Mapping fungal genes to decomposition of soil organic matter

ATP talk Nov 11 2015

Tomas Martin-Bertelsen, CBBP

SOM degradation

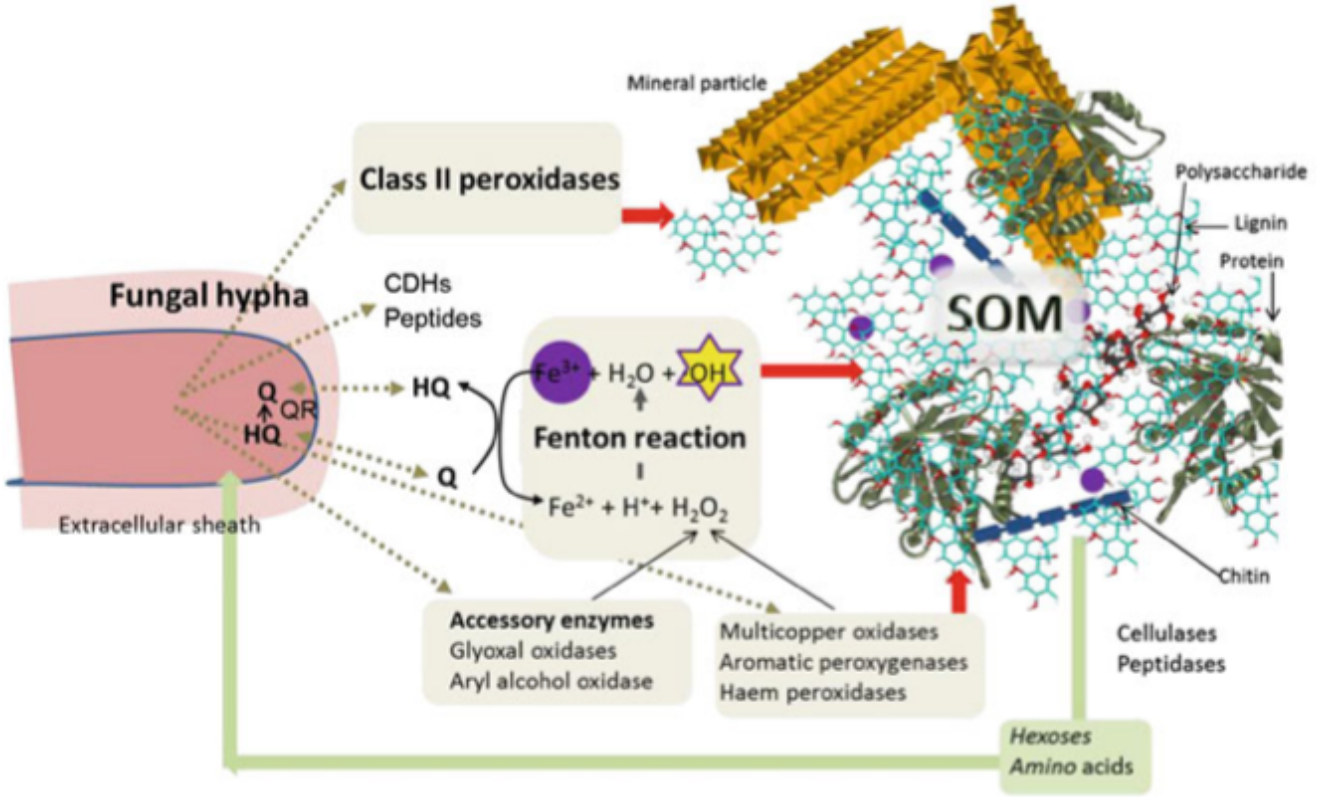


Fig. 8.1 A schema of the complex interactions between fungal hyphae and soil components, which ultimately determine the extent and rate of soil organic matter (SOM) degradation (see main text for explanations)

A. Tunlid et al. 2013, in *Genomics of Soil- and Plant-Associated Fungi*, Soil Biology 36

LUND UNIVERSITY

Introduction

- SOM: Soil Organic Matter
- Major part of global carbon stored in SOM

Aims

- To get closer to a mechanistic understanding we need the components (genes and organic molecules, functional groups).
- Longer perspective: Biomarkers to predict soil qualities, e.g. during field work.

SOM extracts

- Top soil layer of degraded plant-litter
- Collected from spruce forest nearby
- Boiled in water and filtered

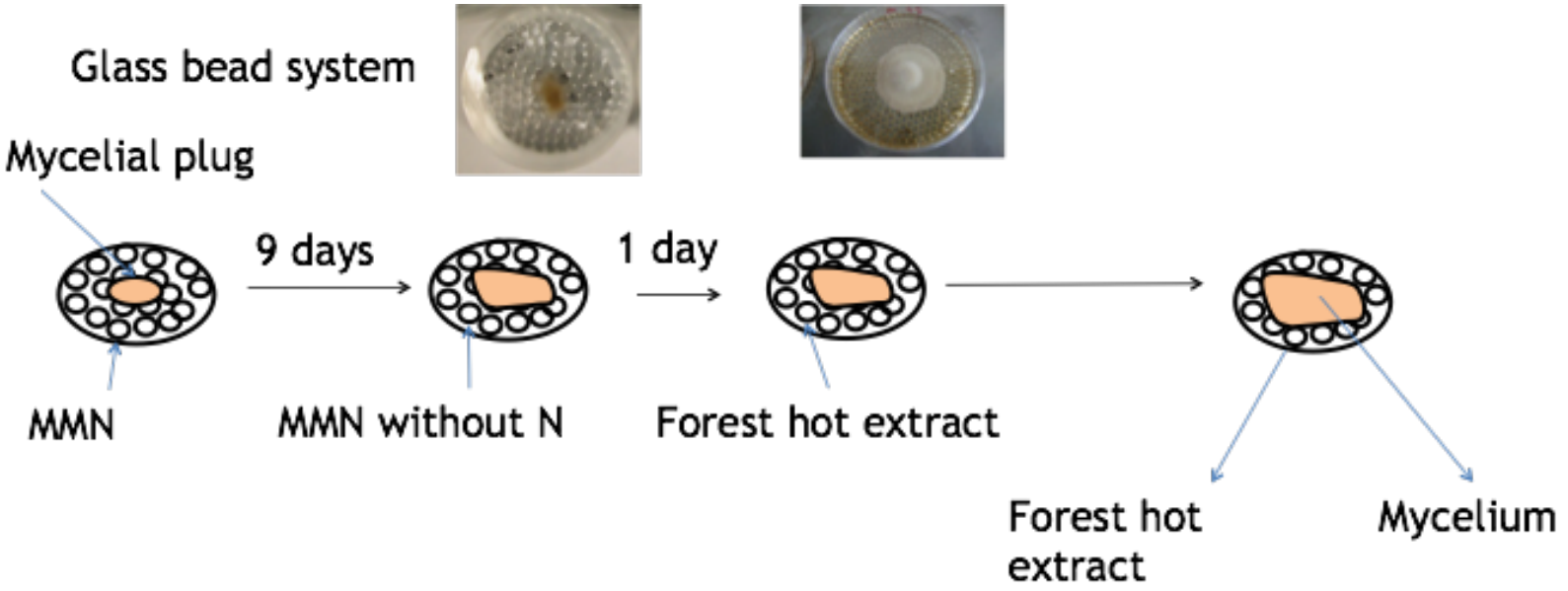
LUND UNIVERSITY

Experimental data

- Several species of litter-decomposing fungi
- Several measurement techniques
  - Transcriptional activity is measured by mRNA sequencing technology
  - chemical modifications quantified by chemical spectra from experimental techniques such as FTIR and Pyrolysis-GC/MS.

- Integration of these diverse data types.

Experimental setup



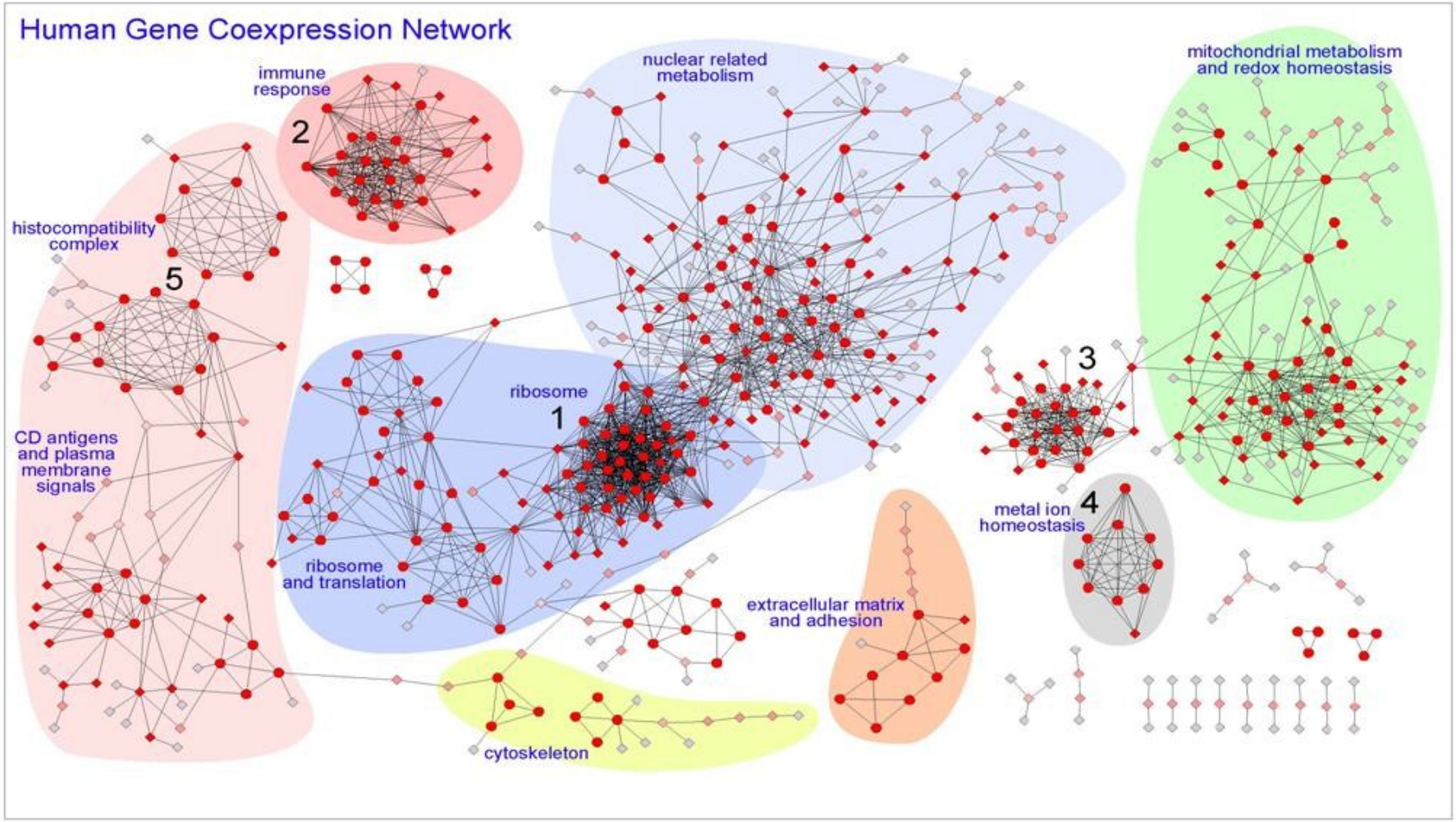
- Comparative experiment: 7 days and 9 different fungi.

- Time series experiment: longer time and plates collected at 4 different time points (2 early, 2 late).

Credits: César Nicolas

Networks and modules

- Biological networks
- Proteins or genes linked together
- Coordinated regulation of genes in biological processes make up functional modules



Prieto et al. 2008, PLOS one

Co-expression network

- Coordinated gene expression due to common function
- Pearson's correlation between pairs of genes
- local rank based on absolute value
  (Ruan et al. 2010 BMC Syst Biol)
  - Connect each gene to top *d* neighbours
    - Sparsely connected network such that edge density varies across network and modules can be identified
    - degree distribution similar to other biological networks

LUND UNIVERSITY

Modularity function

Network with n vertices, m edges defined by adjacency matrix $\mathbf{A}$.
$A_{ij} = 1$ if edge between vertex i and j, otherwise 0.
P_{ij} probability in the null model of edge between vertex i and j.
g_i the assigned module for vertex i.

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij}) \delta(g_i, g_j)$$

A quality score of the module assignments. (Newman and Girvan 2004)

Simulated annealing algorithm for optimization over module assignments. (Reichardt J, Bornholdt S, Phys Rev Lett 2004)

LUND UNIVERSITY

Null model

- Newman null model assigns edges at random with the expected degrees of model vertices constrained to match the degrees in the actual network.

$$k_i = \sum_j A_{ij} \text{ the degree of vertex } i.$$
$$m = \sum_i k_i/2 \text{ number of edges in network.}$$

$$P_{ij} = \frac{k_i k_j}{2m}$$

LUND UNIVERSITY

Orthology

- E. V. Koonin 2005, *Orthologs, Paralogs, and Evolutionary Genomics*
  - Homologs: genes sharing a common origin
  - **Orthologs: genes originating from a single ancestral gene in the last common ancestor of the compared genomes**
  - Paralogs: genes related via duplication
- Orthologous genes often have equivalent functions.
- Makes expression data comparable across species.
- Co-expression network clusters based on orthologous genes.

LUND UNIVERSITY

OrthoClust concept

Yan *et al. Genome Biology* 2014, **15**:R100
http://genomebiology.com/2014/15/8/R100

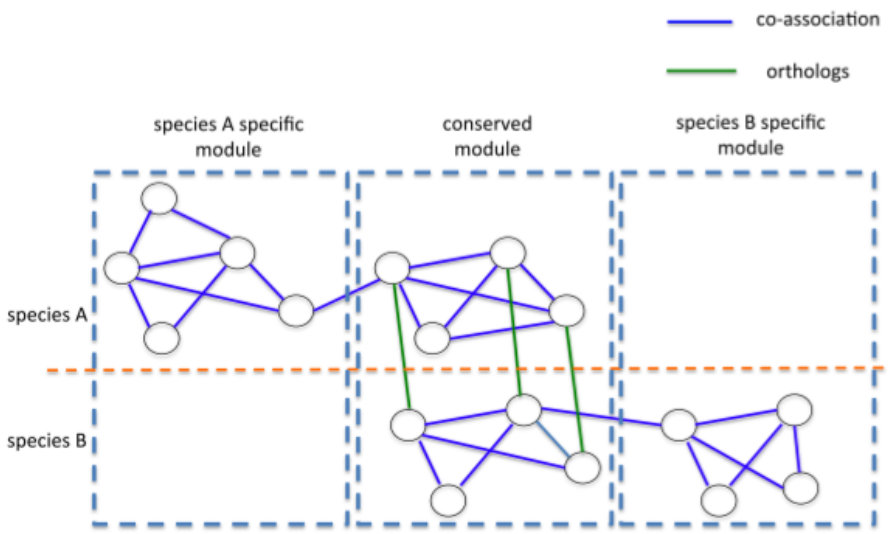


Figure 1 An example to illustrate the idea of modules in a multi-layer network. The co-association networks of species A and B are linked together to form a multi-layer network via orthologous relationship between genes. There are three modules. The middle one is a conserved module with genes from both species, corresponding to fundamental biological functions across different species. The left and right ones are specific modules consisting of genes from species A and B, respectively. They correspond to novel functions that emerged in each of the two species.

OrthoClust modularity function

- Multi-layer network with coupling constant κ
- Each network its own modularity term (species 1 and 2)
- Score increases for Orthologous gene pairs in same module

$$Q = \sum_{i,j \in S_1} B_{ij}^{(1)} \delta(g_i, g_j) + \sum_{i,j \in S_2} B_{ij}^{(2)} \delta(g_i, g_j) + \kappa \sum_{(i,j') \in O(S_1,S_2)} \delta(g_i, g_j)$$

$$B_{ij} = A_{ij} - P_{ij} \text{ the modularity}$$

LUND UNIVERSITY

Multitype data

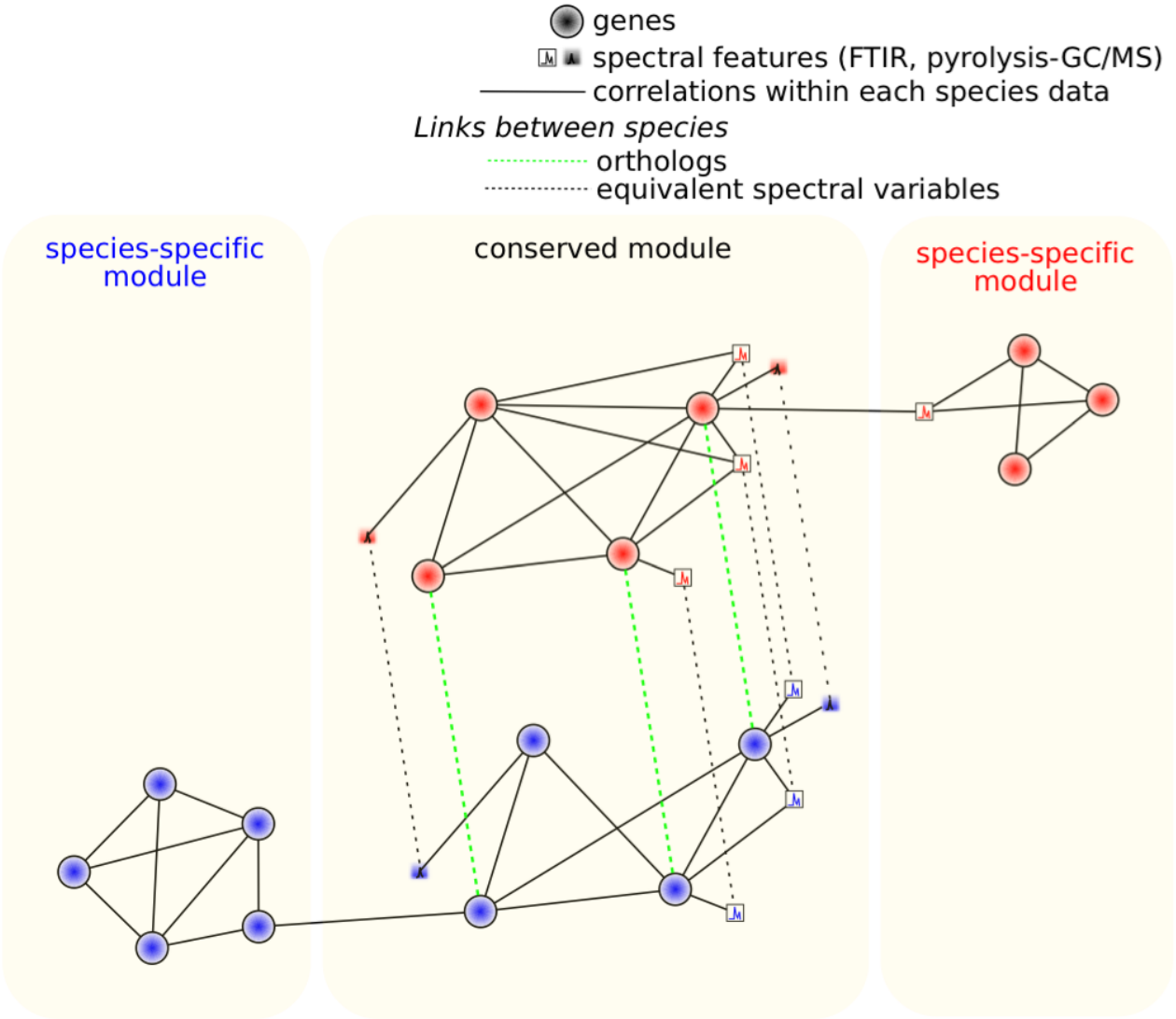
Two parts of sample source material from each growth experiment results in two sets of measurements

- RNA-Seq (gene expression from mycelium part of sample)
- FTIR and pyrolysis-GC/MS chemical spectra of modified SOM extract

Extending OrthoClust

- Multiple data types – each represented as individual networks
- The principle of shared and specific patterns between species (modules, correlations) – now also between different data types
- Modularity term for each data type for each species
- Linking two different data types corresponds to an individual bipartite subnetwork

Extending OrthoClust



LUND UNIVERSITY

Modularity for bipartite networks

- Due to the constraint that edges only occur between nodes of different data types a different null model applies (Barber, Phys. Rev. E, 2007)

$$\tilde{P}_{ij} = \frac{k_i d_j}{m}$$

- Modularity function then becomes

$$Q_{\text{bipartite}} = \frac{1}{m} \sum_{i \in \text{genes}} \sum_{j \in \text{waveno.}} (\tilde{A}_{ij} - \tilde{P}_{ij}) \delta(g_i, g_j)$$

Extended OrthoClust

- Constructing the different correlation networks, adding up the modularity terms and optimize quality function
- In progress …
- Preliminary experiments indicate the need to treat different data types as individual networks as outlined here.

LUND UNIVERSITY

Interpreting identified modules?

- Find enriched biological annotations in the identified modules
  - Does a module contain many genes of certain known function?
  - Secondary metabolite gene clusters perhaps?
- Spectroscopists identify functional groups corresponding to spectral peaks.
- The modules containing genes and spectral variabels may thus elucidate potential mechanism of decomposition.

Future work

- Integrating functional annotation data in the module identification process?
- Alternative methods?

Group factor analysis

- A more generative approach modelling the data directly instead of doing network construction.
- Find latent variables shared between data types as well as latent variables for data type-specific covariations.
- Share latent variables can be used to link gene expression to spectral data.
- Matrix factorization model.
- Allows prediction and simulation of one type of data from another type, e.g. predicting chemical modifications from gene expression alone.

LUND UNIVERSITY

Thank you

Some people from the MICCS project

- Anders Tunlid, Microbial Ecology Group, Department of Biology, PI
- Per Persson, Centre for Environmental and Climate Research & Department of Biology, co-PI
- Carl Troein and Carsten Peterson, CBBP, co-PI
- César Nicolás Cuevas (time series data) postdoc, Microbial Ecology Group, Department of Biology
- Johan Bentzer bioinformatician, Microbial Ecology Group, Department of Biology

Further info about the MICCS project: www.miccs.info

LUND UNIVERSITY

LUND
UNIVERSITY