

Elementi di Bioinformatica

Gianluca Della Vedova

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

2 novembre 2020

Karp-Rabin

Alfabeto binario

- $H(S) = \sum_{i=1}^{|S|} 2^{i-1} H(S[i])$
- sliding window di ampiezza m su T
- $H(T[i+1 : i+m]) =$
 $= (H(T[i : i+m-1]) - T[i]) / 2 + 2^{m-1} T[i+m]$
- operazioni su bit
- $T[i : i+m-1] = P \Leftrightarrow H(T[i : i+m-1]) = H(P)$

Karp-Rabin: problema

Numeri troppo grandi

- Modello RAM: numeri $O(n + m)$
- $\text{mod } p$
- $H(T[i + 1 : i + m]) = ((H(T[i : i + m - 1]) - T[i]) / 2 + 2^{m-1} T[i + m]) \text{ mod } p$
- **NO**
- $2^{m-1} T[i + m] \text{ mod } p$ calcolato iterativamente, $\text{mod } p$ ad ogni passo

Karp-Rabin: falsi positivi

Possibili errori

- Falso positivo (FP): occorrenza non vera
- Falso negativo (FN): occorrenza non trovata
- $H(T[i : i + m - 1]) = H(P) \Leftrightarrow T[i : i + m - 1] = P$
- $H(T[i : i + m - 1]) \bmod p = H(P) \bmod p \Leftarrow T[i : i + m - 1] = P$

Karp-Rabin: falsi positivi

Probabilità di errore

$P[\#FP \geq 1] \leq O(nm/I)$ se il numero primo p è scelto fra tutti i primi $\leq I$

Valori di I

- $I = n^2m \Rightarrow P[\#FP \geq 1] \leq 2.54/n$
- $I = nm^2 \Rightarrow P[\#FP \geq 1] \in O(1/m)$

Abbassare probabilità di errore

Scegliere k primi casuali (indipendenti senza ripetizioni), cambiare primo dopo ogni FP

Las Vegas vs. Monte Carlo

Classificazione algoritmi probabilistici

- Monte Carlo:
 - Sempre veloce
 - Forse non corretto
 - Karp-Rabin
- Las Vegas:
 - Sempre corretto
 - Forse non veloce
 - Quicksort con pivot random

Controllo falsi positivi

L : posizioni iniziali in T delle occorrenze

Run

sequenza $\langle l_1, \dots, l_k \rangle$ di posizioni in L distanti al massimo $m/2$

- $d = l_2 - l_1$
- P semiperiodico con periodo d
- $P = \alpha\beta^{k-1}$, α suffisso di β
- ogni run occupa $\geq n$ caratteri di T
- ogni carattere di T è in max 2 run

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.