Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Matematica

Simmetria e Informazione Genomica: la Regola di Chargaff

Tesi di Laurea in Fisica Matematica

Relatore:
Chiar.mo Prof.
Mirko Degli Esposti

Correlatore:
Dott.
Giampaolo Cristadoro

Presentata da:
Alberto Carmagnini

III Sessione
Anno Accademico 2013/2014

Indice

Abstract

In questo lavoro analizzeremo la generalizzazione ad oligonucleotidi della seconda regola di Chargaff. Ripercorreremo gli approcci matematico-statistici più significativi per quantificare la simmetria reverse-complement all'interno di sequenze genomiche, presenteremo le prove della trasversalità di tale fenomeno e cercheremo di far luce sulle origini evolutive di questa simmetria nascosta nei nostri geni.

In this work we analyze the generalization of Chargaff's second parity rule for oligonucletides. We will follow the most significant mathematical-statistical approaches to quantify reverse-complement symmetry in genomic sequence. We will present evidence for the symmetry phenomenon's universality and we will try to explain the evolutionary origin of this symmetry hidden in our genes.

Capitolo 1

Introduzione

All'inizio degli anni cinquanta Erwin Chargaff e i suoi collaboratori notarono alcune regolarità nella distribuzione delle basi azotate che compongono il DNA e negli anni che seguirono, proseguendo tali studi, arrivarono a formulare quelle che vengono oggi comunemente indicate come le quattro regole di Chargaff enunciate di seguito:

1. **Chargaff first parity rule**
   Le percentuali di Adenina e Timina sono equivalenti nel DNA a doppio filamento così come lo sono quelle di Guanina e Citosina (%A=%T, %G=%C) [11].

2. **Chargaff second parity rule**
   Le percentuali di Guanina e Citosina sono essenzialmente equivalenti tra loro anche nel singolo filamento di DNA e un discorso del tutto analogo vale per Adenina e Timina [31].

3. **The cluster rule**
   Circa il 60% delle pirimidine (T e C) ricorre all'interno di brevi tratti oligonucleotidici e analogo discorso, in virtù delle regole di appaiamento tra i due filamenti, si può fare per le purine [9].

4. **CG rule**
   Il rapporto tra la quantità di Guanina e Citosina (G+C) rispetto alla totalità delle basi (A+T+C+G) è un invariante specie-specifico [10] [8].

Le prime tre regole risultarono essere specie-invarianti mentre la quarta tende ad essere costante per individui appartenenti alla stessa specie sebbene i valori assunti varino da una specie all'altra. La prima regola di appaiamento (CP_I) è alla base del modello della doppia elica proposto da Watson e Crick nel celeberrimo articolo del 1953 [34] ed oltre ad essere stata verificata mediante centinaia di prove dirette e migliaia di indirette è l'unica che sottende interamente ad un principio strutturale legato alla natura chimica del DNA (vedi Appendice B) nonché l'unica ad essere "*incorporated into mainstream biology*" (Donald R. Forsdyke [17]). Le altre regole invece, nonostante il progressivo accumularsi di prove a favore del loro carattere generale, non ottennero un immediato riconoscimento rimanendo ai margini del corpus della genetica molecolare in qualità di curiose osservazioni.

La creazione delle tre banche dati genetiche mondiali[1], quotidianamente interfacciate tra loro, ha reso le sequenze genomiche accessibili a qualunque ricercatore aprendo così le porte all'ingresso "prepotente" dell'informatica e della statistica nel campo della genetica molecolare. Via via che i dati dei sequenziamenti venivano pubblicati alcuni gruppi di ricerca hanno "riscoperto" le regole di Chargaff, in particolar modo la seconda regola (CP_{II}). I dati relativi alla composizione in basi dei genomi sequenziati avvalorano la sorprendente universalità di tale regola: dalla *Rickettsia* allo scimpanzè, dal lievito al mais passando per lo *Streptococcus pneumonie*, l'*Arabidopsis* nonché *Homo sapiens*, le sequenze genetiche di questi organismi verificano tutte CP_{II}.

Lo studio delle frequenze di occorrenza di brevi oligonucleotidi ha inoltre evidenziato un tipo di simmetria non banale presente in tutti i genomi e ha portato ad una generalizzazione della seconda regola di Chargaff, che indicheremo con la sigla CP_{II}^{oligo} e che rappresenta l'argomento centrale di questo lavoro.

Nel prossimo capitolo cercheremo di mostrare in cosa consista questa simmetria che tutti i genomi esibiscono e quale sia il suo significato funzionale. Il capitolo 3 ha invece carattere prettamente matematico-statistico e ha l'obiettivo di presentare gli strumenti più significativi per quantificare il fenomeno della simmetria. Infine nel capitolo 4 indagheremo il rapporto tra simmetria locale e simmetria globale, presentando un approccio

[1] GenBank (NCBI, Bethesda, MD, USA), EMBL (European Nucleotide Archive, Cambridge, UK), DDBJ (DNA Data Bank of Japan, Mishima, Japan)

in grado di restituire una rappresentazione visiva e immediatamente interpretabile della simmetria di sequenza all'interno di un cromosoma.
Per non frammentare eccessivamente l'esposizione si è scelto di corredare la tesi di tre capitoli di appendice che hanno lo scopo di fornire i riferimenti necessari ad una più completa comprensione degli argomenti trattati.
L'architettura della tesi e le modalità con cui verranno esposti gli argomenti risentono innegabilmente del percorso formativo di chi scrive che, prima di cimentarsi con la Matematica, si è laureato in Biotecnologie presso l'Università degli Studi di Firenze. Sebbene quindi i riferimenti alla biologia molecolare e alla genetica costituiscano una parte importante della trattazione, si è cercato di privilegiare gli aspetti più strettamente matematici legati allo studio di CP_{II}^{oligo} e contemporaneamente mostrare come attraverso lo studio della simmetria si possa giungere a considerazioni di carattere generale che hanno un profondo significato evolutivo.

Capitolo 2

Simmetrie di sequenza e il loro ruolo funzionale

Inizialmente scoperta da Chargaff nel 1968 [31] analizzando la composizione in basi di ciascun filamento di DNA costituente il genoma di *Bacillus subtilis*, mediante cromatografia su carta e dunque ben prima che fossero disponibili le tecniche per il sequenziamento genomico, la seconda regola di appaiamento suggerisce la presenza di una certa simmetria nella distribuzione dei nucleotidi all'interno del singolo filamento di DNA.
Indicando con f_N il numero di occorrenze di un determinato nucleotide all'interno di un campione di DNA, possiamo esprimere CP_{II} in termini di frequenze empiriche nel modo seguente:

$$f_A \approx f_T \qquad f_C \approx f_G \quad .$$

Andando alla ricerca di un modello in grado di spiegare l'origine e l'universalità di CP_{II}, si è cominciato a studiare le distribuzioni di frequenza di brevi oligomeri.
Se i genomi fossero frutto di un processo stocastico di tipo bernoulliano in cui la scelta di ciascun nucleotide che compone la sequenza è indipendente dalle altre (vedi Appendice A), potremmo associare all' evento di riscontrare una certa base all'interno del campione genomico, un valore di probabilità. Se volessimo poi che tale filamento di DNA verificasse CP_{II} dovremmo porre alcune condizioni sui valori di probabilità associati a ciascuna base, vale a dire:

$$P(A) = P(T) \qquad P(C) = P(G).$$

Siano dunque $p = P(A) = P(T)$ e $q = P(C) = P(G)$ valori di probabilità fissati, avremo che la probabilità di riscontrare una sequenza ω lunga k in un campione genomico sufficientemente grande sarà data dal prodotto delle probabilità relative ai nucleotidi che la compongono, ovvero:

$$P(\omega) = p^m q^{k-m} \quad \text{dove m = numero di A oppure T presenti in } \omega \text{ e } q = 1 - p$$

In questo caso, le sequenze contenenti esattamente m A o T sarebbero tutte equiprobabili e dunque tali oligonucleotidi dovrebbero essere identicamente distribuiti. Le sequenze genomiche, invece, non sembrano comportarsi in questo modo ma, dagli studi di cui ci occuperemo nei paragrafi seguenti, sono emerse evidenti correlazioni nelle distribuzioni di determinate coppie di sequenze. Prima però di discutere in dettaglio i risultati e i possibili meccanismi biologici alla base di tale fenomeno, è necessario definire preliminarmente alcune relazioni di simmetria tra sequenze della stessa lunghezza.

2.1 Coniugati simmetrici

In virtù della sua struttura chimica (vedi Appendice B) possiamo naturalmente rappresentare l'informazione contenuta nel DNA attraverso sequenze di simboli utilizzando come alfabeto le iniziali delle basi azotate.
Sia quindi $\mathcal{A} = \{A, T, C, G\}$ l'alfabeto del DNA. L'insieme di tutte le possibili sequenze di DNA a singolo filamento composte da k nucleotidi con $k \in \mathbb{N}$ fissato, sarà:

$$\mathcal{A}^k = \{(\omega_1, \ldots, \omega_k) \,|\, \omega_j \in \mathcal{A} \;\forall\, j = 1, \ldots, k\}$$

e ovviamente $card\, \mathcal{A}^k = 4^k$.
Su tale insieme possiamo definire due applicazioni interne biettive:

Reverse symmetry:

$$\rho\colon \omega = (\omega_1, \omega_2, \ldots, \omega_k) \to \omega^{-1} = (\omega_k, \omega_{k-1}, \ldots, \omega_1)$$

Complement symmetry:

$$\delta : \omega = (\omega_1, \omega_2, \ldots, \omega_k) \to \bar{\omega} = (\bar{\omega_1}, \bar{\omega_2}, \ldots, \bar{\omega_k})$$

dove $\bar{\omega_i}$ è il complementare di ω_i secondo le regole di appaiamento di Watson e Crick.
A partire da tali applicazioni è possibile definirne una terza mediante composizione:

Reverse - Complement symmetry:

$$\varphi = (\delta \circ \rho) = (\rho \circ \delta) : \omega = (\omega_1, \omega_2, \ldots, \omega_k) \rightarrow \omega^* = (\bar{\omega}_k, \bar{\omega}_{k-1}, \ldots, \bar{\omega}_1)$$

Chiameremo rispettivamente le coppie (ω, ω^{-1}), $(\omega, \bar{\omega})$, (ω, ω^*), reverse conjugate, complement conjugate e reverse-complement conjugate.
Ad esempio posto $\omega = (ATGC)$ avremo:

$$\rho(\omega) = \omega^{-1} = (CGTA) \quad ; \quad \delta(\omega) = \bar{\omega} = (TACG) \quad ; \quad \varphi(\omega) = \omega^* = (GCAT).$$

2.2 Chargaff second parity rule for oligonucleotides

L'articolo di Prabhu, pubblicato su Nucleic Acids Research nel 1993 [30], riporta il conteggio in basi di tutte le sequenze genomiche allora disponibili in GenBank e rappresenta il punto di partenza per tutti le ricerche successive riguardanti CP_{II}. Tale studio, non solo fornisce le prime prove riguardanti l'universalità della seconda regola di Chargaff ma ne estende la portata. Prabhu si concentrò infatti oltre che sulla composizione in basi anche sulle frequenze di dimeri e trimeri, riscontrate in 4 campioni genomici (verificanti CP_{II} e di lunghezza paragonabile) provenienti da taxa evolutivamente molto distanti tra loro. Notò che tutte le coppie di coniugati reverse-complement avevano frequenze inaspettatamente molto simili in ogni campione analizzato, diversamente da quanto accadeva comparando le frequenze di reverse conjugate o complement conjugate. Decise quindi di rappresentare le occorrenze di coppie reverse-complement come punti del piano e, effettuando una interpolazione mediante metodo dei minimi quadrati (vedi Appendice A), ottenne una retta con pendenza molto vicina ad uno in ogni campione.
Queste osservazioni hanno portato a generalizzare ad oligonucleotidi la seconda regola di Chargaff che, come anticipato nell'introduzione, indicheremo con la sigla CP_{II}^{oligo} e che enunciamo di seguito:
Scelta a piacere una breve sequenza di basi, il numero di copie di tale sequenza all'interno

di un singolo filamento di DNA o RNA sufficientemente grande, è approssimativamente equivalente al numero di copie della sequenza complementare letta in ordine inverso. Espressa in termini di frequenze empiriche e coniugati simmetrici diventa:

$$f_{\omega=(\omega_1,\ldots,\omega_k)} \approx f_{\omega^*=(\bar{\omega}_k,\ldots,\bar{\omega}_1)} \tag{2.1}$$

Negli anni successivi alla pubblicazione di Prabhu, molti altri gruppi di ricerca (ad esempio: [1] [4] [5] [24]) hanno esplorato un numero sempre maggiore di dati genomici, confermando la sostanziale validità di CP_{II}^{oligo} per genomi di organismi eucarioti, eubatteri ed archeobatteri nonché per molti genomi virali. Nel 2006 Mitchell e Bridge hanno riscontrato le prime eccezioni a questa regola in genomi mitocondriali e, come suggerito da Nikolaou e Almirantis, tale deviazione potrebbe essere connessa al particolare meccanismo di replicazione caratteristico di questi organelli cellulari [29].

2.3 Simmetria reverse-complement a livello locale: strutture stem-loop

CP_{II}^{oligo} ha certamente catturato l'attenzione di molti gruppi di ricerca che, con approcci spesso differenti, hanno cercato da un lato di misurare la reverse-complement symmetry (di cui ci occuperemo nel prossimo capitolo), dall'altro di rivelare quali fossero i meccanismi evolutivi che sottendono alla regola.
Una teoria affascinante e ormai universalmente accettata in biologia sull'origine della vita prevede che i ruoli di codifica e utilizzo funzionalmente attivo dell'informazione (oggi separati e ricoperti rispettivamente da DNA e proteine) fossero, in origine, entrambi appannaggio di un'unica classe di molecole: l'RNA [14] [26]. Dunque, seguendo questo ragionamento, anche l'origine di CP_{II}^{oligo} sarebbe da ricercare nella struttura di questa classe di acidi nucleici.
Numerosi studi di cristallografia hanno dimostrato che i t-RNA, i complessi nucleoproteici e i ribozimi (vedi Appendice B) raggiungono la loro struttura terziaria funzionalmente attiva attraverso strutture secondarie denominate stem-loop o strutture "a forcina" [21]. La presenza di strutture secondarie "a forcina" nell'RNA gioca un ruolo fondamentale

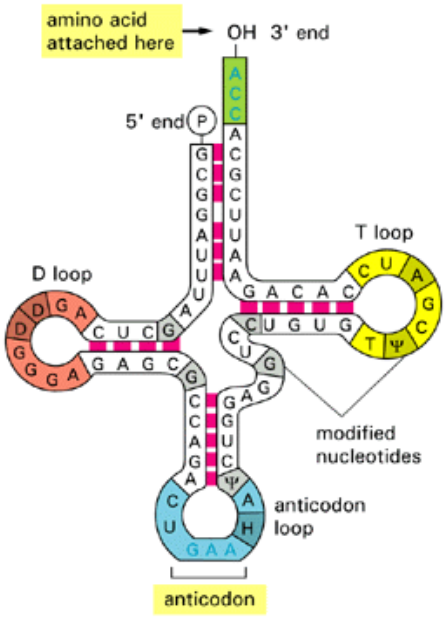

Figura 2.1: Rappresentazione di un tipico t-RNA e relative strutture stem-loop

anche nella regolazione dell'espressione genica sia nei procarioti (come ad esempio in alcuni terminatori[1] ρ indipendenti) sia negli eucarioti: è questo il caso di introni capaci di autosplicing [33].

Strutture secondarie di tipo stem-loop possono originarsi solo quando due coniugati reverse-complement si trovano relativamente vicini tra loro ovvero quando la sequenza è dotata di una certa simmetria interna e forma nel complesso una sorta di palindromo reverse-complement (vedi figura 2.1).

Poiché il trascrittoma[2] è diretta espressione del genoma di un organismo, in quanto le sequenze di RNA sono copie di sequenze di DNA a singolo filamento, se le molecole di RNA sono dotate di questo tipo di simmetria interna che permette loro di ripiegarsi e minimizzare la propria energia mediante formazione di legami idrogeno intracatena, questa proprietà deve essere condivisa anche dal filamento di DNA di cui l'RNA è copia. Gli studi di Forsdyke hanno dimostrato non solo che le sequenze di DNA introniche

[1] Un terminatore è una sequenza in grado di bloccare la trascrizione di un gene

[2] L'insieme delle molecole di RNA presenti in una cellula.

(quindi trascritte in RNA ma non tradotte in proteine) risultano fortemente conservate in caso di selezione darwiniana positiva [18] [19] ma anche che esse sono primariamente responsabili della formazione di stem-loop: ciò che in esse viene conservato è un potenziale di folding. Grazie alla messa a punto del programma di simulazione FORS-D, si è potuto mostrare come il potenziale di folding sia una proprietà diffusa in tutto il genoma e non solo relativa alle sequenze trascritte [20]. Questo potenziale sarebbe stato selezionato nel corso dell'evoluzione poiché la capacità di estrudere stem-loop dalla doppia elica [28] e le interazioni tra loop[3] avrebbero favorito il processo di riconoscimento dei cromosomi omologhi durante la meiosi [17]. Dunque, secondo questa teoria, la validità di CP_{II}^{oligo} a livello genomico sarebbe il risultato di innumerevoli eventi di simmetria locale. Nel loro articolo del 1999 [5] infatti Forsdyke e Bell scrivono:

> "Thus, base pairing in stems provides one possible level of accounting, which would be localized to the region of stem-loop extrusion. It seems unlikely that this relatively short range process could alone explain the precision of single-strand accounting. Base pairing between complementary loops (Tomizawa, 1984; Eguchi et al., 1991), which might occur very efficiently between cis-oriented sequences within one chromosome (Jinks-Robertson et al., 1993), and might operate over long genomic distances (Engels et al., 1994; Henikoff, 1997), might provide another level of accounting. Chargaff's second rule might apply to long genomic segments because of the summation of underlying primary accounting processes involving both stems (short-range accounting) and loops (long-range accounting)."

2.4 Simmetria reverse-complement a livello globale: inversioni ed elementi trasponibili

La posizione di Forsdyke e Bell non è però universalmente condivisa. Sebbene non vi siano dubbi sia sull'importanza funzionale delle strutture stem-loop a livello di trascrittoma, sia sulla possibilità che queste vengano conservate mediante selezione naturale a

[3] Interazione analoga a quella tra l'anticodone dei vari t-RNA e l'm-RNA durante la traduzione: da un punto di vista chimico si tratta di legami idrogeno.

livello genomico, i fenomeni di intrastrand base paring non sembrano poter spiegare da soli CP_{II}^{oligo}. La prima obiezione (immediata ma più debole) riguarda soprattutto gli organismi con bassa densità genica (eucarioti superiori): sebbene il potenziale di folding possa essere diffuso su tutto il genoma, l'effettiva e comprovata formazione di stem-loop funzionalmente attivi riguarda le regioni codificanti o comunque quelle trascritte che costituiscono solo una frazione del patrimonio genetico dell'organismo[4]. Ciò però non esclude che sebbene non siano sufficientemente caratterizzate, tali strutture non si formino anche nelle regioni non trascritte. La seconda invece si basa sul fatto che ciascuna struttura di tipo stem-loop coinvolge poche decine di basi e quindi questo modello non può spiegare simmetrie di sequenze più lunghe. Ad esempio, un recente lavoro di Zhang e Huang su 90 genomi procarioti (genomi ad alta densità genica e tutti verificanti CP_{II}^{oligo}) mostra come il contributo del potenziale di folding alla formazione e al mantenimento della simmetria all'interno del singolo filamento risulti piuttosto limitato. I due ricercatori hanno infatti calcolato la percentuale di sequenze che si trovano a meno di 25 basi di distanza dal proprio reverse-complement conjugate riscontrando un valore medio superiore al 90% per tetranucleotidi, tra il 55% e il 46% per esanucleotidi e solo dell' 1% per decanucleotidi [36].

La scoperta di elementi trasponibili [27] e di lunghe sequenze ripetute quali Alu, Sine e Line presenti in maniera più o meno rappresentativa in molti genomi [25] ha indotto studiosi come Fickett e Baisnée ad indicare invece i fenomeni ricombinativi (vedi Appendice B) quale possibile origine di CP_{II}^{oligo} [4] [16]. Il lavoro di Guenter Albrecht-Buehler del 2006 si inserisce in questa linea di pensiero e propone un modello molto semplice ed elegante per spiegare l'origine di CP_{II}.

Il modello prevede che in origine i genomi non verificassero necessariamente la seconda regola di Chargaff e che dunque fossero presenti asimmetrie nella composizione in basi dei due filamenti di DNA costituenti la doppia elica (tradizionalmente denominati Watson strand e Crick strand). Supponiamo quindi che ad esempio ci fosse un eccesso di Citosine sul Watson strand, ne consegue un eccesso di Guanine sul Crick strand in virtù di CP_I. Sotto l'effetto di trasposizioni invertite in posizioni random del genoma, avremo che ognuno di questi eventi trasferisce alcune delle Citosine soprannumerarie dal Watson

[4]Nel caso del genoma umano costituiscono rispettivamente circa il 5% e 25%.

strand al Crick strand e contemporaneamente ha la stessa azione sulle Guanine sopran-numerarie che trasferisce invece dal Crick strand al Watson strand. Un ragionamento del tutto analogo può essere fatto per Adenina e Timina. Il processo è in pratica irreversibile e auto stabilizzante poiché una volta che un genoma raggiunge lo stadio in cui CP_{II} è verificata, qualsiasi altro evento di inversione non può alterare il livello di compliance di tale genoma alla regola.
Una descrizione quantitativa del fenomeno (nel caso più semplice e irrealistico possibile cioè quello di un genoma in cui il Watson strand è interamente composto da Citosine) è dato dalle seguenti equazioni:

$$f_{watson}(G)_n = \frac{f_{watson}(G)_0}{2}(1 + e^{-2kn}) + \frac{f_{watson}(C)_0}{2}(1 - e^{-2kn})$$
$$f_{watson}(C)_n = \frac{f_{watson}(G)_0}{2}(1 - e^{-2kn}) + \frac{f_{watson}(C)_0}{2}(1 + e^{-2kn})$$

dove n è il numero di inversioni, $f_{watson}(G)_0$ e $f_{watson}(C)_0$ sono rispettivamente il numero iniziale di Guanine e Citosine sul Watson strand mentre $k = \frac{\lambda}{L}$ è una misura di come cambia il filamento dopo ogni evento, è infatti calcolato sulla base della lunghezza media del frammento invertito (λ) e la lunghezza dell'intero genoma (L).
Il modello può anche essere generalizzato a casi più complessi seguendo essenzialmente lo stesso ragionamento. Nel caso specifico, Albrecht-Buehler costruisce una simulazione al computer in cui partendo da genomi che non verificano $CP_{II}^{triplets}$ si arriva, dopo un certo numero di inversioni, a genomi in cui ciascuna tripletta e la relativa reverse-complement conjugate hanno circa lo stesso numero di occorrenze.

> "Thus, the compliance with Chargaff's second parity rules may be interpreted as an inevitable, asymptotic product of (among other causes) numerous inversions and inverted transpositions that occurred in the course of evolution."
>
> (Albrecht-Buehler, 2006)

Capitolo 3

Misurare la simmetria

Nel capitolo precedente abbiamo introdotto la seconda regola di Chargaff (CP_{II}) relativa alla composizione in basi del singolo filamento di DNA che rappresenta il primo ordine di simmetria. Abbiamo anche generalizzato tale regola estendendola ad ordini superiori (CP_{II}^{oligo}) ovvero a coppie di sequenze di tipo reverse-complement conjugate, concentrandoci principalmente sulle possibili spiegazioni biologiche di tale fenomeno. In letteratura sono però altrettanto numerosi e significativi gli articoli in cui vari gruppi di ricerca hanno cercato invece di dare una valutazione quantitativa del fenomeno. Il punto di partenza naturale e comune a tutti gli studi è il conteggio delle frequenze di occorrenza di determinati oligonucleotidi sia in sequenze genomiche sia in campioni generati random. Nei paragrafi che seguono, ci concentreremo quindi sugli approcci e gli strumenti utilizzati dai ricercatori, mutuati dalla teoria delle probabilità e dalla statistica inferenziale o in alcuni casi appositamente costruiti, per "misurare" simmetrie di sequenza.

3.1 L'indice di simmetria S^1

Uno dei primi e maggiormente significativi lavori sull'argomento è quello di Baisnée, Hampson e Baldi pubblicato su Bioinformatics nel 2002. In questo articolo [4] gli autori hanno esaminato sequenze cromosomiali complete di vari organismi ed estratto le frequenze di tutti i possibili oligomeri di lunghezza $k = 1, \ldots, 9$ mediante un software creato appositamente. Il programma infatti crea una finestra sovrapponibile di lunghez-

za k, percorre il campione genomico spostandosi di una base alla volta, legge la sequenza di DNA racchiusa all'interno della finestra, la confronta con l'oligomero scelto e, in caso di identità, incrementa un contatore. Quando la finestra sovrapponibile raggiunge la fine del campione genomico, il valore assunto dal contatore rappresenta appunto la frequenza di occorrenza di quel particolare oligomero. Ovviamente tale procedimento è stato ripetuto per i 4^k possibili oligomeri e per ciascun campione.
Ottenute le frequenze empiriche, per misurare la simmetria di ordine k nei vari campioni di DNA, i ricercatori hanno misurato la similarità tra le distribuzioni di coniugati reverse-complement di lunghezza k attraverso il seguente indicatore di simmetria:

$$S^1 = 1 - \frac{\sum_i |f_i - f_i^*|}{\sum_i f_i + f_i^*} \tag{3.1}$$

dove f_i e f_i^* rappresentano rispettivamente le frequenze relative (espresse in percentuali) dell'oligomero ω_i e del suo reverse-complement conjugate ω_i^*.
Osserviamo esplicitamente che il denominatore vale due quando si considerano distribuzioni complete cioè quando vengono valutate le frequenze di tutti i possibili oligomeri per un certo k fissato. Ciò accade sempre se k è dispari mentre per k pari esistono $4^{\frac{k}{2}}$ oligomeri identici al loro reverse-complement conjugate e quindi si potrebbe decidere di escludere tali elementi auto-simmetrici dall'analisi.
Concentriamoci adesso invece sul numeratore dell'indice di simmetria. Per semplicità, supponiamo che la differenza tra le frequenze di coppie reverse-complement, riscontrate in un dato campione genomico, sia normalmente distribuita con media zero e varianza σ^2. Avremo dunque che il valore atteso della variabile aleatoria $|X| = |f - f^*|$ sarà dato dall'integrale su tutta la retta reale del prodotto tra la variabile e la sua densità di probabilità o più esplicitamente:

$$\begin{aligned} E(|X|) &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}|x|e^{\frac{-|x|^2}{2\sigma^2}}dx \\ &= \frac{1}{\sqrt{2\pi}\sigma}\left(\int_0^{+\infty}xe^{\frac{-x^2}{2\sigma^2}}dx+\int_{-\infty}^{0}-xe^{\frac{-x^2}{2\sigma^2}}dx\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\left(-\sigma^2\left[e^{\frac{-x^2}{2\sigma^2}}\right]_0^{+\infty}+\sigma^2\left[e^{\frac{-x^2}{2\sigma^2}}\right]_{-\infty}^{0}\right) \\ &= \frac{2\sigma}{\sqrt{2\pi}}. \end{aligned}$$

dove nella terza riga si è applicato il teorema fondamentale del calcolo integrale essendo $-\sigma^2\exp(\frac{-x^2}{2\sigma^2})$ una primitiva di $x\exp(\frac{-x^2}{2\sigma^2})$ che è continua e Lebesgue-integrabile.
Per la linearità della previsione e per 3.1 nel caso di distribuzioni complete avremo che il valore atteso dell'indice di simmetria è:

$$E(S^1)\approx 1-\frac{4^k}{\sqrt{2\pi}}\sigma \tag{3.2}$$

Dunque, sotto tali ipotesi, S^1 decresce al crescere dell'ordine di simmetria k rispetto al quale si effettuano le misurazioni.
Per valutare invece il comportamento dell'indice in funzione della grandezza dei dati genomici analizzati è necessario introdurre nelle ipotesi parametri che dipendano dalla lunghezza del campione (L). Intuitivamente possiamo dire che più grande sarà il valore di L maggiore sarà il numero di osservazioni effettuate e minore sarà la deviazione dalla media. L'esempio proposto dagli autori è il seguente:
Supponiamo di aver fissato k e di voler analizzare un campione genomico di lunghezza $L >> k$, supponiamo inoltre che le frequenze degli oligomeri di lunghezza k siano distribuite normalmente con media $\mu = 1/4^k$ e varianza $\sigma_f^2 = \frac{1-\mu}{4^k L}$. Ragionevolmente, la distribuzione della variabile $X = f - f^*$ sarà una normale con media zero e varianza $\sigma^2 = 2\sigma_f^2$ (vedi Appendice A).
Per quanto detto, da 3.2 otteniamo:

$$E(S^1)\approx 1-\frac{4^k}{\sqrt{2\pi}}\frac{\sqrt{2\left(4^k-1\right)}}{4^k\sqrt{L}}=1-\frac{\sqrt{4^k-1}}{\sqrt{\pi L}}. \tag{3.3}$$

Il cui valore chiaramente tende ad uno per $L \to \infty$. Possiamo quindi concludere che l'indicatore S^1 varia tra zero (assenza di simmetria) e uno (perfetta simmetria).
Grazie a questo strumento, i ricercatori sono stati in grado di provare l'universalità di CP_{II} in quanto hanno riscontrato valori di S^1 molto vicini ad uno in tutti i cromosomi eucariotici e procariotici analizzati così come in vari genomi virali mentre i genomi mitocondriali hanno mostrato un livello di simmetria inferiore alla tendenza media. Inoltre lo studio ha evidenziato come le regioni codificanti risultino leggermente più asimmetriche se comparate con quelle non codificanti. Infine, per quanto riguarda gli ordini superiori al primo, i valori di S^1 si mantengono sorprendentemente alti (e anche molto simili tra loro, a parità di lunghezza del campione) in tutti i genomi analizzati indipendentemente dal taxon a cui l'organismo appartiene, mentre tendono universalmente a diminuire all'aumentare dell'ordine $k = 2, \ldots, 9$ considerato nello studio. Ciò giustifica, come già riportato nel capitolo precedente, la generalizzazione della seconda regola di Chargaff ad oligonucleotidi (CP_{II}^{oligo}).
Questi risultati [4] sono stati essenzialmente confermati da studi successivi, portati avanti dal gruppo di ricerca portoghese dell' Università di Aveiro. In un recente articolo interamente incentrato sul genoma umano, oltre all'indice di simmetria S^1 e al coefficiente di correlazione di Pearson (vedi Appendice A), al fine di valutare il fenomeno della simmetria sono state utilizzate anche la divergenza di Kullback-Leibler (vedi Appendice C) ed una normalizzazione della metrica di Ulam[1] definita dagli autori "word symmetry distance"[1].

3.2 Partizione in m-set

Indagare il fenomeno della simmetria in sequenze genomiche presenta diversi tipi di difficoltà dovute sia alla grandezza dei dati sia ad alcuni limiti intrinseci degli strumenti statistici utilizzati. Uno dei limiti dell'indice di simmetria proposto da Baisnée, Hampson e Baldi è quello di restituire valori prossimi all'unità anche per campioni non genomici

[1]Date due sequenze A_1 e A_2 composte dallo stesso numero di simboli appartenenti al medesimo alfabeto, la distanza tra le due sequenze secondo la metrica di Ulam è definita come il numero minimo di spostamenti di simboli necessari a rendere identiche le due sequenze.

ma costruiti mediante un processo stocastico verificante CP_{II} [22].
Nel tentativo di rintracciare, attraverso lo studio della simmetria, le impronte di come i genomi si siano evoluti e in cosa essi si distinguano da sequenze frutto di un processo stocastico, seguiremo l'approccio proposto da un gruppo di ricerca taiwanese che, al fine di costruire nuovi indicatori statistici, ha introdotto una intelligente partizione dell'insieme $\mathcal{A}^k$, sulla base del numero di Adenine e Timine presenti in ciascun oligomero.
La scelta di considerare il numero di Adenine e Timine di una sequenza ha radici termodinamiche in quanto, quando queste due basi si appaiano, formano solo due legami idrogeno laddove Guanina e Citosina ne formano tre (vedi Appendice B). Per questo motivo Adenina e Timina sono dette basi *weak* (Wb) mentre Guanina e Citosina vengono definite *strong* (Sb) in quanto contribuiscono maggiormente alla stabilità della molecola.
Utilizzando la notazione degli autori [12] [13] [22], sia $S = \mathcal{A}^k$ l'insieme delle sequenze lunghe k con $k \in \mathbb{N}$ fissato. Come già visto nel capitolo precedente avremo che $card\ S = 4^k = \tau$. Si possono suddividere i τ elementi di S in $k+1$ classi d'equivalenza (indicate con S_m) nel modo seguente:

$$S_m \subset S \qquad \forall m = 0, \cdots, k$$

$$\bigcup_{m=0}^{k} S_m = S$$

$$S_i \cap S_j = \varnothing \qquad \forall i \neq j$$

$$\omega \in S_m, \Leftrightarrow \omega \text{ è una sequenza lunga k contenente esattamente } m \text{ Wb.}$$

La cardinalità di ciascun elemento della partizione è determinata mediante semplici calcoli combinatori in quanto i modi con cui posso disporre m elementi di tipo Wb e $k-m$ elementi di tipo Sb tenendo conto dell'ordinamento sono esattamente $\binom{k}{m}$. Inoltre, considerando che per ognuna delle k posizioni all'interno della sequenza abbiamo sempre a disposizione due possibilità (A o T nel caso di elementi di tipo Wb e C o G nel caso di elementi di tipo Sb) avremo che:

$$card\ S_m = 2^k \binom{k}{m} = \tau_m$$

mentre utilizzando lo sviluppo del binomio di Newton si trae

$$\sum_{m=0}^{k} \tau_m = 2^k \sum_{m=0}^{k} \binom{k}{m} = 2^k \sum_{m=0}^{k} \binom{k}{m} 1^m \, 1^{k-m} = 2^k (1+1)^k = \tau.$$

A titolo d'esempio consideriamo una sequenza nucleotidica composta da tre basi, avremo quindi $k = 3$ e $\tau = 64$. Dalle formule precedenti ci aspettiamo una partizione composta da quattro sottoinsiemi di S che andiamo di seguito ad elencare:

$$S_0 = \{(\text{CCC}),(\text{CCG}),(\text{CGC}),(\text{CGG}),(\text{GCC}),(\text{GCG}),(\text{GGC}),(\text{GGG})\}$$
$$card\ S_0 = 2^3 \binom{3}{0} = 8 = \tau_0$$

$$S_1 = \{(\text{CCA}),(\text{CCT}),(\text{CAC}),(\text{CTC}),(\text{ACC}),(\text{TCC}),(\text{GGA}),(\text{GGT})\} \cup$$
$$\cup \{(\text{GAG}),(\text{GTG}),(\text{AGG}),(\text{TGG}),(\text{CGA}),(\text{CGT}),(\text{GCA}),(\text{GCT})\} \cup$$
$$\cup \{(\text{GAC}),(\text{GTC}),(\text{CAG}),(\text{CTG}),(\text{AGC}),(\text{TGC}),(\text{ACG}),(\text{TCG})\}$$
$$card\ S_1 = 2^3 \binom{3}{1} = 24 = \tau_1$$

$$S_2 = \{(\text{CAA}),(\text{CAT}),(\text{CTA}),(\text{CTT}),(\text{GAA}),(\text{GAT}),(\text{GTA}),(\text{GTT})\} \cup$$
$$\cup \{(\text{ACA}),(\text{ACT}),(\text{TCA}),(\text{TCT}),(\text{AGA}),(\text{AGT}),(\text{TGA}),(\text{TGT})\} \cup$$
$$\cup \{(\text{AAC}),(\text{ATC}),(\text{TAC}),(\text{TTC}),(\text{AAG}),(\text{ATG}),(\text{TAG}),(\text{TTG})\}$$
$$card\ S_2 = 2^3 \binom{3}{2} = 24 = \tau_2$$

$$S_3 = \{(\text{AAA}),(\text{AAT}),(\text{ATA}),(\text{TAA}),(\text{TTA}),(\text{TAT}),(\text{ATT}),(\text{TTT})\}$$
$$card\ S_3 = 2^3 \binom{3}{3} = 8 = \tau_3$$

Osserviamo esplicitamente che, per ogni sequenza $\omega \in S_m$, tutti i suoi coniugati (reverse, complement o reverse-complement) appartengono allo stesso elemento S_m della

partizione. In altri termini possiamo dire che l' immagine di S_m mediante ciascuna delle tre applicazioni di insiemi definite nel capitolo precedente è proprio S_m

$$\rho(S_m) = S_m \quad ; \quad \delta(S_m) = S_m \quad ; \quad \varphi(S_m) = S_m$$

ed essendo ρ, δ e φ biettive, tali sono le rispettive restrizioni ad S_m vale a dire che i sottoinsiemi della partizione vengono lasciati invariati dalle tre applicazioni di simmetria. Introduciamo adesso altri elementi notazionali che ci saranno utili nel corso della trattazione. Poniamo:

$$\begin{aligned}
f_\omega &:= \text{la frequenza di occorrenza di un certo oligomero } \omega \text{ di lunghezza } k \\
\ddot{f} &:= \text{la frequenza media degli oligomeri lunghi } k \\
\ddot{f}_m &:= \text{la frequenza media degli oligomeri lunghi } k \text{ all' interno di un m-set} \\
L &:= \text{la lunghezza del campione genetico espressa in unità nucleotidiche} \\
p &:= \frac{A+T}{L} \text{ la composizione relativa di Wb nel campione} \\
q &:= 1-p \quad \text{la composizione relativa di Sb nel campione}
\end{aligned}$$

Allora, per ogni k fissato, avremo:

$$\sum_{\omega \in S} f_\omega = L - k + 1 \approx L \qquad \text{mentre} \qquad \sum_{\omega \in S_m} f_\omega = L_m$$

$$\ddot{f} = \frac{L}{\tau} \qquad \text{ed analogamente} \qquad \ddot{f}_m = \frac{L_m}{\tau_m}$$

Utilizzando ancora una volta l'espansione binomiale possiamo parametrizzare L nel seguente modo:

$$L = \tau \ddot{f} = \tau \ddot{f}(p+q)^k = \sum_{m=0}^{k} \left(2^k \binom{k}{m}\right) \left(2^k p^m q^{k-m} \ddot{f}\right) = \sum_{m=0}^{k} \tau_m \left(2^k p^m q^{k-m} \frac{L}{\tau}\right)$$

Conveniamo di porre:

$$\ddot{f}_m^{\infty} := \lim_{L \to \infty} \left(2^k p^m q^{k-m} \frac{L}{\tau}\right) \tag{3.4}$$

che fornisce una stima del valore di $\ddot{f}_m$ per campioni random la cui lunghezza L è di diversi ordini di grandezza superiore alla lunghezza k degli oligomeri di cui misuriamo le frequenze di occorrenza [13].

3.3 Il coefficiente di variazione

La partizione in m-set proposta da Hong-Da Chen e collaboratori, ha il pregio di raggruppare gli oligomeri in classi di equiprobabilità stocastica e ha permesso ai ricercatori di costruire un indicatore, denominato coefficiente di variazione (CV), mediante tecniche di analisi della varianza. Questa tecnica, nota in letteratura con l'acronimo ANOVA, prevede di separare i contributi statistici della varianza interna alle classi da quelli della varianza tra le classi come mostra il ragionamento seguente:

$$\begin{aligned}
\sigma^2 &= \tau^{-1} \sum_{\omega \in S} \left(f_\omega - \ddot{f} \right)^2 \\
&= \tau^{-1} \sum_{\omega \in S} \left[\left(f_\omega - \ddot{f}_m \right) + \left(\ddot{f}_m - \ddot{f} \right) \right]^2 \\
&= \tau^{-1} \sum_{m=0}^{k} \sum_{\omega \in S_m} \left[\left(f_\omega - \ddot{f}_m \right)^2 + \left(\ddot{f}_m - \ddot{f} \right)^2 + 2 \left(f_\omega - \ddot{f}_m \right) \left(\ddot{f}_m - \ddot{f} \right) \right] \\
&= \sum_{m=0}^{k} \frac{\tau_m}{\tau} \left(\ddot{f}_m - \ddot{f} \right)^2 + \sum_{m=0}^{k} \frac{\tau_m}{\tau} \sum_{\omega \in S_m} \left(f_\omega - \ddot{f}_m \right)^2 \tau_m^{-1} \\
&= \sum_{m=0}^{k} \frac{\tau_m}{\tau} \left(\ddot{f}_m - \ddot{f} \right)^2 + \sum_{m=0}^{k} \frac{\tau_m}{\tau} \sigma_m^2
\end{aligned}$$

Quindi abbiamo espresso la varianza totale come somma di due contributi che indicheremo con la seguente notazione [13]:

$$\sigma_{nf}^2 = \sum_{m=0}^{k} \frac{\tau_m}{\tau} \left(\ddot{f}_m - \ddot{f} \right)^2 \tag{3.5}$$

rappresenta la varianza tra gli m-set (non-fluttuante) detta anche varianza "between", mentre

$$\sigma_{fl}^2 = \sum_{m=0}^{k} \frac{\tau_m}{\tau} \sigma_m^2 \tag{3.6}$$

è la media ponderata delle varianze parziali calcolate in ciascun m-set e rappresenta la parte definita fluttuante, detta anche varianza "within".
Seguendo questo procedimento, possiamo ora definire l'indicatore CV:

$$CV^2 = \left(\frac{\sigma}{\ddot{f}} \right)^2 = \frac{\sigma_{nf}^2}{\ddot{f}^2} + \frac{\sigma_{fl}^2}{\ddot{f}^2} = CV_{nf}^2 + CV_{fl}^2 \tag{3.7}$$

Le due componenti di CV^2 hanno comportamenti statistici differenti quando L è molto grande ed inoltre dipendono in modo diverso dal parametro p. Ciò permette, come mostra il ragionamento seguente, di utilizzare CV^2_{fl} per evidenziare caratteristiche proprie di sequenze genomiche che le sequenze random non possiedono.
Supponiamo di costruire una sequenza random di lunghezza paragonabile ad una sequenza genomica e di voler valutare il comportamento del coefficiente di variazione per L molto grande. Sostituendo in 3.7 quanto visto in 3.5 e 3.6 e passando al limite otteniamo:

$$(CV_\infty)^2 := \lim_{L\to\infty} \ddot{f}^{-2} \sum_{m=0}^{k} \frac{\tau_m}{\tau} \left(\ddot{f}_m - \ddot{f}\right)^2 + \lim_{L\to\infty} \ddot{f}^{-2} \sum_{m=0}^{k} \frac{\tau_m}{\tau}\, \sigma_m^2$$

Occupiamoci preliminarmente di CV^2_{nf}. Sostituendo quanto mostrato in 3.4 otteniamo:

$$\begin{aligned}
\left(CV^\infty_{nf}\right)^2 &:= \lim_{L\to\infty} \ddot{f}^{-2} \sum_{m=0}^{k} \frac{\tau_m}{\tau} \left(2^k p^m q^{k-m} \ddot{f} - \ddot{f}\right)^2 \\
&= \lim_{L\to\infty} \sum_{m=0}^{k} \frac{\tau_m}{\tau} \left(2^k p^m q^{k-m} - 1\right)^2 \\
&= \sum_{m=0}^{k} 2^k \binom{k}{m} \left(p^m q^{k-m} - 2^{-k}\right)^2 \\
&= \sum_{m=0}^{k} 2^k \binom{k}{m} p^{2m} q^{2(k-m)} - 2 \sum_{m=0}^{k} \binom{k}{m} p^m q^{k-m} + \sum_{m=0}^{k} \binom{k}{m} 2^{-k} \\
&= 2^k \left(p^2 + q^2\right)^k - 2(p+q)^k + 1 \\
&= 2^k \left(p^2 + q^2\right)^k - 1
\end{aligned}$$

Osserviamo esplicitamente che $\left(CV^\infty_{nf}\right)^2$ non dipende direttamente né da L né da m e si annulla per $p = 0,5$.
Per mostrare il comportamento di CV^2_{fl} e dunque risolvere il secondo dei limiti proposti distinguiamo due casi:

1. Per ($p = 0,5$), L molto grande e k piccolo, tutti gli oligomeri di lunghezza k sono equiprobabili. Possiamo allora approssimare la distribuzione delle frequenze con una distribuzione di Poisson caratterizzata dall'avere media pari alla varianza (vedi Appendice A). Dunque $CV^2 = \frac{\sigma^2}{\ddot{f}^2} = \ddot{f}^{-1} = \frac{\tau}{L}$ che tende a zero per $L \to \infty$.

Da questo risultato e dall'osservazione precedente deduciamo che anche CV_{fl}^2 tende a zero per $L \to \infty$.

2. Per ($p \neq 0,5$), all'interno di ciascun m-set, gli oligomeri sono equiprobabili dunque possiamo approssimare la distribuzione di frequenza in maniera analoga, sostituendo σ_m^2 con $\ddot{f}_m = \frac{L_m}{\tau_m}$ ottenendo:

$$\left(CV_{fl}^{\infty}\right)^2 = \lim_{L\to\infty} \frac{1}{\ddot{f}^2} \sum_{m=0}^{k} \frac{L_m}{\tau} = \lim_{L\to\infty} \frac{\tau}{L}$$

che tende a zero per L grande e non dipende da p.

Siccome CV_{fl} decresce all'aumentare della lunghezza del campione mentre CV_{nf} non si comporta in questo modo, esisterà un valore di L (quando $p \neq 0,5$) oltre il quale CV_{nf} diventa la parte dominante del coefficiente di variazione. Ovviamente tale valore dipende da k e da p ma è stato calcolato essere di diversi ordini di grandezza inferiore rispetto alla lunghezza media di un cromosoma [13]. Isolare il contributo di CV_{fl} si è rivelato fondamentale per riconoscere una sequenza genomica da una generata mediante un processo stocastico. I ricercatori hanno infatti riscontrato valori di CV_{fl} per sequenze random di diversi ordini di grandezza inferiori rispetto a quanto accade per sequenze genomiche come mostra la figura seguente.

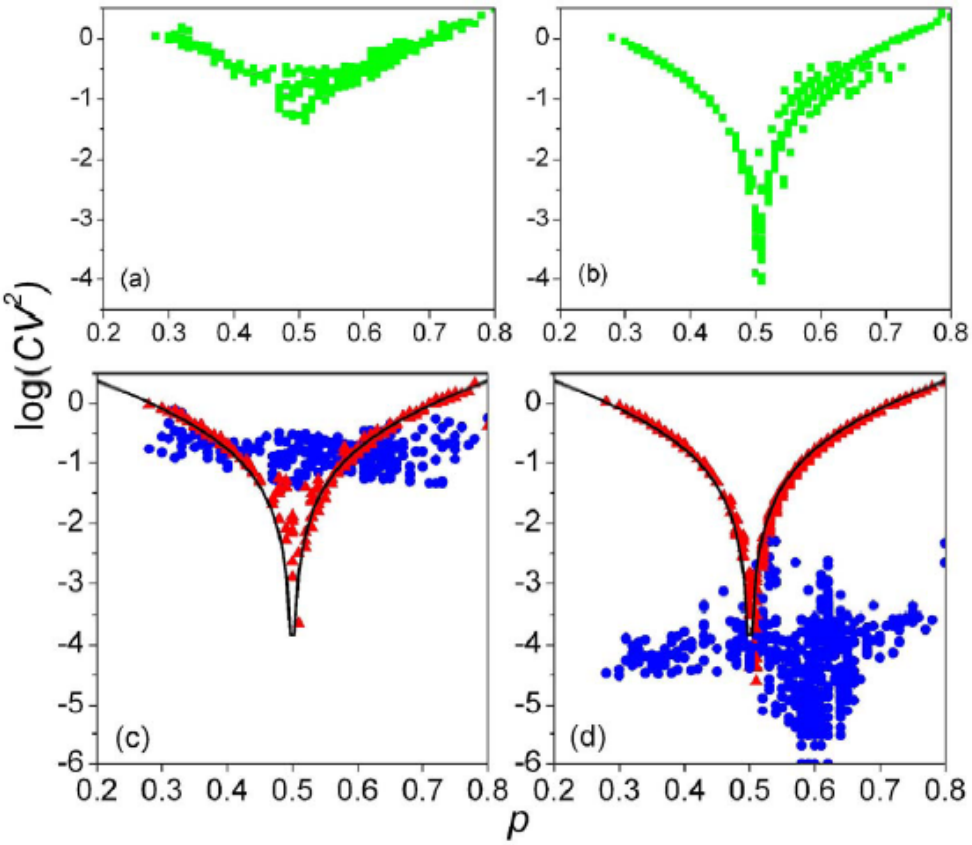

Figura 3.1: Dati di CV^2 su scala logaritmica in funzione del parametro p relativi alla distribuzione di dimeri di 865 sequenze genomiche (pannelli a sinistra) e alle altrettante controparti random di pari lunghezza (a destra). Nei pannelli (c) e (d) sono rappresentati in rosso i valori di CV^2_{nf} mentre in blu quelli di CV^2_{fl}. Figura estratta da [13].

Capitolo 4

Matrici di simmetria a livello cromosomico

Utilizzando la partizione in m-set, introdotta nel capitolo precedente, è possibile definire un nuovo indice di simmetria, la cui applicazione a livello cromosomico ha permesso ai ricercatori di evidenziare le impronte di come i genomi attuali siano stati accresciuti e modificati, nel corso dell'evoluzione, da meccanismi quali duplicazioni segmentali e trasposizioni invertite (vedi Appendice B).

4.1 L'indice di simmetria χ

A partire da $S = \mathcal{A}^k$ insieme delle sequenze lunghe k con $k \geq 2$ fissato, indichiamo con P l'insieme delle coppie non ordinate di reverse-complement conjugate, da cui escludiamo gli elementi auto-simmetrici, e con N la cardinalità di P.
Ad esempio per $k = 2$ avremo:

$$P = \{(\text{AA,TT}), (\text{CC,GG}), (\text{AC,GT}), (\text{AG,CT}), (\text{TG,CA}), (\text{TC,GA})\}$$

$$N = card\, P = 6$$

Utilizzando la notazione introdotta nel capitolo 3, possiamo definire l'indice di simmetria χ nel modo seguente [22]:

$$\chi^2 = \frac{1}{2N} \sum_{(\omega,\omega*)\in P} \left(\frac{f_\omega - f_{\omega*}}{\sigma_{m_\omega}} \right)^2 \tag{4.1}$$

dove, ovviamente, σ_{m_ω} rappresenta la deviazione standard dalla media delle frequenze di occorrenza all'interno dell'm-set a cui sia ω che $\omega*$ appartengono mentre χ è definito come la radice quadrata positiva dell'equazione 4.1.
Osserviamo esplicitamente che un valore di χ pari a zero indica perfetta simmetria mentre un valore prossimo all'unità ne indica l'assenza.
D'altra parte, se supponiamo che la differenza tra le frequenze di coppie reverse-complement sia normalmente distribuita con media zero e varianza σ^2, la distribuzione della variabile $X_i^2 = |f_{\omega_i} - f_{\omega_i*}|^2$ è proprio una distribuzione χ^2 di parametro 1 (vedi Appendice A).
Proseguendo il ragionamento, se consideriamo tutte le possibili coppie non ordinate di sequenze lunghe k del tipo reverse-complement conjugate e non auto-simmetriche, possiamo definire una variabile $Y = \sum_{i=1}^{N} X_i^2$ che avrà distribuzione χ^2 di parametro N: il riferimento a questa distribuzione è reso evidente anche dalla scelta dei simboli utilizzati da Sing-Guan Kong e collaboratori per il loro indice di simmetria.
L'utilizzo della partizione in m-set ha, in questo caso, lo scopo di limitare l'effetto della composizione in basi sulla fluttuazione delle frequenze di occorrenza degli oligomeri: proprio per questo la differenza tra le frequenze di coppie reverse-complement conjugate è pesata secondo $\frac{1}{\sigma_{m_\omega}\sqrt{2}}$.
La comparazione dei valori ottenuti utilizzando l'indice di simmetria S^1 (vedi equazione 3.1) con quelli ottenuti mediante l'indice χ (vedi equazione 4.1) mostra come il secondo abbia un miglior potere risolutivo (vedi tabella 4.1). L'indice di simmetria χ rivela infatti la presenza di simmetria reverse-complement in campioni genomici mentre questa risulta assente in campioni generati random; l'indice S^1 ha invece valori prossimi all'unità in entrambi i casi. Questo non tanto perché la quantità $|f_i - f_i^*|$ sia particolarmente piccola ma piuttosto perché in una sequenza random, la differenza tra le frequenze di occorrenza di qualsiasi coppia di sequenze appartenenti allo stesso m-set è piccola. Ciò illustra l'importanza di misurare $|f_i - f_i^*|$ rispetto a σ_{m_ω}.

Symmetry Index	E. coli	Random	HS1	Random	k
S^1	0,9974	0,9991	0,9992	0,9996	2
χ	0,0345	1,1925	0,0093	1,4425	2
S^1	0,9965	0,9982	0,9992	0,9996	3
χ	0,0255	1.0602	0,0061	1,1587	3
S^1	0,9943	0,9963	0,9989	0,9993	4
χ	0,0307	0,9497	0,0065	1,1097	4
S^1	0,9905	0,9921	0,9984	0,9988	5
χ	0,0399	0,9706	0,0066	1,0207	5
S^1	0,9824	0,9846	0,9973	0,9976	6
χ	0,0611	0,9671	0,0091	1,0082	6

Tabella 4.1: Valori degli indici di simmetria reverse-complement S^1 e χ relativi al genoma di E.coli e al Cromosoma 1 umano. I campioni random hanno pari lunghezza e analoga composizione in basi rispetto alla loro controparte genomica mentre k indica la lunghezza degli oligomeri considerati [22].

4.2 Simmetria globale e locale

L'articolo di Sing-Guan Kong e collaboratori [22] riguardante 786 sequenze cromosomiche complete (356 cromosomi eubatterici, 28 appartenenti ad archeobatteri e 402 cromosomi eucariotici provenienti da 28 specie diverse) conferma sostanzialmente quanto già sapevamo: a livello cromosomico il fenomeno della simmetria reverse-complement è fortemente presente. Il calcolo dell'indice χ, infatti, restituisce valori inferiori a 10^{-1} per tutte le sequenze analizzate e per ogni $k = 2, \ldots, 6$ considerato nello studio. Inoltre, anche utilizzando questo indicatore di simmetria, si è registrata una maggiore deviazione da CP_{II}^{oligo} nelle regioni codificanti rispetto alle regioni introniche (presenti solo negli eucarioti) o intergeniche.

La vera forza del lavoro di Sing-Guan Kong e collaboratori sta nell'utilizzo sistematico dell'indice χ per indagare il fenomeno della simmetria a livello locale oltre che a livello globale. Seguendo la notazione degli autori, conveniamo di indicare con χ_{gl} l'indice di simmetria globale dell'intero cromosoma, con χ_l l'indice di simmetria di un segmento cromosomico lungo l e con $\ddot{\chi}_l$ il valore medio dell'indice di simmetria relativo a tutti i segmenti non sovrapposti di lunghezza l in cui si suddivide l'intero cromosoma.

Per ciascun cromosoma, i ricercatori hanno calcolato $\ddot{\chi}_l$ per valori di $k = 2, \ldots, 6$ ed hanno ripetuto il procedimento utilizzando valori crescenti di l ottenendo così 786 grafici come quello riportato in figura 4.1. I dati, rappresentati in scala logaritmica su entrambi

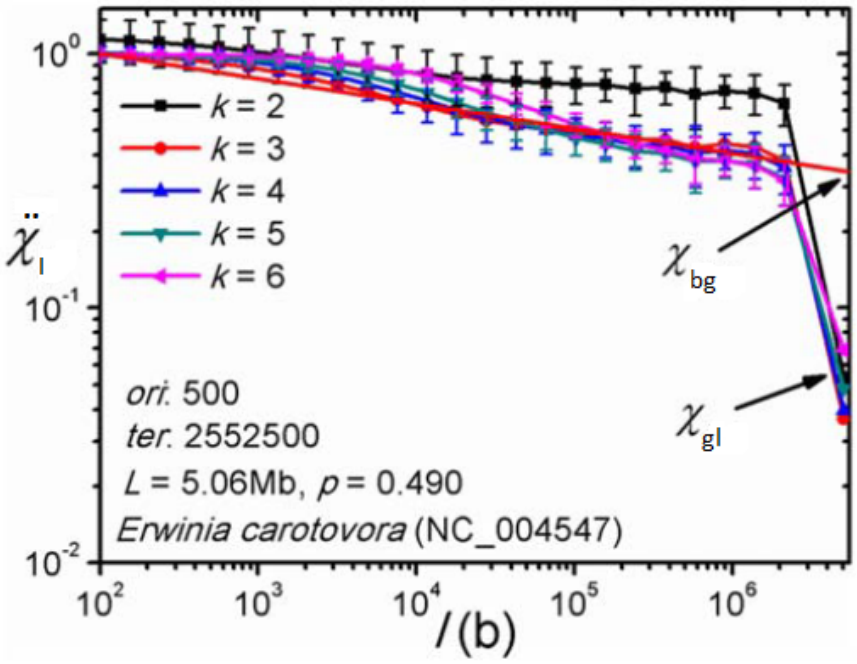

Figura 4.1: Valori di $\ddot{\chi}_l$ in funzione della lunghezza dei segmenti in cui il cromosoma di E. carotovora viene suddiviso. Figura estratta da [22].

gli assi, mostrano l'andamento dei valori di $\ddot{\chi}_l$ in funzione di l. Si nota immediatamente come al crescere di l si assista ad una diminuzione di $\ddot{\chi}_l$ secondo un comportamento approssimabile linearmente, seguito da una drastica caduta quando l si avvicina all'intera lunghezza del cromosoma.
Estrapolando la parte lineare, gli autori hanno definito, per ciascun cromosoma analizzato, la quantità χ_{bg} come il valore che $\ddot{\chi}_l$ assumerebbe se seguisse un comportamento lineare anche quando l è pari alla lunghezza dell'intero cromosoma (vedi figura 4.1).
χ_{bg} rappresenta il valore di background dell'indice di simmetria locale ed è stato utilizzato

per definire il seguente indicatore:

$$r_\chi = \frac{\chi_{bg}}{\chi_{gl}}.$$

Un alto valore di r_χ implica che la simmetria è molto più forte a livello globale di quanto lo sia localmente. Dunque ci aspettiamo che cromosomi che esibiscono un alto valore di r_χ siano caratterizzati prevalentemente da inversioni dovute ad eventi ricombinativi (vedi paragrafo 2.4 e Appendice B.4) piuttosto che da molteplici strutture di tipo stem-loop (vedi paragrafo 2.3).

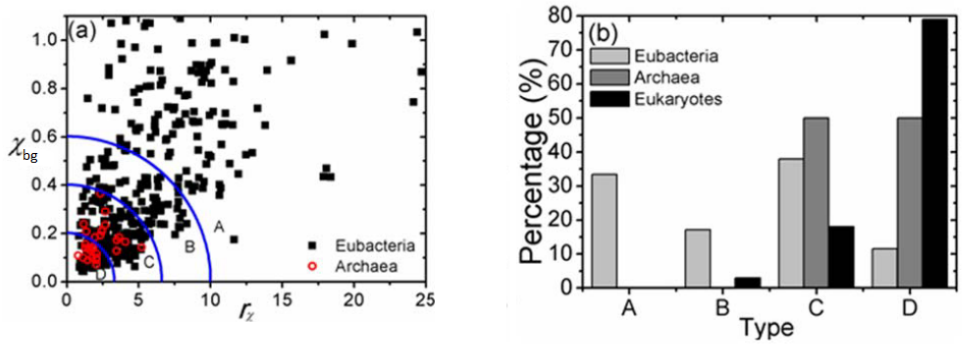

Figura 4.2: (a) Distribuzione dei cromosomi procaritici nel piano r_χ χ_{bg}.
(b) Distribuzione per tipologia dei cromosomi studiati. Figura estratta da [22].

I valori di χ_{bg} in funzione di r_χ per cromosomi di organismi procarioti sono riportati in figura 4.2 a. Anche se i dati non sembrano formare cluster distinti, per semplificare la discussione, gli autori hanno utilizzato la funzione:

$$\mathcal{T} = 0,5(\chi_{bg})^2 + 0,3(r_\chi)^2$$

per suddividere i cromosomi analizzati in quattro classi:

1. Tipo A, $\mathcal{T} > 9$.
2. Tipo B, $4 < \mathcal{T} < 9$.
3. Tipo C, $1 < \mathcal{T} < 4$.
4. Tipo D, $\mathcal{T} < 1$.

Questa classificazione risulta piuttosto arbitraria soprattutto nella scelta dei valori ($\mathcal{T} = 1, 4, 9$) in quanto non ci possiamo aspettare grandi differenze, almeno per quanto concerne la simmetria, tra due cromosomi che si trovano ai lati opposti di una delle linee di demarcazione. Se però analizziamo a quali organismi appartengono le varie tipologie di cromosomi, cominciano ad emergere interessanti correlazioni: i cromosomi di eucarioti superiori ad esempio sono tutti di tipo D mentre i cromosomi di tipo A appartengono esclusivamente ad eubatteri (vedi figura 4.2 b). Inoltre, come mostreremo nel prossimo paragrafo, le caratteristiche di simmetria delle tipologie estreme (tipo A e tipo D) appaiono differenti da un punto di vista qualitativo oltre che quantitativo.

4.3 La matrice χ

Per ottenere una rappresentazione visiva e più facilmente interpretabile del fenomeno, che metta in relazione la simmetria reverse-complement a livello locale con la posizione occupata sul cromosoma dalla sequenza analizzata, Sing-Guan Kong e collaboratori hanno utilizzato il seguente procedimento [22]. Attraverso un software disegnato appositamente, viene generata una finestra sovrapponibile lunga 100 kb, che si sposta lungo il cromosoma di 25 kb alla volta registrando la sequenza di basi che appare all'interno della cornice. Si crea così una serie di n frammenti, tutti della stessa lunghezza e parzialmente sovrapposti, che copre l'intero cromosoma. Per ogni valore di $k = 2, \ldots, 6$ (ordine rispetto al quale si vuole misurare la simmetria reverse-complement) si può costruire una matrice simmetrica n x n dove l'elemento di posto (i, j) è rappresentato dal valore dell'indice di simmetria χ calcolato per la sequenza di 200 kb, ottenuta concatenando il frammento i-esimo col frammento j-esimo, in cui l'intero cromosoma è stato suddiviso. La χ-matrix è pensata per mettere in luce la relazione di simmetria di tipo reverse-complement tra tutte le possibili coppie di segmenti lunghi 100 kb che costituiscono il cromosoma e, attraverso una sua rappresentazione grafica, permette anche di estrapolare informazioni preliminari sulla struttura e la storia evolutiva dei cromosomi analizzati secondo questa metodologia.
In figura 4.3 sono riportate le matrici di simmetria relative a quattro cromosomi procarioti, ciascuno appartenente ad una diversa tipologia, aventi lunghezza paragonabile

(circa 4 Mb) e stesso indice di simmetria globale ($\chi_{gl} \approx 0,05$): le differenze tra i tipi A e D appaiono abbastanza evidenti.

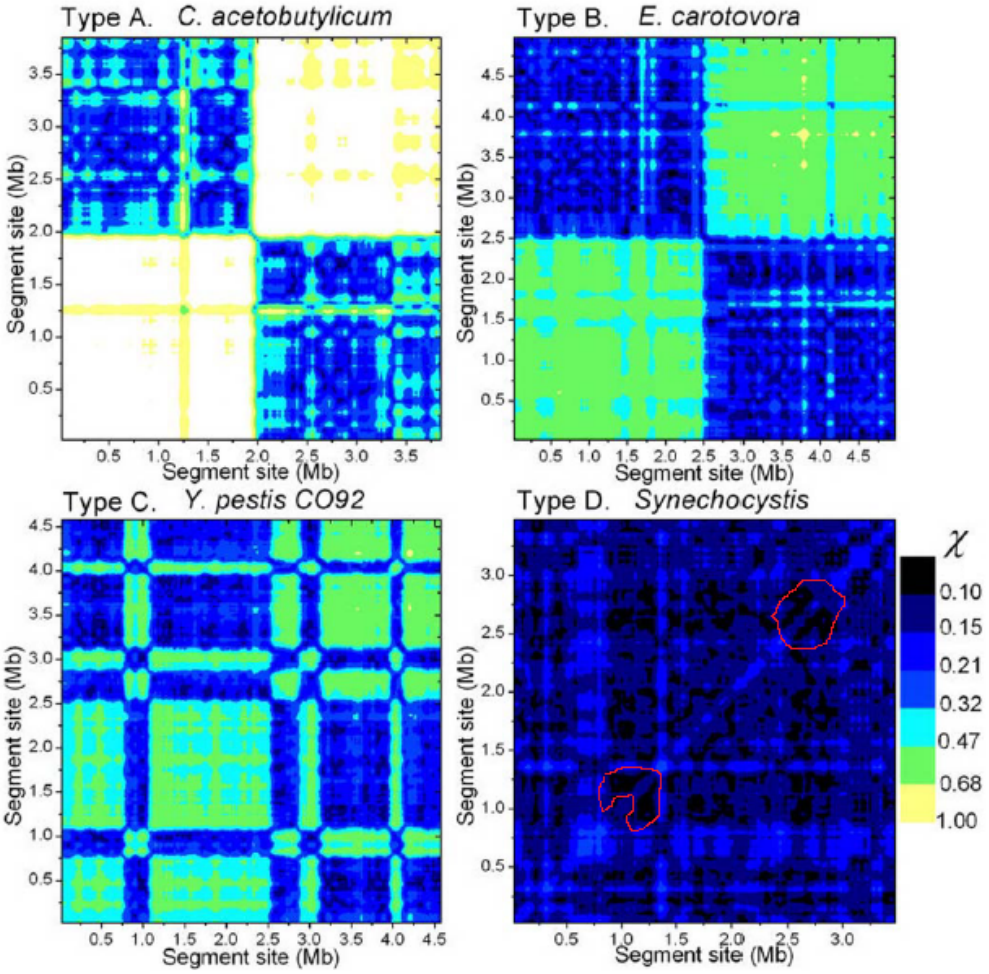

Figura 4.3: Le χ-matrix di quattro cromosomi rappresentativi delle varie tipologie. Il codice di colore è lineare in $\log \chi_{200kb}$ e il valore di χ_{200kb} attribuito a ciascun pixel è la media in $k = 2, \ldots, 5$. Figura estratta da [22].

La χ-matrix di C. acetobutylicum (tipo A) è suddivisa in quattro quadranti. Il colore chiaro della bisettrice primo-terzo quadrante indica assenza di simmetria locale in tutto il cromosoma (almeno su una scala di 100 kb). Osserviamo esplicitamente che la matrice di un cromosoma random sarebbe interamente bianca. La metà del cromosoma

coincide anche col sito di terminazione della duplicazione[1] (*ter*); per questo motivo gli autori hanno chiamato la parte a sinistra di *ter* il filamento veloce (lead-strand), mentre la metà a destra è stata denominata filamento lento (lag-strand). Il colore chiaro del primo e terzo quadrante indica che il contenuto in oligomeri di lunghezza k di segmenti appartenenti allo stesso strand è abbastanza simile, dunque l'indice di simmetria locale del segmento concatenato equivale a quello di uno dei due componenti, che in questo caso è vicino ad uno. Il colore scuro invece del secondo e quarto quadrante indica che qualsiasi coppia di segmenti che si trovano ai lati opposti di *ter* hanno una forte relazione di tipo reverse-complement. Ovviamente non si può dire che le due metà del cromosoma siano una il coniugato reverse-complement dell'altra, poiché in questo caso la figura apparirebbe interamente chiara, con un'unica diagonale scura larga qualche pixel, ma certamente, almeno per quanto riguarda il contenuto in oligomeri, le due metà presentano elevata simmetria. Ciò è probabilmente dovuto alla presenza di famiglie geniche e sequenze regolative simili con orientazione invertita nelle due metà.
La χ-matrix di E. carotovora (tipo B) assomiglia molto alla precedente, eccezion fatta per la lieve ombreggiatura che copre tutti e quattro i quadranti: ciò riflette una maggiore presenza di simmetria locale di tipo reverse-complement in tutto il cromosoma.
La χ-matrix di Y. pestis (tipo C) ha una struttura intermedia tra i tipi B e D. L'esempio proposto non ha una struttura divisa in quadranti e non rivela tracce di bisezione del cromosoma ma presenta una zona (da 0 a 2,7 Mb) di tipo B e una zona (da 2,7 a 4 Mb) molto simile ad un tipo D.
L'ultima χ-matrix riportata è quella di Synechocystis (tipo D), caratterizzato da una simmetria locale diffusa in tutto il cromosoma. La diagonale è relativamente più chiara e contornata da alcune zone di esatta simmetria reverse-complement cerchiate in rosso.
Le χ-matrix di tipo A hanno suggerito agli autori l'ipotesi che tali cromosomi si siano evoluti a partire da un evento di duplicazione invertita che ha coinvolto l'intero cromosoma ancestrale dal momento che esibiscono un alto grado di simmetria reverse-complement globale a fronte di una simmetria reverse-complement locale quasi impercettibile.

[1] I cromosomi batterici sono circolari, la duplicazione semiconservativa avviene simultaneamente in entrambe le direzioni rispetto all'origine di replicazione e su entrambi i filamenti della doppia elica. Per come è costruito il complesso proteico deputato alla sintesi dei nuovi filamenti di DNA e per la loro orientazione antiparallela, la replicazione avviene con velocità diverse.

Capitolo 5

Conclusioni e prospettive

In questo lavoro abbiamo mostrato come la seconda regola di Chargaff, nella sua versione generalizzata ad oligonucleotidi, abbia radici profonde legate alla struttura degli acidi nucleici e alla natura ricombinante delle sequenze genomiche.
Gli strumenti matematico-statistici utilizzati, anche se piuttosto elementari, hanno dato risultati importanti, permettendo di descrivere e quantificare il fenomeno della simmetria all'interno del patrimonio genetico dei vari organismi.
Lo studio della simmetria di tipo reverse-complement può essere un mezzo per indagare la storia stessa dell'informazione genomica, poiché può portare alla luce le tracce di riarrangiamenti cromosomici, eventi di duplicazione e inversioni, che risultano fondamentali per incrementare la variabilità su cui opera la selezione naturale, sia a livello di singoli geni [35], sia su scala più ampia ovvero a livello di interi frammenti cromosomici [3].

> "Nothing in biology makes sense except in the light of evolution."
>
> (Theodosius Dobzhansky)

Appendice A

Note di probabilità e statistica

In questa sezione presenteremo i riferimenti matematici relativi alle distribuzioni di probabilità utilizzate nel testo e forniremo alcune precisazioni che, per non frammentare eccessivamente i ragionamenti esposti, non è stato possibile includere all'interno dei capitoli precedenti. Per chi volesse approfondire può consultare, tra gli altri, [6].

A.1 Schema di Bernoulli

Un numero aleatorio X ha distribuzione discreta se la cardinalità dell'insieme $I(X)$ dei possibili valori x assunti da X è finita o numerabile.
La distribuzione di probabilità sarà del tipo:

$$P(X = x_i) = p(x_i) \qquad \forall x_i \in I(X)$$

e deve verificare inoltre:

$$\sum_{x_i \in I(X)} P(X = x_i) = 1.$$

Una successione di eventi $(E_i)_{i \in \mathbb{N}}$ stocasticamente indipendenti ed equiprobabili cioè tali che $P(E_i) = p \ \forall i \in \mathbb{N}$ prende il nome di schema di Bernoulli.
Data una successione di eventi di questo tipo, la distribuzione del numero di successi S in n prove prende il nome di distribuzione binomiale $B(n, p)$ (detta anche bernoulliana) di parametri n, p ed è caratterizzata dalla seguente equazione:

$$P(S_n = k) = \binom{n}{k} p^k q^{n-k}.$$

A.2 La distribuzione di Poisson

La distribuzione di Poisson è una distribuzione discreta, nota anche come legge degli eventi rari. Un numero aleatorio X ha distribuzione di Poisson di parametro λ se vale:

$$\mathbb{P}(n) \;=\; e^{-\lambda}\,\frac{\lambda^n}{n!} \qquad\qquad \forall\, n \in \mathbb{N}, \quad \lambda \in \mathbb{R}^+ \tag{A.1}$$

Osserviamo esplicitamente che dallo sviluppo in serie di potenze di e^λ otteniamo:

$$e^\lambda = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \qquad \Rightarrow \qquad \mathbb{P}(\mathbb{N}) = 1$$

Il valore atteso di tale distribuzione (vale a dire la media) dopo infinite prove è λ:

$$E[Y] \;=\; \sum_{n=0}^{\infty} n\, e^{-\lambda}\,\frac{\lambda^n}{n!} \;=\; e^{-\lambda} \sum_{n=0}^{\infty} n\,\frac{\lambda\,\lambda^{(n-1)}}{(n-1)!} \;=\; \lambda\, e^{-\lambda}\, e^{\lambda} \;=\; \lambda \tag{A.2}$$

così come la varianza:

$$\mathrm{Var}[Y] \;=\; E\left[Y^2\right] \;-\; (E[Y])^2$$

che per quanto visto in A.2 diventa:

$$\begin{aligned} \mathrm{Var}[Y] \;&=\; \sum_{n=0}^{\infty} n^2\, e^{-\lambda}\,\frac{\lambda^n}{n!} \;-\; \lambda^2 \\ &=\; \lambda\, e^{-\lambda} \sum_{n=1}^{\infty} n\,\frac{\lambda^{(n-1)}}{(n-1)!} \;-\; \lambda^2 \end{aligned}$$

con la sostituzione $n - 1 \;=\; t$ otteniamo:

$$\begin{aligned} \mathrm{Var}[Y] \;&=\; \lambda\, e^{-\lambda} \left(\sum_{t=0}^{\infty} (t+1)\,\frac{\lambda^t}{t!} \right) - \lambda^2 \\ &=\; \lambda\, e^{-\lambda} \left(\sum_{t=0}^{\infty} t\,\frac{\lambda^t}{t!} \;+\; \sum_{t=0}^{\infty} \frac{\lambda^t}{t!} \right) \;-\; \lambda^2 \\ &=\; \lambda\, e^{-\lambda} \left(\lambda\, e^{\lambda} \;+\; e^{\lambda} \right) \;-\; \lambda^2 \\ &=\; \lambda^2 \;+\; \lambda \;-\; \lambda^2 \;=\; \lambda \end{aligned}$$

A.3 Distribuzioni assolutamente continue

Sia X un numero aleatorio, specificare la funzione di ripartizione di X significa assegnare ad X la sua distribuzione di probabilità ovvero:

$$F(x) := P(X \leq x), \qquad x \in \mathbb{R}$$

La funzione di ripartizione è una funzione reale la cui immagine è compresa tra zero e uno ed è monotona. Inoltre si suppone goda delle seguenti proprietà di regolarità:

$$\lim_{y \to x^+} F(y) = F(x)$$

$$\lim_{x \to +\infty} F(x) = 1$$

$$\lim_{x \to -\infty} F(x) = 0.$$

Si dice che X ha distribuzione assolutamente continua se esiste una funzione $f : \mathbb{R} \to \mathbb{R}$ con le seguenti proprietà:

$$\forall x \in \mathbb{R}, \quad f(x) \geq 0$$

$$f \text{ è integrabile}$$

$$\int_{\mathbb{R}} f(s)ds = 1$$

e tale che la funzione di ripartizione di X si possa scrivere come:

$$F(x) = \int_{-\infty}^{x} f(s)ds.$$

Allora tale funzione f si dice densità di probabilità.

A.4 La distribuzione normale

Si dice che una variabile aleatoria X ha distribuzione normale standard (indicata usualmente con $N(0, 1)$) quando è caratterizzata dalla seguente densità di probabilità:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{A.3}$$

Vogliamo adesso mostrare che A.3 è effettivamente una funzione di densità cioè rispetta le proprietà elencate nel paragrafo precedente. Ovviamente (essendo una esponenziale) $f(x)$ è non negativa, continua e dunque integrabile. Vogliamo mostrare che il suo integrale sulla retta reale è pari ad uno. Tale integrale è conosciuto anche come integrale di Gauss e, per risolverlo, consideriamo:

$$\left(\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx\right)^2 = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \frac{1}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\frac{z^2+y^2}{2}} dz dy$$

Introduciamo le coordinate polari

$$z = \rho \cos\theta \qquad\qquad y = \rho \sin\theta$$

e la matrice Jacobiana della trasformazione

$$J_{(\rho,\theta)} = \begin{pmatrix} \cos\theta & -\rho\sin\theta \\ \sin\theta & \rho\cos\theta \end{pmatrix}$$

$$det\left(J_{(\rho,\theta)}\right) = \rho$$

effettuiamo la sostituzione ottenendo:

$$\begin{aligned} \frac{1}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-\frac{z^2+y^2}{2}} \quad dzdy &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{+\infty} \rho e^{-\frac{\rho^2}{2}} \quad d\rho d\theta \\ &= \frac{1}{2\pi} 2\pi \int_0^{+\infty} \rho e^{-\frac{\rho^2}{2}} \quad d\rho \\ &= \left[-e^{\frac{-\rho^2}{2}}\right]_0^{+\infty} = 1. \end{aligned}$$

Dunque anche

$$\int_{-\infty}^{+\infty} f(x) \quad dx = 1$$

e A.3 risulta effettivamente una densità.
Se X ha distribuzione normale standard consideriamo la variabile

$$Y = \mu + \sigma X \qquad\qquad \text{con } \sigma > 0.$$

La funzione di ripartizione di Y sarà:

$$F_Y(y) = P(Y \leq y) = P(\mu + \sigma X \leq y) = P\left(X \leq \frac{y-\mu}{\sigma}\right).$$

Per quanto visto prima allora

$$F_Y(y) = \int_{-\infty}^{\frac{y-\mu}{\sigma}} f(x) \quad dx.$$

Dalla relazione

$$f_Y(y) = \frac{dF_Y(y)}{dy}$$

si ricava

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

e si dice che la variabile Y ha distribuzione normale di parametri μ e σ (dove μ è il valore atteso e σ è la deviazione standard) e la si indica con $N(\mu, \sigma)$.
Concludiamo questo paragrafo con il seguente fondamentale teorema:
Date n variabili aleatorie stocasticamente indipendenti $X_1, \ldots, X_n$, aventi distribuzione gaussiana $N(\mu_i, \sigma_i)$ con $i = 1, \ldots, n$ allora la variabile aleatoria $Z = a_1X_1 + \cdots + a_nX_n$ è una variabile gaussiana con media $\mu_z = a_1\mu_1 + \cdots + a_n\mu_n$ e varianza $\sigma_z^2 = a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$.
Per semplicità di calcolo e notazione dimostriamo il teorema nel caso più elementare ossia quando $Z = X + Y$ tenendo presente che con ragionamenti del tutto analoghi si può ottenere la dimostrazione del caso più generale.
Sia dunque X e Y variabili aleatorie stocasticamente indipendenti e aventi rispettivamente distribuzione $N(0, \sigma_1)$ e $N(0, \sigma_2)$. Avremo che:

$$F_Z(z) = \int_A f_{X,Y}(x, y) \quad dxdy$$

dove $A = \{(x, y) \in \mathbb{R}^2 | x + y \leq z\}$ e $f_{X,Y}(x, y)$ è la densità cogiunta di X e Y.
Col seguente cambio di variabile si ha:

$$\begin{cases} x &= u \\ y &= u - v \end{cases}$$

$$det(J_{(u,v)}) = 1$$

$$F_Z(z) = \int_{\mathbb{R}} du \int_{-\infty}^{z} f_{X,Y}(u, v - u) \quad dv$$

e dalla relazione

$$f_Z(z) = \frac{dF_Z(z)}{dz}$$

ricaviamo

$$f_Z(z) = \int_{\mathbb{R}} f_{X,Y}(u, z-u) \quad du.$$

Poiché X e Y le abbiamo supposte stocasticamente indipendenti, la densità congiunta sarà il prodotto delle densità marginali vale a dire:

$$\begin{aligned} f_Z(z) &= \int_{\mathbb{R}} f_X(x) f_Y(z-x) \quad dx \\ &= \int_{\mathbb{R}} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma_1}\right)^2} \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-x}{\sigma_2}\right)^2} \quad dx \\ &= \frac{1}{\sigma_1\sigma_2 2\pi} \int_{\mathbb{R}} e^{-\frac{1}{2}\left[\left(\frac{\sqrt{\sigma_2^2+\sigma_1^2}}{\sigma_1\sigma_2}x\right)^2 + \frac{z^2}{\sigma_2^2} - \frac{2xz}{\sigma_2^2}\right]} \quad dx \end{aligned}$$

Consideriamo l'esponente, ponendo:

$$\alpha = \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sigma_1\sigma_2}$$

$$\beta = \frac{1}{\sigma_2^2}$$

$$\gamma = \frac{\beta}{\alpha} = \frac{\sigma_1}{\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}$$

avremo che

$$-\frac{1}{2}\left[\left(\frac{\sqrt{\sigma_2^2+\sigma_1^2}}{\sigma_1\sigma_2}x\right)^2 + \frac{z^2}{\sigma_2^2} - \frac{2xz}{\sigma_2^2}\right] = (\alpha x)^2 - 2\beta xz + \beta z^2$$

proseguiamo completando il quadrato aggiungendo e togliendo $\gamma^2 z^2$

$$(\alpha x)^2 - 2\beta xz + \gamma^2 z^2 - \gamma^2 z^2 + \beta z^2 = (\alpha x - \gamma z)^2 + (\beta - \gamma^2) z^2.$$

Allora l'integrale diventa

$$f_Z(z) = \frac{1}{\sigma_1\sigma_2 2\pi} e^{-\frac{1}{2}z^2(\beta-\gamma^2)} \int_{\mathbb{R}} e^{-\frac{1}{2}(\alpha x - \gamma z)^2} \quad dx.$$

con la sostituzione $(\alpha x - \gamma z) = t$ otteniamo un integrale gaussiano standard ovvero:

$$f_Z(z) = \frac{1}{\sigma_1\sigma_2 2\pi} e^{-\frac{1}{2}z^2(\beta-\gamma^2)} \frac{1}{\alpha} \int_{\mathbb{R}} e^{-\frac{1}{2}t^2} \quad dt$$

$$f_Z(z) = \frac{1}{\sigma_1\sigma_2 2\pi} e^{-\frac{1}{2}z^2(\beta-\gamma^2)} \frac{1}{\alpha}\sqrt{2\pi}$$

che in termini di σ_1 e σ_2 diventa:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_2^2+\sigma_1^2}} e^{-\frac{z^2}{2(\sigma_2^2+\sigma_1^2)}}$$

che è proprio la formula della densità gaussiana con media zero e varianza $\sigma_2^2 + \sigma_1^2$.

A.5 La distribuzione χ^2

Sia X una variabile aleatoria con distribuzione normale standard cioè caratterizzata da:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \qquad \forall x \in \mathbb{R}$$

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}}$$

dove f_X e F_X rappresentano rispettivamente densità e funzione di ripartizione.
Vogliamo conoscere la distribuzione della variabile aleatoria $Y = X^2$. Poiché $Y = X^2$ sappiamo già che tale variabile assume valori positivi dunque:

$$F_Y(y) = P(Y < y) = 0 \qquad \text{per valori } y \text{ negativi}$$

Sia dunque $y \geq 0$:

$$\begin{aligned} P(Y < y) &= P(X^2 < y) = P(-\sqrt{y} < X < \sqrt{y}) = \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) = F_X(\sqrt{y}) - (1 - F_X(\sqrt{y})) = \\ &= 2F_X(\sqrt{y}) - 1 = F_Y(y) \end{aligned}$$

dove nella seconda riga si è utilizzata la proprietà di simmetria della normale.
Da questo, con l'usale relazione di derivazione, si può ricavare la densità:

$$f_Y(y) = \frac{d}{dy} F_Y$$

$$f_Y = \frac{1}{\sqrt{2\pi}}\, y^{(\frac{1}{2}-1)}\, e^{\frac{-y}{2}}$$

che è una distribuzione Γ di parametri $\alpha = \lambda = \frac{1}{2}$ e che chiameremo per definizione

$$f_Y := \chi^2 \text{ di parametro } 1 \tag{A.4}$$

Tale distribuzione si può facilmente generalizzare al caso di n campioni aleatori $Y_1, \ldots, Y_n$ stocasticamente indipendenti e con distribuzione χ^2, utilizzando ad esempio le proprietà delle distribuzioni Γ ottenendo una distribuzione $\Gamma(\frac{n}{2}, \frac{1}{2})$ ovvero una distribuzione χ^2 di parametro n.

A.6 Coefficiente di correlazione di Pearson

Date due variabili aleatorie X e Y, il coefficiente di correlazione di Pearson è definito come la loro covarianza divisa per il prodotto delle deviazioni standard delle due variabili:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

dove σ_{XY}, è appunto la covarianza tra X e Y mentre σ_X e σ_Y sono le due deviazioni standard. Tale coefficiente assume sempre valori compresi tra -1 e 1.

A.7 Metodo dei minimi quadrati

Il metodo dei minimi quadrati è una tecnica di regressione che permette di determinare una funzione che meglio approssima un insieme di dati sperimentali. Siano (x_i, y_i) con $i = 1, \ldots, n$ i punti del piano che rappresentano i dati osservati. Si vuole trovare una funzione f che approssimi la successione di punti data minimizzando la distanza (euclidea) tra le due successioni (y_i) e $(f(x_i))$, ovvero la quantità M :

$$M = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

da cui appunto il nome "minimi quadrati".
Il caso più comune è quello lineare (la funzione desiderata è una retta):

$$f(x) = \alpha x + \beta$$

I coefficienti si possono determinare nel modo seguente:

$$\alpha = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

$$\beta = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

Appendice B

Note di genetica

Lo scopo dei paragrafi che seguono non è certamente fornire una trattazione completa di come il materiale genetico sia organizzato in un contesto cellulare, di quale sia il suo reale contenuto informativo o di quali siano i possibili cambiamenti che avvengono ad ogni replicazione all'interno di un genoma. I paragrafi seguenti costituiscono una guida estremamente sintetica per meglio collocare le affermazioni e i risultati dei capitoli precedenti nel contesto della biologia molecolare, fornendo magari, al lettore digiuno di genetica, gli strumenti interpretativi utili a muoversi autonomamente all'interno degli argomenti trattati. Per chi fosse alla ricerca di ulteriori chiarimenti rimandiamo a [33].

B.1 La struttura chimica degli acidi nucleici

Il DNA (acido deossiribonucleico) è una macromolecola organica a doppio filamento. I due filamenti o catene, sono uniti a formare la celeberrima doppia elica mediante interazioni elettrostatiche deboli dette legami idrogeno. Ciascun filamento è un polimero lineare costruito a partire da monomeri detti desossiribonucleotidi legati covalentemente tra loro. Ogni nucleotide è costituito da uno zucchero a cinque atomi di carbonio (deossiribosio), un gruppo fosfato e una base azotata che ne determina la specificità. I legami covalenti che formano la catena sono legami fosfodiesterici tra il gruppo fosfato di un nucleotide e lo zucchero del nucleotide precedente mentre i legami idrogeno si instaurano tra ciascuna base azotata di un filamento e la sua complementare sull'altro filamento se-

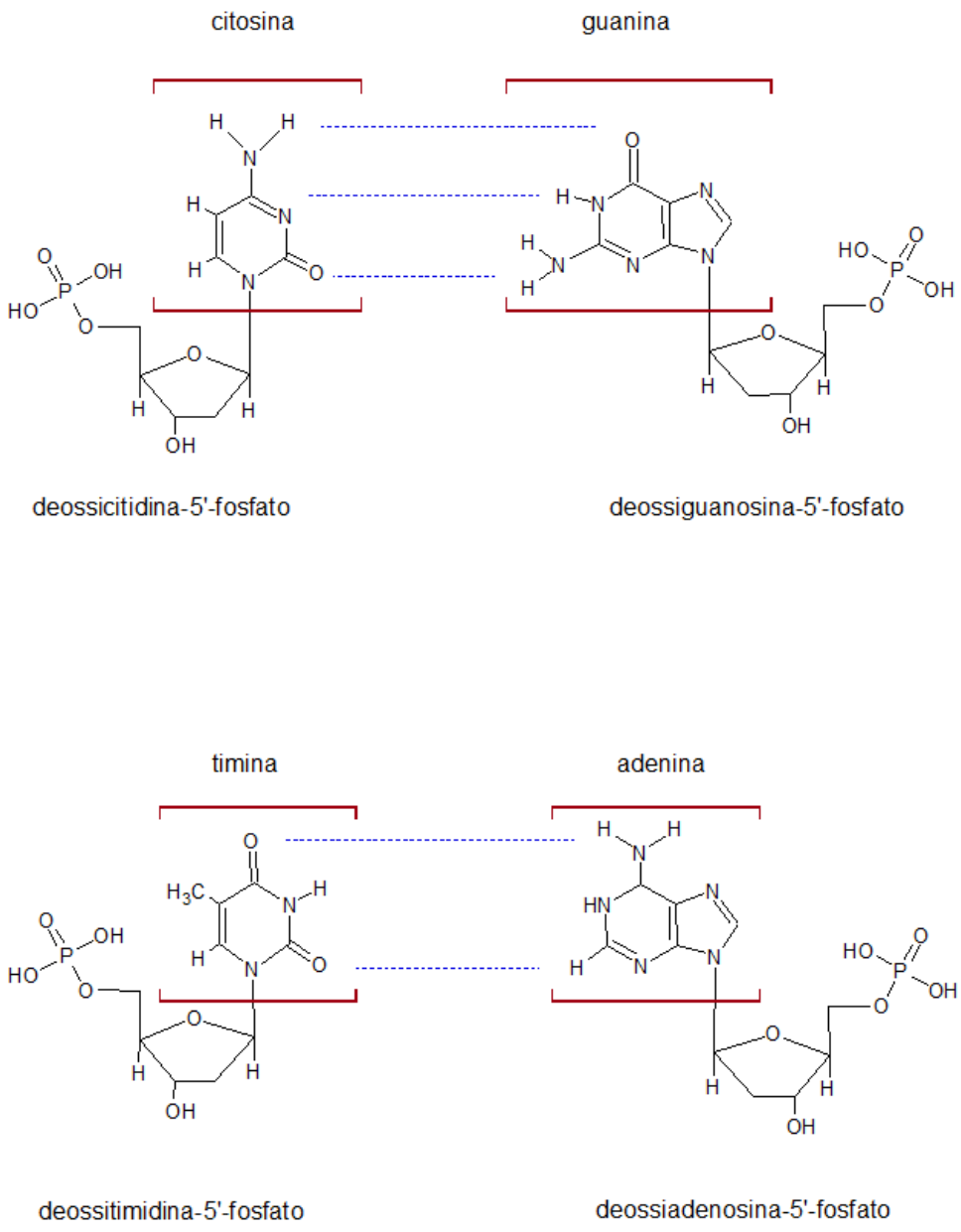

Figura B.1: Formule di struttura di deossiribonucleotidi e relativi appaiamenti secondo le regole di Watson e Crick. Le linee tratteggiate in blu rappresentano i legami idrogeno mentre le parentesi quadre in rosso racchiudono le basi azotate.

condo la regola di Watson-Crick (vedi figura B.1). I legami fosfodiesterici creano quindi una solida architettura ripetitiva zucchero-fosfato che è l'elemento strutturale costante del DNA e conferiscono una precisa polarità al filamento in quanto ciascuna catena avrà un'estremità 5' dove di norma è presente un gruppo fosfato e un' estremità 3' dove di norma troveremo un gruppo ossidrilico (-OH). Pertanto i filamenti della doppia elica hanno un'orientazione antiparallela. I legami idrogeno che stabilizzano la doppia elica invece, essendo più deboli, ne permettono l'apertura (generalmente locale e temporanea) con conseguente separazione dei due filamenti, consentendo così l'accesso al DNA a tutta una serie di complessi proteici che rendono possibili operazioni fondamentali per il funzionamento cellulare quali ad esempio duplicazione e trascrizione.
L'RNA (acido ribonucleico) è invece una molecola organica a singolo filamento. Tale polimero è però capace, ripiegandosi, di formare strutture secondarie stabili attraverso la formazione di legami idrogeno intracatena seguendo la medesima regola di appaiamento tra basi azotate vista per il DNA. Spesso l'RNA è associato a proteine (formando complessi nucleoproteici) o a ioni bivalenti in grado di facilitare la formazione di strutture tridimensionali funzionali. Da un punto di vista chimico, l'RNA differisce dal DNA in quanto lo zucchero che costituisce i nucleotidi è il ribosio e tra le basi azotate non compare la Timina bensì l'Uracile. A seconda della funzione che le molecole di RNA assolvono all'interno della cellula, possono essere raggruppate in varie classi:

- **mRNA**: è la molecola che trasporta l'informazione genica dal nucleo, dove si trova il DNA, al reticolo endoplasmatico rugoso, dove avviene la sintesi delle proteine (RNA messaggero).
- **tRNA**: ha funzione di raccordo tra i codoni presenti sull' m-RNA e gli amminoacidi ovvero i costituenti monomerici delle proteine (RNA di trasporto).
- **rRNA**: queste molecole di RNA, associate con proteine formano il ribosoma ovvero l'organulo cellulare deputato alla sintesi proteica (RNA ribosomiale).
- **Ribozimi**: sono RNA con attività catalitica esattamente come gli enzimi proteici.
- **snRNA**: sono brevi sequenze di RNA, generalmente ricche in Uracile, coinvolte nei processi di regolazione della trascrizione e splicing (piccoli RNA nucleari).

B.2 Il flusso dell'informazione genica

Nel 1956 Francis Crick coniò l'espressione "dogma centrale" per definire il flusso prevalentemente monodirezionale dell'informazione genetica dagli acidi nucleici alle proteine. I processi che permettono di passare da una sequenza di DNA alla costruzione della sequenza di amminoacidi che costituisce la struttura primaria di una proteina sono estremamente complessi ma, senza entrare nei dettagli biochimici, possiamo distinguere due passaggi fondamentali:

$$\textbf{DNA} \longrightarrow \textbf{RNA} \longrightarrow \textbf{proteine}$$

- **Trascrizione**: è il passaggio da DNA ad mRNA, in cui un meraviglioso complesso enzimatico detto RNA-polimerasi, coadiuvato da altre proteine e fattori di trascrizione, effettua la copia della sequenza di un gene.

- **Traduzione**: è il passaggio da una sequenza di basi azotate (mRNA) ad una sequenza di amminoacidi (catena polipeptidica). Questo processo coinvolge i ribosomi e degli adattatori (tRNA).

Il processo che permette invece il passaggio del patrimonio genetico alla generazione successiva è denominato **Duplicazione** o replicazione e ha natura semiconservativa, vale a dire che ciascuna delle due copie di DNA possiede un filamento appartenente alla doppia elica originaria ed uno neosintetizzato.
Ad oggi fanno eccezione al dogma centrale i retrovirus e i retrotrasposoni (vedi paragrafo B.4). Questi virus (come ad esempio HIV) hanno un genoma costituito da RNA e un ciclo di replicazione che ne prevede la retrotrascrizione in DNA per poterlo integrare all'interno del genoma dell'ospite. Fino ad ora, sebbene siano state evidenziate alcune forme alternative di trasmissione dell'informazione come la metilazione, l'editing e lo splicing alternativo o le modificazioni conformazionali generalmente irreversibili tipiche di alcune proteine prioniche, non è stato mai osservato un passaggio inverso alla traduzione.

B.3 L'organizzazione del materiale genetico

Nel contesto cellulare il DNA è associato a proteine (gli istoni) in un complesso detto cromosoma: un sistema di condensazione del DNA che conferisce alla molecola un'organizzazione strutturale di ordine superiore, ne previene eventuali danni e permette di trasmettere efficientemente l'informazione in essa contenuta alle cellule figlie. Questa organizzazione inoltre facilita la regolazione dell'espressione genica e consente la ricombinazione tra i cromosomi parentali introducendo così una fonte di variabilità ulteriore su cui può agire la selezione naturale. I cromosomi eucariotici sono caratterizzati da origini di replicazione multiple (ogni 30-40 kb) e da particolari strutture che ne permettono la mobilitazione e la segregazione quali centromeri e telomeri. Questi elementi, nell'uomo, sono caratterizzati da sequenze ripetute; in particolare nel caso dei telomeri si tratta di 200-400 ripetizioni della sequenza $5'-TTAGGG-3'$. L'insieme dei cromosomi costituisce il corredo genetico di un organismo ovvero il suo genoma. Differenti organismi possiedono genomi di grandezza diversa ed in linea di massima complessità dell'organismo e lunghezza del genoma sono grandezze positivamente correlate. Ad esempio i procarioti hanno genomi di lunghezza media inferiore alle 10 Mb, gli eucarioti unicellulari hanno genomi che si aggirano intorno alle 50 Mb, i protozoi più complessi hanno genomi che arrivano fino a 200 Mb. Esistono però numerose deviazioni da tale tendenza generale ovvero organismi di complessità paragonabile con genomi di grandezza molto diversa: è questo il caso del grano il cui genoma è circa 40 volte più grande di quello del riso. Tali discrepanze possono essere spiegate in termini di densità genica (espressa in numero di geni/Mb). Nella maggioranza dei casi infatti, i genomi di organismi più semplici hanno una maggiore densità genica e per quanto sappiamo fino ad oggi sono due i principali fattori che contribuiscono alla diminuzione della densità genica in rapporto al crescere della complessità degli organismi: l' aumento delle sequenze intergeniche e l'aumento delle dimensioni dei geni. I geni eucarioti sono più lunghi in media essenzialmente per la presenza di introni (sequenze di DNA trascritte ma che vengono eliminate prima della traduzione attraverso un processo di maturazione dell'mRNA chiamato splicing) nonché per l'aumento sia in numero che in estensione delle sequenze regolative necessarie all'espressione genica. Il DNA intergenico invece è la porzione di DNA che non è legato

all'espressione di proteine o RNA strutturali, nel caso dell'uomo costituisce oltre il 60% del genoma ed ha funzione essenzialmente sconosciuta. Queste regioni di DNA comprendono sequenze correlate ai geni come frammenti genici e pseudogeni, sequenze altamente ripetute come i trasposoni e il DNA microsatellite. Una panoramica sull'organizzazione del genoma umano è mostrata in figura B.2.

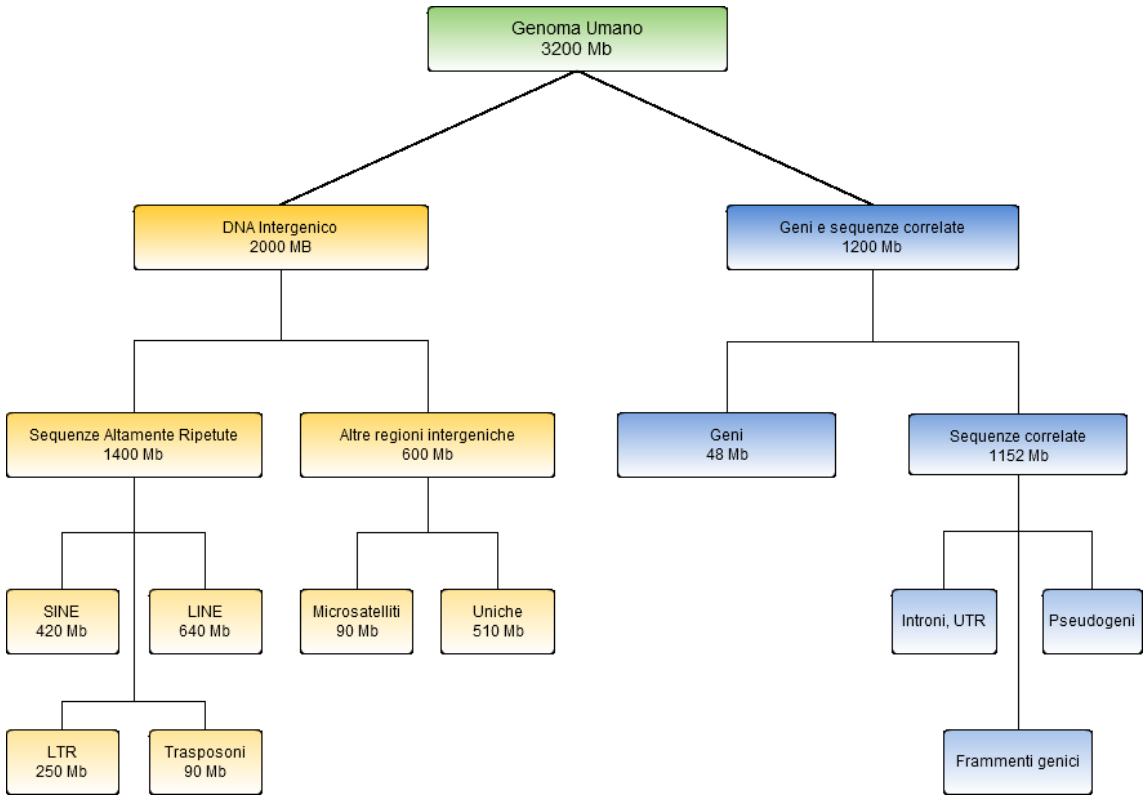

Figura B.2: Organizzazione del genoma umano: tipologie di sequenza e relative lunghezze espresse in Mb.

B.4 Riarrangiamenti cromosomici: ricombinazione e trasposizione

Il DNA è una molecola molto stabile ma non è immune a rotture o a modificazioni. Nella cellula sono presenti numerose proteine che hanno il compito di proteggere il materiale genetico, risaldare eventuali tagli nello scheletro di zucchero-fosfato ed eliminare eventuali basi modificatesi a causa di agenti mutageni. La funzione primaria di tali enzimi è quindi quella di preservare il patrimonio genetico di un organismo per trasferirlo inalterato alle generazioni successive. In questo modo ci si assicura che soluzioni adattative performanti non debbano essere continuamente "riscoperte" attraverso mutazioni puntiformi e casuali ad ogni generazione.
D'altra parte anche la variabilità del pool genico è di fondamentale importanza per l'evoluzione poiché permette di esplorare il paesaggio adattativo e sperimentare nuove soluzioni o semplicemente nuove combinazioni di soluzioni già esistenti: l'esempio più eclatante di un meccanismo in grado di riassortire il pool genico ad ogni generazione è proprio la riproduzione sessuale in quanto ogni nuovo zigote[1] eredita metà del proprio patrimonio genetico da ciascuno dei genitori ed il suo DNA è quindi una combinazione "inedita" di quello materno e paterno.
Esistono però anche altri processi enzimatici che promuovono la variabilità e sono caratterizzati da scambi fisici di materiale genetico tra due molecole di DNA. Tali processi sono detti fenomeni ricombinativi, vengono finemente regolati, necessitano di strutture proteiche ben definite e sono strettamente legati alla duplicazione del DNA. Si distinguono almeno tre tipologie di ricombinazione:

- La **ricombinazione omologa** si ha quando lo scambio di materiale genetico avviene tra due molecole di DNA che presentano una un'alta similarità di sequenza.
- La **ricombinazione sito-specifica** si verifica solo in corrispondenza di determinate sequenze.
- La **trasposizione** riguarda generalmente un breve tratto di DNA particolarmente bravo a replicarsi e capace di spostarsi da una posizione sul cromosoma ad un'altra.

[1] Cellula nata dalla fusione tra il gamete maschile e il gamete femminile.

Nei batteri la ricombinazione omologa ha principalmente la funzione di riparare sequenze di DNA danneggiate o di sbloccare le forcelle di replicazione[2]. Negli eucarioti invece processi di questo tipo si verificano molto frequentemente durante la meiosi e prendono il nome di *crossing over* (vedi figura B.3).

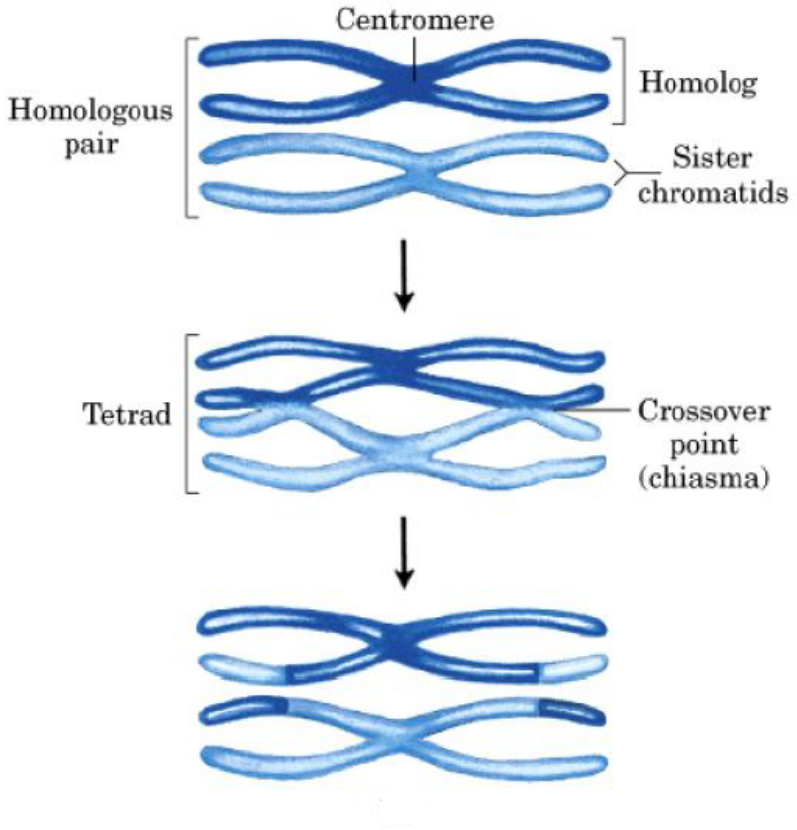

Figura B.3: Rappresentazione schematica del Crossing Over in cui si mostra lo scambio di materiale genetico tra cromosomi omologhi.

La ricombinazione sito-specifica invece necessita di due elementi fondamentali: un enzima denominato ricombinasi che catalizzi il trasferimento del materiale genetico e una breve sequenza di DNA (20-200 bp) detta appunto sito di ricombinazione. Tale processo

[2]Con questo termine si indica la zona in cui i filamenti della doppia elica vengono separati durante la duplicazione del DNA e il complesso macchinario proteico che porta a termine il processo.

avviene in ogni cellula ed ha in ciascuna specie funzioni peculiari tra cui la regolazione dell'espressione genica, il riarrangiamento programmato del DNA durante lo sviluppo o, nel caso di virus, le modificazioni del DNA ospite legate al proprio ciclo di replicazione. Da un punto di vista genomico invece la ricombinazione sito-specifica genera nuove sequenze di DNA in uno dei tre modi mostrati in figura B.4.

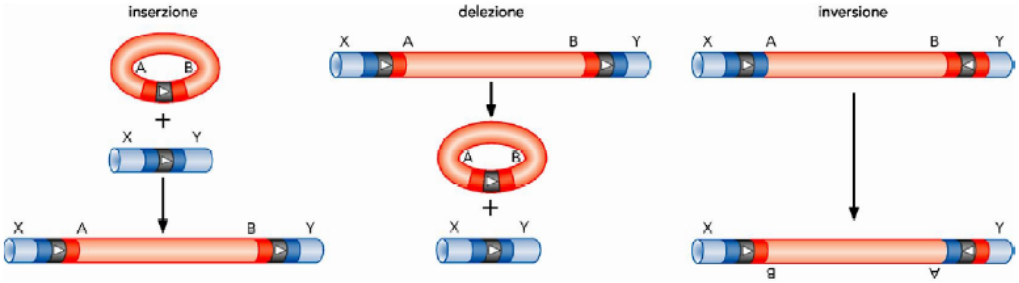

Figura B.4: I possibili esiti di eventi ricombinativi sito-specifici.

Gli elementi trasponibili o trasposoni sono elementi genetici mobili. Lo spostamento avviene mediante un evento di ricombinazione tra le sequenze poste all'estremità di un elemento trasponibile ed un sito bersaglio. Generalmente tale meccanismo è scarsamente selettivo per quanto riguarda la scelta del sito di inserzione che può teoricamente essere posto in una qualunque posizione del genoma. Il risultato è che i trasposoni possono "atterrare" all'interno di geni, distruggendone la funzione ed è stata proprio la compromissione di certe funzioni che ne ha permesso l'identificazione [27]. A seconda della loro generale organizzazione, si distinguono tre classi di elementi trasponibili rappresentate in figura B.5. I trasposoni a DNA e i retrotrasposoni LTR portano sia delle sequenze di DNA che servono da siti per la ricombinazione sia i geni che codificano per le proteine necessarie al processo. La differenza tra le due classi riguarda essenzialmente il meccanismo di replicazione in quanto i trasposoni LTR si replicano mediante un intermedio ad RNA. Questa reazione è catalizzata da una specialissima DNA-polimerasi che utilizza l'RNA come stampo ed è chiamata trascrittasi inversa. I retrotrasposoni poli-A invece non hanno le sequenze ripetute che fiancheggiano i geni necessari alla ricombinazione ma due sequenze dette UTR (untranslated region) e la loro struttura è molto simile a quella

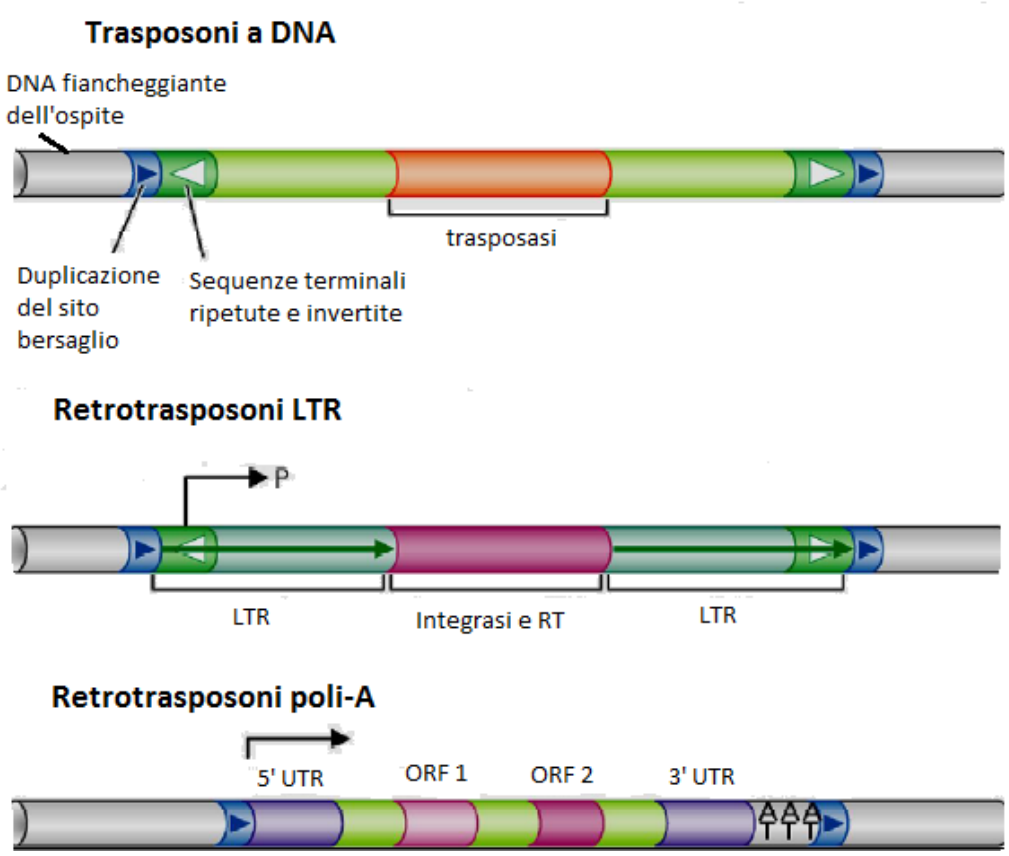

Figura B.5: L'organizzazione delle tre tipologie di elementi genetici mobili.

di un operone[3]. Di questa classe di retrotrasposoni fanno parte le sequenze dette LINE (in grado di spostarsi autonomamente) e SINE (che invece hanno perso la capacità di spostarsi autonomamente ma che utilizzano la trascrittasi codificata da LINE) che sono molto frequenti nel genoma di tutti i vertebrati. In generale, da un punto di vista genomico, il meccanismo di replicazione degli elementi trasponibili può essere di due tipi: copia e incolla oppure taglia e incolla. Nel primo caso si assiste alla duplicazione dell'elemento con conseguente aumento della lunghezza complessiva del genoma dell'organismo mentre nel secondo caso l'elemento semplicemente si sposta lungo il cromosoma. Relativamente alla regola di Chargaff, il meccanismo copia e incolla è particolarmente significativo in quanto spesso provoca l'inversione dell'elemento nella nuova posizione contribuendo quindi alla simmetria tra i due filamenti di DNA.

[3]Organizzazione dell'espressione genica tipica dei procarioti. Col termine operone si indicano un insieme di geni, spesso contigui, regolati in maniera strettamente coordinata

Appendice C

Cenni su Entropia e Informazione

Gli studi sul rapporto tra entropia ed informazione meriterebbero una tesi a sé ed è quindi irrealistico pensare di darne una trattazione esaustiva in queste poche righe. Riportiamo per completezza le definizioni di quanto citato nei precedenti capitoli in modo che il lettore possa avere i riferimenti minimi necessari, mentre per approfondimenti rimandiamo a [32].

C.1 Entropia di Shannon

In teoria dell'informazione si dice entropia il contenuto informativo medio di una certa sorgente S. L'idea di base è che il contenuto informativo di un certo messaggio abbia a che fare con l'incertezza: più si è stupiti nel vedere un certo simbolo o una sequenza di simboli, più alto sarà il loro valore informativo. Formalmente, consideriamo:

X	variabile aleatoria che può assumere un numero finito di valori
$E(X = x_i)$	evento in cui la variabile X assume il valore x_i con $i = 1, \ldots, m$

Definiamo il contenuto informativo dell'evento E:

$$I(E) = -\log(P(E)).$$

Definiamo quindi l'entropia come:

$$H(X) = -\sum_{i=1}^{m} P(X = x_i) \log(P(X = x_i))$$

che è appunto la media del contenuto informativo di ogni simbolo x_i pesata secondo la propria probabilità $P(x_i)$.
Osserviamo esplicitamente che $H(X) \geq 0$ e che gli eventi con probabilità nulla non hanno effetto sul valore dell'entropia coma mostra il seguente limite:

$$\lim_{P(x)\to 0} P(x)\log(P(x)) = \lim_{P(x)\to 0} \frac{\log(P(x))}{\frac{1}{P(x)}}$$

e applicando de l'Hôpital otteniamo:

$$\lim_{P(x)\to 0} \frac{\frac{1}{P(x)}}{\frac{-1}{P(x)^2}} = \lim_{P(x)\to 0} P(x) = 0.$$

C.2 Entropia relativa

Data una variabile aleatoria X che può assumere un numero finito di valori, e date due distribuzioni di probabilità possibili per tale variabile, rispettivamente $p_X(x)$ e $q_X(x)$ si definisce entropia relativa la quantità:

$$D_{KL}(p||q) = \sum_{x\in X} p_X(x)\log\left(\frac{p_X(x)}{q_X(x)}\right) \tag{C.1}$$

dove assumiamo che $0log0 = 0$ e $p_i \log\frac{p_i}{0} = \infty$.
Questa formula è nota come divergenza di Kullback-Leibler ed è una equazione fondamentale della teoria dell'informazione che quantifica la similarità tra due distribuzioni. In altri termini, possiamo dire che C.1 fornisce una stima dell'errore che commettiamo nell'usare come modello la distribuzione p_X quando la distribuzione che realmente genera i dati è q_X.
D_{KL} è usata ad esempio per quantificare, attraverso l'entropia, la dipendenza statistica tra due variabili aleatorie. Infatti, se ci chiediamo quanto la distribuzione congiunta di due variabili X e Y sia simile al prodotto delle distribuzioni marginali, applicando C.1 otteniamo:

$$D_{KL}(p(x,y)||p(x)p(y)) = \sum_i p(x_i,y_i)\log\frac{p(x,y)}{p(x_i)p(y_i)}$$

L'espressione precedente è nota col nome di mutua informazione o informazione reciproca ed è usualmente indicata con la notazione $I(X,Y)$.

Osserviamo esplicitamente che la divergenza di Kullback-Leibler è non negativa, è zero solo se $p_X = q_X$ e non è superiormente limitata. Osserviamo inoltre che D_{KL} non è una metrica in quanto non è simmetrica e non vale la disuguaglianza triangolare [23]. Esiste però un'intera famiglia di divergenze basate sull'entropia di Shannon [7], come ad esempio quella di Jensen-Shannon che è una versione simmetrica di D_{KL}:

$$JSD(p||q) = \frac{1}{2}D_{KL}(p||M) + \frac{1}{2}D_{KL}(q||M) \qquad \text{dove } M = \frac{1}{2}(p+q)$$

La radice quadrata di $JSD(p||q)$ è stato inoltre dimostrato essere una metrica [15].

Bibliografia

[1] V. Afreixo et al., "Genome analysis with inter-nucleotide distances", *Bioinformatics*, vol. 25, no. 23, 2009, pp. 3064-3070.

[2] G. Albrecht-Buehler, "Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions", *PNAS*, vol. 103, no. 47, novembre 2006, pp. 17828-17833.

[3] J.A. Bailey et al., "Recent segmental duplications in the human genome", *Science*, vol. 297, 2002, pp. 1003Ű1007.

[4] P.-F. Baisnée, S. Hampson, P. Baldi, "Why are complementary DNA strand symmetric?", *Bioinformatics*, vol. 18, no. 8, 2002, pp. 1021-1033.

[5] J.S. Bell e D.R. Forsdyke, "Accounting units in DNA", *Journal of Theoretical Biology*, vol. 197, 1999, pp. 51-61.

[6] F. Biagini e M. Campanino, "Elementi di Probabilità e Statistica", *Springer*, 2006, Milano, pp. 236.

[7] S.-H. Cha, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions", *International journal of mathematical model and methods in applied sciences*, vol. 1, no. 4, 2007, pp 300-307.

[8] E. Chargaff, "How genetics got a chemical education", *Annals of the New York Academy of Sciences*, vol 325, 1979, pp.345-360.

[9] E. Chargaff, "Essays on Nucleic Acids", *Elsevier*, 1963, Amsterdam.

[10] E. Chargaff, "Structure and function of nucleic acids as cell constituents", *Federation Proceedings*, vol 10, steembre 1951, pp. 654-659.

[11] E. Chargaff et al., "The composition of the Desoxypentose Nucleic Acids of thymus and spleen", *Journal of Biological Chemistry*, vol 177, 1949, pp. 405-416.

[12] H.-D. Chen et al., "Divergence and Shannon Information in Genomes", *Physical Review Letter*, no. 94, maggio 2005.

[13] H.-D. Chen et al., "Universal Global Imprints of Genome Growth and Evolution - Equivalent Length and Cumulative Mutation Density", *PLoSONE*, vol. 5, no. 4, aprile 2010.

[14] C. De Duve, "Alle origini della vita" *Bollati Boringhieri*, Torino, maggio 2011, pp. 315.

[15] D.M. Endres e J.E. Schindelin, "A New Metric for Probability Distributions", *IEEE Transaction on Information Theory*, vol. 49, no. 7, luglio 2003.

[16] J.W. Fickett et al., "Base compositional structure of genomes", *Genomics*, vol. 13, no. 4, agosto 1992, pp. 1056-1064.

[17] D.R. Forsdyke e J.R. Mortimer, "Chargaff's legacy", *Gene*, no. 261, 2000, pp. 127-137.

[18] D.R. Forsdyke, "Stem-loop potential in MHC genes: a new way of evaluating positive Darwinian selection?", *Immunogenetics*, vol. 43, 1996, pp. 182-189.

[19] D.R. Forsdyke, "Conservation of Stem-Loop Potential in Introns of Snake Venom Phospholipase A2 Genes. An Application of FORS-D Analysis", *Molecular Biology and Evolution*, vol. 12, 1995a, pp. 1157-1165.

[20] D.R. Forsdyke, "A stem loop 'kissing' model for the initiation of recombination and the origin of intron", *Molecular Biology and Evolution*, vol. 12, 1995b, pp. 949-958.

[21] R.F. Gesteland, T.R. Cech e J.F. Atkins, "The RNA World", *Cold Spring Harbor Laboratory Press*, Third Edition, 2006, Cold Spring Harbor, New York.

[22] S.-G. Kong et al., "Inverse Symmetry in Complete Genomes and Whole-Genome Inverse Duplication", *PLoSONE*, vol. 4, no. 11, novembre 2009.

[23] S. Kullback, "Information Theory and Statistics" *Dover Publications*, 1968, New York.

[24] J.R. Lobry e C. Lobry, "Evolution of DNA base composition under no-strand-bias condition when the substitution rates are not constant", *Molecular Biology and Evolution*, no. 16, 1999, pp. 719-723.

[25] E.S. Lander et al., "Initial sequencing and analysis of the human genome", *Nature*, no. 409, 2001, pp 860-921.

[26] J. Maynard Smith ed E. Szathmáry, "The Origins of Life. From the Birth of Life to the Origins of Language" *Oxford University Press*, 1999, Oxford, pp.180.

[27] B. McClintock, "The significance of responses of the genome to challenge", *Science*, no. 226, 1984, pp. 792-801.

[28] A.H. Murchie et al., "Helix opening transitions in supercoiled DNA", *Biochem. Biophys. Acta*, no. 1131, 1992, pp. 1-15.

[29] C. Nikolaou e Y. Almirantis, "Deviations from Chargaff's second parity rule in organellar DNA Insights into the evolution of organellar genomes", *Gene*, no. 381, 2006, pp. 34-41.

[30] V.V. Prabhu, "Symmetry observations in long nucleotide sequences", *Nucleic Acids Research*, vol. 21, no. 12, 1993, pp. 2797-2800.

[31] R. Rudner, J.D. Karkas ed E. Chargaff, "Separation of B. subtilis DNA into complementary strands", *PNAS*, vol. 60, no.3, 1968, pp. 915Ű920.

[32] C.E. Shannon e W. Weaver, "La teoria metematica delle comunicazioni", *ETAS libri*, 1971, Milano.

[33] J.D. Watson et al., "Biologia molecolare del gene" *Zanichelli*, quinta edizione, luglio 2005, Bologna.

[34] J.D. watson e F.H.C. Crick, "Genetical implications of the structure of deoxyribonucleic acid", *Nature*, vol. 171, pp. 964-967.

[35] J. Zhang, "Evolution by gene duplication: an update", *Trends in Ecology and Evolution*, vol. 18, no. 6, giugno 2003, pp. 292-298.

[36] S.-H: Zhang e Y.-Z. Huang, "Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA", *Bioinformatics*, vol. 26, no. 4, 2010, pp. 478-485.

Ringraziamenti

Vorrei ringraziare il professor Mirko Degli Esposti per aver supervisionato la stesura di questo lavoro e soprattutto per avermi dato l'opportunità di trattare un argomento un po' insolito per una tesi in fisica matematica ma che ha rappresentato per me la degna conclusione di un percorso iniziato all'Università di Firenze con la laurea in Biotecnologie molecolari e proseguito presso l'Università di Bologna al Dipartimento di Matematica.

Ringrazio il dottor Giampaolo Cristadoro per i commenti e le critiche costruttive che mi hanno spinto a rendere più rigoroso e leggibile questo lavoro.

Ringrazio la dottoressa Alessia Kogoj, Matteo Allegro e Federico Bucciarelli: i miei maestri di LaTeX.

Ringrazio il dottor Giulio Tralli per le ore spese a dialogare sulla divergenza di Kullback-Leibler e di Jensen-Shannon.

Un grazie ad Antonio Ricciardo per aver controllato i risultati sugli integrali gaussiani.

Grazie a Benedetta Franceschiello per i suoi appunti di teoria dell'informazione

Grazie a Massimiliano Tamburini per il materiale relativo alla chimica supramolecolare degli acidi nucleici.

Grazie a tutti i colleghi di matematica e non che mi hanno supportato in questo percorso, dalla preparazione agli esami alla stesura della tesi, comprese le pause al Caffè Università.

Grazie ai miei genitori per il sostegno morale ed economico e per non aver posto limiti di tempo al mio percorso di studi.

Grazie a Casa Kremlino per avermi ospitato durante la stesura di questo lavoro.

Grazie a Chiara Checcaglini per aver riletto le bozze, per avermi insegnato l'importanza delle virgole e per ogni giorno passato insieme.