Knuth Morris and Pratt Algorithm

It exploits the idea of matching prefix with suffix in a pattern itself.

Key observation

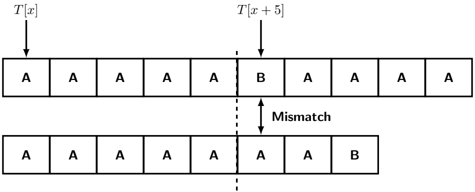
Suppose P has matched k characters with text $T[x, x+1, \ldots, x+k-1]$ and a mismatch occurs at $k+1$, i.e.,

$$P[1..k] = T[x..x+k-1], \text{ and } P[k+1] \neq T[x+k].$$

Then for any $0 < \ell < k$, if $T[x+\ell, \ldots, x+k-1]$ is not a prefix of P, P cannot occur in T at position $x+\ell$.

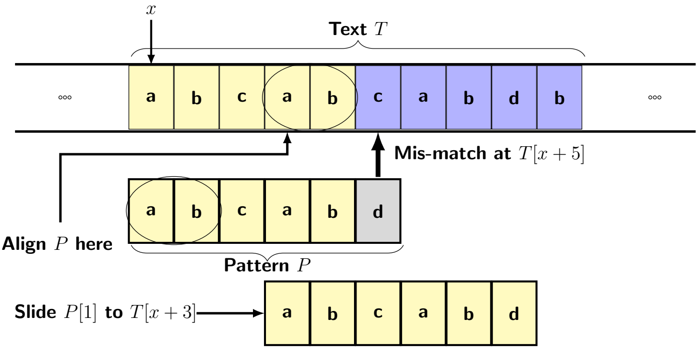
Example

- Consider following situation: P is matched with first five characters of T, and $T[x+5] \neq$ A.



- Shifting P to align $P[1]$ with any of the positions $T[x+1], T[x+2], T[x+3]$, or $T[x+4]$ will not obviously work.

Implication of the Observation



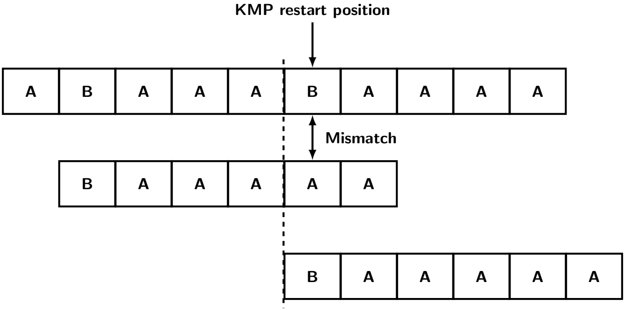
$k = 5$, and $T[x+3, x+4]$ is a prefix of $P[1, \ldots, 6]$. Matching can restart by aligning $P[1..2]$ with $T[x+3, x+4]$

Implication of the Observation

- In general, if first mismatch occurs after match k characters, matching restarts at the leftmost position $x + \ell$ such that $T[x + \ell, \ldots, x + k - 1]$ is a prefix of P
- Equivalently, if T is replaced by P, it also implies m is the smallest index such that:

$$P[\ell + 1, \ldots, k] \text{ is a prefix of } P[1..k].$$

Summary of Observation So Far



- In brute force: every position from $T[2]$ is a restart position.
- Since, none of the proper suffixes: $T[2..5]$, $T[3..5]$, $T[4..5]$, and $T[5..5]$ is a prefix of $P[1..5]$.
- So, matching can only restart at $T[6]$, i.e., after the border.

Knuth Morris Pratt

- The restart position is determined only with respect to already matched positions of T and P.
- This implies that the suffixes of matched portion of T (before the border) are also suffixes of matched part of P.
- Hence, the restart position in a text can be viewed with respect to P itself.
- The underlying idea is: whether any proper suffix of current position of P is a proper prefix of P.

Some Definitions

Definition (*Prefix*)

A prefix of x is a substring u such that $u = x_0 x_1 \cdots x_k$ where $k \in \{0, \ldots, m-1\}$.

Definition (*Suffix*)

A suffix of x is a substring u such that $u = x_{m-k-1} \cdots P_{m-1}$ where $k \in \{0, \ldots, m-1\}$.

Definition (*Proper prefix/suffix*)

A proper prefix (suffix) u of x is called a proper prefix (suffix) respectively, if $u \neq x$, i.e., length of u is less than the length of x.

Examples of Prefix and Suffix

For example, consider the string "**ababa**".

- Its proper prefixes are: "ϵ", "**a**", "**ab**", "**aba**", and "**abab**".
- Its proper suffixes are: "ϵ", "**a**", "**ba**", "**aba**" and "**baba**".
- Only "**a**" and "**aba**" are prefixes that are also suffixes, "**aba**" being the longest.

More on Prefix and Suffix

Definition (*Border*)

A border of x is a substring u is both a proper prefix and a proper suffix of x.

- In other words, u is a border if $u = x_0 x_1 \cdots x_{b-1}$ and $u = x_{k-b} x_{k-b-1} \cdots x_{k-1}$, where $b \in \{0, \cdots, k-1\}$
- E.g., proper prefixes of string **abacab** are: ϵ, **a**, **ab**, **aba**, **abac**, **abaca**
- Proper suffixes are: ϵ, **b**, **ab**, **cab**, **acab**, **bacab**
- Borders are: ϵ, **ab** of widths 0 and 2 respectively.