Lab 8: Intro to Geometric Morphometrics

Introduction

In this lab, we hope to show you how to derive useful variables from geometric morphometrics, by giving you a surface understanding of the theory behind it and experience working with morphometrics software. If you are interested, a valuable morphometrics resource is Jim Rohlf's website at Stony Brook University (http://life.bio.sunysb.edu/morph). The page includes an archive of software (including one program we'll use today), as well as lists of upcoming meetings, a bibliography of morphometrics resources, a glossary of terms, and contact information for people working on morphometrics.

Morphometrics is the mathematical study of shape. There are many times that we are interested in shape for biological studies. For example, when studying allometry or heterochrony, it is often the occurrence of specific shapes relative to other factors such as size or age that is at question. Many of you cold include an analysis of shape in your project. It would be very nice if we could put numbers on shapes and compare those numbers to each other for the statistical analysis of hypotheses about changes in shape.

To study shape first we must define it. We all have an intuitive sense of what the word means, but mathematics requires a more precise definition. According to Kendall (1977) shape is "all the geometric information that remains when **location**, **scale** and **rotational effects** are filtered from an object." As a practical matter we can only study shape by comparing one object to another. That means we have to take a bunch of information about the physical distribution of the different parts of multiple organisms, and then move them into the same place, make them the same size and rotate them, so that they all line up without changing the relative positions of the parts of the organism.

Shape space is a critical concept for understanding how this can be accomplished. The idea is that you take an object and identify the coordinates of a number of homologous points on that object called **landmarks**. That entire object can now be mapped as a single point into a space with as many dimensions as the number of landmarks multiplied by the number of coordinates describing each landmark. For example imagine a line between two points on a two dimensional graph. You could describe that line by assigning an x and a y coordinate to each of those points, giving you four total coordinates. Those four coordinates could now be graphed as one point in a 4-dimensional space. For a triangle you would add one more point and thus you would need a 6-dimensional space to graph it as a single point. If you recorded your original coordinates in three dimensions (x, y and z) you would require a 6D space for a line and a 9D space for a triangle. After mapping this object in our n-dimensional space we transform that space removing position, size and rotation and in the process eliminating several dimensions. What you are left with is shape space.

TPSDIG

The first program we will use today is Jim Rohlf's *TPSDIG*. This is only one of several programs Rohlf has developed, but it is critical for this lab, because it allows us to

enter the data we will be analyzing. We use this program to identify the landmarks on our specimens. Landmarks should be homologous points on our organisms that are consistent, repeatable, and coplanar (i.e., they are all visible in a photograph of a specimen). There is no right number of landmarks, but they should be numerous enough to adequately cover and describe a specimen's shape. The details of what a landmark is and how to choose them is a contentious issue that we will avoid for now.

TPSDIG allows us to define landmarks and save them in a format that the other programs can recognize. To use *TPSDIG*, we need to have photos of all of our specimens in the same view, and it is best if all of the photos have a scale. To save time today, we'll just do two specimens.

First let's take a look at the files that this program uses. Open the folder c:/Documents and Settings>All Users>Start Menu>Programs>IB200A> TPS>Examples. This folder contains three files. Two of them are TIF files with the images of fly wings that we will be analyzing. The third is a TPS file called **test**. Open **test** is a word processor. This file points to the two images, which we will use in our analysis, so that they can be analyzed at the same time. As you can see, both of the pictures in this folder are listed as images.

Open **TPSDIG** (Start – Programs) and go to the file menu. Go to **input source** and choose tps file 'test' in c:/Documents and Settings>All Users>Start Menu>Programs>IB200A> TPS>Examples. Both images are from the left wing of female *Aedes canadenses*.

The image file should appear on your screen. Go to the tool bar and select the **bulls'-eye tool**. You can zoom in on the images using the **+ button**. Each time you click with the **bull's-eye** a red dot appears representing a landmark. Thus, click on places you want to include as landmarks. Remember that landmarks need to be homologous, so pick points that you can clearly identify in both images. In an ideal situation we would have a scale bar in the image, which you could use to define a standard distance by switching to set scale mode, clicking on the endpoints and entering the length, but we don't. If you make a mistake, click on the pointer to switch to edit mode and move your landmarks around. Select '**Label landmarks'** from the options menu to see in what order you made your marks.

For the next image click the **big red arrow**. Now click on the homologous landmarks in this image in the same order. If you forget where you put a landmark you can use the big red arrows to toggle back and forth between the two images.

When you are done, go to the file menu and select save data as. Save your document to the desktop. Type '.tps' at the end of the file name (For some reason the program can't handle this itself). You can open your file in notepad. If you do, you will see it contains the landmark coordinates for your specimens in raw form.

The rest of the programs we'll be using today were developed by David Sheets at Canisius College. They are already on our computers, but they can downloaded at: www2.canisius.edu/~sheets/morphsoft.html. These are probably the most straightforward and easy to use programs available for morphometric analyses.

CoordGen

The first thing we'll need to do before analyzing our data is match all of our specimens up in shape space. In other words, we need to remove differences in location size and orientation from our data. The methods that allow us to do this are called superimposition methods because we are essentially superimposing all of our data on top of each other. CoordGen can use several superimposition methods, although we will focus on two, two-point registration ('Bookstein coordinates) and procrustes.

Two-point registration is the easier of the two methods to understand. Two points are chosen and a line is drawn between them, making a baseline. All the images are then moved, rotated and changed in size, so that their baselines line up exactly. As a consequence you remove 4 dimensions from our original sets of coordinates, two for each of those points.

Open *CoordGen6d* (the application file!). The path is C:\imp\bin\win32. Keep this window open because we'll return to it for other programs and files. When open CoordGen6f, two windows will appear. The one you want to interact with should be obvious.

In the **tangerine** colored box, push the button '**load tps file (no ruler/ no scale factor).'** Select the file you saved from TPSDIG. If you had a scale you could use the light blue boxes to set it, but we don't. A cloud of points will appear on the graph. This shows the position of each of the landmarks for each specimen as they have been superimposed. You are currently looking at a Bookstein coordinate superimposition.

The problem with two-point registration is that it takes all the variance from the two points that make up your baseline and redistributes it over the other points, because it holds those two points as stable all the change in position happens at the other points. This is not realistic, as change really happens throughout an organism, and is not limited to changing relative to two arbitrary points.

Procrustes deals with this problem by rescaling everything around the centroid of the object. The centroid is a point in the middle of the object calculated as the average of all the other points. First all the objects are moved, so that the centroid is located at the origin, thus eliminating location and two coordinates along with it. Next all the objects are rescaled so that their centroid size is one, eliminating size and one more dimension. Finally the objects are rotated, so that the **procrustes distances** between them are minimized, eliminating orientation and one more dimension. The **procrustes distance** is basically the distance between the objects in the new space defined by the first two transformations. I want to emphasize that the point of doing all this is to remove all the factors other than shape, not to reduce the number of dimensions. The dimensions are reduced as a byproduct of these transformations.

Go to the **green box** and toggle between the **show bc** button and **show procrustes** buttons. What differences do you see? Do you know why? Can you identify where your baseline is in the bc?

We could save the results of these superimpositions, in the **lavender box** you can see the buttons used for saving the results. The saved results files could be read by the other programs we'll be using. However, there is no need, because we'll be using other files that have many more samples for the remaining exercises. Our current example dataset with only two mosquito wings is not sufficient for the following statistical packages.

PCAGen

That's all very nice, but there's a problem with the coordinates that come from a procrustes superimposition. When you eliminated size, you eliminated one dimension, but not in the sense that you would normally think of eliminating a dimension. Imagine a three dimensional space. Normally if you eliminated a dimension you would be compressing all the points down to a two-dimensional plane creating a two dimensional graph, that we are all so familiar with. In the procrustes superimposition you don't force all the points onto a plane, but instead onto the surface of a sphere (or technically an n-dimensional hypersphere). The surface of a sphere has two dimensions, like a plane, but the coordinates are not distributed along a rectangular coordinate system, and so conventional statistics do not work. In particular degrees of freedom become very difficult to calculate.

There is a way around this called the **thin-plate spline**. This relies on one final transformation of shape space, but one that does not eliminate dimensions instead it flattens out the shape space creating a new set of coordinates called **partial warp scores**. The basic idea is to treat changes in shape as deformations of an infinitely thin metallic plate. By bending the plate points are moved closer to or further away from each other. The partial warp score is then the bending energy required to generate a deformation from an average specimen. This transformation has two advantages. First the **partial warp scores** can be used with conventional statistical methods. Second, the **thin plate splines** can be visualized as changes in the shape of the organism at a local scale (e.g., enlargement of the head) involving multiple landmarks all moving together.

The next program we'll be using is called PCAGen. This program starts with superimposition data and then computes partial warp scores based on the loaded data. The program then conducts a principle components analysis of the covariance matrix derived from the partial warp scores. The program can display landmark positions for all of the specimens, a plot of the data along different principles component axes, and the deformation implied by a particular principle component vector or thin plate spline.

Open PCAGen6n (c:/Documents and Settings>All Users>Start Menu>Programs> IB200A> IMP). Again, two windows will open and the one you want will be obvious. Push the button labeled '**load file'** and load in the **threepir.bd** file (in the same path as above). Then push the '**no group list'** button. When you do this, a plot of all of our landmark data will appear. These landmarks represent a view of a bunch of fish bodies if you're interested in what you're looking at.

If you push the '**show pca plot'** button in the **purple blox**, a plot of the specimens on pc axes 1 and 2 will be displayed. This gives you an idea of which specimens are similar or different in regards to these axes. You can use the **up** and **down** buttons to examine the other PCs. Next to these buttons is a box showing what percentage of the difference in shape between specimens each axis explains.

Now go to the blue box labeled '**deformation display format'** and select '**vectors on landmarks**.' Then push the '**display pc deformation (bc)**' button in the **grey box**. When you do this, a plot showing you landmarks will appear. The arrows represent the direction and relative amount of the shape change explained by this PC vector. This information is extremely useful as it allows us to see the coordinated

changes in the body of the fish that explain the plurality of change in the shape of the fish body. Note that a PC is an axis not a direction, so that these changes could go either way. For PC 1, we can see that most of the shape change it describes is deepening of the mid-section of the body. You have to use your imagination to visualize the orientation of a fish in this cloud of points and vectors. Examine the deformations for some of the other pc vectors and see what shape changes they describe. You can also view the thin plate splines by clicking the **show def (procrustes)** button.

Two other useful features can be found in the '**statistics**' menu. The '**screen plot (percentages)**' feature shows graphically what percentage of the variation each pc explains. The rule of thumb is to assume meaning for the pc's that occur before the plot flattens out. The '**significant differences in pc components'** feature does pair-wise comparisons for all of the pc's and then reports how many are significantly different. In this example, the first pc is significantly different from the rest, but none of the others are significantly different from each other.

You could save the PCs or the partial warp scores calculated by this program for use with other statistical software.

CVAGen

The final program we'll use today is called CVAGen. The program performs a canonical variates analysis, which allows us to find a set of axes which best discriminate 2 or more groups that may be in our data. For example, this program would be useful if you wanted to test whether a series of named species can be discriminated using geometric morphometrics.

Open CVAGen6j. This time three windows will open. Push the red '**load file'** box and load the file called '**threepir**'. Then go to the '**load group membership list'** and load the file '**threepirgrp**'. When you are done, a plot showing the landmark data will appear, along with a small dialogue box that will tell you that there were two distinct (i.e., significantly different) canonical variates found.

Now push the purple '**show cva plot'** button. This will show you where your different specimens fall out on the canonical variates axes. As you can see, our data forms three very distinct groups. Go to the '**deformation display format'** and again select '**vectors on landmarks**.' Then push the '**display cva deformation (bc)'** button in the **grey box**. When you do this, a plot will appear that will show you the shape change that the selected cv describes.

Another useful feature is the '**show groupings by cva'** function found in the **statistics** menu. Select this function and go to the '**auxiliary results box'** (the smaller colorful box that opened when you started the program). In this box you will find a table that shows how many specimens of your original group fell out in each of the groups determined by the cva analysis. In this case, none of our specimens went into any of the 'wrong' groups. However, this is certainly a possibility.