Applied Mathematical Sciences, Vol. 7, 2013, no. 13, 615 - 622

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 Σ. Let $S \leq T$ if and only if S is a subsequence of T. Let

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

Let

$$X!(S, T)$$

if and only if

$$(X \leq S \wedge X \not\leq T) \vee (X \not\leq S \wedge X \leq T).$$

Let

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

Let

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

THE STRING BARCODING PROBLEM (SBP):
INSTANCE: *Given a set*

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

of strings over some finite alphabet Σ*,*

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

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 **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] – [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 jth letter of $Q[i]$. Let

$$\begin{aligned}
\varphi[1] &= \wedge_{\substack{1 \leq i \leq d, \\ 1 \leq j \leq q}} \vee_{0 \leq s \leq m} x[i,j,s], \\
\varphi[2] &= \wedge_{\substack{1 \leq i \leq d, \\ 1 \leq j \leq q, \\ 0 \leq s[1] < s[2] \leq m}} (\neg x[i,j,s[1]] \vee \neg x[i,j,s[2]]), \\
\varphi[3] &= \wedge_{1 \leq i \leq d} \vee_{1 \leq j \leq k} y[i,j], \\
\varphi[4] &= \wedge_{\substack{1 \leq i \leq d, \\ 1 \leq j[1] < j[2] \leq k}} (\neg y[i,j[1]] \vee \neg y[i,j[2]]), \\
\varphi[5] &= \wedge_{\substack{1 \leq i \leq d, \\ 1 \leq j \leq k, \\ 1 \leq t \leq |Q[j]|, \\ 0 \leq s \leq m, Q[j,t] = a_s}} (\neg y[i,j] \vee x[i,t,s]), \\
\varphi[6] &= \wedge_{\substack{1 \leq i \leq d, \\ 1 \leq j \leq k, \\ |Q[j]| < t \leq q}} (\neg y[i,j] \vee x[i,t,0]), \\
\psi[1] &= \wedge_{\substack{1 \leq i \leq n, \\ 1 \leq j \leq n, \\ i \neq j, \\ 1 \leq s \leq r}} \vee_{1 \leq t \leq d} z[i,j,s,t],
\end{aligned}$$

$$\psi[2] = \bigwedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t[1]<t[2]\le d}} (\neg z[i,j,s,t[1]] \vee \neg z[i,j,s,t[2]]),$$

$$\psi[3] = \bigwedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le t\le d,\\ 1\le s[1]<s[2]\le r}} (\neg z[i,j,s[1],t] \vee \neg z[i,j,s[2],t]),$$

$$\psi[4] = \bigwedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t\le q}} \bigvee_{1\le p\le |S[i]|} u[i,j,s,t,p],$$

$$\psi[5] = \bigwedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t\le q,\\ 1\le p[1]<p[2]\le |S[i]|}} (\neg u[i,j,s,t,p[1]] \vee \neg u[i,j,s,t,p[2]]),$$

$$\psi[6] = \bigwedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t\le d,\\ 1\le p\le k,\\ 1\le b[1]<b[2]\le |Q[p]|,\\ 1\le c[2]\le c[1]\le |S[i]|}} (\neg z[i,j,s,t] \vee \neg y[t,p] \vee \neg u[i,j,s,b[1],c[1]] \vee \neg u[i,j,s,b[2],c[2]]),$$

$$\psi[7] = \bigwedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t\le q}} \bigvee_{1\le p\le |S[j]|} v[i,j,s,t,p],$$

$$\psi[8] = \bigwedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t\le q,\\ 1\le p[1]<p[2]\le |S[j]|}} (\neg v[i,j,s,t,p[1]] \vee \neg v[i,j,s,t,p[2]]),$$

$$\psi[9] = \wedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t\le d,\\ 1\le p\le k,\\ 1\le b[1]<b[2]\le|Q[p]|,\\ 1\le c[2]\le c[1]\le|S[j]|}} (\neg z[i,j,s,t] \vee \neg y[t,p] \vee \neg v[i,j,s,b[1],c[1]] \vee \neg v[i,j,s,b[2],c[2]]),$$

$$\tau[1] = \wedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t[1]\le d,\\ 1\le t[2]\le k,\\ 1\le t[3]\le|Q[t[2]]|,\\ 1\le t[4]\le|S[i]|,\\ 1\le t[5]\le m,\\ S[i,t[4]]=a_{t[5]}}} (\neg z[i,j,s,t[1]] \vee \neg y[t[1],t[2]] \vee \neg w[i,j,s] \vee \neg u[i,j,s,t[3],t[4]] \vee x[t[1],t[3],t[5]]),$$

$$\tau[2] = \wedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t[1]\le d,\\ 1\le t[2]\le k,\\ 1\le p[1]<\ldots<p[|Q[t[2]]|]\le|S[j]|,\\ S[j,p[b]]=a_{c[b]},\\ 1\le c[b]\le m,\\ 1\le b\le|Q[t[2]]|}} (\neg z[i,j,s,t[1]] \vee \neg y[t[1],t[2]] \vee \neg w[i,j,s] \vee (\vee_{1\le c[b]\le m, 1\le b\le|Q[t[2]]|} \neg x[t[1],b,c[b]])),$$

$$\tau[3] = \wedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t[1]\le d,\\ 1\le t[2]\le k,\\ 1\le t[3]\le|Q[t[2]]|,\\ 1\le t[4]\le|S[j]|,\\ 1\le t[5]\le m,\\ S[j,t[4]]=a_{t[5]}}} (\neg z[i,j,s,t[1]] \vee \neg y[t[1],t[2]] \vee$$

$$
\begin{aligned}
&& w[i,j,s] \vee \neg v[i,j,s,t[3],t[4]] \vee x[t[1],t[3],t[5]]),\\
\tau[4] &=& \wedge_{\substack{1\le i\le n,\\ 1\le j\le n,\\ i\ne j,\\ 1\le s\le r,\\ 1\le t[1]\le d,\\ 1\le t[2]\le k,\\ 1\le p[1]<\ldots<p[|Q[t[2]]|]\le|S[i]|,\\ S[i,p[b]]=a_{c[b]},\\ 1\le c[b]\le m,\\ 1\le b\le|Q[t[2]]|}} (\neg z[i,j,s,t[1]] \vee \neg y[t[1],t[2]] \vee w[i,j,s] \vee\\
&& (\vee_{1\le c[b]\le m, 1\le b\le|Q[t[2]]|}\neg x[t[1],b,c[b]])),\\
\xi &=& (\wedge_{i=1}^{6}\varphi[i]) \wedge (\wedge_{i=1}^{9}\psi[i]) \wedge (\wedge_{i=1}^{4}\tau[i]).
\end{aligned}
$$

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

time	average	max	best
OA[1]	47.26 min	9.18 h	14.21 sec
OA[2]	58.13 min	4.71 h	2.83 min

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

[1] V. Yu. Popov, Computational complexity of problems related to DNA sequencing by hybridization, *Doklady Mathematics*, 72 (2005), 642-644.

[2] V. Popov, The approximate period problem for DNA alphabet, *Theoretical Computer Science*, 304 (2003), 443-447.

[3] V. Popov, The Approximate Period Problem, *IAENG International Journal of Computer Science*, 36 (2009), 268-274.

[4] V. Popov, Approximate Periods of Strings for Absolute Distances, *Applied Mathematical Sciences*, 6 (2012), 6713-6717.

[5] V. Popov, Multiple genome rearrangement by swaps and by element duplications, *Theoretical Computer Science*, 385 (2007), 115-126.

[6] V. Popov, Sorting by prefix reversals, *IAENG International Journal of Applied Mathematics*, 40 (2010), 247-250.

[7] A. Gorbenko and V. Popov, Robot Self-Awareness: Occam's Razor for Fluents, *International Journal of Mathematical Analysis*, 6 (2012), 1453-1455.

[8] A. Gorbenko and V. Popov, The Force Law Design of Artificial Physics Optimization for Robot Anticipation of Motion, *Advanced Studies in Theoretical Physics*, 6 (2012), 625-628.

[9] A. Gorbenko, V. Popov, and A. Sheka, Robot Self-Awareness: Exploration of Internal States, *Applied Mathematical Sciences*, 6 (2012), 675-688.

[10] A. Gorbenko, V. Popov, and A. Sheka, Robot Self-Awareness: Temporal Relation Based Data Mining, *Engineering Letters*, 19 (2011), 169-178.

[11] A. Gorbenko and V. Popov, Robot Self-Awareness: Formulation of Hypotheses Based on the Discovered Regularities, *Applied Mathematical Sciences*, 6 (2012), 6583-6585.

[12] A. Gorbenko and V. Popov, Anticipation in Simple Robot Navigation and Finding Regularities, *Applied Mathematical Sciences*, 6 (2012), 6577-6581.

[13] A. Gorbenko and V. Popov, Robot Self-Awareness: Usage of Co-training for Distance Functions for Sequences of Images, *Advanced Studies in Theoretical Physics*, 6 (2012), 1243-1246.

[14] A. Gorbenko and V. Popov, Robot's Actions and Automatic Generation of Distance Functions for Sequences of Images, *Advanced Studies in Theoretical Physics*, 6 (2012), 1247-1251.

[15] A. Gorbenko and V. Popov, Anticipation in Simple Robot Navigation and Learning of Effects of Robot's Actions and Changes of the Environment, *International Journal of Mathematical Analysis*, 6 (2012), 2747-2751.

[16] S. Rash and D. Gusfield, String barcoding: uncovering optimal virus signatures, *Proceedings of the sixth annual international conference on Computational biology*, (2002), 254-261.

[17] M. Dalpasso, G. Lancia, and R. Rizzi, The String Barcoding Problem is NP-Hard, *Lecture Notes in Computer Science*, 3678 (2005), 88-96.

[18] A. Gorbenko and V. Popov, On the Problem of Sensor Placement, *Advanced Studies in Theoretical Physics*, 6 (2012), 1117-1120.

[19] A. Gorbenko and V. Popov, On the Longest Common Subsequence Problem, *Applied Mathematical Sciences*, 6 (2012), 5781-5787.

[20] A. Gorbenko and V. Popov, Computational Experiments for the Problem of Selection of a Minimal Set of Visual Landmarks, *Applied Mathematical Sciences*, 6 (2012), 5775-5780.

[21] A. Gorbenko and V. Popov, The Binary Paint Shop Problem, *Applied Mathematical Sciences*, 6 (2012), 4733-4735.

[22] A. Gorbenko, M. Mornev, V. Popov, and A. Sheka, The Problem of Sensor Placement, *Advanced Studies in Theoretical Physics*, 6 (2012), 965-967.

[23] A. Gorbenko, V. Popov, and A. Sheka, Localization on Discrete Grid Graphs, *Lecture Notes in Electrical Engineering*, 107 (2012), 971-978.

[24] A. Gorbenko and V. Popov, The Problem of Selection of a Minimal Set of Visual Landmarks, *Applied Mathematical Sciences*, 6 (2012), 4729-4732.

[25] A. Gorbenko, M. Mornev, V. Popov, and A. Sheka, The problem of sensor placement for triangulation-based localisation, *International Journal of Automation and Control*, 5 (2011), 245-253.

[26] A. Gorbenko and V. Popov, The Longest Common Parameterized Subsequence Problem, *Applied Mathematical Sciences*, 6 (2012), 2851-2855.

[27] A. Gorbenko and V. Popov, Programming for Modular Reconfigurable Robots, *Programming and Computer Software*, 38 (2012), 13-23.

[28] A. Gorbenko and V. Popov, On the Problem of Placement of Visual Landmarks, *Applied Mathematical Sciences*, 6 (2012), 689-696.

[29] A. Gorbenko, M. Mornev, and V. Popov, Planning a Typical Working Day for Indoor Service Robots, *IAENG International Journal of Computer Science*, 38 (2011), 176-182.

[30] A. Gorbenko and V. Popov, SAT Solvers for the Problem of Sensor Placement, *Advanced Studies in Theoretical Physics*, 6 (2012), 1235-1238.

[31] A. Gorbenko and V. Popov, Clustering Algorithm in Mobile Ad Hoc Networks, *Advanced Studies in Theoretical Physics*, 6 (2012), 1239-1242.

[32] A. Gorbenko and V. Popov, The Problem of Finding Two Edge-Disjoint Hamiltonian Cycles, *Applied Mathematical Sciences*, 6 (2012), 6563-6566.

[33] A. Gorbenko and V. Popov, Hamiltonian Alternating Cycles with Fixed Number of Color Appearances, *Applied Mathematical Sciences*, 6 (2012), 6729-6731.

[34] A. Gorbenko and V. Popov, Footstep Planning for Humanoid Robots, *Applied Mathematical Sciences*, 6 (2012), 6567-6571.

[35] A. Gorbenko and V. Popov, Multiple Occurrences Shortest Common Superstring Problem, *Applied Mathematical Sciences*, 6 (2012), 6573-6576.

[36] A. Gorbenko and V. Popov, The Far From Most String Problem, *Applied Mathematical Sciences*, 6 (2012), 6719-6724.

[37] A. Gorbenko and V. Popov, Multi-agent Path Planning, *Applied Mathematical Sciences*, 6 (2012), 6733-6737.

[38] A. Gorbenko and V. Popov, The c-Fragment Longest Arc-Preserving Common Subsequence Problem, *IAENG International Journal of Computer Science*, 39 (2012), 231-238.

[39] A. Gorbenko and V. Popov, The set of parameterized k-covers problem, *Theoretical Computer Science*, 423 (2012), 19-24.

[40] A. Gorbenko and V. Popov, Task-resource Scheduling Problem, *International Journal of Automation and Computing*, 9 (2012), 429-441.

Received: November 1, 2012