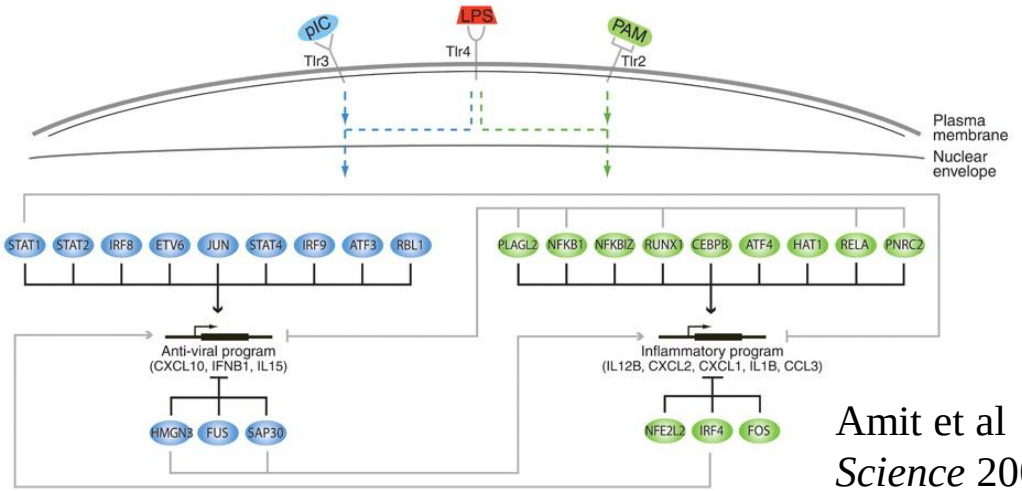
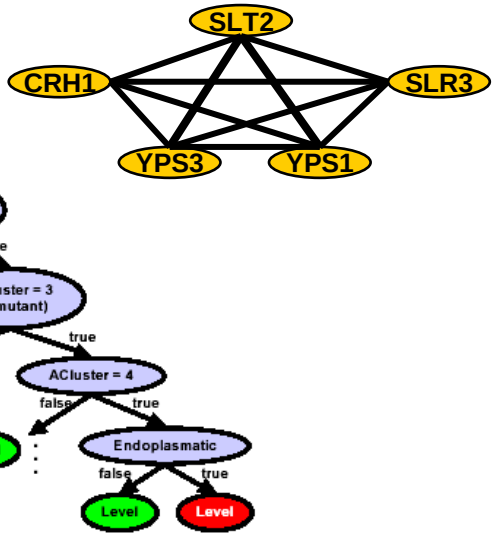
02-710
Computational Genomics

Reconstructing dynamic regulatory networks in multiple species

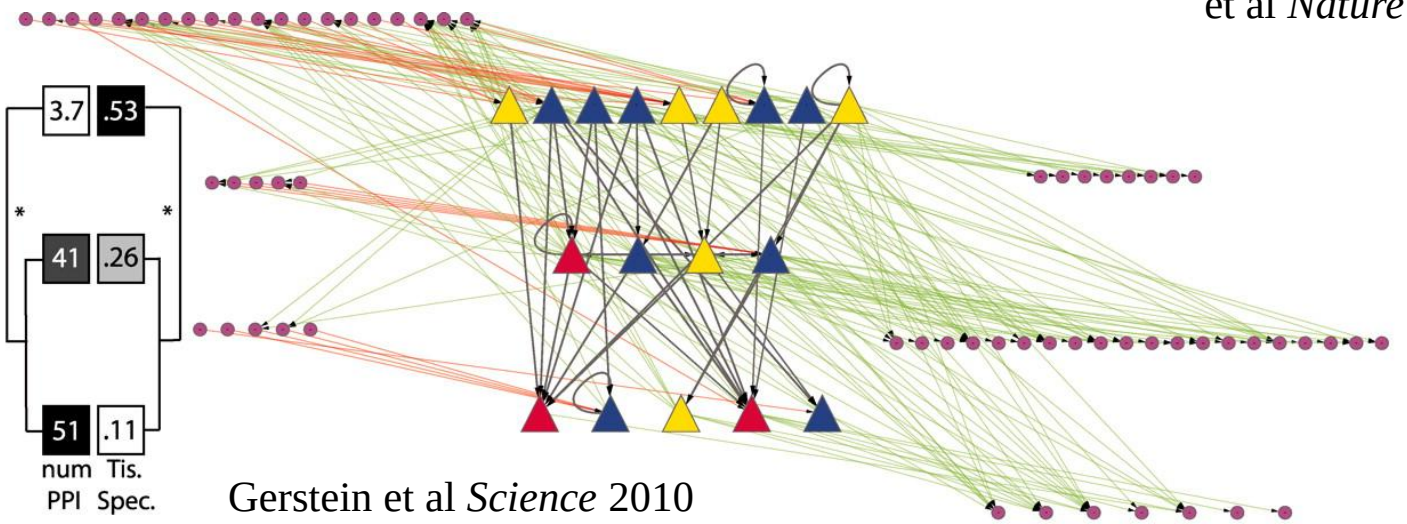
Methods for reconstructing networks in cells



Amit et al *Science* 2009



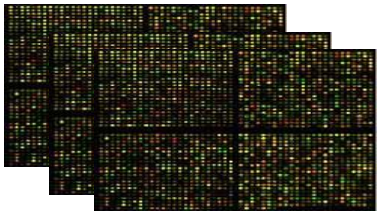
Pe'er et al *Recomb* 2001, Segal et al *Nature Genetics* 2003

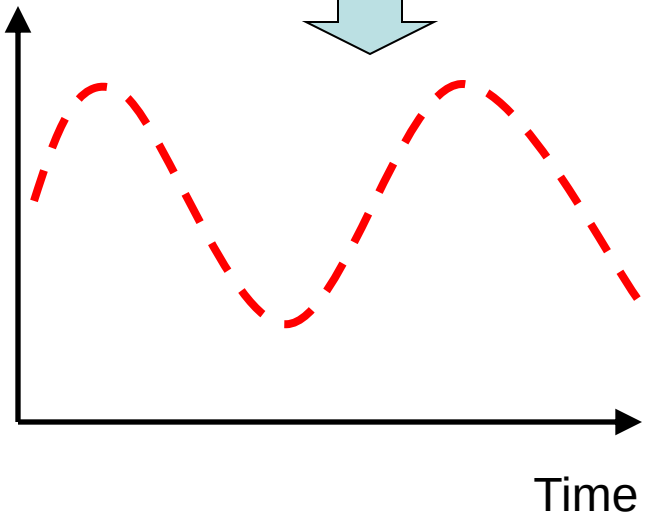


Gerstein et al *Science* 2010

Key problem: Most high-throughput data is static

Time-series measurements



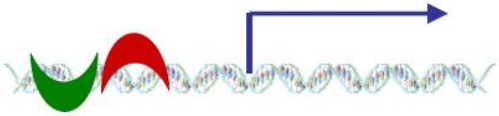


Static data sources

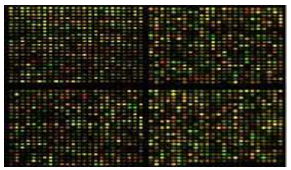
DNA

motif

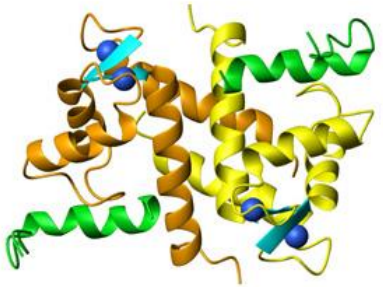
CHIP-chip



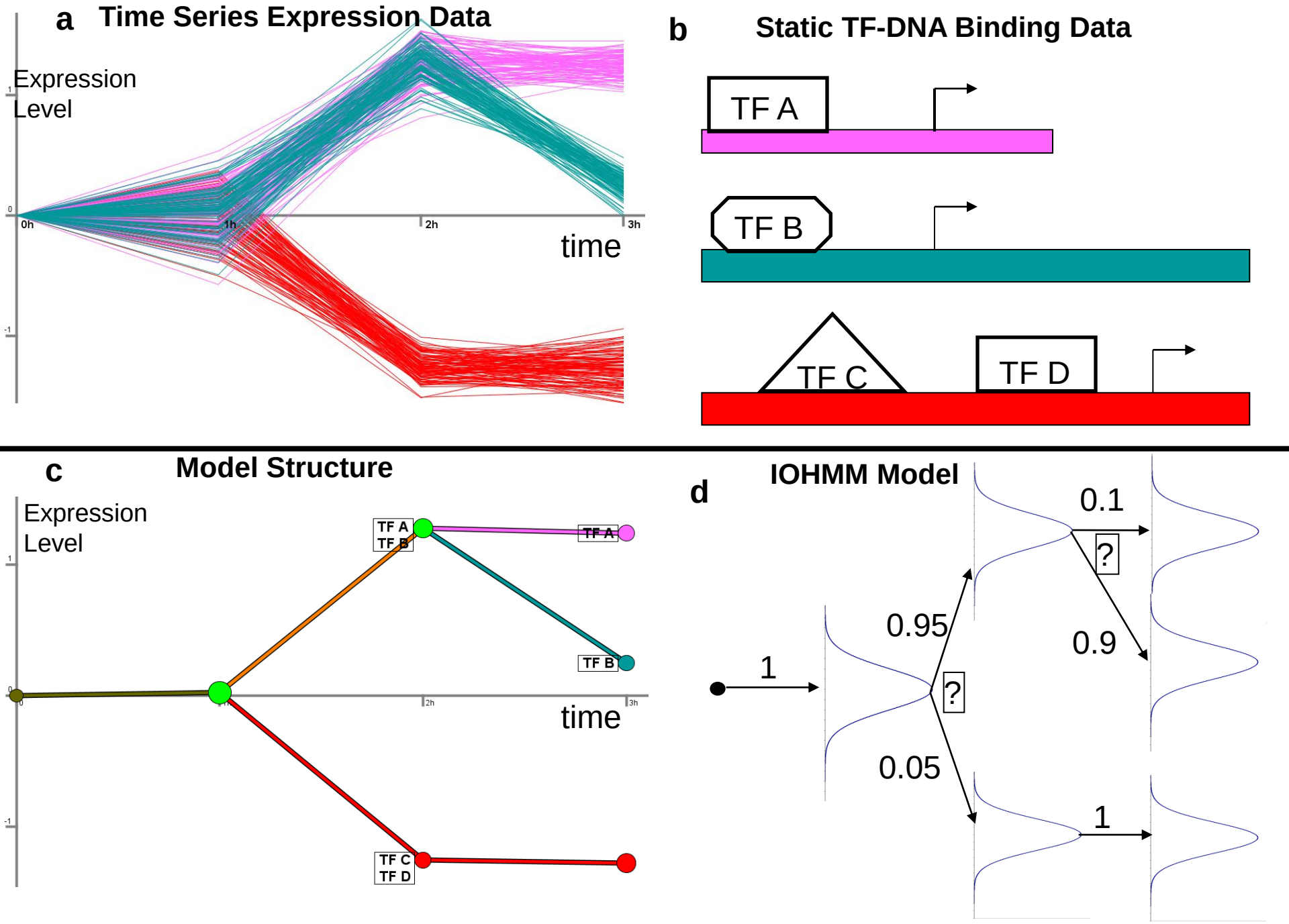
microarray



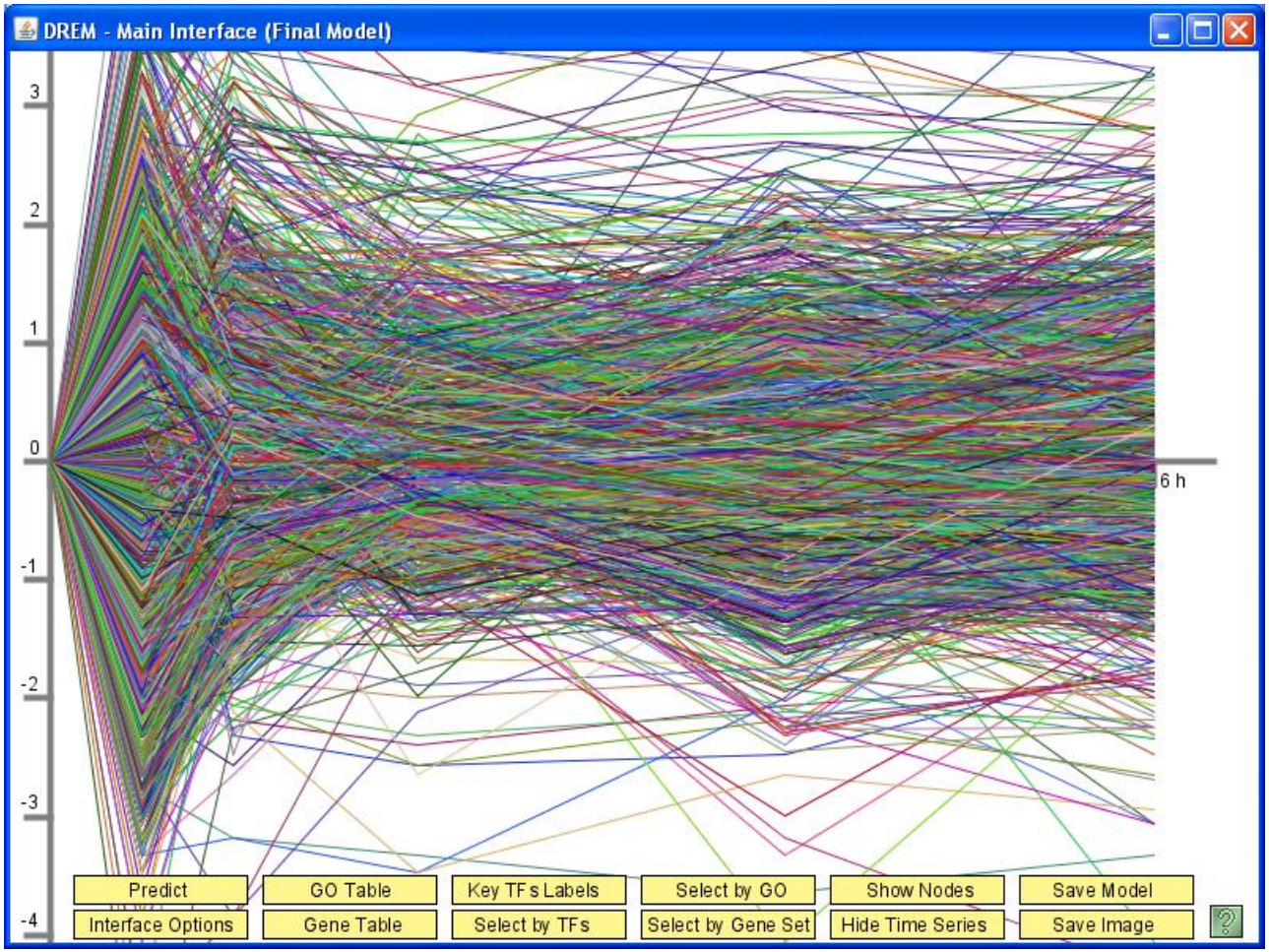
PPI



Method: Integrating time series expression and static protein-DNA interaction data


a
Time Series Expression Data
Expression Level
0h
1h
2h
3h
time
b
Static TF-DNA Binding Data
TF A
TF B
TF C
TF D
c
Model Structure
Expression Level
TF A TF B
TF A
TF B
TF C TF D
time
d
IOHMM Model
1
0.95
0.05
0.1
0.9
1
?
?

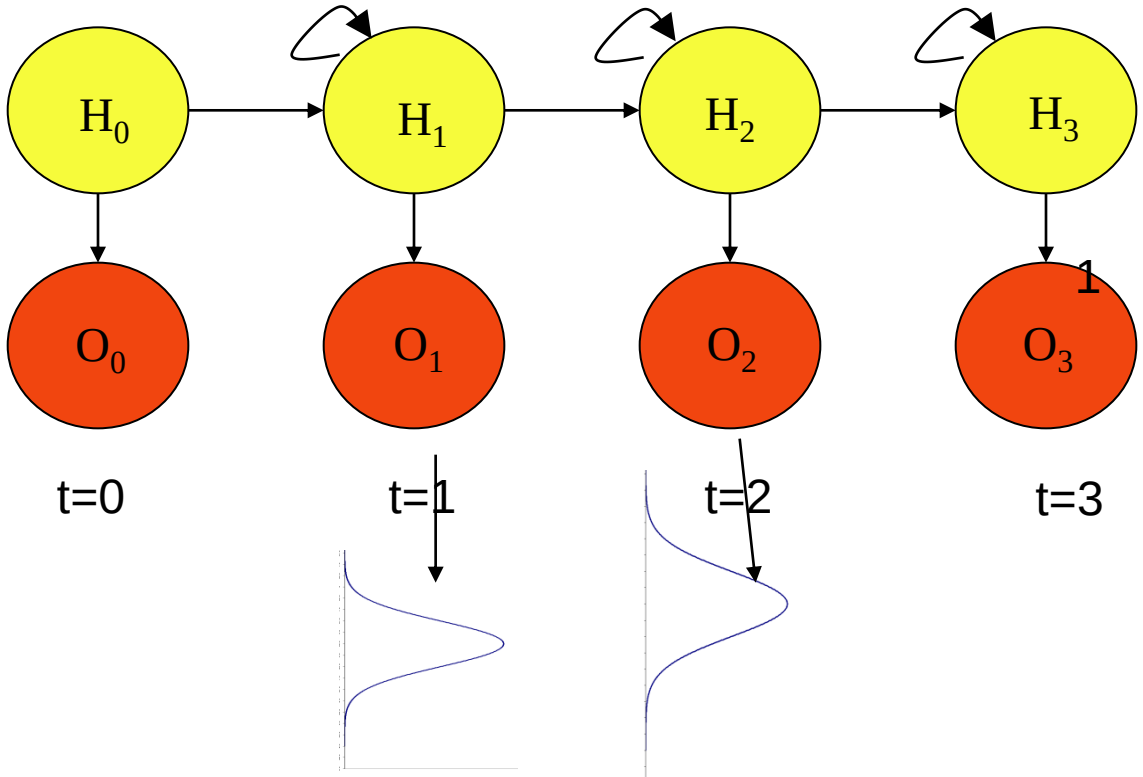
Things are a bit more complicated: Real data



A Hidden Markov Model

Hidden States

Observed outputs
(expression levels)



$$L(H,O;\Theta)=\prod_{i=1}^{n}\left[\prod_{t=1}^{T}p(O_t(i)\mid H_t(i))\right]\left[\prod_{t=2}^{T}p(H_t(i)\mid H_{t-1}(i))\right]$$

Schliep et al *Bioinformatics* 2003

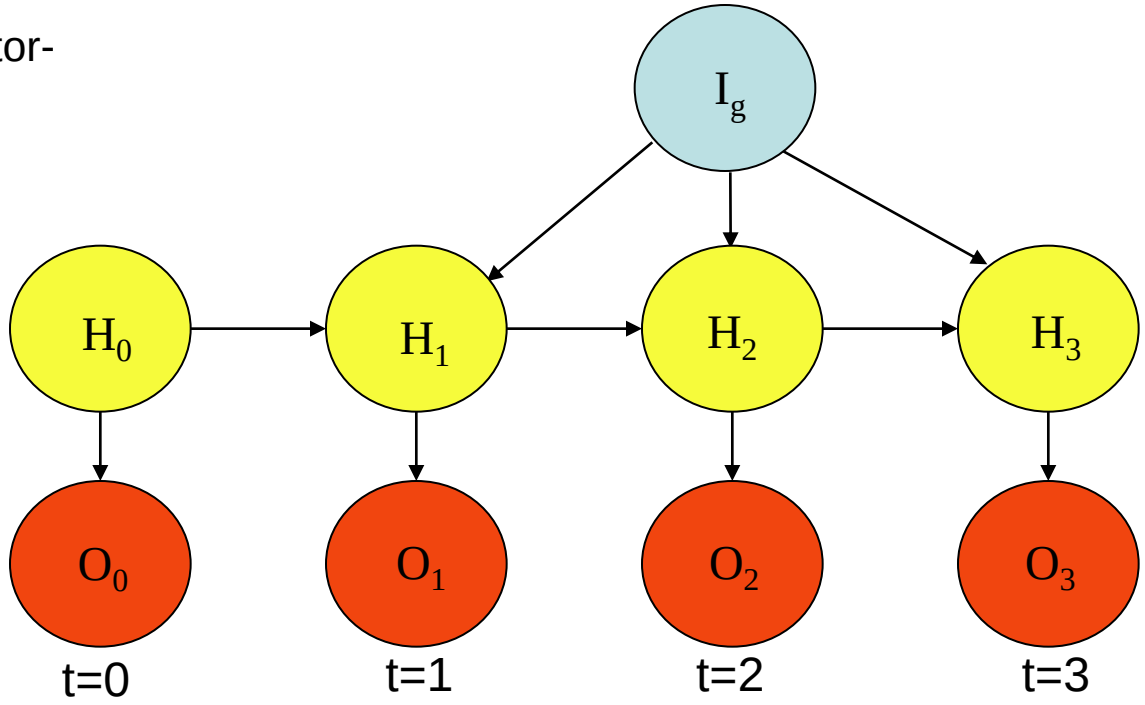
Input – Output Hidden Markov Model

Bengio and Frasconi, *NIPS* 1995

Input (Static transcription factor-gene interactions)

Hidden States Variables
(We constrain transitions between states to form a tree structure)

Output State Variables
(Gaussian distribution for expression values)



Log Likelihood

$$r(G|M) = \sum_{g \in G} \log \sum_{q \in Q} \prod_{t=1}^{n-1} f_{q(t)}(o_g(t)) \prod_{t=1}^{n-1} P(H_t = q(t) | H_{t-1} = q(t-1), I_g)$$

Sum over all genes

Sum over all paths *Q*

Product over all Gaussian emission density values on path

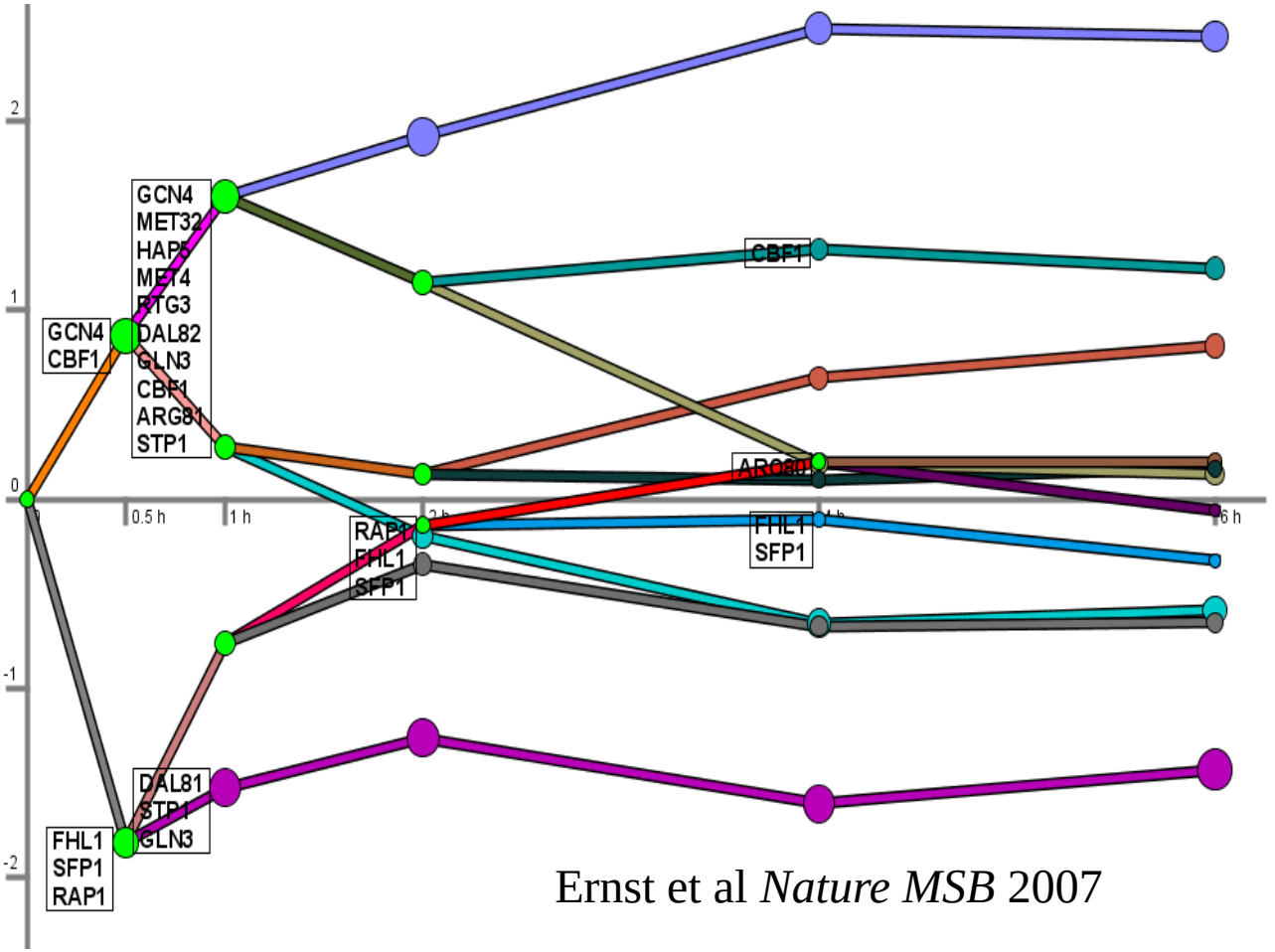
Product over all transition probabilities on path

Results: Yeast response pathways

Application to AA starvation in yeast with condition specific data

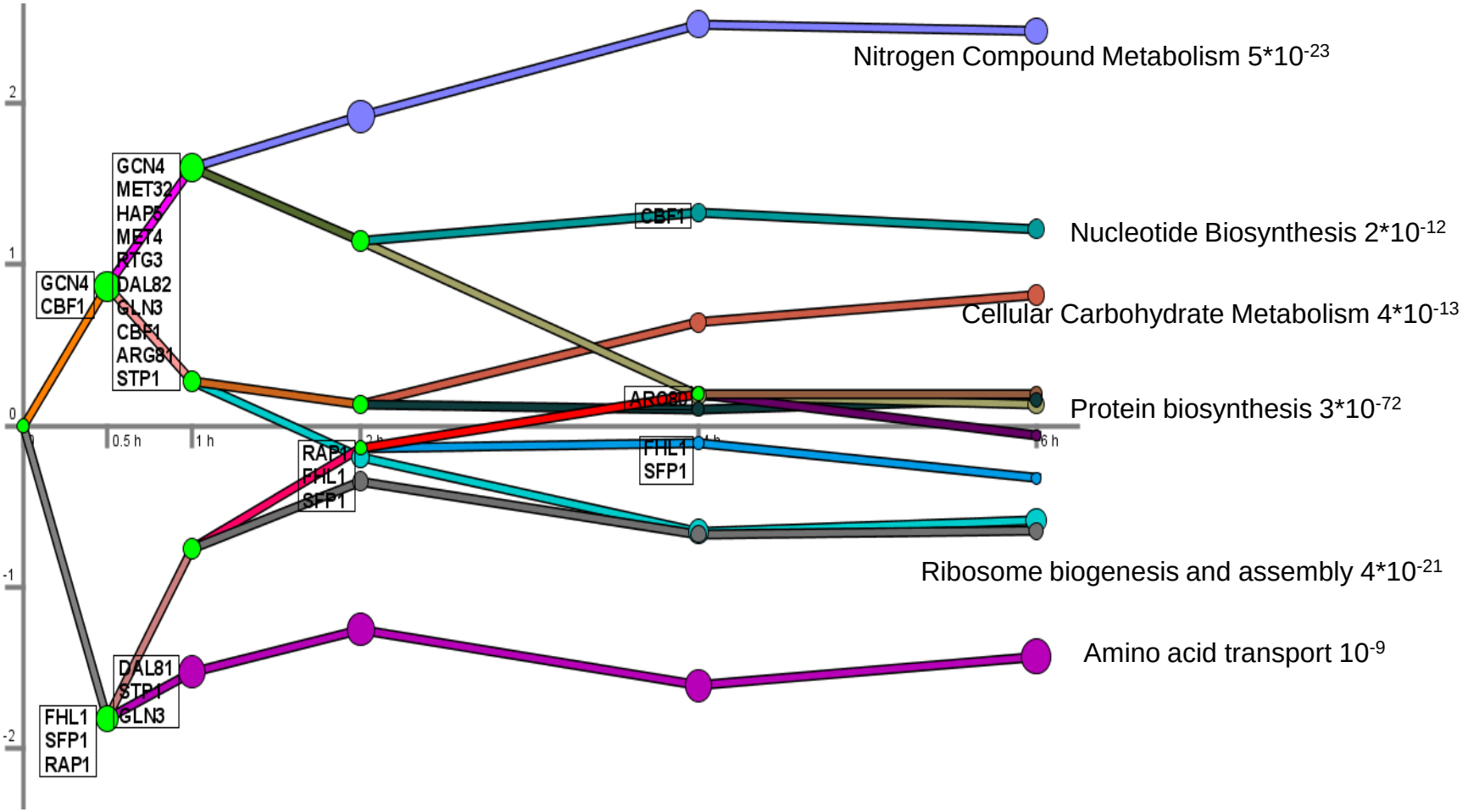
Expression data: Gasch et al Mol. Bio. Cell. 2000,

Chip-chip data: Harbison et al Nature 2004

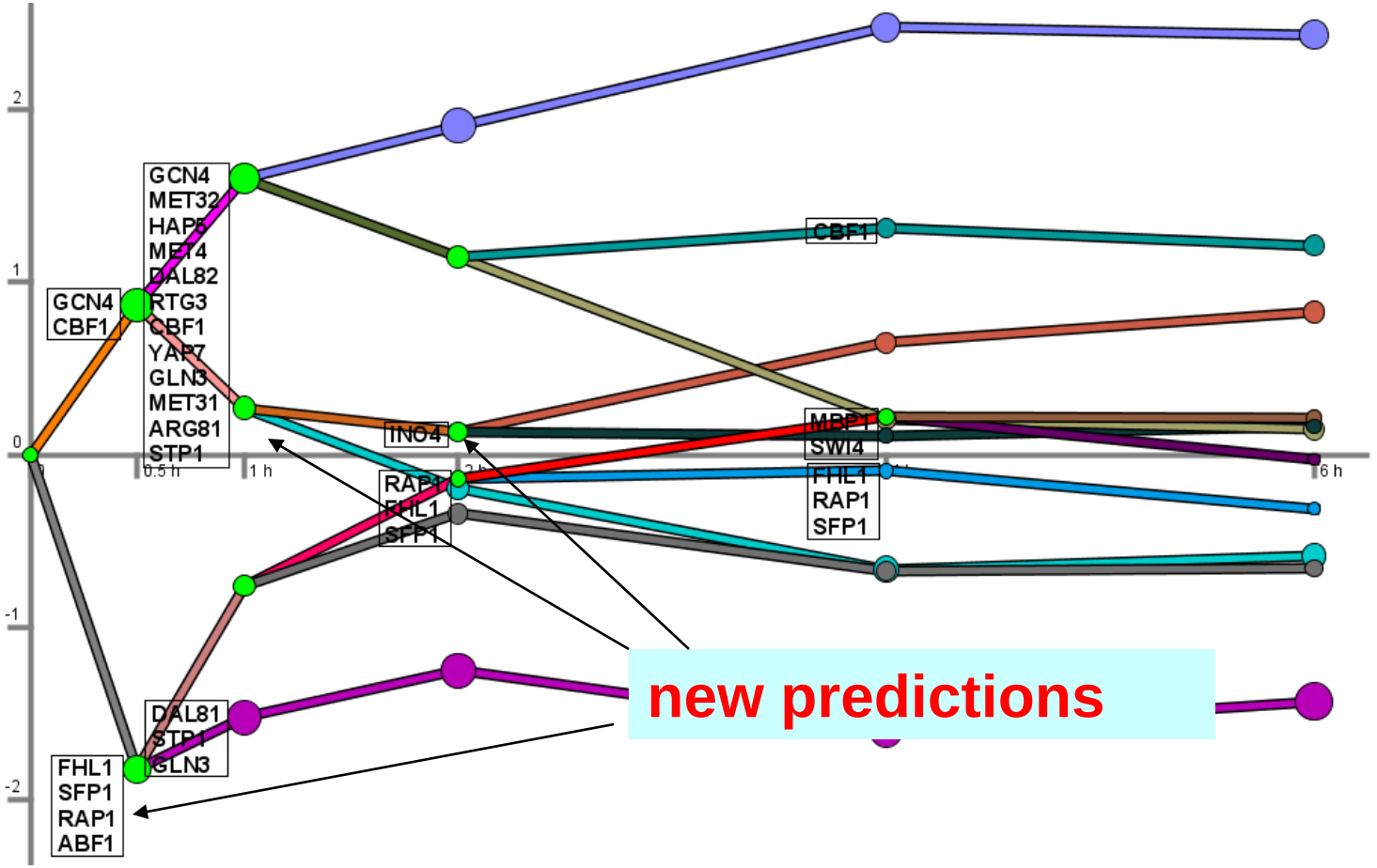


Ernst et al *Nature MSB* 2007

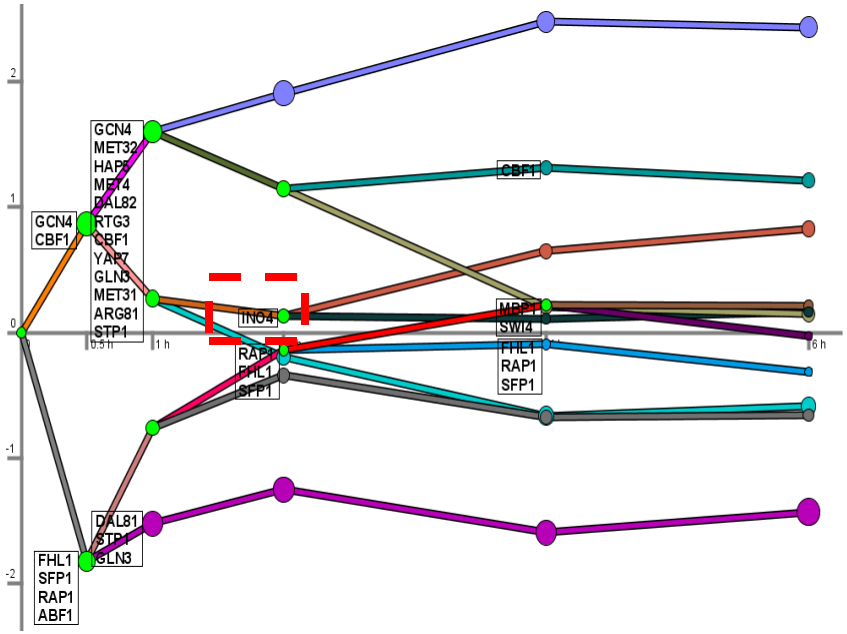
Application to AA starvation in yeast with condition specific data

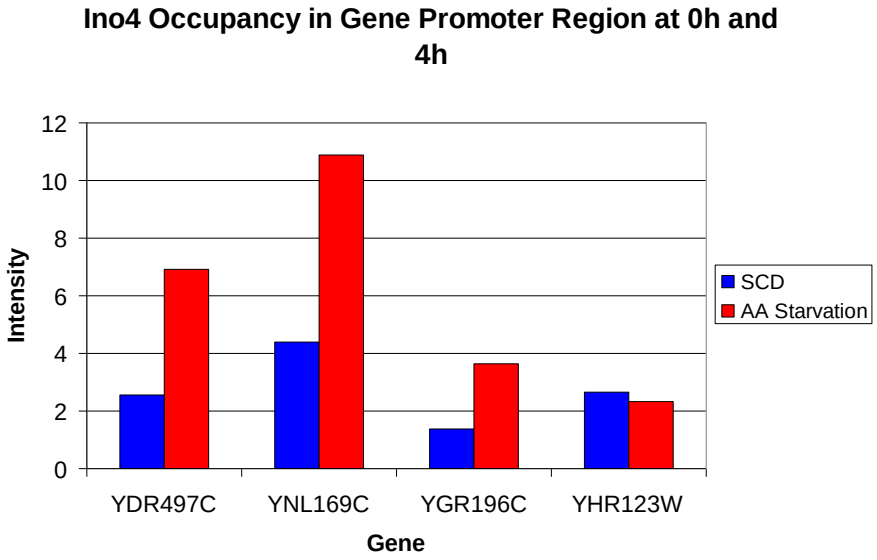


Application to AA starvation in yeast with general binding and motif data

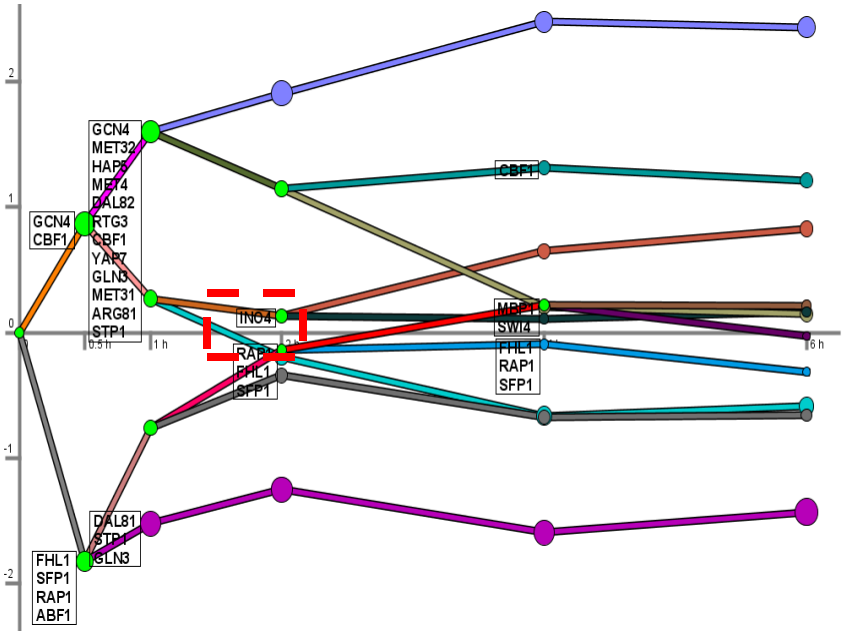


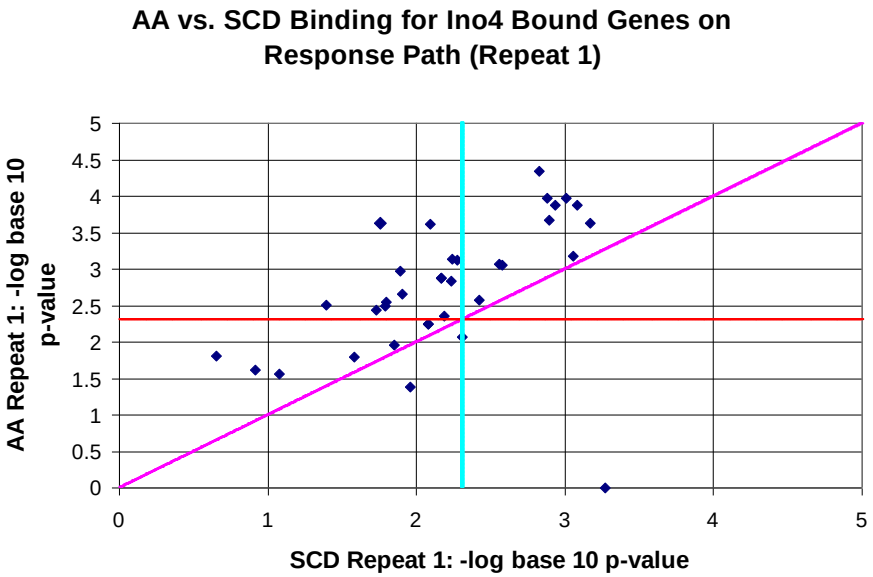
Validating predicted interactions for Ino4





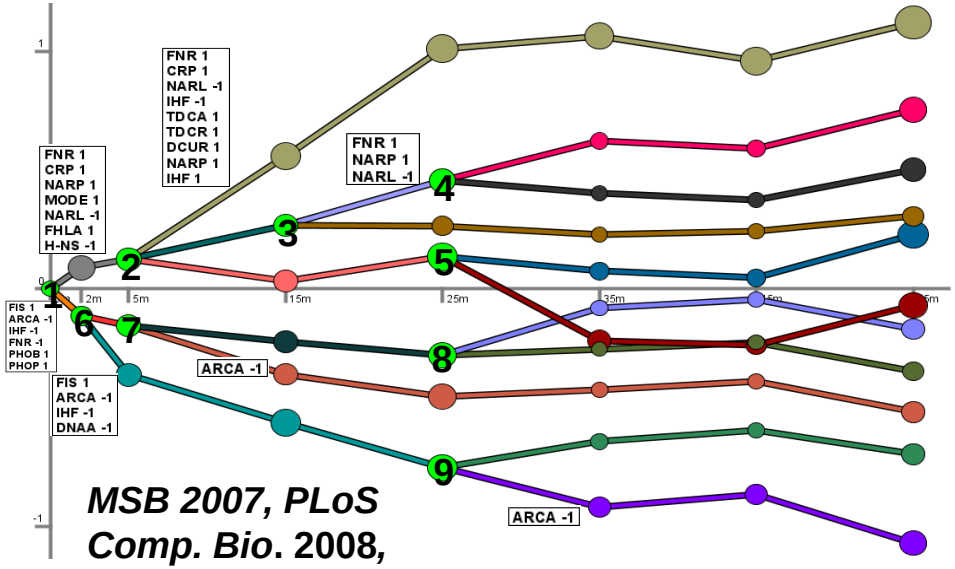
Validating predicted interactions for Ino4





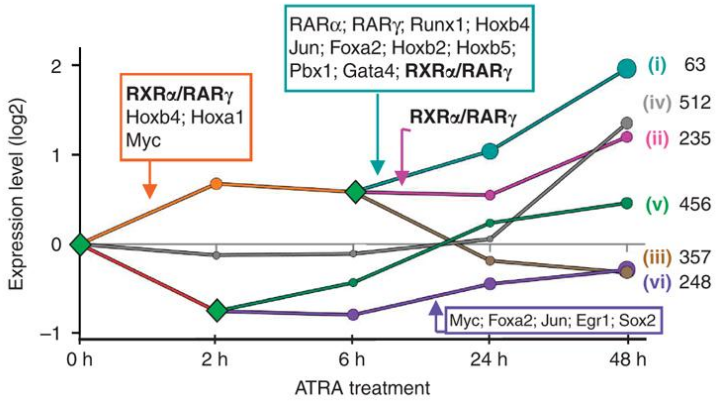
- Ino4 regulates phospholipid biosynthesis
- Many genes in the path are known lipid metabolism genes (GO p-value 6*10-5).
- May be connected to the need of membrane used for the autophagocytosis process which regulates equilibrium between proteins and the diminishing set of amino acids

Stress and hormone response



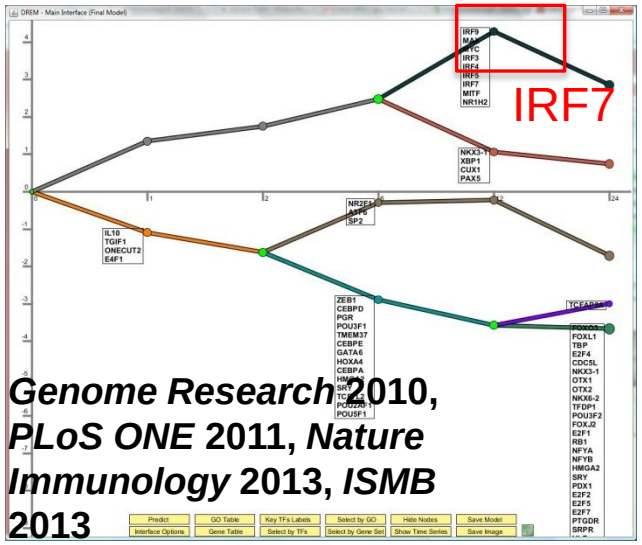
MSB 2007, PLoS Comp. Bio. 2008, *eLife* 2013

Stem cells differentiation



MSB 2011

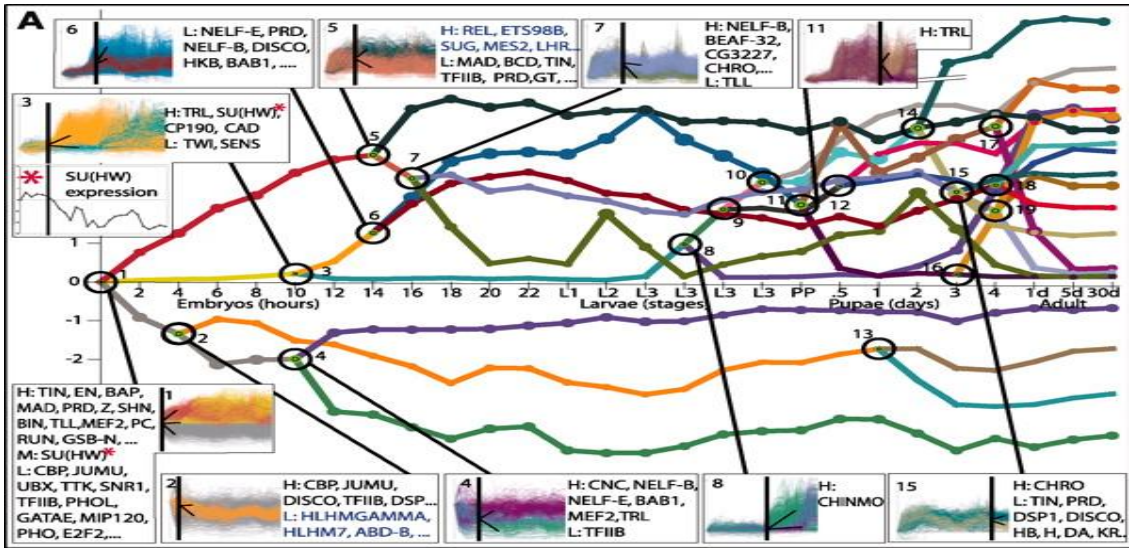
Immune response



Genome Research 2010, *PLoS ONE* 2011, *Nature Immunology* 2013, *ISMB* 2013

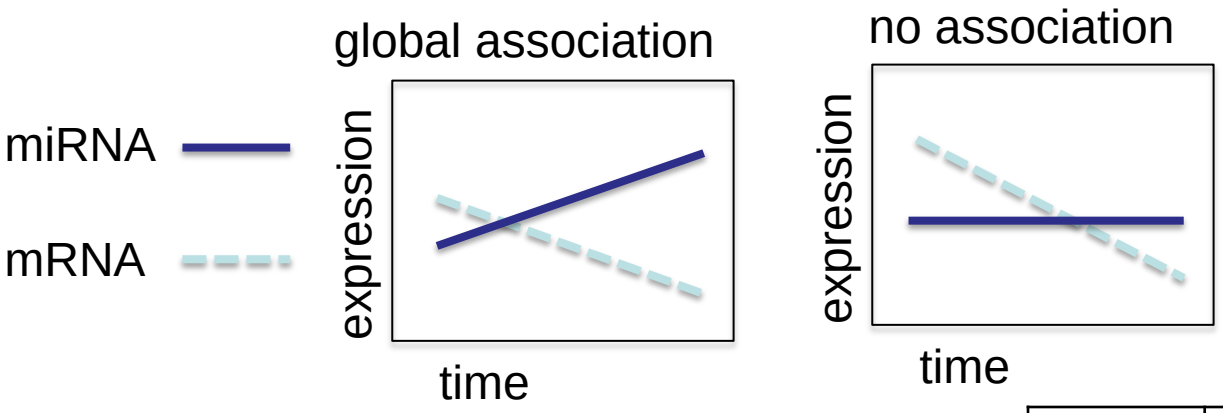
Fly development

Science 2010

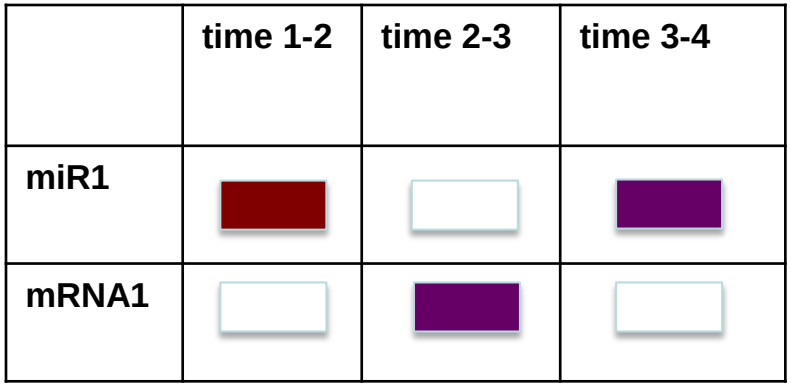


microRNAs

miRNA-target detection using expression data

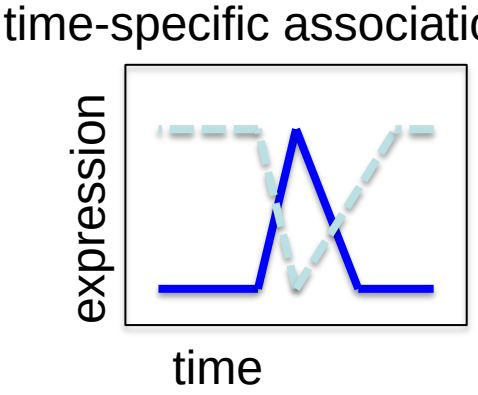


- Pearson-correlation (Cheng et al. 2008, Huang et al. 2011)
- Regression with graphical models, GenMiR++ (Huang et al. 2007)



Disadvantages

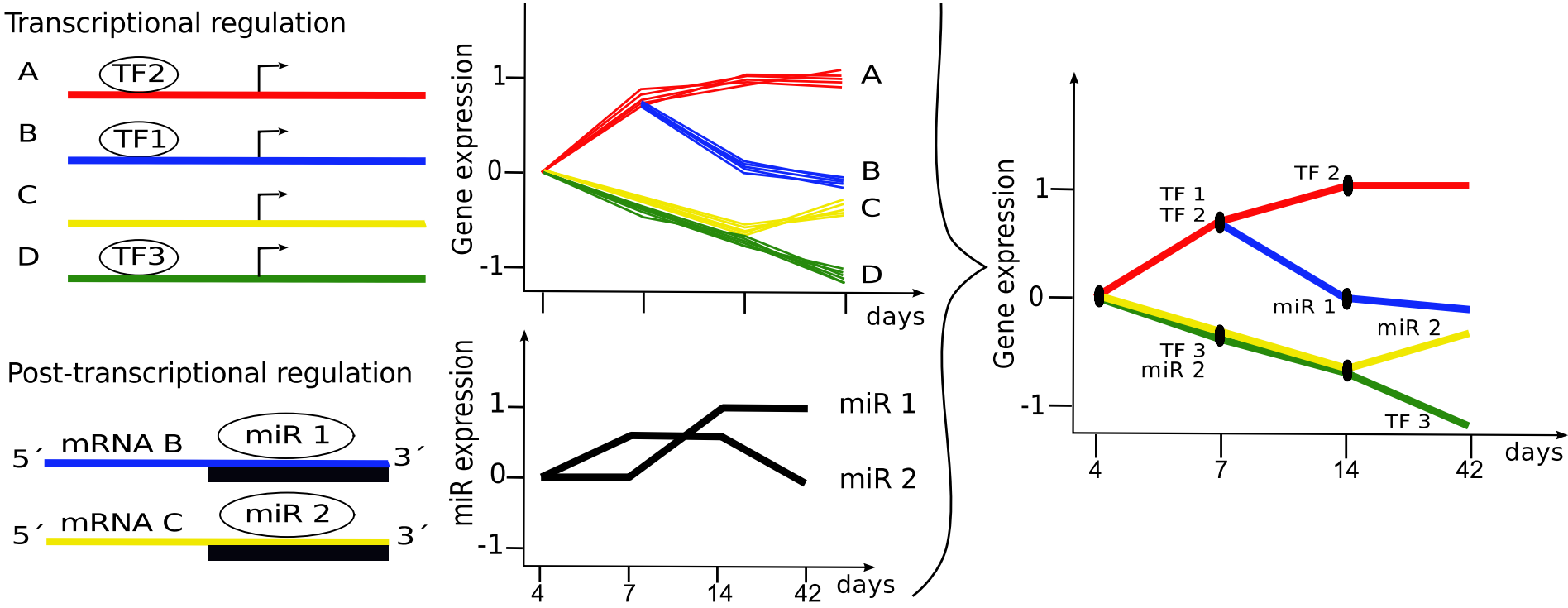
- methods often assume linear correlation
- time is not implicitely modeled
- no prediction of time-specific regulation
- joint modeling of TF regulation seldom possible



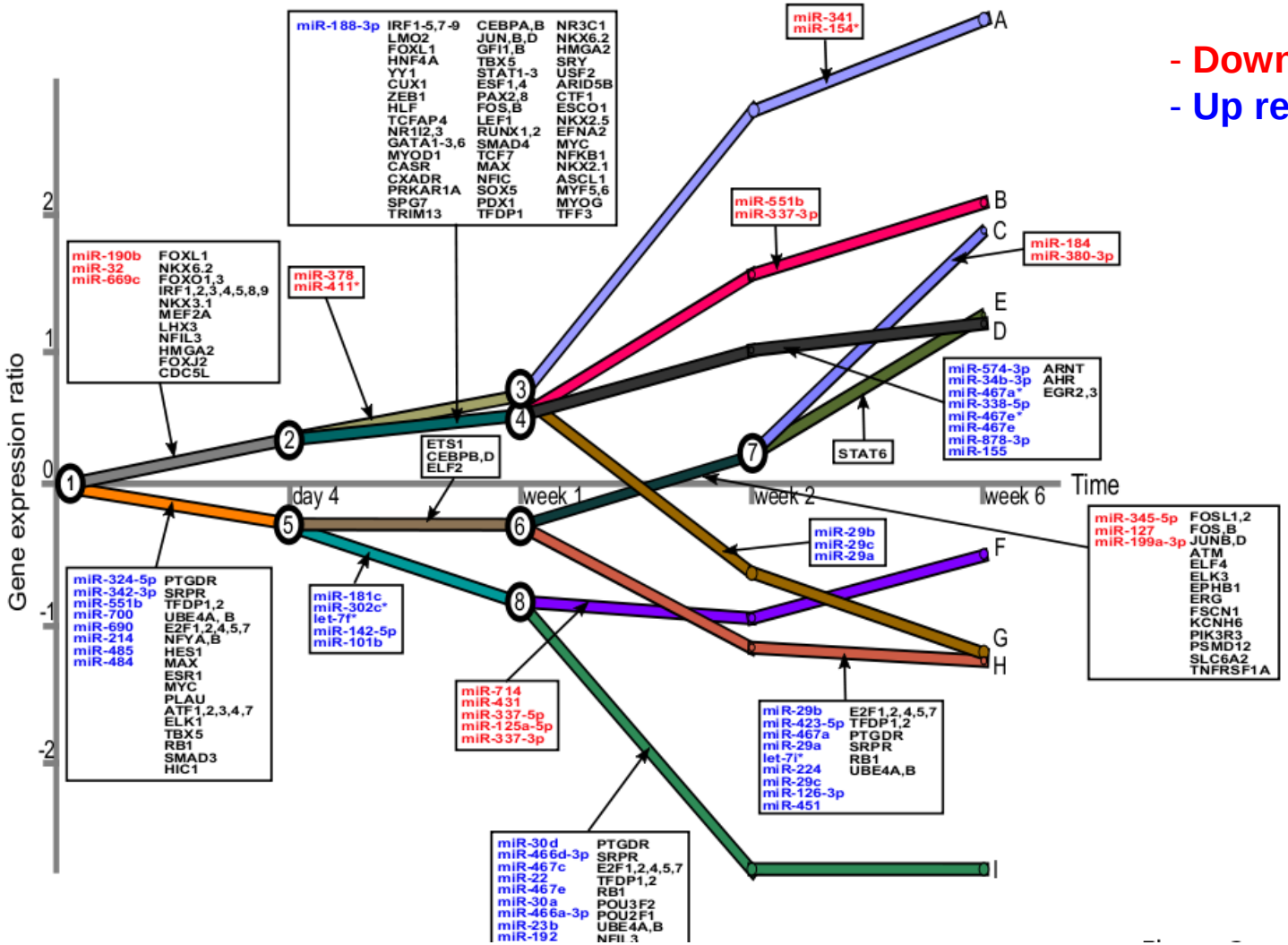
Incorporating miRNAs into DREM

- Incorporate miRNA expression ratios with logistic function to generate dynamic input map for miRNAs

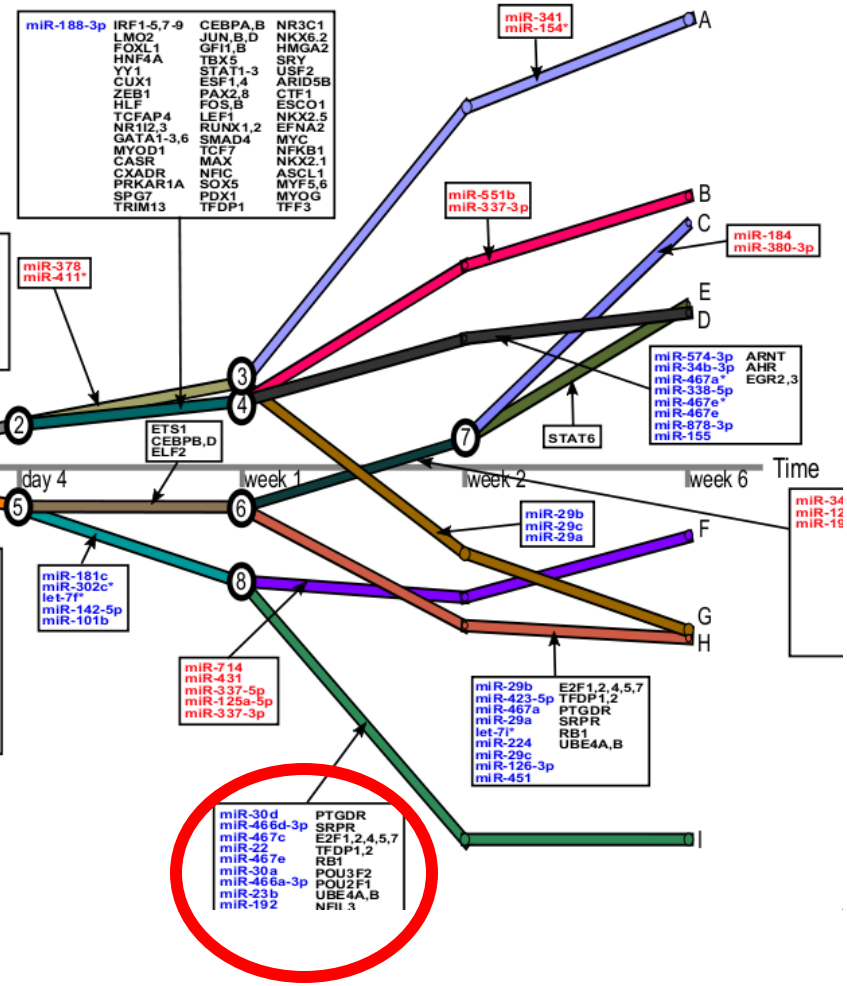
-Enforce positivity constraints for miRNAs coefficients in the logistic regression model (convex optimization, still global optimum)



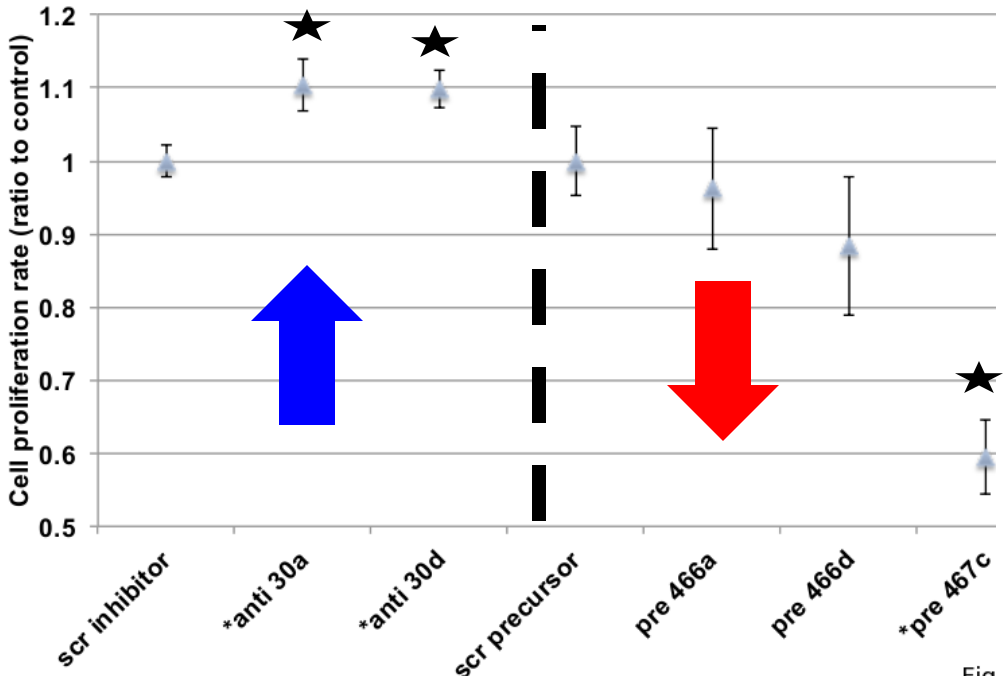
Dynamic regulaotry models for lung development in mice



Validating miRNAs controling Week 1 – Week 2 transition

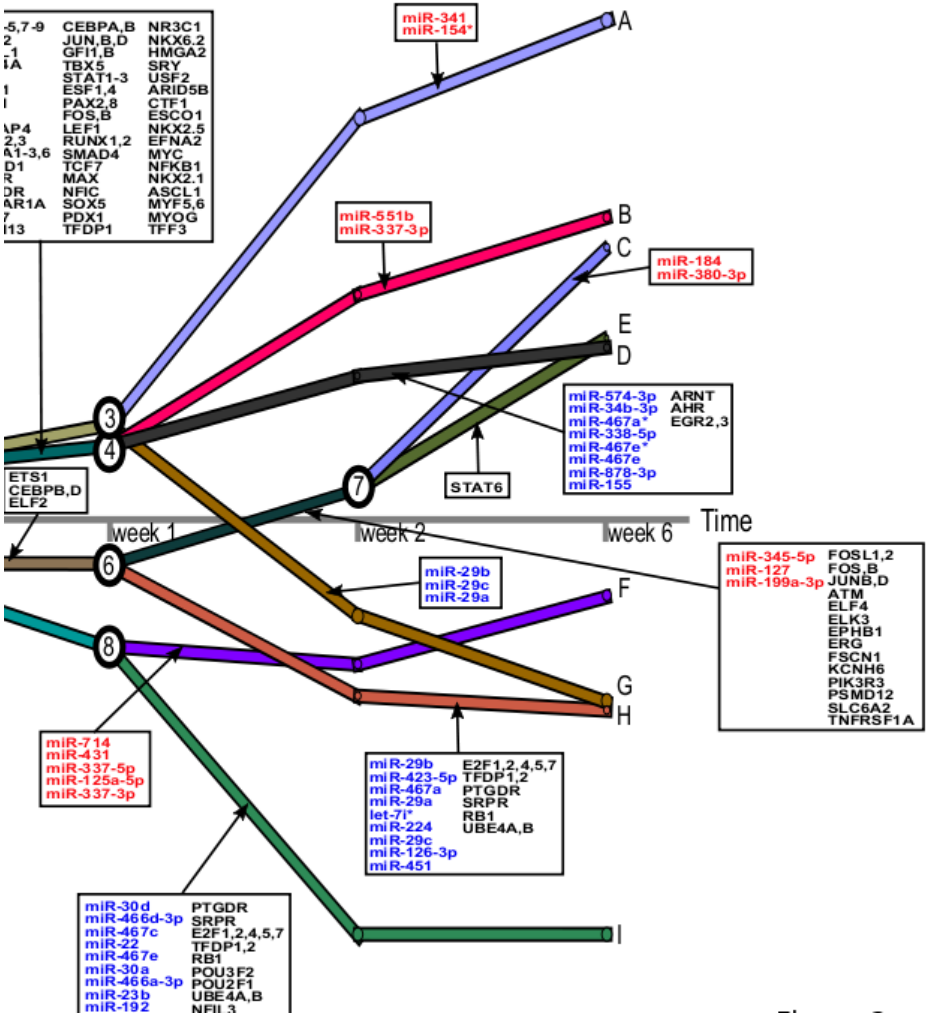


miRNA	sign. paths opposite direction	corrected enrichment p-values
miR-125a-5p	A,B,C+E	0.0108,0.00008,0.03872
miR-337-5p	B,C+E	0.0332, 0.00008
miR-467c	D	$< 10^{-6}$
miR-466a-3p	D	0.05152
miR-466d-3p	D	0.03904
miR-30d	H	0.01456
miR-30a	-	-



Fig

From development to disease

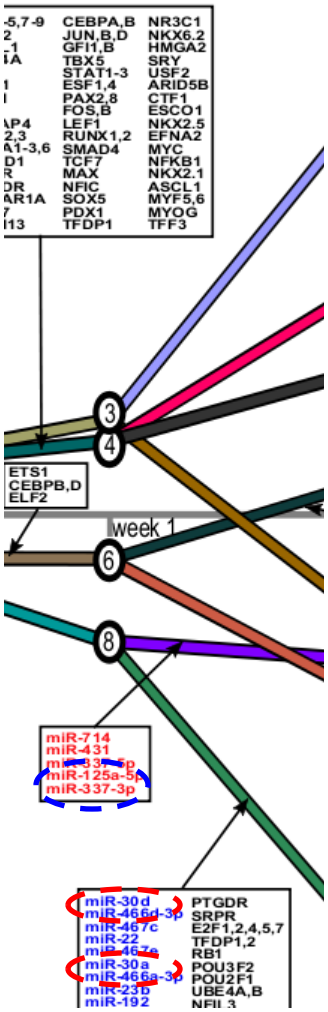


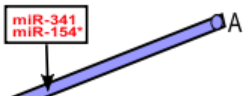
Analysis of significant miRNAs in patients with Idiopathic Pulmonary Fibrosis (IPF)

	10 control vs 10 IPF (TissueBank)	28 control vs 33 IPF (Tissue Bank)	142 control vs 162 IPF (LGRC)	bleomycin at 14d (3 replicates) Liu et al data J. Ex. Med. 2010
miR-125a	down	-	-	down
miR-30a	down	down	down	down
miR-30d	down	down	down	down
miR-467c	-	-	-	-
miR-337	-	up	up	-
miR-466a-3p	-	-	-	down
miR-466d-3p	-	-	-	down

Schulz et al *PNAS* 2013

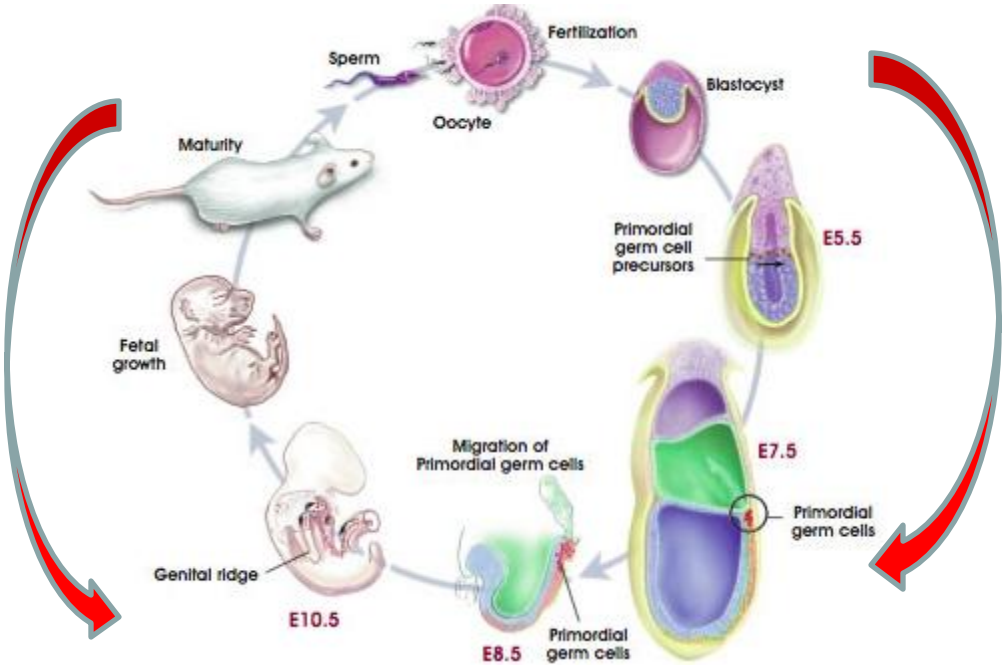
From development to disease





Analysis of significant miRNAs in patients with Idiopathic Pulmonary

Development



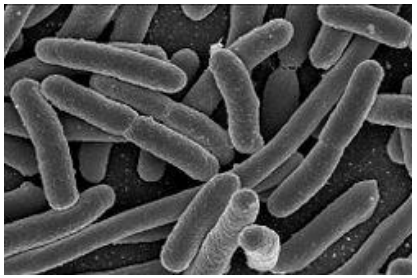
Disease

bleomycin at 14d (3 replicates) Liu et al data J. Ex. Med. 2010
down
down
down
-
-
down
down

Schulz et al *PNAS* 2013

E. coli

Escherichia coli Regulatory Network



- Comprehensive genome wide binding data is currently not available for most transcription factors
- Key transcription factor activate some genes and repress others
  - Direction of interaction should also be part of the input to DREM
- 25% of genes have at least one known regulator based on curated small scale experiments
  - No confirmed negative data
- Expression and motif data also available

E. coli Image from Wikipedia

Using unlabeled data helps supervised learning

Consider setting:

- Set *X* of instances drawn from unknown distribution *P(X)*
- Wish to learn target function *f: X → Y*

Given:

- iid labeled examples $L = \{\langle x_1, y_1 \rangle \ldots \langle x_m, y_m \rangle\}$
- iid unlabeled examples $U = \{x_{m+1}, \ldots x_{m+n}\}$

Determine:

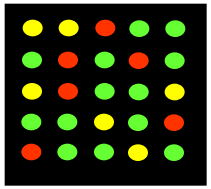
$$\hat{f} \leftarrow \arg\min_{h \in H} \Pr_{x \in P(X)} [h(x) \neq f(x)]$$

SEmi-supervised REgulatory Network Discoverer (SEREND)

	Motif	Exp 1	Exp 2	…	Exp p	Label
Gene 1	8.0	1.2	-0.5	…	0.4	activated
Gene 2	6.2	-0.4	1.0	…	2.0	repressed
Gene 3	7.0	-0.8	1.2		3.2	unknown
…		…	…	…	…	…
Gene N	2.2	0.4	1.4	…	-1.4	unknown

Binary Logistic Regression Classifier

3-Way Logistic Regression Classifier



***P**(regulated)*

***P**(activated)+**P**(repressed)*

"Meta" Classifier

Binary Logistic Regression Classifier

***P**(regulated)*

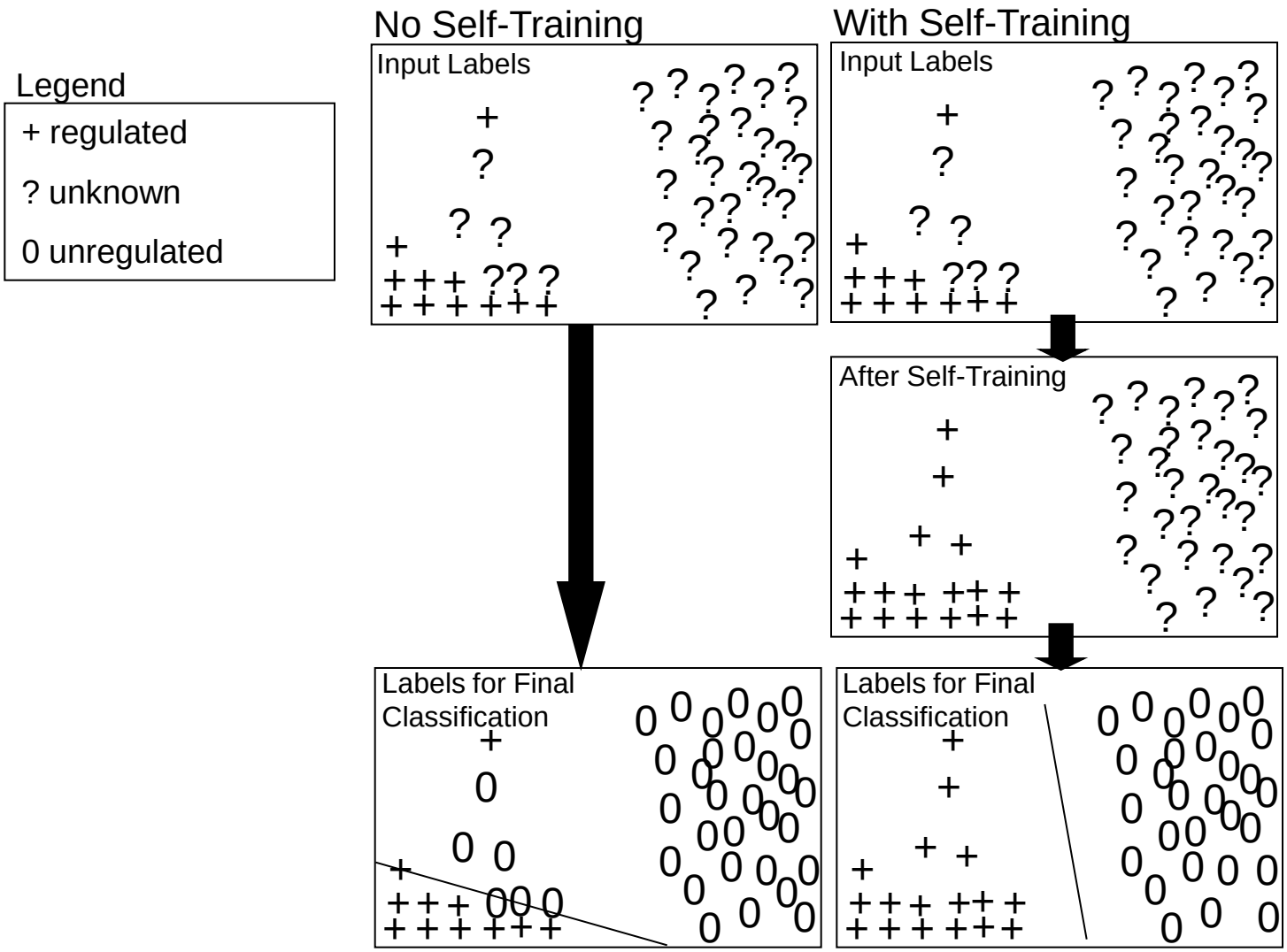
Self-training rule to change 'unknown' labels to either activated or repressed

***P**(regulated) > 2 x (#activated + #repressed)/N*

Ernst et al., PLoS Comput Biol, 2008

Expression Data from (Faith et al, PloS Biology 2007); Curated Interactions from EcoCyc; Motifs from Regulon DB

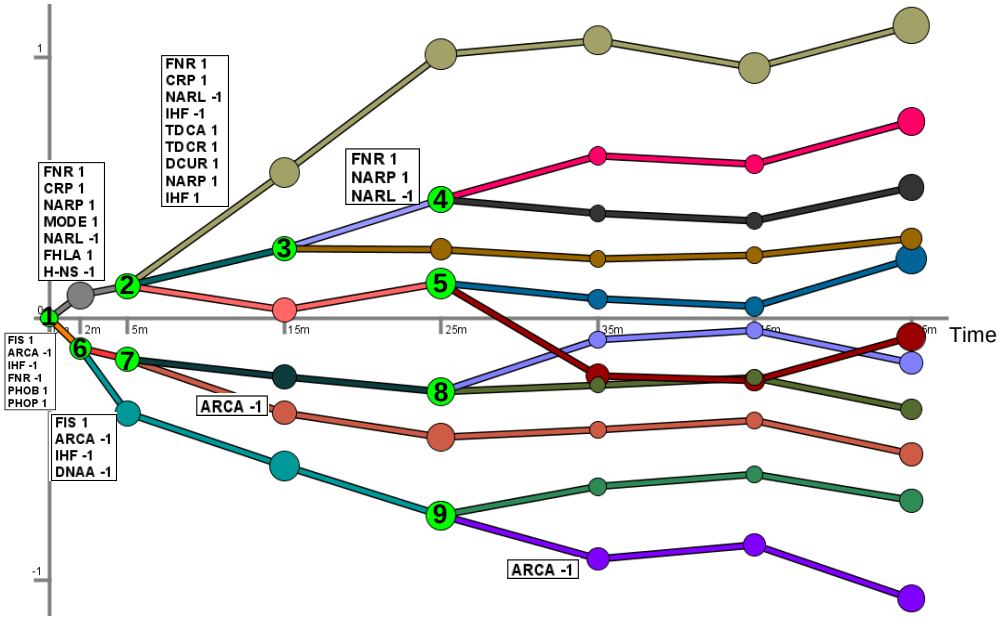
The Self-Training Step

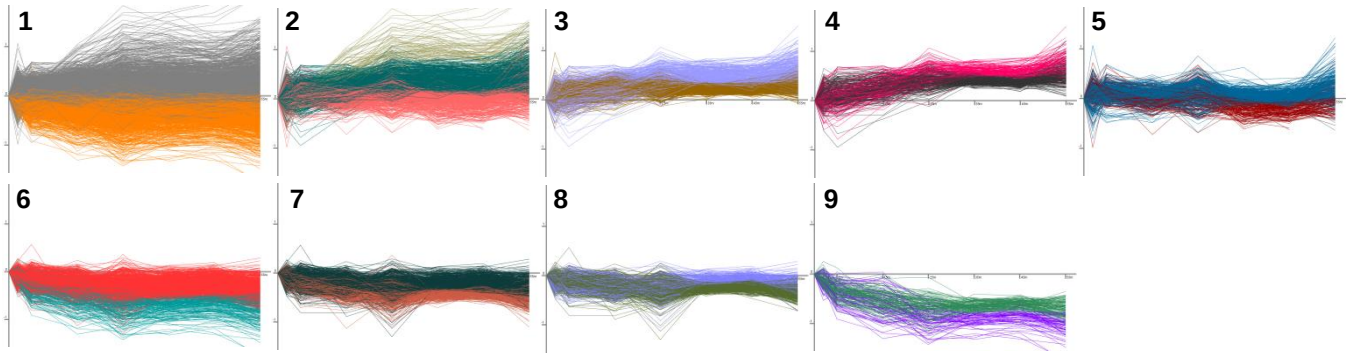


Aerobic-anaerobic shift response in *E. coli*

Expression Level
(log base 10)

1 activator;
-1 repressor





Beg et al *PNAS* 2007, Ernst et al *Plos Comp. Bio.* 2008

DREM is useful, but several questions remain …

Who controls the master regulators?