Journal of Algorithms **34,** 194–201 (2000)
Article ID jagm.1999.1050, available online at http://www.idealibrary.com on IDEAL®

PROBLEMS

Samir Khuller

Problems Section Editor

This column will carry problems arising in the design of algorithms for discrete optimization problems.

Problems are solicited in all areas of algorithm design that are covered by the *Journal of Algorithms*. Some of the problems that appear here may also appear in the "Open Problems" column in *SIGACT News*. Problems should be submitted to me at the Department of Computer Science, University of Maryland, College Park, MD 20742 (E-mail: samir@cs.umd.edu), and if chosen, will appear in this column. Problem submissions should be precise and succinct. Proposals for guest columns, focusing on problems in a specific research area, are also welcome.

Especially welcome are algorithmic problems arising in areas not previously explored by the theoretical computer science community.

This column is a guest column prepared by Tao Jiang, Paul Kearney, and Ming Li.

Some Open Problems in Computational Molecular Biology

Tao Jiang

Department of Computer Science, University of California, Riverside, California 92521
E-mail: jiang@cs.ucr.edu

and

Paul Kearney and Ming Li

Department of Computer Science, University of Waterloo, Ontario N2L 3G1, Canada
E-mail: pkearney@phylogeny.uwaterloo.ca, mli@math.uwaterloo.ca

Computational molecular biology has emerged as one of the most exciting interdisciplinary fields in recent years, riding on the success of the ongoing *Human Genome Project*. The field has not only benefited from the

0196-6774/00 $35.00

many concepts and techniques developed in theoretical computer science, but it has also provided many interesting research problems. In this column, we include five open problems that are concerned with the design of efficient algorithms and computational complexity. Some of these problems have existed in the literature for a while but most are relatively new. The topics covered span several main branches of computational molecular biology such as sequence comparison, the reconstruction of evolutionary trees, physical mapping, and genetic drug target search.

To save space, for each problem, we will give only the necessary mathematical definitions and a brief definition of some relevant existing results. The reader is referred to appropriate references for more detailed information about the problems such as their background, motivation, and relation to other (solved or unsolved) problems. For a general treatment of algorithmic issues in computational molecular biology, see [9, 21, 18].

1. MULTIPLE SEQUENCE ALIGNMENT WITH SP-SCORE

A *sequence* is a string over some finite alphabet Σ. For DNA sequences, the alphabet Σ contains four letters A, C, G, and T representing four distinct nucleotides, and for protein sequences, Σ contains 20 letters, each representing a unique amino acid. Let S be a set of sequences. A *multiple alignment* $\mathscr{A}$ for S is a two-dimensional matrix obtained as follows: spaces are inserted into each sequence to form a row of the matrix so that the resulting sequences have the same length. Figure 1 illustrates a multiple alignment of four sequences.

The *cost* of a multiple alignment $\mathscr{A}$ is defined as follows. A cost function is defined for each column of letters and the cost of $\mathscr{A}$ is the sum of all column costs. An *optimal alignment* for S is one that *minimizes* the cost over all possible multiple alignments for S. The multiple sequence alignment problem in general is to look for an optimal alignment for a given set of sequences.

$$S = \{TGTTTTAA,\ GTAATTA,\ TGCAATTAG,\ TGTAACTTAG\}$$

$$\begin{array}{c} TGT\ \ \ TTTAA \\ GTAA\ \ TTA \\ TGCAA\ \ TTAG \\ TGTAACTTAG \end{array}$$

FIGURE 1

Among the many possible column cost functions, a very popular one is called *sum-of-all-pairs cost* or simply *SP-cost*. Multiple sequence alignment with SP-cost has been studied extensively in the literature. To define SP-cost, let Δ denote a space. For convenience, we will view a space as a special letter. Suppose that s is a function from $(\Sigma \cup \{\Delta\}) \times (\Sigma \cup \{\Delta\})$ to the real numbers. A standard assumption about s is that it is symmetric and forms a metric. That is, it satisfies the following conditions:

1. for any letters a and b, $s(a, b) = s(b, a)$,
2. for any letter a, $s(a, a) = 0$, and
3. (*triangle inequality*) for any three letters a, b, and c, $s(a, b) \leq s(a, c) + s(c, b)$.

Then the SP-cost of a column (with respect to a fixed function s) consisting of letters $a_1, a_2, \ldots, a_k$, is the summation $\sum_{i,j} s(a_i, a_j)$. Hence, we can pose the problem of multiple sequence alignment with SP-cost as the following.

MULTIPLE SEQUENCE ALIGNMENT WITH SP-COST

Instance: Set of sequences $S = \{s_1, s_2, \ldots, s_k\}$.

Goal: Find a multiple alignment for S with the minimum SP-cost.

It is known that the above problem is NP-hard [4, 20]. Gusfield first proposed a polynomial time approximation algorithm for this problem that achieves ratio $2 - \frac{2}{k}$ on k input sequences [8]. Pevzner improved Gusfield's algorithm to obtain a ratio of $2 - \frac{3}{k}$ [17]. Bafna, Lawler, and Pevzner pushed the ratio further to $2 - \frac{l}{k}$ [1] for any fixed l.

Open problem 1. Is multiple sequence alignment with SP-cost MAX SNP-hard for some symmetric and metric cost function s? Is it possible to improve the approximation ratio to some constant smaller than 2?

We note that recently, Just [13] proved that multiple sequence alignment with SP-cost is MAX SNP-hard for some simple nonmetric function s that satisfies the first two conditions but not triangle inequality.

2. ORDINAL REPRESENTATION

An *evolutionary tree* is a tree where the leaves are bijectively labeled by a set S of sequences, each edge e is assigned a positive length $w(e)$, and every internal vertex has degree at least three. In an evolutionary tree the length of an edge is proportional to the number of mutations that have occurred along that edge. Define $P_T(x, y)$ to be the path from sequence x

to sequence y in T and define $d_T(x, y)$ to be the sum of the edge lengths along $P_T(x, y)$. An evolutionary tree T is called *unweighted* if each edge is assigned unit length.

The distance matrix D_T of an evolutionary tree T is defined such that $D_T(x, y) = d_T(x, y)$ for all pairs of sequences labeling T. The distance matrix D_T of an evolutionary tree T is unique to T [10, 22]. In the *distance method* approach to reconstructing evolutionary trees from sequences, a matrix M is obtained by estimating the number of mutations along the evolutionary path between every pair of sequences. M is then an estimate of D_T and is used to reconstruct T.

Due to several factors including multiple mutations at the same sequence site,[1] $M(x, y)$ is often a poor estimate of $D_T(x, y)$ in the absolute sense. However, ordinal relationships among values in D_T are typically preserved by M. In particular, define M and D_T to be *ordinally equivalent* if for all sequences a, b, c, and d,

$$D_T(a, b) \leq D_T(c, d) \qquad \text{if and only if } M(a, b) \leq M(c, d).$$

Open problem 2. Is there a polynomial time algorithm for finding an evolutionary tree T such that the distance matrix of T is ordinally equivalent to a given matrix M?

If such an evolutionary tree exists then it is called an *ordinal representation* of M. Ordinal representations were introduced by Kannan and Warnow [14]. The problem of finding an unweighted evolutionary tree ordinally equivalent to a given matrix M can be solved in $O(n^2 \log^2 n)$ time [15]. This has recently been improved to $O(n^2)$ [23]. It is also known that the unweighted ordinal representation of a matrix M is unique [15].

3. MINIMUM QUARTET INCONSISTENCY

The evolutionary tree reconstruction problem is to find an optimal evolutionary tree, according to some criterion, for a given set of (DNA or protein) sequences. Please see the preceding section for the definition of an evolutionary tree. In recent years *quartet methods* for reconstructing evolutionary trees have received much attention in the computational biology community.

Given a quartet of species $\{a, b, c, d\}$ and an evolutionary tree T, the *quartet topology* induced in T by $\{a, b, c, d\}$ is the path structure connecting a, b, c, and d in T. For a quartet $\{a, b, c, d\}$, if the path in T connecting

[1] A mutation at a site is effectively hidden by a subsequent mutation.

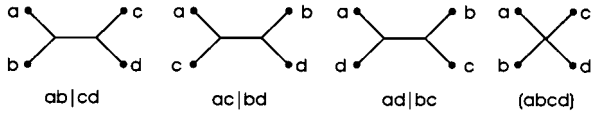

FIGURE 2

labels a and b is disjoint from the path in T connecting c and d, the quartet is said to be *resolved* and is denoted $ab|cd$. Otherwise, the quartet is said to be *unresolved* and is denoted $(abcd)$. The four possible quartet topologies induced by a quartet are depicted in Fig. 2. In the following we consider only binary trees and thus resolved quartets.

Quartet methods proceed by first inferring the quartet topology induced by each quartet and then recombining these quartet topologies to form an estimate of the true evolutionary tree. This approach is based upon the fact that an evolutionary tree T is uniquely characterized by its set Q_T of induced quartet topologies [5] (see Fig. 3).

The computational interest in this paradigm derives from the fact that quartet topology inference methods make mistakes, and so, the set Q of inferred quartet topologies contains *quartet errors*. A quartet $\{a, b, c, d\}$ is a quartet error if $ab|cd \in Q$ but $ab|cd \notin Q_T$. Hence, in this sense, Q is an estimate of Q_T. Consequently, the problem of recombining quartet topologies of Q to form an estimate T' of T can be formulated as an optimization problem.

MINIMUM QUARTET INCONSISTENCY (MQI)

Instance: Set Q containing a (fully resolved) quartet topology for each quartet of labels in S.

Goal: Find an evolutionary tree T' labeled by S such that $|Q_{T'} \oplus Q|$ is minimized.

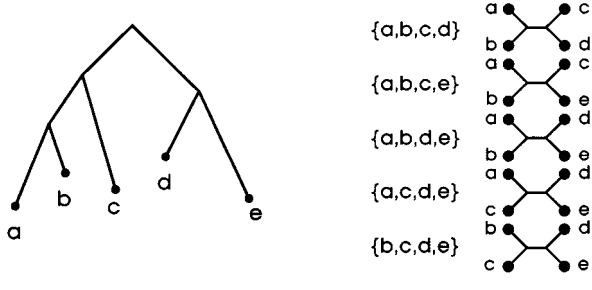

FIGURE 3

It is known that MQI is NP-hard [2]. The results in [2, 3] indicate that one can approximate MQI in polynomial time with ratio n^2, where $n = |S|$. Observe that $|Q| = \theta(n^4)$. No algorithms with better approximation ratios are presently known.

Open problem 3. Can we approximate MQI with ratios at most n? Moreover, is MQI approximable with a constant ratio?

We note that MQI is nonapproximable (with any ratio) if the set Q is not required to contain a topology for every quartet of labels. This follows trivially from the NP-hardness result in [19]. On the other hand, a polynomial time algorithm scheme (PTAS) has been obtained for the complement problem of MQI where we try to minimize $|Q_{T'} \cap Q|$ instead [12]. A crucial fact in this result is that an optimal T' satisfies inequality $|Q_{T'} \cap Q| \geq |Q|/3 = \theta(n^4)$. Unfortunately, there is no analog of such fact for MQI.

4. INTERVAL CUTTING

Our next problem can be easily stated without requiring any background knowledge. The problem arises in the implementation of an algorithmic approach for restriction mapping [6, 11]. Consider closed intervals on the real line with integral end points. An interval $[a, b]$ is *properly contained* in another interval $[c, d]$ if $c < a \leq b < d$. For any interval $[a, b]$, cutting the interval at some point c, where $a < c < b$, produces two intervals $[a, c]$ and $[c, b]$.

INTERVAL CUTTING

Instance: A set S of intervals.

Goal: Find the minimum number of cuts on the intervals in S so that, in the resulting set of intervals, no interval is properly contained in any other interval.

It is known that interval cutting has a PTAS [11]. However, we do not know if the problem is solvable in polynomial time.

Open problem 4. Is interval cutting NP-hard?

Some more general variants of interval cutting are also studied in [11] and some are shown to be NP-hard.

5. CLOSEST SUBSTRING

This problem arises in small molecular drug design where we have to search for conserved regions in many sequences to produce a drug target segment which would bind stably to a site in each of the given sequences.

CLOSEST SUBSTRING

Instance: A set $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ of strings each of length m and over finite alphabet Σ (of size 4 or 20 in practice), and an integer L.

Goal: Find a *median* string s of length L minimizing d such that for each $1 \le i \le n$, there is a length L substring t_i of s_i, satisfying $d_H(s, t_i) \le d$, where d_H is the Hamming distance.

The problem is NP-hard. When $L = m$, the problem in fact has a PTAS [16]. When the objective function is the sum of the Hamming distances between s and t_i's, the problem also has a PTAS [16]. In practice, we usually have $L < m$. The best approximation ratio for this general case is $2 - \frac{2}{2|\Sigma| + 1}$ as proved in [16]. For related literature, see [16].

Open problem 5. Can we achieve a better approximation ratio for the closest substring problem? In particular, is a PTAS possible?

ACKNOWLEDGMENTS

We thank Samir Khuller for several corrections. The research was partially supported by grants from NSERC and CITO. Part of TJ's work was done while visiting at the City University of Hong Kong.

REFERENCES

1. V. Bafna, E. Lawler, and P. Pevzner, Approximation algorithms for multiple sequence alignment, *Theoret. Comput. Sci.* **182** (1997), 233–244.
2. V. Berry, T. Jiang, P. Kearney, M. Li, and T. Wareham, Quartet cleaning: Improved algorithms and simulations, *in* "Proc 7th Annual European Symposium on Algorithms, July, 1999, Prague, Czech Republic."
3. V. Berry, D. Bryant, T. Jiang, P. Kearney, M. Li, T. Wareham, and H. Zhang, A practical algorithm for recovering the best supported edges in an evolutionary tree, *in*, "Proc. 11th Annual ACM-SIAM Symp. on Discrete Algorithms, Jan, 2000," to appear.
4. P. Bonizzoni and G. Della Vedova, The complexity of multiple sequence alignment with SP-score that is metric, *Theoret. Comput. Sci.*, in press.
5. P. Buneman, The recovery of trees from measures of dissimilarity, *in* "Mathematics in the Archaeological and Historical Sciences" (F. R. Hodson, D. G. Kendall, and P. Tautu, Eds.), pp. 387–395, Edinburgh University Press, Edinburgh, 1971.
6. D. Fasulo, T. Jiang, R. Karp, R. Settergren, and E. Thayer, An algorithmic approach to multiple complete digest mapping, *J. Comput. Biol.*, in press.

7. M. R. Garey and D. S. Johnson, ''Computers and Intractabiilty: A Guide to the Theory of NP-Completeness,'' Freeman, New York, 1979.
8. D. Gusfield, Efficient methods for multiple sequence alignment with guaranteed error bounds, *Bull. Math. Biol.* **55** (1993), 141–154.
9. D. Gusfield, ''Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology,'' Cambridge Univ. Press, Cambridge, UK, 1997.
10. S. L. Hakimi and S. S. Yau, Distance matrix of a graph and its realizability, *Quart. Appl. Math.* **22** (1964), 305–317.
11. T. Jiang and R. Karp, Mapping clones with a given ordering or interleaving, *Algorithmica* **21** (1998), 262–284.
12. T. Jiang, P. Kearney, and M. Li, Orchestrating quartets: approximation and data correction, *in* ''Proc. 39th IEEE Symposium on Foundations of Computer Science, Palo Alto, CA, 1998.''
13. W. Just, Reducing gap-0 multiple alignment to multiple alignment, *manuscript*, 1999.
14. S. Kannan and T. Warnow, Tree reconstruction from partial orders, *SIAM J. Comput.* **24** (1995), 511–519.
15. P. Kearney, R. B. Hayward, and H. Meijer, Inferring evolutionary trees from ordinal data, *in* ''Proc. 8th Annual ACM–SIAM Symposium on Discrete Algorithms,'' 1997, pp. 418–426.
16. M. Li, B. Ma, and L. Wang, Finding similar regions in many sequences, *in* ''Proc. 31st ACM Symp. Theory of Computing (STOC'99), 1999.''
17. P. Pevzner, Multiple alignment, communication cost, and graph matching, *SIAM J. Appl. Math.* **56** (1992), 1763–1779.
18. J. Setubal and J. Meidanis, ''Introduction to Computational Molecular Biology,'' PWS–Kent, Boston, 1997.
19. M. Steel, The complexity of reconstructing trees from qualitative characters and subtrees, *J. Classification* **9** (1992), 91–116.
20. L. Wang and T. Jiang, On the complexity of multiple sequence alignment, *J. Comput. Biol.* **1** (1994), 337–348.
21. M. Waterman, ''Introduction to Computational Biology—Maps, Sequences and Genomes,'' Chapman & Hall, London/New York, 1995.
22. M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer, Additive evolutionary trees, *J. Theoret. Biol.* **64** (1977), 199–213.
23. L. Zhang, personal communications.