Longest Common Parameterized Subsequences with Fixed Common Substring

Anna Gorbenko

Department of Intelligent Systems and Robotics
Ural Federal University
620083 Ekaterinburg, Russia
gorbenko.ann@gmail.com

Vladimir Popov

Department of Intelligent Systems and Robotics
Ural Federal University
620083 Ekaterinburg, Russia
Vladimir.Popov@usu.ru

Abstract

In this paper we consider the problem of the longest common parameterized subsequence with fixed common substring (STR-IC-LCPS). In particular, we show that STR-IC-LCPS is NP-complete. We describe an approach to solve STR-IC-LCPS. This approach is based on an explicit reduction from the problem to the satisfiability problem.

Keywords: parameterized pattern matching, satisfiability, NP-complete

Different variants of the problem of the longest common subsequence are extensively used as distance measures for strings. In particular, the following problem was proposed in [1] (see also [2]). STR-IC-LCS:

Given two strings \(S_1\) and \(S_2\) and a constraint pattern \(P\) of length \(n, m,\) and \(r\), respectively, find a longest common subsequence of \(S_1\) and \(S_2\) including \(P\) as a substring.

Another well-studied string comparison measure is that of parameterized matching (basic definitions and results can be found in [3]). It is natural to attempt to accommodate parameterized matching along with some other
distance measures. In this paper we consider a parameterized variant of STR-IC-LCS.

The problem of the longest common parameterized subsequence with fixed common substring (STR-IC-LCPS):

Instance: An alphabet $\Sigma \cup \Pi$, sequences $S_1$ and $S_2$ over $\Sigma \cup \Pi$, a string $P$ over $\Sigma$, and positive integer $k$.

Question: Is there a sequence $T$, $|T| \geq k$, such that $P$ is a substring of $T$ and $T$ is a parameterized subsequence of $S_1$ and $S_2$?

It is clear that there is some connection between longest common parameterized subsequences and longest common parameterized subsequences with fixed common substring. In particular, if $T_1$ is a longest common parameterized subsequence of $S_1$ and $S_2$ and $T_2$ is a parameterized subsequence of $S_1$ and $S_2$ with fixed common substring $P$, then $|T_1| \geq |T_2|$. However, $T_1$ and $T_2$ may significantly differ from each other.

Theorem 1. For any $n$ and $k$, there are sequences $S_1$, $S_2$, $P$, $T_1$, and $T_2$ such that

1. $T_1$ is a longest common parameterized subsequence of $S_1$ and $S_2$;
2. $T_2$ is a longest common parameterized subsequence of $S_1$ and $S_2$ with fixed common substring $P$;
3. $|T_2| \geq n$;
4. $|T_1| \geq |T_2| + k$.

Proof. Let $\Sigma = \{a, b\}$, $\Pi = \emptyset$, $S_1 = a^s b^t$, $S_2 = b^t a^s$, $P = b^t$. We assume that $s > t + k$ and $t > n$. Let $T_1 = a^s$. Since $s > t + k$, it is clear that $T_1$ is a longest common parameterized subsequence of $S_1$ and $S_2$. Let $T_2 = b^t$. It is easy to see that $T_2$ is a parameterized subsequence of $S_1$ and $S_2$. Since $P = b^t$, it is clear that $P$ is a substring of $T_2$. In view of $P = b^t$, it is easy to check that $T_2$ is a longest common parameterized subsequence of $S_1$ and $S_2$ with fixed common substring $P$. By definition of $T_2$, in view of $t > n$, it is clear that $|T_2| \geq n$. Since $s > t + k$, it is easy to see that $|T_1| \geq |T_2| + k$.

Now we consider the complexity of STR-IC-LCPS.

Theorem 2. STR-IC-LCPS is NP-complete.

Proof. It is clear that STR-IC-LCPS is in NP. In order to prove that STR-IC-LCPS is NP-hard, we shall reduce LCPS (see [4]) to STR-IC-LCPS.

LCPS:

Instance: An alphabet $\Sigma \cup \Pi$, sequences $S_1$ and $S_2$ over $\Sigma \cup \Pi$, and positive integer $k$.

Question: Is there a sequence $T$, $|T| \geq k$, that is a parameterized subsequence of $S_1$ and $S_2$?

Let $\Sigma \cup \Pi$ be an alphabet. Let $S_1$ and $S_2$ are sequences over $\Sigma \cup \Pi$.

We assume that $c$ is a letter such that $c \notin \Sigma \cup \Pi$. Let $\Sigma' = \Sigma \cup \{c\}$. Let $P = c$ and $S_i' = cS_i$, $i \in \{1, 2\}$.
It is easy to check that $T$ is a longest common parameterized subsequence of $S_1$ and $S_2$ if and only if $cT$ is a longest common parameterized subsequence of $S'_1$ and $S'_2$ with fixed common substring $P$. Note that LCPS is \textbf{NP}-complete \cite{4}. Therefore, STR-IC-LCPS is \textbf{NP}-complete. \hfill \Box

Encoding different hard problems as Boolean satisfiability and solving them with very efficient satisfiability algorithms has caused considerable interest (see e.g. \cite{5} - \cite{22}). We consider an explicit reduction from STR-IC-LCPS to the satisfiability problem.

Let $\Sigma = \{a_1, a_2, \ldots, a_{|\Sigma|}\}$, $\Pi = \{b_1, b_2, \ldots, b_{|\Pi|}\}$,

- $\varphi[1] = \land_{1 \leq i \leq k} \land_{1 \leq j \leq |\Sigma|, \Pi} x[i, j]$,
- $\varphi[2] = \land_{1 \leq i \leq k} \land_{1 \leq j[1] < j[2] \leq |\Pi|} (\neg x[i, j[1]] \lor \neg x[i, j[2]])$,
- $\varphi[3] = \land_{1 \leq i \leq |\Pi|} \land_{1 \leq j \leq |\Sigma|} u[i, j]$,
- $\varphi[4] = \land_{1 \leq i \leq |\Pi|} \land_{1 \leq j[1] < j[2] \leq |\Sigma|} (\neg u[i, j[1]] \lor \neg u[i, j[2]])$,
- $\varphi[5] = \land_{1 \leq i \leq |\Pi|} \land_{1 \leq j \leq |\Sigma|, \Pi[i]\neq a_j} \neg u[i, j]$,
- $\varphi[6] = \land_{1 \leq i \leq k} \land_{1 \leq j \leq |\Sigma|, \Pi} v[i, j]$,
- $\varphi[7] = \land_{1 \leq i \leq k, 1 \leq j \leq |\Pi|, 1 \leq s \leq |\Sigma|} ((\neg v[i] \lor \neg u[j, s] \lor x[j + i - 1, s]) \land
  (\neg v[i] \lor u[j, s] \lor x[j + i - 1, s]))$,
- $\varphi[8] = \land_{1 \leq i \leq |\Sigma|, \Pi} \land_{1 \leq j \leq |\Sigma|, \Pi} y[i, j]$,
- $\varphi[9] = \land_{1 \leq i \leq |\Sigma|, \Pi} \land_{1 \leq j[1] < j[2] \leq |\Sigma|} (\neg y[i, j[1]] \lor \neg y[i, j[2]])$,
- $\varphi[10] = \land_{1 \leq i \leq |\Sigma|, \Pi[1]} \land_{1 \leq j \leq |\Sigma|} \neg y[i, j]$,
- $\varphi[11] = \land_{1 \leq i[1] < i[2] \leq |\Sigma|, \Pi[1]} \land_{1 \leq j \leq |\Sigma|} \neg y[i[1], j] \land y[i[2], j] \land y[i[1], j] \lor y[i[2], j]$,
- $\varphi[12] = \land_{1 \leq i[1] < i[2] \leq |\Sigma|, \Pi[1]} \land_{1 \leq j \leq |\Sigma|} \neg y[i[1], j] \lor \neg y[i[2], j]$,
- $\varphi[13] = \land_{1 \leq i \leq |\Sigma|, \Pi[1]} \land_{1 \leq j \leq |\Sigma|} \land_{1 \leq k \leq |\Sigma|, \Pi[1]} \land_{1 \leq l \leq |\Sigma|} (\neg z[1, i, j] \lor \neg x[j, l] \lor \neg z[1, i, j])$,
- $\varphi[14] = \land_{1 \leq i \leq |\Sigma|, \Pi[1]} \land_{1 \leq j \leq |\Sigma|} \land_{1 \leq k \leq |\Sigma|, \Pi[1]} \land_{1 \leq l \leq |\Sigma|} (\neg z[2, i, j] \lor \neg x[j, l] \lor \neg z[2, i, j])$,
- $\varphi[15] = \land_{1 \leq i \leq |\Sigma|, \Pi[1]} \land_{1 \leq j \leq |\Sigma|} \land_{1 \leq k \leq |\Sigma|, \Pi[1]} \land_{1 \leq l \leq |\Sigma|} (\neg z[2, i, j] \lor \neg x[j, l] \lor \neg z[2, i, j])$,
- $\varphi[16] = \land_{1 \leq i \leq |\Sigma|, \Pi[1]} \land_{1 \leq j \leq |\Sigma|} (\neg z[2, i, j] \lor \neg y[i, l] \lor x[j, l]) \land
  (\neg z[2, i, j] \lor y[i, l] \lor \neg x[j, l])$,
- $\varphi[17] = \land_{1 \leq i \leq |\Sigma|, \Pi[1]} \land_{1 \leq j \leq |\Sigma|} \land_{1 \leq k \leq |\Sigma|} (\neg z[i, j, l[1]] \lor \neg z[i, j, l[2]]$,
\[ \varphi[18] = \land_{1 \leq i \leq 2, 1 \leq l \leq k} \lor_{1 \leq j \leq |S_i|} z[i, j, l], \]

\[ \varphi[19] = \land_{1 \leq i \leq 2, 1 \leq j[1] \leq |S_i|} (\neg z[i, j[1], l[1]] \lor \neg z[i, j[2], l[2]]), \]

\[ 1 \leq l[1] \leq k, \]


\[ \xi = \land_{i=1}^{10} \varphi[i]. \]

It is easy to check that there is a sequence \( T, |T| \geq k \), such that \( P \) is a substring of \( T \) and \( T \) is a parameterized subsequence of \( S_1 \) and \( S_2 \) if and only if \( \xi \) is satisfiable. It is clear that \( \xi \) is a CNF. So, \( \xi \) gives us an explicit reduction from STR-IC-LCPS to SAT. Now, using standard transformations (see e.g. [23]) we can obtain an explicit transformation \( \xi \) into \( \zeta \) such that \( \xi \iff \zeta \) and \( \zeta \) is a 3-CNF. Clearly, \( \zeta \) gives us an explicit reduction from STR-IC-LCPS to 3SAT.

We have designed generators of natural random instances for STR-IC-LCPS. We have consider our genetic algorithms OA[1] (see [24]), OA[2] (see [25]), OA[3] (see [26]), and OA[4] (see [27]) for SAT. We have used heterogeneous cluster. Each test was runned on a cluster of at least 100 nodes. Note that due to restrictions on computation time (20 hours) we used savepoints. Selected experimental results are given in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>average</th>
<th>max</th>
<th>best</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA[1]</td>
<td>2.7 h</td>
<td>31.72 h</td>
<td>14 min</td>
</tr>
<tr>
<td>OA[2]</td>
<td>2.32 h</td>
<td>26.4 h</td>
<td>18 min</td>
</tr>
<tr>
<td>OA[3]</td>
<td>1.74 h</td>
<td>29 h</td>
<td>26 min</td>
</tr>
<tr>
<td>OA[4]</td>
<td>1.96 h</td>
<td>14.4 h</td>
<td>21.2 min</td>
</tr>
</tbody>
</table>

Table 1: Experimental results for STR-IC-LCPS.

ACKNOWLEDGEMENTS. The work was partially supported by Analytical Departmental Program “Developing the scientific potential of high school” 8.1616.2011.

References


Received: November 1, 2012