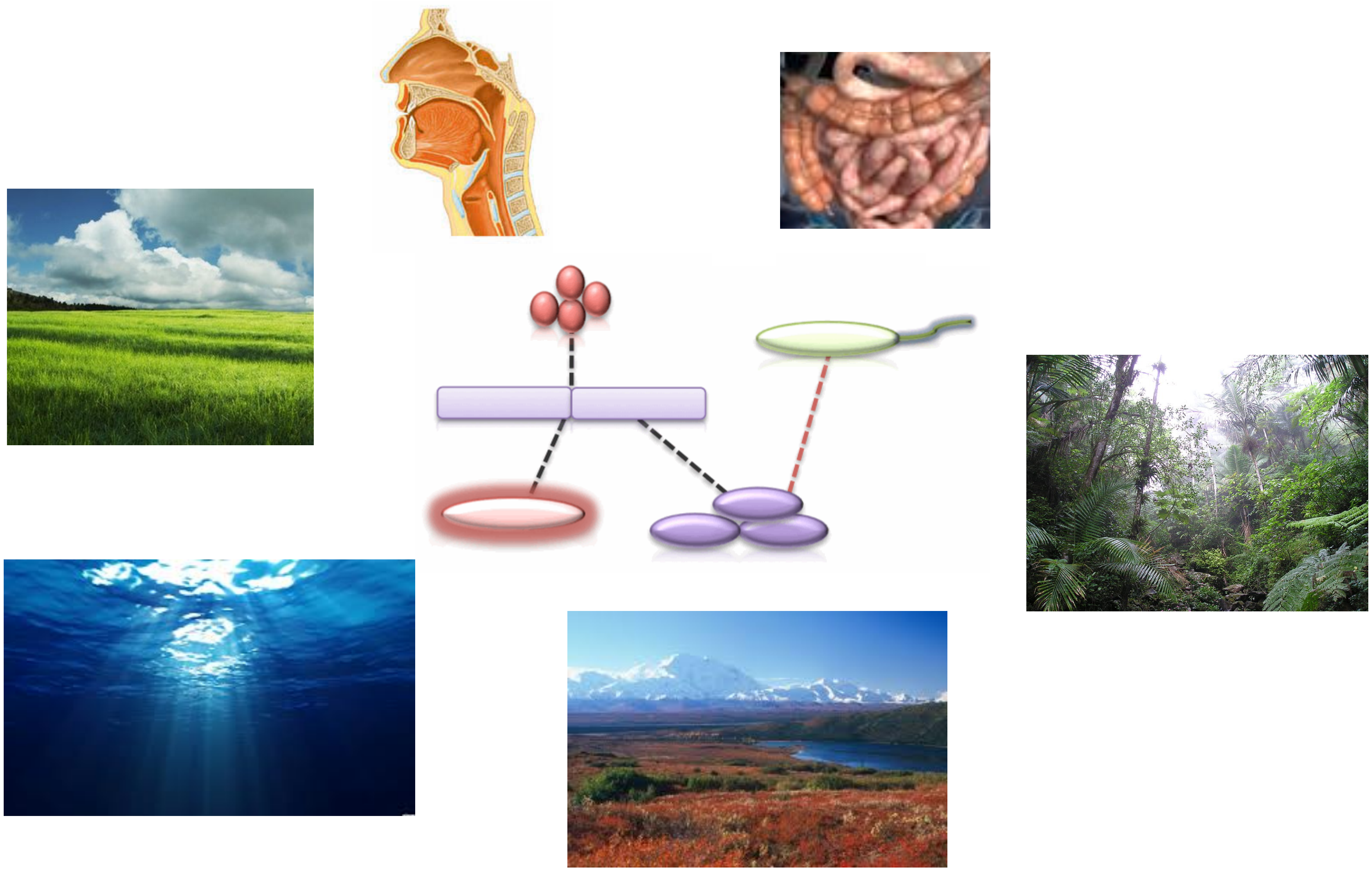
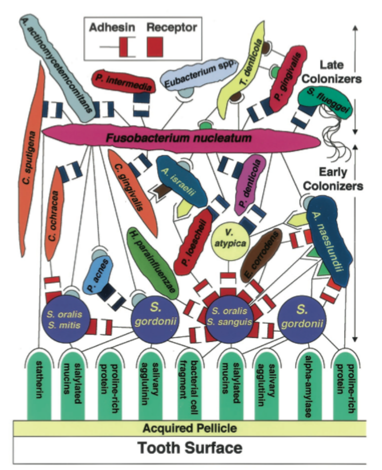
Karoline Faust

Inference of microbial association networks from metagenomic data

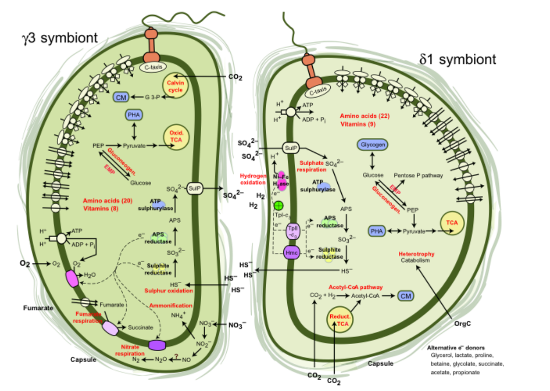


Examples for microbial relationships

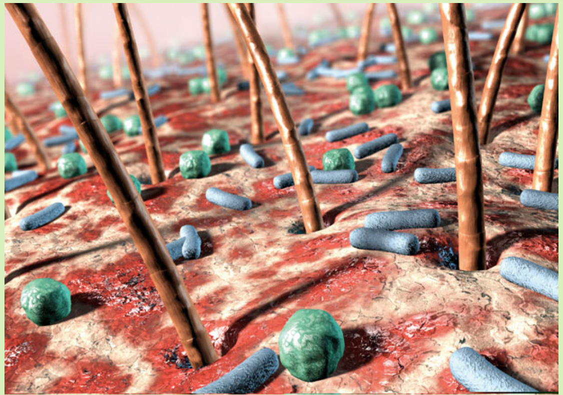


dental plaque formation
(Kolenbrander et al.)

sulfur oxidizer sulfate reducer



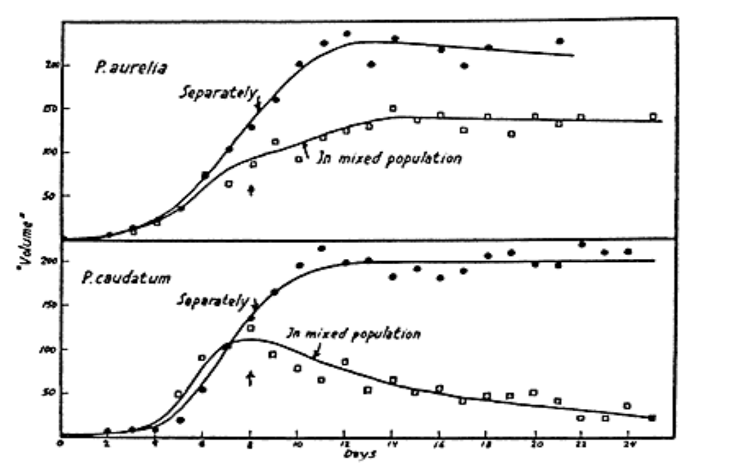
cross-feeding between bacterial symbionts of a marine worm (Woyke et al.)



artist's rendering of human skin bacteria

Amoeba proteus feeding on algae

Bacteriophages infecting a bacterium



competition between two species of Paramecium (Gause)

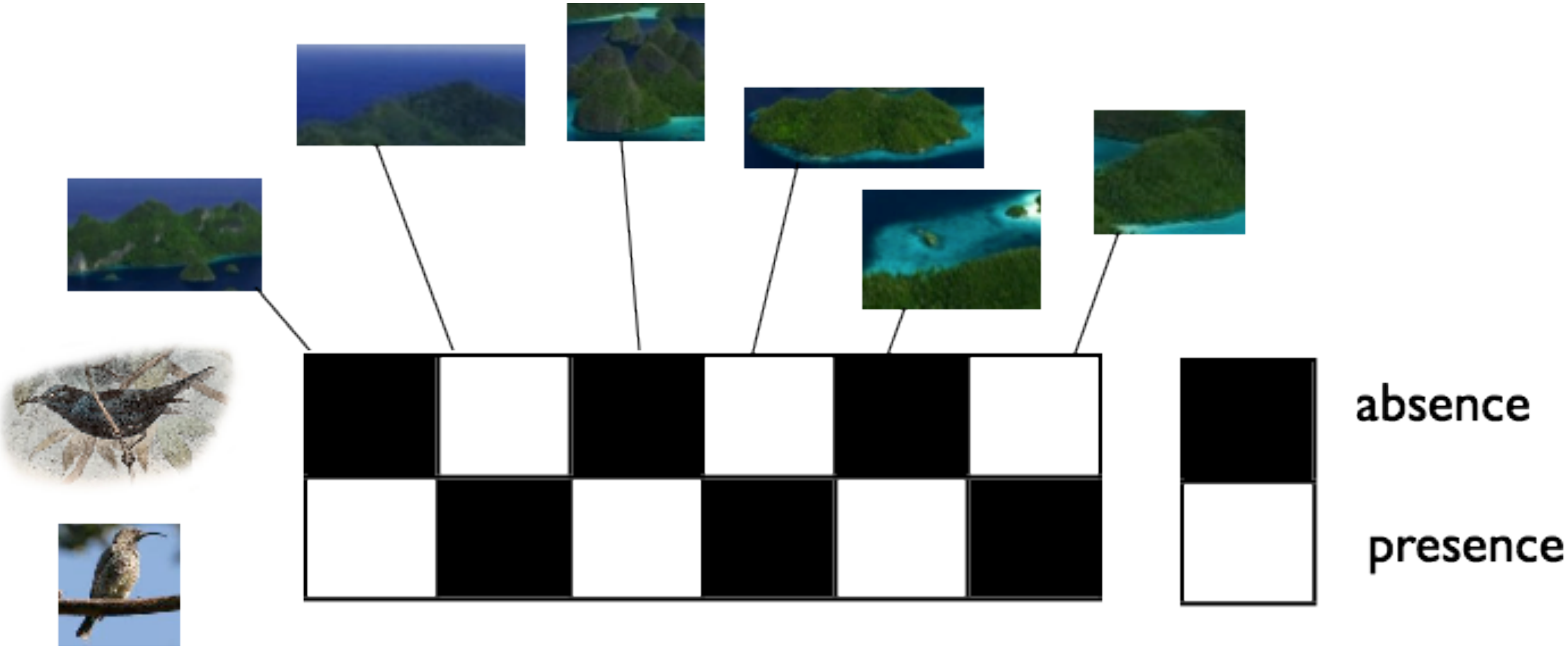
algae bloom killing off other organisms

Gause (1934) "The Struggle for Existence", Williams & Wilkins.
Kolenbrander et al. (2002) "Communication among Oral Bacteria." Microbiol. and Mol. Biol. Reviews 66, 486-505.
Woyke, T. et al. (2006) "Symbiotic insights through metagenomic analysis of a microbial consortium." Nature 443, 950-955.

1. Background

Co-occurrence analysis

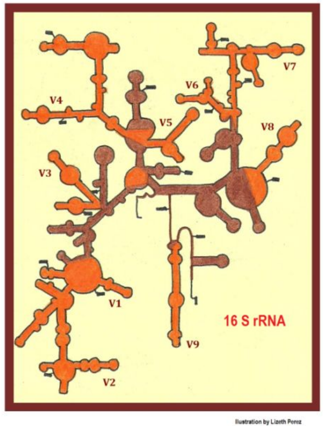
- Jared Diamond suggested that competition between species could be seen from their presences/absences across habitats (checkerboard pattern)
- checkerboard-like co-occurrence patterns have been found for micro-organisms as well (Horner-Devine et al.)



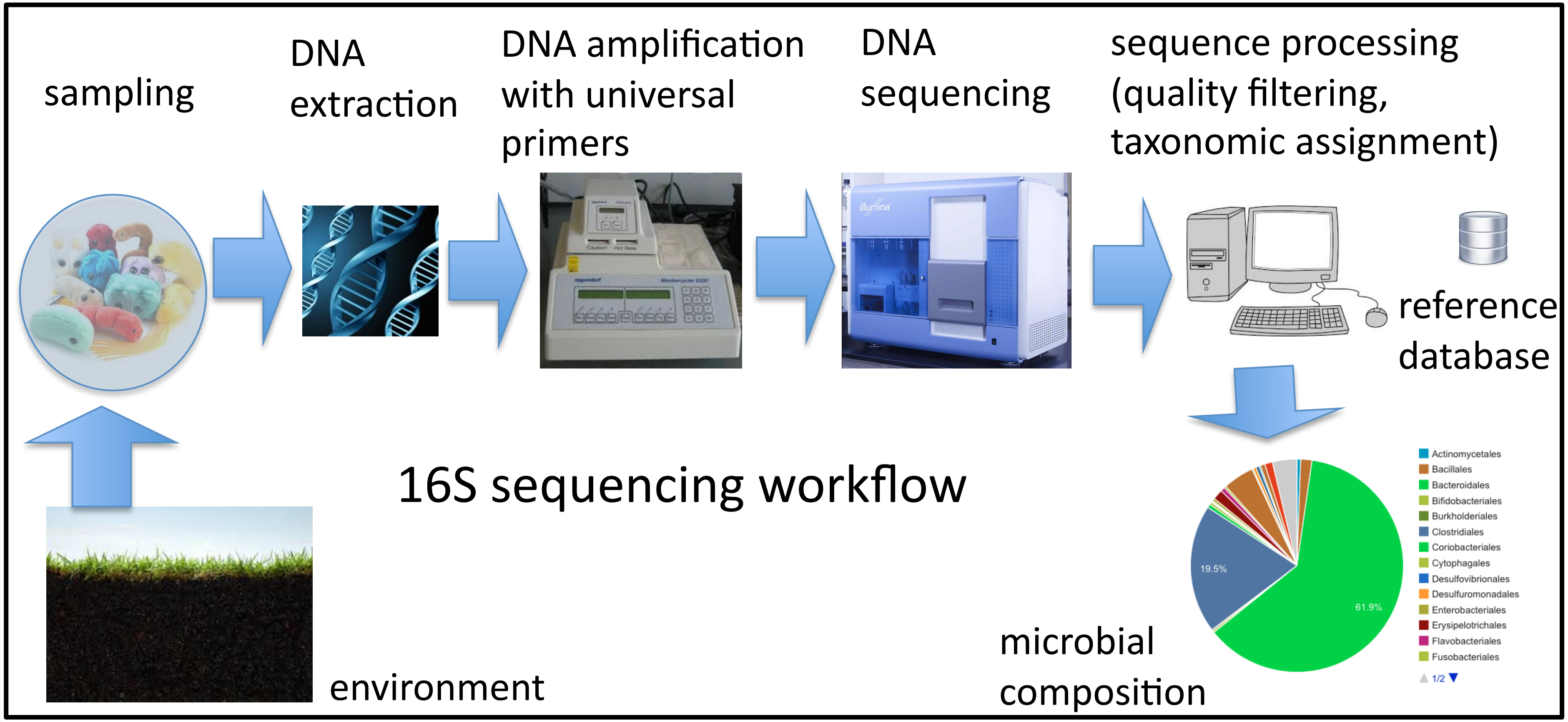
Diamond, J. (1975) “Assembly of species communities”, pp. 342-444 in “Ecology and evolution of communities” edited by Cody and Diamond, Harvard University Press.
Horner-Devine M.C. et al. (2007) “A Comparison Of Taxon Co-Occurrence Patterns For Macro- And Microorganisms” Ecology 88, 1345-1353.

1. Background

Community profiling with 16S sequencing



- 16S ribosomal RNA (coded by 16S rDNA genes) is composed of hypervariable and conserved regions
- hypervariable regions serve as markers for taxonomic classification
- conserved regions serve as binding sites for universal primers during DNA amplification



sampling

DNA extraction

DNA amplification with universal primers

DNA sequencing

sequence processing (quality filtering, taxonomic assignment)

reference database

16S sequencing workflow

environment

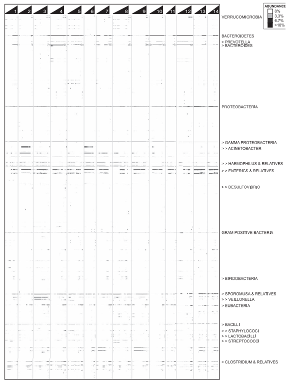
microbial composition

1. Background

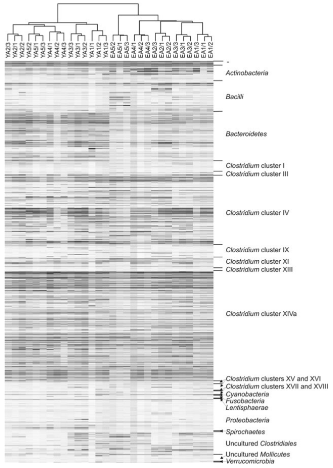
Other community profiling techniques

organism counting techniques (flow cytometry, FlowCam, ZooScan)

phylogenetic microarrays



infant gut (Palmer et al., 2007)



HITChip (Rajilic-Stojanovic et al., 2009)

Reasons for co-occurrence

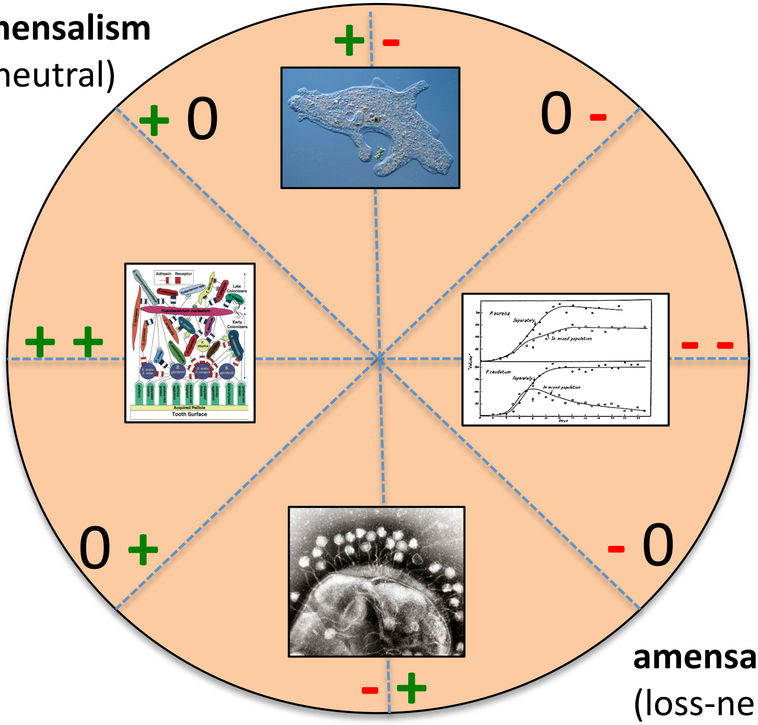
Why would two taxa consistently occur together or avoid each other across samples?

ecological relationships

predator/parasite (win-loss)

commensalism (win-neutral)

mutualism (win-win)

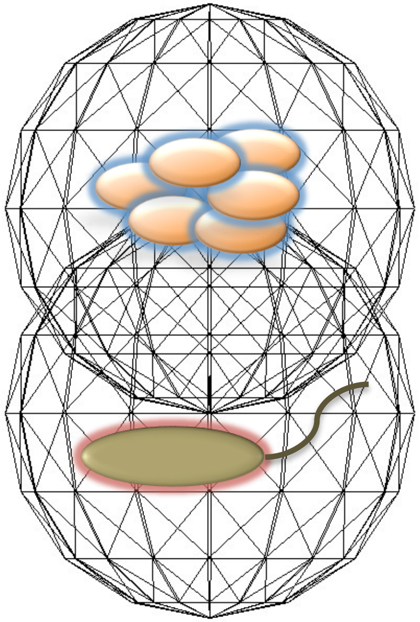


competition (loss-loss)

amensalism (loss-neutral)

prey/host (loss-win)

niche overlap



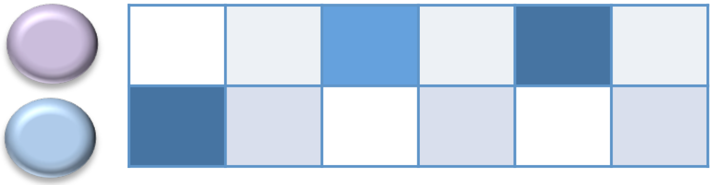
Adapted from Lidicker, W.Z. (1979) “A Clarification of Interactions in Ecological Systems.” BioScience 29, 475-477.

Hutchinson, G.E. (1957) “Concluding remarks” Cold Spring Harbour Symposium on Quantitative Biology 22, 415-427.

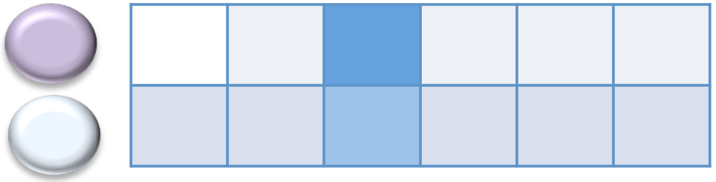
1. Background

1. Background

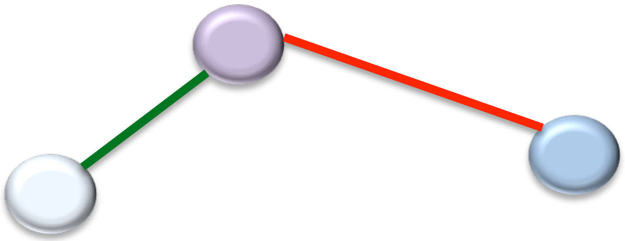
Co-occurrence analysis in a nut shell



mutual exclusion (checker board)/anti-correlation

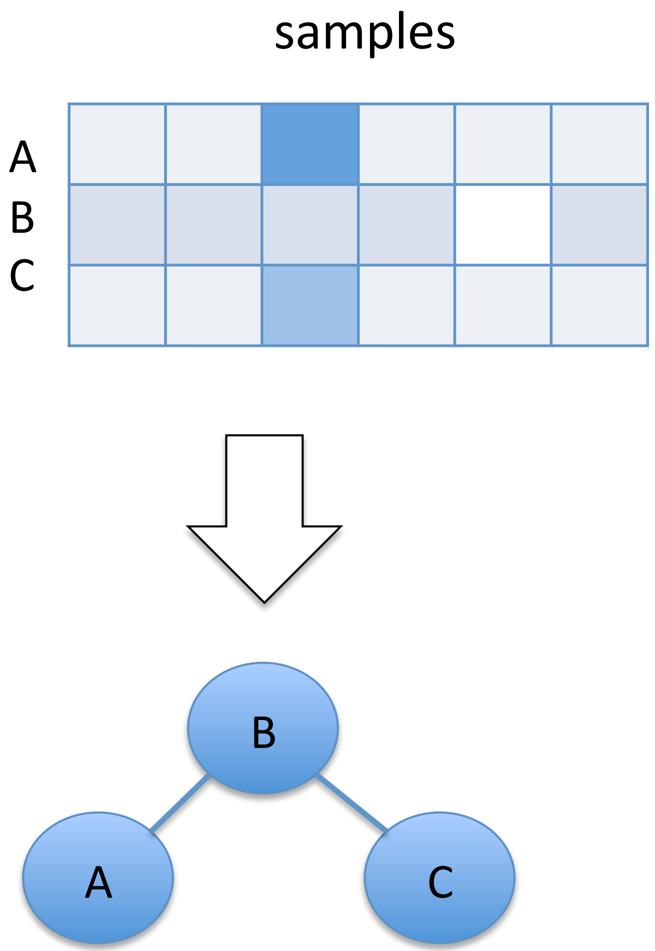


co-occurrence/correlation



1. Background

Co-occurrence analysis is a network inference technique

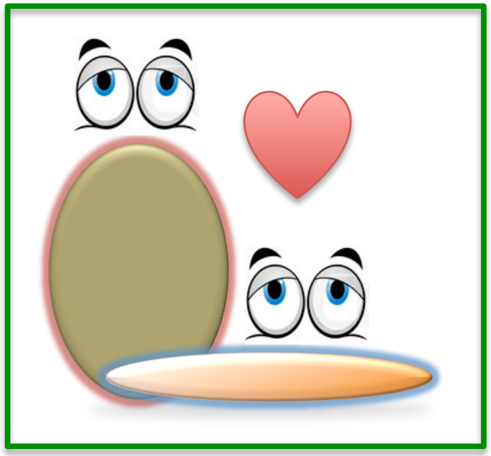


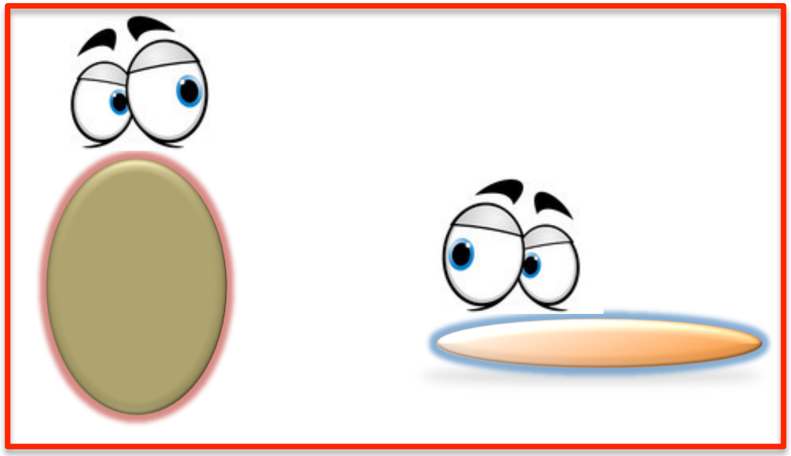
- **network inference (reverse engineering)**: the problem of finding relationships between objects (genes, proteins, metabolites, species...) whose presence/absence or abundance was observed repeatedly
- heavily used in genomics (gene regulatory network inference)

1. Background

Microbial association network inference

- several recent metagenomic data sets measure microbial abundance across a large number of samples
- network inference techniques can identify significant relationships between microorganisms from these data
- significant co-presence (co-occurrence of two microbes across samples) can be interpreted as niche overlap, mutualism, commensalism etc.
- significant mutual exclusion (avoidance of two microbes across samples) can be interpreted as alternative niche preference, competition, amensalism etc.





Network inference technique used in metagenomics and their problems

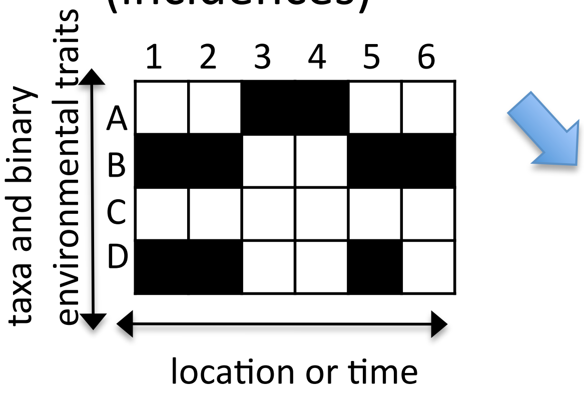
- Which network inference techniques are commonly used in metagenomics and which are their problems?

2. Methods and their problems

Similarity-based network inference

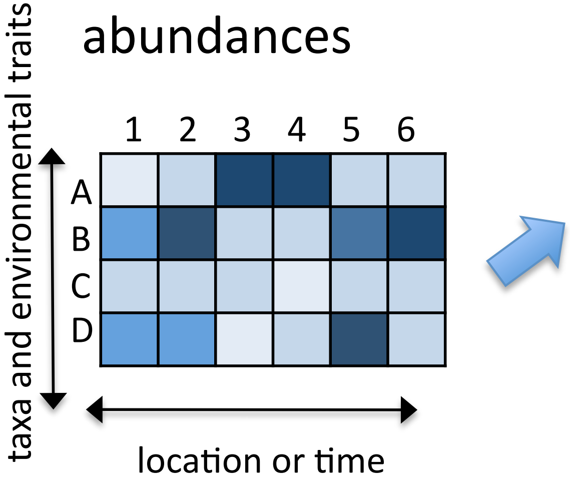
INPUT

presences/absences (incidences)



	1	2	3	4	5	6
A	□	□	■	■	□	□
B	■	■	□	□	■	■
C	□	□	□	□	□	□
D	■	■	□	□	■	□

abundances

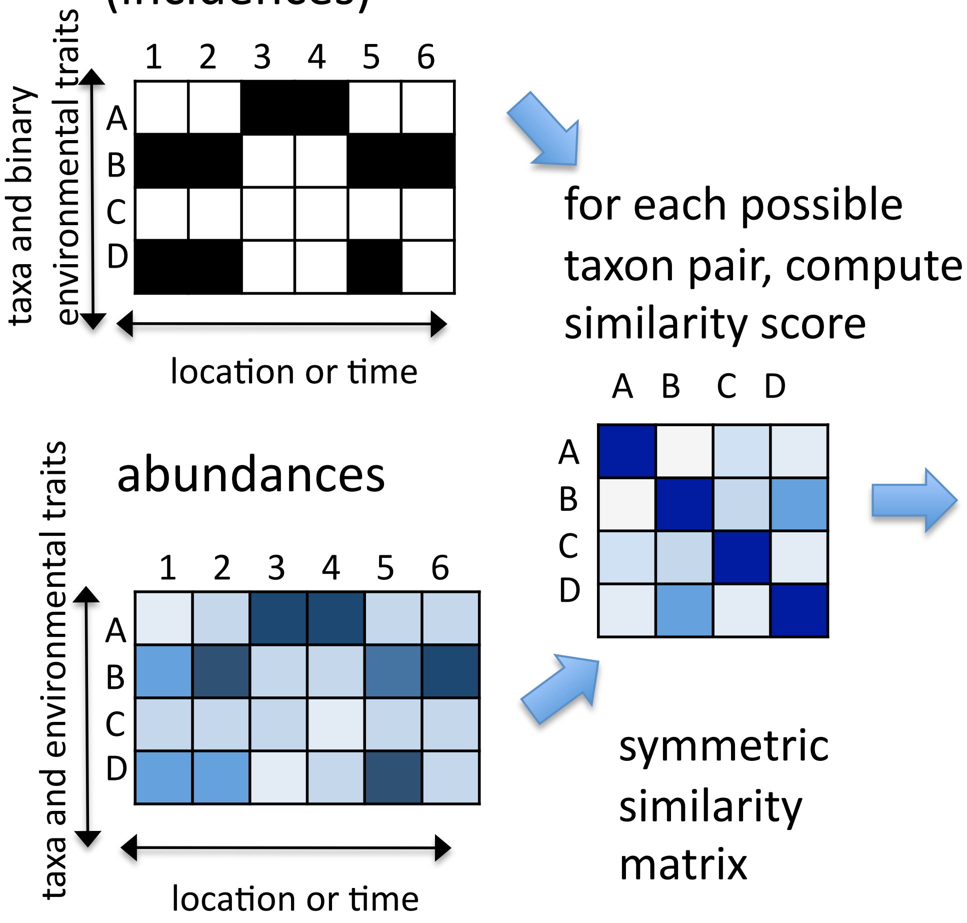


Similarity-based network inference

INPUT

SCORING

presences/absences (incidences)


taxa and binary environmental traits
1
2
3
4
5
6
A
B
C
D
location or time
taxa and environmental traits
1
2
3
4
5
6
A
B
C
D
location or time
A
B
C
D
A
B
C
D

	1	2	3	4	5	6
A	□	□	■	■	□	□
B	■	■	□	□	■	■
C	□	□	□	□	□	□
D	■	■	□	□	■	□

abundances

for each possible taxon pair, compute similarity score

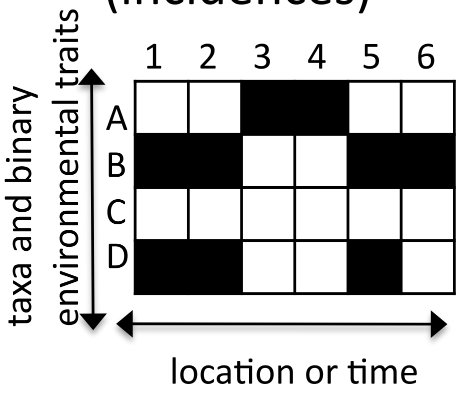
symmetric similarity matrix

2. Methods and their problems

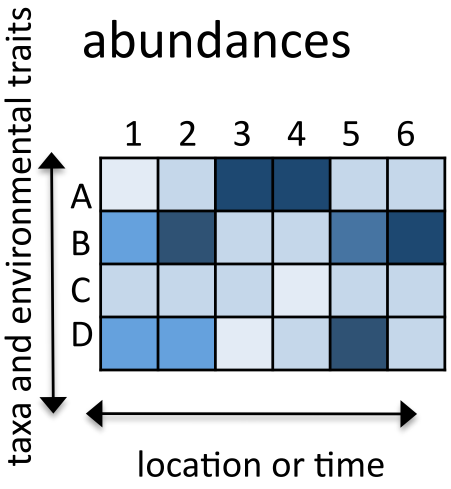
Similarity-based network inference

INPUT

presences/absences (incidences)

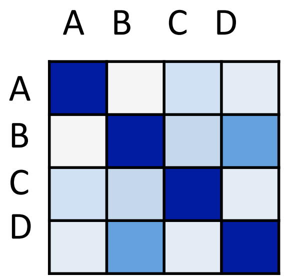


abundances



SCORING

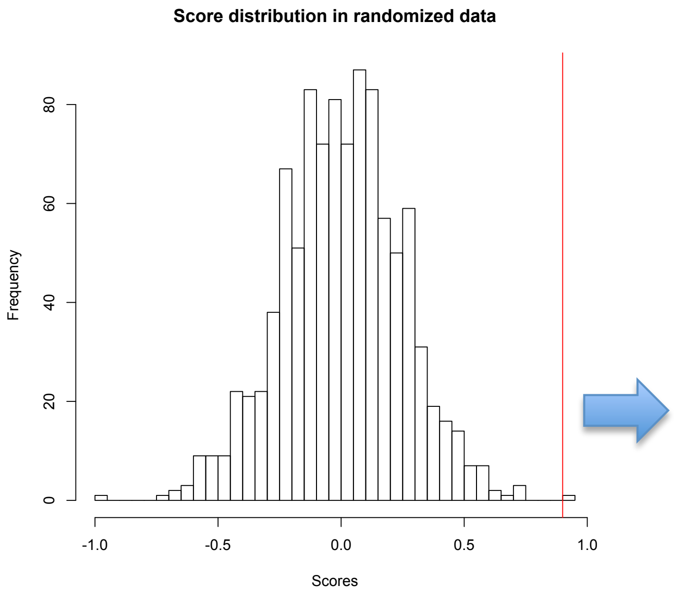
for each possible taxon pair, compute similarity score



symmetric similarity matrix

ASSESSMENT OF SIGNIFICANCE

repeat scoring step many times with randomized data



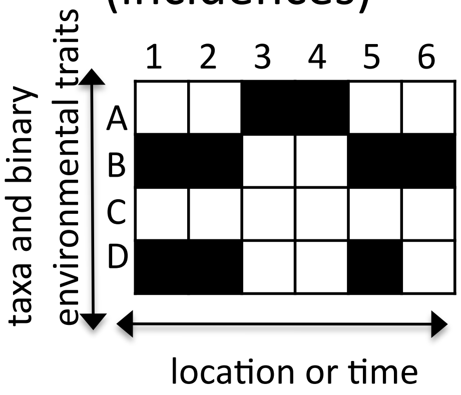
calculate p-values from the random score distribution and discard relationships with p-values above a specified threshold

2. Methods and their problems

Similarity-based network inference

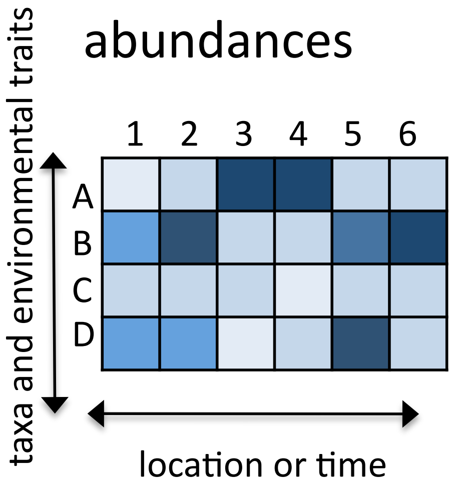
INPUT

presences/absences (incidences)



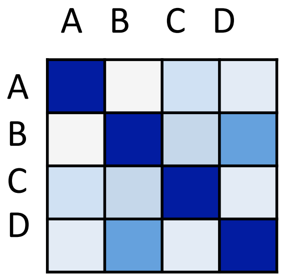
	1	2	3	4	5	6
A	□	□	■	■	□	□
B	■	■	□	□	■	■
C	□	□	□	□	□	□
D	■	■	□	□	■	□

abundances



SCORING

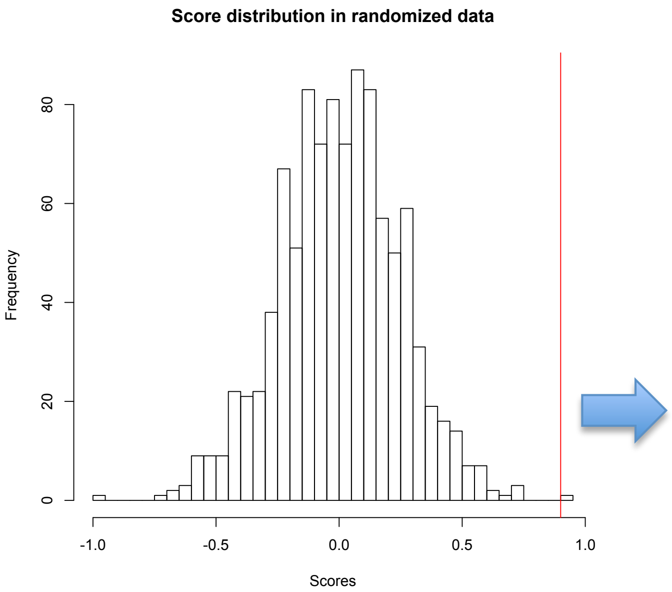
for each possible taxon pair, compute similarity score



symmetric similarity matrix

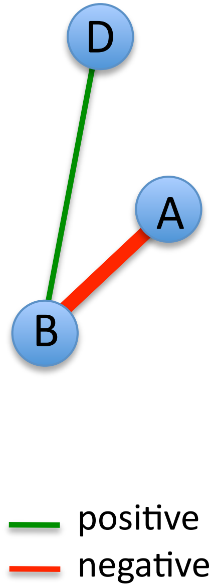
ASSESSMENT OF SIGNIFICANCE

repeat scoring step many times with randomized data



calculate p-values from the random score distribution and discard relationships with p-values above a specified threshold

VISUALIZATION



visualize taxon pairs with significant scores as a network

2. Methods and their problems

Problems of correlation measures – double zeros

- metagenomics data are sparse, i.e. many entries are zero
- when computing a correlation for a vector pair with matching zeros, a high score may result
- Example:

OTU1	303	221	998	826	915	160	924	831	408	795
OTU2	264	172	529	817	576	870	823	533	696	798

- these numbers were sampled from the uniform distribution, with minimum set to 0 and maximum to 1000
- Pearson correlation = 0.33 , Spearman correlation = 0.02 (p-value = 0.97)

2. Methods and their problems

Problems of correlation measures – double zeros continued

- a zero is ambiguous, since it is never clear whether a taxon is really absent or just below detection limit
- rare taxa might co-occur in deeply sequenced samples, but nowhere else
- we should therefore avoid giving a high similarity score to a taxon pair on the basis of their co-absences
- Example: 5 double-zero pairs added to previous vectors

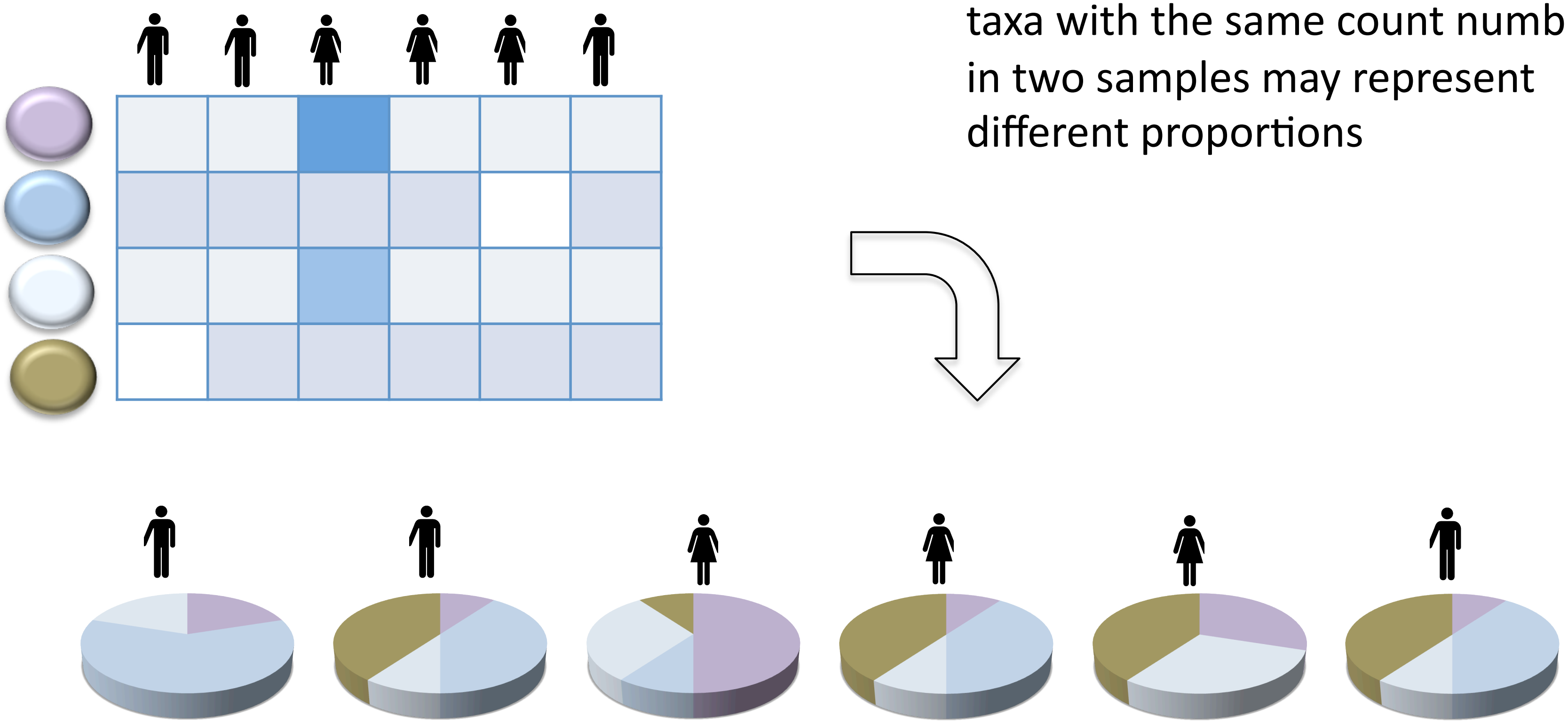
OTU1	303	221	998	826	915	160	924	831	408	795	0	0	0	0	0
OTU2	264	172	529	817	576	870	823	533	696	798	0	0	0	0	0

- Pearson correlation = 0.76, Spearman correlation = 0.7 (p-value = 0.004)

2. Methods and their problems

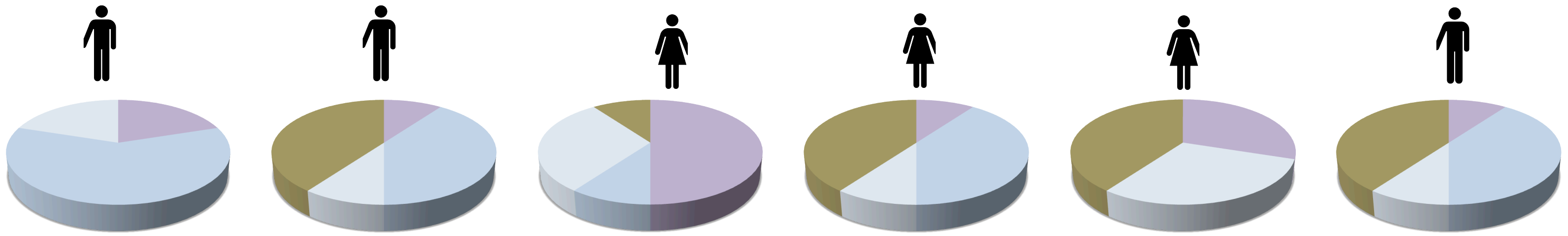
Problems of correlation measures - compositionality

- differences in sequencing depth lead to different total read counts across samples
- sample-wise normalization necessary (e.g. by dividing counts in a sample by the sample count sum or by rarefaction to a common sequencing depth)
- counts are converted into proportions



taxa with the same count number in two samples may represent different proportions

Problems of correlation measures - compositionality

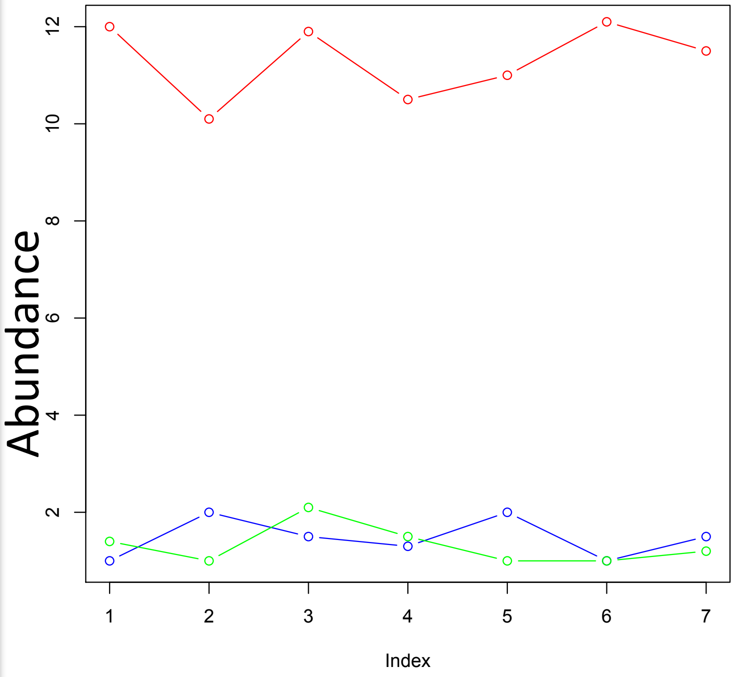


- Pearson and Spearman can be severely distorted, because they consider “absolute” values
- measures based on ratios or log-ratios (KLD, BC) are not affected by data compositionality, since the ratio between two counts in the same sample is not changed by the normalization

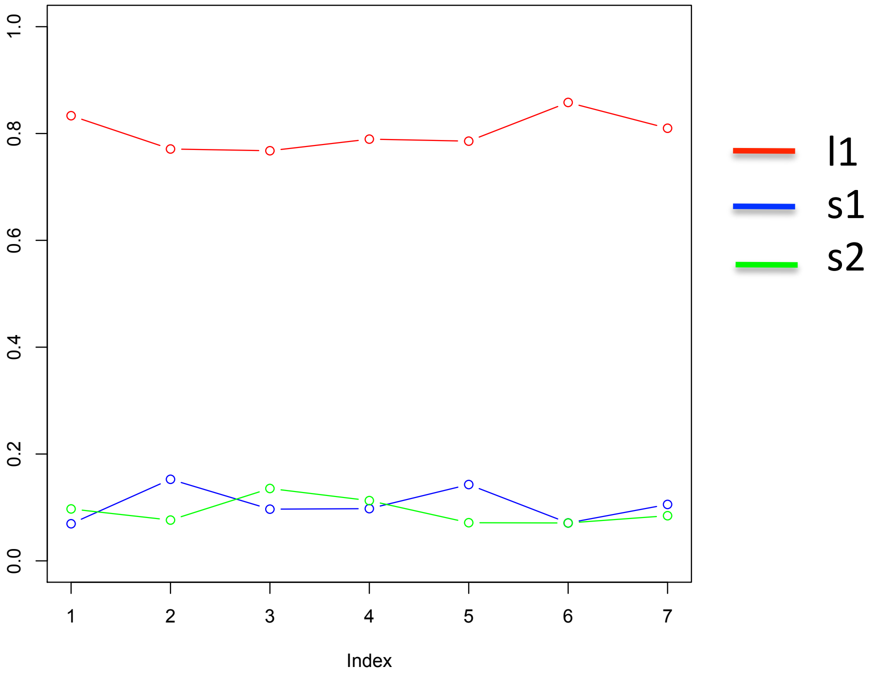
Aitchison J (1982) “The Statistical Analysis of Compositional Data.” Journal of the Royal Statistical Society Series B (Methodological) 44, 139-177.

2. Methods and their problems

Problems of correlation measures - compositionality



normalization



	s1	s2	l1
s1	1	-0.24	-0.69
s2		1	0.31
l1			1

	s1	s2	l1
s1	1	-0.32	-0.73
s2		1	-0.41
l1			1

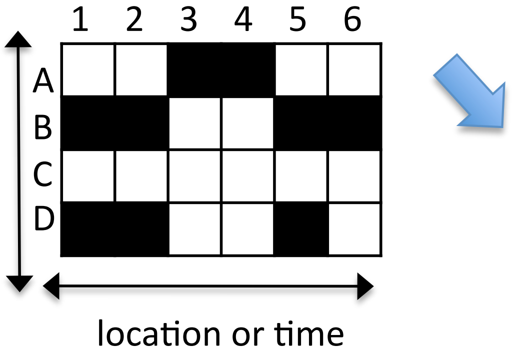
Pearson correlation

2. Methods and their problems

Regression-based network inference

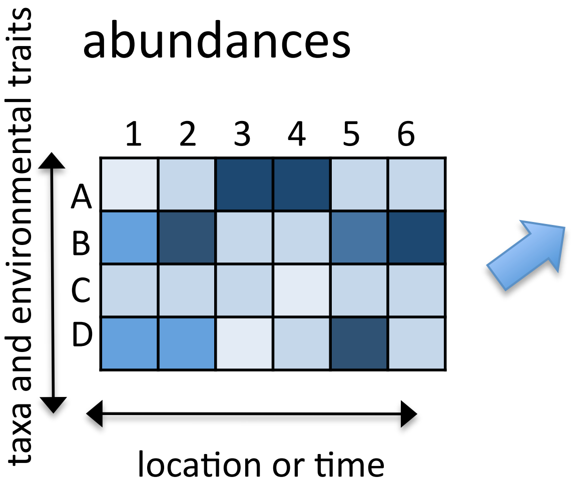
INPUT

presences/absences (incidences)

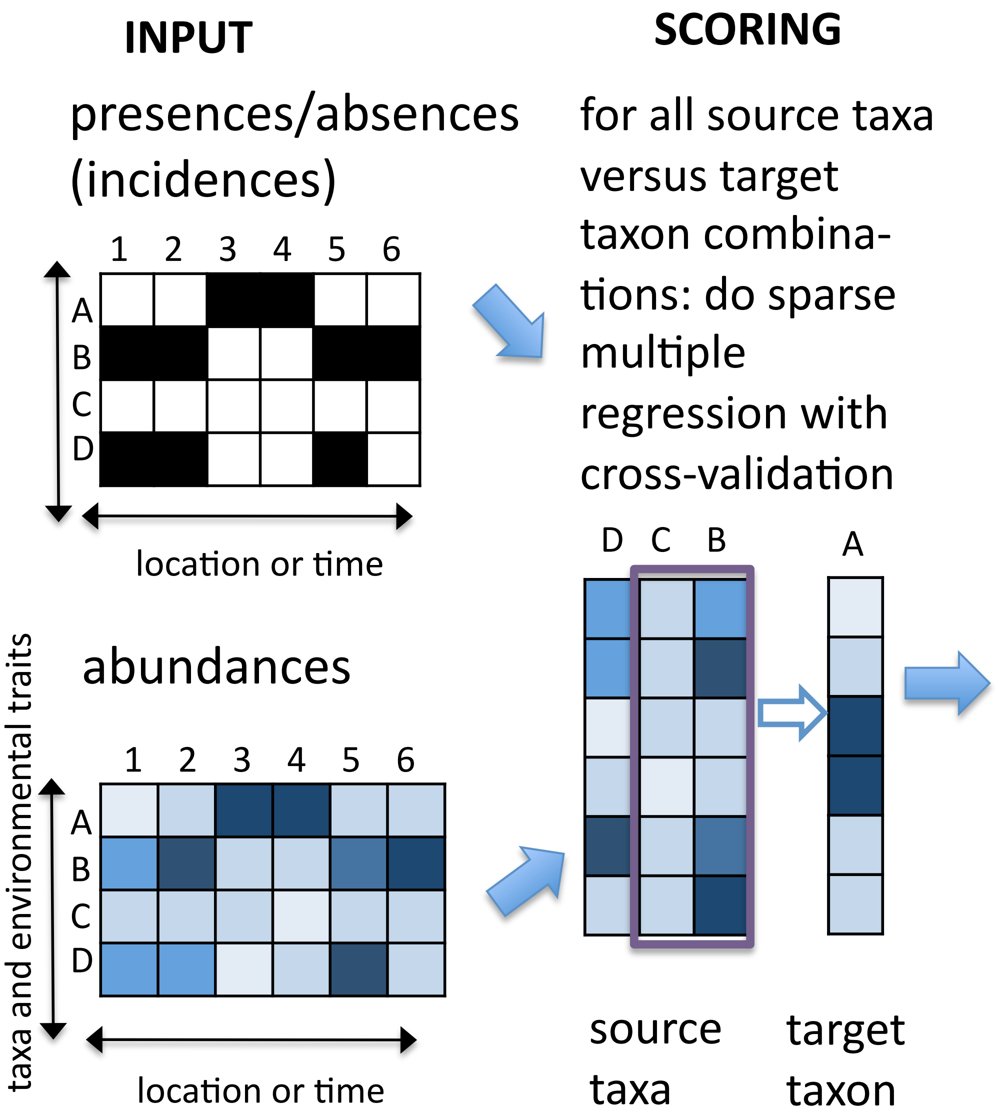


	1	2	3	4	5	6
A	□	□	■	■	□	□
B	■	■	□	□	■	■
C	□	□	□	□	□	□
D	■	■	□	□	■	□

abundances



Regression-based network inference



INPUT

presences/absences (incidences)

abundances

SCORING

for all source taxa versus target taxon combinations: do sparse multiple regression with cross-validation

source taxa

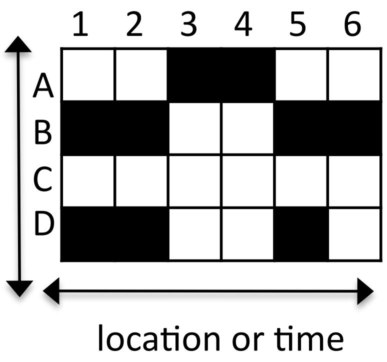
target taxon

2. Methods and their problems

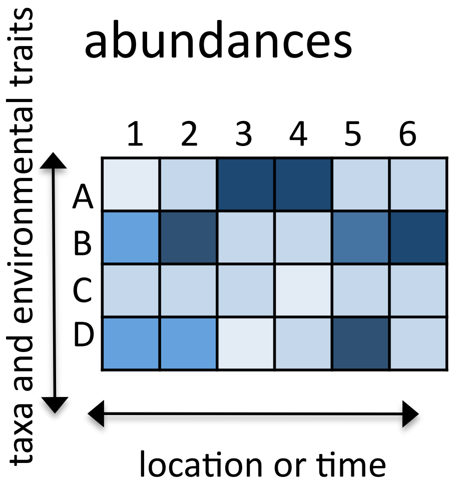
Regression-based network inference

INPUT

presences/absences (incidences)

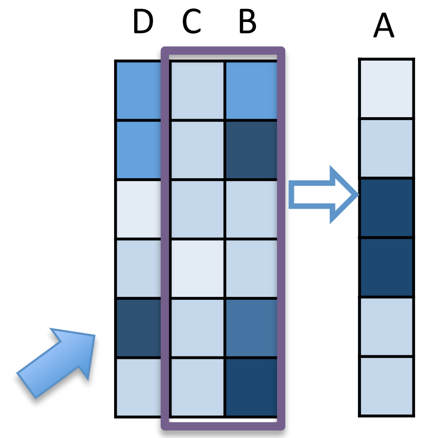


abundances



SCORING

for all source taxa versus target taxon combinations: do sparse multiple regression with cross-validation

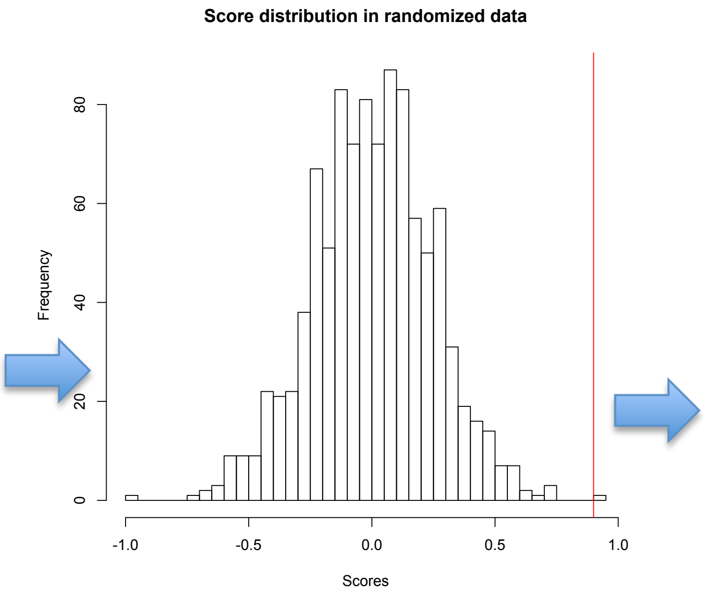


source taxa

target taxon

ASSESSMENT OF SIGNIFICANCE

repeat scoring step many times with randomized data

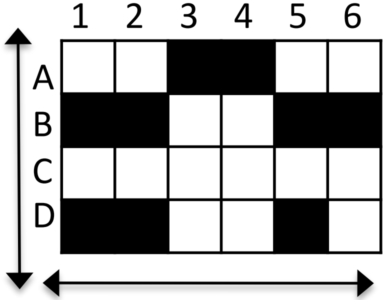


calculate p-values from the random score distribution and discard relationships with p-values above a specified threshold

Regression-based network inference

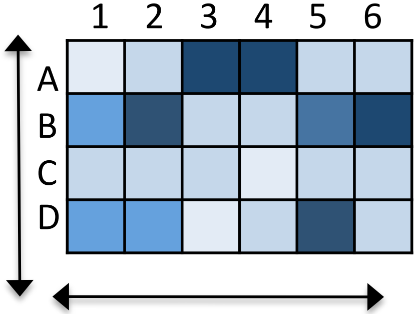
INPUT

presences/absences (incidences)



location or time

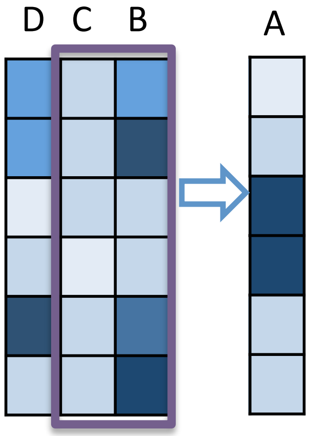
abundances



location or time

SCORING

for all source taxa versus target taxon combinations: do sparse multiple regression with cross-validation

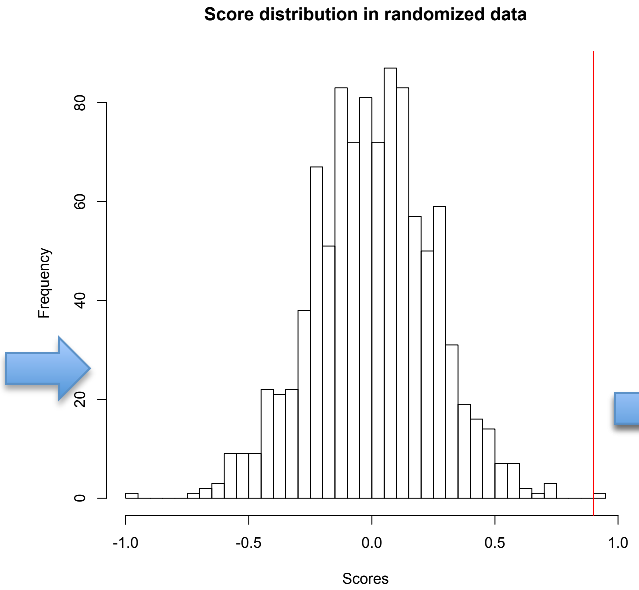


source taxa

target taxon

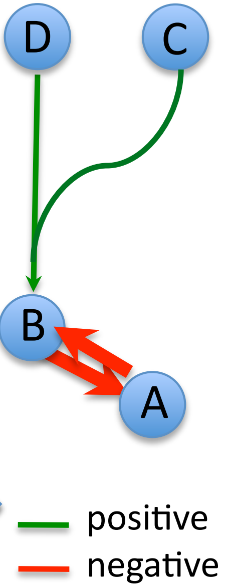
ASSESSMENT OF SIGNIFICANCE

repeat scoring step many times with randomized data



calculate p-values from the random score distribution and discard relationships with p-values above a specified threshold

VISUALIZATION



visualize taxon sets with significant scores as a network (hyper-graph)

2. Methods and their problems

Benefits and problems of regression

- + multiple regression can detect relationships between more than 2 taxa
- + it can predict directed edges and therefore asymmetric relationships (such as commensalism)
- – false positives are more likely with each additional source taxon considered (triplets give many more combinations than pairs do) – need to apply harsher multiple testing correction
- – risk of over-fitting if too many source taxa are considered
- – visualization is difficult (hyperedges)

⇒ recommendation: set sparsity constraint (feature selection) such that only a small number of source taxa is selected for each target taxon

Network inference tools used in metagenomics

- Which tools are out there and how do they work?

Robust correlations with SparCC

- basic idea: use the variance of log ratios (a distance measure robust to compositionality bias introduced by Aitchison)

$$D(x_i, x_j) = \operatorname{var}\left(\log\left(\frac{x_i}{x_j}\right)\right)$$

- the variance of log-ratios is not scaled, i.e. its maximum value is unknown a priori
- starting from the variance of log ratios, an approximation is developed to estimate correlations robustly

$$D(x_i, x_j) = \omega_i^{\,2} - \omega_j^{\,2} - 2\rho_{ij}\omega_i\omega_j$$

where ω is the variance of the log-transformed abundance vector and ρ the covariance

- SparCC estimates covariance ρ for all taxon pairs, assuming that most pairs are only weakly correlated

Friedman & Alm (2012) “Inferring Correlation Networks from Genomic Survey Data.” PLoS Comp Bio 8 (9), e1002687.
Aitchison (2003) “A concise guide to compositional data analysis” In: 2nd Compositional Data Analysis Workshop, Girona, Italy.

3. Tools

SparCC iterations and p-values

Iterations

- SparCC fits a Dirichlet distribution to the observed counts and estimates counts from this distribution
- taxon proportion estimation and robust correlation computation is iterated a number of times
- final correlation is reported as the median of this distribution

P-values

- counts are resampled with replacement for each OTU separately and averaged correlation values are re-computed for these bootstrapped counts
- p-values are computed from this bootstrap distribution as the proportion of bootstrapped correlations that are at least as large as the original correlation value

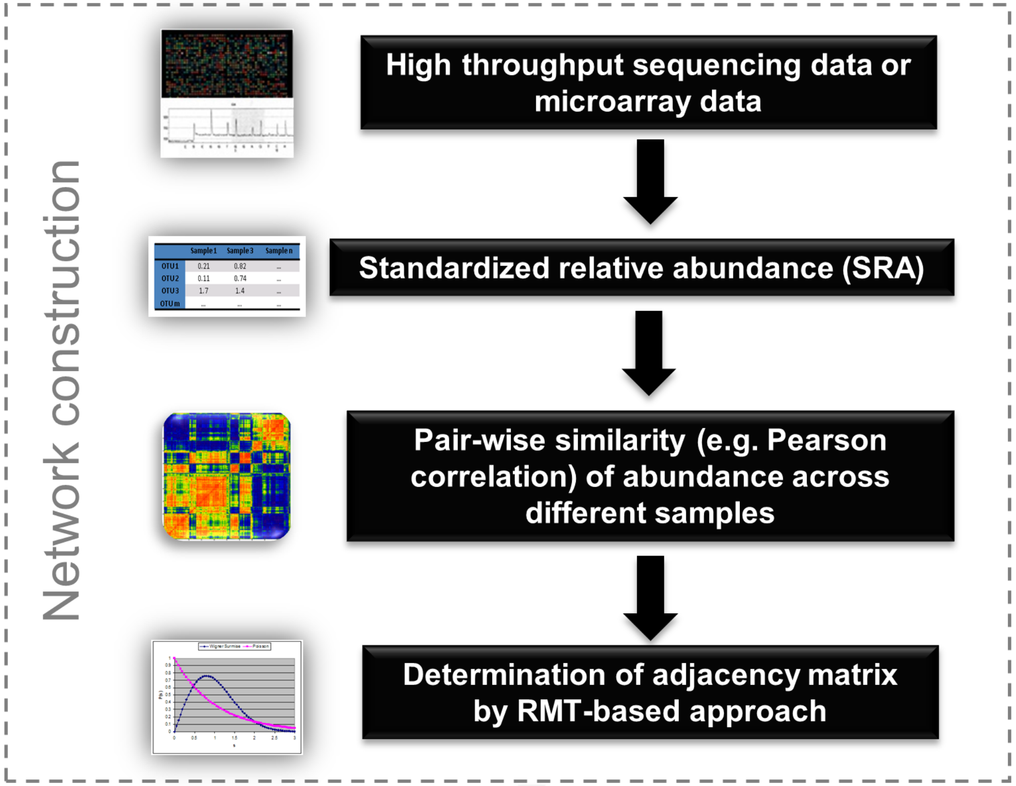
SparCC URL

- **https://bitbucket.org/yonatanf/sparcc** (requires basic python skills)

3. Tools

Alternative threshold computation with MENA

- MENA = molecular ecological network analysis pipeline
- server online: **http://ieg2.ou.edu/MENA/**



High throughput sequencing data or microarray data

Standardized relative abundance (SRA)

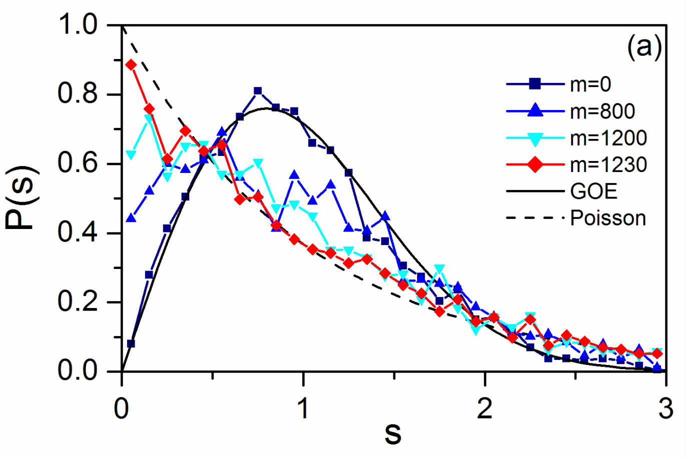
Pair-wise similarity (e.g. Pearson correlation) of abundance across different samples

Determination of adjacency matrix by RMT-based approach

Deng et al. (2012) “Molecular ecological network analyses” BMC Bioinformatics 13, 113.
Zhou et al. (2010) “Functional Molecular Ecological Networks.” mBio 1 (4), e0016910.

MENA's RMT approach

- RMT = random matrix theory
- compute eigenvalue spacing distribution of the Pearson correlation matrix for a given threshold
- do the same for a whole range of thresholds
- retain the threshold where distribution changes from Gaussian to Poisson
- keep all correlations above the threshold



probability of finding eigenvalues with the given spacing

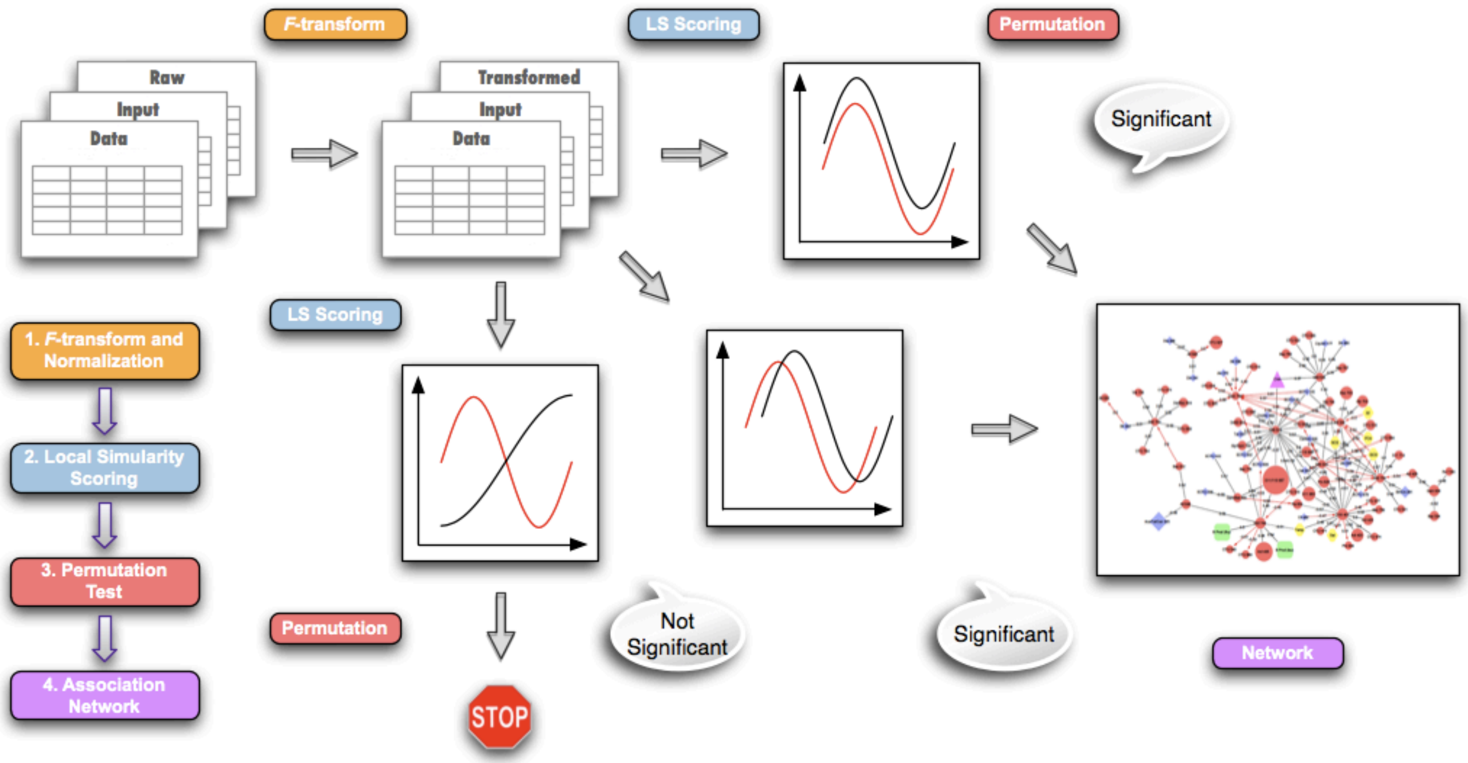
eigenvalue spacing

Lagged time series with LSA

- LSA = local similarity analysis
- command line tool: **http://hallam.microbiology.ubc.ca/fastLSA/install/index.html**
- command line tool: **http://meta.usc.edu/softs/lsa/**
- detects (local) similarity between potentially shifted (lagged) time series
- because it considers lags, LSA returns directed edges (A is shifted with respect to B) as well as undirected edges (A and B are not shifted)
- popular tool in marine and lake metagenomics

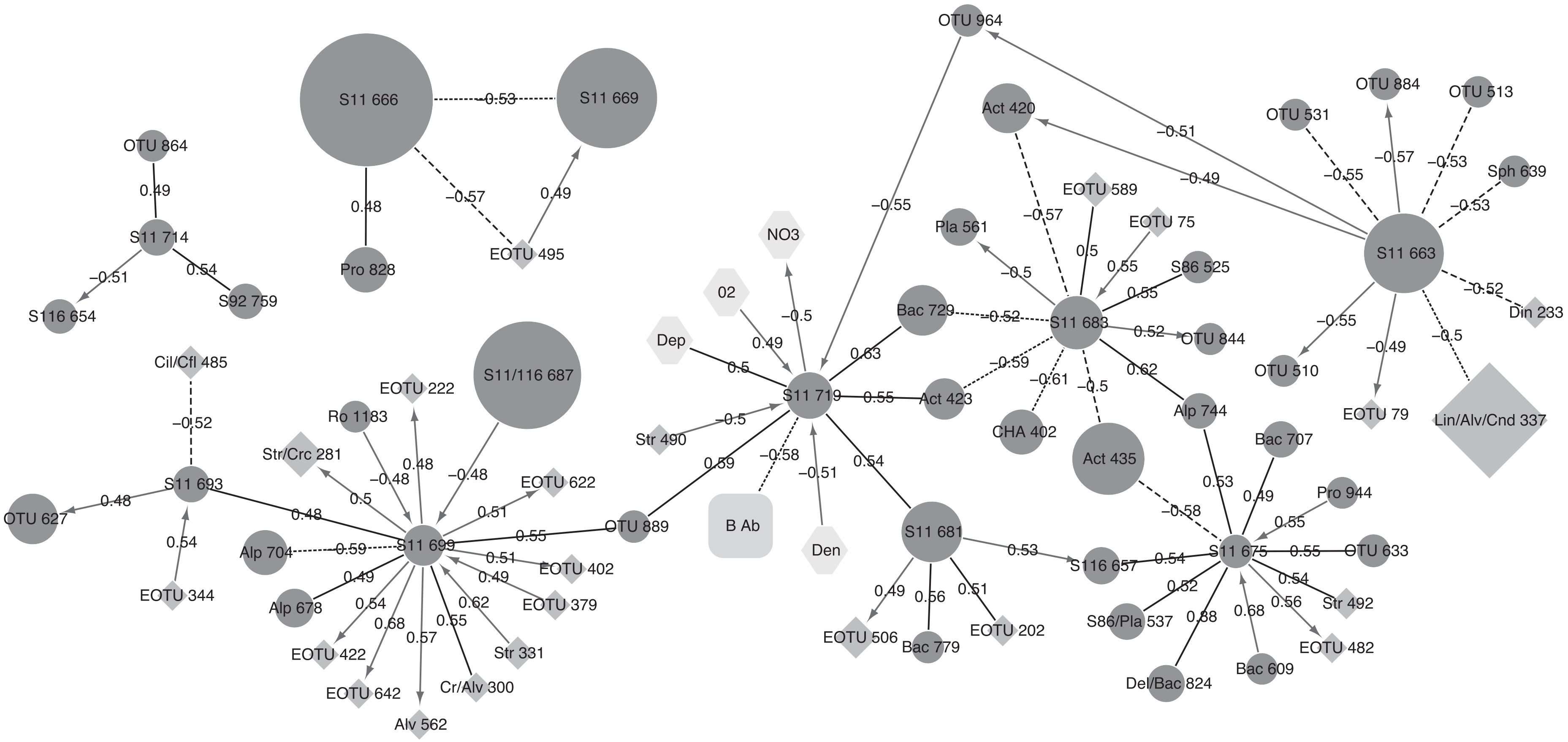
Xia et al. (2013) “Efficient statistical significance approximation for local similarity analysis of high-throughput time series data” Bioinformatics 29 (2), 230-237.
Durno et al. (2013) “Expanding the boundaries of local similarity analysis” BMC Genomics 14 (1), S3.
Xia et al. (2011) “Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates.” BMC Systems Biology 5 (2), S15.
Ruan et el. (2006) “Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors” Bioinformatics 22 (20), 2532-2538.

LSA: pipeline



- time series are transformed to be normally distributed
- for each pair of time series, local similarity score is computed with the dot product using dynamic programming (allowing up to 3 gaps)
- local similarity score is divided by length of time series
- permutation is carried out to assess the significance of the local similarity score
- p-value from permutation is multiple-test corrected (Bonferroni)
- network is constructed from edges with significant similarity scores

LSA: example network

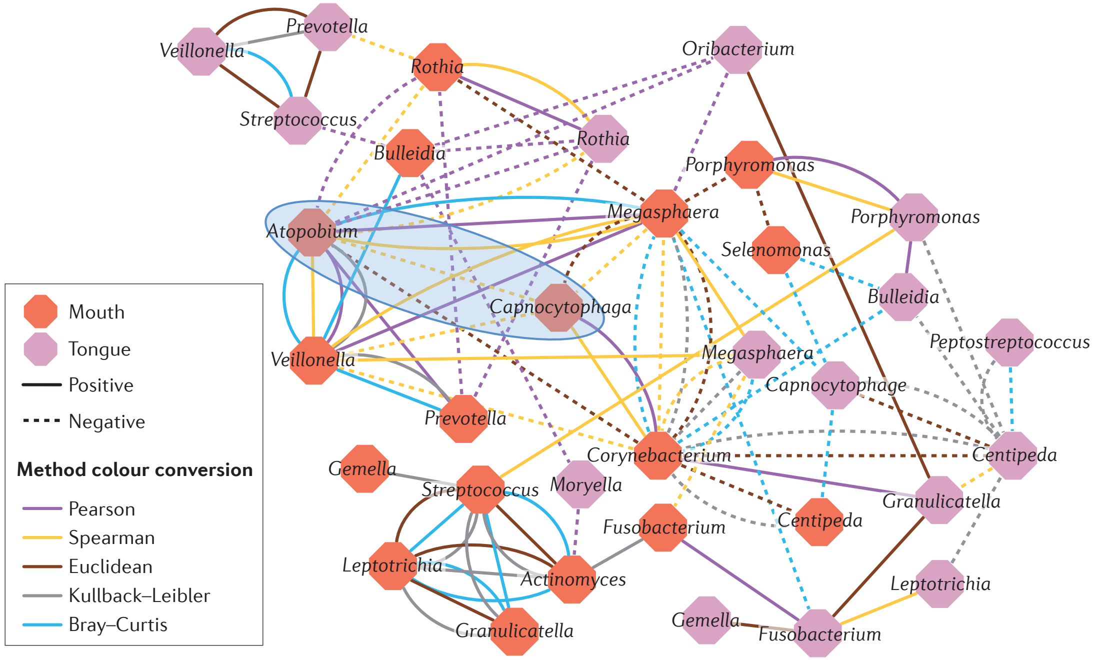


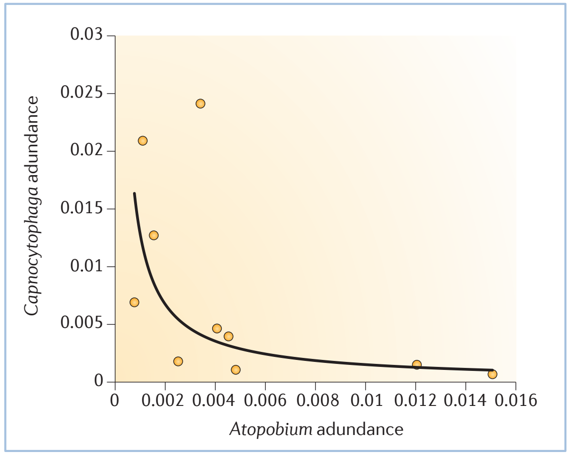
OTUs and chemical properties from water in the southern California coast (sub-surface chlorophyll maximum layer), sampled from August 2000 to March 2004

Steele et al. (2011) “Marine bacterial, archaeal and protistan association networks reveal ecological linkages” ISME 5, 1414-1425.

Ensemble-based network inference with CoNet

- different measures (Pearson, Spearman, Bray Curtis, ...) capture different types of relationships, but they converge when thresholds are increased
- idea of ensemble: measures make different mistakes, but tend to agree on correct result, so combine them





non-linear relationship is missed by Pearson

Faust & Raes (2012) “Microbial interactions: from networks to models.” Nature Reviews Microbiology 10 (8), 538-550.

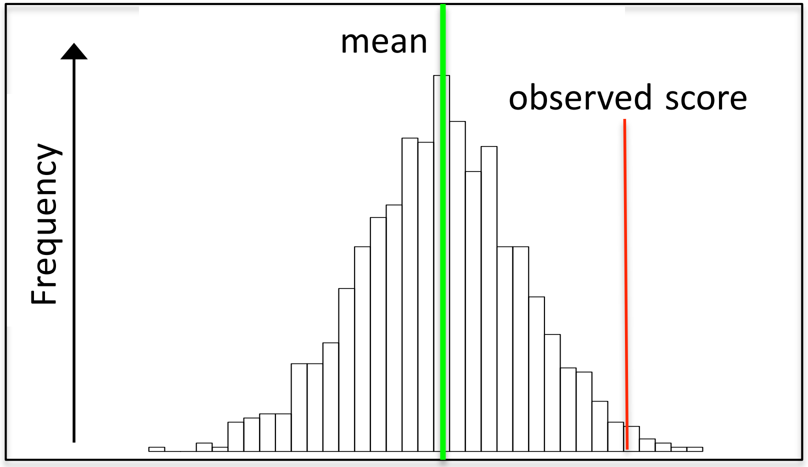
CoNet: Assessing significance

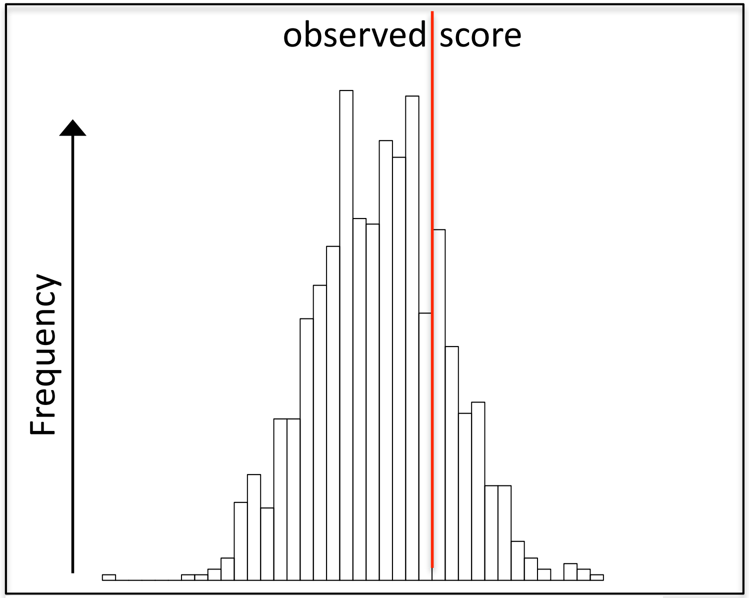
in collaboration with Fah Sathirapongsasuti and Curtis Huttenhower

- for each edge and each of the selected measures, compute permutation and bootstrap distributions

permutation (**null**) **distribution** of method-specific edge score

bootstrap distribution of method-specific edge score (**confidence interval**)



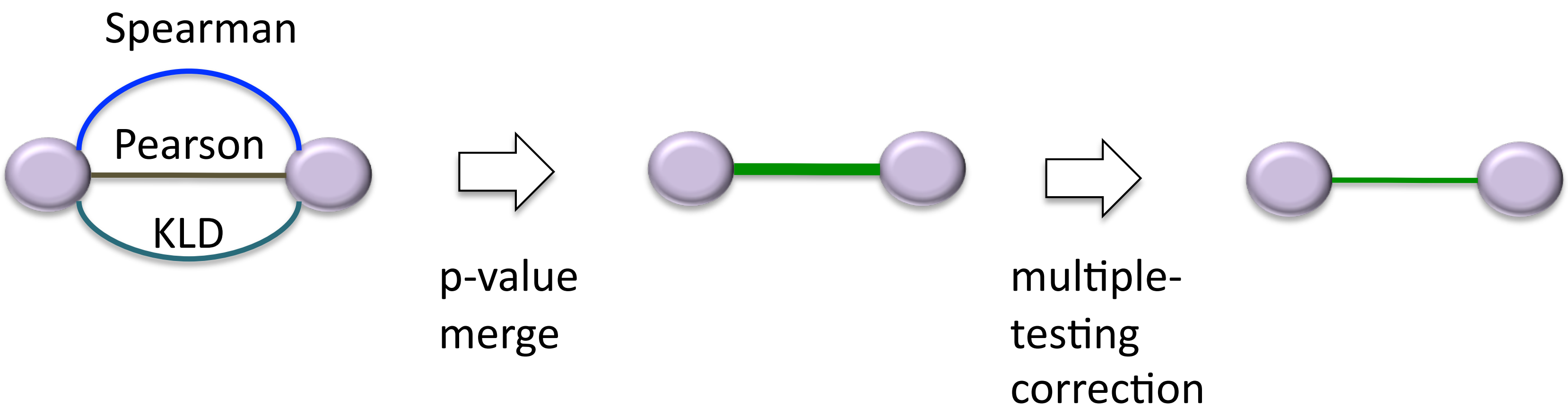


final method-specific p-value is computed as the probability of the null value (mean of the null distribution) under the bootstrap distribution

CoNet: Merging measures

multigraph (network with potentially more than 1 edge between a node pair)

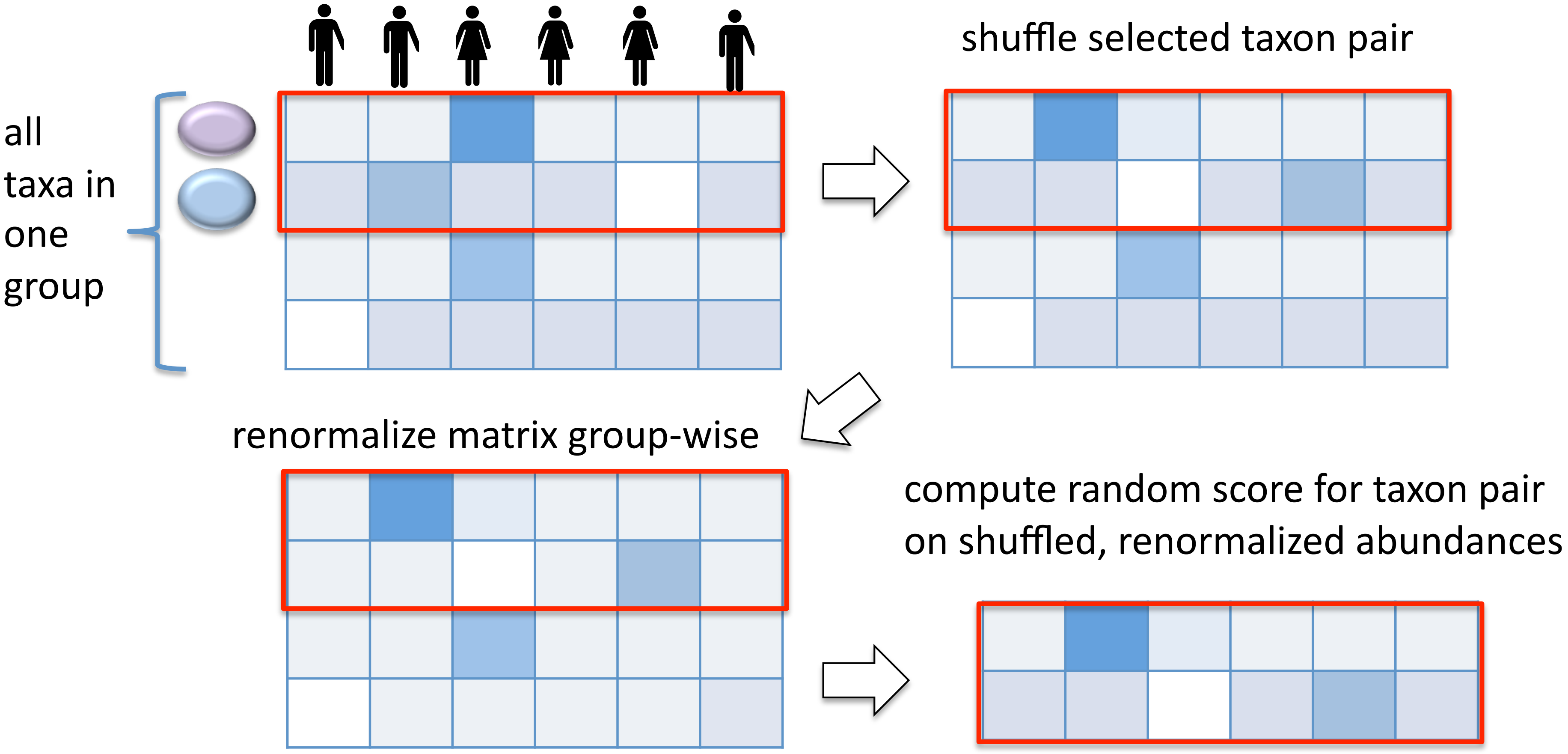
final graph



- measure-specific p-values are merged (e.g. using Fisher's, Brown's or Sime's method)
- merged p-values are corrected for multiple testing (e.g. with Benjamini-Hochberg)
- all edges with p-values above a given threshold are discarded (by default 0.05)

CoNet: Dealing with compositionality

permutation with renormalization (**ReBoot**)

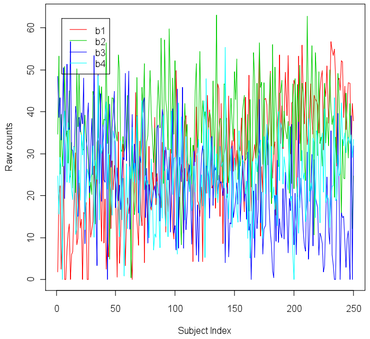

all taxa in one group
shuffle selected taxon pair
renormalize matrix group-wise
compute random score for taxon pair on shuffled, renormalized abundances

Fah Sathirapongsasuti

CoNet: Dealing with compositionality

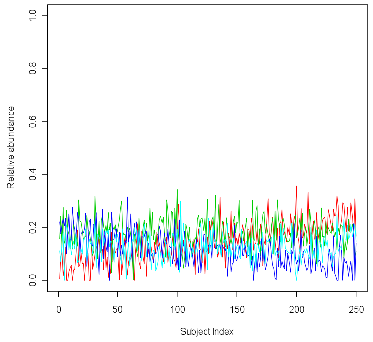
- Permutation test: removes correlation, but also any bias due to compositionality
- Permutation with **renormalization**: shifts null distribution

raw data



true anti-correlation between b1 and b3

normalized data

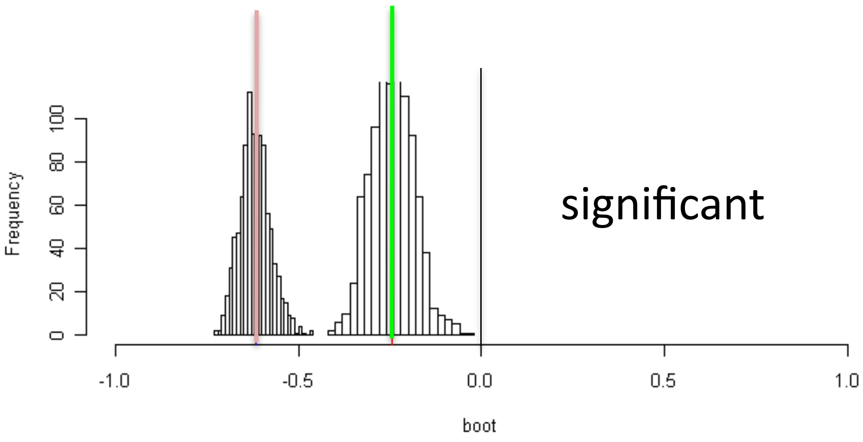


spurious correlation between b2 and b4 introduced by normalization

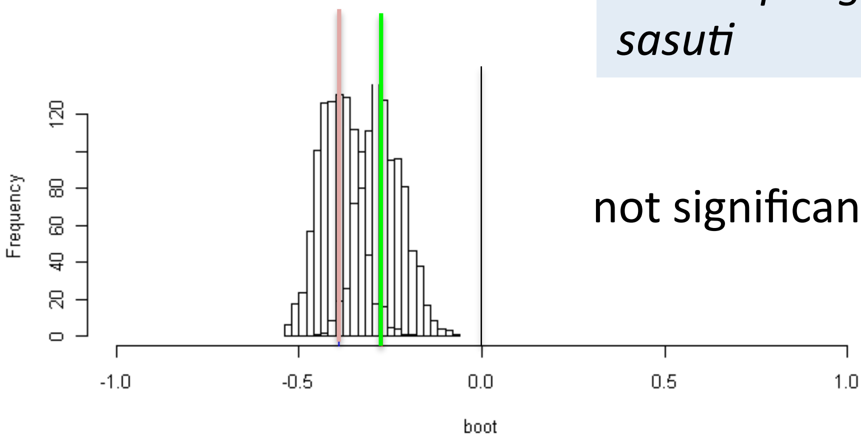
▬ bootstrap distribution mean

🟩 renormalized permutation distribution mean

Fah Sathirapong-sasuti



significant

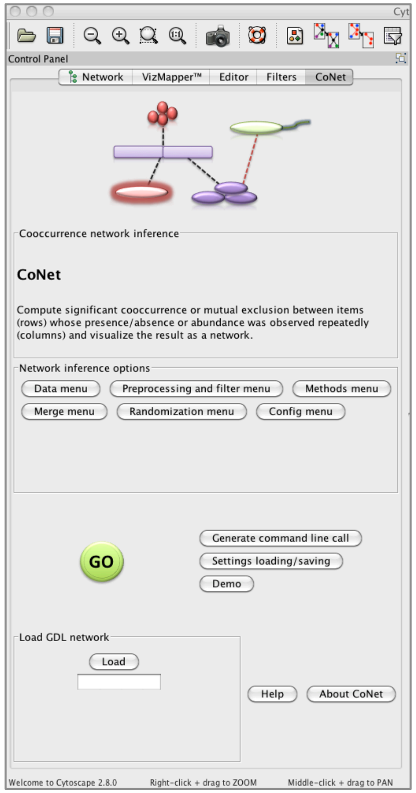


not significant

CoNet: Features

http://systemsbiology.vub.ac.be/conet

- runs as Cytoscape plugin or on command line
- suitable for abundance as well as presence/absence data
- supports QIIME OTU table format
- assigns higher-level taxa from lineages and computes correlations between them
- supports row groups
- supports environmental metadata
- integrates external network inference packages, e.g. minet (mutual information based network inference) and apriori (association rule mining algorithm)
- offers various preprocessing steps, filtering options and missing value treatment
- settings loading/saving
- score distribution plots
- well documented (manual, tutorials, FAQ)


Cyt
Control Panel
Network
VizMapper™
Editor
Filters
CoNet
Cooccurrence network inference

CoNet

Compute significant cooccurrence or mutual exclusion between items (rows) whose presence/absence or abundance was observed repeatedly (columns) and visualize the result as a network.

Network inference options
Data menu
Preprocessing and filter menu
Methods menu
Merge menu
Randomization menu
Config menu
GO
Generate command line call
Settings loading/saving
Demo
Load GDL network
Load
Help
About CoNet
Welcome to Cytoscape 2.8.0
Right-click + drag to ZOOM
Middle-click + drag to PAN

Conclusion on metagenomic network inference tools

- Which one to choose?
- hard to say without benchmark data – we need a database of annotated microbial interactions in specific environments
- on-going: in silico evaluation by a third party (Rob Knight lab)

Limitation of network inference tools

- (lagged) similarity != causality
- taxa can have significant similarity scores because they respond similarly to environmental variables (niche sharing) or because they interact (directly or indirectly) or both
- even if time series are not similar, taxa might interact (e.g. deterministic chaos)