Bioinformatics, 32(5), 2016, 755–763
doi: 10.1093/bioinformatics/btv648
Advance Access Publication Date: 4 November 2015
Original Paper

OXFORD

Bioimage informatics

Shape component analysis: structure-preserving dimension reduction on biological shape spaces

Hao-Chih Lee[1], Tao Liao[2], Yongjie Jessica Zhang[1,2] and Ge Yang[1,3,*]

[1]Department of Biomedical Engineering, [2]Department of Mechanical Engineering, and [3]Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 6, 2015; revised on September 21, 2015; accepted on October 18, 2015

Abstract

Motivation: Quantitative shape analysis is required by a wide range of biological studies across diverse scales, ranging from molecules to cells and organisms. In particular, high-throughput and systems-level studies of biological structures and functions have started to produce large volumes of complex high-dimensional shape data. Analysis and understanding of high-dimensional biological shape data require dimension-reduction techniques.

Results: We have developed a technique for non-linear dimension reduction of 2D and 3D biological shape representations on their Riemannian spaces. A key feature of this technique is that it preserves distances between different shapes in an embedded low-dimensional shape space. We demonstrate an application of this technique by combining it with non-linear mean-shift clustering on the Riemannian spaces for unsupervised clustering of shapes of cellular organelles and proteins.

Availability and implementation: Source code and data for reproducing results of this article are freely available at https://github.com/ccdlcmu/shape_component_analysis_Matlab. The implementation was made in MATLAB and supported on MS Windows, Linux and Mac OS.

Contact: geyang@andrew.cmu.edu

1 Introduction

Geometrical shapes are a fundamental property of biological structures. Quantitative shape analysis is required by many biological studies across diverse scales. For example, at the molecular scale, quantitative analysis of shapes of proteins is essential for understanding their functions and interactions (Chandonia and Brenner, 2006). As another example, at the cellular scale, quantitative analysis of shapes of cells is essential for understanding their morphogenesis and migration (Keren *et al.*, 2008). Recently, high-throughput and systems-level biological studies have started to produce large volumes of complex biological shape data from structural analysis (Chandonia and Brenner, 2006; Oueslati *et al.*, 2015) or image analysis (D'Ambrosio and Vale, 2010; Saito *et al.*, 2004; Sumiya *et al.*, 2011). The biological shape data produced often have high dimensions, which pose a significant challenge for their analysis and understanding.

Dimension reduction is an essential tool for analyzing and understanding high-dimensional data. A wide range of related techniques have been developed (Fodor, 2002). However, for effective dimension reduction of biological shape representations, it is crucial to take into account their specific structures and properties. Specifically, biological shapes are often represented by points on high-dimensional Riemannian spaces (Dryden and Mardia, 1998; Kent, 1994). Differences between distinct shapes are best represented by their Riemannian distances rather than Euclidean distances, which are commonly used in dimension reduction. This can be seen from the example in Figure 1, which shows that Riemannian distances better differentiate between different shapes. Indeed, non-linear Riemannian geometry of shape spaces is proposed as a tool of choice for characterizing geometric differences between shapes (Kendall *et al.*, 1999).

In this study, we described 2D shapes using their landmark representation (Dryden and Mardia, 1998; Kent, 1994) and 3D shapes that

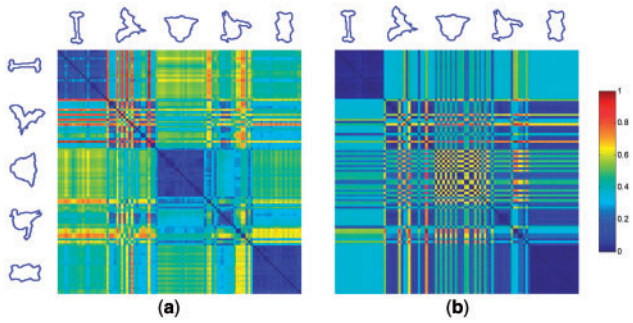

Fig. 1. Riemannian distances versus Euclidean distances in characterizing differences between shapes. Five groups of shapes, with 20 shapes in each group, were selected from the MPEG-7 dataset (Bober, 2001). A representative of each shape group is shown on the top and to the left of the distance map. (**a**) Riemannian distances between different selected shapes; (**b**) Euclidean distances between different selected shapes, calculated using elliptic Fourier descriptors (Kuhl and Giardina, 1982). Each distance is normalized and color coded

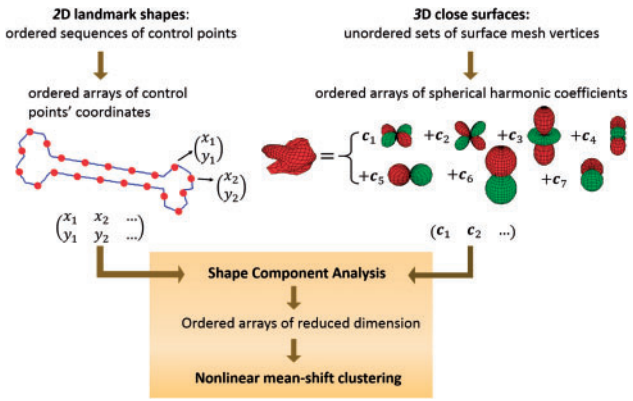

Fig. 2. Overall work flow of the proposed dimension reduction method, with application in mean-shift clustering. Both 2D and 3D shapes are represented as ordered arrays of coefficients to be analyzed by SCA for dimension reduction, followed by nonlinear mean-shift clustering

lack well-defined landmarks using their spherical harmonic representation (SHR) (Kazhdan *et al.*, 2003). We developed a technique for nonlinear dimension reduction of these two shape representations on their Riemannian spaces. A key feature of our technique is that the non-linear distances between shapes are preserved in an embedded low-dimensional shape space. We demonstrate an application of our dimension reduction approach by combining it with non-linear mean-shift clustering on Riemannian spaces (Subbarao and Meer, 2009) for unsupervised clustering of shapes of mitochondria and proteins. Experimental results confirmed that the proposed dimension-reduction technique, when combined with mean-shift clustering, provided equivalent clustering performance but substantially reduced processing time, because the cost of computing the Riemannian distance between two shapes depends linearly on their dimension. The proposed dimension-reduction approach is general and can also be combined with other shape analysis techniques.

In the remaining part of the article, we first outline the theory of shape space and then describe our dimension reduction technique and its integration with unsupervised mean-shift clustering on shape spaces. The overall shape analysis work flow is summarized in Figure 2. We present experimental results on a variety of shape datasets, first on a generic 2D shape dataset, then on a 2D mitochondrial shape dataset and lastly on a 3D protein shape dataset.

2 Methods

2.1 Landmark representation of 2D biological shapes

In this study, we describe 2D biological shapes using their landmark representation (Dryden and Mardia, 1998). Landmark representation can also be used to describe shapes of 3D or higher physical dimensions as long as landmark points are specified.

A shape of physical dimension N is usually represented by a sequence of its D landmark points. In the case of a planar object, its 2D shape can be represented by landmarks along its contour. This ordered sequence of landmark points is referred to as a *configuration*. A configuration can be mathematically represented by a N-by-D real matrix $\mathbf{Y}$, where $\mathbf{Y}_{k,j}$ records kth coordinate of the jth landmark point (Dryden and Mardia, 1998). The *pre-shape* $\mathbf{Z}$ of a configuration $\mathbf{Y}$ has the location and scale information removed. Typically, it is determined by setting $\mathbf{Y}$ to be zero-mean in centroid and unit length in size. The space of pre-shapes can be defined as

$$\{\mathbf{Z} \in \mathbb{R}^{N\times D} : \sum_{j=1}^{D} \mathbf{Z}_{k,j} = 0 \forall k, \quad ||\mathbf{Z}||_{\mathrm{F}}^2 = 1\},$$

where the Frobenius norm $||(\cdot)||_{\mathrm{F}}$ is the square root of the sum of the absolute squares of elements in a matrix. This space is a high-dimensional hyper-sphere in the Euclidean space. While a pre-shape is invariant under the translation and scaling of the original configuration, a shape is invariant under translation, scaling and rotation. The N-dimensional shape space Σ_N^D is defined as the collection of equivalent configurations under translation, scaling and rotation.

2.2 Riemannian geometry of the shape space

The Riemannian geometry of the N-dimensional shape space can be explicitly formulated. Here, we list several related properties that will be used in the remaining part of the article. For ease of implementation, we represent the shape space as a submanifold embedded in $\mathbb{R}^{N\times D}$ with a lifted tangent space.

To define the Riemannian geometry of the shape space, we need to specify: (1) a mathematically defined set containing all possible shapes and (2) the pairwise distance between two shapes. In the theory of landmark representation, a shape is defined as the equivalent class of a pre-shape $\mathbf{Z}$ up to rotations (Dryden and Mardia, 1998). This equivalent class can be represented by

$$[\mathbf{Z}] = \{\mathbf{OZ} : \mathbf{O} \in SO(N)\}, \tag{1}$$

where $SO(N)$ is the group of N-dimensional rotation matrices. The N-dimensional shape space Σ_N^D is a Riemannian manifold containing all possible shapes $[\mathbf{Z}]$ and equipped with a distance $\rho(\cdot,\cdot)$. The Riemannian distance is given by

$$\rho([\mathbf{Z}_0],[\mathbf{Z}_1]) = \cos^{-1}(\max_{\mathbf{O}\in SO(N)} \langle \mathbf{OZ}_0, \mathbf{Z}_1\rangle). \tag{2}$$

Note that this distance is equivalent to the L_2 distance ρ_F after two shapes are rotationally aligned, that is

$$\rho_{\mathrm{F}}([\mathbf{Z}_0],[\mathbf{Z}_1]) = \min_{\mathbf{O}\in SO(N)} ||\mathbf{OZ}_0 - \mathbf{Z}_1||_{\mathrm{F}}. \tag{3}$$

From the computational perspective, the structure of the tangent space is crucial for developing efficient algorithms. Here, we introduce three

concepts: the tangent space of a shape space, the Exponential map and the Log map. Roughly speaking, the tangent space contains information of a function's gradient on a Riemannian manifold, while the Exponential and Log maps are used to exploit the direction of a gradient for searching extrema of the function. The lifted tangent space at $[\mathbf{Z}]$ is represented by

$$H_{\mathbf{Z}}(\Sigma_N^D) = \{\mathbf{X} \in \mathbb{R}^{N\times D} | \langle \mathbf{Z}, \mathbf{X}\rangle = 0, \mathbf{X}\mathbf{Z}^{\mathrm{T}} = \mathbf{Z}\mathbf{X}^{\mathrm{T}}\}. \quad (4)$$

Given an equivalent class $[\mathbf{Z}]$ and a tangent vector $\mathbf{v} \in H_{\mathbf{Z}}(\Sigma_N^D)$, the exponential map on the shape space is given by

$$\mathbf{Exp}_{[\mathbf{Z}]}(\mathbf{v}) = \left[\mathbf{Z}\cos\left(\|\mathbf{v}\|_F\right) + \frac{\mathbf{v}}{\|\mathbf{v}\|_{\mathrm{F}}}\sin(\|\mathbf{v}\|_{\mathrm{F}})\right]. \quad (5)$$

Conversely, when two pre-shapes $\mathbf{Z}_0$ and $\mathbf{Z}_1$ and a rotation $\mathbf{O} \in SO(N)$ satisfy $\mathbf{O}\mathbf{Z}_1\mathbf{Z}_0^T$ being symmetric and positive definite, we obtain the (horizontally lifted) Log map, which is

$$\mathbf{Log}_{\mathbf{Z}_0}(\mathbf{Z}_1) = \frac{s_0}{\sin(s_0)}(\mathbf{O}^T\mathbf{Z}_1 - \mathbf{Z}_0\cos(s_0)), \quad (6)$$

where $s_0 = \rho(\mathbf{Z}_0, \mathbf{Z}_1)$.

Lastly, we note that the Riemannian distance $\rho(\cdot,\cdot)$ and the Log map can be computed using polar decomposition (Golub and Van Loan, 1996; Kendall *et al.*, 1999) as follows:

1. Given two pre-shapes $\mathbf{Z}_0$ and $\mathbf{Z}_1$, compute the singular value decomposition of their covariance matrix

   $$\mathbf{Z}_0\mathbf{Z}_1^T = \mathbf{U}\mathbf{S}\mathbf{V}^T.$$

   Here, $\mathbf{S}$ is a diagonal matrix formed by singular values and U and V are two orthonormal matrices.
2. Compute the distance $\rho(\mathbf{Z}_0, \mathbf{Z}_1)$ as the arc-cosine of sum of absolute value of singular values, i.e. $s = \cos^{-1}\sum_i(|\mathbf{S}_{ii}|)$.
3. Compute the Log map $\mathbf{Log}_{\mathbf{Z}_0}(\mathbf{Z}_1)$ as $\frac{s}{\sin(s)}(\mathbf{O}^T\mathbf{Z}_1 - \mathbf{Z}_0\cos(s))$ where $\mathbf{O}^T = \det(\mathbf{V}\mathbf{U}^T)\mathbf{V}\mathbf{U}^T$.

For 2D shapes, explicit formulas for the Riemannian distance and the Log map have been derived by representing a $2 \times N$ matrix as a $1 \times N$ complex vector (Dryden and Mardia, 1998). This allows a fast implementation without calculating the singular value decomposition.

2.3 Dimension reduction on the shape space

In this section, we define a mapping that projects a high-dimensional shape space into another low-dimensional shape space while the pairwise distance is preserved in the embedded low-dimensional shape space, as defined in Equation (7) or Equation (11). This allows us to adapt, for example, the mean-shift algorithm on the shape space with substantially lower dimensionality.

2.3.1 Low-dimensional shape space and projection error

The essence of principal component analysis is to find the low-dimensional subspace that minimizes projection error in the original Euclidean space. Analogous to this concept, we first define embedded low-dimensional shape spaces and their corresponding projection errors. Given a D-by-r matrix R satisfying $\mathbf{R}^T\mathbf{R} = \mathbf{I}_r$, we define the embedded r-dimensional shape space induced by R as

$$\{[\mathbf{M}\mathbf{R}^T] : \mathbf{M} \in \mathbb{R}^{N\times r}, ||\mathbf{M}||_{\mathrm{F}} = 1\} \quad (7)$$

which is the image of a continuous function

$$f_R : \Sigma_N^r \to \Sigma_N^D, f_R([\mathbf{M}]) = [\mathbf{M}\mathbf{R}^T]. \quad (8)$$

For every shape $[\mathbf{Z}] \in \Sigma_N^D$, we propose to use the cosine of the Riemannian distance as the similarity metric to assess the performance of representing $[\mathbf{Z}]$ in the embedded shape space. That is, we define the similarity measurement as

$$\begin{aligned} E([\mathbf{Z}], [\mathbf{R}]) &\equiv \max_{\mathbf{M}\in R^{n\times r}, ||\mathbf{M}||_{\mathrm{F}}=1} \cos(\rho([\mathbf{Z}], [\mathbf{M}\mathbf{R}^T])) \\ &= \max_{\mathbf{M}\in\mathbf{R}^{n\times r}, ||\mathbf{M}||_{\mathrm{F}}=1, O\in SO(N)} \langle \mathbf{O}\mathbf{Z}, \mathbf{M}\mathbf{R}^T\rangle \\ &= ||\mathbf{Z}\mathbf{R}||_{\mathrm{F}}. \end{aligned} \quad (9)$$

This similarity measurement is continuous in $[\mathbf{Z}]$ and invariant under left multiplication of $SO(N)$ and hence its well-definiteness is guaranteed by the universal property of the quotient (Hungerford, 1980). Note that this similarity measure also reflects the extent of information of a shape data point that is preserved under the projection. And the set of zero similarity scores is an analogy to the set that is perpendicular to a Euclidean subspace. Thus, an embedded submanifold's accuracy to represent a given shape dataset can be easily checked by the similarity scores of individual shape data points.

2.3.2 Structure-preserving dimension reduction

Next we describe the desired mapping. To provide a strong link to the implementation, we consider the shape space Σ_N^D and its tangent space embedded in $\mathbb{R}^{N\times D}$ as described in Section 2.2. The following statement is the cornerstone of our proposed method.

Theorem 1: Given a D-by-r matrix $\mathbf{R}$ satisfying $\mathbf{R}^T\mathbf{R} = \mathbf{I}_r$, the mapping

$$\begin{aligned} T_r : \quad & \Sigma_N^D \cap \{E(\cdot, \mathbf{R}) > 0\} \mapsto \Sigma_N^r \\ & T_r([\mathbf{Z}]) = [\mathbf{Z}\mathbf{R}/||\mathbf{Z}\mathbf{R}||_{\mathrm{F}}] \end{aligned} \quad (10)$$

is well defined. Furthermore, distance is preserved under $\mathrm{T_r}$ in the following embedded r-dimensional shape space induced by $\mathbf{R}$

$$\{[\mathbf{M}\mathbf{R}^T] : \mathbf{M} \in \mathbb{R}^{N\times r}, ||\mathbf{M}||_{\mathrm{F}} = 1\}. \quad (11)$$

Proof: The well-definiteness is followed by verifying that the mapping is continuous and satisfies the universal property of the quotient. The preserving of pairwise distance is ensured since the inner product is preserved in this subset, i.e.

$$\langle \mathbf{M}_1\mathbf{R}^T, \mathbf{M}_2\mathbf{R}^T\rangle = tr(\mathbf{M}_1\mathbf{R}^T\mathbf{R}\mathbf{M}_2^T) = \langle \mathbf{M}_1, \mathbf{M}_2\rangle. \quad (12)$$

Next we address the question of determining the optimal basis matrix $\mathbf{R}$ from a given dataset.

2.3.3 Determining the optimal basis matrix

Given pre-shapes $\mathbf{z}_i$, we propose to determine the optimal basis matrix $\mathbf{R}$ by solving the following optimization problem:

$$\max_{\mathbf{R}^T\mathbf{R}=\mathbf{I}_r} \sum_i E(\mathbf{z}_i, \mathbf{R})^2 = \max_{\mathbf{R}^T\mathbf{R}=\mathbf{I}_r} \sum_i ||\mathbf{z}_i\mathbf{R}||_{\mathrm{F}}^2. \quad (13)$$

Note that the optimal solution is the first r eigenvectors of the covariance matrix $\sum_i \mathbf{z}_i^T\mathbf{z}_i$. The solution to this optimization formulation is similar to the one of 2D principal component analysis (Yang *et al.*, 2004), though we derived this optimization problem from a different perspective. We henceforth refer to this method as shape component analysis (SCA). The overall procedure is summarized in Algorithm 1.

Algorithm 1. Shape component analysis

1: Given pre-shapes $\mathbf{z}_j, j = 1, \ldots, n$
2: specify the reduced dimension r
3: $\mathbf{C} = \sum_j \mathbf{z}_j^T \mathbf{z}_j$
4: Compute the first r eigenvectors of $\mathbf{C}$
5: List the first r eigenvectors in columns of a matrix $\mathbf{R}$
6: $\hat{\mathbf{z}}_j \leftarrow \mathbf{z}_j\mathbf{R}/||\mathbf{z}_j\mathbf{R}||_{\mathrm{F}}$
7: Output $\hat{\mathbf{z}}_j$ for mean-shift clustering

2.4 SHR of 3D biological shapes

Landmark representation can also be used to describe 3D biological shapes, as long as ordered sequences of landmark points are available. However, well-defined landmarks sometimes are lacking in biological studies. In such cases, 3D biological shapes are often represented by unordered sets of surface mesh vertices. But the dimension-reduction technique developed previously for landmark representation can no longer be directly applied. To solve this problem, we adapted SHR to transform unordered sets of mesh vertices of 3D surfaces into an ordered sequence of coefficients in the frequency domain. This approach allows a systematic treatment of SHR similar to the landmark representation.

Spherical harmonics $\{Y_m^l(\theta,\phi)\}$ are an orthonormal system of complex functions for decomposing square-integrable functions into a series of coefficients indexed by integers l and m. Given a closed surface $\mathbf{x}(\theta,\phi) = (x(\theta,\phi), y(\theta,\phi), z(\theta,\phi))^T$ parameterized in polar coordinates, $\mathbf{x}(\theta,\phi)$ can be represented as the linear combination of spherical harmonics

$$\mathbf{x}(\theta,\phi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} \mathbf{c}_m^l Y_m^l(\theta,\phi), \tag{14}$$

where $\mathbf{c}_m^l = (c_{m,l}^x, c_{m,l}^y, c_{m,l}^z)^T$. Each coefficient is calculated by the integration with a specific spherical harmonic, for example,

$$c_{m,l}^x = \int_{S^2} x(\theta,\phi) Y_m^{l*}(\theta,\phi) d\Omega. \tag{15}$$

In practice, the infinite series is truncated by limiting l to $0 \le l \le L$. Detailed explanation of SHR can be found in Kazhdan *et al.* (2003) and Morris *et al.* (2005) and references therein.

Next we would like to point out that the landmark representation and SHR share the same Riemannian geometry. As mentioned in Section 2.2, a Riemannian geometry is specified by the set containing all elements and the pairwise distance between any two elements. We shall define the SHR 'shapes' and show that they (1) form a structurally equivalent set and (2) have the pairwise distance as landmark shapes do. As in the landmark representation, a truncated series of spherical harmonic coefficients can be arranged in an ordered array of coefficients. Such an ordered array is represented as a 3-by-$2L^2$ matrix whose elements record the real and imaginary parts of spherical harmonic coefficients. To define the SHR 'pre-shape', we propose to

1. Set the coefficient $\mathbf{c}_0^0$ to be 0 to make SHR invariant of translation. This operation will set a closed surface to be centered at the origin since
$$\mathbf{c}_0^0 = \left(\int_{S^2} x(\theta,\phi)d\Omega, \int_{S^2} y(\theta,\phi)d\Omega, \int_{S^2} z(\theta,\phi)d\Omega\right)^T$$
is the centroid of the surface.
2. Normalize $\tilde{\mathbf{C}}$ to unit norm to make SHR invariant of scaling.

We denote this normalized matrix as $\tilde{\mathbf{C}}$ to represent SHR pre-shape. Clearly, SHR pre-shapes reside on a high-dimensional hyper-sphere as the landmark pre-shapes. Furthermore, the SHR shape can be defined as the equivalent set of a SHR pre-shape up to rotations, i.e.

$$[\tilde{\mathbf{C}}] = \{\mathbf{O}\,\tilde{\mathbf{C}}, \mathbf{O} \in SO(3)\}. \tag{16}$$

Note that this set is also the equivalent set of a closed surface since rotating the surface is equivalent to applying a rotation matrix to its coefficients array. This can be seen from

$$\mathbf{O}\mathbf{x}(\theta,\phi) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} (\mathbf{O}\mathbf{c}_m^l) Y_m^l(\theta,\phi). \tag{17}$$

So far we have shown that the space of SHR shapes can be represented as the collection of equivalent sets in a hyper-sphere.

The pairwise distance of SHR shapes can be defined as the metric between landmark shapes. The key idea behind the Riemannian distance is to capture the minimum L_2 distance between two rotationally aligned objects. For example, given two closed surfaces, we would compute $\min_{\mathbf{O}\in SO(3)} \int_{S^2} ||\mathbf{O}\mathbf{x}_1 - \mathbf{x}_2||_{\mathrm{F}} d\Omega$ for comparing how different these two surfaces are. Because of the orthonormality of the spherical harmonics, this quantity can be written in terms of SHR coefficient matrices, i.e.

$$\min_{\mathbf{O}\in SO(3)} \int_{S^2} ||\mathbf{O}\mathbf{x}_1 - \mathbf{x}_2||_{\mathrm{F}} d\Omega = \min_{\mathbf{O}\in SO(3)} ||\mathbf{O}\mathbf{C}_1 - \mathbf{C}_2||_{\mathrm{F}}. \tag{18}$$

Note that this is the same distance used in landmark representation in Equation (3). Therefore, SHR shapes and landmark shapes share the same pairwise distance.

Overall, we have shown that SHR shapes and landmark shapes share the same defining set and the same pairwise distance and hence the same Riemannian geometry. Since our dimensional reduction technique in the previous section and the clustering algorithm presented in the following section are designed according to the Riemannian geometry of shape spaces, SHR coefficient arrays can be used as inputs to our algorithm once they are transformed properly. The overall procedure of transforming SHR coefficient arrays to their shape representations is summarized in Algorithm 2.

Algorithm 2. Transformation of an SHR into an SHR shape

1: Given SHR $\mathbf{C}$
2: $\mathbf{C}_0^0 \leftarrow 0$
3: $\tilde{\mathbf{C}} \leftarrow [\mathrm{real}(\mathbf{C})\ \mathrm{imag}(\mathbf{C})]$
4: $\tilde{\mathbf{C}} \leftarrow \tilde{\mathbf{C}}/||\tilde{\mathbf{C}}||_{\mathrm{F}}$
5: Output $\tilde{\mathbf{C}}$ as a pre-shape for subsequent analysis

2.5 Mean-shift clustering on a Riemannian manifold

The mean-shift clustering algorithm is a peak-finding algorithm that searches for representatives supported by the majority of data (Comaniciu and Meer, 2002). A mean-shift clustering algorithm defines modes as local maxima of the underlying probability density distribution estimated by a kernel density estimation and iteratively searches for local maxima using a gradient-based update rule. In Subbarao and Meer (2009), mean-shift clustering is generalized on a Riemannian manifold by replacing the L_2 norm in the kernel

estimation with the Riemannian distance. The update rules are described as follows: given $\mathbf{x}_0$ as a data point, iteratively update

$$\begin{aligned} \mathbf{m}_h(\mathbf{x}_k) &= \frac{\sum_i g(d^2(\mathbf{x}_k, \mathbf{z}_i)/h)\mathbf{Log}_{\mathbf{z}_i}(\mathbf{x}_k)}{\sum_i g(d^2(\mathbf{x}_k, \mathbf{z}_i)/h)}, \\ \mathbf{x}_{k+1} &= \mathbf{Exp}_{\mathbf{x}_k}\left(\mathbf{m}_h(\mathbf{x}_k)\right) \end{aligned} \tag{19}$$

where $\mathbf{z}_i$ are pre-shapes given, h is the bandwidth of the kernel function $g(x) = e^{-x/2}$. Convergence of this algorithm is guaranteed. We implemented mean-shift clustering on the shape space following the procedure shown in Algorithm 3. With the proposed SCA, mean-shift clustering can be performed on the shape space of substantially lower dimensionality without any change in procedure.

Algorithm 3. Mean-shift clustering on the shape space

1: Given pre-shapes $\mathbf{z}_i, i = 1, \ldots, n$
2: **for** $i = 1, \ldots, n$ **do**
3: $\mathbf{x} \leftarrow \mathbf{z}_i$
4: **while** $||\mathbf{m}_h(\mathbf{x})|| < \epsilon$ **do**
5: **for** $j = 1, \ldots, n$ **do**
6: $[U_j, S_j, V_j] = svd(\mathbf{z}_j \mathbf{x}^T)$
7: $O_j = det(V_j U_j^T) V_j U_j^T$
8: $\rho_j = \cos^{-1}(tr(S_j))$
9: $\mathrm{Log}_j = \frac{s_0}{\sin(s_0)}(O_j \mathbf{z}_j - \mathbf{x}\cos(s_0))$
10: **end for**
11: $\mathbf{m}_h(\mathbf{x}) \leftarrow \frac{\sum_i g(\rho_i^2/h)\mathrm{Log}_i}{\sum_i g(\rho_i^2/h)}$
12: $s \leftarrow |\mathbf{m}_h(\mathbf{x})|$
13: $\mathbf{x} \leftarrow \frac{\sin(s)}{s}\mathbf{m}_h(\mathbf{x}) + \cos(s)\mathbf{x}$
14: **end while**
15: Retain $\mathbf{x}$ as a local mode
16: **end for**
17: **Return** distinct local modes

3 Results and discussion

We tested our dimension reduction algorithm on three different shape datasets. All computation was performed on a desktop workstation (2 × Intel Xeon E5503 2.00 GHz and 8GB RAM).

3.1 Application I: 2D generic shapes

3.1.1 Data preparation

We first tested our dimension-reduction algorithm on a generic 2D shape dataset. Specifically, we selected five shape groups from the MPEG-7 dataset (Bober, 2001), with 20 different shapes in each group. The first example of each group is shown in Figure 3. The boundary of each example was fitted by a cubic spline. Then a total of 200 semi-landmark points were sampled from the fitted spline. In this way, each shape was represented by a 2×200 matrix.

3.1.2 Results: geometric approximation

We tested whether our dimension reduction technique could reliably preserve distances between different shapes under reduced dimensions. We first calculated the pairwise Riemannian distances between different shapes and used them as the ground truth. We then calculated Riemannian distance between shapes under different levels of dimension reduction (Fig. 3). We report the approximation accuracy under dimension reduction in root mean square error (RMSE) of the Riemannian distances compared with the ground truth. For comparison, we also calculated the pairwise Euclidean distance on the tangent space of the mean shape. This allowed us to compare our dimension-reduction technique to the principal geodesic analysis (PGA) technique (Fletcher *et al.*, 2004), which is based on the Euclidean distance on the tangent space.

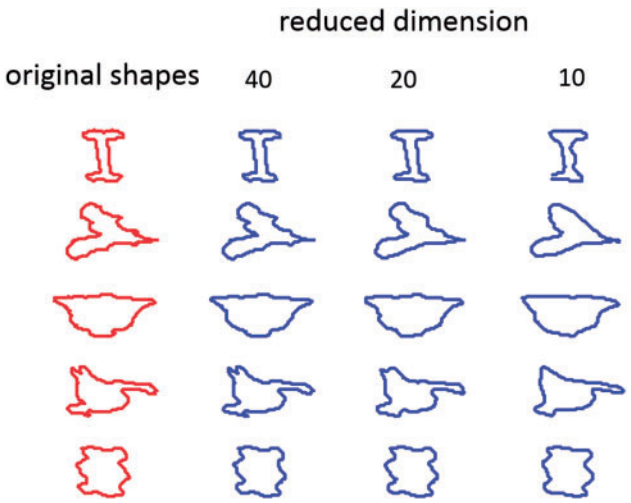

Fig. 3. Reconstructed shapes with different level of dimension reduction using SCA. Examples from different shape groups are shown in rows. Reconstructed shapes with different reduced dimensions are shown in columns

Table 1. Application I: comparison of approximation errors of 2D generic shapes by PGA and SCA

	PGA			SCA		
Dimension	10	20	40	10	20	40
RMSE (%)	5.9	3.0	1.2	11.1	5.3	2.3
Distortion (%)	2.3	2.1	2.2	5.3	1.5	0.4

Furthermore, we calculated the level of distortion under reduced dimension as the normalized error of the distance matrix M compared with the ground truth distance matrix $\mathbf{M}_{\text{truth}}$, i.e.

$$\text{Distortion}(\mathbf{M}, \mathbf{M}_{\text{truth}}) = \frac{||\mathbf{M} - \mathbf{M}_{\text{truth}}||_{\text{F}}}{||\mathbf{M}_{\text{truth}}||_{\text{F}}},$$

where $||\cdot||_{\text{F}}$ is the Frobenius norm. The results were summarized in Table 1. There was a non-vanishing distortion in pairwise distances when PGA and the Euclidean distance were used. In contrast, SCA consistently reduced distortion when the final reduced dimension was increased. In terms of approximation errors, SCA was comparable with PGA under higher final reduced dimensions.

3.1.3 Results: shape clustering

To further assess the performance of our dimension-reduction technique, we combine it with mean-shift clustering and compare it against four other methods. The following is a list of all the methods we tested:

1. Riemannian mean-shift clustering (RMS).
2. RMS with shape component analysis (RMSSCA).
3. Mean-shift clustering on the tangent space, using Euclidean distance (tMS). This method was chosen to compare the difference between using Riemannian and Euclidean distances for shape clustering.

Table 2. Application I: performance of different algorithms on clustering of 2D generic shapes

	2D shapes				
	RMS	RMSSCA	tMS	MSFD	MSLAP
Number of clusters	7	7	7	8	8
Purity	0.8	0.8	0.68	0.64	0.54
NMI	0.73	0.73	0.62	0.53	0.48
AR	0.64	0.64	0.47	0.37	0.26
Run time(s)	2.1	0.6	0.3	0.1	0.06

Twenty shapes from each of the five shape groups were used for testing.

4. Mean-shift clustering with elliptic Fourier descriptors (MSFD) (Kuhl and Giardina, 1982) as inputs. This method was chosen to compare the performance of using rotation-invariant descriptors in shape clustering.
5. Mean-shift clustering with features obtained from Laplacian eigenmap (Belkin and Niyogi, 2003) (MSLAP). This method was chosen to compare the performance of a common strategy of 'flattening the non-linear manifold.' We implemented Laplacian eigenmap with 5-nearest neighbors to project the Riemannian manifold into a 10D Euclidean space, where the mean-shift clustering was performed to find clusters.

For all algorithms, clusters were first identified and then used to classify each data point using one-nearest-neighbor classification. The final clustering results were evaluated by three performance metrics (Vinh *et al.*, 2010), including:

1. Purity $= \frac{1}{N}\sum_k \max_j n_{j,k}$.
2. Normalized mutual Information

$$\mathrm{NMI} = -2\frac{\sum_j\sum_k \frac{n_{j,k}}{N}\log\left(\frac{n_{j,k}}{a_j b_k/N}\right)}{\left(\sum_j \frac{a_j}{N}\log(a_i/N) + \sum_k \frac{b_k}{N}\log(b_k/N)\right)}.$$

3. Adjusted Rand index (AR)

$$\mathrm{AR} = \frac{\sum_{j,k}\binom{n_{j,k}}{2} - \sum_j\binom{a_j}{2}\sum_k\binom{b_k}{2}}{\frac{1}{2}\left(\sum_j\binom{a_j}{2} - \sum_k\binom{b_k}{2}\right) - \left(\sum_j\binom{a_j}{2}\sum_k\binom{b_k}{2}\right)/\binom{n}{2}}.$$

Here, $a_j = \#\{g_j\}$, $b_k = \#\{c_k\}$, $n_{j,k} = \#, \{c_k \cap g_j\}$, c_k and g_j are the kth and jth set of members from clustering and ground truth labeling, respectively, and # denotes the number of elements in a set.

In some cases, these performance metrics could be biased by the number of clusters. To avoid this problem, we adjusted parameters so that each algorithm generated seven to eight clusters. Performance metrics for all the methods tested were summarized in Table 2. The results showed that algorithms using Riemannian distance such as RMS and RMSSCA consistently performed better than algorithms using Euclidean distance such as tMS and MSFD. MSLAP performed the worst among all the algorithms. This confirmed the merit of using non-linear similarity in clustering. Although RMSSCA provided similar clustering performance as RMS, its computational time was significantly reduced, by a factor of ~3.

3.2 Application II: 2D mitochondrial shapes

3.2.1 Experiment design and data preprocessing.

We further tested our SCA algorithm on a mitochondrial shape dataset. Mitochondria within segmental nerves of dissected Drosophila 3rd instar larvae were visualized by fluorescence live imaging. Images were collected on a Nikon Ti-E inverted microscope at 5 frames per second, under a NA of 1.41 and a 100 × magnification. Because of the constraint imposed by the axon geometry, shapes of axonal mitochondria could be represented in 2D instead of 3D. An adaptive active-mask algorithm was used to segment mitochondria (Chen *et al.*, 2012). Individual mitochondria were tracked as in Qiu *et al.* (2012). The boundary of each segmented mitochondrion was fitted by a cubic spline. Then a configuration of 200 semi-landmark points was sampled from the fitted spline.

3.2.2 Results: dimension-reduction and shape clustering

A total of 4000 configurations were collected from 800 mitochondria with five repeats sampled at different time points. Table 3 summarized the performance of SCA and PGA on representing mitochondria morphology. The shape groups identified by the proposed algorithm are shown in Figure 4. The original mean-shift algorithm detected 11 diverse shape clusters. Mean-shift with dimension-reduction provided similar clustering results by identifying eight shape clusters, but the run time was significantly reduced. While the original mean-shift algorithm took 20 756 s to finish the

Table 3. Application II: comparison of approximation errors of 2D mitochondrial shapes by PGA and SCA

	PGA			SCA		
Dimension	8	16	32	8	16	32
RMSE (%)	1.6	0.7	0.2	4.2	1.3	0.3
Distortion (‰)	9.0	2.3	2.4	51	6.4	0.3

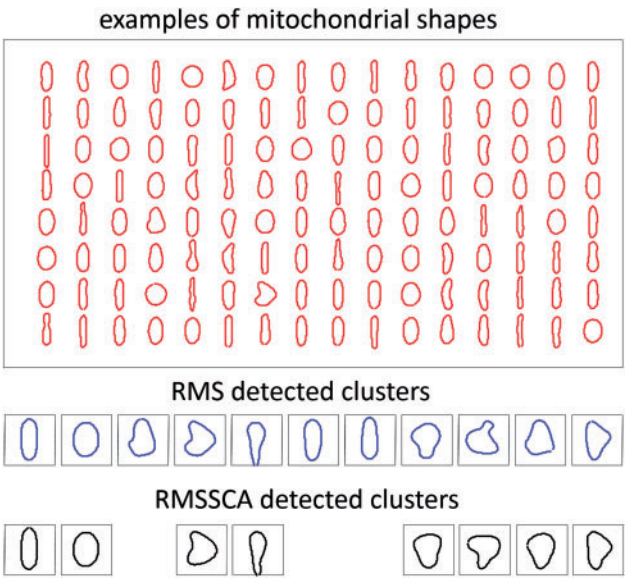

Fig. 4. Clustering of mitochondrial shapes without and with dimension reduction. Red: examples of mitochondrial shapes (selected randomly from 4000 shapes). Blue and black: shape clusters identified by RMS and RMSSCA, respectively

computation, mean-shift with dimension reduction took only 1390 s in total (480 s in dimensional reduction and 910 s in mean-shift clustering, dimension reduced from 2×200 to 2×8), a reduction of computation time by a factor of ~15.

Functions of mitochondria depend critically on their shapes (Campello and Scorrano, 2010). The dimension-reduction technique developed here provides a tool for efficient and multi-resolution representation and analysis of mitochondrial shapes. Combined with unsupervised mean-shift clustering, it supports efficient and unbiased identification of different types of mitochondrial shapes. The identified shape types provide the foundation for further investigation of their cellular functions.

3.3 Application III: 3D protein surfaces

3.3.1 Experiment design

We also tested our algorithm on a 3D protein shape dataset. We selected three series of protein structures from Database of Macromolecular Movements (http://www.molmovdb.org/), including insulin, Ca^{2+} sensor S100 and small G-protein Arf6. These three proteins were selected for their distinct geometry (Fig. 5). Insulin tends to have an ellipsoidal shape, Arf6 is more spherical overall, whereas S100 is approximately *Y*-shaped with a blunt end. Snapshots of the motion sequences of these three proteins were used as examples of each protein class. Overall, 20 examples were collected from each protein class.

3.3.2 Data preprocessing

Surfaces of the selected proteins were generated using Gaussian kernel functions as in Zhang *et al.* (2006) and Liao *et al.* (2013). A multi-level summation of Gaussian kernel functions was employed to generate implicit models from atomic resolution data of the selected proteins. A unique strength of this method is that it allows local resolution control on protein surfaces. Parameters were chosen manually to ensure a genus-zero surface. After triangular meshing of the protein surfaces and output of mesh vertices, spherical harmonics parameterization was computed using SPHARM-MAT software (Brechbühler *et al.*, 1995). We computed spherical harmonics up to order 31 to ensure a 0.2-Å accuracy in approximating original molecule surfaces. In this way, each molecular surface was represented by a matrix of dimension 3×2048.

3.3.3 Results: dimension reduction

First we tested the performance of our algorithm in approximating protein surfaces under reduced dimensions. SCA was performed on the set of 60 protein molecule surfaces and approximations of 10, 20 and 40 dimensions were generated (Fig. 5a). To evaluate the level of surface distortion, we calculated three metrics: surface area distortion, normalized RMSE in spherical harmonic coefficients and normalized RMSE in coordinates of landmark points. Results were summarized in Table 4. In this case, SCA was able to approximate original surfaces with 95% accuracy using 3×40 real matrices (Table 4, rightmost column). Compared with the original 3×2048 coefficient arrays, the dimension of shapes was reduced by a factor of ~51.

3.3.4 Results: shape clustering

Next we tested performance of the proposed algorithm in clustering molecule surfaces. The following is a summary of all the methods we tested:

1. Riemannian mean-shift clustering with dimension reduction (RMSSCA) and without dimension reduction (RMS).
2. Mean-shift clustering with 31 rotation invariant spherical harmonic features (Kazhdan *et al.*, 2003) as inputs (MSSH). From the 31D feature space, principal component analysis was used to select a 6D subspace that account for 99% variance of the data for mean-shift clustering (MSSH + PCA).
3. Mean-shift clustering with features obtained from Laplacian eigenmap (Belkin and Niyogi, 2003) (MSLAP). Laplacian eigenmap was implemented as in the case of 2D shape clustering for

Table 4. Application III: approximation errors of 3D protein surfaces

	SCA		
Reduced dimension	10	20	40
Surface area distortion	10.2 ± 2.44	5.58 ± 1.52	2.43 ± 0.95
nRMSE in SH coefficients	11.6 ± 2.11	7.79 ± 1.40	5.04 ± 0.68
nRMSE in coordinates	11.1 ± 2.09	6.96 ± 1.36	3.98 ± 0.63

Normalized errors in surface area, normalized RMSE in spherical coefficients and normalized RMSE in coordinates of landmark points on the molecule surface are reported. All numbers are reported in percentages.

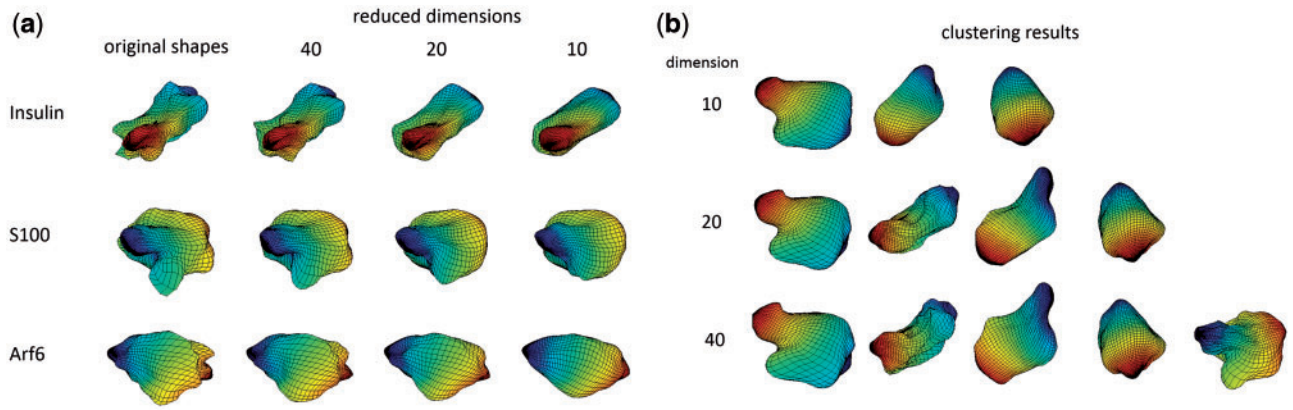

Fig. 5. Dimension reduction and clustering of protein shapes. (**a**) Reconstructed protein surfaces under different reduced dimensions. (**b**) Clusters of protein shapes identified by RMSSCA under different reduced dimensions

Table 5. Application III: performance of different algorithms on 3D shape clustering

	3D shapes				
	RMS	RMSSCA	MSSH	MSSH + PCA	MSLAP
Number of clusters	5	5	5	4	6
Purity	0.73	0.75	0.70	0.70	0.57
NMI	0.489	0.494	0.39	0.39	0.24
AR	0.426	0.421	0.34	0.34	0.06
Run time(s)	33.1	5.6	0.08	0.06	0.13

Twenty examples from each three shape clusters were used for testing in the 3D case.

examining the performance of a common strategy of 'flattening the nonlinear manifold.'

As discussed in Section 3.1, metrics including purity, mutual information and adjusted Rand index were used to evaluate clustering performance. The results were summarized in Table 5. Overall, Riemannian mean-shift clustering provided the best performance. RMSSCA provided similar performance as RMS, but reduced the computation time by a factor of ~5 compared with RMS. Among these algorithms, only RMS and RMSSCA provide backward projection to the original space and thus their clustering results can be visualized. In Figure 5, clustering results via RMSSCA under various reduced dimensions are shown. The three basic structures (ellipsoidal, spherical and Y-shape) were detected under a dimension of 10. As the reduced dimension was increased to 20 and then 40, more details of the protein surfaces were captured by RMSSCA and resulted in more detected shape clusters (Fig. 5b). Overall, the dimension-reduction technique developed here provides a tool for efficient and multi-resolution representation and analysis of 3D proteins shapes. Combined with unsupervised mean-shift clustering, it supports efficient and unbiased classification of proteins shapes and identification of different configuration modes of proteins in their movement.

4 Discussion

In this study, we proposed a technique for dimension-reduction on the Riemannian manifold of 2D and 3D biological shapes. A key advantage of this technique is that it preserves distances between different shapes in an embedded low-dimensional shape space. We showed that although SHR of 3D shapes differs from the landmark representation of 2D shapes, they share the same Riemannian geometry and thus can be processed using the same dimension-reduction technique. We verified the proposed technique on datasets of 2D and 3D shapes. Specifically, we demonstrated an application of the technique by combining it with non-linear mean-shift clustering for unsupervised classification of biological shapes. Our approach is general and provides a tool for analyzing and understanding large sets of high-dimensional shape data. It can also be integrated with shape analysis techniques other than mean-shift clustering.

Our approach is similar to PGA (Fletcher *et al.*, 2004) in that both aim for dimension-reduction of shape representations. PGA relies on principal component analysis on the tangent space of the mean shape. In comparison, our approach maps shapes directly from a high-dimensional shape space to a low-dimensional shape space without making use of the tangent space. In this way, our approach better preserves distances between different shapes by causing less distortion and therefore better retains Riemannian geometry in the dimension-reduced space (Tables 1 and 3). However, our approach also has its limitations. In its current form, it can only handle genus-zero surfaces and does not support local or adaptive approximation of shape features.

Funding

This work was supported by the National Science Foundation [DBI-1052925 and Career Award DBI-1149494 to G.Y., Career Award OCI-1149591 to Y.J.Z.], and by the Department of Defense [PECASE N00014-1401-0234 to Y.J.Z.]. H.-C.L. was supported in part by a Ji-Dian Liang Graduate Research Fellowship and a Bertucci Graduate Research Fellowship.

Conflict of Interest: none declared.

References

Belkin,M. and Niyogi,P. (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, **15**, 1373–1396.

Bober,M. (2001) MPEG-7 visual shape descriptors. *IEEE Trans. Circ. Syst. Video Technol.*, **11**, 716–719.

Brechbühler,C. *et al.* (1995) Parameterization of closed surfaces for 3D shape description. *Comp. Vis. Image Understand.*, **61**, 154–170.

Campello,S. and Scorrano,L. (2010) Mitochondrial shape changes: orchestrating cell pathophysiology. *EMBO Rep.*, **11**, 678–684.

Chandonia,J.-M. and Brenner,S. (2006) The impact of structural genomics: expectations and outcomes. *Science*, **311**, 347–351.

Chen,K.-C. *et al.* (2012) Adaptive active-mask image segmentation for quantitative characterization of mitochondrial morphology. In: *Proceedings of the IEEE International Conference on Image Processing, Orlando, FL*, pp. 2033–2036.

Comaniciu,D. and Meer,P. (2002) Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**, 603–619.

D'Ambrosio,M. and Vale,R. (2010) A whole genome RNAi screen of *Drosophila* s2 cell spreading performed using automated computational image analysis. *J. Cell Biol.*, **191**, 471–478.

Dryden,I. and Mardia,K. (1998). *Statistical Shape Analysis*. John Wiley & Sons, New York.

Fletcher,P. *et al.* (2004) Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Med. Imag.*, **23**, 995–1005.

Fodor,I. (2002) A survey of dimension reduction techniques. *Technical Report UCRL-ID-148494*. Lawrence Livermore National Laboratory.

Golub,G. and Van Loan,C. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD.

Hungerford,T. (1980) *Algebra*. Springer, New York.

Kazhdan,M. *et al.* (2003) Rotation invariant spherical harmonic representation of 3D shape descriptors. In: *Proceedings of the Eurographics Symposium on Geometry Processing, Aachen, Germany,* Vol. 6, pp. 156–165.

Kendall,D. *et al.* (1999) *Shape and Shape Theory*. John Wiley & Sons, New York.

Kent,J. (1994) The complex Bingham distribution and shape analysis. *J. R. Stat. Soc. Ser. B*, 285–299.

Keren,K. *et al.* (2008) Mechanism of shape determination in motile cells. *Nature*, **453**, 475–480.

Kuhl,F. and Giardina,C. (1982) Elliptic Fourier features of a closed contour. *Comp. Graph. Image Proc.*, **18**, 236–258.

Liao,T. *et al.* (2013) Multi-core CPU or GPU-accelerated multiscale modeling for biomolecular complexes. *Mol. Based Math Biol.*, **1**, 164–179.

Morris,R. *et al.* (2005) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics*, **21**, 2347–2355.

Oueslati,E. *et al*. (2015) A new way to visualize DNA's base succession: the *Caenorhabditis elegans* chromosome landscapes. *Med. Biol. Eng. Comput.*, **53**, 1–12.

Qiu,M. *et al*. (2012) Nanometer resolution tracking and modeling of bidirectional axonal cargo transport. In: *Proceedings of the IEEE International Symposium on Biomedical Imaging, Barcelona, Spain*, 992–995.

Saito,T. *et al*. (2004) SCMD: *Saccharomyces cerevisiae* morphological database. *Nucleic Acids Res*, **32**(Suppl 1), D319–D322.

Subbarao,R. and Meer,P. (2009) Nonlinear mean shift over Riemannian manifolds. *Int J. Comp. Vis.*, **84**, 1–20.

Sumiya,E. *et al*. (2011) Cell-morphology profiling of a natural product library identifies bisebromoamide and miuraenamide A as actin filament stabilizers. *ACS Chem. Biol.*, **6**, 425–431.

Vinh,N. *et al*. (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, **11**, 2837–2854.

Yang,J. *et al*. (2004) Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**, 131–137.

Zhang,Y. *et al*. (2006) Quality meshing of implicit solvation models of biomolecular structures. *Comp. Aided Geometr. Des.*, **23**, 510–530.