
Addendum to the article

Jorma Tarhio: Searching Long Patterns with BNDM

In the case of very long patterns, the checking of match candidates is time consuming in SSB and RSSB. We tried a tune-up, which restricts the number of m -grams to be checked. The idea is to consider the offsets of the q -grams of the last segment. For every fingerprint of those q -grams, the smallest and largest offset is stored to arrays lo and hi during preprocessing. When a fingerprint s is found in the last segment during seaching, only m -grams with offsets between $lo[s]$ and $hi[s]$ are checked. The original algorithm checks all of a m -grams in the checking phase.

The tuned variants of SSB and RSSB are called SSBb and RSSBb. Algorithm 4 shows preprocessing of SSBb.

Algorithm 4: SSBb, preprocessing

1. if $m \leq w^2$ then $x \leftarrow 1$ else $x \leftarrow 0$
 2. $a \leftarrow \lfloor (m - q + 1 - x)/w \rfloor + x$; $r \leftarrow \lfloor (m - q + 1)/a \rfloor$
 3. if $r > w$ then $r \leftarrow w$
 4. $y \leftarrow w - r$; $b \leftarrow r \cdot a$; $i \leftarrow m - 1$
 5. for $i \leftarrow 0$ to $2^{q'} - 1$ do
 6. $B[i] \leftarrow 0$; $lo[i] \leftarrow a + 1$; $hi[i] \leftarrow a$
 7. while $i > m - b - q$ do
 8. for $j \leftarrow 1$ to a do
 9. $d \leftarrow F(P, i, q)$
 10. $B[d] \leftarrow B[d] \mid 1 \ll y$
 11. if $i > m - q - a$ then
 12. $j \leftarrow a - (i - (m - q - a))$
 13. if $lo[d] = a + 1$ then $lo[d] \leftarrow j$; $hi[d] \leftarrow j$
 14. if $j > hi[d]$ then $hi[d] \leftarrow j$
 15. $i \leftarrow i - 1$
 16. $y \leftarrow y + 1$
-

In the search algorithm (Algorithm 3), only lines 4 and 13 are changed for SSBb.

4. $s \leftarrow F(T, i, q)$; $d \leftarrow B[s]$
13. for $k = lo[s]$ to $hi[s]$ do

The variants SSBb and RSSBb are faster than the original ones for patterns longer than 15,000 characters in English and DNA texts. The gain is 5–15 % in the case of $m = 50,000$. The gain is slightly better for RSSBb than for SSBb, because RSSBb recognizes the q -grams of the last segment which RSSB does not do.