Alma Mater Studiorum · Università di Bologna

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

Analisi spettrale di matrici Hi-C

Relatrice:
Dott.sa Alessandra Merlotti

Correlatore:
Prof. Daniel Remondini

Presentata da:
Giovanni Raumer

Anno Accademico 2022/2023

Abstract

Le matrici Hi-C rappresentano uno strumento fondamentale per mappare le interazioni spaziali tra regioni del genoma linearmente distanti ma in prossimità all'interno del nucleo cellulare. Sono matrici ad alta dimensionalità che consentono di esaminare la struttura tridimensionale della cromatina e di comprendere meglio il ruolo dell'organizzazione spaziale nella regolazione dei processi genetici. Questo studio di tesi si propone di esaminare tre matrici Hi-C da un punto di vista spettrale. I dati presi in esame riguardano la linea cellulare sana GM12878 e due linee cellulari anormali, KBM7 e T47D. L'analisi sarà condotta utilizzando un approccio a network, con l'obiettivo di caratterizzare le matrici anche nel contesto della Random Matrix Theory. Questo ci consentirà di valutare il grado di casualità delle interazioni genomiche e di stabilire eventuali correlazioni con lo stato biologico delle cellule. Ci proponiamo di identificare eventuali differenze e anomalie spettrali tra il DNA sano e quello affetto da patologie, nonché di individuare eventuali riarrangiamenti spaziali della sequenza di DNA, noti come aberrazioni, analizzando solamente le relative matrici di adiacenza. I risultati evidenziano differenze significative negli autovalori massimali e nel clustering dello spettro, suggerendo una maggiore connessione globale nel network sano. Ma emergono differenze significative soprattutto nella densità spettrale del bulk che mostra una netta divisione tra linea sana e linee aberranti. In secondo luogo, il calcolo dell'Inverse Participation Ratio, unitamente al test di Shapiro-Wilk per tutti gli autovettori ha evidenziato una componente casuale che seppur comune alle 3 reti, risulta più marcata nel network sano. Lo studio ha inoltre portato alla luce la correlazione negativa tra IPR e parametro statistico W associato agli autovettori delle matrici, mostrando come sia necessario considerare anche autovettori relativi ad autovalori piccoli, solitamente trascurati.

Indice

Introduzione

La biologia moderna ha fatto enormi passi avanti nell'analisi del genoma umano e nella comprensione dei processi che regolano la sua struttura e funzione. Uno degli sviluppi più significativi in questo campo è sicuramente la tecnologia Hi-C che ha permesso di mappare l'architettura tridimensionale del genoma con risoluzioni mai raggiunte prima. La tecnica Hi-C è stata introdotta per la prima volta da Lieberman-Aiden et al.[11] nel 2009 e impiega la formaldeide per stabilizzare le interazioni tra regioni genomiche adiacenti all'interno del nucleo ma linearmente lontane nel filamento di cromatina. Successivamente, il DNA viene frammentato e riassemblato in modo che le porzioni con frequenti interazioni siano fisicamente vicine. Questi frammenti sono quindi sequenziati e le loro interazioni vengono identificate e mappate in una matrice detta Hi-C.
Gran parte della letteratura su questa tecnica si concentra sull'individuazione e l'analisi di strutture come i loop, i domini ad associazione topologica (TAD) e i compartimenti cromosomici che sono di grande importanza per comprendere la relazione tra conformazione tridimensionale del genoma, attività genica e stato funzionale della cellula. Un approccio puramente spettrale a questo tipo ti dato è stato esplorato in misura nettamente inferiore, sia per quanto riguarda la caratterizzazione di una singola matrice, sia per il confronto tra linee cellulari sane e aberranti. Tale matrice può rappresentare di fatto un network nel quale i loci genici, ossia i nodi della rete, hanno frequenti interazioni tra essi. Generalmente l'analisi spettrale si concentra su matrici di adiacenza binarie, che rappresentano la presenza o meno di un contatto tra due nodi della rete. Meno materiale è presente sullo studio di matrici pesate, come lo sono quelle Hi-C nel nostro caso. In questa tesi ci concentriamo quindi sull'applicazione di tecniche spettrali per analizzare le caratteristiche strutturali di 3 matrici Hi-C cercando analogie e differenze con modelli ampiamente studiati come le reti Random ER o le reti a invarianza di scala, per poi cercare di far emergere delle differenze sempre di tipo spettrale tra matrici sane e matrici aberranti.
Mentre gli autovalori possono essere interpretati come misure quantitative di complessità strutturale, gli autovettori generalmente forniscono informazioni sulle principali modalità di organizzazione spaziale del genoma. Andremo quindi a calcolare autovalori massimali, strength, distribuzione e densità spettrali per poi passare all'analisi delle componenti degli autovettori e dell'Inverse Participation Ratio. Tutti i network biologici presentano

presentano generalmente un certo grado di casualità, con gli strumenti della Random Matrix Theory andremo quindi a verificare in che misura si manifesta nelle matrici in esame. In letteratura sono presenti alcune prove del fatto che network genetici sani abbiano una più spiccata componente casuale rispetto a reti che presentano aberrazioni[13].

Capitolo 1

Il DNA: struttura 3D e analisi Hi-C

1.1 Struttura di base del DNA

L'acido desossiribonucleico, o DNA, è la molecola polimerica che contiene tutte le informazioni necessarie per l'esistenza e lo sviluppo di un organismo vivente. E' composta da due catene polipeptidiche formate da subunità, i monomeri, chiamati nucleotidi, i quali a loro volta sono formati da uno zucchero, un gruppo fosfato e una base azotata. Le basi che differenziano le 4 tipologie di monomeri sono l'Adenina (A), la guanina (G), la Citosina (C) e la timina (T) (nell'RNA l'uracile (U) sostituisce la timina). In ogni nucleo cellulare sono contenute all'incirca 3,2 miliardi di coppie di basi azotate che si legano sempre secondo gli accoppiamenti A-T e C-G, in modo da rendere efficace il meccanismo di duplicazione che utilizza solo una metà del filamento a doppia elica. Ciò che differenzia un individuo da un altro è proprio la sequenza di queste coppie di basi nella quale sono codificate tutte le istruzioni necessarie per il funzionamento dell'organismo. Una porzione che codifica la produzione di una specifica proteina o di una molecola di RNA è detta gene, e il codice genetico umano solitamente ne contiene tra i 25.000 e i 30.000. Tuttavia, solo all'incirca 21.000 di questi geni codificano le informazioni per la produzione di proteine: come, quando e in che quantità produrle. La totalità del corredo genetico di un individuo viene detta genoma.
Negli organismi eucarioti il materiale genetico è diviso in cromosomi, ossia le unità organizzative tramite le quali il DNA è compattato all'interno delle cellule. Ognuno di questi è una porzione del filamento di DNA contenente i geni relativi a specifiche caratteristiche ereditarie. Sono 23 le coppie totali di cromosomi, di cui 22 sono composte da cromosomi detti omologhi, ossia codificanti per gli stessi caratteri, uno ereditato dalla madre e uno dal padre. I due rimanenti sono i cromosomi sessuali: XX nel caso di individuo femmina e XY nel caso di individuo maschio.
La lunghezza totale del genoma, considerando il filamento di cromatina lineare e non condensato è di approssimativamente 2 m, e il diametro del nucleo che lo contiene è pari

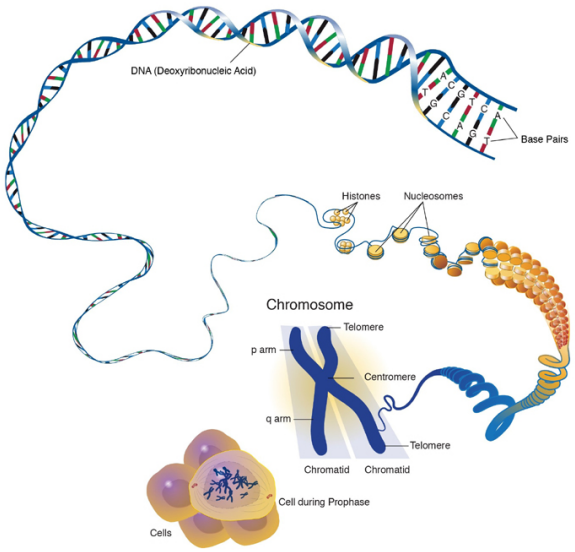


Figura 1.1: vari livelli di organizzazione del filamento di DNA

a 2 μm.
Emerge quindi l'importanza del ripiegamento del DNA: come può un filamento di questa lunghezza compattarsi all'interno nel nucleo rimanendo "accessibile" e decodificabile? Tale processo coinvolge proteine specifiche che si legano e ripiegano il filamento a doppia elica secondo vari livelli di organizzazione, sempre più complessi, in modo da renderla accessibile ad altri enzimi e proteine che la replicano, la riparano e ne esprimono i geni. L'insieme del DNA condensato e delle proteine coinvolte nei vari livelli di avvolgimento viene chiamato cromatina. Proprio questa struttura, altamente condensata, che a sua volta si ripiega per formare quelli che chiamiamo cromosomi, sarà oggetto degli studi presentati in questa tesi, nell'ambito dell'esperimento Hi-C. É da sottolineare il fatto che la struttura cromosomica non sia statica e immutabile ma sia dinamica e durante varie fasi della vita cellulare cambi conformazione.

1.2 Il folding del DNA

Vediamo ora come il DNA raggiunge l'alto livello di condensazione all'interno del nucleo. Sono principalmente due le tipologie di proteine coinvolte nel processo di impacchetamento 3D del DNA: le proteine istoniche, di massa pari a quella del DNA nella massa complessiva di un cromosoma, e le proteine cromosomiche non istoniche.

Il primo stadio di organizzazione consiste nell'avvolgimento del filamento a doppia elica attorno alle proteine istoniche (vedi figura 1.1). La struttura risultante viene chia-

mata nucleosoma ed è l'insieme del materiale genetico avvolto e di 8 istoni, ognuno con una lunghezza di 147 coppie di basi arrotolata ad esso. Ogni agglomerato di questo tipo è separato dagli altri da una porzione di codice genetico che può arrivare a 80 coppie di basi chiamate DNA linker. In figura 1.2 si può osservare il filamento di cromatina prelevato da un nucleo in interfase (il periodo di crescita e sviluppo compreso fra due divisioni cellulari) e in basso la struttura a nucleosomi, osservabili dopo l'unfolding artificiale del filamento di cromatina. Si nota come ogni nucleosoma sia separato dagli altri dal DNA linker.
All'interno di un nucleo cellulare è tuttavia improbabile trovare del materiale genetico con questa conformazione, che possiamo chiamare "a collana di perle". Quello che si osserva al microscopio è quel filamento chiamato cromatina, di diametro 30 μm, che rappresenta un ulteriore livello di organizzazione dei nucleosomi. Tuttavia, il processo che porta alla strutturazione dei nucleosomi in array condensati rimane poco chiaro: il modello a zigzag propone l'agglomerarsi di strutture tetraedriche di 4 nucleosomi come meccanismo per ottenere il filamento di cromatina. Altri modelli, come quello a solenoide propongono strutture meno ordinate. E' sicuramente probabile che il processo alla base dell'organizzazione dei nucleosomi coinvolga le cosiddette code istoniche, filamenti fuoriuscenti dall'agglomerato proteico e una tipologia particolare di istone: l'istone linker H1.[1]

1.3 Regolazione della trascrizione

É importante evidenziare il fatto che non tutto il genoma contenga materiale genetico a noi conosciuto. Il Progetto Genoma Umano, terminato nel 2003, ha portato alla luce molti aspetti del DNA che fino ad allora erano solo delle ipotesi. Ha evidenziato chiaramente come gran parte del codice genetico non codifichi per la produzione di proteine, ma solo l'1,5 % dell'intera sequenza di basi sia adibita a questo (esoni) e il restante 98,5 % non codifichi informazioni genetiche. Molte porzioni di DNA sono adibite alla codifica di informazioni per il controllo dell'espressione dei geni: quando produrre una proteina? in che momento? Sono sequenze dette regolatrici che si trovano in prossimità o all'interno dei geni e ne controllano la trascrizione. Assieme a queste sequenze, gli elementi regolatori, solitamente proteine, facilitano l'espressione genica. Tra le sequenze regolatrici più note troviamo:

- Promotori: sequenza alla quale si lega l'RNA polimerasi per avviare la trascrizione di un gene.
- Enhancer: sequenze di DNA che, se legate con proteine specifiche, aumentano la frequenza di trascrizione di un gene.

Mentre gli elementi regolatori sono:

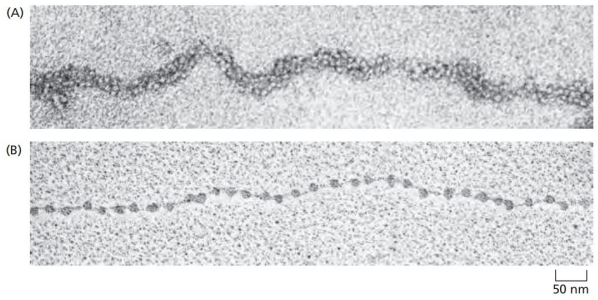


Figura 1.2: Filamento di cromatina(A) e nucleosomi visibili dopo lo srotolamento della cromatina(B)

- Fattori di trascrizione: proteine che si legano a regioni specifiche di DNA, come promotori o enhancer e ne regolano la trascrizione
- Repressori: proteine che legandosi al DNA bloccano la trascrizione di uno specifico gene

1.4 L'importanza della struttura 3D del DNA

Come sottolineato in precedenza, al di là della struttura "a collana di perle" di DNA linker e nucleosomi, la cromatina ha una conformazione altamente dinamica e la sua struttura può variare rapidamente in base alla necessità della cellula e alla sua fase vitale: da una conformazione di compattezza massima durante la mitosi a una struttura più rilassata nel corso dell'interfase. Ma la dinamicità deve necessariamente estendersi a livello locale, per far sì che proteine ed enzimi possano esprimere, riparare e replicare specifiche porzioni di DNA contenuto nei cromosomi.
É ormai chiaro come la struttura dei cromosomi giochi un ruolo cruciale nella regolazione dei processi biologici di un organismo, variando la sua conformazione, modificando la sua struttura rende accessibili diverse porzioni di codice in diversi momenti della vita di una cellula. Una delle componenti più importanti nell'organizzazione ad alto livello della cromatina è la proteina CTCF: è in grado di formare i cosiddetti loop. Come si vede in figura 1.3, essa coordina l'interazione tra enhancer e promotori: secondo il modello a estrusione di loop infatti, i siti CTCF avvicinano tra loro coppie di siti genomici lontani nella catena lineare permettendo agli enhancer di agire sui promotori e quindi attivare la produzione di specifiche proteine. Questo è però solo uno dei diversi meccanismi che agiscono sul filamento di cromatina, è risaputo che i loop si organizzano nei cosiddetti Topological Associated Domains (TAD) che a loro volta vanno a costituire i compartimenti (fig. 1.3). Sono proprio questi i tre meccanismi fondamentali che modulano la struttura 3D della cromatina e influenzano funzioni come la trascrizione e la replicazione, motivo

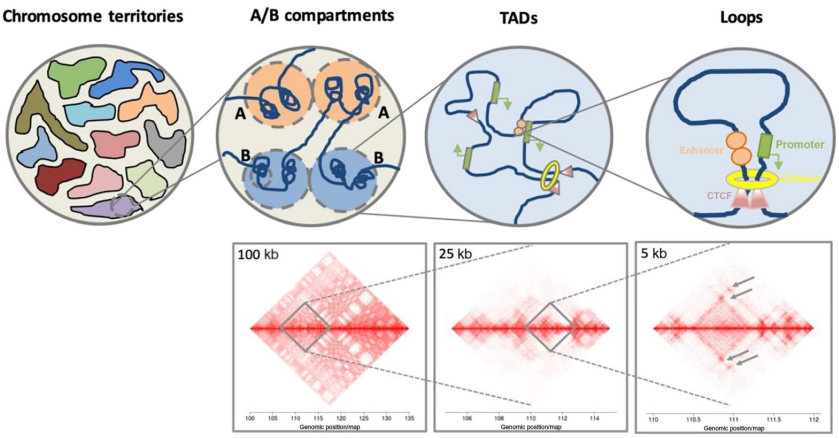


Figura 1.3: Livelli di organizzazione della cromatina nella formazione dei territori cromosomici. In basso le mappe di contatto con la risoluzione necessaria a visualizzare le corrispondenti strutture genomiche

per cui anomalie nell'organizzazione strutturale possono avere potenziali conseguenze patologiche sull'organismo [16][14]. Le aberrazioni colpiscono principalmente loop e TAD andando a creare legami tra promotori e enhancer anomali che mutano o disabilitano la capacità funzionali del codice genetico. Oppure, interruzioni nei domini topologici (TAD) possono far interagire promotori e enhancer solitamente isolati e attivare geni legati a condizioni patologiche. É difficile stabilire se le mutazioni del codice genetico siano causa o conseguenza della patologia, la tentazione è postulare che strutture 3D alterate influenzino le funzionalità del genoma portando a fenotipi malati[10]; tuttavia ciò esula dallo scopo della tesi. L'obiettivo ora non è capire quali meccanismi e quali anomalie costituiscano la struttura delle linee cellulari in esame ma se queste possano riflettersi nelle caratteristiche spettrali della matrice che le rappresenta. Prima di tutto quindi, dobbiamo avere a disposizione dati relativi a quali e quante volte porzioni diverse del filamento di cromatina vengano a contatto nel nucleo cellulare.
Sono molte le tecniche sperimentali nate negli ultimi anni per indagare la conformazione 3D del DNA su vari livelli di organizzazione. Tecniche come la 3C (Chromosome Conformatio Capture), 4C (Circularized Chromosome Conformation Capture) e 5C(Carbon Copy Chromosome Conformation Capture) hanno permesso di indagare in vivo la struttura e le interazioni a lungo range della cromatina a livello molecolare. Tuttavia, queste tecnologie permettono di indagare le interazioni di specifici loci o gruppi di loci. I dati analizzati in questo progetto di tesi provengono invece dalla tecnologia Hi-C (High-throughput Chromosome conformatio capture), in grado di identificare le interazioni cromatiniche su scala genomica.

1.5 L'esperimento Hi-C

La conformazione 3D dei cromosomi porta segmenti della cromatina molto distanti linearmente a trovarsi in grande prossimità all'interno del nucleo. Comprendere come il cromosoma si compatta aiuta a capire la relazione tra la struttura tridimensionale della cromatina e l'attività genica della cellula. L'esperimento Hi-C fa proprio questo, va ad individuare quali porzioni di cromatina, lontane linearmente, vengono a contatto all'interno del nucleo. La problematica principale delle tecnologie 3C, 4C e 5C, è quella di dover identificare a priori i loci da osservare mentre la tecnica Hi-C permette di individuare legami non solo tra loci specifici ma permette di studiare e contare i legami dell'intero genoma. Tali interazioni possono derivare da funzioni biologiche, come le interazioni promotore-enhancer, o da loop polimerici casuali, dove il movimento fisico non diretto della cromatina provoca la collisione dei loci. Il procedimento della tecnica Hi-C per contare le interazioni consiste in 5 step, illustrati in figura 1.4:

1. Cross-linking del DNA: le regioni adiacenti di DNA vengono legate covalentemente con la formaldeide. Questo passaggio conserva le interazioni spaziali in atto nelle cellule.

2. Taglio delle estremità: il DNA viene tagliato in frammenti più piccoli, solitamente attraverso l'uso di una specifica endonucleasi di restrizione, come ad esempio la HindIII. Questo passaggio genera frammenti di DNA contenenti le regioni cross-linkate

3. Etichettatura: le estremità tagliate vengono evidenziate. Questa passaggio consente di identificare gli estremi dei frammenti e di associarli alle sequenze genomiche di origine.

4. Ligazione intermolecolare: vengono uniti i frammenti di DNA provenienti da diverse regioni genomiche che erano in stretta vicinanza spaziale.

5. Purificazione: vengono rimossi i legami covalenti e rimangono solo i filamenti legati e marcati

6. Sequenziamento paired-ends: entrambe le estremità vengono sequenziate in parallelo

Il risultato finale di questo procedimento è una matrice di contatto M dell'intero genoma, diviso in loci il cui numero di basi va a definire la risoluzione. Quindi ogni ingresso della matrice m_{ij} indica il numero di interazioni tra i loci i e j. E' importante sottolineare che la matrice rappresenta una media dell'esperimento ripetuto su più elementi di una linea cellulare[11][17].

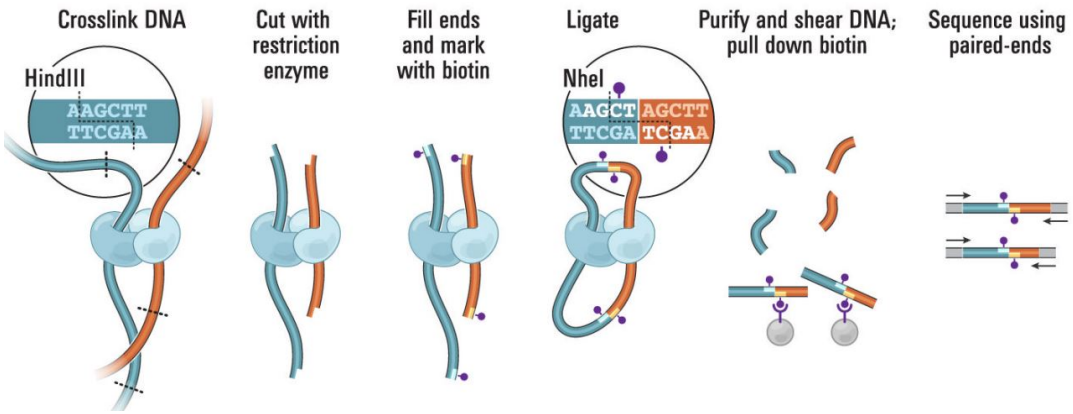


Figura 1.4: Le sei fasi della tecnica Hi-C per il sequenziamento della cromatina

Capitolo 2

Materiali e metodi

Andiamo ora ad isolare il dato Hi-C e a studiarne le caratteristiche spettrali. L'obiettivo di questo elaborato è proprio quello di evidenziare, se presenti, delle differenze tra linea cellulare sana e linee aberranti in termini di informazioni spettrali delle relative matrici. Consideriamo quindi il DNA come un network nel quale i nodi sono le porzioni di cromatina la cui lunghezza è definita dalla risoluzione delle matrici e i pesi delle connessioni sono proprio gli ingressi di tali matrici.

2.1 Linee cellulari

Le matrici oggetto di analisi di questo elaborato provengono dal laboratorio del Prof. Marc Marti-Renom del CNAG-CRG di Barcellona, hanno una risoluzione di 1 Mb e sono relative a tre linee cellulari diverse:

- Linea cellulare sana GM12878: è una linea di linfociti B umani provenienti da donatrice femmina.
- Linea cellulare aberrante KBM7: linea proveniente da donatore maschio affetto da leucemia mieloide cronica (CML) in fase blastica.
- Linea cellulare aberrante T47D: sono cellule epiteliali e provengono da una paziente affetta da tumore mammario.

2.1.1 Introduzione alla teoria dei network

Nell'ambito della teoria dei network possiamo definire un grafo come un insieme di N nodi e N_c connessioni che possono essere rappresentati da una matrice di adiacenza A_{ij} definita come:

$$A_{ij} = \begin{cases} 1 & \text{se } i \sim j \\ 0 & \text{altrimenti} \end{cases} \tag{2.1}$$

Se il grafo è non diretto, ossia i legami tra i loci non hanno un direzione ma indicano solamente un'interazione, allora la matrice sarà simmetrica e avrà autovalori reali. Nel nostro caso, il network è inoltre caratterizzato da connessioni pesate: tra due loci ci possono essere più interazioni e il peso corrisponde proprio a quante volte le due porzioni vengono a contatto nella struttura 3D. É chiaro allora che la matrice di contatto non sarà binaria e avrà come ingressi proprio i pesi delle connessioni. Definiamo il grado di un nodo k come la somma delle connessioni che possiede con gli altri nodi del network. Per reti pesate il grado prende il nome di strength e si calcola nel modo seguente:

$$k = \sum_{j=1}^{N} A_{ij} \tag{2.2}$$

mentre strength medio e massimo sono rispettivamente:

$$\langle k \rangle = \frac{\sum_{j=1}^{N} A_{ij}}{N} \tag{2.3}$$

$$k_{\max} = \max_{1<i<N} k_i \tag{2.4}$$

2.2 Analisi spettrale di matrici Hi-C

Lo spettro di un network consiste nell'insieme degli autovalori della sua matrice di adiacenza A_{ij} e viene indicato con λ_i dove $i = 1, 2, ..., N$. É stato dimostrato come lo spettro fornisca importanti informazioni che possono andare dalle proprietà topologiche del network sottostante a quelle dinamiche relative all'interazione tra i nodi [15].
Il metodo più semplice per visualizzare le matrici consiste nella creazione delle cosiddette mappe di contatto: tabelle che associano ad ogni ingresso della matrice un colore la cui intensità è proporzionale al valore. Per migliorarne la visualizzazione solitamente si calcola il logaritmo degli ingressi riducendo così il range di valori. Le mappe di contatto saranno il punto di partenza dell'analisi: sarà evidente infatti quali matrici presentano più anomalie a livello di contatti intercromosomici e cercheremo un riscontro dal punto di vista spettrale. In figura 2.1 è raffigurato un esempio di mappa di contatto relativa al cromosoma 1 della linea cellulare GM12878. Lungo la diagonale la matrice presenta solo zeri poiché ogni locus genico non presenta interazioni con se stesso.

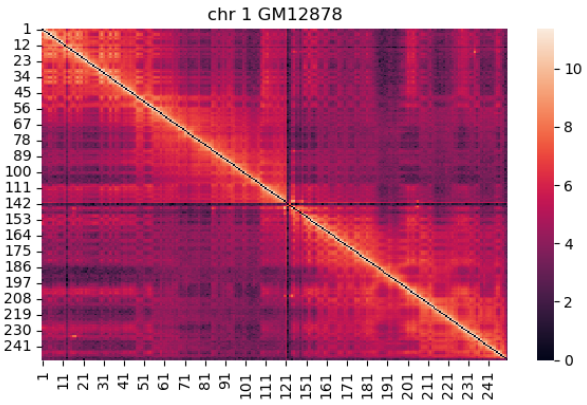


Figura 2.1: mappa di contatto del cromosoma 1 della linea cellulare GM12878

2.2.1 Autovalore principale della matrice di adiacenza

L'autovalore principale o massimale di una matrice di adiacenza, che indichiamo con λ_1 gioca un ruolo chiave nello studio di network. Ha infatti un ruolo molto importante nel descrivere la stabilità dei pattern di interazione della rete sottostante: è una stima della connettività media. Ne è un esempio l'autovalore massimale di reti a invarianza di scala che descrivono l'andamento dei contagi di un'epidemia, la cui soglia di avvio di contagi è determinata proprio da questo valore. Per il teorema di di Perron-Frobenius[18], l'autovalore massimale di una matrice con valori non negativi è reale e positivo mentre per il teorema di Gerschgorin, ogni autovalore di una matrice di adiacenza A sta all'interno di almeno uno dei dischi circolari di centro a_{ii} e raggi $[\sum_{j=1,j\neq i} |a_{ij}|, \sum_{j=1,j\neq i} |a_{ji}|]$ e può essere scritto come:

$$|a_{ii} - \lambda| \leq \sum_{j=1,j\neq i}^{N} a_{ii} \tag{2.5}$$

Solitamente per grafi semplici possiamo considerare $a_{ii} = 0$ e il raggio massimo del disco più grande diventa quindi il massimo strength k_{max}. Di conseguenza tutti gli autovalori sono compresi nell'intervallo $[-k_{max}, k_{max}]$. Per un altro teorema [6], l'autovalore massimale di una matrice simmetrica a valori reali è sempre compresa tra grado medio e grado massimo:

$$\langle k \rangle \leq \lambda_1 \leq k_{max} \tag{2.6}$$

É necessaria però una precisazione, questo teorema è valido per matrici di adiacenza binarie, mentre i dati analizzati in questo elaborato si riferiscono a grafi pesati. In questo caso si parla quindi di strength. Vedremo poi nel prossimo capitolo che i teoremi verranno rispettati anche dagli autovalori dei dati Hi-C. Cerchiamo quindi gli autovalori

delle matrici in esame per verificare l'ipotesi di un autovalore massimale maggiore nel caso della linea cellulare sana rispetto alle due linee aberranti.

2.2.2 Bulk degli autovalori e densità spettrale

Se l'autovalore massimale descrive la stabilità e la dinamica della rete, il bulk degli autovalori restituisce informazioni riguardo al grado di casualità nelle interazioni e connessioni tra i nodi. Il contesto di riferimento per lo studio spettrale degli autovalori è la teoria delle matrici casuali (RMT), proposta da Wigner per spiegare alcune proprietà statistiche dello spettro dei nuclei atomici [12] e rivelatasi estremamente utile nella descrizione di molti modelli reali. Andiamo quindi a studiare la densità spettrale degli autovalori, ossia la densità degli autovalori della matrice di adiacenza. É un'analisi del network che può restituire rapidamente delle differenze globali tra le reti in esame. Possiamo scrivere la densità spettrale come:

$$\rho(\lambda) = \frac{1}{N}\sum_{j=1}^{N}\delta(\lambda - \lambda_j) \tag{2.7}$$

Che converge a una funzione continua per $N \to \infty$, dove N è il numero di nodi della rete. λ_j è il j-esimo autovalore dello spettro, ordinato in modo decrescente. Se A_{ij} è una matrice reale $N \times N$ simmetrica e casuale, la sua densità spettrale, per $N \to \infty$ è:

$$\rho(\lambda) = \begin{cases} (2\sigma^2)^{-1}\sqrt{(4\sigma^2 - \lambda^2)} & \text{se } |\lambda| \leq 2\sigma \\ 0 & \text{altrimenti} \end{cases} \tag{2.8}$$

dove $\langle A_{ij}^2 \rangle = \sigma^2$. Questa funzione, che mostra delle notevoli somiglianze con la legge del semicerchio di Wigner (2.10), descrive la densità spettrale dei network casuali ER, costruiti connettendo ogni coppia di nodi con una probabilità p (figura 2.2). Questa tipologia di grafo ha approssimativamente $pN(N-1)/2$ connessioni distribuite randomicamente e la distribuzione dei gradi $P[k]$ segue una binomiale con valore atteso uguale per tutti i nodi $\langle k \rangle = p(N-1) \approx pN$. La larghezza della distribuzione scala come $\sigma\sqrt{N}$ e le code decadono esponenzialmente. L'autovalore massimale di queste reti inoltre è solitamente molto distanziato dal bulk.
Un altro modello di network molto studiato è il cosiddetto network a invarianza di scala, scoperto da Barabási e Albert nel 1999 [4]. Si costruisce secondo un meccanismo chiamato attaccamento preferenziale. Partendo da un ridotto numero di vertici, ad ogni step si aggiunge un nodo con un certo grado che tende a connettersi con i nodi a grado più alto k_i con una probabilità pari a $\pi(k_i) = k_i / \sum_j k_j$. A differenza delle reti ER, la distribuzione dei gradi di questa rete segue un andamento a legge di potenza $P(k) \sim k^{-\lambda}$ con λ tipicamente compreso tra 2 e 3 mentre la densità spettrale ha un andamento di

tipo triangolare [7] (figura 2.2). Questo indica solitamente la presenza di nodi ad alto grado nel network, utili per rendere la rete robusta nei confronti di perturbazioni esterne [5].

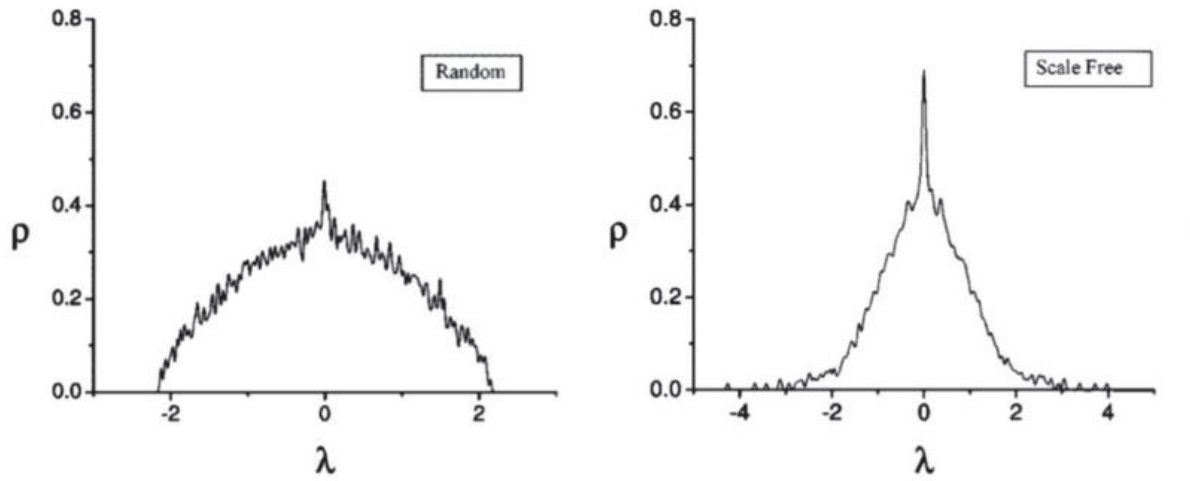


Figura 2.2: A sinistra il tipico andamento della densità spettrale di una rete casuale di tipo ER mentre a destra la densità di una rete a invarianza di scala. Entrambe le reti hanno 1024 nodi [15].

2.2.3 Casualità nei network

La teoria delle matrici casuali RMT descrive le proprietà di matrici i cui elementi sono variabili casuali distribuite secondo gaussiane. l'Ensemble Gaussiano Ortogonale (GOE) è l'insieme statistico più utilizzato per descrivere network complessi che presentano interazioni casuali tra i loro costituenti. É descritto da una misura gaussiana con densità

$$\frac{1}{Z_{GOE_n}} e^{-\frac{n}{4} Tr H^2} \tag{2.9}$$

nello spazio delle matrici reali simmetriche $n \times n$ e la densità degli stati segue la legge del semicerchio di Wigner:

$$\rho(x) = \frac{2}{\pi}\sqrt{1 - x^2} \tag{2.10}$$

La somiglianza con la densità degli autovalori delle reti ER spinge ad approfondire lo studio della componente casuale dei network tramite gli strumenti della RMT. Andremo allora a calcolare la densità della distribuzione dei rapporti tra intervalli di autovalori

consecutivi $P(r)$: siano λ_n gli autovalori ordinati in modo crescente e $s_n = \lambda_{n+1} - \lambda_n$ la distribuzione degli spazi tra autovalori vicini. Calcoliamo quindi i rapporti tra intervalli consecutivi

$$r_n = \frac{s_{n+1}}{s_n} \tag{2.11}$$

o, in alternativa

$$\tilde{r} = \frac{\min(s_n, s_{n-1})}{\max(s_n, s_{n-1})} = \min(r_n, \frac{1}{r_n}) \tag{2.12}$$

Qui utilizzeremo la distribuzione di probabilità di r, che secondo la statistica GOE assume l'andamento della seguente funzione:

$$P(r) = \frac{1}{Z_\beta} \frac{(r + r^2)^\beta}{(1 + r + r^2)(1 + 3/2\beta)} \tag{2.13}$$

dove $Z_\alpha = 8/27$ e $\beta = 1$ nel caso di ensemble gaussiano ortogonale. Andremo a verificare se le reti in esame seguono questa distribuzione o si discostano in maniera significativa. Questo andamento viene solitamente esibito anche da network diversi: sani e aberranti, implicando la comune presenza di una componente casuale [13][2], sia nel caso di modelli random ER sia per reti a invarianza di scala [3].

2.2.4 Inverse Participation Ratio

L'inverse Participation Ratio è una misura utilizzata nell'ambito della teoria dei grafi e delle reti per valutare la distribuzione della connettività dei nodi in una rete. Si tratta di un'importante metrica per comprendere il livello di coinvolgimento o di partecipazione dei nodi nei processi di rete. Viene definito come la somma alla quarta delle componenti dell'autovettore in esame:

$$IPR = \sum_{j=1}^{N} c_j^4 \tag{2.14}$$

Essendo gli autovettori normalizzati con norma euclidea, la somma delle componenti alla seconda potenza non sarebbe informativa in quanto unitaria per tutti. La somma delle componenti alla quarta potenza però fornisce un indice di eterogeneità nella distribuzione delle componenti. Immaginiamo per semplicità due casi estremi:

- Le componenti dell'autovettore sono uniformemente distribuite: se le componenti sono tutte uguali, ognuna ha un peso di $1/\sqrt{N}$ e sommate N volte alla quarta potenza otteniamo un IPR pari a $1/N$.

- una componente pari a 1 e tutte le altre nulle: l'IPR è pari a 1 poiché l'unica componente non nulla è unitaria.

Abbiamo quindi un range di valori per l'IPR che va da $1/N$ a 1. Tanto più si avvicina all'unità, tanto più l'autovettore corrispondente avrà le componenti distribuite eterogeneamente, e ricordando che le componenti di una autovettore relativo ad una matrice di adiacenza sono in corrispondenza con i nodi del grafo, ciò sarà indice di una possibile anomalia nei pattern di interazione del network sottostante. Andremo a calcolare questo parametro per i primi 25 autovettori e successivamente per la totalità degli autovettori.

2.2.5 Autovettori e matrici essenziali

Individuati gli autovettori più significativi a livello di IPR, è utile andare a studiare le componenti dell'autovettore in esame per verificare se le anomalie spettrali riscontrate possano essere rappresentative di anomalie nelle interazioni visibili dalla mappa di contatto. In generale, gli autovettori di una matrice di adiacenza possono essere utilizzati per individuare le comunità del network in quanto rappresentano le direzioni principali di connessione della rete. Gli autovettori relativi agli autovalori più grandi possono indicare nodi più centrali o comunità più fortemente connesse. Infatti ogni componente corrisponde ad un nodo della rete e gruppi di nodi fortemente connessi tra loro hanno una maggiore probabilità di condividere componenti comuni negli autovettori.
Un altro approccio dell'analisi spettrale che fa uso degli autovettori è lo studio della parte essenziale delle matrici di adiacenza [8]. Si definisce la matrice essenziale come

$$A_{ij}{}^{\text{ess}} = \sum_{n=1}^{n^*} \lambda_n a_n{}^{(\text{i})} a_n{}^{(\text{j})} \tag{2.15}$$

dove $a_n{}^{(\text{i})}$ è l'i-esima componente dell'autovettore associato all'autovalore λ_n e $a_n{}^{(\text{i})} a_n{}^{(\text{j})}$ è il proiettore associato all'autovettore $\boldsymbol{a}_n$. n^* è il numero di autospazi che si considerano per costruire la parte essenziale della matrice. In questo modo andiamo ad escludere tutta quella parte dello spettro che in genere si nota essere compatibile con matrici casuali e isoliamo quelle che sono le caratteristiche strutturali specifiche della rete. É un approccio nettamente diverso rispetto a quello di gestire ed eliminare il rumore delle matrici. Si va piuttosto a trascurare la parte non specifica della matrice isolando le proprietà spettrali non locali che differiscono dalle matrici casuali. Emergono perciò più chiaramente i territori cromosomici come le somiglianze tra repliche o le differenze tra linee cellulari diverse. Sono inoltre più stabili rispetto a variazioni nella risoluzione e profondità di sequenziamento dell'esperimento Hi-C. A prova di ciò, in figura 2.3 si vede la distribuzione delle componenti di 4 autovettori: il 1°, 2°, 20° e 100°, ordinati secondo

l'ordine decrescente degli autovalori. Si nota come la parte più casuale della matrice sia rappresentata dagli autovettori di grado più alto, assumono infatti distribuzioni sempre più gaussiane mentre le componenti dei primi autovettori contengano invece la parte caratteristica e rappresentativa della matrice.

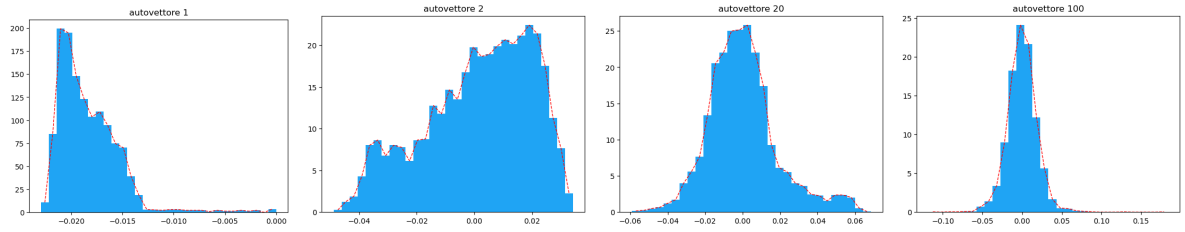


Figura 2.3: Distribuzione delle componenti di diversi autovettori relativi alla matrice di adiacenza della linea GM12878. Man mano che aumenta il grado dell'autovettore si nota come le sue componenti siano distribuite secondo gaussiane sempre più regolari

Capitolo 3

Risultati e discussione

3.1 Preparazione del dato Hi-C

I coefficienti di una matrice Hi-C definiscono quindi un grafo pesato in cui i vertici corrispondono alle porzioni di 1 Mb del genoma ed i pesi del link H_{ij} sono proporzionali alle frequenze di contatto tra queste porzioni. Si può quindi dedurre quali loci siano più prossimi nello spazio 3D osservando quali elementi della matrice presentano più conteggi. Ogni locus genico, nei dati raccolti e qui analizzati, corrisponde a 1 Mb. Le interazioni tra loci non hanno direzione, mettono solo in contatto regioni lineari diverse del genoma. Il grafo è quindi non orientato, la corrispondente matrice Hi-C è simmetrica e presenta zeri lungo tutta la diagonale. Nella regione centromerica, ossia dove i due cromosomi omologhi si intrecciano, ci sono sequenze di DNA ripetuto che rendono il codice genetico mappabile in modo non univoco. Per questo motivo la matrice Hi-C presenta alcune righe e colonne di soli zeri che vanno eliminate poiché non sono informative. Inoltre, prima di effettuare l'analisi spettrale si calcola il logaritmo degli ingressi per ridurre la variabilità tra regioni diverse del genoma, permettendo di attenuare le regioni con più contatti e rendendo più visibili le regione più deboli: la visualizzazione risulta più bilanciata. É stata inoltre rimossa dalla matrici la porzione relativa al cromosoma Y, in quanto gli ingressi che mappano questo cromosoma sono pochi, di conseguenza il segnale risulta molto inficiato dal rumore. Inoltre, le linee GM12878 e T47D sono linee ricavate da donatrici femmine.

3.2 Mappe di contatto

Una volta terminata la fase di pre-processing si possono calcolare le mappe di contatto delle matrici in esame. Questo grafico associa ad ogni elemento della matrice un colore tanto più chiaro quanti sono i contatti del locus e restituisce visivamente le regioni più connesse e quelle più isolate. In figura 3.1 sono raffigurate le mappe relative alle tre

linee cellulari in analisi: GM12878, KBM7 e T47D. Lungo le diagonali sono evidenti i quadrati caratterizzati da un maggior numero di contatti rispetto alle regioni circostanti: rappresentano i territori cromosomici. Nonostante la modalità di acquisizione dati non permetta un confronto diretto, per quanto riguarda le sottostrutture 3-D tra le due linee si può notare come li linee aberranti presentino dei blocchi luminosi al di fuori della diagonale mentre la mappa Hi-C della linea sana sia più "pulita" e regolare. E' evidente come le strutture a blocchi rappresentanti i territori cromosomici siano più compatte nel caso sano mentre nel caso aberrante presentino numerose traslocazioni. Sono state identificate le traslocazioni di entrambe le matrici aberranti individuando i cromosomi coinvolti, in modo da poter riscontrare successivamente, nell'analisi dell'IPR e degli autovettori, delle anomalie rispetto alle componenti relative ai cromosomi catalogati. La mappa di contatto della linea GM12878 non presenta anomalie visibili mentre per la linea KBM7 emergono 3 traslocazioni principali:

- chr 7 - chr 19
- chr 9 - chr 22
- chr 15 - chr 19

La linea cellulare T47D presenta invece le seguenti traslocazioni:

- chr 3 - chr 5
- chr 3 - chr 8
- chr 3 - chr 10
- chr 3 - chr 14
- chr 6 - chr X
- chr 7 - chr 15
- chr 8 - chr 14
- chr 9 - chr 15
- chr 9 - chr 17
- chr 10 - chr 20
- chr 11 - chr 22
- chr 12 - chr 13
- chr 12 - chr 16

- chr 16 - chr 20

Proveremo ad individuare, mediante lo studio degli autovettori, delle corrispondenze tra questi legami e le componenti relative ai rispettivi nodi.

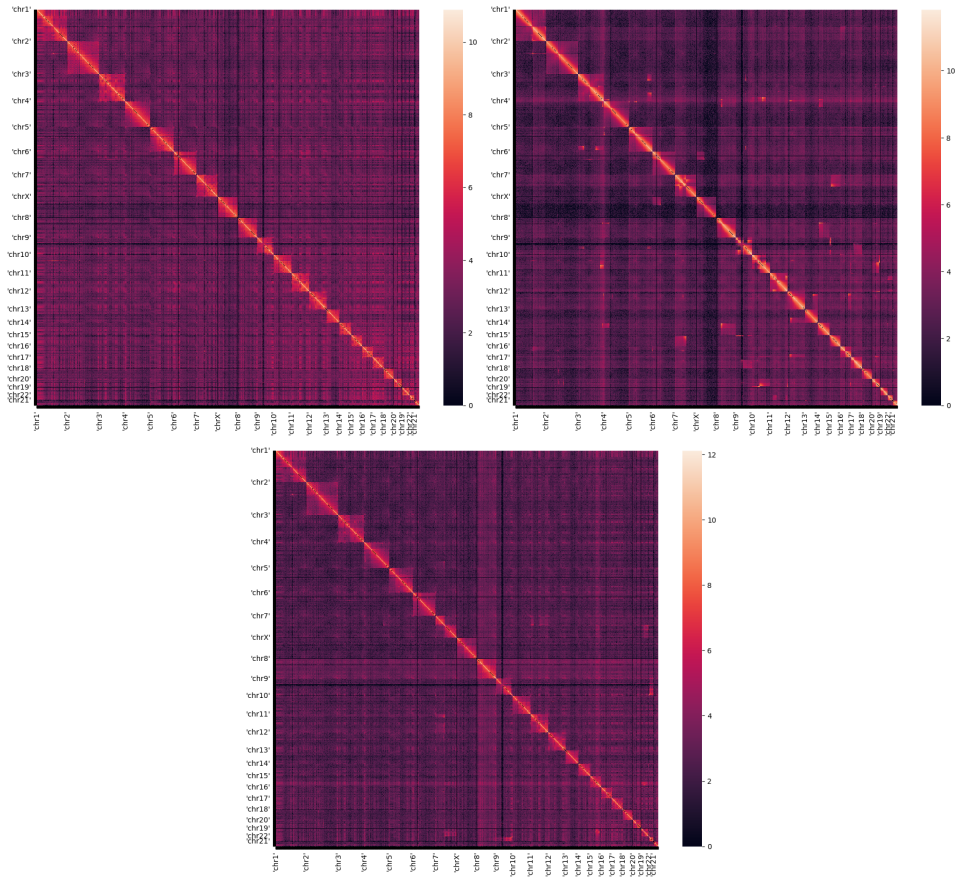


Figura 3.1: In alto a sinistra la mappa di contatto relativa alla linea cellulare GM12878, a destra T47D mentre in basso la mappa relativa alla linea KBM7

3.3 Autovalori e densità spettrale

L'autovalore massimale determina la stabilità dei pattern di interazione e ha inoltre un legame con la connettività media dei nodi: ne è una stima. Ci si può dunque aspettare un autovalore massimale maggiore nel caso sano rispetto ai casi aberranti e ciò si verifica. In tabella 3.1 si possono osservare gli autovalori massimali, il grado medio e il grado massimo delle tre matrici in esame. L'autovalore massimale della linea GM12878 è 8770.9, per la linea KBM7 è 8333.7 mentre quello della linea T47D è 8046.9. Quanto emerge è coerente con un grado di aberrazione crescente, è evidente infatti come la

mappa dei contatti della lina T47D sia quella con più traslocazioni. Si verifica inoltre che gli autovalori massimali siano compresi tra strength medio e strength massimo dei rispettivi network.
Anche il grado medio conferma l'ipotesi di una maggiore connessione globale nella rete sana mentre lo strength massimo presenta il valore più alto in corrispondenza della linea KBM7.
In figura 3.2 si possono vedere le distribuzioni degli autovalori delle tre matrici di adiacenza. Per ottenere una migliore visualizzazione è stato escluso l'autovalore massimale e indicate con una freccia le frequenze degli autovalori al di fuori del bulk. Si nota come all'aumentare del grado di aberrazione della linea la distribuzione diventi più allargata. Tale tendenza risulta ancora più evidente nella densità spettrale in 3.2. Si nota come già nella linea KBM7 la densità assuma una forma più triangolare, ancora più accentuata nella linea T47D. Se da un lato l'autovalore massimale nettamente staccato dal bulk può far pensare a un network casuale ER, dall'altro, l'andamento della densità dei network può suggerire un'altra tipologia dei rete. Infatti, una forma di tipo triangolare nella densità spettrale è tipica dei network a invarianza di scala [7][15].

	λ_1	strength medio	strength max
GM12878	8770.9	8479.6	10687.6
KBM7	8333.7	8122.9	13159.2
T47D	8046.9	7678.6	12242.7

Tabella 3.1

3.4 Clustering degli autovalori

La presenza di raggruppamenti negli autovalori può indicare delle comunità sottostanti nel network. Tramite l'algoritmo DBSCAN è stato effettuato un clustering degli autovalori per le 3 matrici e successivamente sono stati confrontati i cluster emergenti per diversi valori dei parametri dell'algoritmo. Utilizzando come parametri $eps = 1$ e $min\_samples = 1$ si ottengono 55 cluster per la linea GM12878, 74 per la linea T47D e 58 per la linea KBM7. Abbiamo poi reso variabili i parametri dell'algoritmo di clustering per evidenziare quale linea cellulare mostrasse più cluster in linea generale. I parametri variano per epsilon da 1 a 20 e per il numero minimo di punti da 1 a 10 e considerando tutte le combinazioni, la linea cellulare GM12878 nell'84% dei casi presenta un numero minore o uguale di cluster rispetto alla linea T47D (fig. 3.3). Nel 61% dei casi inoltre, il numero di cluster della linea KBM7 è compreso tra il rispettivo valore delle altre due linee. Si può considerare questa come una piccola evidenza di una maggior connettività e maggior compattezza della rete nel caso sano. Si conferma anche una grado di aberrazione e quindi connessione intermedio per quanto riguarda la linea KBM7.

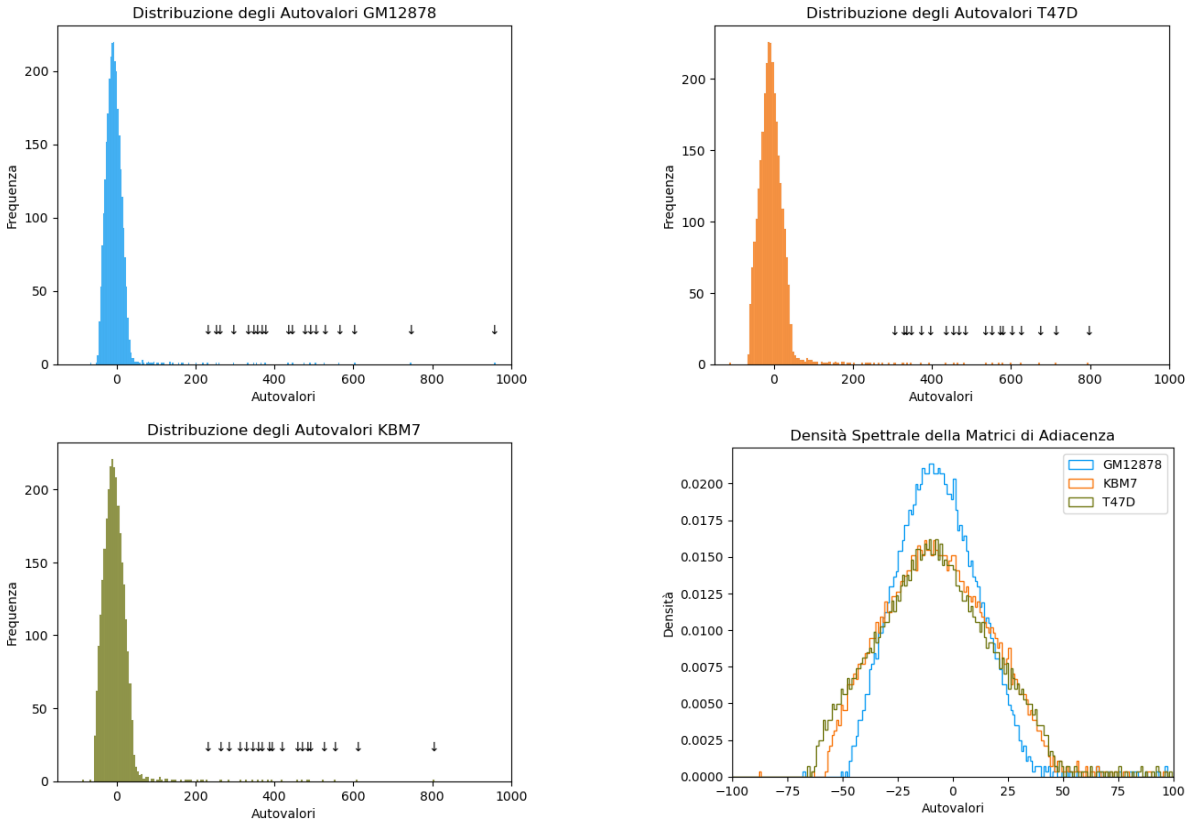


Figura 3.2: In alto a sinistra la distribuzione degli autovalori relativa alla linea cellulare GM12878, a destra T47D. In basso a sinistra gli autovalori della linea KBM7. Ai fini di una migliore visualizzazione è stato tagliato l'autovalore massimale e indicati con delle frecce i 19 autovalori più grandi. In basso a destra le densità spettrali sovrapposte

3.5 Analisi degli autovettori e matrici essenziali

Proviamo ora ad analizzare singoli autovettori per verificare se sia possibile trovare traccia delle traslocazioni. Le componenti degli autovettori infatti sono in corrispondenza con i nodi della rete e possono segnalare delle anomalie nei pattern di interazione del network sottostante. In figura 3.4 sono rappresentate le componenti degli autovettori 1, 9 e 15. Nell'autovettore 1, corrispondente all'autovalore massimale non troviamo caratteristiche sintomatiche di alcune aberrazioni se non il segno invertito delle componenti, che tuttavia non risulta significativo in questi termini. Nell'autovettore 9 e 15 possiamo trovare delle corrispondenze, per la matrice T47D con le aberrazioni chr 3-chr 5 e chr 13-chr 14. É tuttavia evidente come non si possa asserire con certezza che queste caratteristiche siano correlate con le aberrazioni evidenziate. Anche le componenti degli autovettori della linea sana mostrano valori più alti non attribuibili a traslocazioni evidenti, non

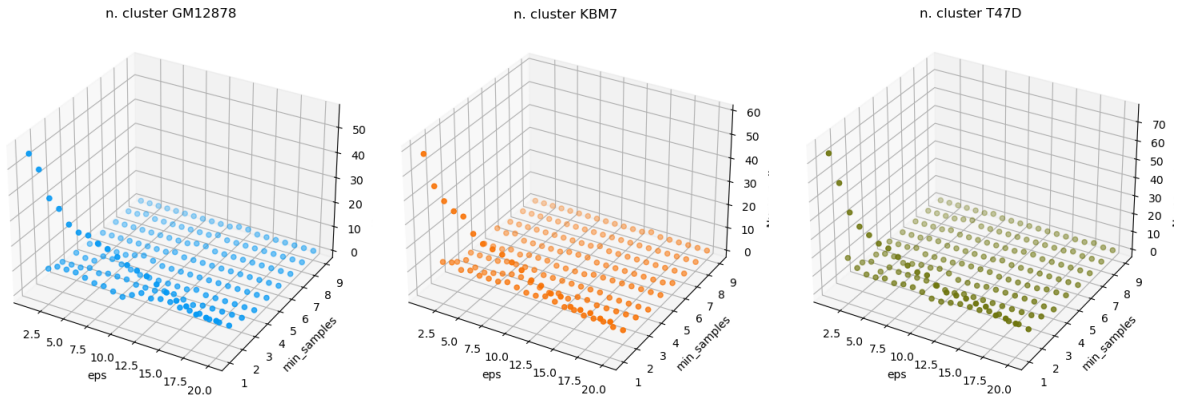


Figura 3.3: In alto a sinistra il clustering degli autovalori relativi alla linea cellulare GM12878, a destra T47D mentre in basso il clustering degli autovalori della linea KBM7.

possiamo quindi trovare una corrispondenza solida tra componenti particolarmente alte e aberrazioni nei contatti della matrice di adiacenza.
Il passo successivo nello studio degli autovettori è il calcolo delle matrici essenziali relative alle 3 linee cellulari utilizzando un numero crescente di proiettori o autospazi. In figura 3.5 sono visibili le 4 matrici essenziali della linea GM12878 costruite con 10,15,20 e 25 autovettori. Lo stesso è stato fatto per le altre due linee, e i risultati sono visibili in figura 3.6 e 3.7. Si nota chiaramente come siano necessari almeno 20 autovettori per ricostruire con chiarezza la struttura cromosomica. Tuttavia, nel caso della linea T47D si nota come l'utilizzo di pochi autovettori generi degli artefatti al di fuori della diagonale che via via vanno attenuandosi nelle matrici successive. Ancora più significativa è la presenza di questi artefatti nella linea sana GM12878, che non mostra traslocazioni evidenti nella mappa intera. Questo suggerisce che l'osservazione di singoli o pochi autovettori può essere fuorviante nell'analisi di singole traslocazioni intercromosomiche e conferma l'impossibilità di utilizzare solamente gli autovettori per individuare spettralmente le aberrazioni nel codice genetico delle linee cellulari.

3.6 Verifica delle ipotesi di RMT

Per verificare e quantificare il grado di casualità di una matrice esistono molti metodi e test tra i quali la rigidità spettrale Δ_3 o il calcolo della NNSD (Nearest Neighbor Spacing Distribution)[15][9]. Per questo studio è stata calcolata la densità di probabilità del rapporto tra spazi di autovalori consecutivi $P(r)$[2]. In figura 3.8 si può vedere un risultato significativo: tutte le 3 linee cellulari, o meglio, le distribuzioni dei rapporti seguono in tutti i casi l'andamento previsto dalla RMT, equazione 2.13, indicando una comune componente casuale nelle 3 reti in esame. Abbiamo usato $Z_\beta = 8/27$ e $\beta = 1$,

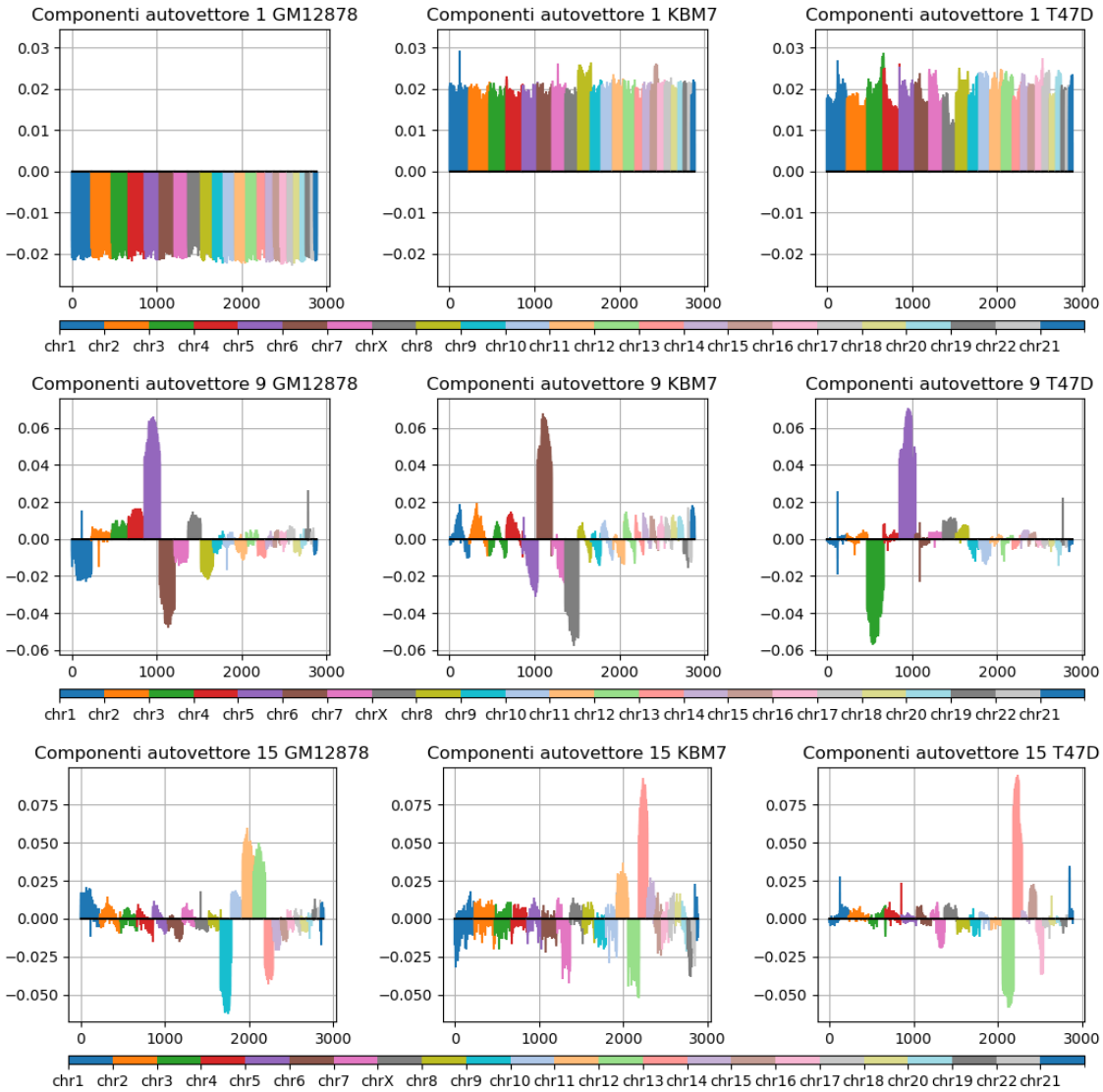


Figura 3.4: Plot delle componenti degli autovalori 1, 9 e 15 dall'alto al basso.

parametri validi per il caso GOE. Questo risultato non ci permette di quantificare la differenza nel grado di casualità tra i network ma fa emergere un'affinità tra le linee sana e malate che sembrava non esserci dalla semplice analisi della distribuzione degli autovalori. Ci si chiede allora il motivo di questa caratteristica comune nelle reti biologiche e ciò potrebbe essere frutto nei network di mutazioni avvenute senza alcun tipo di conseguenza. Secondo alcuni studi infatti, la casualità nelle connessioni è un aspetto essenziale per sistemi e reti che mostrano delle strutture sottostanti[13].

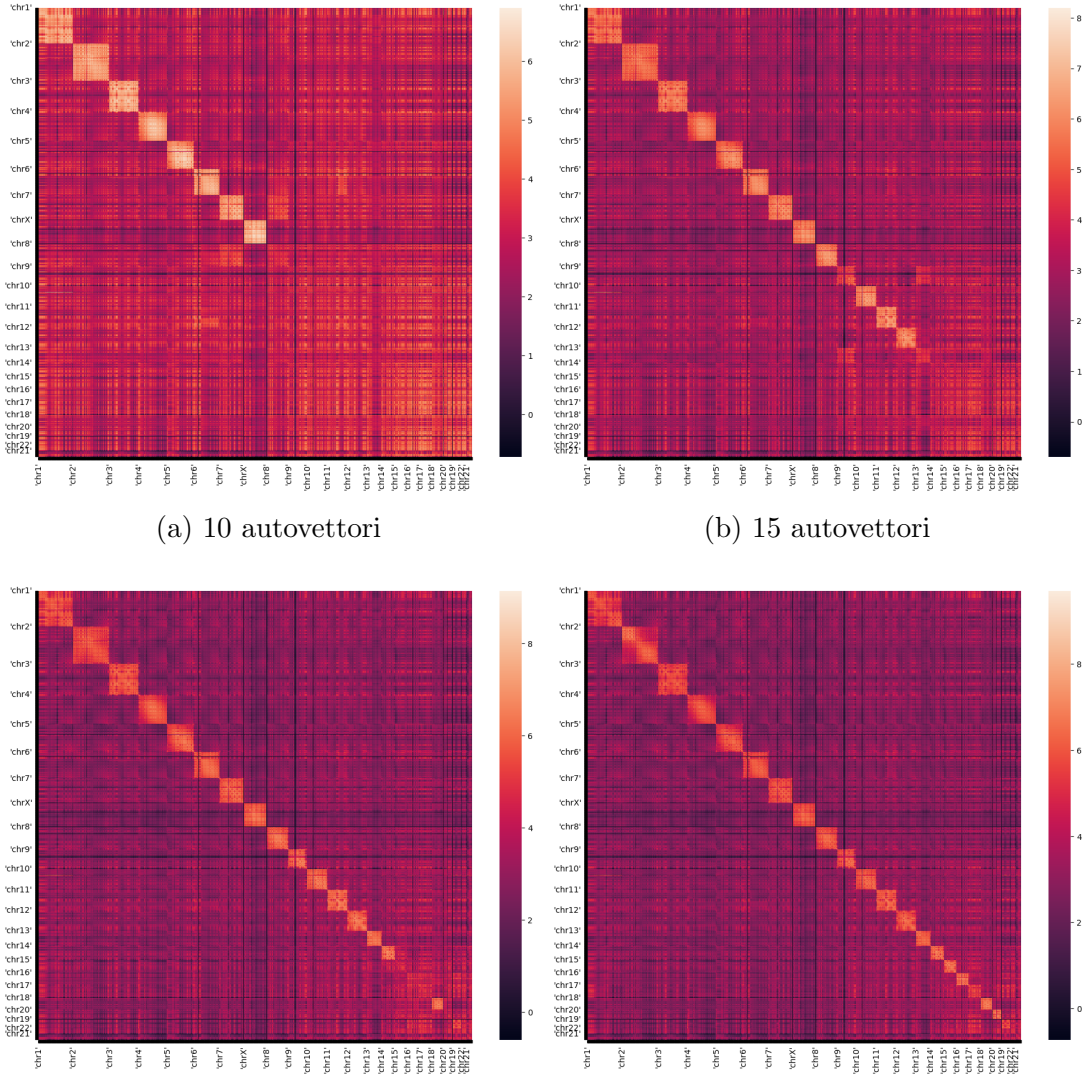
3.7 IPR

Se andiamo ora a calcolare l'Inverse Participation Ratio per i primi 25 autovettori otteniamo i risultati presentati in fig. 3.9. Si nota il massimo dell'IPR per l'autovettore 17, in corrispondenza del quale troviamo anche la maggior differenza tra linea più aberrante e linea sana (fig 3.10). Gli auotovettori che precedono, 15 e 16, confermano l'andamento dell'IPR in questa regione. Tuttavia, valori più alti dell'IPR vengono esibiti anche dalla linea sana e non sembrano esserci delle regioni con tendenze più marcate da parte di una linea in particolare. Più significativo sembra essere invece il calcolo dell'IPR per tutti gli autovettori delle 3 matrici. Sempre in figura 3.9, presentato in scala semilogaritmica, si vede come in questo caso ci sia una divisione più marcata tra le tre linee. In figura 3.9 vediamo gli ingrandimenti delle 3 zone di particolare interesse: fino all'autovettore numero 150, il valore dell'IPR è confrontabile tra le 3 linee con dei picchi da parte della linea T47D. Dall'autovettore 950 al 1200 vediamo una significativa divisione dei 3 campioni, con un primo picco a 1 da parte di GM12878 (una sola componente non nulla), seguita da un picco inferiore di T47D e un altro massimo di KBM7. In figura 3.10 si vedono alcuni tra questi autovettori più significativi: mostrano in modo evidente delle componenti più pronunciate. Infine, in corrispondenza degli autovettori finali notiamo un nuovo aumento dell'IPR, chiaramente più marcato per la linea GM12878. É interessante notare che i cromosomi interessati in queste anomalie sono sempre gli stessi: il 6, il 7 e il cromosoma sessuale X. I risultati emersi innescano un'analisi più approfondita di queste zone dello spettro. Ci saremmo aspettati un comportamento anomalo da parte dei primi autovettori, relativi agli autovalori più grandi mentre anche per autovalori più piccoli emergono delle caratteristiche che distinguono le 3 matrici.

3.8 Test di Shapiro-Wilk

Autovettori relativi ad autovalori molto piccoli sono associati alla parte casuale della matrice e mostrano solitamente una distribuzione delle componenti che tende ad una gaussiana. Per questo si può pensare che l'andamento delle componenti sia per così dire sempre più gaussiano man mano che ci spostiamo verso autovettori relativi ad autova-

lori piccoli. I risultati emersi dal calcolo dell'IPR innescano tuttavia un'indagine più approfondita sulla presenza di possibili autovettori non casuali che non siano tra i primi 25. Ci chiediamo se possa esserci una correlazione tra l'IPR degli autovettori e la gaussianità delle componenti. É possibile che nelle regioni in cui l'IPR è particolarmente elevato risiedano degli autovettori che si discostano nettamente dall'andamento normale, mentre per IPR bassi gli autovettori siano gaussiani. In figura 3.11 i risultati del test di Shapiro-Wilk confermano questa ipotesi: il valore del parametro statistico W si discostano dall'unità proprio in corrispondenza delle regioni a IPR elevato mentre il valore della probabilità dell'ipotesi nulla conferma che nelle altre zone gli autovettori presentano componenti a distribuzione gaussiana. Il parametro statistico W infatti, tanto più di discosta dall'unità e tanto più indica una distribuzione non gaussiana del campione mentre il valore della probabilità p al di sotto di una soglia, in questo caso pari a 0.05, indica il rigetto dell'ipotesi nulla del test, ossia la distribuzione normale del campione. Troviamo però una differenza significativa nell'ampiezza di queste regioni caratterizzate da autovalori con componenti distribuite in modo più casuale. La linea T47D infatti presenta delle zone gaussiane decisamente più ristrette della linea GM12878. Potremmo dire che la linea più aberrante presenta caratteristiche meno casuali della linea sana.
In figura 3.12 sono rappresentati gli scatter plot tra IPR in scala logaritmica e parametro statistico e indicano proprio questa correlazione tra i due parametri: all'aumentare dell'IPR, il parametro W diminuisce e discostandosi dall'unità indica una distribuzione non casuale. Si nota inoltre una biforcazione nei punti, più pronunciata per la linea T47D, che potrebbe essere collegata a due sottoinsiemi di autovettori più gaussiani visibili chiaramente nel grafico del valore p (fig 3.11). Questo ci indica che probabilmente i primi autovettori delle matrici non sono gli unici a contenere informazioni strutturali caratteristiche del network ma devono essere presi in considerazione anche altri autovettori, corrispondenti ad autovalori più piccoli. Questo costringe a rivedere la costruzione delle matrici essenziali che considerano solo i primi autospazi per la ricostruzione della matrice di adiacenza.



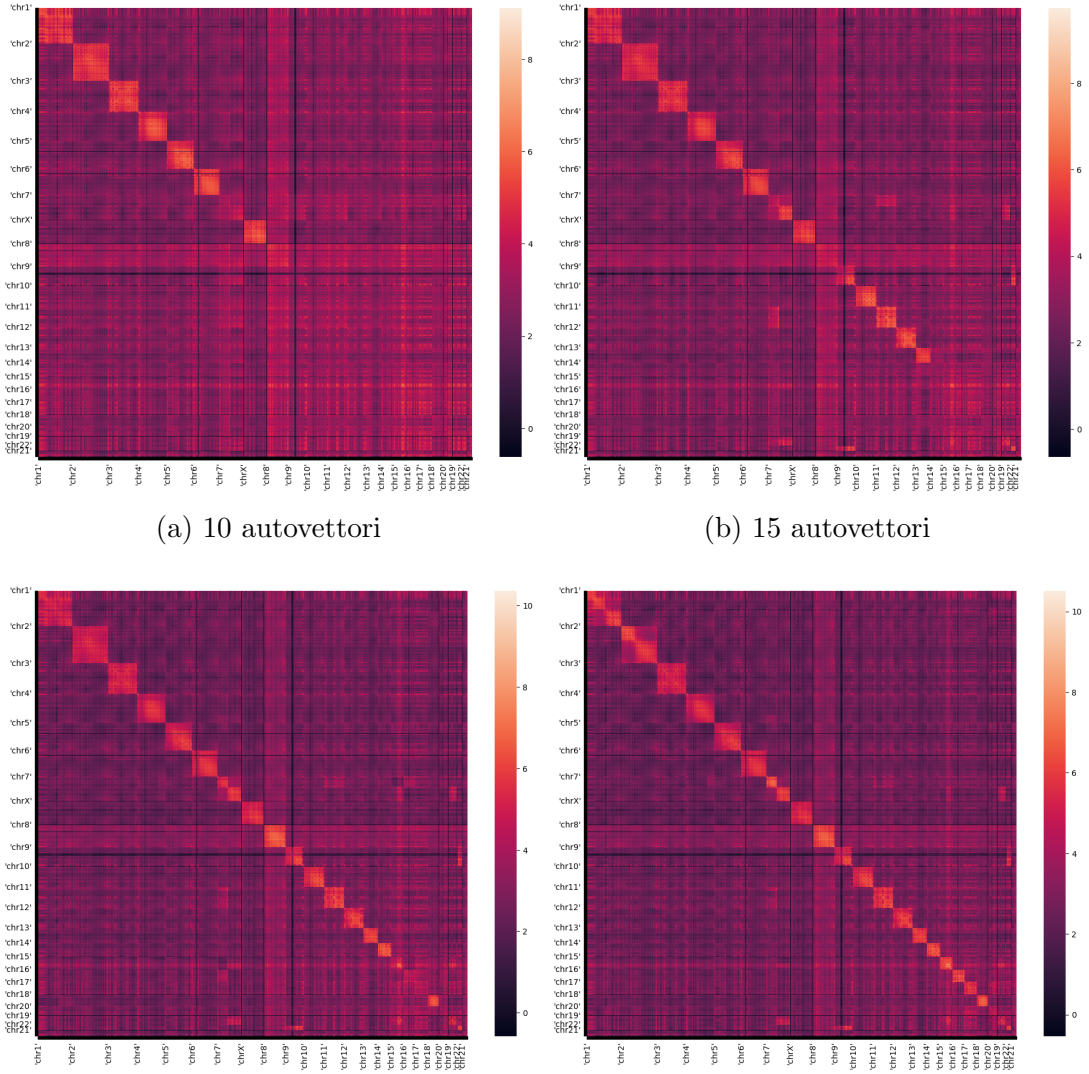
(a) 10 autovettori

(b) 15 autovettori

(c) 20 autovettori

(d) 25 autovettori

Figura 3.5: Matrici di contatto in forma essenziale della linea GM12878 costruite con un numero crescente di autovettori



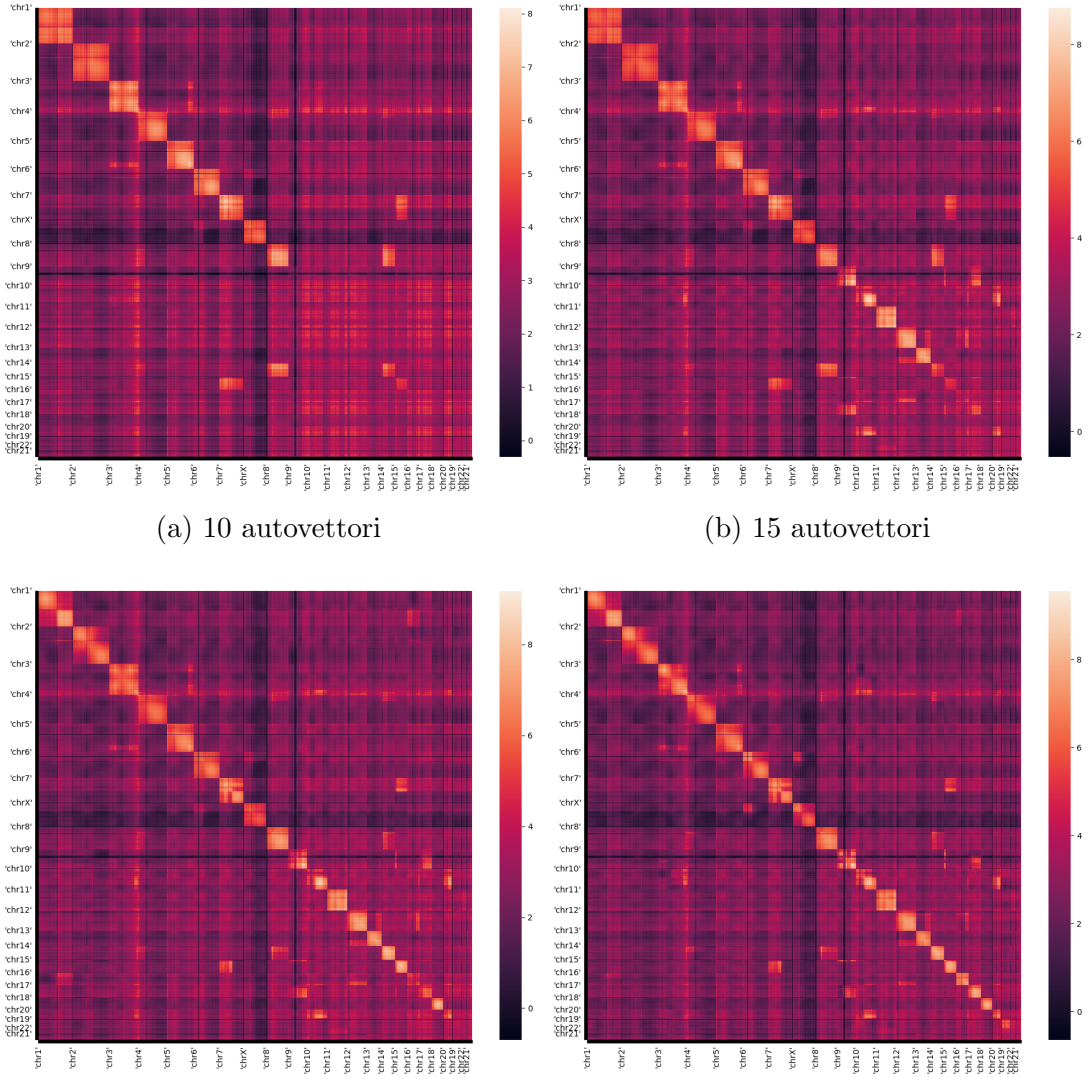
(a) 10 autovettori

(b) 15 autovettori

(c) 20 autovettori

(d) 25 autovettori

Figura 3.6: Matrici di contatto in forma essenziale della linea KBM7 costruite con un numero crescente di autovettori



(a) 10 autovettori

(b) 15 autovettori

(c) 20 autovettori

(d) 25 autovettori

Figura 3.7: Matrici di contatto in forma essenziale della linea T47D costruite con un numero crescente di autovettori

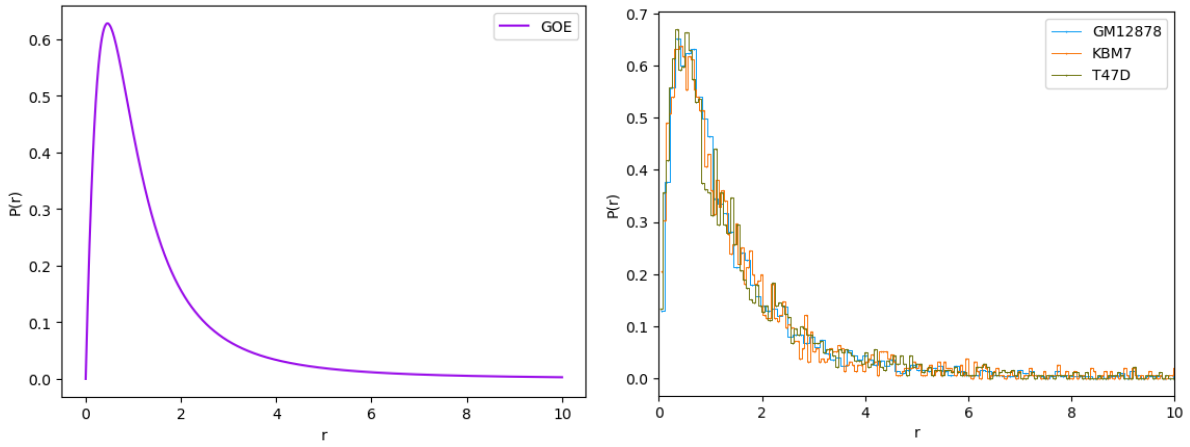


Figura 3.8: A sinistra l'andamento di $P(r)$ secondo la statistica GOE mentre a destra $P(r)$ calcolato per le matrici delle 3 linee cellulari

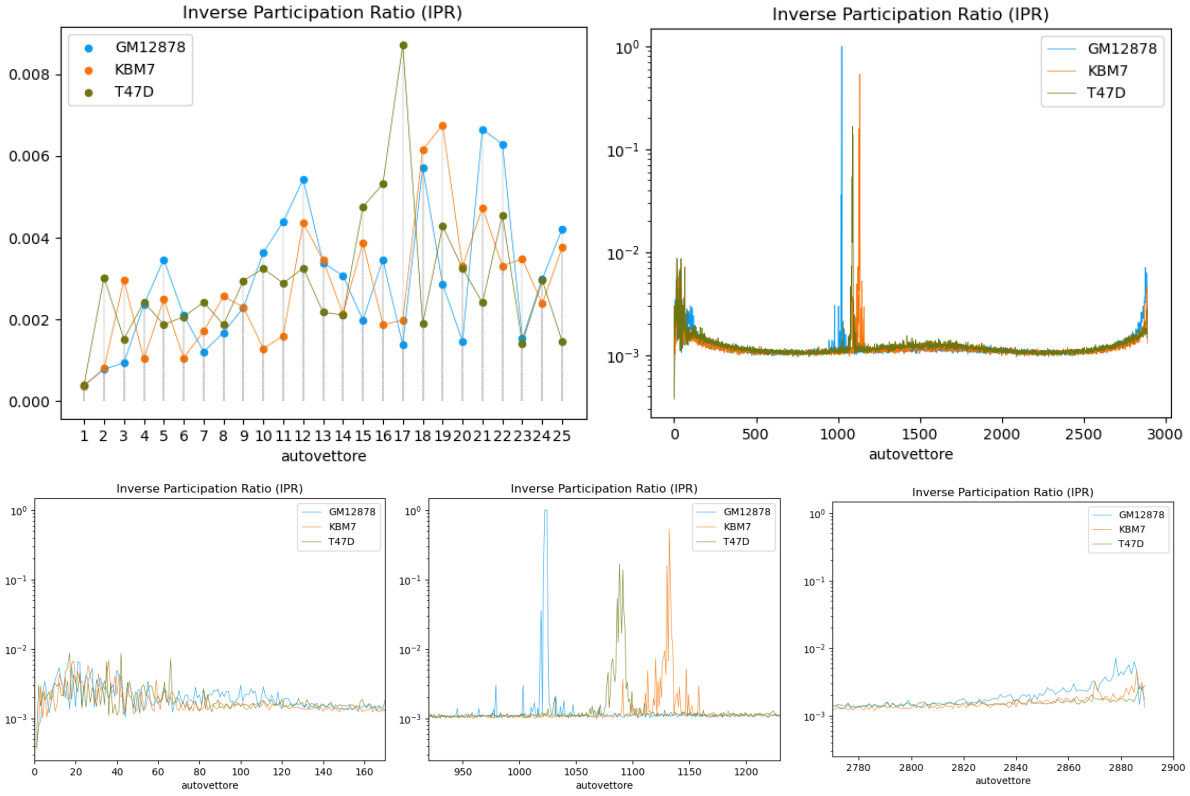


Figura 3.9: In alto a sinistra l'Inverse Participation Ratio calcolato per i primi 25 autovettori delle 3 matrici mentre a destra lo stesso calcolo per tutti gli autovettori. In basso, gli ingrandimenti delle tre zone dove gli autovettori delle matrici di adiacenza mostrano IPR significativi

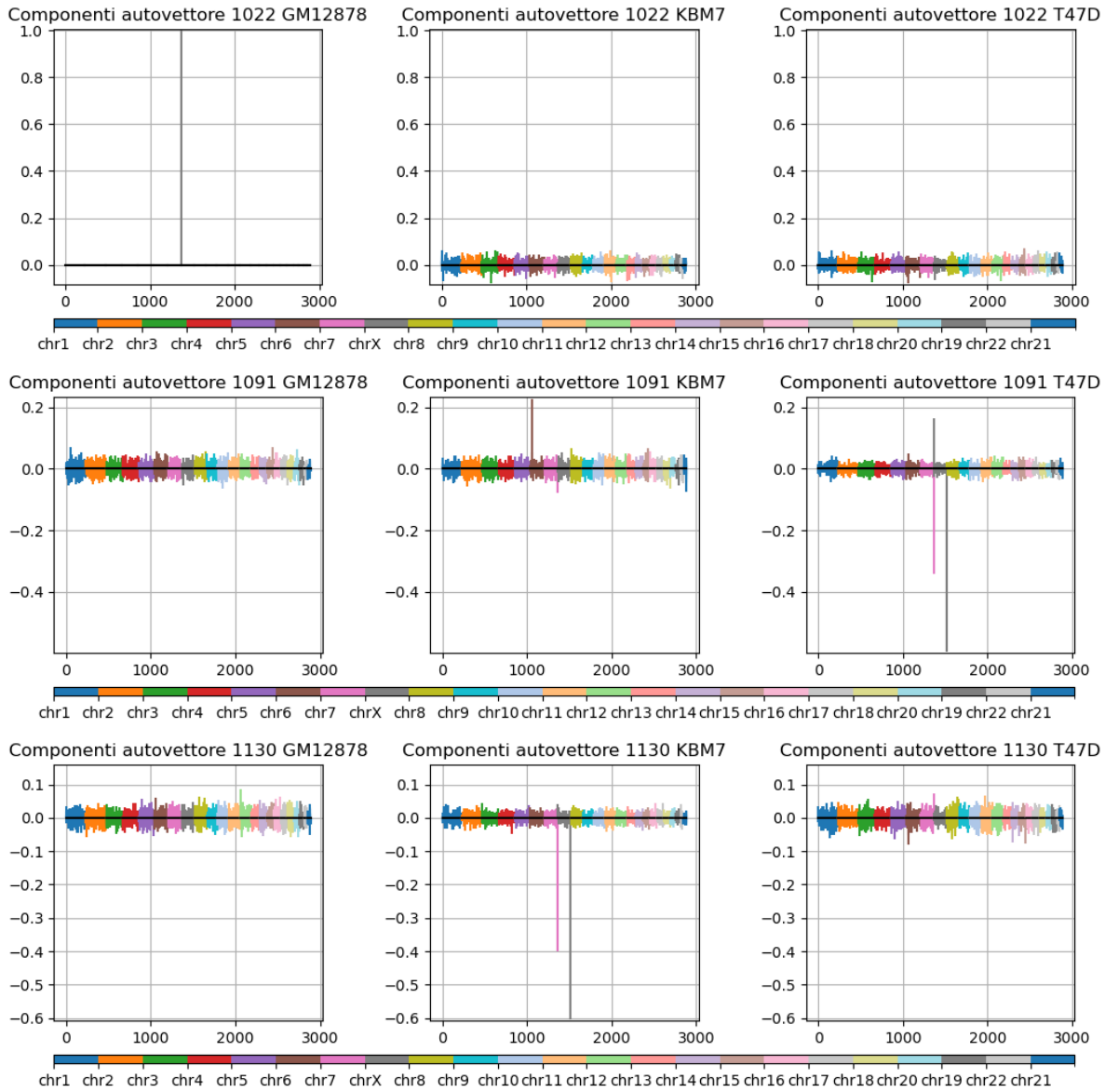


Figura 3.10: Plot delle componenti degli autovalori 1022, 1091 e 1130 dall'alto al basso.

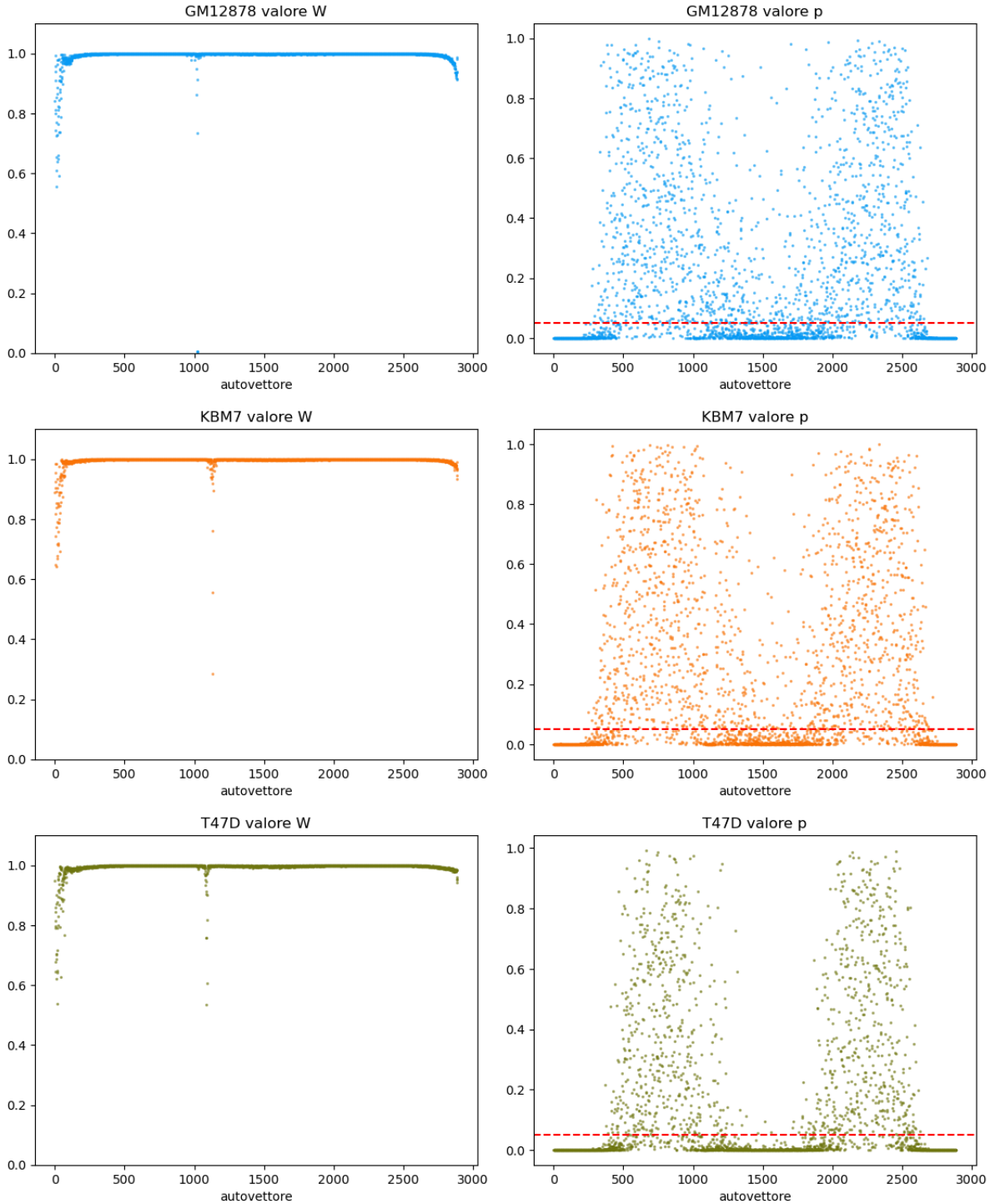


Figura 3.11: Test di Shapiro-Wilk per la gaussianità degli autovettori della linea GM12878, KBM7 e T47D. A sinistra il valore W e a destra il valore p

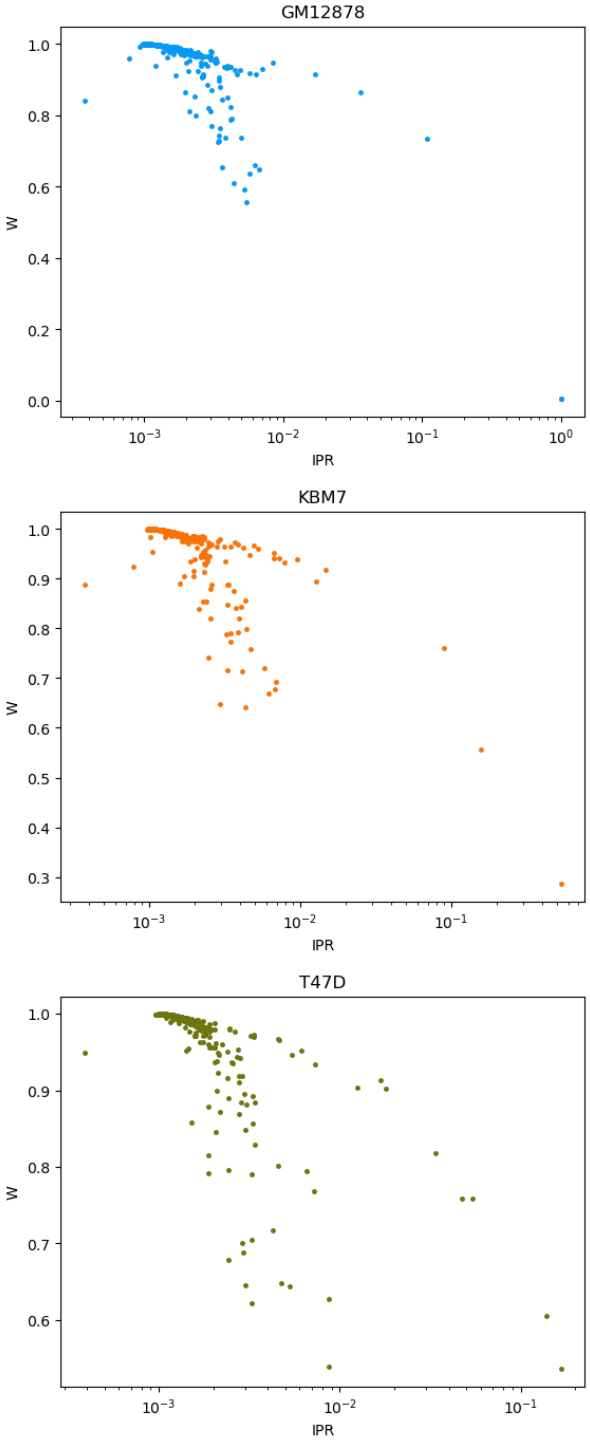


Figura 3.12: Scatter plot tra IPR e il parametro W del test di Shapiro-Wilk associati agli autovettori GM12878, KBM7 e T47D

Conclusioni e sviluppi

Il lavoro qui presentato costituisce quella che può essere una fase preliminare dell'analisi spettrale delle matrici Hi-C. Non si può decretare con certezza l'effettiva efficacia di queste tecniche per identificare la presenza di aberrazioni cromosomiche ma si possono tuttavia sottolineare alcuni aspetti emersi che indicano una possibile direzione per della analisi future.
Nonostante il ridotto numero di matrici a disposizione è evidente dai risultati come un approccio globale allo studio dello spettro sia in grado di restituire differenze non trascurabili tra le linee cellulari. La differenza nella distribuzione del bulk degli autovalori è evidente e differenzia significativamente matrice sana e matrici aberranti. I risultati non sono sufficienti per poter definire con certezza la tipologia di network sottostante, tuttavia pare che le 3 linee esibiscano un bulk degli autovalori triangolare, tipico delle reti a invarianza di scala. Alcune elementi suggeriscono però che non sia un modello scale-free "puro": il bulk infatti non è simmetrico rispetto all'autovalore nullo e non mostra la tipica degenerazione in corrispondenza di tale autovalore, il che fa ipotizzare una possibile componente di tipo hierarchical nei meccanismi di formazione del network. Inoltre, l'autovalore massimale, nettamente distanziato dal bulk è una tipica caratteristica dei network di tipo casuale ER. A conferma della componente randomica presente in tutte e 3 le matrici, la distribuzione dei rapporti tra spazi in autovalori consecutivi segue quella prevista dalla Random Matrix Theory. Ci si può chiedere quindi per quale motivo i network mostrino casualità solo su certi fronti e non per esempio nell'andamento della densità spettrale. Preme sottolineare che la legge del semicerchio di Wigner è una legge limite che si potrebbe paragonare al teorema del limite centrale per variabili casuali. Si può pensare quindi che la convergenza verso caratteristiche puramente casuali abbia rate diversi e che le matrici in esame, seppur mostrino un certo grado di compatibilità con la RMT non abbiano una dimensionalità sufficientemente elevata per una compatibilità totale. O in alternativa, in quell'incompatibilità potrebbe risiedere la parte specifica della matrice che racchiude le caratteristiche strutturali del codice genetico e che diversifica le 3 reti genetiche. L'IPR conferma in una certa misura questa ipotesi: abbiamo visto come le 3 linee mostrino IPR significativi in sole 3 regioni nella sequenza ordinata degli autovettori e in quella centrale siano nettamente separate nel raggiungere il picco. Il test di Shapiro-Wilk sottolinea che in queste regioni le componenti degli autovettori mostrano

bassa gaussianità nella loro distribuzione e che la linea aberrante mostra un grado di casualità inferiore rispetto alla linee aberranti KBM7 e T47D. É come se l'affinità con la RMT, correlata all'Inverse Participation Ratio fosse associata alla salute del network genetico. Urge ripetere un'altra volta che 3 matrici non sono sufficienti a determinare con certezza queste caratteristiche in relazione alle differenze spettrali tra network sani e malati ma già con questi pochi dati a disposizione emergono della differenze da tenere in considerazione.
Non si può dire lo stesso per l'analisi degli autovettori. Questi infatti non mostrano delle corrispondenze significative con le traslocazioni visibili nelle mappe dei contatti, o meglio le mostrano, ma ne esibiscono anche dove le traslocazioni non sono presenti. L'impossibilità di individuare traslocazioni dalle componenti di singoli autovettori viene confermata dalle matrici essenziali. Si nota infatti che ricostruendo le matrici usando un numero non sufficiente di autovettori generi nelle mappe di contatto degli artefatti attribuibili erroneamente a traslocazioni poiché si attenuano quando vengono aggiunti altri autovettori
Da ciò che è emerso pare quindi che gli autovalori siano uno strumento più efficace per evidenziare differenze tra i network sottostanti e che la presenza di caratteristiche tipiche delle matrici causali sia un possibile indice di salute del network.
Si apre da questo studio un ampio spettro di possibili approfondimenti: sicuramente è da indagare l'aspetto della componente casuale nei network e la discrepanza tra densità spettrale e distribuzione dei rapporti tra spazi $P(r)$ nell'esibire caratteristiche tipiche di modelli diversi. Rimane da approfondire anche la rilevanza strutturale di tutti quegli autovettori associati ad autovalori piccoli che solitamente vengono trascurati. Per cercare evidenze più chiare sulla tipologia di network si potrebbe cercare una soglia per rendere binarie le matrici di adiacenza e calcolare la distribuzione dei gradi per tutte le linee.
Sono molti gli scenari che si aprono, possiamo dire però che i risultati emersi qui hanno bisogno in primis di altre numerose convalide ripetendo le analisi per molte altre matrici Hi-C.

Bibliografia

[1] Bruce Alberts. *Molecular biology of the cell*. Garland science, 2017.

[2] YY Atas, Eugene Bogomolny, O Giraud, and G Roux. Distribution of the ratio of consecutive level spacings in random matrix ensembles. *Physical review letters*, 110(8):084101, 2013.

[3] Jayendra N Bandyopadhyay and Sarika Jalan. Universality in complex networks: Random matrix analysis. *Physical Review E*, 76(2):026109, 2007.

[4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[5] Fan Chung, Linyuan Lu, and Van Vu. Spectra of random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 100(11):6313–6318, 2003.

[6] Dragoš Cvetković, Peter Rowlinson, and Slobodan Simić. An introduction to the theory of graph spectra. *(No Title)*, 2009.

[7] Marcus Aloizio Martinez de Aguiar and Yaneer Bar-Yam. Spectral analysis and the dynamic response of complex networks. *Physical Review E*, 71(1):016106, 2005.

[8] Stefano Franzini, Marco Di Stefano, and Cristian Micheletti. esshi-c: essential component analysis of hi-c matrices. *Bioinformatics*, 37(15):2088–2094, 2021.

[9] Sarika Jalan and Jayendra N Bandyopadhyay. Randomness of random networks: A random matrix analysis. *Europhysics Letters*, 87(4):48010, 2009.

[10] Anton Krumm and Zhijun Duan. Understanding the 3d genome: emerging impacts on human disease. In *Seminars in cell & developmental biology*, volume 90, pages 62–77. Elsevier, 2019.

[11] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[12] ML Mehta and Random Matrices. the statistical theory of energy levels, 1967.

[13] Aparna Rai, A Vipin Menon, and Sarika Jalan. Randomness and preserved patterns in cancer network. *Scientific reports*, 4(1):6368, 2014.

[14] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[15] Camellia Sarkar and Sarika Jalan. Spectral properties of complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(10), 2018.

[16] Sang-Hyun Song and Tae-You Kim. Ctcf, cohesin, and chromatin in human cancer. *Genomics & informatics*, 15(4):114, 2017.

[17] Nynke L Van Berkum, Erez Lieberman-Aiden, Louise Williams, Maxim Imakaev, Andreas Gnirke, Leonid A Mirny, Job Dekker, and Eric S Lander. Hi-c: a method to study the three-dimensional architecture of genomes. *JoVE (Journal of Visualized Experiments)*, (39):e1869, 2010.

[18] Piet Van Mieghem. *Graph spectra for complex networks*. Cambridge university press, 2023.