PARAMETERIZED MATCHING

MOTIVATION: Software copying.

DEFINITION:

INPUT: Text $T = t_1, \dots, t_n$

Pattern $P = p_1, \dots, p_m$

OUTPUT: All location j for which $\exists$ bijection $b : \Sigma \to \Sigma$ such that

$$t_j = b(p_1)$$
$$t_{j+1} = b(p_2)$$
$$\vdots$$
$$t_{j+m-1} = b(p_m)$$

EXAMPLE:

$P = ABA$

$T = ABCDCECABAB$

matches:

$b(A) = C$, $b(B) = D$

$b(A) = C$, $b(B) = E$

$b(A) = A$, $b(B) = B$

$b(A) = B$, $b(B) = A$

Amir, Farach, Muthukrishnan (1994):

Parameterized matching can be solved in time $O(n)$.

(using automata methods)

PARAMETERIZED MATCHING WITH DON'T CARES.

Can be solved using convolutions.

In Time $O(|\Sigma|^2 n \log m)$.

For each pair $a, b \in \Sigma$ do:

$$\chi_a(T) \times \chi_b(P)^R$$

Record pairs of non-0 results at every location.

Time: $O(|\Sigma|^2 n \log m)$

$$\boxed{\text{Sort pairs by 1st, then by 2nd element}}$$

Time: $O(n\,|\Sigma|^2 \log |\Sigma|)$

But $|\Sigma| \le O(m)$ so

no worse than

$$O(|\Sigma|^2\, n \log m)$$

Can also be done by binsort in time $O(|\Sigma|^2)$ per text location, so:

Time: $O(|\Sigma|^2 n)$.

Every text location where

at most one pair $\langle a,b \rangle$ $\forall a \in \Sigma$

and at most one pair $\langle a,b \rangle$ $\forall b \in \Sigma$

is a parameterized match.

(Every text letter was matched to at most one pattern letter, and every pattern letter was matched to at most 1 text letter.)

In Time $O(|\Sigma| n \log m)$: Weimin Chen (1999)

Lemma: Let $a_1, a_2, \ldots, a_k \in \mathbb{Z}^+$ then $\left(\sum_{i=1}^{k} a_i\right)^2 \leq k\left(\sum_{i=1}^{k} a_i^2\right)$.

Equality holds iff $a_i = a_j \ \forall i, j$

Proof: We need the following:

Claim: $\forall a, b \in \mathbb{Z}^+$, $2ab \leq a^2 + b^2$

$2ab = a^2 + b^2$ iff $a = b$.

Proof of claim: w.l.o.g. let $b \geq a$

i.e. $b = a + x$

$$2ab = 2a(a+x) = 2a^2 + 2ax \leq (= \text{ iff } x = 0)$$

$$2a^2 + 2ax + x^2 = a^2 + (a+x)^2 = a^2 + b^2 \ \blacksquare$$

Now prove lemma by induction on k.

Base case: $k=1$

$$\left(\sum_{i=1}^{k} a_i\right)^2 = a_1^2 \leq 1 \cdot a_1^2 = k \cdot \sum_{i=1}^{k} a_i^2$$

"=" holds where all elements equal.

Ind. hyp:

$$\left(\sum_{i=1}^{k} a_i\right)^2 \leq k \sum_{i=1}^{k} a_i^2$$

with equality iff all a_i equal.

Prove for $\left(\sum_{i=1}^{k+1} a_i\right)^2$

$$\left(\sum_{i=1}^{k+1} a_i\right)^2 = \left(\left(\sum_{i=1}^{k} a_i\right) + a_{k+1}\right)^2 =$$

$$\left(\sum_{i=1}^{k} a_i\right)^2 + a_{k+1}^2 + 2\left(\left(\sum_{i=1}^{k} a_i\right)\left(a_{k+1}\right)\right) =$$

$$\left(\sum_{i=1}^{k} a_i\right)^2 + a_{k+1}^2 + \sum_{i=1}^{k}\left(2 a_i a_{k+1}\right) \leqslant$$

by claim, "=" iff $a_i = a_{k+1}$
$i = 1, \dots, k$

$$\left(\sum_{i=1}^{k} a_i\right)^2 + a_{k+1}^2 + \sum_{i=1}^{k}\left(a_i^2 + a_{k+1}^2\right) =$$

$$\left(\sum_{i=1}^{k} a_i\right)^2 + a_{k+1}^2 + k\, a_{k+1}^2 + \sum_{i=1}^{k} a_i^2 \leqslant$$

by ind. hyp. "=" iff $a_i = a_j$
$i, j = 1, \dots, k$

$$k\sum_{i=1}^{k} a_i^2 + \sum_{i=1}^{k} a_i^2 + (k+1)\,a_{k+1}^2 =$$

$$(k+1)\sum_{i=1}^{k} a_i^2 + (k+1)\,a_{k+1}^2 =$$

$$(k+1)\sum_{i=1}^{k+1} a_i^2 \quad \blacksquare$$

Use lemma for following algorithm:

Replace text alphabet by $\{1, \ldots, |\Sigma_T|\}$.

T^2 = array where $T^2[i] = (T[i])^2$

$\forall \sigma \in \Sigma_P$ do:

$$M \leftarrow T \times \chi_\sigma(P)^R$$

$$Q \leftarrow T^2 \times \chi_\sigma(P)^R$$

T' = array where $T'[i] = \begin{cases} 1 & T[i] \neq \emptyset \\ 0 & T[i] = \emptyset \end{cases}$

$$K \leftarrow T' \times \chi_\sigma(P)^R$$

$$KQ \leftarrow K \times Q$$

$$(KQ[i] = K[i]Q[i])$$

Note: $$M[i]^2 = \left(\sum_{i=1}^{k} a_i\right)^2$$

$$kQ[i] = k\left(\sum_{i=1}^{k} a_i^2\right)$$

where a_i are all text numbers that match σ.

Conclude: Every location i where $kQ[i] = M[i]^2$

has: all text numbers that match σ are same.

BUT: it could be that same text number ~~σ~~ matches two or more ~~text numbers...~~ σ's...

For every $\sigma \in \Sigma_p$:

record unique text number that matches it (if exists),

i.e. $\sqrt{\dfrac{Q[i]}{k[i]}}$

A location is a parameterized match with don't cares iff all recorded text numbers are **distinct**.

Time: $O(|\Sigma_p|\, n \log m)$.