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 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 Metodi di sequenziamento

1.1.1 Metodi di prima generazione

Le prime tecniche di sequenziamento del genoma furono sviluppate nella seconda metà del Novecento. Nel 1977 infatti vennero pubblicati due metodi di sequenziamento: il metodo Sanger, nel quale la sequenza di nucleotidi viene frammentata grazie a un terminatore di catena [2, 3], e il metodo di Maxam e Gilbert, in cui vengono utilizzati reagenti chimici per tagliare il DNA in frammenti in corrispondenza di basi specifiche [4]. Entrambi i metodi prevedono la misurazione dei frammenti creati tramite elettroforesi su gel di poliacrilammide con una corsia per base, in modo che siano separati in ordine di lunghezza e sia possibile dedurre l'ordine delle basi della sequenza in esame [5]. Mentre il metodo Maxam-Gilbert è stato progressivamente accantonato per la difficoltà tecnica e l'uso di sostanze tossiche che lo caratterizzano, il metodo Sanger è stato affinato con l'utilizzo di marcatori fluorescenti diversi per ogni base, che un lettore laser può distinguere restituendo la sequenza di basi di ogni frammento.

1.1.2 Shotgun assembly

Il sequenziamento del genoma può essere fatto su sequenze più lunghe, come un intero cromosoma, tramite *shotgun assembly*, metodo suggerito già nel 1979 [6]. Il genoma iniziale viene duplicato in modo da produrne più copie identiche, le quali vengono tagliate

in frammenti casuali (*shotgun*) che possono essere letti singolarmente. Si procede quindi con l'*assembly*, cioè la ricostruzione del genoma iniziale. L'assemblamento dei frammenti può avvenire con due metodi diversi, tramite *reference assembly* o con *de novo assembly*.

Reference assembly Per la ricostruzione del genoma viene utilizzata una sequenza di riferimento il più possibile simile alle letture disponibili, che permette di assemblare i frammenti più facilmente tramite allineamento [7].

De novo assembly Se non è disponibile una sequenza di riferimento appropriata, i frammenti vengono legati insieme identificando pattern sovrapponibili e formando più *contig*, che vengono poi combinati con altre tecniche algoritmiche [8].

Ogni frammento viene quindi allineato con gli altri shotgun disponibili, formando una sequenza comune, il *consensus*; ciascuna base della sequenza assemblata ha una certa copertura (*coverage*), che è pari al numero di letture che contribuiscono al posizionamento della base nel consensus. La figura 1.1 mostra un semplice esempio di questo metodo.

Figura 1.1: Esempio di allineamento di shotgun per la formazione del consensus.

1.1.3 Metodi di seconda e terza generazione

I metodi di seconda generazione, comunemente chiamati *NGS - Next Generation Sequencing*, introducono un'elevata parallelizzazione del sequenziamento. Sostanzialmente, i vari metodi disponibili (come ad esempio *454*, *Illumina* o *Ion Torrent*) prevedono inizialmente la frammentazione del genoma da sequenziare, e l'unione di particolari sequenze ai frammenti creati. Essi vengono quindi amplificati tramite *emPCR* o *bridge amplification* [9], e possono essere sequenziati con letture parallele, spesso facendo uso di molecole fluorescenti. Segue quindi l'assemblamento delle letture ottenute per la ricostruzione della sequenza originale.

Il sequenziamento con metodi di terza generazione, quali *Nanopore* o *MinION*, si basano sul sequenziamento a singola molecola [5]. Pur mostrando un'efficienza maggiore rispetto ai metodi precedenti, sono attualmente ancora in fase di sviluppo.

1.2 Stima della dimensione del genoma

Il problema della misurazione della dimensione del genoma costituisce un argomento di interesse, perché oltre a fornire informazioni sulla sua evoluzione [10], permette di approssimare la quantità di dati che verranno prodotti nel sequenziamento e di valutare la complessità delle sequenze assemblate [11].

Di seguito sono elencate le due principali tipologie di metodi per la stima della dimensione del genoma, basate rispettivamente su approcci sperimentali o computazionali.

1.2.1 Metodi sperimentali

Inizialmente la ricerca scientifica ha cercato di stimare la dimensione del genoma con approcci biochimici, come ad esempio i metodi *Feulgen photometry* o *flow cytometry*. Tali metodologie però, oltre ad essere costose dal punto di vista economico e poco efficienti, devono basarsi su genomi specifici di riferimento [11, 12].

1.2.2 Metodi computazionali

Grazie all'aumento delle capacità computazionali, sono stati sviluppati metodi che possono calcolare la lunghezza del genoma utilizzando i dati ricavati dall'assemblaggio di shotgun. Dato che i dati assemblati sono di solito incompleti, è più conveniente cercarne di stimare la dimensione, utilizzando ad esempio i *k-mer*.

In questa trattazione verranno analizzati e confrontati vari approcci algoritmici che utilizzando i k-mer compiono la stima della dimensione del gemoma.

1.3 Sequenze di lunghezza k: i k-mer

I *k-mer* sono tutte le sottostringhe di lunghezza k presenti nella sequenza del genoma [13]. Si prenda ad esempio la sequenza descritta in 1.1.

$$AGATTCGC \tag{1.1}$$

I k-mer di lunghezza $k = 1$ saranno le quattro basi che formano la sequenza: G, T, A, C. Ponendo invece $k = 2$, i k-mer trovati saranno tutte le coppie formate da due basi: $AG, GA, AT, TT, TC, CG, GC$.
Allo stesso modo, per $k = 3$ le sequenze sono: $AGA, GAT, ATT, TTC, TCG, CGC$.

Il listing 1.1 nella pagina seguente mostra una semplice implementazione in pseudocodice per determinare i k-mer di lunghezza `k`, iterando la sequenza di input `seq` e dando in output tutte le sottostringhe di lunghezza `k` presenti in essa.

```
procedure k-mers(seq, k)
  lunghezza = length(seq)
  arr = array di L - k + 1 stringhe vuote

  // La sequenza iniziale viene iterata,
  // salvando il k-mer n-esimo nell'array di output
  for n = 0 to L - k + 1 escluso do
    arr[n] = sottostringa di seq da seq[n] a seq[n+k] escluso

  return arr
```

Listing 1.1: Algoritmo in pseudocodice per la costruzione dei k-mer.

Dato che il numero di k-mer aumenta esponenzialmente all'aumentare del parametro `k`, sono necessari algoritmi più complessi per il calcolo dei k-mer, come ad esempio *Jellyfish* [13] o *KMC2* [14].

1.3.1 K-mer profile

Il *k-mer profile*, detto anche *k-mer spectrum*, rappresenta un indicatore della complessità del genoma preso in esame. Date in input le letture shotgun del genoma, esso mostra il numero di volte che ogni k-mer viene trovato rispetto la quantità di k-mer distinti presenti, ovvero la molteplicità di ciascun k-mer nella sequenza rispetto il numero di k-mer con quella molteplicità [15]. Un esempio di k-mer profile è mostrato dalla figura 1.2 tratta da [16], in cui si può notare come la natura del genoma influenzi direttamente il grafico in ogni sua componente.

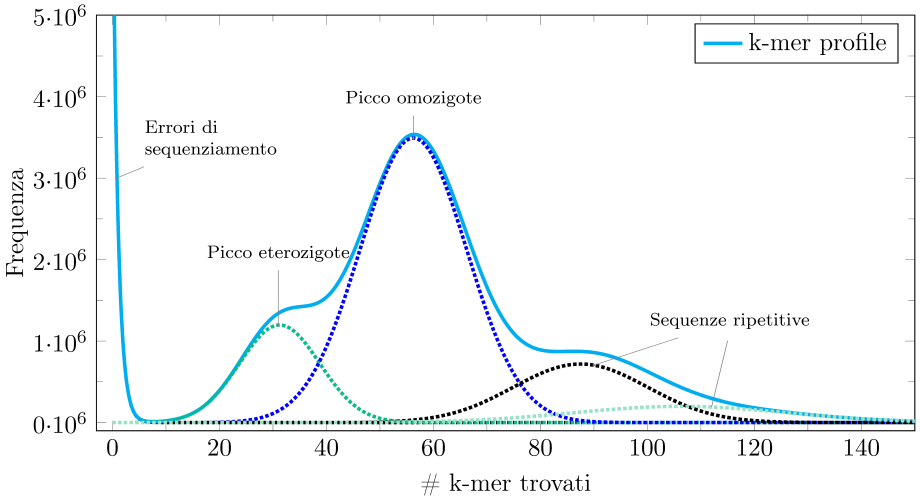

Figura 1.2: Composizione di un k-mer profile.

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

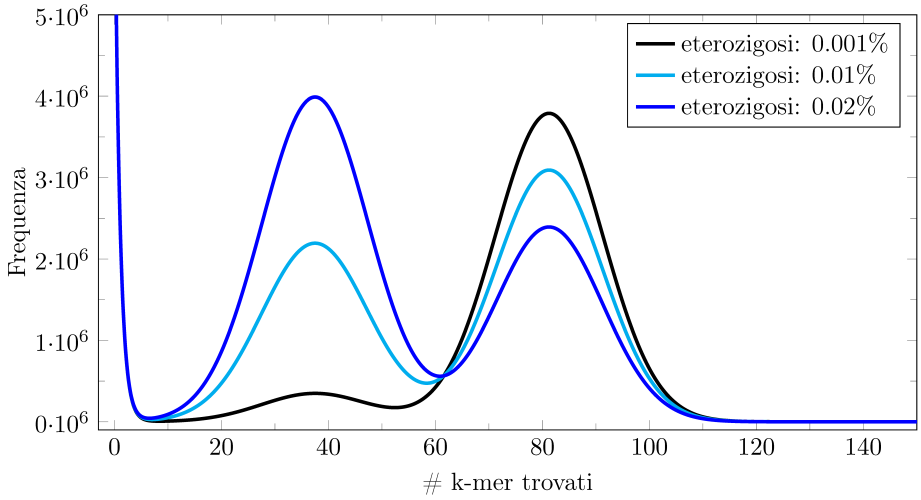

Figura 1.3: Variazione del grafico del k-mer profile al variare dell'eterozigosi.

principali [11]. 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 [16].

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 1.3 tratta da [17] 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.

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 [17]. Il programma per determinare tali caratteristiche utilizza il k-mer profile del genoma preso in esame, descritto nella sezione 1.3.1 a pagina 4.

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, 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 qua-

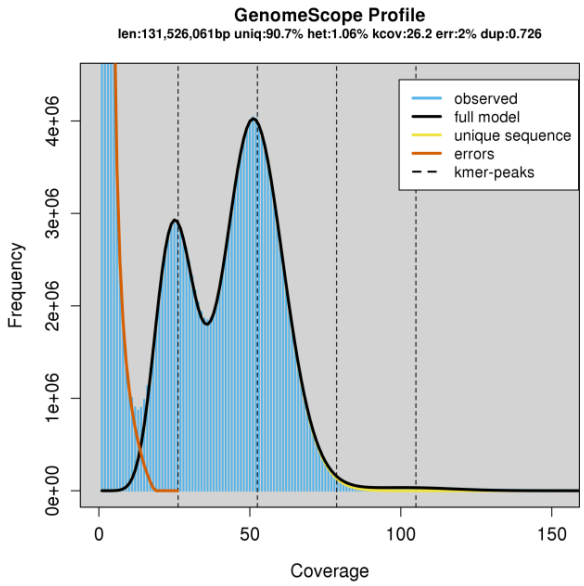

Figura 2.1: TODO + TODO reference a figura.

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* [11] 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* [18, 19]).

3.1 Algoritmo

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 ξ, la deviazione standard ω, l'asimmetria α 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.

Dato un genoma aploide con G basi, il numero di k-mer possibili sarà $G - k + 1$. Ponendo C la copertura media 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 è descritta dall'equazione 3.1.

$$N = C * (G - K + 1) \tag{3.1}$$

Posta la dimensione del genoma molto maggiore del numero di basi utilizzate $G \gg k$, l'equazione 3.2 approssima la dimensione totale del genoma in analisi.

$$G \approx N/C \tag{3.2}$$

A partire sia dal profilo reale che dal modello stimato, il programma calcola quindi il numero totale di k-mer trovati N e la copertura media dei k-mer C, per poi calcolare la dimensione del genoma attraverso l'equazione 3.2.

Capitolo 4

GCE

Capitolo 5

MGSE

Il programma *Mapping-based Genome Size Estimation* (*MGSE*) stima la dimensione del genoma attraverso la mappatura delle letture a un assembly ad alta contiguità [12]. Lo script è open-source e scritto in Python, e processa le informazioni sulla copertura delle letture di input restituendo la dimensione stimata del genoma.

5.1 Algoritmo

Posto che le letture siano distribuite equamente sull'intera sequenza del genoma, il programma TODO

Glossario

Distribuzione normale asimmetrica TODO Una distribuzione normale asimmetrica.. 9

Distribuzione binomiale negativa TODO Una distribuzione.. 7

Distribuzione di Poisson TODO Una distribuzione.. 4

Locus TODO Una distribuzione.. 5

Rapporto di eterozigosi TODO Una distribuzione.. 5

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] 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.

[3] F. Sanger et al. «Nucleotide sequence of bacteriophage ϕX174 DNA». In: *Nature* 265.5596 (feb. 1977), pp. 687–695. ISSN: 1476-4687. DOI: 10.1038/265687a0.

[4] A. M. Maxam e W. Gilbert. «A new method for sequencing DNA.» In: *Proceedings of the National Academy of Sciences* 74.2 (feb. 1977), pp. 560–564. DOI: 10.1073/pnas.74.2.560.

[5] J. Shendure et al. «DNA sequencing at 40: past, present and future». In: *Nature* 550.7676 (ott. 2017), pp. 345–353. DOI: 10.1038/nature24286.

[6] R. Staden. «A strategy of DNA sequencing employing computer programs». In: *Nucleic Acids Res* 6.7 (giu. 1979), pp. 2601–2610.

[7] B. Wajid e E. Serpedin. «Do it yourself guide to genome assembly». In: *Briefings in Functional Genomics* 15.1 (nov. 2014), pp. 1–9. ISSN: 2041-2649. DOI: 10.1093/bfgp/elu042.

[8] C. Noune. «Dynamics, diversity and evolution of Baculoviruses». Tesi di dott. Queensland University of Technology, 2017. DOI: 10.5204/thesis.eprints.113154.

[9] J. M. Heather e B. Chain. «The sequence of sequencers: The history of sequencing DNA». In: *Genomics* 107.1 (gen. 2016), pp. 1–8.

[10] T. R. Gregory. «Synergy between sequence and size in Large-scale genomics». In: *Nature Reviews Genetics* 6.9 (set. 2005), pp. 699–708. ISSN: 1471-0064. DOI: 10.1038/nrg1674.

[11] 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.

[12] B. Pucker. «Mapping-based genome size estimation». In: *bioRxiv* (apr. 2019). DOI: 10.1101/607390.

[13] 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.

[14] 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.

[15] D. Mapleson et al. «KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies». In: *Bioinformatics* 33.4 (feb. 2017), pp. 574–576.

[16] 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.

[17] 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.

[18] 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).

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