ARTICLE IN PRESS

Discrete Applied Mathematics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Relationship between superstring and compression measures: New insights on the greedy conjecture

Bastien Cazaux, Eric Rivals *

LIRMM, CNRS and Université de Montpellier, 161 rue Ada, 34095 Montpellier Cedex 5, France
Institut Biologie Computationnelle, CNRS and Université de Montpellier, 860 rue Saint Priest, 34095 Montpellier Cedex 5, France

ARTICLE INFO

Article history:
Received 25 December 2015
Received in revised form 31 March 2017
Accepted 19 April 2017
Available online xxxx

Keywords:
Approximation algorithm
Shortest Common Superstring Problem
Stringology
Data compression
Assembly
Greedy conjecture

ABSTRACT

A superstring of a set of words is a string that contains each input word as a substring. Given such a set, the Shortest Superstring Problem (SSP) asks for a superstring of minimum length. SSP is an important theoretical problem related to the Asymmetric Travelling Salesman Problem, and also has practical applications in data compression and in bioinformatics. Indeed, it models the question of assembling a genome from a set of sequencing reads. Unfortunately, SSP is known to be NP-hard even on a binary alphabet and also hard to approximate with respect to the superstring length or to the compression achieved by the superstring. Even the variant in which all words share the same length r, called r-SSP, is NP-hard whenever $r > 2$. Numerous involved approximation algorithms achieve approximation ratio above 2 for the superstring, but remain difficult to implement in practice. In contrast the greedy conjecture asked in 1988 whether a simple greedy algorithm achieves ratio of 2 for SSP. Here, we present a novel approach to bound the superstring approximation ratio with the compression ratio, which, when applied to the greedy algorithm, shows a 2 approximation ratio for 3-SSP, and also that greedy achieves ratios smaller than 2. This leads to a new version of the greedy conjecture.

1. Introduction

Given a set of p words $P := \{s_1, s_2, \ldots, s_p\}$ over a finite alphabet Σ, a superstring of P is a string containing each s_i for $1 \leq i \leq p$ as a substring. The **Shortest Superstring Problem (SSP)** asks for a superstring of P of minimal length. SSP is a well studied problem (alias Shortest Common Superstring), with a strong relation to the Asymmetric Travelling Salesman Problem, and is known to be NP-hard even on a binary alphabet [7]. The restriction to instances where all input strings share the same length, say $r > 1$, is denoted r-SSP, becomes polynomial if $r \leq 2$, but remains NP-hard as soon as the strings are of length at least 3 [1]. Two approximation measures can be optimised for SSP: either the length of the superstring is minimised, or the compression is maximised (i.e., the sum of the lengths of the input strings minus that of the superstring). For a word x, $|x|$ denotes the *length* of x. Let $\|P\|$ denote $\sum_{s_i \in P} |s_i|$ and let t be the output superstring, then the compression equals $\|P\| - |t|$. With both measures SSP is hard to approximate (MAX-SNP-hard, see [1]). Since 1991, a long series of elaborated algorithms have improved the approximation ratio for both measures culminating in $2\frac{11}{23}$ for the superstring [13] and in 3/4 for the compression measure [14]. A recent table listing these ratio and the literature, as well as known inapproximability bounds appears in [9]. A detailed survey gives an overview of the numerous application contexts of SSP [8].

* Corresponding author at: LIRMM, CNRS and Université de Montpellier, 161 rue Ada, 34095 Montpellier Cedex 5, France.
E-mail addresses: bastien.cazaux@lirmm.fr (B. Cazaux), rivals@lirmm.fr (E. Rivals).

http://dx.doi.org/10.1016/j.dam.2017.04.017

In 1988, a seminal paper introduced a simple greedy algorithm, consisting in repeatedly merging two words that exhibit the largest (prefix–suffix) overlap until only one string remains [16]. With $P := \{abba, bbaa, aaba\}$ for example, *abba* is first merged with *bbaa* yielding *abbaa* (they share a 3-letter overlap), then, *abbaa* is merged with *aaba* resulting in the superstring *abbaaba* of length 7; as $\|P\| = 12$, the compression obtained equals $\|P\| - |t| = 12 - 7 = 5$. Note that their greedy algorithm, denoted by GREEDY, can be seen as the greedy algorithm of a specific hereditary system [4]. Tarhio and Ukkonen proved in [16] that GREEDY achieves a compression ratio of $1/2$ and formulated the *greedy conjecture*: the greedy algorithm yields a superstring ratio of 2. Despite a lot of research dedicated to SSP, this conjecture has remained open since 1988. A weaker form of this conjecture asks to prove this ratio for r-SSP and some values of r. Blum et al. have shown for GREEDY a superstring ratio of 4 [1], which was later improved to 3.5 in [10]. The greedy conjecture is supported by simulated experiments [18,15]. Moreover, the superstring approximation ratio obtained by the greedy algorithm remains a crucial question, especially since other approximation algorithms are usually less efficient than GREEDY [10].

Recently, it has been proven that in the case where all input words have length 4 (for 4-SSP) the greedy algorithm achieves a superstring ratio of at most 2, as stated by the conjecture [11]. This proof is valid only for words of length 4 and cannot be adapted to words of length 3, for instance. Kulikov and colleagues [11] suggest that the conjecture for 3-SSP follows from the fact that GREEDY achieves 2-approximation of the compression measure, citing [16]. To our knowledge, no proof for the greedy conjecture for words of length 3 has ever been published and there are no mention of it in a recent survey [8]. Here, we study the relationship between the compression ratio and the superstring ratio of an approximation algorithm in general, and derive a bound of the superstring ratio in function of the compression ratio. When applied to GREEDY on words of fixed length (r-SSP), we obtain a superstring approximation ratio of 2 for 3-SSP, and this ratio increases with r to reach for $r = 6$ a value of $7/2$, which is the best known ratio for the greedy algorithm [10]. But we also get a tight superstring ratio of $3/2$ for 2-SSP, thereby demonstrating that the greedy algorithm can achieve a ratio strictly smaller than 2. This shows first that the general relationship between the superstring and compression measures is important and can serve for future research. Second, the ratio smaller than 2 does not contradict known bounds or instances. Indeed, the known examples give a bound that converges towards 2 from below when the length of the input words tends to infinity. Thus, we propose a more precise conjecture for r-SSP, in which the superstring ratio equals $2 - \frac{1}{r}$ instead of 2.

Notation: An *alphabet* Σ is a finite set of *letters*. A linear *word* or *string* over Σ is a finite sequence of elements of Σ. The set of all finite words over Σ is denoted by $\Sigma^\star$, and Σ^r denotes the subset of $\Sigma^\star$ of words of length r for any positive integer r. Given two words x and y, we denote by xy the *concatenation* of x and y.

2. Relation between maximum compression and shortest superstring approximation ratios for SSP

Here, we exhibit for SSP an upper bound of the superstring approximation ratio of an algorithm in function of its compression ratio.

Let $\mathcal{A}$ be a polynomial-time approximation algorithm for SSP. As all approximation algorithms considered here take polynomial time in the input size, we simply omit this characteristic in the sequel. We denote by $s_{\mathcal{A}}(P)$ the output of algorithm $\mathcal{A}$ with input P, and by $s_{opt}(P)$ an optimal superstring for this input. Note that $s_{opt}(P)$ also achieves a maximum compression for P. We only consider approximation algorithms that return a superstring whose length is bounded by $\|P\|$. In other words, we disregard algorithms that insert additional symbols beyond those required by the words of the instance. Without this restriction, the approximation ratio $\text{super}(\mathcal{A})$ would not be defined for any algorithm $\mathcal{A}$, and the ratio $\text{comp}(\mathcal{A})$ could be negative; both ratios are defined a few lines below. Instances where the optimal superstring is the concatenation of all the words of the instance satisfy $|s_{opt}(P)| = \|P\|$. In such cases, for any approximation algorithm $\mathcal{A}$, one has $\|P\| = |s_{opt}(P)| = |s_{\mathcal{A}}(P)| = \|P\|$. Such instances are excluded from Theorem 1. Let us define the superstring approximation ratio of algorithm $\mathcal{A}$, denoted $\text{super}(\mathcal{A})$, as the smallest real value such that for any input P:

$$1 \leq \frac{|s_{\mathcal{A}}(P)|}{|s_{opt}(P)|} \leq \text{super}(\mathcal{A}).$$

Similarly, we define the compression ratio $\text{comp}(\mathcal{A})$ as the largest real value such that, for any input P satisfying $\|P\| \neq |s_{opt}(P)|$, we have

$$0 \leq \text{comp}(\mathcal{A}) \leq \frac{\|P\| - |s_{\mathcal{A}}(P)|}{\|P\| - |s_{opt}(P)|}.$$

Instances where the optimal superstring is the concatenation of all the words of the instance satisfy $|s_{opt}(P)| = \|P\|$. In such cases, for any approximation algorithm $\mathcal{A}$ one has $\|P\| = |s_{opt}(P)| = |s_{\mathcal{A}}(P)| = \|P\|$. Such instances are excluded from Theorem 1.

Theorem 1. *Let P be a set of words satisfying $|s_{opt}(P)| \neq \|P\|$. Let γ be a real such that $0 < \gamma \leq \frac{|s_{opt}(P)|}{\|P\|}$, and let $\mathcal{A}$ be an approximation algorithm for SSP. We have:*

$$super(\mathcal{A}) \leq \frac{(\gamma - 1) \times comp(\mathcal{A}) + 1}{\gamma}.$$

Please cite this article in press as: B. Cazaux, E. Rivals, Relationship between superstring and compression measures: New insights on the greedy conjecture, Discrete Applied Mathematics (2017), http://dx.doi.org/10.1016/j.dam.2017.04.017