Protein Structure Alignment Using Elastic Shape Analysis

Wei Liu
Department of Statistics
Florida State University,
Tallahassee, FL 32310. USA
weiliu@stat.fsu.edu

Anuj Srivastava
Department of Statistics
Florida State University,
Tallahassee, FL 32310. USA
anuj@stat.fsu.edu

Jinfeng Zhang
Department of Statistics
Florida State University,
Tallahassee, FL 32310. USA
jinfeng@stat.fsu.edu

ABSTRACT
In this paper we present a method for flexible protein structure alignment based on elastic shape analysis of backbones, in a manner that can incorporate different characteristics of the backbones. In particular, it can include the backbone geometry, the secondary structures, and the amino-acid sequences in the matching process. As a result, a formal distance can be computed and geodesic paths, showing optimal deformations between conformations/structures, can be computed for any two backbone structures. It can also be used to average shapes of conformations associated with similar proteins. Using proteins from SCOP and PDB databases we demonstrate the matching and clustering of proteins using the backbone geometries, the secondary labels and the primary sequences. We demonstrate almost 92% success rate in automatic clustering of 100 proteins from SCOP database.

1. INTRODUCTION
Structure comparison of proteins is an important tool for understanding the evolutionary relationships between proteins, predicting protein structures and predicting protein functions. There are two types of protein structure comparison problems, comparison of backbone structures (structure alignment) and comparison of the binding or active sites of proteins (surface matching). Proteins are flexible molecules and rigid matching of either backbones or surfaces of proteins, as used by most current methods, has the difficulty of recognizing relatively distant, functionally important similarities. Another well known issue in structure comparison is the lack of rigorous distance metric and comprehensive statistical framework for assessing the statistical significance of similarities between individual protein structures and classes of protein structures. Despite many past studies, protein structure alignment is still a challenging problem, especially for cases where structures undergo significant conformational changes or have large insertion or deletion of unrelated structural fragments. In this paper, we focus on the comparisons of backbone structures (now on referred to simply as structures) and develop methods based on elastic shape analysis. This general framework allows a flexible matching of curves using a combination of stretching and bending of two protein structures. In principle, this alignment can be based on different characteristics of a protein structure including:

1. **Geometric structures:** For a parameterized backbone curve, the geometry is specified by its coordinates \( t \mapsto \beta(t) \in \mathbb{R}^3 \).

2. **Geometric labels:** The local shape of the backbone structure at any point is characterized by certain structural labels such as \( \alpha \)-helix, \( \beta \)-sheet, coil, etc. We can denote such labeling by \( t \mapsto a(t) \in A \), where \( A \) is a discrete set of labels.

3. **Sequences:** The chemical sequence of amino acids along the backbones can be denoted by \( t \mapsto s(t) \in B \), where \( B \) is the set of all amino acid labels.

Automated alignments and comparisons of protein structures are difficult because geometric features (such as \( \alpha \) helices and \( \beta \) sheets) are located in different numbers, at different locations along the protein chain, and packed in many different ways on three-dimensional space.

Many structure alignment methods have been developed in the past that use information derived mostly from structures in different ways. Those methods can be largely divided into several classes based on the specific similarity metrics (distance metrics) they aim to optimize to achieve the best alignment. For example, DALI [9, 8], CE [28], MAMMOTH [23], RAPIDO [19, 20], VAST [7], MASS [4] and MUSTANG [15] break protein structures into short peptides and then use the relationships among the peptide fragments to compute the similarity between two structures. SSAP [29], FAST [34], and SABERTOOTH [30] produce structural alignment based on pairwise residue (or \( C_\alpha \)) distances in structure space. GANGSTA+ [13] uses both residue pair contacts and secondary structure elements. In the TOPOFIT method, similarity of protein structures is analyzed using three-dimensional Delaunay triangulation patterns derived from backbone representation. Other methods such as TM-Align [33] and LGA [32] develop similarity scores incorporating various structure information as the metric for finding optimal structure alignments. Multiple structure alignment (MSA) methods have also been developed such as CBA [5], POSA [31], MultiProt [27], MALECON [22], MASS [4] and MUSTANG [15]. Several studies have been done to comprehensively compare different structure alignment methods [14, 18, 2]. Different criteria tend to rank methods differently and for a particular purpose one method may work better then the others, but in general no method works better than others for all purposes. Andreeva et. al. organized non-trivial cases of structural alignments into database SISYPHUS [1] so that researchers can develop methods focusing on those difficult cases, such as circular permutations, segment-swapping, or...
context-dependent folding. The early methods in structure alignment usually compare the rigid structure proteins without considering the conformational dynamics and the conformational heterogeneity of proteins at different functional states. It is well known that proteins are flexible and undergo significant structural changes as part of their normal function [6]. Hence, protein structural analysis requires algorithms that can deal with molecular flexibility. To address that problem, some flexible structural alignment methods were developed such as POSA [31] and RAPIDO [19, 20] which use graph-based methods. ProtDeform [25] considered different rigid transformations at different sites of proteins, allowing for deformations beyond a global rigid transformation. FlexProt [26] algorithm simultaneously detects the hinge regions and aligns the rigid subparts of the molecules allowing alignment of proteins with conformational changes. Despite these extensive studies in the past, structure alignment, especially flexible structural alignment, continues to be a very challenging problem.

Our approach is to consider the protein backbone as a continuous three-dimensional curve \( \beta(t) \) endowed with an auxiliary function derived from either the secondary structure \( \alpha(t) \) or the amino acid sequence \( s(t) \) along it. Thus, it has two distinct features: (1) the geometry or the shape of its backbone curve and (2) the auxiliary information. Our goal is to develop a comprehensive framework for a statistical analysis of protein backbones using both these pieces of information and this requires tools to compare, match, and deform protein backbones from one to another. This proposed framework will generate:

1. **Matching:** Optimal matching of backbones using the joint shape and auxiliary information.

2. **Deformation:** Optimal deformation of one backbone into another using geodesic paths in the shape or joint shape-sequence space of backbones. The geodesics between two backbone conformations of the same protein may provide useful information on the dynamics of protein structures, i.e. how a protein changes its conformation from one to another.

3. **Comparison:** The length of a geodesic path between a pair of points on a Riemannian manifold provides a proper distance between them. In case of shape manifolds it provides a quantification of dissimilarities between any pair of protein backbones. This computation can be based on different combinations of the three characteristics of the curves: geometric coordinates, geometric labels, and sequence labels.

4. **Statistical Summary:** Compute statistical averages of a collection of backbones in terms of their geometries and the labels. Such tools can be further advanced to define statistical models for capturing variations in protein conformations and for classifying future discoveries into pre-determined classes.

This work is an extension of a recent framework for comparing shapes of curves in Euclidean spaces, called the elastic shape analysis [12, 10]. While these papers are primarily concerned with the shapes of curves, a recent paper has studied the joint use of some auxiliary functions along with shapes [17], in the context of analyzing colored images. Here we utilize a similar framework for protein matching except now the sequence/label information, rather than the color distribution, forms the auxiliary function.

The rest of this paper is organized as follows. Section 2 summarizes the framework for elastic shape analysis of parameterized curves in \( \mathbb{R}^3 \). Section 3 describes the construction of auxiliary functions from the the label information (\( \alpha(t) \) or \( s(t) \)), and their use in joint analysis of protein backbones. Section 4 presents some experimental results involving real protein structures including results on protein matching and clustering. The paper ends with a short summary in Section 5.

2. **ELASTIC SHAPE ANALYSIS**

We will treat the primary structure of a protein as a composite curve, made up of a parametrized curve in \( \mathbb{R}^3 \) and an auxiliary function along it. The curve corresponds to the backbone of protein and the auxiliary function comes from either the amino acid sequence or the secondary labels. Given any two such composite curves, we desire a framework that can quantify the differences in shapes of these two curves, taking into account their shapes and auxiliary functions. Since the comparisons involve shapes, the resulting quantifications should not depend on the rigid motions, global scalings, and parameterizations of these curves.

The basic idea in this approach is the following. We represent each parameterized curve by a special function called the square-root velocity function (SRVF) and restrict to the manifold of such functions under the desired constraints. For example, the rescaling of all curves to a particular length results in a spherical manifold. Then, in order to compare shapes of curves, we remove all shape-preserving transformations from this representation. This is done using an algebraic technique – we form a quotient space of the original manifold with respect to these shape-preserving transformation groups. The most difficult part here is removing the variability introduced by the re-parameterizations of curves since it is an infinite-dimensional manifold. In the resulting quotient space, called the shape space of elastic curves, one can perform statistical analysis of curves as if they are random variables. One can compare, match, and deform one curve into another, or compute averages and covariances of curve populations, and perform hypothesis testing and clustering of curves according to their shapes.

This framework is described next.

2.1 Elastic Representation

Let the backbone of a protein be treated as an parameterized curve in \( \mathbb{R}^3 \), denoted by \( \beta : [0, 1] \rightarrow \mathbb{R}^3 \). In order to analyze its shape, we will represent \( \beta \) by its square-root velocity function: \( q(t) = \beta(t) / \sqrt{\| \beta'(t) \|} \) in \( \mathbb{R}^3 \), where \( \| \cdot \| \) is the standard Euclidean product. The SRVF \( q \) includes both the instantaneous speed \( (\| q(t) \|^2 = \| \beta'(t) \|) \) and direction \( (q(t) / \| q(t) \| = \beta(t) / \| \beta(t) \|) \) of curve \( \beta \) at time \( t \). The use of the time derivative makes SRVF invariant to the translation of curve \( \beta \). Conversely, one can reconstruct the curve \( \beta \) from \( q \) up to a translation. In order for the shape analysis to be invariant to scales, we rescale each curve to length 1. With a slight abuse of notation, we will denote the rescaled curves by \( \beta \). Since \( \int_0^1 \| \beta(t) \| dt = 1 \), we have: \( \int_0^1 \| q(t) \|^2 dt = \int_0^1 \| \beta(t) \| dt = 1 \). In other words, the \( L^2 \) norm of the SRVF is a constant. Restricting to the curves of interest, we obtain the set

\[
\mathcal{C} \equiv \{ q : [0, 1] \rightarrow \mathbb{R}^3 \mid \int_0^1 \| q(t) \|^2 dt = 1 \}.
\] (1)

\( \mathcal{C} \) is called the preshape space and is the set of all SRVFs representing parameterized curves of length 1 in \( \mathbb{R}^3 \). It is actually a Hilbert space in the space \( L^2 \).

We have mentioned four shape preserving transformations – translation, scale, rotation, and re-parameterization. Of these, we have
already eliminated the first two from the representations, but the other two remain. Curves that are within a rotation and/or a re-parameterization of each other result in different elements of \( \mathcal{C} \) despite having the same shape. The removal of the remaining two transformations is performed algebraically as follows. Let \( SO(3) \) be the group of \( 3 \times 3 \) rotation matrices and \( \Gamma \) be the group of all re-parameterizations (they are actually positive diffeomorphisms of \([0, 1]\)). For a curve \( \beta \), a rotation \( O \in SO(n) \) and a re-parameterization \( \gamma \in \Gamma \), the transformed curve is given by \( O(\beta \circ \gamma) \). The SRVF of the transformed curve is given by \( \sqrt{T}O(q \circ \gamma) \). In order to unify all elements in \( \mathcal{C} \) that denote the same shape we define equivalence classes of the type: \( [q] = \{O(q \circ \gamma)\sqrt{T}O \in SO(n), \gamma \in \Gamma\} \). Each such class \([q]\) is associated with a unique shape and vice-versa. The set of all these equivalence classes is called the shape space \( \mathcal{S} \); mathematically, it is a quotient space \( \mathcal{C} / (SO(n) \times \Gamma) \).

### 2.2 Shape Comparisons and Averaging

In order to compare any two shapes, we need a metric. We make the shape space \( \mathcal{S} \) a Riemannian manifold by imposing the \( L^2 \) metric on its tangent spaces. It can be shown that under the SRVF representation this \( L^2 \) metric corresponds to the elastic metric for comparing shapes of curves [10]. In simple words, in a pairwise comparison of curves, it allows us to bend and stretch/compress a curve in order to best match the other. The relative amounts of bending and stretching needed for matching depend on the curves but in general is bounded under this metric. For example, it does not allow infinite stretching of one curve to match the other.

Once we have a Riemannian manifold, we can compute distances between points in that manifold. For any two points, the distance between them is given by the length of the shortest path (called a geodesic) connecting them in that manifold. This is a strength of this approach: it not only provides a distance between two protein conformations, thus quantifying differences between their shapes, but also a geodesic path between them in \( \mathcal{S} \). This path has the interpretation that it provides the optimal deformation of one shape into another. The geodesics are actually computed using the differential geometry of the underlying space \( \mathcal{S} \). Consider two curves \( \beta_1 \) and \( \beta_2 \), represented by their SRVFVs \( q_1 \) and \( q_2 \). In order to compute geodesics between their equivalence classes \([q_1]\) and \([q_2]\), we fix \( q_1 \) and find the optimal rotation and re-parameterization of \( q_2 \) to solve:

\[
(O^*, \gamma^*) = \arg\min_{O \in SO(3), \gamma \in \Gamma} \|q_1 - \sqrt{T}O(q_2 \circ \gamma)\|_2 .
\]

The optimization over rotation is straightforward, using SVD, but the optimization over the re-parameterization requires a dynamic programming algorithm. Please note that the optimal \( \gamma^* \) is the matching function between the two backbones. Define \( q_2^* = \sqrt{T}O^*(q_2 \circ \gamma^*) \) and compute a geodesic path between \( q_1 \) and \( q_2^* \) in \( \mathcal{C} \). Since \( \mathcal{C} \) is a sphere, the geodesic between any two points is given by a great circle whose equation is:

\[
\alpha(\tau) = \frac{1}{\sin(\theta)}(\sin((1-\tau)\theta)q_1 + \sin(\tau\theta)q_2^*). \tag{2}
\]

\( \alpha \) is a geodesic path between the given two shapes such that it is in \([q_1]\) at \( \tau = 0 \) and in \([q_2]\) at \( \tau = 1 \). Here \( \theta = \cos^{-1}(q_1, q_2^*) \) is the distance between the two equivalence classes in \( \mathcal{S} \), i.e. \( d([q_1], [q_2]) = \theta \). This \( \theta \) is a proper distance in the shape space as it satisfies all the three properties of a distance function, including the triangle inequality.

Figure 1 shows two simple examples of this idea using synthetic curves. In each case we take two cylindrical helices, shown in (a) and (b) panels, and compute geodesic paths between them in the shape space \( \mathcal{S} \). The panels (c) and (d) in both cases show the optimal matching between the helices and the panels (d) show the optimal \( \gamma^* \) that resulted in that optimal matching. The bottom rows in each case show the geodesic path between the given curves. One can interpret these paths as the optimal elastic deformations from one shape to other. Note that alignment of helices in the two structures despite their different placements and lengths along the curves.

In studies of protein structures, especially conformational changes, one often gets conformations of the same protein resulted from rotation of one or a few adjacent backbone torsion angles. Such changes may give rise to large RMSD (root-mean-square-deviation) between two conformations, which can be troublesome for methods globally matching rigid protein structures. In such case, a good method should result in a matching of overall structures. We demonstrate the success of our elastic framework in such matching using a simple experiment. We take the backbone of a simple protein 2JVD and distort it by bending it at a fixed point by a random angle. The original curve and three randomly distorted curves are shown in the top left block of Figure 2. Some optimal deformations, obtained as geodesic paths between these curves in \( \mathcal{S} \), are shown in the middle rows. Finally, we demonstrate the matching of corresponding points between the original curve and the distorted curves in the bottom row. We can see that despite significant rotations at a point in the amino acid chain, resulting in large conformational changes, the method can still match residues on both sides of the rotation.

One distinct advantage of this framework is that it allows one to compute statistics of shapes as if they are random variables. For example, given a few sample shapes from a population, this method can produce their average shape in a principled manner. Let \( \beta_1, \beta_2, \ldots, \beta_n \)}
be a given set of backbones, represented by their SRVFs \( q_1, q_2, \ldots, q_n \). Define their mean shape as the quantity:

\[
\mu = \arg\min_{\alpha} \sum_{i=1}^{n} d([q_i], [\alpha])^2.
\]  

(3)

The actual minimizer is found using an iterated gradient-approach that is not repeated here due to the lack of space but has been presented in many papers earlier (see e.g. [12]). Consider the four simulated conformations of the 2JVD backbone (shown in Figure 2). Concentrating only on their geometries, we can compute their conformations between two conformations. In another word, the deformations constrained by the auxiliary information will presumably give more realistic transformations between two conformations.

3. JOINT ANALYSIS

So far we have used only the shapes of backbones but the auxiliary information, i.e. the secondary structure labels \( a(t) \) and/or the amino-acid sequence \( s(t) \), can also play an important role in protein matching and one would like a joint framework for matching. Towards that goal we will construct continuous auxiliary functions along the curves, derived from this additional information. The matchings and deformations are then performed using the higher-dimensional composite curves that are formed by concatenating the geometric and the auxiliary coordinates. Since these matchings are dictated by both geometry and the auxiliary information, it will provide a natural physical interpretations of the subsequent deformations. In another word, the deformations constrained by the auxiliary information will give more realistic transformations between two conformations.

3.1 Using Geometric Labels

We start by describing the construction of the auxiliary functions and then describe their use in matching and comparisons of protein structures. There are two ways of accomplishing this task. One is to take the auxiliary information which is usually available in form of discrete labels and convert it into a corresponding real-valued function. This step, of course requires, a mapping between discrete labels and the continuous values and the choice of this mapping is not obvious. The second idea is to use a label matching program to match the label sequences associated with two proteins, and then construct real-valued functions that have the same local shapes at the matched points. For example, we can construct a function that has Gaussian bumps at those sites and is zero everywhere else, for each of the two proteins. This way the auxiliary functions of the two proteins will match with high probability at the desired matching points dictated by the labels. We will demonstrate these two ideas using two different examples. We will use the first method to incorporate the geometrical labels in shape analysis and the second method to include sequence information.

3.2 Using Amino-Acid Sequence Matching

We know that the secondary structure of a protein can provide important information as it dictates the general three-dimensional form of local segments of proteins. The DSSP [11] definition of a hydrogen bond is adopted for protein secondary structures which are roughly defined as \( \alpha \)-helix, \( \beta \)-sheet and Coil. In Figure 3, we show four proteins with secondary structure marked along their backbones. The colors red, blue and green refer to the labels \( \alpha \)-helix, \( \beta \)-sheet and Coil respectively.

We design the auxiliary function based on the secondary structure information. Let \( \phi \) be a mapping from the set \( \mathcal{A} = \{ \alpha \text{-helix, } \beta \text{-sheet, } \text{coil} \} \rightarrow \mathbb{R} \) given by: \( \phi(\alpha \text{-helix}) = +1, \phi(\beta \text{-sheet}) = -1, \) and \( \phi(\text{coil}) = 0 \). By evaluating the \( \phi \) for each geometric label along the curve, we get a function \( f(t) = \phi(a(t)) \). We smooth \( f(t) \) to function \( f(t) \) using a Gaussian kernel and use \( f(t) \) as the fourth dimension to form \( \hat{\beta}(t) = \left[ \begin{array}{c} \beta(t) \\ \phi(a(t)) \end{array} \right] \), where \( \omega \) is the weight to control the contributions from secondary structure. The composite curves \( \hat{\beta} \) can be viewed as parameterized curves in \( \mathbb{R}^4 \) with the first three coordinates representing the shape property and the last coordinate refer to secondary structure auxiliary function. Figure 3 shows three example of this construction.
In this approach we take any two amino acid sequences (being aligned) as input and design an auxiliary function for each of the curve (structure) that will participate in a joint shape-sequence analysis. There exist numerous programs for matching two amino acid sequences. They take the two sequences of symbols (each symbol denoting an amino acid) and provide a probable match by selecting points in each sequence that are matched to each other. In this paper we have used PIR$^4$, where a few matched residues are selected among all matched residues. We take such a matching and design an auxiliary function for each of the curve as follows. Let $\beta_1$ and $\beta_2$ are the two curves in $\mathbb{R}^3$ (associated with the two backbones) being matched and let $s_1(t)$ and $s_2(t)$ be the corresponding amino-acid sequences. Let $t_1, t_2, \ldots, t_{n_2}$ be the matching points on the $j^{th}$ curve with $j = 1, 2$, obtained through sequence alignment. That is, the amino acids $s_1(t_{i,j})$ and $s_2(t_{j,k})$, $i = 1, 2, \ldots, k$, are matched by the sequence matching algorithm. We call the matched locations $\{t_{i,j}\}$ as landmarks. Define two real-valued functions:

$$f_s(t) = \sum_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}((t - t_{i,j})^2)}$$

for $i = 1, 2, \ldots, k$. Here $\sigma$ is a free parameter that is kept fixed in the analysis. For each protein define a composite curve: $\tilde{\beta}_1(t) = \begin{bmatrix} \beta_1(t) \\ w_1(t) \end{bmatrix}$, $\tilde{\beta}_2(t) = \begin{bmatrix} \beta_2(t) \\ w_2(t) \end{bmatrix}$, where $w$ is a weight that a user can select in order to balance the contributions from the shape and the sequence data. A small value of $w$ puts a smaller weight on the sequence component and vice-versa. Figure 4 shows an example between a pair of proteins.

The composite curves $\tilde{\beta}_i$ can be viewed as parameterized curves in $\mathbb{R}^4$ where the first three coordinates are the shape coordinates and the last coordinate comes from the sequence-related auxiliary function.

### 3.3 Joint Structure Analysis

Now that we have composite curves, how do we define optimal matchings, deformations and distances between them? The answer is same as earlier, except this time curves are in $\mathbb{R}^4$ and not $\mathbb{R}^3$. The original algorithms and ideas apply as earlier. Define the SRVF of the composite curve as $\tilde{q}(t) = \begin{bmatrix} \beta_i(t) \\ \sqrt{\gamma(t)} \end{bmatrix}$ and the pre-shape space of composite curves as: $\tilde{C} = \{q : [0, 1] \rightarrow \mathbb{R}^4 | \int_0^1 \|\tilde{q}(t)\|^2 dt = 1\}$. Since we are interested in removing the rigid rotations of only the shape components, we define the rotation group as: $R = \begin{bmatrix} SO(3) & 0 \\ 0 & 1 \end{bmatrix}$.

Elements of $R$ are $4 \times 4$ rotation matrices that rotate only the shape components while leaving the auxiliary function unchanged. Let $\Gamma$ be the group of re-parametrizations of a curve as earlier. We define the shape space of composite curves as the quotient space $\tilde{S} = \tilde{C}/(R \times \Gamma)$. Under the $L^2$ metric, the shape space becomes a Riemannian space and we can compute geodesics between any two such curves. While these geodesic paths are on the space of curves in $\mathbb{R}^4$, we will simply display their first three components as curves in $\mathbb{R}^3$ and will draw landmarks over them as thick points. We clarify that the cost function for optimal matching is now:

$$(O^*, \gamma^*) = \arg\min_{O \in SO(3), \gamma \in \Gamma} ||\tilde{q}_1 - \sqrt{\gamma} O(q_2 \circ \gamma)||^2$$

Since the $\tilde{q}_i$'s now have both the shape and the sequence components, the resulting optimal matching function $\gamma^*$ solves a joint-optimization problem. Depending on the weight $w$, we can obtain different results as the underlying problem will have different contributions from the two components. For a large $w$, the resulting matching will exactly be that from the auxiliary function matching. For a moderate value of $w$ it provides a matching that is a combination of the shape and the auxiliary function.

### 4. EXPERIMENTAL RESULTS

In this section we demonstrate the use of the elastic shape analysis in improving matching of structures across proteins and in clustering proteins using elastic shape distances, using real protein structures taken from PDB [24] and the SCOP database [21].

#### 4.1 Protein Matching

We study the effects of adding either the geometric labels or the sequence labels to the shape coordinates in protein matching.

**Joint Shape and Structural Labels:** Here we study the influence of the secondary labels in pairwise protein matching. We start with an artificial case in Figure 5, where shown in top is the original protein backbone of 2JVD with secondary structure labels. In the lower left panels, we manually stretch the green part between two $\alpha$-helices further into two different cases and perform matching with the original 2JVD protein. The matching is based either on only the shape coordinates (middle column) or both shape and label coordinates (right column). We have displayed the matching using the geometric labels of each protein along a line and then connecting the corresponding parts. The reader can see a slight improvement in the matching of the geometric labels when the label information is used. Additional examples using real proteins of improvement in matching using the structural labels are shown in Figure 6.

**Joint Shape and Sequence Labels:** We start with an example of the protein pair 1PMC and 1MFC. As a first comparison we compute the geodesic path in the shape space $\tilde{S}$ between the backbone of the two proteins. The resulting geodesic is shown in the top row of the Figure 7. This geodesic represents the optimal deformation, using a bending and stretching, of one backbone into another. The corresponding matching points are shown in the bottom left panel. Since this process did not involve any information from the respective amino acid sequences, the landmarks are not matched well in the process. This confirms the original hypothesis that the geometry of the backbones and the chemistry of the protein...
sequences provide two different sources of information for structural matchings of proteins. If only the geometry is used, it may not respect the matching suggested by the sequence. For the same two proteins we apply the sequence matching and study the shapes of the composite curves resulting from the joint shape-sequence information. The resulting geodesic path in \( \mathcal{S} \) between these backbones is shown in the second row and the corresponding matching is shown in the lower right. Previous registration in Figure 7 is also marked with the landmarks of sequence matching. Comparing these two matching results, we can see that the joint analysis pushes the sequence landmarks to match each other well. In contrast, in shape-only analysis, the matching is based only on the geometric features with no involvement of the sequence landmarks.

In Figure 8, we present a study of changes in matching of two protein backbones when the relative contributions from the auxiliary component (landmarks) are increased from zero to a large value. We use artificial landmarks on two simple proteins: 2JVD (green) and 2ERL (blue) to demonstrate this idea. In the leftmost panel, the matching is purely on the basis of shape and none of the landmarks are matched. As the weight \( w \) associated with the auxiliary component increases, the matchings start improving and finally the matchings are completely dictated by the landmarks.

Figure 6: The registrations of protein structure using shape only and joint analysis

Figure 5: Top shows the original 2JVD backbone with secondary structures color-labeled and left column shows its two distortions. The middle and the right columns show the registrations obtained by elastic shape analysis and the joint shape and secondary structure analysis, respectively.

Figure 7: Geodesic paths and matching between two protein backbones using shape only (top) and the joint shape-sequence (bottom) analysis.

Figure 8: Demonstration of changes in protein matching as the contribution from the auxiliary component (derived from landmarks) is steadily increased.
4.2 Structural Clustering and Classification

Now we consider the problem of clustering and classification of proteins according to the shapes and other structures of their backbones. We describe three different experiments using different protein datasets.

Joint Shape and Sequence Clustering: First we use a set of 20 proteins shown in Figure 9, manually selected from several different classes/families of SCOP database [21]. We would like to automatically group these proteins into clusters of same class/family using the joint metric described earlier. By recording pairwise geodesic distances for the 20 proteins we can form two geodesic distance matrices: one for shape only and one for joint shape-sequence analysis. Shows in the left most panel of Figure 10 is the geodesic distance matrix for the shape analysis. The distance value is relative smaller when the color is darker which means similarity is higher between two proteins. The diagonal of the matrix is black since the geodesic between a protein and itself is zero. Under these distances, one can obtain a clustering of these proteins using any standard algorithm. The second panel in this figure shows the resulting clustering using the shape-only distances. The remaining two panels show the distance matrix and the clustering for the same proteins but using the joint shape-sequence distances. We can see that alignment with landmarks taken from sequence information (Joint clustering) produce much better clustering results than using only structure information (shape clustering). Joint clustering separates alpha proteins completely with beta proteins and classifies alpha proteins into their corresponding subfamily or sub-class correctly. The only exception is 1IXS, which is a large protein containing a small domain, which share similarity with other small proteins. This is a known difficult situation in general (Andreeva 2007). Such treatment of sequence information is mainly for the purpose of method verification. Our goal here is to demonstrate that a very limited number of landmarks (residues) selected on the sequences, when combined with structure alignment method based only on shape, can significantly improve the structure alignment results. In principle, the sequence information can also be used directly without resorting to sequence alignment.

Joint Shape and Sequence Clustering: To evaluate these metrics further, we take another set of five structures – 1RIE, 1BRF.A, 1FQT.B, 1QS.W.A and 1VCK.A, which are fragments of larger proteins as shown in Figure 11. This example is taken from SISYPHUS database as a difficult case for structure alignment. The five structures can be aligned at beginning and ending parts of the structures and there are inserted loops of various lengths in the middle. This figure shows the full length structures and the fragments (red colored parts) that are actually used in the experiment. The goal is to determine if these fragments across structures are structurally similar or not. We start by computing pairwise geodesic distances between these parts in the composite space \( \tilde{S} \). The table of pairwise distances is given in Figure 12. Are these distances small enough to declare the original structures similar? Since we do not have a probability distribution for distances between class or across classes, we estimate it by randomly choosing 30 proteins from different classes in PDB for computing distances. We construct the sample statistics of those distances along with cut-offs for different confidence levels (assuming Gaussian statistics), see the histogram of these distances in the bottom of Figure 12. For instance, we estimate that the probability that a distance value less than \( \mu - 3\sigma \) (\( \mu = 1.0125, \sigma = 0.0678 \)) is obtained for different structures is approximately 0.15%. With this argument, we can see that all the structures in Figure 11 except the second one can be declared similar. The second structure is dissimilar to all others as it leads to larger pairwise distances. Indeed, it can be seen from Figure 11 that the second structure has a large insertion in the middle and quite different from all the other four structures.

Joint Shape and Geometric Label Clustering: In another experiment, we have derived clustering between the following proteins using the joint shape and geometric label distances. These 100 proteins are taken from 19 different classes of the ASTRAL SCOP
We record the time consumed of the experiment on a PC [experiment setups: CPU: Intel Core(TM)2 Duo T9900 @ 3.06GHz ; RAM: 8GB; OS: Windows 7(64-bit); Matlab: ver7.9.0.] The time for each individual case is shown in the following table:

<table>
<thead>
<tr>
<th>Process</th>
<th>Time consumed (in sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read protein backbone from PDB file</td>
<td>0.250455 ~ 62.1008372</td>
</tr>
<tr>
<td>Read secondary structure from DSSP file</td>
<td>0.004774 ~ 0.191958</td>
</tr>
<tr>
<td>Shape analysis for pairwise proteins</td>
<td>( \approx 43 )</td>
</tr>
<tr>
<td>Joint analysis for pairwise proteins</td>
<td>( \approx 43 )</td>
</tr>
</tbody>
</table>

The process of obtaining protein backbone varies depending on the size of the PDB file, and obtaining secondary structure varies depending on the size of DSSP file. For the 100 proteins, the automatic shape analysis process takes approximately 59 hours, while the joint analysis takes the same time.

5. SUMMARY

In summary, we have applied an elastic shape analysis to the problem of structure alignment of proteins. Although quite common in computer vision, this approach has not been applied to this problem as yet. This paper demonstrates the benefit of methodology-crossover from one field to another. We have shown that the elastic shape analysis can effectively deal with conformational changes of proteins; deformation of proteins from one conformation to another can be easily obtained, which may shed light on the dynamics of protein structures; incorporating sequence landmarks can significantly improve the structure alignment results; and average conformations can be calculated for capturing variations among similar conformations and classification of new structures.

6. REFERENCES


[24] RCSB. http://www.rcsb.org/pdb/home/home.do. PDB,