PARAMETERIZED MATCHING

MOTIVATION: Software copying.

DEFINITION:

INPUT: Text $T = t_1, \ldots, t_n$
Pattern $P = \rho_1, \ldots, \rho_m$

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

$$t_j = b(\rho_1)$$
$$t_{j+1} = b(\rho_2)$$
$$\vdots$$
$$t_{j+m-1} = b(\rho_m)$$
EXAMPLE:

\[ P = ABA \]

\[ T = ABCDCCECAABAB \]

matches:

\[ b(A) b(B) \]
\[ c \]
\[ b(A) b(B) \]
\[ c \]
\[ b(A) b(B) \]
\[ c \]
\[ b(A) b(B) \]
\[ a \]
\[ b(A) b(B) \]
\[ c \]
\[ b(A) b(B) \]
\[ c \]

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:

$X_a(T) \times X_b(P)^R$

Record pairs of non-$O$ results at every location.

Time: $O(|\Sigma|^2 n \log m)$
Sort pairs by 1st, then by 2nd element

Time: $O(n |\mathcal{E}|^2 \log |\mathcal{E}|)$

But $|\mathcal{E}| \leq O(m^t)$ so

no worse than

$O(|\mathcal{E}|^2 n \log m)$

Can also be done by bin sort in time $O(|\mathcal{E}|^2)$ per text location, so:

Time: $O(|\mathcal{E}|^2 n)$. 
Every text location where at most one pair $<a, b>$ $\forall a \in \mathcal{Z}$ and at most one pair $<a, b>$ $\forall b \in \mathcal{Z}$ 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 $O(|E| n \log m)$: 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. $b = a + x$

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

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

- P.6 -
Now prove lemma by induction on $k$.

Base case: $k = 1$

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

"=" 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 $\left( \sum_{i=1}^{k+1} a_i \right)^2$
\[
(\sum_{i=1}^{k+1} a_i)^2 = \left(\sum_{i=1}^{k} a_i + a_{k+1}\right)^2 = \\
\left(\sum_{i=1}^{k} a_i\right)^2 + a_{k+1}^2 + 2\left(\sum_{i=1}^{k} a_i\right)(a_{k+1}) = \\
\left(\sum_{i=1}^{k} a_i\right)^2 + a_{k+1}^2 + \sum_{i=1}^{k} (2a_i a_{k+1}) \leq \\
\text{by claim, "=" iff } a_i = a_{k+1} \quad i=1,\ldots, k
\]

\[
\left(\sum_{i=1}^{k} a_i^2\right) + a_{k+1}^2 + \sum_{i=1}^{k} (a_i^2 + a_{k+1}^2) = \\
\left(\sum_{i=1}^{k} a_i^2\right) + a_{k+1}^2 + k a_{k+1}^2 + \sum_{i=1}^{k} a_i^2 \leq \\
\text{by ind. hyp. "=" iff } a_i = a_j \quad i,j=1,\ldots, k
\]
\[ k \sum_{i=1}^{k} a_i^2 + \sum_{i=1}^{k} a_i^2 + (k+1) a_{k+1}^2 = \]

\[ (k+1) \sum_{i=1}^{k} a_i^2 + (k+1) a_{k+1}^2 = \]

\[ (k+1) \sum_{i=1}^{k+1} a_i^2 \]
Use lemma for following algorithm:

Replace text alphabet by \( \Sigma_1, \ldots, |\Sigma_1| \).

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

\[ \forall \sigma \in \Sigma \text{ do:} \]

\[ M \leftarrow T \times X_{\sigma}(P)^R \]

\[ Q \leftarrow T^2 \times X_{\sigma}(P)^R \]

\( T' \) = array where \( T'[i] = \begin{cases} \Sigma_1 & T[i] \neq \emptyset \\ \emptyset & T[i] = \emptyset \end{cases} \)

\[ k \leftarrow T' \times X_{\sigma}(P)^R \]

\[ KQ \leftarrow k \times Q \]

\( (KQ[i] = k[i]Q[i]) \)
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 \( \sigma \).

Conclude: Every location \( i \) where \( kQ[i] = M[i]^2 \) has all text numbers that match \( \sigma \) are same.

BUT: it could be that same text number \( \sigma \) matches two or more text numbers... \( \sigma \)'s...
For every $\sigma \in \Sigma_p$:

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

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

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

Time: $O(1|\Sigma_p| \times n \log m)$. 