Inferring Protein-Signaling Networks

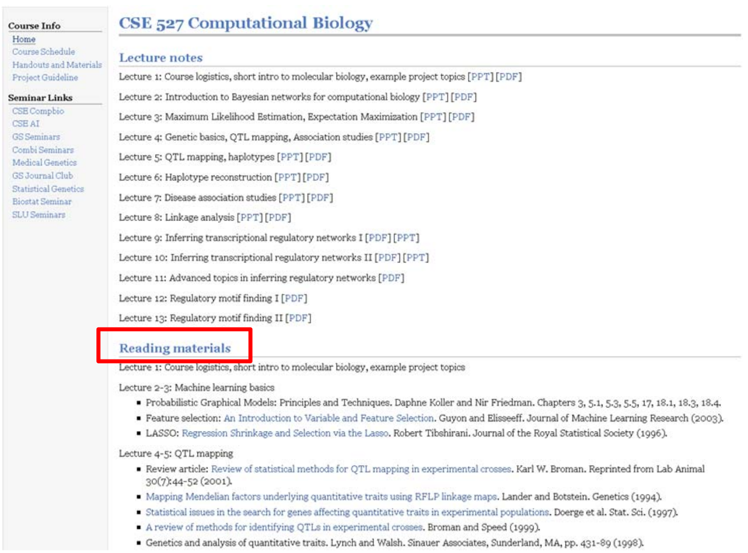
Lectures 14 – Nov 14, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

Course Announcement

- http://www.cs.washington.edu/education/courses/cse527/11au/notes.html



Course Info
Home
Course Schedule
Handouts and Materials
Project Guideline

Seminar Links
CSE Compbio
CSE AI
GS Seminars
Combi Seminars
Medical Genetics
GS Journal Club
Statistical Genetics
Biostat Seminar
SLU Seminars

CSE 527 Computational Biology

Lecture notes

Lecture 1: Course logistics, short intro to molecular biology, example project topics [PPT] [PDF]

Lecture 2: Introduction to Bayesian networks for computational biology [PPT] [PDF]

Lecture 3: Maximum Likelihood Estimation, Expectation Maximization [PPT] [PDF]

Lecture 4: Genetic basics, QTL mapping, Association studies [PPT] [PDF]

Lecture 5: QTL mapping, haplotypes [PPT] [PDF]

Lecture 6: Haplotype reconstruction [PPT] [PDF]

Lecture 7: Disease association studies [PPT] [PDF]

Lecture 8: Linkage analysis [PPT] [PDF]

Lecture 9: Inferring transcriptional regulatory networks I [PDF] [PPT]

Lecture 10: Inferring transcriptional regulatory networks II [PDF] [PPT]

Lecture 11: Advanced topics in inferring regulatory networks [PDF]

Lecture 12: Regulatory motif finding I [PDF]

Lecture 13: Regulatory motif finding II [PDF]

Reading materials

Lecture 1: Course logistics, short intro to molecular biology, example project topics

Lecture 2-3: Machine learning basics
- Probabilistic Graphical Models: Principles and Techniques. Daphne Koller and Nir Friedman. Chapters 3, 5.1, 5.3, 5.5, 17, 18.1, 18.3, 18.4.
- Feature selection: An Introduction to Variable and Feature Selection. Guyon and Elisseeff. Journal of Machine Learning Research (2003).
- LASSO: Regression Shrinkage and Selection via the Lasso. Robert Tibshirani. Journal of the Royal Statistical Society (1996).

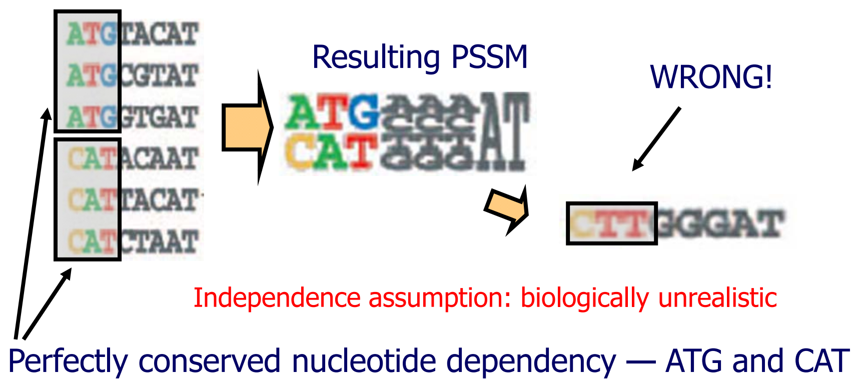
Lecture 4-5: QTL mapping
- Review article: Review of statistical methods for QTL mapping in experimental crosses. Karl W. Broman. Reprinted from Lab Animal 30(7):44-52 (2001).
- Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Lander and Botstein. Genetics (1994).
- Statistical issues in the search for genes affecting quantitative traits in experimental populations. Doerge et al. Stat. Sci. (1997).
- A review of methods for identifying QTLs in experimental crosses. Broman and Speed (1999).
- Genetics and analysis of quantitative traits. Lynch and Walsh. Sinauer Associates, Sunderland, MA, pp. 431-89 (1998).

Outline

- Regulatory motif finding
  - More computational methods
    - Greedy search method (CONSENSUS)
    - Phylogenetic foot-printing method
    - Graph-based methods (MotifCut)
  - Before/ after motif finding
- Inferring signaling networks

3

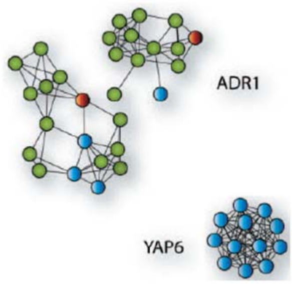
Drawbacks of Existing Methods



4

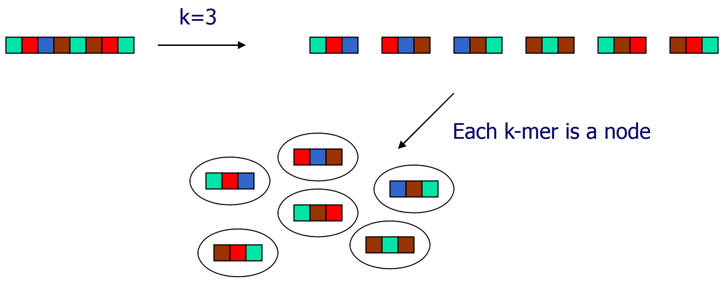
Overview: Graph-Based Representation

- Nodes: k-mers of input sequence
- Edges: pairwise k-mer similarity
- Motif search → maximum density subgraph



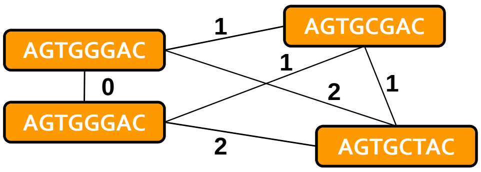
MotifCut Algorithm

- Convert sequence into a collection of k-mers
  - Each overlap/duplicate considered distinct



Motif Graph Representation

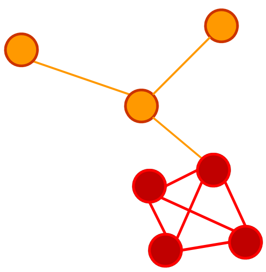
- Nodes are k-mers
- Edge weights are distances between k-mers
  - How the edge weights are determined? (later)



- Same k-mer node can appear multiple times.
  - If a certain k-mer appears frequently in the input sequences, there are many nodes for that k-mer.
- Finding over-represented similar k-mers → Finding maximum density subgraph (MDS)

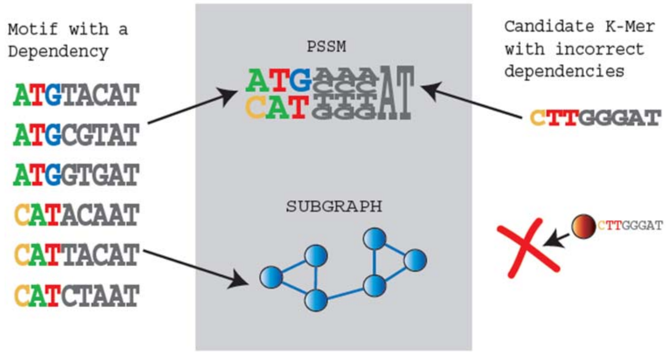
Motif Finding

- Find highest density subgraph



- Density is defined as sum of edge weights per node: graph density $\lambda=|E|/|V|$.
- Find the maximum density subgraph (MDS)

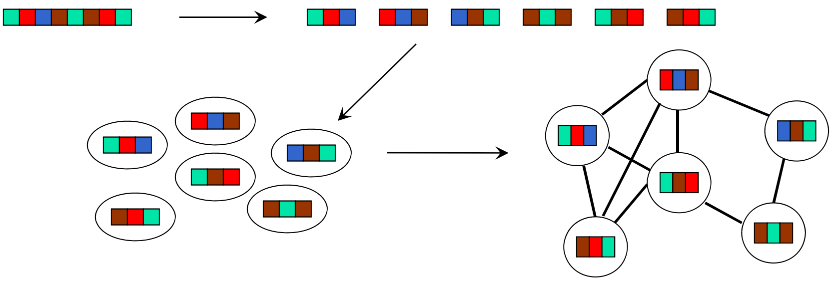
Motif Dependency in MDS



9

MotifCut Algorithm

- Read input sequences
- Generate graph as previously described
  - K-mers are generated by shifting one base pair
  - Each k-mer in the sequence gets a node, including identical k-mers
  - Graph contains as many nodes as there are base pairs
  - Connect edges with weights based on distances between nodes

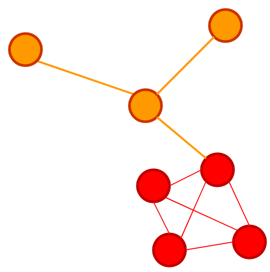


- Find maximum density subgraphs (MDSs)

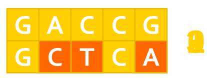
10

Edge Weights

- **Semantics:** Edge weight is the likelihood of two k-mers to be in the same motif



- Use Hamming distance as a way to quantify distance between k-mers



Edge Weights

- Let's make this a bit more precise:
  - For every pair of vertices (v_i, v_j) create an edge with weight w_{ij}
  - w_{ij} = f(Hamming distance between k-mers in v_i, v_j)

$$w_{ij} = \frac{\Pr(v_i \in M \mid v_j \in M) + \Pr(v_j \in M \mid v_i \in M)}{\boxed{\theta(\Pr(v_i \in B)) + \theta(\Pr(v_j \in B))}}$$

↑ Background distribution

M → k-mers of binding site
B → background k-mers

- But how to compute $Pr(v_i \in M \mid v_j \in M)$?

- Simulate it!
  - Way too many variables to account for analytically: Background model, kmer length, hamming distance, etc…

Maximum Density Subgraph

- Standard graph theory method
  - Max-flow / min-cut: simple and easy to implement
  - However, its running time is $O(nm \log(n^2m))$, where n is the number of vertices and m is the number of edges
- Need faster method
- Developed heuristic approach that utilizes max-flow / min-cut method with modifications

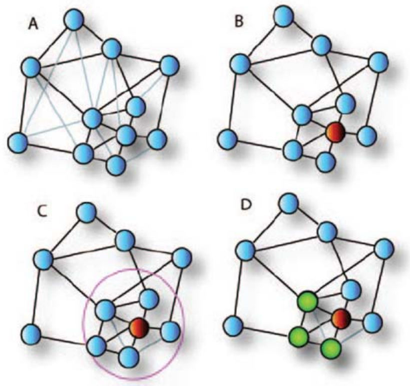
13

MotifCut Algorithm

- Find the maximum density subgraph (MDS)
- MDS optimization

Remove all edges below a certain threshold

Pick one vertex (do this for every vertex)



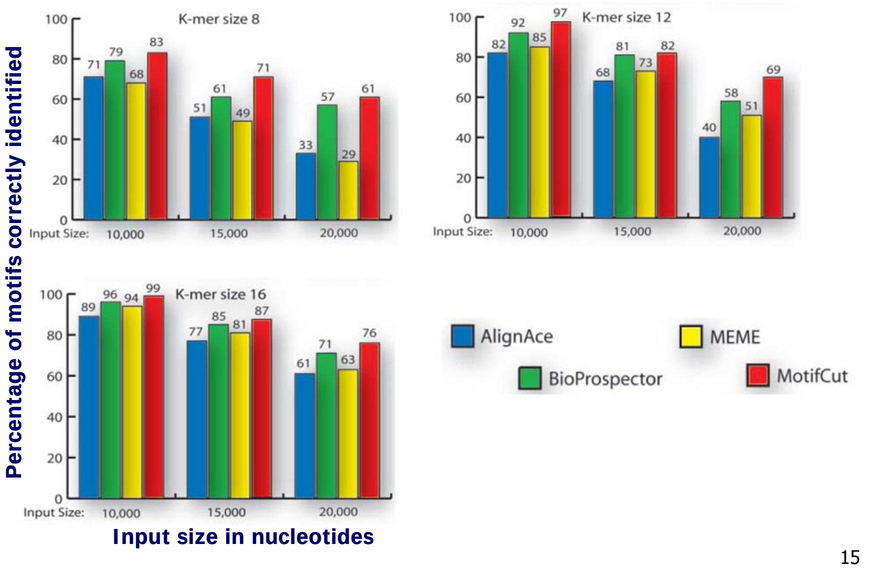
Put back all neighboring edges for that vertex

Use standard algorithm to calculate densest subgraph

Repeat for every vertex

14

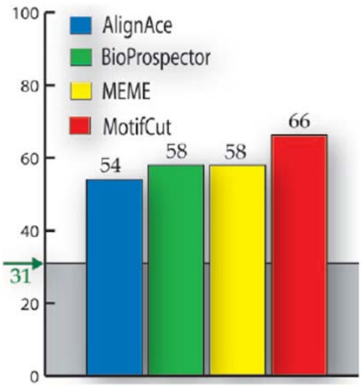
Synthetic Experiment Results



MotifCut: regulatory motifs finding with maximum density subgraphs. Fratkin et al. Bioinformatics (2006).

Yeast Test Results

- Gold standard data (Harbinson et al., 2004)



Outline

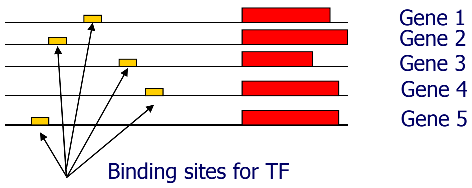
- Regulatory motif finding
  - More computational methods
    - Greedy search method (CONSENSUS)
    - Phylogenetic foot-printing method
    - Graph-based methods (MotifCut)
  - Before/ after motif finding
- Inferring signaling networks

What After Motif Finding ?

- Experiments to confirm results
- DNaseI footprinting & gel-shift assays
- Tells us which subsequences are the binding sites

Before Motif Finding

- How do we obtain a set of sequences on which to run motif finding ?
- In other words, how do we get genes that we believe are regulated by the same transcription factor ?
- Two high-throughput experimental methods: ChIP-chip and microarray.

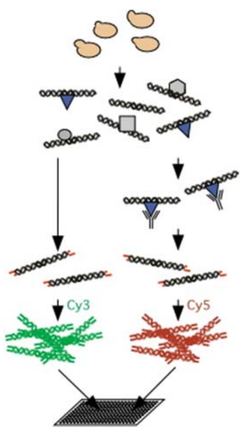


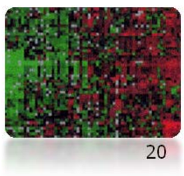
19

Before Motif Finding

- ChIP-chip
  - Take a particular transcription factor TF
  - Take hundreds or thousands of promoter sequences
  - Measure how strongly TF binds to each of the promoter sequences
  - Collect the set to which TF binds strongly, do motif finding on these

- Gene expression data
  - Collect set of genes with similar expression (activity) profiles and do motif finding on these.





20

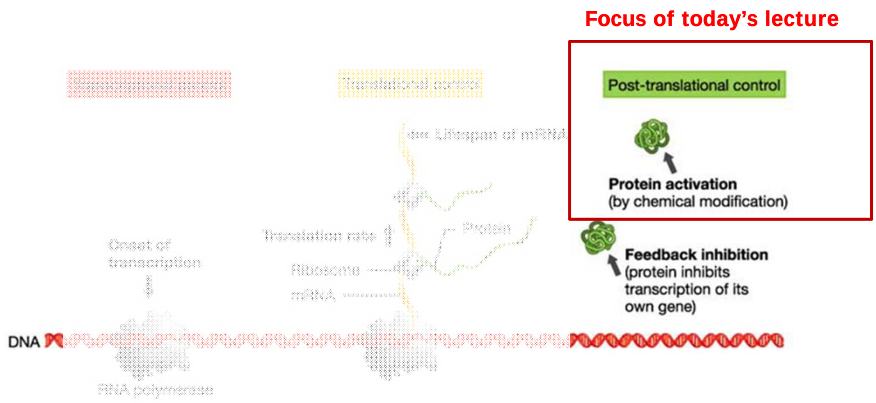
Outline

- Regulatory motif finding
  - More computational methods
  - Before/ after motif finding

- Inferring signaling networks
  - Signaling network
  - Flow cytometry
  - Bayesian networks

Gene Regulation

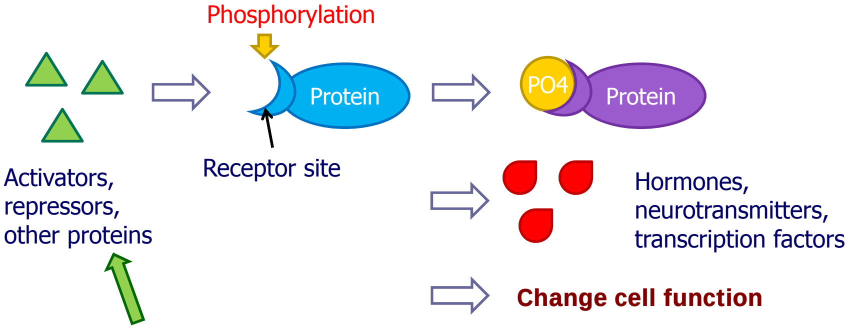
- Transcriptional regulation is one of many regulatory mechanisms in the cell



Source: Mallery, University of Miami

Post-translational Modification

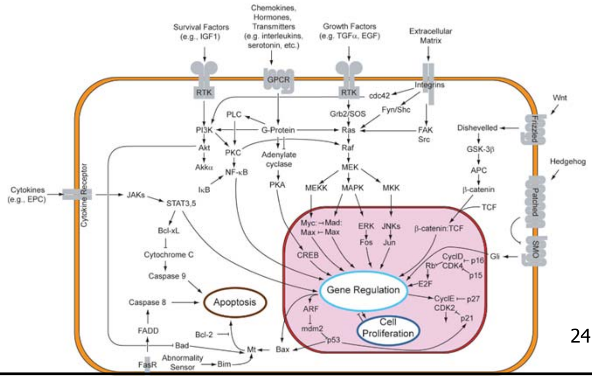
- Most proteins undergo some form of modification following translation.
- Phosphorylation is the most studied and best understood post-translation modification.
  - Addition of a phosphate ($PO4^{3-}$) group to a protein
  - It activates or deactivates many protein enzymes



- Interventions – artificially introducing chemicals which activate/repress the phosphorylation of a protein.

Cellular Signaling Networks

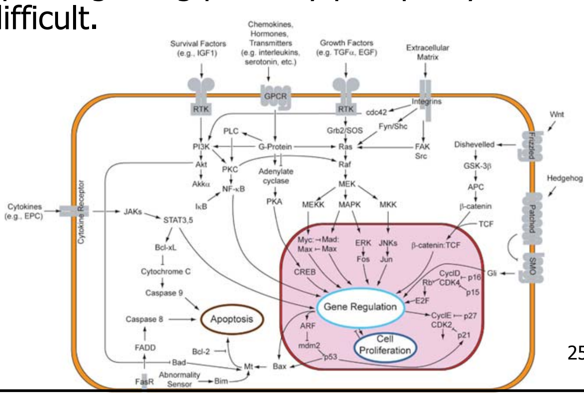
- Cellular signaling
  - Part of a complex system of communication that governs basic cellular activities and coordinates cell actions.
  - The ability of cells to perceive and correctly respond to their microenvironment is the basis of development, tissue repair, and immunity as well as normal tissue homeostasis.



Overview of signal transduction pathways
Source: Wikipedia

Cellular Signaling Networks

- Reversible phosphorylation is a major regulatory mechanism controlling the signaling pathway.
  - Many signaling pathways, including the insulin/IGF-1 signaling pathway, transduce signals from the cell surface to downstream targets via tyrosine kinases and phosphatases.
- Elucidating complex signaling pathway phosphorylation events can be difficult.



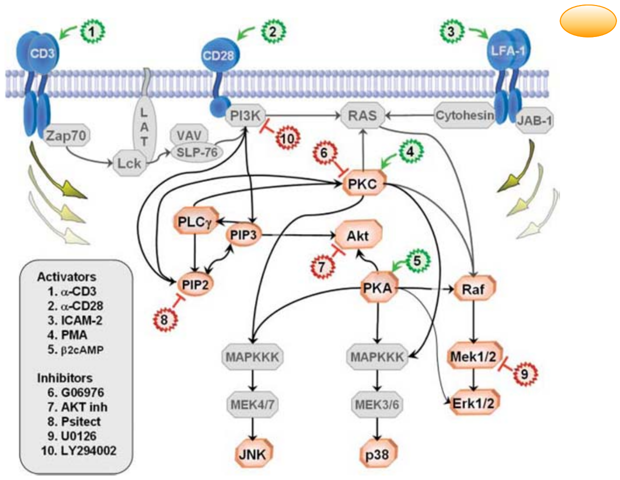
Overview of signal transduction pathways

Source: Wikipedia

25

Signaling Networks – Example

- Classic signaling network and points of intervention
- Human T cell (white blood cell)



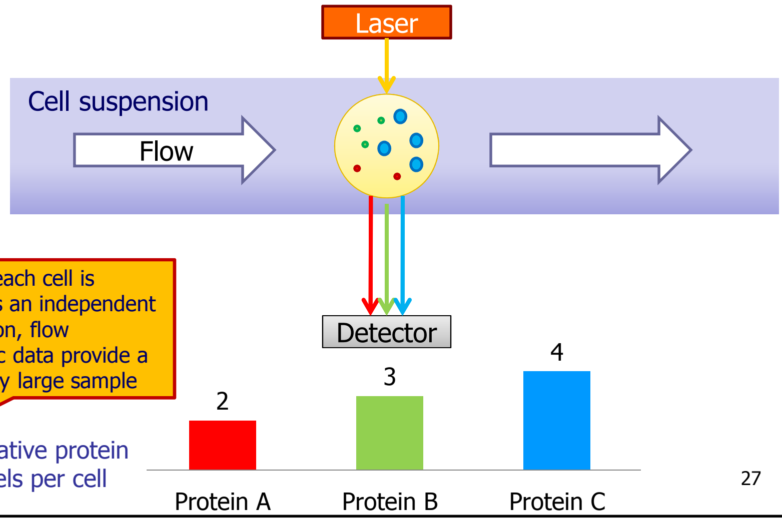
= Measured proteins

26

Source: Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Sachs et al. Science (2005).

Flow Cytometry

- Quantitatively measure as given proteins' expression levels and their phosphorylation states.



Because each cell is treated as an independent observation, flow cytometric data provide a statistically large sample

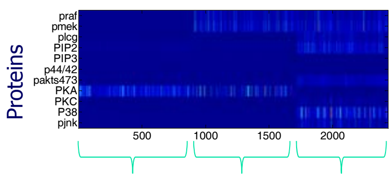
Relative protein levels per cell

27

Flow Cytometry Data

Signal Networks: Flow Cytometry Data

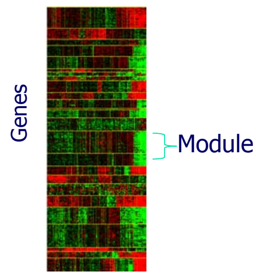
Cells



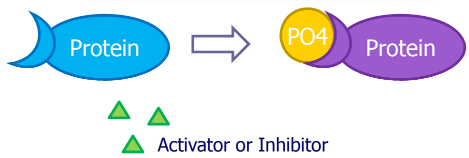
Intervention conditions

Regulatory Networks: Gene Expression Data

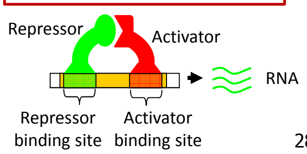
Experiments



Phosphorylated protein levels



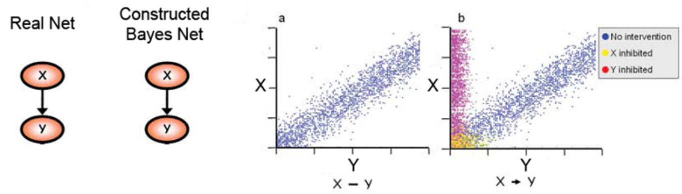
Gene expression levels



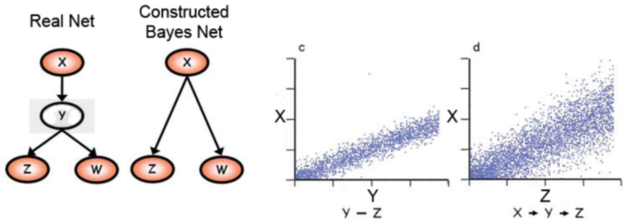
28

Bayesian Networks

- Directionality via intervention



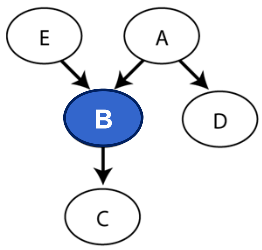
- Structure preservation



29

Bayesian Networks

- Directed Acyclic Graphs (DAGs)



Conditional independence

$P(B \mid D, A, E) = P(B \mid A, E)$

Parents of B (A, E)

$(B \perp D \mid A, E)$

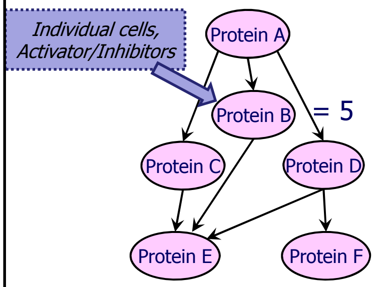
Independent ($\perp$)

30

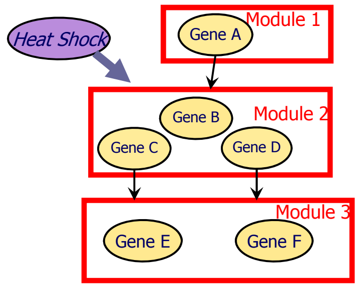
Bayesian network analysis of signaling networks: a primer. Pe'er D. Science STKE (2005).

Bayesian Networks

- Signal network (protein regulation)
- Regulatory networks (gene regulation)

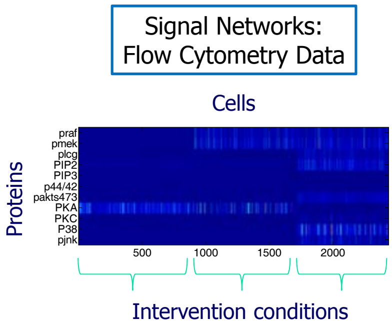


Discrete phosphorylated protein levels

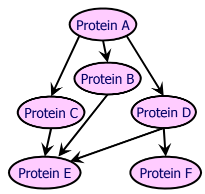


Continuous gene expression levels

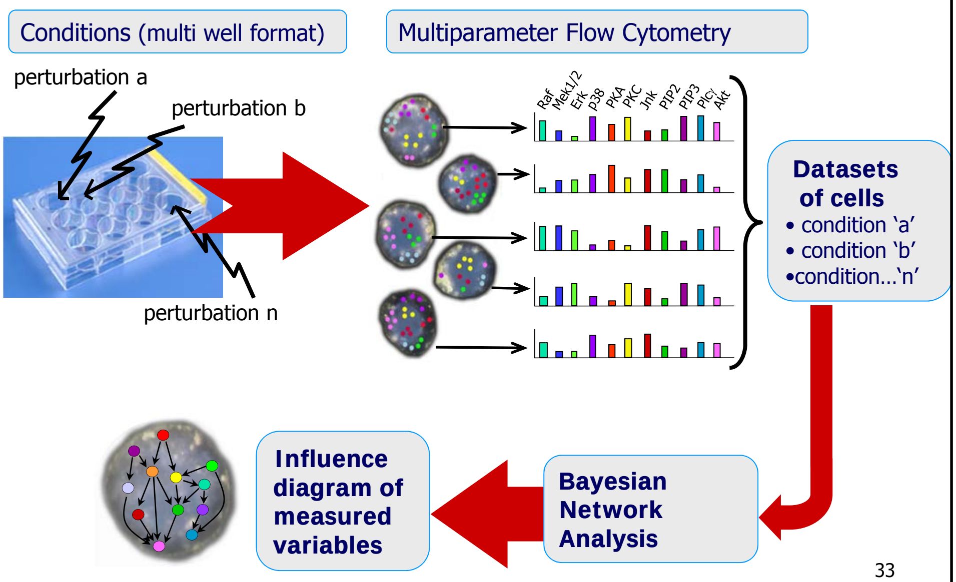
Structure Learning



Learn DAG structure



Overview



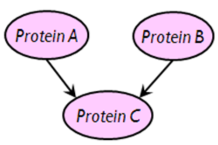
33

Source: Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Sachs et al. Science (2005).

Local Probability Model

- Conditional Probability Tables

D = Data G=Graph
θ = CPT values for each node X
θ_{ijk} = P(Xi=k | Parents(Xi)=j)
N_{ijk} = # times Xi=k and Parents(Xi)=j in the Data



Conditional Probability Table (CPT)

k →, j ↓

A	B	P(C=0|Pa)		P(C=1|Pa)	
0	0	θ_{C10}	+	θ_{C11}	= 1
0	1	θ_{C20}	+	θ_{C21}	= 1
1	0	θ_{C30}	+	θ_{C31}	= 1
1	1	θ_{C40}	+	θ_{C41}	= 1

θ_{ijk}

34

Maximum Likelihood Score

- Find G that maximizes:
  P(Data=D | Graph=G, Θ_{MLE})

D = Data G=Graph
θ = CPT values for each node X
θ_{ijk} = P(Xi=k | Parents(Xi)=j)
N_{ijk} = # times Xi=k and Parents(Xi)=j in the Data

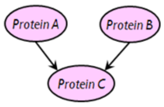
K=#discrete levels of X

N_{ijk} = # times Xi=k and Parents(Xi)=j in the Data

$$= \prod_{i=1}^{\#proteins} \prod_{j=1}^{\#parent\ states} \left(\prod_{k=1}^{K} \theta_{ijk}^{N_{ijk}} \right) \Rightarrow \Theta_{ijk}^{ML} = N_{ijk} / \sum_k N_{ijk}$$

θ_{ijk} = P(Xi=k | Parents(Xi)=j)

Conditional Probability Table (CPT)



k →

j ↓

A	B	P(C=0|Pa)		P(C=1|Pa)	
0	0	Θ_{C10}	+	Θ_{C11}	= 1
0	1	Θ_{C20}	+	Θ_{C21}	= 1
1	0	Θ_{C30}	+	Θ_{C31}	= 1
1	1	Θ_{C40}	+	Θ_{C41}	= 1

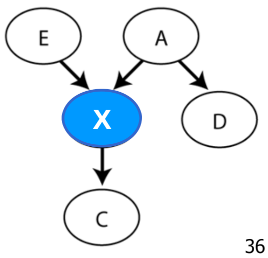
θ_{ijk}

35

Structure Score

- Bayesian score (Structure | Data)
  = log P(Data | Structure) + log P(Structure)

- Decomposability
  log P(Data | Structure)
  = Σ_X FamScore(X, Parents(X) | Data)



36

Bayesian network analysis of signaling networks: a primer. Pe'er D. Science STKE (2005).

Structure Score

D = Data G=Graph
θ = CPT values for each node X
θ_{ijk} = P(Xi=k | Parents(Xi)=j)
N_{ijk} = # times Xi=k and Parents(Xi)=j in the Data

- P(Data=D | Graph=G)
  = ∫ P(D|G,θ) P(θ|G) dθ

P(D,θ|G)

Multinomial (see page 35)

Dirichlet prior ~ Dir(α)

K=#discrete levels of X

N_{ijk} = # times Xi=k and Parents(Xi)=j in the Data

Dirichlet normalizer

$$P(D|G) = \prod_{i=1}^{\#proteins} \prod_{j=1}^{\#parent\ states} \int_{\Theta_{ij}} \left(\prod_{k=1}^{K} \theta_{ijk}{}^{N_{ijk}} \right) \left(\frac{1}{B(\alpha_{ij})} \prod_{k=1}^{K} \theta_{ijk}{}^{\alpha_{ijk}-1} \right) d\theta_{ij}$$

Θ_{ij} = Simplex $\{\sum_k \theta_{ijk} = 1\}$

θ_{ijk} = P(Xi=k | Parents(Xi)=j)

D. Heckerman. A Tutorial on Learning with Bayesian Networks. 1999, 1997, 1995.
G. Cooper E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning, 9, 309-347. 1992.

Structure Score

D = Data G=Graph
θ = CPT values for each node X
θ_{ijk} = P(Xi=k | Parents(Xi)=j)
N_{ijk} = # times Xi=k and Parents(Xi)=j in the Data

Dirichlet normalizer

$$B(\alpha) = \int_{\Delta^K} \prod_{k=1}^{K} \theta_k{}^{\alpha_k - 1} d\theta = \frac{\prod_{k=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{k=1}^{K} \alpha_i)}$$

$$P(D|G) = \prod_{i=1}^{\#proteins} \prod_{j=1}^{\#parent\ states} \frac{1}{B(\alpha_{ij})} \int_{\Theta_{ij}} \left(\prod_{k=1}^{K} \theta_{ijk}{}^{N_{ijk}+\alpha_{ijk}-1} \right) d\theta_{ij}$$

$$P(D|G) = \prod_{i=1}^{\#proteins} \prod_{j=1}^{\#parent\ states} \frac{B(\alpha_{ij} + N_{ij})}{B(\alpha_{ij})}$$

$$P(D|G) = \prod_{i=1}^{\#proteins} \prod_{j=1}^{\#parent\ states} \frac{\Gamma(\sum_{k=1}^{K} \alpha_{ijk})}{\Gamma(\sum_{k=1}^{K} \alpha_{ijk} + N_{ijk})} \prod_{k=1}^{K} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

G. Cooper E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning, 9, 309-347. 1992.

Structure Score

D = Data G=Graph
θ = CPT values for each node X
θ_{ijk} = P(Xi=k | Parents(Xi)=j)
N_{ijk} = # times Xi=k and Parents(Xi)=j in the Data

$$P(D|G) = \prod_{i=1}^{\#proteins} \prod_{j=1}^{\#parent\ states} \frac{\Gamma(\sum_{k=1}^{K} \alpha_{ijk})}{\Gamma(\sum_{k=1}^{K} \alpha_{ijk} + N_{ijk})} \prod_{k=1}^{K} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

$$\text{FamScore}(X_i, Pa_i|D) = \log \prod_{j=1}^{\#parent\ states} \frac{\Gamma(\sum_{k=1}^{K} \alpha_{ijk})}{\Gamma(\sum_{k=1}^{K} \alpha_{ijk} + N_{ijk})} \prod_{k=1}^{K} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

$$\text{Score}(G|D) = \sum_{i=1}^{\#proteins} \text{FamScore}(X_i, Pa_i|D)$$

$$\text{Score}(G|D) = \log P(D|G) + \log P(G)$$

Bayesian Score

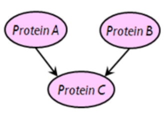
- P(Data=D | Graph=G)
  = ∫ P(D|G,θ) P(θ|G) dθ
  (P(D|G,θ): Multinomial; P(θ|G): Dirichlet prior ~ Dir(α))

D = Data G=Graph
θ = CPT values for each node X
θ_{ijk} = P(Xi=k | Parents(Xi)=j)
N_{ijk} = # times Xi=k and Parents(Xi)=j in the Data

$$= \prod_{i=1}^{\#proteins} \prod_{j=1}^{\#parent\ states} \int_{\Theta_{ij}} \left(\prod_{k=1}^{K} \theta_{ijk}^{N_{ijk}} \right) \left(\frac{1}{B(\alpha_{ij})} \prod_{k=1}^{K} \theta_{ijk}^{\alpha_{ijk}-1} \right) d\theta_{ij}$$

$$\Rightarrow \Theta_{ijk}^{BS} = (N_{ijk} + \boldsymbol{\alpha_{ijk}}) / \sum_k (N_{ijk} + \boldsymbol{\alpha_{ijk}})$$

"Imaginary" counts



Conditional Probability Table (CPT)

k →, j ↓

A	B	P(C=0|Pa)		P(C=1|Pa)	
0	0	Θ_{C10}	+	Θ_{C11}	= 1
0	1	Θ_{C20}	+	Θ_{C21}	= 1
1	0	Θ_{C30}	+	Θ_{C31}	= 1
1	1	Θ_{C40}	+	Θ_{C41}	= 1

θ_{ijk}