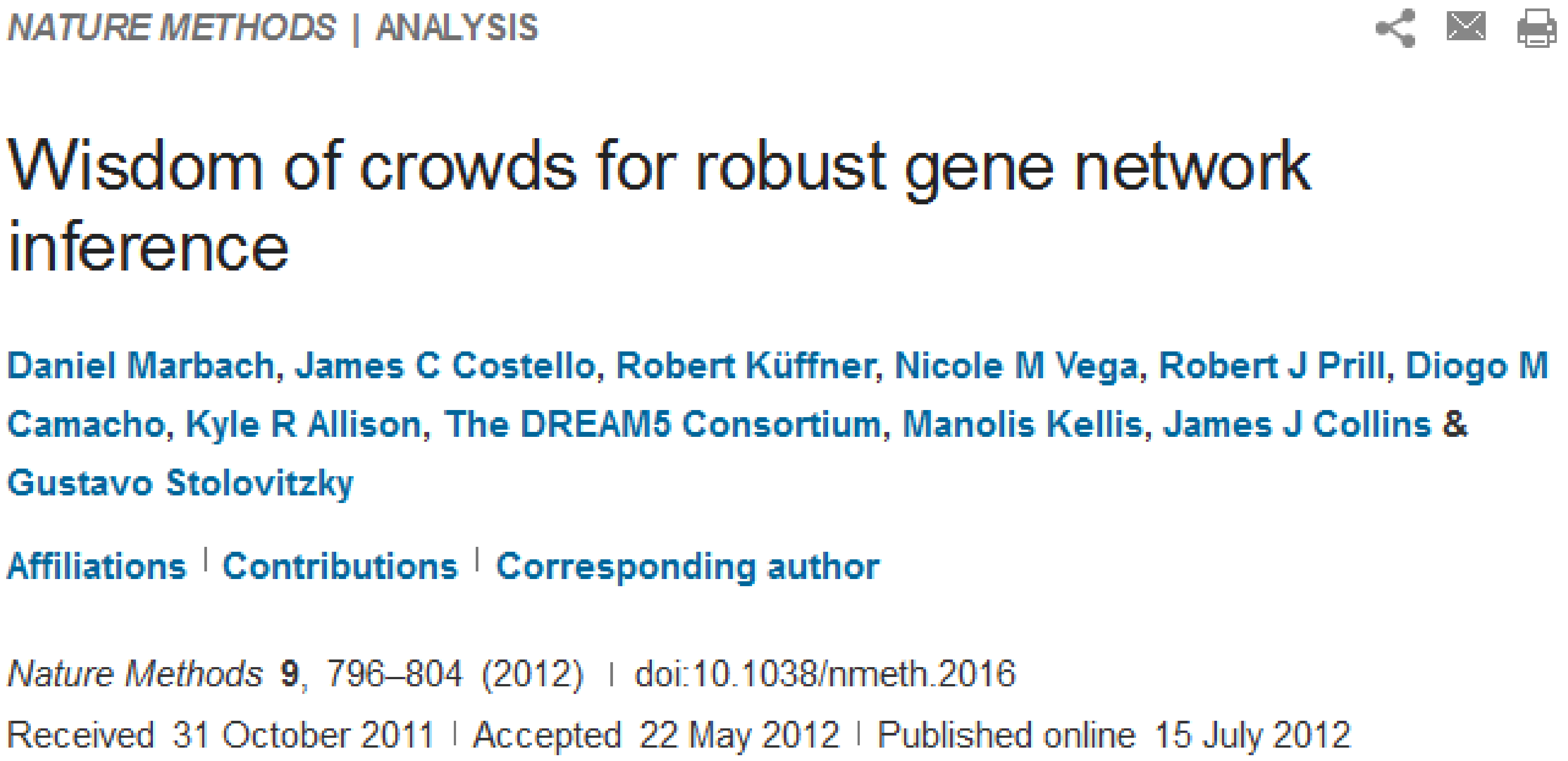
- L12 - Introduction to Protein Structure; Structure Comparison & Classification
- L13 - Predicting protein structure
- L14 - Predicting protein interactions
- L15 - Gene Regulatory Networks
- L16 - Protein Interaction Networks
- L17 - Computable Network Models



NATURE METHODS | ANALYSIS

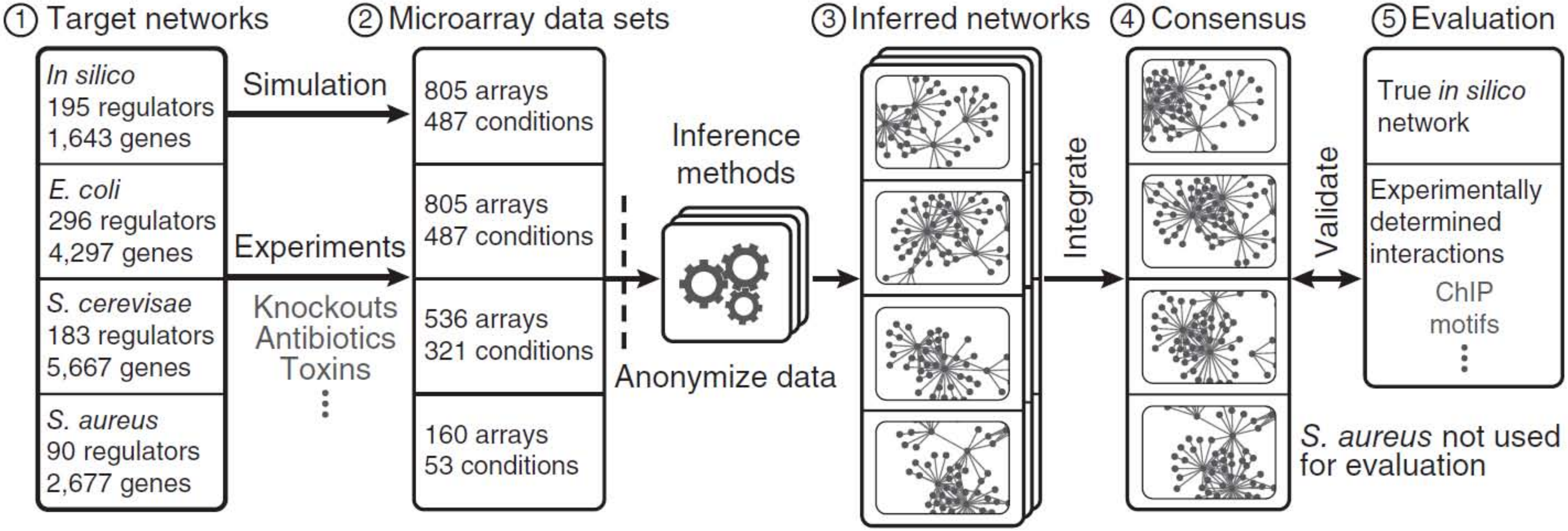
Wisdom of crowds for robust gene network inference

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, The DREAM5 Consortium, Manolis Kellis, James J Collins & Gustavo Stolovitzky

Affiliations | Contributions | Corresponding author

Nature Methods **9**, 796–804 (2012) | doi:10.1038/nmeth.2016

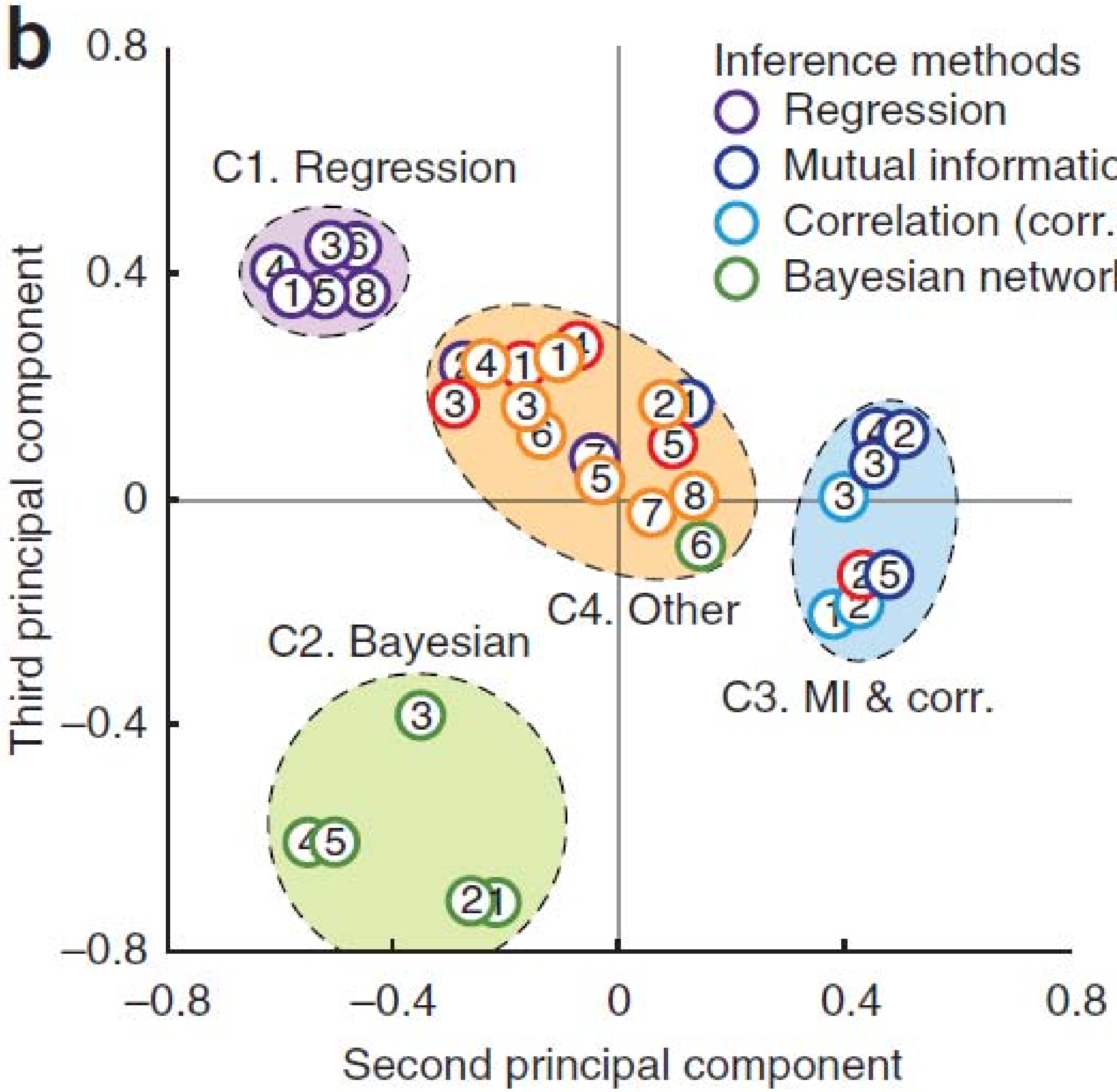
Received 31 October 2011 | Accepted 22 May 2012 | Published online 15 July 2012



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

Wisdom of crowds for robust gene network inference

Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016

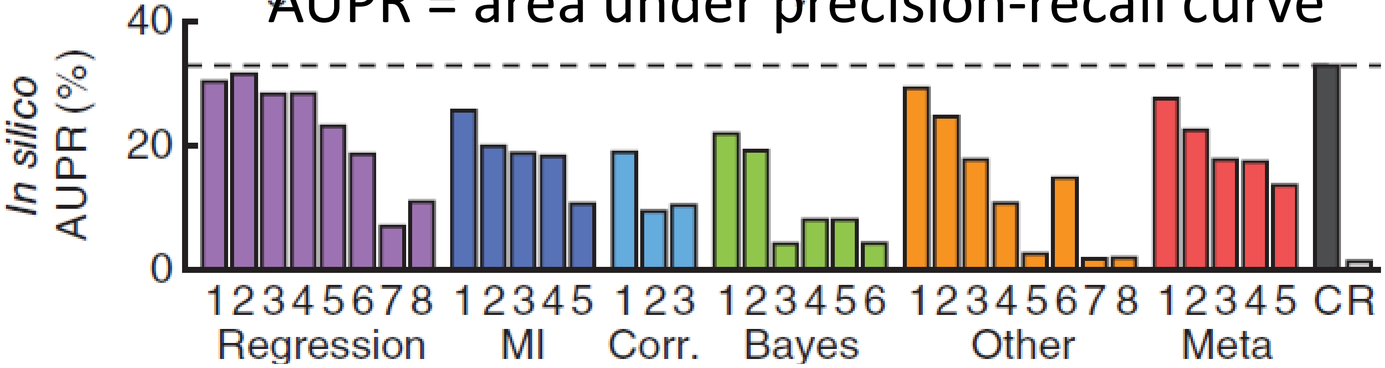


Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

Wisdom of crowds for robust gene network inference

Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016

AUPR = area under precision-recall curve



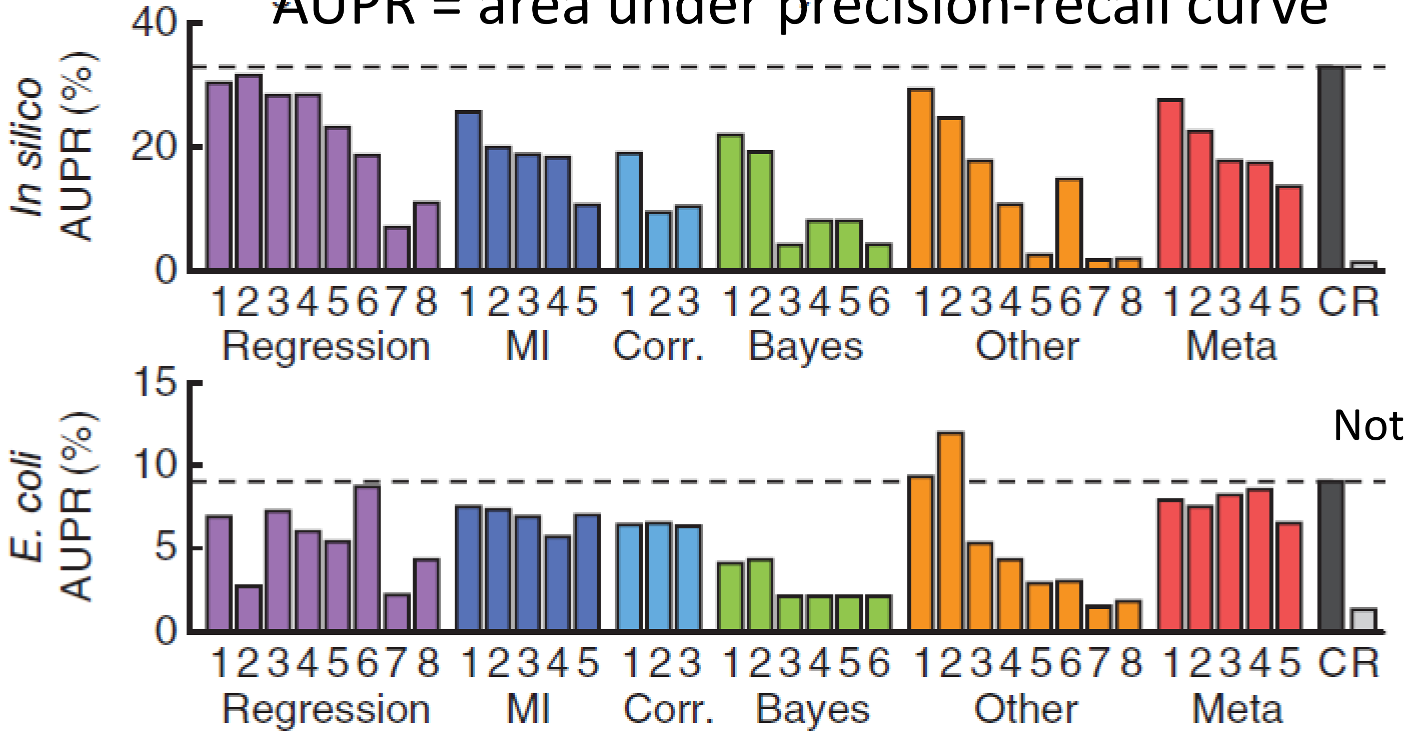
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

Area under precision-recall curve

Wisdom of crowds for robust gene network inference
Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016

AUPR = area under precision-recall curve

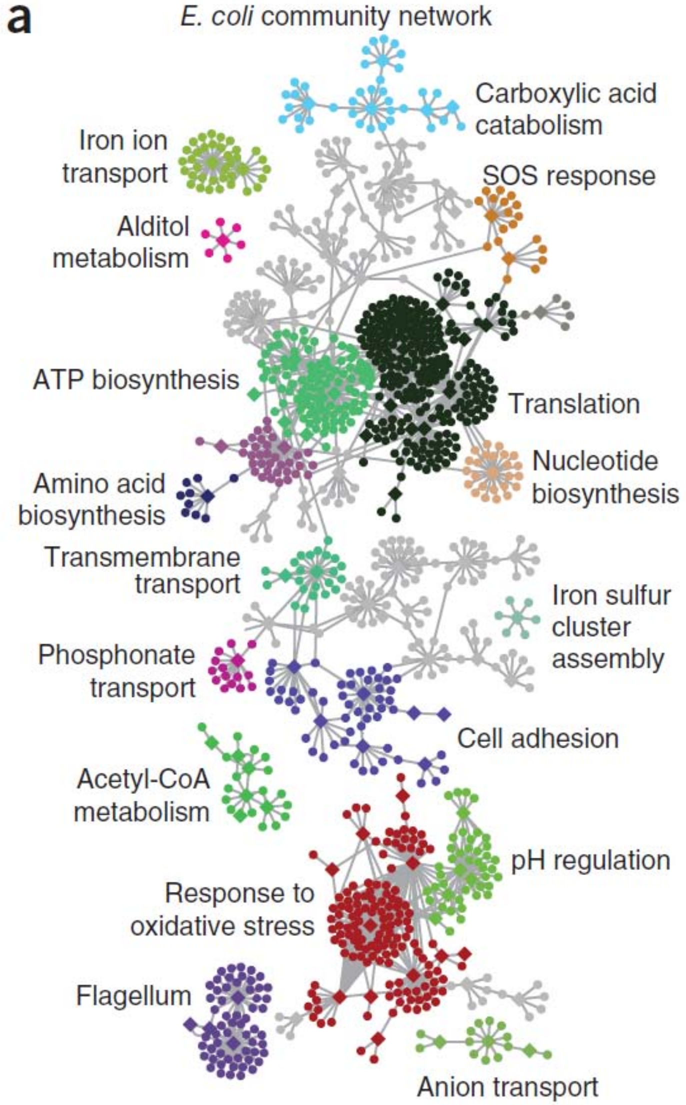
Area under precision-recall curve



Note change of scale!

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

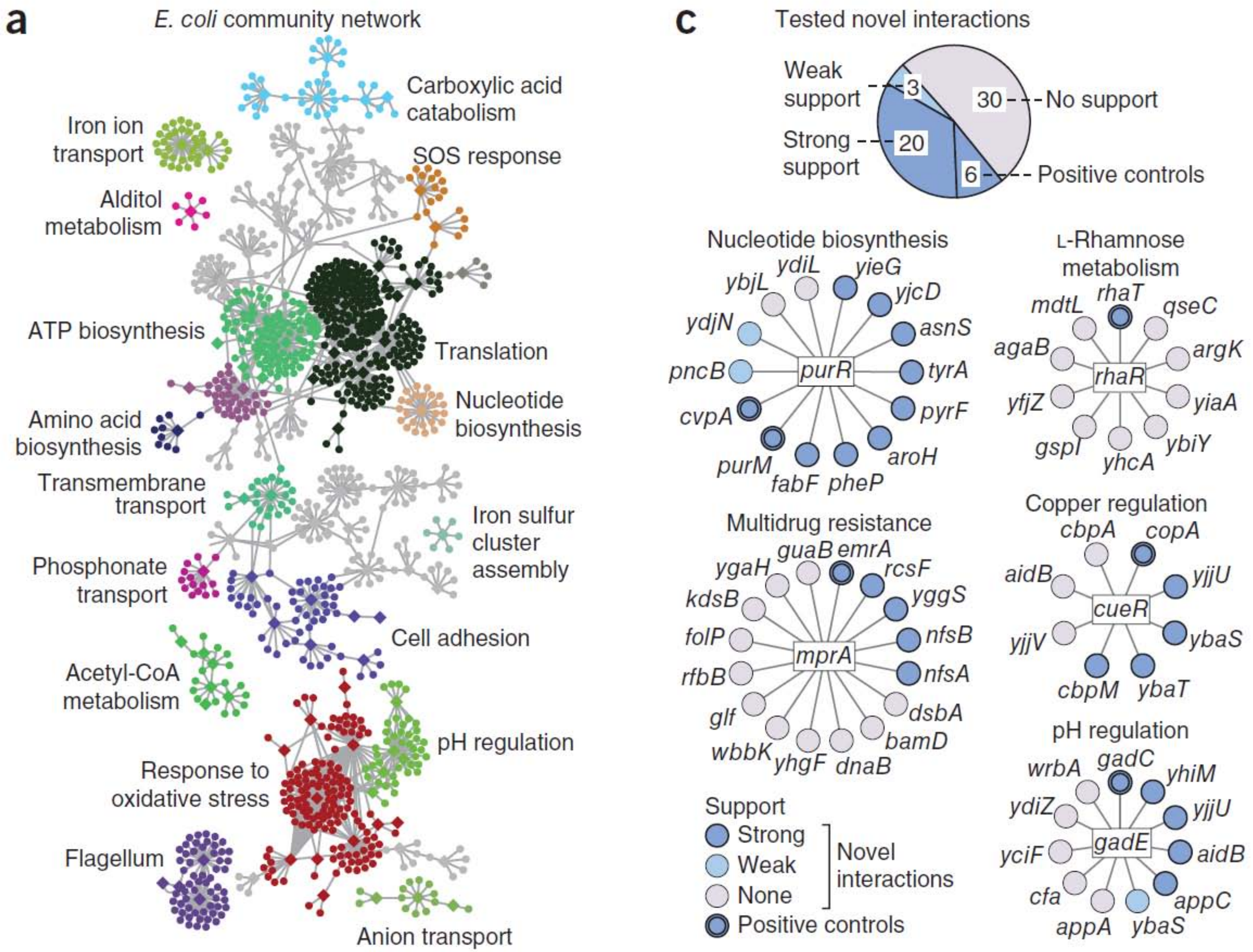
Wisdom of crowds for robust gene network inference
Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

Wisdom of crowds for robust gene network inference

Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016



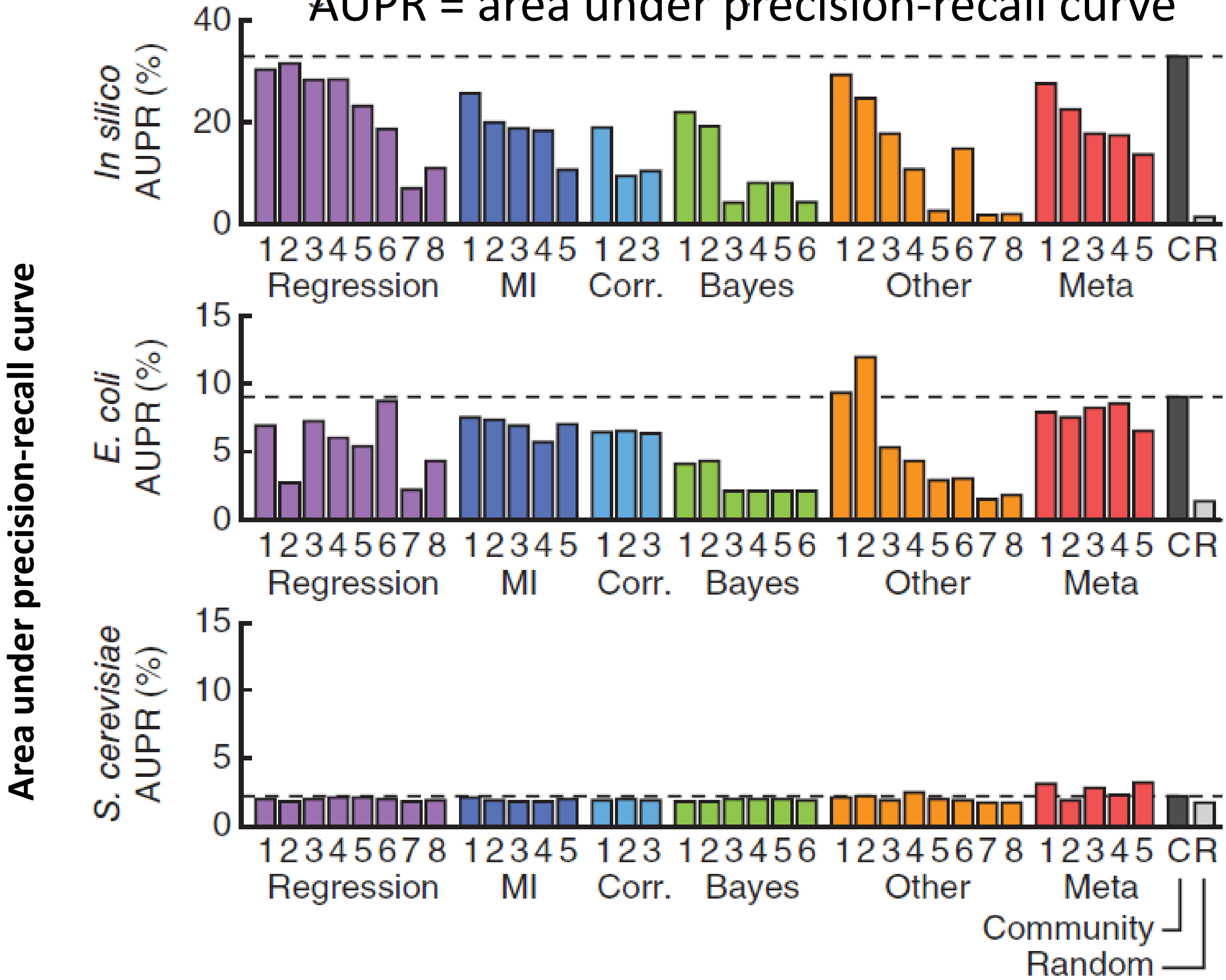
Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

Wisdom of crowds for robust gene network inference

Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016

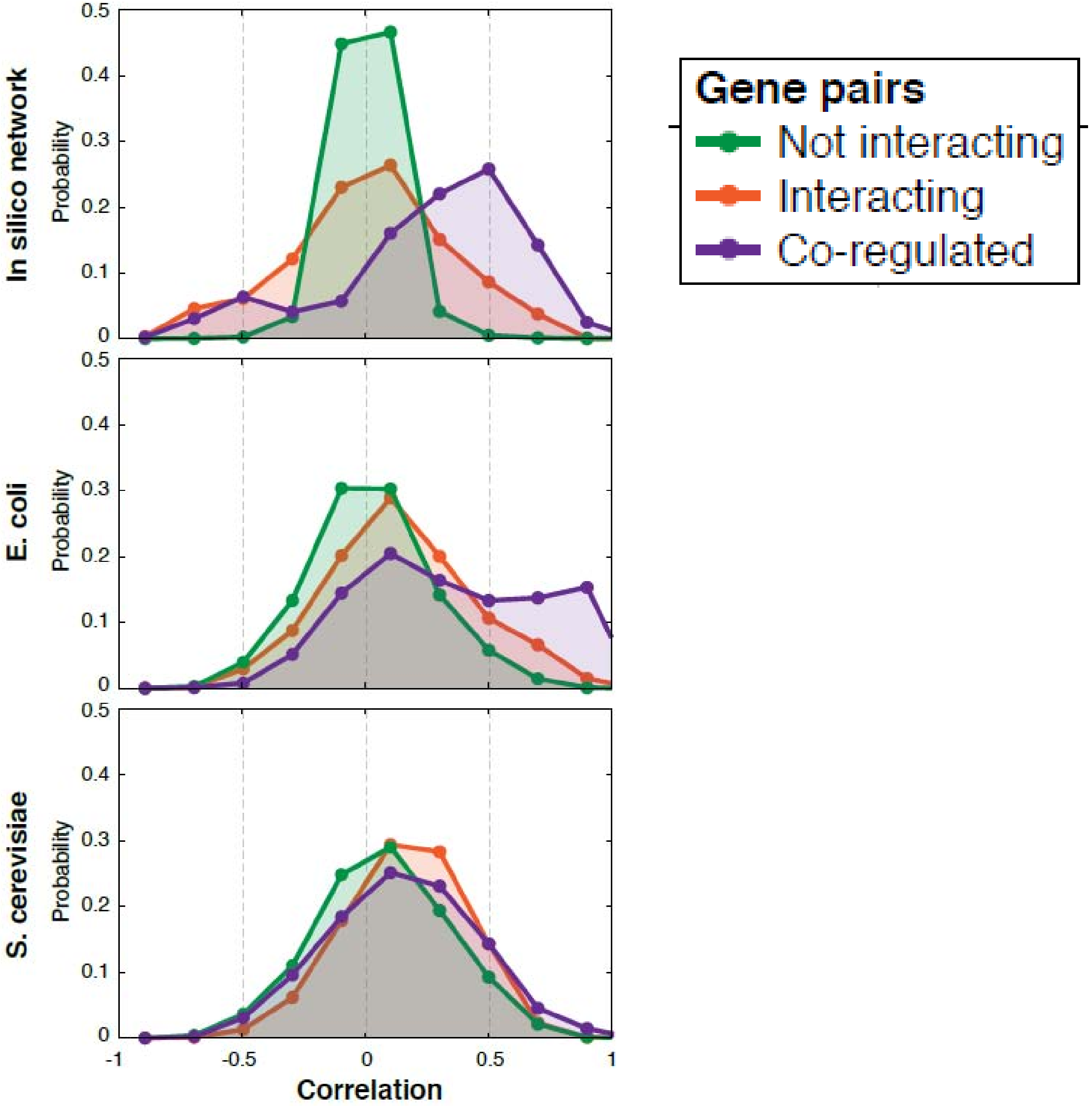
AUPR = area under precision-recall curve



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

Wisdom of crowds for robust gene network inference

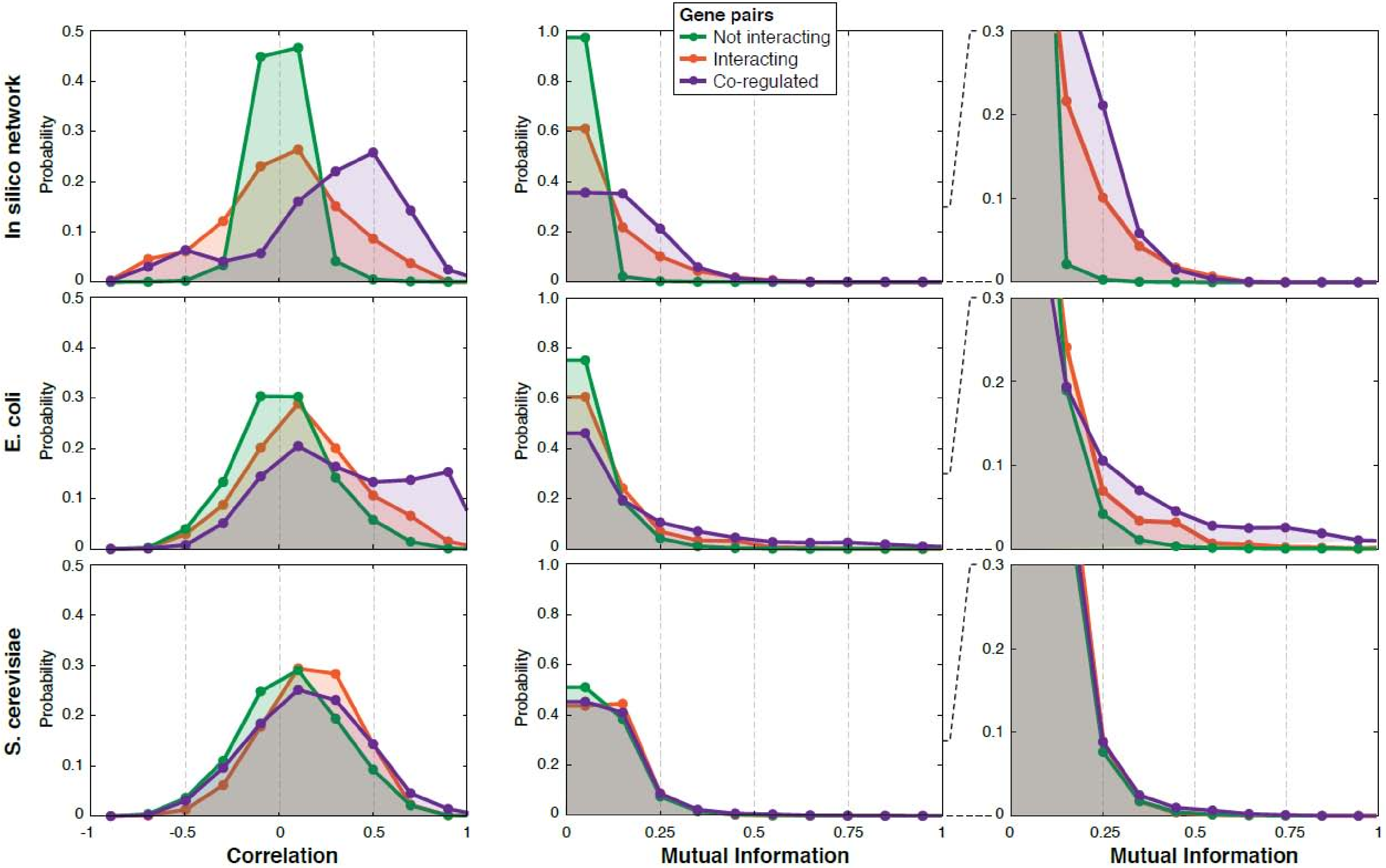
Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

Wisdom of crowds for robust gene network inference

Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

Wisdom of crowds for robust gene network inference

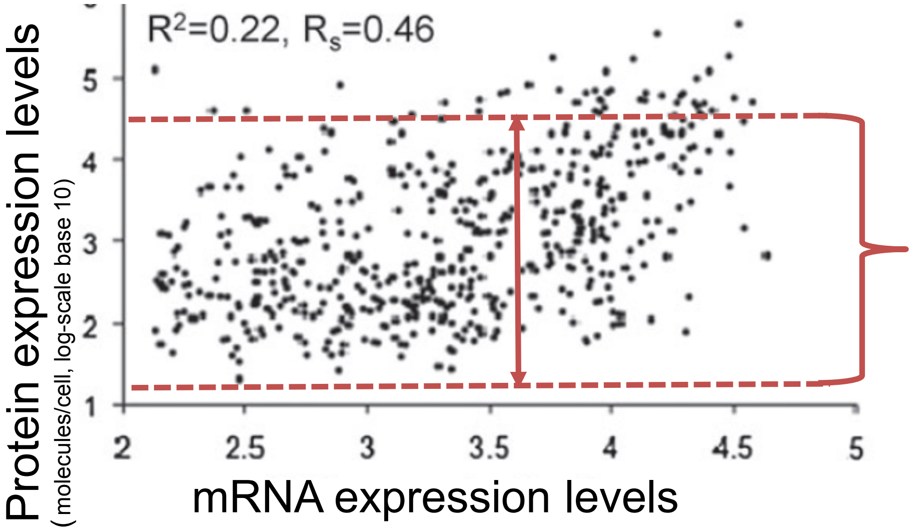
Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016

Thoughts on Gene Expression Data

- Useful for classification and clustering
- Not sufficient for reconstructing regulatory networks in yeast
- Can we infer levels of proteins from gene expression?

Approach

mRNA levels do not predict protein levels

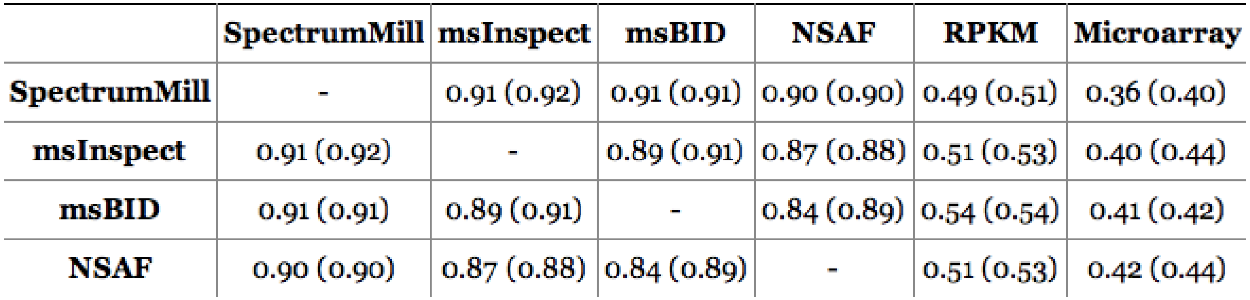


1,000 fold range of protein concentrations

Source: de Sousa Abreu, Raquel, Luiz O. Penalva, et al. "Global Signatures of Protein and mRNA Expression Levels." *Molecular Biosystems* 5, no. 12 (2009): 1512-26.

(arbitrary units, log-scale base 10)

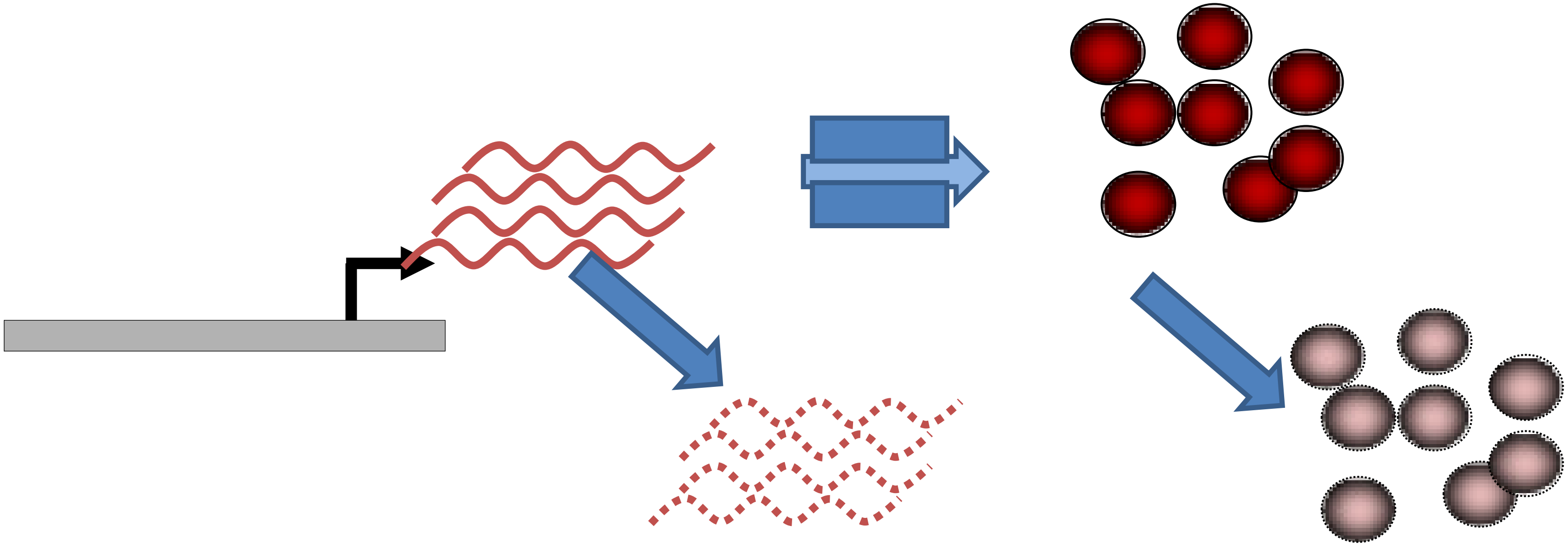
Raquel de Sousa Abreu, Luiz Penalva, Edward Marcotte and Christine Vogel, *Mol. BioSyst.*, 2009 DOI: 10.1039/b908315d

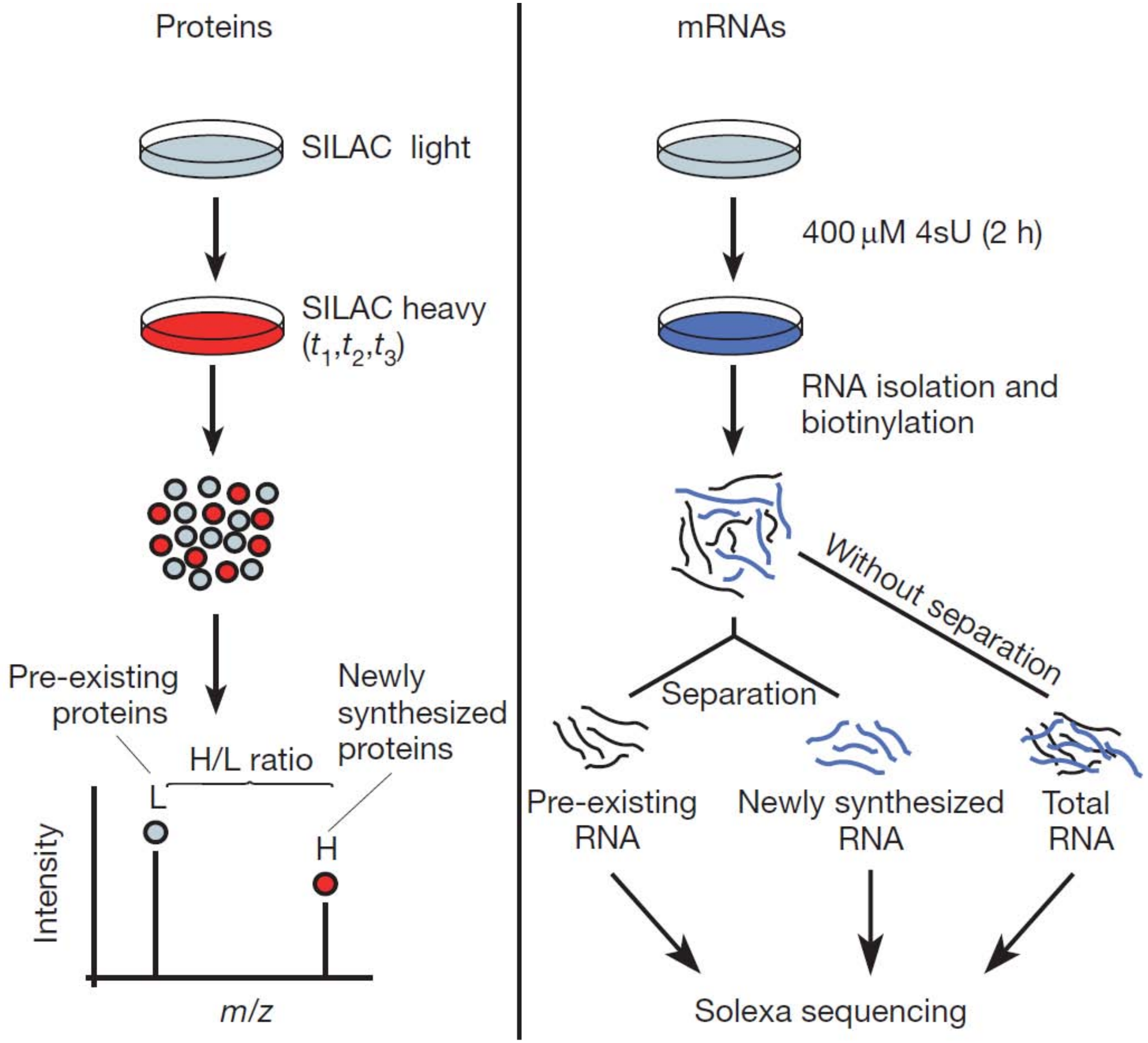


	SpectrumMill	msInspect	msBID	NSAF	RPKM	Microarray
SpectrumMill	-	0.91 (0.92)	0.91 (0.91)	0.90 (0.90)	0.49 (0.51)	0.36 (0.40)
msInspect	0.91 (0.92)	-	0.89 (0.91)	0.87 (0.88)	0.51 (0.53)	0.40 (0.44)
msBID	0.91 (0.91)	0.89 (0.91)	-	0.84 (0.89)	0.54 (0.54)	0.41 (0.42)
NSAF	0.90 (0.90)	0.87 (0.88)	0.84 (0.89)	-	0.51 (0.53)	0.42 (0.44)

Source: Ning, Kang, Damian Fermin, et al. "Comparative Analysis of Different Label-free Mass Spectrometry Based Protein Abundance Estimates and Their Correlation with RNA-Seq Gene Expression Data." *Journal of Proteome Research* 11, no. 4 (2012): 2261-71.

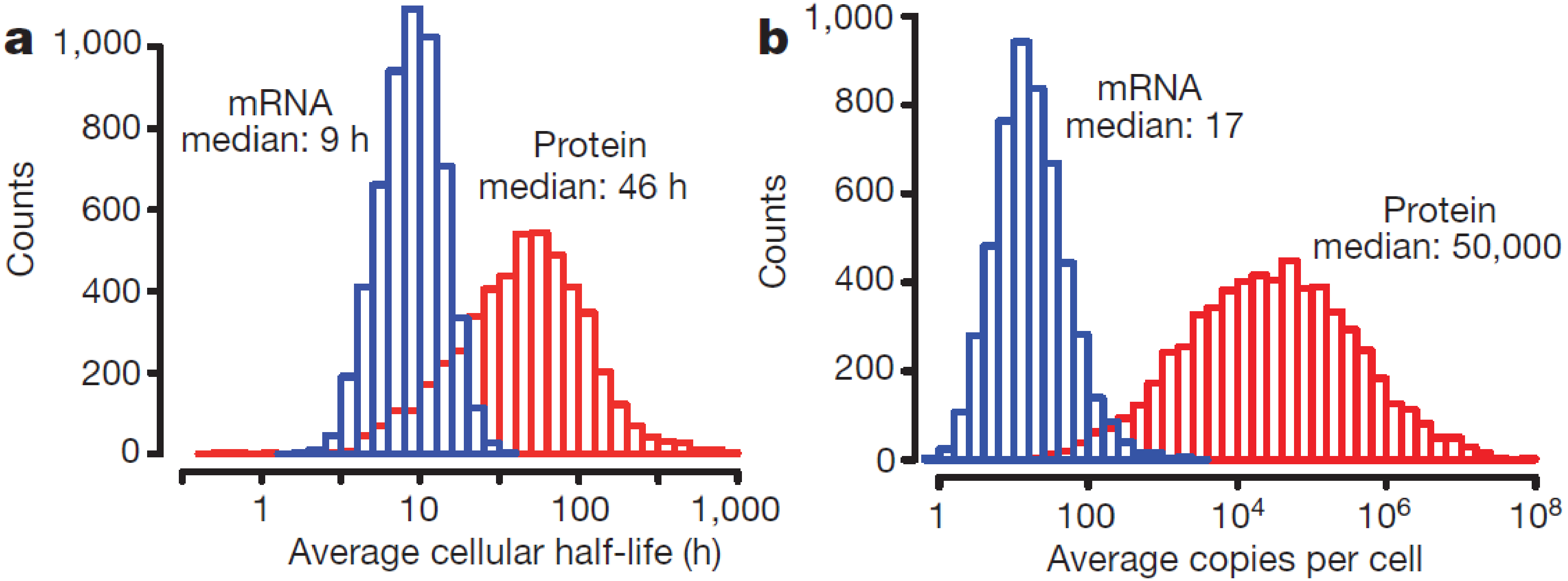
Kang Ning, Damian Fermin, and Alexey I. Nesvizhskii J Proteome Res. 2012 April 6; 11(4): 2261–2271.





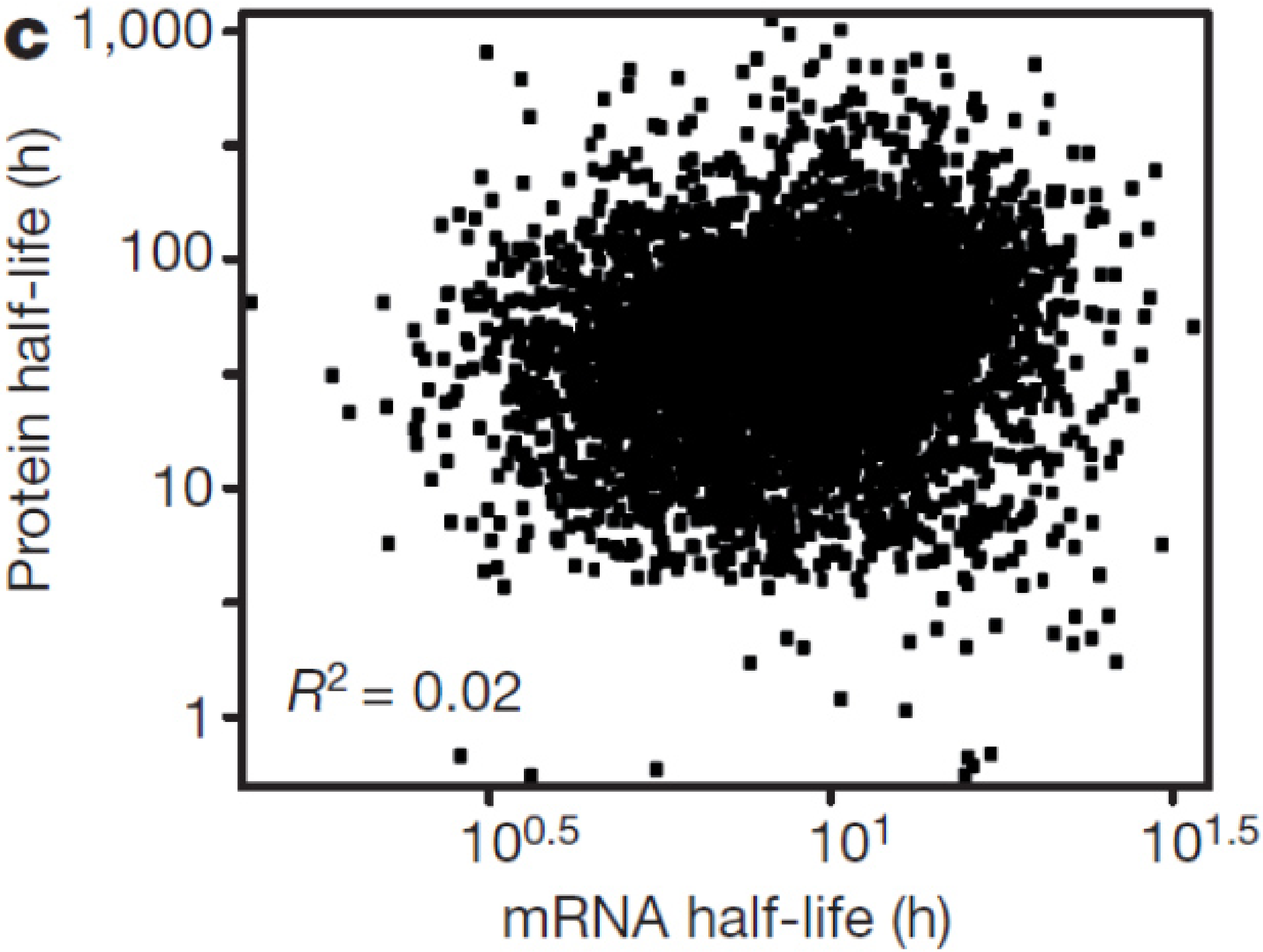
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Schwanhäusser, Björn, Dorothea Busse, et al. "Global Quantification of Mammalian Gene Expression Control." *Nature* 473, no. 7347 (2011): 337-42.

Nature. 2011 May 19;473(7347):337-42. doi: 10.1038/nature10098.
Global quantification of mammalian gene expression control.
Schwanhäusser B1, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M.



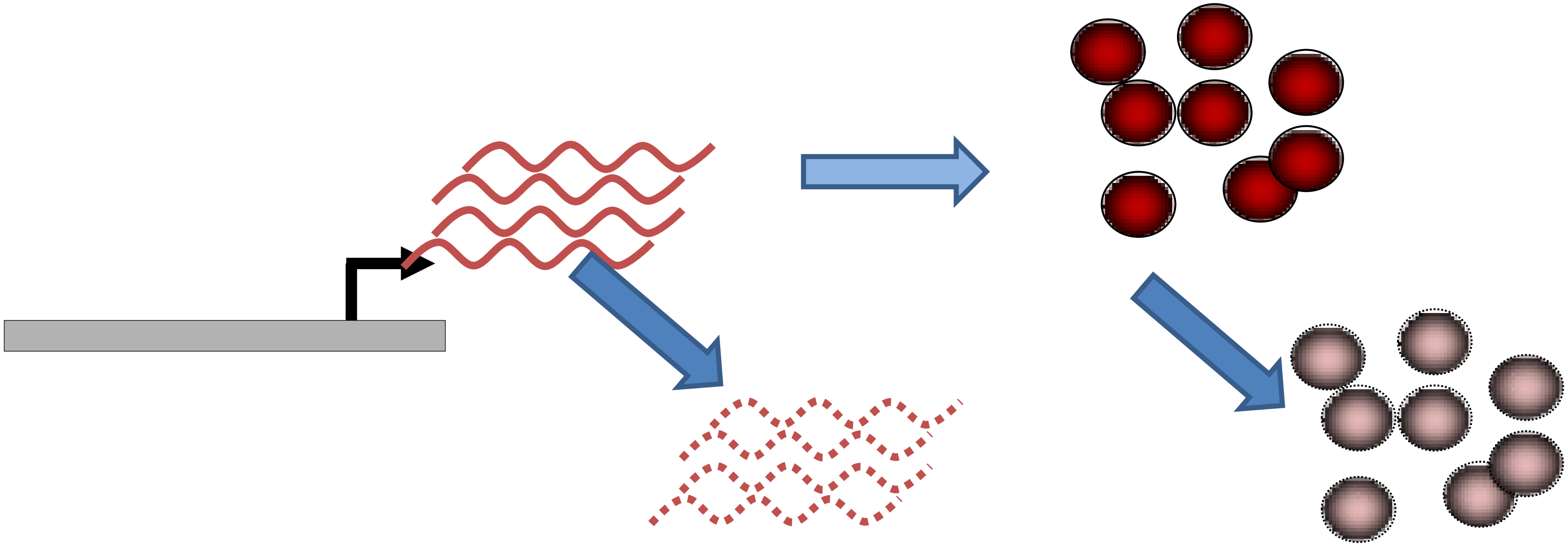
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Schwanhäusser, Björn, Dorothea Busse, et al. "Global Quantification of Mammalian Gene Expression Control." *Nature* 473, no. 7347 (2011): 337-42.

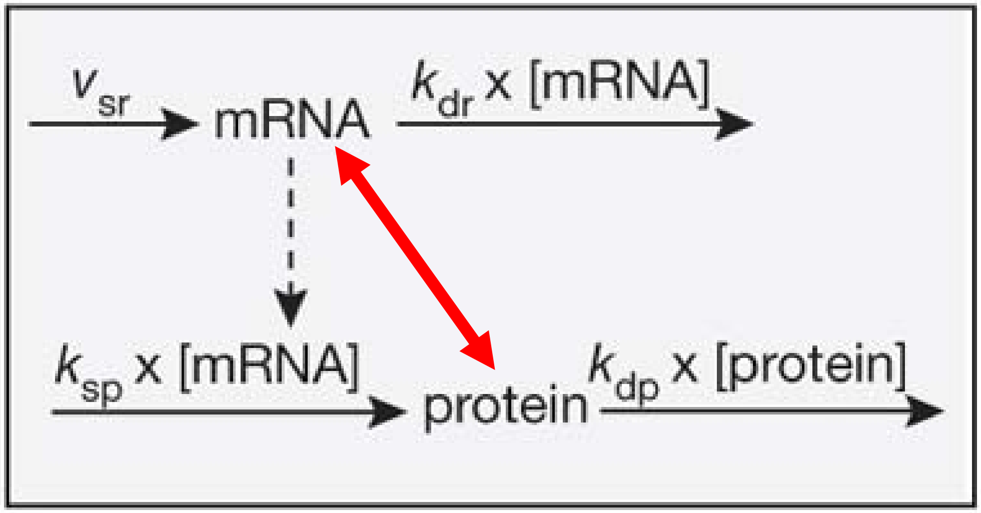
Nature. 2011 May 19;473(7347):337-42. doi: 10.1038/nature10098.
Global quantification of mammalian gene expression control.
Schwanhäusser B1, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M.



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Schwanhäusser, Björn, Dorothea Busse, et al. "Global Quantification of Mammalian Gene Expression Control." *Nature* 473, no. 7347 (2011): 337-42.

Nature. 2011 May 19;473(7347):337-42. doi: 10.1038/nature10098.
Global quantification of mammalian gene expression control.
Schwanhäusser B1, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M.

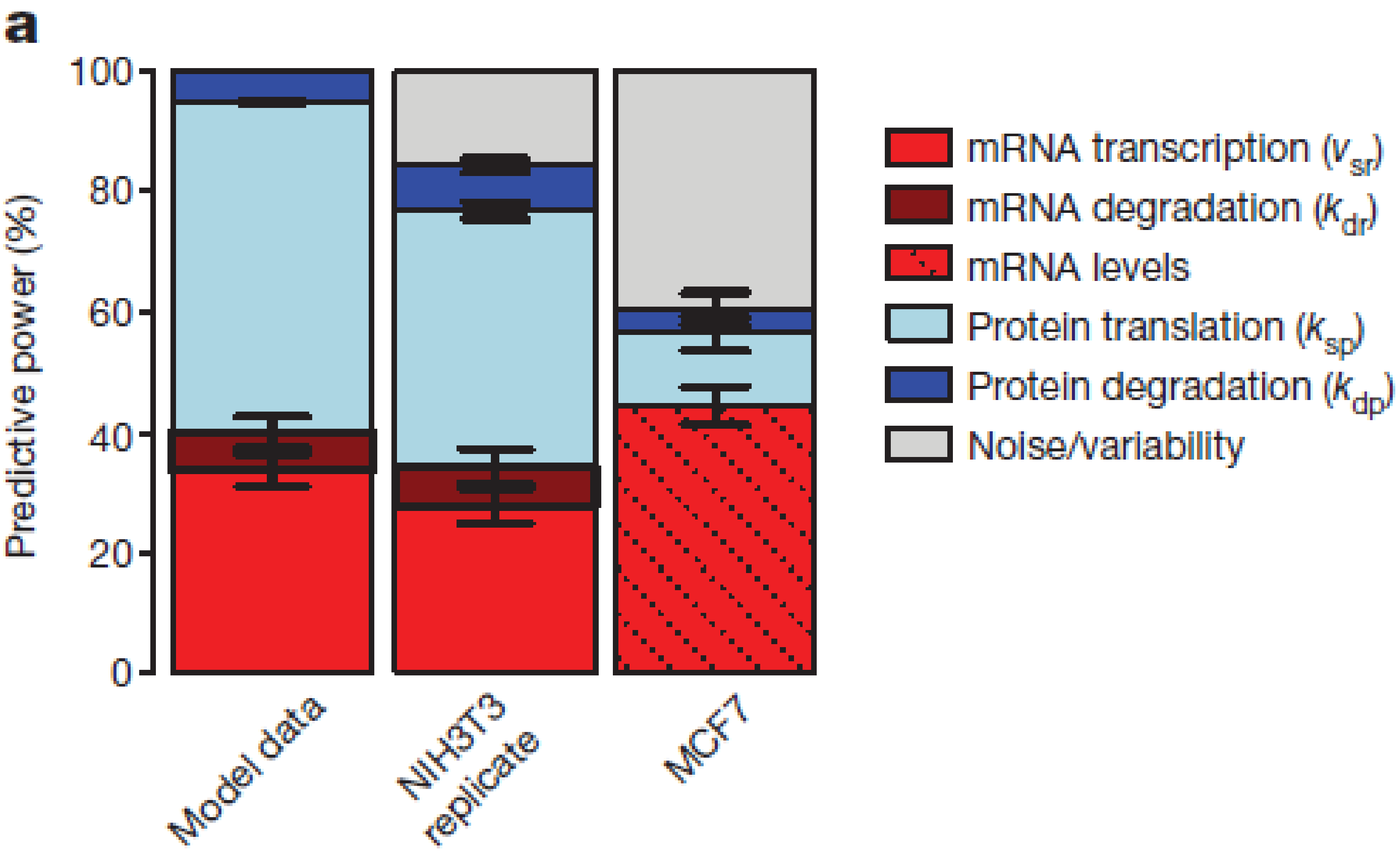




$$\frac{dR}{dt} = v_{sr} - k_{dr} R$$

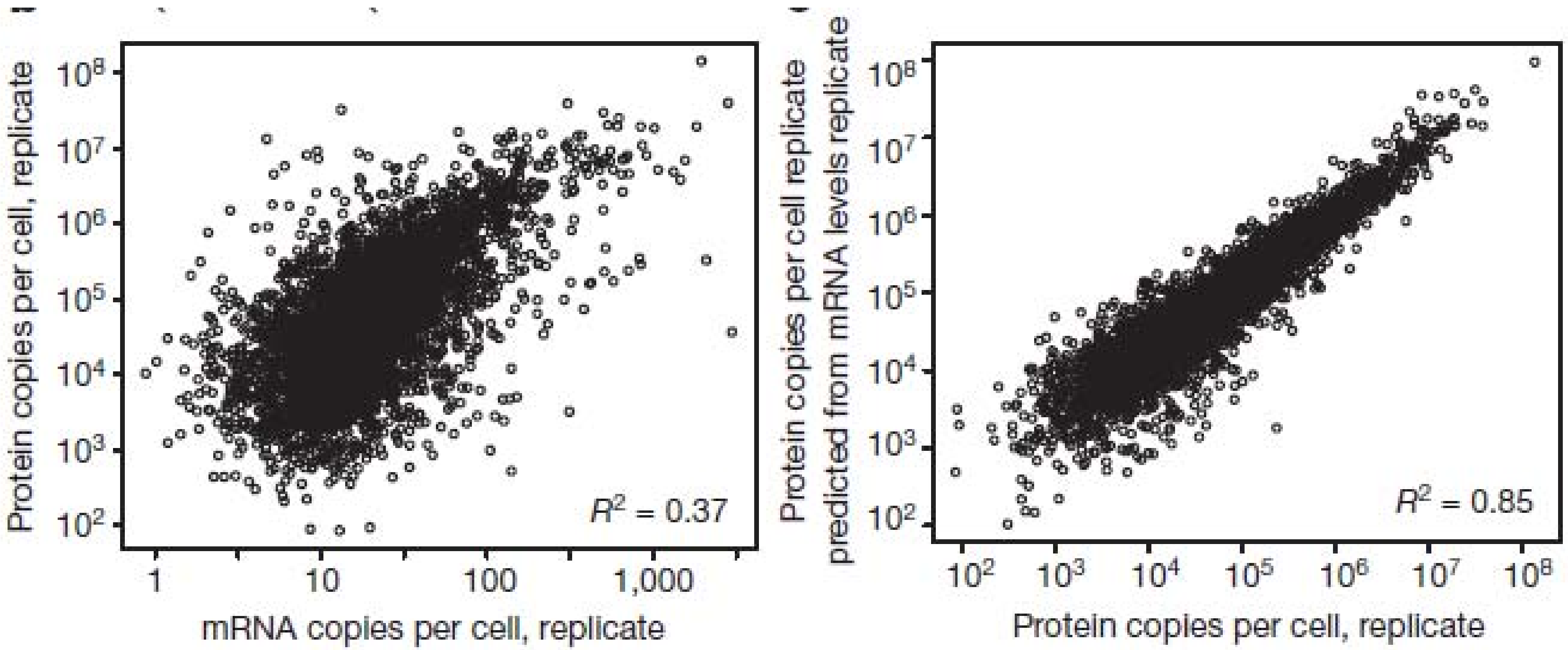
$$\frac{dP}{dt} = k_{sp} R - k_{dp} P$$

Nature. 2011 May 19;473(7347):337-42. doi: 10.1038/nature10098.
Global quantification of mammalian gene expression control.
Schwanhäusser B1, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M.



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Schwanhäusser, Björn, Dorothea Busse, et al. "Global Quantification of Mammalian Gene Expression Control." *Nature* 473, no. 7347 (2011): 337-42.

Nature. 2011 May 19;473(7347):337-42. doi: 10.1038/nature10098.
Global quantification of mammalian gene expression control.
Schwanhäusser B1, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M.

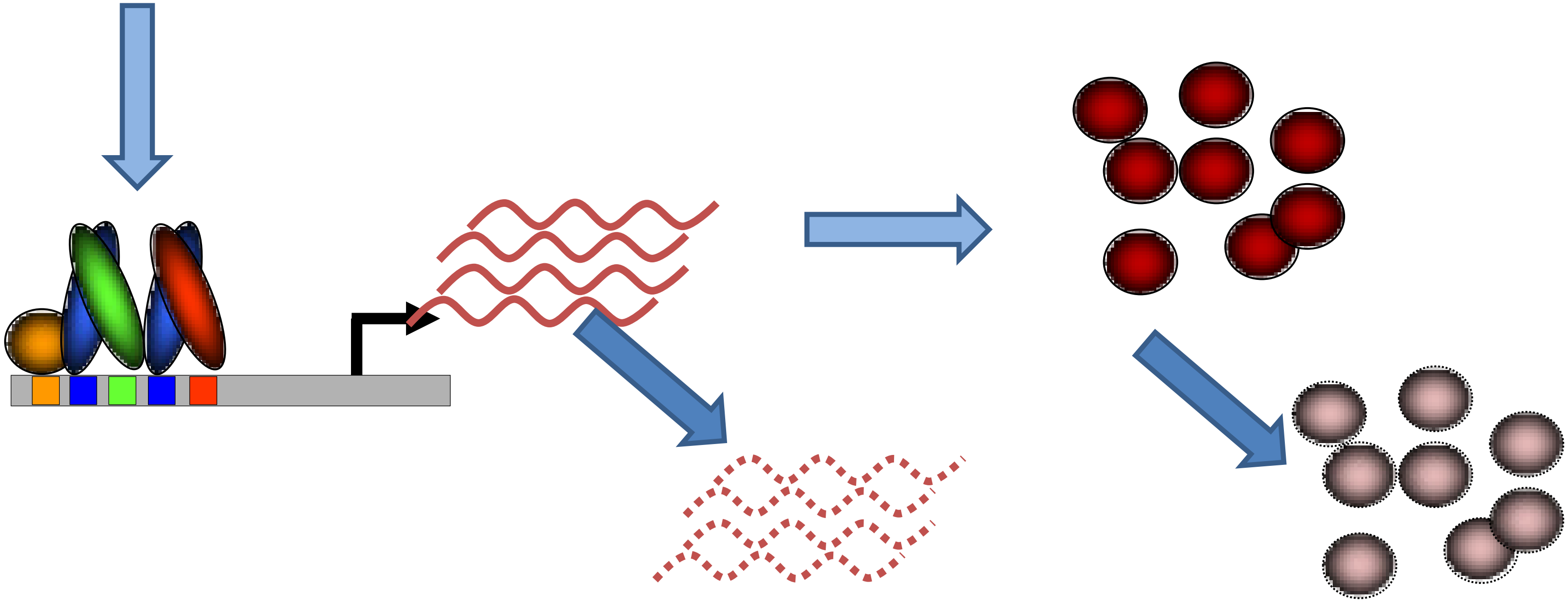


Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Schwanhäusser, Björn, Dorothea Busse, et al. "Global Quantification of Mammalian Gene Expression Control." *Nature* 473, no. 7347 (2011): 337-42.

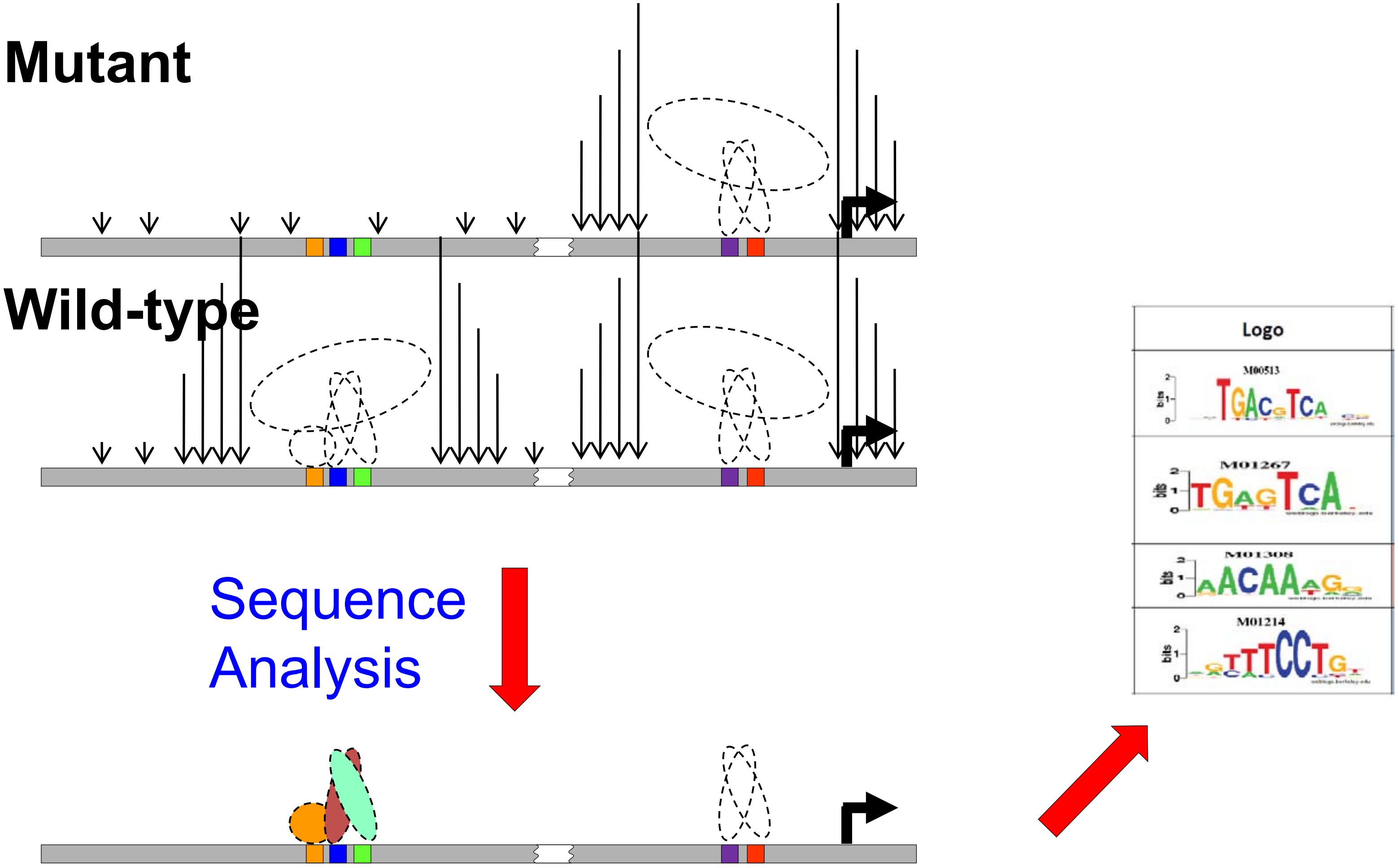
Nature. 2011 May 19;473(7347):337-42. doi: 10.1038/nature10098.
Global quantification of mammalian gene expression control.
Schwanhäusser B1, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M.

Strategies:

1. Use expression to infer upstream events
2. Explicitly model downstream steps



L18 Chromatin and DNase-seq Analysis



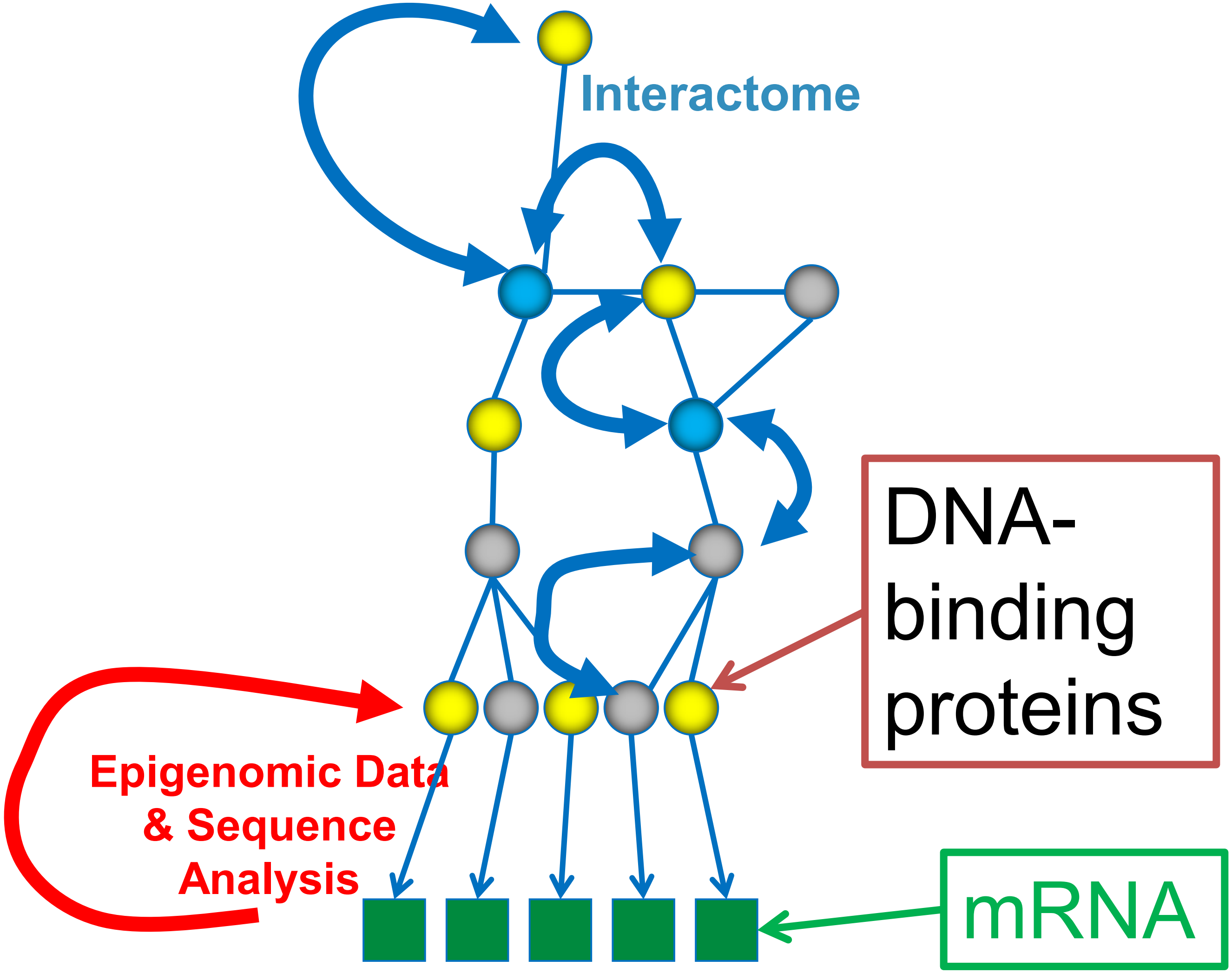
Mutant

Wild-type

Sequence Analysis

Move upstream of transcription

Network integration



Interactome

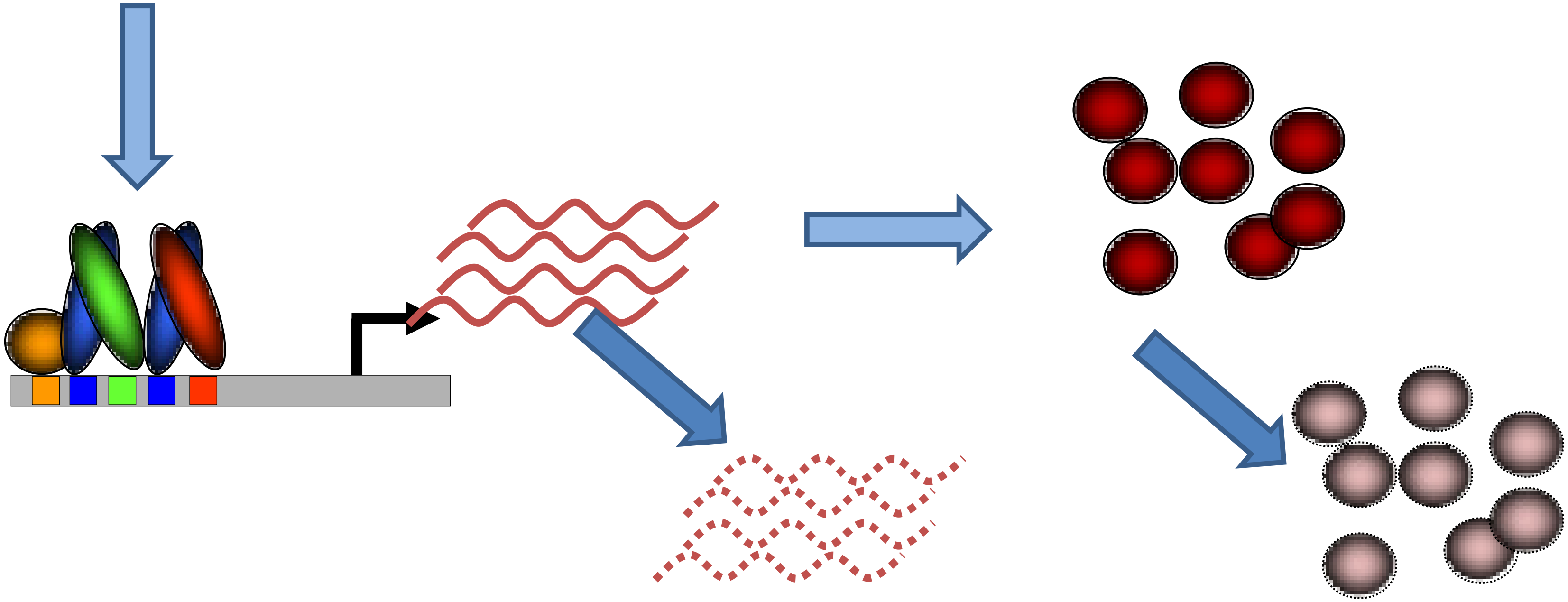
DNA-binding proteins

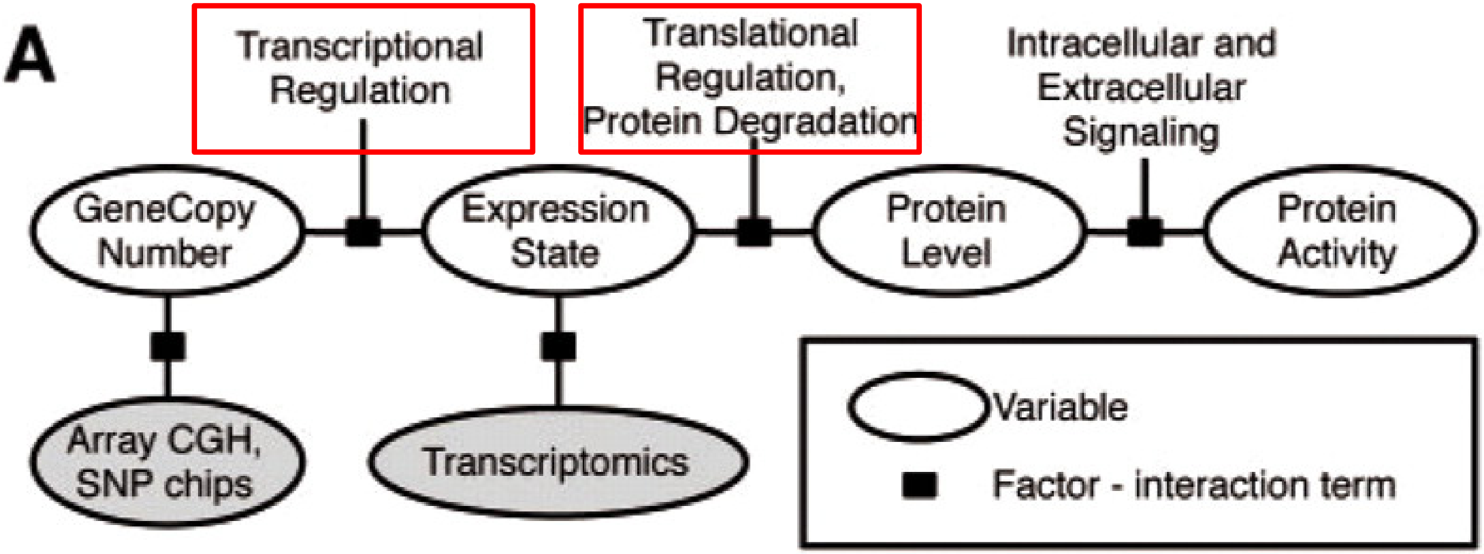
Epigenomic Data & Sequence Analysis

mRNA

Strategies:

1. Use expression to infer upstream events
2. Explicitly model downstream steps




A
Transcriptional Regulation
Translational Regulation, Protein Degradation
Intracellular and Extracellular Signaling
GeneCopy Number
Expression State
Protein Level
Protein Activity
Array CGH, SNP chips
Transcriptomics
Variable
Factor - interaction term

Courtesy of Vaske et al. License: CC-BY.
Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

Vaske C J et al. Bioinformatics 2010;26:i237-i245

BIOINFORMATICS

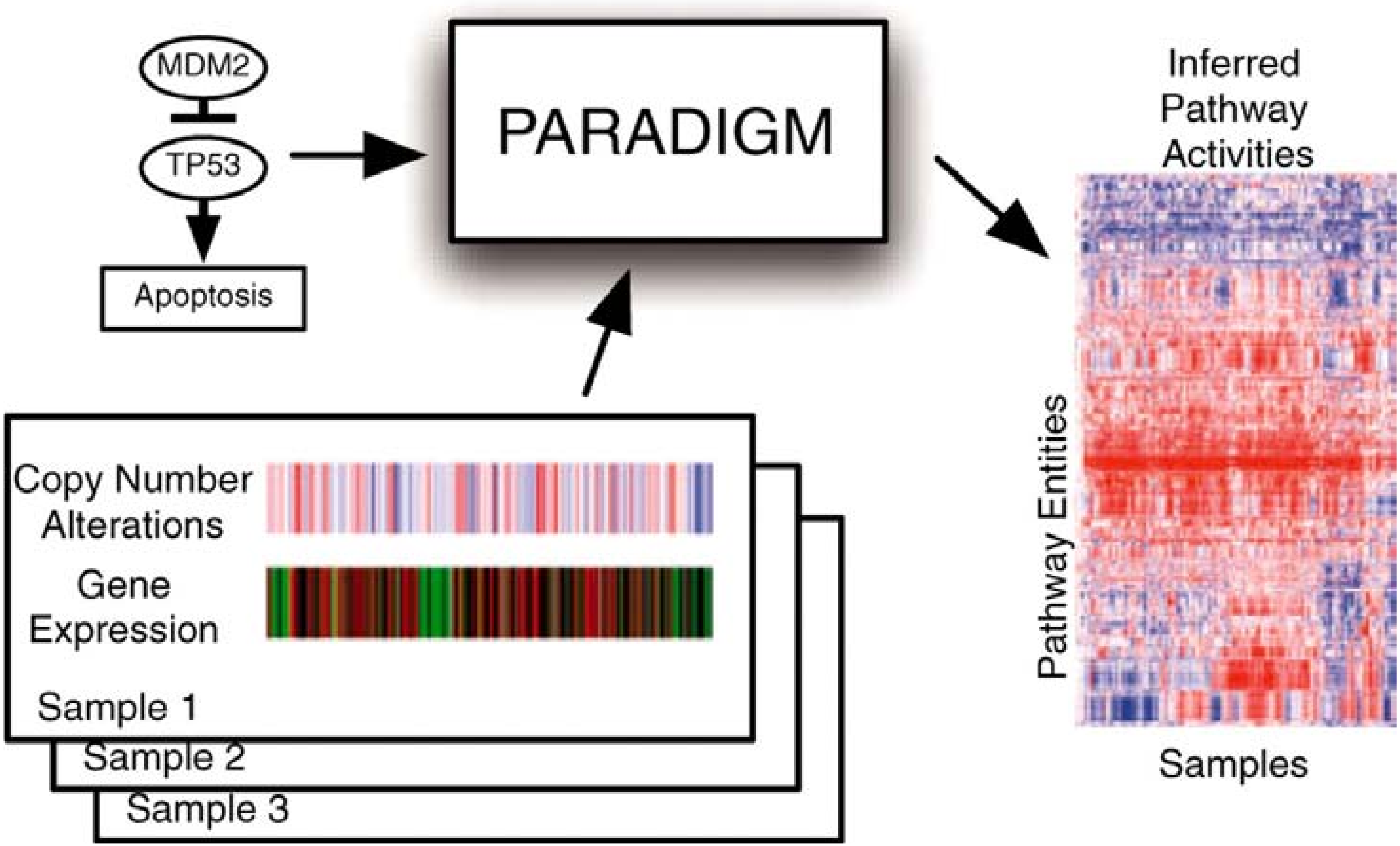
Vol. 26 ISMB 2010, pages i237–i245
doi:10.1093/bioinformatics/btq182

Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM

Charles J. Vaske[1,†], Stephen C. Benz[2,†], J. Zachary Sanborn[2], Dent Earl[2], Christopher Szeto[2], Jingchun Zhu[2], David Haussler[1,2] and Joshua M. Stuart[2,*]

[1]Howard Hughes Medical Institute and [2]Department of Biomolecular Engineering and Center for Biomolecular Science and Engineering, UC Santa Cruz, CA, USA

Overview of the PARADIGM method.

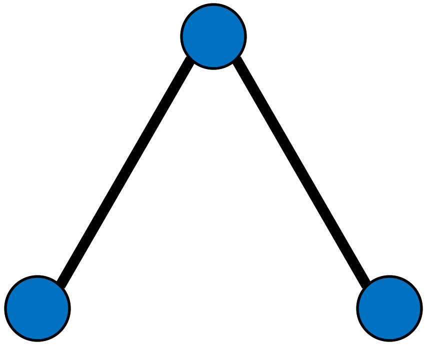


Vaske C J et al. Bioinformatics 2010;26:i237-i245

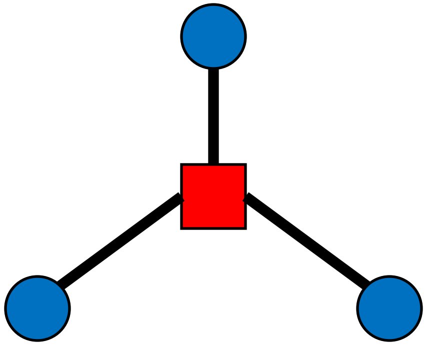
Courtesy of Vaske et al. License: CC-BY.

Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

Factor graphs generalize Bayesian networks



Bayesian network

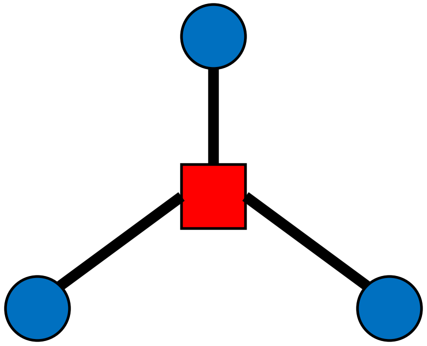


Factor graph

Factor graphs

- Bipartite graph

  (means there are two types of nodes)
- Describes how a global function can be factored into a product of local functions
- Bayesian networks are a type of factor graph



Factor graph

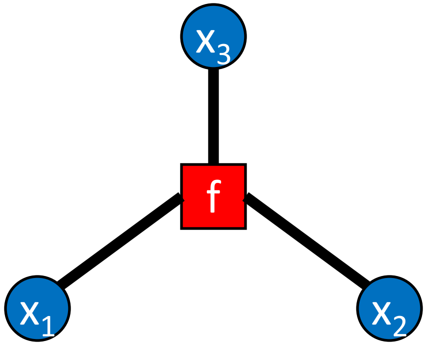
Factor graphs

$$\text{Global function of the variables}: g(x_1, x_2, x_3) = \prod_{j \in J} f_j(X_j)$$

🔵 Variable node, *x*

🟥 Factor node, *f*

→ Edge exists
iff *x* is an argument of *f*



Factor graph

Factor graphs

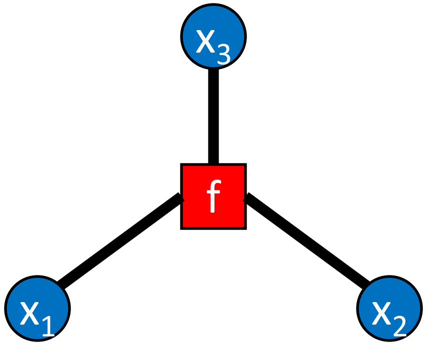
- A node for:

x_1 – every variable and

f – every function f_j (X_j)

- Node x_i is connected to factor f_j iff the variable x_i appears as a term in f_j

$$g(x_1, x_2, x_3) = \prod_{j \in J} f_j(X_j)$$



Factor graph

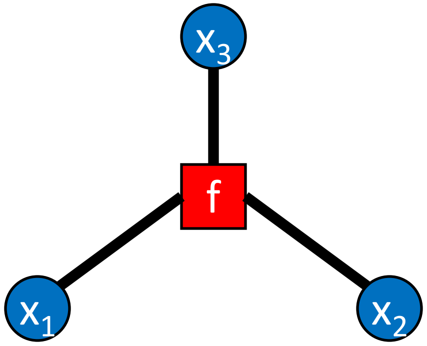
In our setting

$$\text{Joint probabilty function}: P(x_1, x_2, x_3) = \prod_{j \in J} f_j(X_j)$$

🔵 Variable node, *x* = state of gene/protein/pathway

🟥 Factor node, *f* describes relationships

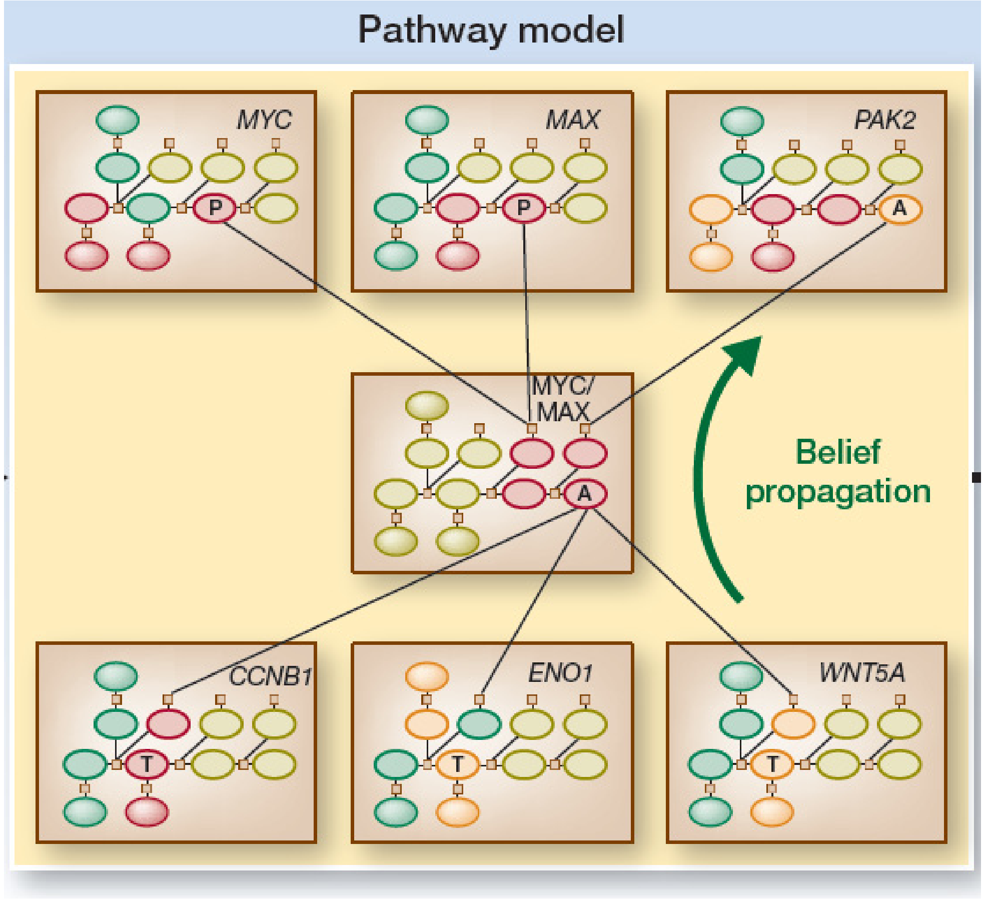
→ Edge exists iff x is an argument of f



Factor graph

Global function: $g(x_1, x_2, x_3, x_4, x_5)$

Marginal $g_i(a)$: sum $g(x_1, x_2, x_3, x_4, x_5)$ over all configurations of the variables with x_i=a



What is the probability that MYC/MAX is active? $P(x_i$=active)

Factor graphs provide a method to compute such marginals

Source: Goldstein, Theodore C., Evan O. Paull, et al. "Molecular Pathways: Extracting Medical Knowledge from High-throughput Genomic Data." *Clinical Cancer Research* 19, no. 12 (2013): 3114-20.

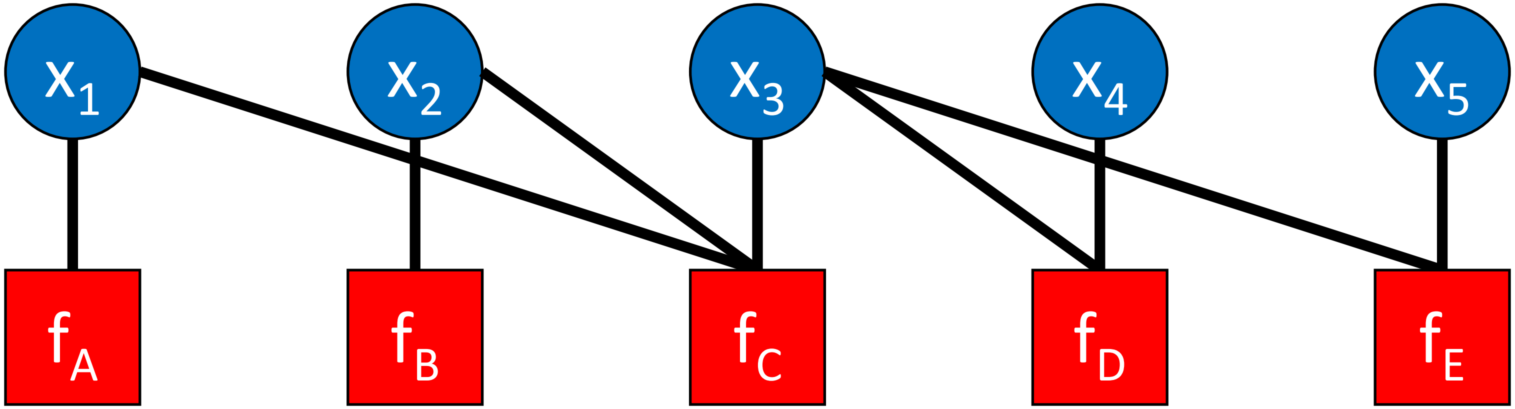
Global function:

$$g(x_1,x_2,x_3,x_4,x_5)=f_A(x_1)f_B(x_2)f_C(x_1,x_2,x_3)f_D(x_3,x_4)f_E(x_3,x_5)$$

Marginal $g_i(a)$: sum $g(x_1,x_2,x_3,x_4,x_5)$ over all configurations of the variables with $x_i=a$

$$g_1(x_1)=f_A(x_1)\times$$

$$\left(\sum_{x_2} f_B(x_2)\left(\sum_{x_3} f_C(x_1,x_2,x_3)\left(\sum_{x_4} f_D(x_3,x_4)\right)\left(\sum_{x_5} f_E(x_3,x_5)\right)\right)\right)$$



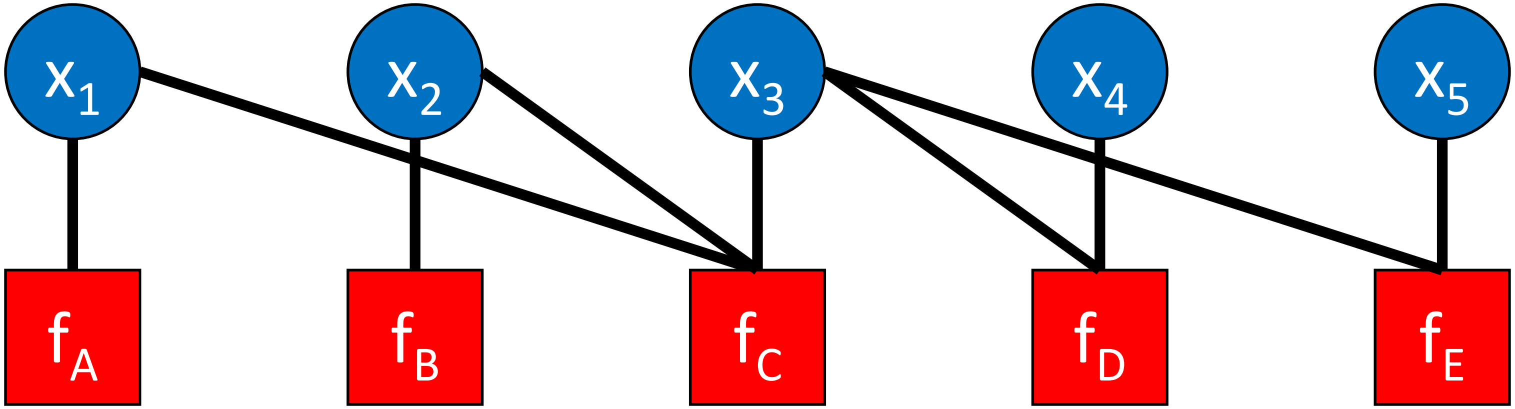
Global function:

$$g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) f_D(x_3, x_4) f_E(x_3, x_5)$$

Marginal $g_i(a)$: sum $g(x_1, x_2, x_3, x_4, x_5)$ over all configurations of the variables with x_i=a

$$g_i(x_i) = \sum_{\sim\{x_i\}} g(x_1, x_2, x_3, x_4, x_5)$$

“not-sum” or summary over all values of $x_{j \neq i}$



Global function:

$$g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) f_D(x_3, x_4) f_E(x_3, x_5)$$

Marginal $g_i(a)$: sum $g(x_1, x_2, x_3, x_4, x_5)$ over all configurations of the variables with x_i=a

$$g_1(x_1) = f_A(x_1) \times$$

$$\left(\sum_{x_2} f_B(x_2) \left(\sum_{x_3} f_C(x_1, x_2, x_3) \left(\sum_{x_4} f_D(x_3, x_4) \right) \left(\sum_{x_5} f_E(x_3, x_5) \right) \right) \right)$$

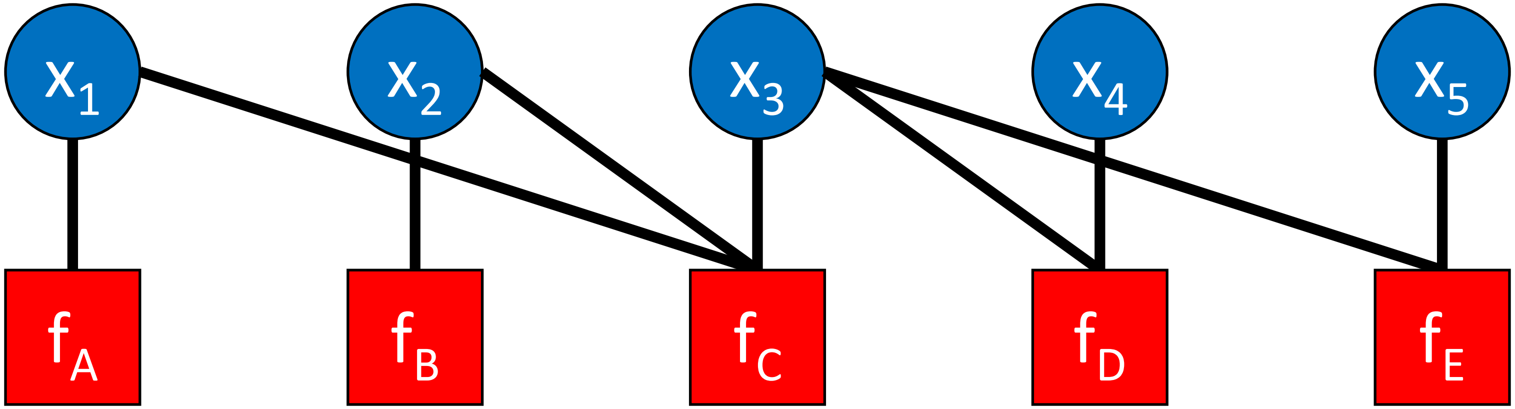
$$g_1(x_1) = f_A(x_1) \times$$

$$\sum_{\sim\{x_1\}} \left(f_B(x_2) f_C(x_1, x_2, x_3) \left(\sum_{\sim\{x_3\}} f_D(x_3, x_4) \right) \left(\sum_{\sim\{x_3\}} f_E(x_3, x_5) \right) \right)$$

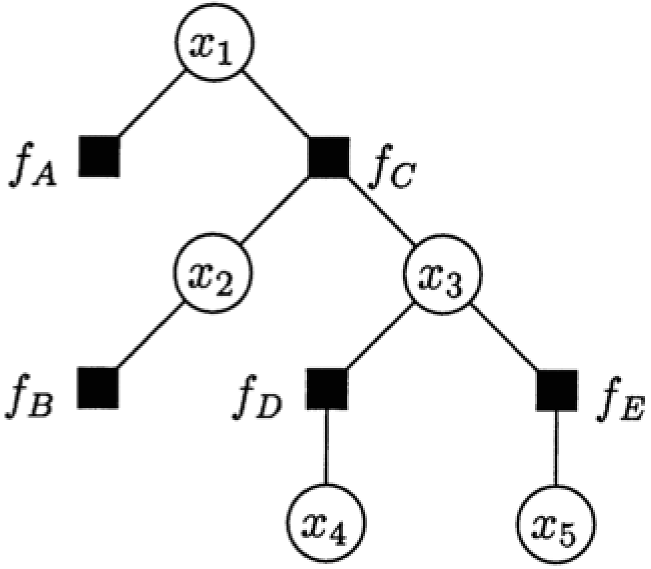
Global function:

$$g(x_1,x_2,x_3,x_4,x_5) = f_A(x_1)f_B(x_2)f_C(x_1,x_2,x_3)f_D(x_3,x_4)f_E(x_3,x_5)$$

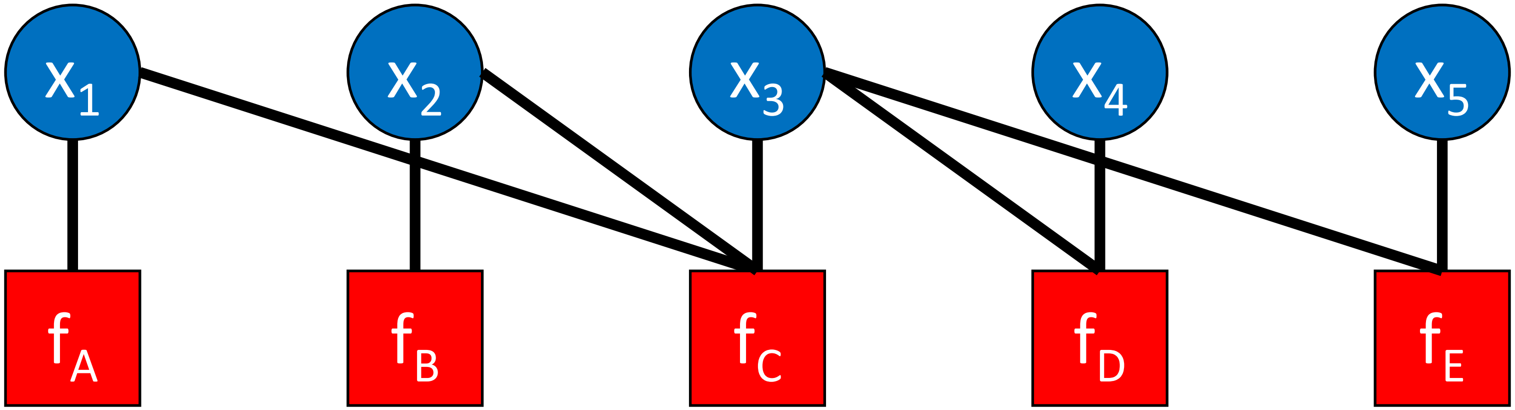
How do we find the marginal for any factor graph?



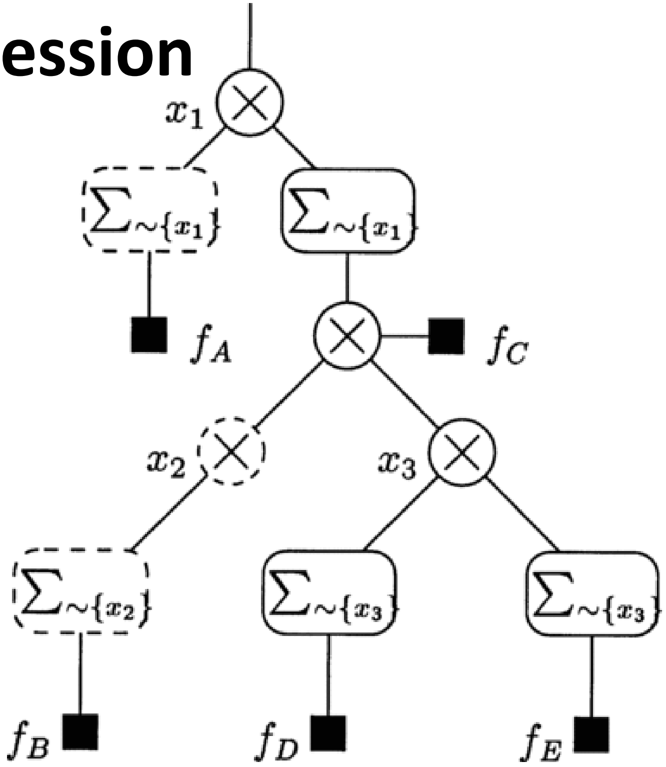
To compute the marginal with respect to variable x_i : draw the factor graph as a tree with root x_i



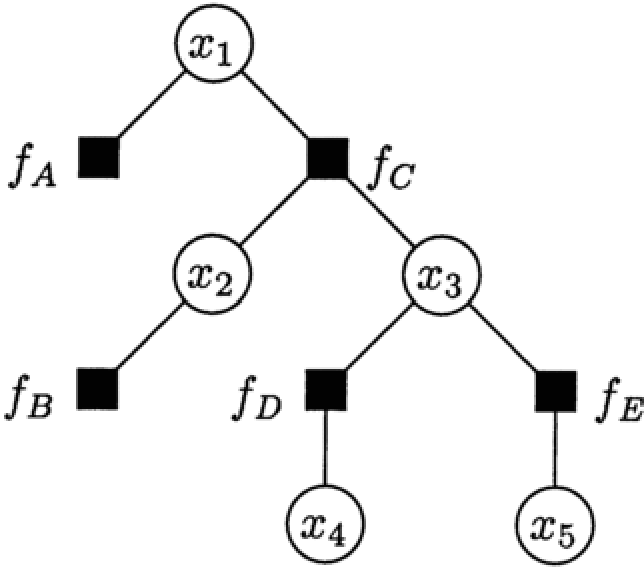
Source: Kschischang, Frank R., Brendan J. Frey,et al. "Factor Graphs and the Sum-product Algorithm." *Information Theory, IEEE Transactions on* 47, no. 2 (2001): 498-519.



Expression Tree



Factor Graph

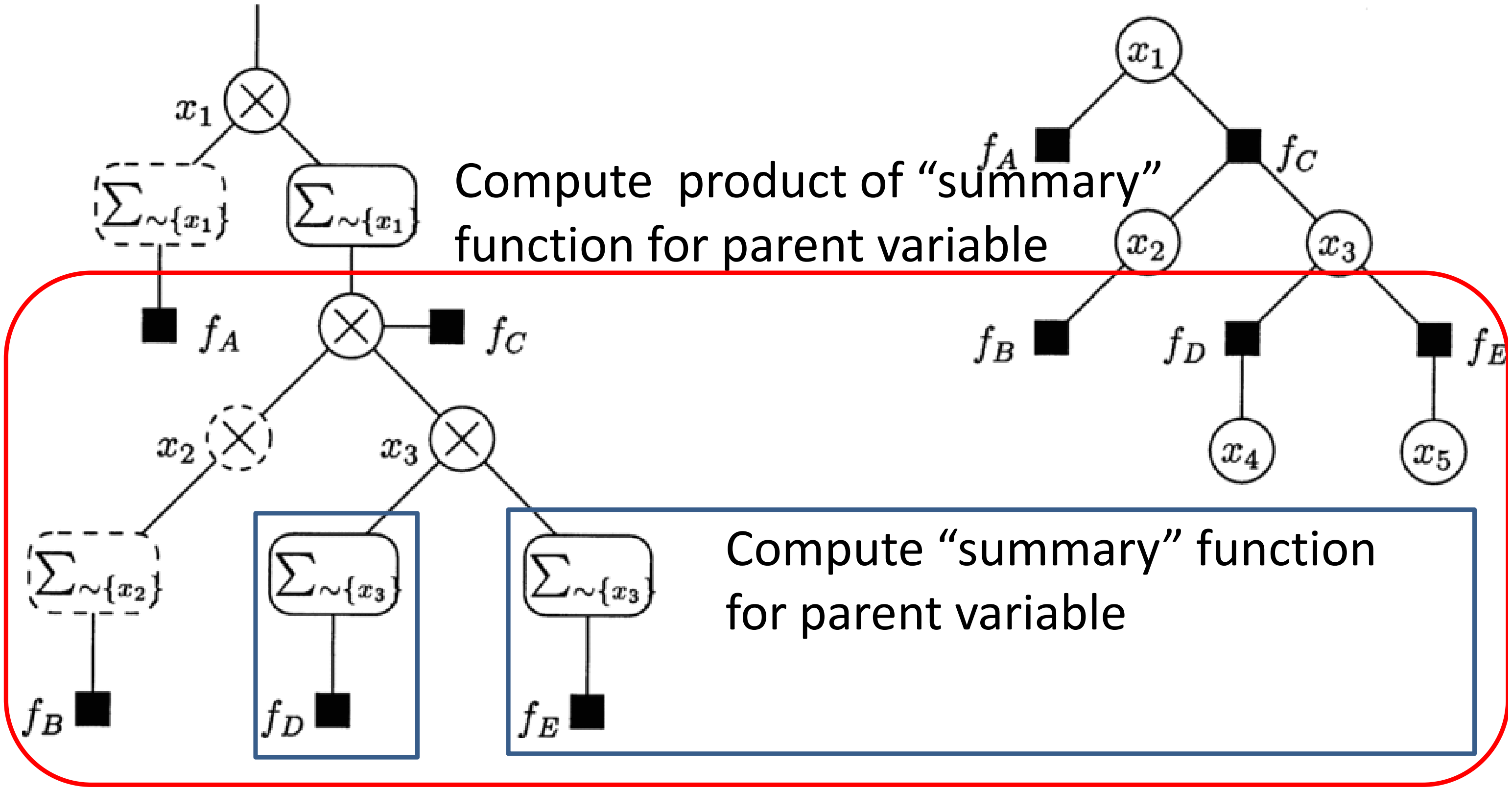


Source: Kschischang, Frank R., Brendan J. Frey,et al. "Factor Graphs and the Sum-product Algorithm." *Information Theory, IEEE Transactions on* 47, no. 2 (2001): 498-519.

Marginal:

$$g_1(x_1) = f_A(x_1) \times$$

$$\sum_{\sim\{x_1\}} \left(f_B(x_2) f_C(x_1, x_2, x_3) \left(\sum_{\sim\{x_3\}} f_D(x_3, x_4) \right) \left(\sum_{\sim\{x_3\}} f_E(x_3, x_5) \right) \right)$$



Compute product of "summary" function for parent variable

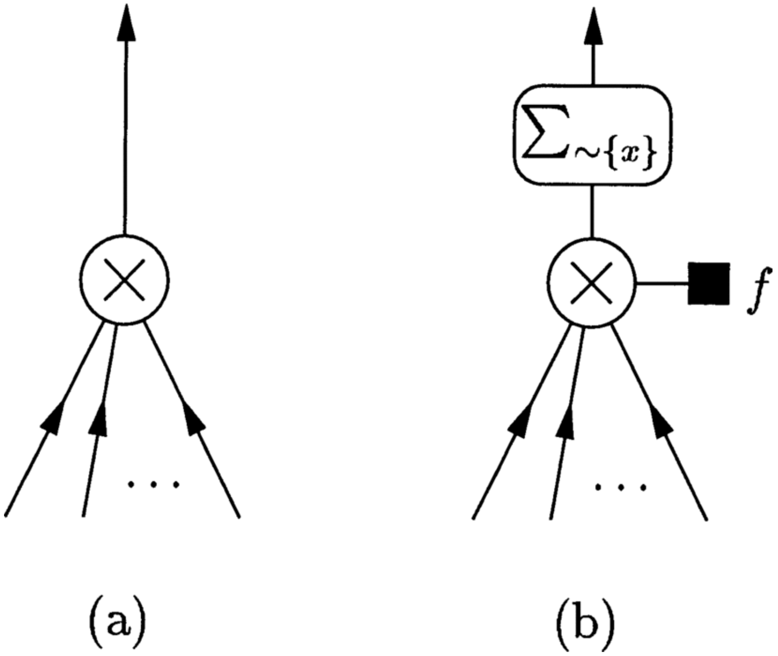
Compute "summary" function for parent variable

Source: Kschischang, Frank R., Brendan J. Frey,et al. "Factor Graphs and the Sum-product Algorithm." *Information Theory, IEEE Transactions on* 47, no. 2 (2001): 498-519.

Marginal:

$$g_1(x_1) = f_A(x_1) \times$$

$$\sum_{\sim\{x_1\}} \left(f_B(x_2) f_C(x_1, x_2, x_3) \left(\sum_{\sim\{x_3\}} f_D(x_3, x_4) \right) \left(\sum_{\sim\{x_3\}} f_E(x_3, x_5) \right) \right)$$



Source: Kschischang, Frank R., Brendan J. Frey,et al. "Factor Graphs and the Sum-product Algorithm." *Information Theory, IEEE Transactions on* 47, no. 2 (2001): 498-519.

Messages flow up from leaves:

•Each vertex waits for messages from all children before computing message to send to parents

•Variable nodes send product of messages from children

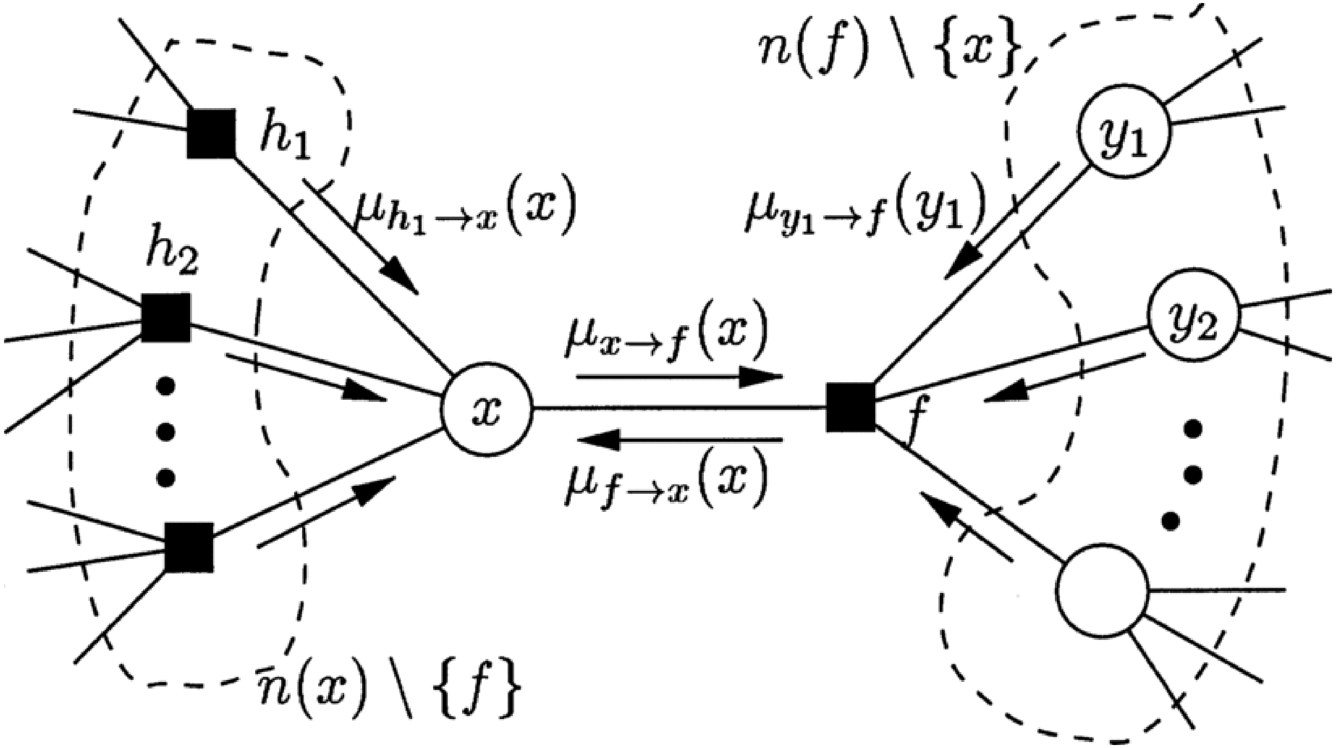
•Factor nodes with parent x send the "summary" for x of the product of the children's functions.

Kschischang, F.R.; Frey, B.J.; Loeliger, H.-A., "Factor graphs and the sum-product algorithm," 2001
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=910572&isnumber=19638

Belief propagation:

An algorithm known as "Sum-Product" can be used to simultaneously compute all marginals!

See citation for details



Source: Kschischang, Frank R., Brendan J. Frey,et al. "Factor Graphs and the Sum-product Algorithm." *Information Theory, IEEE Transactions on* 47, no. 2 (2001): 498-519.

Kschischang, F.R.; Frey, B.J.; Loeliger, H.-A., "Factor graphs and the sum-product algorithm," 2001

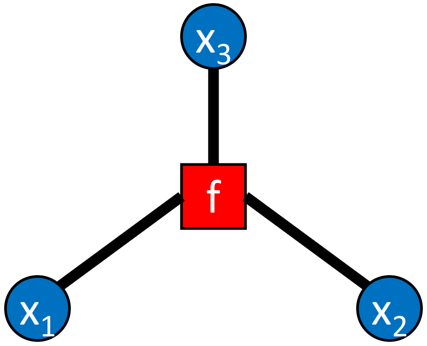
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=910572&isnumber=19638

Factor graphs in PARADIGM

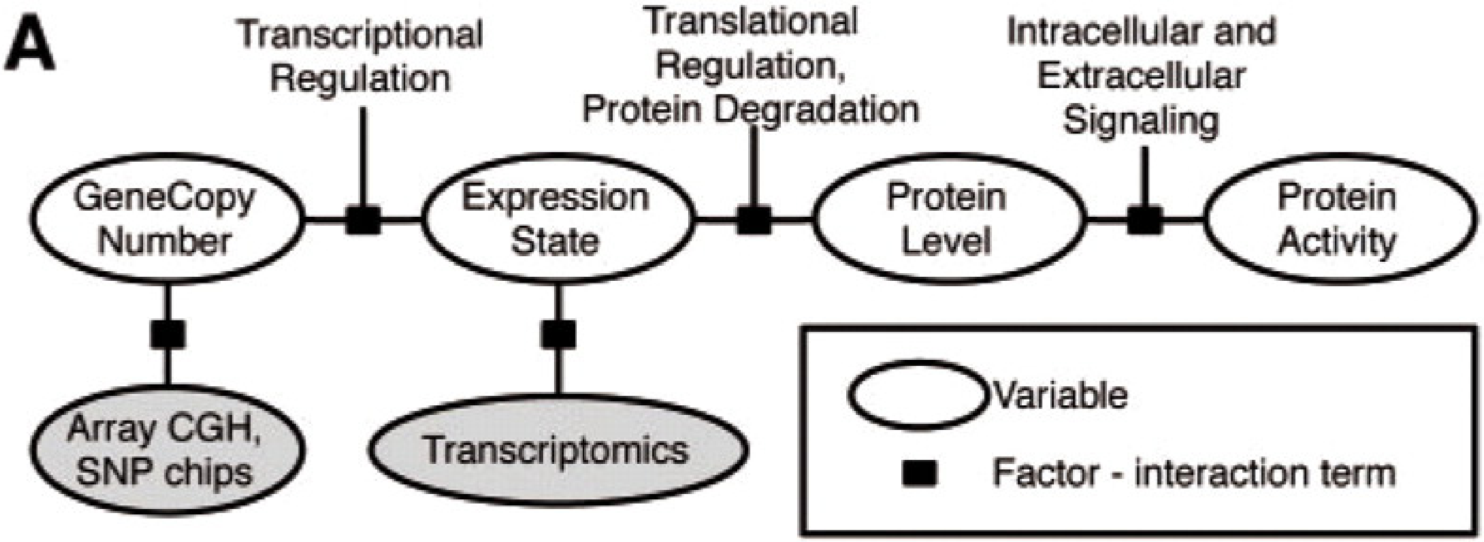
● Variable node, *x:*
three states:
1 activated
0 nominal
-1 deactivated

🟥 Factor node, *f*

→ Edge exists iff x is an argument of f

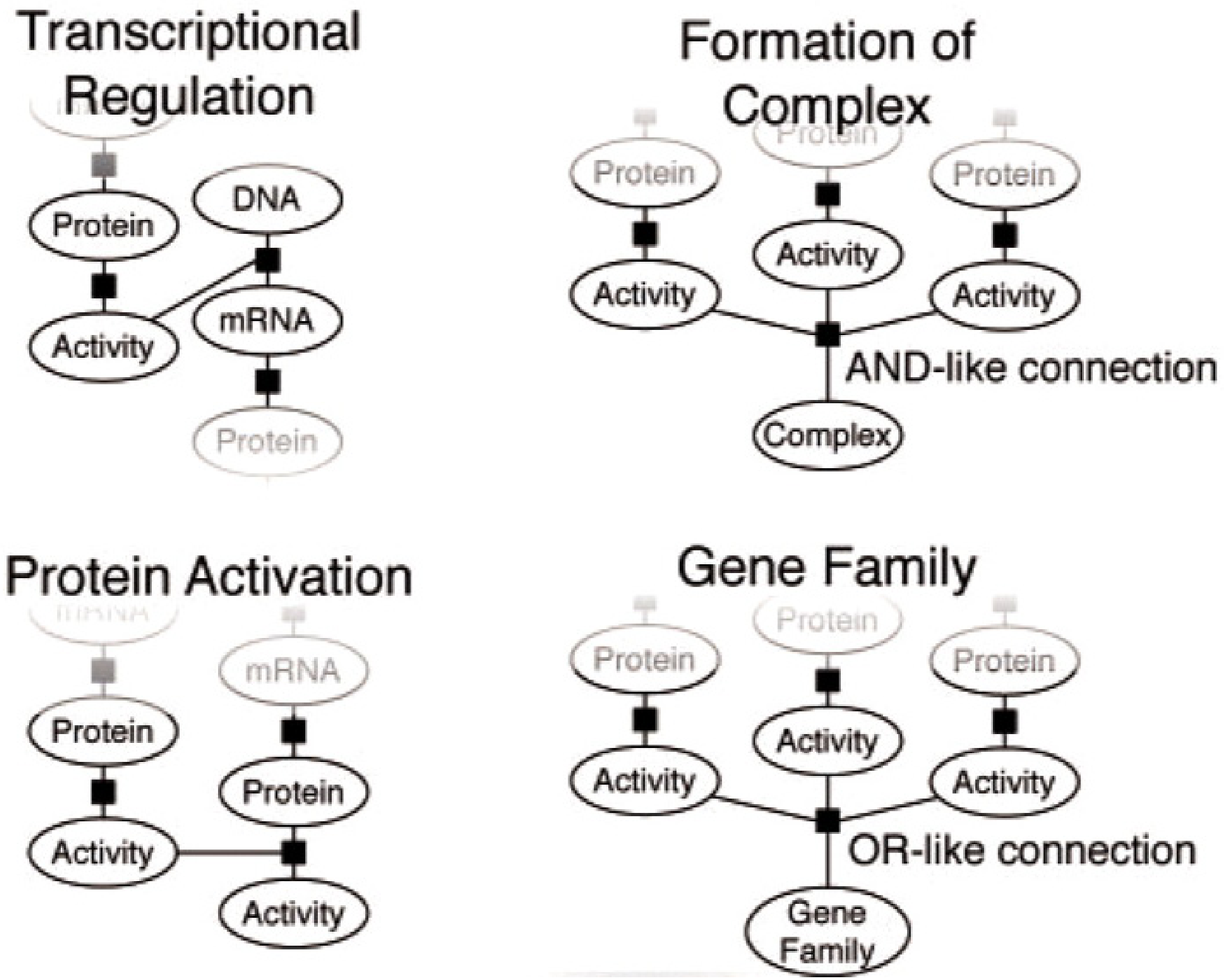


Factor graph


A
Transcriptional Regulation
Translational Regulation, Protein Degradation
Intracellular and Extracellular Signaling
GeneCopy Number
Expression State
Protein Level
Protein Activity
Array CGH, SNP chips
Transcriptomics
Variable
Factor - interaction term

Courtesy of Vaske et al. License: CC-BY.
Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

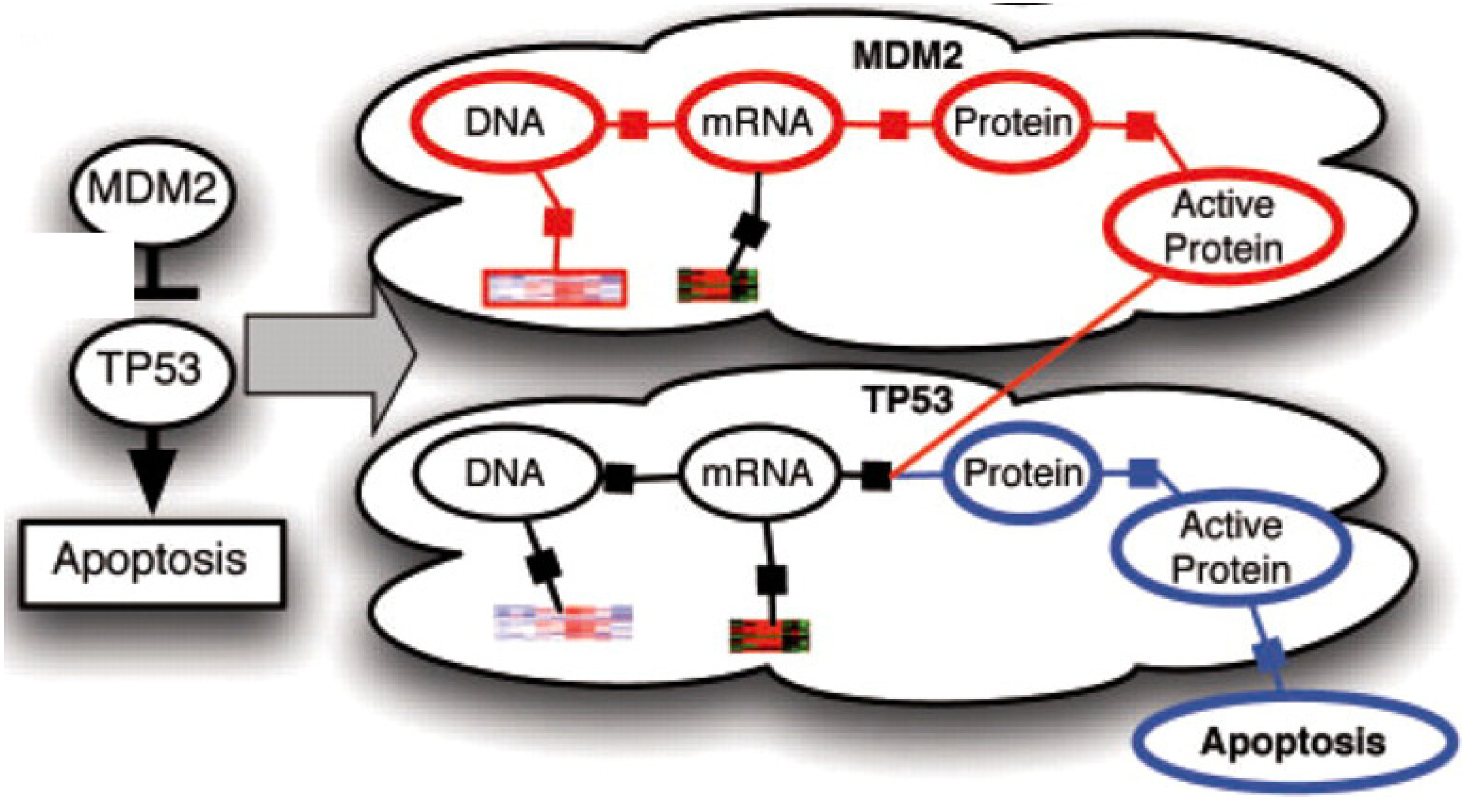
Vaske C J et al. Bioinformatics 2010;26:i237-i245



Courtesy of Vaske et al. License: CC-BY.
Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

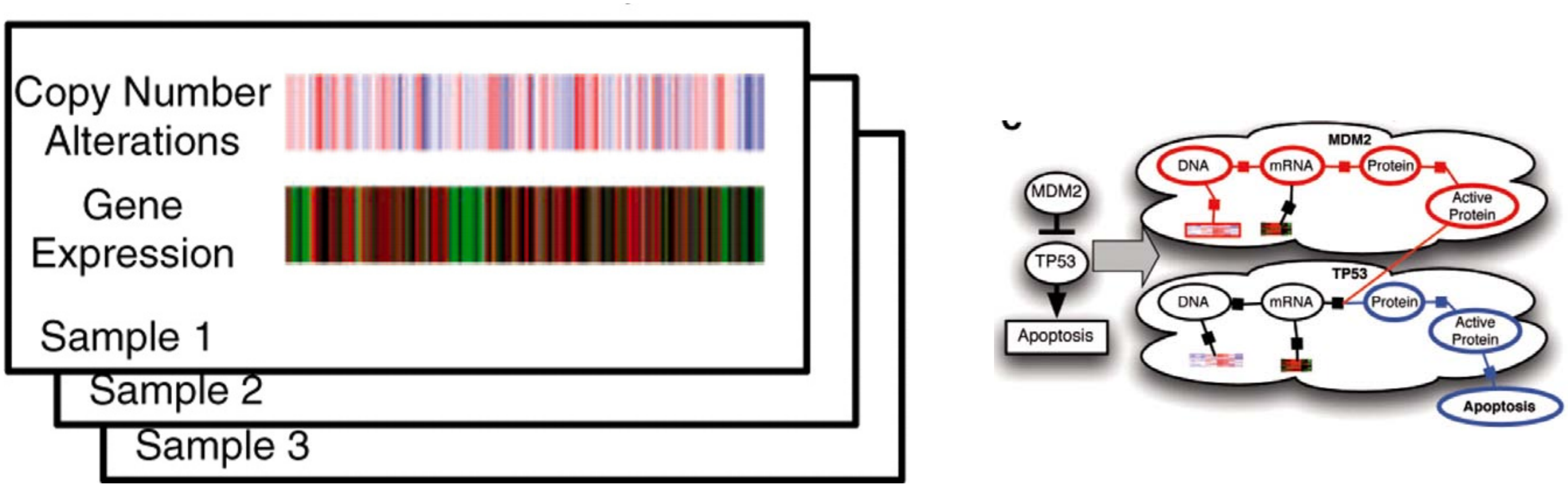
Vaske C J et al. Bioinformatics 2010;26:i237-i245

Bioinformatics



Courtesy of Vaske et al. License: CC-BY.
Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

Vaske C J et al. Bioinformatics 2010;26:i237-i245



Courtesy of Vaske et al. License: CC-BY.
Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

- Goal:
  - Estimate probability that pathways are active
  - Use log likelihood ratio

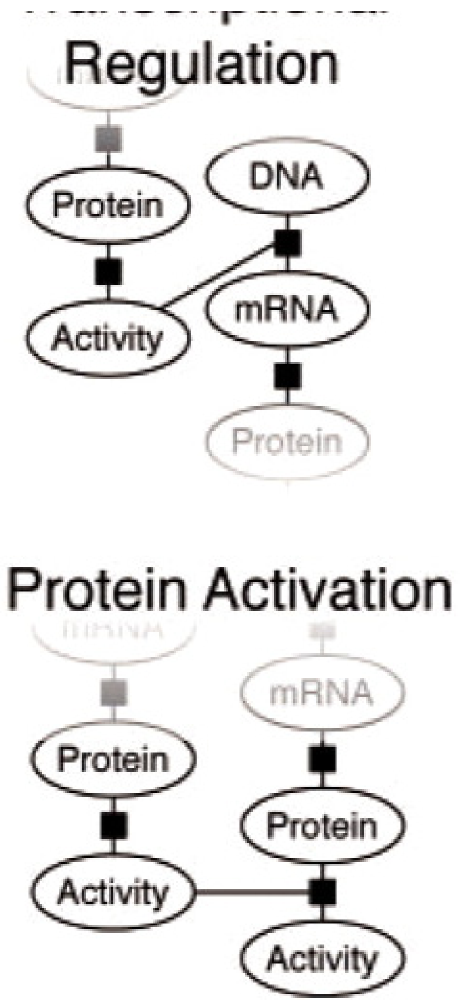
$$L(i,a) = \log\left(\frac{P(D,x_i{=}a|\Phi)}{P(D,x_i{\neq}a|\Phi)}\right) - \log\left(\frac{P(x_i{=}a|\Phi)}{P(x_i{\neq}a|\Phi)}\right)$$

$$= \log\left(\frac{P(D|x_i{=}a,\Phi)}{P(D|x_i{\neq}a,\Phi)}\right).$$

Parameters estimated by EM from experimental data

Vaske C J et al. Bioinformatics 2010;26:i237-i245

Manually constructed

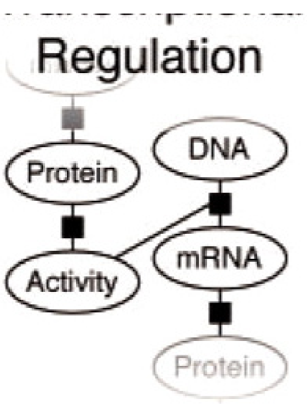


Known pathways:

- Convert to a directed graph
- Each edge is labeled as either positive or negative based on influence
- Define joint probability

Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

Defining joint probability



Courtesy of Vaske et al. License: CC-BY.
Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

Expected state:

- Majority vote of parent variables
- If a parent is connected by a positive edge it contributes a vote of +1 times its own state to the value of the factor.
- If the parent is connected by a negative edge, then the variable votes −1 times its own state.

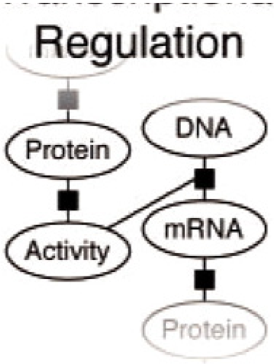
$$\phi_i(x_i, \text{Parents}(x_i)) = \begin{cases} 1-\epsilon & x_i \text{ is the expected state from Parents}(x_i) \\ \frac{\epsilon}{2} & \text{otherwise.} \end{cases}$$

ϵ was set to 0.001

Defining factors manually

$$\phi_i(x_i, \text{Parents}(x_i)) = \begin{cases} 1-\epsilon & x_i \text{ is the expected state from Parents}(x_i) \\ \frac{\epsilon}{2} & \text{otherwise.} \end{cases}$$

ϵ was set to 0.001



Courtesy of Vaske et al. License: CC-BY.
Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

Logic:

•AND: The variables connected to x_i by an edge labeled ‘minimum’ get a single vote, and that vote's value is the minimum value of these variables

•OR: The variables connected to x_i by an edge labeled ‘maximum’ get a single vote, and that vote's value is the maximum value of these variables, creating an OR-like connection.

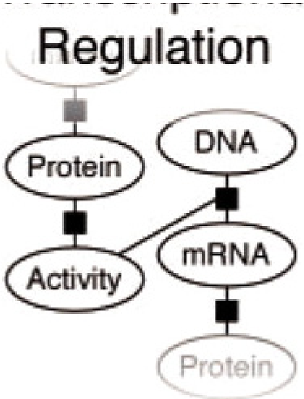
•Votes of zero are treated as abstained votes.

•If there are no votes the expected state is zero. Otherwise, the majority vote is the expected state, and a tie between 1 and −1 results in an expected state of −1 to give more importance to repressors and deletions.

Defining factors manually

$$\phi_i(x_i, \text{Parents}(x_i)) = \begin{cases} 1-\epsilon & x_i \text{ is the expected state from Parents}(x_i) \\ \frac{\epsilon}{2} & \text{otherwise.} \end{cases}$$

ϵ was set to 0.001



Courtesy of Vaske et al. License: CC-BY.
Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

Logic:

•AND: The variables connected to x_i by an edge labeled ‘minimum’ get a single vote, and that vote's value is the minimum value of these variables

•OR: The variables connected to x_i by an edge labeled ‘maximum’ get a single vote, and that vote's value is the maximum value of these variables, creating an OR-like connection.

Compared to Bayesian networks, factor graphs provide an more intuitive way to represent these regulatory steps

Joint probability of graph

$$\phi_i(x_i, \text{Parents}(x_i)) = \begin{cases} 1-\epsilon & x_i \text{ is the expected state from Parents}(x_i) \\ \frac{\epsilon}{2} & \text{otherwise.} \end{cases}$$

$$P(X) = \frac{1}{Z} \prod_{j=1}^{m} \phi_j\left(X_j\right),$$

Product over all *m* factors ϕ_j

$$Z = \prod_j \sum_{\mathbf{S} \sqsubset X_j} \phi_j(\mathbf{S})$$

$\mathbf{S} \sqsubset X$ Setting of variables = possible values

Marginal

$$P(x_i = a|\Phi) = \frac{1}{Z}\prod_{j=1}^{m}\sum_{\mathbf{S} \sqsubset_{A_i(a)} X_j} \phi_j(\mathbf{S})$$

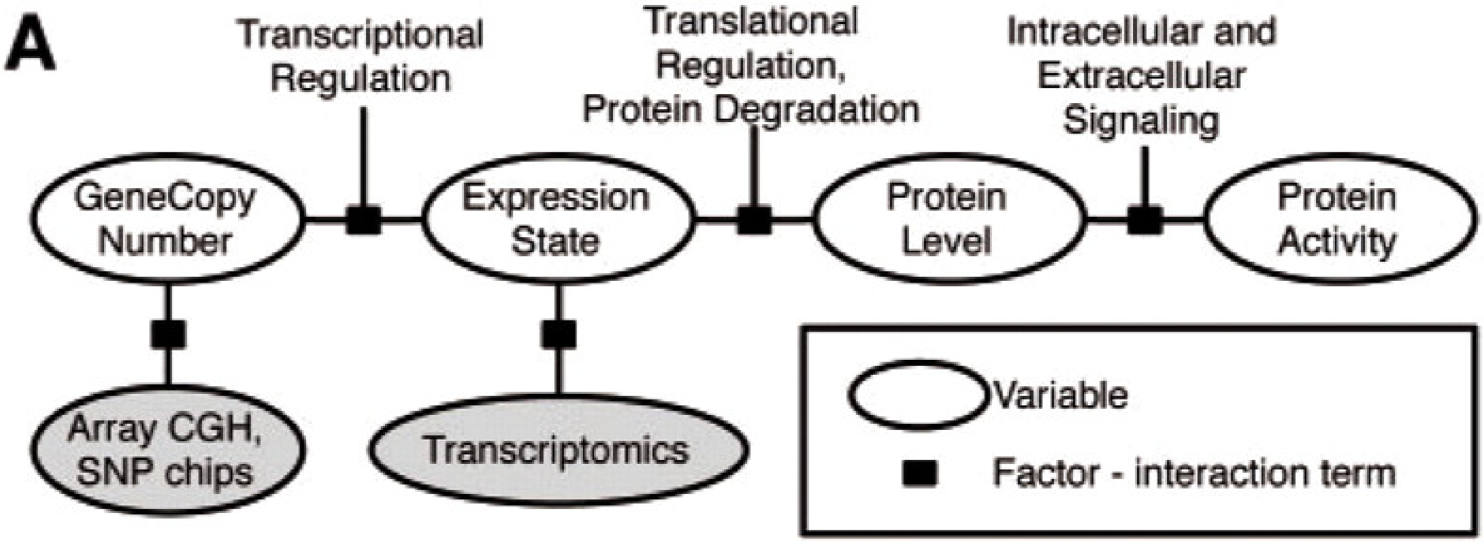
$\{\mathbf{S} \sqsubset_D X\}$ Set of all possible assignments to the variables X consistent with data D

$A_i(a)$ represents the singleton assignment set $\{x_i = a\}$

Φ Full specified factor graph

Likelihood

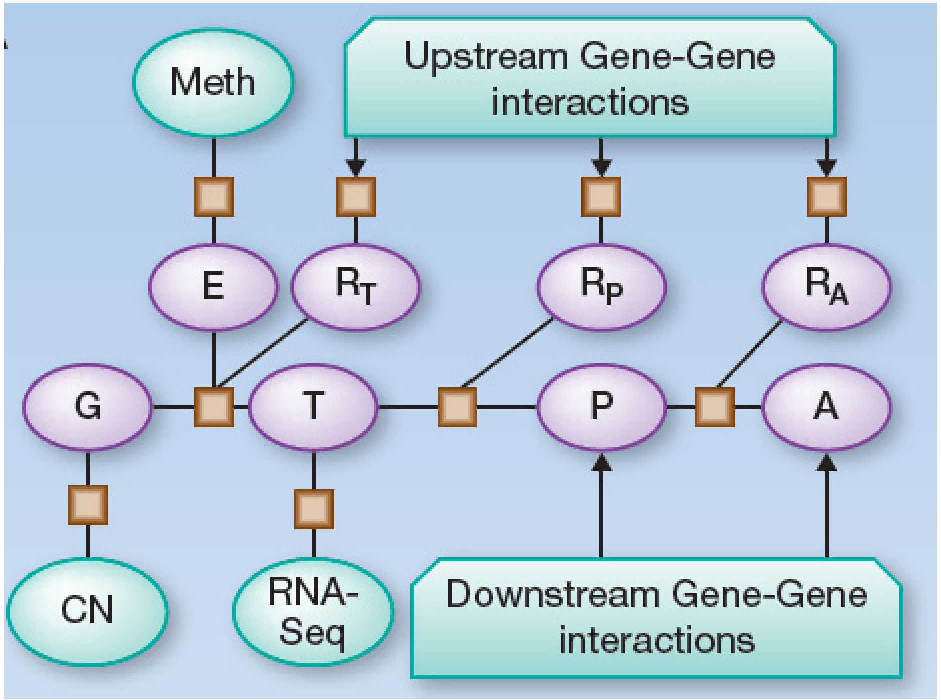
$$P(x_i = a, D|\Phi) = \frac{1}{Z}\prod_{j=1}^{m}\sum_{\mathbf{S} \sqsubset_{A_i(a) \cup D} X_j} \phi_j(\mathbf{S})$$


A
Transcriptional Regulation
Translational Regulation, Protein Degradation
Intracellular and Extracellular Signaling
GeneCopy Number
Expression State
Protein Level
Protein Activity
Array CGH, SNP chips
Transcriptomics
Variable
Factor - interaction term

Courtesy of Vaske et al. License: CC-BY.
Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." Bioinformatics 26, no. 12 (2010): i237-i45.

Vaske C J et al. Bioinformatics 2010;26:i237-i245

- genomic copies (G)
- epigenetic promoter state (E)
- mRNA transcripts (T)
- peptide (P)
- active protein (A).
- Regulation gene expression
  - transcriptional (RT)
  - translational (RP)
  - post-translational (RA)

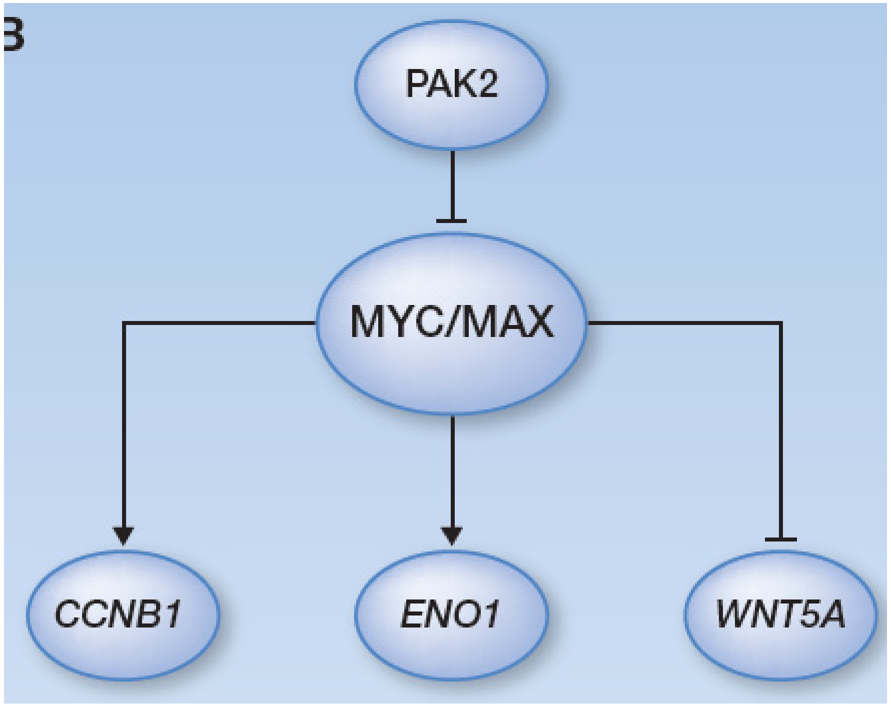

Meth
Upstream Gene-Gene interactions
E
R_T
R_P
R_A
G
T
P
A
CN
RNA-Seq
Downstream Gene-Gene interactions

Source: Goldstein, Theodore C., Evan O. Paull, et al. "Molecular Pathways: Extracting Medical Knowledge from High-throughput Genomic Data." *Clinical Cancer Research* 19, no. 12 (2013): 3114-20.

Molecular Pathways: Extracting Medical Knowledge from High Throughput Genomic Data

Theodore Goldstein, Evan O Paull, Matthew J Ellis, et al.

Clin Cancer Res Published OnlineFirst February 21, 2013.

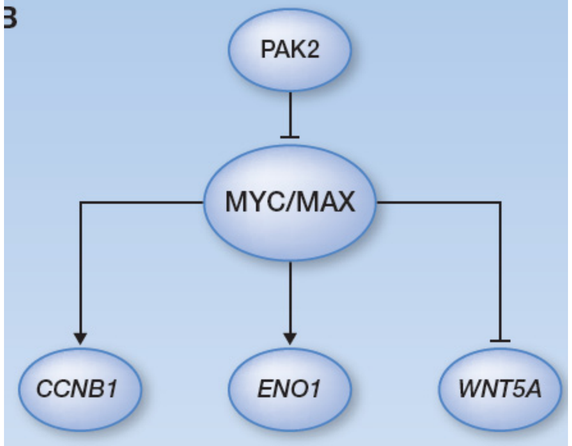


Source: Goldstein, Theodore C., Evan O. Paull, et al. "Molecular Pathways: Extracting Medical Knowledge from High-throughput Genomic Data." *Clinical Cancer Research* 19, no. 12 (2013): 3114-20.

Molecular Pathways: Extracting Medical Knowledge from High Throughput Genomic Data

Theodore Goldstein, Evan O Paull, Matthew J Ellis, et al.

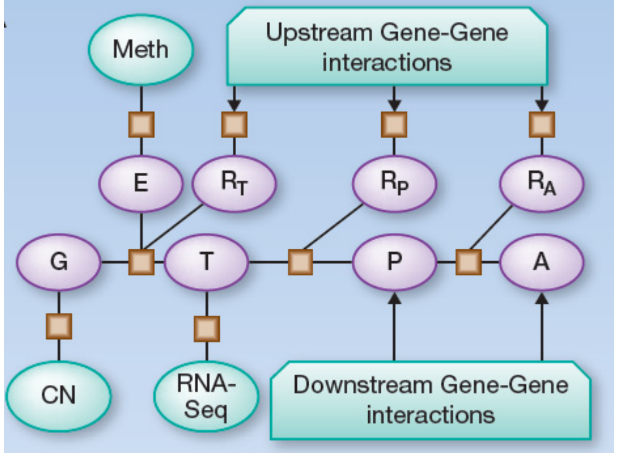
Clin Cancer Res Published OnlineFirst February 21, 2013.

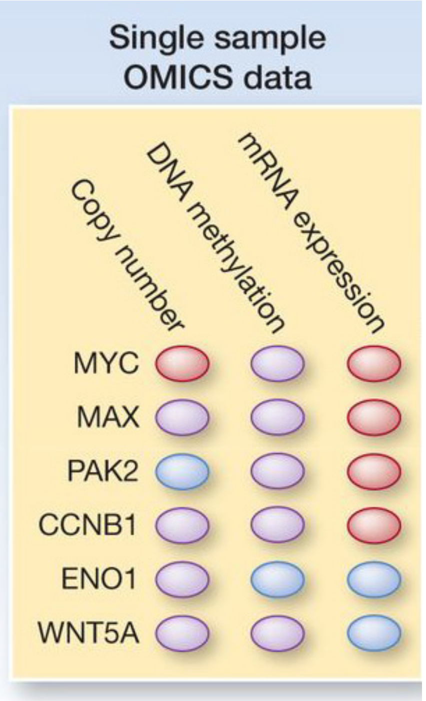


🔴 increased

🟣 normal

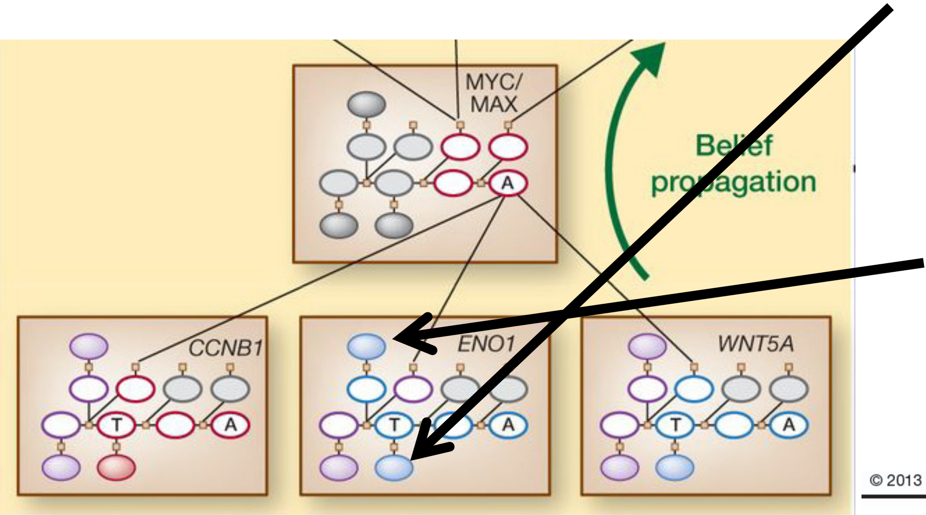
🔵 lower


Meth
Upstream Gene-Gene interactions
E
R_T
R_P
R_A
G
T
P
A
CN
RNA-Seq
Downstream Gene-Gene interactions



"MYC/MAX … is active because one of its known activated targets (CCNB1) is highly expressed while one of its repressed targets (WNT5A) has lower expression"

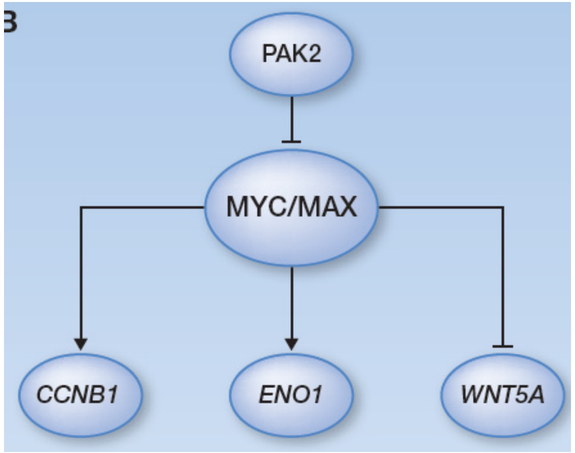
What about ENO1, which should be increasing?



Note lack of epigenetic change

© 2013 American Association for Cancer Research

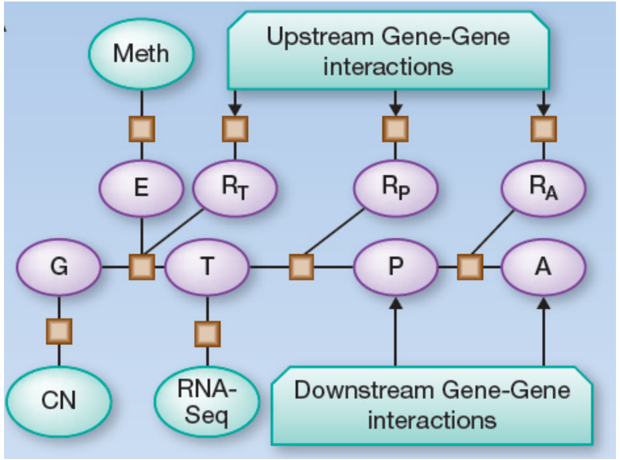
Source: Goldstein, Theodore C., Evan O. Paull, et al. "Molecular Pathways: Extracting Medical Knowledge from High-throughput Genomic Data." *Clinical Cancer Research* 19, no. 12 (2013): 3114-20.

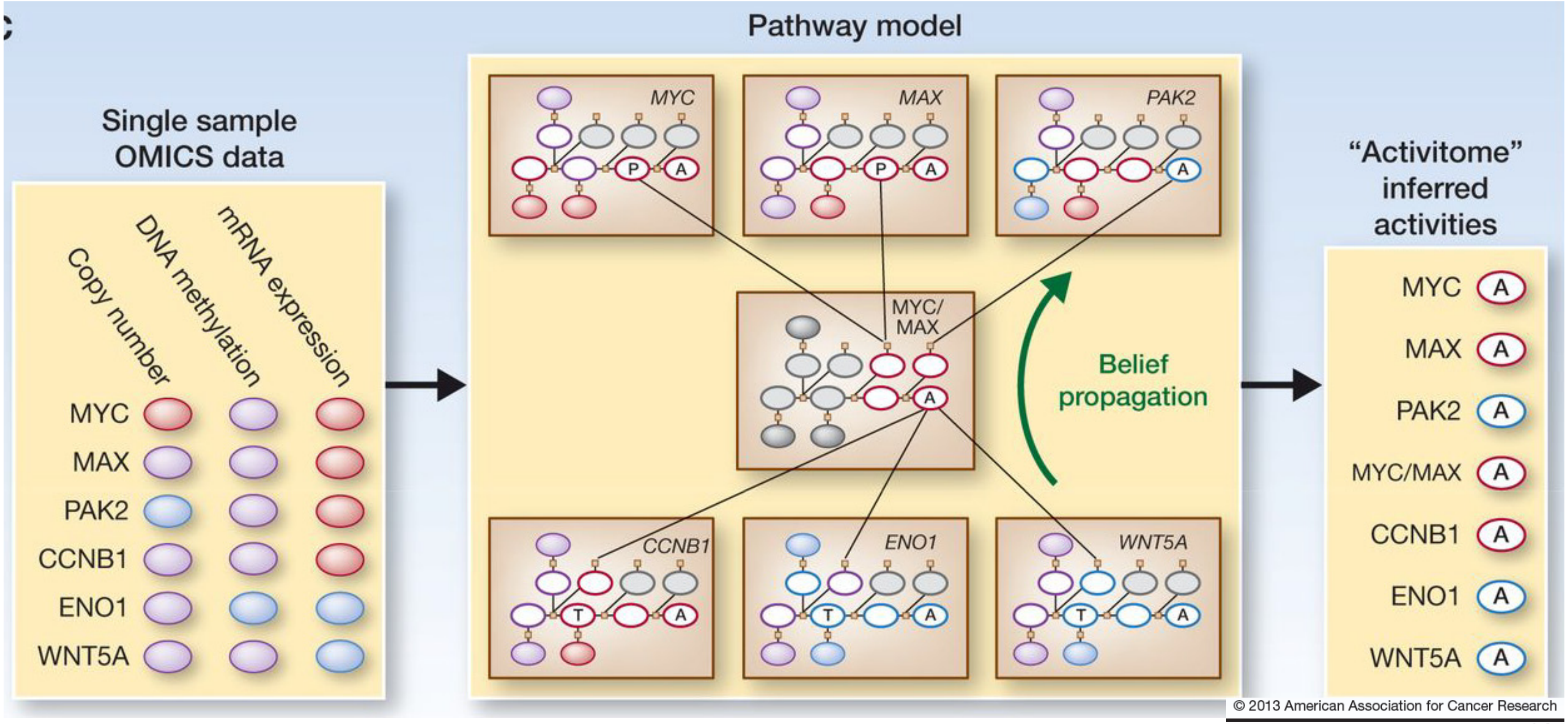


🔴 increased

🟣 normal

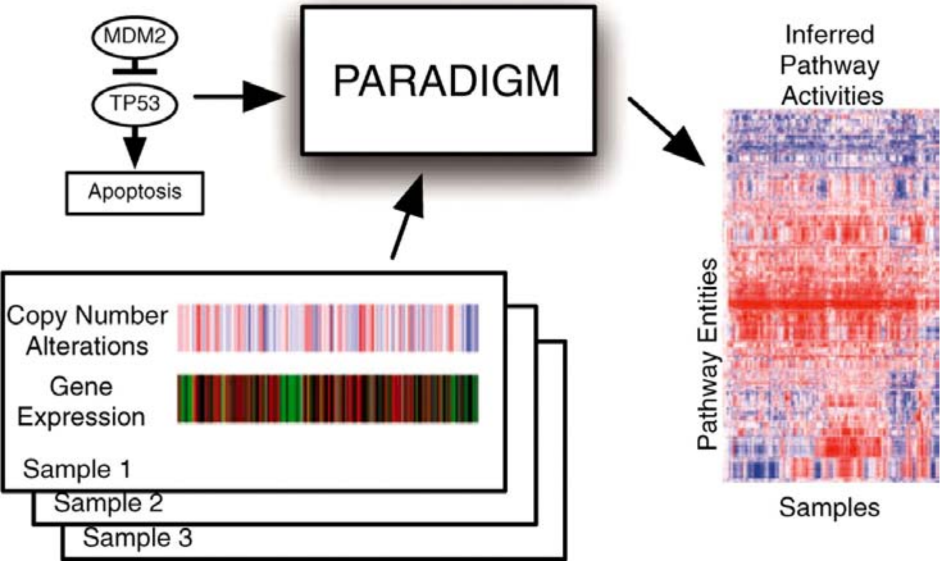
🔵 lower


Meth
Upstream Gene-Gene interactions
E
R_T
R_P
R_A
G
T
P
A
CN
RNA-Seq
Downstream Gene-Gene interactions



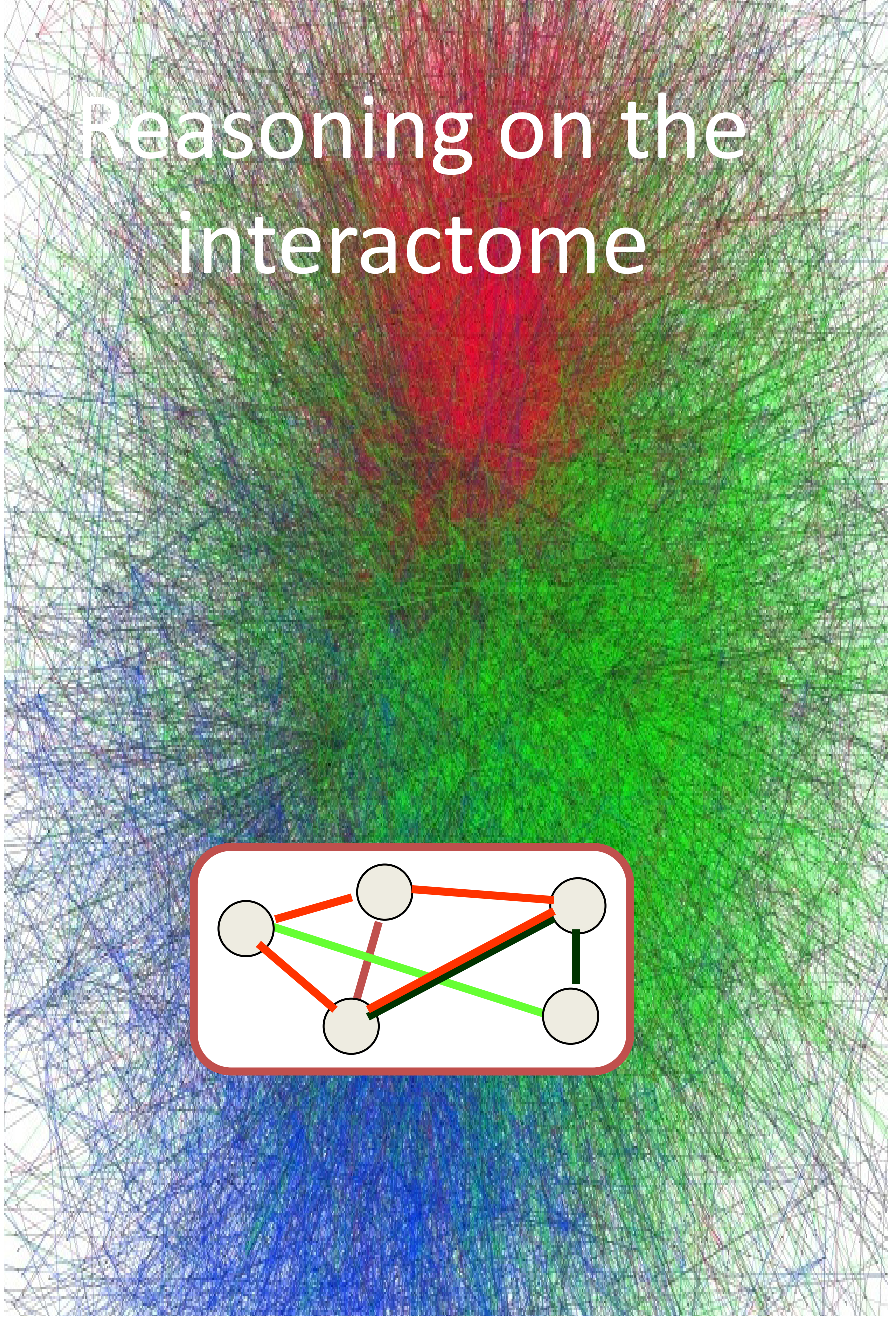
Source: Goldstein, Theodore C., Evan O. Paull, et al. "Molecular Pathways: Extracting Medical Knowledge from High-throughput Genomic Data." *Clinical Cancer Research* 19, no. 12 (2013): 3114-20.

Reasoning on curated pathways



Courtesy of Vaske et al. License: CC-BY.
Source: Vaske, Charles J., Stephen C. Benz, et al. "Inference of Patient-specific Pathway Activities from Multi-dimensional Cancer Genomics Data Using PARADIGM." *Bioinformatics* 26, no. 12 (2010): i237-i45.

Reasoning on the interactome



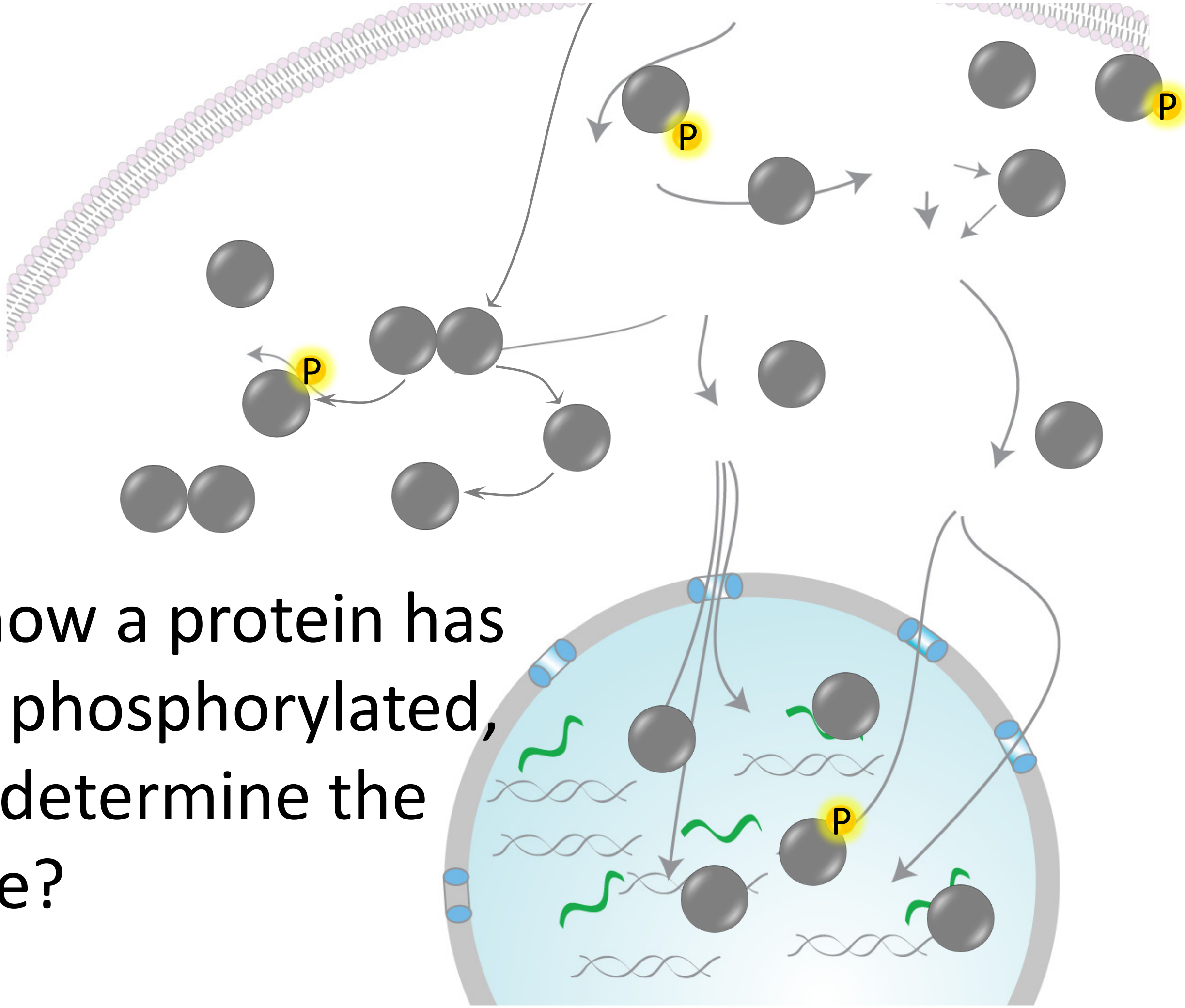
Network Models

- Structure of network
  - Coexpression
  - Mutual information
  - Physical/genetic interactions
- Analysis of network
  - Ad hoc
  - Shortest path
  - Clustering
  - Optimization

Graph Algorithms for Interaction Networks

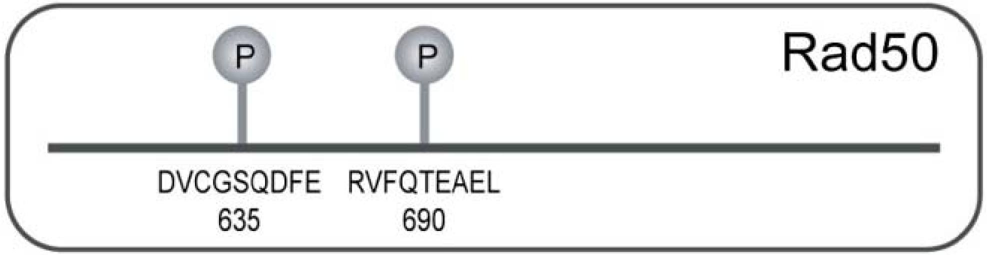
- Rich area of computer science
- Applications to Interaction Networks:
  - Distances:
    - Finding kinase substrates
  - Clustering
    - PPI->Protein complexes, functional annotation
    - Coexpression -> Modules
    - Blast ->Protein families
  - Active subnetworks
    - Finding hidden components of processes

Networkin



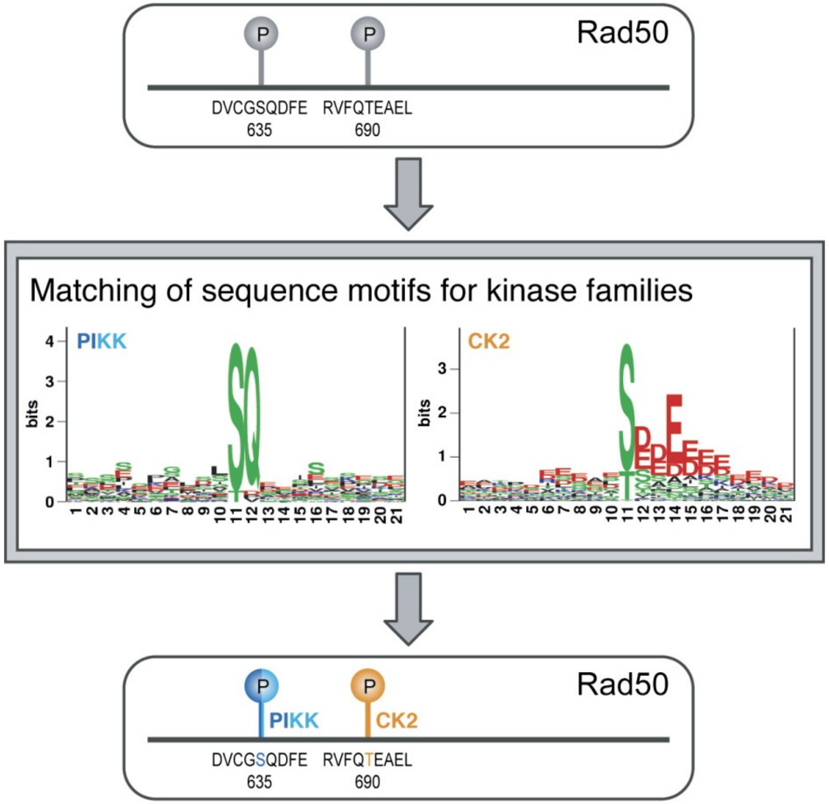
If I know a protein has been phosphorylated, can I determine the kinase?

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Yeger-Lotem, Esti, Laura Riva, et al. "Bridging High-throughput Genetic and Transcriptional Data Reveals Cellular Responses to Alpha-synuclein Toxicity."
Nature Genetics 41, no. 3 (2009): 316-23.



Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Linding, Rune, Lars Juhl Jensen, et al. "Systematic Discovery of in Vivo Phosphorylation Networks." *Cell* 129, no. 7 (2007): 1415-26.

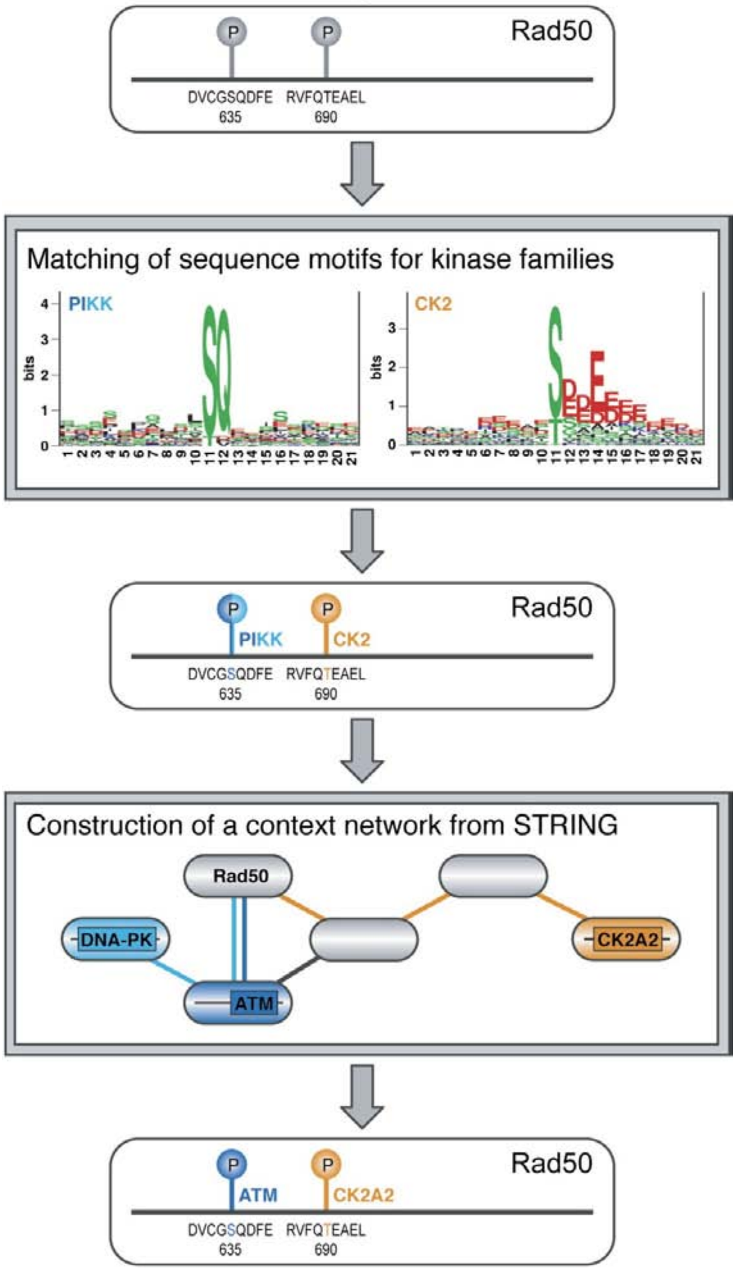
Linding *et al.* (2007) Cell. doi:10.1016/j.cell.2007.05.052



Step 1: Use sequence motifs to determine family of kinase

Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Linding, Rune, Lars Juhl Jensen, et al. "Systematic Discovery of in Vivo Phosphorylation Networks." *Cell* 129, no. 7 (2007): 1415-26.

Linding *et al.* (2007) Cell. doi:10.1016/j.cell.2007.05.052


Rad50
P
P
DVCGSQDFE 635
RVFQTEAEL 690
Matching of sequence motifs for kinase families
PIKK
CK2
bits
Construction of a context network from STRING
Rad50
DNA-PK
ATM
CK2A2
PIKK
CK2
ATM
CK2A2

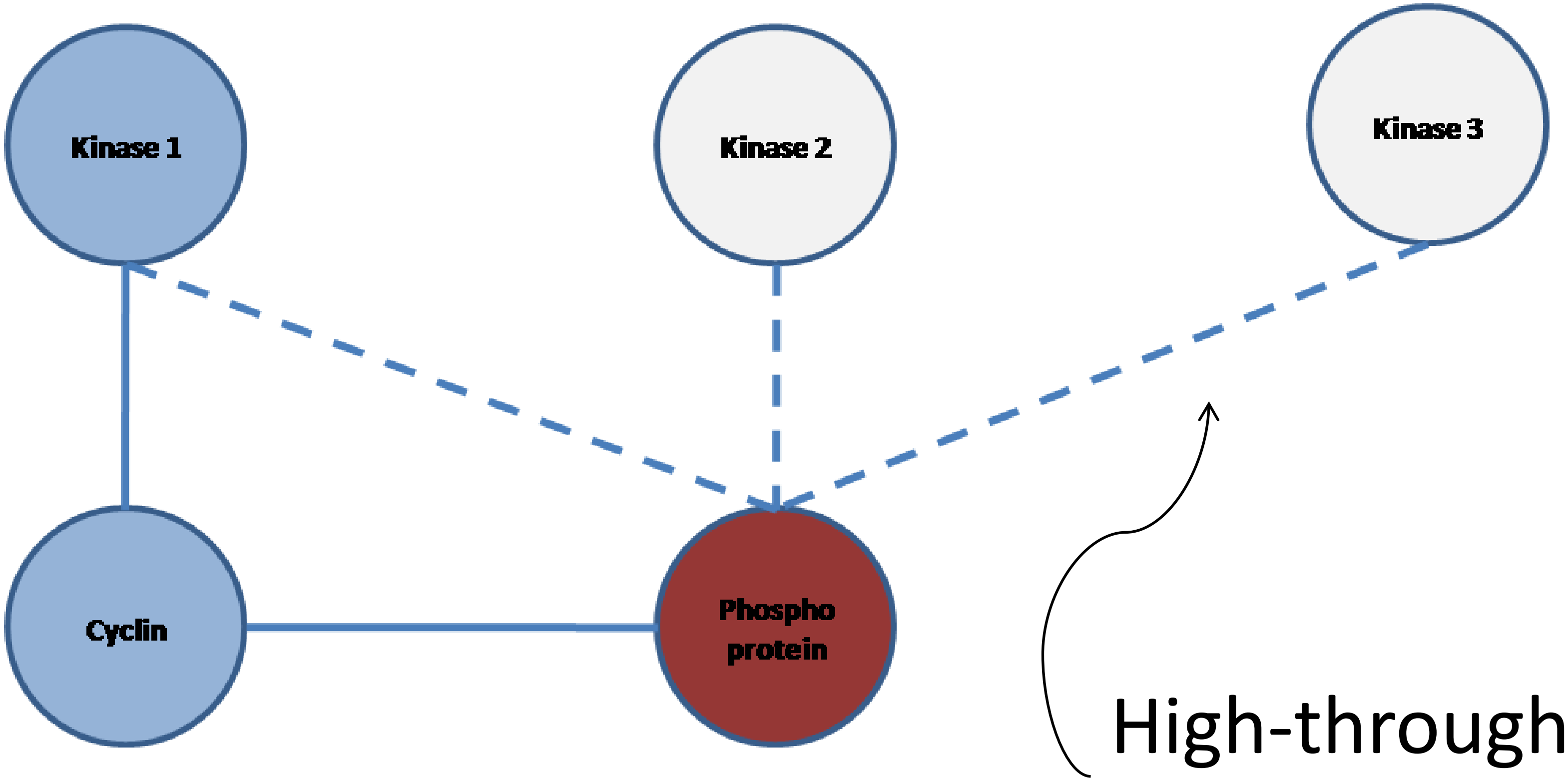
Step 1: Use sequence motifs to determine family of kinase

Step 2: Use Interactome data to find most likely family member

Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Linding, Rune, Lars Juhl Jensen, et al. "Systematic Discovery of in Vivo Phosphorylation Networks." *Cell* 129, no. 7 (2007): 1415-26.

Linding *et al.* (2007) Cell. doi:10.1016/j.cell.2007.05.052

Which is best?



High-throughput Two-hybrid assay

- - - Motif match

— High confidence interaction

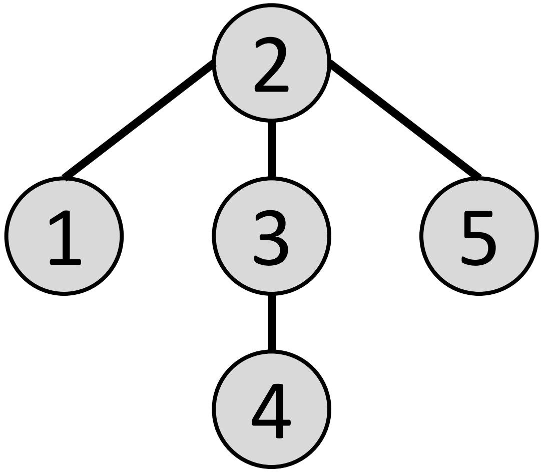
How do we find the closest kinase?

- Many efficient algorithms exist once we treat our problem as one in Graph Theory.

Graph Terminology

- G=(V,E)
- Undirected vs. directed
- Weights – numbers assigned to each edge
- Degree(v) – number of edges incident on v
    - In-degree and out-degree
- Path from a to b is a series of vertices <a, v0, …, b> where edges exist between sequential vertices
- Path length = sum of edges weights (or number of edges) on path.

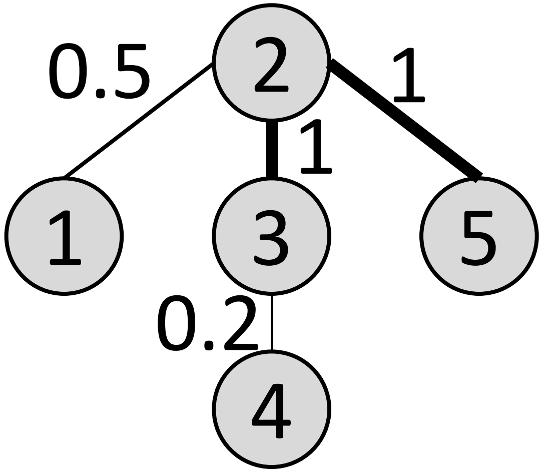
Data Structure



Adjacency Matrix

	1	2	3	4	5
1	0	1	0	0	0
2	1	0	1	0	1
3	0	1	0	1	0
4	0	0	1	0	0
5	0	1	0	0	0

Data Structure



Weights can represent our confidence in the link

Adjacency Matrix

	1	2	3	4	5
1	0	.5	0	0	0
2	.5	0	1	0	1
3	0	1	0	.2	0
4	0	0	.2	0	0
5	0	1	0	0	0

Weighted graph:

$a_{ij}=w_{ij}$ if edge exists; 0 otherwise

Shortest Path Algorithms

- Efficient Algorithms for
  - single pair (u,v)
  - single source/destination to all other nodes
  - all-pairs

Reliability of edges

- Assign weight to each edge based on reliability.
- Total distance in network = sum of edge weights
- If $\text{weight}_{ij}=-\log(P_{ij})$:

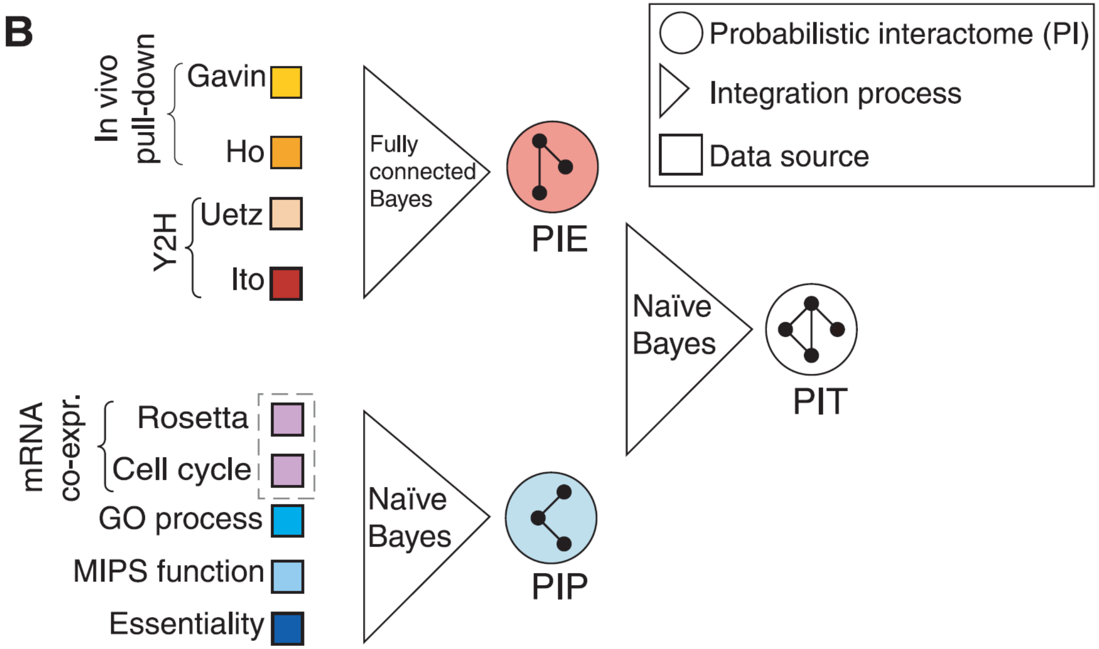
$$\begin{aligned} \min\Sigma w_{ij} &= \min(-\log \Pi P_{ij}) \\ &= \max(\text{joint probability}) \\ &= \text{most probable path} \end{aligned}$$

Interaction Weights

- How do we assign reliability of edges?

A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data

Ronald Jansen,[1*] Haiyuan Yu,[1] Dov Greenbaum,[1] Yuval Kluger,[1] Nevan J. Krogan,[4] Sambath Chung,[1,2] Andrew Emili,[4] Michael Snyder,[2] Jack F. Greenblatt,[4] Mark Gerstein[1,3†]

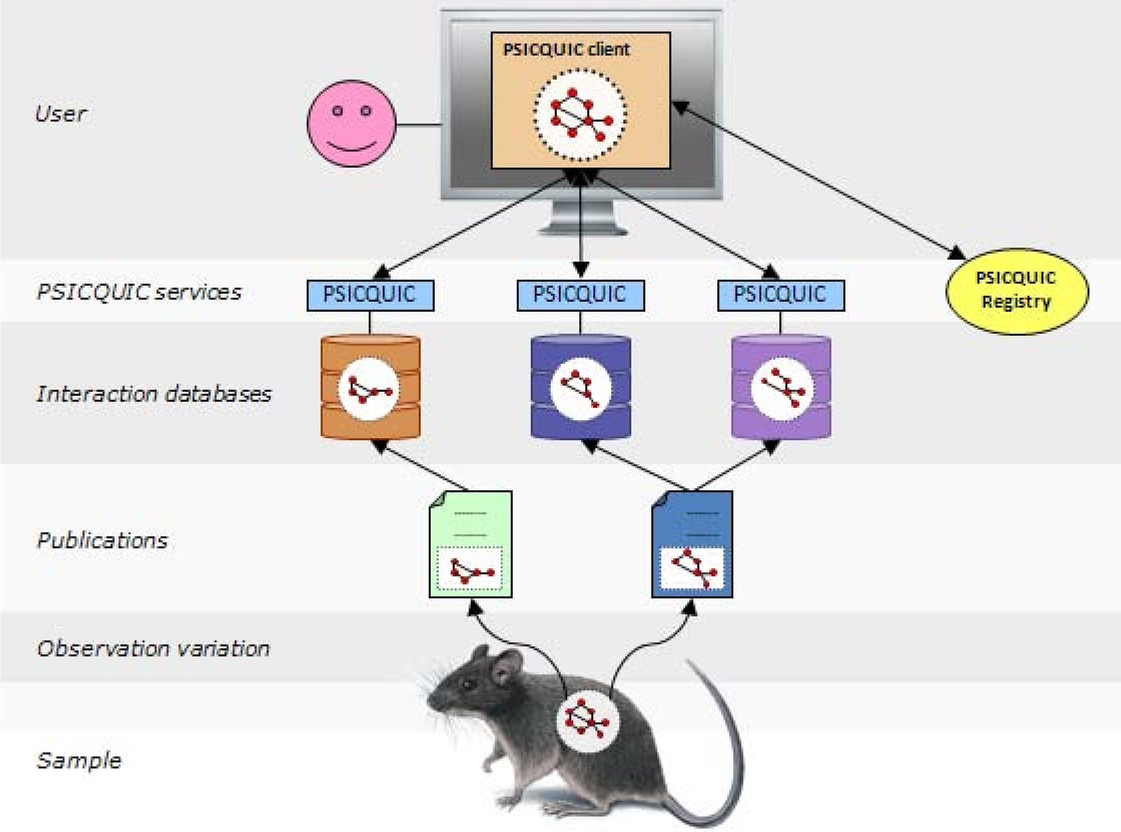


Source: Jansen, Ronald, Haiyuan Yu, et al. "A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data." *Science* 302, no. 5644 (2003): 449-53.

SCIENCE VOL 302 17 OCTOBER 2003

http://www.sciencemag.org/content/302/5644/449.abstract

PSICQUIC and PSISCORE: accessing and scoring molecular interactions
Nature Methods 8, 528–529 (2011)
doi:10.1038/nmeth.1637

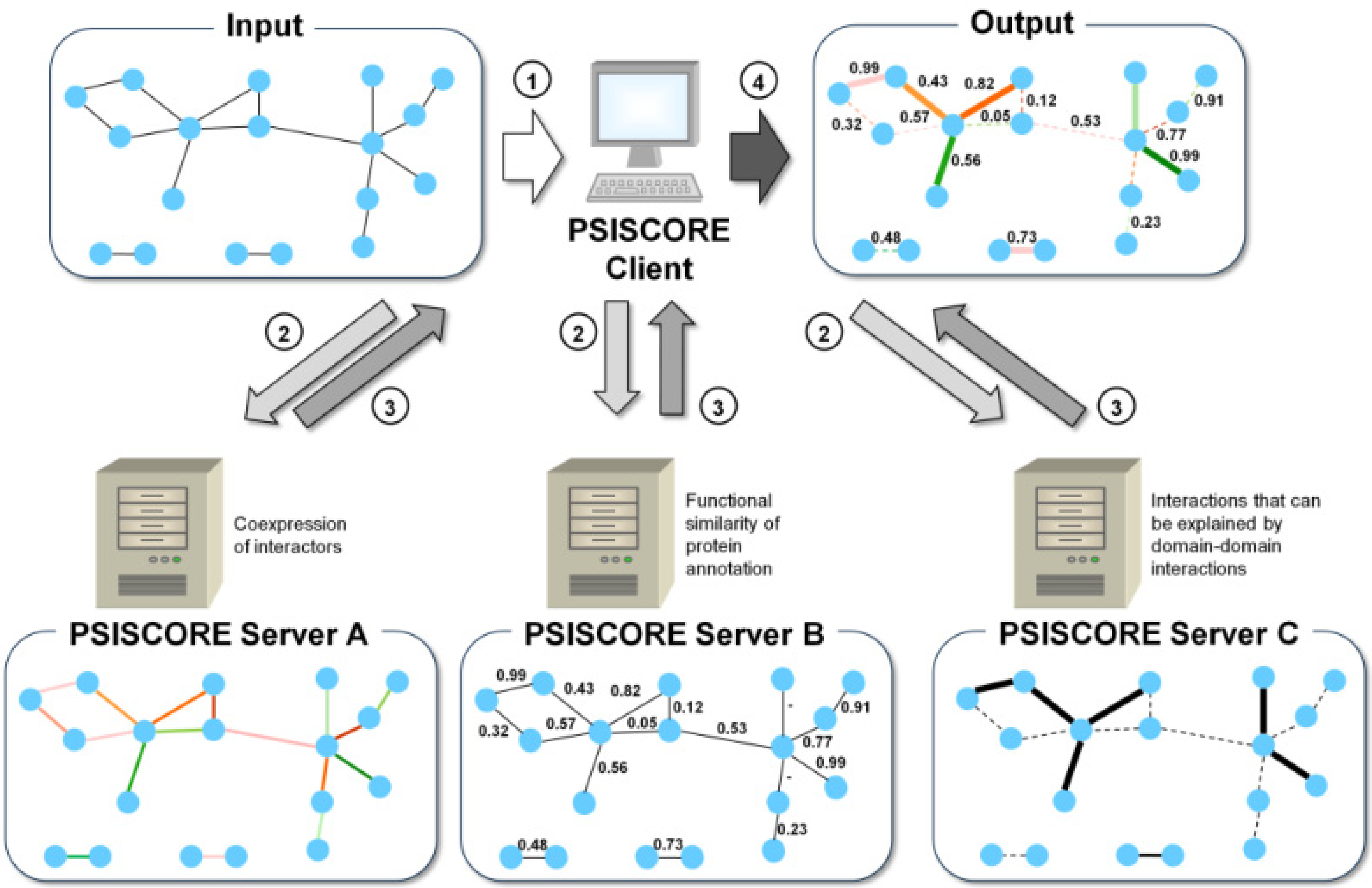


Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Aranda, Bruno, Hagen Blankenburg, et al. "PSICQUIC and PSISCORE: Accessing and Scoring Molecular Interactions." *Nature Methods* 8, no. 7 (2011): 528-9.

Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) released the PSI molecular interaction (MI) XML format

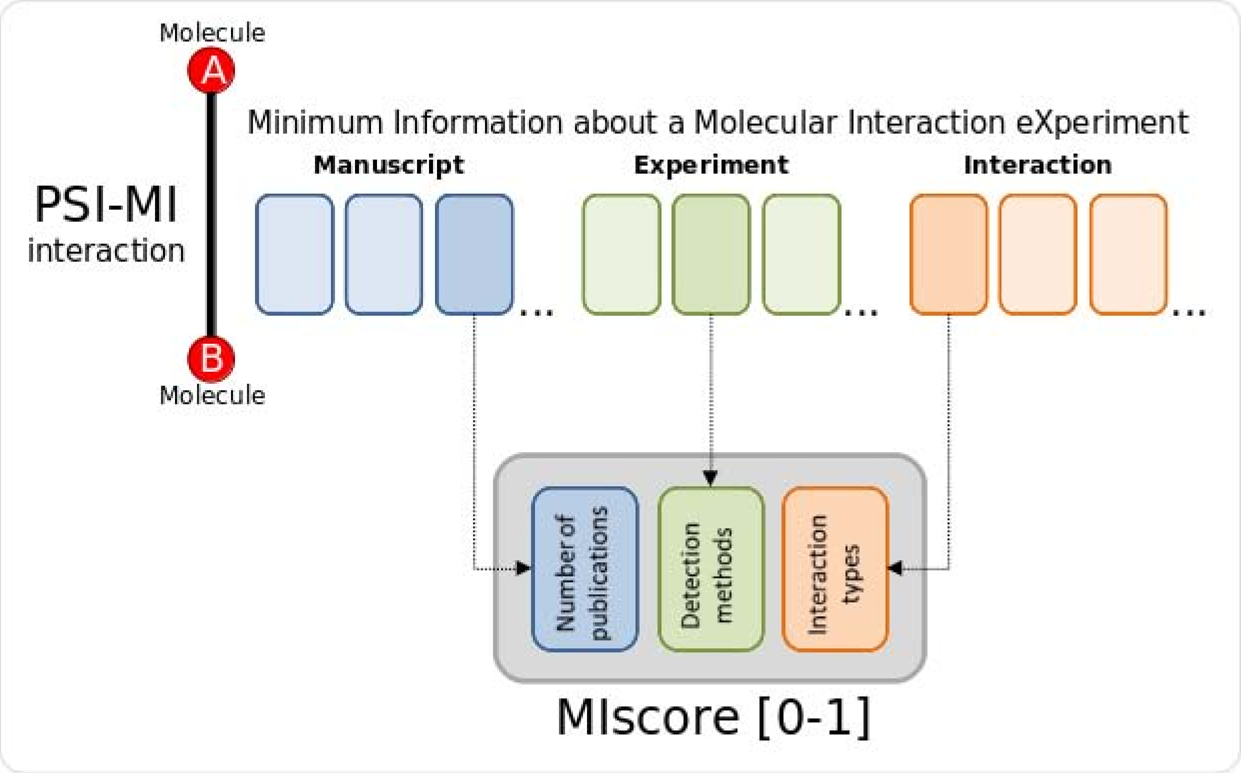
PSI common query interface (PSICQUIC), a community standard for computational access to molecular-interaction data resources.

http://www.nature.com/nmeth/journal/v8/n7/full/nmeth.1637.html



http://www.nature.com/nmeth/journal/v8/n7/full/nmeth.1637.html

Miscore algorithm



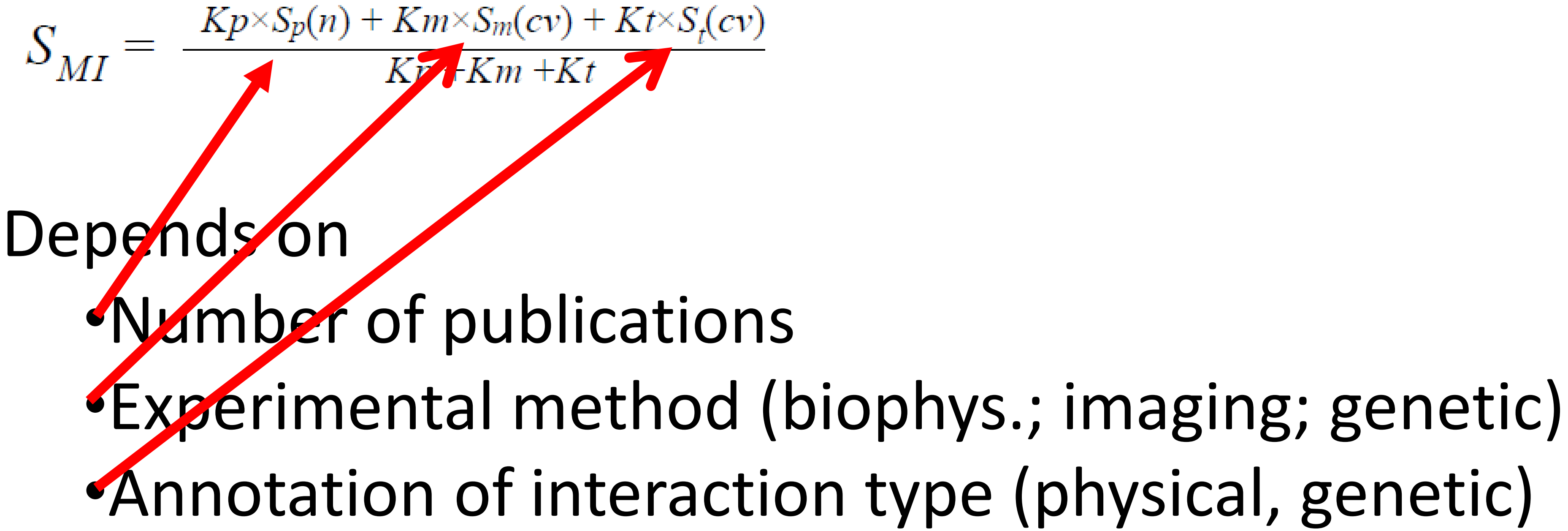
Courtesy of Miscore. Used with permission.

Miscore is a normalized score between 0 and 1 that takes into account several variables:

- Number of publications
- Experimental detection methods found for the interaction
- Interaction types found for the interaction

Each of these variables is also represented by a score between 0 and 1. The importance of each variable in the main equation can be adjusted using a weight factor.

Miscore algorithm

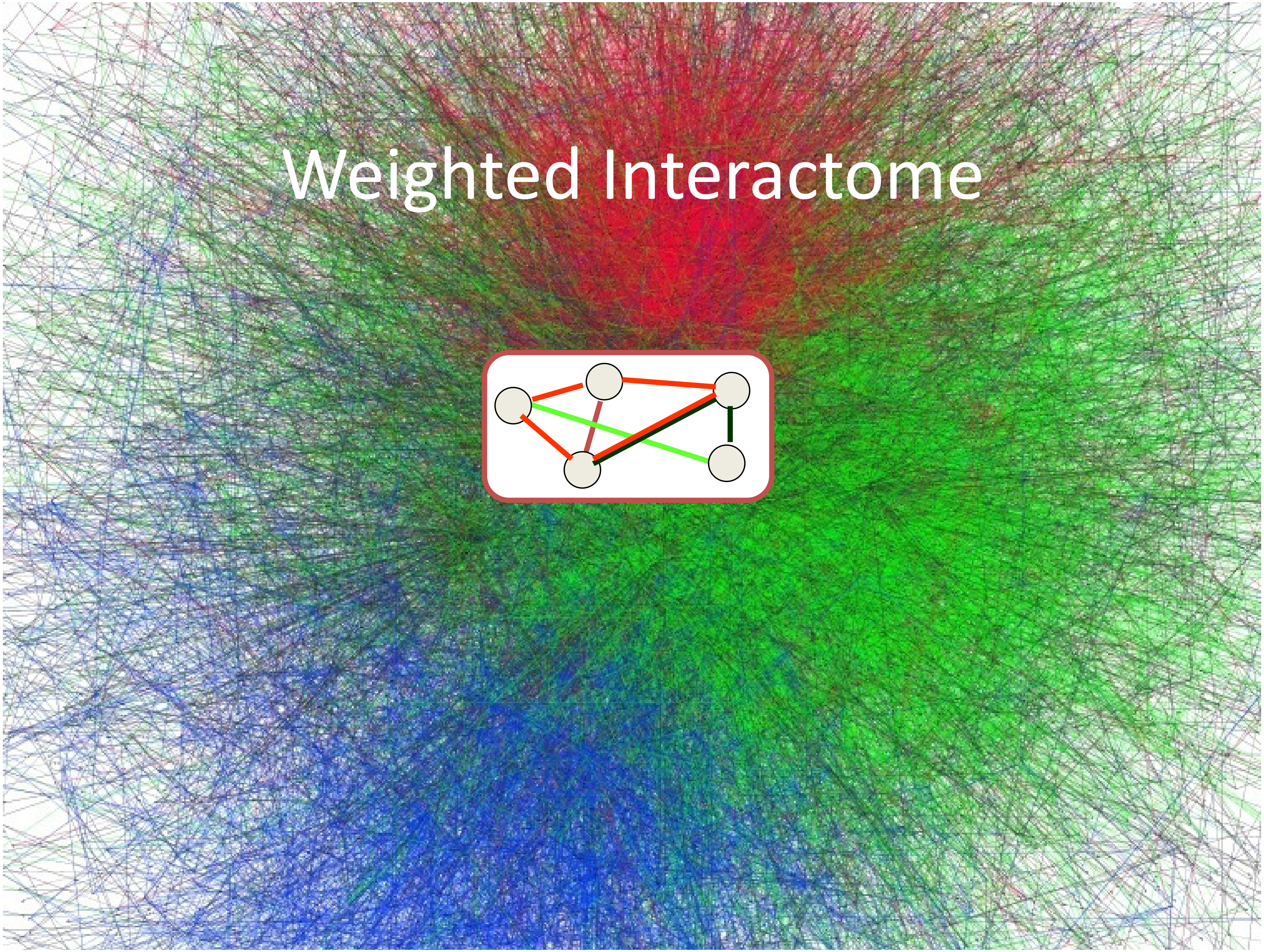


$$S_{MI} = \frac{Kp \times S_p(n) + Km \times S_m(cv) + Kt \times S_t(cv)}{Kp + Km + Kt}$$

Depends on

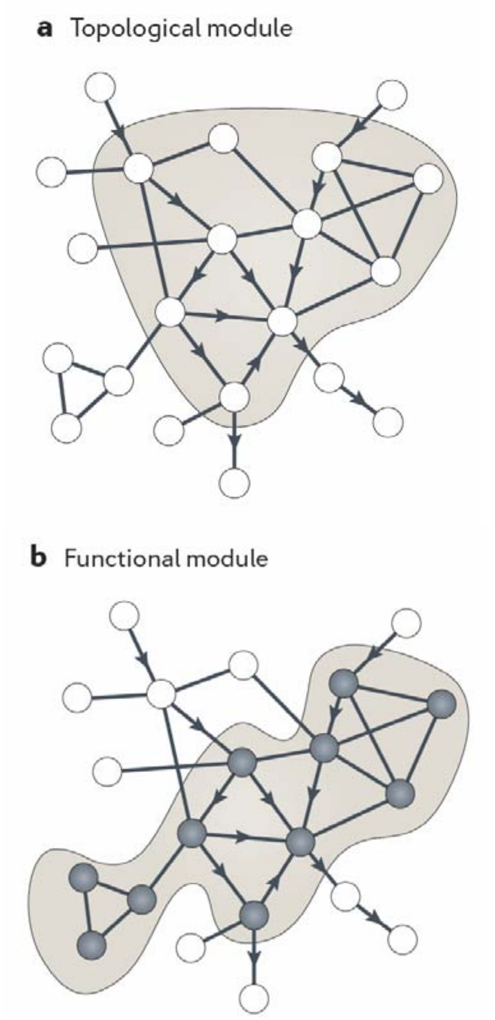
- Number of publications
- Experimental method (biophys.; imaging; genetic)
- Annotation of interaction type (physical, genetic)

Weighted Interactome



Finding Modules

- Topological module:
  - locally dense
  - more connections among nodes in module than with nodes outside module
- Functional module:
  - high density of functionally related nodes

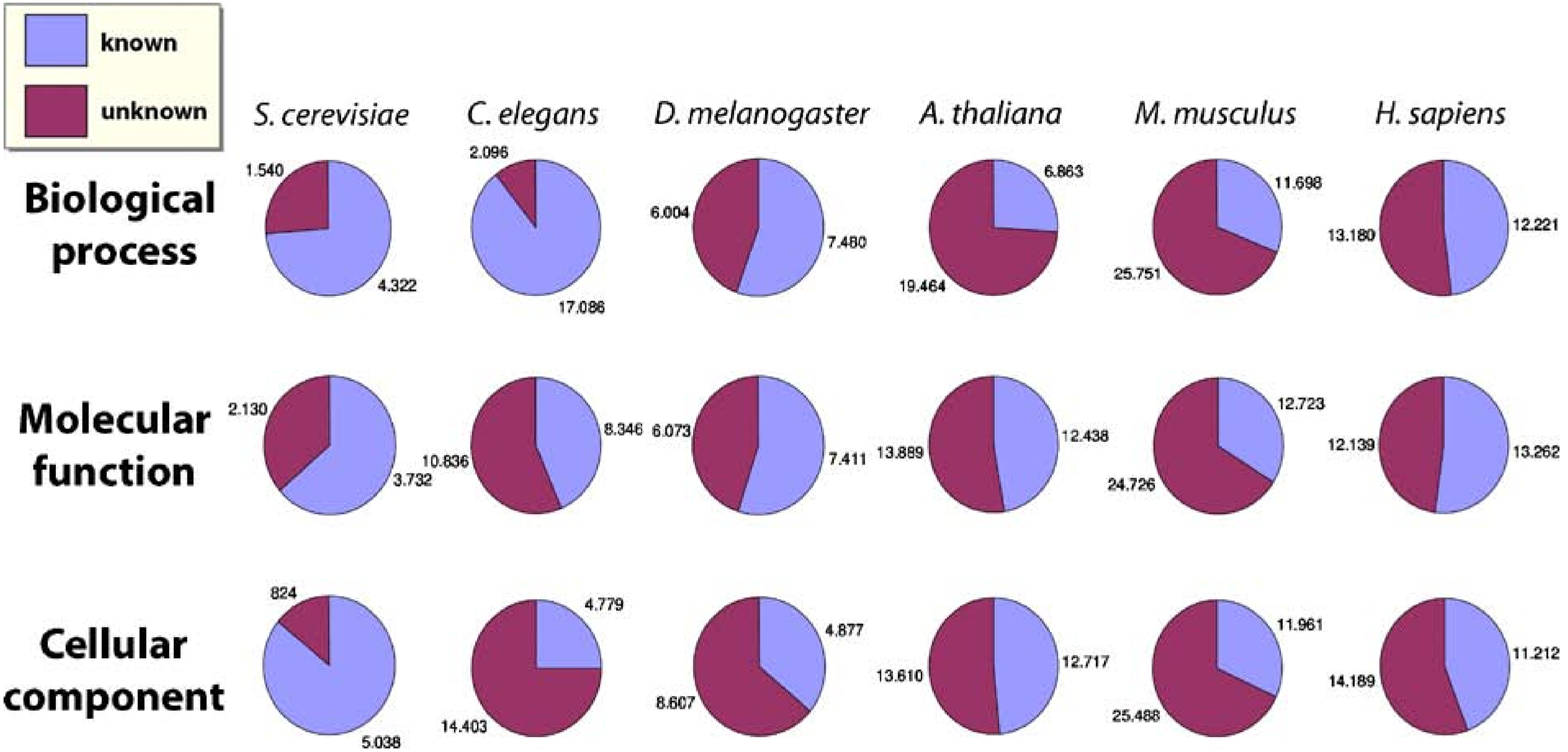


a Topological module

b Functional module

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Barabási, Albert-László, Natali Gulbahce, et al. "Network Medicine: A Network-based Approach to Human Disease." *Nature Reviews Genetics* 12, no. 1 (2011): 56-68.

Can we use networks to predict function

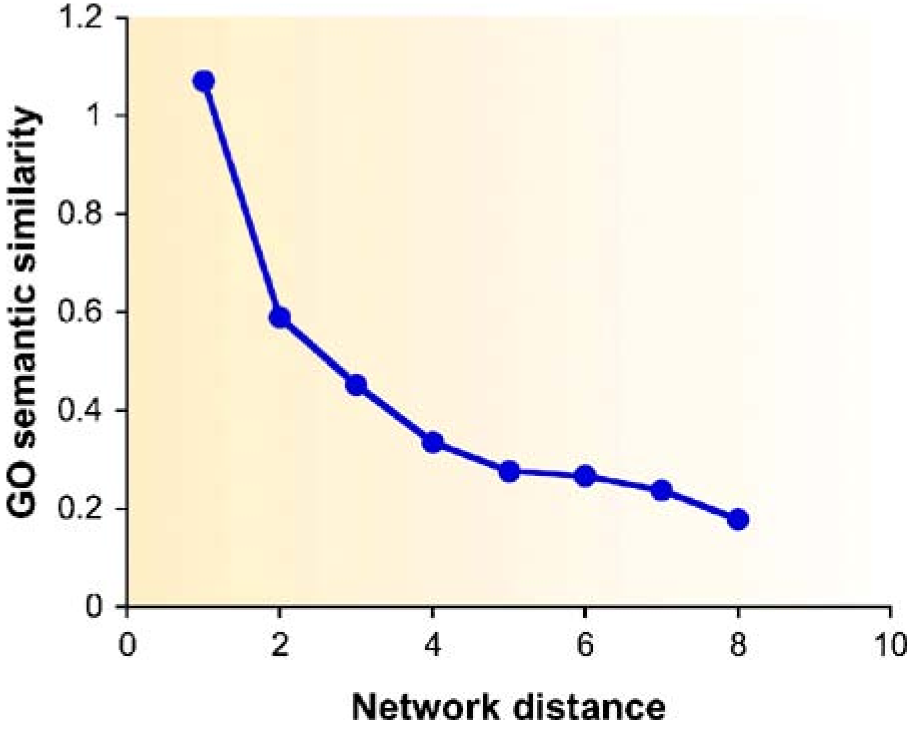


Courtesy of EMBO. Used with permission.
Source: Sharan, Roded, Igor Ulitsky, et al. "Network-based Prediction of Protein Function." *Molecular Systems Biology* 3, no. 1 (2007).

Network-based prediction of protein function
Roded Sharan, Igor Ulitsky & Ron Shamir
doi:10.1038/msb4100129

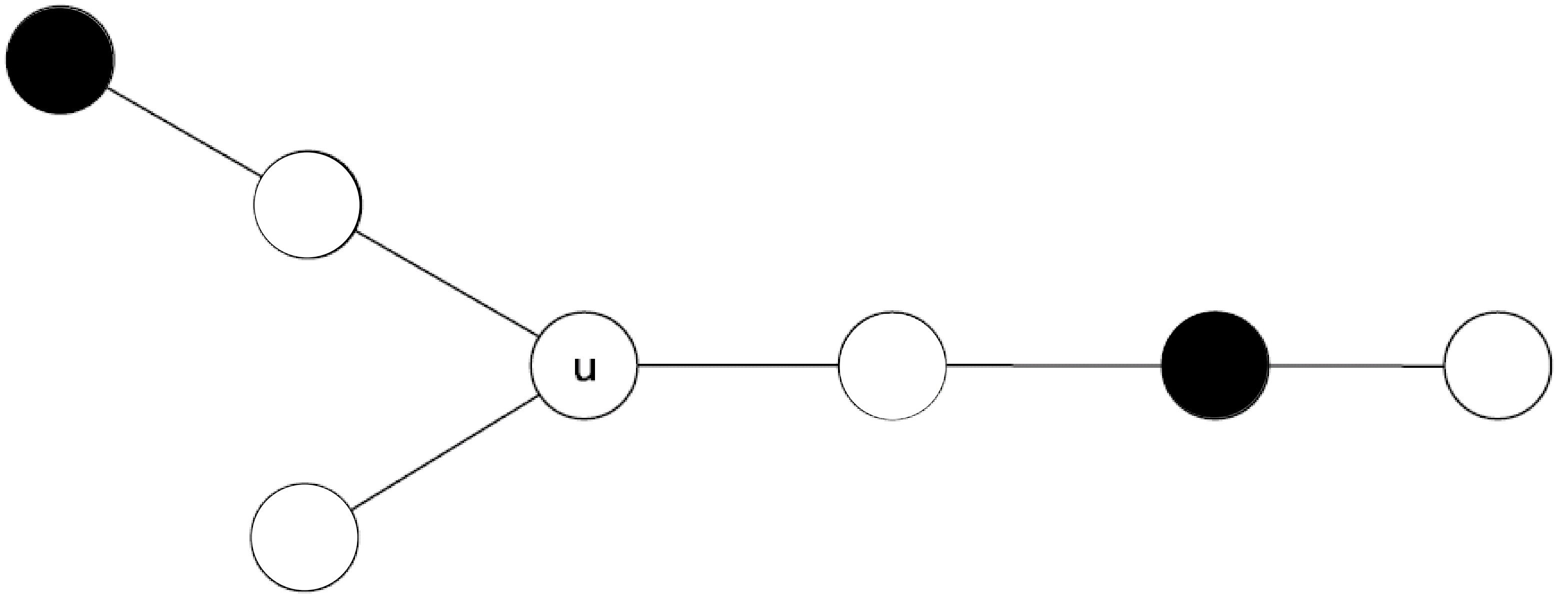
based on the Entrez Gene and the WormBase databases as of September 2006

Can we use networks to predict function

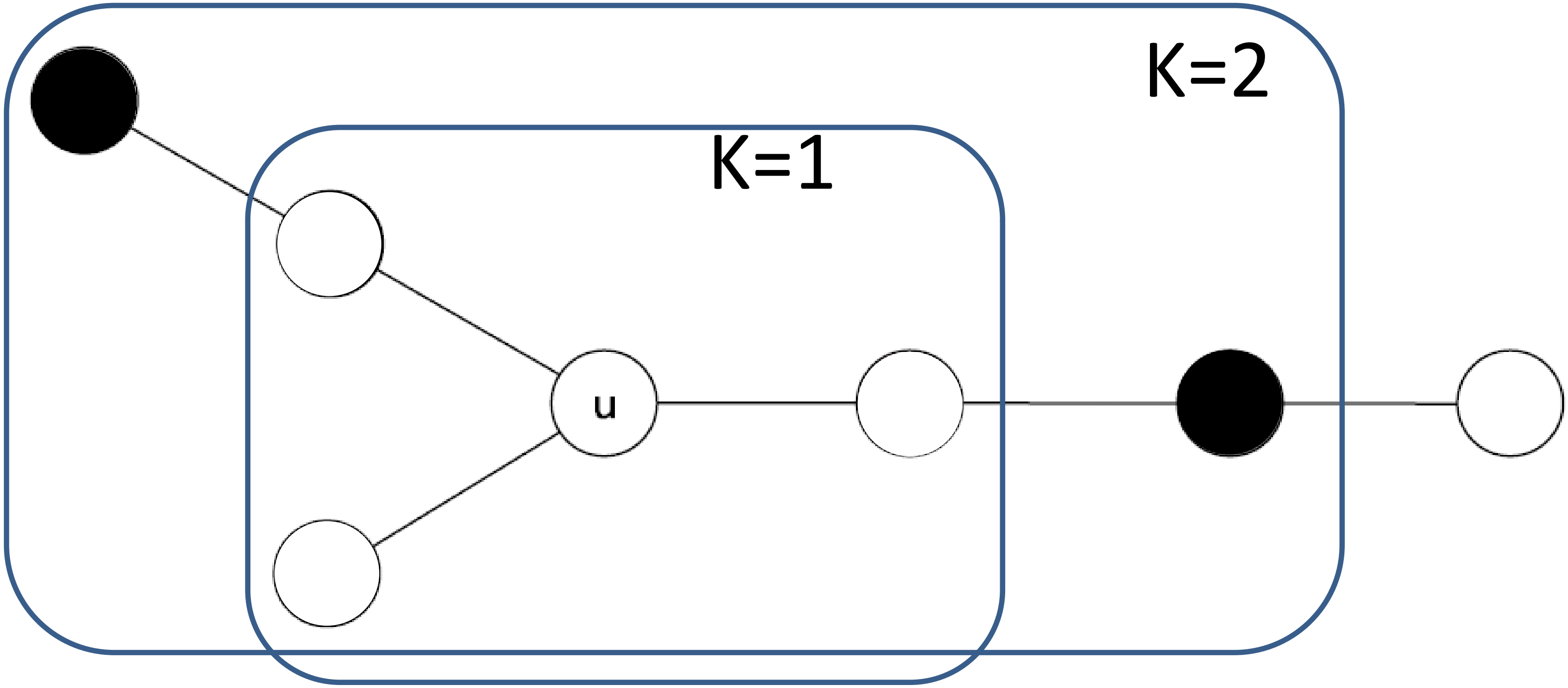


Courtesy of EMBO. Used with permission.
Source: Sharan, Roded, Igor Ulitsky, et al. "Network-based Prediction of Protein Function." *Molecular Systems Biology* 3, no. 1 (2007).

Network-based prediction of protein function
Roded Sharan, Igor Ulitsky & Ron Shamir
doi:10.1038/msb4100129

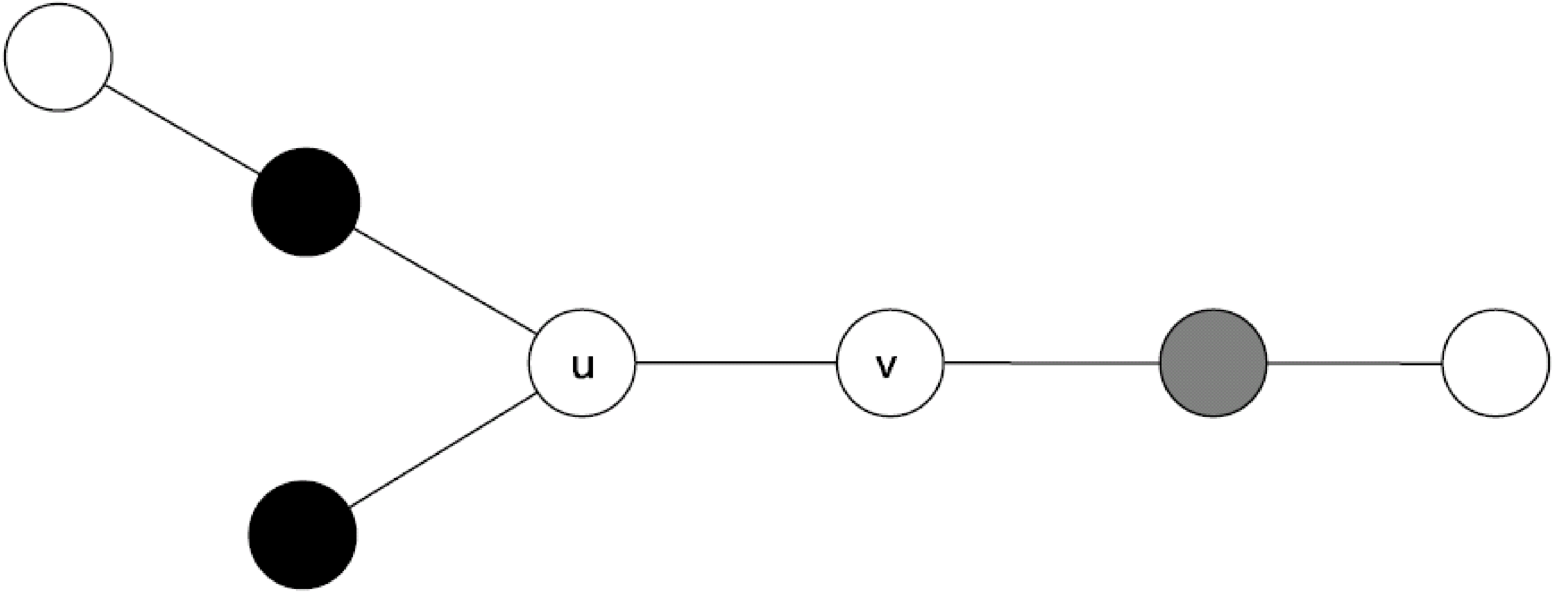


Systematically deduce the annotation of unknown nodes *u* from the known (filled) nodes

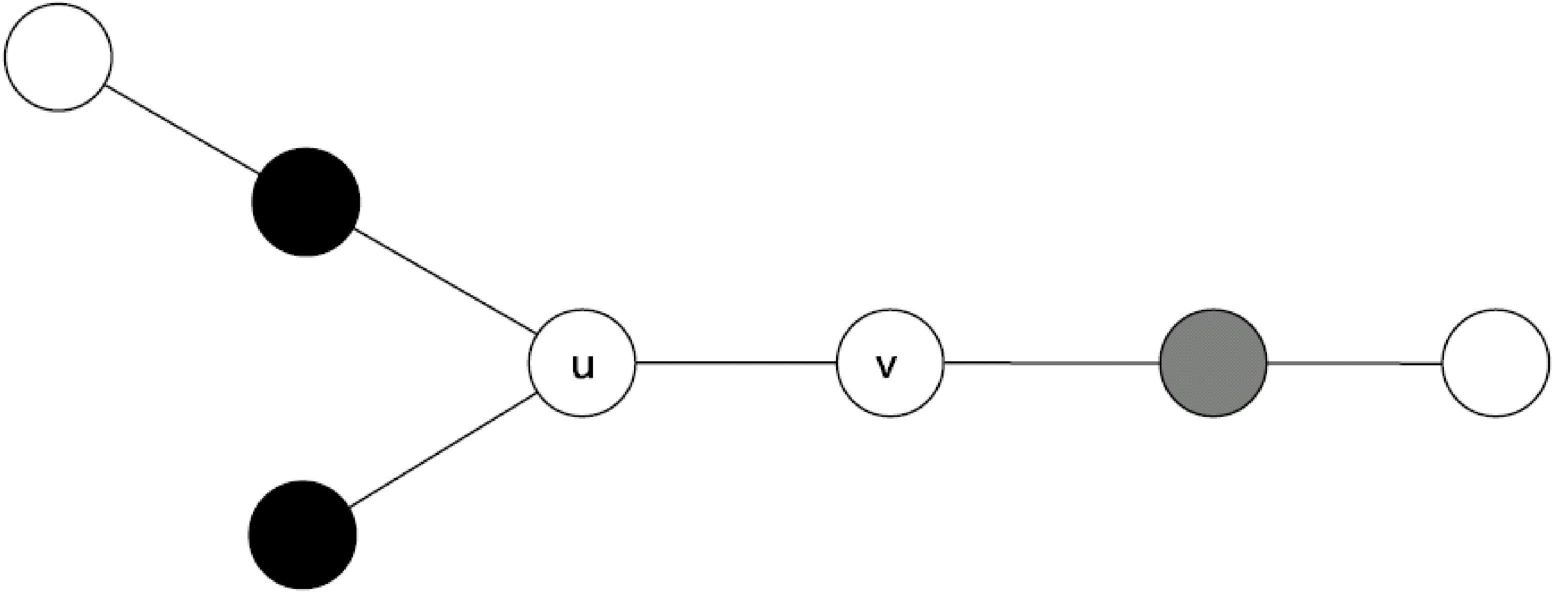


“Direct” method for gene annotation

- K-nearest neighbors
  - assume that a node has the same function as its neighbors



Should *u* and *v* have the same annotation?



Advantages of kNN approach:

- very easy to compute

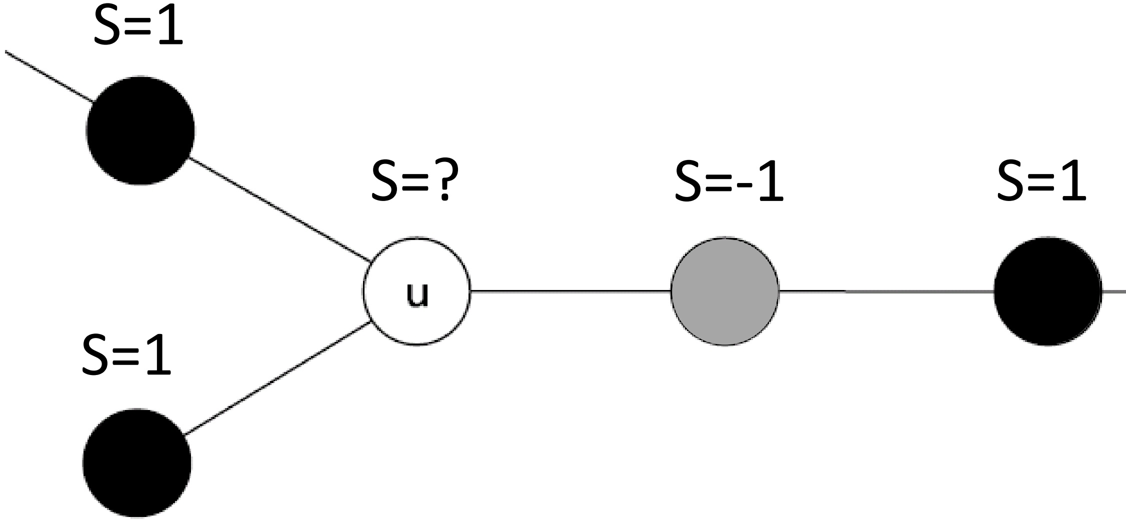
Disadvantages:

- how do you choose the best annotation?

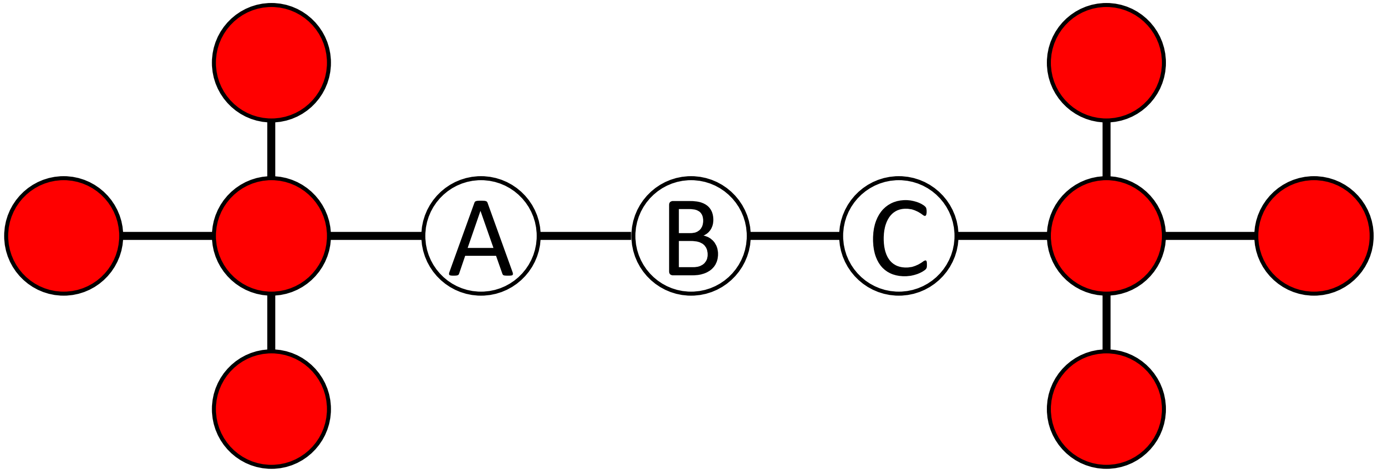
"Direct"

Local search (Karaoz[2004]):

- For each annotation:
  - S_v=1 if v has the annotation, -1 otherwise
  - Procedure: for each unassigned node u, set S_u maximize $\Sigma S_u S_v$ for all edges (u,v)
  - iterate until convergence

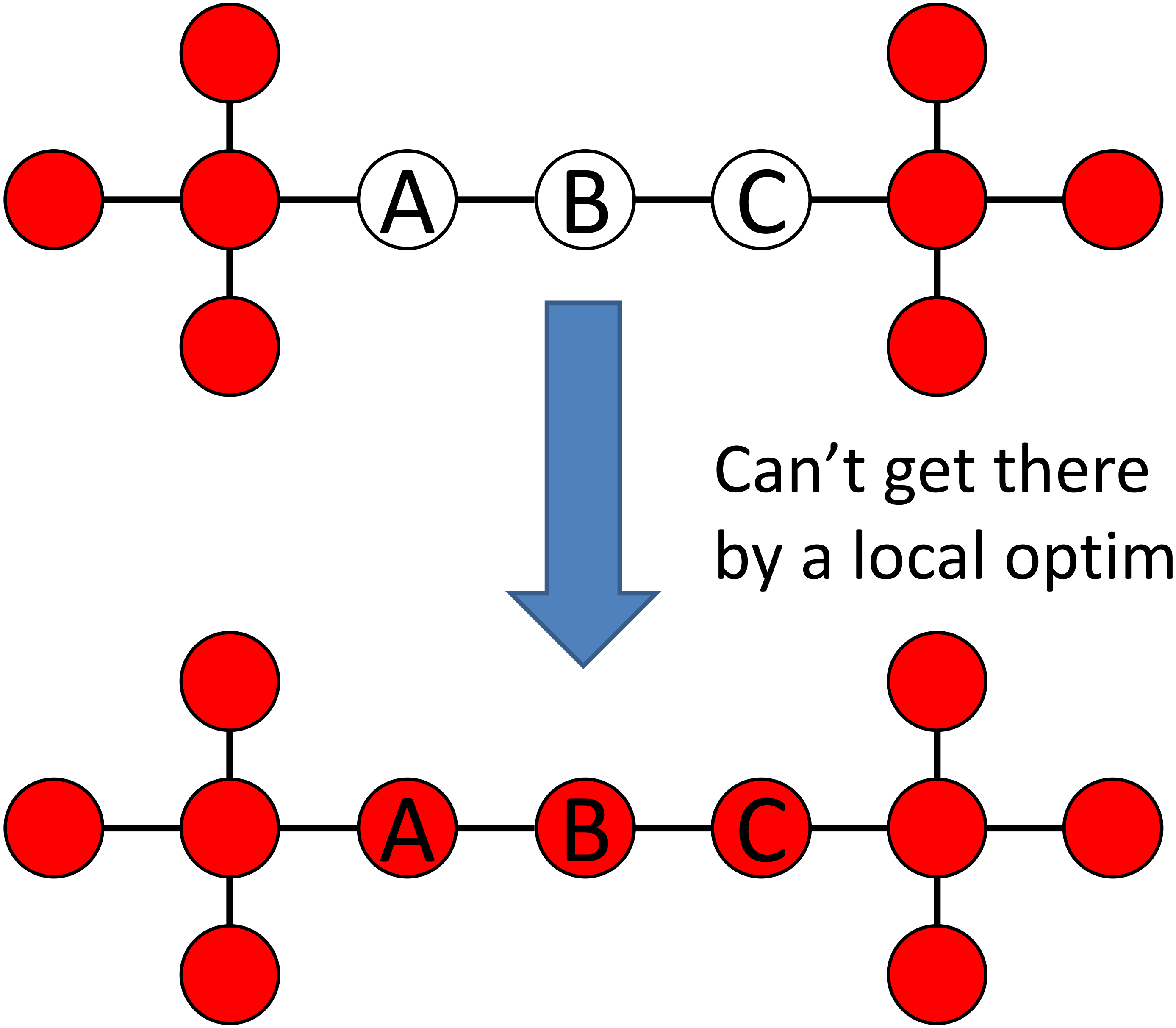


Network-based prediction of protein function
Roded Sharan, Igor Ulitsky & Ron Shamir
doi:10.1038/msb4100129

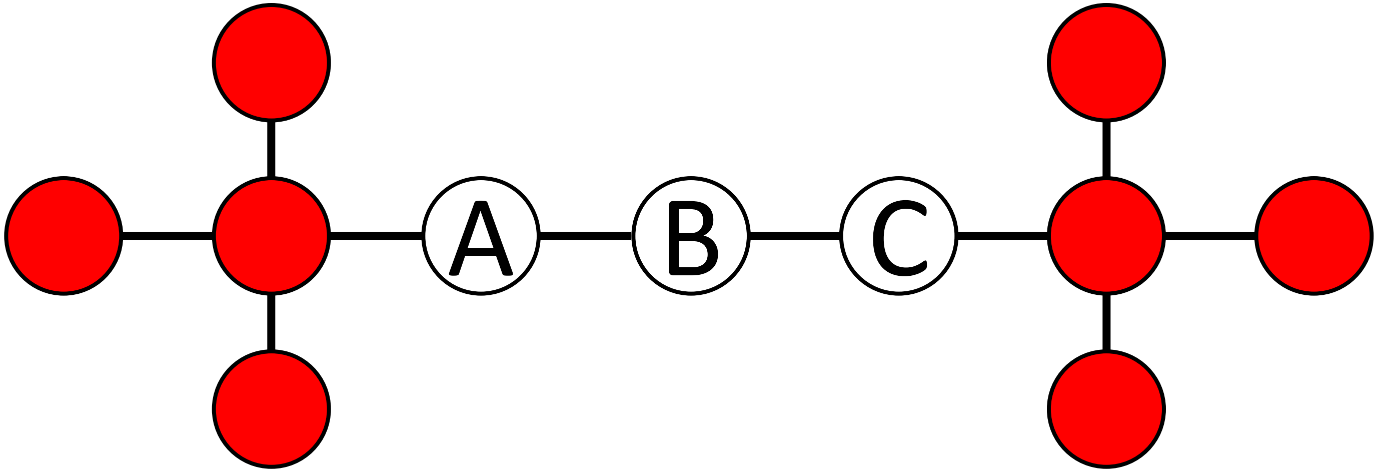


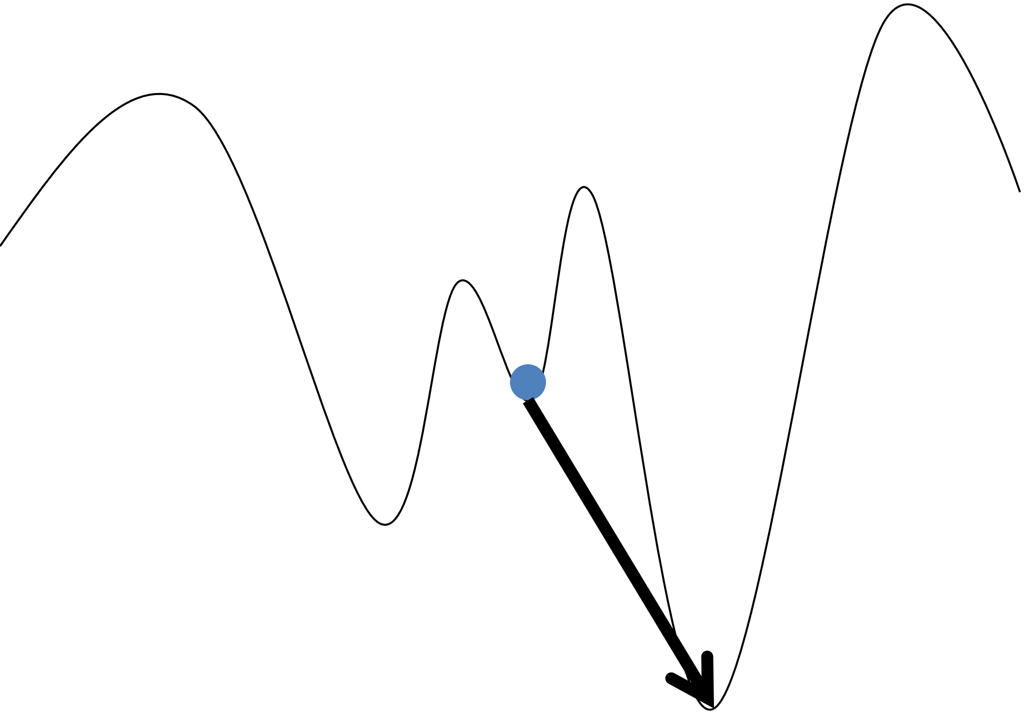
Local search may not find some good solutions.

$\Sigma S_u S_v$ does not improve if I only change A or C. Changing only B makes the score worse.

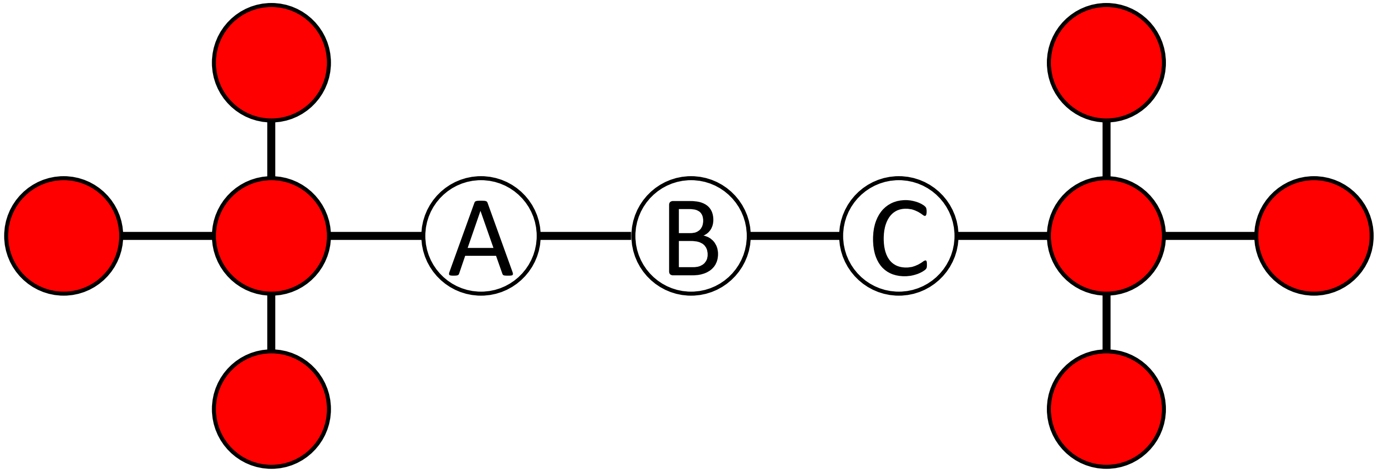


Can't get there
by a local optimization





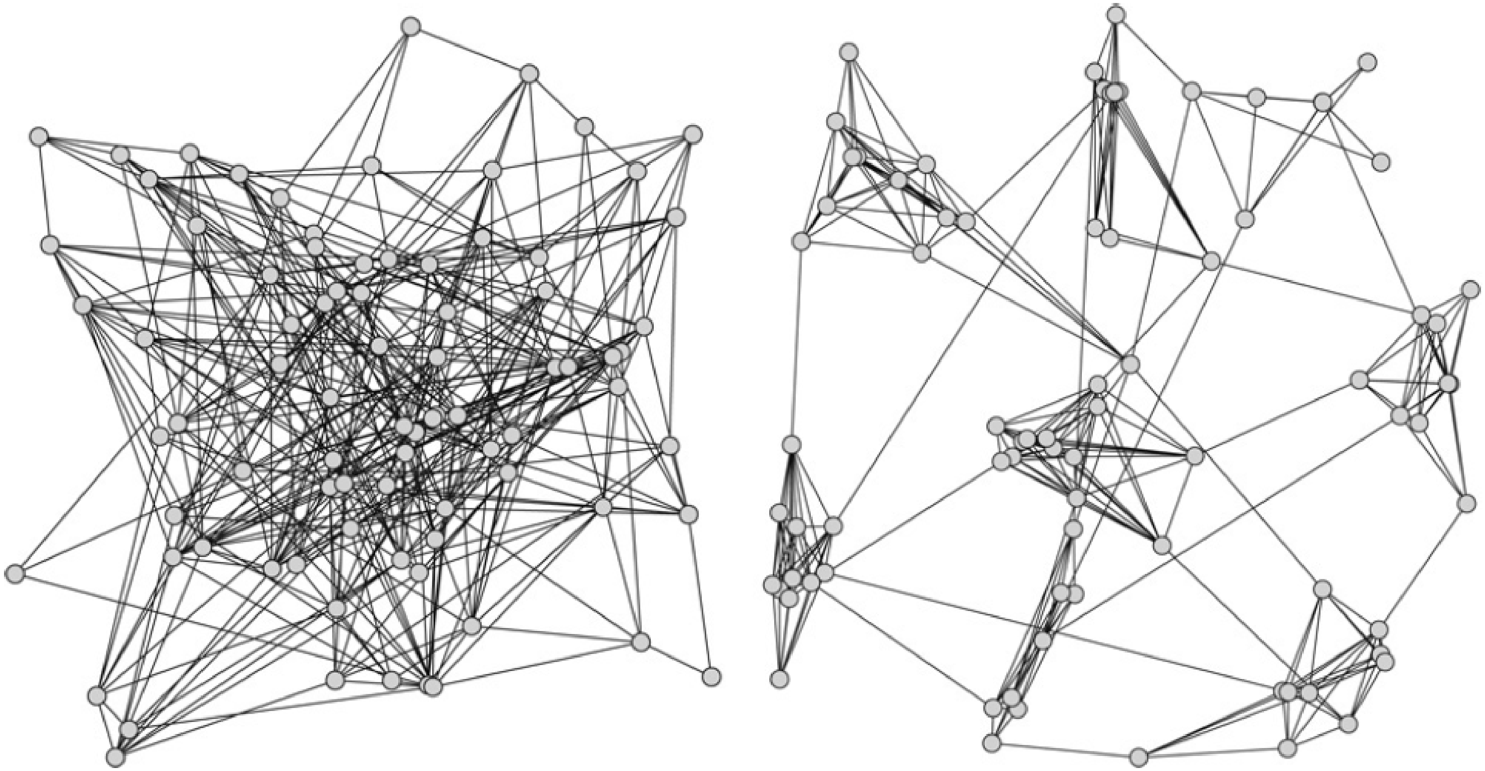
How can we move away from a locally optimal solution?



Simulated Annealing Solution:

```
•Initialize T and  subgraph Gn with score Sn
•Repeat while
    •Pick a neighboring node v to add to the subgraph
    •Score new subgraph -> Stest
    •If Sn<Stest:  keep new subgraph
    •Else keep new subgraph with
                        $P=\exp[-(Stest-Sn)/T]$
    •Modify T according to “cooling schedule.”
```

Clustering Graphs

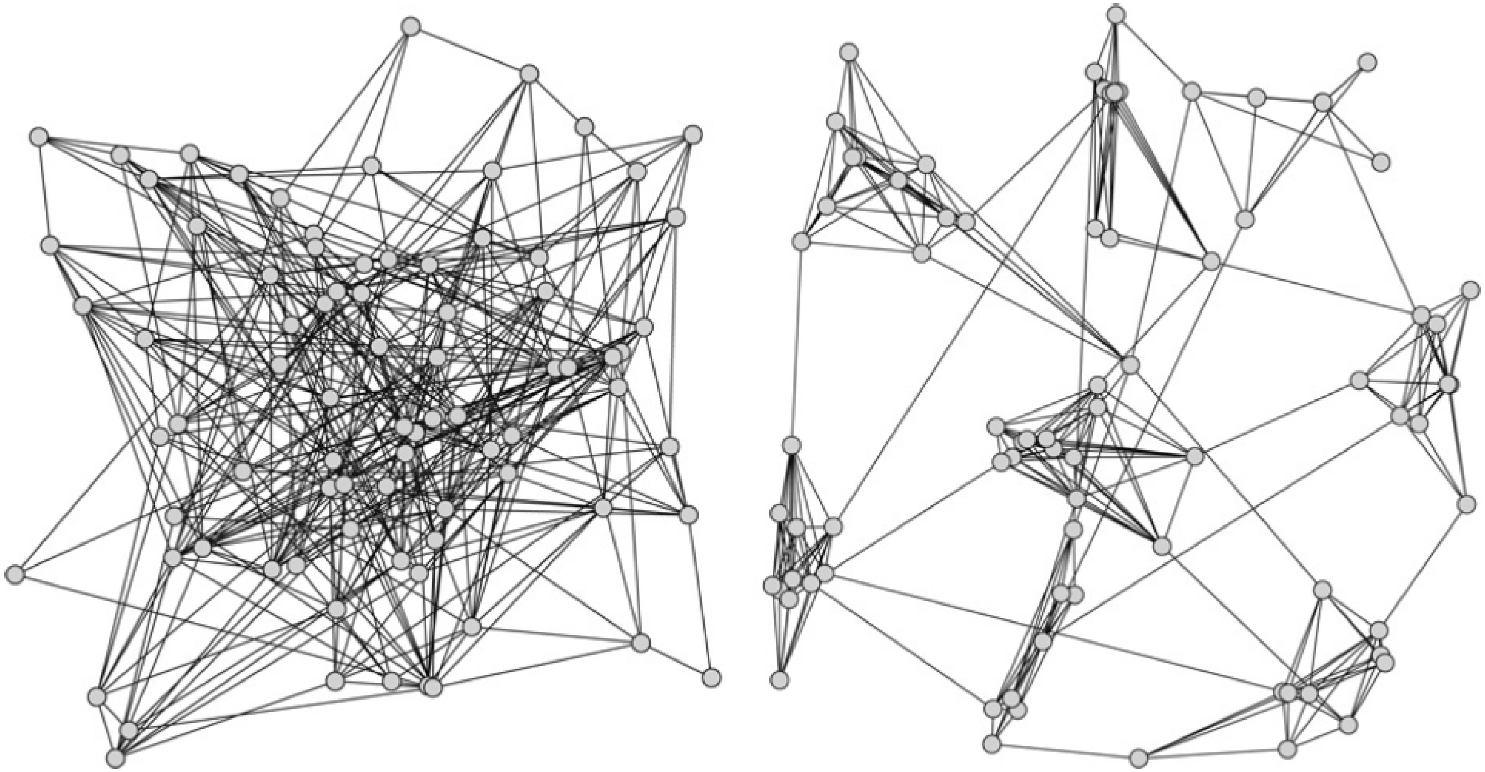


Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Schaeffer, Satu Elisa. "Graph Clustering." *Computer Science Review* 1, no. 1 (2007): 27-64.

Goal: divide the graph into subgraphs each of which has lots of internal connections and few connections to the rest of the graph

Schaeffer Computer Science Review (2007) http://dx.doi.org/10.1016/j.cosrev.2007.05.001

Clustering Graphs



Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Schaeffer, Satu Elisa. "Graph Clustering." *Computer Science Review* 1, no. 1 (2007): 27-64.

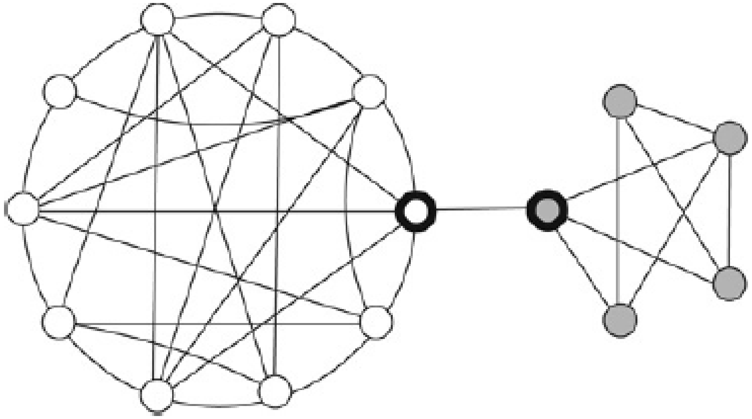
Two algorithms:

edge betweeness
markov clustering

Schaeffer Computer Science Review (2007) http://dx.doi.org/10.1016/j.cosrev.2007.05.001

Betweeness clustering

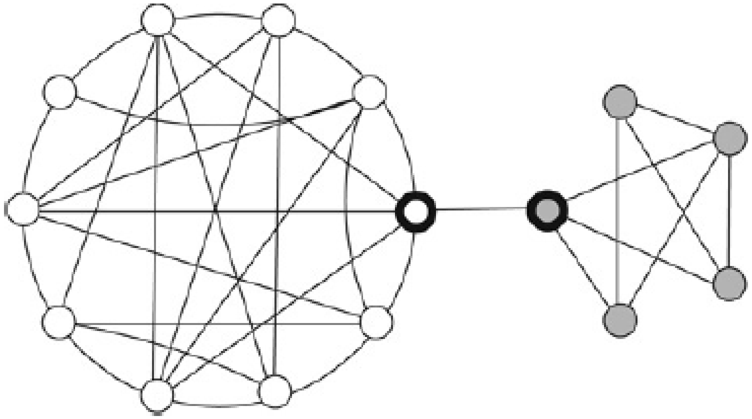
- Edge betweeness = number (or summed weight) of shortest paths between all pairs of vertices that pass through the edge.
  - Take a weighted average if there are >1 shortest paths for the same pair of nodes.



Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Schaeffer, Satu Elisa. "Graph Clustering." *Computer Science Review* 1, no. 1 (2007): 27-64.

Betweeness clustering

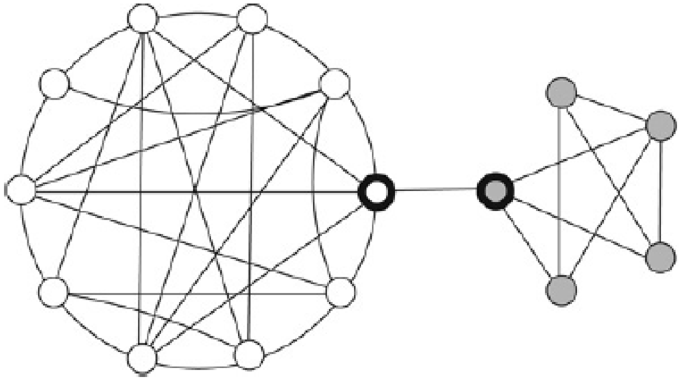
- Repeat until max(betweeness) < threshold:
  - – Compute betweeness
  - – Remove edge with highest betweeness



Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Schaeffer, Satu Elisa. "Graph Clustering." *Computer Science Review* 1, no. 1 (2007): 27-64.

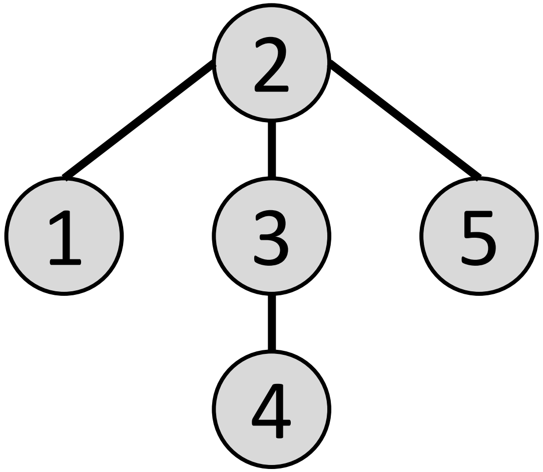
Markov clustering (MCL)

- Goal: produce sharp partitions
- Intuition: A random walk will spend more time within a cluster than passing between clusters.
- Concisely explained here: Enright *et al.* NAR (2002) http://www.ncbi.nlm.nih.gov/pmc/articles/PMC101833



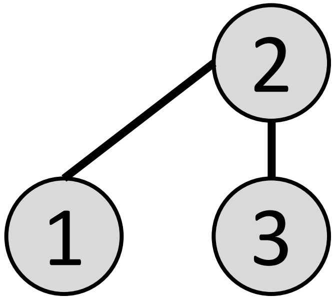
Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Schaeffer, Satu Elisa. "Graph Clustering." *Computer Science Review* 1, no. 1 (2007): 27-64.

Adjacency Matrix



	1	2	3	4	5
1	0	1	0	0	0
2	1	0	1	0	1
3	0	1	0	1	0
4	0	0	1	0	0
5	0	1	0	0	0

Adjacency Matrix

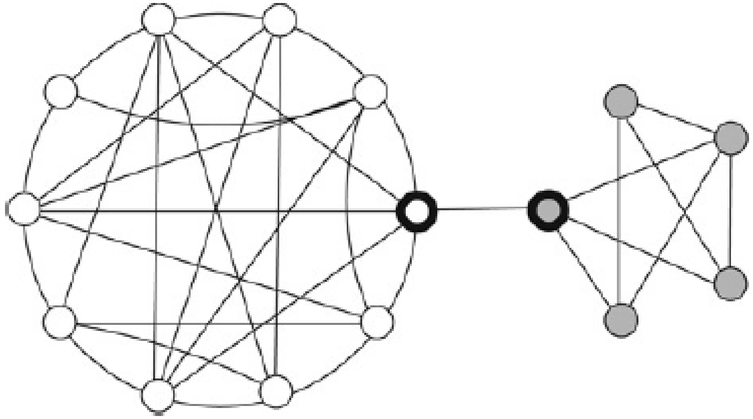


$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 1 \\ 3 & 0 & 1 & 0 \end{array} \times \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 1 \\ 3 & 0 & 1 & 0 \end{array} = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 1 & 0 & 1 \\ 2 & 0 & 2 & 0 \\ 3 & 1 & 0 & 1 \end{array}$$

A^N: a_{ij}= m iff there exist exactly m paths of length N between i and j.

MCL clustering

- Stochastic Matrix: each element Mij represents a probability of moving from i to j (this is a "Column Stochastic Matrix").

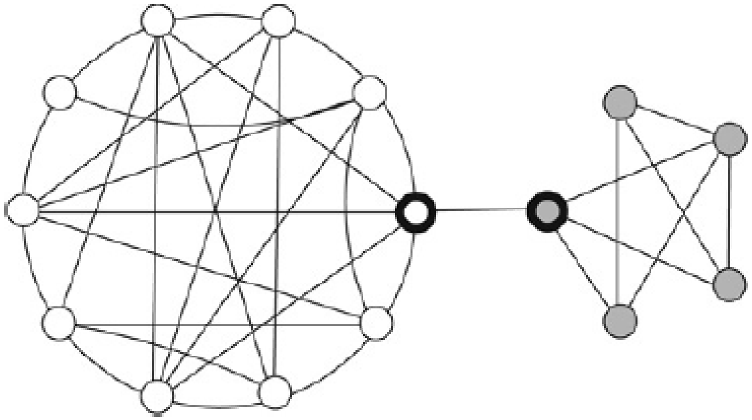


Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Schaeffer, Satu Elisa. "Graph Clustering." *Computer Science Review* 1, no. 1 (2007): 27-64.

MCL clustering

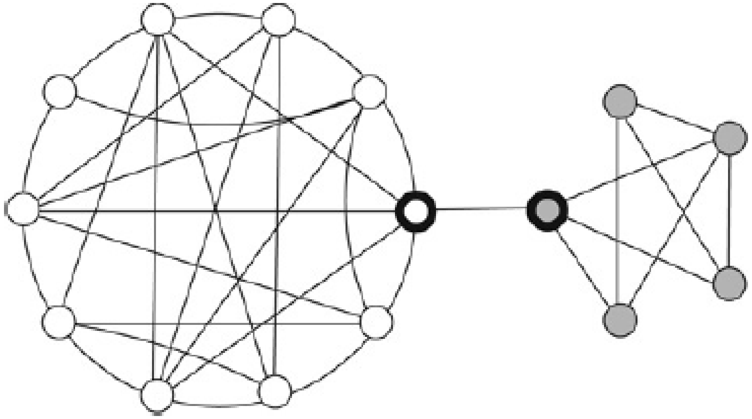
- Stochastic Matrix: each element Mij represents a probability of moving from i to j (this is a "Column Stochastic Matrix").
- Therefore, $\sum_j p_{ij} = 1$
- The probability of moving from i to j in two steps is given by

$$(M^2)_{ij} = \sum_k p_{ik} p_{kj}$$



Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Schaeffer, Satu Elisa. "Graph Clustering." *Computer Science Review* 1, no. 1 (2007): 27-64.

- If we keep multiplying the stochastic matrix by itself, we compute the probabilities of longer and longer walks – we expect that the transitions will occur more frequently within a natural cluster than between them.



Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Schaeffer, Satu Elisa. "Graph Clustering." *Computer Science Review* 1, no. 1 (2007): 27-64.

- This procedure won’t produce discrete clusters, so the algorithm includes an “inflation” step that exaggerates these effects: raise each element of the matrix to the power r and renormalize.

$$p_A = 0.9$$

$$p_B = 0.1$$

$$p_A \rightarrow \frac{.81}{.81 + .01} = .99$$

$$p_B \rightarrow \frac{.01}{.81 + .01} = .01$$

$$(\Gamma_r M)_{pq} = (M_{pq})^r \bigg/ \sum_{i=1}^{k} (M_{iq})^r .$$

Protein Interaction Networks: Computational Analysis

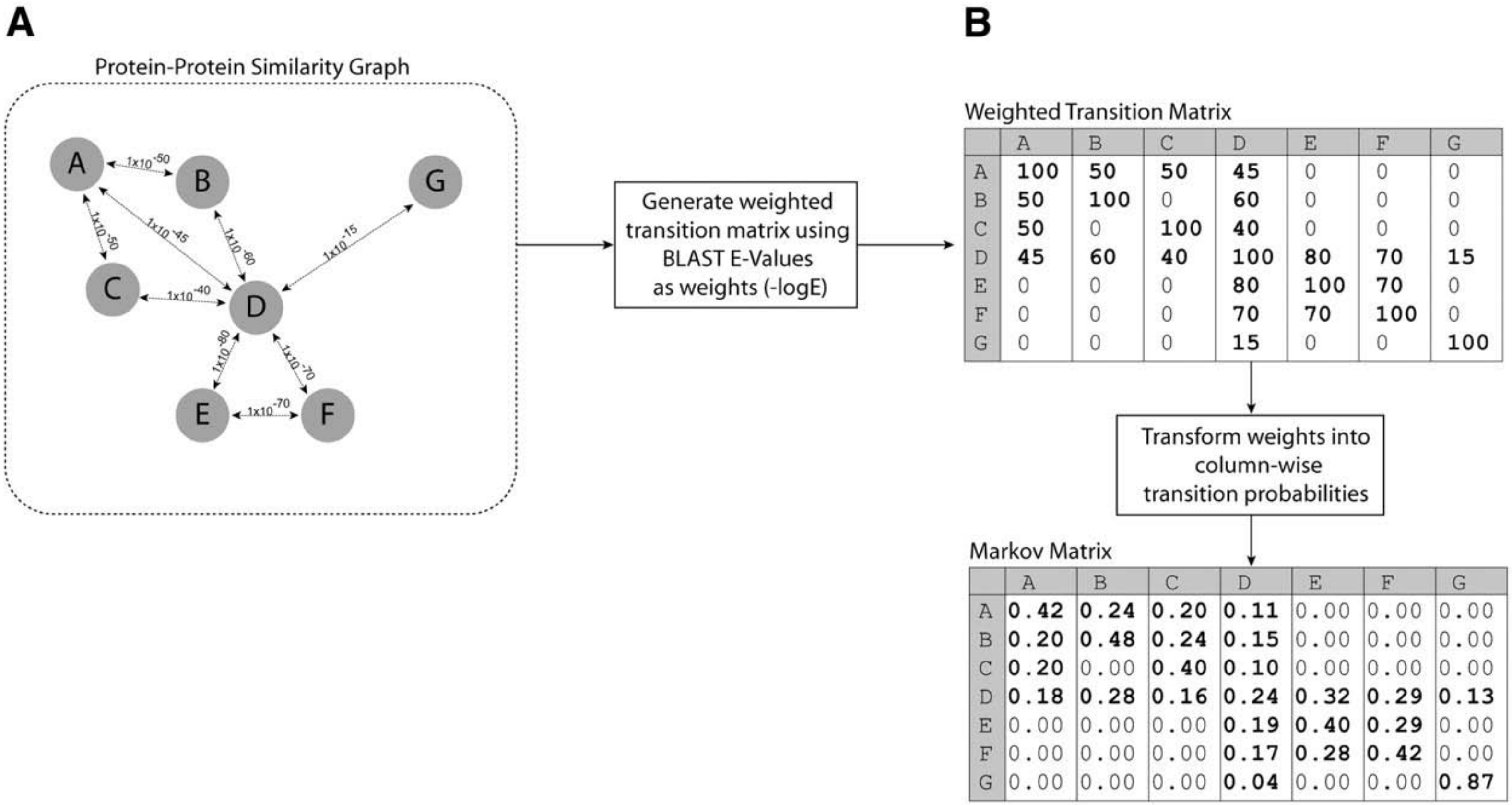
By Aidong Zhang

http://books.google.com/books?id=hOzAUrwW-ZoC&lpg=PA141&ots=Vd0TK0fCAR&dq=mcl%20inflation%20operator&pg=PA142#v=onepage&q&f=true

```
G is a graph
add loops to G          # needed for a prob. of no transition
set Γ to some value          # affects granularity
set M_1 to be the matrix of random walks on G
while (change) {
    M_2 = M_1 * M_1       # expansion
    M_1 = Γ(M_2)     # inflation
    change = difference(M_1, M_2)
    }
set CLUSTERING as the components of M_1
```

Example

- Identifying protein families
- BLAST will identify proteins with shared domains, but these might not be very similar otherwise (eg: SH2, SH3 domains)



Generate weighted transition matrix using BLAST E-Values as weights (-logE)

Weighted Transition Matrix

	A	B	C	D	E	F	G
A	100	50	50	45	0	0	0
B	50	100	0	60	0	0	0
C	50	0	100	40	0	0	0
D	45	60	40	100	80	70	15
E	0	0	0	80	100	70	0
F	0	0	0	70	70	100	0
G	0	0	0	15	0	0	100

Transform weights into column-wise transition probabilities

Markov Matrix

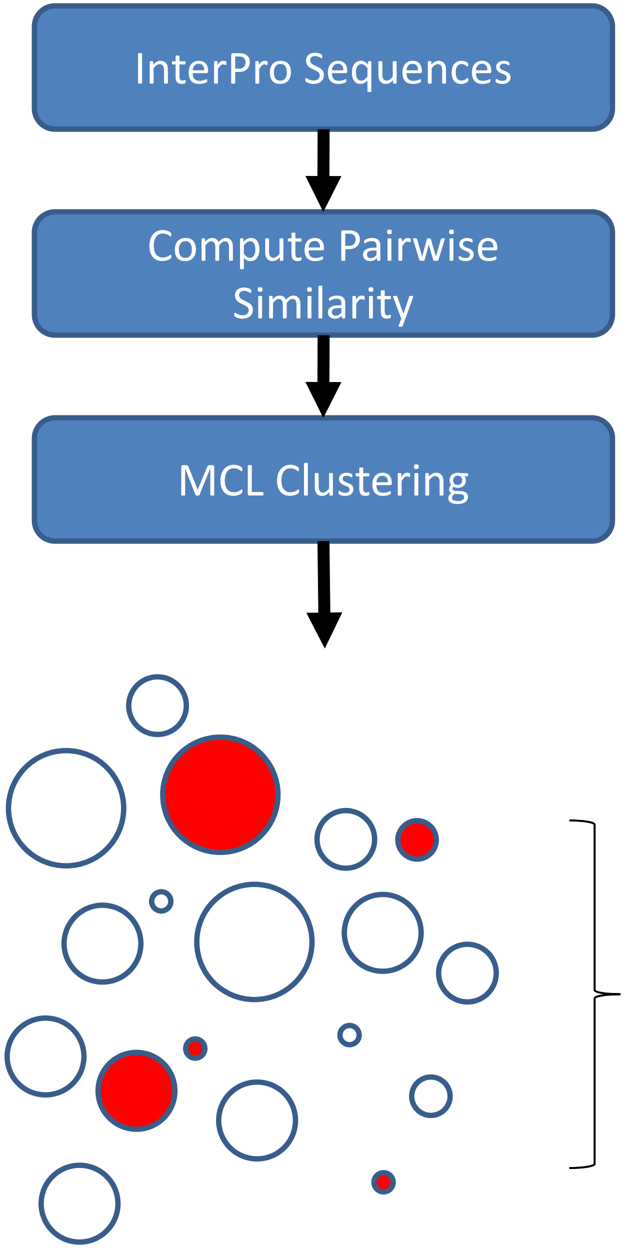
	A	B	C	D	E	F	G
A	0.42	0.24	0.20	0.11	0.00	0.00	0.00
B	0.20	0.48	0.24	0.15	0.00	0.00	0.00
C	0.20	0.00	0.40	0.10	0.00	0.00	0.00
D	0.18	0.28	0.16	0.24	0.32	0.29	0.13
E	0.00	0.00	0.00	0.19	0.40	0.29	0.00
F	0.00	0.00	0.00	0.17	0.28	0.42	0.00
G	0.00	0.00	0.00	0.04	0.00	0.00	0.87

Source: Enright, Anton J., Stijn Van Dongen, et al. "An Efficient Algorithm for Large-scale Detection of Protein Families." *Nucleic Acids Research* 30, no. 7 (2002): 1575-84.

Extremely fast, since it only requires matrix operations

Enright, A. J. et al. Nucl. Acids Res. 2002 30:1575-1584; doi:10.1093/nar/30.7.1575

Nucleic Acids Research



InterPro Sequences

Compute Pairwise Similarity

MCL Clustering

InterPro ID	No. of families	Domain description
IPR001064	141	Crystallin
IPR000504	110	RNA-binding region RNP-1 (RNA recognition motif)
IPR003006	107	Immunoglobulin and major histocompatibility complex domain
IPR000531	97	TonB-dependent receptor protein
IPR003015	96	Myc-type, helix–loop–helix dimerisation domain
IPR001680	76	G-protein β WD-40 repeats
IPR000561	73	EGF-like domain
IPR000169	72	Eukaryotic thiol (cysteine) proteases active sites
IPR001777	42	Fibronectin type III domain

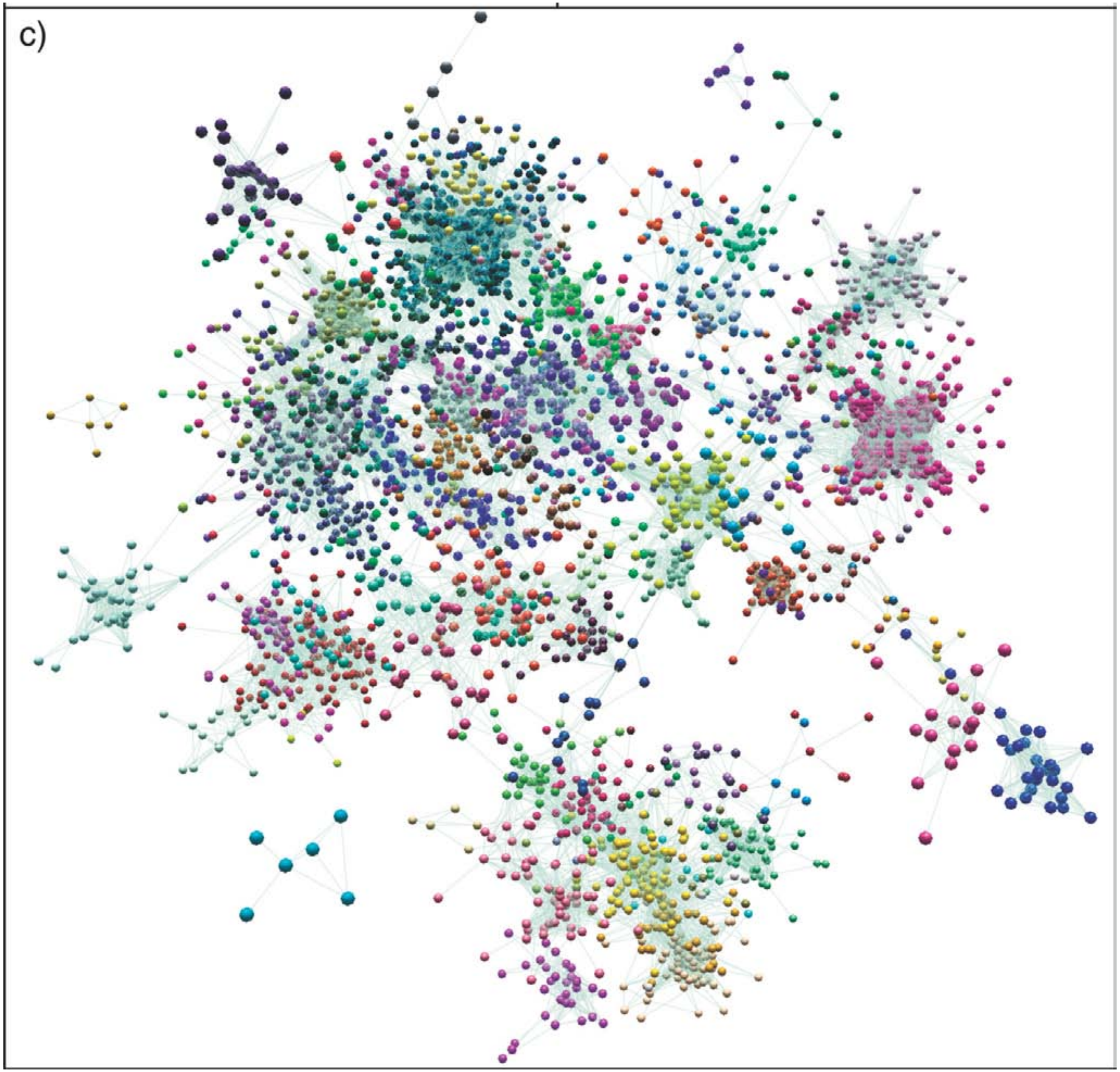
Distinct clusters identified by MCL can still share a common domain

Example

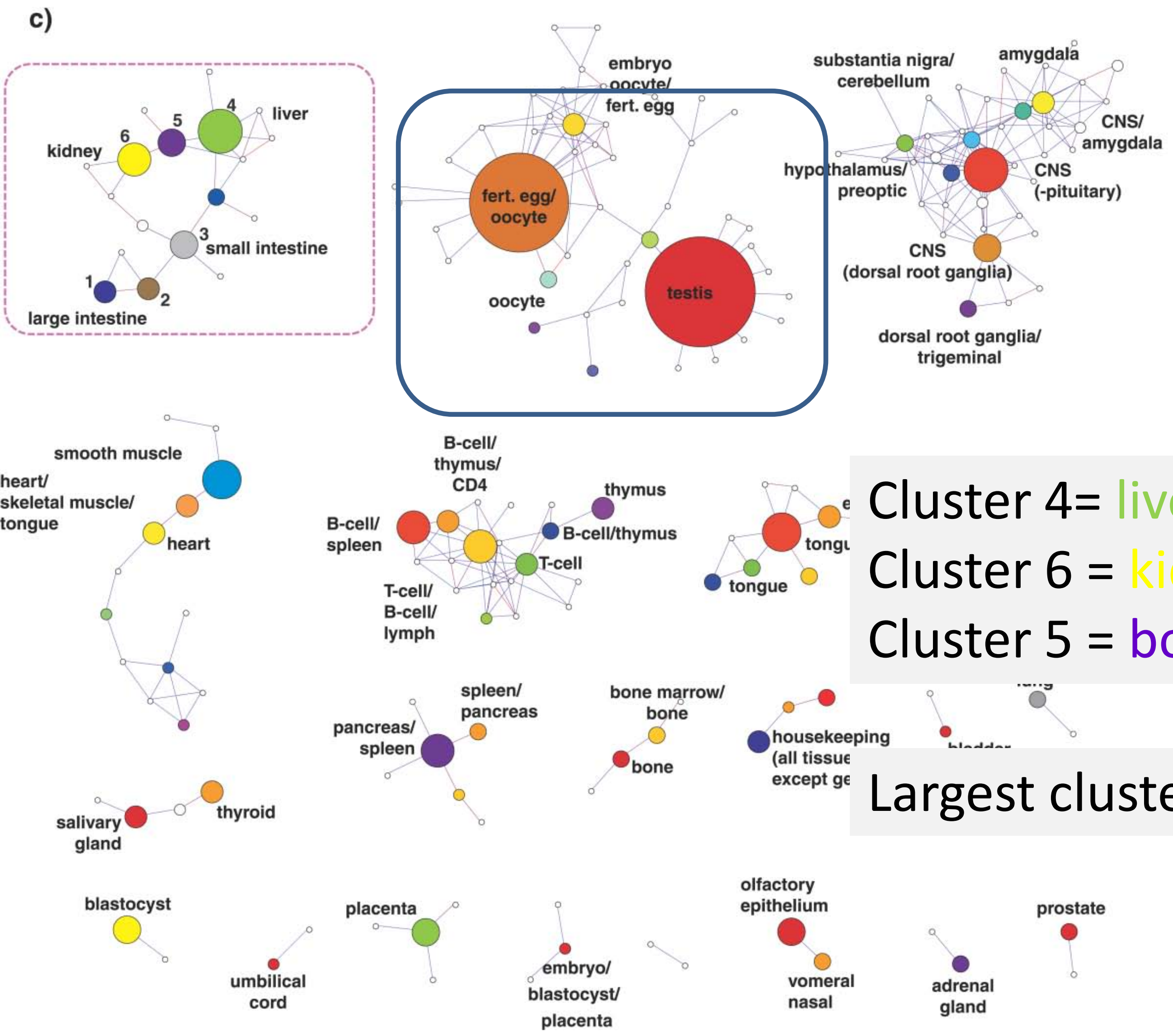
- Clustering expression data for 61 mouse tissues
- Nodes = genes
- Edges = Pearson correlation coefficient > threshold
- Network gives an overview of connections not obvious from hierarchical clustering

Nodes=genes
Edges=pearson correlation of expression in mouse tissues
Clustered by MCL

Freeman, *et al.*(2007) PLoS Comput Biol 3(10): e206. doi:10.1371/journal.pcbi.0030206



Source: Freeman, Tom C., Leon Goldovsky, et al. "Construction, Visualisation, and Clustering of Transcription Networks from Microarray Expression Data." *PLoS Computational Biology* 3, no. 10 (2007): e206.



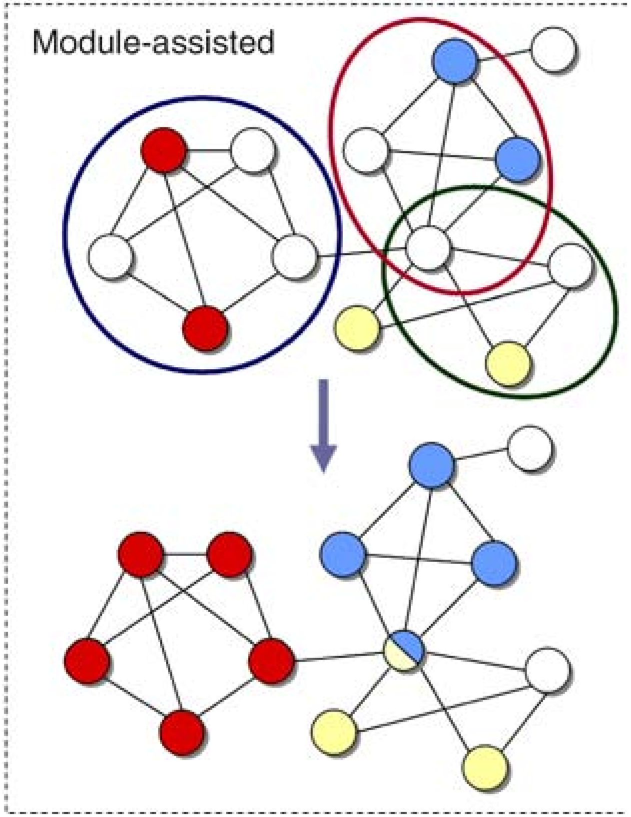
Cluster 4= liver specific

Cluster 6 = kidney specific

Cluster 5 = both liver and kidney

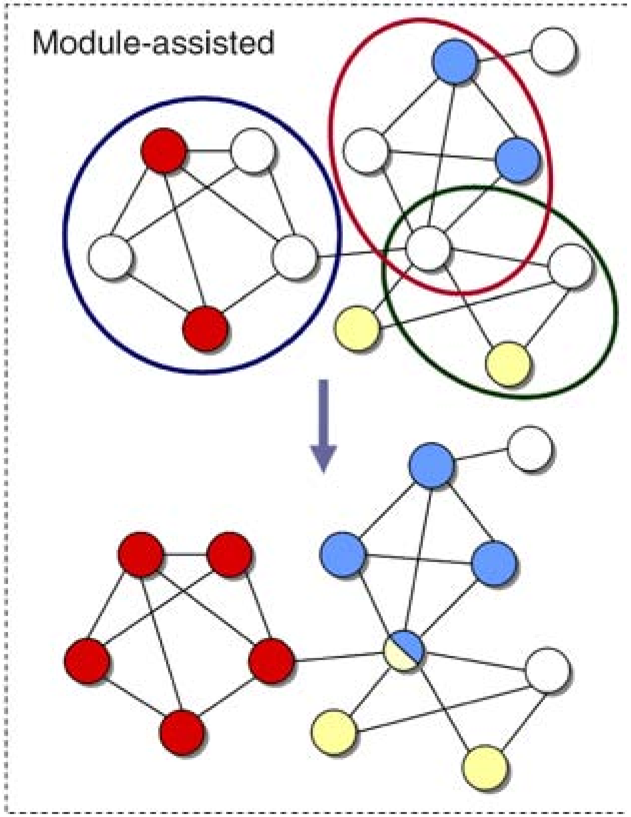
Largest clusters are gamete-specific

Courtesy of Freeman et al. License: CC-BY.
Source: Freeman, Tom C., Leon Goldovsky, et al. "Construction, Visualisation, and Clustering of Transcription Networks from Microarray Expression Data." *PLoS Computational Biology* 3, no. 10 (2007): e206.



Courtesy of EMBO. Used with permission.
Source: Sharan, Roded, Igor Ulitsky, et al. "Network-based Prediction of Protein Function." *Molecular Systems Biology* 3, no. 1 (2007).

How do we decide which function to assign to members of a cluster?



Courtesy of EMBO. Used with permission.

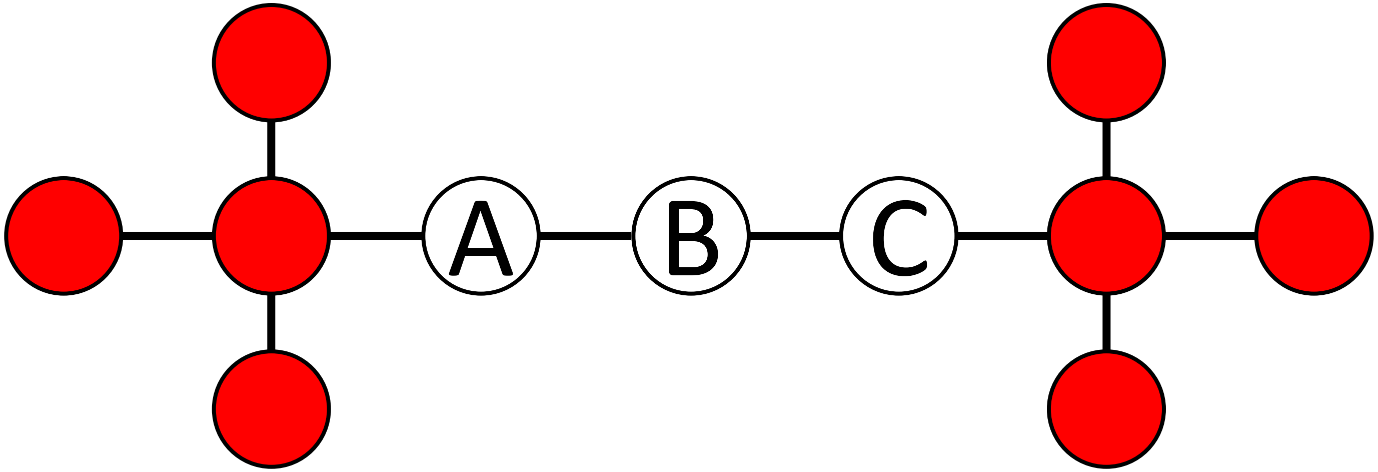
Source: Sharan, Roded, Igor Ulitsky, et al. "Network-based Prediction of Protein Function." *Molecular Systems Biology* 3, no. 1 (2007).

How do we decide which function to assign to members of a cluster?

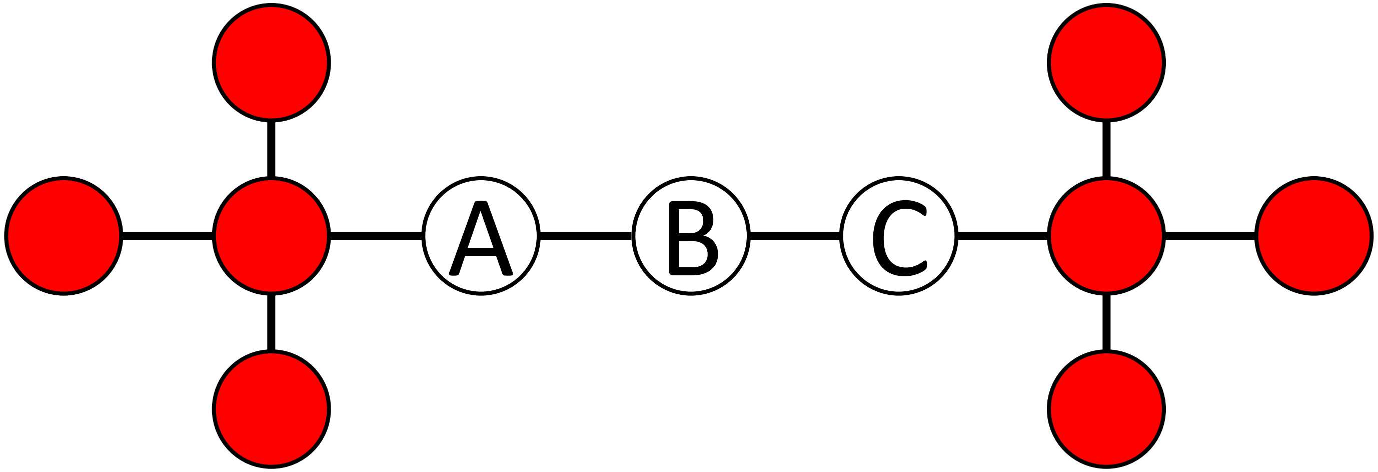
- Consensus
- Significant by hypergeometric

Network Models

- Structure of network
  - Coexpression
  - Mutual information
  - Physical/genetic interactions
- Analysis of network
  - Ad hoc
  - Shortest path
  - Clustering
  - Optimization

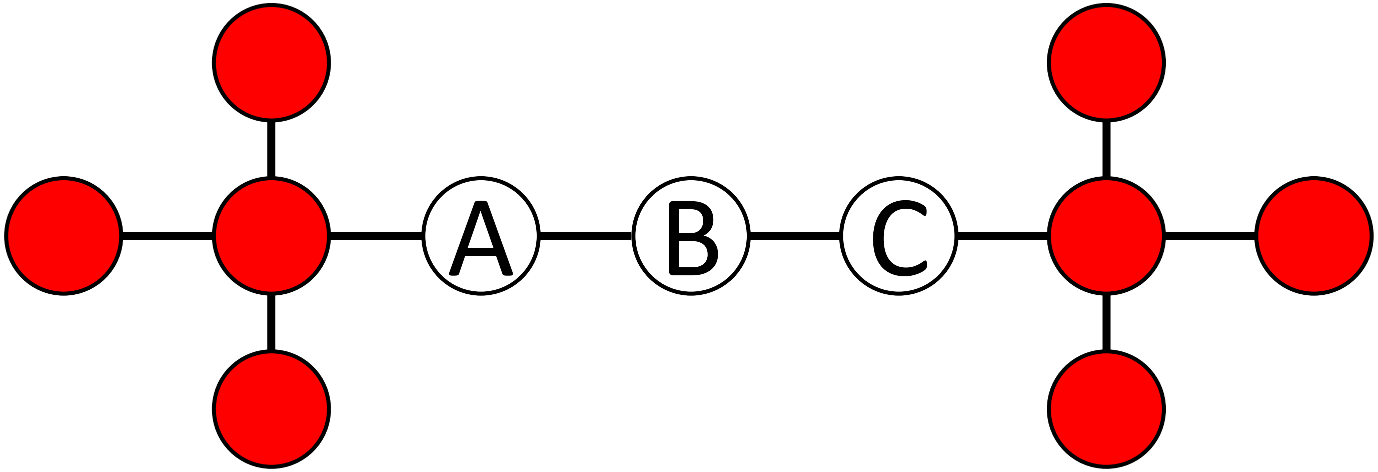


How do we find modules associated with specific data?
Example: paint a PPI network with expression data. Try to find connected components that have overall high expression. (Example: Ideker et al. (2002) Bioinformatics).

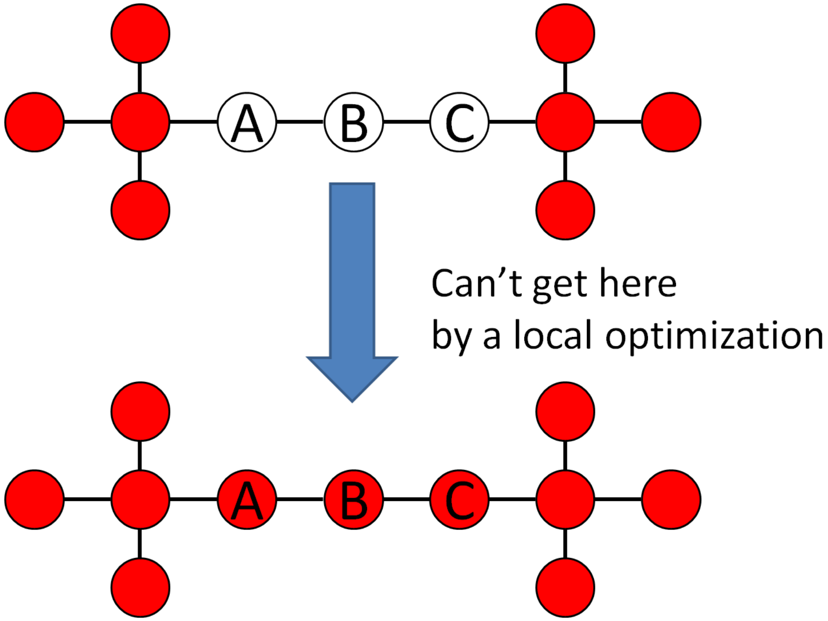


Active subgraph problem:

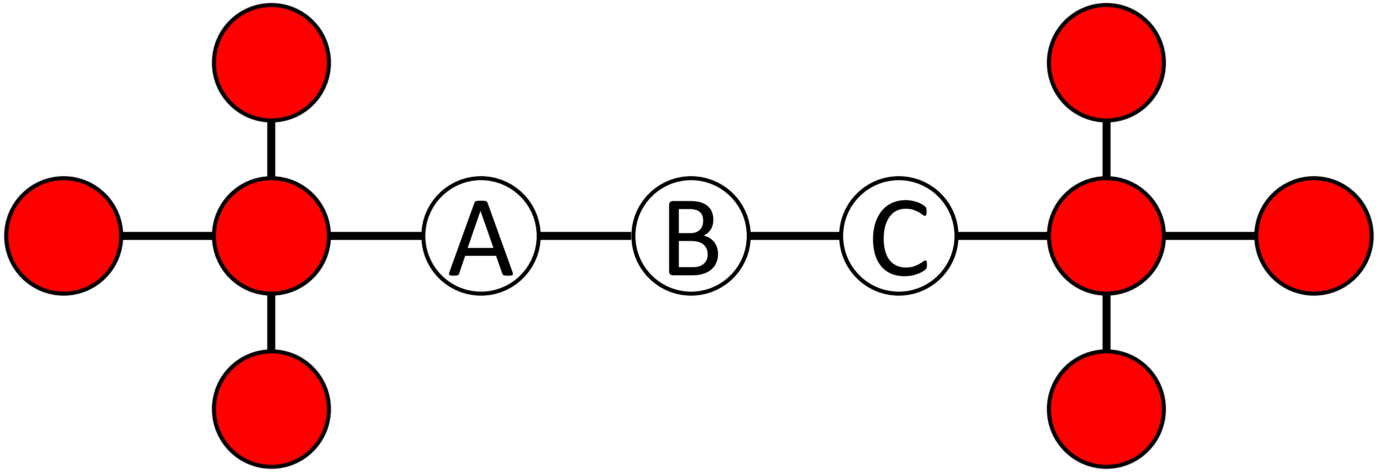
Can reveal hidden components of a biological response.



Where did we see something similar?



- The annotation problem attempts to label the entire graph.
- The active subnet problem searches for a part of the graph that is enriched in a label.

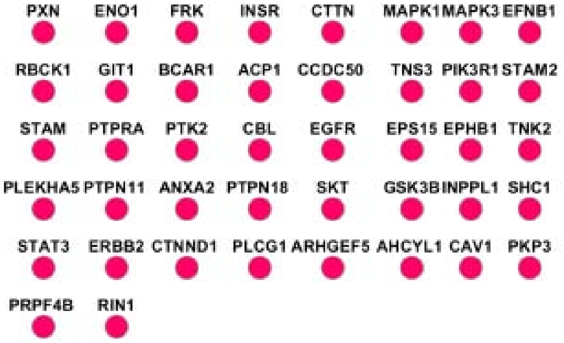


- **Steiner Tree Problem:** Find the smallest tree connecting all the vertices of in a set of interest (terminals).

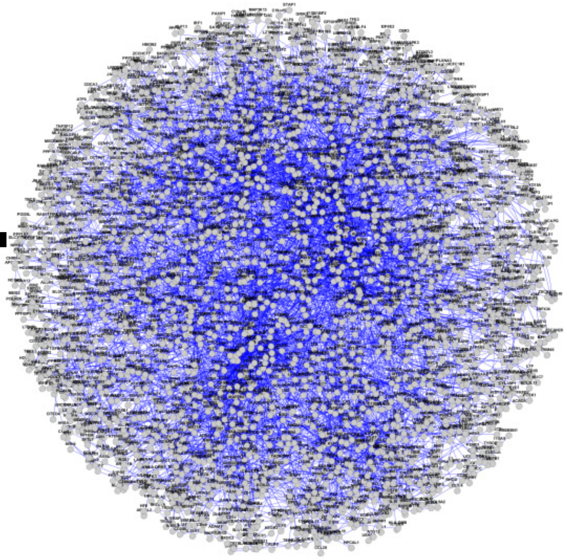
- Downside: will include all terminals, including false positives.

Interactome

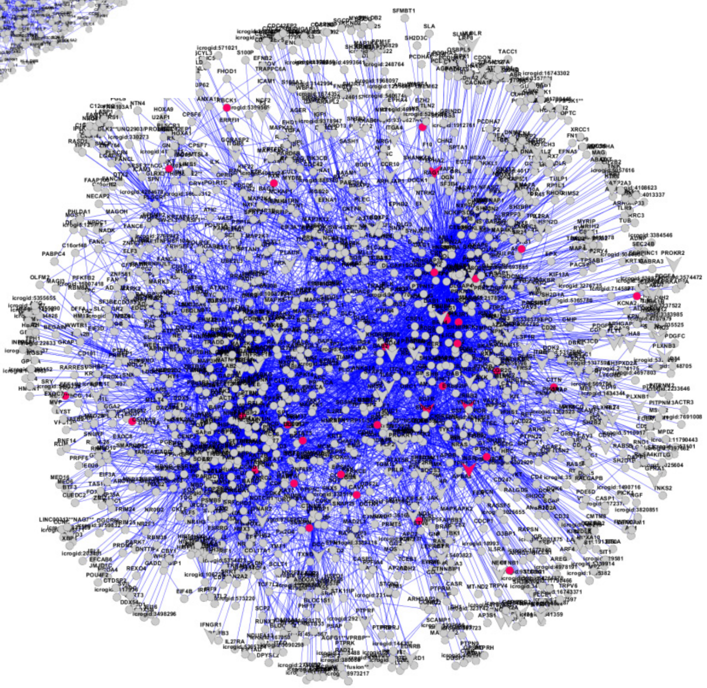
Experimental hits



+

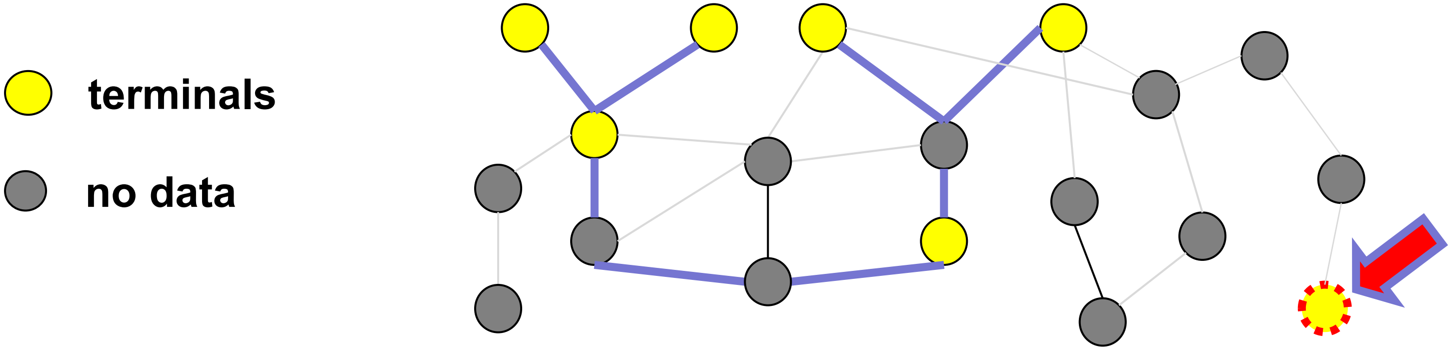


Naïve methods



- Not all hits are real
- Not all edges are real
- Not all edges are known

Avoiding False Positives



● terminals

● no data

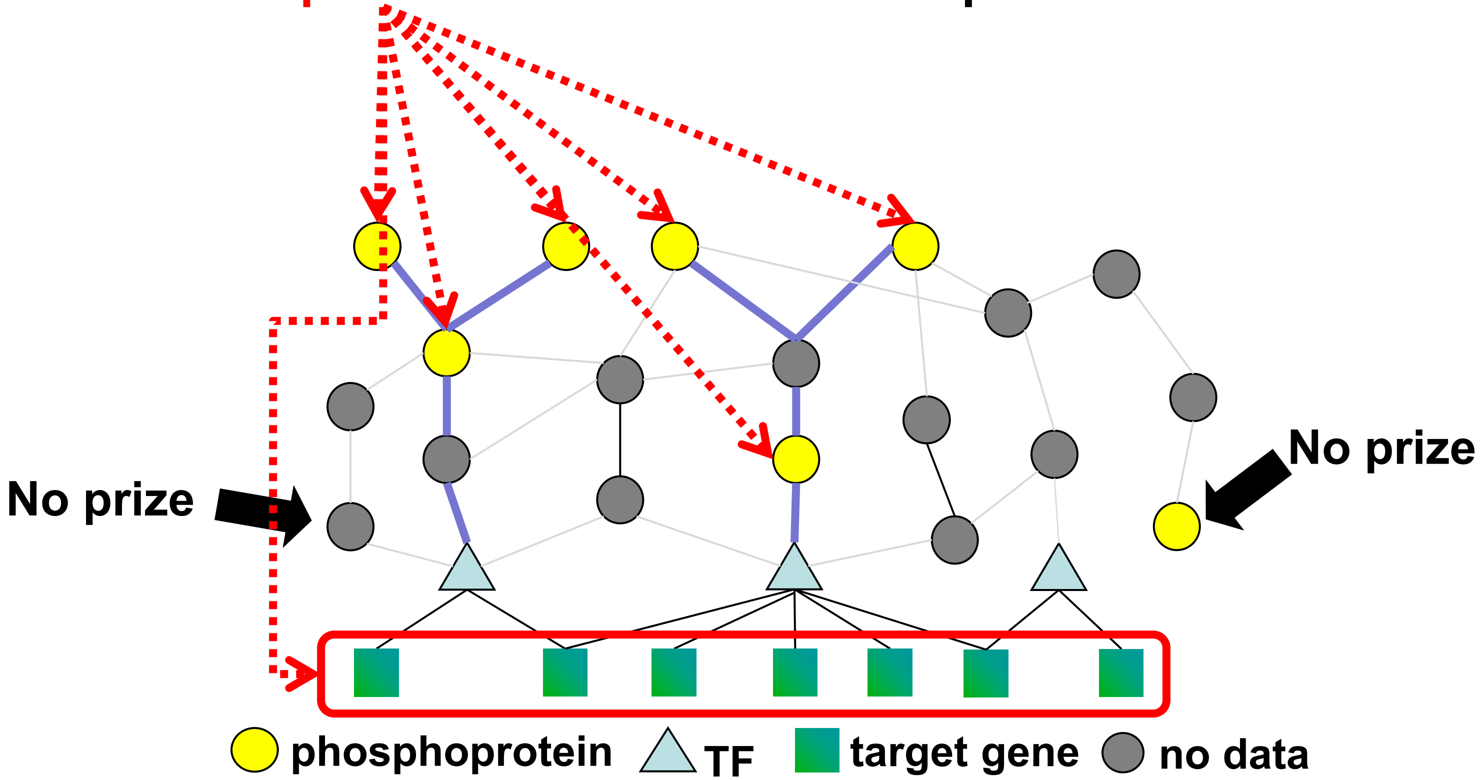
Steiner tree is forced to include this node

Network Models

- Structure of network
  - Coexpression
  - Mutual information
  - Physical/genetic interactions
- Analysis of network
  - Ad hoc
  - Shortest path
  - Clustering
  - Optimization

Prize Collecting Steiner Tree

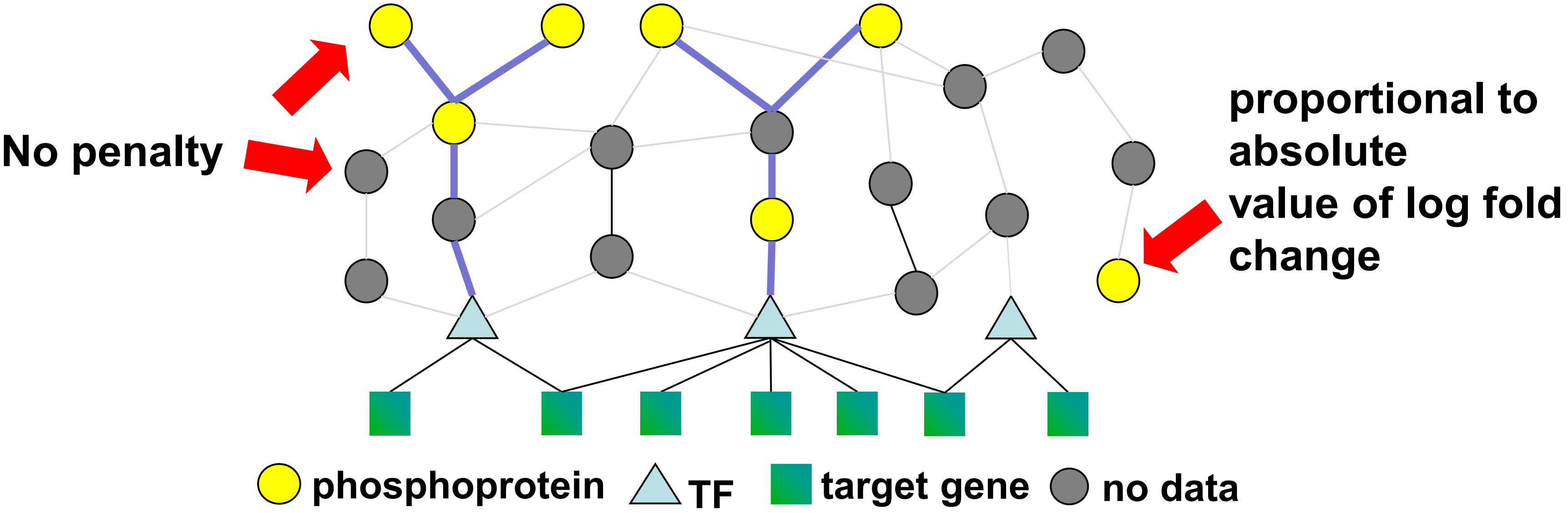
- Collect a prize for each data point included



Courtesy of Huang et al. Used with permission.
Source: Huang, Shao-shan Carol, David C. Clarke, et al. "Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling." *PLoS Computational Biology* 9, no. 2 (2013): e1002887.

Don’t Include All Data

- Pay a penalty for excluding nodes



No penalty

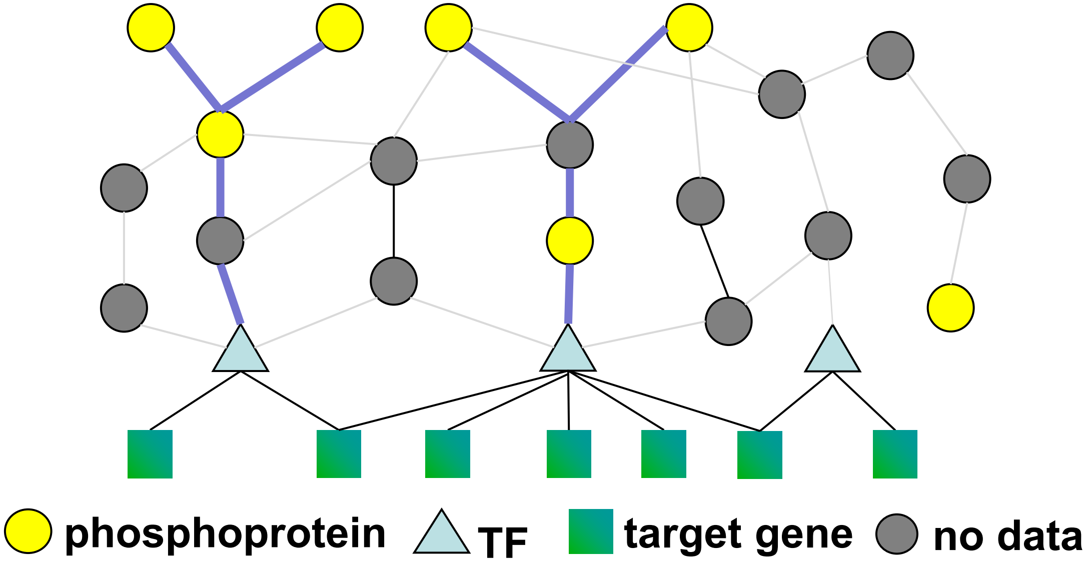
proportional to absolute value of log fold change

Courtesy of Huang et al. Used with permission.
Source: Huang, Shao-shan Carol, David C. Clarke, et al. "Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling." *PLoS Computational Biology* 9, no. 2 (2013): e1002887.

$$\boxed{\sum_{v \text{ not in } T} \beta\,\text{penalty}(v)} + \sum_{e \text{ in } T} \text{cost}(e)$$

Avoid Unlikely Interactions

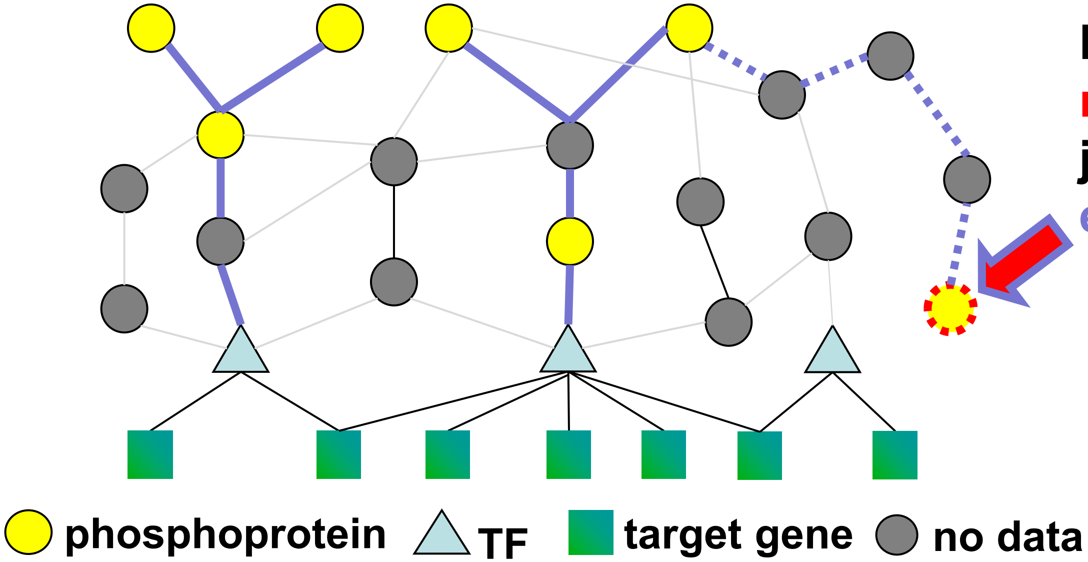
- Pay a cost for including edges based on probability



Courtesy of Huang et al. Used with permission.
Source: Huang, Shao-shan Carol, David C. Clarke, et al. "Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling." *PLoS Computational Biology* 9, no. 2 (2013): e1002887.

$$\sum_{v \text{ not in } T} \beta\, \text{penalty}(v) + \boxed{\sum_{e \text{ in } T} \text{cost}(e)}$$

Balanced Objective Function



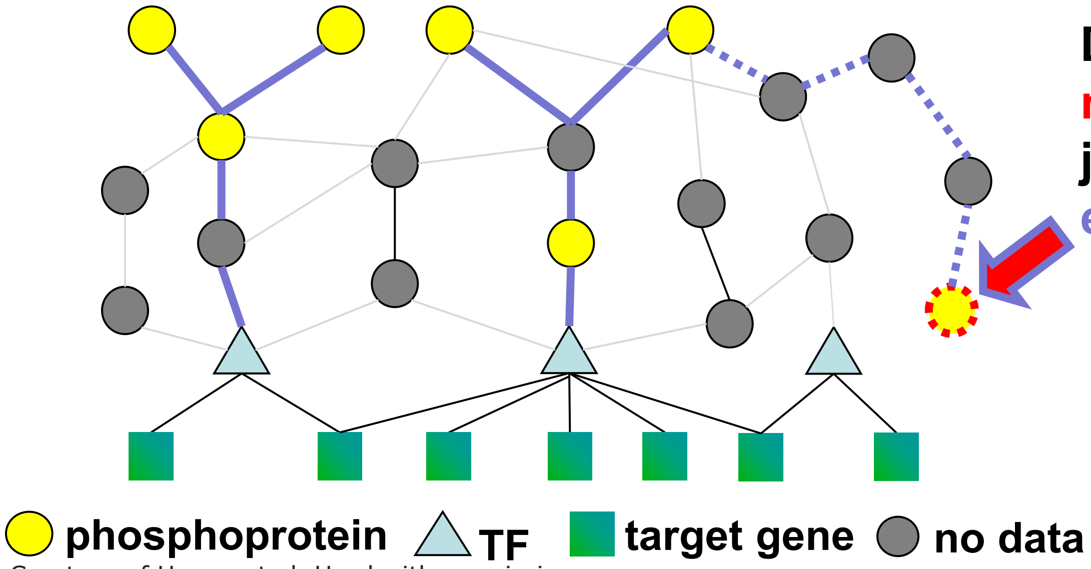
Does the **node penalty** justify the **edge costs**?

Courtesy of Huang et al. Used with permission.
Source: Huang, Shao-shan Carol, David C. Clarke, et al. "Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling." *PLoS Computational Biology* 9, no. 2 (2013): e1002887.

$$\boxed{\sum_{v \text{ not in } T} \beta\,\text{penalty}(v)} + \boxed{\sum_{e \text{ in } T} \text{cost}(e)}$$

Optimization methods:

- Biazzo I, Braunstein A, Zecchina R.
  Phys Rev E Stat Nonlin Soft Matter Phys. 2012 Aug;86(2 Pt 2):026706.
- I. Ljubic, R. Weiskircher, U. Pferschy, G. Klau, P. Mutzel, and M. Fischetti:
  Mathematical Programming, Series B, 105(2-3):427-449, 2006.



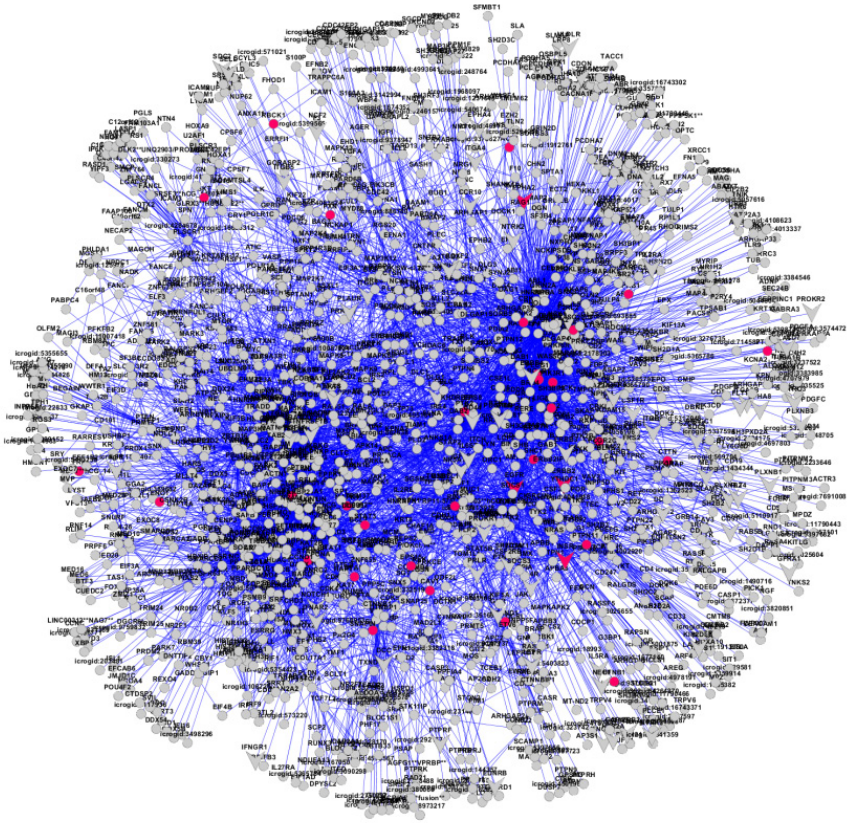
● phosphoprotein △ TF 🟩 target gene ● no data

Does the node penalty justify the edge costs?

Courtesy of Huang et al. Used with permission.
Source: Huang, Shao-shan Carol, David C. Clarke, et al. "Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling." *PLoS Computational Biology* 9, no. 2 (2013): e1002887.

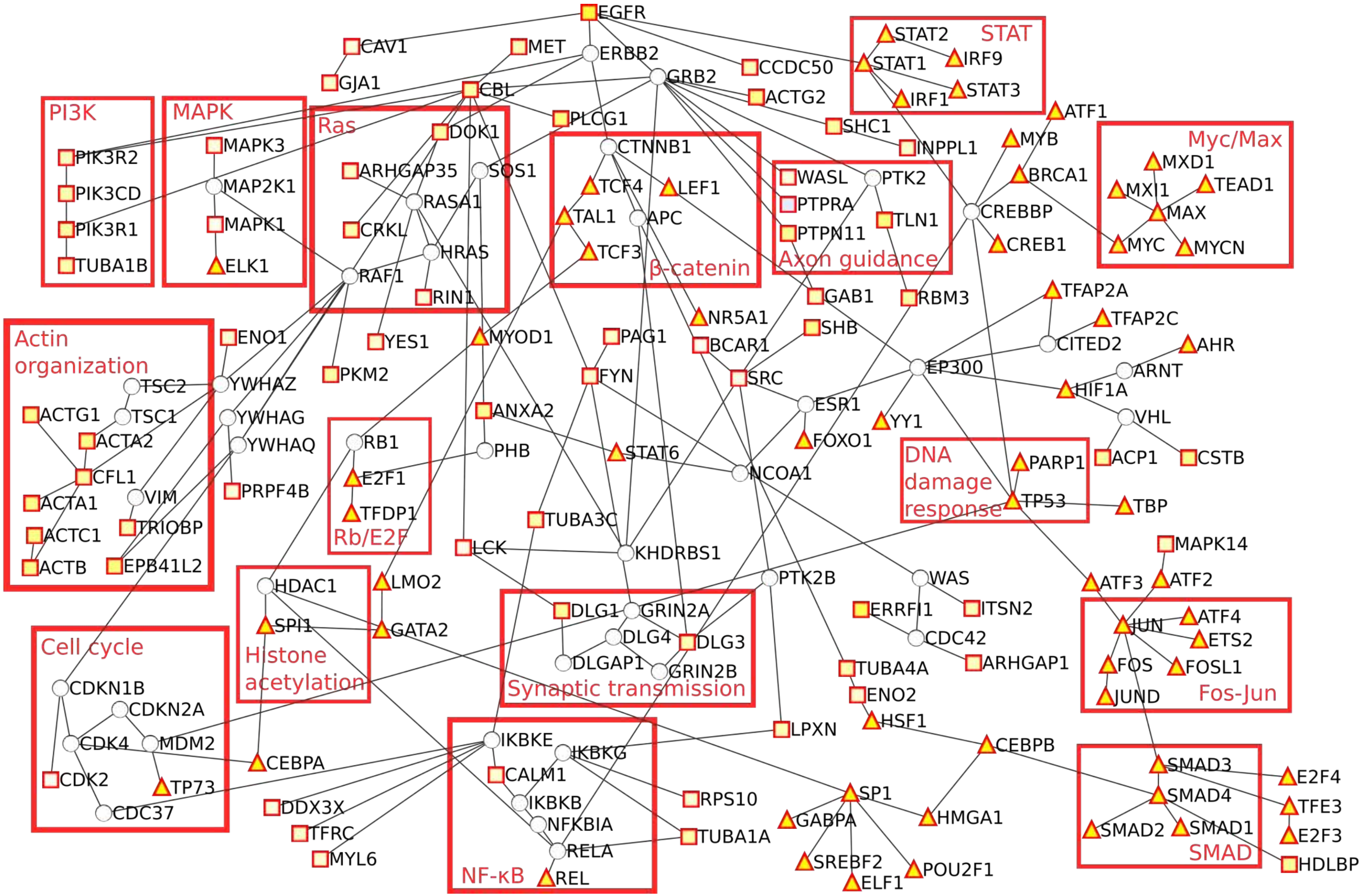
$$\boxed{\sum_{v \text{ not in } T} \beta\, \text{penalty}(v)} + \boxed{\sum_{e \text{ in } T} \text{cost}(e)}$$

Naïve Methods



- >2,500 nearest neighbors of phosphoproteins
- >4,500 nearest neighbors of phosphoproteins +transcription factors

Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling
PLoS Comput Biol 9(2): e1002887. doi:10.1371/journal.pcbi.1002887



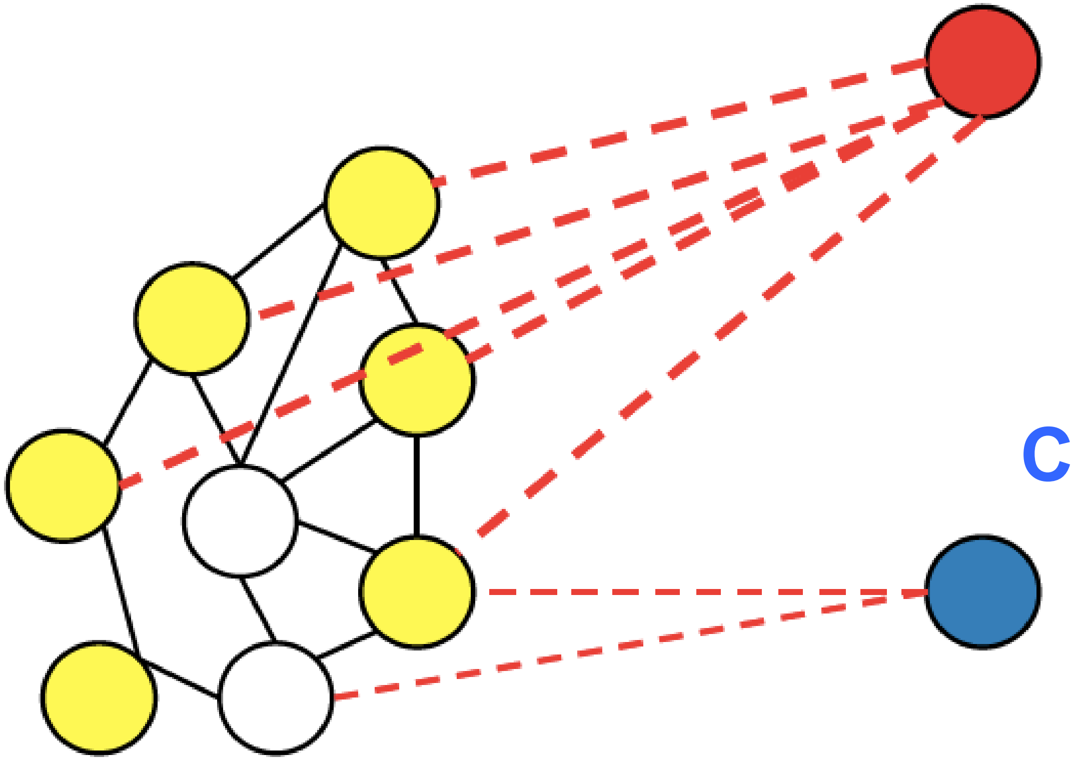
Courtesy of Huang et al. Used with permission.
Source: Huang, Shao-shan Carol, David C. Clarke, et al. "Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling." *PLoS Computational Biology* 9, no. 2 (2013): e1002887.

Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling

PLoS Comput Biol 9(2): e1002887. doi:10.1371/journal.pcbi.1002887

Can we find drug targets?

Rank every node by weighted distance to all prize-collecting Steiner tree nodes



High rank targets

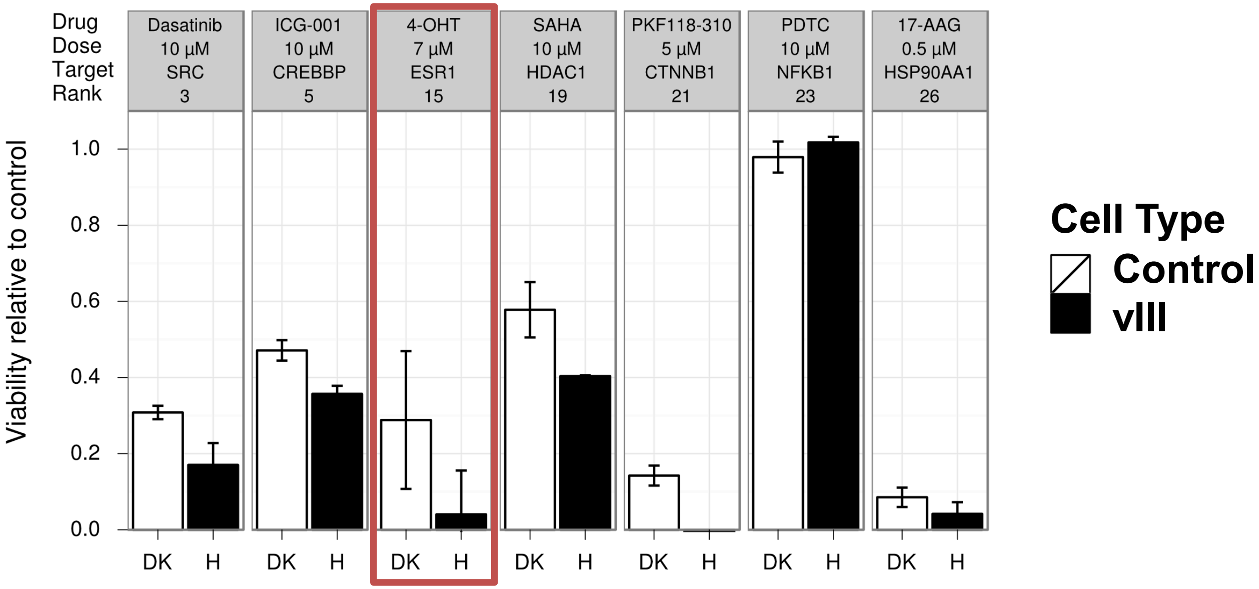
Control targets

Steiner Tree

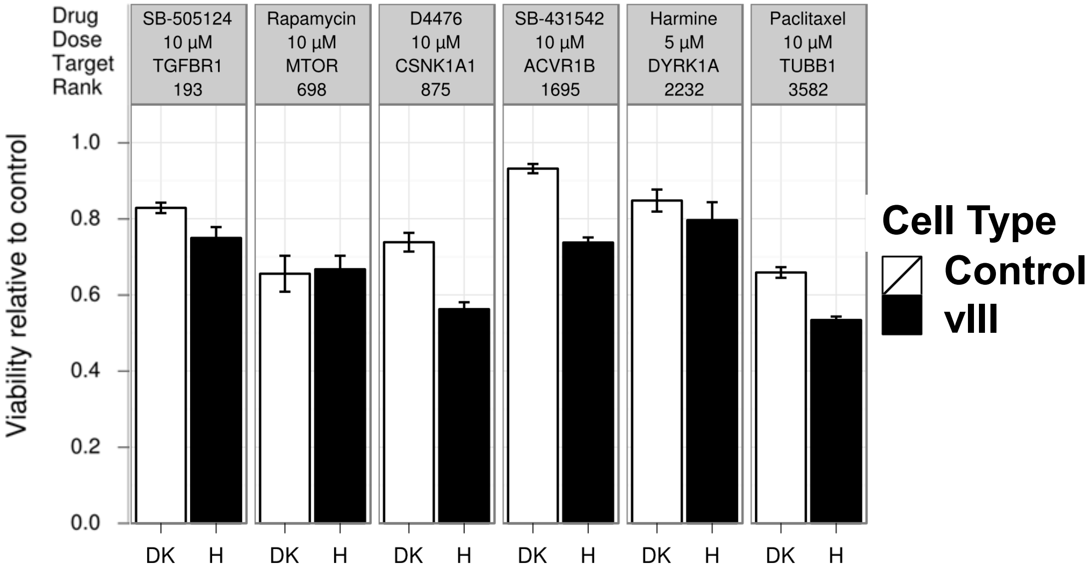
Courtesy of Huang et al. Used with permission.
Source: Huang, Shao-shan Carol, David C. Clarke, et al. "Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling." *PLoS Computational Biology* 9, no. 2 (2013): e1002887.

Rank

<27 out of 11,637



Lower Rank Targets 193 to 3,582 out of 11,637

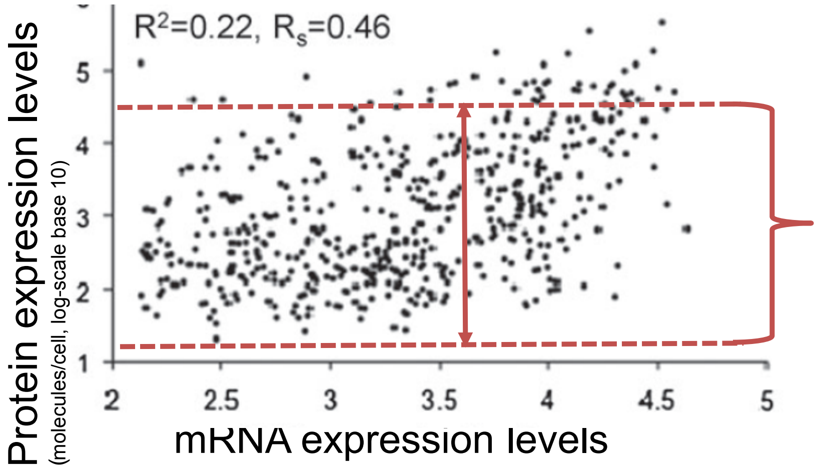


Courtesy of Huang et al. Used with permission.
Source: Huang, Shao-shan Carol, David C. Clarke, et al. "Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling." *PLoS Computational Biology* 9, no. 2 (2013): e1002887.

Data Integration

Approach

mRNA levels do not predict protein levels

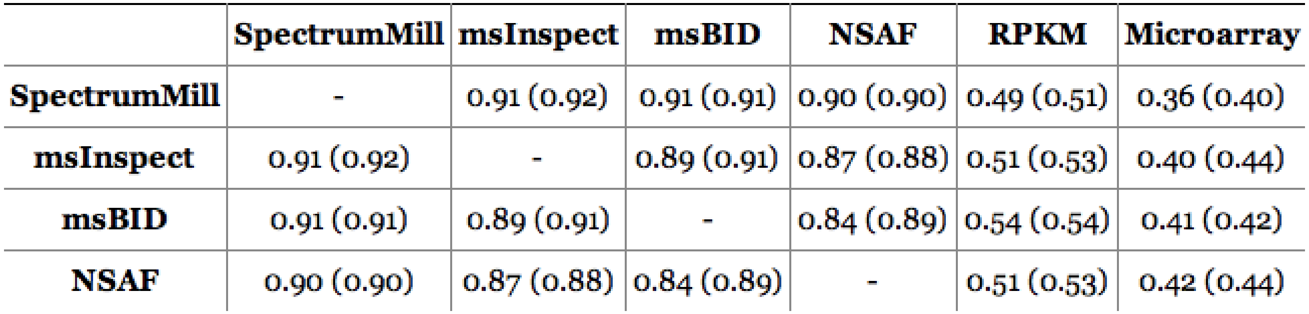


1,000 fold range of protein concentrations

Source: de Sousa Abreu, Raquel, Luiz O. Penalva, et al. "Global Signatures of Protein and mRNA Expression Levels." *Molecular Biosystems* 5, no. 12 (2009): 1512-26.

(arbitrary units, log-scale base 10)

Raquel de Sousa Abreu, Luiz Penalva, Edward Marcotte and Christine Vogel, *Mol. BioSyst.*, 2009 DOI: 10.1039/b908315d

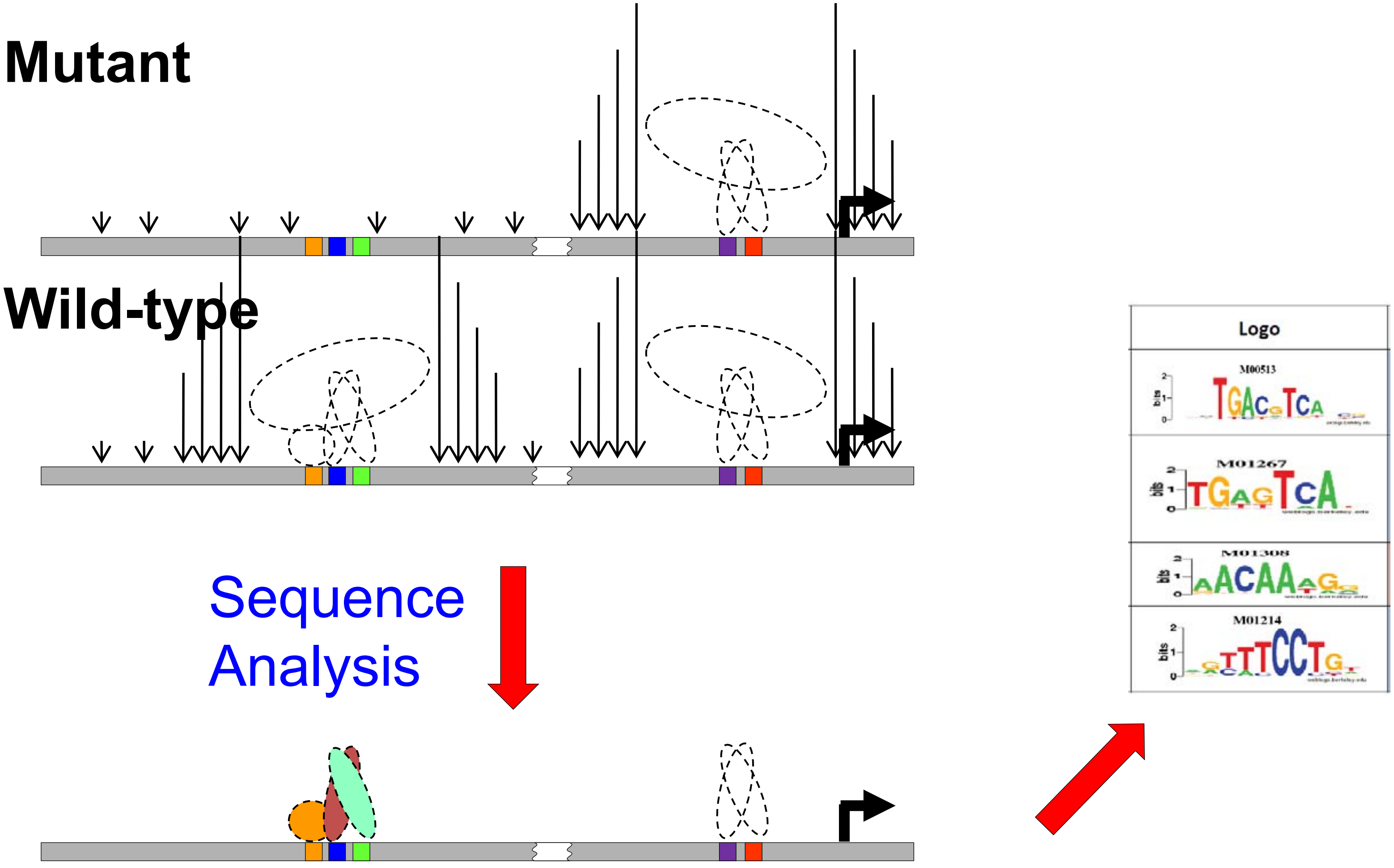


	SpectrumMill	msInspect	msBID	NSAF	RPKM	Microarray
SpectrumMill	-	0.91 (0.92)	0.91 (0.91)	0.90 (0.90)	0.49 (0.51)	0.36 (0.40)
msInspect	0.91 (0.92)	-	0.89 (0.91)	0.87 (0.88)	0.51 (0.53)	0.40 (0.44)
msBID	0.91 (0.91)	0.89 (0.91)	-	0.84 (0.89)	0.54 (0.54)	0.41 (0.42)
NSAF	0.90 (0.90)	0.87 (0.88)	0.84 (0.89)	-	0.51 (0.53)	0.42 (0.44)

Source: Ning, Kang, Damian Fermin, et al. "Comparative Analysis of Different Label-free Mass Spectrometry Based Protein Abundance Estimates and Their Correlation with RNA-Seq Gene Expression Data." *Journal of Proteome Research* 11, no. 4 (2012): 2261-71.

Kang Ning, Damian Fermin, and Alexey I. Nesvizhskii J Proteome Res. 2012 April 6; 11(4): 2261–2271.

L18 Chromatin and DNase-seq Analysis



Mutant

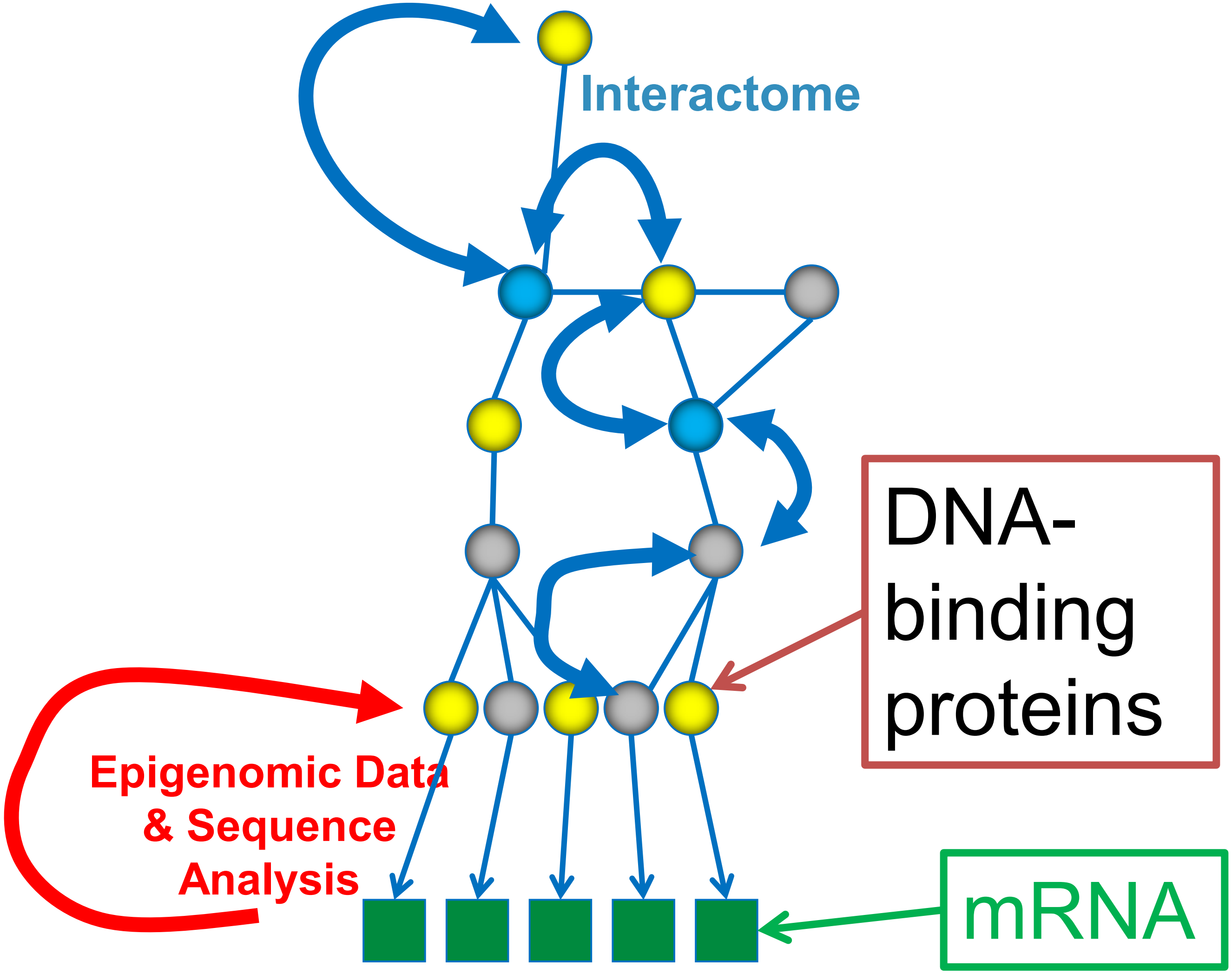
Wild-type

Sequence Analysis

Logo
M00513
M01267
M01308
M01214

Move upstream of transcription

Network integration



Interactome

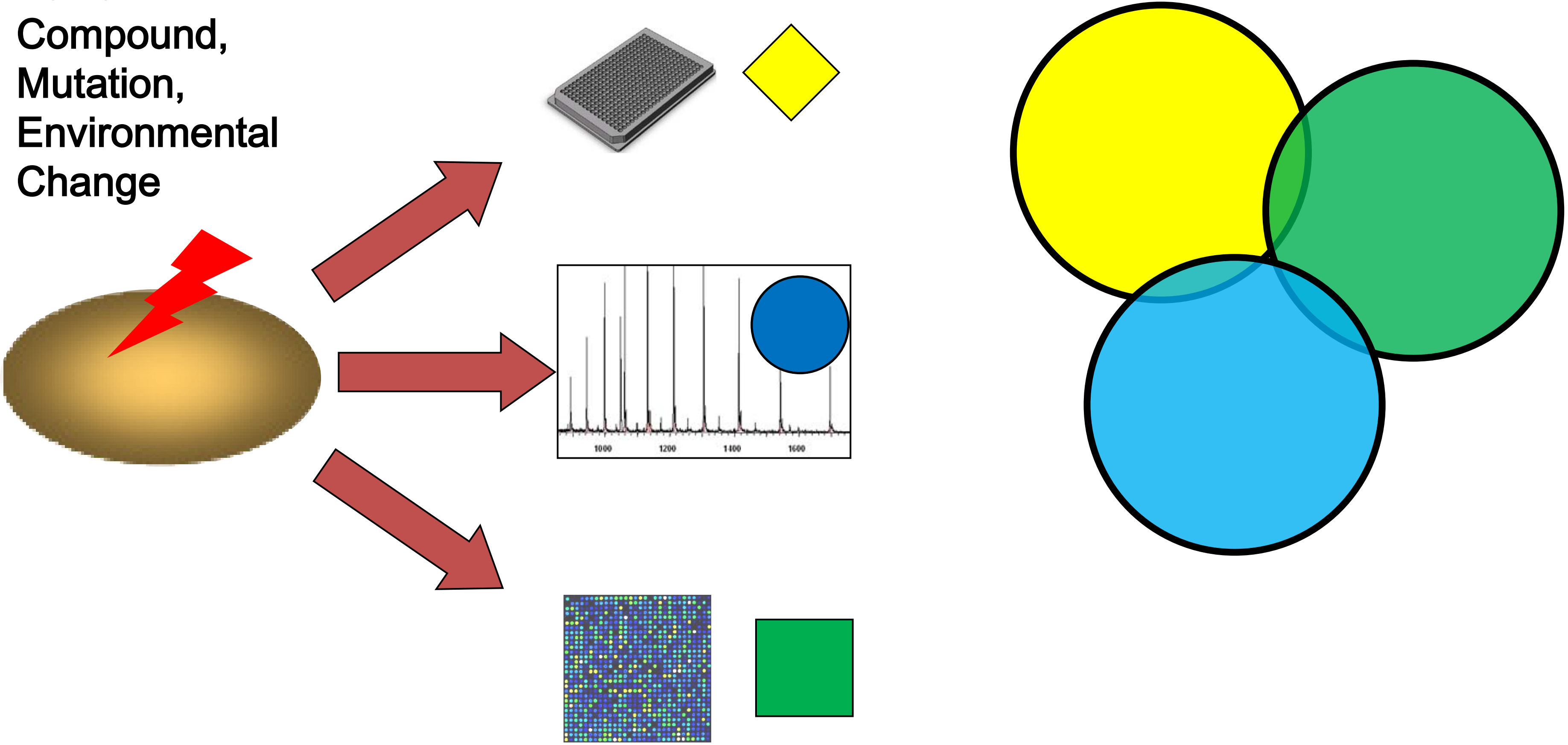
DNA-binding proteins

Epigenomic Data & Sequence Analysis

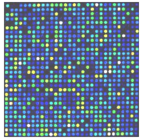
mRNA

‘Omic data don’t agree

Toxic
Compound,
Mutation,
Environmental
Change



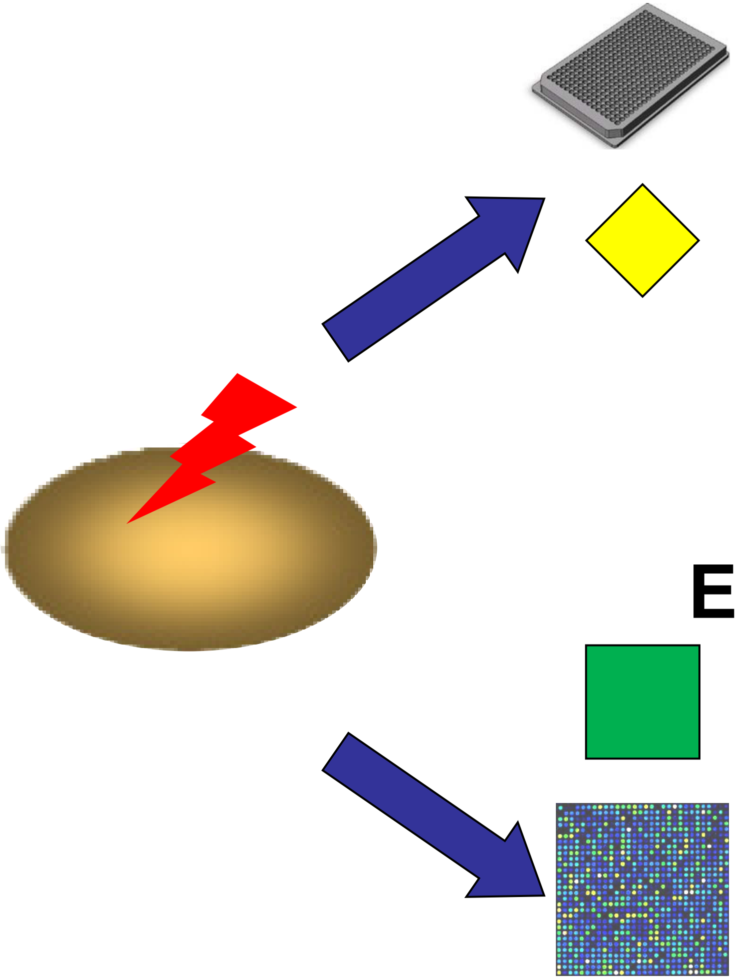
Genetic vs. Expression Data



Perturbation	Differentially expressed genes	Genetic hits	Number of overlapping genes
Growth arrest (Hydroxyurea)	59	86	0
DNA damage (MMS)	198	1448	43
Protein biosynthesis block (Cycloheximide)	20	164	0
ER stress (Tunicamycin)	200	127	5
ATP synthesis block (Arsenic)	828	50	9
Fatty acid metabolism (oleate)	269	103	9
Gene inactivation (24 datasets, median shown)	27	130	0

Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity
Nature Genetics Published online: 22 February 2009

For 156 perturbations:



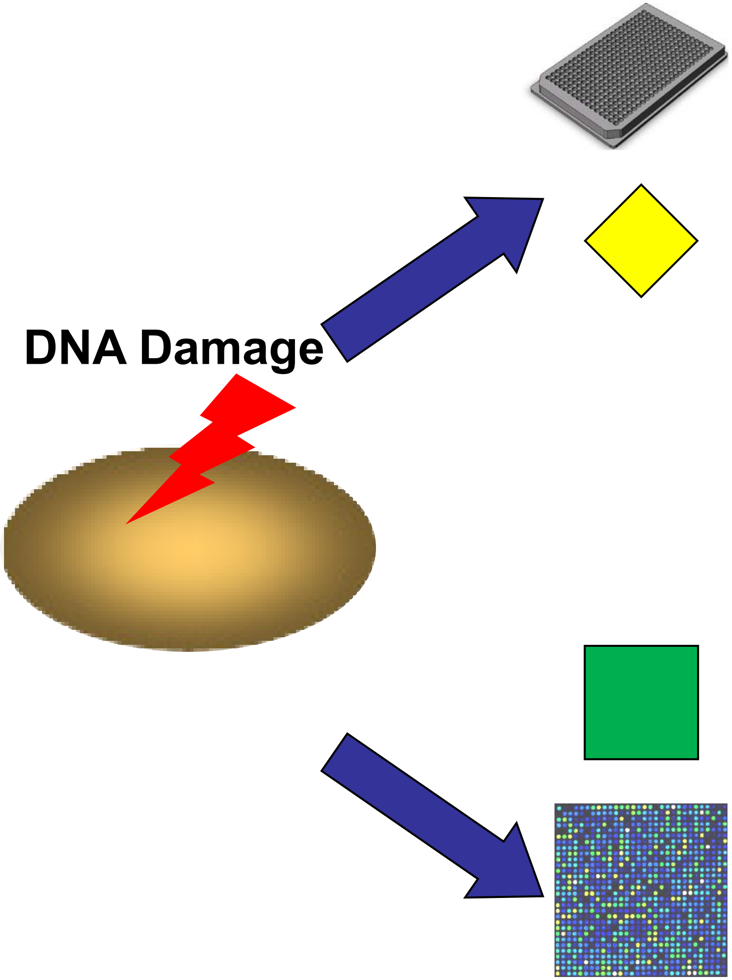
Genetic Data Enriched for:

- Transcriptional regulation
- Signal transduction

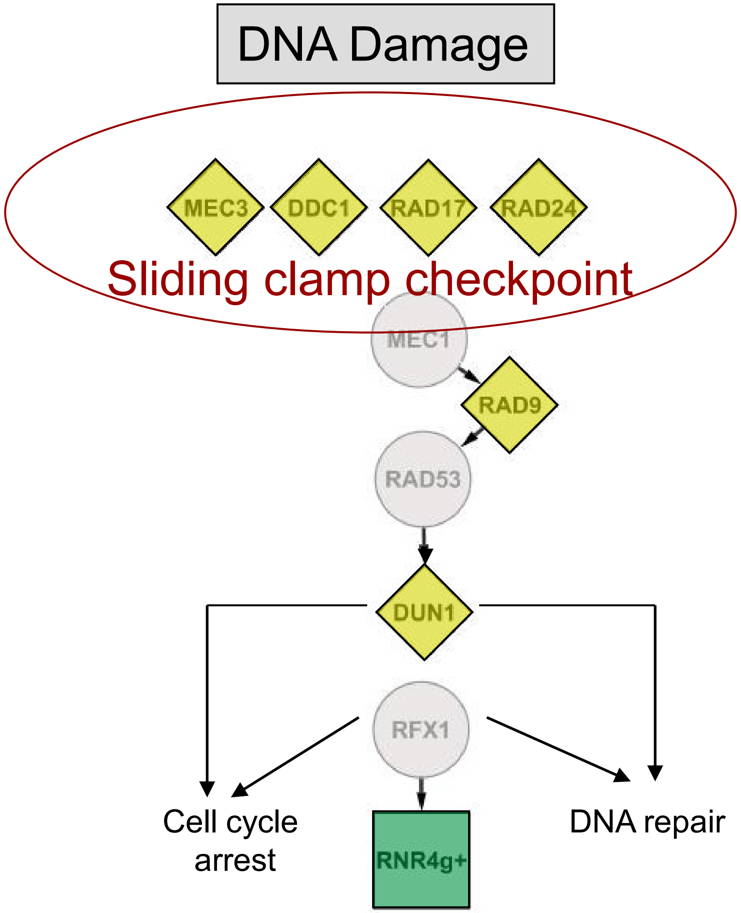
Expression Data Enriched for:

Metabolic Processes
e.g., organic acid
metabolic process,
oxidoreducatse activities

Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity
Nature Genetics Published online: 22 February 2009



DNA Damage



DNA Damage

MEC3 DDC1 RAD17 RAD24

Sliding clamp checkpoint

MEC1

RAD9

RAD53

DUN1

RFX1

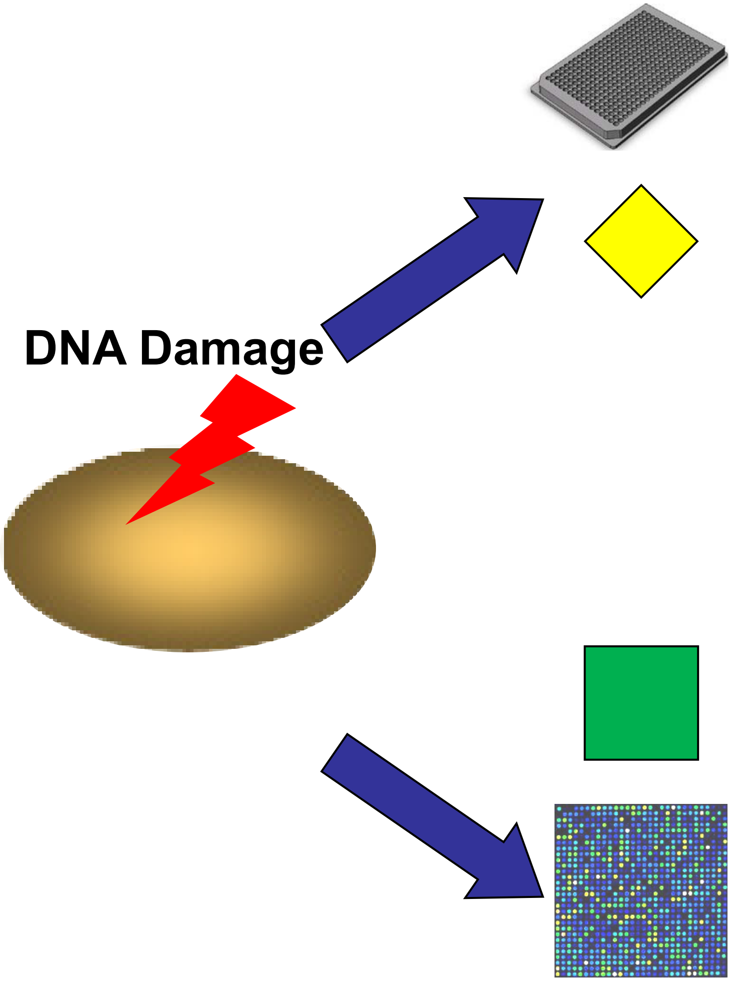
Cell cycle arrest

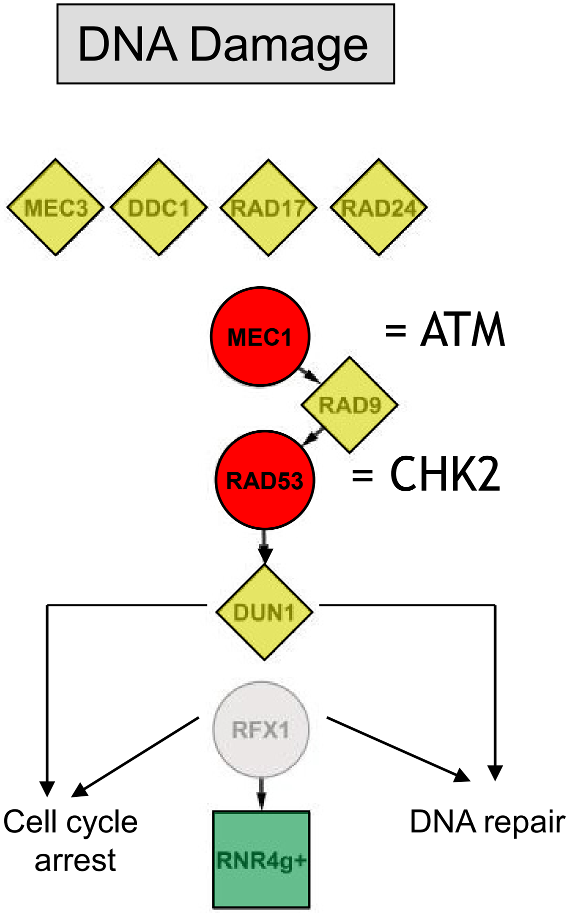
RNR4g+

DNA repair

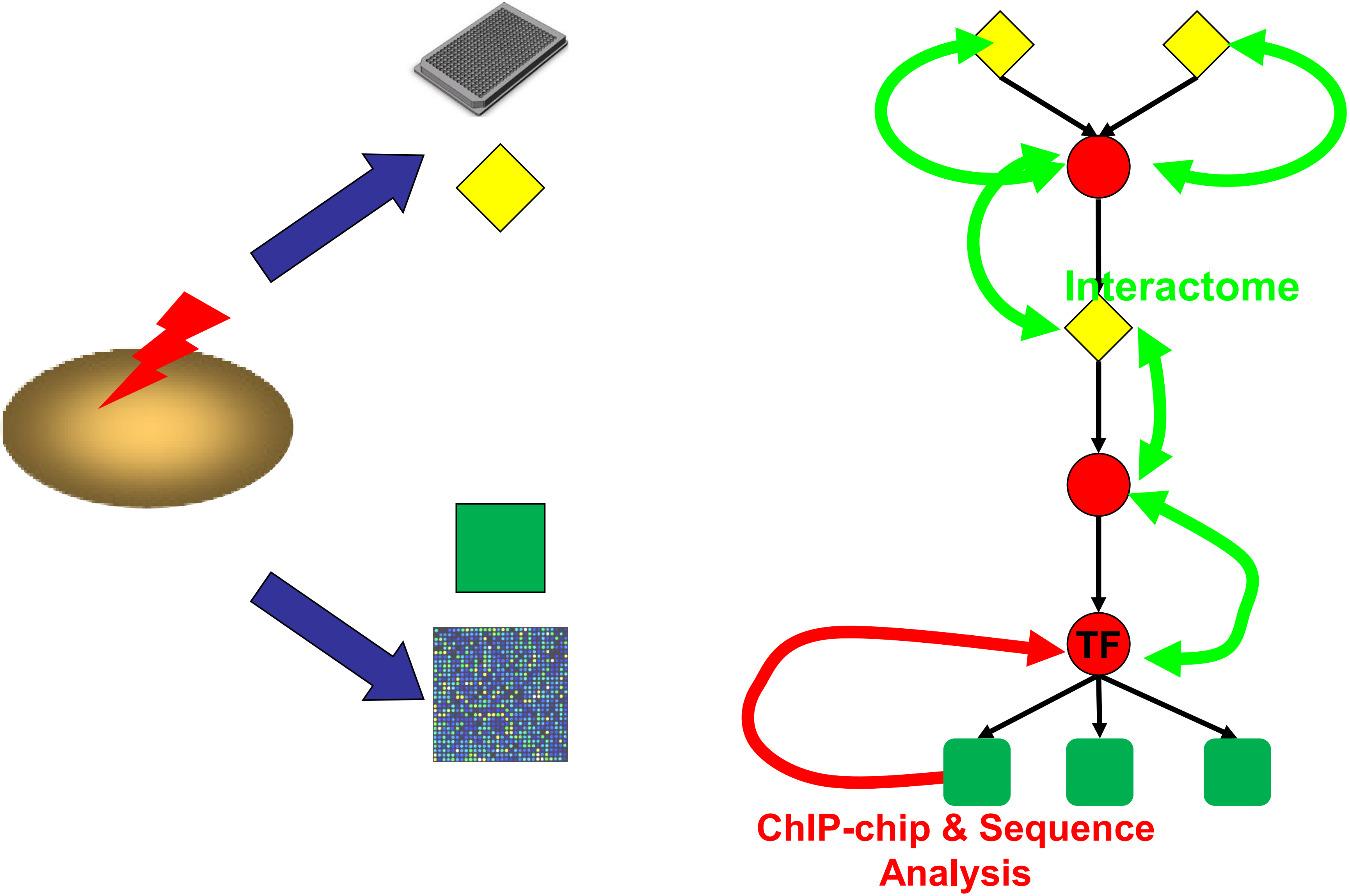
Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity

Nature Genetics Published online: 22 February 2009





Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity
Nature Genetics Published online: 22 February 2009



Interactome

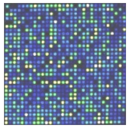
ChIP-chip & Sequence Analysis

Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity
Nature Genetics Published online: 22 February 2009

Test case: Perturbing pheromone response pathway

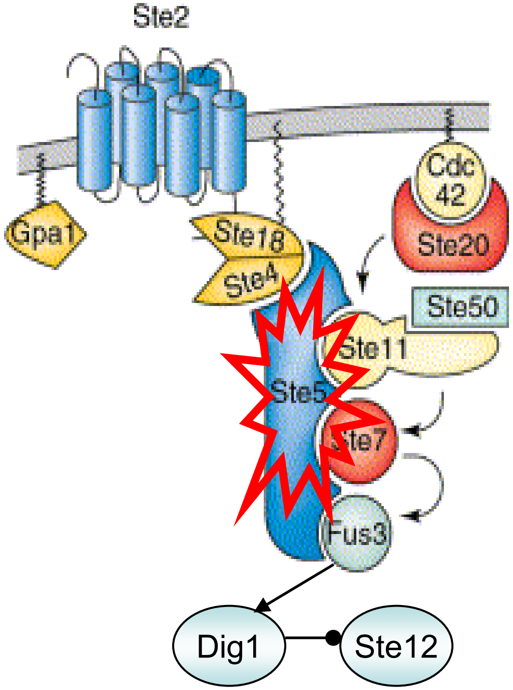
Perturbing Ste5

20 genes rescue mating phenotype (SGD)



12 genes differentially expressed

(Rosetta compendium)

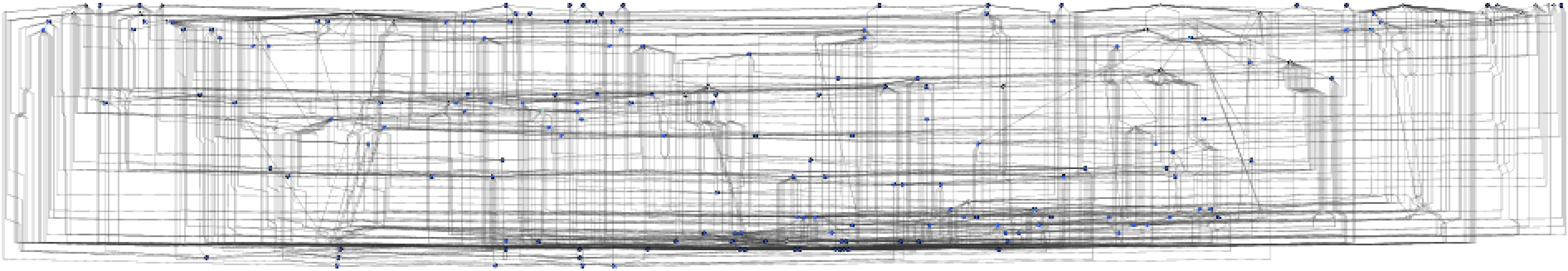


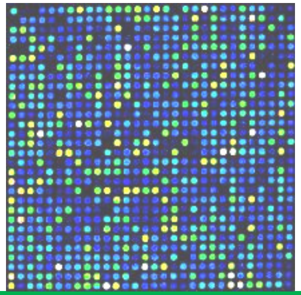
Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity
Nature Genetics Published online: 22 February 2009

Δste5: Naïve approach

Paths limited to length 3

Genetic Data



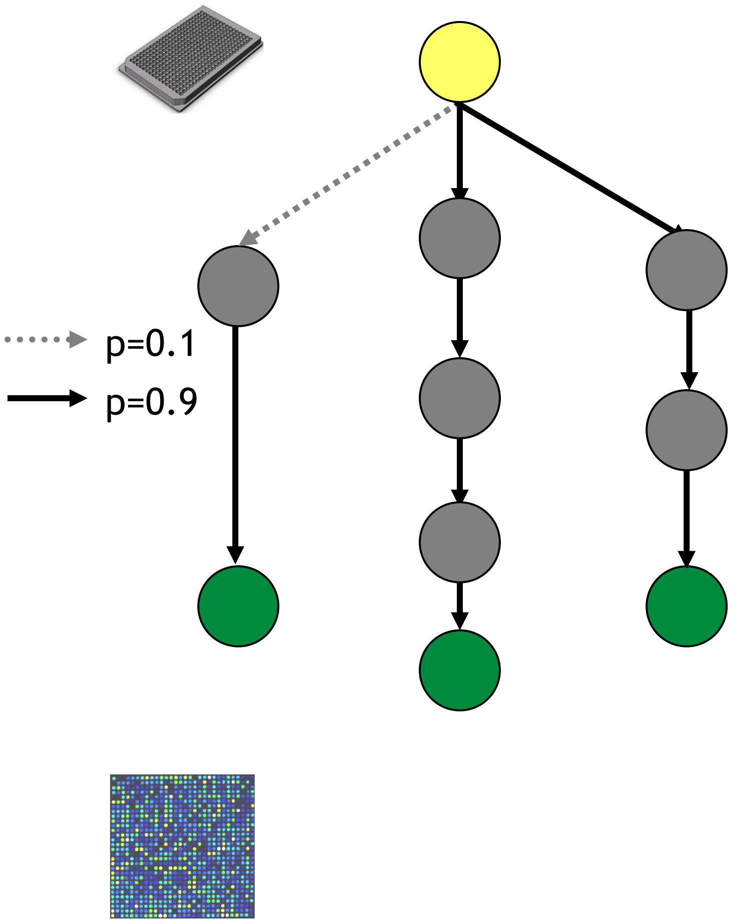


Expression Data

193 nodes, 778 edges

Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity
Nature Genetics Published online: 22 February 2009

Maximize the connectivity via reliable paths



Goal: find paths that maximize product of P_{ij}

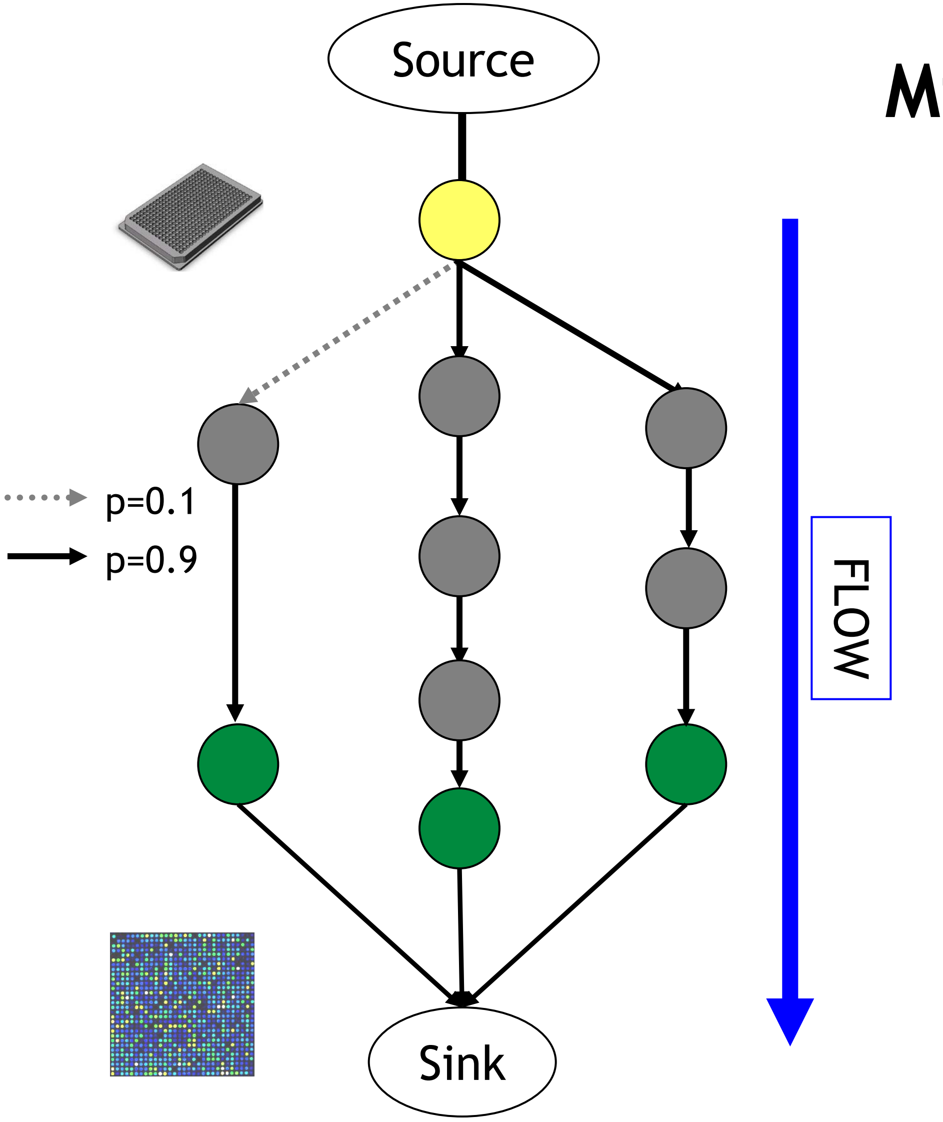
Assign probabilities using a Bayesian approach based on reliability of underlying data type:

Myers, C.L. et al. Genome Biology (2005).

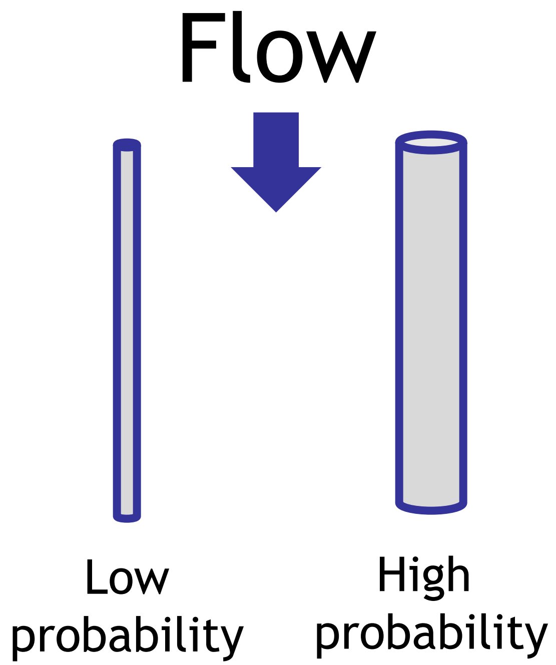
Jansen, R. et al. Science (2003).

Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity
Nature Genetics Published online: 22 February 2009

Maximize the connectivity via reliable paths

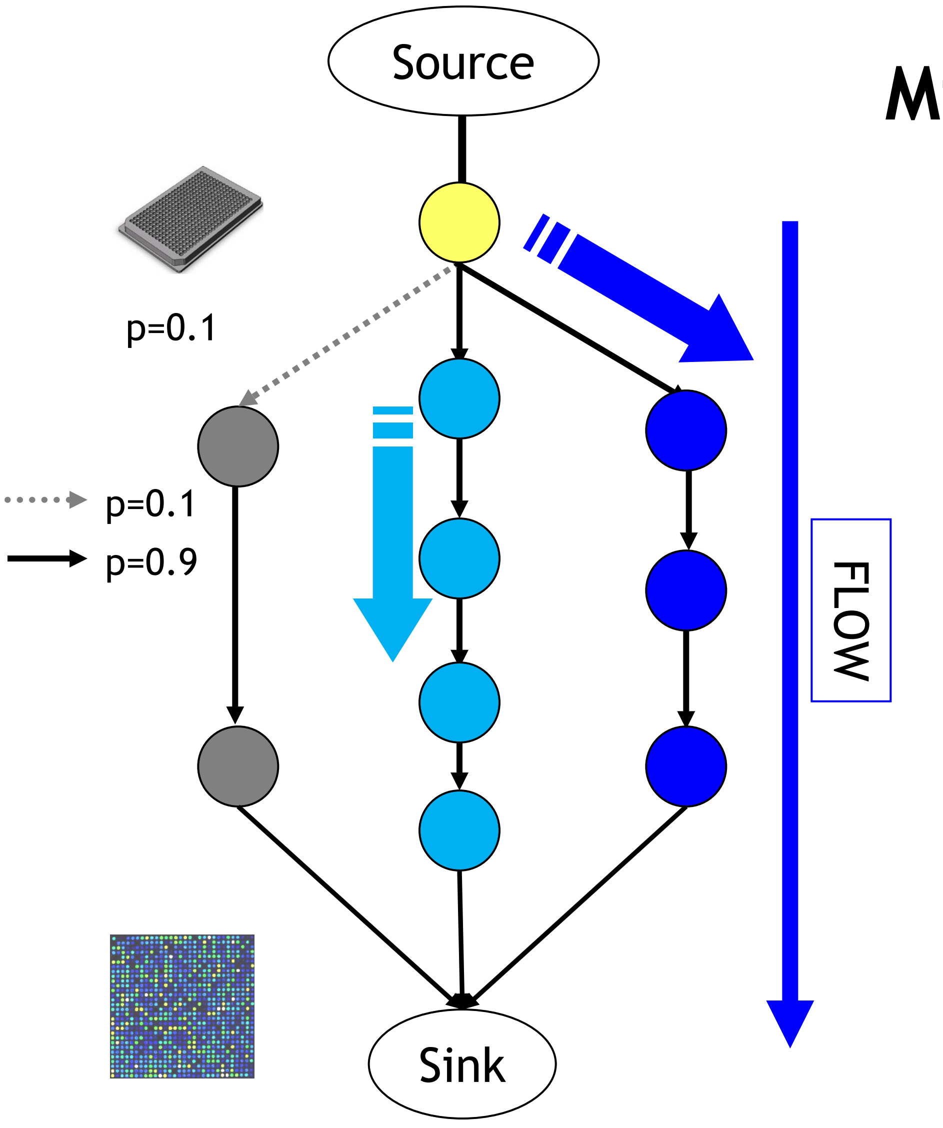


Minimum cost flow problem

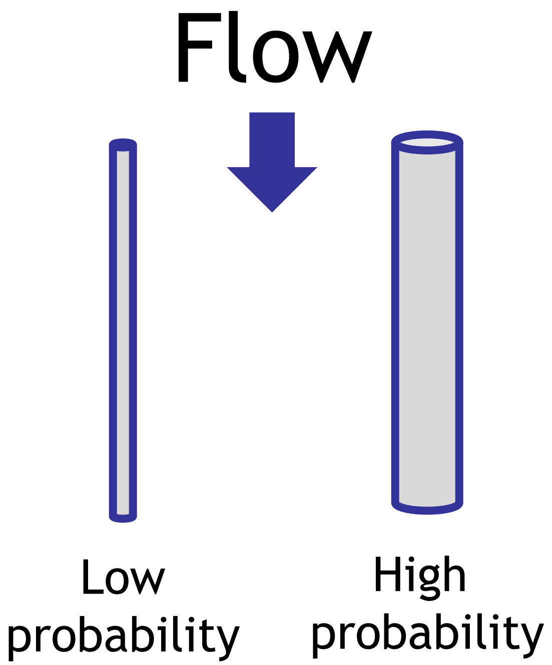


Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity
Nature Genetics Published online: 22 February 2009

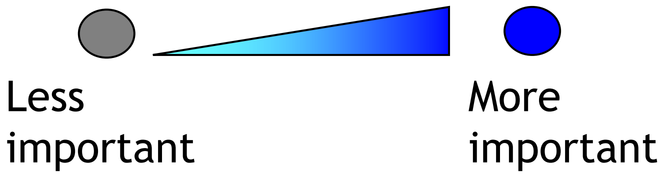
Maximize the connectivity via reliable paths



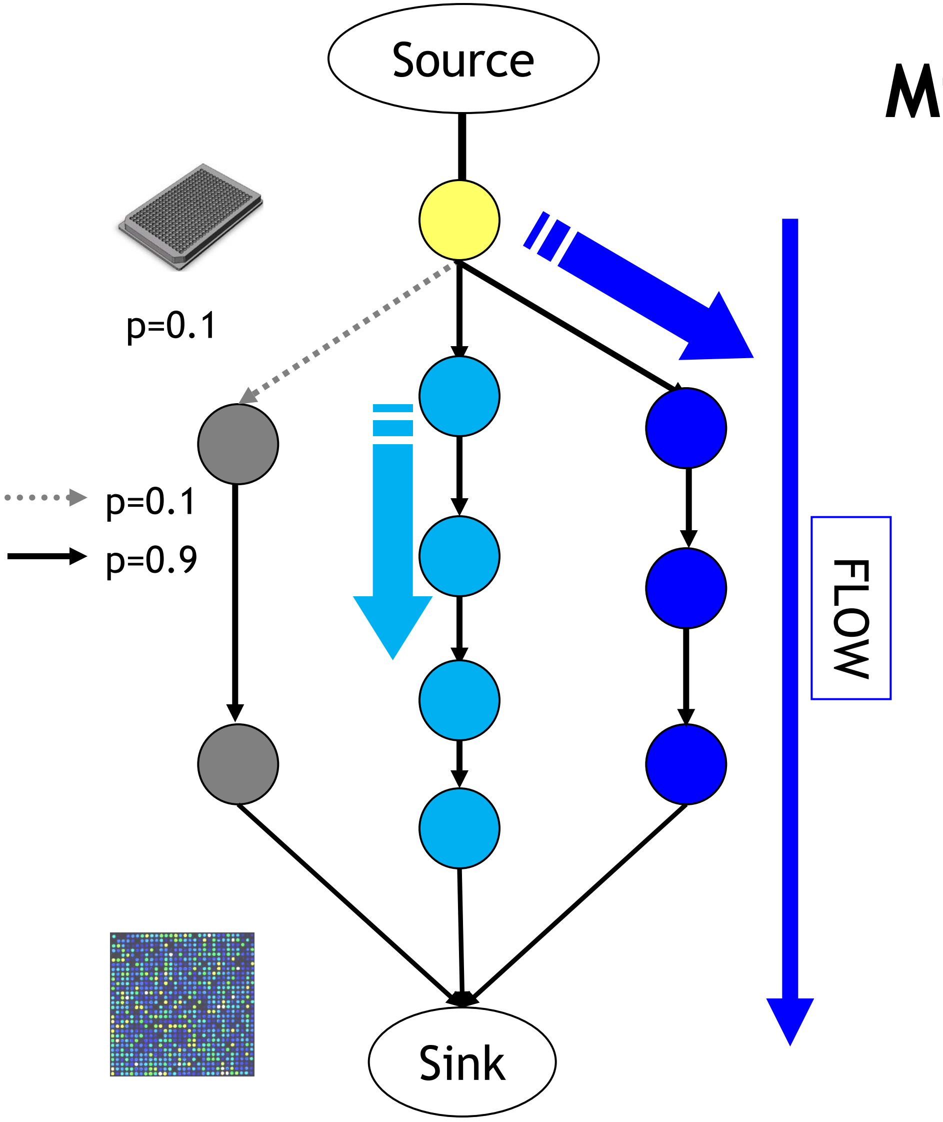
Minimum cost flow problem



Proteins ranked by their incoming flow:



Maximize the connectivity via reliable paths



Minimum cost flow problem

Maximize flow: source to sink

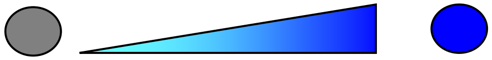
Minimize cost $(e_{ij}) = f_{ij} *(-\log P_{ij})$

min $(\Sigma cost(e_{ij}) - \gamma * \Sigma f_{Sj})$

f_{ij} = flow through e_{ij}

c_{ij} = capacity of e_{ij} = 1 for all e_{ij}

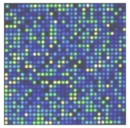
Proteins ranked by their incoming flow:



Test case: Perturbing pheromone response pathway

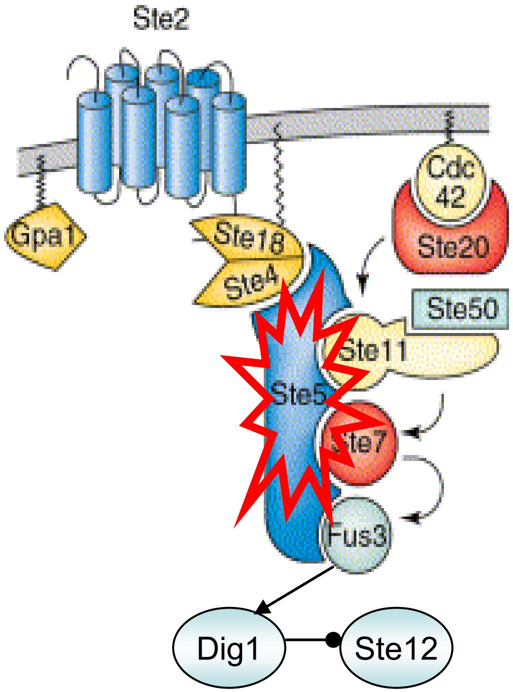
Perturbing Ste5

20 genes rescue mating phenotype (SGD)



12 genes differentially expressed

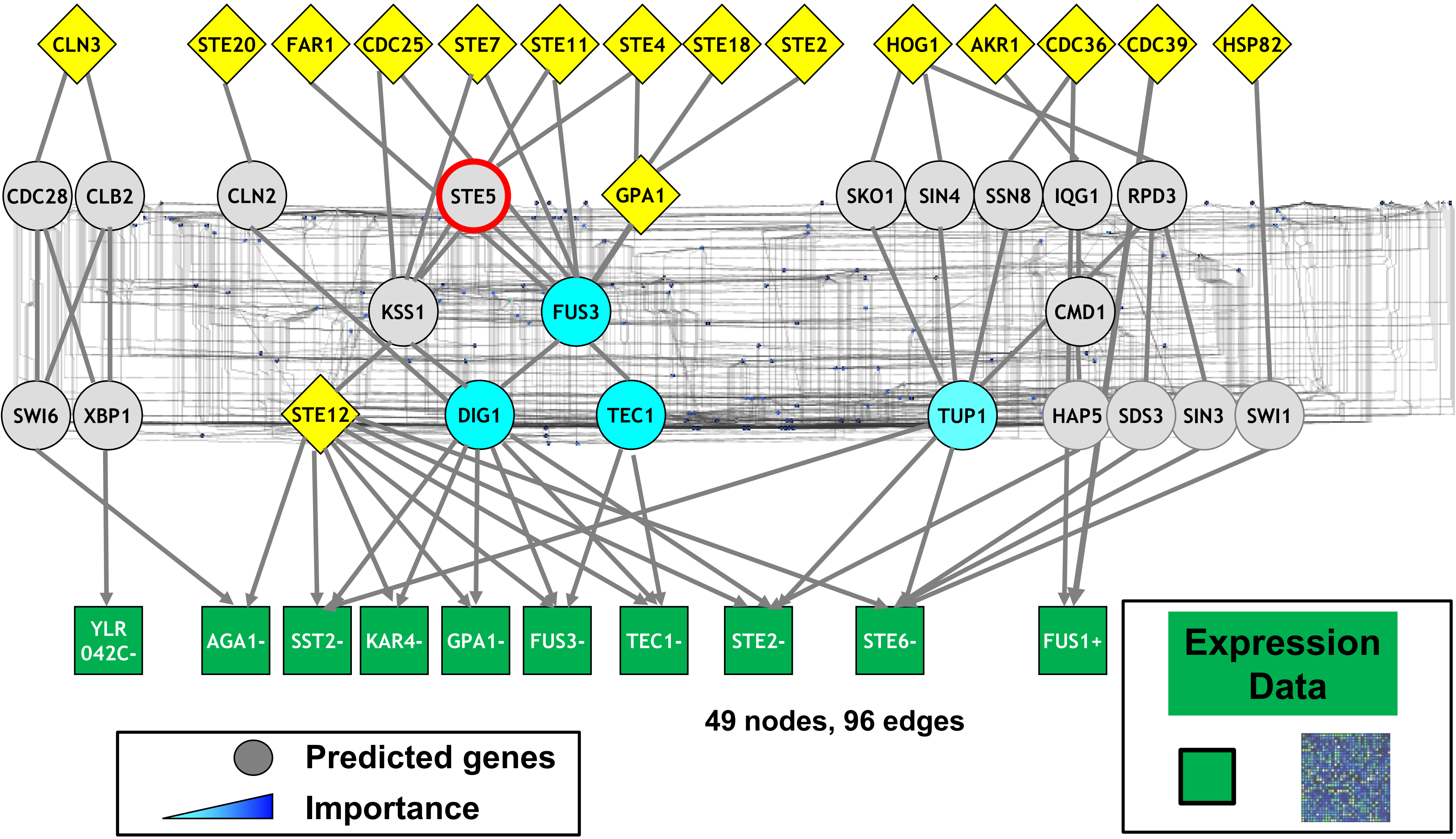
(Rosetta compendium)



Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity
Nature Genetics Published online: 22 February 2009

Enriched for pheromone response $p<10^{-18}$

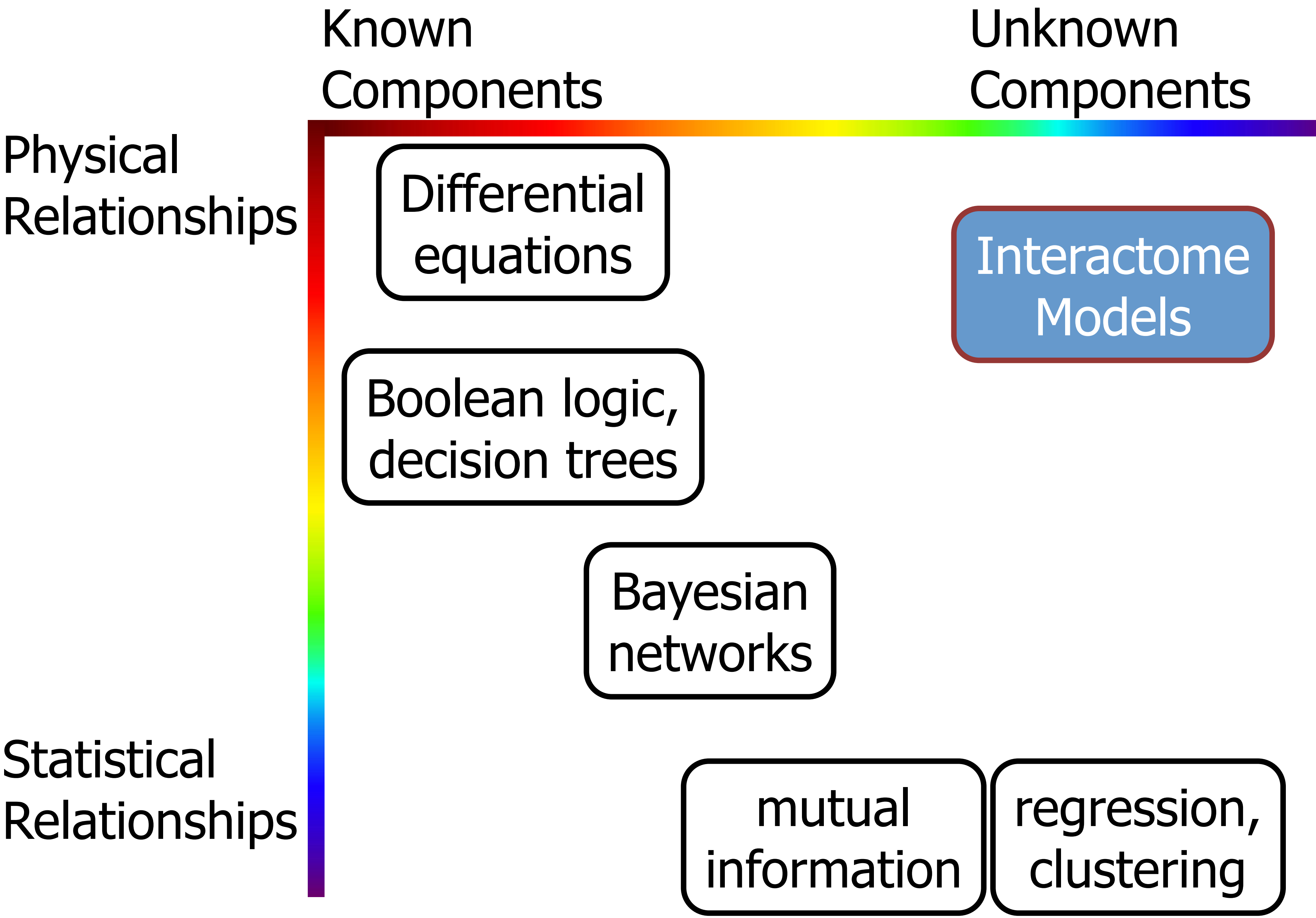
Genetic Data



49 nodes, 96 edges

Network Models

- Structure of network
  - Coexpression
  - Mutual information
  - Physical/genetic interactions
- Analysis of network
  - Ad hoc
  - Shortest path
  - Clustering
  - Optimization



Known Components

Unknown Components

Physical Relationships

Statistical Relationships

Differential equations

Interactome Models

Boolean logic, decision trees

Bayesian networks

mutual information

regression, clustering

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014