

Processing of interval and reporting repeats

First, let's define following functions

$Children((l, i, j))$ - set of all children intervals of lcp-interval (l, i, j)

$Pos((l, i, j)) = \{suftab[r] | i \leq r \leq j\}$ - set of all positions in lcp-interval (l, i, j)

$Pos_c((l, i, j)) = \bigcup_{I \in Children((l, i, j))} Pos(I)$ - set of all positions in children intervals

$Pos'((l, i, j)) = Pos((l, i, j)) - Pos_c((l, i, j))$

$LC(p) = S[p-1]$ if $p > 0$ - returns left context of position p

$LC(0) = \emptyset$

$PartByLC(P, a) = \{p | p \in P \wedge LC(p) = a\}$ - extracts only positions from set P with left context $a \in \Sigma \cup \emptyset$.

Let (l, i, j) be an lcp-interval with $l \geq minLength$ we're about to process.

We divide $Pos'((l, i, j))$ into disjoint and possibly empty sets P'_a according to the characters left to each position (left context).

For each $a \in \Sigma \cup \emptyset$ let's define

$P_a = PartByLC(Pos((l, i, j)), a)$ and

$P'_a = PartByLC(Pos'((l, i, j)), a)$

For $\forall (a, b) \in (\Sigma \cup \{\emptyset\}) \times (\Sigma \cup \{\emptyset\}), a < b, \forall p_1 \in P'_a, \forall p_2 \in P_b$:

we report maximal repeat (p_1, p_2, l) .

It's left maximality comes from different left context of positions p_1 and p_2 and right maximality and match properties come from lcp-table property.

The reason why p_1 is taken only from P'_a is that children intervals already have been processed and their repeats have been reported.

Instead of reporting all repeats, the number of which may be quadratic in length of interval, we report for each interval a special report structure r with following information:

1. $r.lcp$ - lcp value
2. $r.p[a]$ - P'_a for all $a \in \Sigma \cup \emptyset$, organized in buckets indexed by a
3. $r.k$ - number of children reports
4. $r.child[i]$ - links to report structures of children intervals for all $1 \leq i \leq r.k$

By processing lcp-interval tree, we are this way creating match/repeat report tree.
This structure takes up $O(n)$ space.

Later, if we want to make queries of type: "given position p_1 , give me all positions p_2 , so that (p_1, p_2, l) , is maximal match/repeat with $l \geq minLength$ ", we can assign each position p pointer to it's report structure r and find all corresponding matches by following recursive function:

$$Find(p,r)=(\bigcup_{a \neq LC(p)} r.p[a]) \cup (\bigcup_{1 \leq i \leq r.k} Find(p,r.child[i]))$$