582653 Computational Methods of Systems Biology

Lecturers: Markus Heinonen, Leena Salmela
Course assistant: Hongyu Su

16.1.2012
Lecturers: M.Sc. Markus Heinonen and D.Sc.(Tech) Leena Salmela

Course assistant: M.Sc. Hongyu Su


4 credit course
Course topics

▶ Graph theoretical analysis of networks (chapters 3-6)
  ▶ Global properties
  ▶ Centralities, motifs
  ▶ Network clustering

▶ Analysis of biological networks (chapters 8-11)
  ▶ Gene regulation
  ▶ Protein-protein interaction
  ▶ Metabolic networks

▶ Network inference with statistical and kernel methods (not covered by the book)
  ▶ Metabolic networks
  ▶ Protein-protein interaction networks
  ▶ Gene regulation networks
Prerequisites

We will assume that you know at least something about the following

- Introduction to bioinformatics: protein, cell, genes, promoters, etc.
- Data structures: graphs and networks
- Algorithmic notation
- Elementary probability calculus / statistics
- Basic linear algebra / matrix computation
Completing the course

- Exercises (4 sets): 20% of the grade
- 1 Groupwork: 10% of the grade
- Course exam (Friday 2.3., 09.00-12.00 B123): 70 % of the grade
- Grading: 50% of maximum points gives the grade 1/5, 90% gives 5/5
Lectures

Total of eleven lectures:

1. Mon 16.1 introduction
2. Thu 19.1 Network models (Salmela)
3. Mon 23.1 Network clustering (note: lecture given in the exercise slot 10-12!) (Salmela)
   (Thu 26.1 Presentation of group work) (Salmela)
4. Mon 30.1 Motif Discovery (Salmela)
5. Thu 2.2 Transcriptional regulation networks I (Salmela)
6. Mon 6.2 Transcriptional regulation networks II (Salmela)
7. Thu 9.2 Statistical network inference, protein-protein networks (Heinonen)
8. Mon 13.2 Unsupervised network inference (Heinonen)
9. Thu 16.2. Supervised network inference (Heinonen)
10. Mon 20.2 Metabolic networks I (Heinonen)
11. Thu 23.2 Metabolic networks II (Heinonen)
Exercises

- 4 sets of exercises during the course, total 20% of the course grade
- Reviewed in the exercise sessions (30.1, 6.2, 13.2, 20.2)
- To be completed at home, returned in writing at the beginning of exercise session; or prior to the session to hongyu.su@cs.helsinki.fi
  - Note: exercise points will be based on the written answers
  - Don’t return your answers late if you want exercise points!
Groupwork

- Single groupwork assignment, total 10% of the course grade
- Completed in groupwork session Mon 23.1, 12-14
- Groupwork presentations Thu 26.1 10-12 (15 mins)
- Groups of approx. 3 students. For preferred groups, send those to Leena Salmela by email before Thu 19.1. Otherwise assigned randomly.
- Each group reviews and presents the main concepts from a paper:
  - Zhang & Zhang: A big world inside small world networks
  - Ravasz & Barabasi: Hierarchical organization in complex networks
  - Small et al: Scale-free networks which are highly assortative but not small world
  - Callaway et al: Network robustness and fragility: percolation on random graphs
  - Stumpf et al: Subnets of scale-free networks are not scale-free: Sampling properties of networks
Additional reading

- For more broad coverage of the course topics, you may look at the following books
- These books are not required for passing the course
What is systems biology?

Characterizations from the literature:

▶ "A new field of biology that aims at systems-level understanding of biological systems" (Kitano, 2001)
▶ "Systematic investigation of cells, organs and organisms and of cellular processes.." (Klipp et al. 2005)
▶ "The study of the interactions between the components of biological systems, and how these interactions give rise to the function and behavior of that system" (Wikipedia)
What is systems biology?

An antithesis to ’traditional, reductionist biology’:

- Reductionist biology: study of individual components (genes, proteins) of the system, aiming to in depth understanding of how a component behaves
- Systems biology: study of the interplay of a set of components in order to understand how they make up the systems behaviour
- Natural modeling tool a network of individual components with qualitative/quantitative interactions
What is systems biology?

Reductionist and systems approaches need each other for best results:

- **Component knowledge carries systems analysis**: e.g., if functional annotation of a gene is incorrect, the model of a biological system including that gene is likely to give incorrect predictions as well.

- **System analysis for finding crucial components**: e.g., analysis of the biological system can reveal components that are the most promising targets for drugs.
What is systems biology?

Aspects of biological systems (Kitano, 2002):

► System structure: building blocks, interactions
► System dynamics: behaviour, stochasticity
► System control: robustness
  ► adaptation,
  ► parameter insensitivity
  ► feedback loops
  ► redundancy, modularity
► System design
Overview of major research topics in systems biology

- Network modeling: Scale-free networks, etc.
- Network analysis: Clustering, statistical models
- Metabolic networks: reconstruction, flux balance analysis, steady-state analysis, metabolic control analysis
- Transcriptional regulation networks: inference, analysis of motifs, dynamics
- Protein-protein interaction networks
- Dynamic modeling with differential equations
- (Un)Supervised learning of models
- ...
Systems biology and genome-wide measurements; ’omics’

- Systems biology relies on genome-wide measurements
- Yield indirect and direct information on the interactions between the components.

- Genomics: DNA sequences
- Transcriptomics: mRNA expression
- Proteomics: protein concentrations
- Metabolomics: small molecule concentrations
- Fluxomics: enrichment of stable isotopes in metabolites
Examples of direct measurements of interactions:

- Yeast-2-hybrid, Y2H: protein to protein binding ⇒ protein interactions
Examples of direct measurements of interactions:

- ChIP-chip: Chromatin immunoprecipitation, ChIP protein binding to promoter regions of genes followed by DNA microarray \(\Rightarrow\) regulatory interactions
Systems biology and genome-wide measurements; ’omics’

Examples of indirect measurements of interactions:

- Gene expression chips: Two-way experiment (right), Heatmap of expression values (bottom)
'Omics’ databases

Large-scale data repositories for systems biology data:

- Gene expression data: ArrayExpress, NCBI Omnibus. Huge amounts of data, noisy, difficult to compare
- Transcriptional networks: RegulonDB (E. coli), ...
- Y2H: AfCS database.
- Metabolic models: KEGG, BioCyc
- Protein protein interactions: STRING, BioGRID
- ...
Networks and systems biology

- We will concentrate on modelling the interaction networks underlying the biological systems
- Focus on network structure and how it qualitatively reflects in biological function
- Quantitative modelling of dynamic, time dependent behaviour (typically: differential equation systems) is not part of the focus of the course
Physical interactions (typically: binding of molecules, forming a complex) between molecules:

- Protein and DNA: transcription factor proteins, epigenetic silencers, histones . . .
- Protein and RNA: ribosomes, transcriptase proteins, . . .
- Protein and Protein: protein complexes, (some) metabolic pathways, . . .
- Protein and Small molecule: enzymes, metabolic regulation, signaling, . . .
Abstract interactions

We will often look at abstract or logical interactions between the components, rather than mapping physical interactions:

▶ Gene regulatory network: 'gene A' negatively regulates 'gene B'
  ▶ Biologically: transcription factor protein produced by A, binds to the promotor region of B, thus repressing the transcription of B

▶ Metabolite network: metabolite $M_1$ interacts with metabolite $M_3$
  ▶ 'Biologically: protein C catalyzes a chemical reaction $M_1 + M_2 \implies M_3 + M_4$ between metabolites

▶ Correlated behavior: gene A has similar behaviour to gene B in a set of experiments—do not necessarily need to have direct regulatory relationship, although often they have
Graph-theoretical network analysis

Aims:

- To elucidate the structure of networks, in terms of properties computable from the network:
  - degree distribution, connectedness, path length distributions
- Finding frequently occurring patterns or motif from the network:
  - Regulatory cascades, dense overlapping modules, ...

- Landmark paper of network analysis
- Showed that many natural networks (power grid, co-actor relations, neural network of C. elegans) have similar properties
  - Highly clustered, dense local subgraphs
  - Short average path length between and two nodes (hence the term "small world")
- Small-world networks since found in numerous other networks in biology and other fields of life.
Scale-free networks (Barabasi & Albert, 1999)

- Barabasi and Albert created a simple theoretical model for creating small-world networks.
- The resulting network has few nodes with many neighbours and many nodes with few neighbours.
- Scale-freeness refers to the mathematical law governing the degree distribution of the nodes.

(a) Random network
(b) Scale-free network
Network motifs (Alon, 2005)

- Uri Alon and later many others have studied local subnetworks that can be found in biological networks.
- Network motifs are subnetworks that occur in the network much more often than expected by chance.
- Alon also offers functional explanations why the motifs are present.
Networks in biology

Types of biological networks of interest:
- Transcriptional regulation
- Signal transduction
- Protein-protein interactions
- Metabolism
Transcriptional regulation networks

- Often also called 'gene regulatory networks'
- Represented as graphs where nodes correspond to genes
- Directed edges $A \rightarrow B$ correspond to regulatory relations 'product of gene $A$ controls the transcription of gene $B$'
- Positive (enhancer) or negative (repressor) regulation may be indicated by signs or special arrowheads

(Lee et al., Science Vol. 298, 2002)
Transcriptional regulation networks

- In general, not very well known for most organisms
- Fragments of the complete network characterized
- Alon’s network motif work is one of the most notable approach to analyse transcriptional regulation networks

(Lee et al., Science Vol. 298, 2002)
Protein interaction networks

- Protein(-protein) interaction network models two kinds of interactions:
  - Proteins binding to each other and functioning as a complex
  - Proteins catalyzing biochemical reactions sharing a metabolite (enzyme network)
- Represented as undirected networks with proteins as nodes and the interactions as edges
Protein interaction networks

- PPI networks are very much tied to the measurement technology, different network obtained with different methods
- Scale-free properties found in PPIs: network contains few 'hub' proteins with many neighbors
Signal transduction networks

Signal transduction networks model the communication systems by which the cell obtains information from the environment and from other cells (e.g. nutrients, hormones).

Signal transduction involve heterogeneous interactions: protein-small molecule, protein-protein and protein-DNA

(MAPK signaling pathway, Wikipedia Commons)
Metabolic networks

- Metabolism is responsible of providing the cell with energy and building blocks for cell growth
- Metabolic networks are composed of biochemical reactions, catalyzed by enzymes (proteins) and metabolites that participate in the reactions
Metabolic networks

Several representations used, for example

- Hypergraph: Metabolites as nodes, reactions as hyperedges (edges with more than two end-points)
- Bipartite graph: Reactions and metabolites as nodes, edges connect metabolites to reactions and vice versa
- Metabolite graph: metabolites as nodes, edges represent participation in reaction

E.Coli glycolysis, EMP database, www.empproject.com
Metabolic networks

- Study of enzyme kinetics pre-dates the discovery of DNA by a half a century
- Metabolic networks are better characterized for many organisms than other biological networks
- Wide array of dedicated network analysis tools has been developed, together with rigorous theory of metabolic systems
  - Stoichiometric models
  - Functional pathway discovery
  - Analysis of metabolic fluxes (pathway activity in steady state)
  - Metabolic control analysis for modelling metabolic regulation
Dependency networks

The biological networks described before

- some underlying physical interaction (protein-protein, protein-DNA, protein-metabolite) they aim to model
- there typically are dedicated measurements to uncover these interactions
- the networks have undergone significant manual curation
Dependency networks

- High-throughput 'omics' data gives a rise to another approach to network construction
- Correlation networks model statistical dependencies between any components (genes, proteins, metabolites)
- There may or may not be an underlying physical interaction behind the correlation
- For example: gene expression and metabolite concentration may be observed correlate in some set of experiments $\Rightarrow$ new previously unknown interaction?
Dependency networks

- Observe a set of components (genes, proteins, metabolites) in a set of experiments
- Construct a correlation matrix to list all pairwise correlations between components
- Graph is extracted by drawing an edge between components that have correlation higher than threshold
Dependency networks

Other statistical formalisms

- Partial correlations: Are two components still correlated if we remove the common regulatory effects? On complex systems, most variables are correlated by chance alone.
  - Eating ice cream and sunbathing are highly correlated. However, a third (latent) factor (sunny weather) explains most of the correlation, leaving no-to-small direct correlation.

- Information theoretic measures instead of correlation
- Boolean circuits
- (causal) Bayesian networks
Machine learning

Supervised machine learning

- Learn a model from data
- Use the model to predict new interactions
- Support vector machines learn which features of the data are relevant w.r.t. to a prediction target
- E.g. predict new protein protein interactions:
  - Collect data from known PPI’s (protein properties, gene expression profiles, etc.)
  - Learn a model relating two protein’s interaction to their common properties
  - Use the model to predict new interactions
Next steps on the course...

- Next lecture on Thursday 19.1 10-12
- Next week’s timetable
  - Lecture on Mon 23.1 10-12
  - Groupwork right after at 12-14,
  - Groupwork presentations on Thu 26.1 10-12.
- For preferred groups (∼ 3 people) for groupwork send email to leena.salmela@cs.helsinki.fi by Thursday 19.1.
- First exercises in following week at Mon 30.1