Alma Mater Studiorum
Universitá degli Studi di Bologna

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

Modelli statistici per l'organizzazione della struttura tridimensionale del DNA umano

Relatore:
Prof. Gastone Castellani

Tesi di Laurea di:
Francesca Mugianesi

Correlatore:
Prof. Daniel Remondini

16 Dicembre 2016
Anno Accademico 2016/2017

Abstract

Il grande sviluppo della tecnica di genome-wide conformation capture (Hi-C) permette di indagare la complessa ed interessante relazione che intercorre tra la struttura 3D dinamica organizzata gerarchicamente della cromatina e la funzionalità del genoma. In altre parole, questa tecnica consente di avere informazioni tridimensionali sulla struttura dei nuclei delle cellule. Il lavoro di tesi si è incentrato sullo studio multirisoluzione dei domini topologici (TAD) della cromatina di sette tipi cellulari umani. L'algoritmo impiegato è basato sulla segmentazione spettrale iterativa del laplaciano normalizzato associato alla mappa intra-cromosomiale Hi-C. L'analisi dei dati ha rivelato che i TAD boundary tendono ad una distribuzione spaziale regolare, conservata tra tipi cellulari. É possibile che il maggior grado di similarità riscontrato tra alcune linee cellulari abbia basi biologicamente rilevanti. Le dimensioni dei TAD individuati vanno da $\sim$450 kb, con i dati alla risoluzione di 50 kb, fino a $\sim$4.7 Mb, alla risoluzione di 1 Mb, e risultano indipendenti dalla lunghezza specifica del cromosoma.

Indice

Introduzione

La quantità di dati generati tramite la tecnica di *genome-wide chromosome conformation capture* è in rapida crescita e presenta grandi opportunità e sfide volte alla comprensione della struttura del genoma.

Il ruolo dell'organizzazione dell'architettura 3D del genoma umano nella funzionalità dei geni è globalmente riconosciuto dalla comunità scientifica. Infatti, la conformazione 3D della cromatina permette l'avvicinamento spaziale di elementi funzionali che risultano distali nella catena lineare monodimensionale del genoma ed ha quindi un impatto decisivo nella regolazione genica.

La comprensione di come la cromatina sia organizzata alle diverse scale spaziali, dalla scala dei cromosomi a quella della catena di DNA, contribuisce a fare luce sulla complessa relazione che vi è tra la struttura della cromatina, l'attività genica e lo stato funzionale della cellula. Tuttavia, vi sono molte domande fondamentali che al momento non hanno ricevuto risposta, ad esempio in che modo elementi regolatori distali, come gli enhancer, agiscano sui promoter e come i repressori ostacolino simili processi. Si ritiene che questo tipo di fenomeni coinvolga la formazione di loop, mediati da proteine, che avvicinano spazialmente coppie di siti genomici lontani tra loro nella catena cromatinica lineare [1].

Le tecnologie disponibili per indagare l'architettura 3D della cromatina sono diverse e possono operare al livello del singolo locus genico (3C, 4C) [2], di un gruppo di loci (5C, ChIA-PET) [3].

La tecnica Hi-C coinvolge il sequenziamento del genoma e permette la creazione di una libreria della distanza 3D tra tutte le possibile coppie di loci genici.

Le metodologie e gli strumenti per la produzione e l'analisi dei dati Hi-C sono numerosi ed in continua evoluzione, poiché si tratta di un settore in enorme sviluppo. In questo modo, la struttura della cromatina viene integrata a dati di regolazione genica, di alterazioni genetiche ed altro, rivelando caratteristiche strutturali di base che sono considerate i principi organizzativi del folding della cromatina, gettando le basi per applicazioni mediche e farmacologiche estremamente innovative [4].

Capitolo 1

DNA cromosomiale ed il suo packaging nella fibra di cromatina

La funzione più importante del DNA è rappresentata dai geni, le unità funzionali per la produzione di proteine ed RNA. Essi trasportano l'informazione relativa a tutte le proteine che costituiscono un organismo: quando, in quali tipi di cellule ed in quale quantità ogni proteina deve essere prodotta. Un *gene* è un segmento di DNA che contiene le istruzioni per produrre una particolare proteina (o un insieme di proteine strettamente correlate tra loro) e nel genoma umano ve ne sono circa 20,000-25,000.

Il genoma degli eucarioti è diviso in cromosomi ed il suo packaging è uno degli aspetti più complessi e misteriosi della biologia. Basti pensare che ogni cellula umana contiene circa 2 m di DNA, se considerato "srotolato", ed il nucleo di una cellula umana, che contiene il DNA, ha un diametro di soli ~ 6 μm. Ciò sarebbe equivalente ad arrotolare 40 km di corda estremamente sottile in una pallina da tennis!

Il complesso processo del folding del DNA viene mediato da proteine specializzate che legano e piegano il DNA, generando stringhe di solenoidi e loop che costituiscono livelli di organizzazione sempre maggiore, evitando che il DNA si trasformi in un groviglio inutilizzabile (figura 1.1). Infatti, il DNA finemente "impacchettato" risulta facilmente accessibile ai diversi elementi che partecipano alla sua replicazione e riparazione e che utilizzano i suoi geni per la sintesi delle proteine [5].

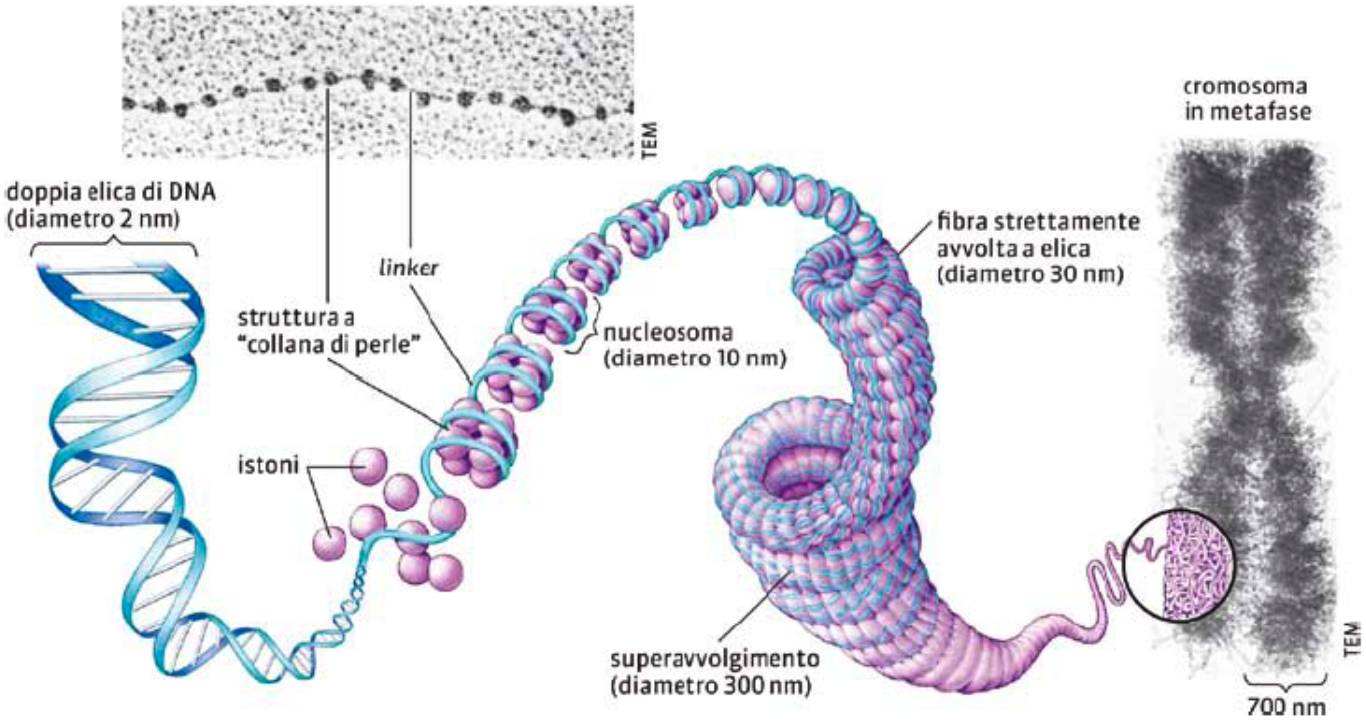

Figura 1.1: Rappresentazione schematica del folding del DNA, dalla scala della catena lineare a quella dei cromosomi.

1.1 Struttura chimica di base Processi fondamentali

Dal punto di vista chimico, l'acido desossiribonucleico (DNA) è un polimero organico i cui monomeri sono chiamati *nucleotidi*. Tutti i nucleotidi hanno tre componenti fondamentali: un gruppo fosfato, il deossiribosio (zucchero pentoso) ed una base azotata. Le basi azotate che possono essere utilizzate nella formazione dei nucleotidi sono quattro: adenina, guanina, citosina e timina. Nell'RNA, al posto della timina si trova l'uracile.

Il DNA è quindi una doppia catena polinucleotidica, complementare, orientata, spiralizzata ed informazionale. L'informazione genetica risiede nell'ordine della disposizione sequenziale dei nucleotidi e dal codice genetico viene tradotta nei corrispondenti amminoacidi, i quali formano le proteine.

La traduzione genetica (*sintesi proteica*) viene mediata dall'RNA, molecola che viene generata per complementarità a partire dalle basi azotate dei nucleotidi del DNA in un processo definito come *trascrizione*.

Al momento della divisione cellulare, l'informazione genetica viene duplicata (*replicazione del DNA*), allo scopo di trasmettere l'informazione genetica alle generazioni cellulari successive. [3].

1.2 Organizzazione in cromosomi

Negli eucarioti, il DNA nucleare è diviso in un certo numero di *cromosomi*.

Il genoma umano, costituito da circa 3.2 x 10^9 nucleotidi, è distribuito in 24 cromosomi diversi. Ogni cromosoma consiste di una molecola lineare di DNA estremamente lunga, associata a proteine che piegano ed impacchettano la sottile catena di DNA in una struttura più compatta. Il complesso di DNA e proteine ha il nome di *cromatina* (dal greco *chroma*, "colore', per le sue proprietà cromatiche). Oltre alle proteine necessarie al folding del DNA, i cromosomi sono associati a molte proteine richieste per lo svolgimento dei processi di espressione genica, replicazione e riparazione del DNA.

I batteri possiedono una singola molecola di DNA, generalmente circolare. Anch'essa è associata a proteine che condensano il DNA, ma sono diverse rispetto alle proteine che svolgono questa funzione negli eucarioti.

Ad eccezione delle cellule germinali e di pochi altri tipi cellulari altamente specializzati che non possono riprodursi e che mancano del tutto del DNA (ad esempio, i globuli rossi), ogni cellula umana contiene due copie di ciascun cromosoma, uno ereditato dalla madre e l'altro dal padre (*cromosomi omologhi*). Gli unici cromosomi non omologhi sono i cromosomi sessuali nel maschio, costituiti da un cromosoma Y ereditato dal padre ed un cromosoma X proveniente dalla madre. Pertanto, ogni cellula umana contiene un totale di 46 cromosomi, 22 coppie comuni al maschio e alla femmina ed una coppia di cromosomi sessuali (X e Y nei maschi, X e X nelle femmine).

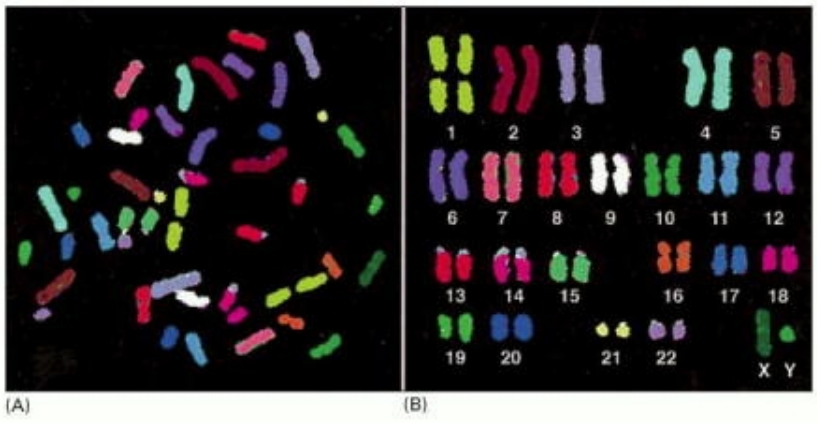

Figura 1.2: Cromosomi umani maschili, isolati da una cellula in procinto di divisione nucleare (mitosi) e perciò molto compatti. Ogni cromosoma è stato evidenziato con coloranti fluorescenti, per permetterne la visualizzazione al microscopio ottico. (A) Cromosomi visualizzati nella configurazione naturale. (B) Cromosomi artificialmente allineati in ordine numerico (cariotipo).

É possibile distinguere i cromosomi umani colorando ognuno con un differente colore (figura 1.2); questa tecnica viene generalmente effettuata nella fase della *mitosi*, uno stage del ciclo cellulare in cui i cromosomi sono particolarmente compatti e facilmente visualizzabili. L'insieme dei 46 cromosomi umani nella fase di mitosi viene chiamato *cariotipo*.

1.3 Mappatura del genoma umano Scoperta della squenza nucleotidica

Il *Progetto Genoma Umano* (HGP, acronimo di Human Genome Project) è stato un progetto di ricerca scientifica internazionale iniziato nel 1990 e completato nel 2003, il cui obiettivo principale è stato quello di determinare la sequenza delle coppie di basi azotate che formano il DNA e di identificare e mappare i geni del genoma umano, dal punto di vista sia fisico sia funzionale. Di questi ne furono previsti circa 200,000 ma ne sono stati trovati 20,000-25,000, di dimensioni medie di 27,000 coppie di nucleotidi (27 kb). Il Progetto Genoma Umano ha inoltre scoperto che soltanto l'1.5% circa della lunghezza totale del DNA si basa su *esoni* (porzioni di geni codificanti proteine) e che il restante 98.5% corrisponde a DNA non codificante.

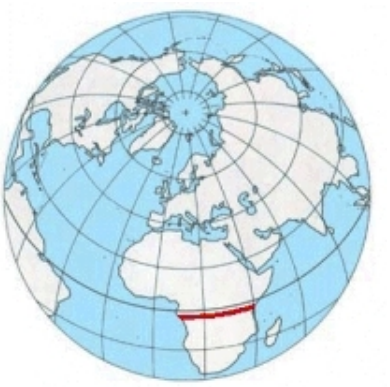

Figura 1.3: Scala del genoma umano. Se ogni coppia di nucleotidi fosse rappresentata delle dimensioni di 1 mm, il genoma umano avrebbe un'estensione di 3,200 km (ampiezza longitudinale dell'Africa centrale).

Se ogni coppia di nucleotidi avesse un'estensione di 1 mm, allora il genoma umano sarebbe lungo 3,200 km, coprendo l'ampiezza longitudinale dell'Africa centrale (figura 1.3). A tale scala, vi sarebbe in media un gene codificante proteine ogni 300 m; un gene medio avrebbe un'estensione di 30 m, ma le sue sequenze codificanti corrisponderebbero soltanto ad 1 m.

	CHROMOSOME 22	HUMAN GENOME
DNA length	48×10^6 nucleotide pairs*	3.2×10^9
Number of genes	approximately 700	approximately 30,000
Smallest protein-coding gene	1000 nucleotide pairs	not analyzed
Largest gene	583,000 nucleotide pairs	2.4×10^6 nucleotide pairs
Mean gene size	19,000 nucleotide pairs	27,000 nucleotide pairs
Smallest number of exons per gene	1	1
Largest number of exons per gene	54	178
Mean number of exons per gene	5.4	8.8
Smallest exon size	8 nucleotide pairs	not analyzed
Largest exon size	7600 nucleotide pairs	17,106 nucleotide pairs
Mean exon size	266 nucleotide pairs	145 nucleotide pairs
Number of pseudogenes**	more than 134	not analyzed
Percentage of DNA sequence in exons (protein coding sequences)	3%	1.5%
Percentage of DNA in high-copy repetitive elements	42%	approximately 50%
Percentage of total human genome	1.5%	100%

* The nucleotide sequence of 33.8×10^6 nucleotides is known; the rest of the chromosome consists primarily of very short repeated sequences that do not code for proteins or RNA.

** A pseudogene is a nucleotide sequence of DNA closely resembling that of a functional gene, but containing numerous deletion mutations that prevent its proper expression. Most pseudogenes arise from the duplication of a functional gene followed by the accumulation of damaging mutations in one copy.

Figura 1.4: Statistiche fondamentali del cromosoma umano 22 e dell'intero genoma umano.

Nel 2001 è stata pubblicata la prima "bozza" del genoma umano, le cui statistiche fondamentali, insieme a quelle specifiche al cromosoma 22, sono nella tabella in figura 1.4 .

Il genoma umano possiede molte differenti *sequenze regolatrici*, cruciali nel controllare l'espressione dei geni. Si tratta solitamente di brevi sequenze in prossimità e all'interno dei geni, le quali assicurano che i geni vengano espressi nel modo ed al momento giusto. La trascrizione differenziale dei geni è uno dei modi con cui gli eucarioti variano la quantità di proteine prodotte a seconda delle necessità cellulari. Soltanto ora, grazie alle potenzialità crescenti dei metodi e degli strumenti disponibili, sta cominciando ad emergere una conoscenza sistematica di tali elementi regolatori e di come essi agiscono insieme in una rete regolatrice genica. Tra gli elementi regolatori più noti vi sono:

- i *fattori di trascrizione*, proteine che legano il DNA in regioni specifiche (presso un promoter o un enhancer) per regolare la trascrizione,

- i *repressori*, proteine silenziatrici che legandosi al DNA bloccano la trascrizione,

- i *promotori*, sequenze di DNA a cui si lega l'RNA polimerasi per iniziare la trascrizione di un gene,

- gli *enhancer*, sequenze di DNA che in seguito al legame di proteine specifiche amplificano (fino a 200 volte) la frequenza di trascrizione del gene che controllano.

A parte i geni (che comprendono gli *introni*, sequenze non codificanti proteine) e le sequenze regolatrici note, il genoma umano contiene ampie regioni di DNA la cui funzione, se esiste, rimane tuttora ignota. Queste regioni comprendono la maggior parte del genoma umano (porzione stimata del 97%) e vi si trovano: *ripetizioni*, *trasposoni* (elementi genetici capaci di spostarsi da una posizione all'altra del genoma), *pseudogeni* (con struttura simile ai geni, ma privi di capacità di espressione), *junk DNA* (residui di processi evolutivi, senza utilità nel presente).

Si riporta il famoso commento riferito ai risultati dell'HGP: *"In some ways it may resemble your garage/bedroom/refrigerator/life: highly individualistic, but unkempt; little evidence of organization; much accumulated clutter; virtually nothing ever discarded; and the few patently valuable items indiscriminately, apparently carelessly, scattered throughout"* [6].

1.4 Diversi stati dei cromosomi nel corso del ciclo cellulare

Per formare un cromosoma, la molecola di DNA, oltre a contenere i geni, deve essere in grado di replicarsi e le copie replicate devono essere separate e divise nelle cellule figlie ad ogni divisione cellulare.

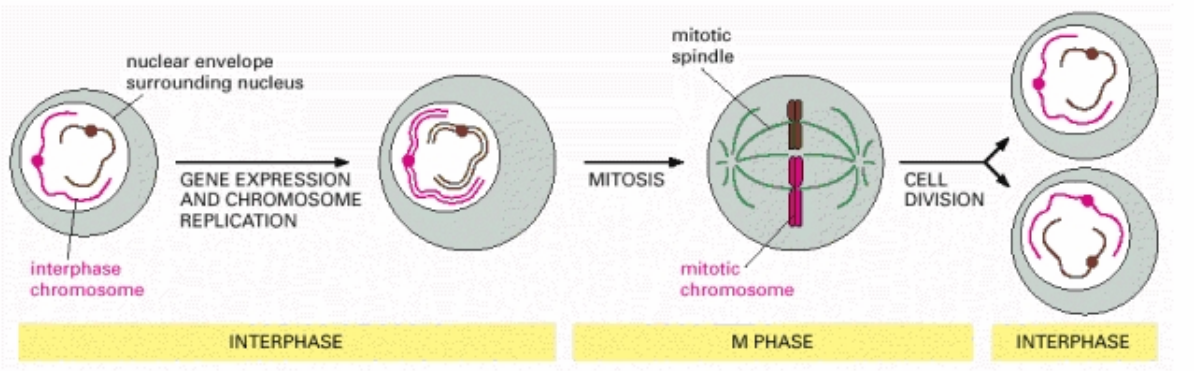

Figura 1.5: Schema semplificato del ciclo cellulare degli eucarioti.

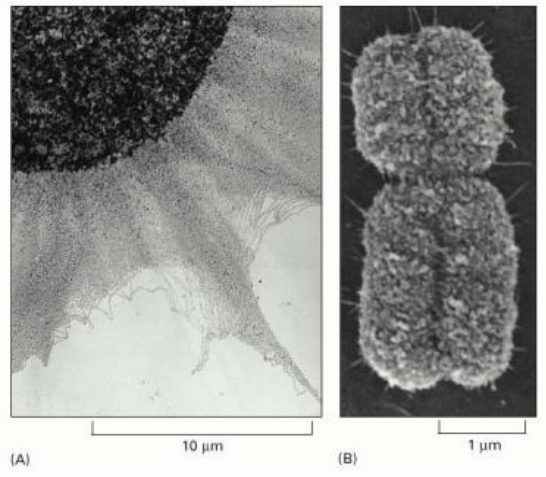

Figura 1.6: Confronto tra la cromatina d'interfase e la cromatina in un cromosoma mitotico. (A) Micrografia elettronica che mostra un groviglio di cromatina uscire da un nucleo d'interfase che ha subito lisi. (B) Micrografia elettronica di un cromosoma mitotico, duplicato e altamente condensato; i due nuovi cromosomi sono ancora legati tra loro; la regione più sottile indica il centromero. Si notino le differenze di scala.

Questo processo si verifica tramite una serie ordinata di fasi, collettivamente note come *ciclo cellulare*. Il ciclo cellulare è schematizzato in modo estremamente sintetico in figura 1.5. Durante l'*interfase*, la cellula sta esprimendo attivamente

i propri geni e sta quindi sintetizzando proteine; in questa fase, prima della divisione cellulare, avviene la replicazione del DNA, con la conseguente duplicazione dei cromosomi. Una volta che la replicazione del DNA è completa, la cellula può entrare nella *fase M*, dove avviene la mitosi ed il nucleo si divide in due nuclei figli. In questo passaggio, i cromosomi sono altamente condensati (si tratta dello stato in cui sono meglio visualizzabili, figura 1.6-B) e vengono separati in set cromosomici completi dal fuso mitotico. Affinché questo processo si realizzi, nei cromosomi sono presenti tre sequenze nucleotidiche specializzate. La prima è l'*origine di replicazione*, che corrisponde al sito in cui ha inizio la duplicazione del DNA. La seconda è il *centromero*, che permette che una copia di ciascun cromosoma condensato sia separata e distribuita in uno dei due nuclei figli; infatti, al centromero si forma un complesso proteico, chiamato *cinetocore*, il quale attacca i cromosomi duplicati al fuso mitotico per separarli. La terza sequenza di DNA specializzata molto importante per il processo considerato costituisce i *telomeri*, ossia le estremità dei cromosomi, evitando che tali estremità vengano riconosciute dalla cellula come DNA danneggiato da riparare.

La maggior parte del tempo nel ciclo cellulare viene trascorso in interfase; in confronto, la fase M è breve e ha la durata di una sola ora in molte cellule mammifere. Durante le fasi del ciclo cellulare in cui la cellula non si sta dividendo, i cromosomi si trovano nella forma di lunghi fili sottili ed apparentemente aggrovigliati, difficilmente distinguibili (*cromosomi d'interfase*, figura 1.6-A).

1.5 Packaging del DNA nei cromosomi Nucleosomi, unità di base della struttura dei cromosomi

Il DNA degli organismi eucarioti si condensa nella struttura dei cromosomi mitotici in modo complesso. Nel caso del cromosoma 22, il rapporto tra la lunghezza del DNA "srotolato" (circa 1.5 cm) e la lunghezza del cromosoma "condensato" (2 μm) è di circa 10,000. Sebbene i cromosomi d'interfase siano meno compatti di quelli mitotici, essi sono comunque fortemente condensati, con un rapporto delle relative dimensioni di circa 1,000.

Tale sorprendente compattazione viene mediata da proteine che arrotolano e ripiegano il DNA in livelli di organizzazione crescente.

É molto importante tenere presente che la struttura dei cromosomi è dinamica; infatti, non solo l'architettura globale dei cromosomi varia il proprio grado di compattazione durante il ciclo cellulare, ma regioni differenti di essi variano il proprio grado di addensamento affinché la cellula abbia accesso a sequenze speci-

fiche per l'espressione genica, il riparo del DNA e la replicazione. Il packaging dei cromosomi deve quindi essere realizzato in modo da permettere l'accesso rapido e localizzato alle regioni necessarie.

Le proteine che legano il DNA per formare i cromosomi eucariotici sono tradizionalmente divise in due classi generali: gli *istoni* e le *proteine cromosomiche non istoniche*. Il complesso di tali proteine e del DNA nucleare delle cellule eucariote forma la cromatina.

Gli istoni sono presenti in quantità talmente elevate che la loro massa totale nella cromatina è circa uguale a quella del DNA. Essi sono fondamentali per la costituzione del primo basilare livello di organizzazione del cromosoma, il *nucleosoma*, il quale fu scoperto nel 1974.

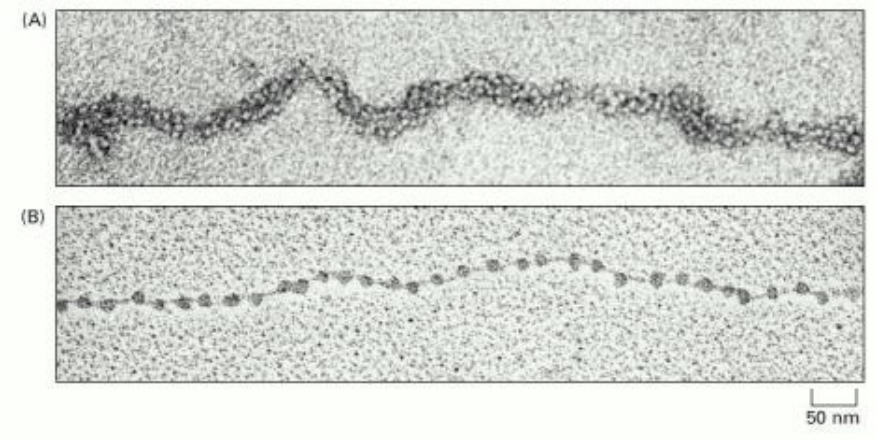

Figura 1.7: Nucleosomi visualizzati al microscopio elettronico. (A) Cromatina isolata direttamente da un nucleo d'interfase; appare come una fibra dello spessore di 30 nm. (B) Tratto di cromatina che è stata sperimentalmente despiralizzata dopo essere stata isolata, per mostrarne i nucleosomi.

Esaminando al microscopio elettronico il contenuto dei nuclei d'interfase, la maggior parte della cromatina si trova in forma di fibra con un diametro di circa 30 nm (figura 1.7-A). Sottoponendo questo tipo di cromatina a trattamenti che ne causando l'unfolding, il microscopio elettronico permette di visualizzare una sorta di *"filo a collana di perle"* (figura 1.7-B). Il filo corrisponde al *DNA linker* e le perle ai nucleosomi, ossia DNA arrotolato su un complesso proteico di otto istoni (detto *ottamero istonico*, costituito da due molecole di ognuno dei seguenti istoni: H2A, H2B, H3 e H4, figura 1.8).

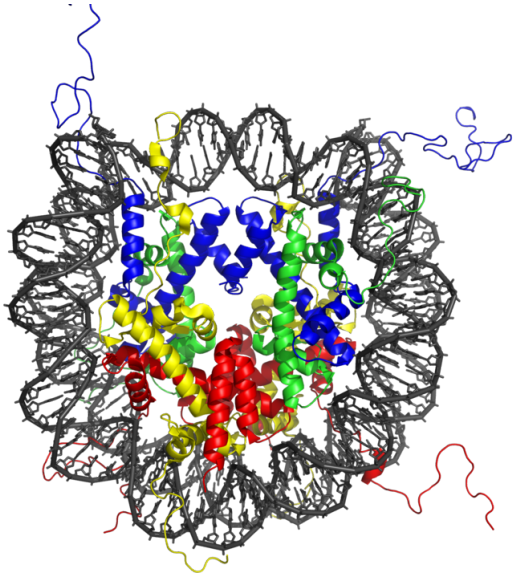

Figura 1.8: Struttura cristallina di un nucleosoma. Gli istoni H2A sono rappresentati in giallo, gli H2B in rosso, gli H3 in blu e gli H4 in verde; il DNA corrisponde alla doppia elica di colore grigio.

1.6 Fibra di cromatina di 30 nm

Il filo a collana di perle rappresenta quindi il primo livello di organizzazione strutturale del DNA. Tuttavia, la cromatina di una cellula vivente adotta raramente tale configurazione. Infatti, i nucleosomi sono generalmente impacchettati tra loro, generando stringhe regolari di DNA condensato che al microscopio elettronico appaiono nella forma di fibra di 30 nm di diametro.

Sono stati proposti diversi modelli per spiegare l'impacchettamento dei nucleosomi in questo stato della cromatina; uno dei più accreditati comprende una serie di varianti strutturali, note collettivamente con il nome di *modello a zig-zag* (figura 1.9) [7].

In realtà, la fibra di 30 nm consiste in un mosaico fluido di diverse configurazioni a zig-zag, a causa della lunghezza variabile del DNA linker tra due nucleosomi adiacenti e della presenza di altre proteine specifiche che si legano al DNA. Un modello alternativo per la formazione della fibra di 30 nm è quello *a solenoide*, che consiste in una superelica contenente circa sei nucleosomi per giro.

Come verrà spiegato nei prossimi capitoli, probabilmente vi sono diversi meccanismi che agiscono assieme nella formazione della fibra di 30 nm da una stringa lineare di nucleosomi; essi coinvolgono l'istone H1, che, come gli istoni dell'ottamero istonico, è carico positivamente e neutralizza la carica negativa del DNA,

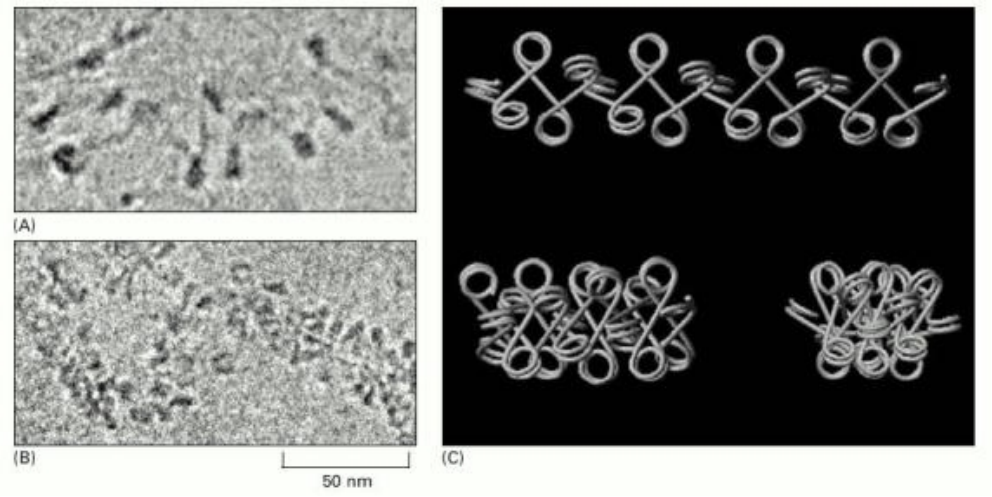

Figura 1.9: Varianti del modello a zig-zag della fibra di cromatina di 30 nm. (A e B) Evidenze di microscopia elettronica per i modelli strutturali che si trovano in (C). (C) Varianti a zig-zag; viene proposta un'interconversione tra queste varianti per spiegare la struttura della fibra cromatinica.

facilitandone quindi la compattazione [5].

1.7 Rimodellamento della cromatina Conseguenze funzionali (Epigenetica)

Vi sono due strategie principali per la variazione reversibile e locale della stuttura della cromatina.

La prima si serve dei *complessi di rimodellamento ATP-dipendenti*; si tratta di macchine proteiche che utilizzano l'energia fornita dall'idrolisi dell'ATP per variare temporaneamente la struttura dei nucleosomi, così che il DNA sia legato meno saldamente al core istonico.

La seconda strategia consiste in modificazioni covalenti delle code istoniche, tra cui l'*acetilazione* e la *metilazione* della lisina, o la *fosforilazione* della serina. Si tratta di processi mediati da enzimi specifici che hanno conseguenze molto importanti, poiché influenzano la stabilità della struttura cromatinica. Infatti, le code istoniche modificate sono in grado di attrarre proteine specifiche che, in base alla precisa modificazione avvenuta, possono causare l'ulteriore compattazione della cromatina o facilitare l'accesso al DNA.

In generale, il rimodellamento della struttura dei nucleosomi permette di regolare l'accesso al DNA nucleosomiale da parte delle proteine che mediano l'espressione genica, la riparazione e la replicazione del DNA.

É in questo contesto che negli ultimi decenni è nata l'*epigenetica*, lo studio dei cambiamenti ereditabili nell'espressione genica che non coinvolgono variazioni della sequenza di DNA vera e propria (ossia del *genotipo*), dando luogo quindi a modificazioni del *fenotipo*.

Grazie a questi meccanismi, la fibra di cromatina tende ad assumere due tipi di configurazioni, con differenze topologiche e funzionali: *eterocromatina* ed *eucromatina*. La prima è tipicamente molto densa, povera di geni e trascrizionalmente inattiva, mentre la seconda è meno condensata, ricca di geni e maggiormente accessibile alla trascrizione.

Vi sono modificazioni nucleosomiche specifiche che distinguono i territori eucromatici da quelli eterocromatici. L'eucromatina è tipicamente ricca degli istoni acetilati H3 e H4 e dell'istone metilato H3K4 (H3K4me); l'eterocromatina invece è associata all'ipoacetilazione degli istoni, alla metilazione dell'istone H3K9 (H3K9me), alla proteina HP1 (Heterochromatin Protein-1) e ad una forma metilata della base azotata citosina detta 5-metilcitosina (5mC).

É possibile impiegare queste caratteristiche per distinguere l'eterocromatina dall'eucromatina, anche se il confinamento dei territori eucromatici ed eterocromatici rimane poco compreso [5].

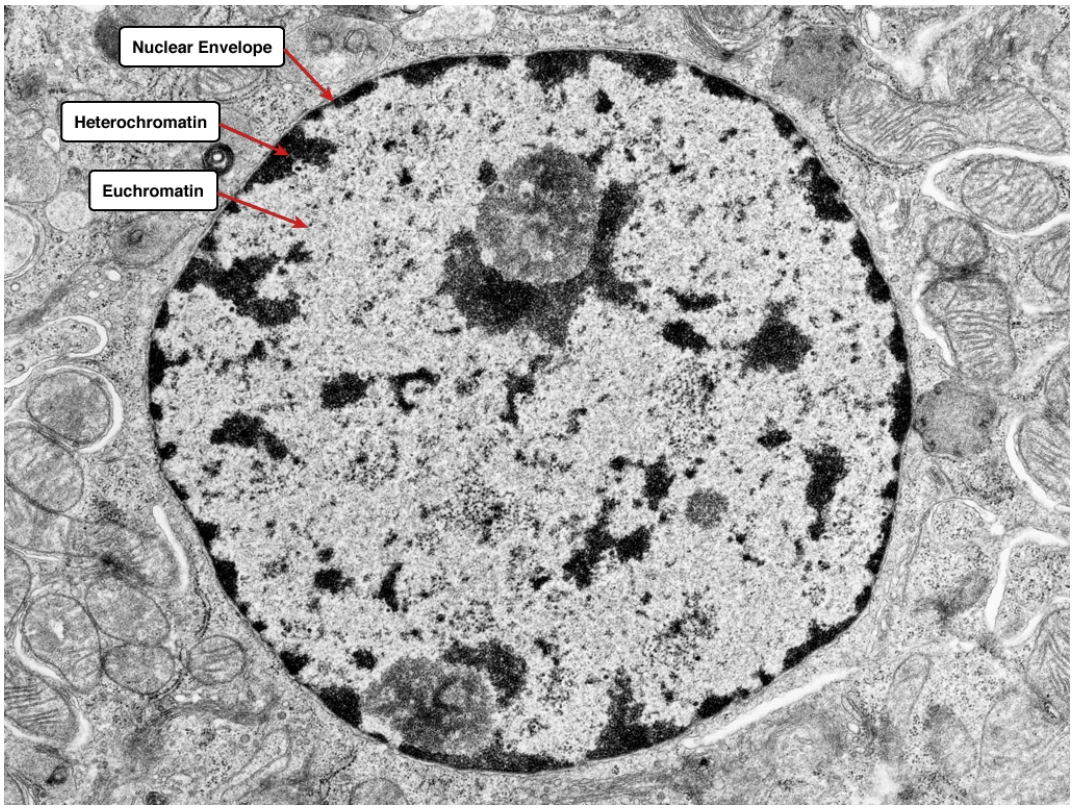

Figura 1.10: Nucleo cellulare. Si distinguono l'eterocromatina, in forma di piccole particelle scure, e l'eucromatina, meno addensata e localizzata.

Come rappresentato in figura 1.10, l'eterocromatina appare in forma di piccole particelle scure, irregolari, sparse nel nucleo o accumulate in prossimità della membrana cellulare; l'eucromatina è più dispersa e non facilmente colorabile.

Entrambe le forme di cromatina contribuiscono a funzioni biologiche molto importanti e riflettono il livello di attività della cellula, poiché l'eucromatina è prevalente nelle cellule trascrizionalmente attive mentre l'eterocromatina è più abbondante nelle cellule inattive.

Capitolo 2

Livelli di organizzazione dinamica della cromatina su diverse scale

La struttura dei cromosomi ha un ruolo essenziale nella regolazione di molti processi biologici, come l'attività dei geni, la replicazione e la riparazione del DNA. Le leggi che determinano l'organizzazione e la dinamica della cromatina nel nucleo d'interfase in vivo ed il loro legame con la regolazione della trascrizione non sono ancora state pienamente comprese.

Come detto, la struttura dei cromosomi presenta diversi livelli di organizzazione gerarchica, corrispondenti a differenti scale: la catena di DNA, i nucleosomi, la fibra di cromatina di 30 nm, strutture risultanti dal folding della fibra di cromatina (dette *strutture di ordine superiore* della cromatina) e, sulla scala dell'intero cromosoma, associazioni intercromosomiche (ad esempio, l'accoppiamento tra cromosomi omologhi). Alcune configurazioni cromosomiche sono accompagnate dall'ancoramento dei centromeri e/o dei telomeri alle strutture nucleari (come la membrana nucleare, il nucleolo) [8].

Negli ultimi anni, questi aspetti sono diventati un'importante oggetto di ricerca nei settori della biochimica, della citologia, della genetica e non solo, dato che fenomeni come il silenziamento dei geni sono legati alla posizione dinamica 3D dei geni coinvolti [9].

Sono stati proposti diversi modelli biofisici per descrivere in modo quantitativo l'organizzazione della cromatina su grande scala e la sua relazione con le dimensioni e la forma del nucleo.

Lo sviluppo della tecnica 3C (Chromosome Conformation Capture) e di tecniche derivate dalla 3C (4C – Circularized Chromosome Conformation Capture, 5C – Carbon Copy Chromosome Conformation Capture, Hi-C) ha permesso di indagare in vivo la struttura e le *interazioni a lungo range* della cromatina, a livello molecolare.

Nei lieviti, l'analisi 3C della cromatina trascrizionalmente attiva ha mostrato variazioni locali nella compattazione della cromatina che non supportano la presenza della fibra di 30 nm [10].

Utilizzando la tecnina Hi-C su scala genomica, che combina il legame di frammenti di DNA spazialmente vicini con il sequenziamento ad alta dimensionalità, si è dimostrato che il genoma è partizionato in numerosi domini che fanno parte di due compartimenti distinti [11].

All'aumentare della risoluzione dei dati, sono stati trovati domini di dimensioni inferiori, da cui si è desunto che i compartimenti sono suddivisi in strutture condensate delle dimensioni di $\sim$1 Mb, chiamate *domini topologici* (Topological Associated Domains - TAD) [12].

Con la disponibilità di set di dati di dimensioni sempre maggiori e di metodi computazionali rigorosi, sono stati individuati decine di migliaia di loop all'interno del genoma umano [1].

Per approfondimenti sulla tecnica Hi-C, dal protocollo sperimentale alle tecniche di analisi dei dati, si rimanda al capitolo 3.

2.1 Mega-domini e modello del globulo frattale

Tramite mappe Hi-C alla risoluzione di 1 Mb, Lieberman-Aiden et al. hanno analizzato tre livelli di organizzazione dell'architettura del genoma, corrispondenti a tre scale spaziali, rappresentati in figura 2.1. Sulla scala del nucleo cellulare, il genoma è suddiviso in due compartimenti distinti (*compartimenti A e B*), corrispondenti all'eucromatina (cromatina aperta, compartimento A) e all'eterocromatina (cromatina chiusa, compartimento B); si tratta di mega-domini delle dimensioni di 5-20 Mb, costituiti da loci che esibiscono lo stesso pattern di contatto e rappresentati da blocchi con maggior frequenza d'interazione nella diagonale delle mappe Hi-C. I cromosomi (raffigurati in figura 2.1 con colori diversi) occupano compartimenti distinti. Anche su scala cromosomica, si alternano regioni di cromatina aperta e chiusa e sulla scala di $\sim$1 Mb il cromosoma consiste di una serie di *globuli frattali*.

Il globulo frattale è un modello di organizzazione locale della cromatina sulla scala di $\sim$1 Mb supportato da diversi studi. Su questa scala infatti, la cromatina è consistente con una conformazione polimerica priva di nodi, che permette il massimo addensamento pur mantenendo la capacità di despiralizzare e riavvolgere facilmente un qualsiasi locus genomico.

Il globulo frattale è ben distinto dal modello di *globulo d'equilibrio* utilizzato precedentemente, che corrisponde ad una struttura molto annodata in cui i loci tra loro vicini nella catena di cromatina monodimensionale non lo sono neces-

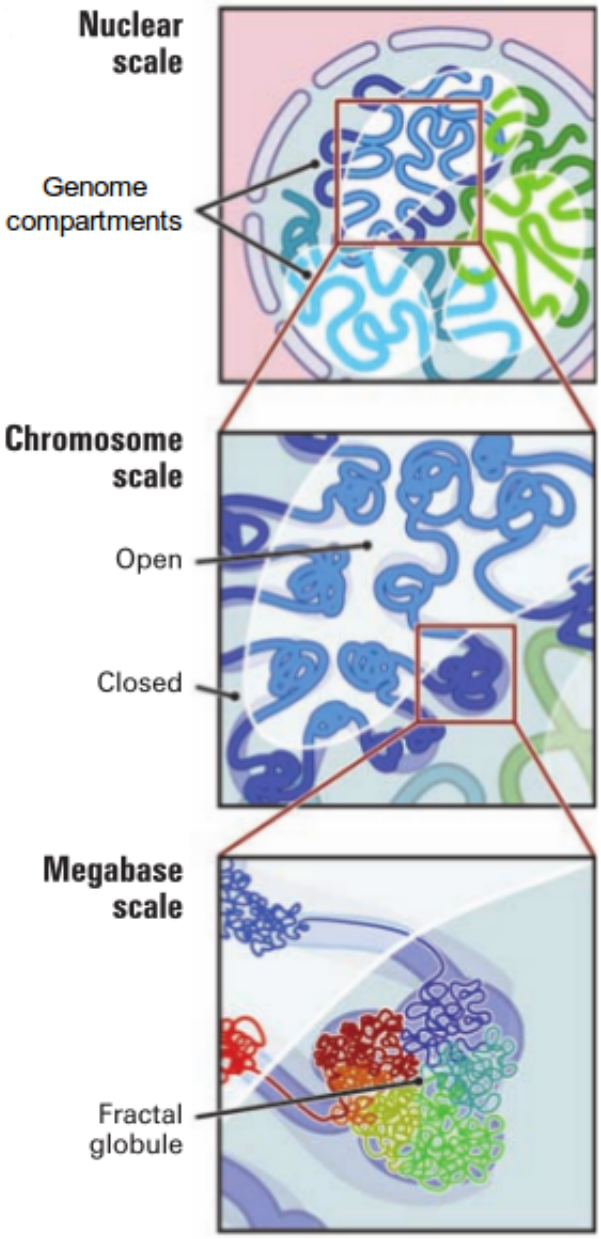

Figura 2.1: Architettura del genoma su tre scale. (In alto) Partizionamento del genoma in due compartimenti, corrispondenti ad eucromatina ed eterocromatina. I cromosomi (blu, verde, azzurro) occupano compartimenti distinti. (Al centro) I singoli cromosomi alternano regioni di cromatina aperta e chiusa. (In basso) Sulla scala del Mb, il cromosoma consiste di una serie di globuli frattali.

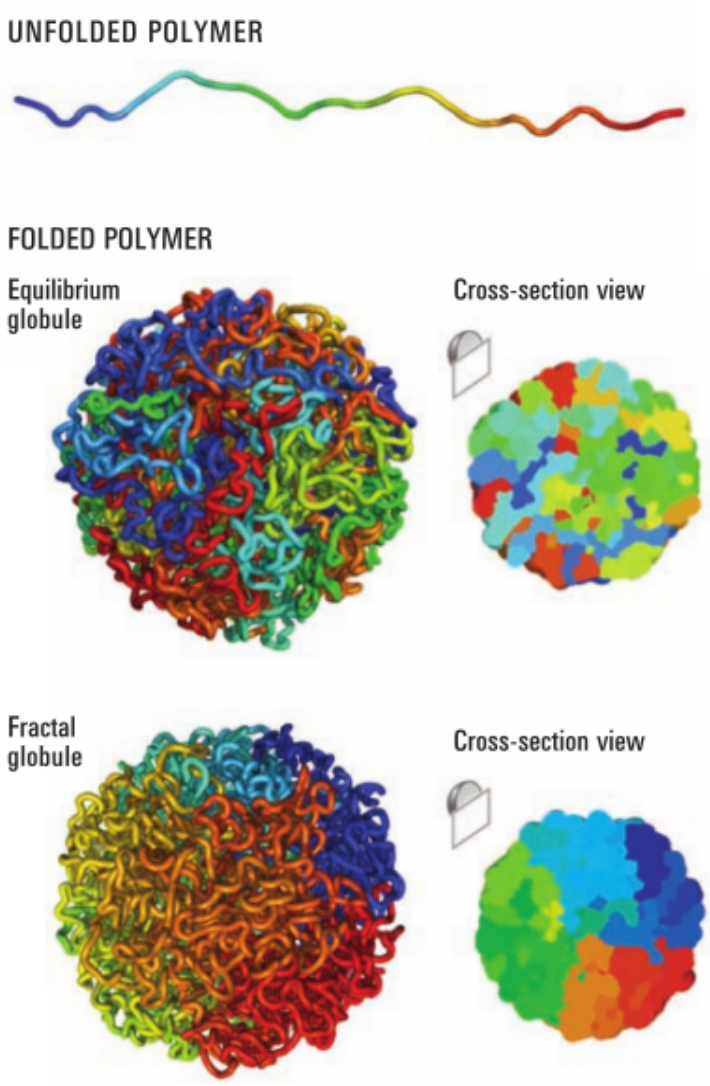

Figura 2.2: (In alto) Catena polimerica distesa; la colorazione corrisponde alla distanza progressiva da un estremo all'altro (blu, azzurro, verde, giallo, arancione, rosso). (Al centro) Globulo d'equilibrio; la struttura è fortemente aggrovigliata e loci che sono tra loro vicini nella catena monodimensionale (di colore simile) non lo sono necessariamente anche in 3D. (In basso) Globulo frattale; i loci tendono ad avere la stessa distanza reciproca nella catena lineare e nella struttura 3D, generando blocchi monocromatici sia alla superficie del globulo sia nella sua sezione.

sariamente anche in 3D. Nel globulo frattale, invece loci prossimi nella catena lineare tendono ad esserlo anche in 3D (figura 2.2).

Per approfondimenti relativi ai modelli 3D della cromatina, si rimanda al testo di Dekker et al. [13].

2.2 Territori cromosomici Topological Associated Domains (TAD)

La ricerca epigenomica ha come obiettivo la comprensione integrata degli aspetti strutturali e funzionali dell'epigenetica e dell'architettura nucleare nel corso della differenziazione delle cellule totipotenti o pluripotenti nei vari tipi cellulari funzionalmente distinti. Negli anni 2000, sono stati fatti enormi progressi riguardo le implicazioni epigenetiche della metilazione del DNA, delle modificazioni istoniche e degli eventi di rimodellamento della cromatina sulla regolazione dei geni. Tuttavia, ciò non è sufficiente per comprendere pienamente come il genoma dia luogo ai diversi epigenomi presenti nei vari tipi cellulari di un organismo pluricellulare. I diversi epigenomi e le loro implicazioni funzionali dipendono anche dall'organizzazione della cromatina su scale superiori e dell'architettura globale del nucleo cellulare.

L'organizzazione topologica dei cromosomi d'interfase in *territori cromosomici* (CTs - Chromosome Territories) è uno dei principi di base dell'architettura nucleare [14].

Questo tipo di indagine ha avuto inizio nel 1885 con Carl Rabl, ma è stato Theodor Boveri ad introdurre nel 1909 il termine "territorio cromosomico". La prima visualizzazione diretta dei CTs è stata possibile tramite tecniche di ibridizzazione in situ, sviluppate a metà degli anni '80.

Si ritiene che le configurazioni topologiche della cromatina non siano casuali e uno dei più importanti e ardui obiettivi della ricerca nell'ambito dell'architettura nucleare è lo studio del meccanismo responsabile della loro formazione e delle loro implicazioni funzionali. Questo settore di ricerca è ancora ai suoi albori, anche se si sta facendo luce su alcuni dei suoi aspetti basilari, tenendo conto anche del carattere dinamico dei territori cromosomici. Sono stati proposti numerosi modelli per spiegare l'assemblamento dei cromosomi d'interfase nei CTs. Tuttavia, essi forniscono soltanto pochi dettagli meccanici riguardo la relazione tra la struttura cromatinica di ordine superiore e la funzionalità genomica.

I recenti sviluppi nella tecnologie genomiche hanno permesso una vera e propria rivoluzione nello studio dell'organizzazione 3D del genoma. In particolare,

la tecnica Hi-C è un metodo che identifica le interazioni cromatiniche di ordine superiore su scala genomica.

In questo modo, Dixon et al. hanno investigato l'organizzazione 3D del genoma umano e di quello del topo in cellule staminali embrioniche ed in cellule differenziate, identificando domini d'interazione locale cromatinica della scala di ∼1 Mb, a cui hanno dato il nome di *domini topologici* (Topological Associated Domains - TAD) [12]. Per fare questo studio, gli autori hanno effettuato un'esperimento Hi-C sulle cellule coinvolte, generando matrici d'interazione a risoluzioni che vanno dai 10 kb a 1 Mb, da cui emergono regioni cromatiniche al cui interno l'interazione è molto alta (TAD) e al cui confine l'interazione è minima (*boundary*).

I domini così individuati sono legati rispetto ai compartimenti A e B descritti da Lieberman-Aiden et al., pur essendone indipendenti, ma soprattutto sono conservati sia tra diversi tipi cellulari sia tra specie differenti, suggerendo la possibilità che si tratti di una proprietà costitutiva dei genomi dei mammiferi, evolutivamente antica. Inoltre, Dixon et al. hanno trovato che i confini tra i domini topologici sono arricchiti di specifici elementi, come siti di legame del fattore di trascrizione CTCF e geni costitutivi (*geni housekeeping*), facendo supporre che questi fattori abbiano un ruolo fondamentale nella determinazione dei domini topologici.

2.3 Aumento della risoluzione Loop cromatinici

Al crescere della risoluzione delle tecniche sperimentali, si ha a che fare con set di dati di dimensioni maggiori e sono necessari metodi computazionali sempre più robusti. Rao et al. sono riusciti a produre mappe Hi-C per nove tipi cellulari, raggiungendo la risoluzione di 1 kb per le cellule lonfoblastoidi umane; hanno generato oltre 5 Tb di dati di sequencing, che registrano all'interno del genoma 15 miliardi di contatti distinti.

Utilizzando queste mappe, gli scienziati hanno trovato che il genoma è partizionato in *domini di contatto* (termine degli autori corrispondente ai domini topologici di Dixon et al.) di dimensioni comprese tra 40 kb-3 Mb (dimensione mediana pari a 185 kb), associati a pattern distinti modificazioni istoniche e confinati in sei sottocompartimenti. Inoltre, essi hanno identificato ∼10,000 loop cromatinici (figura 2.3). Gli studi hanno rivelato che i loop spesso collegano promoter ed enhancer, correlano con l'attivazione genica e sono conservati tra diversi

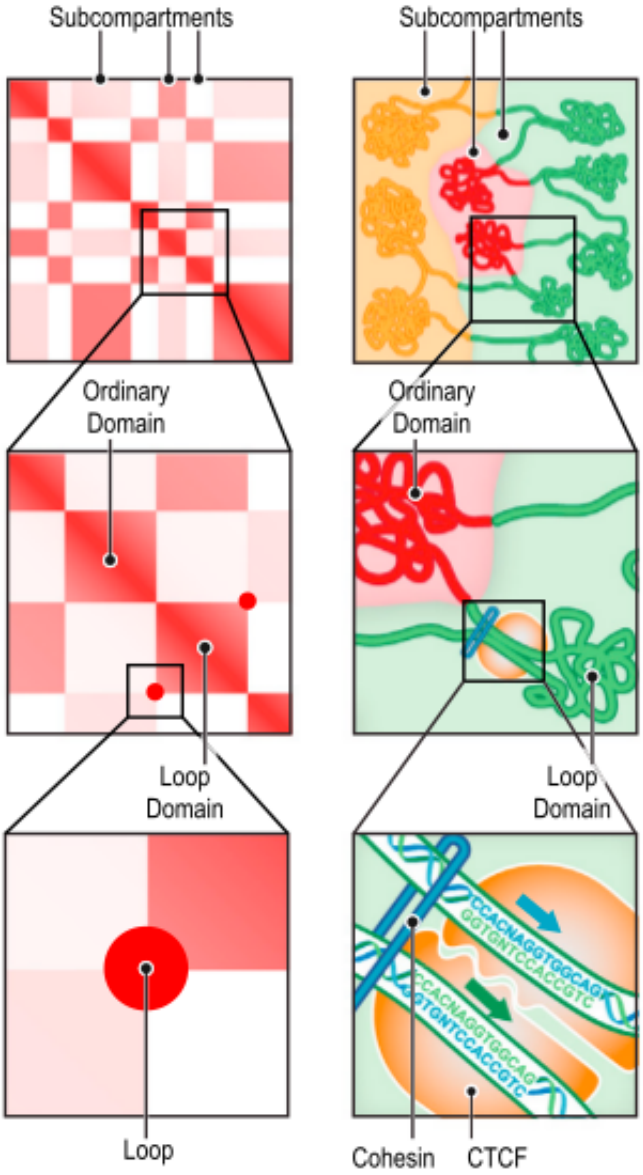

Figura 2.3: Schema dei livelli di organizzazione rivelate dalle mappe Hi-C di Rao et al. (In alto) Il pattern di contatto a lungo range di un locus (a sinistra) indica i loci adiacenti nella struttura nucleare (a destra). Sono stati individuati almeno sei compartimenti, ognuno con un pattern distinto di feature epigenetiche. (Al centro) Blocchi di maggiore frequenza di contatto lungo la diagonale (a sinistra) indicano la presenza di piccoli domini di cromatina condensata, la cui dimensione mediana è di 185 kb (a destra). (In basso) Picchi nella mappa di contatto (a sinistra) indicano la presenza di loop (a destra). I loop tendono a collocarsi ai confini dei domini e legano il fattore CTCF in orientazione convergente (a destra).

tipi cellulari e specie. Gli estremi dei loop tendono ad essere collocati sui confini dei domini di contatto e a legare il fattore di trascrizione CTCF.

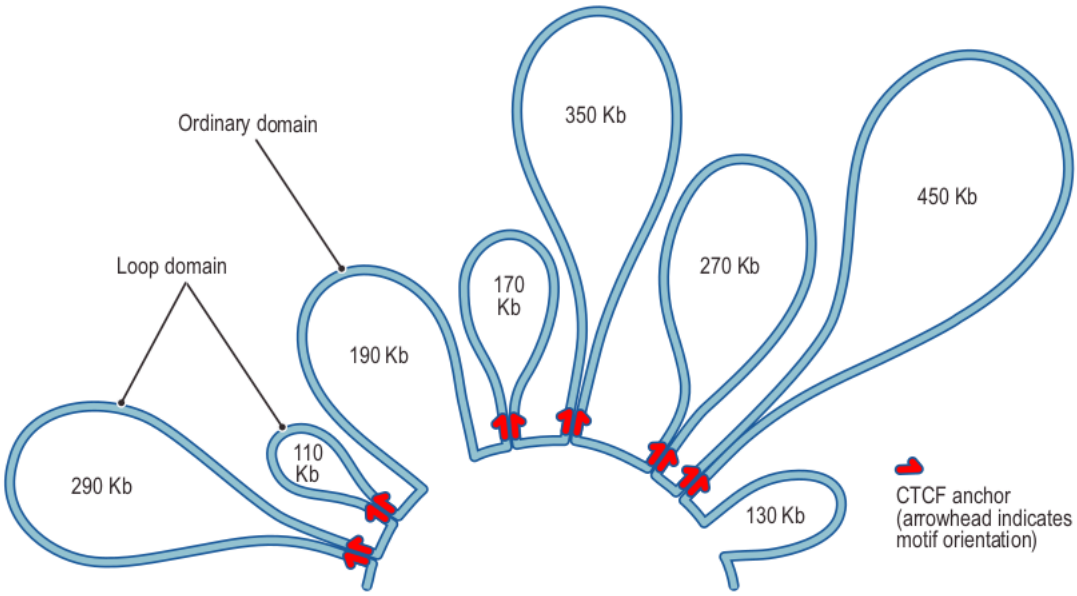

Figura 2.4: Le coppie di motivi CTCF che legano gli estremi dei loop si trovano quasi tutte in orientazione convergente.

I siti CTCF presso gli estremi dei loop si trovano prevalentemente (>90%) in orientazione convergente, con i motivi asimmetrici rivolti uno verso l'altro. Inoltre si è trovato che il cromosoma X è diviso in due macro-domini e contiene ampi loop ancorati a motivi CTCF (figura 2.4).

La risoluzione raggiunta dagli esperimenti Hi-C di Rao et al. ha anche permesso la creazione di mappe diploidi, oltre a quelle aploidi, per analizzare separatamente ciascun cromosoma omologo e rivelare caratteristiche omologo-specifiche.

Per approfondire le tecniche sperimentali utilizzate da Lieberman-Aiden et al., Dixon et al. e Rao et al., si rimanda al capitolo 3.

2.4 Alterazioni dell'organizzazione dell'architettura dei cromosomi in stati patologici

La configurazione 3D della cromatina è un aspetto fondamentale della funzionalità del genoma, poiché porta elementi regolatori e geni in prossimità spaziale per assicurare l'espletamento di appropriati profili di espressione genica specifici al tipo cellulare.

Vi sono evidenze sperimentali dell'alterazione dell'organizzazione 3D della cromatina nei casi patologici. Le cellule che si trovano in questi stati accumulano una varietà di cambiamenti epigenetici e la conseguente organizzazione 3D della loro cromatina è attualmente vivo argomento di ricerca.

La *senescenza cellulare* è coivolta in fenomeni come lo sviluppo dei tumori e l'invecchiamento ed è accompagnata da riarrangiamenti cromatinici su grande scala. La senescenza cellulare è un arresto irreversibile del ciclo cellulare, perciò della proliferazione cellulare, innescato da una varietà di stress che causano danni irreparabili alla catena di DNA (per esempio, la rottura di entrambi i filamenti del DNA).

Chandra et al. hanno mappato le alterazioni dell'organizzazione del genoma in cellule in stato senescente indotto da oncogene. Il confronto tra cellule staminali embrioniche (Embryonic Stem Cells - ESCs), cellule somatiche e cellule senescenti ha mostrato un calo unidirezionale nella connettività locale della cromatina, facendo supporre che la senescenza sia il punto finale di un processo di rimodellamento nucleare continuo che ha luogo nel corso della differenziazione [15].

Neretti et al. hanno recentemente impiegato in combinazione le tecniche Hi-C, FISH (Fluorescence In Situ Hybridization) e metodi in silico per caratterizzare l'architettura 3D dei cromosomi d'interfase in cellule proliferanti, quiescenti e senescenti. Come detto in precedenza, il partizionamento dei cromosomi in eucromatina accessibile ed in eterocromatina compatta è basilare per l'organizzazione del genoma. Sebbene l'organizzazione complessiva della cromatina nei compartimenti A (attivo, eucromatina) e B (repressivo, eterocromatina) e nei TAD sia conservata nelle tre condizioni, si è trovato che un gruppo di TAD cambia compartimento [16].

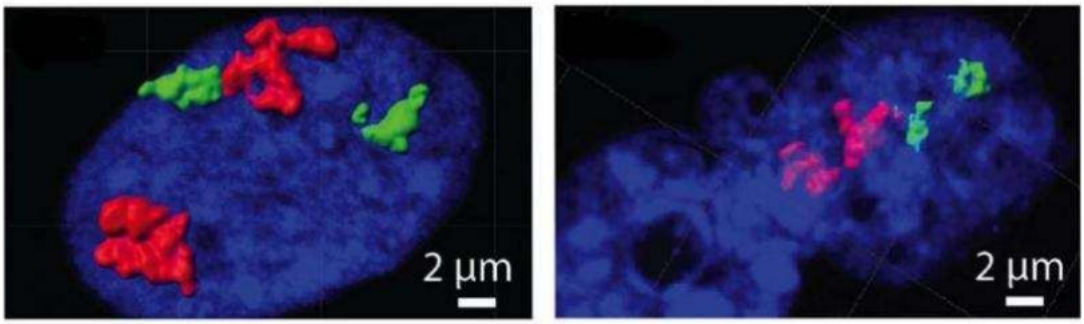

Figura 2.5: I cromosomi 4 (in rosso) e 18 (verde) sono evidentemente più piccoli nel nucleo di una cellula senescente (a destra) che in quello di una cellula non senescente (a sinistra).

A livello globale, le matrici d'interazione Hi-C delle cellule senescenti sono

caratterizzate da un calo relativo delle interazioni a lungo range ed un aumento delle interazioni a corto range all'interno dei cromosomi, fenomeno dovuto ad una riduzione significativa del volume occupato dai singoli bracci cromosomici (evidenziata dalla misura diretta della distanza tra loci genetici, figura 2.5). Nonostante questa compattazione complessiva, i centromeri tendono ad aumentare il proprio volume. Il modello strutturale che deriva dallo studio di Neretti et al. fornisce una visuale ad alta risoluzione della complessa architettura cromosomiale delle cellule senescenti.

Taberlay et al. hanno impiegato la tecnica Hi-C per investigare come la struttura 3D della cromatina venga perturbata in presenza di rimodellamenti epigenetici a lungo range, programmi di espressione genica alterati e *variazioni del numero di copie geniche* (Copy Number Variations - CNVs), ovvero polimorfismi quantitativi di tratti del DNA determinati da delezioni o duplicazioni di uno o più nucleotidi, caratteristici del cancro alla prostata [4]. In generale, questo tipo di alterazioni strutturali sono tipiche del cancro e causano la de-regolazione dell'espressione genica. Gli scienziati hanno trovato che la cellula mantiene la caratteristica segmentazione in TAD sulla scala di $\sim$1 Mb, ma i suoi domini sono generalmente più piccoli a causa della comparsa di nuovi domain boundary. É interessante che gran parte dei confini aggiuntivi che compaiono nelle cellule cancerose si instaurino nelle regioni che hanno CNV.

Tali studi hanno contribuito in modo decisivo a fare luce sulla relazione tra le alterazioni a lungo range genomiche ed epigenomiche e le variazioni nelle interazioni cromatiniche di ordine superiore che si verificano nelle cellule senescenti e cancerose.

Capitolo 3

Metodi di analisi per lo studio dell'architettura 3D del genoma

La quantità di genome-wide chromosome conformation capture data è in rapida crescita e presenta grandi opportunità e sfide stimolanti nel campo del modeling computazionale e dell'interpretazione 3D del genoma. In particolare, nei tempi recenti si è verificato uno sviluppo sorprendente nella produzione di dati Hi-C ad alta risoluzione (*High-throughput Chromosome conformation capture*).

In questo contesto, la varietà e complessità delle ipotesi biologiche che possono essere verificate necessita di metodi computazionali e statistici rigorosi per l'interpretazione dei dati Hi-C.

Nel presente capitolo viene approndita la tecnica Hi-C, dalla produzione sperimentale dei dati agli strumenti computazionali per la loro interpretazione: mapping, filtering, normalizzazione, estrazione dei contatti significativi, individuazione dei domini, visualizzazione e modellizzazione 3D.

Nella pagina web [17] sono riportati i tool bioinformatici relativi all'analisi dei dati 3C, 4C, 5C e Hi-C, specifici per determinati step del processing. Alcuni di questi tool sono più appropriati per gli step iniziali di mapping e filtering (come HiCUP e HiC-inspector), mentre altri sono adatti a passaggi d'analisi successivi (normalizzazione, visualizzazione ed estrazione dei contatti statisticamente significativi).

Ora più che mai, si riconosce che l'organizzazione 3D della cromatina influenza la regolazione genica e la funzionalità genomica. Le tecniche di chromosome conformation capture, prima a livello di loci singoli (tecniche 3C, 4C) o di gruppi di loci (tecniche 5C, ChIA-PET), poi su scala genomica (tecnica Hi-C), hanno permesso di studiare il legame tra struttura della cromatina, regolazione genica, replicazione del DNA ed alterazioni genetiche di vario tipo. Inoltre, studi Hi-C hanno rivelato caratteristiche strutturali conservate che attualmente sono ritenute

i principi organizzativi del folding della cromatina [1][11][12][18].

La conoscenza dei metodi di analisi dei dati Hi-C e delle modalità disponibili per effettuare ogni step d'analisi sta assumendo sempre più importanza, all'aumentare del numero e della varietà dei set di dati Hi-C.

Attualmente, sono disponibili i set di dati Hi-C di un gran numero di organismi diversi, come lieviti, batteri, la mosca, piante, parassiti malarici e numerose linee cellulari del topo e dell'uomo [19].

3.1 Protocollo sperimentale dell'Hi-C

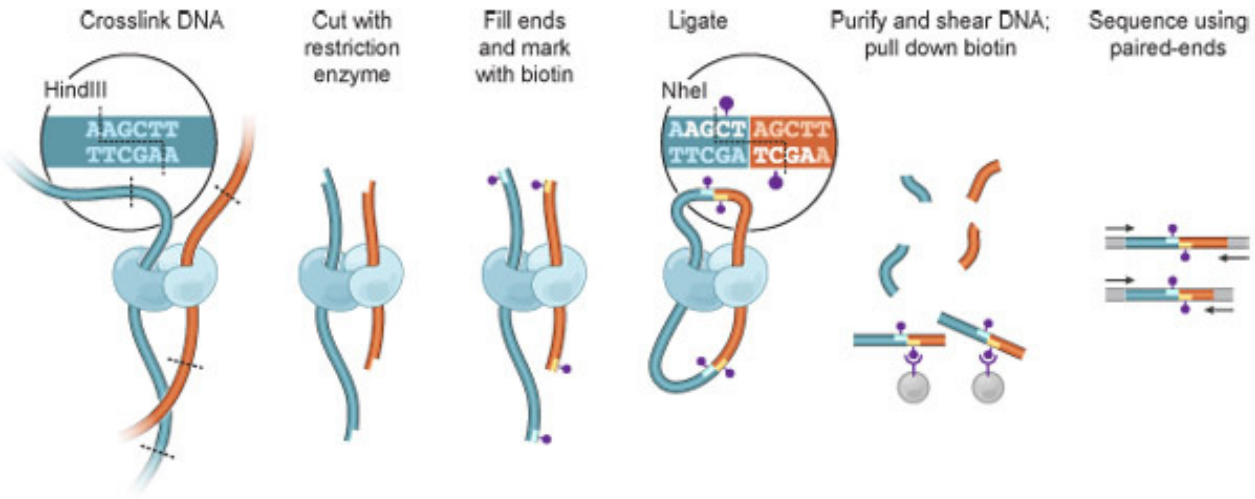

Figura 3.1: Schema sintetico della tecnica sperimentale Hi-C tradizionale (Lieberman-Aiden). Le cellule subiscono il crosslinking con formaldeide, che produce legami covalenti tra segmenti di cromatina spazialmente adiacenti (i frammenti di DNA sono rappresentati in blu e rosso; le proteine che mediano queste interazioni sono colorate di azzurro). La cromatina subisce la digestione ad opera di un enzima di restrizione (HindIII, nella procedura tradizionale di Lieberman-Aiden; i siti di restrizione sono evidenziati da linee tratteggiate). Le estremità libere che ne risultano sono legate da nucleotidi marcati con biotina (punti viola). Si effettua il legame delle estremità in condizioni diluite per creare molecole chimeriche. Si crea il sito di restrizione NheI (si veda l'inserto). Avviene la purificazione del DNA. Le giunzioni biotinilate sono isolate con streptavidina ed identificate con paired-end sequencing.

In sintesi, la procedura sperimentale Hi-C tradizionale consiste di sei step: (1) crosslinking delle cellule tramite formaldeide, (2) digestione del DNA con un enzima di restrizione che lascia le estremità libere, (3) marcatura delle estremità con biotina, (4) legame dei frammenti crosslinked in soluzione diluita, (5) taglio e

purificazione del DNA risultante, estrazione tramite streptavidina dei frammenti contenenti biotina e (6) sequenziamento a paired-end reads dei frammenti estratti (figura 3.1) [11].

La formaldeide è un composto organico che permette il *crosslinking* del DNA, ossia il legame covalente tra segmenti di cromatina prossimi nella struttura 3D. Nel processo denominato *digestione da restrizione*, un *enzima di restrizione* rompe la struttura del DNA in corrispondenza di una specifica sequenza di basi (normalmente 4-6 basi), definita *sito di restrizione*. La biotina è una vitamina idrosolubile che viene legata alle basi azotate per marcarle e la streptavidina è una proteina con un'altissima affinità per la biotina, permettendo quindi la selezione delle molecole che la contengono. In generale, il sequenziamento può essere eseguito a partire da una sola estremità del filamento di DNA che si intende sequenziare (*single-end reads*) o partendo da entrambe le estremità e proseguendo in direzioni opposte (*paired-end reads*).

Nel presente lavoro di tesi, sono stati utilizzati i dati Hi-C generati da Rao et al., la cui metodologia sperimentale Hi-C (denominata dagli autori *in situ Hi-C*) differisce dal protocollo Hi-C originale di Lieberman-Aiden (chiamato *diluition Hi-C*), poiché viene eseguita su nuclei cellulari intatti. I vantaggi di questo approccio dichiarati dagli autori, rispetto al metodo tradizionale, sono: la riduzione di contatti spuri dovuti a legami random tra frammenti nella soluzione diluita, la maggiore velocità del protocollo (richiede 3 giorni anziché 7), il raggiungimento di risoluzioni più alte grazie ad enzimi di restrizione più efficienti. Il confronto effettuato da Rao et al. tra le mappe Hi-C ottenute dalla metodologia in situ Hi-C e da quella di diluition Hi-C ha mostrato che le mappe in situ sono di qualità superiore alla alte risoluzioni, mentre alle risoluzioni inferiori i due metodi producono mappe molto simili [1].

In generale, questo processo produce una *mappa Hi-C*, ossia una libreria di sequencing su scala genomica che permette la misura delle distanze 3D tra tutte le coppie possibili di loci del genoma. La mappa Hi-C consiste in una lista di contatti tra frammenti di DNA prodotti dall'esperimento Hi-C. Segmentando il genoma lineare in loci di dimensione prefissata (ad esempio, in bin di 1 Mb o 1 kb), la mappa Hi-C può essere rappresentata come una *mappa di contatto* H, dove il coefficiente $H_{i,j}$ è il numero di contatti osservati tra i loci i e j. Si definisce *contatto* un accoppiamento tra *read* (corte sequenze di DNA di sintesi che vengono prodotte durante la reazione di sequenziamento) che non viene escluso dall'eliminazione delle read in duplice copia (che corrispondono a frammenti non legati) o che non possono essere allineate in modo unico al genoma.

La mappa di contatto può essere visualizzata come una *heatmap* (figura 3.2), i cui ingressi sono chiamati *pixel*. Un *intervallo* è riferito a una serie di loci

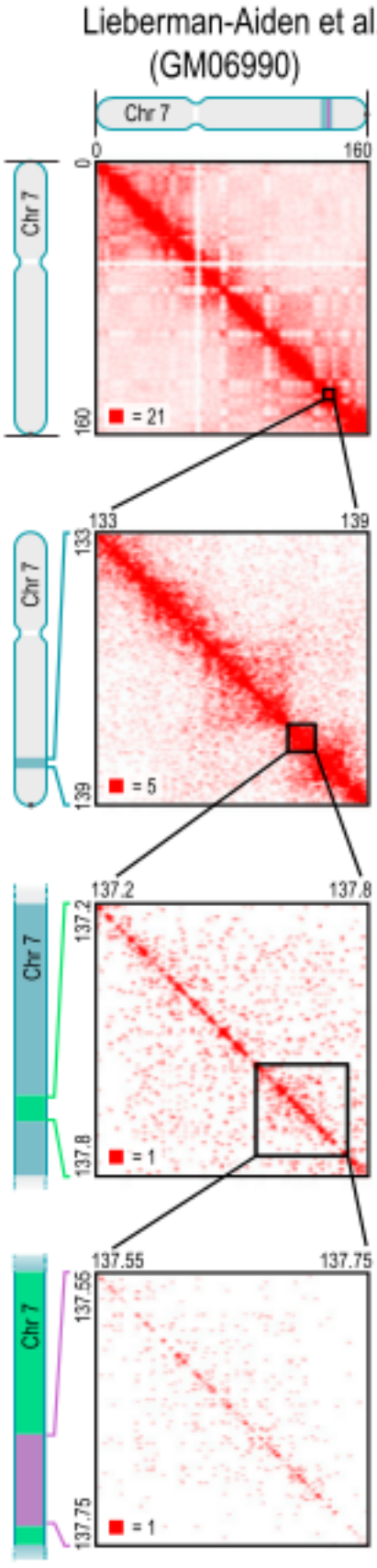

Figura 3.2: Mappa di contatto di Lieberman-Aiden et al. (2009) del cromosoma 7 della linea cellulare linfoblastoide GM06990. Dall'alto in basso: zoom crescente.

consecutivi; i contatti tra due intervalli individuano quindi un rettangolo o un quadrato (un *blocco*). La *risoluzione* di una mappa Hi-C è definita come la dimensione dei loci utilizzati per costruire la matrice dei contatti.

3.2 Mapping, filtering e classificazione delle read Hi-C

Il primo step di processing dei dati Hi-C consiste tipicamente nel mapping delle read al corrispondente genoma di riferimento (con specifiche procedure di pre-processing e post-processing per aumentare la percentuale di read mappate) e nel filtering delle read mappate a vari livelli.

Nei prossimi paragrafi si affrontano alcuni approcci di mapping e di filtering applicabili ai dati Hi-C.

3.2.1 Mapping

Idealmente, le due estremità di una read Hi-C paired-end corrispondono a posizioni distali nel genoma. In altre parole, la maggior parte dei frammenti in una libreria Hi-C ad alta qualità è composta da DNA proveniente da due o più loci non contigui. Tali frammenti vengono chiamati *chimere*. Quando le due estremità di un lungo frammento chimerico vengono sequenziate, se la giunzione è prossima alla metà del frammento, ognuna delle read risultanti sarà mappata in una differente posizione nel genoma. Tuttavia, se la giunzione si trova all'interno di una delle estremità sequenziate del frammento, allora le read stesse saranno chimeriche.

La frequenza delle read chimeriche dipende da molti fattori, tra cui il *size-selection step* (step di selezione dei frammenti in base alla loro lunghezza) e la *read length* utilizzata nel sequencing.

Per questo motivo, sono stati proposti diversi metodi per effettuare il mapping delle read Hi-C.

L'approccio più semplice è quello di filtrare tutte le read che non sono mappabili nel genoma perché chimeriche, ma vi sono almeno quattro metodi alternativi per recuperare l'informazione contenuta nelle read Hi-C chimeriche. Due di questi (*pre-truncation* e *iterative mapping*) pre-processano le read prima del mapping e gli altri due (*allow split alignments* e *split if not mapped*) fanno un post-processing dei risultati dopo il tentativo iniziale di mappare completamente tutte le read.

Per approfondire l'argomento, si rimanda al testo di Ay et al. [20].

3.2.2 Read-level filtering

Una volta che le singole read sono state mappate nel genoma, lo step successivo è decidere quali di queste read mappate siano "affidabili".

Il primo passaggio è l'applicazione di filtri standard ai disallineamenti, come la *mapping quality* (che associa alle read mappate un punteggio, MAPQ) e l'unicità delle read mappate.

Il secondo passaggio è la creazione di una lista di tutti i possibili siti di restrizione nel genoma di riferimento e l'assegnazione di ogni read al sito di restrizione più vicino. Il numero di siti di restrizione può essere elevato, rendendo necessario l'impiego di metodi computazionali robusti, come la ricerca binaria.

Nel terzo passaggio, viene calcolata la distanza tra la coordinata d'inizio di ogni read ed il sito di restrizione più vicino, per filtrare le read che non concordano con lo step di size-selection.

3.2.3 Read-pair level filtering Classificazione delle read

In molte pipeline Hi-C, le coppie di read per cui entrambe le estremità superano con successo i filtri iniziali sono suddivise in varie categorie. L'obiettivo di questa classificazione è far procedere l'analisi utilizzando soltanto le coppie di read che forniscono informazione sulla conformazione 3D della cromatina, al di là della prossimità lineare sulla catena 1D (*coppie di read informative*). Gli approcci relativi a questo livello di filtering delle coppie di read possono essere categorizzati in due gruppi: *strand filter* e *distance filter* [20]. Molte pipeline Hi-C impiegano una combinazione dei due approcci per assicurare l'eliminazione di tutti i possibili artefatti.

3.3 Normalizzazione delle mappe di contatto Hi-C

Idealmente, i coefficienti della matrice raw dei conteggi osservati sarebbero proporzionali alla frequenza di contatto vera tra i loci. Non è però trascorso molto tempo tra la pubblicazione del primo set di dati Hi-C [11] e la scoperta che i conteggi Hi-C hanno bias dovuti a diversi fattori (dipendenti dalla sequenza del DNA). Tra questi vi sono bias associati alle piattaforme di sequencing (come il contenuto di GC) ed all'allineamento delle read (legato alla mappabilità), come pure bias specifici della tecnica Hi-C (ad esempio la frequenza dei siti di restrizione).

La scoperta di questi bias ha portato alla formulazione di diversi metodi di normalizzazione e di correzione dei dati Hi-C [21].

Nel complesso, gli studi dimostrano che la normalizzazione è essenziale negli esperimenti Hi-C. Le mappe Hi-C normalizzate sono visivamente più smooth della loro versione raw, rendendo più efficace l'individuazione di pattern di contatto potenzialmente informativi. Inoltre, la normalizzazione migliora in modo significativo la similarità tra librerie Hi-C create con diversi enzimi di restrizione.

In generale, i conteggi raw e quelli normalizzati sono molto correlati nel caso di dati a bassa risoluzione. Questa correlazione, però, cala all'aumentare della risoluzione, facendo pensare che la normalizzazione è ancora più importante per set di dati Hi-C ad alta risoluzione.

Attualmente, la maggior parte delle implementazioni dei metodi di normalizzazione, per trattare dati Hi-C umani a risoluzioni <10 kb, necessita l'impiego di calcolo parallelo o di unità di elaborazione grafica (GPUs - Graphics Processing Units), più potenti dell' unità di elaborazione centrale standard (CPU - Central Processing Unit) [1].

3.3.1 Explicit-factor correction

I metodi di *explicit-factor correction* richiedono una conoscenza a priori dei fattori che possono generare bias nei dati Hi-C, assumendo che essi siano influenzati da un insieme predeterminato di bias.

Yaffe e Tanay hanno identificato tre fattori di questo genere e hanno sviluppato una procedura di correzione che modella la probabilità di osservare un contatto tra due regioni genomiche date le loro caratteristiche genomiche, come il contenuto di GC, la mappabilità e la lunghezza dei frammenti (noti per influenzare i conteggi Hi-C) [22]. Questo metodo di normalizzazione è quello impiegato da Dixon et al. nello studio dei TAD precedentemente menzionato [12].

Un metodo successivamente sviluppato, HiCNorm [23] fornisce una procedura di correzione esplicita significativamente più veloce, impiegando modelli di regressione (regressione binomiale negativa o di Poisson) e raggiungendo un'accuratezza di normalizzazione simile a quella di Yaffe e Tanay.

3.3.2 Matrix balancing

Un altro approccio alla normalizzazione è la correzione di tutti i fattori che possono causare bias, noti o meno, senza modellizzarli esplicitamente. I metodi di questo tipo si basano sull'assunzione che se non vi fossero bias allora ogni locus del genoma sarebbe "ugualmente visibile", ossia darebbe luogo allo stesso numero di read in un esperimento Hi-C. Questa assunzione rende la normalizzazione un

problema di matrix balancing, in cui l'obiettivo è trovare una decomposizione della mappa di contatto osservata $O = \overrightarrow{b}^T T \overrightarrow{b}$, dove $\overrightarrow{b}$ è un vettore colonna dei termini di bias e T è una mappa di contatto normalizzata, in cui tutte le righe hanno la stessa somma.

Nel contesto dell'Hi-C, Imakaev et al. hanno proposto un metodo iterativo chiamato ICE, che applica l'algoritmo appena descritto ripetutamente per raggiungere la decomposizione desiderata [24]. Questo problema di matrix balancing è stato studiato per decenni in molti diversi contesti (per approfodimenti, si rimanda alla Supplemental Information di [1]).

Il primo metodo di normalizzazione utilizzato è quello impiegato da Lieberman-Aiden et al., denominato *vanilla coverage normalization* (VC). Si calcola un termine di normalizzazione relativo alle righe, R_i, sommando i conteggi di ogni riga e prendendone il reciproco (*vettore VC*). Analogamente, si calcola un termine di normalizzazione relativo alle colonne, C_j, dato dal reciproco della somma dei coefficienti di ogni colonna. Per motivi di simmetria, $R_i = C_i$. Dati i coefficienti della matrice raw $M_{i,j}$, il corrispondente coefficiente della matrice normalizzata $M^*_{i,j}$ è $R_i M_{i,j} C_j$. Si tratta quindi di una normalizzazione molto semplice da implementare, veloce e robusta.

Uno dei problemi della VC normalization è che tende ad una sovra-correzione. Per ridurre questo effetto, un approccio molto semplice consiste nel considerare la radice quadrata del vettore VC (*SQRTVC – Square Root Vanilla Coverage*). Rao et al. hanno trovato che questo metodo fornisce risultati molto simili a quelli di algoritmi più complessi e sofisticati [1].

Più recentemente, Rao et al. hanno impiegato l'algoritmo di matrix balancing di Knight e Ruiz [25], più rapido ed adatto a normalizzare i loro set di dati Hi-C ad alta risoluzione, generati dal sequenziamento di miliardi di read.

Lo sviluppo di tool efficienti per normalizzare mappe Hi-C ad alta risoluzione con l'approccio di matrix balancing costituisce una sfida tuttora aperta.

3.3.3 Joint correction

Il fattore che maggiormente influenza il numero di contatti che si osserva tra una coppia di regioni genomiche di un cromosoma è la distanza 1D (rispetto alla catena lineare del DNA) che le separa. Questo tipo di bias verrà chiamato da qui in poi *effetto della distanza 1D*.

Come è prevedibile, il fenomeno viene riscontrato anche nel folding dei polimeri, in cui regioni adiacenti nella catena 1D non possono risultare eccessivamente distanti nello spazio 3D.

Alcuni metodi estendono i metodi delle tipologie viste per comprendere anche la correzione dell'effetto della distanza 1D (da cui il nome di *joint correction*).

Nello studio di tesi condotto, verrà impiegata la *normalizzazione di Toeplitz*, che consiste nella divisione di ogni elemento della matrice raw per la media dei coefficienti che si trovano alla stessa distanza dalla diagonale. Essa permette di ridurre l'effetto di diversi tipi di bias, tra cui l'effetto della distanza 1D. Tale metodo verrà spiegato nel dettaglio nel capitolo 4.

3.4 Estrazione dei contatti significativi

Un aspetto caratteristico dei chromatin conformation capture data è che permettono l'individuazione di contatti a lungo range, sia tra coppie di loci che si trovano nello stesso cromosoma ma che sono distanti l'uno dall'altro (contatti intracromosomiali a lungo range) sia tra loci in diversi cromosomi (contatti intercromosomiali). L'identificazione dei contatti intercromosomiali statisticamente significativi è abbastanza diretta, poiché, una volta che i bias sono stati eliminati dalla normalizzazione, in assenza di informazioni a priori sulle distanze tra coppie di cromosomi, ci si aspetta che tutte le coppie possibili tra loci intercromosomiali interagiscano allo stesso modo sotto l'ipotesi nulla. Invece, il numero di contatti tra due loci intracromosomiali dipende fortemente dalla distanza genomica lineare tra i loci. Questa dipendenza è dovuta soprattutto al random looping del DNA, più che alla formazione di legami cromatinici specifici. Perciò, nella valutazione della significatività statistica dei conteggi osservati, è necessario tenere conto di tale looping polimerico random [20].

Vi sono diversi approcci per la stima della significatività statistica dei conteggi, i quali tengono conto della dipendenza dalla distanza 1D dei contatti.

In ogni caso, la stima della confidenza ha implicazioni molto importanti nell'identificazione di interazioni funzionali tra enhancer e promoter e tra coppie di siti di legame CTCF che formano i loop cromatinici [1].

3.4.1 Observed/expected ratio

Un metodo per tenere conto della dipendenza dalla distanza 1D dei conteggi è creare un background model dei conteggi che tengono conto dello scaling delle distanze, dell'organizzazione dei domini e di altri bias corretti con la normalizzazione [26]. Questi modelli di background sono poi utilizzati per calcolare il rapporto tra contatti osservati e contatti attesi (*observed/expected ratios*) che sono poi trasformati in p-value o z-score.

3.4.2 Fit parametrici

Un approccio alternativo consiste nell'assumere che una distribuzione specifica descriva la dipendenza dei conteggi dalla distanza 1D e nell'eseguire una stima dei parametri per avere il best fit dei dati. Le distribuzioni precedentemente utilizzate comprendono la legge di potenza [11], la distribuzione di Laplace e la binomiale negativa. Una volta effettuato il fit parametrico dei dati, viene calcolato un enrichment score o una significatività statistica per ogni coppia di loci, in base alla loro distanza genomica ed al loro conteggio.

3.4.3 Fit non parametrici

Invece di assumere una distribuzione specifica, si può risalire alla dipendenza dei conteggi dalla distanza 1D utilizzando metodi non parametrici, direttamente dai conteggi osservati. Rispetto ai fit parametrici, quelli non parametrici sono più generali, poichè assumono che la dipendenza dalla distanza vari sostanzialmente in base a fattori come la risoluzione e al range di distanza genomica. Un metodo recente di questo tipo è Fit-Hi-C; per approfondire, si veda [20].

3.4.4 Peak detection

Un metodo più recente, HiCCUPS, affronta il problema di estrazione dei contatti significativi come un problema di peak detection bidimensionale, calcolando per ogni coppia di loci l'enrichment del loro conteggio rispetto alle regioni circostanti ed individuando quindi i contatti che nella mappa appaiono come picchi rispetto al background locale (problema molto costoso dal punto di vista temporale e computazionale). Questi picchi generalmente corrispondono a punti di anchoring di loop cromatinici molto stabili. In questo modo, si tenta di distinguere i conteggi funzionali da quelli dovuti al looping polimerico random o ad altri fattori non significativi.

3.5 Identificazione dei domini nelle mappe di contatto Hi-C

Come riportato in letteratura, sono stati individuati molti tipi di domini sulla base di caratteristiche epigenetiche specifiche [27], domini associati alla lamina nucleare (*LADs* – Lamina Associated Domains), o associati al nucleolo, o ad una combinazione di questi fattori. Questi domini sono definiti da pattern specifici di segnali 1D. Con la nascita dei dati Hi-C, sono state introdotte nuove tipologie

di domini, che corrispondono a pattern specifici nelle mappe di contatto. Questi domini includono compartimenti di cromatina aperta o chiusa (eucromatina ed eterocromatina) identificati tramite decomposizione spettrale [11], sottocompartimenti individuati dal clustering [28] e TAD (Topologically Associated Domains), corrispondenti a blocchi fortemente interagenti sulla diagonale della mappa di contatto [12].

Recentemente, i TAD sono diventati di particolare interesse per la comunità scientifica ed è stata svilluppata una varietà di metodi per la loro identificazione e caratterizzazione. Alcuni di questi metodi individuano un numero di TAD sostanzialmente diverso con differente distribuzione delle dimensioni. Queste differenze sono in parte dovute a fattori contingenti, come le diverse risoluzioni delle mappe di contatto utilizzate, ma indicano anche che considerare un singolo insieme di domini non sovrapposti può essere una semplificazione, vista l'eterogeneità dell'organizzazione in domini della popolazione cellulare considerata e l'organizzazione gerarchica e dinamica della cromatina (che permette un efficiente folding e unfolding).

Per ulteriori informazioni sull'impatto dell'organizzazione in TAD e delle sue variazioni sulla regolazione genica e la funzionalità genomica, si vedano [29] e [13].

3.5.1 Directionality Index Hidden Markov Model (DI HMM)

Un TAD genera uno "squilibrio" tra i contatti upstream e downstream rispetto ad una regione. Questo squilibrio è un indicatore del fatto che la regione sia all'interno di un TAD, in prossimità del boundary o lontano da un TAD. Dixon et al. hanno quantificato questo squilibrio con una statistica chiamata *directionality index* (DI) ed hanno utilizzato una catena di Markov nascosta (HMM - Hidden Markov Model) per determinare lo stato di bias sottostante di ogni locus (upstream, downstream o nessuno) [12]. Il metodo utilizza poi questi stati HMM per risalire ai TAD; una regione compresa tra due TAD viene identificata come boundary o come cromatina disorganizzata in base alla dimensione della regione stessa.

Anche altri studi utilizzano statistiche basate sui bias di direzionalità per determinare le presenza dei domini e le coordinate dei domini in cellule umane mitotiche [30].

3.5.2 Algoritmo di Arrowhead

In generale, all'aumentare della risoluzione dei dati, l'identificazione dei domini è sempre meno semplice, poiché intervengono fattori sperimentali come il rumore, ma vi sono pure difficoltà intrinseche: la decrescita della frequenza di contatto ai

confini del dominio può essere lieve e può quindi essere confusa con la decrescita molto rapida della probabilità di contatto che si osserva allontanandosi dalla diagonale della mappa di contatto.

Per trattare mappe di contatto a risoluzione molto elevata, Rao et al. hanno proposto un metodo euristico per trovare gli angoli dei blocchi nella diagonale delle matrici Hi-C umane e del topo, corrispondenti a domini 4-5 volte più piccoli dei TAD precedentemente identificati [1].

Vista l'alta risoluzione raggiunta, lungo le diagonali delle mappe vi sono molti piccoli blocchetti di alta frequenza; si verifica un calo della frequenza di contatto per coppie di loci che si trovano in parti opposte rispetto al confine del dominio.

L'algoritmo considerato per prima cosa trasforma la mappa di contatto in una *arrowhead matrix, A,* così definita:

$$A_{i,i+d} = (M *_{i,i-d} - M*_{i,i+d})/(M *_{i,i-d} + M*_{i,i+d}), \tag{3.1}$$

dove M^* è la matrice di contatto normalizzata.

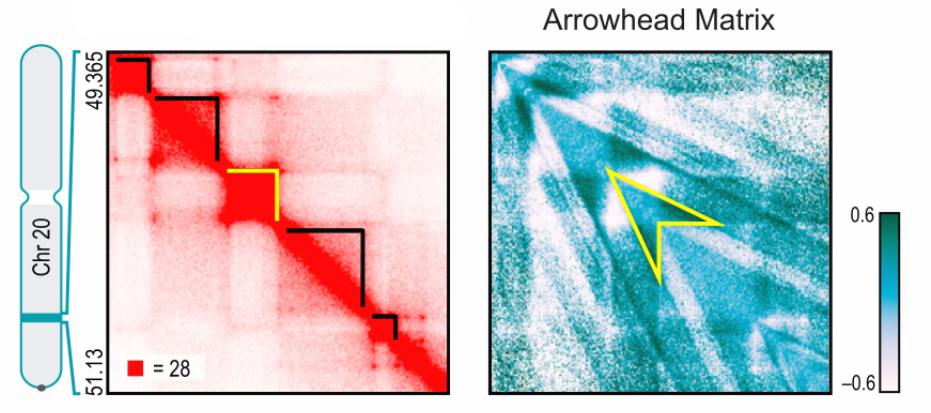

Figura 3.3: Sono stati individuati migliaia di domini nel genoma, corrispondenti a blocchi sulla diagonale della matrice Hi-C (a sinistra, evidenziati in nero). Per farlo, è stata definita la arrowhead matrix, che rimpiazza i domini con motivi a forma di triangolo, che puntano sulla posizione del dominio.

Questa trasformazione fa sì che la matrice risultante tenda a 0 quando due loci sono entrambi interni o esterni ad un dominio e crea motivi a forma di triangolo che puntano sull'angolo in alto a sinistra del dominio (figura 3.3); dalla arrowhead matrix, tramite programmazione dinamica viene creata una matrice "corner-score" che indica la probabilità di ciascun pixel di giacere al confine di un dominio.

I loop corrispondo invece a picchi nella frequenza di contatto, quindi a pixel localmente intensi nella matrice Hi-C; l'algoritmo utilizzato da Rao et al. compara

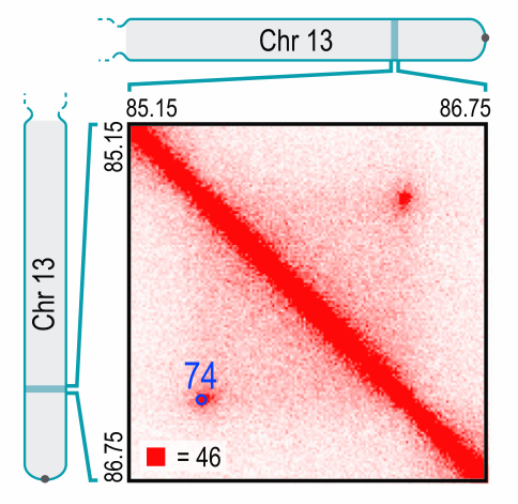

Figura 3.4: I loop vengono individuati nei pixel che presentano una frequenza di contatto significativamente maggiore rispetto alla regione circostante (indicati con cerchietti blu, di raggio di 20 kb, nella parte in basso a sinistra della mappa Hi-C). É indicato il numero di contatti raw del picco. La mappa di contatto rappresentata è alla risoluzione di 10 kb.

il numero di contatti in un pixel al numero di contatti nella regione circostante (figura 3.4).

3.5.3 Domini gerarchici multi-scala

Dall'ispezione visiva di una mappa di contatto Hi-C risulta evidente la presenza di sotto-strutture all'interno dei TAD, che a loro volta possono corrispondere ad unità gerarchiche di regolazione genica o di altre funzioni. É proprio questo il principio su cui si basa l'algoritmo di Chen et al., utilizzato nel presente lavoro di tesi. Esso verrà trattato in modo approndito nel capitolo 4.

3.6 Modellizzazione 3D della struttura della cromatina

In assenza di chromosome conformation capture data, la modellizzazione 3D del genoma può essere effettuata con simulazioni polimeriche basate su un certo numero di assunzioni fisiche e di parametri. Tali approcci polimerici rappresentano i cromosomi come catene polimeriche che diffondono in uno spazio vincolato (spazio nucleare) [31].

Avendo invece a disposizione le mappe di contatto su scala genomica, risalire alla struttura 3D della cromatina, sottostante ai contatti osservati, è diventato un problema fondamentale. I relativi metodi si dividono in due gruppi.

Il primo gruppo, quello dei *consensus methods*, individua la conformazione 3D che meglio descrive i dati Hi-C osservati. Tuttavia, il protocollo sperimentale della tecnica Hi-C impiega milioni di cellule per la creazione di una libreria e quindi potenzialmente si ha a che fare con una varietà di differenti conformazioni.

Per tenere conto dell'eterogeneità cellulare, il secondo gruppo di metodi di ricostruzione 3D, quello degli *ensamble methods*, risale ad un insieme di strutture che rappresentano i dati Hi-C osservati. Infatti, Nagano et al. hanno dimostrato la possibilità di generare dati Hi-C single-cell, portando ad una caratterizzazione e modellizzazione più diretta della variabilità della struttura cromosomica cellulare [18].

3.6.1 Consensus methods

Uno dei metodi più utilizzati per risalire al modello 3D dai chromosome conformation capture data è il *multi-dimensional scaling* (MDS). L'MDS è un metodo statistico classico per cui, date tutte le distanze tra coppie di un insieme di oggetti (matrice delle distanze), si trova la struttura che meglio approssima le distanze tra coppie di elementi. Nel contesto dell'Hi-C, gli oggetti sono frammenti di DNA e le distanze tra coppie di frammenti sono calcolate applicando una trasformata ai conteggi Hi-C. Lo svantaggio dei metodi di MDS è che sono molto sensibili alla scelta della funzione che trasforma i conteggi osservati in distanze spaziali.

Mozziconacci et al. hanno sviluppato l'algoritmo ShRec3D, il quale modellizza la mappa dei conteggi Hi-C come un network pesato, in cui i pesi dei link sono presi come l'inverso della frequenza di contatto tra loci di DNA (nodi); considerando come distanza tra una coppia di nodi lo shortest paths che li separa, calcolato tramite l'algoritmo di Floyd-Warshall, è possibile impiegare il metodo MDS per risalire alle coordinate della struttura 3D sottostante. Per approfondire l'argomento, si rimanda al testo di Mozziconacci et al. [32].

3.6.2 Ensemble methods

Per l'inferenza di un ensemble di modelli 3D, sono stati proposti diversi metodi probabilistici, che producono un insieme di strutture rappresentative dei conteggi osservati. Questi metodi si dividono a loro volta in due categorie: procedure che trovano soluzioni multiple, ognuna delle quali fitta i dati Hi-C, e procedure che individuano un ensemble "vero" che complessivamente descrive in modo ottimale i dati.

Nel primo caso, si è vicini al consensus method, ma invece di risalire ad un modello localmente ottimale, l'ottimizzazione viene lanciata con inizializzazioni multiple, producendo modelli diversi. La variabilità di questi modelli dipende fortemente dalla struttura del problema e dalle inizializzazioni random, rendendo difficile l'interpretazione del legame tra i modelli risultanti e la variabilità cellulare della struttura cromatinica. Rousseau et al. hanno sviluppato un metodo di questo tipo che utilizza un metodo Monte Carlo basato su Catena di Markov (MCMC - Markov Chain Monte Carlo) per approssimare la probabilità a posteriori di ogni modello [33].

Il secondo gruppo di ensamble methods genera simultaneamente, in una singola ottimizzazione, migliaia di strutture, ognuna delle quali è pienamente consistente con i vincoli derivati dai dati Hi-C e che, nel complesso, spiegano meglio i conteggi osservati.

3.7 Visualizzazione dei dati Hi-C

La visualizzazione dei dati genomici è cruciale sia per la formulazione di ipotesi sia per l'individuazione di potenziali artefatti. Nel web vi sono molti browser genomici ed epigenomici volti alla visualizzazione dei dati umani, del topo o di altri organismi. Tuttavia, questi browser sono volti principalmente alla visualizzazione di segnali 1D [34].

Uno dei tool adatti alla visualizzazione di mappe 2D è l'Hi-C Data Browser [35], che utilizza l'UCSC Genome Browser per permettere la visualizzazione simultanea di mappe Hi-C e di segnali 1D. Un'applicazione più recente, Juicebox, consente la visualizzazione di mappe di contatto di diversi set di dati Hi-C (umani o del topo) insieme ad altre feature, come domini, picchi di HiCCUPS e siti di legame CTCF [36]. Ultimamente sono stati sviluppati anche tool per la visualizzazione di modelli 3D della cromatina, tra cui Genome 3D [37] e TADkit [38].

Capitolo 4

Materiali e metodi

Il lavoro di tesi consiste nell'individuazione dei domini topologici (TAD) in cellule umane, al fine di fare un confronto tra sette tipi cellulari diversi, e nello studio topologico del network associato alle matrici Hi-C.

L'obiettivo è la verifica dell'eventuale conservazione dei TAD boundary tra tipi cellulari diversi, l'individuazione dei tipi cellulari con maggior grado di similarità e l'estrazione di informazione topologica tramite misure di centralità sul network.

L'identificazione dei domini è stata effettuata tramite l'algoritmo *TAD_Laplace* presentato da Chen et al. [28], che in modo iterativo segmenta il grafo associato alla mappa di contatto Hi-C tramite il Fiedler vector, ossia l'autovettore relativo al primo autovalore diverso da zero del laplaciano. L'algoritmo considera sia le interazioni locali sia le interazioni a lungo range rappresentate dalla matrice Hi-C.

Nella fase preliminare del lavoro, sono state fatte diverse prove con l'algoritmo *TAD_Laplace* per confrontare i TAD boundary ottenuti con quelli pubblicati da Dixon et al. e Chen et al. [12][28], al fine di verificarne l'attendibilità e prendere confidenza con l'impostazione dei parametri dell'algoritmo.

In questo capitolo, verranno illustrati i dati ed i metodi computazionali e statistici utilizzati nell'analisi condotta.

4.1 Dati utilizzati

TAD_Laplace è stato applicato sui dati Hi-C di sette tipi cellulari, a tre risoluzioni differenti: 50 kb, 500 kb ed 1 Mb. Essi fanno parte del set di dati pubblicato da Rao et al., disponibile sul Gene Expression Omnibus database, con codice di accesso GSE63525 [39]).

In questo set di dati, sono compresi degli archivi tar contenenti le matrici di contatto osservate raw, intra-cromosomiali ed inter-cromosomiali, di ogni tipo

cellulare analizzato nello studio di Rao et al.: GM12878, HMEC, HUVEC, HeLa, IMR90, K562, KBM7, NHEK [1].

Gli archivi delle matrici di contatto intra-cromosomiali contengono le mappe ad otto (nove per la linea GM12878) diverse risoluzioni: 1 Mb, 500 kb, 250 kb, 100 kb, 50 kb, 25 kb, 10 kb, 5 kb (ed 1 kb per GM12878). Per ogni risoluzione, vi sono le subdirectory relative ad ogni cromosoma. Al loro interno, vi sono altre due subdirectory, MAPQG0 e MAPQGE30. La subdirectory MAPQG0 contiene i file delle matrici di contatto costruite con tutte le read pair che sono allineate in modo unico al genoma di riferimento (MAPQ – MAPping Quality >0). La subdirectory MAPQGE30 contiene i file delle matrici di contatto costruite con le read pair che sono mappate nel genoma con una MAPQ≥30.

Le mappe di contatto raw sono rappresentate in notazione sparsa in file di testo. Ogni linea ha tre campi: i, j e H_{ij} (i e j sono espressi come gli estremi inferiori – in *bp*, *base pairs* – dei bin a cui sono riferiti). É riportata solo la matrice triangolare superiore e, siccome la matrice H è simmetrica, vale $H_{ij} = H_{ji}$.

Il set di dati pubblicato contiene anche i vettori di normalizzazione corrispondenti a diversi metodi di normalizzazione, per trasformare le mappe di contatto raw in mappe di contatto normalizzate. Essi sono stati impiegati nella fase di lavoro iniziale, per confrontare i metodi di normalizzazione vanilla coverage (VC), square root vanilla coverage (SQRTVC) e di Knight e Ruiz (KR – metodo utilizzato da Rao et al.), illustrati nel capitolo 3, con il metodo di Toeplitz (impiegato da Chen et al. e nel lavoro di tesi).

I file di testo dei vettori di normalizzazione sono organizzati in modo che la prima riga è il fattore di normalizzazione della prima riga/colonna della matrice di contatto raw, la seconda riga è il fattore di normalizzazione della seconda riga/colonna della matrice di contatto e così via. Per normalizzare il coefficiente H_{ij} della matrice raw, bisogna dividerlo per i corrispondenti fattori di normalizzazione delle righe/colonne i e j.

Nel lavoro di tesi, sono stati utilizzati i dati Hi-C a tre risoluzioni, 50 kb, 500 kb ed 1 Mb, di sette tipi cellulari, GM12878, HMEC, HUVEC, IMR90, K562, KBM7 e NHEK (la linea HeLa non è stata utilizzata poiché richiede dei codici di accesso aggiuntivi), con MAPQ≥30 (buona qualità di allineamento al genoma, evita più rigorosamente i falsi positivi dovuti al cattivo allineamento). Per utilizzare l'algoritmo $TAD\_Laplace$ su queste mappe, nel pre-processing esse sono state convertite in formato full.

4.2 Informazioni di base sulle linee cellulari analizzate

Come detto, sono stati utilizzati i dati Hi-C relativi a sette tipi cellulari umani: GM12878, HMEC, HUVEC, IMR90, K562, KBM7 e NHEK.

GM12878 è una linea cellulare di linfociti B del sangue (sistema immunitario).

La sigla HMEC (Human Mammary Epithelial Cells) indica cellule epiteliali mammarie, isolate dal tessuto mammario di una femmina adulta.

Le cellule HUVEC (Human Umbilical Vein Endothelial Cells) sono cellule endoteliali della vena del cordone ombelicale.

La linea IMR90 è costituita da fibroblasti polmonari (tessuto connettivo) prelevati da un feto normale di 16 settimane.

Le cellule K562 sono linfoblasti prelevati dal midollo osseo di un individuo adulto malato di leucemia mieloide cronica (CML – Chronic Myelogenous Leukemia). I linfoblasti sono cellule immature della serie linfoide dei globuli bianchi, da cui derivano i linfociti B.

Le cellule somatiche KBM-7 sono state isolate da un paziente affetto da leucemia mieloide cronica. La linea è quasi-aploide, ossia è aploide per quasi tutti i cromosomi.

Le cellule NHEK (Normal Human Epidermal Keratinocytes) sono cheratinociti, il tipo cellulare più abbondante nell'epidermide, prelevati da un individuo normale.

4.3 Teoria spettrale dei grafi

La modellizzazione tramite grafi dell'organizzazione spaziale dei cromosomi permette di impiegare metodi spettrali per studiarne quantitativamente le proprietà. La strategia utilizzata per identificare i TAD è basata sulla teoria spettrale dei grafi applicata alla matrice Hi-C. Di seguito, ne vengono illustrati i principi di base.

Un *grafo simmetrico* $\mathcal{G}(V, E)$ è definito come coppie ordinate degli elementi degli insiemi V ed E. $V = \{v_1, v_2, ..., v_N\}$ è un insieme finito di nodi, con cardinalità N, ed E è un insieme di elementi del tipo $\{v_i, v_j\}, i \neq j$.

La *matrice di adiacenza* A è una matrice simmetrica $N \times N$ che rappresenta le relazioni di adiacenza nel grafo $\mathcal{G}$, tale che $A_{ij} = 1$ se $\{v_i, v_j\} \in E$, altrimenti $A_{ij} = 0$ (con A_{ij} si denota il coefficiente (i, j) della matrice A).

Il *connectivity degree* di un nodo, $d(v_i)$, è pari al numero di nodi connessi al nodo v_i ed è dato da

$$d(v_i) = \sum_{j \in N_i} A_{ij}, \tag{4.1}$$

dove N_i denota l'insieme dei nodi connessi a v_i. Nel caso di matrice di adiacenza binaria, il degree dei nodi è dato dalla somma per righe/colonne di A.

La *degree matrix*, D, è definita come una matrice diagonale con i-esimo coefficiente della diagonale pari a $d(v_i)$, ossia

$$D_{ij} = \begin{cases} d(v_i), & i = j \\ 0, & i \neq j. \end{cases} \tag{4.2}$$

Il *laplaciano* del grafo $\mathcal{G}$ è dato da

$$\mathcal{L} = D - A \tag{4.3}$$

e perciò

$$\mathcal{L}_{ij} = \begin{cases} d_i, & i = j \\ -1, & i \sim j \\ 0, & altrimenti, \end{cases} \tag{4.4}$$

dove con $i \sim j$ si intende che i nodi i e j sono connessi.

Il *laplaciano normalizzato* è espresso da

$$\mathcal{L}_{\mathcal{N}} = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}}, \tag{4.5}$$

ossia

$$\mathcal{L}_{\mathcal{N}ij} = \begin{cases} 1, & i = j \\ -\frac{1}{\sqrt{d_i d_j}}, & i \sim j \\ 0, & altrimenti. \end{cases} \tag{4.6}$$

Gli autovettori v di $\mathcal{L}_{\mathcal{N}}$ e gli autovettori u di $\mathcal{L}$ corrispondenti sono legati della relazione di rescaling $u = D^{-1/2}v$.

Per un grafo connesso $\mathcal{G}$, siano $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N$ gli autovalori ordinati in modo crescente di $\mathcal{L}$ (o $\mathcal{L}_{\mathcal{N}}$).

In generale, il numero di autovalori nulli indica il numero di componenti connesse del grafo (un solo autovalore nullo per un grafo connesso) e le componenti diverse da zero dei relativi autovettori indicano i nodi appartenenti alle componenti connesse ($u_1 = (1, 1, ..., 1)$ per un grafo connesso).

Il secondo autovalore più piccolo (λ_2), prende il nome di *Fiedler number*, o *connettività algebrica*, e caratterizza la connettività e stabilità del grafo. Intuitivamente, più un grafo è connesso e maggiore è il valore del suo Fiedler number. Per esempio, un grafo di N nodi completamente connesso, in cui ogni nodo ha connectivity degree pari a $N - 1$, ha il Fiedler number maggiore tra tutti i grafi di N nodi. Perciò, il Fiedler number è una misura appropriata per associare ai dati Hi-C le proprietà di connettività della struttura cromatinica.

L'autovettore associato a λ_2 è chiamato *Fiedler vector*. Il pattern di nodi associati alle componenti positive e negative del Fiedler vector costituisce i *domini nodali* (figura 4.1-d). Dalla teoria spettrale dei grafi è noto che i domini nodali sono molto connessi al loro interno ed hanno meno connessioni con gli altri cluster.

Il laplaciano normalizzato (equazione 4.5) ha diversi vantaggi rispetto a quello non normalizzato (equazione 4.3). Lo spettro del laplaciano non normalizzato $\mathcal{L}$ è maggiormente influenzato dai nodi ad alta connettività. Questo fa sì che i nodi a maggiore connettività mascherino i nodi a bassa connettività, causando la perdita di informazione relativa a strutture complesse. Il laplaciano normalizzato $\mathcal{L}_\mathcal{N}$ riduce l'effetto maschera dei nodi a più alta connettività ed ha maggiore sensibilità per strutture complesse sottostanti, rendendo più omogenee le componenti a maggiore e minore connettività.

Per questo motivo, l'utilizzo del Fiedler number del laplaciano normalizzato risulta più appropriato per l'obiettivo biologico di individuare regioni localmente organizzate di geni co-regolati, cioè per caratterizzare pattern strutturali locali limitando l'influenza di altre regioni ad alta connettività presenti nello stesso cromosoma [28].

Inoltre, a differenza del laplaciano non normalizzato $\mathcal{L}$, il laplaciano normalizzato $\mathcal{L}_\mathcal{N}$ ha un Fiedler number limitato superiormente, ossia

$$\lambda_2 \leq \frac{N}{N-1} \leq 2. \tag{4.7}$$

In generale, invece di considerare connessioni binarie tra coppie di nodi (presenza o assenza del link), si ha a che fare con un *grafo pesato* se si associa un *peso* w_{ij} ad ogni link per caratterizzarne l'"intensità", ossia $A_{ij} = w_{ij}$ se $\{v_i, v_j\} \in E$, altrimenti $A_{ij} = 0$. Le matrici delle connettività D, del laplaciano $\mathcal{L}$ e del laplaciano normalizzato $\mathcal{L}_\mathcal{N}$ sono definite allo stesso modo che nelle equazioni 4.2-4.6.

Si rimanda al testo [40] per approfondimenti sulla teoria spettrale di grafi pesati.

4.4 Metodo computazionale Segmentazione spettrale iterativa del grafo associato alla matrice Hi-C tramite il laplaciano normalizzato

Chen et al. [28] hanno ideato un algoritmo efficiente, denominato *TAD_Laplace*, per individuare i domini topologici dai dati Hi-C. Gli autori hanno trovato che i TAD risultanti hanno una correlazione significativa con i dati di espressione genica (di un esperimento di RNA-seq, tecnica per l'analisi del trascrittoma), poiché i geni appartenenti ad uno stesso dominio si comportano in modo binario, tutti attivi o tutti inattivi, confermando quindi la relazione tra i TAD e la funzionalità del genoma. Effettuando un confronto con gli algoritmi proposti da Dixon et al. [12] e Filippova et al. [41], gli autori hanno trovato che *TAD_Laplace* è computazionalmente vantaggioso ed è maggiormente consistente con i CTCF enrichment peaks (di un esprimento ChIP-seq, metodo per analizzare le interazioni tra proteine e DNA). L'algoritmo individua quindi strutture più significative a livello funzionale.

I coefficienti di una matrice Hi-C definiscono un grafo pesato in cui i vertici corrispondono ai loci del genoma ed i pesi dei link (H_{ij}) sono proporzionali alle frequenze di contatto tra i loci. Si lega così la matrice Hi-C alla distanza 3D dei loci: i loci con alta frequenza di contatto sono associati ad una piccola distanza euclidea nello spazio 3D.

I TAD sono regioni di frequenza di contatto localmente elevata, separate da boundary in cui la frequenza di contatto cala. Per questo motivo, l'identificazione dei TAD può essere direttamente trasformata nel problema di segmentazione spettrale del grafo in componenti debolmente interconnesse. Nell'algoritmo *TAD_Laplace*, la partizione spettrale viene effettuata in modo iterativo, finché la connettività delle componenti del grafo associate ai domini raggiunge un valore impostato dall'utente.

La mappa di contatto Hi-C è una matrice non negativa (perché i coefficienti corrispondono ai conteggi dei contatti tra frammenti di DNA), a diagonale dominante e sparsa (poiché un frammento di DNA ha maggiore probabilità di legarsi a regioni adiacenti piuttosto che a zone distali). Inoltre, i segmenti delle regioni centromeriche non sono mappabili in modo unico nel genoma di riferimento, per la presenza di sequenze ripetute; per questo motivo, nella matrice Hi-C vi sono bande di valore 0, che vengono rimosse perché non sono informative.

Poiché i TAD sono regioni compatte spesso distinguibili come blocchi sulla diagonale della mappa Hi-C, essi corrispondono a componenti connesse del grafo

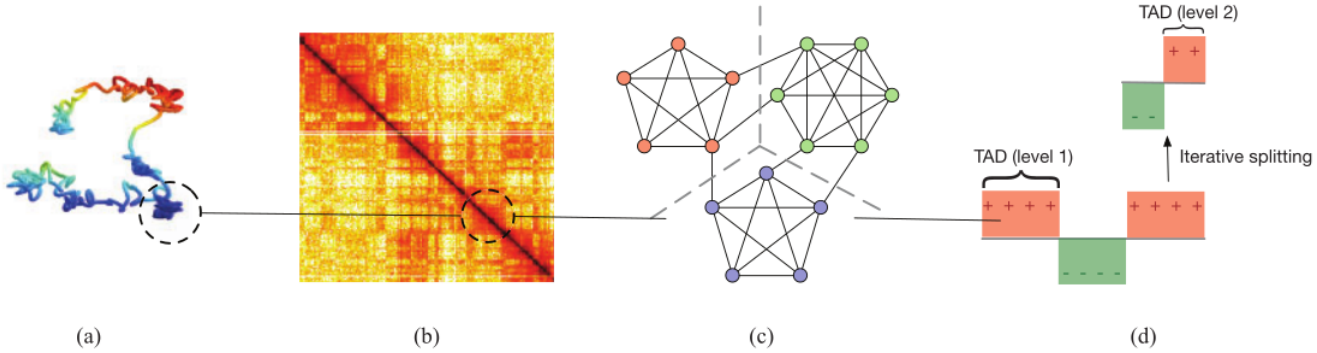

Figura 4.1: Rappresentazione dei TAD in modalità diverse: (a) regioni localmente compatte nella struttura fisica, (b) blocchi diagonali nella mappa Hi-C, (c) componenti altamente connesse nel modello a network (il problema dell'identificazione dei TAD diventa quindi quello di segmentazione del grafo nei punti di scarsa connettività), (d) domini nodali dal Fieldler vector del grafo (principio di base dell'algoritmo utilizzato).

con molte intra-connessioni e poche inter-connessioni (figura 4.1(a)-(c)).

La strategia di base dell'algoritmo *TAD_Laplace* consiste nell'impiego del Fiedler vector del laplaciano normalizzato della matrice Hi-C per segmentare il cromosoma in domini e del Fiedler number di ogni dominio per verificare se il dominio ottenuto è sufficientemente compatto o se necessita di un'ulteriore suddivisione.

Sia H la matrice Hi-C dei contatti osservati relativa ad un cromosoma di lunghezza L, con la diagonale e le regioni non mappabili rimosse.

Si denota con $\bar{H}$, la trasformata logaritmica di H

$$\bar{H}_{ij} = \begin{cases} 0, & i = j \\ log(H_{ij}), & i \neq j, \end{cases} \tag{4.8}$$

la quale riduce il range dinamico della matrice raw H (molto sparsa).

Considerando due loci appartenenti allo stesso cromosoma, la massima distanza possibile tra essi è pari alla lunghezza del DNA che li separa. Perciò, due loci i e j che risultano vicini rispetto alla catena 1D di cromatina, tendono ad avere un conteggio Hi-C H_{ij} elevato, indipendentemente dalla conformazione 3D del cromosoma (*effetto della distanza 1D*, capitolo 3).

Per rimuovere l'effetto della distanza 1D, l'algoritmo impiega la *normalizzazione di Toeplitz*. Essa divide il coefficiente (i, j) della matrice Hi-C per la media di tutti i coefficienti alla stessa distanza $|i-j|$ dalla diagonale. Matematicamente,

questo passaggio è rappresentato da

$$H_T = H \oslash T, \tag{4.9}$$

dove $\oslash$ rappresenta la divisione elemento per elemento dei coefficienti di H e T (*matrice di Toeplitz*); i coefficienti di T sono dati da

$$T_{kl} = \frac{1}{card(\mathcal{I}_{kl})} \sum_{m,n \in \mathcal{I}_{kl}} H_{mn}, \tag{4.10}$$

con l'insieme $\mathcal{I}_{kl} = \{m, n | m-n = k-l, \quad 0 < k, l < L\}$ che indica i coefficienti alla stessa distanza dalla diagonale di (k, l) e $card(\mathcal{I})$ che è pari al numero di elementi appartenenti all'insieme.

La matrice di Toeplitz T rappresenta la frequenza di contatto attesa in funzione della distanza 1D.

Chen et al. affermano che correzioni preliminari ulteriori non hanno generato differenze significative nei risultati [28].

Data una matrice di adiacenza A, si denota con

$$\lambda_2, v_2 \leftarrow Fv(A) \tag{4.11}$$

l'estrazione del Fiedler number λ_2 e del Fiedler vector v_2, effettuata tramite il calcolo dello spettro del laplaciano normalizzato $\mathcal{L}_{\mathcal{N}}$ di A, definito nell'equazione 4.5. Quando il valore del Fiedler number λ_2 non è rilevante, si utilizza la notazione $v_2 \leftarrow Fv(A)$.

4.4.1 Algoritmo *TAD_Laplace*

Per prima cosa, l'algoritmo estrae una prima serie di domini tramite il segno del Fiedler vector, poi suddivide in modo ricorsivo ciascun dominio finché il Fiedler number dei nuovi domini ottenuti supera un certo valore di soglia (λ_0) e la dimensione dei nuovi domini è sufficientemente piccola (minore della soglia ms_0 – *minimum splitting size*). Poiché il Fiedler number è proporzionale alla connettività algebrica, la soglia λ_0 assicura che si trovino domini sufficientemente compatti.

Al primo step (*Step 1*), si considera il grafo pesato associato ad un cromosoma tramite la sua mappa Hi-C e se ne calcola la matrice normalizzata con il metodo di Toeplitz, H_T. Si calcola il Fiedler vector di questo grafo, $v_2^{(1)}$, e si segmenta il network in cluster, in base al segno dei coefficienti del Fiedler vector. Si trova quindi un certo numero di strutture localmente compatte, costituite dagli insiemi di nodi a cui corrispondono componenti del Fiedler vector che hanno consecutivamente lo stesso segno (regione $\mathcal{D}_{i-j}$, definita dalle componenti $v_i^{(1)}$, $v_{i+1}^{(1)}$, ..., $v_j^{(1)}$

che hanno lo stesso segno). Questo primo risultato ottenuto da H_T considera l'organizzazione globale dei contatti della cromatina e la segmentazione risultante è simile ai compartimenti A e B individuati da Lieberman-Aiden [11]. Il confronto è riportato in modo approfondito nella sezione supplementary materials di Chen et al. [28].

Poiché i domini topologici esibiscono strutture organizzate gerarchicamente, dopo la determinazione della prima serie di domini dal Fiedler vector di H_T, tali domini vengono ulteriormente segmentati in sotto-domini di dimensioni inferiori (*Step 2*). Per un dominio iniziale $\mathcal{D}_{i-j}$, si calcola il Fiedler vector del grafo la cui matrice di adiacenza è data dalla sotto-matrice di $\bar{H}$ individuata dagli indici di $\mathcal{D}_{i-j}$; si divide poi $\mathcal{D}_{i-j}$ in sotto-domini in base al segno delle componenti del Fiedler vector. In questo step si utilizza la matrice trasformata logaritmica $\bar{H}$, perché l'individuazioni di questi sotto-domini più piccoli è basata sulle strutture a blocchi nella diagonale di $\bar{H}$, più che sulle interazioni a lungo range evidenziate da H_T. Si calcola il Fiedler number dei sotto-domini ottenuti e lo si confronta con il valore di soglia predefinito λ_0, per determinare se sono sufficientemente compatti o se devono essere ulteriormente suddivisi.

La complessità computazionale dell'algoritmo è dominata dal calcolo dello spettro del laplaciano normalizzato. Lo *Step 1* effettua questa operazione su matrici di dimensioni abbastanza elevate (4,985×4,985 per il cromosoma 1 alla risoluzione di 50 kb, 963×963 per il cromosoma 21 alla risoluzione di 50 kb), ma le iterazioni successive lavorano su matrici di dimensioni molto inferiori, riducendo in modo significativo il tempo necessario per la decomposizione spettrale. Per motivi di efficienza, viene calcolato soltanto lo spettro relativo ai primi due autovalori più piccoli.

Parametri: soglia del Fiedler number λ_0 e limite inferiore per la dimensione dei domini ms_0.

Pre-processing: data la matrice raw Hi-C di un cromosoma (H), si calcolino H_T (equazione 4.9) e $\bar{H}$ (equazione 4.8).

Step 1: calcolo del Fiedler vector della matrice H_T

$$v_2^{(1)} \leftarrow Fv(H_T); \tag{4.12}$$

i TAD iniziali ($\mathcal{D}_{i-j}$) sono dati dai nodi a cui corrispondono componenti contigue di $v_2^{(1)}$ con lo stesso segno.

Step 2: per ogni dominio $\mathcal{D}_{i-j}$ ottenuto, si calcolano il Fiedler number ed il

Fiedler vector della sotto-matrice di $\bar{H}$ associata

$$\lambda_2, v_2 \leftarrow Fv(\bar{H}_{\mathcal{D}_{i-j}}); \tag{4.13}$$

se il Fiedler number è inferiore alla soglia pre-impostata, $\lambda_2 \leq \lambda_0$, si segmenta il dominio tramite il segno delle componenti di v_2.

Iterazione: si ripete lo *Step 2* finché il sotto-dominio ottenuto ha un Fiedler number maggiore del valore di soglia, o se ha dimensioni inferiori al limite ms_0.

L'output dell'algoritmo è un vettore ($TAD\_Boundaries$) che contiene le posizioni (in bin) dei boundary individuati.

4.5 Metodi statistici

4.5.1 Conservazione dei TAD boundary tra i sette tipi cellulari

I TAD boundary di un cromosoma, relativamente ad una linea cellulare, possono essere rappresentati da un vettore binario TB_{bin} di dimensione pari alla lunghezza del cromosoma considerato (L),

$$TB_{bin} = \begin{cases} 0, & \text{se non c'è un boundary} \\ 1, & \text{se c'è un boundary.} \end{cases} \tag{4.14}$$

Quindi, ad una certa risoluzione, per ogni cromosoma si hanno sette vettori del genere (uno per ogni tipo cellulare). Il vettore s_{sum}, dato dalla somma di questi sette vettori binari, è quindi di lunghezza L ed è definito dall'equazione

$$s_{sum} = \begin{cases} 0, & \text{se non c'è nessun boundary in quel bin} \\ 1, & \text{se 1 cellula ha un boundary in quel bin} \\ 2, & \text{se 2 cellule hanno un boundary in quel bin} \\ 3, & \text{se 3 cellule hanno un boundary in quel bin} \\ 4, & \text{se 4 cellule hanno un boundary in quel bin} \\ 5, & \text{se 5 cellule hanno un boundary in quel bin} \\ 6, & \text{se 6 cellule hanno un boundary in quel bin} \\ 7, & \text{se 7 cellule hanno un boundary in quel bin.} \end{cases} \tag{4.15}$$

Il confronto tra la distribuzione di s_{sum} e del suo analogo random (di seguito illustrato) permette di verificare l'eventuale conservazione dei TAD boundary tra diversi tipi cellulari.

4.5.2 Modello nullo
Random reshuffling

Per verificare la significatività dei risultati ottenuti dalle matrici Hi-C dei contatti osservati, si è utilizzato come modello nullo il random reshuffling delle posizioni dei TAD boundary. Ad esempio, se un cromosoma di un tipo cellulare presenta un certo numero di TAD boundary (pari a $\#TB$), il suo analogo nel modello nullo è dato dai seguenti passaggi

- generazione di $\#TB$ numeri interi casuali, compresi tra 1 ed il numero di bin mappabili del cromosoma
- riordinamento dei numeri casuali in modo crescente
- conversione dei numeri generati nel range $[0, L]$, con L lunghezza del cromosoma in esame, facendo in modo che non vengano considerate le regioni non mappabili.

4.5.3 Distribuzione binomiale

Volendo confrontare la distribuzione di s_{sum} (equazione 4.15) con la distribuzione della sua controparte nel modello nullo, per ogni risoluzione e per ogni cromosoma, sono stati generati sette vettori (uno per ogni tipo cellulare) del tipo TB_{bin} (equazione 4.14), la cui somma dà il vettore s_{rand}, descritto analogamente dall'equazione 4.15.

La distribuzione multinomiale è quella che meglio dovrebbe fittare la distribuzione di s_{rand}, poiché tale vettore è modellizzabile come l'esito di una sequenza di prove indipendenti, ciascuna con diverse probabilità di successo p ($p = \frac{\#TB}{L}$, con $\#TB$ numero di TAD boundary nel caso reale, valore diverso per ogni tipo cellulare, e L lunghezza del cromosoma considerato).

Dato che il numero di TAD boundary non varia in modo importante tra i tipi cellulari, si è deciso di lavorare con la distribuzione binomiale per descrivere l'andamento della distribuzione di s_{rand}. Quest'ultima è modellizzabile come un *esperimento di Bernoulli*, ossia come una variabile casuale $s_n = x_1 + x_2 + ... + x_n$, somma di n variabili aleatorie indipendenti, che assumono due soli valori, 0 e 1, detti anche $fallimento$ e *successo*. In generale, si dice esperimento di Bernoulli una sequenza di n prove con le seguenti caratteristiche:

- il risultato di ogni prova può essere solamente successo o fallimento (1 o 0)
- il risultato di ciascuna prova è indipendente dai risultati delle prove precedenti

- la probabilità p di successo, e quindi la probabilità $q = 1 - p$ di fallimento, sono costanti in ciascuna prova.

In realtà, come detto, nel caso considerato l'ultimo requisito non è soddisfatto, perchè $p = \frac{\#TB}{L}$ assume un valore diverso per ogni tipo cellulare. Per questo motivo, per ogni cromosoma, si considera la probabilità media $\overline{p} = \frac{\overline{\#TB}}{L}$, con $\overline{\#TB}$ numero di boundary medio nei vari tipi cellulari.

Poiché si tratta di eventi indipendenti e si applica la regola della probabilità composta, la probabilità che in n prove di un esperimento di Bernoulli si abbiano k successi $(\overbrace{SS\ ...\ S}^{k\ volte}\overbrace{FF\ ...\ F}^{n-k\ volte})$ è data da $\overbrace{pp\ ...\ p}^{k\ volte}\overbrace{qq\ ...\ q}^{n-k\ volte} = p^k q^{(n-k)}$.

Siccome una qualsiasi altra sequenza di n prove con k successi ha la stessa probabilità (cambia soltanto l'ordine dei fattori, che non è rilevante) ed esistono $C_{n,k}$ (coefficiente binomiale) sequenze di n prove con k successi, la probabilità totale è data dall'equazione

$$B(n,p) = C_{n,k}\ p^k q^{(n-k)} = \binom{n}{k} p^k q^{(n-k)} = \frac{n!}{k!(n-k)!}\ p^k q^{(n-k)}, \tag{4.16}$$

con $q = 1-p$. Nel caso in questione, $p = \overline{p} = \frac{\overline{\#TB}}{L}$ ($\overline{\#TB}$ numero di boundary medio nei vari tipi cellulari, per uno specifico cromosoma), e $n = 7$.

4.5.4 Test del χ^2

Per verificare l'appartenenza di due campioni ad una stessa distribuzione, si è utilizzato il test del χ^2.

In generale, questo test statistico verifica il grado di adattamento delle frequenze reali (ricavate da una distribuzione reale) a quelle teoriche (relative ad una distribuzione teorica). Si supponga che in un particolare campione si osservi che un insieme di possibili eventi $E_1, E_2, \ldots, E_k$ si presenti con frequenze $o_1,\ o_2, \ldots, o_k$ (frequenze osservate), e che, secondo le regole della probabilità, ci si attenda che si presenti con frequenze $e_1, e_2, \ldots, e_k$ (frequenze teoriche o attese). La variabile test χ^2 si ottiene sommando per ogni evento E_i il quadrato degli scarti tra le frequenze osservate e quelle teoriche, pesato sulle frequenze teoriche,

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}; \tag{4.17}$$

questa si distribuisce come una variabile χ^2 con $k-1$ gradi di libertà. Ovviamente, se $\chi^2 = 0$, le frequenze osservate coincidono esattamente con quelle teoriche. Se invece $\chi^2 > 0$, più grande è il suo valore e maggiore è la discrepanza

tra frequenze teoriche ed osservate. Fissando un valore critico χ^2_{crit}, corrispondente ad un certo livello di confidenza, se il valore del χ^2 calcolato è superiore a quello critico, le discrepanze tra le frequenze osservate e quelle attese sono significative e le frequenze provengono da popolazioni diverse ($\chi^2 > \chi^2_{crit}$, si accetta l'ipotesi alternativa di distribuzioni significativamente diverse), se è inferiore non è possibile concludere che i campioni non appartengono alla stessa distribuzione ($\chi^2 < \chi^2_{crit}$, non è possibile rifiutare l'ipotesi nulla di appartenenza alla stessa distribuzione).

Nel contesto del lavoro di tesi, il test χ^2 è stato applicato alle distribuzioni dei vettori s_{sum} e s_{rand} (equazione 4.15); quindi, in questo caso il numero di gradi di libertà è 7 ($k-1$, $k=8$ numero di valori che i due vettori possono assumere) ed il valore critico corrispondente al livello di confidenza $\alpha = 0.05$ è $\chi^2_{crit} = 15.51$.

4.5.5 Grado di similarità tra i pattern di TAD di diversi tipi cellulari
Coefficiente di Jaccard

Per misurare in modo quantitativo il grado di similarità tra le posizioni dei TAD boundary tra diversi tipi cellulari, si è utilizzata la metrica definita dal *coefficiente di Jaccard*.

In generale, dati due oggetti, A e B, ciascuno con n attributi binari (0 o 1), il coefficiente di Jaccard è una misura della sovrapposizione di A e B in base ai loro attributi. Il numero totale di occorrenze di ogni combinazione possibile degli attributi di A e B è dato da

- M_{11}, numero totale di attributi dove A e B hanno entrambi valore 1
- M_{10}, numero totale di attributi dove l'attributo di A è 1 e l'attributo di B è 0
- M_{01}, numero totale di attributi dove l'attributo di A è 0 e l'attributo di B è 1
- M_{00}, numero totale di attributi dove A e B hanno entrambi valore 0.

Ogni attributo deve ricadere in una di queste quattro categorie, quindi $M_{11} + M_{01} + M_{10} + M_{00} = n$. Il coefficiente di Jaccard è pari a

$$J = \frac{M_{11}}{M_{01} + M_{10}}. \tag{4.18}$$

Dati due vettori binari x_1 e x_2 di n componenti, il loro coefficiente di Jaccard è espresso da

$$J = \frac{\#\{(x_{1j} = x_{2j}) \cap [(x_{1j} \neq 0) \cup (x_{2j} \neq 0)]\}}{\#[(x_{1j} \neq 0) \cup (x_{2j} \neq 0)]}. \tag{4.19}$$

Ad una certa risoluzione, per ogni cromosoma, le posizioni dei TAD boundary ottenuti per ciascun tipo cellulare possono essere espresse nella forma di vettori binari, TB_{bin}, definiti in equazione 4.14. Come detto, TB_{bin} è un vettore di dimensione pari alla lunghezza del cromosoma considerato, di valore 1 nei bin in cui è presente un boundary.

Il grado di similarità delle posizioni dei boundary tra coppie di tipi cellulari è quindi misurato da J (equazione 4.19), pari alla frazione delle componenti dei vettori TB_{bin} diverse da zero che coincidono.

Tale metrica risulta essere la più appropriata per lo studio in questione, poichè associa a due vettori binari TB_{bin} della stessa lunghezza un grado di similarità tanto maggiore quanto più le posizioni dei boundary sono coincidenti.

Ad una certa risoluzione, per un dato cromosoma, disponendo delle posizioni dei TAD boundary relative ai sette tipi cellulari, si calcolano 21 coefficienti Jaccard, uno per ogni coppia di tipi cellulari ($7 \times 6 \ / \ 2 = 21$).

Alla risoluzione di 50 kb, si è ritenuto opportuno considerare una sorta di variabilità statistica nella posizione dei TAD boundary, poichè a questa scala è possibile che i risultati dell'algoritmo *TAD_Laplace* siano affetti da rumore e che i boundary di due tipi cellulari siano coincidenti anche se non si trovano esattamente nello stesso bin, ma in bin adiacenti (primi vicini). Per questo motivo, si è deciso di associare alla posizione di un boundary un certo numero di bin di incertezza. Il valore scelto a questa risoluzione è Δ=3 bin, che è pari allo 0.1% della lunghezza del cromosoma 1, il più lungo, e allo 0.7% del cromosoma 21, il più corto. Perciò, i boundary di due tipi cellulari che distano al più 3 bin, sono stati giudicati coincidenti.

4.5.6 Z-score

Per verificare la significatività del grado di similarità tra le posizioni dei TAD boundary tra diversi tipi cellulari, i coefficienti di Jaccard ricavati come appena illustrato sono stati confrontati con i valori corrispondenti del modello nullo (random reshuffling).

Per fare ciò, ad una certa risoluzione, per ogni cromosoma e per ogni tipo cellulare, sono stati generati con il metodo precedentemente illustrato dieci vettori di TAD boundary disposti in modo casuale, con numero di boundary pari al numero di boundary del caso reale. In questo modo, per ogni cromosoma e per

ogni coppia di tipi cellulari, si sono ottenuti dieci valori del coefficiente di Jaccard. Allora, calcolando la media di questi dieci valori ($\bar{J}_{rand}$) e la loro deviazione standard ($\sigma_{J_{rand}}$), è possibile associare ad ogni coefficiente di Jaccard reale J uno Z-score, definito come

$$Z = \frac{J - \bar{J}_{rand}}{\sigma_{J_{rand}}}. \tag{4.20}$$

Lo Z-score è quindi una misura di quanto un valore osservato si discosta dalla media della distribuzione del modello nullo, esprimendo in particolare il numero di deviazioni standard comprese tra il valore osservato e la media della distribuzione nulla.

Ad una certa risoluzione, si ottiene una matrice 23×21 (#cromosomi × #coppie di cellule) di Z-score. Mediando sui cromosomi, i valori di Z-score risultanti associati alle coppie di tipi cellulari sono stati utilizzati per fare un ranking delle coppie di linee cellulari più simili tra loro.

La procedura appena illustrata è stata eseguita sui dati alla risoluzione di 50 kb e di 1 Mb, ritenute maggiormente informative.

Capitolo 5

Risultati

Sono stati individuati i TAD boundary tramite l'algoritmo *TAD_Laplace* di Chen et al. [28], per segmentazione spettrale iterativa del laplaciano normalizzato associato alla mappa intra-cromosomiale Hi-C, sui dati di sette tipi cellulari umani (linee cellulari GM12878, HMEC, HUVEC, IMR90, K562, KBM7 e NHEK), a tre risoluzioni, 50 kb, 500 kb ed 1 Mb.

Applicando i metodi computazionali e statistici discussi nel capitolo 4, l'analisi dei dati si è incentrata sulla verifica dell'eventuale conservazione dei TAD boundary tra tipi cellulari diversi e sullo studio delle proprietà topologiche del network associato alla mappa di contatto Hi-C.

5.1 Confronto tra metodi di normalizzazione

Nella fase preliminare del lavoro, l'algoritmo *TAD_Laplace* è stato provato in diverse modalità, sia variando il valore dei parametri di input sia utilizzando diversi metodi di normalizzazione, per studiarne l'effetto sui risultati ottenuti.

Grazie a questa analisi iniziale, si è deciso di seguire il metodo di Chen et al., gli autori dell'algoritmo utilizzato, che utilizza la matrice normalizzata con il metodo di Toeplitz allo *Step 1* e la trasformata logaritmica della matrice Hi-C sul resto delle iterazioni.

É emerso che non utilizzando la normalizzazione di Toeplitz allo *Step 1* l'algoritmo individua un numero minore di TAD boundary. Ciò conferma il ruolo della normalizzazione di Toeplitz, la quale riduce l'effetto della distanza 1D e permette quindi di individuare pattern strutturali 3D locali con rilevanza biologica.

In generale, l'impiego della trasformata logaritmica della matrice Hi-C nelle iterazioni successive allo *Step 1* genera invece l'aumento del numero di TAD boundary identificati. La trasformata logaritmica infatti approccia una sorta di normalizzazione che diminuisce il range dinamico dei dati (molto ampio nelle ma-

trici Hi-C, tipicamente sparse) ed è quindi possibile che tale operazione permetta all'algoritmo di individuare in modo più efficace i pattern locali d'interesse, poiché riduce l'effetto maschera delle regioni a connettività maggiore (ossia con conteggi Hi-C più elevati) a spese delle regioni più debolmente connesse.

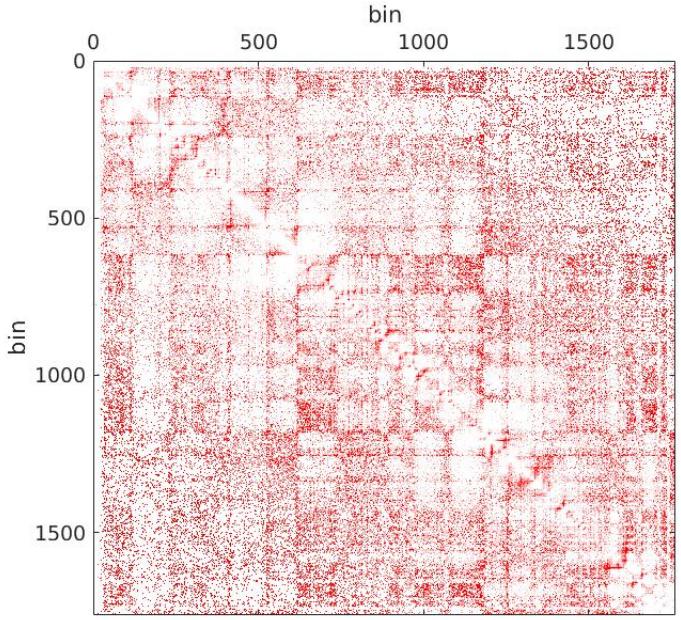

Figura 5.1: Mappa di contatto normalizzata con il metodo di Toeplitz (utilizzato nell'algoritmo *TAD_Laplace*). Cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb.

Nelle figure 5.1-5.4 è rappresentata la mappa Hi-C del cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb, normalizzata con diversi metodi: quello di Toeplitz, utilizzato nell'agoritmo *TAD_Laplace*, quello di Knight e Ruiz, impiegato da Rao et al. [1], il vanilla coverage, utilizzato da Lieberman-Aiden et al.[11], e lo square root vanilla coverage. Tali modalità di normalizzazione sono state approfondite nei capitoli 3 e 4. La mappa normalizzata con il metodo di Toeplitz (figura 5.1) è la sola tra le quattro considerate a non essere stata visualizzata in scala logaritmica.

Risulta evidente che la normalizzazione di Toeplitz è l'unica che riduce drasticamente l'intensità della diagonale della matrice, mettendo quindi in evidenza le interazioni a lungo range, come atteso. Viene quindi confermata la funzione di questo metodo di normalizzazione, che permette di individuare in modo efficace cluster di interazioni a lungo range, riducendo l'effetto della distanza 1D tipico dei polimeri e della catena cromatinica.

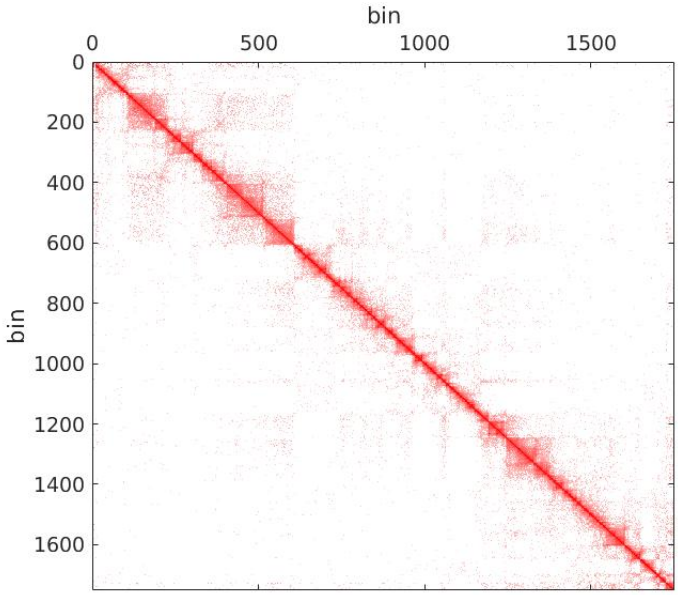

Figura 5.2: Mappa di contatto normalizzata con il metodo di Knight e Ruiz (impiegato da Rao et al.). Cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb.

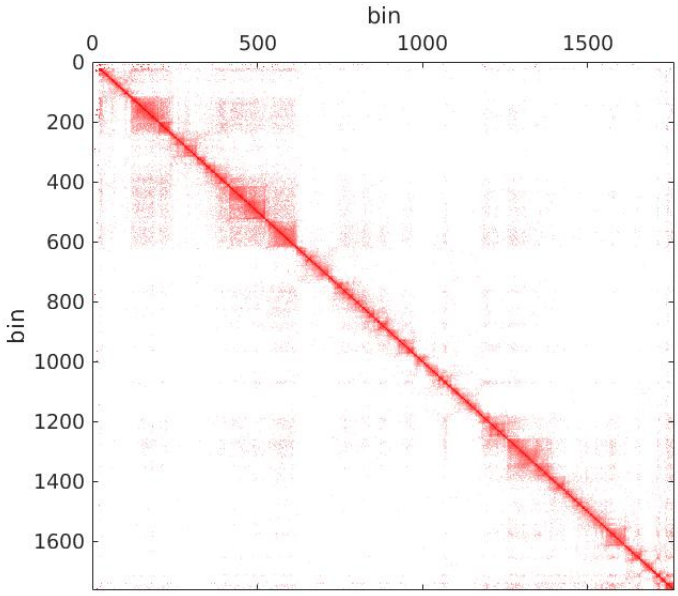

Figura 5.3: Mappa di contatto normalizzata con il metodo vanilla coverage (utilizzato da Lieberma-Aiden et al.). Cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb.

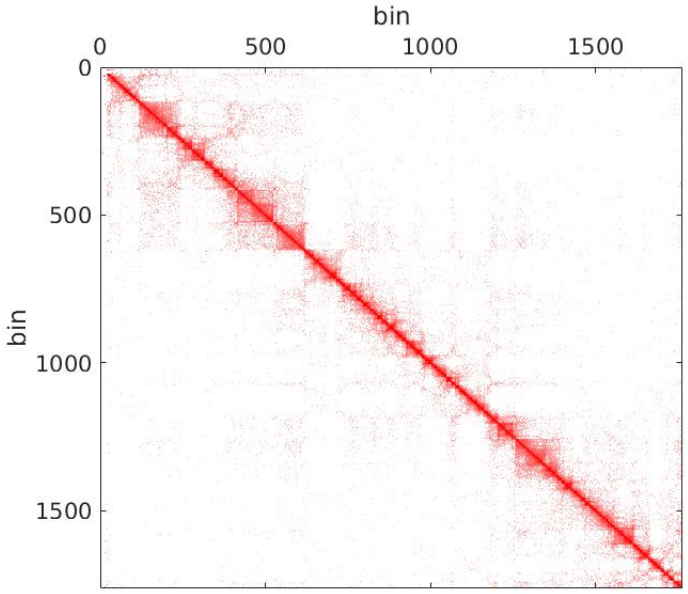

Figura 5.4: Mappa di contatto normalizzata con il metodo square root vanilla coverage (SQRTVC). Cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb.

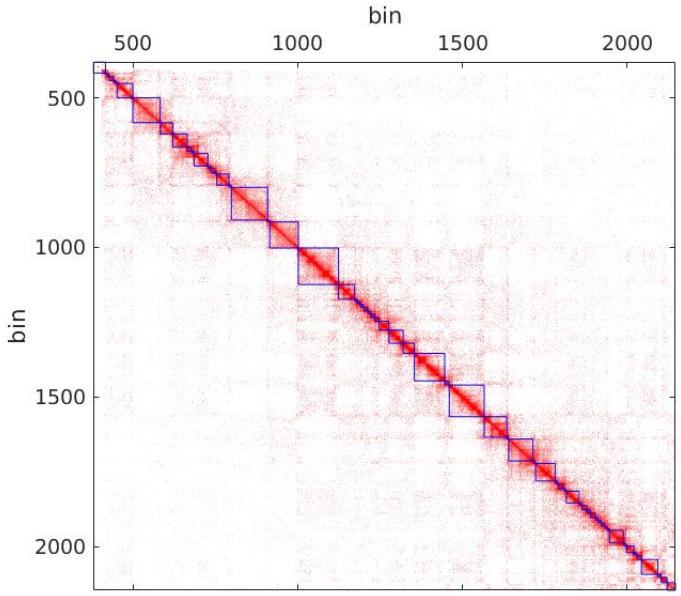

Figura 5.5: Mappa di contatto raw del cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb, con sovrapposti i TAD boundary individuati dall'algoritmo $TAD\_Laplace$ con la normalizzazione di Toeplitz ($\lambda_0 = 0.8$, $ms_0 = 1000$).

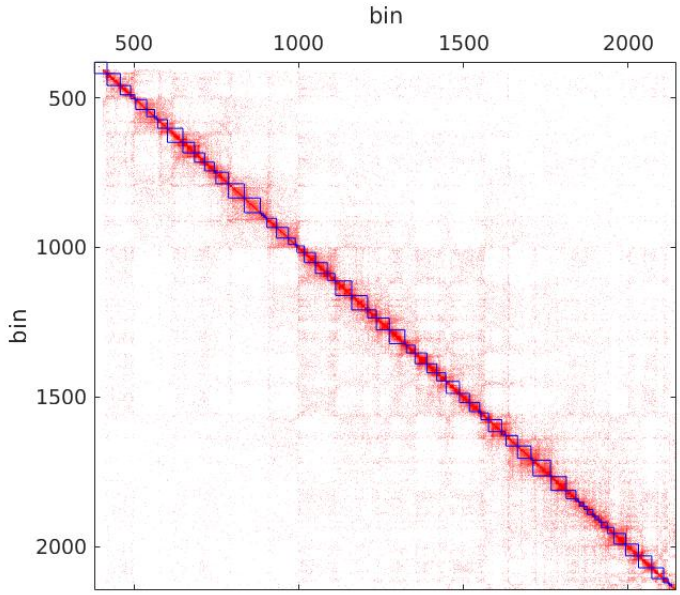

Figura 5.6: Mappa di contatto raw del cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb, con sovrapposti i TAD boundary individuati dall'algoritmo *TAD_Laplace* con la normalizzazione di Knight e Ruiz ($\lambda_0 = 0.8$, $ms_0 = 50$).

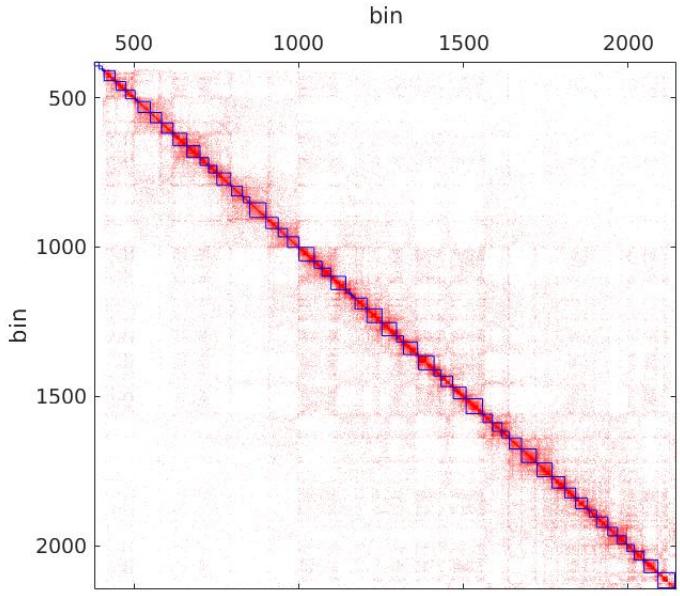

Figura 5.7: Mappa di contatto raw del cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb, con sovrapposti i TAD boundary individuati dall'algoritmo *TAD_Laplace* con la normalizzazione vanilla coverage ($\lambda_0 = 0.8$, $ms_0 = 50$).

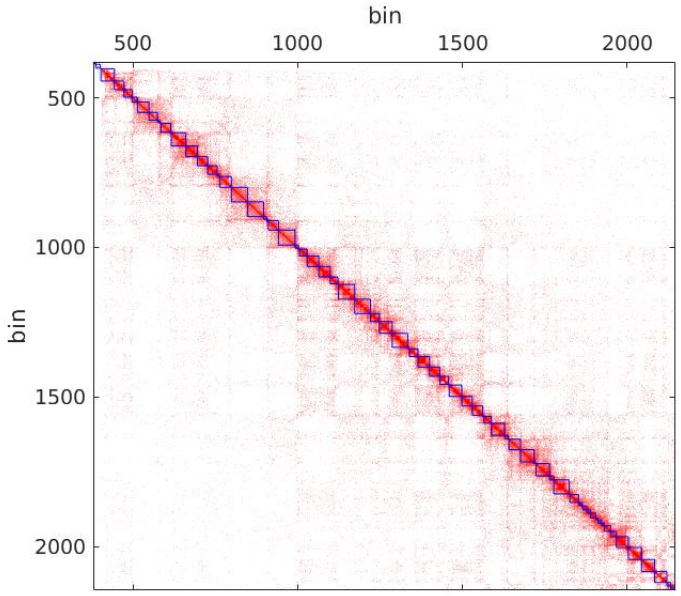

Figura 5.8: Mappa di contatto raw del cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb, con sovrapposti i TAD boundary individuati dall'algoritmo $TAD\_Laplace$ con la normalizzazione square root vanilla coverage ($\lambda_0 = 0.8$, $ms_0 = 50$).

In più, rispetto alle altre normalizzazioni è emerso che, a parità di parametri di soglia (λ_0 e ms_0), l'algoritmo $TAD\_Laplace$ utilizzato sulla matrice normalizzata con il metodo di Toeplitz produce un numero di TAD boundary maggiore, ossia dà luogo ad un clustering più "profondo" che coglie un'organizzazione strutturale più fine, su scale inferiori. Ciò significa che per ottenere lo stesso livello di clustering, con la normalizzazione di Toeplitz sono necessarie meno iterazioni.

Nelle figure 5.5-5.8 è riportata la mappa di contatto raw del cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb, con sovrapposti dei riquadri blu che evidenziano i blocchi diagonali identificati dall'algoritmo $TAD\_Laplace$ come TAD. La differenza nel binning delle figure 5.1-5.4 e 5.5-5.8 è dovuta al fatto che nelle prime sono state rimosse le regioni non mappabili (regioni centromeriche), che nel cromosoma 14 corrispondono ai primi 380 bin.

Sia da un riscontro visivo, sia tramite analisi della distribuzione delle dimensioni dei TAD ottenuti, emerge che rispetto agli altri casi la normalizzazione di Toeplitz genera dei TAD di dimensioni meno omogenee, con la presenza aggiuntiva di TAD di dimensioni molto piccole. Ciò è particolarmente vero in corrispondenza dei bin 430÷480, 1170÷1250 e 1850÷1950, in cui la normalizzazione di Toeplitz dà luogo a serie di TAD di dimensioni molto inferiori rispetto ai blocchi adiacenti (figura 5.5). Infatti, nelle figure 5.9 e 5.10 sono riportati in dettaglio i

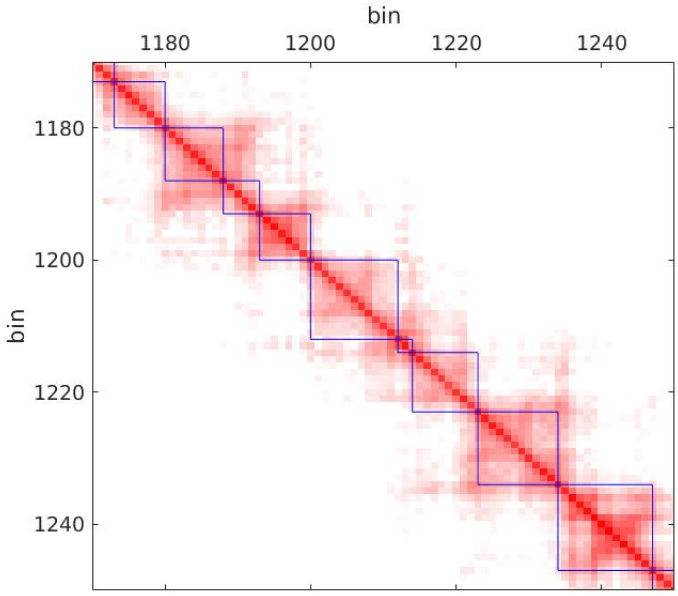

Figura 5.9: TAD boundary individuati dall'algoritmo *TAD_Laplace* con la normalizzazione di Toeplitz ($\lambda_0 = 0.8$, $ms_0 = 1000$), bin 1170÷1250. Cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb.

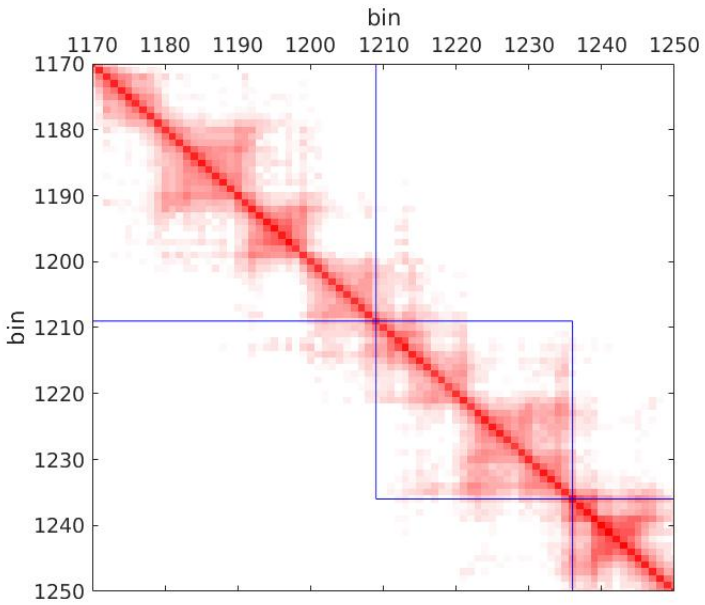

Figura 5.10: TAD boundary individuati dall'algoritmo *TAD_Laplace* con la normalizzazione di Knight e Ruiz ($\lambda_0 = 0.8$, $ms_0 = 50$), bin 1170÷1250. Cromosoma 14 della linea cellulare HMEC, alla risoluzione di 50 kb.

TAD individuati dai metodi di Toeplitz e di Knight e Ruiz nei bin 1170÷1250, da cui si vede chiaramente che la normalizzazione di Toeplitz individua cluster strutturali localmente interagenti di dimensioni inferiori, spingendosi su scale più piccole.

Si noti che nel caso in cui si è utilizzato il metodo di Toeplitz (figura 5.5 e 5.9), il parametro di soglia ms_0 è venti volte più grande ($ms_0 = 1000$, invece che $ms_0 = 50$), diminuendo quindi il numero di iterazioni necessarie a parità di risultato.

Nell'analisi svolta a seguire, sulle orme degli ideatori dell'algoritmo utilizzato, si è impiegata la normalizzazione di Toeplitz allo *Step 1* e la trasformata logaritmica della matrice Hi-C sul resto delle iterazioni. I parametri di soglia λ_0 e ms_0 sono stati variati in base alle dimensioni del cromosoma specifico e del grado di clustering desiderato.

5.2 Conservazione dei TAD boundary tra tipi cellulari

Come detto, sono stati individuati i TAD boundary tramite l'algoritmo di Chen et al. nei 23 cromosomi di sette linee cellulari umane, GM12878, HMEC, HUVEC, IMR90, K562, KBM7 e NHEK, a tre risoluzioni, 50 kb, 500 kb ed 1Mb.

Le mappe Hi-C del tipo cellulare GM12878 sono le uniche ad essere state generate a partire dalla risoluzione di 1 kb, mentre la risoluzione di partenza delle altre linee cellulari è di 5 kb. Per questo motivo, le mappe di GM12878 sono quelle con maggiore densità di conteggi e sulla base di ciò Rao et al. affermano che in generale su di esse si trova un numero di TAD maggiore [1]. Nel lavoro di tesi, questo fenomeno non viene riscontrato, a conferma del fatto che l'algoritmo ed il metodo di normalizzazione utilizzati non risentono del bias sperimentale in questione.

Nelle figure 5.11-5.13 sono rappresentati i TAD boundary del cromosoma 1 dei sette tipi cellulari, alle tre risoluzioni considerate. Visivamente si riscontra che vi sono boundary che tendono ad essere conservati tra diverse linee cellulari.

Per verificare in modo quantitativo la conservazione delle posizioni dei boundary tra tipi cellulari differenti, ad ogni risoluzione si è analizzata la distribuzione del vettore s_{sum}, definito dall'equazione 4.15 e descritto in dettaglio nel capitolo 4, confrontandola con la distribuzione di s_{rand}, vettore analogo ad s_{sum} nel modello nullo di random reshuffling (ridistribuzione casuale delle posizioni dei TAD boundary, si veda il capitolo 4).

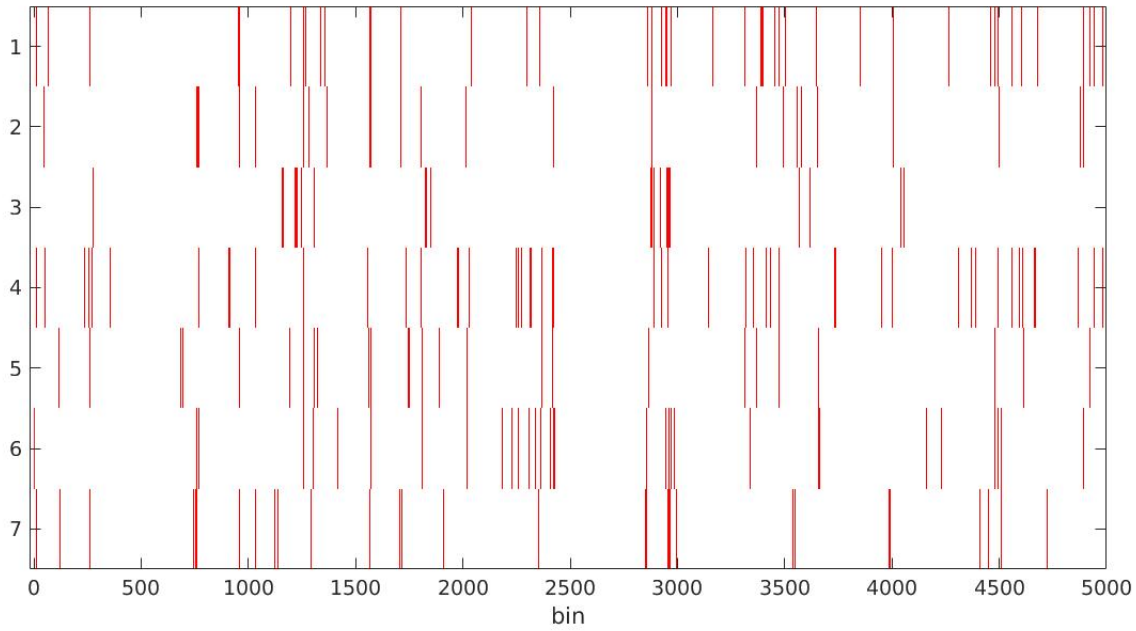

Figura 5.11: Risoluzione di 50 kb. TAD boundary individuati nel cromosoma 1 delle sette linee cellulari analizzate: 1) GM12878, 2) HMEC, 3) HUVEC, 4) IMR90, 5) K562, 6) KBM7 e 7) NHEK.

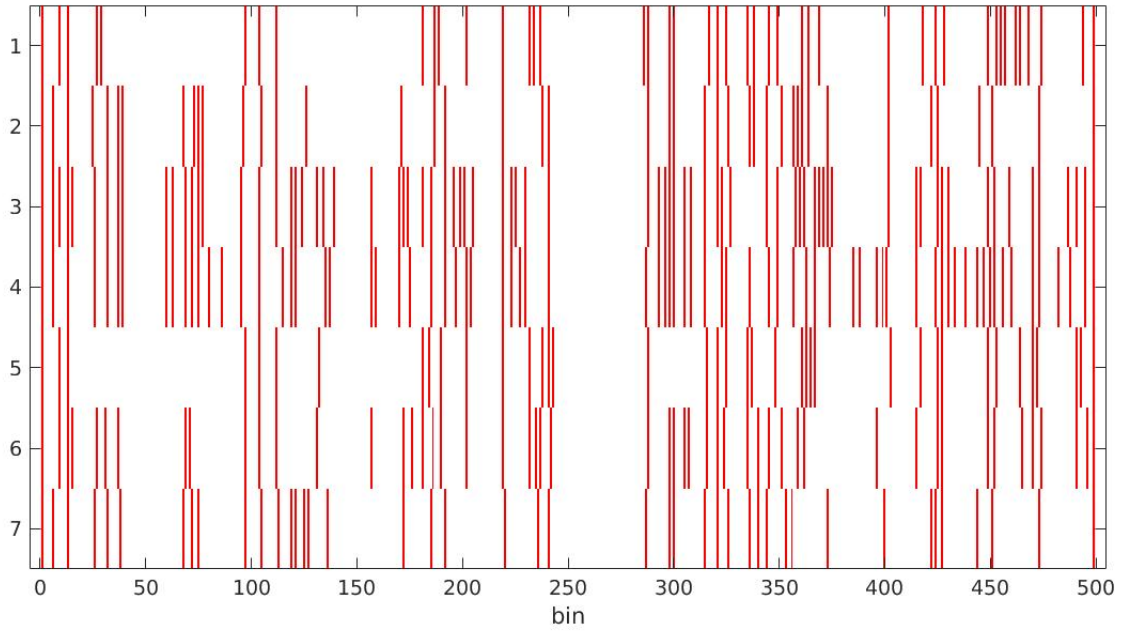

Figura 5.12: Risoluzione di 500 kb. TAD boundary individuati nel cromosoma 1 delle sette linee cellulari analizzate: 1) GM12878, 2) HMEC, 3) HUVEC, 4) IMR90, 5) K562, 6) KBM7 e 7) NHEK.

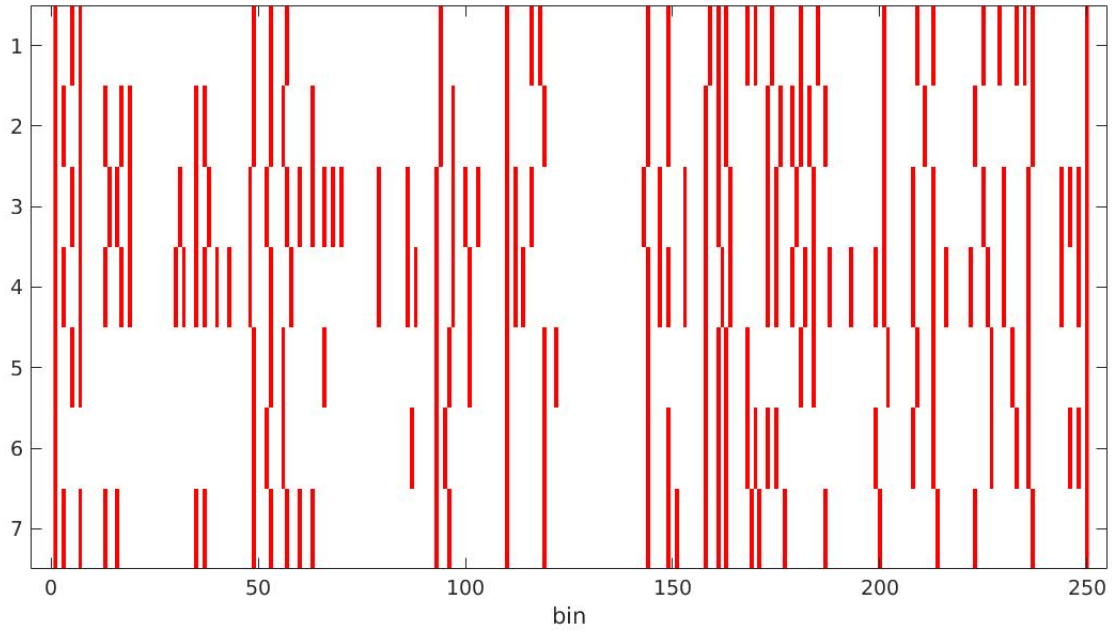

Figura 5.13: Risoluzione di 1 Mb. TAD boundary individuati nel cromosoma 1 delle sette linee cellulari analizzate: 1) GM12878, 2) HMEC, 3) HUVEC, 4) IMR90, 5) K562, 6) KBM7 e 7) NHEK.

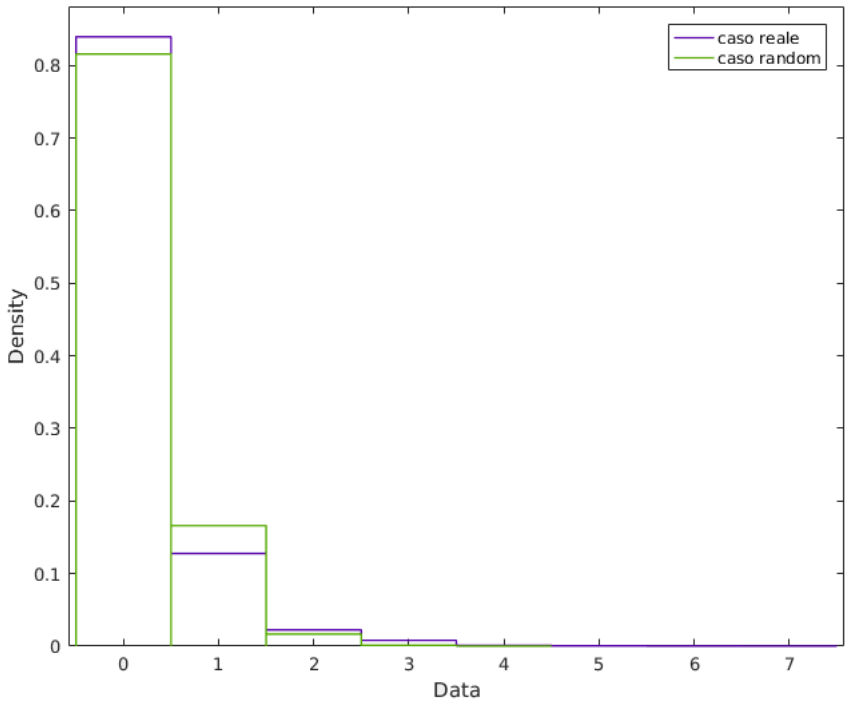

Figura 5.14: Risoluzione di 50 kb. Confronto tra le distribuzioni di s_{sum} ed s_{rand}, per il cromosoma 1.

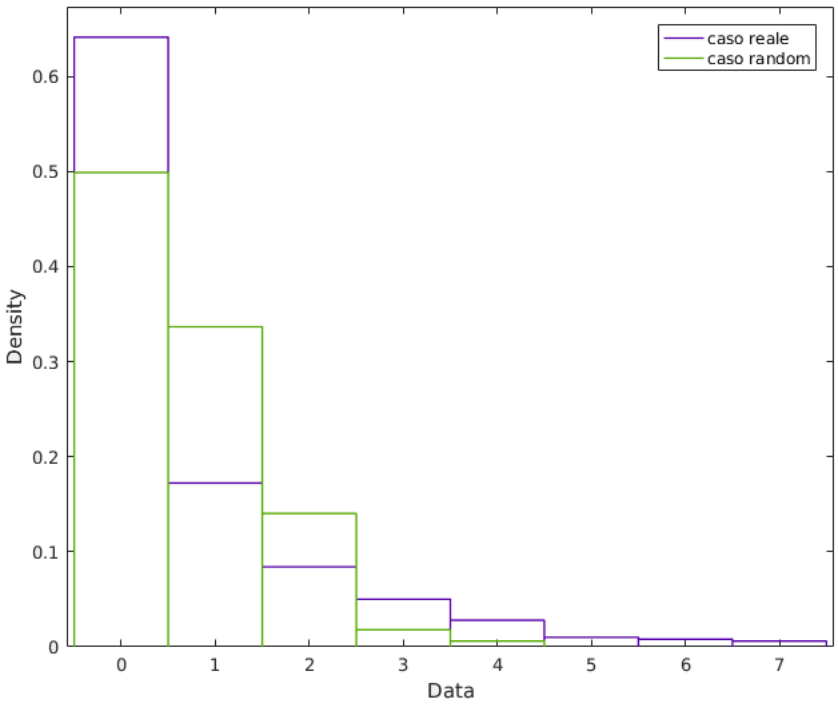

Figura 5.15: Risoluzione di 500 kb. Confronto tra le distribuzioni di s_{sum} ed s_{rand}, per il cromosoma 1.

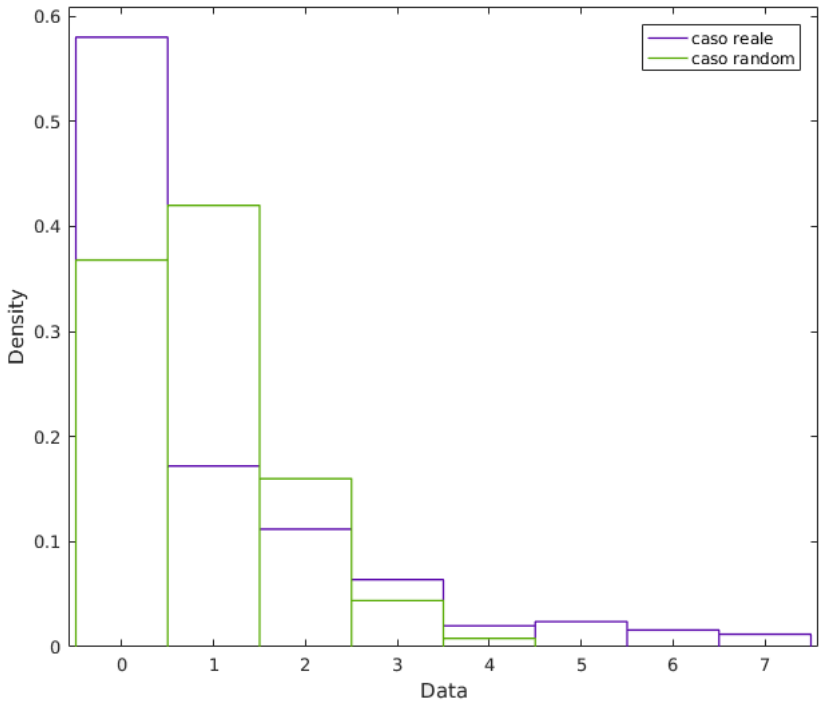

Figura 5.16: Risoluzione di 1 Mb. Confronto tra le distribuzioni di s_{sum} ed s_{rand}, per il cromosoma 1.

Nelle figure 5.14-5.16 è riportato il confronto tra le due distribuzioni in questione (distribuzione reale e distribuzione random) per il cromosoma 1, alle tre risoluzioni studiate. In generale, come atteso, nel caso reale la frequenza dei valori estremi (0 e 4÷7) è maggiore e quella dei valori intermedi (1÷3) è minore, dimostrando che vi è un overlap significativo tra le posizioni dei boundary di diverse linee cellulari.

Effettuando un test del χ^2 tra la distribuzione reale e la distribuzione random, a tutte e tre le risoluzioni e per tutti i cromosomi, si ha che il campione reale non è consistente con la distribuzione del random reshuffling ($\chi^2 > \chi^2_{crit}$, si rigetta l'ipotesi nulla al 5% di significatività).

In più, un test del χ^2 effettuato tra la distribuzione del vettore s_{rand} (equazione 4.15) ed il modello teorico binomiale $B(7, \bar{p})$, descritto nel capitolo 4, ha dimostrato che le due distribuzioni sono consistenti (per tutti i cromosomi, a tutte le risoluzioni, si ha che $\chi^2 > \chi^2_{crit}$ al livello di significatività del 5%). La variabile s_{rand} è quindi modellizzabile come una variabile casuale $s_7 = x_1 + x_2 + ... + x_7$, che somma 7 variabili aleatorie indipendenti che assumono due possibili valori, 0 o 1, con stessa probabilità di successo (ossia di assumere il valore 1). In realtà, a rigore, la distribuzione multinomiale sarebbe stata più appropriata, poichè i tipi cellulari hanno un numero di boundary diverso (quindi la probabilità di successo nei singoli "esperimenti" è diversa). Tuttavia, l'esito del test del χ^2 dimostra che vale l'approssimazione di distribuzione binomiale $B(7, \bar{p})$.

Quindi in sintesi, i risultati ottenuti sono significativamente diversi dal modello nullo di random reshuffling, il quale è consistente con il modello teorico descritto dalla distribuzione binomiale. In tutte le linee cellulari, a tutte le risoluzioni considerate, si riscontra un overlap significativo delle posizioni dei TAD boundary. Ciò è a favore dell'ipotesi per cui i TAD boundary non siano distribuiti casualmente, ma vi sia una tendenza ad una disposizione regolare, conservata tra i tipi cellulari.

5.3 Grado di similarità tra tipi cellulari

Per fare una stima quantitativa del grado di similarità tra le posizioni dei TAD boundary di diversi tipi cellulari si è utilizzata la metrica definita dal coefficiente di Jaccard, descritto nel capitolo 4, pari alla frazione di boundary coincidenti tra coppie di tipi cellulari. I coefficienti di Jaccard ottenuti sono stati confrontati con i valori corrispondenti del modello nullo di random reshuffling, ottenendo degli Z-score che ne verificano la significatività (capitolo 4).

Alla risoluzione di 1 Mb, la percentuale media di coincidenza delle posizioni dei TAD boundary è di 31.1±5.7% (coefficiente di Jaccard medio e sua deviazione

standard).

Alla risoluzione di 50 kb, associando alla posizione dei boundary una variabilità statistica pari a Δ=3 bin come descritto nel capitolo 4, per cui si considerano spazialmente coincidenti i boundary che distano al più 3 bin, si è ottenuta una percentuale media di coincidenza dei boundary del 31.9±5.0%.

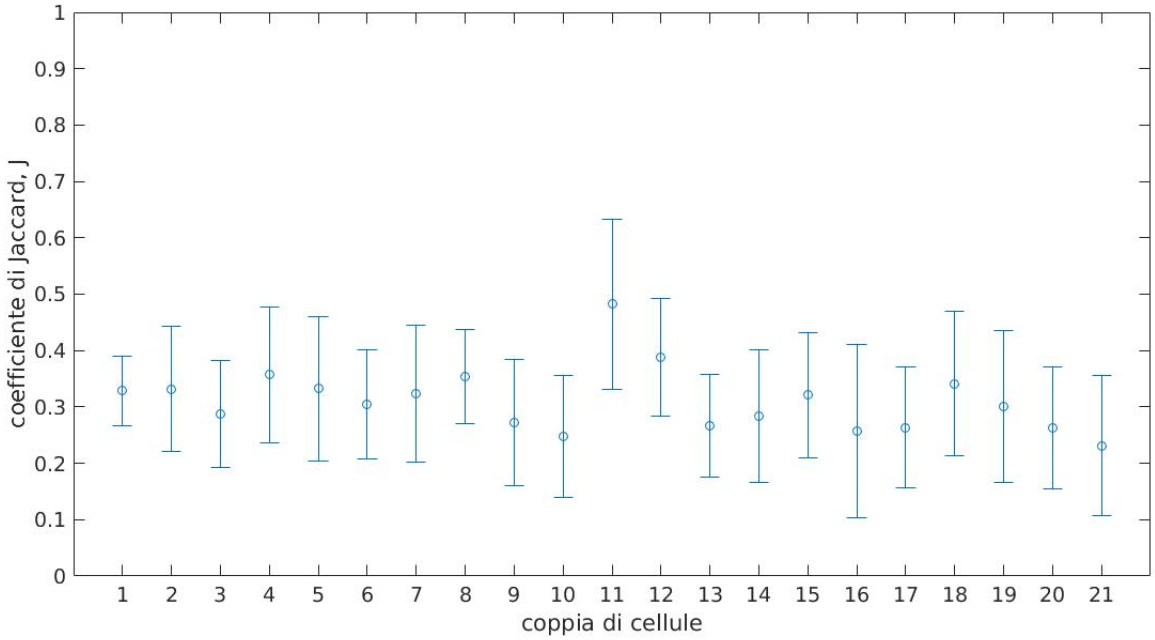

Figura 5.17: Risoluzione di 1 Mb. Coefficiente di Jaccard (frazione) per ogni coppia di tipi cellulari; valore medio sui 23 cromosomi ± deviazione standard. Il significato degli indici dell'asse x è illustrato in tabella 5.1.

Nelle figure 5.17 e 5.18, è rappresentato il coefficiente di Jaccard per ogni coppia di tipi cellulari, rispettivamente alle risoluzioni di 1 Mb e 50 kb.

In tabella 5.1 si specifica per ogni coppia di linee cellulari l'indice corrispondente indicato nell'asse x delle due figure.

In entrambi i casi, si vede immediatamente che le cellule HMEC e NHEK (indice 11), HUVEC e IMR90 (indice 12) sono quelle con maggior grado di similarità.

Confrontando tramite valori Z-score i coefficienti di Jaccard ottenuti con i corrispondenti valori del modello nullo, generato dal random reshuffling come descritto nel capitolo 4, si ha che alla risoluzione di 1 Mb le coppie di cellule più simili tra loro, con Z-score decrescente, sono

1. HMEC e NHEK (indice 11), $J \pm \sigma = 48 \pm 15\%$, $Z = 8.2$
2. GM12878 e K562 (indice 4), $J \pm \sigma = 36 \pm 12\%$, $Z = 6.7$

x label	**coppia di cellule**
1	GM12878 - HMEC
2	GM12878 - HUVEC
3	GM12878 - IMR90
4	GM12878 - K562
5	GM12878 - KBM7
6	GM12878 - NHEK
7	HMEC - HUVEC
8	HMEC - IMR90
9	HMEC - K562
10	HMEC - KBM7
11	HMEC - NHEK
12	HUVEC - IMR90
13	HUVEC - K562
14	HUVEC - KBM7
15	HUVEC - NHEK
16	IMR90 - K562
17	IMR90 - KBM7
18	IMR90 - NHEK
19	K562 - KBM7
20	K562 - NHEK
21	KBM7 - NHEK

Tabella 5.1: Coppie di tipi cellulari corrispondenti agli indici 1÷21 dell'asse x delle figure 5.17 e 5.18.

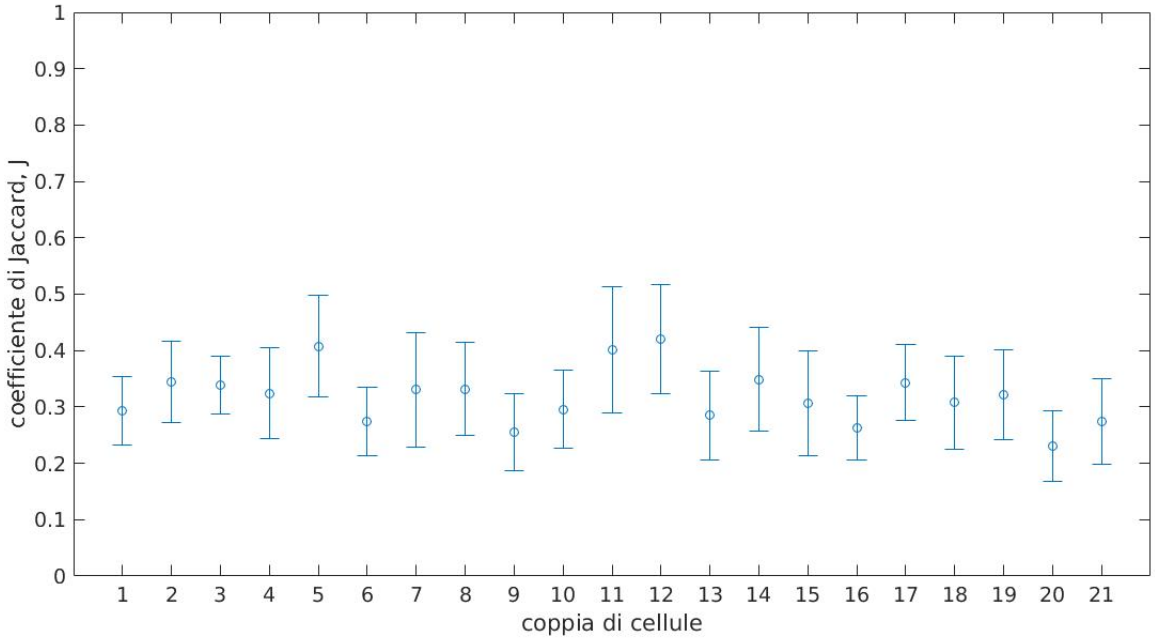

Figura 5.18: Risoluzione di 50 kb. Coefficiente di Jaccard (frazione) per ogni coppia di tipi cellulari; valore medio sui 23 cromosomi ± deviazione standard. Il significato degli indici dell'asse x è illustrato in tabella 5.1.

3. HUVEC e IMR90 (indice 12), $J \pm \sigma = 39 \pm 10\%$, $Z = 6$
4. GM12878 e KBM7 (indice 5), $J \pm \sigma = 33 \pm 13\%$, $Z = 5.8$;

analogamente, alla risoluzione di 50 kb le cellule con maggior grado di similarità, con Z-score decrescente, sono

1. HMEC e NHEK (indice 11), $J \pm \sigma = 40 \pm 11\%$, $Z = 14.3$
2. HUVEC e IMR90 (indice 12), $J \pm \sigma = 42 \pm 10\%$, $Z = 13.3$
3. GM12878 e KBM7 (indice 5), $J \pm \sigma = 41 \pm 9\%$, $Z = 13.3$
4. GM12878 e K562(indice 4), $J \pm \sigma = 32 \pm 8\%$, $Z = 12.1$.

É interessante notare che il maggior grado di similarità tra i pattern di TAD di queste cellule potrebbe avere basi biologiche, in quanto

- HMEC, cellule epiteliali mammarie, e NHEK, cheratinociti, provengono entrambe dall'epitelio
- la linea HUVEC è costituita da cellule endoteliali del cordone ombelicale e le cellule IMR90 sono fibroblasti polmonari prelevati da un feto

- le cellule GM12878 sono linfociti B del sistema immunitario, le cellule K562 sono linfoblasti, cellule staminali precursori dei linfociti B, prelevate da un individuo malato di leucemia mieloide cronica, e le KBM7 sono cellule somatiche quasi aploidi isolate da un paziente affetto dalla stessa patologia.

5.4 Analisi delle dimensioni dei TAD alle diverse risoluzioni

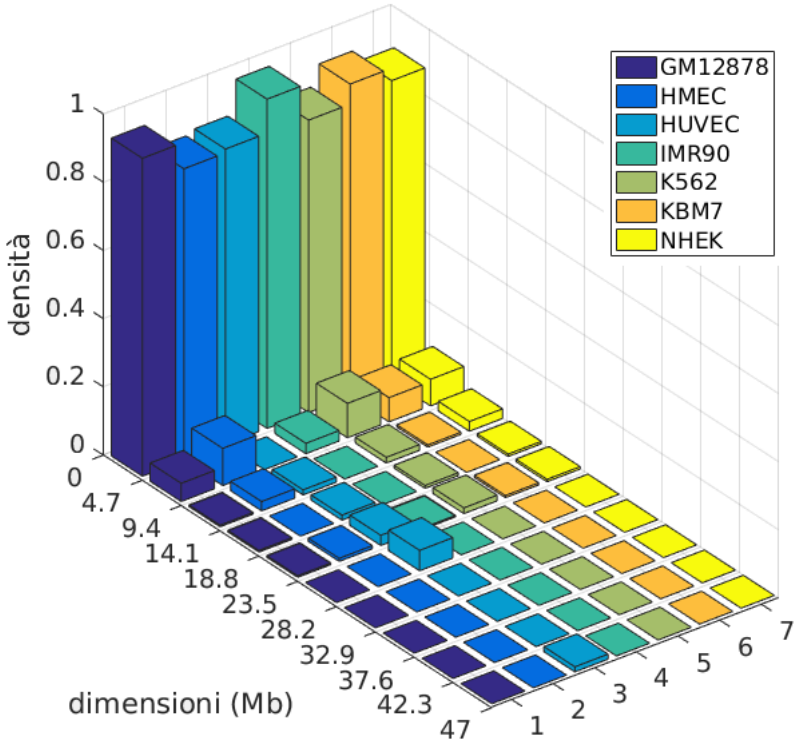

Figura 5.19: Risoluzione a 50 kb. Cromosoma 1, distribuzione delle dimensioni dei domini per i sette tipi cellulari.

Come risulta dalle figure 5.19-5.21, le distribuzioni delle dimensioni dei TAD sono molto skewed, soprattutto alla risoluzione di 50 kb. In generale, si ha una piccola frazione di domini di grandi dimensioni, mentre la maggior parte dei TAD è di dimensioni inferiori.

Si riportano nelle tabelle 5.2 e 5.3 le dimensioni mediane dei cromosomi 1 e 21 (il cromosoma più grande e quello più piccolo) di ciasciun tipo cellulare, alle diverse risoluzioni considerate.

Ovviamente, al crescere della risoluzione vi è la tendenza ad individuare TAD di dimensioni inferiori. Le dimensioni mediane dei domini individuati nel cromosoma 1 sono

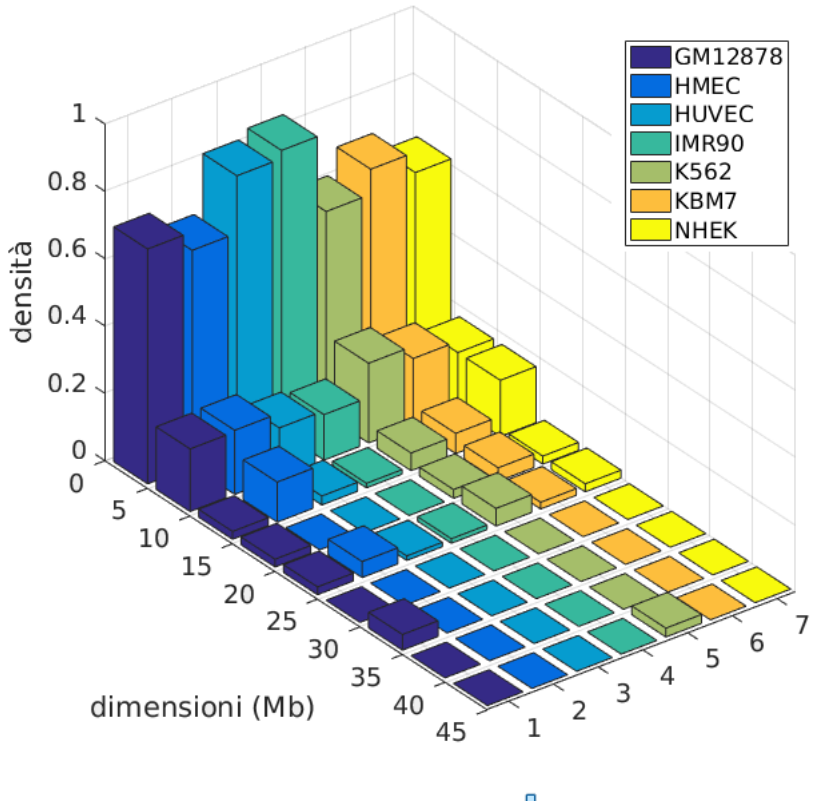

Figura 5.20: Risoluzione a 500 kb. Cromosoma 1, distribuzione delle dimensioni dei domini per i sette tipi cellulari.

cellula	**mediana (Mb)**		
	ris. 50 kb	**ris. 500 kb**	**ris. 1 Mb**
GM12878	0.35	3	4
HMEC	0.7	3.5	5
HUVEC	0.25	2	4
IMR90	0.45	2.5	4
K562	0.5	3.75	5
KBM7	0.65	2.75	5
NHEK	0.4	3.75	6
media (Mb)	**0.47±0.16**	**3±0.8**	**4.7±0.8**

Tabella 5.2: Cromosoma 1, dimensione mediana dei domini nei sette tipi cellulari, alle tre risoluzioni considerate. Nell'ultima riga vi è il valore mediato sulle linee cellulari, con variabilità statistica σ (deviazione standard).

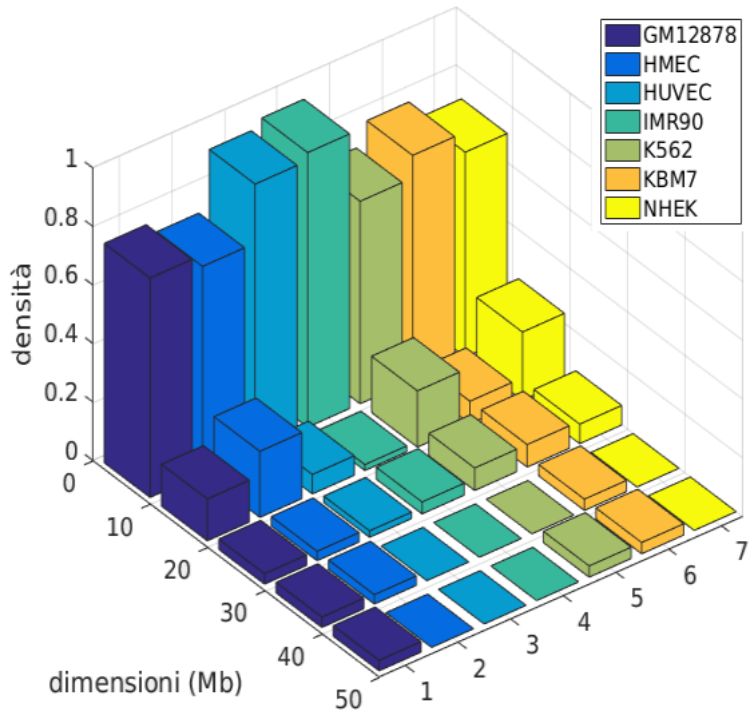

Figura 5.21: Risoluzione a 1 Mb. Cromosoma 1, distribuzione delle dimensioni dei domini per i sette tipi cellulari.

cellula	**mediana (Mb)**		
	ris. 50 kb	**ris. 500 kb**	**ris. 1 Mb**
GM12878	0.35	2	4
HMEC	0.5	2.5	3.5
HUVEC	0.45	2	6.5
IMR90	0.3	2.5	4.5
K562	0.35	3.75	5
KBM7	0.35	3	5
NHEK	0.5	3	4.5
media (Mb)	**0.4±0.08**	**2.7±0.5**	**4.7±1.1**

Tabella 5.3: Cromosoma 21, dimensione mediana dei domini nei sette tipi cellulari, alle tre risoluzioni considerate. Nell'ultima riga vi è il valore mediato sulle linee cellulari, con variabilità statistica σ (deviazione standard).

- 470 kb, con i dati alla risoluzione di 50 kb
- 3 Mb, con i dati alla risoluzione di 500 kb
- 4.7 Mb, con i dati alla risoluzione di 1 Mb,

e nel cromosoma 21

- 400 kb, con i dati alla risoluzione di 50 kb
- 2.7 Mb, con i dati alla risoluzione di 500 kb
- 4.7 Mb, con i dati alla risoluzione di 1 Mb.

É interessante che nonostante il cromosoma 1 sia circa cinque volte più lungo del cromosoma 21, per ogni risoluzione i loro TAD tendono ad avere dimensioni dello stesso ordine di grandezza. Ciò significa che le dimensioni dei TAD tendono ad essere indipendenti dalla lunghezza specifica del cromosoma.

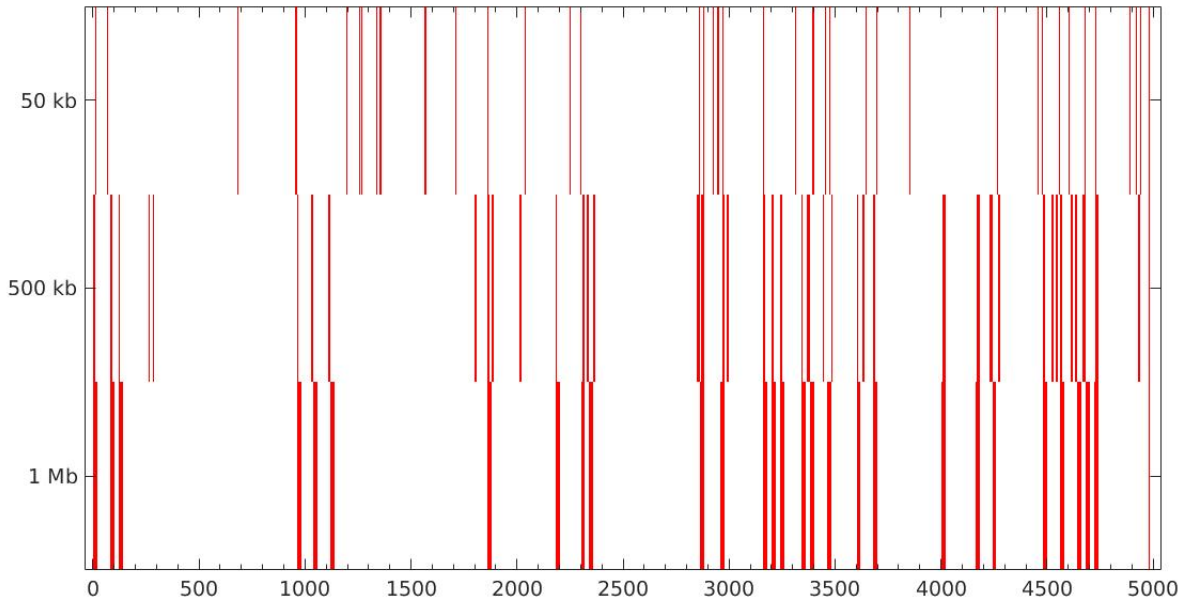

Figura 5.22: Cromosoma 1, TAD boundary individuati alle tre risoluzioni, 50 kb, 500 kb ed 1 Mb. Lo spessore variabile dei segmenti che indicano i boundary è dovuto alla diversa dimensione dei bin alle varie risoluzioni.

Nelle figure 5.22 e 5.23 sono raffigurati i TAD boundary identificati dall'algoritmo $TAD\_Laplace$ alle tre risoluzioni, per i cromosomi 1 e 21.

L'analisi multirisoluzione effettuata permette di individuare pattern di TAD che riflettono la natura gerarchica dell'organizzazione strutturale della cromatina.

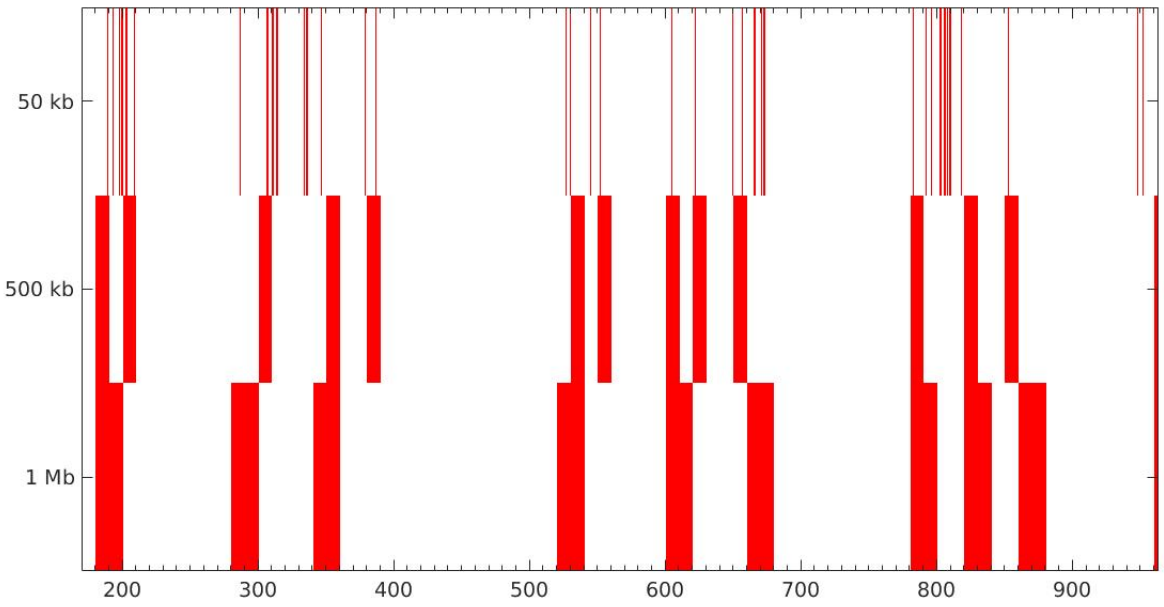

Figura 5.23: Cromosoma 21, TAD boundary individuati alle tre risoluzioni, 50 kb, 500 kb ed 1 Mb. Lo spessore variabile dei segmenti che indicano i boundary è dovuto alla diversa dimensione dei bin alle varie risoluzioni.

Capitolo 6

Conclusioni

Il recente progresso tecnologico, da un punto di vista sia sperimentale, per la generazione di dati tramite la tecnica di genome-wide conformation capture (Hi-C), sia computazionale, per l'integrazione di dati ad alta dimensionalità, permette di indagare l'architettura cromatinica su diverse scale spaziali, facendo luce sulla complessa relazione che intercorre tra conformazione 3D della cromatina, attività genica e stato funzionale della cellula.

É noto che la cromatina possiede una struttura dinamica organizzata in modo gerarchico su diverse scale, che mette in stretta relazione il complesso processo di folding del DNA e la regolazione genica. Si ritiene infatti che le configurazioni topologiche della cromatina non siano casuali ed uno dei più ardui ed importanti obiettivi della ricerca scientifica nell'ambito dell'architettura nucleare è lo studio dei meccanismi responsabili della loro formazione e delle loro implicazioni funzionali.

Su scale spaziali comprese tra 100 kb e 5 Mb circa, la cromatina è organizzata in domini topologici (TAD – Topological Associated Domains) con rilevanza biologica; è stato infatti dimostrato che essi sono correlati all'espressione genica [1][28]. Si tratta di porzioni di DNA ad alta interazione locale, quindi piccola distanza spaziale, corrispondenti a blocchi sulla diagonale della matrice Hi-C (porzioni di matrice con conteggi più elevati).

Il lavoro di tesi si è incentrato sullo studio multirisoluzione dell'organizzazione in TAD della cromatina di cellule umane e sulla verifica dell'eventuale conservazione di questi pattern strutturali tra tipi cellulari diversi.

Per fare ciò, si è utilizzato l'algoritmo *TAD_Laplace* ideato da Chen et al. [28], che identifica i TAD boundary tramite segmentazione spettrale iterativa del laplaciano normalizzato associato alla mappa di contatto intra-cromosomiale Hi-C. I domini individuati dal Fiedler vector del grafo associato, infatti, corrispondono a cluster di nodi altamente connessi, connessi più debolmente al resto

del network, e sono quindi spontaneamente associati ai TAD.

I dati Hi-C utilizzati sono relativi a sette linee cellulari umane, GM12878, HMEC, HUVEC, IMR90, K562, KBM7 e NHEK [39], e sono state considerate tre risoluzioni differenti, 50 kb, 500 kb ed 1 Mb.

Nella fase preliminare del lavoro, è emerso che il metodo di normalizzazione impiegato, ossia la normalizzazione di Toeplitz, riduce il bias dell'effetto della distanza 1D, tipico delle catene polimeriche e della catena cromatinica, ed evidenzia quindi le interazioni a lungo range d'interesse, che hanno rilevanza biologica. Rispetto ad altri metodi di normalizzazione, quello di Toeplitz riduce il numero di iterazioni necessarie, risultando quindi vantaggioso anche da un punto di vista computazionale, e genera un clustering più fine, su scale inferiori. In più, risulta che la trasformata logaritmica della mappa Hi-C riduce l'effetto maschera delle regioni ad alta connettività a spese di quelle più debolmente connesse, migliorando l'identificazione di pattern strutturali 3D locali.

A tutte le risoluzioni analizzate, si è riscontrata una sovrapposizione significativa tra le posizioni dei TAD boundary di tipi cellulari diversi, a conferma dell'ipotesi per cui i TAD boundary non siano distribuiti casualmente ma vi sia la tendenza ad una disposizione regolare, conservata tra tipi cellulari, generata da pattern strutturali funzionalmente rilevanti.

La percentuale media di coincidenza tra le posizioni dei TAD boundary dei diversi tipi cellulari è di 31.1±5.7% (coefficiente di Jaccard ± sua deviazione standard) alla risoluzione di 1 Mb e di 31.9±5.0% alla risoluzione di 50 kb.

In base ai valori di Z-score derivati dal confronto con il corrispondente modello nullo di random reshuffling (ridistribuzione casuale delle posizioni dei boundary), risulta che i i tipi cellulari più simili tra loro sono HMEC-NHEK, HUVEC-IMR90 e GM12878-K562-KBM7. É interessante notare che queste linee cellulari hanno affinità anche da un punto di vista biologico e funzionale.

L'analisi delle dimensioni dei TAD ha mostrato che, come atteso, al diminuire della risoluzione dei dati utilizzati la dimensione mediana dei TAD dei cromosomi aumenta: ∼450 kb alla risoluzione di 50 kb, ∼2.8 Mb alla risoluzione di 500 kb e ∼4.7 Mb alla risoluzione di 1 Mb.

Infine si è trovato che, per ogni risoluzione, i cromosomi presentano TAD di dimensioni dello stesso ordine di grandezza, nonostante la lunghezza dei cromosomi vari in modo significativo. Ciò fa pensare che le dimensioni dei domini siano indipendenti dalla lunghezza specifica del cromosoma.

Diversi studi hanno dimostrato le interessanti possibilità offerte in campo biomedico dall'informazione genomica ricavata tramite la tecnica Hi-C, rivelando che le alterazioni nella conformazione della cromatina e nella regolazione genica sono fortemente legate al cancro, alla differenziazione e allo sviluppo cellu-

lare. Le sfide attualmente intraprese dalla comunità scientifica, che richiedono ulteriori progressi sperimentali e computazionali, coinvolgono la caratterizzazione della variabilità della struttura cromatinica tra singole cellule (single-cell Hi-C) [18], la generazione di mappe di contatto e strutture cromosomiche specifiche per l'aplotipo e l'integrazione di dati genomici ed epigenomici per studiare il legame tra genotipo e fenotipo cellulare [42].

Si pensa di proseguire lo studio iniziato in questa tesi lungo due direzioni principali. (1) Chiarire la rilevanza biologica dei TAD boundary maggiormente conservati tra tipi cellulari diversi e cercare di darne un'interpretazione funzionale anche in relazione alla diversità dei tipi cellulari. (2) Utilizzare metodi della teoria dei network complessi per lo studio e l'analisi dei grafi (matrici di adiacenza) associate alle mappe Hi-C.

Ringraziamenti

Ringrazio i Professori Gastone Castellani e Daniel Remondini, per aver contribuito in modo essenziale alla stesura della mia tesi, al conseguimento del titolo di studi e per avermi dato opportunità che ho sempre desiderato.

Un grazie quotidiano va alla mia famiglia, solida radice. Più di tutti grazie ai miei genitori, che mi hanno permesso di seguire pienamente i miei interessi più veri. Grazie a mia sorella Gaia che quasi ogni giorno si becca la me al naturale, nel bene e nel male, e che nonostante questo sceglie di esserci sempre.

Grazie agli amici e personaggi che hanno costituito il mio panorama in questi ultimi anni, con presenza costante o con brevi comparse. Chi leggendo si sente coinvolto, fa bene a farlo. Lo sai se hai partecipato o stai partecipando, e ti sto ringraziando, sì.

Infine, non che sia meno importante del resto, anzi, ringrazio l'energia che mi spinge a farmi le domande che mi faccio e a muovermi per tentare di rispondervi. Ho l'insaziabile fascino del mistero del mondo e della vita e vivrò per questo.

Bibliografia

[1] S.S.P. Rao, M.H. Huntley, N.C. Durand, E.K. Stamenova, I.D. Bochkov, J.T. Robinson, A.L. Sanbom, I. Machol, A.D. Omer, E.S. Lander, E. Lieberman Aiden, A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping, *Cell 159* (2014) 1665-1680.

[2] J. Dekker, K. Rippe, M. Dekker, N. Kleckner, Capturing Chromosome Conformation, *Science 295* (2002) 1306-1311.

[3] G. Li, L. Cai, H. Chang, P. Hong, Q. Zhou, E.V. Kulakova, Y. Ruan, Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application, *BMC Genomics* (2014) 15(Suppl 12):S11.

[4] P.C. Taberlay, J. Achinger-Kawecka, A.T. Lun, F.A. Buske, K. Sabir, C.M. Gould, E. Zotenko, S.A. Bert, K.A. Giles, D.C. Bauer, G.K. Smyth, C. Stirzaker, S.I. O'Donoghue, S.J. Clark, Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations, *Genome Res.* (2016) 719-731.

[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Chromosomal DNA and Its Packaging in the Chromatin Fiber, *Molecular Biology of the Cell, 4th edition* (2002) 754-756.

[6] G.K. Ferguson, *The human genome: poems on the book of life.*

[7] C. Shekhar, Chromatin Fiber: Zigzag or Solenoid?, *Biomedical Computation Review* (2009).

[8] J. Ostashevsky, A polymer model for large-scale chromatin organization in lower eukaryotes, *Moll Biol Cell* (2002) 2157-2169.

[9] M.R. Hübner, M.A. Eckersley-Maslin, D.L. Spector, Chromatin Organization and Transcriptional Regulation, *Curr Opin Genet Dev* (2013) 89-95.

[10] J. Dekker, Mapping in vivo chromatin interactions in yeast suggests an extended chromatin fiber with regional variation in compaction, *J Biol Chem* (2008) 34532-34540.

[11] E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner, R. Sandstrom, B. Bernstein, M.A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L.A. Mirny, E.S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, *Science* (2009) 289-293.

[12] J.R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, B. Ren, Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions, *Nature 485* (2012) 376–380.

[13] J. Dekker, M.A. Marti-Renom, L.A. Mirny, Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data, *Nature Reviews Genetics 14* (2013) 390–403.

[14] T. Cremer, M. Cremer, Chromosome Territories, *Cold Spring Harb Perspect Biol* (2010).

[15] T. Chandra, P.A. Ewels, S. Schoenfelder, M. Furlan-Magaril, S.W. Wingett, K. Kirschner, J.Y. Thuret, S. Andrews, P. Fraser, W. Reik, Global reorganization of the nuclear landscape in senescent cells, *Cell Rep* (2015) 471-483.

[16] S.W. Criscione, M. De Cecco, B. Siranosian, Y. Zhang, J.A. Kreiling, J.M. Sedivy, N. Neretti, Reorganization of chromosome architecture in replicative cellular senescence, *Science Advances 2* (2016).

[17] *https://omictools.com/3c-4c-5c-hi-c-chia-pet-category.*

[18] T. Nagano, Y. Lubling, T.J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E.D. Laue, A. Tanay, P. Fraser, Single cell Hi-C reveals cell-to-cell variability in chromosome structure, *Nature* (2013).

[19] *GEO* – Gene Expression Omnibus database, *https://www.ncbi.nlm.nih.gov/geo/.*

[20] F. Ay, W.S. Noble, Analysis methods for studying the 3D architecture of the genome, *Genome Biology* (2015).

[21] A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul, J. Mozziconacci, Normalization of a chromosomal contact map, *BMC Genomics* (2012).

[22] E. Yaffe, A. Tanay, Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture, *Nat Genet* (2011) 1059-1065.

[23] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, J.S. Liu, HiCNorm: removing biases in Hi-C data via Poisson regression, *Bioinformatics* (2012) 3131-3133.

[24] M. Imakaev, G. Fudenberg, R.P. McCord, N. Naumova, A. Goloborodko, B.R. Lajoie, J. Dekker, L.A. Mirny, Iterative correction of Hi-C data reveals hallmarks of chromosome organization, *Nature Methods 9* (2012) 999-1003.

[25] P.A. Knight, D. Ruiz, A fast algorithm for matrix balancing, *IMA Journal of Numerical Analysis* (2012).

[26] HOMER, Analyzing Hi-C genome-wide interaction data, *http://homer.salk.edu/homer/interactions.*

[27] Y. Zhu, Z. Chen, K. Zhang, M. Wang, D. Medovoy, J.W. Whitaker, B. Ding, N. Li, L. Zheng, W. Wang, Constructing 3D interaction maps from 1D epigenomes, *Nature Communications 7* (2016).

[28] J. Chen, A.O. Hero, I. Rajapakse, Spectral Identification of Topological Domains, *Bioinformatics* (2015).

[29] T. Sexton, G. Cavalli, The role of chromosome domains in shaping the functional genome, *Cell* (2015) 1049-1059.

[30] N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B.R. Lajoie, L.A. Mirny, J. Dekker, Organization of the mitotic chromosome, *Science* (2013) 948-953.

[31] J. Ostashevsky, A polymer model for large-scale chromatin organization in lower eukaryotes, *Mol Biol Cell* (2002) 2157-2169.

[32] A. Lesne, J. Riposo, P. Roger, A. Cournac, J. Mozziconacci, 3D genome reconstruction from chromosomal contacts, *Nature Methods 11* (2014) 1141-1143.

[33] M. Rousseau, J. Fraser, M.A. Ferraiuolo, J. Dostie, M. Blanchette, Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling, *BMC Bioinformatics* (2011).

[34] Roadmap Epigenomics Project: *http://www.roadmapepigenomics.org/*.

[35] *http://hic.umassmed.edu/welcome/welcome.php*.

[36] *http://www.aidenlab.org/juicebox/*.

[37] *http://genome3d.eu/*.

[38] *http://sgt.cnag.cat/3dg/*.

[39] GEO: *http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525*.

[40] F.R.K. Chung, *Spectral Graph Theory* (1997).

[41] D. Filippova, R. Patro, G. Duggal, C. Kingsford, Identification of alternative topological domains in chromatin, *Algorithms for Molecular Biology* (2014).

[42] M.D. Ritchie, E.R. Holzinger, R. Li, S.A. Pendergrass, D. Kim, Methods of integrating data to uncover genotype–phenotype interactions, *Nature Reviews Genetics 16* (2015) 85-97.