Poster presentation at the 5th International Symposium on Soil Organic Matter, Göttingen, Germany, September 20-24, 2015

Poster 562

Towards linking fungal genes to chemical spectra from soil organic matter using machine learning

Martin-Bertelsen T.[1], Nicolás C. [2], Bentzer J. [2], Persson P. [3], Tunlid A. [2], Troein C. [1]

[1]Lund University, Computational Biology and Biological Physics, Department of Astronomy and Theoretical Physics, Lund, Sweden

[2]Lund University, Microbial Ecology Group, Department of Biology, Lund, Sweden

[3]Lund University, Centre for Environmental and Climate Research & Department of Biology, Lund, Sweden

1. Introduction

Characterization of biological and chemical processes in soil organic matter (SOM) are increasingly being done by complementary high-throughput experimental techniques. Interpreting these diverse high-dimensional data types together poses challenges about data integration and analysis. A bioinformatics merge of machine learning and chemometrics may be a promising avenue to fully explore links between biological and chemical processes. We make here a proposal and demonstrate proof of concept for integrating genome-wide transcriptomic data (RNA-Seq) with chemical spectra (FT-IR, Pyrolysis-GC/MS).

2. Objectives

We aimed to propose computational methods for linking genes to decomposition mechanisms by integrating genome-wide transcription profiling data (RNA-Seq) with chemical changes in the organic compounds occurring during SOM decomposition as measured by chemical spectra (FT-IR, pyrolysis-GC/MS). The methods are intended to enable extraction of patterns that can be recognized and interpreted by domain experts. We also wanted to show proof of concept of integration of genome-wide gene expression data with chemical spectra from a study of two fungal species in a SOM decomposition time series under gradual glucose depletion.

3. Materials & Methods

A controlled experiment was set up to measure effects of decreasing glucose levels over time on the decomposition of SOM. Two ectomycorrhizal fungi with distinct growth rates (Paxillus involutus and Laccaria bicolor) were grown on SOM water extracts under axenic conditions in a time series experiment. To trigger the fungal decomposition activity, the extracts were initially supplemented with glucose. Four time points were selected such that the glucose was almost depleted between the 2nd and 3rd time point. Sampling points were chosen for each species

based on matching glucose levels to define a common time reference scale of the two fungal growth experiments during the time course.
The mycelia was collected for transcriptome profiling using RNA-Seq, while the SOM extracts were analysed by means of FT-IR spectroscopy and Pyrolysis-GC/MS at these four species-specific time points. Species-specific co-expression networks were constructed from pairwise correlations between gene transcriptions across the time points. Genes of similar functions between fungi were identified from orthology detection using the ProteinOrtho software program. Spectral associations akin to chemometric 2D synchronous correlation analysis were computed and correlated with gene transcriptions to create gene-spectral association networks. The OrthoClust software for discovery of conserved as well as species-specific modules of co-expressed genes was modified to include correlated spectral patterns thus enabling characterization of the glucose starvation responses in terms of modules of genes and associated spectral ranges.

4. Results

We propose some computational approaches for integrating biological and chemical data types of genome-wide transcriptional RNA-Seq data and chemical spectra. As proof of concept we demonstrate the use of a clustering method on a time series experiment. [Anticipated:] Modules of genes with similar expression profiles across time points were found and linked to certain spectral ranges of the FT-IR chemical spectra as well as certain compounds found with pyrolysis. Our current work involves more complex statistical modelling of several high-dimensional data types using tailored variants of sparse factor analysis to discover interpretative correlation patterns of SOM decomposition mechanisms. The dataset consist of RNA-Seq and chemical spectra from 10 fungal species in controlled experiments comparing growth medium and SOM extracts. Some of the challenges involve cross-species normalization of RNA-Seq, orthology mapping for functional comparison and identification in the statistical model. However, these statistical methods require large data sets for detailed inference about molecular mechanisms from modern high-throughput data and collaborative efforts are encouraged to reach critical data mass.

5. Conclusion

We have demonstrated a[n anticipated] powerful application of machine learning clustering tools for integration of transcriptomics and chemical spectra to leverage interpretation of data from controlled fungal SOM experiments.