

Part IV

Large network, less details

Infer network from large scale
data



2019, Week 9

How to study a large transcriptional network: many experiments condition, measurements, inferring and experimental validations

OPEN  ACCESS Freely available online

PLOS BIOLOGY

Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation from a Compendium of Expression Profiles

Jeremiah J. Faith¹, Boris Hayete¹, Joshua T. Thaden^{2,3}, Ilaria Mogno^{2,4}, Jamey Wierzbowski^{2,5}, Guillaume Cottarel^{2,5}, Simon Kasif^{1,2}, James J. Collins^{1,2}, Timothy S. Gardner^{1,2*}

1 Bioinformatics Program, Boston University, Boston, Massachusetts, United States of America, **2** Department of Biomedical Engineering, Boston University, Boston, Massachusetts, United States of America, **3** Boston University School of Medicine, Boston, Massachusetts, United States of America, **4** Department of Computer and Systems Science A. Ruberti, University of Rome, La Sapienza, Rome, Italy, **5** Cellicon Biotechnologies, Boston, Massachusetts, United States of America

The task

- Parts lists have benefited from high throughput sequencing
- Further progress requires development and refinement of techniques to determine the dynamic interactions among an organism's parts
- A subset of such interactions is the gene regulatory networks involving transcriptional factors

The challenge

- The total space of possible transcriptional regulatory interactions: **in billions**
 $N_{\text{TFs}} \times N_{\text{Genes}} \times N_{\text{environment context}}$
- Methodology requirement
Finding **thousands** of true regulatory interactions

Current states

- Combination of a large set of expression arrays, ChiP-chip (ChiP-seq etc) and other high throughput methods
- Machine-learning algorithms to identify cis-regulatory motifs and transcriptional factor targets
- Experimental validation of the precision of the methods at the genome scale remains elusive due to the lack of a model organism with a known regulatory structure and compatible experimental data.

Achieved in this study

- Generated and combined large set of affymetrix array data for E coli (445) and 3216 known regulatory interactions (regulonDB)
- Developed a novel and improved context likelihood of relatedness (CLR) algorithm.
- It confirmed known interactions and predicted new interactions
- Experimental confirmed its novel predictions

The schematic overview of the procedure

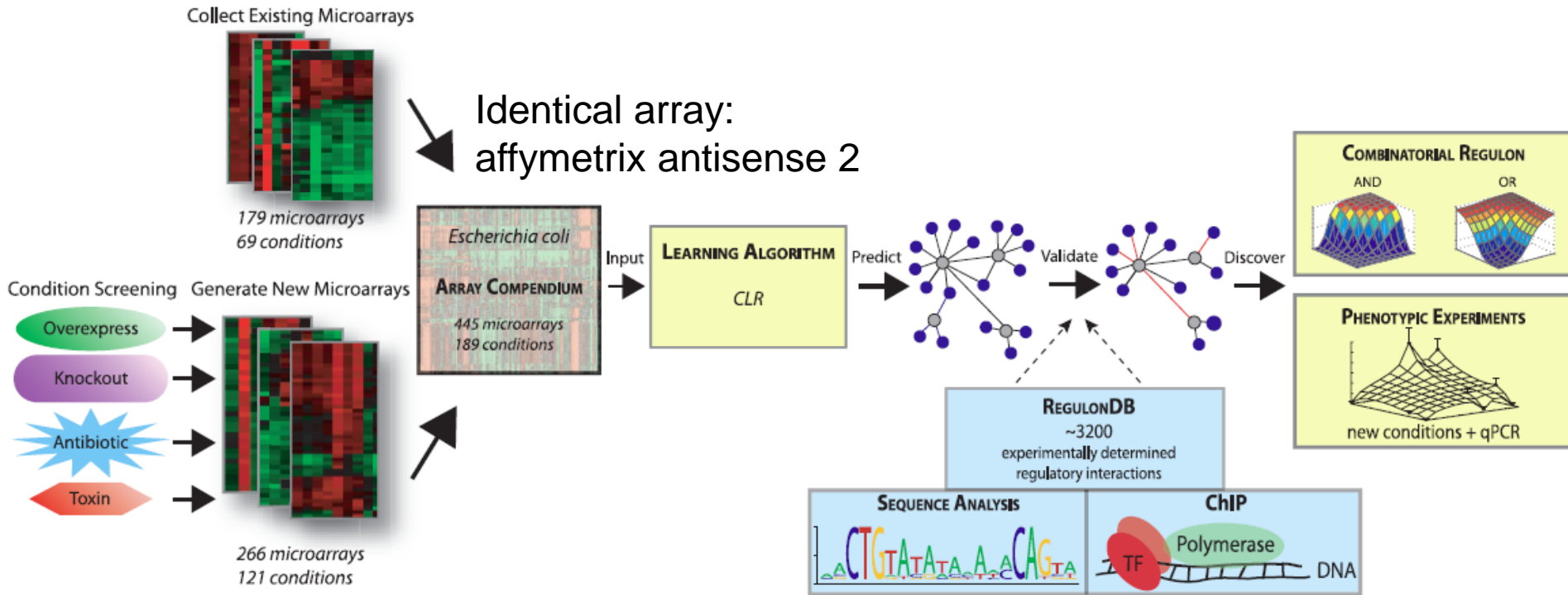


Figure 1. Overview of Our Approach for Mapping the *E. coli* Transcriptional Regulatory Network

Microarray expression profiles were obtained from several investigators. Our laboratory profiled additional conditions, focusing on DNA damage, stress responses, and persistence. These two data sources were combined into one uniformly normalized *E. coli* microarray compendium that was analyzed with the CLR network inference algorithm. The predicted regulatory network was validated using RegulonDB, sequence analysis, and ChIP. The validated network was then examined for cases of combinatorial regulation, one of which was explored with follow-up real-time quantitative PCR experiments.

Summary of the dataset

Table 1. Data Sources for the *Escherichia coli* Microarray Compendium

Publication Title	Arrays	Conditions	Reference
Present study: Large-scale mapping and validation of <i>Escherichia coli</i> transcriptional regulation from a compendium of expression profiles	266	121	(Faith et al.)
Integrating high-throughput and computational data elucidates bacterial networks	43	14	[48]
Genome-scale analysis of the uses of the <i>Escherichia coli</i> genome: Model-driven analysis of heterogeneous data sets	41	20	[49]
Transcriptome profiles for high-cell-density recombinant and wild-type <i>Escherichia coli</i>	32	10	[50]
Amino acid content of recombinant proteins influences the metabolic burden response	16	8	[51]
pH regulates genes for flagellar motility, catabolism, and oxidative stress in <i>Escherichia coli</i> K-12	15	3	[52]
Genome-wide analysis of lipoprotein expression in <i>Escherichia coli</i> MG1655	14	7	[53]
Genome-wide expression analysis indicates that FNR of <i>Escherichia coli</i> K-12 regulates a large number of genes of unknown function	10	3	[54]
Global transcriptional effects of a suppressor tRNA and the inactivation of the regulator <i>frmR</i>	6	2	[55]
Global transcriptional programs reveal a carbon source foraging strategy by <i>Escherichia coli</i>	2	1	[56]

The algorithm comparison

- Relevance Networks
- Algorithm for the Reconstruction of Gene Regulatory Networks (ARACNE)
- Bayesian Network
- Context Likelihood of Relatedness (CLR)
 - Based on Relevance networks and ARACNE

The original Relevance Network

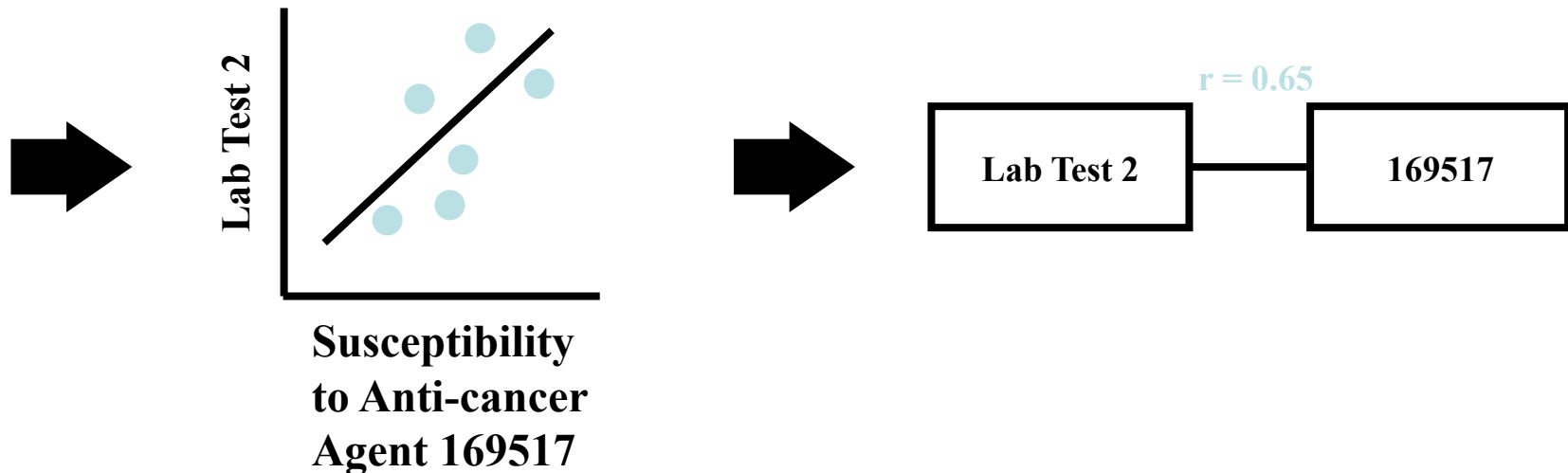
Construction of Relevance Networks

- Patients and cell lines are analyzed as cases
- Clinical parameters, laboratory tests, RNA expression, and susceptibility to anti-cancer agents are all example features of those cases

Patient, Cell Line, Time, etc. ↓	Lab Test 1	Lab Test 2	Clinical Param 1	RNA Expr J02923	Susceptibility to Anti-cancer Agent 169517
	138	3.7	105	0.7	8.1
	134	4.5	99		2.1
	132	5.3	102	7.4	3.3

Construction of Relevance Networks

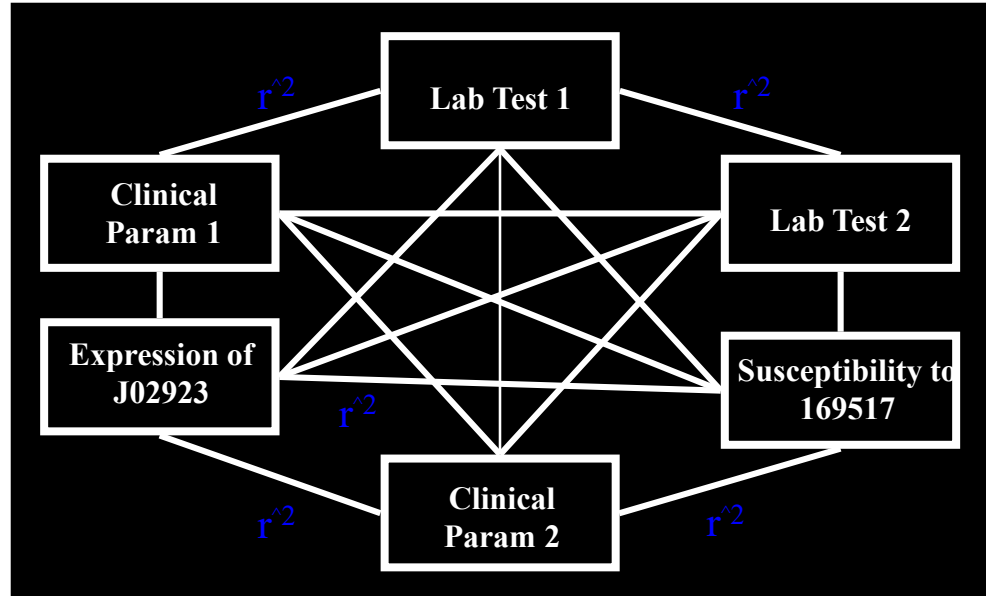
- Perform a pairwise comparison between all features
- For each scatter plot, we fit a linear model and stored
 - Correlation coefficient r



- Every feature is completely connected to every other feature by a linear model of varying quality

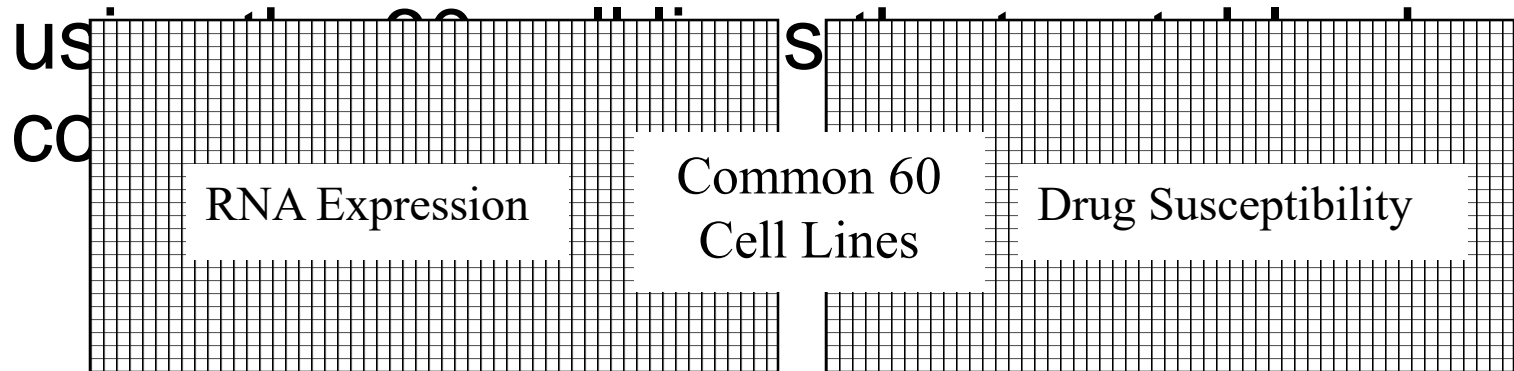
Construction of Relevance Networks

- $r^2 = r^2 * r / \text{abs}(r)$
- Choose a threshold r^2 to split the network
- Drop links with r^2 under threshold
- Breaks the completely connected network into islands where connections are stronger than threshold
- Islands are what we call “relevance networks”
- Display graphically, with thick lines representing strongest links



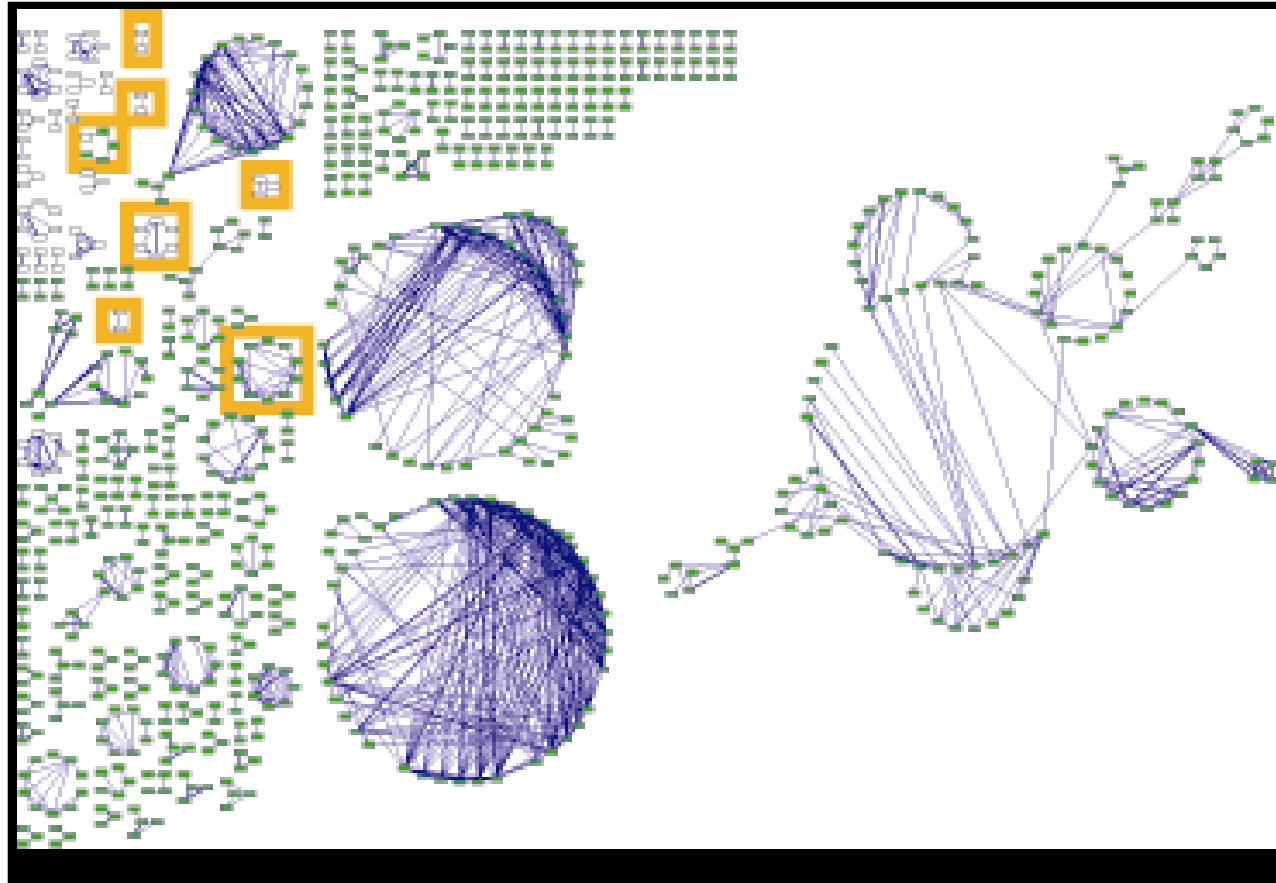
The New Pharmacology

- RNA expression in NCI 60 cell lines was determined using Affymetrix HU6000 arrays
 - 5,223 known genes
 - 1,193 expressed sequence tags
- The RNA expression data set and Anti-cancer drug susceptibility data are integrated, using the Common Cell Lines set, and in



Genes and Anti-Cancer Agents

- Threshold r^2 was 0.8
- 202 networks
- 834 features out of 11,692 (7.1%)
- 1,222 links out of 68,345,586 (.0018%)
- Only one link between a gene and anti-cancer agent



Adapted of relevance network in this study

- Using the correlation between potential target gene y vs transcriptional factor x
- Calculate mutual information instead of r^2
- Mutual information doesn't assume specific properties of the dependence, such as linearity, continuity etc.
- Tradeoffs between true positive and false positive rate in choosing threshold.

Mutual information

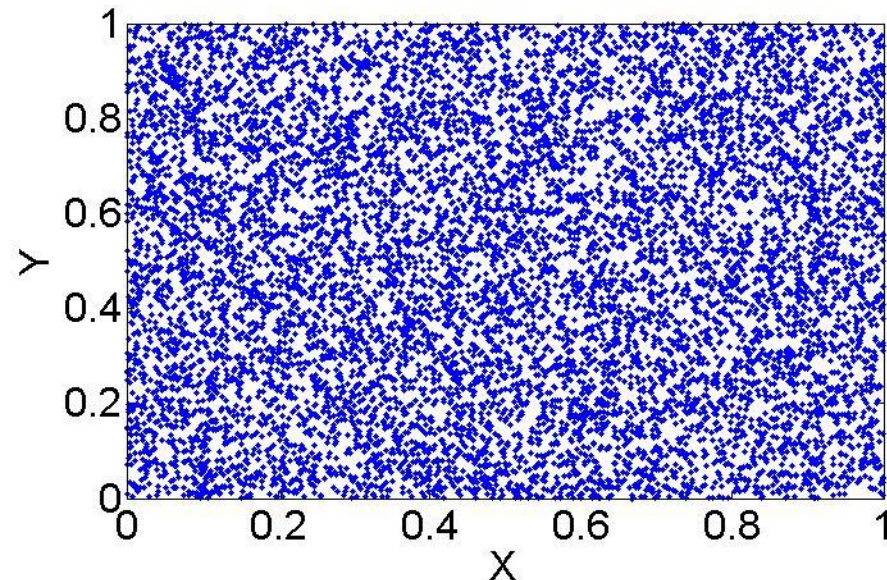
The mutual information for two discrete random variables X and Y is defined as:

$$I(X; Y) = \sum_{i,j} P(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \quad (1)$$

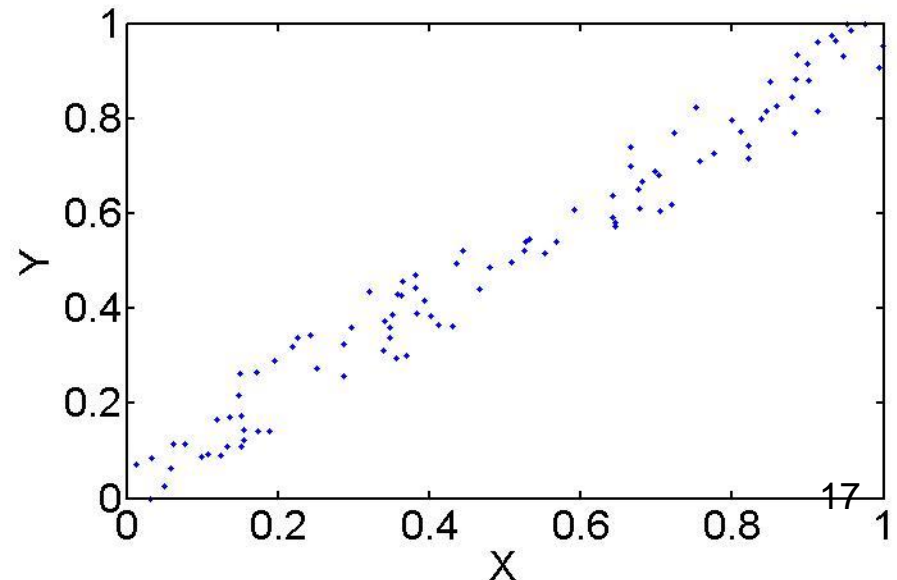
where $P(x_i)$ is the probability that $X = x_i$. For genes, X and Y represent a transcription factor and its potential target gene, and x_i and y_i represent particular expression levels. In the case of continuous

If x and y are not correlated, $I(x,y)=0$, no mutual information

$r^2=0.000001$ $mi=0.005279$



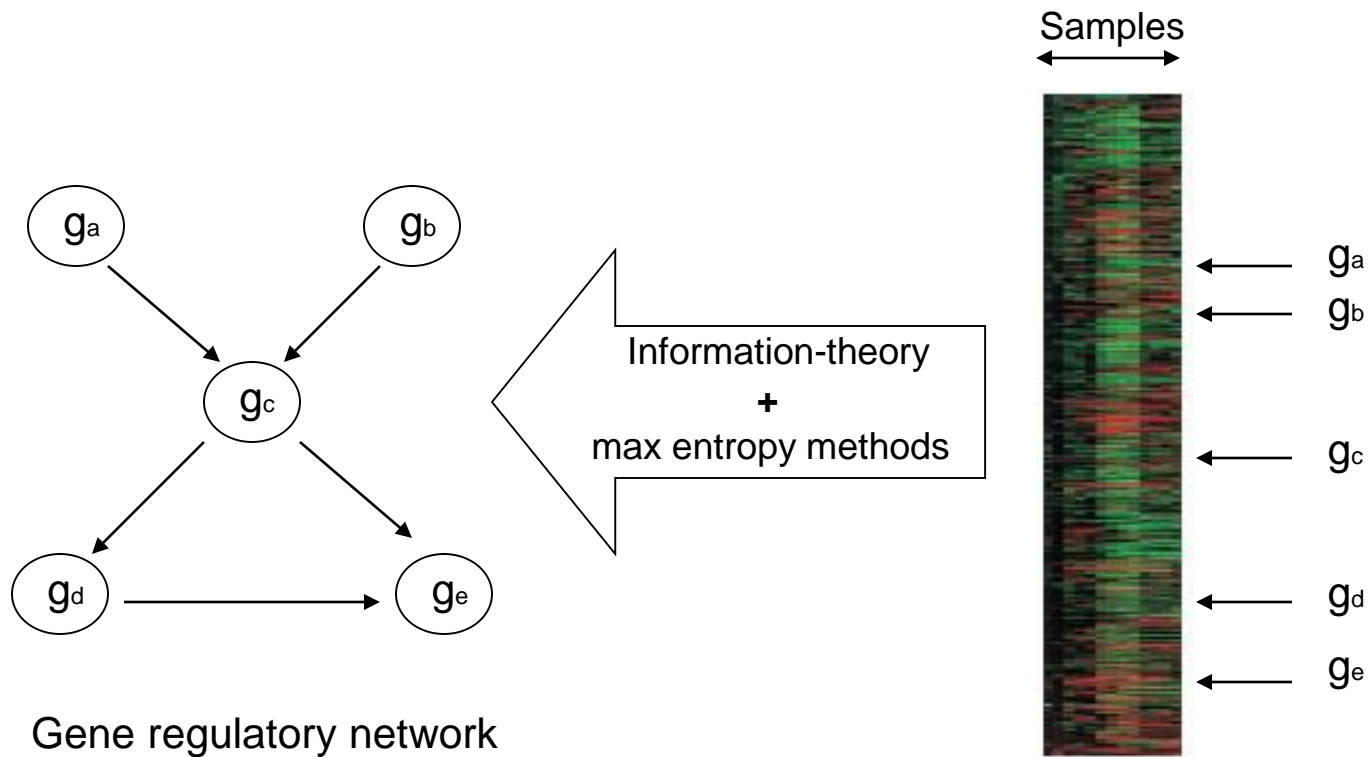
$r^2=0.966635$ $mi=2.126386$



ARACNE

Algorithm for the Reconstruction of Accurate Cellular Networks

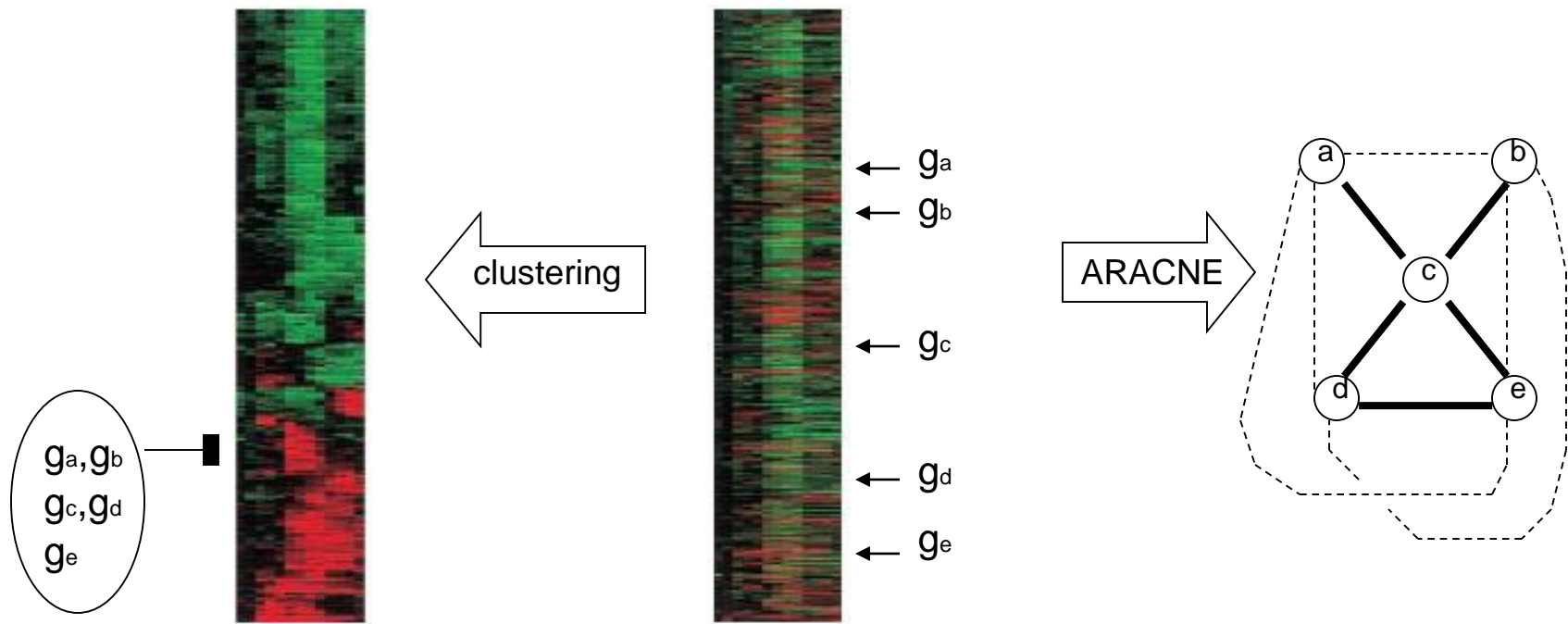
“Reverse engineering” or “deconvolution” problem:



ARACNE vs Clustering

ARACNE recovers specific transcriptional interactions but does not attempt to recover all of them (too complex a problem).

Genome-wide **clustering** of gene expression profiles: cannot discern direct (irreducible) from “cascade” transcriptional gene interactions.



Details (not required)

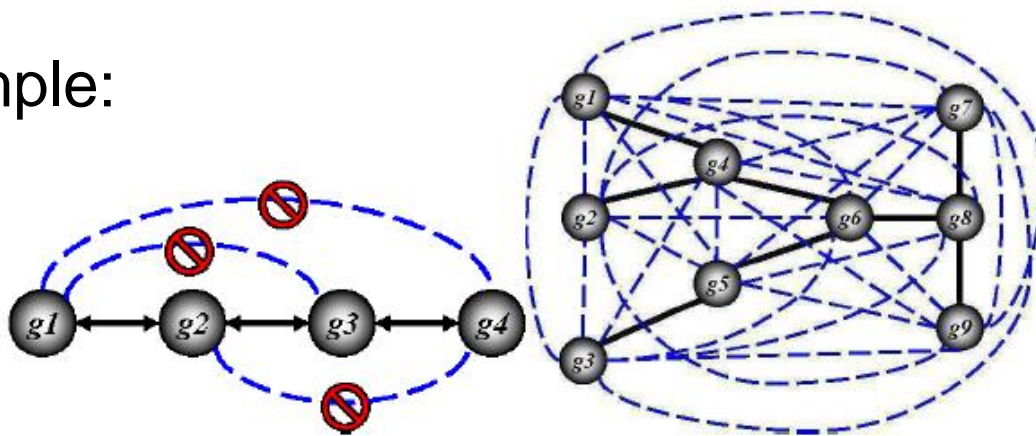
- Assume two-way interaction: pairwise potential determines all statistical dependencies + uniform marginal distributions
- Mutual information (MI) = measure of relatedness

$$I(x, y) = \frac{1}{M} \sum_i \log \left[\frac{p(x_i, y_i)}{p(x_i)p(y_i)} \right]$$

- Independency $I(x_i, y_j) = 0$ if $p(x_i, y_j) = p(x_i)p(y_j)$
- **Data processing inequality**: if genes g_1 and g_3 interact through g_2 then $I(g_1, g_3) \leq \min[I(g_1, g_2); I(g_2, g_3)]$
- **ARACNE starts with network so $I_{ij} > I_0$ for every edge look at gene triplets and remove edge with smallest MI**
- Ignore the direction of the edges
- Reconstruct tree-network topologies exactly
 - higher-order potential interactions will not be accounted for (ARACNE's algorithm will open 3-gene loops).
 - A two-gene interaction will be detected if there are no alternate paths.

ARACNE – Example & Evaluation

- Example:



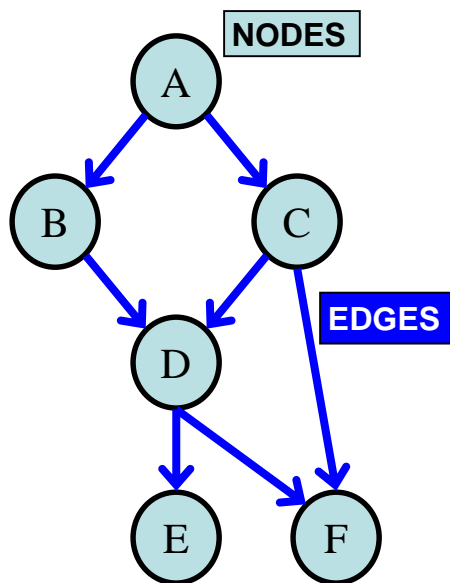
Performance to be assessed via Precision-Recall curves (PRCs)

$$\text{Precision} = \frac{N_{TP}}{N_{TP} + N_{FP}} \quad \text{fraction of true interactions correctly inferred (expected success ratio)}$$

$$\text{Recall} = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad \text{fraction of true interactions among all inferred ones}$$

Bayesian networks

- Marriage between **graph theory** and **probability theory**.



- Directed acyclic graph (**DAG**) representing conditional independence relations.

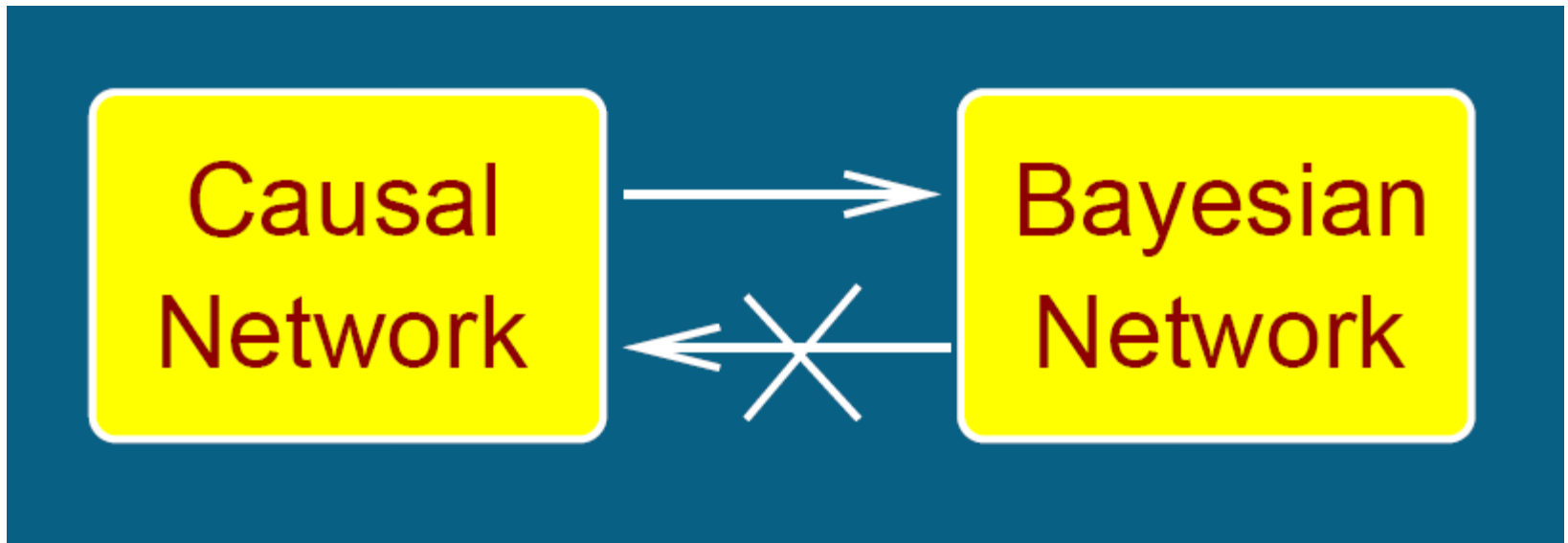
- It is possible to **score** a network in light of the data: $P(D|M)$, D:data, M: network structure.

- We can **infer** how well a particular network explains the observed data.

$$P(A, B, C, D, E, F)$$

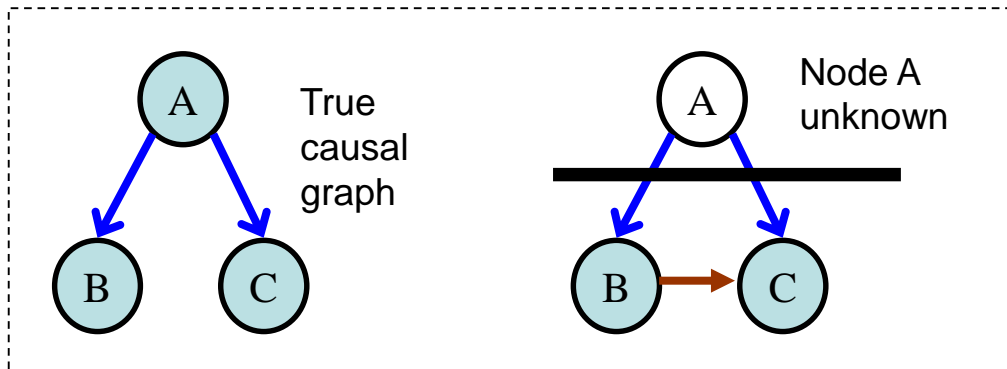
$$= P(A) \cdot P(B | A) \cdot P(C | A) \cdot P(D | B, C) \cdot P(E | D) \cdot P(F | C, D)$$

Bayesian networks versus causal networks

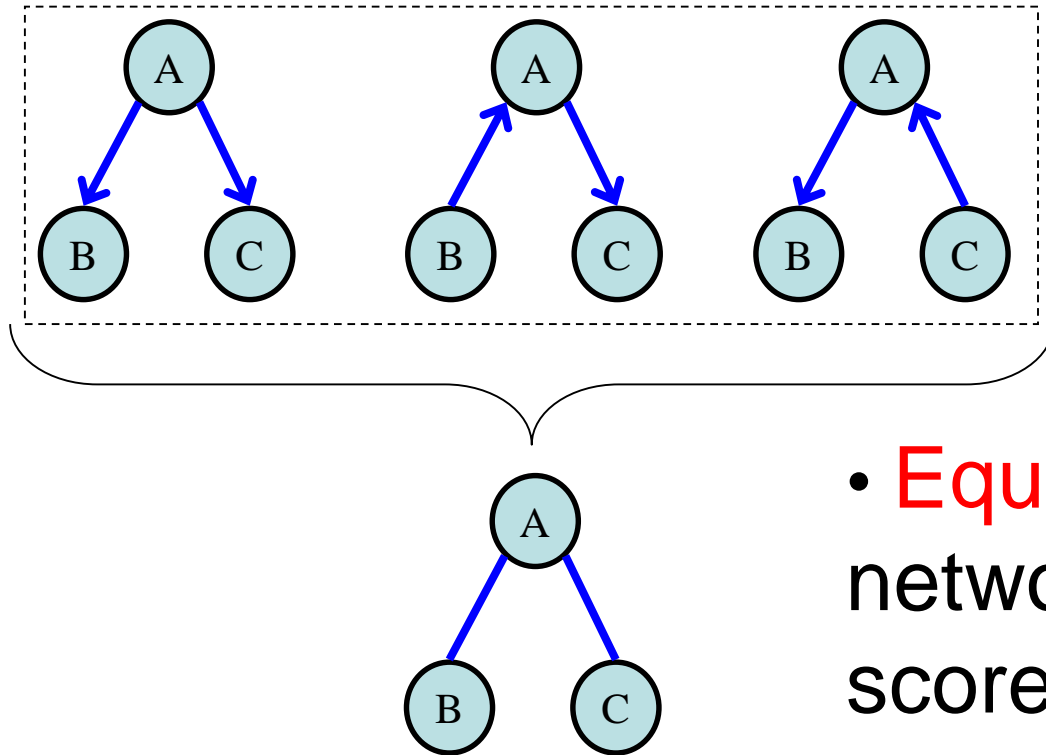


Bayesian networks represent conditional (in)dependence relations - **not** necessarily causal interactions.

Bayesian networks versus causal networks

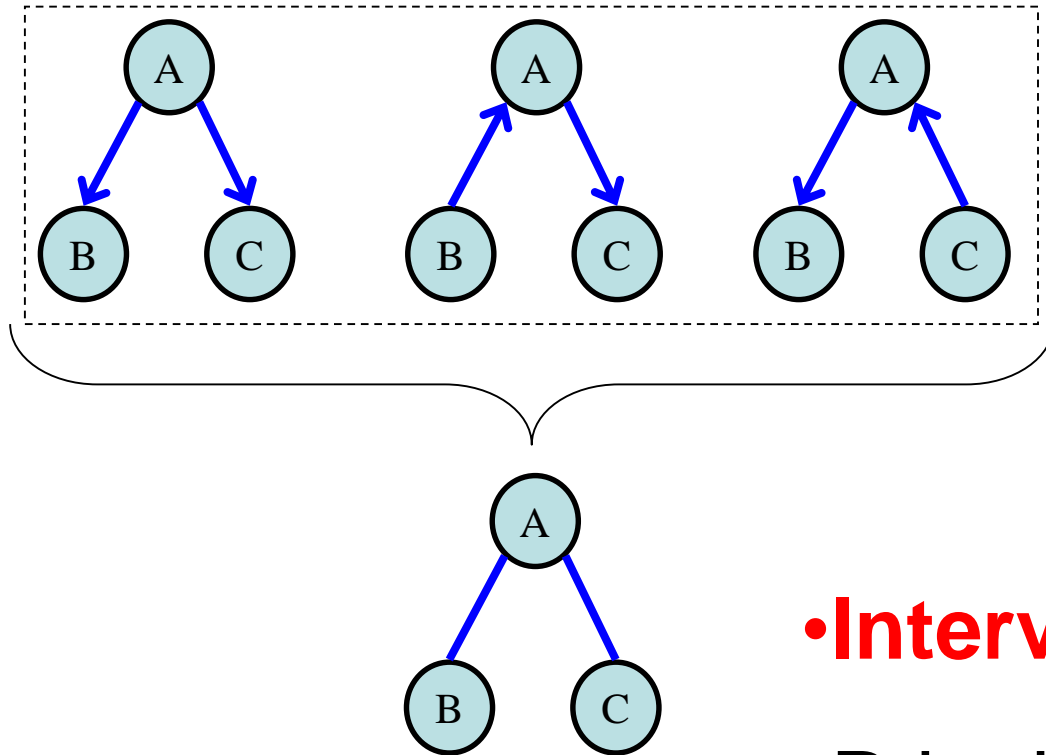


Bayesian networks versus causal networks



- **Equivalence classes:** networks with the same scores: $P(D|M)$.
- Equivalent networks cannot be distinguished in light of the data.

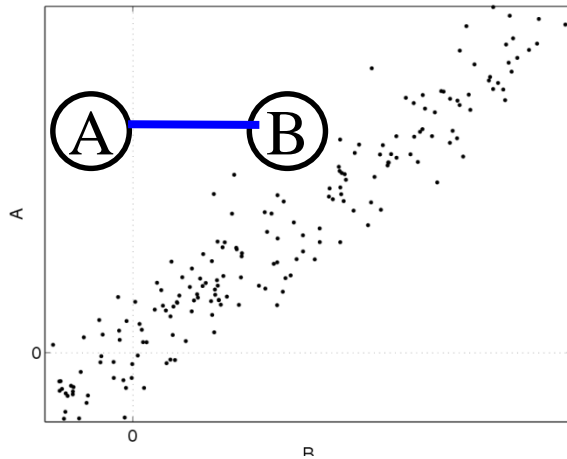
Symmetry breaking



- **Interventions**

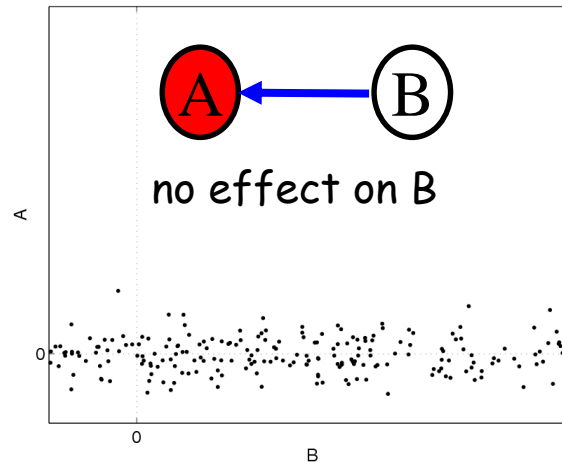
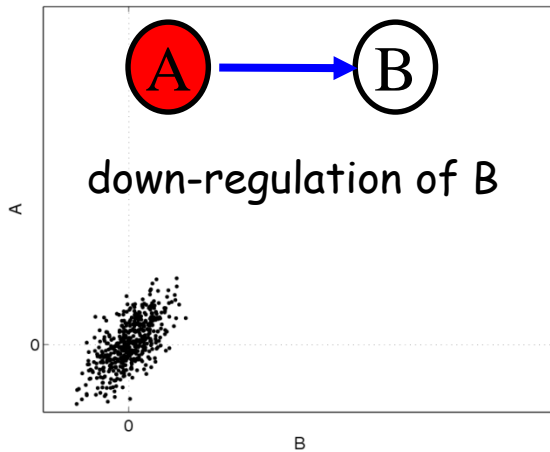
- Prior knowledge

Interventional data



A and B are correlated

inhibition of A

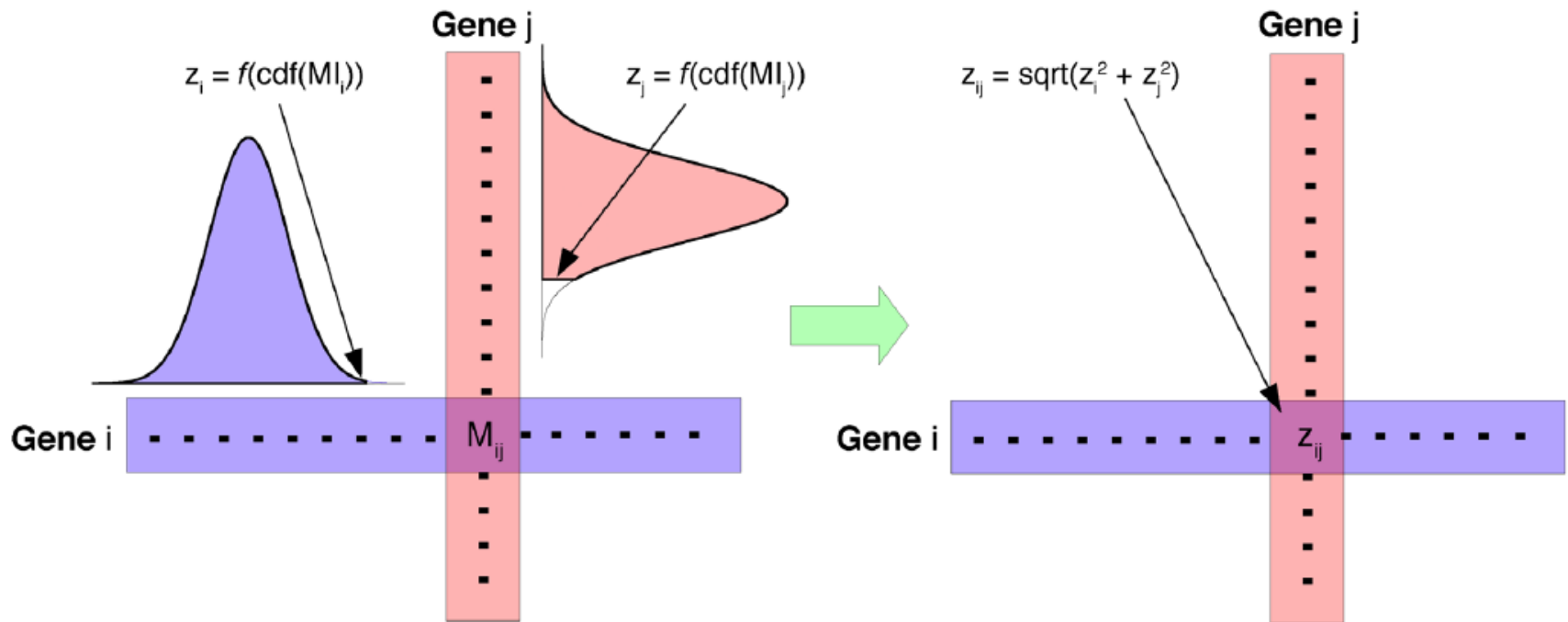


Context Likelihood of Relatedness (CLR)

- Bias from uneven condition sampling and inter-laboratory variations in microarrays complicate network inference.
- indirect regulatory influences and direct (physical) regulatory interactions may not be easily distinguishable from their expression profiles.
- CLR increases the contrast between the physical interactions and the indirect relationships by taking the network context of each relationship into account.

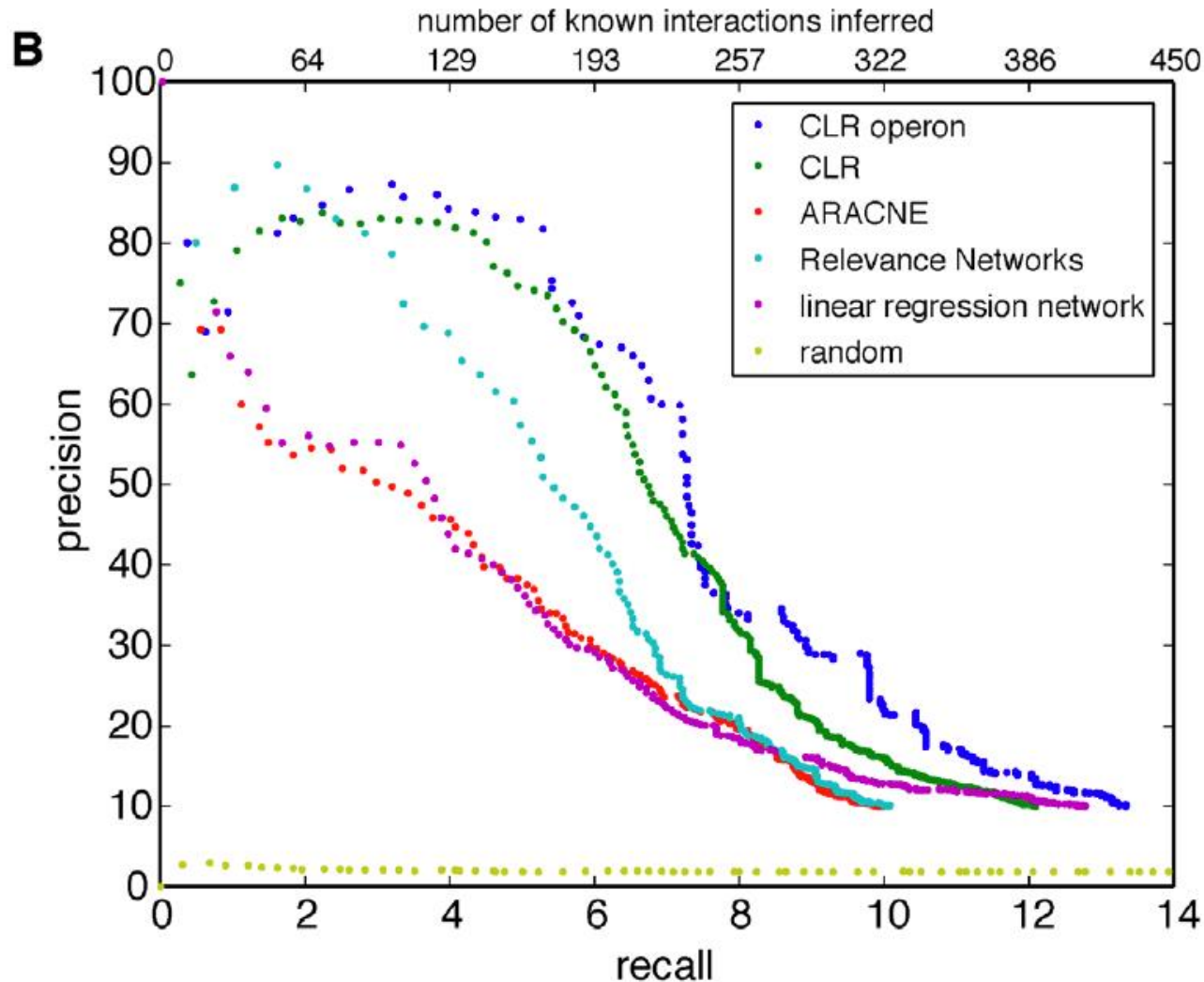
Context Likelihood of Relatedness (CLR)

A



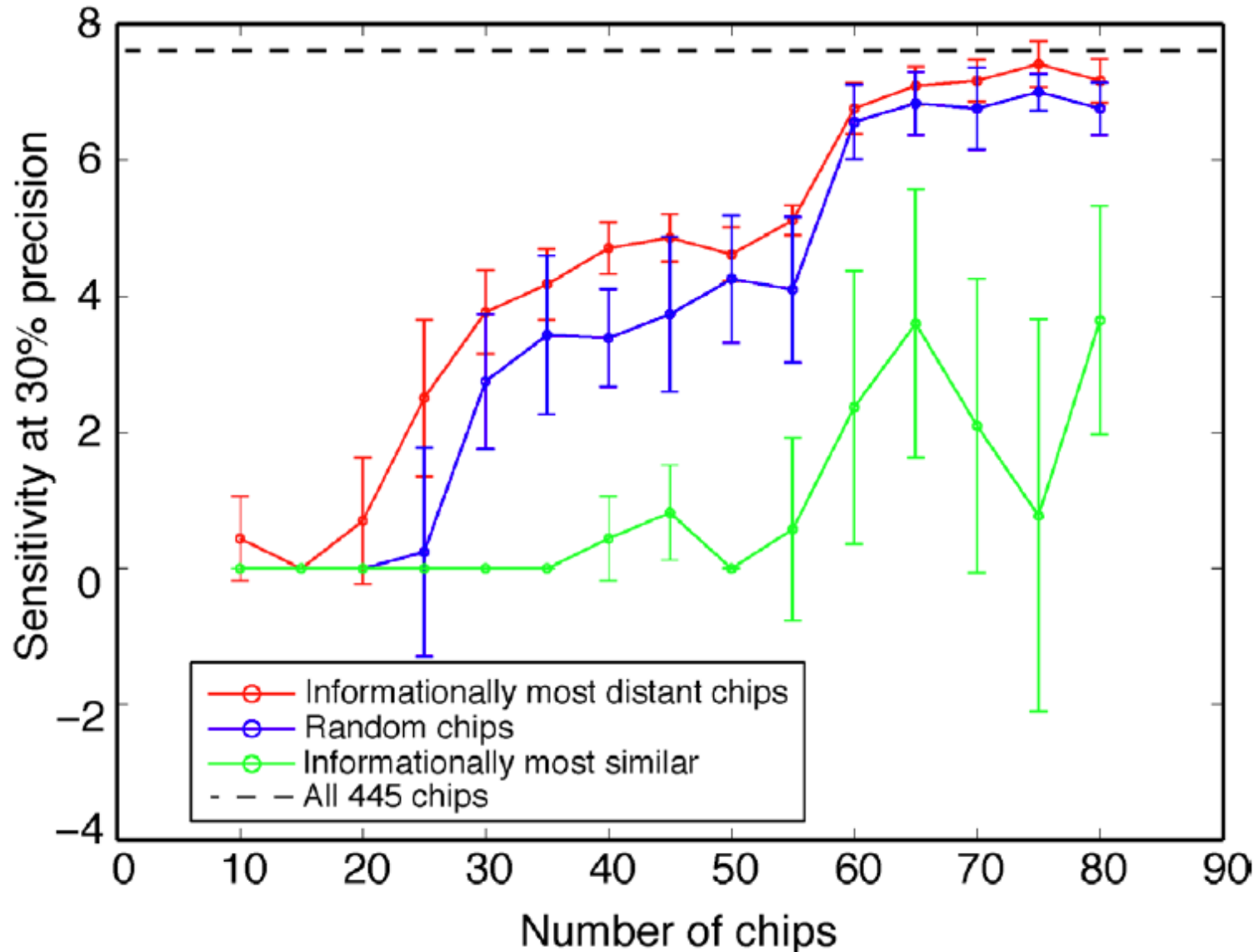
- 1) For gene i , calculate mutual information $MI_{i,k}$ to all the other genes.
- 2) The most of the genes don't interact with gene i , follow normal distribution with mean MM_i and standard deviation s_i .
- 3) The mutual information of gene i and gene j is $MI_{i,j}$, corresponding z-score $Z_i = (MI_{i,j} - MM_i) / s_i$.
- 4) The z-score for interaction between gene i and j is $f(Z_i, Z_j) = \sqrt{Z_i^2 + Z_j^2}$
- 5) This approach takes the variations of the background bias in the data set into consideration, using threshold for z-score, instead of using an uniform threshold for MI.

CLR outperform other algorithms based on precision-recall curve



CLR operon: for genes grouped into the same operon, each gene assigned the highest z-score of those gene.

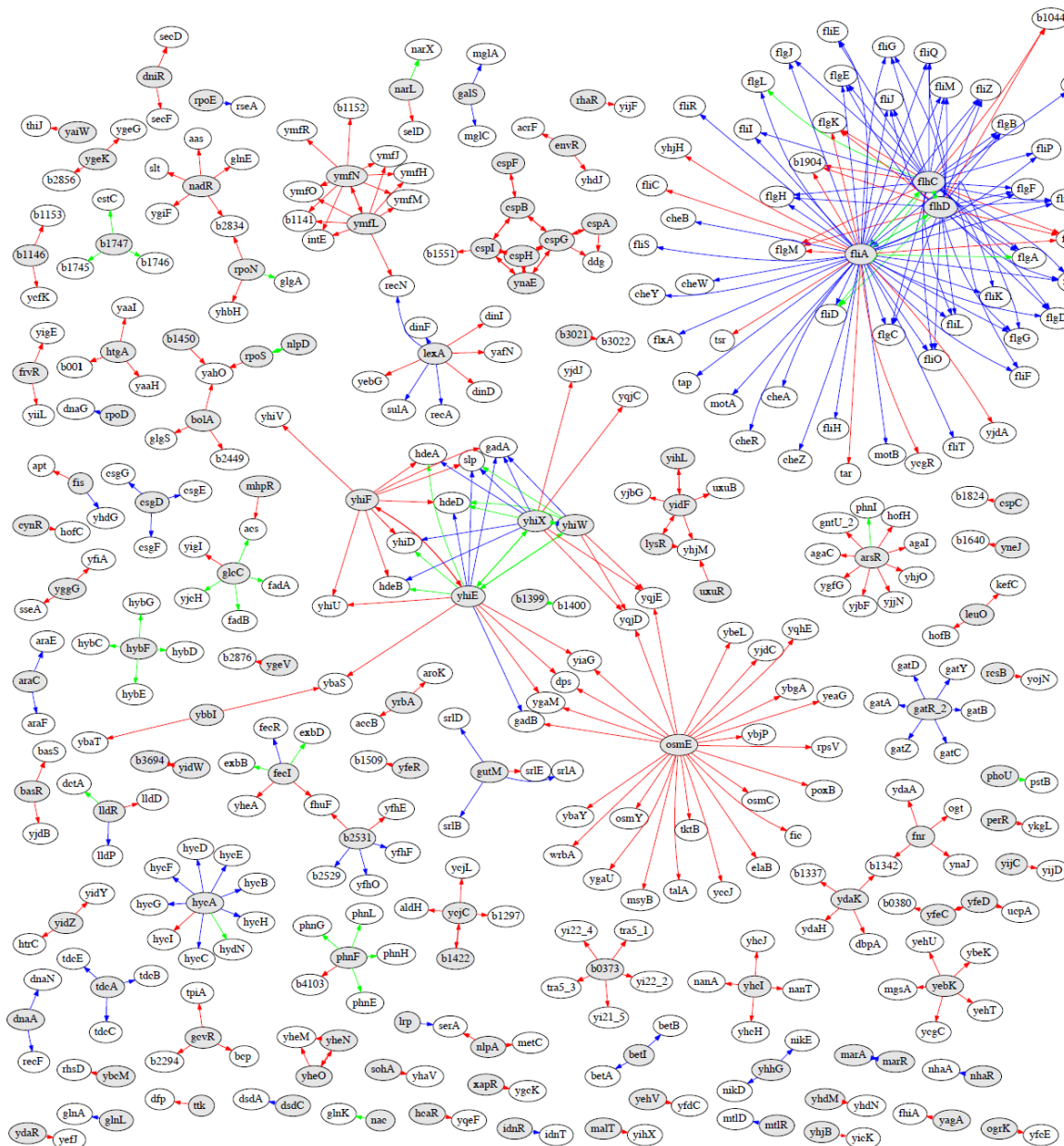
CLR can achieve maximum recall and precision with as little as 60 expression profiles



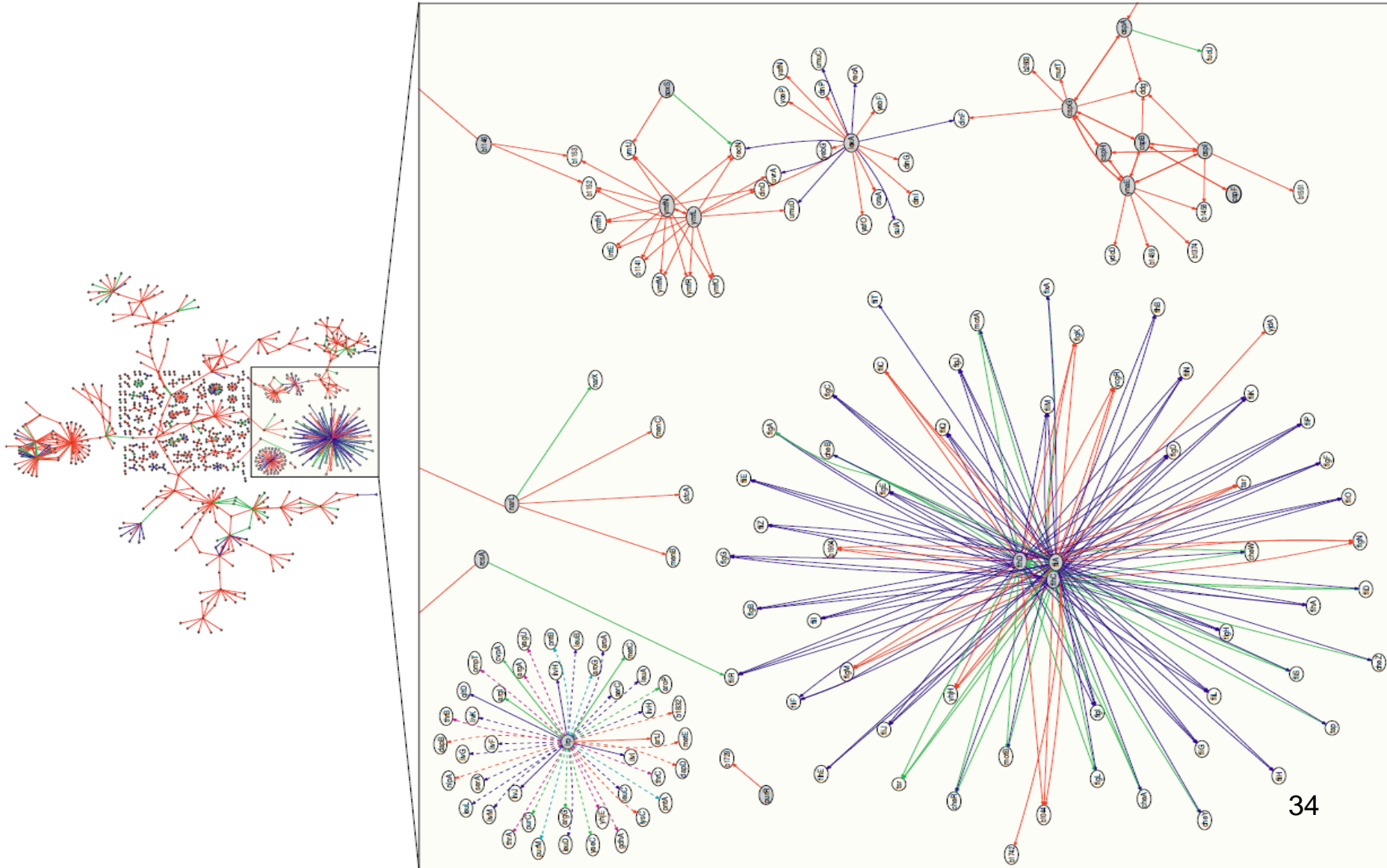
Indication from these 60 chips

- Large environmental perturbations are the most common conditions amongst these 60 profiles, suggesting environmental perturbations are generally more informative than genetic perturbations for network inference.
- The remainder of the profiles in the compendium contribute mainly redundant information.
- the recall achieved by the CLR algorithm appears to be limited largely by the low phenotypic diversity of the dataset.

All the regulations at 80% precision



All the regulations at 60% precision



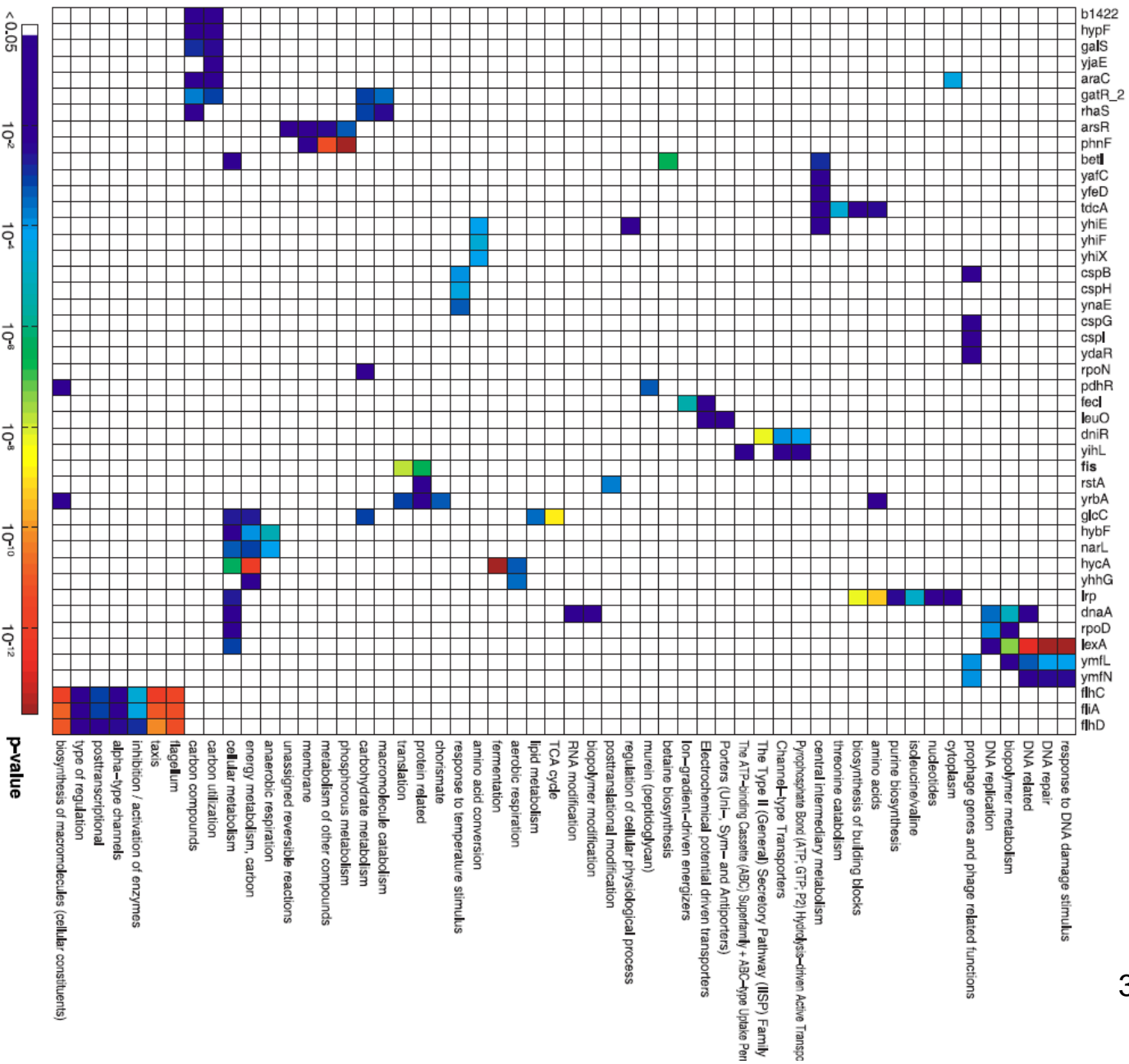


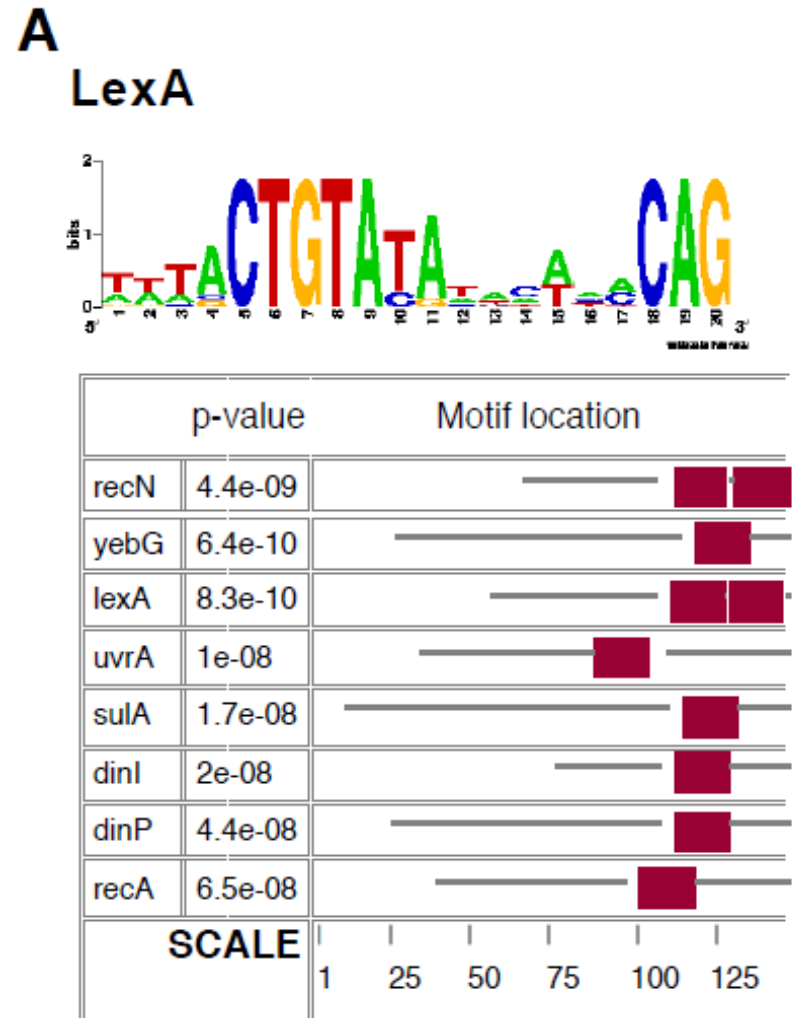
Figure 4. Annotation of Transcription Factor Function by Functional Enrichment Using Predicted Targets from the 60% Precise Network

Discovery of novel regulatory pathways

- Further validation of CLR by sequence analysis of regulatory motifs.
- In vivo confirmation of new regulatory interactions.
- A combinatorial link between central metabolism and iron transport.

Further validation of CLR by sequence analysis of regulatory motifs

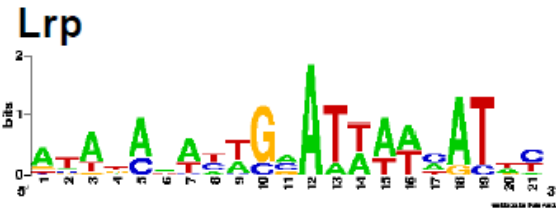
- Using the set of gene targets predict for each TFs, infer the sequence motif bound by the TF 150bp upstream of TSS.
- For those TFs with enough targets, the motif provides a specific location for regulatory interactions.
- LexA: DNA repair, best perturbed regulator, well-predicted.
- Detect a significant binding motif for 28 out of 61 TFs.



The known motif is found in 37
8 out of 13 promoters

Other examples

B



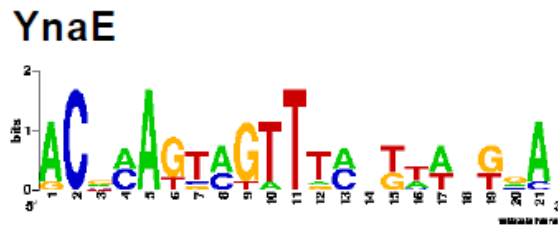
	p-value	Motif location
argI	7e-06	
argG	1.1e-08	
artP	9.4e-08	
nlpA	2e-07	
gltB	7.7e-07	
metC	9.4e-07	
argA	9.4e-07	
metE	1.1e-06	
b1832	1.5e-06	
livJ	1.8e-06	
pntB	6.1e-06	
serC	8e-06	
yhjE	9.8e-06	
SCALE		1 25 50 75 100 125

The known motif is found in
13 out of 25 promoters

a regulator of multiple
biosynthetic operons

Putative novel regulons

C

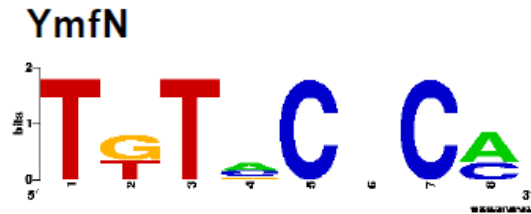


	p-value	Motif location
cspB	3.9e-10	
cspG	6.3e-09	
b1374_s	1.1e-08	
cspH	3.7e-08	
b1459	2.9e-07	
rhsE	6.1e-07	
SCALE		1 25 50 75 100 125

A conserved motif is found in
6 out of 8 promoters

DNA binding protein

D



	p-value	Motif location
ymfJ	1.5e-05	
b1146	0.00064	
recN	0.00033	
intE	0.0003	
b1152	0.00033	
dinD	0.00036	
SCALE		1 25 50 75 100 125 150

A conserved motif is found in
6 out of 6 promoters

DNA binding protein

In vivo confirmation of new regulatory interactions

- Using ChiP and quantitative PCR to obtain physical confirmation of the regulatory interaction inferred by CLR
- Previous studies: validation: a few hand-picked interactions
- This study: unbiased, regulonDB and all targets of particular TFs
- Lrp, PdhR, Fecl: substantial connectivity, each TF, 26-35 operons, totally 93 operon, 244 genes.

Additions to the regulonDB

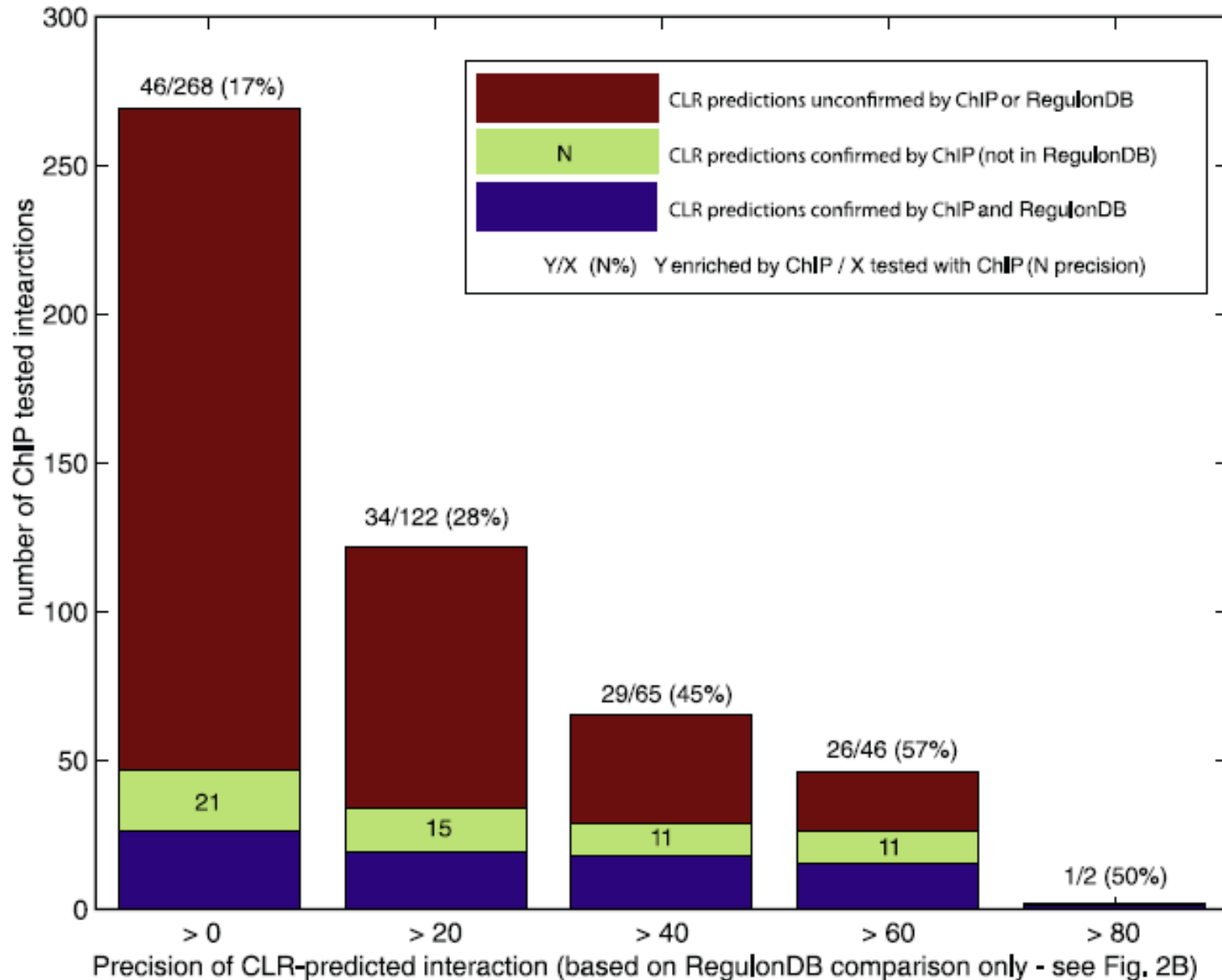
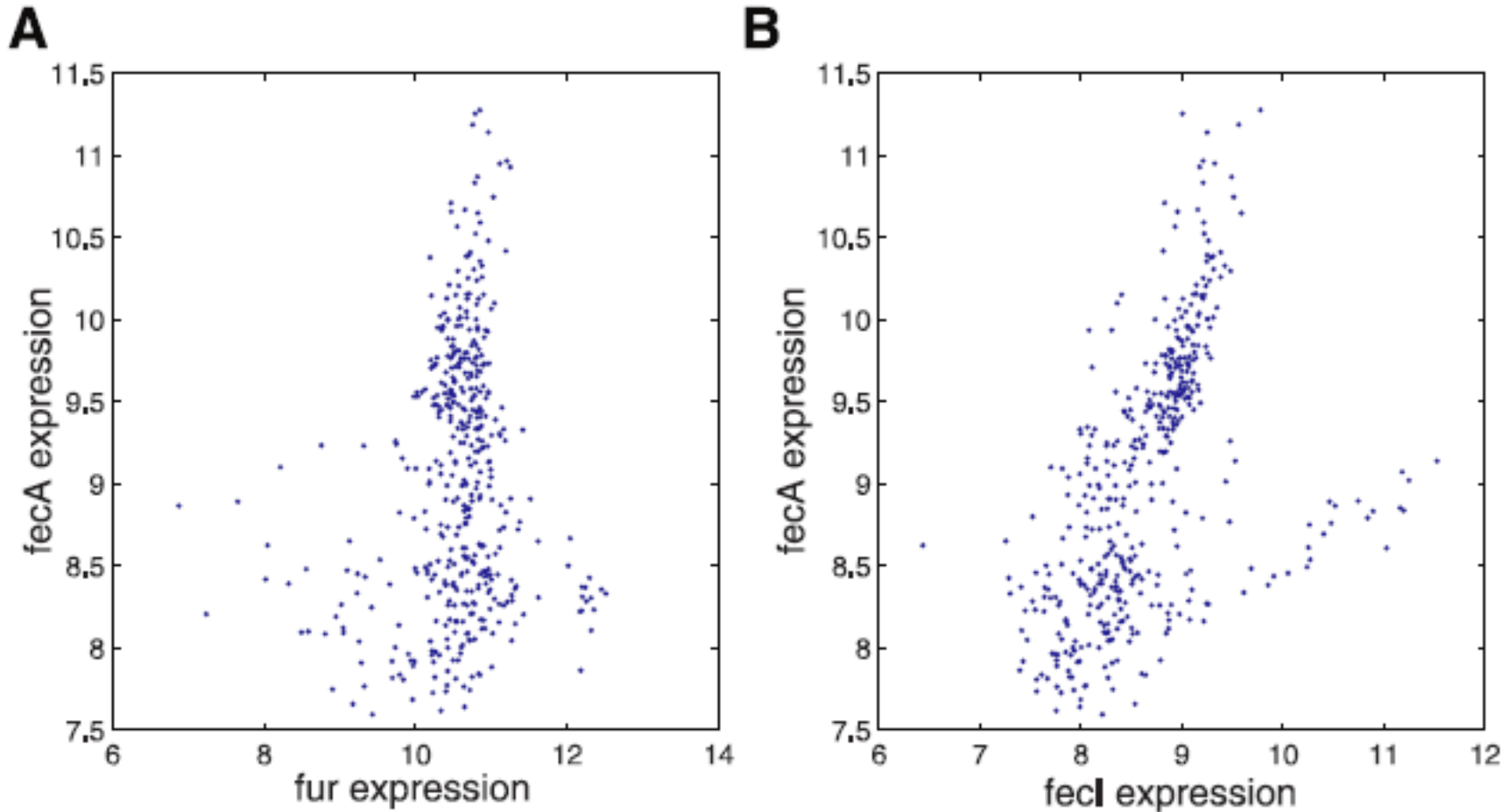


Figure 5. Experimental Validation of Inferred Regulatory Interactions

A combinatorial link between central metabolism and iron transport

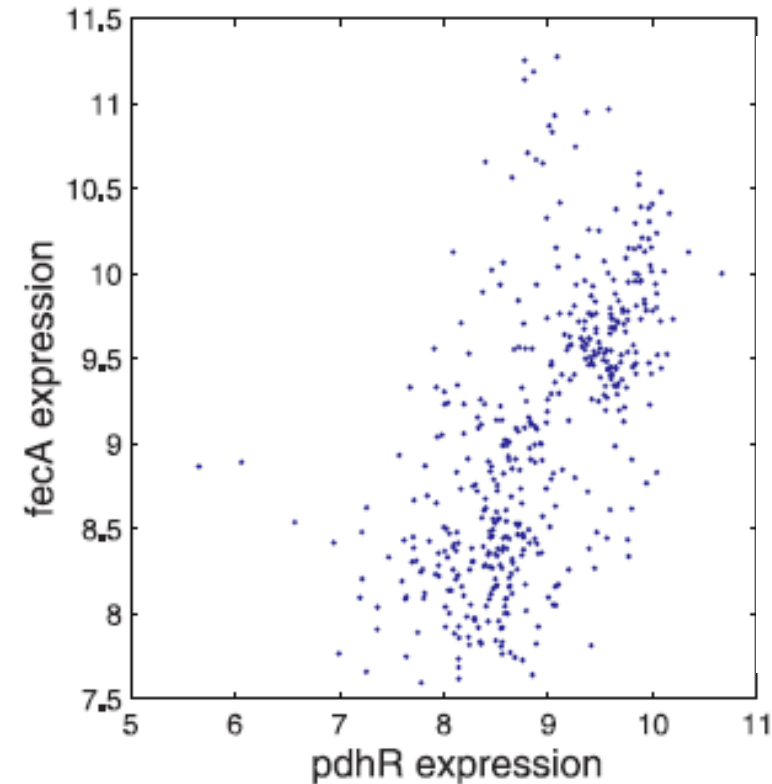
- The inferred regulatory network revealed new combinatorial regulation at many promoters.
- Detailed real-time quantitative PCR analysis of the novel PdhR-*fecA* interaction, which is an interaction that links central metabolism to the control of iron import—a link of potential importance in bacterial virulence and stress protection.
- *fecABCDE* is an operon that encodes a ferric citrate transporter and plays a central role in the import of cellular iron.
- Existing literature described only two regulators of *fecABCDE*—FecI and Fur. The Fur regulation is not apparent in the compendium

fur apparently is not correlated with fecA.
fecI is correlated with fecABCDE, with complex
combinatory regulations



CLR identified PdhR, a pyruvate sensing suppressor and component of energy transduction cascade

C



D

pdhR motif-2 AGTTGTTAAATGTGCA
pdhR motif-1 AATTGGTAAGACCAATT
fecA motif GGAAAAATTAATTCCTTATTTCGATTGTCCTTTTACCCTTCTCGTTGACTCAT

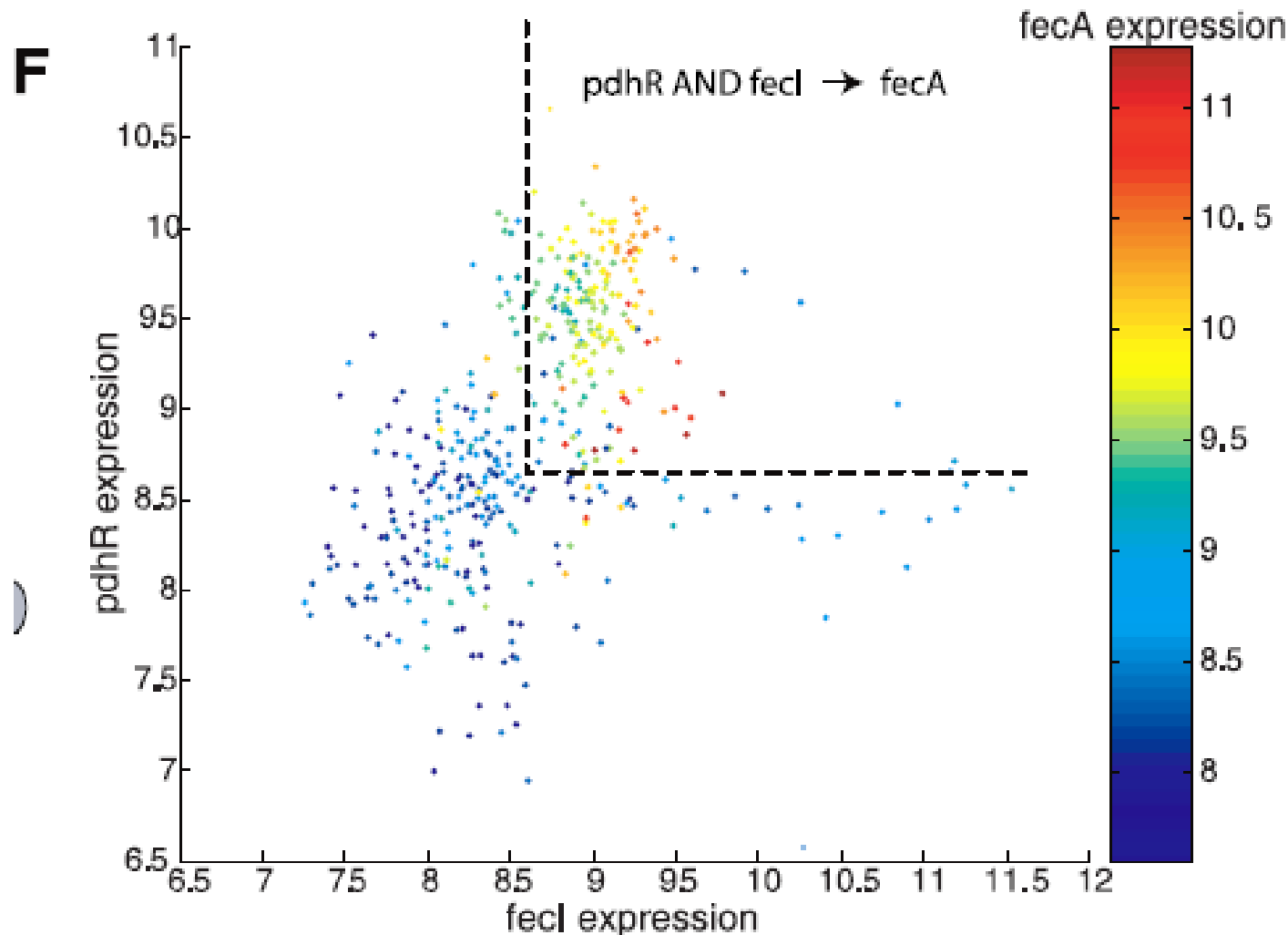
fur binding sites fecI binding site

E

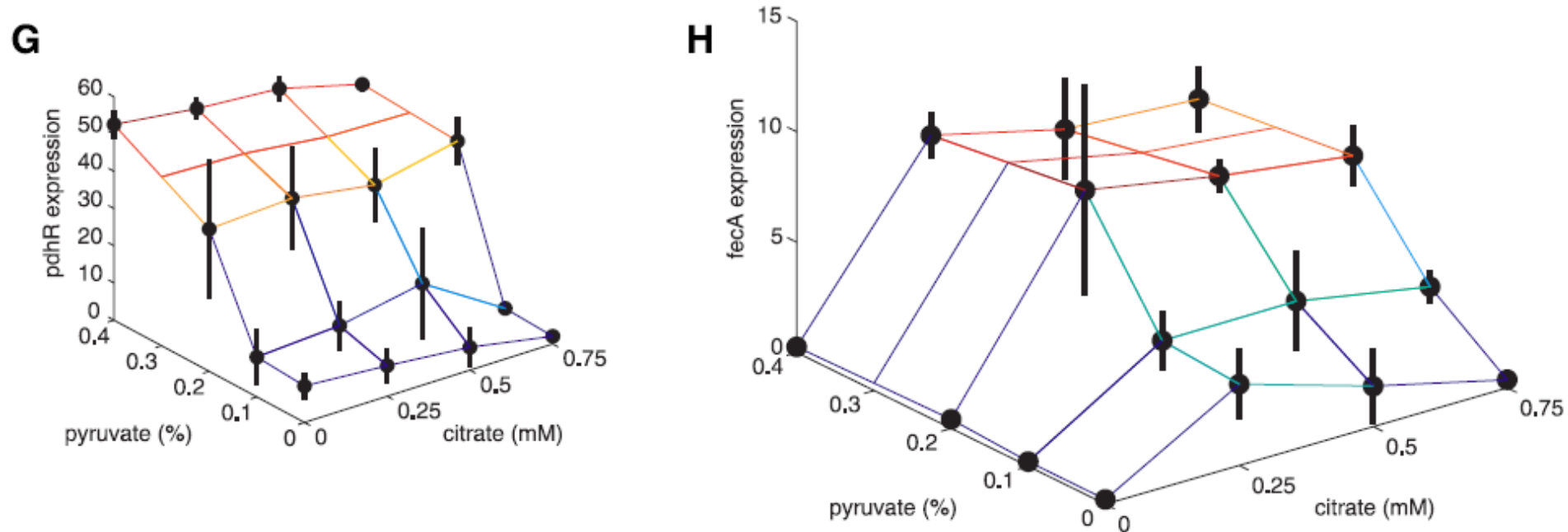


Data shows that *pdhR* correlates with *fecA* expression.
CLR also find *pdhR* binding motif to *fecA* operon.

Compedium data suggest an AND-like logic for combining regulation by pdhR and fecI



Combinatory experiments validate the AND gate regulation of *fecA* by *pdhR* and *fecI*



Pyruvate activate *pdhR* by release of the repression. Citrate activate *fecI*.
qRT-PCR validates the AND gate.

Discovery of new link between central metabolism and iron transport.

Conclusion

- CLR algorithms
- Large integrated dataset, redundant
- Ion transportation and central metabolism coupled at transcription