AMB ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

The difficulty of protein structure alignment under the RMSD

Shuai Cheng Li

Abstract

Background: Protein structure alignment is often modeled as the *largest common point set* (LCP) problem based on the Root Mean Square Deviation (RMSD), a measure commonly used to evaluate structural similarity. In the problem, each residue is represented by the coordinate of the Cα atom, and a structure is modeled as a sequence of 3D points. Out of two such sequences, one is to find two equal-sized subsequences of the maximum length, and a bijection between the points of the subsequences which gives an RMSD within a given threshold. The problem is considered to be difficult in terms of time complexity, but the reasons for its difficulty is not well-understood. Improving this time complexity is considered important in protein structure prediction and structural comparison, where the task of comparing very numerous structures is commonly encountered.

Results: To study why the LCP problem is difficult, we define a natural variant of the problem, called the *minimum aligned distance* (MAD). In the MAD problem, the length of the subsequences to obtain is specified in the input; and instead of fulfilling a threshold, the RMSD between the points of the two subsequences is to be minimized. Our results show that the difficulty of the two problems does not lie solely in the combinatorial complexity of finding the optimal subsequences, or in the task of superimposing the structures. By placing a limit on the distance between consecutive points, and assuming that the points are specified as integral values, we show that both problems are equally difficult, in the sense that they are reducible to each other. In this case, both problems can be exactly solved in polynomial time, although the time complexity remains high.

Conclusions: We showed insights and techniques which we hope will lead to practical algorithms for the LCP problem for protein structures. The study identified two important factors in the problem's complexity: (1) The lack of a limit in the distance between the consecutive points of a structure; (2) The arbitrariness of the precision allowed in the input values. Both issues are of little practical concern for the purpose of protein structure alignment. When these factors are removed, the LCP problem is as hard as that of minimizing the RMSD (MAD problem), and can be solved exactly in polynomial time.

Keywords: Protein Structure, Alignment, RMSD, LCP

Background

A common approach to understand the properties of a protein is to compare it to other proteins. Proteins that are similar, in terms of either their amino acid sequences or 3-dimensional structures, often share similar functions, or are related evolutionarily. The latter, structural comparison, is particularly interesting since protein structures are known to be more evolutionarily conserved than the biological sequences which encode them. Furthermore, proteins of similar structures may have similar functionality, even when their sequences differ [1].

Structural comparison is typically a problem of aligning two sets of 3-dimensional coordinates. (In most of the known structural alignment problems, each point is the 3D coordinates of the Cα atom, one per residue. Hence, a structure can be modeled for structural alignment purpose as a sequence of 3D points.) The alignment usually involves a rigid transformation to superimpose the two sequences of points, and a mapping which specifies the

Correspondence: shuaicli@cityu.edu.hk
Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

matched points. The parameters to optimize in the alignment may differ in different situations, because it is not easy to single out a set of parameters that best captures the similarity between two given structures [2]. In many situations, the alignment needs not match between every point in the two sequences. At present, there is a consensus among molecular biologists in the use of the following two parameters [2-4]:

1. the number of residues (points) or percentage of total residues (points) matched in the alignment.
2. the root mean square deviation (RMSD) of the matched residues (points).

In general, the RMSD need not be minimized. It suffices that it is within a reasonable threshold. Hence, a good alignment is customarily taken to be one which maximizes the number of residue matches, within a given RMSD threshold. Many structural alignment methods are based on this principle. The computational complexity of finding an optimal solution to the problem is not well understood. Shibuya *et al.* formulated a restricted version of the problem, and showed the problem to NP-hard when the dimensionality is arbitrary. It is open whether their problem is NP-hard in 3-dimension [5]. Other problems related to structural comparison based on the RMSD have been found to be difficult. For example, the problem of finding a substructure from multiple 3-dimensional structures which minimizes the total RMSD, is NP-hard [6].

For the variants of the alignment problem that are not based on the RMSD, we have the following results. When the objective is to maximize the number of point matches which are no more than a threshold distance apart, the problem is solvable in $O(n^{32.5})$ time, where n is the number of points [7]. The *contact map overlap* problem, where a graph is created out of each structure, and the problem is one of comparing the two graphs, is NP-hard [8], and remains NP-hard even when we require points that are matchable to be within a threshold distance [9]. These results, together with an early result which shows a related problem called threading to be NP-hard [10], have traditionally led molecular biologists to believe that the structural alignment problem is difficult in general (e.g. [11-13]), even though a PTAS exists for the problem under a broad class of distance measures [14]. Heuristic algorithms have also been proposed for many variants of structural alignment problem [15-23]. While these methods perform reasonably well in general, they provide no guarantee on the quality of their results.

As noted by Shibuya *et al.*, relatively few theoretical results have been obtained on problems defined over the RMSD, and the general problem of structural alignment under the RMSD remains open [5]. At present, whether the problem is intractable or not is not only of theoretical interests but also of practical concerns, due to advances in protein structure prediction which requires the comparison of very numerous structures. In this paper we show mathematical insights and techniques which we hope will lead to practical algorithms for the problem.

We first show that the difficulty of the problem does not lie solely in the individual components of their requirement. More precisely,

- if either a mapping that contains the optimal mapping is known (Theorem 3), or
- if the optimal superposition is known (Lemma 1),

then the problem can be solved in polynomial time.

Our study shows that the difficulty of the LCP problem is also very much due to the two factors: (1) the problem allows the input coordinates to be of any arbitrary precision, and (2) it assumes no limit on the distance between two consecutive Cα atoms.

We consider the case where the input coordinates are integral, and the distance between two consecutive points is restricted. The first requirement is practical since in protein structures, coordinates are typically specified to a fixed precision (e.g. three decimal places in protein structures [24]), and can be trivially scaled up to integral values. Similar assumptions are made in Euclidean problems such as the Euclidean TSP [25]. The second requirement likewise does not add any restriction to the problem of protein structure alignment, since there is a natural upper bound ($\sim$3.8Å) to the distance between two Cα atoms. In this case, the following results hold.

- Given a polynomial time algorithm for finding a largest alignment of RMSD below a threshold d, one can efficiently compute an alignment of a given size ℓ which minimizes the RMSD (Theorem 7). (Since the other direction is easy, this shows that the two problems are of similar difficulty.)
- The structural alignment problem under the RMSD is solvable exactly in polynomial time (Theorem 10).

Preliminary

Notations and definitions

A protein structure for alignment purpose is modeled as a finite, ordered sequence of three dimensional (3D) points. Hence, a structure of n residues is written as $(p_1, p_2, ..., p_n)$, where each point $p_i \in \mathbb{R}^3$. In the 'Results assuming integral coordinates and restricting distance between points' section, we will further assume each p_i to be integral. We write $P' \subseteq P$ iff P' is a subsequence of P, and write $f : P \mapsto Q$ iff f is a mapping which maps points in the sequence P to points in the sequence Q.

Problem statements

We now state our problems. The main problem we consider is the largest common point set (LCP) problem under the RMSD, a well-known problem in protein structure alignment. In the LCP, the objective is to find a mapping of the largest cardinality where the RMSD of the matched points is no more than a given threshold (Table 1).

We do not require the optimal superposition of P and Q in the output, since that can be computed from P', Q', and f in linear time [26]. We refer to f as an *alignment*. An alignment can be *sequential* or *non-sequential*: an alignment is *sequential* iff for any two points p_{i_1}, $p_{i_2} \in P'$, where the corresponding $f(p_{i_1}) = q_{j_1}$ and $f(p_{i_2}) = q_{j_2}$, we have $i_1 < i_2$ iff $j_1 < j_2$. Otherwise the alignment is *non-sequential*. The LCP problem which requires alignments to be sequential is said to be *sequential*, otherwise it is *non-sequential*. We mainly discuss sequential alignment in this paper. The techniques developed can be easily adapted to the non-sequential case. Given two equal-sized sequences $P' = (p_1, \ldots, p_n)$ and Q', together with a bijection f between P' and Q', the root mean square deviation (RMSD) is defined as

$$RMSD(P', Q') = \min_t \sqrt{\frac{\sum_{1 \le i \le n} ||t(f(p_i)) - p_i||^2}{n}}, \quad (1)$$

where t is a rigid transformation. The RMSD, with its corresponding transformation t, can be computed in linear time [26].

A natural variant of the LCP problem is to minimize the RMSD instead of maximizing the size of the mapping, as follows. Given an integer ℓ, find subsequences of size ℓ of the input, such that the RMSD between the points of the subsequences is minimized. We call this problem the minimum aligned distance (MAD) problem (Table 2).

Clearly, if the MAD problem is solvable in polynomial time, then the LCP problem is solvable in polynomial time. However, the other direction is unclear. Theorem 7 will show that for P and Q of integral coordinates, if the LCP problem is solvable in polynomial time, then the MAD problem is solvable in polynomial time.

Table 1 Largest Common Point (LCP) set problem definition under RMSD

	LCP problem by the RMSD measure
Input:	sequences $P = (p_1, \ldots, p_n)$, $Q = (q_1, \ldots, q_m)$ and distance threshold $\theta \in \mathbb{R}$. Without loss of generality assume $m \ge n$.
Output:	(i) subsequences $P' \subseteq P$, $Q' \subseteq Q$, $\|P'\| = \|Q'\|$, and (ii) bijection $f : P' \mapsto Q'$, fulfilling the following conditions: (A) $RMSD(P', f(P')) \le \theta$, (B) the score $l = \|P'\|$ is maximized.

Table 2 Minimum Aligned Distance (MAD) problem definition under RMSD

	MAD problem by the RMSD measure
Input:	sequences $P = (p_1, \ldots, p_n)$, $Q = (q_1, \ldots, q_m)$ and $\ell \in \mathbb{I}$. Without loss of generality assume $m \ge n$.
Output:	(i) subsequences $P' \subseteq P$, $Q' \subseteq Q$, $\|P'\| = \|Q'\|$, and (ii) bijection $f : P' \mapsto Q'$, fulfilling the following conditions: (A) $\|P'\| = \ell$, (B) $d = RMSD(P', f(p'))$ is minimized.

We let $\hat{\mathbf{P}}$, $\hat{\mathbf{Q}}$, $\hat{\mathbf{f}}$, ℓ and d_{opt} denote an optimal P', Q', f, l and d, respectively. The optimal rigid transformation for superimposing $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ is denoted $\mathcal{T}$, and can be computed from $\hat{\mathbf{P}}$, $\hat{\mathbf{Q}}$ and $\hat{\mathbf{f}}$. The symbol c_P^{max} denotes the largest value in the coordinates of P, c_Q^{max} the largest value in the coordinates of Q, and $c_{max} = \max\{c_P^{max}, c_Q^{max}\}$, and we know that $c_{max} = O(n)$ for protein structures.

Results for general LCP and MAD

Complexity of the LCP and MAD when the optimal superposition is known

Since two point sequences with a known mapping can be superimposed optimally under the RMSD in linear time [26], it is natural to ask if the difficulty in LCP or MAD lies solely in the combinatorial complexity of finding the optimal subsets, i.e. $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$. Our results show the contrary: if the optimal superposition $\mathcal{T}$ is known, both problems can be solved in polynomial time.

We first consider the sequential case. Let $d_{p,q} = ||t(p) - q||^2$ and let $M[i, j; k]$ denote the minimum squared sum cost of k pair matches for the point sets $(p_1, p_2, ..., p_i)$ and $(q_1, q_2, ..., q_j)$. If $1 \le k \le \ell$, $2 \le i \le m$ and $2 \le j \le n$, we have a recurrence relation of

$$M[i, j; k] = \max \left\{ \begin{array}{l} M[i-1, j-1; k-1] + d_{p_i, q_j}, \\ M[i, j-1; k], \\ M[i-1, j; k] \end{array} \right\} \quad (2)$$

The base case of the recursion is obvious. Dynamic programming can be employed to fill up the respective $M[i, j; k]$ values. After all the values are filled, one can find the maximum k, such that the squared sum is no more than $k\theta^2$ for the LCP problem. The MAD problem can be solved similarly.

The non-sequential case can be similarly solved using the maximum-flow minimum-cost problem [27]. The following lemma states these results.

Lemma 1. *If an optimal transformation $\mathcal{T}$ is known, both the LCP problem and the MAD problem can be solved in $O(mn\ell)$ time.*

Complexity of the LCP and MAD when the matching between the point sets is known

We next ask if the difficulty in the LCP and MAD could be due to the task of superimposing P and Q in an optimal manner to identify the subsets $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$. To examine this possibility, we remove the combinatorial task of examining each of the possible mapping between the points, by assuming a bijection F which contains the optimal mapping. (Note that this results in the problem known as *model superposition* in structural biology.) Again, our results show the resultant LCP and MAD problems to be solvable exactly in polynomial time.

Assume that F is a bijection that maps points in P to points in Q. Let $P' = (p'_1, \ldots, p'_l)$ be the domain of F and $Q' = (q'_1, \ldots, q'_l)$ be the range of F. Without loss of generality assume that F is sequential, and hence $F(p'_i) = q'_i$.

One can exhaustively evaluate all the subsequences of P' for one with the least RMSD. However, since the number of such subsequences is exponential in l, this does not immediately give us a polynomial time solution.

If a rigid transformation T for Q' is given, and all the pairs (p'_i, q'_i), $1 \leq i \leq l$ are sorted according to the value $||T(p'_i) - q'_i||$, the MAD problem is then to choose the first ℓ pairs from (p'_i, q'_i), $1 \leq i \leq l$, and the LCP problem is to choose the first ℓ pairs, such that $RMSD((p'_1, ..., p'_\ell), (q'_1, ..., q'_\ell)) \leq \theta$ and $\ell = l$ or $RMSD((p'_1, ..., p'_\ell, p'_{\ell+1}), (q'_1, ..., q'_\ell, q'_{\ell+1})) > \theta$. This gives us an incentive to obtain a total ordering of $||T(p'_i) - q'_i||$, which will allow us to solve the MAD problem by selecting the first ℓ pairs in the ordering. The set of transformations which produce the same total according to $||T(p'_i) - q'_i||$, yield the same result for the MAD problem, and therefore these transformations are equivalent. This enables us to design a discrete version of the problem.

For clarity, we first present an algorithm with only translation.

With translations only

Consider two pairs (p'_i, q'_i) and (p'_j, q'_j). The transformations T to separate the two types of transformations that $||T_1(q'_i) - p'_i|| > ||T_1(p'_j) - q'_j||$ and $||T_2(q'_i) - p'_i|| < ||T_2(p'_j) - q'_j||$ are the transformations where

$$||T(q'_i) - p'_i||^2 - ||T(p'_j) - q'_j||^2 = 0. \quad (3)$$

Let $\bullet$ denote dot product. If the transformation is a translation t, we have

$$\begin{aligned} & ||T(q'_i) - p'_i||^2 - ||T(q'_j) - p'_j||^2 \\ = \; & ||q'_i - p'_i - t||^2 - ||q'_j - p'_j - t||^2 \\ = \; & \sum_{v=x,y,z} (v_{q'_i} - v_{p'_i} - v_t)^2 - \sum_{v=x,y,z} (v_{q'_j} - v_{p'_j} - v_t)^2 \\ = \; & ||q'_i - p'_i||^2 - ||q'_j - p'_j||^2 \\ & -2t \bullet ((q'_i - p'_i) - (q'_j - p'_j)) = 0 \end{aligned} \quad (4)$$

Consider the space of all translation vectors in $\mathbb{R}^3$, and consider each vector as a point in this space (not the space that P and Q are in). The values that the variable t in Equation 4 may take form a plane in this translation space. The plane partitions the translation space into two types of translations, T_1 and T_2 say, where $||T_1(q'_i) - p'_i|| > ||T_1(p'_j) - q'_j||$ and $||T_2(q'_i) - p'_i|| < ||T_2(p'_j) - q'_j||$. Since there are l pairs, there are $O(l)$ planes, which partition the space into $O(l^3)$ cells.

The translations in each cell result in the same ordering of the pairs with respect to $||T(p'_i) - q'_i||$. For each cell, this total order can be obtained in $O(l)$ time from any given total order of its neighbor cells, since the change is $O(1)$. Therefore, the MAD solution can be obtained in amortized time $O(l)$ for each cell, and the LCP solutions thus can be obtained in time $O(l^2)$. Hence the total runtime is of $O(\ell^4)$ for the MAD problem, and $O(\ell^5)$ for the LCP problem of translations, with the given mapping F.

With rigid transformations

The rigid transformations which separate the two relations $||T_1(q'_i) - p'_i|| > ||T_1(p'_j) - q'_j||$ and $||T_2(q'_i) - p'_i|| < ||T_2(p'_j) - q'_j||$ are as in Equation 3.

Suppose the rigid transformations T is composed of a rotation R and a translation t.

$$\begin{aligned} & ||T(q'_i) - p'_i||^2 - ||T(q'_j) - p'_j||^2 \\ = \; & ||R(q'_i) - p'_i - t||^2 - ||R(q'_j) - p'_j - t||^2 \\ = \; & ||R(q'_i) - p'_i||^2 - ||R(q'_j) - p'_j||^2 \\ & -2t \bullet [\,(R(q_i) - p_i) - (R(q_j) - p_j)] \end{aligned} \quad (5)$$

A rotation matrix contains three variables, which is specified using three angles, say $\alpha_1, \alpha_2, \alpha_3$, each from $-\pi$ to π. Let $r_i = \cos\alpha_i$, $s_i = \cos\alpha_i$, then Equation 5 can be considered as a polynomial of nine variables in degree six. The nine variables are r_i, s_i and the three variables for translation. In total, there are $O(l)$ such polynomials.

We know the following theorem from the literature.

Theorem 2. *[7,28] Given a set of k polynomials, $\mathcal{P} = \{f_1, ..., f_k\}$, where each polynomial has a maximum degree of s, contains at most r variables, and in addition all the coefficients are rational, then all the sign conditions can be determined by $O(k(k/r)^r s^{O(r)})$ arithmetic operations. A sign condition V is the vector of signs for some point $u \in \mathbb{R}^k$; that is, $V = (sign(f_1(u)), ..., sign(f_k(u)))$. Two points $u, u' \in \mathbb{R}^k$ are equivalent if their sign condition vectors are the same.*

Each sign vector represents the transformations of the cell it belongs to, and it determines a total order of the pairs. Similar as in the case of translation, with Theorem 2, we have

Theorem 3. *Given a bijection $F : P' \to Q'$, where $|P'| = l$, then the MAD problem can be solved in $O(l^{10})$ time and the LCP problem can be solved in $O(l^{11})$ time.*

Results assuming integral coordinates and restricting distance between points

One possible contributing factor to the difficulty of the LCP problem could be its flexibility in allowing input coordinates of any arbitrary precision. This is because intuitively, this arbitrariness in the precision introduces the burden of examining the solution space in an unbounded manner. However, such an exhaustive search is not necessary for the purpose of protein structure comparison, since coordinates of protein structures are specified only to three decimal places in the commonly used PDB format.

In this section, we restrict the precision in which the input coordinates may be specified. Without loss of generality, we assume that the input coordinates are given in integers, since numbers of any fixed precision can be trivially scaled up to integral values. This assumption is used to obtain Lemma 5 and Theorem 7.

We also place an upper bound on the distance between consecutive points according to the structure of proteins. As a result, c_{max} is bounded by n, as follows.

Points drawn from protein structures have upper bounds on their diameters because they are connected, and many are globular.

- For a connected structure, the points are at most $O(n)$ distance apart. That is, c_{max} is of $O(n)$.
- For a globular structure, the points are at most $O(n^{1/3})$ distance apart [14]. That is, c_{max} is of $O(n^{1/3})$.

Given a point p, let the x coordinate of p be denoted x_p. Similarly, we can define y_p and z_p. Without loss of generality, we assume that the first point of a protein structure is at the origin. The largest coordinate of a protein is bounded by $O(n)$, and the largest coordinate of a globular protein structure is bounded by $O(n^{1/3})$; that is

$$\max_{p\in P, v=x,y,z} |v_p| = O(n), \text{if P is a protein structure} \quad (6)$$

$$\max_{p\in P, v=x,y,z} |v_p| = O(n^{1/3}), \text{if P is a globular protein structure} \quad (7)$$

Our results show that, under these two conditions,

1. the LCP problem is of similar difficulty as the MAD problem, and
2. both problems can be solved exactly in polynomial time.

Properties of protein structures

Upper and lower bounds of RMSD

We first establish some bounds to the RMSD. The minimum RMSD is zero if $\mathcal{T}$ brings $\hat{\mathbf{Q}}$ to coincide exactly with $\hat{\mathbf{P}}$. This case is referred to as the exact matching, which can be easily solved by the method in [29]. However, if we assume the RMSD to be non-zero, then a lower bound and an upper bound for it can be computed.

Let π be a permutation of $\{1, ..., \ell\}$. For the sequence $X = (x_1, \ldots, x_n)$, let $d^X_{i,j}$, $1 \le i, j \le n$, denote the Euclidean distance between x_i and x_j. The following results, which are proven in the Appendix, can be obtained.

Lemma 4.

$$\frac{1}{\sqrt{2\ell}} \sqrt{\sum_{i=1}^{\lfloor \ell/2 \rfloor} |d^{\hat{P}}_{\pi(i),\pi(i+\lfloor \ell/2 \rfloor)} - d^{\hat{Q}}_{\pi(i),\pi(i+\lfloor \ell/2 \rfloor)}|^2} \le RMSD(\hat{P}, \hat{Q}).$$

Lemma 5 (Lower bound). *If $RMSD(\hat{P}, \hat{Q}) \neq 0$, then $RMSD(\hat{P}, \hat{Q}) \ge \frac{\sqrt{12c_{max}^2} - \sqrt{12c_{max}^2 - 1}}{\sqrt{2\ell}}$.*

Lemma 6 (Upper bound). *$RMSD(\hat{P}, \hat{Q}) \le 4\sqrt{3}c_{max}$.*

Using an algorithm for the LCP to solve the MAD problem

Suppose there is a polynomial time algorithm for solving the LCP problem. To use it to solve the MAD problem, we assume that $d_{opt} \in [l, u]$, for some real l and u, $l \le u$. We use a binary search strategy in the interval $[l, u]$, as shown in Table 3, to search for the minimum

Table 3 Employing an algorithm for the LCP problem to solve the MAD problem

Input:	sequences $P = (p_1, \ldots, p_n)$, $Q = (q_1, \ldots, q_m)$ and $\ell \in \mathbb{I}$.
	Without loss of generality assume $m \ge n$.
Output:	(i) subsequences $P' \subseteq P$, $Q' \subseteq Q$, $\|P'\| = \|Q'\|$, and
	(ii) mapping $f : P' \mapsto Q'$, fulfilling the following conditions:
	(A) $\|P'\| = \ell$,
	(B) $d = RMSD(P', f(p'))$ is minimized.
1.	$l \leftarrow 0, u \leftarrow \ell c_{max}$
2.	$m \leftarrow 1/2(l + u)$
3.	Call LCP to solve the instance (P, Q, m).
4.	If the LCP solution has size no less than ℓ
	$u \leftarrow m$
	else
	$l \leftarrow m$
5.	If $u - l \le \frac{\sqrt{12c_{max}^2} - \sqrt{12c_{max}^2 - 1}}{\sqrt{2\ell}}$,
	Output the most recent LCP solution of size no less than ℓ.
	Otherwise, repeat Steps 2-5.

value such that the LCP solution size is ℓ. However, the search will not terminate if an arbitrary accuracy of the d_{opt} value is required. We prove below that the accuracy of d_{opt} can be defined by polynomially many bits. Given two threshold t_1 and t_2, assume that we obtain two different LCP solutions, and the RMSD values of the two solutions are θ_1 and θ_2, where $\theta_1 > \theta_2$. Similar to the arguments in Lemma 5, the difference between θ_1 and θ_2 is at least $\frac{\sqrt{12c_{max}^2}-\sqrt{12c_{max}^2-1}}{\sqrt{2\ell}}$. Therefore if two consecutive binary search operators have the difference of the threshold values below $\frac{\sqrt{12c_{max}^2}-\sqrt{12c_{max}^2-1}}{\sqrt{2\ell}}$, the search can be terminated. The values of l and u are the same as in the previous subsection. Hence,

Theorem 7. *Solving the MAD problem is equivalent to solving $O(\log \ell c_{max})$ instances of the LCP problem.*

Since the reduction from the LCP problem to the MAD problem is obvious, we conclude that the two problems are of similar difficulty.

Polynomial time algorithm

We now show that under the two conditions, the LCP and MAD problem can be solved in polynomial time.

As shown in the 'Complexity of the lCP and mAD when the optimal superposition is known' section, when the optimal superposition is known, there are polynomial time algorithms for LCP and MAD. We consider an enumeration of all the possible superpositions. Under the two conditions, we claim that there are at most polynomially many such superpositions.

First, if we know $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$, then optimal superposition can be computed in the following two steps:

1. Translate $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ such that their centroids are at the origin.
2. Then, rotate $\hat{\mathbf{Q}}$ to find the superposition with the minimum distance [26].

Denote the translations to obtain the optimal solution for P and Q as t_P and t_Q, respectively, and denote the optimal rotation by $\hat{\mathbf{R}}$.

We now show that one needs to examine only polynomially many translations and rotation combinations to discover the values for t_P, t_Q, and $\hat{\mathbf{R}}$. These numbers can be effectively bounded by n when properties of protein structures are taken into account. We first describe these properties.

Number of translations

The centroid of $\hat{\mathbf{P}}$ is $\frac{\sum_{p' \in \hat{\mathbf{P}}} p'}{\ell}$. To bring $\hat{\mathbf{P}}$ to origin, the translation is $-\frac{\sum_{p' \in \hat{\mathbf{P}}} p'}{\ell}$. Clearly, all the three coordinates of $-\sum_{p' \in \hat{\mathbf{P}}} p'$ are integers since all the coordinates of the points in $p \in P$ are integers. The value of x-coordinate of $-\frac{\sum_{p' \in \hat{\mathbf{P}}} p'}{\ell}$ is bounded by $-\frac{\sum_{p' \in \hat{\mathbf{P}}} x_{p'}}{\ell} \leq \frac{\sum_{p' \in \hat{\mathbf{P}}} c_p^{max}}{\ell} \leq c_{max}$. Similarly, all the three coordinates of the translation $-\frac{\sum_{p' \in \hat{\mathbf{P}}} p'}{\ell}$ are bounded within the interval $[-c_{max}, c_{max}]$. To obtain an optimal MAD solution, the translation on $\hat{\mathbf{P}}$ must be in the form of $\frac{I}{\ell}$, where I is an integer. Since it is possible to examine all the possible values for I, we have the following result.

Lemma 8. $t_P, t_Q \in \{I/\ell | -\ell c_{max} \leq I \leq \ell c_{max}\}^3$.

Number of rotations

With the centroids of $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ translated to the origin, we proceed to identify the rotation in our algorithm. Let X_P denote the vector $\langle x_{p_1}, ..., x_{p_\ell} \rangle$ for structure P. Similarly we define Y_P and Z_P.

Let $\hat{\mathbf{P}}_t = (p'_1 - t, ..., p'_\ell - t)$ and $\hat{\mathbf{Q}}_t = (q'_1 - t, ..., q'_\ell - t)$, $p'_i \in \hat{\mathbf{P}}$, $q'_i \in \hat{\mathbf{Q}}$.

Given $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$, to compute the rotation $\hat{\mathbf{R}}$, the first step is to create the 3×3 matrix, which is (from [26])

$$M = \begin{pmatrix} X_{\hat{\mathbf{P}}_{t_P}} \bullet X_{\hat{\mathbf{Q}}_{t_Q}} & X_{\hat{\mathbf{P}}_{t_P}} \bullet Y_{\hat{\mathbf{Q}}_{t_Q}} & X_{\hat{\mathbf{P}}_{t_P}} \bullet Z_{\hat{\mathbf{Q}}_{t_Q}} \\ Y_{\hat{\mathbf{P}}_{t_P}} \bullet X_{\hat{\mathbf{Q}}_{t_Q}} & Y_{\hat{\mathbf{P}}_{t_P}} \bullet Y_{\hat{\mathbf{Q}}_{t_Q}} & Y_{\hat{\mathbf{P}}_{t_P}} \bullet Z_{\hat{\mathbf{Q}}_{t_Q}} \\ Z_{\hat{\mathbf{P}}_{t_P}} \bullet X_{\hat{\mathbf{Q}}_{t_Q}} & Z_{\hat{\mathbf{P}}_{t_P}} \bullet Y_{\hat{\mathbf{Q}}_{t_Q}} & Z_{\hat{\mathbf{P}}_{t_P}} \bullet Z_{\hat{\mathbf{Q}}_{t_Q}} \end{pmatrix} \quad (8)$$

Each above matrix is decomposed by the singular value decomposition, and a rotation matrix is produced hereafter.

We know that the coordinate of each point in the protein is within the interval $[-c_{max}, c_{max}]$. This implies that for $U = X, Y, Z$ and $V = X, Y, Z$,

$$U_{\hat{\mathbf{P}}_{t_P}} \bullet V_{\hat{\mathbf{Q}}_{t_Q}} = \sum_{k=1}^{\ell} (p_{i,k} - t_P)(q_{j,k} - t_Q) \leq \sum_{k=1}^{\ell} (2c_{max})^2 \leq 4\ell c_{max}^2.$$

Also, it is clear that $U_{\hat{\mathbf{P}}_{t_P}} \bullet V_{\hat{\mathbf{Q}}_{t_Q}}$ is in the form of I/ℓ^2, where I is an integer. The matrix in Equation 8 has nine elements; we denote $e \in M$ if e is one of these elements. The following lemma follows.

Lemma 9. *For each element $e \in M$ in Equation 8, $e \in \{I/\ell^2 | -4\ell^3 c_{max}^2 \leq I \leq 4\ell^3 c_{max}^2\}$.*

Polynomial time algorithm

To compute the optimal MAD solution, we first enumerate all the possible translations and rotations. A solution is computed for each translation and rotation combination according to Lemma 1. An optimal solution can be chosen from these computed solutions (Table 4).

According to Lemma 8, the optimal translation t_P and t_Q must be within $\{I/\ell | -\ell c_{max} \leq I \leq \ell c_{max}\}^3$. To find the

Table 4 A polynomial time algorithm for the MAD problem

Input:	sequences $P = (p_1, \ldots, p_n)$, $Q = (q_1, \ldots, q_m)$ and $\ell \in \mathbb{I}$. Without loss of generality assume $m \geq n$.
Output:	(i) subsets $P' \subseteq P$, $Q' \subseteq Q$, $\lvert P' \rvert = \lvert Q' \rvert$, and (ii) mapping $f : P' \mapsto Q'$, fulfilling the following conditions: (A) $\lvert P' \rvert = \ell$, (B) $d = RMSD(P', f(p'))$ is minimized.
1.	For each translation $t \in \{l/\ell \rvert - \ell c_{max} \leq l \leq \ell c_{max}\}^3$, For each 3×3 matrix M, where $\forall e \in M, e \in \{l/\ell^2 \rvert -, \{4\ell^3 c_{max}^2 \leq l \leq 4\ell^3 c_{max}^2\}$ Compute rotation matrix R from M. $\mathbb{Q} \leftarrow RQ - t$. Apply an algorithm for the case where the superposition is known to P and $\mathbb{Q}$ (as discussed in the 'Complexity Of The LCP And MAD When The Optimal Superposition Is Known' section), and denote the solution MAD(P, $\mathbb{Q}$).
2.	Output the MAD(P, $\mathbb{Q}$) of the smallest *RMSD* as the solution.

optimal rotation matrix, it suffices that we try all the possible values for each entry in Equation 8. Since there are $\ell^{27} c_{max}^{18}$ matrices, the number of total transformations to examine is bounded by $O(\ell^{33} c_{max}^{34})$. It takes time $O(mn\ell)$ to identify the MAD solution for each transformation. An LCP solution can be obtained by iterating ℓ from 1 to $\min\{m, n\}$ for the MAD problem.

The running time consists of the productions of three parts: the number of possible translations, the number of possible rotation matrix, and the running time for given a rotation matrix and a translation combination (that is, the running time when then transformation is known). These numbers are bounded by c_{max}, which is bounded by m when we consider the properties of protein structures. Likewise, c_{max} is polynomial with respect to the input size if coded in unary.

Theorem 10. *The MAD problem can be solved in $O(\ell^{34} m^{25} n)$ time for protein structures, and in $O(\ell^{34} m^9 n)$ time for globular protein structures. The LCP problem can be solved in $O(\ell^{35} m^{25} n)$ time for protein structures, and in $O(\ell^{35} m^9 n)$ time for globular protein structures. Both the MAD and LCP problems are pseudo-polynomially solvable for general point sets.*

Conclusions

We studied the LCP problem under the RMSD in this paper. As it turns out, the difficulty of the problem does not lie in its combinatoric aspect or its structural superposition aspect alone. That is, if the problem is hard, then it must be a consequence of both aspects. Our results show that if one is allowed to compromise on one of the aspects, then the problem is solvable exactly in polynomial time. Regrettably, we do not see how the optimal solution can be obtained in both cases.

On the other hand, we showed an encouraging result: There is a polynomial time algorithm which solves the problem optimally, if one restricts the input coordinates in the problem to be integral, and places a limit on the distance between consecutive points. These requirements do not pose any restriction to typical uses in the analysis of protein structures, since protein structures are specified only to a fixed precision in practice, and there is an upper bound to the distance between protein residues.

One problem is that our proposed polynomial time algorithm remains high in time complexity. We hope that the present work will provide the foundation for future efforts to obtain algorithms with lower runtime complexities.

Appendix

In this Appendix, we include the proofs of the results in the paper which have been omitted to enhance readability.

Lemma 4.

$$\frac{1}{\sqrt{2\ell}} \sqrt{\sum_{i=1}^{\lfloor \ell/2 \rfloor} |d^{\hat{\mathbf{P}}}_{\pi(i),\pi(i+\lfloor \ell/2 \rfloor)} - d^{\hat{\mathbf{Q}}}_{\pi(i),\pi(i+\lfloor \ell/2 \rfloor)}|^2} \leq RMSD(\hat{\mathbf{P}}, \hat{\mathbf{Q}}).$$

Proof. Without loss of generality, we just show that $\frac{1}{\sqrt{2\ell}} \sqrt{\sum_{i=1}^{\lfloor \ell/2 \rfloor} |d^{\hat{\mathbf{P}}}_{i,i+\lfloor \ell/2 \rfloor} - d^{\hat{\mathbf{Q}}}_{i,i+\lfloor \ell/2 \rfloor}|^2} \leq RMSD(\hat{\mathbf{P}}, \hat{\mathbf{Q}})$.
Let
$r_i = ||\mathcal{T}(q_i) - p_i||^2 + ||\mathcal{T}(q_{i+\lfloor \ell/2 \rfloor}) - p_{i+\lfloor \ell/2 \rfloor}||^2$,
$u_i = \langle p_i, p_{i+\lfloor \ell/2 \rfloor} \rangle$, and
$v_i = \langle q_i, q_{i+\lfloor \ell/2 \rfloor} \rangle$, where $1 \leq i \leq \lfloor \ell/2 \rfloor$.

First, we prove that $r_i \geq |u_i - v_i|^2/2$, for $1 \leq i \leq \lfloor \ell/2 \rfloor$. We first superimpose u_i and v_i to optimize the squared sum; that is, to find transformation T such that $||T(q_i) - p_i||^2 + ||T(q_{i+\lfloor \ell/2 \rfloor}) - p_{i+\lfloor \ell/2 \rfloor}||^2$ is minimized. The centroids have to coincide to minimize the squared distance. Assume the centroids are at the origin and that the angle between $\overrightarrow{(o, p_i)}$ and $\overrightarrow{(o, q_i)}$ is α, where o is the origin, then by trigonometry

$$\begin{aligned} &||T(q_i) - p_i||^2 + ||T(q_{i+\lfloor \ell/2 \rfloor}) - p_{i+\lfloor \ell/2 \rfloor}||^2 \\ &= 2[\,(1/2 \times ||u_i||)^2 + (1/2 \times ||v_i||)^2 \\ &\qquad - 2 \times 1/2||u_i|| \times 1/2||v_i|| \times \cos\alpha] \\ &\geq \frac{(||u_i|| - ||v_i||)^2}{2} \end{aligned}$$

r_i is the squared distance under transformation $\mathcal{T}$, which may not be optimal for superimposing u_i and v_i. Therefore,

$$r_i \geq \frac{(||u_i|| - ||v_i||)^2}{2}$$

Putting things together, we have

$$\begin{aligned} RMSD(\hat{\mathbf{P}}, \hat{\mathbf{Q}}) &\geq \sqrt{\frac{r_1 + \ldots + r_{\lfloor \ell/2 \rfloor}}{\ell}} \\ &\geq \sqrt{\frac{(u_1 - v_1)^2 + \ldots + (u_{\lfloor \ell/2 \rfloor} - v_{\lfloor \ell/2 \rfloor})^2}{2\ell}} \\ &\geq \frac{1}{\sqrt{2\ell}} \sqrt{\sum_{i=1}^{\lfloor \ell/2 \rfloor} |d^{\hat{\mathbf{P}}}_{i,i+\lfloor \ell/2 \rfloor} - d^{\hat{\mathbf{Q}}}_{i,i+\lfloor \ell/2 \rfloor}|^2} \end{aligned}$$

□

Lemma 5. *If* $RMSD(\hat{\boldsymbol{P}}, \hat{\boldsymbol{Q}}) \neq 0$, *then* $RMSD(\hat{\boldsymbol{P}}, \hat{\boldsymbol{Q}}) \geq \frac{\sqrt{12c^2_{max}} - \sqrt{12c^2_{max} - 1}}{\sqrt{2\ell}}$.

Proof. If RMSD($\hat{\mathbf{P}}, \hat{\mathbf{Q}}$) is non-zero, then there is at least a pair of indices i and j, such that $||d^{\hat{\mathbf{P}}}_{i,j} - d^{\hat{\mathbf{Q}}}_{i,j}|| > 0$. According to Lemma 4,

$$\begin{aligned} RMSD(\hat{\mathbf{P}}, \hat{\mathbf{Q}}) &\geq \frac{1}{\sqrt{2\ell}} \sqrt{|d^{\hat{\mathbf{P}}}_{i,j} - d^{\hat{\mathbf{Q}}}_{i,j}|^2} = \frac{|d^{\hat{\mathbf{P}}}_{i,j} - d^{\hat{\mathbf{Q}}}_{i,j}|}{\sqrt{2\ell}} \\ &>= \frac{\sqrt{12c^2_{max}} - \sqrt{12c^2_{max} - 1}}{\sqrt{2\ell}}. \end{aligned}$$

□

Lemma 6. $RMSD(\hat{\boldsymbol{P}}, \hat{\boldsymbol{Q}}) \leq 4\sqrt{3}c_{max}$.

Proof. Denote the furthest point to the origin in P as p_{max}, and the furthest point to the origin in Q as q_{max}. Then,

$$\begin{aligned} &\ell \times RMSD^2(\hat{\mathbf{P}}, \hat{\mathbf{Q}}) \\ &= \sum_{i=1}^{\ell} ||\mathcal{T}(q_i) - p_i||^2 \\ &\leq \sum_{i=1}^{\ell} (||q_i - p_i||)^2 \\ &\leq \sum_{i=1}^{\ell} \max\{||q_{max} - p_{max}||^2, ||q_{max} + p_{max}||^2\} \\ &\leq \ell \max\{||q_{max} - p_{max}||^2, ||q_{max} + p_{max}||^2\} \\ &\leq \ell(2\sqrt{(c_{max}+c_{max})^2 + (c_{max}+c_{max})^2 + (c_{max}+c_{max})^2})^2 \\ &= 48\ell c^2_{max} \end{aligned}$$

□

Abbreviations
LCP: Largest Common Point set; RMSD: Root Mean Square Deviation; MAD: Minimum Aligned Distance.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SCL conceived of the results herein and wrote the manuscript.

Acknowledgements
This work is supported by the CityU SRG 7002731. The author thanks the useful discussions with Prof. Richard M. Karp

Received: 8 December 2011 Accepted: 17 December 2012
Published: 4 January 2013

References

1. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G: **North ACT: Structure of myoglobin: a three-dimensional Fourier synthesis at 5.5 Angstrom resolution.** *Nature* 1960, **185:**416–422.
2. R Kolodny PK, Levitt M: **Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures.** *J Comput Biol* 2005, **346**(4):1173–1188.
3. Sippl MJ: **On distance and similarity in fold space.** *Bioinformatics* 2008, **24**(6):872–873.
4. V A Ilyin CML A Abyzov: **Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point.** *Protein Sci* 2004, **13**(7):1865–1874.
5. Shibuya T, Jansso J, Sadakane K: **Linear-time protein 3-D structure searching with insertions and deletions.** *Algorithms Mol Biol* 2010, **5**(7):1–8.
6. Bu D, Li M, Li SC, Qian J, Xu J: **Finding compact structural motifs.** *Theor Comput Sci* 2009, **410:**2834–2839.
7. Ambühl C, Chakraborty S, Gärtner B: **Computing Largest Common Point Sets under Approximate Congruence.** In *ESA '00: Proceedings of the 8th Annual European Symposium on Algorithms, Volume 1876 of Lecture Notes in Computer Science*. Saarbrücken, Germany: Springer-Verlag ; 2000:52–63.
8. Goldman D, Papadimitriou CH, Istrail S: *Algorithmic Aspects of Protein Structure Similarity*. New York City, NY, USA: IEEE Computer Society; 1999.
9. Li SC, Ng YK: **On protein structure alignment under distance constraint.** In *20th International Symposium on Algorithms and Computation, ISAAC 2009, Proceeedings, Volume 5878 of Lecture Notes in Computer Science*. Honolulu, HI, USA: Springer; 2009:65–76.
10. Lathrop RH: **The Protein Threading Problem With Sequence Amino Acid Interaction Preferences Is NP-Complete.** *Protein Eng* 1995, **7:**1059–1068.
11. Godzik A: **The structural alignment between two proteins: is there a unique answer?** *Protein Sci* 1996, **5**(7):1325–1338.
12. Eidhammer I, Jonassen I, Taylor WR: **Structure Comparison and Structure Patterns.** *J Comput Biol* 2000, **7**(5):685–716.
13. Zhao Z, Fu B, Alanis FJ, Summa CM: **Feedback algorithm and web-server for protein structure alignment.** *J Comput Biol* 2008, **15**(5):505–524.
14. Kolodny R, Linial N: **Approximate Protein Structural Alignment in Polynomial Time.** *Proc Natl Acad Sci* 2004, **101:**12201–12206.
15. Akutsu T, Tashimo H: **Protein structure comparison using representation by line segment sequences.** In *Proc. Pacific Symposium on Biocomputing '96*. Hawaii, USA: World Scientific Press; 1996:25–40.
16. Alexandrov NN: **SARFing the PDB.** *Protein Eng* 1996, **9**(9):727–732.
17. Caprara A, Lancia G: **Structural alignment of large-size proteins via lagrangian relaxation.** In *RECOMB '02: Proceedings of The Sixth Annual International Conference on Computational Biology*. New York, NY, USA: ACM; 2002:100–108.
18. Comin M, Guerra C, Zanotti G: **PROuST: A Comparison Method of Three-Dimensional Structure of Proteins using Indexing Techniques.** *J Comp Biol* 2004, **11:**1061–1072.
19. Gerstein M, Levitt M: **Using Iterative Dynamic Programming to Obtain Accurate Pairwise and Multiple Alignments of Protein Structures.** In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*. St. Louis, MO, USA: AAAI Press; 1996:59–67.

20. Gibrat JF, Madej T, Bryant SH: **Surprising similarities in structure comparison.** *Curr Opin Struct Biol* 1996, **6**(3):377–385.
21. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, **233:**123–138. [http://dx.doi.org/10.1006/jmbi.1993.1489].
22. Lancia G, Carr R, Walenz B, Istrail S: **101 optimal PDB structure alignments: a branch-and-cut algorithm for the maximum contact map overlap problem.** In *RECOMB '01: Proceedings of The Fifth Annual International Conference on Computational Biology*. New York, NY, USA: ACM; 2001:193–202.
23. Singh AP, Brutlag DL: **Hierarchical Protein Structure Superposition Using Both Secondary Structure and Atomic Representations.** In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*. Halkidiki, Greece: AAAI Press; 1997:284–293.
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The protein data bank.** *Nucl Acids Res* 2000, **28:**235–242. [http://nar.oxfordjournals.org/cgi/content/abstract/28/1/235].
25. Papadimitriou C: **The Euclidean Traveling Salesman Problem is NP-Complete.** *Theoretial Compututer Sci* 1977, **4**(3):237–244.
26. Arun KS, Huang TS, Blostein SD: **Least-squares fitting of two 3-D point sets.** *IEEE Trans Pattern Anal Mach Intell* 1987, **9**(5):698–700.
27. Ahuja RK, Magnanti TL, Orlin JB: *Network Flows: Theory, Algorithms, and Applications*. Englewood Cliffs, NJ: Prentice Hall; 1993.
28. Basu S, Pollack R, Roy MF: *A New Algorithm to Find a Point in Every Cell Defined by a Family of Polynomials*(Caviness B, Johnson J, eds.) Springer Vienna: Springer-Verlag; 1998.
29. de Rezende PJ, Lee DT: **Point Set Pattern Matching in d-Dimensions.** *Algorithmica* 1995, **13**(4):387–404.

doi:10.1186/1748-7188-8-1
Cite this article as: Li: The difficulty of protein structure alignment under the RMSD. *Algorithms for Molecular Biology* 2013 **8**:1.