Tight Comparison Bounds for the String Prefix-Matching Problem

Dany Breslauer*
CWI
P.O. Box 4079, 1009 AB
Amsterdam, The Netherlands
dany@cwi.nl

Livio Colussi
Universitá di Padova
Dipartimento di Matematica Pura ed Applicata
Via Belzoni 7, Padova, Italy
colussi@pdmat1.unipd.it

Laura Toniolo†
Universitá di Padova
Dipartimento di Matematica Pura ed Applicata
Via Belzoni 7, Padova, Italy
toniolo@pdmat1.unipd.it

CUCS-049-92

Abstract

In the *string prefix-matching* problem one is interested in finding the longest prefix of a pattern string of length m that occurs starting at each position of a text string of length n. This is a natural generalization of the string matching problem where only occurrences of the whole pattern are sought. The Knuth-Morris-Pratt string matching algorithm can be easily adapted to solve the string prefix-matching problem without making additional comparisons.

In this paper we study the exact complexity of the string prefix-matching problem in the deterministic sequential comparison model. Our bounds do not account for comparisons made in a pattern preprocessing step. The following results are presented:

1. A family of linear-time string prefix-matching algorithms that make at most $\lfloor \frac{2m-1}{m} n \rfloor$ comparisons.
2. A tight lower bound of $\lfloor \frac{2m-1}{m} n \rfloor$ comparisons for any string prefix-matching algorithm.

We also consider the special case when the pattern and the text strings are the same string and all comparisons are accounted. This problem, which we call the *string self-prefix* problem, is similar to the failure function that is computed in the pattern preprocessing of the Knuth-Morris-Pratt string matching algorithm. By using the lower bound for the string prefix-matching problem we are able to show:

3. A lower bound of $2m - \lceil 2\sqrt{m} \rceil$ comparisons for any algorithm that solves the self-prefix problem.

*The work of this author was partially supported by the European Research Consortium for Informatics and Mathematics postdoctoral fellowship.

†The work of this author was done while visiting at Columbia University.

1 Introduction

In the *string prefix-matching* problem one is interested in finding the longest prefix of a pattern string $\mathcal{P}[1..m]$ that starts at each position of a text string $\mathcal{T}[1..n]$. More formally, the required output of the string prefix-matching problem is an integer array $\Pi[1..n]$ such that for each text position i, $\mathcal{T}[i..i+\Pi[i]-1] = \mathcal{P}[1..\Pi[i]]$ and if $\Pi[i] < m$ and $i+\Pi[i] \leq n$, then $\mathcal{T}[i+\Pi[i]] \neq \mathcal{P}[\Pi[i]+1]$.

The string prefix-matching problem is a natural generalization of the standard string matching problem where only complete occurrences of the pattern are sought. The classical linear time string matching algorithm of Knuth, Morris and Pratt [7] can be easily adapted to solve the string prefix-matching problem in the same time bounds without making additional comparisons[1].

In this paper we study the exact number of comparisons performed by algorithms that have access to the input strings by pairwise symbol comparisons that test for equality. This work was motivated by recent results on the exact comparison complexity of the string matching problem [2, 4, 5, 6, 8]: Colussi [4] optimized the Knuth-Morris-Pratt [7] string matching algorithm, which makes $2n-m$ comparisons, using program correctness proof techniques and presented an algorithm that makes $n+\frac{1}{2}(n-m)$ comparisons. His algorithm was later improved by Galil and Giancarlo [6] and further by Breslauer and Galil [2]. Recently, Cole and Hariharan [3] discovered a different algorithm that makes only $n+\frac{c}{m}(n-m)$ comparisons, but requires an expensive pattern preprocessing[2]. Cole and Hariharan [3] also gave a lower bound of $n+\frac{\overline{c}}{m}(n-m)$ comparisons improving lower bounds given by Galil and Giancarlo [5] and Zwick and Paterson [8]. However, there is still a small gap between the lower and upper bounds for string matching.

The string prefix-matching problem is obviously harder than the standard string matching problem since each text symbol must be either compared directly to the first symbol of the pattern or compared successfully to another symbol, while in the string matching problem some text symbols might not be compared at all [1]. Interestingly, this "hardness" introduces more structure that makes the analysis of the string prefix-matching problem easier.

This paper presents matching lower and upper bounds for the string prefix matching problem. In particular we give:

1. A family of linear-time string prefix-matching algorithms that make at most $\lfloor\frac{2m-1}{m}n\rfloor$ comparisons. The pattern preprocessing step of these algorithms is almost identical to that of the string matching algorithm of Knuth, Morris and Pratt [7].

   This upper bound improves on the $2n-1$ comparisons made by the adapted string matching algorithm of Knuth, Morris and Pratt [7].

2. A lower bound of $\lfloor\frac{2m-1}{m}n\rfloor$ comparisons for any string prefix-matching algorithm.

These results show that although the string matching and the string prefix-matching problems are closely related, their exact comparison complexities are inherently different:

[1]Since complete occurrences of the pattern cannot start at text positions larger than $n-m+1$, the string matching algorithm can stop before reaching the end of the text. The prefix-matching algorithm must continue until the end of the text and therefore, it may make at most m extra comparisons.

[2]All bounds for the string matching algorithms mentioned do not account for the comparisons made in a pattern preprocessing step. The pattern preprocessing step of Cole and Hariharan's algorithm makes about m^2 comparisons while the other algorithms make $2m$ comparisons.

- When $m \to \infty$ and $n \gg m$ the comparison complexity of the string matching problem approaches n while the comparison complexity of the string prefix-matching problem approaches $2n$.

- The lower bound proofs of the two problems require different arguments: the pattern string that we use for the lower bound is 'ab^{m-1}' while the lower bounds for the string matching problem require patterns with more complex periodicity structures [3, 8].

Finally, we consider the special case when the text and the pattern strings are the same string and all comparisons are accounted. This problem, which we call the *string self-prefix* problem, is similar to the failure function[3] that is computed in the pattern preprocessing of the Knuth-Morris-Pratt [7] string matching algorithm using $2n-1$ comparisons. We give the following lower bound:

3. $2m - \lceil 2\sqrt{m} \rceil$ comparisons are required by any algorithm that solves the self-prefix problem for a string of length m.

This paper is organized as follows. In Section 2 we describe a family of string prefix-matching algorithms. In Section 3 we give the lower bound and in Section 4 we use this lower bound to prove a lower bound on the self-prefix problem.

2 Upper Bounds

In this section we give a family of linear time on-line string prefix-matching algorithms that make at most $\lfloor \frac{2m-1}{m} n \rfloor$ comparisons. We call a prefix-matching algorithm *on-line* if before comparing $\mathcal{T}[\zeta]$ it has determined if the pattern prefix that starts at text position l terminates before text position ζ for all text positions l, such that $l < \zeta$.

Let $\Psi^\zeta = \{\psi_\epsilon^\zeta | \zeta - m < \psi_1^\zeta < \psi_2^\zeta < \cdots < \psi_{k_\zeta}^\zeta = \zeta\}$ be the set of all text positions for which $\Pi[\psi_\epsilon^\zeta]$ can not be determined without examining $\mathcal{T}[\zeta]$. That is, $\mathcal{T}[\psi_\epsilon^\zeta..\zeta-1] = \mathcal{P}[1..\zeta-\psi_\epsilon^\zeta]$ and $\mathcal{T}[\zeta]$ must be compared to determine whether $\Pi[\psi_\epsilon^\zeta] = \zeta - \psi_\epsilon^\zeta$ or $\Pi[\psi_\epsilon^\zeta] > \zeta - \psi_\epsilon^\zeta$. In this terminology, an on-line prefix-matching algorithm must determine whether $\mathcal{T}[\zeta] = \mathcal{P}[\zeta - \psi_\epsilon^\zeta]$, for all ψ_ϵ^ζ, before examining any text position larger than ζ. Note that $\Psi^{\zeta+1} \subset \Psi^\zeta \cup \{\zeta+1\}$.

In the discussion below we count only comparisons and all other computation is free. We assume that the algorithms have obtained complete information about the pattern in an unaccounted pattern preprocessing step which may compare even all $\binom{m}{2}$ pairs of pattern symbols. We show in Theorem 2.3 how to implement these algorithms efficiently in a standard model.

Comparison efficient on-line prefix-matching algorithms are somewhat restricted with the choices of comparisons they can make. It is easy to see that they gain no advantage by comparing pairs of text symbols. Furthermore, all comparisons at text position ζ must be between $\mathcal{T}[\zeta]$ and some $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ or otherwise can be answered by an adversary as unequal without giving the algorithm any useful information.

In the rest of this section we only consider algorithms that compare $\mathcal{T}[\zeta]$ to some $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ until $\mathcal{T}[\zeta] = \mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ for some ψ_ϵ^ζ, or until $\mathcal{T}[\zeta] \neq \mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ for all ψ_ϵ^ζ, and then move to the

[3] These are essentially different representations of the same information: one can be computed from the other in linear time without extra comparisons.

next text position. The only difference between these algorithms is the order in which the pattern symbols $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ are compared to $\mathcal{T}[\zeta]$.

Lemma 2.1 *Let $\mathcal{A}$ be an on-line prefix-matching algorithm and assume that all the comparisons that it makes are between $\mathcal{T}[\zeta]$ and some $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ and that no $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ is compared more than once. Then, $\mathcal{A}$ makes at most $2n - 1$ comparisons.*

Proof: It is obvious that $\mathcal{A}$ does not need to make more than n comparisons which result in equal answers. In every comparison which results in unequal answer $\mathcal{A}$ determines that at least one prefix of the pattern which starts at some text position ψ_ϵ^ζ terminates at text position ζ. Therefore, $\mathcal{A}$ does not make more than n comparisons which result in unequal answers. However, if all pattern prefixes that start at text positions in Ψ^ζ terminate at text position ζ, then $\mathcal{A}$ moves to the next text position without a comparison that is answered as equal.

Consider the last text position $\zeta = n$. It is clear that if all comparisons at this text position result in unequal answers, then $\mathcal{A}$ got at most $n - 1$ equal answers. On the other hand, if a comparison was answered as equal, then the there is at least one pattern prefix which starts at some text position ψ_ϵ^ζ and was not terminated by an inequality answer and, thus, $\mathcal{A}$ got at most $n - 1$ unequal answers. Therefore, $\mathcal{A}$ makes at most $2n - 1$ comparisons. □

The adapted Knuth-Morris-Pratt [7] prefix-matching algorithm satisfies the conditions of Lemma 2.1 and makes at most $2n-1$ comparisons. There are cases in which it would actually make $2n-1$ comparisons; e.g. $\mathcal{P}[1..2] =$ 'ab' and $\mathcal{T}[1..n] =$ 'a^n'. Note that the Knuth-Morris-Pratt algorithm compares $\mathcal{T}[\zeta]$ to $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ in increasing order of ψ_ϵ^ζ. This order is the worst possible order as we show next:

Theorem 2.2 *Let $\mathcal{A}$ be an on-line prefix-matching algorithm and assume that all the comparisons that it makes are between $\mathcal{T}[\zeta]$ and some $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ and that no $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ is compared more than once. Further, assume that $\mathcal{A}$ may compare $\mathcal{P}[\zeta - \psi_1^\zeta]$ only last. Then, $\mathcal{A}$ makes at most $\lfloor \frac{2m-1}{m} n \rfloor$ comparisons.*

Proof: As in Lemma 2.1, every comparison between $\mathcal{T}[\zeta]$ to $\mathcal{P}[\psi_\epsilon^\zeta]$ which results in an unequal answer detemines that the pattern prefix which starts at text position ψ_ϵ^ζ terminates at text position ζ. We charge such a comparison to text position ψ_ϵ^ζ and charge comparisons that result in equal answer to the text position compared. Using this charging scheme it is obvious that each text position can be charged with at most two comparisons and that comparisons to $\mathcal{T}[\zeta]$ cannot be charged to any text position that is smaller than ψ_1^ζ.

When $\mathcal{A}$ reaches text position ζ, the number of comparisons that are charged to the text positions $\psi_1^\zeta \cdots \zeta - 1$ is at most $2(\zeta - \psi_1^\zeta) - (|\Psi^\zeta| - 1)$. This is so since each of these $\zeta - \psi_1^\zeta$ text positions has a comparison that resulted in equal answer charged to it but at least $|\Psi^\zeta| - 1$ of the text positions do not have a comparison that resulted in unequal answer charged to them.

We show by induction that the number of comparisons charged to text positions smaller than ψ_1^ζ is at most $\lfloor \frac{2m-1}{m}(\psi_1^\zeta - 1) \rfloor$. This is obviously true at the beginning when $\zeta = 1$. The only concern is when $\mathcal{A}$ advances from ζ to $\zeta + 1$ and $\psi_1^\zeta < \psi_1^{\zeta+1}$.

Let $l = \psi_1^{\zeta+1} - \psi_1^\zeta$. The number of comparisons that were charged to the text positions $\psi_1^\zeta \cdots \psi_1^{\zeta+1} - 1$ is at most $2l - 1$ since either at most l text positions were charged with comparisons that resulted in equal answers and ψ_1^ζ was not charged with an unequal answer, or ψ_1^ζ was charged with an unequal answer but then $\psi_1^{\zeta+1} = \zeta + 1$ and text position ζ was not charged with an equal answer. But $l \leq m$ and by simple arithmetic,

$$\lfloor \frac{2m-1}{m}(\psi_1^\zeta - 1) \rfloor + (2l - 1) \leq \lfloor \frac{2m-1}{m}(\psi_1^{\zeta+1} - 1) \rfloor.$$

When $\mathcal{A}$ reaches text position $\zeta = n + 1$, the number of comparisons satisfies,

$$\lfloor \frac{2m-1}{m}(\psi_1^\zeta - 1) \rfloor + 2(\zeta - \psi_1^\zeta) - (|\Psi^\zeta| - 1) \leq \lfloor \frac{2m-1}{m} n \rfloor.$$

□

The Knuth-Morris-Pratt [7] algorithm can be trivially modified to compare $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$ in increasing order of ψ_ϵ^ζ starting from $\epsilon = 2$ and only then compare $\mathcal{P}[\zeta - \psi_1^\zeta]$. This gives a linear time on-line prefix-matching algorithm that makes only $\lfloor \frac{2m-1}{m} n \rfloor$ comparisons and uses the same pattern preprocessing as the Knuth-Morris-Pratt algorithm.

The following theorem has a more general claim.

Theorem 2.3 *Let $\mathcal{A}$ be an on-line prefix-matching algorithm and assume that all the comparisons that it makes are between $\mathcal{T}[\zeta]$ and some $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$. Then, except possibly the step which chooses the order according to which the $\mathcal{P}[\zeta - \psi_\epsilon^\zeta]$'s are compared, $\mathcal{A}$ can be implemented in linear time with a linear preprocessing step.*

Proof: The data structures that are used by Breslauer and Galil [2] to implement a family of string matching algorithms can be used to implement $\mathcal{A}$ in linear time with a linear preprocessing step that makes at most $2m - 1$ comparisons. □

3 Lower Bounds

In this section we show a lower bound on the number of comparisons required by any string prefix-matching algorithm which may have an unaccounted pattern preprocessing step. We describe an adversary that can force such an algorithm to make at least $\lfloor \frac{2m-1}{m} n \rfloor$ comparisons. The techniques we use are somewhat similar to those used by Galil and Giancarlo [5] and Zwick and Paterson [8]. However, the pattern string that is used for the lower bound has a very simple periodicity structure.

Theorem 3.1 *Any prefix-matching algorithm must make at least $\lfloor \frac{2m-1}{m} n \rfloor$ comparisons.*

Proof: Fix the pattern to $\mathcal{P}[1..m]$ = 'ab^{m-1}' and assume that the text alphabet has at least three symbols. We show that an adversary can answer comparisons made by any prefix-matching algorithm in a way that if the algorithm claims to have computed $\Pi[1..n]$ in less than $\lfloor \frac{2m-1}{m} n \rfloor$ comparisons, then it can be fooled.

Consider first algorithms that cannot compare pairs of text symbols. The adversary will maintain each text symbol in one of three states: *unknown*, *potential 'a' or 'b'*, and *fixed 'a' or 'b'*.

Initially the adversary sets all text symbols at positions l, such that $l \equiv 1 \bmod m$, to be potential 'a's and all other text symbols to be unknown. A comparison between an unknown text symbol to 'a' or to 'b' is answered as unequal and the text symbol is set to be a potential 'b' or 'a', respectively. A potential 'a' or 'b' is revealed to the algorithm at the cost of one comparison after which it becomes fixed.

If an algorithm claims it has computed $\Pi[1..n]$ before all text symbols are fixed, the adversary has the freedom of setting one of the unknown or potential symbols to an alphabet symbol other than 'a' or 'b'. Let u be a text position that is not fixed and assume that all other text symbols become fixed. If $T[u]$ is a potential 'b', then there exists v such that $u - m < v < u$ and $\mathcal{T}[v..u-1] =$ 'ab^{u-v-1}', and the adversary can alter $\Pi[v]$ by fixing $\mathcal{T}[u]$ to 'b' or 'c'. Similarly, the adversary can alter $\Pi[u]$ if $T[u]$ is unknown or a potential 'a'.

Thus, any algorithm must make two comparisons at each text position except at the text positions that are set initially to be potential 'a's, where it has to make only one comparison. The total number of comparisons is therefore at least $\lfloor \frac{2m-1}{m} n \rfloor$.

When pairwise comparisons of text symbols are permitted, the lower bound arguments are slightly more complicated. To keep track of the comparisons the adversary maintains a graph with $n+2$ vertices that correspond to the n text symbols and the pattern symbols 'a' and 'b'. The edges of the graph correspond to comparisons and are labeled with their outcome ("equal" or "unequal").

The adversary maintains a two-level representation of the edges. This representation satisfies the following invariants:

1. A subgraph that contains the edges that are labeled "unequal" and all vertices.

   We refer to the connected components in this subgraph as *components*. The adversary will maintain the property that components are bipartite graphs.

2. A subgraph that contains the edges that are labeled "equal" and all vertices.

   We refer to the connected components in this subgraph as *super-vertices*. By transitivity, all vertices in a super-vertex correspond to equal symbols. The adversary will maintain the property that vertices which are in the same super-vertex are always in the same side of a single component.

Initially, the graph has $1 + \lceil \frac{n}{m} \rceil$ edges: between the pattern symbol 'a' and the pattern symbol 'b' and between the pattern symbol 'b' and every text position l, such that $l = 1 \bmod m$. These edges are labeled "unequal"; the invariants are clearly satisfied. The adversary answers comparisons as follows:

- A comparison between vertices that belong to different components is answered as unequal.

  The two components are merged into a single component which is still bipartite.

- A comparison between vertices that belong to the same component is answered as equal if and only if the two vertices are on the same side of the component.

  This may cause two super-vertices to be merged into one. Note that comparisons between vertices that belong to the same component but are on different sides and comparisons between two vertices in the same super-vertex do not contribute anything to the component or super-vertex structure and are practically answered for free.

The invariants are obviously maintained after each comparison is answered. Note that vertices which are in the same super-vertex as one of the pattern symbols correspond to fixed symbols; vertices which are in the same component as the pattern symbols correspond to potential symbols and vertices which are in other components correspond to unknown symbols.

A prefix-matching algorithm can terminate correctly when there is only one component and two super-vertices. Since every connected component with l vertices must have at least $l-1$ edges, there are at least $n+1$ edges labeled "unequal" and at least n edges labeled "equal" at termination. Thus, the total number of comparisons is at least $2n+1-(1+\lceil\frac{n}{m}\rceil) = \lfloor\frac{2m-1}{m}n\rfloor$. $\square$

4 Lower Bounds for the Self-Prefix Problem

In this section we consider the special case where the pattern and the text strings are the same string and all comparisons are accounted. This problem is solved in the preprocessing step of the Knuth-Morris-Pratt [7] string matching algorithm in linear time and $2m-1$ comparisons.

Theorem 4.1 *Fix a positive integer constant k. Then, any algorithm that solves the self-prefix problem for a string of length m, such that $m \geq k$, must make at least $\lfloor\frac{2k-1}{k}m\rfloor - k$ comparisons.*

Proof: The adversary fixes the first k symbols of the string to 'ab^{k-1}' and reveals them to the algorithm for $k-1$ comparisons. By Theorem 3.1 the algorithm must make at least $\lfloor\frac{2k-1}{k}(m-k)\rfloor$ more comparisons. But,

$$\lfloor\frac{2k-1}{k}(m-k)\rfloor + k - 1 = \lfloor\frac{2k-1}{k}m\rfloor - k.$$

$\square$

If the length of the input string is known to the adversary in advance, it can maximize the lower bound as the next corollary shows. In the on-line case, where the string is given a symbol at a time and its length not known in advance, there seems to be a tradeoff between maximizing the number of comparisons in the short term and in the long term.

Corollary 4.2 *The lower bound in Theorem 4.1 reaches its maximal value $2m - \lceil 2\sqrt{m}\rceil$ for $k = \lfloor\sqrt{m}\rfloor$.*

Proof: It is easy to verify that the maximum is achieved for $k = \lfloor\sqrt{m}\rfloor$ and also for $k = \lceil\sqrt{m}\rceil$. $\square$

5 Acknowledgments

We thank Raffaele "The Great Raf" Giancarlo, Matt Franklin and Moti Yung for comments on early versions of this paper.

References

[1] R. S. Boyer and J. S. Moore. A fast string searching algorithm. *Comm. of the ACM*, 20:762–772, 1977.

[2] D. Breslauer and Z. Galil. Closing the Gap in the Exact Complexity of String Matching. manuscript, 1992.

[3] R. Cole and R. Hariharan. Tighter Bounds on The Exact Complexity of String Matching. In *Proc. 33rd IEEE Symp. on Foundations of Computer Science*, 1992. to appear.

[4] L. Colussi. Correctness and efficiency of string matching algorithms. *Inform. and Control*, 95:225–251, 1991.

[5] Z. Galil and R. Giancarlo. On the exact complexity of string matching: lower bounds. *SIAM J. Comput.*, 20(6):1008–1020, 1991.

[6] Z. Galil and R. Giancarlo. The exact complexity of string matching: upper bounds. *SIAM J. Comput.*, 21(3):407–437, 1992.

[7] D. E. Knuth, J. H. Morris, and V. R. Pratt. Fast pattern matching in strings. *SIAM J. Comput.*, 6:322–350, 1977.

[8] U. Zwick and M. S. Paterson. Lower bounds for string matching in the sequential comparison model. manuscript, 1991.