UNIVERSITÁ DEGLI STUDI DI PISA
Facoltà di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea in Informatica

SVILUPPO DI SOFTWARE PER L'ANALISI STATISTICA DI SEQUENZE DI D.N.A.

Candidato:
Daniele Vitale

Tutore Accademico:
Prof. Fabrizio Luccio

Tutori Aziendali:
Dr. Ercan Kuruoglu
Dr. Osman Abul

ANNO ACCADEMICO 2006-2007

Indice

Introduzione

Cinquant'anni dopo la scoperta della doppia elica del DNA da parte di Watson e Crick, nell'aprile del 2003, i responsabili dell'*International Human Genome Sequencing Consortium*, hanno annunciato pubblicamente di aver terminato il sequenziamento dell'intero genoma umano, quindi, si è aperto ufficialmente un nuovo e vasto campo di possibilità per la ricerca scientifica. Tali risultati interessano settori della biologia molecolare, ma anche dell'ingegneria genetica e della ricerca medica e farmacologica. Nonostante l'entusiasmo, sicuramente giustificato, della comunità scientifica, questo evento è solo il primo passo nella comprensione della struttura e della funzionalità del nostro DNA. Il sequenziamento completo del genoma umano ha portato con sé soprattutto nuovi interrogativi. Ci si aspettava infatti che la struttura del nostro DNA e il numero di geni identificati potessero aiutare a spiegare la complessità dell'organismo umano. Invece, una delle più sorprendenti scoperte venuta subito alla luce dai risultati del Progetto Genoma Umano, è stata quella che il patrimonio genetico dell'uomo è formato all'incirca da 30.000 geni, una cifra molto inferiore ai 120.000 che erano stati ipotizzati inizialmente. Questo numero sorprende perché non è molto più grande della quantità di geni che caratterizzano le piante o altri organismi, la cui complessità, non è paragonabile a quella dell'uomo. Inoltre, questa parte rappresentata dai geni non è che una piccolissima percentuale, all'incirca l'1,5%, del nostro genoma. E' stata subito evidente la necessità di sviluppare strumenti potenti e affidabili per l'analisi dei dati sequenziati e di creare un supporto che permetta la condivisione di questo grande patrimonio di informazioni.

In questo contesto si inserisce la bioinformatica, una disciplina scientifica che si occupa di risolvere problemi biologici tramite l'utilizzo di strumenti informatici. Gli obiettivi principali della bioinformatica sono: la creazione di modelli statistici validi per interpretare i dati provenienti dagli esperimenti biologici, in modo da rilevare

tendenze o leggi numeriche; la realizzazione di modelli matematici e algoritmici alla ricerca di motivi o di sequenze rilevanti all'interno del DNA e dell'RNA; infine, l'organizzazione dei dati congiuntamente al loro inserimento in archivi informatici, al fine di renderne l'accesso funzionale e veloce.

L'attività di tirocinio svolta, dall'aprile 2007 al febbraio 2008, presso il Consiglio Nazionale delle Ricerche (CNR), si muove proprio in questo ambito. Ci si è proposto di sviluppare un software che permetta il rilevamento di dipendenze statistiche tra diverse zone delle sequenze di DNA, in particolare di individuare le cosiddette correlazioni a lungo raggio (LRC). Nel primo capitolo verranno presentati alcuni concetti fondamentali di biologia molecolare, verranno spiegate le motivazioni dello studio statistico della sequenza di DNA e verranno, infine, introdotte alcune formule provenienti dalla teoria dell'informazione che ci saranno utili nel prosieguo della relazione. Nel secondo capitolo, si esporranno le scelte implementative e la struttura del software realizzato, non prima però, di avere introdotto alcune informazioni riguardanti il software già esistente per la bioinformatica. Infine, si mostrerà la precisione degli strumenti sviluppati tramite il confronto con i risultati di un lavoro precedente di Stephan Beirer [8]. Nel terzo capitolo si proporrà una possibile estensione a questo precedente lavoro e se ne analizzeranno, nello specifico, i risultati nel quarto e ultimo capitolo.

Capitolo 1

Concetti Introduttivi

1.1 Il DNA: Cenni Biologici

Tutte le cellule viventi, senza alcuna eccezione conosciuta, conservano la loro informazione ereditaria all'interno di molecole a doppio filamento di DNA, una lunga catena polimerica formata sempre dagli stessi 4 tipi di monomeri (A,T,C,G). I monomeri sono legati insieme in una lunga sequenza lineare che codifica l'informazione genetica, proprio come una sequenza di 1 e di 0 codifica l'informazione in un file di computer [1].

1.1.1 La struttura del DNA

Per meglio comprendere i meccanismi che rendono la vita possibile, bisogna comprendere la struttura della molecola a doppio filamento. Ogni monomero su un singolo filamento, comunemente detto nucleotide, é composto di due parti: da uno zucchero (desossiribosio) con attaccato un gruppo fosforico, e da una base azotata, che può essere l'Adenina (A), la Citosina (C), la Guanina (G) o la Timina (T). Generalmente, quando ci si riferisce al singolo nucleotide, lo si indica con il nome della base azotata che lo contraddistingue. Ogni zucchero è legato al successivo tramite il gruppo fosforico, creando così una catena polimerica da cui sporgono le basi azotate. Le basi azotate di un filamento sono collegate a quelle del filamento opposto secondo una regola di complementarietà definita dalla struttura stessa delle basi. L'Adenina si collega alla Timina e la Citosina si collega alla Guanina. I due filamenti ruotano uno

intorno all'altro formando una struttura a doppia elica. I legami fra le coppie di basi sono deboli e ciò permette ai due filamenti di separarsi. Il singolo filamento, quindi, può agire come uno stampo per una nuova molecola di DNA [1].

1.1.2 Trasmissione dell'informazione genetica

Il compito del DNA non si esaurisce nella duplicazione di sé stesso, perché deve esprimere l'informazione contenuta al suo interno. Tale informazione guiderà la sintesi di nuove molecole all'interno della cellula. Questo compito viene svolto da un'altra classe di polimeri, l'RNA. Tramite un processo detto di trascrizione, alcune particolari sequenze di DNA vengono usate come stampo per la sintesi di molecole più corte di acido ribonucleico (RNA). La struttura dell'RNA è formata da uno zucchero leggermente diverso da quello del DNA, il ribosio invece del desossiribosio e al posto della base azotata Timina si trova l'Uracile (U). In questa fase parliamo di mRNA (RNA messaggero), poiché il compito di queste molecole è quello di portare al di fuori del nucleo della cellula l'informazione contenuta nel DNA, in modo che questa venga usata per la sintesi degli amminoacidi, che a loro volta si legheranno insieme formando le proteine. Le proteine possono essere considerate i mattoni degli organismi viventi e hanno diverse funzioni, tra cui dare forma agli organi, trasportare sostanze nutrienti e innescare alcune reazioni chimiche all'interno della cellula. L'informazione nella sequenza di una molecola di RNA messaggero è letta in gruppi di tre nucleotidi alla volta: ciascuna tripletta di nucleotidi, detta anche codone, codifica un singolo amminoacido che andrà ad unirsi insieme ad altri amminoacidi per formare una proteina. In questo modo ci sono $64(= 4x4x4)$ possibili codoni, anche se in realtà gli amminoacidi esistenti sono soltanto 20. Questo perché diversi codoni possono corrispondere ad uno stesso amminoacido. Il codice viene letto da un altro tipo di RNA detto tRNA (RNA transfer). Le molecole di DNA contengono le specifiche di migliaia di proteine e il segmento di sequenza che codifica una singola proteina viene detto gene [1].

1.1.3 Il DNA non codificante

Vista e considerata la sua funzione verrebbe da pensare che il DNA sia totalmente composto di geni. In realtà, non è così. Infatti, una parte considerevole è composta

di sequenze non codificanti che non specificano alcun amminoacido. Il DNA non codificante è presente nel genoma di tutti gli organismi viventi, ma soprattutto in quelli di tipo eucariota, come l'uomo. Nel genoma di quest'ultimo, ben il 98.5 % della sequenza, è non codificante. Mentre sappiamo molto sulle funzionalità del DNA codificante, gran parte di quello non codificante è ancora un mistero, al punto che per molto tempo è stato definito dai biologi molecolari DNA "spazzatura", riferendosi alla sua apparente mancanza di significato. Oggi sappiamo con certezza che parte del DNA non codificante svolge la funzione di regolare la velocità di trascrizione, ma anche di indicare dove iniziano e finiscono le sequenze codificanti. In particolare, sono degne di nota alcune sequenze, dette introniche, che si trovano all'interno di quelle codificanti e che vengono trascritte anche se successivamente eliminate e non utilizzate per la codifica della proteina. Studi hanno dimostrato che la mancanza di queste sequenze introniche provoca errori nella trascrizione e sintesi delle proteine.

Sebbene alcuni studi abbiano messo in luce alcune funzionalità del cosiddetto DNA "spazzatura", il 75% del genoma umano resta ancora privo di significato. Il ruolo del DNA non codificante è di grande interesse e nonostante siano state avanzate diverse teorie, rappresenta una problematica ancora aperta e molto dibattuta. Alcune scoperte, poi, hanno reso più evidente l'importanza di una maggiore comprensione del DNA non codificante, ad esempio, è stato scoperto che, il genoma dello scimpanzé e dell'essere umano, differiscono del solo 1.2 % se prese in considerazione le sequenze codificanti ma del ben 4% prendendo in considerazione anche le parti non codificanti [2].

1.2 Studio statistico delle sequenze di DNA

Come accennato nell'introduzione, lo studio dal punto di vista statistico del DNA, visto come una stringa i cui caratteri appartengono ad un alfabeto di 4 simboli (A,T,C,G), risulta fondamentale per la comprensione della struttura e della funzione del genoma. Le sequenze di DNA hanno dimostrato di essere statisticamente non uniformi e di presentare una variabilità inerente. In questo contesto, lo studio delle fluttuazioni statistiche all'interno della sequenza può rivelare caratteristiche importanti nella sua struttura, portando alla luce proprietà biologiche che altrimenti sarebbero rimaste nascoste.

1.2.1 Correlazioni

Di particolare interesse è lo studio delle correlazioni all'interno della sequenza: considerate due variabili quantitative X e Y si parla di correlazione se esiste tra di esse una relazione di dipendenza. Questa dipendenza può essere di tipo unilaterale, nel senso che una variabile influenza l'altra, oppure di tipo bilaterale, nel senso che le due variabili interagiscono senza che vi sia una relazione di causalità di una rispetto all'altra [6].

Tramite diverse metodologie, è stato possibile quantificare ed individuare all'interno del DNA, correlazioni dette a breve o a lungo raggio, a seconda della distanza su cui tali correlazioni vengono rilevate. Nel corso di questo studio, si è deciso di concentrarsi in particolar modo sulle seconde e sulle conseguenze della loro presenza.

1.2.2 Dipendenza a lungo raggio

Data una sequenza di DNA , considerandola come un processo stocastico stazionario ed ergodico, individuare una correlazione a lungo raggio che abbia un decadimento simile ad una legge di potenza significa aver trovato una dipendenza a lungo raggio, ossia che i valori ad ogni istante sono correlati ai valori di tutti gli istanti successivi.

Essendo una legge di potenza per definizione invariante in scala, le sequenze che presentano dipendenza a lungo raggio mostrano caratteristiche di auto-similarità, che sono di grande interesse poiché implicano proprietà frattali nell'organizzazione del genoma. Studi precedenti hanno dimostrato l'esistenza di correlazioni di questo tipo già nelle zone introniche e di DNA regolatore della trascrizione [7].

1.2.3 Legge di potenza

Una legge di potenza é una relazione polinomiale che presenta proprietà di invarianza di scala. Le leggi di potenza generalmente mettono in relazione due variabili e sono della forma:

$$f(x) = ax^k + o(x^k) \tag{1.1}$$

dove a è una costante, $o(x^k)$ è un infinitesimo di x^k , mentre k ,anch'esso una costante, indica l'ordine del polinomio omogeneo ed è chiamato esponente di scala.

La proprietà di invarianza di scala, caratterizzante una legge di potenza, fa sì che il polinomio soddisfi il seguente criterio:

$$f(cx) = a(cx)^k = c^k f(x) \propto f(x) \tag{1.2}$$

dove c è una costante.

La relazione precedente indica che variando l'ordine di grandezza dell'argomento della funzione, varia il rapporto di proporzionalità in funzione del cambiamento di scala, ma viene preservata la forma della funzione stessa.

Questa relazione appare più chiara se consideriamo il logaritmo da entrambi i lati della funzione 1.1 :

$$log(f(x)) = k\, log(x) + log(a) \tag{1.3}$$

Quest'espressione è lineare rispetto a $log(x)$, questa retta ha pendenza k e intercetta $log(a)$. Se cambiamo l'ordine di grandezza dell'argomento, otteniamo:

$$log(f(cx)) = k\, log(x) + log(a) + k\, log(c) \tag{1.4}$$

Questo mostra chiaramente che la forma della retta è rimasta invariata, ma è cambiata l'intercetta in relazione al cambiamento di scala.

1.3 Misure di Correlazione

Le correlazioni all'interno di sequenze di DNA, possono essere studiate e quantificate seguendo diversi approcci, in questo studio si è scelto di utilizzare concetti e misure che provengono dalla teoria dell'informazione e in particolare la funzione di mutua informazione $I(k)$ e la funzione di autocorrelazione $C(k)$. L'insieme di conoscenze rappresentate dalla teoria dell'informazione è stato spesso utilizzato nello studio delle sequenze di DNA, e tra queste soprattutto il concetto di entropia [5], che offre una misura quantitativa per l'incertezza.

1.3.1 Alcune assunzioni

Una sequenza di DNA può essere considerata come una stringa discreta s di lunghezza l composta di singoli simboli s_n con n=1,2,...,l-1,l. I simboli vengono presi da un alfabeto $A=\{X_1, X_2, ..., X_\lambda\}$ di cardinalità λ . Questo alfabeto può essere, ad esempio, composto dai quattro nucleotidi ed essere quindi di cardinalità 4. Questa sarà la scelta per alcune parti di questo studio ma in seguito verranno fatte anche scelte diverse. Per ogni simbolo di tipo X_i la frequenza con cui compare all'interno della sequenza viene indicata con f_i, mentre la frequenza con cui troviamo un simbolo di tipo X_i ad una distanza k dal simbolo di tipo X_j viene indicata con $f_{ij}(k)$. Facendo l'assunzione, già accennata precedentemente, che la sequenza di DNA sia un processo stocastico ergodico e stazionario, queste due frequenze possono essere prese come stimatori di massima verosimiglianza della probabilità p_i e $p_{ij}(k)$, dove il primo rappresenta la probabilità di trovare il simbolo X_i in una qualsiasi posizione della sequenza, mentre il secondo rappresenta la probabilità congiunta di trovare il simbolo X_i in una qualche posizione della sequenza e il simbolo X_j ad una distanza di $k-1$ da X_i.

1.3.2 La Funzione di mutua informazione

Data una variabile casuale X con una sua funzione di probabilità p la sua entropia viene definita come:

$$H(X) = p(X) log_2 p(X) \tag{1.5}$$

L'entropia di una variabile ne definisce il grado di incertezza. É possibile definire anche un'entropia condizionale $H(X \mid Y)$ che è l'entropia di una variabile casuale X condizionata alla conoscenza di un'altra variabile casuale Y.

Questa riduzione di incertezza, data da un'altra variabile casuale, viene detta mutua informazione. Per due variabili casuali la mutua informazione viene definita come:

$$I(X\,;\,Y) = H(X) - H(X \mid Y) = \sum_{x,y} p(x,y) \cdot log_2 \frac{p(x,y)}{p(x)p(y)} \tag{1.6}$$

La mutua informazione $I(X\,;\,Y)$ è una misura di dipendenza tra le due variabili

casuali. É simmetrica in X e Y, sempre non negativa. Questa è uguale a zero se e solo se X e Y sono indipendenti. Scegliendo il logaritmo in base 2 i risultanti valori saranno espressi in bit[3].

Tenendo conto delle assunzioni fatte nel paragrafo precedente la mutua informazione di due simboli ad una distanza k può essere scritta come:

$$I(k) = \sum_{i,j=1}^{\lambda} p_{ij}(k) \cdot log_2 \frac{p_{ij}(k)}{p_i \cdot p_j} \tag{1.7}$$

In questo modo $I(k)$ può essere interpretata come la distanza tra l'ipotesi di indipendenza statistica $p_i \cdot p_j$e la distribuzione congiunta reale $p_{ij}(k)$ di due simboli X_i e X_j che si trovano ad una distanza k nella sequenza [8].

1.3.3 Funzione di autocorrelazione

Presentiamo un'altra possibile misura di correlazione, che chiameremo funzione di *autocorrelazione*. É importante precisare che in questo caso non ci riferiremo all'autocorrelazione nel senso classico del termine, ma di una misura di dipendenza del secondo ordine che ora andremo a definire. Essendo stata definita nel precedente lavoro di Beirer [8] con questo nome si è deciso di continuare a utilizzarlo all'interno di questa relazione. La funzione di autocorrelazione serve a rilevare dipendenze lineari. Date due variabili di tipo X_i e X_j ad una distanza k la loro dipendenza lineare può essere vista come la deviazione dall'indipendenza statistica:

$$D_{ij}(k) = p_{ij}(k) - p_i\, p_j \tag{1.8}$$

Possiamo definire un vettore $\vec{a}$ di possibili valori da assegnare a X_i. Una specifica funzione di autocorrelazione $C_{\vec{a}}(k)$ può quindi essere definita come forma bilineare della matrice di correlazione $\underline{D}(k)$ (di dimensioni $\lambda \times \lambda$)con gli elementi $D_{ij}(k)$:

$$C_{\vec{a}}(k) = \vec{a} \cdot \underline{D}(k) \cdot \vec{a}^T = \sum_{i,j=1}^{\lambda} a_i \cdot D_{ij}(k) \cdot a_j \tag{1.9}$$

1.3.4 Relazione tra I(k) e C(k)

La funzione di mutua informazione può essere scritta anche come:

$$I(k) = \frac{1}{2\ln 2}\sum_{i,j=1}^{\lambda}\frac{D_{ij}^2(k)}{p_i \cdot p_j} + o(D_{ij}^3) \tag{1.10}$$

questa formula è ottenuta tramite un'espansione di Taylor di $I(k)$ in D_{ij}e ignorando termini più grandi del secondo ordine. Questa relazione mostra come la mutua informazione sia approssimativamente proporzionale alla funzione di correlazione al quadrato. Questo implica che se $C(k)$ mostra una legge di potenza con un decadimento del tipo $k^{-\gamma}$ questo fa sì che ci si aspetti che una legge di potenza di $I(k)$ abbia un decadimento del tipo $k^{-2\gamma}$. Questa constatazione risulterà molto importante perché ci permetterà in seguito di valutare in che modo determinati simboli influenzino l'andamento della mutua informazione.

Capitolo 2

Sviluppo del Software

2.1 Scelte Progettuali

In questa sezione vengono discusse e motivate le principali scelte seguite per lo sviluppo del software utilizzato per l'analisi delle sequenze di DNA. In particolar modo verrà posto l'accento sull'origine dei dati e sulle soluzioni adottate per l'implementazione delle funzioni di mutua informazione e di autocorrelazione.

2.1.1 Basi di dati biologiche

Esistono, ad oggi, più di un migliaio di basi di dati biologiche considerando sia quelle pubbliche che quelle private. I dati in esse contenuti comprendono sia le sequenze di nucleotidi di interi genomi sia le sequenze dei singoli geni e le sequenze di amminoacidi che compongo le varie proteine. Naturalmente, questi dati vengono continuamente aggiornati e queste basi di dati sono in continua crescita. I metodi utilizzati per la catalogazione e i servizi implementati per permettere l'accesso ai dati sono di fondamentale importanza, in considerazione del fatto che una simile mole di dati, se non accessibile in maniera funzionale, sarebbe difficilmente utilizzabile dalla comunità scientifica. Le banche dati si possono dividere in due diverse tipologie: primarie e secondarie.

Le banche dati primarie contengono informazioni e annotazioni delle sequenze nucleotidiche e proteiche, strutture e dati sull'espressione di DNA e proteine [11]. Il punto di riferimento principale per questo tipo di informazioni è sicuramente l'Inter-

national Nucleotide Sequence Database Collaboration (INSDC). Essa è stata formalmente costituita nel febbraio del 1987 e riunisce le informazioni provenienti dalle tre banche dati primarie più importanti: GenBank (National Center for Biotechnology Information) [20], EMBL Nucleotide DB (European Molecular Biology Laboratory) [21], DDBJ (DNA Data Bank of Japan) [22]. In particolare, come si può vedere in figura 2.1, l'apporto maggiore è fornito da GenBank la cui crescita è stata finora esponenziale, raddoppiando le proprie dimensioni ogni 18 mesi. Questa straordinaria crescita sembra possa continuare ancora per diverso tempo.

Le banche dati specializzate si sono sviluppate successivamente e raccolgono insiemi di dati omogenei, dal punto di vista tassonomico e funzionale, disponibili nelle banche dati primarie o derivanti da vari approcci sperimentali, rivisti e annotati con informazioni di valore aggiunto [11]. Esse contengono, quindi, informazioni più specifiche.

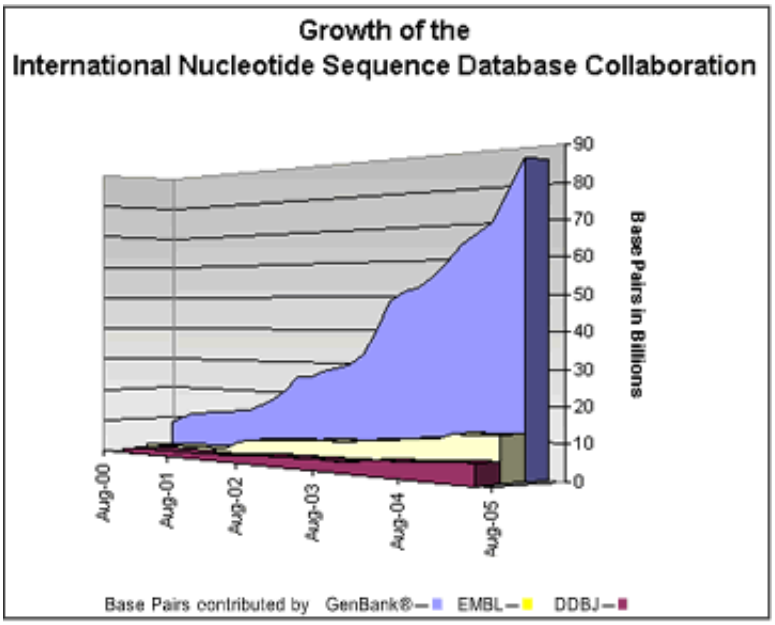

Figura 2.1: Crescita dell'International Nucleotide Sequence Database

2.1.2 Ensembl [15] & Entrez [14]

Nella sezione precedente è stata sottolineata l'importanza delle modalità di accesso alle informazioni. Nel corso del tirocinio si è deciso di affidarsi a due strumenti molto

affidabili: Ensembl ed Entrez.

Ensembl è una base di dati genomica che raccoglie e mantiene aggiornate le sequenze di diversi genomi eucariotici oltre ad una serie di software per la bioinformatica. I dati in essa contenuti provengono da diverse fonti alcune delle quali sono state citate precedentemente. Ensembl è un progetto gestito dal Welcome Trust Sanger Institute e dall'EMBL, esso è totalmente open source e offre quindi libero accesso ai dati. Ensembl è stato utilizzato per reperire le sequenze dei cromosomi utilizzati durante il tirocinio.

Entrez è, invece, quella che si può definire una "meta-base di dati", cioè una base di dati che permette, tramite un particolare motore di ricerca, di reperire informazioni di diversa natura sulle sequenze provenienti da diverse basi di dati biologiche sia primarie che specializzate. Nel corso del tirocinio, l'utilizzo di questo strumento ha permesso di ottenere, in maniera relativamente semplice le informazioni riguardanti le posizioni dei singoli geni sulle sequenze di DNA.

2.1.3 Il formato FASTA [9]

Esistono diversi formati di file che rappresentano sequenze di acidi nucleici o di amminoacidi, nel corso di questo lavoro si è scelto di affidarsi al formato di tipo FASTA. Questa scelta è stata fatta perché la sua semplicità ha reso più agevole la lettura dei dati, evitando problemi di conversione da un formato più complesso. Inoltre, essendo questo formato il più utilizzato, non c'è stato problema nel reperire le informazioni volute in file FASTA.

Una sequenza nel formato FASTA inizia con una singola linea di descrizione, seguita dalle linee di sequenze di dati. La linea di descrizione si distingue dalle altre perché inizia con il simbolo "$>$" , in genere ogni linea di testo non dovrebbe superare gli 80 caratteri. In figura 2.2 si può vedere un semplice esempio. Le sequenze sono rappresentate secondo un codice standard. La tabella 2.1 riporta la codifica per gli acidi nucleici. Esiste una codifica anche per gli amminoacidi ma non viene riportata poiché esula dagli interessi di questo lavoro. La linea di descrizione può contenere alcune informazioni sulle sequenze, ma la loro struttura varia a seconda della base di dati di provenienza. Per semplicità, vista anche l'eterogeneità delle strutture, si è deciso di non tenere conto delle informazioni contenute nella descrizione.

```
>gi|2114300:5865-6152 Homo sapiens immunoglobulin lambda gene locus DNA, clone:92H4
TCCTCTGAGCGGACTCAGTTGCCTGCAGCGTCTGTGGCTTTGGGACAGAGGGCCAGGATCACCTACCAGG
GAGACAGCATAGAATACTTTTATGCAAACTTTTGTACCAGCAGAAGCTAGGACAGGTCCCTGTGCTGGTA
ATCTATGGTGACAGCAACTGGCACTCAGTGATTCCTGAACAACTCTCTGACTCCATATCAGAGAACATGG
CCACCCTGATAATCAATGGGCCCCAGGCTGGAAACAAGGCTATTACTGTCAATCATGAGACAGCACTGAT
ACTCATCT
```

Figura 2.2: Esempio di file FASTA

A	adenosine	**M**	A C (amino)
C	cytidine	**S**	G C (strong)
G	guanine	**W**	A T (weak)
T	thymine	**B**	G T C
U	uridine	**D**	G A T
R	G A (purine)	**H**	A C T
Y	T C (pyrimidine)	**V**	G C A
K	G T (keto)	**N**	A C G T (any)
		-	gap of indeterminate length

Tabella 2.1: Codici per gli acidi nucleici nello standard IUB/IUPAC

2.1.4 Software per Bioinformatica

Qui, di seguito, verranno presentati alcuni software o librerie di cui si è fatto uso durante il corso del tirocinio.

2.1.4.1 BioConductor [16]

BioConductor è un progetto open source che ha come obiettivo la realizzazione di software per l'analisi e la comprensione dei dati genomici. Il progetto, iniziato nel 2001, è gestito principalmente dal Fred Hutchinson Cancer Research Center (FHCRC).

BioConductor si basa prevalentemente sul linguaggio R, anche se esistono integrazioni provenienti da altri linguaggi. Al suo interno esiste una fornita libreria di funzioni statistiche e metodi per la rappresentazione grafica, mentre sono in corso di implementazione funzionalità per la connessione diretta alle basi di dati biologiche più conosciute. Offre, inoltre, strumenti automatici per la documentazione. BioConductor rappresenta sicuramente, soprattutto nel panorama open source, una delle risorse più importanti per lo studio delle sequenze di DNA. Nonostante ciò, non è

stato possibile trovare nelle sue librerie tutte le funzionalità necessarie agli scopi di questo elaborato, a partire dagli strumenti per il calcolo delle frequenze. Naturalmente, è stata vagliata la possibilità di scrivere nuove funzioni per BioConductor al fine di ottenere ciò che si voleva ma, preso atto della necessità di produrre nuovo codice, si è preferito scriverlo in maniera completamente indipendente utilizzando il linguaggio C++. Si è realizzato un insieme di applicazioni, che va sotto il nome di bioUtils. Questa scelta è stata dettata sia dall'esigenza di avere un completo controllo del software utilizzato, sia per sfruttare le maggiori prestazioni di un linguaggio come il C++. R, infatti, ha una buona espressività, il che lo rende molto utilizzato dalla comunità scientifica poco avvezza alla programmazione, ma non unisce a questa una particolare efficienza.

2.1.4.2 Matlab [19] & Octave [18]

Matlab, prodotto dalla MathWorks, è un ambiente per il calcolo scientifico ad alte prestazioni, ma è anche il nome di un linguaggio di semplice utilizzo che opera all'interno dello stesso ambiente. Octave è la versione open source di Matlab. Quest'ultimo rispetto a Octave offre un'interfaccia grafica piuttosto evoluta e un insieme di librerie sicuramente più ampio, anche se si tratta di un problema secondario visto e considerato che le funzioni in Matlab sono utilizzabili anche da Octave. Matlab offre anche degli strumenti specifici per la bioinformatica che supportano l'attività di ricerca di motivi e di allineamento di sequenze, per cui non sono stati necessari per gli scopi di questo lavoro. Inoltre, attraverso alcuni test, si è potuto constatare una certa difficoltà nel lavorare con sequenze molto lunghe. Gli strumenti matematici e la possibilità di visualizzazione grafica dei dati numerici sono stati però utili per trattare i dati ottenuti con i programmi in C++ realizzati durante il tirocinio. Si è fatto uso, per la precisione, di Octave, che, come detto, ha delle caratteristiche simili a MatLab, ma è stato scelto tenendo conto che, al contrario di quest'ultimo è stato realizzato per funzionare su un sistema Unix, che è appunto la piattaforma su cui si è deciso di sviluppare la parte software del lavoro.

2.2 Le bioUtils

Le bioUtils rappresentano un insieme di applicazioni *stand-alone,* realizzate in C++, capaci di conteggiare le frequenze all'interno di una generica sequenza di DNA e, usando quest'ultime, di calcolare le funzioni $I(k)$ e $C^2(K)$.

2.2.1 Il problema del calcolo delle frequenze

Per poter procedere al calcolo delle funzioni di mutua informazione e di autocorrelazione, come è stato detto nel capitolo introduttivo, è necessario ottenere le frequenze dei singoli simboli all'interno della sequenza e quelle delle coppie di simboli che si trovano ad una distanza che va da 1 a k. Sono stati realizzati, a questo proposito, dei programmi il cui compito è di calcolare queste frequenze su una sequenza data e di salvare i risultati all'interno di un file di testo, a cui è stato dato, per esplicitarne il contenuto informativo, estensione *.freq*. I programmi in questione sono $nucfreq$ e $couplefreq$. Si tratta di un calcolo concettualmente semplice, ma che porta con sé difficoltà di tipo computazionale se si prendono in considerazione sequenze di una notevole lunghezza, come nel nostro caso, dove le sequenze sono nell'ordine delle decine di milioni di nucleotidi. I principali problemi riscontrati riguardavano l'occupazione di memoria e il numero di passi necessari per ottenere i valori delle varie frequenze, in particolare le frequenze delle coppie. Se l'occupazione di memoria risulta essere un problema trascurabile date le grandi capacità di memoria RAM disponibili anche su un singolo calcolatore, lo stesso non si può dire per il secondo problema. L'approccio che si è scelto per calcolare le frequenze delle coppie di simboli è stato quello di prendere in considerazione ogni posizione della sequenza e controllare le $k-1$ posizioni successive, ottenendo in questo modo per ogni valore da 1 a k l'occorrenza di una particolare coppia. I risultati per ogni possibile coppia vengono poi salvati all'interno di un file di testo contenente un vettore colonna di $k-1$ righe, dove la i-esima riga contiene il numero di occorrenze di quella particolare coppia ad una distanza i dove naturalmente i è compreso tra 1 e $k-1$. Tutto ciò vuol dire fare $n \cdot k$ letture. Tenendo conto che i valori di k presi in considerazione durante questo lavoro sono nell'ordine di 1×10^6, allora per una sequenza composta da 20 milioni di nucleotidi sono necessarie $20 \cdot 10^{12}$ letture e quindi anche con un milione di letture al secondo si impiegherebbero più di 230 giorni per una sola elaborazione.

Naturalmente, l'approccio utilizzato potrebbe apparire troppo banale, ma in realtà tutti gli altri algoritmi presi in considerazione anche se più eleganti non introducevano migliorie dal punto di vista degli accessi in memoria. Una possibile soluzione, probabilmente la migliore, scaturisce dall'osservazione che il problema è facilmente parallelizzabile. Si potrebbe, infatti, utilizzare una rete di computer che si occupi di eseguire l'analisi su sottoinsiemi della sequenza data, e un software che si occupi di fare il *merge* dei risultati provenienti dalle varie macchine. Non avendo a disposizione risorse hardware di questa entità, si è deciso di affrontare queste difficoltà in un altro modo, ossia, di procedere ad un opportuno campionamento quando le distanze k sono eccessivamente grandi. Questo campionamento può essere definito di volta in volta semplicemente passando al programma $couplefreq$ i parametri desiderati.

2.2.2 MutualInformation & AutoCorrelation

I dati ricavati nella fase precedente, sono utilizzati come input per altri due programmi realizzati per il calcolo delle funzioni di mutua informazione e autocorrelazione. Le applicazioni in questione sono *MutualInformation* e *AutoCorrelation*, che altro non sono se, non l'implementazione rispettivamente delle formule 1.7 e 1.9. Naturalmente, ai due software vengono fornite anche le informazioni che riguardano il campionamento utilizzato nel calcolo delle frequenze in input. *AutoCorrelation,* deve ricevere come input i valori del vettore $\overrightarrow{a}$ che ci permettono di scegliere su quali nucleotidi calcolarla. Entrambe le applicazioni hanno come output un file in un formato leggibile da Octave, che aiuterà nella visualizzazione dei risultati.

2.2.3 Un'estensione per i codoni

I programmi visti fino ad ora sono specifici per l'analisi sui singoli nucleotidi della sequenza, cioè prendiamo in considerazione come alfabeto dei simboli quello composto dai quattro nucleotidi. Naturalmente, questa non è l'unica scelta possibile, in particolare è interessante compiere un'analisi simile sui codoni e quindi prendere come alfabeto quello composto dalle 64 possibili triplette di nucleotidi. A questo proposito, sono stati realizzati anche i programmi *codonfreq*, *couplecodonfreq*, *CodonMutualInformation* e *CodonAutoCorrelation.* Le caratteristiche di questi programmi sono esattamente le stesse delle loro versioni per i singoli nucleotidi, ma in più prevedono

la possibilità di scegliere la modalità da utilizzare per dividere in triplette le sequenze da analizzare. Caratteristiche che saranno mostrate più nello specifico nel terzo capitolo.

2.2.4 Visualizzazione dei risultati

Per poter apprezzare, al meglio, i risultati ottenuti nei passaggi precedenti, di fondamentale importanza è la loro visualizzazione grafica. Per raggiungere questo obiettivo si è fatto uso di Octave, che a sua volta interfacciandosi con Gnuplot, un programma per la realizzazione di grafici di funzioni matematiche in due e tre dimensioni, ha permesso di ottenere in maniera semplice e versatile i grafici di cui avevamo bisogno. L'utilizzo di Octave potrebbe sembrare superfluo, infatti non si tratta di un passaggio indispensabile, visto che ci si potrebbe direttamente servire di Gnuplot [12] per visualizzare i dati, ma l'utilizzo di Octave ci permette, oltre che un interfacciamento più semplice, anche di sfruttare le sue librerie matematiche per successive manipolazioni sui dati.

2.2.5 GeneMasker

GeneMasker è un programma realizzato per il mascheramento delle sequenze di DNA. Per "mascheramento" si intende la sostituzione in alcune particolari zone scelte dall'utente, dei simboli originali con un simbolo di *default*. In particolare, GeneMasker usa come simbolo sostitutivo il carattere 'N'. Il mascheramento permette di mettere in evidenza solo alcune parti della sequenza e su quelle di fare delle analisi specifiche. L'applicazione funziona in tre modalità diverse, si può scegliere di mascherare le zone del tipo prescelto, di mascherare tutta la sequenza tranne le zone del tipo prescelto e infine si può scegliere di estrarre le zone del tipo prescelto e di salvarle su file di tipo FASTA. L'uso, di tale programma, è stato finalizzato a mettere in evidenza i geni sulla sequenza. Per procedere il software deve conoscere le posizioni delle zone da mascherare, queste informazioni sono ottenute dal file *seq_gene.md* [13], un file fornito dall'NCBI che contiene una mappa dettagliata dei geni individuati sul genoma umano. L'ente è in grado di fornire file simili anche per genomi di altre specie. Il programma, come si è detto, è nato per lavorare in riferimento ai geni

ma, poiché, il file *seq_gene.md* contiene anche le posizioni di zone di altro tipo si è reso possibile utilizzare GeneMasker in riferimento a quest'ultime.

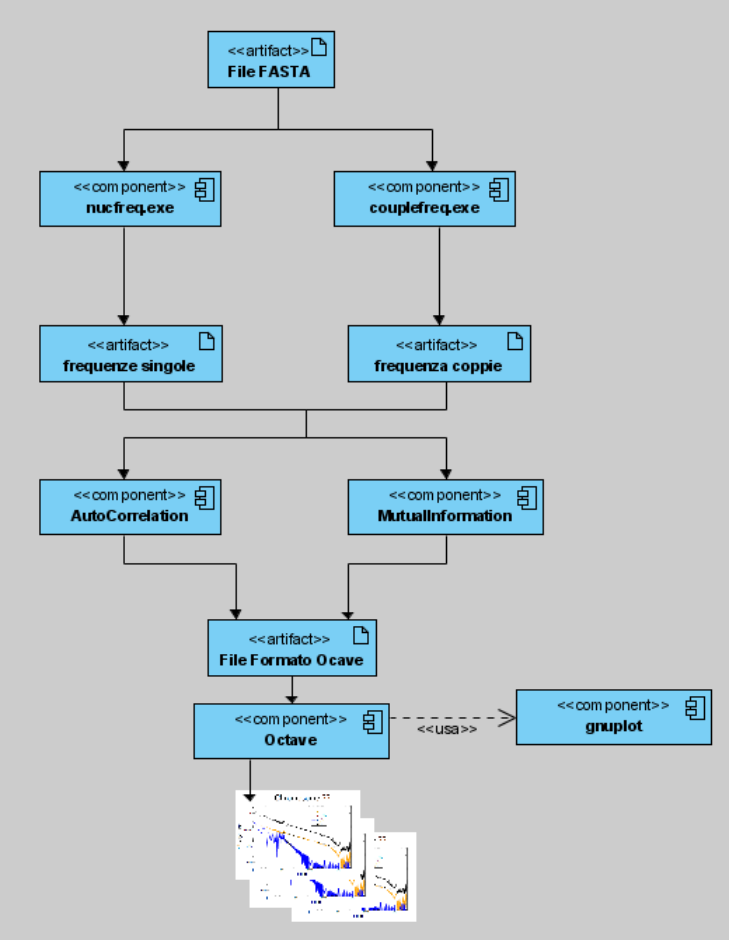

Figura 2.3: Schema del software

2.3 Correlazioni nel genoma umano

Basandoci su un precedente lavoro di Beirer [8], in questa sezione ci proponiamo di ricercare correlazioni in alcuni cromosomi del genoma umano. In particolare, i cromosomi 20, 21 e 22. Nel corso di questo scritto ci si riferirà a questi ed ad altri cromosomi tramite la dicitura ChrX dove X è il numero del cromosoma interessato. Nella tabella 2.2 sono riportate alcune caratteristiche dei cromosomi presi in considerazione. La scelta di questi tre particolari cromosomi del genoma umano era ai

tempi del precedente studio (2003) l'unica scelta ragionevole. Questi tre cromosomi, infatti, erano gli unici rilasciati dall'*International Human Genome Sequencing Consortium* come *finished* dove con questo termine si intende che la sequenza di DNA che li caratterizza è ricostruita in maniera quasi del tutto completa e se ne segnala l'affidabilità. Questo non vuol dire che le sequenze non vengano tutt'ora aggiornate e migliorate, ma questi aggiornamenti non introducono differenze sostanziali rispetto alle precedenti *release*. Il termine *finished* si contrappone al termine *draft* che si usa per indicare quelle sequenza che, seppur complete, non si possono definire ancora affidabili.

Anche nelle sequenze dei cromosomi rilasciati in stato finished però, possono esserci dei nucleotidi ambigui o non sequenziabili, che vengono denotati con il simbolo 'N'. La presenza di questi nucleotidi, di cui non possiamo dire nulla, introduce il problema di come considerarli all'interno dell'analisi statistica svolta. Si tratta di una questione molto importante, soprattutto in considerazione del fatto che questi nucleotidi generalmente non si trovano in singole posizioni della sequenza, ma sono raggruppati in zone che si estendono per migliaia, in alcuni casi per milioni, di basi. Sostituire queste parti con altre sequenze ci porrebbe di fronte al problema di definire una distribuzione dei 4 nucleotidi, con l'implicazione che inserire una sottostruttura potrebbe disturbare il rilevamento delle correlazioni che cerchiamo. D'altra parte eliminarle vorrebbe dire avvicinare parti della sequenza originariamente più distanti, creando la possibilità di rilevare false correlazioni. L'approccio che si è scelto è stato, quindi, di non modificare la sequenza, ma allo stesso tempo di escludere il simbolo N nel rilevamento delle frequenze dei singoli nucleotidi. Mentre, nel caso della frequenza delle coppie, quelle formate da uno o più nucleotidi N, non sono state considerate. I risultati ottenuti in questa fase hanno permesso di valutare, attraverso un confronto con dati attendibili, la bontà degli strumenti per l'analisi realizzati durante il tirocinio.

	Chr20	Chr21	Chr22
dimensione(mb)	62	47	50
Numero di geni	737	352	742
G+C%	44.1	40.8	47.9

Tabella 2.2: Alcune caratteristiche dei cromosomi 20, 21 e 22

2.3.1 Parametri dell'analisi

Per procedere all'analisi sono di fondamentale importanza alcune scelte riguardanti i parametri da utilizzare. Si è deciso di scegliere per la distanza k un valore massimo 10^6. Si tratta di un valore inferiore a quello utilizzato da Beirer ma come vedremo sufficiente per ottenere risultati significativi. Risulta, ad ogni modo, essere troppo grande per essere trattato senza un opportuno campionamento. Per ogni intervallo, rappresentato da due potenze successive di 10, sono stati selezionati gli estremi dell'intervallo e tutti i multipli dell'estremo inferiore compresi nell'intervallo.

2.3.2 Risultati

La figura 2.4 mostra i risultati ottenuti per i tre cromosomi, messi a confronto con i risultati del lavoro precedente. Come si può vedere $I(k)$ e $C^2_{WW}(k)$ mostrano una correlazione a lungo raggio fino ad una distanza di 10^5 nel cromosoma 22 e di 10^4 nei cromosomi 21 e 20. $C^2_{RR}(k)$ mostra invece correlazioni a lungo raggio fino ad una distanza di 10^4 su tutti e tre i cromosomi.

Si può notare come i grafici ottenuti attraverso le bioUtils e quelli del lavoro di Beirer messi a confronto risultino molto simili. Naturalmente i grafici non sono uguali e questo dipende dal campionamento che, naturalmente, ha sottratto una certa quantità di informazione e dal fatto che Beirer ha eseguito i calcoli utilizzando valori di k più grandi. Nonostante questo, si può vedere come, si sia giunti alle medesime conclusioni. Il campionamento ha agito positivamente nel calcolo di $I(k)$ e $C^2_{WW}(k)$, eliminando rumore piuttosto che informazione utile. D'altra parte esso influisce in maniera più negativa nel caso di $C^2_{RR}(k)$, essendo i valori di quest'ultima meno regolari.

L'equivalenza dei risultati ottenuti è ancora più evidente se andiamo a confrontare i coefficienti della legge di potenza calcolati con il metodo dei minimi quadrati ordinari, che è possibile vedere nella tabella 2.4 e nella tabella 2.3. L'intervallo, su cui effettuare la misura, è stato selezionato attraverso delle considerazioni di tipo grafico. Si è scelto di considerare come estremo superiore dell'intervallo quel valore di k dove la pendenza della curva varia in maniera significativa. Si può vedere chiaramente come il coefficiente della legge di potenza se calcolato su tutto il dominio avrebbe valori differenti, di conseguenza questo valore,che abbiamo visto dipendere

fortemente dall'intervallo scelto, può essere preso in considerazione solo come una caratteristica qualitativa della correlazione a lungo raggio [8].

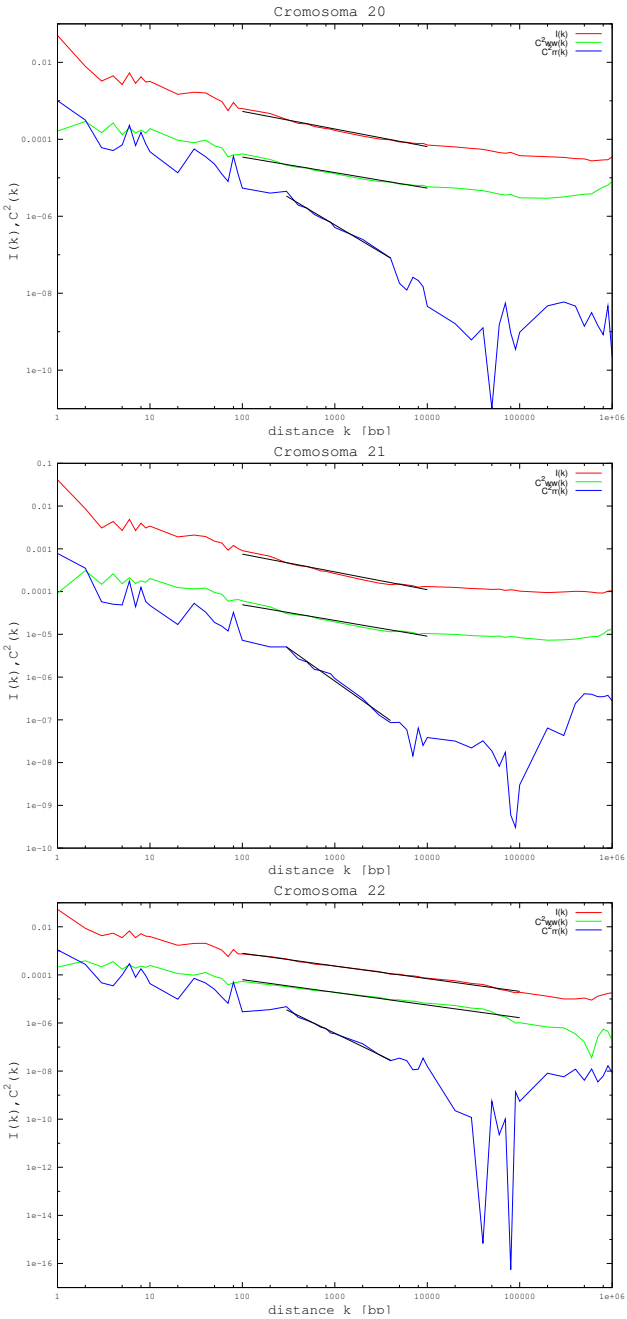

Figura 2.4: Risultati ottenuti con le bioUtils

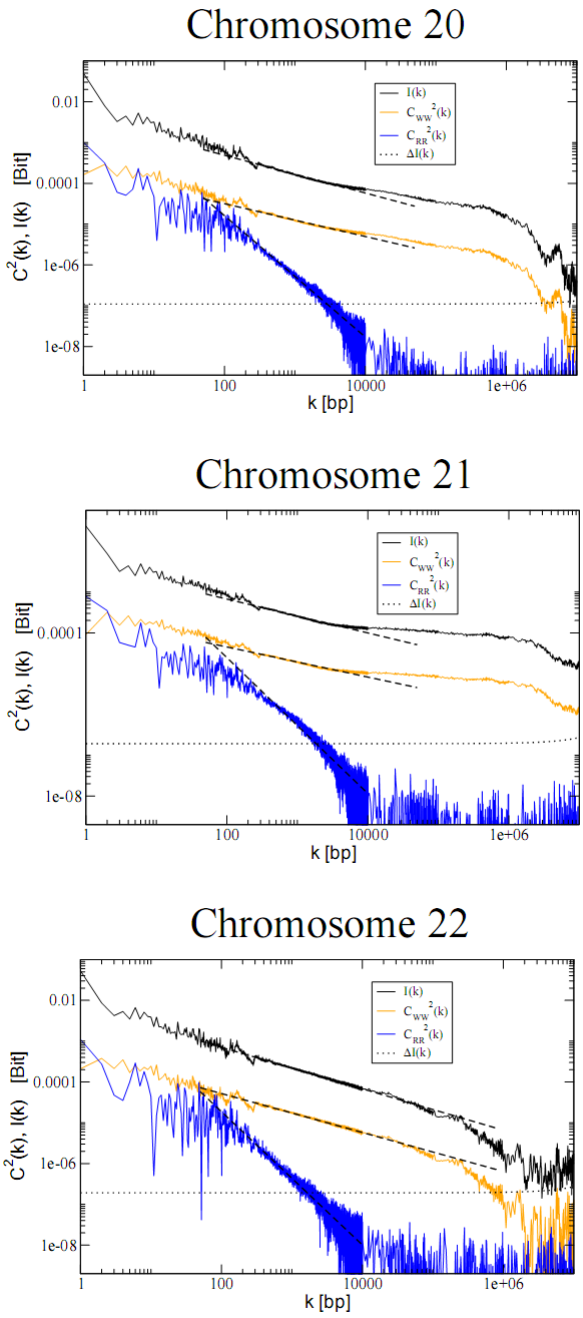

Figura 2.5: Risultati di Beirer [8]

	Chr20	Chr21	Chr22
$2\gamma_I$	0.46	0.41	0.51
range of fit(bp)	$100 .. 10^4$	$100 .. 10^4$	$100 .. 10^5$
$2\gamma_{WW}$	0.40	0.37	0.48
range of fit(bp)	$100 .. 10^4$	$100 .. 10^4$	$100 .. 10^5$
$2\gamma_{RR}$	1.49	1.67	1.63
range of fit(bp)	$100 .. 5 \times 10^3$	$100 .. 5 \times 10^3$	$100 .. 5 \times 10^3$

Tabella 2.3: Coefficienti della legge di potenza ottenuti dall'analisi di Beirer [8]

	Chr20	Chr21	Chr22
$2\gamma_I$	0.457	0.415	0.52
range of fit(bp)	$100 .. 10^4$	$100 .. 10^4$	$100 .. 10^5$
$2\gamma_{WW}$	0.403	0.367	0.474
range of fit(bp)	$100 .. 10^4$	$100 .. 10^4$	$100 .. 10^5$
$2\gamma_{RR}$	1.32	1.28	1.45
range of fit(bp)	$100 .. 5 \times 10^3$	$100 .. 5 \times 10^3$	$100 .. 5 \times 10^3$

Tabella 2.4: Coefficienti della legge di potenza calcolati con l'utilizzo delle bioUtils

Capitolo 3

Dai nucleotidi ai codoni

Nell'ultima sezione ,del capitolo precedente, è stata presentata e discussa la ricerca di correlazioni a lungo raggio all'interno di alcuni cromosomi umani. In questo capitolo si propone una possibile estensione a questo tipo di analisi. In particolare spostando l'attenzione dai singoli nucleotidi ai codoni.

3.1 Codoni

Nel primo capitolo, erano stati definiti i codoni, le triplette di nucleotidi, nell'RNA, che vengono tradotti in amminoacidi durante la sintesi proteica. Si ricorda, inoltre, che esistono 64 possibili triplette ma solo 20 amminoacidi e questa caratteristica è evidenziata dal fatto che codoni diversi possono codificare lo stesso amminoacido. Nella tabella 3.1 si può vedere più nello specifico la codifica. In particolare si può notare come gli unici amminoacidi codificati da un singolo codone siano la Metionina e il Triptofano.

Si rileva anche, che essendo la codifica relativa a sequenze di RNA, al posto del simbolo T che caratterizza il nucleotide contenente la base azotata Timina, troviamo il simbolo U cioè il nucleotide che contiene la base azotata Uracile. Dal punto di vista biologico, questa differenza è motivata dal fatto che la Timina crea legami più stabili rispetto all'Uracile e questa stabilità, che è una prerogativa importante del DNA, sarebbe invece un ostacolo all'interno di una molecola di RNA che per natura viene utilizzata per la sintesi e quindi degradata.

Per quanto riguarda questo studio è importante constatare come questa codifica,

Ala	A	GCU,GCC,GCA,GCG	Leu	L	UUA,UUG,CUU,CUC ,CUA,CUG
Arg	R	CGU,CGC,CGA,CGG,AGA ,AGG	Lys	K	AAA,AAG
Asn	N	AAU,AAC	Met	M	AUG
Asp	D	GAU,GAC	Phe	F	UUU,UUC
Cys	C	UGU,UGC	Pro	P	CCU,CCC,CCA,CCG
Gln	Q	CAA,CAG	Ser	S	UCU,UCC,UCA,UCG,AGU ,AGC
Glu	E	GAA,GAG	Thr	T	ACU,ACC,ACA,ACG
Gly	G	GGU,GGC,GGA,GGG	Trp	W	UGG
His	H	CAU,CAC	Tyr	Y	UAU,UAC
Ile	I	AUU,AUC,AUA	Val	V	GUU,GUC,GUA,GUG
start		AUG,GUG	Stop		UAG,UGA,UAA

Tabella 3.1: Tabella dei 20 amminoacidi ordinari e dei codoni che li codificano

con la U sostituita dalla T, sia corrispondente a quella che possiamo definire su di una sequenza di DNA. Questa corrispondenza è verificabile osservando in figura 3.1 il processo di trasmissione dell'informazione genetica. La molecola di RNA, infatti, viene sintetizzata in modo da risultare complementare ed antiparallela rispetto al filamento di stampo(anticodoni) del DNA [17].

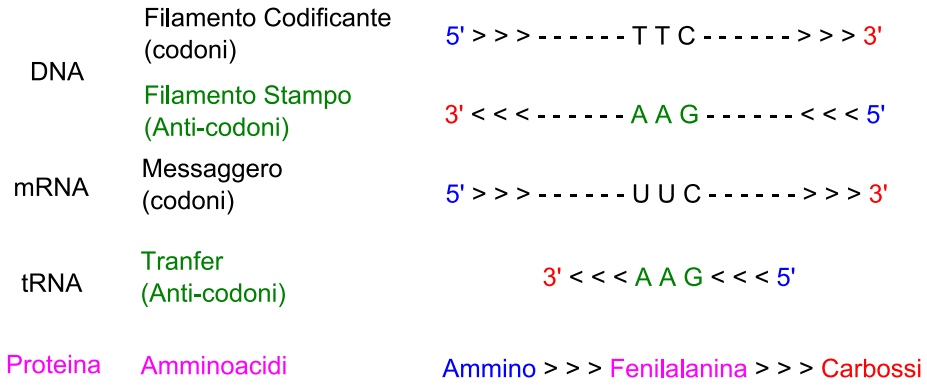

Figura 3.1: Esempio di sintesi Proteica

Nel prosieguo di questo capitolo ci riferiremo ad una generica tripletta di nucleotidi parlando di codoni commettendo in realtà un abuso di notazione. Infatti è evidente come il termine codone indichi non una qualsiasi tripletta della sequenza, ma bensì, una particolare tripletta di nucleotidi che appartiene a zone specifiche

di DNA, le cosiddette coding dna sequence (CDS). Queste zone, che si trovano all'interno dei geni, sono quelle che compongono la sequenza finale di RNA che va a codificare per una proteina.

3.2 Definizione del problema

Ridefiniamo ,in questa sezione, il problema del calcolo della mutua informazione e dell'autocorrelazione. In questo caso l'alfabeto a cui viene fatto riferimento, $A = \{AAA, AAC, ..., TTG, TTT\}$, è quello composto dai 64 possibili codoni, mentre le possibili coppie ad una distanza k sono tutti i possibili accoppiamenti di questi 64 codoni, cioè 4.096 possibili coppie. Se precedentemente però assumevamo di definire la distanza k in coppie di basi (bp), in questo caso, è stato scelto di quantificare la distanza in base al numero di codoni che si trovano tra quelli presi in considerazione. Diremo, quindi, che due codoni si trovano ad una distanza k, se questi sono separati nella sequenza da $k-1$codoni. Si tratta di un'assunzione importante, anche se poteva apparire ovvia visto e considerato che ora riteniamo il codone come elemento base della sequenza. Un'altra differenza rispetto alla definizione data per i nucleotidi riguarda la lunghezza del vettore $\overrightarrow{a}$ che adesso è formato da 64 possibili valori invece che 4.

Come si può immaginare, la complessità del problema da una parte cresce per via, dell'alfabeto dei simboli che ora è 16 volte più grande del precedente, dall'altra la lunghezza della sequenza si riduce di 3 volte. Un problema dal punto di vista computazionale che ha continuato a rappresentare un ostacolo è stato quello del calcolo della frequenza delle coppie. Ma adottando la soluzione del campionamento, già vista prima, è stato possibile ottenere i risultati voluti in un tempo comparabile al caso precedente.

Un'ulteriore decisione di notevole importanza è stata quella sulle regole da seguire per dividere la sequenza in triplette. Non si tratta di una scelta semplice. Infatti per le zone not-coding, che poi sono quelle oggetto di uno studio più approfondito nel corso di questo lavoro non esistono, informazioni che ci dicano che un modo sia migliore di un altro. Mentre, per quanto riguarda le zone coding solo una certa divisione ha significato, vista la loro funzionalità. In considerazione di ciò si è deciso di fare la scelta più semplice, cioè quella di dividere la sequenza originale in triplette

a partire dal primo nucleotide. Anche nella gestione dei nucleotidi di tipo N, si è fatta una scelta molto netta, decidendo di non considerare nel calcolo delle frequenze tutte quelle triplette che contenessero almeno un simbolo 'N'. Tale scelta è in linea con quella fatta in precedenza con i singoli nucleotidi, infatti le triplette sono state lasciate all'interno della sequenza in maniera da non falsare le distanze.

3.3 Le bioUtils per i codoni

Come accennato nel capitolo precedente, per il calcolo delle correlazioni tra codoni, è stata realizzata un'estensione delle bioUtils che offre un insieme di programmi simili a quelli implementati per i singoli nucleotidi. Come si può vedere in figura 3.2 lo schema è praticamente identico tranne che per l'aggiunta di un file ulteriore in input al programma che calcola la funzione di autocorrelazione. Questo file che è stato chiamato acVector, rappresenta il valore del vettore $\overrightarrow{a}$. Il file acVector è un semplice file di testo che ha una riga per ogni possibile codone e un valore zero oppure uno per ognuna di queste righe. Uno o zero indicano, rispettivamente se il codone deve o non deve essere preso in considerazione durante il calcolo della funzione di autocorrelazione. Si può vedere un esempio in figura 3.3. La scelta di creare un file apposito deriva dal fatto che se prima si poteva, essendo in presenza di soli 4 valori, inserire queste informazioni manualmente tramite linea di comando, ora con 64 possibilità una procedura simile sarebbe stata inaccettabile. All'inizio si era creato un solo file acVector, che veniva modificato ogni volta a seconda delle esigenze, poi, per ottenere un'ulteriore comodità sono stati prodotti file acVector già pronti per le correlazioni che ci si proponeva di trovare sui vari cromosomi, in modo da averli sempre pronti all'uso e per evitare che modificando continuamente lo stesso file si potessero introdurre inavvertitamente degli errori.

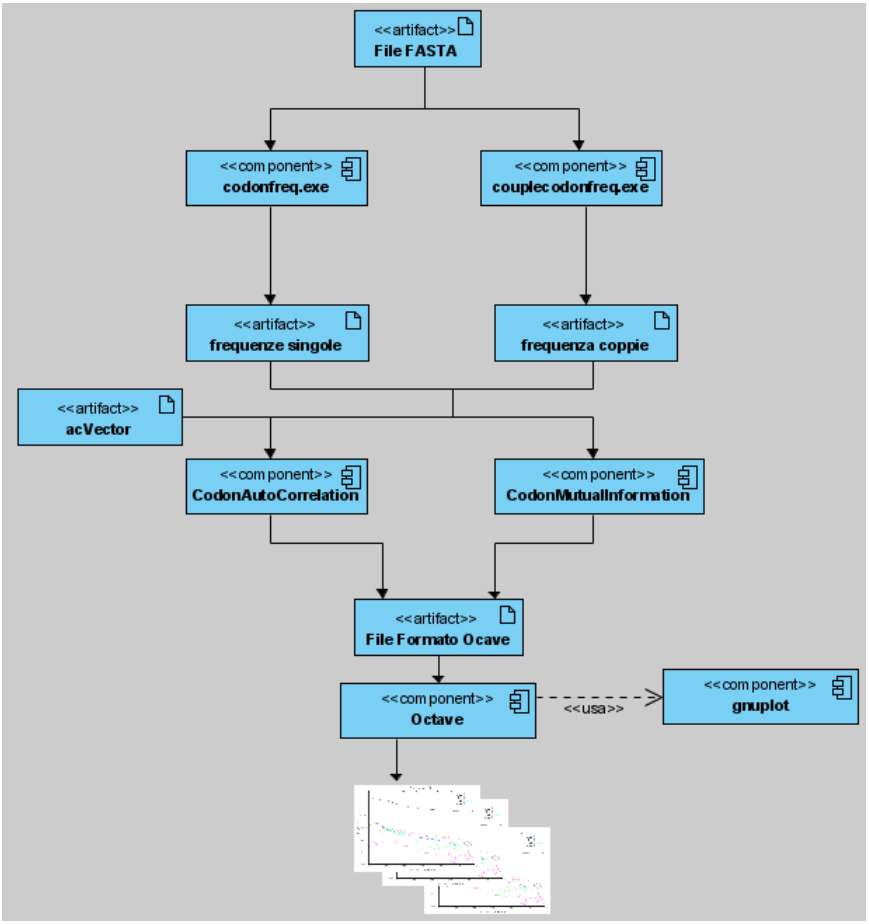

Figura 3.2: Schema software per i codoni

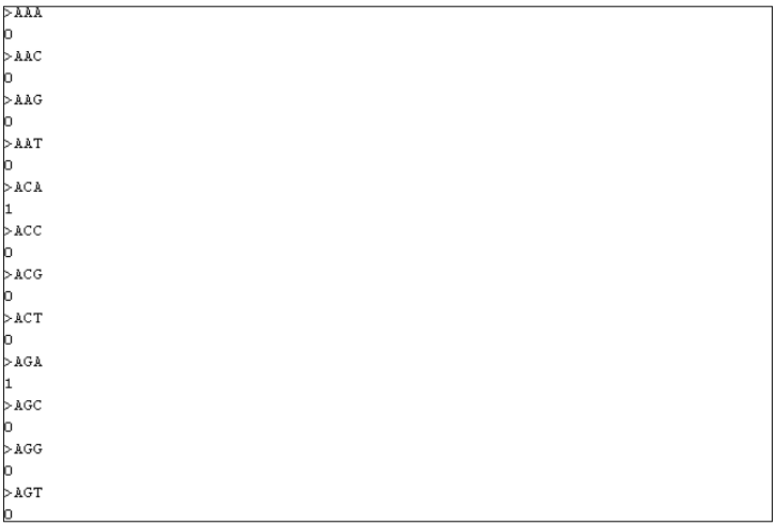

Figura 3.3: Frammento di un file acVector: utilizzando ad esempio un file con questa configurazione otterremo l'autocorrelazione dei codoni ACA e AGA

3.4 Risultati

Nelle figure 3.4, 3.5, 3.6, 3.7, 3.8 e 3.9 sono mostrati i risultati dell'analisi. Sono state calcolate 20 funzioni di autocorrelazione, una per ogni gruppo di codoni che codifica per un particolare amminoacido. Si può constatare che il cromosoma 22 non sembra mostrare correlazioni a lungo raggio, né per $I(k)$ né per $C^2(k)$. Il cromosoma 20 e il cromosoma 21 mostrano, invece, correlazioni a lungo raggio per $I(k)$, il primo fino ad un valore di 10^5 e il secondo fino ad un valore di 10^4. Nella tabella 3.2 è possibile vedere i coefficienti della legge di potenza.

	Chr20	Chr21
$2\gamma_I$	0.21	0.11
range of fit(bp)	10^3 .. 10^5	10^3 .. 10^4

Tabella 3.2: Coefficienti della legge di potenza

Dal punto di vista della funzione di autocorrelazione, nessun cromosoma sembra mostrare correlazioni a lungo raggio. É però interessante notare come alcune curve abbiano un andamento molto instabile, mentre altre hanno un comportamento più

regolare e inoltre si avvicinano maggiormente, dal punto di vista della pendenza, alla funzione di mutua informazione.

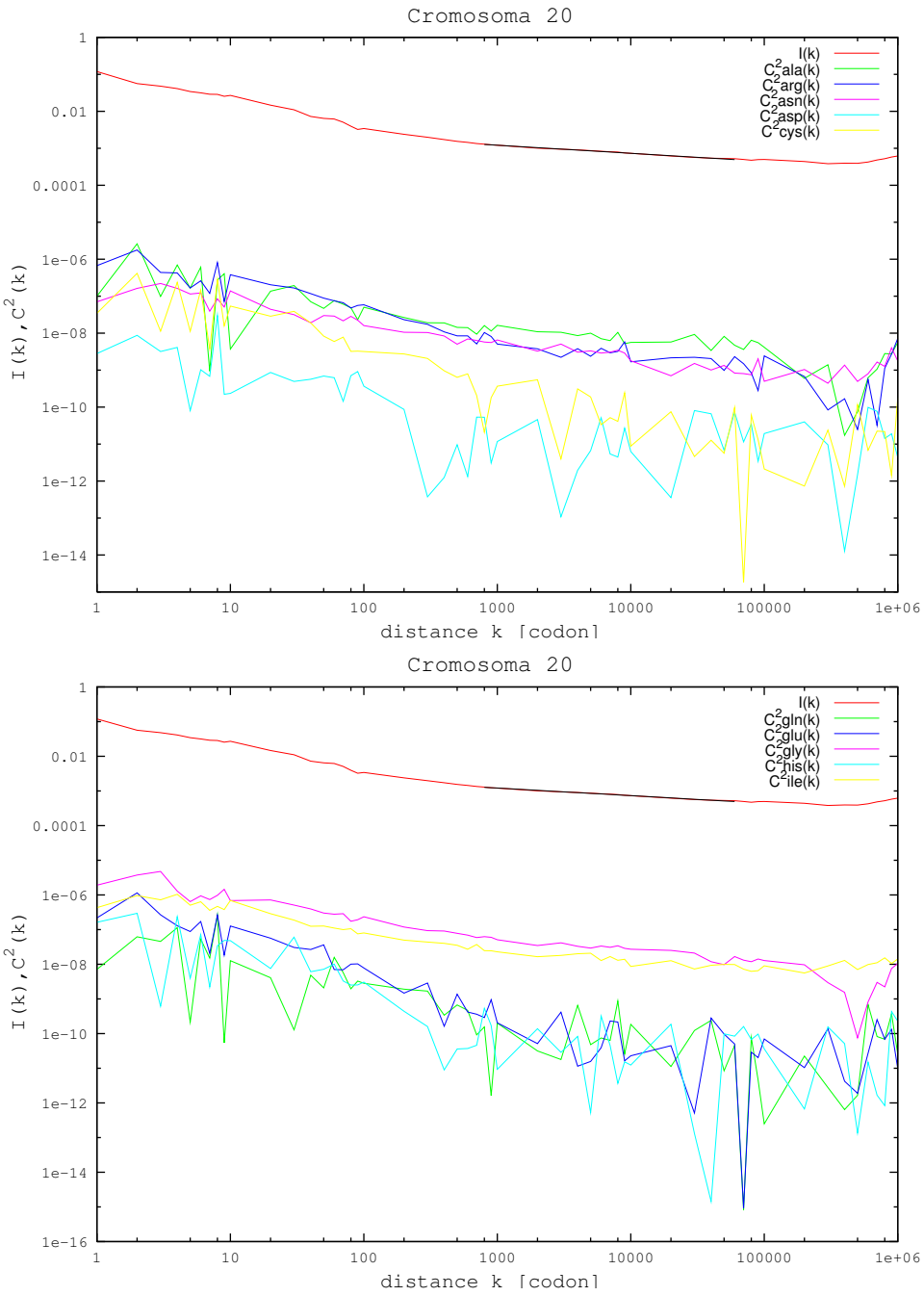

Figura 3.4: Mutua informazione e funzioni di autocorrelazione per il cromosoma 20. In alto: alanina , arginina, asparagina, acido aspartico, cisteina. In basso: glicina, acido glutammico, glutammina, istidina, isoleucina.

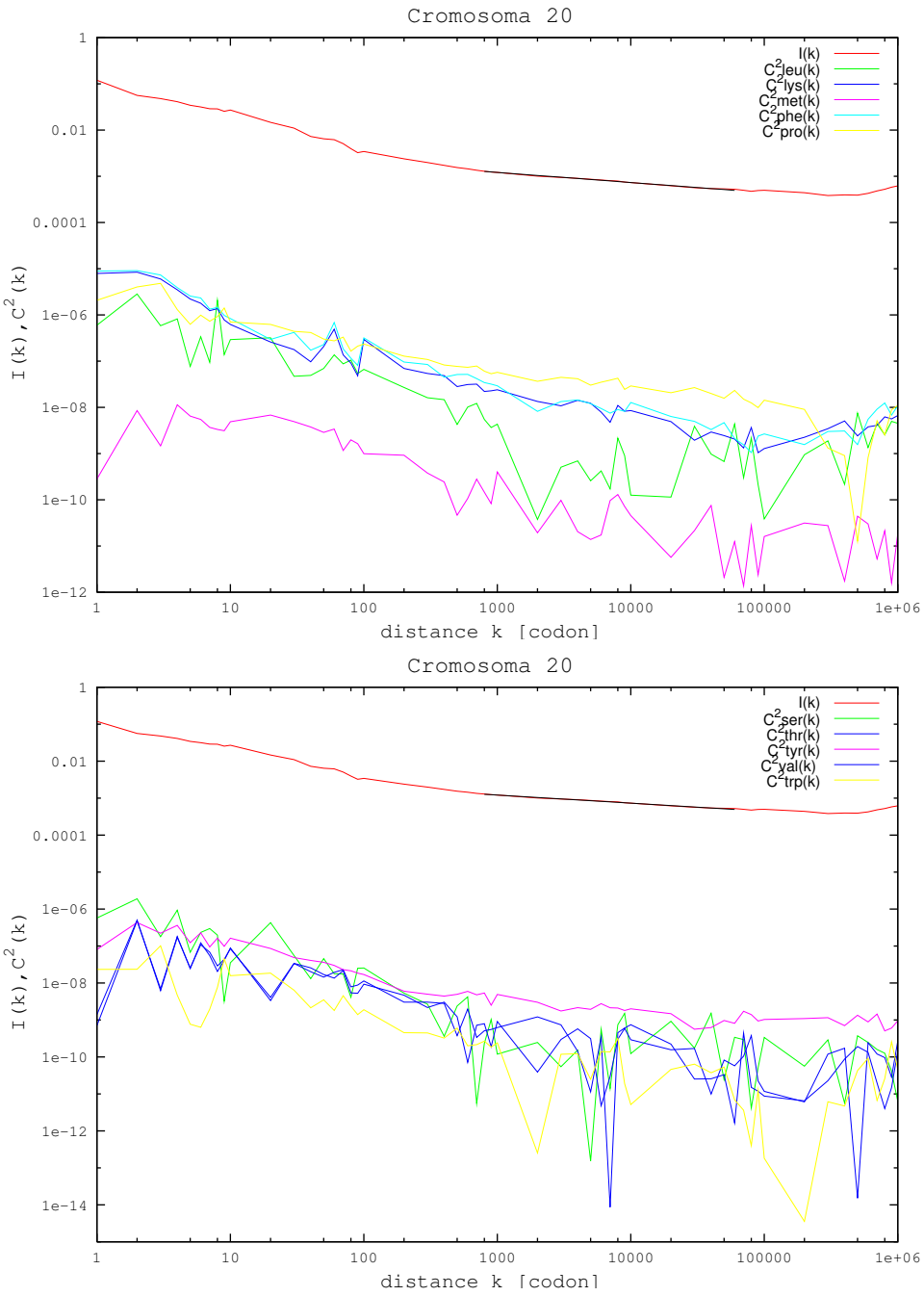

Figura 3.5: Mutua informazione e funzioni di autocorrelazione per il cromosoma 20. In alto: leucina, lisina, metionina, alanina, prolina. In basso: serina, treonina, triptofano, tirosina, valina.

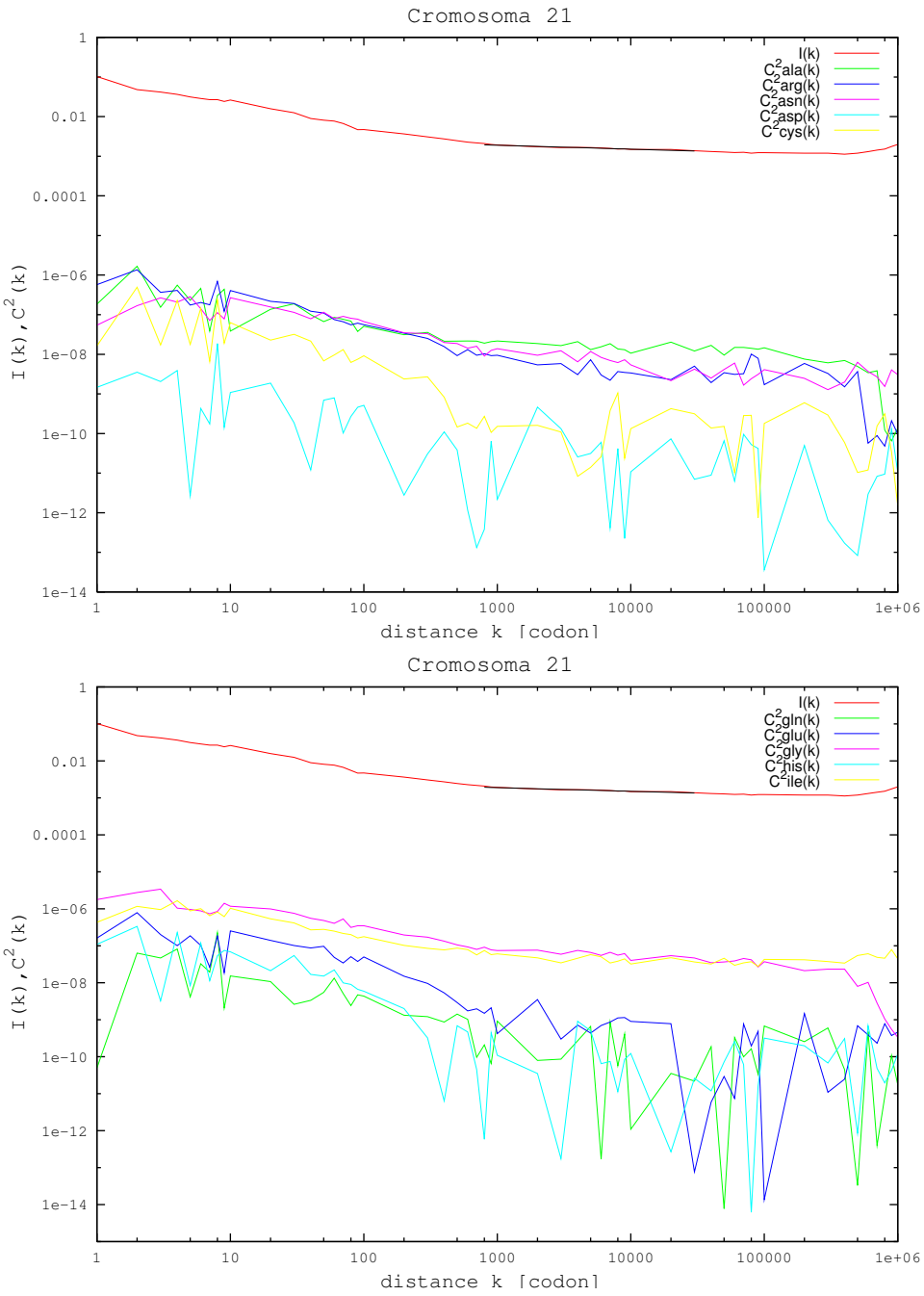

Figura 3.6: Mutua informazione e funzioni di autocorrelazione per il cromosoma 21. In alto: alanina , arginina, asparagina, acido aspartico, cisteina. In basso: glicina, acido glutammico, glutammina, istidina, isoleucina.

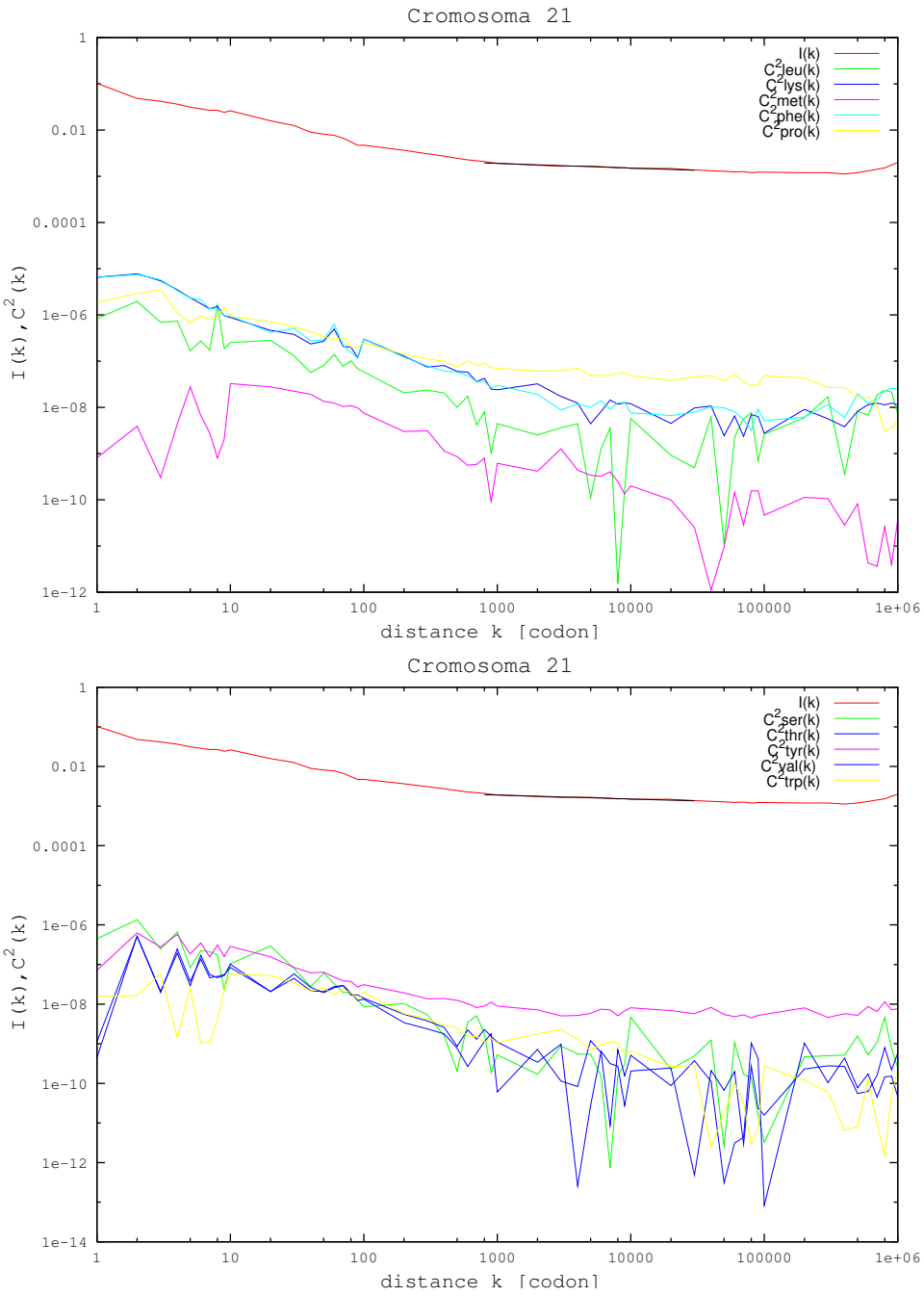

Figura 3.7: Mutua informazione e funzioni di autocorrelazione per il cromosoma 21. In alto: leucina, lisina, metionina, alanina, prolina. In basso: serina, treonina, triptofano, tirosina, valina.

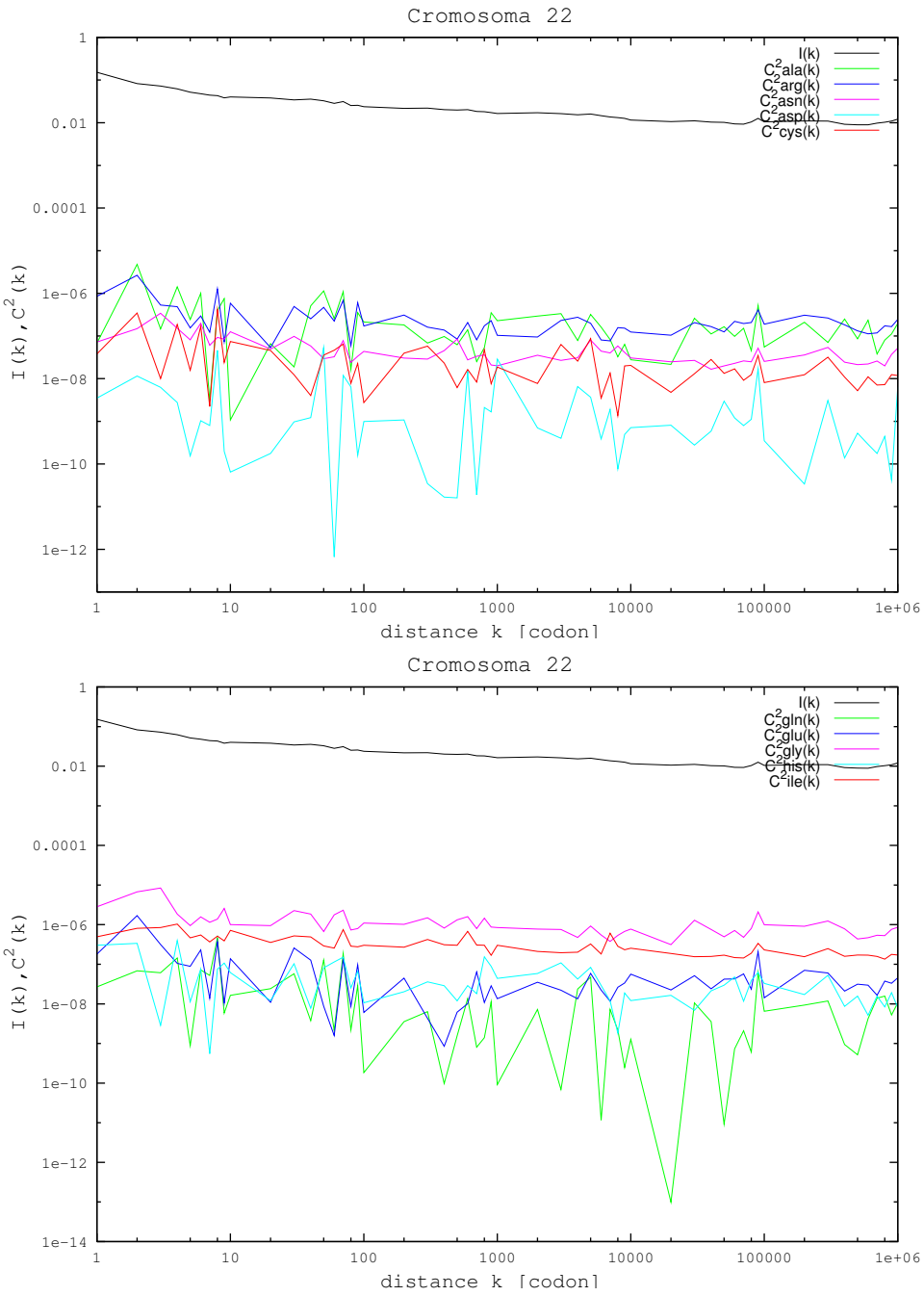

Figura 3.8: Mutua informazione e funzioni di autocorrelazione per il cromosoma 22. In alto: alanina , arginina, asparagina, acido aspartico, cisteina. In basso: glicina, acido glutammico, glutammina, istidina, isoleucina.

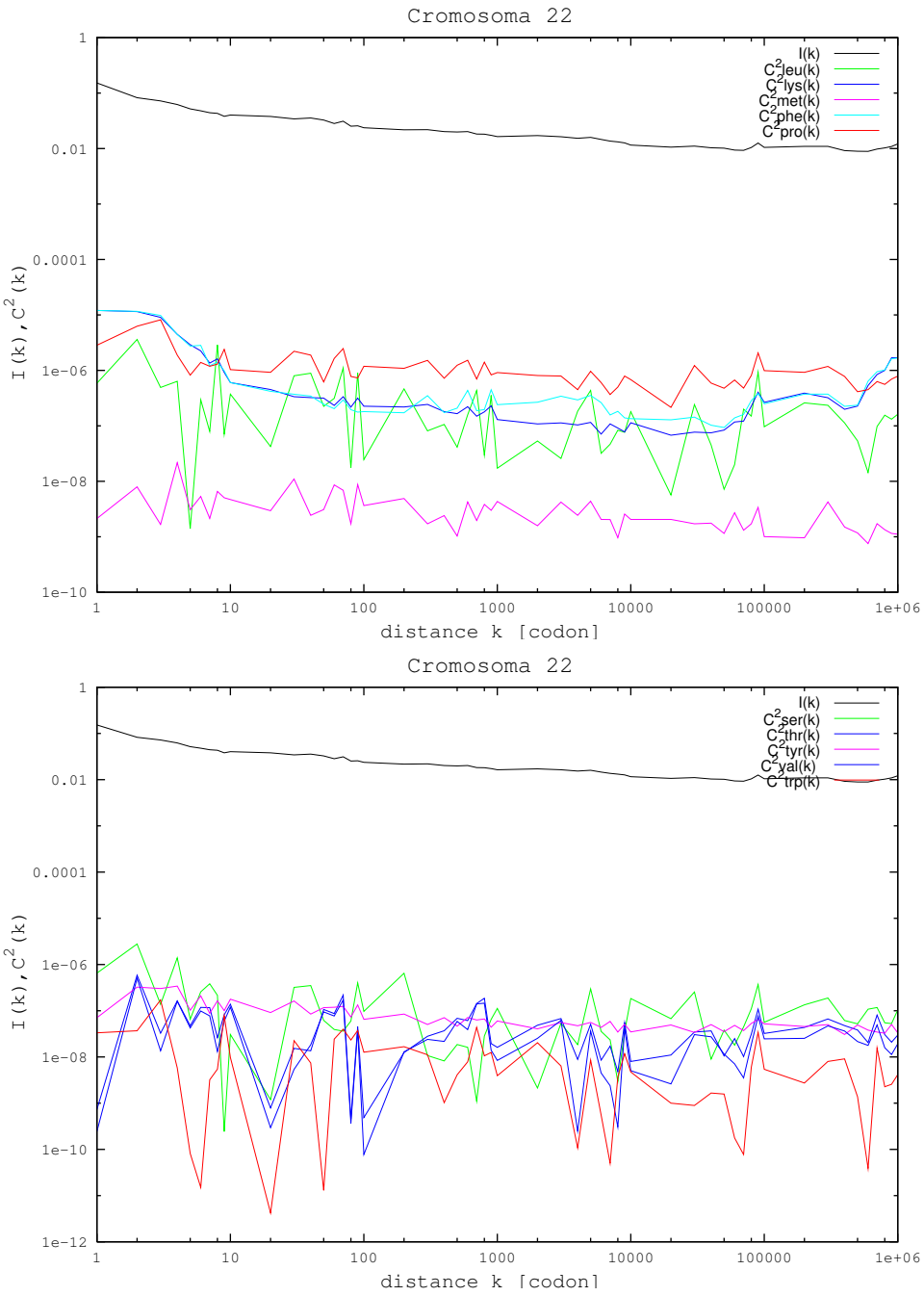

Figura 3.9: Mutua informazione e funzioni di autocorrelazione per il cromosoma 22. In alto: leucina, lisina, metionina, alanina, prolina. In basso: serina, treonina, triptofano, tirosina, valina.

Capitolo 4

Analisi della correlazione tra codoni

In questo capitolo ci si propone di analizzare, più nello specifico, i risultati ottenuti nel precedente capitolo, andando in particolare a studiare le correlazioni individuate tra le coppie di codoni.

4.1 Analisi delle funzioni di autocorrelazione

Come abbiamo visto nel terzo capitolo, diverse curve, pur non presentando correlazioni a lungo raggio, mostrano un andamento simile a quello della mutua informazione. In questa sezione analizzeremo alcune di queste funzioni, per cercare di capire quali componenti, e quindi triplette, influenzino maggiormente il loro comportamento.

Per ragioni di tempo, in questa parte del tirocinio si è deciso di sottoporre, a quest'analisi approfondita, solo uno dei cromosomi in esame. Si è deciso di scartare immediatamente il cromosoma 22 che, come visto, è quello che ha mostrato le caratteristiche meno interessanti. Dovendo scegliere tra i due rimanenti e non avendo oggettive ragioni per preferirne uno all'altro, si è deciso di prendere in considerazione il più piccolo dei due, cioè il 20. Si analizzeranno due funzioni di autocorrelazione che hanno mostrato un andamento particolarmente significativo. Una riguardante i codoni che codificano l'Isoleucina e una per quelli che codificano la Glicina. É doveroso ricordare, anche se potrà apparire ripetitivo, che ogni qual volta parliamo di codoni, non ci riferiamo ai codoni reali ma a generiche triplette di nucleotidi. Inoltre, definirle come codificanti per un particolare amminoacido, in questa situ-

azione, non ha necessariamente una valenza biologica ma ci serve per meglio gestire e classificare le informazioni provenienti dalle varie funzioni.

4.1.1 Isoleucina

Come si può vedere in figura 3.7, la funzione di autocorrelazione dell'Isoleucina ha un andamento molto simile a quello della mutua informazione, anche se, osservandola più da vicino, dimostra comunque di essere molto meno regolare rispetto a quest'ultima. I codoni che codificano per l'Ilenina sono: ATT,ATC,ATA.

Per verificare, quali di questi codoni ne influenza maggiormente l'andamento, calcoliamo le funzioni di autocorrelazione per le possibili coppie formate da queste triplette. Le possibili coppie sono: ATT-ATC, ATT-ATA, ATC-ATA. Ricordiamo che l'ordine non è da considerarsi importante, quindi: ATT-ATC è uguale ad ATC-ATT. Dai risultati visibili in figura 4.1, si può notare come anche con questa scomposizione le curve mostrino un andamento qualitativamente simile, anche se la coppia ATT-ATA mostra sicuramente valori più grandi è un andamento più stabile.

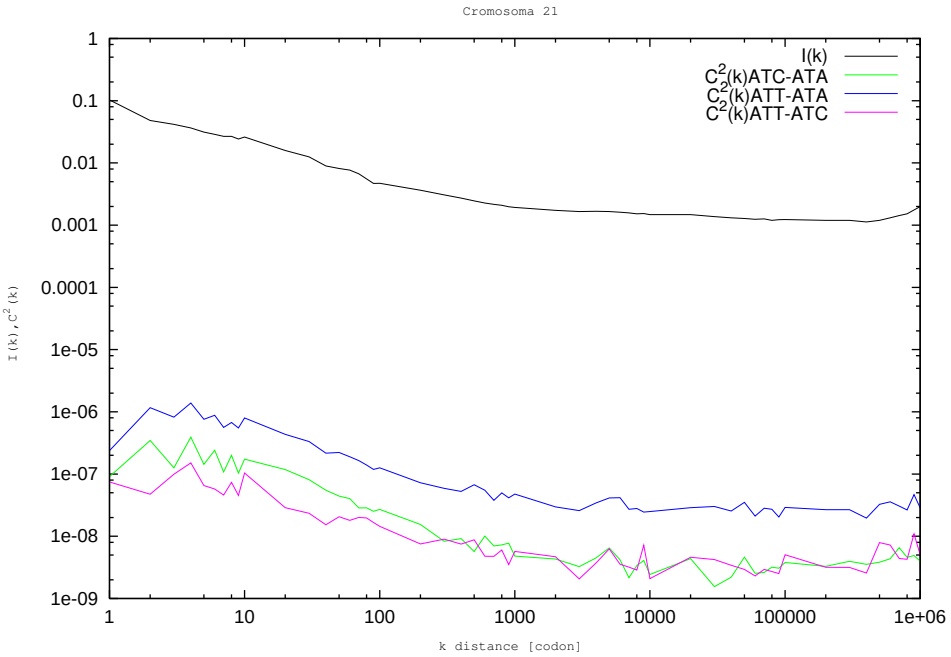

Figura 4.1: Funzioni di autocorrelazione calcolate sulle coppie ATC-ATA, ATT-ATA,ATT-ATC.

Come vedremo successivamente questa coppia presenterà, in generale, caratteristiche rilevanti rispetto alle altre.

4.1.2 Glicina

Partendo dalle stesse considerazioni della sezione precedente, si scompone la funzione di autocorrelazione della Glicina in tutte le possibili coppie formate dai codoni che, codificano per tale amminoacido. In particolare questi codoni sono: GGT,GGC,GGA,GGG. Le possibili coppie sono GGT-GGC, GGT-GGA, GGT-GGG, GGC-GGA, GGC-GGG, GGA-GGG. In questo caso non abbiamo differenze molto marcate. Nonostante ciò, si può notare che la coppia GGT-GGA mostra molta instabilità rispetto alle altre nella parte finale, al contrario la coppia GGT-GGG mostra un andamento molto stabile.

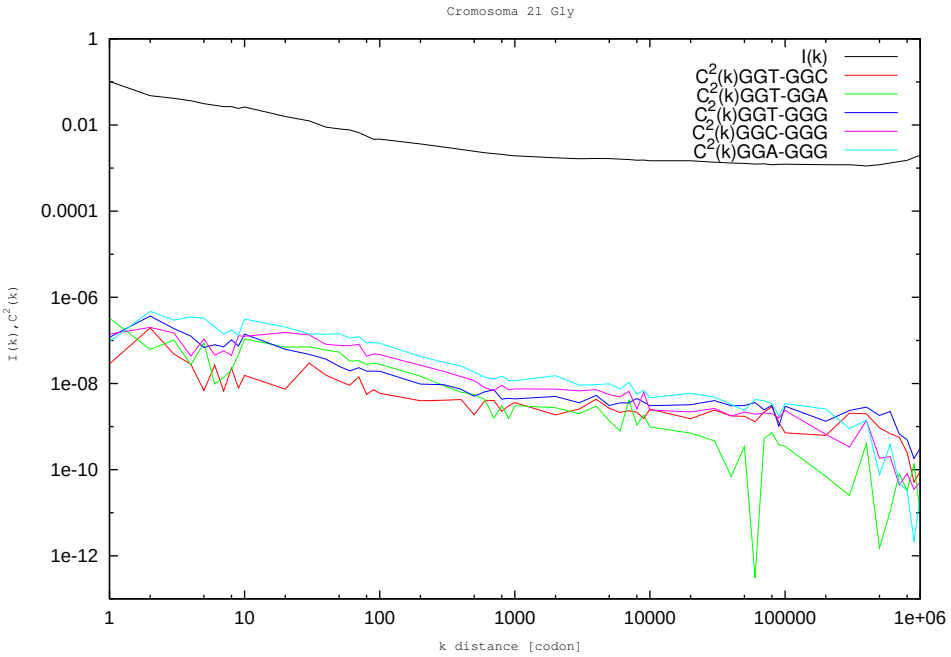

Figura 4.2: Funzioni di autocorrelazione calcolate sulle coppie GGT-GGC, GGT-GGA, GGT-GGG, GGC-GGA, GGC-GGG, GGA-GGG.

4.2 Coppie con forte autocorrelazione

Anche in seguito alle analisi svolte precedentemente, si può constatare come sia più interessante calcolare la correlazione tra le possibili coppie di codoni piuttosto che su altri loro particolari ragruppamenti. Quando si è calcolata l'autocorrelazione tra nucleotidi si è potuto semplicemente calcolare la funzione su tutte le possibili coppie, visto il numero limitato. Nel caso dei codoni è stata necessaria una strategia più raffinata. Si è deciso di effettuare un analisi statistica che ci permettesse di inviduare coppie con una correlazione significativamente più alta rispetto ad altre e di lavorare, successivamente, solo su quelle.

4.2.1 La funzione di autocorrelazione media

Per poter individuare coppie particolari, è fondamentale stabilire in maniera più rigorosa cosa si intende con una "correlazione significativamente più alta". Si è deciso di confrontare le correlazioni delle diverse coppie in termini di distanza dalla correlazione media e si è quindi stabilito un livello di soglia, oltre il quale ritenere questo valore di correlazione più rilevante rispetto ad altri. La correlazione media è stata calcolata, in funzione di k, come la media aritmetica tra i valori di $C(k)$ per tutte le possibili combinazioni di coppie, che nel caso dei codoni sono 2.080. Questo valore scaturisce dall'osservazione che le possibili coppie sono tutte le combinazioni semplici con ripetizioni di 64 valori presi ad insiemi di 2. Con ripetizioni, perché si è tenuto conto anche delle coppie formate dallo stesso codone. In figura 4.3 si può vedere il grafico risultante.

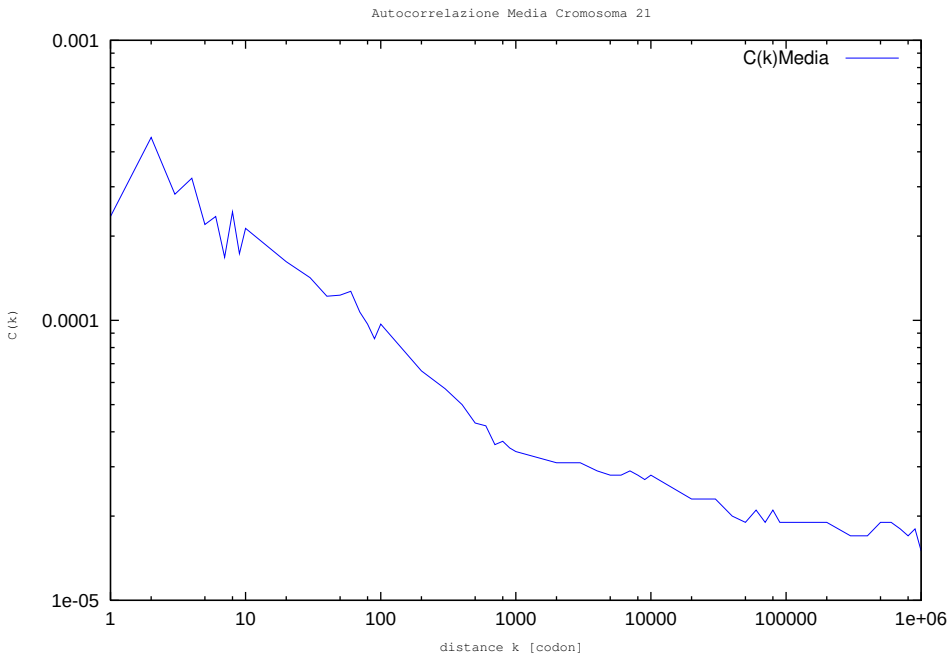

Figura 4.3: Funzione di autocorrelazione media sul cromosoma 21

Si sono quindi calcolate e confrontate tramite un ulteriore programma aggiunto alle bioUtils, *CodonMatrix,* le varie funzioni di autocorrelazione e la loro distanza dalla media per ogni valore di k. Sono quindi state selezionate quelle che rispettavano determinate caratteristiche. In particolare, sono state prese in considerazione solo quelle coppie che avevano valori più alti della media di almeno 10^{-4} per ogni k. Questo valore, è stato scelto in base a considerazioni di tipo empirico. In tabella 4.1 si possono vedere i risultati ottenuti con altri valori di soglia.

soglia	0	10^{-6}	10^{-5}	10^{-4}	10^{-3}
coppie individuate	121	120	100	23	0

Tabella 4.1: Valori di soglia scelti e risultato dell'esecuzione del programma *CodonMatrix.* Si può notare come, per il valore di 10^{-4}, si ottenga una significativa riduzione delle coppie individuate. Questo lo rende un buon valore di soglia per i nostri scopi.

Come si può notare, utilizzando questo valore, otteniamo 23 coppie notevoli. É interessante notare come queste 23 coppie siano formate da un ristretto numero di codoni, che nello specifico sono: AAA, AAT, ATA, ATT, TAA, TAT, TTA, TTT.

Inoltre, questo piccolo insieme è formato da codoni composti solo da Adenina e Timina.

4.2.2 Confronto con la deviazione standard

I risultati, fino ad ora trovati, ci permettono di affermare che le funzioni di autocorrelazione di queste 23 coppie hanno valori superiori alla media, ma non ci bastano a sostenere una loro significatività dal punto di vista statistico. Per poter fare questo ulteriore passo, è necessario introdurre il confronto con un'ulteriore grandezza: la deviazione standard. Fissato un k, avremo 2.080 variabili casuali $C_{ij}(k)$. La loro distribuzione congiunta, per il teorema del limite centrale tende ad una distribuzione normale $N(\mu(k); \sigma^2(k))$ dove nel nostro caso $\mu(k)$ corrisponde all'autocorrelazione media e $\sigma^2(k)$ alla deviazione standard al quadrato. Si consideri che la probabilità di trovare correlazioni che distano dal valor medio per più di 3 volte la deviazione standard è prossima all'1%, come si vede in tabella 4.2 e in figura 4.4. Esprimendo in termini di deviazione standard la differenza tra una particolare $C_{ij}(k)$ e la $C(k)$ media, che indicheremo come $\Delta_{ij}(k)$, si potranno considerare statisticamente significative le correlazioni di tutte le coppie di codoni per le quali , $\frac{\Delta_{ij}(k)}{\sigma(k)} > 3$.

$P[\mu - \sigma \leq x \leq \mu + \sigma]$	0,6826
$P[\mu - 2\sigma \leq x \leq \mu + 2\sigma]$	0,9544
$P[\mu - 3\sigma \leq x \leq \mu + 3\sigma]$	0,9974

Table 4.2: In questa tabella vengono riportati i valori delle probabilità di una variabile con distribuzione normale all'interno di specifici range

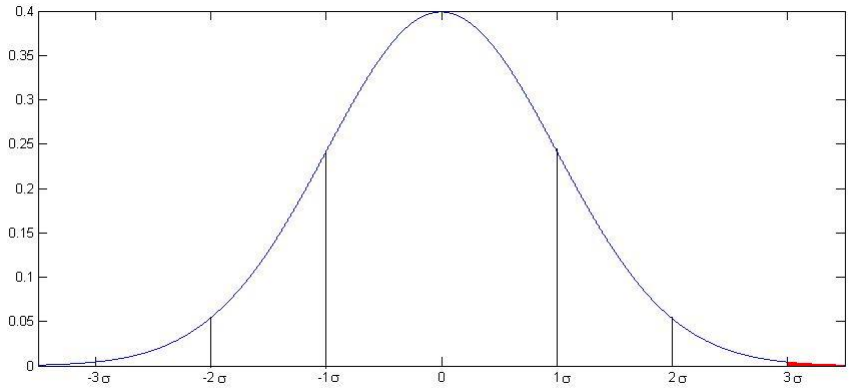

Figura 4.4: Distribuzione normale standard

L'analisi dei risultati in base a questi parametri ha mostrato la significatività delle correlazioni di tutte le coppie individuate, con valori di $\Delta_{ij}(k)$ fino a 10 volte maggiori della deviazione standard. Nelle figure 4.5 e 4.6 si possono vedere , come esempio, questi valori calcolati rispettivamente sulla coppia AAA-ATA e AAT-ATT.

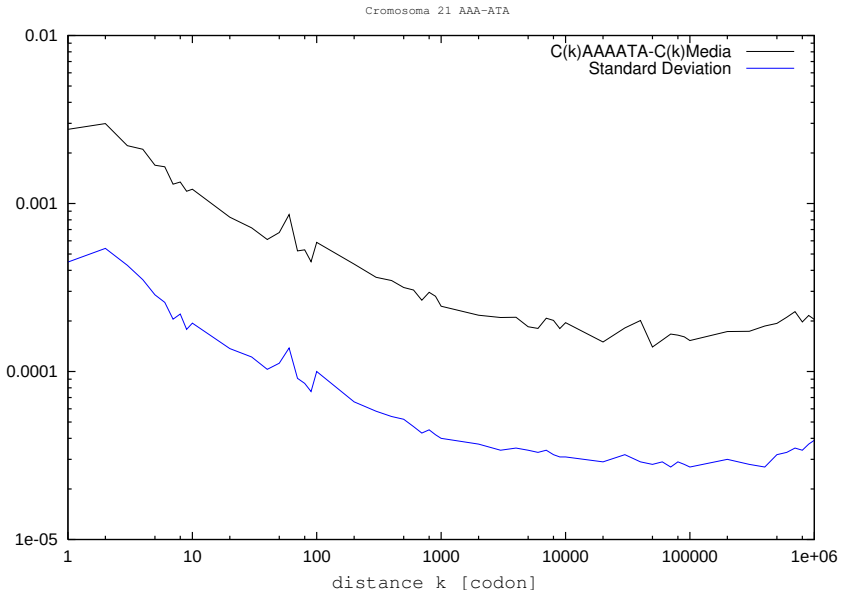

Figura 4.5: Confronto tra la distanza dalla media della funzione di autocorrelazione di AAA-ATA e deviazione standard

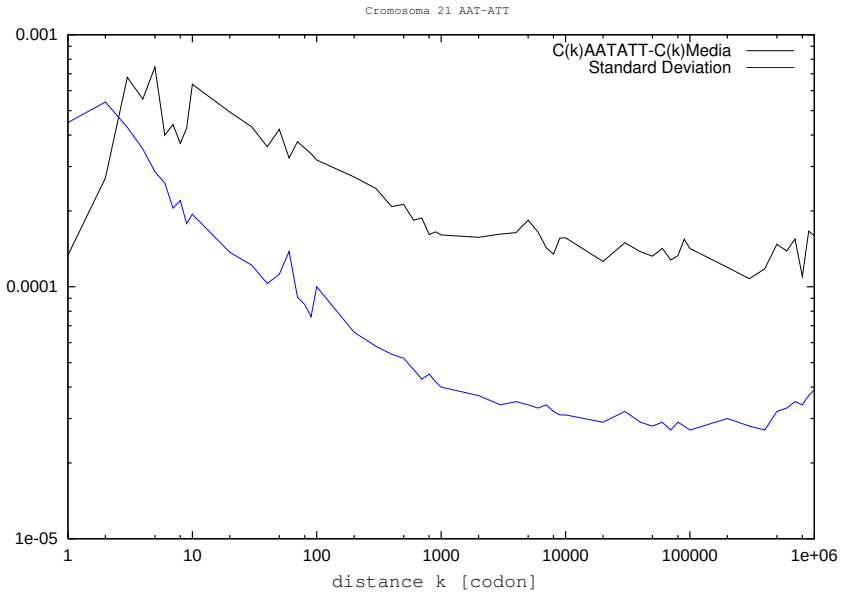

Figura 4.6: Confronto tra la distanza dalla media della funzione di autocorrelazione di AAT-ATT e deviazione standard

Capitolo 5

Conclusioni

Alla fine dell'attività di tirocinio si può dire di avere raggiunto risultati soddisfacenti dal punto di vista dello sviluppo del software. Pur nella sua semplicità, e nonostante l'utilizzo di risorse hardware relativamente povere, il pacchetto di applicazioni realizzato ha permesso di trattare un'enorme mole di dati, ottenendo informazioni significative.

Per quanto riguarda le analisi sui nucleotidi abbiamo confermato i risultati degli studi precedenti di Beirer [8]. Questo ha messo inoltre in evidenza come le migliorie portate alla sequenza del genoma umano dal 2003 ad oggi non hanno modificato proprietà rilevate attraverso studi precedenti.

É stato inoltre introdotto un interessante elemento di novità nello studio del DNA, attraverso un'analisi rivolta alle triplette di nucleotidi sulla parte not-coding della sequenza . Quest'analisi ha permesso di individuare correlazioni a lungo raggio nella mutua informazione calcolata sui cromosomi 20 e 21 del genoma umano.

Inoltre, è stato possibile evidenziare come alcune triplette abbiano una correlazione significativamente più alta rispetto ad altre. Sicuramente è molto rilevante il fatto che queste triplette siano composte soltanto da Adenina e Timina. Per certi versi questo non è del tutto inaspettato, visto che anche nell'analisi dei singoli nucleotidi, Adenina e Timina mostravano una forte correlazione. Non era comunque un risultato banale trovare una simile correlazione anche tra triplette che li contengono. Ciò potrebbe implicare una particolare distribuzione di questi due nucleotidi all'interno della sequenza, non solo singolarmente, ma anche se presi in gruppi che li contengono. Per quanto riguarda gli sviluppi futuri ci sono molte strade da poter

seguire.

In primo luogo un ulteriore sviluppo del software fino ad ora implementato, da una parte rivolto a migliorarne le prestazioni e a offrire nuove caratteristiche, dall'altra sarebbe interessante un intervento volto alla realizzazione di un'interfaccia più *user friendly*. Dal punto di vista dei risultati statistici, sarebbe di grande interesse analizzare più nello specifico le correlazioni trovate, indagando sulle possibili implicazioni biologiche e cercando ulteriori conferme nell'analisi di altre parti del genoma umano per constatare se tali proprietà si ripresentano. Inoltre, uno studio analogo potrebbe essere eseguito sui genomi di altri esseri viventi, in modo da potere proporre un confronto.

Dal punto di vista personale, questa esperienza è stata molto utile per la mia formazione, permettendo di approcciarmi per la prima volta al mondo della ricerca. Inoltre, non solo mi ha permesso di sfruttare le conoscenze teoriche e pratiche maturate durante questi anni di studio, ma anche di utilizzarle per risolvere problemi di tipo interdisciplinare. Infatti, ho potuto provare in prima persona che la ricerca scientifica procede per mezzo dell'effetto sinergico ottenuto superando la difficoltà di integrare le conoscenze provenienti da diverse aree di studio.

Bibliografia

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter. *Biologia molecolare della cellula* (Quarta Edizione) 2004

[2] The Chimpanzee Sequencing and Analysis Consortium. "*Initial sequence of the chimpanze genome and comparison with the human genome*". *Nature* 437 69-87

[3] Thomas M. Cover, Joy A. Thomas. *Elements of Information Theory* (Second Edition) 2006

[4] Karmeshu & A. Krishnamachari. *"Sequence Variability and Long Range Dependence in DNA: An Information Theoretic Perspective".* ICONIP 2004, LNCS 3316, pp. 1354-1361, 2004

[5] C.E. Shannon. *"A mathematical theory of communication*", Bell System Tech. J 27 (1948) 379-423, 623-659

[6] Lorenzo Fattorini. *Dispense per il corso di Statistica,* Università degli studi di Siena, 2001

[7] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons & H. E. Stanley. *Long-range correlations in nucleotide sequences,* Nature 356 (1992) 168-170

[8] Stephan Beirer. *Modelling long-range Correlations in Genomic DNA sequences.* Diploma Thesis, Technische Universität Berlin, Germany, 2003.

[9] I.Grosse, *Statistical Analysis of Biosequences*, Mastersthesis, Humboldt University of Berlin,1995.

[10] NCBI *FASTA format description* http://www.ncbi.nlm.nih.gov/blast/fasta.shtml

[11] *The* DDBJ/EMBL/GenBank *Feature Table Definition* http://www.ncbi.nlm.nih.gov/projects/collab/FT/index.html

[12] Francesca Diella (EMBL), Silvia Boi (Università degli studi di Milano). *Una proteina nella rete: Introduzione alla bioinformatica* http://www.ceebt.embo.org/projects/project18/project18it.html

[13] *Documentazione di Gnuplot*: http://www.Gnuplot.info/documentation.html

[14] ftp://ftp.ncbi.nih.gov/genomes/MapView/Homo_Sapiens/sequence/BUILD.36.2

[15] *Sito web di Entrez*: http://www.ncbi.nlm.nih.gov/sites/entrez

[16] *Sito web di Ensembl:* http://www.ensembl.org/index.html

[17] *Sito web di BioConductor:* http://www.bioconductor.org

[18] *DNA & RNA: i codoni:* http://www.ferrarisvr.it/dna/codons.htm

[19] *Sito web di Octave:* http://www.gnu.org/software/Octave/about.html

[20] *Sito web di Matlab:* http://www.mathworks.com/

[21] *Sito web di GenBank:* http://www.ncbi.nlm.nih.gov/Genbank/

[22] *Sito web di EMBL:* http://www.ebi.ac.uk/embl/

[23] *Sito web della Data Bank of Japan:* http://www.ddbj.nig.ac.jp/

[24] *File seq_ gene.md:*ftp://ftp.ncbi.gov/genomes/MapView/Homo_sapiens/sequence/BUILD.36.2/

Appendice A

HowTo bioUtils

Le bioUtils sono un insieme di applicazioni volte all'analisi statistica di sequenze di DNA. Questa guida contiene le conoscenze essenziali per l'utente. Si tenga comunque conto che il software, se invocato con l'opzione '?' o con i parametri sbagliati, stampa a schermo un messaggio con le modalità d'uso, che contiene informazioni simili a quelle che si possono trovare in questa guida.

1 - Descrizione

Le bioUtils si compongono delle seguenti applicazioni

- *nucfreq:* calcola le frequenze dei singoli nucleotidi all'interno di un file di tipo FASTA.
- *couplefreq*: calcola le frequenze delle coppie di singoli nucleotidi ad una distanza k all'interno di un file di tipo FASTA.
- *MutualInformation:* calcola la funzione di mutua informazione per i singoli nucleotidi.
- *AutoCorrelation:* calcola la funzione di autocorrelazione per i singoli nucleotidi.
- *codonfreq:* calcola la frequenza di codoni all'interno di un file di tipo FASTA.
- *codoncouplefreq:* calcola le frequenze delle coppie di codoni ad una distanza k all'interno di un file di tipo FASTA.

- *CodonMutualInformation:* calcola la funzione di mutua informazione per i codoni.
- *CodonAutoCorrelation:* calcola la funzione di autocorrelazione per i codoni.
- *GeneMasker:* dato un file di tipo FASTA restituisce un altro file nello stesso formato ma con particolari zone mascherate con il simbolo 'N'

2 - Compilazione e Installazione

Per compilare il software bisogna posizionarsi, tramite terminale, all'interno della cartella bioUtils e usare il comando:

>make all

Successivamente usare il comando:

>make install

Esiste la possibilità di compilare e installare solo una delle applicazioni tramite il comando

>make [nome-applicazione]

3 - Utilizzo

In questa sezione vengono specificati i comandi e i parametri per ognuna delle applicazioni.

Lista dei parametri più comuni:

[file-fasta] File di tipo FASTA che contiene la sequenza di DNA da analizzare. Generalmente i file FASTA hanno estensione .fa o .fasta. Ad ogni modo il software non esegue controlli sull'estensione, di conseguenza, affinché il file venga accettato come input, è sufficiente che rispetti il formato standard[9].

[directory-freq] Si intende una directory contenente i file che contengono valori di frequenza che si riferiscono ad una sequenza. Nel caso della frequenza dei singoli nucleotidi il file si chiama nucleotide.freq e nel caso dei codoni codon.freq. Nel caso delle coppie. invece, abbiamo per i singoli nucleotidi file il cui nome è dato dall'unione dei simboli presi in considerazione. Ad esempio nel caso dei singoli nucleotidi aa.freq è il file che contiene le frequenze della coppia Adenina-Adenina. Similmente per i codoni, il nome del file che contiene le frequenze della coppia di codoni AAA e TTT sarebbe AAATTT.freq. Nel caso dei file che contengono le frequenze di coppie le regole sul nome devono essere rispettate altrimenti i file non potranno essere utilizzati correttamente. Mentre, nel caso dei singoli, simboli il nome del file è solo una convenzione, può quindi essere modificato senza influire sul funzionamento dei programmi che li utilizzano come input.

[file-freq] singolo file delle frequenze.

[file-output] il nome del file di testo dove scrivere i risultati, se il file non esiste viene creato. ATTENZIONE se il file già esiste il suo contenuto verrà completamente eliminato e sostituito con il nuovo output.

[file-vector] Un file di tipo vector è un file che contiene un rigo per ogni codone ed un valore uno o zero associato. Tali valori indicano se quel codone deve essere o no preso in considerazione. Questo file viene passato come input alla funzione di autocorrelazione per i codoni.

Comandi:

nucreq:

>nucfreq [file-fasta]
restituisce il file nucleotide.freq.

couplefreq:

>*couplefreq* [file-fasta]

Questo comando restituisce come output 16 file (uno per ogni possibile coppia di nucleotidi), quindi si consiglia di invocarlo da dentro una cartella creata appositamente per salvare i dati sulle frequenze.

MutualInformation

>*MutualInformation* [directory-freq] [file-output]

[directory.freq] deve contenere sia i dati delle singole frequenze che delle frequenze di coppie.

AutoCorrelation

>*AutoCorrelation* [directory-freq] [1/0] [1/0] [1/0] [1/0] [file-output]

[directory.freq] deve contenere sia i dati delle singole frequenze che delle frequenze di coppie. I 4 valori 1/0 da inserire si riferiscono ai simboli da prendere in considerazione per la funzione di autocorrelazione, l'ordine è quello lessicografico quindi: Adeninia, Citosina, Guanina, Timina.

codonfreq

>*codonfreq* [file-fasta]

restituisce il file codon.freq.

codoncouplefreq

>*codoncouplefreq* [file-fasta]

Questo comando restituisce come output 4096 file (uno per ogni possibile coppia di codoni), quindi si consiglia di invocarlo da dentro una cartella creata appositamente per salvare i dati sulle frequenze.

CodonMutualInformation

>*CodonMutualInformation* [directory-freq] [file-freq] [file-output]

Nella directory su devono trovare i file delle coppie di frequenze mentre il singolo file si riferisce alle frequenze dei singoli codoni. ATTENZIONE! quest'ultimo file non deve trovarsi nella stessa cartella che contiene i file della frequenza di coppie.

CodonAutoCorrelation

>*CodonAutoCorrelation* [directory-freq] [file-freq] [file-output] [file-vector]

Nella directory si devono trovare i file delle coppie di frequenze, mentre il singolo file si riferisce alle frequenze dei singoli codoni. ATTENZIONE! quest'ultimo file non deve trovarsi nella stessa cartella che contiene i file delle frequenze delle coppie.

GeneMasker

>*GeneMasker* [opzione] [map] [file-fasta] [numero-cromosoma] [type]

Questo programma accetta 4 possibili opzioni

- -c: maschera le zone di tipo [type]
- -n: maschera le zone non di tipo [type]
- -s: Estrae e salva in file separati le sequenze di tipo [type]
- -l: Fornisce una lista dei possibili valori da inserire come [type]

I possibili valori di type sono : CDS, GENE, RNA.

[map] questo è un file contenente le informazioni riguardanti alcune particolari zone sui cromosomi del genoma umano. Il file è fornito da NCBI ed è liberamente scaricabile da internet [22].

4 - Disinstallazione

Per disinstallare il programma bisogna posizionarsi all'interno della cartella bioUtils e usare il comando:

>make clean

5 - Problemi/Bug noti

Il software è stato sviluppato e testato sui seguenti sistemi operativi:

- Linux Fedora 7 (kernel 2.6.12)
- Linux Fedora 8 (kernel 2.6.21)
- Linux Debian 4.0 (kernel 2.6.8)

Il funzionamento su altre piattaforme di tipo Unixdovrebbe essere garantito dall'implementazione del codice secondo librerie standard C++ ma non è stata testata praticamente. Per la compilazione del codice è stato utilizzato il GNU C Compiler (GCC) versione 4.2.1.

Appendice B

Metodo dei minimi quadrati ordinari

Prima di spiegare come opera il metodo dei minimi quadrati ordinari (in inglese *Ordinary least squares* OLS) si deve definire il modello di regressione lineare. Si considerino due variabili quantitative X ed Y, tra cui esiste una relazione di dipendeza del seguente tipo:

$$Y = f(X)$$

Dove Y sarà la variabile dipendente e X la variabile indipendente (il modello è estendibile anche ad un vettore di $n \times 1$ variabili indipendenti).

Si definisce il modello di regressione lineare come:

$$Y = \alpha + \beta\, X + \epsilon$$

Dove la X è considerata una delle cause rilevanti della Y, mentre tutte le altre variabili che non sono esplicitamente considerate nella relazione vengono inglobate nella variabile ϵ che è detta componente erratica. I parametri α e β vengono stimati sulla base delle osservazioni con il metodo dei minimi quadrati ordinari.

Si considerino le n coppie di osservazioni $(x_1,\ y_1)$, $(x_2,\ y_2)$, ...,$(x_n,\ y_n)$. In base al modello di regressione lineare il valore assunto dalla variabile Y in corrispondenza dell'i- esimo elemento risulta:

$$y_i = \alpha + \beta\, x_i + \epsilon_i,\ i = 1,\ 2,..., n$$

dove $\alpha + \beta x_i$è la previsione di Y sulla base del valore assunto da X mentre, ϵ_iè la differenza fra il valore previsto e il valore osservato, ovvero l'effetto di tutti gli altri fattori non considerati ma che hanno influenza su Y. Siano $\widehat{\alpha}$ e $\widehat{\beta}$due qualunque valutazioni dei parametri incogniti αe β. In questo caso il valore di Y previsto sulla base del valore $x_i$assunto da X risulta:

$$\widehat{y_i} = \widehat{\alpha} + \widehat{\beta}\, x_i,\ i = 1,\ 2, ...,\ n$$

da cui l'errore di previsione, anche detto scarto o residuo, è dato da:

$$e_i = y_i - \widehat{y_i},\ i = 1,\ 2, ...,\ n$$

Il metodo dei minimi quadrati ordinari suggerisce di valutare αe β minimizzando la somma del quadrato degli n residui. La funzione obiettivo è:

$$f(\alpha,\ \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Al fine di minimizzare la $f(\alpha,\ \beta)$si considerino le derivate parziali della stessa rispetto ad αe βe si eguagliano a zero. La soluzione del sistema dà luogo a:

$$\widehat{\beta} = \frac{s_{xy}}{s_X^2}$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\, \overline{x}$$

Dove la prima permette di scrivere $\widehat{\beta}$ come rapporto tra la covarianza di X ed Y e la varianza di X ed evidenzia come, a seconda che la relazione tra X ed Y sia di tipo diretto o inverso, la retta di regressione ottenuta con il metodo dei minimi quadrati risulti rispettivamente crescente o decrescente. Mentre, la seconda equazione implica che

$$\overline{y} = \widehat{\alpha} + \widehat{\beta}\, \overline{x}$$

E quindi la retta di regressione risulta passante per il punto di coordinate $(\overline{x},\ \overline{y})$.

Ringraziamenti

Terrorizzato dall'idea di poter dire delle banalità mi ero ripromesso, solennemente, che avrei evitato di scrivere i ringraziamenti. Di conseguenza ecco, di seguito, una marea di persone che sento il bisogno di ringraziare.

Innanzitutto ringrazio il prof. Ercan Kuruoglu che mi dato la possibilità di svolgere questo lavoro interdisciplinare e che mi ha seguito con constanza e attenzione durante lo svolgimento dello stesso. Naturalmente ringrazio anche il prof. Osman Abul, per la sua disponibilità a darmi chiarimenti e per avere compreso le mie domande nonostante il mio pessimo inglese.

E veniamo ora agli amici e agli affetti.

Un ringraziamento ad Andrea che ormai si era rassegnato all'idea che ogni ora e mezza venissi a bussargli alla porta della stanza per raccontargli qualche fesseria, per chiedergli un consiglio sul tirocino o, quando gli andava male, entrambe le cose.

Un doveroso ringraziamento ad Antonio e Giorgio che tenendomi il computer occupato giocando a Pes6 mi hanno privato di una possibile fonte di distrazione dal mio lavoro.

Grazie anche a Matteo che è un buon amico e si è sempre mostrato disponibile per dare una mano quando ne ho avuto bisogno.

Un grazie a Roberto, il "maestro", l'unico che conosco che sappia apprezzare Woody Allen!

Grazie a (_) (¬ () (Ugo), che pazientemente ha ascoltato le mie paranoie prima dell'ultimo esame.

Grazie a Michele e Francesco che con le loro "pacate" e divertenti discussioni rendono i pasti a mensa più piacevoli, e c'è ne molto bisogno, sopratutto quando servono il manzo all'inglese...

Ringrazio Pietro e Sabine per il continuo supporto e per gli indispensabili rifornimenti di focaccine e cioccolata.

Ringrazio, in modo particolare, i miei genitori che mi hanno sempre dato la possibilità di seguire la mia strada liberamente. Mi hanno aiutato con il loro affetto e con i loro consigli, dandomi gli strumenti per diventare, nei limiti del possibile, una persona serena e soddisfatta della propria vita.

Ringrazio Giuliana, che mi sta sempre vicino e mi ha aiutato costantemente durante il tirocino e durante la stesura di questa relazione. Se questa tesi non è piena di "orrori" grammaticali e anche grazie alla sua paziente rilettura, tutte le imprecisioni rimaste sono solo mia responsabilità. Ma la ringrazio sopratutto perché mi vuole bene e me lo dimostra ogni giorno, e perché se lei non avesse creduto in me, quando neppure io credevo in me stesso, forse non sarei arrivato a scrivere questi ringraziamenti.