

Elementi di Bioinformatica

Gianluca Della Vedova

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

10 ottobre 2019

Gianluca Della Vedova

Elementi di Bioinformatica

1/1

Trie

Trie

- Albero
- Query: parola \in dizionario
- archi etichettati
- Percorso radice-foglia = parola



Dizionario

ABRACADABRA

ARRAY

ABRA

Gianluca Della Vedova

Elementi di Bioinformatica

2/1

Trie

Terminatore

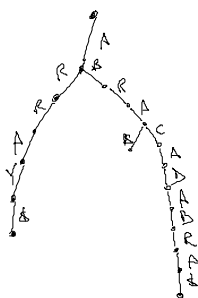
\$ non appartiene all'alfabeto

Dizionario

ABRACADABRA\$

ARRAY\$

ABRA\$



Gianluca Della Vedova

Elementi di Bioinformatica

3/1

Suffix tree

Definizione

- Trie compatto di tutti i suffissi di $T\$$
- Le etichette degli archi uscenti da x iniziano con simboli diversi
- suffissi \Leftrightarrow percorso radice-foglia

BANANA\$

Gianluca Della Vedova

Elementi di Bioinformatica

4/1

Suffix tree 2: Definizione

- foglie etichettate con posizione inizio suffisso
- $\text{path-label}(x)$: concatenazione etichette
- $\text{string-depth}(x)$: lunghezza $\text{path-label}(x)$
- Pattern matching = visita

Problemi

- Spazio $O(n^2)$
- Puntatori al testo (posizioni)
- Spazio $20n$ bytes

Gianluca Della Vedova

Elementi di Bioinformatica

5/1

Suffix array

Definizione

- Array dei suffissi in ordine lessicografico
- Posizioni iniziali del suffisso nell'array
- Spazio $4n$ bytes
- $\text{Lcp}[i]$: lunghezza prefisso comune $\text{SA}[i]$, $\text{SA}[i + 1]$

BANANA\$

i	1	2	3	4	5	6	7
SA	7	6	4	2	1	5	3
Lcp	0	1	3	0	0	2	-

Gianluca Della Vedova

Elementi di Bioinformatica

6/1

Da Suffix tree a Suffix array

- Visita depth-first di ST
- archi uscenti di ogni nodo in ordine lessicografico
- $\text{Lcp}[i] = \text{string-depth di } \text{lca}(i, i + 1)$

Gianluca Della Vedova

Elementi di Bioinformatica

7/1

Da Suffix array a Suffix tree

- $\text{Lcp} = 0$: partizione SA
- corrispondono ai figli della radice
- ricorsione prendendo i

Gianluca Della Vedova

Elementi di Bioinformatica

8/1

Suffix tree generalizzato

s_1 :
BANANA\$
 s_2 :
PANNA\$

Sottostringa comune più lunga di due stringhe

Due stringhe s_1 e s_2

- Suffix tree generalizzato = insieme di stringhe
- $ST(s_1\$s_2\$)$
- Nodo x con foglie di s_1 e s_2
- Sottostringa di s_1 e s_2
- $ST(s_1\$s_2\$)$
- Max string-depth

Licenza d'uso

Quest'opera è soggetta alla licenza Creative Commons:
Attribuzione-Condividi allo stesso modo 4.0.

(<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.