Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Corso di Laurea in Fisica

Caratterizzazione di sequenze di DNA mediante catene di Markov

Relatore:
Prof. Daniel Remondini

Presentata da:
Sara Cerasoli

Sessione II
Anno Accademico 2014/2015

A Tato,
per tutta la strada fatta insieme

Sommario

Questa tesi si inserisce nell'ambito di studio dei modelli stocastici applicati alle sequenze di DNA.
I random walk e le catene di Markov sono tra i processi aleatori che hanno trovato maggiore diffusione in ambito applicativo grazie alla loro capacità di cogliere le caratteristiche salienti di molti sistemi complessi, pur mantenendo semplice la descrizione di questi.
Nello specifico, la trattazione si concentra sull'applicazione di questi nel contesto dell'analisi statistica delle sequenze genomiche.
Il DNA può essere rappresentato in prima approssimazione da una sequenza di nucleotidi che risulta ben riprodotta dal modello a catena di Markov; ciò rappresenta il punto di partenza per andare a studiare le proprietà statistiche delle catene di DNA.
Si approfondisce questo discorso andando ad analizzare uno studio che si ripropone di caratterizzare le sequenze di DNA tramite le distribuzioni delle distanze inter-dinucleotidiche.
Se ne commentano i risultati, al fine di mostrare le potenzialità di questi modelli nel fare emergere caratteristiche rilevanti in altri ambiti, in questo caso quello biologico.

Indice

Introduzione

In questo capitolo introduttivo si cerca di contestualizzare l'ambito in cui questa trattazione si inserisce per poi descrivere più nel dettaglio la struttura stessa della tesi.
La teoria dei processi stocastici riguarda lo studio di sistemi che evolvono nel tempo (ma anche più in generale nello spazio) secondo leggi probabilistiche. Tali sistemi o modelli descrivono fenomeni complessi del mondo reale che hanno la possibilità di essere aleatori. Tali fenomeni sono più frequenti di quanto si possa credere e si incontrano tutte quelle volte in cui le quantità alle quali siamo interessati non sono prevedibili con certezza. I campi in cui i fenomeni sono modellabili con particolari processi stocastici spaziano dalle scienze naturali e biologiche a quelle mediche, dall'economia, all'industria e all'ingegneria.
Ad esempio, le catene di Markov sono utilizzate per modellare processi biologici, nello studio della crescita delle popolazioni, nello studio delle ereditarietà genetiche ed epidemiche. A volte tali modelli vengono molto, a volte troppo, semplificati al fine di poterli trattare matematicamente ma, resta il fatto che questa semplificazione permette di capire quantitativamente il comportamento del fenomeno empirico, e l'utilizzo di tali studi ha portato notevoli contributi verso una migliore comprensione del fenomeno stesso. Inoltre, nel processo del modeling non ci si aspetta che il modello usato sia quello finale: le semplificazioni iniziali portano a comprendere meglio gli aspetti che devono essere di ulteriore studio e quindi a raffinare il modello alla luce di questo.
Le analisi statistiche e in particolare lo studio della distribuzione dei first return time si sono dimostrati uno strumento potente per indagare sulle proprietà delle sequenze genomiche: l'applicazione alle catene di nucleotidi è oggetto di studio già da diversi anni e oggi è un mezzo ben affermato per far emergere caratteristiche statistiche delle sequenze

che possono far luce sulle proprietà della struttura del DNA e possono motivare un più approfondito studio dal punto di vista biologico. Inoltre, individuare proprietà funzionali e strutturali di rilevanza biologica può risultare utile nel classificare gli organismi, nel comprendere la regolazione epigenetica, o gli effetti di patologie che coinvolgono la mutazione genica.

Questa trattazione è strutturata in tre parti principali.
Nella prima, sono presentati da un punto di vista prettamente teorico, i processi stocastici del random walk e della markov chain. Dal momento che di nostro interesse sono i processi stocastici discreti, si analizza nel dettaglio il caso del random walk discreto, dandone la definizione e portando come esempio il caso più semplice del random walk unidimensionale simmetrico. Si prosegue con la trattazione delle catene di Markov, di cui i random walk possono essere un caso particolare: si da la definizione di processo markoviano e se ne descrivono le proprietà; ogni catena può essere descritta interamente dalla distribuzione di probabilità iniziale e dalla sua matrice di transizione; si parla poi del caso di catene stazionarie e si accenna alla loro rappresentazione grafica. Si introduce infine il concetto di first return time e si calcola la sua probabilità in alcuni casi; si osserva poi che per casi regolari come le catene di Markov finite a ricorrenza positiva, la distribuzione di questi segue un andamento fisicamente interessante, quello della distribuzione esponenziale.
Nella seconda parte, si entra nel merito della possibilità di applicare i modelli descritti al caso delle sequenze di DNA. Dopo una piccola parentesi prettamente biologica, necessaria a capire la struttura del DNA si affronta il problema della modellizzazione delle sequenze: fatta una serie di semplificazioni (principio base del modeling), si mostra come la sequenza possa essere vista come una catena di nucleotidi, quindi analizzabile tramite modelli di Markov e random walk.
Nell'ultima parte, sono analizzati i risultati di uno studio che ha caratterizzato le sequenze genomiche basandosi sulle distanze tra dinucleotidi: ciò è fatto a titolo di esempio per mostrare come i modelli descritti sono sfruttati in un caso di analisi di sequenze genomiche. Infatti, la modellizzazione tramite catene di Markov porta a descrivere con buoni risultati le correlazioni tra nucleotidi lungo la sequenza di DNA.

L'obiettivo è mostrare come lo studio effettuato tramite modelli e metodi della fisica sia rilevante per altri ambiti disciplinari: il riscontro di caratteristiche notevoli dal punto di vista fisico spesso può dare indicazioni e nuovi spunti alla direzione di future ricerche in diversi campi.

Capitolo 1

Random Walk e Catene di Markov

Questa prima parte della trattazione è dedicata allo studio dei processi stocastici che saranno sfruttati successivamente nell'ambito dell'applicazione al DNA. Nella prima sezione si espone la teoria dei random walk, con particolare riguardo al caso discreto, e se ne studiano alcuni semplici casi. La seconda parte è invece incentrata su un altro processo stocastico discreto, la catena di Markov, e sulle sue proprietà. Infine, si affronta il concetto di return time e della sua distribuzione statistica.

1.1 Random Walk

Un random walk è la formalizzazione matematica di un cammino costituito da una successione di passi casuali.
I random walks sono usati in diversi campi: ecologia, economia, psicologia, informatica, fisica, chimica e biologia. Infatti, sono in grado di rappresentare il comportamento di molti processi osservati in questi campi, e sono diventati un modello fondamentale per i processi stocastici.
Ci sono diversi tipi di random walk di interesse: spesso sono catene di Markov, ma altri più complessi, sono su grafi, in più dimensioni, o addirittura su superfici curve.
I random walk si differenziano anche rispetto al parametro temporale: il cammino avviene in un tempo discretizzato, indicizzato dai naturali, come X_0, X_1, X_2, ..., oppure tramite passi X_t definiti dalla variabile continua $t \geq 0$.

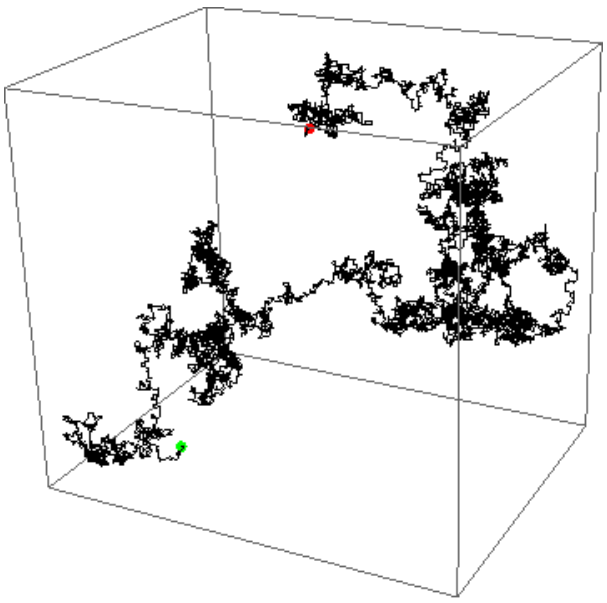


Figura 1.1: Esempio di random walk tridimensionale

I random walks sono collegati ai modelli di diffusione e costituiscono un punto fondamentale nelle discussioni sui processi di Markov: a far aumentare l'interesse nei loro confronti anche certe loro proprietà, in particolare il concetto di return time, su cui ci soffermeremo.
Vediamo in maggiore dettaglio il caso discreto.

1.1.1 Random walk discreto

Definizione . Sia $\{X_k\}_{k=1}^{\infty}$ una successione di variabili random discrete e indipendenti. Per ogni intero positivo n, sia S_n la somma $X_1 + X_2 + \cdots + X_n$. La successione $\{S_n\}_{n=1}^{\infty}$ è detta *random walk*. Se le X_k appartengono a $\mathbb{R}^m$, si dice che $\{S_n\}$ è un random walk in $\mathbb{R}^m$.

Si può pensare alla successione X_k come al risultato di esperimenti indipendenti. Visto che gli X_k sono indipendenti, la probabilità di ottenere una qualsiasi successione (finita) può essere ottenuta moltiplicando le probabilità che ogni X_k ha in quel valore della successione. Queste probabilità sono date dalla distribuzione degli X_k. Tipicamen-

te siamo interessati alle probabilità degli eventi che coinvolgono la successione S_n: tali eventi possono essere quindi descritti in termini degli X_k, e le loro probabilità possono essere calcolate usando l'idea appena esposta.

Ci sono svariati modi di visualizzare un random walk.
Si immagini che una particella sia posizionata nell'origine in $\mathbb{R}^m$ al tempo $n = 0$. La somma $\{S_n\}$ rappresenta la posizione della particella dopo n secondi. Quindi, nell'intervallo di tempo $[n-1, n]$, la particella si muove (o salta) dalla posizione S_{n-1} a S_n. Il vettore che rappresenta questo moto è $S_n - S_{n-1}$, che eguaglia X_k. Ciò significa che in un random walk, i salti sono indipendenti e identicamente distribuiti. Se $m = 1$, ad esempio, si può immaginare una particella sull'asse reale, che parte all'origine, e, alla fine di ogni secondo, salta di un'unità a destra o a sinistra, con probabilità data dalla distribuzione degli X_k. Se $m = 2$, si può visualizzare lo stesso processo che avviene in una città: una persona parte all'intersezione di due strade e può muoversi in una delle quattro direzioni possibili, a seconda della distribuzione degli X_k. Se $m = 3$, si può immaginare la stessa cosa in una giungla, con libertà di movimento in sei direzioni, e così via, sempre ottenendo le probabilità di questi movimenti dalla distribuzione degli X_k.
Un'altro modello utile (usato principalmente nel caso $\mathbb{R}^1$) è pensare a due giocatori che compiono una sequenza di mosse indipendenti e identicamente distribuite: la somma S_n rappresenta il punteggio del primo giocatore dopo n mosse, assumendo che il punteggio del secondo sia $-S_n$. Ad esempio si può pensare a una sequenza di lanci di una moneta, con testa e croce che danno $+1$ o -1, rispettivamente, per il primo giocatore.

Random walk in $\mathbb{Z}^d$

Si consideri come spazio di esistenza $\mathbb{Z}^d$, set di vettori d-dimensionali con coordinate intere. Sia e_j il vettore che ha 1 nella j-sima posizione e 0 in tutte le altre. Data una funzione

$$p(x) \qquad x \in \mathbb{Z}^d,$$

le probabilità di transizione del random walk sono date da

$$p_{x,y} = p(y - x).$$

Si definisce

$$p(x) = \begin{cases} p_j, & \text{se } x = e_j, \quad j = 1, \dots, d \\ q_j, & \text{se } x = -e_j, \quad j = 1, \dots, d \\ 0, & \text{altrimenti} \end{cases}$$

con $p_1 + \cdots + p_d + q_1 + \cdots + q_d = 1$. Un random walk di questo tipo è detto *semplice*. Se $d = 1$, allora

$$p(1) = p, \qquad p(-1) = q, \qquad p(x) = 0 \quad \text{altrimenti},$$

con $p + q = 1$ (the drunkard's walk).

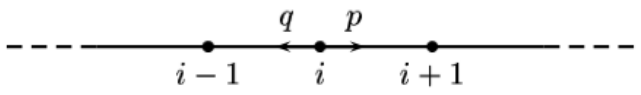


Figura 1.2: Probabilità per un random walk unidimensionale.

Definendo invece

$$p(x) = \begin{cases} \frac{1}{2d} & \text{se } x \in \{\pm e_1, \dots, \pm e_d\} \\ 0 & \text{altrimenti} \end{cases}$$

si ha ancora un random walk semplice, ma con eguali probabilità di muovere da uno stato x a uno stato vicino $x \pm e_1, \dots, x \pm e_d$. Si parla di *random walk semplice simmetrico*.

Random walk semplice simmetrico in una dimensione

Consideriamo il caso più semplice, ma non banale, di random walk in $\mathbb{Z}^1$, ovvero il caso in cui la distribuzione delle variabili random X_k è data dalla funzione di distribuzione

$$p(x) = \begin{cases} \frac{1}{2} & \text{se } x = \pm 1 \\ 0 & \text{altrimenti} \end{cases}$$

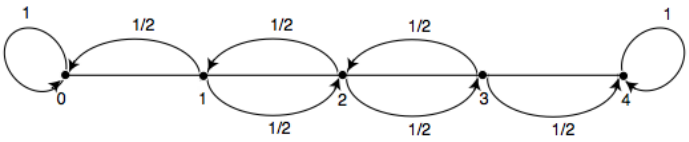


Figura 1.3: Random walk simmetrico unidimensionale.

Quando il cammino parte da 0 si può scrivere

$$S_n = X_1 + \cdots + X_n,$$

dove le X sono le variabili random con $P(X_n = \pm 1) = 1/2$. Se si guarda alla distribuzione dei $(X_1, \ldots, X_n)$, si nota che è un vettore random con valori in $\{-1, +1\}^n$. Chiaramente tutti i possibili valori del vettore sono equiprobabili. Visto che si hanno 2^n elementi in $\{0, 1\}^n$, si ha che

$$P(X_1 = \epsilon_1, \ldots, X_n = \epsilon_n) = \frac{1}{2^n}.$$

Quindi, per n fissato, abbiamo a che fare con una distribuzione uniforme sullo spazio campione $\{-1, +1\}^n$. Gli eventi A che dipendono solo dalle prime n variabili sono sottoinsiemi di questo spazio, e si ha

$$P(A) = \frac{\#A}{2^n}, \qquad A \subset \{-1, +1\}^n,$$

dove $\#A$ è il numero di elementi di A. Quindi, se possiamo contare il numero di elementi di A, possiamo calcolare la sua probabilità. Il principio è banale, ma nella pratica può essere complicato.

A volte è comodo rappresentare il random walk come una poligonale, o un cammino nel piano, dove l'asse delle ascisse rappresenta il tempo e l'asse delle ordinate rappresenta il valore S_n. Data una successione $\{S_n\}$ di somme parziali, si graficano prima i punti (n, S_n), e, successivamente, per ogni $k < n$, si collegano (k, S_k) e $(k + 1, S_{k+1})$ con un segmento. La *lunghezza* di un cammino è la differenza tra i valori temporali dei punti di

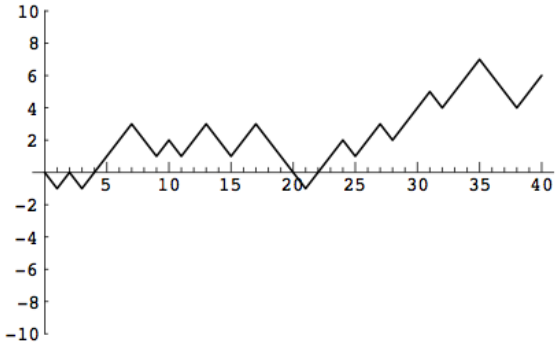


Figura 1.4: Random walk di lunghezza 40.

inizio e fine del cammino.
Il random walk unidimensionale può essere visto come una catena di Markov, argomento che verrà trattato più approfonditamente nella prossima sezione.

1.2 Catene di Markov

Una catena di Markov è un modello matematico di un fenomeno random che evolve nel tempo in un modo tale che il passato influenza il futuro solo tramite il presente. Il "tempo" può essere discreto (variabile intera), continuo (variabile reale), o, più in generale, un insieme totalmente ordinato. Si considerano, in questa trattazione, le sole catene discrete.

1.2.1 Proprietà di Markov

Si consideri una successione di variabili random $\{X_t\}$ con $t \in T$, e T sottoinsieme dei numeri interi. Si dice che la successione ha la *proprietà di Markov* se, per ogni $t \in T$, il processo futuro $(X_{t'}, t' > t, t' \in T)$ è indipendente dal processo passato $(X_{t'}, t' < t, t' \in T)$, condizionatamente su X_t. Si ha quindi che gli X_t assumono certi valori in un set numerabile S, detto *spazio degli stati*. Gli elementi di S sono detti spesso *stati*. Se S è numerabile allora (X_t) è chiamata *catena di Markov*. Per un maggiore formalismo, si riscrive la proprietà di Markov in un modo equivalente.

$(X_t, t \in \mathbb{Z}_+)$ è una catena di Markov se e solo se, per ogni $t \in \mathbb{Z}_+$ e tutti i $i_0, \ldots, i_{t+1} \in S$,

$$P(X_{t+1} = i_{t+1} | X_t = i_t, \ldots, X_0 = i_0) = P(X_{t+1} = i_{t+1} | X_t = i_t).$$

Questa è una *catena di Markov al 1°ordine.*
Una *catena di Markov di ordine m* soddisfa

$$P(X_{t+1} = i_{t+1} | X_t = i_t, \ldots, X_0 = i_0) = P(X_{t+1} = i_{t+1} | X_t = i_t, \ldots, X_{t-m+1} = i_{t-m+1}).$$

Proprietà di omogeneità nel tempo

Dato un tempo t, per il quale il processo è nello stato i_j, la probabilità che nell'istante successivo $t+1$ sia nello stato i_k è indipendente da t. Si parla di *catene omogenee nel tempo*, le quali per definizione, hanno probabilità di transizione $p_{ij}(t, t+1) = p_{ij}$ indipendente da t.

1.2.2 Matrice di Transizione

Supponiamo che al tempo t una variabile random markoviana sia nello stato i_j. La probabilità che al tempo $t+1$ sia nello stato i_k si denota con p_{jk}, detta *probabilità di tranzizione* da i_j a i_k. Scrivendo questa probabilità, si è già fatto uso delle proprietà precedentemente descritte: infatti non compare alcun tipo di dipendenza dagli stati della variabile precedenti a t (perdita di memoria), e il tempo t non appare nella notazione (omogeneità temporale).
É conveniente raggruppare le probabilità di transizione p_{jk} nella cosiddetta *matrice di probabilità di transizione*, o più semplicemente *matrice di transizione* della catena di Markov. Si denota questa matrice con P, e la si scrive,

$$P = \begin{bmatrix} p_{11} & p_{11} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{bmatrix}$$

Ogni riga della matrice corrisponde allo stato *da* cui la transizione viene fatta, ed ogni colonna allo stato *a* cui porta la transizione. Quindi, le probabilità di qualsiasi singola riga della matrice di transizione danno come somma 1, mentre le colonne non hanno particolari condizioni di somma.
Si assume inoltre che esista una qualche distribuzione di probabilità *iniziale* per i vari stati della catena di Markov. Si considera la probabilità π_i che al tempo iniziale la variabile sia nello stato i. Nel caso particolare in cui è noto che la variabile random parta da i, si ha che $\pi_i = 1$, $\pi_j = 0$ per $j \neq i$. In teoria, la distribuzione di probabilità iniziale e la matrice di transizione P determinano insieme tutte le proprietà dell'intero processo. Nella pratica, tuttavia, non è così semplice.

La probabilità che la catena di Markov si muova da uno stato i_i a uno stato i_j dopo due step può essere trovata tramite la moltiplicazione tra matrici.
Sia $p_{ij}^{(2)}$ la probabilità che la variabile markoviana sia nello stato i_i al tempo t, e poi nello stato i_j al tempo $t+2$. Questa è detta probabilità di transizione *two-step*. Visto che la variabile dovrà stare in un qualche stato k nel tempo intermedio $t+1$, la sommatoria su tutti i possibili stati al tempo $t+1$ da

$$p_{ij}^{(2)} = \sum_k p_{ik} p_{kj}.$$

Il lato destro dell'equazione rappresenta l'elemento (i, j) della matrice P^2. Quindi se la matrice $P^{(2)}$ è definita come la matrice il cui elemento (i, j) è $p_{ij}^{(2)}$, allora l'elemento (i, j) di $P^{(2)}$ eguaglia l'elemento (i, j) di P^2. Ciò porta all'identità

$$P^{(2)} = P^2.$$

L'estensione al caso di un numero arbitrario n di step studia la probabilità di transizione di passare da uno stato i a uno stato j in n step: la matrice associata è

$$P^{(n)} = \{p_{ij}^{(n)}\}.$$

Tramite le

$$p_{ij}^{(n+m)} = \sum_k p_{ik}^{(m)} p_{kj}^{(n)} \qquad \textit{Equazioni di Chapman-Kolmogorov}$$

si ottengono le relazioni matriciali

$$P^{(n+m)} = P^{(m)} P^{(n)}$$

e in particolare

$$P^{(n)} = P^{(n-1)} P$$

da cui

$$P^{(n)} = P^n.$$

1.2.3 Processo stazionario

Supponiamo che una catena di Markov abbia matrice di transizione P e che al tempo t la probabilità che il processo sia nello stato j sia φ_j, $j = 1, 2, \ldots, n$. Ciò implica che la probabilità che al tempo $t+1$ il processo sia nello stato j è $\sum_{k=1}^{s} \varphi_k p_{kj}$. Se per ogni j le due probabilità sono uguali, tale che

$$\varphi_j = \sum_{k=1}^{s} \varphi_k p_{kj}, \qquad j = 1, 2, \ldots, s$$

si dice in questo caso che la distribuzione di probabilità $(\varphi_1, \varphi_2, \ldots, \varphi_s)$ è *stazionaria*: ovvero, la probabilità che il processo sia nello stato j non cambia da t a $t+1$, e non cambierà mai. In altre parole, la legge di probabilità che governa un processo stazionario è la stessa per tutti i $t \in \mathbb{Z}_+$. Nonostante ciò, lo stato effettivamente occupato dal processo può ovviamente cambiare da un istante al successivo.
In notazione matriciale, definito il vettore

$$\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_s),$$

si può scrivere

$$\boldsymbol{\varphi} = \boldsymbol{\varphi} P.$$

Il vettore $(\varphi_1, \varphi_2, \ldots, \varphi_s)$ deve inoltre soddisfare $\sum_k \varphi_k = 1$, da cui, in notazione vettoriale,

$$\boldsymbol{\varphi} \mathbf{1} = 1.$$

Per una catena di Markov che raggiunge l'equilibrio stocastico si ha che

$$p_{ij}^{(n)} \quad \rightarrow \quad \varphi_j \qquad \text{con } n \rightarrow \infty$$

e

$$p_{ij}^{(n+1)} = \sum_k p_{ik}^{(n)} p_{kj}$$

quindi, per $n \rightarrow \infty$,

$$\varphi_j = \sum_k \varphi_k p_{kj}.$$

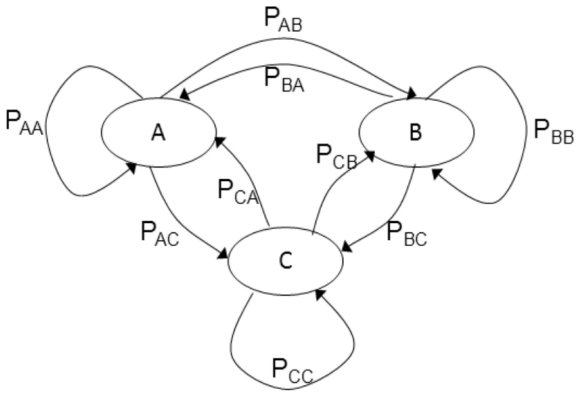
In notazione matriciale,

$$\varphi = \varphi P,$$

problema agli autovalori per trovare la distribuzione stazionaria.

1.2.4 Rappresentazione grafica

Spesso è conveniente rappresentare una catena di Markov direttamente con un grafico. Questo è composto da un insieme di "nodi" e "linee" che li connettono: le linee hanno una direzione, marcata dalle frecce che orientano quindi le linee tra un nodo e l'altro. Si identificano gli stati della catena di Markov con i nodi e le probabilità di transizione con le linee.



$$\mathbf{P}=\begin{bmatrix} p_{AA} & p_{AB} & p_{AC} \\ p_{BA} & p_{BB} & p_{BC} \\ p_{CA} & p_{CB} & p_{CC} \end{bmatrix}$$

Figura 1.5: Esempio di rappresentazione grafica di una catena di Markov.

1.3 Return times

Si consideri una catena di Markov con inizio allo stato i. Si definisce *First Return Time* dello stato i

$$T_i \doteq inf\{n \geq 0 : X_n = i | X_0 = i\}.$$

Siamo interessati a studiare la probabilità che questa quantità sia finita così come a trovare il suo valore di aspettazione. Si definisce la probabilità che si ritorni allo stato i per la prima volta dopo n step

$$f_{ii}^{(n)} \doteq P(T_i = n).$$

Se

$$P(T_i < \infty) = \sum_{n=1}^{\infty} f_{ii}^{(n)} = f_{ii} < 1.$$

lo stato i si dice *transitorio*, mentre se

$$P(T_i < \infty) = \sum_{n=1}^{\infty} f_{ii}^{(n)} = f_{ii} = 1.$$

lo stato i si dice *ricorrente*: quest'ultimo ha la garanzia (probabilità 1) di avere return time finito.
Il valore medio del tempo di ricorrenza (*Mean Recurrence Time*) è il valore atteso del first return time

$$\mu_i = \mathbb{E}[T_i] = \sum_{n=1}^{\infty} n f_{ii}^{(n)}.$$

Se μ_i è finito si dice che lo stato i è a *ricorrenza positiva.*
Si può mostrare inoltre che uno stato i è ricorrente se e solo se il valore atteso di "visite" allo stato è infinito,

$$\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty.$$

Teorema. Per ogni stato i di una catena di Markov

$$\lim_{n\to\infty} p_{ii}^{(n)} = \frac{1}{\sum_{n=0}^{\infty} f_{ii}^{(n)}} = \frac{1}{\mu_i}.$$

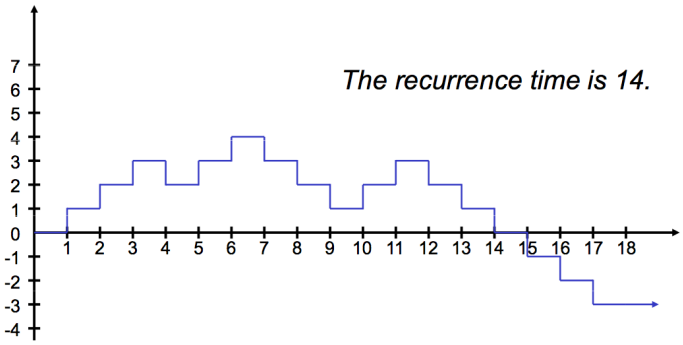


Figura 1.6: Il "recurrence time" è il primo t in cui il processo torna al suo valore iniziale).

Ma ogni catena di Markov finita ha inoltre un'unica distribuzione stazionaria tale che

$$\lim_{n\to\infty} p_{ii}^{(n)} = \varphi_i \quad \text{per ogni } i$$

da cui

$$\varphi_i = \frac{1}{\mu_i},$$

quindi i valori medi per i return time possono essere ricavati alla distribuzione stazionaria.

First return time per un random walk 1-d

Ricordando il processo descritto nella sezione 1.1, il random walk in Z è dato da

$$p_{i,i+1} = p \qquad \text{e} \qquad p_{i,i-1} = 1-p \qquad \text{per ogni } i \in \mathbb{Z}.$$

Se si parte da 0, non c'è possibilità di ritorno dopo un numero dispari di passi, quindi $p_{0,0}^{(2t+1)} = 0$ per ogni $t \in T$. Qualsiasi sequenza di passi di lunghezza $2t$ da 0 a 0 si verifica con probabilità $p^t(1-p)^t$. Ci sono quindi $\binom{2t}{t}$ sequenze possibili. Di conseguenza

$$p_{0,0}^{(2t)} = \binom{2t}{t} p^t (1-p)^t.$$

Usando l'approssimazione di Stirling $n! \simeq \sqrt{2\pi n}(n/e)^n$ si ottiene

$$p_{0,0}^{(2t)} =\simeq \frac{(2t)!}{(t!)^2}(p(1-t))^t \simeq \frac{4p(1-t))^t}{\sqrt{\pi t}}.$$

Caso simmetrico. Se $p = 1 - p$, ovvero $p = 1/2$,

$$p_{0,0} = \sum_{t \in T} p_{0,0}^{(t)} = \sum_{t \in T} p_{0,0}^{(2t)} \geq \sum_{t \in T} \frac{1}{\sqrt{\pi t}} = \infty$$

e lo stato 0 è ricorrente.

Caso asimmetrico. Se $p \neq 1 - p$, allora $4p(1-p) = r < 1$,

$$p_{0,0} = \sum_{t \in T} p_{0,0}^{(t)} = \sum_{t \in T} p_{0,0}^{(2t)} \leq \frac{1}{\sqrt{\pi}} \sum_{t \in T} r^t < \infty,$$

visto che $r < 1$, e lo stato 0 è transitorio.

First return time per un random walk 2-d

Il random walk simmetrico semplice in $\mathbb{Z}^2$ è descritto dalle probabilità di transizione

$$p_{ij} = \begin{cases} \frac{1}{4} & \text{se } |i - j| = 1 \\ 0 & \text{altrimenti} \end{cases}$$

Si ha che

$$p_{0,0} = \sum_{t \in T} p_{0,0}^{(t)} = \sum_{t \in T} p_{0,0}^{(2t)} \sim \sum_{t \in T} \frac{}{\pi t} \sim \sum_{t \in T} \frac{1}{t} = \infty$$

quindi è anch'esso ricorrente.

First return time per un random walk 3-d

Il random walk simmetrico in $\mathbb{Z}^3$, invece, descritto da

$$p_{ij} = \begin{cases} \frac{1}{6} & \text{se } |i - j| = 1 \\ 0 & \text{altrimenti} \end{cases}$$

ha probabilità di return time

$$p_{0,0} \sim \sum_{t \in T} \frac{1}{t^{3/2}} \leq \infty$$

quindi è transitorio.

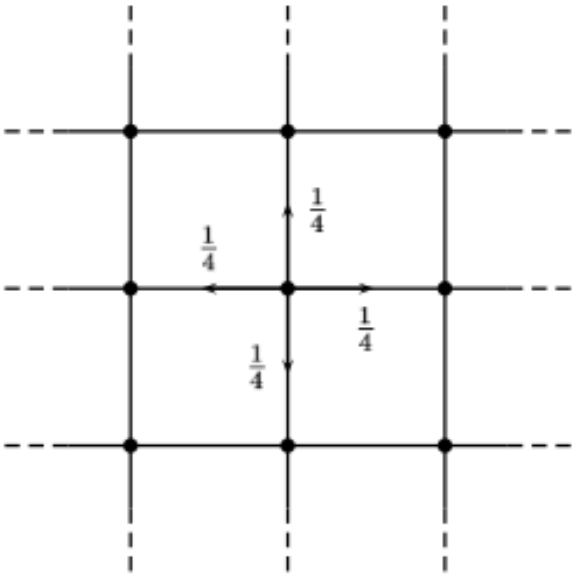


Figura 1.7: Random walk simmetrico 2D.

1.3.1 Statistica dei Return Time

Gli studi sui processi stocastici stazionari di Kac e Harris mostrarono come per molti di questi, la distribuzione dei return times segue un andamento esponenziale.
Harris si concentrò su catene di Markov a ricorrenza positiva (che sono poi quelle di nostro interesse), formate da una sequenza di stati X_n convergente all'infinito, notando appunto che che i first return times, partendo da uno stato fissato, erano distribuiti asintoticamente come un'esponenziale. I return times possiedono questa proprietà solo se la probabilità di un veloce ritorno allo stato iniziale tende a zero; altrimenti le ricorrenze allo stato iniziale posso presentarsi anche in gruppi e avere ogni sorta di distribuzione. Dato un set di variabili X_n, sia T_i il first return time allo stato iniziale i e f_i la probabilità che il return time sia minore di un tempo τ, si ha allora che

$$P[f_i T_i / \mathbb{E}[\tau] > t] \to e^{-t}.$$

In teoria della probabilità, la distribuzione esponenziale è una distribuzione di probabilità

continua che descrive un processo caratterizzato da "perdita di memoria". Infatti, se $P(X > x) = e^{-\lambda x}$,

$$P(X > x_0 + x | X > x_0) = \frac{P(X > x_0 + x, X > x_0)}{P(X > x_0)} = \frac{P(X > x_0 + x)}{P(X > x_0)} =$$

$$= \frac{e^{-\lambda(x_0+x)}}{e^{-\lambda x_0}} = e^{-\lambda x} = P(X > x).$$

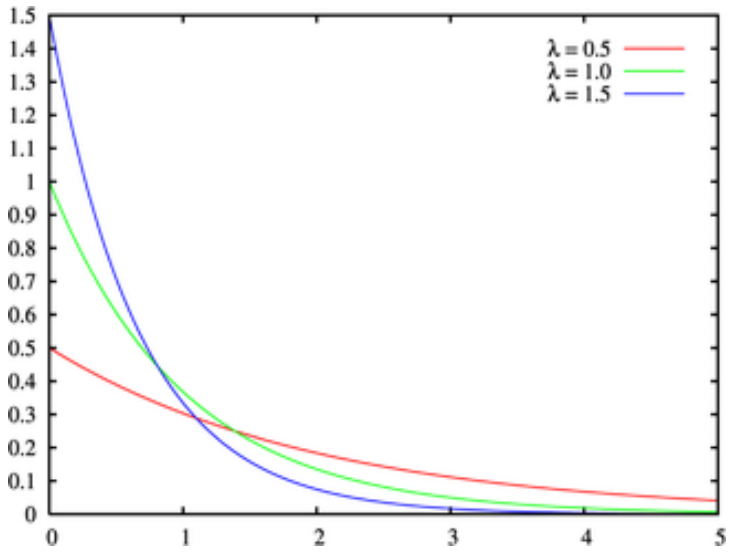


Figura 1.8: Distribuzione esponenziale per diversi valori del parametro λ.

Una variabile aleatoria X con distribuzione esponenziale di parametro λ ha

$$\text{valore atteso} \qquad \mathbb{E}[X] = \frac{1}{\lambda}$$

$$\text{varianza} \qquad Var(X) = \frac{1}{\lambda^2}.$$

La distribuzione esponenziale rappresenta inoltre il limite continuo di un'altra distribuzione con la stessa proprietà, la distribuzione geometrica.
La distribuzione geometrica è infatti una distribuzione di probabilità discreta, sui numeri

naturali: questa descrive la probabilità che un evento, il primo "successo", richieda k prove indipendenti, ognuna di probabilità p:

$$P(X = k) = (1 - p)^{k-1}p \qquad \text{con } k = 1, 2, 3, \ldots$$

dove $1 - p$ indica la probabilità di insuccesso.
Una variabile aleatoria X con distribuzione geometrica di parametro $q = 1 - p$ ha

$$\text{valore atteso} \qquad \mathbb{E}[X] = \frac{1}{p}$$

$$\text{varianza} \qquad Var(X) = \frac{q}{p^2}.$$

Come detto prima, è una distribuzione priva di memoria ed è l'unica discreta con questa proprietà.
Ogni variabile aleatoria a supporto sui numeri naturali e priva di memoria ha distribuzione di probabilità geometrica.

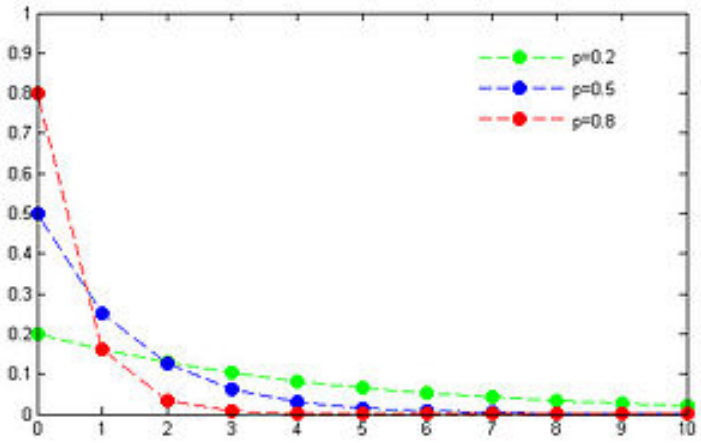


Figura 1.9: Distribuzione geometrica per diversi valori del parametro p.

Capitolo 2

DNA e Modeling

Questo capitolo tratta dell'applicazione dei modelli descritti nella prima parte al caso delle sequenze di DNA. Prima si fornisce il background biologico necessario a comprendere le caratteristiche di interesse per l'applicazione; nella seconda parte si fa luce sul problema della modellizzazione e si descrivono alcuni metodi per l'analisi statistica delle sequenze, che vedono impiegati i processi stocastici del random walk e delle catene di Markov.

2.1 Che cos'è il DNA

In tutti gli organismi viventi l'informazione ereditaria è contenuta, trasmessa e espressa con l'aiuto degli acidi nucleici DNA (acido desossiribonucleico) e RNA (acido ribonucleico). Il DNA è il materiale genetico che l'organismo eredita da propri genitori: questo fornisce le informazioni per la propria replicazione, dirige la sintesi dell'RNA e, tramite quest'ultimo, controlla la sintesi delle proteine.

Composizione e struttura di DNA e RNA
Gli acidi nucleici sono polimeri, costituiti da monomeri detti *nucleotidi*. Ogni nucleotide è composto da un gruppo fosfato, uno zucchero (deossiribosio per il DNA e ribosio per l'RNA) e una base azotata. Ci sono due famiglie di basi azotate: pirimidine e purine. Le prime includono la citosina, la timina e l'uracile (presente nell'RNA al posto della timi-

na) e sono caratterizzate da un anello esagonale composto da atomi di carbonio e azoto. Le purine sono molecole più grandi formate da un anello esagonale legato ad un anello pentagonale e includono l'adenina e la guanina. Nella struttura polimerica, i nucleotidi adiacenti sono uniti da un legame fosfodiesterico: un gruppo fosfato lega gli zuccheri di due nucleotidi. Questo legame ha come risultato un'ossatura con un pattern ripetitivo di unità zucchero-fosfato caratterizzate da un'intrinseca direzionalità: un estremità ha un fosfato attaccato al quinti atomo di carbonio, mentre l'altra ha un gruppo idrossile sul terzo carbonio.

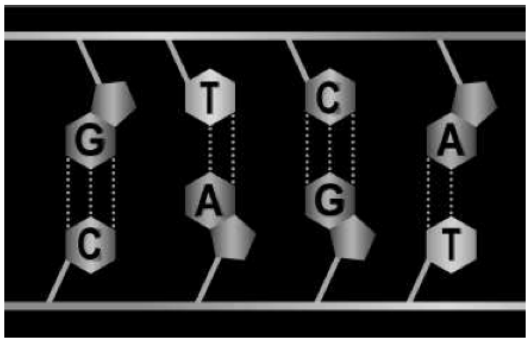


Figura 2.1: Schema dei legami tra basi azotate.

Le molecole di RNA esistono solitamente come singole catene di nucleotidi, mentre le molecole di DNA hanno due catene polinucleotidiche, o filamenti, che si avvolgono a spirale intorno ad un asse immaginario, formando una struttura a doppia elica.

Solo alcune basi nella doppia elica sono compatibili tra di loro: l'adenina si appaia sempre con la timina con due legami idrogeno e la citosina con la guanina tramite tre legami idrogeno. I due filamenti sono quindi complementari: se dovessimo leggere la sequenza di basi lungo uno dei filamenti della doppia elica, conosceremmo la sequenza di basi dell'altro. Questa caratteristica unica del DNA permette la creazione di due copie identiche di ciascuna molecola di DNA nella cellula che si prepara a scindersi, formando cellule figlie

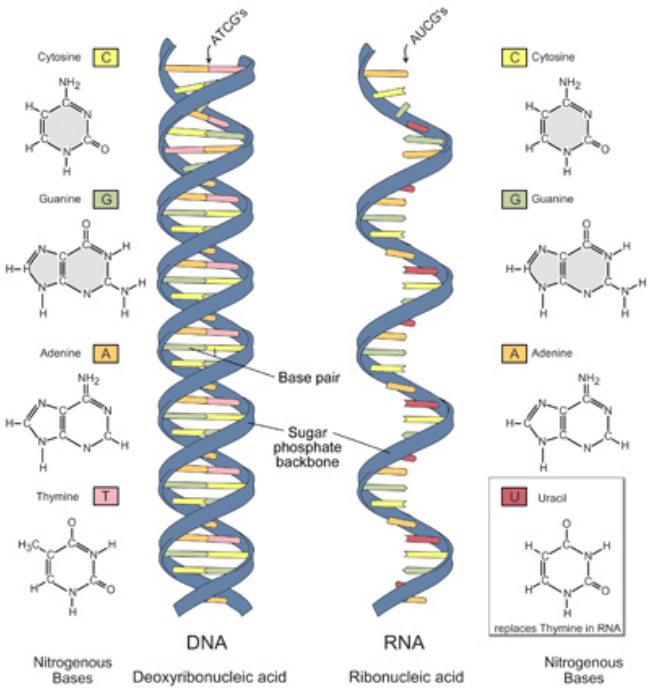


Figura 2.2: Struttura del DNA e del RNA

geneticamente identiche alla cellula genitore. L'accoppiamento delle basi avviene anche nel RNA tra basi di due diverse molecole di RNA o sulla stessa: ad esempio, questo fatto è responsabile della funzionale struttura tridimensionale dell'RNA transfer (tRNA).

2.2 Modeling

Al livello più elementare possibile, la struttura del DNA può essere pensata come una lunga sequenza di nucleotidi. Queste sequenze sono organizzate in parti codificanti, o geni, separate da lunghe regioni intergeniche non codificanti. All'interno di ciascun gene poi, le sequenze di coding (exons) sono spesso interrotte da sequenze noncoding (introni). Le regioni intergeniche e gli introni hanno proprietà statistiche differenti da quelle degli introni. Da qui l'interesse nel trovare un modello che colga queste diverse proprietà, per poter identificare il ruolo del pezzo di sequenza che si vuole analizzare. I modelli statistici per le sequenze genomiche si sono dimostrati molto utili e fondamentali

nelle applicazioni. In questa sezione si studiano i casi più semplici, tra cui appunto il modello basato sulle catene di Markov, nelle quali il concetto di tempo è sostituito dalla posizione lungo la sequenza.
Nel contesto dell'analisi di sequenze di DNA, è noto che le caratteristiche proprie delle catene di Markov non sono rispettate: i dati sperimentali mostrano che la probabilità che un nucleotide sia susseguito da un altro dipende in qualche modo dalla posizione lungo il cromosoma e dai nucleotidi immediatamente precedenti, quindi non sussiste la proprietà di "perdita di memoria".
Tuttavia, vedremo che il modello di Markov riproduce bene molte caratteristiche delle sequenze, se si sono assunte alcune semplificazioni in partenza: la catena di DNA è considerata omogenea, ovvero deve avere la stessa composizione per tutta la sua lunghezza, quindi le probabilità di transizione si ripresentano uguali in ogni punto della sequenza.

2.2.1 Modelli random per le sequenze di DNA

Catena di Markov di ordine 0 o indipendence model

Nelle analisi statistiche delle sequenze genomiche è essenziale comparare i risultati ottenuti con un appropriato *null model*: in questo modo infatti, è possibile rigettare l'ipotesi che la caratteristiche osservate accadano per caso, e confermare invece che sono biologicamente significative.
Ciò si traduce nel generare sequenze di DNA "artificiali", nelle quali i nucleotidi sono posizionati in modo random: è il modello più semplice in assoluto, detto anche *independence model*.
Supponiamo che la sequenza di DNA sia la realizzazione di variabili casuali $X_1, X_2, \ldots, X_n$, con X_t che rappresenta la base nella posizione t lungo la sequenza, che prendono valori nello spazio degli stati $S = \{A, C, G, T\}$. Questo modello assume che le X_t sono indipendenti e che, per $t = 1, 2, \ldots, n$

$$Pr(X_t = A) = p_A, \quad Pr(X_t = C) = p_C, \quad Pr(X_t = G) = p_G, \quad Pr(X_t = T) = p_T.$$

con p_A, p_C, p_G, p_T probabilità per ciascun nucleotide, che soddisfano la relazione $p_A + p_C + p_G + p_T = 1$. La sequenza è costruita quindi scegliendo il nucleotide seguendo

la sua probabilità fissata, determinata a partire dalla relativa frequenza osservata nelle sequenze biologiche. Questo modello può essere visto come caso particolare di catena di Markov, con matrice di transizione

$$\begin{bmatrix} p_A & p_C & p_G & p_T \\ p_A & p_C & p_G & p_T \\ p_A & p_C & p_G & p_T \\ p_A & p_C & p_G & p_T \end{bmatrix}$$

nel quale le probabilità di transizione dipendono dalle precedenti zero basi e per questo si può chiamare catena di Markov di ordine 0.

Il confronto con i dati reali mostra in maniera evidente che le sequenze di DNA non sono affatto stringhe di lettere indipendenti, quindi è necessario un modello più appropriato, che colga le dipendenze che si sanno esistere tra le basi.

Catena di Markov di ordine 1

Uno dei modelli random più largamente utilizzati per le sequenze di DNA è basato sulle catene di Markov: i modelli catena di Markov infatti sono in grado di cogliere le correlazioni a corto raggio tra le basi. Questi, tuttavia, sono pur sempre modelli molto semplificati che non possono riprodurre molte caratteristiche più complesse delle sequenze di DNA, come le correlazioni a lungo raggio.

Una rappresentazione grafica della catena di Markov del DNA (al primo ordine) è riportata in figura: gli stati sono rappresentati dai quattro nucleotidi A, T, C, G, e le frecce rappresentano le probabilità di transizione da uno stato all'altro.
In questo modello la sequenza è costruita considerando le probabilità di transizione al primo ordine, quindi la probabilità di avere un nucleotide nella posizione t dipende dal nucleotide immediatamente precedente in posizione $t-1$:

$$Pr(X_t = C|X_{t-1} = A, X_{t-2}, X_{t-3}, \ldots, X_1) = Pr(X_t = C|X_{t-1} = A) = p(A|C) = p_{AC}.$$

Simili probabilità di transizione descrivono anche le altre possibili combinazioni e sono raccolte tutte all'interno della matrice di transizione.

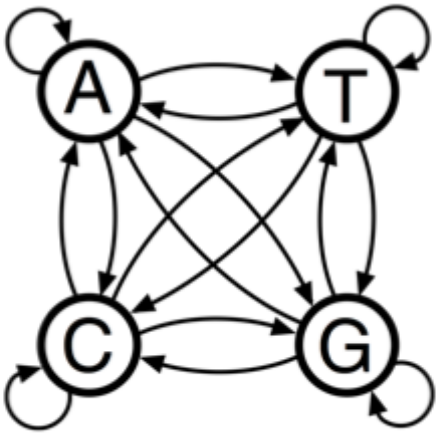


Figura 2.3: Esempio di una catena di markov del DNA

La matrice di transizione per un modello di DNA a catena di Markov al primo ordine è:

$$\begin{bmatrix} p(A|A) & p(C|A) & p(G|A) & p(T|A) \\ p(A|C) & p(C|C) & p(A|G) & p(T|C) \\ p(A|G) & p(C|G) & p(G|G) & p(T|G) \\ p(A|T) & p(C|T) & p(G|T) & p(T|T) \end{bmatrix}$$

La prima riga, ad esempio, descrive come, man mano che ci si muove lungo la sequenza, la base A sia seguita da ciascuna delle quattro possibilità A, C, G, T. Ricordando la proprietà della matrice di transizione per la quale la somma degli elementi di ogni riga è 1, sappiamo che una transizione avverrà. Le probabilità di transizione sono stimate a partire sequenze biologiche: ad esempio, la probabilità $p(A|C)$ di avere A quando nel passo precedente si aveva C è approssimata al rapporto tra le frequenze osservate dei dinucleotidi AC e dei nucleotidi C nella sequenza di DNA considerata:

$$p(A|C) = \frac{\#AC/\#dinucleotides}{\#C/\#nucleotides}$$

Modelli di Markov di ordine più alto darebbero origine a sequenze più "biologicamente" corrette: ad esempio in un modello di ordine due avremmo probabilità di transizione

del tipo $p(A|CT)$, e questa "memoria" permetterebbe la ricostruzione di "parole" da tre lettere nel DNA, che corrispondono ai *codoni* (parti della sequenza che codificano un singolo amino acido). Tuttavia, il primo ordine risulta già sufficiente per una buona descrizione, dal momento che ci interessa compararla a una sequenza costruita in modo random, con solo le percentuali di frequenza di dinucleotide simili.

2.2.2 Modello random per la distanza internucleotidica

Dal null model è possibile modellizzare direttamente anche le distribuzioni delle distanze tra nucleotidi (e dinucleotidi) : infatti, se questi fossero distribuiti in modo del tutto casuale lungo la sequenza, le distanze alle quali si ripresenta lo stesso dinucleotide dovrebbero seguire la distribuzione geometrica:

$$f_x(k) = p_x(1 - p_x)^{k-1}$$

dove p_x è la probabilità di x nella sequenza, stimata a partire dalla sua frequenza nella sequenza biologica (ad esempio, p_{AT}).
La distribuzione geometrica rappresenta quindi il modello random di riferimento per le distanze internucleotidiche. Si noti che questo modello non è indipendente dai modelli random costruiti per le sequenze: infatti, la distribuzione geometrica viene calcolata a partire da una sequenza random in cui si sono fissate unicamente le frequenze dei (di)nucleotidi, partendo dai dati reali, ovvero una catena di Markov di ordine zero.

Il modello di riferimento è confrontato quindi con la distribuzione di probabilità calcolata a partire dai dati reali nel seguente modo. Considerata una sequenza $s = \{s_j\}_{j=1}^{N}$ i cui s_j prendono valori tra {A, C, G, T}, si costruisce la sequenza degli indici per i quali ritroviamo il nucleotide X (o dinucleotide XY) considerato:

$$\{r_j | s_{r_j} = X\} \qquad \text{per le distanze inter-nucleotidiche}$$

$$\{r_j | s_{r_j} s_{r_{j+1}} = XY\} \qquad \text{per le distanze inter-dinucleotidiche.}$$

La sequenza delle inter-distanze $\{\tau_j\}$ è poi calcolata come la differenza tra indici successivi

$\tau_j = r_{j+1} - r_j$. La distribuzione di probabilità sarà data da

$$p(\tau) = \frac{\#\{j|\tau_j = \tau\}}{\#\{\tau_j\}}.$$

Per comparare quantitativamete le distribuzioni che si ottengono per i diversi (di)nucleotidi, si introduce la *divergenza di Kullback-Leibler*: questa è una misura della differenza tra due distribuzioni di probabilità $f(x)$ e $g(x)$. Rappresenta l'informazione persa quando $g(x)$ approssima $f(x)$: più precisamente misura il numero di bit richiesti in più quando si codifica una variabile random con distribuzione $f(x)$ usando una distribuzione alternativa $g(x)$. Per due distribuzioni $f(x)$ e $g(x)$ e una variabile casuale X, la divergenza KL è definita come

$$D(f||g) = \sum_{x \in X} f(x) \log \frac{f(x)}{g(x)}$$

e possiede le seguenti proprietà:

- $D(f||g) > 0$
- $D(f||g) \neq (g||f)$
- $D(f||g) = 0$ se e solo se $f(x) = g(x)$ per ogni x.

Un altro modo di stimare quantitativamente la differenza tra due diverse distribuzioni si ottiene con la distanza di Jensen-Shannon D_{JS}, la versione simmetrizzata della divergenza di Kullback-Leibler D:

$$DJS = \frac{1}{2}D(f||m) + \frac{1}{2}D(g||m) \qquad \text{con} \qquad m : m(x) = \frac{f(x) + g(x)}{2}.$$

Per visualizzare meglio le leggi seguite dalle distribuzioni è comodo utilizzare grafici logaritmici o doppiamente logaritmici, in modo tale da evidenziare un comportamento esponenziale o a legge di potenza.

2.2.3 DNA walk

In questo paragrafo vogliamo discutere dei differenti metodi per rappresentare le sequenze di DNA come random walk.
Uno dei primi lavori che tradussero la sequenza di DNA in un walk, detto *DNA walk*, fu quello di Peng, Buildyrev e altri, sulle correlazioni a lungo raggio delle sequenze genomiche. Questi definirono un DNA walk basato su un random walk unidimensionale, nel quale i movimenti del walker possono essere *su* ($x(k) = +1$) o *giù* ($x(k) = -1$), con lunghezza unitaria (x) per ciascun step k.
Lo spostamento del walker dopo l step è dato dalla somma dei passi unitari:

$$s(l) = \sum_{i=1}^{l} x(i)$$

Nel caso del DNA walk, il walker andrà su nel caso incontri lungo la catena di DNA una pirimidina (nucleotidi C o T), e giù nel caso incontri una purina (nucleotidi A o G).
Un analisi dello scarto quadratico medio della fluttuazione, su una media dello spostamento, enfatizza la presenza di una correlazione a lungo raggio nelle sequenze, specialmente per gli introni.

Sono possibili diversi altri modi di convertire una sequenza di nucleotidi in un walk unidimensionale: ad esempio, si può distinguere tra nucleotidi con legami idrogeno forti (C e G) e deboli (A e T); se invece si fosse interessati alla localizzazione di un particolare dinucleotide, ad esempio il CG, il walker sarà costruito muovendo su nel caso incontri il dinucleotide CG o giù altrimenti.

Sono inoltre possibili walk più complessi e in più dimensioni: ad esempio, si possono mappare i nucleotidi nel piano complesso, facendo corrispondere alla presenza di A, G, T, C nella sequenza i punti $(+1, -1, +j, -j)$. In questa rappresentazione le purine danno valori sull'asse reale, mentre le pirimidine sono limitate all'asse immaginario. Un walk di questo tipo aiuta nell'identificazione di proprietà periodiche nella struttura nucleotidica.

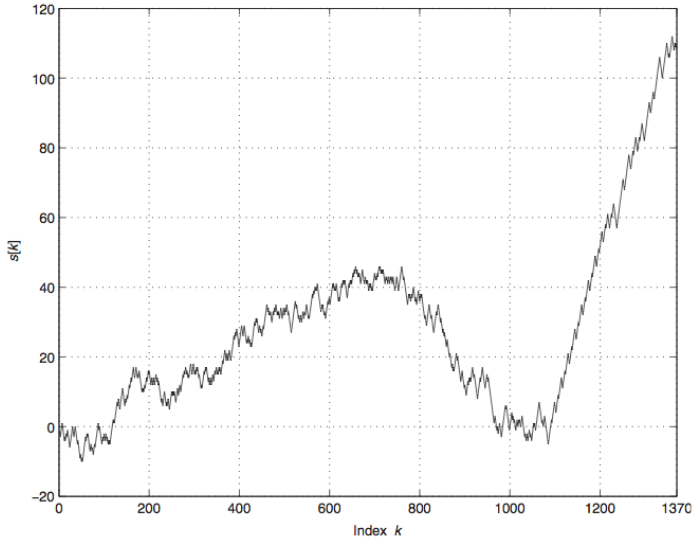


Figura 2.4: DNA walk unodimensionale. Questa rappresentazione illustra i contenuti relativi di purine e pirimidine in una regione non codificante della sequenza genomica dell'*Helicobacter pylori*.

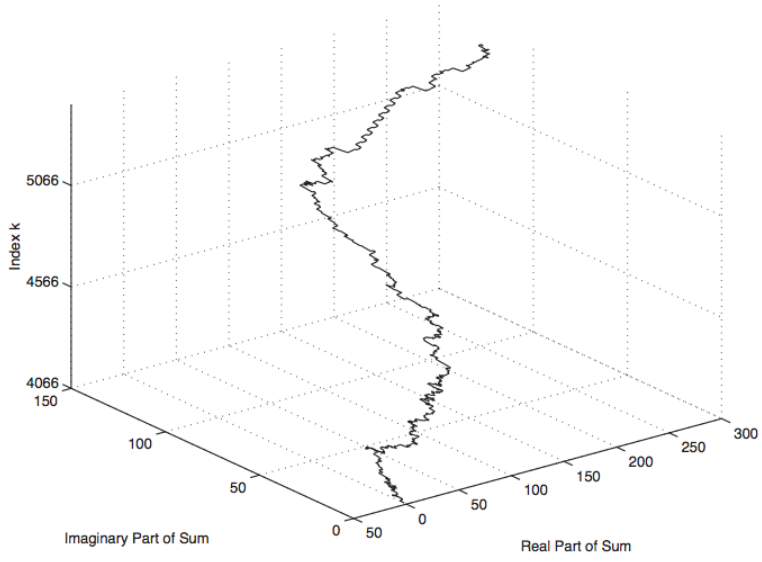


Figura 2.5: DNA walk due-dimensionale. Questa rappresentazione evidenzia l'evoluzione della sequenza e l'andamento della composizione nucleotidica.

Capitolo 3

Analisi delle distanze inter-dinucleotide

In quest'ultima parte della tesi ci si propone di mostrare come i metodi descritti nelle sezioni precedenti possono essere sfruttati per far emergere delle caratteristiche di interesse biologico.
Con questo obiettivo, si fa riferimento a uno studio sulle distanze inter-dinucleotidiche delle sequenze genomiche (vedi [10]) e se ne analizzano i risultati, sottolineando come l'analisi fisica delle proprietà statistiche osservate possa essere messa in relazione ad aspetti biologici rilevanti.

3.1 Studio delle distribuzioni delle distanze interdinucleotidiche nelle sequenze genomiche

Lo studio di nostro interesse ha coinvolto le sequenze biologiche di DNA di 21 differenti organismi, delle quali sono state caratterizzate le distribuzioni dei dinucleotidi.
I genomi sottoposti all'analisi rappresentano l'intera sequenza di DNA di ciascuno degli organismi, ottenuta concatenando tutti i cromosomi.
Il modello sfruttato per studiare le distribuzioni dei dinucleotidi è stato descritto nel capitolo precedente e, ricordiamo, si basa sulle distanze, calcolate in numero di basi, alle quali si ripresenta lo stesso dinucleotide. Per ogni organismo, quindi, si osservano tutte le distribuzioni delle inter-distanze, una per ognuno dei 16 dinucleotidi.

3.1.1 Distribuzione dei dinucleotidi nel DNA umano

L'analisi è dapprima incentrata sulla distribuzione dei dinucleotidi ottenuti per il genoma umano. Il grafico logaritmico, riportato in Figura 3.1, delle distribuzioni dei 16 dinucleotidi all'interno del DNA umano mostra la differenza dell'andamento delle interdistanze: per tutti i dinucleotidi si osserva un decadimento algebrico del tipo $p(\tau) \sim \tau^{-b}$ (coefficiente di regressione lineare $r^2 \geq 0.94$), con esponente $b \sim 3$ simile. Fa eccezione il dinucleotide CG che invece mostra decadimento esponenziale $p(\tau) \sim e^{-d \cdot \tau}$ (con parametro $d = 0.004 \pm 0.001$ e $r^2 = 0.999$).
Come visto nello studio della statistica dei return time, una distribuzione esponenziale descrive un processo stocastico caratterizzato dalla proprietà di perdita di memoria, quindi indipendente dagli stati passati, in questo caso, dalle basi della parte di sequenza precedenti. Inoltre, questo tipo di distribuzione ha una sua lunghezza caratteristica (o meglio una frequenza di ricorrenza dell'evento), che in questo caso rappresenta la distanza alla quale si ripresenta lo stesso dinucleotide (per il DNA umano il valore è $\lambda = 1/d \simeq 250$ basi).
Dall'osservazione del grafico, quindi, il dinucleotide CG sembra essere ben descritto da una variabile aleatoria che esegue una sorta di random walk lungo tutta la sequenza di DNA, con una lunghezza caratteristica di ricorrenza.
Questo andamento è ben differente da quello degli altri dinucleotidi che invece seguono una distribuzione dei return times a legge di potenza, la quale non possiede un'unica scala caratteristica, per cui non è possibile dare un unico valore alla frequenza di ricorrenza per gli altri dinucleotidi.
La differenza di comportamento osservata può essere ricondotta ai diversi ruoli dei dinucleotidi nel DNA umano: il dinucleotide CG, infatti, è sede del processo della metilazione del DNA, un meccanismo epigenetico coinvolto nella regolazione genica e nella conformazione strutturale della cromatina.

3.1.2 Confronto con altri organismi

Dopo aver considerato il caso specifico dell'uomo, si vanno ad osservare le distribuzioni ottenute per gli altri organismi. Sono riportati come esempi i grafici di un mammifero

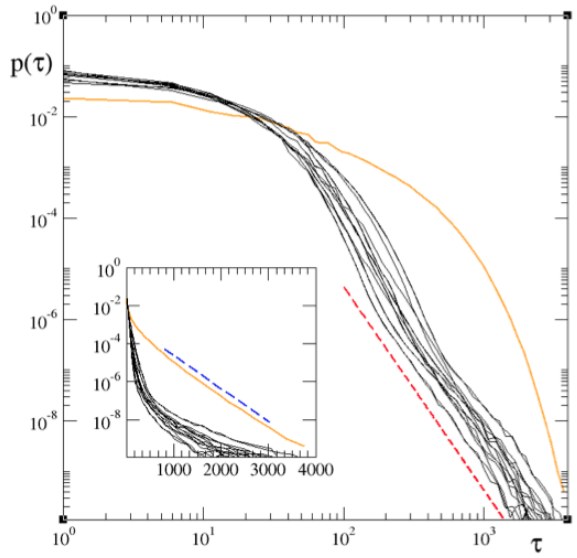


Figura 3.1: Grafico delle distribuzioni dinucleotidiche nel DNA umano. Nel grafico logaritmico doppio e, all'interno, logaritmico semplice, la distribuzione di CG è evidenziata.

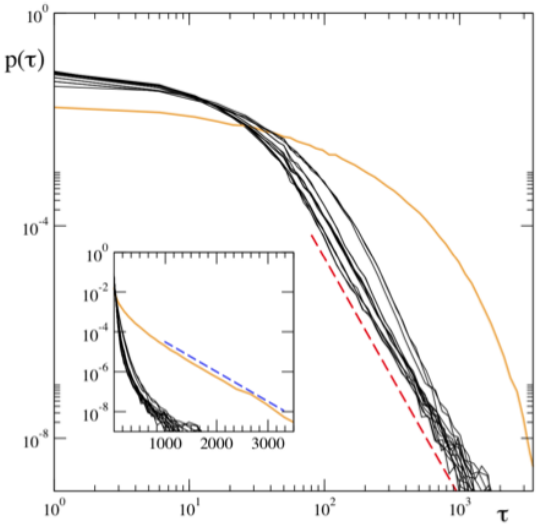


Figura 3.2: Grafico logaritmico doppio delle distribuzioni dei dinucleotidi per *Mus Musculus* con la distribuzione CG evidenziata.

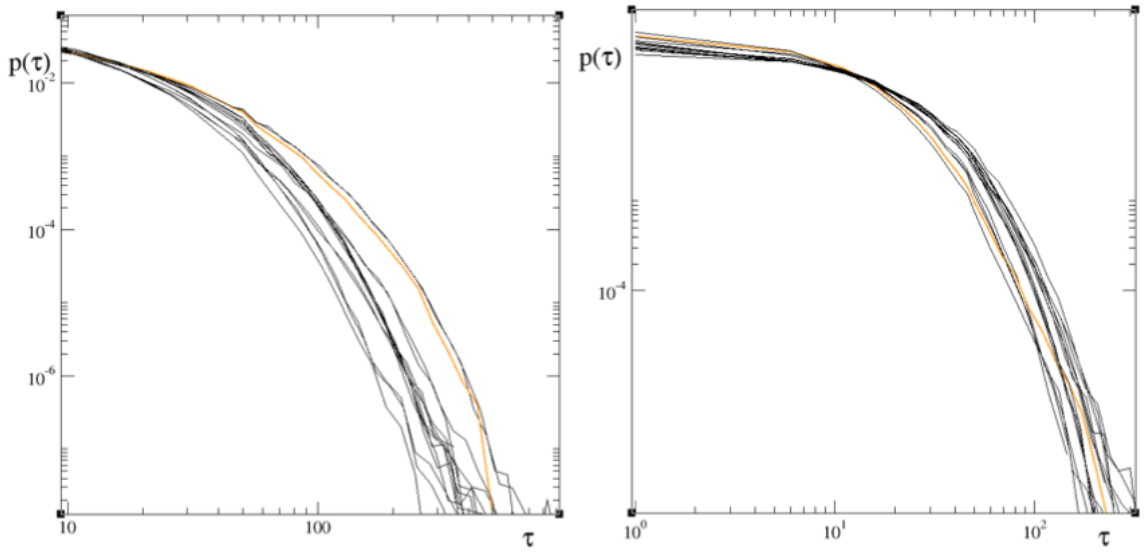


Figura 3.3: Grafico logaritmico doppio delle distribuzioni dei dinucleotidi per *D. melanogaster* e *E.coli*, con la distribuzione CG evidenziata.

(*Mus Musculus*), un insetto (*D. Melanogaster*) e un batterio (*E. coli*)(vedi Figure 3.2 e 3.3).

Il grafico delle distribuzioni per il *Mus* mostra una grande somiglianza con quello dell'uomo, in particolare per il netto discostamento della distribuzione CG. Ciò suggerisce le stesse considerazioni fatte per il DNA umano, e in aggiunta, si deduce che il processo di metilazione che attacca il dinucleotide CG agisca con un meccanismo analogo e che quindi questo possa essere comune a una certa classe di organismi.

I grafici delle distribuzioni di *D. melanogaster* e *E.coli*, invece, presentano un andamento simile per tutti e 16 i dinucleotidi, il che suggerisce che non ci siano particolari processi che coinvolgano un solo dinucleotide o che la metilazione agisca in modo del tutto differente per questi organismi.

Per meglio confrontare i dati ottenuti per i diversi organismi, si comparano le distribuzioni dei 16 dinucleotidi mediante la distanza di Jensen-Shannon. Osservando i grafici (vedi Figura 3.4 e 3.5) delle distanze di JS, è facile accomunare in un'unica categoria un gruppo di 10 organismi che sono tutti i mammiferi tra gli organismi in esame. In

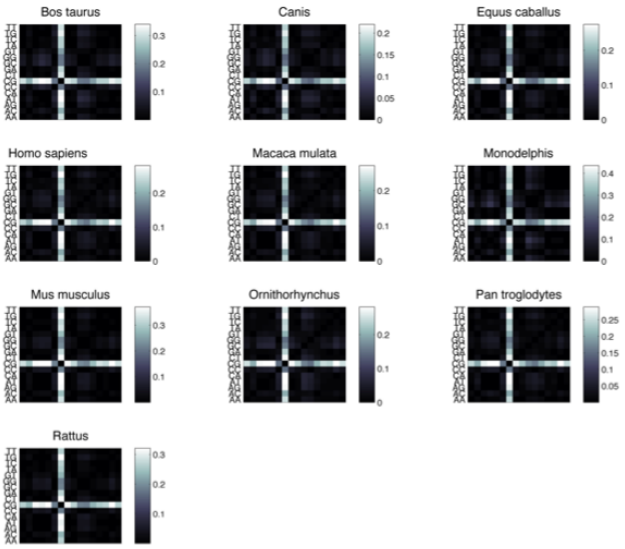


Figura 3.4: Grafico delle distanze di Jensen-Shannon tra le distribuzioni delle distanze inter-dinucleotidiche per i mammiferi

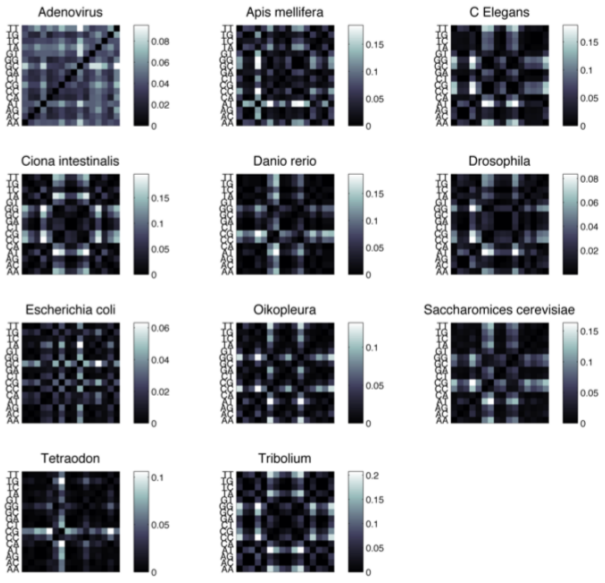


Figura 3.5: Grafico delle distanze di Jensen-Shannon tra le distribuzioni delle distanze inter-dinucleotidiche per i restanti 11 organismi.

Organism	Max	d	λ	r^2
Bos Taurus	3709	0.0037 ± 0.0001	272 ± 1	0.999
Canis Familiaris	3248	0.0036 ± 0.0001	274 ± 1	0.999
Equus Caballus	2927	0.0047 ± 0.0001	214 ± 1	0.999
Homo Sapiens	3760	0.004 ± 0.0001	252 ± 1	0.999
Macaca Mulatta	3907	0.0042 ± 0.0001	240 ± 1	0.999
Monodelphis domestica	8123	0.0022 ± 0.0001	452 ± 1	0.999
Mus Musculus	4617	0.0034 ± 0.0001	295 ± 1	0.999
Ornithorhynchus anatinus	2841	0.0043 ± 0.0001	232 ± 1	0.999
Pan Troglodytes	3376	0.004 ± 0.0001	248 ± 1	0.999
Rattus norvegicus	3845	0.0039 ± 0.0001	257 ± 1	0.998
Adenovirus	517	0.012 ± 0.001	83 ± 3	0.845
Apis Mellifera	6958	0.0033 ± 0.0001	296 ± 2	0.995
Caenorhabditis Elegans	4284	0.0015 ± 0.0001	647 ± 8	0.946
Ciona Intestinalis	3560	0.002 ± 0.0001	490 ± 18	0.688
Danio Rerio	4072	0.0035 ± 0.0001	288 ± 2	0.979
Drosophila melanogaster	568	0.023 ± 0.001	44 ± 1	0.992
Escherichia coli	324	0.037 ± 0.001	27 ± 1	0.973
Oikopleura diotica	679	0.019 ± 0.001	51 ± 1	0.920
Saccharomices Cerevisiae	308	0.027 ± 0.001	37 ± 1	0.995
Tetraodon nigroviridis	1573	0.0032 ± 0.0001	312 ± 7	0.883
Tribolium castaneum	2455	0.0026 ± 0.0001	388 ± 3	0.983

Figura 3.6: Tabella per i valori del fit esponenziale della distribuzione delle inter-distanze del dinucleotide CG, per tutti gli organismi. Per ognuno, è riportata la massima distanza tra CG (max), il parametro di fit d, il coefficiente di regressione r^2 e la lunghezza caratteristica (inverso di d).

particolare si nota che questi presentano la stessa distinzione per il dinucleotide CG, che non è osservata per gli altri. Per i restanti 11 organismi si osserva invece un comportamento molto più eterogeneo. L'unico che sembra presentare una distribuzione CG simile è il *Tetraodon*, e, meno, il *Danio rerio*, dei pesci. Risulta più complicato organizzare in famiglie quest'altro gruppo, dal momento che la gamma di complessità è vasta e ci sono pochi rappresentanti di ciascuna famiglia.
Per meglio caratterizzare la distribuzione CG e verificare l'andamento esponenziale osservato, si riportano anche i valori ottenuti dal fit delle code delle distribuzioni con una funzione esponenziale (vedi Figura 3.6).
Per prima cosa si osserva che i parametri del fit sono molto buoni ($r^2 > 0.998$) per il gruppo di organismi già accomunato dall'analisi delle distanze di JS, che mostravano una differente distribuzione CG. Per questi, si ottengono valori molto simili e in particolare si nota che le lunghezze caratteristiche stanno tutte nell intervallo tra 200 e 300 basi, con l'unica eccezione del *Monodelphis*.

Per l'altro gruppo di organismi la situazione è ancora molto eterogenea: per l'*Apis* e *Danio* il fit esponenziale sembra buono e la lunghezze caratteristiche sono comparabili con quelle dei mammiferi, mentre la *Drosophila* e *Saccharomices* hanno valori di fit addirittura migliori ma le loro lunghezze caratteristiche differiscono di un ordine di grandezza. Per altri organismi invece, come *Adenovirus*, *Ciona*, *Oikopleura*, *Tetraodon* il fit esponenziale non è molto in accordo con la distribuzione originaria. Per i restanti, sono trovati valori intermedi.
Tra le osservazioni da fare, il fatto che alcune lunghezze caratteristiche siano molto piccole può essere spiegato dal fatto che la sequenza genomica è ridotta rispetto alle altre, ma solo per *Adenomavirus*, *E. coli* e *S. cerevisiae*.
Il confronto derivato da ciascuna analisi suggerisce di accomunare, per le similarità riscontrate, un'intera classe di organismi: i mammiferi. Ciò porta a credere che i meccanismi che agiscono sulla distribuzione dei dinucleotidi, in particolare sul CG, siano gli stessi. Da qui, si deduce che il processo di metilazione, che coinvolge il dinucleotide CG, sarà simile per tutti questi organismi, e, data la vicinanza dei valori delle lunghezze caratteristiche, può significare che ciò porti a una struttura del DNA affine.
Il gruppo dei restanti organismi spazia in un ampio range di complessità e la grande eterogeneità dal punto di vista biologico si riflette nei risultati delle analisi. Si può dire che in questi organismi, non c'è un processo epigenetico analogo o comunque non svolge lo stesso ruolo che nei mammiferi, o agisce su porzioni differenti della sequenza. Tuttavia, per un'analisi più approfondita, sarebbe necessario lavorare su più campioni allo stesso grado di complessità per poter osservare meglio similarità tra le distribuzioni e trovare un riscontro con l'appartenenza alla stessa famiglia.
Uno studio di questo tipo risulta quindi produttivo nel classificare gli organismi in alberi filogenetici, ma non solo. Il fit esponenziale eseguito sulle distanze inter-CG ha sottolineato un ruolo strutturale dei dinucleotidi CG negli organismi di maggiore complessità, dal momento che questi "marcano" regolarmente l'intera sequenza di DNA.
Quest'evidenza rappresenta un esempio delle regole geometriche nascoste della struttura del DNA. Infatti, il fatto che l'andamento delle distribuzioni sia in generale a legge di potenza non è di meno rilevanza rispetto al caso particolare della distribuzione esponenziale del CG: la correlazione a legge di potenza è una misura dell'esistenza di uno scaling, che

mostra come ci siano strutture self-simili nella sequenza, analoghe alla geometria frattale. Una correlazione a lungo raggio di questo tipo può trovare giustificazione nel modello evolutivo del DNA, secondo il quale l'attuale sequenza risulta da un piccola catena originaria che ha poi attraversato processi di duplicazione e mutazione.
Un'altra osservazione degna di nota e e che meriterebbe uno studio più approfondito, è che le lunghezze caratteristiche trovate per il gruppo dei mammiferi sono comparabili alla lunghezza tipicamente associata agli *istoni*, proteine complesse che aiutano la modellizzazione della cromatina, e quindi potrebbero essere connesse all'organizzazione della struttura tridimensionale del DNA.

Conclusioni

Al termine di questa trattazione, possiamo affermare che l'utilizzo di modelli basati su processi stocastici, risulta di grande utilità nell'approccio alle sequenze di DNA.
Le catene di Markov sono un modello molto semplice in grado di riprodurre, già al primo ordine, molte delle caratteristiche delle distribuzioni dei nucleotidi lungo la sequenza, con un ottimo rapporto tra semplicità e fedeltà della ricostruzione. Ciò ha permesso, nel corso di numerosi studi che ne vedono l'applicazione già da diversi anni, di caratterizzare le sequenze genomiche e scoprirne informazioni sulla struttura. In questa sede, si può osservare come sfruttando catene di Markov di ordine maggiore di uno, quindi con sempre più "coscienza" degli stati precendenti, si otterrebbe un modello più fedele. Tuttavia, ciò potrebbe portare a una complessificazione della trattazione, che forse non è necessaria dal momento che la correlazione a corto raggio tra nucleotidi è ben riportata dal primo ordine.
Per quanto riguarda invece la correlazione a lungo raggio, la caratterizzazione della catena di DNA e lo studio delle inter-distanze in essa ha dato dei risultati che meritano di essere approfonditi, come ad esempio i valori delle lunghezze caratteristiche delle distanze inter-dinucleotide, comparabili alle dimensioni degli istoni, che hanno ruolo fondamentale nel compattamento della cromatina e quindi nella struttura tridimensionale del DNA. Più in generale, il fatto che la distribuzione dei nucleotidi lungo la sequenza segua leggi geometriche denota regole nascoste, periodicità, la cui interpretazione biologica implicherebbe la comprensione di molti aspetti dell'organizzazione strutturale del DNA.

Bibliografia

[1] N.G. Van Kampen, *Stochastic processes in physics and chemistry.* Elsevier, Amsterdam, 1981

[2] O. Knill, *Probability and stochastic processes with applications* Havard Web-Based, 1994.

[3] T. Konstantopoulos, *Markov Chains and Random Walks.* Lecture notes, 2009.

[4] R. Cogburn, *On the Distribution of First Passage and Return Times for Small Sets.* The Annals of Probability 13.4,1985.

[5] M. Hirata, B. Saussol, and S. Vaienti. *Statistics of Return Times: A General Framework and New Applications.* Communications in Mathematical Physics 206.1 , 1999, 33-55.

[6] L. J. S. Allen *An Introduction to Stochastic Processes with Applications to Biology.* Chapman and Hall/CRC, 2010

[7] W. Ewens, G. Grant. *Statisical Methods in Bioinformatics: An Introduction* . Springer, 2005

[8] Richard Durrett. *Probability : theory and examples.* Cambridge University Press, 2010.

[9] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, *Visualization and analysis of DNA sequences using DNA walks.* Journal of the Franklin Institute, vol. 341, no. 1-2, pp. 37-53, 2004.

[10] G. Paci, G. Cristadoro, B. Monti, M. Lenci, M. Degli Esposti, G. Castellani, D. Remondini. *Biological relevance of dinucleotide inter-distance distributions*. Royal Society Publishing, 2015.

[11] V. Afreixo, C. A. Bastos, A. J. Pinho, S. P. Garcia, and P. J. Ferreira. *Genome analysis with inter-nucleotide distances* Bioinformatics, 25(23) . 3064-3070, 2009.

[12] C. A. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. M. Rodrigues, P. J. Ferreira. *Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions.* J Integr Bioinform, 8(3):172, 2011.

[13] CK Peng. *Statistical properties of DNA sequences.* Physica A, 221:180-92, 1995.

[14] CK Peng, SV Buldyrev, and AL Goldberger. *Long-range correlations in nucleotide sequences.* Nature, 365:168-170, 1992.

[15] P. Allegrini, P. Grigolini, BJ West. *A dynamical approach to DNA sequences*. Physics Letters A, 211 : 217-222, 1996.

[16] C. Cattani, *Fractals and Hidden Symmetries in DNA*, Mathematical Problems in Engineering, vol. 2010, Article ID 507056, 2010.

[17] W. Li, *The study of correlation structures of DNA sequences: a critical review*, Computers and Chemistry, vol. 21, no. 4, pp. 257-271, 1997.