3/23/16 15-750

# String Matching Using FFT

Let $S, T$ be 2 strings of same length.

**Def** $H(S,T)$ = # of mismatches (Hamming Dist)

$\mathrm{Conv}_H(P,S) \equiv H$ a string of length $n-m+1$

$H_i = H(P, S_i^{m+i-1})$

---

Open: Find an $O(n \log m)$ alg for $\mathrm{Conv}_H(P,S)$

---

Assume $\Sigma \subseteq \mathbb{Z}$

$$\langle S,T \rangle = \sum_{i=1}^{n} S_i T_i \equiv S \odot T$$

**Def:** $\mathrm{Conv}_{\langle,\rangle}(P,S)$

Using FFT $\mathrm{Conv}_{\langle,\rangle}$ is $O(n \log n)$ time

Can you see $O(n \log m)$ time?

Consider $\langle S,T\rangle = \sum_{i=1}^{m}(S_i - T_i)^2$

<u>note!</u> $\langle S,T\rangle = 0$ iff $S = T$

$\therefore$ $Conv_{\langle,\rangle}$ solves string matching

---

$$Conv_{\langle,\rangle}(P,S)_i = \sum_{j=1}^{m}(P_j - S_{i+j-1})^2 = \sum_{j=1}^{m} P_j^2 - 2\sum_{j=1}^{m} P_j \cdot S_{i+j-1} + \sum_{j=1}^{m} S_{i+j-1}^2$$

$\sum P_j^2$ = constant

$\sum_{j=1}^{m} P_j S_{i+j-1}$ = FFT conv!

$\sigma_i = \sum_{j=1}^{m} S_{i+j}^2$ = $O(n)$ time using sliding window over all $i$.

$$\underbrace{\begin{array}{l} S_i^2 + \cdots + S_{i+m-1}^2 \\ \quad S_{i+1}^2 + \cdots + S_{i+m}^2 \end{array}}$$

ie $\sigma_{i+1} = \sigma_i - S_i^2 + S_{i+m}^2$

3/13

# String Matching with Wildcards

Input: no wildcards

$\Sigma$ alphabet $|\Sigma|$ large

$P$ pattern $|P| = m$

$T$ text $|T| = n$

---

$*$ $\equiv$ wildcard ($*$ matches everything)

$\Sigma = \{1, \dots, k\} \cup \{*\}$

---

<u>Recall</u> $\Sigma \subseteq \mathbb{Z}$

$$S \oplus T = \sum_{i=1}^{n} S_i \cdot T_i$$

Then $\text{Conv}_{\oplus}(P, T) \equiv C \qquad C_i = P \oplus T_i^{m+i-1}$

is $O(n \log m)$ Time using FFT.

Consider yet another product rule!

$$\langle S,T\rangle_* = \sum_{i=1}^{m} S_i T_i (S_i - T_i)^2 \quad \text{where } S_i, T_i \in \mathbb{R}$$

---

Suppose $\Sigma \subseteq \mathbb{Z}$

$* \equiv 0$

$\sigma \in \Sigma$ are distinct positive integers.

---

<u>Claim</u> $S$ matches $T$ with wild cards.

iff $\langle S,T\rangle_* = 0$

Pf $\quad S_i T_i (S_i - T_i)^2 = \begin{cases} 0 \text{ if } S_i \text{ or } T_i = * \text{ or } S_i = T_i \\ > 0 \quad \text{o.w.} \end{cases}$

---

<u>note</u> $S_i T_i (S_i - T_i)^2 = S_i^3 T_i - 2 S_i^2 T_i^2 + S_i T_i^3$

<u>Def</u> $S^{(k)} = (\alpha_1^k, \ldots, \alpha_m^k)$ where $S = (\alpha_1, \ldots, \alpha_m)$ $\alpha_i \in \mathbb{R}$

$$\langle S,T\rangle_* = \langle S^{(3)},T\rangle - 2\langle S^{(2)},T^{(2)}\rangle + \langle S,T^{(3)}\rangle$$

---

**Def** $\mathrm{Conv}_* \equiv \mathrm{Conv}_{\langle\,,\,\rangle_*}$

$$\mathrm{Conv}_*(P,S) = \mathrm{Conv}_\oplus(P^{(3)},S) - 2\,\mathrm{Conv}_\oplus(P^{(2)},S^{(2)}) + \mathrm{Conv}_\oplus(P,S^{(3)})$$

---

**Thm** $\mathrm{Conv}_*$ is $O(n \log m)$ time $|S|=n$ $|P|=m$

# Fast Hamming Dist for 0/1 strings

**Def** $\bar{H}(S,T)$ = # of matches $= \langle S,T\rangle_{\bar{H}}$

If S and T are zero-one then

**Def** $\langle S,T\rangle_0 \equiv$ # of zero-zero matches

**Def** $\langle S,T\rangle_1 \equiv$ # of 1-1 matches

**Note** $\langle S,T\rangle_{\bar{H}} = \langle S,T\rangle_0 + \langle S,T\rangle_1$

**Note** $\langle S,T\rangle_1 = \langle S,T\rangle_{\oplus}$ when S & T are zero-one strings

$$\bar{S} = (\bar{T}_1, \cdots \bar{T}_m) \quad \bar{T}_i = \begin{cases} 1 & \text{if } T_i = 0 \\ 0 & \text{o.w} \end{cases}$$

$$\langle S,T\rangle_0 = \langle \bar{S},\bar{T}\rangle_{\oplus}$$

**Thm** $\langle\ ,\ \rangle_H$ is $O(n \lg m)$ for fixed size alphabet.