MLSB12

Using PPI Network Autocorrelation in Hierarchical Multi-label Classification Trees for Gene Function Prediction

Daniela Stojanova[1], Michelangelo Ceci[2], Donato Malerba[2], Sašo Džeroski[1]

[1]: Jožef Stefan Institute, Slovenia
[2]: Università degli Studi di Bari, Italy

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

The Task

- Predictive modeling
  - Hierarchical Multi-label Classification (HMC) task
  - Gene Function Prediction

Input: attributes target

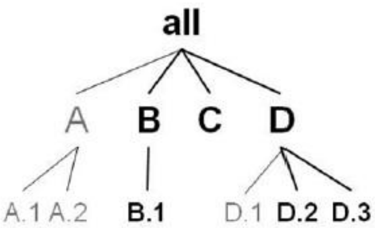
Attribute set -> [ylr216c, 0.595, n, 1.133, 0.255, 0.558, c, 1.193]

GeneID, DESMMS5, BCIP, wtG5, wtG1, 2-deoxyglucose, pRSGAL

Class hierarchy -> all, A, A.1, A.2, B, B.1, C, D, D.1, D.2, D.3

Class vector -> L= [1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1]

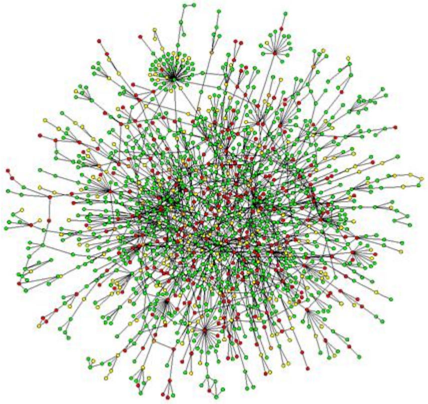
A. METABOLISM
A.1 amino acid metabolism
A.2 nitrogen, sulfur, selenium met.
A.1.3 assimilation of ammonia
A.1.3.1 metabolism of glutamine
A.1.3.1.1 biosynthesis of glutamine
A.1.3.1.2 degradation of glutamine
...
B. ENERGY
B.1 glycolysis and gluconeogenesis
C. CELL CYCLE and DNA PROCESSING
D. TRANSCRIPTION
D.1 RNA synthesis
D.2 RNA processing
D.3 transcriptional control



Output: predictive model

The Context

- Predictive modeling in a network setting
- Networks
  - Protein-protein interaction networks



A DIP PPI network

  - Biological networks (homology, metabolic…)

The Problem

- Multi-label prediction of gene functional classes given:
  - relationships among the classes (instances belonging to multiple hierarchically organized classes)
  - relationships among the instances (instances connected in PPI networks)

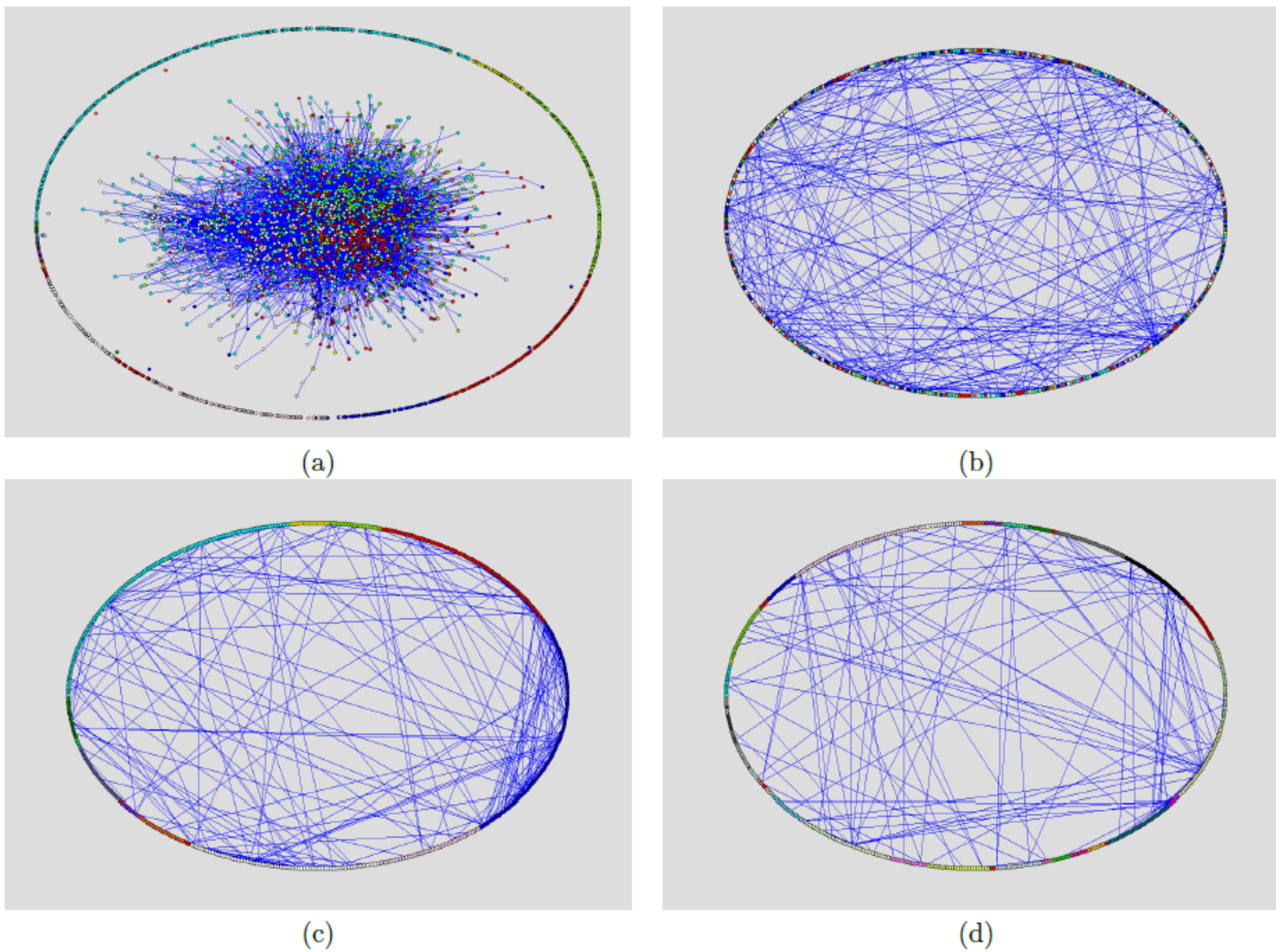
- The latter introduce autocorrelation and violate the i.i.d. assumption

Autocorrelation

- **Correlation** → any statistical relationship between different variables of the same objects
- **Autocorrelation** → any statistical relationships between the same variable on different but related (dependent) objects
- **Network Autocorrelation** in HMC setting→ statistical relationship between observations of a variable on distinct but related nodes in a network where the domain values of the variable are given as subsets of hierarchy
  - In our case nodes are genes and the considered variable represents its biological function

Network autocorrelation of gene functions

DIP Yeast network



(a) Not connected examples arranged along the border (b) Examples are arranged along the border

(c) Examples grouped according to the 1st level of FUN (d) Examples grouped according to the 2nd level of FUN

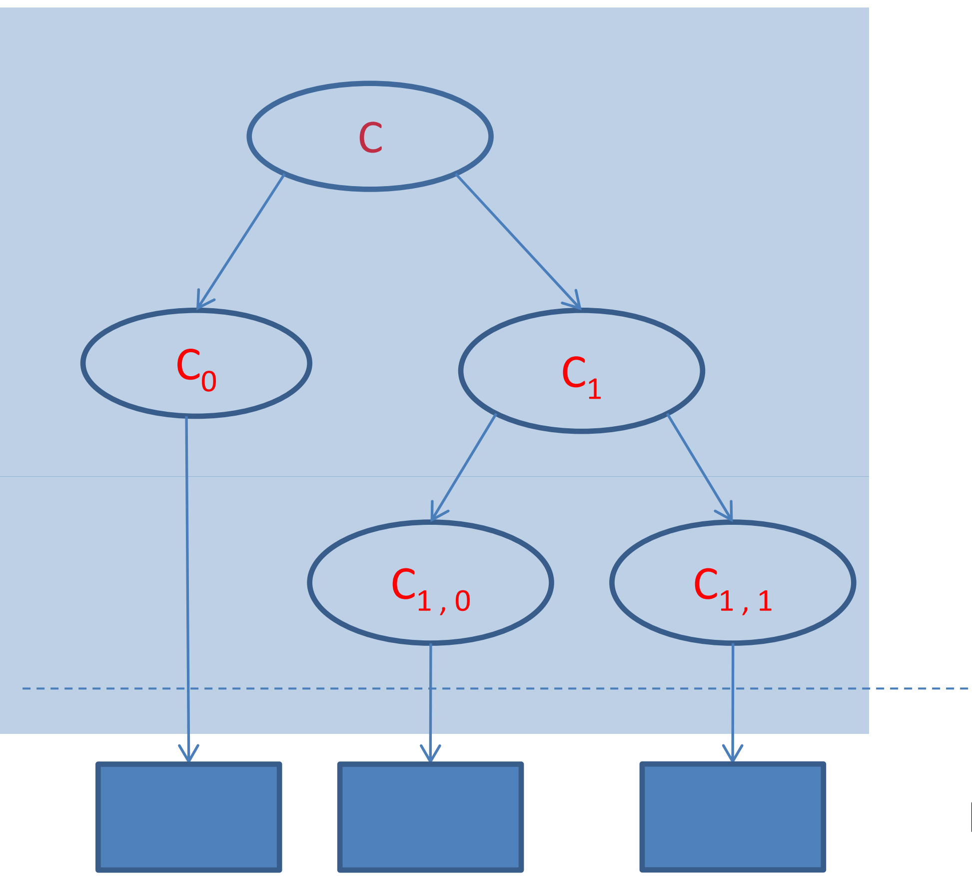
The Basic Idea

- We develop a tree-based algorithm **NHMC** (Network Hierarchical Multi-label Classification) for considering network autocorrelation in the setting of Hierarchical Multi-label Classification (HMC)

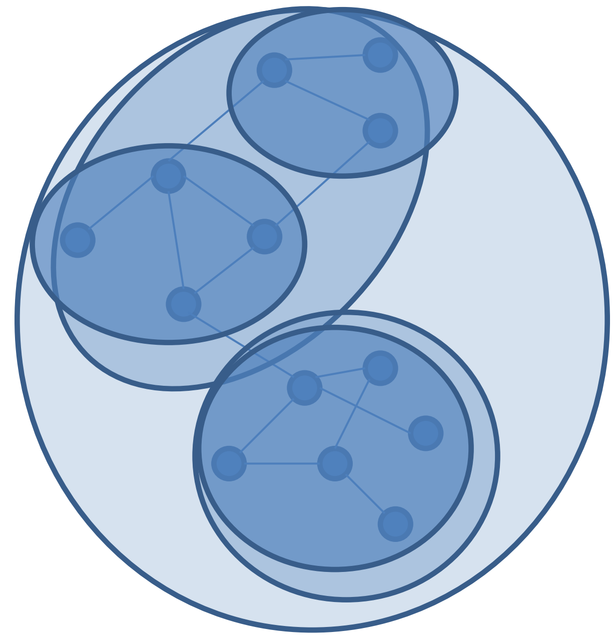
- It is based on the **CLUS-HMC** method that learns Predictive Clustering Trees (PCTs) for HMC. The network is used as background knowledge during training

- PCTs are decision trees viewed as hierarchies of clusters and provide symbolic descriptions of the clusters
  - Predictive clustering combines elements from both prediction and clustering
  - As in clustering, clusters of examples that are similar to each other are identified, but
  - A predictive model is associated to each cluster
- Clustering is based on autocorrelation: each cluster should contain highly autocorrelated entities

The Basic Idea: learning NHMC



Predictive models



Autocorrelation measure

- Geary' C

$$C_Y = \frac{(N-1)\sum_i \sum_j w_{ij}(Y_i - Y_j^2)}{2\sum_{i,j} w_{ij} \sum_i (Y_i - \overline{Y})^2}$$

where Y is a real variable

- New autocorrelation measure for HMC setting

$$A_L = 1 - \frac{(N-1)\cdot \sum_i \sum_j w_{ij} \cdot d(L_i, L_j)^2}{4 \cdot \sum_i \sum_j w_{ij} \cdot \sum_i d(L_i, \overline{L})^2}$$

where

- L is vector representation of labels (satisfies the hierarchical constraint)
- d is weighted Euclidian distance of such vectors

Algorithm outline

Algorithm 1 Top-down induction of NHMC

procedure NHMC($G = (V, E), \eta(\cdot)$) **returns** tree
if stop($V, \eta(\cdot)$) **then**
  return leaf(Prototype($V, \eta(\cdot)$))
else
  $(d^*, h^*, \mathcal{P}^*, \mathcal{P_V}^*) = (null, 0, \emptyset, \emptyset)$
  $D = \{\eta(v) | v \in V\}$
  for each possible Boolean test t according to the values of $\mathbf{X}$ on the dataset D **do**
    $\mathcal{P} = \{D_1, D_2\}$ partition induced by d on D
    $\mathcal{P_V} = \{V_1, V_2\}$ = partition induced by $\mathcal{P}$ on V;
    $h = \boxed{\alpha \cdot \left(\frac{|D_1| \cdot A(D_1) + |D_2| \cdot A(D_2)}{|D|} \right) + (1-\alpha) \cdot \left(Var'(D) - \frac{|D_1| \cdot Var'(D_1) + |D_2| \cdot Var'(D_2)}{|D|} \right)}$
    if $(h > h^*)$ **then**
      $(d^*, h^*, \mathcal{P}^*, \mathcal{P_V}^*) = (d, h, \mathcal{P}, \mathcal{P_V})$
    end if
  end for
  $\{V_1, V_2\} = \mathcal{P_V}^*$
  $tree_1$ = NHMC($G_1 = (V_1, E), \eta(\cdot)$)
  $tree_2$ = NHMC($G_2 = (V_2, E), \eta(\cdot)$)
  return node($d^*, tree_1, tree_2$)
end if

A() network autocorrelation

Var() variance reduction

Datasets

- 12 yeast datasets, from Clare & King, 2003
  - different aspects : sequence statistics, phenotype,
    secondary structure, homology and expression
- Entire set
  - Subset of highly connected genes (>15 connections)

- 2 hierarchies of gene function

  MIPS-FUN

  Gene Ontology (GO)

- 3 PPI networks

  DIP - Database of Interacting Proteins (Deane et al., 2002)

  VM - von Mering (von Mering et al., 2002)

  MIPS - Munich Information Center for Protein Sequences(Mewes et al.,1999)

Results

- Comparison between CLUS-HMC &NHMC
- Comparison to other methods
- Comparison using different PPI networks

Comparison CLUS-HMC &NHMC

Dataset	All genes			Highly connected genes		
	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0$	$\alpha = 1$	$\alpha = 0.5$	$\alpha = 0$
FUN annotated datasets						
seq	0.059	0.054	0.053	0.051	0.094	0.130
pheno	0.036	0.035	0.028	0.068	0.333	0.333
struc	0.030	0.020	0.020	0.093	0.088	0.093
hom	0.073	0.020	0.023	0.149	0.088	0.088
cellcycle	0.032	0.030	0.037	0.047	0.098	0.125
church	0.029	0.020	0.020	0.041	0.091	0.091
derisi	0.027	0.028	0.025	0.048	0.098	0.119
eisen	0.047	0.042	0.025	0.067	0.147	0.183
gasch1	0.036	0.040	0.032	0.060	0.103	0.124
gasch2	0.034	0.034	0.027	0.037	0.108	0.112
spo	0.030	0.029	0.025	0.044	0.049	0.134
exp	0.040	0.030	0.025	0.067	0.091	0.132
Average:	0.039	0.032	0.028	0.064	**0.116**	**0.139**
GO annotated datasets						
seq	0.034	0.032	0.030	0.037	0.072	0.100
pheno	0.019	0.016	0.016	0.051	0.016	0.051
struc	0.018	0.012	0.012	0.078	0.078	0.078
hom	0.040	0.013	0.013	0.047	0.068	0.068
cellcycle	0.019	0.287	0.288	0.027	0.036	0.018
church	0.014	0.015	0.012	0.017	0.025	0.025
derisi	0.017	0.015	0.017	0.078	0.078	0.106
eisen	0.030	0.024	0.024	0.043	0.061	0.146
gasch1	0.024	0.018	0.019	0.051	0.094	0.095
gasch2	0.020	0.021	0.021	0.040	0.088	0.107
spo	0.019	0.018	0.015	0.040	0.078	0.090
exp	0.023	0.017	0.016	0.045	0.036	0.092
Average:	0.022	**0.041**	**0.040**	0.046	**0.058**	**0.081**

NHMC outperforms CLUS-HMC

Comparison to other methods

- $\overline{AUPRC}$ results using HMC-GA (Genetic Algorithm), HMC-LMLP (Artificial Neural Networks), hmAnt-Miner (Ant Colony Optimization) and NHMC_05

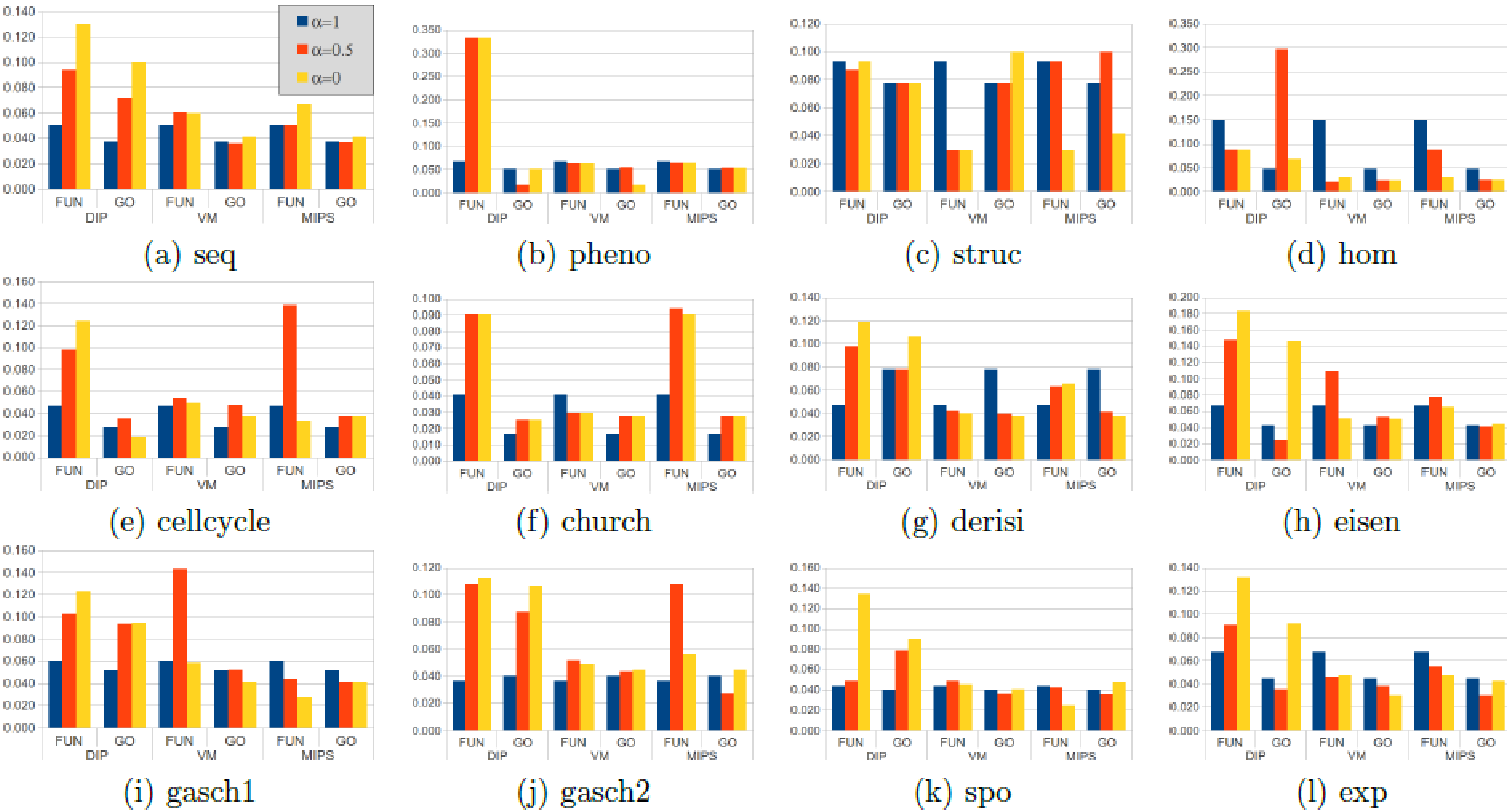
Dataset	HMC-GA	HMC-LMLP	hmAnt-Miner	NHMC_05
pheno	0.148	0.085	0.162	0.241
cellcycle	0.150	0.144	0.154	0.173
church	0.149	0.140	0.168	0.152
derisi	0.152	0.138	0.161	0.172
eisen	0.165	0.173	0.180	0.196
gasch2	0.151	0.132	0.163	0.186
spo	0.151	0.139	0.174	0.181

Best results are obtained using NHMC

Comparison using different PPI networks

$\overline{AUPRC}$ results of CLUS-HMC (α= 1) and NHMC (α = 0.5 & α = 0) for FUN and GO annotations

Best results are obtained using DIP networks



(a) seq (b) pheno (c) struc (d) hom

(e) cellcycle (f) church (g) derisi (h) eisen

(i) gasch1 (j) gasch2 (k) spo (l) exp

Contributions

- Definition of network autocorrelation in HMC setting
- Introduction of an appropriate autocorrelation measure
- Consideration of network autocorrelation in gene function prediction
- Development of method for hierarchical gene function prediction in a PPI network context

Conclusions

- We tackle the problem of HMC prediction of gene functions, when relationships among the classes & the instances exist

- HNMC can predict multiple gene functions, when gene classes are hierarchically organized (in the form of trees or DAGs, according to a classification schemes such as FUN&GO)

- HNMC takes into account PPI networks & network autocorrelation of gene function that arises in this context

- NHMC needs PPI only during training & not for prediction

- Due to the tree structure of the learned models, it is also possible to consider non-stationary autocorrelation

Future Work

- Evaluate our approach on different tasks of gene function prediction
  - additional datasets (organisms)
  - networks
    - homology
    - other similarities
- Different tasks for other biosciences:
  - Predicting community structure from environmental properties in a spatial setting

Q & A