Progetto ASD 2018/19

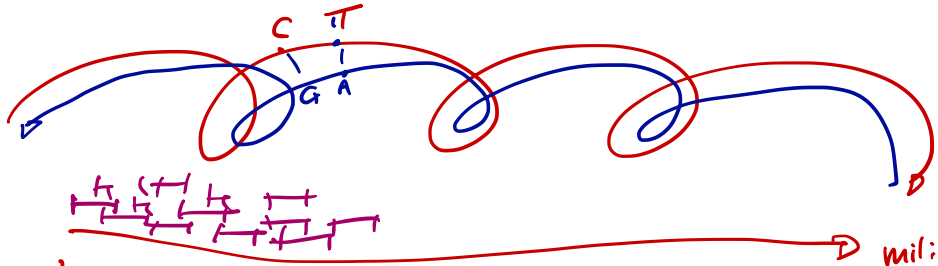
DNA → "sequenziazione"
sequencing

DNA → sequenza sull'alfabeto

$\Sigma = \{A, C, G, T\}$
basi

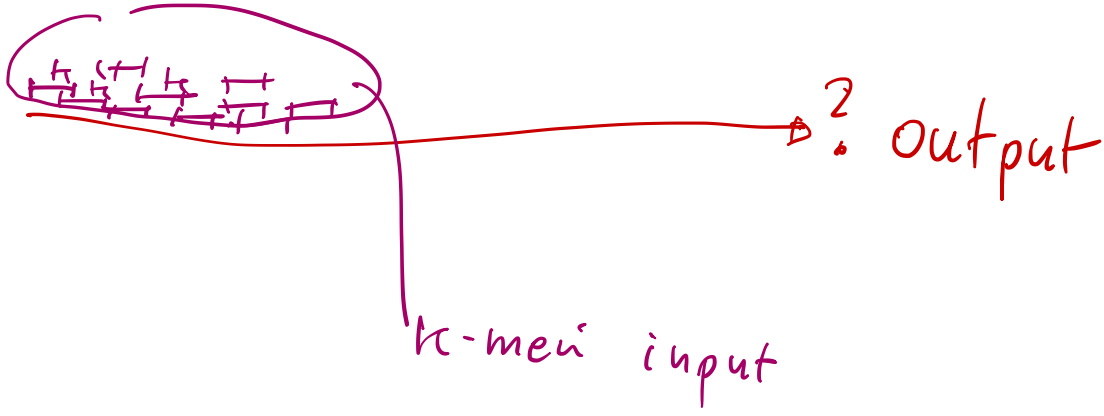
HTS = High-throughput sequencing

k-mero = sequenza di k simboli di Σ (N = simbolo incerto)
k-mer



milioni di basi

$k \sim 10^2 - 10^3$ ordine # basi



Nostra terminologia:

INPUT = file contenente un numero elevato di stringhe sull'alfabeto Σ, ciascuna di lunghezza k

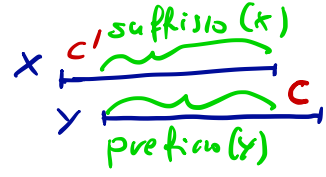
Scopo del progetto:

- grafo di de Bruijn: Σ, k

$G = (V, E)$ orientato

$V = \Sigma^k$

$E = \{ (x,y) : x,y \in \Sigma^k \text{ e } \begin{matrix} \text{suffisso}(x) = \\ \text{prefisso}(y) \\ \text{di lunghezza } k-1 \end{matrix} \}$

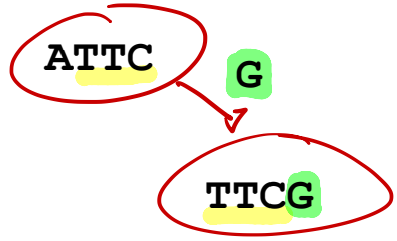


lunghezza $k-1$

$L : E \to \Sigma$

$L(x,y) = c$ t.c. $\begin{matrix} x = c'\alpha \\ y = \alpha c \end{matrix}$, $|\alpha| = k-1$

Esempio:



$K=4$

Grafo regolare in cui ogni nodo ha $|\Sigma|$ archi uscenti e $|\Sigma|$ archi entranti

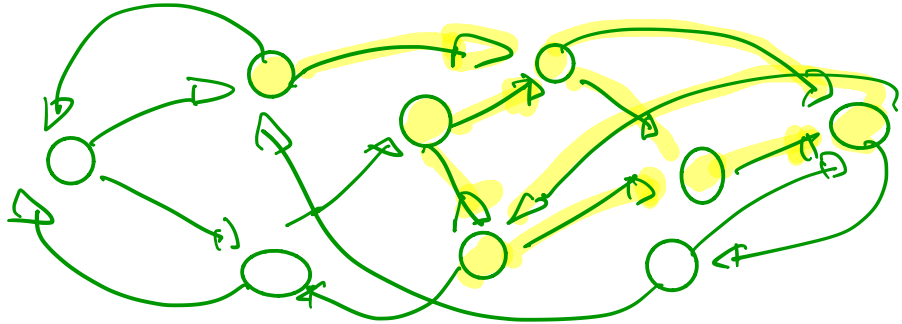
Dato $G=(V,E)$ e un sottoinsieme $V' \subseteq V$

sottografo indotto $G[V'] = (V', E[V'])$

(nota: alcuni nodi potrebbero essere isolati e quindi: possono essere scartati a priori)

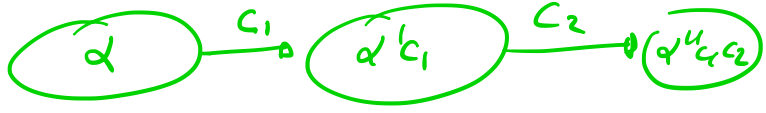
t.c. $E[V'] = \{(x,y) \in E : x,y \in V'\}$

$|\Sigma| = 2$



NOTA: V' è dato dal file di INPUT

Una tecnica comune di sequenziamento percorre un opportuno cammino (con nodi ripetuti) nel grafo, sovrapponendo i rispettivi k-mer.

es. 

$\alpha c_1 c_2$

$|\alpha| = k$
$|\alpha'| = k-1$
$|\alpha''| = k-2$

In realtà ci sono diverse complicazioni.

Task Dato il file di input:

(1) Costruire il grafo di de Bruijn sull'input

(2) Usare opportune strutture di dati per rispondere alle seguenti richieste (query):

2.a. Data una stringa P su Σ, di lunghezza arbitraria m, stabilire se P appare lungo un qualche cammino (il problema è interessante per $m > k$)

2.b. Data P su Σ, trovare il più lungo prefisso di P che soddisfa 2.a (chiaramente se P appare, abbiamo che è lei la risposta)

2.c. Risolvere la 2.a dove P può avere un errore: uno dei simboli di P non corrisponde, ma gli altri sì. ATCC corrisponde al AGCC