Advanced Topics in Learning the Transcriptional Regulatory Networks

Lectures 11 – Nov 2, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

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

1

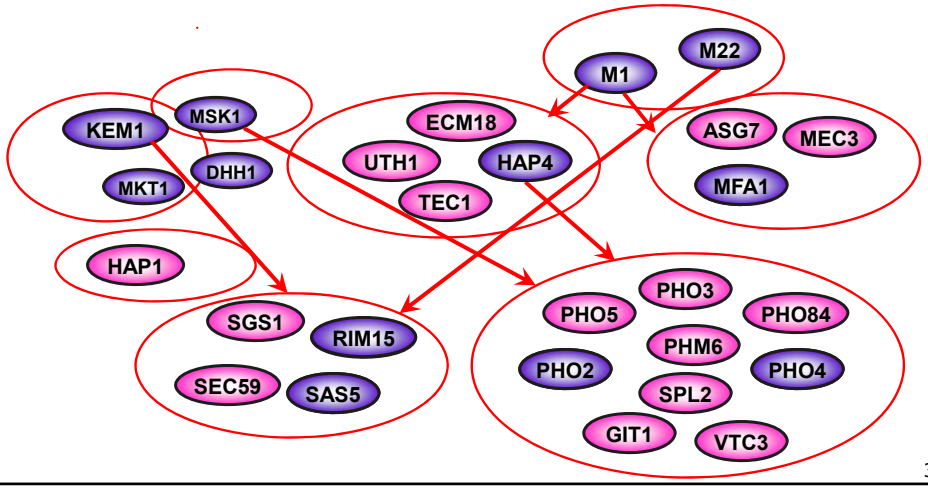
Outline

- Evaluation of the inferred network
  - Functional coherence of gene clusters
  - Predicted regulatory interactions
  - Multiple hypothesis testing
- Advanced topics
  - Structure learning via bootstrapping.
  - Inferring overlapping biological processes.
  - Incorporating prior knowledge.
- Systems genetics
  - Traditional approach
  - Systems biology approach

2

Review – Learning Regulatory Network

- What next?
  - Do gene clusters (modules) make sense?
  - Do predicted regulatory interactions (edges) make sense?



3

Functional Coherence of Gene Clusters

- Gene Ontology (GO) [http://www.geneontology.org/]
  - The GO database provides a controlled vocabulary to describe gene and gene product attribute in any organism.
  - Set of biological phrases (GO terms) which are applied to genes
  - Organized as three separate ontologies
    - Molecular functions
    - Biological processes
    - Cellular components
  - Each gene may
    - Have more than one in molecular function.
    - Take part in more than one biological process.
    - Act in more than one cellular component.

4

Structure of Ontologies

- Shows the relationship between different terms
  - One term may be a more specified description of another more general term.
  - Shows hierarchies of the terms (directed acyclic graph).
  - Each child-term is a member of its parent-term

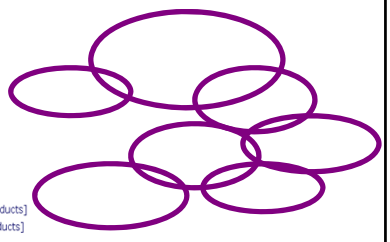
all : all [view gene products]
- GO:0008150 : biological_process [view gene products]
  - GO:0022610 : biological adhesion [view gene products]
  - GO:0065007 : biological regulation [view gene products]
  - GO:0009758 : carbohydrate utilization [view gene products]
  - GO:0015976 : carbon utilization [view gene products]
  - GO:0001906 : cell killing [view gene products]
  - GO:0008283 : cell proliferation [view gene products]
    - GO:0003263 : cardioblast proliferation [view gene products]
    - GO:0071838 : cell proliferation in bone marrow [view gene products]
    - GO:0003295 : cell proliferation involved in atrial ventricular junction remodeling [view gene products]
    - **GO:0035736 : cell proliferation involved in compound eye morphogenesis [view gene products]**
      - GO:2000496 : negative regulation of cell proliferation involved in compound eye morphogenesis [view gene products]
      - GO:2000497 : positive regulation of cell proliferation involved in compound eye morphogenesis [view gene products]
      - GO:2000495 : regulation of cell proliferation involved in compound eye morphogenesis [view gene products]

5

Create Functional Categories

- For each GO term,
  - Genes that have the same GO term form a functional category
- Other gene annotation systems
  - KEGG: Kyoto Encyclopedia of Genes and Genomes [http://www.genome.jp/kegg/]
  - Molecular Signature Database [http://www.broadinstitute.org/gsea/msigdb/index.jsp]

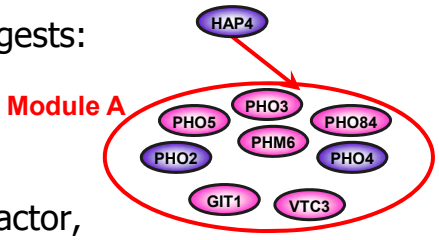
all : all [view gene products]
- GO:0008150 : biological_process [view gene products]
  - GO:0022610 : biological adhesion [view gene products]
  - GO:0065007 : biological regulation [view gene products]
  - GO:0009758 : carbohydrate utilization [view gene products]
  - GO:0015976 : carbon utilization [view gene products]
  - GO:0001906 : cell killing [view gene products]
  - GO:0008283 : cell proliferation [view gene products]
    - GO:0003263 : cardioblast proliferation [view gene products]
    - GO:0071838 : cell proliferation in bone marrow [view gene products]
    - GO:0003295 : cell proliferation involved in atrial ventricular junction remodeling [view gene products]
    - **GO:0035736 : cell proliferation involved in compound eye morphogenesis [view gene products]**
      - GO:2000496 : negative regulation of cell proliferation involved in compound eye morphogenesis [view gene products]
      - GO:2000497 : positive regulation of cell proliferation involved in compound eye morphogenesis [view gene products]
      - GO:2000495 : regulation of cell proliferation involved in compound eye morphogenesis [view gene products]



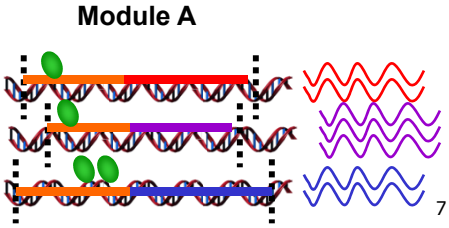
Functional categories

Predicted Regulatory Interaction I

- Say that your network suggests:



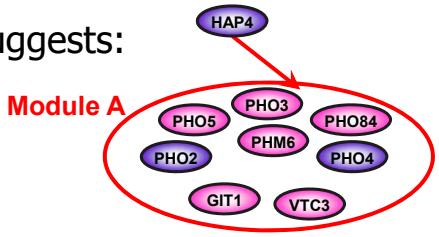
- If HAP4 is a transcription factor,
  - Targets should have a binding site for HAP4.
  - Or there should be different kind of evidence that HAP4 binds to genes in Module A (chip-chip or chip-seq data).



7

Predicted Regulatory Interaction II

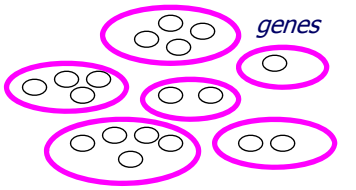
- Say that your network suggests:

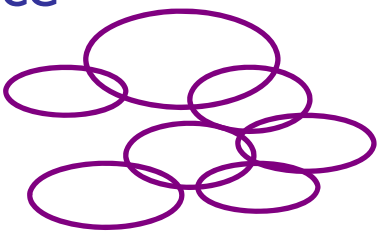


- If HAP4 really regulates module A, deletion (or overexpression) of HAP4 should lead to significant up/down- regulation of genes in module A.
  - There are many publicly available gene expression data that measure expression of genes after deleting/over-expressing a certain gene.

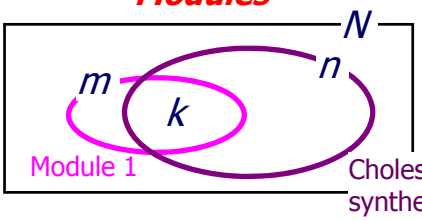
8

Functional Coherence





Modules **Known functional categories**



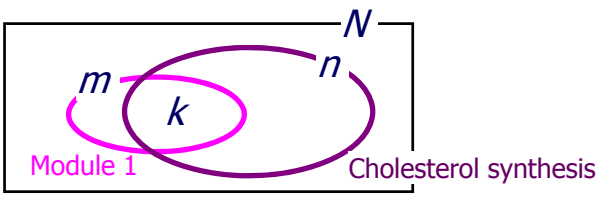
Gene ontology (GO)
http://www.geneontology.org/
Predicted targets of regulators
Sharing TF binding sites
:

- How significant is the overlap?
  - Calculate P(# overlap ≥ k | m, n, N; two groups are independent) based on the hypergeometric distribution

9

Examples

- Say N=1000, m=100, n=200 genes
  - If k = 40 genes in the intersection, p-value = 2.7410e-07.
  - If k = 30, p-value = 0.0039
  - If k = 20, p-value = 0.4394.

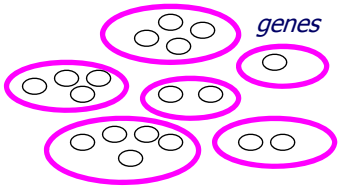


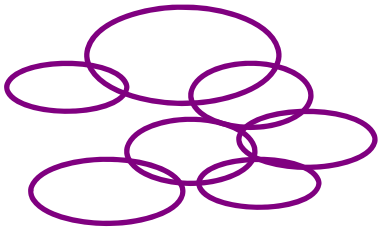
- How significant is the overlap?
  - Calculate p-value = P(# overlap ≥ k | m, n, N; two groups are independent), based on the hypergeometric distribution
  - What p-values are considered to be significant?

10

Multiple Hypothesis Testing

- Say that there are 200 modules and 3000 functional categories





Modules ***Known functional categories***

- How many hypotheses are we testing?
  - 200 x 3000 = 600,000
  - Is p-value of 0.001 significant? (p-value=0.001: frequency of observing the # genes in intersection by random.)
- P-values should be "corrected"
  - Bonferroni correction: min(1, p-value x # hypotheses)
  - FDR correction: control false discovery rate

11

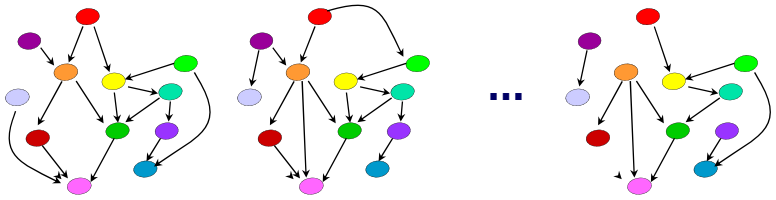
Outline

- Evaluation of the inferred network
  - Functional coherence of gene clusters
  - Predicted regulatory interactions
  - Multiple hypothesis testing
- Advanced topics
  - Structure learning via bootstrapping.
  - Inferring overlapping biological processes.
  - Incorporating prior knowledge.
- Systems genetics
  - Traditional approach
  - Systems biology approach

12

Structure Learning Via Boostrapping

- Many networks that achieve similar scores



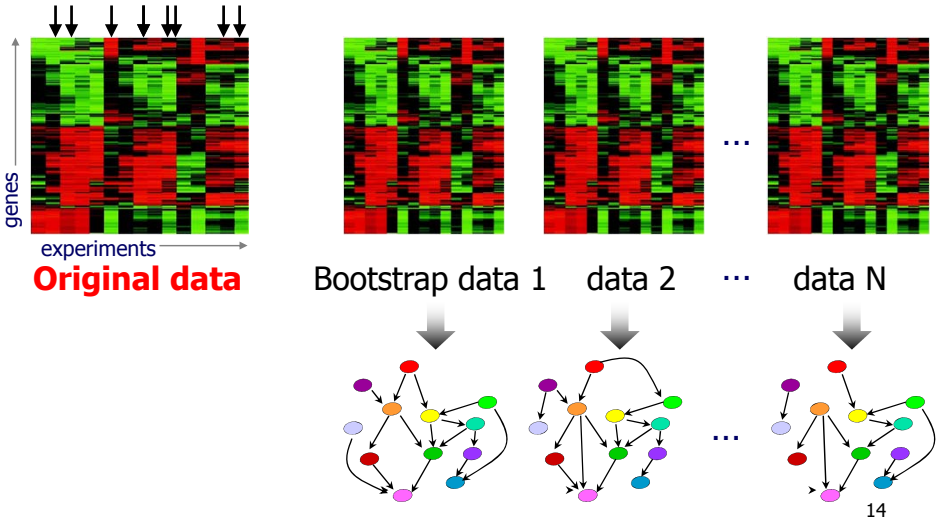
- Which one would you choose?
  - Estimate the robustness of each network or each edge.
  - How?? Learn the networks from **multiple datasets**.

13

Inferring sub-networks from perturbed expression profiles, Pe'er et al. Bioinformatics 2001

Bootstrapping

- Sampling with replacement



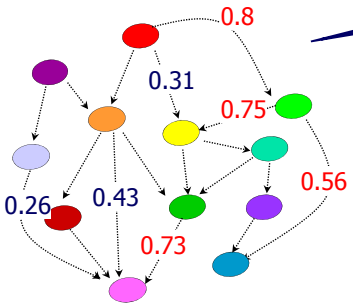
14

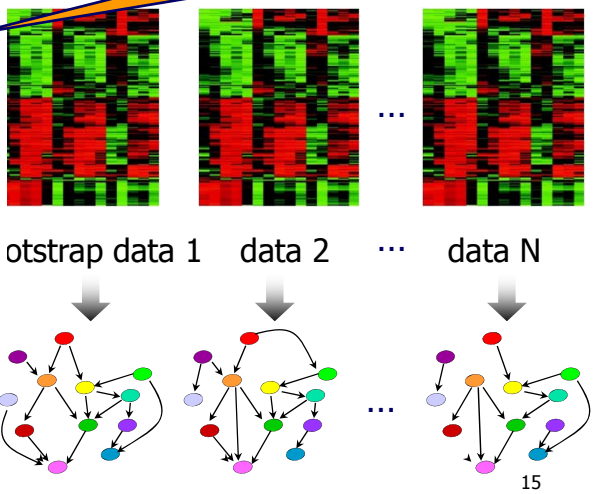
Inferring sub-networks from perturbed expression profiles, Pe'er et al. Bioinformatics 2001

Bootstrapping

- Sampling with replacement

- Estimated confidence of each edge *i*
  = # networks that contain the edge / total # networks (N)



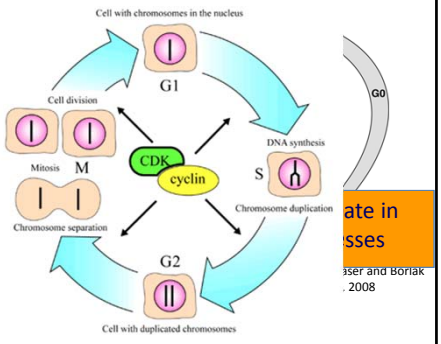


15

Inferring sub-networks from perturbed expression profiles, Pe'er et al. Bioinformatics 2001

Overlapping Processes

- The living cell is a complex system
  - Example, the cell cycle
    - Cell cycle: the series of events that take place in a cell leading to its division and duplication.
  - Genes functionally relevant to cell cycle regulation in the specific cell cycle phase



ate in
sses

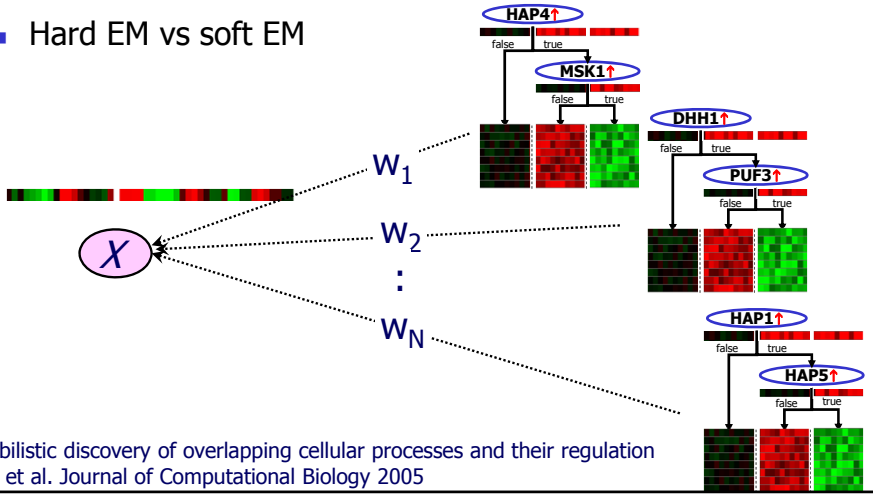
ıser and Borlak
. 2008

- Mutually exclusive clustering as a common approach to analyzing gene expression
  - (+) genes likely to share a common function
  - (-) group genes into mutually exclusive clusters
  - (-) no info about genes relation to one another

16

Decomposition of Processes…

- Model an expression level of a gene as **a mixture of regulatory modules**.
- Hard EM vs soft EM



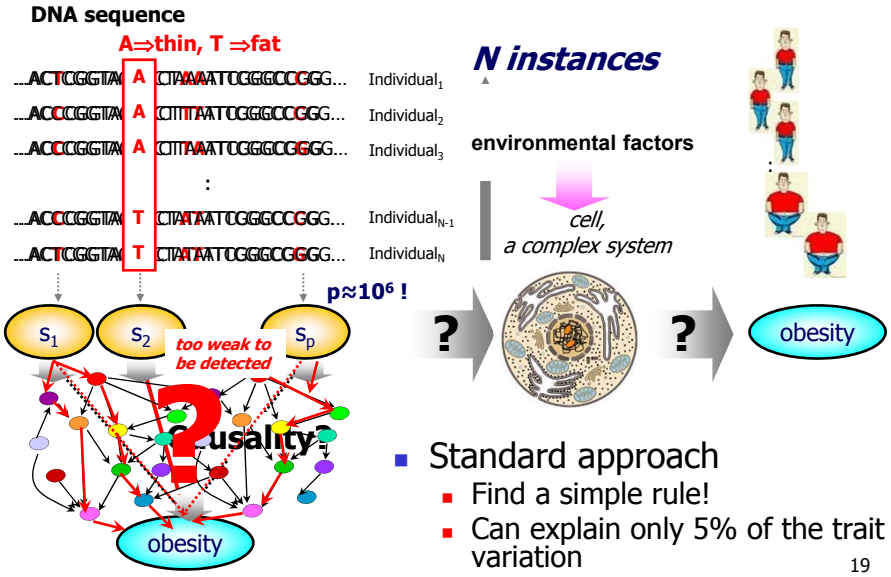
Probabilistic discovery of overlapping cellular processes and their regulation
Battle et al. Journal of Computational Biology 2005

Outline

- Evaluation of the inferred network
  - Functional coherence of gene clusters
  - Predicted regulatory interactions
  - Multiple hypothesis testing
- Advanced topics
  - Structure learning via bootstrapping.
  - Inferring overlapping biological processes.
  - Incorporating prior knowledge.
- Systems genetics
  - Traditional approach
  - Systems biology approach (example application, Lee et al. PLoS Genetics 2009)

18

Motivation: Genotype to Phenotype

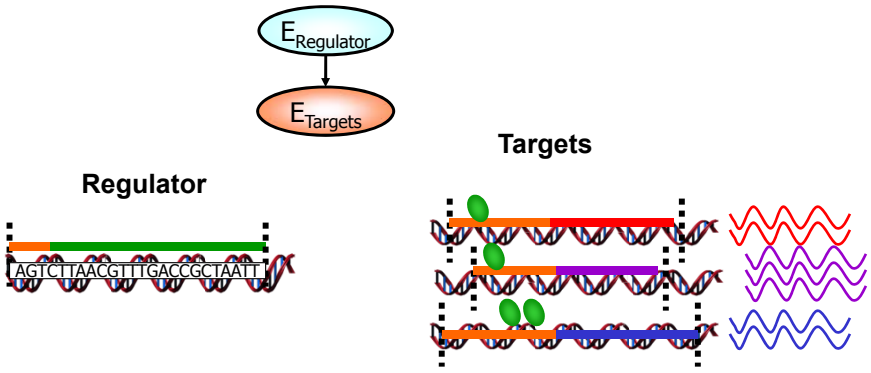


- Standard approach
  - Find a simple rule!
  - Can explain only 5% of the trait variation

19

Genetic Variation and Regulation

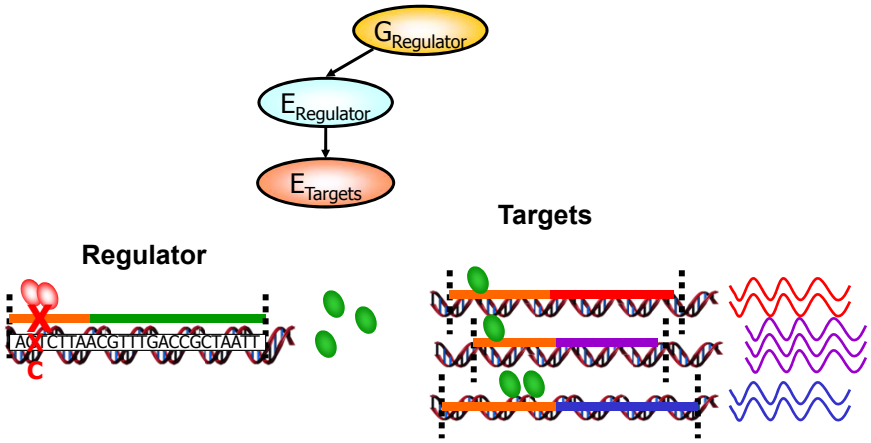
- Activity level of Regulator changes the expression levels of Targets it binds to.
- Regulator's expression is predictive of Targets' expression



Segal et al., Nature Genetics 2003; Lee et al., PNAS 2006

Regulation Variation & Mechanisms

- Regulator SNPs ⇒ change in binding site



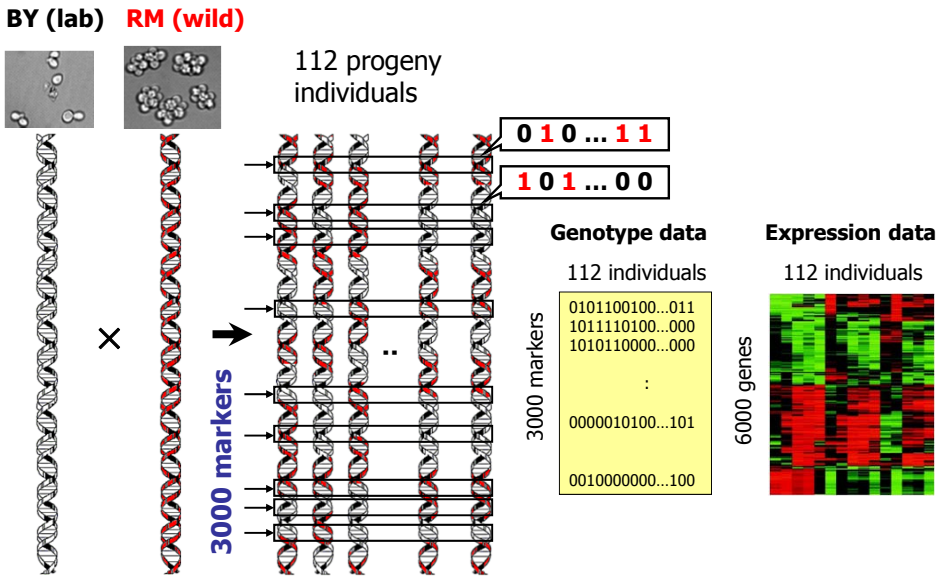
Regulation Variation & Mechanisms

- Regulator SNPs ⇒ change in regulator function

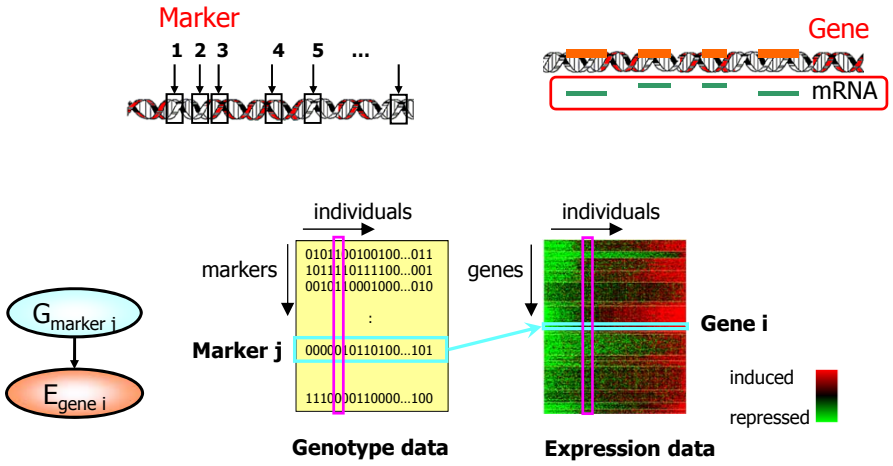
- Regulator's genotype is predictive of Targets' expression

22

eQTL Data [Brem et al. (2002) Science]



Traditional Approach: Single Marker eQTL mapping



Goals

"Genotype data" – binary values

...ACTCGGTAGACCTAAATTCGGCCCGG...
...ACCCGGTAGACCTTTATTCGGCCCGG...
...ACCCGGTAGACCTTAATTCGGCCGGG...
:
...ACCCGGTAGTCCTATATTCGGCCCGG...
...ACTCGGTAGTCCTATATTCGGCCGGG...

s_1 s_2 ... s_P

$p \approx 10^6$!

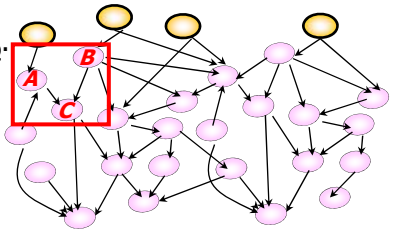
"Expression data" – measurement of mRNA levels of all genes

$Individual_1$
$Individual_2$
$Individual_3$
:
$Individual_{N-1}$
$Individual_N$

e_1 e_6 ... e_Q

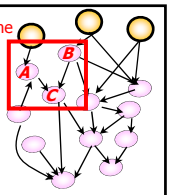
$Q \approx 2x10^4$

- Causality relationships among s_{1-P} and e_{1-Q}:
  - Gene regulatory network
    A and ***B*** regulate the expression of ***C***
    (A and B are regulators of C)
- Construction of the network
  - Multiple regression problems

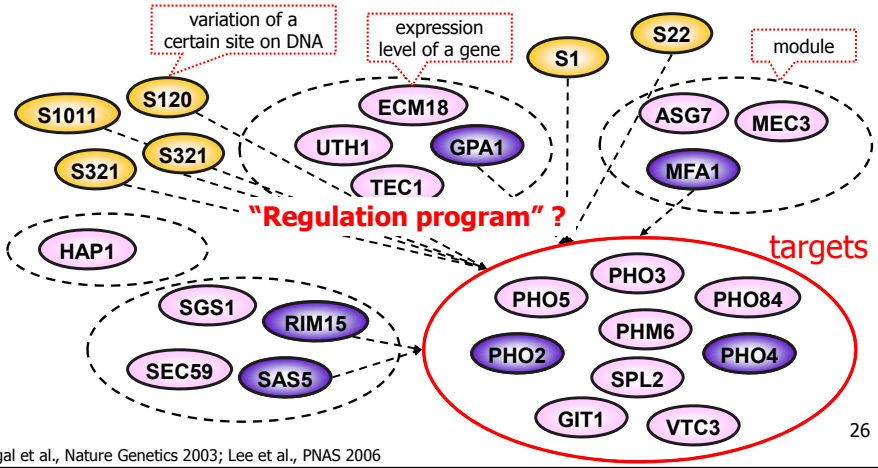


Regulatory Network

A and ***B*** regulate the expression of ***C*** **(A and B are regulators of C)**



- Candidate regulators (x_1,..,x_N):
  - Sequence variations
  - Expression levels of genes that have regulatory roles

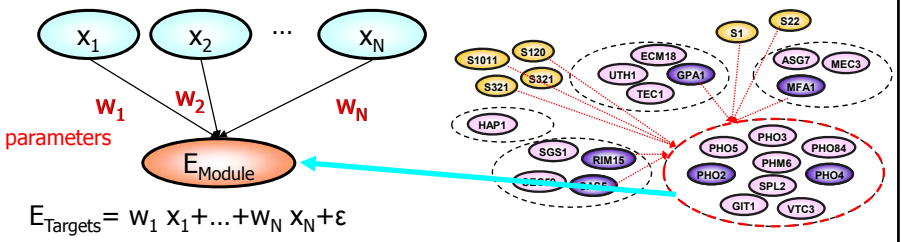


Segal et al., Nature Genetics 2003; Lee et al., PNAS 2006

26

Regulation as Linear Regression

$$\text{minimize}_{\mathbf{w}} (w_1x_1 + \ldots w_Nx_N - E_{Targets})^2$$



$E_{Targets} = w_1 x_1 + \ldots + w_N x_N + \varepsilon$

- But we often have very large *N*
- … and linear regression gives them all nonzero weight!

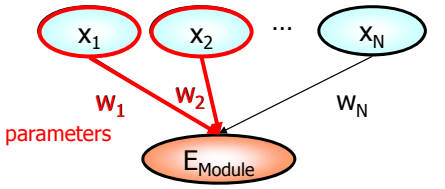
Problem: This objective learns too many regulators

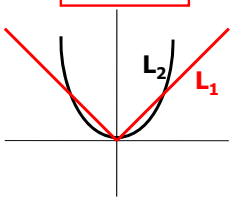
27

Lasso* (L_1) Regression

L_1 term

$$\text{minimize}_{\mathbf{w}} (w_1x_1 + \ldots w_Nx_N - E_{Module})^2 + \Sigma\, \mathbf{C}\, |w_i|$$



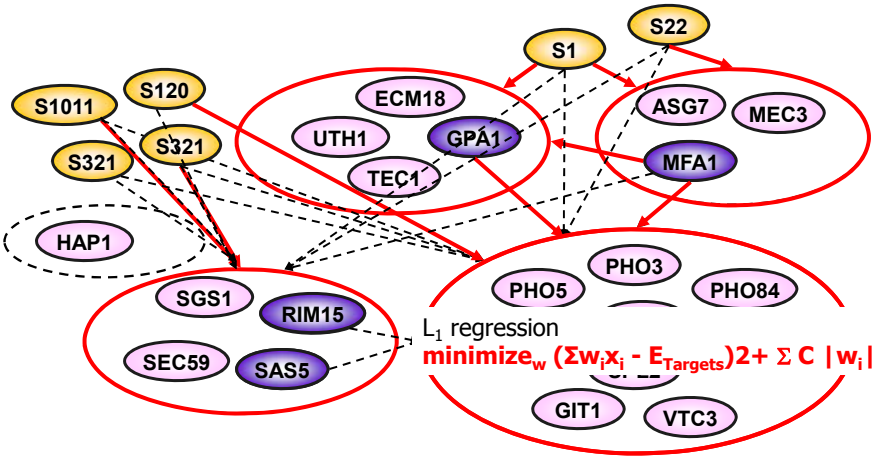


- Induces sparsity in the solution **w** (many w_i's set to zero)
  - Provably selects "right" features when many features are irrelevant
- Convex optimization problem
  - No combinatorial search
  - Unique global optimum
  - Efficient optimization

* Tibshirani, 1996

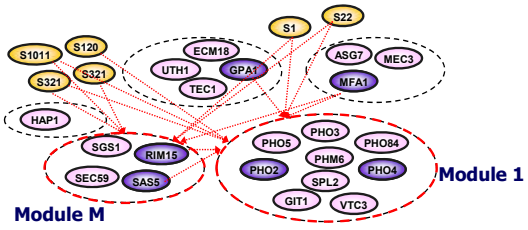
Learning Regulatory Network

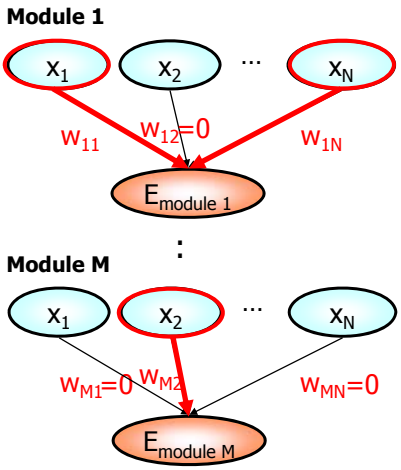
- Cluster genes into modules
- Learn a regulatory program for each module



Lee et al., PLoS Genet 2009

Learning the Regulatory Network





- Multiple regression tasks

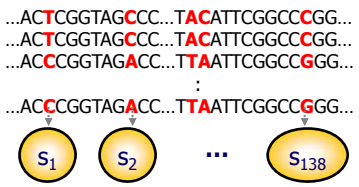
  minimize$_{w1}$ (Σ $w_{1n}x_n - E_{module1}$)^2+ Σ C|w_{1n}|

  :

  minimize$_{wn}$ (Σ $w_{Mn}x_n - E_{moduleM}$)^2+ Σ C|w_{Mn}|

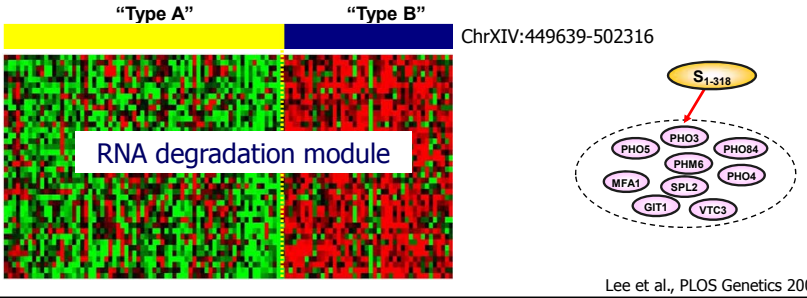
- Challenges?
  - Too large N!
    - # regulatory genes + # sequence variations
    - For human: 2000+1,000,000
  - Redundant features
    - $\{x_i, x_j, \ldots, x_k\}$ are perfectly correlated

Challenge: Redundant Features!



All individuals have either
TC...ACC (Type A) or
CA...TAG (Type B) for S_1~S_{318}

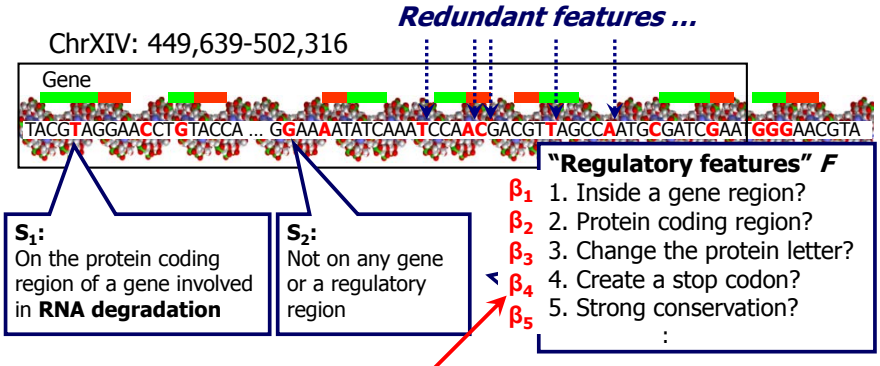
- Selected 318 sequence variations perfectly correlated
- Which of 318 is real causative variation?
- **Experiments for all 318 variations not feasible!**



Lee et al., PLOS Genetics 2009

Motivation

- Not all sequence variations are equally likely to be causal.



- Idea: Prioritize SNPs that have "good" regulatory features
- Problem: How much weight do we give to different regulatory features
  - Too many weights to estimate using cross-validation

Metaprior Model [Lee et al. ICML 2007]

- Multiple regression tasks

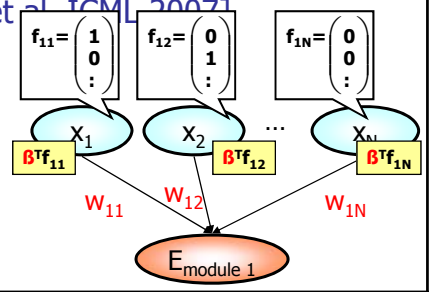
$\text{minimize}_{\mathbf{w1}} (\Sigma\, w_{1n}x_n - E_{module1})^2 + \Sigma\, \mathbf{C_{1n}}|\mathbf{w_{1n}}|$

:

$\text{minimize}_{\mathbf{wn}} (\Sigma\, w_{Mn}x_n - E_{moduleM})^2 + \Sigma\, \mathbf{C_{Mn}}|\mathbf{w_{Mn}}|$

Regulatory features
Protein coding region? (1=YES, 0=NO)
Known to be related to the module?
:

How about $C_{mn}=g(1/\beta^T f_{mn})$?



"Regulatory potential" (relevance score)=
β_1 x Protein coding region? +
β_2 x Strong conservation? + ...

- Learning
  - Learn **w**'s: for given ß,

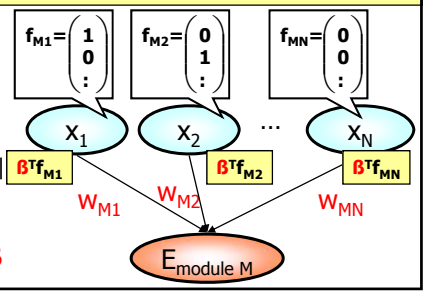
$\text{minimize}_{\mathbf{w1}} (\Sigma\, w_{1n}x_n - E_{module1})^2 + \Sigma\, g(1/\beta^T f_{1n})\,|w_{1n}|$

:

$\text{minimize}_{\mathbf{wn}} (\Sigma\, w_{Mn}x_n - E_{moduleM})^2 + \Sigma\, g(1/\beta^T f_{Mn})\,|w_{Mn}|$

  - Learn **ß** : for given **w**'s,

$\text{minimize}_{\boldsymbol{\beta}} \Sigma_m \Sigma_n\, g(1/\boldsymbol{\beta}^T f_{mn})\,|w_{mn}| + D\,\boldsymbol{\beta}^T\boldsymbol{\beta}$



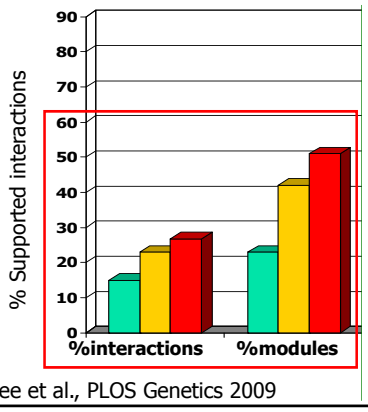
Transfer Learning

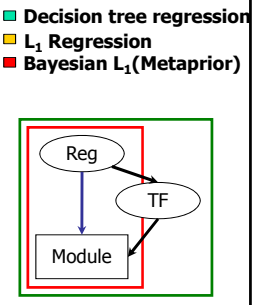
- What do regulatory potentials $\boldsymbol{\beta^T f_{mn}}$ do?
  - They do **not** change selection of "strong" regulators – those where prediction of targets is clear
  - They only help disambiguate between weak ones
- Strong regulators help teach us what to look for in other regulators

Transfer of knowledge
between different regression tasks

Biological Evaluation I

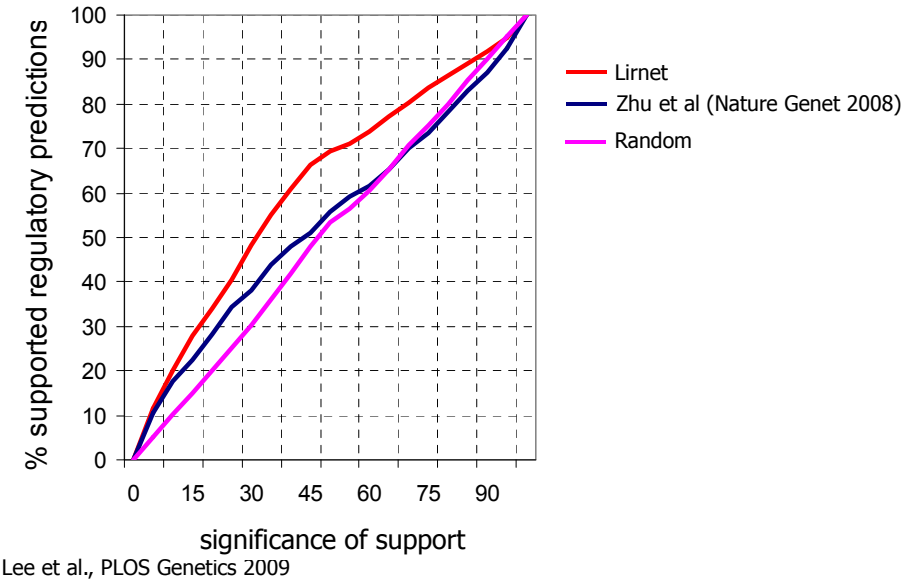
- How many predicted interactions have support in **other** data?
  - Deletion/ over-expression microarrays [Hughes et al. 2000; Chua et al. 2006]
  - ChIP-chip binding experiments [Harbison et al. 2004]
  - Transcription factor binding sites [Maclsaac et al. 2006]
  - mRNA binding pull-down experiments [Gerber et al. 2004]
  - Literature-curated signaling interactions





Lee et al., PLOS Genetics 2009

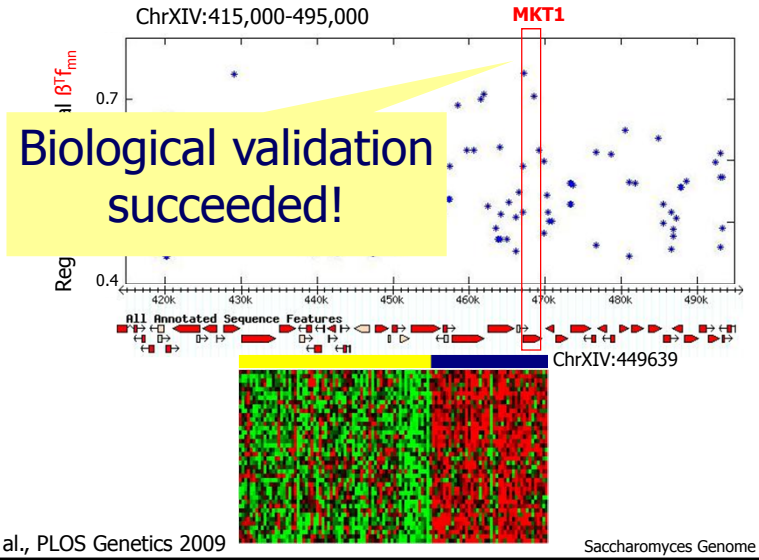
Biological Evaluation II



Lee et al., PLOS Genetics 2009

What Regulates the RNA degradation module?

- The regulatory potential over all 318 variations in the region



Biological validation succeeded!

Lee et al., PLOS Genetics 2009

Saccharomyces Genome Database (SGD)

Summary

- Motivation
  - Why are we interested in inferring the regulatory network?
- Algorithms for learning regulatory networks
  - Tree-CPDs with Bayesian score
  - Linear Gaussian CPDs with regularization
- Evaluation of the method
  - Statistical evaluation
  - Biological interpretation
- Advanced topics
  - Structure learning via bootstrapping.
  - Inferring overlapping biological processes.
  - Incorporating prior knowledge.
- Systems genetics
  - Traditional approach
  - Systems biology approach

38