複数の点集合の最大共通部分集合の近似可能性について

阿久津 達也
機械技術研究所

本稿では、複数の点集合が与えられた時、各点集合の部分集合となるような点集合で点数が最大のもの（最大共通部分点集合）を求める問題について考察を行なう。なお、各点集合は回転および平行移動により適当な位置にずらすことが許されるものとする。主な結果として、点集合の個数に制約が無いときに１次元以上の任意の次元のユークリッド空間において、最大共通部分点集合を近似することが最大クリークの近似と少なくとも同程度以上に困難であることを示す。なお、この結果は頂点が座標値を持つグラフに対しても拡張できる。

ON APPROXIMABILITY OF THE LARGEST COMMON POINT SET OF MULTIPLE POINT SETS

Tatsuya AKUTSU

Mechanical Engineering Laboratory,
1-2 Namiki, Tsukuba, Ibaraki, 305 Japan.
e-mail: akutsu@mel.go.jp

This paper considers the following problem: Given a set of point sets, find the largest point set which is a subset of each point set where each point set may be transformed by any isometric transformation. It is proved that approximating the largest common point set is at least as hard as approximating the maximum clique if the number of input point sets is not bounded. A similar result holds for the problem of approximating the largest common connected subgraph in space, too.

1 Introduction

In genome information processing and chemical information processing, it is important to extract a common part of data from multiple data [4, 8, 13]. In particular, extracting a common patterns from multiple amino acid sequences automatically has been studied extensively [4, 12]. However, a few works have been done for extracting a common patterns from multiple three dimensional protein structures while three dimensional patterns of proteins are considered to have very important information [8]. Motivated by these situations, this paper considers the following problem:

INSTANCE: A collection of D-dimensional point sets $S = \{S_1, S_2, \cdots, S_h\}$.

PROBLEM: Find a set of point C in D-dimensions such that

- $|C|$ is the maximum,
- there is a set of isometric transformations $T = \{T_1, T_2, \cdots, T_h\}$ such that $C = T_1(S_1) \bigcap T_2(S_2) \bigcap \cdots \bigcap T_h(S_h)$.

We call this problem as the *largest common point set problem* (LCP, in short). Relating to LCP, we consider another problem. Instead of point sets, we consider graphs such that a vertex corresponds to a point in D-dimensions and an edge corresponds to a line segment which connects its endpoints. Chemical structures with fixed three dimensional structures and solid models in mechanical CAD can be considered as such graphs. Then, the problem is, given a set of such graphs $\{G_1, G_2, \cdots, G_h\}$, to find the connected graph G_c which is congruent with a subgraph of G_i for every G_i and the number of edges of G_c is the maximum, where each G_i is allowed to be transformed by an isometric transformation. We call this problem as the *largest common geometric subgraph problem* (LCGS, in short).

This paper shows that approximating LCP as well as approximating LCGS is at least as hard as approximating the maximum clique. It is also shown that LCP (resp. LCGS) can be solved in polynomial time if the number of input sets (resp. graphs) is bounded by a constant.

Relating to LCP and LCGS, several studies have been done. The congruity of point sets and graphs in three dimensions was studied by Atkinson [6] and Alt, Mehlhorn, Wagener and Welzl [3]. Moreover, Alt et al. studied the congruity of point sets and graphs in higher dimensions [3]. We also studied a parallel algorithm for the congruity in three dimensions [1]. Sugihara studied the congruity and the partial congruity of polyhedra [15]. While these works are concerned with exact matchings, approximate matchings of point sets in two dimensions have been studied extensively [3, 10, 11].

2 Hardness of Approximating the Largest Common Point Set

In this section, we show that approximating LCP is at least as hard as approximating the maximum clique. Before describing details, we briefly overview recent results about approximation algorithms.

An approximation algorithm for a maximization or minimization problem is said to approximate the optimal value $opt(X)$ within a factor of $f(n)$ if, for all instances X of the problem of size n, $\frac{1}{f(n)} < \frac{g(X)}{opt(X)} < f(n)$ holds where $g(X)$ is the value found by the approximation algorithm. An optimization problem is said to have a polynomial-time approximation scheme if, for any $c > 1$, there exists a polynomial-time algorithm that approximates the optimal solution within a factor of c [9]. Recently, the following two results were proved [5]: MAXSNP-hard problems [14] do not have polynomial time approximation schemes unless $P = NP$; For some $\varepsilon > 0$, the size of the maximum clique in a graph can not be approximated within a factor of n^{ε} in polynomial time unless $P = NP$.

Note that the maximum clique problem (MAX-CLIQUE, in short) is defined as follows: given an undirected graph $G(V, E)$, find the maximum subgraph $G'(V', E')$ of G such that G' is a complete graph (i.e. $(\forall v, w \in V')(\{v, w\} \in E')$). A complete subgraph of G is called as a clique of G. Let $opt_{CLIQUE}(G)$ be the size (the number of vertices) of the maximum clique of a graph G. We assume without loss of generality (w.l.o.g.) that each graph G has at least one edge.

Here, we consider the original problem. For simplicity, we consider the case of 1-dimensional space. The discussions can be trivially generalized to any dimensions. Since we consider the 1-dimensional space, we identify each point with its coordinate value. Moreover, each isometric transformation is specified by a pair (s, l) where s is either '+' or '−' and l is a real number. If s is '+', it denotes the transformation such that each point x is transformed to $x + l$. If s is '−', it denotes the transformation such that each point x is transformed to $-x + l$. For an instance S of LCP, $opt_{LCP}(S)$ denotes the size (the number of elements) of the largest common point set. For LCP, k denotes $\max_i |S_i|$. Theorem 1 describes the main result of this paper by reducing MAX-CLIQUE to LCP. Similar reductions were used to prove the hardness of computing the largest common subtree of bounded vertex degree [2] and the longest common subsequence [12].

[Theorem 1] If $opt_{LCP}(S)$ is approximated within a factor of $O(f(k, h))$ in $O(T(k, h))$ time, then $opt_{CLIQUE}(G)$ can be approximated within a factor of $O(f(2n, n + 1))$ in $O(T(2n, n + 1) + n^2)$ time where n is the number of vertices of G.

(Proof) We reduce MAX-CLIQUE to LCP as follows.

Let $G(V, E)$ be the input for MAX-CLIQUE where $V = \{v_1, \cdots, v_n\}$. For $v \in V$, $\Gamma(v)$ denotes the set of adjacent vertices of v (i.e. $\Gamma(v) = \{w | \{v, w\} \in E\}$). We construct a point set $Q = \{P_1, \cdots, P_{2n}\}$ in 1-dimensional space as follows (see Fig.1). Let L_1 and L_2 be sufficiently large numbers such that $L_1 \gg n^2$ and $L_2 \gg nL_1$ hold, respectively. For example, $L_1 = 100n^2$ and $L_2 = 100nL_1$ are all right. Then, P_i is defined as follows:

$$P_i = \begin{cases} 0, & i = 1 \\ P_{i-1} + L_1 + i - 2, & 1 < i \leq n \\ L_2 + P_{i-n}, & i > n \end{cases}$$

Each of P_i and P_{n+i} corresponds to a vertex v_i.

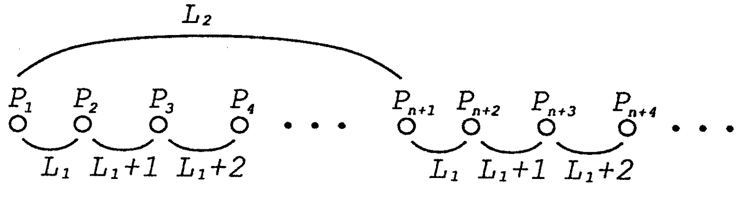

Figure 1: An example of a point set Q.

An instance of LCP is a collection of point sets $S = \{S_1, \cdots, S_n, S_{n+1}\}$, where $S_{n+1} =$

$\{P_1, \cdots, P_n\}$ and, for $i < n+1$, S_i is defined as follows (see Fig.2).

$$S_i = \{ P_j \mid (j \le n) \land (j = i \ \lor \ v_j \in \Gamma(v_i)) \} \bigcup \{ P_j \mid j > n \ \land \ j \neq n+i \} .$$

For $i < n+1$, S_i corresponds to a vertex v_i. Note that this construction can be done in $O(n^2)$ time.

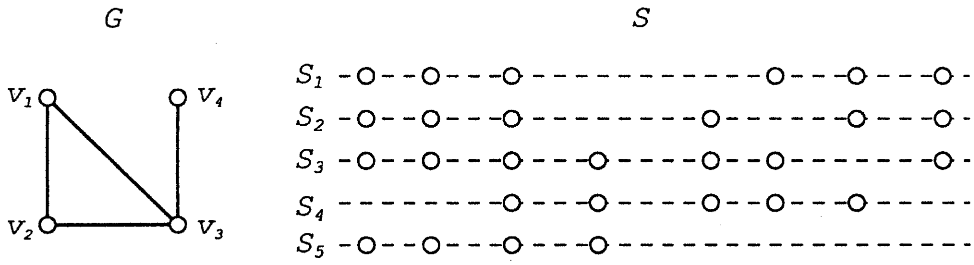

Figure 2: Reduction from MAX-CLIQUE to LCP.

First, we show that, if there is an m-clique (a clique with m vertices), there is a common point set C of size m. Let $W = \{v_{i_1}, \cdots, v_{i_m}\}$ be the point set of the clique. Then, C is defined as $\{P_{i_1}, \cdots, P_{i_m}\}$. For each S_i, C coincides with a subset of S_i by the following transformations. For S_{n+1}, C is trivially a subset of S_{n+1} and then T_{n+1} is specified by (+,0). If $i < n+1$ and $v_i \in W$, C is a subset of S_i and then T_i is specified by (+,0), too. If $i < n+1$ and $v_i \notin W$, C coincides with a subset of S_i by translating C with length L_2 and then T_i is specified by $(+, L_2)$.

Next, we show that, if there is a common point set C of size m, an m-clique can be constructed in $O(n^2)$ time. We assume $m > 2$ since G is assumed to have at least one edge. We can assume w.l.o.g. that $C = \{P_{i_1}, \cdots, P_{i_m}\}$ is a subset of S_{n+1}. Let $\{T_1, \cdots, T_n\}$ be the set of transformations such that $T_1(S_1) \bigcap \cdots \bigcap T_n(S_n) \bigcap S_{n+1} = C$.

Claim 1: For $i < n+1$, if $P_i \in C$, T_i is specified by (+,0) and, if $P_i \notin C$, T_i is specified by $(+, L_2)$.

(Proof) It is sufficient to prove that each transformation is specified by either (+,0) or $(+, L_2)$.

First, we assume that, for some T_i, T_i is specified by $(+, l)$ such that $l \neq 0$ and $l \neq L_2$. Then, it is easy to see that $|T_i(S_i) \cap C| < 2$ holds since $L_1 \gg n^2$ and $L_2 \gg nL_1$ are assumed. Thus, T_i should not be specified by $(+, l)$ such that $l \neq 0$ and $l \neq L_2$.

Next, we assume that, for some T_i, T_i is specified by $(-, l)$. Then, $|T_i(S_i) \cap C| < 3$. It is proved as follows. We assume w.l.o.g. that P_s and P_t $(n \ge s > t)$ in S_i correspond to $P_{s'}$ and $P_{t'}$ $(s' < t' \le n)$ in C by T_i, respectively. That is, $T_i(P_s) = P_{s'}$ and $T_i(P_t) = P_{t'}$ hold. Note that $s - t = t' - s'$ holds since $L_1 \gg n^2$ is assumed. Then, any other point $P_u \in S_i$ can not coincides with $P_q \in C$ by T_i since $|T_i(P_{t-j}) - T_i(P_{t-j-1})| < |P_{t'+j} - P_{t'+j+1}|$ holds for all $n - t' > j > 0$ and $|T_i(P_{s+j}) - T_i(P_{s+j+1})| > |P_{s'-j} - P_{s'-j-1}|$ holds for all $n - s > j > 0$ (see Fig.3). Therefore, T_i should not be specified by $(-, l)$ and the claim is proved.

We select $V' \subset V$ as follows. If T_i is specified by (+,0) (i.e. $P_i \in C$), then $v_i \in V'$. Otherwise, $v_i \notin V'$. It is easy to see that the subgraph of G induced by V' is a clique.

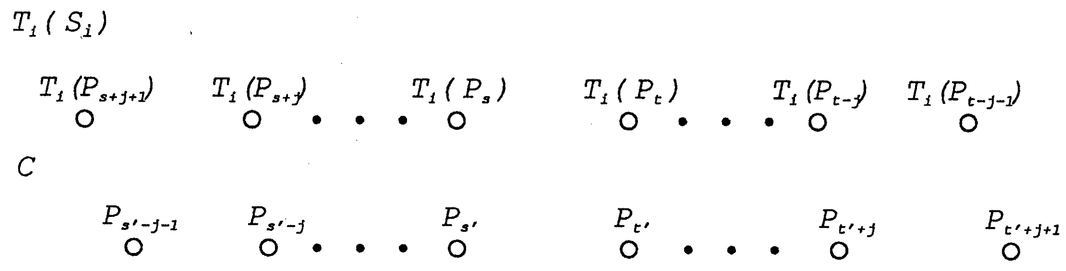

Figure 3: Correspondence of point sets in a case of $(-, l)$.

From the above discussions, it is shown that $opt_{CLIQUE}(G) = opt_{LCP}(S)$ holds and an m-clique can be constructed from a common point set C of size m in $O(n^2)$ time. Therefore, the theorem is proved. □

It follows from Theorem 1 and Ref.[5] that, for some $\varepsilon > 0$, the largest common point set can not be approximated within a factor of n^ε in polynomial time unless $P = NP$ where $n = \min(\{k, h\})$. Note that NP-hardness of LCP follows from the NP-hardness of MAX-CLIQUE, too.

It is easy to see that the same results holds for LCGS in D-dimensions ($D > 1$) considering such graphs as in Fig.4-(a). By the way, vertex degree is bounded by a constant in chemical structures. In such a case, the transformation of Fig.4-(a) is not adequate. In this case, we consider such graphs as in Fig.4-(b) and then we can show that the problem is MAXSNP-hard. It is proved by reducing the independent set problem with bounded vertex degree, which is one of the well-known MAXSNP-hard problems [14]. Since the reduction is similar to one described in Ref.[2], we omit the proof here.

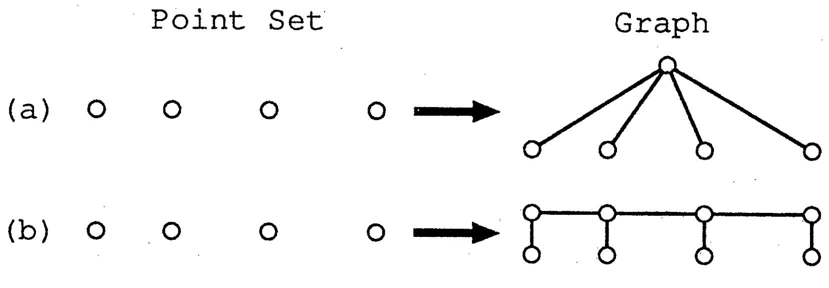

Figure 4: Graphs for the largest common geometric subgraph problem.

3 The Case where the Number of Input Sets is Bounded

While LCP is proved to be NP-hard in Section 2, this section shows that LCP can be solved in polynomial time if the number of input sets is bounded by a constant. The algorithm is very simple and is based on exhaustive search. While we consider the three dimensional case, it seems that the discussion can be extended to any fixed higher dimensions.

Let $S = \{S_1, \cdots, S_h\}$ be an instance of LCP where h is a constant. For simplicity, we assume that the largest common point set is not on a plane. For each point set S_i, (P_i^1, P_i^2, P_i^3) denotes an arbitrary triplet such that each P_i^j belongs to S_i and they do not lie on the same plane. From each of S_i, such a triplet is selected and they are ordered as a sequence. For each sequence $((P_1^1, P_1^2, P_1^3), \cdots, (P_h^1, P_h^2, P_h^3))$, the following procedure is executed: First, each S_i $(i > 1)$ is moved by isometric transformation so that P_i^1, P_i^2 and P_i^3 coincide with P_1^1, P_1^2 and P_1^3, respectively. Next, $C = S_1 \bigcap S_2' \bigcap S_3' \bigcap \cdots \bigcap S_h'$ is computed where S_i' denotes the transformed set of S_i. Finally, the largest C is the largest common point set.

Since the correctness of the algorithm is almost obvious, we consider the time complexity. The number of sequences is $O(k^{3h})$ since the number of triplets is $O(k^3)$ for each S_i. The time required for testing each sequence is $O(kh)$. Thus, the total time required is $O(hk^{3h+1})$. Since h is a constant, the algorithm works in polynomial time. Note that the algorithm can be modified for LCGS.

4 Concluding Remarks

This paper shows that approximating the largest common point set is at least as hard as approximating the maximum clique. It is also shown that the largest common point set problem can be solved in polynomial time if the number of input sets is bounded by a constant.

While the hardness result for approximating the maximum clique was known, a polynomial time algorithm which approximates the maximum clique within a factor of $O(n/\log^2 n)$ [7] is known. However, we did not find an approximation algorithm for LCP even within a factor of $O(n/\log n)$. Thus, it is interesting to study whether or not such an approximation algorithm exists.

Although this paper shows a negative result for finding the largest common point set, it does not mean that common substructures of multiple protein structures can not computed efficiently. Since atoms in the backbone chain of a protein can be regarded as a sequence of points, a consecutive portion of the sequence might be regarded as a substructure. In such a case, techniques in approximate point or string matching seem to be useful. We are developing practical pattern matching algorithms for three dimensional protein structures based on such techniques. Details will be presented elsewhere.

References

[1] T. Akutsu. "A parallel algorithm for determining the congruity of point sets in three dimensions". Technical Report AL31-3, Information Processing Society of Japan, 1992.

[2] T. Akutsu. "On the largest common subtree of multiple trees". In *Proceedings of the 6th Karuizawa Workshop on Circuits and Systems*, pp. 109–114, 1993.

[3] H. Alt, K. Mehlhorn, H. Wagener, and E. Welzl. "Congruence, similarity, and symmetries of geometric objects". *Discrete and Computational Geometry*, Vol. 3, pp. 237–256, 1988.

[4] S. Arikawa, S. Kuhara, S. Miyano, Y. Mukouchi, A. Shinohara, and T. Shinohara. "Machine discovery of a negative motif from amino acid sequences by decision trees over regular patterns". In *Proc. International Conference on Fifth Generation Computer Systems*, pp. 618–625, 1992.

[5] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. "Proof verification and hardness of approximation problems". In *Proceedings of IEEE Symposium on Foundations of Computer Science*, pp. 14–23, 1992.

[6] M. D. Atkinson. "An optimal algorithm for geometrical congruence". *Journal of Algorithms*, Vol. 8, pp. 159–172, 1987.

[7] R. B. Boppana and M. M. Halldórsson. "Approximating maximum independent set by excluding subgraphs". *BIT*, Vol. 32, pp. 180–196, 1992.

[8] C. Branden and J. Tooze, editors. *"Introduction to Protein Structure"*. Garland Publishing, Inc., New York, 1991.

[9] M. R. Garey and D. S. Johnson. *"Computers and Intractability: A Guide to the Theory of NP-completeness"*. Freeman, San Francisco, 1979.

[10] P. J. Heffernan and S. Schirra. "Approximate decision algorithm for point sets congruence". In *Proceedings of ACM Symposium on Computational Geometry*, pp. 93–101, 1992.

[11] K. Imai, S. Sumino, and H. Imai. "Minimax geometric fitting of two corresponding sets of points". In *Proceedings of ACM Symposium on Computational Geometry*, pp. 266–275, 1989.

[12] R. Maier. "The complexity of some problems on subsequences and supersequences". *Journal of the ACM*, Vol. 25, pp. 322–336, 1978.

[13] M. Ooki, S. Sasaki, and H. Chihara, editors. *"Chemistry and Information"*. Iwanami Shoten, Japan, 1981. (in Japanese).

[14] C. Papadimitriou and M. Yannakakis. "Optimization, approximation, and complexity classes". *Journal of Computer and System Sciences*, Vol. 43, pp. 425–440, 1991.

[15] K. Sugihara. "An $n \log n$ algorithm for determining the congruity of polyhedra". *Journal of Computer and System Sciences*, Vol. 29, pp. 36–47, 1984.