Gene Regulatory Networks II

02-710 Computational Genomics

Seyoung Kim

Goal: Discover Structure and Function of Complex systems in the Cell

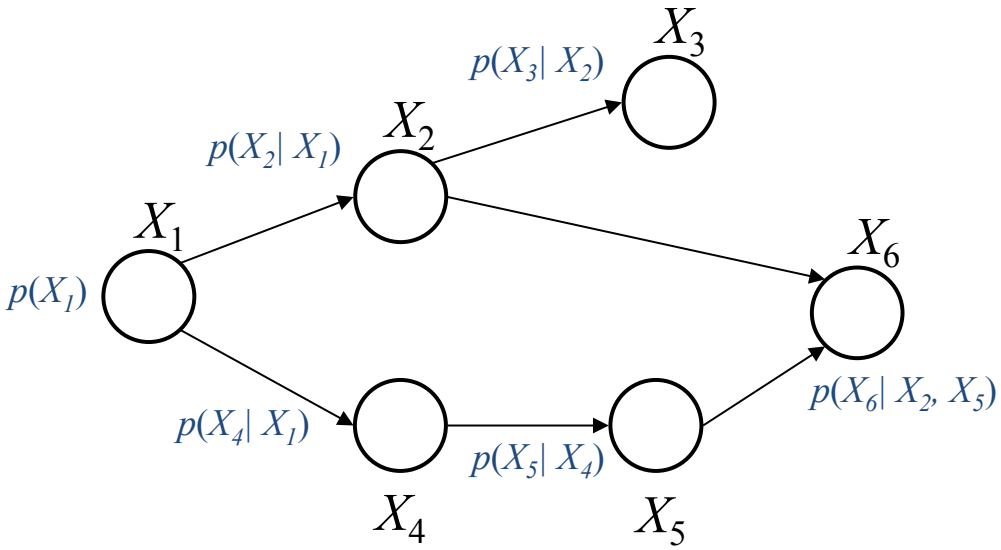
- Identify the different regulators and their target genes that are involved in the system.
- Represent the relationship between regulators and their target genes as a network
  - Nodes: entities (regulators, target genes)
  - Edges: regulatory relationship

High-level goal: Use high throughput data to discover patterns of combinatorial regulation and to understand how the activity of genes involved in related biological processes is coordinated and interconnected.

Overview

- Bayesian networks (network with directed edges): Module networks and their extensions
  - Module network (Segal et al., Nature Genetics 2003): Gene module's activity is determined by their expression levels of regulator genes
  - Geronemo (Lee et al., PNAS 2006): Gene module's activity is determined by their expression levels of regulator gene and SNPs
  - Lirnet (Lee et al., PLoS Genetics 2009): incorporates prior knowledge
  - CONEXIC (Akavia et al., Cell 2010): cancer data analysis for copy number variation and gene expression data
- Gaussian graphical models (network with undirected edges) and their extensions

Probabilistic Graphical Models

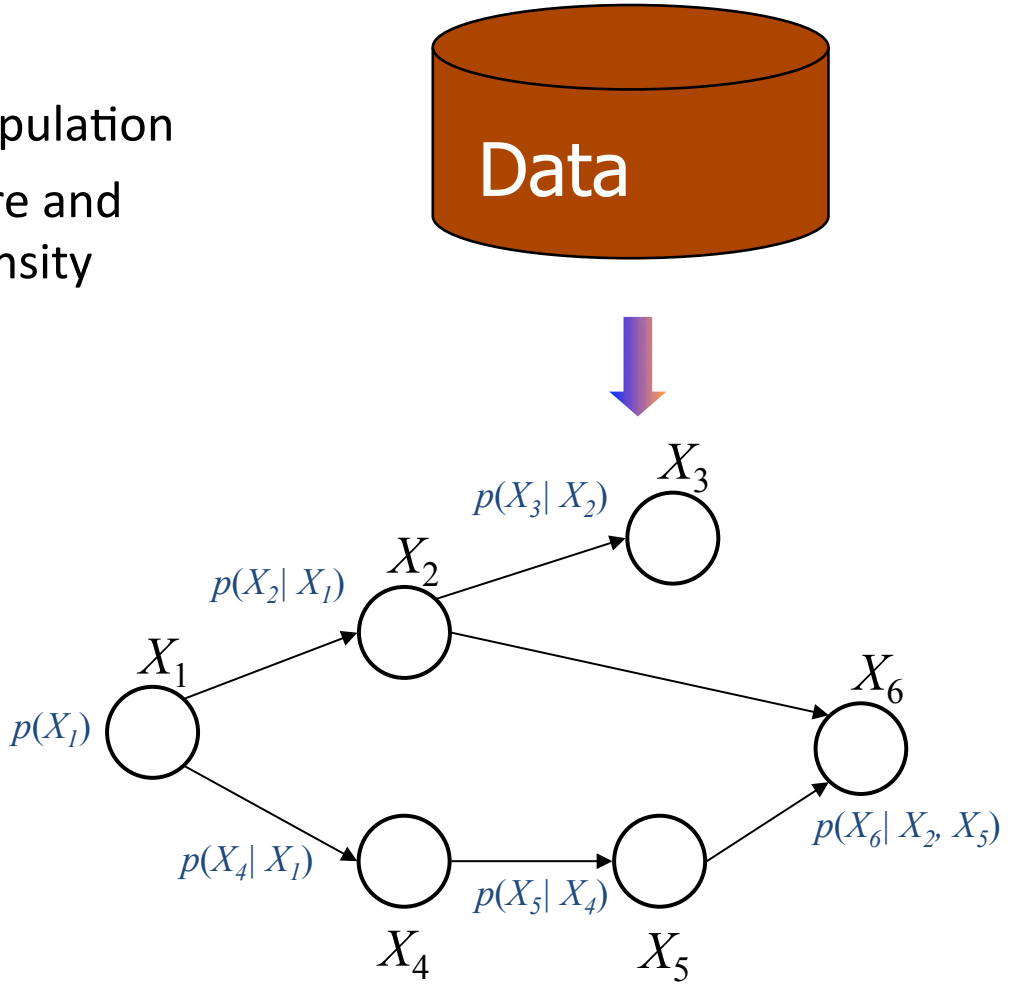


- The joint distribution on (X_1, X_2,…, X_N) factors according to the "parent-of" relations defined by the edges E :

$$p(X_1, X_2, X_3, X_4, X_5, X_6) = p(X_1)\, p(X_2|\ X_1) p(X_3|\ X_2)\, p(X_4|\ X_1) p(X_5|\ X_4) p(X_6|\ X_2, X_5)$$

Learning Bayesian Networks

- Density estimation
  - Model data distribution in population
  - Learn both the graph structure and the associated probability density
  - Probabilistic inference:
    - Prediction
    - Classification



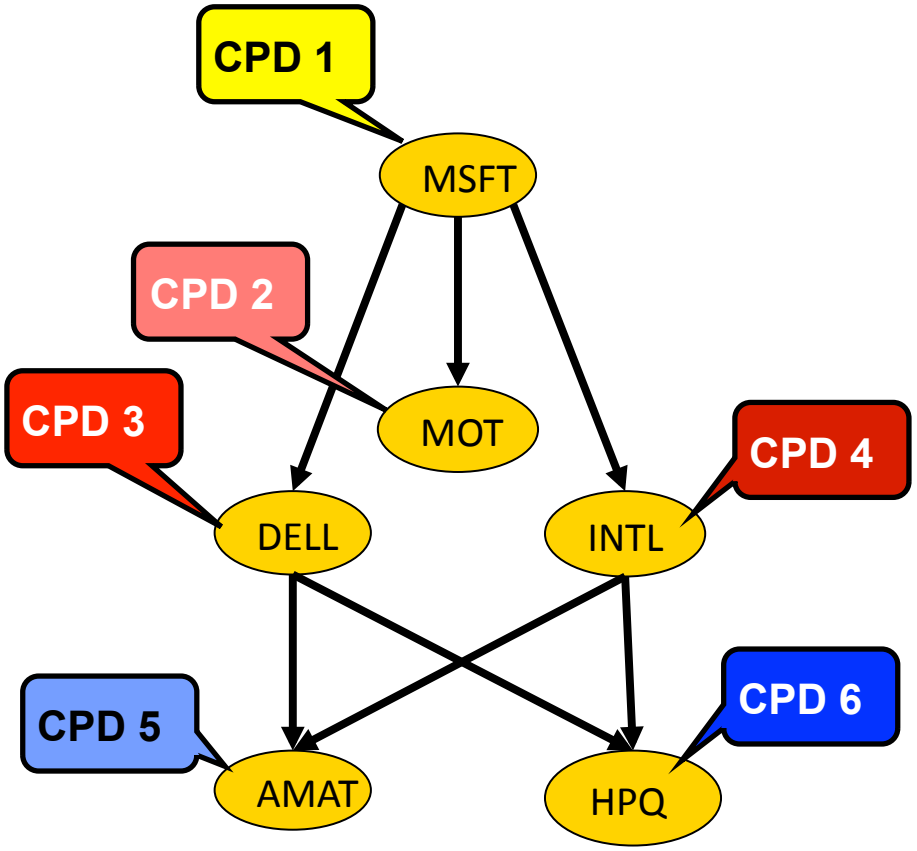
$$p(X_1, X_2, X_3, X_4, X_5, X_6) = p(X_1)\ p(X_2|\ X_1)p(X_3|\ X_2)\ p(X_4|\ X_1)p(X_5|\ X_4)p(X_6|\ X_2, X_5)$$

The Module Network Idea

- Unlike methods that represent individual genes, module-based methods explicitly model modular structure. This helps in:
  - Reducing dependence on possibly noisy measurements for individual genes, by combining information among genes in the same module
  - Elevated statistical significance

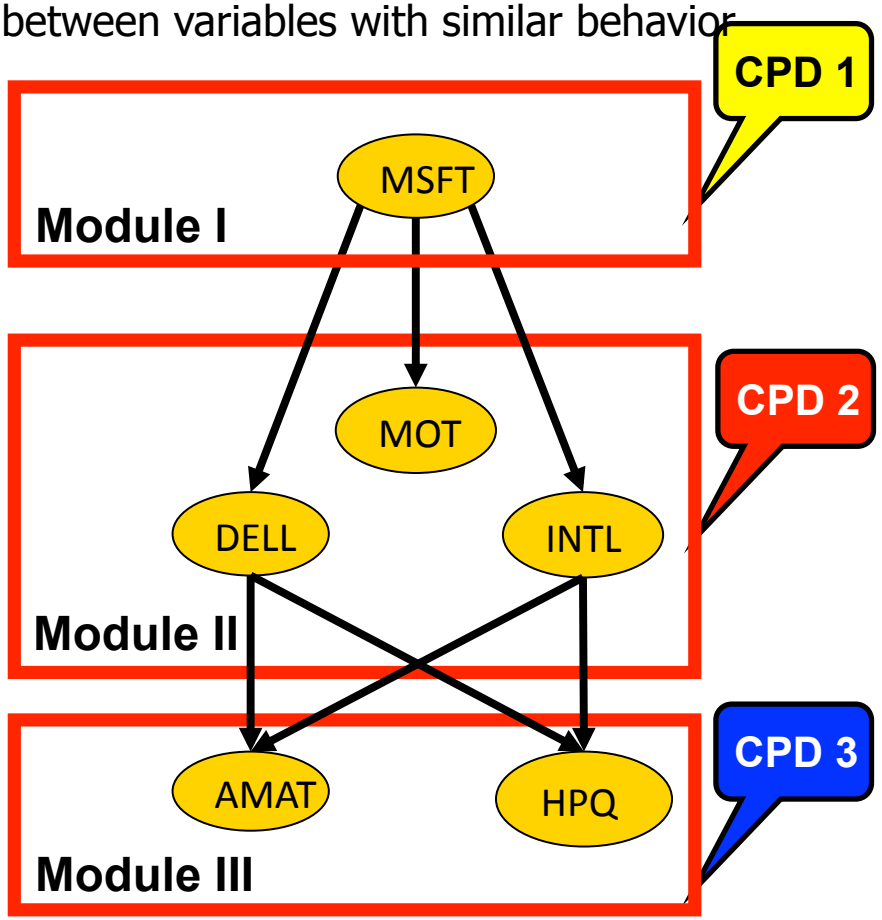
The Module Network Idea

Bayesian Network



Module Network

Share parameters and dependencies between variables with similar behavior



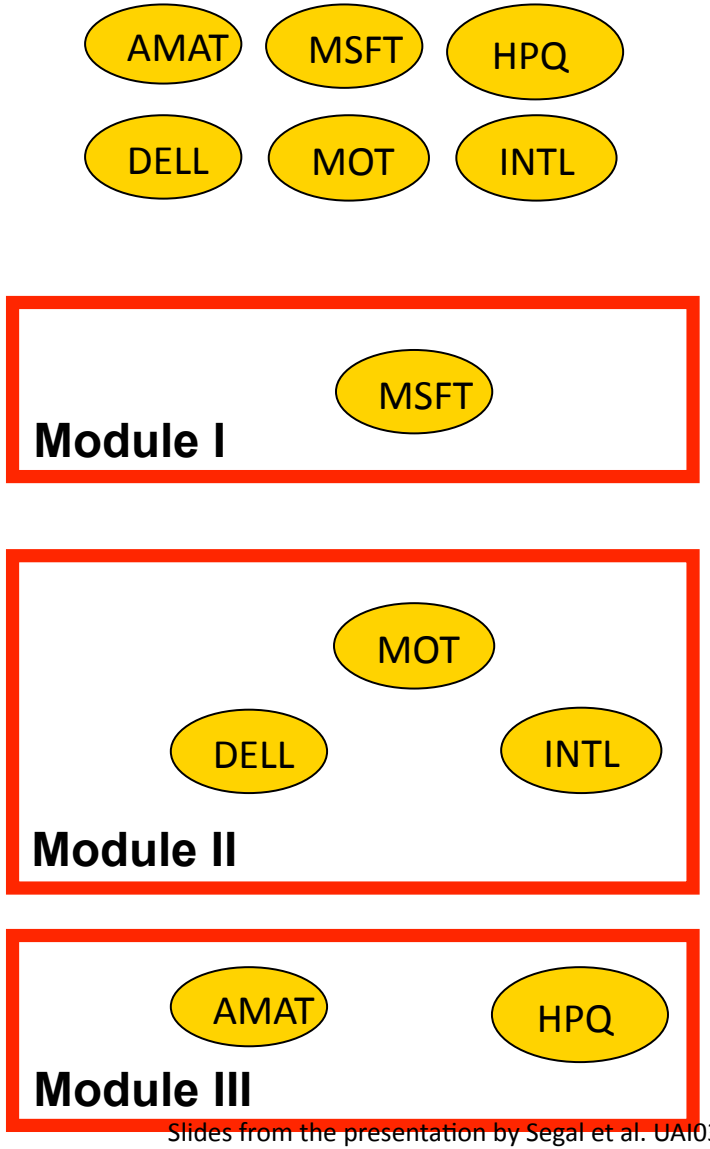
Slides from the presentation by Segal et al. UAI03

Learning Module Network

- Module Network
  - Model definition
  - Learning the model
- Experimental results

Module Network Components

- Module Assignment Function A(•)
  - A(MSFT)=M_I
  - A(MOT)=A(DELL)=A(INTL) =M_{II}
  - A(AMAT)= A(HPQ)=M_{III}

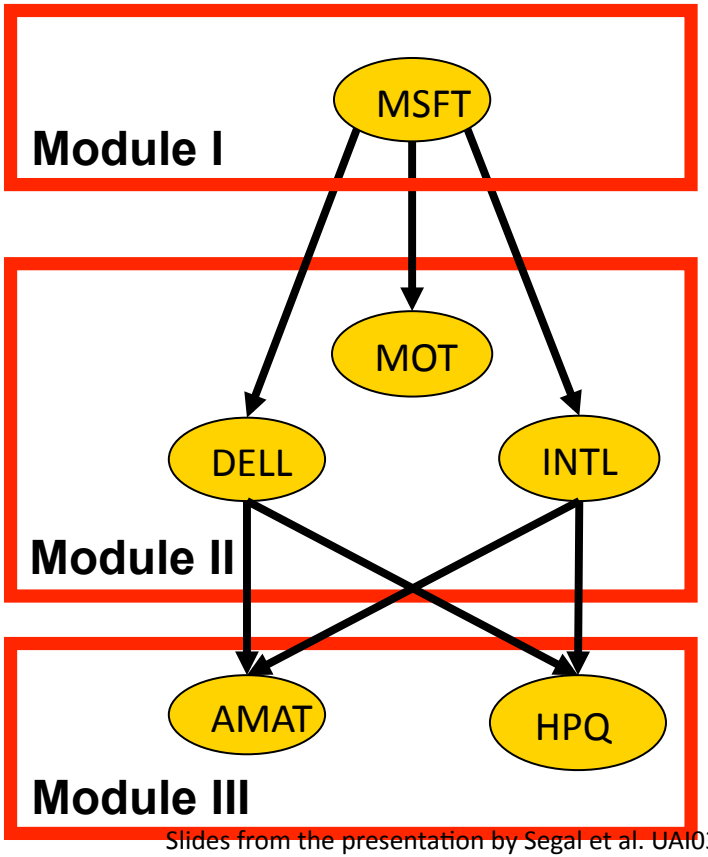


Slides from the presentation by Segal et al. UAI03

Module Network Components

- Module Assignment Function

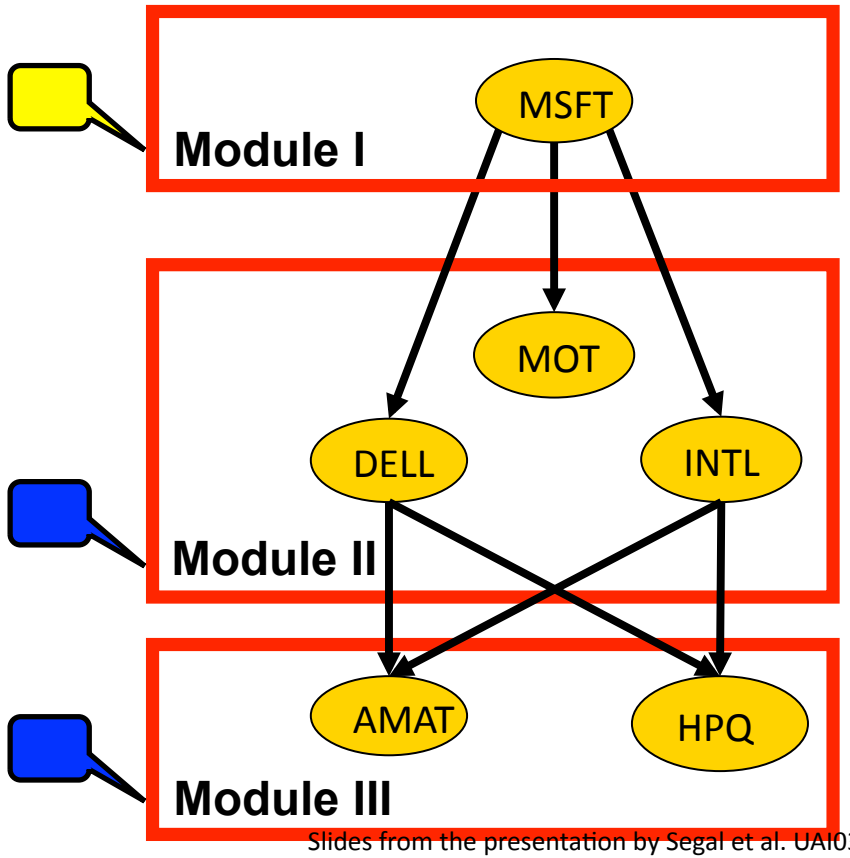
- Set of parents for each module
  - $Pa(M_I)=\emptyset$
  - $Pa(M_{II})=\{MSFT\}$
  - $Pa(M_{III})=\{DELL, INTL\}$



Slides from the presentation by Segal et al. UAI03

Module Network Components

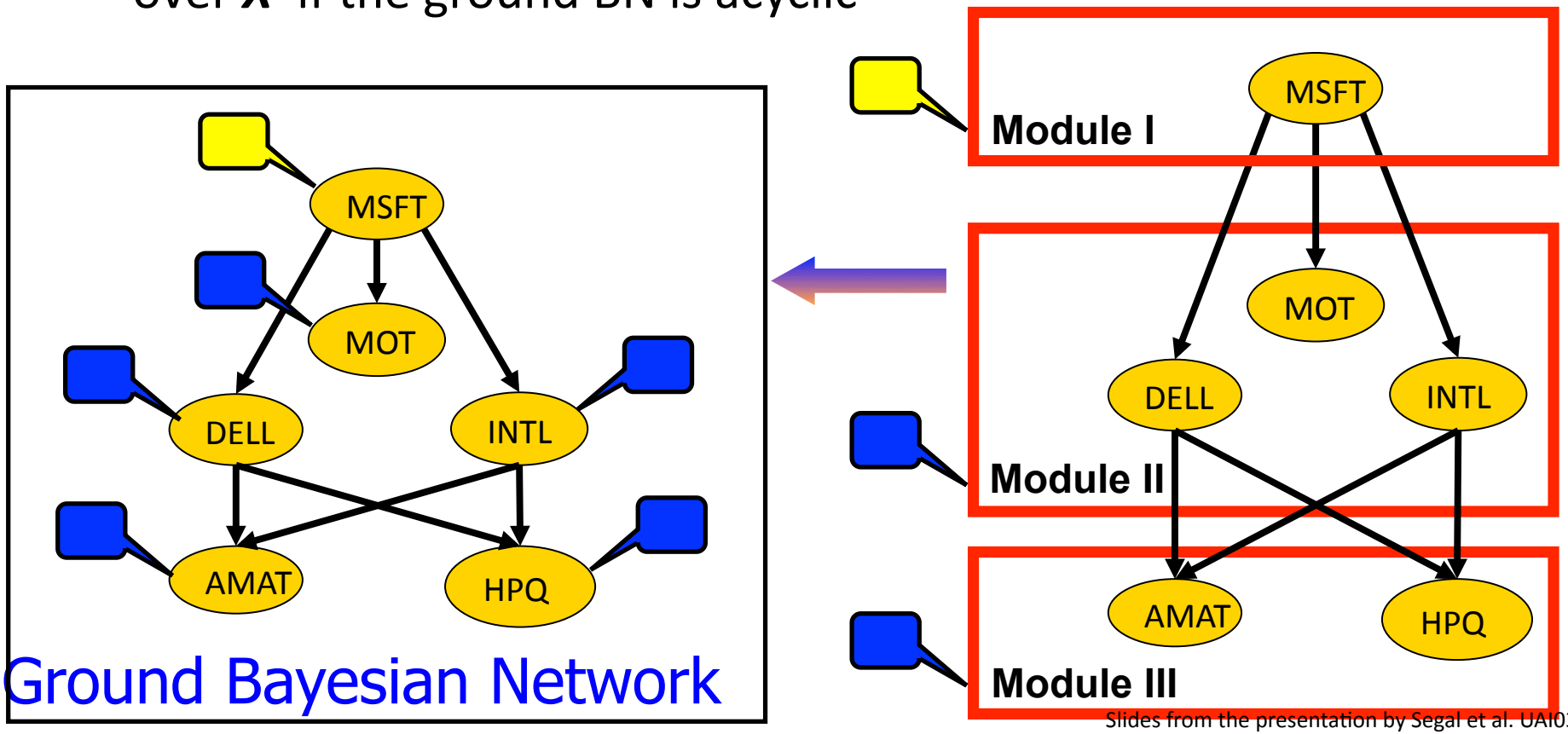
- Module Assignment Function
- Set of parents for each module
- Conditional probability density (CPD) template for each module



Slides from the presentation by Segal et al. UAI03

Ground Bayesian Network

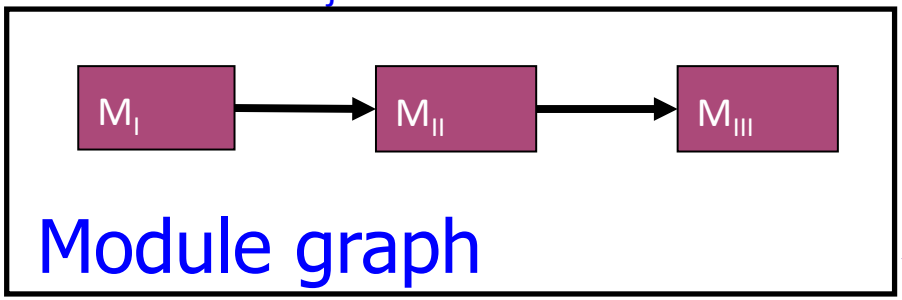
- A module network induces a *ground BN* over ***X***
- A module network defines a coherent probabilty distribution over ***X*** if the ground BN is acyclic

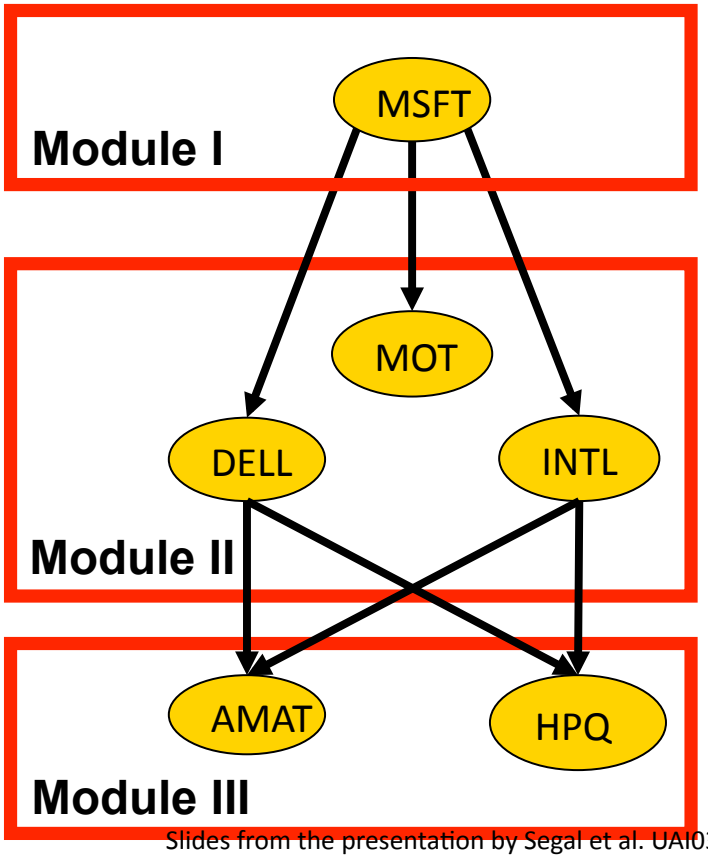


Slides from the presentation by Segal et al. UAI03

Module Graph

- Nodes correspond to modules
- $M_i \rightarrow M_j$ if at least one variable in M_i is a parent of M_j





Theorem:

The ground BN is acyclic if the module graph is acyclic

Acyclicity checked efficiently using the module graph

Slides from the presentation by Segal et al. UAI03

Learning Module Network

- Module Network
  - Model definition
  - Learning the model
- Experimental results

Learning Overview

- Given data ***D***, find assignment function ***A*** and structure ***S*** that maximize the Bayesian score

$$\text{Score}(S,A:D) = \underbrace{\log P(D \mid S,A)}_{\text{Marginal likelihood}} + \underbrace{\log P(S,A)}_{\text{Assignment / structure prior}}$$

- Marginal data likelihood

$$P(D \mid S,A) = \int \underbrace{P(D \mid S,A,\theta)}_{\text{Data likelihood}} \underbrace{P(\theta \mid S,A)}_{\text{Parameter prior}} d\theta$$

Slides from the presentation by Segal et al. UAI03

Bayesian Score Decomposition

- Bayesian score decomposes by modules

$$score(S,A:D) = \sum_{j=1}^{k} score_{M_j}(Pa_{M_j}, X^j : D)$$

Module j parents

Module j variables

Bayesian Score Decomposition

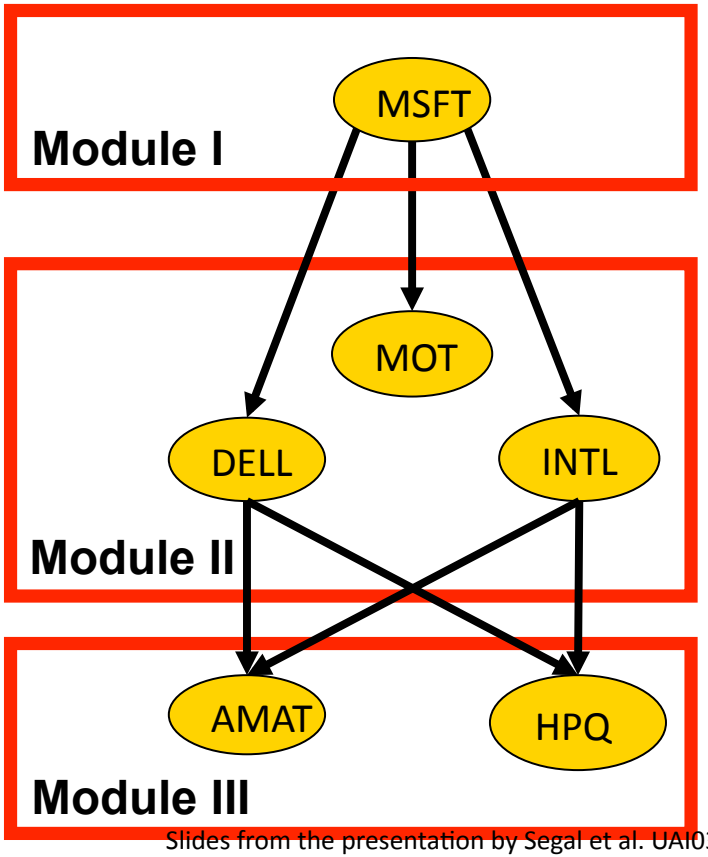
- Bayesian score decomposes by modules

$$score(S,A:D) = \sum_{j=1}^{k} score_{M_j}(Pa_{M_j}, X^j : D)$$

$$score = score_{M_1}(\varnothing, X^1 : D) +$$

$$score_{M_2}(MSFT, X^2 : D) +$$

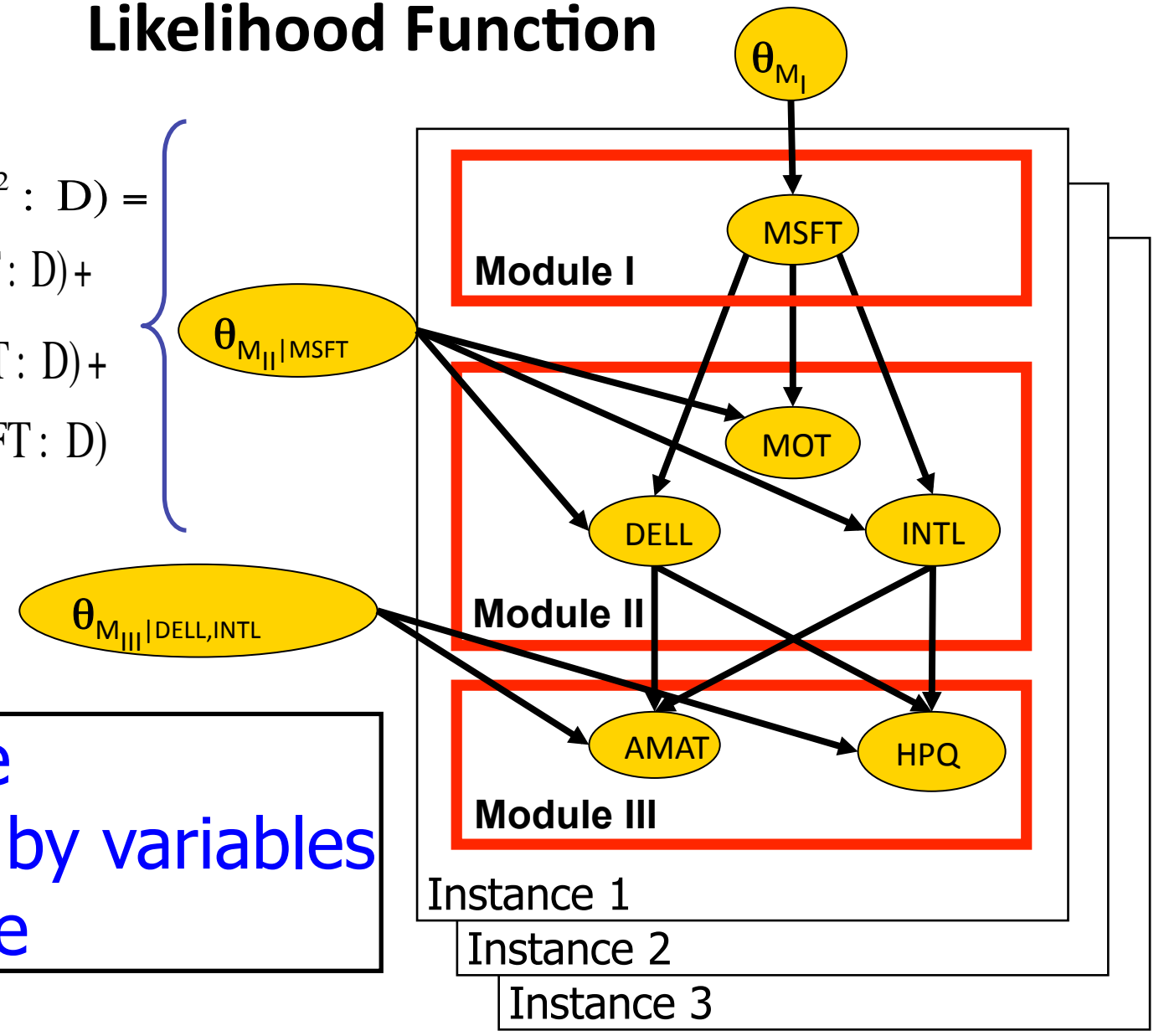
$$score_{M_3}(\{DELL, INTL\}, X^3 : D)$$



Slides from the presentation by Segal et al. UAI03

Likelihood Function

$$\text{Score}_{M_2}(\text{MSFT}, X^2 : D) =$$
$$\text{Score}(\text{DELL},\text{MSFT} : D) +$$
$$\text{Score}(\text{MOT},\text{MSFT} : D) +$$
$$\text{Score}(\text{INTL},\text{MSFT} : D)$$



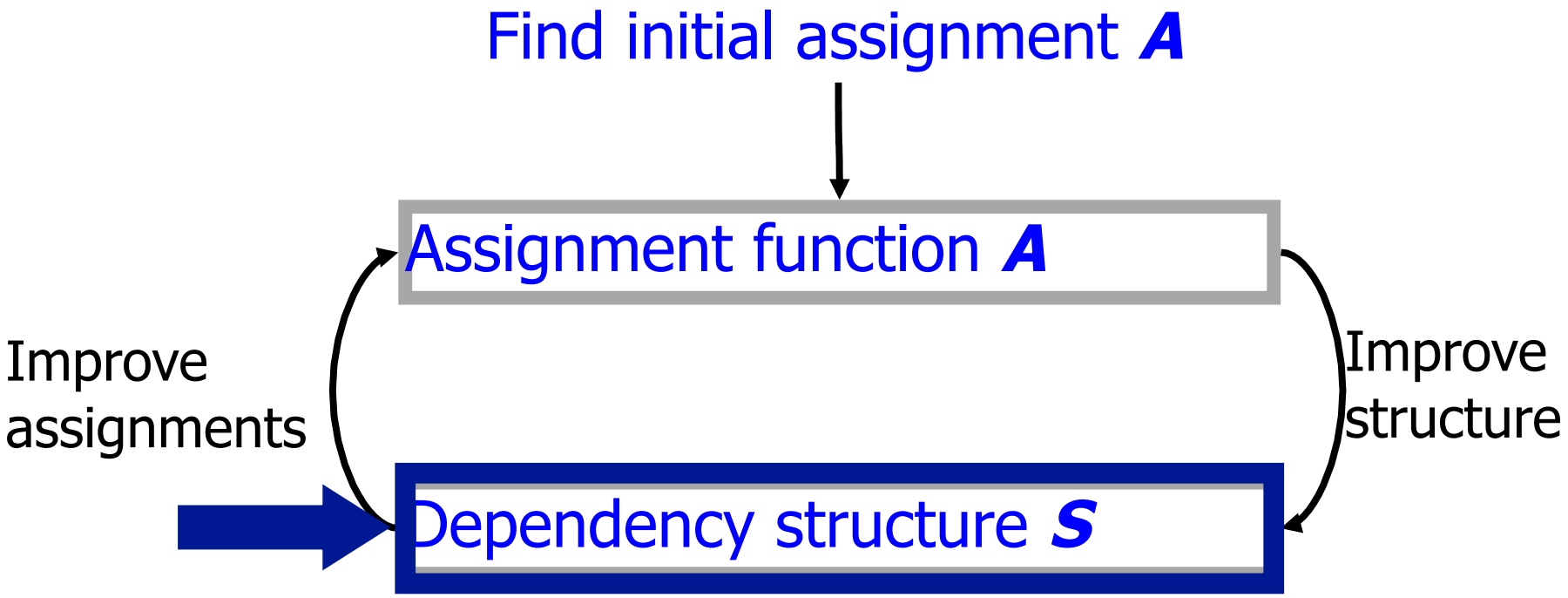
Module score decomposes by variables in the module

Slides from the presentation by Segal et al. UAI03

Algorithm Overview

- Find assignment function ***A*** and structure ***S*** that maximize the Bayesian score

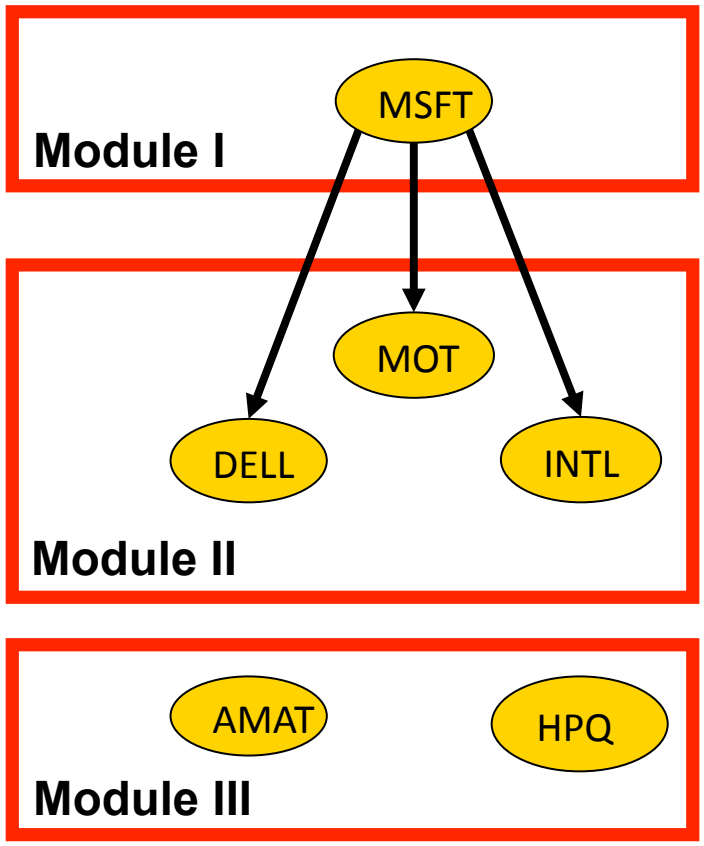
$$Score(S,A : D) = logP(D \mid S,A) + logP(S,A)$$



Slides from the presentation by Segal et al. UAI03

Learning Dependency Structure

- Heuristic search with operators
  - Add/delete parent for module

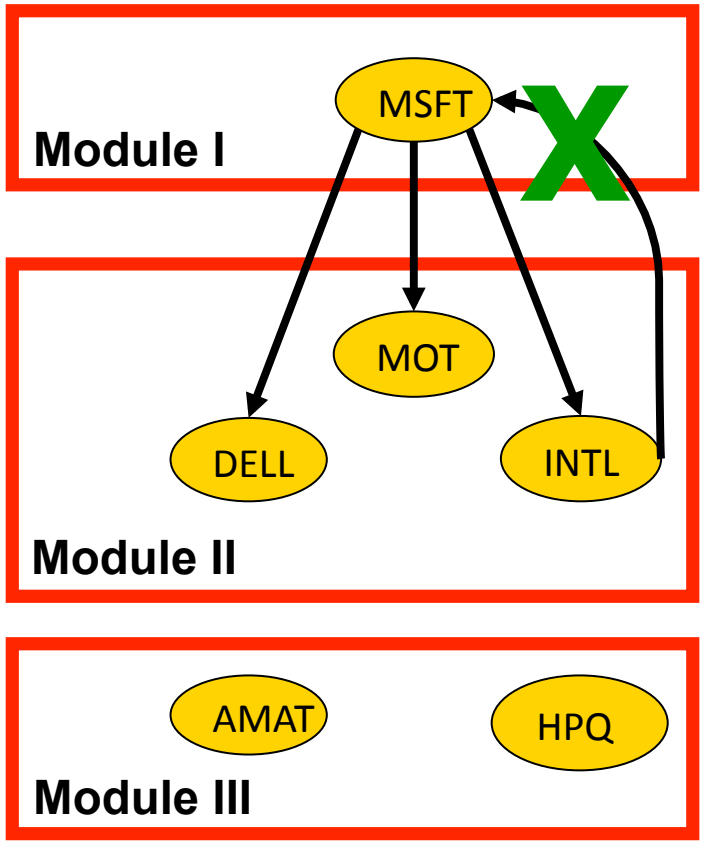


Slides from the presentation by Segal et al. UAI03

Learning Dependency Structure

- Heuristic search with operators
  - Add/delete parent for module
- Handle acyclicity
  - Can be checked efficiently on the module graph

X $\text{INTL} \rightarrow \text{Module}_I$



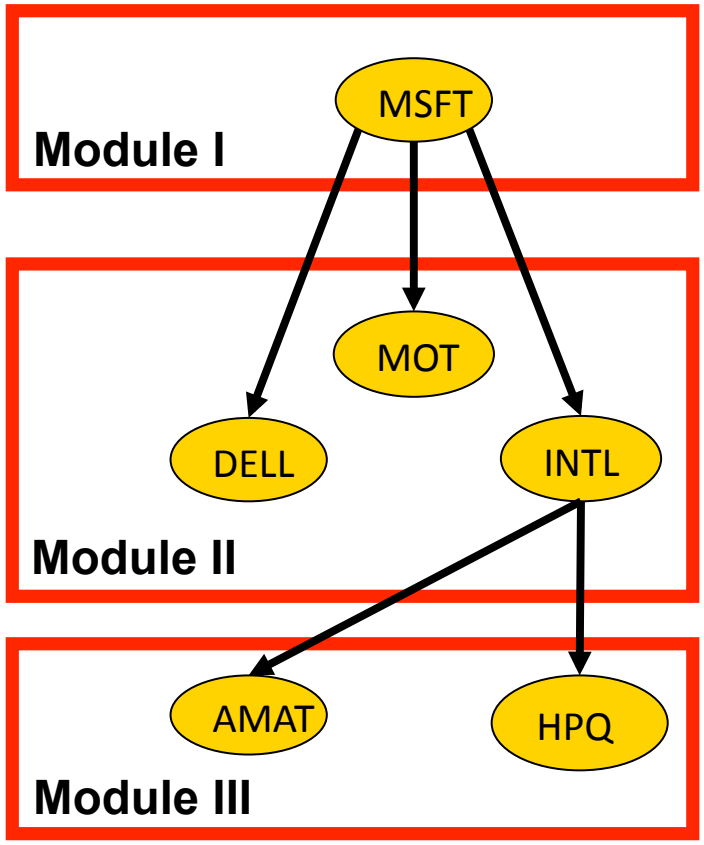
Slides from the presentation by Segal et al. UAI03

Learning Dependency Structure

- Heuristic search with operators
  - Add/delete parent for module
- Handle acyclicity
  - Can be checked efficiently on the module graph

✗ INTL → $Module_I$

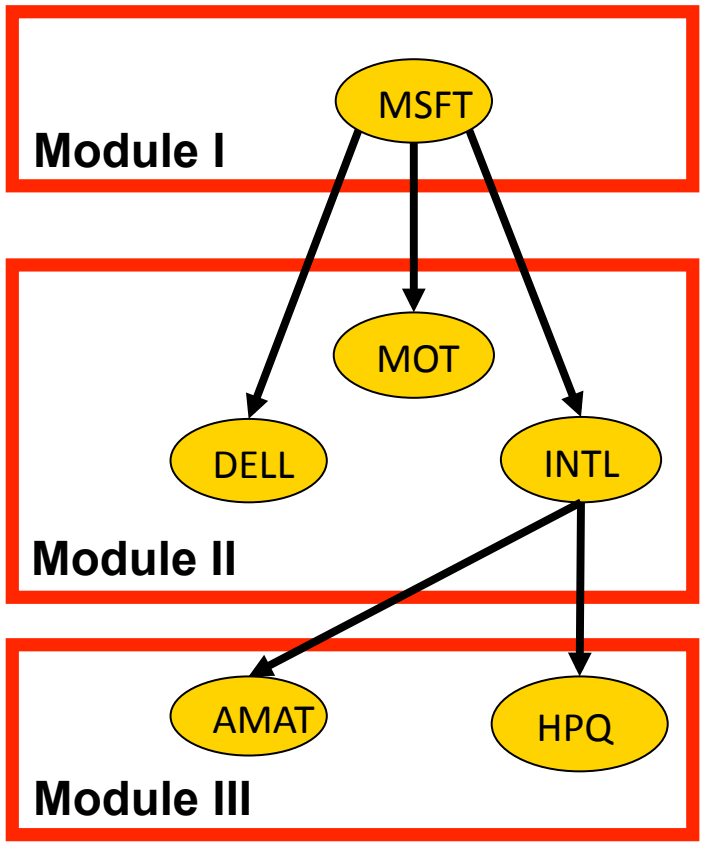
✓ INTL → $Module_{III}$



Slides from the presentation by Segal et al. UAI03

Learning Dependency Structure

- Heuristic search with operators
  - Add/delete parent for module
- Handle acyclicity
  - Can be checked efficiently on the module graph
- Efficient computation
  - After applying operator for module M_j, only update score of operators for module M_j

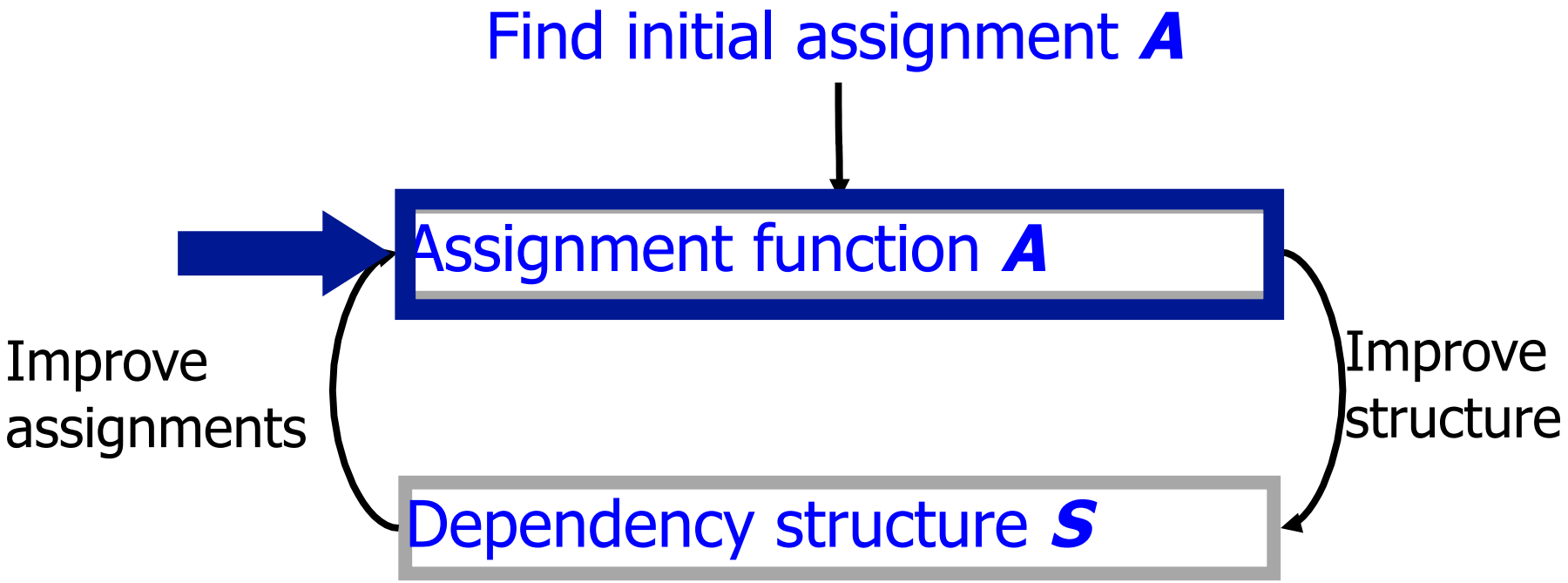


Slides from the presentation by Segal et al. UAI03

Algorithm Overview

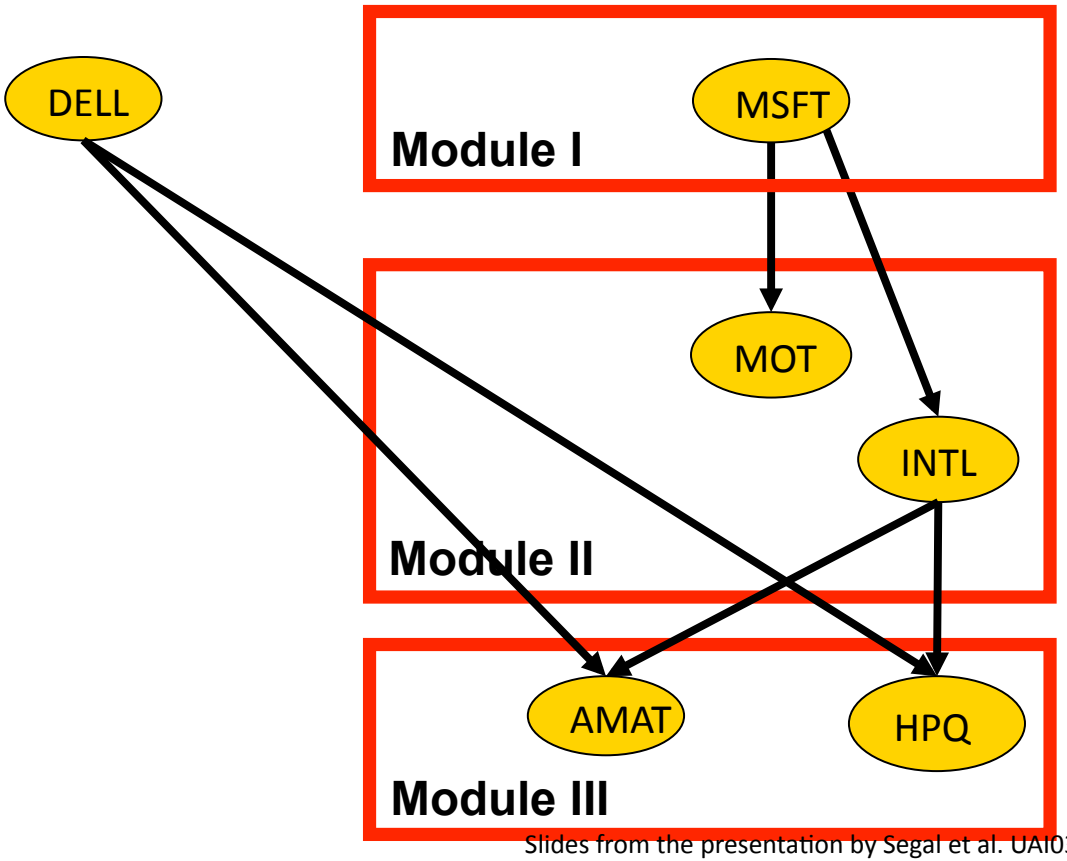
- Find assignment function ***A*** and structure ***S*** that maximize the Bayesian score

$$Score(S,A : D) = logP(D \mid S,A) + logP(S,A)$$



Slides from the presentation by Segal et al. UAI03

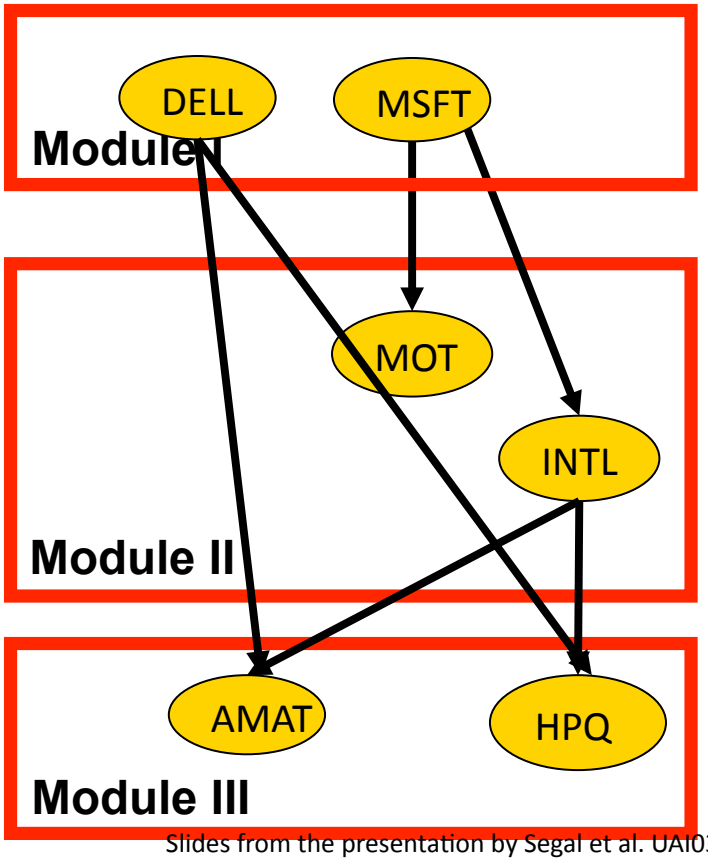
Learning Assignment Function



Slides from the presentation by Segal et al. UAI03

Learning Assignment Function

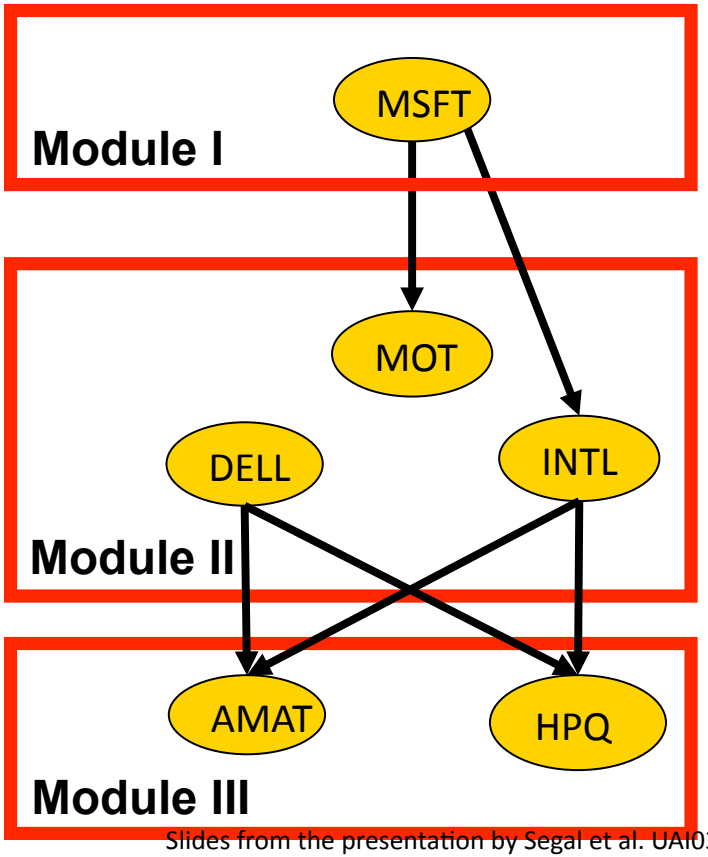
- A(DELL)=M_I
  - Score: 0.7



Slides from the presentation by Segal et al. UAI03

Learning Assignment Function

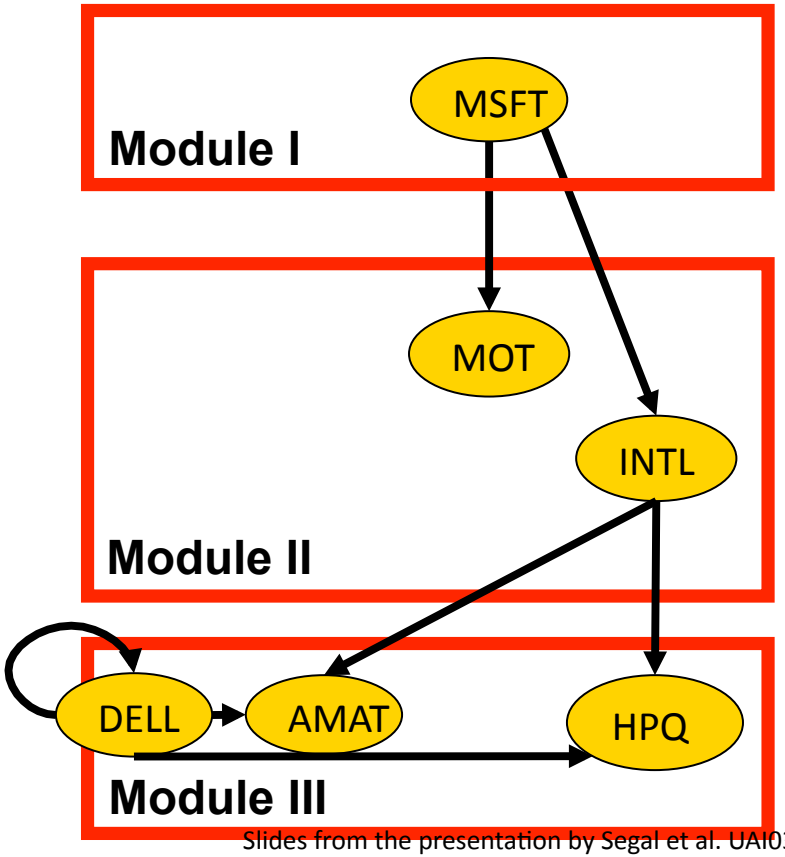
- A(DELL)=M_I
  - Score: 0.7
- A(DELL)=M_{II}
  - Score: 0.9



Slides from the presentation by Segal et al. UAI03

Learning Assignment Function

- A(DELL)=M_I
  - Score: 0.7
- A(DELL)=M_{II}
  - Score: 0.9
- A(DELL)=M_{III}
  - Score: cyclic!



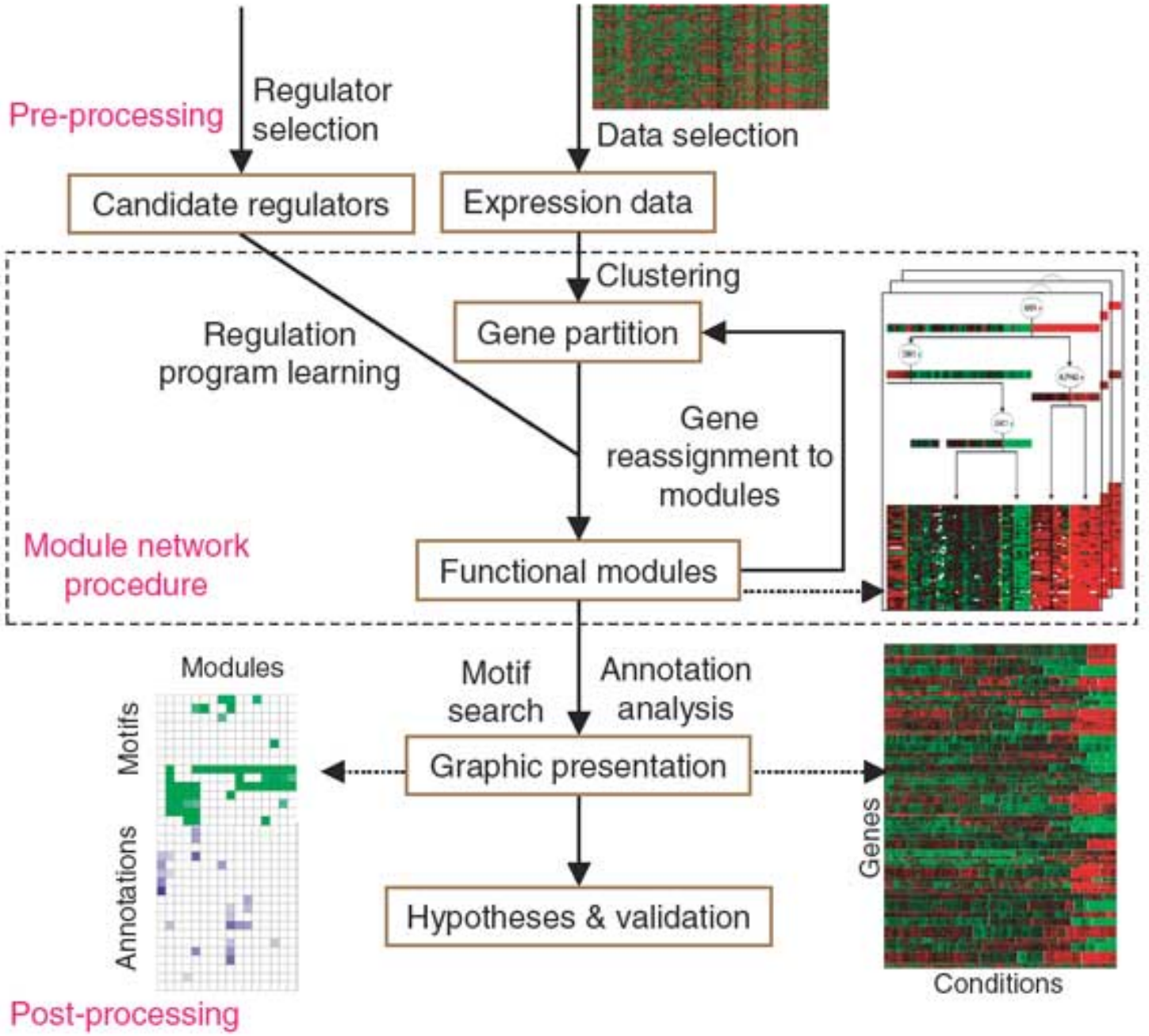
Slides from the presentation by Segal et al. UAI03

Learning Module Network

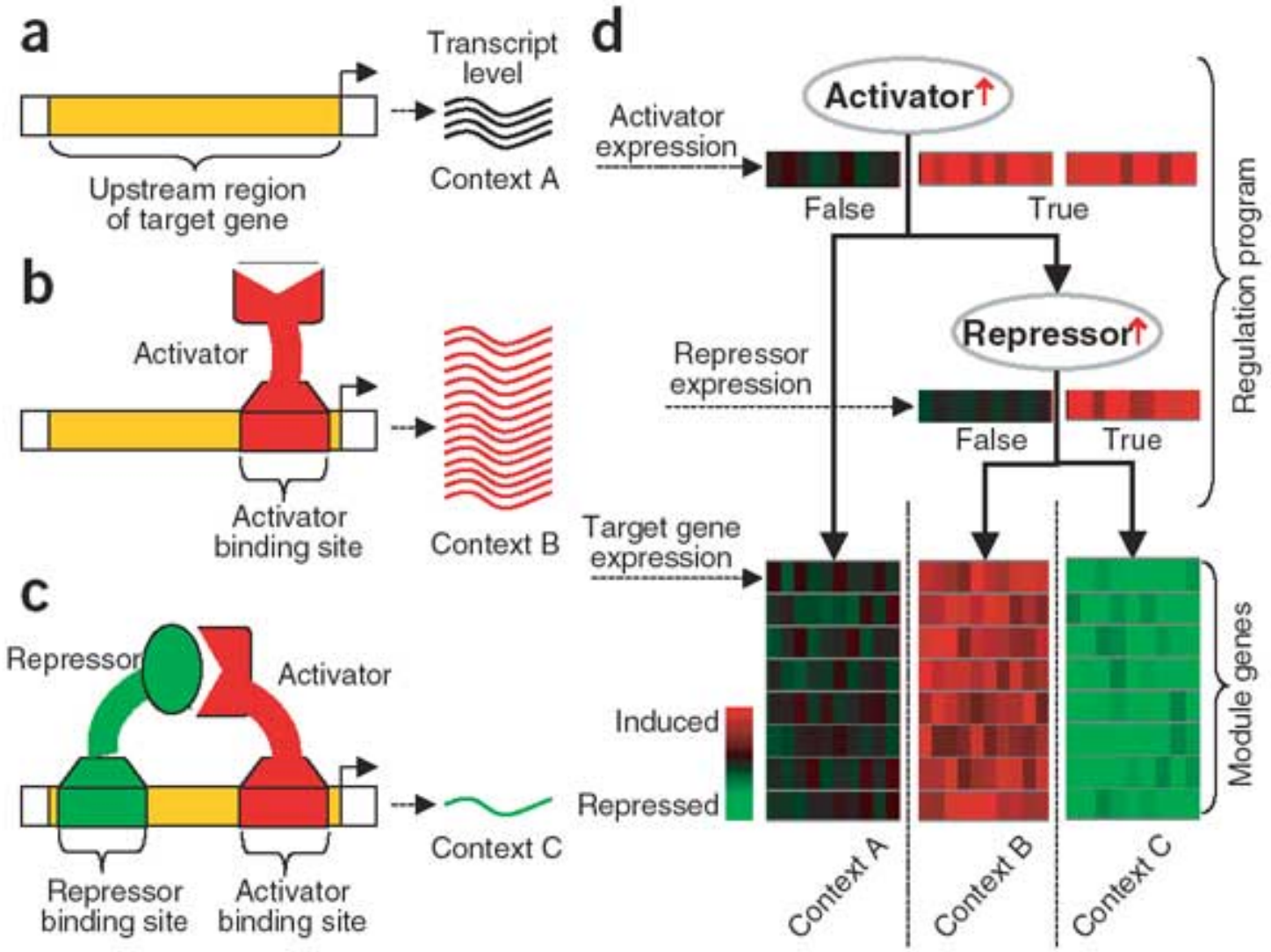
- Module Network
  - Model definition
  - Learning the model

- Experimental results

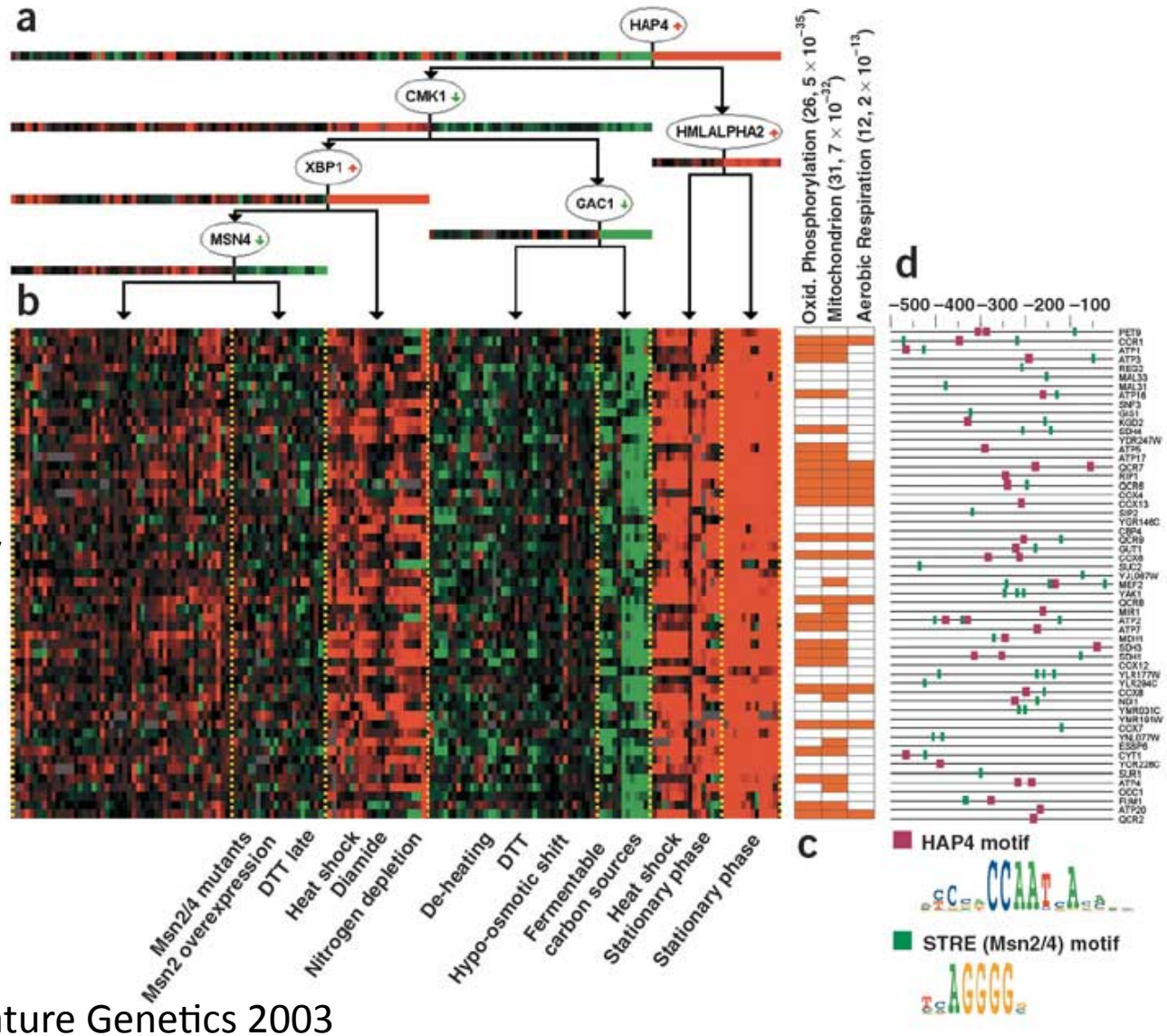
Learning Module Network from Gene Expression Data



Determine Combinatorial Control From Gene Expression data Under Multiple Conditions



Resulting Module

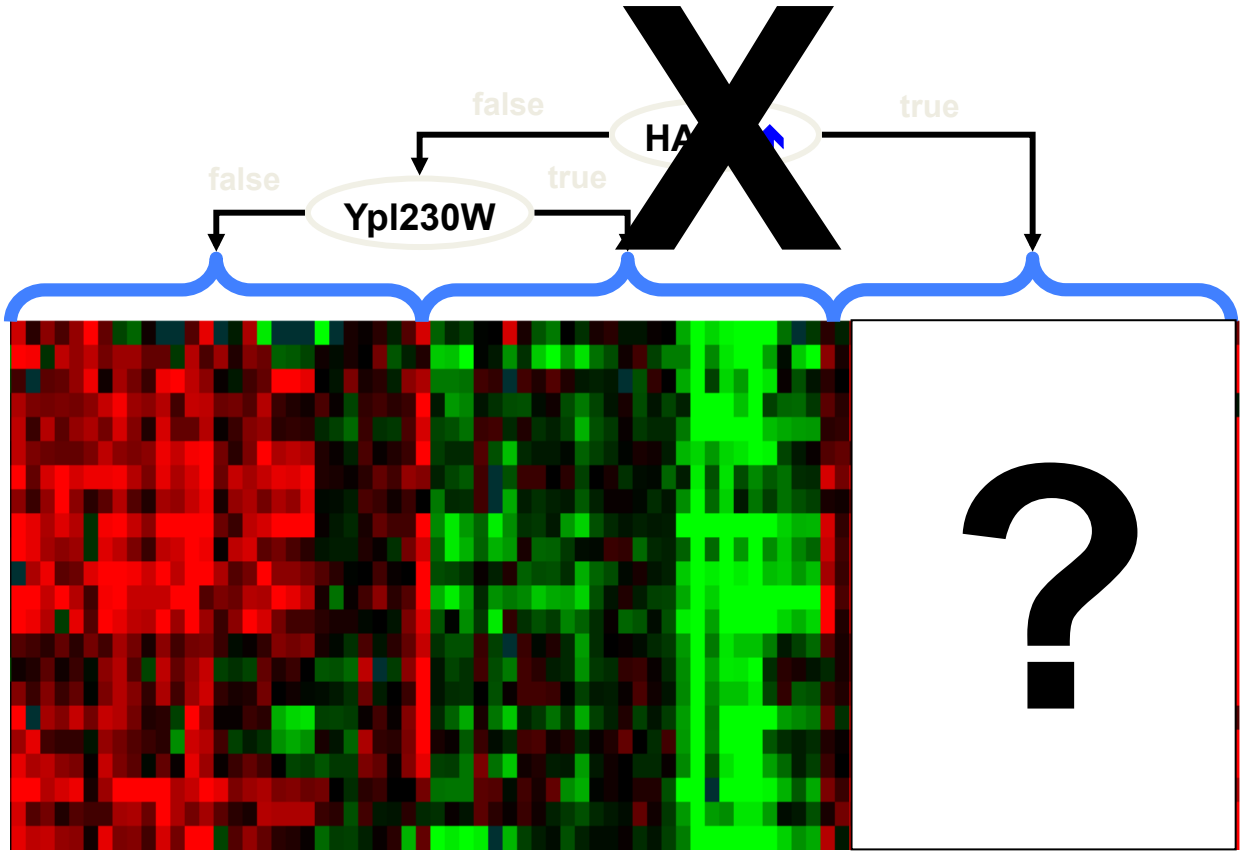


Row: genes
Columns: array (condition)

Segal et al Nature Genetics 2003

Experimental Design

- Hypothesis: Regulator ‘X’ **activates** process ‘Y’
- Experiment: Knock out ‘X’ and repeat experiment



Biological Experiments Validation

- Were the differentially expressed genes predicted as targets?
- Rank modules by enrichment for diff. expressed genes

△ Ypl230w

#	Module	Significance
39	Protein folding	7/23, 1e-4
29	Cell differentiation	6/41, 2e-2
5	Glycolysis and folding	5/37, 4e-2
34	Mitochondrial and protein fate	5/37, 4e-2

△ Ppt1

#	Module	Significance
14	Ribosomal and phosphate metabolism	8/32, 9e 3
11	Amino acid and purine metabolism	11/53, 1e 2
15	mRNA, rRNA and tRNA processing	9/43, 2e 2
39	Protein folding	6/23, 2e 2
30	Cell cycle	7/30, 2e 2

△ Kin82

#	Module	Significance
3	Energy and osmotic stress I	8/31, 1e 4
2	Energy, osmolarity & cAMP signaling	9/64, 6e 3
15	mRNA, rRNA and tRNA processing	6/43, 2e 2

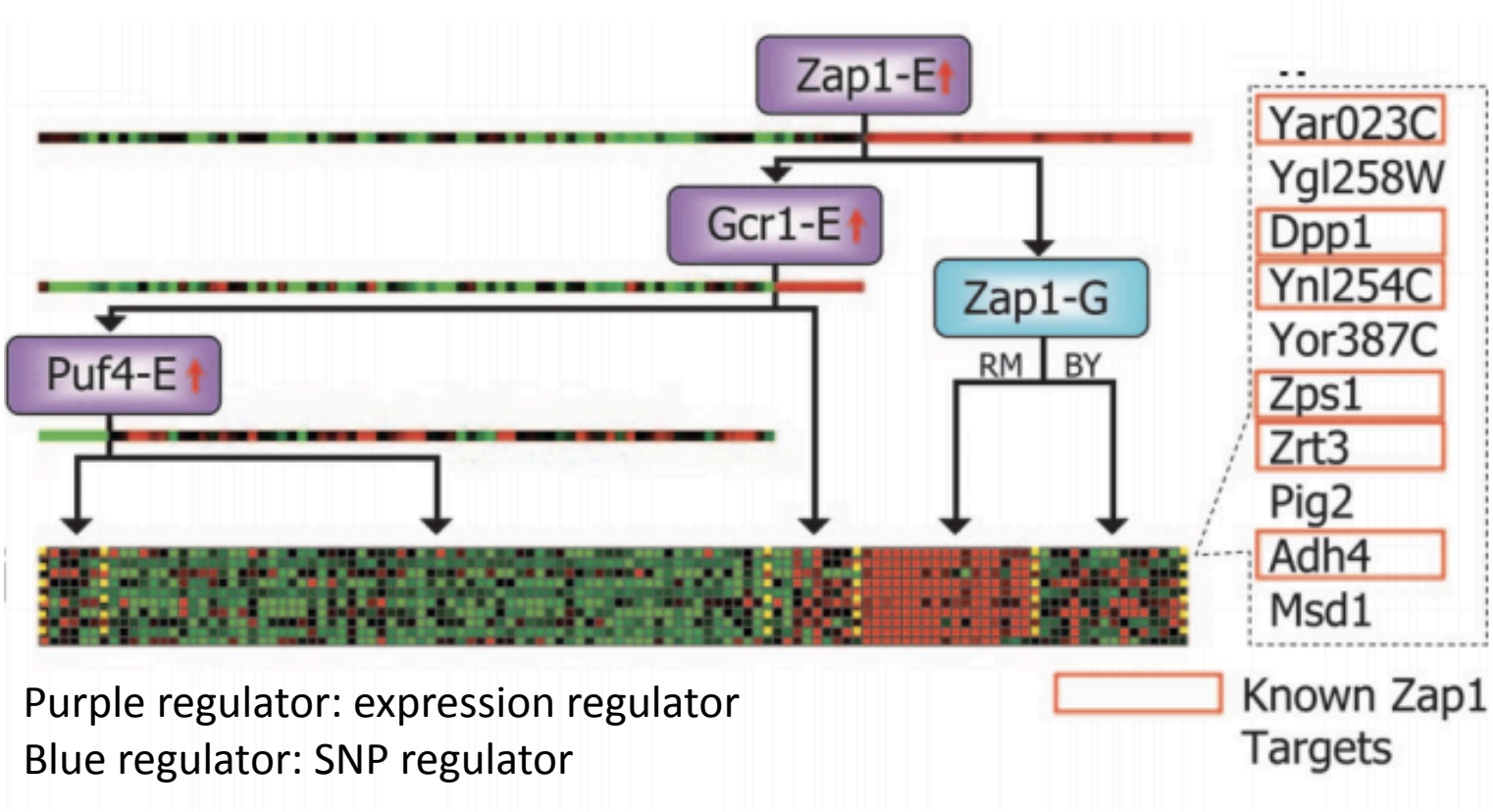
- All regulators regulate predicted modules

Segal et al., Nature Genetics, 2003

Extensions of Module Network: Geronemo

(Lee et al., PNAS 2006)

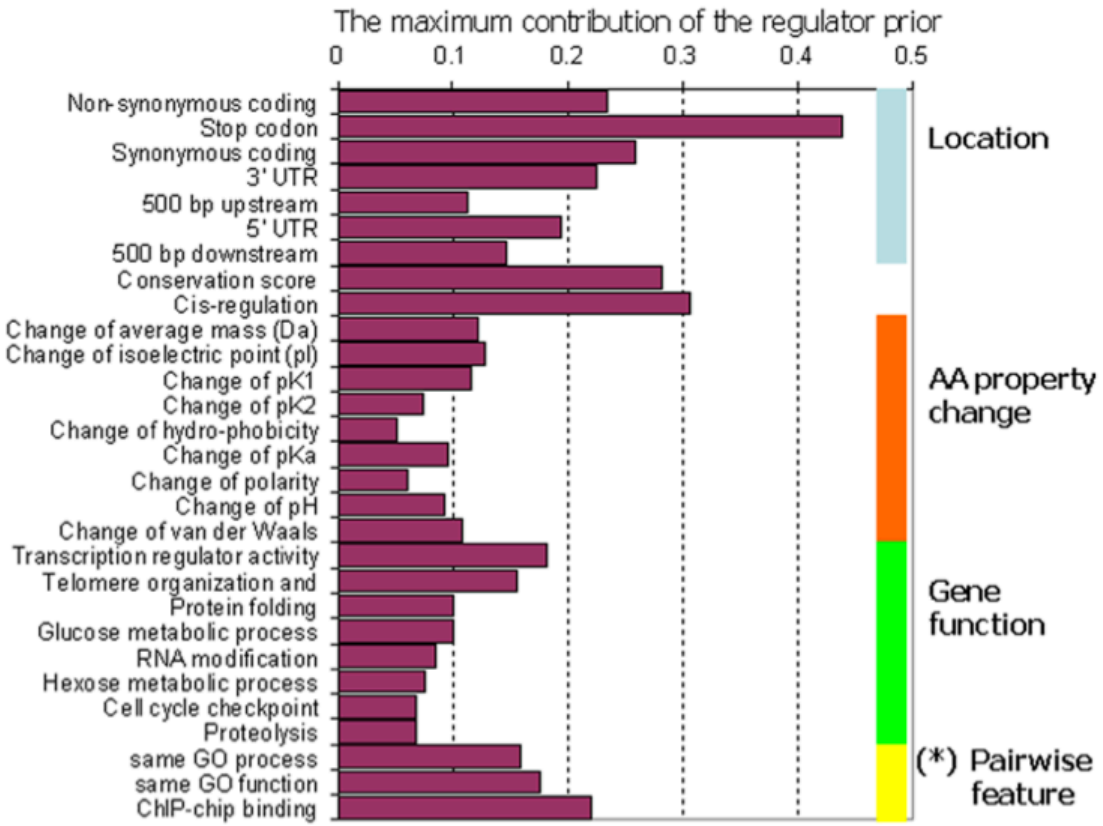
- Both gene expression levels and SNPs can be regulators for gene expression levels of other genes



Purple regulator: expression regulator
Blue regulator: SNP regulator

Extensions of Module Network: Lirnet

(Lee et al., PLoS Genetics 2009)

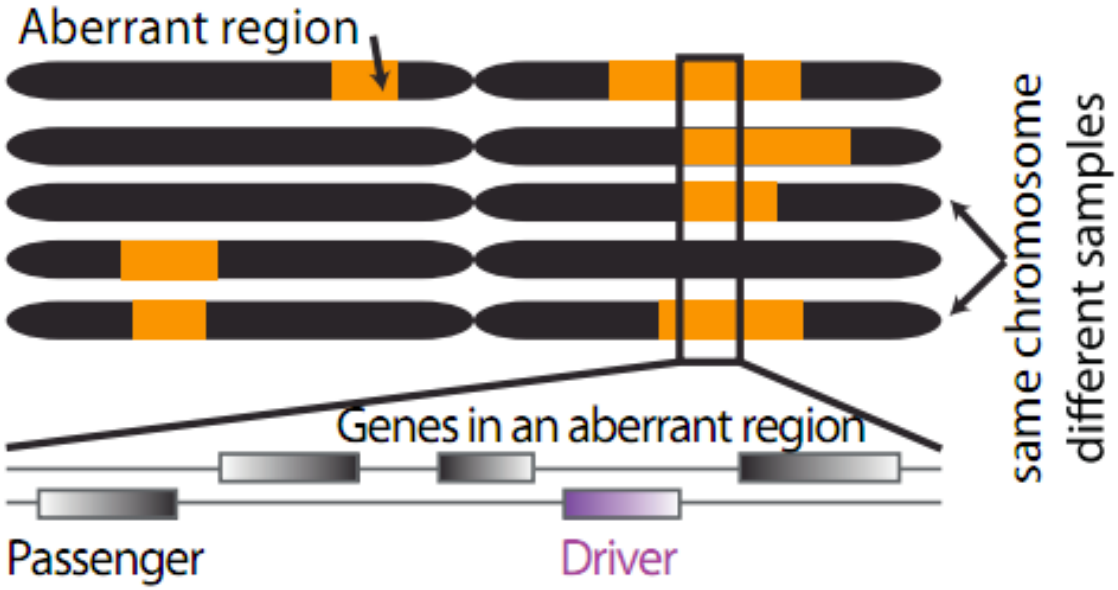


- Incorporate prior knowledge (regulatory features) on gene-expression regulators and SNP regulators
- Regulatory features (on the left) and learned regulatory priors (purple bar)

Extensions of Module Network: CONEXIC

(Akavia et al., Cell 2010)

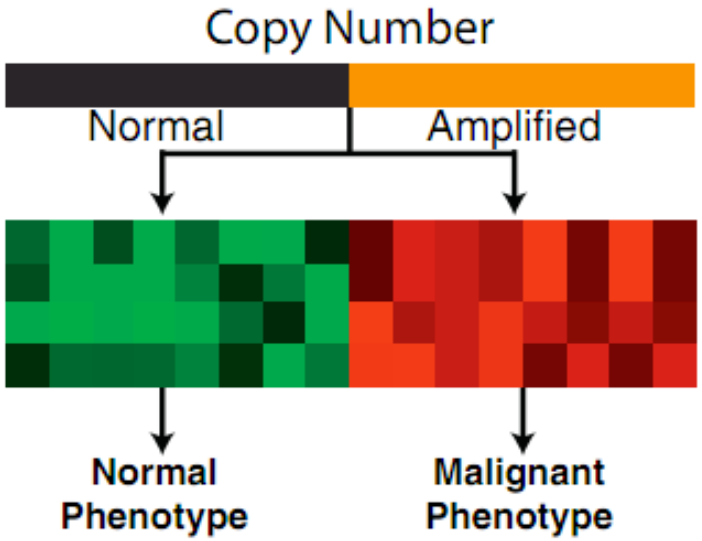
- Extends module networks to handle cancer copy number variation and gene expression data to find cancer-causing (driver) mutation
- A driver mutation should occur in multiple tumors more often than would be expected by chance



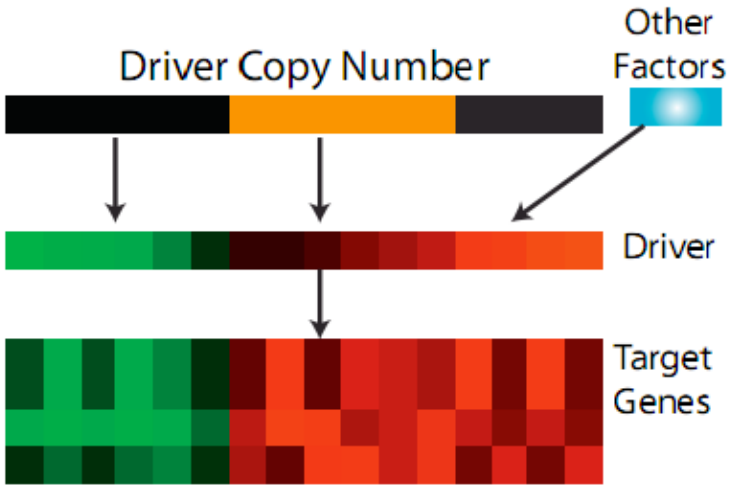
Extensions of Module Network: CONEXIC

(Akavia et al., Cell 2010)

- A driver mutation may be associated (correlated) with the expression of a group of genes that form a module



- A driver may be over-expressed due to amplification of the DNA encoding it or the action of other factors. The target genes correlate with driver gene expression rather than driver copy number



Overview

- Bayesian networks (network with directed edges): Module networks and their extensions
  - Module network (Segal et al., Nature Genetics 2003): Gene module's activity is determined by their expression levels of regulator genes
  - Geronemo (Lee et al., PNAS 2006): Gene module's activity is determined by their expression levels of regulator gene and SNPs
  - Lirnet (Lee et al., PLoS Genetics 2009): incorporates prior knowledge
  - CONEXIC (Akavia et al., Cell 2010): cancer data analysis for copy number variation and gene expression data

- Gaussian graphical models (network with undirected edges) and their extensions

Gaussian Graphical Models

- The gene expressions for K genes $Y=\{y_1, \ldots, y_K\}$ are Gaussian distributed:

$$Y \sim N(0_K, \Theta^{-1})$$

  - 0_K: vector of K zeros
  - Θ: K by K inverse covariance matrix

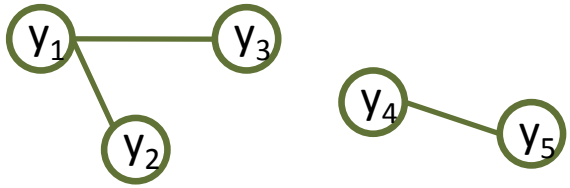
- Then, the inverse covariance matrix Θ encodes a Gaussian graphical model
  - Non-zero elements in Θ correspond to edges

Gaussian Graphical Models

- Non-zero elements in Θ correspond to edges

Gaussian graphical model encoded by **Θ**

y_1	1.3	0.3	0.8	0	0
y_2	0.3	1.0	0	0	0
y_3	0.8	0	1.2	0	0
y_4	0	0	0	1.5	1.1
y_5	0	0	0	1.1	0.9

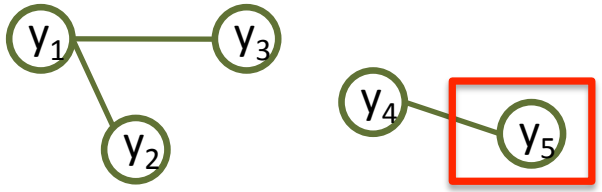


Gaussian Graphical Models

- Non-zero elements in Θ correspond to edges

Gaussian graphical model encoded by **Θ**

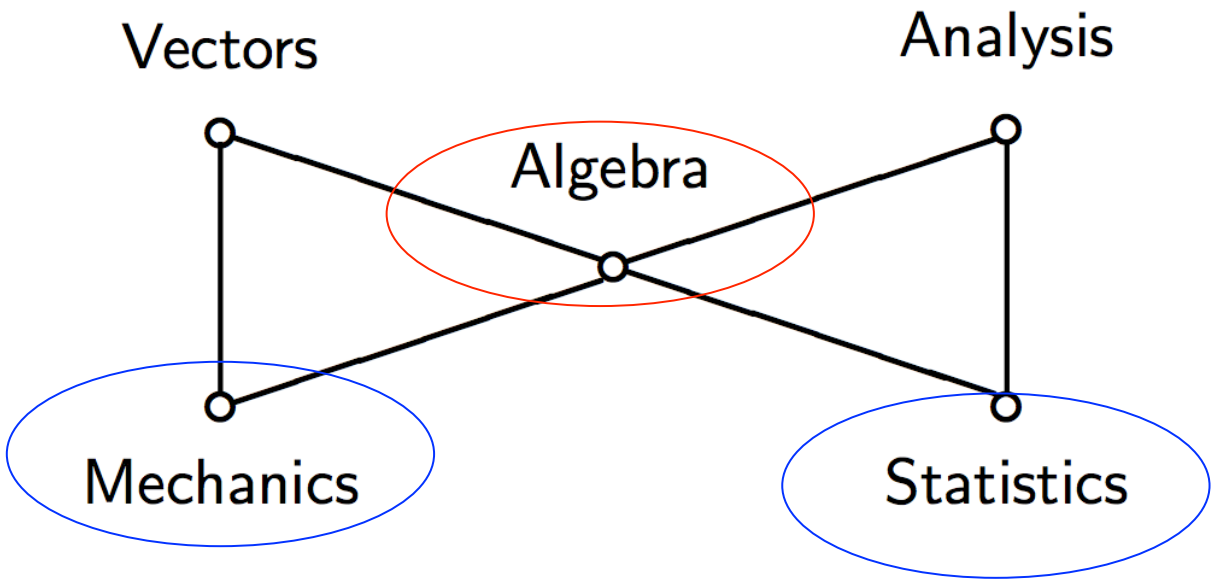
y_1	1.3	0.3	0.8	0	0
y_2	0.3	1.0	0	0	0
y_3	0.8	0	1.2	0	0
y_4	0	0	0	1.5	1.1
y_5	0	0	0	1.1	0.9



Nonzero/zero pattern of the y_5's column matches the neighbors of the node y_5

Probabilistic Graphical Models

- Statistics and Mechanics are independent of each other conditional on Algebra



Learning a Sparse Gaussian Graphical Models

- Minimize negative log likelihood of data with L_1 penalty

$$\arg\min \ \log\det\Theta - tr(S\Theta) - \lambda \|\Theta\|_1$$

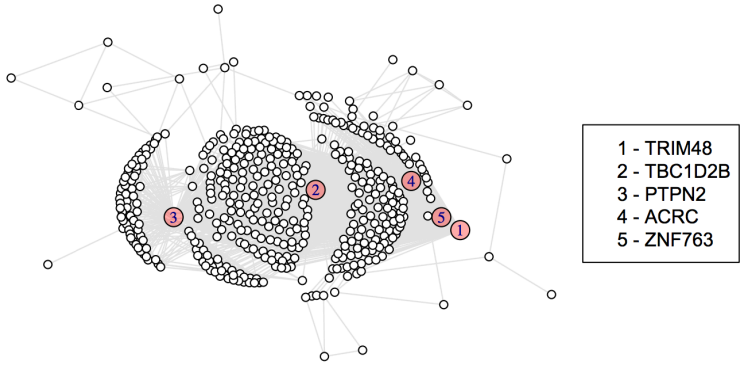
where

- tr(A) is the trace of matrix A
- S is a *K* by *K* sample covariance
- $||\Theta||_1$ is an L_1 regularization

- The optimization problem is **convex**!
  - Many software packages exist (e.g., BIG&QUIC, Hsieh et al., NIPS 2013; FastGGM, Wang et al., Plos Comp Bio 2016)

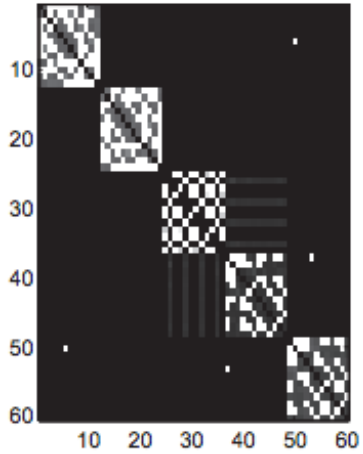
Extensions of Gaussian Graphical Models

- Network with hubs
  (Tan et al., JMLR 2014)



Learned from Glioblastoma expression data
(Hubs as pink nodes)

- Network with block structures (Tan et al., UAI 2009)



Summary

- Modeling gene networks with Bayesian networks
  - Probabilistic model for learning modules of variables and their structural dependencies
  - Module networks have improved performance over Bayesian networks
    - Statistical robustness
    - Interpretability
  - Reconstruction of many known regulatory modules and prediction of targets for unknown regulators

- Modeling gene networks with undirected networks
  - Gaussian graphical models are extremely popular: fast learning methods are available (more efficient than Bayesian network learning)