Advanced Studies in Biology, Vol. 4, 2012, no. 8, 373 - 380

The Longest Common Subsequence 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

Algorithms on sequences of symbols have been studied for a long time and now form a fundamental part of computer science. One of the very important problems in analysis of sequences is the longest common subsequence problem. For the general case of an arbitrary number of input sequences, the problem is **NP**-hard. We describe an approach to solve this problem. This approach is based on constructing a logical model for the longest common subsequence problem.

Keywords: the longest common subsequence problem, logical model, satisfiability problem, **NP**-complete

1 Introduction

Algorithms on sequences of symbols have been studied for a long time and now form a fundamental part of computer science. One of the very important problems in analysis of sequences is the longest common subsequence (LCS) problem. The LCS problem is to find the longest subsequence common to all sequences in a set of sequences (often just two).

The computational problem of finding the longest common subsequence of a set of k strings has been studied extensively over the last thirty years (see [1, 2, 3] and references). This problem has many applications. When $k = 2$, the longest common subsequence is a measure of the similarity of two strings and is thus useful in molecular biology, pattern recognition, and text compression [4, 5, 6]. The version of LCS in which the number of strings is unrestricted is also useful in text compression [5], and is a special case of the multiple sequence alignment and consensus subsequence discovery problem in molecular biology [7, 8, 9].

The k-unrestricted LCS problem is **NP**-complete [5]. If the number of sequences is fixed at k with maximum length n, their longest common subsequence can be found in $O(n^{k-1})$ time, through an extension of the pairwise algorithm [3]. Suppose $|S_1| = n$ and $|S_2| = m$, the longest common subsequence of S_1 and S_2 can be found in time $O(nm)$ [10, 11, 12].

2 Problem Definition

Given two sequences S and T over some fixed alphabet Σ, the sequence T is a subsequence of S 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. 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, and $S[i, j]$ to denote the substring of S consisting of the ith letter through the jth letter.

Given sequences S_i, $1 \leq i \leq m$, over some fixed alphabet Σ, the LCS problem asks for a longest sequence T that is a subsequence of S_i for all $i \in \{1, 2, \ldots, m\}$.

In the decision version LCS can be formulated as following.

LCS:

INSTANCE: *A fixed alphabet Σ, sequences S_i, $1 \leq i \leq m$, and positive integer k.*

QUESTION: *Is there a sequence T, $|T| \geq k$, that is a subsequence of S_i for all $i \in \{1, 2, \ldots, m\}$?*

3 Logical model

The propositional satisfiability problem (PSAT) is a core problem in mathematical logic and computing theory. Propositional satisfiability is the problem of determining if the variables of a given boolean function can be assigned in such a way as to make the formula evaluate to true. PSAT was the first known **NP**-complete problem, as proved by Stephen Cook in 1971.

Encoding problems as Boolean satisfiability and solving them with very efficient satisfiability algorithms has recently caused considerable interest. From a practical point of view, to further investigations of the LCS of interest to find explicit reductions to the satisfiability problem. Such reduction will provide a basis for construction of an effective intelligent algorithm for LCS. In this paper we consider reductions from LCS to PSAT.

Let

$$\varphi[i,j] = \Big(\bigvee_{1 \le l \le |\Sigma|} x[i,j,l]\Big) \wedge \Big(\bigwedge_{\substack{1 \le l[1] \le |\Sigma| \\ 1 \le l[2] \le |\Sigma| \\ l[1] \ne l[2]}} (\neg x[i,j,l[1]] \vee \neg x[i,j,l[2]])\Big),$$

$$\varphi = \bigwedge_{\substack{1 \le i \le m \\ 1 \le j \le |S_i|}} \varphi[i,j],$$

$$\psi[1,i,j] = \bigwedge_{\substack{1 \le l[1] \le k \\ 1 \le l[2] \le k \\ l[1] \ne l[2]}} (\neg y[i,j,l[1]] \vee \neg y[i,j,l[2]]),$$

$$\psi[1] = \bigwedge_{\substack{1 \le i \le m \\ 1 \le j \le |S_i|}} \psi[1,i,j],$$

$$\psi[2,i,j,l] = \bigwedge_{\substack{1 \le j[1] \le |S_i| \\ j[1] \ne j}} (\neg y[i,j,l] \vee \neg y[i,j[1],l]),$$

$$\psi[2] = \bigwedge_{\substack{1 \le i \le m \\ 1 \le j \le |S_i| \\ 1 \le l \le k}} \psi[2,i,j,l],$$

$$\psi[3,i,j,l] = \bigwedge_{\substack{1 \le j[1] \le |S_i| \\ 1 \le l[1] \le k \\ j[1] < j \\ l[1] \ge l}} (\neg y[i,j,l] \vee \neg y[i,j[1],l[1]]),$$

$$\psi[3] = \bigwedge_{\substack{1 \le i \le m \\ 1 \le j \le |S_i| \\ 1 \le l \le k}} \psi[3,i,j,l],$$

$$\psi[4,i,l] = \bigvee_{1 \le j \le |S_i|} y[i,j,l],$$

$$\psi[4] = \bigwedge_{\substack{1 \le i \le m \\ 1 \le l \le k}} \psi[4,i,l],$$

$$\xi[i[1], i[2], j[1], j[2], p] = \bigwedge_{1 \leq l \leq |\Sigma|} (y[i[1], j[1], p] \wedge y[i[2], j[2], p]) \rightarrow$$

$$x[i[1], j[1], l] = x[i[2], j[2], l],$$

$$\xi = \bigwedge_{\substack{1 \leq i[1] \leq m \\ 1 \leq i[2] \leq m \\ i[1] \neq i[2] \\ 1 \leq j[1] \leq |S_{i[1]}| \\ 1 \leq j[2] \leq |S_{i[2]}| \\ 1 \leq p \leq k}} \xi[i[1], i[2], j[1], j[2], p],$$

$$\tau = \varphi \wedge (\bigwedge_{1 \leq q \leq 4} \psi[q]) \wedge \xi.$$

Theorem. *Given sequences S_i, $1 \leq i \leq m$, over some fixed alphabet Σ, and positive integer k. There is a sequence T, $|T| \geq k$, that is a subsequence of S_i for all $i \in \{1, 2, \dots, m\}$, if and only if τ is satisfiable.*

Proof. Suppose that there is a sequence T, $|T| \geq k$, that is a subsequence of S_i for all $i \in \{1, 2, \dots, m\}$. Clearly, we can assume that $|T| = k$. In this case,

$$T = S_1[j[1,1]]S_1[j[1,2]] \dots S_1[j[1,k]] =$$

$$S_2[j[2,1]]S_2[j[2,2]] \dots S_2[j[2,k]] =$$

$$\dots$$

$$S_m[j[m,1]]S_m[j[m,2]] \dots S_m[j[m,k]]$$

for some

$$1 \leq j[1,1] < j[1,2] < \dots < j[1,k] \leq |S_1|,$$

$$1 \leq j[2,1] < j[2,2] < \dots < j[2,k] \leq |S_2|,$$

$$\dots$$

$$1 \leq j[m,1] < j[m,2] < \dots < j[m,k] \leq |S_m|.$$

Let $\Sigma = \{a_1, a_2, \dots, a_p\}$; $x[i,j,l] = 1$ where $1 \leq i \leq m$, $1 \leq j \leq |S_i|$, $1 \leq l \leq |\Sigma|$, $S_i[j] = a_l$; $x[i,j,l] = 0$ where $1 \leq i \leq m$, $1 \leq j \leq |S_i|$, $1 \leq l \leq |\Sigma|$, $S_i[j] \neq a_l$; $y[i,j,l] = 0$ where $1 \leq i \leq m$, $1 \leq j \leq |S_i|$, $1 \leq l \leq k$, $j \notin \{j[i,1], j[i,2], \dots, j[i,k]\}$; $y[i,j,l] = 0$ where $1 \leq i \leq m$, $1 \leq j \leq |S_i|$, $1 \leq l \leq k$, $j = j[i,r]$, $l \neq r$; $y[i,j,l] = 1$ where $1 \leq i \leq m$, $1 \leq j \leq |S_i|$, $1 \leq l \leq k$, $j = j[i,r]$, $l = r$.

Suppose that $S_i[j] = a_{l_0}$. In this case, $x[i,j,l_0] = 1$. Therefore,

$$\bigvee_{1 \leq l \leq |\Sigma|} x[i,j,l] = 1. \quad (1)$$

Since $S_i[j] = a_{l_0}$, it is easy to see that $x[i,j,l] = 0$ where $i \neq l_0$. In view of $l[1] \neq l[2]$, it is clear that $\neg x[i,j,l[1]] \vee \neg x[i,j,l[2]] = 1$ where $1 \leq l[1] \leq |\Sigma|$, $1 \leq l[2] \leq |\Sigma|$, $l[1] \neq l[2]$. In view of (1), hence we get $\varphi[i,j] = 1$. Therefore, by arbitrariness of i and j we obtain that $\varphi = 1$.

If $j \notin \{j[i,1], j[i,2], \ldots, j[i,k]\}$, then $y[i,j,l] = 0$ for all l. Hence $\neg y[i,j,l] = 1$. So, $\psi[1,i,j] = 1$. If $j = j[i,r]$ and $l \neq r$, then $y[i,j,l] = 0$. In view of $l[1] \neq l[2]$, we obtain $\neg y[i,j,l[1]] \vee \neg y[i,j,l[2]] = 1$. Therefore, $\psi[1] = 1$.

Consider $\psi[2,i,j,l]$. If $y[i,j,l] = 0$ or $y[i,j[1],l] = 0$, then $\psi[2,i,j,l] = 1$. Suppose that $y[i,j,l] = 1$ and $y[i,j[1],l] = 1$. By definition, if $y[i,j,l] = 1$, then $j = j[i,r]$, $l = r$. Similarly, $j[1] = j[i,r]$, $l = r$. Hence $j = j[1]$. Therefore, $\psi[2] = 1$.

Consider $\psi[3,i,j,l]$. If $y[i,j,l] = 0$ or $y[i,j[1],l[1]] = 0$, then $\psi[3,i,j,l] = 1$. Suppose that $y[i,j,l] = 1$ and $y[i,j[1],l[1]] = 1$. By definition, if $y[i,j,l] = 1$, then $j = j[i,r]$, $l = r$. Similarly, $j[1] = j[i,r[1]]$, $l[1] = r[1]$. Since $j[1] < j$, $j[i,r[1]] < j[i,r]$. Hence $r[1] < r$. By definition, $l[1] \geq l$. In view of $l = r$ and $l[1] = r[1]$, we obtain $r[1] \geq r$. Therefore, $\psi[3] = 1$.

Note that $y[i,j[i,l],l] = 1$ for all $1 \leq i \leq m$, $1 \leq l \leq k$. Therefore, $\psi[4] = 1$.

Consider ξ. If $y[i[1],j[1],p] = 0$ or $y[i[2],j[2],p] = 0$, then

$$((y[i[1],j[1],p] \wedge y[i[2],j[2],p]) \rightarrow x[i[1],j[1],l] = x[i[2],j[2],l]) = 1.$$

Suppose that $y[i[1],j[1],p] = 1$ and $y[i[2],j[2],p] = 1$ where $i[1] \neq i[2]$. By definition, $j[1] = j[i[1],r[1]]$, $p = r[1]$, $j[2] = j[i[2],r[2]]$, $p = r[2]$. Note that

$$S_{i[1]}[j[i[1],1]]S_{i[1]}[j[i[1],2]] \ldots S_{i[1]}[j[i[1],k]] =$$

$$S_{i[2]}[j[i[2],1]]S_{i[2]}[j[i[2],2]] \ldots S_{i[2]}[j[i[2],k]].$$

Hence $S_{i[1]}[j[i[1],p]] = S_{i[2]}[j[i[2],p]]$. By definition, this implies $x[i[1],j[1],l] = x[i[2],j[2],l]$ for all $1 \leq l \leq |\Sigma|$. Therefore, $\xi = 1$.

Since $\varphi = \psi[1] = \psi[2] = \psi[3] = \psi[4] = \xi = 1$, it is clear that $\tau = 1$.

Now suppose that $\tau = 1$. By definition, $\varphi = \psi[1] = \psi[2] = \psi[3] = \psi[4] = \xi = 1$. Since $\varphi = 1$, it is easy to see that $\varphi[i,j] = 1$ for all $1 \leq i \leq m$, $1 \leq j \leq |S_i|$. Hence $\neg x[i,j,l[1]] \vee \neg x[i,j,l[2]] = 1$ for all $1 \leq l[1] \leq |\Sigma|$, $1 \leq l[2] \leq |\Sigma|$ such that $l[1] \neq l[2]$. Therefore, for given i and j there is no more than one value of l such that $x[i,j,l] = 1$. In view of $\varphi[i,j] = 1$, we have (1). Hence there exists at least one value of l such that $x[i,j,l] = 1$. Therefore, we can assume that $S_i[j] = a_l$ where $1 \leq i \leq m$, $1 \leq j \leq |S_i|$, $1 \leq l \leq |\Sigma|$, $x[i,j,l] = 1$. Since $\psi[1] = 1$, it is easy to see that $\psi[1,i,j] = 1$ for all $1 \leq i \leq m$, $1 \leq j \leq |S_i|$. Hence $\neg y[i,j,l[1]] \vee \neg y[i,j,l[2]] = 1$ for all $1 \leq l[1] \leq k$, $1 \leq l[2] \leq k$ such that $l[1] \neq l[2]$. Therefore, for given i and j there is no more than one value of l such that $y[i,j,l] = 1$. Since $\psi[2] = 1$, it is easy to see that $\psi[2,i,j,l] = 1$ for all $1 \leq i \leq m$, $1 \leq j \leq |S_i|$, $1 \leq l \leq k$.

Hence $\neg y[i,j,l] \vee \neg y[i,j[1],l] = 1$ for all $1 \leq j[1] \leq |S_i|$ such that $j[1] \neq j$. Therefore, for given i, j, and l either $y[i,j,l] = 0$ or

$$y[i,j,l] = 1 \wedge (\bigwedge_{j[1] \neq j} y[i,j[1],l] = 0). \tag{2}$$

Since $\psi[3] = 1$, it is easy to see that $\psi[3,i,j,l] = 1$ for all $1 \leq i \leq m$, $1 \leq j \leq |S_i|$, $1 \leq l \leq k$. Hence $\neg y[i,j,l] \vee \neg y[i,j[1],l[1]] = 1$ for all $1 \leq j[1] \leq |S_i|$, $1 \leq l[1] \leq k$ such that $j[1] < j$ and $l[1] \geq l$. Therefore, for given i, j, and l either $y[i,j,l] = 0$ or

$$y[i,j,l] = 1 \wedge (\bigwedge_{\substack{j[1] < j \\ l[1] \geq l}} y[i,j[1],l[1]] = 0). \tag{3}$$

Since $\psi[4] = 1$, it is easy to see that $\psi[4,i,l] = 1$ for all $1 \leq i \leq m$, $1 \leq l \leq k$. Hence for given i and l there exists at least one value of j such that $y[i,j,l] = 1$. Since for given i and j there is no more than one value of l such that $y[i,j,l] = 1$, we can suppose that

$$y[i,j[i,1],l[i,1]] = 1, y[i,j[i,2],l[i,2]] = 1, \ldots, y[i,j[i,p[i]],l[i,p[i]]] = 1 \tag{4}$$

and $y[i,j,l] = 0$ for given i where $1 \leq p[i] \leq |S_i|$,

$$(i,l) \notin \{([i,1],l[i,1]),(j[i,2],l[i,2]),\ldots,(j[i,p[i]],l[i,p[i]])\}.$$

Since for given i and l there exists at least one value of j such that $y[i,j,l] = 1$, we can suppose that $p[i] \geq k$. In view of (2), it is clear that $p[i] \leq k$. So, $p[i] = k$. Without loss of generality, we can suppose that

$$j[i,1] < j[i,2] < \ldots < j[i,k].$$

In view of (3), it is easy to see that if $j[i,t_1] < j[i,t_2]$, then $l[i,t_1] < l[i,t_2]$. Therefore, (4) can be represented as follows:

$$y[i,j[i,1],1] = 1, y[i,j[i,2],2] = 1, \ldots, y[i,j[i,k],k] = 1.$$

Let

$$T = S_1[j[1,1]]S_1[j[1,2]]\ldots S_1[j[1,k]].$$

Since $\xi = 1$, it is easy to see that

$$S_1[j[1,1]]S_1[j[1,2]]\ldots S_1[j[1,k]] =$$

$$S_2[j[2,1]]S_2[j[2,2]]\ldots S_2[j[2,k]] =$$

$$\ldots$$

$$S_m[j[m,1]]S_m[j[m,2]]\ldots S_m[j[m,k]].$$

□

So, in view of theorem, we obtain an explicit reduction from LCS to PSAT.

4 Conclusion

In papers [13, 14, 15, 16, 17, 18, 19] the authors considered some algorithms to solve logical models. Our computational experiments have shown that these algorithms can be used to solve the logical model for LCS.

References

[1] H. Bodlaender, R. Downey, M. Fellows, and H. Wareham, The parameterized complexity of sequence alignment and consensus, *Theoretical Computer science*, 147 (1995), 31-54.

[2] D. Hirschberg, Recent results on the complexity of common subsequence problems, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, (1983), 325-330.

[3] R. Irving and C. Fraser, Two algorithms for the longest common subsequence of three (or more) strings, *Proceedings of the Third Annual Symposium on Combinatorial Pattern Matching*, (1992), 214-229.

[4] S. Lu and K. Fu, A sentence-to-sentence clustering procedure for pattern analysis, *Transactions on Systems, Man, and Cybernetics*, 8 (1978), 381-389.

[5] D. Maier, The complexity of some problems on subsequences and supersequences, *Journal of the ACM*, 25 (1978), 322-336.

[6] D. Sankoff, Matching comparisons under deletion/insertion constraints, *Proceedings of the National Academy of Sciences of the United States of America*, 69 (1972), 4-6.

[7] W. Day and F. McMorris, Discovering consensus molecular sequences, *Information and Classification – Concepts, Methods, and Applications*, (1993), 393-402.

[8] W. Day and F. McMorris, The computation of consensus patterns in DNA sequences, *Mathematical and Computer Modelling*, 17 (1993), 49-52.

[9] P. Pevzner, Multiple alignment, communication cost, and graph matching, *SIAM Journal on Applied Mathematics*, 52 (1992), 1763-1779.

[10] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, Introduction to Algorithms, Third edition. The MIT Press, Cambridge, Massachusetts, 2009.

[11] D. Hirschberg, The longest common subsequence problem, PhD Thesis, Princeton University, Princeton, 1975.

[12] R. Wagner and M. Fischer, The string-to-string correction problem, *Journal of the ACM*, 21 (1974), 168-173.

[13] 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.

[14] 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.

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

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

[17] A. Gorbenko and V. Popov, On the Optimal Reconfiguration Planning for Modular Self-Reconfigurable DNA Nanomechanical Robots, *Advanced Studies in Biology*, 4 (2012), 95-101.

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

[19] A. Gorbenko, V. Popov, and A. Sheka, Localization on Discrete Grid Graphs, *Proceedings of the CICA 2011*, (2012), 971-978.

Received: May, 2012