UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II

Facoltà di Scienze
Matematiche Fisiche e Naturali

corso di laurea in fisica
anno accademico 2001/2002
tesi di laurea

MODELLI STOCASTICI DI MUTAZIONI SPONTANEE DEL DNA

Relatore *Prof. Luca Peliti*

Candidato *Osvaldo Zagordi*
matr. *60/572*

Indice

Introduzione

Negli ultimi anni le nuove tecniche di sequenziamento del DNA hanno reso disponibile una enorme mole di dati riguardanti il genoma di diversi organismi. Il primo organismo il cui genoma è stato interamente pubblicato fu l'*Haemophilus Influenzae*, in seguito venne il turno di alcuni eucarioti e il genoma umano è stato sequenziato quasi per intero. Era ovvio che questo avrebbe dato un nuovo impulso alle discipline nell'area della genetica, tra cui quella dell'evoluzione molecolare.

Questa disciplina si propone di indagare come l'evoluzione abbia portato alle relazioni che intercorrono attualmente tra gli organismi mediante l'analisi del patrimonio genetico di questi.

Un'opportunità particolarmente stimolante è rappresentata dai virus. Questi infatti evolvono tanto velocemente da poter essere seguiti *real-time* nell'arco di tempi umanamente ragionevoli[1], cosa che ha portato ad esempio a degli studi sorprendenti sul virus HIV. È stato possibile ad esempio stabilire con certezza in un gruppo di malati quali di questi fossero stati infettati dallo stesso portatore [1]. Per dare un'idea di quanto velocemente evolvano i virus si pensi che il virus presente nel sangue di un singolo paziente ha presentato notevoli cambiamenti nell'arco di soli sette anni di osservazione [2].

Se da 40 anni ormai i progressi in biologia molecolare hanno influenzato fortemente gli studi sull'evoluzione, solo da poco è vero anche il contrario. Negli ultimi anni infatti i biologi molecolari hanno cominciato ad utilizzare un approccio "evolutivo" per studiare più a fondo campi quali la biologia

[1] Mentre per gli altri organismi le scale dei tempi sono nell'ordine dei milioni di anni.

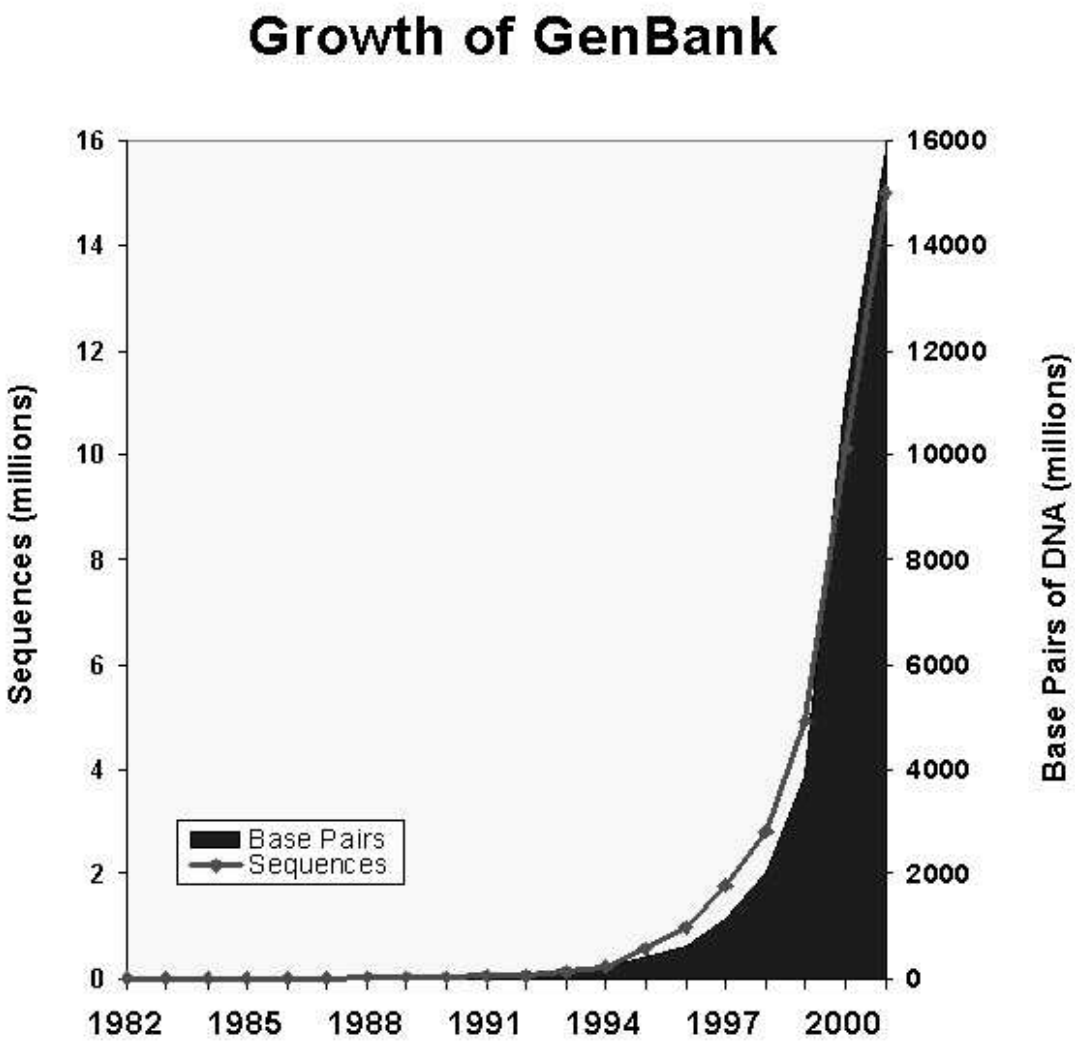


Figura 1: Crescita delle sequenze disponibili su GenBank.

evolutiva e l'immunologia, dove è più marcata l'interazione tra queste due discipline. La biologia molecolare ha effettivamente permesso l'unificazione di molte delle discipline biologiche, in quanto un gran numero di processi vitali sono identici a livello molecolare per tutti gli organismi. Da questo discende che un'altro campo in cui varie branche della biologia si unificano è la biologia evolutiva. Si sa infatti che tutti i viventi discendono da un unico progenitore che ha fatto la sua comparsa sulla terra circa 4 miliardi di anni fa; la mutazione e la selezione naturale sono state poi la causa dell'enorme biodiversità presente attualmente. Le relazioni evolutive fra gli organismi però sono ancora largamente sconosciute, così come i processi che hanno dato origine agli organismi complessi quali i mammiferi. Questi problemi si studiano al giorno d'oggi in maniera quantitativa, ed è per questo che gli strumenti statistici hanno un ruolo essenziale nelle scienze biologiche.

L'interesse verso i modelli di mutazione è sorto verso la fine degli anni '60, quando si sono rese disponibili le prime sequenze geniche. Nel corso degli anni sono stati proposti e risolti modelli a un numero sempre crescente di parametri. Nel contempo la quantità enorme di dati disponibili (vedi figura 1) ha permesso di indagare in maniera sempre più estesa le relazioni di "parentela" fra gli organismi.

I modelli di mutazione hanno ricevuto l'attenzione dei fisici in quanto sono stati assimilati a catene di *Markov*. In questo lavoro ci proponiamo di studiare le conseguenze che ha nei modelli di mutazione la regola di **bilancio dettagliato** che compare nella teoria dei processi stocastici.

Nel primo capitolo vengono descritti alcuni concetti utili a farsi un'idea di cosa significhi la dinamica dei geni in una popolazione e alcune teorie fondamentali per giustificare il tipo di analisi che si fa.

Nel secondo si introduce il modello generale di sostituzione e le grandezze che lo caratterizzano. Viene spiegato come si può definire una distanza tra le sequenze e perchè è utile introdurre la *reversibilità temporale*. Viene infine dimostrata l'equivalenza fra questa proprietà e il *bilancio dettagliato*.

Nel terzo capitolo si formalizza il problema nel modo più generale possibile

e si dimostra che bastano **sei** parametri a descriverlo. Con una procedura analitica si risolvono le equazioni che descrivono l'andamento temporale delle osservabili.

Nel quarto viene spiegato perché abbiamo fatto ricorso a una procedura numerica per la stima dei parametri evolutivi e come viene effettuata. Si spiega inoltre perché non si possono ricavare più di **cinque** parametri indipendenti.

Nel quinto si utilizza la procedura studiata per calcolare le distanze fra delle sequenze reali e si confronta il risultato ottenuto con quello dato dal modello più semplice (*Jukes-Cantor*).

Capitolo 1

Le mutazioni del DNA

Circa mezzo secolo fa la scoperta del DNA (*Watson & Crick* [3]) diede inizio a una nuova era per la biologia. Era inevitabile che questa svolta coinvolgesse drasticamente anche il campo della biologia evolutiva. In che modo la teoria di Darwin si raccorda con le scoperte più recenti della biologia molecolare? Come si può utilizzare quest'ultima per far luce sulla storia evolutiva delle specie viventi? Nel capitolo che segue introdurremo brevemente la cosiddetta *teoria neo-darwinista*. Presenteremo alcuni concetti utili a capire come i geni si distribuiscono in una popolazione e come i fattori aleatori giochino un ruolo essenziale in tali fenomeni. Negli ultimi due paragrafi descriveremo la *teoria neutrale dell'evoluzione molecolare* e una sua importante conseguenza. Entrambe sono di importanza fondamentale per giustificare tutta l'analisi seguente. Alcuni dei concetti necessari per la comprensione di questo capitolo e dei successivi sono riportati in appendice.

1.1 L'evoluzione

Nel 1859 Darwin, nella sua opera più famosa "L'origine delle specie" (per un'edizione *on-line* si veda [4]), propose l'idea che gli organismi viventi evolvono grazie all'azione combinata di due fattori:

la variabilità, un processo caratteristico delle forme di vita che dà origine alla nascita di individui con caratteristiche diverse dai loro progenitori,

la selezione naturale, esercitata dall'ambiente in cui gli organismi vivono e che fa sì che individui con caratteristiche più adatte a un determinato ambiente tendano ad avere una progenie più numerosa, così che tali caratteristiche diventino più frequenti nella specie.

La combinazione di queste due forze fa sì che gli organismi evolvano, ovvero si trasformino nel tempo, tendendo a un sempre migliore adattamento all'ambiente in cui si trovano. Darwin non proponeva quale fosse il meccanismo che dava origine a questa differenziazione intrinseca. Oggi che il ruolo del DNA è stato chiarito, la teoria *neo-darwinista* afferma che alla base di questa variabilità ci sono le **mutazioni** di tale molecola. Per capire meglio il ruolo delle mutazioni introduciamo un concetto su cui si basa tutta la teoria *neo-darwinista*. Tale visione sostiene che l'informazione può passare dal DNA alle proteine (espressione) dal DNA al DNA (replicazione) ma **non** dalle proteine al DNA. Quest'osservazione è conosciuta come **dogma centrale** della biologia molecolare[1]. Solo le informazioni contenute nel DNA possono essere ereditate, e questo è il motivo per cui tutta la storia evolutiva di un organismo risiede in questa molecola. Tale "visione" nega la possibilità di un azione delle proteine (fattori "funzionali") sul materiale genetico (fattori ereditari), in altre parole è il *genotipo* (insieme delle caratteristiche genetiche di un individuo) a influenzarle il *fenotipo* (caratteristiche morfologiche) e non viceversa. In questo modo si precisa anche il ruolo delle mutazioni: poiché avvengono sul DNA si possono trasmettere alla progenie; inoltre, grazie all'influenza che hanno sul fenotipo, determinano la sopravvivenza dell'organismo che le ha subite e quindi la possibilità che siano trasmesse.

[1]Come tutte le buone regole anche questa ha un'eccezione: la *trascrittasi inversa*, il cui ruolo è sinteticamente esposto in figura 1.1.

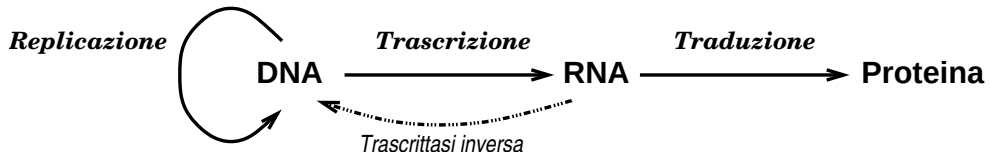


Figura 1.1: Il dogma centrale della biologia molecolare. È mostrato anche l'azione della *trascrittasi inversa*, che "viola" la regola effettuando copie di DNA usando RNA come stampo

1.2 Mutazioni

Le mutazioni, come discusso sopra, rappresentano uno dei fattori essenziali che intervengono nel processo di evoluzione, sarebbe quindi importante osservarle direttamente. Non è possibile però osservare direttamente il fenomeno *biofisico* di mutazione su individui nel loro ambiente naturale a causa della sua estrema rarità. Quello che si osserva è l'effetto di un *background* di mutazioni che hanno agito su una specie per un tempo sufficientemente lungo. Da queste osservazioni si cerca di inferire l'entità del processo che è alla base. Come vedremo in questa analisi si incontrano non pochi problemi.
Definiamo le due grandezze:

tasso di mutazione: la probabilità che il genoma di un individuo, nel trasmettersi alla progenie, presenti delle differenze;

tasso di sostituzione: la probabilità di osservare, in una popolazione che discende da una specie antenata, un certo numero di differenze rispetto a quella ancestrale.

Se quello che si osserva è infatti un tasso di sostituzione, come facciamo a essere sicuri che questo corrisponda al tasso di mutazione? Fattori importanti come la selezione naturale possono aver causato la morte di tutti gli individui che presentavano una particolare mutazione che quindi noi non riusciremmo a vedere. Si perderebbe in questo modo persino la possibilità di fare una trattazione basata su modelli stocastici, che è quella che ci accingiamo a fare. All'estremo opposto, delle mutazioni che non hanno effetto

funzionale si distribuirebbero a caso nella popolazione, risentendo enormemente di fattori aleatori. Nel seguito introdurremo alcune nozioni su come i geni si distribuiscano in una popolazione che evolve e il legame fra mutazioni e sostituzioni.

1.3 Dinamica dei geni nelle popolazioni

1.3.1 Frequenza di diversi alleli

La posizione di un gene su di un cromosoma ovvero all'interno del genoma è detta *locus*; ogni forma alternativa di gene (quando ve ne siano più d'una) in un determinato locus è detta *allele*. In una popolazione possono esserci numerosi alleli, ognuno con frequenza diversa. Se la frequenza di un gene mutante cresce fino a che esso diventa l'unico presente allora si dice che l'allele si è fissato. Per determinare l'evoluzione di una popolazione si studia l'effetto di fattori quali l'ambiente e l'interazione di questo con i diversi alleli, la *deriva genica casuale*(che definiremo in seguito), la selezione naturale (che abbiamo già introdotto), e altri come la *ricombinazione* o *la migrazione*. Due approcci sono possibili:

l'approccio deterministico, in cui si considera la popolazione infinita e l'effetto dell'ambiente costante nel tempo o soggetto a variazioni deterministiche. In questo modo si possono scrivere delle equazioni deterministiche e l'evoluzione della popolazione è determinata univocamente note le condizioni iniziali. È evidente che tale approccio, sebbene più semplice, non è molto aderente alla realtà. Molti sono infatti gli elementi aleatori che entrano in gioco e, come vedremo, la loro entità è tutt'altro che trascurabile.

l'approccio stocastico, in cui data la condizione iniziale non si pretende di conoscere deterministicamente l'evoluzione successiva ma solo la probabilità che questa vada in una certa direzione.

1.3.2 Selezione naturale

La selezione naturale è l'effetto della **differente** capacità di riproduzione di organismi geneticamente diversi causata da vari fattori come mortalità, fertilità e altri caratteri legati alla riproduzione. Per questo motivo la selezione naturale non può avvenire in una popolazione i cui individui abbiano tutti i caratteri identici. Il principale effetto della selezione naturale è il cambiamento in frequenza di alleli tra una generazione e l'altra, ma un tale effetto può essere causato anche da altri fattori come la *deriva genetica casuale*. La selezione naturale ci porta a introdurre il concetto di *fitness*, ovvero la capacità di un individuo di sopravvivere e riprodursi. Dato che la dimensione totale di una popolazione è sempre limitata dalle risorse disponibili conviene riferirsi sempre alla fitness *relativa* come possibilità di successo di un individuo all'interno di una popolazione. Il trattamento più semplice considera la fitness come funzione costante del tempo e tale che tutti i loci contribuiscano indipendentemente al successo di un individuo.

Come vedremo, la maggior parte delle mutazioni che possono emergere hanno un'influenza negativa sulla fitness del portatore. La selezione naturale agirà rimuovendo dalla popolazione questi alleli che a lungo andare andranno a scomparire: si parla allora di **selezione negativa**. Alcune delle mutazioni cui un individuo va incontro possono essere neutre, ovvero la fitness del portatore è identica a quella del miglior allele della popolazione: la sorte di una tale mutazione sarà determinata così da fattori casuali. Raramente una mutazione può avere un effetto positivo sul portatore, agirà su questo una **selezione positiva** che potrebbe portare questo allele a rimpiazzare nel tempo tutti gli altri.

1.3.3 Equilibrio di *Hardy-Weinberg*

Proprio al termine del diciannovesimo secolo, quando Darwin pubblicava la sua famosa opera, Gregor Mendel effettuava i famosi esperimenti che possono considerarsi come la nascita della genetica. Mendel, oltre a concepire

l'esistenza di questi caratteri che oggi chiamiamo geni, riconobbe che vi erano tra questi alcuni che risultavano *dominanti* e altri *recessivi*. Per chiarire facciamo un esempio:

Consideriamo un organismo diploide[2] come l'uomo e prendiamo in considerazione un carattere, ad esempio il colore degli occhi[3]. Per semplicità consideriamo che in una popolazione sono presenti solo due caratteri: occhi neri (A) e occhi blu (a). Dire che il carattere "occhi neri" è dominante su quello "occhi blu" significa che un individuo che eredita entrambi i geni esprimerà quello dominante, ovvero le sue caratteristiche fisiche saranno determinate da questo gene: in questo caso avrà gli occhi neri. Abbiamo così che, sebbene le caratteristiche di ogni generazione siano un riarrangiamento delle caratteristiche di quella precedente, la variabilità genetica è dimezzata ogni volta. Come si mantiene quindi l'ammontare necessario di variabilità in una popolazione? Perché gli alleli recessivi non si perdono completamente? Una risposta a queste domande è data dal teorema di *Hardy-Weinberg* che qui discutiamo. Alla base di questo teorema sta il modo in cui individui genotipicamente diversi contribuiscono mediante la riproduzione all'insieme di genotipi della generazione successiva. Un individuo diploide fornisce alla generazione successiva entrambi i geni con probabilità 50% ciscuna: in altri termini un gene è dominante *nel fenotipo, non nel genotipo*. Questo concetto trova una sua interpretazione grafica nel cosiddetto *quadrato di Punnett*, raffigurato in figura 1.2 e 1.3. Consideriamo **un** individuo *eterozigote* Aa (occhi neri) che si accoppia con **un** individuo *omozigote* AA (occhi neri). Poiché ogni gene di ciascun individuo ha il 50% di probabilità di trasmettersi alla prole, in questa saranno presenti il genotipo Aa con probabilità un mezzo e quello AA con uguale probabilità, sebbene tutti abbiano gli occhi neri. In figura 1.2 sono mostrati i diversi tipi di accoppiamento di geni.

Consideriamo ora una popolazione di individui diploidi composta di maschi

[2]Un organismo si dice diploide quando ogni gene è presente in due copie per ciascuna cellula, di solito una ereditata dal padre e una dalla madre.

[3]In realtà questo è un carattere multigenico, ovvero determinato da diversi geni che contribuiscono in diversi stadi del processo di pigmentazione.

e femmine, quindi che si riproducono in maniera sessuata, e focalizziamo la nostra attenzione su un carattere per il quale siano presenti solo due alleli. Definiamo p la frequenza dell'allele A e q la frequenza dell'allele a. Il discorso appena esposto giustifica le seguenti formule che danno la frequenza degli alleli in funzione di quella dei genotipi:

$$\begin{aligned} p &= f_{\mathsf{AA}} + \frac{1}{2} f_{\mathsf{Aa}} \\ q &= f_{\mathsf{aa}} + \frac{1}{2} f_{\mathsf{Aa}}. \end{aligned} \tag{1.1}$$

dove f_{AA} e f_{aa} sono le frequenze dei genotipi omozigoti AA e aa rispettivamente, e f_{Aa} la frequenza degli eterozigoti. Le formule appena scritte esprimono il fatto che alla diffusione di un allele contribuiscono gli individui che sono omozigoti per questo allele più gli individui eterozigoti al 50%. Valgono ovviamente le condizioni di normalizzazione

$$\begin{aligned} f_{\mathsf{AA}} + f_{\mathsf{Aa}} + f_{\mathsf{aa}} &= 1 \\ p + q &= 1 \end{aligned} \tag{1.2}$$

Il risultato di una generazione in cui tutti gli individui di una popolazione hanno fornito con la stessa probabilità il loro patrimonio genetico si può rappresentare in termini del quadrato di Punnett rappresentato in fig.1.3.

Il quadrato di Punnett in figura 1.3 illustra le seguenti equazioni:

$$\begin{aligned} f_{\mathsf{AA}} &= p^2 \\ f_{\mathsf{aa}} &= q^2 \\ f_{\mathsf{Aa}} &= p \times q + q \times p = 2pq. \end{aligned} \tag{1.3}$$

Osserviamo che per descrivere una popolazione abbiamo una sola frequenza per gli alleli (essendo l'altra imposta dalla normalizzazione) e due frequenze per i genotipi (la terza sarà imposta come sopra). Le equazioni

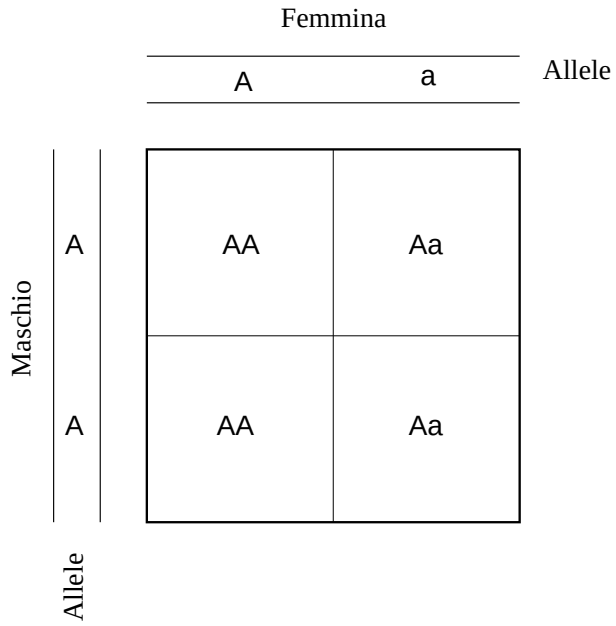


Figura 1.2: Quadrato di Punnett per un singolo incrocio. All'interno dei quadrati sono rappresentate le frequenze di genotipi che emergono nella prole di una coppia in cui vi sia un eterozigote e un omozigote. Ogni quadrato corrisponde a una probabilità del 25% (0.5×0.5).

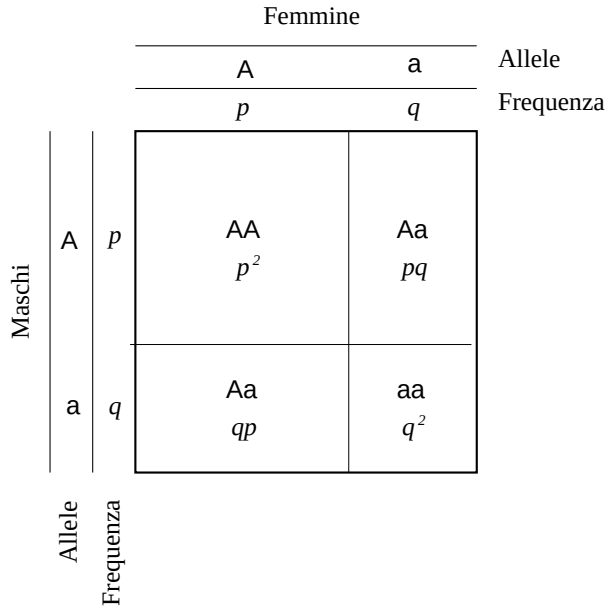


Figura 1.3: Quadrato di Punnett per una popolazione. All'interno dei quadrati è rappresentato il risultato, in termini di frequenza di genotipi, di una generazione in una popolazione in cui ogni individuo ha uguale probabilità di accoppiarsi con un individuo del sesso opposto.

(1.1) si riducono quindi a una sola equazione in due incognite. Ne consegue che pur mantenendo fissa la frequenza degli alleli vi possono essere delle generazioni in cui i genotipi si arrangiano in maniera differente, privilegiando gli omozigoti oppure gli eterozigoti. Effettivamente si possono formulare delle ipotesi che prevedono la costanza nel tempo degli alleli (popolazione infinitamente grande, assenza di selezione etc.), ma che non verificano necessariamente la costanza dei genotipi. Se poi si aggiungono altri vincoli, come l'ipotesi che gli accoppiamenti avvengano con probabilità uniforme fra gli individui di sesso opposto (in inglese *panmictic*), allora la proporzione dei **genotipi** non cambia da una generazione alla successiva e la popolazione viene detta all'**equilibrio di Hardy-Weinberg**. Le frequenze all'equilibrio sono quelle riportate nelle formule (1.3) che garantiscono la presenza di un certo numero di individui per ogni genotipo. Vale la pena di osservare che, partendo da una generazione in cui i genotipi non sono all'equilibrio, basta **una sola generazione** "panmictic" per ottenere i valori predetti dalle formule precedenti.

1.3.4 Deriva genica casuale

Abbiamo già accennato prima che la selezione naturale non è l'unico fattore che può determinare le sorti di un allele, esistono anche fattori casuali come la deriva genica. Tale effetto trova le sue motivazioni nell'inevitabile finitezza di una popolazione. Se un allele è presente in una certa generazione con frequenza p, non è detto che si trasmetta alla generazione successiva con la stessa frequenza a causa degli inevitabili errori di campionamento. Tali errori sono tanto più grandi quanto più è piccola la popolazione. È facile mostrare anche con semplici simulazioni numeriche che per una popolazione di poche decine di individui in cui sono presenti due alleli vi può essere la perdita di uno di questi e il "fissaggio" dell'altro anche nel giro di poche generazioni. Vale la pena di osservare che, sebbene una popolazione di poche decine di individui possa sembrare troppo piccola per tentare di modellizzare un qualunque caso reale di interesse, in realtà non va considerata sempre come popolazione il

numero di tutti gli individui presenti. Le ipotesi che di solito si formulano per trattare la deriva genetica sono infatti alquanto restrittive; non tutti gli individui forniscono il proprio patrimonio genetico con la stessa probabilità, le generazioni si sovrappongono, il numero di maschi non è uguale a quello delle femmine etc. Si utilizza quindi il concetto di *popolazione effettiva*, che è solitamente più piccola di quella reale (a volte molto più piccola). Alcuni stime sostengono che per la specie umana la popolazione effettiva sia un terzo di quella reale.

1.3.5 Sostituzione genica

La sostituzione genica è il processo mediante il quale un allele mutante che emerge a un certo punto della storia evolutiva di un organismo sostituisce interamente quello inizialmente dominante. La comparsa di nuovi mutanti è continua e naturalmente non tutti diventeranno dominanti, bensì saranno a loro volta rimpiazzati da altri mutanti. Si può quindi parlare di *tasso di sostituzione genica* (*rate*) come il numero di sostituzioni per unità di tempo. Altre grandezze utili a descrivere i processi di sostituzione genica sono la *probabilità* e il *tempo di fissaggio*. La prima è la probabilità che un mutante appena emerso sostituisca tutti gli alleli fino a divenire l'unico presente, la seconda quantità indica invece il tempo necessario affinché si verifichi un tale evento. Vedremo nel seguito da cosa dipendono le prime due.

Probabilità di fissaggio

La probabilità di fissaggio di un allele dipende essenzialmente da tre fattori;

- la sua frequenza iniziale q,
- il suo vantaggio selettivo s,
- la grandezza effettiva della popolazione N_{e}.

Il vantaggio selettivo è indicato come la differenza in fitness relativa tra un individuo che possiede questo allele e uno che non lo possiede. Si può di-

mostrare, sebbene sotto ipotesi abbastanza stringenti quali il fatto che il contributo di un singolo gene è mediato su tutti gli altri, che tale probabilità vale (si veda *Kimura* [5])

$$P = \frac{1 - e^{-4N_e sq}}{1 - e^{-4N_e s}}. \tag{1.4}$$

È evidente che per $s \to 0$ si ha $P \to q$, ovvero un allele che non dà alcun vantaggio selettivo ha una probabilità di fissarsi pari alla sua frequenza. Questo si spiega considerando che un mutante neutro si fissa per semplice deriva genetica, che favorisce tutti gli alleli allo stesso modo. Consideriamo un allele che appare in singola copia in una popolazione di N individui diploidi: la sua frequenza iniziale sarà quindi $1/2N$. Se la popolazione effettiva è pari alla popolazione reale l'equazione (1.4) si riduce a

$$P = \frac{1 - e^{-2s}}{1 - e^{-4Ns}}, \tag{1.5}$$

che, se s è positivo (mutazione vantaggiosa) e piccolo, e se $N \gg 1/s$, si riduce a $P = 2s$. In altre parole, mutazioni selettivamente vantaggiose che emergono in popolazioni grandi hanno una probabilità di fissarsi che è circa il doppio del loro vantaggio selettivo.

Rate di sostituzione genica

Fin qui abbiamo distinto fra le **mutazioni**, che rappresentano la comparsa di nuovi mutanti e le **sostituzioni** che indicano invece l'avvenuto fissaggio di un mutante. Ricordiamo ancora una volta che il rate di sostituzione K indica il numero di mutanti che si sono fissati per unità di tempo. Consideriamo di nuovo delle mutazioni neutrali che avvengono a un rate μ (numero di mutazioni per gene per unità di tempo). Il numero di mutanti che emergono a un determinato locus in una popolazione **diploide** di dimensione N è $2N\mu$ per generazione. Come abbiamo visto prima la probabilità che ha una mutazione selettivamente neutra di fissarsi è $P = q = 1/2N$. Moltiplicando questa per il numero di mutazioni che compaiono abbiamo il numero di

mutazioni che si fissano, ovvero di sostituzioni:

$$K_0 = 2N\mu P = \mu. \tag{1.6}$$

Per mutazioni neutre quindi il rate di sostituzioni è pari al rate di mutazioni. Tale risultato è stato notato per la prima volta da Kimura nel 1968 (cfr. [6]). Detto in maniera intuitiva il numero di mutanti che emergono in una popolazione piccola è di conseguenza piccolo, ma la probabilità di fissarsi che hanno è alta. In una popolazione grande invece emergeranno molti mutanti, ma con piccola probabilità di fissarsi, il risultato finale è indipendente dalla taglia della popolazione.
Per una mutazione selettivamente vantaggiosa invece si moltiplica il numero di mutanti che compaiono per la probabilità di fissarsi e si ottiene:

$$K_+ = 4N\mu s. \tag{1.7}$$

In definitiva il rate di sostituzioni dipende da tre fattori:

- la taglia della popolazione N,
- il vantaggio selettivo s,
- il rate di mutazioni μ.

1.4 La teoria neutrale dell'evoluzione molecolare

Abbiamo detto all'inizio del capitolo del ruolo fondamentale della selezione naturale nell'evoluzione degli organismi viventi. In seguito abbiamo descritto altri fenomeni che intervengono nel modificare il patrimonio genetico degli organismi, come la deriva genica casuale o le mutazioni. Ma qual è l'entità dell'apporto di ciascun fattore all'evoluzione?
Secondo i *selezionisti* la selezione è il fattore predominante nel "modellare" gli organismi sulla base della loro fitness, mentre i fattori casuali hanno un'influenza estremamente piccola. A partire dagli anni '60 però, quando si sono

resi disponibili i primi dati sulle strutture delle proteine provenienti da diverse specie, ci si è accorti che l'ammontare di variabilità genica era estremamente più grande di quanto ci si aspettava.

Infatti, se la selezione naturale rimuove dalla popolazione i portatori di alleli inferiori, vi è un limite al numero di mutazioni che possono emergere senza far sì che la popolazione si estingua a causa di queste "morti selettive"? Il genetista *Haldane* stimò che il numero massimo di mutazioni che potevano fissarsi in una popolazione senza che questa si estinguesse era di una ogni 300 generazioni. Kimura, analizzando le differenze fra le proteine di diversi organismi, nel '68 stimò che i mammiferi nella loro storia evolutiva erano andati incontro a circa una sostituzione genica ogni due anni. Troppo per qualunque specie.

Questo lo portò a ipotizzare che la maggior parte dei cambiamenti che si incontrano in un processo evolutivo sono dovuti a mutazioni neutrali o quasi neutrali che si sono fissate (si veda *Kimura* [6]), teoria formulata anche da King e Jukes nel '69 (*King & Jukes* [7]). La *teoria neutrale dell'evoluzione molecolare* sostiene che **a livello molecolare** la maggioranza dei cambiamenti evolutivi e della variabilità delle specie è dovuta a mutazioni selettivamente neutre o quasi neutre che si sono fissate, e non da fenomeni di selezione positiva.

È importante insistere sul fatto che tale teoria non nega il ruolo della selezione naturale nel "modellare" le caratteristiche degli organismi viventi, semplicemente sostiene che questo fattore non è abbastanza forte da dominare su quelli casuali. I *neutralisti* non affermano che tutte le mutazioni sono neutre, né quindi che tutti gli alleli danno luogo alla stessa fitness; affermano piuttosto che la maggioranza delle mutazioni cui un organismo va incontro nella sua storia evolutiva danno un vantaggio selettivo (oppure uno svantaggio) inferiore a $1/2N_{\mathrm{e}}$, ovvero inferiore all'"intensità" dei fattori casuali. Il "cuore" della disputa tra selezionisti e neutralisti sta nell'assegnare le proporzioni tra mutazioni svantaggiose, neutre e vantaggiose. I primi sostengono che poche delle mutazioni non svantaggiose sono neutre e molte invece sono

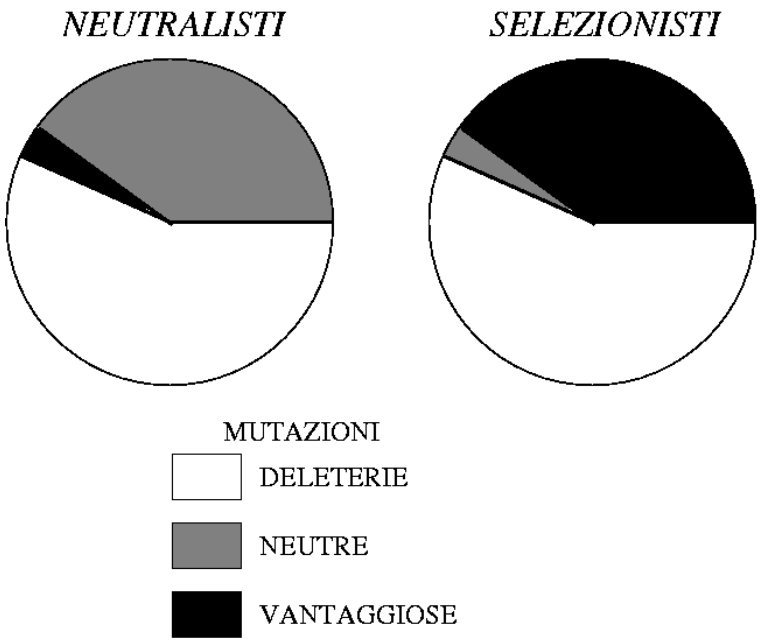


Figura 1.4: Confronto fra il tipo di mutazioni che **compaiono** in un gene secondo la visione neutralista e selezionista. Per entrambe il maggior numero di mutazioni è perso in quanto deleterio, ma i selezionisti sostengono che la maggioranza delle rimanenti è rappresentata da mutazioni selettivamente vantaggiose. Le proporzioni rappresentate sono schematiche e non rappresentano la reale suddivisione per ciascun modello. I neutralisti probabilmente darebbero alle mutazioni vantaggiose una percentuale ancora minore di quella rappresentata.

quelle vantaggiose. Gli altri sostengono che la stragrande maggioranza delle mutazioni non svantaggiose sono selettivamente neutre. I neutralisti danno al *caso* la parte più grande della responsabilità nell'evoluzione, i selezionisti alla *necessità*. Un'idea di queste differenze è illustrata in figura 1.4.

1.5 L'orologio molecolare

Una delle proteine più studiate è l'*emoglobina*. Nella maggior parte dei vertebrati questa consiste di quattro catene polipeptidiche, due codificate dai geni della famiglia delle α-globine e due da quelli delle β-globine. Un'analisi piut-

tosto semplice che si può fare su questa proteina è contare quanti aminoacidi differenti si trovano confrontando catene di organismi differenti. Quello che si scopre è che confrontando specie più lontanamente correlate si trova un numero maggiore di sostituzioni. Se poi si considerano i dati provenienti dalla datazione di reperti fossili si trova qualcosa di straordinario.

Consideriamo ad esempio le differenze in aminoacidi dell'α-globina tra l'uomo e la mucca: questi presentano differenze in 17 aminoacidi su 149, e si sa che le due specie si sono separate circa 80 milioni di anni fa. L'uomo, d'altra parte, presenta 57 differenze con gli alligatori, ovvero 3.4 volte quelle con la mucca. Questo suggerisce che la divergenza dell'uomo dall'alligatore è avvenuta circa 270 milioni di anni or sono (3.4×80). I reperti fossili in effetti confermano questa stima, indicando in 300 milioni di anni il tempo di divergenza. Se si ripete la stessa cosa con altre specie, a parte alcune eccezioni, si trova una grossa correlazione fra la divergenza stimata sulla base dei reperti fossili e quella stimata dai dati delle proteine, ovvero l'α-globina si comporta come un **orologio molecolare**.

Il fatto che il rate di mutazione genica sia pressoché costante nel tempo riveste una doppia importanza. I neutralisti infatti l'hanno interpretato come una conferma della loro teoria e soprattutto l'esistenza di un tale meccanismo giustifica un'analisi dei processi evolutivi basata sui dati molecolari. Confrontare specie diverse mediante l'analisi del loro patrimonio genetico si presta innanzitutto ad analisi quantitative; inoltre ci affranca dalle estreme variazioni della velocità di mutazione a livello morfologico[4]. Riscriviamo le formule (1.6) e (1.7):

$$\begin{aligned} K_0 &= 2N\mu P = \mu \\ K_+ &= 4N\mu s, \end{aligned}$$

che indicano rispettivamente il tasso di sostituzioni neutre e vantaggiose. Il rate di sostituzioni in caso di vantaggio selettivo dipende, come già evidenziato, da tre grandezze. L'esistenza di un orologio molecolare richiederebbe

[4]Si pensi ai cosiddetti "fossili viventi": organismi che hanno lasciato le loro caratteristiche invariate anche da milioni di anni a questa parte.

quindi che tre grandezze, legate ad eventi ecologici, selettivi e casuali, siano costanti oppure si combinino in modo tale da rendere il loro prodotto costante. È chiaro che entrambe queste eventualità sono da ritenersi estremamente improbabili, e che tale quindi risulta l'esistenza di un orologio molecolare in presenza di selezione naturale. L'orologio molecolare è una caratteristica fondamentale del processo evolutivo ed è necessario per condurre un'analisi basata sui dati provenienti dal patrimonio genetico.

Capitolo 2

Modelli di sostituzione

In questo capitolo, dopo aver introdotto il modello generale di sostituzione di nucleotidi nel DNA, presenteremo alcuni dei modelli più semplici e famosi. Discuteremo poi come si possa definire una distanza evolutiva e perché si preferisce ricorrere alla cosiddetta *reversibilità temporale* per stimarla. Dimostreremo infine che tale proprietà è equivalente al *bilancio dettagliato* che si incontra nella teoria dei processi stocastici.

2.1 Il modello generale a 4 ipotesi (*G4H*)

Il modello *G4H* assume che il processo di sostituzione dei nucleotidi sia un processo di Markov in cui :

- i rates di sostituzione non dipendono dal sito;
- sono costanti nel tempo;
- sono identici per le due linee evolutive;
- le frequenze delle basi sono all'equilibrio nella sequenza antenata[1], quindi (per le ipotesi precedenti) si mantengono tali in quelle derivate.

[1]Notare che le frequenze di equilibrio sono quelle dettate dai rates, ovvero quelle ottenute risolvendo le *master equations* con gli stessi rates di mutazione

Nella realtà nessuna di queste ipotesi è soddisfatta esattamente, ma rappresentano comunque un punto di partenza necessario per portare avanti un'analisi. Introduciamo ora alcune grandezze utili nello studio dei modelli di sostituzione:

matrice di divergenza $\mathsf{X}_{[4,4]}(t)$ i cui elementi $x_{ij}(t)$ indicano la mutua probabilità di avere al tempo t il nucleotide i nella prima sequenza e il nucleotide j nella seconda[2], osserviamo che a $t = 0$ tale matrice si riduce a una matrice diagonale i cui elementi sono le frequenze di equilibrio delle basi;

matrice di sostituzione $\mathsf{R}_{[4,4]}$ i cui elementi $r_{ij} = r_{i\leftarrow j}$ rappresentano i tassi di sostituzione del nucleotide j col nucleotide i;

matrice evolutiva $\mathsf{P}_{[4,4]}(t)$ dove gli elementi $p_{ij}(t) = p_{i\leftarrow j}(t)$ sono le probabilità di avere, in un certo sito, il nucleotide i al tempo t dato j a $t = 0$.

Con le grandezze scritte ora possiamo scrivere le equazioni fondamentali del modello *G4H*.

La matrice di divergenza è data da

$$\mathsf{X}(t) = \mathsf{P}(t)\mathsf{X}(t=0)\mathsf{P}^T(t); \quad x_{ij}(t) = \sum_{k=1}^{4} p_{ik}(t) f_k p_{jk}(t), \tag{2.1}$$

dove la matrice evolutiva è la soluzione dell'equazione differenziale

$$\frac{d\mathsf{P}(t)}{dt} = \mathsf{P}(t)\mathsf{R}; \quad \frac{dp_{ij}(t)}{dt} = \sum_{k=1}^{4} p_{ik}(t) r_{kj} \tag{2.2}$$

ovvero

$$\mathsf{P}(t) = \exp\{\mathsf{R}t\}. \tag{2.3}$$

Le quantità sperimentalmente accessibili in maniera diretta sono gli elementi della matrice $\mathsf{X}(t)$. Da questi elementi si può tentare di ricavare i rates

[2]Ciò implica che $\sum_j x_{ij} = f_i$, dove f_i è la frequenza d'equilibrio del nucleotide i, ergo $\sum_{ij} x_{ij} = \sum_i f_i = 1$

e quindi la distanza evolutiva come verrà spiegato in seguito. La matrice di divergenza però non ha 16 parametri indipendenti, infatti la simmetria che si evince dall'equazione (2.1) riduce gli elementi indipendenti a 10, e le 4 equazioni

$$2x_{ii}(t) = 2f_i - \sum_{i(\neq j)} x_{ij}(t) - \sum_{j(\neq i)} x_{ij}(t) \tag{2.4}$$

li riducono ulteriormente a 6. Questo argomento è stato utilizzato da *Rodriguez et al.* [8] per dimostrare che i rates di mutazione non possono essere ricavati in modelli che hanno più di 6 parametri indipendenti.

Osserviamo che

$$x_{ij}(t \to \infty) = f_i f_j. \tag{2.5}$$

Le equazioni (2.4) sono scritte "mediando" sulle righe e sulle colonne perché la matrice di divergenza **osservata** non è in generale simmetrica. Tale asimmetria discende innanzitutto dagli errori di campionamento, inoltre può essere indice del fatto che il fenomeno biologico non segue esattamente le ipotesi formulate per costruire il modello come l'uguaglianza dei tassi lungo le due linee. Se ad esempio vogliamo ricostruire i rates di mutazione a partire da due sequenze osservate, facendo l'ipotesi che l'evoluzione di queste sia avvenuto sullo schema di un dato modello, può accadere di trovarsi in un caso di inapplicabilità di questo (per esempio perché è negativo un radicando o l'argomento di un logaritmo); la minore o maggiore presenza di questi casi è indice della *robustezza* di un modello.

Introduciamo ora alcuni semplici modelli di sostituzione prima di passare ad una formulazione più generale.

2.1.1 Il modello di *Jukes - Cantor* (JC)

Introdotto da *Jukes* e *Cantor* nel 1969 [9], tale modello assume che le mutazioni avvengano con probabilità identica α per ogni coppia di basi diverse.

La matrice dei rates è dunque

$$\mathsf{R} = \begin{pmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{pmatrix}$$

e l'ordine delle basi è A, C, G, T.
Il vettore delle frequenze d'equilibrio è banalmente

$$\vec{f} = (1/4, 1/4, 1/4, 1/4).$$

Mentre la matrice evolutiva $\mathsf{P}(t)$ è data da

$$p_{ij}(t) = \begin{cases} \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} & \text{se } i \neq j; \\ \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} & \text{se } i = j. \end{cases}$$

Ricordiamo che dalla semplice osservazione delle sequenze non è possibile ricavare i rates di mutazione, bensì il prodotto di questi per il tempo; infatti è evidente che moltiplicare tutti i rates per una costante e dividere il tempo per la stessa lascia immutate le quantità osservabili. Indicazioni sulla scala temporale vengono solitamente dalla datazione di reperti fossili.

È particolarmente semplice nel modello JC ricavare le formule che esprimono i rates di mutazione in funzione degli osservabili. Dal momento che in questo modello tutte le basi sono equivalenti l'unica quantità indispensabile è il numero di siti occupati dallo stesso nucleotide, da cui banalmente si ricava il numero di siti in cui i nucleotidi sono diversi.

Immaginiamo (come al solito) che da una sequenza ancestrale discendano due sequenze che noi osserviamo dopo un tempo t. L'elemento di matrice $p_{ij}(t)$ rappresenta la probabilità che il nucleotide j sia mutato in i, quindi la probabilità che un sito sia occupato dallo stesso nucleotide in entrambe le sequenze sarà data da

$$I = p_{AX}^2 + p_{CX}^2 + p_{GX}^2 + p_{TX}^2 \tag{2.6}$$

indipendentemente dal nucleotide X che lo occupava in quella ancestrale. Nel nostro modello tale quantità si ricava facilmente (scegliendo un nucleotide qualunque al posto di X per effettuare il calcolo) e vale

$$I(t) = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}, \tag{2.7}$$

da cui è immediato ricavare la relazione

$$\alpha t = -\frac{1}{8}\ln(\frac{4I-1}{3}). \tag{2.8}$$

2.1.2 Il modello *Kimura a 2 parametri (K2)*

La più semplice generalizzazione del modello precedente è quello proposto da Kimura nel 1980 [10] in cui la matrice di sostituzione ha due parametri indipendenti, uno per le *transizioni* e un altro per le *trasversioni*. Le quattro basi azotate in questo modello non sono più tutte equivalenti, ma si distinguono sulla base della struttura chimica in *purine* (adenina e guanina) e in *pirimidine* (citosina e timina). Le transizioni sono le mutazioni che trasformano una purina in un'altra purina o una pirimidina in un'altra pirimidina (e quindi $A \leftrightarrow G$ e $C \leftrightarrow T$) mentre le trasversioni trasformano una purina in pirimidina o viceversa ($A \leftrightarrow C, A \leftrightarrow T, C \leftrightarrow G, C \leftrightarrow T$). Il modello *K2* assegna un rate α alle transizioni e un diverso rate β alle trasversioni[3], è evidente che uguagliando questi due si torna al caso precedente. La matrice dei rates sarà dunque

$$\mathsf{R} = \begin{pmatrix} 1-\alpha-2\beta & \beta & \alpha & \beta \\ \beta & 1-\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & 1-\alpha-2\beta & \beta \\ \beta & \alpha & \beta & 1-\alpha-2\beta \end{pmatrix}$$

con il solito ordine delle basi.

Anche in questo caso il vettore delle frequenze d'equilibrio è

$$\vec{f} = (1/4, 1/4, 1/4, 1/4).$$

[3]È importante notare che β è il rate per una specifica trasversione, essendo ogni nucleotide soggetto a due trasversioni distinte

La matrice evolutiva sarà caratterizzata da tre quantità,

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t} & \text{se } i = j; \\ \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t} & \text{se } i, j \text{ differiscono per una transizione;} \\ \frac{1}{4} - \frac{1}{4}e^{-4\beta t} & \text{se } i, j \text{ differiscono per una specifica trasversione.} \end{cases}$$

Anche in questo caso si possono ottenere i rates di mutazione in maniera piuttosto semplice dal conteggio dei siti che differiscono per una transizione o per una trasversione.

2.2 Distanza genetica

2.2.1 Multiple hits

Ma a cosa serve in pratica conoscere i rates di mutazione che governano un processo di sostituzione di nucleotidi? Abbiamo già detto che siamo interessati a valutare la distanza evolutiva fra due sequenze osservate, e che le sequenze di DNA si prestano a un'analisi quantitativa. In altre parole è più esatto cercare di inferire la distanza fra gli umani e gli uccelli analizzando il loro genoma piuttosto che confrontando un braccio con un'ala.

Ma come si conduce una tale analisi? A prima vista confrontare due sequenze ci dice solo se un determinato sito è occupato da nucleotidi uguali oppure no. Eppure se osserviamo una differenza fra le sequenze questa può essere dovuta a diversi tipi di sostituzione:

sostituzione singola: è avvenuta una sola sostituzione in una delle due sequenze, da cui si osserva la differenza;

sostituzione multipla: più sostituzioni sulla stessa sequenza, ma si osserva solo un cambiamento (E.G. $A \to T \to C$);

sostituzione coincidente: una sostituzione su ciascuna sequenza, ma verso basi diverse ($A \to T$ su una e $A \to G$ sull'altra).

Inoltre, anche siti che presentano lo stesso nucleotide in entrambe le sequenze potrebbero mascherare delle avvenute sostituzioni:

sostituzione parallela: la stessa sostituzione si verifica in entrambe le sequenze, così non si hanno differenze;

sostituzione convergente: prima si hanno sostituzioni diverse sulle due sequenze, poi su una si ha un'altra sostituzione che *fa convergere* questa verso l'altra sequenza, si hanno ben 3 sostituzioni, ma zero differenze (E.G. $A \to C$ sulla prima e $A \to G \to C$ sulla seconda);

sostituzione all'indietro: una sequenza va incontro a una mutazione, poi a un'altra che la corregge (E.G. $A \to C \to A$).

Tutto questo mostra come il semplice conteggio delle differenze tra le sequenze tenda a sottostimare la distanza evolutiva, tranne per sequenze molto simili tra loro[4]. In effetti la maggior parte delle differenze vengono generate dalle prime mutazioni, come si può evincere dall'andamento esponenziale degli elementi della matrice evolutiva, in seguito poi le sostituzioni multiple allo stesso sito (*multiple hits*) rallentano questo "allontanamento" fino a farlo saturare quando le sequenze diventano totalmente "random".

2.2.2 La regola di Jukes

Ma come si corregge allora la stima della distanza evolutiva? In letteratura si fa ricorso a questo punto alla **reversibilità temporale** (time reversibility), ovvero la proprietà di un modello di soddisfare le seguenti relazioni:

Reversibilità temporale 1 *Date le frequenze d'equilibrio per ciascun nucleotide f_i e le probabilità p_{ij} che il nucleotide j muti in i in un tempo t, si dice che un modello di sostituzione soddisfa la reversibilità temporale se*

$$p_{ij} f_j = p_{ji} f_i \qquad \forall i, j.$$

Nei modelli che soddisfano questa proprietà si fa quindi ricorso alla *regola di Jukes* che qui introduciamo. Immaginiamo come sempre che al tempo $t = 0$

[4] In tal caso effettivamente la probabilità di sostituzioni multiple allo stesso sito è bassa a causa del breve tempo trascorso

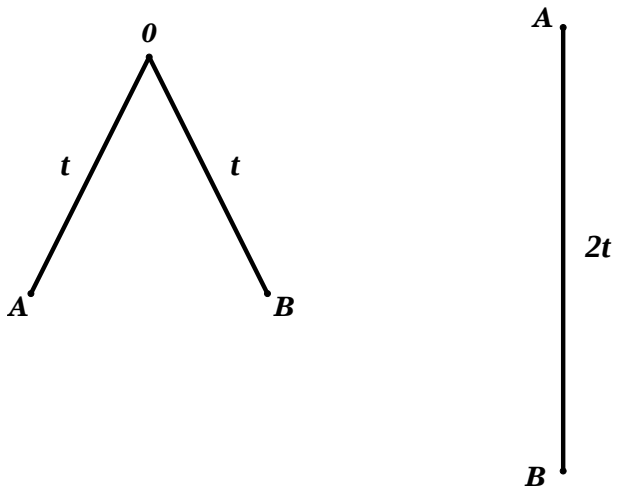


Figura 2.1: Rappresentazione grafica della regola di Jukes

due sequenze si separino da una sequenza progenitrice e che continuino ad allontanarsi accumulando sostituzioni nel tempo seguendo un determinato schema. Il numero di sostituzioni cui queste sequenze sono andate incontro in un tempo t è lo stesso di quelle che una singola sequenza accumula in un tempo $2t$. Immaginiamo che dalla sequenza *0* discendano dopo un tempo t le sequenze A e B. La distanza fra queste due è quella necessaria a percorrere all'indietro nel tempo il ramo da A fino a *0* e poi in avanti fino a B, come mostrato in figura 2.1. La stima così ottenuta è valida in quanto la proprietà (1) ci dice che se nell'andare avanti nel tempo la sequenza A ha mutato i propri nucleotidi Y in X in proporzione di $p_{XY}f_Y$ allora nell'andare indietro il flusso di X in Y, che vale $p_{YX}f_X$, è lo stesso.

A questo punto è semplice valutare la distanza fra le due sequenze, questa sarà infatti

$$d = 2t \sum_i f_i \mu_i = 2t \sum_i f_i \sum_{j(\neq i)} r_{ji} \tag{2.9}$$

dove $\mu_i = \sum_{j(\neq i)} r_{ji}$ è il tasso totale di sostituzione del nucleotide i. La formula (2.9) rappresenta il reale numero di sostituzioni avvenute nel tempo, in quanto i rates di mutazione sono ottenuti con le formule di inversione

presentate in precedenza che correggono già gli effetti di multiple hits. A titolo di esempio scriviamo allora la distanza per il modello di *Jukes-Cantor*

$$d_{JC} = 2t(3\alpha) = -\frac{3}{4}\ln(1 - \frac{4}{3}P), \tag{2.10}$$

dove P è la proporzione di siti differenti
e per quello di *Kimura a 2 parametri*

$$d_{K2} = 2t(\alpha + 2\beta) = \frac{1}{2}\ln\frac{1}{1-Q-2P} + \frac{1}{4}\ln\frac{1}{1-2Q}, \tag{2.11}$$

dove P e Q sono la proporzione di siti che differiscono per una transizione e per una trasversione rispettivamente. È interessante notare che per alcuni autori la reversibilità temporale è una condizione necessaria per la valutazione delle distanze evolutive (*Barry & Hartigan* [11]).

2.3 Bilancio dettagliato

La proprietà di *time reversibility* si incontra con un altro nome nella teoria dei processi stocastici; la proprietà di *bilancio dettagliato*. Abbiamo visto come i modelli di mutazione siano riconducibili a catene di Markov le cui matrici sono le matrici dei rates. Un processo di Markov soddisfa il bilancio dettagliato quando

Bilancio dettagliato 1 *Date le frequenze d'equilibrio per ciascuno stato f_i e i rates di mutazione r_{ij} dello stato j in i, sussistono le uguaglianze*

$$r_{ij}f_j = r_{ji}f_i \qquad \forall i, j.$$

Nel nostro caso, ripetiamo, gli stati sono la presenza di una determinata base in un dato punto della sequenza. Dimostriamo che questa proprietà è equivalente alla reversibiltà temporale.

BILANCIO DETTAGLIATO ⇒ REVERSIBILITÀ TEMPORALE

Abbiamo visto nel paragrafo 2.1 che la matrice evolutiva si ottiene dalla matrice dei rates mediante la formula (2.3)

$$\mathsf{P}(t) = \exp\{\mathsf{R}t\}.$$

Sviluppando in serie si ha

$$\mathsf{P}(t) = \mathbb{I} + \mathsf{R}t + \frac{1}{2}\mathsf{R}^2t^2 + \cdots, \tag{2.12}$$

ovvero

$$p_{ij} = \delta_{ij} + r_{ij}t + \frac{1}{2}r_{ik}r_{kj}t^2 + \cdots \tag{2.13}$$

e ovviamente

$$p_{ji} = \delta_{ji} + r_{ji}t + \frac{1}{2}r_{jk}r_{ki}t^2 + \cdots. \tag{2.14}$$

Scriviamo la (2.13) diversamente:

$$p_{ij} = \delta_{ij} + \sum_{n=1}^{\infty} \frac{(s_{ij})^{(n)}}{n!}t^n, \tag{2.15}$$

dove definiamo

$$\begin{aligned} s_{ij}^{(n)} &= \sum_{k_1k_2\cdots k_{n-1}} r_{i,k_1}r_{k_1,k_2}\cdots r_{k_{n-2},k_{n-1}}r_{k_{n-1},j} \quad \text{per } n \geq 2 \\ s_{ij}^{(n)} &= r_{ij}, \qquad\qquad\qquad\qquad\qquad\qquad \text{per } n = 1. \end{aligned} \tag{2.16}$$

Dimostreremo ora che, se vale il bilancio dettagliato, sussiste l'identità

$$s_{ij}^{(n)}f_j = s_{ji}^{(n)}f_i, \qquad \forall i, j, n. \tag{2.17}$$

Scriviamo per esteso

$$s_{ij}^{(n)}f_j = \sum_{k_1\cdots k_{n-1}} r_{i,k_1}\cdots r_{k_{n-1},j}f_j, \tag{2.18}$$

che diventa, applicando ripetutamente il bilancio dettagliato,

$$\sum_{k_1\cdots k_{n-1}} r_{i,k_1}\cdots r_{j,k_{n-1}} f_{k_{n-1}} = \sum_{k_1\cdots k_{n-1}} r_{i,k_1}\cdots r_{k_{n-1},k_{n-2}} r_{j,k_{n-1}} f_{k_{n-2}} = \cdots \tag{2.19}$$

fino a diventare

$$\cdots = \sum_{k_1\cdots k_{n-1}} r_{k_1,i} r_{k_2,k_1}\cdots r_{j,k_{n-1}} f_i. \tag{2.20}$$

Riordinando i fattori si ottiene

$$\sum_{k_1\cdots k_{n-1}} r_{k_1,i} r_{k_2,k_1}\cdots r_{j,k_{n-1}} f_i = \sum_{k_1\cdots k_{n-1}} r_{j,k_{n-1}} r_{k_{n-1},k_{n-2}} r_{k_{n-2},k_{n-3}}\cdots r_{k_1,i} f_i. \tag{2.21}$$

Poiché gli indici da k_1 a k_{n-1} sono muti, la quantità scritta in (2.21) è uguale a $s^n_{ji} f_i$, per tutti gli $n \geq 2$. Otteniamo così la (2.17) per $n \geq 1$, mentre essa discende direttamente dalla proprietà di bilancio dettagliato per $n = 1$. Poiché inoltre $\delta_{ij} f_j = \delta_{ji} f_i$, abbiamo $p_{ij} f_j = p_{ji} fi$ che è quanto si voleva dimostrare.

BILANCIO DETTAGLIATO ⇐ REVERSIBILITÀ TEMPORALE

Riscriviamo per comodità la formula (2.2):

$$\frac{d\mathsf{P}(t)}{dt} = \mathsf{P}(t)\mathsf{R}; \quad \frac{dp_{ij}(t)}{dt} = \sum_k p_{ik}(t) r_{kj}. \tag{2.22}$$

Calcoliamo la derivata rispetto al tempo di $p_{ij}f_j$; per la proprietà 1 questa sarà uguale alla derivata di $p_{ji}f_i$. Dalla formula (2.22) possiamo scrivere, poiché le frequenze d'equilibrio ovviamente non dipendono dal tempo,

$$\frac{d}{dt}(p_{ij}(t)f_j) = f_j \frac{dp_{ij}(t)}{dt} = \sum_k p_{ik}(t) r_{kj} f_j. \tag{2.23}$$

La (2.22) può scriversi però anche come

$$\frac{dp_{ij}(t)}{dt} = \sum_k r_{ik} p_{kj}(t),$$

dato che P e R commutano come si evince dalla soluzione (2.3). Il secondo membro della (2.23) si può scrivere quindi come

$$\sum_k p_{ik}(t) r_{kj} f_j = \sum_k r_{ik} p_{kj}(t) f_j. \tag{2.24}$$

Usando la reversibiltà temporale l'ultimo membro della (2.24) diventa

$$\sum_k r_{ik} p_{kj}(t) f_j = \sum_k r_{ik} p_{jk}(t) f_k. \tag{2.25}$$

Infine scriviamo

$$\frac{d}{dt}(p_{ji}(t)f_i) = f_i \frac{dp_{ji}(t)}{dt} = \sum_k p_{jk}(t) r_{ki} f_i. \tag{2.26}$$

Sottraendo dalla (2.25) la (2.26), quantità uguali come detto sopra, e mettendo in evidenza $p_{jk}(t)$ abbiamo

$$\sum_k p_{jk}(t)(r_{ik}f_k - r_{ki}f_i) = 0, \tag{2.27}$$

da cui segue il bilancio dettagliato.

Capitolo 3

Violazione del bilancio dettagliato

In questo lavoro ci siamo proposti di studiare un modello più generale possibile che fosse applicabile a sequenze stabili di DNA e che non presupponesse la proprietà di bilancio dettagliato. Descriveremo brevemente cosa sono le sequenze pseudogeniche e perché è utile fare analisi su queste. Vedremo cosa implica guardare DNA stabile e stabiliremo esattamente di quanti e quali parametri avremo bisogno.

3.1 Applicazione a sequenze di pseudogeni

Gli *pseudogeni* sono sequenze di DNA estremamente simili ai normali geni che rimangono però *inespressi*; non vengono cioè trascritti in RNA né tantomeno tradotti in proteine, sono quindi privi di funzionalità. Esistono almeno due meccanismi mediante i quali delle sequenze pseudogeniche possono inserirsi nel genoma:

duplicazione - quando delle modificazioni (come mutazioni o in-del) compaiono durante il processo di duplicazione del DNA e agiscono sulla sequenza in modo tale da non permettere più la produzione della proteina. Dato che questa copia non ha più funzione può essere "disatti-

vata" a livello di trascrizione o di traduzione. Tali copie vengono dette **non processate** oppure **duplicate**.

retrotrasposizione - quando la trascrizione inversa di un segmento di RNA genera una sequenza di DNA che successivamente è inserita nel genoma. Queste copie vengono chiamate **processate**.

Oltre alla mancanza di funzionalità ci interessa la capacità di queste sequenze di andare incontro a mutazioni nel corso dell'evoluzione. Le mutazioni che avverranno su tali sequenze infatti non subiranno l'effetto di bias della selezione naturale e per questo possono essere utilizzate come tracciatori più esatti della storia evolutiva degli organismi.

3.2 Il modello

3.2.1 Sequenze di DNA stabile

Sono ben note ormai le regole di *Watson-Crick* per l'accoppiamento delle basi nel DNA. Queste affermano che la struttura del DNA è tale da consentire tra i due filamenti solo due tipi di accoppiamento tra basi: l'accoppiamento $A = T$ e quello $C \equiv G$[1] Nel corso dei processi di replicazione una mutazione su uno dei filamenti "costringerebbe" il nucleotide corrispondente dell'altro filamento a cambiare verso quello complementare, "stabilizzando" in tal modo la catena. Chiamiamo *stabile* dunque una sequenza di DNA in cui siano rispettate le regole di accoppiamento fra le basi.

3.2.2 Dimensionalità dello spazio dei parametri

Abbiamo detto prima che il modello è rivolto a studiare sequenze di DNA stabile. Quale sarà l'effetto di questo sui parametri? Le mutazioni, ovvero il

[1]La scelta di due simboli diversi rispecchia la diversità dei due legami: il legame fra adenina e timina consiste di due ponti idrogeno, quello fra citosina e guanina di tre. Tale comportamento ha un'origine *quantistica*.

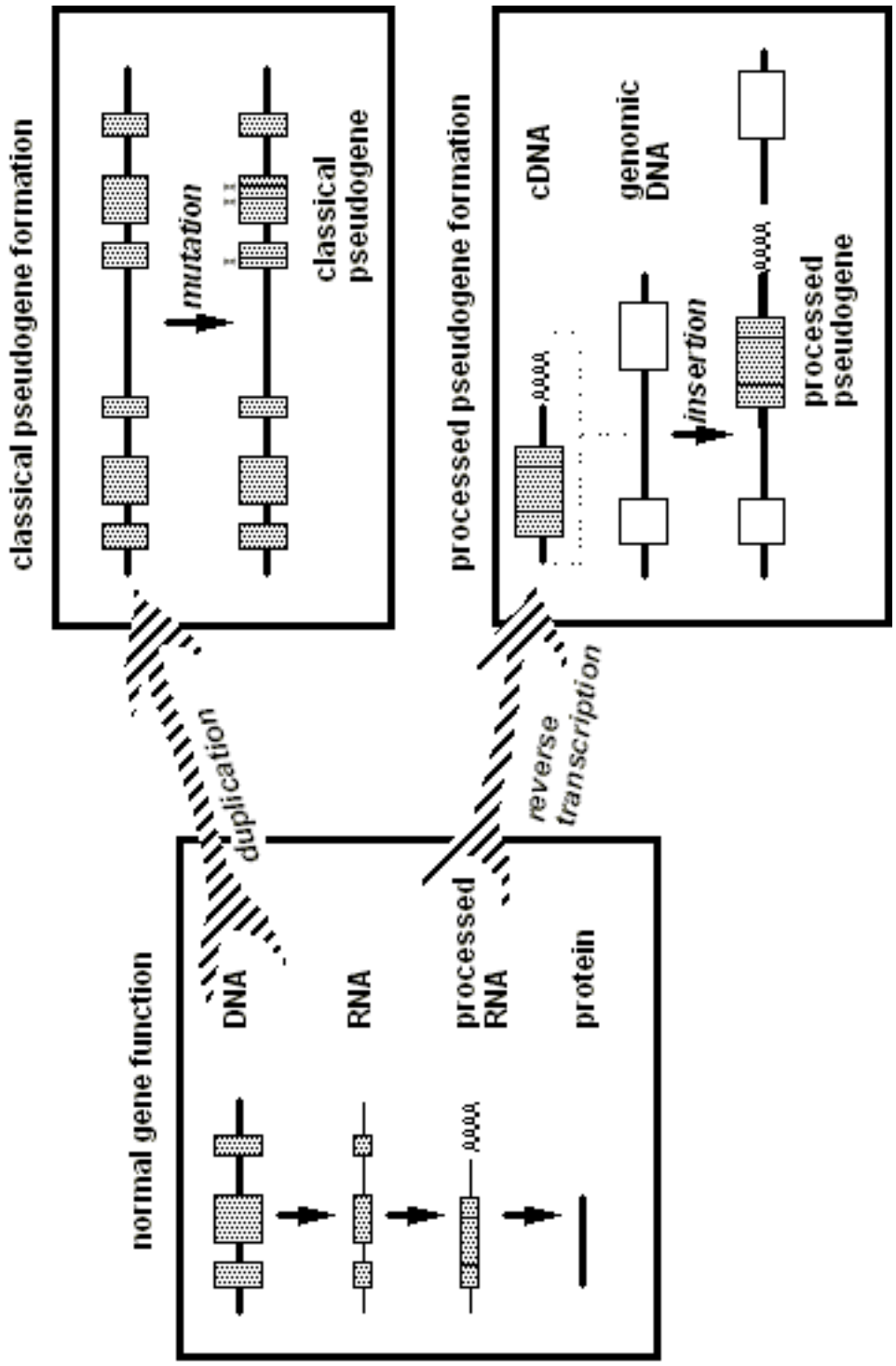


Figura 3.1: Meccanismi mediante i quali gli pseudogeni compaiono nel genoma

fenomeno biofisico di "errore" nei processi di replicazione o di riparo del DNA, che è alla base della nostra analisi, non seguirà necessariamente uno schema diverso da quello che si avrebbe in altre condizioni, ma dal punto di vista del modello da introdurre il guardare un segmento di DNA stabile ha una grande importanza. L'analisi infatti non si può condurre che su sequenze che hanno "nascosto" le cose fino a identificare alcuni rates di mutazione: quelli fra due nucleotidi e i loro complementari. Ad esempio, se in un punto del genoma si verifica la sostituzione $A \to C$, questa è per noi indistinguibile dalla sostituzione nel sito corrispondente del filamento complementare $T \to G$, è chiaro quindi che nel nostro modello i rates di queste due mutazioni dovranno esser posti uguali. In generale $r_{ij} = r_{\overline{ij}}$, o, in dettaglio,

$$\begin{aligned} \mu_1 &\equiv r_{AC} = r_{TG} \\ \mu_2 &\equiv r_{AG} = r_{TC} \\ \mu_3 &\equiv r_{AT} = r_{TA} \\ \mu_4 &\equiv r_{CA} = r_{GT} \\ \mu_5 &\equiv r_{CG} = r_{GC} \\ \mu_6 &\equiv r_{CT} = r_{GA}. \end{aligned} \tag{3.1}$$

Abbiamo così che per la presenza di questa simmetria 6 parametri sono sufficienti a rappresentare tutte le mutazioni possibili, come indicato in figura 3.2.

3.2.3 Frequenze d'equilibrio e regole di Chargaff

Come si trovano le frequenze d'equilibrio dei quattro nucleotidi? Riprendiamo la *master equation*

$$\dot{q_i} = \sum_j (r_{ij} q_j - r_{ji} q_i)$$

che ci fornisce la derivata rispetto al tempo della popolazione dello stato i. Uguagliando a zero $\dot{q_i}$ troviamo le frequenze d'equilibrio. Si giunge a un

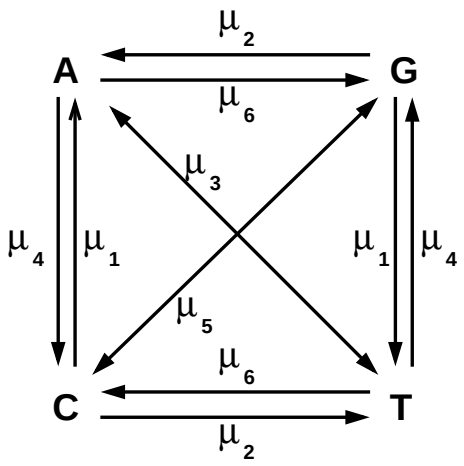


Figura 3.2: Schema di sostituzione applicato

sistema in quattro incognite di quattro equazioni dipendenti. Per risolverlo sostituiamo a una delle equazioni la condizione di normalizzazione

$$\sum_i q_i = 1. \tag{3.2}$$

La soluzione del sistema sarà estremamente semplificata se osserviamo che per ogni nucleotide A è presente un nucleotide T, e per ogni C un nucleotide G. In simboli

$$q_A = q_T \quad e \quad q_C = q_G.$$

Tali regole sono conosciute col nome di *regole di Chargaff* e discendono dal fatto che osserviamo sequenze di DNA stabile. La soluzione sarà quindi

$$\begin{aligned} f_1 &\equiv f_A = f_T = \frac{1}{2}\frac{\mu_1 + \mu_2}{\mu_1 + \mu_2 + \mu_4 + \mu_6} \\ f_2 &\equiv f_C = f_G = \frac{1}{2}\frac{\mu_4 + \mu_6}{\mu_1 + \mu_2 + \mu_4 + \mu_6} \end{aligned} \tag{3.3}$$

Risulta evidente che le frequenze d'equilibrio **non dipendono** da μ_3 né da μ_5, ovvero dai rates di mutazione tra un nucleotide e il suo complementare.

3.2.4 Bilancio dettagliato

Nella derivazione delle frequenze d'equilibrio non abbiamo fatto uso del bilancio dettagliato, perché abbiamo detto che cerchiamo un modello in cui tale proprietà sia violata. Ma cosa cambierebbe applicandolo?
Nella *master equation* si annullerebbero tutti i termini della somma, rendendo più immediata la soluzione. Si avrebbe così

$$\begin{aligned} 0 = r_{AC} f_C - r_{CA} f_A \quad &\Rightarrow \quad \frac{f_A}{f_C} = \frac{\mu_1}{\mu_4} \\ 0 = r_{TG} f_G - r_{GT} f_T \quad &\Rightarrow \quad \frac{f_T}{f_G} = \frac{\mu_1}{\mu_4} \\ &\dots \\ 0 = r_{AG} f_G - r_{GA} f_A \quad &\Rightarrow \quad \frac{f_A}{f_G} = \frac{\mu_2}{\mu_6} \quad , \end{aligned} \tag{3.4}$$

e, poiché abbiamo detto che $f_A = f_T = f_1$, $f_C = f_G = f_2$ e $f_1 + f_2 = \frac{1}{2}$, risulta che

$$\begin{aligned} f_1 \quad &= \frac{1}{2} \frac{\mu_2}{\mu_2 + \mu_6} \\ f_2 \quad &= \frac{1}{2} \frac{\mu_6}{\mu_2 + \mu_6}. \end{aligned} \tag{3.5}$$

Dividendo membro a membro le espressioni scritte sopra per $\frac{f_A}{f_C}$ e di $\frac{f_A}{f_G}$ si giunge alla relazione

$$\mu_1 \mu_6 = \mu_2 \mu_4 \tag{3.6}$$

che si ritrova anche in altro modo.

La teoria dei processi di Markov ci dice infatti che il bilancio dettagliato si può esprimere anche nella seguente forma:

Bilancio dettagliato 2 *Dati tre stati i, j, k, si dice che il bilancio dettagliato è soddisfatto se sussistono le seguenti relazioni fra i rates di mutazione*

$$r_{ik} r_{kj} r_{ji} = r_{ij} r_{jk} r_{ki} \qquad \forall i, j, k.$$

In altri termini se, preso un "triangolo" di stati, il prodotto dei rates fra gli stati incontrati percorrendo il triangolo in senso orario è pari a quello ottenuto percorrendolo in senso antiorario. Applichiamo questo risultato al nostro modello. Osserviamo la figura 3.2, consideriamo la terna di basi ATC. Percorrendola in senso orario (ATCA) troviamo i rates $\mu_3\mu_6\mu_1$, mentre in senso contrario invece (ACTA) $\mu_4\mu_2\mu_3$. Uguagliando queste due quantità ritroviamo la relazione (3.6). La stessa si ritrova considerando una qualunque terna dello schema.

3.3 Risoluzione esatta del modello

Abbiamo trovato che un modello simile al nostro ma leggermente meno generale è quello proposto nel 1981 da *Takahata* e *Kimura* [13], che qui indicheremo con *TK5*. Tale modello differisce dal nostro (che infatti ha un parametro in più) in quanto pone uguali i rates di sostituzione fra le due coppie di nucleotidi complementari, ovvero $A \leftrightarrow T$ e $C \leftrightarrow G$. Nell'articolo originale però non si evidenziano chiaramente le simmetrie del sistema, oltre a fare un'assunzione che introduce *ad hoc* il bilancio dettagliato. Viene inoltre introdotta una procedura molto efficace per trovare gli elementi della matrice di divergenza, che descriviamo di seguito.
Calcoliamo la derivata rispetto al tempo dell'elemento x_{ij} della matrice di divergenza, prendiamo come esempio x_{AC}. Questa sarà

$$\frac{dx_{AC}}{dt} = q_C\dot{q}_A + q_A\dot{q}_C, \tag{3.7}$$

dove q_i al solito esprime la popolazione dello stato i e le derivate rispetto al tempo $\dot{q}_A$ e $\dot{q}_A$ sono da esprimere mediante la *master equation* e valgono

$$\begin{aligned} \dot{q}_A &= (\mu_1 q_C + \mu_2 q_G + \mu_3 q_T) - (\mu_3 + \mu_4 + \mu_6) q_A \\ \dot{q}_C &= (\mu_4 q_A + \mu_5 q_C + \mu_6 q_T) - (\mu_1 + \mu_2 + \mu_5) q_C. \end{aligned} \tag{3.8}$$

Moltiplicando $\dot{q}_A$ per q_C, $\dot{q}_C$ per q_A e sommando abbiamo la derivata dell'elemento di matrice x_{AC} in funzione di x_{AA} e di x_{AC} dove i coefficienti sono combinazioni lineari dei rates di mutazione.

A proposito dell'equazione (3.7) bisogna fare alcune precisazioni. Abbiamo indicato con q_i la popolazione dell'i-esimo stato, nelle ipotesi del modello G4H abbiamo detto che le frequenze di ciascuna base sono all'equilibrio, quindi si dovrebbe avere $\dot{q}_i = 0$. Eppure nel metodo esposto sopra tali quantità sono diverse da zero. Perché?
Dire che le frequenze delle basi sono all'equilibrio vuol dire che **campionando l'intera sequenza** le q_i sono all'equilibrio, ovvero la proporzione di ciascuna rispetto al totale non cambia: abbiamo chiamato f_i tali quantità. Ciò non implica affatto che su ogni sito le basi non subiscano mutazioni, altrimenti non vi sarebbe fenomeno da studiare: ciascuna base ha una probabilità di mutare in un'altra in proporzione che è data dalla *master equation*. Ecco perché possiamo esprimere le derivate degli elementi della matrice di divergenza in funzione delle grandezze q_i e $\dot{q}_i$.

Si arriva a un sistema di 10 equazioni differenziali ordinarie nel tempo accoppiate e dipendenti fra loro. Le equazioni sono 10 a causa della simmetria già evidenziata $x_{ij} = x_{ji}$, sono dipendenti perché le quantità necessarie a descrivere il fenomeno sono in numero inferiore grazie all'intrinseca simmetria del modello. Per come è stato costruito il modello, infatti, sostituendo a un nucleotide i il suo complementare $\overline{\imath}$ non cambiano le equazioni che governano il processo di mutazione, ovvero $\dot{q}_i = \dot{q}_{\overline{\imath}}$. Questo implica dunque che

$$x_{ij} = x_{ji} = x_{\overline{\imath}\overline{\jmath}} = x_{\overline{\jmath}\overline{\imath}}.$$

Abbiamo infine che la matrice di divergenza sarà caratterizzata da un numero inferiore di grandezze, quattro più una frequenza d'equilibrio. Sinteti-

camente:

$$\begin{aligned} P &\equiv x_{AG} = x_{GA} = x_{TC} = x_{CT} \\ R &\equiv x_{AC} = x_{CA} = x_{TG} = x_{GT} \end{aligned} \tag{3.9}$$

$$\begin{aligned} Q_1 &\equiv x_{AT} = x_{TA} \\ Q_2 &\equiv x_{CG} = x_{GC}, \end{aligned}$$

quantità che definiscono dodici elementi della matrice X e quelli lungo la diagonale

$$\begin{aligned} S_1 &\equiv x_{AA} = x_{TT} \\ S_2 &\equiv x_{CC} = x_{GG} \end{aligned}$$

che si ottengono dalla condizione di normalizzazione (2.4) in cui compaiono le frequenze d'equilibrio.

Le quantità P, R, Q_1, Q_2, insieme con una frequenza d'equilibrio[2], sono necessarie e sufficienti a scrivere la matrice di divergenza. Per ottenerle introduciamo le sei quantità

$$\begin{aligned} X_\pm &\equiv 2S_1 \pm 2Q_1 \\ Y_\pm &\equiv 2S_2 \pm 2Q_2 \\ Z_\pm &\equiv 2P \pm 2R \end{aligned} \tag{3.10}$$

che riducono il sistema di 10 equazioni a un sistema di 6 equazioni diagonalizzabile a blocchi. Le soluzioni di questo sono

$$\begin{aligned} X_+ &= \omega[\omega + (1-\omega)e^{\lambda_0 t}] \\ Y_+ &= (1-\omega)(1-\omega+\omega e^{\lambda_0 t}) \\ Z_+ &= 2\omega(1-\omega)(1-e^{\lambda_0 t}) \end{aligned} \tag{3.11}$$

[2]Ricordiamo che l'altra frequenza d'equilibrio si ricava dalla condizione di normalizzazione.

$$
\begin{aligned}
X_- &= \frac{1}{g^2}\{2b[a\omega - b(1-\omega)]e^{\lambda_1 t} + [\zeta\omega + b^2(1-\omega)]e^{(\lambda_1+g)t} + \\
&\quad +[\eta\omega + b^2(1-\omega)]e^{(\lambda_1-g)t}\} \\
Y_- &= \frac{1}{g^2}\{-2a[a\omega - b(1-\omega)]e^{\lambda_1 t} + [a^2\omega + \eta(1-\omega)]e^{\lambda_2 t} + \\
&\quad +[a^2\omega + \zeta(1-\omega)]e^{\lambda_3 t}\} \qquad (3.12) \\
Z_- &= \frac{1}{g^2}\{-2(d-c)[a\omega - b(1-\omega)]e^{\lambda_1 t} + \\
&\quad +[a(d-c+g)\omega - b(d-c-g)(1-\omega)]e^{\lambda_2 t} + \\
&\quad -b(d-c+g)(1-\omega)]e^{\lambda_3 t}\}
\end{aligned}
$$

dove le quantità introdotte valgono

$$
\begin{aligned}
a &\equiv \mu_6 - \mu_4 \\
b &\equiv \mu_2 - \mu_1 \\
c &\equiv 2\mu_3 + \mu_4 + \mu_6 \\
d &\equiv \mu_1 + \mu_2 + 2\mu_5 \\
\omega &\equiv 2f_1 = 2f_A = 2f_T \\
\lambda_0 &\equiv -2(\mu_1 + \mu_2 + \mu_4 + \mu_6) \\
\lambda_1 &\equiv -(\mu_1 + \mu_2 + 2\mu_3 + \mu_4 + \mu_5 + \mu_6) \\
\lambda_2 &\equiv \lambda_1 + g \\
\lambda_3 &\equiv \lambda_1 - g \\
g^2 &\equiv \sqrt{(d-c)^2 + 4ab} \\
\zeta &\equiv \frac{1}{2}(d-c)(d-c+g) + ab \\
\eta &\equiv \frac{1}{2}(d-c)(d-c-g) + ab
\end{aligned}
$$

Vale la pena ricordare ancora una volta che questa soluzione non fa richiesta del bilancio dettagliato.

Capitolo 4

Stima dei parametri evolutivi

Non sempre è possibile ricavare analiticamente i tassi di mutazione dalle osservabili. Nel nostro caso le espressioni sembrano troppo complesse per essere invertite, e quindi si preferisce fare ricorso a metodi numerici. È ovvio però che anche con questi non è possibile stimare parametri indipendenti in numero maggiore delle osservabili. Da questo risulta che è necessario fare un'assunzione che "limita" la generalità del modello.

4.1 Metodi statistici

Takahata e Kimura nel loro articolo [13], al fine di semplificare e invertire le espressioni della matrice di divergenza per ottenere i rates di mutazione, introducono un'ipotesi aggiuntiva sui parametri che di fatto riduce il numero di parametri indipendenti, e che risulta essere proprio l'ipotesi di bilancio dettagliato. La nostra scelta invece rende estremamente complesse le espressioni suddette e abbiamo quindi scelto una strada diversa. La stima dei parametri evolutivi può avvenire con due metodi statistici[1]:

la massima somiglianza (*maximum likelihood*);

i minimi quadrati (*least squares*).

[1] Si veda l'articolo di rassegna di *Zharkikh* [14]

Il primo si basa sulla massimizzazione delle probabilità di ottenere i dati osservati sotto lo schema di mutazione considerato. Il secondo minimizza i quadrati delle differenze tra i dati predetti e quelli osservati. Una soluzione analitica con questi metodi è solitamente molto complessa ed è stata fatta per i modelli più semplici, facendo ritrovare le quantità riportate in precedenza. Il vantaggio di questi metodi è che si prestano a un approccio numerico, ovvero si possono trovare i parametri con degli algoritmi di minimizzazione.

Nella nostra analisi ci siamo avvalsi del metodo dei minimo quadrati nella versione "pesata", che corrisponde a minimizzare il χ^2. Cerchiamo quindi il minimo della quantità

$$f(\mu_1, \dots, \mu_6) = \sum_{ij} \frac{(x_{ij} - \bar{x}_{ij})^2}{\bar{x}_{ij}}, \tag{4.1}$$

dove x_{ij} e $\bar{x}_{ij}$ sono rispettivamente gli elementi della matrice di divergenza teorica e osservata. Le quantità x_{ij} sono calcolate dalle relazioni (3.9) e (3.10). L'equazione (4.1) può essere sviluppata e, sfruttando la condizione di normalizzazione, condotta nella forma più semplice

$$\sum_{ij} \left(\bar{x}_{ij} - 2x_{ij} + \frac{x_{ij}^2}{\bar{x}_{ij}} \right) = \sum_{ij} \frac{x_{ij}^2}{\bar{x}_{ij}} - 1.$$

Notiamo che, poiché $\bar{x}_{ij} \geq 0$, il funzionale si annulla se e solo se $x_{ij} = \bar{x}_{ij}$ per ogni i, j.

4.2 L'algoritmo

Data la complessità delle equazioni che compaiono è stata scelta la strada dell'ottimizzazione numerica, ovvero abbiamo implementato un algoritmo di minimizzazione multidimensionale che desse una stima dei parametri evolutivi. L'algoritmo usato è il cosiddetto *downhill simplex*, ideato da *Nelder & Mead* [15]. Tale procedura richiede il calcolo dei soli valori della funzione e non delle sue derivate, e data la sua semplicità è possibile spiegarla in maniera estremamente naturale. Un simplesso è, in N dimensioni, la figura

geometrica descritta da $N+1$ punti. Per intenderci esso è un segmento in una dimensione, un triangolo in due e un tetraedro in tre. Consideriamo sempre un simplesso non degenere, ovvero che abbia un volume N-dimensionale finito. In questo modo se si prende un vertice del simplesso come origine, i vettori che congiungono questa agli altri vertici generano l'intero spazio di N dimensioni. Riferiamoci per semplicità al caso tridimensionale.

L'algoritmo valuta ad ogni passo la funzione nei 4 vertici del simplesso e confronta i valori ottenuti. A questo punto sono possibili diverse trasformazioni, la più semplice delle quali consiste nel prendere il vertice del simplesso dove la funzione è più grande e *riflettere* questo punto rispetto alla faccia opposta, nel tentativo di raggiungere un punto in cui la funzione è minore. Se lo trova, allora può effettuare un'*espansione* che cerca un punto ancora minore aumentando il volume del tetraedro. Ancora il programma può *contrarre* tutto il simplesso verso il punto più basso.

È necessario fornire al programma un "punto di partenza", ovvero $N+1$ punti in cui il programma posiziona il simplesso la prima volta e da cui parte la procedura di ottimizzazione. Una possibilità è quella di assegnare un'origine $\mathbf{P_0}$ e definire gli altri N punti come

$$\mathbf{P_i} = \mathbf{P_0} + \lambda_i \mathbf{e_i},$$

dove $\mathbf{e_i}$ sono gli N vettori di base. In questo modo oltre ad ottenere un simplesso sicuramente non degenere si può, scegliendo opportunamente le costanti λ_i, fornire all'algoritmo le dimensioni caratteristiche del problema da trattare. L'algoritmo effettua le trasformazioni spostando il simplesso e lo ferma dove crede di aver trovato un minimo. Arresta cioè la procedura quando i valori della funzione calcolati nei vertici differiscono tra loro meno di un valore di tolleranza scelto da noi. Tale valore potrebbe essere anche dell'ordine della precisione della macchina, sebbene non sia sempre consigliabile spingersi così lontano. È sempre utile inoltre far ripartire la procedura di minimizzazione da dove l'algoritmo si è fermato una prima volta. A tal scopo ad esempio si possono reinizializzare N punti, lasciando l'ultimo dove

era stato posizionato dall'algoritmo nel *run* precedente. Questa attenzione permette di assicurarsi che il programma ha trovato un minimo assoluto.

4.3 Numero massimo di parametri

Prima di passare all'analisi di sequenze reali l'algoritmo è stato testato con matrici di divergenza scelte da noi.
Abbiamo quindi scelto dei valori arbitrari per i tassi di mutazione e per il tempo. Con questi, mediante le formule presentate nel capitolo precedente, abbiamo calcolato gli elementi della matrice di divergenza "teorica". Tali valori sono stati utilizzati nella procedura di ottimizzazione descritta in precedenza per ritrovare i valori dei parametri. Quello che è stato notato è che l'algoritmo così utilizzato **non è stabile.** In altre parole inizializzando il simplesso in punti leggermente differenti, il programma diversi trova punti di minimo completamente scorrelati. Questo comportamento si può spiegare col fatto che cerchiamo di stimare 6 grandezze indipendenti usando solo 5 osservabili. Come abbiamo detto in precedenza tante sono le quantità che definiscono univocamente la matrice di divergenza (errori di campionamento a parte). Dobbiamo quindi accontentarci di stimare fino a un massimo di 5 parametri indipendenti, ovvero imporre una relazione fra i sei. Quale relazione scegliere? Abbiamo molte possibilità, per economia di tempo ne sono state provate solo due, che corrispondono a modelli di mutazione studiati in modo approfondito in letteratura:

Modello di *Takahata & Kimura* [13]:

corrisponde considerare uguali i tassi di mutazione fra ogni nucleotide e il suo complementare (quindi $A \leftrightarrow T$ e $C \leftrightarrow G$). Nel nostro modello questo corrisponde a porre $\mu_3 = \mu_5$.

Modello reversibile:

ovvero imporre ai parametri la proprietà di bilancio dettagliato. Nel nostro modello questa equivale alla relazione (3.6), ovvero $\mu_1\mu_6 = \mu_2\mu_4$.

Per confrontare queste due possibilità abbiamo generato osservabili con sei parametri indipendenti e abbiamo minimizzato il funzionale di somiglianza con i cinque che ci derivavano da ciascuna scelta.

Per quanto riguarda la prima scelta bisogna dire che non esiste un'evidenza biologica che possa giustificarla, né tantomeno esiste una ragione analitica. Inoltre abbiamo notato che, cercando di stimare quantità generate con sei parametri, tale vincolo porta l'algoritmo di minimizzazione a trovare sì un minimo, ma che tale minimo non è zero. Non si riesce a trovare quindi un set di parametri che diano esattamente le stesse osservabili.

Per testare la seconda possibilità abbiamo imposto $\mu_6 = \mu_2\mu_4/\mu_1$. Tale scelta presenta due vantaggi: innanzitutto rappresenta una situazione di "simmetria" che ha il vantaggio di essere più semplice da trattare. Cercando poi di "fittare" quantità generate da sei parametri senza bilancio dettagliato con un insieme di cinque parametri che soddisfano tale proprietà, si riesce a trovare il minimo del funzionale di somiglianza e tale minimo vale zero. Un modello reversibile quindi sembra riuscire a ricostruire le osservabili anche se queste sono state prodotte mediante un modello non reversibile.

Capitolo 5

Risultati

La procedura descritta sopra viene applicata a due sequenze reali. Prima di procedere all'analisi le sequenze vanno allineate, per farlo si utilizzano alcuni strumenti presenti in rete. Calcoliamo inoltre la distanza secondo la formula di *Jukes-Cantor* e la confrontiamo con quella valutata da noi.

5.1 L'allineamento delle sequenze

Prima di procedere all'analisi delle sequenze è necessario procedere con il loro allineamento (si veda in appendice). Al giorno d'oggi sono disponibili in rete numerosi strumenti che si rivelano fondamentali per chi si occupa di biologia molecolare. Si hanno così a disposizione banche dati in cui si può esplorare il genoma umano "cliccando" sui cromosomi, programmi che allineano una sequenza data con quelle presenti nei suoi database e infine programmi che mostrano l'allineamento ottimale tra due sequenze fornite da noi[1].
Il problema dell'allineamento di due sequenze consiste, in breve, nell'assegnare delle *penalties*, ovvero un'energia positiva alle *gap* (siti in cui si è verificato un *in-del*) e ai *mismatch* (siti in cui si è verificata una sostituzione), e un'energia negativa ai siti identici. La realizzazione che corrisponde al mini-

[1]Un ottimo punto di partenza è senz'altro http://www.ncbi.nlm.nih.gov/. Un'altro elenco di link utili è all'indirizzo http://matisse.ucsd.edu/itp-bioinfo/links.html

mo dell'energia viene scelta come allineamento e da esso si ricava la matrice di divergenza. L'allineamento di stringhe di caratteri appartenenti a un alfabeto è stato trattato da un punto di vista statistico da *Hwa & Lässig* [16]. Per ottenere dei dati su cui applicare la nostra analisi siamo partiti dalla sequenza di *Rattus Norvegicus* identificata dal codice U33544 (*accession number*), corrispondente al pseudogene del citocromo P450[2]. Questa è stata data al programma Fasta che confronta una sequenza fornita con un enorme numero di banche dati e fornisce i risultati più rilevanti, ovvero quelli che hanno presentato una somiglianza (*score*) più rilevante. Fra queste la nostra scelta è caduta su una sequenze di *Mus Musculus* (*accession number* AF129405) corrispondente al pseudogene Cyp2b10. Nella figura (5.1) è mostrato il risultato dell'allineamento di queste due sequenze ottenuto col programma Blast2. Si vede come il programma abbia individuato due regioni omologhe, la prima di 173 basi che inizia sulla base 451 del pseudogene del *Rattus* e la seconda di 117 basi che inizia alla base 1138, con un grado di omologia molto simile (88% la prima e 87% la seconda). Scegliamo il primo allineamento per costruire la matrice di divergenza in quanto ci permette di avere una statistica lievemente maggiore.

In figura (5.2) viene mostrato l'allineamento fra la stessa sequenza di *Rattus Norvegicus* citata sopra e la sequenza umana **genica** corrispondente al citocromo P450-IIB (*accession number* M29873). Come prima il programma individua due regioni di omologia, la prima lunga 88 basi che comincia in corrispondenza della base 1131 della sequenza del *Rattus* (nell'allineamento precedente cominciava alla base 1138) e la seconda lunga 175 basi che comincia alla base 451, esattamente come sopra. Non c'è dubbio quindi che la regione da analizzare per un confronto con il caso precedente sia quest'ultima.

Nel prossimo paragrafo ricaviamo la matrice di divergenza per entrambi

[2]I citocromi sono pigmenti, presenti in quasi tutti gli organismi viventi, la cui presenza è essenziale per i meccanismi di trasporto di elettroliti nelle cellule. Il citocromo P450 in particolare è una famiglia di circa 60 geni, i meccanismi di espressione dei quali sono particolarmente importanti per l'azione di alcuni farmaci. Si veda anche http://drnelson.utmem.edu/CytochromeP450.html.

questi casi e calcoliamo la distanza fra le sequenze, prima con la formula di *Jukes-Cantor* e poi con il modello sviluppato da noi.

5.2 Matrice di divergenza

Consideriamo il primo allineamento fornito dal programma.
Esso consiste in una sequenza di 173 nucleotidi con 153 identità. Si vede immediatamente, usando la formula (2.10), che la distanza stimata *à la Jukes-Cantor* vale

$$d_{1-JC} = 0.1256 \pm 0.0004. \tag{5.1}$$

L'errore è stato stimato dalla formula

$$\sigma_{JC} = \frac{p(1-p)}{L(1-\frac{4p}{3})^2} \tag{5.2}$$

che è stata ricavata da *Kimura & Ohta* [17]. È stato utilizzato un valore di L pari a $346 = 173 \times 2$ in quanto confrontiamo entrambi i filamenti. Per valutare la distanza nel nostro modello dobbiamo determinare la matrice di divergenza, sempre confrontando sia il filamento mostrato sia quello complementare (non mostrati). Contando le occorrenze di ciascun accoppiamento sulle due sequenze che compaiono in figura e inferendo quelle sulle sequenze complementari abbiamo la seguente matrice di divergenza

$$\mathsf{X} = \frac{1}{346}\begin{pmatrix} 79 & 4 & 6 & 2 \\ 2 & 74 & 2 & 4 \\ 4 & 2 & 74 & 2 \\ 2 & 6 & 4 & 79 \end{pmatrix}.$$

Sono ben visibili gli effetti degli errori di campionamento, per cui la matrice non sempre verifica la simmetria $x_{ij} = x_{ji}$, ma sono altrettanto visibili le simmetrie discusse in precedenza, e cioè $x_{ij} = x_{\bar{i}\bar{j}}$, che non sono affette da errori di campionamento. I valori da inserire nella procedura di ottimizzazione numerica sono ottenuti mediando i contributi dei diversi elementi della matrice.

Abbiamo così i seguenti valori osservati (definiti nelle formule (3.9)):

$$\begin{aligned} P = &\quad 5/346 = &\quad 0.0144509 \\ R = &\quad 3/346 = &\quad 0.0086705 \\ Q_1 = &\quad 2/346 = &\quad 0.005780 \\ Q_2 = &\quad 2/346 = &\quad 0.005780 \\ f_1 = &\quad 91/346 = &\quad 0.263006. \end{aligned}$$

L'algoritmo fornisce i seguenti valori per i parametri di mutazione:

$$\begin{aligned} \mu_1 = &\quad 1.94159 \\ \mu_2 = &\quad 3.32436 \\ \mu_3 = &\quad 1.14463 \\ \mu_4 = &\quad 1.83252 \\ \mu_5 = &\quad 1.22975 \\ \mu_6 = &\quad 3.13761 \quad (= \frac{\mu_2 \mu_4}{\mu_1}). \end{aligned}$$

La distanza si ottiene dalla formula (2.9) e vale:

$$d_1 = 0.1115. \tag{5.3}$$

N.B. Mentre è possibile dare una stima abbastanza esatta per l'errore sulla distanza calcolata dal modello di *Jukes-Cantor*, non è possibile fare altrettanto per la distanza ricavata col modello discusso da noi, dato che il nostro risultato è basato su un metodo numerico. Guardando il comportamento dell'algoritmo possiamo stimare approssimativamente un errore di circa $10^{-3} \sim 10^{-4}$.

Vediamo ora l'allineamento della sequenza del *Rattus Norvegicus* con la sequenza umana. La figura 5.2 mostra il risultato. Consideriamo, come detto prima, la seconda regione di omologia, che mostra identità in 131 siti su 175 (74.8%). La distanza col metodo di *Jukes-Cantor* vale

$$d_{2-JC} = 0.3062 \pm 0.0012. \tag{5.4}$$

La matrice di divergenza risulta essere

$$\mathsf{X} = \frac{1}{350}\begin{pmatrix} 67 & 10 & 7 & 2 \\ 9 & 64 & 2 & 14 \\ 14 & 2 & 64 & 7 \\ 2 & 7 & 10 & 67 \end{pmatrix}.$$

Analogamente a prima abbiamo le seguenti quantità da inserire nell'algoritmo

$$\begin{aligned} P = &\quad 10.5/350 = &\quad 0.03 \\ R = &\quad 9.5/350 = &\quad 0.027143 \\ Q_1 = &\quad 2/350 = &\quad 0.005714 \\ Q_2 = &\quad 2/350 = &\quad 0.005714 \\ f_1 = &\quad 86/350 = &\quad 0.245714. \end{aligned}$$

La minimizzazione numerica fornisce i seguenti valori per i parametri di mutazione:

$$\begin{aligned} \mu_1 = &\quad 7.121386 \\ \mu_2 = &\quad 7.895639 \\ \mu_3 = &\quad 0.412934 \\ \mu_4 = &\quad 7.369823 \\ \mu_5 = &\quad 0.350738 \\ \mu_6 = &\quad 8.171086 \quad (= \frac{\mu_2\mu_4}{\mu_1}). \end{aligned}$$

da cui la distanza è pari a

$$d_2 = 0.3026. \tag{5.5}$$

Blast 2 Sequences results

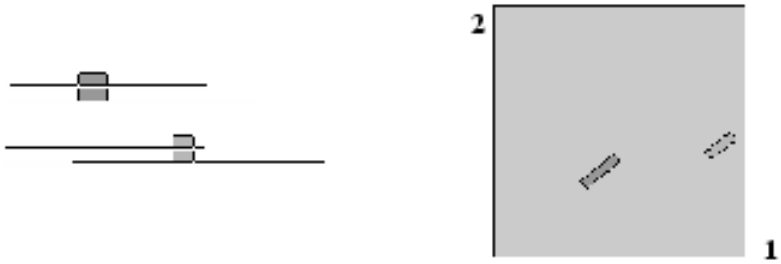
PubMed Entrez BLAST OMIM Taxonomy Structure

BLAST 2 SEQUENCES RESULTS VERSION BLASTN 2.2.3 [Apr-24-2002]

Match: 1 Mismatch: -2 gap open: 5 gap extension: 2
x_dropoff: 30 expect: 10.000000 wordsize: 11 Filter Align

Sequence 1 gi 1173608 Rattus norvegicus cytochrome P450 (CYP2B16P) pseudogene, exons 5 and 6. **Length** 1270 (1 .. 1270)

Sequence 2 gi 4761570 Mus musculus Cyp2b10-like pseudogene, mRNA sequence **Length** 1609 (1 .. 1609)



NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database

NOTE:If protein translation is reversed, please repeat the search with reverse strand of the query sequence

```
Score =  217 bits (113), Expect = 2e-53
Identities = 153/173 (88%)
 Strand = Plus / Plus
```

```
Query: 451 tttgaactcttccctggtgtcctgaagtactttcctggtgcccacagacaaatctccaga 510
           ||||| |||||| ||||  ||||||||||||||||||||||||||||||||||||||| |
Sbjct: 472 tttgagctcttctctggcttcctgaagtactttcctggtgcccacagacaaatctccaaa 531


Query: 511 aacctccatgaaatcctggacttcattggccagagtgtggagaagcacagggccactttg 570
           ||||| || ||| |||| |||| ||||||||| ||||||||||||||||||||||| |||
Sbjct: 532 aacctgcaggaactcctcgactacattggccatagtgtggagaagcacagggccaccttg 591


Query: 571 gacccaaatgctccacgagactttatatatacttaccttctgcacatggagaa 623
           ||||| | || |||||||||||| ||  ||| |||||||||| |||||||||
Sbjct: 592 gaccccagtgttccacgagacttcattgatatttaccttctgcgcatggagaa 644

Score =  160 bits (83), Expect = 5e-36
Identities = 117/134 (87%)
 Strand = Plus / Plus
```

```
Query: 1138 tcttctttgctggcactgagactagcagcaccacactccgctatggcttcctgatcatg
            |||||||||||||||| ||||| |||||||||| ||||||||||||||||||| |||||
Sbjct: 707  tcttctttgctggcaccgagaccagcagcaccacgctccgctatggcttcctgctcatg


Query: 1198 tcaagtaccctcat 1211
            |||||||||| |||
Sbjct: 767  tcaagtaccccat 780

CPU time:     0.07 user secs.       0.04 sys. secs          0.11 total

Lambda      K       H
    1.33    0.621     1.12

Gapped
Lambda      K       H
    1.33    0.621     1.12


Matrix: blastn matrix:1 -2
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 5
Number of Sequences: 0
Number of extensions: 5
Number of successful extensions: 2
Number of sequences better than 10.0: 1
length of query: 1270
length of database: 6,799,009,920
effective HSP length: 25
effective length of query: 1245
effective length of database: 6,799,009,895
effective search space: 8464767319275
effective search space used: 8464767319275
T: 0
A: 30
X1: 6 (11.5 bits)
X2: 15 (28.8 bits)
S1: 12 (23.8 bits)
S2: 21 (41.1 bits)
```

Figura 5.1: Allineamento di una sequenza pseudogenica del *Rattus Norvegicus* con una del *Mus Musculus* ottenuto col programma Blast2, disponibile all'indirizzo http://www.ncbi.nlm.nih.gov/gorf/bl2.html. Si notano due regioni di omologia di diversa lunghezza.

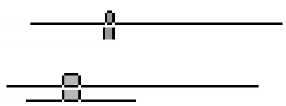
Blast 2 Sequences results

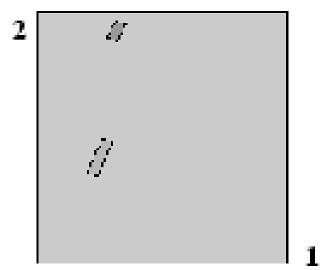
PubMed Entrez BLAST OMIM Taxonomy Structure

BLAST 2 SEQUENCES RESULTS VERSION BLASTN 2.1.2 [Oct-19-2000]

Match: 1 Mismatch: -2 gap open: 5 gap extension: 2
x_dropoff: 50 expect: 10.000000 wordsize: 11 Filter Align

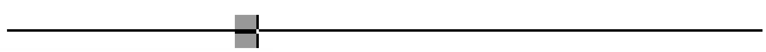
Sequence 1	gi 181293	Human cytochrome P450-IIB (hIIB3) mRNA, complete cds.	**Length** 2907 (1 .. 2907)
Sequence 2	gi 1173608	Rattus norvegicus cytochrome P450 (CYP2B16P) pseudogene, exons 5 and 6.	**Length** 1270 (1 .. 1270)





NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database

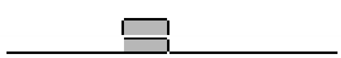
```
Score =  117 bits (61), Expect = 2e-23
Identities = 79/88 (89%)
 Strand = Plus / Plus
```



```
Query: 883  ctctcgctcttctttgctggcactgagaccaccagcaccactctccgctacggcttcctg 942
            ||||| ||||||||||||||||||||||| | ||||||||| |||||||| |||||||||
Sbjct: 1131 ctctctctcttctttgctggcactgagactagcagcaccacactccgctatggcttcctg 1190
CDS    1131 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Query: 943  ctcatgctcaaataccctcatgtcgcag 970
             |||||||||| ||||||||| || |||
Sbjct: 1191 atcatgctcaagtaccctcatatcacag 1218
CDS    1191 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

```
Score = 83.4 bits (43), Expect = 6e-13
Identities = 131/175 (74%)
 Strand = Plus / Plus
```



```
Query: 655 tttgagctcttctctggcttcttgaaatactttcctggggcacacaggcaagtttacaaa 714
           ||||| |||||| ||||  || |||| ||||||||||| || ||||| ||| | | || |
Sbjct: 451 tttgaactcttccctggtgtcctgaagtactttcctggtgcccacagacaaatctccaga 510
```

```
Query: 775 gaccccagcgccccccagggacctcatcgacacctacctgctccacatggaaaaaag 829
           ||||| |  || ||  | ||| | ||  | || ||||| || |||||||| ||||
Sbjct: 571 gacccaaatgctccacgagactttatatatacttaccttctgcacatggagaaag 625
CDS    571 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

CPU time:     0.15 user secs.       0.03 sys. secs          0.18 total

Gapped
Lambda      K       H
    1.33    0.621     1.12

Gapped
Lambda      K       H
    1.33    0.621     1.12


Matrix: blastn matrix:1 -2
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 8
Number of Sequences: 0
Number of extensions: 8
Number of successful extensions: 3
Number of sequences better than 10.0: 1
length of query: 2907
length of database: 2,385,885,539
effective HSP length: 25
effective length of query: 2882
effective length of database: 2,385,885,514
effective search space: 6876122051348
effective search space used: 6876122051348
T: 0
A: 0
X1: 6 (11.5 bits)
X2: 26 (50.0 bits)
S1: 12 (23.8 bits)
S2: 21 (41.1 bits)
```

Figura 5.2: Allineamento di una sequenza pseudogenica del *Rattus Norvegicus* con una sequenza genica umana ottenuto col programma Blast2, disponibile all'indirizzo http://www.ncbi.nlm.nih.gov/gorf/bl2.html. Anche qui si notano due regioni di omologia, corrispondenti a quelle mostrate in figura 5.1.

Conclusioni

Il lavoro di tesi si proponeva di indagare le conseguenze della proprietà di bilancio dettagliato nei modelli di sostituzione del DNA.

Il primo passo è stato dimostrare l'equivalenza tra il bilancio dettagliato e la proprietà dei modelli di sostituzione nota come reversibilità temporale. La discussione sulle regole di appaiamento delle basi ha portato a individuare i parametri necessari per modellizzare in maniera **completa** il fenomeno di mutazione. Abbiamo visto quindi come sei parametri siano sufficienti a descrivere il fenomeno, senza spingersi in complicati modelli a dodici parametri.

Un discorso analogo ha portato a riconoscere che il numero massimo di osservabili indipendenti è cinque, cosa che rende impossibile la stima di tutti i tassi di sostituzione. A questo punto è stato necessario dare un vincolo ai parametri che permettesse alla procedura di minimizzazione di ottenere un risultato. La scelta di un modello reversibile è premiata dal fatto che si riescono a ricostruire le osservabili generate anche con un modello non reversibile e dal fatto che la distanza calcolata con un modello a un parametro viene sì corretta dal modello sviluppato, ma non stravolta.

In realtà la scelta di un modello reversibile non è l'unica che permette di ricostruire esattamente le osservabili. Definiamo una misura della rottura del bilancio dettagliato

$$\Delta = \frac{\mu_1 \mu_6}{\mu_2 \mu_4}.$$

Quando $\Delta = 1$ vale il bilancio dettagliato, ovvero il modello è reversibile. Si osserva che anche modelli con Δ diverso da uno riescono a ricostruire le

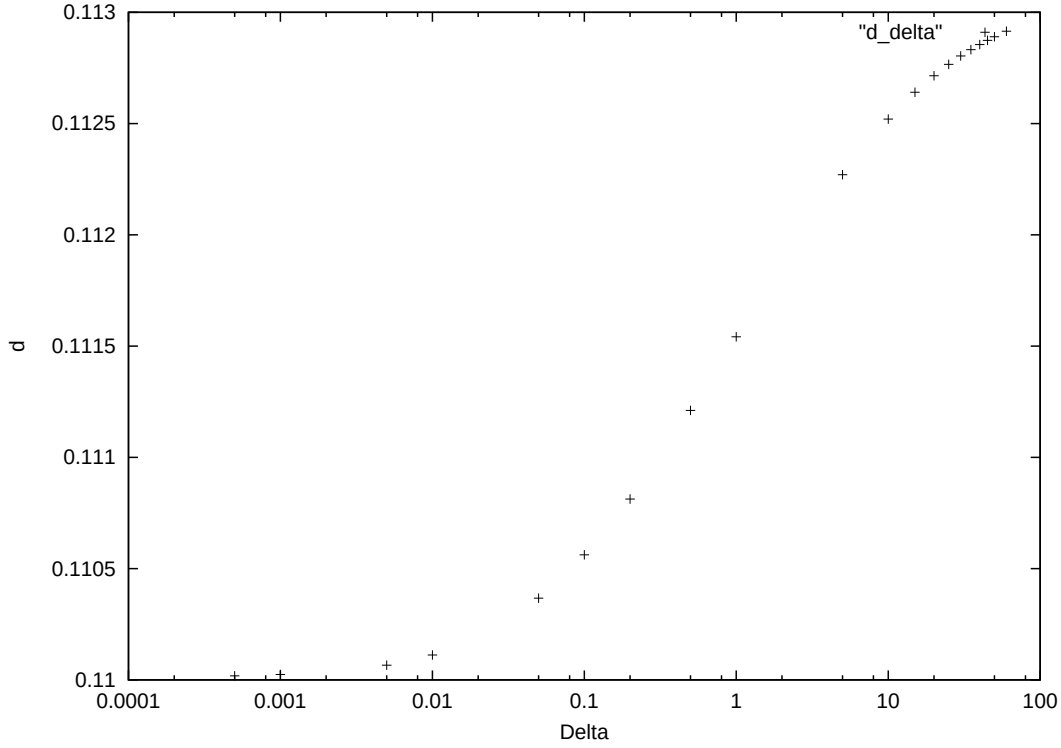


Figura 5.3: Distanza fra le due sequenze pseudogeniche stimata in funzione del parametro di violazione del bilancio dettagliato. Per $\Delta = 1$ il bilancio dettagliato è soddisfatto, ovvero il modello è reversibile. Sono riportati solo risultati in cui il modello riesce a ricostruire le osservabili.

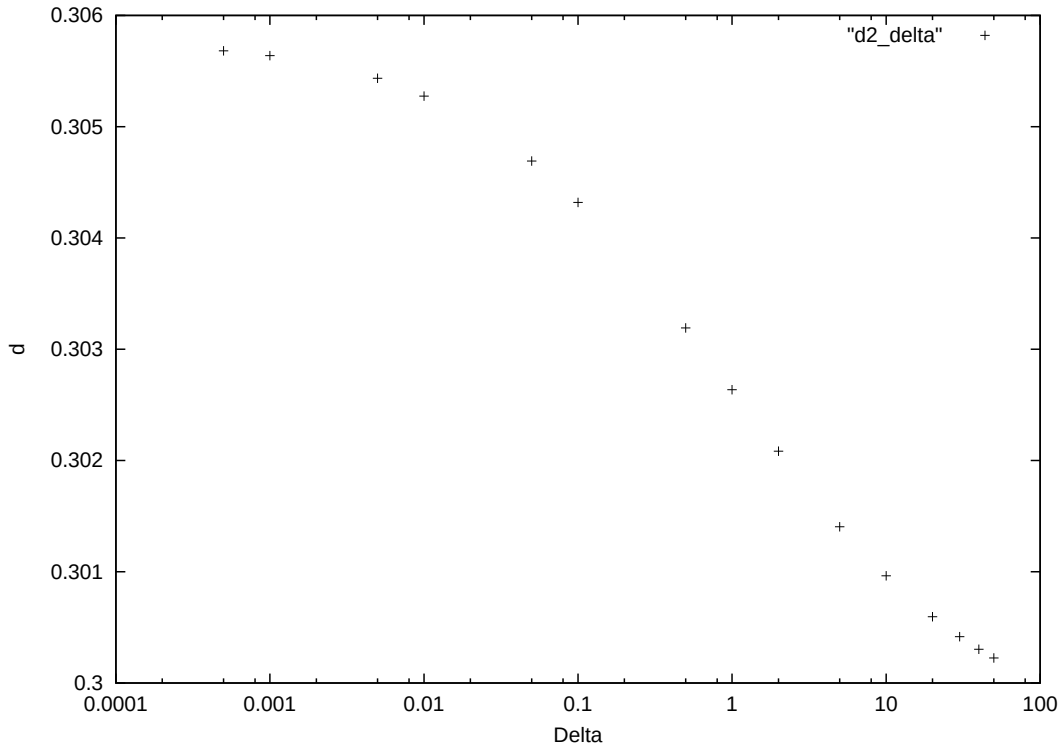


Figura 5.4: Distanza fra la sequenza pseudogenica del *Rattus* e quella genica umana stimata in funzione del parametro di violazione del bilancio dettagliato.

osservabili con esattezza (il funzionale si annulla), ma cambia ovviamente la distanza calcolata. In figura 5.3 e 5.4 riportiamo la distanza tra le coppie di sequenze già esaminate in precedenza ricostruita secondo modelli con diversi valori di Δ. Si riportano solo i dati dell'intervallo in cui il modello riesce a ricostruire le osservabili.

Come si può pensare allora di "scegliere" un modello? Pensiamo a due sequenze progenitrici che si dividono ciascuna in due sequenze discendenti e ammettiamo che lo schema di mutazione cui sono soggette sia lo stesso. Anche il tempo di divergenza sarà lo stesso, quindi si hanno a disposizione quattro sequenze che possiamo confrontare due a due. Abbiamo così a disposizione due matrici di divergenza che dovrebbero dare la stessa distanza (a meno degli errori).

Ancora si può effettuare un'altra prova. Supponiamo che la sequenza antenata fosse l'unione delle due sequenze considerate prima e che valgano le ipotesi precedenti. In ogni discendente avremmo una sequenza più lunga da confrontare con quella dell'altro. Anche da queste possiamo calcolare la distanza e confrontarla con quelle ricavate prima.

In questo modo è come se aumentassimo il numero di osservabili senza aumentare quello dei parametri.

Analisi di questo tipo però non sono prive di rischio. Non tutti i geni di due organismi che dividono un antenato comune noto discendono dal gene corrispondente di questo antenato. Il tempo di divergenza potrebbe quindi essere diverso, e tale sarebbe la distanza tra le sequenze.

È importante notare che un'analisi del genere è da compiersi su sequenze non funzionali, quali quelle pseudogeniche. I vincoli che agiscono sulle sequenze codificanti sono tali da rendere estremamente difficile giustificare l'assunzione secondo cui lo schema di mutazione è lo stesso.

In generale quando si confrontano sequenze funzionali si distingue la distanza calcolata sul primo, secondo o terzo nucleotide di ogni codone. Poiché la *degenerazione del codice genetico* (vedi appendice) è maggiore sul terzo nucleotide si vede che anche la distanza stimata su questo è maggiore (la

sequenza ha subito il *bias* della selezione naturale, per cui i portatori di una mutazione negativa non sono sopravvissuti e non hanno potuto trasmetterla).

Il fatto che nell'analisi riportata le sequenze più distanti siano le seconda dipende dal tempo di divergenza che è estremamente più piccolo per il *Rattus* e il *Mus* che per il *Rattus* e l'uomo.

Appendice A

Concetti di base

A.1 Geni

A.1.1 Il DNA

Le molecole di acido **deossiribonucleico** (DNA) sono per tutti gli organismi viventi (ad eccezione di alcuni virus) i portatori dell'informazione ereditaria. Queste molecole consistono in due filamenti complementari attaccati l'uno all'altro e avvolti a formare un'elica destrogira. Ciascun filamento è un polinucleotide lineare di lunghezza variabile fatto di quattro nucleotidi, anche detti basi azotate: adenina (A), citosina (C), guanina (G) e timina (T). I due filamenti sono tenuti insieme grazie all'accoppiamento tra le basi: esistono solo due tipi di accoppiamento: l'accoppiamento $A = T$ (che consiste in due legami idrogeno e perciò è detto legame debole, e quello $C \equiv G$ (che è fatto da tre legami idrogeno e quindi è detto legame forte). Tali regole di accoppiamento sono dette *regole di Watson-Crick*. La molecola di DNA è polare, questo permette di definire un verso. Ci si riferisce alle due direzioni sfruttando la numerazione degli atomi di carbonio delle basi: il verso in cui viene *trascritta* (ovvero trasformata in RNA) è detto $5' - 3'$. È estremamente importante specificare un verso sulla catena in quanto l'informazione genetica risiede, come vedremo, nella sequenza ordinata di nucleotidi di cui è fatta

la molecola.

A.1.2 Definizione di gene

Nella definizione tradizionale un gene è un segmento di DNA che "codifica" per una catena polipeptidica. Negli ultimi anni però la definizione di gene è cambiata e include adesso una qualunque sequenza di DNA o RNA che effettua una specifica funzione. Tali funzioni però non richiedono necessariamente che la sequenza sia tradotta (in proteina) né tantomeno trascritta (in RNA). Si distinguono così tre tipi di geni:

geni che codificano per proteine ovvero geni prima trascritti in RNA e poi tradotti in proteine,

geni che specificano molecole di RNA ovvero geni che vengono trascritti ma non tradotti,

geni regolatori che non sono neanche trascritti.

A.2 Aminoacidi, proteine, codice genetico

Gli aminoacidi sono le strutture chimiche fondamentali di cui sono costituiti gli organismi. Tutte le proteine di cui sono composti gli esseri viventi, dall'uomo ai batteri, sono fatte da 20 aminoacidi sistemati in una o più catene dette catene polipeptidiche. Ogni aminoacido è composto da un atomo di carbonio (detto C_α) cui sono legati un'ammina (NH_2), un gruppo carbossile ($COOH$) e un **gruppo R** che distingue un aminoacido dall'altro. Il *legame peptidico* che si instaura tra il gruppo NH_2 di un aminoacido e il gruppo $COOH$ di un aminoacido adiacente rende possibile la formazione di lunghe catene: le **proteine**. Tali catene si dispongono nello spazio assumendo forme anche molto complesse, si parla di struttura **primaria, secondaria, terziaria e quaternaria** delle proteine per distinguere i vari livelli di complessità di queste.

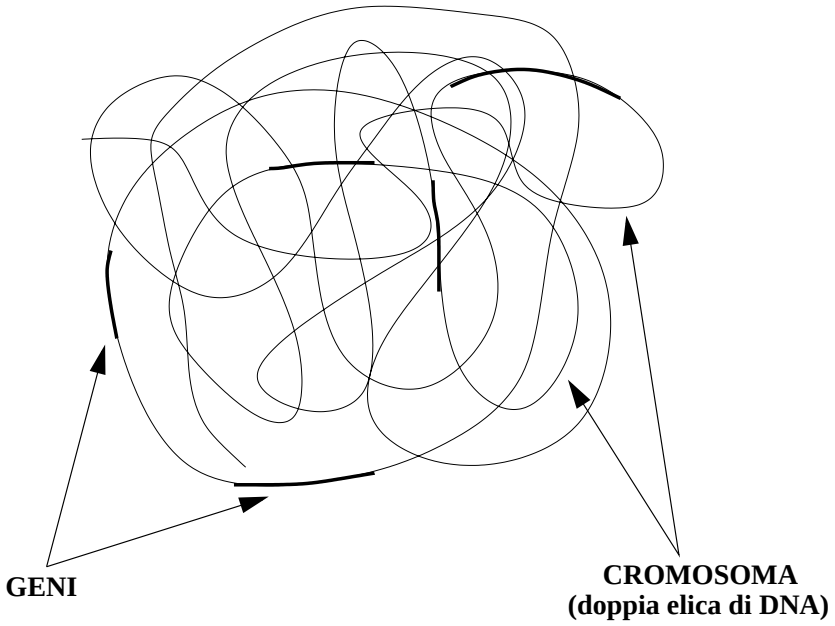


Figura A.1: Illustrazione schematica della posizione dei geni

La struttura primaria è la semplice sequenza di aminoacidi lungo la catena, data questa però risultano determinate tutte le altre.

Abbiamo detto che il DNA contiene l'informazione ereditaria necessaria allo sviluppo di un organismo, ma il DNA stesso è pressoché privo di funzionalità diretta nei processi cellulari; la parte funzionale è svolta dalle proteine. Per esempio non esiste una sequenza di DNA capace di trasportare l'ossigeno alle cellule, esiste però un gene capace di "ordinare" la produzione di una proteina che lo faccia (l'*emoglobina*). Ma come l'informazione contenuta nel DNA viene trasformata in "ordine" di produrre una proteina? Il DNA viene trascritto in RNA, una molecola con struttura molto simile ma che consiste di un solo filamento in cui la timina viene sostituita con l'*uracile*. L'RNA viene poi "letto" e tradotto in proteina[1]. Esiste una precisa corrispondenza tra le sequenze di basi e le proteine prodotte; tale corrispondenza è il **codice**

[1]Sono le regole di Watson-Crick a permettere che l'informazione passi da DNA a RNA. Nel processo di trascrizione infatti il segmento $5' - 3'$ viene usato come stampo ed è quindi univoca la sequenza di RNA prodotta.

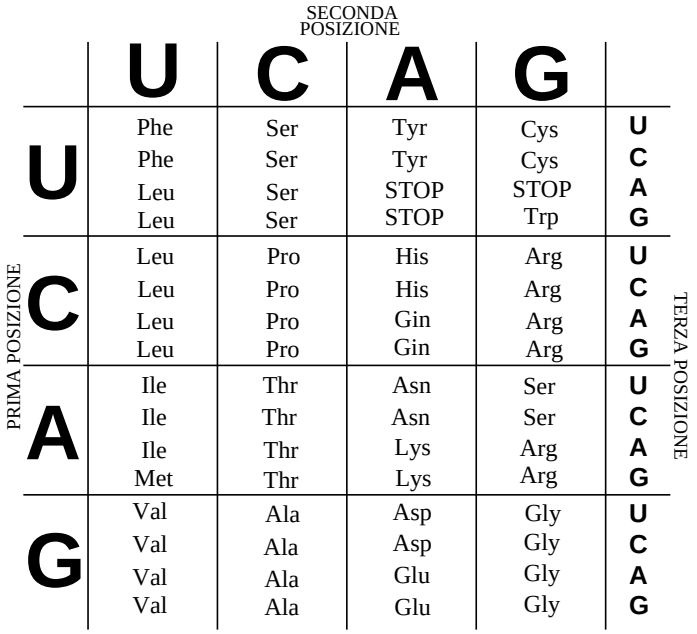


Figura A.2: Codice genetico. Viene usata la lettera U al posto della T in quanto sulla molecola di RNA la timina viene sostituita dall'*uracile*.

genetico. Grazie a questo codice ogni tripletta di basi consecutive (detta **codone**) corrisponde ad uno e un solo aminoacido. La sequenza di codoni permette di identificare la catena polipeptidica e quindi la proteina. Dato che esistono quattro basi azotate esistono $64(=4^3)$ diverse triplette, ma solo 20 aminoacidi. Ne deriva che più codoni corrispondono allo stesso aminoacido, ovvero, come si dice spesso, il codice genetico è **degenere**[2]. Come si vede dalla figura A.2 la maggiore degenerazione si ha sulla terza posizione, dove in molti casi qualunque nucleotide specifica lo stesso aminoacido.

[2]Notare che vi sono tre combinazioni che non specificano alcun aminoacido, bensì sono codoni di *stop*; servono cioè a indicare in che punto bisogna arrestare la lettura della sequenza

A.3 Mutazioni

Il materiale genetico viene sottoposto di continuo all'azione di numerosi agenti capaci di provocare *mutazioni*, come possono essere alcune sostanze chimiche o radiazioni ionizzanti. Spesso però sono i normali processi cellulari le cause di queste mutazioni. Nella stragrande maggioranza dei casi tali "errori" sono riparati mediante dei meccanismi estremamente efficienti, eppure esiste una piccola percentuale di mutazioni non corrette. Queste possono avere effetto su un solo nucleotide (in tal caso si parla di mutazione *puntuale*) oppure diversi nucleotidi adiacenti. Fra quelle puntuali distinguiamo le *sostituzioni di nucleotidi*, che consistono nella sostituzione di un nucleotide con un altro, le *inserzioni*, in cui un nucleotide viene aggiunto alla catena nucleotidica, e le *delezioni* che consistono nella cancellazione di un nucleotide dalla sequenza. È evidente che, sebbene tutte riguardino un solo nucleotide alla volta, le ultime due hanno effetto su tutta la porzione di DNA che segue, in quanto verrebbe a cambiare lo schema di lettura di tutti i codoni successivi. L'esistenza di mutazioni come le inserzioni e le delezioni (cui spesso ci si riferisce indistintamente con l'espressione *in-del*) rende necessario l'*allineamento* di sequenze omologhe prima di passare all'analisi comparativa. Si cerca cioè lo schema che realizzi il numero più alto di siti corrispondenti fra due sequenze supponendo il numero minore di in-del (per una descrizione di alcuni metodi si veda *Li* [18]). Quando si parla di modelli di sostituzione si intendono modelli che studiano mutazioni puntuali, presupponendo che le sequenze siano state già allineate e che si possano inferire i rates di mutazione dalle sostituzioni osservate (si veda la discussione nel paragrafo 1.3.5).

Appendice B

Catene di Markov

B.1 Concetti preliminari

B.1.1 Processi stocastici

Per definire una variabile stocastica $\mathbf{Y}$ bisogna specificare:
- l'insieme dei possibili valori assunti
- la distribuzione di probabilità in questo insieme.

Una volta definita una variabile stocastica se ne possono derivare un'infinità di altre mediante una funzione f che mappi $\mathbf{Y}$ in queste e sia funzione anche di una variabile addizionale t. Chiameremo allora

$$\mathbf{X_Y}(t) = f(\mathbf{Y}, t) \tag{B.1}$$

una funzione casuale, ovvero (poiché t indica di solito il tempo) un processo stocastico.

B.1.2 Distribuzioni di probabilità

Sia $\mathbf{X}$ una variabile casuale avente r componenti $X_1 \cdots X_r$. La probabilità $P_r(x_1, \ldots, x_r)$ è chiamata la *distribuzione di probabilità congiunta* delle r variabili $X_1 \cdots X_r$. Indichiamo con $P_{s|r-s}(x_1, \ldots, x_s | x_{s+1}, \ldots, x_r)$ la *probabilità* delle s variabili $X_1, \ldots, X_s$, *condizionata* alla realizzazione delle $r - s$

variabili $X_{s+1} = x_{s+1}, \ldots, X_r = x_r$. La *regola di Bayes* lega la probabilità condizionata alle probabilità congiunte dei due gruppi di variabili mediante la seguente formula:

$$P_{s|r-s}(x_1, \ldots, x_s | x_{s+1}, \ldots, x_r) = \frac{P_r(x_1, \ldots, x_r)}{P_{r-s}(x_{s+1}, \ldots, x_r)} \quad . \tag{B.2}$$

Aggiungiamo ora la dipendenza dal tempo.
Definiamo la probabilità condizionata $P_{1|1}(x_2, t_2|x_1, t_1)$ come la densità di probabilità che X assuma il valore x_2 al tempo t_2, una volta assunto il valore x_1 al tempo t_1. Vale la pena di osservare che tale probabilità è normalizzata, I.E.

$$\int P_{1|1}(x_2, t_2|x_1, t_1) dx_2 = 1. \tag{B.3}$$

Si può estendere il discorso fatto fissando k istanti di tempo e chiedendosi la probabilità di avere l realizzazioni successive a questi, ovvero considerare $P_{l|k}$ e legarla tramite la regola di Bayes a P_l e P_k, abbiamo così

$$\begin{gathered} P_{l|k}(x_{k+1}, t_{k+1}; \ldots; x_{k+l}, t_{k+l} | x_1, t_1; \ldots; x_k, t_k) = \\ \frac{P_{k+l}(x_1, t_1; \ldots; x_k, t_k; x_{k+1}, t_{k+1}; \ldots; x_{k+l}, t_{k+l})}{P_k(x_1, t_1; \ldots; x_k, t_k)} \quad . \end{gathered} \tag{B.4}$$

Queste relazioni ci torneranno utili quando introdurremo l'equazione di *Chapman-Kolmogorov*.

B.2 Processi di Markov

Un processo stocastico è detto di *Markov* se vale la seguente proprietà :

Proprietà di Markov 1 *Per ogni insieme di n istanti di tempo successivi (I.E. $t_1 < t_2 < \cdots < t_n$) si ha:*

$$P_{1|n-1}(x_n, t_n; | x_1, t_1; \ldots; x_{n-1}, t_{n-1}) = P_{1|1}(x_n, t_n | x_{n-1}, t_{n-1}) \quad .$$

In altre parole la probabilità condizionata di avere il sistema nello stato x_n al tempo t_n dato il valore x_{n-1} al tempo t_{n-1} è univocamente determinato

e non dipende dai valori agli istanti precedenti, si dice perciò che *il sistema non ha memoria*. Una volta note la probabilità ad un istante iniziale $P_1(x_1, t_1)$ e la probabilità condizionata $P_{1|1}(x_2, t_2; x_1, t_1)$ si può ricostruire l'intera evoluzione del sistema da questi, ad esempio

$$P_3(x_1, t_1; x_2, t_2; x_3, t_3) = P_{1|2}(x_3, t_3|x_1, t_1; x_2, t_2)P_2(x_1, t_1; x_2, t_2) = \\ = P_{1|1}(x_3, t_3; x_2, t_2)P_{1|1}(x_2, t_2; x_1, t_1)P_1(x_1, t_1) \quad \text{(B.5)}$$

con $t_1 < t_2 < t_3$. Iterando l'algoritmo si ottengono gli altri valori di P_n.

B.2.1 Il moto browniano

Il processo di Markov più famoso in fisica è il moto browniano. Introdurlo ci dà la possibilità di fare alcune precisazioni sui processi markoviani.

Consideriamo il moto di una particella pesante immersa in un fluido di particelle più leggere. Per semplicità consideriamo il caso unidimensionale. Le particelle leggere collideranno numerose e in maniera casuale con la particella browniana, causandone il cambiamento di velocità. Se ad esempio questa è dotata di una certa velocità v saranno gli urti frontali quelli più probabili, ma la probabilità di un cambiamento di direzione δv dipenderà solo da v e non dalla velocità agli istanti precedenti; concludiamo che la velocità di una particella browniana è un processo di Markov. Eppure le prime osservazioni sperimentali non confermavano questa tesi, finché Einstein e Smoluchowski non fecero notare che in realtà quello che si osserva non è la posizione dopo ognuno di questi cambiamenti, bensì dopo molti. Il moto di una tale particella infatti possiede un tempo di autocorrelazione, ovvero il tempo necessario all'equilibrio affinché una velocità iniziale sia stata completamente "smorzata". Questo tempo è in generale molto più piccolo della risoluzione temporale, quindi quello che si osserva è lo spostamento netto della particella dopo molti cambiamenti di velocità.

Se si considerano le posizioni successive della particella $x_1, x_2, \ldots$ abbiamo che lo spostamento $x_n - x_{n-1}$ rappresenta un altro processo di Markov, in quanto non dipende dagli spostamenti $x_{n-1} - x_{n-2}, x_{n-2} - x_{n-3}, \ldots$. Su una

scala dei tempi non troppo raffinata si ha così che non solo la velocità è un processo di Markov ma anche lo spostamento.

Facciamo ora alcune considerazioni.
La proprietà di Markov vale solo in prima approssimazione, se un certo spostamento s_k è stato abbastanza grande, sarà favorita una grande velocità "in uscita", tale velocità sopravviverà un tempo dell'ordine del tempo di autocorrelazione e quindi a sua volta favorirà uno spostamento s_{k+1} grande. La presenza stessa di un tempo di autocorrelazione diverso da zero darà origine a una lieve dipendenza tra spostamenti successivi, I.E. rappresenterà una memoria del sistema. Scegliere un'opportuna scala temporale sarà quindi necessario per rendere la descrizione più esatta. Lo stesso discorso vale per la velocità; gli urti con le particelle del fluido sono brevi, ma non istantanei, quindi una conoscenza del passato ci dice qualcosa sul tipo di urti e sul cambiamento di velocità successivo. Se poi si considera che la particella browniana crea un flusso nel fluido, questo si comporta come una riserva di memoria che viola la proprietà di Markov.

B.2.2 Caveat

Un processo di Markov può anche essere un processo a più componenti; le tre componenti della velocità nel moto browniano ad esempio. Se un processo stocastico ad r componenti rimane tale se si guardano $s < r$ componenti, altrettanto non può dirsi per un processo di Markov. La conoscenza di tutte le componenti all'istante t potrebbe essere necessaria per determinare lo stato all'istante $t + 1$. Prendiamo come esempio una miscela di gas di molecole binarie che si dissociano, nota la probabilità di dissociazione di ciascun gas, la composizione della miscela al tempo $t+1$ dipenderà da *tutti* i gas presenti al tempo t.

Al contrario, un processo che non sia markoviano può diventarlo prendendo in considerazione un numero più alto di variabili. Prendiamo il moto di una particella browniana in un campo di forze non omogeneo, il processo

che consideri solo la velocità o solo gli spostamenti non è markoviano, ma quello a due componenti che li considera entrambi lo è.

In sostanza ogni sistema fisico isolato è un processo di Markov se si considerano *tutte* le variabili microscopiche come componenti di questo, ciò perché il moto microscopico nello spazio delle fasi è deterministico. Lo scopo della fisica è quello di trovare un numero piccolo di componenti tali da poter descrivere il sistema come un processo di Markov, almeno approssimativamente. La giustificazione di tale "riduzione" è ancora oggi oggetto di discussione e rappresenta uno dei problemi fondamentali per la meccanica statistica.

Concludiamo questo paragrafo sottolineando alcune cose:

- Spesso in fisica si definisce un processo come un *fenomeno* dipendente in qualche modo dal tempo. In tal caso non ha senso chiedersi se questo sia o no markoviano senza specificare quali variabili si considerano. La difficoltà sta proprio nel trovare il numero minimo di variabili che rendono tale fenomeno almeno approssimativamente markoviano.

- La proprietà 1 deve valere per tutte le distribuzioni di probabilità P_n, non è possibile affermare che il processo è markoviano se tale proprietà vale solo per le prime distribuzioni. Sapendo che è markoviano poi si può ricostruire come mostrato l'intera gerarchia.

- Consideriamo l'equazione differenziale

$$\dot{P}(x,t) = \Omega[P(x,t)] \quad , \tag{B.6}$$

  dove Ω è un operatore che agisce su P come funzione dipendente da x. Nota la condizione iniziale $P(x,t_0)$ l'equazione si può risolvere e si conosce così univocamente $P(x,t)$ a $t > t_0$. Ciò non implica però che il processo $X(t)$ sia di Markov. Se infatti $P(x,t)$ è la probabilità che $X(t) = x$ allora l'equazione (B.6) ci dice solo che tale probabilità soddisfa un'equazione differenziale, senza nulla garantire circa le altre distribuzioni che compaiono nella 1. Ad esempio se Ω è l'operatore nullo abbiamo solo che tale probabilità è costante e quindi

il fenomeno stazionario, ma non tutti i fenomeni stazionari sono di Markov. Torneremo su un'equazione del genere quando parleremo di *master equation*.

B.2.3 L'equazione di *Chapman-Kolmogorov*

Se si integrano entrambi i membri dell'identità (B.5) rispetto a x_2 si ottiene (sempre per $t_1 < t_2 < t_3$)

$$P_2(x_1, t_1; x_3, t_3) = P_1(x_1, t_1) \int P_{1|1}(x_3, t_3|x_2, t_2) P_{1|1}(x_2, t_2|x_1, t_1) dx_2 \ . \quad \text{(B.7)}$$

Dividendo entrambi i membri per $P_1(x_1, t_1)$ e sfruttando la regola di Bayes si ottiene l'equazione di *Chapman-Kolmogorov*

$$P_{1|1}(x_3, t_3|x_1, t_1) = \int P_{1|1}(x_3, t_3|x_2, t_2) P_{1|1}(x_2, t_2|x_1, t_1) dx_2 \quad . \quad \text{(B.8)}$$

Abbiamo visto che, come descritto nell'equazione (B.5), da P_1 e da $P_{1|1}$ si può ricostruire l'intera gerarchia del processo, queste due quantità però non possono essere arbitrarie, bensì scelte in maniera da soddisfare la (B.8) e, ovviamente, la relazione

$$P_1(x_2, t_2) = \int P_{1|1}(x_2, t_2|x_1, t_1) P_1(x_1, t_1) dx_1 \quad . \quad \text{(B.9)}$$

B.2.4 Processi stazionari

Consideriamo un sistema fisico isolato, descritto da un insieme di variabili $\mathbf{X}(t)$ tali da poter essere considerate un processo di Markov. Se il sistema è all'equilibrio allora il processo di Markov si dice stazionario, le distribuzioni di probabilità non dipendono dal tempo e si determinano con le regole usuali della meccanica statistica dell'equilibrio. Per tali processi la probabilità $P_{1|1}$, che chiamaremo *probabilità di transizione*, non dipende dai valori di t ma solo dalla loro differenza, riscriviamo allora con una nuova notazione

$$P_{1|1}(x_2, t_2|x_1, t_1) = T_\tau(x_2|x_2); \qquad \tau = t_2 - t_1 \ . \quad \text{(B.10)}$$

La *Chapman-Kolmogorov* diventa quindi $(\tau, \tau' > 0)$

$$T_{\tau+\tau'}(x_3|x_1) = \int T_{\tau'}(x_3|x_2)T_\tau(x_2|x_1)dx_2 \quad , \tag{B.11}$$

che può essere interpretata come un prodotto di matrici e diventare

$$T_{\tau'+\tau} = T_{\tau'}T_\tau \quad . \tag{B.12}$$

B.3 Catene di Markov

Fra i processi di Markov si distinguono per la loro semplicità le cosiddette *catene di Markov*, definite come segue

Catena di Markov 1 *Un processo di Markov si dice catena di Markov quando*

1. *l'insieme dei possibili stati è un insieme discreto*
2. *la variabile temporale è discreta e assume solo valori interi*
3. *il processo è stazionario o almeno omogeneo, così che la probabilità di transizione dipende solo dalle differenze temporali.*

Nel nostro caso l'insieme degli stati non solo è discreto, ma anche finito, si parla quindi di *catena di Markov finita*. Indichiamo con T (matrice $N \times N$) la probabilità di transizione e $P(t)$ il vettore a N componenti che rappresenta la distribuzione di probabilità all'istante t. Data la distribuzione di probabilità all'istante iniziale $P(t = 0)$, la proprietà (B.5) insieme con la (B.12) ci dicono che all'istante $t = \tau$ avremo $P(t = \tau) = T^\tau P(t = 0)$. Lo studio delle catene di Markov richiede quindi l'analisi delle potenze $n-$esime della matrice di transizione. Questa è caratterizzata essenzialmente da due proprietà:

- gli elementi sono non negativi
- la somma degli elementi di ciascuna colonna è uno.

Nello studio che faremo tratteremo un sistema di quattro stati, ognuno di questi rappresenta la presenza in un particolare punto del genoma di una delle quattro basi azotate; adenina, citosina, guanina e timina (indicate d'ora in avanti rispettivamente con A, C, G, T). L'informazione "statistica" verrà dal campionare una sequenza per tutta la sua lunghezza che, essendo finita, darà luogo a possibili errori di campionamento.

B.4 La *master equation*

L'equazione di *Chapman-Kolmogorov* è una relazione funzionale che le probabilità di transizione devono rispettare, risulta però difficile da maneggiare e non sempre il suo significato fisico è chiaro. Da questa si ricava la *master equation*, più direttamente legata al fenomeno fisico, dalla quale appare evidente il significato del bilancio dettagliato e si calcolano in maniera più diretta le frequenze d'equilibrio. In questo paragrafo deriveremo la forma generale della *master equation*.

Facciamo di nuovo riferimento alla probabilità di transizione $P_{1|1}$, per $t_2 - t_1 = 0$ questa si riduce a

$$P_{1|1}(n_2, t_1|n_1, t_1) = \delta_{n_2,n_1} \quad . \tag{B.13}$$

Per stati continui anziché discreti la delta di Kronecker diventa di Dirac. Consideriamo un processo di Markov stazionario o almeno omogeneo, in modo da poter scrivere T_τ come probabilità di transizione. Sviluppando al primo ordine in $\tau = t_2 - t_1$ abbiamo

$$T_\tau(x_2|x_1) = \delta(x_2 - x_1)(1 - \beta\tau) + \tau W(x_2|x_1) + \mathcal{O}(\tau^2) \tag{B.14}$$

dove $W(x_2|x_1)$ è la probabilità di transizione per unità di tempo da x_1 a x_2 ed è quindi maggiore o uguale a zero. $1 - \beta\tau$ invece è la probabilità che non vi siano transizioni nel tempo τ, si spiega così perché compare davanti alla delta di Dirac. È chiaro che β indica la probabilità di transizione totale da

x_1 a tutti gli altri stati nel tempo unitario, e vale dunque[1]

$$\beta(x_1) = \int W(x_2|x_1)dx_2 \quad . \tag{B.15}$$

Introduciamo ora la (B.14) nell'equazione di *Chapman-Kolmogorov* (B.8), abbiamo

$$T_{\tau+\tau'}(x_3|x_1) = [1 - \beta(x_3)\tau']T_\tau(x_3|x_1) + \tau' \int W(x_3|x_2)T_\tau(x_2|x_1)dx_2 \ . \tag{B.16}$$

Portando $T_\tau(x_3|x_1)$ a sinistra, dividendo per τ', prendendo il limite per $\tau' \to 0$ e usando la definizione di β (B.15) si giunge all'espressione

$$\frac{\partial}{\partial\tau}T_\tau(x_3|x_1) = \int [W(x_3|x_2)T_\tau(x_2|x_1) - W(x_2|x_3)T_\tau(x_3|x_1)]dx_2 \quad . \tag{B.17}$$

Tale versione differenziale dell'equazione di *Chapman-Kolmogorov* è chiamata *master equation*.

Ricordando che $T_\tau(x_2|x_1)$ è la funzione di distribuzione $P_1(x_2)$ degli stati aventi per valore iniziale x_1 scriveremo

$$\frac{\partial P(x,t)}{\partial t} = \int [W(x|x')P(x',t) - W(x'|x)P(x,t)]dx' \quad , \tag{B.18}$$

che nel caso discreto diventa

$$\frac{dp_i(t)}{dt} = \sum_j (W_{ij}p_j(t) - W_{ji}p_i(t)) \quad . \tag{B.19}$$

È facile vedere che la distribuzione d'equilibrio si ottiene risolvendo il sistema

$$\sum_j (W_{ij}p_j^\infty(t) - W_{ji}p_i^\infty(t)) = 0 \quad \forall i \quad . \tag{B.20}$$

[1] β inoltre è legato ai cosiddetti *jump moments*, si veda *van Kampen* [19] cap.V

Bibliografia

[1] *Ou et al. - Laboratory Investigation Group & Epidemiologic Investigation Group* Molecular epimediology of HIV transmission in a dental practice, *Science* (1992) **256**, 1165-1171

[2] *Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, Leigh Brown AG* Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a syngle infected patient, *Proceedings of the National Academy of Sciences of the USA* (1992) **89**, 4835-4839

[3] *Crick FHC, Watson J* Molecular structure of nucleic acids *Nature* (1953) **171**, 737-738

[4] http://www.literature.org/authors/darwin-charles/the-origin-of-species/

[5] *Kimura M* On the probability of fixation of mutant genes in population, *Genetics* (1962) **47**, 713-719

[6] *Kimura M* Evolutionary rate at the molecular level, *Nature* (1968) **217**, 624-626

[7] *King JL, Jukes TH* Non-Darwinian evolution, *Science* (1969) **164**, 788-798

[8] *Rodriguez F, Oliver JL, Marín A, Medina JR* The general stochastic model of nucleotide substitution, *J. of Theoretical Biology* (1990) **142**, 485-501

[9] *Jukes TH, Cantor CR*: Evolution of protein molecules, in: Munro NH (ed.) *Mammalian Protein Metabolism*. Academic Press, New York

[10] *Kimura M* A simple method for estimating evolutionary rates of basesubstitution through comparative studies of nucleotide sequences, *J. of Molecular Evolution* (1980) **16**, 111-120

[11] *Barry D, Hartigan JA* Statistical analysis of hominoid molecular evolution, *Statistical Science* (1987) **2**, 191-210

[12] *Page RDM, Holmes EC*: Molecular Evolution *A phylogenetic approach*. Blackwell Science, Oxford

[13] *Takahata N, Kimura M* A model of evolutionary base substitutions and its application with special reference to rapid changes of pseudogenes, *Genetics* (1981) **98**, 641-657

[14] *Zharkikh A* Estimation of evolutionary distances between nucleotide sequences, *J. of Molecular Evolution* (1994) **39**, 315-329

[15] *Nelder JA, Mead R Computer Journal* (1965) **7**, 308-313

[16] *Hwa T, Lässig M* Similarity detection and localization, *Physical Review Letters* (1996) **76**, 2591-2594

[17] *Kimura M, Ohta T* On the stochastic model for estimation of mutational distance between homologous proteins *J. of Molecular Evolution* (1972) **2**, 87-90

[18] *Li WH*: Molecular Evolution. Sinauer Associates, Sunderland Massachussets

[19] *van Kampen NG*: Stochastic processes in physics and chemistry. North-Holland, Amsterdam