Računanje najduljeg zajedničkog prefiksa temeljenog na BWT

Projekt iz bioinformatike

Tonko Čupić Zvonimir Jurelinac Tomislav Živec

Fakultet elektrotehnike i računarstva

26.01.2018.

Zadatak

Izgradnja polja najduljh zajedničkih prefiksa

Koraci u računanju LCP polja ulaznog niza:

1. Izračun sufiksnog polja ulaznog niza
2. Određivanje Burrows-Wheelerove transformacije ulaza
3. Izgradnja stabla valića nad BW-transformatom
4. Provedba algoritama 1 i 2 (iz rada Beller et al. (2013)) koji kao rezultat daju LCP polje

Sažeto:

$$Ulaz \Rightarrow Sufiksno\ polje \Rightarrow BWT \Rightarrow Stablo\ valića \Rightarrow LCP\ polje$$

Rezultati

Što je sve postignuto?

1. Parsiranje ulaza u **FASTA** formatu
2. Izračun **sufiksnog polja** pomoću gotove `sais` biblioteke
3. Određivanje **Burrows-Wheelerove** transformacije prema jednostavnom izrazu: $BWT[i] = S[SA[i] - 1]$
4. Izgradnja **stabla valića** (vlastita implementacija pomoću niza bitvektora)
5. Implementacija **algoritama 1 i 2** (iz rada *Beller et al. (2013)*)
6. Pohrana rezultata u datoteku

Rezultati

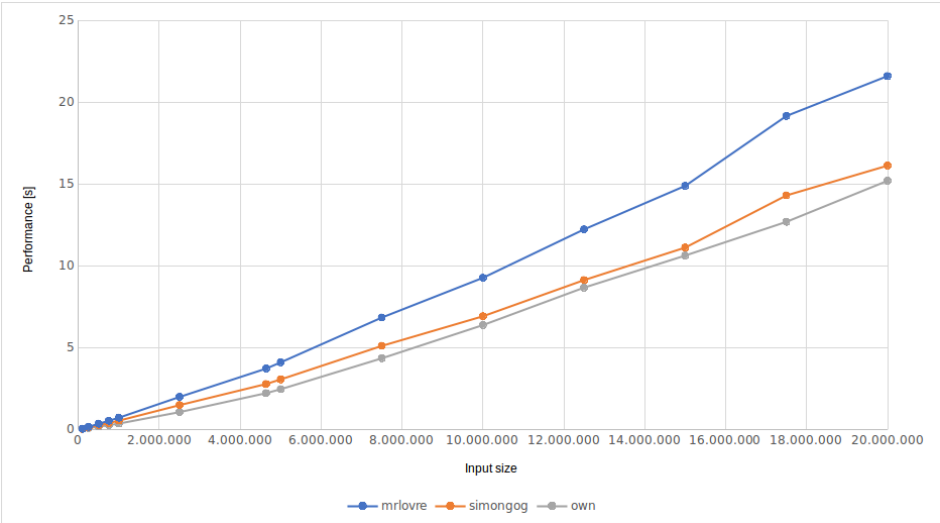
Ostvarene optimizacije

1. **Uklanjanje rekurzije** (prisutna u konstrukciji stabla valića i algoritmu 1) — zamjena s redom i stogom
2. **Optimiziranje korištenih struktura podataka** — obično polje umjesto hash-tablice, vlastita minimalna implementacija stoga
3. **Brzi bitvektori** — koncept *bucketa* (kutija) – veliki bucketi (256 bitova) za spremanje prefiksnih suma, mali (64 bita) za pohranu bitova i brzo brojenje (`popcount`)

Rezultati

Konačne performanse

Vrijeme izvođenja



Zauzeće memorije