In: MONDAINI, R. (Org.). Proceedings of the Second Brazilian Symposium on Mathematical and
Computational Biology. Rio de Janeiro, 2003, v. 1, p. 179-207.

Shape spaces, metrics and linear models for evolutionary rates

LEANDRO R. MONTEIRO[1], LUIS H. GUILLERMO[2], LUIS A. RIVERA[2]

[1]Laboratório de Ciências Ambientais - LCA -
[2]Laboratório de Ciências Matemáticas - LCMAT -
Universidade Estadual do Norte Fluminense - UENF -
Av. Alberto Lamego, 2000, CEP 28015-620, Campos dos Goytacazes, RJ, Brasil
lrmont, guillerm, rivera@uenf.br

Abstract. The use of geometric techniques and concepts in combination with multivariate statistical techniques for the analysis of variation has been a major advance for the statistical analysis of shape during the last decade. Fundamental to the formalism of what has become known as geometric morphometrics, we find a Riemannian shape manifold with a particular metric called Procrustes distance. This metric measures shape differences in configurations of reference points, represented by sets of rectangular coordinates. The use of Procrustes distances in substitution to regular sums of squares in general linear models is a established procedure of the morphometric toolkit. In this paper, we review the geometric formalism and propose the application of Procrustes distances as replacement for parameters in quantitative genetic models for the assessment of evolutionary processes influencing biological shape changes. A biological case study is examined, where conflicting hypotheses of natural selection and genetic drift could be held responsible for an observed difference in body shape and size of fish populations in different environments of a coastal lagoon. The geometric methods in combination with the evolutionary models allow for a deep analysis of the problem and indicate natural selection as the probable cause for the observed shape difference.

Keywords: Procrustes distances, shape spaces, evolutionary rates, natural selection, linear models.

1 Introduction

The study of biological shape variation and its causes has been a topic of interest for several centuries [1]. For a long time, two methodological schools have provided tools for the analysis and quantification of shape differences between organisms. The most widely used were the methods derived from the biometrical school of genetic analysis, which were based on extracting factors from the covariance structure of collections of distance measurements between two reference points (called landmarks) in biological structures [2]. Such techniques allowed for the study of variation within and between defined groups, but failed to provide a geometrical interpretation of shape differences [1]. On the other hand, a second school based on the tradition of the works of D'Arcy Thompson [3] used techniques for the study of shape changes that preserved the geometric relations among reference points in biological structures, but failed to provide variables that allowed for the study of variation [1]. The solution to this problem was reached after the so called morphometric synthesis, which was the combination of techniques that preserved the geometry of a biological structure's shape, and allowed for the analysis of variation [2].

One of the important developments were precise definitions of "morphometric", "shape" and "size". Morphometrics is defined by Bookstein [2] as the study of covariances of shape and causal factors. Therefore, the objects of morphometric study are not the biological shapes alone, but their variation, causes and consequences. In this context, shape is defined as the geometric properties of an object, that are invariant to rotation, translation and scaling [4]. Indeed the shape space used can be viewed as a set of equivalence classes $[Z] = \{Z\Gamma \ / \ \Gamma \in O(k)\}$; where $Z = Y/||Y||$, $||Y||^2 = trace(Y.Y^T)$ and Y is the matrix obtained from the original data configuration $X = (x_{ij})_{p \times k}$ in the following way: its i-th line Y_i is of the form $Y_i = X_i - \bar{X}$, being $\bar{X}$ the centroid of the p-gone determined by the points $X_i \in \mathbb{R}^k$ (the lines of the matrix X). In such way that the elements of $[Z]$ have centroids at the origin, are normalized and are identified through an element Γ of the group $O(k)$ of orthogonal transformations in $\mathbb{R}^k$ (linear transformations preserving angles). Therefore in the shape space we have represented the original data sets X passing over location, scaling and orientation. Size is defined as any positive real valued function $g(X)$ of a configuration of landmarks that satisfies $g(aX) = ag(X)$ [4]. After the incorporation of geometric concepts, the field of morphometrics has experienced an enormous theoretical development during the last decade [4, 5, 6, 7], Leading to the most statistically powerful methods currently available [8, 9, 10].

The metric subjacent to geometric morphometric methods is a chord distance in a Riemannian space called Procrustes distance. Because it is defined as a sum of squares, it can replace regular univariate sums of squares in general linear models such as regression [11], analysis of variance [12, 13], and can be used for the estimation of important biological parameters such as the heritability of shape [14]. The substitution of regular sums of squares by Procrustes distances reduces the measure of shape variation of multidimensional objects to a scalar and allows for the direct quantification and interpretation of biological processes that cause shape variation without the need of techniques for extracting the latent structure from covariance matrices. Some important linear models have not yet incorporated Procrustes distances as a measure of shape differences. One such example is the calculation of evolutionary rates for shape. Although there are well established methods for the calculation of evolutionary rates [15] and rate tests used for discriminating among possible evolutionary processes influencing univariate characters [16], the calculation of evolutionary rates for multidimensional features, such as shape, is yet to be examined in detail, and the methods so far proposed are based on covariance or correlation matrices calculated from sets of "size" measures, a feature that cannot be separated from shape in such non-geometric data sets [2]. Because the rate tests for natural selection [16] are based on among population mean squares, the Procrustes distances can be readily implemented on the formulas to assess the evolutionary process most likely to be responsible for the observed shape divergence. Our aim in this study is to review the geometric formalism and biological models and problems to which the Procrustes metric can be applied, proposing the incorporation of geometric estimates of shape divergence into evolutionary rate tests for natural selection.

2 Geometric formalism

From the geometrical point of view, each landmark can be considered as a point $X_i = (x_{i1}, ..., x_{ik}) \in \mathbb{R}^k$ in the k-dimensional Euclidean space $\mathbb{R}^k$ (usually, $k = 2$ or $k = 3$; the case $k \geq 4$ is of interest in multivariate statistics, where the shapes of multivariate data sets carry information about normality, linearity and correlation between variables).

A set of p landmarks $\{X_1, ..., X_p\}$ determine a $p \times k$-matrix $X = (x_{ij})_{p\times k}$ whose i-th line is given by the coordinates of the point X_i, $1 \leq i \leq p$. In order to study the shape of the original configurations X, we have to construct certain topological spaces.

2.1 Pre-shapes and the shape space Σ_k^p

We begin with the set of the original data set configurations $X = (x_{ij})_{p\times k}$, where each configuration X is viewed as a real matrix with p lines and k columns. Formally, this space $\mathcal{M}^{p\times k}$ is characterized by

$$X \in \mathcal{M}^{p\times k} \iff X = (x_{ij})_{p\times k} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pk} \end{pmatrix}.$$

For each $i = 1, ..., p$ the i-th line can be identified with the point $X_i = (x_{i1}, ..., x_{ik}) \in \mathbb{R}^k$. Also, for each $j = 1, ..., k$ we denote $X^j = (x_{1j}...x_{pj})^T$ the j-th column of X, which can be identified as a point in $\mathbb{R}^p$.

The second space E_2, called the *preform space*, is obtained after the standardization of location of the configurations X in such way that their centroids $C(X)$ are translated to the origin. The centroid is defined as the *geometric center*

$$C(X) = (\bar{x}^1, ..., \bar{x}^k) := \frac{1}{p}\sum_{i=1}^{p} X_i$$

of the p-gone in $\mathbb{R}^k$ determined by the points $X_1, ..., X_p \in \mathbb{R}^k$, where

$$\bar{x}^j = \frac{1}{p}\sum_{i=1}^{p} x_{ij}.$$

This space E_2 can be defined in the following form

$$E_2 := \{X \in \mathcal{M}^{p\times k} \ / \ C(X) = 0\}.$$

The dimension of the preform space E_2 is $pk - k$.

A third space E_3, the so called *form space*, is obtained after the standardization of location and orientation of the configurations X, i.e., through the identification of two elements Y and $\tilde{Y}$ of the preform space when one of them is applied into the other using an adequate orthogonal transformation $\Gamma \in O(k)$.

In order to define this manifold E_3, we consider the linear group $O(k)$ of the orthogonal linear transformations of the Euclidean space $\mathbb{R}^k$, acting on E_2 throughout the mapping

$$\Phi : O(k) \times E_2 \to E_2$$

defined by

$$(\Gamma, Y) \mapsto \Phi(\Gamma, Y) := Y.\Gamma.$$

Then we define E_3 as being the quotient space $\frac{E_2}{O(k)}$ whose elements are the $O(k)$-fibres determined for the action Φ. More precisely,

$$E_3 := \{[Y] \; / \; Y \in E_2\};$$

where

$$[Y] := \{Y.\Gamma \; / \; \Gamma \in O(k)\}.$$

We have that the dimension of this manifold E_3 is $pk - k - \binom{k}{2}$. [Recall that $O(k)$ is a submanifold of $\mathcal{M}^{k\times k}$ of dimension $\binom{k}{2} = k(k-1)/2$.]

The **pre-shape space** E_4 can be viewed as the submanifold obtained as being the inverse image $f^{-1}(1)$ of the regular value 1 of the function

$$f\colon E_2 \to \mathbb{R}, \; Y \mapsto f(Y) = ||Y||^2 = \sum_{i=1}^{p} ||X_i - C(X)||^2. \tag{1}$$

This space E_4 has dimension $pk - k - 1$. Each element Y of the preform space determines the pre-shape

$$Y/||Y|| \in E_4.$$

The **shape space** Σ_k^p is the quotient space obtained from the pre-shape space, identifying two elements of E_4 which differ by a rotation. This means that each element of this shape space is the equivalence class

$$[Z] := \{Z.\Gamma^T \; / \; \Gamma \in O(k)\};$$

where Z is an element of the pre-shape space.

It follows that

$$dim(\Sigma_k^p) = pk - k - 1 - k(k-1)/2. \tag{2}$$

2.2 The Procrustes Distances

It is of interest in Morphology to have a measure of shape similarity between objects. For this we present the following Procrustes Distances on the above spaces.

In the following, we consider two configurations $X_1, X_2 \in \mathcal{M}^{p\times k}$ and its representing Z_i, $i = 1, 2.$ in the pre-shape space E_4.

The *Full Procrustes Distance* between two configuration matrices $X_1, X_2 \in \mathcal{M}^{p\times k}$ is given by

$$d_F(X_1, X_2) := \min\{||Z_2 - tZ_1\Gamma|| \; /t > 0, \Gamma \in O(k)\}. \tag{3}$$

The *Partial Procrustes Distance* $d_P(X_1, X_2)$ between two configuration matrices $X_1, X_2 \in \mathcal{M}^{p\times k}$ is obtained by matching their pre-shapes Z_1, Z_2 as closely as possible over rotations but not scale. More precisely

$$d_P(X_1, X_2) := \min\{||Z_2 - Z_1\Gamma|| \; /\Gamma \in O(k)\}. \tag{4}$$

The *Procrustes Distance* $\rho(X_1, X_2)$ between two configuration matrices $X_1, X_2 \in \mathcal{M}^{p\times k}$ is the length of the closest arc of great circle (spherical geodesic) between the Z_1 and Z_2 on the pre-shape sphere, i.e.,

$$\rho(X_1, X_2) := 2\arcsin[\frac{1}{2}d_F(X_1, X_2)]. \tag{5}$$

It is a fact that ρ is a metric in the manifold Σ_k^p. This shape space Σ_k^p, provided with the Procrustes Distance ρ, is a Riemannian manifold. This pair (Σ_k^p, ρ) is the so called **Kendall manifold**, which is denoted simply as Σ_k^p.

These three distances are related in the following way

$$d_F(X_1, X_2) = \sin\rho, \quad d_P(X_1, X_2) = 2\sin\frac{\rho}{2}. \tag{6}$$

2.3 Smallest distance between configurations

Two configurations X_1 and X_2 (in figure space) can be compared by similarity when they are in pre-shape space. So, we can rotate efficiently the second configuration by placing the respective points as nearly as possible to the first one.

Let Z_1 and Z_2 be the respective pre-shape configurations of X_1 and X_2. We must obtain the minimizing rotation Γ of Z_1 using partial Procrustes distance as

$$\begin{aligned} d_P^2(X_1, X_2) &= \min\{trace[(Z_2 - Z_1.\Gamma)^T(Z_2 - Z_1.\Gamma)]\} \\ &= \min\{||Z_2||^2 + ||Z_1.\Gamma||^2 - 2\max\{trace(Z_2^T.Z_1.\Gamma)\}\} \\ &= \min\{||Z_2||^2 + ||Z_1||^2 - 2\max\{trace\left(Z_2^T.Z_1.\Gamma\right)\}\}, \end{aligned}$$

where $||Z_2|| = ||Z_1|| = 1$ because Z_1 and Z_2 are elements of pre-shape space, and $||Z_1.\Gamma|| = ||Z_1||$ because the size of Z_1 is preserved after rotation.

The rotation of Z_1 using the minimizing rotation Γ, considered as the best rotation, permit us to have the Procrustes distance that give us a parameter of comparison between configurations X_1 and X_2. We implemented the best rotation operation in C++ programming language to compute the Procrustes distance.

2.3.1 The best rotation

We wish to find the best rotation matrix Γ that maximize $trace(Z_2^T.Z_1.\Gamma)$. This problem we could solve as $R = VSU^T$, where V and U are square orthonormal matrices computed by the singular values decomposition method [17] as $U\Sigma V^T = Z_2^T Z_1$. The matrix S is the identity transformed from Σ preserving the respective sign of diagonal elements, see [4]. Another form best elegant to find the best rotation is by using quaternions [18].

The rotation matrix Γ is expressed by a vector $\mathbf{q} = (s, \mathbf{v})$ called quaternions representation [19, 20], where s is a scalar related to the angle of rotation and $\mathbf{v}$ is the axis, vector, of rotation. Let Z_{1k} be the k-th line of matrix Z_1 (similar case Z_{2k} for Z_2). Then, the vector Z_{1k} must be expressed as $\tilde{Z}_{1k} = (0, Z_{1k})$ to rotate with quaternion $\mathbf{q}$. The rotation of $\tilde{Z}_{1k}$ is given by

$$(0, Z_{1k}.\Gamma) = \mathbf{q}.\tilde{Z}_{1k}.\mathbf{q}^{-1}, ||\mathbf{q}|| = 1. \tag{7}$$

With this, $trace(Z_2^T.Z_1.\Gamma)$ is expressed as

$$\begin{aligned} trace(Z_2^T.Z_1.\Gamma) &= trace((Z_1.\Gamma)^T.Z_2) \\ &= \sum_{k=1}^{p} (\mathbf{q}.(0, Z_{1k}).\mathbf{q}^{-1})^T.(0, Z_{2k}) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{p} (\mathbf{q}.\tilde{Z}_{1k})^T.(\tilde{Z}_{2k}.\mathbf{q}) \\
&= \sum_{k=1}^{p} \mathbf{q}^T.(\mathbb{Z}_{1k}^T.\mathbb{Z}_{2k}).\mathbf{q}) \\
&= \mathbf{q}^T.\sum_{k=1}^{p} (\mathbb{Z}_{1k}^T.\mathbb{Z}_{2k}).\mathbf{q} \\
&= \mathbf{q}^T.\mathbb{N}.\mathbf{q},
\end{aligned}
$$

where $\mathbb{Z}_{1k}^T$ is the expand orthogonal 4×4 matrix of $\tilde{Z}_{1k}$ such that

$$
\mathbf{q}.\tilde{Z}_{1k} = \begin{pmatrix} 0 & -x_{1k} & -y_{1k} & -z_{1k} \\ x_{1k} & 0 & z_{1k} & -y_{1k} \\ y_{1k} & -z_{1k} & 0 & x_{1k} \\ z_{1k} & y_{1k} & -x_{1k} & 0 \end{pmatrix} .\mathbf{q} = \mathbb{Z}_{1k}.\mathbf{q}
$$

and

$$
\tilde{Z}_{2k}.\mathbf{q} = \begin{pmatrix} 0 & -x_{2k} & -y_{2k} & -z_{2k} \\ x_{2k} & 0 & -z_{2k} & y_{2k} \\ y_{2k} & z_{2k} & 0 & -x_{2k} \\ z_{2k} & -y_{2k} & x_{2k} & 0 \end{pmatrix} .\mathbf{q} = \mathbb{Z}_{2k}.\mathbf{q},
$$

considering $\tilde{Z}_{rk} = (0, x_{rk}, y_{rk}, y_{rk})$, for $r = 1, 2$.

The problem of maximizing $\mathbf{q}^T.\mathbb{N}.\mathbf{q}$ with $||\mathbf{q}|| = 1$ has solution when $\mathbf{q}$ is the unit eigenvector corresponding to the greater eigenvalue of matrix $\mathbb{N}$ [18], that is easily computed.

3 Evolutionary rate tests

The tests of evolutionary rates and evolutionary divergence are based on the assumption that, under neutral evolution (genetic drift) models, continuous phenotypic characters have an expected mean at generation t that is approximated by a normal distribution. The parameters of this distribution depend on the choice of a particular model [16]. There are two models available: the constant-heritability of Lande [21, 22] and the mutation-drift equilibrium of Turelli et al. [23]. Under the constant-heritability (CH) model, the expectation for the mean of a continuous character z at a time t is

$$
\bar{z}(t) \sim N\left[\bar{z}(0), \frac{\sigma^2 h^2 t}{N_e}\right], \tag{8}
$$

where $\bar{z}(0)$ is the mean of character z when $t = 0$, σ^2 is the variance of the character, h^2 is the heritability of the character, and N_e is the effective population size. In this model, deviation from the mean at $t = 0$ should increase for larger variances, heritabilities and the number of generations passed, but should decrease for large effective population sizes (as expected in models of genetic drift). Whenever the observed deviation from the initial mean is larger than expected by the neutral model, we can infer that a process of directional selection is taking place. When the observed deviation is smaller than expected, we infer a process of stabilizing selection maintaining the similarity among generations. If the observed deviation is within the expected interval, we accept the null hypothesis of neutral evolutionary processes. The test statistics for evolutionary rates follow an F-distribution with $n-1$ degrees of freedom in the numerator (n is the number of lineages studied at once) and infinite degrees of freedom in the denominator. The statistic for the test of directional selection in the constant-heritability model is

$$F = \frac{z^2}{\sigma^2 h^2 t / N_e}, \tag{9}$$

where $z = \bar{z}(t) - \bar{z}(0)$, the difference between mean values at generations t and 0. If there is no information regarding the ancestral mean, we can also compare populations descended from the same ancestor. In this case, the evolutionary rate test can be transformed into a divergence rate test as

$$F = \frac{S^2_{\bar{z}(t)}}{\sigma^2 h^2 t / N_e}, \tag{10}$$

where the difference between generations is substituted by an among population mean square

$$S^2_{\bar{z}(t)} = \frac{\sum \left[\bar{z}_i(t) - \bar{\bar{z}}(t)\right]^2}{(n-1)}, \tag{11}$$

where the term $\bar{\bar{z}}(t)$ corresponds to the grand average over all populations.

The mutation-drift equilibrium (MDE) model expects the value of the continuous character under neutral evolution to approximate

$$\bar{z}(t) \sim N\lfloor \bar{z}(0), 2t\sigma_m^2 \rceil, \tag{12}$$

where σ_m^2 is the mutational variance, the rate by which new variation is added to the continuous character by mutation. Accordingly, the test statistic for directional selection in divergence rates for the MDE model will be

$$F = \frac{S^2_{\bar{z}(t)}}{2t\sigma_m^2}, \tag{13}$$

The tests statistics for stabilizing selection are obtained by the inversion of (9), (10) and (13). Because of the inversion the new degrees of freedom in the numerator are infinite and in the denominator are equal to $n-1$.

The Procrustes distances have successfully replaced sums of squares in general linear models of multivariate regression [11, 24], analysis of variance for testing group differences [12], for testing fluctuating asymmetry [13], and quantitative genetics [14]. To address questions related to the evolution of multidimensional shape using the models for evolutionary rates, we can replace (11) by

$$S^2_{\bar{z}} = \frac{\sum d_p^2(\bar{Z}_i, \bar{\bar{Z}})}{(n-1)(dim\Sigma_k^p)}, \tag{14}$$

where the numerator corresponds to the sum of squared partial Procrustes distances between $\bar{Z}_i(t)$ (the mean shape of population i) and $\bar{\bar{Z}}(t)$ (the grand average calculated from n populations). The inclusion of the term $dim\Sigma_k^p$ in the denominator represents the dimensionality of shape space, as defined in (2).

The choice of which model to use depends on the effective population size (N_e) and the number of generations (t) passed between ancestors and descendants or the number of generations as separate populations for divergence rate tests. The CH model is designed for populations recently separated which have not yet achieved mutation-drift equilibrium. It is appropriate if $t < N_e/5$. For populations in equilibrium, one should use the MDE, which will be appropriate if $t > 4N_e$ [23]. In general, one should use the CH model for populations recently separated and the MDE model for populations or species separated for a long time (see Spicer [16] for a discussion).

The estimation of test parameters may be problematic, particularly those related to quantitative genetics (heritatibilities, mutational variances, effective population sizes, generation times and number of generations passed). A large number of parameter estimates for various characters and organisms can be found in the literature [16]. Usually, reasonable and conservative estimates should be preferred [26], but because of the uncertainty related to parameter estimation, these tests should not be considered rigorous quantitative tests, but qualitative indications of evolutionary processes [23].

4 Application to a biological problem: evolutionary processes driving body shape divergence between fish populations from a coastal lagoon

The plains at northern Rio de Janeiro State were formed by sediment deposition from the Paraíba do Sul River during the Holocene [27]. The process of sediment deposition formed a large number of lagoons with various environmental conditions. Among the most recent (2000 years or less) are the coastal lagoons formed by closure (a sand bar by river bound sediment) of fluvial channels connecting the Paraíba do Sul River and the sea [28].One such lagoon is the Iquipari lagoon, which is narrow but very long (about 10 km). Because of its nature, the Iquipari lagoon presents an environmental gradient along its longest axis (from the ocean inwards). The environmental gradient [29] is observable in water salinity, nutrients, fauna (freshwater species are not found in the sand bar region), and flora. A large bank of submerged macrophytes and macroalgae is observable in the sand bar region (the margin is clear of plants), whereas the interior of the lagoon is dominated by macrophytes, particularly *Typha dominguensis*, that grow on marginal shallow waters. This environmental gradient creates contrasting situations where animal populations inhabiting the same lagoon may experience different environments regarding water chemistry (with physiological implications), structural complexity (favoring different types of locomotion), and predation regimes. The livebearing fish *Poecilia vivipara* is widespread on the lagoons of northern Rio de Janeiro. Its high tolerance to environmental changes, particularly salinity and temperature [30], makes it one of the few species abundant in all lagoon environments of the region. Because in poeciliine fishes, females have lower mobility than males [31, 32], they experiment little environmental variation during their lifetime, and thus, are expected to be adapted by natural selection to the particular environmental conditions of the region they inhabit.

Body shape evolution in fish is mostly driven by locomotion needs, relating to the structure of the environment [33, 34]. Fish that swim in open areas for longer periods of time have slender bodies, which provide less thrust or acceleration, but are more efficient in long range, high speed swimming [34]. Fish that live in structurally complex environments (such as a coral reef or dense banks of macrophytes) benefit from higher, laterally compressed bodies, which provide more thrust and more manoeuvrability [34]. Given the nature of environmental differences observed within coastal lagoons in northern Rio de Janeiro, we can expect that populations inhabiting different environments (even within the same lagoon) will present body shape differences.

In order to test this hypothesis, we conducted a study where we collected *Poecilia vivipara* specimens from two different sites in Iquipari lagoon: one population from the sand bar region and one population from the interior of the lagoon. From each site, 60 females were collected fixed and stored in 10% formalin. The specimens were photographed by a high resolution digital camera and the coordinates of 12 landmarks (Figure 1) were registered for each, using the program TpsDig [36].

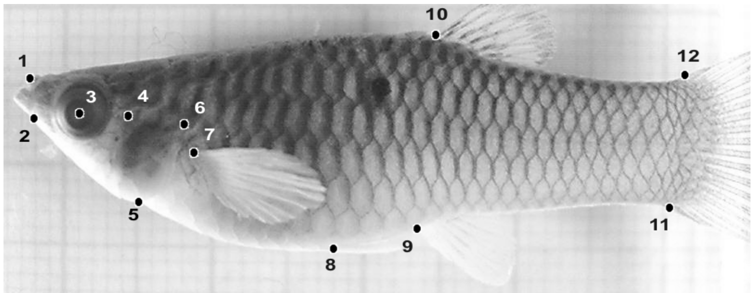

Figure 1: Landmarks used in this study: 1- snout tip; 2- articulation of lower jaw; 3- center of eye; 4- anterodorsal extremity of opercle; 5- ventral conjunction of opercula; 6- posterodorsal extremity of opercle; 7- dorsal insertion of pectoral fin; 8- anterior insertion of ventral fin; 9- anterior insertion of anal fin; 10- anterior insertion of dorsal fin; 11- ventral insertion of caudal fin; 12- dorsal insertion of caudal fin.

The landmark configurations were superimposed by the generalized Procrustes analysis described above, and the visualization of resulting aligned configurations in figure space indicate a difference between mean shapes in the two populations (Figure 2).

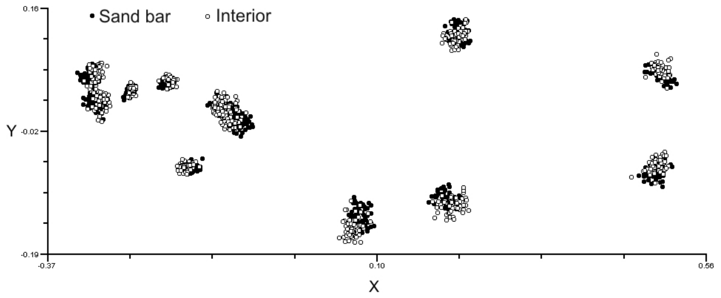

Figure 2: Superimposed landmark configurations in figure space for specimens colected in the sand bar and interior of Iquipari lagoon.

The next step is to determine whether the observed mean shape differences are statistically significant. This test can be performed by calculation of an F-test derived by Goodall [12] which uses the sum of squared Procrustes distances between group means and an overall mean as an among groups sums of squares, and the sum of squared Procrustes distances between each individual and its group mean as a within-group sums of squares. After division by appropriate degrees of freedom, the ratio of the among group and within-group mean squares is distributed as an F-statistic with $pk - k - k(k-1)$ degrees of freedom in the numerator and $pk - k - k(k-1)\sum_{i=1}^{a}(n_i - 1)$ degrees of freedom in the denominator, where p = number of landmarks, k = number of dimensions (2 for planar figures, 3 for three-dimensional objects) and n_i is the sample size for group i, which is summed over a groups. For the particular case of testing the significance of shape difference between *Poecilia vivipara* populations in the Iquipari lagoon, the F-test was significant ($F = 20.0739, df_1 = 20, df_2 = 2360 : P < 0.00001$).

After determining that there is a significant shape difference between the two populations, the next step would be to use one of the evolutionary models explicited above to assess whether the amount of shape difference can be explained by the neutral model of evolution. In deciding which model should be used, we refer to the quantities referred by Turelli et al. [23]. If $t < N_e/5$, we should use the constant-heritability model, if $t > 4N_e$, we should use the mutation-drift equilibrium model. In our case, the easiest quantity to determine is t, the number of generations since divergence. A conservative estimate of time first has to consider that the populations were established at the same time as the formation of the lagoon. From geological dating results, we can infer that the lagoon of Iquipari is under 2000 years of age, for it is very close to the ocean [27]. We conservatively set the time of separation of the populations to 2000 years and translate it to number of generations using an estimate of 2 generations/year [35]. As a result we can set a maximum limit to 4000 generations passed since divergence. The next step would be to calculate the effective population size(N_e), which can be estimated from total population size (N), for the smallest N_e/N ratios are around 0.5 [16]. From density estimates of number of fishes by marginal area, we can conservatively estimate the population size in the Iquipari lagoon to be around 300000 specimens (but it is probably larger). Considering the N_e/N ratio to be 0.5, we can calculate N_e to be 150000. Because $4000 < (150000/5)$, we should use the constant heritability model for the test of directional selection. The remaining parameters of the model that have to be estimated are: between group shape difference ($S^2_{\bar{z}(t)}$), shape variance (σ^2) and shape heritability (h^2). The estimate of between-group shape difference using Procrustes distances in (14) is 0.00002739033. The within-group variance calculated using Procrustes distances is 0.00008166949. Heritability estimates are more difficult to obtain directly, for they represent the proportion of total variance that is caused by additive effects of genes (as opposed to dominance, epistasis and environmental variance) [25]. Heritability estimates for morphological characters in fish species vary from 0.14 to 0.80, where even the lowest values are statistically significant [37, 38, 39, 40, 41].

The F-value in 10 can be thought of as a function and the uncertainty of parameter estimation can be dealt with by considering sensible intervals over which the parameters may vary to check which particular combinations of parameters may lead to acceptance or rejection of the neutral model. Using this method, it is possible to deal with uncertainty in at least two parameters and we can visualize the effect of varying parameters as a surface, where the parameters are designated to X and Y axes, and the F-statistic is the Z axis (Figure 3).

Using the constant-heritability model from 10, but substituting the estimates of between group and within group variances by sums of Procrustes distances, we can test for directional selection over a defined interval of uncertain parameters. In our case, uncertainty is most present in estimates of heritability and effective population size, therefore. Considering that the one-tailed F-value (because we are testing only for directional selection) on the statistical table, for 20 and infinite degrees of freedom, and alpha 0.01 is 1.88, we can assess which region of the parameter space would lead to acceptance of the null hypothesis (which combinations produce Fs smaller than 1.88). In Figure 3, we see that it takes a maximum effective population size of 25000 for the null model of genetic drift to be responsible for the observed magnitude of morphometric differentiation. Based on rough, but conservative approximations, we can estimate the effective population size in the lagoon to be much larger than that (probably larger than 150000). If we are to consider the estimated population size as true, any heritability value different of zero (the existence of significant heritability is one of the conditions for natural selection to operate) will lead to rejection of the null hypothesis.

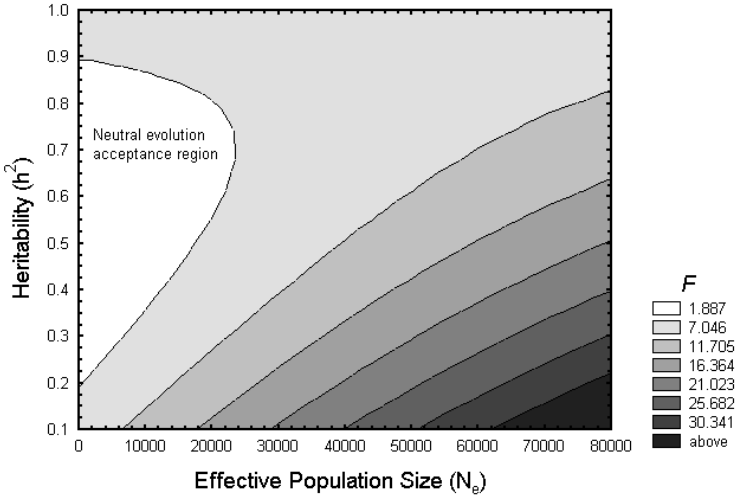

Figure 3: Contour plot of the F-surface for heritability and effective population size.

The analysis of size differences presented similar results to the shape analysis. In geometric studies of shape such as the one we performed, size is measured by the square root of summed squared differences between each landmark and the mean landmark, also called centroid. This size measure is called centroid size [2]. Because size is inherently univariate, it can be treated by standard methods (equations 10 and (11) can be used directly). The surface obtained for the test of directional selection on size (not shown) is remarkably similar to that obtained for shape. Hendry and Kinnison [15] review evolutionary rates in microevolutionary studies. These authors identify two measures of evolution with different implications and interpretations: darwins and haldanes. Evolutionary rates measured in darwins, $d = (\ln z_1 - \ln z_2)/t$, are simply the mean difference in natural logarithms between two populations divided by the time of separation in millions of years. The rates in haldanes, $h = [(z_1/s_p) - (z_2/s_p)]/g$, are mean differences standardized by pooled within group standard deviations (s_p) divided by the number of generations since divergence. The biological and statistical differences between the two measures of evolution are discussed in depth by Hendry and Kinnison [15]. Although the body shape differences in Procrustes distances do not fit directly in the formulas for evolutionary rates, the evolution in body size can be measured by them. In our study the observed evolutionary rates for body size divergence were 110 darwins and 0.000397 haldanes. These values allow for direct comparison with evolutionary rates published for other species and characters.

5 Discussion

Our results have shown that the use of Procrustes distances in substitution to regular sums of squares can be valuable and highly informative for the understanding of evolutionary processes influencing shape variation in biological structures. The use of Procrustes distances in general linear models has been suggested for regression [11], analysis of variance [12, 13], and quantitative genetics [14]. The present study is, however, the first proposal to the use of Procrustes distances as a measure of differentiation in the calculation of rate tests for natural selection. In addition, the simple method of calculating a F-surface for dealing with uncertainty in estimated parameters had never been used before, and should be very useful when the parameters cannot be accurately estimated, but sensible and conservative intervals of variation can be ascertained.

The combination of geometric methods with evolutionary models was highly informative regarding possible processes influencing shape variation of body shape in the populations of *Poecilia vivipara*. Given that the estimates of uncertain parameters were very conservative, we can indicate natural selection as a probable process responsible for the observed shape differences in populations inhabiting different environments. The magnitude of the observed shape difference was too large for genetic drift to explain the observed pattern. If the effective population size was smaller than 25000 individuals, we would also accept the null model of genetic drift. However, that population size is much lower than the conservative estimates we provided and the null model should be rejected.

Although the Procrustes distances do not apply directly to existing evolutionary rates (darwins or haldanes) for the comparative assessment of relative speeds of evolutionary modifications, it is possible to devise an evolutionary rate based on Procrustes distances, which would lead to the comparison of rates between different structures or species. One of the major concerns regarding the comparison of evolutionary rates is the scale where each rate is measured [15]. Dividing the Procrustes distance between two mean shapes (two different populations or the same population in different generations) by the appropriate degrees of freedom (the dimensionality of shape space), and dividing that by the number of generations, we would obtain a measure of the rate of change in shape space per generation that could be compared with estimates of the same rate for shapes in different structures or different populations or species.

The difference observed in body size was also remarkably high, and the resulting patterns and processes inferred were the same as those for the analysis of body shape. Because size is a univariate character, its evolutionary rates could be compared with that for other species and characters. The evolutionary rate observed falls within the interval of evolutionary rates observed in different studies regarding other Poeciliine species, such as *Poecilia reticulara* and *Gambusia affinis*, even though the number of generations in our study was considerably larger than those reported in the literature [15].

The field of morphological evolution has benefited greatly from the introduction of geometric methods in the analysis of shape variation, particularly because the characteristic spatial localization of factors generated by geometric morphometric analyses are readily intepreted by models for the development and evolution of complex morphological structures [42]. The use of Procrustes distances for the calculation of evolutionary rates and rate tests is an important methodological advance that should improve our understanding of the processes causing the diversity of shapes observed in the natural world.

References

[1] Bookstein, F. L. A brief history of the morphometric synthesis. *Marcus, L. F., E. Bello and A. Garcia-Valdecasas (eds.). Contributions to Morphometrics. Monografias*, Museo Nacional de Ciencias Naturales, Madrid, pp. 15-40, 1993.

[2] Bookstein, F. L. Morphometric Tools for Landmark Data. Geometry and Biology. *New York; Cambridge University Press*, 1991.

[3] Thompson, D'A. W. On Growth and Form. *MacMillan, London*, 1917.

[4] Dryden, I. L. and Mardia, K. V. Statistical shape analysis. *John Wiley and Sons*, New York, NY, 1998.

[5] Rohlf, F. J. and Marcus, L. F. A revolution in morphometrics. *TREE*, 8(4): 129-132, 1993.

[6] Bookstein, F. L. Biometrics, biomathematics and the morphometric synthesis. . *Bull. Math. Biol.*, 58(2): 313-365, 1996.

[7] Marcus, M. Corti, A. Loy, G. J. P. Naylor and D. E. Slice. Advances in morphometrics. *NATO ASI Series A: Life Sciences*, vol. 284. Plenum Press, New York, 1996.

[8] Monteiro, L. R.; Bordin, B. and Reis, S. F. Shape distances, shape spaces and the comparison of morphometric methods. *Trends in Ecology and Evolution*, v. 15, p. 217-220, 2000.

[9] Rohlf, F. J. Statistical power comparisons among alternative morphometric methods. *Amer. J. Phys. Anthropol.*, 111:463-478, 2000.

[10] Rohlf, F. J. On the use of shape spaces to compare morphometric methods. *Hystrix It. J. Mamm. (N.S.)*, 11(1):8-24, 2000.

[11] Monteiro, L. R. Multivariate regression models and geometric morphometrics: the search for causal factors in the analysis of shapeo. *Systematic Biology*, 48: 192-199, 1999.

[12] Goodall, C. R. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society B*, v. 53, p. 285-339, 1991.

[13] Klingenberg, C. P. and McIntyre G. S. Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with Procrustes methods. *Evolution*, 52: 1363-1375, 1998.

[14] Monteiro, L. R.; Diniz-Filho, J. A. F.; Reis, S. F.; Arajo, E. D. Geometric estimates of heritability in biological shape. *Evolution* , 56: 563-572, 2002.

[15] Hendry, A.P. and Kinnison M.T. Perspective: The pace of modern life: Measuring rates of contemporary microevolution. *Evolution*, 53: 1637-1653, 1999.

[16] Spicer, G. S. Morphological evolution of the Drosophila virilis species group as assessed by rate tests for natural selection on quantitative characters. *Evolution*, v 47, p. 1240-1254, 1993.

[17] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T. and Flannery, B. P. Numerical Recipes in C: The Art of Scientific Computing. *Cambridge*, 1992.

[18] Horn, B. K. P. Closed-form solution of absolute orientation using unit quaternios. *Opt. Soc. Am. A*, v.4, n.4, 629-642, 1987.

[19] Pletinckx, D. Quaternio calculus as a basic tool in computer graphics. *The Visual Computer*, 5:2-13, 1989.

[20] Watt, A. and Watt, M. Advanced Animation and Rendering Techniques. *Addison-Wesley*, 1992.

[21] Lande, R. Natural selection and random genetic drift in phenotypic evolution. *Evolution*, v. 30, p. 314-334, 1976.

[22] Lande, R. Statistical tests for natural selection on quantitative characters. *Evolution*, v 31, p. 314-334, 1977.

[23] Turelli, M.; Gillespie, J. H.; Lande, R. Rate tests for selection on quantitative caracters during macroevolution and microevolution. *Evolution*, v 42, p. 1085-1089. 1988.

[24] Monteiro, L. R., Abe, A. S. Functional and historical determinants of shape in the scapula of xenarthran mammals: the evolution of a complex morphological structure. *J. Morphol.*, 241: 251-263, 1999.

[25] Lynch, M. and Walsh, M. Genetics and analysis of quantitative traits. *Sinauer, Sunderland, MA*, 1998.

[26] Diniz-Filho, J. A. F. Métodos filogenéticos comparativos. *Ribeirão Preto; Holos Editora*, 2000.

[27] Martin, L., Suguio, K., Dominguez, J. M. L. and Flexor, J. M. Geologia do quaternário costeiro do litoral norte do Rio de Janeiro e do Espírito Santo. *Belo Horizonte: CPRM*, 1997.

[28] Soffiati, A. Aspectos históricos das lagoas do Norte do Estado do Rio de Janeiro. *In Esteves, F. A. (ed). Ecologia das Lagoas Costeiras do Parque Nacional da Restinga de Jurubatiba e do Município de Macaé, RJ. Rio de Janeiro: NUPEM-UFRJ*, p. 1-35, 1998.

[29] Suzuki, M.S., Figueired, R.O., Castro, S.C., Silva, C.F., Pereira, E.A., Silva, J.A. and Aragon, G.T. Sand bar opening in a coastal lagoon (Iquipari) in the northern region of Rio de Janeiro State: hydrological and hydrochemical changes. *Braz. J. Biol.*, 62: 51-62, 2002.

[30] Trexler, J. C. Phenotypic plasticity in poeciliid life histories. *Meffe, A.; Snelson, F. F. (eds.). The ecology and evolution of poeciliid fishes (Poeciliidae)*, New Jersey: Prentice Hall, p. 201-213, 1989.

[31] Becher, S. A. and Magurran, A. E. Gene flow in Trinidadian guppies. *Journal of Fish Biology 56*, 241-249, 2000.

[32] Johnson, J. B. Hierarchical organization of genetic variation in the Costa Rican livebearing fish Brachyrhaphis rhabdophora (Poeciliidae). *Biological Journal of the Linnean Society*, 72: 1637-1653, 2001.

[33] Pakkasmaa, S. and Pironen, J. Morphological differentiation among local trout (Salmo trutta) populations. *Biological Journal of the Linnean Society*, 72: 231-239, 2001.

[34] Walker, J. A. Ecological morphology of lacustrine threespined stickleback Gasterosteus aculeatus L. (Gasterosteidae) body shape. *Biological Journal of the Linnean Society*, 61, 3-50, 1997.

[35] Reznick, D.N., Shaw, F.H., Rodd, F.H. and Shaw, R.G. Evaluation of the rate of evolution in natural populations of guppies (Poecilia reticulata). *Science*, 275: 1934-1937, 1997.

[36] Rohlf, F. J. TPSDig. Version 1.17. Stony Brook: Department of Ecology and Evolution. *State University of New York at Stony Brook*, 1998.

[37] Smoker, W. W.; Gharret, A. J.; Stekoll, M. S.; Joyce, J. E. Genetic-analysis of size in an anadromous population of pink salmon. *Canadian Journal of Fisheries and Aquatic Sciences*, v. 51, p. 9-15, 1994.

[38] Taniguchi, N.; Yamasaki, M.; Takagi, M.; Tsujimura, A. Genetic and environmental variances of body size and morphological traits in communally reared clonal lines from gynogenetic diploid ayu, Plecoglossus altivelis. *Aquaculture*, v. 140, p. 333-341, 1996.

[39] Jonasson, J. and Gjedrem, T. Genetic correlation for body weight of Atlantic salmon grilse between fish in sea ranching and land-based farming. *Aquaculture*, v. 157, p. 205-214, 1997.

[40] Choe, M. K.; Yomozaki, F. Estimation of heritabilities of growth traits, and phenotypic and genetic correlations in juvenile masu salmon Oncorhynchus masou. *Fisheries Science*, v. 64, p. 903-908, 1998.

[41] Nakajima, M. and Taniguchi, N. Genetic control of growth in the guppy (Poecilia reticulata). *Aquaculture*, v. 204, p. 393-405, 2002.

[42] Atchley, W. R. and Hall, B. K. A model for development and evolution of complex morphological structures. *Biol. Rev.* 66: 101-157, 1991.