

# Bayesian Flexible Shape Matching with Applications to Structural Proteomics

Scott C. Schmidler\*

## Abstract

We introduce a method for flexible shape registration using Bayesian change-point analysis. Our approach is particularly suitable for shapes containing “hinge”-like flexibility, and is motivated by problems in structural proteomics and bioinformatics. We define a class of flexible shape spaces and an associated Procrustes-type metric, along with a highly efficient algorithm for computing flexible shape distance. We use this distance to define distributions over flexible shapes, and develop Bayesian models for several variants including both affine and rigid-body component transformations. We demonstrate the approach on several examples arising from protein structure analysis, including structure alignment and function discovery.

**Keywords:** Statistical shape analysis, changepoint analysis, Bayesian methods, bioinformatics, structural proteomics, protein structure alignment

---

\*Scott C. Schmidler is Assistant Professor, Department of Statistical Science, Duke University, Durham, NC 27708-0251. Tel: (919) 684-8064; Fax: (919) 684-8594; Email: schmidler@stat.duke.edu

# 1 Introduction

Data collected on geometric shapes arise in many diverse fields, including computer vision, archeology, astronomy, CAD design, anatomy and morphology, and molecular and cellular biology. Statistical analysis and comparison of shapes is an important challenge throughout these applications. In recent years, these problems have benefited from a unified treatment using techniques of statistical shape analysis and stochastic geometry (Stoyan and Stoyan, 1994; Stoyan et al., 1995; Small, 1996; Dryden and Mardia, 1998; Kendall et al., 1999).

In this paper we introduce methods for the analysis of *flexible* shapes, where sampled configurations from the same geometric object or a common population may represent significantly different mathematical shapes due to inherent variability in the object conformation, irrespective of measurement error. Our study of such problems is motivated especially by problems in molecular and structural biology and polymer chemistry. However, our approach has broader potential applications to many problems arising from the biological and physical sciences and the analysis of physical simulations.

We restrict our attention to landmark data, with objects represented by a set of corresponding loci (*landmarks*) chosen in a meaningful domain-specific way. Let  $X_{n \times d} \in \mathbb{R}^{nd}$  be a *configuration* matrix of  $n$  landmark coordinates in  $d$  dimensions representing an observation on an object, and  $Y_{n \times d}$  similarly another observation obtained either by repeated measurement of the same object or a distinct object. Denote by  $x_i \in \mathbb{R}^d$  the  $i^{th}$  row of  $X$  representing the  $i^{th}$  landmark coordinates, and  $y_i$  is the corresponding landmark on object  $Y$ . In our example of Section 2, the objects are protein molecules and  $x_i$  are 3-dimensional coordinates of atoms in space. In biological applications, corresponding landmarks often represent features sharing a

common evolutionary ancestor, and such points are referred to as *homologous*.

## 2 Bayesian Shape Matching and Protein Structure Alignment

In this and other work, we have developed a Bayesian framework for shape matching and alignment, motivated by problems in structural bioinformatics and proteomics. In applications in molecular biology, bioinformatics, and proteomics, often the landmark correspondences (homology relationships) are themselves unknown and subject to inference, considerably complicating the analysis. Previously we have developed a Bayesian framework for landmark matching in statistical shape analysis (Rodriguez and Schmidler, 2006; Schmidler, 2006, 2008). This work has applications to other areas such as object recognition in computer vision, where the landmark correspondences may also be unknown. We briefly outline this approach here before focusing on the particular component addressed in this paper. (A related approach has been developed independently in parallel by (Green and Mardia, 2006); similarities and differences between the approaches are discussed in (Rodriguez and Schmidler, 2006).)

### 2.1 Protein Structure Alignment

Protein structure alignment involves the comparison of two possibly related protein structures to determine common ancestry, function, fold, or substructural motifs Eidhammer et al. (2000). Structure alignment has a number of applications in structural and functional proteomics. It is used to search databases of molecular structures in order to identify potential homologous (evolutionarily related) proteins or other proteins with shared structural similarity. It is also used to classify newly observed

structures into known structural families. Both of these techniques are major tools for identification of protein function and mechanism. Structure alignment can also be used to estimate evolutionary distances for phylogenetic reconstruction over much larger timescales than those obtained from DNA and protein sequences. This is because the structure of a protein is much more strongly conserved than sequence, being more directly related to function. Often the homology of protein sequences which lie in the “twilight zone” threshold of statistically significant sequence similarity can be unambiguously determined through structural comparison.

Pairwise and multiple structure alignment are also key tools in studying conservation and variability of gene products at the molecular level in terms of detailed physical interactions, where the effects of genetic mutations are realized. Understanding inter- and intra-species variability of molecular structures can be a key component to designing new drugs with high specificity and low toxicity, for example.

The problem of protein structure alignment has a long history (see (Eidhammer et al., 2000) for a recent review), but has primarily been addressed from an algorithmic rather than statistical perspective. Here and in related work (Rodriguez and Schmidler, 2006; Schmidler, 2006, 2008; Wang and Schmidler, 2008) we introduce a formal Bayesian framework for shape matching and apply it to the problem of protein structure alignment.

## 2.2 Bayesian Pairwise Structure Alignment

Let  $X_{n_x \times 3}$  and  $Y_{n_y \times 3}$  be matrices of three-dimensional landmark coordinates on two proteins. By far the most common choice of landmarks for aligning proteins are the  $\alpha$ -carbons ( $C_\alpha$ 's) of the peptide backbone, but other possibilities include additional backbone atoms, sidechain centroids, key active site groups, or points sampled

from a Van der Waals or electrostatic surface representation (Connolly, 1983; Honig and Nicholls, 1995).

We define an *alignment* of  $X$  to  $Y$  as a pair  $(M, T(\cdot; \theta))$ , where  $T(X; \theta) \in \mathcal{T}$  is a registration transformation in some family  $\mathcal{T}$  parametrized by  $\theta$ , and  $M = (m_{ij})$  is an  $n_x \times n_y$  adjacency matrix for a bipartite graph representing a (possibly incomplete) *matching* between the landmarks  $X$  and  $Y$ :

$$m_{ij} = \begin{cases} 1 & \text{if } X_i \text{ is matched with } Y_j \\ 0 & \text{otherwise} \end{cases}$$

with  $\text{rank}(M) = n \leq \min(n_x, n_y)$  the number of matches. Where there is a possibility of ambiguity, we use  $M^{XY}$  to denote the matching for  $X$  to  $Y$ , otherwise we suppress the superscript. Choices for  $\mathcal{T}$  considered in this paper include  $SE(3)$  (rigid body) with parameters  $\theta = (\mu, R)$  and  $GL(3)$  (affine) with parameters  $\theta = (\mu, A)$  in Section 4, and piecewise-constant combinations of these (see Section 5).

We have developed a Bayesian approach to structure alignment, by defining a prior distribution on alignments  $P(M, \theta)$  and basing inferences on the posterior distribution:

$$P(M, \theta | X, Y) = \frac{P(X, Y | M, \theta)P(M, \theta)}{\sum_M \int_{\theta} P(X, Y | M, \theta)P(M, \theta)}$$

where the marginal likelihood  $P(X, Y)$  involves a sum over all possible matchings. Because the number of possible matchings is exponential in  $n_x$  and  $n_y$ , the posterior distribution may not be represented explicitly. However, expectations under the posterior may be approximated by Monte Carlo sampling from  $P(M, \theta | X, Y)$  by iteratively sampling from the conditional posteriors

$$P(M | X, Y, \theta) \quad \text{and} \quad P(\theta | X, Y, M)$$

to form a Gibbs sampler, or by reversible-jump Metropolis algorithm (Gilks et al., 1996; Green, 1995). Exact forms for the conditional distribution of  $M$  and efficient algorithms for sampling are provided elsewhere (Rodriguez and Schmidler, 2006; Wang and Schmidler, 2008).

In this paper, we focus on the other aspect of this problem, the inference of  $T(; \theta)$  conditional on a matching  $M$  and the submatrices  $X_M$  and  $Y_M$  whose corresponding rows consist of matched landmarks. As our focus here is on inferring  $T(; \theta)$  given  $M$ , we will without loss of generality suppress the subscripts and simply denote by  $X$  and  $Y$  the  $n \times 3$  submatrices  $X_M$  and  $Y_M$ . We develop an approach for Bayesian inference of  $T(; \theta)$  using a class of *flexible* transformations, and demonstrate this approach to the problem of protein structure alignment and analysis.

### 3 Shape Analysis

The *shape* of a configuration  $X$  is defined to be those (geometric) aspects of  $X$  which are invariant under rotation, translation, and scaling (Small, 1996; Dryden and Mardia, 1998; Kendall et al., 1999). Formally, the shape of  $X$  is the orbit of  $X$  under operations in the Euclidean similarity group, and is denoted by:

$$[X] = \{T(X) \mid T \in S(d)\} \tag{1}$$

where  $T(X) = \beta X R + \mathbf{1}_n^T \mu$  with  $R \in SO(d)$  a rotation matrix ( $RR^T = I$ ,  $\det R = 1$ ), and  $\mu \in \mathbb{R}^d$  a translation vector, and  $\beta \in \mathbb{R}^+$  a scale factor. The *shape space* is then the quotient space  $\mathbb{R}^{nd}/S(d)$ , denoted by  $\Sigma_n^d$  in the notation of (Kendall et al., 1999).

The *size-and-shape* of  $X$  removes rotation and translation but not scale, and is

given by:

$$[X]_S = \{T(X) \mid T \in SE(d)\} \quad (2)$$

where  $SE(d)$  is the special Euclidean isometry group of rigid-body transformations  $T(X) = XR + \mathbf{1}^T \mu$  without scaling. The size-and-shape space  $\mathbb{R}^{nd}/SE(d)$  is denoted by  $S\Sigma_n^d$ . Our applications in this paper are motivated primarily by the analysis of molecular structures, which are measured on a natural physical scale (as opposed to say, objects in images). Thus in later sections we focus primarily on size-and-shape here, often dropping the qualifier “size-and”, although our definitions and methods are general.

We will also make use of the *affine shape*, the orbit of  $X$  under the general affine group:

$$[X]_A = \{T(X) \mid T \in GA(d)\}$$

where  $T(X) = XA + \mathbf{1}^T \mu$  with  $A \in GL(d)$  a non-singular affine transformation. In addition to rigid motion, affine transformations also allow shear, scaling, and reflection.

More generally, we may define a notion of  $G$ -shape as an equivalence class of configurations under the action of a general transformation group  $G$  (see e.g. (Dryden, 1999)):

$$[X]_G = \{g(X) \mid g \in G\} \quad (3)$$

To date most work has restricted attention to  $G$  in  $S(d)$ ,  $SE(d)$ , or  $GA(d)$ . However, other classes of transformations (which need not be a group) have been explored, including thin-plate splines for planar shapes (Bookstein, 1989) and other smooth deformations (Amit et al., 1991). In Section 5 we will see that piecewise combinations of these transformations can yield an interesting new class of transformations.

### 3.1 Procrustes analysis and shape distance

Under the above definitions, two configurations  $X$  and  $Y$  have the same shape if  $\exists \theta : T(X, \theta) = Y$ . When  $[X] \neq [Y]$  or in the presence of population variability and measurement error, a useful notion of distance between shapes is given by the *Procrustes distance*, which is the minimum RMSD over all possible rigid-body transformations:

$$d_P^2(X, Y) = \min_{\substack{\mu \in \mathbb{R}^3 \\ R \in \mathbf{SO}(3)}} \|Y - (XR + \mathbf{1}\mu^T)\|^2 = \|X_c\|^2 + \|Y_c\|^2 - 2\text{tr}(D) \quad (4)$$

Here  $\|X\|^2$  denotes Frobenious norm  $\|X\|_F = \text{tr}(X^T X)^{\frac{1}{2}}$ ,  $X_c = CX$  denotes the centered coordinates with  $C = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ , and  $Y_c^T X_c = UDV^T$  is a *pseudo*-singular-value decomposition (SVD) with  $U, V \in SO(d)$  (Kendall et al., 1999).  $d_P$  is the Riemannian distance in the size-and-shape space  $S\Sigma_n^d$ . The minimizing registration parameters  $(R, \mu)$  are given by:

$$\hat{R} = VU^T \quad \text{and} \quad \hat{\mu} = \mu_y - \mu_x \quad (5)$$

$\mu_x = (I - C)X$  is the centroid (column means) of  $X$ . The result (5) is known as ordinary (partial) Procrustes analysis, and related techniques for least-squares minimization over the various transformation groups above are collectively referred to as Procrustes analysis in the statistical shape literature (Goodall, 1991; Dryden and Mardia, 1998).

The analogous affine shape distance is the minimum RMSD under affine transfor-

mations:

$$d_A^2(X, Y) = \min_{\substack{\mu \in \mathbb{R}^3 \\ A \in \mathbf{GL}(3)}} \|(XA + \mathbf{1}\mu^T) - Y\|^2 = \|Y\|^2 - \|Q^T Y\|^2 \quad (6)$$

and is simply the residual sum-of-squares from a multivariate regression obtained by writing the  $(d + 1) \times d$  matrix  $\tilde{A} = [\mu : A]^T$  and  $\tilde{X} = [\mathbf{1} : X]$  so  $d_A(X, Y)^2 = \|\tilde{X}\tilde{A}^* - Y\|^2$ . The minimizing transformation is then given by  $\hat{A} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T Y$ , or in the form of (5),  $\hat{A} = (X^T X)^{-1} X^T Y$  and  $\hat{\mu} = \mu_y - \mu_x$ . The optimal affine registration may be obtained efficiently using standard least-squares calculations, e.g. by taking the decomposition  $X = QR$  and solving  $RA = Q^T Y$  (see Section A.2). Note that  $d_A$  is not symmetric, but may be easily symmetrized.

We will see below that the notions of shape and shape distance can greatly simplify the representation of probability densities on molecular structures.

## 4 Bayesian Shape Registration

When analyzing data from populations of shapes or when object configurations are measured with error, shape comparison becomes a statistical question, with questions of significance, quantification of uncertainty, and prediction. The statistical analysis of shape data has developed rapidly in recent years (Small, 1996; Dryden and Mardia, 1998; Lele and Richtsmeier, 2001). Distribution theory for shape spaces has been studied extensively for planar shapes ( $d = 2$ ), but many of these results are difficult to generalize to  $d \geq 3$  due to the reliance on representation of landmarks as complex variables. We consider here two alternative approaches to specifying probability densities for shape analyses for  $d \geq 3$ , where our problems of interest lie.

## 4.1 Statistical models for registration

The least-squares registration (Procrustes analysis) techniques of the previous section have a natural interpretation as statistical models. Let  $\theta$  denote the parameters of registrations  $g \in G$ ; e.g.  $\theta = (\mu, \beta, R)$  for shape,  $(\mu, R)$  for size-and-shape,  $(\mu, A)$  for affine shape, and so on. Let  $Y$  be a random perturbation of  $X$  which has been transformed by  $g \in G$ :

$$Y = T(X; \theta) + \varepsilon \quad \text{vec}(\varepsilon) \sim N(0, \Sigma) \quad (7)$$

where  $\varepsilon$  is an  $n \times d$  matrix of random errors. For example, for size-and-shape ( $G = SE(d)$  and  $\theta = (\mu, R)$ ) this becomes

$$Y = XR + \mathbf{1}\mu^T + \varepsilon$$

with  $Y$  a rotated and translated perturbation of  $X$ . If  $\Sigma = \sigma^2 I_{nd}$  then  $\hat{\mu}, \hat{R}$ , and  $\hat{A}$  above are maximum likelihood estimates (MLEs). For more general  $\Sigma$  factored as  $\Sigma_n \otimes \Sigma_d$  with  $\Sigma_d = \sigma_d^2 I$ , MLEs are obtained from a weighted least squares procedure which may be solved by ordinary least squares by first pre-multiplying the coordinate matrices by  $\Sigma_n^{-\frac{1}{2}}$ . For general  $\Sigma_d$  Procrustes methods may still be used, but require more care (Goodall, 1991). Note that  $X$  given  $Y$  is also of the form  $X = Y\tilde{R} + \mathbf{1}\tilde{\mu}^T + \tilde{\varepsilon}$ , with  $\tilde{R} = R^{-1}$ ,  $\tilde{\mu} = -\mu R^{-1}$ , and  $\text{vec}(\tilde{\varepsilon}) \sim N(0, \tilde{\Sigma})$ , and if  $\Sigma = \sigma^2 I_{nd}$  then  $\tilde{\Sigma} = \Sigma$ .

A conditional density for  $Y | X$  may be obtained by conditioning on the MLEs to obtain the conditional (or *profile*) likelihood:

$$P(Y | X) = (2\pi|\hat{\Sigma}|)^{-\frac{3}{2}} e^{-\|Y - T(X; \hat{\theta})\|_{\hat{\Sigma}}^2}$$

e.g. for size-and-shape the exponent becomes  $\|Y - (X\hat{R} + \mathbf{1}\hat{\mu}^T)\|_{\hat{\Sigma}}^2$ .

#### 4.1.1 Marginal shape densities and Bayesian registration

A Bayesian approach treats the registration parameters  $\theta = (\mu, R, \Sigma)$  or  $(\mu, A, \Sigma)$  as random variables, placing prior distributions  $\pi_0(\theta)$  on these parameters to obtain posterior distributions  $P(\theta | X, Y)$ . A (conditional) density for  $Y$  is then given by the *predictive* distribution:

$$P(Y | X) = \int P(Y | X, \theta)\pi_0(\theta)d\theta$$

If  $\pi_0$  is uniform, this is closely related to the marginal “offset” shape densities of Dryden and Mardia (1998):

$$P(Y) = \int P(Y | X, \theta)d\theta$$

where  $Y \sim N(\mu, \Sigma)$ . Offset distributions have been studied extensively in  $d = 2$  dimensions, but are much more difficult to work with analytically in higher dimensions. Here we will always condition on both  $X$  and  $Y$ , so  $P(Y | X) \propto P(X, Y)$  and the marginal  $P(X)$  need not be specified.

**Bayesian affine registration** The predictive density approach is particularly simple in the case of affine shape, where the registration problem reduces to multivariate linear regression. Convenient priors are given by standard conjugate Bayesian analysis for the multivariate linear model (Box and Tiao, 1973; Gelman et al., 2004). Suppose  $\Sigma = \sigma^2 I_{nd}$ , then the non-informative uniform prior on  $A$  and  $\log \sigma$  given by

$P(A, \sigma^2) \propto \sigma^{-2}$  yields posterior

$$\begin{aligned} \text{vec}(A) | X, Y, \sigma^2 &\sim N(\text{vec}(\hat{A}), ((X \otimes \mathbf{1}_d)^T (X \otimes \mathbf{1}_d))^{-1} \sigma^2) \\ \sigma^2 | X, Y &\sim \text{Inv-}\chi^{-2}(\nu, d_A(X, Y)/\nu) \end{aligned}$$

for the registration parameters, where  $\nu = (n - d)d$ . When the number of landmarks  $n$  is large relative to dimension  $d$ , the non-informative prior will perform well.

To obtain a predictive distribution for  $Y | X$ , the prior distribution must be proper. Adopting the standard informative prior  $P(A, \sigma^2) \sim N(A_0, \Sigma_A) Ga(\frac{\nu}{2}, \frac{\nu\lambda}{2})$  yields posterior

$$\begin{aligned} A | X, Y, \sigma^2 &\sim N(\hat{A}, (\Sigma_0^{-1} + ((X \otimes \mathbf{1}_d)^T (X \otimes \mathbf{1}_d))^{-1} \sigma^2)^{-1}) \\ \sigma^2 | X, Y &\sim Ga(\nu, d_A(X, Y)/\nu) \end{aligned}$$

and resulting conditional density of  $Y$  obtained by integrating out  $\sigma^2$  is multivariate-t:

$$Y | X \sim t_{d^2}(\nu, XA_0, \lambda\Sigma_0).$$

Note that while  $A_0 \equiv 0$  is standard in regression,  $A_0 = I$  seems more sensible here.

**Bayesian rigid-body registration** Bayesian rigid-body shape or size-and-shape registration is more difficult, due to the necessity of integrating over  $R \in SO(d)$ , which cannot be done analytically in general (Goodall and Mardia, 1993). However, for prior  $P(R, \mu, \sigma^2) \propto \sigma^{-2} \pi_0(R)$  which is non-informative on  $\sigma^2$  and  $\mu$  the posterior distributions

$$P(R, \mu | X, Y, \sigma^2) \propto \frac{1}{(2\pi\sigma)^{\frac{dn}{2}}} e^{-\frac{1}{2\sigma^2} \|XR + \mathbf{1}\mu^T - Y\|^2} \quad R \in SO(d)$$

may be computed by Monte Carlo sampling. Direct sampling from this distribution appears difficult but can be achieved by Metropolis or rejection sampling.

Under informative priors on  $R$ , the predictive distribution may be obtained by sampling directly from the prior as long as  $\pi_0(R)$  not too concentrated, as  $SO(d)$  is compact and low-dimensional. There are many ways to parametrize  $SO(d)$ . For  $\pi_0$  diffuse the axis-angle representation  $(r, \phi)$  is convenient for  $d = 3$ , where taking  $\pi_0(R) = f(r)h(\phi)$  with  $f(r) = 1/4\pi$  uniform on the unit sphere  $S^2$  and  $h(\phi) = 1/2\pi$  is easily sampled. For informative priors, Wang and Schmidler (2008) represent  $R$  by *unit quaternions*.

## 4.2 Distributions on shapes

An alternative and in some ways simpler approach is to define rotationally symmetric distributions directly on shape space using the Riemannian metric given by the appropriate shape distance:

$$P([X]_G) = Z(G, n, d)^{-1} e^{-\kappa d_G^2(X, \gamma)} \quad (8)$$

For planar shape space, using the partial Procrustes distance (4) the normalization  $Z(S(2), n, 2)$  can be calculated (Dryden, 1991; Mardia and Dryden, 1999) by:  $Z_2 = 1 + \sqrt{\pi}(2\kappa)^{\frac{5}{2}-k}(k-2)!(I_{k-\frac{3}{2}}(4\kappa) + L_{k-\frac{3}{2}}(4\kappa))$  where  $I_\nu()$  is the modified Bessel function of the first kind and  $L_\nu()$  is the modified Struve function. However, normalization for shape or size-and-shape in  $d \geq 3$  appears to be difficult. Unfortunately this normalizing constant is needed for estimation of  $\sigma^2$  as well as for model selection problems such as those involved in landmark selection (Section 2) and flexible matching (Section 5).

However, we can circumvent this problem by normalizing (8) over configuration

space rather than shape space, which yields the multivariate normal distribution

$$P(Y | X, \sigma^2) = (2\pi\sigma)^{-\frac{d^*}{2}} e^{-\frac{1}{2\sigma^2} d_P^2(X,Y)} \quad (9)$$

when  $\Sigma = \sigma^2 I$ , where  $d^* = d(2n - d - 1)/2$  is the dimension of the tangent space. The Procrustes residuals may be interpreted as approximate tangent space coordinates, and so this distribution may be viewed as a density on the tangent space approximation to shape space located at pole  $X$ . Detailed discussions of tangent-space approximations to shape space are given by (Small, 1996; Dryden and Mardia, 1998).

Note that (9) is defined *conditional* on optimal transformation parameters  $\hat{\theta}$ , and therefore can be interpreted as a profile likelihood (say for  $\sigma$ ,  $M$  of Section 2, or  $(k, C^k)$  of Section 5) which maximizes over nuisance parameters  $\theta$ . Thus definition of likelihoods in shape space gives a profile likelihood in configuration space, and may suffer from previously identified problems of profile approaches. Nevertheless, for complex problems densities defined on (tangent space approximations to) shape space appear to provide the most straightforward approach to inference.

### 4.3 Protein structure alignment: rigid-body and affine

As described in Section 2, the protein structure alignment problem is a combination of landmark matching and registration. The registration problem may be addressed in a least-squares or maximum-likelihood framework using Procrustes techniques. Least-squares techniques for structure alignment have a long history, and recent reviews are given in (Eidhammer et al., 2000; Rodriguez and Schmidler, 2006). Figure 1 shows the use Procrustes analysis to superimpose two strongly related proteins, human deoxy-hemoglobin  $\beta$ -chain and sperm whale myoglobin, using rigid-body transformations.

Previous work (Wu et al., 1998) noted that affine superposition has computational

advantages for multiple structure alignment (see Section 6), yet performs similarly to rigid-body superposition. Since a major application of structure alignment is to search large databases such as the Protein Data Bank (Bernstein et al., 1977), or to perform cluster analysis on large numbers of structures, computational efficiency is a key consideration. We have also seen above that affine transforms also provide easier derivations of predictive shape distributions, and Appendix A shows they permit fast algorithms for flexible shape registration. Thus while rigid body transformations are more physically plausible, affine transformations have computational advantages and may be preferred if they yield results similar to rigid-body transforms.

As noted in Section 3, affine transformations allow shear, scaling, and reflection in addition to rigid motion when registering  $X$  to  $Y$ . An affine transform may be decomposed into  $\hat{A} = RDZ$  where a  $R$  is a pure rotation,  $D$  is a diagonal scale matrix, and  $Z$  is a unit upper-triangular shear matrix. These components may be obtained from  $\hat{A}$  by taking the Cholesky factor  $G$  of  $\hat{A}^T \hat{A}$  and setting  $D = \text{diag}(G)$ ,  $Z = D^{-1}G$ , and  $R = AG^{-1}$  (Wu et al., 1998).

Figure 1d shows the scale, reflection, and shear terms introduced in the pairwise affine globin alignment. It can be seen that that very little rescaling and shear is actually introduced in the affine registration: the diagonal elements are very nearly one and the off-diagonal elements nearly zero. Posterior credible intervals indicate these terms are not significant. This is likely due to data-imposed constraints maintaining the physical scale of distances between successive alpha-carbons along the backbone. While affine and rigid-body transforms may not be expected to yield identical results in many shape problems, for molecular structure we can take advantage of this to obtain increased computational efficiency.

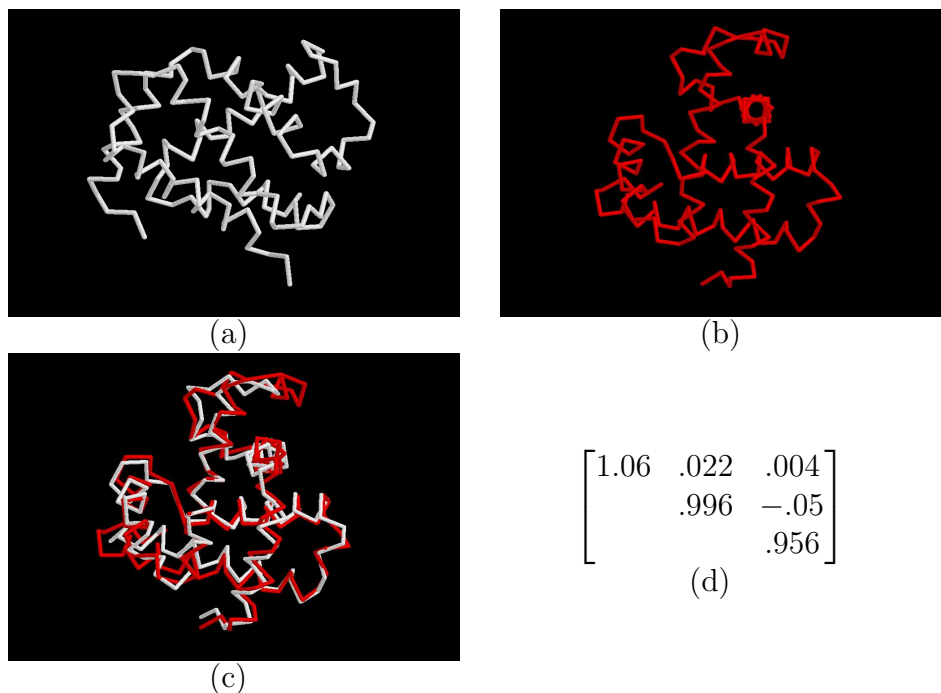


Figure 1: Procrustes techniques in protein structure alignment. Shown are (a) human hemoglobin  $\beta$ -chain (4HHBA), (b) sperm whale myoglobin (5MBN), (c) superimposition using rigid-body transformation. The affine superimposition (not shown) is visually indistinguishable from the rigid-body, as evidenced by (d) the associated scaling and shear components.

## 5 Flexible Shape and Bayesian Matching

It is often the case that geometric objects encountered in biological and engineering applications exhibit some inherent flexibility or internal degrees of freedom, allowing the adoption of alternative conformations. For example, proteins often have distinct domains separated by “hinge”-points or rotational degrees of freedom (Gerstein and Echols, 2004). The mathematical definitions of traditional shape analysis would imply that each such conformation has a distinct “shape”, as the orbits of the configuration coordinates under Euclidean transformations are distinct. However, in many applications it makes sense to consider that these configurations represent the same shape; hence here we extend the notion of shape by defining a notion of the *flexible*

*shape* of an object. Our notion of flexibility is motivated by problems in structural bioinformatics, but may be extended appropriately to other applications. Variance decomposition using principal warps (Bookstein, 1989) also allows a different type of flexibility, though here we use flexibility to define equivalence rather than as a descriptor of variation. Flexible transforms may still lie in the class of general transformation groups discussed by Dryden (1999).

In this paper we consider a specific notion of flexible shape based on changepoint processes. As with shapes above, flexible shape equivalence and comparison must be done in the presence of population variability and measurement error, requiring a statistical notion of flexible shape matching. The result is a statistical changepoint analysis problem on shape spaces.

## 5.1 Piecewise transformations and flexible shape

We begin by considering the transformation  $T(X)$  as a *sequence* of functions

$$T(X) = (T_1(x_1), T_2(x_2), \dots, T_n(x_n))$$

applied to each landmark coordinate vector  $x_i \in \mathbb{R}^3$ , where until now we have required that  $T_i(\cdot) \equiv T(\cdot)$  for each  $i$ . A *changepoint* will be an index  $j$  such that  $T_j(\cdot) \neq T_{j+1}(\cdot)$ , so that the transformation applied to  $X_{i \leq j}$  differs from that applied to  $X_{i > j}$ . Thus for a single changepoint in a rigid-body transform, we require two sets of rotation and translation parameters  $(\mu_1, R_1)$  and  $(\mu_2, R_2)$ . The resulting transformation is piecewise-constant as a function of  $i$ :

$$T_i(x_i) = \begin{cases} T_1(x_i) & i \leq j \\ T_2(x_i) & i > j \end{cases}$$

For aligning protein molecules, a transformation changepoint allows a hinge-like motion in the protein between amino acids  $j$  and  $j + 1$ . A flexible transformation may have multiple changepoints  $C^k = (c_1, \dots, c_k)$ , yielding a transformation with  $k + 1$  piecewise constant components. Denote by  $G^k$  the set of all transformations obtained by taking  $k$  piecewise constant component transformations from group  $G$ .

**Definition.** The *flexible shape* of a landmark configuration matrix  $X$  is the quotient space of  $X$  under  $G^k$ , i.e. the set  $\{g(X) : g \in G^k\}$  of configurations of  $X$  obtained under transformations in  $G^k$ .

Note that  $G^0$  is the traditional notion of shape given in (3). When  $G$  is a group,  $G^k$  is still a group, so we have simply expanded the class of transformation groups considered in (3), and  $G^0$  is a subgroup.

**Definition.** The *k-flexible G-shape distance* between  $X$  and  $Y$  is

$$d_{F_G^k}^2(X, Y) = \min_{\substack{C^k \in \mathcal{C}^k \\ g \in G}} \sum_{j=1}^k \|Y_j - g(X_j)\|^2 \quad (10)$$

where  $X_j = X_{[c_{j-1}+1:c_j]}$  and  $Y_j = Y_{[c_{j-1}+1:c_j]}$ , and  $\mathcal{C}^k$  are all  $k$ -subsequences of  $\mathbb{Z}_n$ .

Note that, as with the affine and similarity group distances, these distances are in general asymmetric, and similar procedures can be used to symmetrize them. However, the *k-flexible size-and-shape distance* is symmetric, and is given by

$$d_{F_S^k}^2(X, Y) = \min_{C^k, \mathbf{u}, \mathbf{R}} \sum_{j=1}^k \|Y_j - (X_j R_j + \mathbf{1} \mu_j^T)\|^2 = \min_{C^k} \sum_{j=1}^k \|Y_j\|^2 + \|X_j\|^2 - 2\text{tr}(D_j)$$

where  $\mathbf{u} = (\mu_1, \dots, \mu_k)$ ,  $\mathbf{R} = (R_1, \dots, R_k)$ , and  $Y_j^T X_j = U_j D_j V_j^T$ .

Certain restrictions are required to make this definition practical. In particular, if  $|c_i - c_{i-1}| < p/d$  for  $p$  the degrees of freedom in transformations  $g \in G$ , then  $g_k$  is un-

derdetermined, thus we define  $\mathcal{C}_k \ni C^k$  to be the set of  $C^k$  such that  $|c_i - c_{i-1}| \geq p/d$ . Note also that any two configurations  $X$  and  $Y$  have the same  $k$ -flexible shape for  $k = n/dp$ , so in practice we may wish to define a bounding value  $K$  to speak of  $K$ -flexible shape equivalence. In the presence of population variability and measurement error, testing equivalence becomes a statistical problem which inherently involves model selection. This requires a prior distribution or penalty on  $k$  and/or  $C^k$ , since otherwise  $d_{F_G^k} \leq d_{F_G^{k'}}$  whenever  $k' < k$ . It may also be desirable to add a smoothness penalty on the sequence of transformations  $\{T_i\}$  to (10); this is discussed in Section 5.2.

## 5.2 Bayesian changepoint analysis

In order to compare two or more observed configurations under flexible transformations  $G^k$ , we must also account for measurement error and population variability in the population (flexible) shape. Identification of the existence and location of such flexion points, along with the associated transformation parameters  $(\mu_1, R_1, \mu_2, R_2)$ , from only the observed data ( $X$  and  $Y$ ) becomes a problem in changepoint analysis.

Bayesian analysis of changepoint problems (Carlin et al., 1992; Barry and Hartigan, 1993; Stephens, 1994) proceeds by defining a prior distribution  $P(k, C^k)$  over the number and location of changepoints, and obtaining the associated posterior distribution. Many variations on changepoint analysis exist, and are sometimes referred to as *segmentation* problems (Schmidler et al., 2000, 2004). Adopting here the model of Section 4.2, we define the shape space likelihood for a set of changepoints by:

$$P(X, Y | k, C^k) = Z_{F_G^k} e^{-d_{F_G^k}^2(X, Y)} = \prod_{j=0}^k Z_j^{-1} e^{-d_G^2(X_{[s_j:c_j]}, Y_{[s_j:c_j]})}$$

ignoring constant  $P(X)$ , to obtain the posterior distribution over changepoints

$$P(k, C^k | X, Y) = \frac{P(X, Y | k, C^k)P(k, C^k)}{\sum_{k'} \sum_{C^{k'}} P(X, Y | k', C^{k'})P(k', C^{k'})}.$$

Quantities of interest in Bayesian changepoint analysis include the marginal posterior distribution of the number of changepoints  $k$ , the marginal posterior probability of a changepoint between landmarks  $i$  and  $i+1$ , and the *maximum a posteriori* (MAP) set of changepoints and associated transformation parameters:

$$\arg \max_{k, C^k, \theta^k} P(X, Y | k, C^k)P(C^k | k)P(k)$$

where  $\theta^k = (\theta_1, \dots, \theta_k)$  are the transformation parameters for each region  $\theta_i = (\mu_i, R_i)$ . For flexible shape matching and database searching, we may also be interested in the marginal posterior mean-squared deviation:

$$P(d_{P_f}(X, Y) < \epsilon) = \sum_{k=0}^K \sum_{C^k} P(X, Y | k, C^k)P(C^k | k)P(k)$$

Note that the predictive distributions of Section 4.1.1 may also be applied here. If independent priors are assigned for each  $\theta_i$ , the posteriors are conditionally independent given  $(k, C_k)$  and follow the analysis in Section 4; the individual posterior means are obtained by Procrustes analysis on  $X_{[s_j, c_j]}, Y_{[s_j, c_j]}$ , and in the affine case the conditional predictive distribution for  $Y$  becomes simply

$$\begin{aligned} P(Y | X, k, C_k) &= \int \cdots \int P(Y | X, k, C_k, \theta) \pi_0(\theta) d\theta_1 \dots d\theta_k \\ &= \prod_{j=1}^k t_{d^2}(\nu, X_j A_0, \lambda \Sigma_0) \end{aligned}$$

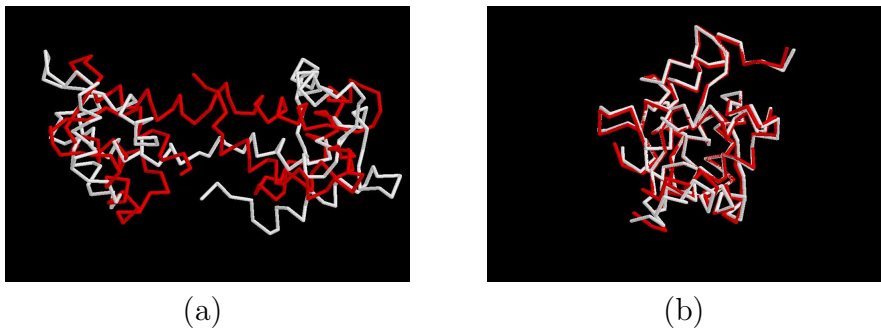


Figure 2: (a) Alignment of two Calmodulin structures using rigid-body transformations results in a minimum RMSD of 11.97Å, a non-significant match. (b) Alignment of the same structures using Bayesian flexible alignment yields an RMSD of 0.7Å with a single changepoint, a highly significant match.

and the marginal predictive distribution for  $Y$  is given by

$$P(Y | X) = \sum_{k=1}^K \sum_{C^k} \left[ \prod_{j=1}^k t_{d^2}(\nu, X_j A_0, \lambda \Sigma_0) \right] P(C_k | k) P(k)$$

which can be computed efficiently using the recursions of Appendix A.

Figure 2 shows an example of protein structure alignment using two different solution NMR structures of calmodulin (1cfc). The coordinates represent two different conformations of the same molecule so the matching matrix  $M$  is known. Figure 2a shows the optimal rigid body transformation between the two structures, while Figure 2b shows the MAP single-changepoint transformation. The flexible model provides a clear improvement, yielding a highly significant RMSD where the rigid approach yields a highly insignificant RMSD.

**Enforcing continuity with hierarchical and smoothing priors** In some cases it may be desirable to enforce continuity on transformations  $\theta_1, \dots, \theta_k$ . For example, we may allow only a single global translation  $\mu$  with rotations  $R_1, \dots, R_k$ . More generally, we may wish to impose dependent prior distributions on the transformations  $\theta_1, \dots, \theta_k$ . For example, hierarchical priors arise naturally under assumptions

of exchangeability:

$$P(\theta^k | k) = \int \prod_{j=1}^k \pi_0(\theta_j | \omega) f(\omega) d\omega$$

and require only minor modification of the computational algorithms of Appendix A. Alternatively, smoothing priors to avoid abrupt changes in transforms may be desirable. For affine transformations, the latter will yield a Kalman-filter-like modification to the recursions of Appendix A. For rigid-body rotations, recent work on quaternion Kalman filters can yield similarly efficient calculations.

**Algorithms** Applications of shape analysis often require large numbers of shape comparisons in real time; examples include searching databases of images or molecules, or object tracking and recognition in computer vision. It is therefore critical that flexible matching be nearly as fast as rigid matching. As the Bayesian flexible alignment approach requires performing a large number of registrations, highly efficient algorithms for exact calculation of the necessary quantities are given in the Appendix.

## 6 Bayesian Flexible Mean Shape

A common task in shape analysis is to estimate the (shape of) the mean of a population of shapes, denoted by  $[M]$ , from multiple observations  $X^1, \dots, X^m$ . Here we generalize the notion of mean shape to *mean flexible shape*, allowing each observed configuration to be a *flexible* perturbation of the mean.

## 6.1 Mean shape and mean flexible shape

A mean shape estimate  $[\hat{M}]$  for observed configurations  $X^1, \dots, X^m$  can be obtained by minimization of

$$[\hat{M}]_G = \arg \min_{\theta, M} \sum_{j=1}^m \|T(X^j; \theta_j) - M\|^2 = \arg \min_M \sum_{j=1}^m d_G^2(X^j, M) \quad (11)$$

which is the profile maximum likelihood estimate under the error model  $X^j = T(M, \theta_j) + \varepsilon_j$ , the population model analog of (7).  $\hat{M}$  may be obtained by *generalized Procrustes analysis* (GPA) (Goodall, 1991; Dryden and Mardia, 1998) for the various transformation groups of Section 3, but is particularly efficient for affine shape, where  $[\hat{M}]_A = \arg \min_M \sum_{j=1}^m d_A^2(X^j, M)$  is given subject to  $M^T M = I$  by maximizing

$$\sum_{j=1}^m \text{tr}(M^T (I - H_j) M) = m \text{tr}(M^T (I - \bar{H}) M)$$

where  $H_j = X_j(X_j^T X_j)^{-1} X_j^T$  is the projection operator for the  $j^{\text{th}}$  shape, with solution  $\hat{M}$  the three largest eigenvectors of  $\bar{H}$  (Hastie and Kishon, 1991; Wu et al., 1998). For rigid shape or size-and-shape in  $d > 2$ , GPA requires an iterative algorithm alternating between computing  $\{(\hat{R}_j, \hat{\mu}_j)\}_{j=1}^m$  using (5), and averaging  $\hat{M} = m^{-1} \sum_{j=1}^m \hat{X}_j$ .

We now generalize (11) to define the mean *k-flexible G-shape* estimate by

$$[\hat{M}]_{F_G^k} = \arg \min_M \sum_{j=1}^m d_{F_G^k}^2(X^j, M) \quad (12)$$

which is the profile maximum likelihood estimate under the error model  $X^j = g(M, C^k, \theta) + \varepsilon_j$  with  $C^k$  and  $\theta_1, \dots, \theta_k$  as nuisance parameters.

Although each configuration may exhibit flexibility at different changepoints, the flexibility of the mean shape must be the union of all individual changepoints. Thus

change points have the interpretation as hinges or rotatable ball-and-socket joints in the population mean (flexible) shape, and each observed configuration may or may not exhibit a flexion at any given change point.

## 6.2 Bayesian flexible multiple alignment

### 6.2.1 Bayesian estimation of mean shape

Bayesian estimation of mean shape involves placing a prior distribution on  $M$ . If  $P(M)$  is specified as an exponential of a quadratic function, the smoothing prior

$$P(M) = Z^{-1} \exp(-\lambda M^T \Omega M)$$

corresponds to a Gaussian process prior on  $M$ . Then Procrustes analysis techniques may be used with minor modification to find the MAP estimate of  $M$ . More general priors such as nonparametric or structure priors may be desirable when few configuration matrices are available for estimation or some landmarks are frequently unobserved; this can be accomplished by Monte Carlo estimation (Schmidler, 2008).

### 6.2.2 Bayesian flexible mean shape

The Bayesian flexible alignment approach of Section 5 is equally applicable to mean flexible shape estimation. The likelihood becomes

$$P(X^1, \dots, X^m \mid k, C^k) = \prod_{j=1}^k Z_j^{-1} e^{-\sum_{j=1}^m d_P^2(X_{[s_j^k:c_j^k]}^1, \bar{X}_{[s_j^k:c_j^k]})}$$

Note that when  $m = 2$ , this offers yet another alternative shape density to the conditional forms given in Section 4, specifying the joint distribution  $P(X, Y \mid M)$  instead of factoring  $P(X, Y) = P(Y \mid X)P(X)$  and canceling  $P(X)$  as in the regression case.

The posterior distribution of changepoints and the MAP alignment may be computed recursively exactly as before. Again the affine averaging yields a substantial computational advantage over the rigid-body case: estimation of  $\hat{M}$  is non-iterative, and updating of  $(A_{[i,j]}^1, \dots, A_{[i,j]}^m)$ , and  $\bar{X}_{[i,j]}$  to  $(i+1, j)$  or  $(i, j+1)$  is obtained quickly via the least-squares updating techniques of Appendix A.

### 6.3 Flexible multiple protein structure alignment

Multiple protein structure alignment involves the simultaneous alignment of a set of structures  $X^1, X^2, \dots, X^m$  (Wu et al., 1998; Wang and Schmidler, 2008). Multiple alignment is used to analyze families of related proteins for studying conservation and function, to analyze simulation output, to cluster protein structure databases, and to develop models for classification of structures into families and folds.

The issues discussed previously giving rise to the need for flexible pairwise alignment apply equally to multiple alignment. If several pairwise flexible alignments are obtained from a database search, it is desirable to multiply align them to build a family model. And when clustering databases, we often need to compute the mean of a subset of structures, and may wish to allow flexible alignments while doing so. As above, we assume that the matchings are given, and treat only the registration problem here. Methods for matching in the multiple alignment problem are given by (Wang and Schmidler, 2008; Schmidler, 2008).

Application of the Bayesian flexible mean shape estimation to the problem of multiple flexible protein structure alignment is shown in Figures 3a and 3, where the MAP flexible alignments are shown using both rigid and affine multiple alignments.

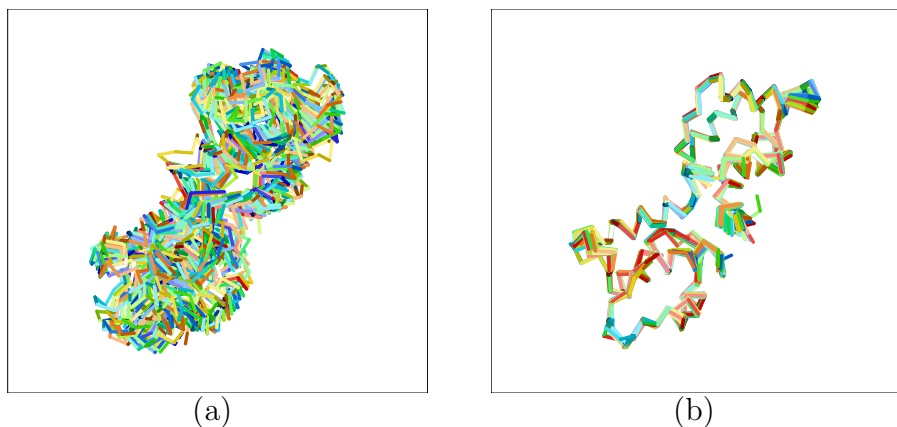


Figure 3: Multiple alignment of Calmodulin structures using (a) rigid-body mean-shape estimation, and (b) Bayesian estimation of mean flexible shape (MAP estimate).

## 6.4 Identification of disordered, flexible, and hinge regions

Our Bayesian flexible shape matching approach may also be applied to the analysis of multiple structures in order to determine regions of flexibility and mobility. Such regions in a protein are often closely associated with function and mechanism, and can provide key insights into the activity of the protein.

Figure 4 shows an analysis of two structures of a triose phosphate isomerase (TIM) barrel enzyme using our approach. While the two structures align quite well with a rigid body transform, the posterior distribution on  $k$  the number of changepoints shows a peak at  $k = 2$ , and closer inspection shows the joint posterior identifies two hinge sites responsible for the flexible loop “flap”. This flap is located at the enzyme active site and plays a key role in substrate binding and dissociation. Location of such flexible regions, or more subtle allosteric flexibility involving multiple flexion points, can provide key insights into the function and mechanism of a protein.

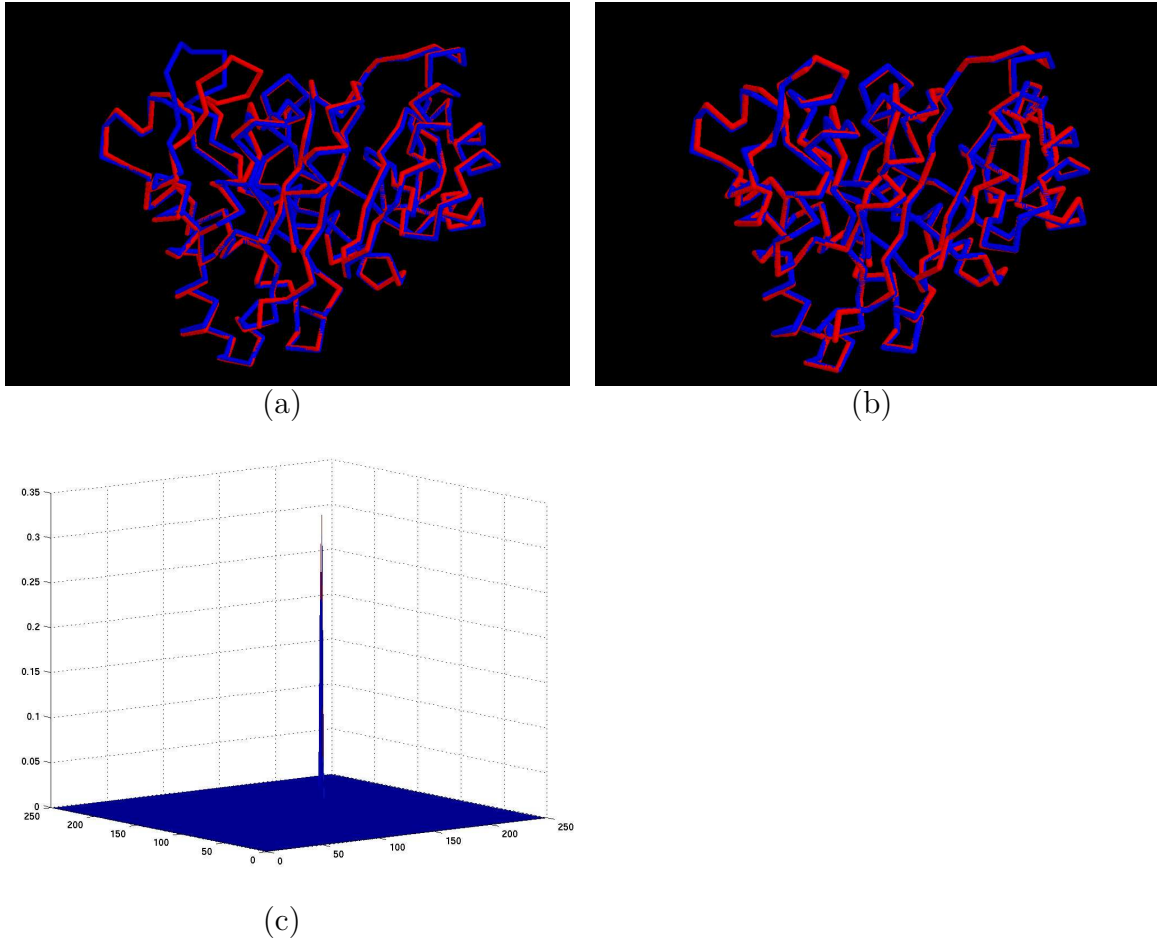


Figure 4: Alignment of two TIM barrel structures using (a) rigid-body transformations and (b) Bayesian flexible alignment. The resulting marginal posterior joint distribution for two hinge-points (c) clearly identifies the flexible loop at the active site of the enzyme (upper left-hand corner of (a)).

## 6.5 Distance-matrix estimates of mean shape

In some situations Procrustes methods lack of asymptotic consistency, and alternative estimators of mean shape have been proposed (Kent and Mardia, 1997; Dryden and Mardia, 1998; Lele and Richtsmeier, 2001). Lele *et al.* have proposed an alternative estimator obtained from a method-of-moments bias correction to the multidimensional scaling (MDS) estimate. The MDS estimate is obtained by constructing the  $n \times n$  inter-landmark (squared) distance matrix  $D^r = (d_{ij}^r)$  for  $r = 1, \dots, m$  with

$d_{ij}^r = \|x_i^r - x_j^r\|^2$ , and computing a  $d$ -dimensional MDS of the mean distance matrix:

$$\bar{D} = \frac{1}{m} \sum_{r=1}^m D^r \quad \text{and} \quad \hat{M} = [\lambda_1 \gamma_1 \quad \lambda_2 \gamma_2 \quad \dots \quad \lambda_d \gamma_d]$$

where  $\lambda_1, \dots, \lambda_n$  are the eigenvalues of  $\bar{D} = \Gamma \Lambda \Gamma^T$  in decreasing order, and  $\gamma_1, \dots, \gamma_n$  the associated unit eigenvectors.  $\bar{D}$  is biased, and the EDMA estimator (Lele and Richtsmeier, 2001) uses a bias corrected  $\tilde{d}_{ij} = \frac{1}{m} \sum_{r=1}^m d_{ij}^r + b_{ij}$  before performing MDS on  $\tilde{D}$  to obtain the mean-shape estimate. ( $\hat{D}$  and  $\tilde{D}$  estimate of *reflection-size-and-shape* rather than *size-and-shape*, as conversion to inter-landmark distances loses information on handedness. Since chirality of molecules critically important, we prefer Procrustes techniques for this application, but the empirical results of Section 4.3 for affine transforms may carry over to EMDA. The definitions, shape-space models, and algorithmic machinery of flexible shape given in previous sections apply independently of which estimator for mean shape is used.

## 7 Extensions and Discussion

**Bayesian Flexible Structure Alignment** The probabilistic framework for structure alignment given in Section 2 can be extended to perform flexible structure alignments. Flexibility is a crucial aspect of protein function (Gerstein and Echols, 2004). Many proteins exist in the database in different conformations, making them difficult to match by traditional methods - even two structures of the same protein may be difficult to align (see Figure 2). Alignment of structures allowing for such flexibility is a difficult task, and until very recently no fully automated algorithms existed for this problem. An important issue is permitting sufficient flexibility without allowing too much, since any backbone configuration may be transformed to any other using

sufficient degrees of freedom. The Bayesian approach provides a natural formalism for balancing the prior preference for low flexibility against observed fit to the data. Work is ongoing to combine the Bayesian alignment methods of Section 2 with the flexible-shape models described here.

**Flexible shape classification and clustering** From the definitions of flexible shape and shape distance, distributions on flexible shape populations, and statistical inference procedures introduced in Sections 5 and 6, it is straightforward to develop Bayes classifiers and model-based clustering of flexible shapes using mixture models

$$P(x) = \sum_c P(X | c)w_c = \sum_c w_c Z_G^k(X) e^{-d_{FG}^k(X, \gamma_c)} \quad (13)$$

where  $c$  is a class label. The notion of flexible shape distance  $d_{FG}^k$  given by (10) may also be useful in more complex kernel classification methods; classification and clustering of protein structures is an active area of research.

**Other forms of shape flexibility** The notion of shape flexibility introduced in this paper centers around hinge-like flexion in ordered landmark sets, appropriate for the applications studied here. When landmarks have no natural ordering, the change-point hinges must be defined as (hyper)planes in  $\mathbb{R}^d$  defined (perhaps) by subsets of landmarks. However the efficient algorithms described are no longer applicable and analysis will be significantly more computationally demanding. This is an interesting avenue for future research, as is exploring flexible shape under other forms of flexibility beyond hinge-like flexion, where related work on smooth and stochastic deformations of geometric objects exists (Bookstein, 1989; Amit et al., 1991).

# Acknowledgments

This work was partially supported by NSF grant DMS-0605141 (SCS).

## A Appendix: Algorithms

### A.1 Recursions for marginal posteriors

Posterior quantities of Section 5.2 such as marginal posterior distributions of  $k$  and  $z_i$ , or MAP estimates  $G_{\text{MAP}} = (k_{\text{MAP}}^*, C_{\text{MAP}}^k, \theta_{\text{MAP}}^k)$ , may be calculated exactly and efficiently using dynamic programming recursions related to stochastic segment models (Schmidler et al., 2004). Denote by  $\psi(i, j)$  any of the (normalized) densities  $\int_{\sigma^2} P(X_{[i:j]}, Y_{[i:j]} | \sigma^2) P(\sigma^2)$  described in Section 4. Note that if the  $\theta_i$ 's are independent *a priori*, then

$$P(X, Y | k, C_k) = \int_{\theta} \prod_{j=1}^k P(X_{[s_j:e_j]}, Y_{[s_j:e_j]} | \theta_i) P(\theta) = \prod_{j=1}^k \psi(s_j, e_j)$$

We first construct the matrix  $\Psi_{i < j} = (\psi(i, j))$  by computing all  $\binom{n}{2}$  sub-sequence Procrustes distances (Section A.2), then compute the forward/backward variables:

$$\alpha(j, k) = \sum_{i < j} \alpha(i, k-1) \psi(i+1, j) \quad \beta(j, k) = \sum_{i > j} \beta(i, k-1) \psi(j+1, i)$$

Then  $P(X, Y) = \sum_k p(k) \alpha(n, k)$  and  $P(k | X, Y) = \frac{p(k) \alpha(n, k)}{P(X, Y)}$ , and

$$P(z_i | X, Y) = \sum_k P(z_i | k, X, Y) P(k | X, Y) = \sum_k P(k | X, Y) \sum_{j=1}^k \alpha(i, j) \beta(i, k-j)$$

The calculation of  $P(z_i | X, Y)$  may be done in  $O(K)$  where  $K \leq n/p$  is the maximum allowed number of changepoints, by recursively computing  $\tilde{\beta}(i, k) = \sum_{j=0}^{K-k} \beta(i, j) = \tilde{\beta}(i, k+1) + \beta(i, k)$  and then  $P(z_i | X, Y) = \sum_k P(k | X, Y) \alpha(i, k) \tilde{\beta}(i, k)$  and hence the entire algorithm takes  $O(nK)$ . The MAP flexible registration may be computed by replacing sums with maximizations in the above recursions. Bayes factors may also be computed trivially by  $P(k_1 | X, Y)/P(k_2 | X, Y)$

When  $\theta_i$  are not *a priori* independent, e.g. when regions share a common  $\sigma^2$  or in the hierarchical setup of Section 5.2, these recursions may be adapted to perform conditional sampling yielding a highly efficient blocked Gibbs sampler. A special case is the smoothed rotation model of Section 5.2, which may be computed exactly using a Kalman filter-like modification of the above recursions.

## A.2 Least-squares updating of flexible registrations

**Affine registration updating** In addition to analytical tractability of shape densities, affine transformations also permit computational speedups through use of least-squares updating methods. To obtain affine shape distance  $d_A$  we may decompose  $X = QR$  with  $Q_{n \times 3}$  orthogonal and  $R_{3 \times 3}$  upper triangular, and obtain  $\hat{A}$  by solving  $R\hat{A} = Q^T Y$  by back-substitution (Golub and Van Loan, 1996). The residuals are given by  $(I - H)Y$  where  $H = QQ^T$  is the projection operator which gives  $X\hat{A} = HY$ .

*QR decomposition updating:* To compute flexible shape distances via the recursions of Section A, we require the matrix  $\Psi$  of all  $\binom{n}{2}$  sub-sequence Procrustes distances. Rather than perform  $\binom{n}{2}$  separate Procrustes analyses, we can use  $QR$  updating to compute  $\psi(i, j)$ 's sequentially. First we construct  $X_c^T X_c = \sum_{i=1}^n x_i x_i^T$  sequentially by

$$X_{[1:i+1]}^T X_{[1:i+1]} = X_{[1:i]}^T X_{[1:i]} + x_{i+1} x_{i+1}^T \quad \text{and} \quad X_{[i+1:j]}^T X_{[i+1:j]} = X_{[i:j]}^T X_{[i:j]} - x_i x_i^T$$

We then wish to obtain the  $QR$  decomposition of  $X_{[i:j+1]}^T X_{[i:j+1]}$  or  $X_{[i+1:j]}^T X_{[i+1:j]}$  from that of  $X_{[i:j]}^T X_{[i:j]}$ , i.e. when a row is added to or deleted from  $X$ . This differs slightly from the typical setup of variable selection in regression, in that we need to update the  $QR$  decomposition of  $X^T X$  rather than  $X$ . The  $QR$  updating may be done efficiently by techniques given in (Lawson and Hanson, 1995) without loss of generality, let  $X = X_{[i:j]}$  and  $\tilde{X} = X_{[i:j+1]}$ . If  $X = QR$ , then

$$\tilde{X} = J_{j,d} J_{j,d-1} \dots J_{j,1} Q [R^T : 0 : x_j]^T = \tilde{Q} \tilde{R}$$

for Givens rotations  $J_{a,b}$  yielding  $\tilde{R}$  upper triangular (Golub and Van Loan, 1996). Related but less-developed techniques exist for updating SVDs which may be applicable shape and size-and-shape flexible registration, which we are exploring. In practice, flexible alignment of two 300 residue proteins is very fast.

## References

- Amit, Y., Grenander, U., and Piccioni, M. (1991). Structural image restoration through deformable templates. *J. Amer. Statist. Assoc.*, 86(414):376–387.
- Barndorff-Nielsen, O. E., Kendall, W. S., and van Lieshout, M. N. M., editors (1999). *Stochastic Geometry: Likelihood and Computation*, volume 80 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.
- Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.*, 88:309–319.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein

- Data Bank: a computer-based archival file for macro-molecular structures. *J. Mol. Biol.*, 112:535–542.
- Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intel.*, 11(6):567–585.
- Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Wiley.
- Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Appl. Statist.*, 41(2):389–405.
- Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221:709–13.
- Dryden, I. L. (1991). Discussion to 'Procrustes methods in the statistical analysis of shape' by C. R. Goodall. *J. Roy. Stat. Soc. B*, 53:327–328.
- Dryden, I. L. (1999). *General Shape and Registration Analysis*, pages 333–364. Volume 80 of Barndorff-Nielsen et al. (1999).
- Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. Wiley.
- Eidhammer, I., Jonassen, I., and Taylor, W. R. (2000). Structure comparison and structure patterns. *J. Comp. Biol.*, 7(5):685–716.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall, 2<sup>nd</sup> edition.
- Gerstein, M. and Echols, N. (2004). Exploring the range of protein flexibility, from a structural proteomics perspective. *Curr. Opin. Chem. Biol.*, 8:14–19.

- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
- Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins University Press, 3<sup>rd</sup> edition.
- Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *J. Roy. Stat. Soc. B*, 53(2):285–339.
- Goodall, C. R. and Mardia, K. V. (1993). Multivariate aspects of shape theory. *Ann. Statist.*, 21(2):848–866.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–32.
- Green, P. J. and Mardia, K. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. (*preprint*).
- Hastie, T. and Kishon, E. (1991). Discussion to Goodall (1991). *J. Roy. Stat. Soc. B*, 53:330–331.
- Honig, B. and Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, 268:1144–9.
- Kendall, D. G., Barden, D., Carne, T. K., and Le, H. (1999). *Shape and Shape Theory*. Wiley.
- Kent, J. T. and Mardia, K. V. (1997). Consistency of Procrustes estimators. *J. Roy. Stat. Soc. B*, 59(1):281–290.
- Lawson, C. L. and Hanson, R. J. (1995). *Solving Least Squares Problems*. SIAM.

- Lele, S. R. and Richtsmeier, J. T. (2001). *An Invariant Approach to Statistical Analysis of Shapes*. Chapman & Hall.
- Mardia, K. V. and Dryden, I. L. (1999). The complex Watson distribution and shape analysis. *J. Roy. Stat. Soc. B*, 61:913–926.
- Rodriguez, A. and Schmidler, S. C. (2006). Bayesian protein structure alignment. (submitted to *Annals of Applied Statistics*).
- Schmidler, S. C. (2006). *Fast Bayesian Shape Matching Using Geometric Algorithms (with discussion)*., pages 471–490. Oxford University Press, Oxford.
- Schmidler, S. C. (2008). Bayesian landmark matching and mean shape estimation. (*in preparation*).
- Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2000). Bayesian segmentation of protein secondary structure. *J. Comp. Biol.*, 7(1):233–248.
- Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2004). Stochastic segment interaction models for biological sequence analysis. (submitted to *J. Amer. Statist. Assoc.*).
- Small, C. G. (1996). *The Statistical Theory of Shape*. Springer.
- Stephens, D. A. (1994). Bayesian retrospective multiple-change-point identification. *Appl. Statist.*, 43(1):159–178.
- Stoyan, D., Kendall, W. S., and Mecke, J. (1995). *Stochastic Geometry and its Applications*. Wiley, 2<sup>nd</sup> edition.
- Stoyan, D. and Stoyan, H. (1994). *Fractals, Random Shapes and Point Fields: Methods of Geometric Statistics*. Wiley.

Wang, R. and Schmidler, S. C. (2008). Bayesian multiple protein structure alignment and analysis of protein families. (in preparation).

Wu, T. D., Schmidler, S. C., Hastie, T., and Brutlag, D. L. (1998). Regression analysis of multiple protein structures. *J. Comp. Biol.*, 5(3):597–607.