生物情報ソフトウェア論 課題 3-5

理学部生物情報科学科 3 年
05-135502
今田雄太郎

課題 3

source code: larssonSadakane3.cc

実行例

$ ./larssonSadakane3 < arrayInput5.txt > larssonSadakane3Output5.txt

課題 4

4.1

source code: inducedSort4.cc

実行例

$ ./inducedSort4 < arrayInput5.txt &> inducedSort4Output5.txt

4.2

source code: inducedSort4Check.cc

実行例

$ ./inducedSort4Check

4.3

source code: inducedSort4time.cc

実行例

$ ./inducedSort4time inducedSort4time.dat

配列の長さと、計算時間との関係は、図 1 のようになった。

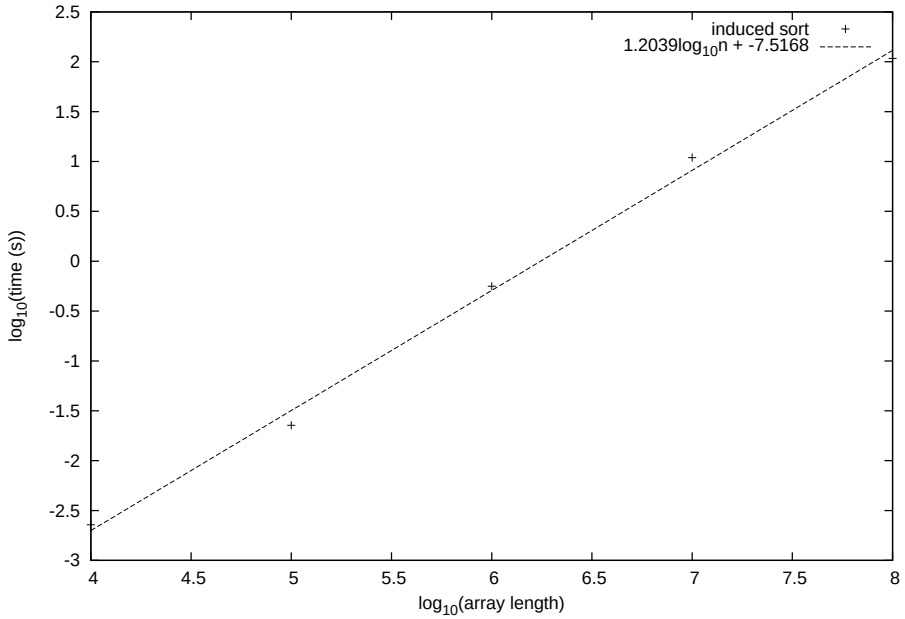

図 1　配列の長さと、suffix array 構築に要する計算時間の関係

従って、配列の長さ n に対してかかる計算時間は、

$$\begin{aligned}\log_{10} \text{time (s)} &= 1.2039 \log_{10} n - 7.5168 \\ &= \log_{10} \left(3.0422 \times 10^{-8} n^{1.2039}\right)\end{aligned}$$

よって、

$$\text{time (s)} = O(n^{1.2039}) \sim O(n)$$

であるから、計算時間は、配列の長さに対してほぼ線形に増えていくことがわかる。

課題 5

5.1

source code: burrowsWheeler2.cc

実行例

$ ./burrowsWheeler2 < burrowsWheeler2In.txt &> burrowsWheeler2Out.txt

5.2

source code: burrowsWheeler2time.cc

実行例

$ ./burrowsWheeler2time burrowsWheeler2time.dat

ランダム配列に対して問い合わせを行っているため、問い合わせ配列が含まれるのは、確率的に高々 10 ~ 15 程度の長さまででであるが、問い合わせ時間が線形比例しているかどうかわかりにくいため、長さ $k(= 10 \sim 10^5)$ の問い合わせ配列を用いて検索時間の測定を行った。問い合わせ配列の長さと検索時間の関係は、図 2 のようになった。

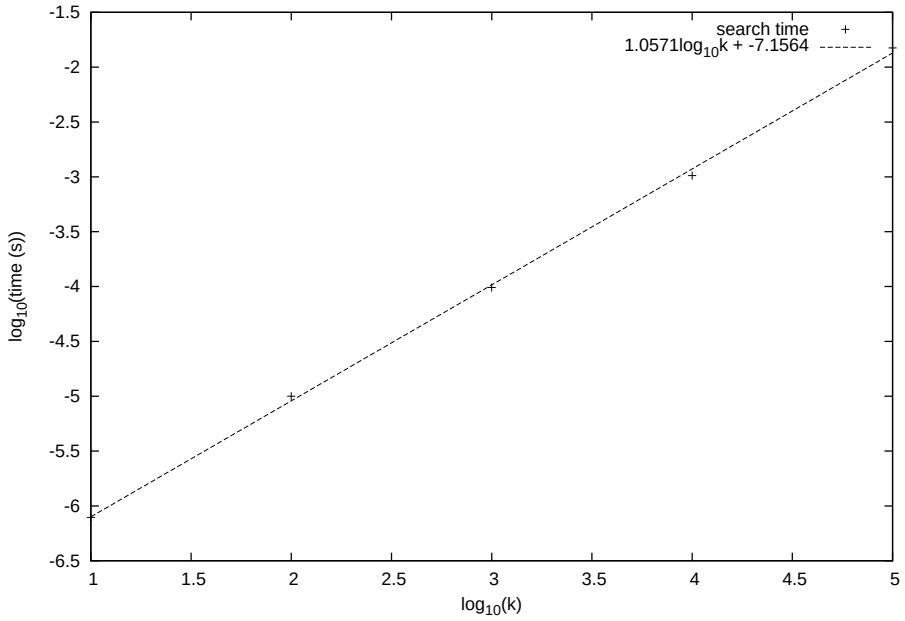

図 2　長さ 10^6 の配列に対して、長さ $k(= 10 \sim 10^5)$ の配列を検索するのにかかった時間。(ここでは、最適化オプション-O3 を除いてコンパイルして実行した。)

従って、問い合わせ配列の長さ k に対してかかる計算時間は、

$$\begin{aligned}\log_{10} \text{time (s)} &= 1.0571 \log_{10} k - 7.1564 \\ &= \log_{10} \left(6.9758 \times 10^{-8} k^{1.0571} \right)\end{aligned}$$

よって、

$$\text{time (s)} = O(k^{1.0571}) \sim O(k)$$

であるから、計算時間は、問い合わせ配列の長さに対してほぼ線形に増えていくことがわかる。計算量 $O(n)$ の Induced Sorting によって suffix array を構築し、suffix array から計算量 $O(n)$ で Burrows Wheeler Transform を行い、更に補助表を作成するという前準備を行っておけば、問い合わせ配列の線形時間で検索を行うことができる。