Notes on Sequence Binary Decision Diagrams: Relationship to Acyclic Automata and Complexities of Binary Set Operations

Shuhei Denzumi¹, Ryo Yoshinaka², Hiroki Arimura¹, and Shin-ichi Minato¹,²

¹Graduate School of IST, Hokkaido University, Japan
²ERATO MINATO Discrete Structure Manipulation System Project, JST, Japan
{denzumi,ry,arim,minato}@ist.hokudai.ac.jp

Abstract. Manipulation of large sequence data is one of the most important problems in string processing. Recently, Loekito et al. (Knowl. Inf. Syst., 24(2), 235-268, 2009) have introduced a new data structure, called Sequence Binary Decision Diagrams (SeqBDDs, or SDDs), which are descendants of both acyclic DFAs (ADFAs) and binary decision diagrams (BDDs). SDDs can compactly represent sets of sequences as well as minimal ADFAs, while SDDs allow efficient set operations inherited from BDDs. A novel feature of the SDDs is that different SDDs can share equivalent subgraphs and duplicated computation in common to save the time and space in various operations. In this paper, we study fundamental properties of SDDs. In particular, we first present non-trivial relationships between sizes of minimum SDDs and minimal ADFAs. We then analyze the complexities of algorithms for Boolean set operations, called the binary synthesis. Finally, we show experimental results to confirm the results of the theoretical analysis on real data sets.

1 Introduction

1.1 Backgrounds

Compact string indexes for storing sets of strings are fundamental data structures in computer science, and have been extensively studied in the decades [2, 4–6, 9, 19]. Examples of compact string indexes include: tries [1, 5], finite automata and transducers [6, 10], suffix trees [15], suffix arrays [14], DAWGs [2], and factor automata (FAs) [19]. By the rapid increase of massive amounts of sequence data such as biological sequences, natural language texts, and event sequences, these compact string indexes have attracted much attention and gained more importance [5, 9]. In such applications, an index have not only to compactly store sets of strings for searching, but also have to efficiently manipulate them with various set operations, e.g., merge, intersection, and subtraction.

Minimal acyclic deterministic finite automata (ADFAs) [5, 6, 10] are one of such index structures that fulfill the above requirement based on finite automata theory, and have been used in many sequence processing applications [13, 18]. However, they have drawback of complicated procedures for minimization and various set operations caused by multiple branching of the underlying directed acyclic graph structure. To overcome this problem, Loekito et al. [12] proposed the class of sequence binary decision diagrams (sequence BDDs, or abbreviated as SDDs in this paper), which is a compact representation for sets of strings that allows a variety of operations for sets of strings. An SDD is a node-labeled graph structure, which resembles to an acyclic DFA in binary form, but with the minimization rule which is different from one for
a minimal DFA. A novel feature of the SDDs is their ability to share equivalent subgraphs and results of similar intermediate computation between different SDDs, which avoids redundant generation of nodes and computation.

1.2 Main results

In this paper, we present theoretical analysis of two fundamental problems on sequence binary decision diagrams, which have not been studied before, the relationship to acyclic automata and the complexities of binary set operations as follows.

The relationship to acyclic automata. The structure of SDDs apparently resembles that of acyclic Deterministic Finite Automata (DFAs), which are a classical model for representing string sets. While a state of a DFA may have many outgoing edges, a node of an SDD always has two outgoing edges, which can be seen as just the “first-child next-sibling” representation of a branching with many edges. Indeed one can find a straightforward translation from an acyclic DFA to an SDD and vice versa. However, there are subtle differences between those formalisms and actually SDD can be even more compact. We first study the relationship between the sizes of an SDD and an equivalent acyclic DFA. As results, we show that the minimal SDD is as small as the minimum ADFA for the same language. Conversely, we show that the minimal ADFA is at most $|\Sigma|$ time larger than an equivalent minimum SDD on a given alphabet $\Sigma$. As corollary, for every language, the sizes of SDDs differ at most factor $|\Sigma|$ for any letter ordering on $\Sigma$.

The computational complexities of binary set operations. Next, we study the complexity of the binary synthesis, which are binary operations for minimal SDDs, such as union, intersection, and subtraction, followed by minimization of the output. Specifically, we study upper and lowerbounds for the time complexity of of the binary synthesis algorithms, Union, for union and subtraction proposed by Loeckito et al. [12], where Union is similar to the apply algorithm by Bryant [3], and conjectured to run in input linear-time [12]. Then, we first present a generalization, called Meld$\diamond$, of Union in the style of Knuth’s melding operation for BDDs [11], where Meld$\diamond$ implements eight set operations specified with a binary Boolean function $F_\diamond$ as a parameter. We show upperbounds that its time complexity is quadratic in the input size, and linear in the size of non-reduced version of the output size. Moreover, we show a lowerbound that Meld$\diamond$ actually requires quadratic time in input size for some infinite series of inputs using a technique recently devised for BDD [3], giving matching upperbound.

Experimental results. Finally, we run experiments on real data sets. We first observed that minimal SDDs were Superior to minimal DFAs when large subgraphs were shared in inputs and outputs due to the node-sharing across multiple SDDs. We also observed that each binary synthesis operation took less than seconds to take set operations $\cup$, $\cap$, and $\setminus$ of two input SDDs with around three to four thousands of nodes each, which were relatively smaller than the running time for set construction [7].

1.3 Related works

There have been a number of researches on manipulation of finite automata in automata theory and string algorithms. The textbook [10] gives classic examples of a quadratic-time algorithm for computing the union, intersection, and subtraction of
two DFAs, and a state-minimization algorithm for a given DFA. Daciuk, Mihov, Watson, and Watson [6] presented an incremental algorithm for constructing the minimal ADFA for a set of strings. Blumer et al. [2] and Crochemore [4] gave linear-time algorithms for construction of the minimal state ADFAs for the set of all factors of an input string. Compared with a straightforward two-stage algorithm for binary synthesis for ADFAs using product followed by state-minimization [10], the advantages of the proposed Meld, are its simplicity and efficiency that it directly computes the output by applying the on-the-fly minimization [7]. Using binary synthesis, Denzumi et al. [7] presented a simple linear-time algorithm for incremental construction of SDDs, and a recursive top-down algorithm for construction of factor SDDs.

The class of SDDs inherited many of its features from binary decision diagrams (BDDs), which are compact representation for storing and manipulating combinatorial structures developed in logic design community [3, 11, 16, 21]. Especially, BDDs equipped with binary synthesis operation were invented by Bryant [3] in 80s for dealing with Boolean functions, while their variant with node-sharing and zero-suppress rules, called zero-suppressed BDDs (ZDDs), were proposed by Minato [16] in 90s for sparse combinatorial sets. On their early history, reduced BDDs were constructed from tree-like circuits through offline minimization. After the invention of the binary synthesis algorithm by Bryant [3], it became popular to build large BDDs on-the-fly in real applications. Loekito et al. [12] discovered that if we remove the ordering constraint on the 1-edges from ZDDs, then the resulting variant of ZDDs, which actually are SDDs, have similar structure to ADFA in binary form and suitable to storing and manipulating sets of strings. This observation led to the invention of SDDs [12].

**Organization of this paper.** In Section 2, we prepare basic notions and notations on sequence BDDs. In Section 3, we gives the size bounds for SDDs and DFAs. In Section 4, we gives the time and space complexities of binary synthesis procedures for SDDs. In Section 5, we show some experimental results. In Section 6, we conclude this paper. For details of basic properties and algorithms related to SDDs not described in this paper, please consult the companion paper [7].

## 2 Preliminaries

In this section, we give basic definitions and notations in strings and sequence BDDs according to [11, 12, 16]. For the details of results not found here, please consult the companion paper [7]. An ordered alphabet is a pair \( \langle \Sigma, \prec \rangle \) where \( \Sigma \) is a finite alphabet and \( \prec \) is a total order on \( \Sigma \). The order \( \prec \) associated with \( \Sigma \) is often denoted by \( \prec_{\Sigma} \) and the ordered alphabet is simply written \( \Sigma \) for legibility. A string on \( \Sigma \) is a sequence \( s = s_1 \cdots s_n \) of letters \( s_i \in \Sigma \) \((1 \leq i \leq n)\), where \( |s| = n \) denotes the length. If \( s = xyz \) for some \( x, y, z \in \Sigma^* \), then we say that \( x \) is a prefix, \( y \) is a factor, and \( z \) is a suffix of \( s \). A string set (or a language) is any finite \( S \subseteq \Sigma^* \). We denote by \( |S| \) the cardinality.

For any \( x \in \Sigma \), we define \( x \cdot S = \{ xy \mid y \in S \} \).

**Sequence BDDs.** Let \( \text{dom} \) be a countable domain of the nodes. A sequence binary decision diagram or a sequence BDD (abbreviated as SDD\(^{1}\) here) is a directed

\(^1\) Note that the abbreviation SeqBDD was used to denote sequence BDD in the original paper by Loekito et al. [12]. We also note that the abbreviation SDD was also used for the set decision diagrams (Couvreur, Thierry-Mieg, Proc. FORTE 2005, LNCS 3731, 443–457, 2005) and the spectral decision diagrams (Thornton, Drechsler, Proc. DATE’01, IEEE, 713–719, 2001).
acyclic graph (DAG) \( B = \langle \Sigma, V, \tau, r, 0, 1 \rangle \) where \( V = V(B) \subseteq \text{dom} \) is a finite set of nodes, \( r \in V \) is called the root of \( B \) and \( 0 \) and \( 1 \in V \) are distinct nodes called the 0- and 1-terminals, resp. The nodes in \( V_N = V \setminus \{0, 1\} \) are called nonterminals. Each node \( v \in V_N \) of \( B \) is labeled by a symbol \( v.\text{lab} \) in \( \Sigma \) and has possibly identical two children, the 0-child and the 1-child, denoted by \( v.0 \) and \( v.1 \), resp. We call the edge from \( v \) to \( v.0 \) and \( v.1 \) the 0- and 1-edge of \( v \), resp. The information is formally described by a function \( \tau : V_N \rightarrow \Sigma \times V^2 \) that assigns the triple \( \tau(v) = (v.\text{lab}, v.0, v.1) \) to each \( v \in V_N \). An SDD must be acyclic, that is, one may assume a strict partial order \( \succ \) on \( V \) such that \( v \succ v.0 \) and \( v \succ v.1 \) hold for any \( v \in V_N \). The 1-child and 0-child of a node correspond to the leftmost-child and the right-sibling in a DAG in binary form \([1, 11]\). Siblings are deterministically ordered from left to right according to the order \( \prec \). That is, we always have \( v.\text{lab} \prec (v.0).\text{lab} \) unless \( v.0 \) is a terminal node. We assume that any SDD \( B \) is well-defined meaning that \( B \) is both acyclic and deterministic. We define the size of \( B \) by \( |B| = |V_N| + |V| - 2 \), the number of non-terminals in \( B \).

To each node \( v \in V \), we inductively (w.r.t. \( \succ \)) assign a language \( L_B(v) \) as follows:

(i) \( L_B(0) = \emptyset \);
(ii) \( L_B(1) = \{ \varepsilon \} \);
(iii) \( L_B(v) = L_B(v.0) \cup (v.\text{lab}) \cdot L_B(v.1) \).

Equivalently, \( s \in L_B(v) \) iff there is a path from \( v \) to \( 1 \) such that one obtains \( s \) by concatenating the labels of the nodes whose 1-edges appear in the path, where, hence, there are just \( n \) 1-edges on the path. We say that two SDDs \( B \) and \( B' \) are equivalent if \( L(B) = L(B') \). An SDD \( B \) is said to be minimal if it has the smallest number of nodes among the equivalent SDDs, i.e., \( |B| \leq |B'| \) for any SDD \( B' \) such that \( L(B') = L(B) \).

Figure 1 illustrates examples of SDDs together with the minimum deterministic finite automaton for the same language.

**Reduced SDDs.** A reduced SDD is a normal form of SDDs. An SDD is said to be reduced if it satisfies the following two conditions:

1. For any \( u, v \in V_N, \tau(u) = \tau(v) \) implies \( u = v \) (node-sharing rule).
2. For any \( v \in V_N, v.1 \neq 0 \) holds (zero-suppress rule).

The above rules say that no distinct non-terminal nodes have the same triple, and the 1-child of any non-terminal node \( v \) is not the 0-terminal. For any finite set of

---

**Figure 1.** Examples of three index structures on \( \Sigma = \{a, b, c\} \) for the same string set \( S_1 = \{aa, aab, acc, ab, acc, bb, bbb, bc, bcc, c, cc\} \): a minimal DFA \( G_1 \) (left), a minimal DFA as a non-reduced SDD \( G_2 \) (middle), and a reduced SDD \( G_3 \) (right). In the figure, solid and dotted arrows indicate the 1- and 0-edge. The edges to the 0-terminal are omitted.
Global variable: uniqtable: hash table for triples.

Proc Getnode(x: letter, P₀, P₁: SDD):
1: if (P₁ = 0) return P₀; /* zero-suppress rule */
2: else if ((R ← uniqtable[(x, P₀, P₁)]) exists) return R; /* node-sharing rule */
3: else
4: R ← a new node with τ(R) = (x, P₀, P₁); uniqtable[(x, P₀, P₁)] ← R;
5: return R;

Figure 2. The Getnode procedure for on-the-fly minimization.

strings \( L \subseteq \Sigma^* \), we can construct the canonical SDD for \( L \) [7] in a way similar to the minimal DFA in Myhill-Nerode theorem (e.g.,[10, 20]). Actually, the next theorem gives a characterization of minimal SDDs in terms of a reduced SDD and the canonical SDD. See the companion paper [7] for the details.

**Theorem 1 (Denzumi et al. [7]).** For any SDD \( B \) with the language \( L = L(B) \), the following (1)–(3) are equivalent to each other.

(1) \( B \) is a reduced SDD.
(2) \( B \) is a canonical SDD for \( L \) up to isomorphism.
(3) \( B \) is a minimal SDD.

Due to Theorem 1, for a fixed language \( L \), we may call a reduced SDD for \( L \) the reduced SDD for \( L \) when we work modulo isomorphism. In order to satisfy the node-sharing rule, we maintain a hash table, called uniqtable, which is the inverse of \( \tau \). That is, it gives the unique node \( v \) such that \( \tau(v) = (x, v₀, v₁) \) (if exists) for the key \( (x, v₀, v₁) \). As is the case for BDDs and ZDDs, usually we consider only reduced SDDs. Hereafter we assume that all SDDs are reduced unless otherwise noted.

While an SDD represents a set of strings, we often would like to manipulate two or more sets of strings. In our shared SDD environment, the terminals 0 and 1, the function \( \tau \) and thus the hash \( \text{uniqtable} : \Sigma \times \text{dom} \times \text{dom} \rightarrow \text{dom} \) are shared by more than one SDD in common so that we can have a compact representation of a family of sets of strings. One may think of the shared SDD environment as a single SDD with multiple roots. By picking up a node \( v \) as the root, one can extract a subgraph as an SDD consisting of all the nodes that are reachable from \( v \). For convenience, we often identify a node \( v \) with the SDD rooted by \( v \) extracted from the shared SDD environment. Hence \( |v| \) represents the number of nonterminal nodes reachable from the node \( v \).

Figure 2 shows the node allocation procedure Getnode, which is used as a subroutine in algorithms on SDDs. Throughout this paper, we assume that the hash table uniqtable is a global variable, and a look-up for it takes \( O(1) \) time; We have to add additional \( O(\log n) \) term if we use balanced binary tree dictionary [1]. In the shared and reduced SDD environment studied here and in [7, 12], we use write-only construction, similarly to [8], such that any new SDD is constructed by adding a new node on the top of already constructed SDDs using a call of Getnode given existing nodes as its arguments. The next lemma guarantees that we always have reduced SDDs as long as we solely use Getnode to obtain a new node.

**Lemma 2 (Denzumi et al. [7]).** Let \( B \) be any reduced SDD. For any symbol \( x \in \Sigma \) and nodes \( v₀, v₁ \in V(B) \) in \( B \) such that \( v₀ \not\in V_N \) or \( x \prec_\Sigma v₀.lab \), if we invoke
\( v = \text{Getnode}(x, v_0, v_1) \) on \( B \) and add the result \( v \) to \( V(B) \), then the resulting SDD \( B' \) with root \( v \) obtained from \( B \) is well-defined and reduced, too.

Based on the procedure \text{Getnode} above, for example, we can implement an off-line minimization (i.e., reduction) algorithm \text{Reduce} for SDDs in linear time and space [7]. Factually it simply makes a copy of an input SDD using \text{Getnode}, which merges equivalent nodes in \text{uniqtable}. See [7, 12] for details.

3 Space-Bounds for Sequence Binary Decision Diagrams and Acyclic Automata

The structure of SDDs apparently resembles that of acyclic deterministic finite automata (ADFAs). There is a straightforward translation from an ADFA to an SDD and the other way around. However, we should note subtle differences between those formalisms. Actually SDD can be even more compact. This section discusses their relationship in detail.

3.1 Finite Automata

We presume a basic knowledge of the automaton theory. For a comprehensive introduction to the automaton theory, see [10] for example. A (partial) DFA is represented by a tuple \( A = \langle \Sigma, \Gamma, \delta, q_0, F \rangle \), where \( \Sigma \) is the input alphabet, \( \Gamma \) is the state set, \( \delta \) is the partial transition function from \( \Gamma \times \Sigma \) to \( \Gamma \), \( q_0 \in \Gamma \) is the initial symbol and \( F \subseteq \Gamma \) is the set of acceptance states. The partial function \( \delta \) can be regarded as a subset \( \delta \subseteq \Gamma \times \Sigma \times \Gamma \). We define the size of a DFA \( A \), denoted by \(|A|\), as the number of labeled edges in \( \delta \) as a partial function, i.e., \(|A| = |\delta|\).

The set of strings that lead the automaton \( A \) from a state \( q \) to an accept state is denoted by \( L_A(q) \). The language \( L(A) \) accepted by \( A \) is \( L_A(q_0) \). A minimum DFA has no state \( q \) such that \( L_A(q) = \emptyset \) and no distinct states \( q' \) and \( q'' \) such that \( L_A(q') = L_A(q'') \). Since we are concerned with finite sets of strings, all DFAs discussed in this section are acyclic (ADFA). We say that \( A \) and \( B \), which can be an ADFA or an SDD, are equivalent if \( L(A) = L(B) \).

3.2 From ADFAs to SDDs

We first give a straightforward translation from an ADFA to an equivalent SDD, which may be non-reduced, and compare the sizes of them.

**Theorem 3.** For any ADFA \( A = \langle \Sigma, \Gamma, \delta, q_0, F \rangle \), there is an equivalent SDD \( B = \langle \Sigma, V, \tau, 0, 1, r \rangle \) such that \(|V_N| \leq |\delta|\). Moreover, for every positive integer \( n \geq 1 \), there is an ADFA \( A \) that admits no equivalent SDD \( B \) such that \(|V_N| < |\delta| = n\).

**Proof.** For an ADFA \( A = \langle \Sigma, \Gamma, \delta, q_0, F \rangle \), we construct an equivalent SDD \( B(A) \). Let

\[
\deg(q) = |\{ a \in \Sigma \mid \delta(q, a) \text{ is defined} \}|
\]

We define \( B(A) = \langle \Sigma, V, \tau, 0, 1, r \rangle \) as follows. The set of nodes is given by

\[
V = \{0, 1\} \cup \{[q, i] \mid q \in \Gamma \text{ and } 1 \leq i \leq \deg(q) \}.
\]

For each \( q \in \Gamma \) with \( \deg(q) = k \geq 1 \), let \( a_1, \ldots, a_k \in \Sigma \) and \( q_1, \ldots, q_k \in \Gamma \) be such that
\[ \delta(q, a_i) = q_i \text{ for } i = 1, \ldots, k, \]
\[ a_1 < a_2 < \cdots < a_k. \]

Define \( \tau \) by
\[ \tau([q, i]) = \begin{cases} 
\langle a_i, [q, i + 1], \hat{q}_i \rangle & \text{if } i < k, \\
\langle a_k, 1, \hat{q}_k \rangle & \text{if } i = k \text{ and } q \notin F, \\
\langle a_k, 0, \hat{q}_k \rangle & \text{if } i = k \text{ and } q \in F,
\end{cases} \]
where
\[ \hat{q}' = \begin{cases} 
[q', 1] & \text{if } \deg(q') > 0, \\
1 & \text{if } \deg(q') = 0 \text{ and } q' \in F, \\
0 & \text{if } \deg(q') = 0 \text{ and } q' \notin F.
\end{cases} \]

The root \( r \) of \( B(A) \) is \( \hat{q}_0 \).

It is easy to see that \( L_A(q) = L_{B(A)}(\hat{q}) \) for all \( q \in \Gamma \). We note that the above construction can be done in linear time in \(|\delta|\).

The first claim of the theorem can be verified by the above construction of \( B(A) \). The second claim is established by observing the minimum ADFA and the reduced SDD that accept the singleton \( \{a^n\} \) for each positive integer \( n \). For the detail of construction of the reduced (canonical) SDD from a string set, consult [7]. \( \square \)

We remark that \( B(A) \) in the proof is not necessarily reduced for a minimum ADFA \( A \).

**Example 4.** Let us compare the minimum ADFA \( A \) and the constructed SDD \( B(A) \) for the set \( \{ab, b\} \) with \( a < b \):

<table>
<thead>
<tr>
<th>Transition rules of ( A )</th>
<th>Corresponding nodes of ( B(A) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta(q_0, a) = q_1 )</td>
<td>( \tau([q_0, 1]) = \langle a, [q_0, 2], [q_1, 1] \rangle )</td>
</tr>
<tr>
<td>( \delta(q_0, b) = q_2 )</td>
<td>( \tau([q_0, 2]) = \langle b, 0, 1 \rangle )</td>
</tr>
<tr>
<td>( \delta(q_1, b) = q_2 )</td>
<td>( \tau([q_1, 1]) = \langle b, 0, 1 \rangle )</td>
</tr>
</tbody>
</table>

\( B(A) \) is not reduced since \( \tau([q_0, 2]) = \tau([q_1, 1]) \) for \( [q_0, 2] \neq [q_1, 1] \).

In this example, \( A \) has two distinct edges that are labeled with \( b \) and come into \( q_2 \), which should be merged into the same node in a reduced SDD. Hence the reduced SDD can be more compact than the minimum ADFA for the same language. We next discuss how much an SDD can be smaller than an ADFA through a translation from an SDD into an ADFA.

### 3.3 From SDDs to ADFA\( s\)

We next discuss how much an SDD can be smaller than an ADFA through a translation from an SDD into an ADFA. Let an SDD \( B = \langle \Sigma, V, \tau, 0, 1, r \rangle \) be given. We construct an ADFA \( A(B) = \langle \Sigma, \Gamma, \delta, q_0, F \rangle \) such that \( L(A(B)) = L(B) \). We assume that \( r \neq 0 \). Otherwise, the translation is trivial.

For each \( P \in V_N \), let \( \tilde{P} = [P_1, \ldots, P_k] \) be such that \( P_1 = P \) and \( \tau(P_i) = \langle a_i, P_{i+1}, R_i \rangle \) for some \( R_i \in V \) for \( i \leq k \) and \( P_{k+1} \in \{0, 1\} \). We define \( \tilde{1} \) to be \( [1] \). Let
\[ \Gamma = \{\tilde{r}\} \cup \{\tilde{P}_1 \mid P_1 \neq 0 \text{ is the 1-child of some } P \in V_N \}, \]
Proof. The first claim, inequality holds.

It is easy to see that \( L_B(P) = L_{A(B)}(\tilde{P}) \) for all \( \tilde{P} \in \Gamma \). This implies that if \( B \) is reduced, \( A(B) \) is minimum. Contrary to the translation from an ADFA into an equivalent SDD, this construction takes \( O(|\Sigma||V_N|) \) time. In fact, this is optimal. The following theorem implies that the reduced SDD can be about \(|\Sigma|\) times more compact than the minimum ADFA for the same set of strings.

**Theorem 5.** For any SDD \( B = \langle \Sigma, V, \tau, r, 0, 1 \rangle \), there is an equivalent ADFA \( A = \langle \Sigma, \Gamma, \delta, q_0, F \rangle \) such that \(|\Gamma| \leq |V_N| + 1\) and

\[
|\delta| \leq \begin{cases} 
|V_N|(|V_N| + 1)/2 & \text{if } |V_N| \leq |\Sigma|; \\
|\Sigma|(2|V_N| - |\Sigma| + 1)/2 & \text{if } |V_N| > |\Sigma|.
\end{cases}
\]

Moreover, there is an SDD \( B \) that admits no equivalent ADFA \( A \) for which the strict inequality holds.

**Proof.** The first claim, \(|\Gamma| \leq |V_N| + 1\), clearly holds by the conversion.

In order to establish the second part of the theorem, we give a variant of the construction of \( A(B) \). We define \( C(B) \) from \( B \) by replacing the definition of \( \Gamma \) in \( A(B) \) with \( \Gamma = \{ P \mid P \in V - \{0\} \} \). For \( C(B) = \langle \Sigma, \Gamma, \delta, q_0, F \rangle \), we prove the inequality by induction on \(|V_N|\). Clearly \( A(B) \) is not bigger than \( C(B) \), and thus this claim implies the theorem. In the following discussion, we ignore the root of \( B \) and the initial state of \( C(B) \), because it does not affect the discussion of their description size. For \(|V_N| = 1\), it is easy to see that the claim holds. Suppose that \(|V_N| > 1\). Let \( B' \) be obtained from \( B \) by deleting an arbitrary nonterminal node \( P \) that has no incoming edge.

If \(|V_N| \leq |\Sigma|\), we have \(|\delta'| \leq (|V_N| - 1)|V_N|/2 \) by the induction hypothesis, where \( \delta' \) denotes the transition set of \( C(B') \). By definition, \( C(B) \) can be obtained from \( C(B') \) by adding one state \( \tilde{P} \) and at most \(|V_N|\) outgoing edges from it. Hence

\[
|\delta| \leq |\delta'| + |V_N| \leq (|V_N| - 1)|V_N|/2 + |V_N| = |V_N|(|V_N| + 1)/2.
\]

If \(|V_N| > |\Sigma|\), we have \(|\delta'| \leq |\Sigma|(2|V_N| - |\Sigma| - 1)/2 \) by the induction hypothesis. By definition, \( C(B) \) can be obtained from \( C(B') \) by adding one state \( \tilde{P} \) and at most \(|\Sigma|\) outgoing edges from it. Hence

\[
|\delta| \leq |\delta'| + |\Sigma| \leq |\Sigma|(2|V_N| - |\Sigma| - 1)/2 + |\Sigma| = |\Sigma|(2|V_N| - |\Sigma| + 1)/2.
\]

We have proven the inequality.

In order to see that the above bound is tight, consider the reduced SDD and the minimum ADFA for the language \( L_n = \{ a_0a_1 \ldots a_i \mid 0 \leq k \leq n - |\Sigma|, 0 \leq j \leq \min\{m, n\}, 1 \leq i_1 < \cdots < i_j \leq m \} \) over \( \Sigma = \{a_0, \ldots, a_m\} \) with \( a_0 < a_1 < \cdots < a_m \).

We note that if \( a_m < \cdots < a_1 < a_0 \), we have \(|V_N'| = |\delta| \) for the node set \( V' \) of the reduced SDD \( B' \) for \( L_n \) and the transition set \( \delta \) of the minimum ADFA \( A \) for \( L_n \) in the proof of Theorem 5. Hence an order on \( \Sigma \) induces a reduced SDD that has asymptotically \(|\Sigma|\) times more nodes than the one induced by another order on \( \Sigma \).
Corollary 6. For an order $\pi$ on $\Sigma$ and a finite language $L$ over $\Sigma$, let $B^\pi(L) = \langle \langle \Sigma, \pi \rangle, V^\pi, \pi^\pi, S, 0, 1 \rangle$ be the reduced SDD for $L$ that respects the order $\pi$ over $\Sigma$. For any order $\pi, \rho$ on $\Sigma$, we have $|V^\pi_N| \leq |\Sigma||V^\rho_N|$.

Proof. Let $\delta$ be the transition set of the minimum automaton for $L$. By Theorem 3, $|V^\pi_N| \leq |\delta|$. By Theorem 5, $|\delta| \leq |\Sigma||V^\rho_N|$. Hence $|V^\pi_N| \leq |\Sigma||V^\rho_N|$.

Through the conversion techniques presented above between ADFAs and SDDs and/or by Theorems 3 and 5, most known results on the size of minimum ADFAs can be translated into those on SDDs. A special case is where the set of all factors of a string is in concern. Let

$$Fact(w) = \{ y \in \Sigma^* | w = xyz \text{ for some } x, z \in \Sigma^* \}.$$ 

The literature has intensively studied the factor automata for the set $Fact(w)$.

Theorem 7 (Blumer et al. [2], Crochemore [4]). For $w \in \Sigma^*$, let $\Gamma$ and $\delta$ be the state set and the transition set of the minimum A DFA for $Fact(w)$. Then $|\Gamma| \leq 2|w| - 2$ and $|\delta| \leq 3|w| - 4$.

Corollary 8. For $w \in \Sigma^*$, let $V$ be the node set of the reduced SDD for $Fact(w)$. Then $|w| \leq |V_N| \leq 3|w| - 4$.

Proof. By Theorem 7 and Theorem 5.

For $w = cb^a$ with $a < b < c$, we have $|V_N| = 3|w| - 4$.

Corollary 9. For $w \in \Sigma^*$ and order $\pi$ on $\Sigma$, let $V^\pi$ be the node set of the reduced SDD for $Fact(w)$. Then $|V^\pi_N| \leq |V^\rho_N| + |w| - 1$. Moreover, there are $w, \pi$ and $\rho$ for which the equality holds.

Proof. Let $\delta$ be the transition set of the minimum automaton for $Fact(w)$. We have $|V^\pi_N| \leq |\delta|$ and $|\Gamma| \leq |V^\rho_N| + 1$ by Theorems 3 and 5, respectively. Blumer et al. [2, Lemma 1.6] show that $|\delta| \leq |\Gamma| + |w| - 2$. Hence

$$|V^\pi_N| \leq |\delta| \leq |\Gamma| + |w| - 2 \leq |V^\rho_N| + |w| - 1.$$ 

In fact for $w = a^n b$, $\pi = \langle b < a \rangle$, $\rho = \langle a \prec b \rangle$, we have $|V^\pi_N| = 2n - 1$ and $|V^\rho_N| = n - 1$.

4 Input- and Output-Sensitive Time-bounds for Binary Synthesis Operations

In this section, we consider time complexity of set operations on SDDs. In particular, given a binary set operation $\diamond \in \{ \cup, \cap, \setminus, \cdot \}$, we consider the synthesis problem that receives two reduced SDDs $P, Q$ and computes $R = P \diamond Q$, where $P \diamond Q$ denotes the reduced SDD such that $L(P \diamond Q) = L(P) \diamond L(Q)$. Bryant [3] presented in his seminal paper on BDDs, a recursive synthesis algorithm for all Boolean operations. Loekito et al. [12] gave its string-counterpart for union $\cup$ and difference $\setminus$, Below, we generalize the algorithm in [12] for a family of set operations, called melding, in the style of Knuth [11].
4.1 The Family of Melding Operations

We give a family of binary set operations ⋄ called melding below. A terminal operation table is a binary Boolean function \( F : \{0, 1\}^2 \rightarrow \{0, 1\} \) such that \( F[0, 0] = 0 \). Clearly, there are exactly eight such tables. Let \( O = \{\cup, \cap, \setminus, /, \oplus, \emptyset, \text{LHS}, \text{RHS}\} \) be a set of names of set operations ⋄. Let \( \text{uniqtable} \) be any SDD-terminal operation table \( F_0 \). A specified terminal operation table \( F_0 \) uniquely determines melding operation \( P \diamond Q \). In what follows, we assume that the inputs \( P \) and \( Q \) and the output \( R \) are built by using the same hashtable \( \text{uniqtable} \), where \( \text{uniqtable} \) is initialized with the empty relation before constructing \( P \) and \( Q \). Moreover, our algorithm uses a hashtable \( \text{cache} : \text{op} \times \text{dom}^2 \rightarrow \text{dom} \) that stores invocation patterns of operations for avoiding redundant computation, where \( \text{op} \) is the set of operation names. By a similar discussion in Knuth [11], we establish the following theorem. Meld_0 directly computes the output without producing redundant nodes.

**Theorem 10 (correctness).** Let ⋄ \( \in O \) be any of the eight operations. Given \( F_0 \), the algorithm \( \text{Meld}_0 \) in Fig. 4 correctly computes the reduced SDD for \( R = P \diamond Q \) exactly eight string set operations \( P \diamond Q \), where the set operation \( P \diamond Q \) is defined as follows:

- the union \( P \cup Q \),
- the intersection \( P \cap Q \),
- the difference \( P \setminus Q \),
- the inverse difference \( P / Q = Q \setminus P \),
- the symmetric difference \( P \oplus Q = (P \setminus Q) \cup (Q \setminus P) \),
- the empty set \( \emptyset \),
- the left hand side \( \text{LHS}(P, Q) = P \).
- the right hand side \( \text{RHS}(P, Q) = Q \).

4.2 Input-Sensitive Complexity of Binary Synthesis

First, we start with input-sensitive analysis of the time complexity for the melding procedure. We prepare some necessary notations. Consider the algorithm \( \text{Meld}_0 \) of Fig. 4. Let us denote by \( \text{Meld}_0^0 \) and \( \text{Meld}_0^1 \) the first and second parts of the algorithm, that is, the toplevel if-clause and else-clause consisting of Lines 1 to 5 and Lines 6 to 12, respectively. For a procedure \( \alpha \), \( \#\alpha(P, Q) \) denotes the number of times that \( \alpha \)
the style of Wegener [21], which analyzed the time complexity of Boolean operations.

4.3 Pseudo Output Sensitive Complexity of Binary Synthesis

Next, we present output-sensitive analysis of the time complexity of the merging in the style of Wegener [21], which analyzed the time complexity of Boolean operations...
for BDDs based on the size of non-reduced BDDs. We define \( R^* = P \circ_s Q \) to be the (possibly non-reduced) SDD computed by \( \text{Meld}_o \) equipped with the modification of \( \text{Getnode} \) in Fig. 4 by removing Line 1 and 2 for node-sharing and zero-suppress rules. Clearly, the non-reduced output size \( |R^*| \) is bounded from above by \( O(|P| \cdot |Q|) \).

**Theorem 13 (output-sensitive complexity w.r.t. non-reduced output).** The reduced SDD for \( R = P \circ_s Q \) can be computed in \( O(|R^*|) \) time and space by the algorithm \( \text{Meld}_o \) in Fig. 4, where \( R^* \) is the non-reduced SDD for \( P \circ_s Q \).

**Proof.** Consider the computation of \( \text{Meld}_o \) of Fig. 4 equipped with \( \text{Getnode}^* \). Since each call of \( \text{Getnode}^* \) increases the output size by at least one, we have \( \#\text{Getnode}^* \leq |R^*| \) (Claim 4). Since exactly one of Line 8, 9, and 10 is executed in \( \text{Meld}_o \) and it contains at least one call for \( \text{Getnode} \), we have \( \#\text{Meld}_o \leq \#\text{Getnode}^* \) (Claim 5). From the proof for Theorem 11, we have \( \#\text{Meld}_o \leq 2 \cdot \#\text{Meld}_o + 1 \) (Claim 3). Combining Claims 3, 4, and 5 above, we now have \( \#\text{Meld}_o \leq 2 \cdot \#\text{Meld}_o + 1 \leq 2 \cdot |R^*| + 1 = O(|R^*|). \) Since \( \text{uniqtable} \) and \( \text{cache} \) contain at most \( \#\text{Getnode}^* \) and \( \#\text{Meld}_o \) entries, resp., the space complexity follows from a similar argument to the proof for Theorem 11.

**4.4 A Lowerbound for the Time Complexity of Binary Synthesis**

In BDD community, there has been a strong belief that the quadratic input-sensitive complexities of the binary synthesis procedures for a number of variants of BDDs, including the BDDs and ZDDs, is output-linear time for most input instances, and there has been no super-linear lowerbounds for its time complexity. Recently, Yoshinaka et al. [22] show that this conjecture is not true for BDDs and ZDDs; They constructed an infinite sequence of input BDDs that demonstrated the quadratic lowerbounds for the time complexity of the melding for BDDs and ZDDs. Based on their discussion, below we show that the above quadratic input-sensitive complexity of the melding is terms of input size is optimal for SDDs in reality.

**Theorem 14.** Let \( \circ \) be any melding operations. The algorithm \( \text{Meld}_o \) of Fig. 4 requires \( \Omega(|P| \cdot |Q|) \) time and space regardless of the output size, where \( P \) and \( Q \) are the input SDDs.

**Proof.** Our example that the binary synthesis takes \( O(|P| \cdot |Q|) \) time to compute \( R = P \circ Q \) where \( |R| \) is linear in \( |P| + |Q| \) is just a straightforward translation of the one from [22]. The theorem can be shown in a way similar to [22]. Here we give a rough sketch of the proof. Let \( \Sigma = \{0, 1\} \). For a fixed positive integer \( n \), we define

\[
S = \{x_1y_1 \ldots x_n y_n z_1 \ldots z_m \in \{0, 1\}^{2n+m} | x_{\beta(z_1 \ldots z_m)} = 1\},
\]

\[
T = \{x_1y_1 \ldots x_n y_n z_1 \ldots z_m \in \{0, 1\}^{2n+m} | y_{\beta(z_1 \ldots z_m)} = 1\},
\]

where \( m = \lceil \log n \rceil \) and

\[
\beta(z_1 \ldots z_m) = \begin{cases} 1 + \sum_{k=1}^{m} 2^{-k} z_k & \text{if } \sum_{k=1}^{m} 2^{-k} z_k < n; \\ 1 & \text{otherwise.} \end{cases}
\]

We have

\[
S \circ T = \{x_1y_1 \ldots x_n y_n z_1 \ldots z_m \in \{0, 1\}^{2n+m} | F_o[x_{\beta(z_1 \ldots z_m)}, y_{\beta(z_1 \ldots z_m)}] = 1\}.
\]
Let P and Q be the reduced SDD for S and T, resp.

We first show that $|P|, |Q|, |R| = O(2^n)$. It is easy to see that every node in P and Q represents a set of strings of a fixed length, since all strings in S and T have the same length $2n + m$. We define the level of a node to be $2n + m - k$ if the node represents a set of strings of length $k$. Since the membership of $x_1 y_1 \ldots x_n y_n z_1 \ldots z_m$ to S does not depend on any of $y_i$, it is not hard to see that there are at most $O(2^k)$ nodes of level $2k$ for $0 \leq k < n$. The number of nodes of level $2k + 1$ is at most twice as big as that of level $2k$. On the other hand, since there are at most $2^k$ distinct sets of strings of length $k$, there are at most $|\Sigma| \cdot 2^{2k}$ nodes of level $2n + m - k$ for $0 \leq k \leq m = \lceil \log n \rceil$. All in all, $|P| = O(2^n)$. Similarly $|Q| = O(2^n)$. It is easy to see that for any $x_1, y_1, x_1', y_1' \in \{0, 1\}$ such that $F_{o}[x_1, y_1] = F_{o}[x_1', y_1']$, we have $w_1 x_1 y_1 w_2 \in S \circ T$ iff $w_1 x_1' y_1' w_2 \in S \circ T$ for any $w_1 \in \{0, 1\}^{2k}, w_2 \in \{0, 1\}^{2n+m-k-2}$ with $k < n$. Hence we have $|R| = O(2^n)$ by a discussion similar to the one for $|P|, |Q| = O(2^n)$.

Second we show that $\#Meld_o \geq 2^{2n}$. For $w \in \{0, 1\}^{2n}$, let $P_w$ denote the node of P such that $L(P_w) = \{ w' | w w' \in S \}$. In fact P has such a node for each w. Similarly we let $Q_w$ be such that $L(Q_w) = \{ w' | w w' \in T \}$. By definition, the algorithm calls $Meld_o(P_w, Q_w)$ for each w. Moreover, $P_{x_1 \ldots y_n} \neq P_{x_i' \ldots y_n}$ whenever $x_i \neq x_i'$ for some i and $Q_{x_1 \ldots y_n} \neq Q_{x_i' \ldots y_n}$ whenever $y_i \neq y_i'$ for some i. Therefore, for distinct $w, w' \in \{0, 1\}^{2n}$, the pairs $\langle P_w, Q_w \rangle$ and $\langle P_{w'}, Q_{w'} \rangle$ are distinct. This means that $\#Meld_o \geq 2^{2n}$. □

5 Experiments

This section presents our experimental results on SDDs. Our first experiment has constructed SDDs and DFAs for the same sets of strings of real data and compared their sizes. Secondly we have implemented the binary synthesis algorithm Meld_o and computed different binary operations on sets over SDDs.

Setting. The data sets used in our experiments are summarized in Table 1. BibleAll and BibleBi are sets of all sentences and all word bi-grams drawn from an English text bible.txt and Ecoli is a single DNA string in ecoli.txt in Canterbury corpus². We implemented our shared and reduced SDD environment on the top of the SAPPORO BDD package [17] for BDDs and ZDDs written in C and C++, where each node is encoded in a 32-bit integer and a node triple occupies approximately 30

² http://corpus.canterbury.ac.nz/resources/
bytes in average including hash entries in *uniqtable*. We also used another implementation of SDD environment in functional language *Erlang*. Experiments were run on a PC (Intel Core i7, 2.67 GHz, 3.25 GB memory, Windows XP SP3). About 1.5 GB of memory was allocated to the SDD environment in maximum.

**Exp 1: Comparison of the size of indexes.** Figure 5 shows the sizes of SDDs and DFAs for different sets of strings, where **BibleAll (Fac)**, **BibleBi (Fac)** and **Ecoli (Fac)** mean the sets of all factors of sequences in the respective input files. We see that a minimal SDD is 0 to 23 percent more succinct than the equivalent minimal ADFA in binary format. In particular, the size ratio for factor sets is even smaller than that for the original string data.

**Exp 2: Binary synthesis.** We divided the source texts into two parts, the first half $H_1$ and the second $H_2$, and then performed $\text{Meld}_\diamond$ on those parts for $\diamond \in \{\cup, \cap, \setminus, /\}$. The results are presented in Table 2. It took less than seconds to compute set operations $\diamond$ on two SDDs with around three to four millions of nodes each. The output size of $H_1 \cup H_2$ is much larger than that of $H_1 \cap H_2$, but the running time is not different that much.

Overall, we conclude that the shared and reduced SDD environment with the above algorithms is a practical choice for storing and manipulating string sets in large-scale string applications.

### 6 Conclusion

In this paper, we consider the class of sequence binary decision diagrams (SDDs) proposed by Loekito *et al.* [12], and studied two fundamental problems on sequence binary decision diagrams, the relationship to acyclic automata and the complexities of the binary synthesis operation. In Sec. 4, we showed the quadratic time complexity of the $\text{Meld}_\diamond$ algorithm. In [7], it is shown that the $\text{Meld}_\diamond$ runs in input linear time if one of the argument is the minimal SDD of linear shape corresponding to a string. Therefore, it would be an interesting future problem to study special cases that $\text{Meld}_\diamond$ has input linear time complexity. It would be another problem to apply SDDs for studying the dynamic versions of sequence analysis problems such as the maximal repeat problem and the consistent string problem [9].
References

17. S. Minato. SAPPORO BDD package. DCS, Hokkaido University, unreleased, 2011.