# Data import

## Type of morphological data

Categorical morphological data (discrete characters) should be treated as factors when imported to calculate character distances, as the symbols used to represent different states are arbitrary (e.g., could be equally represented by letters, such as for DNA data). If continuous variables are used as phylogenetic characters, those should be read in from a separate file and treated as numeric data, since input values for each state (e.g., 0.234; 2.456; 3.567; etc) represent true distance between data points.

## Treatment of inapplicable and missing data

Categorical data including symbols for inapplicable and missing data (typically “-“ and “?”, respectively) will be read in and treated as separate categories of data relative to numerical symbols for different character states (“0”, “1”, “2”, etc.). Two options are available: option 1—converting inapplicables/missing to “NA” (as done by default in the package Claddis) or option 2—keeping the original symbols.

In the example provided below, option 1 will convert unknown conditions to “NA” and thus will ignore the respective taxa with inapplicable/missing data to calculate inter-character distances. The resulting distance matrix will introduce “NaN” to every pairwise comparison involving character two characters with “NA” (all comparisons including character five, as well as any pairwise comparisons involving characters 4, 5 and 7) (Table 2-in blue). Statistical tests and clustering methods cannot utilize such matrices with “NaN” as data entries, and removal of observations contributing to excessive “NaN” have to be performed— such as taxon removal during inter-taxon distance matrix construction using Claddis.

Besides, in comparisons between characters inclusive of states with “NA”, the latter will contribute 0 difference to the distance matrix. For instance, distance between characters 6 (0,0) and 7 (“NA”, 0) and in the table below is 0 (Table 2-in red). The implicit assumption with option 1 is that unknown characters contribute 0 distance. Therefore, this approach biases the distance matrix by minimizing the overall distance between characters to the lowest possible values. It assumes that, whatever the true condition represented by the unknown state, it is always assumed to be equal to the known character states (e.g., character states scored as “1” for Taxa A and B).

With option 2, inapplicables/missing data will be treated as a distinct categorical variable relative to numeric symbols. As a result, pairwise comparisons with characters with unknown data will avoid the introduction of “NaN”, allowing all characters to be considered (Table 3-in blue). However, by considering the unknown states (“-“ and “?”) as always distinct from all character states with information provided (“0”, “1”, “2”, etc.), this approach will bias the distance matrix by maximizing the overall distance between characters to the maximum possible values— the opposite pattern to option 1. As a result, in contrast to option 1 that introduces a distance of 0 between characters six and seven, option 2 will introduce a distance of 0.5 (Table 3-in red).

Contrary to approaches to create inter-taxon distance matrices to estimate morphospace, removing observations with excessive inapplicable/missing data for the final distance matrix is not possible because each character must be assigned to at least one partition (regardless of the amount of missing or inapplicable data). Therefore, here we recommend following the approach in option 2, which maximizes the distance between characters with inapplicables/missing data and avoids inapplicable outputs in the distance matrix. However, as characters with excessive missing data are likely to bias the results of phylogenetic analyses (besides biasing the estimation of the distance matrix), we suggest avoiding or removing such characters for morphological phylogenetic analysis. This will avoid the several phylogenetic analytical issues for characters with excessive inapplicable/missing data, besides the biases introduced in construction of distance matrices and the assessment of character partitioning.

Table 1. Example dataset.

|  |  |  |
| --- | --- | --- |
|  | Taxon A | Taxon B |
| Character 1 | 0 | 0 |
| Character 2 | 1 | 1 |
| Character 3 | 0 | 0 |
| Character 4 | 0 | ? |
| Character 5 | ? | ? |
| Character 6 | 1 | 1 |
| Character 7 | ? | 1 |
| Character 8 | 0 | 0 |
| Character 9 | 1 | 1 |
| Character 10 | 1 | 1 |

Table 2. Distance matrix (option 1):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Char1 | Char2 | Char3 | Char4 | Char5 | Char6 | Char7 | Char8 | Char9 | Char10 |
| Char1 | 0 | 1 | 0 | 0 | NA | 1 | 1 | 0 | 1 | 1 |
| Char2 | 1 | 0 | 1 | 1 | NA | 0 | 0 | 1 | 0 | 0 |
| Char3 | 0 | 1 | 0 | 0 | NA | 1 | 1 | 0 | 1 | 1 |
| Char4 | 0 | 1 | 0 | 0 | NA | 1 | NA | 0 | 1 | 1 |
| Char5 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| Char6 | 1 | 0 | 1 | 1 | NA | 0 | 0 | 1 | 0 | 0 |
| Char7 | 1 | 0 | 1 | NA | NA | 0 | 0 | 1 | 0 | 0 |
| Char8 | 0 | 1 | 0 | 0 | NA | 1 | 1 | 0 | 1 | 1 |
| Char9 | 1 | 0 | 1 | 1 | NA | 0 | 0 | 1 | 0 | 0 |
| Char10 | 1 | 0 | 1 | 1 | NA | 0 | 0 | 1 | 0 | 0 |

Table 3. Distance matrix (option 2):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Char1 | Char2 | Char3 | Char4 | Char5 | Char6 | Char7 | Char8 | Char9 | Char10 |
| Char1 | 0 | 1 | 0 | 0.5 | 1 | 1 | 1 | 0 | 1 | 1 |
| Char2 | 1 | 0 | 1 | 1 | 1 | 0 | 0.5 | 1 | 0 | 0 |
| Char3 | 0 | 1 | 0 | 0.5 | 1 | 1 | 1 | 0 | 1 | 1 |
| Char4 | 0.5 | 1 | 0.5 | 0 | 0.5 | 1 | 1 | 0.5 | 1 | 1 |
| Char5 | 1 | 1 | 1 | 0.5 | 0 | 1 | 0.5 | 1 | 1 | 1 |
| Char6 | 1 | 0 | 1 | 1 | 1 | 0 | 0.5 | 1 | 0 | 0 |
| Char7 | 1 | 0.5 | 1 | 1 | 0.5 | 0.5 | 0 | 1 | 0.5 | 0.5 |
| Char8 | 0 | 1 | 0 | 0.5 | 1 | 1 | 1 | 0 | 1 | 1 |
| Char9 | 1 | 0 | 1 | 1 | 1 | 0 | 0.5 | 1 | 0 | 0 |
| Char10 | 1 | 0 | 1 | 1 | 1 | 0 | 0.5 | 1 | 0 | 0 |