Algorithms **2008**, *1*, 43-51; DOI: 10.3390/a1020043

OPEN ACCESS

algorithms

ISSN 1099-4300

Article

www.mdpi.com/journal/algorithms

A PTAS For The k-Consensus Structures Problem Under Squared Euclidean Distance

Shuai Cheng Li [1,⋆], Yen Kaow Ng [2] and Louxin Zhang [3]

[1] David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada N2L 3G1.
[2] Department of Computer Science and Communication Engineering, Kyushu University, Fukuoka 819-0395, Japan.
[3] Department of Mathematics, National University of Singapore, Singapore 117543.
E-mails: scli@cs.uwaterloo.ca; kalngyk@tcslab.csce.kyushu-u.ac.jp; matzlx@nus.edu.sg
⋆ Author to whom correspondence should be addressed.

Received: 5 September 2008; in revised form: 1 October 2008; Accepted: 9 October 2008; Published: 9 October 2008

Abstract: In this paper we consider a basic clustering problem that has uses in bioinformatics. A *structural fragment* is a sequence of ℓ points in a 3D space, where ℓ is a fixed natural number. Two structural fragments f_1 and f_2 are equivalent if and only if $f_1 = f_2 \cdot R + \tau$ under some rotation R and translation τ. We consider the *distance* between two structural fragments to be the sum of the squared Euclidean distance between all corresponding points of the structural fragments. Given a set of n structural fragments, we consider the problem of finding k (or fewer) structural fragments $g_1, g_2, \ldots, g_k$, so as to minimize the sum of the distances between each of $f_1, f_2, \ldots, f_n$ to its nearest structural fragment in $g_1, \ldots, g_k$. In this paper we show a polynomial-time approximation scheme (PTAS) for the problem through a simple sampling strategy.

Keywords: Clustering 3D point sequences; squared Euclidean distance; algorithm; polynomial-time approximation scheme.

---

1. Introduction

In this paper we consider the problem of clustering similar sequences of 3D points. Two such sequences of points are considered the same if they are equivalent under rotation and translation. The scenario which we consider is as follows. Suppose there is an original sequence of points that gave rise

to a few variations of itself, through slight changes in some or all of its points. Now given these variations of the sequence, we are to reconstruct the original sequence. A likely candidate for such an original sequence would be a sequence which is "nearest" in terms of some distance measure, to the variations.

A more complicated scenario involves k original sequences of the same length. Formally, we formulate the problem as follows. Given n sequences of points $f_1, f_2, \ldots, f_n$, we are to find a set of k sequences $g_1, \ldots, g_k$, such that the sum of distances

$$\sum_{1 \le i \le n} \min_{1 \le j \le k} dist(f_i, g_j) \tag{1}$$

is minimized. In this paper we consider the case where $dist$ is the *minimum* sum of squared Euclidean distances between each of the points in the two sequences f_i and g_k, *under all possible rigid transformations on the sequences of points*. A cost function in the form of the squared Euclidean distance is used in many techniques for clustering 3D points [1]. Since our clustering problem is quite different from those previously studied, it calls for a new technique. (The "square" in the distance measure is to fulfill a condition needed by the method in this paper. The method does not work, for example, in the case of the root mean squared Euclidean distance. On the other hand, the method easily adapts to other distance measures that fulfill the required condition.)

Such a problem has potential use in clustering protein structures. A protein structure is typically given as a sequence of points in 3D space, and for various reasons, there are typically minor variations in their measured structures. The problem can be considered a model of the situation where we have a set of measurements of a few protein structures, and are to reconstruct the original structures.

In this paper, we show that there is a polynomial-time approximation scheme (PTAS) for the problem, through a sampling strategy. More precisely, we show that an optimal solution obtained by sampling smaller subsets of the input suffices to give us an approximate solution, and the approximation ratio improves as we increase the size of the subsets we sample.

2. Preliminaries

Throughout this paper we let ℓ be a fixed non-zero natural number. A *structural fragment* is a sequence of ℓ 3D-points. The *mean square distance (MS)* between two structural fragments $f = (f[1], \ldots, f[\ell])$ and $g = (g[1], \ldots, g[\ell])$, is defined to be

$$MS(f, g) = \min_{R \in \mathcal{R}, \tau \in \mathcal{T}} \sum_{i=1}^{\ell} \parallel f[i] - (R \cdot g[i] + \tau) \parallel^2 \tag{2}$$

where $\mathcal{R}$ is the set of all rotation matrices, $\mathcal{T}$ the set of all translation vectors, and $\parallel x - y \parallel$ is the Euclidean distance between $x, y \in \mathbb{R}^3$.

The root of the MS measure, $RMS(f, g) = \sqrt{MS(f, g)}$ is a measure that has been extensively studied. Note that $R \in \mathcal{R}$, $\tau \in \mathcal{T}$ that minimize $\sum_{i=1}^{\ell} \parallel f[i] - (R \cdot g[i] + \tau) \parallel^2$ to give us $MS(f, g)$ will also give us $RMS(f, g)$, and vice versa. Since given any f and g, there are closed form equations [2, 3] for finding R and τ that give $RMS(f, g)$, $MS(f, g)$ can be computed efficiently for any f and g.

Furthermore, it is known that to minimize $\sum_{i=1}^{\ell} \parallel f[i] - (R \cdot g[i] + \tau) \parallel^2$, the centroid of f and g must coincide [2]. Due to this, without loss of generality we assume that all structural fragments have

centroids at the origin. Such transformations can be done in $O(n\ell)$ time. After such transformations, in computing $MS(f,g)$, only the parameter $R \in \mathcal{R}$ need to be considered, that is,

$$MS(f,g) = \min_{R\in\mathcal{R}} \sum_{i=1}^{\ell} \parallel f[i] - R \cdot g[i] \parallel^2 \tag{3}$$

Suppose that given a set of n structural fragments $f_1, f_2, \ldots, f_n$, we are to find k structural fragments $g_1, \ldots, g_k$, such that each structural fragment f_i is "near", in terms of the MS, to at least one of the structural fragments in $g_1, \ldots, g_k$. We formulate such a problem as follows:

k-CONSENSUS STRUCTURAL FRAGMENTS PROBLEM UNDER MS	
Input:	n structural fragments $f_1, \ldots f_n$, and a non-zero natural number $k < n$.
Output:	k structural fragments $g_1, \ldots g_k$, minimizing the cost $\sum_{i=1}^{n} \min_{1\le j\le k} MS(f_i, g_j)$.

In this paper we will demonstrate that there is a PTAS for the problem.

We use the following notations: Cardinality of a set A is written $|A|$. For a set A and non-zero natural number n, A^n denotes the set of all length n sequences of elements of A. Let elements in a set A be indexed, say $A = \{f_1, f_2, \ldots, f_n\}$, then $A^{m!}$ denotes the set of all the length m sequences $f_{i_1}, f_{i_2}, \ldots, f_{i_m}$, where $1 \le i_1 \le i_2 \le \ldots \le i_m \le n$. For a sequence S, $S(i)$ denotes the i-th element in S, and $|S|$ denotes its length.

3. PTAS for the k-Consensus Structural Fragments

The following lemma, from [4], is central to the method.

Lemma 1 ([4]) *Let $a_1, a_2, \ldots, a_n$ be a sequence of real numbers and let $r \in N$, $1 \le r \le n$. Then the following equation holds:*

$$\frac{1}{n^r} \sum_{1\le i_1, i_2, \ldots, i_r \le n} \sum_{i=1}^{n} \left(\frac{a_{i_1} + a_{i_2} + \cdots + a_{i_r}}{r} - a_i\right)^2 = \frac{r+1}{r} \sum_{i=1}^{n} \left(\frac{a_1 + a_2 + \cdots + a_n}{n} - a_i\right)^2 \tag{4}$$

Let $P_1 = (x_1, y_1, z_1), P_2 = (x_2, y_2, z_2), \ldots, P_n = (x_n, y_n, z_n)$ be a sequence of 3D points.

$$\begin{aligned}
& \frac{1}{n^r} \sum_{1\le i_1, i_2, \ldots, i_r \le n} \sum_{i=1}^{n} \parallel \frac{P_{i_1} + P_{i_2} + \cdots + P_{i_r}}{r} - P_i \parallel^2 \\
= & \frac{1}{n^r} \sum_{1\le i_1, \ldots, i_r \le n} \sum_{i=1}^{n} \left(\frac{x_{i_1} + \ldots + x_{i_r}}{r} - x_i\right)^2 + \left(\frac{y_{i_1} + \ldots + y_{i_r}}{r} - y_i\right)^2 + \left(\frac{z_{i_1} + \ldots + z_{i_r}}{r} - z_i\right)^2 \\
= & \frac{r+1}{r} \sum_{i=1}^{n} \left(\frac{x_1 + \ldots + x_n}{n} - x_i\right)^2 + \left(\frac{y_1 + \ldots + z_n}{n} - z_i\right)^2 + \left(\frac{z_1 + \ldots + z_n}{n} - z_i\right)^2 \\
= & \frac{r+1}{r} \sum_{i=1}^{n} \parallel \frac{P_1 + P_2 + \cdots + P_n}{n} - P_i \parallel^2
\end{aligned} \tag{5}$$

One can similarly extend the equation for structural fragments. Let $f_1, \ldots, f_n$ be n structural fragments, the equation becomes:

$$\frac{1}{n^r} \sum_{1 \leq i_1, \ldots, i_r \leq n} \sum_{i=1}^{n} \parallel \frac{f_{i_1} + \cdots + f_{i_r}}{r} - f_i \parallel^2 = \frac{r+1}{r} \sum_{i=1}^{n} \parallel \frac{f_1 + \cdots + f_n}{n} - f_i \parallel^2 \tag{6}$$

The equation says that there exists a sequence of r structural fragments $f_{i_1}, f_{i_2}, \ldots, f_{i_r}$ such that

$$\sum_{i=1}^{n} \parallel \frac{f_{i_1} + \cdots + f_{i_r}}{r} - f_i \parallel^2 \leq \frac{r+1}{r} \sum_{i=1}^{n} \parallel \frac{f_1 + \cdots + f_n}{n} - f_i \parallel^2 \tag{7}$$

Our strategy uses this fact —in essentially the same way as in [4]— to approximate the optimal solution for the k-consensus structural fragments problem. That is, by exhaustively sampling every combination of k sequences, each of r elements from the space $\mathcal{R}' \times \{f_1, \ldots, f_n\}$, where $f_1, \ldots, f_n$ is the input and $\mathcal{R}'$ is a fixed selected set of rotations, which we next discuss.

3.1. Discretized Rotation Space

Any rotation can be represented by a normalized vector u and a rotation angle θ, where u is the axis about which an object is rotated by θ. If we apply (u, θ) to a vector v, we obtain vector $\hat{v}$, which is:

$$\hat{v} = u(v \cdot u) + (v - w(v \cdot w)) \cos\theta + (v \times w) \sin\theta \tag{8}$$

where $\cdot$ represents dot product, and $\times$ represent cross product.

By the equation, one can verify that a change of ϵ in u will result in a change of at most $\alpha_1 \epsilon |v|$ in $|\hat{v}|$ for some computable $\alpha_1 \in \mathbb{R}$; and a change of ϵ in θ will result in a change of at most $\alpha_2 \epsilon |v|$ in $|\hat{v}|$ for some computable $\alpha_2 \in \mathbb{R}$. Now any rotation along an axis through the origin can be written in the form $(\theta_1, \theta_2, \theta_3)$, where $\theta_1, \theta_2, \theta_3 \in [0, 2\pi)$ are respectively a rotation along each of the x, y, z axes. Similarly, changes of ϵ in θ_1, θ_2 and θ_3 will result in a change of at most $\alpha\epsilon|v|$, for some computable $\alpha \in \mathbb{R}$.

We discretize the values that each θ_i, $1 \leq i \leq 3$ may take within the range $[0, 2\pi)$ into a series of angles of angular difference ϑ. There are hence at most $O(1/\vartheta)$ of such values for each θ_i, $1 \leq i \leq 3$. Let $\mathcal{R}'$ denote the set of all possible discretized rotations $(\theta_1, \theta_2, \theta_3)$. Note that $|\mathcal{R}'|$ is of order $O(1/\vartheta^3)$.

Let $\boldsymbol{d}$ be the diameter of a ball that is able to encapsulate each of $f_1, f_2, \ldots, f_n$. Hence any distance between two points among $f_1, \ldots, f_n$ is at most $\boldsymbol{d}$. In this paper we assume $\boldsymbol{d}$ to be constant with respect to the input size. Note that for a protein structure, $\boldsymbol{d}$ is of order $O(\ell)$ [5]. For any $b \in \mathbb{R}$, we can choose ϑ so small that for any rotation R and any point $p \in \mathbb{R}^3$, there exists $R' \in \mathcal{R}'$ such that $\parallel R \cdot p - R' \cdot p \parallel \leq \alpha\vartheta\boldsymbol{d} \leq b$.

3.2. A Polynomial-time Algorithm With Cost $((1+\epsilon)D_{opt} + c)$

Our algorithm for the k-consensus structural fragments problem is summarized in Table 1.

This is what the algorithm does: In (2), we explore m distinct subsets $A_1, \ldots, A_m$ from $f_1, \ldots, f_n$, in the hope that each subset is from a distinct cluster in the optimal clustering. Since we explore all possible such subsets this is bound to happen. We then try to evaluate the score of each subset A_j by sampling up to r structural fragments (allowing repeats) from it (from (2.1) onwards). Such an evaluation is possible

due to Equation 7. The evaluation also requires us to exhaustively try out all possible transformations in $\mathcal{R}'$, which is what we try to do in (2.2). Each of these samplings of A_j produces a consensus structural fragment u_j for A_j in (2.3), the score of which is evaluated in (2.4). Finally in (3), we output the consensus patterns $u_1, \ldots, u_m$ which give us the best score.

We now analyze the runtime complexity of the algorithm. Consider the number of $F_1, F_2, \ldots, F_m$ in (2.1) that are possible. Let each F_j be represented by a length r string of $n+1$ symbols, n of which each represents one of $f_1, \ldots, f_n$, while the remaining symbol represents "nothing". It is clear that for any A_j, any $F_j \in A_j^{r!}$, or $F_j \in A_j^{|A_j|!}$ (where $|A_j| \leq r$), can be represented by one such string. Furthermore, any $F_1, F_2, \ldots, F_m$ can be completely represented by k such strings — that is, to represent the case where $m < k$, $k - m$ strings can be set to "nothing" completely. From this, we can see that there are at most $(n+1)^{rk} = O(n^{rk})$ possible combinations of $F_1, F_2, \ldots, F_m$.

For each of these combinations, there are $|\mathcal{R}'|^{rk}$ possible combinations of $\Theta_1, \Theta_2, \ldots, \Theta_m$ at (2.2), hence resulting in $O((n|\mathcal{R}'|)^{rk})$ iterations to run for (2.3) to (2.5). Since (2.3) can be done in $O(rk\ell)$, (2.4) in $O(nk|\mathcal{R}'|\ell)$, and (2.5) in $O(n)$ time, the algorithm completes in $O(k\ell(r+n|\mathcal{R}'|)(n|\mathcal{R}'|)^{rk})$ time.

We argue that D_{min} eventually is at most $(r+1)/r$ of the optimal solution plus a factor. Suppose the optimal solution results in the $m \leq k$ disjoint clusters $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m \subseteq \{f_1, \ldots, f_n\}$.

For each $\mathbf{A}_j$, $1 \leq j \leq m$, let $\boldsymbol{u}_j$ be a structural fragment which minimizes $\sum_{f \in \mathbf{A}_j} MS(\boldsymbol{u}_j, f)$. Fur-

Approximation Algorithm k-CONSENSUS STRUCTURAL FRAGMENTS

Input: structural fragments $f_1, \ldots f_n$, natural numbers $k < n$ and $r \geq 1$.

Output: up to k structural fragments $u_1, \ldots u_m$, $m \leq k$.

(1) Let $D_{min} = \infty$, Consensus$= \emptyset$. (Consensus will contain the output.)

(2) For every possible set of $m \leq k$ disjoint sets $A_1, \ldots, A_m \subseteq \{f_1, \ldots, f_n\}$.

(2.1) For every possible set $F_1, F_2, \ldots, F_m$ of sequences where

$F_j \in A_j^{r!}$ if $|A_j| > r$, otherwise

F_j is the (unique) sequence in $A_j^{|A_j|!}$ that contains all the elements of A_j.

(Note that every distinct set of F_1, …, F_m needs to be considered only once.)

(2.2) For every possible sequence $\Theta_1, \Theta_2, \ldots, \Theta_m$, where $\Theta_j \in \mathcal{R}'^{|F_j|}$ for $1 \leq j \leq m$.

(2.3) For $j = 1$ to m, find u_j, the average structural fragment

for $\Theta_j(1) \cdot F_j(1)$,

$\Theta_j(2) \cdot F_j(2)$,

$\vdots$

$\Theta_j(|F_j|) \cdot F_j(|F_j|)$.

(2.4) For $i = 1$ to n, find $d_i = \min\{\| u_j - R \cdot f_i \|^2 | \ 1 \leq j \leq m, R \in \mathcal{R}'\}$.

(2.5) If $\sum_{i=1}^{n} d_i < D_{min}$,

set D_{min} to $\sum_j d_j$ and set Consensus to $\{u_1, \ldots, u_m\}$.

(3) Output Consensus.

Table 1. Polynomial-time algorithm for the problem.

thermore, for each $f \in \mathbf{A}_j$, let $\boldsymbol{R}_f$ be a rotation where

$$\boldsymbol{R}_f \in \arg\min_{R \in \mathcal{R}} \parallel \boldsymbol{u}_j - R \cdot f \parallel^2 \tag{9}$$

and let

$$\mathbf{D}_j = \sum_{f \in \mathbf{A}_j} \parallel \boldsymbol{u}_j - \boldsymbol{R}_f \cdot f \parallel^2 \quad \text{(Hence the optimal cost, } \mathbf{D} = \textstyle\sum_{j=1}^{m} \mathbf{D}_j\text{.)} \tag{10}$$

By the property of the MS measure, it can be shown that $\boldsymbol{u}_j$ is the average of $\{\boldsymbol{R}_f \cdot f \mid f \in \mathbf{A}_j\}$. For each $\mathbf{A}_j$ where $|\mathbf{A}_j| > r$, by Equation 6,

$$\frac{1}{|\mathbf{A}_j|^r} \sum_{F_j \in \mathbf{A}_j^r} \sum_{f \in \mathbf{A}_j} \parallel \frac{\boldsymbol{R}_{F_j(1)} \cdot F_j(1) + \cdots + \boldsymbol{R}_{F_j(r)} \cdot F_j(r)}{r} - \boldsymbol{R}_f \cdot f \parallel^2 \;=\; \frac{r+1}{r} \mathbf{D}_j \tag{11}$$

For each such $\mathbf{A}_j$, let $\boldsymbol{F}_j \in \mathbf{A}_j^r$ be such that

$$\sum_{f \in \mathbf{A}_j} \parallel \frac{\boldsymbol{R}_{\boldsymbol{F}_j(1)} \cdot \boldsymbol{F}_j(1) + \cdots + \boldsymbol{R}_{\boldsymbol{F}_j(r)} \cdot \boldsymbol{F}_j(r)}{r} - \boldsymbol{R}_f \cdot f \parallel^2 \;\leq\; \frac{r+1}{r} \mathbf{D}_j \tag{12}$$

Without loss of generality assume that each $\boldsymbol{F}_j \in \mathbf{A}_j^{r!}$. Let

$$\boldsymbol{\mu}_j = \begin{cases} \frac{\boldsymbol{R}_{\boldsymbol{F}_j(1)} \cdot \boldsymbol{F}_j(1) + \cdots + \boldsymbol{R}_{\boldsymbol{F}_j(r)} \cdot \boldsymbol{F}_j(r)}{r} & \text{if } |\mathbf{A}_j| > r \\ \frac{\boldsymbol{R}_{\boldsymbol{F}_j(1)} \cdot \boldsymbol{F}_j(1) + \cdots + \boldsymbol{R}_{\boldsymbol{F}_j(|\mathbf{A}_j|)} \cdot \boldsymbol{F}_j(|\mathbf{A}_j|)}{|\mathbf{A}_j|} & \text{otherwise} \end{cases} \tag{13}$$

Then we may write,

$$\sum_{j=1}^{m} \sum_{f \in \mathbf{A}_j} \parallel \boldsymbol{\mu}_j - \boldsymbol{R}_f \cdot f \parallel^2 \;\leq\; \frac{r+1}{r} \mathbf{D} \tag{14}$$

For each rotation $\boldsymbol{R}_f$, let R_f be a closest rotation to $\boldsymbol{R}_f$ within $\mathcal{R}'$. Also, let

$$\mu_j = \begin{cases} \frac{R_{\boldsymbol{F}_j(1)} \cdot \boldsymbol{F}_j(1) + \cdots + R_{\boldsymbol{F}_j(r)} \cdot \boldsymbol{F}_j(r)}{r} & \text{if } |\mathbf{A}_j| > r \\ \frac{R_{\boldsymbol{F}_j(1)} \cdot \boldsymbol{F}_j(1) + \cdots + R_{\boldsymbol{F}_j(|\mathbf{A}_j|)} \cdot \boldsymbol{F}_j(|\mathbf{A}_j|)}{|\mathbf{A}_j|} & \text{otherwise} \end{cases} \tag{15}$$

Since we exhaustively sample all possible $F_j \in A_j^{r!}$ for all possible A_j and for all $R \in \mathcal{R}'$, it is clear that:

$$D_{min} \;\leq\; \sum_{j=1}^{m} \sum_{f \in \mathbf{A}_j} \parallel \mu_j - R_f \cdot f \parallel^2 \tag{16}$$

We will now relate the LHS of Equation 14 with the RHS of Equation 16. The RHS of Equation 16 is

$$\begin{aligned}
& \sum_{j=1}^{m}\sum_{f\in\mathbf{A}_j} \| \mu_j - R_f \cdot f \|^2 \\
= & \sum_{j=1}^{m}\sum_{f\in\mathbf{A}_j} \| \mu_j + (\boldsymbol{\mu}_j - \boldsymbol{\mu}_j) + (\boldsymbol{R}_f \cdot f - \boldsymbol{R}_f \cdot f) - R_f \cdot f \|^2 \\
\leq & \sum_{j=1}^{m}\sum_{f\in\mathbf{A}_j} (\| \boldsymbol{\mu}_j - \boldsymbol{R}_f \cdot f \| + (\| \mu_j - \boldsymbol{\mu}_j \| + \| \boldsymbol{R}_f \cdot f - R_f \cdot f \|))^2 \\
= & \sum_{j=1}^{m}\sum_{f\in\mathbf{A}_j} \| \boldsymbol{\mu}_j - \boldsymbol{R}_f \cdot f \|^2 + (\| \mu_j - \boldsymbol{\mu}_j \| + \| \boldsymbol{R}_f \cdot f - R_f \cdot f \|)^2 \\
& \qquad + 2 \| \boldsymbol{\mu}_j - \boldsymbol{R}_f \cdot f \| (\| \mu_j - \boldsymbol{\mu}_j \| + \| \boldsymbol{R}_f \cdot f - R_f \cdot f \|) \\
\leq & \sum_{j=1}^{m}\sum_{f\in\mathbf{A}_j} \| \boldsymbol{\mu}_j - \boldsymbol{R}_f \cdot f \|^2 + 8n\ell b
\end{aligned} \tag{17}$$

Hence by Equation 14, D_{min} is at most $(r+1)/r = 1 + 1/r$ of the optimal solution plus a factor $c = 8n\ell b$. Let $\epsilon = 1/r$,

Theorem 2 *For any $c, \epsilon \in \mathbb{R}$, a $((1+\epsilon)D_{opt} + c)$-approximation solution for the k-consensus structural fragments problem can be computed in*

$$O(k\ell(\frac{1}{\epsilon} + n|\mathcal{R}'|)(n|\mathcal{R}'|)^{\frac{k}{\epsilon}})$$

time.

The factor c in Theorem 2 is due to error introduced by the use of discretization in rotations. If we are able to estimate a lower bound of D_{opt}, we can scale this error by refining the discretization such that c is an arbitrarily small factor of D_{opt}. To do so, in the next section we show a lower bound to D_{opt}.

3.3. A Polynomial-time 4-approximation Algorithm

We now show a 4-approximation algorithm for the k-consensus structural fragments problem. We first show the case for $k = 1$, and then generalizes the result to all $k \geq 2$.

Let the input n structural fragments be $f_1, f_2, \ldots, f_n$. Let f_a, $1 \leq a \leq n$ be the structural fragment where

$$\sum_{1\leq j\leq n \wedge j\neq a} MS(f_a, f_j)$$

is minimized. Note that f_a can be found in time $O(n^2\ell)$, since for any $1 \leq i, j \leq n$, $MS(f_i, f_j)$ (more precisely, $RMS(f_i, f_j)$) can be computed in time $O(\ell)$ using closed form equations from [3].

We argue that f_a is a 4-approximation. Let the optimal structural fragment be f_{opt}, the corresponding distance D_{opt}, and let f_b ($1 \leq b \leq n$) be the fragment where $MS(f_b, f_{opt})$ is minimized.

We first note that the cost of using f_a as solution, $\sum_{i\neq a} MS(f_a, f_i) \leq \sum_{i\neq b} MS(f_b, f_i)$. To continue we first establish the following claim.

Claim 1 $MS(f, f') \leq 2(MS(f, f'') + MS(f'', f'))$.

PROOF. In [6], it is shown that

$$RMS(f, f') \leq RMS(f, f'') + RMS(f'', f') \tag{18}$$

Squaring both sides gives

$$MS(f, f') \leq MS(f, f'') + MS(f'', f') + 2RMS(f, f'')RMS(f'', f') \tag{19}$$

Since

$$2RMS(f, f'')RMS(f'', f') \leq MS(f, f'') + MS(f'', f') \tag{20}$$

we have $MS(f, f') \leq 2(MS(f, f'') + MS(f'', f'))$. ∎

By the above claim,

$$\begin{aligned}
\sum_{i \neq b} MS(f_b, f_i) &\leq 2\sum_{i \neq b}(MS(f_b, f_{opt}) + MS(f_{opt}, f_i)) && (21)\\
&= 2\sum_{i \neq b} MS(f_b, f_{opt}) + 2\sum_{i \neq b} MS(f_i, f_{opt}) && (22)\\
&\leq 2\sum_{i \neq b} MS(f_b, f_{opt}) + 2D_{opt} && (23)\\
&\leq 2\sum_{j \neq b} MS(f_j, f_{opt}) + 2D_{opt} && (24)\\
&\leq 2D_{opt} + 2D_{opt} = 4D_{opt} && (25)
\end{aligned}$$

Hence $\sum_{i \neq a} MS(f_a, f_i) \leq 4D_{opt}$. We now extend this to k structural fragments.

4-Approximation Algorithm k-CONSENSUS STRUCTURAL FRAGMENTS
Input: structural fragments $S = \{f_1, \ldots f_n\}$, natural number $k < n$.
Output: up to k structural fragments A.
(1) For every set $A \subseteq S$ of up to k structural fragments, do
(2) Compute cost$(A) = \sum_{f \in S-A} \min_{f' \in A} MS(f, f')$
(3) Output A with the least cost(A).

We first pre-compute $MS(f, f')$ for every pair of $f, f' \in S$, which takes time $O(n^2\ell)$. Then, at step (1), there are at most $O(n^k)$ combinations of A, each which takes $O(nk)$ time to compute at step (2). Hence in total we can perform the computation in $O(n^2\ell + kn^{k+1})$ time. To see that the solution is a 4-approximation, let $S_1, S_2, \ldots, S_m$ where $m \leq k$ be an optimal clustering. Then, by our earlier argument, there exists $f_{i_1} \in S_1$, $f_{i_2} \in S_2$, …, $f_{i_m} \in S_m$ such that each f_{i_x} is a 4-approximation for S_x, and hence $f_{i_1}, f_{i_2}, \ldots, f_{i_m}$ is a 4-approximation for the k-consensus structural fragments problem. Since the algorithm exhaustively search for every combination of up to k fragments, it gives a solution at least as good as $f_{i_1}, f_{i_2} \ldots, f_{i_m}$, and hence is a 4-approximation algorithm.

Theorem 3 *A 4-approximation solution for the k-consensus structural fragments problem can be computed in $O(n^2\ell + kn^{k+1})$ time.*

3.4. A $(1+\epsilon)$ Polynomial-time Approximation Scheme

Recall that the algorithm in Section 3.2 has cost $D \leq (1+\epsilon)D_{opt} + 8n\ell b$ where $b = \alpha\vartheta\boldsymbol{d}$. From Section 3.3 we have a lower bound $\boldsymbol{D}_{opt}$ of D_{opt}. We want $8n\ell b \leq \epsilon\boldsymbol{D}_{opt} \leq \epsilon D_{opt}$. To do so, it suffices that we set $\vartheta \leq \epsilon\boldsymbol{D}_{opt}/(8n\ell\alpha\boldsymbol{d})$. This results in an $|\mathcal{R}'|$ of order $O(1/\vartheta^3) = O((n\ell\boldsymbol{d})^3)$. Substituting this in Theorem 2, and combining with Theorem 3, we get the following.

Theorem 4 *For any $\epsilon \in \mathbb{R}$, a $((1+\epsilon)D_{opt})$-approximation solution for the k-consensus structural fragments problem can be computed in*

$$O(n^2\ell + kn^{k+1} + k\ell(\frac{2}{\epsilon} + n\lambda)(n\lambda)^{\frac{2k}{\epsilon}})$$

time, where $\lambda = (n\ell\boldsymbol{d})^3$.

4. Discussions

The method in this paper depends on Lemma 1. For this reason, the technique does not extend to the problem under distance measures where Lemma 1 cannot be applied, for example, the RMS measure. However, should Lemma 1 apply to a distance measure, it should be easy to adapt the method here to solve the problem for that distance measure.

One can also formulate variations of the k-consensus structural fragments problem. For example,

k-CLOSEST STRUCTURAL FRAGMENTS PROBLEM UNDER MS	
Input:	n structural fragments $f_1, \ldots f_n$, and a non-zero natural number $k < n$.
Output:	k structural fragments $g_1, \ldots g_k$, minimizing the threshold $\max_{1\leq i\leq n} \min_{1\leq j\leq k} MS(f_i, g_j)$.

While the cost function of the k-consensus structural fragments problem resembles that of the k-means problem, the cost function of the k-closest structural fragments resembles that of the (absolute) k-center problem. One interesting problem for future study is whether this problem has a PTAS or not. It is not clear how to generalize the technique employed in this paper to k-closest structural fragments problem under MS.

References

1. Jain, A. K.; Murty, M. N.; Flynn, P. J. Data clustering: a review. *ACM Computing Surveys* **1999**, *31*(3), 264–323.
2. Arun, K. S.; Huang, T. S.; Blostein, S. D. Least-squares fitting of two 3-d point sets. *IEEE Trans. Pattern Anal. Mach. Intell.* **1987**, *9*(5), 698–700.
3. Umeyama, S. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **1991**, *13*(4), 376–380.
4. Qian, J.; Li, S. C.; Bu, D.; Li, M.; Xu, J. Finding compact structural motifs. In *Combinatorial Pattern Matching, 18th Annual Symposium, CPM 2007, London, Canada, July 9-11, 2007, Proceedings*; B. Ma, K.Z. Zhang Eds.; Springer, 2007; Vol. 4580 of *Lecture Notes in Computer Science*, pp 142–149.

5. Hao, M.; Rackovsky, S.; Liwo, A.; Pincus, M.; Scheraga, H. Effects of compact volume and chain stiffness on the conformations of native proteins. *Proc. Natl. Acad. Sci.* **1992**, *89*, 6614–6618.
6. Boris, S. A revised proof of the metric properties of optimally superimposed vector sets. *Acta Crystallographica Section A* **2002**, *58*(5), 506.