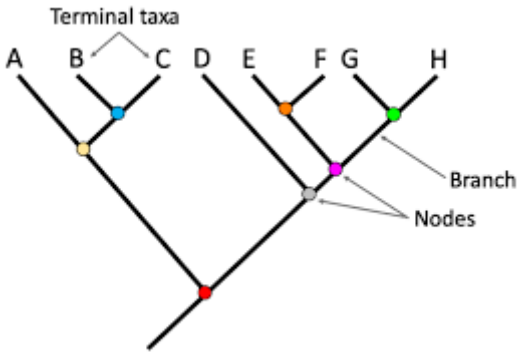
Standard and pivot approach for analyzing compositional data

A set of variables, x, is called composition when the sum of these variables for each observation is constant. Each variable in the composition x is called a part. There are D parts in each composition.

Since, the compositions don't live in Euclidean space, people believe that the usual statistical methods are not applicable to the compositional data. Therefore we have to transform them before using them in statistical analysis. Several transformation have been developed. However the most reliable one is isometric log ratio transformation, ILR. Once you have ILR variable, you can apply the standard statistical tools.

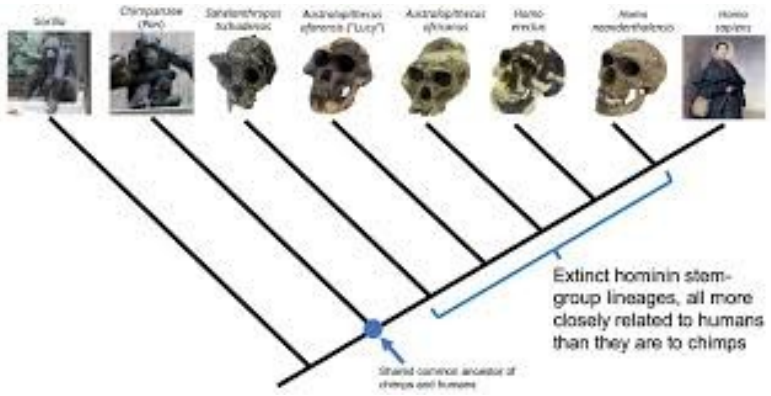
At the heart of ILR transformation, lies a serial binary partition (SBP). Considering the parts in x as biological species, SBP is binary evolutionary tree. There are multiple way to construct trees, all of them are statistically equivalent but not biologically.



This image shows one possible of SBP where A-H are parts (original variables) in x. Once you decided for the SBP, you encode the tree in a binary matrix in R and convert it to ILR variables using the function BASIS (defined in the R script). From now on, everything is as normal i.e. apply the standard statistical tools on ILR variables.

SBP construction could be supervised (guided) or unsupervised. For supervised SBP construction, we use CLR coefficients and also our knowledge from the field to carefully construct a SBP which better answers our research questions. CLR coefficients are measures of influence of each part on the response variable.

A shortcut for SBP construction is ignoring all guides and using the following tree:



This is called pivot ILR.