

# Linear Time Inference of Strings from Cover Arrays using a Binary Alphabet

Tanaeem M. Moosa\*, Sumaiya Nazeen, M. Sohel Rahman, and Rezwana Reaz

A $\ell$ EDA Group

Department of CSE, BUET

Dhaka-1000, Bangladesh

{tanaeem,nazeen,msrahman,rimpi}@cse.buet.ac.bd

**Abstract.** Covers being one of the most popular form of regularities in strings, have drawn much attention over time. In this paper, we focus on the problem of linear time inference of strings from cover arrays using the least sized alphabet possible. We present an algorithm that can reconstruct a string  $x$  over a two-letter alphabet whenever a valid cover array  $C$  is given as an input. This algorithm uses several interesting combinatorial properties of cover arrays and an interesting relation between border array and cover array to achieve this. Our algorithm runs in linear time.

## 1 Introduction

A substring  $w$  of string  $x$  is called a *cover* of  $x$  if  $x$  can be constructed by concatenation and/or superposition of  $w$ . Though  $x$  is always a cover of itself, we do not consider so, in this paper. We limit our focus on the so-called *aligned covers* where the cover  $w$  needs to be a proper substring and also a *border* (i.e., a prefix and a suffix) of  $x$ . For example, the string  $x = abcababcbcab$  is constructed by the concatenation (at position 6) and superposition (at positions 9 and 12) of  $w = abcab$ . Thus  $x$  has a proper cover,  $w$  which is also a border. A string that has a proper cover is called *coverable* or *quasiperiodic*, otherwise it is *superprimitive* [1]. The array  $C$  is called the *minimal-cover* (resp. *maximal-cover*) *array* of the string  $x$  of length  $n$ , if for each  $i$ ,  $1 \leq i \leq n$ ,  $C[i]$  stores either the length of the shortest (resp. longest) cover of  $x[1..i]$ , when such a cover exists, or zero otherwise. The array  $B[1..n]$  is the *border array* of the string  $x$  if  $B[i]$  stores the length of the longest border of  $x[1..i]$ ,  $1 \leq i \leq n$ .

---

\* Currently working at Google Inc., USA.

Repetitions in strings like periods, borders, covers etc. have always been a subject of great interest for the computer scientists because of its diverse applications in fields like molecular biology, probability theory, coding theory, data compression and formal language theory. In fact, in the last two decades string periodicity has drawn a lot of attention from different disciplines of science. The famous KMP [2] pattern matching algorithm depends on the *failure function* which is nothing but the *border array*. Another well-known pattern matching algorithm namely the BOYER-MOORE algorithm [3] makes use of similar kind of repetitions in strings. Such repetitions in strings are often encoded in data structures like graphs and integer arrays [4]. Thus, researchers have shown interest not only in finding repetitions in strings but also in reconstructing strings from those repetitive information. Apostolico *et al.* [5] gave an online linear runtime algorithm computing the minimal-cover array of a string. Smyth *et al.* [6] provided an online linear runtime algorithm for computing the maximal cover array which describes all the covers of a string. The problem of reverse engineering a string was first introduced by Franěk *et al.* [7]. They proposed a method to check if an integer array is a *border array* for some string. Border arrays are better known as *failure functions* [8]. They showed an online linear time algorithm to verify if a given integer array is a border array for some string  $w$  on an unbounded alphabet. Duval *et al.* [9] gave an online linear time algorithm for bounded alphabet to solve this problem. Bannai *et al.* [4] solved the problem of inferring a string from a given suffix array on minimal sized alphabet by proposing a linear time algorithm. Smyth *et al.* discussed a possible solution of string inference problem from prefix arrays in [10].

Crochemore *et al.* [11] presented a constructive algorithm checking if an integer array is the minimal-cover or maximal-cover array of some string. When the array is valid, their algorithm produces a string over an unbounded alphabet whose cover array is the input array. All these algorithms run in linear time. Very recently, Tomohiro *et al.* [12] proposed a way to verify whether a given integer array is a valid parameterized border array (p-border array) for a binary alphabet. They further extended their work in [13] by giving an  $O(n^{1.5})$ -time  $O(n)$ -space algorithm to verify if a given integer array of length  $n$  is a valid p-border array for an unbounded alphabet.

In this paper, we address the open problem stated in [11]. We present a linear time algorithm for reconstruction of a string from cover array using least sized alphabet. Our algorithm is closely analogous to the MINARRAYTOSTRING

algorithm in [11]. We achieve the least possible size of alphabet by incorporating an interesting relation between border array and cover array of a string presented in [6]. In fact, our algorithm is able to reconstruct strings from valid cover arrays using an alphabet consisting of no more than two characters.

The rest of this paper is organized as follows. Section 2 gives an account of definitions and notations used throughout the paper. Section 3 presents the addressed problem formally and lists important properties and lemmas used later. In Section 4 we describe our algorithm and main findings. Section 5 provides some experimental analysis of our algorithm. Finally, Section 6 gives the conclusions.

## 2 Preliminaries

A string  $x$  is a finite sequence of symbols drawn from an alphabet  $\Sigma$ , where  $\Sigma[i]$  denotes the  $i$ -th symbol of  $\Sigma$ . The set of all strings over  $\Sigma$  is denoted by  $\Sigma^*$ . The *length* of a string is denoted by  $|x|$ . The *empty string*, the string of length zero, is denoted by  $\epsilon$ .

A string  $w$  is a *factor* of string  $x$  if  $x = uvw$  for two strings  $u$  and  $v$ . It is a *prefix* of  $x$  if  $u$  is empty and *suffix* of  $x$  if  $v$  is empty. It is a *proper prefix* of  $x = uv$  when  $v$  is nonempty and a *proper suffix* of  $x = uw$  when  $u$  is nonempty. For example,  $w = abc$  is a *factor* of  $x = pqabcmn$ , a *proper prefix* of  $x = pqabc$  and a *proper suffix* of  $x = abcmn$ , where  $u = pq$ ,  $v = mn$  and  $w, u, v, x \in \Sigma^*$ .

A string  $u$  is a *period* of  $x$  if  $x$  is a prefix of  $u^k$  for some positive integer  $k$ , or equivalently if  $x$  is a prefix of  $ux$ . The *period* of  $x$  is the shortest period of  $x$ . For example, if  $x = abcabcab$ , then  $abc$ ,  $abcabc$  and the string  $x$  itself are periods of  $x$ , while  $abc$  is the *period* of  $x$ .

A string  $u$  is a *border* of  $x$  if  $u$  is a *prefix* and a *suffix* of  $x$  and  $u \neq x$ . A *border*  $u$  of  $x[1..i]$  with  $i > 0$  has one of the two following forms:

- $u = \epsilon$
- $u = x[1..j]x[j+1]$  with  $j+1 < i$  and where  $x[1..j]$  is a border of  $x[1..i-1]$  and  $x[i] = x[j+1]$

Thus, a *border*  $u$  of a regular string  $x = x[1..n]$  is a proper *prefix* of  $x$  that is also a *suffix* of  $x$ ; thus  $u = x[1..b] = x[n-b+1..n]$  for some  $b \in 0..n-1$ .

The *border array* of a *regular string*  $x = x[1..n]$  is an integer array  $B = B[1..n]$  such that, for every  $i \in 1..n$ ,  $B[i]$  is the length of the longest border of  $x[1..i]$ .

A string  $w$  of length  $m$  is a *cover* of string  $x[1..n]$  if both  $m < n$  and there exists a set of positions  $P \subseteq \{1, \dots, n - m + 1\}$  satisfying  $x[i..i + m - 1] = w$  for all  $i \in P$  and  $\bigcup_{i \in P} \{i, \dots, i + m - 1\} = \{1, \dots, n\}$ . Therefore, if substring  $w$  of string  $x$  is a *cover* of  $x$ , then  $x$  can be constructed by concatenation and/or superposition of  $w$ . Though  $x$  is always a cover of itself, we do not consider so, in this paper. We limit our focus on the so-called *aligned covers* where the cover  $w$  needs to be a *proper substring* and also a *border* (i.e., a prefix and a suffix) of  $x$ . For example, the string  $x = abcababcbcab$  has proper cover  $w = abcab$  which is also a border. A string that has a proper cover is called *coverable* or *quasiperiodic*, otherwise it is *superprimitive*.

The *minimal-cover array*  $C$  of  $x$  is the array of integers  $C[1..n]$  for which  $C[i]$ ,  $1 \leq i \leq n$ , stores the length of the shortest cover of the prefix  $x[1..i]$ , if such a cover exists, or zero otherwise. The *maximal-cover array*  $C^M$  stores longest cover at each position instead. An example is given below. In what follows, we mean by *cover array*  $C$ , the minimal cover array unless otherwise specified. An example of minimal and maximal cover array is given in Figure 2.

$i$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
$x[i]$	a	b	a	a	b	a	b	a	a	b	a	a	b	a	b	a	a	b	a	b	a	b	a
$C[i]$	0	0	0	0	0	3	0	3	0	5	3	0	5	3	0	3	0	5	3	0	3	0	3
$C^M[i]$	0	0	0	0	0	3	0	3	0	5	6	0	5	6	0	8	9	10	11	0	8	0	3

**Fig. 1.** Illustration of *minimal* and *maximal cover array*.

Adopting the graphical approach described in [11], we define the cover graph as follows:

**Definition 1** A *cover graph*  $G = (V, E)$  is an undirected graph where  $V = \{1..n\}$  and each vertex  $i$ ,  $1 \leq i \leq n$  corresponds to index  $i$  of string  $x[1..n]$ . The edge set  $E$  is defined as follows based on the equivalence relation of indices of  $x$ :

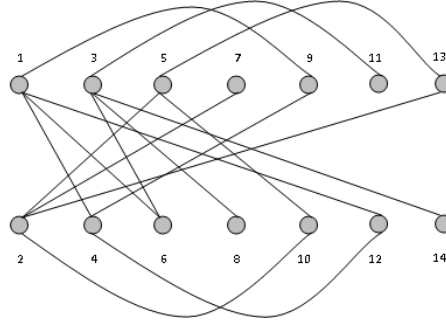
$$E = \bigcup_{i=1, \dots, n} \bigcup_{j=1, \dots, \gamma[i]} (j, i - \gamma[i] + j),$$

where  $\gamma$  is any valid cover array.

Figure 2 shows a *Cover Graph* constructed from given cover array  $C$ .

$i$	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$C[i]$	0	0	0	0	3	0	3	0	5	3	0	5	3	3

(a)



(b)

**Fig. 2.** Illustration of a *Cover Graph*. (a) Input cover array  $C$ , and (b) Corresponding *Cover Graph*.

### 3 Problem Definition & Important Properties

We start with a formal definition of the problem handled in this paper.

**Problem 1** *Linear time inference of strings using the least sized alphabet from cover arrays.*

**Input:** A valid cover array  $C$ , of length  $n$ .

**Output:** A string  $x$  of length  $n$  on a minimum sized alphabet.

Before presenting our algorithm, we mention some important properties related to the cover array and border array which will be used later.

**Property 1** (*Transitivity property of a cover* [11]) *If each of  $u$  and  $v$  covers  $x$  and  $|u| < |v|$ , then  $u$  covers  $v$ .*

**Property 2 (Totally covered position in cover array [11])** A position  $j \neq 0$  of a cover array  $C$  is called *totally covered*, if there is a position  $i > j$  for which  $C[i] \neq 0$  and  $i - C[i] + 1 \leq j - C[j] + 1 < j$ .

**Property 3 (Pruned minimal cover array [11])** Let  $C^P$  be obtained from  $C$  by setting  $C[i] = 0$  for all totally covered indices  $i$  on  $C$ . We call  $C^P$  the *pruned minimal cover array* of  $x$ . Figure 3 shows an example of pruned minimal cover array.

$i$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
$x[i]$	a	b	a	a	b	a	b	a	a	b	a	b	a	a	b	a	a	b	a	b	a	a	b	a
$C[i]$	0	0	0	0	0	3	0	3	0	5	3	7	3	9	5	3	0	5	3	0	3	9	5	3
$C^P[i]$	0	0	0	0	0	3	0	0	0	0	0	0	0	9	5	0	0	0	0	0	0	9	5	3

**Fig. 3.** Illustration of *minimal* and *pruned minimal cover array*.

**Property 4 (Border constraint on cover array [11])** The nonzero values in  $C$  induce an equivalence relation on the positions of every string that has the minimal-cover array  $C$ . More precisely, if we find the value  $l \neq 0$  in position  $i$  of  $C$ , then this imposes the constraints

$$x[k] = x[i - l + k]$$

for  $k = 1, \dots, l$ . The positions  $k$  and  $i - l + k$  are *bidirectionally linked*.

**Property 5 ([11])** Let  $i$  and  $j$  be positions such that  $j < i$ ,  $j - C[j] \geq i - C[i]$ ,  $C[i] \neq 0$  and  $C[j] \neq 0$ . Furthermore, let  $r = j - (i - C[i] + 1)$ . If  $i - C[i] = j - C[j]$ , then  $C[r] = 0$ , otherwise if  $i - C[i] < j - C[j]$ , then  $C[r] = C[j]$ .

**Property 6 ([11])** Let  $i$  and  $j$  be positions such that  $j < i$  and  $j - C[j] < i - C[i]$ . Then  $(i - C[i]) - (j - C[j]) > C[j]/2$ .

**Property 7 ([11])** The sum of the elements of  $C^P$  does not exceed  $2n$ .

**Property 8 ([6])** For every integer  $i \in 1..n - 1$ , if  $B[i] \leq B[i - 1]$ , then  $C[i] = 0$

## 4 Our Algorithm

In this section, we present an efficient algorithm, which reconstructs a string  $x$  from a cover array  $C[1..n]$  on a binary alphabet in linear time. We call this algorithm, Algorithm SIMA (**S**tring **I**nference using **M**inimum-sized **A**lphabet). We assume that a valid cover array will be given as input. The validity of a cover array can be easily checked by Property 2 [11] and Property 6 [11] using the same approach used in [11] without changing the running time of our algorithm.

The algorithm uses the following arrays:

- $C[1..n]$ : valid cover array.
- $B[1..n]$ : border array keeping track of the lengths of longest borders.
- $x[1..n]$ : string constructed by the algorithm.

We solve the stated problem in three steps.

- *Step 1 (Array Transformation)*: Adopting the same strategy used in [11], convert the input cover array to a minimal cover array  $C$  using procedure MAXTOMIN [11] in case a maximal cover array is given as input.
- *Step 2 (Pruning)*: Covert the (minimal) cover array  $C$  to a pruned (minimal) cover array  $C^P$  by applying procedure PRUNE [11].
- *Step 3 (String Inference)*:
  - i) Construct a cover graph  $G(V, E)$  from  $C^P$ . This graph  $G$  has the same connected components as the graph directly constructed from  $C$  [11].
  - ii) Compute connected components of  $G$ . Decide which character to assign to the first position of each component as follows : Let,  $i$  be the first position of any component. If  $x[B[i-1]+1] = a$ , then assign  $b$  to  $x[i]$ . Otherwise, assign  $a$  to  $x[i]$ .

For each position  $j$  in string  $x$ , the algorithm also computes  $B[j]$  online, according to the well-known “Failure Function Algorithm” described in [8].

The procedure MAXTOMIN described in [11], works as follows:

Given, a cover array  $C[1..n]$ , it checks each value  $C[i]$ ,  $1 \leq i \leq n$  as follows.

- if  $C[i] = 0$ , then leaves it unchanged.
- if  $C[i] \neq 0$ , then substitutes  $C[i]$  with  $C[C[i]]$ , provided  $C[C[i]]$  is *nonzero*. Otherwise,  $C[i]$  is kept unchanged.

The procedure PRUNE described in [11], works as follows:

Given, a cover array  $C[1..n]$ , it finds each totally covered position  $i$  and substitutes  $C[i]$  by 0.

- The procedure scans  $C[1 \dots n]$  from large to small indices.
- Keeps a variable  $l$ , initially made 0. If at any instant  $l$  is larger than  $C[i]$ , then  $i$  is a totally covered position. So,  $C[i]$  is made 0.
- At each iteration, next value of  $l$  is computed.

For ease of understanding the procedures MAXTOMIN and PRUNE are given in Figure 5 and Figure 6. The algorithm SIMA is given in Figure 4. And its execution steps for a given cover array is illustrated in Figure 7.

Now, we state and prove the main findings.

**Theorem 1** *Let  $C^P$  be a pruned cover array of input cover array  $C$ , which resulted from Step 2 of the Algorithm SIMA. Let  $x$  be the word which is a result of the Algorithm SIMA. Let  $C_x$  is the (minimal) cover array for  $x$ . Then  $C = C_x$ .*

*Proof.* We just need to show that each assignment of a character to position  $i$  of the string  $x$  does not violate any constraints set by the values of  $C^P[i]$ .

Here we first construct the cover graph  $G$  from  $C^P$ . Then the nonzero values in  $C^P$  state that, the letters at positions  $i$  and  $j$  of  $x$  need to be equal, if  $i$  and  $j$  are connected in  $G$ . Since pruning does not reduce vertex connectivity [11], the cover graph induced by  $C^P$  has the same connected components as the one induced by  $C$ . The number of edges in the graph induced by  $C^P$  is bounded by  $2n$  according to Property 7 [11].

After constructing the graph, we compute the connected components of the constructed graph and at the same time assigns characters to the output string  $x$ . It also computes the value of longest border  $B[i]$  for string  $x[1 \dots i]$  for each  $i$  as the iterations advances. Computation of connected component is done to assign same character to those positions in the string which correspond to member vertices of a connected component.

We take decision only to assign a character to the first member (from left) of a component, and assign the same character to the remaining members of that component. That is, we can consider the following two cases:

1. When  $C^P[i] = k, 0 < k < i$ . This means,  $i$  has an edge with  $k$ , hence both  $i$  and  $k$  belong to the same component. So, whenever a character is assigned to  $x[k]$  it is also assigned to  $x[i]$ . Thus we do not need to take a decision about which character to assign to  $x[i]$  when  $C^P[i]$  is nonzero.
2. When  $C^P[i] = 0$ . If position  $i$  corresponds to the first member of a component, we check the value of  $B[i - 1]$ . Let,  $B[i - 1] = k$ . We can satisfy



```

SIMA( $C, n$ )
1   $C \leftarrow \text{MAXTO MIN}(C, n)$ ;
2   $C \leftarrow \text{PRUNE}(C, n)$ ;
3   $\triangleright$  Produce Edges
4  for  $i \leftarrow 1$  to  $n$ 
5  do
6       $E[i] \leftarrow \text{empty list}$ ;
7  for  $i \leftarrow 1$  to  $n$ 
8  do for  $j \leftarrow 1$  to  $C[i]$ 
9      do  $E[i - C[i] + 1 + j].\text{add}(j)$ ;
10      $E[j].\text{add}(i - C[i] + 1 + j)$ ;
11  $\triangleright$  Compute connected components by DFS and assign characters to output string
12  $S \leftarrow \text{empty stack}$ ;
13  $ch \leftarrow 'a'$ ;
14 for  $i \leftarrow 1$  to  $n$ 
15 do
16     if  $x[i] = \text{NIL}$ 
17     then  $S.\text{PUSH}(i)$ ;
18         if  $i > 1$  and  $C[i] = 0$ 
19         then if  $x[B[i - 1] + 1] = 'a'$ 
20         then  $ch \leftarrow 'b'$ ;
21         else  $ch \leftarrow 'a'$ ;
22         while not  $S.\text{EMPTY}()$ 
23         do  $p \leftarrow S.\text{POP}()$ ;
24              $x[p] \leftarrow ch$ ;
25             for each element  $j$  of  $E[p]$ 
26             do
27                 if  $x[j] = \text{NIL}$ 
28                 then  $S.\text{PUSH}(j)$ ;
29     if  $i > 1$ 
30     then  $l \leftarrow B[i - 1] + 1$ ;
31         while  $l \neq 0$ 
32         do
33             if  $x[i] = x[l]$ 
34             then  $B[i] \leftarrow l$ ;
35             Break;
36             else  $l \leftarrow B[l - 1] + 1$ ;
37     if  $l = 0$ 
38     then if  $x[i] = x[1]$ 
39     then  $B[i] \leftarrow 1$ ;
40     else  $B[i] \leftarrow 0$ ;
41
42 return  $x$ ;

```

**Fig. 4.** Algorithm SIMA.

```

MAXTOMIN( $C, n$ )
1  for  $i \leftarrow 1$  to  $n$ 
2  do
3      if  $C[i] \neq 0$  and  $C[C[i]] \neq 0$ 
4          then  $C[i] \leftarrow C[C[i]]$ 

```

**Fig. 5.** Procedure MAXTOMIN.

```

PRUNE( $C, n$ )
1   $l \leftarrow 0$ 
2  for  $i \leftarrow n$  to 0
3  do
4      if  $l \geq C[i]$ 
5          then  $C[i] \leftarrow 0$ 
6       $l \leftarrow \text{MAX}(0, \text{MAX}(l, C[i]) - 1)$ 
7  return  $C$ 

```

**Fig. 6.** Procedure PRUNE.

$C^P[i] = 0$ , if we can ensure  $B[i] \leq B[i - 1]$ , as stated in Property 8 [6]. Thus we assign  $x[i]$  a character different from  $x[k + 1]$  so that no border of length greater than  $k$  is possible for  $x[1 \dots i]$ . This obviously keeps  $C^P[i] = 0$ . Again, if  $i$  does not correspond to the first member of a component, then it is already assigned a valid character according to the component condition (i.e., all other members of a component receive the same character as the first one).

Thus the resultant string  $x[1 \dots n]$  satisfies the pruned cover array  $C^P$  at every position.

**Theorem 2** *Any string constructed by the algorithm SIMA uses an alphabet comprising no more than two characters.*

*Proof.* We prove this claim by induction on the length of cover array.

Without loss of generality, let,  $C[1 \dots n]$  be a valid (minimal) cover array of string  $x$  of length  $n$ . Let, the two characters to be assigned to infer the output string  $x$  be in  $\{a, b\}$ .

**Basis:** When  $n = 1$ , for a valid cover array  $C[1] = 0$ . In this case,  $x$  constitutes of a single character ‘ $a$ ’ and  $B[1] = 0$ .

When  $n = 2$ , two values of  $C[2]$  are possible for a valid cover array  $C$ . One is  $C[2] = 1$ . In this case,  $x[2]$  must be ‘ $a$ ’ to obtain  $x = aa$ . Otherwise,  $C[2] = 0$ .

$i$	1	2	3	4	5	6	7	8	9	10	11	12	13
$C[i]$	0	1	0	0	0	0	0	0	0	0	0	6	0
$C^P[i]$	0	1	0	0	0	0	0	0	0	0	0	6	0

(a)

Components:  $\{1, 2, 7, 8\}$   $\{3, 9\}$   $\{4, 10\}$   $\{5, 11\}$   $\{6, 12\}$   $\{13\}$

(b)

$i$	1	2	3	4	5	6	7	8	9	10	11	12	13	Comment
$x[i]$	a	a					a	a						$x[1] \leftarrow a$ and $x[2], x[7], x[8]$
$B[i]$	0													are assigned $x[1]$
$x[i]$	a	a					a	a						
$B[i]$	0	1												
$x[i]$	a	a	b				a	a	b					$x[B[2] + 1]$ is a. So $x[3] \leftarrow b$ .
$B[i]$	0	1	0											$x[9] \leftarrow x[3]$
$x[i]$	a	a	b	b			a	a	b	b				$x[B[3] + 1]$ is a. So $x[4] \leftarrow b$ .
$B[i]$	0	1	0	0										$x[10] \leftarrow x[4]$
$x[i]$	a	a	b	b	b		a	a	b	b	b			$x[B[4] + 1]$ is a. So $x[5] \leftarrow b$ .
$B[i]$	0	1	0	0	0									$x[11] \leftarrow x[5]$
$x[i]$	a	a	b	b	b	b	a	a	b	b	b	b		$x[B[5] + 1]$ is a. So $x[6] \leftarrow b$ .
$B[i]$	0	1	0	0	0	0								$x[12] \leftarrow x[6]$
$x[i]$	a	a	b	b	b	b	a	a	b	b	b	b		
$B[i]$	0	1	0	0	0	0	1							
$x[i]$	a	a	b	b	b	b	a	a	b	b	b	b		
$B[i]$	0	1	0	0	0	0	1	2	3					
$x[i]$	a	a	b	b	b	b	a	a	b	b	b	b		
$B[i]$	0	1	0	0	0	0	1	2	3	4				
$x[i]$	a	a	b	b	b	b	a	a	b	b	b	b		
$B[i]$	0	1	0	0	0	0	1	2	3	4	5			
$x[i]$	a	a	b	b	b	b	a	a	b	b	b	b		
$B[i]$	0	1	0	0	0	0	1	2	3	4	5	6		
$x[i]$	a	a	b	b	b	b	a	a	b	b	b	b	b	$x[B[12] + 1]$ is a. So $x[13] \leftarrow b$ .
$B[i]$	0	1	0	0	0	0	1	2	3	4	5	6	0	

(c)

**Fig. 7.** An example run of Algorithm SIMA. (a) Input cover array  $C$  before and after pruning, (b) Connected components of corresponding cover graph, and (c) String Inference by Algorithm SIMA.

In this case,  $x[2]$  must be 'b' to obtain  $x = ab$ . In both case, value of  $B[2]$  is computed.

**Induction:** Let  $n > 2$ . We assume that up to length  $n - 1$ ,  $B[1..n - 1]$  and  $x[1..n - 1]$  have been computed and  $x[1..n - 1]$  needs an alphabet consisting of two characters. We consider the assignment of character to  $x[n]$ .

*Case 1:*  $C[n] = 0$

According to Property 8, for every integer  $1 \leq i \leq n - 1$ , if  $B[i + 1] \leq B[i]$  then  $C[i + 1] = 0$ .

Let,  $B[n - 1] = k$ . Now, if  $x[k + 1] = 'a'$  then we assign 'b' to  $x[n]$  so that  $B[i + 1]$  cannot become greater than  $k$ . Or, if  $x[k + 1] = 'b'$  then we assign 'a' to  $x[n]$  for the same reason. This maintains the constraint  $C[n] = 0$ . So  $x[1..n]$  uses a two-character alphabet.

*Case 2:*  $C[n] = k$ ,  $1 \leq k < n$

Position  $n$  has an edge with position  $k$ . Our algorithm assigns into  $x[n]$  the same character that it assigns into  $x[k]$ . Since  $k < n$ , so  $x[k]$  is either 'a' or 'b'. Thus, we do not need to introduce any new character for  $x[n]$  here.

Thus algorithm SIMA produces a string  $x[1..n]$  which uses an alphabet of no more than two characters.

**Theorem 3** *Algorithm SIMA runs in linear time.*

*Proof.* The each of the two procedures MAXTOMIN [11] and PRUNE [11] runs in linear time [11]. The step of producing edges  $E$  of graph  $G$  induced by  $C^P$  is also linear because the number of edges is bounded by  $2n$  according to Property 5 [11].

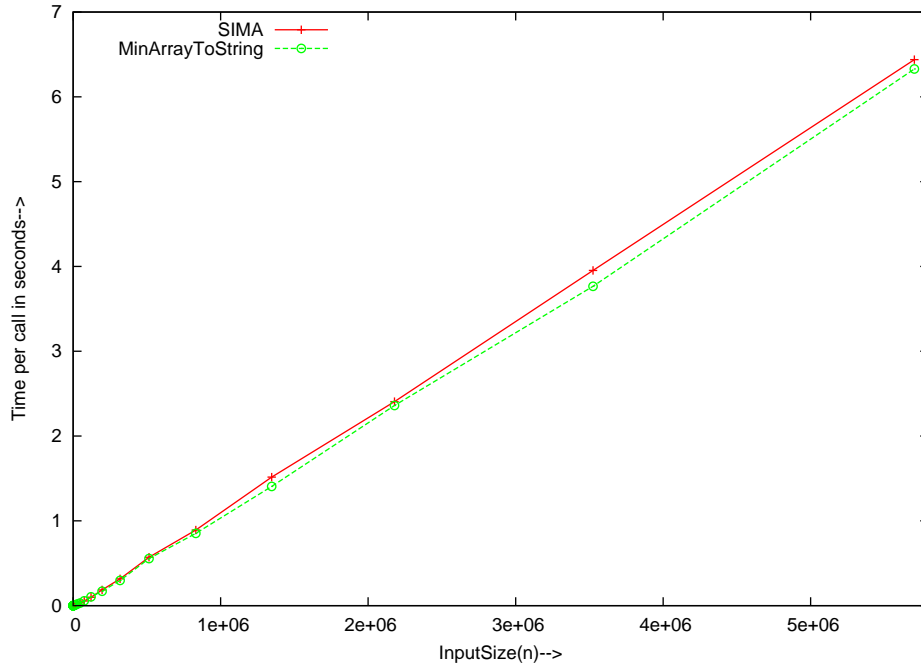
The third for loop computes the connected components in the graph by depth first search and assigns letters to the output string. This computation is linear in the number of edges which is bounded by  $2n$ . Also the overall on-line computation of the border array  $B$  runs in linear time [2]. Hence our algorithm runs in linear time.

## 5 Experimental Results

We have investigated the practical performance of Algorithm SIMA on various datasets. The experiments were performed on a computer with 4 GB of main memory and 3.1 GHz Intel Pentium 4 processor, running the Windows XP Service Pack 3 operating system. All programs were compiled with Visual Studio 6.0.

The investigated data includes, all valid cover arrays for length 8 to 14 and cover arrays generated from *Fibonacci words* of different sizes. The experimental results are summarized below.

- We have been able to verify the linear runtime of our algorithm experimentally. Figure 8 shows the timing diagram of our algorithm for *fibonacci word dataset*. For hardware limitations we restricted our test from *fibonacci word* size 4 to 34.



**Fig. 8.** Verification of Linear runtime of Algorithm SIMA.

- We have also compared our algorithm with the implementation of MINARRAYTOSTRING available at [14]. In every case, our algorithm was able to infer valid strings with no more than two letters which is a sure improvement over MINARRAYTOSTRING. The comparative results for all valid cover arrays of length 8 is shown in Table 1. Table 2 shows the comparison of the two algorithms for several genome sequences available at [15].

Input Cover Array	String Inferred By SIMA	String Inferred By MINARRAYTOSTRING
0 0 0 0 0 0 0 0	<i>a b b b b b b b</i>	<i>a b c d e f g h</i>
0 0 0 0 0 0 0 4	<i>a b b b a b b b</i>	<i>a b c d a b c d</i>
0 0 0 0 0 3 0 0	<i>a b b a b b b b</i>	<i>a b c a b c d e</i>
0 0 0 0 0 3 0 3	<i>a b a a b a b a</i>	<i>a b a a b a b a</i>
0 0 0 0 0 3 4 0	<i>a b b a b b a a</i>	<i>a b c a b c a d</i>
0 0 0 0 0 3 4 5	<i>a b b a b b a b</i>	<i>a b c a b c a b</i>
0 0 0 2 0 0 0 0	<i>a b a b b b b b</i>	<i>a b a b c d e f</i>
0 0 0 2 3 0 0 0	<i>a b a b a a a a</i>	<i>a b a b a c d e</i>
0 0 0 2 3 0 0 3	<i>a b a b a a b a</i>	<i>a b a b a a b a</i>
0 0 0 2 3 2 0 0	<i>a b a b a b b b</i>	<i>a b a b a b c d</i>
0 0 0 2 3 2 3 0	<i>a b a b a b a a</i>	<i>a b a b a b a c</i>
0 0 0 2 3 2 3 2	<i>a b a b a b a b</i>	<i>a b a b a b a b</i>
0 1 0 0 0 0 0 0	<i>a a b b b b b b</i>	<i>a a b c d e f g</i>
0 1 0 0 0 0 0 4	<i>a a b b a a b b</i>	<i>a a b c a a b c</i>
0 1 0 0 0 3 0 0	<i>a a b a a b b b</i>	<i>a a b a a b c d</i>
0 1 0 0 0 3 4 0	<i>a a b a a b a b</i>	<i>a a b a a b a c</i>
0 1 0 0 0 3 4 5	<i>a a b a a b a a</i>	<i>a a b a a b a a</i>
0 1 1 0 0 0 0 0	<i>a a a b b b b b</i>	<i>a a a b c d e f</i>
0 1 1 0 0 0 0 4	<i>a a a b a a a b</i>	<i>a a a b a a a b</i>
0 1 1 1 0 0 0 0	<i>a a a a b b b b</i>	<i>a a a a b c d e</i>
0 1 1 1 1 0 0 0	<i>a a a a a b b b</i>	<i>a a a a a b c d</i>
0 1 1 1 1 1 0 0	<i>a a a a a a b b</i>	<i>a a a a a b c</i>
0 1 1 1 1 1 1 0	<i>a a a a a a a b</i>	<i>a a a a a a b</i>
0 1 1 1 1 1 1 1	<i>a a a a a a a a</i>	<i>a a a a a a a a</i>

**Table 1.** Comparison on *Inferred String* between algorithms SIMA and MINARRAYTOSTRING.

Genome Sequence	SIMA	MINARRAYTOSTRING
<i>Acidovorax citrulli</i> AAC00-1	2	5352783
<i>Buchnera aphidicola</i> 5A	2	642133
<i>Ca. Blochmannia floridanus</i>	2	705649
<i>Dickeya dadantii</i> 3937	2	4922813
<i>Edwardsiella ictarluri</i> 93-146	2	3812326
<i>Klebsiella pneumonia</i> 342	2	5920281

**Table 2.** Comparison on *Alphabet Size* between algorithms SIMA and MINARRAYTOSTRING.

- Finally we have observed an interesting fact that the set of distinct valid cover arrays is generated from  $m$ -alphabet string for a certain length, where  $m \geq 2$ . We generated all possible strings for length of 8 with alphabet sizes 2, 3, 4, 5, 6, 7 and 8, and computed cover arrays for all of them. For each alphabet size we got same set of distinct cover arrays.

## 6 Conclusion

In this paper, we have presented a linear time algorithm to solve the problem of inference of strings using the least sized alphabet (i.e., binary alphabet) from valid cover arrays. We achieved the least possible bound on alphabet size by incorporating an interesting relation between cover array and border array of a string. The main finding of this paper is that, from any valid cover array of length  $n$ , it is possible to infer a string over an alphabet that consists only two distinct characters unless the cover array is of the form  $01^{k-1}$ ,  $1 \leq k \leq n$ . In that particular case, our algorithm infers a string over an alphabet consisting only of a single character.

## References

1. A. Apostolico and A. Ehrenfeucht, “Efficient detection of quasiperiodicities in strings,” *Theor. Comput. Sci.*, vol. 119, no. 2, pp. 247–265, 1993.
2. D. E. Knuth, J. H. M. Jr., and V. R. Pratt, “Fast pattern matching in strings,” *SIAM J. Comput.*, vol. 6, no. 2, pp. 323–350, 1977.
3. R. S. Boyer and J. S. Moore, “A fast string searching algorithm,” *Commun. ACM*, vol. 20, no. 10, pp. 762–772, 1977.
4. H. Bannai, S. Inenaga, A. Shinohara, and M. Takeda, “Inferring strings from graphs and arrays,” in *MFCS*, pp. 208–217, 2003.
5. A. Apostolico and D. Breslauer, “Of periods, quasiperiods, repetitions and covers,” in *Structures in Logic and Computer Science*, pp. 236–248, 1997.
6. Y. Li and W. F. Smyth, “Computing the cover array in linear time,” *Algorithmica*, vol. 32, no. 1, pp. 95–106, 2002.
7. F. Franěk, W. Lu, P. J. Ryan, W. F. Smyth, Y. Sun, and L. Yang, “Verifying a border array in linear time,” *Journal on Combinatorial Mathematics and Combinatorial Computing*, vol. 42, pp. 223–236, 2002.
8. A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*. Addison-Wesley, 1974.

9. J.-P. Duval, T. Lecroq, and A. Lefebvre, "Border array on bounded alphabet," *Journal of Automata, Languages and Combinatorics*, vol. 10, no. 1, pp. 51–60, 2005.
10. W. F. Smyth and S. Wang, "New perspectives on the prefix array," in *SPIRE*, pp. 133–143, 2008.
11. M. Crochemore, C. S. Iliopoulos, S. P. Pissis, and G. Tischler, "Cover array string reconstruction," in *CPM*, pp. 251–259, 2010.
12. T. I, S. Inenaga, H. Bannai, and M. Takeda, "Counting parameterized border arrays for a binary alphabet," in *LATA*, pp. 422–433, 2009.
13. T. I, S. Inenaga, H. Bannai, and M. Takeda, "Verifying a parameterized border array in  $O(n^{1.5})$  time," in *CPM*, pp. 238–250, 2010.
14. <http://www.kcl.ac.uk/staff/tischler/src/recovering-0.0.0.tar.bz2> (Last accessed on December 12, 2010).
15. [https://asap.ahabs.wisc.edu/asap/download\\_Source.php?LocationID=&SequenceVersionID=&GenomeID=](https://asap.ahabs.wisc.edu/asap/download_Source.php?LocationID=&SequenceVersionID=&GenomeID=) (Last accessed on December 18, 2010).