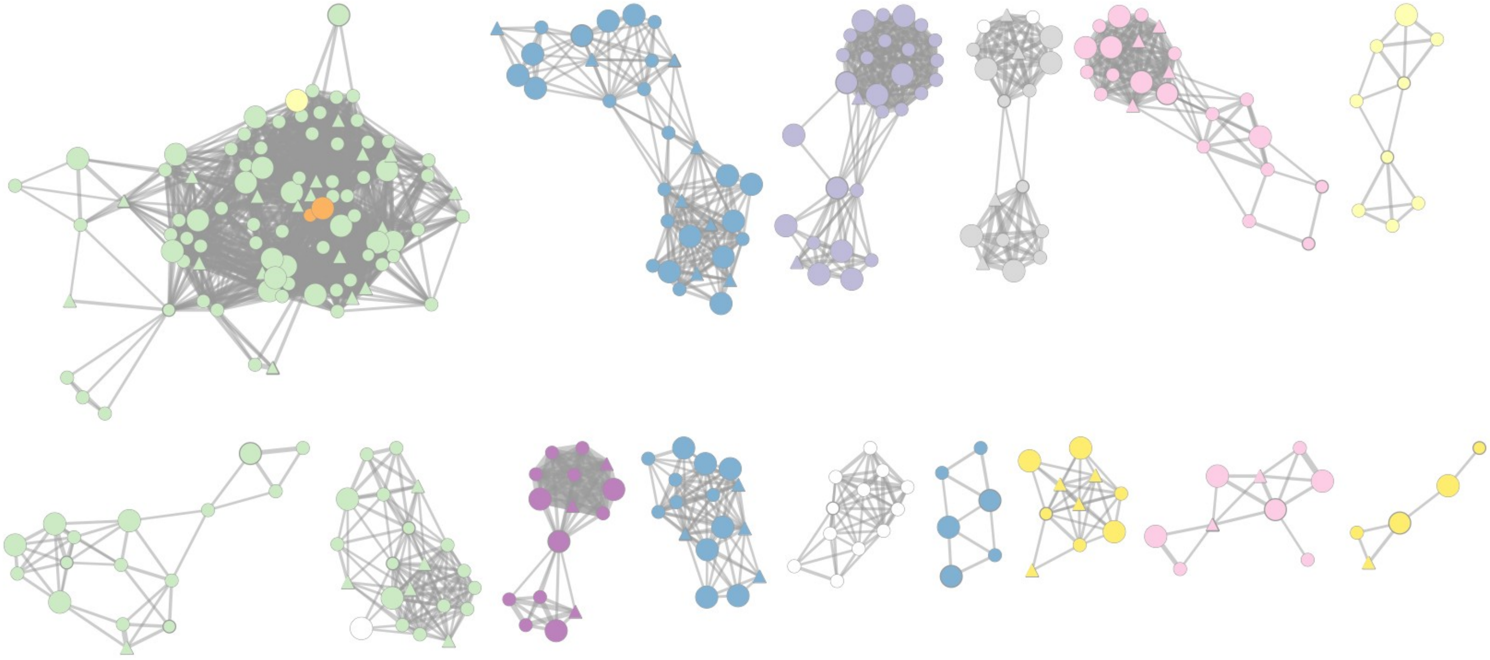
Seminar on biological networks

Teemu Kivioja
Teemu.Kivioja@helsinki.fi
Taipale laboratory

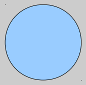
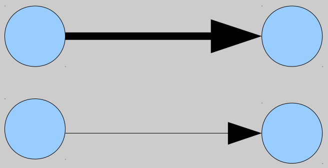
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI



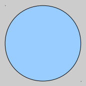
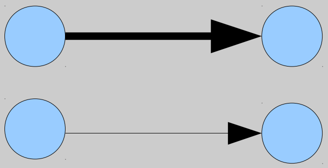
"Network science deals with complexity by "simplifying" complex systems, summarizing them merely as ***components (nodes)*** *and* ***interactions (edges)*** *between them. In this simplified approach, the functional richness of each node is lost. Despite or even perhaps because of such simplifications, useful discoveries can be made."*

Marc Vidal, Michael E. Cusick, and Albert-László Barabási, ***Cell****, 144, 986-998 (2011).*

Biological networks: weighted graphs with a biological interpretation

	Mathematical $G = (V, E)$	Biological	Visual
Nodes (vertices)	A finite set V	Components of the system: genes, proteins, metabolites ...	
Edges	A set E of, unordered or ordered, node pairs	Pairwise interactions of the components: physical, regulatory, genetic...	
Edge weights	Function from the set E to real numbers	Interaction strength	

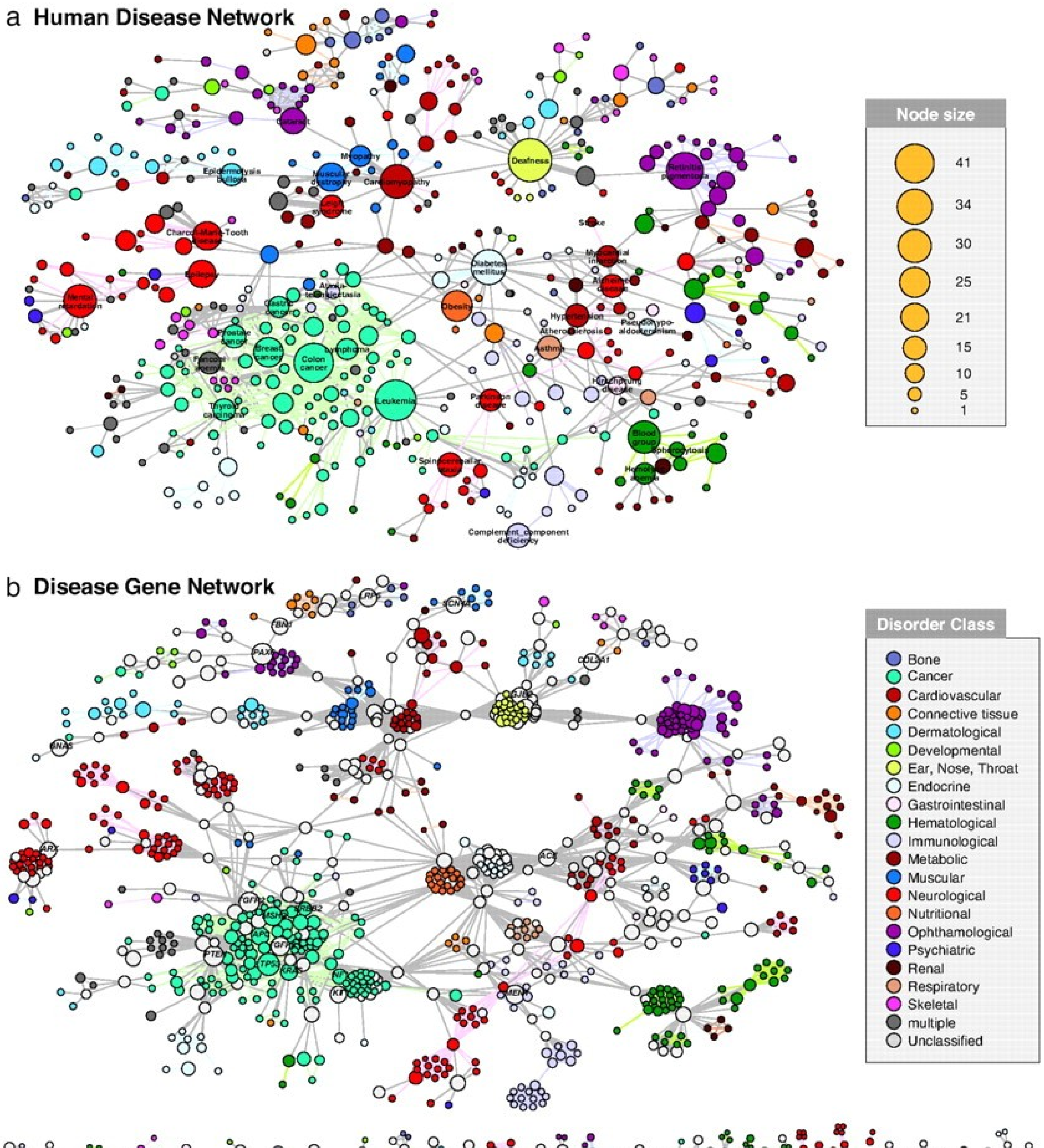
Biological networks: weighted graphs with a biological interpretation

	Mathematical $G = (V, E)$	Biological	Visual
Nodes (vertices	A finite set V	Components of the system: genes, proteins, metabolites ...	
Edges			
Edge weights	Function from the set E to real numbers	Interaction strength	

Typical process in bioinformatics:
move from **biological** (concrete) to **computational** (abstract) and back
In a way that produces **useful knowledge**

Nodes: disorders
Edge if disorders share genes in which mutations are associated with both disorders

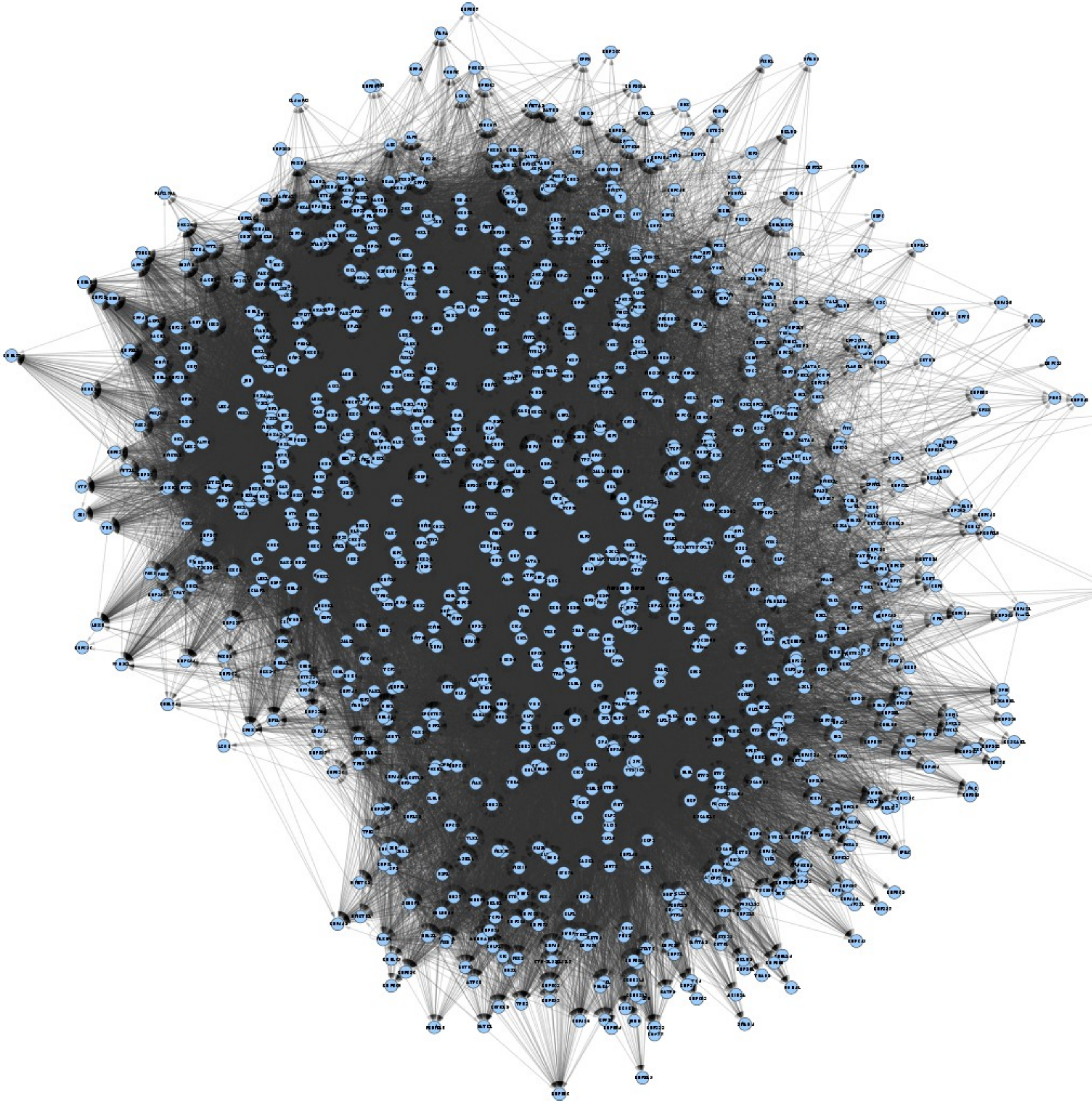
Nodes: genes
Edge if they are associated with the same disorder



Goh K. et al.
The human disease network
PNAS 2007; 104:8685-8690

Nodes:
transcription factors (TFs)
Edge if a TF is predicted to regulate another by binding to an enhancer element

Only about
1300 nodes and
50,000 edges



Course requirements

- Written report, 10-15 pages
- Oral presentation, about 45 minutes
- Active participation in the meetings including
    - this meeting
    - one personal meeting with me
    - the oral presentations of all the participants
- Absence from maximum one presentation allowed (not your own)

Grading

- The report and the presentation have equal weights
- Constructive participation (or the lack of it) can raise (lower) the grade by one
- The difficulty of the material is taken into account
- Meeting the deadlines counts
- Check the department “joint objectives for all seminars”!

General guidelines (1)

- The presentation and the report need to understandable to your peers
  - Introduce the field
  - Expand the concise presentation of original articles
  - Biology: Explain everything beyond very basic molecular biology (gene no, ortholog yes)
  - Computation: Explain all graph theoretical concepts needed
- Use supplementary material, review articles, cited articles, books, Wikipedia … to first understand yourself and then to explain to others
- No need to cover everything in the articles, choose the main points

General guidelines (2)

- In the report emphasize scientific precision
  - Use formal definitions, formulas, references …
  - Follow the basic conventions of scientific writing (as taught on the course)
- In the oral presentation emphasize clarity
  - Use schematic figures, simple examples …
  - Omit technical details, long sentences ...
  - Go slowly, especially in difficult parts: better to get one algorithmic idea across than rush through five
  - Think what you are going to say (and rehearse it)

Make sure that everybody

- understands
  - The biological problem
  - The abstract computational problem
  - How they are connected
- knows what the nodes and edges represent in your networks
- what is the input to the computation and its source
- what is the output of the computation and its interpretation

Schedule

- Personal meetings: January 28, **room D211**

- 10:00
- 10:30
- 11:00
- 11:30
- 12:00

- **Written report deadline: March 14** (Friday)
  - Distributed to all participants. Read!
- Oral presentations: at 10:15 here

- March 18
- March 25
- April 1
- April 8
- April 15

- Done before Easter

Personal meeting

- Ready
  - article selection
  - broad understanding of the topic and the selected material
  - outline of the report with
    - section headings
    - a couple bullet points per heading
- Aim
  - to go through the material and the story that you plan to present
  - to check that its scope and size are appropriate for the course
  - To discuss problematic parts (if any)
- Details can change during the work

What is a good topic for this seminar?

- Should have something to do with biology and networks ...
- Has both biologically and computationally relevant aspects
  - My suggestions tend to be algorithmic but statistical or mathematical are OK too
- However, the computational method does not need to be the newest, fastest and the most shining version
  - The goal is to communicate essential concepts and ideas, not detailed solutions
- At least two papers to be covered (but not everything in them)
  - Method and its application
  - Two competing methods
  - A simpler special case and a more general one ..
- Interesting!

How to choose one?

1. Choose one of my suggestions today after my presentation
2. Choose one of my suggestions not already chosen within a week and tell me by email
3. Choose your own topic and send it and the two main articles you are going to cover for my approval within a week.
    - See the review articles on the course web site

My suggestions

- Topics that I think
    - are interesting
    - have suitable material
- The article(s) mentioned just the starting point
    - Many more to choose from
    - Follow references backward and forward (Google scholar)
    - Choose based on your interests and knowledge

Topic 1: Detection of protein families by network clustering

- Network clustering identifies dense neighborhoods
    - Sets of nodes have a higher tendency to link to nodes within the set than to nodes in other sets
- It can be used to assign proteins to families based on their sequence similarity
- MCL is network clustering algorithm that has been widely used in protein family detection and in other tasks
- You can emphasize either
    - The MCL algorithm (mathematical)
    - Protein family detection (recent PFAM clustering)
- Or choose another application

Enright et al. An efficient algorithm for large-scale detection of protein families.
Nucl. Acids Res. (2002) 30 (7): 1575-1584.

Topic 2: Identifying conserved pathways by network alignment

- In principal one can extend the search from individual conserved proteins to conserved pathways using protein-protein interaction data but …
- Computationally the task is not an easy one
- Many papers available

Kelley et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. PNAS (2003) 100 (20): 11394–11399

Topic 3: Network motif discovery by color coding

- Network motifs are small subnetworks with characteristic topologies that appear in biological networks more often than expected by change
- It has been shown that specific types of motifs carry out specific dynamic functions (like electronic circuits)
- Counting occurrences of motifs that are larger than few nodes is algorithmically challenging
- Color coding is an elegant algorithmic technique that can be applied to many network problems (also alignment)

Alon. Network motifs: theory and experimental approaches.
Nature Reviews Genetics 8, 450-461 (2007)

Alon et al. Biomolecular network motif counting and discovery by color coding.
Bioinformatics (2008) 24 (13): i241-i249.

Topic 4: Inferring regulatory interactions by network deconvolution

- Correlation of expression between a gene coding a transcription factor and another gene suggests potential regulatory relationship
- However, indirect relationships also cause correlation
- Two recent papers propose “network convolution” as a means to separate direct from indirect effects
- Such methods could have large impact as vast amounts correlative data is produced by high-throughput methods

---

Barzel et al. Network link prediction by global silencing of indirect correlations.
Nature Biotechnology (2013) 31, 720–725

Feizi et al. Network deconvolution as a general method to distinguish direct dependencies in networks.
Nature Biotechnology (2013) 31, 726–733

Topic 5: Linking genes and diseases via network propagation

- Genes related to same disease tend to lie close to each other in various biological networks
- This observation can used to find novel disease-causing genes or to divide diseases into subtypes

Vanunu et al. Associating Genes and Protein Complexes with Disease via Network Propagation
PLoS Comput Biol 6(1): e1000641. doi:10.1371/journal.pcbi.10

Hofree et al. Network-based stratification of tumor mutations. Nature Methods 10, 1108–1115 (2013)

- See also

Eronen et al. Biomine: predicting links between biological entities using network models of heterogeneous databases. BMC Bioinformatics 2012, 13:119

Topic assignment

Topic 1:	Detection of protein families by network clustering
Topic 2:	Identifying conserved pathways by network alignment
Topic 3:	Network motif discovery by color coding
Topic 4:	Inferring regulatory interactions by network deconvolution
Topic 5:	Linking genes and diseases via network propagation

Schedule, your name in two places!

- Personal meetings: January 28, **room D211**
- **Written report deadline: March 14** (Friday)
    - Distributed to all participants. Read!
- Oral presentations: at 10:15 here
- Done before Easter

- 10:00
- 10:30
- 11:00
- 11:30
- 12:00

- March 18
- March 25
- April 1
- April 8
- April 15

Teemu.Kivioja@helsinki.fi