PD Dr. Markus Göker
DSMZ – German collection of microorganisms and cell cultures
Inhoffenstraße 7B
38124 Braunschweig
www.dsmz.de

Defining biologically meaningful molecular operational taxonomic units

A reliable taxonomy is crucial for the assessment of biodiversity and for the comparison of habitats based on their species composition. Determining taxon boundaries is challenging in the case of organisms for which often only molecular data are available, such as bacteria, fungi, and many unicellular eukaryotes. Even in the case of organisms with well-established microscopical characteristics, molecular taxonomy is necessary to determine misidentified and mislabelled GenBank sequences, to identify incompletely known specimens and cryptic species, and last but not least to analyse sequences directly sampled from the environment as in metagenomics studies.

As taxon boundaries cannot directly be inferred from a phylogenetic tree, researchers mostly apply clustering algorithms in combination with predefined thresholds to pair-wise genetic distances in order to assign sequences to molecular operational taxonomic units. However, the chosen thresholds differ in the literature, even if applied to the same organisms and molecular markers, and are often based on subjective criteria or just on tradition. Furthermore, the clustering algorithm also may have a significant impact on the outcome, but it is seldom addressed which algorithm is most appropriate for molecular taxonomy. Further potential sources of biases are alignment ambiguity and rate heterogeneity between sites.

We here introduce a simple yet effective and flexible clustering optimization method that addresses all these issues. Using biologically sensible reference partitions, our method automatically distinguishes between within-taxon and between-taxon sequence heterogeneity in the course of identifying optimal clustering settings. Usage examples for clustering optimization with alternative types of biological data are provided, including approaches for calculating the degree of uncertainty in estimating the optimal parameters. The novel algorithm is discussed as a tool for the automated self-correction of GenBank data and as general method for molecular taxonomy that results in taxonomic units which optimally account for both traditional species concepts and genetic divergence.