Modeling soil microbial communities using artificial neural networks

Marcio R. Lambais
Department of Soil Science
University of São Paulo
Piracicaba, Brazil
Outline

✓ Modeling soil microbial communities based on FAME profiles
✓ Modeling soil microbial communities based on 16S rRNA gene sequencing
✓ Potential use of ANN for modeling phytobiomes
Atlantic Forest

- Biodiversity hotspot (UNESCO Biosphere Reserve)
- Estimated number of plant species is more than 20,000
- Half of the plant species are endemic
- Covered approximately 15% of the Brazilian land in the past (1.3 million km²)
- More than 90% of the total area has been devastated
- Microbial diversity in soils and/or associated with plants and animals is mostly unknown
Microbial communities in soils under the Atlantic forest

- Rhizospheric soils may harbor more than 1,000 bacterial and 100-200 fungal species
- Bacterial community structure in the rhizosphere is dependent on the plant taxon
- The factors driving bacterial community assembling in space and time are not known
Spatial variability in different forests

✓ 3 distinct forests (CB, GA, AS)
  10 ha permanent plots
  256 sub-plots sampled
  20 x 20 m grid
  0-5 cm depth

✓ Physical and chemical attributes

✓ Microbial community analyses: Fatty Acid Methyl Ester (FAME) profiles
Canonical redundancy analysis (RDA)
How to make sense of the complex dataset

✓ Dataset is multidimensional

✓ Dataset has no linear structure

✓ Data distribution is unknown

✓ Data show no apparent relationships

✓ Scientist has limited training on sophisticated statistical methods
An artificial neural network consists of several simple parallel processing units which communicate by sending signals to each other over a large number of weighted connections.

Analogous to a neuron system.
Artificial Neural Network

**Advantages**
- Predictive
- Pattern recognition
- Do not assume data distribution models
- Analyzes complex data with non-linear relationships
- Learn from data
- Requires less formal statistics training to develop

**Disadvantages**
- Black boxes
- Overtraining
- Limited ability to explicitly identify possible causal relationships
- Model development is empirical and many methodological issues remain to be resolved
Kohonen Self-Organizing Maps (K-SOM)

✓ unsupervised learning

✓ no expected output

✓ non-linear generalization of PCA (dimensionality reduction)

✓ samples clustered by similarity

Synapse (Peltarian Corporation)
Santos et al. 2014, Soil Biol Biochem
ANN Modeling

Microbial Biomass

Supervised learning
Training 85% dataset; Validation 15% dataset
Optimum number of iterations based on minimum error
Synapse (Peltarian Corporation)
ANN Predictions

Training 85% dataset
Validation 15% dataset
Synapse (Peltarian Corporation)
Artificial Neural Network

✓ can help us predicting the effects of environmental variables on specific microbial biomarkers

✓ could be used for hypothesis driven experiments in the field

✓ could help monitoring the impacts of anthropogenic factors in natural forests
Genesis of spodosols (podzols)

**Occurrence:** temperate, sub-tropical and tropical regions

**Characteristics:** eluviation of OM from horizon E to Bh; some have individualized **bleached mottles** depleted in organic matter, as compared to the surrounding soil;

**Previous studies:** microbial community structure in the mottles differs from surrounding soil; dominance of Pseudomonas or Acidobacteria in the mottles, depending on the profile location; bacterial communities in the mottles might be associated with drainage and or specific organic compounds
Genesis of spodosols (podzols)

**Occurrence:** temperate, sub-tropical and tropical regions

**Characteristics:** eluviation of OM from horizon E to Bh; some have individualized **bleached mottles** depleted in organic matter, as compared to the surrounding soil;

**Previous studies:** microbial community structure in the mottles differs from surrounding soil; dominance of Pseudomonas or Acidobacteria in the mottles, depending on the profile location; bacterial communities in the mottles might be associated with drainage and or specific organic compounds
Soil profiles studied (Cardoso Island, São Paulo)

- 11 profiles sampled for massive sequencing of the 16S rRNA gene V4 region
- 3 drainage types: Well Drained (5 profiles); Intermediary Drainage (3 profiles); Poorly drained (3 profiles)
- Samples collected from each horizon and from within the mottles and immediate vicinity
- Organic matter analyzed by pyrolysis coupled to GC/MS
✓ Pipeline from the *Brazilian Microbiome Project* (approximately 20 million valid 16S rRNA gene sequences)

✓ **USEARCH7 (Uparse)**: sequence quality filtering and clustering into Operational Taxonomic Units (OTUs)

✓ **Qiime 1.8** – OTU taxonomic identification, phylogenetic relationships and bacterial diversity
Bacterial community structure comparisons (PCA)

Horizons

PC1 (58.5%)
PC2 (11.6%)

M + S
Predominant phyla in the horizons
Actinobacteria and Acidobacteria abundance is the main difference between A- and Bh-horizon bacterial communities.
Kohonen Self-Organizing Maps
<table>
<thead>
<tr>
<th>Clusters</th>
<th>Unified Distance Matrix</th>
<th>Description</th>
<th>n-alkaC7.25</th>
<th>n-alkaC26-35</th>
<th>n-alkaC7.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>methylk</td>
<td>lignins</td>
<td>methyle</td>
<td>N-comp</td>
<td>polym</td>
<td>phenols</td>
</tr>
<tr>
<td>prist</td>
<td>polysac</td>
<td>sterols</td>
<td>Sodium</td>
<td>Potassium</td>
<td>N_total</td>
</tr>
<tr>
<td>C_total</td>
<td>Unassigned</td>
<td>AD3</td>
<td>Acidobacteria</td>
<td>Aotrobacteria</td>
<td>Armatimonadetes</td>
</tr>
</tbody>
</table>

The image shows a K-SOM (Knowledge-based Self-Organizing Map) with different clusters and unified distance matrix values. The description column includes various chemical and biological data represented in heatmaps with color gradients ranging from blue to red.
ANN modeling

Confidence Plot: Output + 0.03124874 is within desired with 95% confidence

Confidence Plot: Output + 0.04307419 is within desired with 95% confidence

Acidobacteria

Actinobacteria
ANN predictions
ANN predictions
Conclusions

- ANNs are useful to analyze complex soil microbial community datasets, in addition to the usual statistical methods.

- ANN can be used for detecting similarity patterns in complex soil microbial communities.

- ANNs can be used for predicting quantitative changes in soil microbial community structure.
ANN can be used for modeling phytobiomes

✓ better understanding the effects of plant-microbe interactions on crop productivity (K-SOM, prediction models)

✓ forecast crop production based on the microbial community and environmental conditions (time series predictions)

✓ predict increases in crop production due to microbial inoculations (time series predictions); or inoculation efficiency (classification)

✓ pathogen outbreaks (anomaly detection)
Acknowledgments

- Eder Costa, Elisa R. Matos, Kelly J. Silva
- David Crowley (UC-Riverside)
- Jorge Rodrigues (UC-Davis)
- Pablo Vidal (USP) and Peter Buurman (WU Environmental Sciences)