Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Corso di Laurea in Fisica

ANALISI E CONFRONTO DI SEQUENZE DI DNA MEDIANTE MODELLI MARKOVIANI

Relatore:
Prof. Daniel Remondini

Presentata da:
Maria Francesca Morrone

Sessione II
Anno Accademico 2014/2015

“ Un esperto è un uomo
che ha fatto tutti gli errori
che sia possibile compiere
in un campo molto ristretto.”
Niels Bohr

Alla mia famiglia e
ai miei amici di facoltà

Abstract

Lo scopo di questa tesi è quello di evidenziare, attraverso varie analisi statistiche ed applicazione di modelli stocastici, il comportamento strutturale e funzionale dei dinucleotidi che compongono le sequenze di DNA di diversi organismi. Gli organismi che abbiamo scelto di prendere in considerazione sono l'uomo, il topo e l'Escherichia coli. Questa scelta non è stata casuale, ma oculata, al fine di mettere in risalto alcune differenze tra organismi eucarioti, quali l'uomo e il topo, ed organismi procarioti come il batterio E.coli. Nella prima parte del nostro studio, abbiamo computato le distanze che intercorrono tra occorrenze successive dello stesso dinucleotide lungo la sequenza, usando un metodo di non sovrapposizione, ed abbiamo iterato il calcolo per tutti i 16 dinucleotidi. Dopodiché ci siamo preoccupati di graficare le distribuzioni di distanza dei 16 dinucleotidi per l'E.Coli, il topo e l'uomo; gli istogrammi evidenziano un comportamento anomalo della distribuzione di CG che accomuna gli organismi eucarioti e di cui, invece, è esente l'organismo procariote esaminato. Questo dato statistico trova una spiegazione nei processi biologici di metilazione che possono innescarsi sul dinucleotide CG nelle sequenze eucariotiche. In seguito, per determinare quanto ciascuna delle 16 distribuzioni si discosti dalle altre abbiamo usato la divergenza di Jensen-Shannon. Per quantificare le differenze sostanziali tra le distribuzioni di CG dei 3 organismi considerati abbiamo deciso di verificare quale fosse il miglior fit per tali curve tra un esponenziale ed una power-law. L'esponenziale rappresenta un buon fit per le code delle distribuzioni di CG del topo e dell'uomo; ciò rivela la presenza di una lunghezza caratteristica per entrambi gli organismi.

Nella seconda parte dello studio, i risultati vengono confrontati con modelli markoviani: sequenze random generate con catene di Markov di ordine zero (basate sulle frequenze relative dei nucleotidi) e uno (basate sulle probabilità di transizione tra diversi nucleotidi). Quest'ultima riproduce abbastanza fedelmente la sequenza biologica di partenza, per cui abbiamo scelto di utilizzare la catena Markov del 1° ordine per altre analisi statistiche riguardanti le distribuzioni dei nucleotidi, dinucleotidi, ed anche dei trinucleotidi con particolare interesse per quelli in cui è contenuto CG, in modo da verificare se l'anomalia si ripercuote anche in essi.

Riteniamo pertanto che metodi basati su questo approccio potrebbero essere sfruttati per confermare le peculiarità biologiche e per migliorare l'individuazione delle aree di interesse, come le isole CpG, ed eventualmente promotori e Lamina Associated Domains (LAD), nel genoma di diversi organismi.

Indice

Introduzione

In questa tesi noi affrontiamo un problema di ambito biologico con un approccio fisico, e precisamente probabilistico/statistico, che prevede l'applicazione di modelli stocastici, utilizzati come termine di paragone.

Nel primo capitolo ci preoccupiamo di descrivere il background biologico necessario per addentrarsi nelle successive analisi statistiche. Le nozioni biologiche che presentiamo rigurdano la descrizione della struttura del DNA con brevi accenni alla composizione chimica, la spiegazione dei meccanismi di metilazione e delle isole CpG, la cui definizione formale è stata data da Gardner, Gardiner e Frommer nel 1987. Alla fine del suddetto capitolo presentiamo brevemente gli organismi oggetto del nostro studio, con le loro principali caratteristiche genomiche.

Nel secondo capitolo descriviamo nel dettaglio il metodo statistico impiegato, quindi spieghiamo dal punto di vista teorico il metodo di non sovrapposizione con il quale abbiamo letto le sequenze di DNA, il calcolo delle interdistanze e delle relative distribuzioni, la divergenza di Jensen-Shannon ed i modelli stocastici markoviani, con particolare interesse per le catene di Markov di ordine 0 e 1.

Nel terzo capitolo, riportiamo i risultati ottenuti dall'impiego dei metodi sopra elencati, attraverso grafici, come istogrammi ed heatmaps, matrici e tabelle che contengono i dati numerici.

Nel quarto capitolo, traiamo le conclusioni che sono state dedotte da questa trattazione e gli eventuali sviluppi futuri che potrebbero migliorare la veridicità delle nostre riflessioni; in particolare suggeriamo di impiegare i medesimi principi e modelli statistici su altri organismi con lo scopo di scoprire ulteriori differenze che dividono il mondo procariote da quello eucariote.

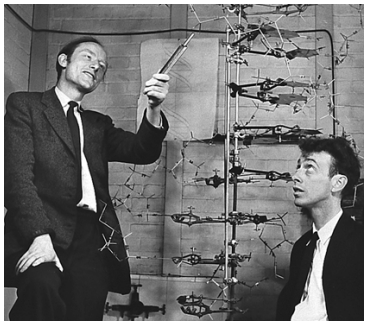
Capitolo 1

DNA: struttura e funzione nei diversi organismi

Dal momento che questa tesi si basa sull'applicazione di metodi ed analisi statistiche su sequenze di DNA di organismi procarioti, come l'Escherichia Coli, ed eucarioti, quali sono il topo e l'uomo, riteniamo opportuno fornire qualche spiegazione preliminare di ambito biologico riguardante la struttura e funzionalità di un genoma, la differenza fra organismi procarioti ed eucarioti e le principali caratteristiche genomiche degli organismi scelti come oggetto di questa trattazione statistica.

1.1 La struttura del DNA : un modello di atomi di cartone e legami di filo di ferro

La curiosità di indentificare la composizione chimica del nucleo cellulare spinse il biologo svizzero Johann Friedrich Miescher verso la metà dell' 800 a raccogliere cellule dal pus delle ferite aperte; fortuitamente queste particolari cellule hanno pochissimo citoplasma e, pertanto, permettono di facilitare l'analisi e la separazione del materiale contenuto nel nucleo. Nel 1868 egli riuscì ad isolare un composto organico dotato delle proprietà di un acido contenente una significativa quantità di fosforo e lo chiamò nucleina, totalmente ignaro di aver scoperto la sostanza che parecchi anni dopo sarebbe divenuta nota come acido desossiribonucleico, o in forma abbreviata, DNA. La scoperta di Miescher, paradossalmente, non destò molto interesse nella comunità scientifica ma stuzzicò l'arguzia di due noti scienziati, Watson e Crick, che raccogliendo le informazioni relative alle dimensioni e ai tipi di legami delle subunità del DNA ritagliarono nel cartoncino sagome di queste subunità e le unirono tramite fili di ferro, che costituivano i legami tra le subunità. Questo modello costruito da Watson e Crick nel 1953 riproduceva abbastanza fedelmente la struttura odierna del DNA.



Oggi, infatti, sappiamo che la molecola di DNA è composta da quattro tipi di subunità, dette nucleotidi. Ogni nucleotide consiste di uno zucchero a cinque atomi di carbonio (desossiribosio nel DNA e ribosio nel RNA), di un gruppo fosfato e di una delle seguenti basi azotate : adenina (A) , citosina (C) , guanina (G) , timina (T). Le prime due sono purine, ovvero basi azotate con struttura a doppio anello, mentre le seconde sono pirimidine, basi azotate più piccole con struttura ad anello singolo. Tra le basi pirimidiniche troviamo anche l'uracile, che è la base azotata presente nel RNA al posto della timina.

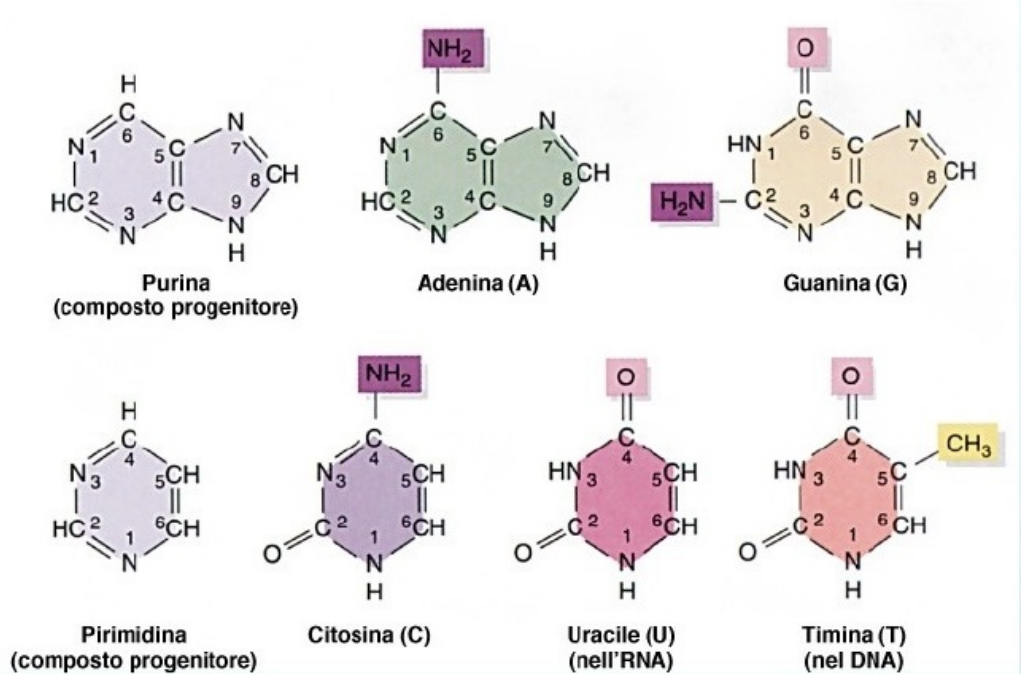


Figura 1.1: Struttura chimica delle basi azotate

Grazie all'intuitivo modello creato da Watson e Crick, possiamo affermare che la molecola di DNA consiste di due filamenti di nucleotidi avvolti l'uno sull'altro a formare una doppia elica, con le basi azotate rivolte verso l'interno della molecola. I due filamenti sono costituiti da catene in cui si alternano unità di zucchero (desossiribosio) e gruppi fosfato e sono orientati in direzioni opposte: uno in direzione $5' \longrightarrow 3'$ e l'altro in direzione $3' \longrightarrow 5'$ (i numeri sono quelli identificativi dell'atomo di carbonio).

Le basi azotate di un filamento sono connesse a quelle dell'altro filamento mediante legami idrogeno. Per tutta la lunghezza della molecola di DNA, l'adenina si appaia sempre alla timina e la citosina sempre alla guanina, secondo lo schema : A-T e G-C.

La struttura a doppia elica del DNA spiega il processo di duplicazione della molecola dell'ereditarietà che realizza una propria replica prima che la cellula si divida. Gli enzimi riescono facilmente a rompere i legami idrogeno tra i due filamenti nucleotidici del DNA; quando questi enzimi agiscono sulla molecola, un filamento si può svolgere dall'altro, lasciando così esposto un tratto di basi nucleotidiche. Le cellule contengono una riserva di nucleotidi liberi che possono essere appaiati con queste basi. Non appena un nuovo tratto di filamento si forma su quello parentale, i due filamenti si avvolgono in forma di doppia elica, ogni "nuova" molecola di DNA è quindi, in realtà, per metà vecchia

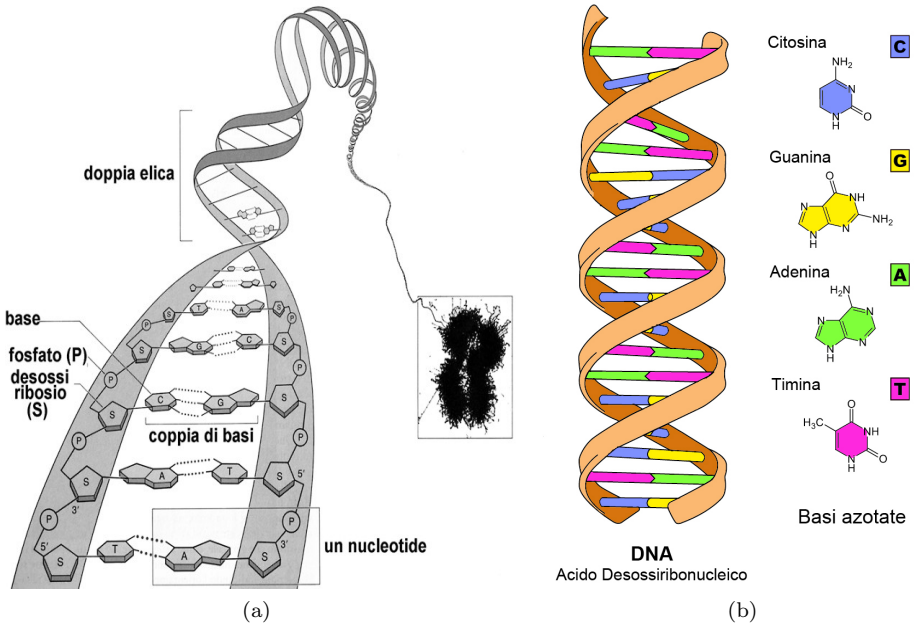


Figura 1.2: Struttura a doppia elica del DNA delle basi azotate

(in quanto contiene un filamento parentale) e per metà nuova. Gli enzimi chiamati DNA polimerasi possono agganciare brevi porzioni di nucleotidi liberi (nel citoplama) a segmenti srotolati del filamento parentale, che funge da "stampo". Gli enzimi DNA ligasi riempiono i piccoli vuoti tra i nuovi segmenti brevi per ottenere un filamento continuo; quindi gli enzimi avvolgono il filamento stampo al suo complementare per formare la doppia elica del DNA.

1.2 Il DNA nella cellula eucariote e procariote

La principale differenza tra organismi eucarioti e procarioti è che i primi presentano un nucleo interno ben definito e isolato dal resto della cellula tramite la membrana nucleare, nel quale è racchiusa la maggior parte del materiale genetico rappresentato dal DNA (una parte minore è contenuta nei mitocondri); i secondi invece hanno il DNA sparso nel citoplasma. Gli eucarioti si distinguono dai procarioti anche per numerose caratteristiche a livello molecolare. Ad esempio, gli eucarioti hanno:

- diverse proprietà delle sequenze genomiche regolatrici;
- geni organizzati in introni ed esoni con conseguente processamento (splicing) del trascritto primario;
- trascrizione e traduzione di un trascritto sono eventi separati nello spazio e nel tempo;
- i trascritti eucariotici non sono (quasi) mai policistronici, ossia portano una sola ORF;
- percentuale di DNA non codificante molto più elevata;

- DNA associato a istoni;
- diversa percentuale di G-C nel genoma;
- presenza di colesterolo nella membrana cellulare, tranne nei funghi, nelle piante e in alcuni protisti che, pur essendo eucarioti, non presentano colesterolo nella membrana.

Le nostre analisi statistiche forniscono un'evidenza sperimentale del penultimo punto di questo elenco, come spiegheremo nel capitolo successivo, e come si può notare nelle figure che inseriremo nel capitolo 3.

1.3 La metilazione del DNA

La metilazione del DNA è una modificazione epigenetica del DNA. Il processo consiste nel legame di un gruppo metile (-CH3) ad una base azotata. Differenti basi azotate possono subire questo tipo di modificazione per diverse funzioni. L'adenina può essere metilata a livello delle sequenze GmeATC del genoma di alcuni batteri come Escherichia coli dall'enzima dam (DNA adenina metilasi). La funzione di questa metilazione è quella di proteggere il genoma della cellula dall'attacco delle endonucleasi di restrizione da lei stessa prodotte, come mezzo di resistenza all'attacco di fagi.

Un altro esempio di metilazione del DNA è la metilazione della citosina, che costituisce la quasi totalità della metilazione di DNA eucariotico. Nei mammiferi, la 5-metil-citosina (5meC o 5mC) si trova quasi esclusivamente nel dinucleotide CpG (citosina seguita da una guanina) nella regione codificante dei geni. Le isole CpG sono regioni genomiche che contengono un'elevata densità di siti CpG. Il dinucleotide CpG è poco rappresentato nel DNA degli eucarioti poiché soggetto a mutazione secondaria a metilazione. Il genoma dei mammiferi è quasi del tutto metilato, a eccezione di alcune zone ricche del dinucleotide CpG che per questo vengono chiamate isole CpG, solitamente abbondanti in regioni regolative e promotori dei geni eucariotici; la metilazione di queste sequenze aumenta in particolari patologie come la sindrome di Prader-Willi. La metilazione fisiologica del DNA in queste regioni dipende da proteine chiamate DNA-metil-transferasi o DNMTs ed è un fenomeno che interviene nel controllo dell'espressione genica, nell'inattivazione del cromosoma X, nella struttura cromatinica e nell'imprinting genomico. La metilazione della citosina è anche un importante fattore di mutazione. Infatti, mentre la deaminazione della citosina produce uracile - una base azotata che non appartiene al DNA (bensì all'RNA) ed è immediatamente riconosciuta come estranea- la deaminazione della 5-metil citosina (5meC) la trasforma in timina, generando un mismatch, nel quale il sistema di riparazione dei mismatch (mismatch repair) non sempre preserva la corretta base azotata. Si sospetta che sia implicata anche nel taglio degli introni, e nella modifica degli esoni.

1.3.1 La metilazione nei mammiferi: CpG island

La metilazione del DNA è essenziale per il normale sviluppo ed è associata con alcuni processi chiave, tra cui l'imprinting genomico, l'inattivazione del cromosoma X, la soppressione di elementi ripetitivi e la carcinogenesi. Tale metilazione è una modificazione biochimica e coinvolge l'aggiunta di un gruppo metile al livello del carbonio-5 della

citosina, quasi esclusivamente nel contesto del dinucleotide CpG. La connessione tra metilazione del DNA e trascrizione genica è stata intensamente studiata negli ultimi 40 anni; il modello che attualmente gode di consenso trasversale nella comunità scientifica vede la metilazione del DNA associata a repressione della trascrizione. In particolar modo, la metilazione di un promotore a monte di un gene causerebbe la sua repressione, che a sua volta può essere invertita o alleviata tramite demetilazione della stessa sequenza. In altre parole, i diversi pattern di metilazione, dunque, regolano l'accensione e lo spegnimento di alcuni geni. È chiaro che un errore nella metilazione del DNA determina un cambiamento nell'organizzazione spaziale della cromatina e ciò determina, a sua volta, delle "stonature"; situazioni di ipo- o iper-metilazione, possono portare rispettivamente all'accensione o spegnimento di geni che agiscono come oncosoppressori o nei meccanismi di riparo del DNA. Epimutazioni di questo tipo sono state identificate in molti tipi di tumori. La metilazione del DNA a livello di isole CpG è associata con repressione genica. Tali isole CpG, sono preferenzialmente localizzate al promotore dei molti geni, in particolar modo di geni house-keeping e in alcuni geni tessuto-specifici. La 5-metil-citosina, è più instabile, più soggetta a mutazioni e tende a deaminare rispetto alla citosina non modificata. La deaminazione della citosina causa la formazione di uracile (una base azotata normalmente non presente nel DNA ma nell'RNA), mentre la deaminazione della 5mC porta alla timina che non è più appaiata alla guanina dell'altro filamento. Quest'ultima è più frequente della prima anche a causa dell'efficienza dei meccanismi di riparo nel riconoscere i due tipi di mutazione. Una volta stabilito lo schema di metilazione, questo viene mantenuto ad opera di DNA-metil-transferasi di mantenimento, le quali hanno una particolare affinità per le sequenze emi-metilate, e tendono dunque a metilare il nuovo filamento formatosi su uno stampo già metilato. Nelle cellule dei mammiferi, esistono inoltre, degli elementi cis-agenti che sono dispersi in tutto il genoma e fungono da elemento segnale o limite per la propagazione della metilazione.

1.3.2 DNA metiltransferasi

Nelle cellule di mammifero, la metilazione del DNA si verifica principalmente nella posizione C5 della citosina nel contesto di dinucleotidi CpG ed è effettuata da una famiglia di enzimi chiamata DNMT (DNA-metil-transferasi). In particolar modo DNMT1 è responsabile del mantenimento del pattern di metilazione sui filamenti sintetizzati ad ogni ciclo di replicazione. In assenza di DNMT1, l'apparato di replicazione produce filamenti di DNA che non vengono metilati causando una "diluizione" della metilazione e di conseguenza un processo di demetilazione "passiva" (che cioè non comporta la rimozione attiva del gruppo metile dalla citosina già metilata). Si pensa che DNMT3a e DNMT3b siano le metiltransferasi coinvolte nella metilazione de novo. In aggiunta ad esse, DNMT3L è una proteina che non ha attività catalitica ma si pensa abbia attività regolatrice delle metiltransferasi de novo, aumentando la loro capacità di legarsi al DNA e stimolando la loro attività. Infine, DNMT2 (TRDMT1) è stato identificato come un omologo delle DNA metiltransferasi, e contiene tutti i 10 motivi di sequenza comuni a tutte le DNA metiltransferasi; tuttavia, DNMT2 (TRDMT1) non metila il DNA, ma metila citosina-38 nel loop dell'anticodone dell'acido aspartico tRNA.

1.3.3 Metilazione del DNA nel cancro

La metilazione del DNA è un importante regolatore della trascrizione genica ed è stato dimostrato che la metilazione anormale del DNA è associata al silenziamento genico non programmato, ed i geni con alti livelli di 5-metilcitosina nella regione del promotore, sono trascrizionalmente silenziati. La metilazione del DNA è essenziale durante lo sviluppo embrionale, e nelle cellule somatiche, i pattern di metilazione del DNA sono generalmente trasmessi alle cellule figlie con un'alta fedeltà. I pattern anormali di metilazione del DNA sono stati associati ad un gran numero di neoplasie umane e si trovano in due forme distinte: ipermetilazione e ipometilazione rispetto al tessuto normale. L'ipermetilazione è una delle principali modifiche epigenetiche che reprimono la trascrizione attraverso la regione promotrice di un gene oncosoppressore. L'ipermetilazione si verifica in genere nelle isole CpG, nella regione del promotore ed è associata con l'inattivazione genica. L'ipometilazione globale è stata anche coinvolta nello sviluppo e nella progressione del cancro attraverso meccanismi diversi.

1.4 Tre diversi organismi a confronto : il batterio delle feci, il topo e l'uomo

1.4.1 Escherichia Coli

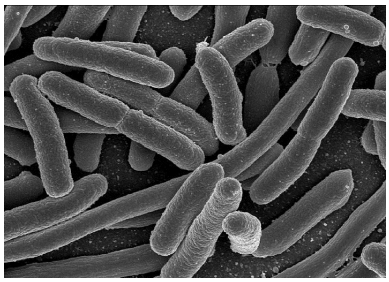


Figura 1.3: Il batterio Escherichia Coli.

Escherichia coli è un batterio Gram-negativo ed è la specie più nota del genere Escherichia: se ne distinguono almeno 171 sierotipi, ognuno con una diversa combinazione degli antigeni O, H, K, F. Il nome deriva dal suo scopritore, il tedesco-austriaco Theodor Escherich. Appartiene al gruppo degli enterobatteri ed è usato come organismo modello dei batteri. È una delle specie principali di batteri che vivono nella parte inferiore dell'intestino di animali a sangue caldo (uccelli e mammiferi, incluso l'uomo). Sono necessari per la digestione corretta del cibo. La sua presenza nei corpi idrici segnala la presenza di condizioni di fecalizzazione (è il principale indicatore di contaminazione fecale, insieme con gli enterococchi). Il numero di cellule di E. coli nelle feci che un umano espelle in un giorno va dai 100 miliardi ai 10 trilioni. Il genere Escherichia, insieme ad altri generi (Enterobacter, Klebsiella, Citrobacter, Serratia, ecc.), viene raggruppato sotto il nome di coliformi. Tecnicamente il "gruppo dei coliformi" comprende batteri aerobi e anaerobi non sporigeni.

Nel gruppo dei coliformi la specie Escherichia coli è ampiamente rappresentata ed è in esclusivo rapporto col tratto gastrointestinale dell'uomo e degli altri animali a sangue caldo, a differenza dei microrganismi appartenenti a diversi generi, tra cui Enterobacter, Klebsiella e Citrobacter (che si caratterizzano per una potenziale capacità di ricrescita una volta pervenuti nell'ambiente). La specie Escherichia coli è un microrganismo a forma di bastoncello, gram-negativo, aerobio e anaerobio facoltativo, non sporigeno, che cresce alla temperatura di $44{,}5°C$, lattosio-fermentante. Anche se rappresenta un comune

abitante dell'intestino e ha un ruolo nel processo digestivo, ci sono situazioni in cui E. coli può provocare malattie nell'uomo e negli animali. Alcuni ceppi di E. coli sono l'agente eziologico di malattie intestinali ed extra-intestinali, come infezioni del tratto urinario, meningite, peritonite, setticemia e polmonite. Alcuni ceppi di E. coli sono tossigenici, producono cioè tossine che possono essere causa di diarrea. La dissenteria da E. coli è una comune tossinfezione alimentare, poiché viene contratta principalmente da alimenti contaminati. La contaminazione può avvenire da carni infette non adeguatamente cotte, da latte non pastorizzato e formaggi derivati, e da altri alimenti contaminati da feci.

1.4.2 Mus Musculus

Il Mus musculus, ovvero il topo comune, è un piccolo mammifero roditore della famiglia dei Muridi. Il genoma del topo venne sequenziato completamente verso la fine del 2002, ha una parte aploide che misura circa 3000 megabasi (più o meno come quella umana) ed è distribuita su 20 cromosomi. Tuttavia, è difficile fare una conta attendibile dei geni contenuti nel genoma del topo: una stima recente, accettata dalla maggior parte di studiosi, parla di 23786 geni, contro i 23686 dell'uomo. Questi numeri ci suggeriscono che, virtualmente, ciascun gene di topo trova un omologo nel genoma umano, il ché permette di effettuare esperimenti su di essi con il fine di ricavare informazioni, per via indiretta, anche sull'uomo.

1.4.3 Homo sapiens

Il genoma umano, cioè il genoma dell'Homo sapiens, è il termine con il quale ci si riferisce al DNA nucleare e non comprende il DNA mitocondriale. Ha un corredo di circa 3,2 miliardi di paia di basi di DNA contenenti all'incirca 20000-25000 geni. Il Progetto Genoma Umano ha scoperto più DNA non codificante di quanto previsto , ben il 98,5%, con solo circa l'1,5% della lunghezza totale che si basa su esoni codificanti proteine.

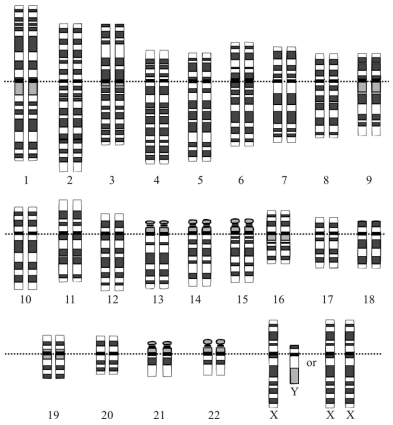


Figura 1.4: Rappresentazione grafica del cariotipo umano femminile e maschile normale.

Il DNA nucleare umano si raggruppa in 24 tipologie di cromosomi : 22 autosomi , più due cromosomi che determinano il sesso: cromosoma X e cromosoma Y. I cromosomi 1-22 sono numerati in ordine di lunghezza decrescente. Le cellule somatiche hanno due copie dei cromosomi 1-22 provenienti ognuna da un genitore, più un cromosoma X della madre e un cromosoma X o Y del padre, per un totale di 46 cromosomi. È stata ipotizzata l'esistenza di 20000 - 25000 geni codificanti proteine, ma il numero preciso non si può determinare con certezza. Oltre ai geni codificanti proteine, il genoma umano contiene diverse migliaia di geni codificanti un RNA incluso il tRNA, l'RNA ribosomico, microRNA. Tuttavia, è necessario specificare che i geni codificanti proteine (specificatamente, codificanti esoni) costituiscono meno dell'1,5 % dell'intero genoma umano.

Capitolo 2

Metodi statistici per l'analisi delle sequenze di DNA

In questo capitolo entriamo nel cuore dell'analisi statistica attuata su tre diversi organismi: uno procariote, il batterio Escherichia Coli, ed altri due eucarioti, il topo e l'uomo. L'aspetto interessante di questo studio è che ci condurrà a dei risultati rilevanti dal punto di vista biologico e che trovano una loro spiegazione nei processi di metilazione che si innescano in un genoma di un organismo eucariote, e che, come si vedrà, invece sono assenti nei procarioti. Per mettere in risalto le ragioni che ci hanno portato a compiere questo studio, forniamo in questo capitolo una descrizione del motodo statistico utilizzato, portando a conoscenza il lettore delle variabili più rilevanti impiegate nel corso dell'analisi.

2.1 La lettura delle sequenze ed il calcolo delle interdistanze tra dinucleotidi

Per prima cosa abbiamo letto le sequenze di DNA da un file in formato FASTA contenente una riga d'intestazione che inizia con il seguente simbolo $>$ e, dopo un'andata a capo, segue la sequenza stessa; in genere, la prima linea d'intestione è esplicativa di ciò che è contenuto nel file FASTA, essa contiene il nome dell'organismo da cui è stata prelevata la sequenza e specifica il particolare cromosoma, qualora il genoma non fosse completo. Nel nostro caso, la sequenza di DNA dell'Escherichia Coli è completa, cioè corrisponde all'intero genoma del batterio; mentre per quanto riguarda gli altri due organismi eucarioti scelti, il topo e l'uomo, abbiamo analizzato la sequenza di DNA del cromosoma 1 dell'intero corredo genetico contenuto nel nucleo. Per dare un'idea della differenza quantitativa di corredo genetico tra un batterio ed un mammifero, basta pensare che la sequenza del solo cromosoma 1 del topo è 40 volte più lunga di quella dell'intero genoma di E.coli, mentre quella dell'uomo lo è 60 volte. La sequenza è composta dalla successione di nucleotidi A, C, G, T, ma in alcuni tratti di cromosoma sono state riscontrate sostanze sconosciute e sono state contrassegnate con la lettera N; quest'ultima è ignorata durante la lettura della sequenza e, pertanto, non viene conteggiata. Abbiamo letto le sequenze di DNA

e calcolato, attaraverso un'apposita funzione scritta in codice Matlab, le interdistanze che intercorrono tra un dinucleotide scelto e il successivo, iterando il calcolo su tutti i 16 dinucleotidi. Per fare ciò, abbiamo impiegato un approccio di non sovrapposizione, secondo il quale la distanza minima tra dinucleotidi adiacenti è 1: infatti, se in un tratto della sequenza troviamo una successione di nucleotidi del tipo CCCC, questa verrebbe interpretata come composta da 2 dinucleotidi CC alla distanza di 2; per evitare ciò, alla fine del processo di conteggio delle posizioni che intercorrono tra 2 medesimi dinucleotidi, sottraiamo 1 ad ogni distanza calcolata. Con il metodo di sovrapposizione, invece, una successione CCCC viene letta come composta da 2 dinucleotidi CC alla distanza 0.

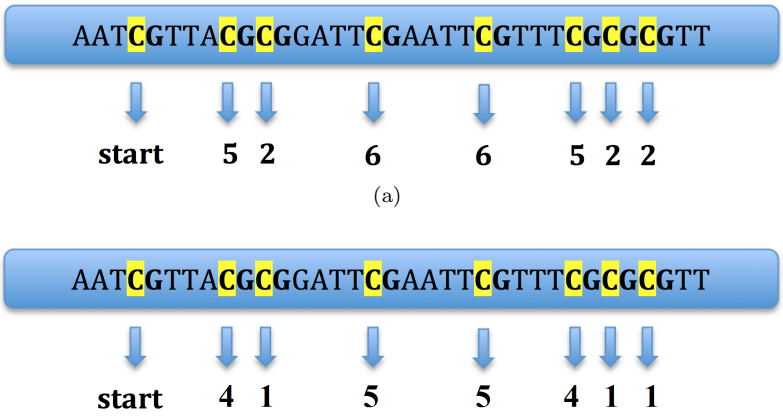


Figura 2.1: Lettura della sequenza conteggiando le interdistanze tra dinucleotidi CG (a) e sottraendo 1 usando il metodo di non sovrapposizione (b).

La lunghezza di ciascuno dei 16 vettori distanza dipende, ovviamente, dalla lunghezza della sequenza stessa, per dare un'idea di quanti dati sono stati raccolti, riportiamo qualche numero:

-nel caso dell'E.Coli, la cui sequenza completa di DNA è lunga 4639675 bp , abbiamo ottenuto dei vettori distanza contenenti mediamente 230000- 330000 elementi ;

- nel caso del genoma del topo, il cui cromosoma 1 è lungo 195471971 bp, i vettori distanza hanno circa 7000000 - 14000000 elementi;

- nel caso del genoma umano, il cromosoma 1 è di 224384768 bp, ed i 16 vettori distanza calcolati contengono circa 10000000 - 22000000 elementi.

L'unità di misura utilizzata per la lunghezza fisica di sequenze ad acidi nucleici a doppio filamento è il base pair (coppie di basi), abbreviato come bp o bps. Diamo qui di seguito una lista delle abbreviazioni comunemente usate per la lunghezza di una molecola di D/RNA :

-bp (=base pair), un bp corrisponde a circa 3,4 Å(340 pm) di lunghezza lungo il filo;

-kb (=kbp) = kilo base pair= 1000 bp ;

-Mb (=Mbp) = mega base pair = 1000000 bp;

-Gb = giga base pair = 1000000000 bp.

Una volta calcolati i vettori distanza, abbiamo costruito gli istogrammi con un massimo di 50 bin in cui grafichiamo la distribuzione di distanze, dopodiché ne abbiamo studiato l'andamento.

2.2 La divergenza di Jensen-Shannon

La divergenza di Jensen-Shannon D_{JS} è un metodo che si usa in statistica ed in teoria della probabilità per misurare la vicinanza tra due distribuzioni di probabilità P e Q, oppure per verificare quanto bene l'una approssima l'altra. Essa è una versione simmetrizzata della divergenza di Kullbach-Leibler D_{KL}, anche detta divergenza di informazione, entropia relativa, o KLIC), e che è indicata con $D_{\mathrm{KL}}(P||Q)$, fornisce la misura dell'informazione persa quando Q è usata per approssimare P. Date due distribuzioni di probabilità, la divergenza di Kullbach-Leibler è definita come:

$$D_{\mathrm{KL}}(P||Q) = \sum_i \log \frac{P_{\mathrm{i}}}{Q_{\mathrm{i}}} P_{\mathrm{i}} \tag{2.1}$$

con le seguenti proprietà:

- positività $D_{\mathrm{KL}}(P||Q) > 0$;
- asimmetria $D_{\mathrm{KL}}(P||Q) \neq D_{\mathrm{KL}}(Q||P)$;
- $D_{\mathrm{KL}}(P_{\mathrm{i}}||Q_{\mathrm{i}}) = 0$ se e solo se $P_{\mathrm{i}} = Q_{\mathrm{i}}$.

La divergenza di Jensen-Shannon, come abbiamo già detto, è la versione simmetrizzata di quella di Kullbach-Leibler, ed è definita come:

$$D_{\mathrm{JS}}(P||Q) = \frac{1}{2} D(P|M) + \frac{1}{2} D(Q|M) \tag{2.2}$$

in cui P e Q sono due distribuzioni discrete:
$P = p_{\mathrm{i}}$, i= 1, ... N e $Q = q_{\mathrm{i}}$ i= 1, ... N tali che $\sum_i p_{\mathrm{i}} = 1$ e $\sum_i q_{\mathrm{i}} = 1$;
M è la distribuzione media di P e Q.

$$M = (P + Q)/2, m_{\mathrm{i}} = (p_{\mathrm{i}} + q_{\mathrm{i}})/2 \tag{2.3}$$

Le proprietà della divergenza di Jensen-Shannon richiamano quelle della Kullbach-Leibler con l'unica differenza che in questo caso vale la simmetria:

- positività $D_{\mathrm{JS}}(P||Q) > 0$,
- simmetria $D_{\mathrm{JS}}(P_{\mathrm{i}}||Q_{\mathrm{i}}) = D_{\mathrm{JS}}(Q_{\mathrm{i}}||P_{\mathrm{i}})$,
- $D_{\mathrm{JS}}(P_{\mathrm{i}}||Q_{\mathrm{i}}) = 0$ se e solo se $P_{\mathrm{i}} = Q_{\mathrm{i}}$.

La Jensen-Shannon ci permette di confrontare la frequenza relativa con cui compare nella sequenza un dinucleotide scelto con ciascuna frequenza relativa dei restanti quindici e vedere quanto le due frequenze si discostino tra loro. In seguito le matrici di divergenza per i tre organismi sono state rappresentate con le relative heatmaps. Quest'ultime forniscono una rappresentazione grafica ottimale che associa ad ogni singolo elemento della matrice un colore, secondo una scala opportuna.

2.3 La distribuzione di distanze nei tre diversi organismi

Dopo aver calcolato i vettori distanza di ciascuno dei 16 dinucleotidi XY con $XY \in \{A, C, G, T\}$, ci serviamo di una particolare funzione scritta in codice Matlab attarverso la quale :

- generiamo una serie di indici τ_{j} con valori compresi tra 1 e 350- 370 bp (base pair = coppia di basi) e che saranno le interdistanze tra dinucleotidi.

- contiamo quante volte ciscuna delle 16 interdistanze τ_{j} si trova nei vettori distanza attraverso la nota funzione histc(), questo conteggio ci permette di valutare quali sono le interdistanze tra dinucleotidi più frequenti per ogni singolo dinucleotide. In altre parole, ciascun $p(\tau_{\mathrm{j}})$ ci dice quanto vale la probabilità che due medesimi dinucleotidi si trovino ad una distanza τ_{j} nella sequenza.

In definitiva otteniamo una densità di probabilità discreta che ci pare opportuno, per una migliore rappresentazione grafica, normalizzare. La condizione di normalizzazione che abbiamo imposto tiene conto del fatto che i vari bin non hanno tutti la stessa larghezza, per cui per normalizzare ogni singolo τ_{j} si tiene conto del binsize $s_{\mathrm{j}} = \tau_{\mathrm{j}+1} - \tau_{\mathrm{j}}$. La distribuzione di distanze $p(\tau)$ avrà pertanto le seguenti proprietà:

- $0 \leq p(\tau_{\mathrm{j}}) < 1$;
- $\sum_{i}^{N} p(\tau_{\mathrm{j}}) = 1$ con N che va da 1 fino ad un massimo di 50, infatti abbiamo scelto una soglia di 50 bin per evitare valori nulli della distribuzione .

2.4 Il fit delle code delle distribuzioni

Infine, abbiamo studiato quale è la curva che fitta meglio le code delle distribuzioni di distanza ed abbiamo dedotto dalle figure in scala logaritmica su uno degli assi o su entrambi gli assi che le code delle distribuzioni hanno un comportamento asintoticamente algebrico o esponenziale. Di conseguenza, abbiamo esaminato quale curva fitta meglio le code tra una power-law ed un esponenziale.

2.5 I modelli markoviani

Un processo di Markov è un processo stocastico nel quale la probabilità di transizione che determina il passaggio ad uno stato di sistema dipende unicamente dallo stato di sistema immediatamente precedente (proprietà di Markov) e non dal come si è giunti a tale stato (in quest'ultima ipotesi si parla di processo non markoviano). Tale processo prende il nome dal matematico russo Andrej Andreevič Markov che per primo ne sviluppò la teoria. Modelli di tipo markoviano vengono utilizzati nel progetto di reti di telecomunicazioni; la teoria delle code che ne consegue trova applicazione in molti ambiti: dalla fila alle poste ai pacchetti in coda in un router.

Il processo markoviano è stato applicato, nel nostro caso, per generare sequenze random di DNA. Infatti, in qualsiasi studio che coinvolge genomi di diversi organismi è

necessario un modello adatto per la generazione di sequenze casuali di DNA con lo scopo di confrontare e convalidare qualsiasi risultato ricavato dalla sequenza originale con ciò che si otterebbe se la sequenza dovesse essere generata casualmente. Naturalmente, esistono molte possibili definizioni di sequenza casuale e diversi modelli (dai più semplici a quelli più complicati): è quindi necessario considerare attentamente il metodo migliore per ogni studio, cercando un buon compromesso tra complessità ed attendibilità. Uno dei modelli casuali più diffuso per sequenze di DNA si basa sulla generazione di catene di Markov.

Una sequenza $\{X_n\}$ di variabili random discrete è detta catena di Markov se soddisfa la proprietà di Markov:

per tutti gli $n \geq 1$ e $(x_1, x_2, \ldots x_n)$ vale che

$$P(X_{\mathrm{n}+1} = x_{\mathrm{n}+1} \mid X_1 = x_1, \ldots, X_{\mathrm{n}} = x_{\mathrm{n}}) = P(X_{\mathrm{n}+1} = x_{\mathrm{n}+1} \mid X_{\mathrm{n}} = x_{\mathrm{n}}). \tag{2.4}$$

La proprietà di Markov espressa in (2.4) afferma che la probabilità condizionata allo stato n+1, data dal valore X_n allo step n, è univocamente determinata ed è indipendente dalla conoscenza dei valori di probabilità degli stati precedenti. Una catena di Markov è definata da :

- uno spazio degli stati discreto $S = \{x_1, x_2, \ldots x_n\}, \forall n$;
- uno stato iniziale X_0 ;
- le probabilità di transizione $P(X_{n+1} = j \mid X_n = i)$ con i,j ∈ S.

Se per tutti gli n e i , j ∈ S si ha $P(X_{n+1} = j \mid X_n = i) = p_{ij}$ indipendente da n, la catena è detta omogenea ed abbiamo la seguente matrice delle probabilità di transizione:

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{12} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{1m} & p_{2m} & \cdots & p_{mm} \end{bmatrix}$$

Le probabilità di transizione determinano interamente il comportamento della catena: la loro conoscenza, insieme a quella dello stato iniziale, è sufficiente per generare l'intera sequenza. Questo concetto può essere generalizzato a catene di Markov di ordine m, dove ciascun stato n-esimo dipende dagli stati m immediatamente precedenti e non da altri stati. Le catene di Markov sono state ampiamente impiegate nel contesto dell'analisi di sequenze genomiche in quanto esse sono in grado di catturare le correlazioni a corto raggio fra le basi, tuttavia dobbiamo dire che questi sono semplici modelli che non possono riprodurre molti aspetti complessi delle sequenze di DNA, come le relazioni a lungo-raggio. Una rappresentazione grafica della catena di DNA di Markov del primo ordine è presente in Figura 2.2, essa mostra che lo spazio degli stati consiste di quattro nucleotidi e le frecce mostrano le probabilità di transizione da uno stato all'altro, ovvero dallo stato A ad A stesso e agli tre nucleotidi e così via.

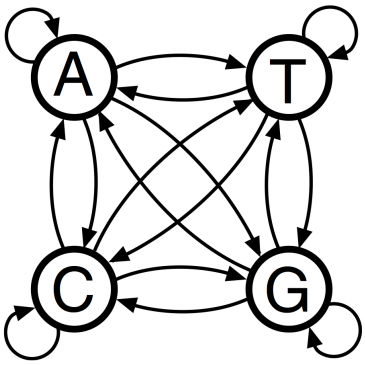


Figura 2.2: Schema di funzionamento di un modello di Markov di ordine 1 applicato alle sequenze genomiche

I due modelli markoviani utilizzati in questa tesi per la generazione random delle sequenze sono :

• **la catena di Markov di ordine zero o modello di Bernoulli**: è il modello più semplice in assoluto, nel quale la sequenza è costruita assegnando a ciascuna delle quattro lettere A, C ,G ,T una probabilità fissata, indipendente dai valori precedenti. Le probabilità di comparsa di ciascun nucleotide sono determinate dalle frequenze relative delle basi nella sequenza di DNA.

• **La catena di Markov di ordine uno**: in questo modello la sequenza sintetica di DNA è generata prendendo in considerazione le probabilità di transizione del primo ordine, ovvero la probabilità di trovare un nucleotide X dopo aver trovato Y nel precedente stato. I valori della matrice di transizione sono ricavati dalle frequenze relative dei nucleotidi e dinucleotidi nella sequenza.

La matrice di transizione per una sequenza di DNA simulata tramite la catena di Markov di ordine 1 è data da:

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

Le probabilità di transizione sono stimate dalle sequenze biologiche come segue: la probabilità $p(C \mid A)$ di trovare un C dopo aver trovato A nello stato precedente è ben approssimata dal rapporto (2.5) delle frequenze osservate di dinucleotidi CA e del nucleotide A nella sequenza di DNA.

$$p(C|A) = \frac{\#CA/\#dinucleotidi}{\#A/\#nucleotidi} \tag{2.5}$$

Modelli markoviani di ordine superiore darebbero luogo a sequenze che approssimano meglio la sequenza originale e che, pertanto, sono biologicamente più corrette: per esempio, in un modello del secondo ordine avremmo probabilità di transizione del tipo $p(A \mid AT)$, costruendo così sequenze di DNA con parole di tre lettere, che corrispondono ai codoni (pezzi di sequenza che specificano un singolo amminoacido, e che entrano in gioco durante la sintesi proteica nel processo di trascrizione). Tuttavia, ci siamo limitati

alla generazione di modelli markoviani di primo ordine perchè ci interessa confrontare i nostri risultati con quelli ottenuti con una sequenza di riferimento che sia costituita da pezzi di dinucleotidi, ma con i diniculeotidi distribuiti casualmente lungo la sequenza.

2.6 Analisi statistiche sulla sequenza markoviana di ordine 1

Come ulteriore approfondimento, abbiamo deciso di ripetere una parte delle analisi statistiche leggendo le sequenze generate con il modello markoviano di ordine 1 che riproduce abbastanza similmente la sequenza originaria. A causa dei limiti computazionali del server su cui sono state effettutate le analisi, è stato impossibile generare sequenze di Markov di ordine 1 della stessa lunghezza delle sequenze biologiche, eccetto per il batterio Escherichia coli. Invece le sequenze simulate del topo e dell'uomo sono state ridotte rispettivamente a 10000000 e 15000000 caratteri rispetto ai 195471971 e 224384768 caratteri della sequenza del cromosoma 1.

Quello che ci siamo proposti di fare è stato riprodurre gli istogrammi delle densità di probabilità $p(\tau)$ delle interdistanze τ dei dinucleotidi, e poi, anche dei nucleotidi e dei trinucleotidi. Gli istogrammi sono stati creati sempre mantenendo la soglia massima di 50 bin. I trinucleotidi sono i codoni, che si trovano lungo l'mRNA (RNA messaggero) per codificare l'informazione per l'inserimento di uno specifico amminoacido durante la sintesi proteica o per la fine della stessa (definito codone di stop). Il codone è, dunque, un'unità codificante alla base del codice genetico.

Vogliamo verificare tramite un confronto delle $p(\tau)$ se il comportamento anomalo del dinucleotide CG, che avremo modo di valutare nel successivo capitolo, si ripercuote nei trinucleotidi in cui si trova il CG, quindi ci siamo focalizzati sulle distribuzioni di distanza di ACG, CCG, CGA, CGC, CGG, CGT, GCG e TCG.

Capitolo 3

I risultati: istogrammi ed heatmaps

Dopo aver illustrato nel precedente capitolo i diversi metodi statistici che abbiamo utilizzato nel corso dell'analisi, riporteremo nel capitolo seguente i risultati ottenuti in tabelle e grafici, istogrammi ed heatmaps, dai quali è stato possibile ricavare un gran numero di osservazioni, evidenze sperimentali fisiche, matematiche, nonché diverse considerazioni che sono prova di nozioni di ambito biologico. Infine, abbiamo giustificato la scelta di una rappresentazione grafica piuttosto che un'altra, avendo come obiettivo la maggiore leggibilità dei dati .

3.1 Confronto delle distribuzioni di distanza tra dinucleotidi

In questo paragrafo presentiamo e descriviamo i risultati evinti dal calcolo delle interdistanze τ dei 16 dinucleotidi e delle relative distribuzioni di probabilità discrete $p(\tau)$ nel genoma dell'E.coli, del topo e dell'uomo. Gli istogrammi delle distribuzioni di distanza dei tre organismi diversi permettono un interessante confronto che ci porterà a scoprire le eventuali analogie e differenze. Abbiamo scelto di realizzare gli istogrammi in scala log-log in modo da facilitare l'analisi delle code delle distribuzioni. Inoltre, per ridurre il rumore nelle code delle $p(\tau)$ abbiamo adoperato un binning parzialmente logaritmico per distanze sopra una determinata soglia, questo ha permesso di migliorare la qualità e la leggibilità delle Figure 3.1 , 3.2 e 3.3. Per ognuna di queste figure è stata inserita una tabella che esprime le proprietà quantitative che caratterizzano i vettori distanza, come la lunghezza, il valore massimo, medio e mediano. Tutti i valori di distanza sono espressi in bp(base pairs - coppie di basi).

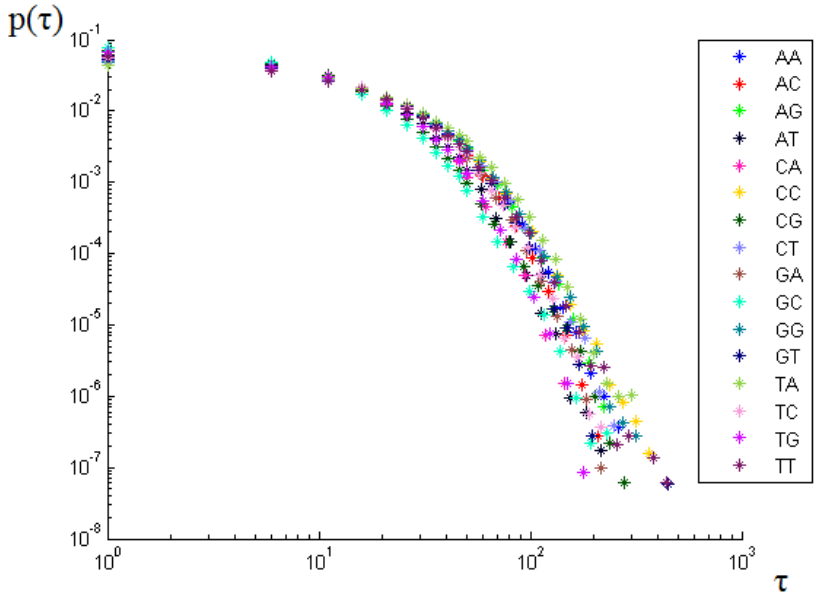


Figura 3.1: Istogramma delle distribuzioni di distanza dei dinucleotidi dell'E.coli

Dinucleotide	lunghezza vettore distanza	distanza massima	distanza media	distanza mediana
AA	255199	351	17,18	11
AC	256661	251	17,08	12
AG	237876	265	18,50	12
AT	309818	253	13,98	10
CA	325148	180	13,27	9
CC	231434	420	19,05	13
CG	346669	324	12,38	9
CT	236060	293	18,65	13
GA	267246	256	16,36	11
GC	383930	272	11,08	8
GG	230092	364	19,16	13
GT	255607	511	17,15	12
TA	211960	524	20,89	14
TC	267287	553	16,36	11
TG	322238	216	13,40	9
TT	256225	504	17,11	11

Tabella 3.1 : Proprietà dei vettori distanza computati per ciascuno dei 16 dinucleotidi nel genoma dell'E.coli.

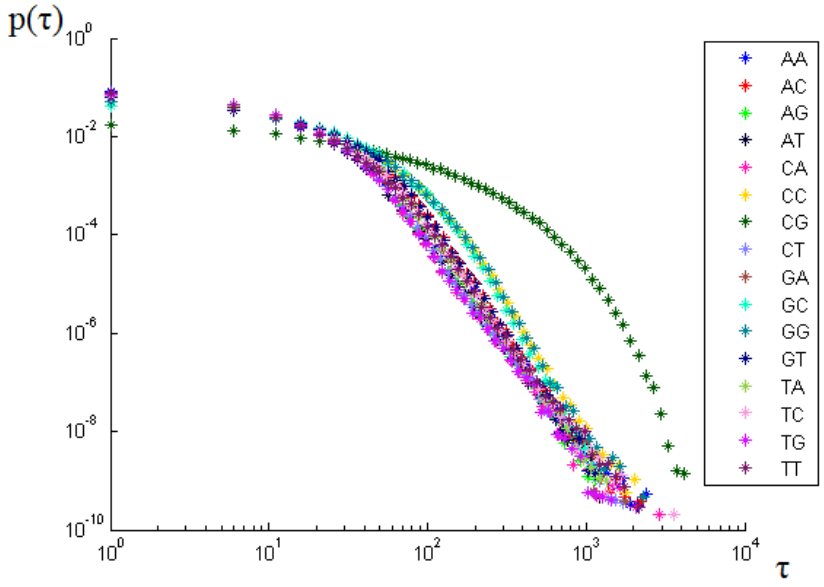


Figura 3.2: Istogramma delle distribuzioni di distanza dei dinucleotidi del topo.

Dinucleotide	lunghezza vettore distanza	distanza massima	distanza media	distanza mediana
AA	13029985	2671	13,73	8
AC	10225227	2454	17,77	11
AG	14026881	3517	12,68	8
AT	14450279	2128	12,28	8
CA	14258404	3236	12,46	8
CC	7765298	2267	23,71	14
CG	**1477560**	**4617**	**128,88**	**62**
CT	14017986	3316	12,69	8
GA	11916164	2222	15,10	9
GC	7618025	1746	24,19	15
GG	7767563	3586	23,71	14
GT	10190846	1630	17,83	11
TA	12527810	2128	14,32	9
TC	11910703	3990	15,11	9
TG	14220597	2802	12,50	8
TT	12982263	2439	13,78	8

Tabella 3.2 : Proprietà dei vettori distanza computati per ciascuno dei 16 dinucleotidi nel cromosoma 1 del topo.

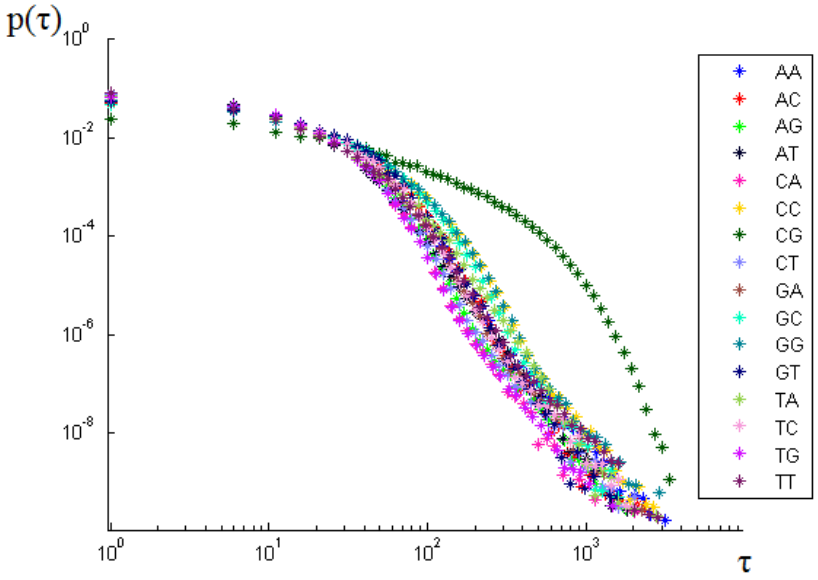


Figura 3.3: Istogramma delle distribuzioni di distanza dei dinucleotidi dell'uomo

Dinucleotide	lunghezza vettore distanza	distanza massima	distanza media	distanza mediana
AA	15427907	3565	13,60	8
AC	11315441	2629	18,91	12
AG	16057269	3118	13,03	8
AT	16786342	2025	12,42	8
CA	16383697	2269	12,75	9
CC	9733138	3009	22,15	12
CG	**2284469**	**3760**	**97,61**	**40**
CT	16081767	2272	13,01	8
GA	13465813	2524	15,73	9
GC	9950941	1815	21,64	12
GG	9727312	3236	22,16	12
GT	11335698	1980	18,87	12
TA	14309539	3122	14,74	8
TC	13483578	2264	15,71	9
TG	16410687	1668	12,73	9
TT	15462753	3148	13,57	8

Tabella 3.3 : Proprietà del vettori distanza computati per ciascuno dei 16 dinucleotidi nel cromosoma 1 dell'uomo.

Immediatamente, si evince dalle Figure 3.2 e 3.3 che negli organismi eucarioti (topo e uomo) il dinucleotide CG ha un andamento molto diverso rispetto a quello dei restanti dinucleotidi. Infatti, anche nelle corrispettive tabelle i valori medi e mediani delle interdistanze di CG nel genoma umano e del topo sono più alti rispetto a quelli dei restanti dinucleotidi. Questo suggerisce un'interpretazione biologica che spieghi il com-

portamento diverso dei dinucleotidi CG: essi sono i siti in cui può innescarsi un processo di metilazione.

3.2 La matrice divergenza JS e relative heatmaps

La divergenza di Jensen-Shannon ci permette di confrontare la frequenza relativa con cui compare nella sequenza un dinucleotide scelto con ciascuna frequenza relativa dei restanti quindici e vedere quanto le due frequenze si discostino tra loro. Dopo aver calcolato attraverso una funzione scritta in codice Matlab la frequenza con cui compare ciascuno dei 16 nucleotidi nella sequenza, ciascun elemento della risultante matrice di divergenza simmetrica indica quanto diverge la frequenza di un dinucleotide da quella di tutti gli altri 15, di conseguenza lungo la diagonale della matrice compariranno degli zeri. Riportiamo qui di seguito le matrici e le relative heatmaps per i tre organismi.

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	0	0,0119	0,0096	0,0162	0,0316	0,0158	0,0256	0,0098	0,0132	0,0371	0,0160	0,0119	0,0139	0,0136	0,0306	0,0005
AC	0,0119	0	0,0026	0,0098	0,0153	0,0034	0,0229	0,0027	0,0032	0,0359	0,0035	0,0004	0,0093	0,0036	0,0144	0,0118
AG	0,0096	0,0026	0	0,0147	0,0219	0,0016	0,0313	0,0004	0,0042	0,0456	0,0017	0,0025	0,0039	0,0045	0,0209	0,0099
AT	0,0162	0,0098	0,0147	0	0,0052	0,0186	0,0063	0,0156	0,0060	0,0129	0,0194	0,0103	0,0283	0,0062	0,0050	0,0165
CA	0,0316	0,0153	0,0219	0,0052	0	0,0239	0,0093	0,0230	0,0098	0,0142	0,0247	0,0157	0,0397	0,0099	0,0003	0,0320
CC	0,0158	0,0034	0,0016	0,0186	0,0239	0	0,0365	0,0015	0,0053	0,0533	0,0005	0,0032	0,0037	0,0055	0,0229	0,0165
CG	0,0256	0,0229	0,0313	0,0063	0,0093	0,0365	0	0,0326	0,0191	0,0039	0,0376	0,0234	0,0488	0,0194	0,0094	0,0255
CT	0,0098	0,0027	0,0004	0,0156	0,0230	0,0015	0,0326	0	0,0046	0,0472	0,0016	0,0026	0,0035	0,0049	0,0219	0,0102
GA	0,0132	0,0032	0,0042	0,0060	0,0098	0,0053	0,0191	0,0046	0	0,0321	0,0057	0,0033	0,0125	0,0003	0,0091	0,0142
GC	0,0371	0,0359	0,0456	0,0129	0,0142	0,0533	0,0039	0,0472	0,0321	0	0,0544	0,0366	0,0678	0,0327	0,0146	0,0362
GG	0,0160	0,0035	0,0017	0,0194	0,0247	0,0005	0,0376	0,0016	0,0057	0,0544	0	0,0034	0,0036	0,0060	0,0236	0,0166
GT	0,0119	0,0004	0,0025	0,0103	0,0157	0,0032	0,0234	0,0026	0,0033	0,0366	0,0034	0	0,0091	0,0037	0,0147	0,0119
TA	0,0139	0,0093	0,0039	0,0283	0,0397	0,0037	0,0488	0,0035	0,0125	0,0678	0,0036	0,0091	0	0,0129	0,0382	0,0147
TC	0,0136	0,0036	0,0045	0,0062	0,0099	0,0055	0,0194	0,0049	0,0003	0,0327	0,0060	0,0037	0,0129	0	0,0093	0,0148
TG	0,0306	0,0144	0,0209	0,0050	0,0003	0,0229	0,0094	0,0219	0,0091	0,0146	0,0236	0,0147	0,0382	0,0093	0	0,0310
TT	0,0005	0,0118	0,0099	0,0165	0,0320	0,0165	0,0255	0,0102	0,0142	0,0362	0,0166	0,0119	0,0147	0,0148	0,0310	0

Tabella 3.1: La matrice di divergenza di Jensen-Shannon della distribuzione di distanze dell'E.coli.

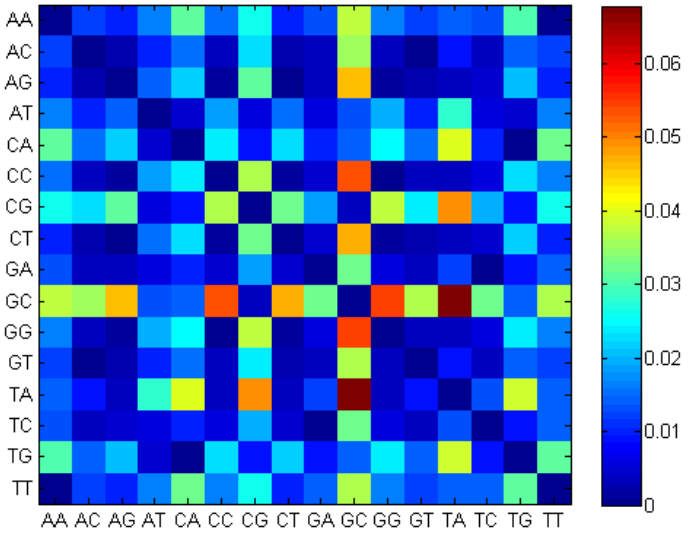


Figura 3.4: Heatmap delle distribuzioni di distanza dei dinucleotidi dell'E.coli

La rappresentazione grafica della matrice di divergenza dell'E.coli evidenzia tre dinucleotidi CC, CG e GG che potrebbero avere una differente distribuzione rispetto a quella

degli altri, ma, come vedremo in seguito, è difficile distinguere se la loro distribuzione sia esponenziale o una legge di potenza.

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	0	0,0127	0,0047	0,0071	0,0061	0,0383	0,3356	0,0047	0,0037	0,0771	0,0382	0,0130	0,0038	0,0038	0,0060	0,0001
AC	0,0127	0	0,0208	0,0256	0,0213	0,0134	0,2833	0,0207	0,0062	0,0437	0,0135	0,0000	0,0112	0,0061	0,0210	0,0125
AG	0,0047	0,0208	0	0,0012	0,0013	0,0473	0,3573	0,00002	0,0054	0,0753	0,0473	0,0212	0,0024	0,0054	0,0013	0,0047
AT	0,0071	0,0256	0,0012	0	0,0013	0,0525	0,3650	0,0012	0,0086	0,0790	0,0523	0,0260	0,0039	0,0086	0,0013	0,0073
CA	0,0061	0,0213	0,0013	0,0013	0	0,0506	0,3641	0,0013	0,0066	0,0790	0,0506	0,0217	0,0038	0,0067	0,00002	0,0062
CC	0,0383	0,0134	0,0473	0,0525	0,0506	0	0,2163	0,0473	0,0253	0,0188	0,0002	0,0132	0,0312	0,0253	0,0502	0,0378
CG	0,3356	0,2833	0,3573	0,3650	0,3641	0,2163	0	0,3573	0,3169	0,2238	0,2165	0,2822	0,3280	0,3170	0,3634	0,3346
CT	0,0047	0,0207	0,00002	0,0012	0,0013	0,0473	0,3573	0	0,0054	0,0752	0,0472	0,0211	0,0024	0,0054	0,0013	0,0047
GA	0,0037	0,0062	0,0054	0,0086	0,0066	0,0253	0,3169	0,0054	0	0,0548	0,0254	0,0064	0,0019	0,00003	0,0065	0,0036
GC	0,0771	0,0437	0,0753	0,0790	0,0790	0,0188	0,2238	0,0752	0,0548	0	0,0189	0,0434	0,0575	0,0547	0,0784	0,0766
GG	0,0382	0,0135	0,0473	0,0523	0,0506	0,0002	0,2165	0,0472	0,0254	0,0189	0	0,0133	0,0312	0,0253	0,0502	0,0378
GT	0,0130	0,00003	0,0212	0,0260	0,0217	0,0132	0,2822	0,0211	0,0064	0,0434	0,0133	0	0,0115	0,0063	0,0214	0,0128
TA	0,0038	0,0112	0,0024	0,0039	0,0038	0,0312	0,3280	0,0024	0,0019	0,0575	0,0312	0,0115	0	0,0019	0,0037	0,0037
TC	0,0038	0,0061	0,0054	0,0086	0,0067	0,0253	0,3170	0,0054	0,00003	0,0547	0,0253	0,0063	0,0019	0	0,0065	0,0036
TG	0,0060	0,0210	0,0013	0,0013	0,00002	0,0502	0,3634	0,0013	0,0065	0,0784	0,0502	0,0214	0,0037	0,0065	0	0,0061
TT	0,0001	0,0125	0,0047	0,0073	0,0062	0,0378	0,3346	0,0047	0,0036	0,0766	0,0378	0,0128	0,0037	0,0036	0,0061	0

Tabella 3.2: La matrice di divergenza di Jensen-Shannon della distribuzione di distanze del Mus musculus.

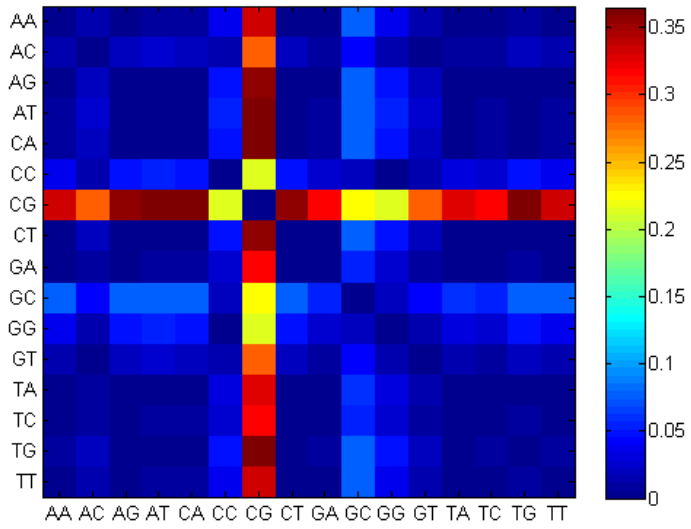


Figura 3.5: Heatmap delle distribuzioni di distanza dei dinucleotidi del topo

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	0	0,0262	0,0122	0,0125	0,0168	0,0359	0,2506	0,0122	0,0088	0,0718	0,0358	0,0260	0,0093	0,0088	0,0169	0,0001
AC	0,0262	0	0,0259	0,0336	0,0282	0,0115	0,1909	0,0262	0,0087	0,0262	0,0113	0,00002	0,0207	0,0088	0,0284	0,0264
AG	0,0122	0,0259	0	0,0030	0,0024	0,0351	0,2626	0,00002	0,0081	0,0512	0,0356	0,0257	0,0048	0,0081	0,0024	0,0123
AT	0,0125	0,0336	0,0030	0	0,0038	0,0421	0,2713	0,0029	0,0118	0,0589	0,0424	0,0333	0,0054	0,0117	0,0038	0,0124
CA	0,0168	0,0282	0,0024	0,0038	0	0,0420	0,2717	0,0024	0,0114	0,0521	0,0426	0,0279	0,0080	0,0113	0,00001	0,0169
CC	0,0359	0,0115	0,0351	0,0421	0,0420	0	0,1544	0,0353	0,0172	0,0207	0,0016	0,0115	0,0236	0,0173	0,0423	0,0361
CG	0,2506	0,1909	0,2626	0,2713	0,2717	0,1544	0	0,2629	0,2209	0,1706	0,1548	0,1912	0,2329	0,2212	0,2721	0,2509
CT	0,0122	0,0262	0,00002	0,0029	0,0024	0,0353	0,2629	0	0,0082	0,0514	0,0358	0,0260	0,0048	0,0081	0,0024	0,0123
GA	0,0088	0,0087	0,0081	0,0118	0,0114	0,0172	0,2209	0,0082	0	0,0382	0,0180	0,0086	0,0060	0,00003	0,0116	0,0090
GC	0,0718	0,0262	0,0512	0,0589	0,0521	0,0207	0,1706	0,0514	0,0382	0	0,0217	0,0263	0,0428	0,0383	0,0524	0,0722
GG	0,0358	0,0113	0,0356	0,0424	0,0426	0,0016	0,1548	0,0358	0,0180	0,0217	0	0,0113	0,0240	0,0181	0,0429	0,0360
GT	0,0260	0,00002	0,0257	0,0333	0,0279	0,0115	0,1912	0,0260	0,0086	0,0263	0,0113	0	0,0205	0,0087	0,0282	0,0262
TA	0,0093	0,0207	0,0048	0,0054	0,0080	0,0236	0,2329	0,0048	0,0060	0,0428	0,0240	0,0205	0	0,0060	0,0081	0,0093
TC	0,0088	0,0088	0,0081	0,0117	0,0113	0,0173	0,2212	0,0081	0,00003	0,0383	0,0181	0,0087	0,0060	0	0,0115	0,0089
TG	0,0169	0,0284	0,0024	0,0038	0,00001	0,0423	0,2721	0,0024	0,0116	0,0524	0,0429	0,0282	0,0081	0,0115	0	0,0169
TT	0,0001	0,0264	0,0123	0,0124	0,0169	0,0361	0,2509	0,0123	0,0090	0,0722	0,0360	0,0262	0,0093	0,0089	0,0169	0

Tabella 3.3: La matrice di divergenza di Jensen-Shannon della distribuzione di distanze dell'uomo.

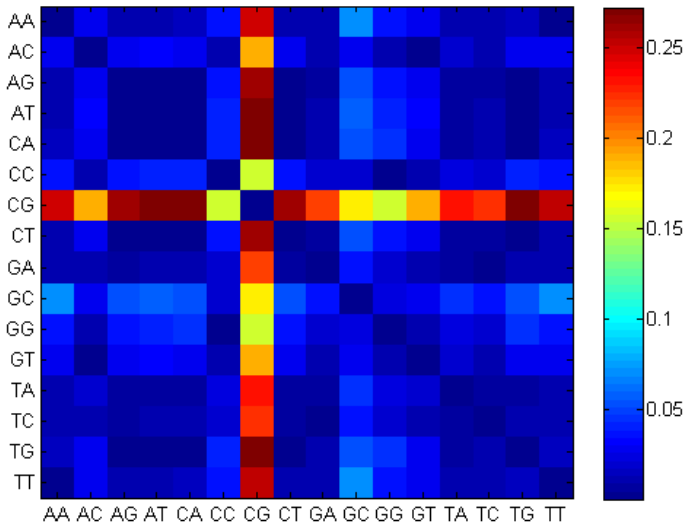


Figura 3.6: Heatmap delle distribuzioni di distanza dei dinucleotidi dell'uomo

Come si vede in Figura 3.5 e 3.6 la distribuzione del dinucleotide CG si discosta in modo evidente da quella degli altri dinucleotidi.

3.3 Il fit delle code delle distribuzioni

Dall'osservazione degli istogrammi e delle heatmaps dei paragrafi precedenti, si potrebbe ipotizzare che il dinucleotidide CG abbia negli organismi eucarioti, come il topo e l'uomo, un ruolo peculiare che invece è assente nell'organismo procariote E.Coli. L'evidenza sperimentale di questo risultato desta subito l'attenzione, per cui ci è sembrato opportuno approfondire questo aspetto che accomuna il topo e l'uomo attraverso un fit delle code delle distribuzioni di distanza dei tre organismi. Abbiamo studiato se tali code mostrano un decadimento algebrico del tipo $y \sim x^{-|b|}$ o un decadimento di tipo esponenziale $y \sim e^{-|d|\cdot|x|}$, e ci siamo soffermati sul valore dei parametri b e d, a seconda dei casi. Nel caso dell'E.Coli è difficile distinguere se le code siano fittate meglio da un esponenziale o una legge di potenza.

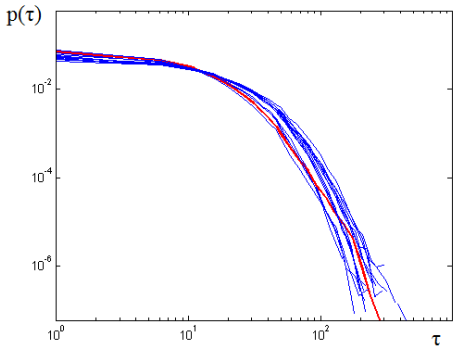


Figura 3.7: Le distribuzioni di distanza dei dinucleotidi dell'E.coli, la distribuzione del dinucleotide CG è in rosso.

Per quanto riguarda il topo, abbiamo già visto la marcata differenza di andamento della distribuzione di distanza del dinucleotide CG in figura 3.8 e ne abbiamo dato una giustificazione riferendoci all'azione degli enzimi DNMT. Il miglior fit per la coda della distribuzione di distanza:

- di CG è un esponenziale della forma $y \approx a \cdot e^{d \cdot x}$ ed i parametri del fit sono

  $a = 0.00047 \pm 0.00017$ e $d = -(0.00333 \pm 0.00028)$. Un'ulteoriore conferma che si tratti di un buon fit è data dai coefficienti di determinazione $R^2 = 0.9963$ ed aggiustato (o corretto) $\bar{R}^2 = 0.9959$.

  Possiamo notare che la distribuzione esponenziale segnala la presenza di una "lunghezza caratteristica" tra consecutive comparse del medesimo dinucleotide nella sequenza, per cui dal parametro d ricaviamo tale lunghezza caratteristica :

$$\lambda = \frac{1}{\mid d \mid} = \frac{1}{0.00333} = 300,3bp = 300bp \tag{3.1}$$

- dei restanti dinucleotidi è una legge di potenza del tipo $y \approx a \cdot x^b$.

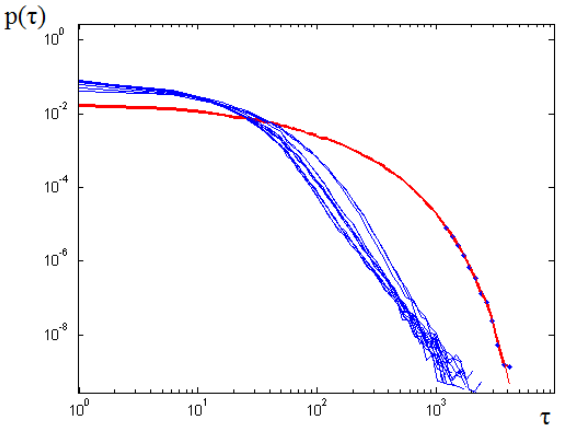


Figura 3.8: Le distribuzioni di distanza dei dinucleotidi del topo, la distribuzione del dinucleotide CG è in rosso.

In modo analogo, nell'uomo abbiamo notato in figura 3.9 che la distribuzione di distanze di CG si discosta molto da quella degli altri 15 dinucleotidi, e la spiegazione di questo anomalo comportamento deve essere ricercata nei processi di metilazione. Esaminando le code di tali distribuzioni, si vede che il miglior fit per la distribuzione:

- di CG è sempre un esponenziale della forma $y \approx a \cdot e^{d \cdot x}$, i parametri del fit sono $a = 0.00051 \pm 0,00015 \quad d = -(0.0040 \pm 0,0002)$.

  I valori del coefficiente di determinazione $R^2 = 0.9982$ e del coefficiente di determinazione aggiustato (o corretto) $\bar{R}^2 = 0.998$ danno prova della bontà del fit.

  La lunghezza caratteristica per l'uomo si ricava dal parametro d del fit:

$$\lambda = \frac{1}{\mid d \mid} = \frac{1}{0.0040} = 250bp. \tag{3.2}$$

- dei restanti dinucleotidi è una legge di potenza $y \approx a \cdot x^b$.

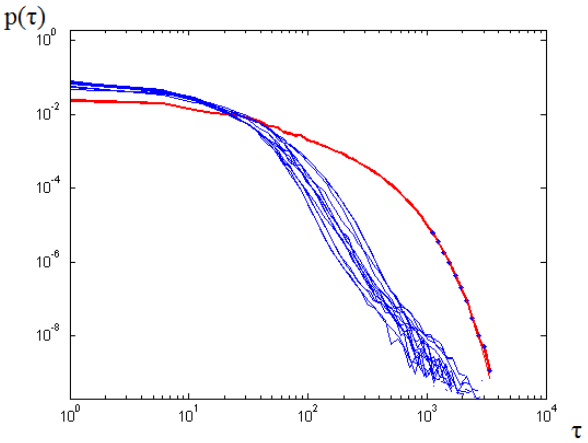


Figura 3.9: Le distribuzioni di distanza dei dinucleotidi dell'uomo, la distribuzione del dinucleotide CG è in rosso.

Il dinucleotide CG mostra dunque una sorta di costante spaziatura lungo l'intera sequenza di DNA, a differenza degli altri dinucleotidi, le cui code della distribuzione sono fittate da una power-law, suggerendoci l'assenza di una singola lunghezza caratteristica. Questo aspetto sembra evidenziare un ruolo strutturale del CG, che si ripresenta nella sequenza con una certa regolarità, indicata dalla lunghezza caratteristica λ.

Il decadimento esponenziale, piuttosto che algebrico, del dinucleotide CG negli eucarioti suggerisce che essi sono i siti dove si attacca un gruppo metile per azione di una specifica famiglia di enzimi, la DNA metiltrasferasi di cui abbiamo parlato nel paragrafo 1.3 del capitolo 1.

In prossimità delle regioni codificanti (geni) , questi dinucleotidi possono essere trovati in gruppi, che prendono il nome di CpG islands che hanno un ruolo fondamentale nella regolazione epigenetica. Dobbiamo inoltre sottolineare che, siccome le regioni codificanti (che si occupano della sintesi proteica) del genoma del topo e dell'uomo costituiscono solo una piccola parte, circa l'1 % , dell'intera sequenza, le nostre statistiche riguardano

principalmente le regioni non codificanti (per il 25% introni), che sono candidate per avere ruoli funzionali nella struttura tridimensionale della cromatina e nella regolazione di elementi trasponibili.

3.4 Confronto con i modelli markoviani

Dopo aver comparato le differenti distribuzioni di distanza tra loro, abbiamo applicato dei modelli random, come quelli markoviani, per un ulteriore confronto. Abbiamo generato le sequenze simulate di Markov di ordine zero ed uno e calcolato quali delle due approssima meglio la sequenza biologica di partenza.

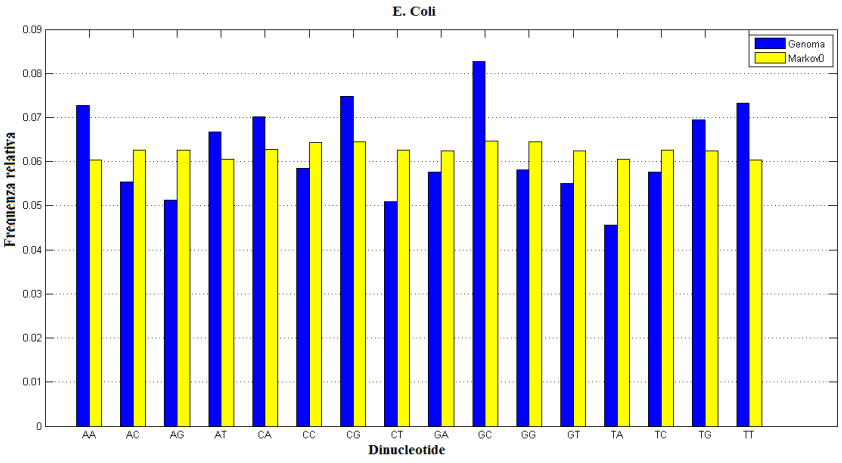


Figura 3.10: Confronto tra le frequenze relative di comparsa dei dinucleotidi nel genoma dell'E.coli e in una sequenza random generata con una catena di Markov di ordine 0.

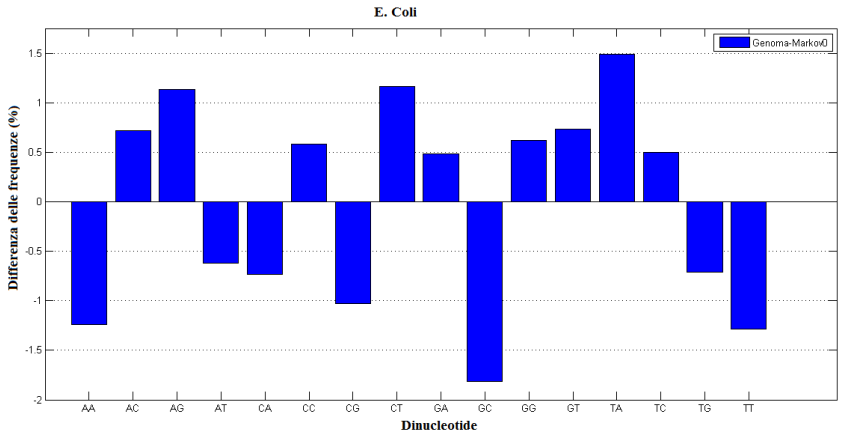


Figura 3.11: Differenza percentuale di frequenza relativa dei dinucleotidi tra il genoma dell'ecoli e la sequenza simulata Markov0.

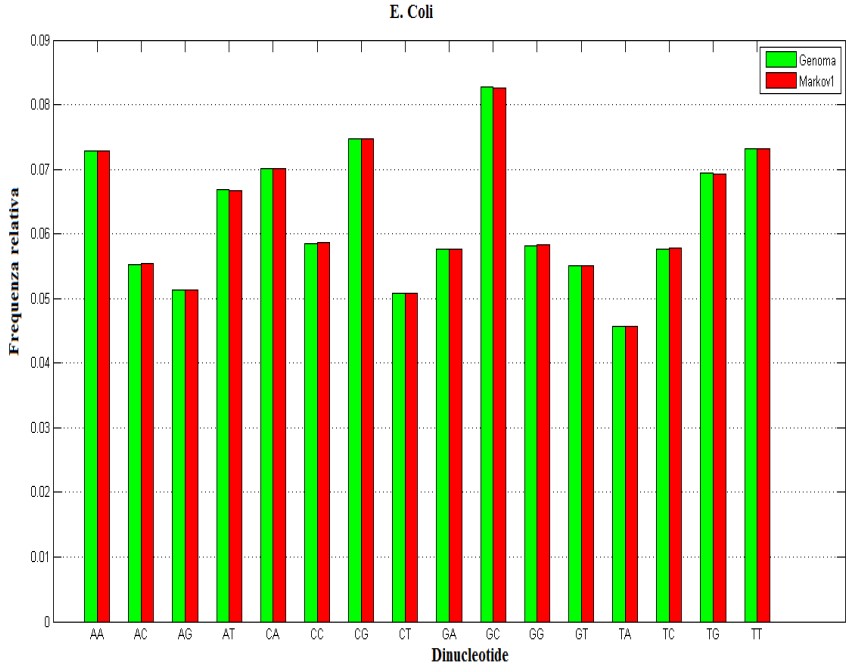


Figura 3.12: Confronto tra le frequenze relative di comparsa dei dinucleotidi nel genoma dell'E.coli e in una sequenza random generata con una catena di Markov del 1° ordine.

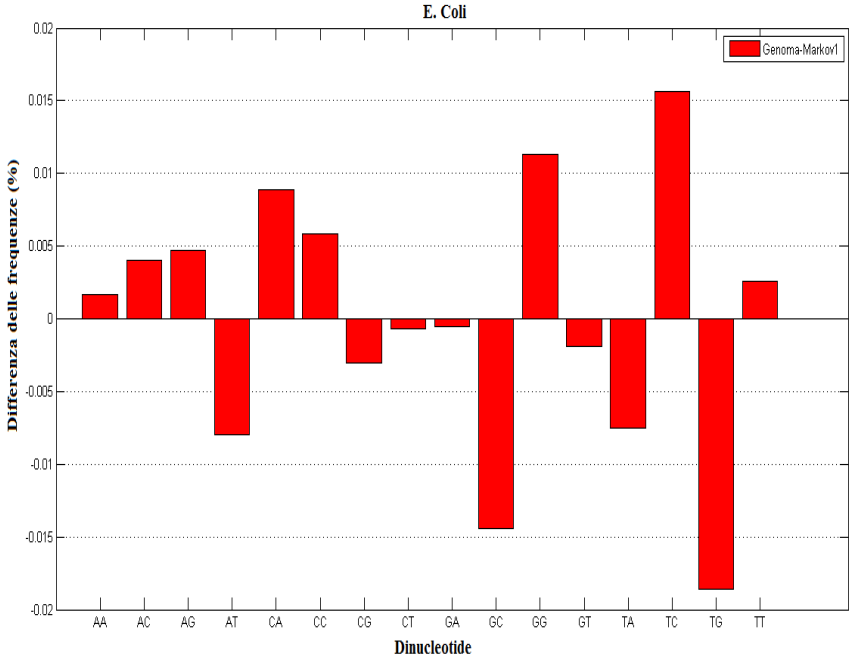


Figura 3.13: Differenza percentuale di frequenza relativa dei dinucleotidi tra il genoma dell'ecoli e la sequenza simulata Markov1.

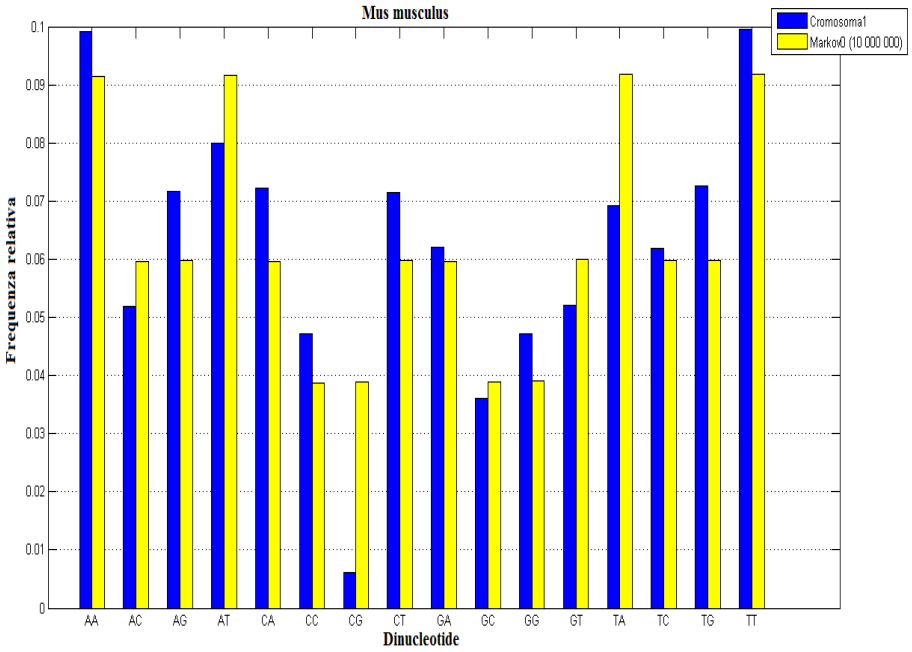


Figura 3.14: Confronto tra le frequenze relative di comparsa dei dinucleotidi nel cromosoma 1 del topo e in una sequenza random generata con una catena di Markov di ordine 0.

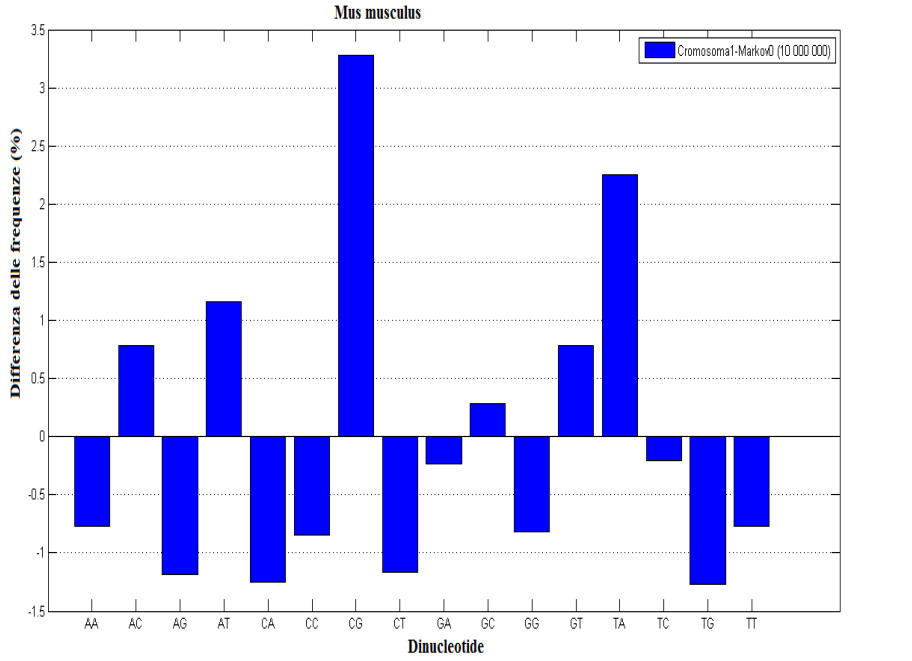


Figura 3.15: Differenza percentuale di frequenza relativa dei dinucleotidi tra il cromosoma 1 del topo e la sequenza simulata Markov0.

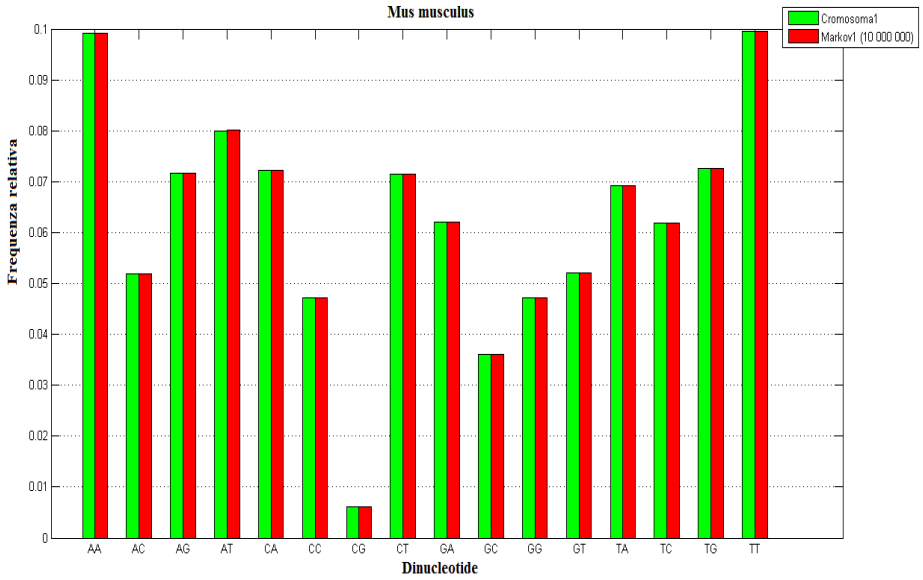


Figura 3.16: Confronto tra le frequenze relative di comparsa dei dinucleotidi nel cromosoma 1 del topo e in una sequenza random generata con una catena di Markov del 1° ordine.

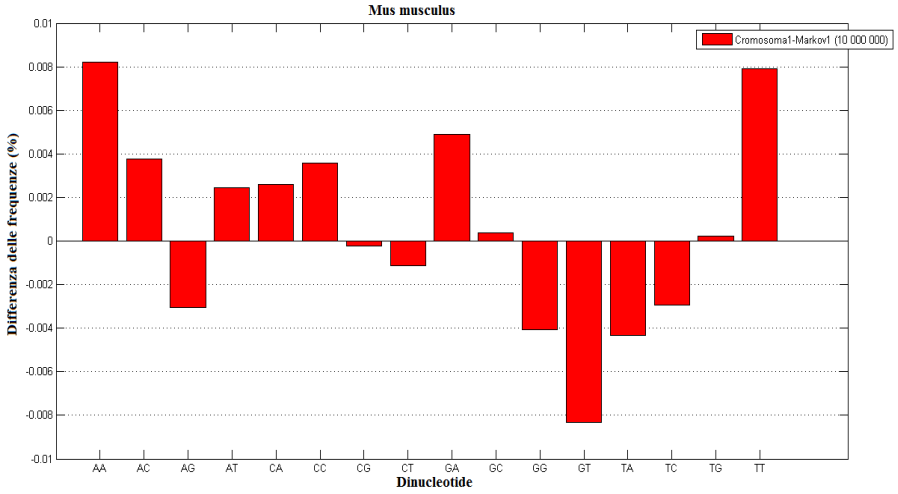


Figura 3.17: Differenza percentuale di frequenza relativa dei dinucleotidi tra il cromosoma 1 del topo e la sequenza simulata Markov1.

Come si può notare dalle leggende degli istrogrammi, le sequenze di Markov di ordine 0 e 1 generate sono ridotte a 10000000 caratteri, al fronte dei 195471971 caratteri della sequenza originaria. Abbiamo operato una tale modifica a causa di problemi computazionali che non hanno permesso la generazione di sequenze random così lunghe.

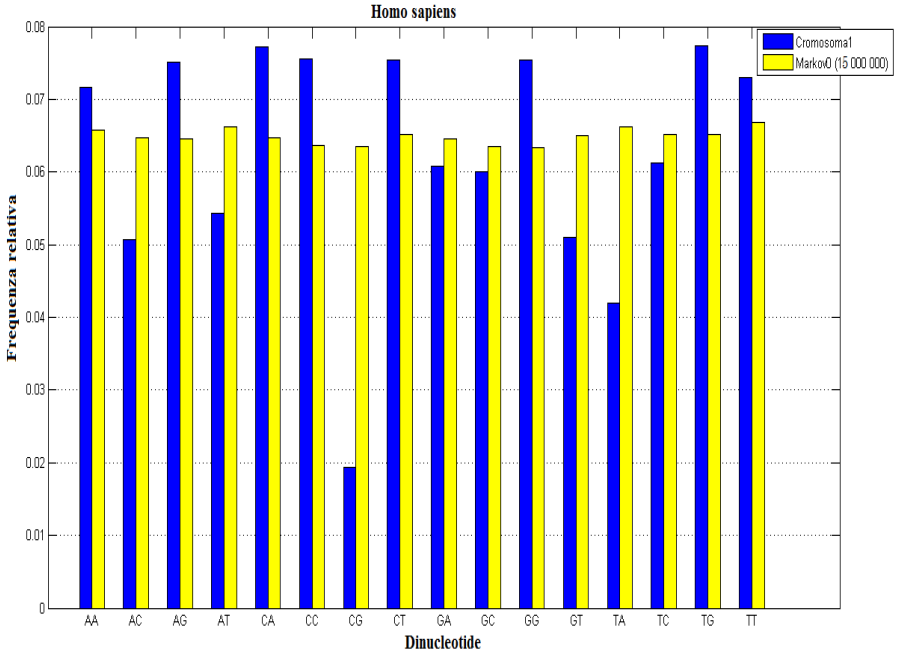


Figura 3.18: Confronto tra le frequenze relative di comparsa dei dinucleotidi nel cromosoma 1 dell'uomo e in una sequenza random generata con una catena di Markov di ordine 0.

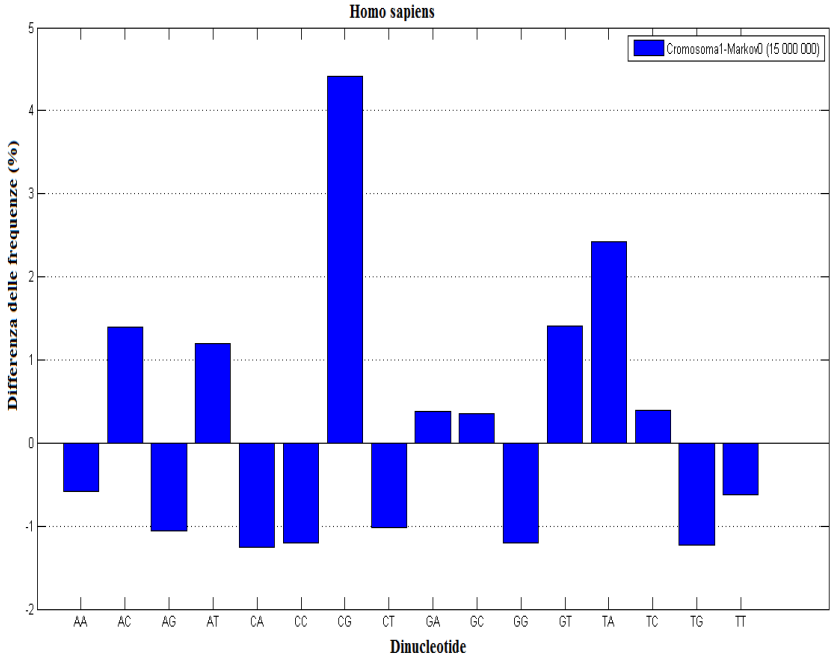


Figura 3.19: Differenza percentuale di frequenza relativa dei dinucleotidi tra il cromosoma 1 dell'uomo e la sequenza simulata Markov0.

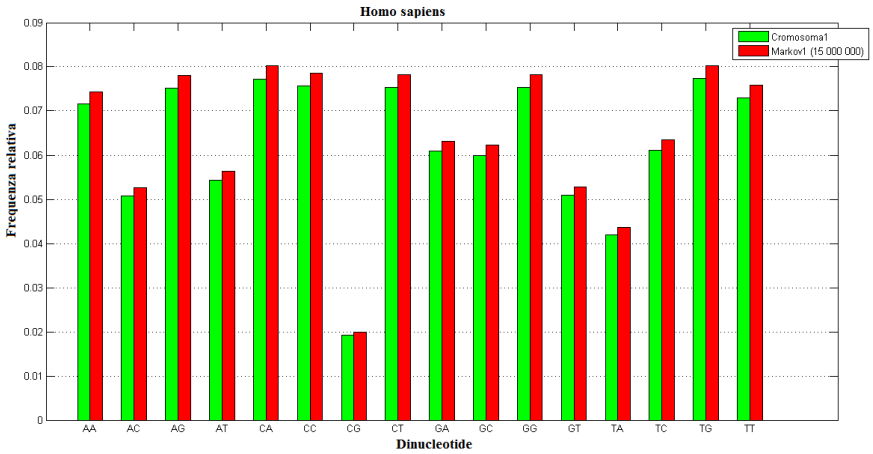


Figura 3.20: Confronto tra le frequenze relative di comparsa dei dinucleotidi nel cromosoma 1 dell'uomo e in una sequenza random generata con una catena di Markov del 1° ordine.

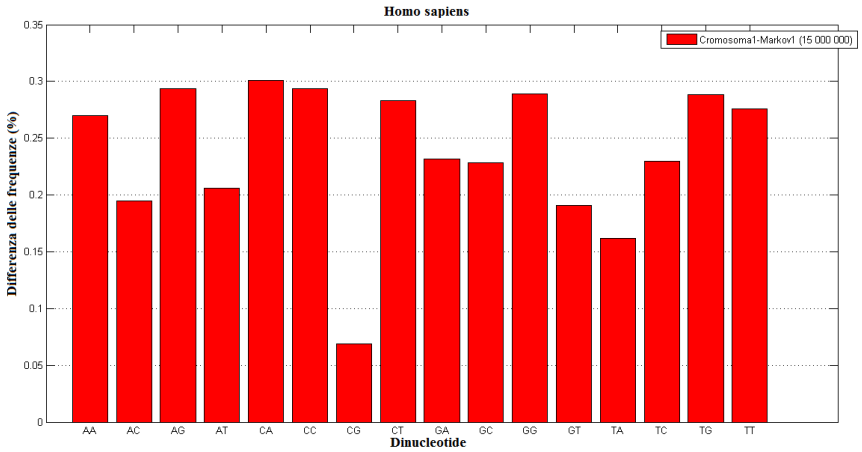


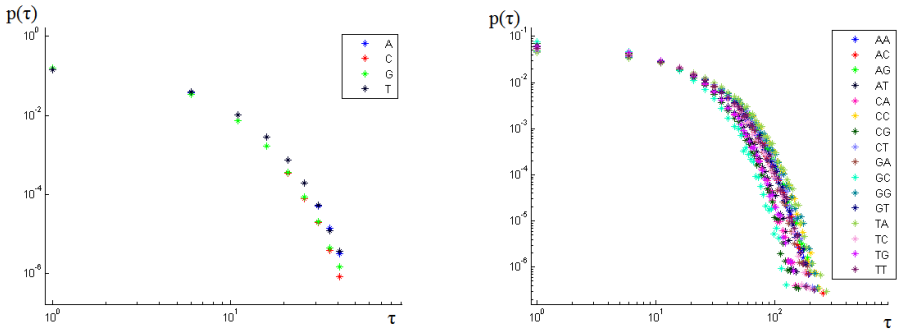
Figura 3.21: Differenza percentuale di frequenza relativa dei dinucleotidi tra il cromosoma 1 dell'uomo e la sequenza simulata Markov1.

Nel caso delle catene di Markov generate per l'uomo abbiamo operato una riduzione della sequenza a 15000000 caratteri, come è evidenziato nelle leggende delle figure, al fronte dei 224384768 caratteri della sequenza del cromosoma 1 dell'uomo.

3.5 Analisi statistiche sulla catena di Markov di 1° ordine

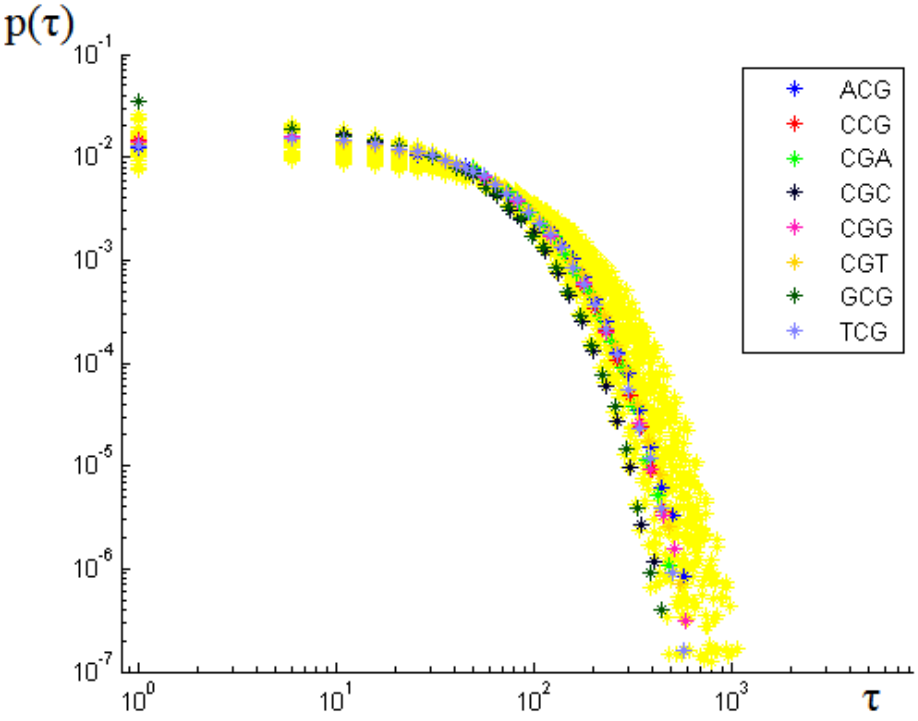
Infine, abbiamo scelto di ripetere ed eseguire alcune analisi statistiche leggendo la sequenza generata con il modello Markov di 1° ordine. Come si evince dalle figure la catena Markov0, che non utilizza la matrice probabilità di transizione, genera una sequenza molto differente da quella biologica originaria; invece la catena Markov1 è una buona approssimazione della sequenza originaria. Tuttavia, a causa dei limiti computazionali

del server utilizzato, non è stato possibile generare catene Markov1 del topo e dell'uomo della stessa lunghezza del cromosoma 1 di tali organismi. Dopo aver letto le sequenze Markov1 dell'E.coli, del topo e dell'uomo, abbiamo calcolato le interdistanze τ e, successivamente, la distribuzione delle distanze $p(\tau)$ tra nucleotidi, dinucleotidi e trinucleotidi in ciaascuna delle tre sequenze generate. Riportiamo qui di seguito gli istogrammi, realizzati sempre con un binning parzialmente logaritmico e con una soglia massima di bin pari a 50.



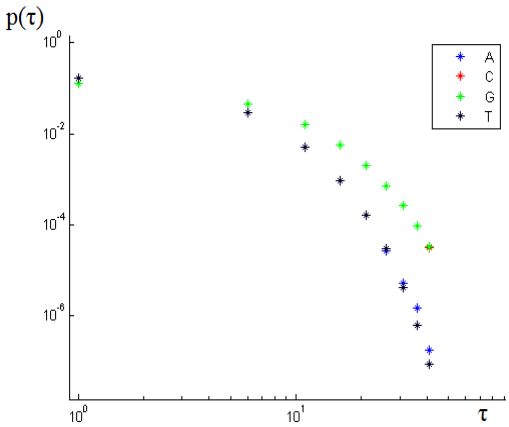
(a)

(b)

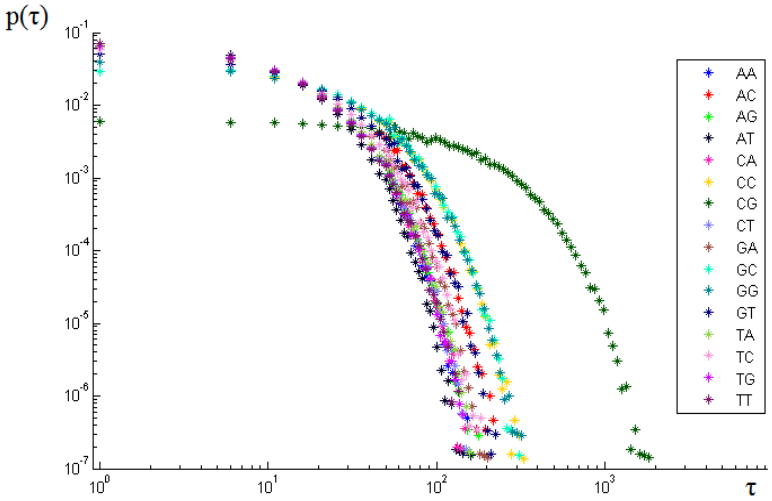


(c)

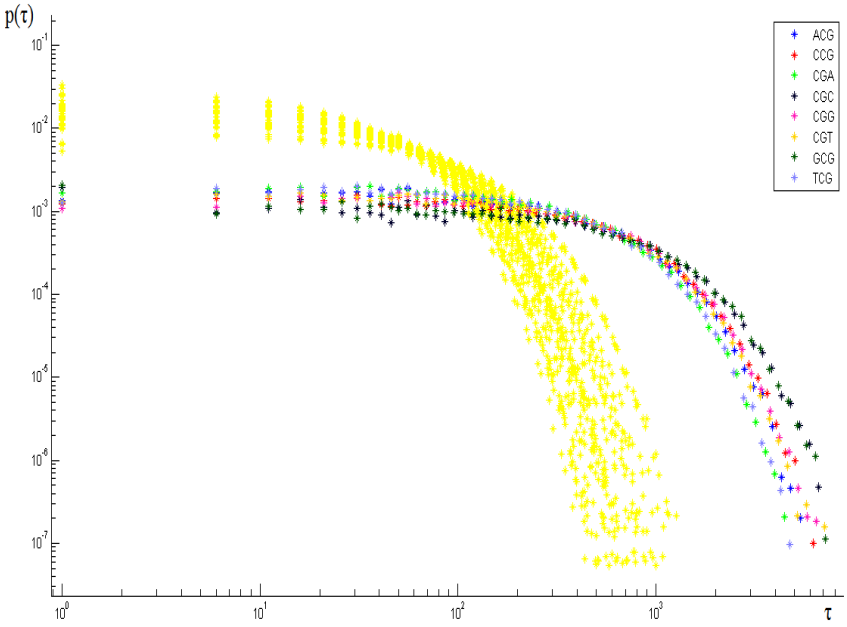
Figura 3.22: Distribuzione di distanza dei 4 nucleotidi (a), 16 dinucleotidi (b), 64 trinucleotidi (c) nella sequenza random Markov1 dell'E.Coli. Nella (c) le curve in giallo sono le distribuzioni di distanza dei 56 trinucleotidi in cui è assente CG.



(a)

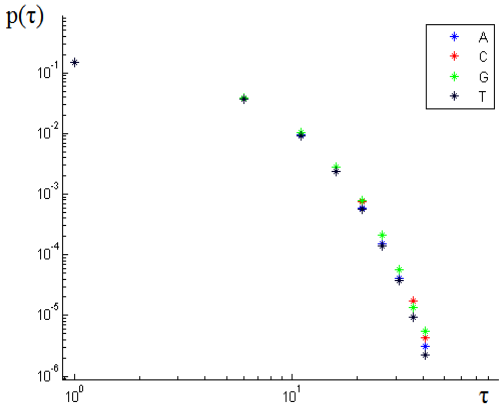


(b)

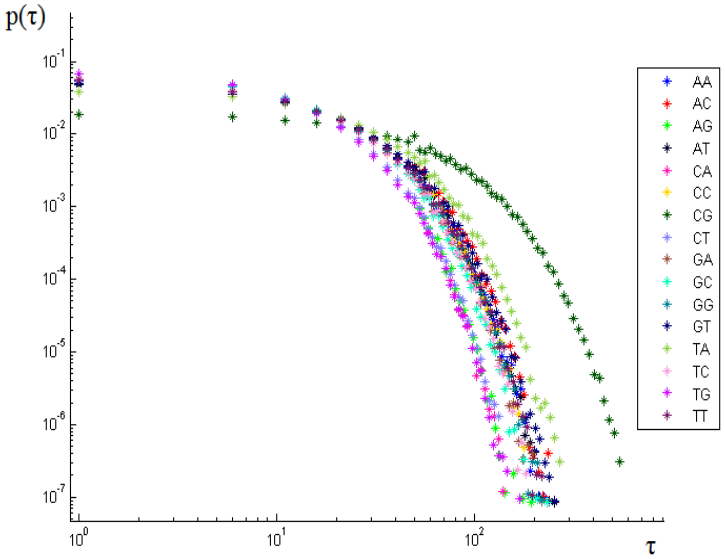


(c)

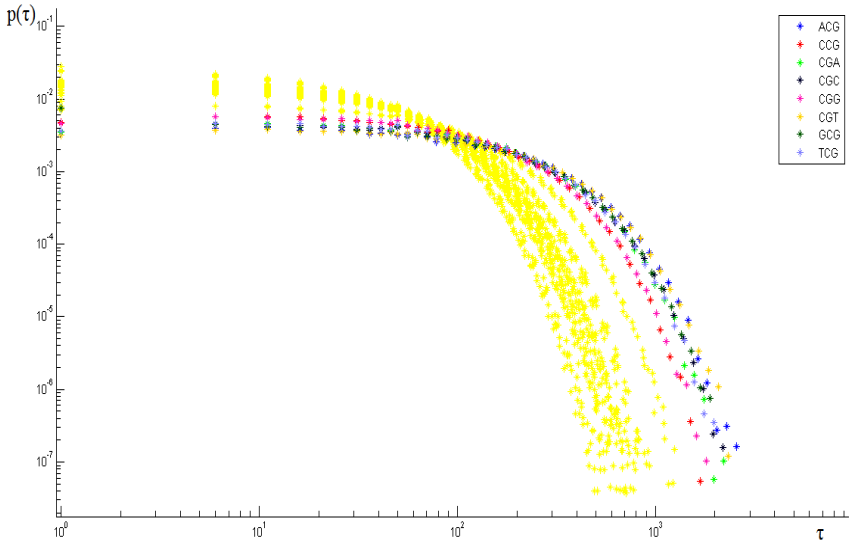
Figura 3.23: Distribuzione di distanza dei 4 nucleotidi (a), 16 dinucleotidi (b), 64 trinucleotidi (c) nella sequenza random Markov1 del topo. Nella (c) le curve in giallo sono le distribuzioni di distanza dei 56 trinucleotidi in cui è assente CG.



(a)



(b)



(c)

Figura 3.24: Distribuzione di distanza dei 4 nucleotidi (a), 16 dinucleotidi (b), 64 trinucleotidi (c) nella sequenza random Markov1 dell'uomo. Nella (c) le curve in giallo sono le distribuzioni di distanza dei 56 trinucleotidi in cui è assente CG.

Come si può notare nelle Figure 3.23 (b) e 3.24 (b) il comportamento anomalo del dinucleotide CG si ripresenta nelle sequenze Markov1 generate per il topo e per l'uomo. L'aspetto più interessante di questa analisi è stato quello di spostare le nostre attenzioni anche ai trinucleotidi, che costituiscono i codoni, fondamentali per la sintesi proteica. Dalle figure 3.23 (c) e 3.24 (c) emerge che le distribuzioni di distanza dei trinucleotidi in cui è presente il CG si discostano dalle distribuzioni dei restanti 56 trinucleotidi (curve in giallo) del topo e dell'uomo. Potremmo spiegarci questo fenomeno pensando che il CG che compone la tripletta potrebbe essere oggetto di metilazione, ovvero ad esso potrebbe attaccarsi un gruppo metile -CH3. Dal confronto delle figure 3.22 (c), 3.23 (c) e 3.24 (c) si nota, nuovamente, che il comportamento anomalo dei trinucleotidi in cui è presente CG riguarda solamente gli organismi di tipo eucariote.

Capitolo 4

Conclusioni

Alla fine di questa tesi ci proponiamo di trarre qualche conclusione dall'osservazione dei risultati grafici e numerici ottenuti con analisi qualitative e quantitative. In questo capitolo spieghiamo lo scopo per cui abbiamo affrontato questo studio e gli eventuali sviluppi futuri a partire da ciò che è stato trattato in questa tesi.

4.1 Considerazioni sulle distribuzioni di distanza e sui fit delle code

Abbiamo visto negli istogrammi delle distribuzioni di distanza del topo e dell'uomo la notevole differenza della distribuzione di CG, ed abbiamo capito che l'anomalo comportamento di tale curva poteva trovare una spiegazione nel ruolo strutturale e funzionale del DNA all'interno di questi organismi. Infatti, nel topo e nell'uomo, i dinucleotidi CG sono i siti in cui potrebbe avvenire il processo di metilazione, ovvero un meccanismo epigenetico che riveste un ruolo fondamentale nella regolazione genica e anche nella formazione strutturale della cromatina. Un'analisi più dettagliata delle distribuzioni rivela che nel topo e nell'uomo la distribuzione di distanza del dinucleotide CG è di tipo esponenziale, quindi tale distribuzione presenta una lunghezza caratteristica e può essere confrontata con le distribuzioni degli altri dinucleotidi, le cui code sono fittate da una legge di potenza che pertanto è priva di una scala caratteristica. Inoltre, i parametri associati alle distribuzioni del topo e dell'uomo, come la lunghezza caratterestica λ e l'esponente b della power-law, sono consistentemente simili (stesso ordine di grandezza) e questo suggerisce una somiglianza strutturale dei DNA dei due organismi a livello di distanze tra dinucleotidi. Possiamo concludere che sia gli istrogrammi delle distribuzioni che i rispettivi fit annessi confermano un'evidente analogia tra topo e uomo e, soprattutto, un elevato grado di somiglianza di possibile innesco dei processi di metilazione dell'uno e dell'altro.

Abbiamo eseguito le medesime analisi statistiche anche sull'organismo procariote Escherichia Coli, un batterio che non possiede meccanismi epigenetici simili, e che probabilmente non sfrutta i processi di metilazione del DNA per gli stessi scopi degli organismi pluricellulari. Quest'ultima ipotesi è confermata dalla mancanza di significative differenze delle distribuzioni di distanza, sia per quanto riguarda la loro pendenza che i parametri

dei fit delle loro code. Inoltre, sottolineamo ancora che in questo caso non si riesce a distinguere chiaramente la pendenza delle code della distribuzione, per cui né il fit power-law né quello esponenziale possono ritenersi soddisfacenti. Tutte queste osservazioni ci portano a pensare al differente ruolo della metilazione di DNA nei batteri ed anche che solo una piccola porzione di sequenza di DNA del batterio è coinvolta in questo processo, ma il nostro studio non si presta a quantificare differenze di questo tipo.

Infine, possiamo concludere che questo studio conferma le differenze sostanziali che esistono tra organismi eucarioti e procarioti a livello di struttura e funzionalità dei rispettivi DNA. Sarebbe interessante ripetere l'analisi su altri organismi, sia di tipo eucariote che procariote, in modo da poter confermare quanto è stato già detto, o eventualmente, smentirlo, basandosi sulle intuizioni delle nozioni biologiche che si nascondano dietro questi risultati.

4.2 Considerazioni sulle catene di Markov di ordine 0 e 1

Dopo aver computato le frequenze dei dinucleotidi nelle sequenze di DNA, abbiamo notato nel topo e nell'uomo, la scarsa percentuale del dinucleotide CG rispetto a quella degli altri dinucleotidi. Di seguito, abbiamo generato sequenze random usando i modelli della catena di Markov di ordine zero (basata sulle frequenze relative dei nucleotidi) ed uno (basata sulle probabilità di transizione tra diversi nucleotidi). Abbiamo visto che il modello Markov 0 sovrastima il contenuto di CG nella sequenza di oltre il 3% e il 4 % rispettivamente per il topo e per l'uomo (come si osserva nelle Figure 3.15 e 3.19).

Invece, il modello Markov 1 è capace di riprodurre le abbondanze dei dinucleotidi abbastanza fedelmente, con differenze percentuali di frequenza inferiore allo 0,02 % per l'E.coli, 0,01% per il topo , 0,3 % per l'uomo (come si osserva nelle Figure 3.13 , 3.17 e 3.21). Inoltre dalle figure emerge che nella catena Markov 1 del topo e dell'uomo l'abbondanza di CG è molto inferiore rispetto a quella degli altri nucleotidi, così come lo era nel cromosoma 1 di questi organismi. Per questi motivi, abbiamo deciso di servirci delle catene generate Markov1 per altre analisi statistiche che hanno evidenziato ulteriormente la validità delle intuizioni biologiche.

Appendice A

Files in formato FASTA

Il codice utilizzato per le analisi statistiche è scritto in Matlab e il Bioinformatic Toolbox, di cui è provvisto il software, fornisce utili funzioni, come per esempio quella di poter leggere i file in formato FASTA. Un file in formato FASTA è ampiamente usato per rappresentare sequenze nucleotidiche o sequenze peptidiche. Una sequenza in formato FASTA inizia con una riga di descrizione, seguita dalla sequenza stessa. La riga di intestazione è distinta dalla sequenza grazie al simbolo nella prima colonna. La parola che segue il simbolo è un identificatore, mentre la frase successiva fornisce una descrizione del contenuto della sequenza.

Per esempio, per il genoma dell'E.coli:

>U00096.2 Escherichia coli K-12 MG1655 complete genome (4639675 bp) ;

per il cromosoma 1 del Mus Musculus:

>1dna:chromosome chromosome:GRCm38:1:1:195471971:1 REF

per il cromosoma 1 dell'Homo Sapiens:

>gi—224384768—gb—CM000663.1— Homo sapiens chromosome 1, GRCh37 primary reference assembly .

Database delle sequenze genomiche:

I siti internet da cui abbiamo prelevato le sequenze di DNA sono i seguenti:

- Homo Sapiens (human) genome sequence release hg19
  (`http://hgdownload.cse.ucsc.edu/downloads.html#human`);
- Mus Musculus(mouse) genome sequence release 76
  (`http://ftp.ensembl.org/pub/release-76/fasta/mus_musculus/dna/`);
- Escherichia coli K-12 strain genome
  (`http://www.ncbi.nlm.nih.gov/nuccore/U00096`).

Appendice B

Implementazione delle analisi

In questa appendice descriviamo in dettaglio le principali parti del programma implementato per l'analisi dei dinucleotidi nelle sequenze di DNA. Il codice di scrittura è Matlab, e sono state impiegate alcune funzioni del Toolbox Bioinformatics.

La distribuzione di distanze tra dinucleotidi

La prima parte del programma è dedicata all'analisi delle distribuzioni di distanza dei dinucleotidi nella sequenza. Le interdistanze sono calcolate con il metodo di non sovrapposizione tramite la funzione calc_dist2, i cui argomenti in input sono la sequenza dei nucleotidi e il dinucleotide scelto, essa restituisce in output i valori di distanza computati in un array. Il processo viene interato 16 volte, per cui alla fine le interdistanze di ciascuno dei 16 dinucleotidi sono immagazzinate in un cell array 1x16.

Le distribuzioni sono poi graficate con una scala log-log, usando un binning parzialmente logaritmico per ridurre il rumore nelle code. La funzione partial_log_bin prende in input le interdistanze calcolate e le suddivide nei diversi bin, permettendo di controllare alcuni parametri come la soglia massima di bin e la larghezza del bin.

Per confrontare in modo quantitativo le diverse distribuzioni, calcoliamo la matrice divergenza di Jensen Shannon, una versione simmetrizzata della Kullback-Leibler, usando la funzione KL nel case specifico js. La matrice è poi rappresentata graficamente in modo conveniente da una heatmap.

Modelli random di riferimento

I modelli scelti per la generazione di sequenze random di riferimento sono quelli markoviani e, precisamente, le catene di Markov di ordine 0 e 1. La catena markoviana di ordine nullo è generata usando la funzione randseq, che prende in input la lunghezza desiderata della sequenza ed i pesi probabilistici dei 4 nucleotidi. Questi sono stati stimati dalle sequenze biologiche di partenza come le frequenze relative dei nucleotidi computate con la funzione basecount.

Per la generazione della catena markoviana del 1° ordine abbiamo implementato la funzione hmmgenerate (hidden markovian model generate), usando una matrice di emissione uguale a 1 ed ottenendo una sequenza di stati (variabile states) che è poi convertita nelle 4 lettere A, C, G, T. La scelta di impiegare una tale funzione è dettata da ragioni di efficienza. Le probabilità di transizione sono calcolate dalle frequenze relative dei nu-

cleotidi e dinucleotidi nelle sequenze del cromosoma 1, attraverso le funzioni basecount e dimercount rispettivamente (si noti che i conteggi dei dinucleotidi sono sovrapposti, cioè un tratto di sequenza AAA è conteggiato come due dinucleotidi).

Successivamente, abbiamo computato le interdistanze tra nucleotidi attraverso la funzione calc_dist1 dandole come input la sequenza Markov1 ed un nucleotide , quelle tra dinucleotidi con la funzione calc_dist2 descritta sopra, quelle tra trinucleotidi con la funzione calc_dist3 i cui argomenti input sono la catena Markov1 ed un trinucleotide. Alla fine di questi processi iterati per il numero di volte opportuno, abbiamo graficato le distribuzioni di distanza tramite istogrammi, servendoci della funzione partial_log_bin descritta precedentemente.

Ringraziamenti

Al termine di questo lavoro di tesi desidero ringraziare il mio relatore, il Prof.re Daniel Remondini, per avermi suggerito un argomento così stimolante e che ritengo affine alle mie curiosità personali. Un ringraziamento speciale anche il supporto dimostratomi in questi mesi, e per avermi incitato ad applicarmi nei momenti in cui credevo di non riuscire a portare a compimento il lavoro.
Un grazie di cuore a tutti i componenti della mia splendida famiglia, in particolare a mia mamma Rosanna, che nonostante la sua assenza fisica, ha saputo offrirmi il supporto psicologico ed emotivo, che mi ha dato la forza di giungere al termine di questo corso di laurea. Un'infinità di ringraziamenti alle mie coinquiline e sorelle, Giorgia, Angela e Carmela che non hanno esitato ad essere presenti nei diversi momenti di difficoltà, convincendomi a non mollare; insieme a loro ho condiviso dei momenti memorabili di vita e di gioia, grazie a loro per tutte le sorprese e le risate che hanno saputo regalarmi e che sono state il trampolino di lancio per un nuovo inizio. Ringrazio anche mio fratello Natale, o meglio Natalino, con cui invece ho ironizzato a distanza sulle mie notti insonni pre-esame. Ringrazio mio padre Francesco, di cui ho avvertito la vicinanza, capisco che è anche grazie ai suoi sacrifici e a quelli di mia madre che ho potuto permettermi gli studi universitari.
Infine, un enorme grazie va ai miei amici di facoltà, Chiara, Marta, Sonia, Eleonora, Federico, Walid, Sofia, Anna, per i momenti di ilarità e di svago che abbiamo condiviso, ma anche per le giornate di studio che abbiamo trascorso sostenendoci vicendevolmente, attraverso lo scambio di idee, il ripasso dei contenuti, le correzioni reciproche; abbiamo combattuto insieme quella che abbiamo soprannominato durante tutto il corso di studio "depre" (depressione), e da cui siamo usciti vittoriosi.
Ho deciso di dedicare questa tesi alla mia famiglia e ai miei amici di facoltà, perchè ritengo siano le persone che con il supporto, ma anche la sopportazione delle mie lamentele, sono riuscite a convincermi ad addentrarmi in questa faticosa, ma avvincente, avventura che è la fisica.

Bibliografia

[1] Giulia Paci, Giampaolo Cristadoro, Barbara Monti, Marco Lenci, Mirko Degli Esposti, Gastone Castellani, Daniel Remondini, *Biological relevance of dinucleotide inter-distance distributions*, The Royal Society.

[2] Cecie Starr, *Biologia: Ereditarietà ed evoluzione*, Torino, 2002.