### Elementi di Bioinformatica

### Gianluca Della Vedova

Univ. Milano-Bicocca http://gianluca.dellavedova.org

26 novembre 2018

Grafi di assemblaggio

Gianluca Della Vedova Elementi di Bioinformatica

# Assemblaggio di genomi

## Tecnologie

- Porzioni di genoma chiamate read
- 50–10000bp (base pairs)
- spesso in coppie (mate pairs)
- o posizione originaria ignota

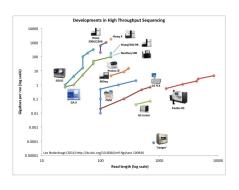
#### Obiettivo

1/1

Ricostruire il genoma: circa 3 miliardi bp

Gianluca Della Vedova Elementi di Bioinformatica 2/1

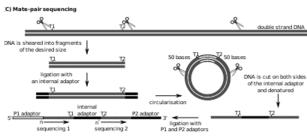
# Evoluzione tecnologica



Gianluca Della Vedova

Elementi di Bioinformatica

Mate pairs



Gianluca Della Vedova

Elementi di Bioinformatica

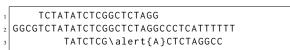
Regola 1

Suffisso di una read può essere prefisso di un'altra read: overlap

### Overlap — sovrapposizione



### Probabile motivo



Errore oppure organismi diploidi

Gianluca Della Vedova Elementi di Bioinformatica

# Grafo di overlap

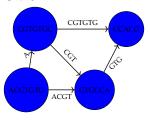
### Read

ACGTGTG

CGTGTGC GTGCCA CCACG

Arco fra tutte le coppie di read con overlap abbastanza lungo

#### Grafo



Gianluca Della Vedova Elementi di Bioinformatica

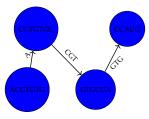
# String Graph

### Read

CGTGTGC GTGCCA CCACG

Si rimuovono gli archi transitivi dal grafo di overlap

#### Grafo



## Shortest superstring

#### Istanza

Insieme  $S = \{s_1, \ldots, s_n\}$  di stringhe

### Soluzioni ammissibili

Superstring T di S. Ogni  $s_i$  è sottostringa di T

### Funzione obiettivo

T è il genoma assemblato,  $\mathcal{S}$  le read

#### Problema

Regioni ripetute

Gianluca Della Vedova Elementi di Bioinformatica Gianluca Della Vedova Elementi di Bioinformatica

# Algoritmo ingordo

### Algoritmo

- Fondere le due stringhe con massimo overlap
- ② Finchè non rimane una stringa sola

Gianluca Della Vedova Elementi di Bioinformatica

# Esempio: a\_long\_long\_long\_time

- ng\_lon \_long\_ a\_long long\_l ong\_ti ong\_lo long\_t g\_long g\_time ng\_tim
- a ng\_time ng\_lon \_long\_ a\_long long\_l ong\_ti ong\_lo long\_t g\_long
- a ng\_time g\_long\_ng\_lon a\_long long\_l ong\_ti ong\_lo long\_t
- ng\_time long\_ti g\_long\_ng\_lon a\_long long\_long\_lo
- ng\_time ong\_lon long\_ti g\_long\_ a\_long long\_l
- ong\_lon long\_time g\_long\_ a\_long long\_l
- long\_lon long\_time g\_long\_ a\_long
- long\_lon g\_long\_time a\_long
- ② long\_long\_time a\_long a\_long\_long\_time

Gianluca Della Vedova Elementi di Bioinformatica

10/1

# Problema del commesso viaggiatore (TSP)

#### Istanza

Grafo orientato  $G = \langle V, A \rangle$ , con archi pesati  $w : A \mapsto \mathbb{Q}^+$ 

### Soluzioni ammissibili

Permutazione  $\Pi = \langle \pi_1, \dots, \pi_n \rangle$  of V

#### Funzione obiettivo

 $w(\pi_n, \pi_1) + \sum_{i=1}^n w(\pi_i, \pi_{i+1})$ 

- Una soluzione è un percorso che tocca ogni città esattamente una volta e torna al punto di partenza
- Il costo è il peso totale di tutti gli archi attraversati
- NP-completo, ma risolvibile in pratica

Gianluca Della Vedova Elementi di Bioinformatica

## Superstringa più corta e TSP

#### Similarità

1 read = 1 città

#### Differenze

- assemblaggio ≠ ciclo
- lunghezza stringa ≠ costo percorso TSP

# Proprietà

 $|S|=\sum_{i=1}^n|s_i|-\sum_{i=1}^{n-1}|ov(s_i,s_{i+1})|,$ dove $ov(\cdot,\cdot)$ è la lunghezza della sovrapposizione fra le stringhe

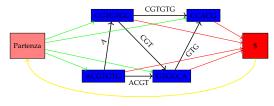
Gianluca Della Vedova Elementi di Bioinformatica

## Grafo di overlap — TSP





### Grafo



Gianluca Della Vedova Elementi di Bioinformatica

13/1

11/1

9/1

# Overlay — Layout — Consensus

## Passi

- Overlap: calcolare le sovrapposizioni e costruire il grafo. Usare suffix array (esatto) o programmazione dinamica
- 2 Layout: Fondere i cammini per ottenere i contigs. Le ripetizioni (branching nodes) vengono rimosse.
- 3 Consensus: calcola i nucleotidi

Gianluca Della Vedova Elementi di Bioinformatica

14/1

## Reverse and complement

- Non si conosce lo strand
- Versione canonica (minima fra x e revcomp(x)
- o complica il calcolo degli overlap

### **SBH**

# DNA array

- Tecnologia vecchia
- Per ogni k-mero, si conosce se appare nel genoma
- $k \approx 8$

# Procedura

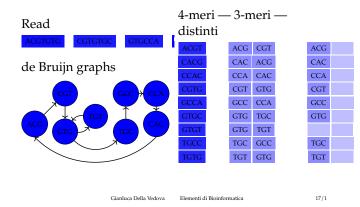
- Ogni k-mero viene diviso in (k − 1)-meri
- ② Un vertice per ogni (k − 1)-mero
- Un arco per ogni k-mero

### Adesso

Stessa procedura, a partire dai read

Gianluca Della Vedova Elementi di Bioinformatica 15/1 Gianluca Della Vedova Elementi di Bioinformatica 16/1

# Grafo di de Bruijn



### Grafi Euleriani

### Definizione

Sia  $G = \langle V, A \rangle$  un grafo orientato. G è semi-euleriano se esistono due vertici s, t tali che  $N_G^-(s) = N_G^+(s) + 1$ ,  $N_G^-(t) = N_G^+(v) - 1$ , mentre per ogni altro vertice w,  $N_G^-(w) = N_G^+(w)$ .

#### Definizione

Sia  $G = \langle V, A \rangle$  un grafo orientato. G è euleriano se  $N_G^-(w) = N_G^+(w)$ . per ogni vertice.

### Teorema

Un grafo connesso  $G = \langle V, A \rangle$  ha un cammino euleriano se e solo se *G* è semi-euleriano. *G* ha un ciclo euleriano se e solo se *G* è euleriano.

Gianluca Della Vedova Elementi di Bioinformatica

# Altre fasi

- bubble popping
- tip removal

Gianluca Della Vedova Elementi di Bioinformatica

21/1

# Licenza d'uso

Quest'opera è soggetta alla licenza Creative Commons: Attribuzione-Condividi allo stesso modo 3.0. https://creativecommons.org/licenses/by-sa/4.0/ Sei libero di riprodurre, distribuire, comunicare al pubblico, esporre in pubblico, rappresentare, eseguire, recitare e modificare quest'opera alle seguenti condizioni:

- Attribuzione Devi attribuire la paternità dell'opera nei modi indicati dall'autore o da chi ti ha dato l'opera in licenza e in modo tale da non suggerire che essi avallino te o il modo in cui tu usi l'opera.
- Condividi allo stesso modo Se alteri o trasformi quest'opera, o se la usi per crearne un'altra, puoi distribuire l'opera risultante solo con una licenza identica o equivalente a questa.

# Problemi su grafi

### Ciclo Euleriano

- 1 Un assemblaggio valido è un cammino che attraversa ogni arco esattamente una volta
- 2 Cammino Euleriano

#### Ciclo Hamiltoniano

- É un cammino che attraversa ogni vertice esattamente una
- ② Caso particolare di TSP

#### Confronto

Qual è più difficile da risolvere?

Gianluca Della Vedova Elementi di Bioinformatica

18/1

22/1

### Grafi Euleriani 2

#### Teorema

Sia  $G = \langle V, A \rangle$  un grafo semi-euleriano e sia P un cammino da s a t. Sia G<sub>1</sub> il grafo ottenuto da G togliendo tutti gli archi di P. Allora  $G_1$  è euleriano.

#### Teorema

Sia  $G = \langle V, A \rangle$  un grafo euleriano e sia C un ciclo di G. Sia  $G_1$  il grafo ottenuto da G togliendo tutti gli archi di C. Allora G<sub>1</sub> è euleriano.

Gianluca Della Vedova Elementi di Bioinformatica

# Scaffolding

- Fondere contigs in scaffolds
- usando mate pairs
- anche con revcomp

Gianluca Della Vedova Elementi di Bioinformatica

Gianluca Della Vedova Elementi di Bioinformatica