Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Studio delle interdistanze dei dinucleotidi CG e TA nei cromosomi umani

Relatore:
Prof. Daniel Remondini

Correlatore:
Dott.ssa Alessandra Merlotti

Presentata da:
Jacopo D'Alberto

Anno Accademico 2018/2019

Alla mia famiglia

Sommario

Il lavoro svolto in questa tesi si colloca nell'ambito delle analisi statistiche delle sequenze di DNA. In particolare sono state studiate le distribuzioni delle interdistanze dei dinucleotidi CG e TA all'interno del genoma umano. La prima parte dello studio si é occupata di effettuare una valutazione della qualitá dei dati utilizzati per le analisi. Questo perché, all'interno dei vari cromosomi che compongono il DNA, sono presenti alcuni nucleotidi che non sono stati correttamente sequenziati e le cui basi azotate sono per il momento sconosciute. É emerso che tali nucleotidi, indicati con la lettera N, sono presenti in grandi blocchi e localizzati alle estremitá ed al centro dei vari cromosomi. Si é osservato che, nonostante le numerose occorrenze, questi nucleotidi possono essere trascurati e che l'effetto della loro rimozione dalle sequenze di DNA é statisticamente irrilevante. La seconda parte delle analisi é riservata allo studio delle distribuzioni delle interdistanze del dinucleotide CG. Quello che si é visto é che tali distribuzioni sono ben descritte da una legge di potenza traslata con cutoff esponenziale. I parametri, ottenuti dal fit, assumono valori simili ad eccezione dei cromosomi 16, 17, 19, 20, 22 e Y. Particolare attenzione é stata riposta sui valori anomali del parametro del cutoff esponenziale. Nel tentativo di spiegare l'andamento di tali valori, é stata formulata un'ipotesi basata su di un possibile legame tra il parametro del cutoff esponenziale e la densitá del dinucleotide CG all'interno dei vari cromosomi. L'ultima sezione é dedicata invece allo studio delle distribuzioni delle interdistanze del dinucleotide TA. Si é visto nel cromosoma 1 che questo dinucleotide é quello che ha un andamento piú simile al dinucleotide CG. Tali distribuzioni sono state poi fittate con la legge di potenza traslata con cutoff esponenziale e con una legge di potenza traslata senza cutoff ed é emerso che sono meglio descritte dalla seconda.

Indice

Introduzione

A partire dagli anni duemila nello studio del DNA, sono state proposte diverse metodologie di analisi volte ad individuare ed a comprendere eventuali strutture di correlazione nelle sequenze genomiche degli esseri viventi (vedere articoli [1], [2], [3]). Una svolta si ha con la pubblicazione dell'articolo [4] dal quale emerge un nuovo metodo per processare le sequenze di DNA basato sulla interdistanza tra nucleotidi, ovvero la distanza rispetto alla successiva occorrenza dello stesso nucleotide. Questa nuova metodologia ha cosí permesso di ottenere un'intera caratterizzazione di genomi completi di esseri viventi appartenenti a specie diverse. Successivamente, come emerge dall'articolo [5], si é passati allo studio delle interdistanze dei dinucleotidi. Ció é motivato dal fatto che i dinucleotidi hanno un ruolo rilevante nella biologia del genoma, quindi uno studio delle distribuzioni delle loro interdistanze puó essere la chiave per comprendere meglio il DNA. Infatti, grazie a questo potente strumento é stato possibile identificare le isole CpG [6] e caratterizzare la potenziale suscettibilitá dei diversi cromosomi alle modificazioni epigenetiche [7]. I lavori [8], [9], [10] e [11] hanno mostrato, attraverso uno studio condotto sulle sequenze piú lunghe di DNA di diversi esseri viventi, che esistono delle caratteristiche peculiari nelle sequenze genomiche dei mammiferi: le code delle distribuzioni delle interdistanze dei dinucleotidi CG mostrano un decadimento esponenziale a differenza delle code delle distribuzioni delle interdistanze di tutti i dinucleotidi diversi da CG che mostrano un decadimento simile ad una legge di potenza. Ció é stato motivato sottolineando il ruolo specifico assunto dalle CG all'interno del genoma dei mammiferi, dal momento che rappresentano i siti preferenziali della metilazione (meccanismo epigenetico fondamentale coinvolto nella regolazione dei geni [12], [13]). Altri risultati importanti che emergono da questi articoli riguardano la distribuzione delle interdistanze dei dinucleotidi CG: é stato infatti verificato che la densitá di probabilitá che meglio fitta questa distribuzione é la distribuzione Gamma, il cui parametro di scala puó essere associato alla complessitá della categoria dell'organismo in questione.

Partendo da questi esiti e limitando l'attenzione al genoma umano, si vuole analizzare nel dettaglio l'andamento di queste distribuzioni, cromosoma per cromosoma ed estendere i risultati ottenuti per il primo cromosoma umano a tutta la restante parte del DNA. In particolare, si vuole analizzare le distribuzioni delle interdistanze dei dinucleotidi CG e TA. Nel caso del primo dinucleotide, a differenza dei precedenti lavori in

cui le distribuzioni sono state fittate con una funzione gamma, si propone di utilizzare come modello una legge di potenza traslata con cutoff esponenziale in quanto quest'ultima descrive meglio l'andamento delle distribuzioni alle piccole interdistanze. Per quanto riguarda invece il dinucleotide TA si vuole mostrare il legame con il dinucleotide CG ed individuare quale tra una legge di potenza traslata con cutoff esponenziale e una legge di potenza traslata senza cutoff descriva meglio il comportamento delle distribuzioni.

Inoltre si vuole riporre particolare attenzione allo studio delle parti non sequenziate dei cromosomi,argomento ben approfondito in apposite sezioni.

Capitolo 1

Metodi statistici per l'analisi dei dati

In questo capitolo verranno illustrate nel dettaglio tutte le analisi che sono state svolte sulle sequenze di DNA.
In particolare, in una prima sezione verranno descritti il tipo di dati utilizzati e come da essi sia possibile estrarre la distribuzione delle interdistanze di un determinato dinucleotide. Nelle successive tre sezioni verranno invece rispettivamente trattati tre diversi aspetti dell'analisi, mentre l'ultima sezione sará dedicata alla bontá dei fit utilizzati.

1.1 Dati

1.1.1 FASTA file

I 22 autosomi (cromosomi che non partecipano alla determinazione del sesso) piú i 2 cromosomi sessuali sono stati scaricati dal database della NCBI sotto formato di file FASTA. In particolare, la versione che é stata utilizzata per le analisi é la release 12 aggiornata al 31/03/2018.
Questo tipo di formato é molto importante nell'ambito della biochimica e della bioinformatica, in quanto é prettamente utilizzato per la rappresentazione di sequenze di nucleotidi o di amminoacidi, nelle quali questi ultimi sono rispettivamente indicati attraverso singole lettere. In genere, un file di tipo FASTA contiene una riga d'intestazione in cui viene specificato il nome dell'organismo, dal quale la sequenza é stata estratta, ed il cromosoma in questione.

Una sequenza di nucleotidi é caratterizzata dalla ripetizione di 5 lettere: A, C, G, T, N. Le prime quattro si riferiscono rispettivamente alle quattro basi azotate presenti nel DNA, ossia Adenina, Timina, Citosina e Guanina; mentre l'ultima lettera si riferisce

alle parole inglesi "aNy base" (in italiano "qualunque base") ed indica un nucleotide sconosciuto che non é stato correttamente classificato.

La lunghezza fisica delle sequenze di acidi nucleici a doppio filamento (e.g. DNA) é misurabile in "coppie di basi" o "paia di basi"(abbreviate come pb o, dall'inglese "base pair",bp o bps). Maggiore é la lunghezza della sequenza di nucleotidi, maggiore é il numero di lettere utilizzate nel file FASTA; fatto che comporta un aumento delle dimensioni del file in questione. Ogni cromosoma ha una diversa lunghezza, come si puó facilmente notare osservando le differenti dimensioni dei vari file FASTA (vedere Tab. 1.1).

Cromosoma	**Dimensione (MB)**	**Cromosoma**	**Dimensione (MB)**
1	252,5	13	116,0
2	245,7	14	108,6
3	201,1	15	103,4
4	192,9	16	91,6
5	184,1	17	84,4
6	173,2	18	81,5
7	161,6	19	59,5
8	147,2	20	65,4
9	140,4	21	47,4
10	135,7	22	51,5
11	137,0	X	158,3
12	135,2	Y	58,0

Tabella 1.1: *Dimensione dei file FASTA corrispondenti ai diversi cromosomi umani.*

1.1.2 Distribuzione delle interdistanze

Come detto precedentemente la sequenza di nucleotidi é formata dalla ripetizione di 5 lettere: A, C, G, T, N. Si suppone, per il momento, di poter ignorare tutte le "N" (fatto motivato esaustivamente nelle sezioni 1.2 e 2.1). Per ottenere la distribuzione delle interdistanze (distribuzione di probabilitá) $p(\tau)$ di un particolare dinucleotide si deve seguire la seguente procedura: si rimuovono dalla sequenza tutti i nucleotidi sconosciuti, si trova la posizione dei dinucleotidi lungo la sequenza e si calcola la distanza tra due occorrenze consecutive dello stesso dinucleotide, ottenendo un vettore di valori di interdistanze τ, contate in termini di numero di basi. Successivamente si calcola la frequenza di ciascun valore di interdistanza e dividendola poi per la somma delle frequenze si ottiene in questo

modo $p(\tau)$ (vedere eq. 1.1).

$$p(\tau) = \frac{\#\{j = 1, 2, ...|\tau_j = \tau\}}{\#\{j = 1, 2, ...|\tau_j\}}\,, \tag{1.1}$$

Le distribuzioni delle interdistanze sono state poi plottate in scala semi-logaritmica rispetto all'asse delle Y (vedere Fig. 1.1 e 1.3). Tale procedura é stata implementata in MATLAB grazie all'utilizzo del pacchetto *Bioinformatics*.

1.2 Analisi delle N

Nella sezione precedente, al momento del calcolo delle interdistanze, sono state ignorate le basi sconosciute indicate dalla lettera N. Dal momento che la percentuale di queste basi, non correttamente classificate, varia da cromosoma a cromosoma, andando a costituire in alcuni casi valori non irrilevanti (vedere Tab. 1.2), é necessario e di fondamentale importanza accertarsi che la rimozione delle N dalle sequenze analizzate abbia effetti trascurabili.

Prima di tutto si osserva che nel calcolo delle interdistanze tra i dinucleotidi, qualora si incontri una N, sono possibili tre diverse azioni:

1. il conteggio della distanza continua come se si fosse incontrato uno qualsiasi dei nucleotidi noti (A,C,G,T),

2. il conteggio della distanza non viene aggiornato ignorando la N incontrata: ció equivale a rimuovere la N dalla sequenza,

3. il conteggio della distanza viene calcolato come descritto dal punto 1 tuttavia, alla fine, la distanza viene scartata dalla distribuzione.

Una soluzione al problema di quale sia la migliore azione da eseguire é data dallo studio [8] secondo il quale é possibile rimuovere le N dalla sequenza analizzata, in quanto l'effetto dovuto allo scarto di queste lettere é trascurabile. In particolare queste lettere sono generalmente presenti in blocchi influenzando cosí solo pochi conteggi di distanze. Inoltre, nei cromosomi con un alto contenuto di N, queste lettere sono maggiormente distribuite nella parte iniziale e nella parte finale del cromosoma [8]. Questo approccio é stato anche adottato nei lavori [10] e [11].

Partendo da questi risultati, si é voluto prima di tutto caratterizzare in modo dettagliato la posizione delle basi sconosciute nelle sequenze di nucleotidi dei vari cromosomi umani. Per fare ció é stato necessario implementare un programma in MATLAB in grado di leggere le sequenze di nucleotidi, contenute nei file FASTA, e di individuare tutte le posizioni delle N, poi memorizzate in un vettore. In seguito, partendo da questi vettori, sono stati creati degli istogrammi in modo da permetterne una facile visualizzazione (vedere Fig. 2.1, 2.2 e 2.3).

Un'ulteriore analisi, sulla distribuzione delle N lungo le sequenze di DNA, é stata compiuta al fine di verificare ed eventualmente migliorare i risultati ottenuti nel lavoro [8]. In particolare, quello che si é voluto verificare é se,all'interno di ogni cromosoma, le N siano raggruppate in grandi gruppi oppure siano presenti lungo la sequenza sotto forma di tanti piccoli blocchi. Questo fatto é di fontamentale importanza in quanto maggiore é il numero di blocchi maggiore é il numero di interdistanze tra dinucleotidi influenzate dalle N. Per fare questo é stato necessario individuare tutte le sequenze costituite da sole N ripetute consecutivamente e calcolarne la lunghezza in termini di numero di basi. Sono stati in questo modo costruiti tanti vettori, contenenti le lunghezze di queste sequenze, quanti il numero dei cromosomi analizzati e sono stati successivamente utilizzati per creare altrettanti istogrammi (vedere Fig. 2.4, 2.5, 2.6 e 2.7 e Tab. 2.1 e 2.2).

Cromosoma	**Percentuale N (%)**	**Cromosoma**	**Percentuale N (%)**
1	7,4	13	14
2	0,68	14	15
3	0,099	15	17
4	0,24	16	9,4
5	0,15	17	0,41
6	0,43	18	0,35
7	0,24	19	0,30
8	0,26	20	0,78
9	12	21	14
10	0,40	22	23
11	0,41	X	0,74
12	0,10	Y	54

Tabella 1.2: *Percentuali di basi sconosciute (N) nei vari cromosomi umani.*

1.3 Analisi del dinucleotide CG

Seguendo il procedimento descritto nella sottosezione 1.1.2 é quindi possibile ricavare, cromosoma per cromosoma, le distribuzioni delle interdistanze del dinucleotide CG. Nelle pagine seguenti sono riportati i grafici rappresentanti le distribuzioni delle interdistanze del nucleotide CG per tutti i 24 cromosomi umani (Fig. 1.1 e 1.2).

Guardando le Fig. 1.1, 1.2 si puó vedere come l'andamento delle distribuzioni risulta simile per tutti i cromosomi. Si puó inoltre notare che all'aumentare dei valori delle interdistanze si ha un incremento delle fluttuazioni, fatto che puó influenzare negativamente i risultati del fit.

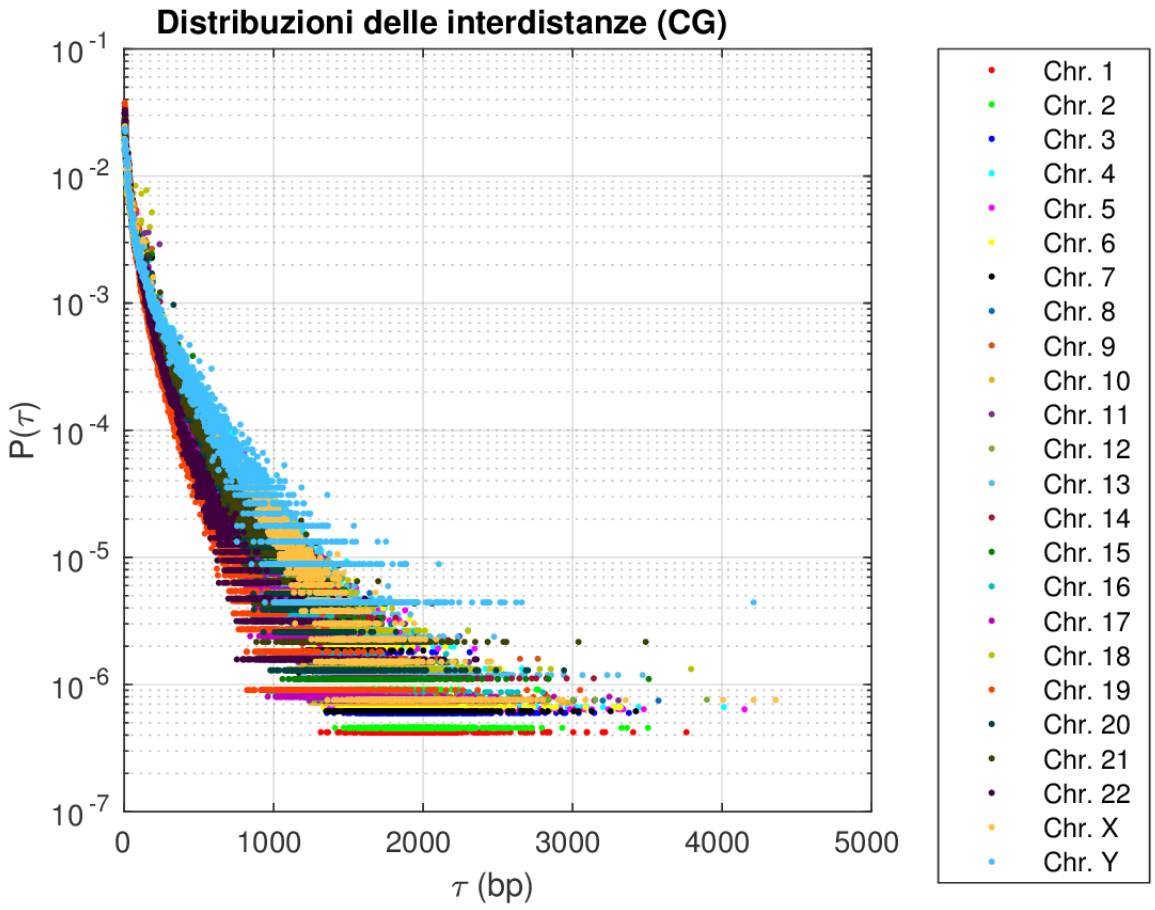

Figura 1.1: *Distribuzioni delle interdistanze del dinucleotide CG, in scala semilogaritmica rispetto all'asse delle Y, di tutti i 24 cromosomi umani.*

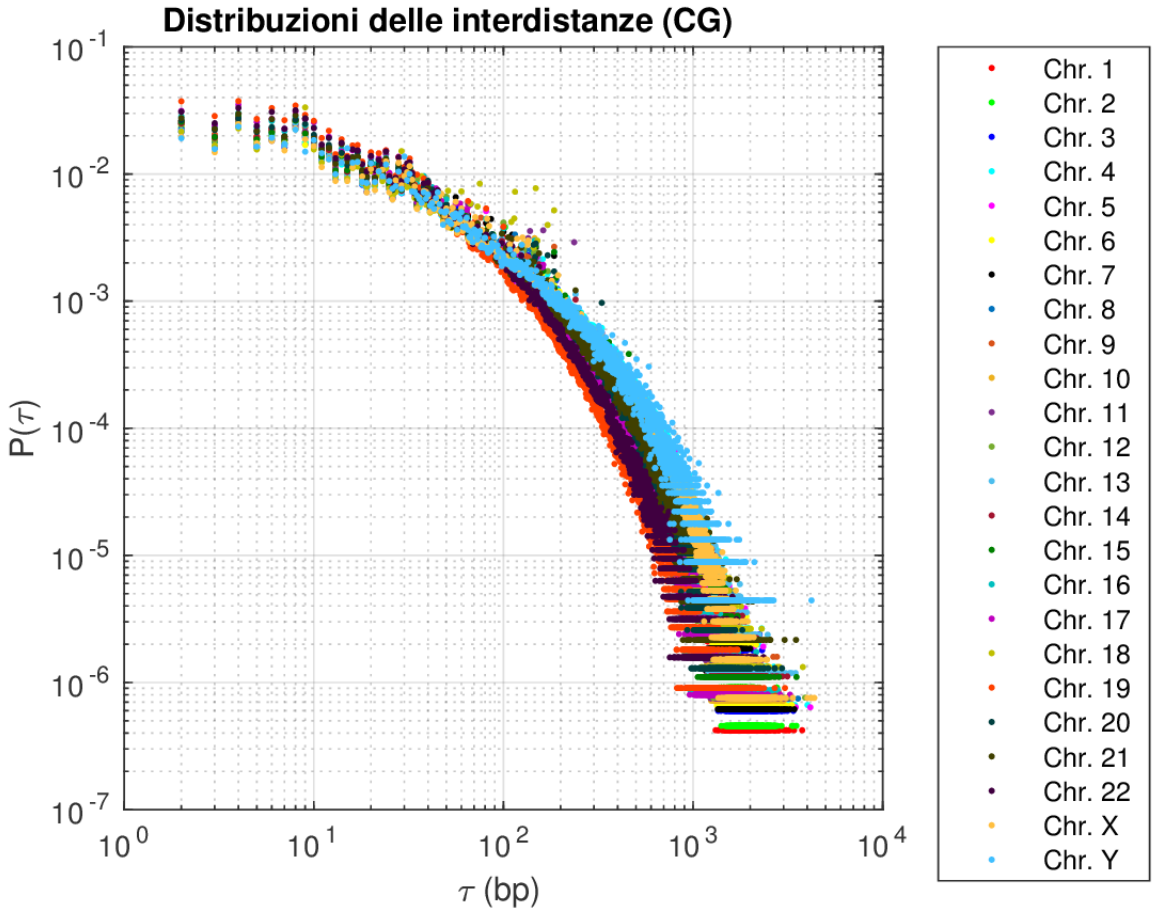

Figura 1.2: *Distribuzioni delle interdistanze del dinucleotide CG, in scala doppio logaritmica, di tutti i 24 cromosomi umani.*

Si é deciso quindi di rimuovere le parti piú rumorose delle code dall'analisi, scegliendo come valore di cutoff il 90° percentile (vedere Tab. 1.3). Questo taglio ha inoltre il vantaggio di permettere l'analisi della stessa quantitá di dati per tutti i cromosomi.

Cromosoma	**90° percentile (bp)**	**Cromosoma**	**90° percentile (bp)**
1	1864	13	1728
2	1899	14	1667
3	1887	15	1521
4	1958	16	1406
5	1866	17	1346
6	1832	18	1579
7	1811	19	1285
8	1780	20	1426
9	1715	21	1420
10	1664	22	1192
11	1745	X	1862
12	1771	Y	1513

Tabella 1.3: *Valori di cutoff per le distribuzioni di CG corrispondenti al 90° percentile.*

Si sceglie quindi di utilizzare come rappresentazioni di dati su cui effettuare il fit le distribuzioni delle interdistanze in scala semilogaritmica rispetto all'asse Y (Fig. 1.1) a cui sono state rimosse le code secondo i valori di cutoff riportati precedentemente (Tab. 1.3). Per quanto riguarda il modello di funzione da utilizzare per il fit, si é optato per una legge di potenza traslata di un parametro d con un cutoff esponenziale $e^{-\frac{x}{b}}$:

$$p(x) = (c(x+d)^{-a}e^{-\frac{x}{b}}). \tag{1.2}$$

Nel lavoro [10] si é visto che é possibile ottenere un fit migliore se anziché fittare le distribuzioni delle interdistanze con l' eq. 1.2 si fittano le stesse distribuzioni in scala semilogaritmica, rispetto all'asse delle Y, con il logaritmo della funzione 1.2, cioé con l'equazione:

$$p(x) = \log(c(x+d)^{-a}e^{-\frac{x}{b}}). \tag{1.3}$$

Il fit é stato realizzato utilizzando la funzione *fit* di MATLAB, opportunamente implementata, la quale sfrutta il metodo non lineare dei minimi quadrati tramite un algoritmo

di trust-region (letteralmente regione di confidenza). In particolare, dopo aver fissato i limiti superiori ed inferiori dei coefficienti da trovare grazie al fit, sono state effettuate diverse prove, facendo variare i valori di partenza dei coefficienti stessi, ottenendo alla fine sempre lo stesso risultato. Inoltre,utilizzando la funzione di MATLAB *confint*, sono stati calcolati gli errori sui parametri del fit, stimati ad un intervallo di confidenza del 95%. I risultati sono mostrati nella sezione 2.2.

1.4 Analisi del dinucleotide TA

Come giá accennato nell'introduzione, l'analisi del dinucleotide TA puó essere divisa in due sezioni separate. La prima sezione si occupa di studiare le distribuzioni delle inter-distanze di tutti e 16 i dinucleotidi nel primo cromosoma umano, al fine di evidenziare un'eventuale correlazione tra i dinucleotidi CG e TA. Per fare questo é necessario ottenere, seguendo il procedimento descritto nella sottosezione 1.1.2, i grafici delle distribuzioni di probabilitá di tutti e 16 i dinucleotidi (vedere Fig. 1.3).
Anche in questo caso, si é deciso di rimuovere le parti piú rumorose delle code dall'analisi utilizzando lo stesso criterio scelto in precedenza. I valori di cutoff, corrispondenti al 90° percentile, sono raccolti in Tab. 1.4.

Dinucleotide	**90° percentile (bp)**	**Dinucleotide**	**90° percentile (bp)**
AA	866	GA	719
AC	558	GC	715
AG	689	GG	873
AT	620	GT	575
CA	477	TA	705
CC	866	TC	755
CG	1864	TG	531
CT	745	TT	918

Tabella 1.4: *Valori di cutoff per le distribuzioni di tutti i 16 dinucleotidi del primo cromosoma umano corrispondenti al 90° percentile.*

Analogamente al caso del dinucleotide CG, si sceglie come funzione da utilizzare per il fit l'eq. 1.3 e si esegue il fit sui grafici in scala semilogaritmica a cui sono state tolte le code piú rumorose attraverso un programma scritto in MATLAB, in cui sono presenti le stesse funzioni citate precedentemente. I risultati sono mostrati nella sezione 2.3.1.

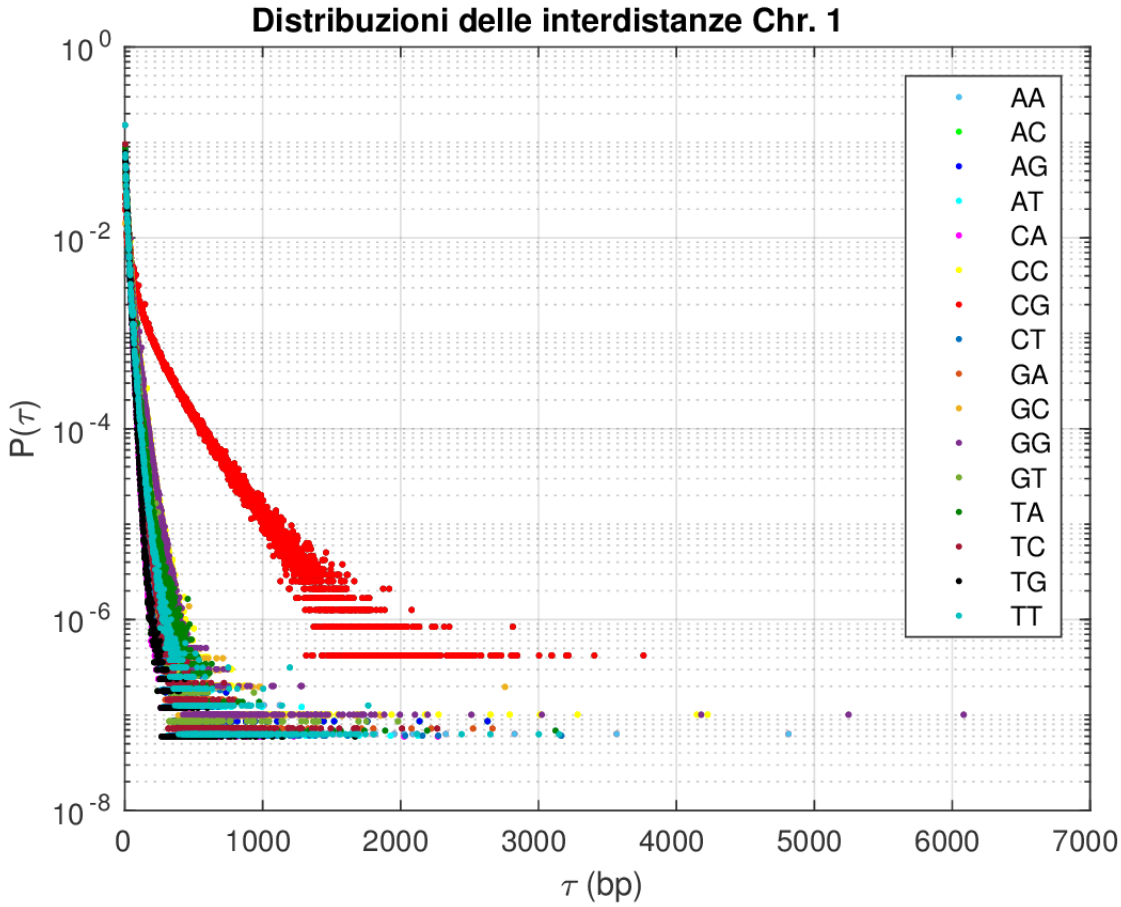

Figura 1.3: *Distribuzioni delle interdistanze di tutti i 16 dinucleotidi, in scala semilogaritmica rispetto all'asse delle Y, del primo cromosoma umano.*

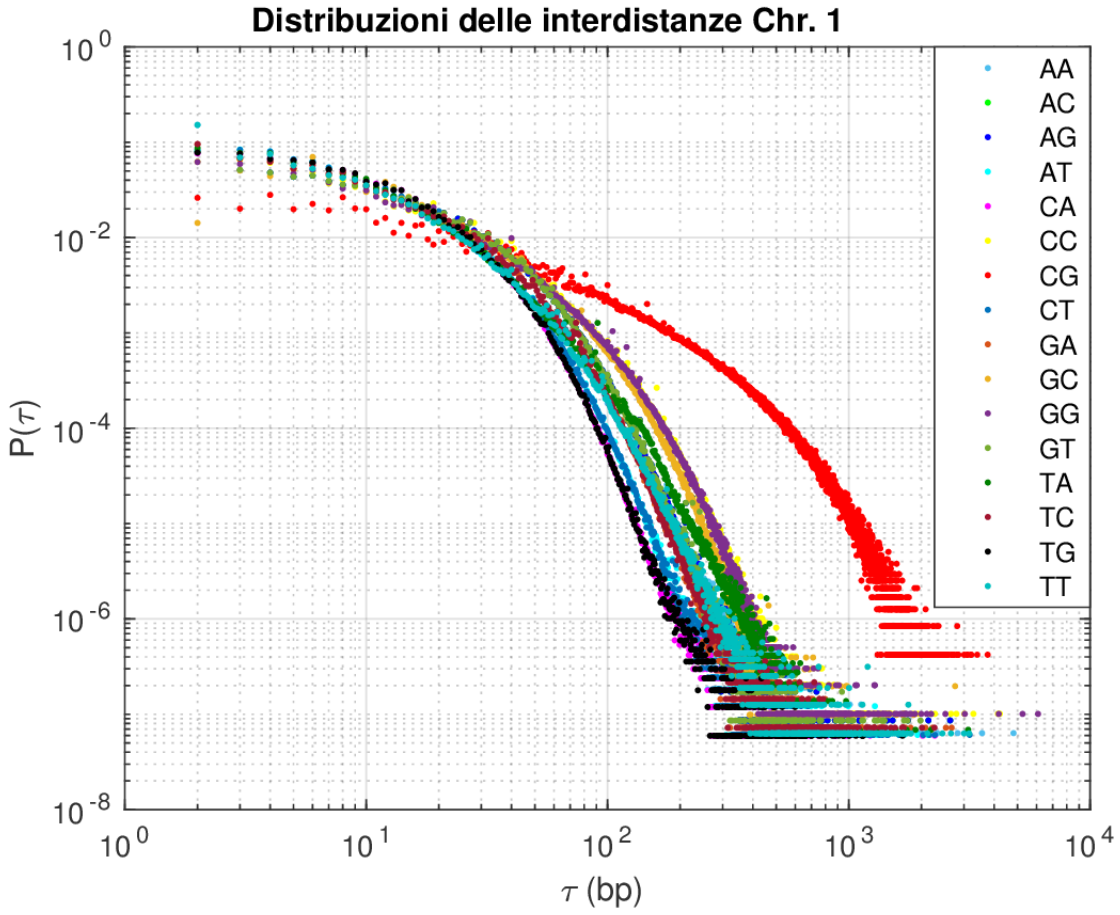

Figura 1.4: *Distribuzioni delle interdistanze di tutti i 16 dinucleotidi, in scala doppio logaritmica, del primo cromosoma umano.*

La seconda sezione si concentra sullo studio delle distribuzioni delle interdistanze del dinucleotide TA, al fine di individuare quale funzione descriva meglio i dati sperimentali. Per fare ció si procede in maniera analoga a quella descritta nella sezione 1.3: si individuano le distribuzioni delle interdistanze del dinucleotide TA per tutti i cromosomi umani e si rappresentano in scala semilogaritmica (vedere Fig. 1.5). Si rimuovono le parti piú rumorose delle code dall'analisi, scegliendo come valore di cutoff il 90° percentile (vedere Tab. 1.5).
Per quanto riguarda invece le funzioni da utilizzare nel fit, oltre alla giá citata eq. 1.3 si é deciso di utilizzare una legge di potenza traslata di un parametro d:

$$p(x) = (c(x+d)^{-a}) \tag{1.4}$$

Anche in questo caso,si é deciso di fittare le distribuzioni in scala semilogaritmica rispetto all'asse delle Y utilizzando il logaritmo della funzione 1.4, cioé attraverso l'equazione:

$$p(x) = (c(x+d)^{-a}) \tag{1.5}$$

Per realizzare il tutto sono state utilizzate le funzioni di MATLAB *fit* e *confint*, opportunamente implementate. I risultati sono mostrati nella sezione 2.3.2.

Cromosoma	**90° percentile (bp)**	**Cromosoma**	**90° percentile (bp)**
1	705	13	586
2	695	14	612
3	555	15	560
4	707	16	716
5	662	17	713
6	630	18	605
7	783	19	719
8	715	20	624
9	686	21	631
10	705	22	592
11	660	X	644
12	687	Y	579

Tabella 1.5: *Valori di cutoff per le distribuzioni di TA corrispondenti al 90° percentile.*

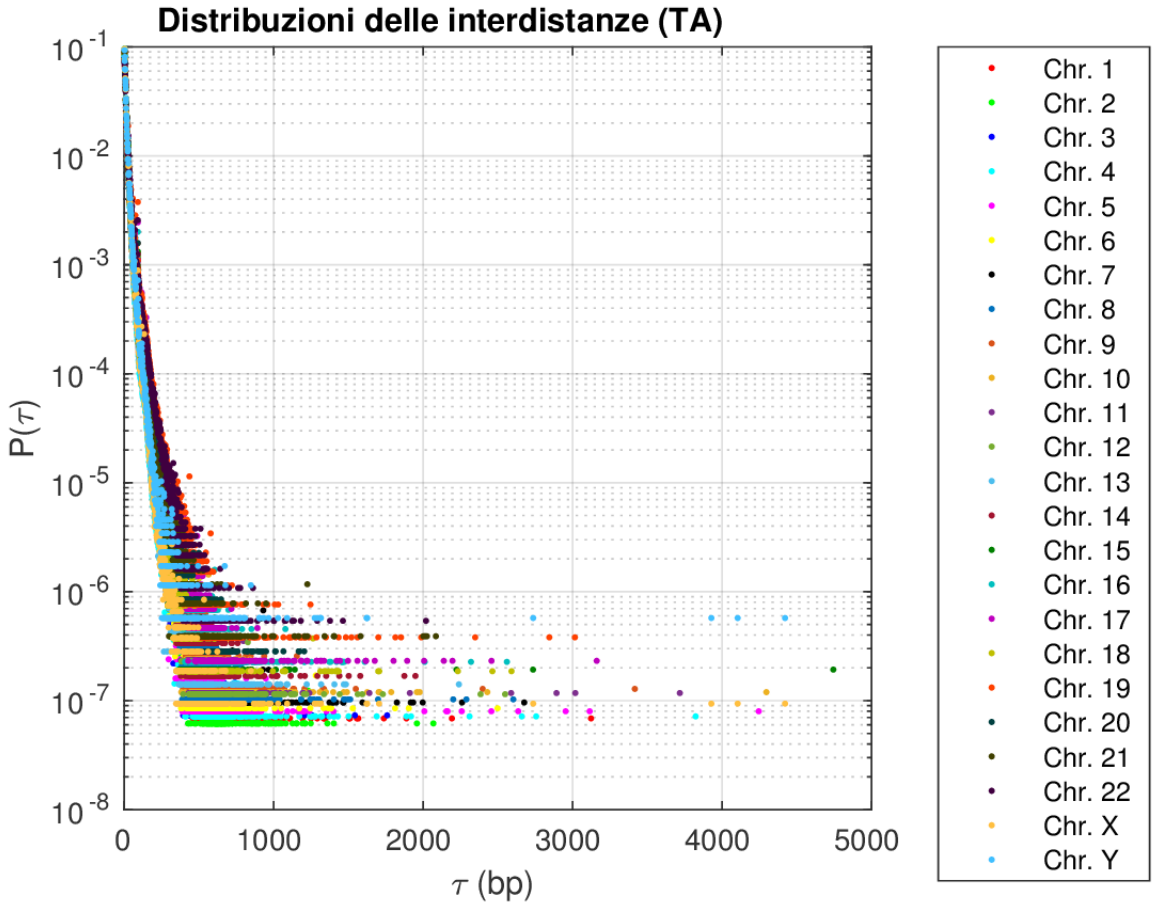

Figura 1.5: *Distribuzioni delle interdistanze del dinucleotide TA, in scala semilogaritmica rispetto all'asse delle Y, di tutti i 24 cromosomi umani.*

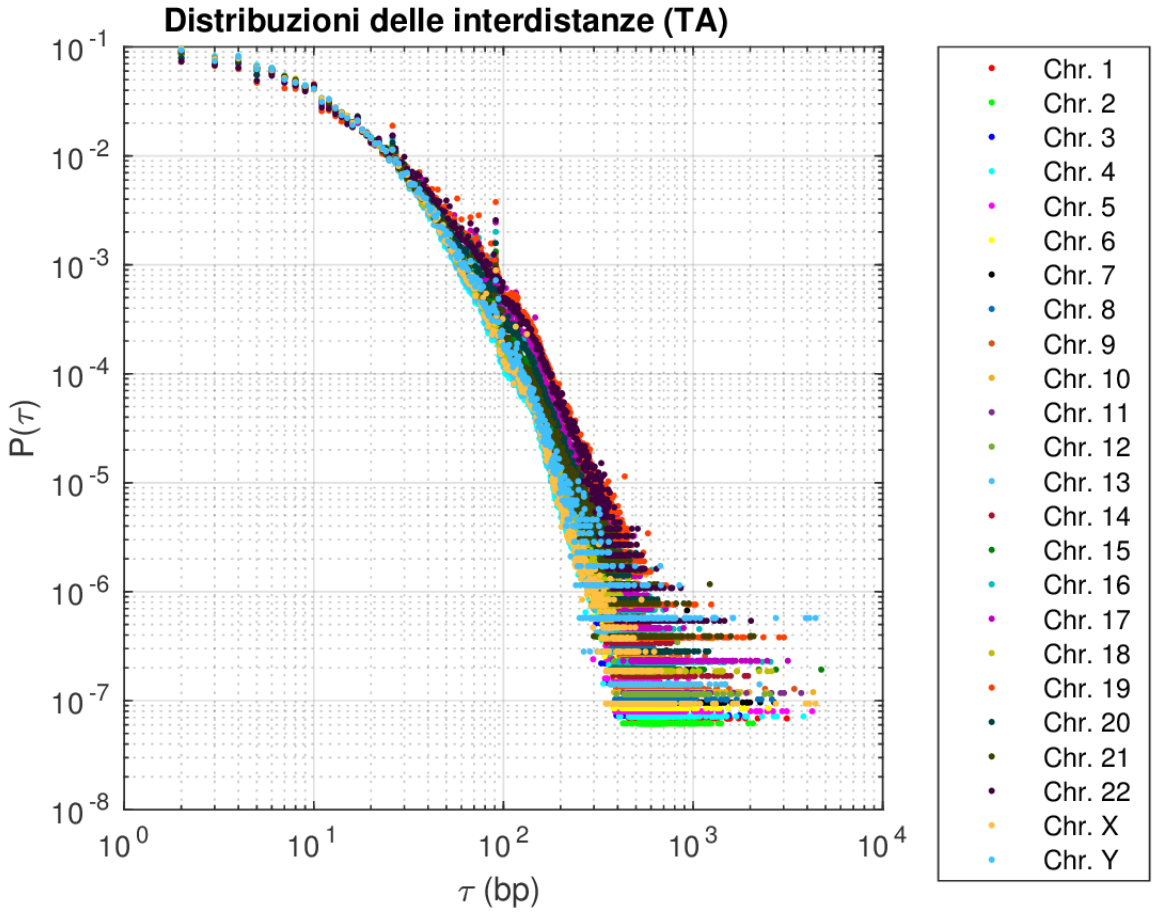

Figura 1.6: *Distribuzioni delle interdistanze del dinucleotide TA, in scala doppio logaritmica, di tutti i 24 cromosomi umani.*

1.5 Bontá dei fit

Per valutare la bontá dei fit effettuati si é tenuto conto di diversi fattori:

- parametri statistici quali SSE, R-quadro aggiustato e RMSE;
- analisi dei residui;
- intervalli di confidenza.

Per quanto riguarda i parametri statistici,la bontá del fit aumenta quando i valori del SSE e del RMSE tendono a 0 ed il valore del R-quadro aggiustato tende a 1.
L'analisi dei residui é uno studio qualitativo che si compie osservando i grafici dei residui ottenuti grazie a MATLAB. Assumendo che il modello scelto per il fit sia corretto, i residui approssimano gli errori casuali. Quindi se i residui si dispongono in modo casuale, ció significa che il modello fitta i dati correttamente.
Per quanto riguarda invece l'ultimo punto, maggiore é l'ampiezza dell'intervallo maggiore é l'incertezza con cui si individuano i coefficienti del fit.

Capitolo 2

Risultati

In questo capitolo sono esposti i risultati delle analisi descritte precedentemente. Anche in questo caso, si ha una suddivisione in diverse sezioni, ognuna delle quali tratta un aspetto particolare dell'analisi. La prima parte affronta i risultati ottenuti dalla caratterizzazione delle N, ovvero tutti quei nucleotidi non sequenziati correttamente, mentre la seconda e la terza mostrano rispettivamente i risultati ottenuti dalle analisi dei dinucleotidi CG e TA. L'ultima sezione si occupa di mostrare i valori dei parametri utilizzati per valutare la bontá dei fit.

2.1 Studio delle N

In questo paragrafo sono riportati i risultati delle analisi condotte sulle N. In particolare in una prima sottosezione verranno illustrati gli esiti dello studio della posizione delle basi sconosciute lungo le sequenze di DNA, mentre in una seconda sottosezione saranno riportati in tabelle i dati ottenuti dallo studio delle sequenze di N ripetute consecutivamente.

2.1.1 Posizione delle N

Si é scelto di rappresentare i dati ottenuti dall'analisi sulle posizioni delle N lungo le sequenze di DNA mediante istogrammi. Tali istogrammi sono stati realizzati attraverso la funzione di MATLAB *hist* a cui sono stati dati, come parametri d'ingresso, il vettore contenente tutte le posizioni delle N ed il numero di bin totali (fissato a 100). In questo modo, la lunghezza totale, in numero di basi, del cromosoma analizzato é stata divisa in 100 parti. Le altezze dei rettangoli dell'istogramma si riferiscono al numero di basi sconosciute presenti all'interno di queste sezioni, della sequenza di nucleotidi, precedentemente individuate.

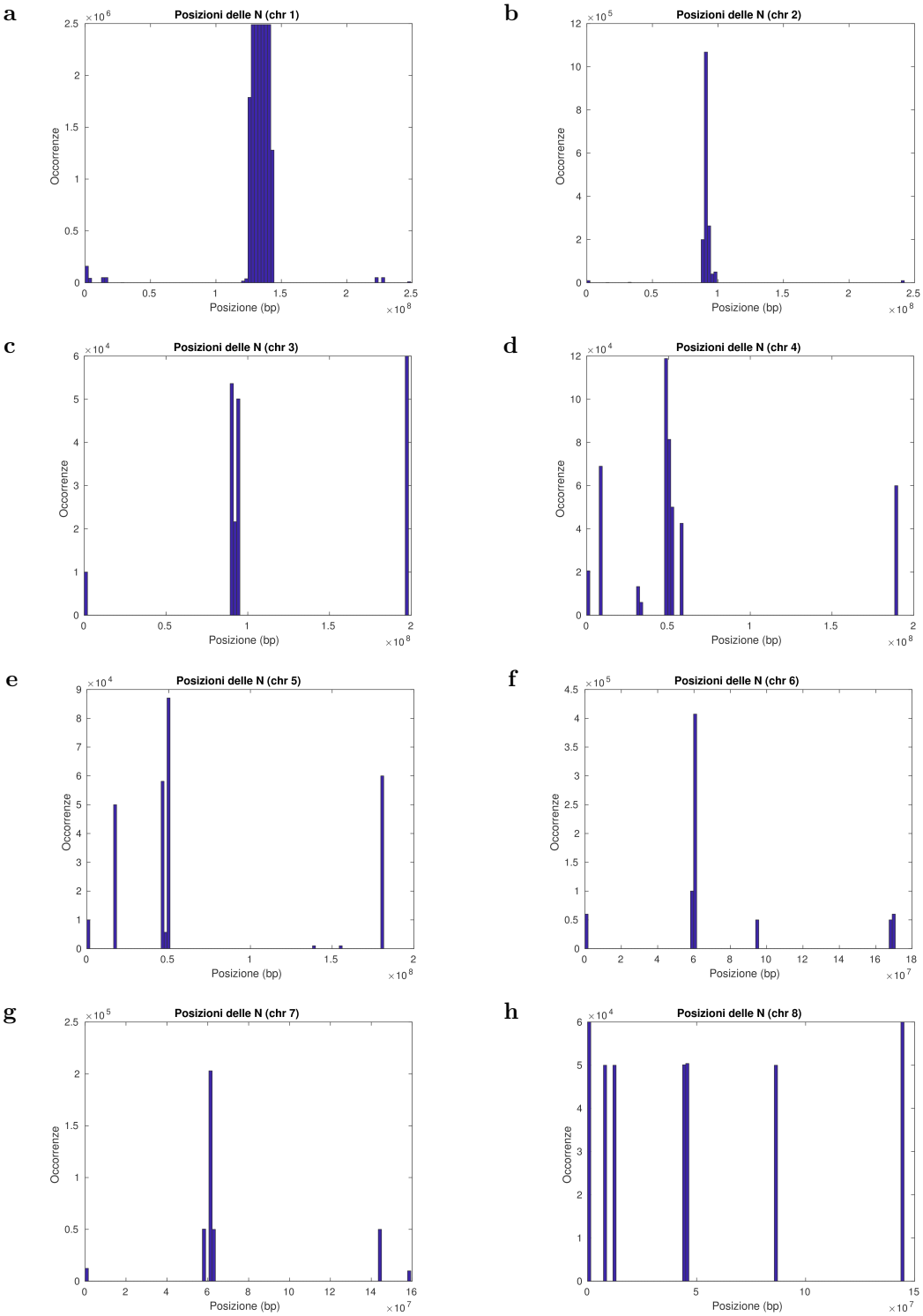

Figura 2.1: Parte 1. *Posizioni delle N, lungo le sequenze di nucleotidi, per i vari cromosomi umani.*

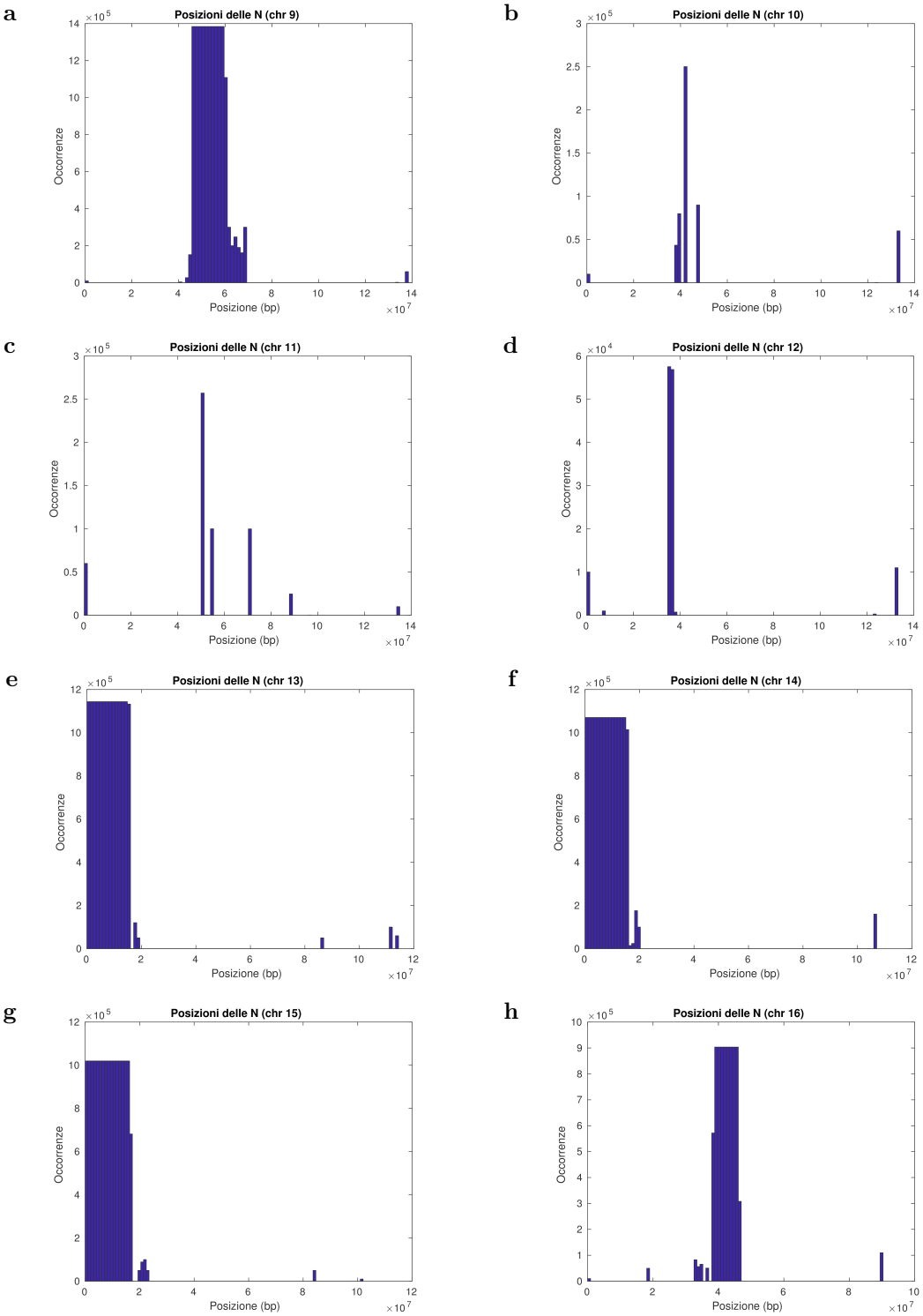

Figura 2.2: Parte 2. *Posizioni delle N, lungo le sequenze di nucleotidi, per i vari cromosomi umani.*

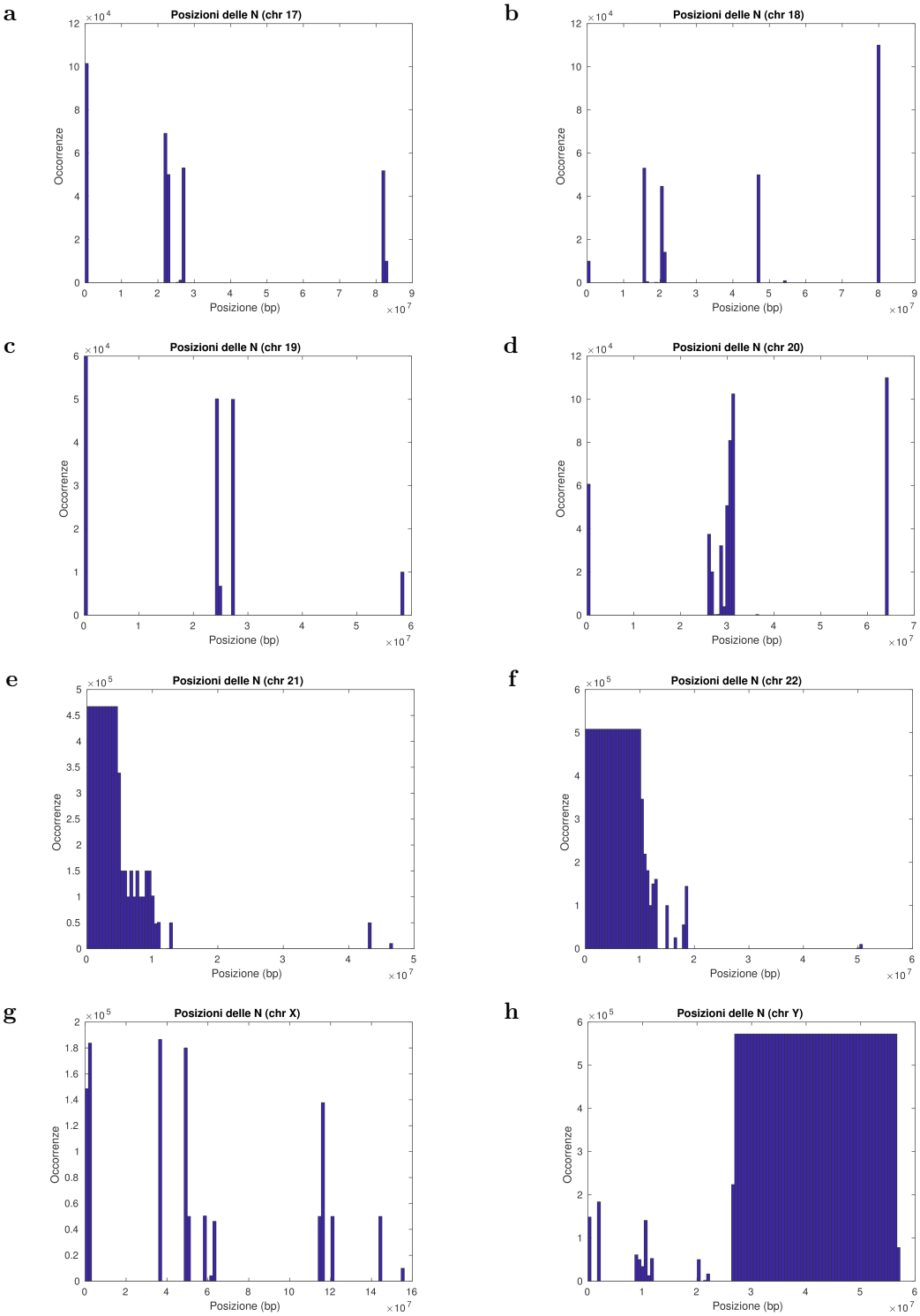

Figura 2.3: Parte 3. *Posizioni delle N, lungo le sequenze di nucleotidi, per i vari cromosomi umani.*

Quello che emerge dalle figure 2.1, 2.2 e 2.3 é che in generale le N, all'interno dei cromosomi, sono presenti in blocchi. Piú nello specifico, si osserva che le N sono distribuite generalmente nella parte iniziale, centrale e finale del cromosoma, fatto accentuato quando si ha a che fare con cromosomi con un alto contenuto di N.

2.1.2 Lunghezza delle sequenze di N

Come descritto nella sezione 1.2, nella seconda parte dello studio sulle basi sconosciute, sono state analizzate le sequenze di N ripetute consecutivamente. I risultati di questo studio sono raccolti in due tabelle (vedere Tab. 2.1, 2.2).
La prima tabella contiene informazioni riguardanti il numero di sequenze di N consecutive individuate, cromosoma per cromosoma. La seconda tabella invece illustra le dimensioni, in termini di numero di basi, delle varie sequenze, dividendole per ordini di grandezza.

Chr	**Numero N**	**Numero seq. N**	**Chr**	**Numero N**	**Numero seq. N**
1	$1,85 \times 10^7$	166	13	$1,64 \times 10^7$	20
2	$1,65 \times 10^6$	26	14	$1,65 \times 10^7$	25
3	$1,95 \times 10^5$	22	15	$1,73 \times 10^7$	19
4	$4,62 \times 10^5$	18	16	$8,53 \times 10^6$	21
5	$2,73 \times 10^5$	37	17	$3,37 \times 10^5$	36
6	$7,27 \times 10^5$	15	18	$2,84 \times 10^5$	61
7	$3,76 \times 10^5$	17	19	$1,77 \times 10^5$	9
8	$3,70 \times 10^5$	12	20	$4,99 \times 10^5$	90
9	$1,66 \times 10^7$	43	21	$6,62 \times 10^6$	49
10	$5,34 \times 10^5$	44	22	$1,17 \times 10^7$	44
11	$5,53 \times 10^5$	17	X	$1,15 \times 10^6$	29
12	$1,37 \times 10^5$	27	Y	$3,08 \times 10^7$	56

Tabella 2.1: *Numero di basi sconosciute e di sequenze di N ripetute consecutivamente nei vari cromosomi umani.*

Come si puó osservare dalla tabella 2.1 il numero di queste sequenze é estremamente basso ad eccezione dei cromosomi 1 e 20 che assumono valori piú elevati. Tuttavia confrontando questi valori con il numero di interdistanze, del dinucleotide d'interesse, individuate nell'analisi (generalmente dell'ordine di grandezza di qualche milione) si osserva che sono completamente trascurabili. Ció comporta che il numero di interdistanze influenzate dalla presenza delle N (e quindi da scartare) é irrilevante.

Cromosoma	**Occorrenze lunghezze sequenze N**								
	$\leq 10^1$	$\leq 10^2$	$\leq 10^3$	$\leq 10^4$	$\leq 10^5$	$\leq 10^6$	$\leq 10^7$	$> 10^7$	(bp)
1	89	3	59	3	11	0	0	1	
2	4	0	10	3	6	2	1	0	
3	2	4	8	3	5	0	0	0	
4	0	1	4	3	9	1	0	0	
5	0	0	20	12	5	0	0	0	
6	1	0	7	1	4	2	0	0	
7	0	1	5	2	9	0	0	0	
8	0	0	5	0	7	0	0	0	
9	0	7	6	6	19	4	0	1	
10	7	9	13	7	6	2	0	0	
11	0	0	11	0	3	3	0	0	
12	2	3	13	5	4	0	0	0	
13	0	0	12	0	6	1	0	1	
14	0	0	17	3	1	3	0	1	
15	2	1	7	0	8	0	0	1	
16	0	0	10	2	7	1	1	0	
17	0	4	23	1	8	0	0	0	
18	25	1	28	2	4	1	0	0	
19	0	0	4	1	4	0	0	0	
20	10	16	48	7	8	1	0	0	
21	1	5	13	0	26	3	1	0	
22	0	0	17	4	20	2	0	1	
X	0	1	13	0	11	4	0	0	
Y	0	6	22	12	14	1	0	1	

Tabella 2.2: *Occorrenze delle lunghezze, in termini di numero di basi, delle sequenze di N ripetute consecutivamente, nei vari cromosomi umani. Le colonne individuano diversi range di lunghezze.*

La tabella 2.2 mostra come le N , all'interno dei cromosomi, si distribuiscono in grosse sequenze costituite da N ripetute consecutivamente. Si puó infatti vedere come in alcuni cromosomi (ad esempio chr. 1) esistono sequenze di N lunghe, in termini di numero di basi, piú di dieci milioni di nucleotidi. Gli stessi risultati sono riportati nei seguenti istogrammi (Fig.) É bene rimarcare, in seguito a questi risultati, che la rimozione delle N é statisticamente ininfluente sui risultati ottenuti dallo studio delle interdistanze dei dinucleotidi CG e TA.

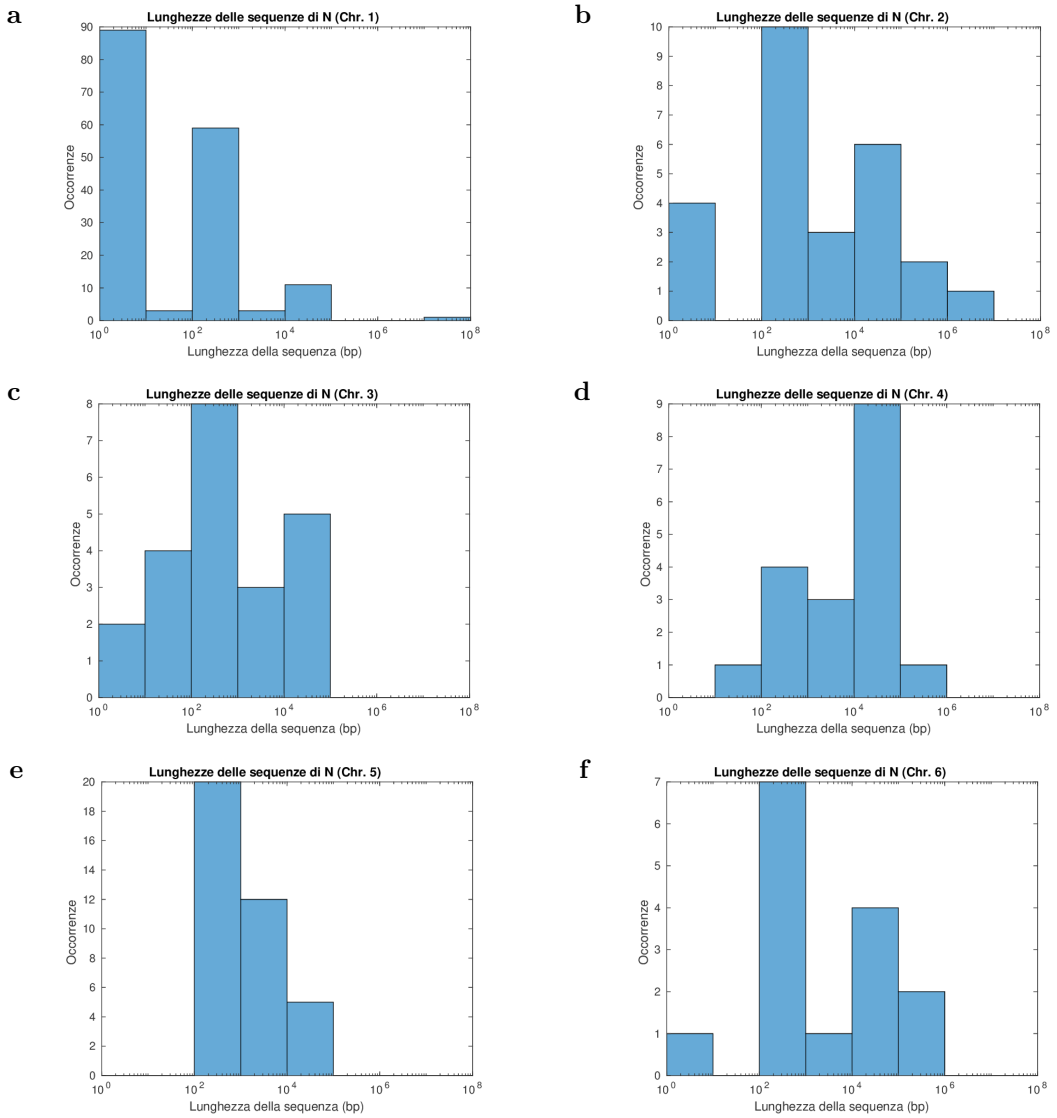

Figura 2.4: Parte 1. *Istogrammi contenenti le occorrenze delle lunghezze delle sequenze di N ripetute consecutivamente, nei vari cromosomi umani .*

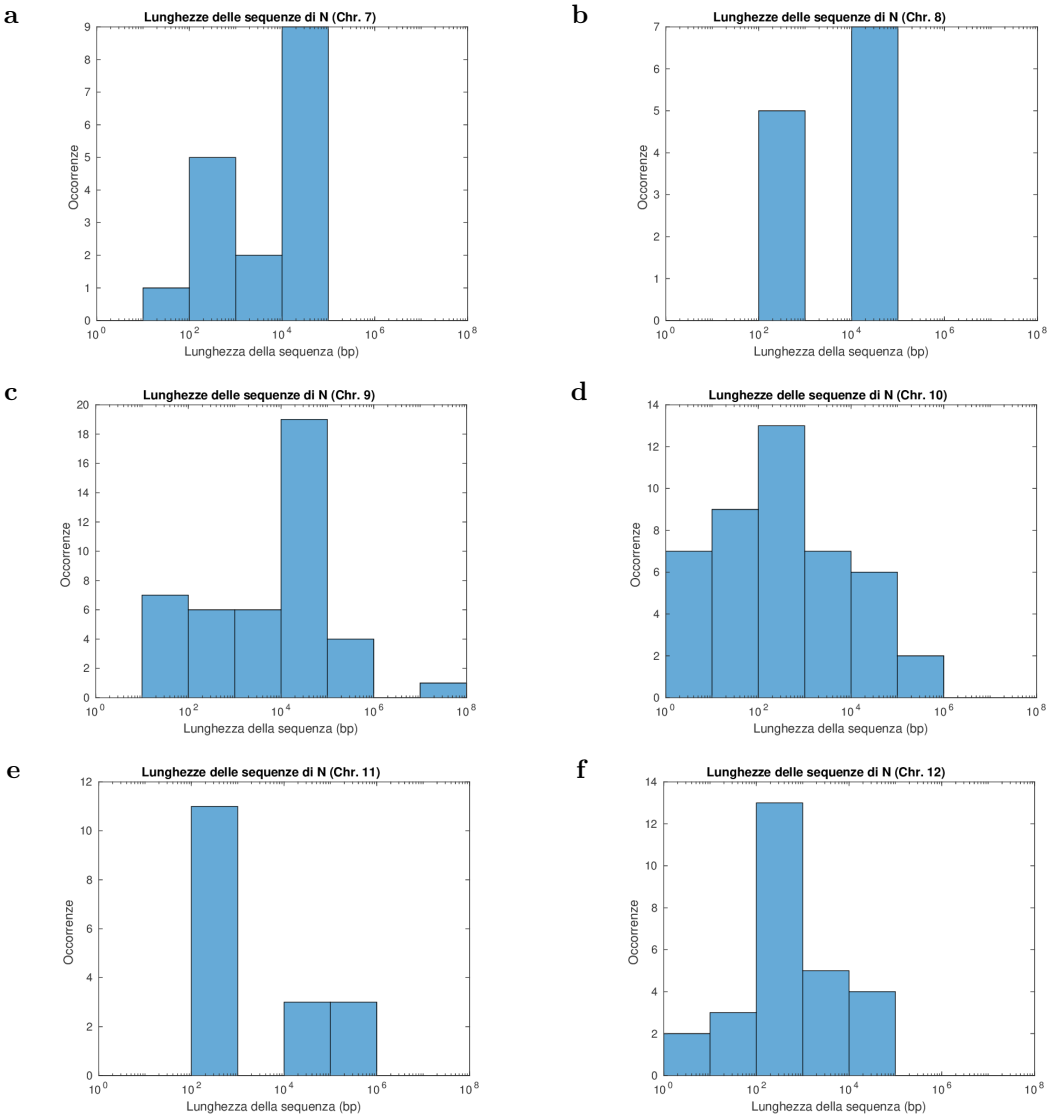

Figura 2.5: Parte 2. *Istogrammi contenenti le occorrenze delle lunghezze delle sequenze di N ripetute consecutivamente, nei vari cromosomi umani .*

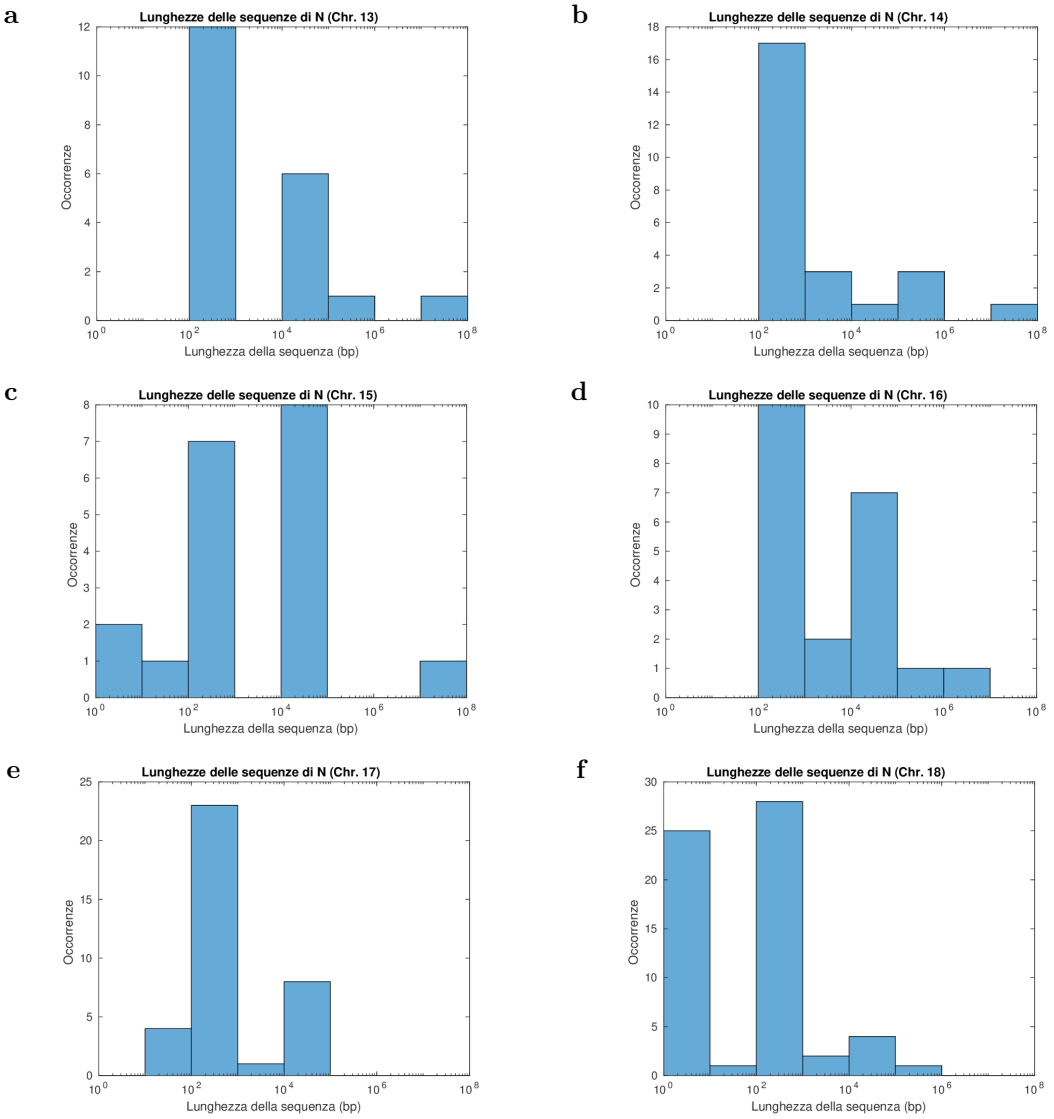

Figura 2.6: Parte 3. *Istogrammi contenenti le occorrenze delle lunghezze delle sequenze di N ripetute consecutivamente, nei vari cromosomi umani .*

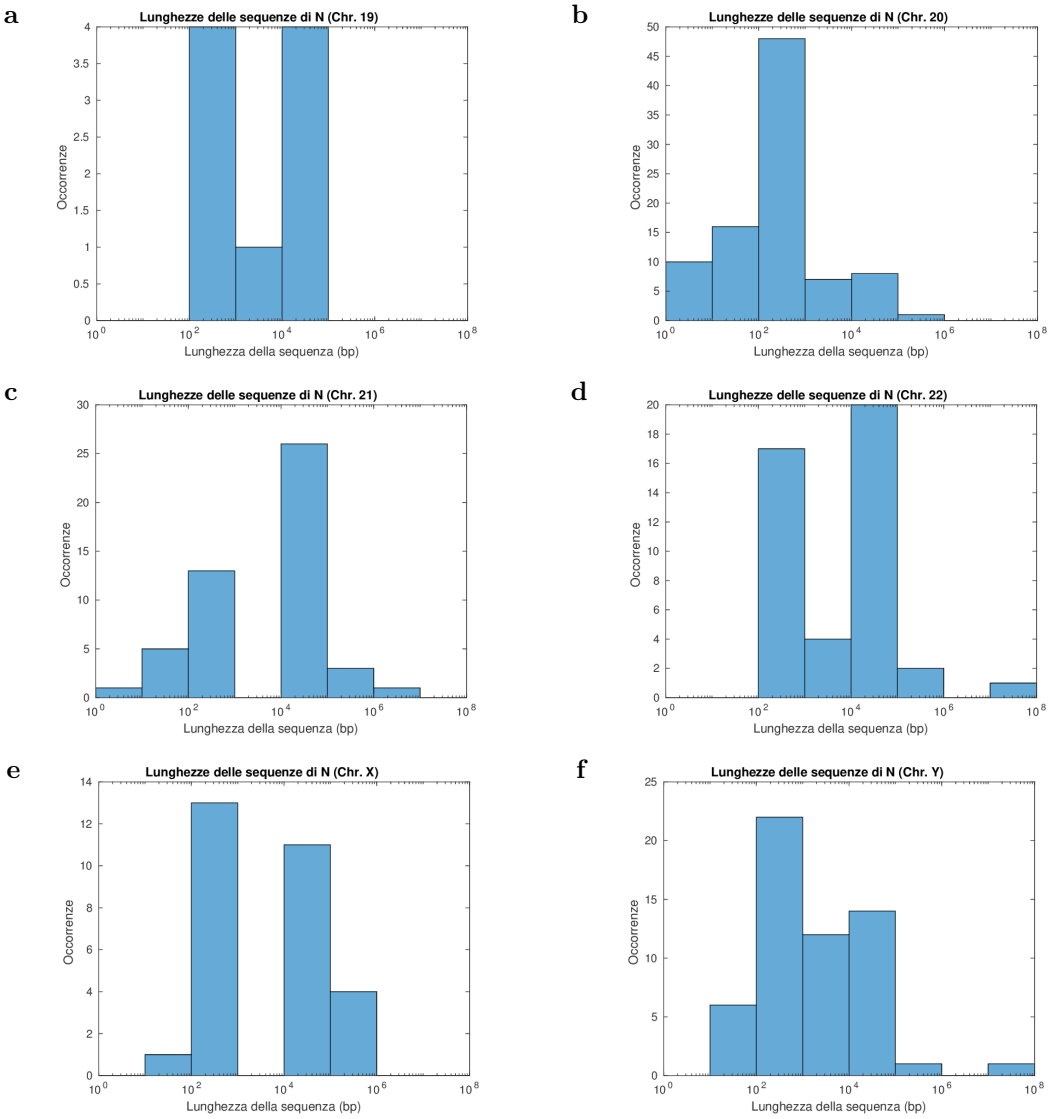

Figura 2.7: Parte 4. *Istogrammi contenenti le occorrenze delle lunghezze delle sequenze di N ripetute consecutivamente, nei vari cromosomi umani .*

2.2 Studio del dinucleotide CG

Questo paragrafo é suddiviso in tre sezioni distinte: le prima due si occupano di illustrare gli esiti dei fit fatti con l'eq. 1.3 sulle distribuzioni delle interdistanze, mentre la terza di analizzare e commentare i risultati ottenuti. Per completezza il logaritmo della legge di potenza traslata con cutoff esponenziale, descritto dall'eq. 1.3, é richiamato qui di seguito:

$$p(x) = \log(c(x+d)^{-a}e^{-\frac{x}{b}}).$$

2.2.1 Risultati dei fit: figure

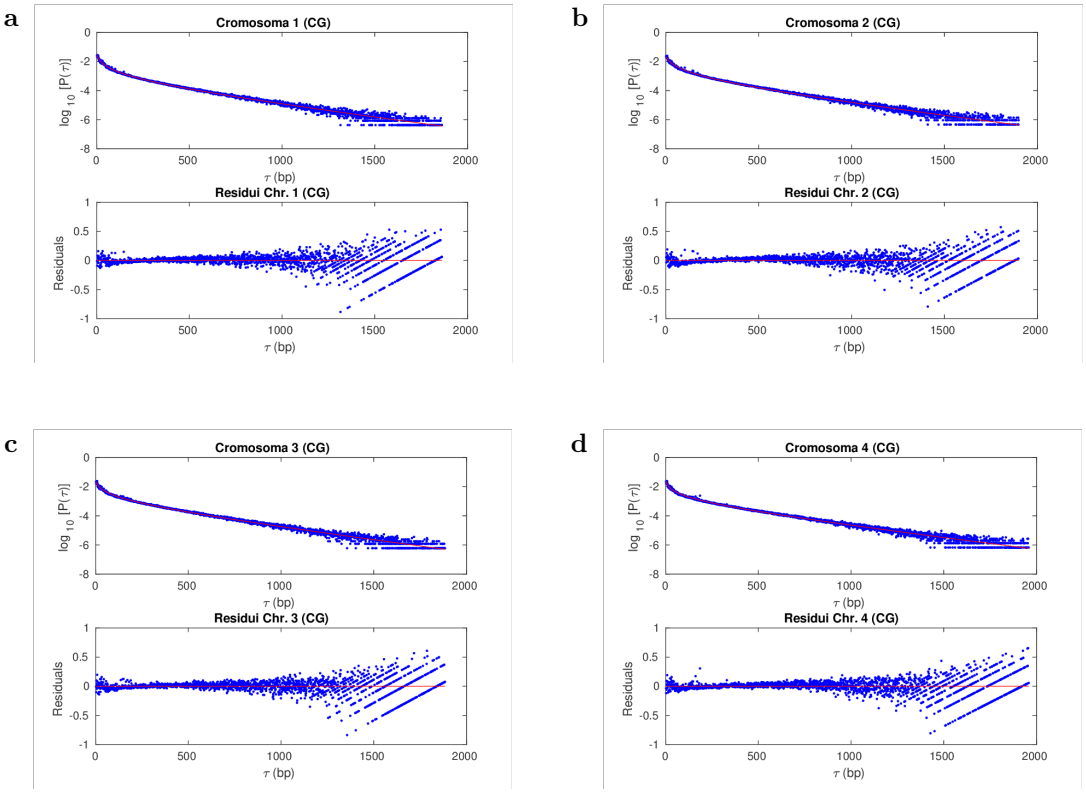

Figura 2.8: Parte 1. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide CG fittate con l'eq. 1.3, insieme ai rispettivi grafici dei residui.*

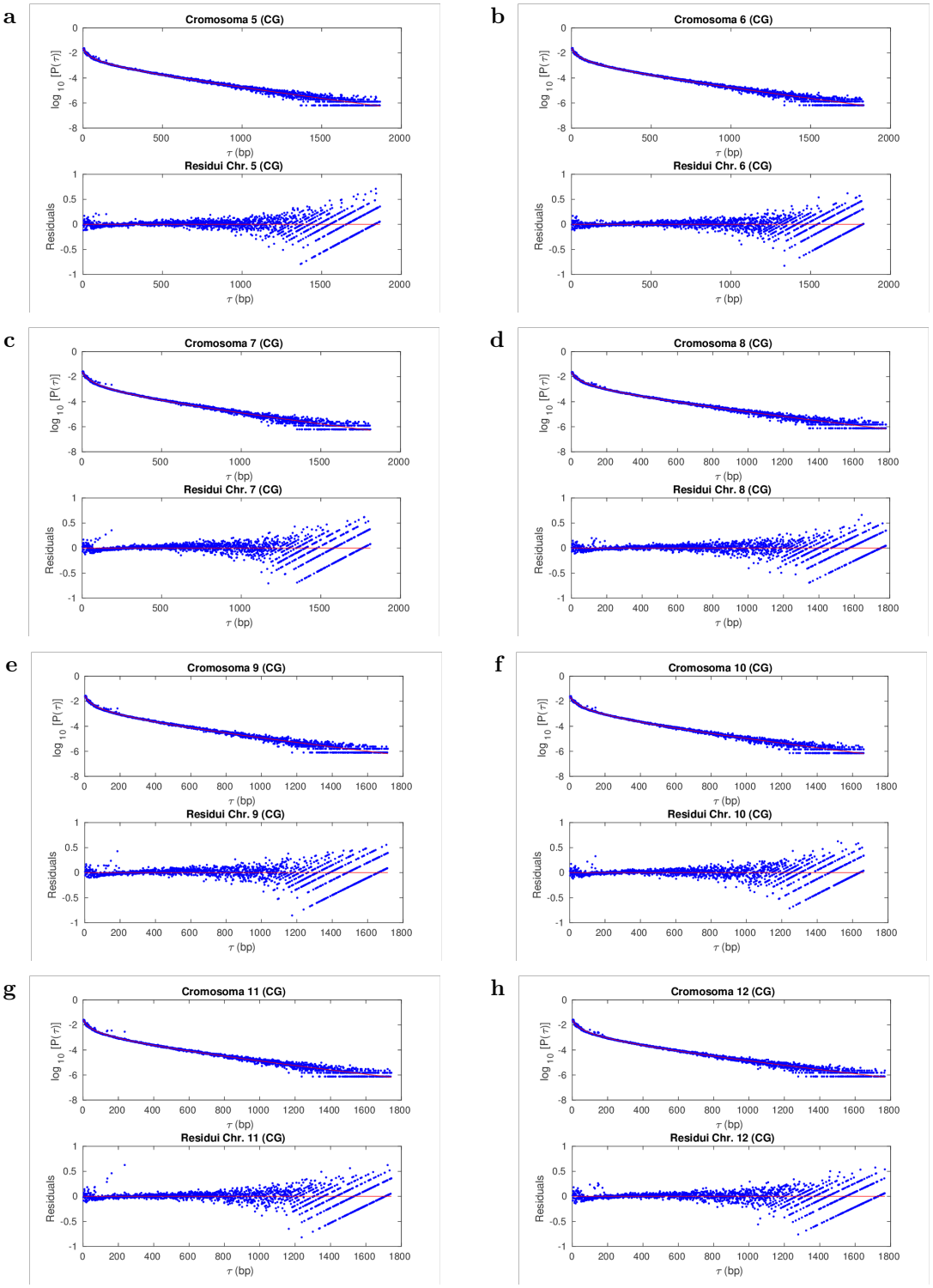

Figura 2.9: Parte 2. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide CG fittate con l'eq. 1.3, insieme ai rispettivi grafici dei residui.*

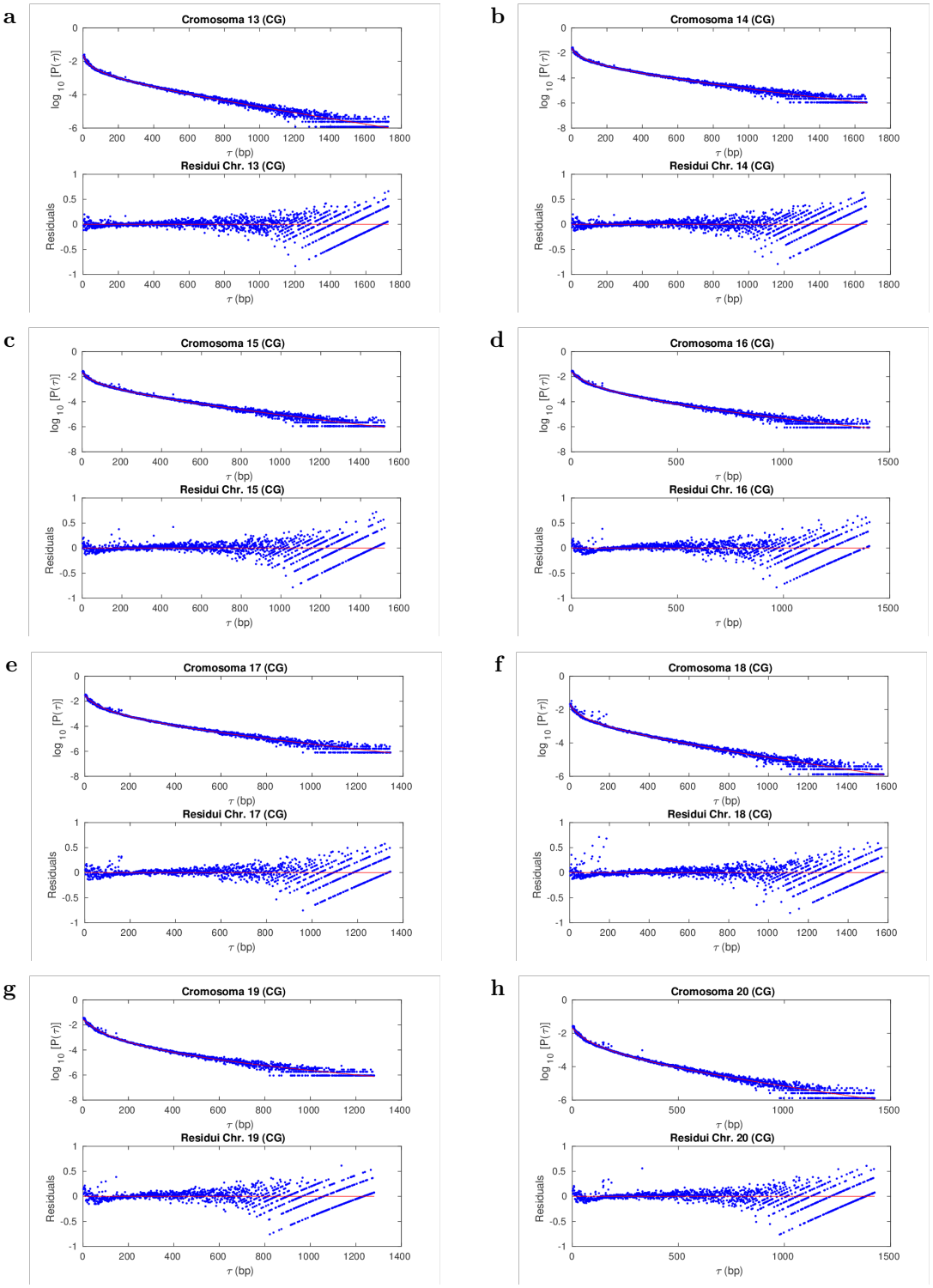

Figura 2.10: Parte 3. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide CG fittate con l'eq. 1.3, insieme ai rispettivi grafici dei residui.*

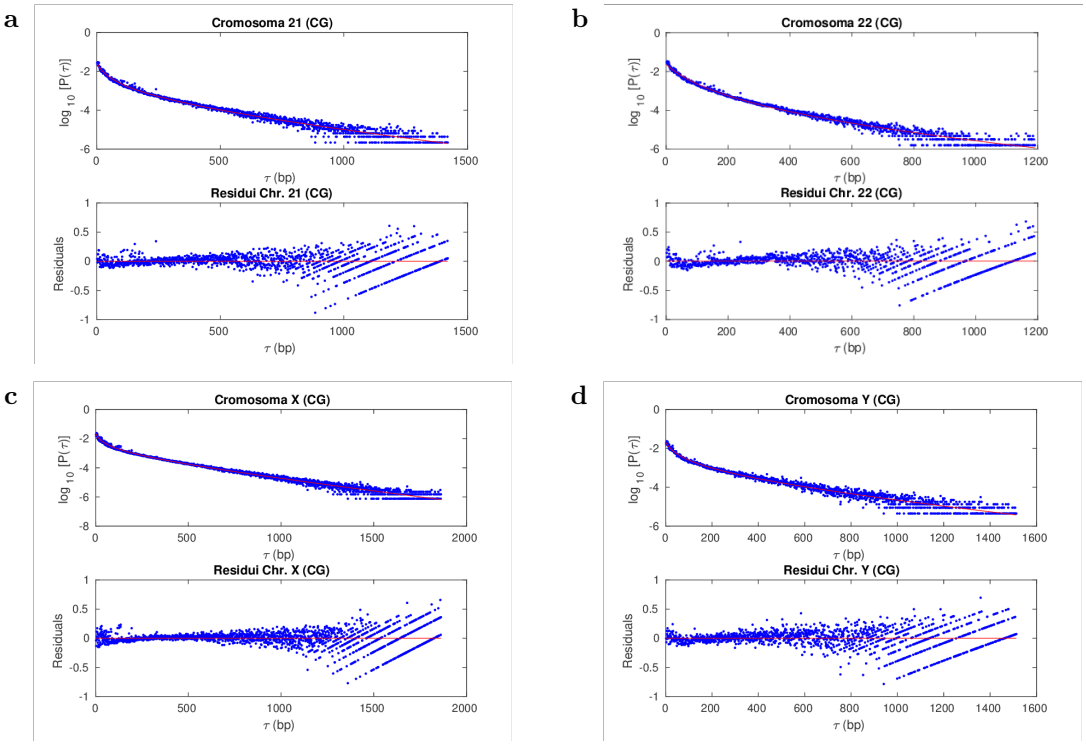

Figura 2.11: Parte 4. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide CG fittate con l'eq. 1.3, insieme ai rispettivi grafici dei residui.*

2.2.2 Risultati dei fit: parametri

I parametri a, b, c, d (vedere eq. 1.3) insieme ai rispettivi intervalli di confidenza, ottenuti mediante le funzioni MATLAB *fit* e *confint*, sono riportati nelle seguenti tabelle e figure (vedere Tab. 2.3 e Fig. 2.12, 2.13).

Osservando le figure 2.12 e 2.13 si possono trarre informazioni molto importanti. In particolare, guardando la figura 2.12, si puó notare un raggruppamento di punti centrale ed alcuni punti visibilmente distaccati dal resto del gruppo. Dal momento che un punto,in questo tipo di grafico, caratterizza un particolare cromosoma (avendo come coordinate i rispettivi parametri a,b e d ottenuti dal fit), si puó osservare che esiste un insieme, molto numeroso, di cromosomi con parametri simili. Inoltre vi sono, sebbene in un numero ridotto, anche cromosomi con parametri che assumono valori "anomali", lontani dalla media: é il caso dei cromosomi 16, 17, 19, 20, 22 e Y.

Chr.	**a**	**b (bp)**	**c (bp^a)**	**d (bp)**
1	$1,02 \pm 0,08$	298 ± 8	$0,4 \pm 0,2$	12 ± 6
2	$0,99 \pm 0,08$	302 ± 9	$0,4 \pm 0,2$	18 ± 8
3	$0,81 \pm 0,07$	287 ± 7	$0,16 \pm 0,06$	19 ± 6
4	$0,76 \pm 0,06$	302 ± 7	$0,12 \pm 0,04$	6 ± 5
5	$0,83 \pm 0,07$	293 ± 8	$0,18 \pm 0,07$	9 ± 6
6	$0,88 \pm 0,07$	300 ± 8	$0,21 \pm 0,08$	10 ± 6
7	$1,11 \pm 0,09$	317 ± 10	$0,7 \pm 0,4$	18 ± 8
8	$0,97 \pm 0,08$	302 ± 9	$0,3 \pm 0,2$	15 ± 7
9	$1,07 \pm 0,09$	303 ± 10	$0,5 \pm 0,3$	15 ± 7
10	$1,1 \pm 0,1$	297 ± 10	$0,7 \pm 0,4$	20 ± 8
11	$0,99 \pm 0,08$	301 ± 10	$0,4 \pm 0,2$	12 ± 6
12	$1,07 \pm 0,09$	315 ± 10	$0,5 \pm 0,3$	16 ± 7
13	$0,85 \pm 0,07$	299 ± 9	$0,19 \pm 0,08$	10 ± 6
14	$1,1 \pm 0,1$	315 ± 10	$0,5 \pm 0,3$	17 ± 8
15	$1,2 \pm 0,1$	285 ± 10	$1,2 \pm 0,9$	28 ± 12
16	$\mathbf{1,9 \pm 0,2}$	$\mathbf{360 \pm 40}$	$\mathbf{51 \pm 70}$	$\mathbf{52 \pm 16}$
17	$\mathbf{1,7 \pm 0,2}$	$\mathbf{308 \pm 20}$	$\mathbf{10 \pm 10}$	$\mathbf{33 \pm 11}$
18	$1,0 \pm 0,1$	309 ± 13	$0,5 \pm 0,3$	20 ± 10
19	$\mathbf{2,9 \pm 0,3}$	$\mathbf{631 \pm 150}$	$\mathbf{(0,6 \pm 1) \times 10^4}$	$\mathbf{70 \pm 17}$
20	$\mathbf{1,6 \pm 0,2}$	$\mathbf{333 \pm 30}$	$\mathbf{11 \pm 13}$	$\mathbf{45 \pm 16}$
21	$1,4 \pm 0,1$	353 ± 30	2 ± 2	24 ± 10
22	$\mathbf{2,8 \pm 0,4}$	$\mathbf{493 \pm 120}$	$\mathbf{(0,6 \pm 2) \times 10^4}$	$\mathbf{86 \pm 30}$
X	$0,88 \pm 0,07$	306 ± 9	$0,22 \pm 0,09$	12 ± 7
Y	$\mathbf{1,2 \pm 0,2}$	$\mathbf{404 \pm 30}$	$\mathbf{1,1 \pm 0,9}$	$\mathbf{26 \pm 13}$

Tabella 2.3: *Valori dei parametri estratti dal fit con l'eq. 1.3 fatto sulle distribuzioni delle interdistanze del dinucleotide CG in scala semilogaritmica. Il parametro a é adimensionale. Gli errori sui parametri dei fit sono stimati ad un intervallo di confidenza del 95 %.*

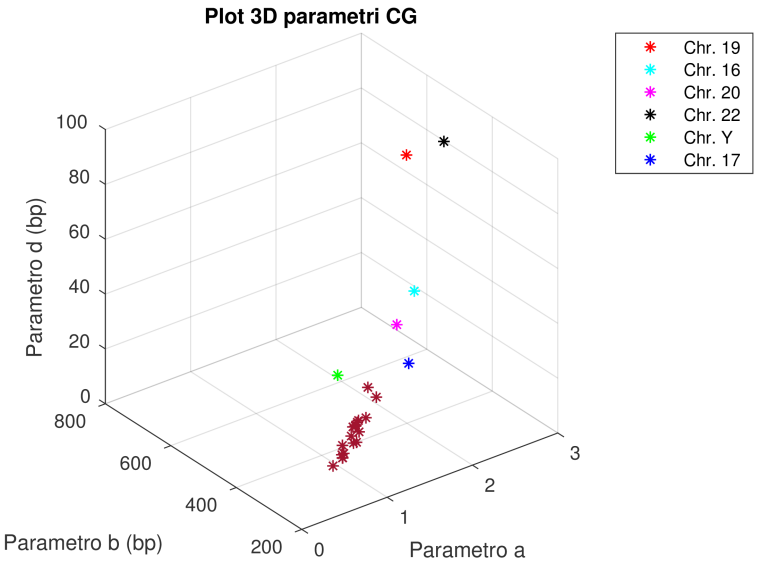

Figura 2.12: *Plot 3D dei parametri a,b,d riportati nella tabella 2.3.*

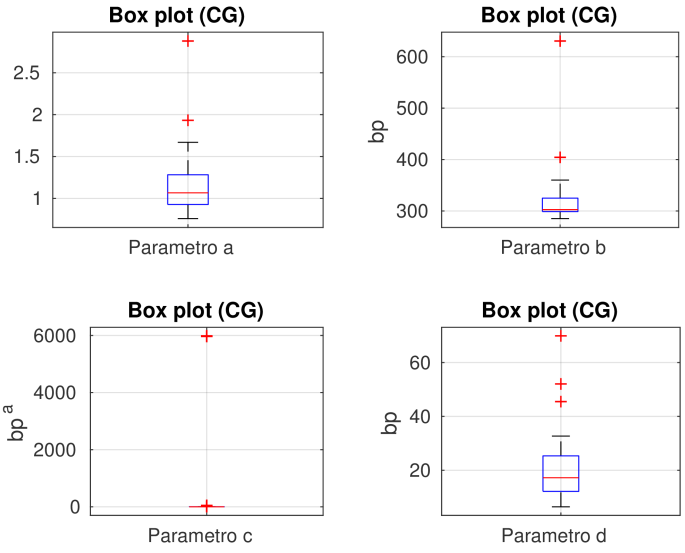

Figura 2.13: *Box Plot dei quattro parametri riportati nella tabella 2.3.*

2.2.3 Relazione tra il parametro b e la densitá di CG

Particolare attenzione é stata posta sui valori assunti dal parametro b. Infatti, secondo quanto riportato da [10] e [11], il parametro b dell' eq. 1.3 sembrerebbe essere legato ai valori medi di interdistanza del dinucleotide CG lungo la sequenza di DNA in considerazione. Dalla figura 2.13 e dalla tabella 2.3 emerge che i valori del parametro b dei cromosomi 19 e 22 sono outlier: essi infatti assumono valori, rispettivamente, di circa 600 e 500 bp a discapito della media che é di circa 300 bp.

Per fornire una spiegazione a questi valori anomali é stata proposta un'ipotesi: la distanza media tra dinucleotidi dovrebbe decrescere proporzionalmente all'inverso della densitá di dinucleotidi all'interno della sequenza. Per verificare tale ipotesi é stato condotto uno studio sulle percentuali del dinucleotide CG e dei nucleotidi C e G all'interno di ogni cromosoma, al fine di verificare la presenza di una percentuale inferiore alla media qualora il valore di b risulti superiore ai 300 bp.

Per il calcolo delle percentuali si é deciso di procedere in due modi diversi: un primo prevede di includere, all'interno del numero totale di basi che costituiscono il cromosoma, anche le basi non sequenziate (N), al contrario del secondo nel quale vengono ignorate. Inoltre é stato calcolato, in entrambi i casi, il rapporto tra la percentuale del dinucleotide CG ed il prodotto delle percentuali dei singoli nucleotidi C e G.

In tabella 2.4 sono mostrati i risultati ottenuti dallo studio condotto escludendo, dal numero totale di basi, le N. Come si puó osservare, nei cromosomi in cui il valore del parametro b é maggiore rispetto alla media generale, le percentuali dei nucleotidi C e G e del dinucleotide CG sono maggiori rispetto ai valori ottenuti negli altri cromosomi. Inoltre si osserva che il rapporto tra le percentuali assume per questi cromosomi un valore sensibilmente piú elevato.

In tabella 2.5 sono invece mostrati i risultati ottenuti dallo studio condotto includendo nel numero totale di basi anche le N. Anche in questo caso, nei cromosomi in cui il valore del parametro b é maggiore rispetto alla media, le percentuali dei nucleotidi C e G e del dinucleotide CG sono piú alte. Tuttavia, a differenza del caso precedente, il rapporto tra la percentuale del dinucleotide e il prodotto delle percentuali dei nucleotidi non fornisce alcuna informazione. Ció é da imputare al fatto che le percentuali delle N variano di molto da cromosoma a cromosoma, andando a costituire in un caso anche il 50% delle basi totali.

Si puó osservare che l'andamento delle percentuali di CG previsto dall'ipotesi fatta per motivare i valori anomali del parametro b, non si presenta. Infatti, secondo tale ipotesi, nei cromosomi con un valore alto di b (vedere ad esempio chr. 19 e chr. 22) ci si aspettava una bassa percentuale di CG rispetto alla media degli altri cromosomi. Quello che si osserva é, invece, l'esatto contrario.

Chr.	**%C**	**%G**	**%CG**	**%CG/%C%G**
1	20,85	20,87	1,03	0,00237
2	20,09	20,14	0,91	0,00225
3	19,80	19,86	0,84	0,00215
4	19,10	19,15	0,79	0,00217
5	19,71	19,79	0,86	0,00222
6	19,78	19,82	0,89	0,00227
7	20,33	20,37	1,02	0,00247
8	20,05	20,10	0,92	0,00229
9	20,61	20,67	1,03	0,00242
10	20,74	20,80	1,04	0,00242
11	20,74	20,80	0,99	0,00230
12	20,35	20,42	0,99	0,00238
13	19,23	19,32	0,88	0,00231
14	20,34	20,49	0,99	0,00237
15	20,97	21,06	1,07	0,00242
16	**22,21**	**22,37**	**1,41**	**0,00283**
17	**22,58**	**22,73**	**1,51**	**0,00293**
18	19,72	20,05	0,94	0,00239
19	**23,88**	**24,06**	**1,89**	**0,00329**
20	**21,76**	**22,04**	**1,21**	**0,00252**
21	20,42	20,52	1,15	0,00275
22	**23,39**	**23,61**	**1,62**	**0,00293**
X	19,71	19,82	0,85	0,00219
Y	**20,01**	**20,01**	**0,86**	**0,00214**

Tabella 2.4: *Valori delle percentuali dei nucleotidi C e G e del dinucleotide CG rispetto al numero totale delle basi azotate (escluse le N).*

Chr.	**%C**	**%G**	**%CG**	**%CG/%C%G**
1	19,30	19,33	0,95	0,00256
2	19,95	20,01	0,91	0,00227
3	19,79	19,84	0,84	0,00215
4	19,05	19,10	0,79	0,00217
5	19,68	19,76	0,86	0,00222
6	19,70	19,74	0,88	0,00227
7	20,28	20,32	1,02	0,00247
8	20,00	20,05	0,92	0,00230
9	18,14	18,19	0,91	0,00275
10	20,66	20,72	1,04	0,00243
11	20,66	20,71	0,99	0,00231
12	20,33	20,40	0,99	0,00238
13	16,47	16,56	0,74	0,00270
14	17,21	17,34	0,84	0,00281
15	17,41	17,48	0,89	0,00292
16	**20,12**	**20,26**	**1,27**	**0,00313**
17	**22,49**	**22,64**	**1,50**	**0,00294**
18	19,65	19,98	0,94	0,00240
19	**23,81**	**23,99**	**1,89**	**0,00330**
20	**21,59**	**21,87**	**1,20**	**0,00254**
21	17,52	17,61	0,99	0,00321
22	**18,03**	**18,19**	**1,25**	**0,00381**
X	19,56	19,67	0,85	0,00220
Y	**9,240**	**9,241**	**0,39**	**0,00463**

Tabella 2.5: *Valori delle percentuali dei nucleotidi C e G e del dinucleotide CG rispetto al numero totale delle basi azotate (incluse le N).*

2.3 Studio del dinucleotide TA

Il paragrafo é stato diviso in due sezioni, ognuna delle quali riporta i risultati della rispettiva analisi. Per completezza sono riportate qui di seguito l'equazione 1.3:

$$p(x) = \log(c(x+d)^{-a}e^{-\frac{x}{b}}),$$

e l'equazione 1.5:

$$p(x) = log(c(x+d)^{-a}).$$

2.3.1 Studio del cromosoma 1

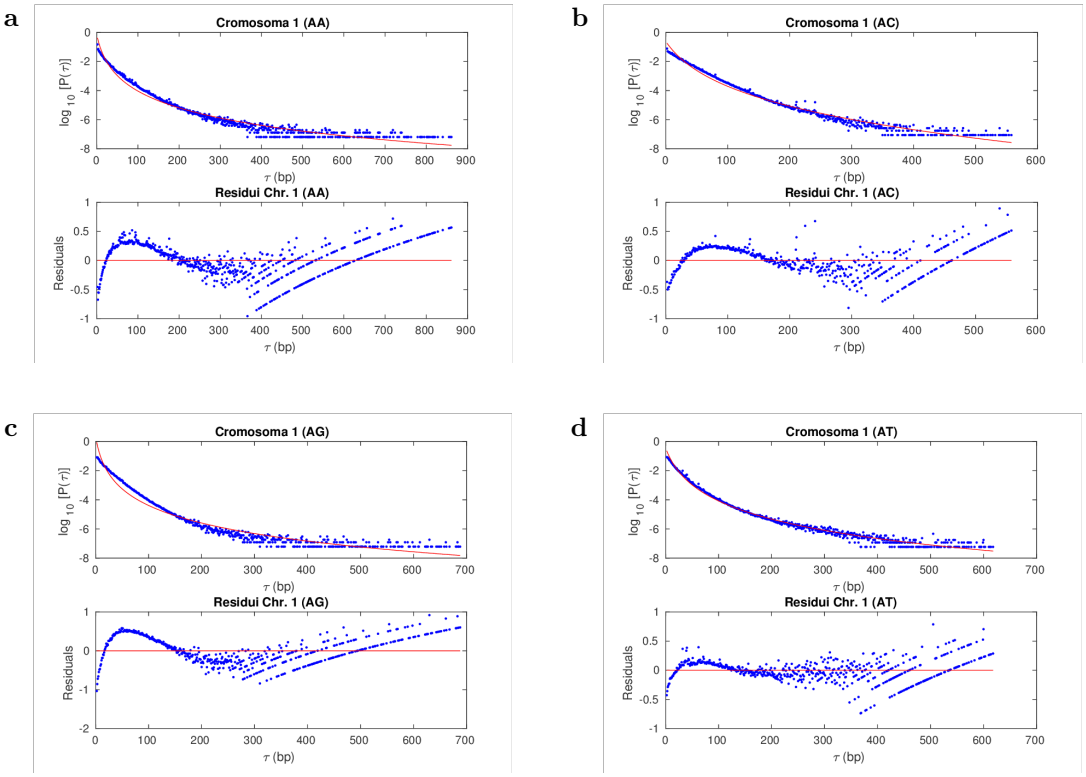

Figura 2.14: Parte 1. *Fit con eq. 1.3 delle distribuzioni delle interdistanze, in scala semilogaritmica rispetto all'asse delle Y, di tutti e 16 i dinucleotidi nel cromosoma 1, insieme ai rispettivi grafici dei residui.*

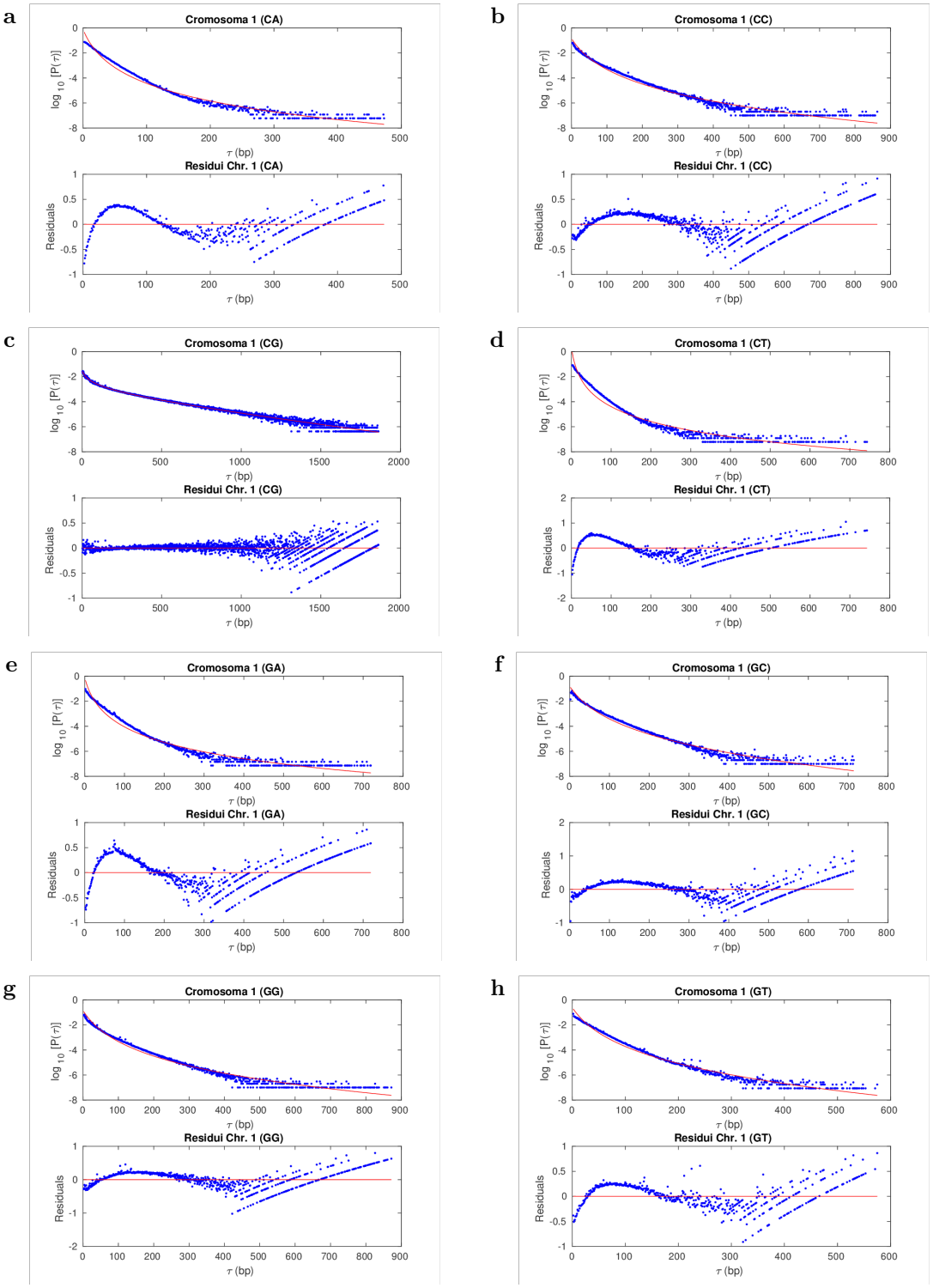

Figura 2.15: Parte 2. *Fit con eq. 1.3 delle distribuzioni delle interdistanze,in scala semilogaritmica rispetto all'asse delle Y, di tutti e 16 i dinucleotidi nel cromosoma 1, insieme ai rispettivi grafici dei residui.*

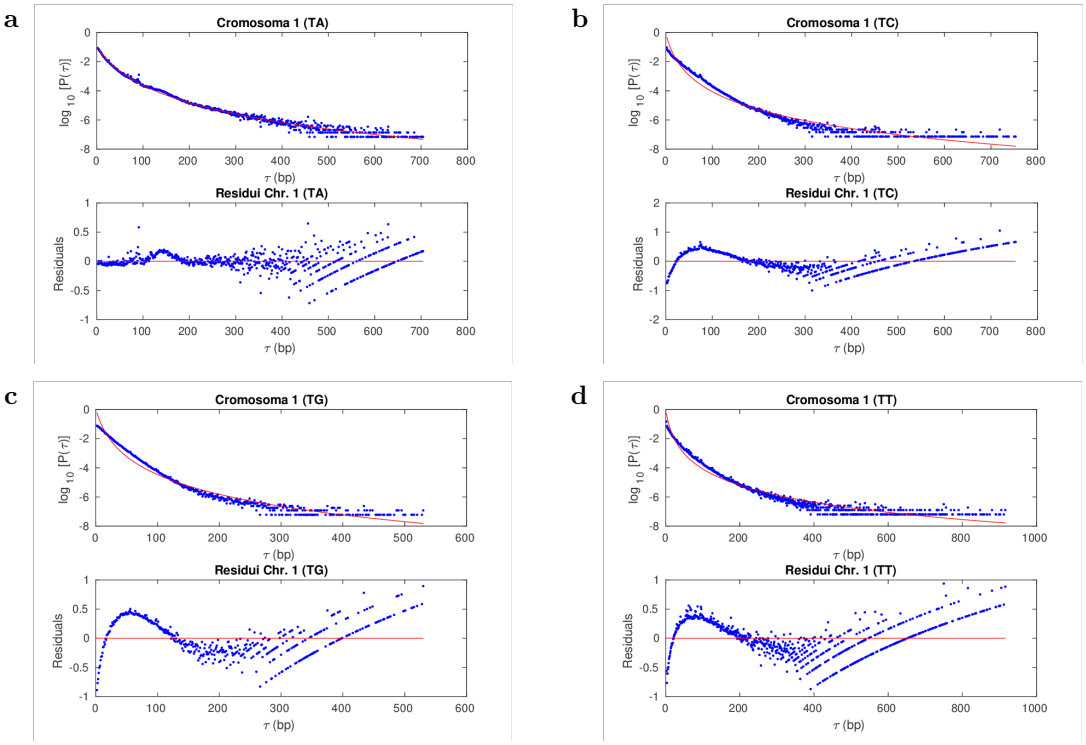

Figura 2.16: Parte 3. *Fit con eq. 1.3 delle distribuzioni delle interdistanze, in scala semilogaritmica rispetto all'asse delle Y, di tutti e 16 i dinucleotidi nel cromosoma 1, insieme ai rispettivi grafici dei residui.*

Osservando i grafici delle interdistanze dei vari dinucleotidi fittati con l'eq. 1.3 e i rispettivi grafici dei residui, si puó notare che il dinucleotide TA (Fig. 2.15g) é quello che si comporta nel modo piú simile al dinucleotide CG (Fig. 2.15a). A parte per una piccola gobba presente tra i 100-200 bp l'eq. 1.3 descrive in modo corretto anche il comportamento del dinucleotide TA (fatto sostenuto dal grafico dei residui). CG e TA sono entrambi ben descritti dallo stesso tipo di funzione matematica, perció si deduce un loro possibile collegamento.

2.3.2 Caratterizzazione del dinucleotide TA

Di seguito sono riportati i grafici contenenti le distribuzioni delle interdistanze del dinucleotide TA, in scala semilogaritmica rispetto all'asse delle Y, fittate con l'eq. 1.3.

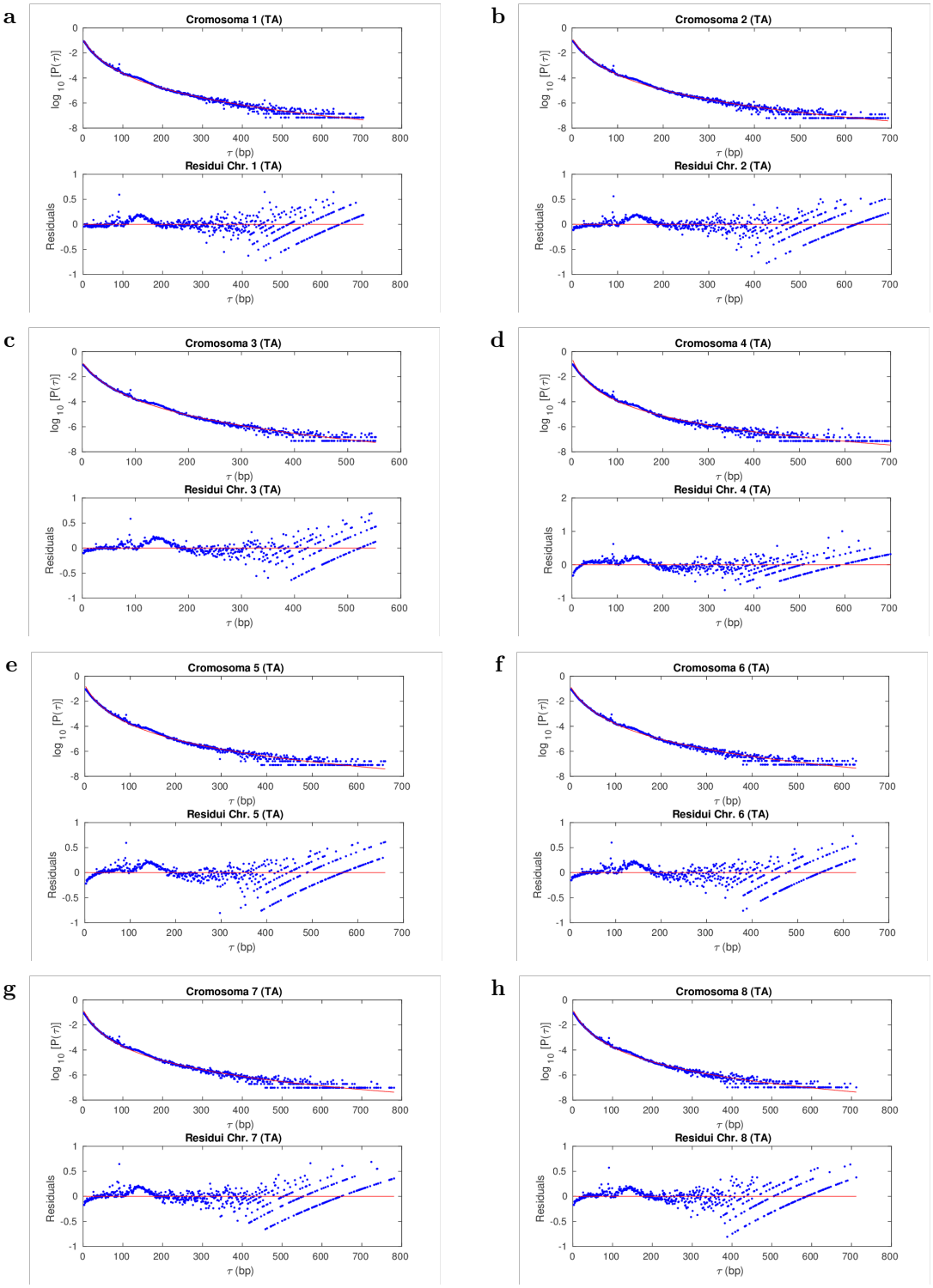

Figura 2.17: Parte 1. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide TA fittate con l'eq. 1.3, insieme ai rispettivi grafici dei residui.*

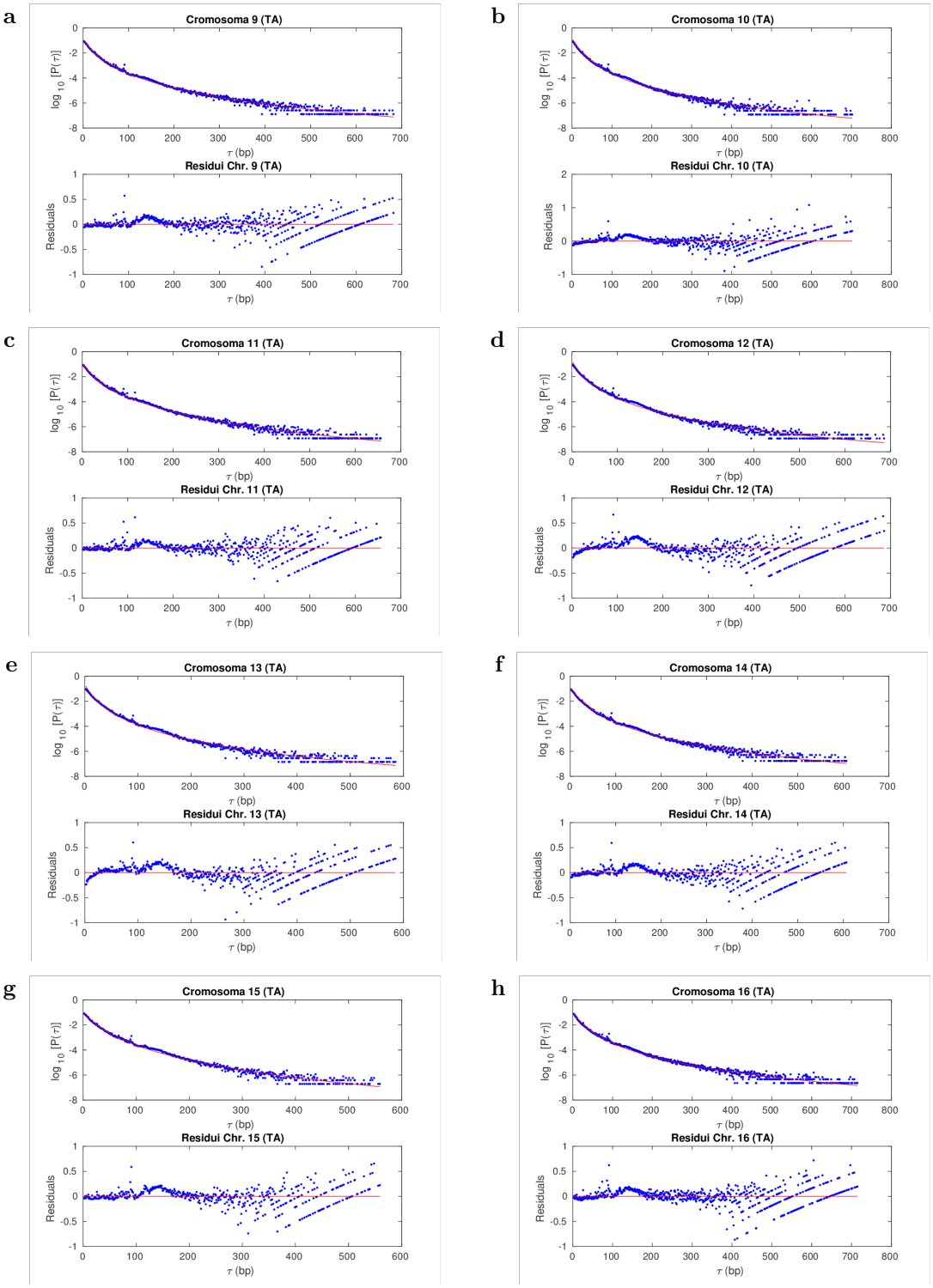

Figura 2.18: Parte 2. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide TA fittate con l'eq. 1.3, insieme ai rispettivi grafici dei residui.*

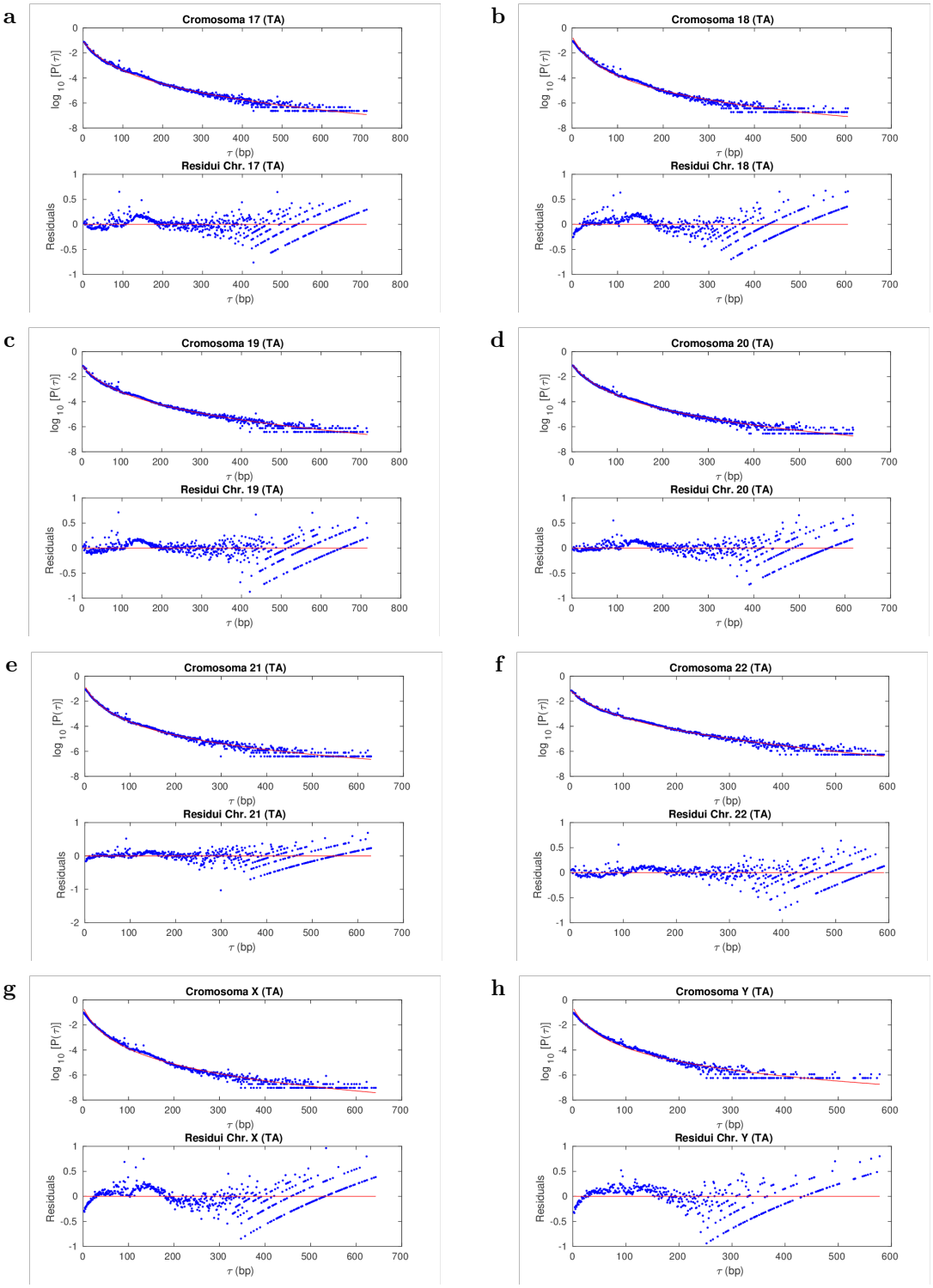

Figura 2.19: Parte 3. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide TA fittate con l'eq. 1.3, insieme ai rispettivi grafici dei residui.*

I parametri a, b, c, d (vedere eq. 1.3) insieme ai rispettivi intervalli di confidenza, ottenuti mediante le funzioni MATLAB *fit* e *confint*, sono riportati nelle seguenti tabelle (vedere Tab. 2.6 e 2.7).

Chr.	**a**	**b (bp)**	**c (bp^a)**	**d (bp)**
1	$4,9 \pm 0,5$	$(2 \pm 4) \times 10^3$	$(1 \pm 3) \times 10^7$	40 ± 9
2	$5,1 \pm 0,5$	$(0,08 \pm 7) \times 10^6$	$(1 \pm 3) \times 10^7$	36 ± 8
3	$5,2 \pm 0,7$	$(1 \pm 2) \times 10^3$	$(2 \pm 8) \times 10^7$	37 ± 10
4	$4,5 \pm 0,4$	$(0,005 \pm 2) \times 10^8$	$(3 \pm 6) \times 10^5$	21 ± 5
5	$4,9 \pm 0,5$	$(0,005 \pm 3) \times 10^8$	$(0,3 \pm 1) \times 10^7$	29 ± 8
6	$5,1 \pm 0,6$	$(0,001 \pm 1) \times 10^9$	$(1 \pm 3) \times 10^7$	34 ± 8
7	$4,6 \pm 0,4$	$(0,004 \pm 2) \times 10^8$	$(0,9 \pm 2) \times 10^6$	29 ± 7
8	$4,8 \pm 0,5$	$(0,004 \pm 2) \times 10^8$	$(2 \pm 5) \times 10^6$	30 ± 7
9	$4,8 \pm 0,5$	$(0,1 \pm 2) \times 10^5$	$(3 \pm 9) \times 10^6$	36 ± 8
10	$4,8 \pm 0,6$	$(0,005 \pm 3) \times 10^8$	$(0,4 \pm 1) \times 10^7$	35 ± 9
11	$4,9 \pm 0,5$	$(3 \pm 8) \times 10^3$	$(0,6 \pm 2) \times 10^7$	38 ± 9
12	$4,8 \pm 0,5$	$(0,006 \pm 4) \times 10^8$	$(2 \pm 7) \times 10^6$	31 ± 8

Tabella 2.6: Parte 1: *Valori dei parametri estratti dal fit con l'eq. 1.3 fatto sulle distribuzioni delle interdistanze del dinucleotide TA in scala semilogaritmica. Il parametro a é adimensionale. Gli errori sui parametri dei fit sono stimati ad un intervallo di confidenza del 95 %.*

Chr.	a	b (bp)	c (bp^a)	d (bp)
13	$4,7 \pm 0,5$	$(0,003 \pm 1) \times 10^8$	$(0,9 \pm 3) \times 10^6$	25 ± 7
14	$4,8 \pm 0,6$	$(0,005 \pm 3) \times 10^8$	$(0,4 \pm 1) \times 10^7$	35 ± 9
15	$4,9 \pm 0,8$	$(0,8 \pm 1) \times 10^3$	$(1 \pm 5) \times 10^7$	41 ± 12
16	$\mathbf{4,6 \pm 0,5}$	$\mathbf{(0,6 \pm 4) \times 10^4}$	$\mathbf{(2 \pm 7) \times 10^6}$	$\mathbf{40 \pm 10}$
17	$\mathbf{4,5 \pm 0,6}$	$\mathbf{(0,9 \pm 1) \times 10^3}$	$\mathbf{(3 \pm 9) \times 10^6}$	$\mathbf{45 \pm 12}$
18	$4,7 \pm 0,6$	$(0,006 \pm 5) \times 10^8$	$(1 \pm 3) \times 10^6$	27 ± 8
19	$\mathbf{4,0 \pm 0,6}$	$\mathbf{(6 \pm 4) \times 10^2}$	$\mathbf{(2 \pm 8) \times 10^5}$	$\mathbf{42 \pm 13}$
20	$\mathbf{4,7 \pm 0,6}$	$\mathbf{(2 \pm 6) \times 10^3}$	$\mathbf{(0,4 \pm 2) \times 10^7}$	$\mathbf{42 \pm 11}$
21	$4,1 \pm 0,5$	$(0,004 \pm 2) \times 10^8$	$(1 \pm 3) \times 10^5$	25 ± 7
22	$\mathbf{3,7 \pm 0,6}$	$\mathbf{(3 \pm 1) \times 10^2}$	$\mathbf{(0,8 \pm 2) \times 10^5}$	$\mathbf{40 \pm 13}$
X	$4,9 \pm 0,6$	$(0,003 \pm 2) \times 10^8$	$(2 \pm 6) \times 10^6$	26 ± 8
Y	$\mathbf{4,2 \pm 0,7}$	$\mathbf{(0,003 \pm 2) \times 10^8}$	$\mathbf{(0,009 \pm 2) \times 10^5}$	$\mathbf{20 \pm 8}$

Tabella 2.7: Parte 2: *Valori dei parametri estratti dal fit con l'eq. 1.3 fatto sulle distribuzioni delle interdistanze del dinucleotide TA in scala semilogaritmica. Il parametro a é adimensionale. Gli errori sui parametri dei fit sono stimati ad un intervallo di confidenza del 95 %.*

Sono riportati ora i grafici contenenti le distribuzioni delle interdistanze del dinucleotide TA, in scala semilogaritmica, fittate con l'eq. 1.5. I parametri a, c, d (vedere eq.1.5) insieme ai rispettivi intervalli di confidenza, ottenuti mediante le funzioni MATLAB *fit* e *confint*, sono riportati nella tabella (vedere Tab. 2.8).

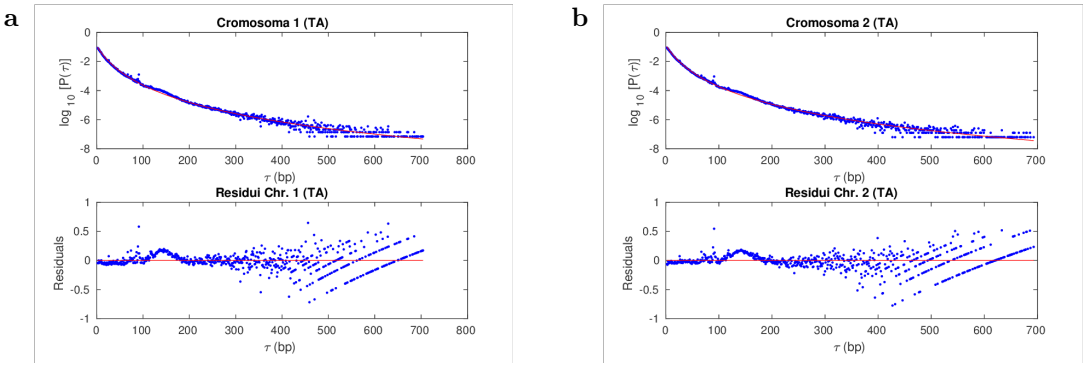

Figura 2.20: Parte 1. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide TA fittate con l'eq. 1.5, insieme ai rispettivi grafici dei residui.*

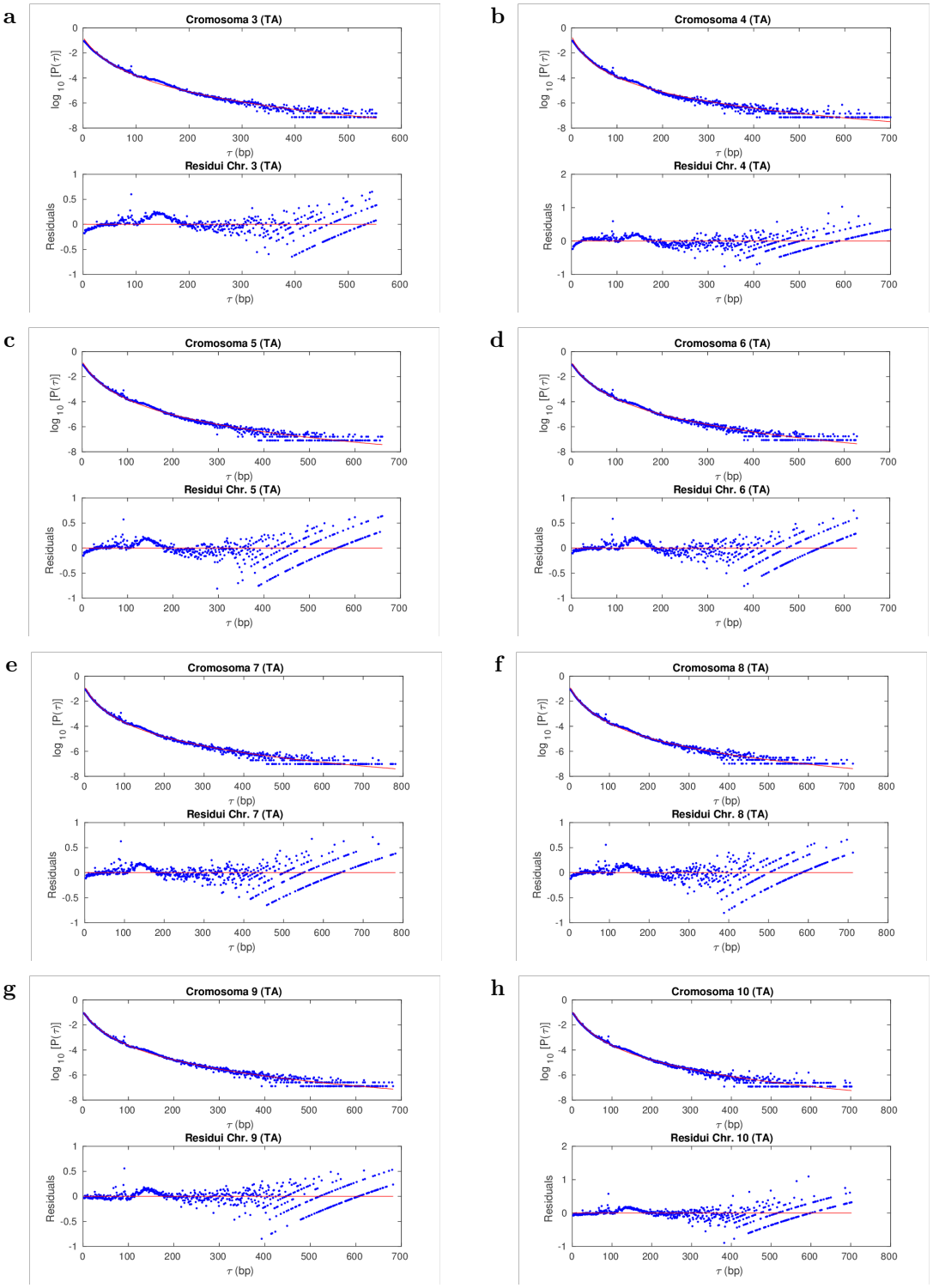

Figura 2.21: Parte 2. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide TA fittate con l'eq. 1.5, insieme ai rispettivi grafici dei residui.*

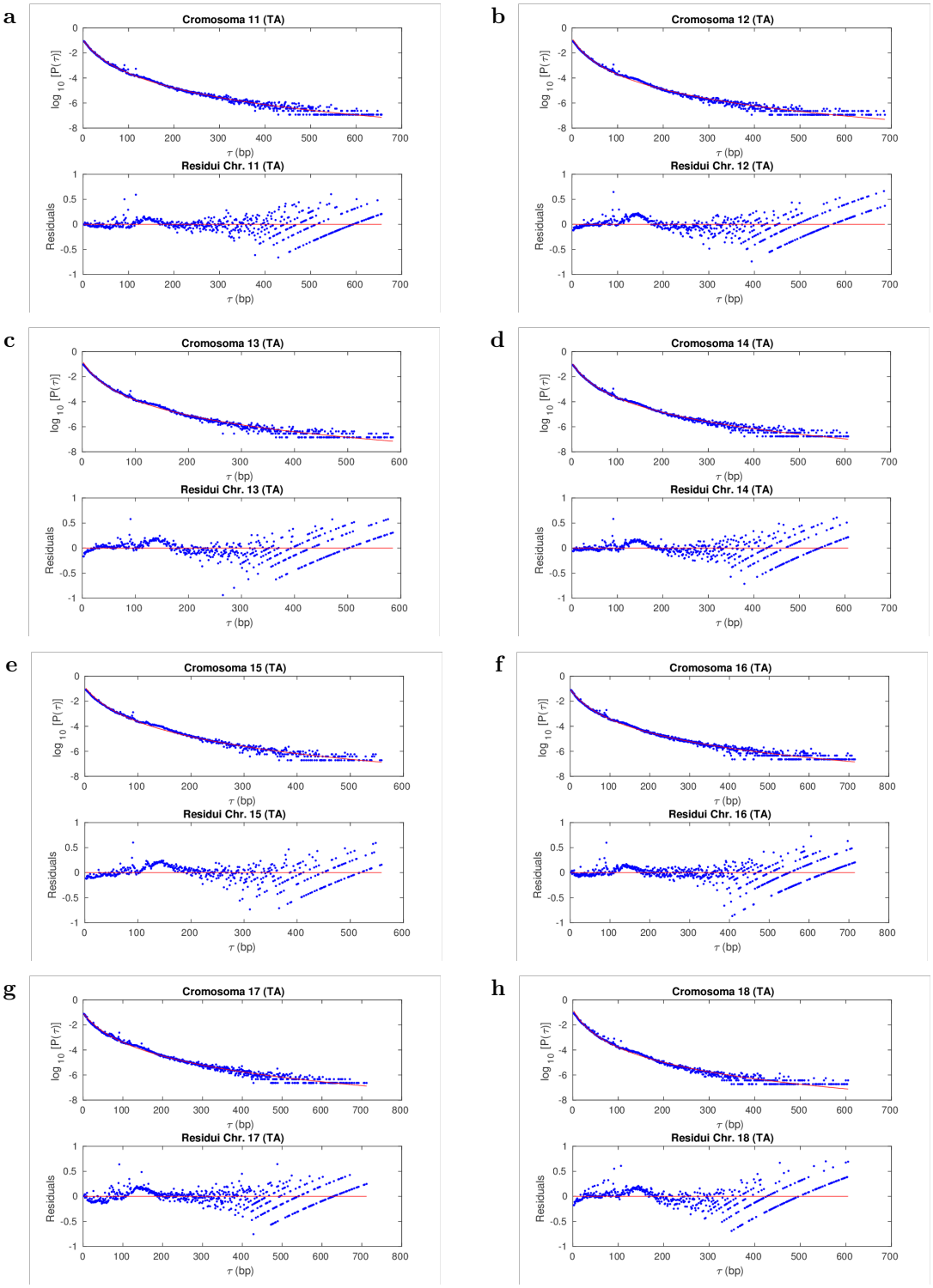

Figura 2.22: Parte 3. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide TA fittate con l'eq. 1.5, insieme ai rispettivi grafici dei residui.*

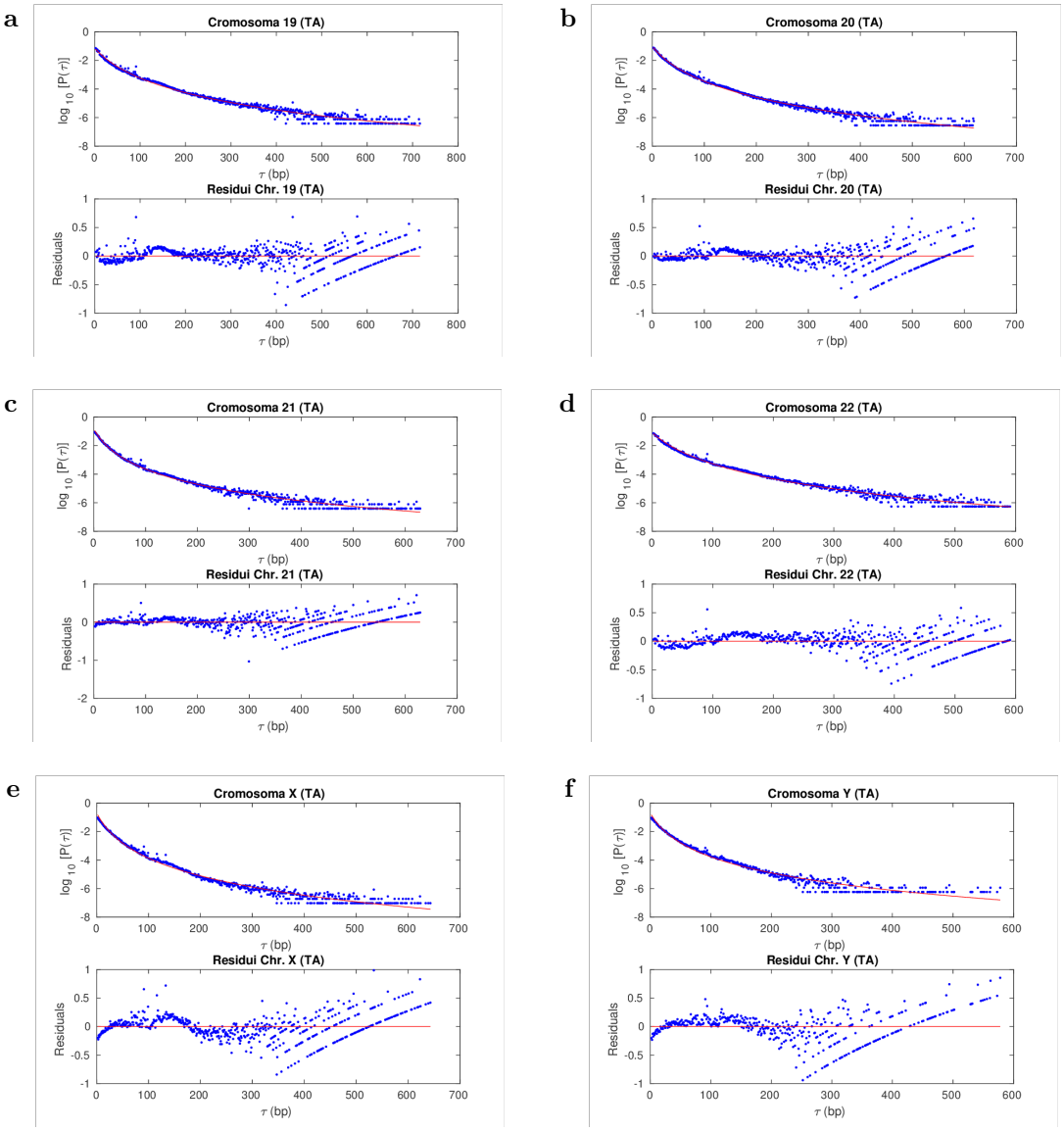

Figura 2.23: Parte 4. *Distribuzioni, in scala semilogaritmica rispetto all'asse delle Y, delle interdistanze del dinucleotide TA fittate con l'eq. 1.5, insieme ai rispettivi grafici dei residui.*

Chr.	**a**	**c (bp[a])**	**d (bp)**
1	$5,2 \pm 0,1$	$(3 \pm 2) \times 10^7$	43 ± 4
2	$5,1 \pm 0,1$	$(2 \pm 1) \times 10^7$	39 ± 4
3	$5,3 \pm 0,1$	$(3 \pm 3) \times 10^7$	36 ± 4
4	$4,6 \pm 0,1$	$(6 \pm 4) \times 10^5$	24 ± 3
5	$5,1 \pm 0,1$	$(8 \pm 7) \times 10^6$	33 ± 4
6	$5,2 \pm 0,1$	$(2 \pm 2) \times 10^7$	37 ± 4
7	$4,7 \pm 0,1$	$(2 \pm 1) \times 10^6$	32 ± 4
8	$4,9 \pm 0,1$	$(4 \pm 3) \times 10^6$	33 ± 4
9	$4,9 \pm 0,1$	$(6 \pm 5) \times 10^6$	38 ± 4
10	$5,0 \pm 0,1$	$(9 \pm 9) \times 10^6$	39 ± 5
11	$5,1 \pm 0,1$	$(2 \pm 2) \times 10^7$	43 ± 4
12	$4,9 \pm 0,1$	$(6 \pm 5) \times 10^6$	35 ± 4
13	$4,8 \pm 0,1$	$(2 \pm 2) \times 10^6$	29 ± 4
14	$4,9 \pm 0,1$	$(7 \pm 6) \times 10^6$	37 ± 4
15	$5,2 \pm 0,2$	$(3 \pm 3) \times 10^7$	41 ± 5
16	$\mathbf{4,8 \pm 0,1}$	$\mathbf{(7 \pm 6) \times 10^6}$	$\mathbf{45 \pm 5}$
17	$\mathbf{5,0 \pm 0,1}$	$\mathbf{(3 \pm 2) \times 10^7}$	$\mathbf{51 \pm 5}$
18	$4,8 \pm 0,1$	$(3 \pm 3) \times 10^6$	31 ± 4
19	$\mathbf{4,9 \pm 0,2}$	$\mathbf{(3 \pm 3) \times 10^7}$	$\mathbf{59 \pm 7}$
20	$\mathbf{5,0 \pm 0,1}$	$\mathbf{(3 \pm 2) \times 10^7}$	$\mathbf{48 \pm 5}$
21	$4,2 \pm 0,1$	$(2 \pm 1) \times 10^5$	27 ± 4
22	$\mathbf{4,9 \pm 0,2}$	$\mathbf{(2 \pm 3) \times 10^7}$	$\mathbf{55 \pm 6}$
X	$5,0 \pm 0,2$	$(6 \pm 6) \times 10^6$	30 ± 4
Y	$\mathbf{4,4 \pm 0,2}$	$\mathbf{(3 \pm 4) \times 10^5}$	$\mathbf{25 \pm 5}$

Tabella 2.8: *Valori dei parametri estratti dal fit con l'eq. 1.5 fatto sulle distribuzioni delle interdistanze del dinucleotide TA in scala semilogaritmica. Il parametro a é adimensionale. Gli errori sui parametri dei fit sono stimati ad un intervallo di confidenza del 95 % e arrotondati alla prima cifra significativa.*

2.4 Bontá dei fit

In questo paragrafo sono riportati, nelle apposite tabelle, i valori del R-quadro aggiustato, del SSE e del RMSE relativi ai fit fatti alle distribuzioni delle interdistanze dei dinucleotidi CG e TA. Nella prima sezione sono riportati i valori che si riferiscono alla distribuzione delle CG fittata con l' eq. 1.3, mentre nella seconda i valori che si riferiscono alla distribuzione delle TA fittata prima con l'eq. 1.3 e poi con l'eq. 1.5.

2.4.1 Dinucleotide CG

I grafici dei residui delle distribuzioni delle CG (vedere Fig. 2.8, 2.9, 2.10 e 2.11) mostrano un comportamento analogo tra i vari cromosomi umani, con una piccola deviazione sistematica in corrispondenza di valori di interdistanza bassi. Si nota inoltre un aumento del rumore nelle code delle distribuzioni, che peró non inficia una corretta stima dei valori dei parametri [11].

Ad eccezione dei cromosomi 19 e 22 (in particolar modo per il parametro b) l'incertezza con cui si individua i parametri non é molto grande, pertanto le ampiezze degli intervalli di confidenza sono abbastanza limitate. Escludendo il parametro SSE, i cui valori sono leggermente alti se confrontati con quelli ottenuti nei fit del dinucleotide TA, i parametri R-quadro aggiustato e RMSE assumono valori molto vicini ai limiti ideali.

Chr.	**R^2 a.**	**SSE (bp^2)**	**RMSE(bp)**	**Chr.**	**R^2 a.**	**SSE (bp^2)**	**RMSE(bp)**
1	0,9855	34,1812	0,1381	13	0,9811	32,7443	0,1410
2	0,9860	32,3350	0,1334	14	0,9812	32,0930	0,1423
3	0,9854	32,4511	0,1339	15	0,9806	31,6047	0,1483
4	0,9832	37,4389	0,1408	**16**	**0,9805**	**30,7237**	**0,1525**
5	0,9826	37,2682	0,1443	**17**	**0,9817**	**28,2647**	**0,1485**
6	0,9845	32,2655	0,1348	18	0,9794	32,3555	0,1463
7	0,9824	36,1905	0,1456	**19**	**0,9821**	**26,5734**	**0,1495**
8	0,9855	28,5058	0,1295	**20**	**0,9802**	**29,1422**	**0,1479**
9	0,9814	35,6097	0,1477	21	0,9753	30,7656	0,1525
10	0,9843	29,7961	0,1367	**22**	**0,9789**	**26,3092**	**0,1559**
11	0,9821	34,8528	0,1443	X	0,9835	34,5292	0,1385
12	0,9844	30,0610	0,1342	**Y**	**0,9633**	**40,0605**	**0,1691**

Tabella 2.9: *Parametri ottenuti dal fit con l'eq. 1.3 delle distribuzioni delle interdistanze, in scala semilogaritmica rispetto all'asse delle Y, del dinucleotide CG per tutti i cromosomi umani.*

2.4.2 Dinucleotide TA

Chr.	R^2 a.	SSE(bp^2)	RMSE(bp)	Chr.	R^2 a.	SSE(bp^2)	RMSE(bp)
1	0,9884	17,15	0,1645	13	0,9830	19,15	0,1948
2	0,9885	17,64	0,1668	14	0,9865	15,05	0,1683
3	0,9871	16,25	0,1776	15	0,9848	15,92	0,1804
4	0,9834	23,13	0,1973	**16**	**0,9842**	**19,51**	**0,1747**
5	0,9833	23,30	0,2008	**17**	**0,9857**	**17,67**	**0,1690**
6	0,9860	18,55	0,1828	18	0,9814	20,93	0,2020
7	0,9851	20,99	0,1814	**19**	**0,9822**	**19,85**	**0,1771**
8	0,9852	19,72	0,1831	**20**	**0,9851**	**16,29**	**0,1706**
9	0,9873	16,47	0,1649	21	0,9805	18,77	0,1869
10	0,9820	20,70	0,1996	**22**	**0,9846**	**14,27**	**0,1615**
11	0,9883	15,00	0,1600	X	0,9793	27,55	0,2242
12	0,9847	20,03	0,1858	**Y**	**0,9674**	**26,37**	**0,2512**

Tabella 2.10: *Parametri ottenuti dal fit con l'eq.1.3 delle distribuzioni delle interdistanze, in scala semilogaritmica rispetto l'asse delle Y, del dinucleotide TA per tutti i cromosomi umani.*

Chr.	R^2 a.	SSE(bp^2)	RMSE(bp)	Chr.	R^2 a.	SSE(bp^2)	RMSE(bp)
1	0,9886	16,87	0,1630	13	0,9831	18,99	0,1938
2	0,9885	17,57	0,1663	14	0,9865	15,01	0,1680
3	0,9871	16,33	0,1779	15	0,9847	16,01	0,1807
4	0,9836	22,88	0,1961	**16**	**0,9845**	**19,26**	**0,1735**
5	0,9834	23,13	0,1999	**17**	**0,9861**	**17,22**	**0,1667**
6	0,9861	18,47	0,1823	18	0,9816	20,72	0,2008
7	0,9852	20,86	0,1807	**19**	**0,9831**	**18,91**	**0,1727**
8	0,9853	19,62	0,1825	**20**	**0,9854**	**15,93**	**0,1685**
9	0,9874	16,39	0,1643	21	0,9806	18,70	0,1864
10	0,9821	23,58	0,1989	**22**	**0,9844**	**14,49**	**0,1626**
11	0,9885	14,75	0,1585	X	0,9796	27,26	0,2228
12	0,9849	19,86	0,1849	**Y**	**0,9680**	**25,98**	**0,2490**

Tabella 2.11: *Parametri ottenuti dal fit con l'eq.1.5 delle distribuzioni delle interdistanze, in scala semilogaritmica rispetto l'asse delle Y, del dinucleotide TA per tutti i cromosomi umani.*

Per facilitare il confronto dei parametri ottenuti dai due diversi fit fatti sulla distribuzione delle TA sono stati realizzati dei box plot (vedere Fig. 2.24, 2.25 e 2.26).

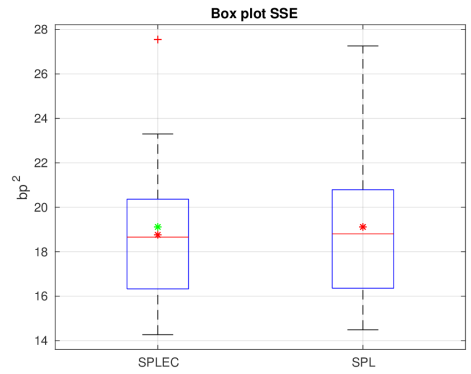

Figura 2.24: *Box plot dei valori di SSE riportati nelle tabelle 2.10 e 2.11. Le stelle rosse, presenti nei due box, indicano rispettivamente le medie dei valori utilizzati. La stella verde, invece, rappresenta la proiezione della media del Box di destra sul box di sinistra, facilitando il confronto tra i due valori. La sigla SPLEC si riferisce all'eq. 1.3 e la sigla SPL all'eq. 1.5.*

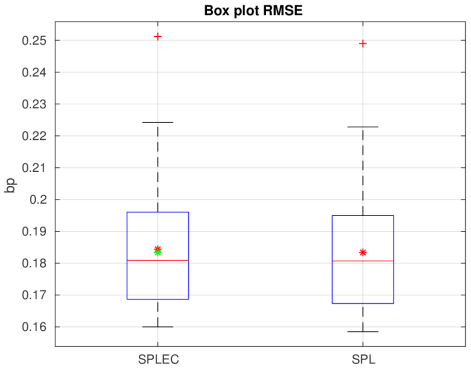

Figura 2.25: *Box plot dei valori di RMSE riportati nelle tabelle 2.10 e 2.11. Le stelle rosse, presenti nei due box, indicano rispettivamente le medie dei valori utilizzati. La stella verde, invece, rappresenta la proiezione della media del Box di destra sul box di sinistra, facilitando il confronto tra i due valori. La sigla SPLEC si riferisce all'eq. 1.3 e la sigla SPL all'eq. 1.5.*

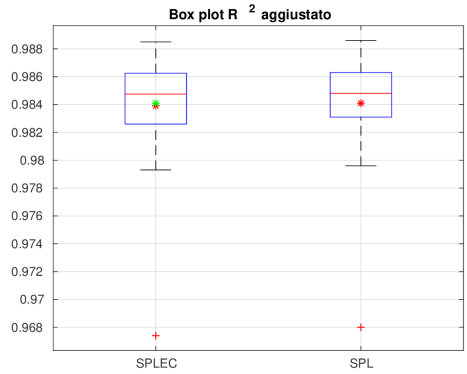

Figura 2.26: *Box plot dei valori di R^2 aggiustato riportati nelle tabelle 2.10 e 2.11. Le stelle rosse, presenti nei due box, indicano rispettivamente le medie dei valori utilizzati. La stella verde, invece, rappresenta la proiezione della media del Box di destra sul box di sinistra, facilitando il confronto tra i due valori. La sigla SPLEC si riferisce all'eq. 1.3 e la sigla SPL all'eq. 1.5.*

Osservando i valori medi dei vari parametri riportati nelle figure 2.24, 2.25 e 2.26 e confrontando valore per valore i parametri riportati nelle tabelle 2.10 e 2.11 si puó osservare che:

- I valori di SSE ottenuti dal fit della distribuzione del dinucleotide TA con l'eq. 1.3 sono migliori di quelli ottenuti con l'eq. 1.5 in quanto piú vicini allo zero;
- I valori di RMSE ottenuti dal fit con l'eq. 1.5 sono migliori di quelli ottenuti dal fit con l'eq. 1.3 per lo stesso motivo del punto precedente;
- I valori di R^2 aggiustato ottenuti dal fit con l'eq. 1.5 sono migliori di quelli ottenuti con l'eq. 1.3, in quanto piú vicini ad uno.

L'analisi dei grafici dei residui non fornisce alcuna informazione utile riguardo a quale delle due funzioni descriva meglio i dati, in quanto i residui dovuti al fit con l'eq. 1.3 sono pressocché uguali a quelli ottenuti dal fit con l'eq. 1.5. Le ampiezze degli intervalli di confidenza dei parametri ottenuti dal fit con l'eq. 1.5 sono leggermente piú piccoli rispetto a quelli ottenuti dal fit con l'eq. 1.5.
Pertanto si puó concludere, rimanendo nell'ambito dei cromosomi umani, che l'eq. 1.5 descrive piú accuratamente rispetto all'eq. 1.3 il comportamento delle distribuzioni delle interdistanze, in scala semilogaritmica, del dinucleotide TA.

Capitolo 3

Conclusioni

Dopo aver analizzato la qualitá dei dati utilizzati attraverso le analisi condotte sui nucleotidi non correttamente sequenziati (N) nei cromosomi umani, descritte ampiamente nei paragrafi 1.2 e 2.1, si puó concludere che l'effetto dovuto alla rimozione delle N dalle sequenze di nucleotidi é statisticamente trascurabile. Piú nello specifico, in un primo studio si é osservato che la percentuale di occorrenze delle N lungo la sequenza di nucleotidi varia in maniera evidente da cromosoma a cromosoma, arrivando a raggiungere anche valori particolarmente elevati, come nel caso dei cromosomi 22 e Y. Tuttavia si é visto che questi nucleotidi si dispongono generalmente in grossi blocchi localizzati agli estremi e nella parte centrale dei cromosomi. In una seconda e piú attenta analisi condotta cromosoma per cromosoma, al fine di quantificare il numero di interdistanze di dinucleotidi influenzate dalla presenza di queste N, sono state individuate, contate e misurate (in termini di numero di basi) tutte le sequenze costituite da N ripetute consecutivamente. Si é osservato che il numero di queste sequenze é estremamente basso (generalmente al di sotto del centinaio di sequenze) rapportato al numero di interdistanze considerate ogni volta per i vari cromosomi (dell'ordine di qualche milione) e che generalmente la lunghezza di queste sequenze é superiore al migliaio di basi. Quello che emerge é che le N si dispongono tendenzialmente in lunghe sequenze, influenzando cosí un numero di interdistanze basso e quindi trascurabile.

La seconda parte delle analisi é dedicata allo studio delle distribuzioni delle interdistanze del dinucleotide CG all'interno dei cromosomi umani. Tali distribuzioni, dopo essere state private delle code, in quanto molto rumorose, sono state fittate attraverso delle leggi di potenza traslate con cutoff esponenziale (vedere eq. 1.2). Lo studio dei grafici dei residui, unito ai risultati forniti dai parametri utilizzati per valutare la bontá del fit, permettono di concludere che la legge di potenza traslata con cutoff esponenziale, descrive in modo accurato le distribuzioni delle interdistanze del dinucleotide CG per tutti i cromosomi umani. Tuttavia si osserva che i parametri dell'eq. 1.3 restituiti dai fit, non assumono sempre valori simili. Piú nello specifico si é visto che la maggior parte dei cromosomi é caratterizzata da parametri i cui valori differiscono di poco. Vi sono

peró cromosomi i cui parametri assumono valori anomali, come nel caso dei cromosomi 16, 17, 19, 20, 22 e Y. Particolare attenzione é stata riposta sul parametro b dell'equazione 1.3 (termine di cutoff esponenziale). Per motivare i valori decisamente alti di questo parametro nei cromosomi 16, 19, 22 e Y si é ipotizzato che l'aumento dei valori di b fosse dovuto ad una minore presenza, in termine di occorrenze, del dinucleotide CG lungo le sequenze di DNA. Grazie ad uno studio volto ad individuare le percentuali dei nucleotidi C e G e del dinucleotide CG all'interno delle sequenze di nucleotidi che costituiscono i vari cromosomi non é stato osservato nessun andamento di tale tipo.

L'ultima sezione della tesi si é occupata invece di studiare le distribuzioni delle interdistanze del dinucleotide TA per tutti i cromosomi umani. In una prima sezione sono stati plottati i grafici delle distribuzioni delle interdistanze di tutti i 16 dinucleotidi del primo cromosoma umano, fittate con la legge di potenza traslata con cutoff esponenziale, e i loro rispettivi grafici dei residui. Si é osservato che il dinucleotide TA ha un comportamento molto simile al dinucleotide CG. Questo fatto, evidenziato in particolare dall'analisi dei grafici dei residui, ha suggerito lo studio di questo dinucleotide in particolare. In una seconda sezione,sono state fittate con la legge di potenza traslata con cutoff esponenziale (vedere eq. 1.2) e con una legge di potenza traslata senza cutoff esponenziale (vedere eq. 1.4) tutte le 24 distribuzioni di TA. Grazie all'analisi dei parametri utilizzati per valutare la bontá dei fit é stato possibile concludere che la legge di potenza traslata senza cutoff esponenziale, rappresentata dall'eq. 1.4, descrive il comportamento delle distribuzioni delle interdistanze di TA in modo migliore. Inoltre, a differenza del dinucleotide CG, i cromosomi 16, 17, 19, 20, 22 e Y non presentano parametri con valori anomali fatto che suggerisce che questa irregolaritá non dipenda dai cromosomi bensí dal dinucleotide CG stesso.

Bibliografia

[1] D. Anastassiou (2001). *Genomic signal processing*. IEEE, 18, 8-20.

[2] A. K. Brodzik, O. Peters (2005). *Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences*. IEEE, 5, 373-376.

[3] T. Mahalakshmi, Achuth Nair (2005). *Visualization Of Genomic Data Using Inter-Nucleotide Distance Signals*. Proceedings of IEEE Genomic Signal Processing, 408.

[4] Afreixo V., Bastos C. A., Pinho A. J. et all. (2009). *Genome Analysis with inter-nucleotide distances*. Bioinformatics, 25, 23, 3064-3070.

[5] Bastos C. A., Afreixo V., Pinho A. J., et all. (2011). *Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions*. Journal of Integrative Bioinformatics, 8, 31-42.

[6] Afreixo V., Bastos C. A., Rodrigues J.M and Silva R. (2015). *Identification of DNA CpG Islands Using Inter-dinucleotide Distances*. Optimization in the Natural Sciences, 162-172.

[7] Moghaddasi H., Khalifeh K., Darooneh A. H. (2017). *Distinguishing Functional DNA Words; A Method for Measuring Clustering Levels*. Sci Rep, 7, 41543.

[8] Paci G. (2014). *Statistical methods for the analysis of DNA sequences: appli- cation to dinucleotide distribution in the human genome*. Tesi di laurea magistrale in Fisica, Universit di Bologna, Relatore: Remondini D.; Correlatore: Cristadoro G.

[9] Paci G., Cristadoro G., Monti B., Lenci M., Degli Esposti M., Castellani G. and Remondini D. (2016). *Characterization of DNA methylation as a function of biological complexity via dinucleotide inter-distances*. Philosophical transactions A, 374.

[10] Merlotti A. (2016). *DNA sequence analysis: a statistical characterization of di-nucleotides interdistances across multiple organisms.* Tesi di laurea magistrale in Fisica, Universit di Bologna, Relatore: Remondini D.; Correlatore: talo Faria do Valle

[11] Merlotti A., Faria do Valle I., Castellani G. and Remondini D. (2018). *Statistical modelling of CG interdistance across multiple organisms.* BMC Bioinformatics, 19, 355

[12] Adrian P. Bird (1986). *CpG-rich islands and the function of DNA methylation.* Nature, 321, 209-213.

[13] Bock C., Walter J., Paulsen M., Lengauer T. (2007). *CpG island mapping by epigenome prediction.* PLoS Comput. Biol., 3, e110.