DETERMINISTIC SAMPLING

IDEA:
Find a small sample of \( P \) such that \( \forall i \)

If the sample positions do not match then no occurrence of \( P \) in \( T \).
(Easy to satisfy—any position can do that)

BUT
If the sample positions match
Then

\[ T \]

\[ p \]

\[ i \]

Guaranteed a large area
where pattern cannot start-dead zone

\[ \text{large} = \frac{m}{c} \]

where \( c \) is
some constant.
PATTERN MATCHING ALGORITHM

For $i = 1$ to $n - m$ do
    check sample
    If positive then set location $i$ as candidate
end For

For every candidate, kill all candidates in its dead zone.

For remaining candidates - verify in naive way whether pattern occurs.

Time: $O(ns)$: "For" loop, where

$s = \text{sample size.}

\begin{align*}
O\left(\frac{m \cdot n}{\left(\frac{3}{c}\right)}\right) &= O(nc) = O(n) \\
\text{c is a constant}
\end{align*}

for verification

ds-3
VISHKIN (1990):

It is possible to construct a sample of size \( \log m \) for non-periodic patterns.

**Definition:** \( P \) is periodic if

\[
P = U^v U^r
\]

where \( v > 1 \) and \( U^r \) is a prefix of \( U \).

\( U \) is a period of \( P \).
EXAMPLE:

ABCAB is not periodic

ABCABCAB is periodic

ABC ABC ABC ABC ABC ABC ABC ABC A

period 1

period 2

period 3

period 4
PERIODICITY LEMMA: (Fine & Wilf)

Let \( U_1, U_2 \) be periods of \( P \)
Let \( |U_1| = x_1, \ |U_2| = x_2 \).

Then \( P \) has a period \( U_3 \), where
\[ |U_3| = \gcd(x_1, x_2) \]

Proof: Generalize the concept of period to allow \( \nu \geq 1 \)
\((P = UU^\nu \text{ where } U^\nu \text{ is a non-empty prefix of } U, \text{ makes } U \text{ a period})\).

Alternate view of periodicity:
Let \( |P| = m, \ |U| = x \)
\[ \forall i \quad 1 \leq i \leq m-x \]
\[ P[i] = P[i+x] \]
Claim: If $b$ and $c$ are co-prime then $b-c$ and $c$ are co-prime (for $b > c$).

Proof: \[ b-c = x \cdot w \]
\[ c = x \cdot z \quad x \neq 1 \]
\[ b = x \cdot w + x \cdot z = x \cdot (w+z) \]
contradicting co-primality of $b/c$.

Conclude: \[ \gcd(p, q) = \gcd(p-q, q) \]
(for $pq$).

Proof: \[ p = a \cdot b \]
\[ q = a \cdot c \quad \text{where } b, c \text{ co-prime.} \]
\[ p-q = a \cdot (b-c) \]
By claim $(b-c)$ and $c$ are co-prime.
Return to Proof of Periodicity Lemma:

By induction on $|P|$. $|P|=1$ is obvious.

Assume lemma true for strings of length less than $n$. Prove for $|P|=n$.

\[
\begin{array}{c}
U_1 \\
U_2 \\
U_2 \\
U_3 \\
U_2 \\
\end{array}
\]

Consider $U_1$.

$U_1[i] = U_1[i+x_2]$ \quad $\forall 1 \leq i \leq x_1-x_2$

But because $P$ has period $U_2$ of length $x_2$

$P[i] = P[i+x_1] = P[i+x_1-x_2]$ \quad $\forall 1 \leq i \leq m-x_1$.
Conclude: $U_1$ has period $(x_1 - x_2)$

We have: $U_1$ (of length $< 1\pi$)
  has period of length $x_2$
  and period of length $(x_1 - x_2)$

By ind hyp it has period of
  length $\gcd (x_1 - x_2, x_2)$.

But by claim, this equals
  $\gcd (x_1, x_2)$

So: $P: \boxed{U_1 U_1}$

also has period of length
  $\gcd (x_1, x_2)$. 
How is SAMPLE CONSTRUCTED?

Consider $P$ shifted and stacked $\frac{m}{2}$ times.

Because non-periodic, exists a column that has at least 2 different symbols. Choose column $j$ where symbol $a$ occurs $< \frac{1}{2}$ times. Discard all rows where not $a$ in column $j$. Repeat until only one row left.
EXAMPLE:

A B A B A B A B
A B A B B A A B A B
A B A B B A A B
A B A B B A A B
A B A B B A A B

SAMPLE:
7, A
8, B

Cancels from i-1 to i+3 except i.
SAMPLE SIZE: $O(\log m)$

Why? At every iteration, at least half of the rows are eliminated.

WHAT ABOUT PERIODIC PATTERNS?

1. Find smallest period.

2. Look for consecutive repetitions of period.