DEVELOPMENT OF A TAXON MIXING MODEL TO INFER SOURCES AND MIXING OF MICROBIAL TAXA FROM MOLECULAR SEQUENCE DATASETS.

Christopher Thornton[1] and William J. Brazelton[1]
[1]christopher.thornton@utah.edu, 257 South 1400 East, Department of Biology, University of Utah, Salt Lake City, UT 84103

Any biological investigation into subsurface habitats requires some effort to distinguish between *bona fide* subsurface organisms and contaminant organisms that are introduced from the surface or during sample handling. Currently, no quantitative tools are available to detect and correct for contamination in large DNA sequence datasets. We suggest that all kinds of 'contamination' can be evaluated against a null model of physical mixing of parent community compositions to produce a daughter community composition. Moreover, this mixing should be evident from observed patterns of relative abundances of taxa across samples. Therefore, from a dataset of environmental DNA sequences, one can detect and remove contaminants and thereby infer the hypothetical 'true' community composition of all environmental samples from the system. This approach relies on the same logic employed by geochemical fluid mixing models.

We are developing open-source software that will automate these inferences from environmental distributions of 16S rRNA deep-sequencing datasets. This software will test all possible pairwise relationships between taxa and then categorize each taxon according to its inferred source. The community compositions of all inferred parent sources will be reported, and any taxa that are not clearly derived from physical mixing of the parent sources will be highlighted as potential consequences of biogeochemical interactions (e.g. growth at the mixing front). Some taxa may be derived from physical mixing in some samples and exhibit growth in other samples, and these inevitable complexities are the motivation for our development of software that can automate the statistical analyses and iteratively test potential models.

Eventually, this modeling approach will be extended to include biological and biogeochemical interactions as potential explanations for observed deviations from simple mixing ratios. Applying the same principles as geochemical fluid mixing models, we should be able to not only infer the composition of two parent sources but also to predict the community composition of a mixed fluid that would result from the biogeochemical processes occurring during the formation of the mixed fluid. Clearly, achieving this goal will depend on obtaining much greater insight into the biogeochemical processes of subsurface environments, which is the long-term goal of the research program. In general, this research direction aims to harness large DNA sequence datasets in a rigorous manner in order to move beyond mere stamp-collecting and to link biological data with models of biogeochemical activity.

Furthermore, the ultimate goal of this research direction to detect contaminating organisms and to infer the true source of a given organism has clear applications for astrobiology and life detection efforts.