02/12/08

# Boyer-Moore Proof

Proving that Boyer Moore runs in linear time


When running the algorithm
the pattern is matched to the text until a mismatch found and then shift the pattern to the right. The goal is to shift by the least amount of characters.

Definitions

**|α|** - period

**Periodic string** S = α α α α.... ( $\alpha^i$)
many strings are not fully periodic

**Semi-periodic** S = suf( α) $\alpha^i$

e.g.


**Prefix semi-periodic** S = $\alpha^i$pref( α)

Every semi-periodic string is also prefix semi-periodic. A is different, but both definitions work for such a string.

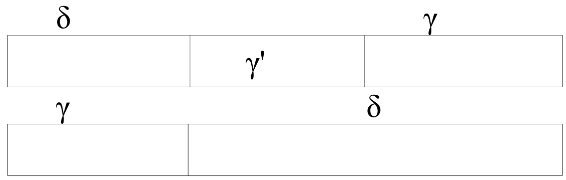
Lemma:

$S = \delta\gamma = \gamma\delta \quad \Rightarrow \quad \delta = \alpha^i,\ \gamma = \alpha^j$

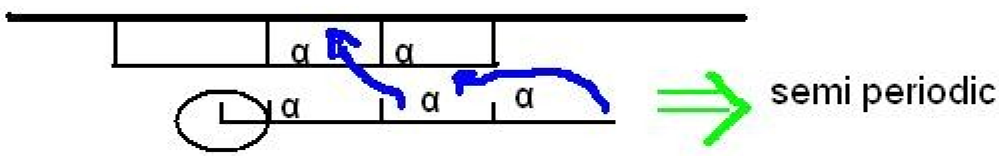
assume $|S| = n$ and $|\delta| > |\gamma|$

$\delta\gamma = \gamma\delta$

$\cancel{\delta'}\gamma'\gamma = \gamma\gamma'\cancel{\delta'}$ => $\gamma'\gamma = \gamma\gamma'$ => by induction δ is periodic so γ is also periodic

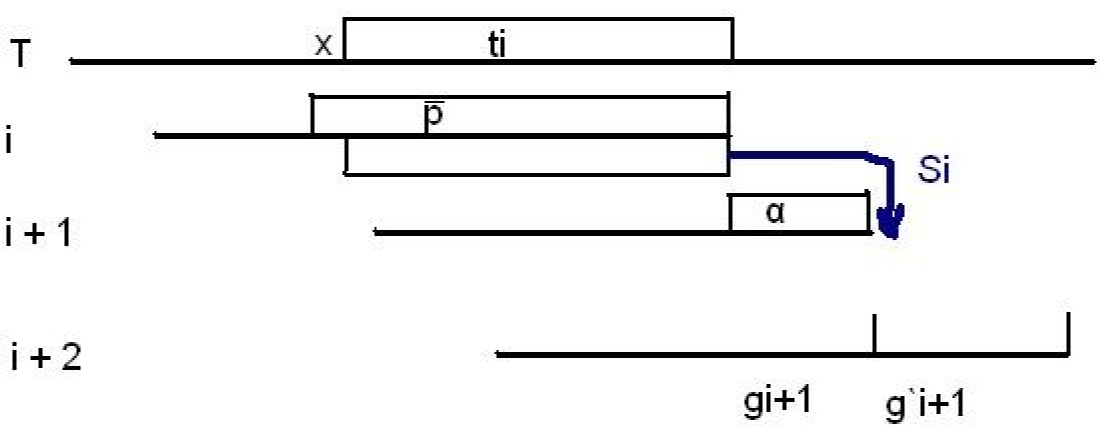



If P matches at positions p and p' in text and $p - p' < |p|/2$ then p is semi-periodic with period $p' - p$




<u>Definitions</u>

$t_i$ – set of characters that were matched at phase i
p – suffix of pattern that contains both $t_i$ and one more mismatched character $|p| = |t_i| + 1$
$S_i$ – # of characters that I jumped at phase i
β is the smallest possible period of α
α – $\alpha = \beta^l$ – smallest β such that $\alpha = \beta^l$
$g_{i+1}$ – # of characters matched in phase i + 1 not for the first time




$|t_i| + 1 = g_{i+1} + g'_i$

We will prove that $g_i < 3S_i$

$$\sum_i ( g_i + g'_i) \le m + \sum_i g_i \le m + 3\sum_i S_i \le m + 3m = 4m$$

if $S_i \ge (|t_i| + 1)/ 3$ than $g_i < 3S_i$ trivially

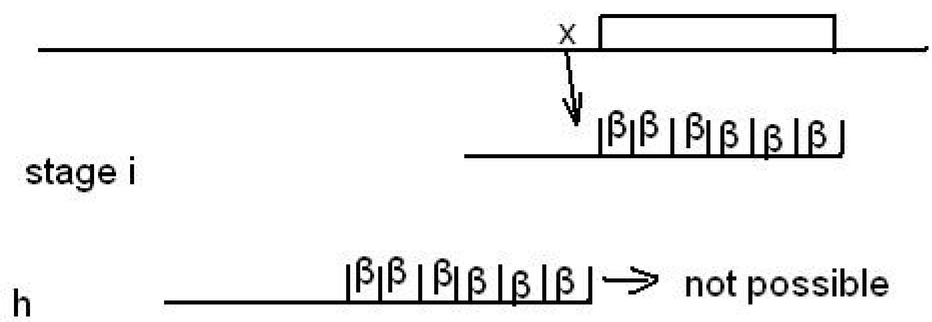
assume $S_i \ge (|t_i| + 1)/ 3$

I

If $S_i \ge (|t_i| + 1)/ 3$ then p & $t_i$ are semi-periodic with period $\alpha$. The proof is the same as Lemma (shifting strings)

II

At stage h < i end of P cannot coincide with boundary of β unit

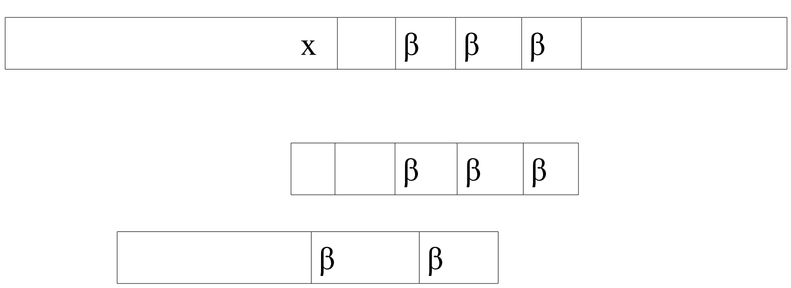



we know that after stage h we shifted pattern somewhere.
We have two possibilities:

1. Pattern matched the boundary of β -> clearly we could not shifted beyond i -> this option is not possible
2. Shifted such that we hit somewhere inside β boundary -> not possible either since β is the smallest possible shift and if such shift happened it would contradict that β is the smallest.

III

At any stage h < i "work < | β| in other words
th overlaps $t_i < |\beta|$

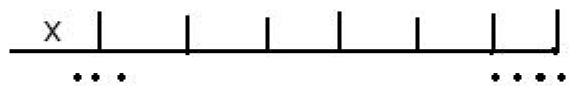



=> this implies that β is not smallest => contradiction => at any stage our work does not overlap

IV

At stage $h < i$ the rightmost end of pattern can only line up with the rightmost $|\beta| - 1$ characters of $t_i$ or leftmost $|\beta|$ characters of $t_i$.

We prove that it is impossible to escape the boundaries of $\beta$.




We show that $g_i < 3\beta$, we know that $\beta \le S_i \Rightarrow g_i \le 3S_i$

# of characters I saw in past is bounded by shifts I do and # of shifts is bounded by m $\Rightarrow g_i \le m$