UNIVERSITÀ DEGLI STUDI DI PADOVA

**UNIVERSITÀ DEGLI STUDI DI PADOVA**

**Dipartimento di Ingegneria dell'Informazione**

Corso di Laurea in

**INGEGNERIA INFORMATICA**

# Stima della dimensione del genoma tramite k-mers: confronto tra metodi computazionali.


Relatore:
**Prof. Matteo Comin**

Laureando:
**Mattia Tamiazzo**


Data di laurea xx/09/2022

**Anno Accademico 2021-2022**

# Sommario

La dimensione del genoma è la quantità totale di DNA nucleare aploide presente nelle cellule di un organismo. La determinazione della dimensione del genoma costituisce un argomento di interesse, perché non esistono valori di riferimento assoluti che permettono di stabilire quale approccio sia più efficace, e perché i metodi sperimentali per la sua misurazione sono attualmente costosi dal punto di vista temporale ed economico. Una soluzione alla stima della dimensione del genoma con metodi computazionali è l'utilizzo di k-mers, sottostringhe di DNA di lunghezza k. Questa trattazione si pone l'obiettivo di analizzare e comparare vari approcci algoritmici pubblicati in letteratura per la stima della dimensione del genoma.

# Indice

# Capitolo 1

# Introduzione

Il sequenziamento del DNA costituisce una tecnica fondamentale per lo studio del genoma di una specie, perché permette di determinare l'ordine delle basi azotate dei nucleotidi che costituiscono il DNA. Tale processo trova applicazione in molti studi biologici che riguardano vari ambiti, come ad esempio la medicina riproduttiva, l'oncologia o l'infettivologia, attraverso indagini tra cellule diverse dello stesso individuo o lo studio delle mutazioni genetiche tra individui di una stessa specie [1].

## 1.1 Storia

Lo studio approfondito del DNA si sviluppa a partire dal 1953, con la scoperta della sua struttura tridimensionale ad opera di James Watson e Francis Crick [2], contribuendo all'analisi dell'azione degli acidi nucleici nella sintesi proteica. Solo nel 1977 però, vennero sviluppate le prime strategie sperimentali per il sequenziamento, come il famoso metodo Sanger [3, 4] TODO

## 1.2 Sequenze di lunghezza k: i k-mer

TODO

### 1.2.1 K-mer profile

Il *k-mer profile*, detto anche *k-mer spectrum*, conta la frequenza dei k-mer trovati nelle letture di input, non assemblate o allineate. Esso può rappresentare un indicatore della complessità del genoma preso in esame [5], e mostra la quantità di k-mer distinti trovati ad una certa frequenza. Un esempio di k-mer profile è mostrato dalla figura 1.1 nella pagina seguente tratta da [6], in cui si può notare come la natura del genoma influenzi direttamente il grafico.

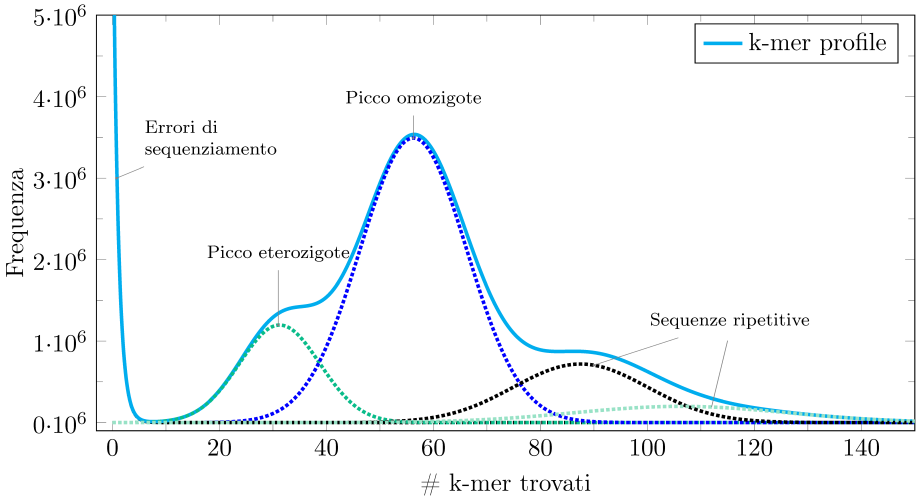


Figura 1.1: TODO.

Ipotizzando che il genoma sia ideale, omozigote e senza ripetizioni, e che le letture siano state fatte senza errori con una certa copertura, il grafico del k-mer profile sarà una distribuzione di Poisson centrata sulla copertura media disponibile.

In casi reali invece, il genoma sarà eterozigote con una certa percentuale di eterozigosi e saranno presenti errori di sequenziamento; il k-mer profile presenterà tre picchi principali [7]. Il primo picco del grafico corrisponde ai k-mer derivati da errori di sequenziamento, che accadono spesso ma che hanno bassa frequenza perché presentano poche occorrenze nelle letture di input; il secondo invece, rappresenta i k-mer eterozigoti e il terzo quelli omozigoti, presenti quindi su uno o entrambi gli alleli del set di cromosomi. I k-mer eterozigoti devono essere trattati più attentamente, perché possono risultare simili a quelli del primo picco, derivanti da errori di sequenziamento [6].

La lunga coda della distribuzione rappresenta invece le sequenze ripetitive, che occorrono con alta frequenza e sono presenti in un elevato numero di locus. Eventuali ripetizioni aggiungono al grafico ulteriori picchi, mentre errori nelle letture aumentano la varianza e producono distorsioni nel grafico.

La figura 1.2 nella pagina successiva mostra come all'aumentare del rapporto di eterozigosi la quantità di k-mer eterozigoti del secondo picco diventi dominante rispetto ai k-mer omozigoti del terzo picco, che invece diminuiscono.

Il k-mer profile può essere calcolato tramite programmi specifici date delle letture del genoma di input, quali *Jellyfish* [8] o *KMC2* [9].

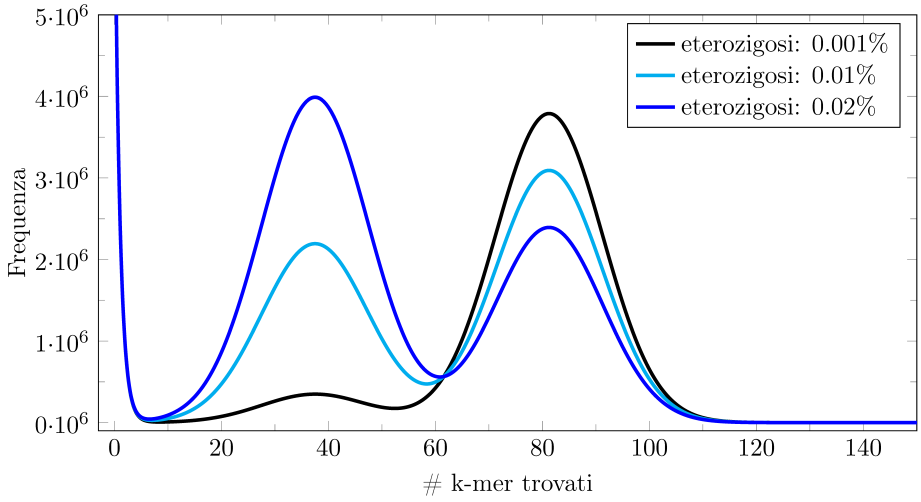


Figura 1.2: TODO.

# Capitolo 2

# GenomeScope

Il progetto open source *GenomeScope* cerca sia di stimare le caratteristiche del genoma completo, come la sua lunghezza o il rapporto di eterozigosi, sia di determinare le proprietà delle letture di DNA che prende in input, come la copertura (*read coverage*) o l'error rate [5]. Il programma per determinare tali caratteristiche utilizza il k-mer profile del genoma preso in esame, descritto nella sezione 1.2.1 a pagina 1.

## 2.1 Algoritmo

Il programma effettua una regressione non lineare dei dati iniziali, generando un profilo che cerca di approssimare il k-mer profile reale. Prendendo in input le letture del genoma che si vuole studiare, esso crea un modello che approssima il più possibile il k-mer profile. La funzione $f(X)$ scelta per l'interpolazione delle frequenze dei k-mer trovati è la somma di quattro distribuzioni binomiali negative $\mathcal{NB}(X; p, n)$, rispettivamente per rappresentare k-mer eterozigoti trovati nel genoma diploide una volta (unici) o tre volte (duplicati), e k-mer omozigoti di cui si trovano due occorrenze (unici) o trovati quattro volte (duplicati). La funzione $f(X)$ è descritta dall'equazione 2.1, in cui $G$ rappresenta un coefficiente di scala legato alla dimensione del genoma, $\lambda$ e $\rho$ sono rispettivamente la media e la varianza della distribuzione.

$$f(X) = G * (\alpha\mathcal{NB}(X; \lambda, \lambda/\rho) + \beta\mathcal{NB}(X; 2\lambda, 2\lambda/\rho) + \gamma\mathcal{NB}(X; 3\lambda, 3\lambda/\rho) + \delta\mathcal{NB}(X; 4\lambda, 4\lambda/\rho)) \quad (2.1)$$

I coefficienti $\alpha, \beta, \gamma$ e $\delta$ dipendono dai parametri $r$ e $d$, che rappresentano rispettivamente il rapporto di eterozigosi, cioè la percentuale di basi che sono specifiche a uno o due cromosomi omologhi, e la percentuale del genoma che è presente in due copie.

Lo scopo del programma è quindi determinare i coefficienti $r, d, \lambda$ e $\rho$, oltre alla dimensione totale del genoma $G$. La funzione scelta $f(X)$, tramite cui poi può essere calcolata la dimensione del genoma, è quella che restituisce la minore somma dei quadrati degli errori residui (*Residual Sum of Square Error - RSSE*), cioè che minimizzi la somma tra i qua-

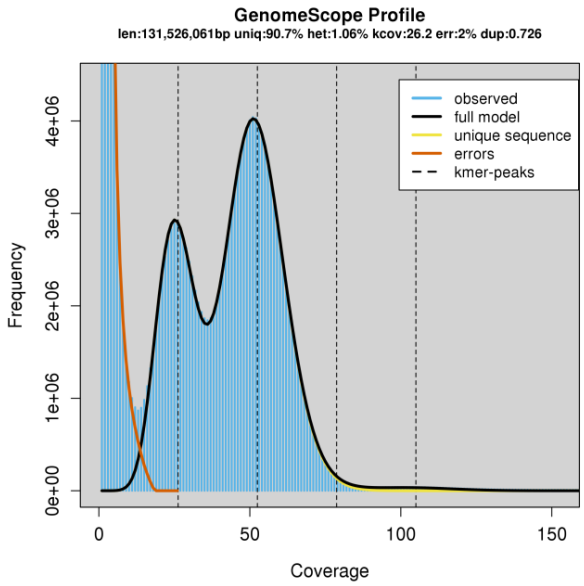


Figura 2.1: TODO + TODO reference a figura.

drati degli errori tra i valori osservati e quelli stimati, come descritto dall'equazione 2.2. Per dedurre i valori dei coefficienti, viene utilizzata la funzione `nls` del linguaggio di programmazione R, che compie la regressione non lineare dei dati alla funzione obiettivo.

$$RSSE = \sum_{x=E}^{+\infty} (kmer_{obs}[x] - kmer_{pred}[x])^2 \tag{2.2}$$

Al termine, il programma mostra all'utente i dati relativi al genoma trovati, come il rapporto di eterozigosi, la media e la varianza della distribuzione, l'indice RSSE, che rappresenta la percentuale di k-mer non considerati dal modello, e la dimensione stimata del genoma.

Eventuali errori di sequenziamento, ad esempio dovuti a duplicazioni con PCR o a sequenze contaminate, sono determinati solo empiricamente: dopo varie iterazioni del software in cui viene abbassata la soglia di copertura richiesta, i k-mer che non riescono ad essere rappresentati dal modello vengono identificati come errori di sequenziamento.

# Capitolo 3

# findGSE

Il programma *findGSE* [7] ha come obiettivo principale la stima della lunghezza del genoma. Utilizzando le frequenze dei k-mer trovati nelle letture a disposizione, il programma compie una regressione non lineare dei dati utilizzando come funzione una distribuzione normale asimmetrica (*skew normal distribution* [10, 11]).

## 3.1 Algoritmo

Dato un genoma aploide con $G$ basi, il numero di k-mer possibili sarà $G-k+1$. Ponendo $C$ la copertura dei k-mer, cioè che in media ogni k-mer sia trovato in $C$ letture diverse, e $N$ il numero di k-mer trovati nelle letture, la quantità di k-mer presenti nel genoma sarà $N = C * (G - K + 1)$. Dall'equazione si deduce che $G \approx N/C$ se $G \gg k$.

Nel programma viene assunto che le frequenze dei k-mer possano essere approssimate da una distribuzione normale asimmetrica $SN(\xi, \omega^2, \alpha)$. Presa in input la distribuzione delle frequenze dei k-mer (k-mer profile), l'algoritmo effettua la regressione determinando i quattro parametri che descrivono una distribuzione normale asimmetrica, la media $\xi$, la deviazione standard $\omega$, l'asimmetria $\alpha$ e un fattore di scala $s$. Ad ogni iterazione, il programma cerca di minimizzare l'errore tra i dati di input e la funzione stimata, in modo da approssimare il più possibile il k-mer profile reale.

# Glossario

**Distribuzione normale asimmetrica** TODO Una distribuzione normale asimmetrica.. 7

**Distribuzione binomiale negativa** TODO Una distribuzione.. 5

**Distribuzione di Poisson** TODO Una distribuzione.. 2

**Locus** TODO Una distribuzione.. 2

**Rapporto di eterozigosi** TODO Una distribuzione.. 2

# Bibliografia


[1] J. Shendure e E. L. Aiden. «The expanding scope of DNA sequencing». In: *Nature Biotechnology* 30.11 (nov. 2012), pp. 1084–1094. ISSN: 1546-1696. DOI: 10.1038/nbt.2421.

[2] J. D. Watson e F. H. C. Crick. «Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid». In: *Nature* 171.4356 (apr. 1953), pp. 737–738. ISSN: 1476-4687. DOI: 10.1038/171737a0.

[3] F. Sanger, S. Nicklen e A. R. Coulson. «DNA sequencing with chain-terminating inhibitors». In: *Proceedings of the National Academy of Sciences* 74.12 (gen. 1977), pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

[4] F. Sanger et al. «Nucleotide sequence of bacteriophage $\phi$X174 DNA». In: *Nature* 265.5596 (feb. 1977), pp. 687–695. ISSN: 1476-4687. DOI: 10.1038/265687a0.

[5] G. W. Vurture et al. «GenomeScope: fast reference-free genome profiling from short reads». In: *Bioinformatics* 33.14 (lug. 2017), pp. 2202–2204. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx153.

[6] J. Sohn e J. Nam. «The present and future of de novo whole-genome assembly». In: *Briefings in Bioinformatics* 19.1 (ott. 2016), pp. 23–40. ISSN: 1477-4054. DOI: 10.1093/bib/bbw096.

[7] H. Sun et al. «findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies». In: *Bioinformatics* 34.4 (ott. 2017), pp. 550–557. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx637.

[8] G. Marçais e C. Kingsford. «A fast, lock-free approach for efficient parallel counting of occurrences of k-mers». In: *Bioinformatics* 27.6 (gen. 2011), pp. 764–770. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr011.

[9] S. Deorowicz et al. «KMC 2: fast and resource-frugal k-mer counting». In: *Bioinformatics* 31.10 (gen. 2015), pp. 1569–1576. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv022.

[10] A. Azzalini. «A Class of Distributions Which Includes the Normal Ones». In: *Scandinavian Journal of Statistics* 12.2 (1985), pp. 171–178. ISSN: 03036898, 14679469. URL: http://www.jstor.org/stable/4615982 (visitato il 05/08/2022).

[11] A. Azzalini. «The Skew-normal Distribution and Related Multivariate Families». In: *Scandinavian Journal of Statistics* 32.2 (mag. 2005), pp. 159–188. DOI: https://doi.org/10.1111/j.1467-9469.2005.00426.x.