# 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 |

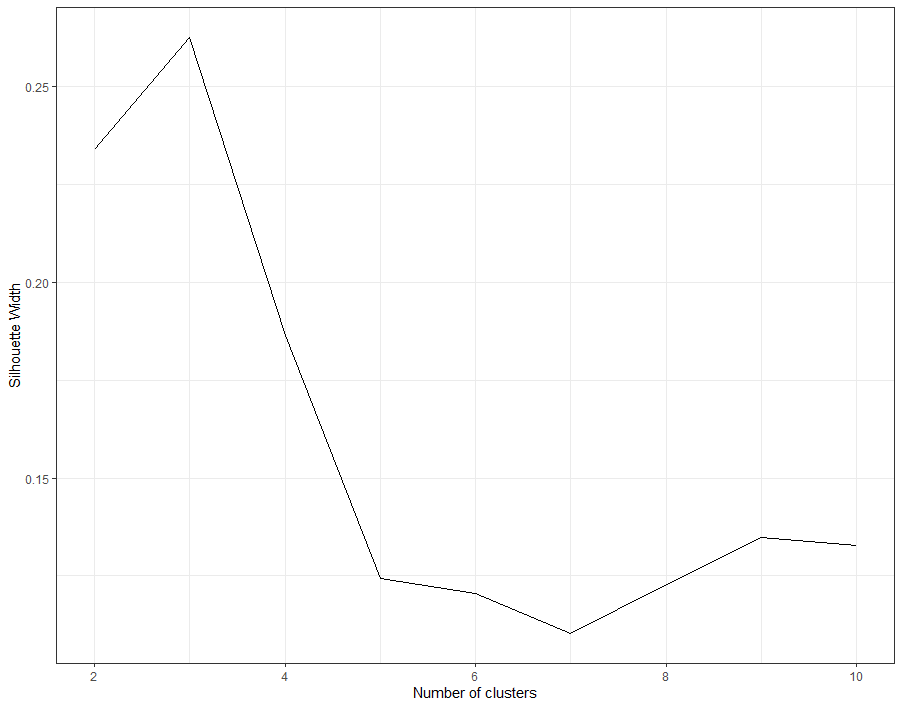
# Cluster identification

## Using PAM and silhouettes width index (Si)

Euclidean distances are extremely sensitive to missing data, and alternative choices such as Gower distances provide more suitable alternatives for the handling of missing data(Lloyd, 2016; Lehmann et al., 2019). This issue creates a subsequent problem for estimating clusters using K-means, perhaps the most popular clustering approach, as it depends on an Euclidean-based distance matrix. Further, K-means are based on measuring the distance between samples and cluster centroids (i.e., the center of mass or mean vector of the cluster). The mean vector is particularly sensitive to outliers (as any other mean estimate) (Rencher and Christensen, 2012), which can be particularly problematic for small-sized clusters or clusters of drastically different sizes, which are to be expected from most standard sized morphological datasets. Those limitation from Euclidean distances and K-means thus limit the quality of the clustering analysis based on such approaches.

Therefore, this package uses Gower distances to create the inter-character distance matrix “D”, following the original description of this approach in Simões and Pierce (2021). For the clustering analysis, we utilize here partitioning around medoids (PAM, or K-medoids), which can estimate clusters using Gower distances. PAM is analogous to K-means, but it has its clusters centered around medoids instead of centered around centroids, which are less prone to the impact from outliers and heterogeneous cluster sizes (Rencher and Christensen, 2012; Budiaji and Leisch, 2019).

To define how many clusters the data could be partitioned into, various PAM partitioning schemes are tested and the quality of each clustering scheme using the silhouette index (Si) approach (Rousseeuw, 1987), which determines how well an object falls within their cluster compared to other clusters. PAM partitioning schemes tested range from the minimum number of possible partitions (K=2) to a large number of partitions (user-defined , default K = 10).



**Fig. X.** Si plot indicating the higher quality of clustering when the number of partitions (k) = 3.

## t-SNEs

As an additional and independent test of the quality of the chosen partitioning schemes, we also provide a graphic visualization approach based on t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten and Hinton, 2008). This has become a popular ordination and visualization tool in machine learning given its ability to reduce exceptionally large number of dimensions into only two or three dimensions. More traditional ordination procedures, such as principal components analysis (PCA, for continues data) or principal coordinate analysis (PCoA, for discrete data), can preserve the linear relationship between data points at a lower dimensionality. However, as those metrics try to preserve the local distances between data points they become less efficient on characterizing the overall structure of high dimensional data—it is more important to reduce the local linear distance between similar (neighboring) data points while maximizing the distance between distant datapoints (Van Der Maaten and Hinton, 2008). For such cases, nonlinear ordination procedures are preferred to observe the overall data structure in a reduced number of dimensions. t-SNE has been demonstrated to be more efficient on preserving both local and global structures when reducing high dimensional data into only two or three dimensions compared to other nonlinear ordination procedures (Van Der Maaten, 2009), thus offering an important advantage over previously utilized graphic approaches to determine morphological clusters such as PCoA.

## The best candidate clustering scheme

The resulting clustering scheme obtained from PAM+Si is directly mapped on top of the graphics produced by t-SNEs (Fig. X) in order to test the congruence of cluster size and composition between both approaches (Simões and Pierce, 2021). If major discrepancies occur, it is suggested that other values of K considered suboptimal by Si may be attempted and compared against t-SNEs. As an additional alternative, the group limitation for t-SNEs can be based on anatomical subdivisions instead of the clustering scheme obtained by PAM+Si. If there is a closer correspondence between tSNEs and anatomical partitioning as compared to PAM+Si and tSNEs, it might be more reasonable to follow anatomical partitioning.

Chart, scatter chart

Description automatically generated

**Fig. X.** tSNE plot of the first two dimensions with data points colored according to the partitioning scheme determined by PAM+Si.

# Strength of selection (using morphological clock rates)

The strength of natural selection operating upon particular regions of the phenotype (e.g., phenotypic partitions) across distinct clades in a phylogeny can be indirectly measured by comparing the variation of evolutionary rates relative to the background evolutionary rates. When rates are significantly accelerated, it provides support for positive phenotypic selection in analogy with the / ratio in molecular evolution, whereas strongly decelerating rates represent an instance of stabilizing selection, stasis or constraint (Yang, 2014; Baker et al., 2016). This concept first applied to phenotypic traits using continuous data in phylogenetic comparative methods by Baker et al. (2016) was extended to discrete data and evolutionary rates estimated with Bayesian molecular or morphological clocks (Simões and Pierce, 2021).

The clock rate on every tree branch (Δv) is compared to the background rate of evolution (Δb), forming the rate scalar ratio (*r* = Δv/Δb), as defined by Baker et al. (2016). This measure is equivalent to the interpretation of relative rates of character evolution produced by relaxed Bayesian clocks, in which estimates > 1 indicate rates above background rate levels (the base of the clock rate) and are therefore accelerating. In contrast, relative branch rate values < 1 indicate values below background rate levels, implying a decrease in the rates of evolution in that branch (Ronquist et al., 2019).

In its original implementation, when evolutionary rates are at least twice as fast as the background rates (*r* >2) that would be interpreted as a positive phenotypic selection(Baker et al., 2016). In our implementation here we utilize a flexible threshold, taking into account the dispersion of the distribution for the background rates. We specifically test for branch rate values that are statistically significantly different from the mean background rate value, thus taking into account the standard deviation (sd) of the background rate value (obtained from the Bayesian posterior). Simões and Pierce (2021) established 1sd from the background mean rate as the threshold: when the mean rate of evolution on a given branch is greater than the background rate plus one standard deviation (Δv > it is interpreted as an instance of positive selection. When the mean rate of evolution on a branch is less than the main background rate minus one standard deviation (Δv > it is interpreted as an instance of stabilizing selection or stasis.

Here we expand on that by enabling users to compute multiple threshold levels across the tree: using the 95% confidence interval (CI), or 1, 2, or more standard deviations. Users can plot only one of those thresholds or all of those thresholds combined onto the evolutionary tree to assess the degree upon which particular clades are evolving faster or slower compared to background rates, with direct implications to the strength of selection operating upon the morphological traits used to calculate the respective rate values.

# Statistical analysis available

Users can assess the normality of the distribution for each time bin using the Shapiro-Wilk normality test and visual assessment of data distribution. Additionally, users can use the Bartlett test of homogeneity of variances to assess homoscedasticity in the data.

For statistical hypothesis testing, we provide fast outputs of parametric (pairwise t-tests) and nonparametric (pairwise Wilcoxon rank sum (Mann-Whitney) tests) between phenotypic partitions, time bins, or specific kind of analyses.

# References

Baker, J., Meade, A., Pagel, M., Venditti, C. 2016. Positive phenotypic selection inferred from phylogenies. Biol. J. Linn. Soc. 118, 95-115.

Budiaji, W., Leisch, F. 2019. Simple K-Medoids Partitioning Algorithm for Mixed Variable Data. Algorithms 12, 177.

Lehmann, O. E. R., Ezcurra, M. D., Butler, R. J., Lloyd, G. T. 2019. Biases with the Generalized Euclidean Distance measure in disparity analyses with high levels of missing data. Palaeontology 62, 837-849.

Lloyd, G. T. 2016. Estimating morphological diversity and tempo with discrete character-taxon matrices: implementation, challenges, progress, and future directions. Biol. J. Linn. Soc. 118, 131-151.

Rencher, A. C., Christensen, W. F. 2012. Methods of Multivariate Analysis. John Wiley & Sons, Hoboken.

Ronquist, F., Huelsenbeck, J., Teslenko, M., Nylander, J. A. A. 2019. MrBayes version 3.2 manual: tutorials and model summaries. <https://nbisweden.github.io/MrBayes/manual.html>.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comp. Appl. Math. 20, 53-65.

Simões, T. R., Pierce, S. E. 2021. Sustained High Rates of Morphological Evolution During the Rise of Tetrapods. Nat. Ecol. Evol. 5, 1403–1414.

Van Der Maaten, L. 2009. Learning a parametric embedding by preserving local structure. Artificial Intelligence and Statistics, pp. 384-391.

Van Der Maaten, L., Hinton, G. 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579-2605.

Yang, Z. 2014. Molecular evolution: a statistical approach. Oxford University Press.