# Geometric morphometrics

Julius Gelšvartas

The main goal of morphometrics is to study how shapes vary and their covariance with other variables. Even though morphometrics can be used to describe the form of any object it is mostly used in biology to describe organisms. Morphometrics is very important in biology because it allows quantitative descriptions of organisms. Quantitative approach allowed scientists to compare shapes of different organisms much better and they no longer had to rely on word descriptions that usually had the problem of being interpreted differently by each scientist. This shift to quantitative descriptions was caused by advances in statistical analysis methods that allowed to interpret collected data.

First method of morphometrics called "Traditional" morphometrics was done by measuring linear distances (such as length, width, and height) and multivariate statistical tools were used to describe patterns of shape variation within and among groups. This approach also sometimes used counts, ratios, areas and angles measures. The biggest advantage of this method was that it was very simple, however it had several difficulties. The biggest problem was that linear distance measurements are usually highly correlated with size and this makes shape analysis difficult. Another problem was that measurements taken from two different shapes could produce equal results because the data did not include the location of where the measurements were taken relative to each other. And it was also not possible to reconstruct graphical representation of the shape from taken measurements. Figure 1 illustrates the problems of "Traditional" morphometrics. To overcome these problems a more sophisticated method called Geometric morphometrics was created.

Landmark-based geometric morphometrics uses a set of landmarks to describe shape. Landmark is a two- or three-dimensional point described by a tightly defined set of rules. The results that are generated by this method directly depend on the quality of landmarks. A lot of efforts have to be put to choose landmarks that would have high evolutionary significance. Each landmark also has to be present on every studied organism. If a landmark is not present on at least one of studied organisms it either has to be marked approximately or it can not be used at all. The number of landmarks selected should not exceed the number of specimen samples because extra landmarks will be redundant. Usually number of landmarks is approximately equal to the number of specimen samples. There are three types of landmarks that can be used. True landmarks that have some biological significance. Pseudo-landmarks are defined by relative locations e.g. "the point of highest curvature of this bone". Semi-landmarks are defined by a location relative to other landmarks e.g. "midway between landmarks X and Y". Landmarks can sometimes have weighted value in analysis according to their importance.

Extracted landmark data has a lot of variations in position, orientation and scale between specimens. These non-shape variations have to be removed before further analysis. There are several methods used to superimpose landmarks each of them having different optimization criteria. The most simple one is two-point registration. This method translates, scales and rotates all landmarks such that two named landmarks are in the same place in all specimens. The biggest disadvantage of this method is that it removes all the data from those two landmarks. Another popular method is Generalized Procrustes analysis (GPA, also sometimes called Generalized least squares). This method first calculates the centroid of landmark configurations and translates it to the origin. This is done by taking $k$ points in two dimensional space

$$\left(\left(x_1, y_1\right), \left(x_2, y_2\right), ..., \left(x_k, y_k\right)\right)$$

The mean of these points is $(\bar{x}, \bar{y})$ where

$$\bar{x} = \frac{x_1 + x_2 + ... + x_k}{k}, \bar{y} = \frac{y_1 + y_2 + ... + y_k}{k}$$

And all points are translated to the origin

$$(x, y) \rightarrow (x - \bar{x}, y - \bar{y})$$

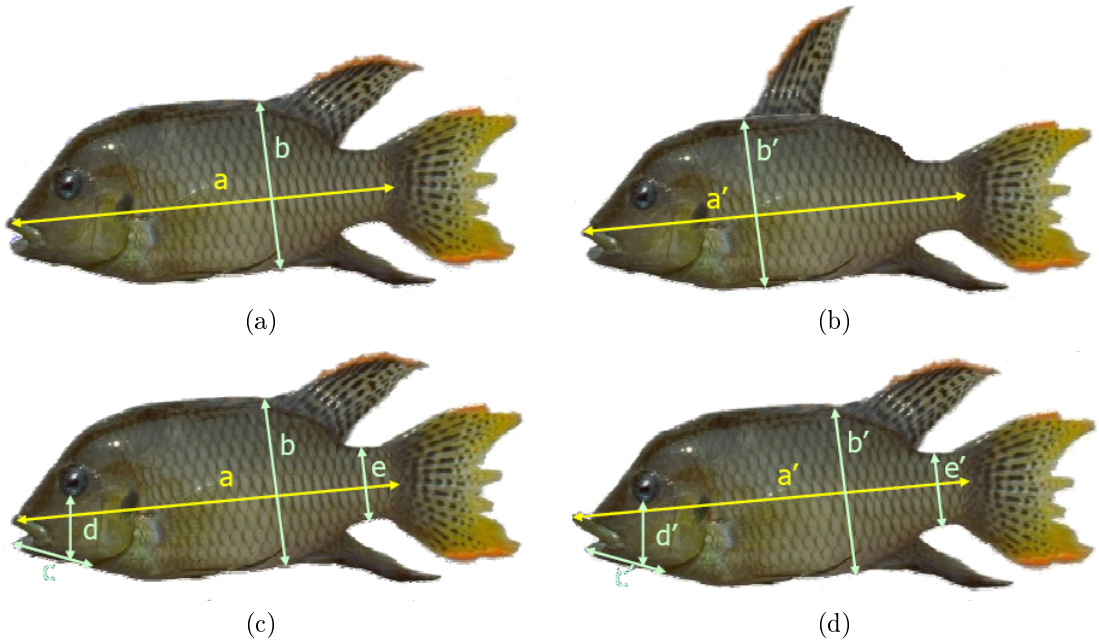


Figure 1: Size variables can be insufficient model of shape as seen in (a) and (b) $a = a'$, $b = b'$ that would lead to conclusion that shapes are equal. Length ratios can be used as seen in (c) and (b) the problem than is that $a \neq a'$and lengths are compared $b/a \neq b'/a'$, $c/a \neq c'/a'$, $d/a \neq d'/a'$, $e/a \neq e'/a'$ conclusion shapes are completely different even though $b = b'$, $c = c'$, $d = d'$, $e = e'$. Data from [2].

giving points

$$\left(\left(x_1 - \bar{x}, y_1 - \bar{y}\right), \left(x_2 - \bar{x}, y_2 - \bar{y}\right), ..., \left(x_k - \bar{x}, y_k - \bar{y}\right)\right)$$

Points are than scaled to a unit size. First centroid size is found

$$s = \sqrt{\left(x_1 - \bar{x}\right)^2 + \left(y_1 - \bar{y}\right)^2 + ... + \left(x_k - \bar{x}\right)^2 + \left(y_k - \bar{y}\right)^2}$$

and all the points are scaled

$$\left(\left(\left(x_1 - \bar{x}\right)/s, \left(y_1 - \bar{y}\right)/s\right), ..., \left(\left(x_k - \bar{x}\right)/s, \left(y_k - \bar{y}\right)/s\right)\right)$$

Finally, the rotation is calculated by minimizing the sum of squared distance between corresponding landmarks. This step is more complicated. We have two objects with point coordinates

$$\left(\left(\left(x_1, y_1\right), ..., \left(x_k, y_k\right)\right), \left(\left(w_1, z_1\right), ..., \left(w_k, z_k\right)\right)\right)$$

One object is fixed and the other one is rotated around the origin so that the squared distance between points is minimized. Rotation by $\theta$ angle gives coordinates

$$(u, v) = (cos\theta w - sin\theta z, sin\theta w + cos\theta z)$$

distance between all points is calculated

$$d = \sqrt{\left(u_1 - x_1\right)^2 + \left(v_1 - y_1\right)^2 + ... + \left(u_k - x_k\right)^2 + \left(v_k - y_k\right)^2}$$

we want to minimize this distance using least squares method. In order to do this we want to find $\theta$ which gives the minimum squared distance. This requires to take the derivative of $d^2$ with respect to $\theta$ and solving for $\theta$ when derivative is equal to zero. This gives

$$\theta = tan^{-1}\left(\frac{\sum_{i=1}^{k}\left(w_i v_i - u_i z_i\right)}{\sum_{i=1}^{k}\left(u_i w_i + v_i z_i\right)}\right)$$

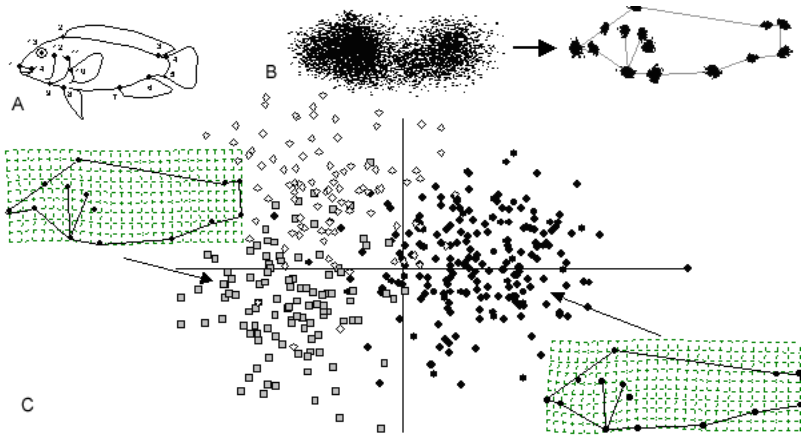


Figure 2: Graphical representation of the four-step morphometric protocol. A: Quantify raw data (landmarks recorded on body of cichlid fish), B: Remove non-shape variation (landmarks of 412 specimens before and after GPA), C: statistical analysis (CVA) and graphical presentation of results. Deformation grids for mean specimen for (right) Eretmodus cyanostictus and (left) Spathodus erythrodon (magnified by 3X to emphasize shape differences). Data from [3].

all points of second object can than be rotated by $\theta$ angles with respect to the origin.

After performing GPA to all samples shape differences can be extracted by calculating differences in coordinates of corresponding landmarks. This data is than used in multivariate analysis to compare shape variations. Principal component analysis (PCA), canonical variates analysis (CVA) and factor analysis are some of commonly used tools. An alternative method is to use thin-plate spline that allows to map the deformation in shape from one object to another. This method calculates the transformation grid that shows how one object can be deformed into another. Usually object is compared to the mean shape. Parameters of these deformations can than be used to statistically compare variations in shape between populations. The biggest advantage of geometric morphometrics is that it captures geometry of analyzed objects, and preserves this information throughout the analysis. This allows to see results visualized not only as statistical scatter plots but also as configurations of landmarks points. Figure 2 show an example of geometric morphometrics analysis.

The biggest disadvantage of landmark-based geometric morphometrics methods is that a number of landmarks available can sometimes be insufficient to capture the shape of an object. An alternative is to use outline analysis method. This method first extracts a boundary around an object. Points are than digitized along the boundary. These points are than fitted with a mathematical function (usually some form of Fourier analysis). Different curves are than compared using coefficients of the functions as shape variables in multivariate analysis. This approach however also has some limitations. This method is not capable of capturing shape changes inside an object (landmark-based method can have landmarks not only on the boundary). This method is also hard to apply when analyzed data is three dimensional.

Geometric morphometrics is an active research area and there are new methods proposed to address the limitations of methods introduced above. One particularly promising method aims to create a method that would join landmark-based and outline analysis into one method taking advantages from both. This method proposes the use of sliding semi-landmarks that are created between real landmarks but on the boundary of an object. This method can also be adopted to be used with three dimensional data. The biggest problem of these new methods is that there still is no well approved technique on how to use them. More information can be found on [1] review and also on [4] website.

# References

[1] D.C. Adams, F.J. Rohlf, and D.E. Slice. Geometric morphometrics: ten years of progress following the 'revolution'. *Italian Journal of Zoology*, 71(1):5–16, 2004.

[2] D. Adriaens. Workshop "deformities in fish larvae". In *Geometric morphometrics as a useful tool for visualising and analysing deformities in fish*, March 11, 2005.

[3] L. Rüber and D. C. Adams. Evolutionary convergence of body shape and trophic morphology in cichlids from lake tanganyika. *Journal of Evolutionary Biology*, 14:325–332(8), March 2001.

[4] F. James Rohlf. Morphometrics [http://life.bio.sunysb.edu/morph/], 02 2010.