Politecnico di Milano

Scuola di Ingegneria Industriale e dell'Informazione - Milano Leonardo

Corso di Studi in Ingegneria Matematica

Analisi di forma dei profili ChIP-Seq

Relatore:
Prof. Piercesare Secchi
Correlatore:
Dott.ssa Marzia Cremona

Tesi di Laurea di:
Alice Parodi
Matr. n. 784245

Anno Accademico 2012-2013

Alla mia famiglia

Indice

Sommario

Questa Tesi si inserisce nel contesto dell'indagine epigenetica e riguarda la caratterizzazione di forma dei dati derivanti dall'analisi mediante ChIP-Sequencing dell'interazione proteina-DNA. In particolare in questo lavoro si descrive nel dettaglio la tecnica di raccolta ed elaborazione dei dati per procedere poi ad un'analisi statistica del segnale ottenuto. In primo luogo si prevede un'indagine sul modello statistico alla base dei dati, con la valutazione delle ipotesi proposte in letteratura e la formulazione di un buon modello di generazione. Entrando nel dettaglio dell'analisi statistica, poi, si indaga sulla possibilità di distinguere i diversi tipi di interazione proteina-DNA mediante tecniche di classificazione non supervisionata, basate sulla valutazione di forma dei dati. L'algoritmo alla base di questa clusterizzazione è l'algoritmo del *k-mean* valutato in primo esame su indici di forma opportunamente scelti per rappresentare i dati e in un secondo tempo sui dati funzionali nel loro complesso. A tale scopo è necessario introdurre il *k-mean alignment*, ovvero l'adattamento dell'algoritmo ai dati di tipo funzionale che ammette anche la registrazione dei dati in esame. La tesi si conclude con alcune considerazioni legate all'interpretazione biologica dei risultati proposti e con la presentazione di alcuni interessanti spunti di riflessione e obiettivi per l'ulteriore sviluppo del progetto.

Abstract

This Thesis, introduced in the epigenetics investigation, concerns the characterization of the shape of the data resulting from the ChIP-Sequencing analysis. In particular in this work we describe in detail the technique of collection and processing of data and proceed, then, to an analysis of these data. First, we provide a survey on the statistical model underlying the data, with the evaluation of the hypotheses proposed in the literature, and then we propose the formulation of a good generation model. Then, going into detail of the statistical analysis, this project investigates the possibility of distinguishing the different types of protein-DNA interaction by unsupervised classification based on the evaluation of the shape of the data. The algorithm at the basis of this clustering is the algorithm of *k-mean* evaluated, in the first examination, on shape indices suitably chosen to represent the data and, in a second time, on the overall structure of the data. In this last case the algorithm of the *k-mean* is appropriately adapted (*k-mean alignment*) to identify the components of variability interesting for the classification of functional data. In the last part of the Thesis we propose some biological interpretation of the results and we suggest some interesting ideas and objectives for the further development of the project.

Introduzione

Con la conclusione nel 2003 dell'*Human Genome Project*, che ha presentato il sequenziamento completo del DNA umano e la definizione delle regioni associate alla codifica delle proteine, la ricerca in ambito genetico ha concentrato i suoi interessi sull'espressione dei geni codificanti. Si è riscontrato uno stretto legame tra l'espressione genica, la differenziazione cellulare e la presenza di proteine attorno alla molecola di DNA. In questi anni, dunque, un ambito di ricerca in fase di sviluppo riguarda lo studio della connessione tra la presenza di proteine attorno alla molecola di DNA e la manifestazione fenotipica della cellula. Il progetto congiunto IEO-IIT-PoliMi denominato *Genomic Computing* si colloca proprio in tale contesto. La Tesi di Laurea si inserisce in questo progetto e si pone l'obiettivo di valutare gli esiti di esperimenti di ChIP-Seq (*Chromatine ImmunoPrecipitation Sequencing*): vengono proposte opportune tecniche statistiche con lo scopo di determinare le caratteristiche di forma significative dei dati per permettere una loro classificazione e valutarne un possibile significato biologico. La tecnica del ChIP-Seq permette, infatti, di valutare le interazioni tra le proteine che circondano la molecola di DNA e la molecola stessa: risulta biologicamente interessante analizzare i possibili cambiamenti di questa interazione, in particolare focalizzando l'attenzione sulle alterazioni di forma dei dati ottenuti dal ChIP-Seq che allo stato dell'arte non vengono considerate dalle analisi.
Il lavoro, dunque, si sviluppa nel suo complesso in una definizione introduttiva degli esperimenti, con la collocazione nell'ambito della ricerca epigenetica, e in una conseguente analisi di forma dei dati.

In particolare nel Capitolo 1 si propone una definizione riassuntiva dell'ambito di indagine necessaria per la comprensione dei dati e per la caratterizzazione concreta degli obiettivi del lavoro: oltre alla definizione del processo di sequenziamento, infatti, si presentano nel dettaglio le possibili interazioni tra la forma dei dati di ChIP-Seq e il

legame DNA-proteina.

Si procede, poi, con una descrizione del processo di raccolta dati (Capitolo 2) analizzando in generale i passaggi per la definizione degli esisti di esperimenti, descrivendo una panoramica quanto più possibile generale, che poi si focalizza sui tool utilizzati in questo lavoro. Si vuole spiegare l'origine dei dati come funzioni in $\mathbb{Z}$; dalla loro completa comprensione, infatti, è possibile trarre spunto per le successive analisi.

A partire dal Capitolo 3 ci si addentra nell'analisi vera e propria dei dati. Si esamina, in primo luogo, la loro struttura alla ricerca di un modello statistico valido che li definisca; in letteratura, infatti, sono proposti differenti interpretazioni e si vuole valutare la loro coerenza. Si osserva che molti dei modelli proposti, che valutano le funzioni in $\mathbb{Z}$ come conteggi generati da una distribuzione di Poisson o Binomiale Negativa, non sono propriamente adatti a definire l'andamento dei dati. É, dunque, necessaria un'analisi più accurata per definire un buon modello teorico. Si introduce, pertanto, una definizione alternativa dei dati, basata sull'albero sotteso alla funzione e dall'analisi della struttura di questo albero si comprende che rientra nella particolare categoria degli alberi di Galton-Watson con distribuzione generatrice di Poisson e pertanto si definisce questo modello per la caratterizzazione dei dati.

Si procede, quindi, allo sviluppo dello scopo vero e proprio della Tesi, ovvero all'analisi della forma dei dati utilizzando differenti metodologie di classificazione e applicandole a un dataset reale di interazione tra il fattore di trascrizione GATA1 e il DNA.

In particolare nel Capitolo 4 ci si concentra su un'indagine basata sulla caratterizzazione di forma dei dati mediante la definizione di grandezze caratteristiche, alcune connesse alle funzioni vere e proprie, altre agli alberi sottesi precedentemente definiti. Questa caratterizzazione porta, quindi, alla presentazione di una prima classificazione dei dati, basata sull'algoritmo del *k-mean*.

Nel Capitolo 5 si procede al raffinamento della clusterizzazione precedentemente proposta; si valuta, infatti, una classificazione basta direttamente sulla struttura funzionale dei dati, che non necessiti della definizione degli indici di forma e che pertanto possa essere generalizzata a dati per esempio non derivanti dall'indagine dello specifico legame fattore di trascrizione-DNA. A seguito di una definizione introduttiva dei dati funzionali, si applica l'algoritmo del *k-mean alignment* ai profili, concentrandosi su diverse tipologie di allineamento e differenti valutazioni della similarità tra i dati. L'analisi completa

basata sulla valutazione complessiva dei dati, tuttavia, pur essendo efficace risulta computazionalmente problematica e pertanto si propone un'analisi di tipo *bootstrap* per definire una classificazione sufficientemente ampia.

Si conclude, infine, (Capitolo 6) con la presentazione di alcuni interessanti risultati biologici ricavati da questo tipo di indagine, ma anche con la definizione dettagliata dei molti sviluppi che questo lavoro richiede. Risulta necessaria, infatti, una validazione dei metodi proposti, che sembrano consistenti, dal momento che presentano tutti risultati confrontabili, ma che devono essere avvalorati da altri riscontri di robustezza. In particolare, risulta necessaria l'analisi di repliche tecniche e/o biologiche degli esperimenti ChIP-Seq; inoltre, i positivi risultati biologici ottenuti nell'esperimento analizzato in questa Tesi devono ricevere una conferma anche per altre proteine, per diverse tipologie di interazioni epigenetiche o con la scoperta di ulteriori connessioni con il fenotipo dei tessuti analizzati.

Capitolo 1

Epigenetica e ChIP-Seq

1.1 Dalla genetica all'epigenetica

Il termine ***genetica*** viene utilizzato per la prima volta dal biologo inglese William Bateson[1] in occasione di una conferenza sull'ibridazione tenutasi alla Royal Horticultural Society di Londra nel 1906; in questa sede lo scienziato definisce questa nuova *scienza dell'eredità e della variazione*, come lo *studio scientifico dei fattori responsabili delle somiglianze e delle differenze osservabili tra individui imparentati per discendenza*.

L'inizio dello studio del legame tra la manifestazione dei caratteri e i fattori genetici, tuttavia, risale a molti anni prima: si deve, infatti, a Gregor Mendel[2] la prima analisi sistematica e rigorosa dell'ereditarietà biologica; egli viene considerato il precursore della genetica moderna grazie alla formalizzazione delle leggi di segregazione e assortimento indipendente (Tabella 1.1), che tuttora sono alla base di tutti gli studi di genetica (1866: *Esperimenti sugli ibridi vegetali*).

Successive analisi hanno portato al raffinamento delle conclusioni sperimentali del naturalista ceco con la definizione delle unità caratteristiche della trasmissione ereditaria, i cromosomi, cioè le *strutture individuali e differenziate che contengono i fattori mendeliani* (Walter Sutton[3] e Theodor Boveri[4] nel 1902), e poi dei geni, cioè *le particelle che*

[1] **William Bateson**: genetista britannico (1861-1926), ebbe il merito di divulgare le ricerche di Mendel in lingua inglese.

[2] **Gregor Mendel**: frate agostiniano, naturalista e matematico ceco (1822-1884) considerato il precursore della genetica moderna per le sue ricerche sui caratteri ereditari.

[3] **Walter Sutton**: biologo statunitense (1877-1916).

[4] **Theodor Boveri**: biologo tedesco (1862-1915).

Leggi di Mendel	
Legge della Segregazione	Ogni individuo ha coppie di fattori per ciascun carattere. Quando due gameti si uniscono nella fecondazione il discendente riceve un allele da ognuno dei due genitori, pertanto quando due organismi omozigoti (uno per l'allele dominante e uno recessivo) si incrociano si generano quattro individui eterozigoti che manifestano l'allele dominante, mentre alla seconda generazione il rapporto fenotipico tra dominante e recessivo è 3:1.
Legge dell'Assortimento Indipendente	Alleli di un gene segregano indipendentemente da alleli di altri geni.

Tabella 1.1: Prima e Seconda legge di Mendel.

posseggono le proprietà mendeliane di segregazione e ricombinazione (1908: Vilhelm Johannsen[5]).
Un'ulteriore svolta nell'analisi genetica si ha quando nel 1935 il genetista George W. Beadle[6] e il biochimico Edward L. Tatum[7] dimostrano che la sintesi delle proteine, i costituenti fondamentali delle cellule, dipende dai geni: la mutazione di un gene influenza, infatti, la sintesi della proteina corrispondente (la formula *un gene - un enzima* che varrà agli scienziati il premio Nobel per la medicina e la fisiologia nel 1958). Si inizia, quindi, a delineare il meccanismo di definizione del fenotipo di un organismo, basato sul trasferimento del materiale genetico dei cromosomi tra generazioni e sulla codifica delle proteine a partire dai geni espressi.

A partire dagli anni '50 si intuisce l'esistenza di un'unica molecola che contiene tutte le informazioni genetiche, il DNA; la struttura a doppia elica di questa molecola (Figura 1.1) viene definita nel dettaglio da James Watson[8], Francis Crick[9] e Maurice Wilkins[10] nel 1953, anno che segna dunque l'inizio dell'era della *genetica del DNA*, ovvero dell'analisi della struttura di questa molecola, del suo ruolo nella codifica delle proteine e delle sue possibili mutazioni connesse ad alterazioni fenotipiche.

[5]**Vilhelm Johannsen**: botanico e genetista danese (1857- 1927).
[6]**George W. Beadle**: biologo statunitense (1903-1989), premio Nobel per la medicina nel 1958.
[7]**Edward L. Tatum**: genetista statunitense (1909-1975), premio Nobel per la medicina nel 1958.
[8]**James Watson**: biologo statunitense (1928) premio Nobel per la medicina nel 1962.
[9]**Francis Crick**: scienziato britannico (1916-2004) premio Nobel per la medicina nel 1962.
[10]**Maurice Wilkins**: biologo neozelandese naturalizzato britannico (1916-2004), premio Nobel per la medicina nel 1962.

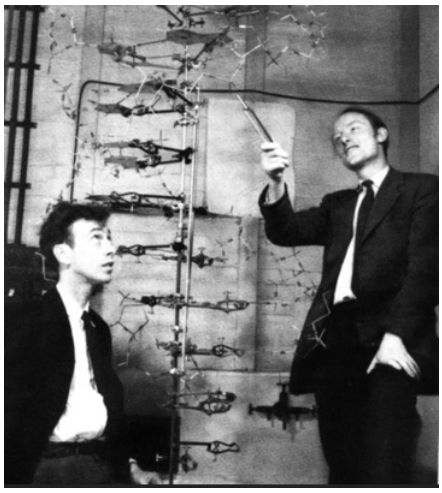


Figura 1.1: Struttura del DNA definita da Watson e Crick.

La ricerca sull'analisi del DNA culmina con l'avvio nel 1990 del progetto internazionale *Human Genome Project*[11] [7], che ha l'obiettivo di determinare la sequenza di nucleotidi che compongono il DNA umano e di definire i geni caratterizzanti la nostra specie, attraverso la definizione delle funzionalità e della regione del DNA associate a ciascuno di essi. Il progetto si è concluso con successo nel 2003 con il sequenziamento completo del DNA umano e con la definizione dei nuovi obiettivi di ricerca: ci si propone ora di analizzare i fattori esterni al DNA che sono responsabili della differenziazione cellulare. É noto, infatti, che la struttura della molecola è praticamente inalterata in tutte le cellule dell'organismo, ma la conformazione delle cellule varia a seconda della loro collocazione e dalla funzione che svolgono. Si avvia, pertanto, nell'ambito della ricerca *postgenomica* un'analisi ***epigenomica***, concernente cioè tutto ciò che è sopra ($\epsilon\pi i$) alla sequenza DNA, al fine di determinare quali sono i fattori rilevanti nella differenziazione e nella crescita delle cellule, con evidenti riscontri nell'analisi e nella cura di particolari patologie.

Per analizzare nel dettaglio gli obiettivi e le peculiarità dell'analisi epigenetica è necessario definire le caratteristiche tridimensionali della molecola di DNA. La doppia elica ha una lunghezza complessiva che può raggiungere i due metri, pertanto per essere contenuta nel nucleo di ogni cellula deve esse avvolta su se stessa in modo da occupare regioni sferiche di diametro al più pari a $10\mu m$. Gli avvolgimenti sono compiuti attorno a

[11] Progetto fondato dal National Institutes of Health e finanziato dall' US Department of Energy.

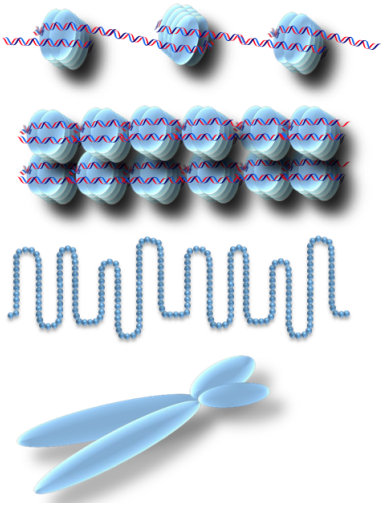


Figura 1.2: Avvolgimenti della molecola attorno a istoni e compattamento della cromatina per la costituzione dei cromosomi.

agglomerati proteici, detti istoni, individuati nel 1974 da Kornberg[12] e connessi al DNA, con cui vanno a formare i nucleosomi, attraverso legami idrogeno. Questo addensamento rappresenta il primo livello di compattamento della cromatina, che si avvolge ancora su se stessa per formare la caratteristica struttura dei cromosomi, come rappresentato in Figura 1.2.

L'analisi epigenomica si propone, dunque, di studiare i meccanismi molecolari mediante i quali la cellula altera il grado di attività dei geni senza modificare l'informazione in essi contenuta, ovvero senza modificare la sequenza di DNA, ma alterando ad esempio la struttura tridimensionale della molecola o il livello di espressione dei geni . Le alterazioni possono riguardare:

- cambiamenti del DNA senza variazioni in sequenza: metilazioni di basi come la citosina che provocano il silenziamento della regione dove avvengono;

- cambiamenti negli istoni dei nucleosomi, ovvero alterazioni nelle proteine attorno alle quali si addensa la cromatina;

[12] **Roger Kornberg**: biochimico e biologo strutturale statunitense (1947), studioso della forma tridimensionale del DNA e della connessione della struttura della molecola con la regolazione dei meccanismi di trascrizione, ricerche per le quali è stato insignito del premio Noble per la chimica nel 2006.

- variazioni nelle posizioni dei nucleosomi rispetto alla sequenza e quindi modifiche degli avvolgimenti dell'elica e conseguenti cambiamenti delle regioni espresse della sequenza (rimodellamento della cromatina).

Dai primi anni XXI secolo è possibile svolgere effettivamente ricerche epigenetiche grazie allo sviluppo di tecnologie all'avanguardia (NGS: *Next Generation Sequencing*) che permettono di superare gli evidenti limiti della tecnica dell'elettroforesi capillare (CE) utilizzata per lo sviluppo dell'Human Genome Project che permetteva di ottenere solo limitate informazioni genetiche da un dato sistema biologico. L'avvento del NGS permette di analizzare il genoma, il trascrittoma e l'epigenoma di tutte le specie in maniera più dettagliata e precisa. Queste tecniche sono basate sulla frammentazione della sequenza in analisi (ad esempio cDNA) in piccoli segmenti (*reads*) che possono essere facilmente sequenziati e quindi identificati; i *reads* sono poi riallineati su un genoma di riferimento noto (in alternativa può essere usato l'assemblaggio de novo), così che l'insieme dei *reads* allineati possa evidenziare le regioni di cDNA in analisi. [16]

In questo lavoro si vogliono analizzare i dati ottenuti con una particolare tecnica NGS chiamata *Chromatine ImmunoPrecipitation Sequencing* (ChIP-Seq) al fine di definire le interazioni tra proteine e la molecola di DNA.

1.2 ChIP-Sequencing

La tecnica del ChIP-Seq (*Chromatine ImmunoPrecipitation Sequencing*) si inserisce nel contesto delle tecniche NGS e permette di studiare l'interazione proteine-DNA e di selezionare le regioni della sequenza nucleotidica in cui tale interazione avviene; essa prevede diversi passi che culminano con l'allineamento dei *reads* con il genoma di riferimento, come tipico delle NGS.
La procedura di analisi consiste, come rappresentato in Figura 1.3 nelle seguenti fasi:

1. rafforzare il legame tra le proteine che circondano il DNA e la sequenza stessa (tramite formaldeide);

2. rompere la sequenza in frammenti tramite un processo di sonicazione;

3. selezionare i frammenti di interesse, quelli cioè connessi alla proteina scelta, tramite il legame con uno specifico anticorpo e l'immunoprecipitazione dei frammenti così selezionati;

4. purificare i frammenti di DNA rompendo i legami precedentemente fortificati;

5. sequenziare la parte iniziale (*single-end*) o le parti iniziale e finale (*paired-end*) dei frammenti, ottenendo i *reads*[13];

6. allineare i reads così ottenuti alla sequenza di DNA di riferimento;

7. selezionare le regioni di interesse.

Analizzando nel dettaglio i dati derivanti dal sequenziamento (Figura 1.5) si può ipotizzare che la sonicazione permetta la rottura della sequenza di DNA solo nei tratti esterni alla proteina, ma che, a causa della selezione di frammenti di interesse con l'immunoprecipitazione, devono comprendere la regione a cui la proteina è connessa. Si assume, quindi, che i *reads* ricoprano la parte di DNA circondata dalla proteina di interesse. Per quanto riguarda la lettura dei *reads* è necessario precisare che le tecniche di lettura delle sequenze di DNA, data la struttura di questa molecola, prevedono la possibilità di leggere soltanto lungo una direzione della sequenza. La doppia elica, infatti, è costituita da zuccheri a cui sono legate le basi azotate (Figura 1.4), a seconda dell'estremità libera degli zuccheri si definisce l'estremità 3' o 5' del filamento della doppia elica ed è possibile leggere soltanto da 5' a 3'. Per riconoscere le basi azotate che compongono i *reads* isolati dall'immunoprecipitazione è necessario separare i filamenti di doppia elica e, a seconda di quale viene selezionato per la decodifica, si procede alla lettura dal suo inizio o della sua fine[14]. In particolare si leggerà il filamento che inizia con l'estremità 5' libera e, a seconda che questo sia collocato all'inizio del *reads* o alla fine, questo va a comporre l'insieme dei frammenti marcati con positivi (frammenti blu dell'immagine) o negativi (rossi). A seguito della lettura di un numero sufficiente di basi (generalmente pari a 30-50 basi) si hanno i *reads* pronti per l'allineamento alla sequenza completa di DNA.
Sul genoma si riconoscono, dunque, le regioni in cui sono allineati i *reads* positivi e

[13] Qui si concludono le fasi del ChIP-Seq, i punti successivi sono per l'analisi dei risultati.

[14] La lettura di una sola estremità del frammento, come per i dati in esame in questo lavoro, è tipica del *single-end sequencing*, mentre nel caso del *paired-end sequencing* vengono letti entrambi.

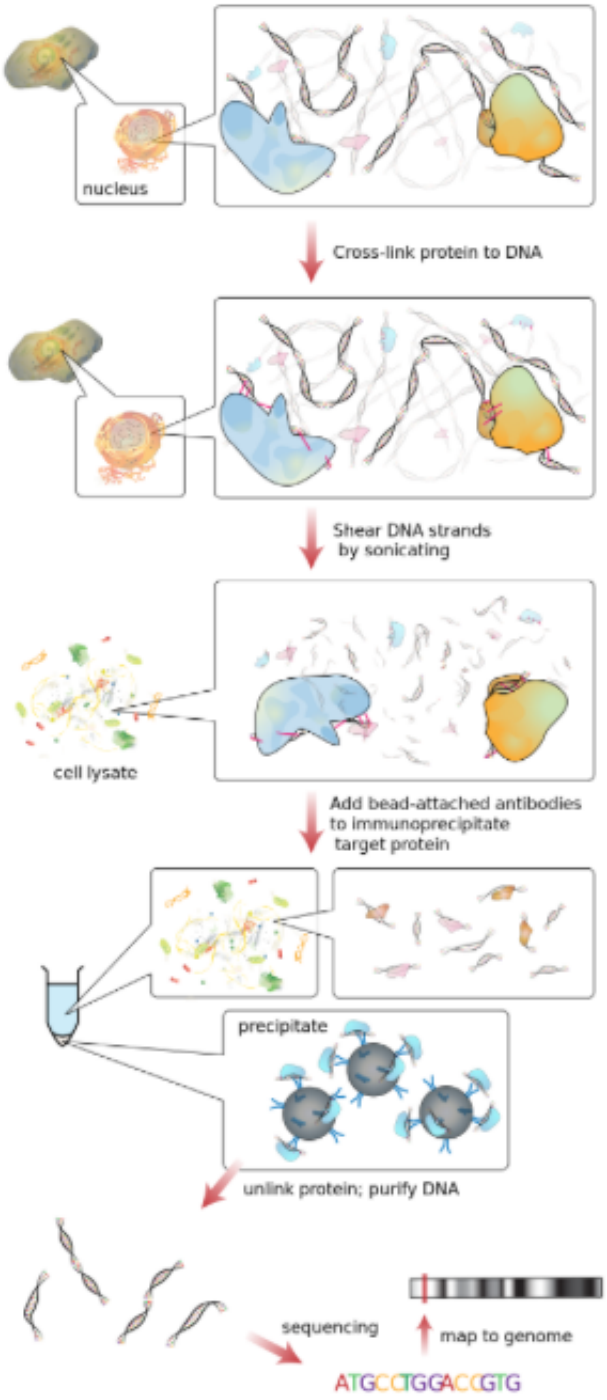


Figura 1.3: Fasi del processo di ChIP-Sequencing.

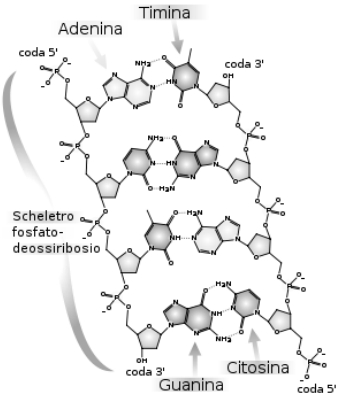


Figura 1.4: Struttura doppia elica DNA, in evidenza le estremità 3' e 5'.

negativi e per ogni base della sequenza del DNA si procede al conteggio dei *reads* allineati che coprono questa base per delineare i due picchi sulla sequenza (picco rosso originato dai frammenti negativi e picco blu dai frammenti positivi). Una volta definita la struttura completa dei picchi sulla sequenza è necessario ricostruire la forma complessiva del picco di origine, non costituito, cioè, dai *reads* positivi e negativi distinti, ma da tutto il frammento. A tale scopo è necessario determinare quale sia la lunghezza dei frammenti iniziali e, noto questo parametro, ciascun *read* viene allungato o traslato a sufficienza e la forma del picco è definita [4] (per il dettaglio della procedura di allineamento e definizione del picco si veda il Capitolo 2).

I dati a disposizione per l'analisi della struttura proteica che circonda il DNA sono, pertanto, i conteggi del numero di frammenti allineati su ciascuna base della sequenza del genoma umano, ovvero una funzione detta *coverage function* che per ogni base della sequenza definisce il conteggio associato.

Opportuni software, detti *peak callers*[15], permettono poi l'analisi dell'insieme dei frammenti allineati associati a una particolare proteina e, mediante il confronto con un opportuno controllo, definiscono quali sono le regioni del genoma considerate "anomale" e definiscono, quindi, i picchi significativi per l'interazione genoma-proteina selezionando le regioni della funzione dei conteggi. Non è sufficiente la semplice analisi della sequenza dei conteggi per definire i picchi della funzione, in quanto il processo di sonicazione e immunoprecipitazione per la selezione dei frammenti associati alla proteina è influenzato

[15]Per la raccolta dei dati analizzati in questo lavoro si utilizza il software MACS [9] esaminato nel dettaglio nella Sezione 2.4.

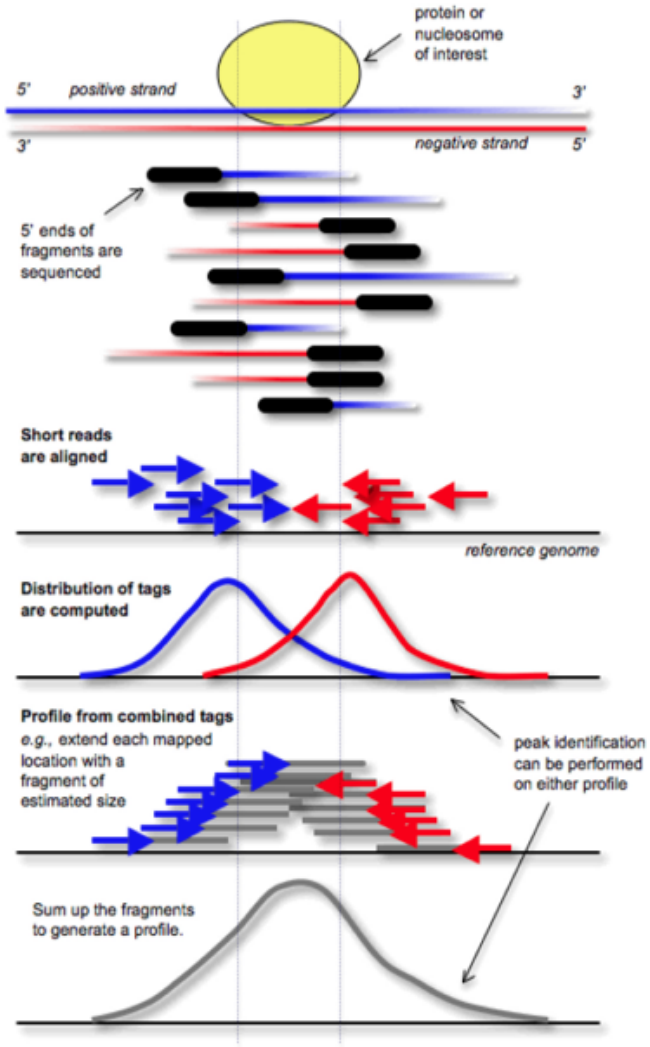


Figura 1.5: Definizione dei picchi a partire dalla selezione dei *reads*.

anche da fattori indipendenti dalla proteina stessa, come ad esempio la struttura della cromatina. La molecola di DNA, infatti, non è uniformemente distribuita nel nucleo della cellula, ma è raggruppata in agglomerati, pertanto alcune regioni risultano più esposte e quindi facilmente selezionabili dal processo di sonicazione, indipendentemente dalla presenza o meno della proteina. Possono, dunque, emergere delle regioni di picco derivanti solo dalla conformazione della molecola, ma non connesse alla presenza della proteina. Per ovviare a questo problema di utilizza dunque una sequenza di controllo generata in modo da rispettare le caratteristiche biologiche della molecola come i ripiegamenti della cromatina, con cui confrontare la sequenza e così definire le regioni che si discostano dall'andamento di base della funzione. Diverse sono le tecniche di generazione delle sequenze di controllo, ad esempio si utilizzano immunoprecipitazioni con anticorpi generici (Mock ImmunoPrecipitation [1]) o selezioni casuali di frammenti senza l'utilizzo di anticorpi.

L'obiettivo di questo lavoro, una volta noti i picchi significativi, è quello di analizzare le caratteristiche di forma che li definiscono; è noto, infatti, che a seconda della tipologia di proteina l'ingombro che questa occupa sul DNA è differente, ad esempio i fattori di trascrizione identificano regioni molto brevi e specifiche che si mostrano in picchi stretti e alti, mentre gli istoni dei nucleosomi definiscono regioni meno dettagliate e più vaste, associate, dunque, a picchi più frastagliati e ampi. La possibilità di caratterizzare la forma di queste regioni anche per proteine non note e l'eventuale riconoscimento per le proteine note di comportamenti anomali può portare alla presentazione di caratteristiche significative della struttura epigenetica delle cellule nel caso di particolari mutazioni fenotipiche dell'organismo, con l'evidente ruolo nell'indagine su patologie per le quali una semplice analisi genetica non può dare risultati soddisfacenti.

1.3 Analisi di forma e presentazione del progetto congiunto IEO-IIT-PoliMi

Come si è già sottolineato, risulta interessante l'analisi dei picchi della *coverage function* non solo in merito alla loro collocazione sul genoma, ma anche riguardo alla loro forma. L'analisi della localizzazione delle regioni, ad esempio associate a specifici istoni dei nucleosomi può essere utile per la definizione dei ripiegamenti della cromatina

e per la conseguente caratterizzazione della struttura tridimensionale della molecola di DNA nelle diverse cellule. In merito alla forma dei picchi, invece, si nota come già prime valutazioni qualitative di dati ChIP-seq hanno portato a osservare molte variazioni nella forma dei picchi per diversi esperimenti, in particolare (come proposto in Figura 1.6, tratta da dati del progetto ENCODE[16]) si sono riscontrare:

- variazioni causate da diverse localizzazioni dei siti di legame tra la proteina o l'istone e la molecola di DNA; analizzando i diversi legami lungo una stessa sequenza, infatti, si possono manifestare differenti profili;
- differenze connesse alla tipologia di legame che si analizza, come si è già osservato, ad esempio, fattori di trascrizione (ORC1) e istoni (H3K4me3 o H3K79me2) manifestano profili molto diversi;
- differenze connesse alla tipologia di anticorpo utilizzato: per riconoscere lo stesso fattore di trascrizione si possono utilizzare diversi anticorpi che possono portare alla presentazione di profili differenti; oltre all'anticorpo specifico per la proteina in esame si possono collegare, infatti, alla proteina differenti catene di amminoacidi, per poi utilizzare l'anticorpo che le riconosce per selezionare la regione;
- variazioni nel confronto tra più esperimenti di ChIP-Seq nella forma dell'espressione del legame tra diversi fattori di trascrizione e la doppia elica; può accadere, infatti che fattori differenti manifestino interazioni simili, questo può derivare dalla possibile presenza di cofattori che, interagendo con le diverse proteine, portano a manifestazioni comuni.

Oltre a questi evidenti motivi di diversità nell'andamento dei picchi, tuttavia, esistono molti elementi che causano variabilità nell'espressione, elementi che non sono identificabili dalla sola analisi dell'esperimento, ma che devono essere considerati attraverso l'indagine specifica dell'andamento della *coverage function*, in particolare:

- ciascun esperimento di ChIP-Seq è condotto su una molteplicità di cellule quanto più omogenee possibile, ma che hanno naturalmente delle diversità, come la

[16]**Progetto ENCODE (*The Encyclopedia of DNA Elements*):** progetto sviluppato da un gruppo di ricerca internazionale e fondato dal *National Human Genome Research Institute* (NHGRI) per definire gli elementi funzionali del genoma umano, inclusi anche gli elementi esterni alla molecola che ne regolano l'espressione [6].

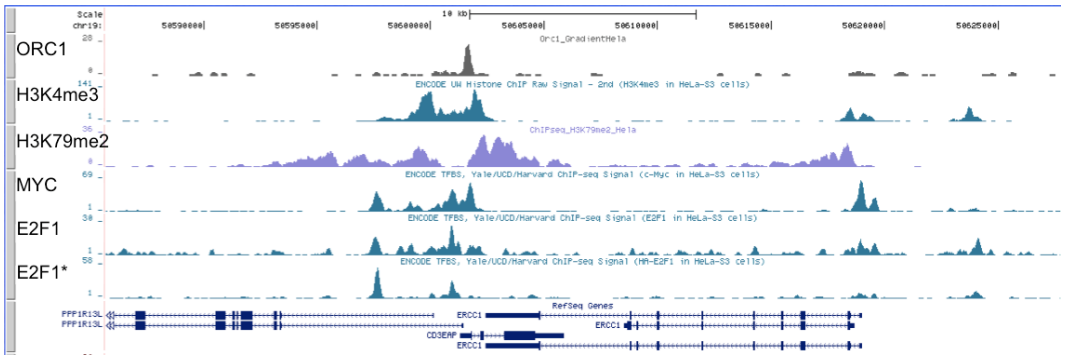


Figura 1.6: Presentazione di diversi profili di picco con forme differenti a seconda della localizzazione o della tipologia di proteina.

fase nel ciclo cellulare o la contaminazione da tessuti vicini. Per selezionare un numero sufficiente di frammenti con l'immunoprecipitazione, infatti, è necessario utilizzare il materiale genetico di più cellule il più simili possibile al fine di non avere risultati non confrontabili; questo insieme di cellule, tuttavia, contiene al suo interno delle diversità fisiologiche che possono portare a *coverage function* non perfettamente definite. La collezione simultanea di questi dati e la definizione di un'unica *coverage function*, infatti, può portare alla presentazione di dati distorti ed è, pertanto, necessaria un'analisi statistica per individuare e separare queste componenti;

- possono esistere dei legami multipli di un'unica proteina a più regioni del DNA, che ovviamente si possono manifestare in diverse strutture della *coverage function*;

- si possono presentare interazioni tra differenti proteine nella regione che circonda il DNA e quindi manifestarsi regioni della sequenza anche non direttamente connesse alla proteina in esame, ma legate a interazioni di una proteina collegata a quella in considerazione.

L'obiettivo dell'indagine, quindi, risulta quello di discernere, grazie a opportune analisi statistiche, i contributi di questi ultimi tre fattori, essendo note le caratteristiche generali dell'esperimento e quindi limitando la variabilità di forma connessa agli elementi descritti precedentemente.

Il progetto IEO-IIT-PoliMi *Genomic Computing*, nella sezione di analisi della forma dei picchi ha proprio l'obiettivo di discernere queste componenti di variabilità. In particolare gli obiettivi prefissati sono i seguenti:

1. associare la forma dei picchi al loro significato biologico, ad esempio focalizzandosi sulle diversità delle sequenze di nucleotidi sottese alle regioni attraverso un'analisi dei motivi per valutarne le eventuali differenze anche a livello di alterazioni epigenomiche;

2. determinare picchi significativamente diversi in due differenti esperimenti, ad esempio associati a colture di cellule patologiche o sane;

3. sviluppare dei metodi per decomporre la *coverage function* distinguendo i profili associati a raggruppamenti omogenei di cellule all'interno della moltitudine che viene analizzata nel singolo esperimento così da permettere anche un corretto confronto tra i dati; risulta fondamentale, infatti, scindere i diversi contributi delle differenti sottopopolazioni di cellule per poter poi effettuare comparazioni tra gli esperimenti;

4. decomporre picchi complessi che possono denotare legami multipli proteina-DNA;

5. identificare legami tra le interazioni delle proteine e la struttura della cromatina.

Una prima fase del progetto prevede la valutazione di dati relativi a un fattore di trascrizione, la proteina GATA1 e la determinazione di caratteristiche significative dei picchi selezionati con un opportuno *peak caller* al fine di proporre possibili risposte per gli obiettivi richiesti.

1.4 Proteina GATA1

I dati raccolti per questa analisi si riferiscono all'interazione della sequenza di DNA con la proteina GATA1, un fattore di trascrizione che ha un ruolo fondamentale nella regolazione di geni, molti dei quali connessi alla composizione del sangue. GATA1, infatti, regola la crescita e la divisione dei globuli rossi e la differenziazione delle cellule del midollo per la costituzione dei megacariociti, ovvero le cellule responsabili della

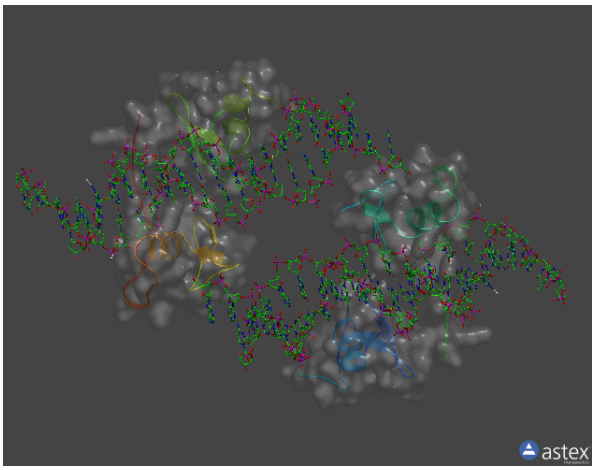


Figura 1.7: Proteina GATA1 connessa alla doppia elica di DNA.

produzione delle piastrine. Questa proteina è fondamentale anche per la regolazione della maturazione di numerosi tipi di globuli bianchi.

Questo fattore di trascrizione, rappresentato nella sua struttura terziaria in Figura 1.7, è costituito tra tre domini proteici ed è connesso al DNA con il caratteristico motivo composto da Guanina-Adenina-Timina-Adenina.

Capitolo 2

Raccolta dati

Attraverso la tecnica del ChIP-Seq, come descritto in dettaglio nel Capitolo 1, si può ottenere, mediante immunoprecipitazione, un insieme di frammenti relativi a regioni del genoma che interagiscono con una specifica proteina in esame. Questi frammenti vengono sequenziati al fine di discernere la sequenza che li caratterizza e i *reads* che si ottengono vengono allineati sul genoma; a questo punto è possibile definire la forma della *coverage function* ed analizzare i risultati con opportuni software al fine, ad esempio, di definire le regioni di picco interessanti dal punto di vista biologico.

In questo capitolo ci si occupa dell'analisi di questo percorso di definizione dei dati a partire dalla lettura dei *reads* con opportune verifiche di qualità (Sezioni 2.1, 2.2), per poi concludere con la descrizione della tecnica di allineamento utilizzata (Sezione 2.3) e con la definizione dei picchi (Sezione 2.4).

2.1 Lettura dei frammenti

Una moderna tecnologia (ideata nel 2006) molto utilizzata per la lettura dei *reads* di DNA è il sequenziamento mediante sintesi di filamenti uguali a quelli da analizzare. Questa tecnologia, sviluppata dalla compagnia americana *Illumina* e pertanto definita *Illumina Genome Analyzer*, è basata sull'amplificazione delle sequenze per generare raggruppamenti omogenei e quindi procedere alla lettura ordinata delle basi grazie alla connessione con opportuni coloranti.
Si distinguono le seguenti fasi:

1. i campioni di DNA sono posizionati su una cella a flusso in modo da permettere l'accesso agli enzimi utili per le successive analisi. Questi campioni sono connessi in specifiche regioni della cella grazie al collegamento di opportuni adattatori aggiunti alle estremità dei frammenti con i complementari ancorati alla cella di flusso;

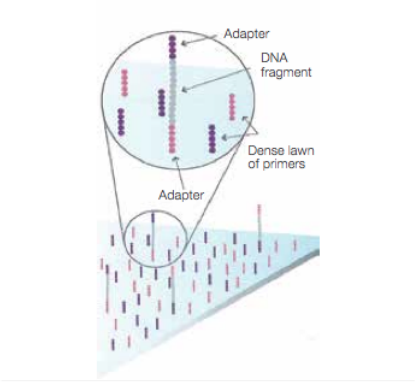


Figura 2.1: Rappresentazione della cella a flusso con i frammenti di DNA ancorati.

2. i frammenti di DNA sono amplificati nella cella a flusso. La *bridge amplification* permette di costituire attraverso la tecnica della PCR (*Polymerase Chain Reaction*) raggruppamenti di circa 1000 frammenti uguali;

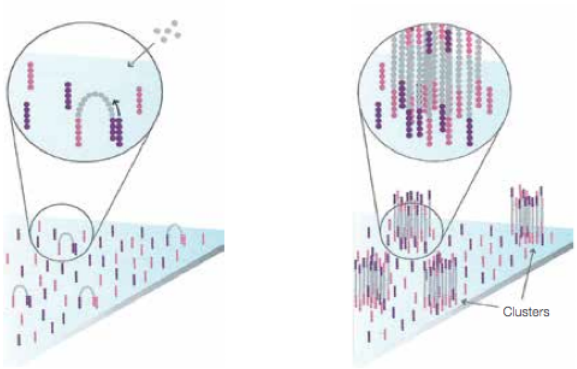


Figura 2.2: *Bridge amplification* e definizione dei raggruppamenti.

3. i raggruppamenti, formati da singoli filamenti dell'elica, per essere sequenziati subiscono un'ibridazione di un frammento complementare alla sequenza dell'adattatore e successivamente sulla cella a flusso viene aggiunta una miscela di DNA polimerasi e dei quattro nucleotidi marcati con fluorofori differenti. Questi si connettono per complementarietà alla base libera della sequenza di ciascun filamento, all'interno di ciascun raggruppamento clonale. Dopo questa incorporazione i reagenti in eccesso vengono eliminati e si rileva la fluorescenza relativa a ogni cluster. Si elimina, quindi, la marcatura fluorescente e si procede con il sequenziamento successivo, fino alla completa registrazione di tutte le fluorescenze;

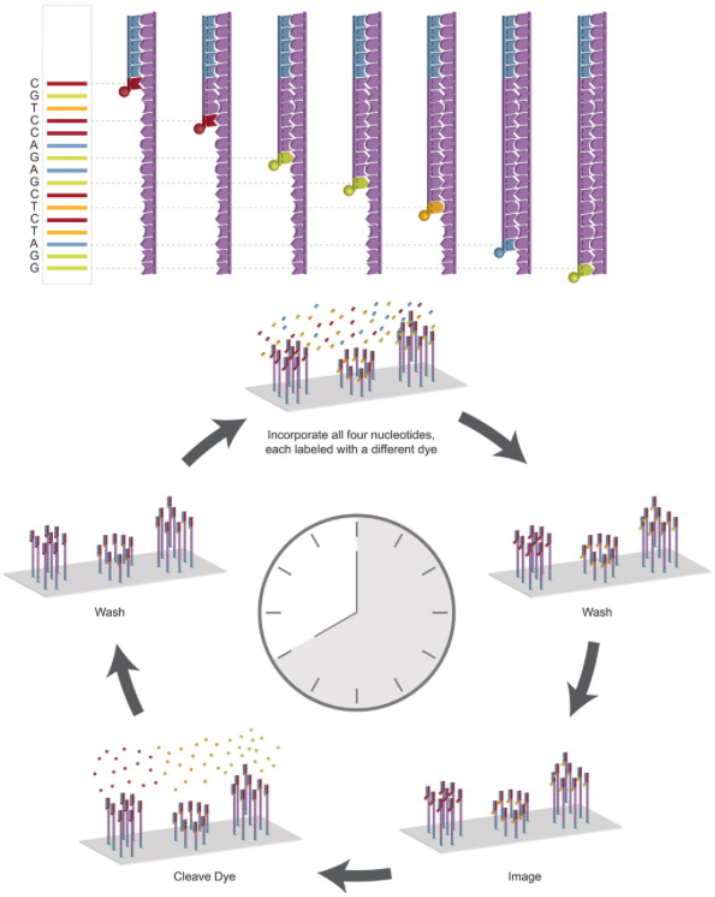


Figura 2.3: Sequenziamento per sintesi.

4. le immagini associate a ciascun ciclo di ibridazione sono note; si procede, pertanto, alla lettura delle basi associate a ogni frammento. Questo procedimento avviene

mediante opportune analisi di immagini ed è possibile a seguito dell'amplificazione e conseguente definizione dei cluster, così che ogni fluorescenza sia ben dettagliata e riconoscibile;

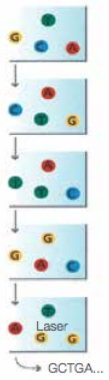


Figura 2.4: Immagini per la lettura delle diverse basi.

Attraverso questo procedimento vengono quindi riconosciute tutte le sequenze che sono state selezionale dall'immunoprecipitazione. A ciascun nucleotide letto in una di queste sequenze, inoltre, è possibile associare un codice di qualità della lettura connesso alla probabilità p di errore nella definizione della base. Nei file in cui è raccolto l'elenco delle sequenze (formato .fastq) è indicato per ogni base anche il valore Q, attraverso una codifica dal formato ASCII (*American Standard Code for Information Interchange*), che è connesso alla probabilità p attraverso la formula

$$Q = -10 \log_{10} p.$$

Il codice ASCII è una rappresentazione numerica dei caratteri: i primi 32 simboli di questa codifica sono di controllo, mentre dal trentatreesimo ("!") inizia la corrispondenza biunivoca tra carattere e numero intero. A seconda del formato utilizzato dal sequenziatore (Sanger, Solexa o Ilumina) si può valutare l'indice di qualità Q di ogni nucleotide della sequenza attraverso la decodifica del codice ASCII associato a ciascuna base (Figura 2.5), e quindi calcolare p.

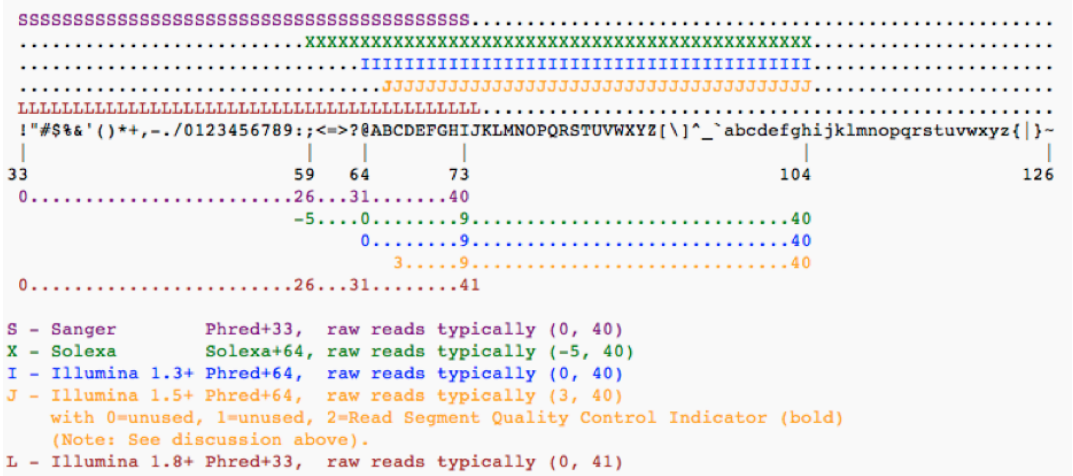


Figura 2.5: Definizione dei diversi formati per la decodifica del carattere ASCII come indicatore di qualità della lettura.

2.2 Controllo qualità dei frammenti

Prima di analizzare le sequenze lette, utilizzarle per definire la *coverage function* e trarne le adeguate considerazioni biologiche è necessario compiere semplici controlli di qualità che confermino che i frammenti siano effettivamente letti correttamente. A partire dalla valutazione degli indici p associati a ciascuna base dei frammenti è possibile stilare un rapporto che valuti le possibili manifestazioni di errore nella lettura; numerosi sono i software utilizzati per questo tipo di analisi, e uno dei più completi è FastQC [14] che permette una visualizzazione delle possibili criticità con dettaglio sufficiente da spiegare anche le possibili cause della scarsa affidabilità dei dati.
Questo programma valuta differenti aspetti per il controllo della qualità complessiva dei frammenti, dalla qualità della lettura delle singole basi, calcolata con opportuni algoritmi, alla verosimiglianza dell'assegnamento di ciascuna base della sequenza al nucleotide scelto, alle duplicazioni delle sequenze e a considerazioni sulle loro lunghezze. Queste informazioni sono descritte dettagliatamente in Appendice A e per ognuna di esse sono indicate le eventuali criticità. Per ogni voce, poi, il programma, secondo opportuni standard, assegna automaticamente un simbolo che indichi il livello di qualità associato:

 ottima qualità del dataset rispetto al parametro in esame;

 qualità scarsa che necessita di verifiche;

 qualità del tutto insufficiente: è necessario prestare attenzione soprattutto nel caso in cui ci sia più di un parametro con tale indicazione.

Per ogni insieme di frammenti, pertanto, si propone una valutazione complessiva della qualità con le indicazioni della bontà di lettura secondo i differenti parametri. I file che sono utilizzati per l'analisi sono i file .fastq generati dal processo di lettura in cui sono elencati tutti i *reads*. Le analisi vengono compiute o sulla qualità complessiva delle diverse sequenze o sulla valutazione della qualità media tra tutte le sequenze nella lettura delle diverse basi; generalmente a causa del processo di lettura, infatti, la qualità varia a seconda di quante basi del frammento sono già state lette (in particolare, le prime basi lette vengono identificate univocamente mentre più procede il processo più la qualità si abbassa).
In Figura 2.6 si propone a titolo di esempio la valutazione complessiva del dataset relativo alla proteina GATA1 utilizzato per le successive analisi.

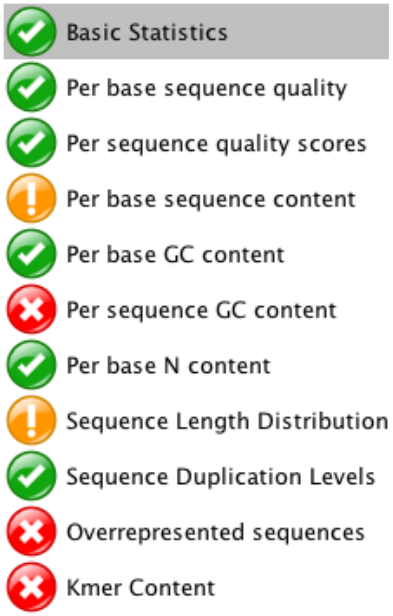


Figura 2.6: Valutazione della qualità complessiva del dataset.

Si presenta, ora l'analisi dettagliata della qualità del dataset introdotto in precedenza relativo all'esperimento di ChIP-Seq per la proteina GATA 1, come da schema riassuntivo proposto in Figura 2.6.
L'output di FastQC per le singole grandezze definite in tabella è il seguente:

Statistiche introduttive

Basic sequence stats	
Measure	Value
Filename	wgEncodeYaleChIPseqRawDataRep1K562Gata1.fastq
File type	Conventional base calls
Encoding	Illumina <1.3
Total Sequences	7855375
Filtered Sequences	0
Sequence length	28-35
%GC	44

Figura 2.7: FastQC: statistiche introduttive.

Qualità per ogni base della sequenza

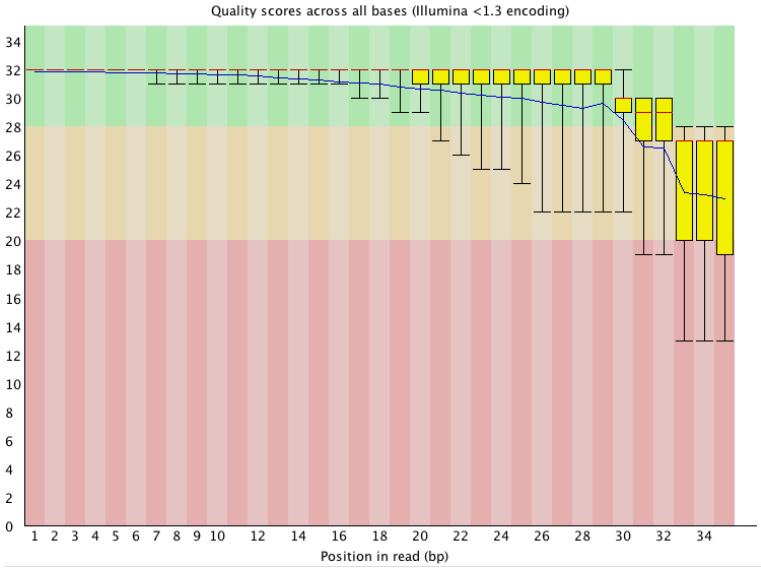


Figura 2.8: FastQC: qualità per ogni base della sequenza.

Qualità per sequenze

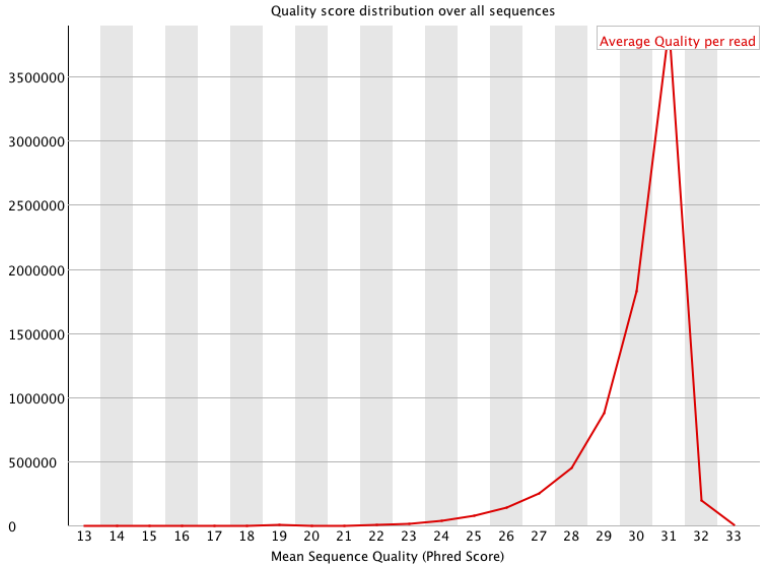


Figura 2.9: FastQC: qualità per sequenze.

Contenuto di nucleotidi nelle basi

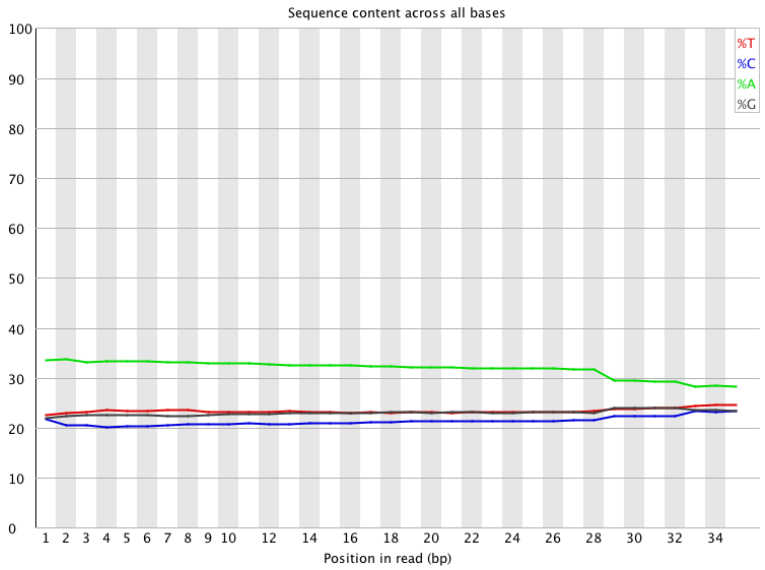


Figura 2.10: FastQC: contenuto di nucleotidi nelle basi.

Contenuto di C e G nelle basi

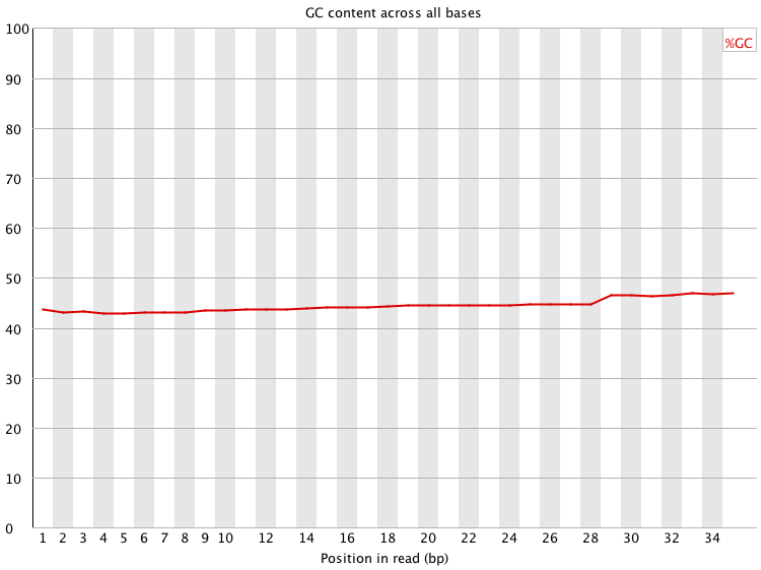


Figura 2.11: FastQC: contenuto di C e G nelle basi.

Contenuto di C e G per ogni sequenza

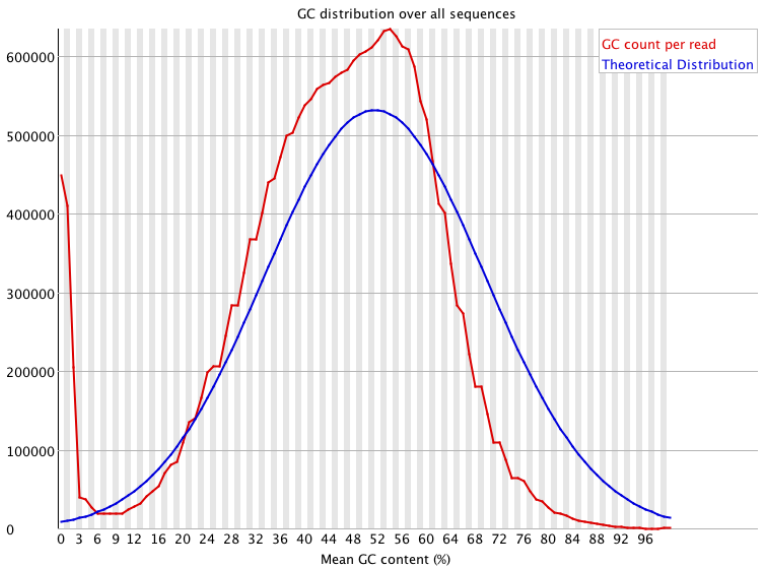


Figura 2.12: FastQC: contenuto di C e G per ogni sequenza.

Contenuto di N nelle basi

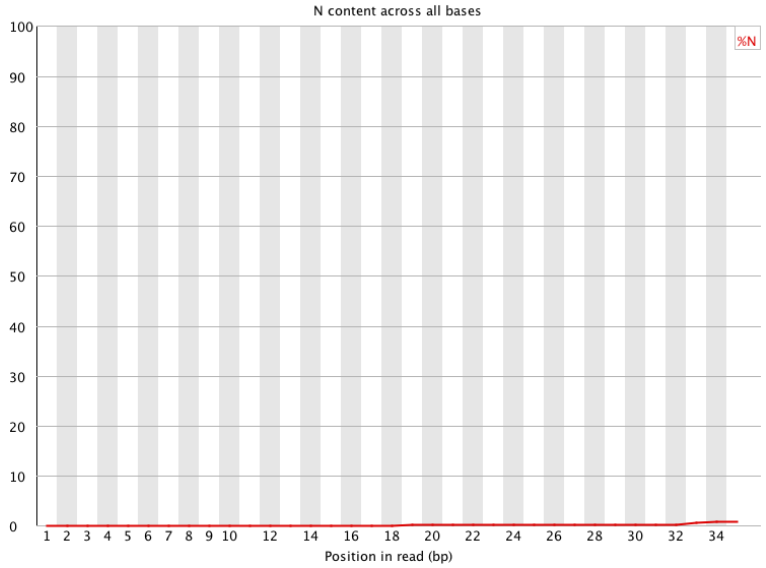


Figura 2.13: FastQC: contenuto di N (nucleotidi non letti) nelle basi.

Distribuzione delle lunghezze delle sequenze

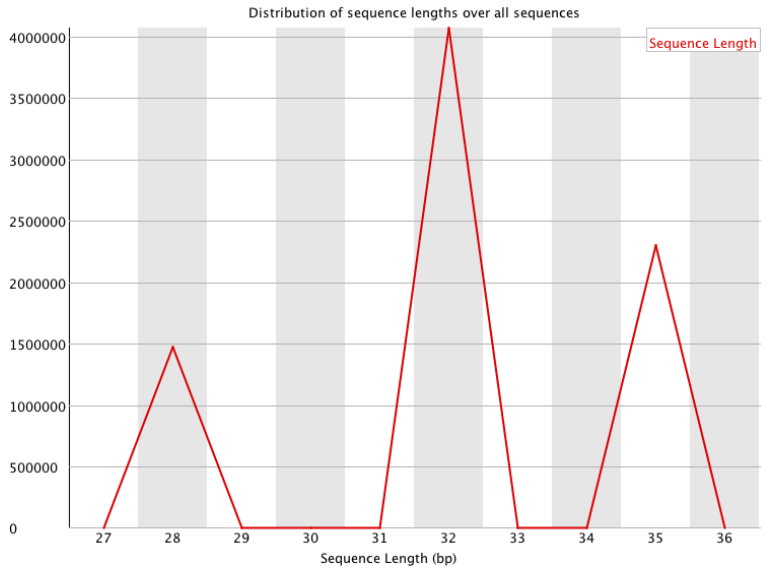


Figura 2.14: FastQC: istribuzione delle lunghezze delle sequenze.

Sequenze duplicate

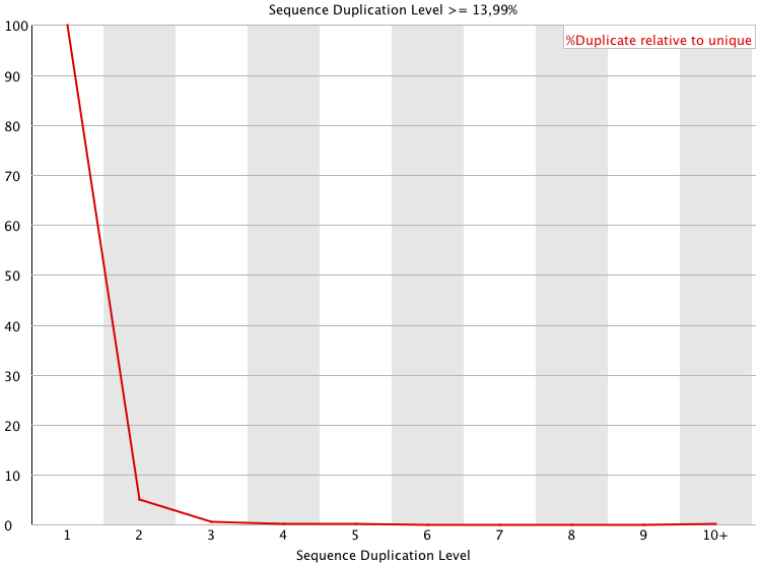
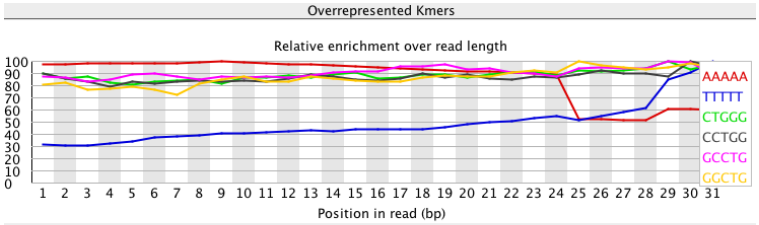


Figura 2.15: FastQC: sequenze duplicate.

k-meri sovraespressi



Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Position
AAAAA	14210280	18,51	20,54	9
TTTTT	884350	5,72	12,552	31
CTGGG	575160	4,263	4,834	29
CCTGG	485280	3,898	4,496	30
GCCTG	480200	3,857	4,254	29
GGCTG	519115	3,848	4,468	25
GCTGG	518820	3,846	4,489	28
CCCAG	603625	3,812	4,31	27
CCTCC	392600	3,702	5,014	30
CCCCC	355280	3,681	9,925	31
GGTGG	508940	3,482	3,788	9
CTCCC	354620	3,344	5,009	30
CCAGG	558620	3,256	3,624	25
CCAGC	508235	3,21	3,757	29
CCTGT	395475	3,132	3,625	19
TGGTG	460720	3,108	3,517	8
GCCTC	349855	3,044	3,515	29
CTGCC	347835	3,027	3,673	31
GGGGG	318195	2,208	6,315	31

Figura 2.16: FastQC: *k*-meri sovraespressi.

Sequenze sovraespresse

Overrepresented sequences			
Sequence	Count	Percentage	Possible Source
AAAAAAAAAAAAAAAAAAAAA...	81793	1,041	No Hit

Figura 2.17: FastQC: sequenze sovraespresse.

Questi risultati portano a concludere che il dataset inerente alla proteina GATA1 ha una buona qualità complessiva; pur presentando alcune criticità, infatti, nel complesso risulta affidabile.

Si osserva che la presenza di una sequenza in più dell'1% del dataset porta alla manifestazione dei problemi inerenti anche agli altri fattori di qualità. Questa sequenza, infatti, è costituita da tutte Adenine, pertanto anche la quantità di basi A sarà nel complesso alterata e conseguentemente anche il contenuto di CG, si discosterà di molto dall'andamento atteso dal momento che esistono un numero significativo di sequenze in cui non sono presenti queste basi. La motivazione della presenza di questa sequenza sovraespressa, tuttavia, non è motivata, infatti non è presente una descrizione della possibile causa di sovraespressione, tuttavia risulta necessaria la rimozione di questa sequenza dal dataset. Per quanto concerne la presenza di *k*-meri sovraespressi, inoltre, si osserva che quelli più evidenti, ovvero con un rapporto complessivo tra valore reale e atteso superiore a 10 sono i *k*-meri associati alle sequenze AAAAA e TTTTT, ovvero ancora connessi al problema precedentemente introdotto.
Si osserva, infine, che il dataset contiene sequenze di lunghezza differente, la maggior parte si attesta su lunghezze pari a 32 basi, ma circa il 30% del dataset è costituito da sequenze più lunghe (35 basi) e il 20% da sequenze più corte (28 basi).

2.3 Allineamento al genoma

Dopo aver raccolto, sequenziato i frammenti di DNA e verificato la qualità della lettura, si procede all'allineamento dei *reads* al genoma di riferimento, per poi riconoscere le regioni di arricchimento.

L'esigenza di un allineamento locale efficace ha portato allo sviluppo di numerose tecniche di allineamento migliori della semplice ricerca di stringhe con la valutazione

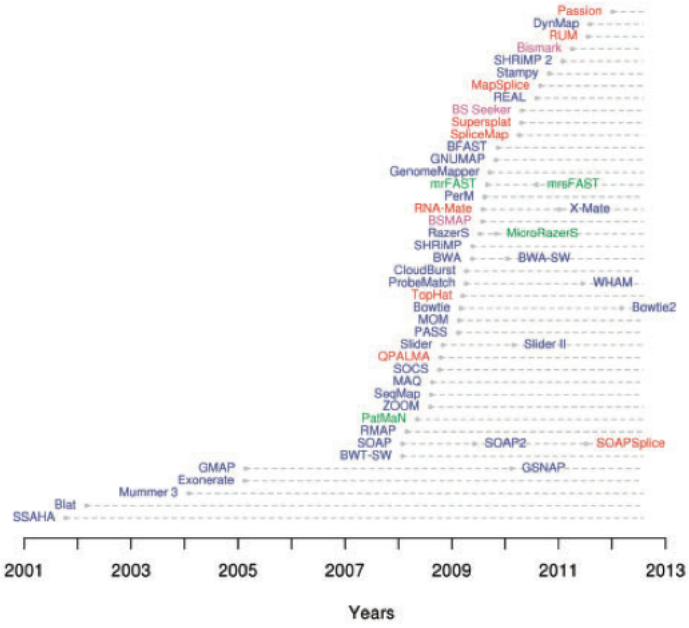


Figura 2.18: Schema riassuntivo degli algoritmi di allineamento locale.

in sequenza di tutto il genoma. Il numero di operazioni per compiere questa ricerca (che eventualmente può essere estesa, con la definizione della distana di Levenshtein, alla ricerca con *mismatch* o *gap*), infatti, è molto elevato e dell'ordine di $O(L_g L_r N_r)$ con L_g lunghezza del genoma, L_r lunghezza dei trammenti e N_r numero di frammenti da analizzare; valore che risulta proibitivo nel caso, ad esempio, del genoma umano che conta 3.3B di basi.

A partire dagli anni 2000 sono stati sviluppati numerosi algoritmi di allineamento locale (come presentato in Figura 2.18) che magari non permettono di trovare tutte le possibili regioni di allineamento, come permette, invece il costoso algoritmo di ricerca esatta di Smith-Waterman, ma certamente hanno un'efficienza molto superiore. Alcuni di questi sono basati ancora sulla scansione di tutto il genoma (come MAQ), altri, invece, sulla definizione della trasformata di Burrows-Wheeler (come SOAP2, Bowtie o BWA). La trasformata di Burrows-Wheeler (definita nel dettaglio in Appendice B.1) oltre ad essere molto adatta alla compressione di stringhe e pertanto molto utilizzata in programmi come bzip2, è molto efficace per l'allineamento di brevi sequenze al genoma soprattutto perché consente di memorizzare sinteticamente molte caratteristiche della stringa da

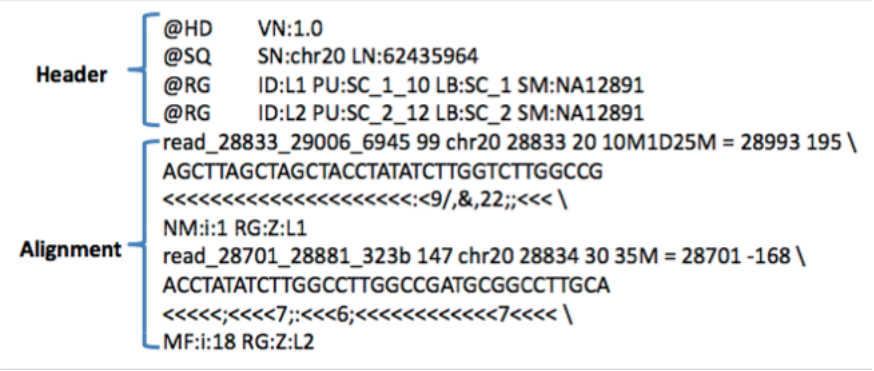


Figura 2.19: Esempio di file .sam.

analizzare, che altrimenti occuperebbero un'eccessiva quantità di memoria.

In questa sezione si considera la tecnica del *Burrows-Wheeler Alignemnt* (BWA), sviluppata da Heng Li e Richard Dublin [10] che permette di allineare brevi frammenti alla sequenza del genoma umano consentendo *mismatches* e introduzione di spazi. Questo algoritmo, analizzato nel dettaglio in Appendice B.1, permette di allineare efficacemente i frammenti al genoma; dalla simulazione condotta dagli stessi H. Li e R. Durbin e riportata nel dettaglio in [10] sull'allineamento di frammenti lunghi 30 basi si ottiene, per l'allineamento *single-end* un'accuratezza dell'80.6%, risultato migliore degli altri algoritmi basati sulla trasformata di Burrows-Wheeler e paragonabile a quello ottenuto con algoritmi basati sull'esplorazione dell'intera sequenza. L'ottimo risultato si ottiene, inoltre, in un tempo molto inferiore rispetto agli algoritmi come MAQ (sempre sulla stessa simulazione si ottiene il risultato in meno di 1/20 del tempo impiegato da MAQ); questo algoritmo, inoltre, risulta facilmente parallelizzabile e quindi potenzialmente ancora più efficace. La memoria utilizzata, inoltre è indipendente dal numero di frammenti da allineare, a differenza degli algoritmi basati sull'esplorazione della sequenza che, invece, sono lineari in questo parametro.

I programmi di allineamento di *reads* sul genoma generano in output un file di tipo .sam (*sequence alignment map*), o .bam in versione binaria, che contiene tutte le informazioni sul procedimento di allineamento e sul risultato.
In Figura 2.19 si presenta un esempio di file .sam in cui si evidenzia la struttura che lo compone. Esiste, infatti, una prima parte di intestazione (*header*) e caratterizzazione complessiva dell'allineamento e una seconda in cui sono presentati nel dettaglio i risultati della procedura(*alignment*).

La sezione di intestazione comprende generalmente:

- la definizione dell'indice di riferimento dell'allineamento (`@HD`);
- alcune informazioni sulla sequenza di riferimento, come la localizzazione sul genoma e la lunghezza (`@SQ`);
- alcune caratterizzazioni dei frammenti allineati, come l'identificativo del centro che li ha prodotti e la piattaforma con cui sono stati generati (`@RG`);
- le caratteristiche del tool che è stato utilizzato per l'allineamento, come il nome del software e la versione (`@PG`).

La sezione di allineamento, presentata nel dettaglio in [12], contiene undici campi obbligatori che contengono informazioni generali sull'allineamento, ma anche sulle caratteristiche dettagliate dei frammenti e ella sequenza. L'ultimo campo, infine, è un codice di qualità identificato ancora con la codifica ASCII, che valuta nel complesso l'esito del processo di allineamento.

Data la complessità del formato esistono numerosi tool (*samtools*) che permettono l'estrazione di informazioni rilevanti dai file .sam o .bam o la loro conversione in formati più leggibili, come il formato .bed, che contiene semplicemente l'elenco dei frammenti allineati e la loro localizzazione sul genoma di riferimento.

Si sottolinea che esistono molte raccolte di file di tipo .sam che registrano gli esperimenti condotti con tecniche NGS, come il ChIP-Seq. Per questo progetto di Tesi si utilizzano i dati raccolti da ENCODE contenenti frammenti già letti e allineati, su cui è stata condotta la sola analisi di qualità mediante FastQC. Questi dati sono poi analizzati con gli strumenti proposti nelle prossime sezioni per trarre gli opportuni riscontri biologici.

2.4 *Peak-caller*: MACS

In questa ultima sezione si analizzano le operazioni compiute dai software detti *peak-callers* per definire, una volta note le regioni di allineamento per i *reads*, le zone del genoma realmente associate alla proteina in esame. Si ricorda che il dato in ingresso per la definizione dei picchi è la sola struttura di allineamento di una ridotta porzione dei

frammenti selezionati; sono stati allineati, infatti, segmenti di solo circa trenta basi corrispondenti alle parti iniziale o finale dei frammenti isolati con l'immunoprecipitazione. La ricerca delle regioni significative si basa, come principio generale, sui seguenti passi:

1. definizione di un profilo di segnale (*coverage function*) lungo ogni cromosoma;
2. definizione di un modello di *background*;
3. identificazione delle regioni di picco attraverso il confronto tra funzione e *background*;
4. eventuale filtraggio dei picchi erroneamente definiti;
5. valutazione e ordinamento dei picchi in base al livello di significatività della loro definizione.

Osservazione: Risulta fondamentale il punto 2. e quindi la definizione delle regioni di significatività non solo valutando il segnale, ma confrontandolo opportunamente con il *background* in quanto la struttura della cromatina non è omogenea su tutta la lunghezza del DNA, pertanto alcune regioni possono essere selezionate dall'immunoprecipitazione anche se non significative per la proteina in esame. Esistono differenti metodologie per definire la struttura di controllo per il confronto, dall'immunoprecipitazione con anticorpo generico (anticorpo che non riconosce le proteine che interagiscono con il DNA), alla semplice sonicazione della cromatina senza alcun anticorpo. In questa analisi i risultati proposti sono stati ricavati con un controllo generato dall'immunoprecipitazione di frammenti selezionati da un anticorpo generico (IgG Mock-IP).

Diversi *peak-callers* sviluppano in maniera differente i passi di definizione delle regioni di arricchimento: dalla caratterizzazione della *coverage function* a partire dai frammenti allineati, al processo di confronto tra le funzioni, al criterio di ordinamento degli esiti ottenuti. In Figura 2.20 si propone uno schema riassuntivo delle differenziazioni dei principali software utilizzati [11].

Focalizzandosi sul software MACS (*Model-based Analysis of ChIP-Seq*) [9], [13], in Figura 2.21 si presentano gli aspetti fondamentali della ricerca delle regioni di arricchimento.

Entrando nel dettaglio dell'analisi si osserva che il primo passo da compiere sia per la sequenza di controllo che per la sequenza di partenza è la rimozione dei *reads* in eccesso.

	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based/ duplicate[e]	Refs.
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
ERANGE v3.1	Tag aggregation	1: Height cutoff Hiqh quality peak estimate, per-region estimate, or input	Hiqh quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally *P* values	*P* value	1: None 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Optional peak height, ratio to background	Yes / No	4,18
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes	19
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No	14
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Target FDR, number nearest neighbors for clustering	No / No	17
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No	5

Figura 2.20: Caratteristiche fondamentali di alcuni *peak-callers*.

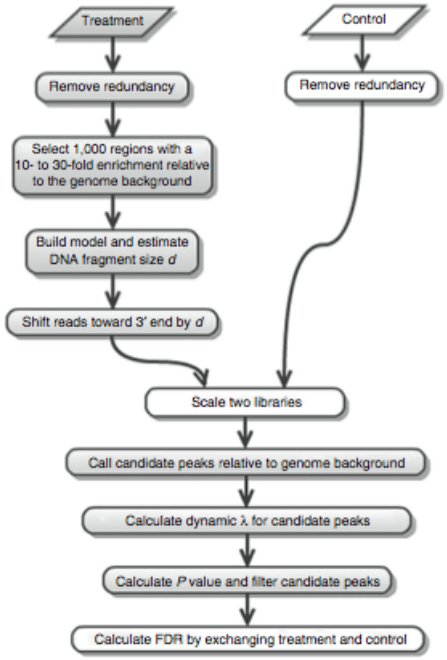


Figura 2.21: Procedimento per la definizione delle regioni di arricchimento con MACS.

Il procedimento di lettura dei frammenti infatti, come si è visto nel dettaglio in Sezione 2.1, prevede la necessità di duplicare artificialmente (mediante PCR) le sequenze, in modo da permetterne la agevole lettura; questo procedimento, necessario per costituire i raggruppamenti sulla piastra di lettura può, però, portare all'errata costituzione di ulteriori cluster con le medesime caratteristiche e dunque la duplicazione dell'informazione iniziale. É necessario, pertanto, rimuovere dalla sequenza con i *reads* allineati tutti i frammenti duplicati che sono, quindi, localizzati esattamente sulle stesse basi.

Come si è già osservato nell'introduzione iniziale sulla tecnica del ChIP-Seq, i *reads* allineati sul genoma sono la prima o l'ultima porzione dei frammenti selezionati dall'immunoprecipitazione. Si ricorda, infatti, che è possibile procedere alla lettura solo in una specifica direzione del frammento e si costituiscono, pertanto, due conteggi sul genoma, il primo per gli inizi dei frammenti (conteggio positivo) e il secondo con le estremità conclusive (conteggio negativo), come rappresentato in Figura 2.22.

Per ricostruire l'informazione completa è necessario selezionare la regione effettivamente associata al frammento originale. Per compiere questa analisi, MACS seleziona

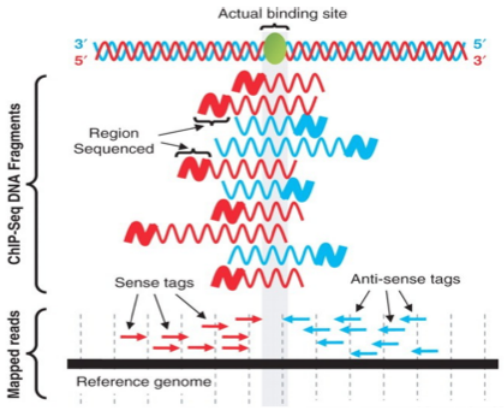


Figura 2.22: Identificazione dei frammenti letti per la definizione dei picchi.

un numero sufficiente di regioni (circa mille) associate a un significativo arricchimento rispetto alla funzione di *background* sia per i picchi positivi (detti anche segmenti di Watson) sia per quelli negativi (o segmenti di Crick); ricerca, poi, la distanza tra i punti di massimo delle differenti regioni positive e negative e, valutando il valore più frequente, definisce il parametro di distanza d (come rappresentato in Figura 2.23). I brevi frammenti iniziali sono, quindi, traslati di $d/2$ alla loro estremità 3' per definire così la più verosimile regione di picco, come schematizzato in Figura 2.24.

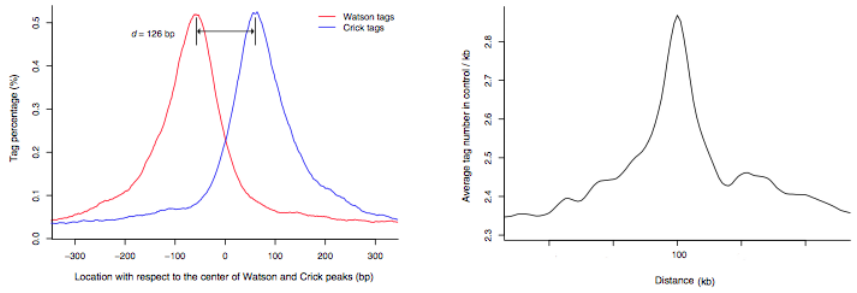


Figura 2.23: Rappresentazione di due picchi di Watson e Crick associati e della distribuzione della distanza di traslazione ottima.

Il procedimento descritto, tuttavia, può apparire poco accurato in quanto si basa sulla selezione di una ridotta quantità di regioni di interesse e sulla valutazione della distanza tra due punti di riferimento; la particolare regolarità del procedimento di selezione dei

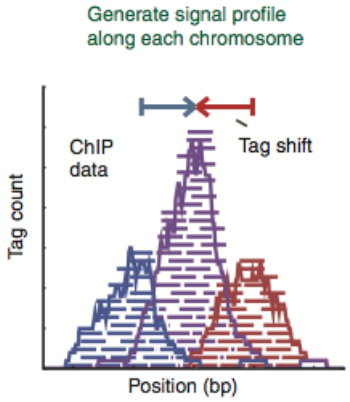


Figura 2.24: Presentazione del procedimento di traslazione dei frammenti per definire la forma del picco.

frammenti, tuttavia fa sì che questo metodo risulti ugualmente efficace. Procedere ad un'analisi globale sulle funzioni dei picchi positivi e negativi non conduce a risultati significativamente differenti. Si riporta, ad esempio, il caso della *coverage function* derivante dall'analisi della proteina GATA1. Il parametro d calcolato da MACS con la tecnica precedentemente introdotta vale 112 bp, mentre, compiendo un'analisi globale, il risultato ottimo risulta 108 bp. La tecnica di analisi complessiva, richiede la valutazione della distanza $\|\cdot\|_{\mathcal{L}^2}$ tra la funzione complessiva dei picchi positivi $f_p(x)$ e la funzione dei picchi negativi $f_n(x)$ opportunamente traslata

$$d(\tau) = \|f_p(x) - f_n(x-\tau)\|_{\mathcal{L}^2}$$

Si ricerca, dunque, il τ per cui questa funzione risulta minima e il valore ottimo è $d^* = \min d(\tau)$.
Il risultato proposto da MACS, dunque, risulta sufficientemente accurato e pertanto verrà utilizzato nelle pessime analisi.
Al temine di questo procedimento di traslazione è necessario procedere all'effettivo confronto tra il segnale di interesse e il controllo al fine di selezionare le regioni che risultano significativamente arricchite.
La prima operazione da compiere è quella di rendere le due funzioni effettivamente confrontabili, ovvero è necessario scalare linearmente la funzione di controllo con un

opportuno k pari a:

$$k = \frac{\text{conteggio totale della funzione}}{\text{conteggio totale del controllo}}$$

per avere un numero di frammenti uguagli a quella della funzione di interesse.

É, dunque, possibile definire un modello per la selezione delle regioni di arricchimento.
Per quanto concerne la funzione di controllo $f_{BG}(x)$, si assume che il suo andamento possa essere modellizzato con una distribuzione di Possion nel parametro λ_{BG} che racchiude tutta l'informazione di media e varianza. Si precisa che esistono, però, delle variazioni locali nell'andamento della f_{BG} causate delle note differenze biologiche nella struttura tridimensionale della cromatina e pertanto anche la funzione di *background* può avere delle alterazioni nelle diverse regioni della *coverage function*.
Il parametro λ_{BG} deve, quindi, essere opportunamente corretto al variare della regione in analisi, definendo λ_{loc}:

$$\lambda_{\text{loc}} = \max(\lambda_{BG}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k})$$

dove $\lambda_{1k}, \lambda_{5k}, \lambda_{10k}$ sono i parametri caratteristici della distribuzione di Poisson per regioni della funzione di *background* ampie rispettivamente 1kbase, 5kbase, 10kbase e centrate attorno al punto di interesse.

Il parametro λ_{loc} così definito viene, quindi, utilizzato per il confronto con il valore medio del conteggio complessivo delle regioni della *coverage function*. Definendo, infatti, un'opportuna ampiezza per una finestra mobile sulla funzione si valuta λ_{loc} per il *background* e λ_c per la funzione di interesse. Si definisce, poi, una soglia che come default per MACS è pari a 10^{-8} per il p-value del test:

$$\begin{cases} H_0 : \lambda_c \leq \lambda_{\text{loc}} \\ H_1 : \lambda_c > \lambda_{\text{loc}} \end{cases}$$

Se per la regione di analisi si trova un p-value inferiore alla soglia fissata si può affermare che esiste sufficiente evidenza per assumere l'arricchimento della regione in confronto al controllo e quindi si deduce la presenza della proteina in esame.

É, infine, possibile valutare la bontà del metodo di ricerca delle regioni di arricchimento calcolando un *False Discovery Rate* empirico per l'arricchimento; si valutano, infatti, oltre ai picchi di arricchimento per la *coverage function* del ChIP-Seq rispetto

al controllo, anche i picchi del controllo rispetto alla funzione in esame. Calcolando il rapporto tra la numerosità dei due insiemi di picchi

$$\text{FDR} = \frac{\text{n}^{\cdot} \text{ picchi del controllo}}{\text{n}^{\cdot} \text{ picchi ChIP-Seq}};$$

si osserva che più risulta ridotto questo rapporto più il metodo è efficace per la definizione delle regioni di arricchimento.

Capitolo 3

Presentazione dei dati

In questa sezione ci si pone l'obiettivo di analizzare i dati sperimentali ottenuti con la tecnica del ChIP-Seq precedentemente descritta rispetto alla proteina GATA1 al fine di determinare se è possibile identificare un modello statistico che sia alla base della loro generazione. Si procede, pertanto, all'esplorazione dei picchi della *coverage function* (alcuni esempi in Figura 3.1) per osservare se il campione deriva da distribuzioni note, come la distribuzione di Poisson o la distribuzione Binomiale Negativa. Osservando, però, che non è possibile associare questi modelli ai dati in esame, si definisce una rappresentazione alternativa dei picchi, basata, in particolare, su alberi sottesi ad essi; questa nuova visione permette, infatti, di proporre un buon modello di generazione dei dati.

3.1 Distribuzione di Poisson e Binomiale Negativa

La costituzione della *coverage function* come conteggi di eventi quali allineamenti di frammenti sul genoma porta alla formulazione dell'ipotesi, come in [19], [18], che i conteggi siano originati da una distribuzione di Poisson o da una Binomiale Negativa, eventualmente associata ad una componente puntuale concentrata in 0, data la elevata proporzione di zeri nella funzione stessa. Analizzando dati su differenti esperimenti di ChIP-Seq [3] si osserva, infatti, (Figura 3.2) che la percentuale di zeri nella *coverage function* si attesta mediamente oltre il 20%, risultato che, ad esempio, per la distribuzione di Poisson non è raggiungibile; la media della distribuzione, infatti, deve attestarsi su

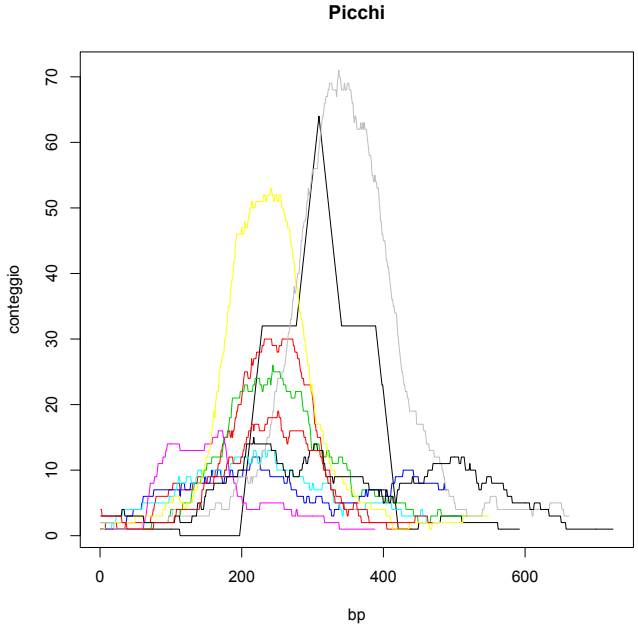


Figura 3.1: Esempio di 10 picchi della *coverage function*.

valori sufficientemente elevati da poter definire anche le massime quote della *coverage function*, riducendo così la probabilità di assumere valori nulli. Il modello di mistura tra una distribuzione nota e una delta nell'origine (*Zero-inflated* Poisson o Binomiale Negativa), quindi risulta quello più adeguato per la modellizzazione di questo tipo di funzioni.

Concentrandosi, tuttavia, sull'analisi delle regioni di significativa attività della *coverage function*, quelle cioè identificate dal *peak caller* come picchi, si nota che l'introduzione della componente concentrata in 0 non è necessaria; la percentuale di valori nulli, infatti, risulta significativamente minore rispetto all'andamento medio della *coverage function* e tale da non giustificare l'introduzione della componente concentrata in zero. Ci si può soffermare, dunque, sulla verifica dei modelli di generazione di Poisson o Binomiale Negativa; si presentano, quindi, le definizioni delle distribuzioni in esame.

Definizione 1. Distribuzione di Poisson: la variabile aleatoria X ha distribuzione di Poisson di parametro λ, $X \sim \mathcal{P}(\lambda)$, se:

$$\mathbb{P}(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}.$$

Observed vs. expected percentage of zero counts					
Cell line	**Observed**	**P**	**NB**	**ZIP**	**ZINB**
Gm10847InputIggrab	22.07%	9.85%	20.89%	23.513%	23.239%
Gm10847MusiggMusigg	22.013%	9.94%	20.868%	23.361%	23.22%
Gm12878InputIggrab	14.21%	1.08%	9.841%	15.709%	15.659%
Gm12878MusiggMusigg	26.532%	15.703%	25.846%	27.858%	27.64%
Gm12891InputIggrab	28.477%	9.49%	28.89%	29.57%	30.463%
Gm12891MusiggMusigg	23.583%	11.168%	22.71%	25.336%	25.536%
Gm12892InputIggrab	27.801%	13.15%	27.818%	29.03%	28.909%
Gm12892MusiggMusigg	30.705%	13.986%	31.292%	31.8%	29.211%
Gm15510InputIggrab	22.977%	10.883%	21.917%	24.467%	24.256%
Gm15510MusiggMusigg	20.134%	4.635%	19.06%	21.562%	21.43%
Gm18505InputIggrab	21.63%	9.213%	20.438%	22.899%	22.775%
Gm18505MusiggMusigg	27.304%	15.044%	27.647%	28.588%	28.727%
Gm18526InputIggrab	22.975%	11.938%	21.904%	24.475%	24.282%
Gm18526MusiggMusigg	25.478%	11.245%	25.425%	26.829%	26.825%
Gm18951InputIggrab	24.975%	12.554%	24.188%	26.483%	26.113%
Gm19099InputIggrab	20.983%	9.691%	19.188%	22.744%	22.456%
Gm19099MusiggMusigg	19.31%	4.581%	17.648%	20.762%	20.718%
Gm19193InputIggrab	24.35%	11.492	23.484%	25.86%	25.521%
Gm19193MusiggMusigg	21.558%	8.187%	20.037%	23.115%	22.884%
Helas3LargeFragment	19.1%	4.153%	16.580%	21.492%	21.274%
Helas3MnaseV2	18.885%	3.37%	18.504%	21.282%	21.854%
Helas3MouseiggV2	14.241%	.956%	10.421%	16.322%	16.158%
Helas3Nakeddna	15.007%	1.895%	11.678%	17.094%	17.125%
Hepg2ControlForskln	32.54%	15.266%	32.787%	33.927%	34.918%
Hepg2ControlInskln	45.481%	22.747%	46.602%	46.522%	48.245%
Hepg2InputV2	48.906%	36.024%	48.949%	50.9%	50.611%

Figura 3.2: Numeri di zeri effettivi e stimati per differenti esperimenti di ChIP-Seq.

La distribuzione di Poisson definisce la probabilità che in certo intervallo di tempo fissato si realizzino esattamente n eventi rari, sapendo che mediamente se ne realizzano λ. Nel caso specifico, si valuta come realizzazione dell'evento nell'intervallo di tempo fissato l'allineamento di un frammento isolato dall'immunoprecipitazione su una specifica base del genoma.
Si osserva, infine, una caratteristica peculiare della distribuzione di Possion ovvero il significato del parametro λ che risulta essere uguale sia al suo valore atteso $\mathbb{E}[X]$ che alla sua varianza $\mathrm{Var}[X]$.

Definizione 2. Distribuzione Binomiale Negativa: la variabile aleatoria X ha distribuzione di Binomiale Negativa di parametri p e r, $X \sim \mathcal{BN}(p, r)$, se:

$$\mathbb{P}(X = k) = \binom{k + r - 1}{k} (1 - p)^r p^k.$$

Questa distribuzione viene utilizzata in alternativa alla distribuzione di Poisson nel caso in cui il modello empirico presenti una varianza maggiore rispetto al valore medio, infatti, a differenza della distribuzione di Poisson in cui $\mathbb{E}[X] = \mathrm{Var}[X]$, per la distribuzione Binomiale Negativa vale:

$$\mathbb{E}[X] = \frac{pr}{1 - p} \quad \text{e} \quad \mathrm{Var}[X] = \frac{pr}{(1 - p)^2}$$

da cui si conclude, come ipotizzato, che $(1 - p) = \mathbb{E}[X]/\mathrm{Var[X]} \in [0, 1]$.
La distribuzione Binomiale Negativa, inoltre, valuta il numero di successi di una sequenza di prove di Bernoulli, ciascuna con probabilità di successo p, precedenti all'r-esimo insuccesso (r fissato). Osservando il significato dei parametri si nota che, al tendere di r all'infinito si elimina il vincolo sul numero massimo di insuccessi e pertanto il conteggio si riduce al semplice numero di successi delle prove di Bernoulli senza alcun vincolo, come nel caso della distribuzione di Poisson. Questa considerazione qualitativa è avvalorata dalla convergenza in legge della distribuzione Binomiale Negativa alla distribuzione di Poisson, infatti:

$$\mathbb{P}_{\mathcal{BN}}(X = k) \underset{r \uparrow +\infty}{\longrightarrow} \mathbb{P}_{\mathcal{P}}(X = k)$$

Questo legame con la distribuzione di Poisson, quindi, giustifica la ricerca del modello

nell'ambito di queste due distribuzioni, al fine di definire se è possibile trovarne una che ben rappresenti le caratteristiche dei dati in esame.

Si osserva, infine, che è possibile raggruppare i risultati di un esperimento generato da una distribuzione Binomiale Negativa in categorie e valutare le frequenze assolute e relative di occorrenza delle differenti categorie, ovvero se $X_1, X_2, \ldots$ sono variabili aleatorie iid da $\mathcal{BN}(p, r)$ che assumono valori discreti $\{0, 1, 2, \ldots\}$, si può definire il numero di occorrenze di ogni valore k con $k = \{0, 1, 2, \ldots\}$: $f_0, f_1, \ldots$ e valutare, poi, le loro frequenze relative $\psi_k = f_k/(\sum_{k\geq 0} f_k)$.
Questa diversa visione è alla base delle considerazioni proposte in Sezione 3.1.2 per valutare il modello di generazione dei dati in esame.

3.1.1 Test per l'analisi della distribuzione di Poisson

In questa sezione si vuole verificare se i conteggi di ciascun picco della *coverage function* possano derivare da una distribuzione di Poisson con opportuna media λ. Per la verifica di questa ipotesi si precisa che non è necessario assumere che tra i differenti picchi si mantenga lo stesso valore di media λ. Questa ipotesi, infatti sarebbe restrittiva e del tutto inadeguata dal momento che, come si è visto, differenti fattori biologici influenzano l'altezza della *coverage function* che, pertanto, può non risultare uniforme nemmeno nelle regioni con significativa attività. Si procede, pertanto, a un test separato per ciascuna regione di picco P_i al fine di valutare se i conteggi $X_1^i, X_2^i, \ldots, X_n^i$ che la definiscono derivano da una generica distribuzione di Poisson. Si conclude con un'opportuna correzione dei p-value dettata dalla molteplicità di questi test e con la conseguente verifica dell'evidenza statistica per accettare o rifiutare l'ipotesi nulla complessiva. Si considera, quindi, per il picco i-esimo il test:

$$\text{test } i = \begin{cases} H_0 : X_1^i, X_2^i, \ldots, X_{n_i}^i \overset{iid}{\sim} \mathcal{P}(\lambda_i) \\ H_1 : \text{dati associati a } P_i \text{ non derivano da distribuzione di Poisson} \end{cases}$$

con n_i lunghezza del picco P_i.

Si presentano ora due differenti test per verificare l'ipotesi nulla per ciascun picco, ovvero assumendo come dati i conteggi $X_1, X_2, \ldots, X_n$:

- *likelihood ratio test*: per questo test, basato sulla verosimiglianza associata alla

distribuzione scelta, si definisce la statistica test:

$$T_{LR} = 2 \sum_{j=1}^{n} X_j \ln\left(\frac{X_j}{\overline{X}_n}\right)$$

con $\overline{X}_n$ media campionaria degli x_j; sotto l'ipotesi nulla questa statistica si distribuisce asintoticamente (fissato λ, per $n \to +\infty$) come una variabile aleatoria Chi-quadro con $n-1$ gradi di libertà; da cui si ha evidenza a livello α per rifiutare H_0 se

$$T_{LR} > \chi^2_{1-\alpha}(n-1)$$

dove $\chi^2_{1-\alpha}(n-1)$ è il quantile di ordine $1-\alpha$ della distribuzione Chi-quadro con $n-1$ gradi di libertà. É possibile calcolare anche il p-value del test come la probabilità che una generica variabile aleatoria T distribuita come $\chi^2(n-1)$ assuma valori più estremi di T_{LR}, ovvero:

$$\text{p-value} = \mathbb{P}(T > T_{LR});$$

- *conditional Chi-squared test*: questo test è basato sull'indice di dispersione:

$$T_{CC} = \sum_{j=1}^{n} \frac{(X_j - \overline{X}_x)^2}{\overline{X}_n} = \frac{(n-1)S^2}{\overline{X}_n}$$

con S^2 stimatore non distorto della varianza di X. Sotto l'ipotesi nulla questa statistica si distribuisce per $n \to +\infty$ come una variabile Chi-quadro con $n-1$ gradi di libertà, da cui si determina la condizione per il rifiuto di H_0 per un test di livello α:

$$T_{CC} > \chi^2_{1-\alpha}(n-1)$$

e il p-value del test risulta, come nel caso precedente:

$$\text{p-value} = \mathbb{P}(T > T_{CC}).$$

É possibile, dunque, testare i picchi della *coverage function* per osservare se si può accettare l'ipotesi nulla di distribuzione di Poisson.
Valutando i p-value per questi test si ottengono valori molto ridotti che, anche a seguito

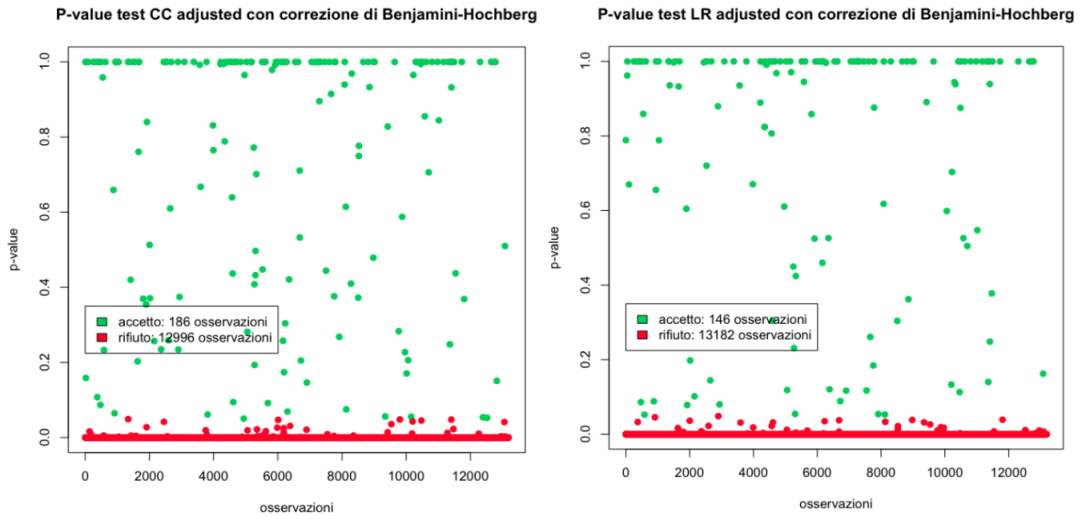


Figura 3.3: P-value per i test sulla distribuzione di Poisson.

di opportune correzioni per la molteplicità dei test (Figura 3.3), danno forte evidenza statistica per rifiutare l'ipotesi nulla: non si può, pertanto, concludere che i dati siano generati da una distribuzione di Poisson.

3.1.2 Test per l'analisi della distribuzione Binomiale Negativa

Per valutare la possibilità che i conteggi associati a ciascun picco siano generati da una distribuzione Binomiale Negativa $\mathcal{BN}(p, r)$ si confrontano, attraverso la statistica di Pearson, le frequenze assolute osservate (f_i) per i conteggi con quelle previste (θ_i) nel caso di distribuzione Binomiale Negativa. Si può definire la statistica test:

$$T = \sum_{i=0}^{k} \frac{(f_i - \theta_i)^2}{\theta_i}.$$

Questa statistica sotto l'ipotesi nulla di generazione dei dati da distribuzione Binomiale Negativa ha distribuzione Chi-quadro con $k - (2 + 1)$ gradi di libertà, con k numero di classi a cui si sottraggono i due gradi di libertà utilizzati per la stima dei parametri della distribuzione utili per definire le frequenze θ_i. Si procede, in primo luogo, alla definizione degli stimatori non distorti per media e varianza, che in termini di frequenze

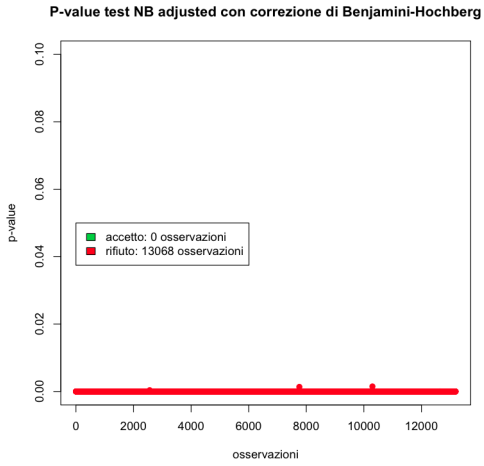


Figura 3.4: P-value per il test sulla distribuzione Binomiale Negativa.

assolute f_i, associate ai valori x_i, risultano:

$$\overline{X}_n = \frac{1}{\sum_{i=0}^{k} f_i} \sum_{i=0}^{k} (f_i x_i);$$

$$S^2 = \frac{1}{\sum_{i=0}^{k} f_i - 1} \left[\sum_{i=0}^{k} (f_i x_i^2) - \frac{\left(\sum_{i=0}^{k} (f_i x_i) \right)^2}{\sum_{i=0}^{k} f_i} \right].$$

Si trovano, dunque, per $q = 1 - p$:

$$\hat{q} = \overline{X}_n / S^2 \quad \text{e} \quad \hat{r} = \overline{X}_n \ \hat{q}/\hat{p}.$$

Calcolando quindi le probabilità di occorrenza di ciascun valore $\mathbb{P}(x_i)$ secondo la definizione di distribuzione Binomiale Negativa, si trovano le frequenze assolute stimate $\theta_i = (\sum_{i=0}^{k} f_i)\mathbb{P}(x_i)$ e si valuta, dunque, la statistica test precedentemente definita.
Analizzando i p-value (Figura 3.4), opportunamente corretti, si trova che vi è forte evidenza per rifiutare l'ipotesi nulla. Dunque, anche in questo caso, non è possibile definire la distribuzione di origine dei dati.

3.2 Rappresentazione dei picchi come alberi

Non potendo definire alcun modello di generazione dei dati a partire dai valori della *coverage function* si procede alla ricerca di una rappresentazione alternativa dei dati, che possa condurre alla caratterizzazione di un modello di origine.
Si introduce, quindi, per ogni picco l'albero con radice sotteso ad esso e si osserva che questi alberi sono alberi di Galton-Watson con una fissata densità generatrice π; sulla base di questa considerazione, quindi, si può proporre il modello di generazione.

3.2.1 Definizione rigorosa degli alberi

A partire dalla *coverage function*, definita come $f : \{1, \dots, N\} \to \mathbb{Z}^+$ con $f(x)$ numero di frammenti selezionati dall'immunoprecipitazione allineati alla x-esima base del genoma è possibile definire l'insieme di escursione $\mathcal{U}_h$

$$\mathcal{U}_h = \{(x, f(x)) | f(x) \geq h\} \quad \text{con} \quad h \in \mathbb{Z}^+$$

Questo insieme comprende le posizioni x e le rispettive immagini secondo f associate ad un livello di allineamento superiore ad h. La definizione di $\mathcal{U}_h$, associata alla valutazione della sua struttura al variare del parametro h permette di caratterizzare l'albero $\mathcal{T}$ associato al picco della funzione f.
Per la definizione rigorosa dell'albero è necessario introdurre per ogni $h \in \mathbb{Z}^+$ l'insieme C_h delle componenti connesse dell'insieme di escursione $\mathcal{U}_h$ (rappresentato in un esempio in Figura 3.5)

$$C_h = \{(a, b) \,|\, \forall\, x \in (a, b),\ f(x) \geq h \text{ e } \forall (c, d),\ c < a;\ d > b\ \ \exists\, y \in (c, d) \,|\, f(y) < h\}$$

C_h è l'unione di tutti gli intervalli associati a una quota superiore ad h, ovvero gli insiemi (a, b) t.c. $\{(a, b), (f(a), f(b)) \in \mathcal{U}_h\}$ per i quali una generica estensione (c, d) non appartiene più ad $\mathcal{U}_h$. Avendo definito l'insieme C_h al variare del parametro h in $\mathbb{Z}^+$ si può procedere alla definizione rigorosa dell'albero $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ caratterizzandone i vertici $\mathcal{V}$ e gli archi $\mathcal{E}$ ovvero:

- i vertici $\mathcal{V}$ corrispondono a tutte le componenti connesse dell'insieme C_h, al variare di h in $\mathbb{Z}^+$;

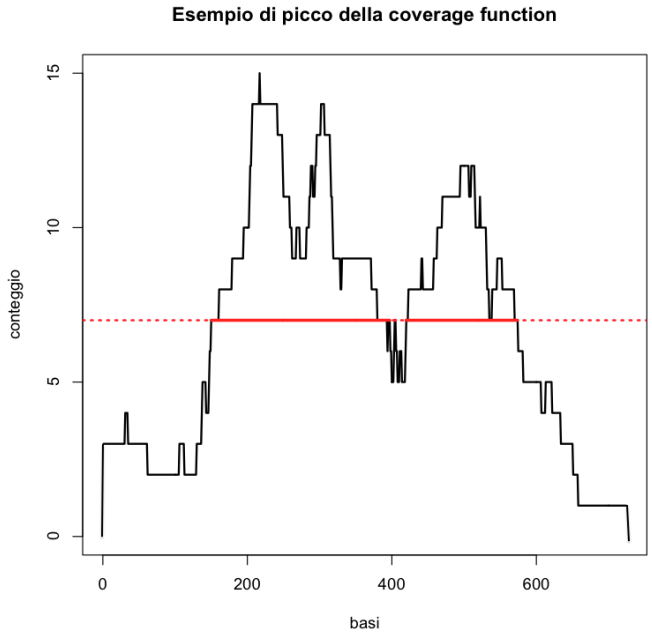


Figura 3.5: Insieme delle componenti connesse C_h (segmenti rossi con linea continua).

- gli archi $(i, j) \in \mathcal{E}$ connettono $i, j \in \mathcal{V}$ se le corrispondenti componenti connesse c_i e c_j, associate rispettivamente a un livello h_i e h_j (con $h_i < h_j$) sono tali che $h_i = h_j - 1$ e $c_i \subseteq c_j$.

Si trova, quindi, un albero (esempio in Figura 3.6) caratterizzato dalla radice corrispondente all'unico elemento dell'insieme delle componenti connesse C_0 le cui foglie sono associate ai massimi locali di f.

3.2.2 Definizione implementativa degli alberi

Una caratterizzazione equivalente, ma più agevole da implementare per l'albero $\mathcal{T}$ prevede la definizione, a partire dalla *coverage function* f di una funzione che consideri soltanto le variazioni di quota di f e non la lunghezza dei tratti orizzontali di quest'ultima; si definisce, quindi $g(x) : \{1, \ldots, m\} \to \mathbb{Z}^+$ tale che:

$$g(k)-g(k-1) = \begin{cases} 1 & \text{se } k\text{-esima variazione della } \textit{coverage function} \text{ è un incremento} \\ -1 & \text{se } k\text{-esima variazione della } \textit{coverage function} \text{ è un decremento} \end{cases}$$

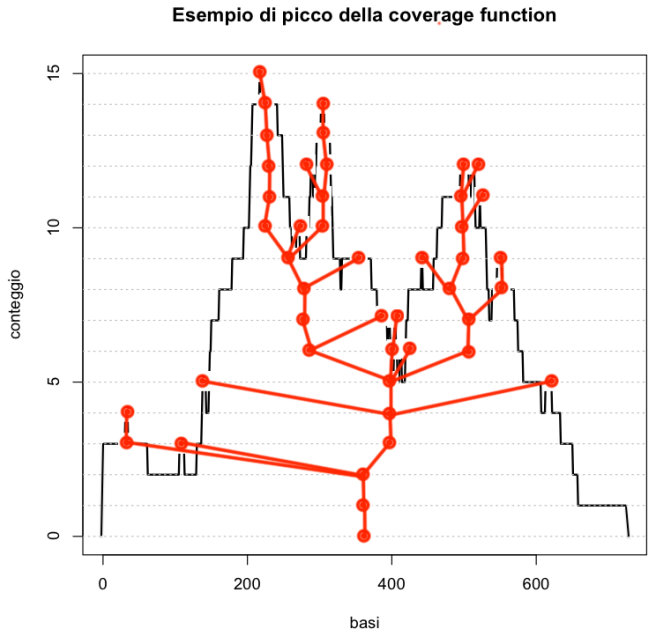


Figura 3.6: Rappresentazione dell'albero $\mathcal{T}$ associato al picco della *coverage function*.

con $g(0) = 0$.
La funzione rappresentata in Figura 3.7, quindi, a differenza di f, tiene conto solo delle variazioni di quota dell'allineamento e non della lunghezza dei tratti alla quota fissata.
Si procede, poi alla definizione dell'albero $\mathcal{T}$ con le seguenti regole:

- si definisce la radice dell'albero, associata al vetrice in (0,0) di g;
- per ogni aumento della funzione g, indicato in blu in Figura 3.7, si costruisce un nuovo nodo nell'albero allontanandosi dalla radice;
- per ogni diminuzione di g, punti rossi nella Figura 3.7, si ripercorre all'indietro l'albero avvicinandosi alla radice.

Si sottolinea che la procedura proposta prevede la generazione di un nuovo ramo dell'albero ogni qual volta si conclude un tratto discendente della funzione g e si inizia un nuovo tratto ascendente; a ogni incremento deve, infatti, corrispondere la creazione di un nuovo nodo dell'albero, non essendo possibile ripercorrere allontanandosi dalla radice i percorsi già creati. Per il picco in esame si trova con il metodo descritto, a partire

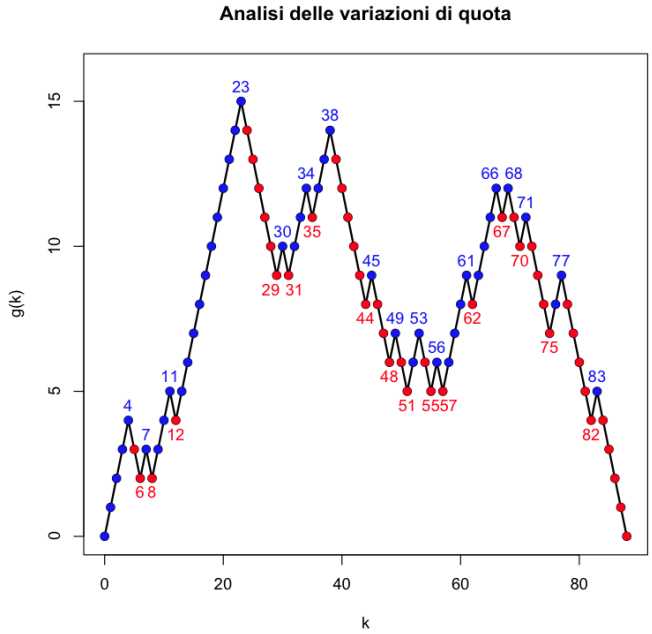


Figura 3.7: Funzione g definita a partire dalla *coverage function*.

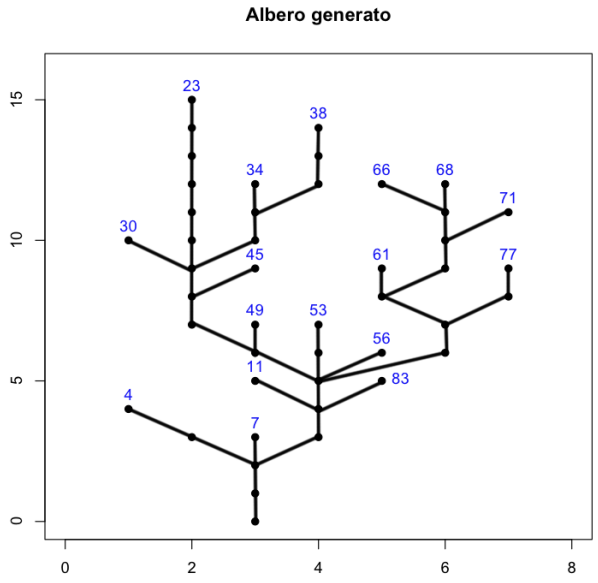


Figura 3.8: Albero generato a partire dalla funzione g.

da g rappresentata in Figura 3.7, l'albero proposto in Figura 3.8.

Osservazione: la definizione dell'albero associato al picco risulta una buona rappresentazione del dato, infatti ne caratterizza bene la forma, tuttavia è necessario precisare che non vi è corrispondenza univoca tra albero e picco; come è facilmente intuibile dalla definizione implementativa dell'albero esiste una corrispondenza univoca nella costruzione dell'albero a partire dalla funzione g, mentre f non è univocamente associata a g. Come si è già affermato, infatti, questa rappresentazione non tiene conto della lunghezza di eventuali tratti orizzontali della *coverage function*.
La rappresentazione dei picchi mediante alberi associati, dunque, non può essere considerata una rappresentazione equivalente a quella canonica, ma da questa possono essere ugualmente dedotte caratteristiche significative, utili per i differenti scopi di questo progetto, come la definizione di un modello di generazione di dati con struttura analoga a quelli proposti o la presentazione di caratteristiche di forma rilevanti per le successive analisi (Capitolo 4).

3.2.3 Alberi di Galton-Watson

Attraverso l'algoritmo proposto nella Sezione 3.2.2 è possibile definire per ogni picco P_i un albero $\mathcal{T}_i$. Questi alberi possono essere caratterizzati da vettori contenenti il numero di figli di ogni nodo dell'albero, valutato attraverso un'esplorazione ordinata dell'albero stesso. Ad esempio, esplorando a partire dalla radice e, a parità di quota, da destra a sinistra, gli alberi in Figura 3.9, si definisce il vettore cartterizzante:

```
Albero 1:
> 2 0 2 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
Albero 2:
> 1 2 0 1 1 1 1 1 2 0 2 0 1 1 1 1 1 1 3 0 0 1 1 1 1 1 1 1 1 2 0 2 0 1 2 1
1 0 0
Albero 3:
> 1 1 1 3 0 1 1 1 0 2 2 0 1 0 1 1 2 2 0 2 0 1 0 1 1 1 1 1 2 3 0 0 1 2 0 0
1 1 0
```

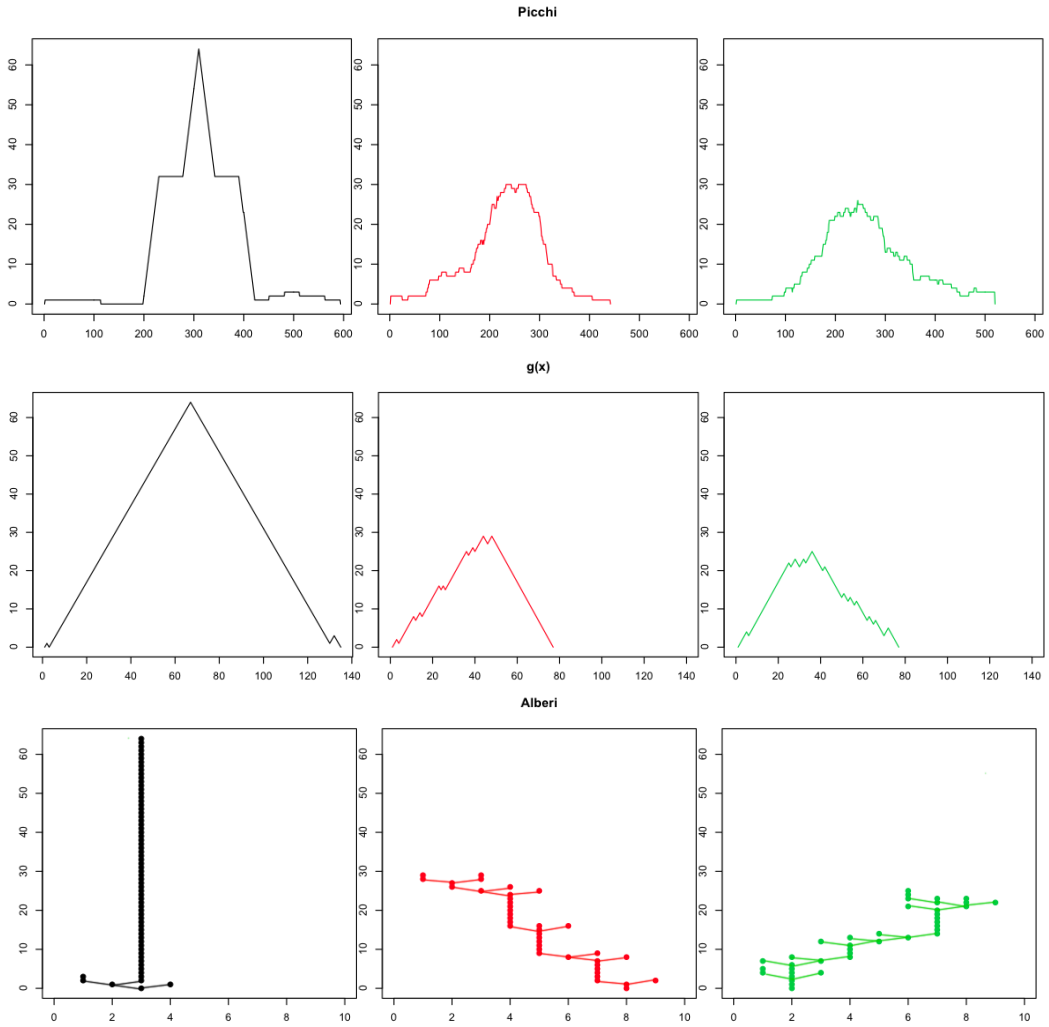


Figura 3.9: Costruzione dell'albero associato a diversi picchi della *coverage function* mediante la definizione di g.

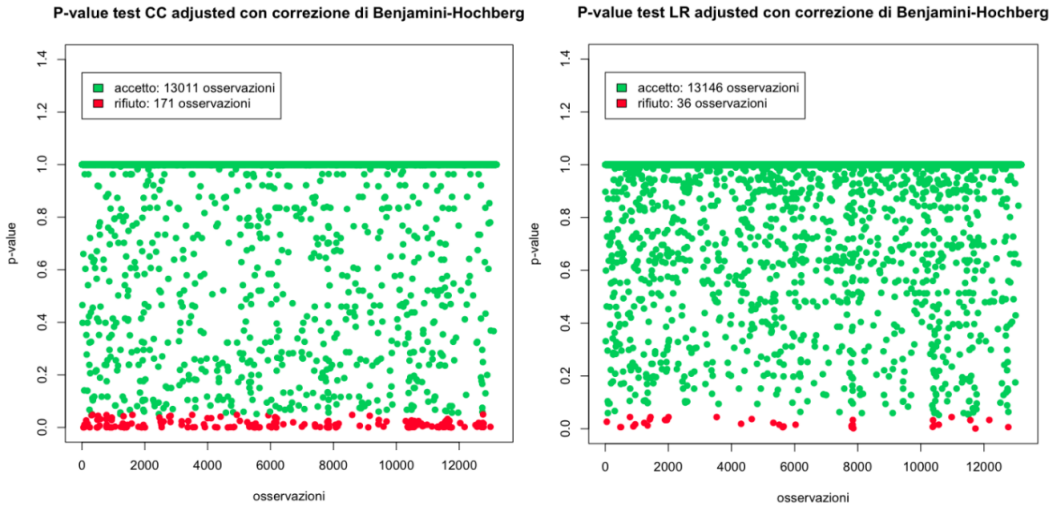


Figura 3.10: P-value per i test sulla distribuzione di Poisson dei vettori caratterizzanti gli alberi $\mathcal{T}_i$.

Avendo questo dato connesso ad ogni picco del dataset di partenza è possibile verificare l'assunzione proposta in [19] in cui si associa agli alberi sottesi un modello di Galton-Watson.

Definizione 3. Albero di Galtorn-Watson: data una generica distribuzione di probabilità $\pi(x)$ si può costruire un albero in maniera ricorsiva, ovvero partendo dalla radice si assegna a ciascun nodo un numero di figli generato dalla distribuzione π; gli alberi così definiti sono detti alberi di Galton-Watson con distribuzione generatrice π.

Nel caso specifico degli alberi generati dai picchi della *coverage function* si trova che effettivamente l'assunzione di albero di Galton-Watson è verificata e in particolare si prova che la distribuzione generatrice è la distribuzione di Poisson.
Questa affermazione è validata da test condotti come in Sezione 3.1.1. Si presentano a questo proposito, i dati relativi ai p-value dei test proposti opportunamente corretti con l'algoritmo di Benjamini-Hochberg per i test multipli e si osserva (Figura 3.10) come effettivamente ci sia forte evidenza per accettare l'ipotesi nulla di distribuzione generatrice di Galton-Watson di Poisson.
Per il calcolo del parametro λ caratteristico di questa distribuzione si procede a una

sua stima con il metodo di massima verosimiglianza, eguagliandolo, cioè, allo stimatore ottenuto dalla massimizzazione delle verosimiglianza dei dati, ovvero alla media campanaria.
Valutando questo parametro per tutti gli alberi $\mathcal{T}_i$ si osserva (Tabella 3.1) che i dati hanno un parametro associato λ_i molto simile tra loro, che si attesta attorno al valore 1.

minimo	1° quantile	mediana	media	3° quantile	massimo
0.2722	0.9667	0.9762	0.9745	0.9839	1.0040

Tabella 3.1: Statistiche riassuntive per il parametro λ.

Capitolo 4

Analisi multivariata dei dati ChIP-Seq

Come si è già presentato, i dati utilizzati in questo lavoro sono i picchi ottenuti dall'analisi mediante *peak-caller* della *coverage function*. L'obiettivo del progetto è l'analisi della forma di questi picchi al fine di determinarne le caratteristiche significative. A tale scopo è possibile definire, come si è visto nel dettaglio nel Capitolo 3, una differente rappresentazione dei picchi, basata sulla definizione degli alberi con radice da essi generati.
Lo scopo di questo capitolo è, dunque, proporre un'analisi di forma dei dati che si basi sulla definizione di opportuni indici, introdotti dall'analisi delle due rappresentazioni dei dati, al fine di determinare, poi, una prima divisione dei picchi in raggruppamenti omogenei.

4.1 Introduzione di indici di forma

A partire dalle due rappresentazioni dei dati già introdotte e rappresentate in Figura 4.2, è, dunque, possibile definire delle grandezze discrete che caratterizzino i dati stessi. Questi indici devono tenere conto di tutte le caratteristiche dei picchi, sia di quelle evidenti con la sola rappresentazione classica ma non rappresentabili in termini di alberi, sia delle caratteristiche evidenti in entrambe le visualizzazioni, sia di quelle proprie della nuova rappresentazione, al fine di avere una caratterizzazione il più completa possibile.

Le grandezze visualizzabili solo a partire dalla rappresentazione originale sono quelle connesse alla dimensione del picco, infatti la presenza di tratti orizzontali che non è

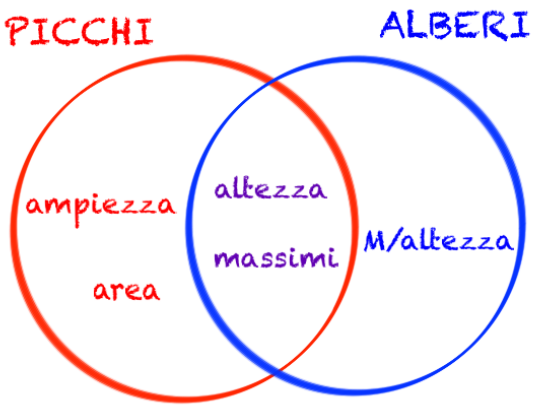


Figura 4.1: Distinzione tra gli indici di forma definiti a partire dai picchi o dagli alberi.

evidente nella rappresentazione ad alberi può variare la larghezza del picco e l'area sottesa da esso; si introducono, dunque, per la caratterizzazione di forma nei termini di dimensioni:

- l'area sottesa dal picco, che nel caso in esame quantifica il numero di frammenti allineati in quella regione del genoma;
- l'ampiezza a metà altezza del picco, che caratterizza la grandezza del picco non considerando possibili influenze del rumore che può variare significativamente l'ampiezza alla base del picco, ma non a metà della sua altezza.

Esistono, poi, indici osservabili sia a partire dall'albero che a partire dalla semplice visualizzazione originale, ovvero:

- l'altezza del picco, corrispondente nei termini della rappresentazione ad albero alla distanza massima tra le foglie e la radice;
- il numero di massimi locali del picco, ovvero il numero di foglie dell'albero.

Si precisa che per la definizione dei numero di massimi locali è necessario compiere un'analisi più dettagliata del semplice conteggio dei punti di massimo o delle foglie degli alberi, infatti la conformazione dei dati fa sì che spesso siano presenti piccole variazioni di altezza associate, ad esempio alla lunghezza dei *reads* che non sono in realtà associate a effettive variazioni di forma e che quindi non devono essere considerate per la

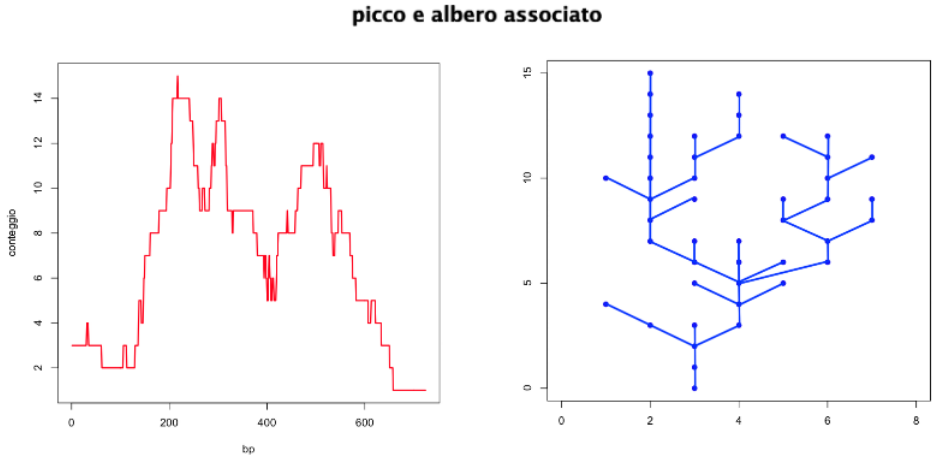


Figura 4.2: Esempio di picco della *coverage function* e albero associato.

valutazione del numero di massimi locali inteso come grandezza caratterizzante la forma del dato. Si introduce, quindi, una rappresentazione con attenuazione del rumore dei picchi attraverso la definizione delle *spline* sottese ai picchi (si veda il Capitolo 5), per poi procedere al conteggio dei massimi locali restanti che si distinguono significativamente dai vicini (ovvero con una differenza di quota sufficiente, con una soglia legata all'altezza del picco) e che, pertanto, sono indici di effettive variazioni nella conformazione. Per quanto riguarda la corrispondente analisi in termini di alberi, non verranno considerate tutte le foglie, ma solo quelle associate a rami di sufficiente lunghezza, dove anche per l'albero si definisce la soglia di lunghezza di interesse in base all'altezza complessiva.

La nuova rappresentazione dei dati attraverso gli alberi con radice sottesi permette, inoltre, di aggiungere informazioni all'analisi; si introduce, infatti, una misura che valuti la complessità dell'albero associato al fine di distinguere, ad esempio, un picco perfetto da un picco caratterizzato solo da rumore:

- l'indice di forma M, ovvero la cardinalità dell'insieme di *matching massimo*.

  Definizione 4. ***Matching massimo***: si dice, in particolare *matching* per un albero $\mathcal{T}$ un sottoinsieme $\mathcal{M}$ degli archi di $\mathcal{T}$ con la proprietà che nessuna coppia di archi in $\mathcal{M}$ ha in comune vertici di $\mathcal{T}$. Un *matching* si dice *massimo* se contiene almeno tanti archi quanti ne contengono tutti gli altri *matching*.

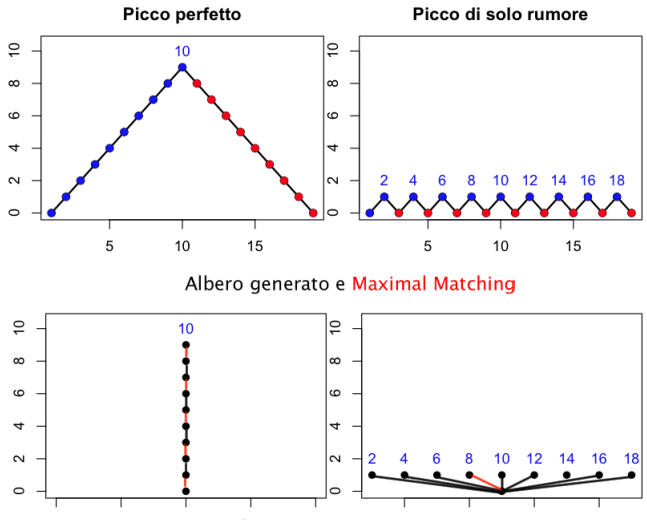


Figura 4.3: Confronto tra "picco perfetto" e "picco di solo rumore": funzione g e albero associato

L'importanza di questo valore può essere osservata analizzando due conformazioni opposte degli alberi.
Fissando il numero di m (nel caso particolare $m = 19$) di punti della funzione g (che, come si è visto nel Capitolo 3, è facilmente definibile a partire dal picco) si possono definire la funzione associata al "picco perfetto" e quella per il "picco di solo rumore", rappresentate in Figura 4.3. Si definiscono, poi, gli alberi generati da queste g con un numero di nodi n, in tal caso, pari a 10.

Per alberi così ridotti, inoltre, è facile calcolare direttamente l'indice M (algoritmo generale: *algoritmo blossom* di Jack Edmonds [8]) e si può dimostrare (come proposto dallo stesso Edmonds in [8]) che nel caso di "picco perfetto" questo assume valore massimo pari a $\lfloor n/2 \rfloor$ (per un albero composto da $n = 10$ nodi vale, dunque, $M = 5$), mentre assume valore minimo nel caso di "picco di solo rumore" ($M = 1$). L'indice, quindi, risulta effettivamente un buon indicatore della complessità del picco.
Un'altra peculiarità si questo indice, che lo rende molto adatto all'analisi degli alberi associati ai picchi della *coverage function*, è la sua stabilità all'introduzione di rumore. Come si è osservato valutando l'indice per un "picco di solo rumore", il numero di nodi dell'albero associati al rumore risulta ininfluente nella valutazione di M, pertanto anche per gli alberi della nostra analisi variazioni rumorose del picco non portano a cambiamenti nell'indice. Si osserva che per rumore nella g si intendono aumenti di un livello della

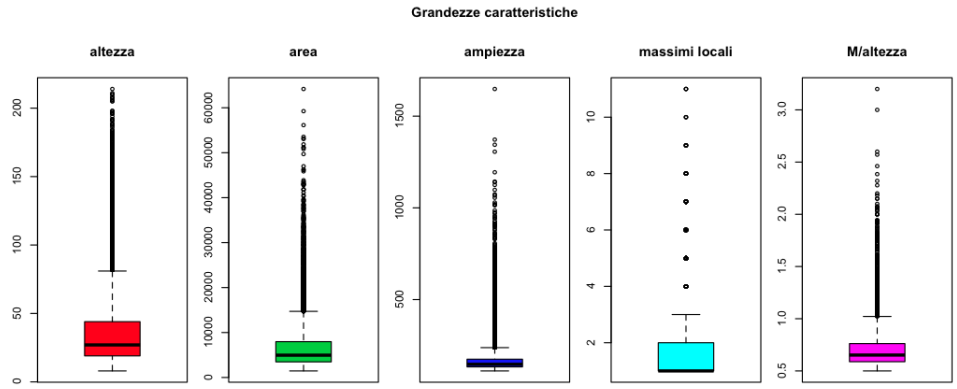


Figura 4.4: Boxplot degli indici di forma.

funzione seguiti da immediata diminuzione, corrispondenti cioè all'introduzione di un nuovo ramo nell'albero. La presenza di questi salti è associata alla variazione di un solo frammento nell'allineamento sul genoma e quindi a un cambiamento di aumento e successiva immediata diminuzione della *coverage function*, queste variazioni risultano poco significative nella valutazione complessiva, pertanto un buon indice di forma come M non deve variare alla loro presenza.
Si precisa, infine, che l'indice, data la sua definizione attraverso l'introduzione dei *maximal matching*, risulta, ovviamente, molto infulenzato dall'altezza del picco; nelle analisi successive, quindi, si utilizzerà l'indice diviso per l'altezza, così che risulti effettivamente un indice di forma, ovvero caratterizzi la complessità del picco, come la presenza di significative diramazioni nell'albero, senza diventare prevalentemente un indice di altezza.

Nel caso di alberi con struttura non semplice come quelli proposti in Figura 4.3, non è possibile un'immediata valutazione dell'indice M ed è, pertanto, necessaria l'introduzione di un algoritmo per il calcolo del *matching massimo*. Questo algoritmo (*algoritmo blossom*), introdotto da Edmonds [8], è definito nel dettaglio in Appendice B.2.

4.2 Rappresentazione e analisi degli indici

Valutando i diversi indici, si analizzare i loro andamenti complessivi (Boxplot in Figura 4.4) e le interazioni tra essi, in Figura 4.7.

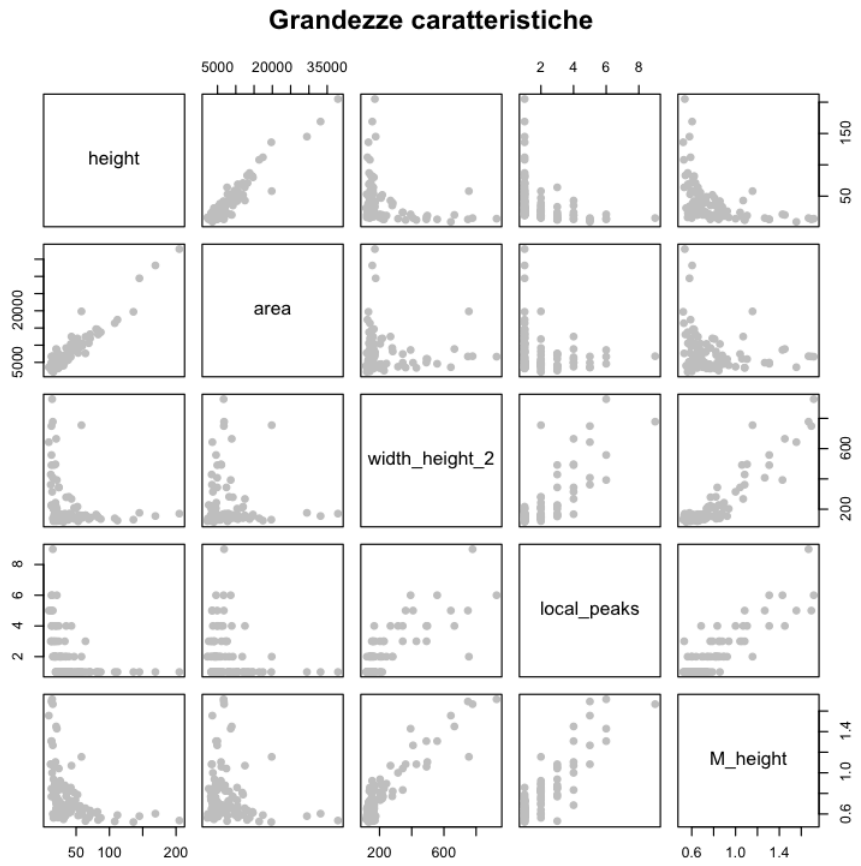


Figura 4.5: Analisi complessiva degli indici.

Osservando i boxplot si nota come complessivamente le grandezze risultino molto concentrate tra il primo e il terzo quantile empirico, ma siano presenti molti dati con comportamenti anomali; caratteristica che si auspica possa trovare una spiegazione con le analisi successive.

Relativamente ai legami tra le grandezze caratteristiche introdotte, si osserva che vi sono molte correlazioni positive; ad esempio si può rilevare, come previsto, che le due grandezze associate alle dimensione del picco (altezza e area) sono strettamente collegate. In merito all'indice di forma M, invece, si nota che la normalizzazione con l'altezza ha effettivamente eliminato il legame tra questo indice e l'altezza del picco; l'indice rappresenta, quindi, una buona caratterizzazione che ovviamente è connessa sia al numero di massimi locali (correlazione positiva), ovvero alla complessità di picco, sia all'ampiezza (correlazione positiva).

Si può concludere, dunque, che almeno nel caso di picchi associati a fattori di trascrizione, ovvero con andamenti simili a quelli proposti in questa analisi, gli indici introdotti caratterizzano abbastanza bene la forma dei dati.
Si procede, quindi, a una prima classificazione dei picchi sfruttando questa rappresentazione.

4.3 Algoritmo di classificazione *k-mean*

In questa sezione si presenta l'algoritmo di classificazione non supervisionata che verrà utilizzato per l'analisi dei dati, considerati puntualmente nello spazio dei cinque indici di forma introdotti: il *k-mean*. Questo è un algoritmo di classificazione non supervisionata non gerarchico, infatti non si procede, come nel caso degli algoritmi gerarchici al raffinamento della classificazione all'aumentare del numero di cluster, ma, per ogni k fissato definisce dei nuovi raggruppamenti.
La classificazione è definita secondo il seguente algoritmo:

dato un insieme di punti $\mathbf{x}_1, \dots, \mathbf{x}_N$ di $\mathbb{R}^d$ e fissato il numero di cluster k, si stabilisce la distanza utilizzata per confrontare i dati, generalmente la distanza euclidea:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_2 = \sqrt{\sum_{p=1}^{d} (x_1^p - x_2^p)^2}.$$

L'algoritmo poi si sviluppa:

0. se non sono assegnati all'inizio, si scelgono casualmente tra le osservazioni i centroidi iniziali dei k cluster: $\mathbf{c}_1, \dots, \mathbf{c}_k$;

1. si assegna ciascuna osservazione $\mathbf{x}_i$ al cluster j^* identificato dal centroide più vicino

$$j^* = \operatorname{argmin}_{j=1:k} d(\mathbf{x}_i, \mathbf{c}_j)$$

e si definiscono così i nuovi cluster C_1, C_k;

2. si definisce il nuovo centroide per il cluster:

$$\mathbf{c}_j = \text{argmin}_{\mathbf{c}\in\mathbb{R}^d} \sum_{\mathbf{x}_i \in C_j} d(\mathbf{x}_i, c);$$

poiché che la ricerca della soluzione di questo problema di minimizzazione in tutto lo spazio $\mathbb{R}^d$ può essere costosa, si può stabilire di limitare la ricerca all'insieme dei dati di partenza: $c \in \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ (algoritmo dei k-medoidi).

Si iterano i punti 1. e 2. fino a che non vi sono più modifiche nelle assegnazioni dei dati ai cluster.

Questo procedimento permette di identificare la migliore suddivisione avendo fissato il numero k di cluster. É necessario, quindi, definire quale è il numero ottimale di cluster per la classificazione. Questa decisione di basa generalmente su un indice calcolato a partire dalla scomposizione della varianza di un insieme di dati $\mathbf{x}_1, \ldots, \mathbf{x}_N$ raggruppati in k cluster di dimensioni $n_1, \ldots, n_k$. La varianza totale SS_{tot} si può, infatti, scomporre in una parte associata alla variabilità tra i gruppi $\text{SS}_{\text{between}}$ e in una associata alla variabilità all'interno dei gruppi $\text{SS}_{\text{within}}$

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}})^2}_{\text{SS}_{\text{tot}}} = \underbrace{\sum_{i=1}^{k} n_i(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^2}_{\text{SS}_{\text{between}}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^2}_{\text{SS}_{\text{within}}}.$$

A partire da questo risultato si può stimare la bontà della suddivisione valutando la percentuale di variabilità spiegata dai raggruppamenti, ovvero si calcola il rapporto tra la varianza interna ai gruppi e quella totale:

$$\text{SS}_{\text{W}\%} = \frac{\text{SS}_{\text{within}}}{\text{SS}_{\text{tot}}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_j)^2}{\sum_{i=1}^{g} n_i(\mathbf{x}_i - \overline{\mathbf{x}})^2}.$$

Avendo questo risultato per ogni raggruppamento e dunque in funzione del numero di cluster k si sceglie come valore ottimale il k associato a un $\text{SS}_{\text{W}\%}$ ridotto, o almeno connesso a una netta diminuzione dell'indice rispetto al caso di $k-1$ raggruppamenti,

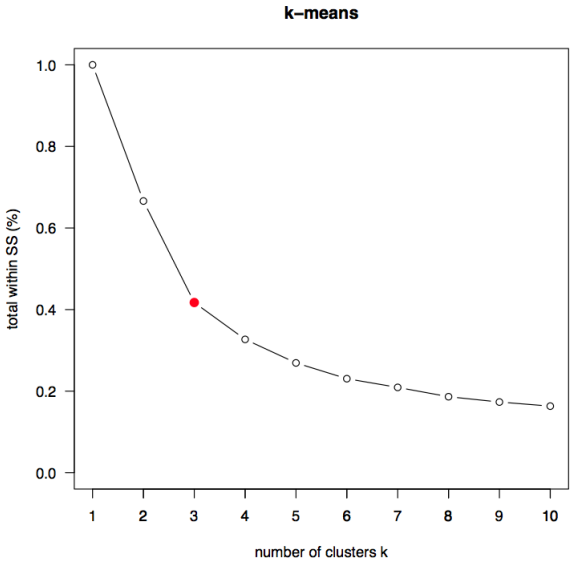


Figura 4.6: Andamento della proporzione di varianza spiegata dai raggruppamenti in funzione del numero di cluster k.

posto che poi il valore dell'indice di mantenga pressochè costante per numero di cluster maggiori di k.

4.4 *k-mean* per la classificazione dei dati

Analizzando i picchi come punti nello spazio dei cinque indici caratteristici mediante l'algoritmo del *k-mean* si ricerca, in primo luogo, il numero ottimale di raggruppamenti, valutando l'andamento di $SS_{W\%}$ in funzione del numero di cluster k. Come si osserva dalla Figura 4.6, in corrispondenza del valore $k = 3$ si nota una buona diminuzione della proporzione di varianza spiegata rispetto al caso $k = 2$, mentre per valori superiori non si attestano analoghi miglioramenti. Si sceglie, quindi, di procedere con l'analisi con tre cluster.

Si precisa che, data la particolare struttura degli indici, per applicare coerentemente l'algoritmo del $k - mean$ è necessario effettuare la standardizzazione dei dati. Senza questo accorgimento la loro disomogeneità negli ordini di grandezza, infatti, porterebbe a una classificazione eccessivamente influezata dagli indici più elevati.

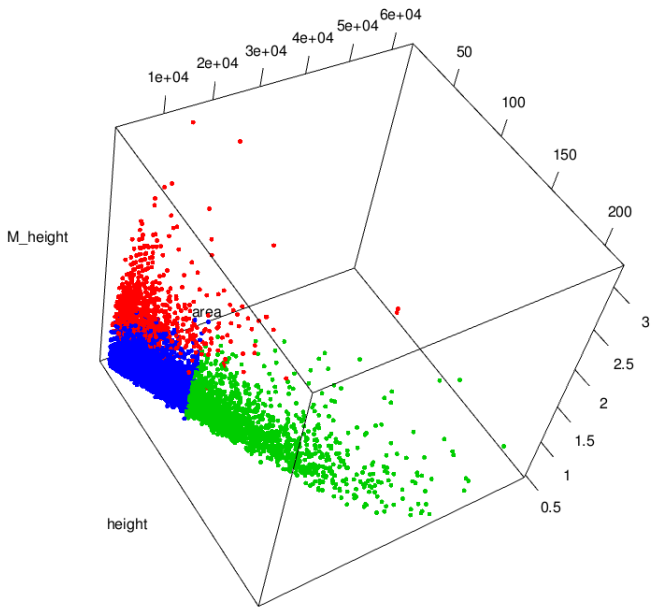


Figura 4.7: Raggruppamenti nello spazio definito dagli indici di forma area, altezza e M/altezza.

Valutando i risultati, si osservano dei buoni raggruppamenti nello spazio degli indici di forma: si propone in Figura 4.7 la rappresentazione nello spazio delle grandezze area, altezza e coefficiente M/altezza, spazio in cui sono ben evidenti le distinzioni tra i cluster.

Valutando oltre alle grandezze, anche i raggruppamenti dei dati (Figura 4.8), eventualmente anche allineati rispetto al punto di massimo (Figura 4.9) per agevolarne la lettura, si deducono le caratteristiche significative dei differenti cluster.

In particolare il primo raggruppamento comprende i picchi con un andamento molto frastagliato, associati cioè a un coefficiente M/altezza elevato e di dimensioni ridotte; gli altri due cluster, invece, comprendono i picchi con andamento regolare. Il secondo cluster, in particolare, comprende i picchi più alti e ampi, mentre il terzo picchi più bassi e regolari.

I dati raggruppati in questi ultimi due cluster sono probabilmente quelli associati alla regolare presenza della proteina attorno al filamento del DNA, mentre risulta interessante approfondire l'analisi del terzo raggruppamento, in particolare si introdurranno successivamente (Capitolo 5) tecniche di analisi differenti che proporranno ulteriori distinzioni e caratterizzazioni di questa categoria di dati.

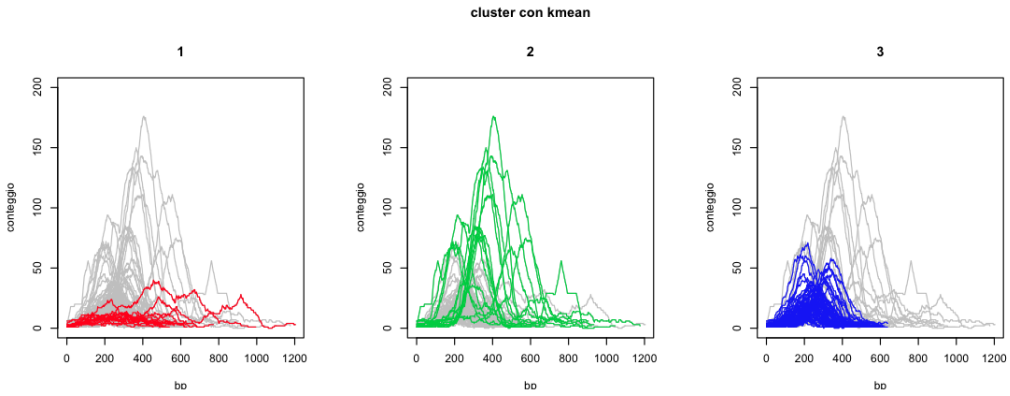


Figura 4.8: Classificazione nei tre cluster.

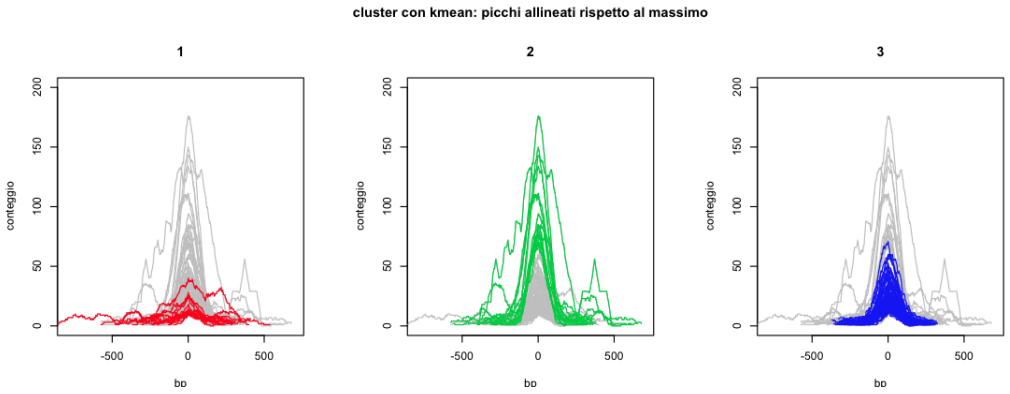


Figura 4.9: Classificazione nei tre cluster con allineamento dei picchi rispetto al punto di massimo.

Capitolo 5

Analisi funzionale dei dati ChIP-Seq

Si presenta, in questo capitolo, un altro approccio di analisi dei picchi di ChIP-Seq che non prevede la loro rappresentazione multivariata attraverso la definizione di indici di forma, ma consiste nell'analisi dei picchi P_i della *coverage function* nel loro complesso, ovvero come funzioni $f_i : \{0, 1, \ldots, L_i\} \to \mathbb{Z}^+$.
Si considera questa visione in quanto con tale rappresentazione può essere considerata la forma nel suo complesso e si prospettano risultati maggiormente approfonditi, in particolare si presenteranno ulteriori dettagli nella caratterizzazione dei picchi regolari e considerazioni interessanti sugli altri dati.

L'obiettivo di questa sezione risulta quindi la definizione di un nuovo metodo di classificazione da cui ricavare una più dettagliata caratterizzazione di forma dei picchi attraverso il loro raggruppamento in insiemi dalle caratteristiche funzionali simili.
Un efficace metodo di classificazione non supervisionata di dati funzionali è l'algoritmo del *k-mean* che, tuttavia, necessita di una generalizzazione rispetto al caso multivariato data la particolare struttura dei dati in esame.

5.1 Analisi dei dati funzionali

In questa sezione si presenta una descrizione introduttiva della tipologia di dati che si andranno ad analizzare, in particolare si propone una definizione di variabile aleaoria funzionale e se ne descrivono le principali caratteristiche, per poi delineare gli ambiti di indagine associati a questo tipo di dati [22].

Definizione 5. Funzione aleatoria: si dice funzione aleatoria $X : \Omega \to Y$ una variabile aleatoria definita su uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ a valori in un generico spazio metrico Y infinito dimensionale; in generale, poi, si predilige la scelta di uno spazio di Hilbert come L^2 per semplicità nella definizione della metrica.

Per questo genere di funzioni X è possibile definire la funzione di ripartizione associata

Definizione 6. Funzione di ripartizione: data la funzione aleatoria X definita sull'intervallo (a, b) si può valutare in corrispondenza di ciascuna realizzazione $\tilde{x}$ la funzione di ripartizione:

$$F_X(\tilde{x}) = \mathbb{P}(X(t) \leq \tilde{x}(t) \ \forall t \in (a, b))$$

Data l'infinita dimensionalità dello spazio delle variabili aleatorie funzionali non è possibile definire la densità di probabilità in uno specifico valore $\tilde{x}$, ma solo la probabilità che la variabile appartenga a un intorno della valutazione $\tilde{x}$ considerata. Avendo definito la metrica $\|\cdot\|_Y$ associata allo spazio delle funzioni, infatti, si definisce l'intorno

$$\mathcal{B}_r(\tilde{x}) = \{x \in Y : \|x - \tilde{x}\|_Y \leq r\}$$

e da questo la probabilità di appartenenza della variabile aleatoria a $\mathcal{B}_r(\tilde{x})$: $\mathbb{P}(X \in \mathcal{B}_r(\tilde{x}))$.

Oltre alla generalizzazione al caso funzionale degli aspetti caratterizzanti di una variabile aleatoria, è possibile definire anche valore atteso e varianza, per poi introdurre stimatori adatti alla loro previsione, ottenuti dalla semplice generalizzazione del caso di variabile aleatoria discreta. Concentrandosi sul caso particolare di variabili aleatorie funzionali in $L^2(a, b)$ si introducono:

Definizione 7. Valore atteso: si dice valore atteso di una generica variabile aleatoria $X : \Omega \to L^2(a, b)$ una funzione $\mu \in L^2(a, b)$ tale che

$$\mu(t) = \mathbb{E}[X(t)]$$

Definizione 8. Covarianza: si dice covarianza di una variabile aleatoria $X : \Omega \to L^2(a, b)$ una funzione $\Sigma \in L^2((a, b) \times (a, b))$ tale che

$$\Sigma(t, s) = \mathrm{Cov}(X(s), X(t)) = \mathbb{E}[(X(s) - \mu(s))(X(t) - \mu(t))].$$

Si definisce poi la varianza di una variabile aleatoria come Var(t)=Cov(t, t).

Avendo, quindi, un insieme di variabili aleatorie funzionali indipendenti e identicamente distribuite

$$X_1, X_2, \dots X_N : \Omega \to L^2(a, b)$$

si possono introdurre le statistiche riassuntive:

- media campionaria
$$\overline{X}(t) = \frac{1}{N} \sum_{i=1}^{N} X_i(t);$$
- covarianza campionaria
$$S(t, s) = \frac{1}{N-1} \sum_{i=1}^{N} \left[(X_i(t) - \overline{X}(t))(X_i(s) - \overline{X}(s)) \right].$$

I principali obiettivi e ambiti di indagine che si perseguono nell'analisi dei dati funzionali appena presentati sono, ad esempio:

- analisi esplorative dei dati associate ad una riduzione dimensionale non guidata dalle informazioni specifiche sul dataset o all'eliminazione del rumore associato ai dati stessi;
- caratterizzazione della variabilità di un insieme di dati funzionali;
- ricerca di raggruppamenti tra i dati con caratteristiche qualitative simili.

5.1.1 Esplorazione dei dati

Analizzando esempi di dati funzionali, si osserva che questi sono generalmente affetti da rumore che ne altera il valore puntuale, ma che non ne varia l'andamento complessivo; le analisi di forma dei dati, dunque, non devono essere eccessivamente influenzate da questi contributi puntuali esterni, ma devono riguardare unicamente l'andamento complessivo delle funzioni.
Per tale ragione si introduce un'approssimazione dei dati con funzioni sufficientemente regolari, ad esempio di tipo polinomiale o periodico. Questo accorgimento permette non solo di definire un'approssimazione adeguata dei dati, ma anche, data la possibilità

di introdurre una base per lo spazio delle funzioni approssimanti, di definire ciascuna funzione in esame unicamente attraverso i coefficienti associati a ciascun elemento della base.
Si può ottenere, così, una riduzione dimensionale dei dati, che non risultano più caratterizzati da infiniti valori puntuali al variare di t in (a, b), ma solo attraverso un numero sufficientemente rappresentativo di coefficienti per gli elementi della base introdotta.

Obiettivo 1. ***Approssimazione****: dato un insieme di funzioni $x_1, \dots x_N$ si ricercano delle buone funzioni approssimanti $s_1, \dots, s_N$ per eliminare la componente di rumore e definire una rappresentazione ridotta di questi dati funzionali.*

La migliore approssimazione per funzioni non periodiche è definita mediante le funzioni *spline*, che combinano la semplicità computazionale dei polinomi con una maggiore flessibilità, in modo da raggiungere migliori livelli di adattamento ai dati rispetto all'approssimazione polinomiale riducendo addirittura il numero di elementi della base.

Per la rappresentazione di una funzione mediante *spline* è necessario suddividere il dominio di analisi (a, b) in L regioni uguali, definendo un valore di ampiezza della suddivisione τ_L e $l = 1, \dots, L-1$ nodi. Su ogni sottointervallo, poi, la *spline* è un polinomio di ordine fissato m; l'ordine del polinomio (che coincide con quello della *spline* nel suo complesso, data l'uguaglianza dell'ordine dei polinomi in tutti i sottointervalli) è il numero di vincoli necessari per definirlo, ovvero il grado del polinomio più uno. Un vincolo fondamentale per la regolarità delle funzioni di tipo *spline* è la connessione dei polinomi dei sottointervalli nei nodi: le derivate fino al grado $m-2$ dei polinomi degli intervalli adiacenti devono, infatti, raccordarsi con continuità nei nodi.
Avendo delineato le caratteristiche delle *spline* è, dunque, possibile definire il numero di gradi di libertà dell'approssimazione, ovvero il numero di parametri che caratterizzano univocamente la curva complessiva.
Tenendo conto dei gradi di libertà in ognuno degli L intervalli (m) e dei $m-1$ vincoli di regolarità in ciascuno dei $(L-1)$ nodi interni, si trova che il numero di gradi di libertà complessivo è semplicemente l'ordine della *spline* sommato al numero di nodi interni scelti

$$k = m + L - 1.$$

La scelta dei valori dei k parametri necessari per definire la spline è basata sul criterio dei minimi quadrati, ovvero, data la generica funzione $x : (a, b) \to \mathbb{R}$, si ricerca la *spline* $s : [a, b] \to \mathbb{R}$ univocamente caratterizzata dai k parametri, $s = s(k)$ attraverso la minimizzazione del funzionale

$$D(k) = \|x - s\|_{L^2} = \int_a^b (s(t) - x(t))^2 dt.$$

É possibile introdurre un altro criterio per determinare la *spline* ottimale che tenga conto anche della regolarità della funzione s ottenuta; si introduce, quindi, nel funzionale obiettivo un termine di penalità associato generalmente alla derivata seconda della funzione

$$D_{\text{pen}} = \|x - s\|_{L^2} + \lambda \int_a^b (s''(t))^2 dt.$$

Il parametro λ è scelto in modo da bilanciare il contributo associato alla fedeltà ai dati, $\|x - s\|_{L^2}$, con la regolarità della funzione: un buon indicatore di irregolarità, infatti, è proprio l'integrale su tutto il dominio della derivata seconda della funzione. L'obiettivo risulta, dunque, ricercare una s che, oltre a rappresentare bene i dati, sia nel complesso sufficientemente regolare, quindi con un termine $\int_a^b (s''(t))^2 dt$ ridotto.

A conclusione di questa introduzione sulle *spline* si sottolinea che in generale si predilige l'utilizzo di *spline* cubiche, ovvero caratterizzate da polinomi di terzo grado nei sottointervalli (ordine $m = 4$), che nel complesso devono risultare $C^2((a, b))$; la scelta, poi, del criterio per la definizione della *spline* ottimale e quindi dei parametri che la caratterizzano, è strettamente connessa alla tipologia di dati in esame.

5.1.2 Analisi della variabilità per dati funzionali

Un ulteriore obiettivo di indagine è, come consuetudine per l'analisi statistica di variabili aleatorie, la valutazione della variabilità dei dati; in particolare per dati funzionali si osserva come sia possibile presentare due distinte tipologie di variabilità. Si può distinguere una variabilità nell'espressione del fenomeno (variabilità in ordinata o in ampiezza) e una connessa al tempo in cui il fenomeno si manifesta (variabilità in ascissa o in fase), come rappresentato in Figura 5.1.

Nell'analisi di un dataset funzionale è necessario, pertanto, distinguere tra queste

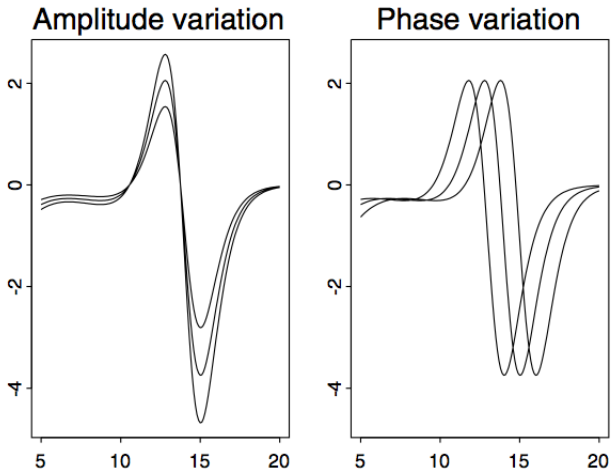


Figura 5.1: Variabiltà in ampiezza e in fase.

due tipologie di variabilità e, se possibile, scinderle al fine di discernere quali sono le componenti di un fenomeno che ne alterano l'evoluzione temporale e quali l'intensità (processo di registrazione di dati funzionali).

Obiettivo 2. ***Allineamento tra curve****: considerando un dataset* $x_1, x_2, x_3, \dots x_n$ *con* $x_i : (a, b) \to \mathbb{R}$ *si ricerca l'insieme di funzioni, dette warping function,* $h_1, h_2, \dots h_n$ *con* $h_i : (a, b) \to (a_0, b_0) \subseteq (a, b)$ *tali che le funzioni trasformate* $x_i^*(t) = x_i(h_i^{-1}(t))$ *per ogni* $i \in \{1, 2, \dots n\}$ *presentino solo variabilità in ampiezza.*

Analizzando le x_i^* è, dunque, possibile valutare le differenze in ampiezza del campione, mentre si osserva come le h_i racchiudano in sè le componenti di variabilità di fase. Per la scelta delle *warping function* si possono definire molteplici possibilità, tra cui:

- traslazioni: $h_i(t) = t - \delta_i$.
  Il caso descritto è quello più semplice, infatti la funzione registrata x_i^* è ottenuta dalla funzione di partenza come una semplice traslazione orizzontale. La scelta di questa categoria di *warping function* risulta molto efficace nel caso di processi in cui può esserci uno sfasamento dell'inizio della misurazione tra le diverse osservazioni, ma poi queste procedono allo stesso modo.

  In questo caso il problema dell'allineamento tra curve si riduce, dunque, alla semplice ricerca del parametro δ_i ottimale; esistono differenti metodi per raggiungere questo obiettivo:

– **metodo locale: registrazione attraverso *landmark***. Una volta definita una caratteristica specifica, o *landmark* delle funzioni in esame (ad esempio il punto di massimo globale) si procede alla loro traslazione affinché questo punto vada a coincidere per tutte le curve; il parametro δ_i, sarà, quindi, semplicemente la differenza tra l'ascissa del punto caratteristico della funzione x_i e quello di riferimento (ottenuto ad esempio dalla semplice media dei *landmark* delle diverse funzioni). Questo metodo di allineamento presenta delle evidenti criticità dovute, ad esempio, alla possibile ambiguità nella caratterizzazione del *landmark* o alla presenza di variazioni nelle regioni non scelte come *landmark* che sono ignorate.

– **metodo globale: registrazione con funzione target**. Questo metodo considera le funzioni nel loro complesso e non solo in uno specifico punto come nel caso di registrazione con *landmark*, tuttavia necessita della definizione di una funzione target x_0 alla quale allineare tutte le funzioni. Una volta nota la funzione di riferimento si stabiliscono i parametri δ_i dalla minimizzazione del funzionale globale:

$$\text{REGSSE} = \sum_{i=1}^{N} \int_{\tau} [\underbrace{x_i(t+\delta_i)}_{x_i^*(t)} - x_0(t)]^2 \, dt.$$

La criticità di questo metodo risulta essere come già osservato la definizione della funzione di riferimento.

- generiche funzioni continue, derivabili e crescenti: $h_i(t)$.
Si sceglie di utilizzare funzioni crescenti nel dominio per evitare di ammettere possibili inversioni tra le funzioni utili soltanto al miglioramento dell'allineamento. Generalmente vengono scelte, inoltre, come possibili *warping function* famiglie di funzioni parametrizzabili, come ad esempio le affinità, così che l'indagine si riduca alla ricerca di parametri che ottimizzino un funzionale come il REGSSE precedentemente definito. A differenza delle sole traslazioni, la scelta di funzioni più complesse può consentire il confronto tra processi che non presentano variabilità in fase dovuta solo alle differenze nell'istante di inizio, ma anche tra processi che presentano evoluzioni temporali differenti, ad esempio connesse a differenti

velocità di sviluppo.

Una volta completato il processo di registrazione è effettivamente possibile caratterizzare le due differenti variabilità nel datastet. L'analisi delle *warping function* h_i permette, infatti, di valutare la variabilità in fase attraverso il confronto tra i diversi processi di registrazione delle funzioni valutando, ad esempio, se prevedono accelerazioni, decelerazioni o semplici traslazioni temporali. L'analisi delle funzioni registrate x_i^*, invece, permettere di analizzare la variabilità in ampiezza, essendo l'unica componente rimasta che distingue le diverse realizzazioni.

5.1.3 Classificazione non supervisionata di dati funzionali

Avendo delineato le caratteristiche complessive della variabilità di questa categoria di dati, è possibile introdurre un metodo efficace per la divisione del dataset in componenti con andamenti simili [23].

Obiettivo 3. *Dato un insieme di funzioni* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ *che, nel caso più generale sono funzioni* $\mathbf{x}_i : \mathbb{R} \to \mathbb{R}^d$*, si vuole procedere alla loro suddivisione in classi omogenee.*
L'obiettivo risulta, dunque, l'introduzione di una divisione in raggruppamenti in base alle distinzioni in ampiezza delle funzioni, una volta che queste hanno comportamento in fase confrontabile.

Questo obiettivo viene perseguito attraverso l'estensione del metodo standard del *k-mean* (proposto in Sezione 4.3) prevedendo anche la possibilità che le funzioni vengano allineate (*k-mean alignment*). L'allineamento considerato in questo caso è basato su un insieme di *warping function* $\mathcal{W}$ sufficientemente generale, purchè siano mantenute le proprietà di regolarità e monotonia precedentemente richieste; si ipotizza, inoltre, un confronto tra funzioni di tipo globale, basato cioè sull'introduzione di un indice ρ che consideri le funzioni nel loro complesso e non unicamente in punti particolari. Si definiscono, dunque, le proprietà richieste per $\mathcal{W}$ e ρ:

1. l'indice di similarità ρ deve essere limitato, con valore massimo pari a 1 e maggiore risulta l'indice, più le curve in analisi sono simili. Le semplici proprietà che deve rispettare sono:

   - riflessività: $\rho(\mathbf{x}, \mathbf{x}) = 1 \quad \forall \mathbf{x}$

- simmetria: $\rho(\mathbf{x}_1, \mathbf{x}_2) = \rho(\mathbf{x}_2, \mathbf{x}_1) \ \ \forall \mathbf{x}_1, \mathbf{x}_2$
- transitività: se $\rho(\mathbf{x}_1, \mathbf{x}_2) = 1$ e $\rho(\mathbf{x}_2, \mathbf{x}_3) = 1$ allora $\rho(\mathbf{x}_1, \mathbf{x}_3) = 1 \ \ \forall \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$;

2. lo spazio delle *warping function* $\mathcal{W}$ deve possedere una struttura di gruppo rispetto all'operazione di composizione $\circ$, ovvero gli elementi di $\mathcal{W}$ rispettano la proprietà associativa

$$(\mathbf{x}_1 \circ \mathbf{x}_2) \circ \mathbf{x}_3 = \mathbf{x}_1 \circ (\mathbf{x}_2 \circ \mathbf{x}_3)$$

ed esistono l'elemento nullo e quello inverso rispetto alla stessa operazione $\circ$;

3. l'indice ρ e lo spazio $\mathcal{W}$ sono consistenti nel senso che se due curve $\mathbf{x}_1$ e $\mathbf{x}_2$ hanno un indice di similarità $\rho(\mathbf{x}_1, \mathbf{x}_2)$, la simultanea trasformazione rispetto a una qualunque delle funzioni in $\mathcal{W}$ non altera l'indice:

$$\rho(\mathbf{x}_1, \mathbf{x}_2) = \rho(\mathbf{x}_1 \circ h, \mathbf{x}_2 \circ h) \ \ \forall h \in \mathcal{W}.$$

Questa proprietà è fondamentale dal momento che garantisce che non sia possibile ottenere un aumento dell'indice di similarità semplicemente applicando una determinata funzione h a entrambe le curve di partenza.

Gli indici di similarità e gli insiemi di trasformazioni ammissibili sono molti; nel caso specifico del lavoro in esame, ci si concentra su un indice che confronta due curve analizzandone l'andamento complessivo, ovvero il coseno tra le curve.

Definizione 9. Coseno tra $\mathbf{x}_1$ e $\mathbf{x}_2$: date due generiche funzioni $\mathbf{x}_1, \mathbf{x}_2 : \mathbb{R} \to \mathbb{R}^d$ si dice coseno tra di esse la media dei coseni tra gli angoli sulle componenti omologhe delle curve $\mathbf{x}_1$ e $\mathbf{x}_2$,ovvero sulle componenti $x_{i,p}$ del vettore $\mathbf{x}_i = (x_{i,1}, ...x_{i,d})$.

$$\rho(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{d} \sum_{p=1}^{d} \frac{\int_{\mathbb{R}} x_{1,p}(t) x_{2,p}(t) dt}{\sqrt{\int_{\mathbb{R}} (x_{1,p}(t))^2 dt} \sqrt{\int_{\mathbb{R}} (x_{2,p}(t))^2 dt}}.$$

Per quanto riguarda lo spazio $\mathcal{W}$ si sceglie per semplicità lo spazio delle trasformazioni affini crescenti, ovvero l'insieme:

$$\mathcal{W} = \{h : h(t) = at + b \text{ con } a \in \mathbb{R}^+, b \in \mathbb{R}\}.$$

L'algoritmo del *k-mean alignment* prevede, dunque:

dato un insieme di funzioni $c_1, \ldots, c_n : \mathbb{R} \to \mathbb{R}^d$ e il numero di cluster k:

0. definizione casuale dei centri inziali dei cluster $\varphi_1, \ldots \varphi_k$ (se non già assegnati in origine)

1. per ogni osservazione c_i:

   (a) per ogni cluster j ricerca della $h_i^j(t)$ in $\mathcal{W}$ che massimizzi la similarità $\rho(\varphi_j, c_i \circ h_i^j)$, e definizione dell'indice di similarità ottimo ρ_{ij}^*

   (b) assegnamento dell'osservazione i al cluster con indice di similarità massimo $j^* = \text{argmax}_{j=1:k}(\rho_{ij}^*)$

2. avendo definito la suddivisione delle osservazioni nei cluster, definizione dei nuovi centri di riferimento per ogni cluster

3. Ripetizione dei punti 1. e 2. fino a che non ci sono alterazioni nell'assegnamento delle osservazioni ai cluster o viene raggiunto il numero massimo di iterazioni ammissibili.

Si osserva che la versione dell'algoritmo presentata prevede, come di consueto per un algoritmo di tipo *k-mean*, la definizione iterazione per iterazione dei centri di ciascun cluster aggiornandoli a seconda degli elementi presenti nel cluster.
É possibile prevedere anche una versione semplificata nel caso di centri già fissati; il problema, in questo caso, si riconduce a un'analisi di allineamento ottimo e assegnamento al cluster corrispondente. Infatti in questo caso risulta sufficiente valutare, per ogni funzione, la similarità ottima con ciascun centro, ovvero quella corrispondente alla migliore trasformazione all'interno di $\mathcal{W}$ e decidere poi l'appartenenza al cluster con similarità massima (punto 1. dell'algoritmo completo).

Osservazione: nel caso di un gran numero di dati, il procedimento proposto può risultare computazionalmente oneroso, pertanto è necessario introdurre una procedura

che renda più efficace la definizione dei raggruppamenti finali. Si considera la procedura della *bootstrap aggregation* in cui si stabilisce di selezionare casualmente dall'insieme dei dati complessivi dei campionamenti di un numero limitato di dati. Per ciascuno di questi campionamenti è possibile poi applicare l'algoritmo del *k-mean alignment* che, avendo ridotto la numerosità dei dati, risulta computazionalmente accettabile, e poi raccogliere tutti i dati in un'unica classificazione complessiva raggruppando i cluster omogenei e ottenendo la suddivisione cercata.

5.2 Applicazione ai ChIP-Seq

L'analisi dei dati proposti con un approccio funzionale ha l'obiettivo di definire una nuova classificazione per i picchi della *coverage function* che non tenga conto della forma solo in termini degli indici proposti nel Capitolo 4, ma che consideri i dati nel loro complesso.
Questo metodo risulta applicabile anche a dati differenti da quelli proposti nell'esperimento in esame, ad esempio ottenuti da indagini non su fattori di trascrizione, ma su istoni, che manifestano picchi della *coverage function* molto differenti e che pertanto potrebbero non risultare ben caratterizzati dalle grandezze discrete proposte in precedenza.

L'utilizzo di un classificatore funzionale, basato in particolare sull'algoritmo del *k-mean* con allineamento, permette, inoltre, di raccogliere informazioni sulle eventuali trasformazioni in fase che celano un comportamento analogo dei dati. Risulta determinante valutare queste trasformazioni, che permettono ai diversi dati di appartenere allo stesso raggruppamento: da questi risultati, come si osserverà nel dettaglio nelle prossime sezioni, si possono ipotizzare significativi risvolti biologici.

In questo Capitolo si conduce un'analisi funzionale sui dati valutando le loro diversità e somiglianze con l'indice proposto in Definizione 9, ovvero il coseno calcolato direttamente sui dati; si introduce, poi, una differente classificazione basata sull'andamento della derivata dei dati al fine di focalizzarsi maggiormente sull'andamento qualitativo trascurando il valore puntuale delle funzioni. Questa classificazione proporrà dei raffinamenti della prima analisi, ma per essere compiuta necessita della definizione di un'opportuna funzione approssimante s con una regolarità tale da permettere questo tipo di valutazione, ovvero una *spline* che per risultare una buona approssimazione sia in termini di *fitting* che di regolarità sarà scelta nell'ambito delle *spline* cubiche con

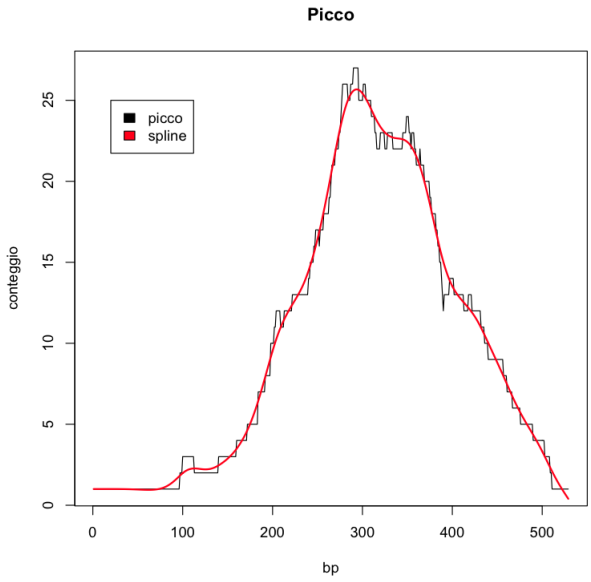


Figura 5.2: Esempio di picco e approssimazione con *spline* con penalizzazione.

penalizzazione sulla derivata seconda; si vuole definire, infatti, un buon compromesso tra l'influenza dei dati e la regolarità della funzione, evitando l'introduzione di eccessive variazioni nella concavità della curva approssimante. A titolo di esempio si presenta, in Figura 5.2, l'andamento della *spline* approssimante per un generico picco del dataset.

Il lavoro riguardante l'analisi funzionale dei dati è suddiviso, quindi, in un'analisi dei picchi con l'algoritmo descritto in precedenza sia considerando come insieme di *warping function* le traslazioni (Sezione 5.2.2) sia le affinità (Sezione 5.2.3). In entrambi i casi si esaminano i risultati ottenuti sia valutando la similarità tra i dati, sia tra le loro derivate, per definire in conclusione la migliore classificazione. In primo luogo si compiono analisi esplorative su un insieme ridotto di dati al fine di determinare i parametri ottimali; attraverso la procedura di *bootstrap* si procede, poi, all'applicazione del metodo ottimale a un insieme più ampio di dati, per dedurre una classificazione complessiva e valutare i possibili riscontri a essa connessi.

5.2.1 *k-mean alignment* per i dati ChIP-Seq

L'algoritmo del *k-mean alignment*, come descritto nel dettaglio in Sezione 5.1.3, porta alla distinzione in k raggruppamenti dei dati in base alla loro variabilità in ampiezza,

tenendo conto, attraverso le *warping function*, delle possibili alterazioni in fase.

Per questa analisi si utilizza come indice di similarità il coseno ρ, come introdotto nella Definizione 9 calcolato in primo esame tra le funzioni e in secondo luogo tra le derivate delle loro approssimazioni.
Attraverso un'analisi introduttiva si ricercano il numero di cluster ottimale per la distinzione e la classe di *warping funcition* $\mathcal{W}$ migliore per il dataset proposto. Per questo tipo di analisi esplorativa si assumono come trasformazioni ammesse le traslazioni, le dilatazioni, ovvero trasformazioni dell'asse dei tempi del tipo $h(t) = \alpha t$ con $\alpha \in \mathbb{R}^+$, e le affinità, ovvero una combinazione di queste due categorie $h(t) = \alpha t + \beta$, con $\alpha \in \mathbb{R}^+, \beta \in \mathbb{R}$; si considera, inoltre, il caso di semplice classificazione senza allineamento. Si assume, infine, la possibilità di raggruppare i dati in un numero di cluster variabile tra uno e sei; la possibilità di definire un unico cluster permette di indagare il ruolo del solo allineamento per le funzioni in esame: in questo caso, infatti, si valuta la similarità complessiva dei dati senza introdurre raggruppamenti, considerando solo la variazione in fase.
La valutazione di questi casi viene compiuta considerando l'indice

$$\rho^* = \frac{1}{n} \sum_{i=1}^{n} \rho^*_{i,j^*}$$

ovvero la media degli indici di similarità ottimi ρ^*_{i,j^*}, rispetto, cioè, alla migliore *warping function* nella classe, per ogni osservazione $i = 1, \dots n$ rispetto al centroide del cluster di appartenenza j^*.
In Figura 5.3 si presenta, quindi, l'andamento dell'indice di similarità tra i dati

$$\rho(x_1, x_2) = \frac{\int_{\mathbb{R}} x_1(t) x_2(t) dt}{\sqrt{\int_{\mathbb{R}} (x_1(t))^2 dt} \sqrt{\int_{\mathbb{R}} (x_2(t))^2 dt}}$$

descritto al variare della tipologia di trasformazioni $\mathcal{W}$ e del numero di cluster. Si ricercano, in corrispondenza di ogni tipologia di trasformazione, quei valori di k che permettono un significativo miglioramento nella similarità, ovvero dove si assiste a un netto cambiamento nel valore di similarità globale, che poi all'aumentare successivo del numero di cluster non varia più (gomiti nelle curve in Figura 5.3).
Analizzando le diverse categorie di trasformazioni ammissibili, si rileva che:

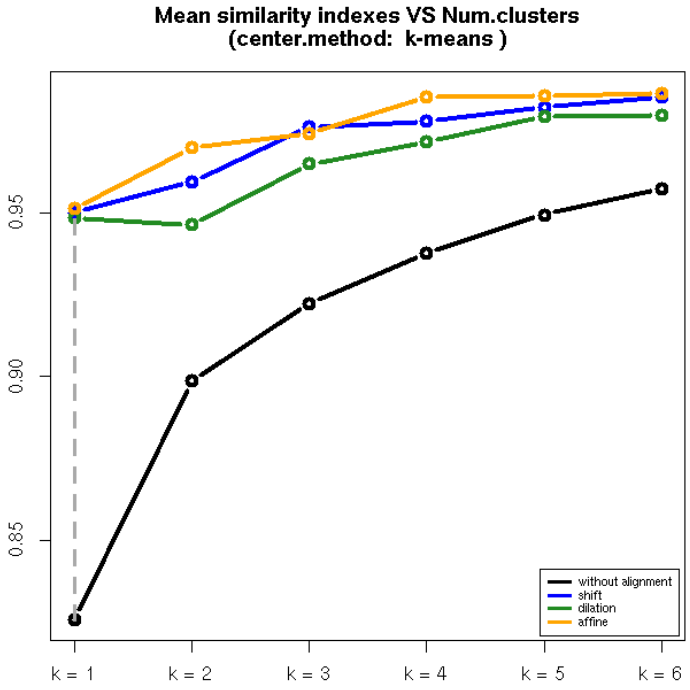


Figura 5.3: Analisi dei diversi tipi di classificazione al variare di $\mathcal{W}$ e del numero di cluster.

- rispetto alla categoria delle traslazioni, il numero di cluster appropriato risulta essere $k = 3$, infatti rispetto a $k = 2$ si osserva un netto aumento dell'indice ρ^*, mentre all'aumentare di k oltre a 3, le variazioni non sono significative;

- introducendo anche le dilatazioni, analizzando, cioè le trasformazioni affini si trova che il numero di cluster adeguato è $k = 4$;

Ulteriori considerazioni, invece possono essere dedotte valutando la similarità tra le derivate:

$$\rho(x_1, x_2) = \frac{\int_{\mathbb{R}} s_1'(t)s_2'(t)dt}{\sqrt{\int_{\mathbb{R}}(s_1'(t))^2dt}\sqrt{\int_{\mathbb{R}}(s_2'(t))^2dt}}$$

con s_1 e s_2 le curve approssimanti precedentemente introdotte.

Si precisa che la possibilità di utilizzare questo indice è dettata dall'approssimazione mediante *spline* dei dati. Questa tipologia di funzioni, infatti, ha caratteristiche di regolarità fondamentali tali da permettere di considerare come indice di similarità non solo il coseno tra le funzioni, ma anche quello tra le derivate; come si è già sottolineato, infatti le *spline* cubiche sono C^2 nel dominio di definizione.

Come nel caso precedente si considera la possibilità di un numero di raggruppamenti che varia da due a sei e un insieme di *warping function* che varia tra traslazioni, dilatazioni o affinità. Si presenta in Figura 5.4 il grafico riassuntivo riguardante l'indice di similarità medio ρ^* al variare dei parametri descritti. Valutando anche in questo caso l'andamento dell'indice ρ^* si osserva che:

- per quanto riguarda l'insieme $\mathcal{W}$ delle traslazioni l'analisi risulta interessante in corrispondenza di $k = 4$, infatti per $k = 5$ il livello di similarità risulta invariato, mentre per valori minori la similarità risulta troppo ridotta;

- introducendo anche la possibilità di dilatazioni e quindi considerando l'insieme delle affinità, non si assiste a un significativo miglioramento della classificazione. Vi è solo un'alterazione nella definizione dei cluster: sono raccolti in un unico raggruppamento i picchi più regolari, ma non si introducono ulteriori caratterizzazioni dei dati.

Osservazione: Si omette l'analisi delle sole dilatazioni e della classificazione senza

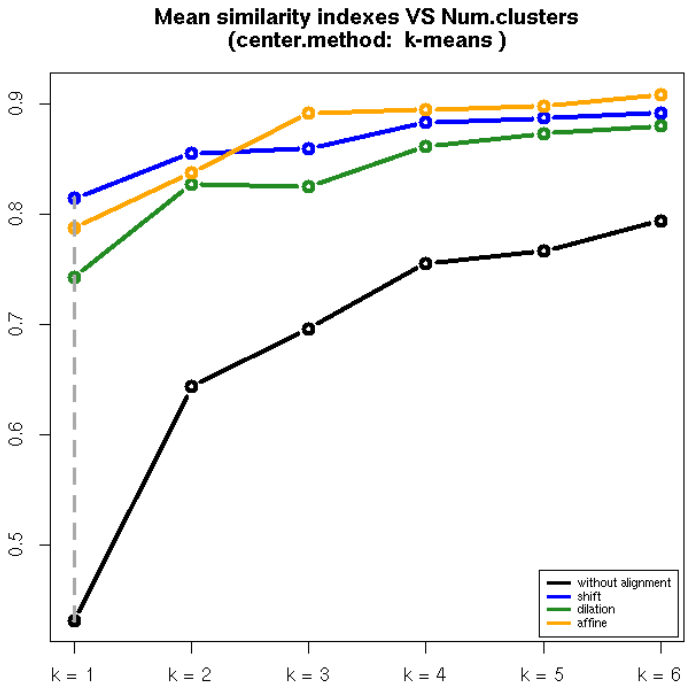


Figura 5.4: Analisi dei diversi tipi di classificazione al variare di $\mathcal{W}$ e del numero di cluster per le *spline* associate ai dati con indice di similarità basato sulle derivate prime.

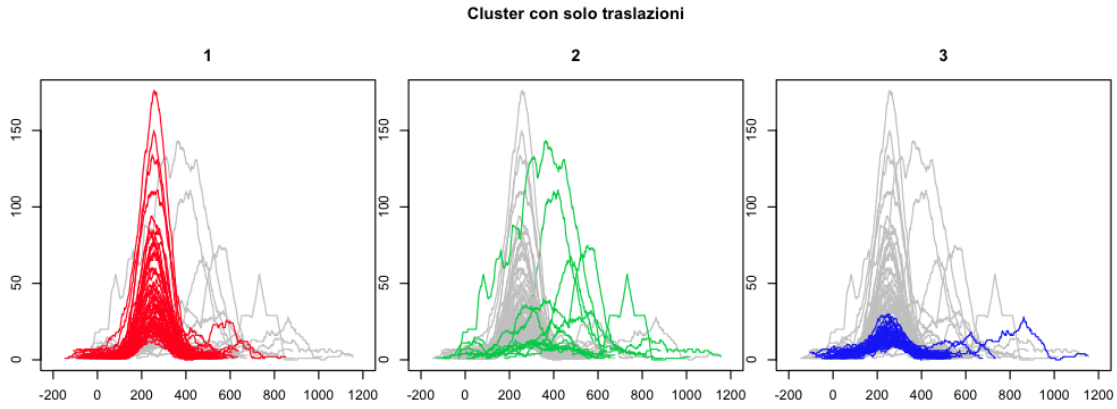


Figura 5.5: Classificazione in tre cluster con sole traslazioni.

allineamento in quanto non risulta interessante prevedere un raggruppamento dei dati indagando sulla variabilità in ampiezza impedendo traslazioni tra i picchi. Come giustificato successivamente, infatti, questi possono avere un andamento molto simile, ma che si manifesta soltanto in posizioni differenti. Lo scopo d'indagine risulta proprio l'analisi di queste somiglianze; impedendo l'introduzione di traslazioni queste similarità verrebbero erroneamente trascurate.

5.2.2 Analisi del *k-mean* nella classe delle traslazioni

Valutazione dei raggruppamenti identificati dall'indice di similarità ρ tra le curve

Come prima analisi si presentano i risultati ottenuti analizzando 100 picchi selezionati casualmente dal dataset distinti in 3 cluster; in Figura 5.5 sono rappresentati i picchi suddivisi nei tre cluster e allineati con i coefficienti di traslazione ottimi.

Al fine di interpretare i risultati ottenuti con l'analisi funzionale si può rappresentare la classificazione ottenuta in funzione degli indici di forma introdotti nel Capitolo 4 in particolare in Figura 5.6 si possono valutare i raggruppamenti in funzione delle 3 grandezze che risultano maggiormente significative per l'interpretazione del risultato, ovvero l'altezza, l'area e il coefficiente di forma M/altezza.

Valutando questi esiti, si osserva che il primo cluster è composto da picchi di dimensioni variabili, ma regolari (infatti sono caratterizzati da un indice M/altezza ridotto), mentre il secondo e il terzo cluster sono composti da picchi più frastagliati, ad esempio picchi bimodali. Il secondo cluster, in particolare, è composto da picchi molto frastagliati

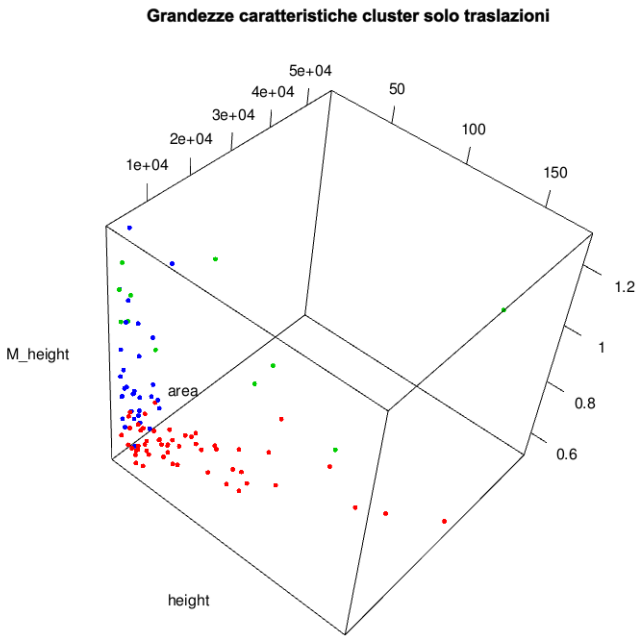


Figura 5.6: Rappresentazione dei cluster in funzione degli indici di forma.

(valori estremi di M/altezza), oppure da picchi molto ampi, elevata area o altezza. Il terzo, invece, è composto dai picchi più ridotti in area e altezza, con una regolarità intermedia tra quella dei due cluster precedenti.

Risulta particolarmente significativo il confronto con la classificazione basata sui soli indici introdotti nel Capitolo 4. Come proposto in Figura 5.8, si osserva che il primo cluster che si definisce in questa nuova analisi contiene picchi che appartenevano ai due raggruppamenti di picchi regolari introdotti dall'analisi precedente, che in questo caso uniti in un unico insieme, dato l'andamento funzionale simile con la regione di picco molto evidente. Altri picchi che nel caso multivariato appartenevano al terzo cluster, quello più numeroso e con picchi regolari, vengono invece accomunati, nell'attuale terzo cluster, a dati che precedentemente venivano considerati molto rumorosi. Questa unione è dettata dalla forma funzionale simile, in quanto questi dati hanno una ridotta altezza, ma comunque un andamento abbastanza regolare. Il secondo cluster, invece, racchiude, come si è già osservato, tutti i picchi complessivamente frastagliati e di dimensioni variabili.

Come ulteriore interpretazione da questo nuovo raggruppamento con allineamento,

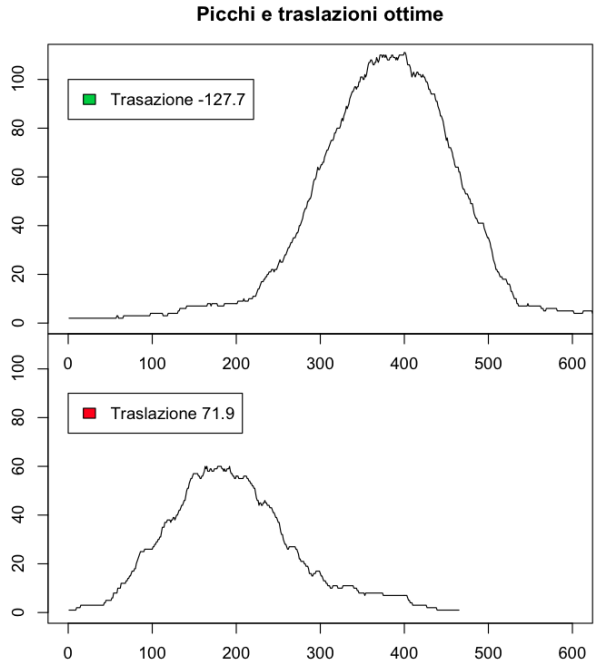


Figura 5.7: Confronto tra picchi con traslazioni ottime positive e negative.

inoltre, può risultare interessante analizzare le trasformazioni che hanno fatto corrispondere i differenti dati nei cluster. Il coefficiente di traslazione, infatti, potrebbe assumere un interessante significato biologico.
Sapendo che le regioni in analisi sono quelle associate all'interazione tra la proteina di interesse e la molecola di DNA, si ricorda che la conformazione di queste regioni di picco è ottenuta mediante un procedimento di sonicazione e poi di immunoprecipitazione. La sonicazione e l'immunoprecipitazione sono connesse, oltre che alla presenza del fattore di trascrizione, alla struttura della molecola di DNA. La selezione dei frammenti che vengono allineati al genoma, infatti, è influenzata dai ripiegamenti della cromatina: regioni molto compatte hanno una bassa probabilità di venire selezionate e isolate.

I frammenti isolati attorno alla regione di picco, che risulta essere quella determinante per l'allineamento e per la classificazione, denotano, dunque, la presenza di cromatina poco ripiegata, infatti questi frammenti sono comunque in numero superiore a quello atteso, essendo selezionati da MACS, ma potrebbero non essere strettamente connessi a regioni di legame della proteina di interesse.
Come si osserva dalla Figura 5.7, dunque, i picchi con valori di traslazione ottima

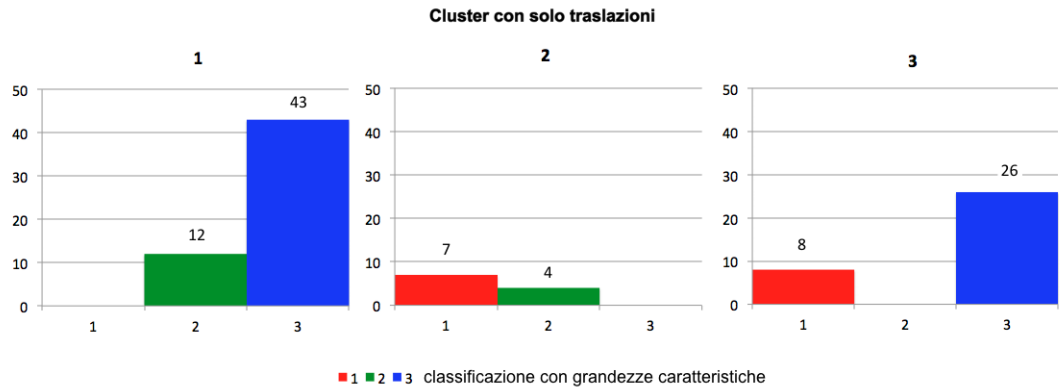


Figura 5.8: Composizione dei cluster delle traslazioni in funzione dei cluster multivariati.

per l'allineamento negativo sono preceduti da una regione di DNA forse caratterizzata da cromatina poco densa, mentre i picchi con traslazione negativa potrebbero essere preceduti da cromatina fitta e seguiti da regioni molto aperte. Qualora, invece, la traslazione si attesti su valori poco elevati, il picco potrebbe inserirsi in una struttura omogenea.

Questa analisi è lecita soprattutto per il primo cluster (i valori di traslazione ottima sono rappresentati in Figura 5.9), infatti è composto dai picchi meglio definiti, associati cioè con molta evidenza alla proteina in esame. Data la loro semplice struttura, inoltre, sono verosimilmente associati alla proteina senza nessuna interazione particolare con altri fattori di trascrizione o complessi proteici e pertanto l'informazione nella parte non specificatamente definita come picco potrebbe riguardare la struttura della molecola di DNA sottostante.

Questa interpretazione è, tuttavia, solo un'ipotesi qualitativa e risultano necessarie ulteriori valutazioni con opportuni strumenti di indagine sui ripiegamenti della cromatina, attraverso tecniche NGS che permettono di localizzare la posizione dei nucleosomi (MNase-Seq) e delle zone aperte della cromatina (DNase-Seq).

Valutazione dei raggruppamenti identificati dall'indice di similarità ρ tra le derivate

Analizzando, invece, i dati utilizzando come indice di similarità il coseno tra le derivate sempre considerando come insieme di *warping function* le traslazioni e, come si era

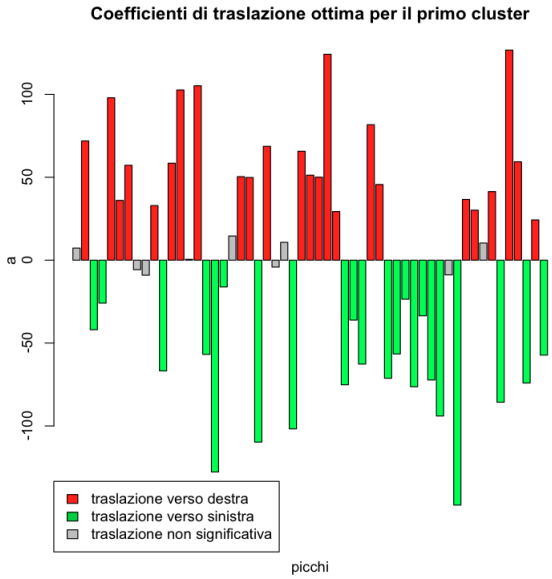


Figura 5.9: Valore dei coefficienti delle traslazioni ottime per il primo cluster.

definito in precedenza, un numero di raggruppamenti pari a quattro, si può valutare oltre alla rappresentazione dei cluster dopo l'allineamento (Figura 5.11) anche l'andamento delle derivate prime dei dati (Figura 5.12).

L'indagine condotta su questi dati risulta un raffinamento dell'analisi proposta nel caso precedente; dalla Figura 5.10, infatti, si nota che il primo e il quarto raggruppamento sono composti dai picchi con un andamento ben definito: per la maggior parte dati che nell'analisi precedente appartenevano al primo cluster, cioè quello dei picchi regolari. Si osserva che le derivate di questi dati hanno nel loro complesso un andamento molto regolare con una regione a derivata positiva seguita da una zona con derivata negativa, ovvero definiscono bene un'unica regione di picco.

La distinzione tra questi due raggruppamenti è legata alla pendenza della derivata, che per il quarto cluster è più accentuata, ad indicare la presenza di picchi più stretti e alti, mentre per il primo cluster è più ridotta, con picchi meno compatti. Questa distinzione, sebbene risulti effettivamente più approfondita della valutazione precedente, risulta forse trascurabile. É più che sufficiente, infatti, definire un solo raggruppamento di picchi regolari, mentre sarebbe più interessante dettagliare con maggiore accuratezza l'insieme dei dati anomali, che celano probabilmente dei comportamenti più difficili da interpretare, ma maggiormente interessanti dal punto di vista biologico.

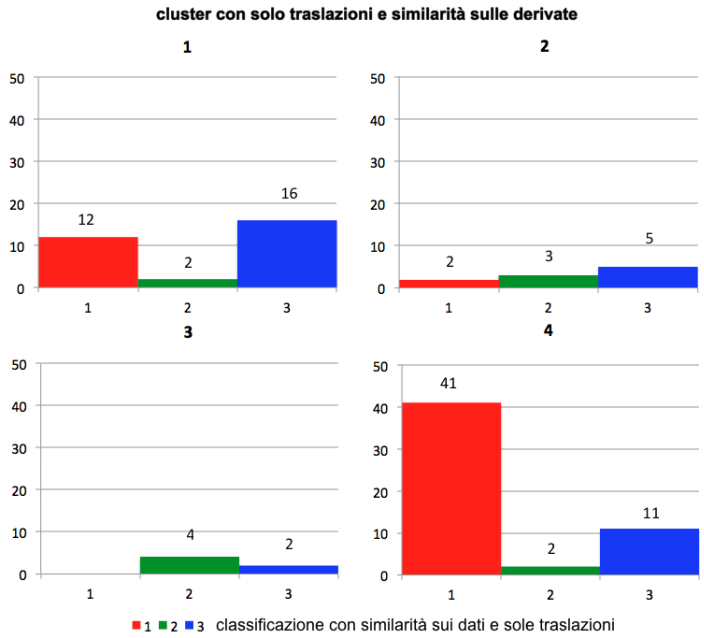


Figura 5.10: Composizione dei quattro cluster ottenuti dalla valutazione delle derivate in funzione dei raggruppamenti ottenuti dalla classificazione in base al coseno delle funzioni

Per questi due cluster si conferma comunque l'analisi dei coefficienti di traslazione ottima; i coefficienti di traslazione, nonostante la variazione del metodo di allineamento risultano avere valori analoghi rispetto ai dati rappresentati in Figura 5.9, anche se in questa analisi i dati vengono distinti in due raggruppamenti (Figura 5.13 in cui si evidenziano in giallo gli unici dati per cui si ha una variazione dell'indice da positivo a negativo rispetto ai risultati ottenuti con il *k-mean* con similarità sui dati).

Il secondo e il terzo raggruppamento sono costituiti principalmente da dati appartenenti all'insieme dei picchi irregolari precedentemente definiti, in cui si presenta una distinzione tra i dati caratterizzati dalla presenza di una regione di picco molto accentuata, preceduta da un'altra regione distinta di significativo aumento dei valori della *coverage function* (secondo cluster) e quei dati che invece rappresentano una regione di attività complessiva, ma che non definisce la canonica forma di picco (terzo cluster).

Pur rappresentando una buona classificazione, in alcuni aspetti più accurata di quella definita considerando la sola similarità tra i dati, questa soluzione presenta delle carenze: propone, ad esempio, dei raggruppamenti distinti per picchi ben definiti e non introduce ulteriori dettagli nella classificazione dei dati anomali.

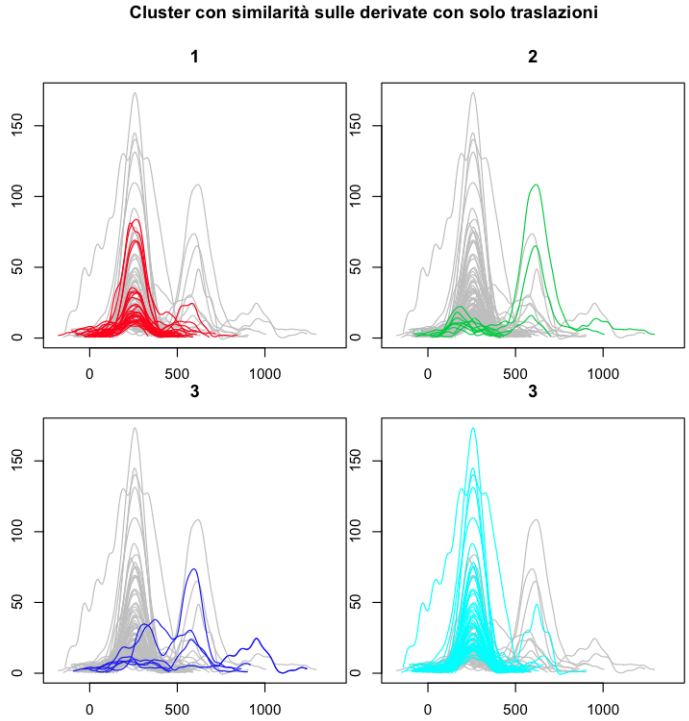


Figura 5.11: Classificazione delle *spline* in quattro cluster con traslazioni

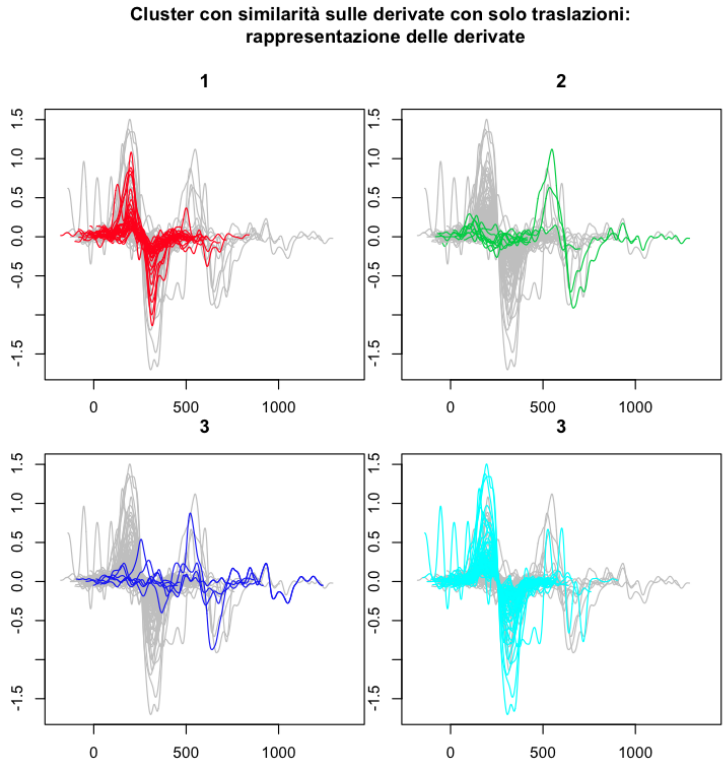


Figura 5.12: Andamento delle derivate nei quattro cluster

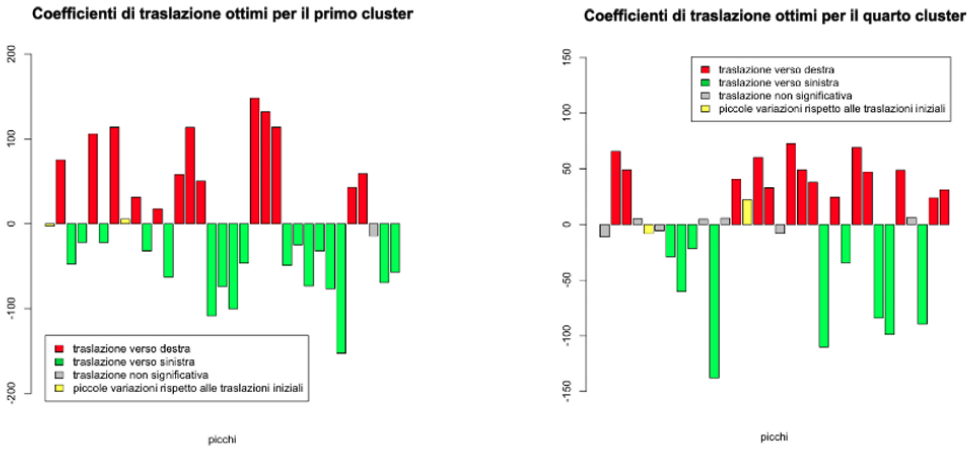


Figura 5.13: Coefficienti di traslazione per il primo e il quarto cluster.

Si deduce, quindi, la necessità di una nuova classificazione che risulti effettivamente un raffinamento di quella precedente, ma mirata maggiormente alla distinzione dei dati irregolari. L'utilizzo dell'indice di similarità basato sulle derivate non ha condotto al raffinamento auspicato, forse per la struttura dell'indice o per la necessaria introduzione dello *smoothing* e pertanto nelle successive analisi si stabilisce di focalizzarsi sull'insieme delle *warping function* delle affinità considerando la similarità direttamente in termini di coseno tra i dati.

5.2.3 Analisi del *k-mean* nella classe delle affinità

Valutazione dei raggruppamenti identificati dall'indice di similarità ρ tra le curve

Si analizza ora il caso più generale di *warping function* affini, composte, cioè da una componente traslatoria e da una di dilatazione; un possibile significato biologico del coefficiente di dilatazione verrà illustrato in sede di analisi dei risultati.

Come si è già osservato commentando la Figura 5.3, il numero di cluster da considerare per questo tipo di insieme $\mathcal{W}$ è quattro; i risultati ottenuti dalla classificazione e allineamento sono presentati in Figura 5.14.

Anche in questo caso si può presentare una rappresentazione dei raggruppamenti nello spazio delle grandezze caratteristiche precedentemente introdotte, tuttavia una migliore caratterizzazione dei cluster ottenuti è deducibile dall'analisi qualitativa degli

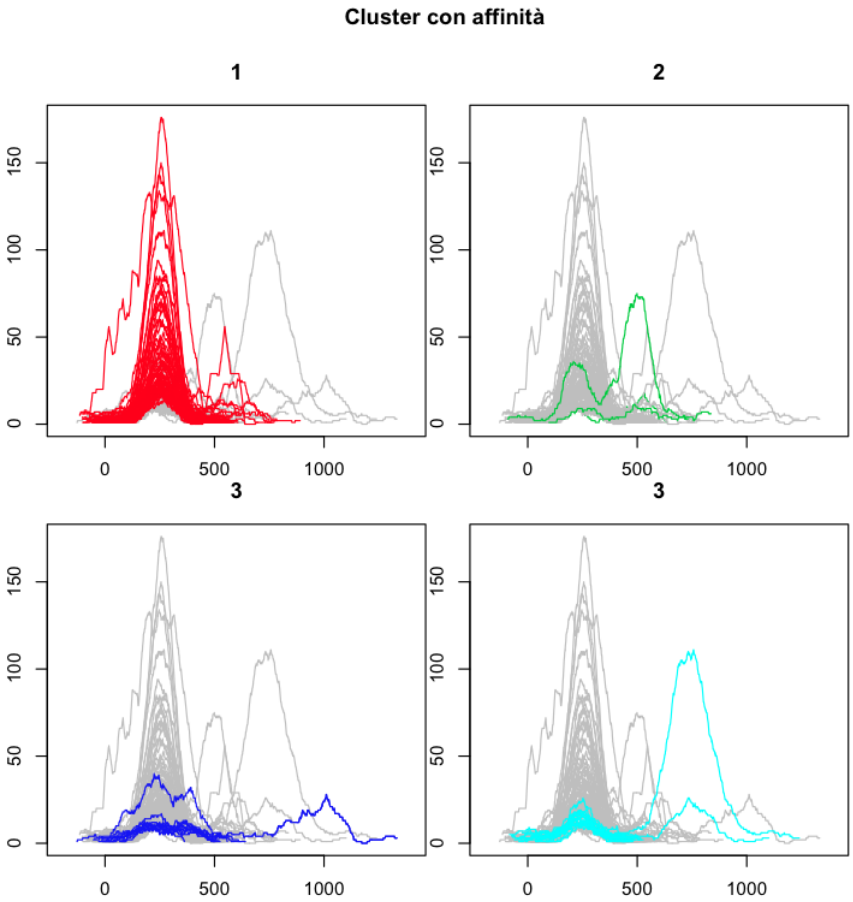


Figura 5.14: Classificazione in quattro cluster con trasformazioni affini.

insiemi (Figura 5.14) e dalla semplice analisi delle variazioni di collocazione attuale dei dati rispetto alle sole traslazioni ammesse sempre nell'ambito delle similarità tra funzioni (Figura 5.15).

Risulta interessante osservare come, ammettendo anche la componente di dilatazione in $\mathcal{W}$, l'insieme dei picchi ben caratterizzati, ovvero il primo cluster della classificazione con sole traslazioni non subisca stravolgimenti: tutti i picchi che ne facevano parte in prima analisi ne fanno parte anche modificando l'insieme delle *warping function*, mentre risultano inseriti in esso anche alcuni picchi che facevano parte dei cluster 2 e 3, ovvero quei picchi che risultavano troppo ampi per fare parte del primo cluster e che, però, a seguito dell'introduzione della dilatazione, si adattano bene alla forma canonica.
Per quanto riguarda gli altri tre cluster che si vanno a costituire, l'ultimo è formato in prevalenza da picchi del terzo cluster, quelli che avevano un coefficiente di complessità M/altezza intermedio; quelli più regolari, infatti, sono rientrati a fare parte del primo cluster, mentre questi sono andati ora (associati a due picchi che erano nel secondo cluster sempre con complessità intermedia) a costituire l'ultimo raggruppamento.
Il terzo, invece, è formato da picchi molto frastagliati e, tranne due, con altezze molto ridotte; ha infatti raccolto i picchi più ridotti e frastagliati che in precedenza erano parte del secondo e terzo cluster.
Il secondo raggruppamento, invece, è costituito da soli tre picchi, che non sembrano avere caratteristiche comuni, almeno tra quelle facilmente osservabili con un'analisi quantitativa, ma hanno solamente un comportamento anomalo rispetto agli altri tre raggruppamenti descritti.

Sarebbe, comunque, necessaria un'analisi più dettagliata di questi tre ultimi raggruppamenti per non limitarsi a una caratterizzazione solo qualitativa e comunque limitata dalla numerosità dei dati. Questa analisi verrà proposta in Sezione 5.2.4 e commentata a seguito dell'applicazione di questa tecnica di classificazione ad un insieme più ampio di dati.

Certamente più agevole risulta, invece, l'analisi del primo raggruppamento; come si è già sottolineato, questo raccoglie tutte le regioni caratterizzate dalla diretta interazione della proteina con la catena di DNA (i picchi risultano ben definiti e regolari).
L'osservazione interessante riguarda quei dati che dall'analisi con sole traslazioni ammesse non risultavano fare parte di questo raggruppamento, ma che, grazie alla possibilità di dilatazioni, nel caso specifico contrazioni, ora sono considerati come regolari.

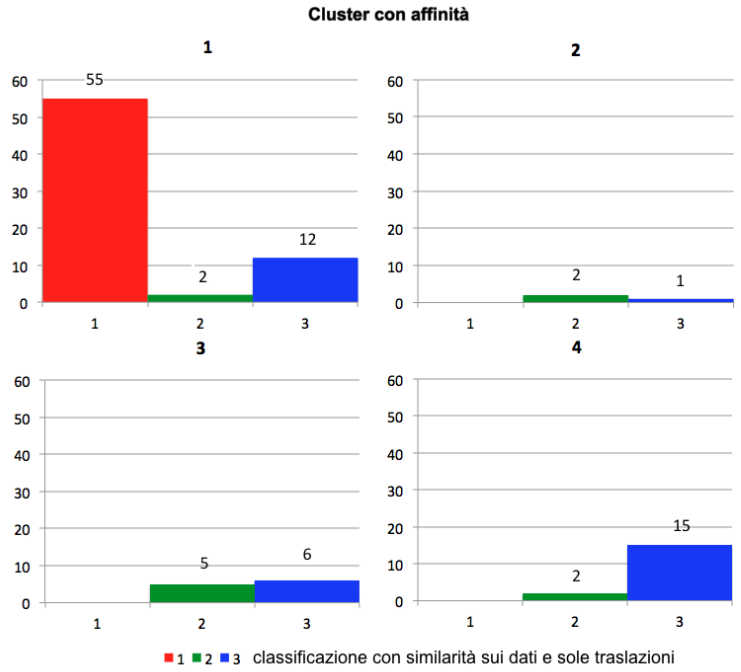


Figura 5.15: Composizione dei cluster delle affinità in funzione dei cluster delle sole traslazioni.

L'introduzione della possibilità di applicare una dilatazione con coefficiente minore di uno, infatti, permette di identificare come picchi associati alla semplice interazione proteina-sequenza anche quei dati che risultano più ampi, ma la cui forma non lascia intendere altre tipologie di legame o interazioni esterne.
Bisogna ricordare, infatti, che le indagini sui dati di tipo ChIP-Seq prevedono l'utilizzo di una moltitudine di cellule in un unico esperimento, cellule selezionate in modo da essere il più omogenee possibile, ma chepossono distinguersi per brevi distanze evolutive e dunque avere delle piccole differenze a livello delle interazioni epigenomiche. Queste differenze si potrebbero semplicemente manifestare in traslazioni di poche basi delle regioni di legame sulla sequenza; raccogliendo tutti i dati di questo insieme di cellule in un'unica *coverage function*, possono dunque emergere picchi più ampi causati soltanto dalla sovrapposizione di queste differenti regioni di interazione non perfettamente coincidenti.
Secondo questa ipotetica interpretazione biologica dei coefficienti di dilatazione, quindi, dalla semplice raccolta dei dati relativi al coefficiente di dilatazione ottimo si possono identificare queste regioni di differenziazione tra le diverse cellule nella stessa coltu-

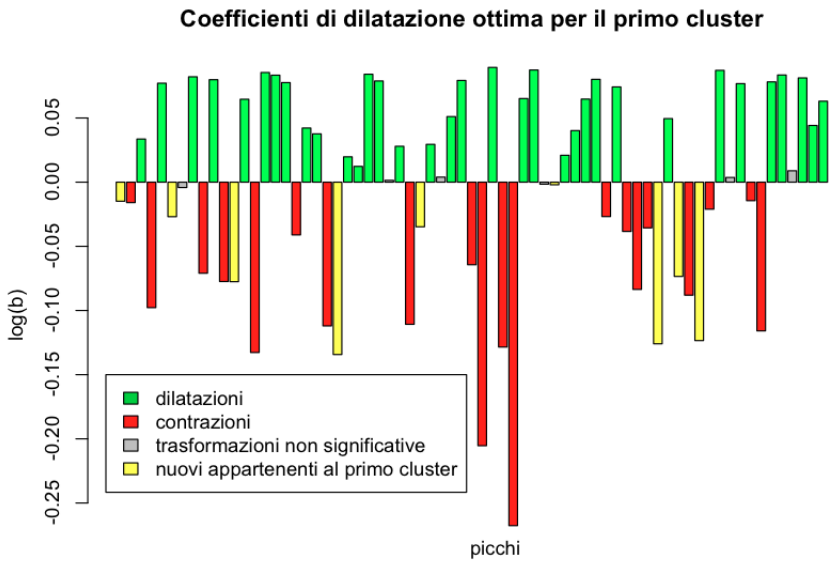


Figura 5.16: Valore dei coefficienti delle dilatazioni ottime per il primo cluster.

ra, ovvero le regioni associate a dati del cluster dei picchi regolari con coefficiente di dilatazione minore di uno (valori in rosso del grafico riassuntivo in Figura 5.9).

Per quanto concerne la valutazione con indice di similarità sulle derivate si ricorda che l'analisi è omessa in quanto l'introduzione della componente di dilatazione non porta a un significativo miglioramento della similarità complessiva e inoltre si è già concluso che l'indice non può comprendere tutte le caratteristiche dei dati, che sono meglio caratterizzati dalla valutazione diretta del coseno tra le funzioni.

5.2.4 Analisi complessiva dei dati

In questa sezione si analizzano i dati nel loro complesso, non concentrandosi unicamente su 100 picchi, ma ampliando l'insieme in esame.

In primo luogo si confermano i risultati introduttivi proposti ampliando l'insieme dei dati: su un campione di 1.000 picchi si conduce un'indagine complessiva dei metodi di classificazione funzionale proposti. Si presentano sia la valutazione dei risultati nel caso similarità calcolata direttamente sui dati con insieme di *warping function* le traslazioni, sia alcune considerazioni dettate dalla generalizzazione dell'insieme delle

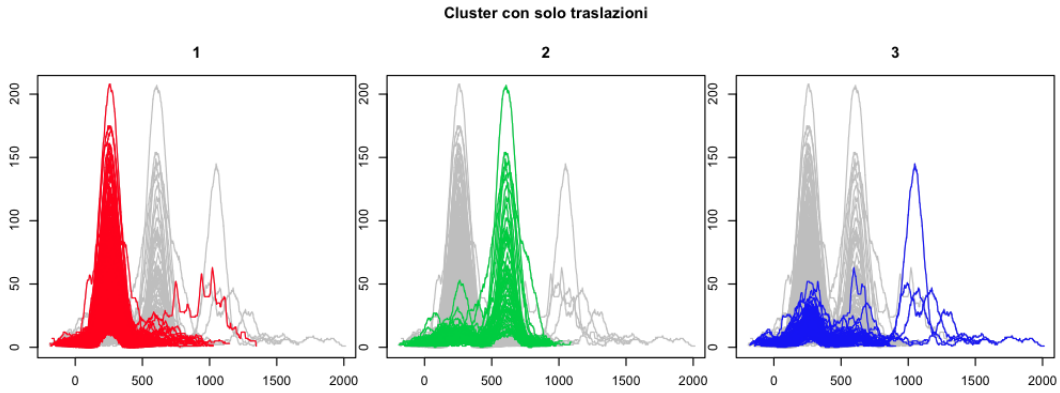


Figura 5.17: Classificazione in tre cluster con sole traslazioni per un campionamento composto da 1.000 picchi scelti casualmente.

warping function con l'introduzione della componente di dilatazione (trasformazioni affini).

A seguito della conferma della bontà dell'analisi introdotta in precedenza, come risultato definitivo si propone per un vasto insieme di dati l'analisi compiuta con l'algoritmo del *k-mean alignment* con *warping function* scelte nella classe delle trasformazioni affini, indice di similarità il coseno tra le curve e numero di raggruppamenti pari a quattro. A tale scopo si utilizza un algoritmo basato sul principio della *bootstrap aggregation* con raggruppamenti costituti da 1.000 campioni scelti casualmente tra i 13.182 disposizione e procedendo con un'analisi di dieci campionamenti.

Analisi delle classificazioni funzionali per un campione di 1.000 dati

In questa prima sezione si confermano i risultati preliminari introdotti a seguito dell'ampliamento dell'insieme in esame. Valutando i raggruppamenti ottenuti dal *k-mean alignment* con le sole traslazioni (Figura 5.17) è confermata la precedente distinzione, con la definizione di un cluster composto da dati regolari (primo cluster in figura), un secondo con dati frastagliati, ma ampi e l'ultimo composto da picchi non ben definiti.

La generalizzazione nel caso di trasformazioni affini porta a confermare il ruolo del coefficiente di dilatazione; valutando, infatti, la composizione dei raggruppamenti ottenuti (Figura 5.19) e la loro struttura (Figura 5.18) si osserva che:

- il raggruppamento azzurro, composto da dati che nell'analisi precedente erano considerati picchi frastagliati o non ben definiti, racchiude quei dati che hanno

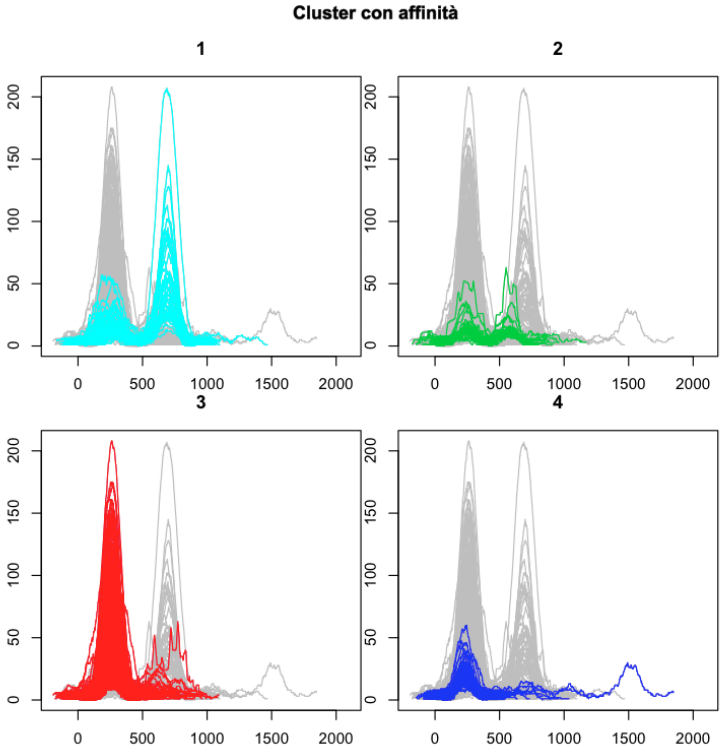


Figura 5.18: Classificazione per il campionamento composto da 1.000 picchi.

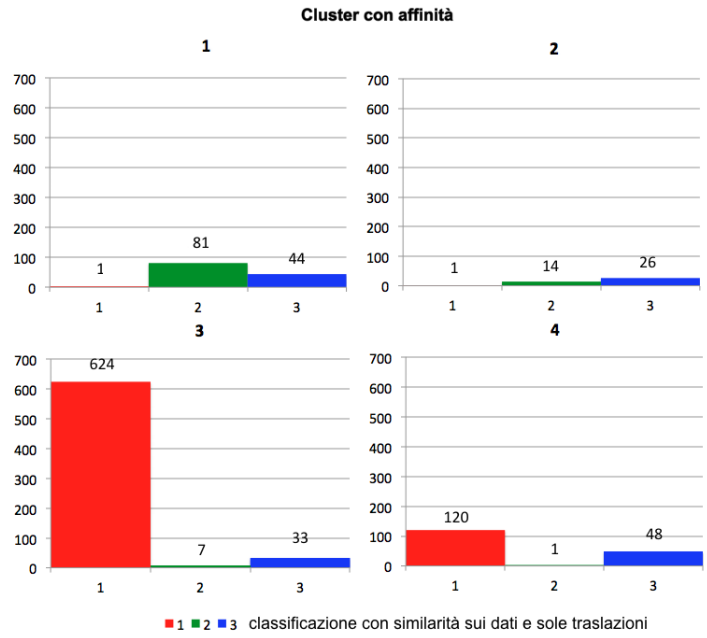


Figura 5.19: Composizione dei cluster delle affinità in funzione dei cluster delle sole traslazioni per un campionamento di 1.000 picchi.

una forma a picco abbastanza definita, ma circondata da un'ampia regione ad alta attività, regione che non permette la loro classificazione come picchi regolari;

- il cluster verde comprende dati che non hanno una forma ben definita; raccoglie, infatti quei dati dei due precedenti raggruppamenti irregolari che hanno una dimensione ridotta e una struttura frastagliata;

- il più numeroso raggruppamento (rosso), composto dalla maggior parte dai dati già assunti come regolari, contiene i dati con la struttura di picco ideale. A quelli appartenenti al primo cluster precedente si aggiungono quei dati che attraverso l'introduzione del termine di dilatazione (nel caso specifico contrazione) vanno ad assumere la forma ideale;

- nel cluster blu sono, invece, raccolti tutti quei dati che hanno una struttura abbastanza regolare, ma un'altezza troppo ridotta per essere introdotti nel raggruppamento dei picchi ideali.

A seguito di queste considerazioni si può affermare che l'analisi proposta in precedenza risulta effettivamente accettabile e robusta anche a seguito dell'aumento della numerosità dei dati in esame. Si procede, quindi, all'analisi dell'insieme più ampio dei dati con la procedura della *bootstrap aggregation*.

Bootstrap aggregation per la classificazione di un campione più ampio

La suddivisione proposta per un campionamento esemplificativo composto da 1.000 dati risulta robusta anche al variare dei dati considerati. Si procede, infatti, a un'analisi di dieci campionamenti costituiti dalla scelta casuale di 1.000 picchi e si ottengono classificazioni qualitativamente analoghe a quelle proposte nel caso dall'unico insieme introdotto in precedenza.
In Figura 5.20 si presentano alcuni dati scelti casualmente tra i cluster che si definiscono per i dieci campionamenti. Nell'immagine si presentano i picchi già suddivisi in maniera omogenea tra i differenti campionamenti (e coerente con la caratterizzazione del campionamento esemplificativo di Sezione 5.2.4) in modo che i dati rossi rappresentino i cluster di picchi ideali, quelli verdi gli insiemi di dati privi di una significativa caratterizzazione di forma, quelli blu identifichino i raggruppamenti di picchi regolari, ma con un'altezza ridotta e, infine, quelli azzurri con una regione di picco associata da una regione di significativo scostamento della *coverage function* dal *background*.

L'identificazione di raggruppamenti omogenei tra i differenti campionamenti è avvalorata dall'analisi qualitativa dei dati e dalla valutazione della similarità tra gli andamenti medi dei dati appartenenti a ciascun cluster (Figura 5.21). La similarità complessiva tra tutti i rappresentati dei quattro cluster, calcolata in termini di coseno tra gli andamenti medi, è riportata in Tabella 5.1. Si osserva che questo valore, elevato per tutti i cluster, risulta meno significativo per l'ultimo cluster: infatti i dati appartenenti a questo raggruppamento hanno un andamento variabile, con la caratteristica struttura a picco circondata da regioni differentemente localizzate di attività (in alcuni casi picchi bimodali), per cui questo risultato è abbastanza naturale.

cluster	1	2	3	4
	0.9987	0.9600	0.9385	0.8652

Tabella 5.1: Similarità complessiva tra i cluster dei diversi campionamenti.

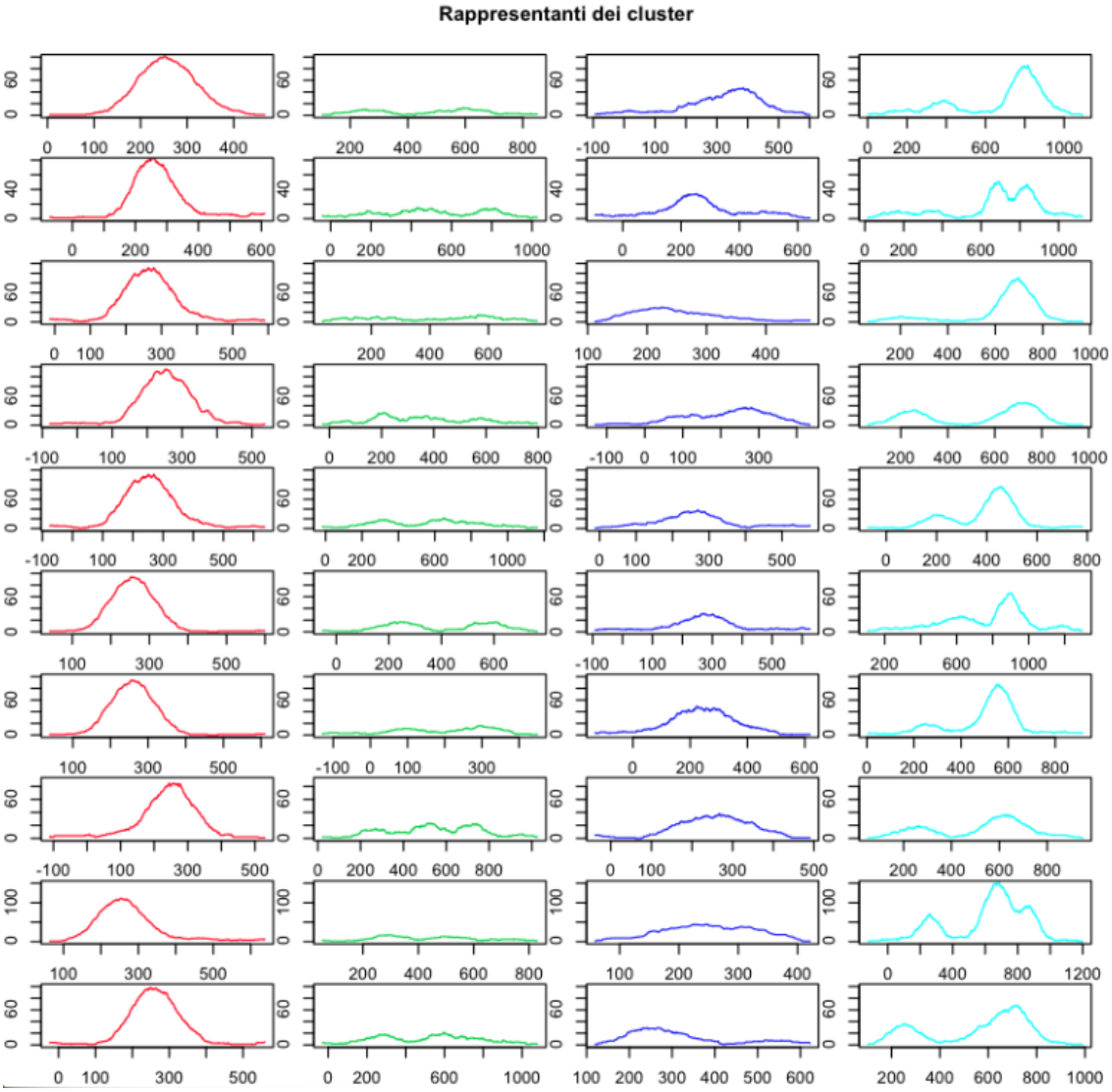


Figura 5.20: Rappresentanti dei cluster per i 10 raggruppamenti.

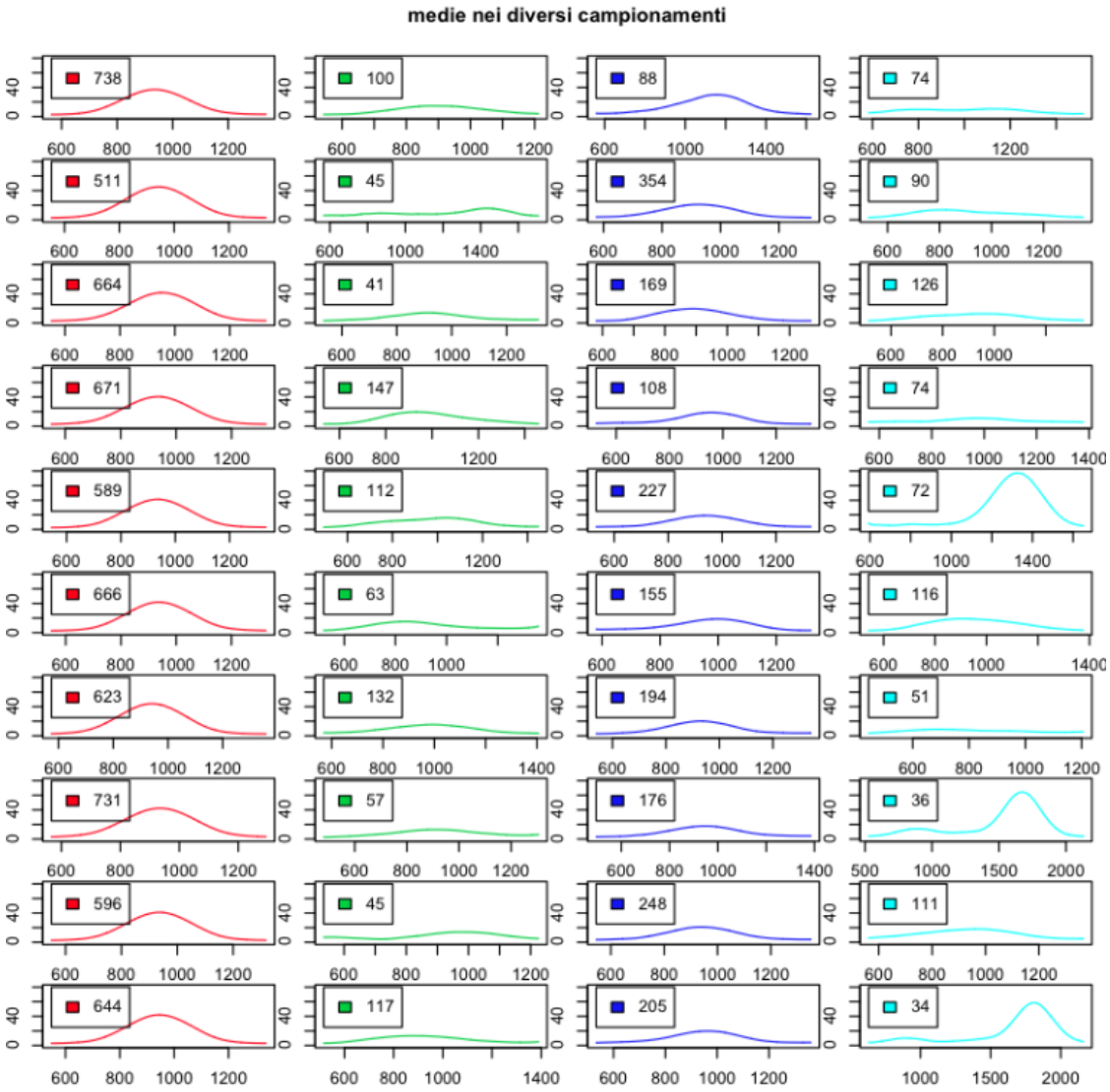


Figura 5.21: Andamento della media dei dati dei cluster nei differenti campionamenti con indicazione delle numerosità di ciascun cluster.

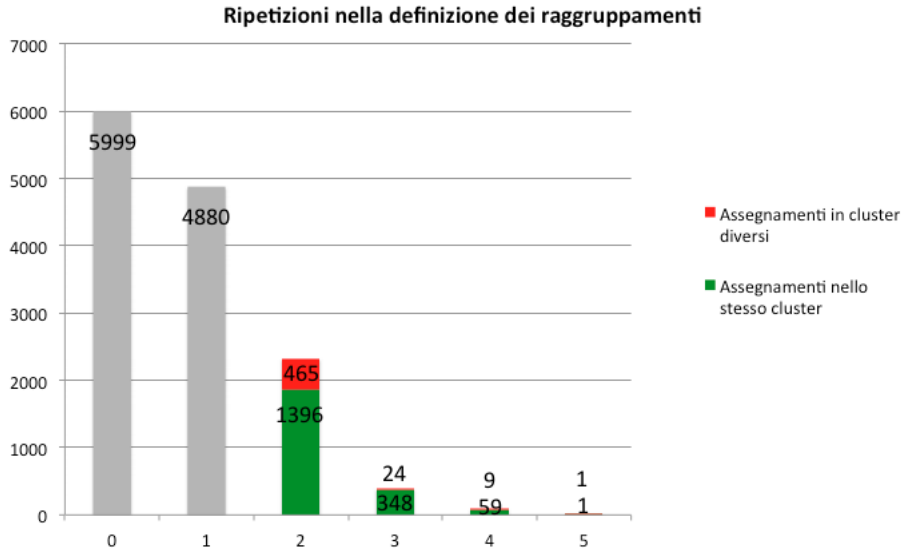


Figura 5.22: Analisi dei dati presenti in più raggruppamenti.

Analizzando nel dettaglio gli esiti di questa analisi, si osserva che il metodo di indagine mediante *bootstrap aggregation* prevede una scelta casuale per ogni campionamento dei dati da considerare e pertanto è possibile che alcuni di questi vengano considerati più volte in diversi raggruppamenti. Nel caso della simulazione in esame, che coinvolge nel complesso 7.138 dati, si osserva che 2.303 sono considerati in più di un campionamento; in Figura 5.22 sono riportati i dati relativi al numero di campioni considerati in più raggruppamenti e si osserva che meno del 30% di essi è classificata diversamente nei diversi insiemi in cui è presente. Questo risultato, pur non ottimale, è in gran parte associato a errori nella classificazione del terzo cluster (blu), che è costituito da dati che possono facilmente essere confusi con dati prevalentemente rumorosi (secondo cluster) o dati con regioni di attività poco definite (quarto cluster).

Analizzando i risultati complessivi dell'algoritmo di classificazione si definisce una suddivisione per i 7.183 picchi considerati in almeno uno dei dieci campionamenti. Oltre ai picchi assegnati a un unico raggruppamento e pertanto definiti come appartenenti al cluster corrispondente, si associano a uno specifico cluster tutti quei dati che vengono considerati in più campionamenti ma che corrispondono allo stesso cluster, e quei dati che pur essendo associati a cluster diversi hanno evidenza per appartenere a uno specifico. Si

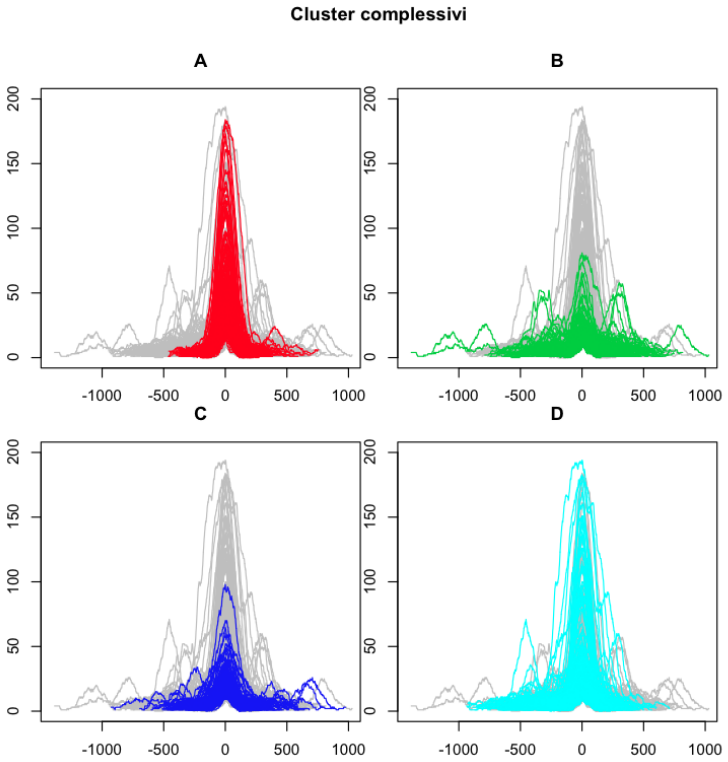


Figura 5.23: Rappresentazione di 200 picchi esemplificativi per ogni cluster complessivo

permette, infatti per ogni dato un errore di classificazione: per i dati in tre campionamenti, ad esempio, si sceglie di associare il picco ad un cluster qualora almeno due assegnamenti risultino concordi, e così per quattro raggruppamenti è sufficiente che tre siano uguali. Con queste assunzioni, a partire dai dati riportati nell'istogramma di Figura 5.22 si trova che dei 7.183 meno del 7% non può essere assegnato a un cluster, di cui la maggior parte dei quali è costituita dai picchi considerati in due soli campionamenti con rappresentazioni diverse, mentre molto interessante è il dato riguardante i dati considerati in tre raggruppamenti per i quali solo il 6.7% non riesce a essere classificato.

Si delineano, dunque, le suddivisioni presentate in Figura 5.23, in cui si rispecchia la suddivisione precedentemente proposta per un solo raggruppamento: i picchi appartenenti al primo cluster ben definiti e compatti, mentre quelli dell'ultimo circondati da una regione significativamente diversa dal controllo. Il secondo e il terzo cluster, infine, sono associati a andamenti meno accentuati e meno regolari (soprattutto per il secondo raggruppamento è effettivamente difficile definire una regione di picco vera e propria).

Capitolo 6

Interpretazioni biologiche

6.1 Validazione dei cluster

In questa sezione si analizzano i cluster definiti in conclusione del Capitolo 5. Questi, come si è dettagliatamente descritto, sono ottenuti al termine del percorso di raffinamento di analisi a partire dalla definizione dei raggruppamenti basati sugli indici di forma, fino alla definizione dei parametri ottimi per l'algoritmo del *k-mean alignment*.
Si ricorda che sono risultati quattro cluster:

- il primo costituito dai dati regolari con la forma del picco molto ben definita;
- il secondo che racchiude dati con un'altezza ridotta e senza alcuna forma caratteristica;
- il terzo che comprende dati con una forma a picco non ben definita, spesso di dimensione ridotta, ma pur sempre identificabile;
- il quarto con dati con una forma a picco definita, ma non chaiamente delimitata, con una significativa regione limitrofa caratterizzata da un'elevata attività.

Avendo presentato questa distinzione in termini di forma, si ricercano alcune possibili implicazioni biologiche di questa divisione.
Come prima osservazione si ricorda che i dati riguardano l'interazione proteina-DNA ed in particolare il dataset in esame riguarda il fattore di trascrizione GATA1; come si è introdotto in Sezione 1.4 questa proteina generalmente si lega al DNA in regioni definite

dal motivo GATA (Guanina-Adenina-Timina-Adenina). Con il software MEME-ChIP [24], [26] si ricercano, dunque, gli eventuali motivi sottesi dai picchi suddivisi nei cluster precedentemente introdotti. I risultati per i quattro raggruppamenti sono presentati rispettivamente nelle Figure 6.1, 6.2, 6.3, 6.4. In queste immagini si presenta il *logo* associato al motivo rinvenuto e il corrispondente valore dell'indice di significatività E; questo indice risulta minore quanto più la significatività del motivo è elevata.

Valutando l'esito di questa ricerca si osserva che possono essere proposte alcune interessanti osservazioni sui differenti cluster:

- in merito al primo cluster si osserva che esiste forte evidenza per affermare che vi è realmente un motivo sotteso e pertanto le regioni evidenziate dai picchi regolari celano effettivamente il legame diretto della proteina GATA1 con il DNA;

- per quanto concerne il secondo raggruppamento si osserva che non viene rilevato il motivo caratteristico di GATA1, tuttavia esiste una discreta evidenza per affermare che la regione è associata a fattori di trascrizione, infatti si rileva il motivo proposto in Figura 6.2 che è associato ai fattori di trascrizione. La presenza di questa connessione potrebbe indicare che queste zone sono state erroneamente considerate arricchite da MACS oppure che la proteina GATA1 interagisce con il DNA in modo indiretto, connettendosi a fattori di trascrizione che sono poi a diretto contatto con il DNA;

- in merito al terzo raggruppamento si osserva che vengono rilevati due motivi, uno associato alla proteina specifica GATA1 e un secondo connesso ai fattori di trascrizione; il primo motivo, inoltre, è identificato con un indice di significatività inferiore al caso del primo cluster. Questi due elementi confermano che i dati appartenenti al terzo raggruppamento hanno una forma che è probabilmente connessa alla proteina in esame, ma non così marcata da affermare con sufficiente certezza che il legame sia diretto e non associato ad altri fattori;

- il quarto raggruppamento, infine, è composto da picchi di forma meno regolare, anche se comunque con altezze tali da poter affermare che esiste il legame con la proteina in analisi. Questa ipotesi è confermata dalla ricerca dei motivi sottesi, che determina GATA, anche se con livello di significatività ridotto.

Figura 6.1: Risultato della ricerca dei motivi per il primo cluster.

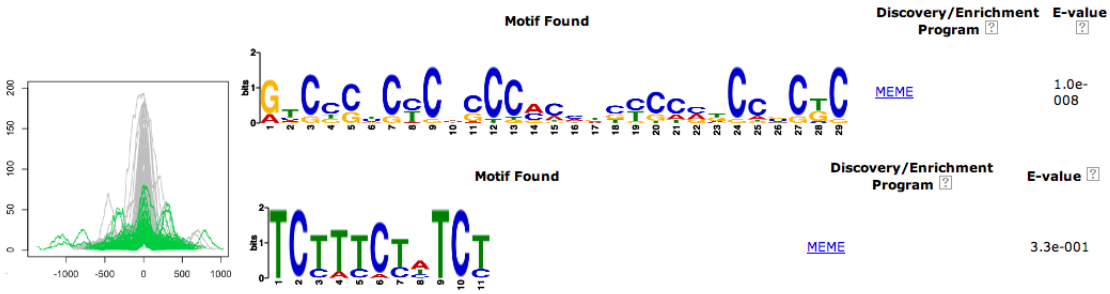


Figura 6.2: Risultato della ricerca dei motivi per il secondo cluster.

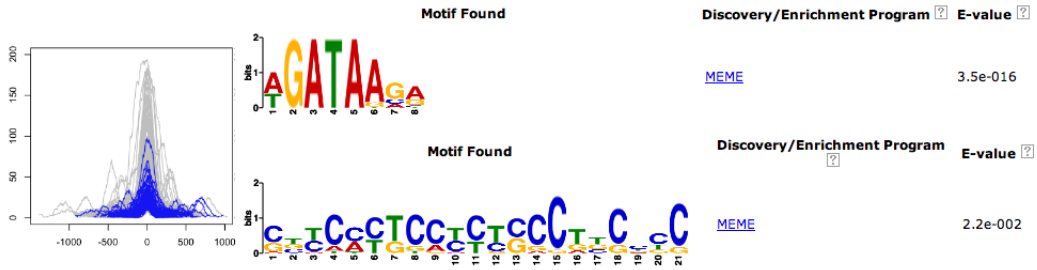


Figura 6.3: Risultato della ricerca dei motivi per il terzo cluster.

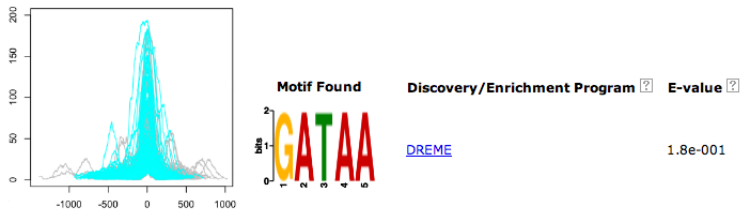


Figura 6.4: Risultato della ricerca dei motivi per il quarto cluster.

Si può concludere, dunque, che i raggruppamenti proposti in questo lavoro, basati sulla sola analisi statistica della forma dei dati risultano coerenti e portano a interessanti interpretazioni biologiche.

6.2 Sviluppi futuri

In questa sezione si analizzano le ulteriori analisi che possono essere condotte per approfondire i risultati presentati.

Le tecniche proposte in questo lavoro sono state confrontate tra loro evidenziando nel dettaglio i vantaggi di ciascuna nei confronti delle altre e hanno presentato nel complesso risultati coerenti e avvalorati da buone considerazioni biologiche. Come ogni procedimento di indagine statistica, tuttavia, questa classificazione necessita di ulteriori conferme; è necessario, infatti, dimostrarne la robustezza in termini di repliche dell'esperimento.
Ci si propone, dunque, di analizzare i dati ottenuti con una replica tecnica di questo esperimento, ovvero da una analoga analisi di ChIP-Seq sullo stesso campione, e quelli ottenuti come replica biologica, ovvero dal sequenziamento di un campione differente, ma costituito da cellule con le medesime caratteristiche. La necessità di questa operazione, resa possibile dalla varietà di dati ENCODE, è legata all'esigenza di confermare le conclusioni proposte. Si ipotizza, infatti, che la suddivisione in raggruppamenti e la loro caratterizzazione sia effettivamente connessa all'interazione proteina-DNA e pertanto si auspica che questa classificazione non subisca variazioni nel caso di repliche sia biologiche che tecniche. Ci si aspetta, inoltre, che le regioni selezionate nei diversi esprimenti vadano a costituire gli stessi raggruppamenti, ovvero ad esempio che le regioni associate a picchi regolari, che non lasciano presagire nessun particolare legame complesso tra proteine, restino tali anche nei successivi esperimenti.

Parallelamente a questi spunti di riflessione connessi alla stessa proteina considerata nella Tesi, risulta certamente interessante l'analisi per dati inerenti ad altri fattori di trascrizione, o ad altre proteine per esempio di tipo istonico. A tale proposito si potrebbero riproporre tutte queste analisi e valutazioni, per verificare la reale applicabilità di questa tecnica ai generici dati di tipo ChIP-Seq.

Da ultimo, dopo aver confermato l'affidabilità del metodo di classificazione sia mediante conferma degli esiti con le repliche, sia mediante la validazione con altre

tipologie di proteine, potrebbe essere interessante valutare se questa analisi permette di discernere le variazioni del legame proteina-DNA tra popolazioni differenti di cellule, ad esempio cellule con diverse manifestazioni fenotipiche o cellule sane e patologiche. Si vuole, ad esempio, proporre la stessa classificazione per dati derivanti da esperimenti su cellule patologiche che quindi presentano delle manifestazioni geniche diverse al fine di determinare le regioni del DNA con alterazioni epigenetiche. Nel caso del fattore di trascrizione considerato in questo lavoro (GATA1) si possono, ad esempio considerare cellule patologiche connesse a un disordine mieloproliferativo transitorio, come la leucemia melodie cronica. Per queste patologie, infatti, è stata ipotizzata [25] una connessione con un'alterazione nella proteina GATA1 e quindi una variazione di espressione genica. Qualora si riscontrassero effettive variazioni di forma dei legami proteina-DNA, si potrebbe verificare realmente questa connessione e quindi sviluppare ulteriori metodologie di indagine.

Appendice A

Valutazione della qualità dei frammenti

In questa sezione si analizzano nel dettaglio le caratteristiche del dataset considerate nella valutazione complessiva della qualità.
Come già introdotto nel Capitolo 2, il programma FastQC permette una analisi complessiva delle caratteristiche dei *reads* letti, sia valutando la qualità di lettura delle singole basi, sia le caratteristiche globali delle sequenza lette.

In Tabella A.1 si precisano i parametri considerati e si propongono gli standard utilizzati del software per la valutazione automatica del dataset; in particolare si indicano i casi in cui viene presentata l'indicazione di parametro alterato  o di parametro molto distante dal comportamento ottimale  .

Sezione	Descrizione	Warning	Errore
Statistiche introduttive	Elenco delle caratteristiche del file analizzato: • nome del file; • tipologia di file; • tipologia di codice per il controllo qualità; • numero totale di sequenze; • lunghezza delle sequenze (indicata la lunghezza della minore e della maggiore nel caso di sequenze differenti); • percentuale complessiva di basi C e G nelle sequenze.	Mai.	Mai.
Qualità per ogni base della sequenza	Rappresentazione attraverso boxplot dei valori di qualità di ciascuna base nel file. Per ogni posizione della sequenza si valutano i livelli di qualità e si rappresentano i relativi boxplot e la media (linea blu); sono indicati, inoltre, i livelli ottimi (regioni a fondo verde), accettabili (arancione) e scarsi (rosso), secondo gli standard associati a ciascun criterio di valutazione. Si sottolinea come, in generale, la qualità degradi dall'inizio alla fine dei *reads* a causa del meccanismo di lettura.	Se esiste un estremo inferiore di un boxplot (primo quantile) minore di 10 o una mediana (livello centrale del boxplot) minore di 25.	Se esiste un estremo inferiore di un boxplot (primo quantile) minore di 5 o una mediana (livello centrale del boxplot) minore di 20.

Qualità per sequenze	Il grafico indica per ogni qualità il numero di sequenze aventi qualità media pari a quella indicata in ascissa e permette, pertanto, di valutare se un sottoinsieme significativo delle sequenze ha qualità troppo ridotta, errore causato magari da un problema sistematico nella misurazione.	Se il valore più frequentemente osservato è inferiore a 27.	Se il valore più frequentemente osservato è inferiore a 20.
Contenuto di nucleotidi nelle basi	Rappresentazione delle proporzioni di ogni nucleotide nelle diverse posizioni delle sequenze; in un insieme casuale di sequenze ci si aspetta che non ci siano significative differenze tra le diverse basi e che, quindi, le linee nel grafico siano parallele. La percentuale costante di ciascuna base, poi, deve riflettere le proporzione di basi nell'intero genoma. Se ci sono forti cambiamenti probabilmente esiste una contaminazione esterna che danneggia i risultati dell'esperimento. Proprio per l'uguaglianza delle proporzioni nel genoma si osserva che nei singoli frammenti deve esistere un legame tra le basi accoppiate: A,T e C,G devono essere presenti in proporzioni uguali.	Se esiste almeno una base in cui la differenza tra A e T o G e C è maggiore del 10%.	Se esiste almeno una base in cui la differenza tra A e T o G e C è maggiore del 20%.
Contenuto di C e G nelle basi	Rappresentazione della proporzione complessiva di C e G nelle diverse basi. Anche in questo	Se esiste almeno una base la cui proporzione si	Se esiste almeno una base la cui proporzione si

	caso si deve mantenere costante lungo tutta la lunghezza delle sequenze. Se si osservano variazioni significative per un ridotto insieme di basi si può ipotizzare che vi sia una sequenza sovrarappresentata nei frammenti selezionati, o un semplice errore nella lettura.	discosta di oltre il 5% dal valore medio del contenuto di basi C e G.	discosta di oltre il 10% dal valore medio del contenuto di basi C e G.
Contenuto di C e G per ogni sequenza	Valutazione della proporzione di basi C e G nelle sequenze, comparata con la distribuzione ipotizzata; come distribuzione ipotizzata si sceglie una distribuzione normale con media e mediana pari alla mediana empirica poiché non è nota la proporzione effettiva di basi C e G.	Se la somma delle deviazioni dall'andamento ipotizzato supera il 15% dei *reads*.	Se la somma delle deviazioni dall'andamento ipotizzato supera il 30% dei *reads*.
Contenuto di N nelle basi	Nel caso in cui non sia possibile in fase di lettura dell'immagine definire con sufficiente confidenza il nucleotide, questo viene indicato con N, evenienza non inusuale soprattutto alla fine della sequenza; si contano, dunque, per ogni base il numero di N presenti. Nel caso in cui si manifesti un numero eccessivo di N nelle sequenze, probabilmente il metodo di lettura non è adatto.	Se in almeno una base la percentuale supera il 5%.	Se in almeno una base la percentuale supera il 20%.
Distribuzione delle lunghezze delle	Alcuni programmi di lettura generano sequenze di lunghezza costante, altri invece variabile; è	Se i frammenti non hanno tutti la stessa lunghezza.	Se esistono sequenze di lunghezza nulla.

sequenze	necessario, quindi, valutare l'andamento delle lunghezze dei frammenti; si visualizza, quindi, per ogni intervallo di valori di lunghezza il numero di sequenze corrispondenti.		
Sequenze duplicate	La presenza di frammenti uguali potrebbe indicare l'ottima selezione delle sequenze simili a un certo obiettivo, ma più verosimilmente l'uguaglianza è dovuta a distorsioni sistematiche nel processo di duplicazione artificiale mediante PCR per la lettura. Il grafico rappresenta il numero relativo di sequenze presenti in copie. Il livello associato a 10 copie racchiude in sè la somma di tutti i livelli associati a un numero di copie maggiori o uguali a 10, pertanto è usuale assistere a lievi aumenti nel valore delle frequenze relative, anche se fisicamente questa quantità dovrebbe essere decrescente. Se il valore associato a 10 è molto alto significa che un gran numero di sequenze è associato a un alto livello di duplicazione, risultato che deve essere considerato negativamente.	Se le sequenze non uniche rappresentano più del 20 % del totale.	Se le sequenze non uniche rappresentano più del 50 % del totale.
Sequenze sovraespresse	La presenza di una sequenza sovraespressa in un insieme può significare che è molto importante	Se esiste almeno una sequenza nell'elenco di	Se esistono sequenze presenti in più dell'1% del

	dal punto di vista biologico, ma più verosimilmente che il dataset è contaminato. Si presenta, quindi, una lista di sequenze che sono uguali in più del 0.1% del totale del dataset e per ognuna di esse viene controllato in un opportuno database di sequenze contaminate quali possono essere le cause della savraespressione.	quelle presenti in più dello 0.1 % del totale.	totale.
***k*-meri sovraespressi**	L'analisi della sola sovrarappresentazione delle sequenze può non mettere in luce tutti problemi di sovrarappresentazione, infatti alcune sequenze possono risultare diverse solo a causa di errori puntuali nella lettura, e quindi non essere identificate come uguali; è necessario valutare, dunque, la presenza di raggruppamenti di k elementi uguali (k-meri) con k generalmente pari a 5. Per ogni raggruppamento di k elementi, quindi, si determina un valore atteso per il livello del raggruppamento nel dataset complessivo e in ogni posizione della base.	Se esiste almeno un k-mero con rapporto tra presenza reale e attesa complessivamente superiore a 3 o in uno specifico punto superiore a 5. In questo caso viene anche presentato un grafico per i k-meri critici in modo che si possa localizzare dove sono maggiormente concentrati e quindi le regioni mal lette.	Se esiste almeno un k-mero presente con rapporto tra il valore reale e atteso in più di una base pari ad almeno 10. Anche in questo caso viene presentato l'andamento della proporzione dei k-meri critici.

Tabella A.1: Valutazione della qualità dei *reads* con FastQC.

Appendice B

Presentazione di algoritmi

In questa sezione si presentano nel dettaglio gli algoritmi introdotti nei capitoli precedenti, in particolare nella Sezione B.1 si introduce l'algoritmo dell'allineamento locale BWA, mentre nella Sezione B.2 si presenta l'algoritmo per il calcolo del coefficiente di forma M.

B.1 Algoritmo di allineamento BWA

Si ricorda che l'algoritmo BWA, introdotto nel Capitolo 2, viene utilizzato per allineare i frammenti dei *reads* ottenuti dall'immunoprecipitazione per definire la *coverage function* e quindi dedurre le caratteristiche epigenomiche ricercate con il processo di ChIP-Seq.

Per la caratterizzazione dell'algoritmo è necessario introdurre delle definizioni preliminari a partire da una stringa $X = a_0 a_1 \dots a_{n-2}$ formata da caratteri ordinabili appartenenti a un alfabeto Σ.
Per condurre le successive analisi è necessario aggiungere all'alfabeto il carattere ausiliario \$ che assume il valore più piccolo nell'ordine lessicografico degli elementi di Σ. Questo carattere ha il ruolo di indicare, durante tutto lo sviluppo dell'algoritmo, il punto iniziale e finale della stringa di partenza, infatti la stringa X viene modificata in $X = a_0 \dots a_{n-2} a_{n-1}$ con $a_{n-1} =$ \$ e $a_0, \dots, a_{n-2} \neq$ \$. Si definisce, inoltre, la lunghezza di X come $|X|$ ed è, in questo caso, pari a n.

Si introducono le caratteristiche essenziali di questo algoritmo attraverso la presentazione di un esempio:

Esempio

Per la caratterizzazione del BWA si utilizza la stringa `googol` ovvero

$$X = \texttt{g o o g o l \$}$$

con $\Sigma = \{\texttt{g, o, l, \$}\}$ e $|X| = n = 7$.

Per chiarire lo sviluppo successivo dell'algoritmo si definiscono alcuni elementi e operazioni preliminari nella classe delle stringhe:

Definizione 10. Estrazione del singolo carattere: il carattere in posizione i è

$$X[i] = a_i \;\; i = 0, \ldots n-1.$$

Definizione 11. Estrazione della sottostringa: la sottostringa da i a j è

$$X[i, j] = a_i, \ldots, a_j.$$

Definizione 12. Suffisso: il suffisso a partire dalla i-esima posizione è

$$X_i = X[i, n-1],\ X[0, i-1].$$

Definizione 13. Vettore suffisso: il vettore suffisso S di X è la permutazione degli interi $0 : n-1$ t.c. $S(i)$ è la posizione dell'i-esimo suffisso più piccolo, dove l'ordinamento è definito attraverso l'ordine lessicografico.

Definizione 14. Prefisso: la stringa Y di lunghezza $|Y| = m$ è un prefisso per X se

$$X[0, m-1] = Y.$$

Esempio

Per la stringa X precedentemente introdotta si possono definire i suffissi indicati in Tabella B.1

X_0	g o o g o l $
X_1	o o g o l $ g
X_2	o g o l $ g o
X_3	g o l $ g o o
X_4	o l $ g o o g
X_5	l $ g o o g o
X_6	$ g o o g o l

Tabella B.1: Suffissi per la stringa `googol`.

da cui il vettore suffisso S a partire dall'ordinamento dei suffissi risulta essere quello proposto in Tabella B.2.

X_6	$ g o o g o l	6
X_3	g o l $ g o o	3
X_0	g o o g o l $	0
X_5	l $ g o o g o	5
X_2	o g o l $ g o	2
X_4	o l $ g o o g	4
X_1	o o g o l $ g	1

Tabella B.2: Suffissi ordinati per la stringa `googol` e vettore suffisso.

Si introduce, infine la trasformata di Burrows-Wheeler che è alla base di questo algoritmo di allineamento:

Definizione 15. Trasformata di Burrows-Wheeler: la BTW B è una stringa di lunghezza $|B| = |X| = n$ definita a partire da S come:

$$B[i] = \begin{cases} \$ & \text{se } S(i) = 0 \\ X[S(i) - 1] & \text{se } S(i) \neq 0 \end{cases}$$

Esempio
poiché il vettore suffisso di $X =$ `'g o o g o l $'` è $S = (6, 3, 0, 5, 2, 4, 1)$
la BTW risultante è

$$B = \texttt{l o \$ o o g g}.$$

Questa trasformata racchiude in sé tutte le informazioni relative all'insieme dei suffissi, infatti da essa è facilmente ricostruibile l'insieme presentato in Tabella B.2 ed è pertanto molto utile nella memorizzazione snella di tutti questi elementi che risultano fondamentali nell'algoritmo di allineamento.

A partire dalla trasformata di Burrows-Wheeler esistono numerosi algoritmi per la ricerca locale di stringhe, in particolare esiste un algoritmo di ricerca esatta basato sul semplice ordinamento lessicografico di B; in questa sezione, tuttavia, si presenta un algoritmo iterativo che sfrutta la trasformata per la definizione del vettore dei conteggi e della matrice di occorrenza, che verranno poi direttamente utilizzati nell'agoritmo BWA.

Esempio
A conferma dell'utilità della trasformata di Burrows-Wheeler nella memorizzazione delle caratteristiche delle stringhe, si presenta il semplice procedimento di ricostruzione dei suffissi a partire dalla sola stringa B.
Il procedimento è basato solamente sull'ordinamento lessicografico e sulla concatenazione di stringhe.
Questo procedimento è presentato in dettaglio nella Tabella B.3, in cui la prima colonna è la trasformata di Burrows-Wheeler che viene ordinata (seconda colonna); si concatenano poi i caratteri della trasformata con i caratteri della colonna ordinata. Procedendo con questo processo di ordinamento e concatenazione con la stringa B si ottengono i suffissi introdotti precedentemente.

Trasformata	Ordino	Concateno	Ordino	Concateno	Ordino
l	$	l$	$g	l$g	$go
o	g	og	go	ogo	gol
$	g	$g	go	$go	goo
o	l	ol	l$	ol$	l$g
o	o	oo	og	oog	ogo
g	o	go	ol	gol	ol$
g	o	go	oo	goo	oog

Concateno	Ordino	Concateno	Ordino	Concateno	Ordino
l$go	$goo	l$goo	$goog	l$goog	$googo
ogol	gol$	ogol$	gol$g	ogol$g	gol$go
$goo	goog	$goog	googo	$googo	googol
ol$g	l$go	ol$go	l$goo	ol$goo	l$goog
oogo	ogol	oogol	ogol$	oogol$	ogol$g
gol$	ol$g	gol$g	ol$go	gol$go	ol$goo
goog	oogo	googo	oogol	googol	oogol$

Concateno	Ordino ⇒ Suffissi
l$googo	$googol
ogol$go	gol$goo
$googol	googol$
ol$goog	l$googo
oogol$g	ogol$go
gol$goo	ol$goog
googol$	oogol$g

Tabella B.3: Generazione a partire dalla trasformata di Burrows-Wheeler dei suffissi della stringa X.

Al termine del procedimento si conferma, quindi, la definizione dei suffissi introdotti nella Tabella B.2.

Si introducono, inoltre,

Definizione 16. Intervallo del vettore suffisso e Insieme delle posizioni: per una data sottostringa W di X, si definisce l'intervallo del vettore suffisso, come l'intervallo compreso tra $\underline{R}(W)$ e $\overline{R}(W)$ con:

$$\underline{R}(W) = \min\{k\text{: } W \text{ è un prefisso di } X_{S(k)}\}$$
$$\overline{R}(W) = \max\{k\text{: } W \text{ è un prefisso di } X_{S(k)}\}.$$

da cui si poi si può associare l'insieme delle posizioni di tutte le occorrenze di W in X come

$$P = \{S(k) : \underline{R}(W) \leq k \leq \overline{R}(W)\}$$

L'insieme P, dunque, rappresenta le esatte posizioni iniziali di tutte le localizzazioni di W in X.

Esempio
Se $W =$ `go` si trova che W è un prefisso di:

X_3	`g o l $ g o o`
X_0	`g o o g o l $`

che, con $S = (6, 3, 0, 5, 2, 4, 1)$, porta alla definizione di

$$\underline{R}(W) = 1 \quad \text{e} \quad \overline{R}(W) = 2$$

e l'insieme delle posizioni è

$$P = \{3, 0\}.$$

Avere definito l'insieme delle posizioni, quindi, equivale ad avere determinato l'insieme dove si localizza la stringa W nella stringa complessiva X, sempre tenendo conto che la prima posizione di X è indicizzata con 0.

Definiendo ancora

Definizione 17. Vettore dei conteggi: per ogni $a \in \Sigma$, $a \neq \$$ si definisce $C(a)$ come il numero di simboli in $X[0, n-2]$ lessicograficamente inferiori ad a.

Definizione 18. Matrice delle occorrenze: per ogni $a \in \Sigma$, $a \neq \$$ e con $i = 0, \ldots, n-1$ si può definire $O(a, i)$ come il numero si occorrenze di a in $B[0, i]$.

Esempio

poiché l'ordinamento lessicografico di Σ è {`g l o`} si trova

$$C(\texttt{'g'}) = 0 \qquad C(\texttt{'l'}) = 2 \qquad C(\texttt{'o'}) = 3$$

e quindi

$$O = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 & 3 & 3 & 3 \end{bmatrix}$$

Esiste una stretta connessione tra il vettore dei conteggi, la matrice delle occorrenze e gli estremi dell'intervallo del vettore suffisso, si dimostra [10], infatti, che

$$\underline{R}(aW) = C(a) + O(a, \underline{R}(W) - 1) + 1$$

$$\overline{R}(aW) = C(a) + O(a, \overline{R}(W))$$

e $\underline{R}(aW) \leq \overline{R}(aW)$ se e solo se aW è una sottostringa di X; queste considerazioni rendono possibile valutare se W è una sottostringa di X e contarne le occorrenze calcolando iterativamente $\underline{R}(W)$ e $\overline{R}(W)$ a partire dall'ultimo carattere di W (Algoritmo di Ferragina-Manzini).

É possibile, tuttavia, definire un algoritmo ricorsivo che semplifichi la ricerca e permetta la localizzazione in X della stringa W consentendo anche l'introduzione di z

errori tra *mismatches* e *gaps*. Si osserva che nell'esempio proposto $\Sigma = \{$`g, l, o, $`$\}$, mentre nel caso del dataset di ChIP-Seq $\Sigma = \{A, C, T, G, \$\}$, ovvero da tutti i possibili nucleotidi che compongono i frammenti.

0. Valutazioni preliminari:

   (a) calcolare la BWT B di X;

   (b) calcolare il vettore C e la matrice O per tutti gli elementi di B.

1. Calcolo di D a partire da W

```
j ← 0
t ← 0
for i = 0 to |W| − 1
    if W[j, i] non è una sottostringa di X
        t ← t + 1
        j ← i + 1
    D[i] ← t.
```

2. Calcolo degli intervalli del vettore suffisso I associati a W in X con la funzione ricorsiva INDICI_RICORSIVA

```
INDICI_RICORSIVA (W, i, z, k, l)
  se i < 0 restituisci {[k, l]}
  se z < D[i] restituisci ∅
 I ← ∅
 I ← I ∪ INDICI_RICORSIVA (W, i − 1, z − 1, k, l)      ⋆
  per ogni b ∈ Σ
      k ← C(b) + O(b, k − 1) + 1
      l ← C(b) + O(b, l)
       se k ≤ l
           I ← I ∪ INDICI_RICORSIVA (W, i, z − 1, k, l)      ⋆
            se b = W[i]
                I ← I ∪ INDICI_RICORSIVA (W, i − 1, z, k, l)
```

```
            altrimenti
                I ← I ∪ INDICI_RICORSIVA (W, i − 1, z − 1, k, l)
    restituisci I.
```

La prima parte dell'algoritmo prevede il calcolo del vettore $D(i)$, questo per ogni i in $0 : |W| - 1$ rappresenta il limite inferiore del numero di differenze in $W[0, i]$, ovvero il numero minimo di errori che si devono compiere per trovare una corrispondenza tra $W[0, i]$ e X; ovviamente un limite inferiore accettabile è $D(i) = 0 \ \ \forall i \in \{0, \ldots, |W| - 1\}$, tuttavia una limitazione così poco stringente risulta computazionalmente poco adeguata. Esistono numerosi metodi per il calcolo del parametro D; si presenta quello più semplice che prevede la facile ricerca di sottostringhe nella sequenza di partenza.

Esempio

Per la ricerca nella X precedentemente introdotta della stringa $W =$ `go` si trova che

$$D = (0, 0)$$

infatti:

- $i = 0$ $W[0, 0] =$ `g` è una sottostringa di X e dunque $D(0) = 0$;
- $i = 1$ $W[0, 1] =$ `go` è ancora una sottostringa di X e quindi $D(1) = 0$.

Per una sottostringa diversa $W =$ `lol` , invece, si trova

$$D = (0, 1, 1)$$

infatti:

- $i = 0$ $W[0, 0] =$ `l` è una sottostringa di X e pertanto $D(0) = 0$;
- $i = 1$ $W[0, 1] =$ `lo` non è una sottostringa di X, quindi si incrementano t e j, da cui $t \leftarrow 1$ e $j \leftarrow i + 1 = 2$ e $D(1) = 1$;
- $i = 2$ $W[2, 2] =$ `l` è una sottostringa di X e quindi non sono necessari incrementi di z e $D(2) = 1$.

La seconda parte dell'algoritmo, invece, permette di trovare tutti gli intervalli di X contenenti la stringa W a meno di z differenze, con z scelto opportunamente tenendo conto del vettore D. In questa prima implementazione si considera l'introduzione di *gap* (in particolare la possibilità di introdurre inserimenti o cancellazioni è prevista dalle righe dell'algoritmo indicate con $\star$) o *mismatch* allo stesso modo, ma è possibile introdurre dei coefficienti di penalizzazione per privilegiare la sostituzione dei caratteri o l'introduzione del *gap* nell'allineamento.

Gli intervalli cercati per la corrispondenza di W in X si trovano, quindi, con la funzione

$$\text{INDICI\_RICORSIVA}\ (W, |W| - 1, z, 1, |X| - 1).$$

B.2 Algoritmo per il calcolo dell'indice di forma M

In questa Sezione si propone nel dettaglio l'algoritmo di Edmonds per il calcolo dell'indice di forma M per un albero dalla struttura generica.
Questo algoritmo, o *algoritmo blossom* costruisce, a partire da un generico grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, l'insieme di *matching massimo* migliorando iterativamente il *matching* precedentemente trovato fino a che questo non diventa massimo, con una complessità computazionale polinomiale nel numero n di nodi dell'albero $O(n^3)$.
Per presentare l'algoritmo è necessario definire per un dato grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ e un *matching* $\mathcal{M}$:

Definizione 19. Vertice esposto: $v \in \mathcal{V}$ si dice esposto per il *matching* $\mathcal{M}$ se nessun arco di $\mathcal{M}$ è incidente in v.

Definizione 20. Cammino alternante: un cammino è alternante per $\mathcal{M}$ se i suoi archi appartengono alternativamente a $\mathcal{M}$ o non vi appartengono.

Definizione 21. Cammino aumentante: un cammino alternante è aumentante se inizia e finisce in due vertici esposti distinti.

Definizione 22. Aumento del *matching* $\mathcal{M}$ lungo un cammino aumentante $\mathcal{P}$: operazione di sostituzione del *matching* $\mathcal{M}$ con un nuovo *matching* $\mathcal{M}_1$, ottenuto come

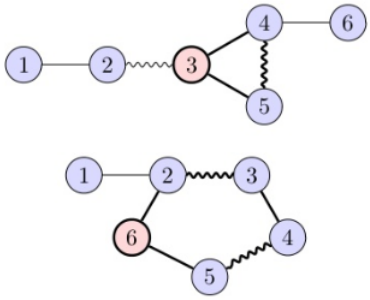


Figura B.1: Esempi di *blossom*.

differenza simmetrica di $\mathcal{M}$ e $\mathcal{P}$, ovvero composto da tutti gli archi che appartengono solo a $\mathcal{M}$ o solo a $\mathcal{P}$, cioè:

$$\mathcal{M}_1 = \mathcal{M} \ominus \mathcal{P} = (\mathcal{M} - \mathcal{P}) \cup (\mathcal{P} - \mathcal{M})$$

Si presenta, inoltre il risultato che è alla base dell'algoritmo iterativo di Edmond:

Teorema 1. ***Teorema di Berge:*** *Un grafo $\mathcal{G}$ con mathcing $\mathcal{M}$ ha un cammino aumentante se e solo se $\mathcal{M}$ non è massimo.*

Si conclude, pertanto, che un *matching* è massimo se in $\mathcal{G}$ non esiste nessun cammino aumentante per $\mathcal{M}$; la definizione del *matching massimo* prevede, quindi, partendo da un insieme iniziale, la definizione di cammini aumentanti fino a che non si raggiunge questa condizione di massimalità.
L'algoritmo proposto per la definizione di *matching massimo* prevede, in primo luogo la definizione di *blossom* e la presentazione del Lemma di Edmonds.

Definizione 23. ***Blossom***: dati $\mathcal{G}$ e $\mathcal{M}$ si dice *blossom B* un ciclo in $\mathcal{G}$ contenente $2k + 1$ nodi di cui esattamente k appartengono a $\mathcal{M}$ e dove uno dei nodi del ciclo v (base del *blossom*) è tale che esiste un cammino alternato che connette v a un vertice esposto della componente connessa di v.

Due esempi di *blossom* con evidenziata la base sono presentati in Figura B.1.
Si può, inoltre, dimostrare [8]:

Lemma 1. ***Lemma di Edmonds:*** *siano $\mathcal{G}'$ e $\mathcal{M}'$ il grafo e il matching ottenuti contraendo un blossom B di $(\mathcal{G}, \mathcal{M})$ nella sua radice, il matching $\mathcal{M}$ di $\mathcal{G}$ è massimo se e solo se $\mathcal{M}'$ è massimo in $\mathcal{G}'$.*

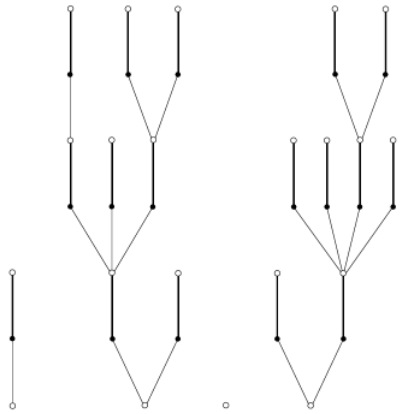


Figura B.2: Foresta alternante associata a un *matching* $\mathcal{M}$.

L'algoritmo di ricerca del *matching* massimo per un grafo $\mathcal{G}$ prevede ad ogni iterazione la definizione di un cammino aumentante o la contrazione di un *blossom* e conseguentemente la modifica dell'ambito di ricerca in $\mathcal{G}'$. Questo procedimento continua fino a che non è più possibile contrarre *blossom* o definire cammini aumentanti; ovvero si è raggiunto il *matching* di cardinalità massima. Per questa ricerca è necessario definire la struttura ausiliaria della foresta alternante associata al grafo $\mathcal{G}$ rispetto al *matching* $\mathcal{M}$ (in Figura B.2 si presenta un esempio di foresta alternarne con il *matching* evidenziato in grassetto).
Si definiscono, quindi:

Definizione 24. Foresta in $\mathcal{G}$: insieme di grafi corrispondenti ciascuno a una porzione specifica del grafo $\mathcal{G}$.

Definizione 25. Foresta $\mathcal{F}$ alternante rispetto a $\mathcal{M}$ in $\mathcal{G}$: foresta in $\mathcal{G}$ con le seguenti proprietà:

- ogni componente connessa di $\mathcal{F}$, ovvero ogni grafo della foresta, contiene esattamente un vertice non coperto da $\mathcal{M}$, cioè la radice del grafo;
- per ogni vertice v di $\mathcal{F}$, l'unico cammino che connette v alla sua radice è un cammino $\mathcal{M}$-alternante.

In una foresta alternante si possono distinguere i vertici esterni e interni: i vertici esterni sono quei nodi che hanno distanza pari dalla radice della componente connessa

che li contiene (nodi vuoti in Figura B.2), mentre quelli interni hanno distanza dispari (nodi pieni in Figura B.2).

La procedura di definizione del cammino aumentante considera i vertici $x \in \mathcal{V}$ e gli archi $\{x, y\} \in \mathcal{E}$ per modificare adeguatamente la foresta $\mathcal{F}$ e quindi il *matching*.
Una volta costruita la foresta $\mathcal{M}$ alternante, per ogni vertice esterno x, si considerano tutti gli $y \in \mathcal{V}$, tale che $\{x, y\}$ non appartiene a $\mathcal{M}$ ed è, quindi, possibile:

1. qualora y non appartenga a $\mathcal{F}$, **modificare la foresta** aggiungendo all'albero di $\mathcal{F}$ associato a x gli archi $\{x, y\}$ e $\{y, z\}$ con z vertice di $\mathcal{G}$ connesso a y e $\{y, z\}$ arco di $\mathcal{M}$;

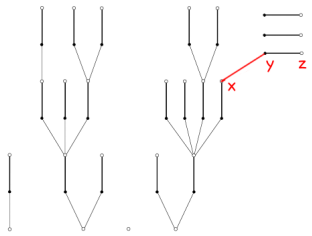


Figura B.3: Esempio di crescita della foresta.

2. se y appartiene a $\mathcal{F}$, ma ad un grafo di $\mathcal{F}$ diverso da quello di x ed è vertice esterno per la foresta, **definire un cammino aumentante** per poi aumentare il *matching*. Il cammino è costituito dal cammino dalla radice $v(x)$ dell'albero di x a x, dall'arco $\{x, y\}$ e dal cammino da y alla radice $v(y)$;

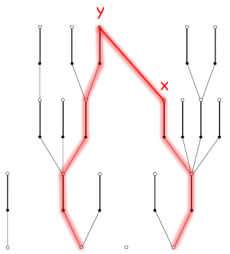


Figura B.4: Esempio di cammino aumentate.

3. se y appartiene allo stesso albero di x in $\mathcal{F}$ ed è vertice esterno, allora deve esistere un *blossom* costituito da $\{x, y\}$ e dagli archi che collegano x e y nell'albero da cui si può **contrarre il *blossom*** e quindi ricercare il cammino aumentante nel grafo $\mathcal{G}'$ a partire dal *matching* $\mathcal{M}'$[1].

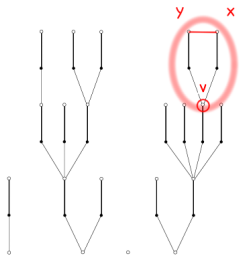


Figura B.5: Esempio di *blossom* per la foresta.

Complessivamente si può riassumere l'algoritmo di Edmonds:

dato un grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

1. definizione di un matching $\mathcal{M}$;
2. definizione della foresta $\mathcal{M}$-alternante $\mathcal{F}$;
3. per ogni vertice esterno x di $\mathcal{M}$:

   3.1 per ogni vertice y tale che esiste una connessione $\{x, y\}$ in $\mathcal{E}$:

   - se $y \notin \mathcal{F}$ $\Rightarrow$ Crescita della foresta
   - se $y \in \mathcal{F}$, y è vertice esterno di $\mathcal{F}$ e la componente connessa di y è diversa dalla componente connessa di x $\Rightarrow$ Aumento del matching, ovvero $\mathcal{M}' = \mathcal{M} \oplus P(x) \cup \{x, y\} \cup P(y)$ e punto 1.
   - se $y \in \mathcal{F}$, y è vertice esterno di $\mathcal{F}$ e la componente connessa di y è uguale alla componente connessa di x

[1] Si precisa che, data la struttura ad albero dei dati analizzati in questa tesi, non si presentano mai *blossom* e quindi non si incorre mai nella necessità si modificare in questo modo i grafi della foresta.

$\Rightarrow$ Contrazione del blossom e quindi modifica di $\mathcal{F}, \mathcal{G}, \mathcal{M}$ e punto 1.

- se $y \in \mathcal{F}$ e y è vertice interno di $\mathcal{F}$ $\Rightarrow$ non si effettuano modifiche a $\mathcal{F}$ o $\mathcal{M}$

Si aggiorna così il matching fino a che non è possibile apportare più modifiche a $\mathcal{M}$ o alla foresta e si ottiene il matching massimo.

Bibliografia

[**Capitolo 1**]

[1] A. F. Bardet, Q. He, J. Zeitlinger, A. Stark: *A computational pipeline for comparative ChIP-Seq analyses*, Nature Protocols, 2011, vol. 7 n. 1

[2] A. Barski, K. Zhao: *Genomic Location Analysis by ChIP-Seq*, Journal of Cellular Biochemistry, 2009, 107

[3] A. Diaz, K. Park, D. A. Lim, J. S. Song: *Normalization, bias correction and peak calling for ChIP-seq*, Statistical Applications in Genetics and Molecular Biology, 2012; 11(3)

[4] P. J. Park *ChIP-seq: advantages and challenges of a maturing technology*, Nature Reviews Genetics, ottobre 2009; 10(10)

[5] R. E. Thurman et al.: *The accessible chromatin landscape oh the human genome*, Nature: international weekly journal of science, 2012, n. 489

[6] www.genome.ucsc.edu/ENCODE

[7] www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[**Capitolo 2**]

[8] J. Edmonds: *Path, trees and flowers*, Canadian Journal of Mathematics, 1965, 17

[9] J. Feng, T. Liu, B. Qin, Y. Zhang, X. S. Liu: *Identifying ChIP-seq enrichment using MACS*, Nature Protocols, 2012, vol. 7 n. 9

[10] H. Li, R. Durbin: *Fast and accurate short read alignment with BurrowsÐWheeler transform*, Bioinformatics, 2009, vol. 25 n. 14

[11] S. Pepke, B. Wold, A. Mortazavi: *Computation for ChIP-seq and RNA-seq studies*, Nature America, novembre 2009, vol. 6 n. 11

[12] The SAM/BAM Format Specification Working Group: *Sequence Alignment/Map Format Specification*, maggio 2013

[13] Y. Zhang, T. Liu, C. A. Meyer et al.: *Model-based Analysis of ChIP-Seq (MACS)*, Genome Biology, 2008

[14] www.bioninformatics.babraham.ac.uk/projects/fastqc

[15] www.illumina.com, *Illumina Sequencing technology*

[16] www.illumina.com/NGS, *An introduction to Next Generation Sequencing*

[**Capitolo 3**]

[17] L. D. Brown, L. H. Zhao: *A test for the Poisson distribution*, The Indian Journal of Statistics, 2002; vol. 64, serie A

[18] S. N. Evans, V. Hower, L. Pachter: *Coverage statistics for sequence census methods*, BMC Bioinformatics, 2010, 11

[19] V. Hower, S. N. Evans, L. Pachter: *Shape-based peak identification for ChIP-Seq*, BMC Bioinformatics, 2011, 12

[20] S. Janson: *Simply generated trees, conditioned GaltonÐWatson trees, random allocations and condensation* , Probability Surveys, 2012, vol. 9

[**Capitolo 4**]

[21] B. Bernahrd, H. Korte, J. Vygen: *Ottimizzazione combinatoria: teoria e algoritmi* Springer, 2011

[**Capitolo 5**]

[22] J. O. Ramsay, B. W. Silverman: *Functional Data Analysis* Springer, 2005

[23] L. M. Sangalli, P. Secchi, S. Vantini, V. Vitelli: *k-mean alignment for curve clustering*, Computational Statistics and Data Analysis, 2010, 54

[**Capitolo 6**]

[24] P.Machanick, T. L. Bailey: *MEME-ChIP: motif analysis of large DNA datasets* Bioinformatics 2011, vol.27 n.12

[25] S. Malinge, T. Chlon, L. C. Doré, R. C. Ketterling et al.: *Development of acute megakaryoblastic leukemia in Down syndrome is associated with sequential epigenetic changes*, Blood, agosto 2013

[26] www.http://meme.nbcr.net/meme/cgi-bin/meme-chip.cgi

Elenco delle figure

Elenco delle tabelle

Ringraziamenti

Un sentito ringraziamento al Professor Piercesare Secchi, per i preziosi consigli, il tempo che mi ha dedicato e i numerosi spunti di riflessione che mi ha suggerito; un particolare grazie a Laura Sangalli e Simone Vantini per la loro pazienza, il loro entusiasmo, le loro sagge idee e il loro interessamento per questo lavoro. Grazie anche a Mariza Cremona, già parte attiva del progetto che mi ha guidato nell'analisi di questo ambito nuovo e a Giancarlo Ferrari che con la sua "mania per gli alberi" ci ha convinto a seguire anche questo percorso.

Desidero ringraziare in questo momento di conclusione del percorso universitario la mia famiglia che mi ha sempre sostenuto nelle mie scelte e fatto sì che mantenessi in ogni occasione l'entusiasmo e la serenità, anche nei momenti di stanchezza e sconforto, e che mi ha insegnato a contare su me stessa e a dare il meglio di me in ogni circostanza; insegnamenti che certamente mi hanno aiutato in questo percorso al Politecnico, ma che terrò sempre preziosi.

Ma anche grazie a tutti gli amici, compagni di questa avventura al Poli e non. Grazie a Stefano che si è sempre dimostrato un amico in questi ultimi anni sopportando le mie paranoie e che . . . ha una bellissima casa a Madrid; grazie a Marco, con cui ogni tanto si riesce a fare anche un discorso serio (anche se finge sin dal primo anno di essere la "più stupida persona stupida" che conosciamo); a Giacomo che dal ragazzo in prima fila con la MontBlanc ora è diventato un grande compagno di studio e risate; a Teresa (e Nabla) senza le quali le vacanze non sarebbero state le stesse; a Fabio che con i racconti delle sue conquiste ha reso incredibilmente divertenti le pause caffè; a Anna e Fabrizio per le mitiche torte che hanno conquistato anche altri dipartimenti del Poli (forse il contributo di Fabrizio per le torte non è fondamentale, ma è certamente un buon amico su cui contare);

a Giancarlo e Giorgio, immancabili compagni di vacanze (in realtà Giancarlo ha dovuto, purtroppo per lui, sopportarmi per un po' anche qui in università) e ad Alessandro, amico sin dal liceo con cui finalmente (devo ringraziare il Politecnico anche per questo) ci siamo rivisti. Grazie, infine, agli amici di sempre Alessandro, Pietro e Alessandra, che in un modo o nell'altro so che ci saranno sempre.