

**UNIVERSITÀ
DEGLI STUDI
DI PADOVA**



**DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE**

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

1	Introduzione	1
1.1	Storia	1
2	GenomeScope	3
2.1	K-mer profile	3
2.2	Algoritmo	3
	Bibliografia	7

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 [5].

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 [7], 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]

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 [6]. Il programma per determinare tali caratteristiche fa uso del *k-mer profile*, che può essere calcolato tramite programmi specifici, quali *Jellyfish* [2] o *KMC2* [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, come descritto dalla figura 2.1 nella pagina seguente tratta da TODO.

Ipotizzando che le letture siano state fatte senza errori con una certa copertura, se il genoma iniziale è omozigote e senza ripetizioni, il grafico del k-mer profile sarà una distribuzione di Poisson centrata sulla copertura media disponibile; se il genoma è eterozigote invece, il grafico del k-mer profile avrà due picchi in corrispondenza di una certa frequenza e del doppio del suo valore, che rappresentano rispettivamente i k-mer eterozigoti o omozigoti trovati, presenti su uno o su entrambi gli alleli.

Eventuali ripetizioni aggiungono al grafico ulteriori picchi, mentre errori nelle letture aumentano la varianza e producono distorsioni nel grafico.

2.2 Algoritmo

Il programma effettua una regressione non lineare dei dati del k-mer profile, generando un nuovo 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 approssimi 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

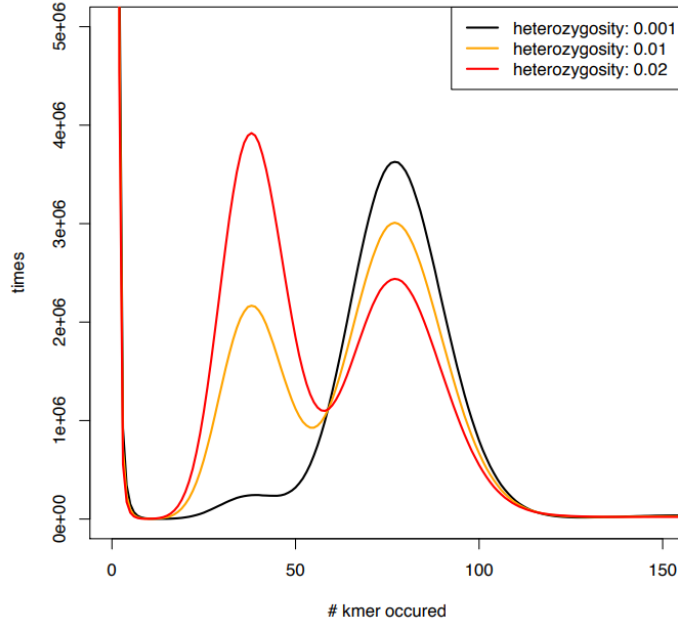


Figura 2.1: TODO.

(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, λ e ρ 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 α, β, γ e δ 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, λ e ρ , 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 quadrati 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.

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

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

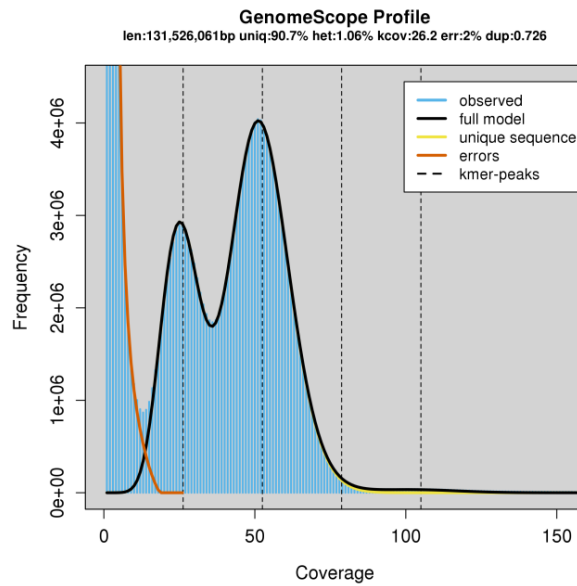


Figura 2.2: TODO + TODO reference a figura.

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.

Bibliografia

- [1] 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](https://doi.org/10.1093/bioinformatics/btv022).
- [2] 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](https://doi.org/10.1093/bioinformatics/btr011).
- [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 (1977), pp. 5463–5467. DOI: [10.1073/pnas.74.12.5463](https://doi.org/10.1073/pnas.74.12.5463).
- [4] F. Sanger et al. «Nucleotide sequence of bacteriophage ϕ X174 DNA». In: *Nature* 265.5596 (feb. 1977), pp. 687–695. ISSN: 1476-4687. DOI: [10.1038/265687a0](https://doi.org/10.1038/265687a0).
- [5] 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](https://doi.org/10.1038/nbt.2421).
- [6] 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](https://doi.org/10.1093/bioinformatics/btx153).
- [7] 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](https://doi.org/10.1038/171737a0).