W_h}$ nelle foglie $W_1, \ldots, W_h$ è

$$p(\omega_{W_1}, \ldots, \omega_{W_h}) := p(\omega_{W_1}) \cdot \cdots \cdot p(\omega_{W_h})$$

Questo è un POLINOMIO NEI PARAMETRI p_i E t_{ij}. Ciò consente di usare metodi algebrico-geometrici.

Modelli algebrici e invarianti filogenetici

Dato un modello probabilistico di questo tipo relativo ad un albero con m foglie e ad un alfabeto con n caratteri, si ottengono

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

polinomi che calcolano le probabilità delle osservazioni.

Questi polinomi dipendono dalle variabili p_i, t_{ij} che sono i parametri del modello. Tra questi quelli indipendenti sono

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

Estendendo questi polinomi a valori complessi, possiamo considerare un' applicazione

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

la cui immagine è una varietà algebrica, cioè è definita da un sistema di equazioni polinomiali. I relativi polinomi sono detti invarianti filogenetici.

Alcune di queste varietà, come i prodotti di Segre, sono ben note. Altre non sono mai state studiate in precedenza.

Utilità dell'approccio algebrico

L'approccio algebrico-geometrico può essere estremamente utile.

Uno dei problemi fondamentali della geometria algebrica è quello di studiare l'insieme dei polinomi che si annullano su una data varietà.

In questo contesto ciò equivale a determinare gli invarianti filogenetici del modello.

Utilità degli invarianti filogenetici

Se il modello è adeguato 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.

I metodi dell'algebra computazionale fondati sulle basi di Gröbner rendono fattibile il calcolo di invarianti filogenetici.

In conclusione...

... spero di aver fatto intuire, con gli argomenti trattati (altri ancora se ne sarebbero potuti scegliere!), come, pur essendosi munite nel corso dei loro sviluppi di approcci astratti, concetti sottili e tecniche raffinate, algebra e geometria non hanno tradito la loro origine e natura concrete, riuscendo ad essere adoperate per affrontare problemi di grande utilità.

Anzi, è mia convinzione che sia proprio la natura astratta della matematica a renderla estremamente efficace per abbordare problemi difficili e diversi.

Ma è anche utile sottolineare come spesso sia proprio dai problemi pratici che vengono alla matematica stimoli e suggerimenti per sviluppi di grande rilevanza per il progresso di questa disciplina.