Multivariate Data Analysis in Microbial Ecology - New Skin for the old Ceremony

Jean Thioulouse

The molecular biology revolution has a particularly strong impact in microbial ecology, as molecular methods are now giving access to data that were previously impossible to obtain. Soil microbial ecology studies are a good example of this situation. Knowledge of soil bacterial diversity is of great interest, both from an applied agronomical perspective, and in the framework of theoretical ecological models like the species-area relationship. Previously, only culturable species could be studied, which represented an extremely low part of the total bacterial community. Today, tools like DNA fingerprints, DNA microarrays, and transcriptomic methods can be used directly on DNA extracts from bulk soil samples, providing new insights into the diversity and functioning of bacterial soil communities.

However, while molecular tools have been rapidly apropriated by microbiologists, the statistical methods needed to analyse the resulting huge amounts of numerical information still represent an obstacle for many microbial ecologists. The R environment already plays an important role in genomic data analysis thanks to the bioconductor project, but the statistical methods needed for multivariate ecological data analysis are part of standard R packages, like vegan and ade4. Although these packages are designed for plant or animal ecology, many of their functions can be used to analyse molecular biology data sets. Furthermore, other packages, like seqinr and made4 are very useful to bridge standard packages and genomic data structures. Lastly, graphical user interfaces are also needed to help biologists master the intricacies of some R functions.