Finding Optimal Alignment and Consensus of Circular Strings

Taehyung Lee[1,*], Joong Chae Na[2,**],
Heejin Park[3,***,†], Kunsoo Park[1,*], and Jeong Seop Sim[4,‡]

[1] Seoul National University, Seoul 151-742, South Korea
[2] Sejong University, Seoul 143-747, South Korea
[3] Hanyang University, Seoul 133-791, South Korea
[4] Inha University, Incheon 402-751, South Korea

Abstract. We consider the problem of finding the optimal alignment and consensus (string) of circular strings. Circular strings are different from linear strings in that the first (leftmost) symbol of a circular string is wrapped around next to the last (rightmost) symbol. In nature, for example, bacterial and mitochondrial DNAs typically form circular strings. The consensus string problem is finding a representative string (consensus) of a given set of strings, and it has been studied on linear strings extensively. However, only a few efforts have been made for the consensus problem for circular strings, even though circular strings are biologically important. In this paper, we introduce the consensus problem for circular strings and present novel algorithms to find the optimal alignment and consensus of circular strings under the Hamming distance metric. They are $O(n^2 \log n)$-time algorithms for three circular strings and an $O(n^3 \log n)$-time algorithm for four circular strings. Our algorithms are $O(n/\log n)$ times faster than the naïve algorithm directly using the solutions for the linear consensus problems, which takes $O(n^3)$ time for three circular strings and $O(n^4)$ time for four circular strings. We achieved this speedup by adopting a convolution and a system of linear equations into our algorithms to reflect the characteristics of circular strings that we found.

[*] This work was supported by NAP of Korea Research Council of Fundamental Science & Technology. The ICT at Seoul National University provides research facilities for this study.

[**] This work was supported by the Korea Research Foundation(KRF) grant funded by the Korea government(MEST) (No. 2009-0069977).

[***] This work was supported by 21C Frontier Functional Proteomics Project from Korean Ministry of Education, Science and Technology (FPR08-A1-020).

[†] Corresponding author, E-mail: hjpark@hanyang.ac.kr

[‡] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0090441).

A. Amir and L. Parida (Eds.): CPM 2010, LNCS 6129, pp. 310–322, 2010.

1 Introduction

A *circular* (or *cyclic*) string is the string that is constructed by linking the beginning and end of a (linear) string together, which can be often found in nature. Gusfield emphasized "Bacterial and mitochondrial DNA is typically circular, both in its genomic DNA and in plasmids, even some true eukaryotes contain plasmid DNA. Consequently, tools for handling circular strings may someday be of use in those organisms." (see [1], page 12)

Finding a representative string of a given set $\mathbb{S} = \{S_1, \ldots, S_m\}$ of m strings of equal length, called a *consensus string* (or *closest string* or *center string*), is a fundamental problem in multiple sequence alignment, which is closely related to the motif recognition problem. Among the conditions that a string should satisfy to be accepted as a consensus, the two most important conditions are

1. to minimize the sum of (Hamming) distances from the strings in $\mathbb{S}$ to the consensus, and
2. to minimize the longest distance (or radius) from the strings in $\mathbb{S}$ to the consensus.

In this paper we consider four different types of consensus problems, CS, CR, CSR, and BSR: Problem CS is finding the optimal consensus minimizing the distance sum, Problem CR is finding the optimal consensus minimizing the radius, Problem CSR is finding the optimal consensus minimizing both distance sum and radius if one exists, and finally Problem BSR is finding a consensus whose distance sum and radius are smaller than given thresholds.

There has been substantial research to solve the problems for linear strings. Problem CS is easy to solve. We can find a string that minimizes the distance sum by selecting the symbol occurring most often in each position of the strings in $\mathbb{S}$. However, Problem CR is hard in general. Given a parameter r, the problem of asking the existence of a string X such that $\max_{1 \le i \le m} d(X, S_i) \le r$ is NP-complete for general m, even when the symbols of the strings are drawn from a binary alphabet [2]. Thus, attention has been restricted to approximation solutions [3,4,5,6,7,8] and fixed-parameter solutions [8,9,10,11]. Furthermore, there have been some algorithms for a small constant m. Gramm et al. [9] proposed a direct combinatorial algorithm for Problem CR for three strings. Sze et al. [12] showed a condition for the existence of a string whose radius is less than or equal to r. Boucher et al. [13] proposed an algorithm for finding a string X such that $\max_{1 \le i \le 4} d(X, S_i) \le r$ for four binary strings. Problems CSR and BSR were considered by Amir et al [14]. They considered the problems for three strings. However, there have been only a few results on multiple alignment of circular strings, even though circular strings are biologically important. Our algorithms differ from the existing multiple alignment algorithms for circular strings [15,16], which use the sum-of-pairs score and some general purpose multiple sequence alignment techniques, such as `clustalW` [17].

In this paper, we introduce the consensus problem for circular strings and present novel algorithms to find the optimal alignment and consensus of circular strings. The consensus problem for circular strings is different from the consensus