# MULTIPLE STRUCTURE ALIGNMENT BY OPTIMAL RMSD IMPLIES THAT THE AVERAGE STRUCTURE IS A CONSENSUS

Xueyi Wang* and Jack Snoeyink

*Department of Computer Science, University of North Carolina at Chapel Hill*
*Chapel Hill, NC, 27599-3175, USA*
*Email: {xwang*, snoeyink}@cs.unc.edu*

Root mean square deviation (RMSD) is often used to measure the difference between structures. We show mathematically that, for multiple structure alignment, the minimum RMSD (weighted at aligned positions or unweighted) for all pairs is the same as the RMSD to the average of the structures. Thus, using RMSD implies that the average is a consensus structure. We use this property to validate and improve algorithms for multiple structure alignment. In particular, we establish the properties of the average structure, and show that an iterative algorithm proposed by Sutcliffe and co-authors can find it efficiently — each iteration takes linear time and the number of iterations is small. We explore the residuals after alignment and assign weights to positions to identify aligned cores of structures. Observing this property also calls into question whether global RMSD is the right way to compare multiple protein structures, and guides the search for more local techniques.

## 1. INTRODUCTION

Although protein structures are uniquely determined by their sequences[1], protein structures are better conserved through evolution than the sequences[2]. Proteins with similar 3D structures may have similar functions and are often evolved from common ancestors[3]. As structural biologists classify proteins, how should they compare structures?

Pairwise comparisons are commonly performed by measuring the root mean squared deviation (RMSD) between corresponding atoms in two structure, once a suitable correspondence has been chosen and the molecules have been translated and rotated as rigid bodies to the best match[4–6]. Corresponding atoms may also be given weights so that core atoms have the greatest influence on the matching and weighted RMSD score.

Pairwise comparison can be extended to multiple structure alignment in several ways. In this paper we look at ways to extend RMSD (weighted at aligned positions or unweighted) after a correspondence between atoms has already been chosen. Multiple structure alignment is an important tool to identify structurally conserved regions, to provide clues for building evolutionary trees and finding common ancestors, and to determine consensus structures for protein families.

For multiple structure alignment, first we need to choose a score function to measure the goodness of the alignment. Examples from the literature include the sum of all pairwise squared distances[7,8], which we also use, or the average RMSD per aligned position[9]. If we consider the protein structures as rigid bodies, then problem of multiple structure alignment is to translate and rotate these structures to minimize the score function. Several methods also choose a *consensus structure* to represent the whole alignment.

Many algorithms have been presented to solve this multiple structure alignment problem. Some first do pairwise structure alignments and then use heuristic methods to integrate the structures. Gerstein and Levitt[10] choose the structure that has minimum total RMSD to all other structures as the consensus structure and aligns other structures to it. Ochagavia and Wodak[9] and Lupyan *et al.*[7] present a progressive algorithm that chooses one structure at a time and minimizes the total RMSD to all the already aligned structures until all the structures are aligned. Other researchers use non-deterministic methods. Sali and Blundell[11] use simulated annealing to determine the optimal structure alignments and Guda *et al.*[12] use Monte Carlo optimization.

Other algorithms align all the structures together instead of aligning each pair separately. Two iterative algorithms by Sutcliffe *et al.*[8] align protein structures to their average structure, also done by Verboon and Gabriel[13] and Pennec[14]. We will focus most of our attention on this approach. MUSTA[15] use geometric

[*] Corresponding author.

hashing and finds a consensus structure of Cα atoms. MultiProt[16] iteratively chooses each structure as a consensus structure, aligns all other structures to the consensus structure, and detects the largest core among aligned molecules. MASS[17] and CBA[18] first align secondary structure and then align tertiary structure.

In this paper, we show that if you use the root of total squared deviation to score multiple structure alignment, then mathematically you obtain the same result by taking the average structure as a consensus structure, and doing pairwise alignment to this consensus. We can use this to establish properties of the Sutcliffe *et al.*[8] algorithms, including a better stopping condition. In our tests on protein families from HOMSTRAD[19], this algorithm quickly reaches the optimum alignment and consensus structure. By modeling deviations from the average positions as 3-dimensional Gaussian distributions, we can also determine weights for well-aligned positions that can determine the aligned core. We also raise the question, "If the average is not the right consensus structure then what scoring function should replace wRMSD?"

## 2. METHODS

We define the average of structures and weighted RMSD for multiple structures for position weights, and then establish the properties of wRMSD.

### 2.1. Average structure and weighted root mean square deviation

We assume there are $n$ structures each having $m$ points (atoms), so that structure $S_i$ for ($1 \le i \le n$) has points $p_{i1}, p_{i2}, \ldots, p_{im}$. For a fixed position $k$, the $n$ points $p_{ik}$ for $(1 \le i \le n)$ are assumed to correspond. We define the average structure $\overline{S}$ to have points $\overline{p}_k = \frac{1}{n}\sum_{i=1}^{n} p_{ik}$ for ($1 \le k \le m$).

We may assign a position weight $w_k \ge 0$ to each aligned position $k$ and define the weighted root mean squared deviation (wRMSD) as the weighted sum of all squared pairwise distances between structures, i.e. wRMSD = $\sqrt{\frac{2}{mn(n-1)}\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m} w_k \left\| p_{ik} - p_{jk} \right\|^2}$ . Weights allow us to emphasize some positions in the alignment (e.g., an aligned core) and reduce or eliminate the influence of other positions; we obtain the standard RMSD by setting $w_k = 1$ for ($1 \le k \le m$).

Note there are $n(n-1)/2$ structure pairs, and each structure pair has $m$ squared distances. If we want to transform the atom positions to minimize wRMSD, then, because $m$ and $n$ are fixed and the square root function is monotone increasing, we can instead minimize the weighted sum of all squared pairwise distances $\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m} w_k \left\| p_{ik} - p_{jk} \right\|^2$ .

The following technical lemma on weighted sums of squares allows us to make several observations about the average structure under wRMSD.

*Lemma 1.* For any aligned position $k$, the total squared distance from $p_{1k}, p_{2k}, \ldots, p_{nk}$ to any point $q_k$ equals the total to the average point $\overline{p}_k$ plus from $\overline{p}_k$ to $q_k$:

$$\sum_{i=1}^{n}\left\| p_{ik} - q_k \right\|^2 = \sum_{i=1}^{n}\left\| p_{ik} - \overline{p}_k \right\|^2 + \sum_{i=1}^{n}\left\| q_k - \overline{p}_k \right\|^2$$

*Proof.* To establish the Lemma, we subtract the second term from both sides, expand the difference of squares, then apply the definition of $\overline{p}_k$ in the penultimate step.

$$\sum_{i=1}^{n}\left[\left\| p_{ik} - q_k \right\|^2 - \left\| p_{ik} - \overline{p}_k \right\|^2\right]$$

$$= \sum_{i=1}^{n}\left[p_{ik} - q_k + p_{ik} - \overline{p}_k\right]\cdot\left[p_{ik} - q_k - p_{ik} + \overline{p}_k\right]$$

$$= \left[\overline{p}_k - q_k\right]\cdot\sum_{i=1}^{n}\left[2p_{ik} - \overline{p}_k - q_k\right]$$

$$= \left[\overline{p}_k - q_k\right]\cdot\sum_{i=1}^{n}\left[\overline{p}_k - q_k\right] = \sum_{i=1}^{n}\left\| q_k - \overline{p}_k \right\|^2 \text{ . } \square$$

Our first theorem says that if wRMSD is used to compare multiple structures, then what is really happening is that all structures are being compared to the average structure – that the average structure $\overline{S}$ is a consensus, whether we recognize it or not. It is better computationally to recognize this, because it reduces the number of pairs of structures that must be compared from $n(n-1)/2$ to $n$.

*Theorem 1.* The weighted sum of squared distances for all pairs equals the weighted sum of squared distances to the average structure $\overline{S}$ :

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m} w_k \left\| p_{ik} - p_{jk} \right\|^2 = n\sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik} - \overline{p}_k \right\|^2 \text{ .}$$

*Proof.* In Lemma 1, replace $q_k$ by $p_{jk}$, then multiply by the weight $w_k$, and sum over all $j$ and $k$ to obtain:

$$\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{i=1}^{n} w_k \left\| p_{ik} - p_{jk} \right\|^2 = 2\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j=1}^{n} w_k \left\| p_{ik} - \overline{p}_k \right\|^2 .$$

We can re-arrange the order of summation on the left, noticing that terms with $i = j$ cancel and every other term appears twice. The resulting equation gives the desired result after dividing out the extra factor of two:

$$2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m} w_k \left\| p_{ik} - p_{jk} \right\|^2 = 2\sum_{k=1}^{m}\left[ w_k \sum_{i=1}^{n}\sum_{j=1}^{n} \left\| p_{ik} - \overline{p}_k \right\|^2 \right]$$

$$= 2n\sum_{k=1}^{m}\sum_{i=1}^{n} w_k \left\| p_{ik} - \overline{p}_k \right\|^2 . \square$$

Two more theorems suggest how to choose the structure closest to a given set of structures. If you can choose any structures, then chose the average $\overline{S}$ ; if you must choose from a limited set, then choose the structure closest to the average $\overline{S}$ .

*Theorem 2*. The average structure $\overline{S}$ minimizes the weighted sum of squared distances from all the structures, *i.e.* for any structure $Q$ with points $q_1$, $q_2$, …, $q_m$, $\sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik} - q_k \right\|^2 \geq \sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik} - \overline{p}_k \right\|^2$ and equality holds if and only if $q_k = \overline{p}_k$ for all positions with $w_k > 0$.

*Proof*. This follows immediately from Lemma 1 since $\sum_{i=1}^{n} w_k \left\| q_k - \overline{p}_k \right\|^2 \geq 0$ with equality if and only if $q_k = \overline{p}_k$ or $w_k = 0$. □

*Theorem 3*. The structure from a set $Q_1$, …, $Q_m$ that minimizes the weighted sum of squared distances from all the structures $S_i$ is the one whose wRMSD is closest to $\overline{S}$ .

*Proof*. In Lemma 1 $\sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik} - \overline{p}_k \right\|^2$ is fixed by the set of structures, so it is both necessary and sufficient to minimize $\sum_{i=1}^{n} w_k \left\| q_k - \overline{p}_k \right\|^2$ . □

## 2.2. Minimizing wRMSD

In structure alignment, we translate and rotate structures in 3D space to minimize wRMSD. We define $R_i$ as a 3×3 rotation matrix and $T_i$ as a 3×1 translation vector for structure $S_i$. We aim to find the optimal $T_i$ and $R_i$ for each structure to minimize the wRMSD. The target function is:

$$\underset{R,T}{arg\,min}\left( \sum_{i=2}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{m} w_k \left\| R_i p_{ik} - T_i - R_j p_{jk} + T_j \right\|^2 \right).$$

We can fix one of the rotations to be the identity, and one of the translations to be zero. When there are only two structures, then the minimization reduces to a linear equation in $R$ and $T$. Horn[5] showed that these can be found separately: the minimum wRMSD for the pair can be found by translating each structure so its origin is the weighted center of mass, i.e. $\sum_{1\leq k\leq m} w_k p_{ik} = 0$ , then applying an optimum rotation found with quaternions.

To minimize wRMSD with more than two structures, we can combine Theorem 1 with Horn's analysis to show that wRMSD is minimized when the centroids (weighted centers of mass) of each structure is the centroid of the average structure, i.e. each structure may be translated so the origin is the centroid.

Finding optimum rotations for several structures is harder than for a pair because the minimization problem no longer reduces to a linear equation. We can use the fact that the average is the best consensus (Theorem 1), and modify a simple iterative algorithm of Sutcliffe *et al.*[8] to converge to the minimum wRMSD. Instead of directly finding the optimal rotation matrices, we align each structure to the average structure separately to minimize wRMSD. Because rotating structures also changes the average structure, we repeat until the algorithm converges to a local minimum of wRMSD.

*Algorithm*: Given $n$ structures with $m$ points (atoms) each and weights $w_k$ at each position, minimize wRMSD to within a threshold value $\varepsilon$ (e.g. $\varepsilon = 1.0\times10^{-5}$).

1. Translate the weighted centroid of each structure $S_i$ for ($1 \leq i \leq n$) to the origin. (optionally align each structure to a randomly chosen $S_i$ for a good initial average.)
2. Calculate the average $\overline{S}$, with points $\overline{p}_k = \frac{1}{n}\sum_{i=1}^{n} p_{ik}$ , and $SD = \sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik} - \overline{p}_k \right\|^2$ .
3. For each ($1 \leq i \leq n$), align $S_i$ to $\overline{S}$ using Horn's method to calculate optimum rotation matrix $R_i$ that minimizes $\sum_{k=1}^{m} w_k \left\| R_i p_{ik} - \overline{p}_k \right\|^2$ and replace $S_i = R_i \cdot S_i$ .
4. Calculate new average $\overline{S}^{\text{new}}$ and deviation $SD^{\text{new}} = \sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik}{}^{\text{new}} - \overline{p}_k{}^{\text{new}} \right\|^2$ .
5. If $SD - SD^{\text{new}} < \varepsilon$, then the algorithm terminates; otherwise, set $SD = SD^{\text{new}}$ and $\overline{S} = \overline{S}^{\text{new}}$ and go to step 3.

Horn's method and our theorems imply that the deviation *SD* decreases monotonically in each iteration. From theorem 1, we know that minimizing the deviation *SD* to the average minimizes the global wRMSD. From Horn[5], in step 3 we have

$$\sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik}{}^{\text{new}} - \bar{p}_k \right\|^2 \le \sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik} - \bar{p}_k \right\|^2 = SD$$

From theorem 2, in step 4 we have

$$SD^{\text{new}} = \sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik}{}^{\text{new}} - \bar{p}_k{}^{\text{new}} \right\|^2$$

$$\le \sum_{i=1}^{n}\sum_{k=1}^{m} w_k \left\| p_{ik}{}^{\text{new}} - \bar{p}_k \right\|^2$$

So $SD^{\text{new}} \le SD$ and *SD* decreases in each iteration. We stop when this decrease is less than the threshold $\varepsilon$; this will be a local minimum of *SD*.

Horn’s method calculates the optimal rotation matrix for two *m*-atom structures in O(*m*) operations, so initialization and each iteration take O(*n m*) operations. Our experiments show that for any start positions of all *n* structures, the algorithm converges in a maximum of 4–6 iterations when $\varepsilon = 1.0\times10^{-5}$. The number of iterations is one fewer when the proteins start with a preliminary alignment from the optional initialization in step 1. Because the lower bound for aligning *n* structures with *m* points per structure is O(*n m*), this algorithm is close to the optimum.

We must make two remarks about the paper of Sutcliffe *et al.*[8], which proposed the algorithm above. First, they actually give different weights to individual atoms, which they change during the minimization. We can establish analogues of Theorems 1–3 for individual atom weights if the weight of a corresponding pair of atoms is the half-normalized product of the individual weights. To minimize wRMSD for such weights, however, we have observed that it is no longer sufficient to translate the structure centroids to the origin. We believe that this may explain why Sutcliffe's algorithm can take many iterations for convergence — the weights are not well-grounded in mathematics. We plan to explore atom weights more thoroughly in a subsequent paper.

Second, their termination condition was when the deviation between two average structures was small, which is actually testing only the second inequality on the decrease of *SD* above. It is a stronger condition to terminate based on the deviation of *SD*.

While preparing the final version of this paper, we found two papers with similar iterative algorithms[13,14]. Both algorithms use singular value decomposition (SVD) as the subroutine for finding an optimal rotation matrix; quaternions should be used instead because they preserve chirality. Pennec[14] presented an iterative algorithm for unweighted multiple structure alignment and our work can be regarded as the extension of his work. Verboon and Gabriel[13] presented their iterative algorithm as minimizing wRMSD with *atom weights* (different atoms having different weights), but in fact it works only for *position weights* because the optimization of translation and of rotation cannot be separated with atom weights.

## 3. RESULTS AND DISCUSSION

### 3.1. Performance

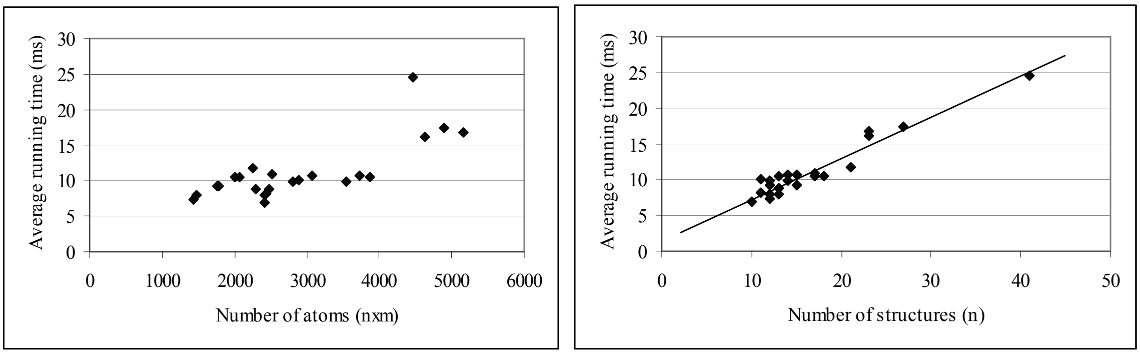
We test the performance of our algorithm by minimizing the RMSD for 23 protein families from HOMSTRAD[19], which are all the families that contain more than 10 structures with total aligned length longer than 100. We set $\varepsilon = 1.0\times10^{-5}$ and run the experiment on a 1.8GHz Pentium M laptop with 768M memory. The code is written in MATLAB and is downloadable at http://www.cs.unc.edu/~xwang/.

We run our algorithm 5,000 times for each protein family. Each time we begin by randomly rotating each structure in 3D space and then minimize the RMSD. We expect that the changes in RMSD will be small, since these proteins were carefully aligned with a combination of tools, but want to make sure that our algorithm does not become stuck in local minima that are not the global minimum. The results are shown in Table 1.

For each protein family’s 5,000 tests, the difference between maximum RMSD and minimum RMSD is less than $1.0\times10^{-8}$, so they converge to the same local minimum. Moreover, the optimal RMSD values found by our algorithm are less than the original RMSD from the alignments in HOMSTRAD in all cases. In three cases the relative difference is greater than 3%; in each of these cases there is an aligned core for all proteins in the family, but some disordered regions allow our algorithm to finds alignments with better RMSD. These cases clearly call for weighted alignment.

**Table 1.** Performance of the algorithm on different protein families from HOMSTRAD. We report n, the number of proteins, m, the number of atoms aligned, RMSD from the HOMSTRAD Alignment (HA), the RMSD for the optimal alignment from our algorithm, statistics on iterations and time (milliseconds) for 5,000 runs of each alignment.

| Protein family | *n* | *m* | RMSD HA(Å) | optim. RMSD | % rel. diff | Iterations avg,med,max | Time (ms) avg,median,max |
|---|---|---|---|---|---|---|---|
| immunoglobulin domain - V set - heavy chain | 21 | 107 | 1.224 | 1.213 | 0.91 | 3.8, 4, 4 | 11.7, 10, 30 |
| globin | 41 | 109 | 1.781 | 1.747 | 1.95 | 4.0, 4, 5 | 24.4, 20, 40 |
| phospholipase A2 | 18 | 111 | 1.492 | 1.478 | 0.95 | 3.9, 4, 4 | 10.5, 10, 41 |
| ubiquitin conjugating enzyme | 13 | 114 | 1.729 | 1.714 | 0.88 | 4.0, 4, 5 | 7.9, 10, 11 |
| Lipocalin family | 15 | 118 | 2.881 | 2.873 | 0.28 | 4.0, 4, 5 | 9.3, 10, 30 |
| glycosyl hydrolase family 22 (lysozyme) | 12 | 119 | 1.357 | 1.342 | 1.12 | 3.9, 4, 4 | 7.3, 10, 11 |
| Fatty acid binding protein-like | 17 | 122 | 1.825 | 1.824 | 0.05 | 4.0, 4, 5 | 10.5, 10, 40 |
| Proteasome A-type and B-type | 17 | 148 | 3.302 | 3.032 | 8.91 | 4.8, 5, 6 | 9.3, 10, 21 |
| phycocyanin | 12 | 148 | 2.188 | 2.077 | 5.34 | 4.0, 4, 5 | 11.0, 10, 40 |
| short-chain dehydrogenases/reductases | 13 | 177 | 1.971 | 1.954 | 0.87 | 4.0, 4, 5 | 8.8, 10, 11 |
| serine proteinase - eukaryotic | 27 | 181 | 1.454 | 1.435 | 1.32 | 3.8, 4, 4 | 17.4, 20, 40 |
| Papain fam cysteine proteinase | 13 | 190 | 1.396 | 1.383 | 0.94 | 3.9, 4, 5 | 8.9, 10, 30 |
| glutathione S-transferase | 14 | 200 | 2.336 | 2.315 | 0.91 | 4.0, 4, 5 | 9.8, 10, 20 |
| Alpha amylase, catalytic dom. | 23 | 201 | 2.327 | 2.293 | 1.48 | 4.0, 4, 5 | 16.1, 20, 40 |
| legume lectin | 12 | 202 | 1.302 | 1.287 | 1.17 | 3.8, 4, 4 | 8.0, 10, 30 |
| Serine/Threonine protein kinases, catalytic domain | 15 | 205 | 2.561 | 2.503 | 2.32 | 4.0, 4, 5 | 10.6, 10, 21 |
| subtilase | 11 | 222 | 2.279 | 2.268 | 0.49 | 4.0, 4, 5 | 8.1, 10, 30 |
| Alpha amylase, catalytic and C-terminal domains | 23 | 224 | 2.668 | 2.602 | 2.54 | 4.0, 4, 5 | 16.6, 20, 40 |
| triose phosphate isomerase | 10 | 242 | 1.398 | 1.386 | 0.87 | 3.7, 4, 4 | 7.0, 10, 11 |
| pyridine nucleotide-disulphide oxidoreductases class-I | 11 | 262 | 3.870 | 3.420 | 13.16 | 4.7, 5, 6 | 10.1, 10, 21 |
| lactate/malate dehydrogenase | 14 | 266 | 2.036 | 2.024 | 0.59 | 4.0, 4, 5 | 10.9, 10, 21 |
| cytochrome p450 | 12 | 295 | 2.872 | 2.861 | 0.38 | 4.0, 4, 5 | 9.8, 10, 30 |
| aspartic proteinase | 13 | 297 | 1.932 | 1.877 | 2.93 | 4.0, 4, 4 | 10.5, 10, 30 |



(a) Average running time vs. number of atoms

(b) Average running time vs. number of structures

**Fig. 1**. Average running time vs. the number of atoms or the number of structures

The maximum number of iterations is 6 and the average and median number of iterations is around 4, so *I* is a small constant and the algorithm achieves the lower bound of multiple structure alignment, which is $\Theta(n\ m)$. All of the average running time is less than 25 milliseconds and all of the maximum running time is less than 40 milliseconds, which means our algorithm is highly efficient.

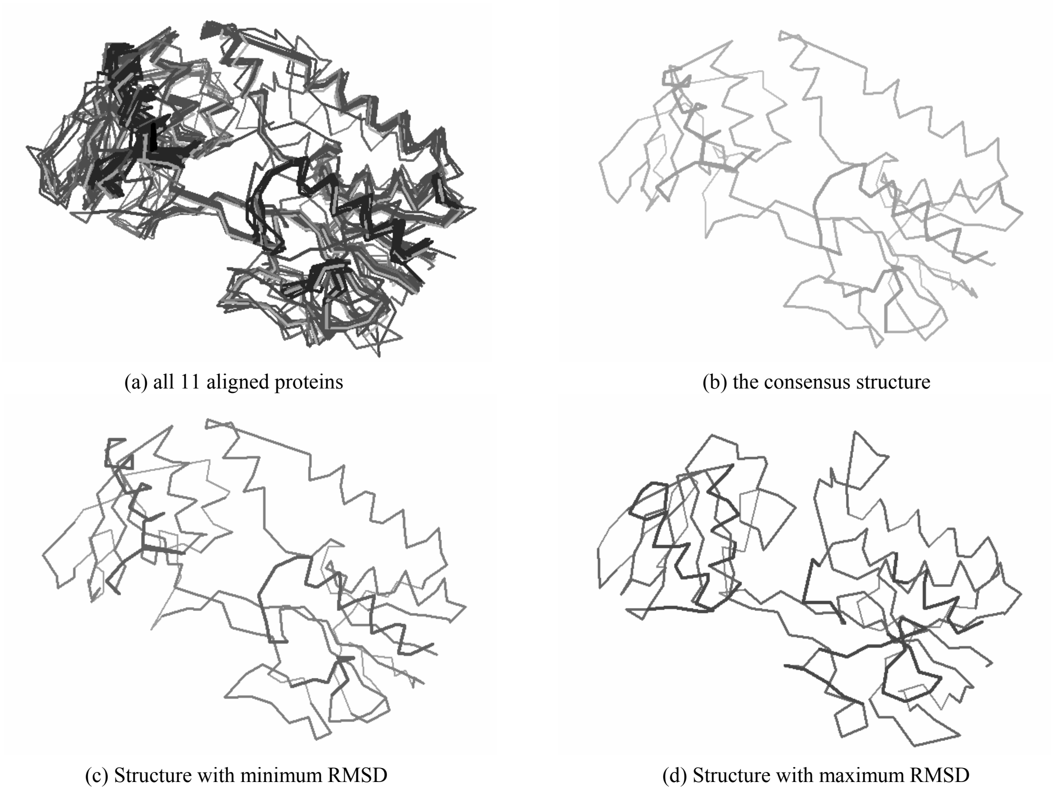
Figure 1a and 1b show the relationship between the average running time and the number of atoms ($n \times m$) and the number of structures ($n$) in each protein family. The average running time shows linear relation with the number of structures but not the number of atoms, because the most time-consuming operation is computing eigenvectors and eigenvalues of a 4×4 matrix in Horn's method, which takes O($n$) in each iteration.

## 3.2. Consensus structure

For a given protein family, one problem is to find a consensus structure to summarize the structure information. Altman and Gerstein[20] and Chew and Kedem[21] propose to use the average structure of the conserved core as the consensus structure. In fact, by Theorems 1 and 2, the wRMSD is minimized by aligning to the average structure, and no other structure has better wRMSD with all structures. Thus, we claim that the average structure is the natural candidate for the consensus structure.

One objection to this claim is that the average structure is not a true protein structure – it may have physically unrealizable distances or angles due to the averaging. This depends on the intended use for the consensus structure — in fact, some other proposed consensus structures are even more schematic: Taylor *et al.*[22], Chew and Kedem[21], and Ye and Janardan[23] use vectors between neighboring $C_\alpha$ atoms to represent protein structures and define a consensus structure as a collection of average vectors from aligned columns.

But a more significant answer comes from Theorem 3: if you do have a set of structures from which you wish to choose a consensus, including the proposal of Gerstein and Levitt[10] to use the true protein structure that has the minimum RMSD to all other structures, or POSA of Ye and Godzik[24], which builds a consensus structure by rearranging input structures based on alignments of partial order graphs based on



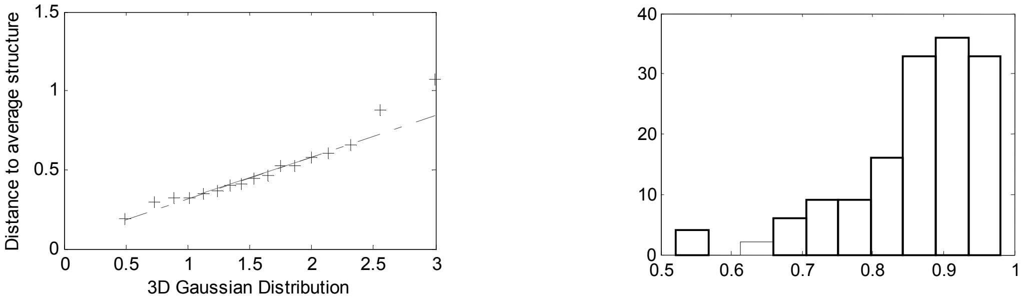
(a) all 11 aligned proteins

(b) the consensus structure

(c) Structure with minimum RMSD

(d) Structure with maximum RMSD

**Fig. 2**. Multiple structure alignment for pyridine nucleotide-disulphide oxidoreductases class-I



(a) Distribution of the best aligned position

(b) histogram of $R^2$ for all aligned positions

**Fig. 3**. 3D Gaussian Distribution analysis of the distances from each atom to corresponding points on the average structure

these structures, then you should choose from this set the structure with minimum wRMSD to the average.

Figure 2 shows the alignment of conserved core of protein family pyridine nucleotide-disulphide oxidoreductases class-I, the consensus structure, the consensus protein structure with the minimum RMSD to all other structures, and the structure with maximum RMSD to other structures.

## 3.3. Statistical analysis of deviation from consensus in aligned structures

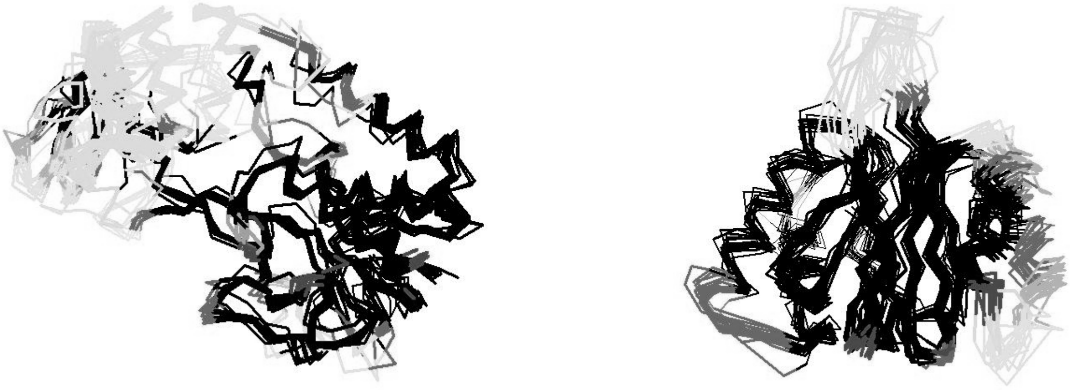
Deriving the statistical description of the aligned protein structures is an intriguing question that has significant theoretical and practical implications. As a first step, we investigate the following question concerning the spatial distribution of aligned positions in a protein family. More specifically, we want to test the null hypothesis that, at a fixed position $k$, the distances the $n$ atoms can be found from the average $\overline{p}_k$ , especially those that are in the "core" area of protein structures, are consistent with distances from a 3D Gaussian distribution. We chose the Gaussian not only because it is the most widely used distribution function, due to the central limit theorem of statistics, but also because previous studies hint that Gaussian is the best model to describe the aligned structures[25]. If, by checking our data, we can establish the fact that aligned positions are distributed according to the Gaussian distribution in 3D, the set of aligned protein structures can be conveniently described by a concise model that is composed by the average structure and the covariance matrix specifying the distribution of the positions.

To test the fitness of our data to the hypothesized 3D Gaussian model, we adopted the Quantile-Quantile Plot (q-q plot) procedure[26], which is commonly used to determine whether two data sets come form a common distribution. In our procedure, the y-axis is the distances from each structure to the average structure for each aligned position, and the x-axis is the quantile data from 3D Gaussian. Figure 4a shows the q-q plot for the best aligned position. The correlation coefficient $R^2$ is 0.9632, which suggests that the data fits the 3D Gaussian model pretty well. We carried the same experiments for all the aligned positions and the collected the histogram of the correlation coefficient $R^2$ is shown in figure 4b. We identify that more than 79% of the positions we check have $R^2 > 0.8$.

The types of curves in q-q plots reveal information that can be used to classify whether a position should be deemed part of the core. The illustrated q-q plot has the last two curves above the line, which indicates that the two corresponding structures have larger errors in this position than would be predicted by a Gaussian distribution. Most position produce curves like this, or with all or almost all points on a line through the origin. Low slope indicates that they align well, and that the residuals may fit a 3D Gaussian distribution with a small scale. A few plots begin above the line and come down, or stay on a line of higher slope, indicating that such positions are disordered and should not be considered part of the core.

## 3.4. Determining and weighting the core for aligned structures

There are many ways in which we can potentially use this model of the alignment in a family to determine the structurally conserved core of the family, and help biologist to compare protein structures. Due to space constraints, we briefly demonstrate one heuristic for determining position weights to identify and align the conserved core of two of our structure families.



(a) pyridine nucleotide-disulphide oxidoreductases class-I (b) proteasome A-type and B-type

**Figure 4**. Aligned protein families using position weights. The black colored positions satisfy $a_k \le \bar{a} + \sigma$, dark-gray colored atoms satisfy $\bar{a} + \sigma < a_k \le \bar{a} + 2\sigma$, gray colored atoms satisfy $\bar{a} + 2\sigma < a_k \le \bar{a} + 3\sigma$, and light-gray colored atoms satisfy $a_k > \bar{a} + 3\sigma$.

We use the following iterative algorithm to assign weights.

1. Align the protein structures by RMSD using the algorithm of Section 3.2.
2. For each aligned position *k*, calculate distance $d_{ik} = \|p_{ik} - \bar{p}_k\|$ for (1 ≤ *i* ≤ *n*), and the correlation coefficient $R_k^2$ by assuming that deviations have a 3D Gaussian distribution, and the average squared distance $a_k = \frac{1}{n}\sum_{i=1}^{n} d_{ik}^{\ 2}$. Then calculate the mean $\bar{a}$ and standard deviation $\sigma$ of $a_k$.
3. If all $a_k \le \bar{a} + 3\sigma$, then exit the algorithm; Otherwise set all weights $w_k = \begin{cases} R_k^{\ 2} / a_k & if\ a_k \le \bar{a} + 3\sigma \\ 0 & otherwise. \end{cases}$, align structures by wRMSD, and go to step 2.

The term $1/a_k$ in the weights encourages the alignment in the positions where the average squared deviations are small, and the term $R_k^2$ encourages those positions where the distances to the average structure are close to 3D Gaussian distribution. Figure 3 shows two examples of alignments, where the black is the core and gray are portions that are eliminated by being given weight zero, often due to divergence in of some or all members in the family.

## 4. CONCLUSION

In this paper, we analyzed the problem of minimizing the multiple structure alignment using weighted RMSD with weights at aligned positions, which includes RMSD as a special case. While directly minimizing wRMSD is hard in multiple structure alignment, we show that this problem is the same as minimizing the wRMSD to the average structure. Thus, the average structure is the natural choice for a consensus structure.

Based on this property, we create an efficient iterative algorithm for minimizing the wRMSD and prove its convergence and other properties. Each iteration takes time proportional to the number of atoms in the structures. We tested the algorithm on the protein families from HOMSTRAD database that have more than 10 proteins with total aligned length longer than 100 atoms. The results show our algorithm minimizes the wRMSD in less than 50 milliseconds in Matlab for any protein family. Regardless of the starting positions of structures, the tests show that the algorithm converges to the same local minimum, which is most probably the global minimum. The tests also show that the number of iterations is a small constant whenever the input does not have near symmetry, so the algorithm achieves the linear lower bound for multiple structure alignment.

The algorithm in the paper is for aligning protein structures after sequence alignment. We plan to extend our work to weighted multiple structure alignment with atom weights at different atoms (which includes gapped structure alignment as a special case). We plan to devise new algorithms to achieve better aligned structures for multiple structure alignment by combining the sequence and structure alignments together and build 3D Hidden Markov Models for protein structure classification.

## Acknowledgments

We thank Jun Huan for helpful discussion and advice during this work, and the reviewers for their comments. This work is partially supported by NIH grant 127666 and a UNC Bioinformatics and Computational Biology training grant.

## References

1. Sela M, White FH Jr, Anfinsen CB. Reductive cleavage of disulfide bridges in ribonuclease. *Science* 1957; **125**: 691-692.
2. Holm L, Sander C. Mapping the protein universe. *Science* 1996; **273**: 595-602.
3. Branden C, Tooze J. *Introduction to protein structure*, 2nd ed. Garland Publishing, Inc., New York. 1999.
4. Kabsch W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica A* 1978; **34**: 827–828.
5. Horn BKP. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A: Optics, Image Science, and Vision* 1987; **4(4)**: 629-642.
6. Coutsias EA, Seok C, Dill KA. Using quaternions to calculate RMSD. *Journal of Computational Chemistry* 2004; **25(15)**: 1849-1857.
7. Lupyan D, Leo-Macias A, Ortiz AR. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* 2005; **21(15)**: 3255-3263.
8. Sutcliffe MJ, Haneef I, Carney D, Blundell TL. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Engineering* 1987; **1(5)**: 377-384.
9. Ochagavia ME, Wodak S. Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins. *Proteins* 2004; **55(2)**: 436-454.
10. Gerstein M, Levitt M. Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Science* 1998; **7**: 445-456.
11. Sali A, Blundell TL. The definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *Journal of Molecular Biology* 1990; **212**: 403-428.
12. Guda C, Scheeff ED, Bourne PE, Shindyalov IN. A new algorithm for the alignment of multiple protein structures using Monte Carlo optimization. *Proceedings of Pacific Symposium on Biocomputing* 2001: 275-286.
13. Verboon P, Gabriel KR. Generalized Procrustes analysis with iterative weighting to achieve resistance, *Br. J. Math. Statist. Psychol*, 1995; **48**:57-74.
14. Pennec X. Multiple registration and mean rigid shapes: Application to the 3D case. *Proceedings of the 16th Leeds Annual Statistical Workship*, 1996: 178-185.
15. Leibowitz N, Nussinov R, Wolfson HJ. MUSTA--a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins. *Journal of Computational Biology* 2001; **8(2)**: 93-121.
16. Shatsky M, Nussinov R, ,Wolfson HJ. A method for simultaneous alignment of multiple protein structures. *Proteins* 2004; **56(1)**: 143-156.
17. Dror O, Benyamini H, Nussinov R, Wolfson HJ. Multiple structural alignment by secondary structures: Algorithm and applications. *Protein Science* 2003; **12**: 2492-2507.
18. Ebert J, Brutlag D. Development and validation of a consistency based multiple structure alignment algorithm. *Bioinformatics* 2006; **22(9)**: 1080-1087.
19. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Science* 1998; **7**: 2469-2471.
20. Altman RB, Gerstein M. Finding an average core structure: application to the globins. *Proc. Int. Conf. Intelligent Systems for Molecular Biology* 1994; **2**: 19-27.
21. Chew LP, Kedem K. Finding the Consensus Shape for a Protein Family. *Algorithmica* 2002; **38(1)**: 115-129.
22. Taylor WR, Flores TP, Orengo CA. Multiple protein structure alignment. *Protein Science* 1994; **3**: 1858-1870.
23. Ye J, Janardan R. Approximate multiple protein structure alignment using the sum-of-pairs distance. *Journal of Computational Biology* 2004; **11(5)**: 986-1000.
24. Ye Y, Godzik A. Multiple flexible structure alignment using partial order graphs. *Bioinformatics* 2005; **21(10)**: 2362-2369.
25. Alexandrov V, Gerstein M. Using 3D Hidden Markov Models that explicitly represent spatial coordinates to model and compare protein structures. *BMC Bioinformatics* 2004; **5**:2.
26. Evans M, Hastings N, Peacock B. *Statistical Distributions*. 3rd ed. New York, Wiley. 2000.