Cytometry
PART A
Journal of the International Society for Analytical Cytology

Deformation-Based Nuclear Morphometry: Capturing Nuclear Shape Variation in HeLa Cells

Gustavo K. Rohde,[1]* Alexandre J. S. Ribeiro,[1] Kris N. Dahl,[1,2] Robert F. Murphy[1,3,4]

[1]Center for Bioimage Informatics and Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

[2]Bioimage Informatics, Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

[3]Bioimage Informatics, Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

[4]Bioimage Informatics, Department of Machine Learning, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

Received 7 September 2007; Accepted 12 November 2007

Grant sponsor: NSF; Grant number: EF-0331657.

*Correspondence to: Gustavo K. Rohde, Department of Biomedical Engineering, Carnegie Mellon University, HH C 122, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA.

Email: gustavor@cmu.edu

Published online 28 December 2007 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/cyto.a.20506

• Abstract

The empirical characterization of nuclear shape distributions is an important unsolved problem with many applications in biology and medicine. Numerous genetic diseases and cancers have alterations in nuclear morphology, and methods for characterization of morphology could aid in both diagnoses and fundamental understanding of these disorders. Automated approaches have been used to measure features related to the size and shape of the cell nucleus, and statistical analysis of these features has often been performed assuming an underlying Euclidean (linear) vector space. We discuss the difficulties associated with the analysis of nuclear shape in light of the fact that shape spaces are nonlinear, and demonstrate methods for characterizing nuclear shapes and shape distributions based on spatial transformations that map one nucleus to another. By combining large deformation metric mapping with multidimensional scaling we offer a flexible approach for elucidating the intrinsic nonlinear degrees of freedom of a distribution of nuclear shapes. More specifically, we demonstrate approaches for nuclear shape interpolation and computation of mean nuclear shape. We also provide a method for estimating the number of free parameters that contribute to shape as well as an approach for visualizing most representative shape variations within a distribution of nuclei. The proposed methodology can be completely automated, is independent of the dimensionality of the images, and can handle complex shapes. Results obtained by analyzing two sets of images of HeLa cells are shown. In addition to identifying the modes of variation in normal HeLa nuclei, the effects of lamin A/C on nuclear morphology are quantitatively described. © 2007 International Society for Analytical Cytology

• Key terms

nuclear morphometry; shape statistics; shape models; image registration

COMPUTATIONAL analysis of cellular and subcellular structures aims to provide quantitative information (such as the measurement of physical quantities) that can be used to generate and test hypotheses related to normal and pathological eukaryotic cell characterization. Such studies have long been a major topic of biomedical research (see, for example (1,2)) and advances in microscope image acquisition systems and sophisticated image processing algorithms over the past decade have established computational analysis of cell images as a important component of cell biology research (3–6). Amongst many other interesting topics, image-based analysis of nuclear morphometry is a key problem due to the important roles that the cell nucleus plays in biology. Nuclear morphology, and associated changes, have been studied in conjunction with cellular movements (7), cancer (8,9), Hutchinson-Gilford progeria (10), as well as gene expression and protein synthesis (11), to name a few.

Both visual and computational approaches have been applied in characterizing nuclear morphology. For example, nuclear morphology can be visually rated on an objective scale of "normal" and "dysmorphic" (12) but this limits both reproducibility and the number of samples that can be tested. Alternatively, quantitative descriptors of nuclear morphology can be computed from images. Since it is difficult to fully control all physical and biological sources of variation in common experimental setups (e.g., cell cycle phase, focal plane position) most studies are statistical in na-

ISAC
International Society for Analytical Cytology

ture: quantitative nuclear shape and size information is analyzed for significant, broad trends. This information is then analyzed in conjunction with different properties of cells or tissues with the goal of elucidating important relationships and increasing our understanding of fundamental biological concepts. To date, the vast majority of nuclear morphology studies have been based on the extraction of parameters related to shape and size and statistical analysis of their respective means, variation, and covariation (see (2,8,11,13–16) for examples). While such approaches have produced useful results in distinguishing healthy and pathological tissues, as well as providing useful representations of shape distributions, important recent advances in the theory of shape statistics (17,18) could increase the accuracy of the computations.

One of the key concepts arising from such theory is that shape spaces are inherently nonlinear and standard formulae often used for computing sample means, variances, etc. need to be modified to account for the nonlinearities. We use the following example to illustrate this concept. We first construct a distribution of shapes based on a medial axis parametric representation and show that the simple (Euclidean) average of medial axis coordinates does not necessarily represent the correct mean.

Let a represent a real valued random variable uniformly distributed in the closed interval [0,1/2]. A medial axis (a set of 2D coordinates representing a curve on the plane) is constructed based on the random variable a as

$$y(s) = \begin{pmatrix} s \\ \sin(2\pi s) \end{pmatrix}, \qquad (1)$$

with $s \in [0,a]$. The boundary of each object is constructed by traveling a constant distance d in the normal direction from the medial axis $y(s)$. Part A of Figure 1 shows a sampling of shapes created from such a model. Each shape is represented by the medial axis as well as its boundary. Morphometric studies aim to recover information about the shape distribution by extracting and analyzing information from tens, hundreds, or thousands of images containing the shapes of interest. Following this approach, one could be tempted to simply extract the medial axis model by fitting such a model to each shape. Note that for the purposes of this demonstration we do not consider algorithms for extracting medial axis representations, but rather assume these are given. Let $z^k(s)$ represent the medial axis extracted from the kth shape. Assuming that the underlying geometry is a Euclidean vector space, an "average" medial axis is simply given by

$$\bar{y}(s) \sim 1/N \sum_{k=1}^{N} z^k(s)$$

where N is the number of figures or shapes available. The Euclidean average of the medial axis distribution defined in Eq. (1) is shown in Figure 1. For comparison purposes, the known mean shape (defined by the medial axis representation in Eq. (1), with $s \in [0,E\{a\}]$) is also shown in Figure 1, part B. It is clear that the average shape computed by assuming an Euclidean vector space as the underlying geometry is incorrect; in fact, it produces a shape which cannot be represented using the model defined in (1). In portions where the medial axis is approximately linear, both the Euclidean and the correct mean are close to each other. In parts where the medial axis does not closely approximate a straight line, however, the Euclidean average can produce large errors. This is due to the fact that medial axis parameters are not elements of an Euclidean vector space and therefore standard formulae for computing means, variances, covariances, etc., do not apply (19). In fact it can be shown that the elements of medial axis representations belong to a nonlinear manifold (the Riemannian symmetric space) and standard statistical analysis methods such as principal component analysis (PCA) have to be modified to account for an appropriate notion of distance within the manifold (19). A more detailed explanation of this particular example is provided in the appendix together with the description of an alternative method soon to be described.

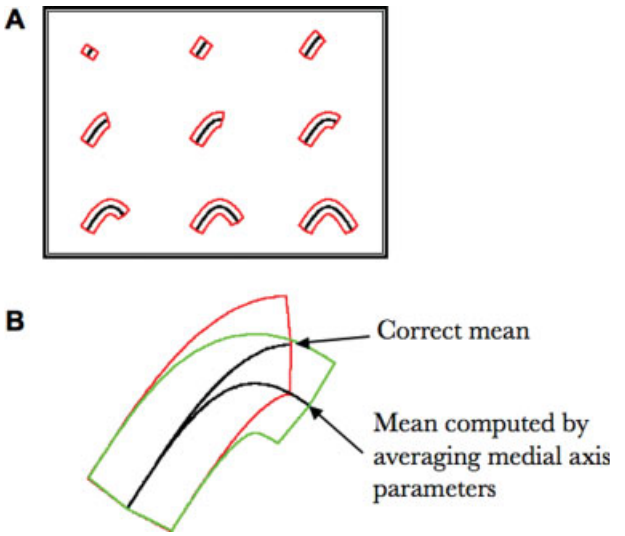

Figure 1. Simulated nuclear shapes demonstrating the nonlinearity of shape distributions. A medial axis-based shape distribution is shown on top (see text for details). The mean shape computed through Euclidean averaging the medial axis parameters is shown in conjunction with the known mean shape. Where the medial axis is approximately linear, the Euclidean average approximates the correct known mean shape well. Where the medial axis is not linear, however, significant errors can arise. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

The field of statistical shape analysis (17,18) has long provided important tools for medicine and biology. We mention briefly a few of the major research directions in the area and their potential applications to nuclear morphometry. The landmark-based work pioneered by Kendall (20) and Bookstein (21) can yield valuable results, but it is not directly applicable to nuclei because corresponding landmarks between different nuclei are difficult to ascertain (although recent advances may circumvent such difficulties (22)). Shape analysis methods via medial axis representations (23) is also popular, and recent work by Fletcher et al. (19) has provided a mathematical basis

for performing PCA based on medial axes extracted from image data. Medial axis representations, however, can be cumbersome to extract, especially for complex shapes (shapes with numerous "blobs" may require more than one medial axis) or for three-dimensional shapes.

An attractive alternative for statistical analysis of shapes is provided by the computational anatomy (CA) framework, where the goal is to quantify shape differences by analyzing the spatial transformations that map different elements of a population (24–26). Here the definition of a shape space is linked to the orbit of a template image (that is, the set of images composed through deformations of a template image) under smooth and invertible spatial transformations (diffeomorphisms). The framework can be extended to handle unlabeled landmarks, contours, as well as dense imagery in arbitrary dimensions and is therefore a viable candidate for modeling distributions of nuclear morphology. Here we show how tools derived from the CA framework can be used to characterize important features of nuclear shape. More specifically, using the large deformation metric mapping (LDMM) framework of Miller and coworkers (25,27), combined with multidimensional scaling (MDS) (28), we offer methods for performing interpolation between two nuclear shapes, measuring "geodesic" distances between them, as well as computing the most representative (mean) shape from a distribution of nuclei. Although this methodology has been previously applied to brain imaging studies (see, for example, (29,30)), we believe the work described here is the first to investigate the application of similar methods to nuclear morphology. Diffeomorphic methods have been recently applied to register nuclei in sets of either 2D or 3D images (41), but not as an approach to characterize nuclear shape distributions. Issues particular to nuclear morphology study, such as the lack of a standardized orientation, and initialization, are discussed. In addition, by combining classical MDS with distance measurements originating from the LDMM framework we provide methods for estimating the intrinsic dimension (number of free parameters), as well as methods for visualizing the most significant variations, of a nuclear shape distribution. The combination of the LDMM-MDS frameworks constitutes a novel approach for characterizing the nonlinear properties of biological shape distributions and are in stark contrast to previous methods based on the analysis of deformation models using PCA (see, for example, (31)).

Methodology and Results

Our goal is to measure important aspects of a given distribution of nuclear shapes automatically from a set of N two- or three-dimensional images $I_k(x)$, $k = 1, \ldots, N$, each containing one nucleus from a fixed population, and with x belonging to a fixed domain Ω (the fixed grid of pixels in the images). Following the approach put forth by Grenander and Miller (24,25) we aim to understand shape distribution-related quantities by analyzing the spatial deformations that map one (image) shape to another. More specifically, we study the set of forms generated by diffeomorphisms (smooth invertible mappings) g acting on different morphological exemplars $I_k(g(x))$. Provided we are able to find algorithms for computing meaningful spatial transformations between different anatomical exemplars, this approach avoids needing to compute medial axis representations or other shape parameterizations, which can be difficult to do with complex shapes in three dimensions, for example.

The spatial mapping $g(x)$ between different morphologies is computed via integration of an ordinary differential equation

$$\begin{cases} \frac{dg(x,t)}{dt} = v(g(x,t),t) \\ g(x,0) = x \end{cases} \tag{2}$$

with $t \in [0,T]$, and integrating the velocity field $v(g(x,t),t)$ (computed as described below) over time. Following the LDMM framework of Miller and coworkers (24,27,32) we choose v to satisfy the following minimization problem

$$v = \arg\min_{v(x,t)} \left(\int_0^T \|v(x,t)\|_V^2 dt + \|I_n(x) - I_m(g(x,T))\|_{L^2}^2 \right) \tag{3}$$

where $\|f\|_{L^2} = \sqrt{\int_{x\in\Omega} |f(x)|^2 dx}$ is the standard L^2 norm for square integrable functions on Ω, and $\|f\|_V$ is simply $\|Lf\|_{L^2}$ with L being a differential operator described in detail below. Intuitively, the minimization problem defined earlier can be understood as trying to find the spatial transformation g, as computed through Eqs. (2) and (3) that matches the images I_m and I_n in the sense of least squares, while at the same time minimizing the amount of incremental "effort" (stretching, bending, deformation, etc.) required to do so. As shown by Miller et al. (32), the quantity

$$d(I_m, I_n) = \int_0^T \|v(x,t)\|_V dt \tag{4}$$

defines a true "geodesic" distance (length) on the manifold of diffeomorphisms in that it satisfies all three required properties: it is positive, symmetric, and satisfies the triangle inequality. We note that the minimization problem (3) is computationally demanding. While algorithms for its minimization based on Euler-Lagrange equations exist (see (27), for example) our work below is based on the so-called fast "greedy algorithm" proposed previously (33). In short, if operator L does not differentiate in time, the space-time $\Omega \times T$ domain can be discretized into a sequence of locally optimal velocities v and the final solution is computed by integrating forward the solution. The partial differential equation associated with the locally in-time optimal solutions can be shown to be (24,27):

$$L^{\#} Lv(x,t_k) + b(g(x,t_k)) = 0 \tag{5}$$

with $L^{\#}$ representing the adjoint of L (conjugate transpose for the case when L is a matrix) and $b(g(x,t_k))$ representing the first variation of an image force term $F(I_m, I_n, g(x,t_k)) = \|I_n(x) - I_m(g(x,T))\|_{L^2}^2$:

$$b(g(x,t_k)) = -[I_m(g(x,t_k)) - I_n(x)]\nabla I_m(g(x,t_k)).$$

Thus $v(x,t_k) = -(L^{\#}L)^{-1}b(g(x,t_k))$ represents a vector pointing from the current shape configuration in the direction of the target shape. Following Beg et al. (27) we choose $L = \alpha\Delta^2 + \gamma$ and as in Joshi et al. (30) we use the following formula for updating the solution: $g(x,t_{k+1}) = g(x + \varepsilon v(x,t_k),t_k)$, with ε representing the time step size. In our implementation we have used the following parameter definitions: $\alpha = 0.8$, $\gamma = 0.05$, $\varepsilon = 0.025$ with 100 iterations, yielding $T = 2.5$. The velocity field v is computed by Eq. (5) estimating the inverse of $L^{\#}L$ as in Beg et al. (27). Briefly, the solution of $L^{\#}L\, g = f$ is computed by taking the discrete Fourier transform of g and f, denoted as $\hat{f}$ and $\hat{g}$ respectively. g can be computed by dividing $\hat{f}$ (with $k = (k_1,k_2)$) by $A(k)^2$ where

$$A(k) = \gamma + 2\alpha\sum_{i=1}^{2}\frac{1-\cos(2\pi\Delta x_i k_i)}{\Delta x_i^2}$$

with Δx_i the pixel resolution in dimension i, and then taking the inverse Fourier transforming the result of this division. The image derivatives necessary for computing $b(g(x,t_k))$ were estimated using the centered finite difference formula (34).

In short, the framework above seeks to characterize differences in shape by measuring the minimum amount of incremental "effort" necessary to deform one shape into another. The amount of effort is measured by Eq. (4) and it provides a distance on the manifold generated by the orbit of image data $I_k(x)$, $k = 1, \ldots, N$ with N being the number of training images, under the spatial transformations computed as described earlier.

Image Data Acquisition

In our experiments we use previously acquired images of HeLa cell nuclei (total of 87 cell nuclei), obtained as described (35), as well as HeLa cells expressing lamin modifications. Lamin modifications in HeLa cells were studied by either overexpression of lamin A, mutant lamin A proteins (such as progerin) or, in the case presented here, knockdown of lamin A/C. More specifically, the lamin A/C gene lmna was knocked down in HeLa cells using a pG-SUPER shRNA to lmna and a GFP-reporter (36) transfected into cells using Lipofectamine 2000. Cells were fixed at different time points after transfection, permeabilized, blocked, and labeled with an antibody against lamin A/C (Novocastra, Vector Laboratories, Burlingame, CA) and DNA was labeled with DAPI (Invitrogen, Eugene, OR). Labeling of lamin A/C allowed a measure of heterogeneity of lamin labeling. Cells were imaged at 63× (1.4 NA) on an inverted fluorescence microscope with a CCD camera (Leica, Bannockburn, IL). As a negative control of transfection and exposure to shRNA, knockdown of luciferase (a nonexistent gene in HeLa cells) using the same vector was used. In total, we have used about 120 cell nuclei per time point (three time points as shown below) in our lamin A/C knockdown study.

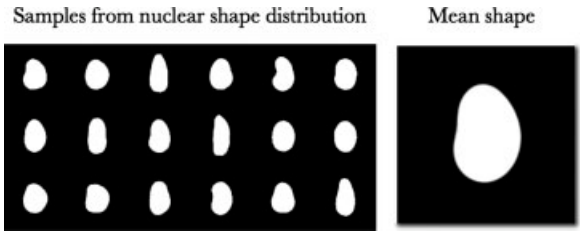

Figure 2. Left panel: sample images of normal HeLa cell nuclei after initialization. Here the images were normalized to account for variations in overall size, orientation, as well as for coordinate reversals (image flips). Right panel: Mean shape computed with LDMM-based algorithm (see text for details).

Preprocessing and Initialization

As our primary concern in this preliminary work is in characterizing the overall shape of cell nuclei, we use a binary version (obtained via the thresholding method described in (37)) of each image in conjunction with the framework described earlier. In the case of lamin knockdown studies we have used the DNA channel for initialization (as well as subsequent morphological analysis). Each set of images was manually inspected and nuclei for which the segmentation process did not work well, due to imprecise boundaries, were discarded.

We note that the deformation-based image matching framework mentioned earlier is not invariant to rigid body transformations. Since the concept of shape is normally understood to be the study of geometric forms modulo variations in position, orientation, and size, we initialize the set of images by minimizing the following functional

$$\Psi(A_1,\ldots,A_N,r_1,\ldots,r_N) = \sum_{m=1}^{N-1}\sum_{n=m+1}^{N}\int_{\Omega}|I_m(A_m x + r_m) - I_n(A_n x + r_n)|^2\,dx \qquad (6)$$

with respect to matrices A_m (each parameterized by rotation and isotropic scaling) and translation vectors r_m. The minimization of such functional is computationally expensive and therefore we resort to using the following approximation. Each binarized image is first scaled so that its foreground (portion that defines the nucleus) has the same area. The translation vectors are computed simply by translating the center of mass of the object to the center of the field of view of each image. All images are then rotated to have the same orientation through a principal axis (Hotteling) transform. Finally, each image is then "flipped" left to right, and up and down, simply by reversing the coordinates of its pixel values, until the functional in (6) is minimized. Figure 2 (left portion) shows a few sample images taken from the normal HeLa cell nuclei population after initialization. We note that although the images used here are binary and two-dimensional, the framework we use can easily accommodate grayscale three-dimensional images (at the cost of an increase in computational complexity).

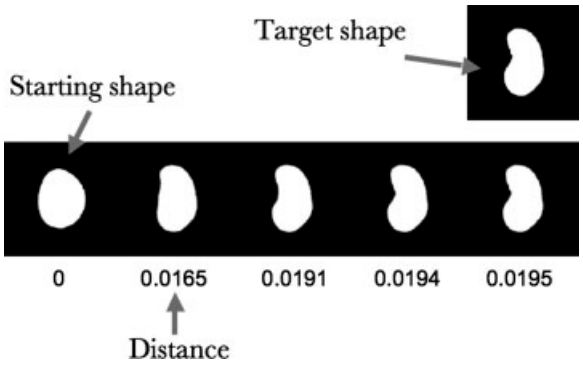

Figure 3. Large deformation metric mapping demonstration. The image on the far left is deformed so as to match the image on the top right while at the same time providing a metric distance between the two shapes.

A Matching Example

The deformation-based image analysis framework described earlier can be more easily understood through visual analysis of an example deformation between two nuclear shapes. An example is provided in Figure 3, where the image of one nucleus is deformed so as to match the image of another. Several intermediary images are shown as the framework seeks to match the image on the far left, to the image on the top right. The example also serves to demonstrate that the framework can be used to perform shape interpolation, while at the same time defining a distance (shown at the bottom of each figure) between shapes.

Mean Nuclear Shape

The ability to compute moments from sample data is fundamental for understanding statistical distributions as a large class of distributions can be expanded as a function of its moments (38). For unimodal distributions, the first moment (mean) is one of the most important parameters describing a sample population. In nuclear morphometry applications, mean shapes can be used as templates, on which comparisons between different populations (healthy or diseased) can be based, or teaching tools (39,40), as well as in generative models that aim to summarize the information contained in a given population of cells (15,16).

As explained earlier, linear averaging cannot be directly applied to image data, nor to parametric descriptions of image data, in an effort to produce a mean shape whenever these do not belong to an Euclidean space. We follow the approach described in Joshi et al. (30) and define the mean shape as the solution to the following minimization problem:

$$\{g_k^*, \bar{I}\} = \arg\min_{S,g_k} \sum_{k=1}^{N} F(I_k, S, g_k) + \int_0^T \|Lv_k(x,t)\|^2 dt \quad (7)$$

with $S(x) = 1/N \sum_{K=1}^{N} I_k(g_k(x,T))$ and subject to

$$g_k(x,T) = \int_0^T v_k(g(x,t),t)dt.$$

A more detailed explanation of the motivation behind the minimization approach defined in (7) is shown in the appendix. We again use the greedy algorithm discussed earlier in our minimization of (7). In essence, the algorithm described earlier aims to estimate a mean shape by finding the set of transformations that align all images in the set, simultaneously, with minimum effort, where the notion of effort is provided by the operator L and the distance function (4). The resulting mean nuclear structure for the normal HeLa cell population described earlier is shown on the right side of Figure 2. As shown here, perhaps contrary to common intuition, the mean shape is not strictly symmetric. A slight concavity on the left of the shape exists while the top of the shape seems to be more pointed than the bottom portion.

The mean shape can be used to establish broad trends and differences between cell populations. Here, we use this concept to examine the effects of lamin knockdown on nuclear morphology. Several studies have shown a dysmorphic shape of nuclei related to loss of lamins (9). However, previous studies were limited in determining the presence or absence of dysmorphic "blebs" in a given population. Using the framework discussed earlier, we are able to obtain more specific quantitative information about the interdependency of nuclear morphology and lamin knockdown concentration by comparing both degree of knockdown (measured by lamin A antibody) and time after expression. Sample images (after initialization) of extracted cell nuclei are shown in Figure 4 and the results of mean shape computations are shown in Figure 5, where synthetic images combining the mean shape and lamin A/C concentration information are plotted on a time axis. In this case the intensity value (in arbitrary units) of each coordinate in each image is directly proportional to the average lamin concentration at that location, within the given time point. The evidence supports that, on average, lower lamin concentration is associated with an increase in bending (concavity) of the cell nucleus. In addition, a slight increase in overall area (in this experiment size was not accounted for in the initialization procedure as it was a parameter of interest) is also associated with a decrease in lamin A/C concentration.

Nonlinear Dimension Reduction

The LDMM framework discussed earlier can be used in conjunction with the classical MDS technique for the purposes

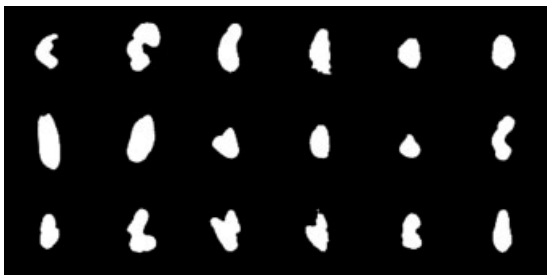

Figure 4. Sample images, after initialization, of nuclear shape morphology in HeLa cells after knockdown of lamin A/C (see text for details).

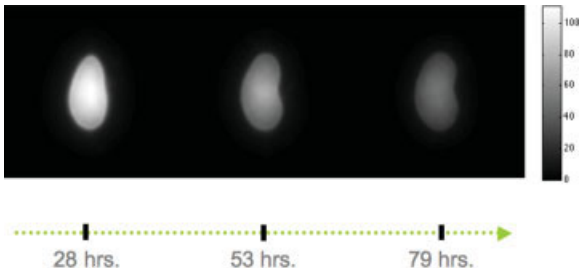

Figure 5. Relationship between nuclear morphology and lamin concentration. Each image in the time line represents the mean shape. The intensity value of each pixel and each image represents the average A/C lamin concentration at that location, in arbitrary units. As time progresses, lamin concentration diminishes and overall bending of the nuclei increase. A slight increase in overall area is also associated with a decrease in lamin concentration. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

of detecting the intrinsic degrees of freedom of a shape distribution dataset. Given a dataset of images $I_k(x)$, $k = 1, \ldots, N$, each containing one nucleus as described earlier (in the following computations we have used the set of images originating from normal HeLa cells), define D to be a matrix of distances between the images in the dataset: i.e. $D_{m,n} = d^2(I_m, I_n)$, with d defined in Eq. (4). Our goal in using MDS is to find a set of coordinates w_k, $k = 1, \ldots, N$, in Euclidean space that best preserves the notion of distance imparted by the LDMM framework described earlier. As described in (28), this task can be achieved by choosing the top eigenvalues and corresponding eigenvectors of $G = -0.5(Id - uu^T)\, D(Id - uu^T)$, with $u^T = 1/\sqrt{N}(1, \ldots, 1)$, and Id representing the identity matrix. Let $\lambda_1, \lambda_2, \ldots, \lambda_N$, represent the sequence of eigenvalues of G, organized in descending magnitude, and with corresponding eigenvectors $b_1, b_2, \ldots, b_N$. Then the ith component of vector w_k is equal to $\sqrt{\lambda_i} b_i^k$. The true dimensionality of the data (the number of free parameters responsible for the variation in shape) can be estimated from the decrease in error (residual variance) between D and $\tilde{D}$, where $\tilde{D}_{m,n} = \|w_m - w_n\|$, as a function of the number of components used in approximating each vector w_k is increased. As in (42), we define the residual variance to be $1 - R^2(\tilde{D}, D)$, with R denoting the standard correlation coefficient between the entries of both matrices.

As shown in Figure 6, the intrinsic dimensionality of the nuclear shape distribution (for the set of normal HeLa cells) seems to be close to three parameters. For comparison, we also plot the residual variance obtained by MDS on a standard Euclidean distance matrix where the distance between two images is simply given by the square root of the sum of the squared differences of their intensity values. In this case the outputs of MDS are equivalent to those of standard PCA (43) on the preprocessed set of nuclear images. Note that in both cases, the error decreases. However, the PCA (E-MDS) error is noticeably higher and it does not imply any particular intrinsic dimension (number of free parameters) for the dataset. This is to be expected since PCA simply finds the linear subspace that minimizes the error between the reconstruction and the data, while the data may lie in a nonlinear submanifold.

In addition to the number of free parameters, the combined LDMM and MDS framework described earlier can be used as a tool for visualizing the most representative modes of shape variation in a distribution of nuclei. Figure 7 contains a two-dimensional representation of the dataset computed through PCA (or MDS on Euclidean distances) and LDMM-MDS, where only the first two components of vectors w_k of each image are plotted in each case. For both point distributions the points labeled with diamonds correspond to the images on the left, stacked vertically. These indicate that the vertical dimension (the second coefficient in both E-MDS and LDMM-MDS) is associated with differences in concavity in the shape distribution. The images corresponding to the points labeled with black squares are shown in a horizontal strip at the bottom of the figure. The variation in these appears to be related to elongation of the nuclei, which in the LDMM-MDS coordinates corresponds to variations in the first coordinate, while in PCA (E-MDS) coordinates corresponds to variations in both the first and second coordinates of w_k. In this case, a low-dimensional representation computed with LDMM-MDS appears to be more effective for elucidating modes of shape variations since it is able to differentiate them into separate coordinates.

Finally, we note that the low-dimensional representation of the image data provided by the combined LDMM-MDS framework discussed earlier can be used to visualize and explore by inspection extreme cases of shapes and images in a distribution. Again we note the difference in extreme shapes provided by E-MDS and LDMM-MDS. While the extreme in

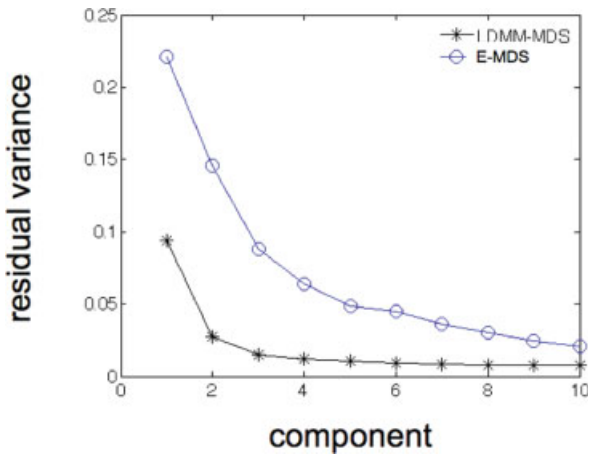

Figure 6. Residual variance (see text for definition) between original and reconstructed images using both Euclidean distance MDS (E-MDS) and LDMM-MDS. Both curves demonstrate that by including more components in the dimension reduction operation, the residual variance can be decreased. However, for any fixed number of components, the residual variance for standard E-MDS is always larger than that of LDMM-MDS. In addition, the LDMM-MDS framework suggests that the number of free parameters responsible for variation in the shape distribution is approximately three, while E-MDS suggests a larger number of free parameters is present in the distribution. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

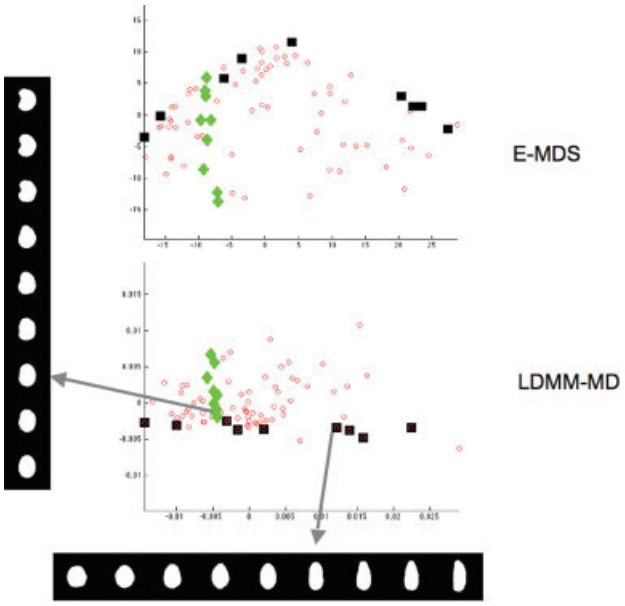

Figure 7. Plot of the first two components of the low-dimensional representation of the image data computed both with Euclidean distance MDS (E-MDS) as well as LDMM-MDS. Each small circle (red online) corresponds to one image in the dataset. Images associated with specific data points are shown on the left (diamonds) or across the bottom (squares). Each dark square corresponds, in order, to each image shown in the horizontal bottom series of images. Likewise, each light triangle corresponds to each image stacked vertically. While LDMM-MDS separates different modes of shape variation (bending and elongation) into separate coordinates, E-MDS seems to mix them. See text for more details. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

the horizontal coordinate in both these methods coincides, this is not true of the second (vertical) dimension: LDMM-MDS identifies a concave shape in its extreme top, while E-MDS identifies a symmetric one.

Discussion and Conclusions

The LDMM framework, together with the shape averaging algorithm described in (30), can be used to compute important features of distributions of nuclear shapes such as means and variances. We note a couple of important considerations to keep in mind when interpreting the results of nuclear shape calculations. First, since nuclei do not have a standardized coordinate system, the position and orientation of the nucleus within the computed mean image is arbitrary and the result shown in Figures 4 and 5 is dependent on the way we have chosen to implement the initialization procedure. Unlike image-based analysis of brain shape, where a standardized coordinate system is easily established through anatomical considerations, the initialization procedure plays a crucial role in studying nuclear shape. The initialization procedure we have chosen aims at removing all differences between the shapes represented in the images by searching for the rigid body, isotropic scaling, and coordinate reversals (image flips) that minimize the sum of squared errors between each shape, thus removing the effect of such transformations on subsequent analysis. We note that these are common considerations in shape analysis methods (17,18).

Although nuclear shapes are usually assumed to have a symmetric shape configuration, the resulting mean shape shown in Figure 2 (right side) is asymmetric. In addition, the means computed from the lamin A/C knockdown studies are also asymmetric, with the level of asymmetry being dependent on overall lamin concentration. We believe this will nearly always be the case for empirical estimates of the mean nuclear shape, as defined by differences in form modulo rigid body transformations and isotropic scaling, due to the initialization step described earlier. Lastly, although we have used the computationally efficient algorithm proposed by Joshi et al. (30) for computing means, we note alternative algorithms are available (see for example (29)), as shape averaging continues to be an active area of biomedical image processing research.

As we have shown earlier, the combination of MDS with distances computed from the LDMM technique can provide a powerful framework for analyzing nuclear morphology. By considering the error between the MDS reconstruction of the geodesic distances produced with LDMM, one is able to estimate the intrinsic dimensionality, or the number of free parameters, that contribute to the shape variations within a dataset. In addition, by mapping each image to a low-dimensional Cartesian coordinate system, one is also able to easily visualize the most significant differences between shapes in a distribution of nuclei. Our analysis of normal HeLa cell nuclei indicates that, although each image contained 196×196 pixel intensity values, approximately three parameters can account for a large amount of shape variations. The first two primary modes of shape variation were determined to be differences in elongation and differences in bending (concavity).

We note that the LDMM-MDS technique is in sharp contrast with previous attempts to characterize morphological variations in biological datasets (see (31) as well as (44–46)), where PCA analysis on the deformation fields that map each exemplar shape, or parameters that describe the shapes, was performed. Although PCA has been successfully used to infer important information related to cell shape (44–46) comparisons using PCA performed directly on preprocessed (initialized as in the description above) is not as informative as LDMM-MDS since, by design, it is optimal for extracting structure from linear subspaces. As we have demonstrated earlier, shape spaces can contain significant nonlinearities. We further clarify that our purpose in using such a framework is not only to overcome the difficulties associated with extraction of parametric descriptions of shape features such as medial axis, but also to introduce methods (through the combination of the LDMM and MDS techniques) for estimating quantities based on the nonlinear geometric space to which the data belong. For example, even though medial axis parameters can be readily extracted from the shapes in Figure 1 (or Fig. 2), their treatment as vectors embedded in a linear vector space is not appropriate (19). The LDMM-MDS technique we described earlier is designed to handle the nonlinearities present in the data.

The reader may note that the LDMM framework above depends on several constants α, γ, and ε and the choice of these is based on prior work (see for example, (27)) as well as experimentation with the data at hand. More specifically, for the application of nuclear shape analysis, sample images are chosen as test images for the algorithms to match. Step sizes ε, and the strength of α, γ are chosen, through trial and error, until the algorithm is able to match these reliably. The choice of these parameters, at this point in our research, is not related to any biophysical properties of HeLa cell nuclei as our goal here is not to infer information about nuclear shape dynamics, but the statistical characterization of shapes in a distribution of nuclei.

Comparisons between different distributions of nuclei are also possible under the framework described earlier. The mean shape was used for establishing relationships between morphology and lamin concentration. As the effects on single cell measurements may not be representative due to uncontrolled sources of variation that influence typical experiments, an average trend can be used to establish the connections between morphology and lamin concentration. Our experiments provide evidence supporting an average increase in bending and concavity associated with an overall decrease in lamin concentration. In the future we plan to study the relationship between different distributions by comparing the number of free parameters associated with each, as well as the most representative modes of shape variation. These could provide important insight into the effects of different phenomena on nuclear morphology.

We also envision methods for performing inference based on the nonlinear low-dimensional representation of the image data described earlier. Modeling of the low-dimensional representation of the shape distribution can be performed by Parzen windowing approximations, for example. Thus the information content can potentially be summarized by a few exemplar images and weighting coefficients that approximate the low-dimensional point distribution obtained by LDMM-MDS (such as the one displayed in Fig. 7). The shape interpolation framework discussed earlier can be used to estimate the image associated with any coordinate in the low-dimensional Cartesian coordinate system, so long as it is a convex combination of points present in the distribution. These techniques can be used for summarizing the important features of nuclear distributions in the context of generative models (15,16).

Finally we note that our use of binarized images is not a requirement and in fact the algorithms described earlier are prepared to handle more complex grayscale images without modification. The segmentation and binarization steps may contain inaccuracies that if accentuated, may propagate through to the image registration process. These may be avoided by choosing not to segment the images prior to registration. In addition, more complex structural information can be obtained by including chromatin information and we plan to investigate including such information in our analysis.

Literature Cited

1. Mendelsohn ML, Kolman WA, Perry B, Prewitt JM. Computer analysis of cell images. Postgrad Med 1965;38:567–573.
2. Prewitt JM, Mendelsohn ML. The analysis of cell images. Ann N Y Acad Sci 1966;128: 1035–1053.
3. Eils R, Athale C. Computational imaging in cell biology. J Cell Biol 2003;161:477–481.
4. Wang YL, Hahn KM, Murphy RF, Horwitz AF. From imaging to understanding: Frontiers in Live Cell Imaging, Bethesda, MD, April 19-21, 2006. J Cell Biol 2006;174:481–484.
5. Chen X, Velliste M, Murphy RF. Automated interpretation of subcellular patterns in fluorescence microscope images for location proteomics. Cytometry A 2006;69A: 631–640.
6. Boland MV, Markey MK, Murphy RF. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry 1998;33:366–375.
7. Wessels D, Voss E, Von Bergen N, Burns R, Stites J, Soll DR. A computer-assisted system for reconstructing and interpreting the dynamic three-dimensional relationships of the outer surface, nucleus and pseudopods of crawling cells. Cell Motil Cytoskeleton 1998;41:225–246.
8. Lesty C, Raphael M, Nonnenmacher L, Binet JL. Two statistical approaches to nuclei shape and size in a morphometric description of lymph node sections in non-Hodgkin's lymphoma. Cytometry 1989;10:28–36.
9. Zink D, Fischer AH, Nickerson JA. Nuclear structure in cancer cells. Nat Rev Cancer 2004;4:677–687.
10. Dahl KN, Scaffidi P, Islam MF, Yodh AG, Wilson KL, Misteli T. Distinct structural and mechanical properties of the nuclear lamina in Hutchinson-Gilford progeria syndrome. Proc Natl Acad Sci USA 2006;103:10271–10276.
11. Thomas CH, Collier JH, Sfeir CS, Healy KE. Engineering gene expression and protein synthesis by modulation of nuclear shape. Proc Natl Acad Sci USA 2002; 99:1972–1977.
12. Goldman RD, Shumaker DK, Erdos MR, Eriksson M, Goldman AE, Gordon LB, Gruenbaum Y, Khuon S, Mendez M, Varga R, Collins FS. Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson-Gilford progeria syndrome. Proc Natl Acad Sci USA 2004;101:8963–8968.
13. Ghani AM, Krause JR. Investigation of cell size and nuclear clefts as prognostic parameters in chronic lymphocytic leukemia. Cancer 1986;58:2233–2238.
14. Gil J, Wu H, Wang BY. Image analysis and morphometry in the diagnosis of breast cancer. Microsc Res Tech 2002;59:109–118.
15. Zhao T, Murphy RF. Automated learning of generative models for subcellular location: Building blocks for systems biology. Cytometry A 2007;71A:978–990.
16. Zhao T, Chen S-C, Murphy RF. Location proteomics. In: Choi S, editor. Introduction to Systems Biology. Totowa, NJ: Humana Press; 2007. pp 196–214.
17. Small C. The Statistical Theory of Shape. New York: Springer; 1996.
18. Dryden IL, Mardia KV. Statistical Shape Analysis. Chichester: Wiley; 1998.
19. Fletcher PT, Lu C, Pizer SM, Joshi S. Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans Med Imaging 2004;23:995–1005.
20. Kendall DG. Shape manifolds, procrustean metrics, and complex projective spaces. Bull Lond Math Soc 1984;16:81–121.
21. Bookstein FL. Size and shape spaces for landmark data in two dimensions. Stat Sci 1986;1:181–242.
22. Chui H, Rangarajan A, Zhang J, Leonard CM. Unsupervised learning of an Atlas from unlabeled point-sets. IEEE Trans Pattern Anal Mach Intell 2004;26:160–172.
23. Blum H. Biological shape and visual science. I. J Theor Biol 1973;38:205–287.
24. Grenander U, Miller MI. Computational anatomy: An emerging discipline. Q Appl Math 1998;56:617–694.
25. Miller MI. Computational anatomy: Shape, growth, and atrophy comparison via diffeomorphisms. Neuroimage 2004;23 (Suppl 1):S19–S33.
26. Thompson D. On Growth and Form. Cambridge: Cambridge University Press; 1917.
27. Beg MF, Miller MI, Trouve A, Younes L. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int J Comput Vis 2005;61:139–157.
28. Cox T, Cox M. Multidimensional Scaling. London: Chapman & Hall; 1994.
29. Avants B, Gee JC. Geodesic estimation for large deformation anatomical shape averaging and interpolation. Neuroimage 2004;23 (Suppl 1):S139–S150.
30. Joshi S, Davis B, Jomier M, Gerig G. Unbiased diffeomorphic atlas construction for computational anatomy. Neuroimage 2004;23 (Suppl 1):S151–S160.
31. Rueckert D, Frangi AF, Schnabel JA. Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration. IEEE Trans Med Imaging 2003;22:1014–1025.
32. Miller MI, Trouve A, Younes L. On the metrics and euler-lagrange equations of computational anatomy. Annu Rev Biomed Eng 2002;4:375–405.
33. Christensen GE, Rabbitt RD, Miller MI. Deformable templates using large deformation kinematics. IEEE Trans Image Process 1996;5:1435–1447.
34. Heath MT. Scientific Computing: An Introductory Survey. New York: McGraw-Hill; 1996.
35. Boland MV, Murphy RF. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics 2001;17:1213–1223.
36. Kojima S, Vignjevic D, Borisy GG. Improved silencing vector co-expressing GFP and small hairpin RNA. Biotechniques 2004;36:74–79.
37. Otsu N. Threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 1979;9:62–66.
38. McCullagh P. Tensor Methods in Statistics. London: Chapman & Hall; 1987.
39. Ko KC, Egbelu PJ. Virtual cell formation. Int J Prod Res 2003;41:2365–2389.

40. Schaff JC, Slepchenko B, Moraru II, Fortin D, Loew LM. The virtual cell project. Mol Biol Cell 2002;13:274A.
41. Yang S, Kohler S, Teller K, Cremer T, Le Baccon P, Heard E, Eils R, Rohr K. Non-rigid registration of 3D multi-channel microscopy images of cell nuclei. LNCS 2006;190: 907–914.
42. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290:2319–2323.
43. Saul LK, Weinberger KQ, Ham JH, Sha F, Lee DD. Spectral methods for dimensionality reduction. In: Chapelle O, Shoelkopf B, Zie A, editors. Semisupervised Learning. Cambridge, MA: MIT Press; 2006.
44. Dunn GA, Brown AF. Alignment of fibroblasts on grooved surfaces described by a simple geometric transformation. J Cell Sci 1986;83:313–340.
45. Lacayo CI, Pincus Z, VanDuijn MM, Wilson CA, Fletcher DA, Gertler FB, Mogilner A, Theriot JA. Emergence of large-scale cell morphology and movement from local actin filament growth dynamics. PLoS Biol 2007;5:e233.
46. Pincus Z, Theriot JA. Comparison of quantitative methods for cell-shape analysis. J Microsc 2007;227 (Part 2):140–156.

Appendix

The goal of this appendix is to facilitate a more intuitive understanding of the concepts and computations described earlier. More specifically, we analyze the example shown in Figure 1 more closely and contrast the linear averaging and LDMM frameworks discussed earlier.

We begin by analyzing the example provided in Figure 1 in more detail, by focusing exclusively on the average of two medial axis representations taken from the shape distribution described in the introductory paragraphs and displayed in Figure 1. The top portion of Figure A1 displays two sample shapes composed of their medial axis and border. Note that the two shapes are superimposed, and the only difference between them is their termination point (the smaller shape appears with a grid texture). An Euclidean (linear) average between these shapes is defined as the mid point of the straight line connecting two corresponding points (see top portion of Fig. A1). Note, however, that the mid point of such straight line is not a point that belongs to the set of shapes in our distribution. The end result of such averaging operation is shown at the bottom of Figure A1. For comparison purposes,

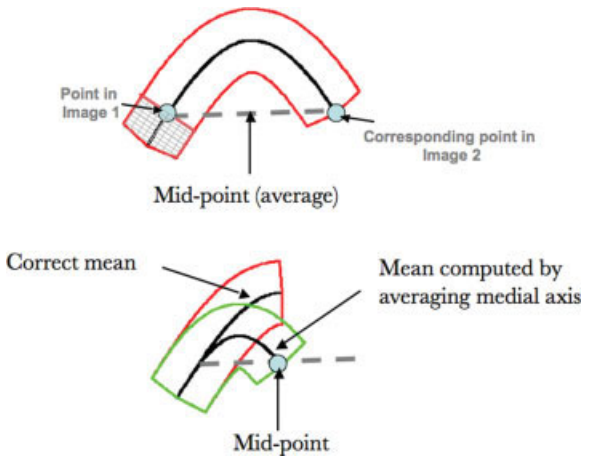

Figure A1. Detailed explanation of medial axis-based linear averaging computation. The mid-point distance between corresponding medial axis coordinates is not related to the nonlinear structure of the shapes. As a consequence, medial axis linear averages do not produce a correct mean estimate. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

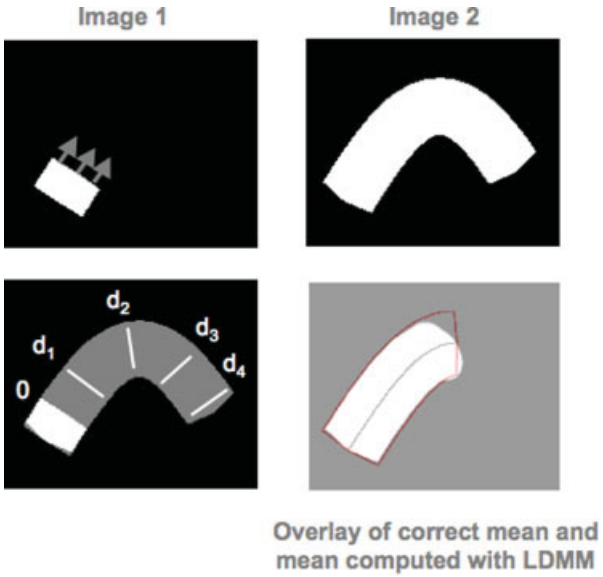

Figure A2. Schematic description of the greedy LDMM and shape averaging algorithms. Given two shapes (top row) LDMM matches image 1 by incrementally deforming it so as to match the two images. The forces guiding the deformation are composed of the difference image, the gradient of image 1, as well a smoothing operation defined by differential operator L described in the text. Gray arrows indicate the direction of the incremental deformation. The distance between the original image and the deformed image can be computed at each iteration. The ''average'' shape is computed by choosing the image associated with half the distance between the two images. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

the correct mean (see introductory paragraphs) is overlaid in this figure as well.

In contrast, let us now consider the nonlinear framework provided by the large deformation metric mapping (LDMM) framework. Binarized versions of the two shapes displayed in Figure A1 are shown in Figure A2 (top row). Image 1 can then be deformed so as to match image 2 according to the LDMM framework discussed earlier. Briefly, the greedy algorithm works by incrementally deforming image 1 with the deformation field $g(x,t_{k+1}) = g(x + \varepsilon v(x,t_k),t_k)$, with ε representing the time step size, and velocity field $v(x,t_k)$ computed according to the differential Eq. (5). The arrows in image 1 represent the direction of deformation in the first iteration. The process is repeated recursively until the two shapes match.

Given a geodesic distance, defined in Eq. (4), the mean of a set of shapes (in this case two shapes represented in the images shown in Fig. A2) can be defined as

$$\bar{I} = \arg\min_{s} \frac{1}{N} \sum_{k=1}^{N=2} d^2(S_k, I) \qquad \text{(A1)}$$

where d is given in Eq. (4). For the case of the two images shown in Figures A1 and A2, this can be accomplished by matching image one to image two via the LDMM framework discussed earlier, yielding the geodesic distance $d(S_1,S_2)$

between the images. The bottom left image in Figure A2 shows the border of the deformed image 1, and assosciated "hypothetical" distances d_1, d_2, d_3, and d_4. The mean image can then be computed by deforming S_1 so that it goes half way the distance $d(S_1,S_2)$. More precisely, the velocity field $v(x,t)$, $t \in [0,T]$, computed from matching images S_1 and S_2 can then be used to generate $\bar{I}(x) = S_1(g(x))$ where $g(x)$ is deformation field for which Eq. (A1) is minimized by integrating

$$g(x) = \int_0^{\xi} v(g(x,t),t)dt$$

and choosing ξ such that

$$\frac{d(S_1,S_2)}{2} = \int_0^{\xi} \|v(x,t)\|_V dt.$$

The result is shown on the bottom right panel of Figure A2. The correct mean is overlaid on the same image for comparison purposes. As shown here, the LDMM framework is capable of generating a better approximation to the mean image by introducing a geodesic distance measure.

We note that the greedy algorithm discussed earlier should be interpreted as approximate solution of Eq. (3) and the geodesic distances obtained from its application may not be optimal (27). Moreover, the image alignment process described earlier is not invariant to an exchange in the order of the images. However, the mean shape computational algorithm described in the methods section earlier, summarized in Eq. (7) and discussed in detail in (30), is symmetric (invariant with respect to an exchange in the order of the images), and is designed to approximate this solution for a large number of images, in a computationally efficient manner.