# SWeeP and machine learning in supertree construction: family Formicidae analysis

Monique Schreiner, Roberto Tadeu Raittz, Mariane Golçalves Kulik

*Universidade Federal do Paraná*

**Abstract**

Phylogenies including all taxa of a large group, the supertrees, are essential for research in macroevolution, biogeography and conservation. In groups with lack of data, the construction of super trees becomes complex. Particularly, the ants (family Formicidae) compose a group with genetic data distribution quite heterogenous wherein few species have thousands of sequences registered on biological databases while most of the species have few or none sequences registered, fact that make a global phylogenetic analysis difficult. The present research proposes the construction of a supertree for all extant ants, including the ones missing genetic data available, using vectorial techniques and artificial intelligence. To perform this, all ant protein sequences available on NCBI were downloaded and clustered. For prospection, the largest cluster were analyzed. The sequences of the largest cluster were vectorized using the algorithm SWeeP which transforms sequences of amino acids in compact vectors preserving comparability of the sequences and allowing big data analysis. One matrix per protein was generated for the most frequent proteins on the cluster. A Principal Component Analysis (PCA) was performed on the SWeeP matrixes. The result showed that it was possible to distinguish the main subfamilies of ants using a few of the principal components. Multilayer Perceptron Neural Networks (MLP) were trained to classify the subfamilies using the first one hundred principal components of the matrixes. The mean accuracy of the neural networks was 0.9576 (SD=0.064). The neural network with the best performance was for the protein Cytochrome C Subunit I with 0.994 accuracy. The preliminary results showed that it is possible to classify organisms using molecular data without the use of alignment techniques. Alignment techniques are computationally expensive and hence limit the amount of data that can be used. The classification learning model corroborates the potential of the proposed vectoral model. The learning model will be applied in the prediction of missing elements and will allow the use of multiple proteins. At last, the developed model will be used in the construction of the Formicidae supertree.

Funding: CAPES