NETWORK ANALYSIS TOOLS IN FOOD MICROBIAL ECOLOGY

Prof. Eugenio Parente
Dipartimento di Scienze
Università degli Studi della Basilicata

In this presentation

- What is network analysis?
  - definitions
  - what can I use it for?
  - node, edge and topological properties
  - (simple) tools for network visualization and analysis
  - suggested readings
- Host-parasite networks
  - bacteriophages of lactic acid bacteria
  - structure of phage-bacteria interaction matrices
  - PBIM in *S. thermophilus*
- OTU networks: analysis and visualisation of food microbial communities
  - molecular tools for the analysis of food microbial communities
  - FoodMicrobionet
  - OTU – sample networks
 - Co-occurrence networks
- Conclusions

Dynamic networks will not be considered here although they are a major topic in network science

Königsgberg bridges and Euler (1736)
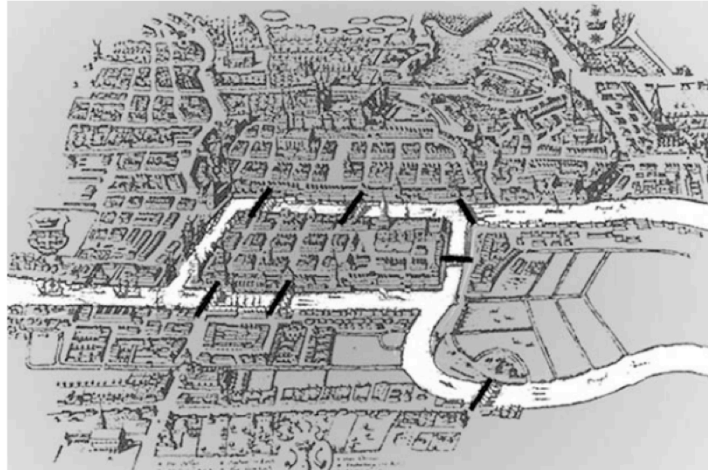
Esiste un modo per raggiungere tutte le masse di terra (vertici o nodi) attraversando ogni ponte (lato, margine, connessione) esattamente una volta? La risposta è no: perché questo sia possibile ci dovrebbero essere al massimo 2 nodi (quello iniziale e quello finale) con un numero dispari di connessioni; la formulazione di Eulero è "*Un qualsiasi grafo è percorribile se e solo se ha tutti i nodi di grado pari, o due di essi sono di grado dispari; per percorrere un grafo "possibile" con due nodi di grado dispari, è necessario partire da uno di essi, e si terminerà sull'altro nodo dispari.*"

Graph (and networks) terminology
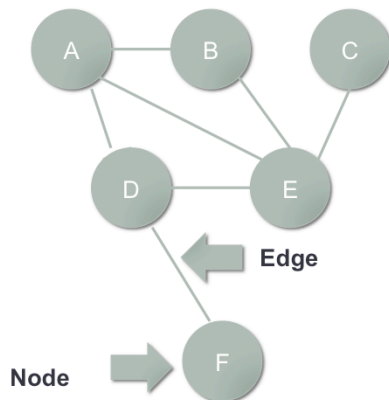
- A graph is an ordered pair of a set of nodes (or vertices) and edges (or links)
- G=(V,E)
- V={A,B,C,D,E,F}
- E={{A,B},{A,D},{A,E}, {B,E},{C,E},{D,E}, {D,F}}

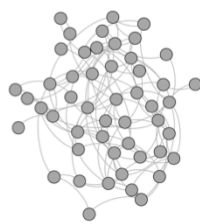In italiano Graph theory -> teoria dei grafi
Node, vertex ->nodo, vertice; Edge, link->archi (usato più spesso per i grafi diretti), lati, spigoli, cammini
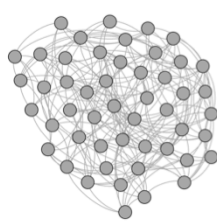
4

Models of networks: the random network

N=50, E/N=0.05

N=50, E/N=0.1

N=50, E/N=0.2

- defined by number of nodes and number of edges
- every node has the same probability p of being connected to another node (otherwise the network is said to be biased)
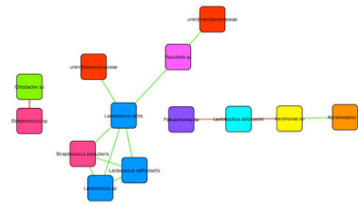- properties can be calculated (on average) analytically

random network of same size, but with different average degree, made with Gephi

Bipartite and unipartite undirected graphs
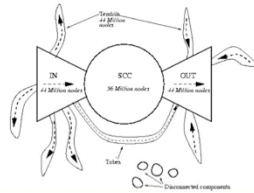
A bipartite OTU-Food sample undirected network

An unipartite, undirected microbial co-occurrence network
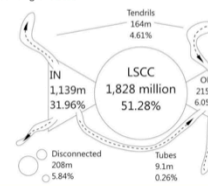
The world wide web as a directed graph

Boder et al., 2000

WDC Hyperlink graph 2012

- Balanced size of IN and OUT: 21%
- Size of LSCC: 27%

- IN much larger than OUT: 31% vs. 6%
- LSCC much larger: 51%

Size: 541 TB
Source: Common Crawl Foundation - http://commoncrawl.org
Created On: February 15, 2012
Last Updated: November 24, 2015

http://www.slideshare.net/bizer/graph-structure-in-the-web-revisited-www2014-web-science-track

Multigraphs

Here different edges, representing different methods for the evaluation of cooccurrence/coexclusion, are shown for each node; the graph is bipartite; two other special cases: a set of directed edges connecting two nodes in a multigraph is called a quiver; self-loops (i.e. edges connecting a node with itself) may be admissible or not; in some instances a disconnected graph (i.e. a graph with subnetworks which are not connected among them) can be generated

Datasets

- Wanna play? An extensive set of data is available at
  https://github.com/gephi/gephi/wiki/Datasets
- look at biological networks
  - **Diseasome**: a network of disorders and disease genes linked by known disorder–gene associations, indicating the common genetic origin of many diseases.
  - **Elegans neural network**: A directed, weighted network representing the neural network of C. *elegans*
  - **Yeast**: Protein-Protein interaction network in yeast
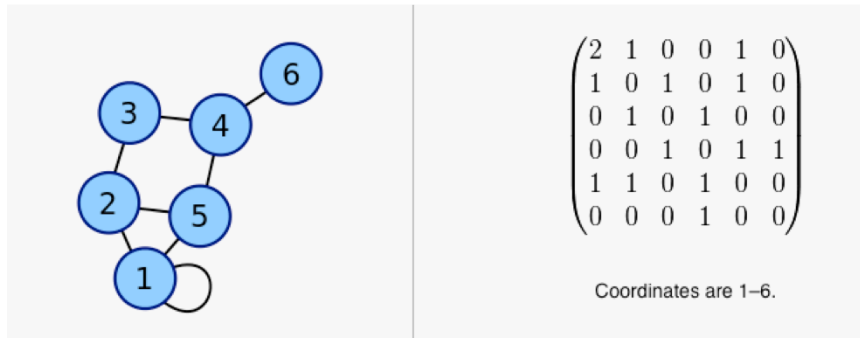
Data formats: the adjacency matrix

$$\begin{pmatrix} 2 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Coordinates are 1–6.

Figure from Wikipedia

Data formats: the adjacency matrix

Directed Cayley graph of S_4

Coordinates are 0–23.
As the graph is directed, the matrix is not symmetric.
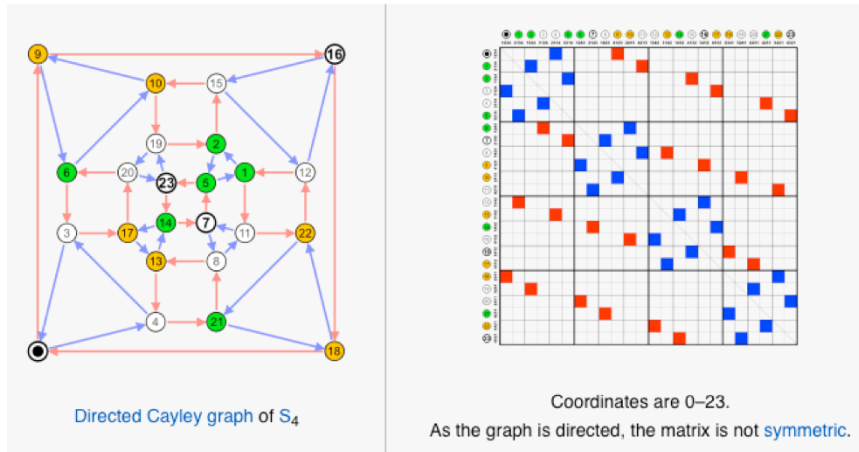
Figure from Wikipedia

Adjacency matrix for a bipartite graph

Host															Phage												
	p574_A	p575_A	p576_A	p577_B	p591_C	p596_D	p603_B	p604_E	p605_F	p607_G	p611_H	p616_I	p620_L	p641_M	p642_N	p654_O	p671_P	p1027_Q	p1028_Q	p1032_Q	p1033_R	p1034_R	p1036_R	p1040_S	p1041_S	p1042_S	DT1
CHCC3063	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0
CHCC2070	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
CHCC2130	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHCC2133	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
CHCC2134	1	0	1	0	0	1	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
CHCC2389	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1
CHCC3048	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CHCC3049	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
CHCC3050	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
CHCC3046	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
CHCC4323	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
CHCC4325	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
CHCC4327	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
CHCC4460	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
CHCC4895	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
CHCC6592	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
SMQ-301	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Wide format

An abundance table: wide format

Taxon	CL4.4	CL12.81	CL7.8	CL4.2	CL8.8	CL10.8	CL4.8
Root;Other;Other;Other;Other;Other	0.00023898	0	0.00021062	0	0	0.00095946	0
Root;k__Bacteria__Actinobacteria;c__Actinobacteria;f__Microbacteriaceae;g__Microbacterium;s__	0	0	0	0	0	0	0
Root;k__Bacteria__Actinobacteria;c__Actinobacteria;f__Microbacteriaceae;g__Pseudoclavibacter;s__Pseudoclavibacterhelvoli	0	0	0	0	0	0	0
Root;k__Bacteria__Bacteroidetes;c__Flavobacteria;f__Flavobacteriaceae;Other;Other	0	0	0	0	0	0	0
Root;k__Bacteria__Bacteroidetes;c__Flavobacteria;f__Flavobacteriaceae;g__Chryseobacterium;Other	0	0	0	0.00035323	0	0	0.0002173
Root;k__Bacteria__Bacteroidetes;c__Flavobacteria;f__Flavobacteriaceae;g__Chryseobacterium;s__	0	0.00043328	0	0	0	0	0
Root;k__Bacteria__Bacteroidetes;c__Flavobacteria;f__Flavobacteriaceae;g__Chryseobacterium;s__Chryseobacteriumbovis	0	0	0	0	0	0	0
Root;k__Bacteria__Bacteroidetes;c__Flavobacteria;f__Flavobacteriaceae;g__Flavobacterium;s__Flavobacteriumfrigidarium	0	0.00043328	0	0	0.00015552	0	0
Root;k__Bacteria__Bacteroidetes;c__Flavobacteria;f__Flavobacteriaceae;g__Haloanella;s__	0	0.00086655	0	0	0	0	0
Root;k__Bacteria__Bacteroidetes;c__Flavobacteria;f__Flavobacteriaceae;g__Myroides;s__	0	0	0	0	0.00015552	0	0
Root;k__Bacteria__Firmicutes;Other;Other;Other;Other	0	0	0	0	0	0	0
Root;k__Bacteria__Firmicutes;c__Bacilli;Other;Other;Other	0.00095591	0	0	0	0.00062208	0.00011993	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Aerococcaceae;g__;s__	0	0	0	0	0	0	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Carnobacteriaceae;g__Carnobacterium;s__	0.0031067	0.00129983	0.00021062	0	0	0.00059966	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Enterococcaceae;g__Enterococcus;Other	0	0	0	0	0.00015552	0.00011993	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Enterococcaceae;g__Enterococcus;s__Enterococcusfaecalis	0	0	0	0	0	0	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Lactobacillaceae;g__Lactobacillus;s__Lactobacillusdelbrueckii	0.00011949	0.59662045	0.8258214	0.34528435	0.86500778	0.68073879	0.85180356
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Leuconostocaceae;g__Leuconostoc;s__	0.00011949	0	0	0.0010597	0	0	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Listeriaceae;g__Brochothrix;s__	0	0.00043328	0.00021062	0	0.00015552	0	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Staphylococcaceae;g__Macrococcus;s__Macrococcuscaseolyticus	0	0	0	0	0	0	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Staphylococcaceae;g__Staphylococcus;Other	0.00035847	0	0.00063185	0.0010597	0.00077761	0.00059966	0.00043459
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Streptococcaceae;Other;Other	0.01278528	0.00259965	0.0168492	0.01660191	0.00559876	0.00047973	0.00282486
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Streptococcaceae;g__Lactococcus;Other	0.05281396	0.00086655	0.01032014	0.00088308	0.00404355	0.00143919	0.00369405
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Streptococcaceae;g__Lactococcus;s__	0.0063329	0.00043328	0.00042123	0.00017662	0.00062208	0.00179899	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Streptococcaceae;g__Lactococcus;s__Lactococcusgarvieae	0.00382364	0.00346621	0.01474305	0.00900742	0.01041991	0.00287839	0.00391134
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Streptococcaceae;g__Lactococcus;s__Lactococcuslactis	0.00143386	0	0.00400169	0.00229601	0.00139969	0.00011993	0.00065189
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Streptococcaceae;g__Streptococcus;Other	0	0	0.00042123	0.00141293	0	0.0003598	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Streptococcaceae;g__Streptococcus;s__Streptococcusparauberis	0.00561596	0.00129983	0.00379107	0.0044154	0.00155521	0.0010794	0.00130378
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Streptococcaceae;g__Streptococcus;s__Streptococcuspluranimalium	0	0	0	0	0	0	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Streptococcaceae;g__Streptococcus;s__Streptococcussuis	0.00071693	0	0	0	0.00015552	0.00059966	0
Root;k__Bacteria__Firmicutes;c__Bacilli;f__Streptococcaceae;g__Streptococcus;s__Streptococcusthermophilus	0.9057235	0.33838822	0.11120472	0.61691982	0.10326594	0.29779324	0.13276836

This is a dense table, with a lot of 0; the meaning of 0 values in unclear: they might mean absence or presence below the detection limit; this again would depend on the filtering options; were singleton and doubletons discarded? etc. If you want to run a correlation analysis rows with many 0 would inflate the correlation; different library sizes may seriously affect
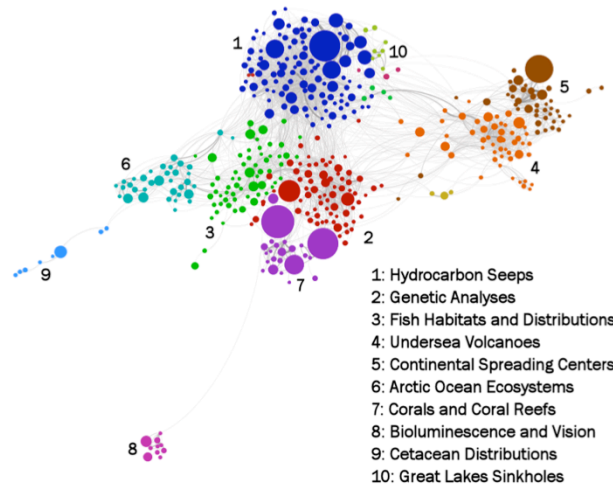
An edge table (long format)

	A	B	C	D	E	F	G	H	I	J
1	Source	Target	Type	Id	Label	Weight	Gtarget1	Gtarget2	Bioproject	Note
2	A1A	Root;k__;c__	Undirected	1		0.84033614	16S_DNA	V1-V3	SRP052240	--
3	A1A	Root;k__Bac	Undirected	2		0.05252101	16S_DNA	V1-V3	SRP052240	--
4	A1A	Root;k__Bac	Undirected	3		0.15756303	16S_DNA	V1-V3	SRP052240	--
5	A1A	Root;k__Bac	Undirected	4		0.05252101	16S_DNA	V1-V3	SRP052240	--
6	A1A	Root;k__Bac	Undirected	5		0.6565126	16S_DNA	V1-V3	SRP052240	--
7	A1A	Root;k__Bac	Undirected	6		0.0262605	16S_DNA	V1-V3	SRP052240	--
8	A1A	Root;k__Bac	Undirected	7		1.1554621	16S_DNA	V1-V3	SRP052240	--
9	A1A	Root;k__Bac	Undirected	8		0.18382353	16S_DNA	V1-V3	SRP052240	--
10	A1A	Root;k__Bac	Undirected	9		1.3392857	16S_DNA	V1-V3	SRP052240	--
11	A1A	Root;k__Bac	Undirected	10		0.0262605	16S_DNA	V1-V3	SRP052240	--
12	A1A	Root;k__Bac	Undirected	11		3.335084	16S_DNA	V1-V3	SRP052240	--
13	A1A	Root;k__Bac	Undirected	12		0.05252101	16S_DNA	V1-V3	SRP052240	--
14	A1A	Root;k__Bac	Undirected	13		0.07878152	16S_DNA	V1-V3	SRP052240	--
15	A1A	Root;k__Bac	Undirected	14		0.0262605	16S_DNA	V1-V3	SRP052240	--
16	A1A	Root;k__Bac	Undirected	15		85.05777	16S_DNA	V1-V3	SRP052240	--
17	A1A	Root;k__Bac	Undirected	16		0.0262605	16S_DNA	V1-V3	SRP052240	--
18	A1A	Root;k__Bac	Undirected	17		0.13130252	16S_DNA	V1-V3	SRP052240	--
19	A1A	Root;k__Bac	Undirected	18		0.0262605	16S_DNA	V1-V3	SRP052240	--
20	A1A	Root;k__Bac	Undirected	19		0.5514706	16S_DNA	V1-V3	SRP052240	--
21	A1A	Root;k__Bac	Undirected	20		0.05252101	16S_DNA	V1-V3	SRP052240	--
22	A1A	Root;k__Bac	Undirected	21		0.10504202	16S_DNA	V1-V3	SRP052240	--
23	A1A	Root;k__Bac	Undirected	22		0.0262605	16S_DNA	V1-V3	SRP052240	--
24	A1A	Root;k__Bac	Undirected	23		3.7289915	16S_DNA	V1-V3	SRP052240	--
25	A1A	Root;k__Bac	Undirected	24		0.0262605	16S_DNA	V1-V3	SRP052240	--
26	A1A	Root;k__Bac	Undirected	25		1.8644958	16S_DNA	V1-V3	SRP052240	--
27	A1A	Root;k__Bac	Undirected	26		0.15756303	16S_DNA	V1-V3	SRP052240	--
28	A1A	Root;k__Bac	Undirected	27		0.26260504	16S_DNA	V1-V3	SRP052240	--
29	A1B	Root;k__;c__	Undirected	28		0.67805123	16S_DNA	V1-V3	SRP052240	--
30	A1B	Root;k__Bac	Undirected	30		0.05022602	16S_DNA	V1-V3	SRP052240	--
31	A1B	Root;k__Bac	Undirected	31		0.12556504	16S_DNA	V1-V3	SRP052240	--
32	A1B	Root;k__Bac	Undirected	32		0.15067805	16S_DNA	V1-V3	SRP052240	--
33	A1B	Root;k__Bac	Undirected	33		0.75339025	16S_DNA	V1-V3	SRP052240	--
34	A1B	Root;k__Bac	Undirected	34		0.92918134	16S_DNA	V1-V3	SRP052240	--
35	A1B	Root;k__Bac	Undirected	35		0.9040683	16S_DNA	V1-V3	SRP052240	--

14
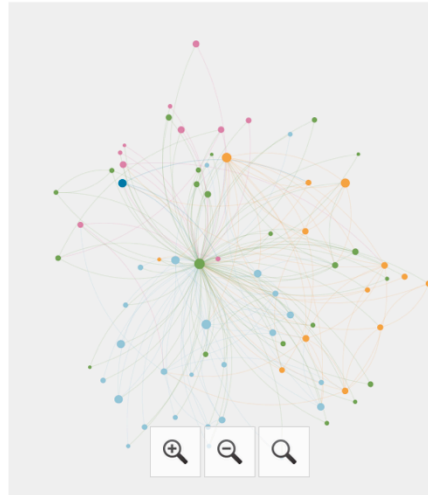
mappa degli account di twitter seguiti da @OIIOxford Oxford Internet Institute
University of Network http://oxfordinternetinstitute.github.io/InteractiveVis/
network/; in general in social network analysis
nodes are called actors and edges ties

Bibliographic networks

1: Hydrocarbon Seeps
2: Genetic Analyses
3: Fish Habitats and Distributions
4: Undersea Volcanoes
5: Continental Spreading Centers
6: Arctic Ocean Ecosystems
7: Corals and Coral Reefs
8: Bioluminescence and Vision
9: Cetacean Distributions
10: Great Lakes Sinkholes

http://www.lib.noaa.gov/bibliometrics/ This is a bibliographig coupling network; two papers are connected if both cite the same paper

Co-author networks (applied microbiology in Universities of Southern Italy)
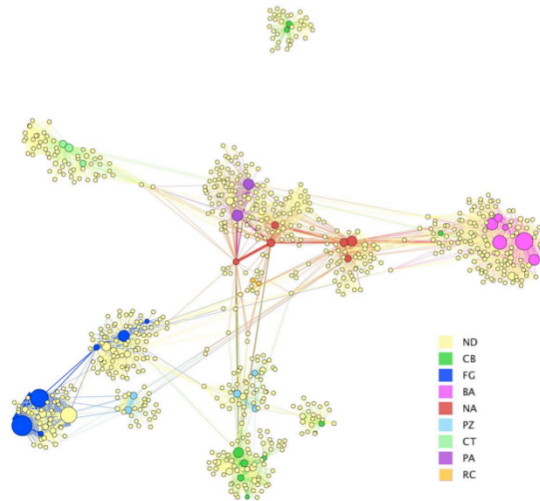
Again an undirected network; here links are paper coauthored by two authors (the nodes) and the strength/weigth is the number of papers; size of nodes proportional to number of papers

Biological networks

- Metabolic networks
- Protein-protein interaction networks
- DNA-protein interaction networks
- Neural networks
- Ecological networks
  - food webs
  - co-occurrence/co-exclusion networks
  - pollinator networks

A schematic representation of metabolic networks (1)

(a)   (b)
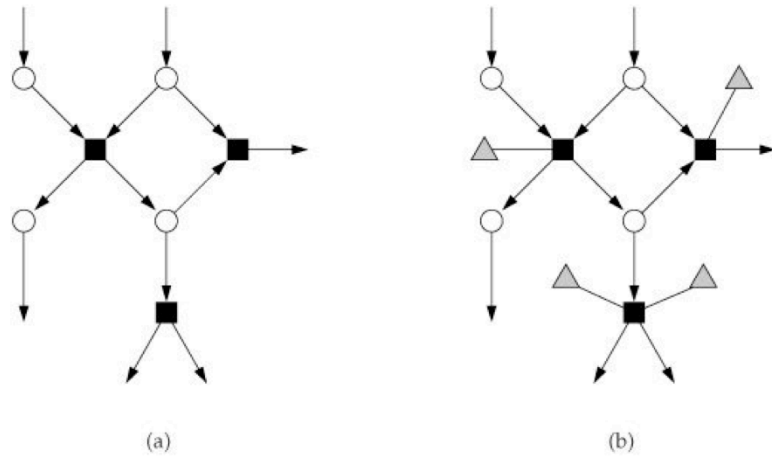
a. bipartite representation: circles are metabolites, squares are reactions; b. tripartite representations, enzymes are triangles
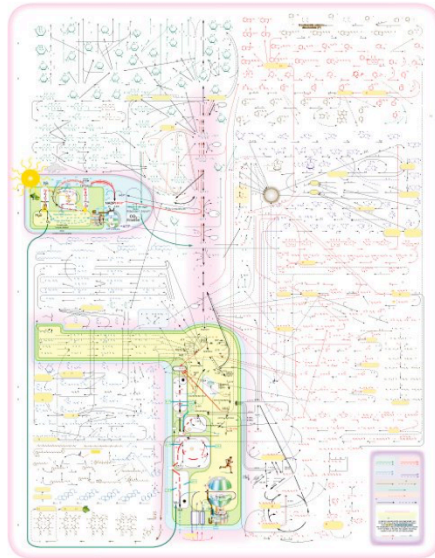
A schematic representation of metabolic networks (2)

here usually the reactions/enzymes are on the edges, but this may cause some loss of information

The citrate metabolism network for L. casei ATCC334 from KEGG
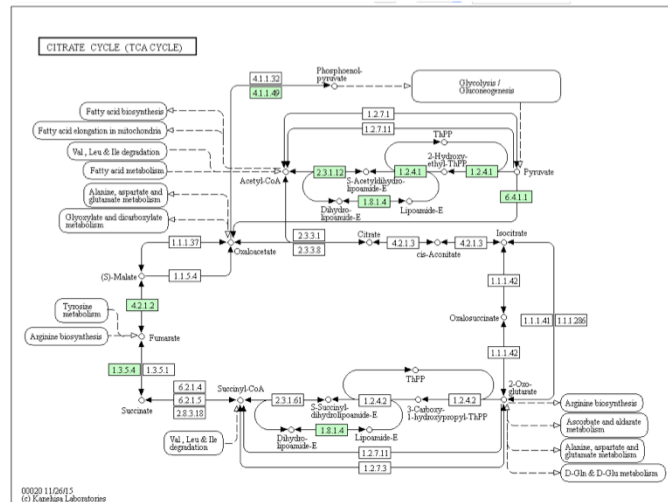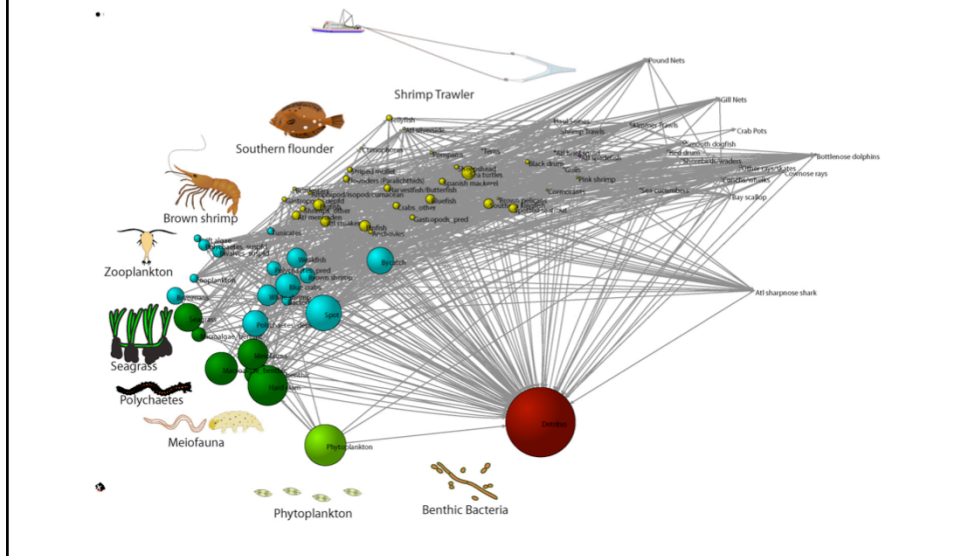
http://core.ecu.edu/BIOL/luczkovichj/core_sound/core_sound.html. Conventionally direction of arrows is with energy flow
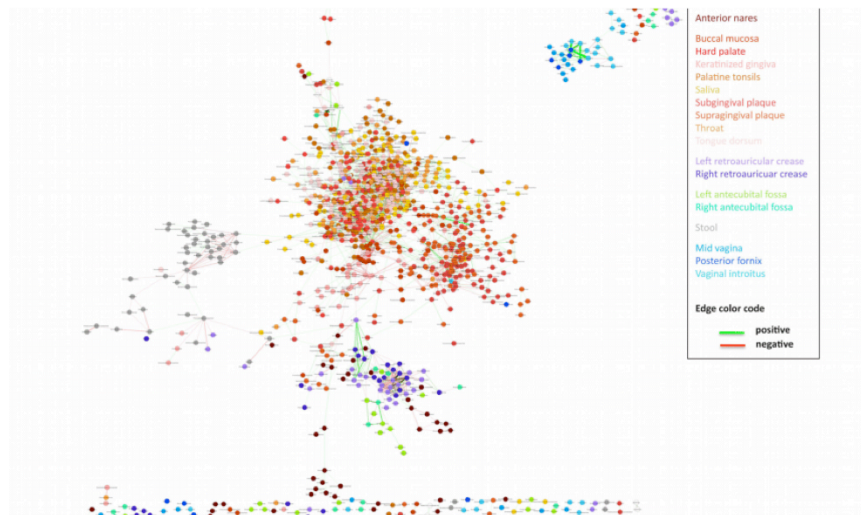
figura supplementare presa da Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., & Huttenhower, C. (2012). Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Computational Biology*, *8*(7), e1002606. doi: 10.1371/journal.pcbi.1002606

Figure S1 Significant co-occurrence and co-exclusion relationships among the abundances of clades in the human microbiome. The network displays all significant phylotype associations within and across the 18 body sites sampled by the HMP. Nodes represent phylotypes (colored according to the body site in which they occur) whereas edges represent significant relationships between phylotypes. Edge thickness reflects the strength of the relationship, and edge color its directionality (green co-occurrence, red co-exclusion).

Beyond pretty graphs: what can Network analysis do for me

- make the exploration of the data (much) easier
- compute network and node stats
  - explore networks (small world? scale free?)
  - compare different networks
  - identify nodes with special properties (large degree or weighted degree, centrality measures)
  - explore relationships between nodes
  - find communities / partition the network
  - extract sub-networks by filtering
- explore dynamic behaviours
  - how does the network change in time?

Node properties

- **qualitative metadata and features**
- **degree**: the number of edges of a node (in-degree and out-degree are computed separately for directed networks)
- **weighted degree**: the sum of the weights of the edges of a node
- **clustering coefficient**: a measure of the degree to which nodes in a graph tend to cluster together
- **centrality measures**: measures of the "importance" of a node in a network
  - degree centrality
  - closeness centrality
  - betweeness centrality
  - eigenvector centrality

hub a node with a lot of connections, assortative if tends to connect with other hubs, disassortative if not; authority: a node with a lot of incoming connections

Edge properties

- **qualitative metadata and features**
- **source and target nodes**
- **weight**
- **type** (directed vs undirected)

Centrality measures (undirected networks)
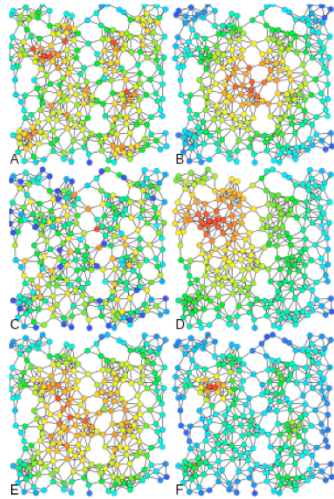
A. **Degree centrality**: the number of links incident upon a node
B. **Closeness centrality**: a measure of how long it will take to spread information from s to all other nodes sequentially (using shortest paths)
C. **Betweenness centrality**: the number of shortest paths which pass through the given vertex
D. **Eigenvector centrality**: a measure of the influence of a node in a network (used in Google search!)
E. **Katz centrality**: measures the number of all nodes that can be connected through a path, while the contributions of distant nodes are penalized
F. **Alpha centrality**: related to eigenvector centrality

da Wikipedia centrality refers to indicators which identify the most important vertices within a graph; degree centrality and disease: a node with a high index has a high probablity of catching a disease flowing through the network;

Network properties

- **size**: the number of nodes (N) of a graph
- **number of components**
- **connectivity (or connectance)**: the ratio between the number of edges and the the possible number of edges, i.e. the number of edg
- **average (weighted) degree**: 2E/N (for weighted degree $2(\Sigma w_i E_i)/N$
- es that would make the network a complete network, N(N-1)/2
- **average path length**: the average length of shortest paths between all nodes
- **clustering coefficient**: the average of the clustering coefficients of all nodes
- **modularity**: a measure of the strength of division of a network in modules/clusters (range -0.5 to 1, the higher the more modular)
- **fit and exponent of the power law (or other distributions)**: in many cases large networks are scale-free and a log-log plot of number of nodes vs. their degree is linear $P(k) \sim k^{-\gamma}$

Scale free networks contain hubs (i.e. nodes which have a degree largely exceeding the average) and are relatively robust to failure (i.e. several edges and nodes must be destroyed before the network becomes disconnected or fragmented; the network is resistant to the destruction of one hub but if several fail at the same time…). Nodes with a low degre usually have a high clustering coefficient, and clustering coefficient distribution also follows power law. the exponent gamma is usually in the range 2-3. This is also commected to the small world phenomenon: i.e. small communities of people which are highly connected but in which one member is connected to a hub. High degrees of clustering and hierarchic structures are also associated with power law distributions od the clustering coefficient of nodes

Node degree distribution
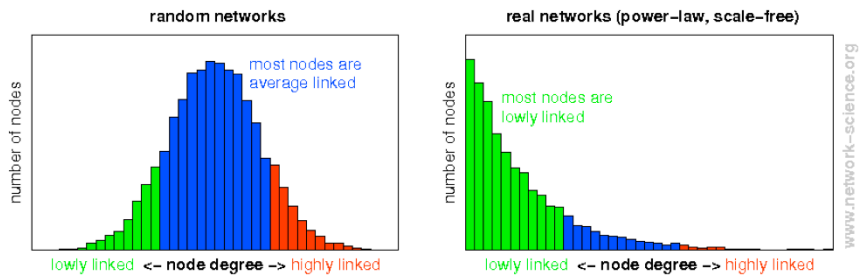
In random networks the node degree distribution follows Poisson law

Properties for "famous" networks
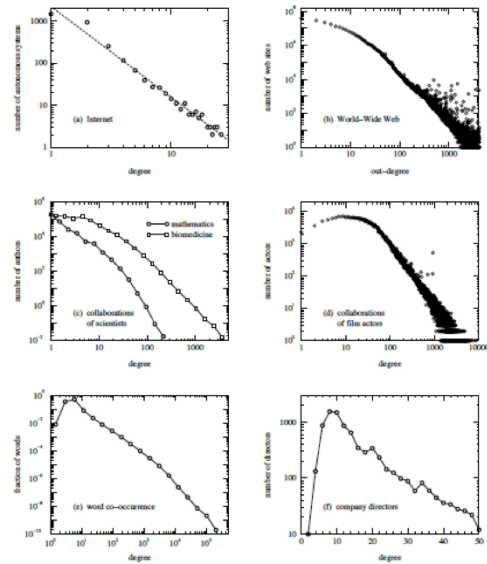
http://www.santafe.edu/media/workingpapers/02-02-005.pdf Note that the fit is not always perfect. There are several reasons for this and models of combined distributions

Properties for "famous" networks

network	n	z	clustering coefficient C measured	random graph
Internet (autonomous systems)[a]	6 374	3.8	0.24	0.00060
World-Wide Web (sites)[b]	153 127	35.2	0.11	0.00023
power grid[c]	4 941	2.7	0.080	0.00054
biology collaborations[d]	1 520 251	15.5	0.081	0.000010
mathematics collaborations[e]	253 339	3.9	0.15	0.000015
film actor collaborations[f]	449 913	113.4	0.20	0.00025
company directors[f]	7 673	14.4	0.59	0.0019
word co-occurrence[g]	460 902	70.1	0.44	0.00015
neural network[c]	282	14.0	0.28	0.049
metabolic network[h]	315	28.3	0.59	0.090
food web[i]	134	8.7	0.22	0.065

http://www.santafe.edu/media/workingpapers/02-02-005.pdf
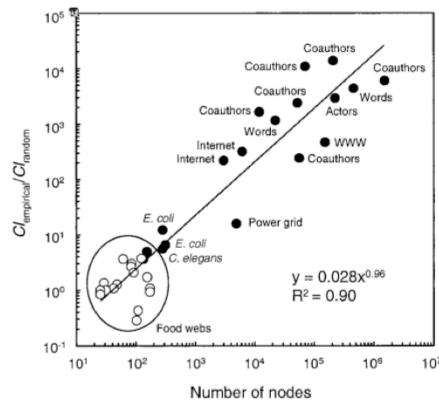
Fig. 1. Log–log plot of the clustering coefficient ratios (empirical/random web values) as a function of size of the network. Open circles represent data from 16 trophic food webs from the current analysis. Dark circles represent data from previous studies of 18 scale-free small-world networks summarized in ref. 19: 2 taxonomic food webs (22); *E. coli* substrate and reaction graphs (40); *C. elegans* neural network, movie actors, and power grid (17); 4 science coauthorship data sets (41, 42); 2 math and science coauthorship data sets (43); low and high estimates for Internet domains (44, 45); world wide web sites (46); and concurrence and synonomy of words (47, 44).

Comparing network properties: clustering coefficients for food webs versus scale-free, small world networks

Food-web structure and network theory:
The role of connectance and size

Jennifer A. Dunne*†‡, Richard J. Williams*, and Neo D. Martinez*

*Romberg Tiburon Center, San Francisco State University, 3152 Paradise Drive, Tiburon, CA 94920; and †Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

The authors conclude that food webs do not necessarily display the characteristics of other small world scale free networks, and that show smaller size and higher complexity, in terms of connectivity, and that degree distributions are typically related to size and connectance. Food webs with relatively high connectance typically display uniform distributions, webs with middle connectance tend to have exponential distributions, and webs with very low connectance display power–law or partial power–law distributions.

Suggested readings
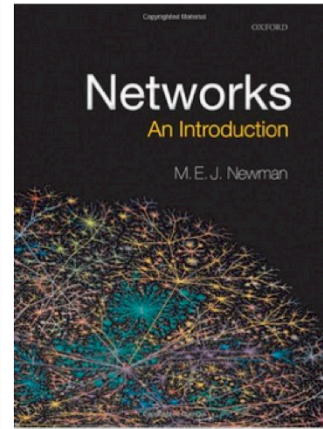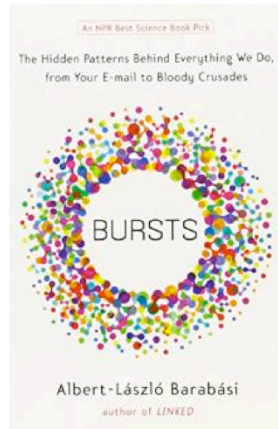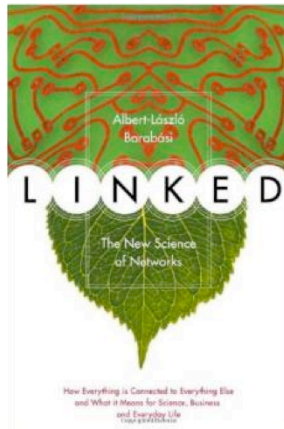
(More) Suggested readings

- Dunne, J. A., Williams, R. J., & Martinez, N. D. (2002). Food-web structure and network theory: The role of connectance and size. Proceedings of the National Academy of Sciences of the United States of America, 99(20), 12917–12922. doi:10.1073/pnas.192407699
- Deng, Y., Jiang, Y.-H., Yang, Y., He, Z., Luo, F., & Zhou, J. (2012). Molecular ecological network analyses. BMC Bioinformatics, 13(1), 113. doi:10.1186/1471-2105-13-113
- Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., & Huttenhower, C. (2012). Microbial Co-occurrence Relationships in the Human Microbiome. PLoS Computational Biology, 8(7), e1002606. doi:10.1371/journal.pcbi.1002606
- Flores, C. O., Meyer, J. R., Valverde, S., Farr, L., & Weitz, J. S. (2011). Statistical structure of host-phage interactions. Proceedings of the National Academy of Sciences, 108(28), E288–97. doi:10.1073/pnas.1101595108

Network visualization and analysis tools 1: Gephi
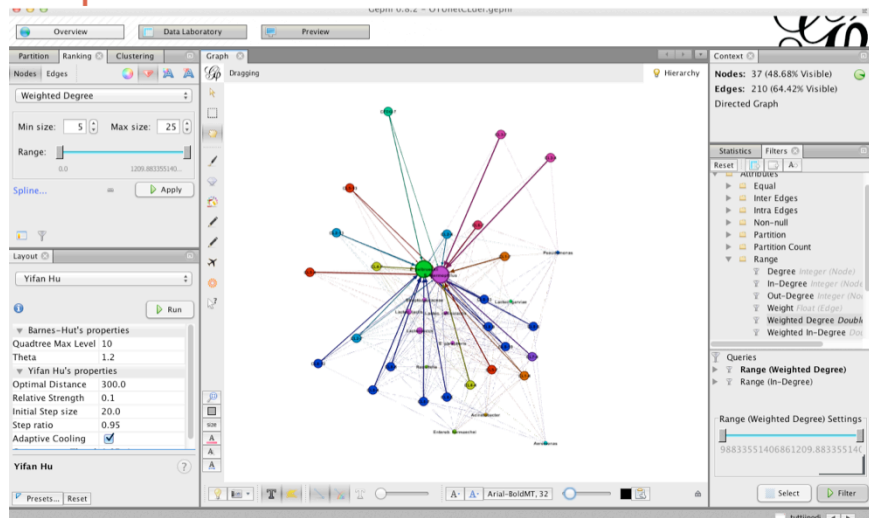
A few words on the Gephi project, its development etc.

Network visualization and analysis tools 2: Cytoscape
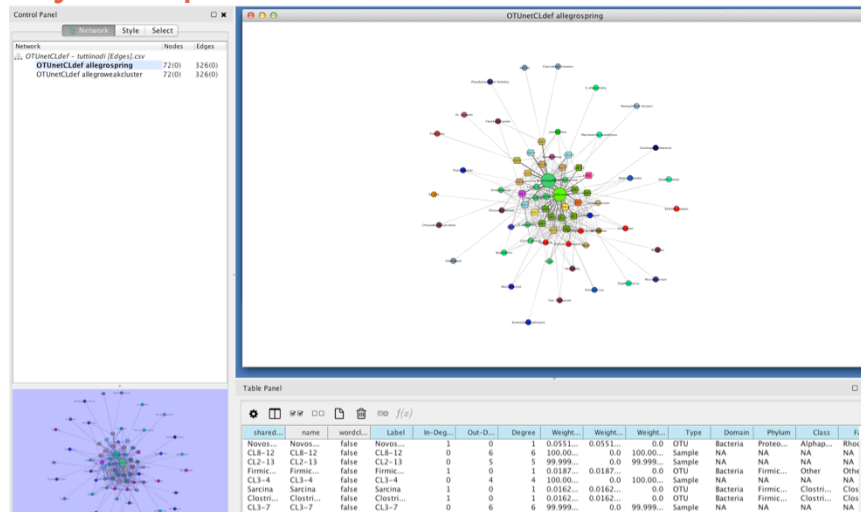
What can I do with them?

- import/export networks in a variety of formats
- visually explore networks
- calculate node and network statistics
- annotate networks by applying styles
- apply layouts to facilitate visualization/interpretation
- filter the networks to select and operate on subnetworks
- and much more (using plugins/apps)

Fare qualche esempio dopo

Other tools

- CoNet: a Cytoscape app for the generation and analysis of co-occurrence/co-exclusion networks http://psbweb05.psb.ugent.be/conet/index.php
- MENA: an online tool for the analysis of molecular ecological networks http://129.15.40.240/mena/
- BiMat: a Matlab package for the analysis of bipartite networks http://arxiv.org/abs/1406.6732
- and many others… (including R libraries such as igraph and bipartite)

Two examples of network analysis in food microbial ecology

- **Phage-Bacteria interaction networks**: how to use interaction matrices (weighted or unweighted) to improve visual representation and extract information for comparison among studies

- **OTU-sample and OTU-OTU networks**: use of data on the composition of food microbial communities generated by amplicon (16S DNA/RNA) targeted High Throughput Sequencing (HTS) to detect core microbial communities, cluster samples and detect significant microbe-microbe interactions (the FoodMicrobionet initiative)

PHAGE – HOST INTERACTIONS

Bacteriophages of lactic acid bacteria
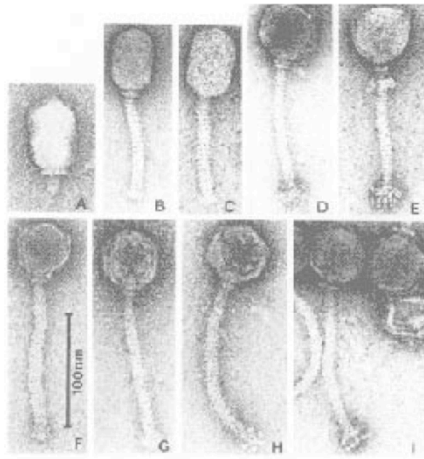
Fig 4. Electron micrographs of negatively stained bacteriophages active against lactic streptococci (A) phage P034; (B) P106; (C) P001; (D) P013; (E) P029; (F) P091; (G) P008; (H) P219; (I) P14 Thermagglutination is the scale for all pictures (see bars). (From Lembke et al. 1980.)

- all belong to the order *Caudovirales* (families *Myoviridae*, *Siphoviridae* and *Podoviridae*)
- ds DNA genome, 30-45 kb, many have been sequenced
- can be both lytic and temperate
- are the main cause of fermentation failure in the cheese industry

Phage-host relationships
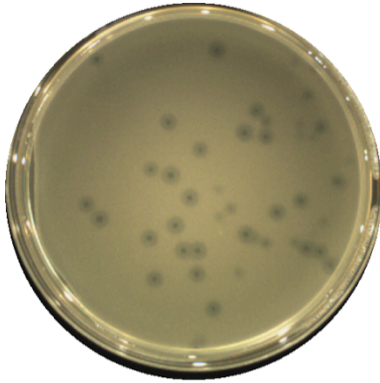
- **Phage sensitivity** is evaluated by a variety of methods providing quantitative (plaque counts,activity tests, etc.) or qualitative results (spot test, activity tests, etc.)
- the **host spectrum** of bacteriophages may be narrow (typical in *S. thermophilus*) or wide (some *Lactoc. lactis* phages) and depends on a variety of factors
- knowledge of phage host relationships is essential for **starter cultures management**

Potential structures of Phage Bacteria Interaction Matrices
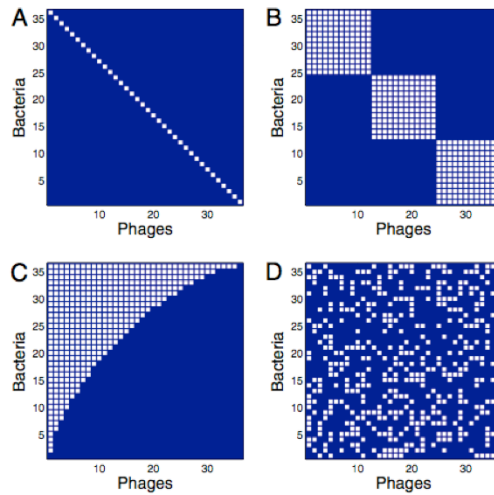
Flores, C. O., Meyer, J. R., Valverde, S., Farr, L., & Weitz, J. S. (2011). Statistical structure of host-phage interactions. *Proceedings of the National Academy of Sciences*, *108*(28), E288–97. doi:10.1073/pnas.1101595108

Expectations from evolution in bacteria and bacteriophages

Nestedness and modularity vs. random expectations
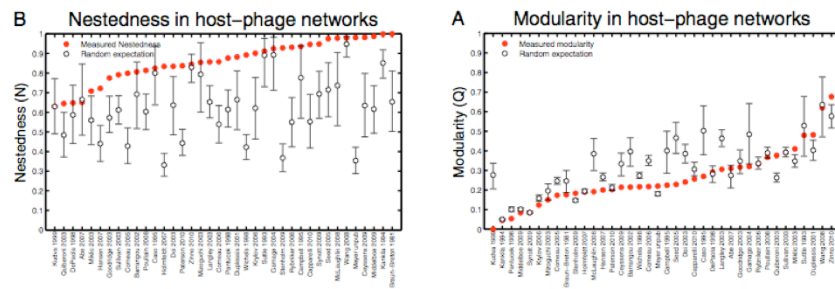
B   Nestedness in host–phage networks

A   Modularity in host–phage networks

Many S. thermophilus matrices are more modular than random expectations, this is probably due to the CRISPR-CAS system

Unsorted PBIM (Zinno et al., 2010)

Host	p574_A	p575_A	p576_A	p577_B	p591_C	p596_D	p603_B	p604_E	p605_F	p607_G	p611_H	p616_I	p620_L	p641_M	p642_N	p654_O	p671_P	p1027_Q	p1028_Q	p1032_Q	p1033_R	p1034_R	p1036_R	p1040_S	p1041_S	p1042_S	DT1
CHCC3063	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0
CHCC2070	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
CHCC2130	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHCC2133	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
CHCC2134	1	0	1	0	0	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
CHCC2389	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1
CHCC3048	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CHCC3049	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
CHCC3050	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
CHCC3046	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
CHCC4323	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
CHCC4325	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
CHCC4327	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
CHCC4460	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
CHCC4895	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
CHCC6592	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
SMQ-301	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Many S. thermophilus matrices are more modular than random expectations, this is probably due to the CRISPR-CAS system

Parameters from unsorted PBIMs

- H number of hosts;
- P number of phages;
- S number of species (H+P);
- M size (HxP);
- I number of interactions;
- C connectance (I/M);
- LH mean number of interactions, host (I/H);
- LP mean number of interactions, phage (I/P)

Many S. thermophilus matrices are more modular than random expectations, this is probably due to the CRISPR-CAS system

Sorting PBIMs by matrix clustering

Differences in isolation strategy may contribute

Network visualisations
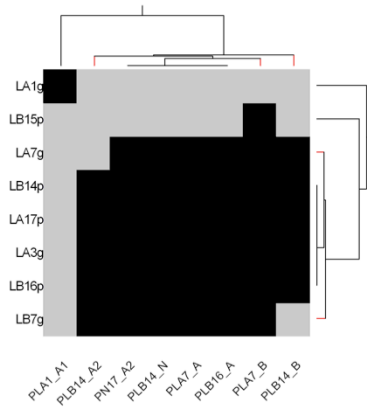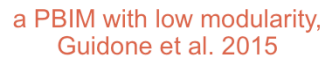
Network visualisations
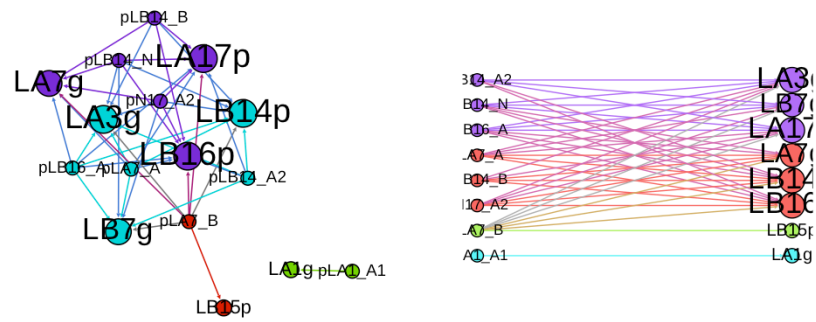
A PCA analysis from network properties
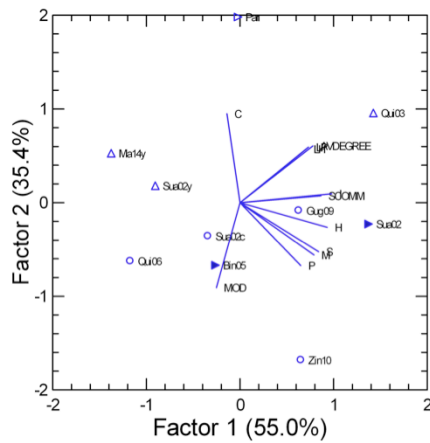
- an analysis was carried out on all published studies on *S. thermophilus* bacteriophages with PBIMs
- both parameters from the unsorted PBIMS and network parameters were used
- the analysis clearly separate different studies and, sometimes, subsets within studies

OTU - NETWORKS

Evolution of the microbiological approaches used to study microbial diversity in food ecosystems

Heterogeneity/biases
- different extraction methods
- different targets (16S RNA or RNA gene, variable region)
- different platforms
- different bioinformatics pipelines
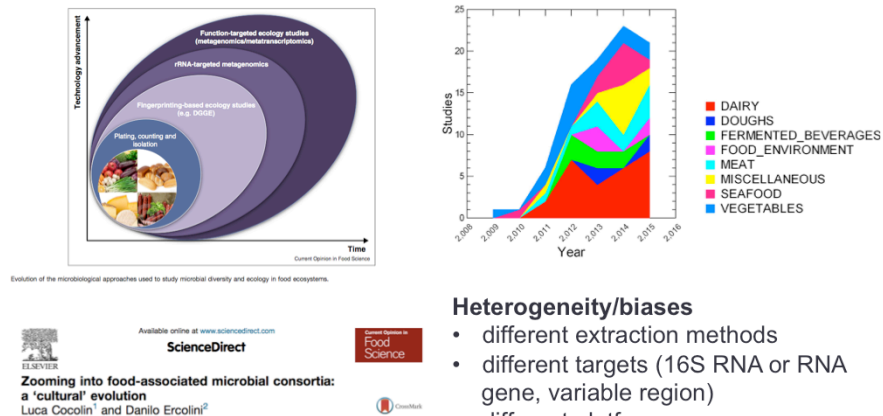
Number of papers rising steadily, more than 100 at the time of writing this presentation. evolution of methods in food microbial ecology, from counts+isolation to cultivation independent methods (DGGE, tRFLP, clone libraries) to amplicon targeted NGS to metagenomics. Number of studies published on amplicon targeted metagenomics 87+, 44+ in last two years

Evaluating the structure and function of microbial communities in foods

- The (more or less recent) past
  - **cultivation dependent methods** (count, isolate, identify/type): bias from non-culturability and or lack of selectivity, low throughput, low sensitivity (only species present at >1-5% detected)
  - **cultivation independent methods**: T/DGGE, ARISA, t-RFLP, analysis of clone libraries, microarrays, etc.: method dependent biases, often semiquantitative nature, sometimes lack of identification at the species level

- the present and (near) future
  - **amplicon targeted HTS**: targets the 16S RNA gene (active + inactive) or RNA (active only), high throughput (generally 2000-5000 sequences/sample needed for good coverage), several potential biases, resolution at the genus/species level
  - **shotgun sequencing of RNA libraries**: high throughput, identifies both genera/species and pathways active at different stages of evolution of a community

Advantages and disadvantages of HTS tools
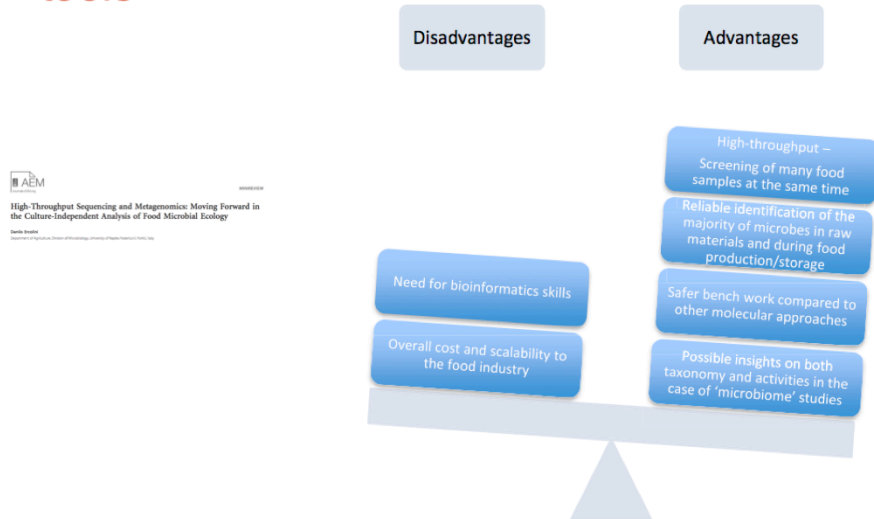
FIG 2 Advantages and disadvantages of the use of HTS to study food-associated microbial ecology.

NGS and third generation sequencing platforms

- **Next generation sequencing**
  - **Ion Torrent**: semiconductor sequencing; 200 bp; 0.02-10 Gb throughput; 1×10^7-7×10^8 reads/run, 4.5 h
  - **Illumina**: reversible terminator sequencing. 100-500 bp (increasing); 6-600 Gb throughput; 1×10^7-3×10^9 reads/run, 20 h to several days
  - **454**: pyrosequencing. 400-700 bp; 35-700 Mb throughput, 1×10^5-1×10^6 reads/run, 99-99.99% accuracy; 10-23 h
  - **SOLiD**: sequencing by ligation, 75 bp; 90-180 Gb throughput, 1.5-6$\times10^6$ reads/run, 99.99% accuracy; 7-12 d
- **Third generation sequencing**
  - PacBio RS II: Single Molecule, Real-Time (SMRT®) DNA Sequencing. 14-40 gbp; 1 Gb throughput, 70,000 reads/run, 99.999% accuracy (genomes); 30 min
  - Helicos, Oxford Nanopore

this and the following slides from http://users.ugent.be/~avierstr/nextgen/ Next_generation_sequencing_web.pdf; visit also http:// www.molecularecologist.com/next-gen-fieldguide-2014/

Flow chart for pyrosequencing

Flow	Notes
Food sample	design of the experiment
Extraction and purification of nucleic acids	DNA vs RNA, potential biases
Library preparation	
Emulsion PCR	DNA vs cDNA, potential biases amplicon vs shotgun
Pyrosequencing	potential biases, homopolymer stretches, quality, multiplexing, number of reads
Bioinformatic analysis	different level of choices for quality control, filtering, matching/grouping, identification
Data tables	

only some of this steps may be performed in your lab; generally sent out to services (darker boxes); issues with extraction from different matrices, extracellular DNA, contamination with mitochondrial/chloroplast, active vs inactive, bacteria resistant to lysis

The output: diversity

- **alpha diversity**: the species (OTUs) composition in a subunit (at different taxonomic levels), with diversity indices (commonly used indices: number of species, Simpson, Shannon, Chao1) and measurements of coverage
- **beta diversity**: measures of compositional dissimilarity among samples (two common measures are Bray-Curtis, which treats taxa as unrelated and Weighted Unifrac, which is based on phylogenetic distance among taxa and their abundance); these are used as an input for ordination methods (clustering, PCoA = MDS, etc.) or for inferential methods often together with metadata on samples

This is a dense table, with a lot of 0; the meaning of 0 values in unclear: they might mean absence or presence below the detection limit; this again would depend on the filtering options; were singleton and doubletons discarded? etc. If you want to run a correlation analysis rows with many 0 would inflate the correlation; different library sizes may seriously affect

Typical output 1 – Diversity indices

label	seq. before cleanup	seq. after cleanup	dairy	cycle	observed species	shannon	chao1	Goods' ESC
CL1-4	4148	3759	1	4	42	1.67	72.00	0.9909
CL1-7	4082	3206	1	7	35	0.49	44.10	0.9956
CL2-4	7063	5662	2	4	54	1.57	82.88	0.9961
CL2-7	8038	6746	2	7	39	1.32	49.11	0.9979
CL2-13	4696	3987	2	13	21	1.20	26.25	0.9982
CL3-4	6604	5334	3	4	13	0.07	31.00	0.9983
CL3-7	7610	6160	3	7	16	0.07	71.00	0.9982
CL4-4	9838	8369	4	4	57	0.98	111.17	0.9969
CL4-7	6241	4930	4	7	51	1.31	75.43	0.9961
CL5-4	6914	5510	5	4	19	0.07	41.00	0.9978
CL5-7	6135	5274	5	7	51	1.17	58.80	0.9975
CL5-13	5318	4190	5	13	16	1.15	19.00	0.9986
CL6-7	9250	7483	6	7	52	0.34	119.67	0.9961

Typical output 2 – Rarefaction curves
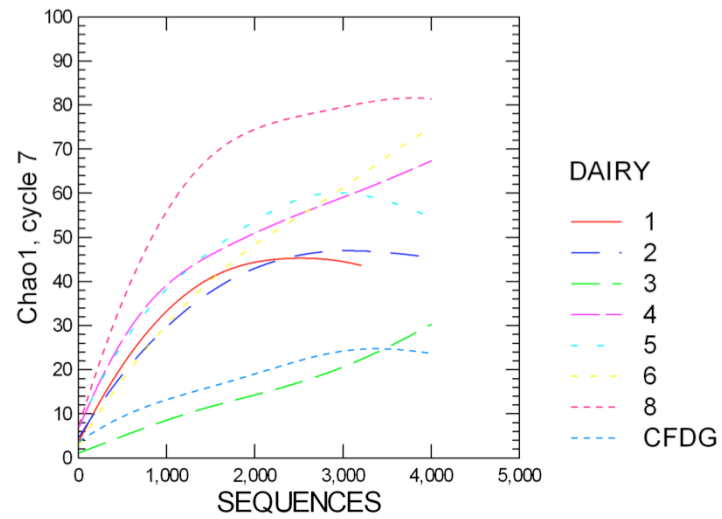
Typical output 3 – Pseudo-heat map
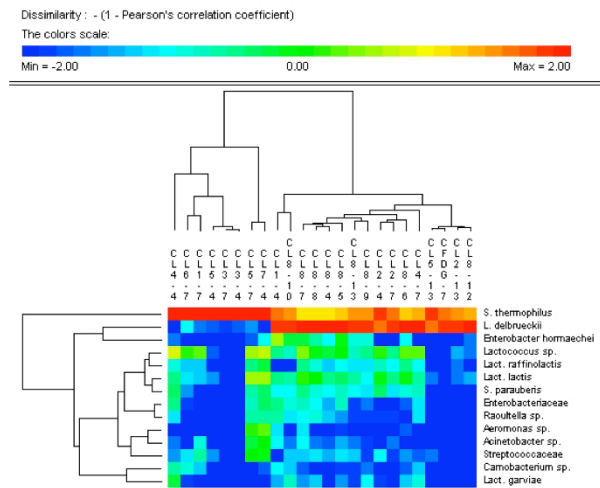
Typical output 4 – PcoA plot, weighted Unifrac
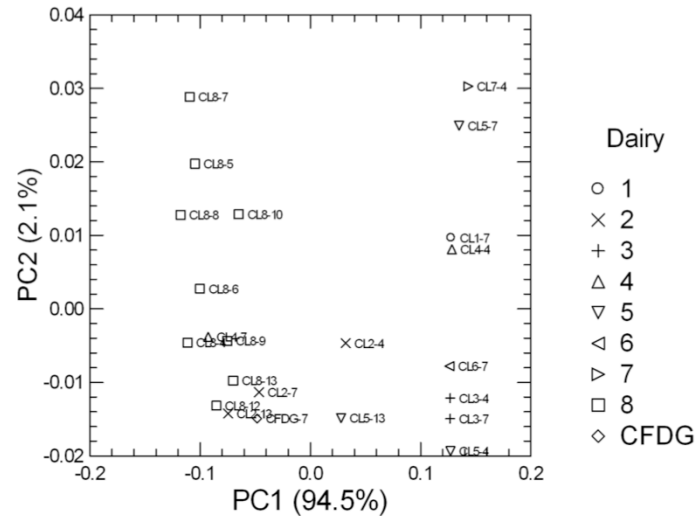
Beyond species: NGS monitoring of *S. thermophilus* biotypes based on *serB*
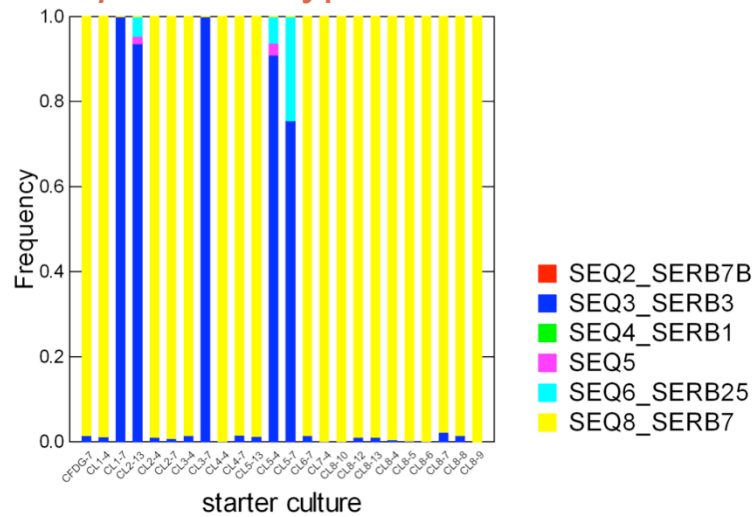
Mining next-generation sequencing studies in food microbial ecology

Quantitative information on the structure of food microbial communities is rapidly accumulating but extracting this information (for the purpose of writing reviews, designing new experiments, for food process development) is far from simple.

Alternatives:

- process sequences from SRA with own bioinformatics pipeline
- extract data from published papers
- **use abundance tables provided by authors**

point 2 needs unplotting, labour intensive, cannot get the exact data, difficult to compare. So, approach 1 leaves the biases/differences related to target,extraction/amplification, platform/protocol, 2 almost useless, 3 affected by several biases but at least biases are known

includes the equivalent of almost 20% of published studies, so it is reasonably representative

FoodMicrobionet flowchart
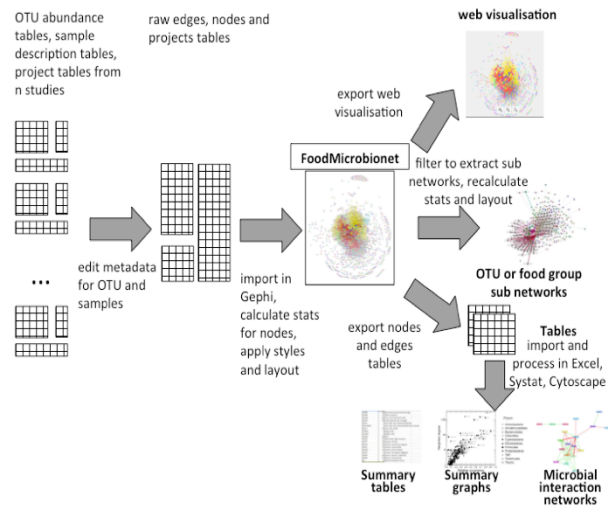
The rationale behind FoodMicrobionet. FoodMicrobionet as a database designed for network analysis tools FoodMicrobionet tools, examples in the following slides

Structure of edges and nodes tables

- **Edges**
  - source node label (OTU lineage)
  - target node label (food sample)
  - type (undirected)
  - weight (OTU abundance in sample)
  - administrative fields (gene target, etc.)

- **Nodes** (metadata)
  - node type (OTU, sample)
  - "taxonomy"
    - OTU lineage for OTUs
    - FoodID and FoodEx 2 classification
  - other administrative fields (outlinks to NCBI and dx.doi.org, custom fields, etc.)

Network representation of the abundance matrix

Web visualisation
http://www.foodmicrobionet.org/fmbn1_0_3web/
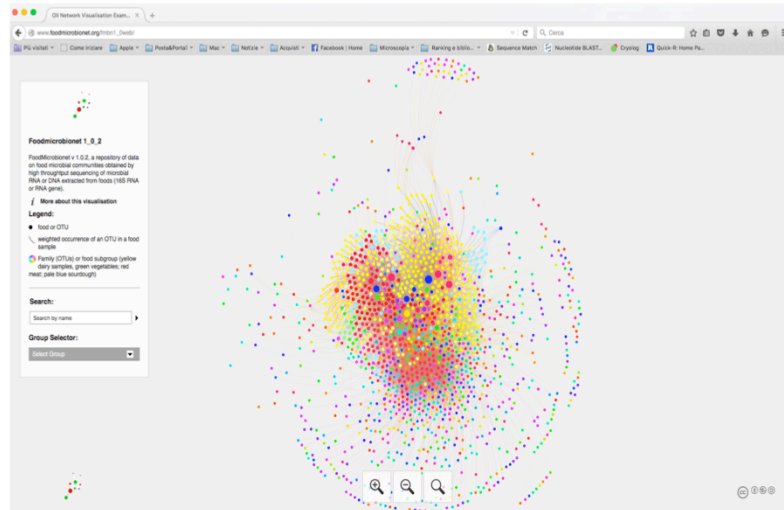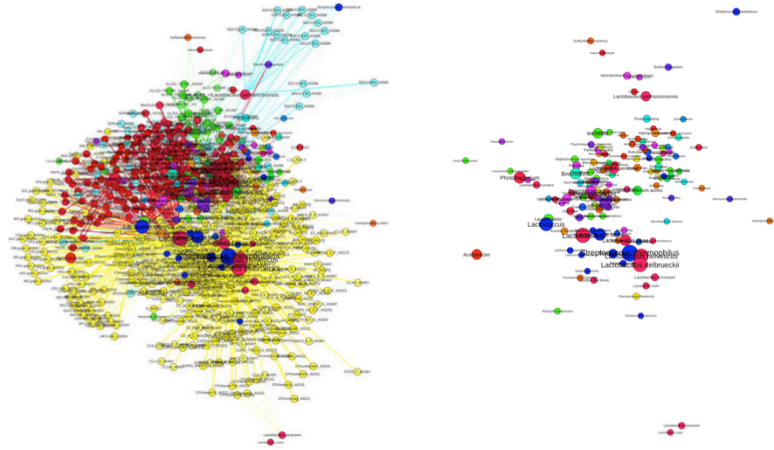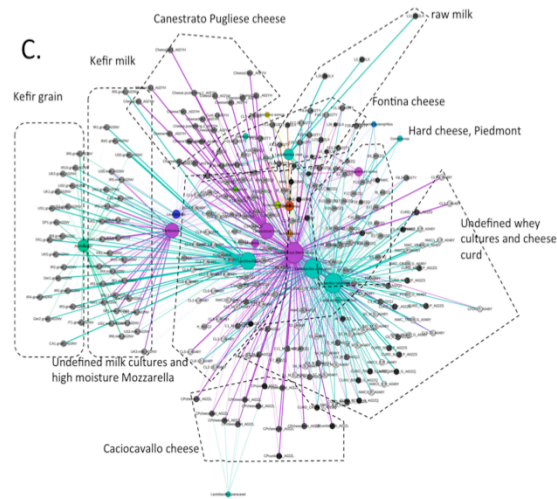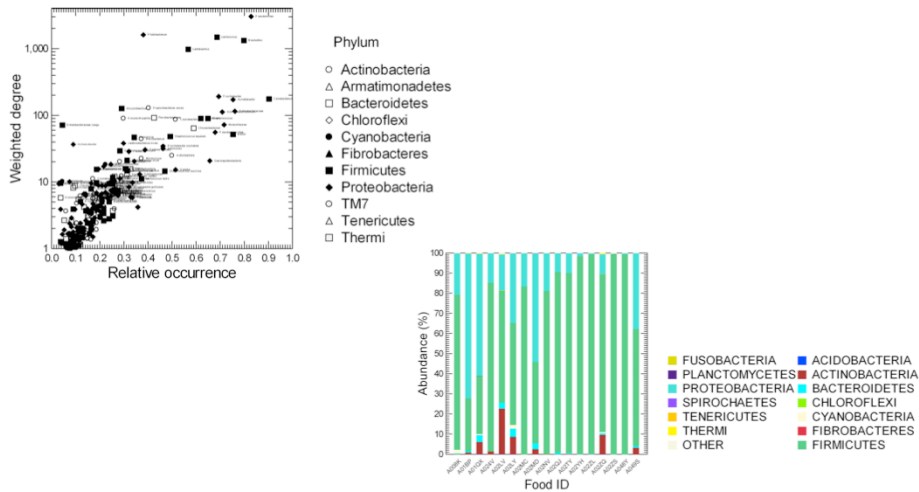
The flowchart in network representation: filter

Food specific networks, dairy

on the left a graph showing weighted degree (sum of abundances) as a function of relative occurrence (frequency in samples) for meat products; on the right abundance of different phyla in different food groups A02xxx are cheeses and milks, A48 starters, A049S meat products, A01QX fresh meat

Co-occurrence/co-exclusion networks
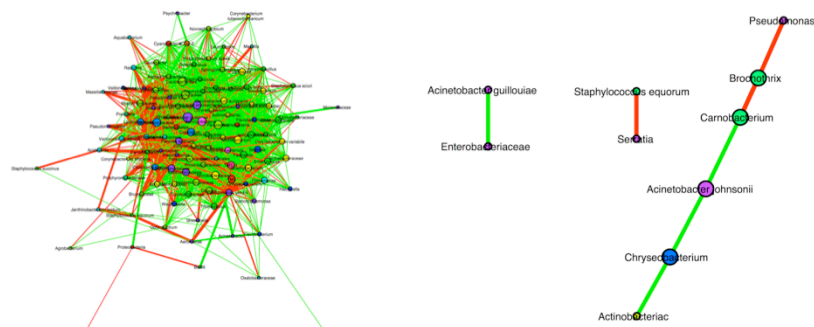
Exploring the Sources of Bacterial Spoilers in Beefsteaks by Culture-Independent High-Throughput Sequencing

Francesca De Filippis*, Antonietta La Storia*, Francesco Villani*, Danilo Ercolini*

left beef all samples (similar to time 0): 187 nodes, diameter 5, ave degree 35.15, right beef, spoiled; obtained using conet/cytoscape, measure co-occurrence/co-exlusion, not necessarily interactions,

Microbial interaction network analysis on selected datasets showed that the complexity (in terms of network size, average path length and modularity) of OTU-OTU networks increased with the complexity of the microbial community. It was lowest for kefir and for undefined starters and fresh cheeses, increased in surface ripened cheese, and was largest for raw meat samples. However, it was lower than that found in other microbial communities (human microbiome, soil and other environmental microbial communities), with no fit of the potewr law, clustering coefficients usually higher than corresponding random networks. As a comparison: Deng et al., provided stats for networks from environmental and human sources; network size ranged from 107 to 254 nodes, the node degree distribution showed a good fit of the power law for all networks and the modularity (which measures the occurrence of modules which are strongly interconnected) was significantly higher than that of random networks, while the occurrence of a hierarchical structure was variable

Network properties: food vs other environments

- **soil microbial communities** (Deng et al., 2012): OTU-OTU networks tend to be small world and scale-free, with small APL (similar order of magnitude to logN) and a hierarchical organization
- **human microbiome** (Faust et al., 2012): scale-free, short APL, high modularity, niche-specific relationships with a few hubs connecting different body areas; strong co-occurrence among phylogenetically similar organisms and exclusion among distant ones, no hierarchical organization
- **food** ???: nothing is known on food microbial communities studies are generally small, and diversity is limited

Faust data from the initial cohort of the Human microbiome project, 239 individuals 18 habitats, network with 197 clades (several taxonomic levels) with 2005 significant relationships
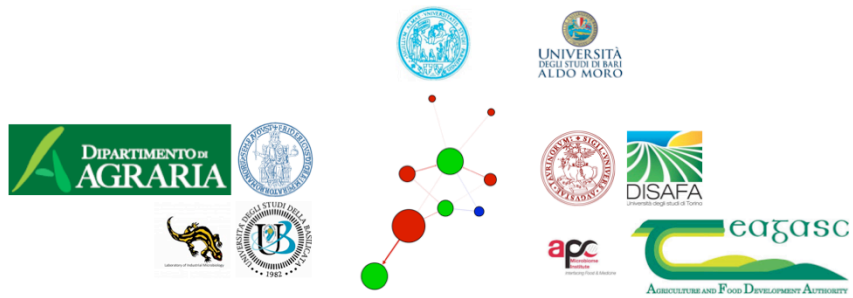
Co-occurrence/co-exclusion networks: stats

Dataset	Reference	Samples	OTU	Nodes	Components	Diameter	Average Degree	Average Path Length	Clustering Coefficient	Clustering Coefficient random network
Kefir	Marsh et al., 2013	48	61	7	2	1	3.14	1	0.712	0.449
Mozzarella	Guidone et al., 2015	29	77	22	2	5	4.18	2.19	0.453	0.19
Fontina, all samples	Dolci et al., 2014	27	296	27	2	4	5.19	1.92	0.235	0.192
Fontine, cheese	Dolci et al, 2014	18	158	11	2	4	2	2.95	0.152	0.181
NWC and curd	De Filippis et al., 2014	50	49	2	1	1	1	1	0	0
NWC and curd, Mozzarella	De Filippis et al., 2014	24	34	2	1	1	1	1	0	0
NWC and curd, Grana type	De Filippis et al., 2014	26	33	4	2	1	1	1	0	0
Piedmont hard cheese, all	Cocolin et al., unpublished	39	344	9	2	2	3.11	1.36	0.463	0.282
Beef, all	De Filippis et al., 2013	108	827	118	1	5	35.15	1.83	0.544	0.297
Beef, spoiled	De Filippis et al., 2013	45	108	11	3	5	1.46	2.35	0	0
Hamburger	Cocolin et al., unpublished	52	129	10	1	3	3	1.60	0.709	0.300
Fermented meat	Greppi et al., 2015	30	74	12	1	4	3.67	1.86	0.462	0.305

As a comparison: Deng et al., provided stats for networks from environmental and human sources; network size ranged from 107 to 254 nodes, the node degree distribution showed a good fit of the power law for all networks and the modularity (which measures the occurrence of modules which are strongly interconnected) was significantly higher than that of random networks, while the occurrence of a hierarchical structure was variable
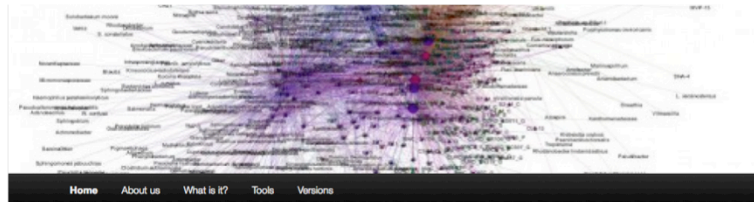
Conclusions

- network analysis is a powerful tool for studying emerging properties in microbial communities
- although network analysis tools have been mostly used for the representation of sample-OTU relationships or for the study of co-occurrence/co-exclusion, rapid accumulation of high resolution data on molecular characterization of microbial communities is extremely promising for formulating hypotheses on the structure and dynamics of microbial networks under a range of situations

Contributors, FMBN 1.0

Are you interested?

Visit our web site: http://www.foodmicrobionet.org
Contact us: foodmicrobionet@gmail.com

Some rights reserved

- This presentation was created by Eugenio Parente (2015) and modified in February 2016.
- with the exception of material downloaded from other sites you can:
  - Share — copy and redistribute the material in any medium or format
  - Adapt — remix, transform, and build upon the material
  - The licensor cannot revoke these freedoms as long as you follow the license terms.
- Under the terms of the Creative Commons Deed BY-NC-SA