The String Barcoding Problem

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 an approach to solve the string barcoding problem. This approach is based on an explicit reduction from the problem to the satisfiability problem.

Keywords: string barcoding problem, satisfiability, NP-complete

Investigation of different regularities can be used to identify various important knowledge (see e.g. [1] – [15]). In particular, the string barcoding problem was proposed for rapid identification of unknown pathogens [16].

Given sequences $S$ and $T$ over some finite alphabet $\Sigma$. Let $S \leq T$ if and only if $S$ is a subsequence of $T$. Let

$$S(\{S[1], \ldots, S[n]\}) = \{X \mid \exists i \in \{1, \ldots, n\}(X \leq S[i])\}.$$ 

Let

$$X!(S, T)$$

if and only if

$$(X \leq S \land X \not\leq T) \lor (X \not\leq S \land X \leq T).$$

Let

$$P \subseteq S(\{S[1], \ldots, S[n]\}).$$
Let

\[ P_{i,j} = \{X \mid (X \in P) \land X!((S[i], S[j])) \}. \]

**The string barcoding problem (SBP):**

**Instance:** Given a set

\[ \{S[1], \ldots, S[n]\} \]

of strings over some finite alphabet \( \Sigma \),

\[ Q \subseteq S(\{S[1], \ldots, S[n]\}), \]

and positive integers \( d \) and \( r \).

**Question:** Is there a set \( P \subseteq Q \) such that \( |P| \leq d \) and \( |P_{i,j}| \geq r \), for any \( i \neq j \), \( i, j \in \{1, \ldots, n\} \)?

Note that SBP is \( \text{NP} \)-complete \([17]\). Encoding different hard problems as instances of SAT and solving them with efficient satisfiability algorithms has caused considerable interest (see e.g. \([18] \text{–} [37]\)). In this paper, we consider an approach to solve the SBP problem. Our approach is based on an explicit reduction from the problem to the satisfiability problem.

Let \( \Sigma = \{a_1, a_2, \ldots, a_m\} \), \( Q = \{Q[1], \ldots, Q[k]\} \), \( q = \max_{i=1}^{k} |Q[i]| \). We assume that \( Q[i, j] \) is the \( j \)th letter of \( Q[i] \). Let

\[
\phi[1] = \bigwedge_{1 \leq i \leq d, 1 \leq j \leq q, 0 \leq s \leq m} x[i, j, s],
\]

\[
\phi[2] = \bigwedge_{1 \leq i \leq d, 1 \leq j \leq q} \left( \neg x[i, j, s[1]] \lor \neg x[i, j, s[2]] \right),
\]

\[
\phi[3] = \bigwedge_{1 \leq i \leq d} \bigvee_{1 \leq j \leq k} y[i, j],
\]

\[
\phi[4] = \bigwedge_{1 \leq i \leq d, 1 \leq j \leq q, 1 \leq t \leq d} \left( \neg y[i, j[1]] \lor \neg y[i, j[2]] \right),
\]

\[
\phi[5] = \bigwedge_{1 \leq i \leq d, 1 \leq j \leq k} \left( \neg y[i, j] \lor x[i, t, s] \right),
\]

\[
\phi[6] = \bigwedge_{1 \leq i \leq d, 1 \leq j \leq k, 1 \leq t \leq \|Q[j]\|, 0 \leq s \leq m, Q[j][t] = a_s} \neg y[i, j] \lor x[i, t, s],
\]

\[
\psi[1] = \bigwedge_{1 \leq i \leq n, 1 \leq t \leq d} z[i, j, s, t].
\]
\[ \psi[2] = \land_{1 \leq i \leq n, 1 \leq j \leq n, i \neq j, 1 \leq s \leq r, 1 \leq t[1] \leq t[2] \leq d} (\neg z[i, j, s, t[1]] \lor \neg z[i, j, s, t[2]]), \]

\[ \psi[3] = \land_{1 \leq i \leq n, 1 \leq j \leq n, i \neq j, 1 \leq t \leq d, 1 \leq s[1] \leq s[2] \leq r} (\neg z[i, j, s[1], t] \lor \neg z[i, j, s[2], t]), \]

\[ \psi[4] = \land_{1 \leq i \leq n, 1 \leq j \leq n, i \neq j, 1 \leq s \leq r, 1 \leq t \leq q} \lor 1 \leq p \leq |S[i]| u[i, j, s, t, p], \]

\[ \psi[5] = \land_{1 \leq i \leq n, 1 \leq j \leq n, i \neq j, 1 \leq s \leq r, 1 \leq t \leq q} \lor 1 \leq p[1] \leq p[2] \leq |S[i]| (\neg u[i, j, s, t, p[1]] \lor u[i, j, s, t, p[2]]), \]

\[ \psi[6] = \land_{1 \leq i \leq n, 1 \leq j \leq n, i \neq j, 1 \leq s \leq r, 1 \leq t \leq d, 1 \leq p \leq k, 1 \leq b[1] \leq b[2] \leq |Q[p]|, 1 \leq c[1] \leq c[2] \leq |S[i]|} (\neg y[t, p] \lor u[i, j, s, b[1], c[1]] \lor u[i, j, s, b[2], c[2]]), \]

\[ \psi[7] = \land_{1 \leq i \leq n, 1 \leq j \leq n, i \neq j, 1 \leq s \leq r, 1 \leq t \leq q} \lor 1 \leq p \leq |S[j]| v[i, j, s, t, p], \]

\[ \psi[8] = \land_{1 \leq i \leq n, 1 \leq j \leq n, i \neq j, 1 \leq s \leq r, 1 \leq t \leq q} \lor 1 \leq p[1] \leq p[2] \leq |S[j]| (\neg v[i, j, s, t, p[1]] \lor v[i, j, s, t, p[2]]), \]
\[ \psi[9] = \land_{1 \leq i \leq n, \atop 1 \leq j \leq n, \atop i \neq j, \atop 1 \leq s \leq r, \atop 1 \leq t \leq d, \atop 1 \leq p \leq k, \atop 1 \leq b[1] < b[2] \leq |Q[p]|, \atop 1 \leq c[2] \leq c[1] \leq |S[j]|} \land_{1 \leq i \leq n, \atop 1 \leq j \leq n, \atop i \neq j, \atop 1 \leq s \leq r, \atop 1 \leq t[1] \leq d, \atop 1 \leq t[2] \leq k, \atop 1 \leq t[3] \leq |Q[t[2]]|, \atop 1 \leq t[4] \leq |S[i]|, \atop 1 \leq t[5] \leq m, \atop S[i][t[4]] = a_{t[5]}} \\ (\neg z[i, j, s, t] \land \neg y[t, p] \land \neg v[i, j, s, b[1], c[1]] \land \neg v[i, j, s, b[2], c[2]]), \]

\[ \tau[1] = \land_{1 \leq i \leq n, \atop 1 \leq j \leq n, \atop i \neq j, \atop 1 \leq s \leq r, \atop 1 \leq t[1] \leq d, \atop 1 \leq t[2] \leq k, \atop 1 \leq t[3] \leq |Q[t[2]]|, \atop 1 \leq t[4] \leq |S[i]|, \atop 1 \leq t[5] \leq m, \atop S[i][t[4]] = a_{t[5]}} \\ \neg w[i, j, t] \land \neg u[i, j, s, b[1], c[1]] \land \neg t[i, t[1]] \land \neg x[t, t[3], t[4]] \land \neg y[t[1], t[2]], \]

\[ \tau[2] = \land_{1 \leq i \leq n, \atop 1 \leq j \leq n, \atop i \neq j, \atop 1 \leq s \leq r, \atop 1 \leq t[1] \leq d, \atop 1 \leq t[2] \leq k, \atop 1 \leq p[1] \leq \ldots \leq p[|Q[t[2]]|] \leq |S[j]|, \atop S[j][p[b]] = a_{c[b]}, \atop 1 \leq c[b] \leq m, \atop 1 \leq b \leq |Q[t[2]]|} \\ (\lor_{1 \leq c[b] \leq m, 1 \leq b \leq |Q[t[2]]|} \neg x[t[1], b, c[b]]), \]

\[ \tau[3] = \land_{1 \leq i \leq n, \atop 1 \leq j \leq n, \atop i \neq j, \atop 1 \leq s \leq r, \atop 1 \leq t[1] \leq d, \atop 1 \leq t[2] \leq k, \atop 1 \leq t[3] \leq |Q[t[2]]|, \atop 1 \leq t[4] \leq |S[j]|, \atop 1 \leq t[5] \leq m, \atop S[j][t[4]] = a_{t[5]}} \\ \neg z[i, j, t[1]] \land \neg y[t[1], t[2]], \]
The string barcoding problem

\[ w[i, j, s] \lor \neg v[i, j, s, t[3], t[4]] \lor x[t[1], t[3], t[5]], \]
\[ \tau[4] = \land_{1 \leq i \leq n,} \left( \land_{1 \leq j \leq n,} \left( \land_{i \neq j,} \right) \left( \land_{1 \leq s \leq r,} \left( \land_{1 \leq t[1] \leq d,} \left( \land_{1 \leq t[2] \leq k,} \left( \land_{1 \leq p[1] \leq \ldots \leq p[i] \leq |Q[t[2]]|} S[i, p[b]] = a[b]), \right) \right) \right) \right) \left( \land_{1 \leq c[b] \leq m,} \left( \land_{1 \leq b \leq |Q[t[2]]|} \neg x[t[1], b, c[b]] \right) \right), \]
\[ \xi = (\land_{i=1}^6 \varphi[i]) \land (\land_{i=1}^9 \psi[i]) \land (\land_{i=1}^4 \tau[i]). \]

It is easy to check that there is a set \( P \subseteq Q \) such that \( |P| \leq d \) and \( |P_{i,j}| \geq r \), for any \( i \neq j \), \( i, j \in \{1, \ldots, n\} \), if and only if \( \xi \) is satisfiable. Clearly, \( \xi \) is a CNF. So, \( \xi \) gives us an explicit reduction from SBP to SAT. Now, using standard transformations (see e.g. [38]) we can obtain an explicit transformation \( \xi \) into \( \zeta \) such that \( \xi \iff \zeta \) and \( \zeta \) is a 3-CNF. It is easy to see that \( \zeta \) gives us an explicit reduction from SBP to 3SAT.

For computational experiments, we have designed a generator of natural instances for SBP. We have considered our genetic algorithms OA[1] (see [39]) and OA[2] (see [40]) for SAT. We have used heterogeneous cluster. Each test was runned on a cluster of at least 100 nodes. Selected experimental results are given in Table 1.

<table>
<thead>
<tr>
<th>time</th>
<th>average</th>
<th>max</th>
<th>best</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA[1]</td>
<td>47.26 min</td>
<td>9.18 h</td>
<td>14.21 sec</td>
</tr>
<tr>
<td>OA[2]</td>
<td>58.13 min</td>
<td>4.71 h</td>
<td>2.83 min</td>
</tr>
</tbody>
</table>

Table 1: Experimental results for SBP.

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