Applied Mathematical Sciences, Vol. 7, 2013, no. 48, 2373 - 2380
HIKARI Ltd, www.m-hikari.com

The Shortest Common Parameterized Supersequence 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 the problem of the shortest common parameterized supersequence. In particular, we consider an explicit reduction from the problem to the satisfiability problem.

Keywords: parameterized supersequence, satisfiability, **NP**-complete

The well-known problem of the shortest common supersequence (SCS) is a classical distance measure for strings. Another well-studied string comparison measure is that of parameterized matching, where two equal-length strings are a parameterized-match if there exists a bijection on the alphabets such that one string matches the other under the bijection (see e.g. [1, 2]). For instance, it is considered the periodicity of parameterized strings. These results and some other studies about the periodicity of parameterized strings showed considerable differences between parameterized strings and ordinary strings.

Nevertheless, binary parameterized strings behave in a very similar way as ordinary strings with respect to repetitions. It is interesting to note that most of the works associated with parameterized pattern matching present polynomial time algorithms. There have been several attempts to accommodate parameterized matching along with other distance measures. In this paper we consider the problem of the shortest common parameterized supersequence (SCPS) which combines the SCS measure with parameterized matching.

A model of parameterized pattern matching was introduced in [3]. The main motivation for this scheme lies in software maintenance, where programs are to be considered "identical" even if variable names are different. Therefore, strings under this model are comprised of symbols from two disjoint sets Σ and Π containing fixed symbols and parameter symbols, respectively. Formally, parameterized pattern matching is as follows (e.g. [4]). A parameterized string is a string over $\Sigma \cup \Pi$. Two parameterized strings S_1 and S_2 of same length are said to parameterized match if there exists a bijection $f : \Pi_1 \to \Pi_2$, where Π_1 and Π_2 are the symbols from Π in S_1 and S_2 respectively, such that the following holds: S_1 (S_2, respectively) equals S_2 (S_1, respectively) when any occurrence $x \in \Pi_1$ (Π_2, respectively) is replaced by $f(x)$ ($f^{-1}(x)$, respectively).

Given two sequences S and T over some fixed alphabet Ξ, the sequence S is a supersequence of T if T can be obtained from S by deleting some letters from S. Notice that the order of the remaining letters of S bases must be preserved. Respectively, T is a subsequence of S if T can be obtained from S by deleting some letters from S. The length of a sequence S is the number of letters in it and is denoted as $|S|$. For simplicity, we use $S[i]$ to denote the ith letter in sequence S.

Given two sequences S and T over some fixed alphabet $\Sigma \cup \Pi$, the sequence S is a parameterized supersequence of T if T parameterized match U where U can be obtained from S by deleting some letters from S. Similarly, given two sequences S and T over some fixed alphabet $\Sigma \cup \Pi$, the sequence T is a parameterized subsequence of S if T parameterized match U where U can be obtained from S by deleting some letters from S.

Given sequences S_1 and S_2 over some fixed alphabet $\Sigma \cup \Pi$, the SCPS problem asks for a shortest sequence T that is a parameterized supersequence of S_1 and S_2. In the decision version SCPS can be formulated as following:

THE SHORTEST COMMON PARAMETERIZED SUPERSEQUENCE PROBLEM (SCPS):

INSTANCE: *Given an alphabet* $\Sigma \cup \Pi$, *sequences* S_1 *and* S_2, *and positive integer* k.

QUESTION: *Is there a sequence* T, $|T| \leq k$, *that is a parameterized supersequence of* S_1 *and* S_2*?*

The algorithmic properties of different problems of finding regularities are thoroughly studied in theoretical computer science (see e.g. [5] – [10]). In particular, encoding different hard problems as instances of different variants of the satisfiability problem and solving them with very efficient satisfiability algorithms has caused considerable interest (see e.g. [11] – [29]). Given sequences S_1 and S_2 over some fixed alphabet $\Sigma \cup \Pi$, the longest common parameterized subsequence (LCPS) problem asks for a longest sequence T that is a parameterized subsequence of S_1 and S_2. In [30] proved that LCPS is **NP**-hard. Note that by inserting the uncommon symbols (taking into account bijections) while preserving the symbol order, we can get a shortest common parameterized supersequence from longest common parameterized subsequence. Therefore, SCPS is **NP**-hard. Since $|S_1| + |S_2| = |T_1| + 2|T_2|$ where T_1 is a shortest common parameterized supersequence of S_1 and S_2 and T_2 is a longest common parameterized subsequence of S_1 and S_2, it is obvious that to solve SCPS we can use some algorithms for LCPS. However, in order to obtain a solution of decision version of SCPS we need multiple runs of a solver for decision version of LCPS. In view of complexity of SCPS, development of a direct solver for SCPS is preferred. In this paper, we consider an explicit reduction from SCPS to the satisfiability problem. Let

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

$$\varphi_{1,1}[i] = \vee_{1\leq j\leq k} x[i,j],$$

$$\varphi_{1,2}[i] = \wedge_{1\leq j[1]<j[2]\leq k}(\neg x[i,j[1]] \vee \neg x[i,j[2]]),$$

$$\varphi_1 = \wedge_{1\leq i\leq |S_1|}(\varphi_{1,1}[i] \wedge \varphi_{1,2}[i]),$$

$$\varphi_2[j] = \wedge_{1\leq i[1]<i[2]\leq |S_1|}(\neg x[i[1],j] \vee \neg x[i[2],j]),$$

$$\varphi_2 = \wedge_{1\leq j\leq k}\varphi_2[j],$$

$$\varphi_3[i,j] = \wedge_{1\leq i[1]\leq |S_1|, 1\leq j[1]\leq k, i[1]>i, j[1]<j}(\neg x[i,j] \vee \neg x[i[1],j[1]]),$$

$$\varphi_3 = \wedge_{1\leq i\leq |S_1|, 1\leq j\leq k}\varphi_3[i,j],$$

$$\varphi_{4,1}[i] = \vee_{1\leq j\leq k} y[i,j],$$

$$\varphi_{4,2}[i] = \wedge_{1\leq j[1]<j[2]\leq k}(\neg y[i,j[1]] \vee \neg y[i,j[2]]),$$

$$\varphi_4 = \wedge_{1\leq i\leq |S_2|}(\varphi_{4,1}[i] \wedge \varphi_{4,2}[i]),$$

$$\varphi_5[j] = \wedge_{1\leq i[1]<i[2]\leq |S_2|}(\neg y[i[1],j] \vee \neg y[i[2],j]),$$

$$\varphi_5 = \wedge_{1\leq j\leq k}\varphi_5[j],$$

$$\varphi_6[i,j] = \wedge_{1\leq i[1]\leq |S_2|, 1\leq j[1]\leq k, i[1]>i, j[1]<j}(\neg y[i,j] \vee \neg y[i[1],j[1]]),$$

$$\varphi_6 = \wedge_{1\leq i\leq |S_2|, 1\leq j\leq k}\varphi_6[i,j],$$

$$\psi_{1,1}[i] = \vee_{1\leq j\leq |\Pi|} u[i,j],$$

$$\psi_{1,1} = \wedge_{1\leq i\leq |S_1|, S_1[i]\in\Pi}\psi_{1,1}[i],$$

$$\psi_{1,2}[i] = \wedge_{1\leq j[1]<j[2]\leq|\Pi|}(\neg u[i,j[1]] \vee \neg u[i,j[2]]),$$

$$\psi_{1,2} = \wedge_{1\leq i\leq|S_1|,S_1[i]\in\Pi}\psi_{1,2}[i],$$

$$\psi_1 = \psi_{1,1} \wedge \psi_{1,2},$$

$$\psi_{2,1}[i[1],i[2],j] = \neg u[i[1],j] \vee u[i[2],j],$$

$$\psi_{2,2}[i[1],i[2],j] = u[i[1],j] \vee \neg u[i[2],j],$$

$$\psi_2[i[1],i[2]] = \wedge_{1\leq j\leq|\Pi|,l\in\{1,2\}}\psi_{2,l}[i[1],i[2],j],$$

$$\psi_2 = \wedge_{1\leq i[1]<i[2]\leq|S_1|,S_1[i[1]]=S_1[i[2]]\in\Pi}\psi_2[i[1],i[2]],$$

$$\psi_3 = \wedge_{1\leq i[1]<i[2]\leq|S_1|,S_1[i[1]]\neq S_1[i[2]],S_1[i[1]],S_1[i[2]]\in\Pi,1\leq j\leq|\Pi|}(\neg u[i[1],j] \vee \neg u[i[2],j]),$$

$$\psi_{4,1} = \wedge_{1\leq i\leq|S_2|,S_2[i]\in\Pi} \vee_{1\leq j\leq|\Pi|} v[i,j],$$

$$\psi_{4,2} = \wedge_{1\leq i\leq|S_2|,S_2[i]\in\Pi,1\leq j[1]<j[2]\leq|\Pi|}(\neg v[i,j[1]] \vee \neg v[i,j[2]]),$$

$$\psi_4 = \psi_{4,1} \wedge \psi_{4,2},$$

$$\psi_{5,1}[i[1],i[2],j] = \neg v[i[1],j] \vee v[i[2],j],$$

$$\psi_{5,2}[i[1],i[2],j] = v[i[1],j] \vee \neg v[i[2],j],$$

$$\psi_5[i[1],i[2]] = \wedge_{1\leq j\leq|\Pi|l\in\{1,2\}}\psi_{5,l}[i[1],i[2],j],$$

$$\psi_5 = \wedge_{1\leq i[1]<i[2]\leq|S_1|,S_2[i[1]]=S_2[i[2]]\in\Pi}\psi_5[i[1],i[2]],$$

$$\psi_6 = \wedge_{1\leq i[1]<i[2]\leq|S_2|,S_2[i[1]]\neq S_2[i[2]],S_2[i[1]],S_2[i[2]]\in\Pi,1\leq j\leq|\Pi|}(\neg v[i[1],j] \vee \neg v[i[2],j]),$$

$$\rho_1[i[1],i[2],j] = \wedge_{1\leq l\leq|\Pi|}(\neg x[i[1],j] \vee \neg y[i[2],j] \vee \neg u[i[1],l] \vee \neg v[i[2],l]),$$

$$\rho_1 = \wedge_{1\leq i[1]\leq|S_1|,1\leq i[2]\leq|S_2|,1\leq j\leq k,S_1[i[1]],S_2[i[2]]\in\Pi}\rho_1[i[1],i[2],j],$$

$$\rho_2 = \wedge_{1\leq i[1]\leq|S_1|,1\leq i[2]\leq|S_2|,1\leq j\leq k,S_1[i[1]]\in\Sigma,S_2[i[2]]\in\Pi}(\neg x[i[1],j] \vee \neg y[i[2],j]),$$

$$\rho_3 = \wedge_{1\leq i[1]\leq|S_1|,1\leq i[2]\leq|S_2|,1\leq j\leq k,S_1[i[1]]\in\Pi,S_2[i[2]]\in\Sigma}(\neg x[i[1],j] \vee \neg y[i[2],j]),$$

$$\rho_4[j] = \wedge_{1\leq j\leq k}(\neg x[i[1],j] \vee \neg y[i[2],j]),$$

$$\rho_4 = \wedge_{1\leq i[1]\leq|S_1|,1\leq i[2]\leq|S_2|,S_1[i[1]]\in\Sigma,S_2[i[2]]\in\Sigma,S_1[i[1]]\neq S_2[i[2]]}\rho_4[j],$$

$$\xi = (\wedge_{i=1}^{6}\varphi_i) \wedge (\wedge_{i=1}^{6}\psi_i) \wedge (\wedge_{i=1}^{4}\rho_i).$$

Theorem. *Given a fixed alphabet $\Sigma\cup\Pi$, sequences S_1 and S_2, and positive integer k. There is a sequence T, $|T| \leq k$, that is a parameterized supersequence of S_1 and S_2 if and only if ξ is satisfiable.*

Proof. Given a fixed alphabet $\Sigma \cup \Pi$, sequences S_1 and S_2, and positive integer k. Suppose that there is a sequence T, $|T| \leq k$, that is a parameterized supersequence of S_1 and S_2. Without loss of generality we can assume that $|T| = k$. Let $x[i,j] = 1$ where $1 \leq i \leq |S_1|$, $1 \leq j \leq k$, and image of $S_1[i]$ is located in T at position j. Respectively, let $y[i,j] = 1$ where $1 \leq i \leq |S_2|$, $1 \leq j \leq k$, and image of $S_2[i]$ is located in T at position j.

Since T is a parameterized supersequence of S_1 and S_2, there are bijections f_1 and f_2 which transform S_1 and S_2 into sequences U_1 and U_2 such that T is a supersequence of U_1 and U_2. Let $u[i,j]=1$ where $1\le i\le |S_1|$, $S_1[i]\in\Pi$, $1\le j\le|\Pi|$, and $f_1(S_1[i])=b_j$. Respectively, let $v[i,j]=1$ where $1\le i\le|S_2|$, $S_2[i]\in\Pi$, $1\le j\le|\Pi|$, and $f_2(S_2[i])=b_j$.

Assume that all other variables are equal to 0.

It is easy to see that $\varphi_{1,1}[i]=1$ if and only if there is j such that $x[i,j]=1$. Since T is a parameterized supersequence of S_1, there is j such that image of $S_1[i]$ is located in T at position j. Therefore, by definition of $x[i,j]$, $\varphi_{1,1}[i]=1$ for all i. Clearly, we can suppose that there is only one value of j such that image of $S_1[i]$ is located in T at position j. Note that $\varphi_{1,2}[i]=1$ if and only if for given value of i there is no more then one value of j such that $x[i,j]=1$. Thus, $\varphi_{1,2}[i]=1$ for any i. So, $\varphi_1=1$. By definition of supersequence, if $i[1]\neq i[2]$, then images of $S_1[i[1]]$ and $S_1[i[2]]$ are located in T at different positions. Thus, by definition of $x[i,j]$, for any j there is no more then one value of i such that $x[i,j]=1$. Therefore, $\varphi_2[j]=1$ for all j. So, $\varphi_2=1$. Suppose that $T[j]$ is the image of $S_1[i]$ and $T[j[1]]$ is the image of $S_1[i[1]]$. By definition of supersequence, if $i[1]>i$, then $j[1]>j$. Thus, by definition of $x[i,j]$, it is easy to see that if $x[i,j]=1$, then $x[i[1],j[1]]=0$ for any i, j, $i[1]$, $j[1]$ such that $i[1]>i$ and $j[1]<j$. Therefore, $\varphi_3[i,j]=1$ for all i and j. So, $\varphi_3=1$. Similarly, we can show that $\varphi_4=\varphi_5=\varphi_6=1$. Using the same arguments and definitions of $u[i,j]$ and $v[i,j]$, we can check that $\wedge_{i=1}^{6}\psi_i=1$. Also it is easy to verify that satisfiability of $\wedge_{i=1}^{4}\rho_i$ follows from definition of f and definitions of $x[i,j]$, $y[i,j]$, $u[i,j]$, and $v[i,j]$. Therefore, $\xi=1$.

Now suppose that $\xi=1$. Since $\varphi_1=1$, it is easy to see that, for all i, there is only one value of j such that $x[i,j]=1$. Similarly, in view of $\varphi_4=1$, it is clear that, for all i, there is only one value of j such that $y[i,j]=1$. So, we can consider values of $x[i,j]$ and $y[i,j]$ as a definition of positions of $S_1[i]$ and $S_2[i]$ in T. In view of $\varphi_2=1$ and $\varphi_5=1$, each letter has a unique position. Since $\varphi_3=1$ and $\varphi_6=1$, it is easy to check that order of letters is preserved.

Similarly, we can consider $\wedge_{i=1}^{6}\psi_i$ as a definition of bijections f_1 and f_2. Using this assumption by direct verification we can check that T is a parameterized supersequence of S_1 and S_2. □

Clearly, ξ is a CNF. So, ξ give us an explicit reduction from SCPS to SAT.

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

References

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

[2] A. Gorbenko and V. Popov, Longest Common Parameterized Subsequences with Fixed Common Substring, *Applied Mathematical Sciences*, 7 (2013), 645-650.

[3] B. S. Baker, Parameterized Pattern Matching: Algorithms and Applications, *Journal of Computer and System Sciences*, 52 (1996), 28-42.

[4] A. Amir, M. Farach, and S. Muthukrishnan, Alphabet Dependence in Parameterized Matching, *Information Processing Letters*, 49 (1994), 111-115.

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

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

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

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

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

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

[11] A. Gorbenko and V. Popov, The Minimum Test Collection Problem, *Applied Mathematical Sciences*, 7 (2013), 1191-1193.

[12] A. Gorbenko and V. Popov, The Farthest Substring Problem, *Applied Mathematical Sciences*, 7 (2013), 1209-1212.

[13] A. Gorbenko and V. Popov, Computational Experiments for the Problem of Hamiltonian Path with Fixed Number of Color Repetitions, *Advanced Studies in Theoretical Physics*, 7 (2013), 121-126.

[14] A. Gorbenko and V. Popov, On Hamilton Paths in Grid Graphs, *Advanced Studies in Theoretical Physics*, 7 (2013), 127-130.

[15] A. Gorbenko and V. Popov, Computational Experiments for the Problem of Sensor-Mission Assignment in Wireless Sensor Networks, *Advanced Studies in Theoretical Physics*, 7 (2013), 135-139.

[16] A. Gorbenko and V. Popov, The Swap Common Superstring Problem, *Applied Mathematical Sciences*, 7 (2013), 609-614.

[17] A. Gorbenko and V. Popov, The String Barcoding Problem, *Applied Mathematical Sciences*, 7 (2013), 615-622.

[18] A. Gorbenko and V. Popov, On Multiple Occurrences Shortest Common Superstring Problem, *Applied Mathematical Sciences*, 7 (2013), 641-644.

[19] A. Gorbenko and V. Popov, The Hamiltonian Alternating Path Problem, *IAENG International Journal of Applied Mathematics*, 42 (2012), 204-213.

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

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

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

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

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

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

[26] V. Popov, Partially Distinguishable Guards, *Applied Mathematical Sciences*, 6 (2012), 6587-6591.

[27] A. Gorbenko and V. Popov, The Problem of Selection of a Set of Partially Distinguishable Guards, *Applied Mathematical Sciences*, 7 (2013), 651-654.

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

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

[30] O. Keller, T. Kopelowitz, and M. Lewenstein, On the Longest Common Parameterized Subsequence, *Theoretical Computer Science*, 410 (2009), 5347-5353.

Received: February 12, 2013