Computing matching statistics on Wheeler DFAs

Alessio Conte[1], **Nicola Cotumaccio**[2,3], **Travis Gagie**[3], **Giovanni Manzini**[1], **Nicola Prezza**[4], **Marinella Sciortino**[5]

[1]University of Pisa, Italy

[2]GSSI, L'Aquila, Italy

[3]Dalhousie University, Halifax, Canada

[4]Ca' Foscari Unversity, Venice, Italy

[5]University of Palermo, Italy

Abstract

Matching statistics were introduced to solve the approximate string matching problem, which is a recurrent subroutine in bioinformatics applications. In 2010, Ohlebusch et al. [SPIRE 2010] proposed a time and space efficient algorithm for computing matching statistics which relies on some components of a compressed suffix tree - notably, the longest common prefix (LCP) array. In this paper, we show how their algorithm can be generalized from strings to Wheeler deterministic finite automata. Most importantly, we introduce a notion of LCP array for Wheeler automata, thus establishing a first clear step towards extending (compressed) suffix tree functionalities to labeled graphs.

Introduction

Given a string T and a pattern π, the classical formulation of the pattern matching problem requires to decide whether the pattern π occurs in the string T and, possibly, count the number of such occurrences and report the positions where they occur. The invention of the FM-index [1], which is based on the Burrows-Wheeler transform [2], opened a new line of research in the pattern matching field. The indexing and compression techniques behind the FM-index deeply rely on the idea of suffix sorting, and over the years have been generalized from strings to trees [3], De Brujin graphs [4, 5], Wheeler graphs [6, 7] and arbitrary graphs [8, 9]. In particular, the class of Wheeler graphs is probably the one that captures the intuition behind the FM-index in the simplest way, and indeed the notion of Wheeler order has relevant consequences in automata theory [7, 10].

However, in bioinformatics we are not only interested in exact pattern matching, but also in a myriad of variations of the pattern matching problem [11]. In particular, *matching statistics* were introduced to solve the approximate pattern matching problem [12]. A powerful data structure that is able to address the variations of the pattern matching problem at once is the

alessio.conte@unipi.it .

suffix tree [13]. The main drawback of the suffix tree is its space consumption, which is non-negligible both in theory and in practice. As a consequence, the suffix tree has been replaced by the *suffix array* [14]. While suffix arrays do not have all the functionalities of suffix trees, it has been shown that they can be augmented with some additional data structures — notably, the longest common prefix (LCP) array — so that it is possible to retrieve the full functionalities of a suffix trees [15]. All these components can be successfully compressed, leading to the so-called *compressed suffix trees* [16].

The natural question is whether it is possible to provide suffix tree functionalities not only to strings, but also to graphs, and in particular Wheeler graphs. In this paper, we provide a first partial affirmative answer by considering the problem of computing matching statistics. In 2010, Ohlebusch et al. [17] proposed a time and space efficient algorithm for computing matching statistics which relies on some components of a compressed suffix tree. In this paper, we show how their algorithm can be generalized from strings to Wheeler deterministic finite automata. Most importantly, we introduce a notion of longest common prefix (LCP) array for Wheeler automata, thus establishing an important step towards extending (compressed) suffix tree functionalities to labeled graphs.

Notation and first definitions

Throughout the paper, we consider an alphabet Σ and a fixed total order $\preceq$ on Σ. We denote by Σ^* the set of all finite strings on Σ and by Σ^ω the set of all (countably) infinite strings on Σ. The empty word is ϵ. If $\alpha \in \Sigma^*$, then α^R is the reverse string of α. We extend the total order $\preceq$ from Σ to $\Sigma^* \cup \Sigma^\omega$ lexicographically. If i and j are integers, with $i \leq j$, define $[i, j] = \{i, i + 1, ..., j - 1, j\}$. If T is a string, the i-th character of T is $T[i]$, and $T[i .. j] = T[i] .. T[j]$.

We will consider deterministic automata $\mathscr{A} = (Q, E, s_0, F)$, where Q is the set of states, $E \subseteq Q \times Q \times \Sigma$ is the set of labeled edges, $s_0 \in Q$ is the initial state and $F \subseteq Q$ is the set of final states. The definition implies that for every $u \in Q$ and for every $a \in \Sigma$ there exists at most one edge labeled a leaving u. Following [7, 10], we assume that s_0 has no incoming edges, and every state is reachable from the initial state; moreover, all edges entering the same state have the same label (*input-consistency*), so that for every $u \in Q \smallsetminus \{s_0\}$ we can let $\lambda(u)$ be the label of all edges entering u. We define $\lambda(s_0) = \#$, where $\# \notin \Sigma$ is a special character for which we assume $\# \prec a$ for every $a \in \Sigma$ (the character $\#$ is an analogous of the termination character $\$$ used for suffix trees and suffix arrays). As a consequence, an edge (u', u, a) can be simply written as (u', u), because it must be $a = \lambda(u)$.

We assume familiarity with the notions of suffix array (SA), Burrows Wheeler transform (BWT), FM-index and backward search [1].

The *matching statistics* of a pattern $\pi = \pi[1 .. m]$ with respect to a string $T = T[1 .. n]$ are defined as follows. Assume that $T[n] = \$ \notin \Sigma$, where $\$ \prec a$ for every $a \in \Sigma$. Determining the matching statistics of π with respect to T means determining, for $1 \leq i \leq m$, (i) the longest prefix π' of $\pi[i .. m]$ which occurs in T, and (ii) the interval corresponding to the set of all

strings starting with π' in the list of all lexicographically sorted suffixes. We can describe (i) and (ii) by means of three values: the length ℓ_i of π', and the endpoints l_i and r_i of the interval considered in (ii). For example, let $T = mississippi\$$ (see Figure 1), and $\pi = stpissi$. For $i = 1$, we have $\pi' = s$, so $\ell_1 = 1$ and $[l_1, r_1] = [9, 12]$ (suffixes starting with s). For $i = 2$, we have $\pi' = \epsilon$, so $\ell_2 = 0$ and $[l_2, r_2] = [1, n] = [1, 12]$ (all suffixes start with the empty string). For $i = 3$, we have $\pi' = pi$, so $\ell_3 = 2$, and $[l_3, r_3] = [7, 7]$ (suffixes starting with pi). For $i = 4$, we have $\pi' = issi$, so $\ell_4 = 4$, and $[l_4, r_4] = [4, 5]$ (suffixes starting with $issi$). One can proceed analogously for $i = 5, 6, 7$.

Computing matching statistics for strings

We will first describe the algorithm by Ohlebusch et al. [17], emphasizing the ideas that we will generalize when switching to Wheeler DFAs. The algorithms computes the matching statistics using a number of iterations linear in m by exploiting the backward search. We start from the end of π, and we use the backward search (starting from the interval $[1, n]$ which corresponds to the set of suffixes prefixed by the empty string) to find the interval of all occurrences of the last character of π in T (if any). Then, starting from the new interval, we use the backward search to find all the occurrences of the suffix of length 2 of π in T (if any), and so on. At some point, it may happen that for some $i \le m + 1$ we have that $\pi[i \mathinner{.\,.} m]$ occurs in T, but the next application of the backward search returns the empty interval, so that $\pi[i - 1 \mathinner{.\,.} m]$ does not occur in T (the case $i = m + 1$ corresponds to the initial setting when $\pi[i \mathinner{.\,.} m]$ is the empty string). We distinguish two cases:

- (Case 1) If $l_i = 1$ and $r_i = n$, this means that all suffixes of T are prefixed by $\pi[i \mathinner{.\,.} m]$. This may happen in particular if $i = m + 1$: this means that the first backward search has been unsuccessful. We immediately conclude that character $\pi[i - 1]$ does not occur in T, so $\ell_{i-1} = 0$ and $[l_{i-1}, r_{i-1}] = [1, n]$ (because all suffixes start with the empty string). In this case, in the following iterations of the algorithm, we can simply discard $[i - 1, m]$: when for $i' \le i - 2$ we will be searching for the longest prefix of $\pi[i', m]$ occurring in T, it will suffice to search for the longest prefix of $\pi[i', i - 2]$ occurring in T.

- (Case 2) If $l_i > 1$ or $r_i < n$, this means that the number of suffixes of T starting with $\pi[i \mathinner{.\,.} m]$ is less than n. Now, every suffix starting with $\pi[i \mathinner{.\,.} m]$ also starts with $\pi[i \mathinner{.\,.} m - 1]$. If the number of suffixes starting with $\pi[i \mathinner{.\,.} m - 1]$ is equal to the number of suffixes starting with $\pi[i \mathinner{.\,.} m]$, then also $\pi[i - 1 \mathinner{.\,.} m - 1]$ does not occur in T. More in general, for $j \le m - 1$ we can have that $\pi[i - 1 \mathinner{.\,.} j]$ occurs in T only if the number of suffixes starting with $\pi[i \mathinner{.\,.} j]$ is larger than the number of suffixes starting with $\pi[i \mathinner{.\,.} m]$. Since we are interested in maximal matches, we want j to be as large as possible: we will show later how to compute the largest integer j such that the number of suffixes starting with $\pi[i . j]$ is larger than the number of suffixes starting with $\pi[i \mathinner{.\,.} m]$. Notice that j always exists, because all n suffixes start with the empty string, but less than n suffixes start with $\pi[i \mathinner{.\,.} m]$. After determining j we discard $\pi[j + 1 \mathinner{.\,.} m]$ (so in the following iterations of the algorithm we will simply consider $[1 \mathinner{.\,.} j]$), and we recursively apply the

backward search starting from the interval associated with the occurrences of $\pi[i \mathinner{.\,.} j]$ — we will also see how to compute this interval.

Let us apply the above algorithm to $T = mississippi\$$ and $\pi = stpissi$. We start with the interval $[1, n] = [1, 12]$, corresponding to the empty pattern, and character $\pi[7] = i$. A backward step yields the interval $[l_7, r_7] = [2, 5]$ (suffixes starting with i), so $\ell_7 = 1$. Now, we apply a backward step from $[2, 5]$ and $\pi[6] = s$, obtaining $[l_6, r_6] = [9, 10]$ (suffixes starting with si), so $\ell_6 = 2$. Again, we apply a backward step from $[9, 10]$ and $\pi[5] = s$, obtaining $[l_5, r_5] = [11, 12]$ (suffixes starting with ssi), so $\ell_5 = 3$. Again, we apply a backward step from $[11, 12]$ and $\pi[4] = i$, obtaining $[l_4, r_4] = [4, 5]$ (suffixes starting with $issi$), so $\ell_4 = 4$. We now apply a backward step from $[4, 5]$ and $\pi[3] = p$, and we obtain the empty interval. This means that no suffix starts with $pissi$. Notice in Figure 1 that the number of suffixes starting with $issi$ is equal to the number of suffixes starting with iss or is, but the number of suffixes starting with i is bigger. As a consequence, we consider the interval of all suffixes starting with i — which is $[2, 5]$ — and we apply a backward step with $\pi[3] = p$. This time the backward step is successful, and we obtain $[l_3, r_3] = [7, 7]$ (suffixes starting with pi), and $\ell_3 = 2$. We now apply a backward step from $[7, 7]$ and $\pi[2] = t$, obtaining the empty interval. This means that no suffix starts with tpi. Notice in Figure 1 that the number of suffixes starting with p is bigger than the number of suffixes starting with pi. The corresponding interval is $[7, 8]$, but a backward step with $\pi[2] = t$ is still unsuccessful (so no suffix starts with tp). The number of suffixes starting with p is smaller than the number of suffixes starting with the empty string (which is equal to $n = 12$), so we apply a backward step with $[1, 12]$ and $\pi[2] = t$. Since the backward step is still unsuccessful, we conclude that $\pi[2] = t$ does not occur in S, so $[l_2, r_2] = [1, n] = [1, 12]$ and $\ell_2 = 0$. Finally, we start again from the whole interval $[1, 12]$, and a backward step with $\pi[1] = s$ returns $[l_1, r_1] = [9, 12]$ (suffixes starting with s), so $\ell_1 = 1$.

It is easy to see that the number of iterations is linear in m. Indeed, every time we apply a backward step, either we move to the left across π to compute a new matching statistic, or we increase by at least 1 the length of the suffix of π which is forever discarded. This implies that the number of iterations is bounded by $2|\pi| = 2m$.

We are only left with showing (i) how to compute j and (ii) the interval of all suffixes starting with $\pi[i \mathinner{.\,.} j]$ in Case 2 of the algorithm. To this end, we introduce the longest common prefix (LCP) array $\text{LCP} = \text{LCP}[2, n]$ of T. We define $\text{LCP}[i]$ to be the length of the longest common prefix of the $(i - 1)$-st lexicographically smallest suffix of T and the i-th lexicographically smallest suffix of T. In Figure 1 we have $\text{LCP}[5] = 4$ because the fourth lexicographically smallest suffix of T is $issippi\$$, the fifth lexicographically smallest suffix of T is $ississippi\$$, and the longest common prefix of $issippi\$$ and $ississippi\$$ is $issi$, which has length 4. Remember that in the example the backward search starting from $[4, 5]$ (suffixes starting with $issi$) and p was unsuccessful, so computing j means determining the longest prefix of $issi$ such that the the number of suffixes starting with such a prefix is bigger than 2. This is easy to compute by using the LCP array: the longest such prefix is the one of length $\max\{\text{LCP}[4], \text{LCP}[6]\} = \max\{1, 0\} = 1$, so that the desired prefix is i. As a consequence, we

are only left with showing how to compute the interval of all suffixes starting with the prefix i — which is [2, 5]. Notice that in order to compute this interval, it is enough to expand the interval [4, 6] in both directions as long as the LCP value does not go below 1. Since $\text{LCP}[4] = 1, \text{LCP}[3] = 1$, and $\text{LCP}[2] = 0$, and we already know that $\text{LCP}[6] = 0$, we conclude that the desired interval is [2, 5]. In other words, given a position t, we must be able to compute the biggest integer k less than t such that $\text{LCP}[k] < \text{LCP}[t]$, and the smallest integer k bigger than t such that $\text{LCP}[k] < \text{LCP}[t]$ (in our case, $t = 4$). These queries are called PSV ("previous smaller value") and NSV ("next smaller value") queries. The LCP array can be augmented in such a way that PSV and NSV queries can be solved efficiently: different space-time trade-offs are possible, we refer the reader to [17] for details.

Matching statistics for Wheeler DFAs

Let us define Wheeler DFAs [7].

Definition 1. *Let $\mathscr{A} = (Q, E, s_0, F)$ be a DFA. A* Wheeler order *on $\mathscr{A}$ is a total order on Q such that $s_0 \leq u$ for every $u \in Q$ and:*

(Axiom 1) If $u, v \in Q$ and $u < v$, then $\lambda(u) \leq \lambda(v)$.

(Axiom 2) If $(u', u), (v', v) \in E, \lambda(u) = \lambda(v)$ and $u < v$, then $u' < v'$.

A DFA $\mathscr{A}$ is Wheeler *if it admits a Wheeler order*.

It is immediate to check that this definition is equivalent to the one in [7], where it was shown that if a DFA $\mathscr{A}$ admits a Wheeler order , then is uniquely determined (that is, is *the* Wheeler order on $\mathscr{A}$). In the following, we fix a Wheeler DFA $\mathscr{A} = (Q, E, s_0, F)$, where we assume $Q = \{u_1, \dots, u_n\}$, with $u_1 < u_2 < \cdots < u_n$ in the Wheeler order, and u_1 coincides with the initial state s_0. See Figure 2 for an example.

We now show that a Wheeler order can be seen of as a permutation of the set of all states playing the same role as the suffix array of a string. In the following, it will be expedient to (conceptually) assume that s_0 has a self-loop labeled # (this is consistent with Axiom 1, because # $\prec a$ for every $a \in \Sigma$). This implies that every state has at least one incoming edge, so for every state u_i there exists at least one infinite string $\alpha \in \Sigma^\omega$ that can be read starting from u_i and following edges in a backward fashion (for example, in Figure 2 for u_9 such a string is $ce\#\#\#\dots$). We denote by I_{u_i} the set of all such strings. Formally:

Definition 2. *Let $i \in [1, n]$. For every state $u_i \in Q$ define:*

$$I_{u_i} = \left\{ \alpha \in \Sigma^\omega \mid \text{there exist integers } f_1, f_2, \dots \text{ in } [1, n] \text{ such that } (i)\, f_1 = i, \right.$$
$$\left. (ii)\, (u_{f_{k+1}}, u_{f_k}) \in E \text{ for every } k \geq 1 \text{ and } (iii)\, \alpha = \lambda(u_{f_1})\lambda(u_{f_2})\dots \right\}.$$

For example, in Figure 2 we have $I_{u_3} = \{abdg\#\#\#\dots, abeh\#\#\#\dots, acei\#\#\#\dots\}$.

The following lemma shows that the permutation of the states defined by the Wheeler order is the one lexicographically sorting the strings entering each state, just like the permutation

defined by the suffix array lexicographically sorts the suffixes of the strings (a suffix is seen as a string "leaving" a text position).

Lemma 3. *Let $i, j \in [1, n]$, with $i < j$. Let $\alpha \in I_{u_i}$ and $\beta \in I_{u_j}$. Then, $\alpha \preceq \beta$.*

Proof. Let $f_1, f_2, \ldots$ in $[1, n]$ be such that (i) $f_1 = i$, (ii) $(u_{f_{k+1}}, u_{f_k}) \in E$ for every $k \geq 1$ and (iii) $\alpha = \lambda(u_{f_1})\lambda(u_{f_2})\ldots$. Analogously, let $g_1, g_2, \ldots$ in $[1, n]$ be such that (i) $g_1 = j$, (ii) $(u_{g_{k+1}}, u_{g_k}) \in E$ for every $k \geq 1$ and (iii) $\beta = \lambda(u_{g_1})\lambda(u_{g_2})\ldots$ Let $\alpha \neq \beta$. We must prove that $\alpha \prec \beta$. Let $p \geq 1$ be the smallest integer such that the p-th character of α is different than the p-th character of β. In other words, we know that $\lambda(u_{f_1}) = \lambda(u_{g_1}), \lambda(u_{f_2}) = \lambda(u_{g_2}), \ldots, \lambda(u_{f_{p-1}}) = \lambda(u_{g_{p-1}})$, but $\lambda(u_{f_p}) \neq \lambda(u_{g_p})$. We must prove that $\lambda(u_{f_p}) \prec \lambda(u_{g_p})$. Since $\lambda(u_{f_1}) = \lambda(u_{g_1}) f_1 = i < j = g_1$, and $(u_{f_2}, u_{f_1}), (u_{g_2}, u_{g_1}) \in E$, from Axiom 2 we obtain $f_2 < g_2$. Since $\lambda(u_{f_2}) = \lambda(u_{g_2})$, $f_2 < g_2$, and $(u_{f_3}, u_{f_2}), (u_{g_3}, u_{g_2}) \in E$, from Axiom 2 we obtain $f_3 < g_3$. By iterating this argument, we conclude $f_p < g_p$. By Axiom 1, we obtain $\lambda(u_{f_p}) \preceq \lambda(u_{g_p})$. Since $\lambda(u_{f_p}) \neq \lambda(u_{g_p})$, we conclude $\lambda(u_{f_p}) \prec \lambda(u_{g_p})$. □

If we think of a string as a labeled path, then the suffix array sorts the strings that can be read from each position by moving forward (that is, the suffixes of the string), while the Wheeler order sorts the strings that can be read from each position by moving backward towards the initial state. The underlying idea is the same: the forward vs backward difference is only due to historical reasons [6]. To compute the matching statistics on Wheeler DFA we reason as in the previous section replacing backward search with the *forward search* [6] defined as follows: given an interval $[i, j]$ in $[1, n]$ and $a \in \Sigma$, find the (possibly empty) interval $[i', j']$ in $[1, n]$ such that a state $v_{k'}$ is reachable from some state v_k, with $i \leq k \leq j$, through an edge labeled a, if and only if $i' \leq k' \leq j'$ (this easily follows by using the axioms of Definition 1). For a constant size alphabet, given $[i, j]$ and a then $[i', j']$ can be determined in constant time. Given a string $\pi \in \Sigma^*$, if we start from the whole set of states and repeatedly apply the forward search we reach the set of all states u_i for which there exists $\alpha \in I_{u_i}$ prefixed by π^R; this is an interval with respect to the Wheeler order: in the following we call this interval $T(\pi)$.

Because of the forward vs backward difference the problem of matching statistics will be defined in a symmetrical way on Wheeler DFAs. Given a pattern $\pi = \pi[1 . . m]$, for every $1 \leq i \leq m$ we want to determine (i) the longest suffix π' of $\pi[1 . . i]$ which occurs in the Wheeler DFA $\mathscr{A}$ (that is, that can be read somewhere on $\mathscr{A}$ by concatenating edges), and *(ii)* the endpoints of the interval $T(\pi')$.

Broadly speaking, we can apply the same idea of the algorithm for strings, but in a symmetrical way. We start from the *beginning* of π (not from the end of π), and initially we consider the whole set of states. We repeatedly apply the *forward* search (not the backward search), until the forward search returns the empty interval for some $i \geq 0$. This means that $\pi[1 . . i + 1]$ does not occur in $\mathscr{A}$. Then, if $T(\pi[1 . . i])$ is the whole set of states, we conclude that the character $\pi[i + 1]$ labels no edge in the graph. Otherwise, we must find the smallest j such that $T(\pi[1 . . i])$ is strictly contained in $T(\pi[j . . i])$ (that is, we must determine the longest

suffix $\pi[j \mathinner{\ldotp\ldotp} i]$ of $\pi[1 \mathinner{\ldotp\ldotp} i]$ which reaches more states than $\pi[1 \mathinner{\ldotp\ldotp} i]$). Then we must determine the endpoints of the interval $T(\pi[j \mathinner{\ldotp\ldotp} i])$ so that we can go on with the forward search.

The challenge now is to find a way to solve the same subproblems that we identified in Case 2 of the algorithm for strings. In other words, we must find a way to determine j and find the endpoints of the interval $T(\pi[j \mathinner{\ldotp\ldotp} i])$. We will show that the solution is not as simple as the one for the algorithm on strings.

The LCP array and matching statistics for Wheeler DFAs

We start observing that I_{u_i} may be an infinite set. For example, in Figure 2, we have

$$I_{u_2} = \{ aaaaa\ldots, abdf\,\#\,\#\,\#\,\ldots, aabdf\,\#\,\#\,\#\,\ldots, aaabdf\,\#\,\#\,\#\,\ldots, \ldots \}$$

In general, an infinite set of (lexicographically sorted) strings in Σ^ω need not admit a minimum or a maximum. For example, the set $\{ baaaa\ldots, abaaa\ldots, aabaa\ldots, aaaba\ldots \}$ does not admit a minimum (but only the *infimum* string $aaaaa\ldots$). Nonetheless, Lemma 3 implies that each I_{u_i} admits both a minimum and a maximum. For example, the minimum is obtained as follows. Let $f_1 = i$, and for every $k \geq 1$, recursively let f_{k+1} be the smallest integer in $[1, n]$ such that $(u_{f_{k+1}}, u_{f_k}) \in E$. Then, the minimum of I_{u_i} is $\lambda(u_{f_1})\lambda(u_{f_2})\ldots$, and analogously one can determine the maximum.

In the following, we will denote the minimum and the maximum of I_{u_i} by $\min_i$ and $\max_i$, respectively (for example, in Figure 2 we have $\min_2 = aaaaa\ldots$, and $\max_2 = abdf\,\#\,\#\,\#\,\ldots$). Lemma 3 implies that:

$$\min_1 \preceq \max_1 \preceq \min_2 \preceq \max_2 \preceq \cdots \preceq \max_{n-1} \preceq \min_n \preceq \max_n.$$

This suggests to generalize the LCP array as follows. Given $\alpha, \beta \in \Sigma^* \cup \Sigma^\omega$, let $\mathrm{lcp}(\alpha, \beta)$ be the length of the longest common prefix of α and β (if $\alpha = \beta \in \Sigma^\omega$, define $\mathrm{lcp}(\alpha, \beta) = \infty$).

Definition 4. *The* LCP-array *of a Wheeler automaton $\mathscr{A}$ is the array* $\mathrm{LCP}_{\mathscr{A}} = \mathrm{LCP}_{\mathscr{A}}[2, 2n]$ *which contains the following $2n - 1$ values in this order:* $\mathrm{lcp}(\min_1, \max_1), \mathrm{lcp}(\max_1, \min_2), \mathrm{lcp}(\min_2, \max_2), \ldots, \mathrm{lcp}(\max_{n-1}, \min_n), \mathrm{lcp}(\min_n, \max_n).$

From the above characterization of $\min_i$ and $\max_i$, one can prove that for every entry either $\mathrm{LCP}_{\mathscr{A}}[i] = \infty$ or $\mathrm{LCP}_{\mathscr{A}}[i] < 3n$ (it follows from Fine and Wilf Theorem [18, 19]), and one can design a polynomial time algorithm to compute $\mathrm{LCP}_{\mathscr{A}}$.

Unfortunately, the array $\mathrm{LCP}_{\mathscr{A}}$ alone is not sufficient for computing matching statistics. Assume that $T(\pi) = \{ u_r, u_{r+1}, \ldots, u_{s-1}, u_s \}$, and that when we apply the forward search by adding a character c, we obtain $T(\pi c) = \varnothing$. We must then determine the largest suffix π' of $T(\pi)$ such that $T(\pi)$ is *strictly* contained in $T(\pi')$. Suppose that *every* string in I_{u_r} is prefixed by π^R, and *every* string in I_{u_s} is prefixed by π^R. In particular, both $\min_r$ and $\max_s$ are prefixed

by π^R. In this case, we can proceed like in the algorithm for strings: the desired suffix π' is the one having length $\max\{\mathrm{lcp}(\max_{r-1}, \min_r), \mathrm{lcp}(\max_s, \min_{s+1})\}$, which can be determined using $\mathrm{LCP}_{\mathscr{A}}$. However, in general, even if some string in I_{u_r} must be prefixed by π^R, the string $\min_r$ need not be prefixed by π^R, and similarly $\max_s$ need not be prefixed by π^R. The worst-case scenario occurs when $r = s$. Consider Figure 2, and assume that $\pi = heba$. Then, we have $r = s = 3$ (note that $abeh\,\#\,\#\,\#\,\ldots$ is a string in I_{u_3} prefixed by π^R). However, both $\min_3 = abdg\,\#\,\#\,\#\,\ldots$, and $\max_3 = acei\,\#\,\#\,\#\,\ldots$, are not prefixed by π^R. Notice that $\mathrm{lcp}(\max_2, \min_3) = 3$ and $\mathrm{lcp}(\max_3, \min_4) = 3$, but π' is not the suffix of length 3 of π. Indeed, since $\min_3$ is only prefixed by the prefix of π^R of length 2, and $\max_3$ is only prefixed by the prefix of π^R of length 1, we conclude that it must be $|\pi'| = 2$. In general, the desired suffix π' is the one having length $|\pi'|$ given by:

$$\max\left\{\min\left\{\mathrm{lcp}(\max_{r-1}, \min_r), \mathrm{lcp}\left(\min_r, \pi^R\right)\right\}, \min\left\{\mathrm{lcp}\left(\pi^R, \max_s\right), \mathrm{lcp}(\max_s, \min_{s+1})\right\}\right\}.$$

(1)

The above formula shows that, in order to compute π', in addition to $\mathrm{LCP}_{\mathscr{A}}$ it suffices to know the values $\mathrm{lcp}\left(\min_r, \pi^R\right)$ and $\mathrm{lcp}\left(\pi^R, \max_s\right)$ (π' is a suffix of π, so it is determined by its length). We now show how our algorithm can efficiently maintain the current pattern π, the set $T(\pi) = \{u_r, u_{r+1}, \ldots, u_{s-1}, u_s\}$ and the values $\mathrm{lcp}\left(\min_r, \pi^R\right)$ and $\mathrm{lcp}\left(\pi^R, \max_s\right)$ during the computation of the matching statistics. We assume that the input automaton is encoded with the rank/select data structures supporting the execution of a step of forward search in $O(\log|\Sigma|)$ time, see [6] for details. In addition, we will use the following result.

Lemma 5. *Let $A[1, n]$ be a sequence of values over an ordered alphabet Σ. Consider the following queries: (i) given $i, j \in [1 \,.\, .\, n]$, compute the minimum value in $S[i \,.\, .\, j]$, and (ii) given $t \in [1 \,.\, .\, n]$ and $c \in \Sigma$, determine the biggest $k < t$ (or the smallest $k > t$) such that $A[k] < c$. Then, A can be augumented with a data structure of $2n + o(n)$ bits such that query (i) can be answered in constant time and query (ii) can be answered in $O(\log n)$ time.*

Proof. There exists a data structure of $2n + o(n)$ bits that allows to solve range minimum queries in constant time [20], so using A we can solve queries (i) in constant time. Now, let us show how to solve queries (ii). Let f_1 be the answer of query (i) on input $i = \lceil t/2 \rceil$ and $j = t - 1$. If $f_1 < c$, then we must keep searching in the interval $[\lceil t/2 \rceil, t - 1]$, otherwise, we must keep searching in the interval $[1, \lceil t/2 \rceil - 1]$. In other words, we can answer a query (ii) by means of a binary search on $[1, t - 1]$, which takes $O(\log t)$ (and so $O(\log n)$) time. $\square$

Notice that query (ii) can be seen as a variant of PSV and NSV queries. In the following, we assume that the array $\mathrm{LCP}_{\mathscr{A}}$ has been augmented with the data structure of Lemma 5.

At the beginning we have $\pi = \epsilon$, so $T(\epsilon) = \{1, 2, \ldots, n\}$ and trivially $\mathrm{lcp}\left(\min_r, \pi^R\right) = \mathrm{lcp}\left(\pi^R, \max_s\right) = 0$. At each iteration we perform a step of forward search computing $T(\pi c)$ given $T(\pi)$; then we distinguish two cases according to whether $T(\pi c)$ is empty or not.

Case 1. $T(\pi c) = \{u_{r'}, u_{r'+1}, \ldots, u_{s'-1}, u_{s'}\}$ is not empty. In that case πc will become the pattern at the next iteration. Since we already have $T(\pi c)$ we are left with the task of computing $\mathrm{Icp}(\min_{r'}, c\pi^R)$ and $\mathrm{Icp}(c\pi^R, \max_{s'})$. We only show how to compute $\mathrm{Icp}(\min_{r'}, c\pi^R)$, the latter computation being analogous. Let k be the smallest integer in $[1, n]$ such that $(u_k, u_{r'}) \in E$. Notice that we can easily compute k by means of standard rank/select operations on the compact data structure used to encode $\mathcal{A}$. Since $u_{r'} \in T(\pi c)$, it must be $k \leq s$. Moreover, the characterization of $\min_{r'}$ that we described above implies that $\min_{r'} = c\min_k$, hence $\mathrm{Icp}(\min_{r'}, c\pi^R) = \mathrm{lcp}(c\min_k, c\pi^R) = 1 + \mathrm{lcp}(\min_k, \pi^R)$. To compute $\mathrm{Icp}(\min_k, \pi^R)$ we distinguish two subcases:

  a. $k > r$, hence $r < k \leq s$. Since $u_r, u_s \in T(\pi)$, there exist $\alpha \in I_{u_r}$ and $\beta \in I_{u_s}$ both prefixed by π^R. But $\alpha \preceq \max_r \preceq \min_k \preceq \min_s \preceq \beta$, so $\min_k$ is also prefixed by π^R, and we conclude $\mathrm{lcp}(\min_k, \pi^R) = |\pi|$.

  b. $k \leq r$. In this case, we have $\min_k \preceq \max_k \preceq \min_{k+1} \prec \max_{k+1} \preceq \cdots \preceq \min_r \prec \pi^R$, and therefore $\mathrm{Icp}(\min_k, \pi^R)$ is equal to

$$\min\left\{\mathrm{lcp}(\min_k, \max_k), \mathrm{lcp}(\max_k, \min_{k+1}), \mathrm{lcp}(\min_{k+1}, \max_{k+1}), \ldots, \mathrm{lcp}(\min_r, \pi^R)\right\}$$

With the above formula we can compute $\mathrm{Icp}(\min_k, \pi^R)$ using query (i) of Lemma 5 over the range $\mathrm{LCP}_{\mathcal{A}}[2k, 2r-1]$ and the value $\mathrm{Icp}(\min_r, \pi^R)$.

Case 2. $T(\pi c)$ is empty. In this case at the next iteration the pattern will be largest suffix π' of π such that $T(\pi)$ is *strictly* contained in $T(\pi') = \{u_{r''}, \ldots, u_{s''}\}$. We compute $|\pi'|$ using (1); if $|\pi'| > \mathrm{lcp}(\min_r, \pi^R)$ we set $r'' = r$, otherwise we apply query (ii) of Lemma 5 to find the rightmost entry r'' in $\mathrm{LCP}_{\mathcal{A}}[2, 2r-1]$ smaller than $|\pi'|$. Computing s'' is analogous.

Given $T(\pi') = \{u_{r''}, u_{r''+1}, \ldots, u_{s''-1}, u_{s''}\}$, where $r'' \leq r, s \leq s''$, and at least one inequality is strict, we want to compute $\mathrm{Icp}(\min_{r''}, (\pi')^R)$ and $\mathrm{Icp}((\pi')^R, \max_{s''})$. We only consider $\mathrm{Icp}(\min_{r''}, (\pi')^R)$, the latter computation being analogous. We distinguish two subcases:

  a. $r'' = r$. Then $\mathrm{Icp}(\min_{r''}, (\pi')^R) = \mathrm{Icp}(\min_r, (\pi')^R) = \min\{\mathrm{lcp}(\min_r, \pi^R), |\pi'|\}$.

  b. $r'' < r$. In particular, since $u_{r''}$ is the left endpoint of $T(\pi')$ and $|T(\pi')| \geq 2$, one can prove like in Case 1a) that $\max_{r''}$ is prefixed by $(\pi')^R$. We immediately conclude that $\mathrm{lcp}(\min_{r''}, (\pi')^R) = \min\{\mathrm{lcp}(\min_{r''}, \max_{r''}), |\pi'|\}$, which can be immediately computed since $\mathrm{Icp}(\min_{r''}, \max_{r''})$ is a value stored in $\mathrm{LCP}_{\mathcal{A}}$.

We can summarize the above discussion as follows.

Theorem 6. *Given a Wheeler DFA $\mathcal{A}$, there exists a data structure occupying $O(|\mathcal{A}|)$ words which can compute the pattern matching statistics of a pattern P in time $O(|P|\log|\mathcal{A}|)$.*

Funding

References

[1]. Ferragina P and Manzini G, "Opportunistic data structures with applications," in Proc. 41 st Annual Symposium on Foundations of Computer Science (FOCS'00), 2000, pp. 390–398.

[2]. Burrows M and Wheeler DJ, "A block-sorting lossless data compression algorithm," Tech. Rep, 1994.

[3]. Ferragina P, Luccio F, Manzini G, and Muthukrishnan S, "Structuring labeled trees for optimal succinctness, and beyond," in proc. 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05), 2005, pp. 184–193.

[4]. Bowe A, Onodera T, Sadakane K, and Shibuya T, "Succinct de Bruijn graphs," in Algorithms in Bioinformatics, Berlin, Heidelberg, 2012, pp. 225–235, Springer Berlin Heidelberg.

[5]. Mäkinen V, Välimäki N, and Sirén J, "Indexing graphs for path queries with applications in genome research," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 11, pp. 375–388, 2014. [PubMed: 26355784]

[6]. Gagie T, Manzini G, and Sirén J, "Wheeler graphs: A framework for BWT-based data structures," Theoret. Comput. Sci, vol. 698, pp. 67–78, 2017. [PubMed: 29276331]

[7]. Alanko J, D'Agostino G, Policriti A, and Prezza N, "Regular languages meet prefix sorting," in Proc. of the 31st Symposium on Discrete Algorithms, (SODA'20). 2020. pp. 911–930, SIAM.

[8]. Cotumaccio N and Prezza N, "On indexing and compressing finite automata," in Proc. of the 32nd Symposium on Discrete Algorithms, (SODA'21). 2021. pp. 2585–2599, SIAM.

[9]. Cotumaccio N, "Graphs can be succinctly indexed for pattern matching in $O(|E|^2+|V|^{5/2})$ time," in 2022 Data Compression Conference (DCC), 2022, pp. 272–281.

[10]. Alanko J, D'Agostino G, Policriti A, and Prezza N, "Wheeler languages," Information and Computation, vol. 281, pp. 104820, 2021.

[11]. Gusfield D, Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge University Press, 1997.

[12]. Chang WI and Lawler EL, "Sublinear approximate string matching and biological applications," Algorithmica, vol. 12, pp. 327–344, 2005.

[13]. Weiner P, "Linear pattern matching algorithms," in Proc. 14th IEEE Annual Symposium on Switching and Automata Theory, 1973, pp. 1–11.

[14]. Manber U and Myers G, "Suffix arrays: A new method for on-line string searches," SIAM J. Comput, vol. 22, no. 5, pp. 935–948, 1993.

[15]. Abouelhoda MI, Kurtz S, and Ohlebusch E, "Replacing suffix trees with enhanced suffix arrays," J. of Discrete Algorithms, vol. 2, no. 1, pp. 53–86, 2004.

[16]. Sadakane K, "Compressed suffix trees with full functionality," Theor. Comp. Sys, vol. 41, no. 4, pp. 589–607, 2007.

[17]. Ohlebusch E, Gog S, and Kügell A, "Computing matching statistics and maximal exact matches on compressed full-text indexes," in Proceedings of the 17th International Conference on String Processing and Information Retrieval (SPIRE'10), Berlin, Heidelberg, 2010, p. 347–358, Springer-Verlag.

[18]. Fine NJ and Wilf HS, "Uniqueness theorem for periodic functions," Proc. Amer. Math. Soc, no. 16, pp. 109–114, 1965.

[19]. Mantaci S, Restivo A, Rosone G, and Sciortino M, "An extension of the Burrows-Wheeler transform," Theor. Comput. Sci, vol. 387, no. 3, pp. 298–312, 2007.

[20]. Fischer Johannes, "Optimal succinctness for range minimum queries," in LATIN 2010: Theoretical Informatics, López-Ortiz Alejandro, Ed., Berlin, Heidelberg, 2010, pp. 158–169, Springer Berlin Heidelberg.

i	Sorted suffixes	LCP	SA	BWT
1	$		12	i
2	i$	0	11	p
3	ippi$	1	8	s
4	issippi$	1	5	s
5	ississippi$	4	2	m
6	mississippi$	0	1	$
7	pi$	0	10	p
8	ppi$	1	9	i
9	sippi$	0	7	s
10	sissippi$	2	4	s
11	ssippi$	1	6	i
12	ssissippi$	3	3	i

Figure 1:
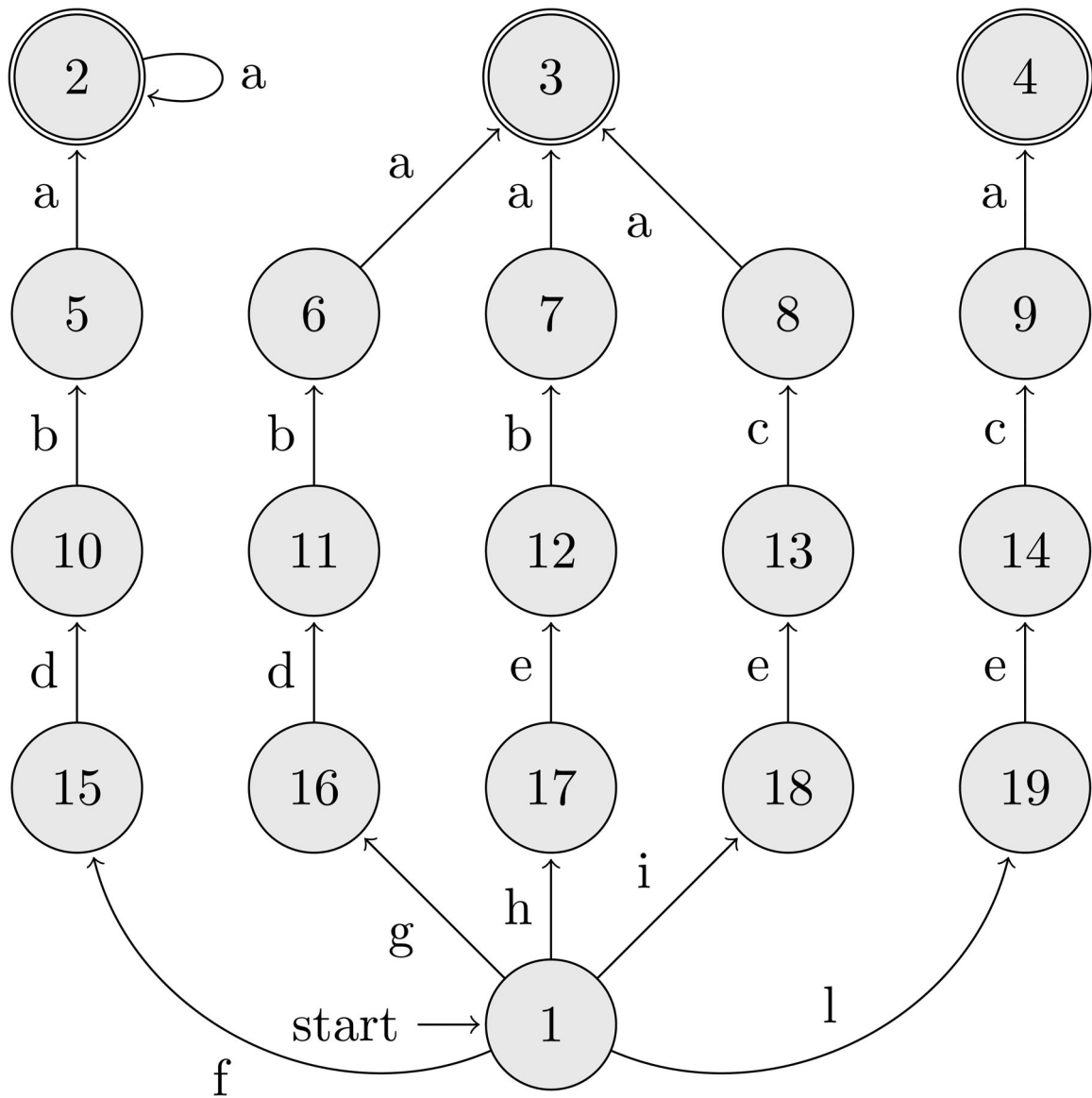The sorted suffixes of "mississippi$" and the LCP, SA, and BWT arrays.

Figure 2:
A Wheeler DFA. States are numbered according to their positions in the Wheeler order.