How to Structure and Name Data Files

Overview:

This is a set of suggestions for how to structure data files and name those files in a way that will allow maximum flexibility for interfacing with other programs, avoid errors, etc. The emphasis is on Microsoft Excel, but most of the principles are general.

Feel free to email me with suggested changes or additions. Thanks to Doug Altshuler for helping to get this together.

Format:

- Excel files should be saved in .xls format rather than .xlsx.
- File (document) names should not have spaces, parentheses, etc. Underscores and dashes are OK.
- As you work on the data file and code file(s), save new versions frequently that have a higher number appended at the end. For example: "Hummingbird_Data_1.xls"
    - This way, you should be able to sort by name within the folder and always easily find the newest version.
    - Version numbers are better than appending dates to the names because dates can be written in many different formats. Also, you may often edit a file more than once on a given day.
- Column headers must have eight or fewer characters, with no spaces, slashes, %, etc. Stick with letters and numbers. Underscores and dashes are OK.
    - It is best to use fewer than 8 characters so that, for example, if you later create a log-transformed version of a given column then you can just add an "L" to the front of the name.
- All column names should be defined. One method is to have an additional sheet in your Excel file with those definitions. Another method is to use the “insert comment” function to have notes associated with each column header.
- Do not include semi-redundant columns, such as log transforms. These are better done once you have your data into an analysis program. You do not want to garbage-up your original data file.
- If data are coming from a lab or field notebook, then it is good practice to include a column that indicates the page number from which the entry comes.
- Enter dates as three separate columns for year, month, day.
- Enter times as separate columns for hours since midnight (military time) and minutes.
- Code sex as 0 for females and 1 for males. That is anatomically correct.

Additional formatting requirements for phylogenetic analyses:

- For ease of entry to phylogenetic analyses, the first six columns must be in the

following sequence:

1. Two-character species code.
    - Note that codes of 01, 02, etc. are problematic because some programs (including Excel) do not recognize the leading zero.
    - Also, some programs recognize upper and lower case letters as different, but others do not, so do not rely on that difference.
2. An underscore, as in "_"
3. Genus
4. Another underscore
5. Species
6. Another underscore
7. Subspecies, named as "Subspeci" to keep within 8 characters

- The eighth column might have phylogeny position, perhaps named as "PhylPosi". For the legend or inserted comment, make note of which phylogeny (i.e., the filename and date created) these numbers come from.

Some general points to keep in mind:

- Always proofread carefully. Two people are suggested, one to read from the computer data file and the other to check in the original data source, e.g., lab notebook. Do not proofread the other way around or error rate increases.
- If possible, proofread twice.
- Make notes in the file indicating who did the proofreading and when.
- Make sure it is clear if the problems noted were corrected. This is efficiently done by inserting a comment into individual cells that have been corrected.
- Many journals are now requesting that raw data be published, usually as online supplemental material, but also sometimes in an online repository such as Dryad.
    - Thus, the raw data file must be kept in good order all the way through. It will become "public" at some point and you do not want to embarrass yourself or your collaborators. Nor do you want errors.
- In addition to not calculating semi-redundant columns, such as logs, do not use pivot table functions in Excel to calculate species means and standard errors.
- Instead, use code in R or SPSS to calculate log-transformed data, and the means and standard errors from both the raw and the log-transformed data.
    - Make sure to save the code (syntax) that you use for doing this, and to insert comments liberally to indicate what you are doing and why.
    - In SPSS, the AGGREGATE function is used to compute means and so forth, and write them to a new file.
    - It is also good practice to publish the code along with the raw data.
- The goal is to require only two files to recreate all of your analyses:
    1. The "golden master" Excel file, which contains only data without calculations.
    2. A code or syntax file (e.g., for R or SPSS) that does manipulations of that file, including creating other files in specific formats required for various analysis programs (e.g., the Matlab Regresisonv2.m program of Lavin et al. 2008)

If you have data on multiple individuals within multiple species or strains:

- Calculate the means and standard errors for the male and female data separately.
    - Those means can then be averaged to obtain a mean for a species or strain that is not influenced by different sample sizes in each sex. Standard errors are more complicated.
    - It is also good practice to publish the code (e.g., for R or SPSS or SAS) along with the raw data. Just make sure it includes plenty of clear comments.

12 March 2013

Theodore Garland, Jr.
Professor
Department of Biology
University of California, Riverside
Riverside, CA 92521
U.S.A.

Office Phone: (951) 827-3524 [2366 Spieth Hall]
Facsimile: (951) 827-4286 = Dept. office (not confidential)
Email: tgarland@ucr.edu

Main Departmental page:
http://www.biology.ucr.edu/people/faculty/Garland.html
Lab page:
http://www.biology.ucr.edu/people/faculty/Garland/Garland2.html
List of all Publications:
http://www.biology.ucr.edu/people/faculty/Garland/GarlandPublications.html
Google Scholar Citations Profile:
http://scholar.google.com/citations?user=iSSbrhwAAAAJ