An Efficient Multiple Protein Structure Comparison Method and its Application to Structure Clustering and Outlier Detection

Wei Wu, Anuj Srivastava, Jose Laborde, Jinfeng Zhang
Department of Statistics, Florida State University, Tallahassee, FL, USA.
Emails: wwu, anuj, laborde, jinfeng@stat.fsu.edu.

Abstract—Despite many years of research, comparing multiple protein structures simultaneously (multiple structure comparison) is still a challenging problem. Most of the previous studies have focused on similarities among subsets of residues (or atoms) from a group of proteins (local structure alignment) by minimizing some similarity scores based on root mean square deviation (RMSD) of the aligned residues. In this paper, we designed a novel mathematical and statistical framework for multiple global structure comparison (MGSC). Under this framework, a formal geodesic distance is defined for any pair of protein structures and a mean structure can be estimated for a group of protein structures. The multiple structure comparison is then conveniently performed by comparing each individual structure with the mean structure. The formal distance facilitates consistent and accurate clustering of protein structures. An efficient clustering algorithm was designed based on the developed method. Probability models can be built for groups of protein structures and used in hypothesis testing. A robust outlier detection algorithm was designed to illustrate the potential applications of the framework.

I. INTRODUCTION

Comparison of protein structures is an essential tool for detecting conserved structural motifs, understanding the evolutionary relationships of proteins, and predicting the functions and structures of proteins. Many multiple structure alignment methods have been developed in the past [1], [2], [3], [4], [5], [6]. Most of the these methods aimed to solve local structure alignment (LSA) problem. The goal of LSA is to detect similar substructures in different proteins, which can be superposed and give small distances, such as RMSDs. Global structure comparison (GSC), on the other hand, has been paid much less attention in the past, despite that it can also be very useful [7], [8]. When multiple protein structures need to be compared, multiple structure comparison methods that compare all the structures simultaneously are more desirable than pairwise comparison methods. However, to the best of our knowledge, there have been no methods developed in the past for multiple global protein structure comparison. Availability of such method would add a valuable tool to the repertoire of protein structure analysis.

In this study, we aim to develop a rigorous mathematical and statistical framework for multiple global protein structure comparison, which should have the following capabilities: (1) The distance calculated for any pair of protein structures should be a formal distance. This is important for defining the mean structures, building probability distributions, and performing clustering for a group of protein structures; (2) The framework should allow us to estimate some essential properties of the population, such as mean structures and covariances given certain assumption of the underlying probability models; (3) Probabilistic models can be constructed for a group of protein structures, which characterize the structure variations within these protein structures; and (4) Outliers, which do not share significant global similarities with the rest of the proteins in a sample, can be automatically detected using the probabilistic model.

II. METHODS

A. Mathematical Framework

In this section we present a mathematical framework for shape analysis of curves that achieves goals laid out in the Introduction section. In this framework objects of interest are treated as continuous parameterized curves rather than a sequence of discrete points. We briefly summarize the framework here and refer the reader to the papers [9], [10] for details.

1) Representation Space of Functions: In this framework each protein structure is treated as a parametrized curve $f : [0, 1] \to \mathbb{R}^3$. In case that only a set of discrete points along a curve are available, one can linearly interpolate between them to obtain a full curve. For any two such curves, say $f_1$ and $f_2$, we will assume that the corresponding points $f_1(t)$ and $f_2(t)$, for all $t \in [0, 1]$, are matched to each other. Let $\gamma : [0, 1] \to [0, 1]$ be a
smooth invertible function with $\gamma(0) = 0$ and $\gamma(1) = 1$. It is called a re-parametrization function because for any curve $f$, $f \circ \gamma$ is a re-parametrization of $f$. The curve $f(\gamma(t))$ goes through the same set of points as $f(t)$, but at different rates.

Now let $\mathcal{F}$ denote the set of curves in $\mathbb{R}^3$ given by $\mathcal{F} = \{ f : [0, 1] \to \mathbb{R}^3 | f$ is absolutely continuous $\}$. In this approach, the shapes of curves are analyzed using a new function, called the Square-Root Velocity Function or SRVF, defined as: $q(t) = \frac{f(t)}{\sqrt{|f(t)|}}$, where $| \cdot |$ is the 2-norm of a vector. One can show that for any $f \in \mathcal{F}$, $q$ lies in the set $L^2([0, 1], \mathbb{R}^3)$ or simply $\mathbb{L}^2$.

Let $\Gamma$ be the set of all re-parametrizations: $\Gamma = \{ \gamma : [0, 1] \to [0, 1] | \gamma(0) = 0, \gamma(1) = 1, \gamma > 0 \}$. Elements of $\Gamma$ form a group with function composition, and the identity in this group is the self-mapping $\gamma_{id}(t) = t$. For any function $h$, we will use $\| h \|$ to denote its $L^2$ norm $(\int_0^1 \| h(t) \|^2 h(t) dt)^{1/2}$. One can easily find that the SRVF of $f \circ \gamma$ is given by: $\tilde{q}(t) = (q \circ \gamma)(t) \sqrt{|\gamma(t)|}$. We will denote this transformation by $(q, \gamma) = (q \circ \gamma) \sqrt{|\gamma|}$. Similarly, if a curve is rotated by a rotation matrix $O$, an element of $SO(3)$, then its SRVF is also rotated by the same matrix, i.e. the SRVF of the curve $O f(t)$ is $O q(t)$. Here, $SO(3)$ is the set of all $3 \times 3$ rotation matrices.

2) Elastic Shape Metric: To compare two protein structures, one needs to measure the metric distance between them. We consider the problem of registration and shape comparison for two curves, as given in the following definition.

**Definition 1:** [Shape Metric] For any two functions $f_1$, $f_2 \in \mathcal{F}$ and the corresponding SRVFs, $q_1, q_2 \in \mathbb{L}^2$, we define the shape metric $d_s$ to be:

$$d_s(f_1, f_2) = \inf_{\gamma \in \Gamma, O \in SO(3)} \| q_1 - O(q_2 \circ \gamma) \sqrt{|\gamma|} \|.$$ 

This minimization over re-parametrization and rotation is exactly the problem of global alignment of $f_1$ and $f_2$.

Now that we have a tool for pairwise alignment of curves, we extend it to the problem of multiple alignment as follows.

**B. Mean shape and multiple curve comparison**

At first, for a given set of re-parametrization functions $\gamma_1, \gamma_2, \ldots, \gamma_n \in \Gamma$, their Karcher mean is computed as $\overline{\gamma}_n(t) = \frac{1}{n} \sum_{i=1}^n \gamma_i(t)$. Similarly, for a set of rotation matrices $\{ O_i \}$, viewed as elements of $SO(3)$, we compute $A = \frac{1}{n} \sum_{i=1}^n O_i$. The projection of $A$ on $SO(3)$, given by $U V^T$, where $A = U \Sigma V^T$ is the modified SVD of $A$ (with $U, V \in SO(3)$) is considered as the desired mean rotation. We will use $\overline{O}_n$ to denote the mean of a set of rotation matrices.

Next we consider the problem of finding the mean shape for a set of curves given by $f_1, f_2, \ldots, f_n$. Let their SRVFs be given by $q_1, q_2, \ldots, q_n$.

**Definition 2:** Define the mean shape as a local minimum of the sum of squares of the distance $d_s$:

$$\mu_f = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n d_s(f, f_i)^2.$$  \hspace{1cm} (1)

In terms of the SRVFs, the mean shape is given by:

$$\mu_q = \arg \min_{q \in \mathbb{L}^2} \left( \sum_{i=1}^n \left( \inf_{\gamma \in \Gamma, O \in SO(3)} \| q - O(q_i \circ \gamma) \sqrt{|\gamma|} \|^2 \right) \right).$$

The algorithm for computing the mean shape class is as follows:

**Algorithm 1:** Mean Shape Class

1) Initialization Step: Select $\mu_q = q_j$, where $j$ is any index in $\arg \min_{1 \leq i \leq n} \| q_i - \frac{1}{n} \sum_{k=1}^n q_k \|$.  

2) For each $q_i$, find $\gamma_i^*$ and $O_i^*$ by solving: $\{ \gamma_i^*, O_i^* \} = \arg \min_{\gamma \in \Gamma, O \in SO(3)} \| q - O(q_i \circ \gamma) \sqrt{|\gamma|} \|$.  

3) Compute the aligned SRVFs using $\tilde{q}_i \rightarrow O_i^*(q_i \circ \gamma_i^*) \sqrt{|\gamma_i^*|}$.  

4) If $\frac{1}{n} \sum_{i=1}^n \tilde{q}_i - \mu_q$ is small, then stop. Else, update $\mu_q \rightarrow \frac{1}{n} \sum_{i=1}^n \tilde{q}_i$ and return to step 2.

Now that we have computed the mean shape class, the next task is to find a particular element of this class so that it can be used as a template to align the given curves to this template. This element is chosen by imposing additional constraints on the mean curve. We define the template to be an element, say $\mu_0$ of the shape class such that the re-parametrization functions and rotations that match individual SRVFs $(q_i)$ to $\mu_0$ have means $\gamma_{id}$ and $I_i$, respectively. That is, there exist a set of the re-parametrization functions and rotations $\{ \gamma_i^*, O_i^* \} \in \arg \min_{\gamma \in \Gamma, O \in SO(3)} \| O - O(q_i \circ \gamma) \|$, $i = 1, \ldots, n$, that satisfy

$$\frac{1}{n} \sum_{i=1}^n \gamma_i(t) = t,$$

and $\overline{O}_n = I_i$.  \hspace{1cm} (2)

The algorithm to construct this template from the mean shape class is as follows:

**Algorithm 2:** Template Selection in Mean Shape Class

Let $\mu_q$ be any element of the mean shape class, obtained using Algorithm 1.

1) For each $q_i$, find $\gamma_i^*$ and $O_i^*$ by solving: $\{ \gamma_i^*, O_i^* \} = \arg \min_{\gamma \in \Gamma, O \in SO(3)} \| q - O(q_i \circ \gamma) \sqrt{|\gamma|} \|$.  

2) Compute the aligned SRVFs using $\tilde{q}_i \rightarrow O_i^*(q_i \circ \gamma_i^*) \sqrt{|\gamma_i^*|}$.  

3) Compute the aligned SRVFs using $\tilde{q}_i \rightarrow O_i^*(q_i \circ \gamma_i^*) \sqrt{|\gamma_i^*|}$.  

4) If $\frac{1}{n} \sum_{i=1}^n \tilde{q}_i - \mu_q$ is small, then stop. Else, update $\mu_q \rightarrow \frac{1}{n} \sum_{i=1}^n \tilde{q}_i$ and return to step 2.
2) Compute the mean $\gamma_n$ of all $\{\gamma_i^*\}$ and mean $\mathcal{O}_n$ of all $\{\mathcal{O}_i^*\}$. Then compute the template $\mu_0 = \mathcal{O}^{-1}(\mu_q, \gamma_n^{-1}) = \mathcal{O}^{-1}(\mu_q \circ \gamma_n^{-1})$.

We can show that $\mu_0$ resulting from this procedure satisfies Eqn. 2. Now we can use Algorithms 1 and 2 together to present the full procedure for finding a template $\mu_0$ that is used to align the individual functions.

**Curve Alignment Algorithm:** Given a set of curves $f_1, f_2, \ldots, f_n$, let $q_1, q_2, \ldots, q_n$ denote their SRVFs, respectively.

1) Compute the mean shape class of $q_1, q_2, \ldots, q_n$ using Algorithm 1. Denote it by $\mu_q$.

2) Select the preferred element $\mu_0$ of the class of $\mu_q$ using Algorithm 2 (Note that this algorithm requires a step for computing the means of the pre-parametrization functions and rotations).

3) For $i = 1, 2, \ldots, n$, find $\gamma_i^*$ and $\mathcal{O}_i^*$ by solving: $\{\gamma_i^*, \mathcal{O}_i^*\} = \arg\min_{\gamma, \mathcal{O} \in SO(3)} \|\mu_0 - \mathcal{O}(q_i, \gamma)\|_2$.

4) Compute the aligned SRVFs $\tilde{q}_i = \mathcal{O}^*(q_i, \gamma_i^*)$ and aligned curves $\tilde{f}_i = \mathcal{O}^*(f_i \circ \gamma_i^*)$.

5) Return the template $\mu_0$, the pre-parametrization functions $\{\gamma_i^*\}$, the rotations $\{\mathcal{O}_i^*\}$, the aligned SRVFs $\{\tilde{q}_i\}$, and the aligned curves $\{\tilde{f}_i\}$.

**III. Results**

We test the clustering and outlier detection method using three protein families taken from the SCOP database [11]: globin-like (alpha proteins, SCOP classification: a.1.1), PDZ domain (beta proteins, b.36.1.1), and phosphate binding protein-like (alpha/beta proteins, c.94.1.1).

A subset of proteins with sequence identity smaller than 50% are selected from each family using the PISCES server [12]. This results in 45 proteins in the globin-like family (called Dataset 1), 23 proteins in the PDZ domain family (Dataset 2), and 50 proteins in the phosphate binding protein-like family (Dataset 3). Protein sequence lengths range from 110 to 176, 82 to 123, and 200 to 492, in the three datasets, respectively.

The pre-processing is performed similarly to an earlier study [10]. Basically, the alignment and rotation is conducted in the SRVF space. To have a robust estimate of the SRVFs, we smooth the resampled signals with a Gaussian kernel function. We also add one residue at both N and C terminal of each protein chain by extrapolating from the two terminal residues to allow some degree of freedom on matching boundary residues. The added residues are deleted after matching.

**A. Multiple global structure comparison**

In this section, we report the results of multiple global structure alignment of the above datasets.

**Globin-like family (dataset 1).** In our algorithm, we iteratively identify a template $\mu$ (mean structure in the SRVF space), and rotate and align each curve $\{q_i\}$ calculated from protein $i$ to $\mu$. These optimal re-parametrization functions are shown in Fig. 1A. These functions are near their center, the identity $\gamma_{id}(t) = t$.

This indicates the alignment is not drastic, with only some local adjustment. Note that optimal rotations are $3 \times 3$ matrices. To intuitively view them in a graph, we denote their center $I_3$ as the fixed point $(1, 1, 1)^T / \sqrt{3}$ in the unit sphere, and show the rotation from each sequence to $\mu$ as one point on the sphere. Using this notation, the rotation from each function to $\mu$ is identical to the rotation from this point to $(1, 1, 1)^T / \sqrt{3}$ on the sphere. Fig. 1B shows how the representative points are distributed over the sphere, indicating significant rotations need to be made to these proteins in registration.

The transformed data (after re-parametrization and rotation) are shown in Fig. 1C and D. To make the result more readable, we centralize the data by subtracting their sample mean. We can see the three coordinates (x, y and z) are well aligned (except for a few outliers), and the 3-D structures are well superimposed.

**PDZ domain (dataset 2) and phosphate binding protein-like family (dataset 3).** The global alignment results on the other two datasets are shown in Fig. 2 (A: dataset 2, B: dataset 3), where the x, y and z coordinates after alignment are shown. Most of 23 structures in dataset 2 align very well with one another.
In contrast, the 50 structures in dataset 3 are not so well aligned. In dataset 2, two structures appear to be outliers compared to the rest of the structures in the dataset. It would be desirable to identify the possible outliers during alignment so that the set of structures can be realigned without outliers to further improve the alignment.

### B. Outlier Detection

Outliers are the structures with significantly different shapes from the majority of structures in a sample. In the comparison result for Datasets 1 and 2 (see Figs. 1D and 2A), we find that most proteins can be well aligned except a few outliers. In this section, we design a method to detect and remove these outliers from a given set of protein structures. Once removed, the set can be re-aligned to obtain better alignment.

The outlier removal algorithm is designed based on the same statistical framework outlined earlier. Let \{\tilde{q}_i\} be the aligned SRVFs after the alignment. In practice, \tilde{q}_i is discretized as a matrix. We concatenate the three components (x, y and z) as one high-dimensional vector, denoted by \(a_i \in \mathbb{R}^D, i = 1, \cdots, n\), where \(D\) is the dimension and \(n\) is the sample size. A two-step procedure is designed as follows:

**Step 1: Pre-screening:** Finding all potential outliers

1. Perform a standard Principal Component Analysis (PCA) on \(\{a_i\}\) with 95% cumulative variance. Denote the transformed data by \(b_i, i = 1, \cdots, n\).
2. Estimate the mean, \(\mu\), and covariance, \(\Sigma\), of \(\{b_i\}\).
   
   Let \(c_i = \Sigma^{-1/2}(b_i - \mu), i = 1, \cdots, n\).
3. Let \(F\) denote the C.D.F. of \(\chi^2(n)\). For given \(\alpha \in [0, 1]\), the \(i\)th protein is potentially an outlier if \(1 - F(|c_i|^2) < \alpha, i = 1, \cdots, n\).

**Step 2: Verification:** Selection within 5% tail probability

1. Denote the output of Step 1 (potential outliers) as \(A\), and the rest in the main cluster as \(B\).
2. Estimate the mean, \(\mu_B\), and covariance, \(\Sigma_B\), in \(B\).
   
   Perform a SVD Decomposition on \(\Sigma_B\) such that \(\Sigma_B = USU^T\) with \(S = diag(s_1, \cdots, s_D)\).
3. Assume the 95% cumulative variance is reached as the \(d\)th component. Then do a shrinkage approximation of covariance by letting \(S \mapsto S + \epsilon s_{d+1} I_D\).
4. Let \(e_i = S^{-1/2}(a_i - \mu_B), i \in A\). The \(i\)th protein is considered as an outlier if \(1 - F(|e_i|^2) < 0.05\). This is equivalent to a p-value < 0.05.

---

Fig. 2. Transformation on Data 2 and 3. A: The transformed 3D-coordinates of Dataset 2. B: The transformed 3D-coordinates of Dataset 3.

Outlier detection on Dataset 1. We first perform outlier detection on Dataset 1. Fig. 3A shows the numbers of detected outliers in Step 1 (blue circle) and Step 2 (red star) as a function of prescreening parameter \(\alpha\) in the range \([0.1, 0.3]\). We note that the number of potential outliers increases approximately with the value of \(\alpha\); a small \(\alpha\) cannot detect any outliers (e.g. 0 potential outliers when \(\alpha \leq 0.11\)), whereas many proteins can be identified as potential outliers for a large \(\alpha\), e.g. 22 (out of 45) proteins are identified when \(\alpha = 0.29\). By applying Step 2 with threshold 0.05 (with respect to 95% confidence region), most falsely detected outliers are returned to the main cluster. There are about 3 to 4 outliers that consistently appear with respect to different values of \(\alpha \in [0.13, 0.3]\). In Fig. 3B, the four outliers detected (red) have very different structures as the rest of proteins in the class.

Outlier detection on Dataset 2. We then perform the same detection on Dataset 2 with 23 proteins. Fig. 3C shows the numbers of detected outliers in Step 1 (blue circle) and Step 2 (red star) as a function of \(\alpha\) in \([0.1, 0.3]\).
0.3]. Here we find the number of potential outliers also increase with the value of $\alpha$. However, by applying Step 2 with threshold 0.05, we will consistently get the same two outliers with respect to different values of $\alpha \in [0.23, 0.3]$. In Fig. 3D, we can see that the two detected outliers (red) are quite different from the rest.

C. Clustering and outlier detection

An examination of the alignment for Dataset 3 (Fig. 2B) indicates that the structures fall into two groups with significant structural differences. Due to the small sample size, we use a K-means clustering on the aligned SRVF $\{\tilde{q}_i\}$ for number of clusters $K = 2$. The result is shown in Fig. 4A. We can see that the data are properly separated as two different classes (with outliers present in both). There are 24 proteins in Cluster 1, and 26 in Cluster 2.

Once the data are clustered, we then apply the same outlier detection procedure on each cluster individually. In this case, 9 sequences are identified as outliers in Cluster 1, and 8 in Cluster 2 (Fig. 4B). The data are now separated as two classes, and we should realign them in their own class. We call this alignment as posterior alignment. The result on the posterior alignment on both clusters are shown in Fig. 4C and D, one for each class. In contrast to the overall alignment (see Fig. 2D), the pattern in each cluster is very clear, and structural difference is observed between two clusters.

IV. Conclusion

In this study, we have designed the first multiple global structure alignment method based on a rigorous mathematical and statistical framework we developed recently [10]. The method calculates formal distances, geodesic distances, for any pairs of protein structures. Mean and covariances can be estimated for a set of protein structures. Probability models can be further built for groups of protein structures and used in downstream applications such as outlier detection, classification, and clustering.

Acknowledgment

Funding: NIH 1R21GM101552.

References


Fig. 4. Clustering and outlier detection in Dataset 3. **A:** Clustering using K-means method with $K = 2$. Sequences in Clusters 1 and 2 are denoted in blue and red colors, respectively. **B:** Two main clusters after the outliers are removed. **C, D:** Posterior alignment on the outlier-removed cluster 1 and 2, respectively.