SUPPLEMENTARY MATERIAL FOR
Compact, Efficient and Unlimited Capacity: Language Modeling with Compressed Suffix Trees

Algorithm 6 Compute one-sided occurrence counts, $N^{1+}(\bullet\,\alpha)$ or $N^{1+}(\alpha\,\bullet)$ for pattern α

Precondition: node n in CST t matches α

function N1P(t, n, α)
  $o \leftarrow 1$
  if string-depth(n) $= |\alpha|$ **then**
    $o \leftarrow$ degree(n)
  return o

Algorithm 7 Compute backward occurrence counts, $N^{1+}(\bullet\,\alpha)$, using only forward CST

Precondition: v_{F} is the node in the forward CST t_{F} matching pattern α
Precondition: the CSA component, a_{F} of t_{F} is a wavelet tree

function N1PBACK1($t_{\mathrm{F}}, v_{\mathrm{F}}, \alpha$)
  $S \leftarrow$ int-syms(a_{F}, [lb(v_{F}), rb(v_{F})])
  return $|S|$

Function/Constant	Description	Complexity
SAS	sample rate of the suffix array. determines the number of jumps in $\mathcal{T}^{bwt}$ required before a suffix array value can be accessed	8 (in our exp.)
SA[i]	access the i-th element of the suffix array	$O(SAS \log \sigma)$
leaf(n)	tests if node n is a leaf of the t	$O(1)$
string-depth(n)	pattern length for the path from root to n (inclusive). Requires *SA*[i] access if leaf	$O(1)$ non-leaf; $O(SAS \log \sigma)$ leaf
edge(n, k)	k^{th} symbol in the edge label from root for node n. Requires *SA*[i] access	$O(SAS \log \sigma)$
degree(n)	number of child nodes under parent n	$O(\sigma/64)$
children(n)	list of all d child nodes under n	$O(\sigma/64 + d)$
back-search($[l, r], s$)	finds the node $v = [l', r']$ from parent node $\alpha = v' = [l, r]$ matching the pattern $s\alpha$. Requires 2 RANK operations on the wavelet tree	$O(\log \sigma)$
fw-search($[l, r], s$)	finds the node $v = [l', r']$ from parent node $\alpha = v' = [l, r]$ matching the pattern αs. Requires $\log \sigma$ accesses to *SA* and one LCP access	$O(SAS \log^2 \sigma + LCP_C)$
int-syms($a, [l, r]$)	finds the set of symbols $P(\alpha)$ preceeding pattern α matched by $[l, r]$; returns a list of tuples describing the bounds and the preceding symbol $\langle l, r, s \rangle$	$O(\lvert P(\alpha) \rvert \log \sigma)$

Table 1: Summary of CSA and CST functions used and their time complexity of inference. The above assumes that n or (equivalently) $[l, r]$ matches α in the CSA a and/or CST t.