Testing Equality of Cell Populations Based on Shape and Geodesic Distance

Charles Hagwood*, Javier Bernal, Michael Halter, John Elliott, and Tegan Brennan

***Abstract*—Image cytometry has emerged as a valuable *in vitro* screening tool and advances in automated microscopy have made it possible to readily analyze large cellular populations of image data. The purpose of this paper is to illustrate the viability of using cell shape to test equality of cell populations based on image data. Shape space theory is reviewed, from which differences between shapes can be quantified in terms of geodesic distance. Several multivariate nonparametric statistical hypothesis tests are adapted to test equality of cell populations. It is illustrated that geodesic distance can be a better feature than cell spread area and roundness in distinguishing between cell populations. Tests based on geodesic distance are able to detect natural perturbations of cells, whereas Kolmogorov–Smirnov tests based on area and roundness are not.**

***Index Terms*—Cells, energy, geodesics, hypothesis tests, minimum spanning tree, nearest neighbor, shape space.**

I. Introduction

MANY cell-based studies rely on the determination of an effect to various treatments to assess and to compare experimental results, e.g., in cancer studies, drug screening, toxicology studies, and genomic research. These studies depend on biological, as well as statistical tools to process the data [1]. Two issues involved are 1) determination of a homogeneous population of treatment units, cells, and 2) differentiating between treated cells and untreated cells. The isolation and characterization of homogeneous cell populations are necessary requirements, because a valid comparison requires treatment groups to be made up of homogeneous units, see [2]. Determination of a homogeneous population can be an issue in cell-based studies. For example, it is known that there are potential gains from stem cell treatment. Transplantation of autologous and allogeneic stem cells offer substantial promise for treating a number of diseases, but determination of a homogeneous population of cells is an issue. Many current methods/sorters for differentiation of cells result in possibly mixed cell populations. Statistical hypothesis tests are applied to determine exactly when the results are truly homogeneous and once a homogeneous population of cells has been treated, hypothesis testing is used to determine if an effect exists.

Typically statistical inference regarding a treatment effect is based on extracted attributes, such as cell volume, cell area, cell perimeter, or other attributes. Halter *et al.* [3] in a study used volume to compare populations. Others have used more involved attributes, such as, extracted Fourier frequency components from cell images. Moon and Javidi [4] used SEOL digital holographic microscopy, Gabor wavelet analysis, to extract multiple features from an image. However, most extracted attributes like those described above, do not uniquely identify a cell population. Our goal is to provide statistical tests for issues 1) and 2) based on cell shape, because shape provides a unique characterization.

Shape governs a cell's motility, division, taxonomy and, oftentimes, shape changes are a precursor to disease. What is shape? Shape is generated by the outline or silhouette of an object. Although related, the boundary contour and shape of an object are different. Contours may possess several features, such as being closed, smooth and simple, but shape requires additional attributes, viz, scale invariance, translational invariance, and rotational invariance to describe it. That is, shape is unchanged by a translation, a scale change, or a rotation. Furthermore, the ability to reconstruct, up to these invariants, the object's boundary contour from its shape is an additional requirement we make. This is something that cannot be done for many other shape descriptors, such as perimeter, area, aspect ratio, elongation, eccentricity, etc.

The space of shape representations associated with a space of contours is called its shape space. Shape space is not a linear space, like Euclidean space, for the sum of two closed curves is not closed under addition. Therefore, classical statistical techniques cannot be applied directly to the contours themselves. Srivastava *et al.* [5] developed a theoretical foundation for shape space where calculus and other analysis can be performed on tangent spaces at points in the space. Shape space has the structure of a Riemannian manifold. A manifold is a topological space that locally looks like Euclidean space [6]. A Riemannian manifold is a manifold that assigns a metric to each of the available linear spaces, the tangent spaces at points on the manifold. With this metric on the tangent spaces, a notion of velocity can be defined and thus the length of a curve and the geodesic distance between points in the manifold can be computed. In particular, in shape space, the geodesic distance between elements

Manuscript received May 09, 2013; revised August 01, 2013; accepted August 08, 2013. Date of publication August 26, 2013; date of current version November 25, 2013. *Asterisk indicates corresponding author.*

*C. Hagwood is with the Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (e-mail: hagwood@nist.gov).

J. Bernal is with the Applied and Computational Mathematics Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (e-mail: bernal@nist.gov).

M. Halter and J. Elliott are with the Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (e-mail: michael.halter@nist.gov; john.elliott@nist.gov).

T. Brennan is with the Institute for Applied Computational Science, Harvard University, Cambridge, MA 02138 USA (e-mail: tbrennan01@fas.harvard.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2013.2278467

of the space can be computed. This distance is used as a dissimilarity measure in our statistical tests.

In statistics, determination of equality of populations is usually formulated as a hypothesis testing problem about distribution functions of attributes of the populations, e.g., $H_0 : F(x) = G(x)$ versus $H_1 : F(x) \neq G(x)$ where $F(x)$ is the distribution function for the attributes of the treated population and $G(x)$ the distribution function for the attributes of the untreated population, with data $x \in R^d$. In the univariate case, $d = 1$, there are several nonparametric tests, such as, the Wald–Wolfowitz runs tests, the Kolmogorov–Smirnov test and the Wilcoxon rank sum test. Rank and runs tests are based on ordering the data and it is difficult to generalize these to higher dimensions. When these tests are generalized to dimensions $d > 1$, some form of dissimilarity measure or interpoint distance metric between observations is required. Three multivariate nonparametric tests where shape space methodology can be applied are: 1) the Minimum Spanning Tree test of Friedman and Rafsky [7], 2) Schilling's [8] Nearest Neighbor Distance test, and 3) Szekely and Rizzo's [9] and Aslan and Zech's [10] statistical Energy test. Here, our atttibute is shape. In Section II, the methodolgy of shape space is reviewed and because of the non-Euclidean nature of shape data, a probability space is created on shape space where statistical inference can be performed. In Section III, these three tests are reviewed, it is demonstrated how they apply to cell populations using geodesic distance and then, are compared based on power calculations.

A test procedure is no good if it cannot clearly distinguish between distinctly different shapes and between clearly alike shapes. Our analysis is based on a population of DLEX-p46 cells, a replicate population of DLEX-p46 cells and a population of NIH-3T3 cells. DLEX-p46 cells are round, whereas, NIH-3T3 cells have an elongated, spindly appearance. From these populations, an inhomogenous population of cells can be formed by mixing the two populations and a population more similar to DLEX-p46 or NIH-3T3 in shape can be formed by choosing a shape along the geodesic paths between theses two cell lines, see Figs. 2 and 3. The image preparation process and imaging for these populations are described in Section II of [11]. K-means (with k $= 4$) was used as the segmentation procedure. In Hagwood *et al.* [11], it was shown that k-means is a very competitive algorithm compared to the Otsu, Watershed, and Canny algorithms for the segmentation of fluorescently labeled cells. The cell images evaluated in that study were representative of cell images used for high content screening. In this study, k-means was applied as the segmentation algorithm based on the findings of that previous study and because the cell density and fluorescent labeling protocol used to generate these data sets were similar. After image preparation, processing and segmentation were completed, all cells are represented as their 2-D projected boundaries. In Fig. 1, the boundaries of two segmented cells are illustrated. For each cell, the raw boundary points coming from segmentation are interpolated to produce a smooth boundary curve, as required in the shape space construction. These smooth curves obtained in this manner are the data used to compute shape representations of the cells described in Section II. A MATLAB software package

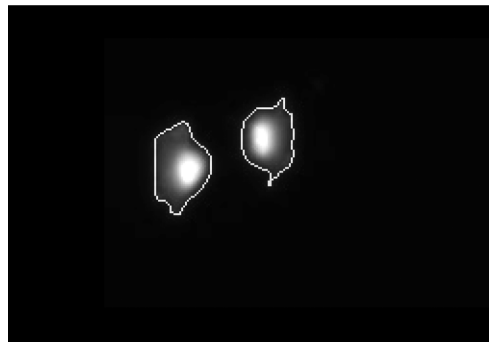

Fig. 1. Boundary curves of two segmented cells.

by Srivastava (http://ssamg.stat.fsu.edu/software) was used to compute geodesic paths and geodesic distances between shapes.

II. Shape Space

Shape-based analysis of objects has a long legacy in statistics [12], computer vision and object recognition [13], biology [14], etc. In these fields, shape similarity measures, $d(A, B)$, between two shapes A and B are usually defined [15] and the tests presented in Section III are useful for these fields. A large part of past analyses was based on using landmarks as summary points on an object, see [16] and [17]. Shape analysis based just on boundary data of a smooth curve or surface was developed by [5], [18], [19], among others. Shape theory based on just bounday data is used here.

Let $\gamma(t) : [0, 2\pi] \rightarrow R^2$ be a parametric representation of a curve, assumed to be smooth and closed, $\gamma(0) = \gamma(2\pi)$. For example, a parametrization of the unit circle is $\gamma(t) = (\sin(t), \cos(t))$. Srivastava *et al.* [5] defined the shape of the curve represented by γ as the result of a function, the shape function, acting on γ. As required, this shape function acting on smooth, simple closed curves is invariant under scaling, translation, rotation and reparametrization. First the shape function scales all curves to have length 2π. This produces scale invariance. Let $\| \ \|$ denote the Euclidean norm on R^2 and $\mathcal{S}^c = \{q \in L^2([0, 2\pi], R^2) : q(t) = \dot{\gamma}(t)/\sqrt{\|\dot{\gamma}(t)\|}, \gamma \text{ a smooth, closed curve of length } 2\pi\}$, the precursor to shape space. Because of differentiation, any translation $\gamma(t) + C$ will correspond to the same representation $q(t)$ in $\mathcal{S}^c$. One may recover $\gamma(t)$ from $q(t)$, since $\gamma(t) = \int_0^t q(s)\|q(s)\|ds$ up to a constant. The required invariance with respect to rotations and reparametrizations is taken care of by identifying their actions on the pre-shape space $\mathcal{S}^c$. The corresponding quotient space $\mathcal{S} = \mathcal{S}^c/(SO2 \times \Gamma)$ is the shape space, where SO2 represents all rotation actions and Γ all reparameterizations. Points in $\mathcal{S}$ are just equivalence classes generated by rotations and reparametrizations that contain curves with the same shape, [20].

The mathematical structure needed to study $\mathcal{S}$ may be obtained from the mathematical structure of $\mathcal{S}^c$. The space $\mathcal{S}^c$ is a subset of the space of square integrable functions on R^2, $L^2([0, 2\pi], R^2)$. Furthermore, $\int_0^{2\pi} \|q(t)\|^2 dt = \int_0^{2\pi} \|\dot{\gamma}(t)\| dt = 2\pi$, therefore points

Fig. 2. Eight points along the geodesic path between a DLEX-p46 cell and a NIH-3T3 cell, geodesic distance between cells is 0.92.

Fig. 3. Eight points along the geodesic path between two DLEX-p46 cells, geodesic distance between cells is 0.62.

in $\mathcal{S}^c$ lie on a sphere in $L^2([0,2\pi],R^2)$. The condition $\gamma(t)$ is closed implies, from the recovery condition $\gamma(t) = \int_0^t q(s)\|q(s)\|ds$, that $\int_0^{2\pi} q(t)\|q(t)\|dt = 0$. Thus, $\mathcal{S}^c = \{q(t)|\int_0^{2\pi}\|q(t)\|^2dt = 2\pi, \int_0^{2\pi} q(t)\|q(t)\|dt = 0\}$. The sphere is a differential manifold and as a level surface of a sphere in $L^2([0,2\pi],R^2)$, $\mathcal{S}^c$ is a differential manifold, Do Carmo [6]. A manifold is a space such that each of its points has a neighborhood about it that is homeomorphic to a ball in some Euclidean space, R^d. When this homeomorphism is differentiable the space is called a differentiable manifold, Do Carmo [6]. Its coordinate mappings provide the means to perform calculus on it. The space $\mathcal{S}^c$ is a Riemannian manifold. A Riemannian manifold is a manifold that assigns a metric to its tangent spaces [21]. With this metric on the tangent spaces, a notion of velocity can be defined and thus the length of a curve can be computed. Paths between points in a Riemannian manifold of minimum length are called geodesics. In the case of $\mathcal{S}^c$, these geodesics are used to determine how close the shapes of curves are to each other. Dissimilarity between cells can then be quantified as the length of the geodesic path connecting them. Figs. 2 and 3 show points on two geodesics.

Under the Riemannian structure that the shape space $\mathcal{S}$ inherits from the pre-shape space $\mathcal{S}^c$, the geodesic distance between two equivalence classes $[q_1]$ and $[q_2]$ in $\mathcal{S}$, $q_1, q_2 \in \mathcal{S}^c$, as given by Srivastava *et al.* [5], is

$$d_s([q_1],[q_2]) = \inf_{O\in SO2,\gamma\in\Gamma} d_c\left(q_1, \sqrt{\dot{\gamma}}O(q_2\circ\gamma)\right)$$

where d_c is the geodesic distance function in the pre-shape space $\mathcal{S}^c$. Srivastava *et al.* [5] provide algorithms for solving the optimization problems associated with computing d_c and d_s.

Finally, in our procedure, the concept of an orientable differentiable manifold is used to define a volume element of $\mathcal{S}^c$. A manifold is orientable if it is not twisted like the Mobius strip [6]. That is, one can consistently choose a "clockwise" orientation for all loops around the manifold. Spheres, planes and tori are orientable. Thus, in such manifolds a measure can be defined, as well as, Borel sets. Because $\mathcal{S}^c$ is a level curve in an orientable manifold it is orientable [22]. Therefore, $\mathcal{S}^c$ can be made into a probability space with measure P, where P is defined by its value on the volume elements of $\mathcal{S}^c$.

III. Homogeneity Tests

The general notion of how to test equality of populations of vector space data has been around for some time. Many such tests rely on using the vector space metric as a dissimilarity measure to distinguish between populations. On the other hand, the notion of homogeneity tests for data in more general spaces is relatively new. Using results from Section II, the framework of these vector space tests can be generalized to test equality of populations in shape space, with geodesic distance as the dissimilarity measure. For two cell populations C and G, let $\{C_1, C_2, \ldots, C_{n_1}\}$ and $\{G_1, G_2, \ldots, G_{n_2}\}$ be samples of boundary contours from each population and let $n = n_1 + n_2$ be the total number of cells. Let $\mathcal{S}$ be shape space and let $\mathcal{S}_C$ be the subspace of shapes associated with C and $\mathcal{S}_G$ the subspace of shapes associated with G. As shown in Section II, a probability space can be created on the shapes $\mathcal{S}_C$ with probability measure P_C and a probability space can be created on the shapes S_G with probability measure P_G. Via a standard construction, see [23], a probability measure can be created on the product of these two spaces, making it possible to define $d(C_i, G_j)$, the geodesic distance from C_i to G_j, as a well defined random variable. With this we can test $H_0 : P_C = P_G$ against the alternative $H_1 : P_C \neq P_G$, that is, we can test if the shapes of these cell populations are statistically identical.

Data

We used as data a sample D_1 of 100 DLEX-p46 cells, a sample D_2 of 100 cells from a replicate population of DLEX-p46 cells and a sample N_1 of 100 NIH-3T3 cells. Fig. 4 shows cells from these populations. We used D_1 and D_2 to test whether two replicate populations of DLEX-p46 cells are indeed replicates, and we used D_1 and N_1 to test equality of DLEX-p46 and NIH-3T3 populations. Then, we formed a mixed sample by taking D_1 and mixing a certain percentage r of N_1 and looked to see how large r has to be before our tests detect the N_1 cells. The best test is that one that detects the NIH-3T3 cells with smallest r. This last test is just a power comparison of the three tests. The data for our tests are six 100 $\times$ 100 geodesic distance matrices, between D_1 and D_1, D_2 and D_2, N_1 and N_1, D_1 and D_2, D_1 and N_1, and D_2 and N_1. We tested

$$H_0 : P_{D_1} = P_{D_2} \quad \text{versus} \quad H_1 : P_{D_1} \neq P_{D_2} \tag{1}$$

$$H_0 : P_{D_1} = P_{N_1} \quad \text{versus} \quad H_1 : P_{D_1} \neq P_{N_1} \tag{2}$$

$$H_0 : P_{D_1} = P_{\text{mixed}} \quad \text{versus} \quad H_1 : P_{D_1} \neq P_{\text{mixed}} \tag{3}$$

P_K is the probability measure associated with population K and P_{mixed} is associated with the mixed population of DLEX-p46 cells and NIH-3T3 cells.

A parametric test of H_0 against H_1, which requires providing distribution functions for P_C and P_G is difficult to formulate at this level of development of shape theory. Multivariate nonparametric tests are available that can be generalized to shape data.

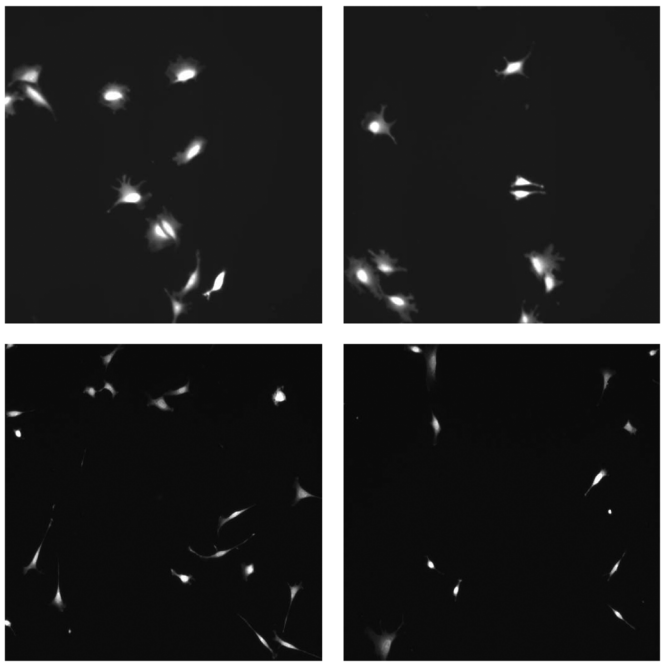

Fig. 4. Top row: Cells from two Replicate populations of DLEX-p46 cells. Bottom Row: Cells from two Replicate populations of NIH-3T3 cells.

Three popular tests for determining whether two populations originate from the same source are the Friedman–Rafsky Minimum Spanning Tree Test [7], the Shilling Nearest Neighbor Test [8], and the Energy Test [9], [10]. These tests have been used frequently, and in this section they have been adapted to test H_0 against H_1.

Friedman–Rafsky test

The Friedman–Rafsky [7] nonparametric test is a multivariate generalization of the Wald–Wolfowitz runs test [24] for testing equality of two univariate populations. In the Wald–Wolfowitz runs test, independent observations on two populations $X_1, X_2, \ldots, X_{n_1}$, and $Y_1, Y_2, \ldots, Y_{n_2}$ are pooled and ordered in ascending order. Runs are defined as sequences of consecutive observations from the same sample in the sequence of ordered pooled observations. A small number of runs indicates the populations are different.

Friedman and Rafsky generalized order to $R^d, d > 1$, by replacing the ordering for univariate samples by an ordering on weighted spanning trees, with each observation in the pooled sample being a tree node. Edges are weighted by the distance between the two nodes forming an edge. Trees are ordered by comparing the sum of their edge weights. The Minimum Spanning Tree (MST) is the tree with minimum total weight and it is used to redefine runs in R^d, as follows. Break the MST at edges where defining nodes are from different populations, count the number of disjoint trees, T_n that results. Small T_n indicates the populations are different. Friedman and Rafsky showed as $n \to \infty$

$$T_n^* = \frac{(T_n - \mu_n)}{\sigma_n} \Rightarrow Z \tag{4}$$

where Z is a standard normal random variable and

$$\mu_n = \frac{2n_1n_2}{n} + 1 \tag{5}$$

$$\sigma_n^2 = \frac{2n_1n_2}{n(n-1)}\left(\frac{2n_1n_2 - n}{n} + \frac{C - n + 2}{(n-2)(n-3)} \times [n(n-1) - 4n_1n_2 + 2]\right) \tag{6}$$

where C equals the number of edge pairs that share a common node, which equals $1/2\sum_{i=1}^{n} d_i(d_i - 1)$, d_i the degree of the ith node. For small values of T_n^*, the null hypothesis is rejected, i.e., H_0 rejected if

$$T_n^* < z_{1-\alpha} \quad \text{(rejection region)} \tag{7}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution. The p-value at the data is

$$\Phi(T_n^*) \quad \text{(p value)}. \tag{8}$$

For our problem, the shapes of the contours, $\{C_1, C_2, \ldots, C_{n_1}, G_1, \ldots, G_{n_2}\}$ are the nodes and the edge weights are the geodesic distances.

Schilling Nearest Neighbor test

The Nearest Neighbor test was developed by Schilling [8], also see [25], as a nonparametric procedure to test equality of multivariate populations. To test $P_C = P_G$ form the pooled sample, $\{C_1, C_2, \ldots, C_{n_1}, G_{n_1+1}, \ldots, G_n\}$ and let

$$Z_i = C_i \quad i = 1, \ldots, n_1 \tag{9}$$

$$= G_i \quad i = n_1 + 1, \ldots, n. \tag{10}$$

Define the rth Nearest Neighbor, Z_{i_r}, of Z_j as rth largest, $Z_i, i \neq j$, in the ordering of $\|Z_1 - Z_j\|, \|Z_2 - Z_j\|, \ldots, \|Z_n - Z_j\|$ from smallest to largest, where in our case $\|X - Y\| = d(X, Y)$, the geodesic distance. Define

$$I_j(r) = 1 \quad \text{if } Z_{i_r} \text{ belongs to the same sample as } Z_j \tag{11}$$

$$= 0 \quad \text{otherwise} \tag{12}$$

and for predetermined k, form the statistics

$$T_{k,n} = \frac{1}{nk}\sum_{j=1}^{n}\sum_{r=1}^{k} I_j(r). \tag{13}$$

$T_{k,n}$ is the proportion of all k Nearest Neighbor comparisons in which a point and its neighbor are members of the same sample. For large values of $T_{k,n} H_0$ is rejected. Schilling showed that as $n_1, n_2 \to \infty$ with n_i/n tending to some constant, λ_i for $i = 1, 2$, then

$$T_{nk}^* = \frac{(nk)^{\frac{1}{2}}(T_{k,n} - \mu_k)}{\sigma_k} \Rightarrow Z \tag{14}$$

where Z is the standard normal distribution and

$$\mu_k = \lambda_1^2 + \lambda_2^2 \tag{15}$$

$$\sigma_k^2 = \lambda_1\lambda_2 + 4\lambda_1^2\lambda_2^2 k\bar{p}_1 - \lambda_1\lambda_2(\lambda_1 - \lambda_2)^2 k(1 - \bar{p}_2). \tag{16}$$

Here, $\bar{p}_i, i = 1, 2$ are constants described in [8]. Schilling's original paper was written for points in $R^d, d = 1, \ldots, \infty$. If

the $n_1 = n_2$, then $\lambda_i = 1/2$. We used $d = \infty$ and for our example $\lambda_1 = \lambda_2 = 0.5$. Schilling showed in this case

$$\sigma_k^2 = \lambda_1\lambda_2 + 4\lambda_1^2\lambda_2^2\left(1 - \binom{2k}{k}2^{-2k}\right). \quad (17)$$

One rejects H_0 in favor of H_1 if

$$T_{k,n} > u_{k,n} + \frac{\sigma_k z_{1-\alpha}}{\sqrt{nk}} \quad \text{(rejection region)} \quad (18)$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution. And, the p-value for this test is

$$1 - \Phi\left(\frac{(nk)^{\frac{1}{2}}(T^o_{k,n} - \mu_k)}{\sigma_k}\right) \quad \text{(p value)} \quad (19)$$

where $T^o_{k,n}$ is the observed Schilling proportion and Φ is the standard normal cumulative distribution function.

Energy test

The energy test was proposed by Zech and Aslan [10], [26]. Also, see Szekely and Rizzo [9]. The concept of statistical energy is based on the principles of potential energy in physics. The potential energy of a continuous charge distribution $\rho(x)$ is

$$\phi = \frac{1}{2}\int\int \frac{\rho(x)\rho(y)}{|x-y|}dxdy. \quad (20)$$

If the charge distribution is split into negative and positive charges, $\rho(x) = \rho_+(x) - \rho_-(x)$, then (20) becomes

$$\phi = \frac{1}{2}\int\int \frac{\rho_+(x)\rho_+(y)}{|x-y|}dxdy - \int\int \frac{\rho_-(x)\rho_+(x)}{|x-y|}dxdy + \frac{1}{2}\int\int \frac{\rho_-(x)\rho_-(y)}{|x-y|}dxdy. \quad (21)$$

From electrostatics the potential energy is minimized when $\rho_+(x) = \rho_-(x)$. If the integrals are discretized and a sample $X : X_1, X_2, \ldots, X_{n_1}$ is considered a sample of generalized positive charges of charge $1/n_1$ and a sample $Y : Y_1, Y_2, \ldots, Y_{n_2}$ as a sample of negative charges of charge $-1/n_2$, then the statistical energy in the pooled sample is

$$T_n = \frac{n_1 n_2}{n_1+n_2}\left(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}R(X_i, Y_j) - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}R(X_i, X_j) - \frac{1}{n_2^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}R(Y_i, Y_j)\right) \quad (22)$$

where $1/|x-y|$ has been generalized to $R(x, y)$, any measure of distance between x and y. We use the geodesic distance between cell shapes x and y as $R(x, y)$.

For large positive values of the test statistic T_n, the equality hypothesis should be rejected. T_n is a degenerate V-statistic and its asymptotic distribution is known not to be normal.

Permutation Distributions

Oftentimes, an asymptotic normal test requires a large sample for the asymptotic distribution to be accurate and hence useful. When the asymptotic distribution is not accurate, a permutation distribution provides an alternative. A permutation distribution is an approximate distribution for the test statistic T_n based on resampling the pooled data $Z_1, \ldots, Z_{n_1}, Z_{n_1+1}, Z_{n_1+n_2}$ from the two populations. The first n_1 Z_i's being the first population. Let $\pi = (\pi_1, \ldots, \pi_{n_1}, \pi_{n_1+1}, \ldots, \pi_{n_1+n_2})$ be a permutation of $\{1, 2, \ldots, n_1+n_2\}$. T_n is applied to each of the $(n_1+n_2)!/n_1!n_2!$ possible two samples, $\{Z_{\pi_1}, Z_{\pi_2}, \ldots, Z_{\pi_{n_1}}\}$, $\{Z_{\pi_{n_1+1}}, \ldots, Z_{\pi_{n_1+n_2}}\}$. In practice, the permutation distribution is approximated by the without replacement bootstrap distribution of T_n, (see [27, p. 207]). Let $t_1^*, \ldots, t_B^*$ be B independent with replacement bootstrap replicates. Histogram these results and use this as the distribution for T_n. The permutation distribution p-value is

$$ALS(T_n) = \begin{cases} \#\{t_k^* \le t_n\} & H_0 \text{ rejected for small } t_n \\ \#\{t_k^* > t_n\} & H_0 \text{ rejected for large } t_n \end{cases} \quad (23)$$

where t_n is T_n evaluated at the original sampled data. One may think of the p-values as the likelihood that a test statistic of this magnitude occurred merely by chance. Small p-values indicate that the data provide evidence that the null hypothesis is false.

IV. Test Results

Test I:

$$H_0 : P_{D_1}(x) = P_{D_2}(x) \quad \text{versus} \quad H_1 : P_{D_1}(x) \neq P_{D_2}(x)$$

Fig. 5 contains the test results of Test I for testing for equality of a population of DLEX-p46 and a replicate population of DLEX-p46 cells. Here, we expect H_0 not be rejected by each test. Permutation histograms for the energy, Nearest Neighbor, and MST test statistics are shown, as well as, the asymptotic normal test statistic distributions for the MST and Nearest Neighbor tests are shown. For the Nearest Neighbor test statistic, we took $\mathrm{k} = 20$. We did not find much difference between $\mathrm{k} = 2$ to $\mathrm{k} = 20$. Fifteen hundred bootstrap samples were used for the energy and Nearest Neighbor tests. Because, the MST algorithm is a very time consuming algorithm, only 500 bootstrap replicates were used to calculate its permutation distribution. With sample sizes $n_1 = n_2 = 100$, the Nearest Neighbor asymptotic normal and permutation histogram densities differ in their tails. Sample sizes of 100 are not large enough for the Nearest Neighbor asymptotic normal distribution to be accurate. The asymptotic MST normal distribution is a good approximation to its permutation distribution. The tests have ALS p-values 0.25, 0.26,0.78. Thus, the hypothesis H_0 cannot be rejected at significance levels of 25% or better in each case.

Test II

$$H_0 : P_{D_1}(x) = P_{N_1}(x) \quad \text{versus} \quad H_1 : P_{D_1}(x) \neq P_{N_1}(x).$$

Fig. 6 shows the permutation distributions for the three test statistics for testing equality of DLEX-p46 and NIH-3T3 cell populations, as well as, the asymptotic normal test statistic distributions for the MST and Nearest Neighbor tests are shown. Here, each test is expected to reject H_0. All the p-values were of the order of 10^{-6} and were set to zero. H_0 is strongly rejected by all tests. Again, in the tails the asymptotic and permutation distributions differ.

Test III:

$$H_0 : P_{D_1}(x) = P_{\text{mixed}}(x) \quad \text{versus} \quad H_1 : P_{D_1}(x) \neq P_{\text{mixed}}(x)$$

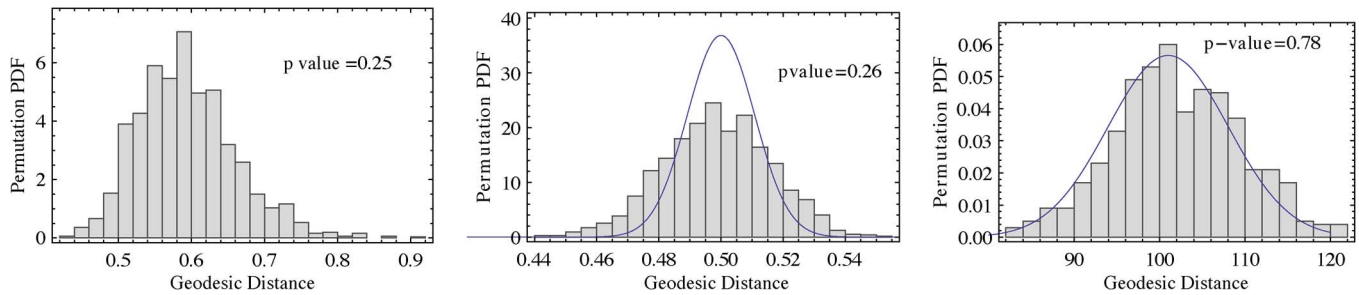

Fig. 5. Test I (DLEX-p46 versus DLEX-p46 Replicate). (a) Energy test statistic. (b) Nearest neighbor test statistic $(\mathrm{k} = 20)$. (c) Minimum spanning tree test statistic.

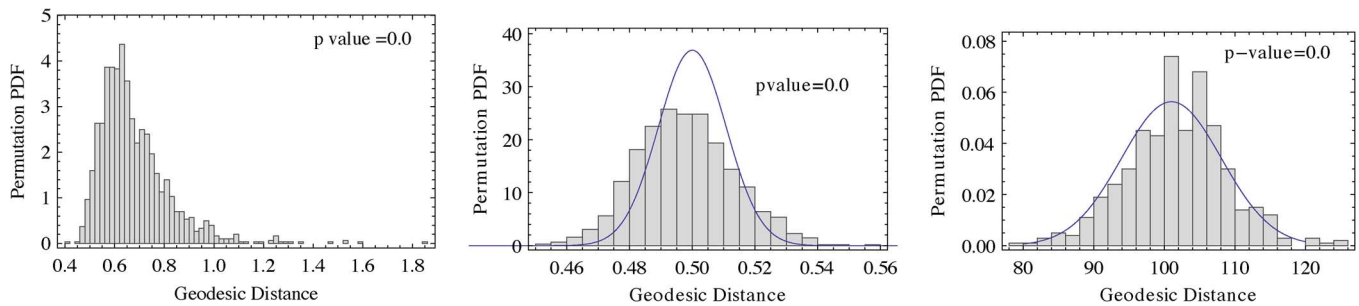

Fig. 6. Test II. (a) Energy test statistic. (b) Nearest neighbor test statistic $(\mathrm{k} = 20)$. (c) Minimum spanning tree test statistic.

In Fig. 7, the powers of the three tests against a specific mixture alternative are shown. From Test I, we know that DLEX-p46 and its replicate are equivalent populations and from Test II we know that DLEX-p46 and the NIH-3T3 populations are completely different. The power alternative we use is to contaminate the replicate DLEX-p46 sample with NIH-3T3 cells. This mixture is compared to the other DLEX-p46 population. We see what mixing proportion is needed for the tests to detect the contamination. Seven populations are created with mixing proportions starting at 10% and going to 28% by 3% increments. Then, seven runs of Test III at each mixture proportion are performed and p-values of the tests are recorded at each run (see Fig. 7). The Energy test has the most power to detect the contamination quickest. It detects the contamination at the 0.05 significance level at about a 14% mixing proportion and it detects the contamination at the 0.03 significance level at a mixture of 16%. The other tests are less powerful and take up to a 25% mixing proportion to detect the contamination at the 5% significance level.

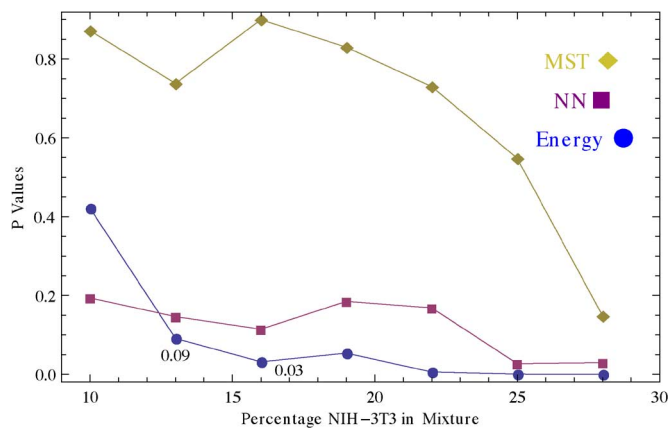

Fig. 7. Power plots: p-values versus contamination proportions for $H_0 : P_{D_1}(x) = P_{\mathrm{mixed}}(x)$ versus $H_1 : P_{D_1}(x) \neq P_{\mathrm{mixed}}(x)$.

V. Other Descriptors

It was stated in the Introduction that typically univariate cell attributes or descriptors have been used to describe a cell population. An often used attribute is cell spread area, [28], [29]. When using cell spread area, the population is identified by its cell areas and equality of populations becomes the two sample problem in statistics. A popular nonparametric statistical test used for the two sample problem is the Kolmogorov–Smirnov statistic. Let $F(x)$ and $G(x)$ denote the distribution functions of the populations of spread areas. To test

$$H_0 : F(x) = G(x) \quad \forall x \geq 0 \quad \text{versus} \quad H_1 : F(x) \neq G(x) \exists x \geq 0$$

the Kolmogorov–Smirnov test rejects H_0 at significance level α if $T = \sup_x |F_n(x) - G_n(x)| > c_\alpha$ where $F_n(x), G_n(x)$ are the empirical distribution functions of $F(x), G(x)$ and c_α is the $1 - \alpha$ quantile for the two sided Kolmogorov–Smirnov test statistic. This test statistic and its quantiles can be found in [30]. Fig. 8 contains a plot of the empirical distribution functions of spread areas for the three populations discussed in the Section IV.

For Test I, with cell area as attribute, the Kolmogorov–Smirnov test resulted in the p-value 0.11, and for Test II the test resulted in the p-value 8×10^{-25}. Thus, cell area distinguishes these populations as well.

The data in Table I show for cell area, the univariate Kolmogorov–Smirnov test is not as powerful as the energy test based on shape. As in Section IV, when a DLEX-p46 population is mixed with NIH-3T3 cells at various proportions and Test III is performed, the Kolmogorov–Smirnov tests detects

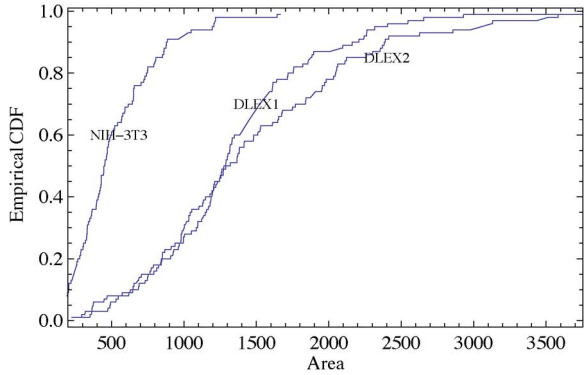

Fig. 8. Empirical cumulative distribution functions for the populations of DLEX1, DLEX2 (DLEX-p46 and its replicate), and NIH-3T3 cell spread areas.

TABLE I
P-VAUES FOR DLEX-P46 VERSUS MIXTURE OF A REPLICATE DLEX-P46 AND NIH-3T3 FOR KOLMOGOROV–SMIRNOV WITH CELL SPREAD AREAS

mixing %	10	13	16	19	22	25	28
p-values	0.23	0.20	0.15	0.1	0.05	0.025	0.01

Fig. 9. Seven points along a geodesic path between a cell from a DLEX-p46 population (first cell on left) and a cell from a NIH-3T3 population (first cell on right).

contamination of the DLEX-p46 population with NIH-3T3 cells after 22% contamination. The energy test detects the contamination at about 14% contamination. Also, we performed the Kolmogorov–Smirnov test with roundness ($4\pi\,\mathrm{area}/\mathrm{perimeter}^2$) as a descriptor. For Test I, with cell roundness as attribute, the Kolmogorov–Smirnov test resulted in the p-value 0.24, and for Test II the test resulted in the p-value 8×10^{-46}. Also, roundness as a descriptor distinguishes equality of these populations.

VI. PERTURBATION OF CELLS

The goal of this section is to compare the previous tests based on shape versus those based on area and roundness when the populations are closer in shape. We do this by comparing a DLEX-p46 population with a perturbation of it. Cells in the DLEX-p46 population are perturbed using the following procedure. We pair cell i from the file of DLEX-p46 cells with cell i from the file of NIH-3T3 cells, $i = 1, \ldots, 100$ (file order). Then, the geodesic path with endpoints from the ith pair is determined. Fig. 9, shows one such path. The second cell from the left on the path is taken as the perturbed DLEX-p46 cell. The total geodesic distance along the path in Fig. 9 is 0.71 and the distance from first cell on left to the second cell is 0.12. This new population contains 100 perturbed cells. The mean path length over all 100 paths is 0.62 and the mean distance over all paths from the second cell to the DLEX-p46 endpoint cell is 0.102. We tested to determine equality of the DLEX-p46 population and its perturbation. The three tests based on shape rejected equality of the populations with p-values smaller than 10^{-6}. Kolmogorov–Smirnov tests based on area and roundness gave p-values 0.23 and 0.07, respectively. Thus, a Kolmogorov–Smirnov test with $\alpha = 0.05$ significance level and

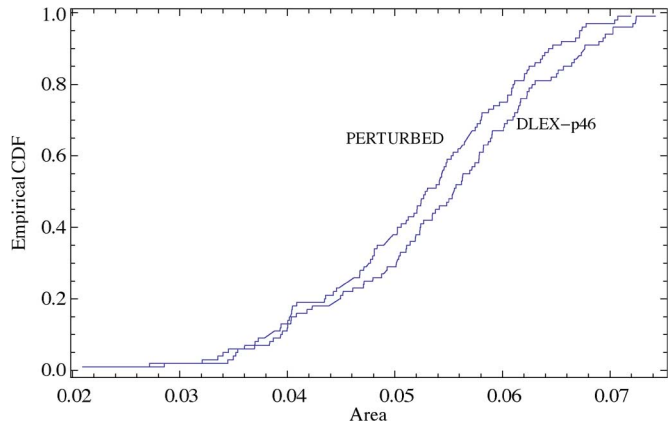

Fig. 10. Empirical cumulative distribution function of a population of DLEX-p46 cell areas and the empirical cumulative distribution function of a population of cell areas for a perturbed population of DLEX-p46 cells.

with either cell spread area as data or cell roundness as data would not reject the hypothesis that the perturbed population of DLEX-p46 differs from the population of DLEX-p46 cells. Fig. 10 contains a plot of the area empirical distribution functions of the two populations (the geodesic algorithm scales all curves to have length one). This example provides evidence that to detect subtle, but significant changes in shape, geodesic distance is a better descriptor than area or roundness.

VII. SUMMARY

We illustrated the viability of using cell shape to test equality of two cell populations. In order to test if two cell populations are statistically identical, at least three steps are required: 1) image processing and segmentation, 2) defining an appropriate cell descriptor and 3) defining a test statistic. This paper concentrated on steps 2) and 3). Shape was emphasized as our descriptor. In addition to the fact that shape serves as an important surrogate for many biological responses, there is a one-to-one (up to invariants) correspondence between a cell's shape and its boundary contour. We presented the methodology of shape space theory that allows one to define a probability measure on shape space and to compute the geodesic distance between cells. With this theory, one can associate with a shape population, C, a probability measure P_C. The problem of equality of two cell populations then can be formulated as the hypothesis testing problem $H_0 : P_C = P_G$ where C and G are shapes of two cell populations. To test H_0, three test statistics, the Energy, the Nearest Neighbor, Minimum Spanning Tree tests from multivariate statistics for data in Euclidean space, were applied to shape data using geodesic distance instead of Euclidean distance.

We tested, using samples of size 100, 1) whether a population of DLEX-p46 cells has the same shape distribution as a population of replicate DLEX-p46 cells and 2) whether a population of DLEX-p46 cells has the same shape distribution as a population NIH-3T3 cells. All three tests showed that the shapes of the DLEX-p46 are statistically equivalent to the shapes of the a replicate population of DLEX-p46 cells. They rejected the hypothesis that the shapes of DLEX-p46 cells are similar to shapes of NIH-3T3 cells. The three tests were compared based

on their power to reject equality of a DLEX-p46 population and a replicate DLEX-p46 population mixed with k% NIH-3T3 cells. The Energy test is the most powerful in detecting the contamination of the replicate DLEX-p46 population with NIH-3T3 cell. Because these three cell populations are either far apart or very close, the same conclusions were reached by the Kolmogorov–Smirnov test applied to cell spread area. When a shape perturbation is performed on the DLEX-p46 cells, the tests based on geodesic distance detected the perturbation, whereas the test based on area failed to do so. Both the Nearest Neighbor test and the MST test are computer time consuming (at least for the MATLAB program we used). The Energy test runs quite fast.

In conclusion, all three tests are viable tests for testing equality of cell populations using shape. The Energy test is the most accurate and requires less computer time than the others.

ACKNOWLEDGMENT

The authors would like to thank A. Srivastava for allowing us to use his MATLAB program for computing geodesics and the reviewers for their valuable comments.

REFERENCES

[1] A. Plant, J. Elliott, A. Tona, D. McDaniel, and K. Langenbach, "Tools for quantitative and validated measurements of cells," *Methods Molecular Biol.*, vol. 356, pp. 95–107, 2006.

[2] G. Juan, E. Hernando, and C. Cordon-Cardo, "Separation of live cells in different phases of the cell cycle for gene expression analysis," *Cytometry*, vol. 49, pp. 170–175, 2002.

[3] M. Halter, J. Elliott, J. Hubbard, A. Tona, and A. Plant, "Cell volume distributions reveal cell growth rates and division times," *J. Theor. Biol.*, vol. 257, pp. 124–130, 2009.

[4] I. Moon and B. Javidi, "Three-dimensional identification of stem cells by computational holographic imaging," *J. R. Soc. Interface*, vol. 4, pp. 305–313, 2007.

[5] A. Srivastava, E. Klassen, S. Joshi, and I. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1415–1428, Jul. 2011.

[6] M. do Carmo, *Differential Geometry of Curves and Surfaces*. Upper Saddle River, NJ: Prentice Hall, 1976.

[7] J. H. Friedman and L. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *Ann. Stat.*, vol. 7, pp. 697–717, 1979.

[8] M. Schilling, "Multivariate two-sample tests based on nearest neighbors," *J. Acoust. Soc. Am.*, vol. 81, pp. 799–806, 1986.

[9] G. Szekely and M. Rizzo, "Testing for equal distributions in high dimension," *InterStat*, vol. 5, Nov. 2004.

[10] G. Zech and B. Aslan, "Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding," *Nucl. Instrum. Methods Phys. Res. A*, vol. 537, pp. 626–636, 2005.

[11] C. Hagwood, J. Bernal, J. Elliott, and M. Halter, "Evaluation of segmentation algorithms on cell populations using CDF curves," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 380–390, Feb. 2012.

[12] F. Bookstein, "The measurement of biological shape and shape change," in *Lecture Notes on Biomathematics*. New York: Springer, 1978, vol. 24.

[13] A. Frome, Y. Singer, F. Sha, and J. Malik, "Learning globallly-consistent local distance function for shape-based image retrieval and classification," in *Proc. IEEE 11th Int.Conf. Comput. Vis.*, Oct. 2007, pp. 1550 –5499.

[14] K. D. G. Rohde, A. Ribeiro, and R. Murphy, "Deformation-based nuclear morphometry: Capturing nuclear shape variation in hela cells," *Cytometry Part A*, vol. 73, pp. 341–350, 2007.

[15] E. Arkin, P. Chew, D. Huttenlocher, K. Kedem, and J. Mitchell, "An efficiently computable metric for comparing polygonal shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 209–216, Mar. 1991.

[16] D. Kendall, "Shape manifolds, procrustean metrics, and complex projective spaces," *Bull. Lond. Math. Soc.*, vol. 16, pp. 81–121, 1984.

[17] I. Dryden and K. Mardia, *Statistical Shape Analysis*. New York: Wiley, 1998.

[18] L. Younes, "Computable elastic distance between shapes," *SIAM J. Appl. Math.*, vol. 58, pp. 565–586, 2008.

[19] E. Sharon and D. Mumford, "2d-shape analysis using conformal mapping," *Int. J. Comput. Vis.*, vol. 70, pp. 55–75, 2006.

[20] I. Herstein, *Topics in Algebra*. Waltham, MA: Blaisdell, 1964.

[21] M. do Carmo, *Riemannian Geometry*. Boston, MA: Birkhauser, 1992.

[22] L. Tu, *An Introduction to Manifolds*. New York: Springer, 2007.

[23] E. Hewitt and K. Stromberg, *Real and abstract analysis*. New York: Springer-Verlag, 1965.

[24] A. Wald and J. Wolfowitz, "Two-sample tests for multivariate distribution," *Ann. Math. Stat.*, vol. 11, pp. 147–162, 1940.

[25] N. Henze, "A multivariate two-sample test based on the number of Nearest Neighbor type coincidences," *Ann. Stat.*, vol. 16, p. 772, 1988.

[26] G. Zech and B. Aslan, "A multivariate two-sample test based on the concept of minimum energy," in *PHYSTAT2003*, Stanford, CA, Sep. 8–11, 2003, pp. 97–100.

[27] B. Efron and R. Tibsirani, *An Introduction to the Bootstrap*. New York: Chapman Hall, 1993.

[28] J. Elliottt, M. Halter, A. Plant, J. Woodward, K. Langenbach, and A. Tona, "Evaluating the performance of fibrillar collagen films formed at polystrene surfaces as cell culture substrates," *Biointerphases*, vol. 3, pp. 19–28, 2008.

[29] K. Bhadriraju, M. Young, S. Ruiz, D. Pirone, J. Tan, and C. Chen, "Activation of rock by rhoa is regulated by cell adhesion, shape, and cytoskeletal tension," *Cell Res.*, vol. 313, pp. 3616–3623, 2007.

[30] W. Conover, *Practical Nonparametric Statistics*. New York: Wiley, 1999.

[31] R. Serfling, *Approximation Theorems of Mathematical Statistics*. New York: Wiley, 1980.

[32] , D. L. Taylor, J. R. Haskins, and K. G. , Eds., *Methods in Molecular Biology*. Totowa, NJ: Humana, 2006, vol. 356.