International Workshop On Combinatorial Algorithms
IWOCA

http://www.iwoca.org—Problems section.

Sorting Strings by Reversals

Guillaume Fertin
LINA, UMR CNRS 6241, University of Nantes, France
guillaume.fertin@univ-nantes.fr

presented at IWOCA 2016
18 Aug. 2016

Introduction. The objects we consider are strings of length n, built on an alphabet Σ. Given a string S, let $S[i, j]$ be the substring of S between positions i and j (both included). A *reversal* $\rho(i, j)$ applied on S consists in taking $S[i, j]$, reversing it, and replacing it at the same location. For instance, if $S = abb\underline{cba}bcc$, then $\rho(4, 6)$ gives $S' = abb\underline{abc}bcc$.

Two strings S and T are said to be *compatible* if the multiset used to build S is the same as the one used to build T. For instance, $S = abbcbabcc$ and $T = cacbabcbb$ are compatible because both are built on the multiset $\{a, a, b, b, b, b, c, c, c\}$.

A *block* in a string S is a maximal substring of S built on only one letter. The number of blocks in S is denoted $b(S)$. For instance, if $S = a\underline{bb}cb\underline{a}b\underline{cc}$, then the three underlined substrings are blocks, and $b(S) = 7$.

Finally, for any two compatible strings S and T, we let $b_{max} = \max\{b(S), b(T)\}$.

One problem, three questions. The problem we consider is the following optimization problem:

Given two compatible strings S and T of length n, what is the minimum number of reversals (called $rd(S, T)$) needed to obtain T from S ?

Note that, because reversals are involutive, for any compatible strings S and T, $rd(S, T) = rd(T, S)$, thus identifying the start and end strings is of no importance.

Here are three open questions:

1. **Reversal diameter**

The reversal diameter $D(n,k)$ is the maximum over all $rd(S,T)$ for all compatible strings of length n with $|\Sigma| = k$. If $|\Sigma| = n$, then S and T are permutations, and in that case we know that $D(n,n) = n - 1$ [1]. Since $b_{max} = n$ when strings are permutations, can we generalize this to any value of k by saying that $D(n,k) = b_{max} - 1$? (this is a bold conjecture!)

2. **Number of blocks in intermediate sequences**

   Let S and T be two compatible strings, and consider any shortest reversal sequence (or SRS) $(S_1, S_2, ...S_p)$ where $S_1 = S$, $S_p = T$ and $p = rd(S,T)$. The two following properties can be easily shown [2]:

   - in any SRS, $b(S_i) = O(b_{max})$ for any i in $[1;p]$
   - there are examples of compatible strings (S,T) for which in any SRS, $b(S_i) > b_{max}$ for at least one i in $[1;p]$

   We have the following conjecture:

   For any compatible strings S and T, there exists an SRS such that $b(S_i) = b_{max} + O(1)$ for any i in $[1;p]$.

   Can we prove/disprove this conjecture?

3. **Approximability of computing $rd(S,T)$**

   We know that computing $rd(S,T)$ is NP-hard, even for some very constrained strings built on a binary alphabet [2].

   Is the problem approximable within a constant ratio on binary alphabets? Same question when $|\Sigma| = O(1)$.

References

[1] Vineet Bafna, Pavel A. Pevzner: Genome Rearrangements and Sorting by Reversals SIAM J. Computing 25(2): 272289 (1996)

[2] Laurent Bulteau, Guillaume Fertin, Christian Komusiewicz: (Prefix) reversal distance for (signed) strings with few blocks or small alphabets. J. Discrete Algorithms 37: 44-55 (2016)