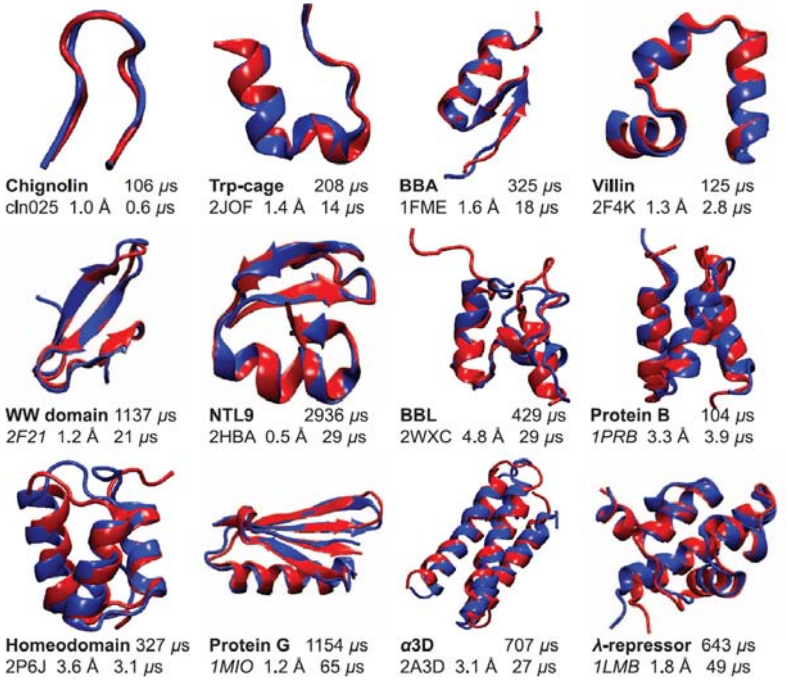
- L12 - Introduction to Protein Structure; Structure Comparison & Classification
- L13 - Predicting protein structure
- L14 - Predicting protein interactions
- L15 - Gene Regulatory Networks
- L16 - Protein Interaction Networks
- L17 - Computable Network Models

Outline

- Bayesian Networks for PPI prediction
- Gene expression
  - Distance metrics
  - Clustering
  - Signatures
  - Modules
    - Bayesian networks
    - Regression
    - Mutual Information
    - Evaluation on real and simulated data

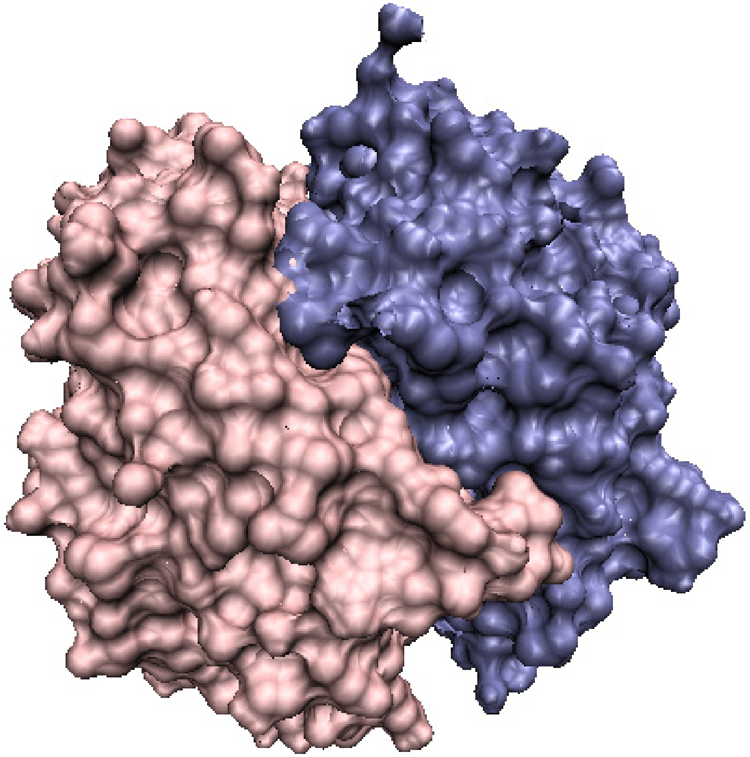
Predictions

Last time: protein structure

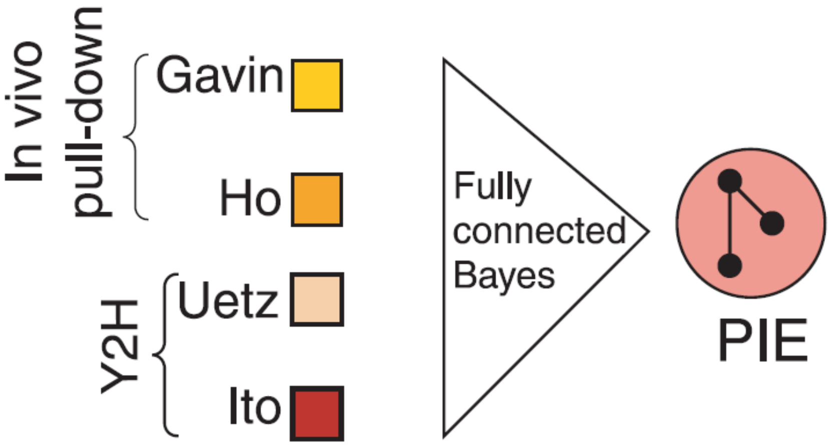


Source: Lindorff-Larsen, Kresten, Stefano Piana, et al. "How Fast-folding Proteins Fold." *Science* 334, no. 6055 (2011): 517-20.

Now: protein interactions

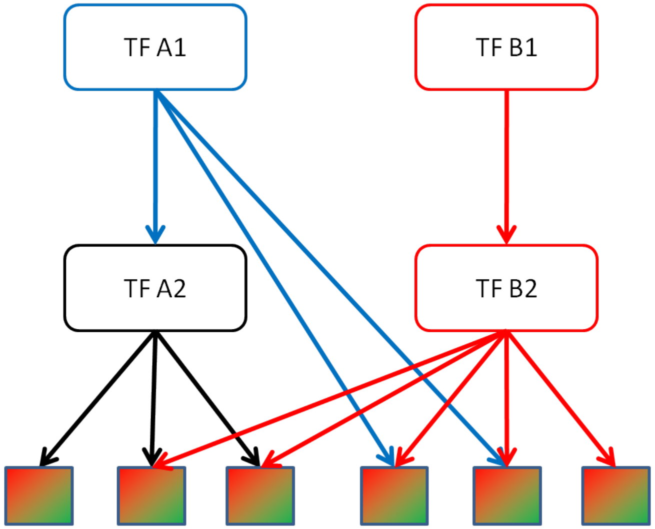


Bayesian Networks



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.

Predict unknown variables from observations

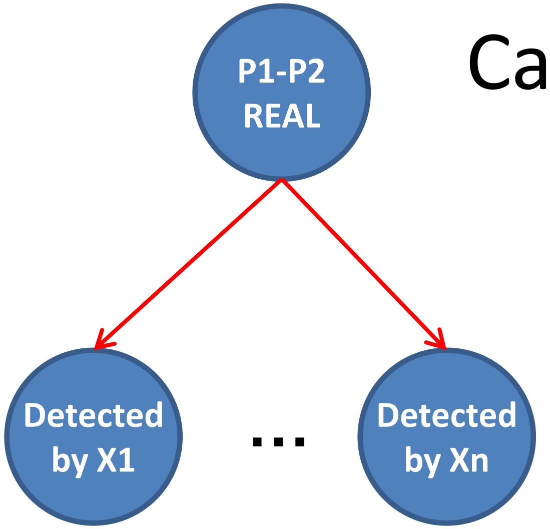


A "natural" way to think about biological networks.

Bayesian Networks

- Bayesian Networks are a tool for reasoning with probabilities
- Consist of a graph (network) and a set of probabilities
- These can be “learned” from the data

Graphical Structure Expresses our Beliefs



P1-P2 REAL

Detected by X1

…

Detected by Xn

Cause (often “hidden”)

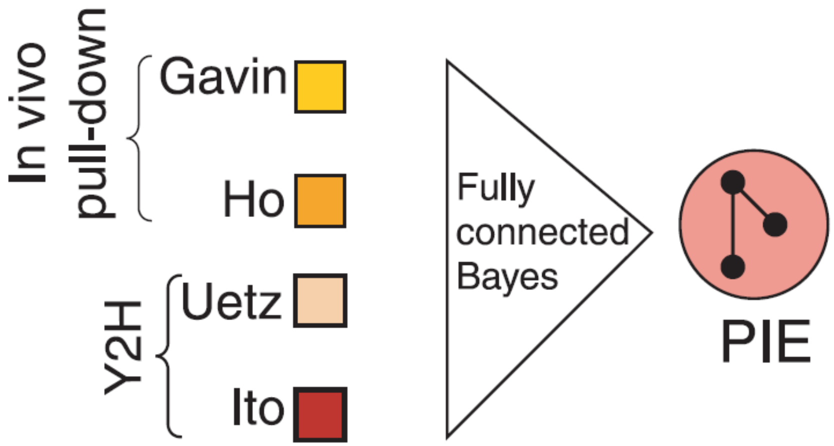
Effects (observed)

How do we obtain a BN?

- Two problems:
  - learning graph structure
    - NP-complete
    - approximation algorithms
  - probability distributions

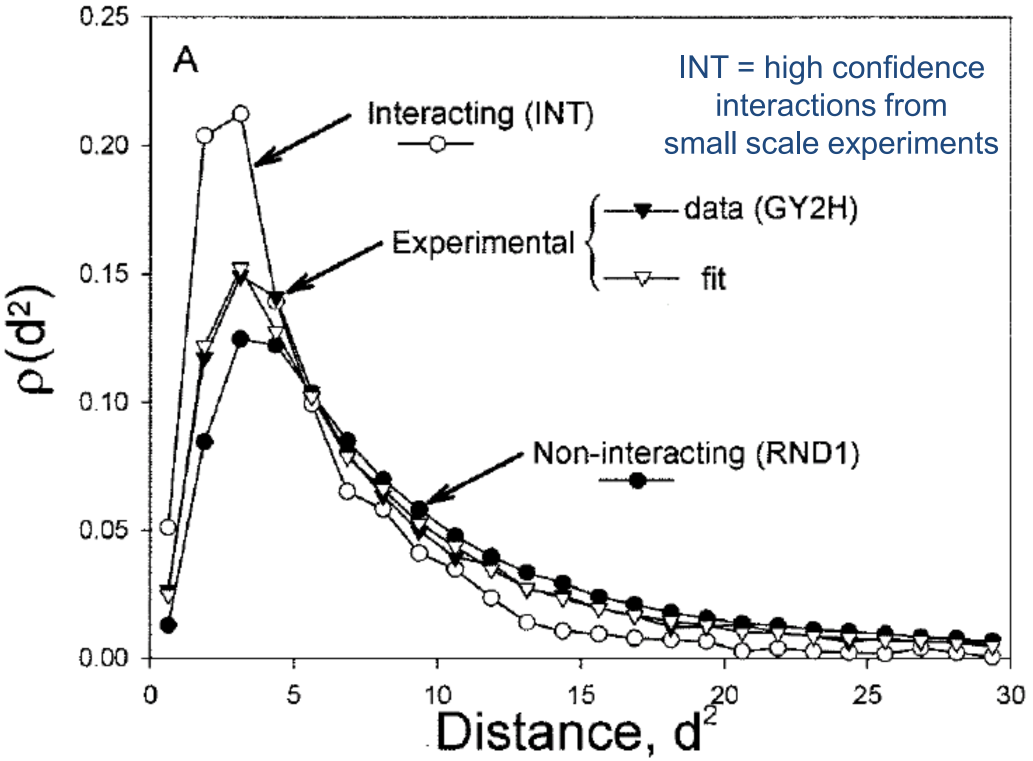
Goal

- What other data could help?



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.

Properties of real interactions: correlated expression
Expression Profile Reliability (EPR)



INT = high confidence interactions from small scale experiments

Source: Deane, Charlotte M., Łukasz Salwiński, et al. "Protein Interactions Two Methods for Aassessment of the Reliability of High Throughput Observations." *Molecular & Cellular Proteomics* 1, no. 5 (2002): 349-56.

d = "distance" that measures the difference between two mRNA expression profiles

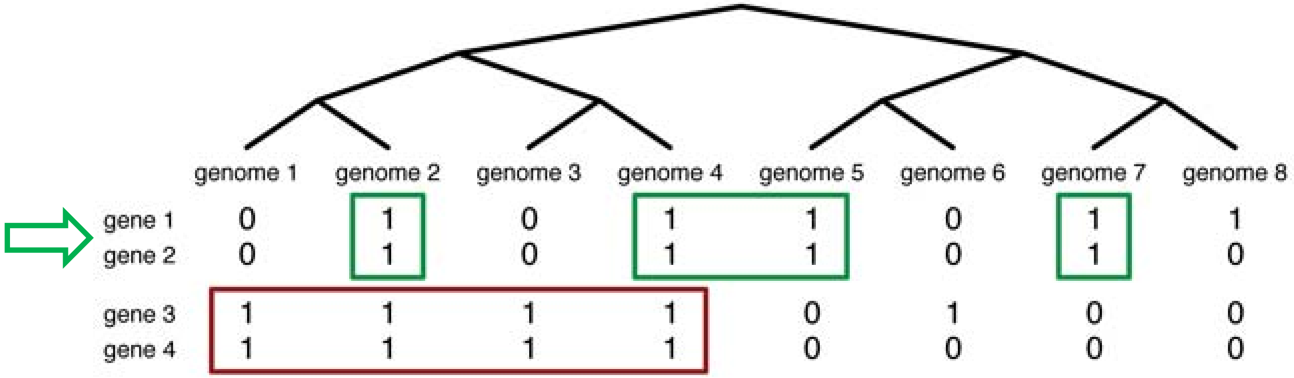
Note: proteins involved in "true" protein-protein interactions have more similar mRNA expression profiles than random pairs. Use this to assess how good an experimental set of interactions is.

Deane et al. Mol. & Cell. Proteomics (2002) 1.5, 349-356

Co-evolution

Which pattern below is more likely to represent a pair of interacting proteins?

More likely to interact



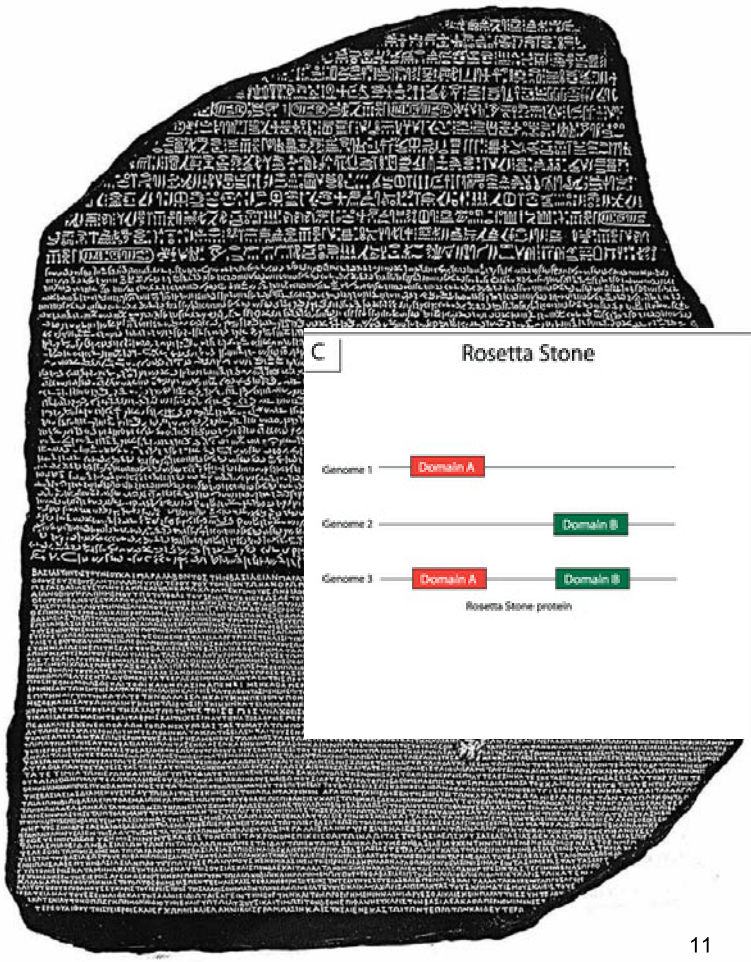
	genome 1	genome 2	genome 3	genome 4	genome 5	genome 6	genome 7	genome 8
gene 1	0	1	0	1	1	0	1	1
gene 2	0	1	0	1	1	0	1	0
gene 3	1	1	1	1	0	1	0	0
gene 4	1	1	1	1	0	0	0	0

Source: Cokus, Shawn, Sayaka Mizutani, et al. "An Improved Method for Identifying Functionally Linked Proteins Using Phylogenetic Profiles." *BMC Bioinformatics* 8, no. Suppl 4 (2007): S7.

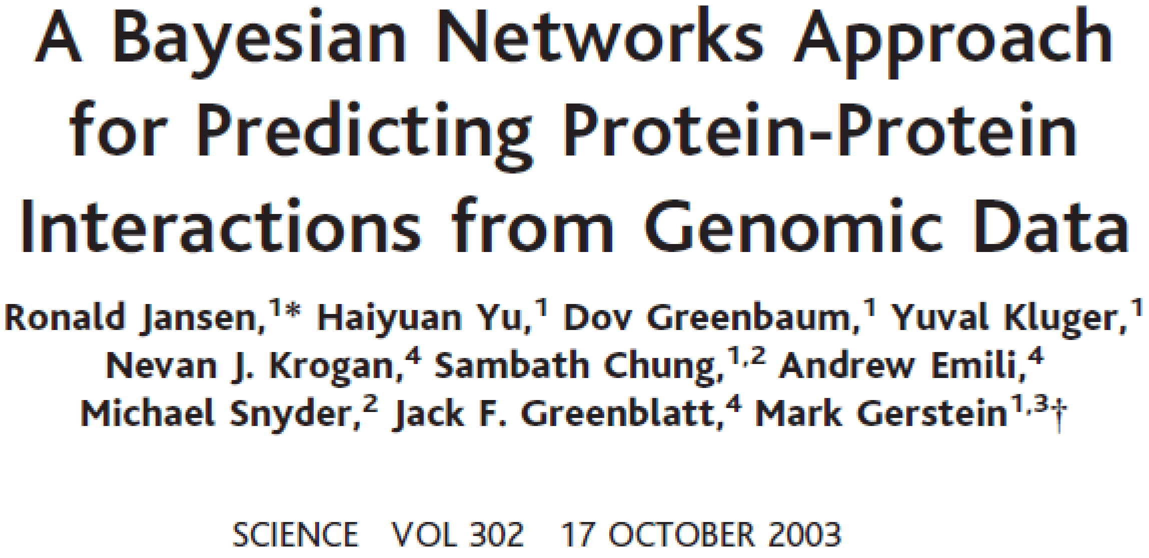
Cokus *et al. BMC Bioinformatics* 2007 **8**(Suppl 4):S7 doi:10.1186/1471-2105-8-S4-S7

Rosetta Stone

- Look for genes that are fused in some organisms
  - Almost 7,000 pairs found in *E. coli.*
  - >6% of known interactions can be found with this method
  - Not very common in eukaryotes



Integrating diverse data



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†]

SCIENCE VOL 302 17 OCTOBER 2003

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

Requirement of Bayesian Classification

- Gold standard training data
  - Independent from evidence
  - Large
  - No systematic bias

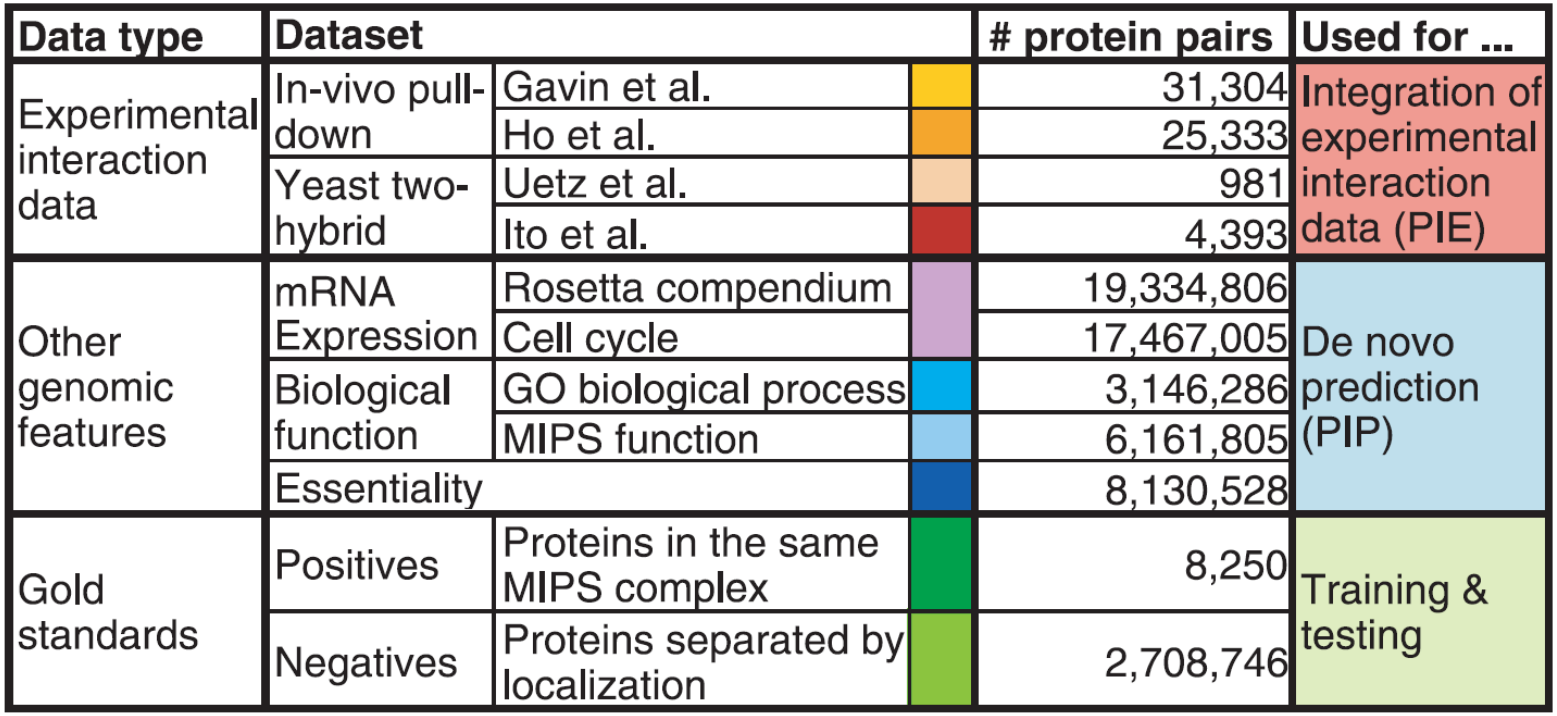
Positive training data: MIPS

- Hand-curated from literature

Negative training data:

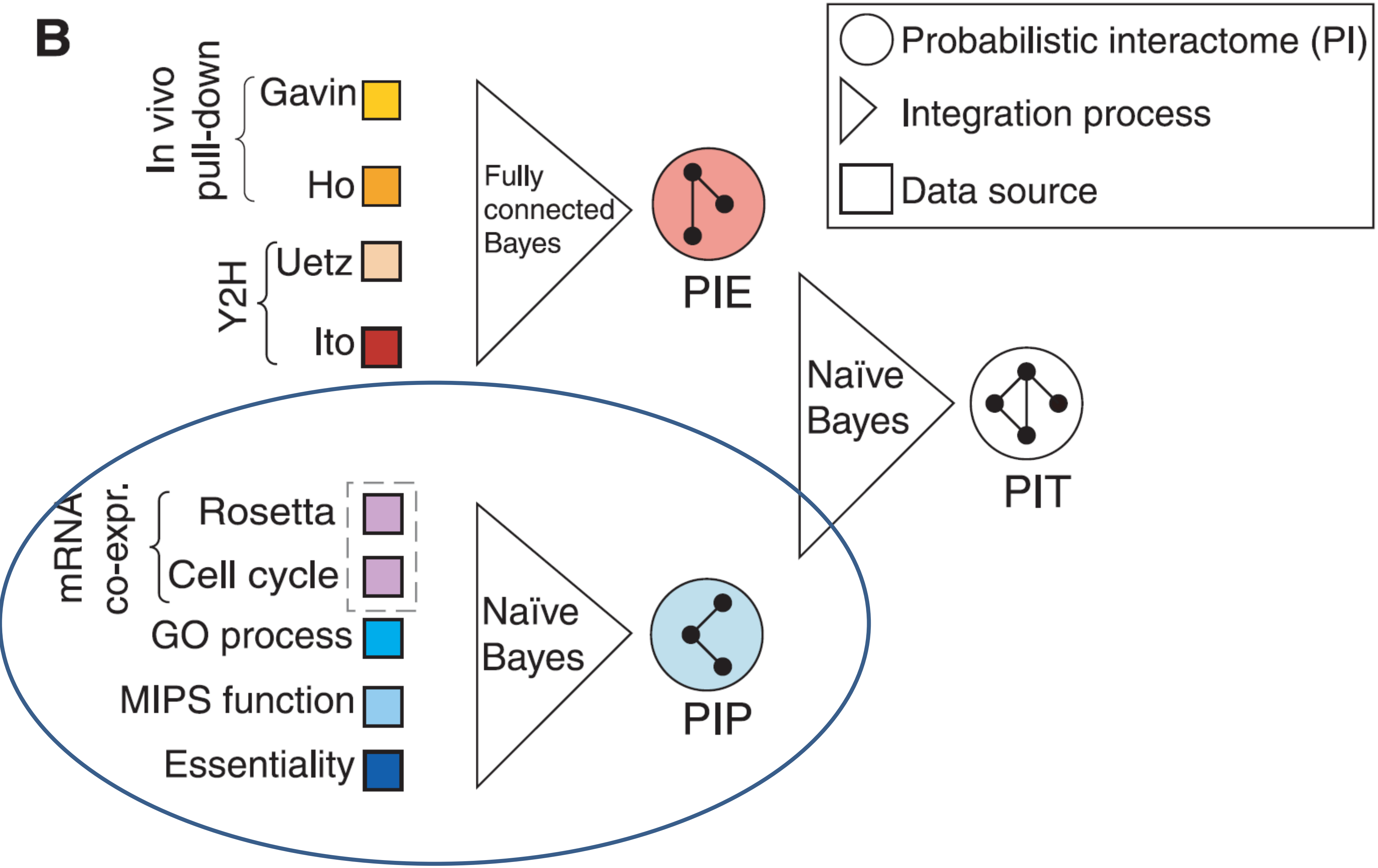
- Proteins in different subcellular compartments

Integrating diverse data



Data type	Dataset			# protein pairs	Used for ...
Experimental interaction data	In-vivo pull-down	Gavin et al.	🟨	31,304	Integration of experimental interaction data (PIE)
		Ho et al.	🟧	25,333	
	Yeast two-hybrid	Uetz et al.	🟧	981	
		Ito et al.	🟥	4,393	
Other genomic features	mRNA Expression	Rosetta compendium	🟪	19,334,806	De novo prediction (PIP)
		Cell cycle	🟪	17,467,005	
	Biological function	GO biological process	🟦	3,146,286	
		MIPS function	🟦	6,161,805	
	Essentiality		🟦	8,130,528	
Gold standards	Positives	Proteins in the same MIPS complex	🟩	8,250	Training & testing
	Negatives	Proteins separated by localization	🟩	2,708,746	

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.



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.

likelihood ratio =

if > 1 classify as true
if < 1 classify as false

$$\frac{P(\text{true\_PPI}|Data)}{P(\text{false\_PPI}|Data)} = \frac{P(Data|\text{true\_PPI})P(\text{true\_PPI})}{P(Data|\text{false\_PPI})P(\text{false\_PPI})}$$

log likelihood ratio =

$$\log\left[\frac{P(\text{true\_PPI}|Data)}{P(\text{false\_PPI}|Data)}\right] = \boxed{\log\left[\frac{P(\text{true\_PPI})}{P(\text{false\_PPI})}\right]} + \log\left[\frac{P(Data|\text{true\_PPI})}{P(Data|\text{false\_PPI})}\right]$$

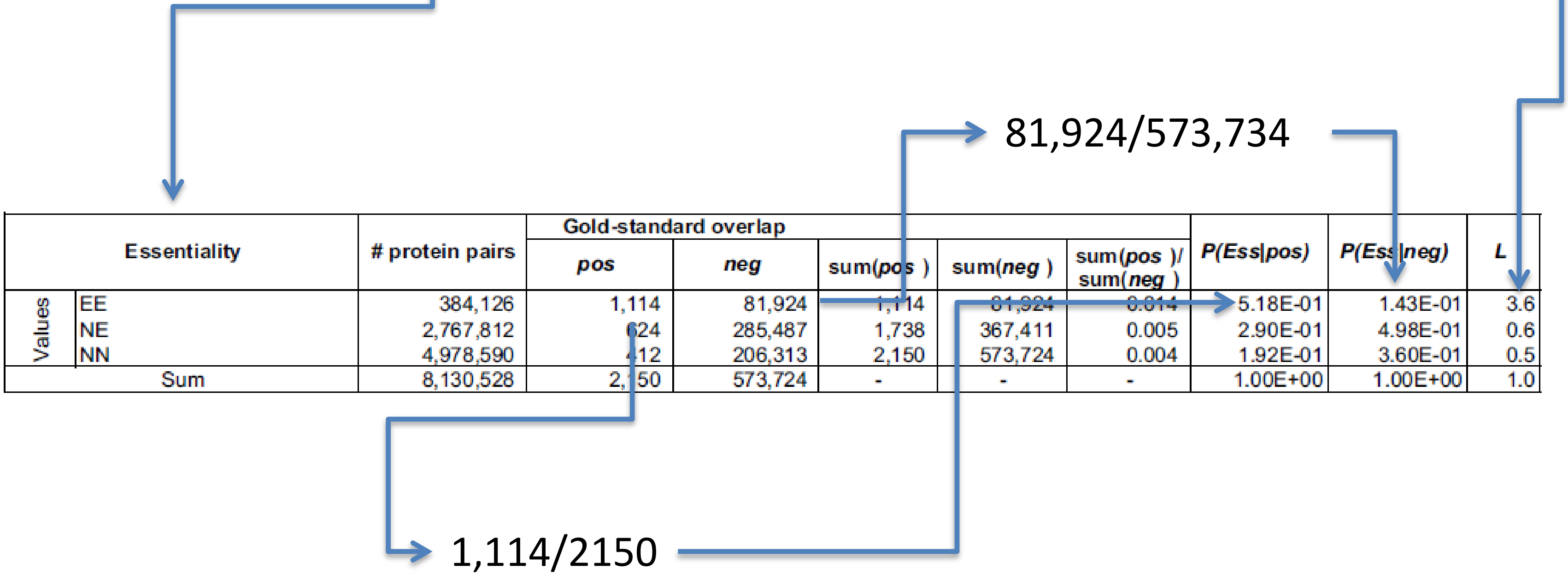
Prior probability is the same for all interactions
--does not affect ranking

Ranking function =

$$\log\left[\frac{P(Data\,|\,true\_PPI)}{P(Data\,|\,false\_PPI)}\right] = \prod_{i}^{M} \frac{P(Observation_i\,|\,true\_PPI)}{P(Observation_i\,|\,false\_PPI)}$$

Protein pairs in the essentiality data can take on three discrete values (EE, both essential; NN, both non-essential; and NE, one essential and one not)

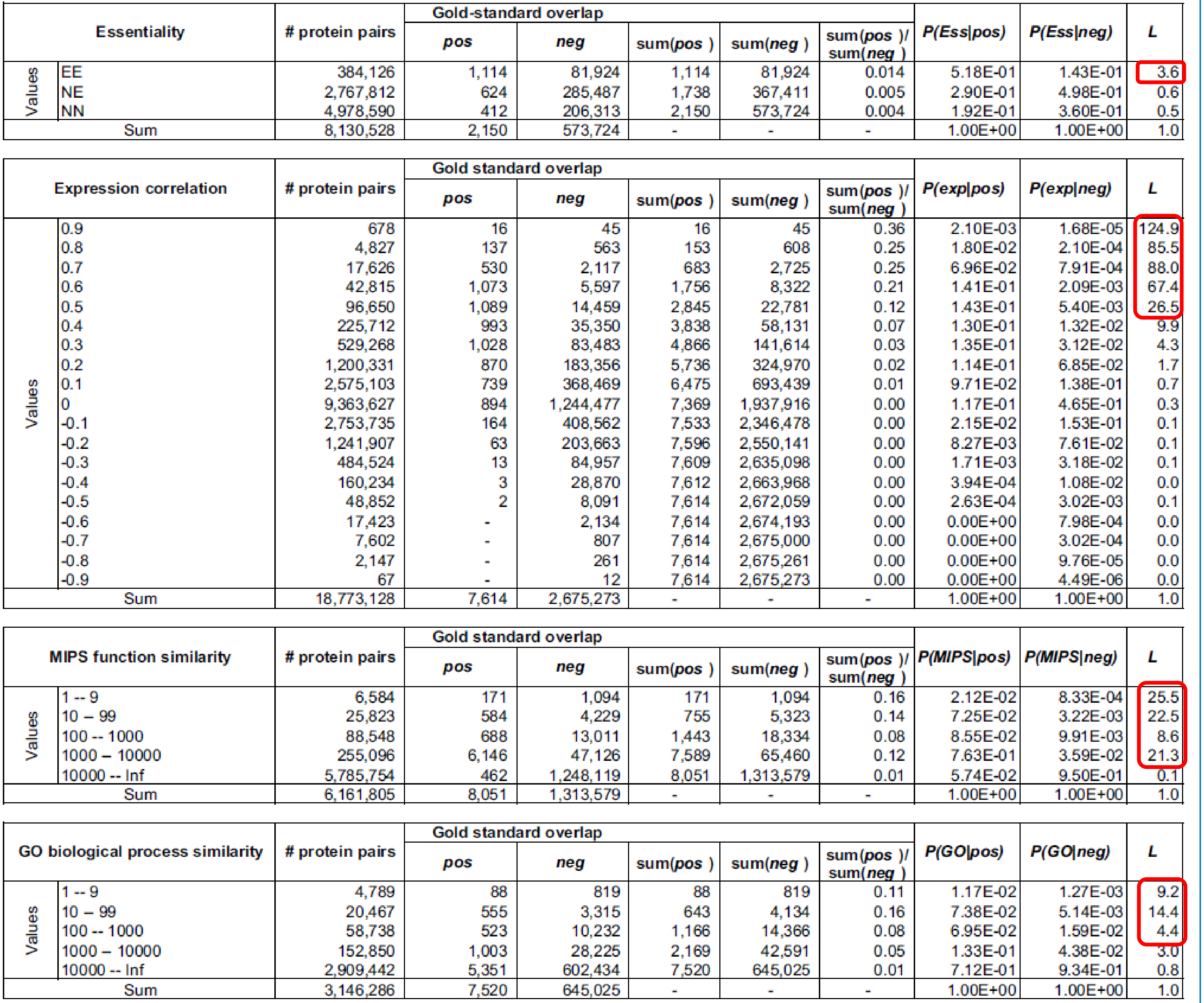
Likelihood=L= $\dfrac{P(f \mid pos)}{P(f \mid neg)}$



81,924/573,734

	Essentiality	# protein pairs	Gold-standard overlap					P(Ess|pos)	P(Ess|neg)	L
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	EE	384,126	1,114	81,924	1,114	81,924	0.014	5.18E-01	1.43E-01	3.6
	NE	2,767,812	624	285,487	1,738	367,411	0.005	2.90E-01	4.98E-01	0.6
	NN	4,978,590	412	206,313	2,150	573,724	0.004	1.92E-01	3.60E-01	0.5
Sum		8,130,528	2,150	573,724	-	-	-	1.00E+00	1.00E+00	1.0

1,114/2150

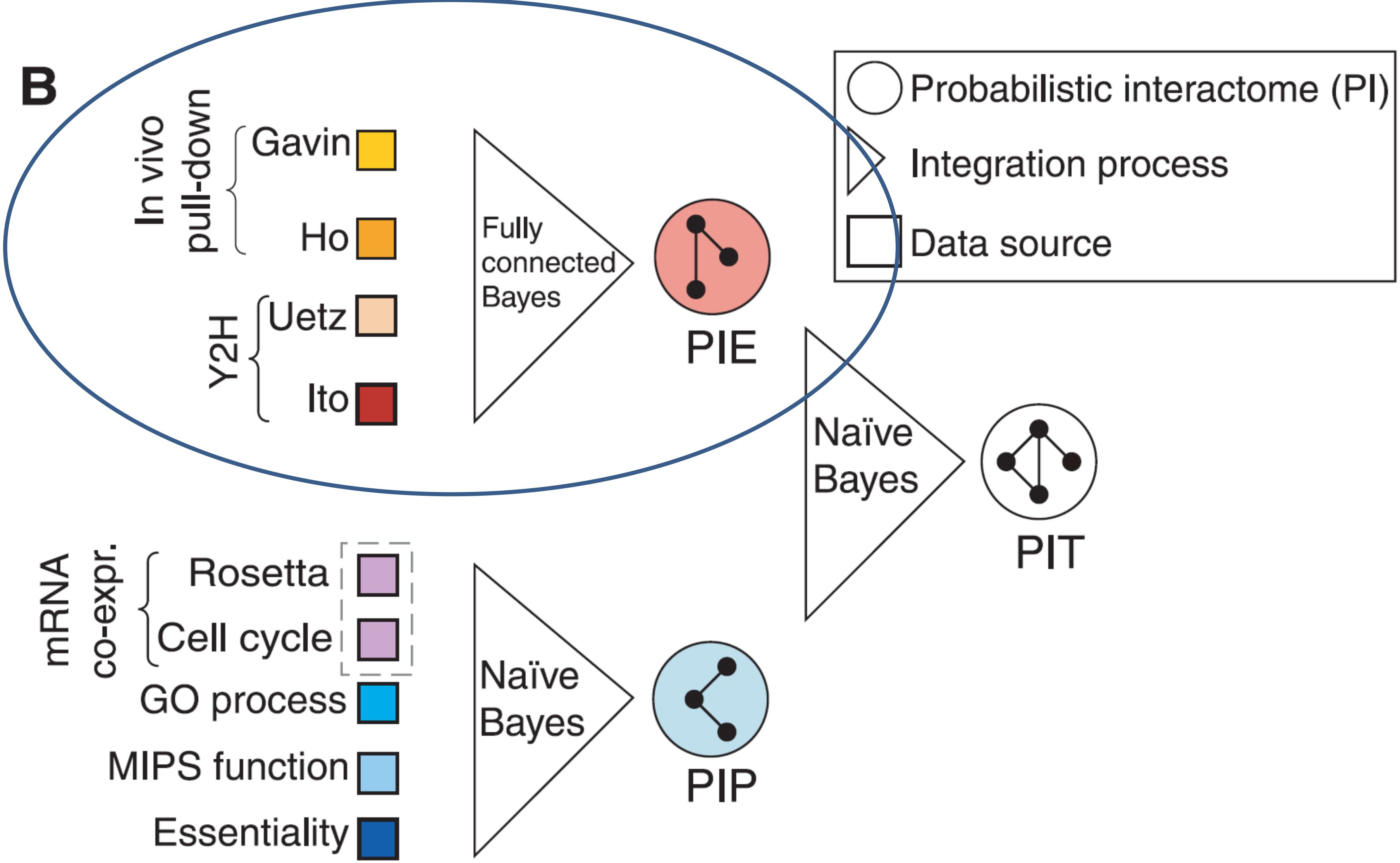


	Essentiality	# protein pairs	Gold-standard overlap					*P(Ess|pos)*	*P(Ess|neg)*	*L*
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	EE	384,126	1,114	81,924	1,114	81,924	0.014	5.18E-01	1.43E-01	3.6
	NE	2,767,812	624	285,487	1,738	367,411	0.005	2.90E-01	4.98E-01	0.6
	NN	4,978,590	412	206,313	2,150	573,724	0.004	1.92E-01	3.60E-01	0.5
	Sum	8,130,528	2,150	573,724	-	-	-	1.00E+00	1.00E+00	1.0

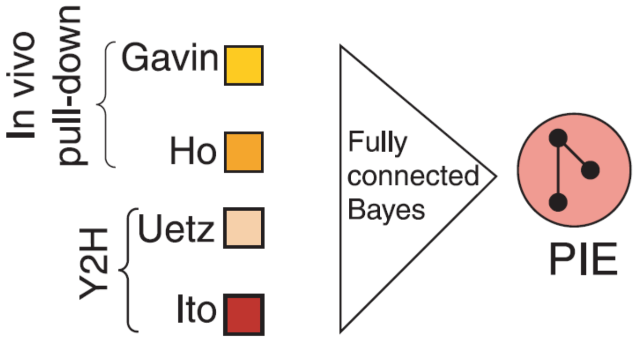
	Expression correlation	# protein pairs	Gold standard overlap					*P(exp|pos)*	*P(exp|neg)*	*L*
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	0.9	678	16	45	16	45	0.36	2.10E-03	1.68E-05	124.9
	0.8	4,827	137	563	153	608	0.25	1.80E-02	2.10E-04	85.5
	0.7	17,626	530	2,117	683	2,725	0.25	6.96E-02	7.91E-04	88.0
	0.6	42,815	1,073	5,597	1,756	8,322	0.21	1.41E-01	2.09E-03	67.4
	0.5	96,650	1,089	14,459	2,845	22,781	0.12	1.43E-01	5.40E-03	26.5
	0.4	225,712	993	35,350	3,838	58,131	0.07	1.30E-01	1.32E-02	9.9
	0.3	529,268	1,028	83,483	4,866	141,614	0.03	1.35E-01	3.12E-02	4.3
	0.2	1,200,331	870	183,356	5,736	324,970	0.02	1.14E-01	6.85E-02	1.7
	0.1	2,575,103	739	368,469	6,475	693,439	0.01	9.71E-02	1.38E-01	0.7
	0	9,363,627	894	1,244,477	7,369	1,937,916	0.00	1.17E-01	4.65E-01	0.3
	-0.1	2,753,735	164	408,562	7,533	2,346,478	0.00	2.15E-02	1.53E-01	0.1
	-0.2	1,241,907	63	203,663	7,596	2,550,141	0.00	8.27E-03	7.61E-02	0.1
	-0.3	484,524	13	84,957	7,609	2,635,098	0.00	1.71E-03	3.18E-02	0.1
	-0.4	160,234	3	28,870	7,612	2,663,968	0.00	3.94E-04	1.08E-02	0.0
	-0.5	48,852	2	8,091	7,614	2,672,059	0.00	2.63E-04	3.02E-03	0.1
	-0.6	17,423	-	2,134	7,614	2,674,193	0.00	0.00E+00	7.98E-04	0.0
	-0.7	7,602	-	807	7,614	2,675,000	0.00	0.00E+00	3.02E-04	0.0
	-0.8	2,147	-	261	7,614	2,675,261	0.00	0.00E+00	9.76E-05	0.0
	-0.9	67	-	12	7,614	2,675,273	0.00	0.00E+00	4.49E-06	0.0
	Sum	18,773,128	7,614	2,675,273	-	-	-	1.00E+00	1.00E+00	1.0

	MIPS function similarity	# protein pairs	Gold standard overlap					*P(MIPS|pos)*	*P(MIPS|neg)*	*L*
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	1 -- 9	6,584	171	1,094	171	1,094	0.16	2.12E-02	8.33E-04	25.5
	10 – 99	25,823	584	4,229	755	5,323	0.14	7.25E-02	3.22E-03	22.5
	100 -- 1000	88,548	688	13,011	1,443	18,334	0.08	8.55E-02	9.91E-03	8.6
	1000 – 10000	255,096	6,146	47,126	7,589	65,460	0.12	7.63E-01	3.59E-02	21.3
	10000 -- Inf	5,785,754	462	1,248,119	8,051	1,313,579	0.01	5.74E-02	9.50E-01	0.1
	Sum	6,161,805	8,051	1,313,579	-	-	-	1.00E+00	1.00E+00	1.0

	GO biological process similarity	# protein pairs	Gold standard overlap					*P(GO|pos)*	*P(GO|neg)*	*L*
			pos	*neg*	sum(*pos*)	sum(*neg*)	sum(*pos*)/ sum(*neg*)			
Values	1 -- 9	4,789	88	819	88	819	0.11	1.17E-02	1.27E-03	9.2
	10 – 99	20,467	555	3,315	643	4,134	0.16	7.38E-02	5.14E-03	14.4
	100 -- 1000	58,738	523	10,232	1,166	14,366	0.08	6.95E-02	1.59E-02	4.4
	1000 – 10000	152,850	1,003	28,225	2,169	42,591	0.05	1.33E-01	4.38E-02	3.0
	10000 -- Inf	2,909,442	5,351	602,434	7,520	645,025	0.01	7.12E-01	9.34E-01	0.8
	Sum	3,146,286	7,520	645,025	-	-	-	1.00E+00	1.00E+00	1.0

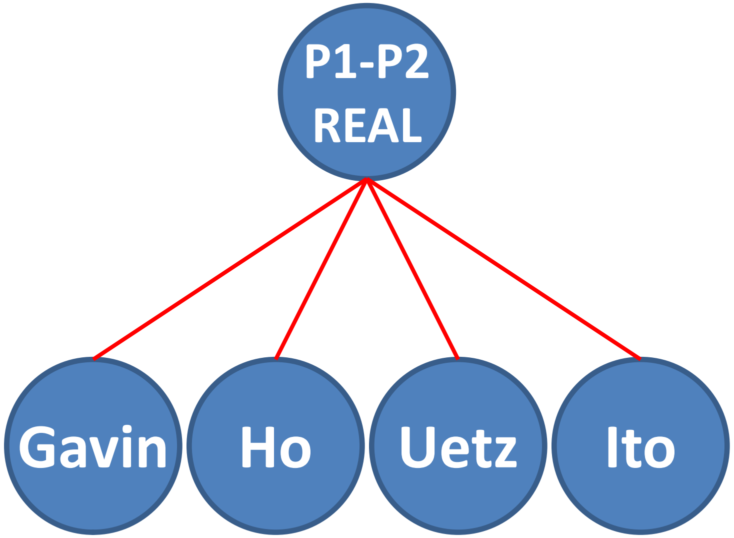


Source: Jansen, Ronald, Haiyuan Yu, et al. "A Bayesian Networks Approach for Predicting Protein-protein Interactions from Genomic Data." *Science* 302, no. 5644 (2003): 449-53.

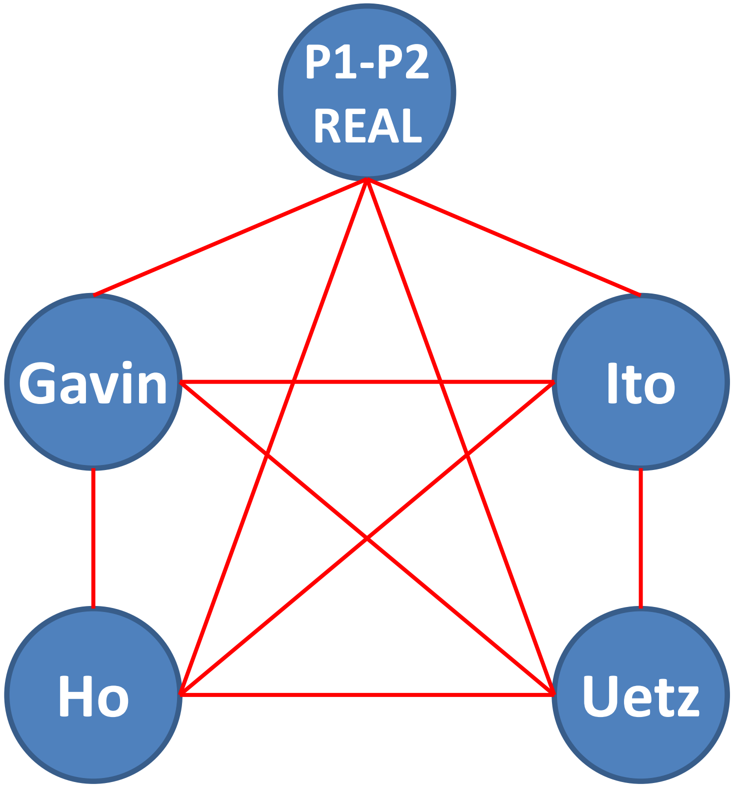


Source: Jansen, Ronald, Haiyuan Yu, et al. "A Bayesian Networks Approach for Predicting Protein-protein Interactions from Genomic Data." *Science* 302, no. 5644 (2003): 449-53.

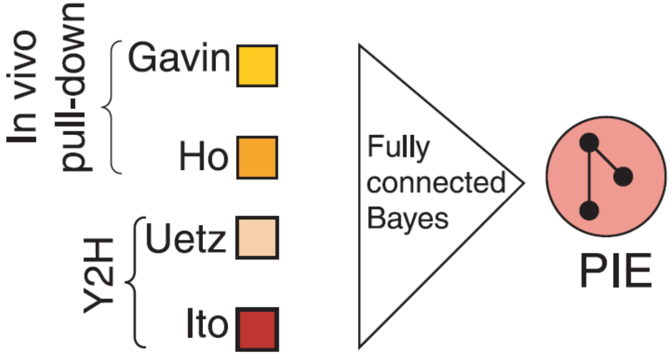
What do we mean by fully connected?



$$P(X_1 \ldots X_n|\text{PPI}) = \prod_i [P(Xi|\text{PPI})]$$

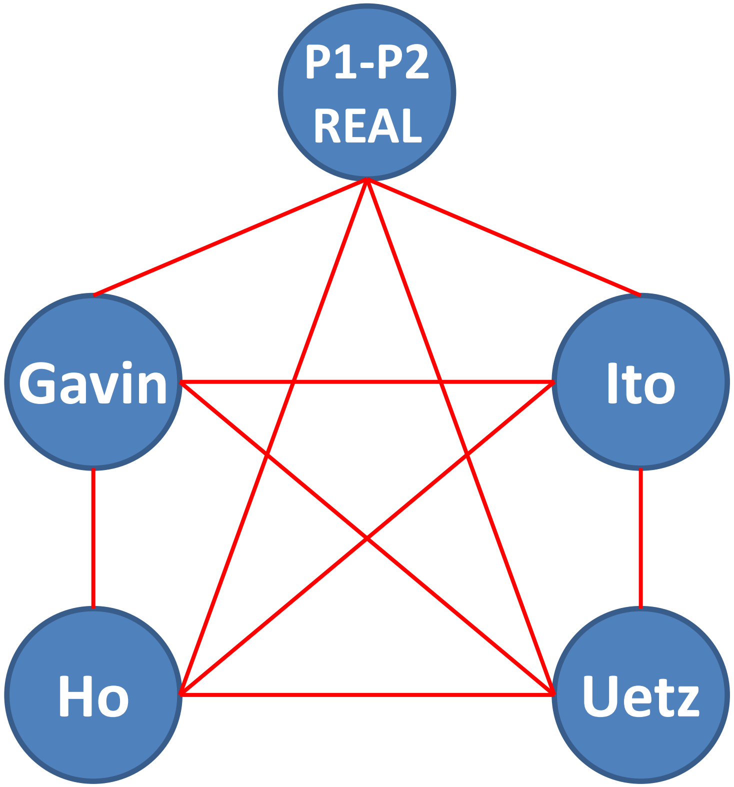


$$P(X_1 \ldots X_n|\text{PPI}) \neq \prod_i [P(Xi|\text{PPI})]$$

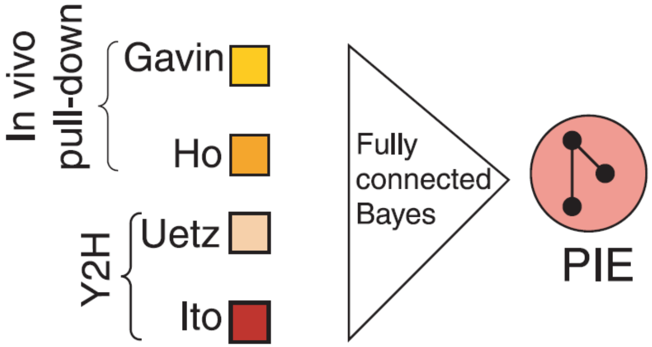


Source: Jansen, Ronald, Haiyuan Yu, et al. "A Bayesian Networks Approach for Predicting Protein-protein Interactions from Genomic Data." *Science* 302, no. 5644 (2003): 449-53.

Fully connected →
Compute probabilities for all 16 possible combinations

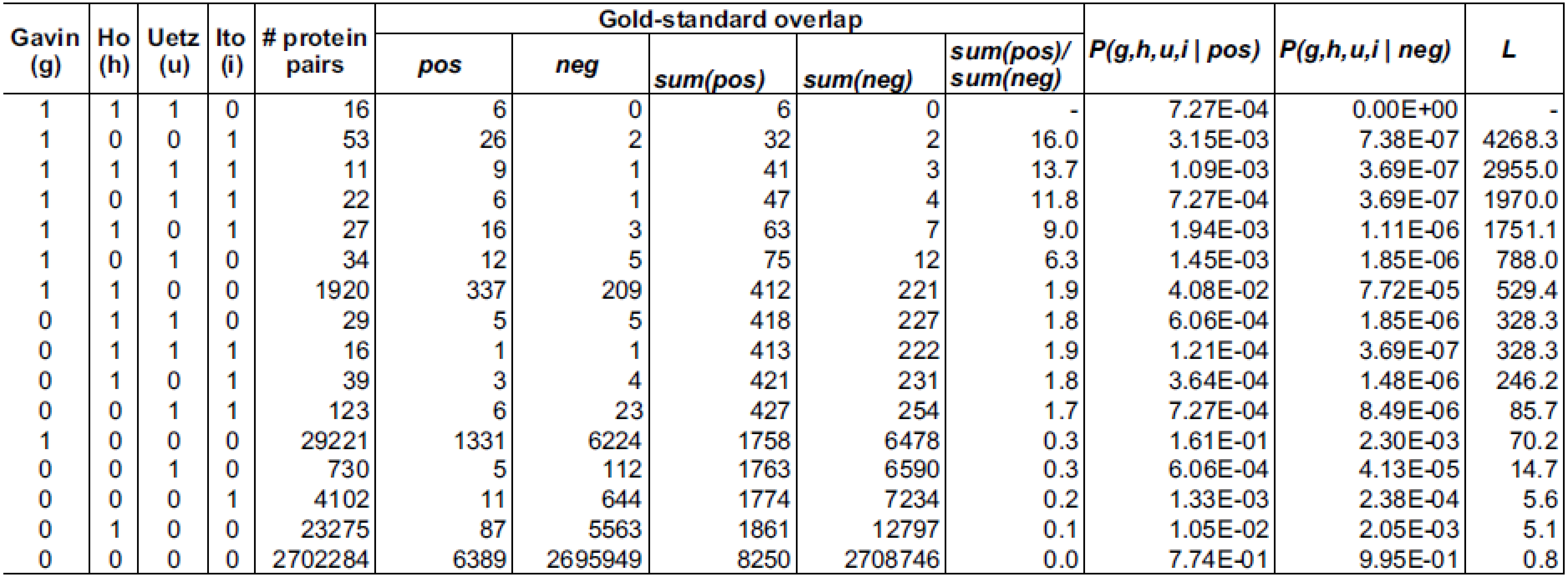


$$P(X_1 \ldots X_n|\text{PPI}) \neq \prod_i [P(Xi|\text{PPI})]$$

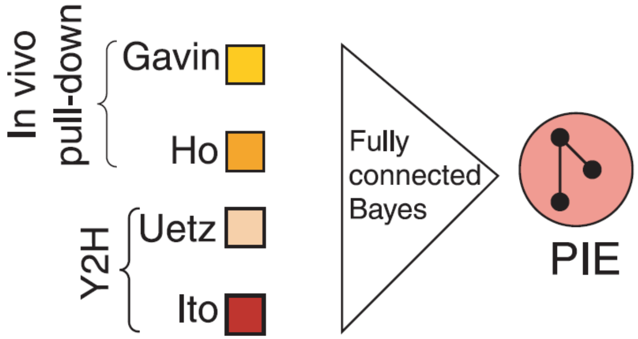


Fully connected →
Compute probabilities for all 16 possible combinations

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.



Gavin (g)	Ho (h)	Uetz (u)	Ito (i)	# protein pairs	Gold-standard overlap					P(g,h,u,i | pos)	P(g,h,u,i | neg)	L
					pos	neg	sum(pos)	sum(neg)	sum(pos)/ sum(neg)			
1	1	1	0	16	6	0	6	0	-	7.27E-04	0.00E+00	-
1	0	0	1	53	26	2	32	2	16.0	3.15E-03	7.38E-07	4268.3
1	1	1	1	11	9	1	41	3	13.7	1.09E-03	3.69E-07	2955.0
1	0	1	1	22	6	1	47	4	11.8	7.27E-04	3.69E-07	1970.0
1	1	0	1	27	16	3	63	7	9.0	1.94E-03	1.11E-06	1751.1
1	0	1	0	34	12	5	75	12	6.3	1.45E-03	1.85E-06	788.0
1	1	0	0	1920	337	209	412	221	1.9	4.08E-02	7.72E-05	529.4
0	1	1	0	29	5	5	418	227	1.8	6.06E-04	1.85E-06	328.3
0	1	1	1	16	1	1	413	222	1.9	1.21E-04	3.69E-07	328.3
0	1	0	1	39	3	4	421	231	1.8	3.64E-04	1.48E-06	246.2
0	0	1	1	123	6	23	427	254	1.7	7.27E-04	8.49E-06	85.7
1	0	0	0	29221	1331	6224	1758	6478	0.3	1.61E-01	2.30E-03	70.2
0	0	1	0	730	5	112	1763	6590	0.3	6.06E-04	4.13E-05	14.7
0	0	0	1	4102	11	644	1774	7234	0.2	1.33E-03	2.38E-04	5.6
0	1	0	0	23275	87	5563	1861	12797	0.1	1.05E-02	2.05E-03	5.1
0	0	0	0	2702284	6389	2695949	8250	2708746	0.0	7.74E-01	9.95E-01	0.8



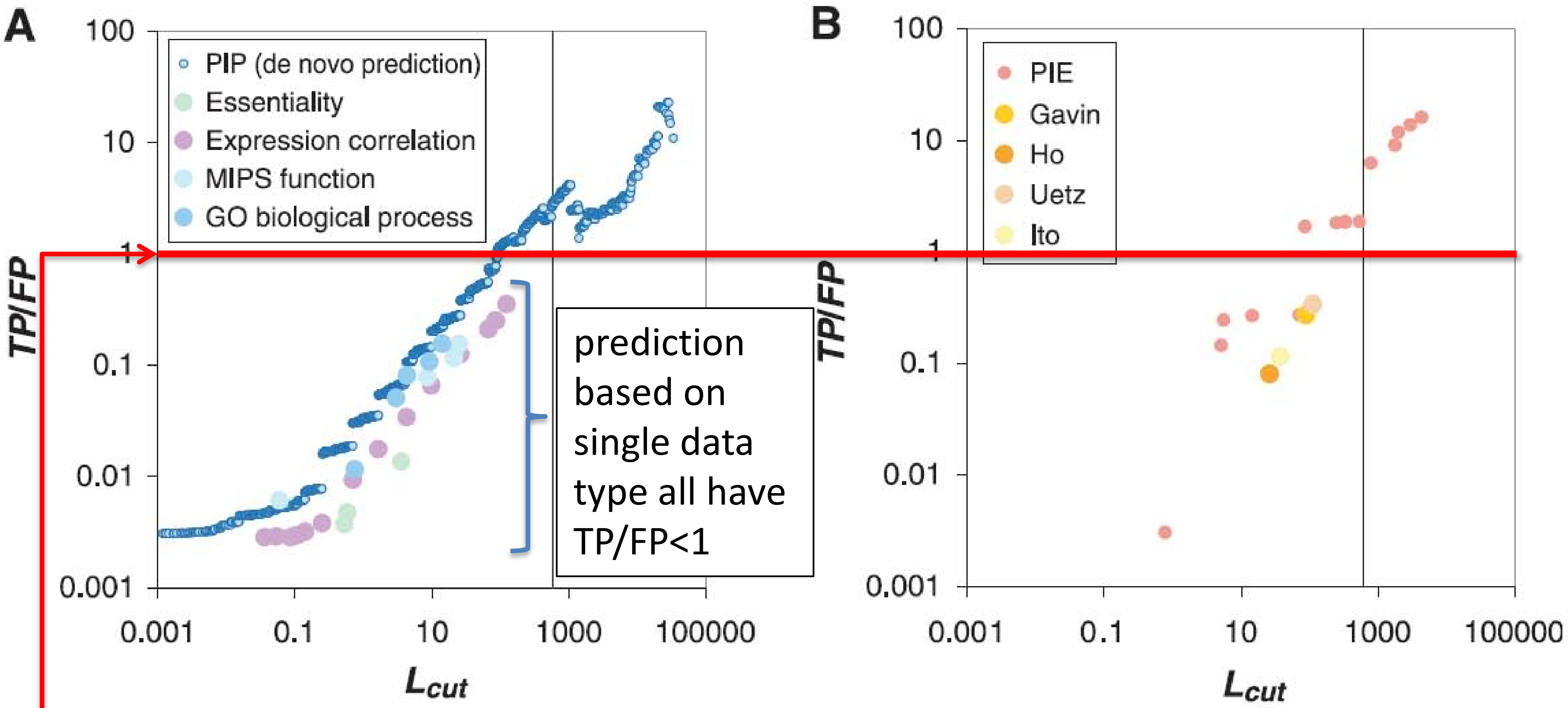
Interpret with caution, as numbers are small

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.

Gavin (g)	Ho (h)	Uetz (u)	Ito (i)	# protein pairs	Gold-standard overlap							
					pos	neg	sum(pos)	sum(neg)	sum(pos)/ sum(neg)	*P(g,h,u,i | pos)*	*P(g,h,u,i | neg)*	*L*
1	1	1	0	16	6	0	6	0	-	7.27E-04	0.00E+00	-
1	0	0	1	53	26	2	32	2	16.0	3.15E-03	7.38E-07	4268.3
1	1	1	1	11	9	1	41	3	13.7	1.09E-03	3.69E-07	2955.0
1	0	1	1	22	6	1	47	4	11.8	7.27E-04	3.69E-07	1970.0
1	1	0	1	27	16	3	63	7	9.0	1.94E-03	1.11E-06	1751.1
1	0	1	0	34	12	5	75	12	6.3	1.45E-03	1.85E-06	788.0
1	1	0	0	1920	337	209	412	221	1.9	4.08E-02	7.72E-05	529.4
0	1	1	0	29	5	5	418	227	1.8	6.06E-04	1.85E-06	328.3
0	1	1	1	16	1	1	413	222	1.9	1.21E-04	3.69E-07	328.3
0	1	0	1	39	3	4	421	231	1.8	3.64E-04	1.48E-06	246.2
0	0	1	1	123	6	23	427	254	1.7	7.27E-04	8.49E-06	85.7
1	0	0	0	29221	1331	6224	1758	6478	0.3	1.61E-01	2.30E-03	70.2
0	0	1	0	730	5	112	1763	6590	0.3	6.06E-04	4.13E-05	14.7
0	0	0	1	4102	11	644	1774	7234	0.2	1.33E-03	2.38E-04	5.6
0	1	0	0	23275	87	5563	1861	12797	0.1	1.05E-02	2.05E-03	5.1
0	0	0	0	2702284	6389	2695949	8250	2708746	0.0	7.74E-01	9.95E-01	0.8

How many gold-standard events do we score correctly at different likelihood cutoffs?

$$\log\left[\frac{P(Data \mid true\_PPI)}{P(Data \mid false\_PPI)}\right]$$



prediction based on single data type all have TP/FP<1

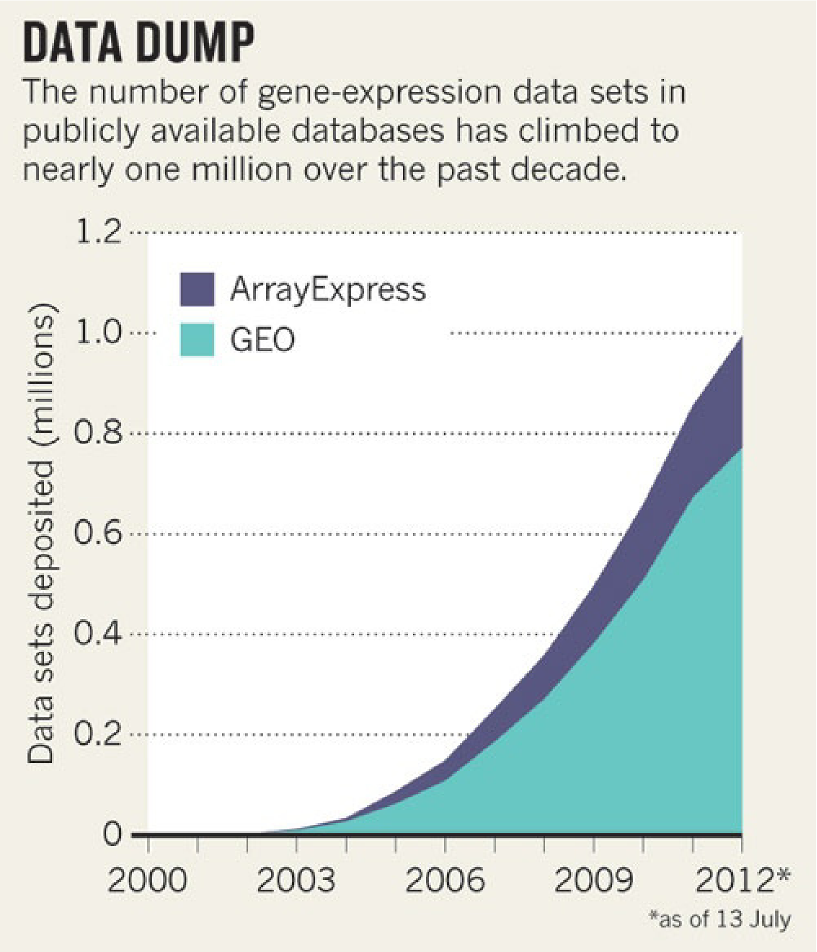
TF=FP

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.

Outline

- Bayesian Networks for PPI prediction
- Gene expression
  - Distance metrics
  - Clustering
  - Signatures
  - Modules
    - Bayesian networks
    - Regression
    - Mutual Information
    - Evaluation on real and simulated data

Gene Expression Data



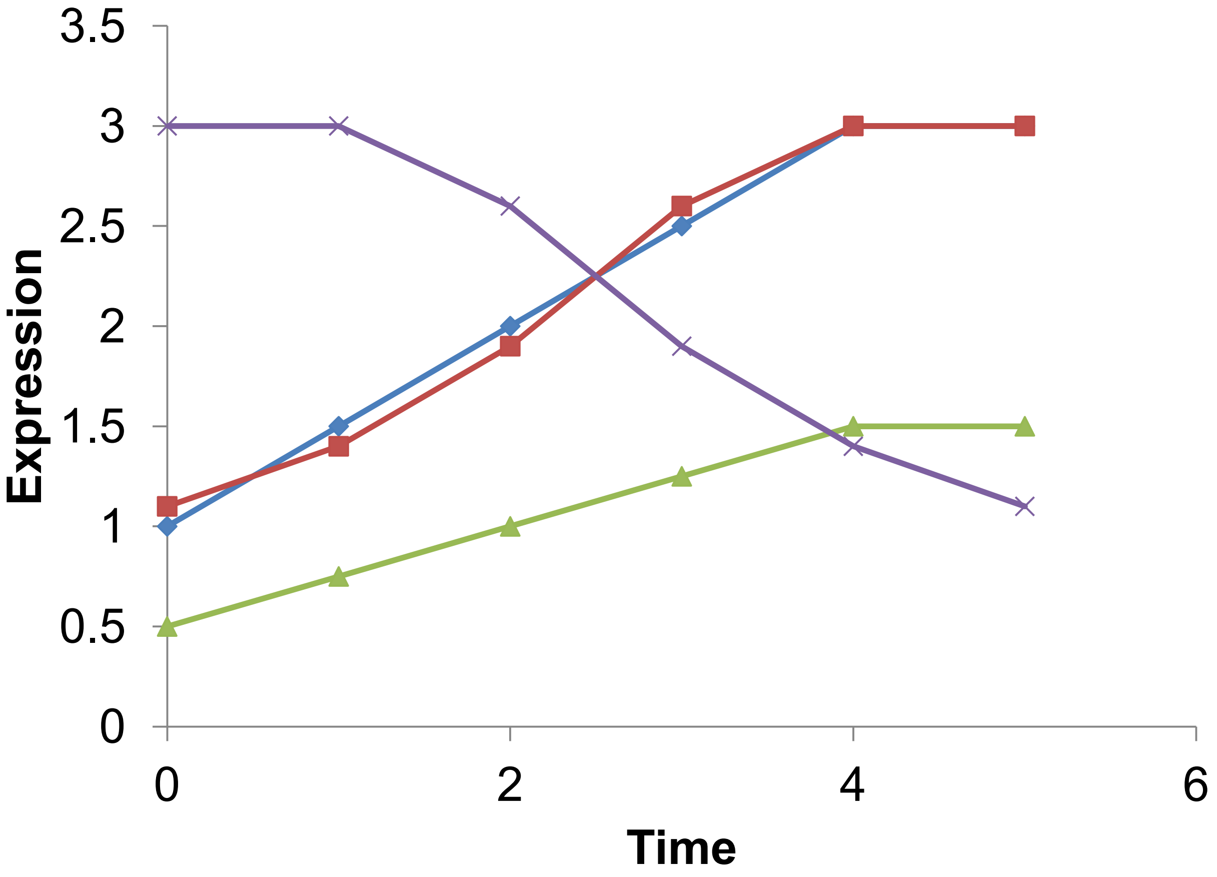
Courtesy of Macmillan Publishers Limited. Used with permission. Source: Baker, Monya. "Gene Data to Hit Milestone." *Nature* 487, no. 7407 (2012): 282-3.

- Identify co-expressed genes
- Classify new datasets
- Discover regulatory networks

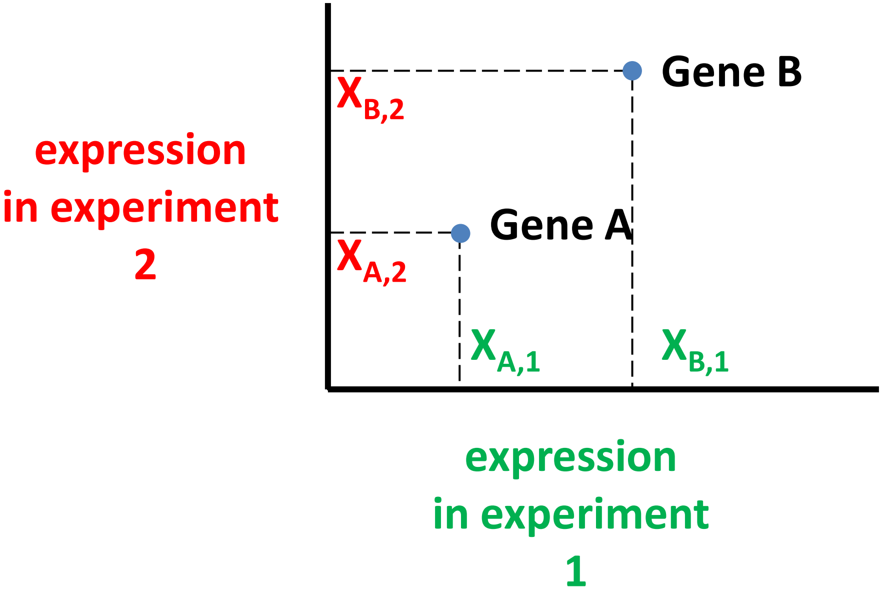
Clustering

- Text Section 16.2
- Multiple mechanisms could lead to up-regulation in any one condition
- Goal: Find genes that have “similar” expression over many condition.
- How do you define “similar”?

Distance Metrics



Expression data as multidimensional vectors



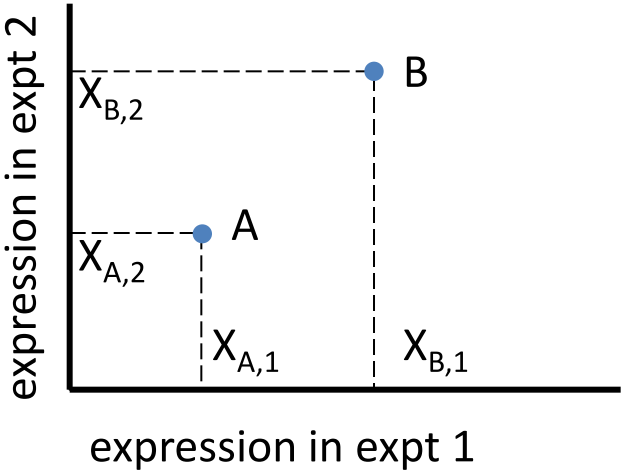
X_A =(1, 0.5,   -1, 0.25, …)
X_B =(0.2, 0.4, -1.2, 0.05, …)
…

What is a natural way to compare these vectors?

Euclidean

- $X_{i,j}$ = Expression of gene *i* in condition *j*

$$d(X_A, X_B) = \sqrt{\sum_{k=1}^{N} (X_{A,k} - X_{B,k})^2}$$

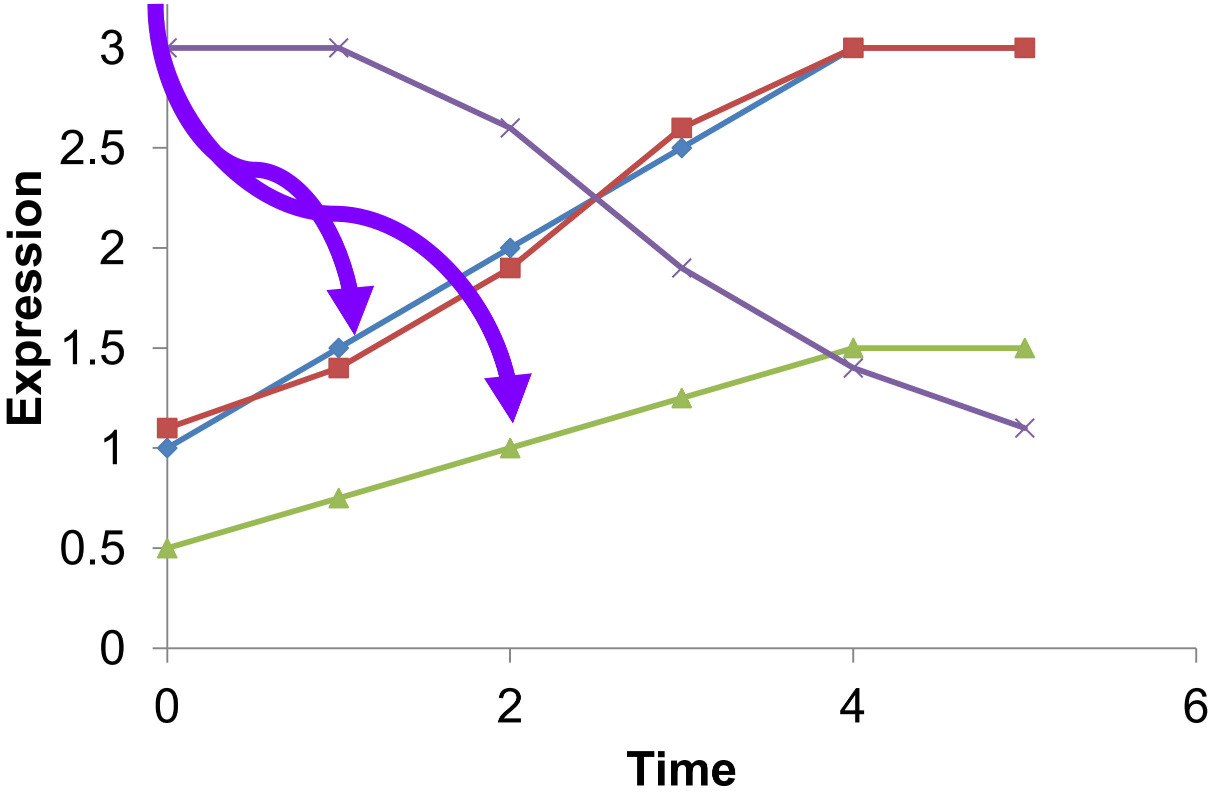


Distance

- Metrics have a formal definition:
  - $d(x, y) \geq 0$
  - $d(x, y) = 0$ if and only if $x = y$
  - $d(x, y) = d(y, x)$
  - Triangle inequality:
    $d(x, z) \leq d(x, y) + d(y, z)$
- The triangle inequality need not hold for a measure of "similarity."
- Distance ~ Dissimilarity = 1 - similarity

Distance Metrics

Can we capture the similarity of these patterns?



Pearson Correlation

- $X_{i,j}$ = Expression of gene *i* in condition *j*
- Z_i = z-score of gene *i* one experiment:

$$Z_A = \frac{X_A - \overline{X}_A}{\sigma} \qquad \sigma^2 = \frac{\sum (X - \overline{X})}{N}$$

Pearson Correlation

- $X_{i,j}$ = Expression of gene *i* in condition *j*
- Z_i = z-score of gene *i* one experiment:

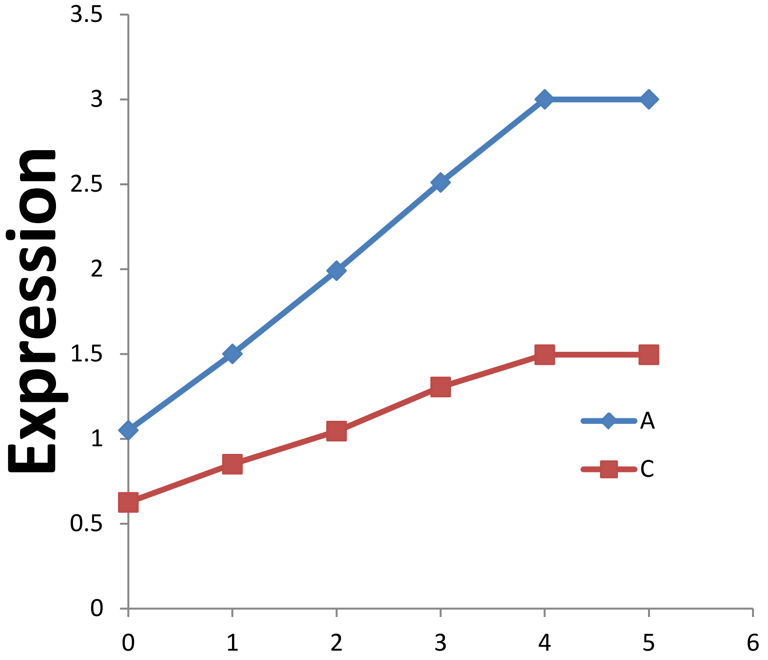
- Pearson correlation

$$r_{A,B} = \frac{\sum Z_A Z_B}{N}$$

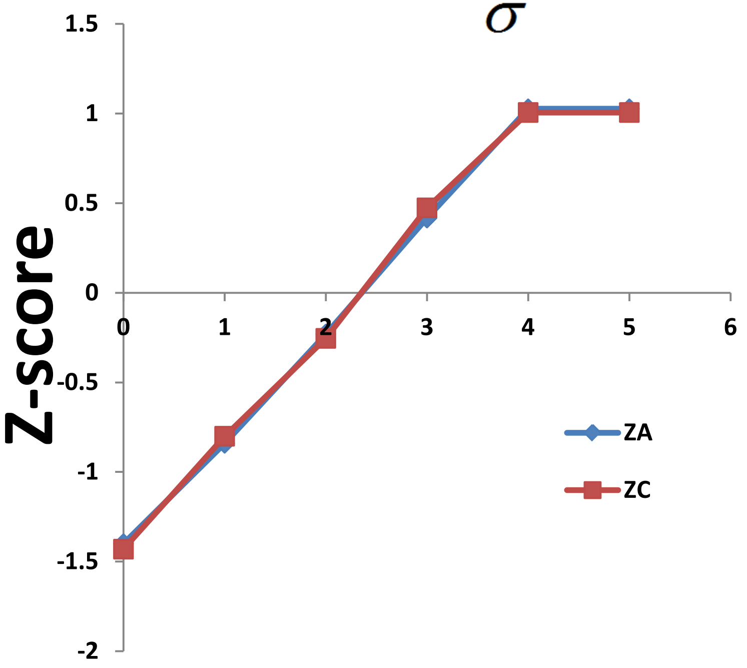
over all experiments

  - from +1 (perfect correlation) to -1 (anti-correlated)
- Distance = 1-$r_{A,B}$

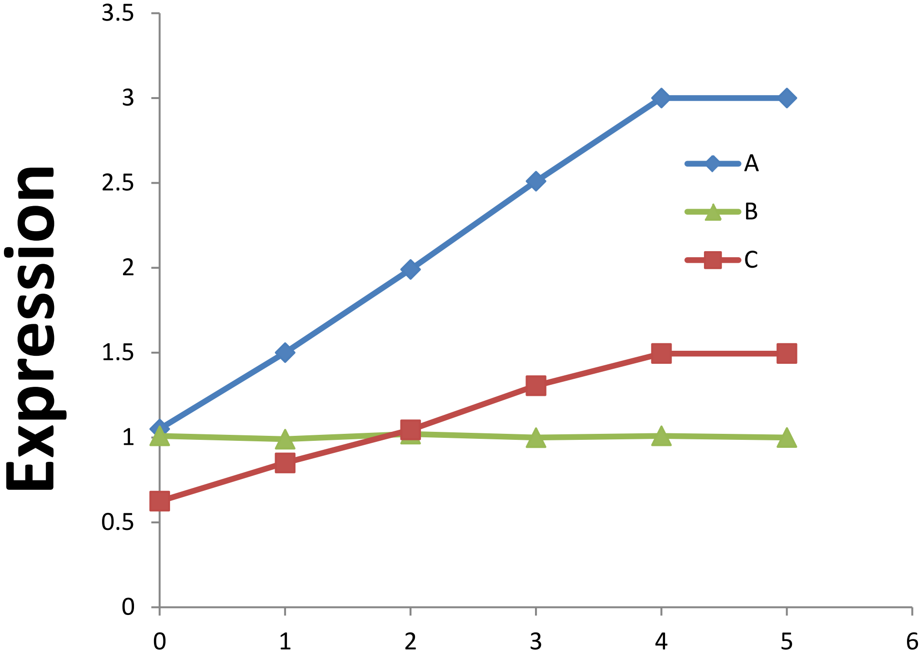
$$Z_A = \frac{X_A - \overline{X}_A}{\sigma} \qquad \sigma^2 = \frac{\sum (X - \overline{X})}{N}$$

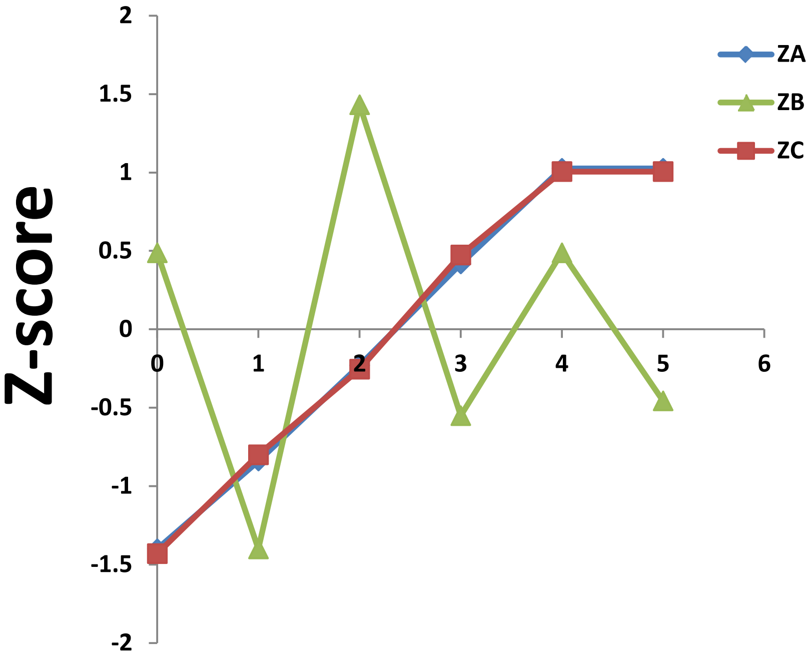


$$Z_A = \frac{X_A - \overline{X}_A}{\sigma}$$



$$r_{A,B} = \frac{\sum Z_A Z_B}{N}$$





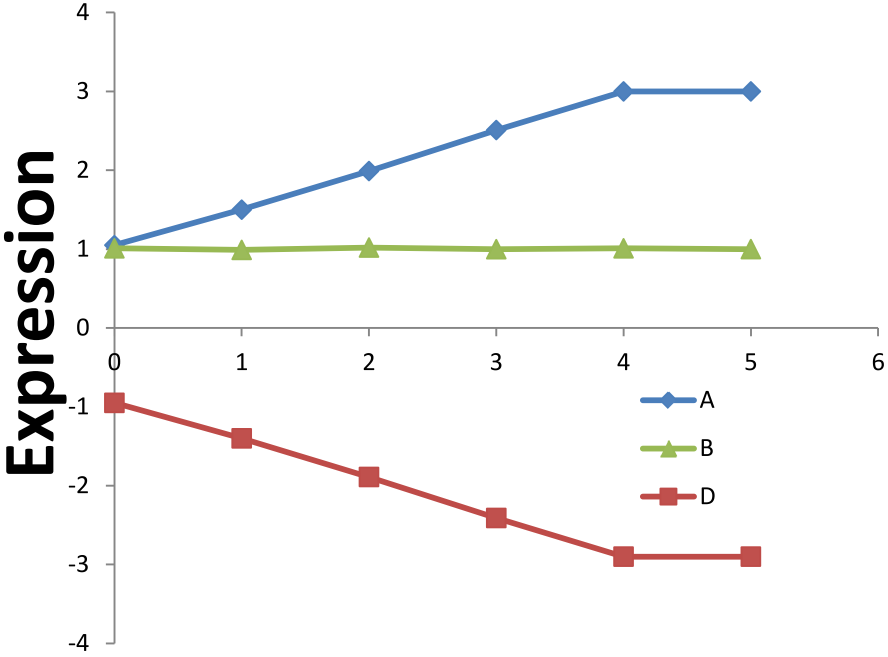
$$Z_A = \frac{X_A - \bar{X}_A}{\sigma}$$

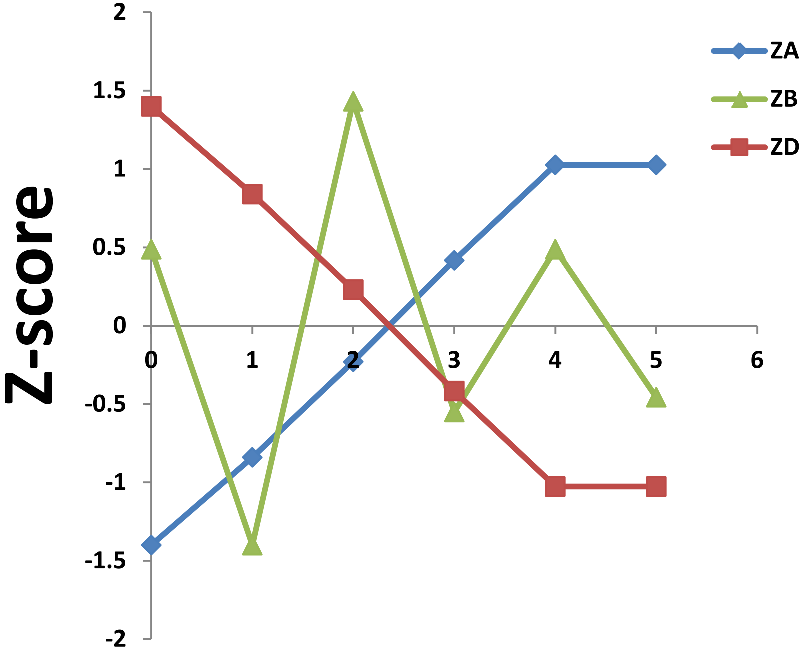
$R_{A,B}$= -0.01

$R_{A,C}$= 0.999

$R_{B,C}$= -0.03

$$r_{A,B} = \frac{\sum Z_A Z_B}{N}$$





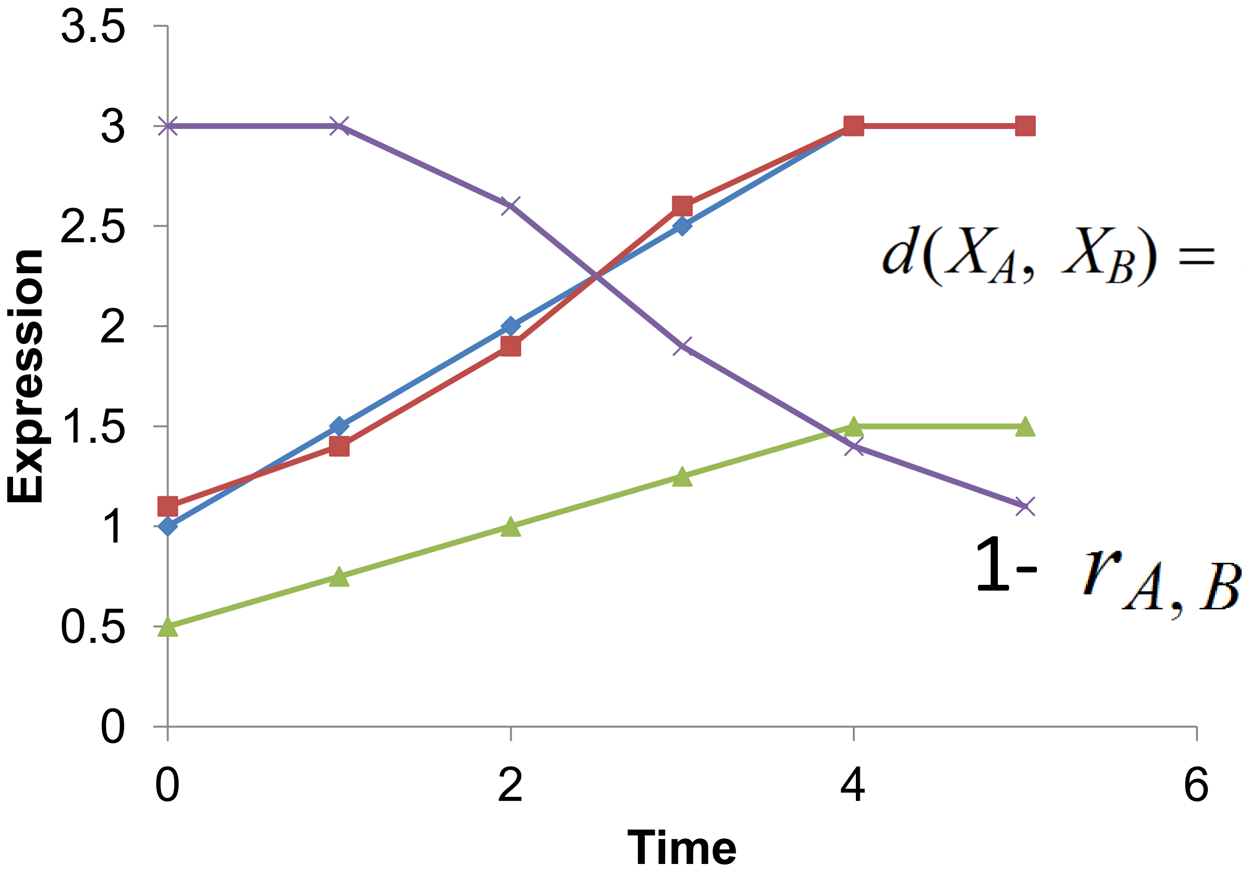
$R_{A,B}$= -0.01

$R_{A,D}$= -1.0

$R_{B,D}$= 0.007

$$r_{A,B} = \frac{\sum Z_A Z_B}{N}$$

Distance Metrics



$$d(X_A,\ X_B) = \sqrt{\sum_{k=1}^{N} (X_{A,k} - X_{B,k})^2}$$

1- $r_{A,B} = \dfrac{\sum Z_A Z_B}{N}$

Missing Data

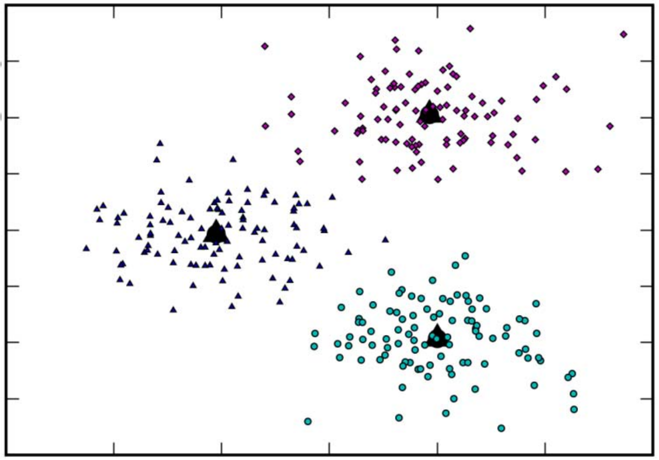
- What if a particular data point is missing? (Back in the old days: there was a bubble or a hair on the array)
  - ignore that gene in all samples
  - ignore that sample for all genes
  - replace missing value with a constant
  - “impute” a value
    - example: compute the K most similar genes (arrays) using the available data; set the missing value to the mean of that for these K genes (arrays)

Outline

- Bayesian Networks for PPI prediction
- Gene expression
  - Distance metrics
  - Clustering
  - Signatures
  - Modules
    - Bayesian networks
    - Regression
    - Mutual Information
    - Evaluation on real and simulated data

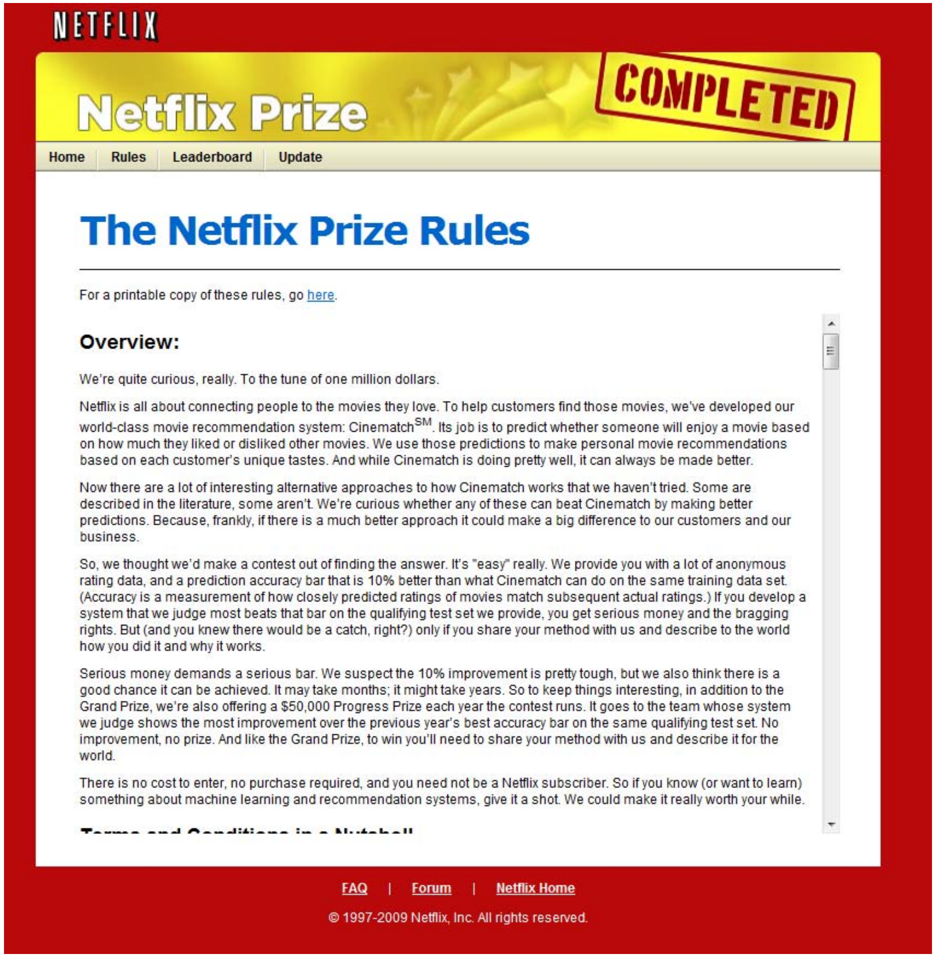
Clustering

- Intuitive idea that we want to find an underlying grouping
- In practice, this can be hard to define and implement.



- An example of unsupervised learning

Unsupervised Learning



The Netflix Prize Rules

For a printable copy of these rules, go here.

Overview:

We're quite curious, really. To the tune of one million dollars.

Netflix is all about connecting people to the movies they love. To help customers find those movies, we've developed our world-class movie recommendation system: Cinematch^{SM}. Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. We use those predictions to make personal movie recommendations based on each customer's unique tastes. And while Cinematch is doing pretty well, it can always be made better.

Now there are a lot of interesting alternative approaches to how Cinematch works that we haven't tried. Some are described in the literature, some aren't. We're curious whether any of these can beat Cinematch by making better predictions. Because, frankly, if there is a much better approach it could make a big difference to our customers and our business.

So, we thought we'd make a contest out of finding the answer. It's "easy" really. We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set. (Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.) If you develop a system that we judge most beats that bar on the qualifying test set we provide, you get serious money and the bragging rights. But (and you knew there would be a catch, right?) only if you share your method with us and describe to the world how you did it and why it works.

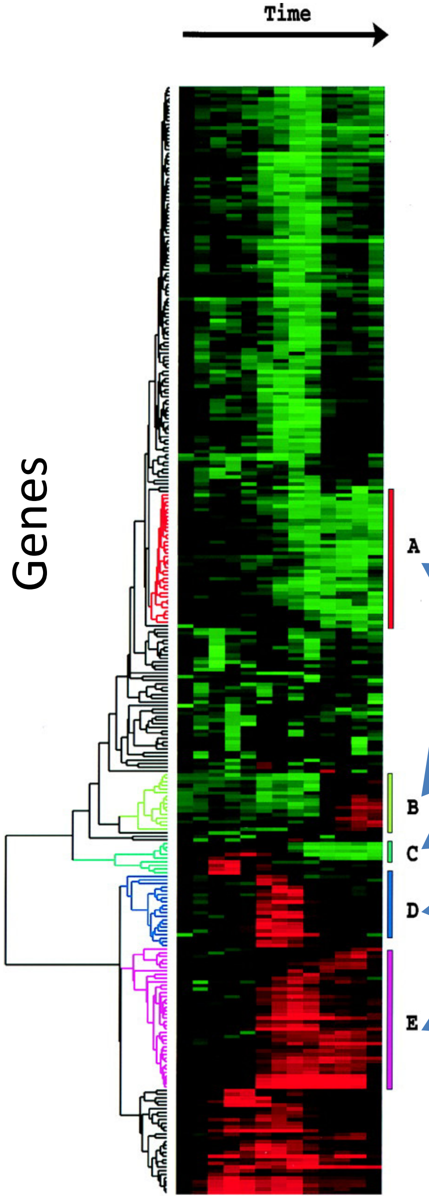
Serious money demands a serious bar. We suspect the 10% improvement is pretty tough, but we also think there is a good chance it can be achieved. It may take months; it might take years. So to keep things interesting, in addition to the Grand Prize, we're also offering a $50,000 Progress Prize each year the contest runs. It goes to the team whose system we judge shows the most improvement over the previous year's best accuracy bar on the same qualifying test set. No improvement, no prize. And like the Grand Prize, to win you'll need to share your method with us and describe it for the world.

There is no cost to enter, no purchase required, and you need not be a Netflix subscriber. So if you know (or want to learn) something about machine learning and recommendation systems, give it a shot. We could make it really worth your while.

Terms and Conditions in a Nutshell

© 1997-2009 Netflix, Inc. All rights reserved.

Clustering 8600 human genes based on time course of expression following serum stimulation of fibroblasts



Key: Black = little change Green = down Red = up

(relative to initial time point)

(A) cholesterol biosynthesis

(B) the cell cycle

(C) the immediate-early response

(D) signaling and angiogenesis

(E) wound healing and tissue remodeling

Source: Iyer, Vishwanath R., Michael B. Eisen, et al. "The Transcriptional Program in the Response of Human Fibroblasts to Serum." *Science* 283, no. 5398 (1999): 83-7.

Iyer et al. *Science* 1999

Why cluster?

- Cluster genes (rows)
  - Measure expression at multiple time-points, different conditions, etc.

Similar expression patterns may suggest similar functions of genes

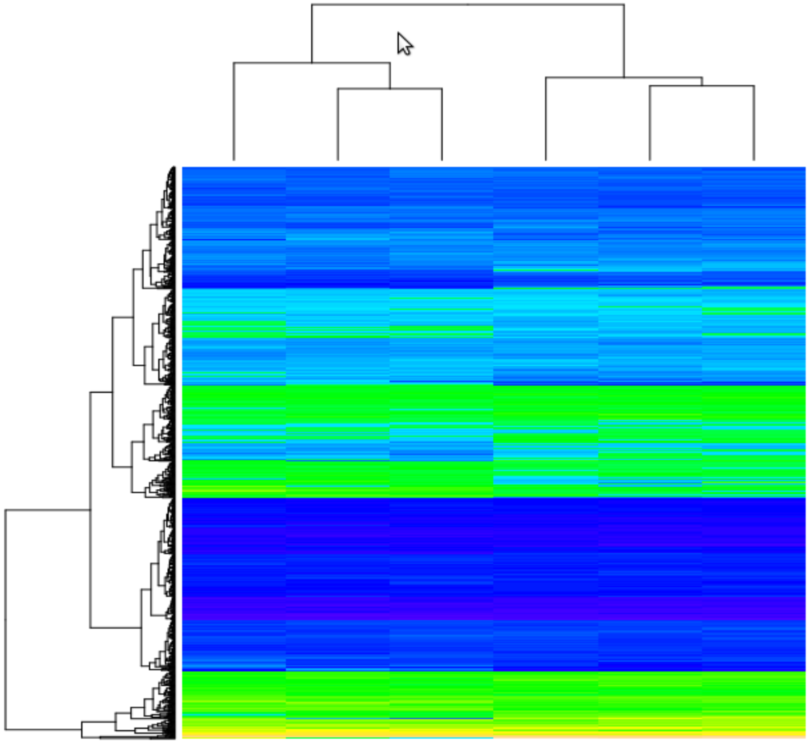
- Cluster samples (columns)
  - e.g., expression levels of thousands of genes for each tumor sample

Similar expression patterns may suggest biological relationship among samples

Hierarchcial clustering

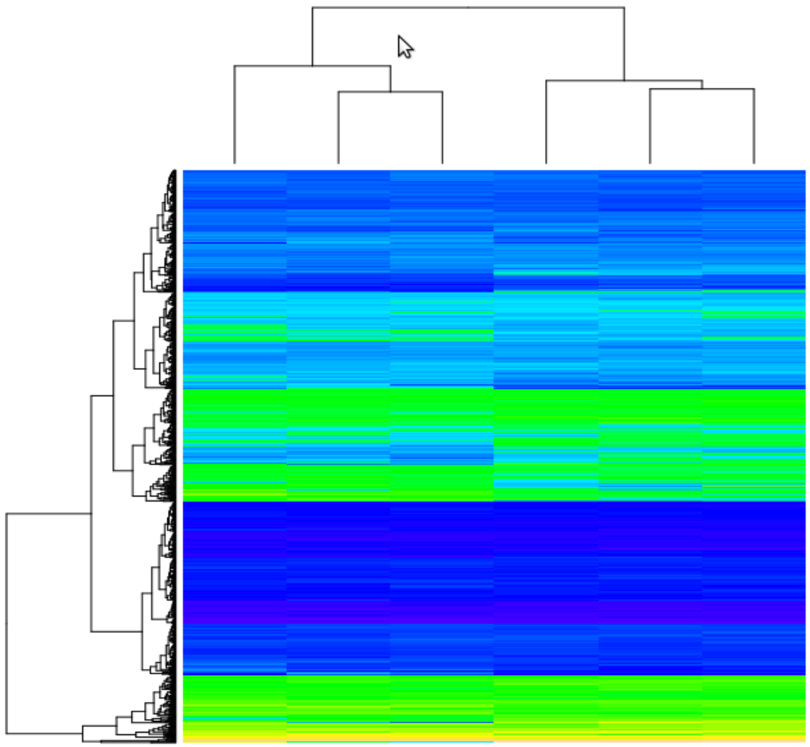
Two types of approaches:

- Agglomerative
- Divisive



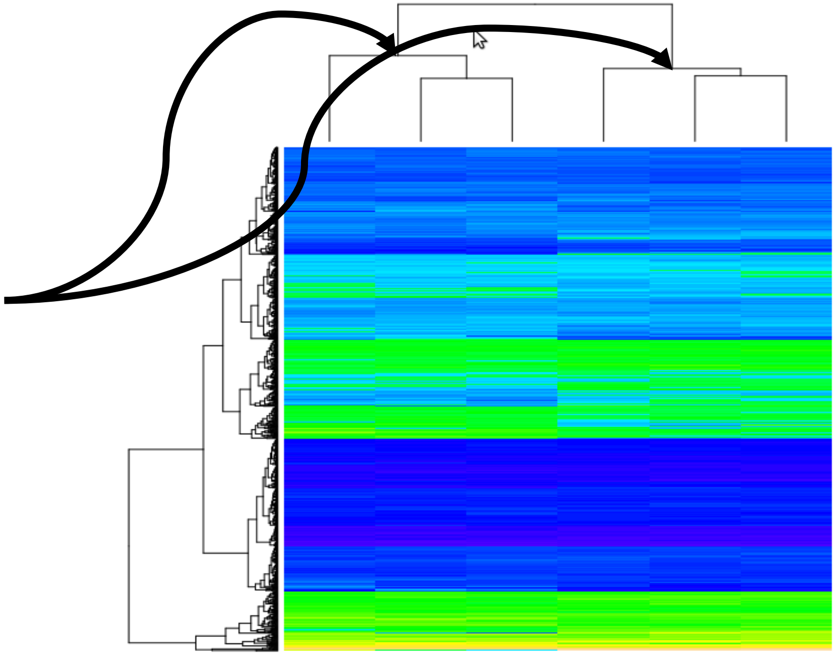
Agglomerative Clustering Algorithm

- Initialize: Each data point is in its own cluster
- Repeat until there is only one cluster:
  - Merge the two most similar clusters.



Agglomerative Clustering Algorithm

- Initialize: Each data point is in its own cluster
- Repeat until there is only one cluster:
  - Merge the two most similar clusters.

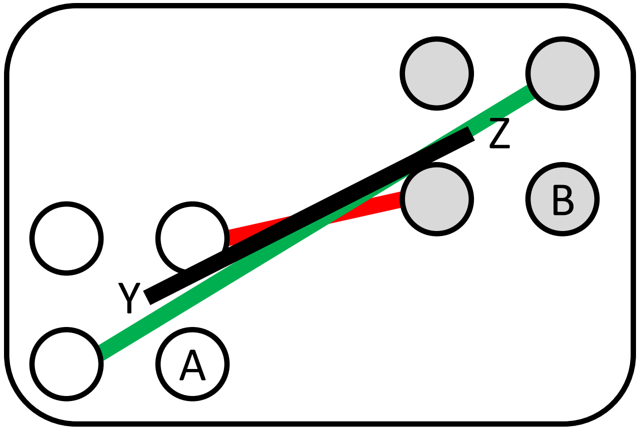


If distance is defined for a vector, how do I compare clusters?

- Clusters Y, Z with A in Y and B in Z
- Single linkage = min{$d_{A,B}$}
- Complete linkage = max{$d_{A,B}$}
- UPGMC (Unweighted Pair Group Method using **Centroids**

$$\text{centroid} = \hat{Y} = \frac{1}{N_Y} \sum_{i \in Y} X_{i,j}$$

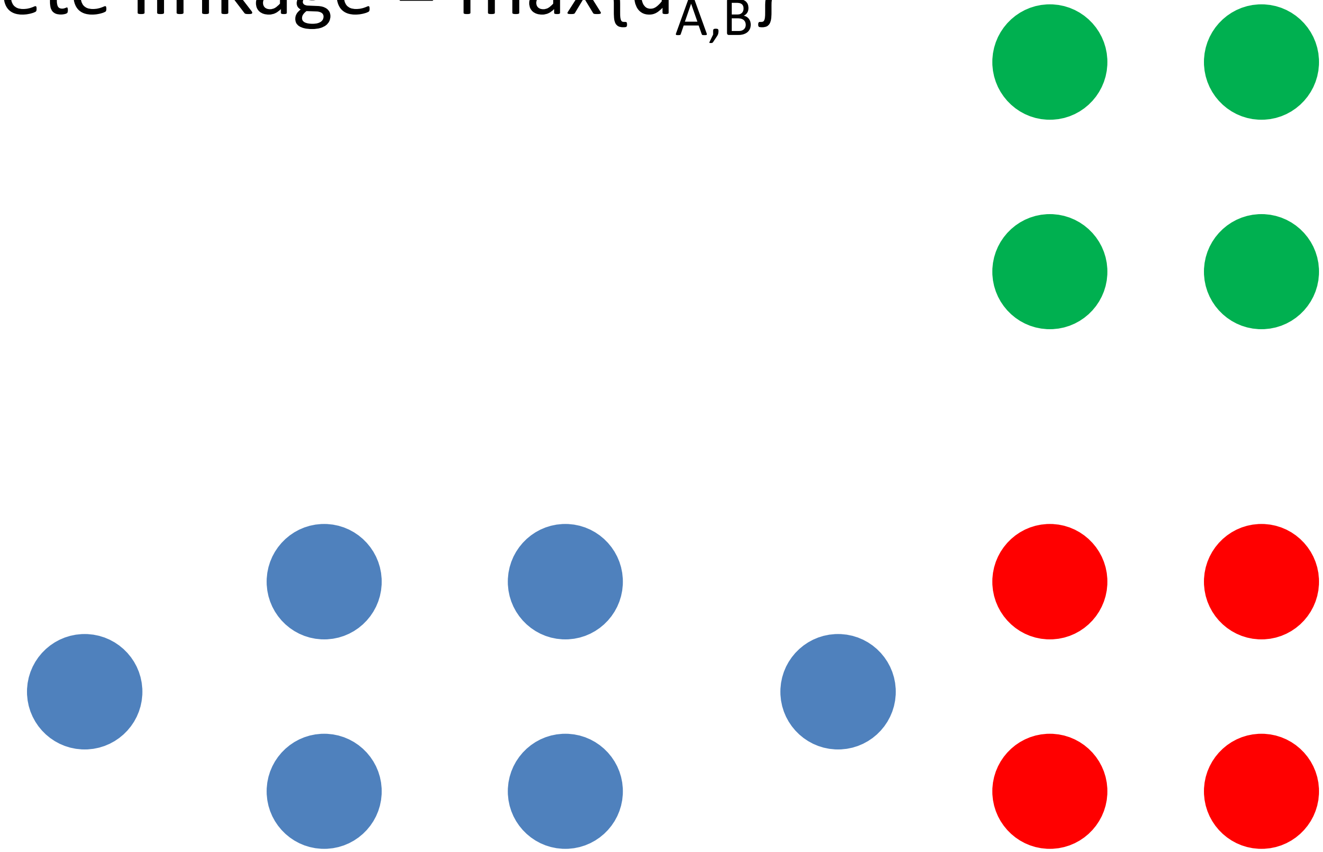
  - Define distance as $\delta_{Y,Z} = d_{\hat{Y},\hat{Z}}$



- UPGMA (Unweighted Pair Group Method with Arithmetic **Mean**) average of pairwise distances:

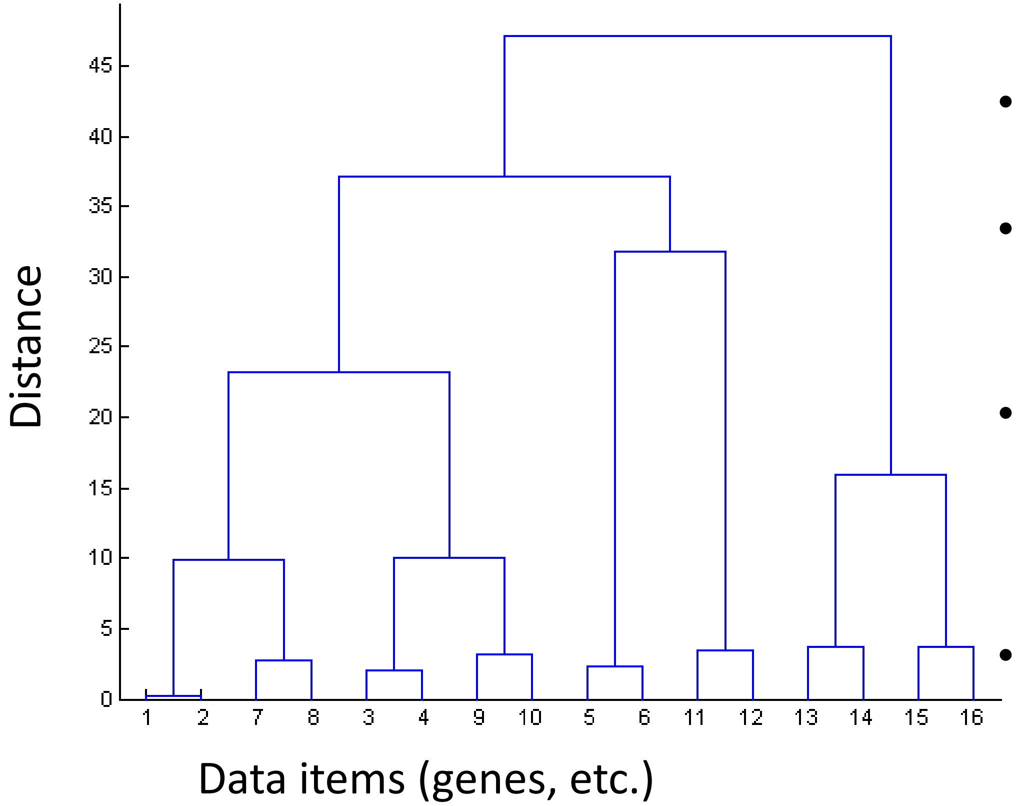
$$\delta_{Y,Z} = \frac{1}{N_Y N_Z} \sum_{i \in Y} \sum_{j \in Z} d_{i,j}$$

- Single linkage = min{$d_{A,B}$}
- Complete linkage = max{$d_{A,B}$}



- If clusters exist and are compact, it should not matter.
- Single linkage will “chain” together groups with one intermediate point.
- Complete linkage will not combine two groups if even one point is distant.

Interpreting the Dendogram



- This produces a binary tree or ***dendrogram***
- The final cluster is the root and each data item is a leaf
- The heights of the bars indicate how close the items are
- Can ‘slice’ the tree at any distance cutoff to produce discrete clusters
- Dendogram represents the results of the **<u>clustering</u>**; its usefulness in representing the **<u>data</u>** is mixed.
- The results will always be hierarchical, even if the data are not.

K-means clustering

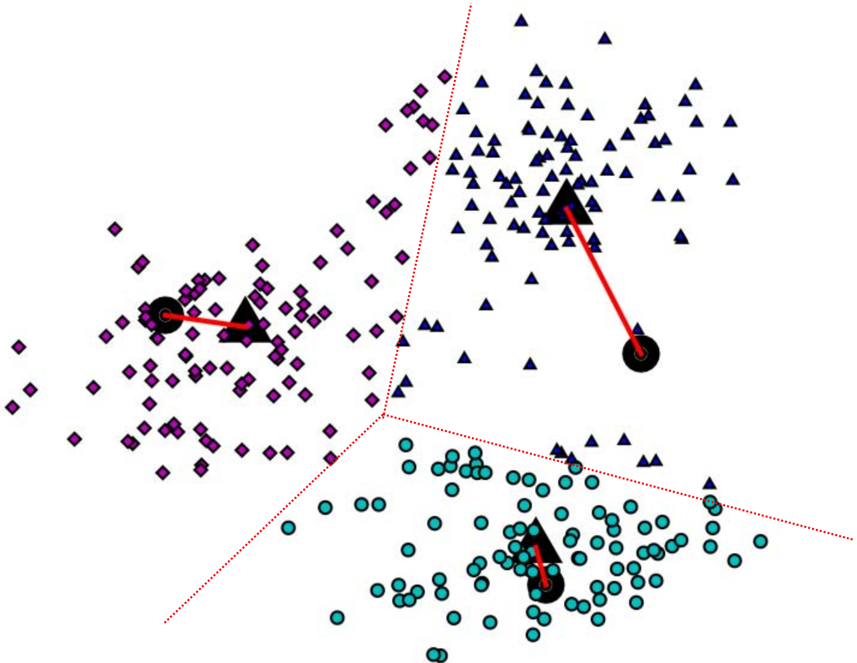
- Advantage: gives sharp partitions of the data
- Disadvantage: need to specify the number of clusters (K).
- Goal: find a set of k clusters that minimizes the distances of each point in the cluster to the cluster mean:

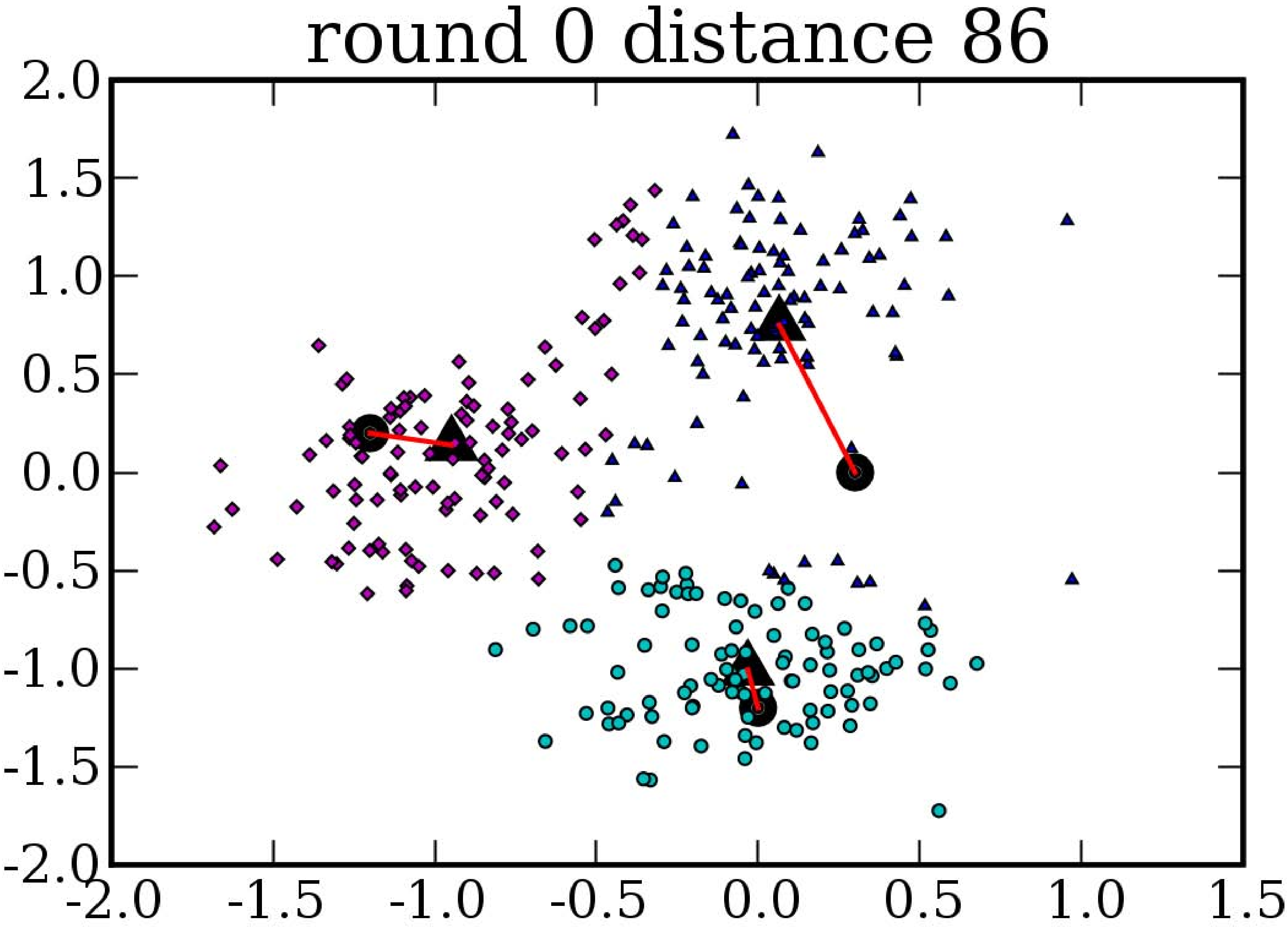
$$\text{centroid}_j = \hat{Y}_j = \frac{1}{N_{Y_j}} \sum_{i \in Y_j} X_i$$

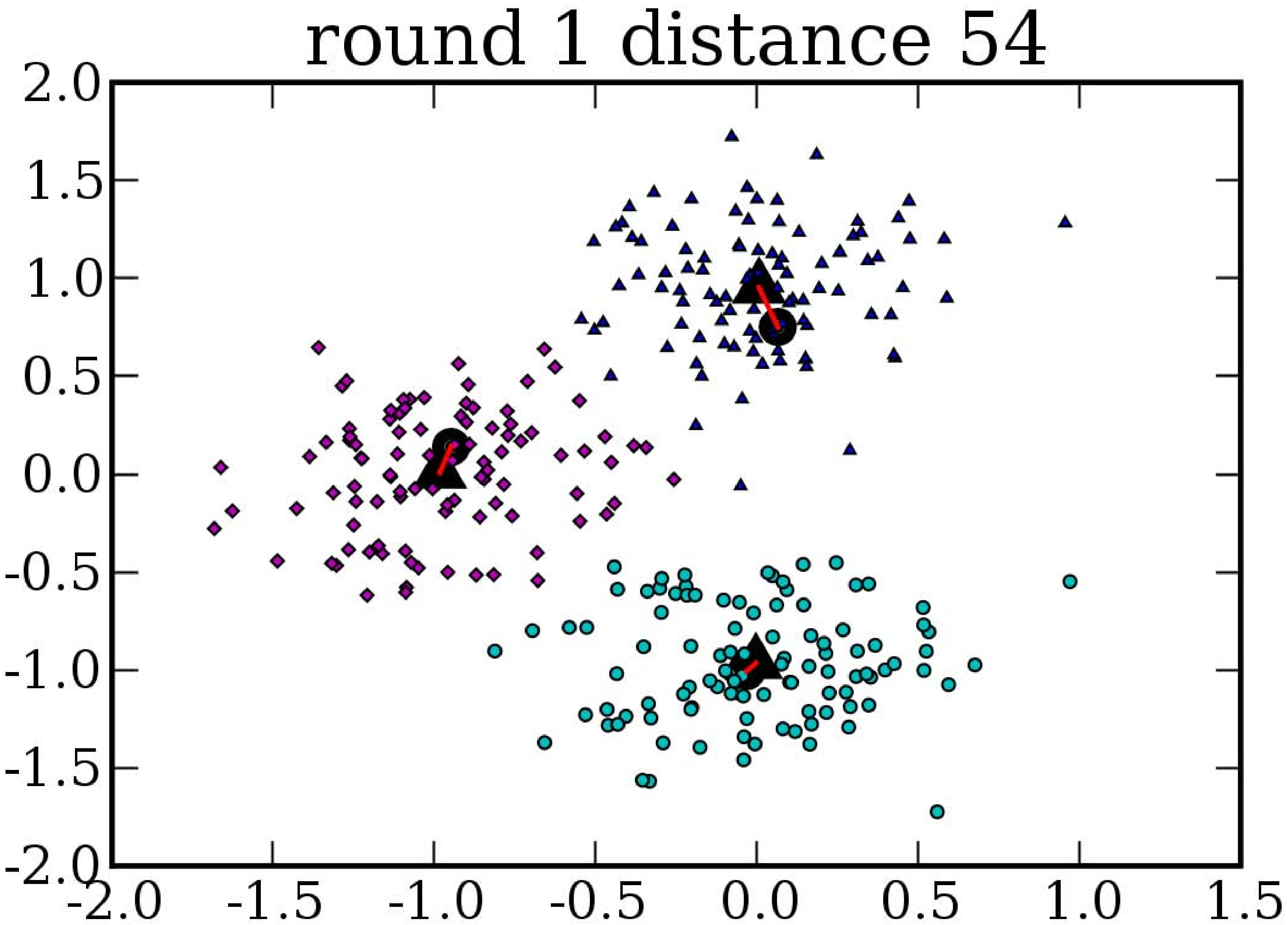
$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in C(i)} \left| X_j - \hat{Y}_i \right|^2$$

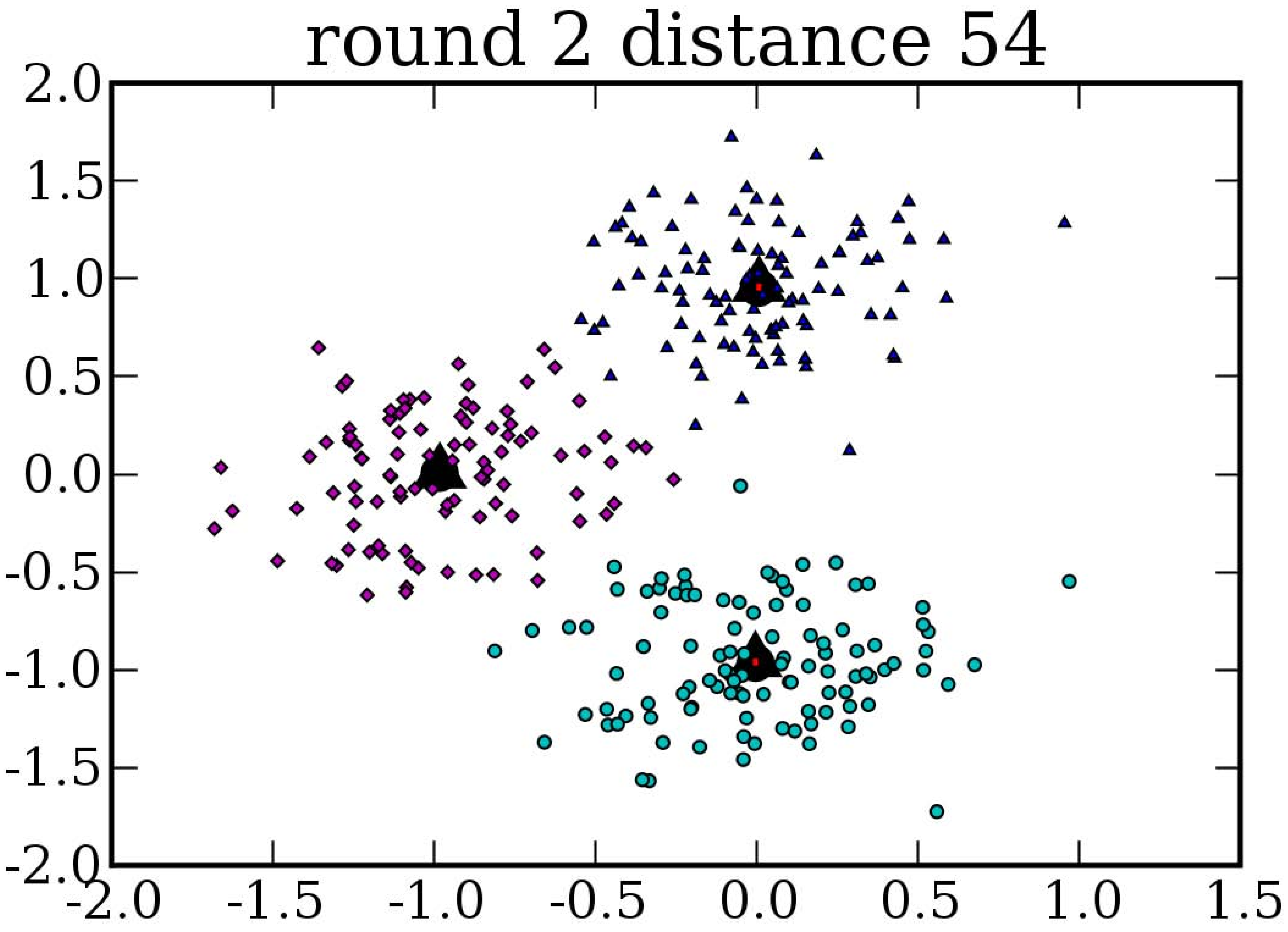
K-means clustering algorithm

- Initialize: choose k points as cluster means
- Repeat until convergence:
  - Assignment: place each point X_i in the cluster with the closest mean.
  - Update: recalculate the mean for each cluster

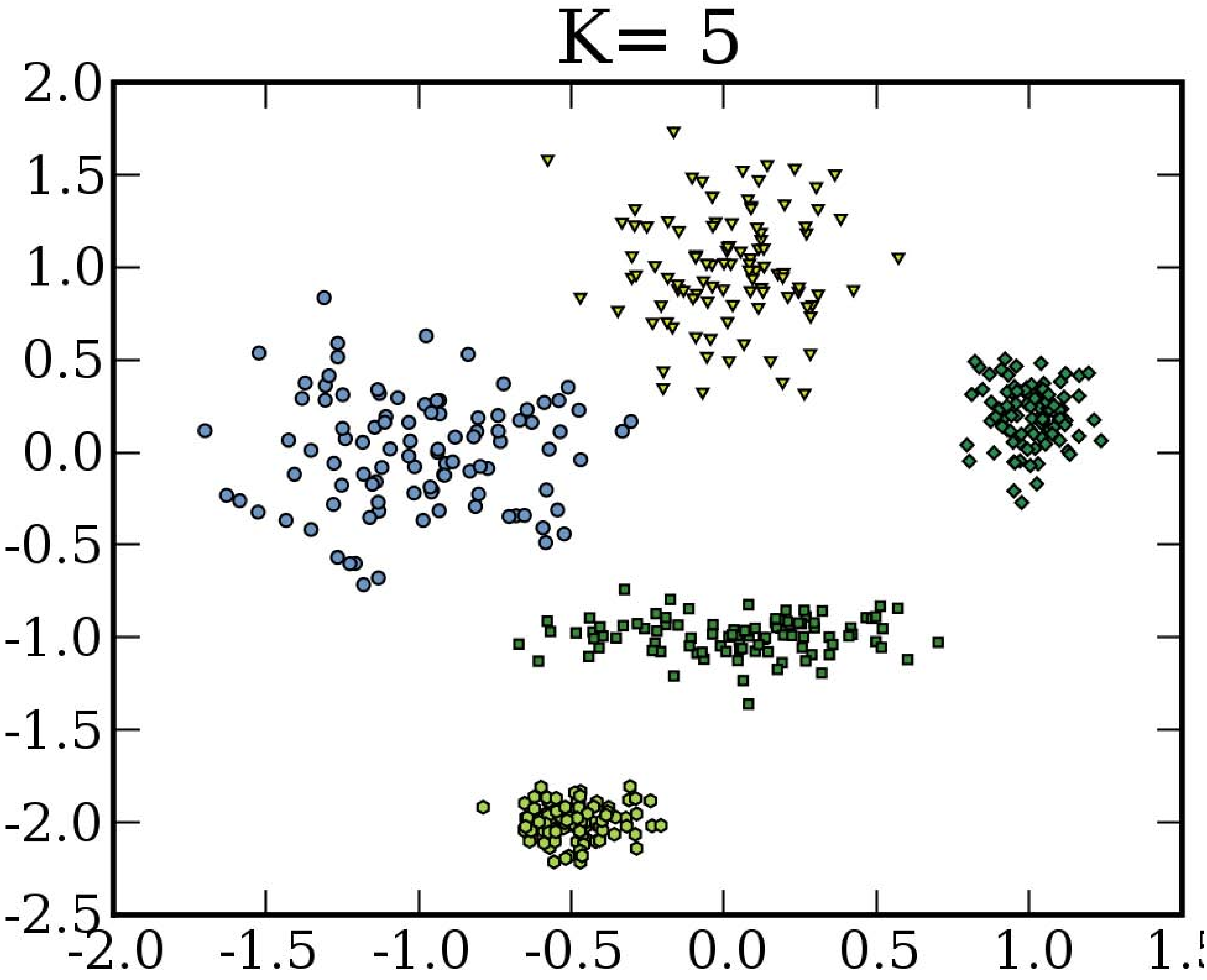



round 0 distance 86
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

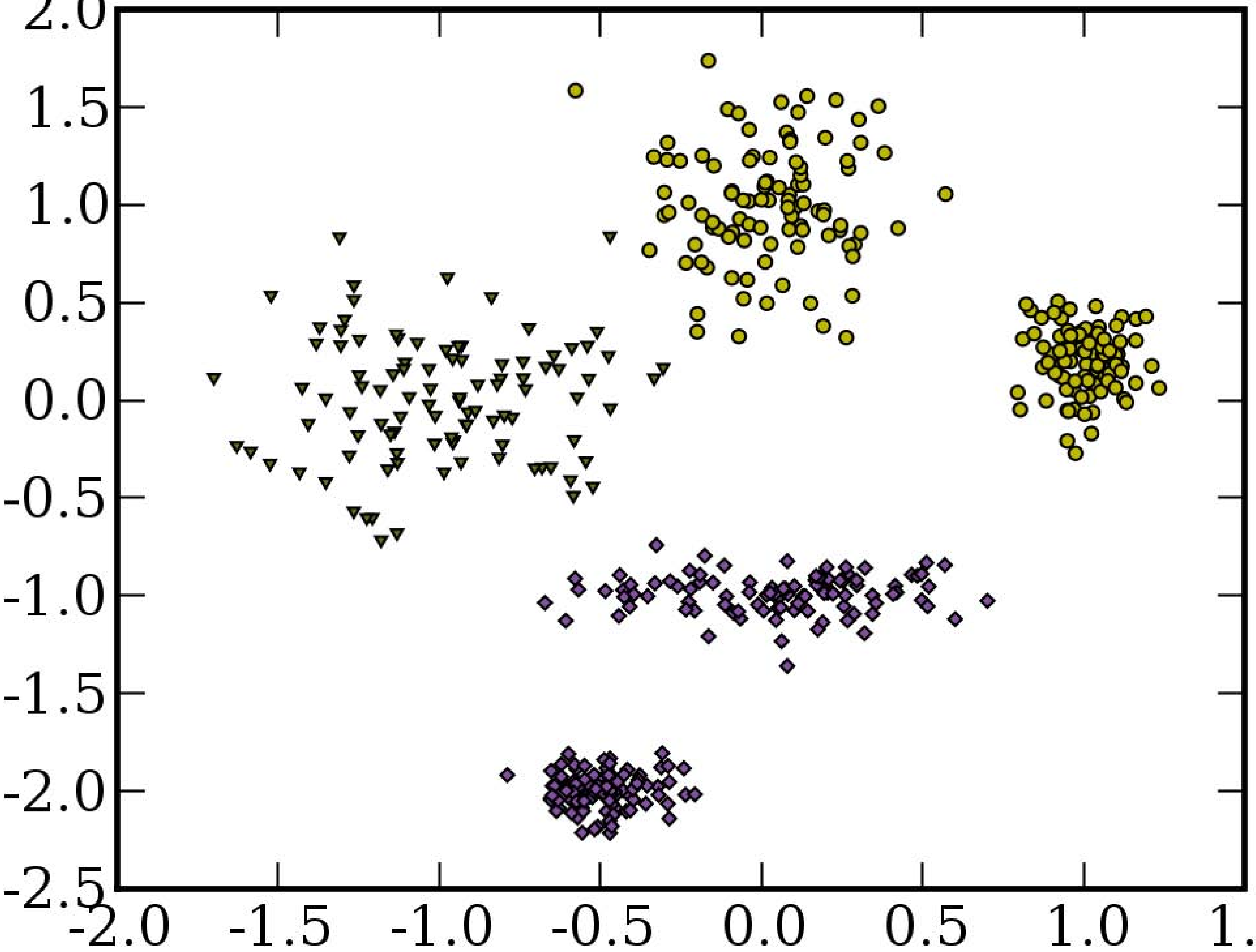

round 1 distance 54
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


round 2 distance 54
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

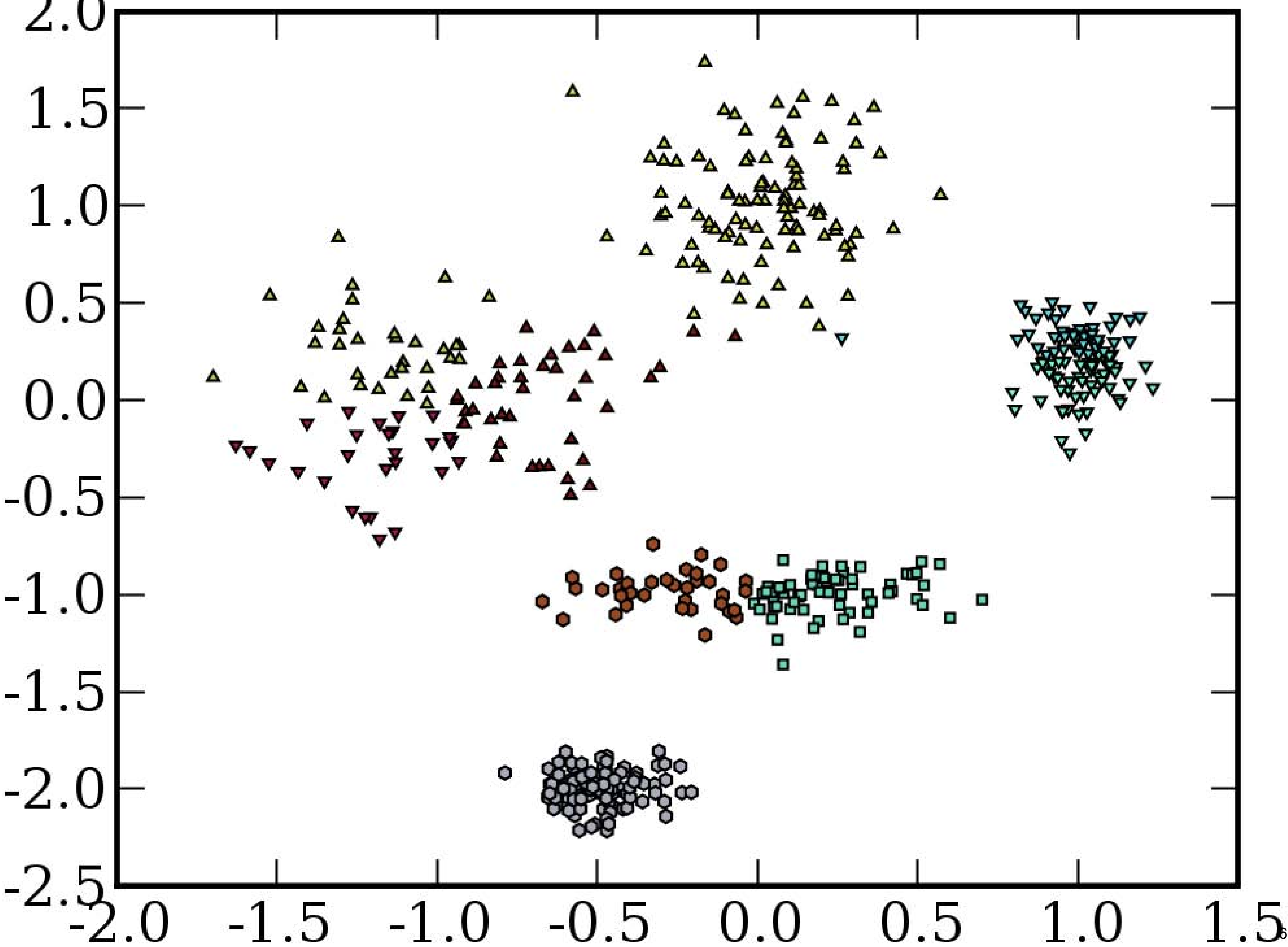
What if you choose the wrong K?

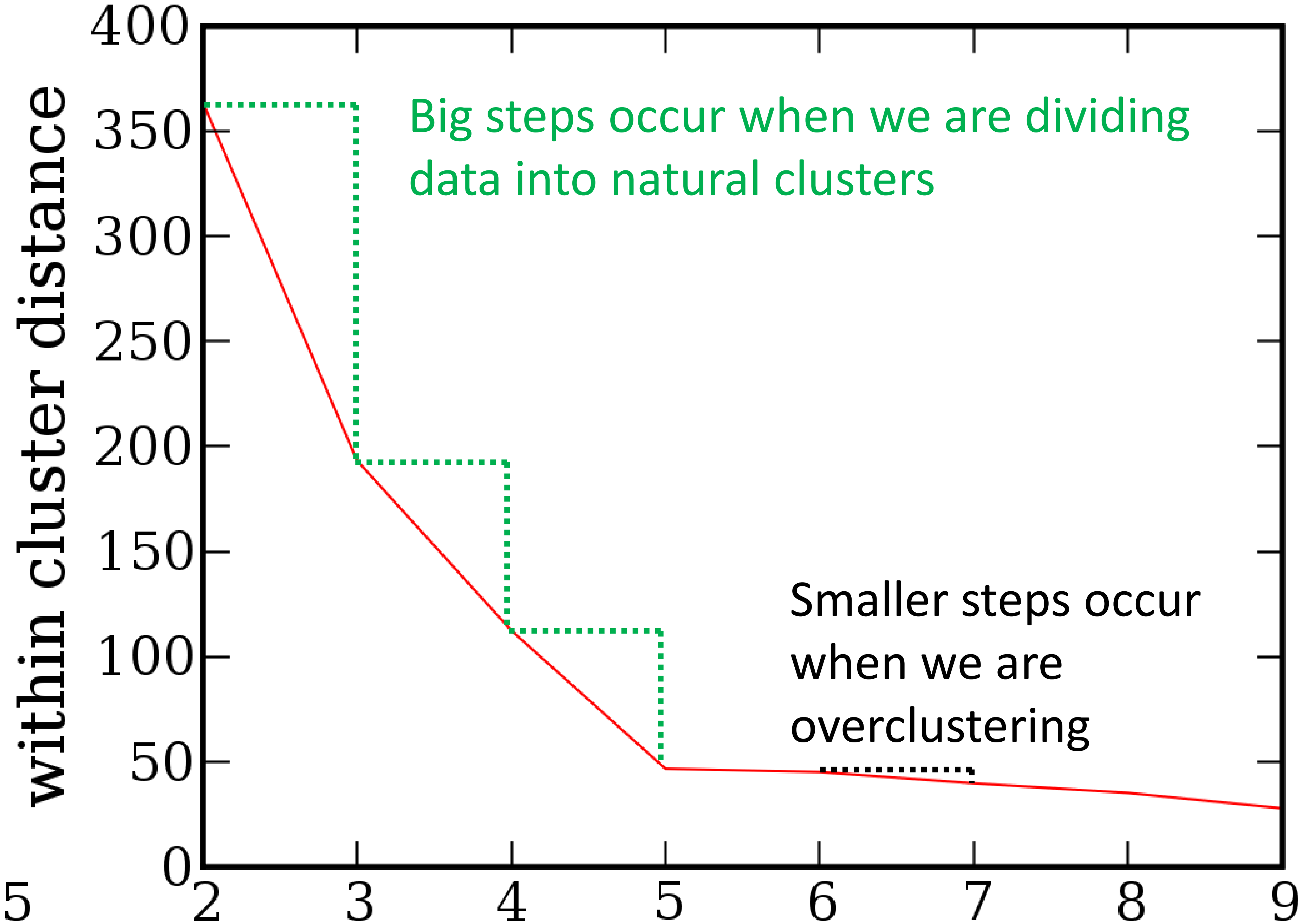

K= 5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

K= 3



K= 9



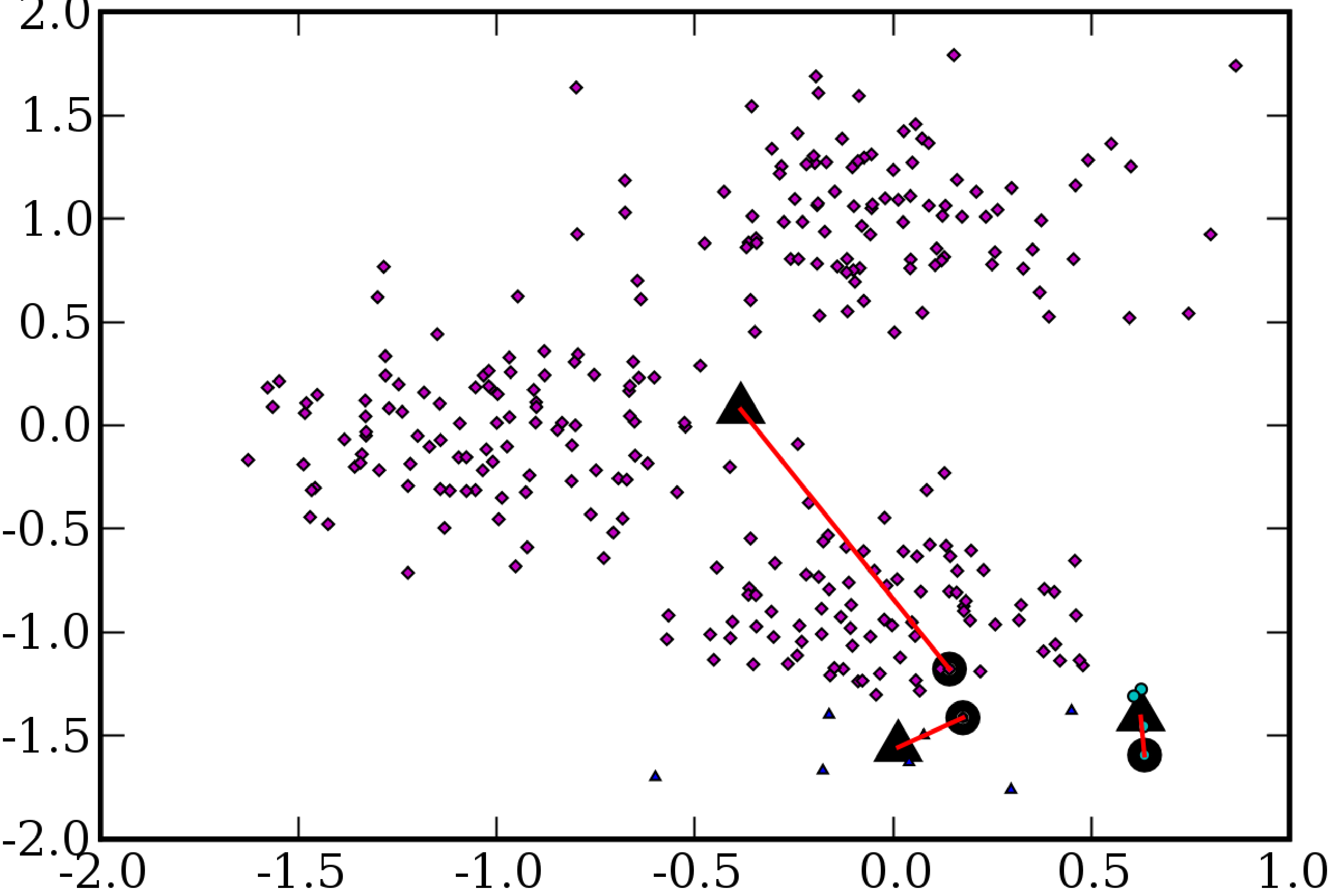


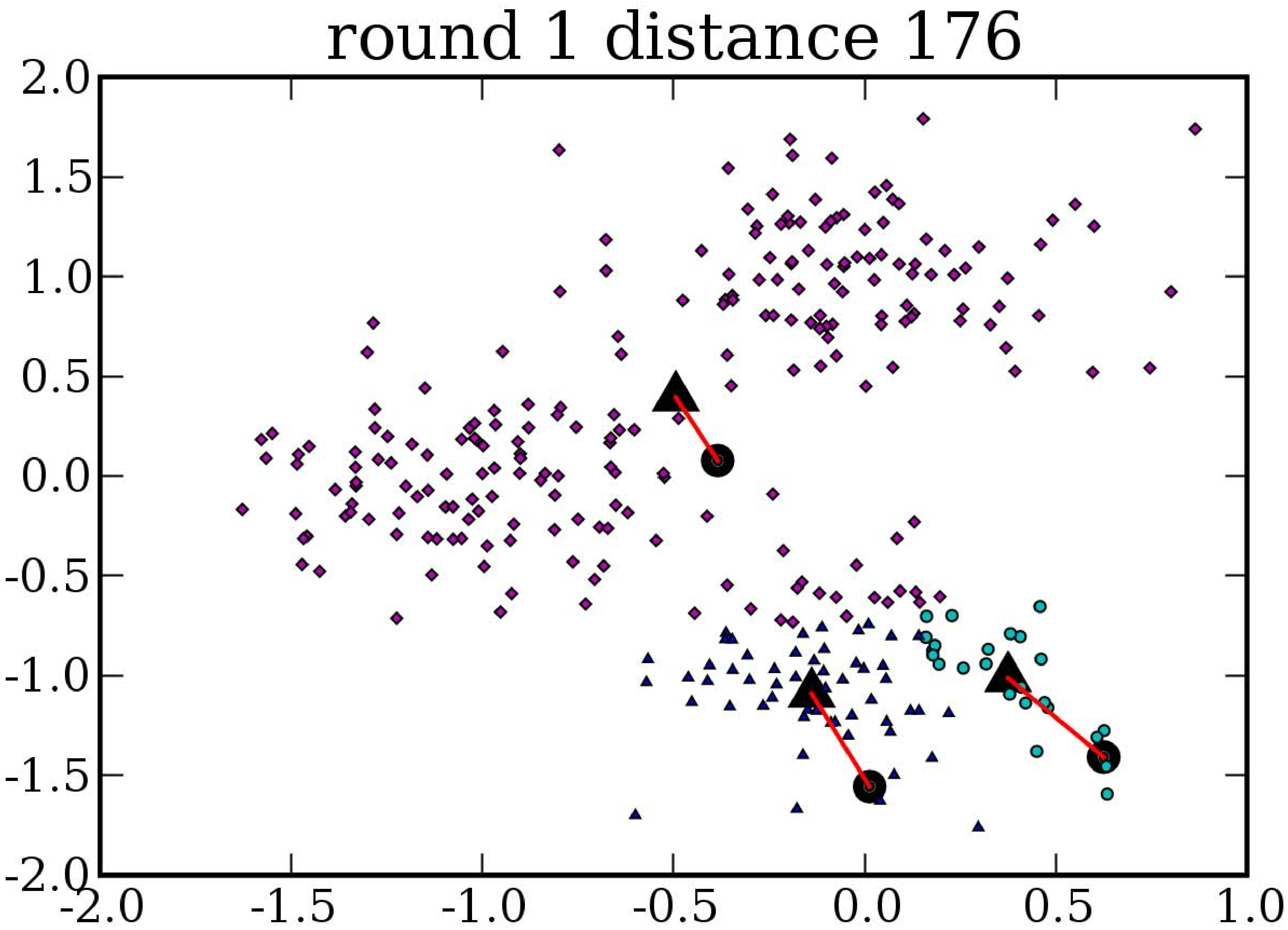
Big steps occur when we are dividing data into natural clusters

Smaller steps occur when we are overclustering

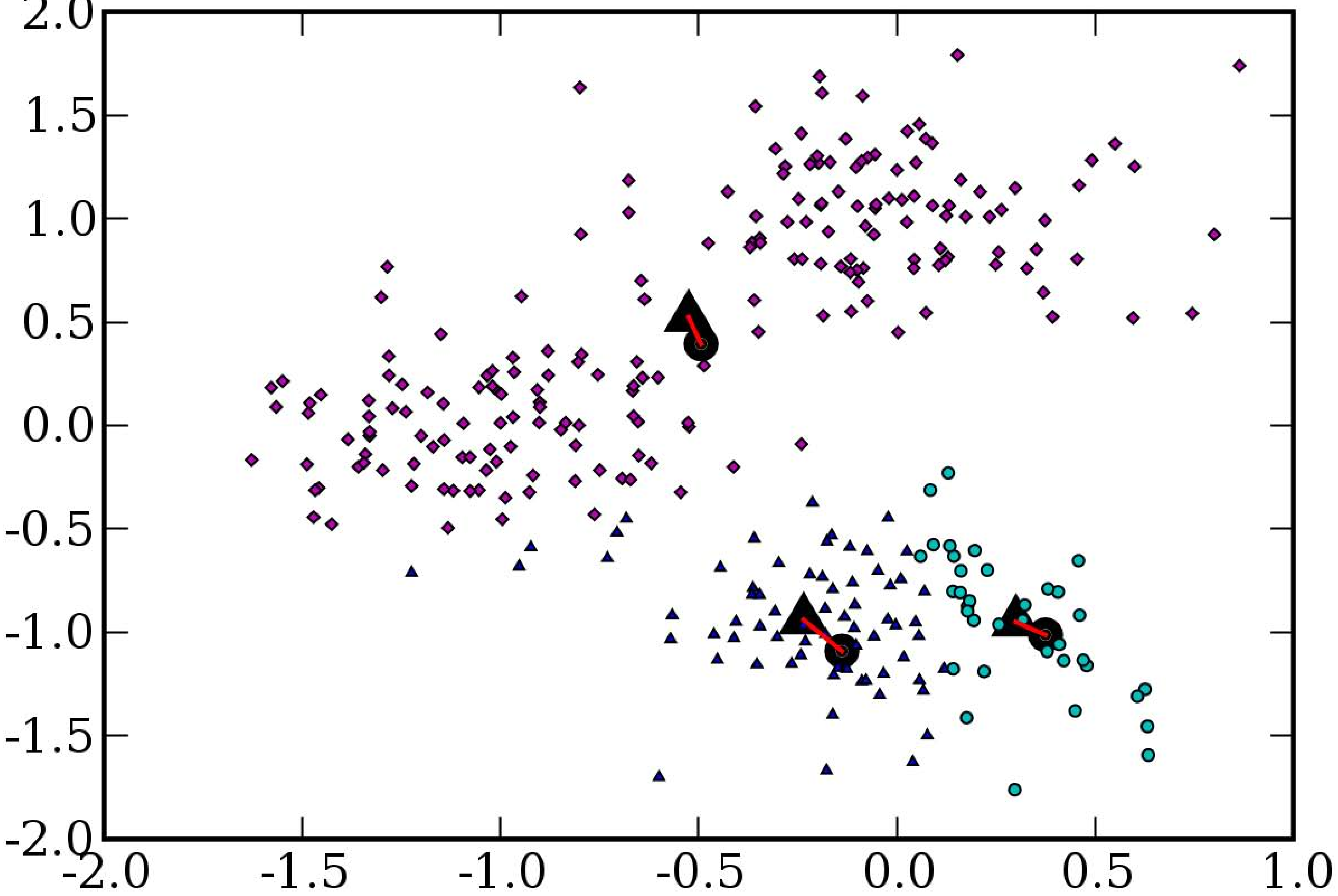
What if we choose pathologically bad initial positions?

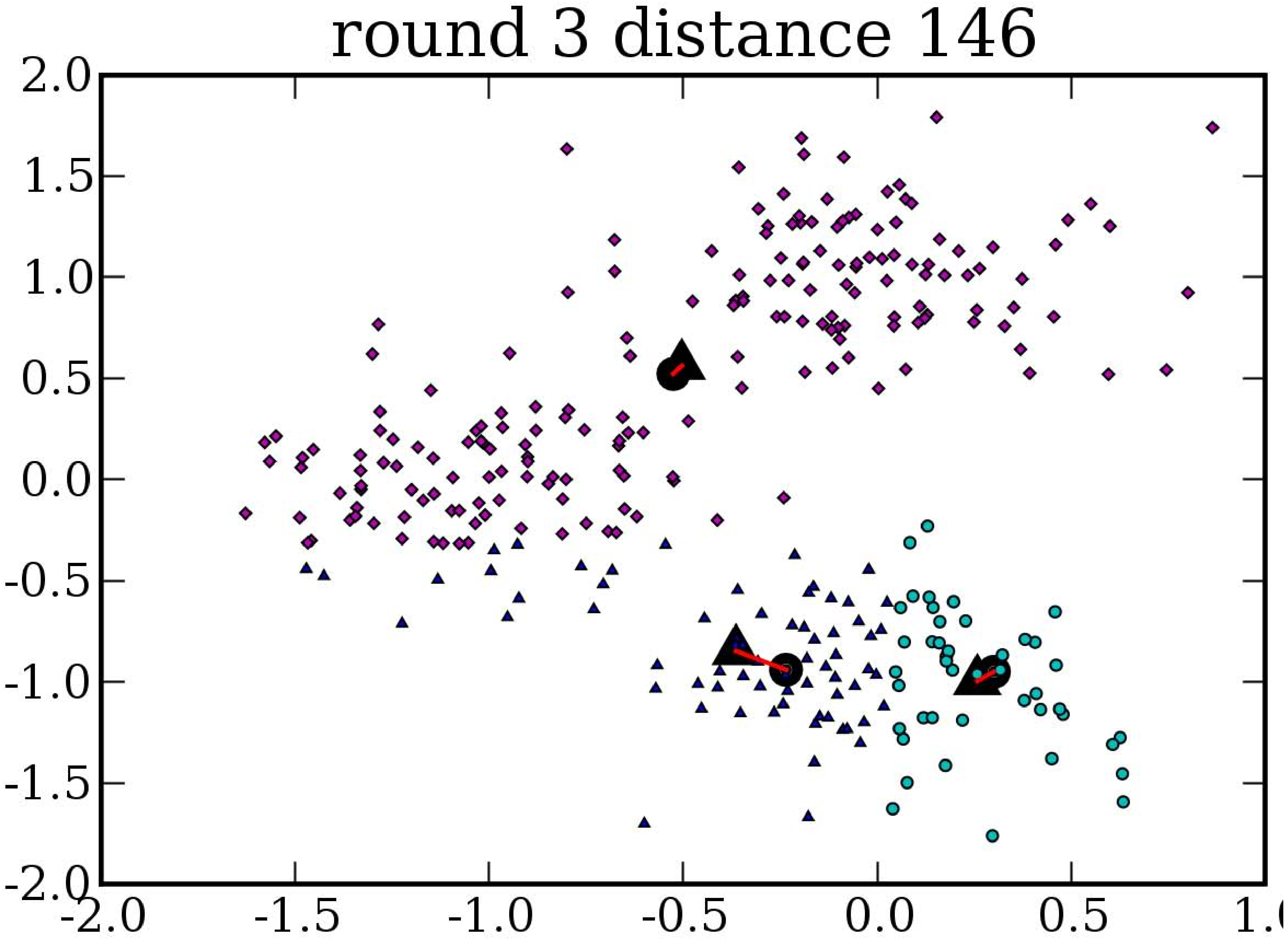
round 0 distance 285

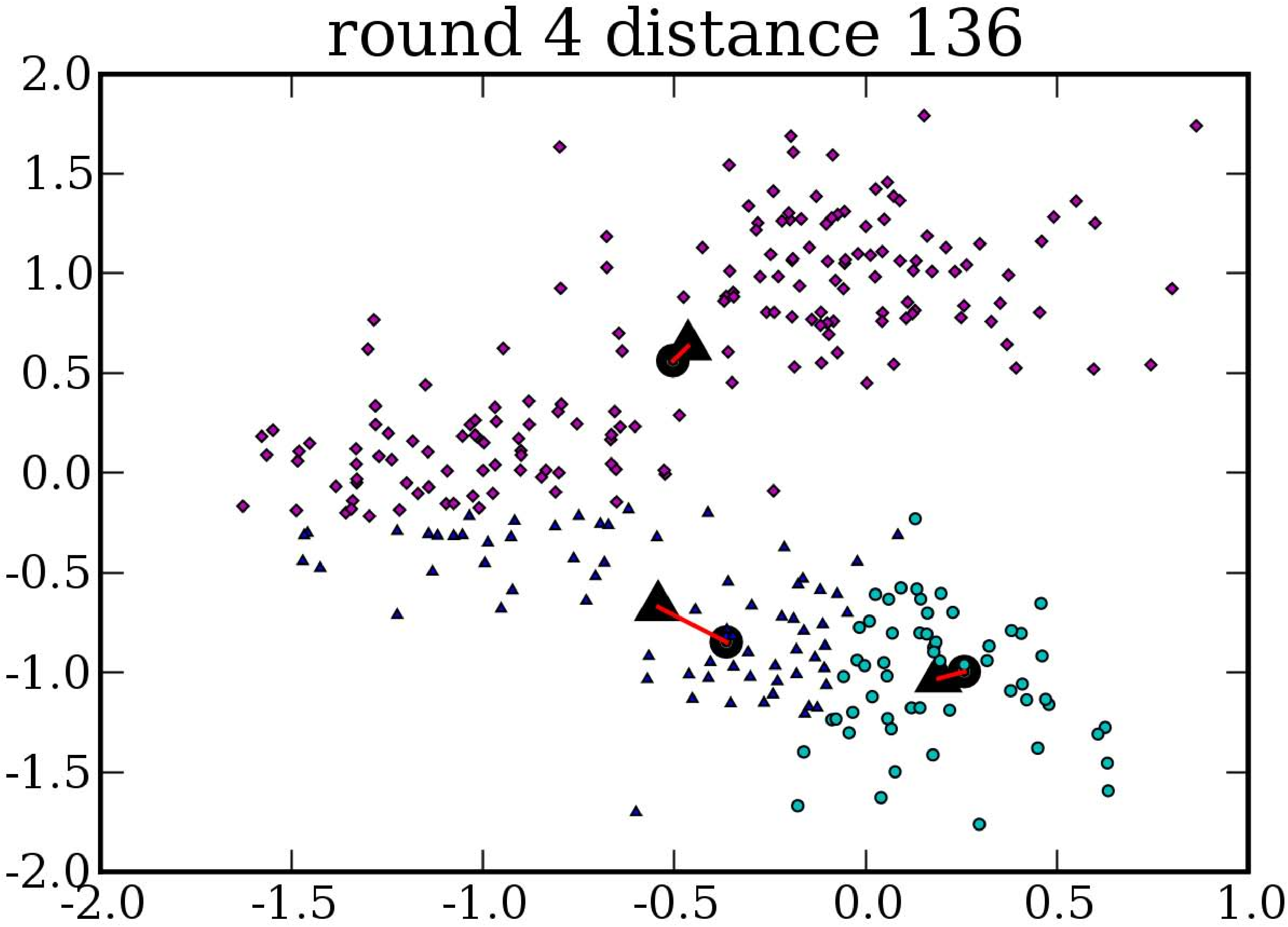



round 1 distance 176
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

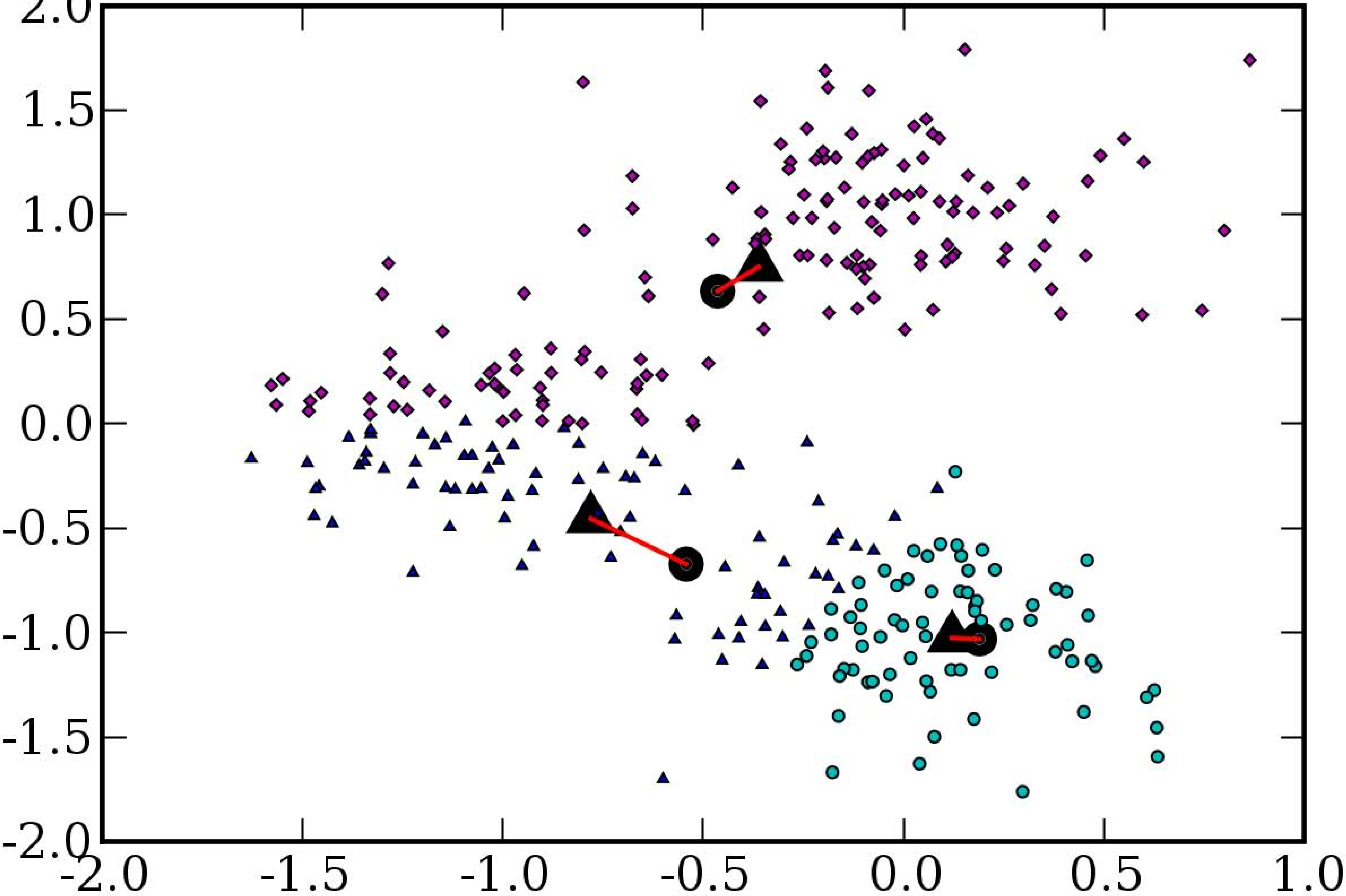
round 2 distance 150

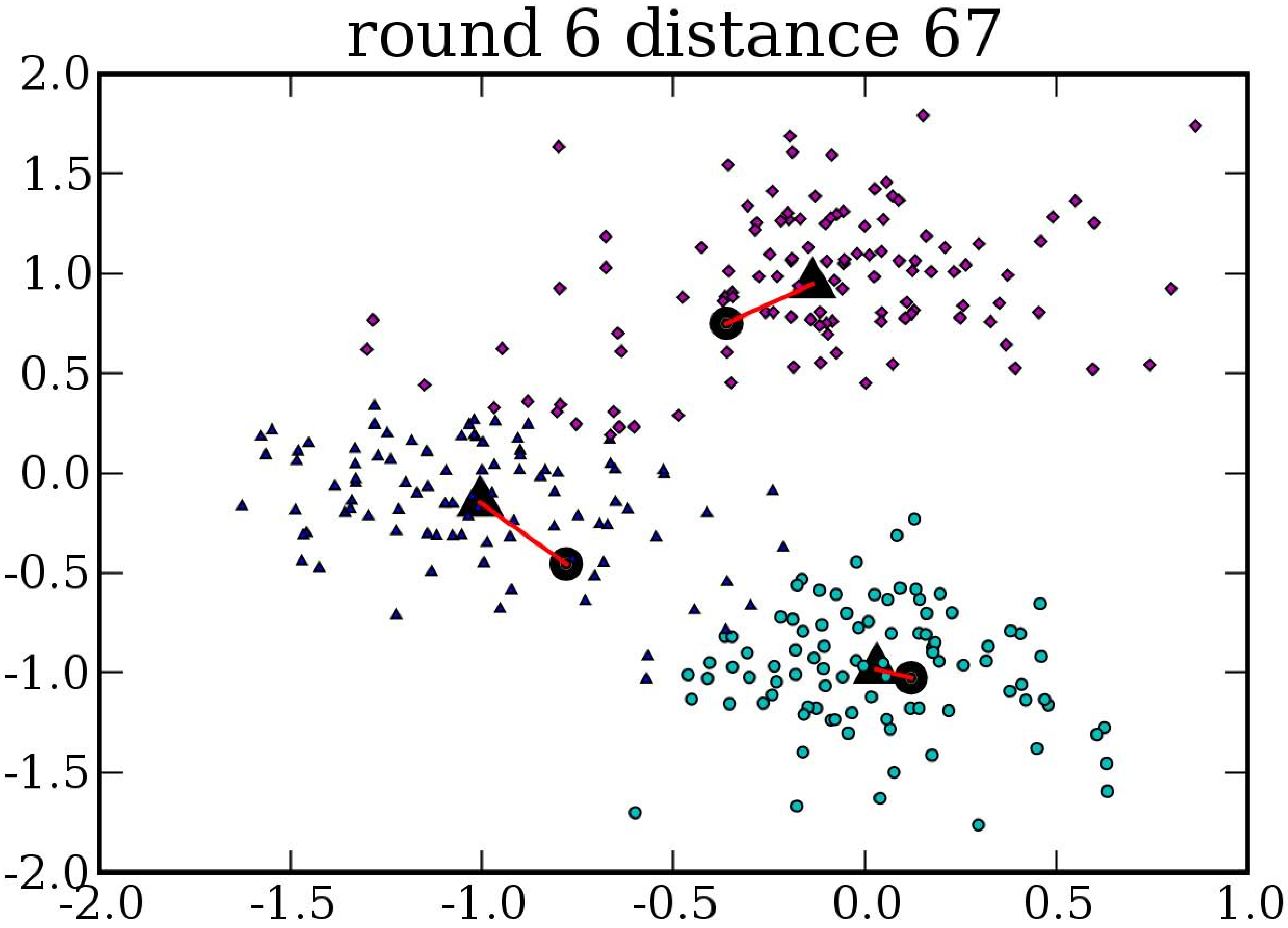



round 3 distance 146
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.(

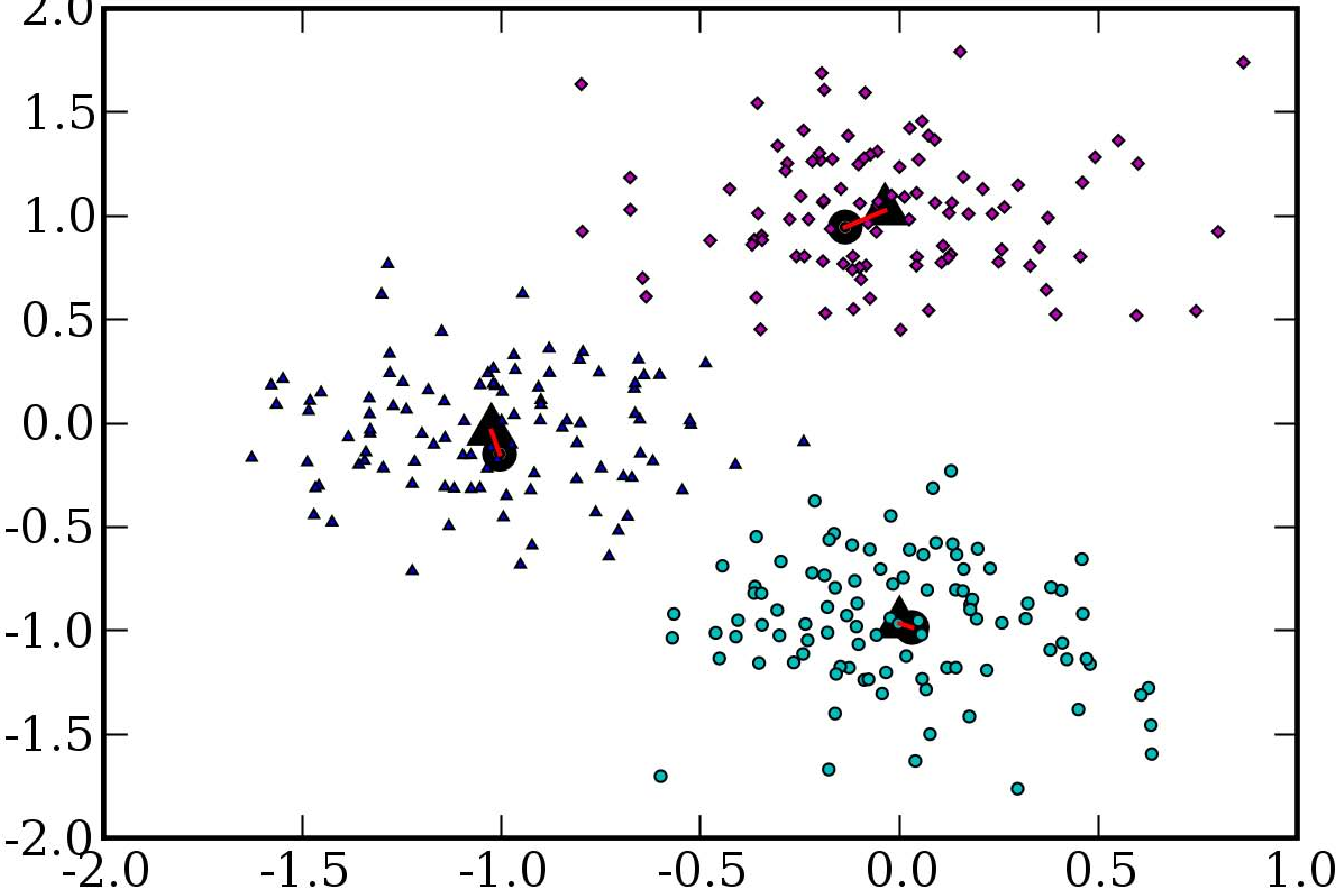

round 4 distance 136
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

round 5 distance 114

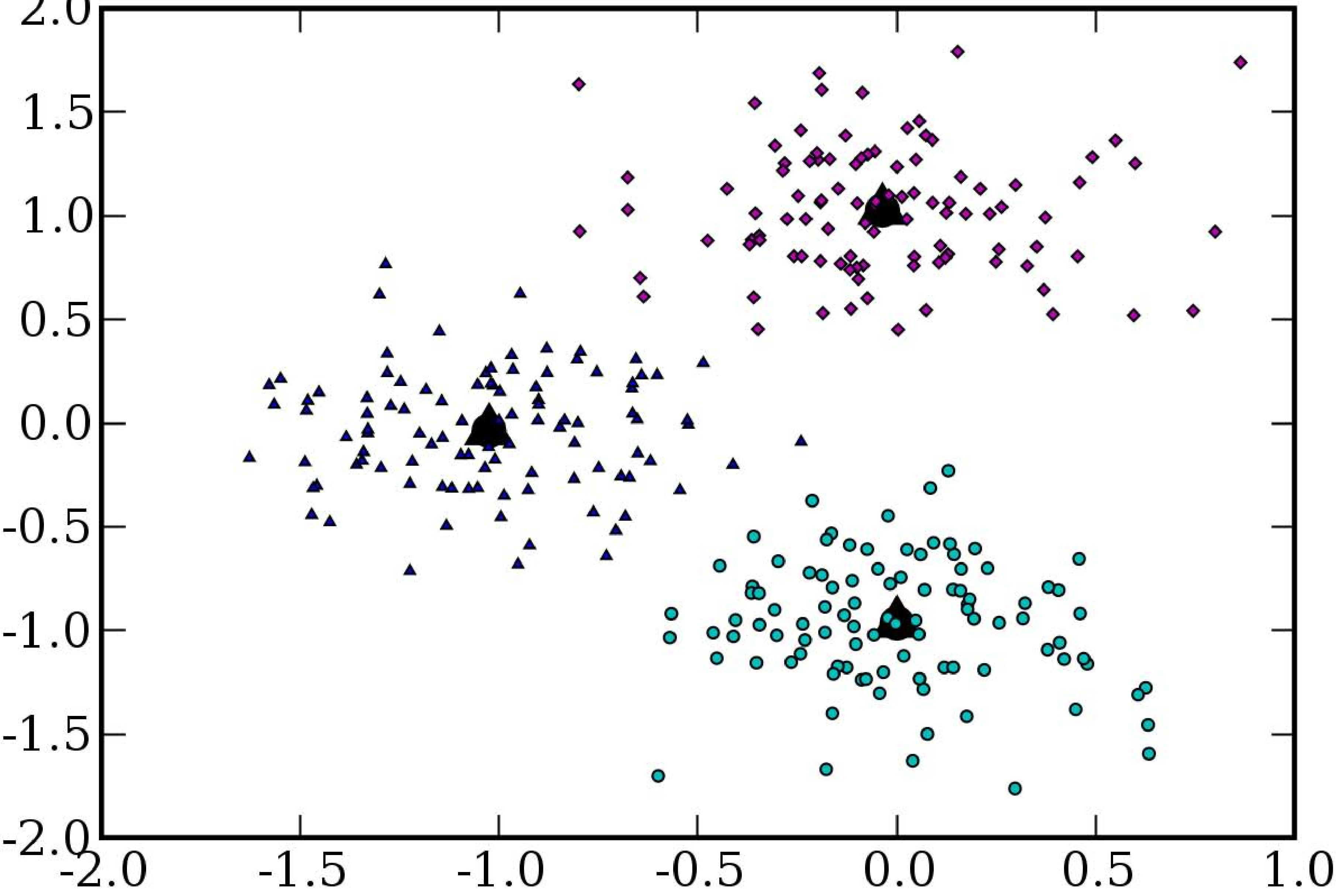



round 6 distance 67
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

round 7 distance 54



round 8 distance 54



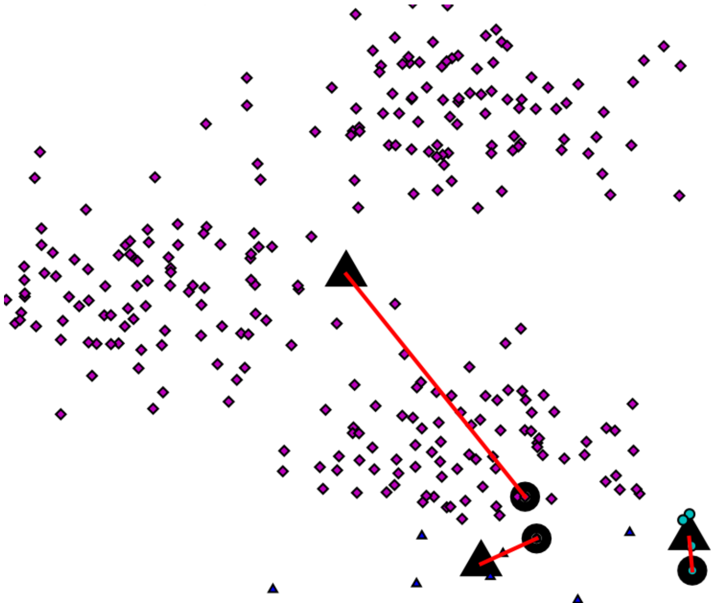
What if we choose pathologically bad initial positions?

Often, the algorithm gets a reasonable answer, but not always!

Convergence

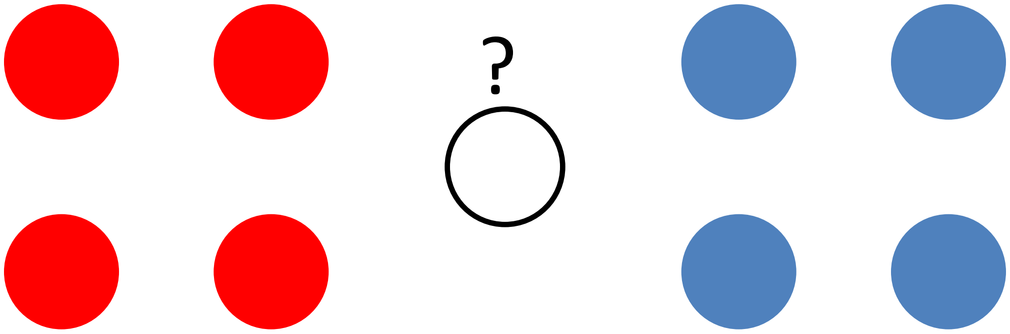
- K-means always converges.
- The assignment and update steps always either reduce the objective function or leave it unchanged.

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in C(i)} \left| X_j - \hat{Y}_i \right|^2$$



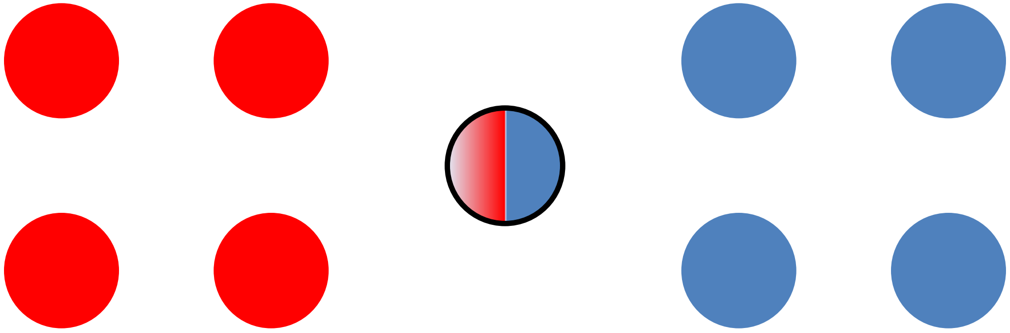
Convergence

- However, it doesn't always find the same solution.



K=2

Fuzzy K-means



K=2

K-means

- Initialize: choose k points as cluster means
- Repeat until convergence:
  - Assignment: place each point X_i in the cluster with the closest mean.
  - Update: recalculate the mean for each cluster

Fuzzy k-means

- Initialize: choose k points as cluster means
- Repeat until convergence:
  - Assignment: calculate probability of each point belonging to each cluster.
  - Update: recalculate the mean for each cluster using these probabilities

K-means

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in C(i)} \left| X_j - \hat{Y}_i \right|^2$$

$$\text{centroid}_j = \hat{Y}_j = \frac{1}{N_{Y_j}} \sum_{i \in Y_j} X_i$$

Fuzzy k-means

$$\underset{\mu, Y}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j=1}^{N} \mu_{i,\,j}^{r} \left| X_j - \hat{Y}_i \right|^2$$

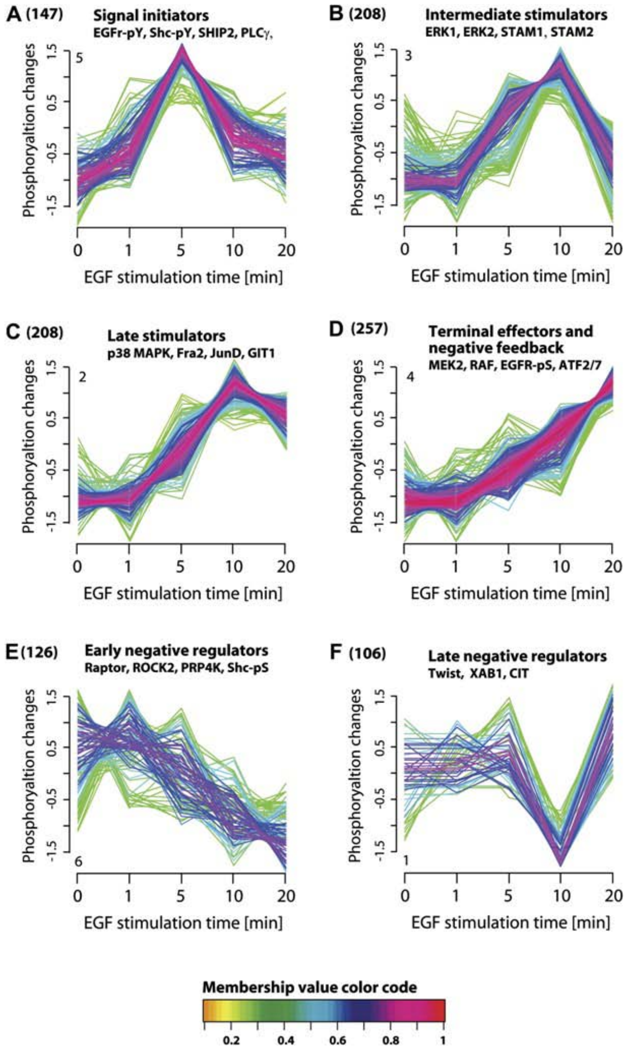
$$\text{centroid}_j = \hat{Y}_j = \frac{\sum_{i=1}^{N} \mu_{i,\,j}^{r} X_i}{\sum_{i=1}^{N} \mu_{i,\,j}^{r}}$$

$\mu_{i,\,j}^{r}$= membership of point j in cluster i

Larger values of r make the clusters more fuzzy.

Relationship to EM and Gaussian mixture models

Example of Fuzzy K-means



Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: Olsen, Jesper V., Blagoy Blagoev, et al. "Global, In Vivo, and Site-specific Phosphorylation Dynamics in Signaling Networks." *Cell* 127, no. 3 (2006): 635-48.

Olsen, *et al.* (2006) Cell. doi:10.1016/j.cell.2006.09.026

Limits of k-means

K-means uses Euclidean distance

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in C(i)} \left| X_j - \hat{Y}_i \right|^2$$

$$\text{centroid}_j = \hat{Y}_j = \frac{1}{N_{Y_j}} \sum_{i \in Y_j} X_i$$

- Gives most weight to largest differences
- Can’t be used if data are qualitative
- Centroid usually does not represent any datum

K-means

- Best clustering minimizes within-cluster Euclidean distance of from centroids

$$\text{centroid} = \hat{Y} = \frac{1}{N_Y} \sum_{i \in Y} X_{i,\,j}$$

K-medoids

- Best clustering minimizes within-cluster dissimilarity from medoids (exemplar)

$$\text{medoid}_k = \underset{i}{\text{argmin}} \sum_{i' \in C(k)} D(X_i,\ X_i')$$

K-medoids clustering

- Initialize: choose k points as cluster means
- Repeat until convergence:
  - Assignment: place each point X_i in the cluster with the closest medoid.
  - Update: recalculate the medoid for each cluster

Other approaches

- SOM (Text 16.3)
- Affinity Propagation
  - Frey and Dueck (2007) Science.

So What?

- Clusters could reveal underlying biological processes not evident from complete list of differentially expressed genes
- Clusters could be co-regulated. How could we find upstream factors?

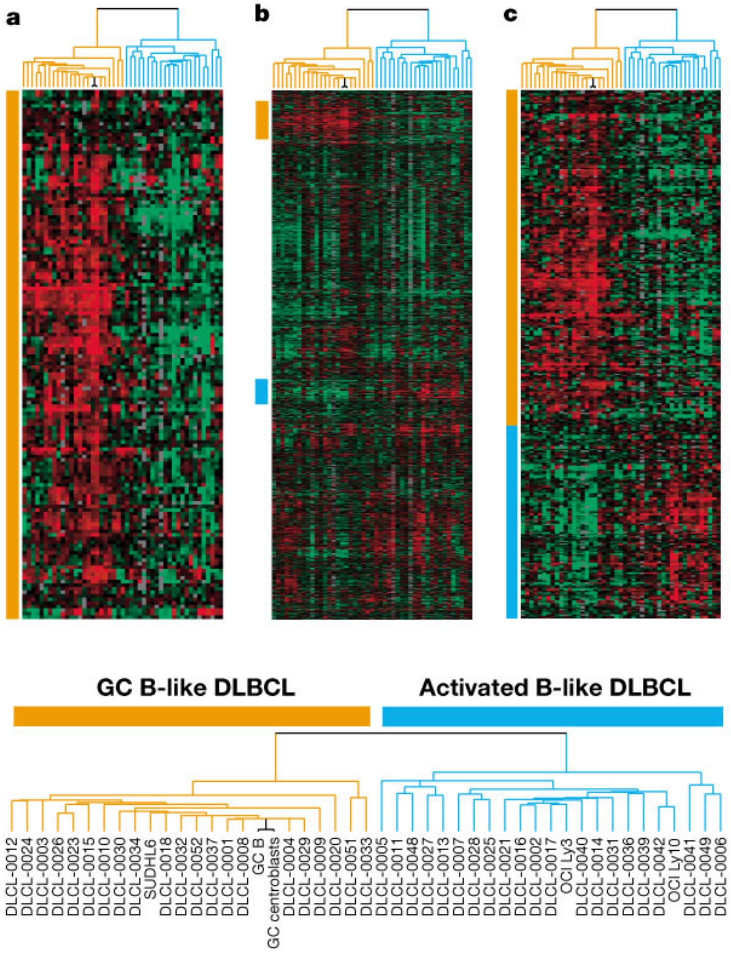
Outline

- Bayesian Networks for PPI prediction
- Gene expression
  - Distance metrics
  - Clustering
  - Signatures
  - Modules
    - Bayesian networks
    - Regression
    - Mutual Information
    - Evaluation on real and simulated data

Personalized Medicine

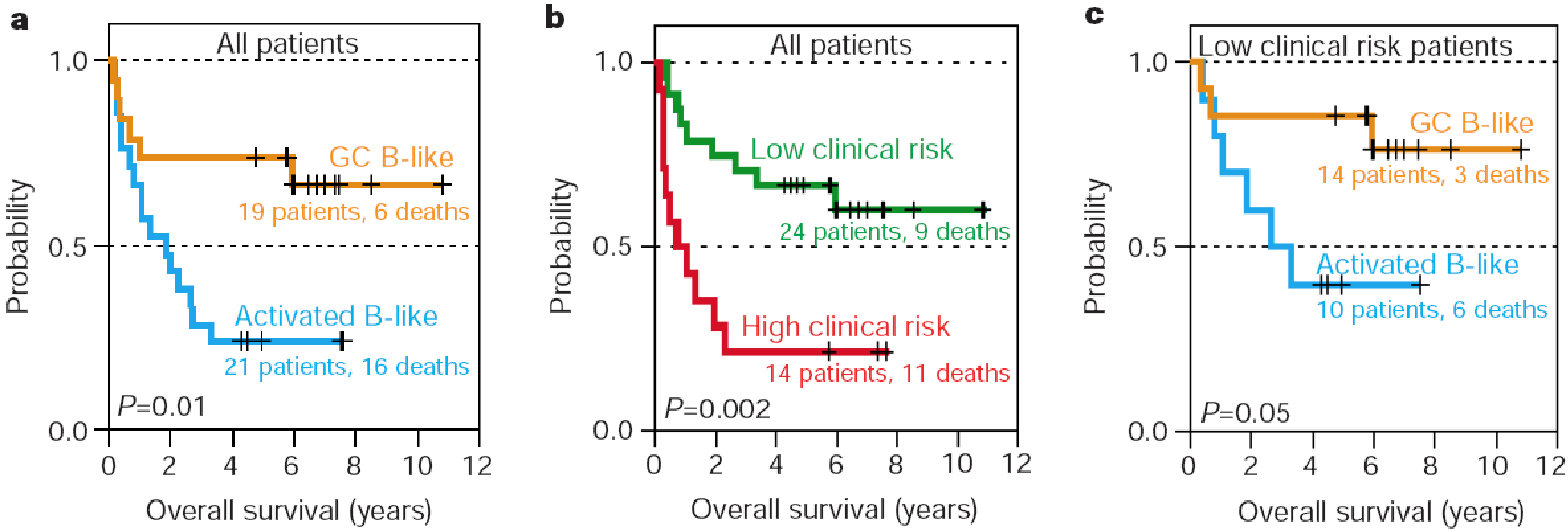
- Can gene expression be used for diagnosis and to determine the best treatment?

Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Alizadeh, Ash A., Michael B. Eisen, et al. "Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling." *Nature* 403, no. 6769 (2000): 503-11.

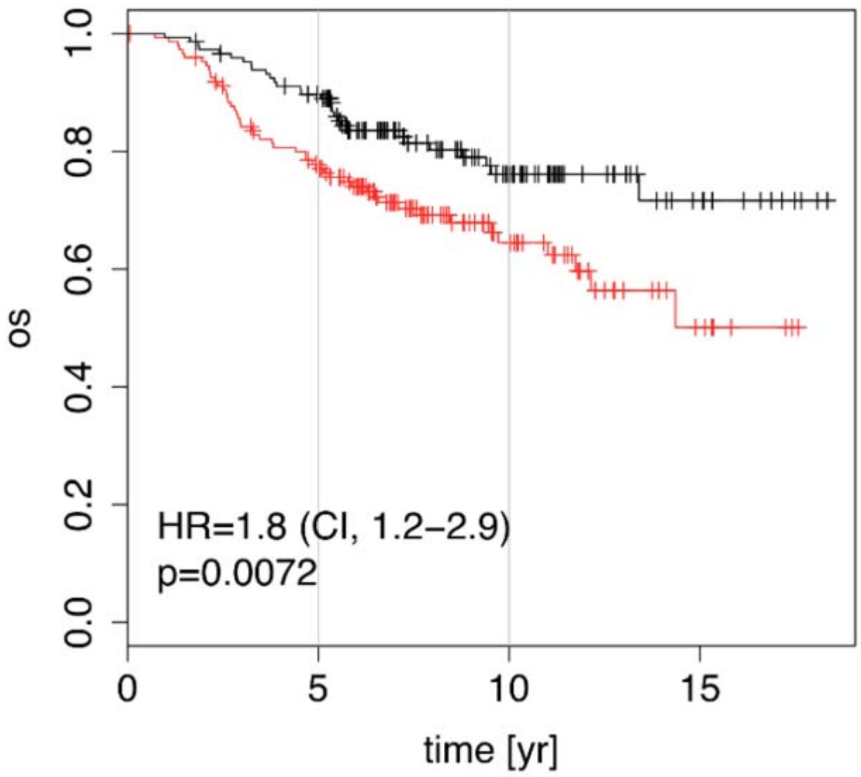
Nature **403**, 503-511(3 February 2000)



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Alizadeh, Ash A., Michael B. Eisen, et al. "Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling." *Nature* 403, no. 6769 (2000): 503-11.

Alizadeh *et al.*(2000) Nature.

post-prandial laughter

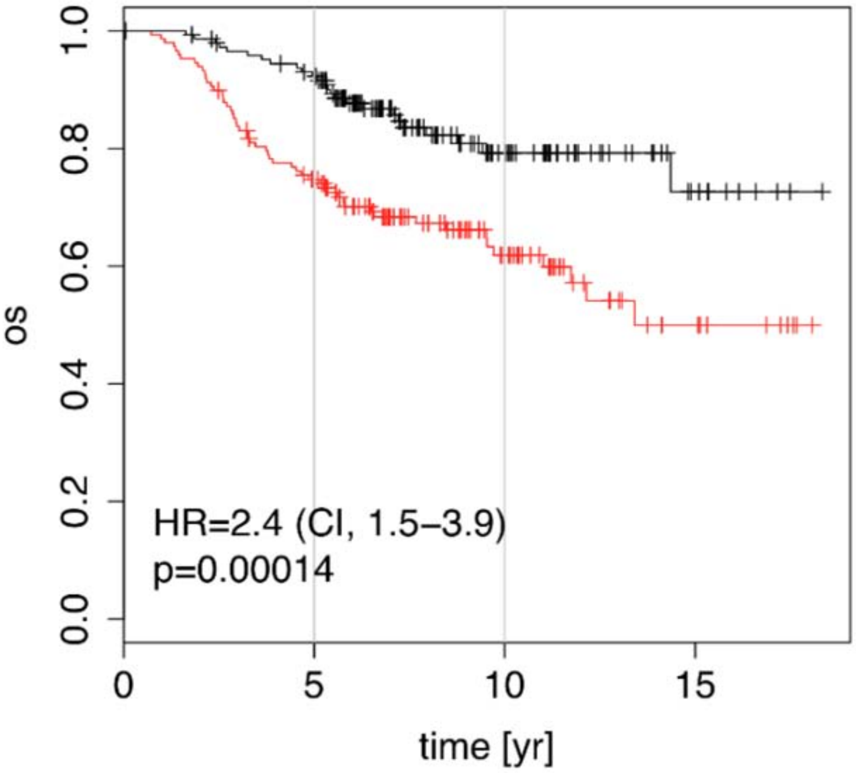


Courtesy of Venet et al. License: CC-BY.
Source: Venet, David, Jacques E. Dumont, et al. "Most Random Gene Expression Signatures are Significantly Associated with Breast Cancer Outcome." *PLoS Computational Biology* 7, no. 10 (2011): e1002240.

Testing whether laughter IS the best medicine

OS= the fraction of patients alive (overall survival)
Hazard Ratio= Death rate vs. control

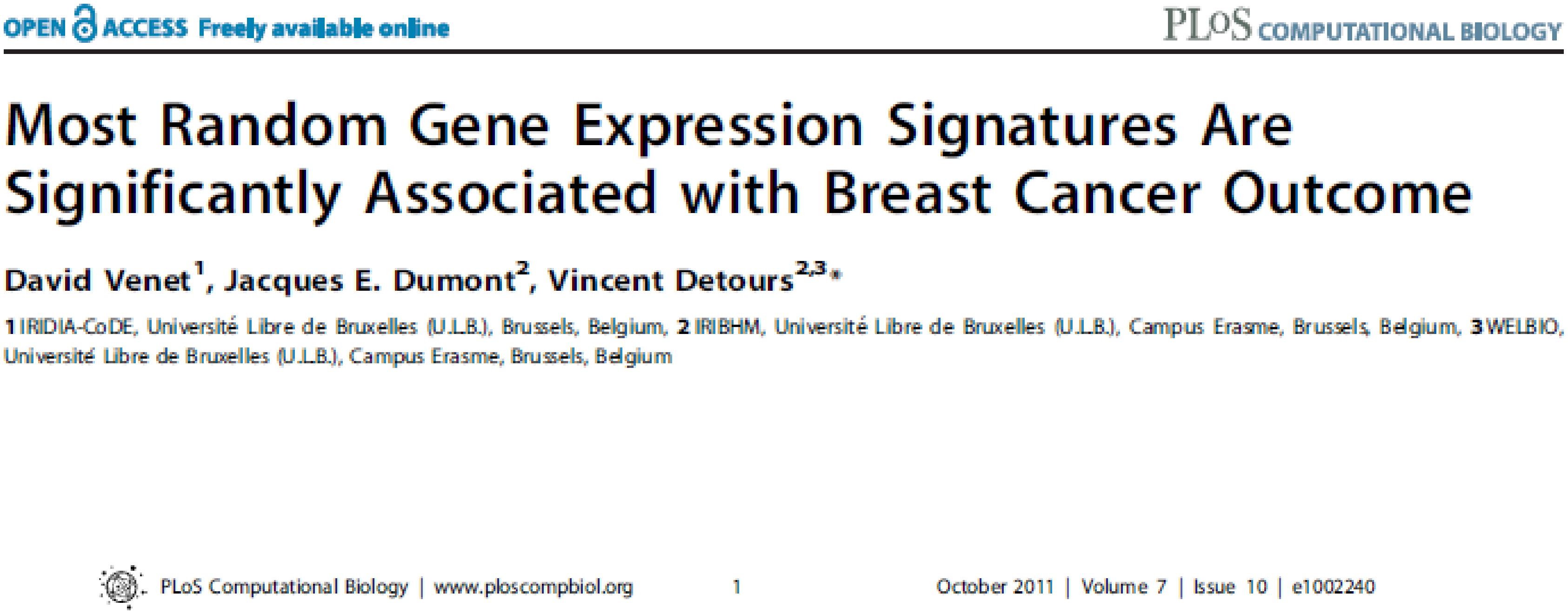
social defeat in mice



Courtesy of Venet et al. License: CC-BY.
Source: Venet, David, Jacques E. Dumont, et al. "Most Random Gene Expression Signatures are Significantly Associated with Breast Cancer Outcome." *PLoS Computational Biology* 7, no. 10 (2011): e1002240.

Testing whether laughter IS the best medicine

OS= the fraction of patients alive (overall survival)
Hazard Ratio= Death rate vs. control



OPEN ACCESS Freely available online

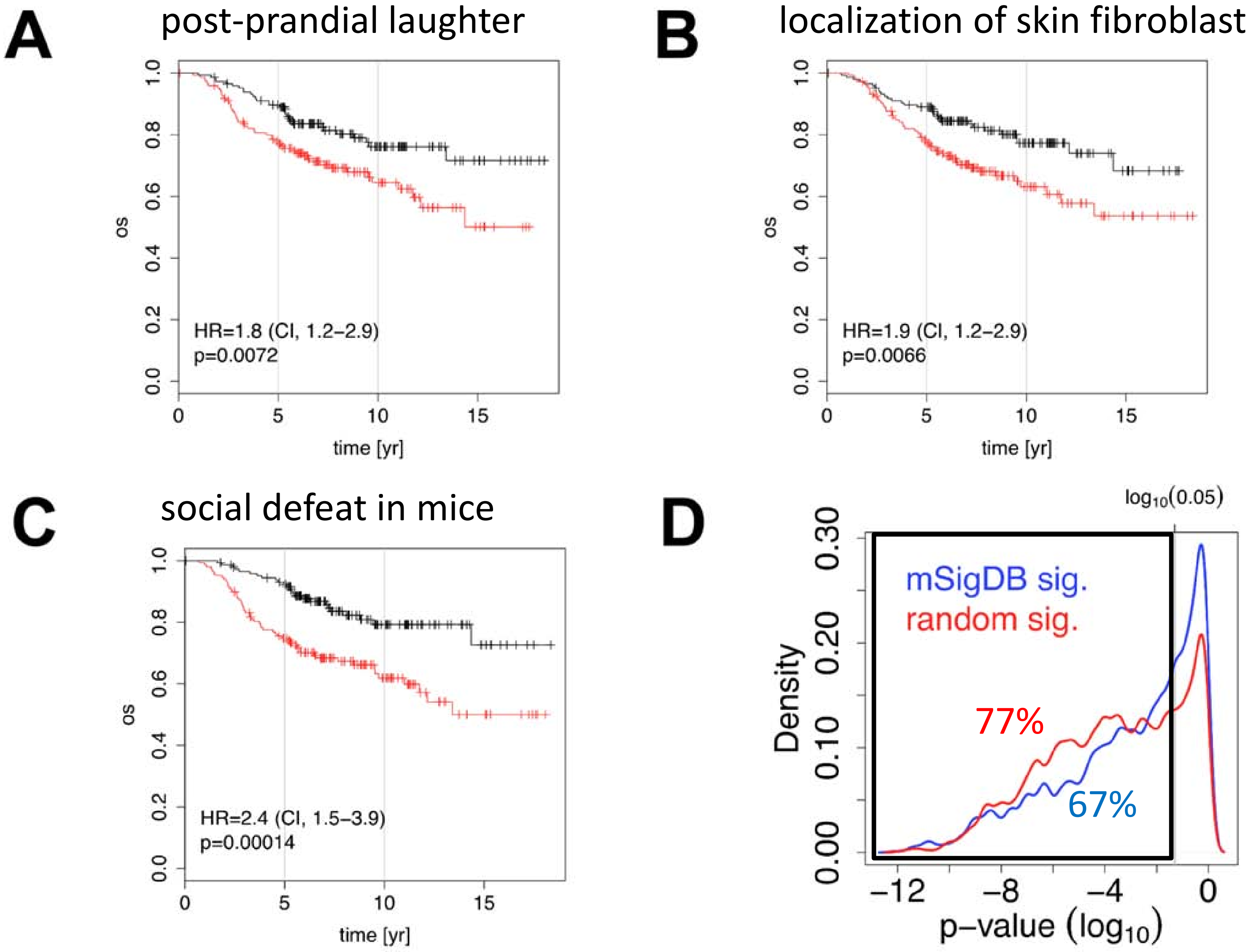
PLoS COMPUTATIONAL BIOLOGY

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

[1] IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, [2] IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, [3] WELBIO, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium

PLoS Computational Biology | www.ploscompbiol.org 1 October 2011 | Volume 7 | Issue 10 | e1002240

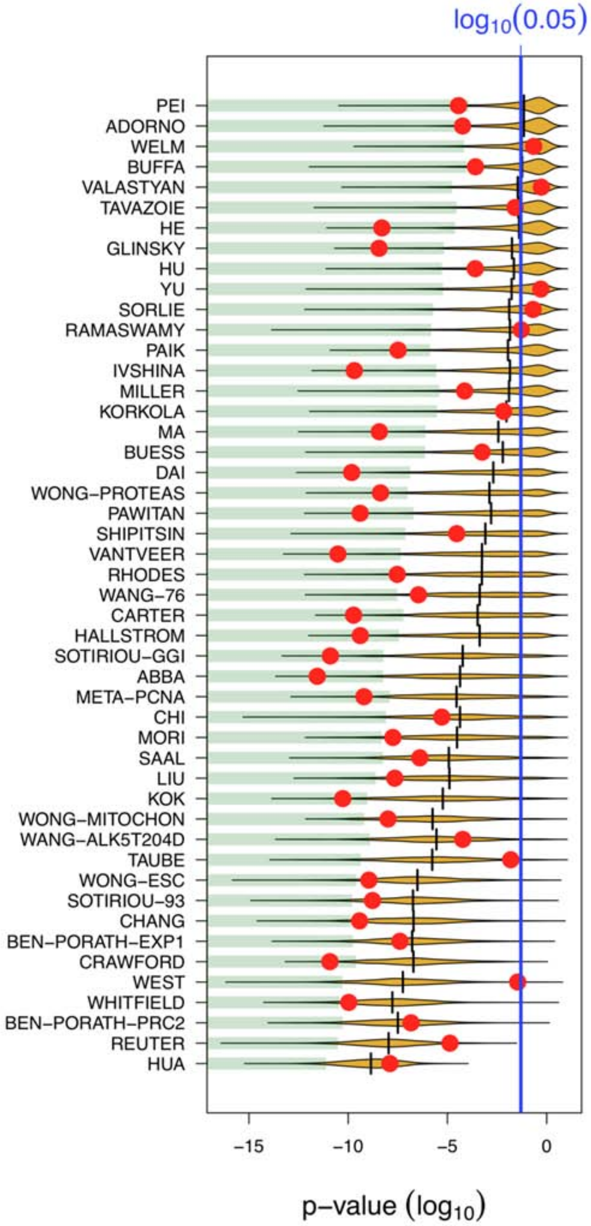


Courtesy of Venet et al. License: CC-BY.

Source: Venet, David, Jacques E. Dumont, et al. "Most Random Gene Expression Signatures are Significantly Associated with Breast Cancer Outcome." *PLoS Computational Biology* 7, no. 10 (2011): e1002240.

OS= the fraction of patients alive (overall survival)

Hazard Ratio= Death rate vs. control



● Published Signature

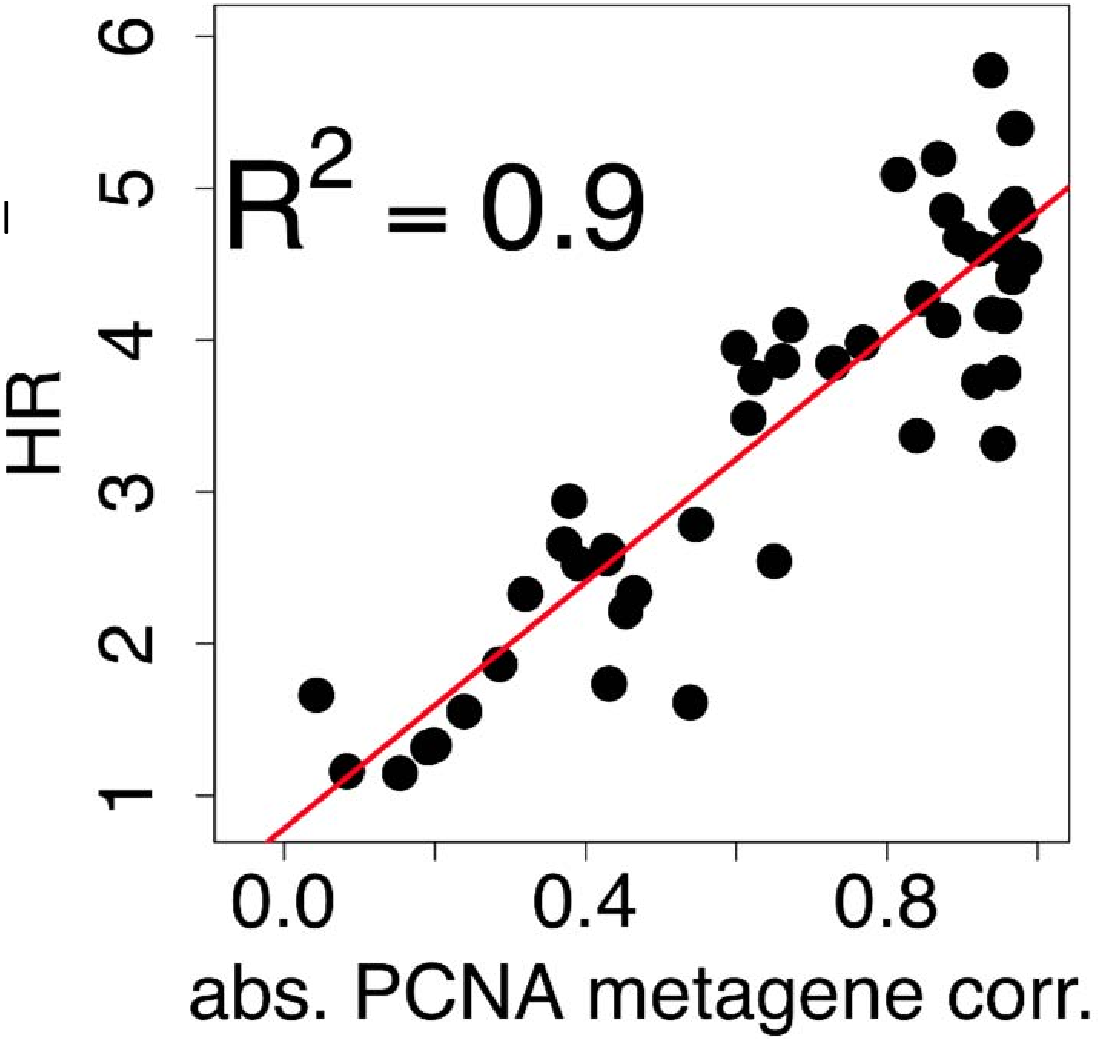
Distribution for random signatures

▭ Best 5% of random signatures

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

Source: Venet, David, Jacques E. Dumont, et al. "Most Random Gene Expression Signatures are Significantly Associated with Breast Cancer Outcome." *PLoS Computational Biology* 7, no. 10 (2011): e1002240.

Hazard Ratio=
Death rate vs. control



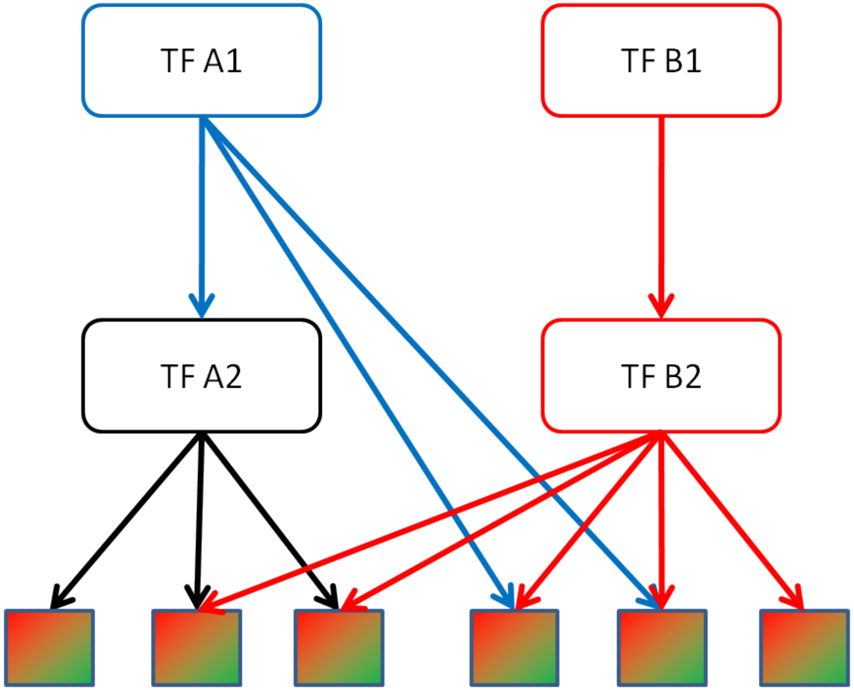
Courtesy of Venet et al. License: CC-BY.
Source: Venet, David, Jacques E. Dumont, et al. "Most Random Gene Expression Signatures are Significantly Associated with Breast Cancer Outcome." *PLoS Computational Biology* 7, no. 10 (2011): e1002240.

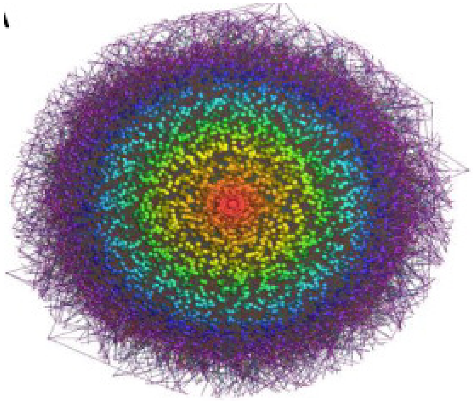
PCNA metagene = 1% genes the most positively correlated with expression of PCNA (proliferating cell nuclear antigen, a known marker) across 36 tissues

Outline

- Bayesian Networks for PPI prediction
- Gene expression
  - Distance metrics
  - Clustering
  - Signatures
  - Modules
    - Bayesian networks
    - Regression
    - Mutual Information
    - Evaluation on real and simulated data

Reconstructing Regulatory Networks

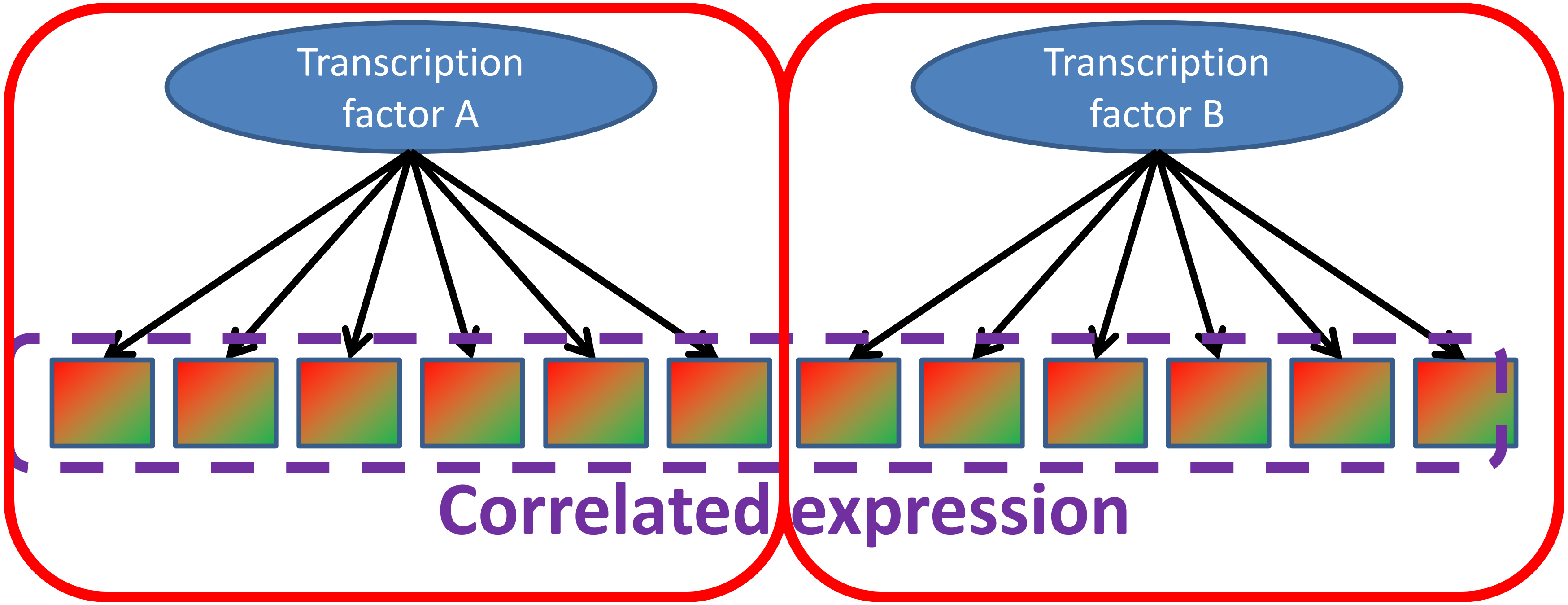


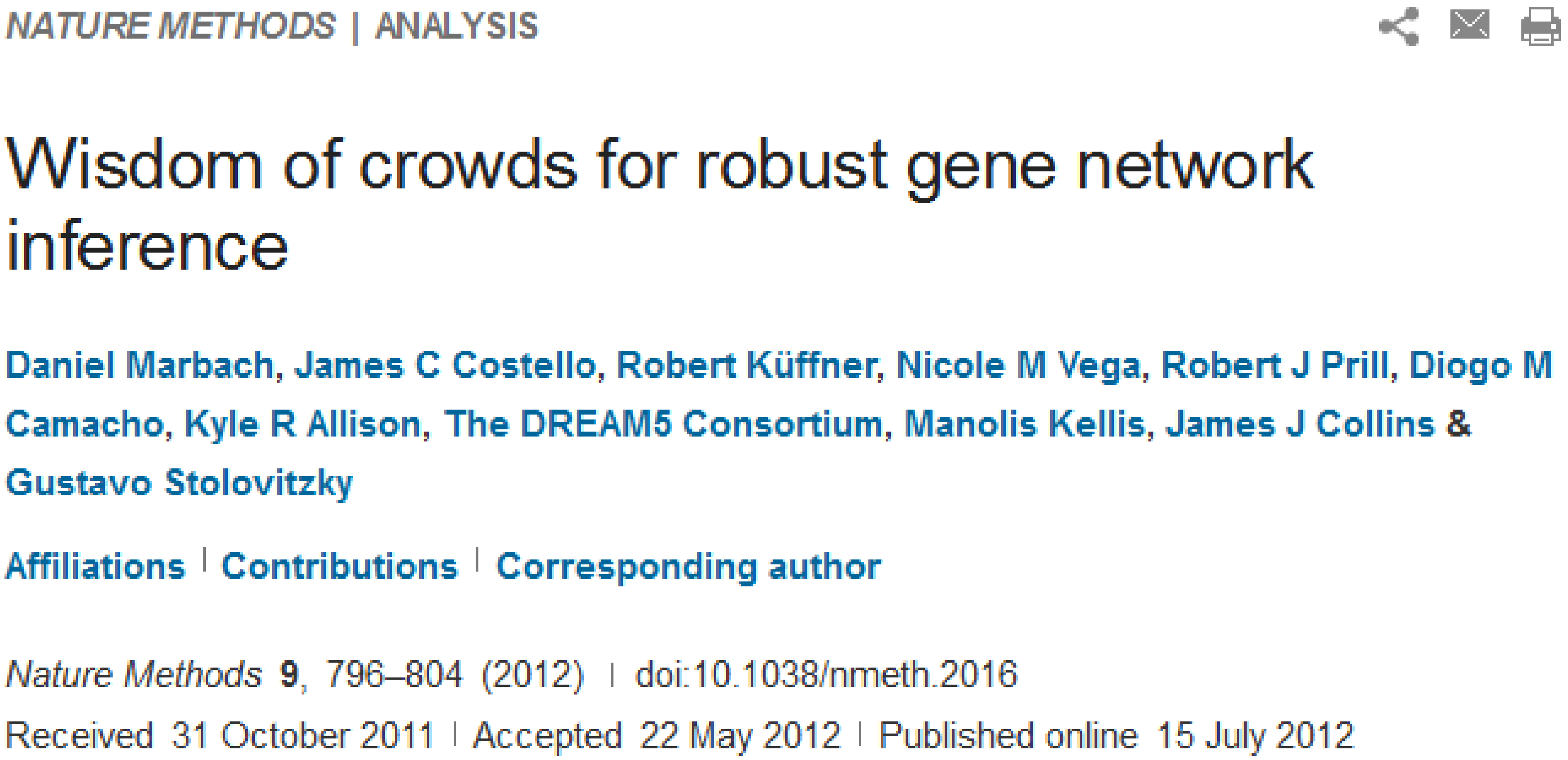


Courtesy of Elsevier B.V. Used with permission.
Source: Sumazin, Pavel, Xuerui Yang, et al. "An Extensive MicroRNA-mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma." *Cell* 147, no. 2 (2011): 370-81.

Clustering vs. "modules"

- Clusters are purely phenomenological – no claim of causality
- The term "module" is used to imply a more mechanistic connection





NATURE METHODS | ANALYSIS

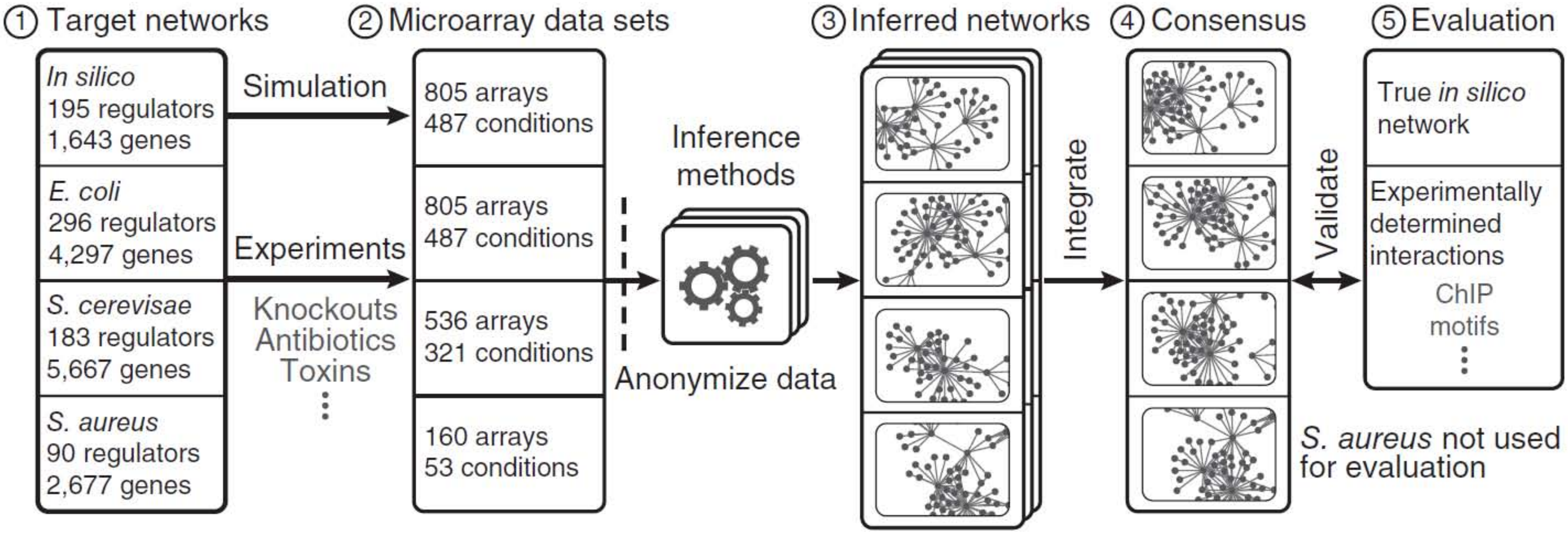
Wisdom of crowds for robust gene network inference

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

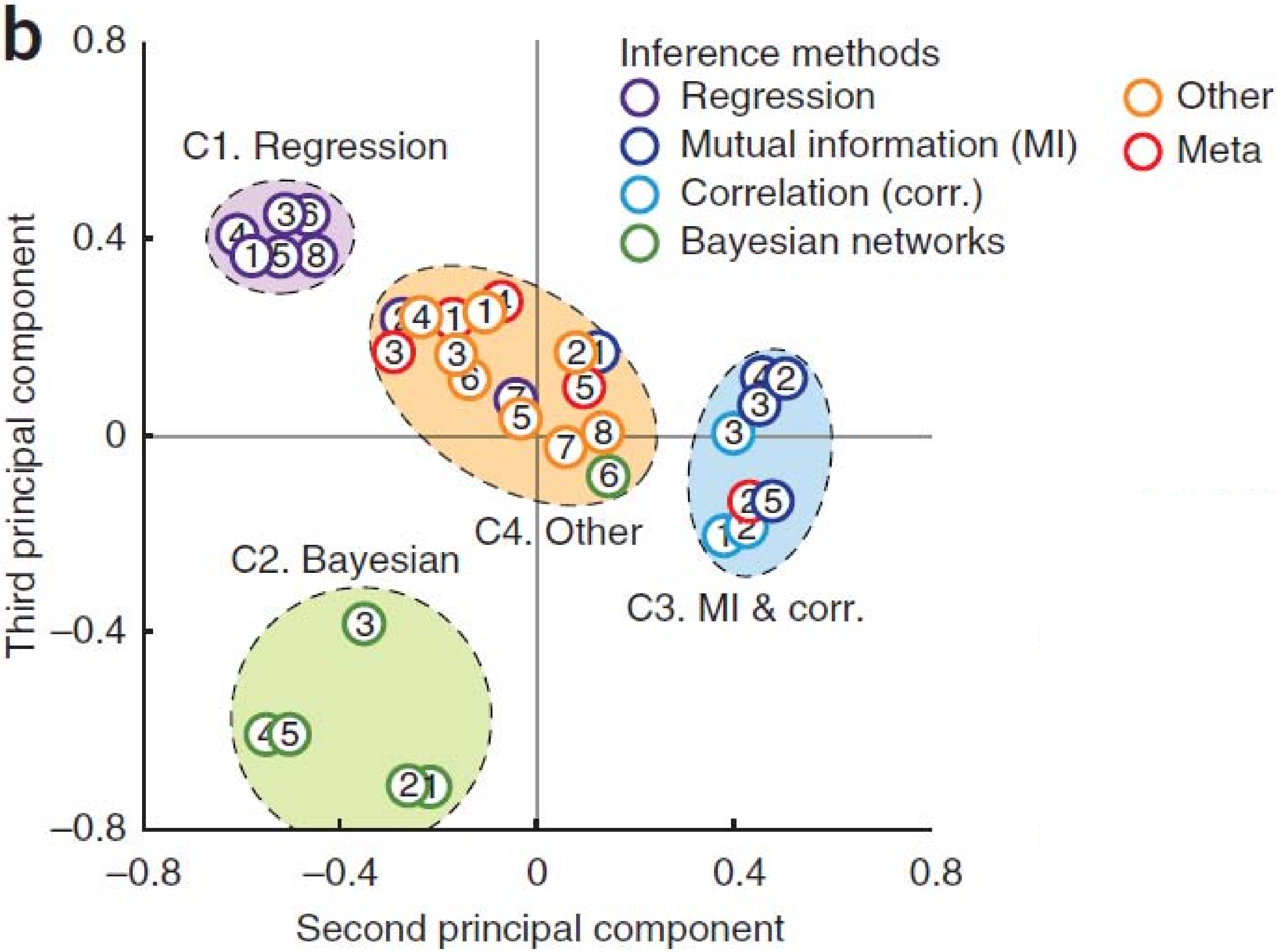
Received 31 October 2011 | Accepted 22 May 2012 | Published online 15 July 2012



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.

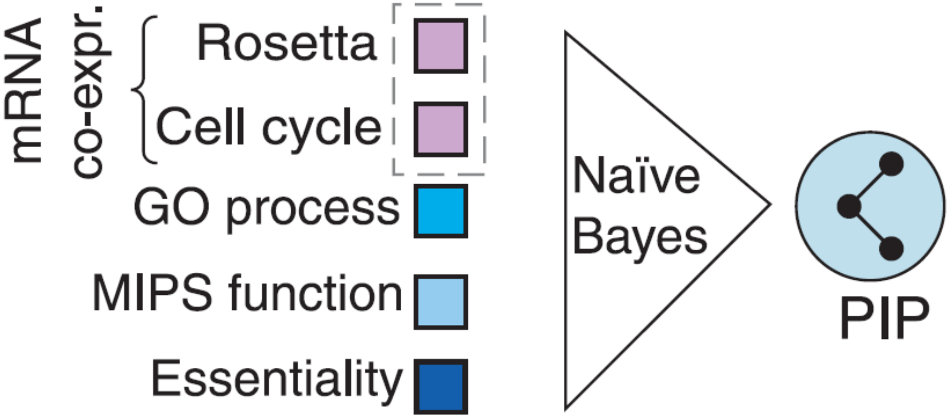
Wisdom of crowds for robust gene network inference

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

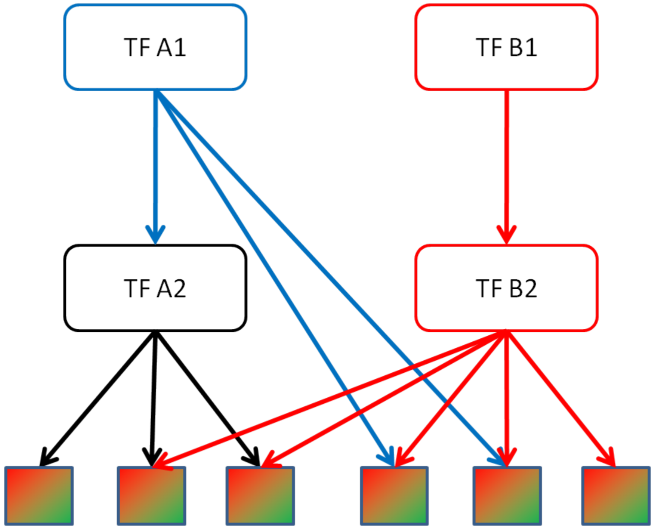
Outline

- Bayesian Networks for PPI prediction
- Gene expression
  - Distance metrics
  - Clustering
  - Signatures
  - Modules
    - Bayesian networks
    - Regression
    - Mutual Information
    - Evaluation on real and simulated data

Bayesian Networks

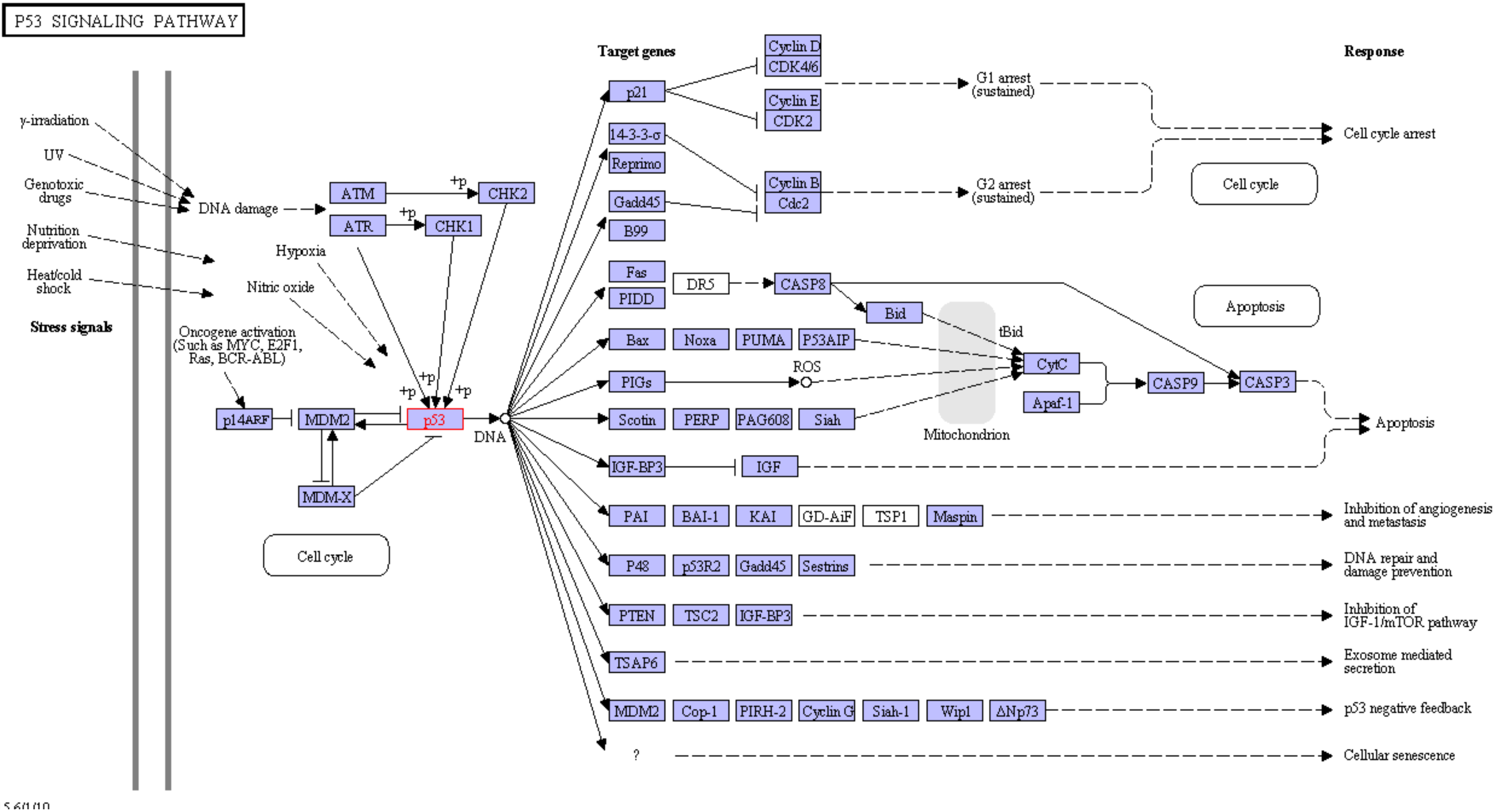


Predict unknown variables from observations



A “natural” way to think about biological networks.

Is the p53 pathway activated?



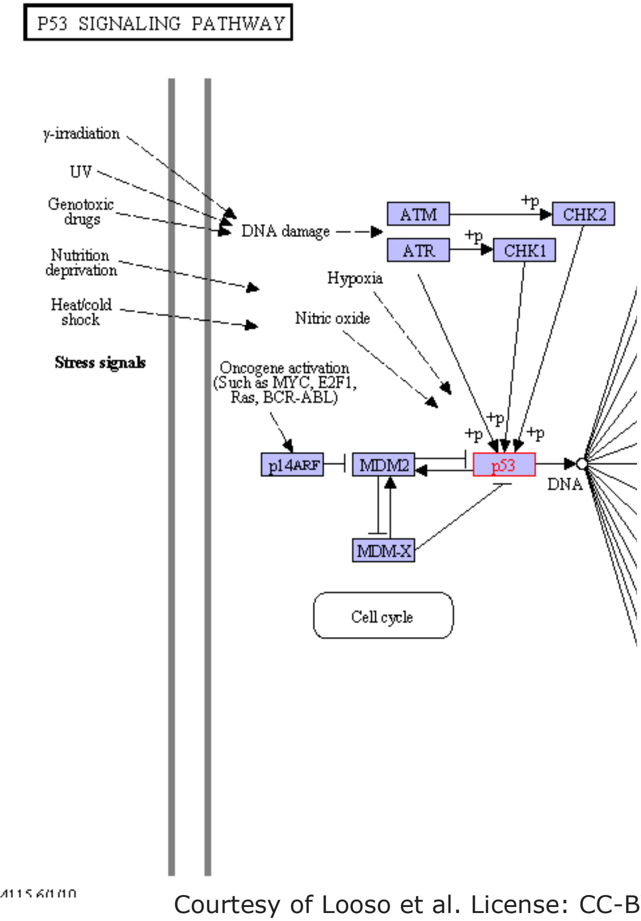
Courtesy of Looso et al. License: CC-BY.

Source: Looso, Mario, Jens Preussner, et al. "A De Novo Assembly of the Newt Transcriptome Combined with Proteomic Validation Identifies New Protein Families Expressed During Tissue Regeneration." *Genome Biology* 14, no. 2 (2013): R16.

Is the p53 pathway activated?

Possible Evidence

- Known p53 targets are up-regulated

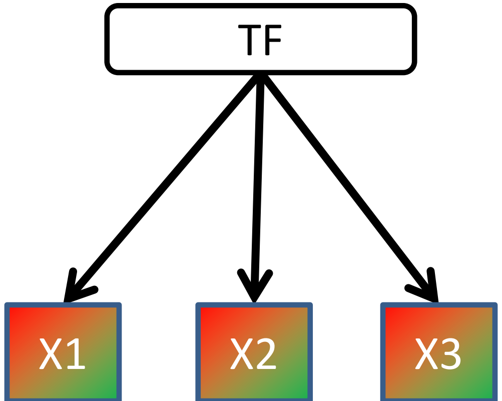


Courtesy of Looso et al. License: CC-BY.

Source: Looso, Mario, Jens Preussner, et al. "A De Novo Assembly of the Newt Transcriptome Combined with Proteomic Validation Identifies New Protein Families Expressed During Tissue Regeneration." *Genome Biology* 14, no. 2 (2013): R16.

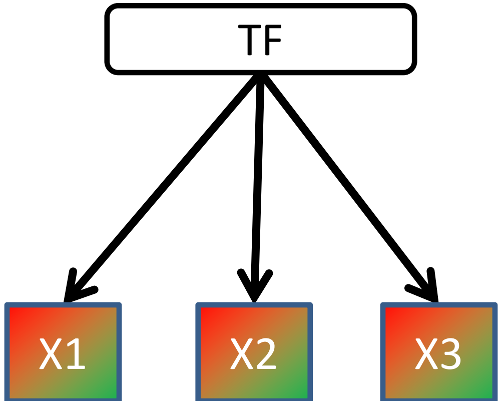
Is the p53 pathway activated?

- Formulate problem probabilistically
- Compute
  - P(p53 pathway activated| data)
- How?
  - Relatively easy to compute p(X up | TF up)
  - How?



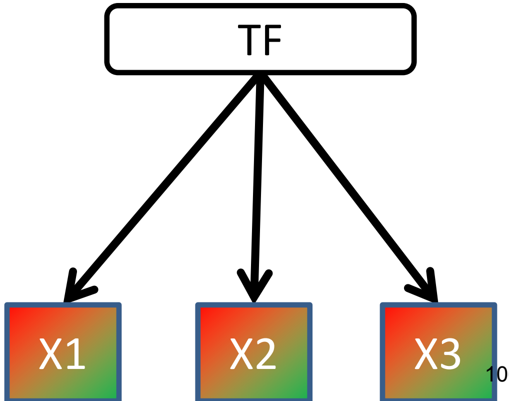
Is the p53 pathway activated?

- Formulate problem probabilistically
- Compute
  - P(p53 pathway activated| data)
- How?
  - Relatively easy to compute p(X up | TF up)
  - Look over lots of experiments and tabulate:
    - X1 up & TF up
    - X1 up & TF not up
    - X1 not up & TF not up
    - X1 not up & TF up



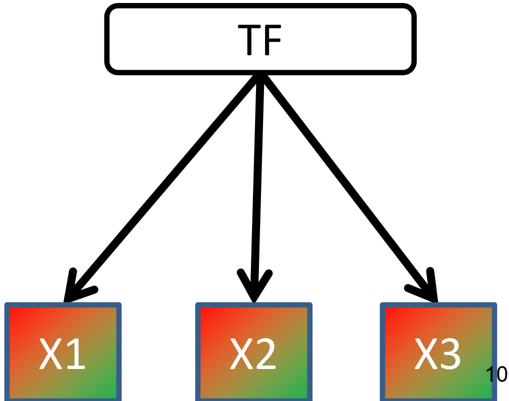
Is the p53 pathway activated?

- Formulate problem probabilistically
- Compute
  - P(p53 pathway activated| data)
- How?
  - Relatively easy to compute p(X up | TF up)
  - P(TF up|X up) = p(X up | TF up) p(TF up)/p(X up)

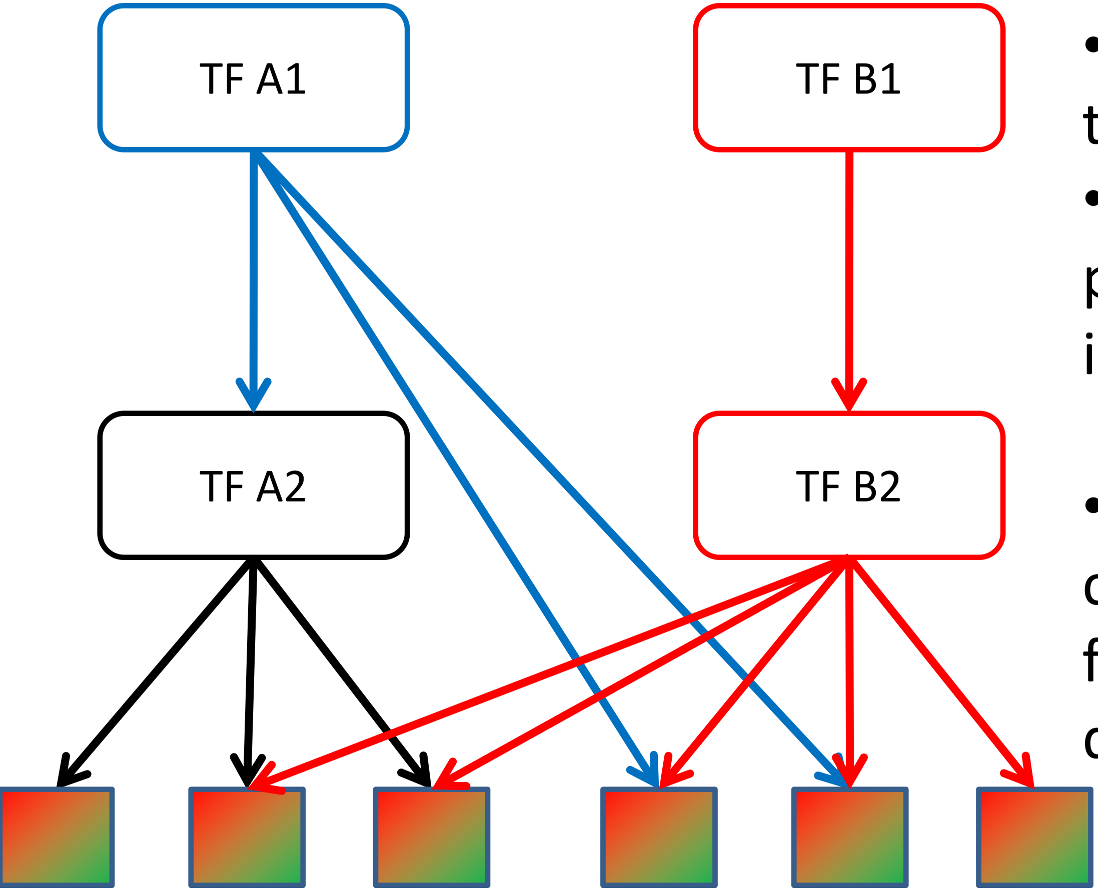


Is the p53 pathway activated?

- Formulate problem probabilistically
- Compute
  - P(p53 pathway activated| data)
- How?
  - Even with p(TF up | X up) how do we compare this explanation of the data to other possible explanations?
  - Can we include upstream data?



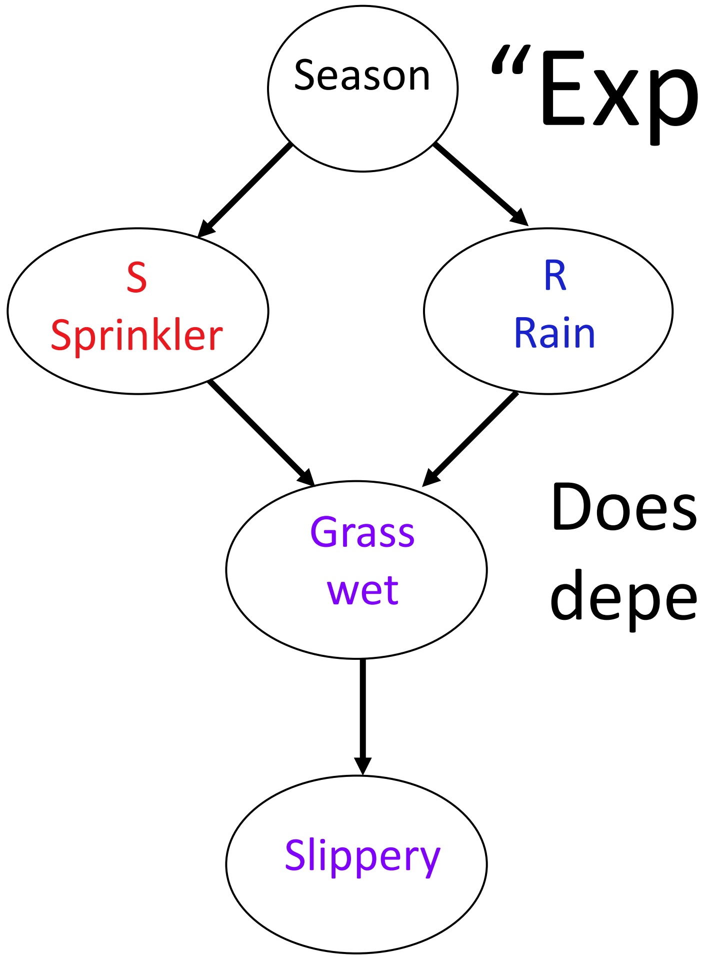
Application to Gene Networks



- Which pathway activated this set of genes?
- Either A or B or both would produce similar but not identical results.

- Bayes Nets estimate conditional probability tables from lots of gene expression data.
  - How often is TF B2 expressed when TF B1 is expressed, etc.

“Explaining Away”

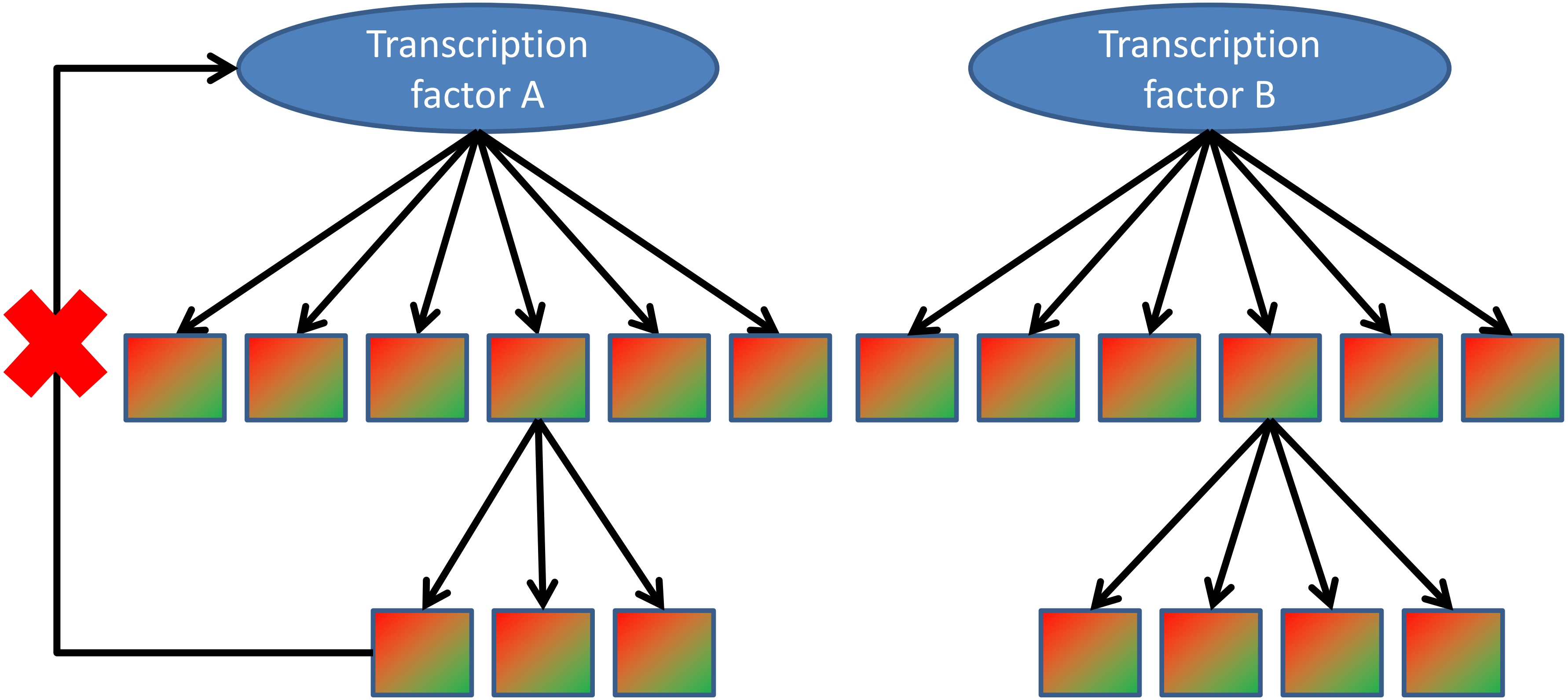


Does the probability that it’s raining depend on whether the sprinkler is on?

In a causal sense, clearly not.

But in a probabilistic model, the knowledge that it is raining influences our beliefs.

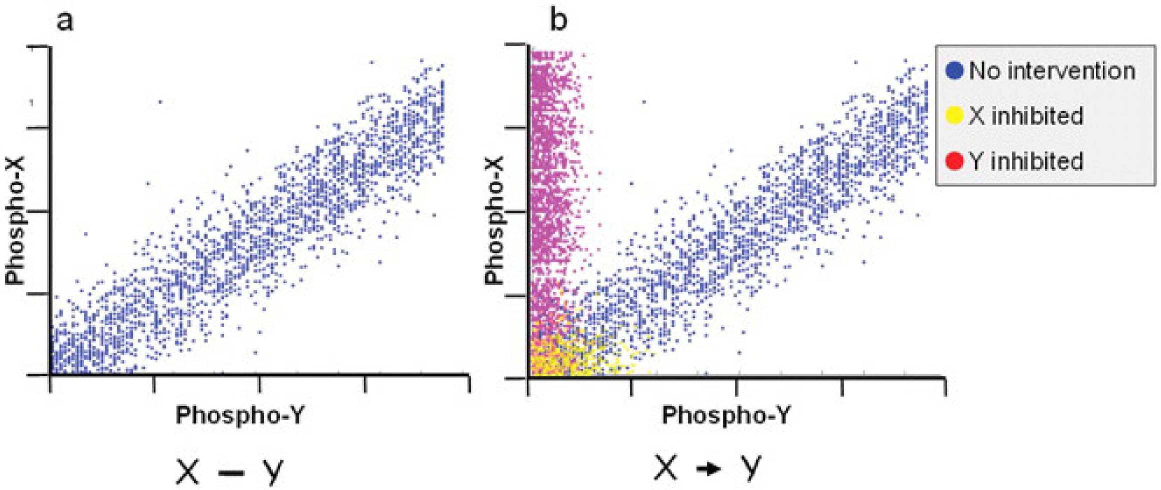
Application to Gene Networks



Multi-layer networks are possible, but feedback is not

Learning Models from Data

- Searching for the BN structure: NP-complete
  - Too many possible structures to evaluate all of them, even for very small networks.
  - Many algorithms have been proposed
  - Incorporated some prior knowledge can reduce the search space.
    - Which nodes should regulate transcription?
    - Which should cause changes in phosphorylation?
  - Intervention experiments help

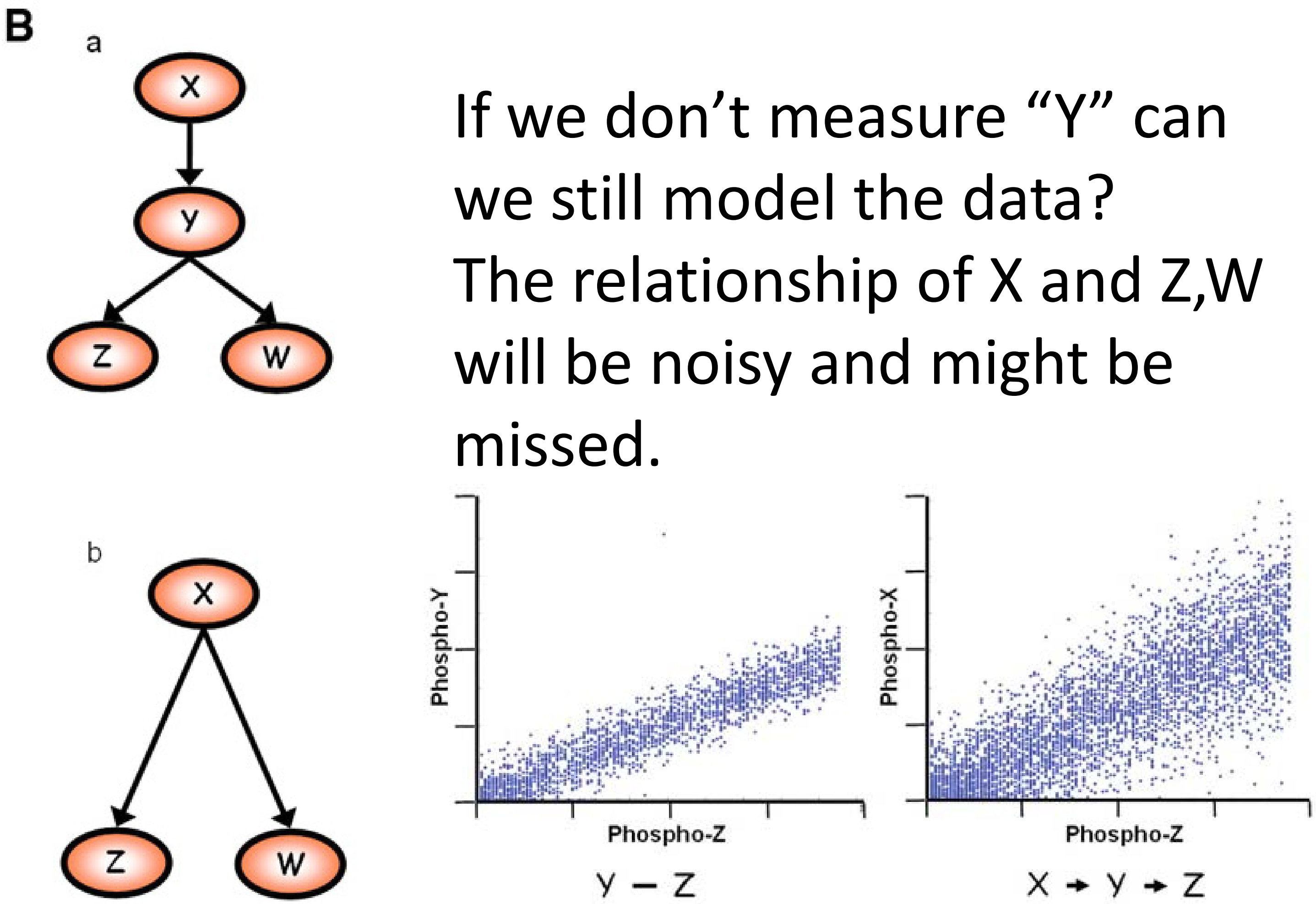


Source: Sachs, Karen, Omar Perez, et al. "Causal Protein-signaling Networks Derived From Multiparameter Single-cell Data." *Science* 308, no. 5721 (2005): 523-9.

- Without interventions, all we can say is that X and Y are correlated
- Interventions allow us to determine which is the parent.

K. Sachs et al., Science 308, 523 -529 (2005)

Fig. 1. Bayesian network modeling with single-cell data



If we don’t measure “Y” can we still model the data?
The relationship of X and Z,W will be noisy and might be missed.

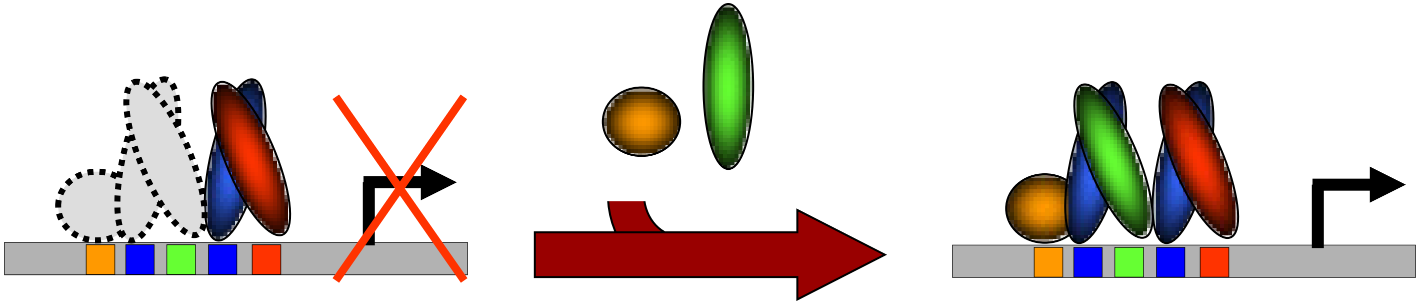
Source: Sachs, Karen, Omar Perez, et al. "Causal Protein-signaling Networks Derived From Multiparameter Single-cell Data." *Science* 308, no. 5721 (2005): 523-9.

K. Sachs et al., Science 308, 523 -529 (2005)

Outline

- Bayesian Networks for PPI prediction
- Gene expression
  - Distance metrics
  - Clustering
  - Signatures
  - Modules
    - Bayesian networks
    - Regression
    - Mutual Information
    - Evaluation on real and simulated data

Regression-based models



Predicted expression : $Y_g = f_g(X_{Tg}) + \varepsilon$

Assume that expression of gene X_g is some function of the expression of its transcription factors $X_{Tg} = \{X_t, t \in T_g\}$

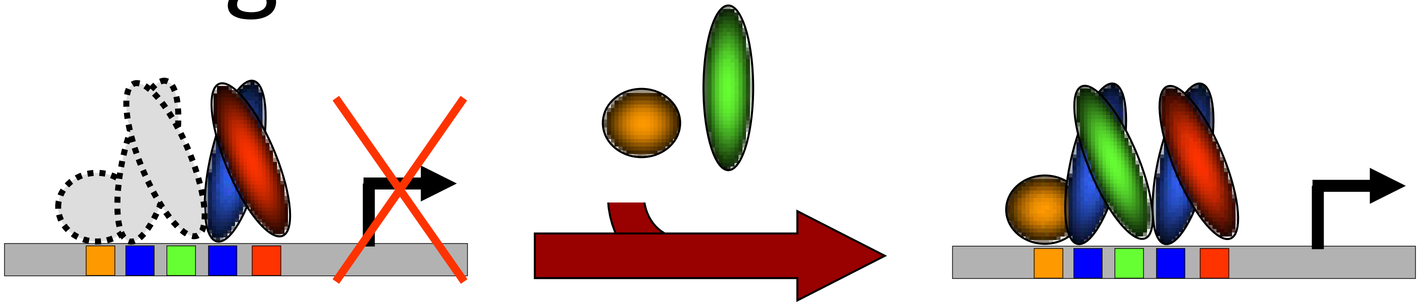
X_i = measured expression of i-th gene

X_{Ti} = measured expression of a set of TFs potentially regulating gene i

f_g is an arbitrary function

ϵ is noise

Regression-based models



$$f_g(X_{Tg}) = \sum_{t \in T_g} \beta_{t,g} X_t$$

f_g is frequently assumed to be a linear function

The values of the $\beta_{t,g}$ reflect the influence of each TF on gene *g*

How do we discover the values of the $\beta_{t,g}$?

BMC Syst Biol. 2012 Nov 22;6:145. doi: 10.1186/1752-0509-6-145.
TIGRESS: Trustful Inference of Gene REgulation using Stability Selection.

Regression-based models

$$Y_g = \sum_{t \in T_g} \beta_{t,g} X_t + \varepsilon$$

Define an objective function:

Sum over M training data sets and N genes

Find parameters that minimize “residual sum of squares” between observed (X) and predicted (Y) expression levels.

$$RSS = \sum_{j=1}^{M} \sum_{i=1}^{N} (X_{i,j} - Y_{i,j})^2$$

Regression-based models

$$Y_g = \sum_{t \in T_g} \beta_{t,g} X_t + \varepsilon \qquad RSS = \sum_{j=1}^{M} \sum_{i=1}^{N} (X_{i,j} - Y_{i,j})^2$$

<u>Problems:</u>

Standard regression will produce many very small values of β, which makes interpretation difficult

β values can be unstable to changes in training data

Solutions:

Subset Selection and Coefficient Shrinkage

- see Section 3.4 of Hastie Tibshirani and Friedman "The elements of statistical learning" for general approaches and "TIGRESS: Trustful Inference of Gene REgulation using Stability Selection" for a successful DREAM challenge doi: 10.1186/1752-0509-6-145.

Outline

- Bayesian Networks for PPI prediction
- Gene expression
  - Distance metrics
  - Clustering
  - Signatures
  - Modules
    - Bayesian networks
    - Regression
    - Mutual Information
    - Evaluation on real and simulated data

Quick Review of Information Theory

Information content of an event E

$$I(E) = \log_2 \frac{1}{P(E)}$$

Rare letters have higher information content

Quick Review of Information Theory

Information content of an event E

$$I\left(E\right)=\log_2\frac{1}{P\left(E\right)}$$

Entropy is evaluated over all possible outcomes

$$H\left(S\right)=\sum_i p_i I(s_i)=\sum_i p_i \log_2\frac{1}{p_i}$$

$$H(f)=-\int f(x)\ln f(x)dx.$$

Mutual Information

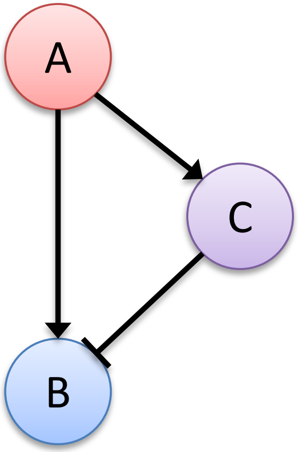
- Does knowing variable X reduce the uncertainty in variable Y?
- Example:
  - P(Rain) depends on P(Clouds)
  - P(target expressed) depends on P(TF expressed)

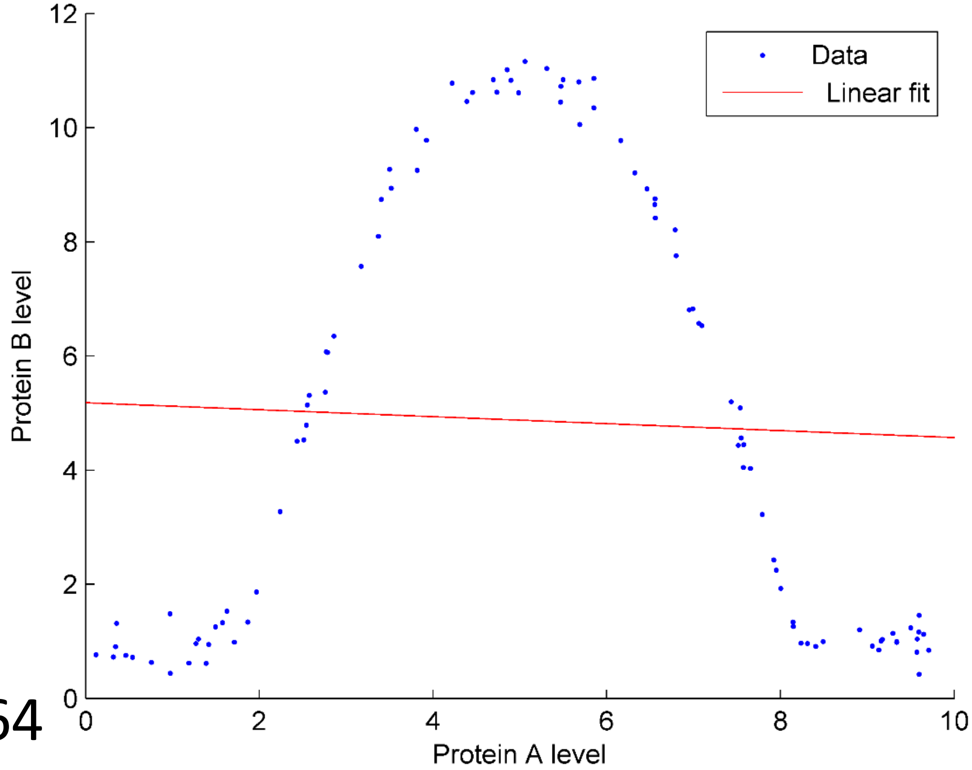
$$I\left(x,y\right)=H(x)+H(y)-H(x,y)$$

- I(x,y) = 0 means variables are independent
- Reveals non-linear relationships that are missed by correlation.

Mutual information detects non-linear relationships

Incoherent feed-forward loop (FFL)





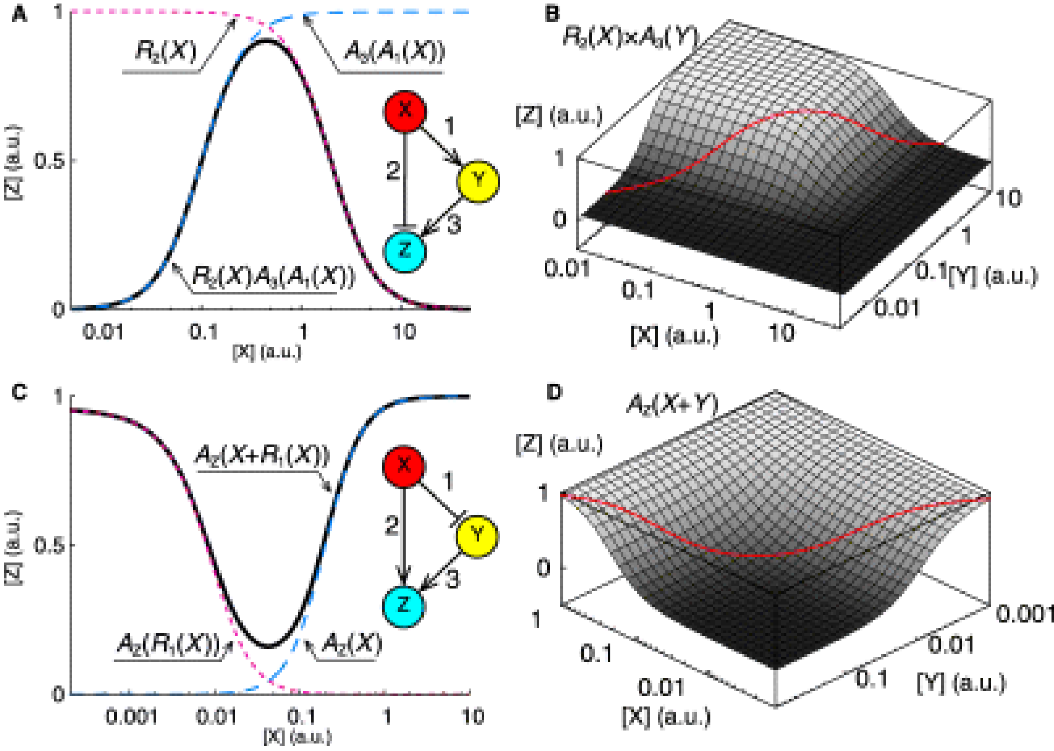
Mutual information = 1.7343

Correlation coefficient = -0.0464

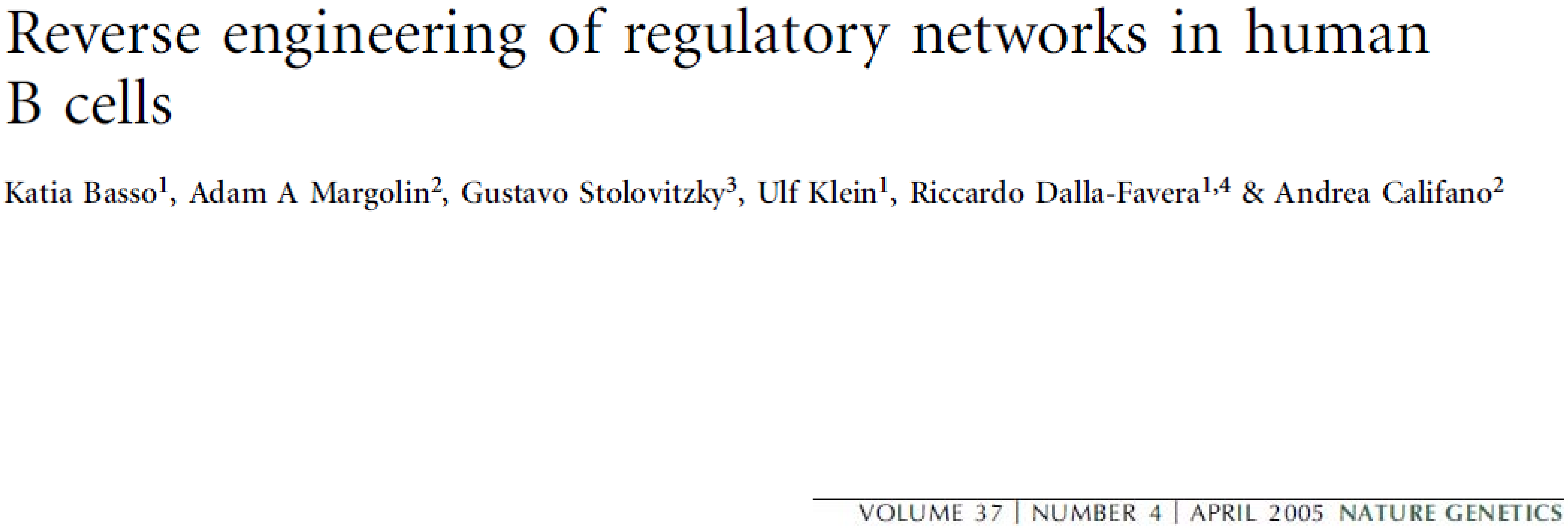
No correlation, but knowing A reduces the uncertainty in the distribution of B

Mutual information detects non-linear relationships

- Complex regulatory network structure => complex relationships between protein levels
- Example: incoherent feed-forward loop (FFL)



ARACNe



Reverse engineering of regulatory networks in human B cells

Katia Basso[1], Adam A Margolin[2], Gustavo Stolovitzky[3], Ulf Klein[1], Riccardo Dalla-Favera[1,4] & Andrea Califano[2]

VOLUME 37 | NUMBER 4 | APRIL 2005 NATURE GENETICS

ARACNe

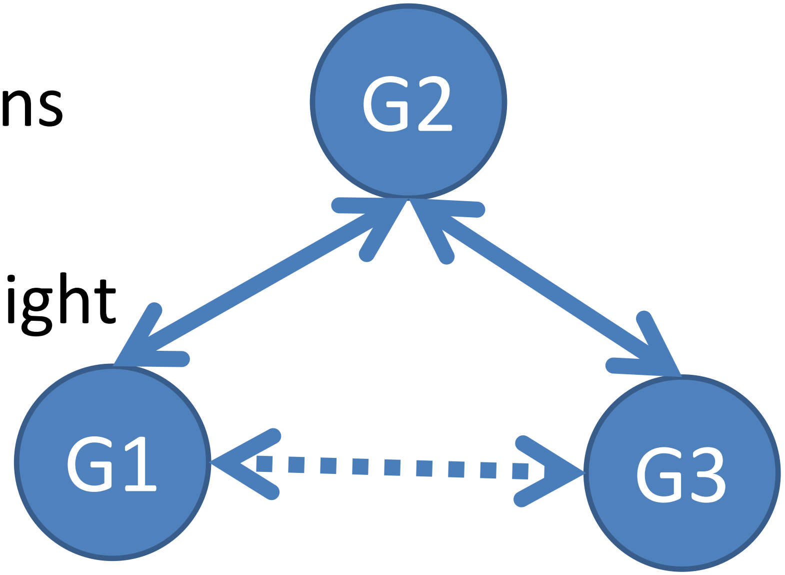
- Find TF-target relationships using mutual information

$$H(f) = -\int f(x) \ln f(x) dx.$$

- How do you recognize a significant value of MI?
  - – randomly shuffle expression data
  - – compute distribution of Mutual information

ARACNE

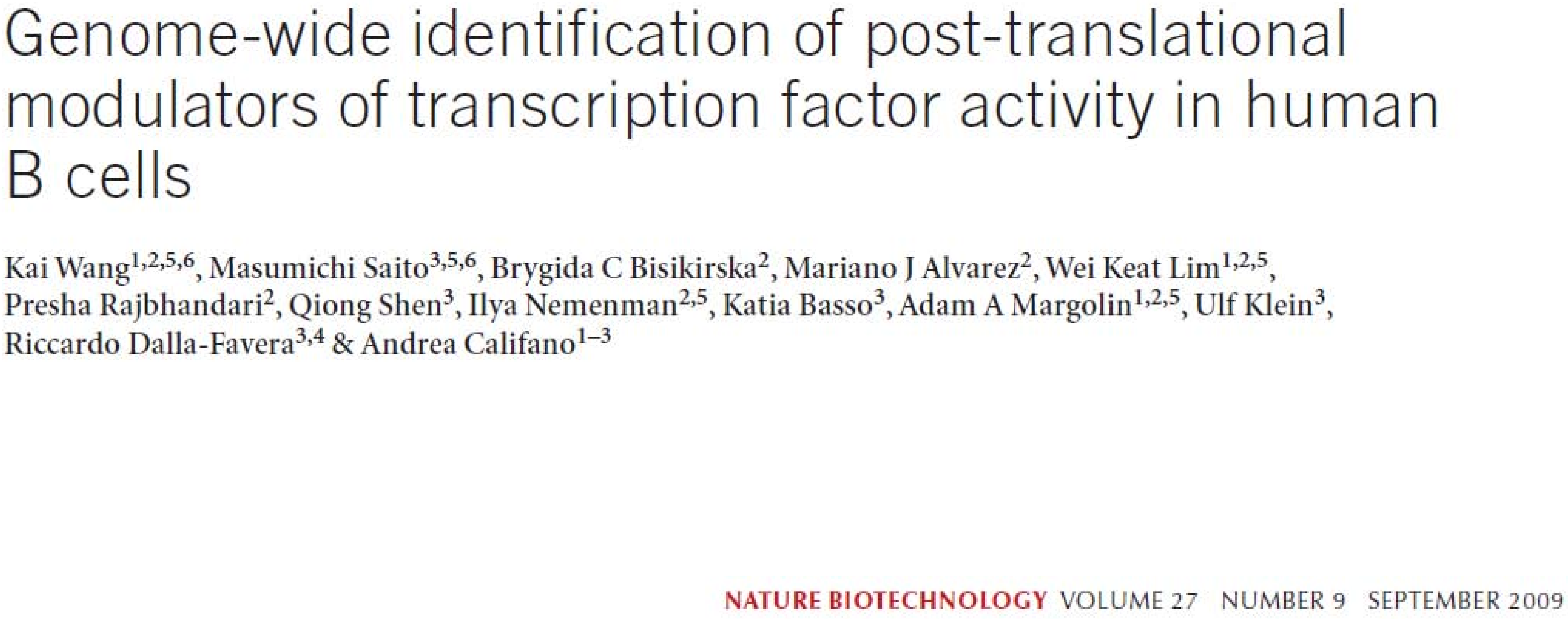
- Data processing inequality
  - Eliminate indirect interactions
  - If G2 regulates G1,G3 I(G1,G3)>0 but adds no insight
  - Remove edge with smallest mutual information in each triple



$$I(g_1, g_3) \leq \min\left[I(g_1, g_2); I(g_2, g_3)\right]$$

MINDy

- Identify proteins that modulate TF function
  - Other TFs



Genome-wide identification of post-translational modulators of transcription factor activity in human B cells

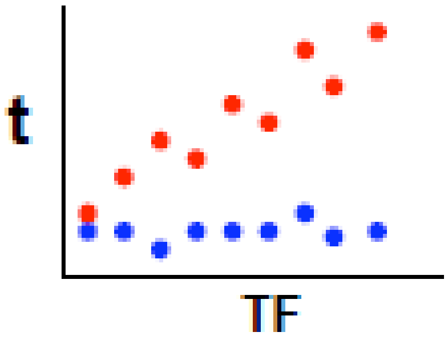
Kai Wang[1,2,5,6], Masumichi Saito[3,5,6], Brygida C Bisikirska[2], Mariano J Alvarez[2], Wei Keat Lim[1,2,5], Presha Rajbhandari[2], Qiong Shen[3], Ilya Nemenman[2,5], Katia Basso[3], Adam A Margolin[1,2,5], Ulf Klein[3], Riccardo Dalla-Favera[3,4] & Andrea Califano[1–3]

NATURE BIOTECHNOLOGY VOLUME 27 NUMBER 9 SEPTEMBER 2009

Model

- Assumes that expression of target T is determined by TF and modulator (M)

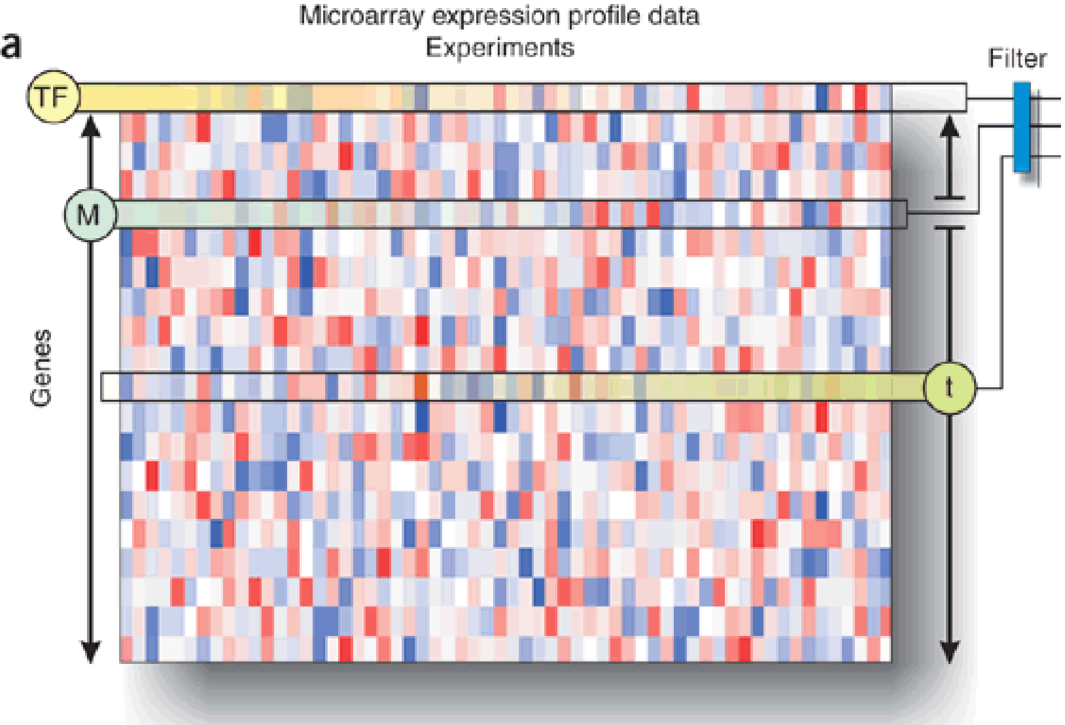
$$[T] = C \cdot [TF]^{h} \cdot [M]^{g}$$



Modulator present at highest levels

Modulator present at lowest levels

-> Suggests M is an activator

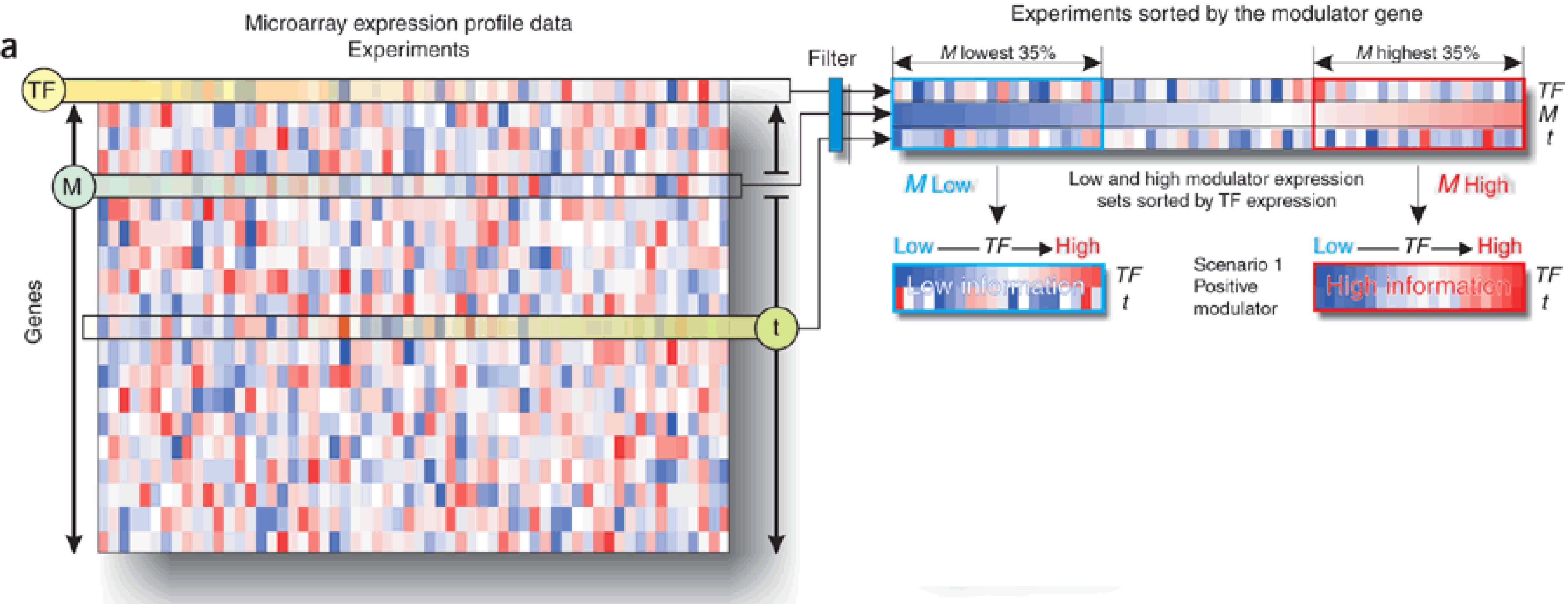


Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Wang, Kai, Masumichi Saito, et al. "Genome-wide Identification of Post-translational Modulators of Transcription Factor Activity in Human B cells." *Nature Biotechnology* 27, no. 9 (2009): 829-37.

Filters

1. expression of the modulator and of the TF must be statistically independent
2. the modulator expression must have sufficient range
3. may be filtered by additional criteria—for example, molecular functions.

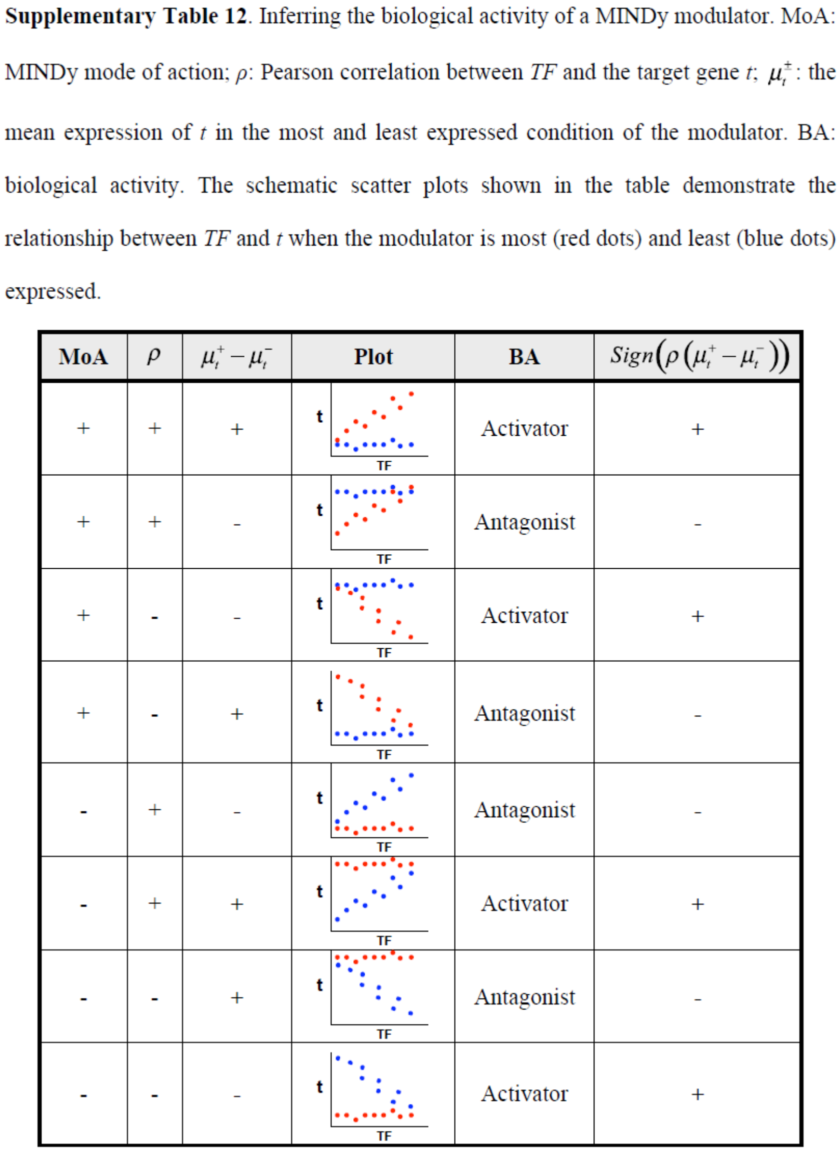
Genome-wide identification of post-translational modulators of transcription factor activity in human B cells
Kai Wang, Masumichi Saito, Brygida C Bisikirska, Mariano J Alvarez, Wei Keat Lim, Presha Rajbhandari, Qiong Shen, Ilya Nemenman, Katia Basso, Adam A Margolin, Ulf Klein, Riccardo Dalla-Favera & Andrea Califano
Nature Biotechnology **27**, 829 - 837 (2009) Published online: 9 September 2009
doi:10.1038/nbt.1563



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Wang, Kai, Masumichi Saito, et al. "Genome-wide Identification of Post-translational Modulators of Transcription Factor Activity in Human B cells." *Nature Biotechnology* 27, no. 9 (2009): 829-37.

Estimate conditional mutual information

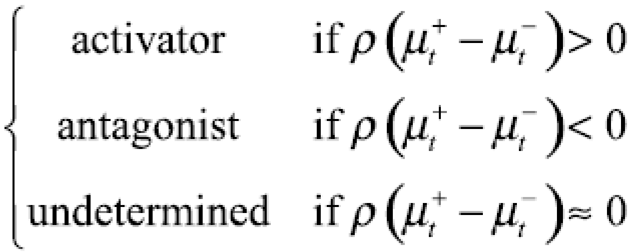
Genome-wide identification of post-translational modulators of transcription factor activity in human B cells
Kai Wang, Masumichi Saito, Brygida C Bisikirska, Mariano J Alvarez, Wei Keat Lim, Presha Rajbhandari, Qiong Shen, Ilya Nemenman, Katia Basso, Adam A Margolin, Ulf Klein, Riccardo Dalla-Favera & Andrea Califano
Nature Biotechnology **27**, 829 - 837 (2009) Published online: 9 September 2009
doi:10.1038/nbt.1563



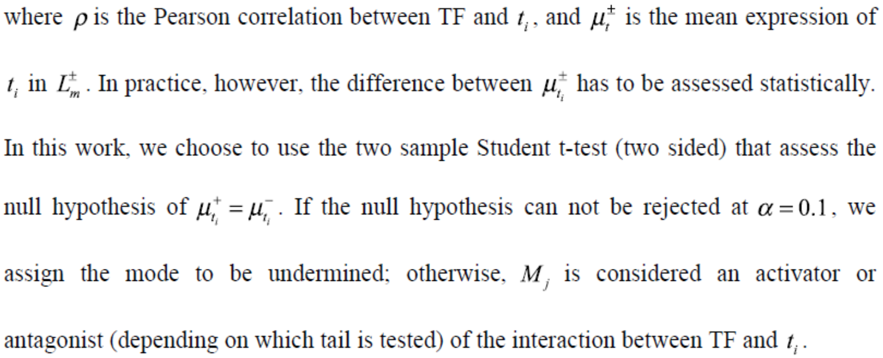
Supplementary Table 12. Inferring the biological activity of a MINDy modulator. MoA: MINDy mode of action; ρ: Pearson correlation between *TF* and the target gene *t*; $\mu_t^{\pm}$: the mean expression of *t* in the most and least expressed condition of the modulator. BA: biological activity. The schematic scatter plots shown in the table demonstrate the relationship between *TF* and *t* when the modulator is most (red dots) and least (blue dots) expressed.

MoA	ρ	$\mu_t^+ - \mu_t^-$	Plot	BA	$Sign(\rho(\mu_t^+ - \mu_t^-))$
+	+	+		Activator	+
+	+	-		Antagonist	-
+	-	-		Activator	+
+	-	+		Antagonist	-
-	+	-		Antagonist	-
-	+	+		Activator	+
-	-	+		Antagonist	-
-	-	-		Activator	+

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Wang, Kai, Masumichi Saito, et al. "Genome-wide Identification of Post-translational Modulators of Transcription Factor Activity in Human B cells." *Nature Biotechnology* 27, no. 9 (2009): 829-37.



$$\begin{cases} \text{activator} & \text{if } \rho\left(\mu_t^+ - \mu_t^-\right) > 0 \\ \text{antagonist} & \text{if } \rho\left(\mu_t^+ - \mu_t^-\right) < 0 \\ \text{undetermined} & \text{if } \rho\left(\mu_t^+ - \mu_t^-\right) \approx 0 \end{cases}$$



where ρ is the Pearson correlation between TF and t_i, and $\mu_t^{\pm}$ is the mean expression of t_i in $L_m^{\pm}$. In practice, however, the difference between $\mu_{t_i}^{\pm}$ has to be assessed statistically. In this work, we choose to use the two sample Student t-test (two sided) that assess the null hypothesis of $\mu_{t_i}^{+} = \mu_{t_i}^{-}$. If the null hypothesis can not be rejected at $\alpha = 0.1$, we assign the mode to be undermined; otherwise, M_j is considered an activator or antagonist (depending on which tail is tested) of the interaction between TF and t_i.

Note than none of these curve saturate

What regulates MYC?

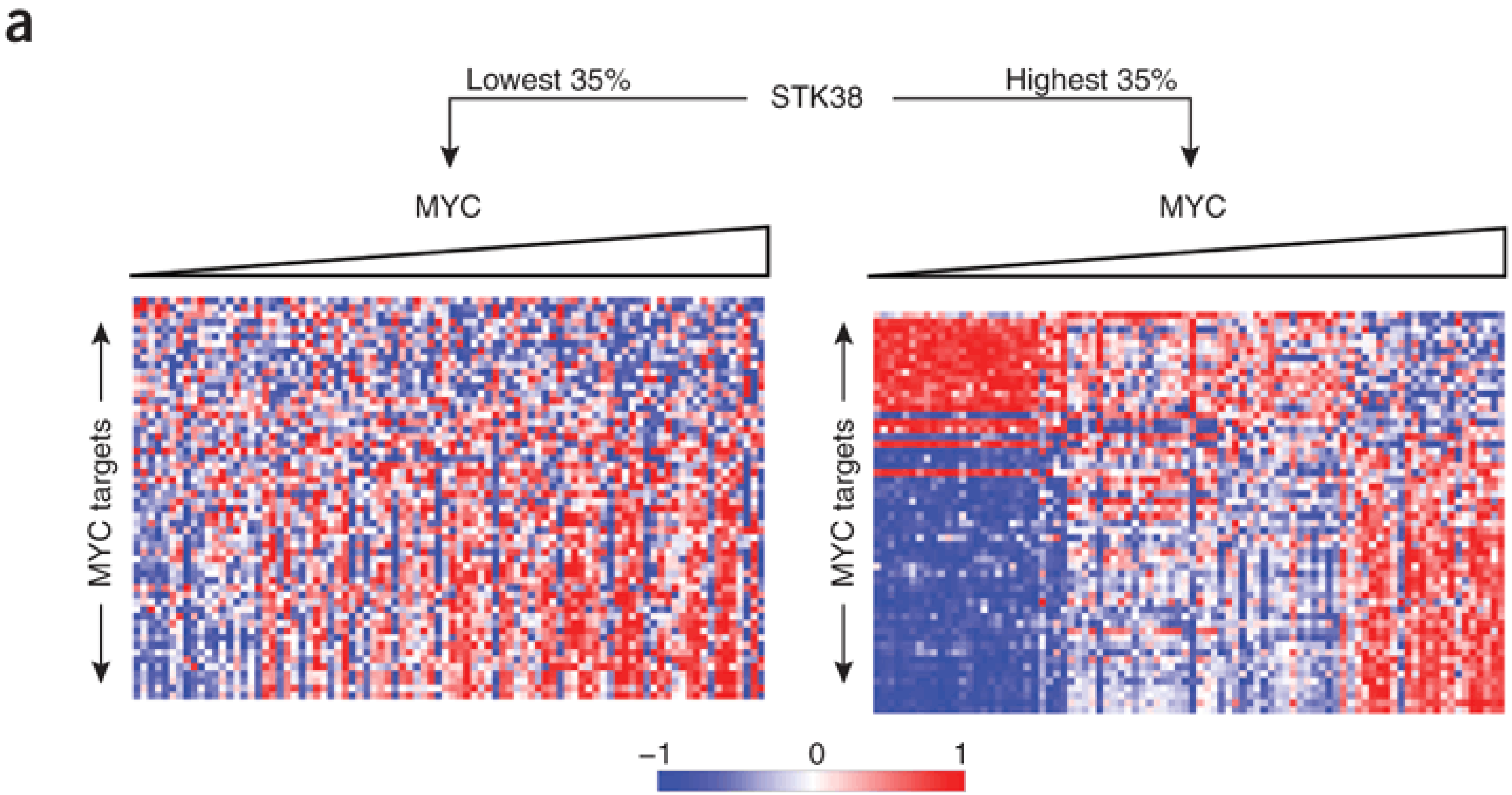
Input:

254 expression profiles in B cells
(normal and tumor)
various sets of candidate regulators

Evaluation:

1. comparison to known modulators
2. experimental tests of four candidates

What regulates MYC?

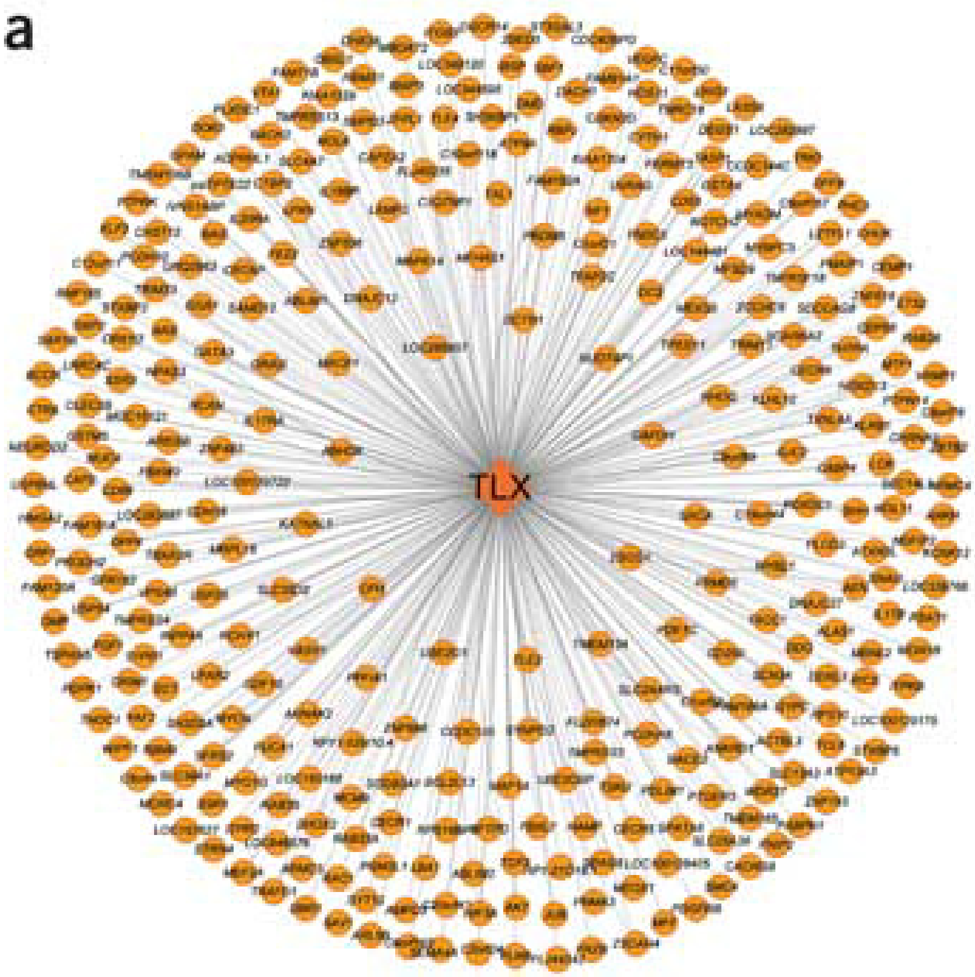


Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Wang, Kai, Masumichi Saito, et al. "Genome-wide Identification of Post-translational Modulators of Transcription Factor Activity in Human B cells." *Nature Biotechnology* 27, no. 9 (2009): 829-37.

Limitations

- Need huge expression datasets
- Can’t find:
  - modulator that do not change in expression
  - modulator that are highly correlated with target
  - modulators that both activate and repress

Huge networks!

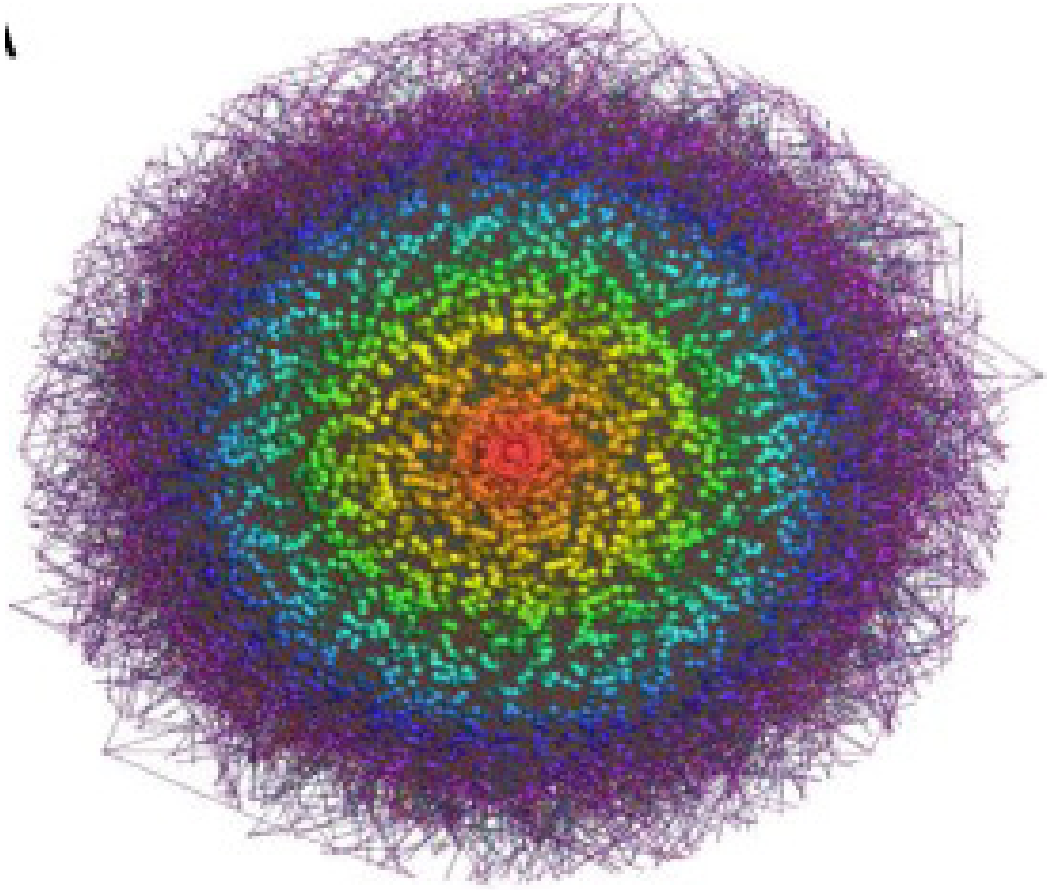


This is just the nearest neighbors of one node of interest from ARACNe!

Nature Medicine 18, 436–440 (2012) doi:10.1038/nm.2610

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Della Gatta, Giusy, Teresa Palomero, et al. "Reverse Engineering of TLX Oncogenic Transcriptional Networks Identifies RUNX1 as Tumor Suppressor in T-ALL." *Nature Medicine* 18, no. 3 (2012): 436-40.

Huge networks!



Conditional MI network of miR modulators 248,000 interactions

http://www.sciencedirect.com/science/article/pii/S0092867411011524

Courtesy of Elsevier B.V. Used with permission.
Source: Sumazin, Pavel, Xuerui Yang, et al. "An Extensive MicroRNA-mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma." *Cell* 147, no. 2 (2011): 370-81.

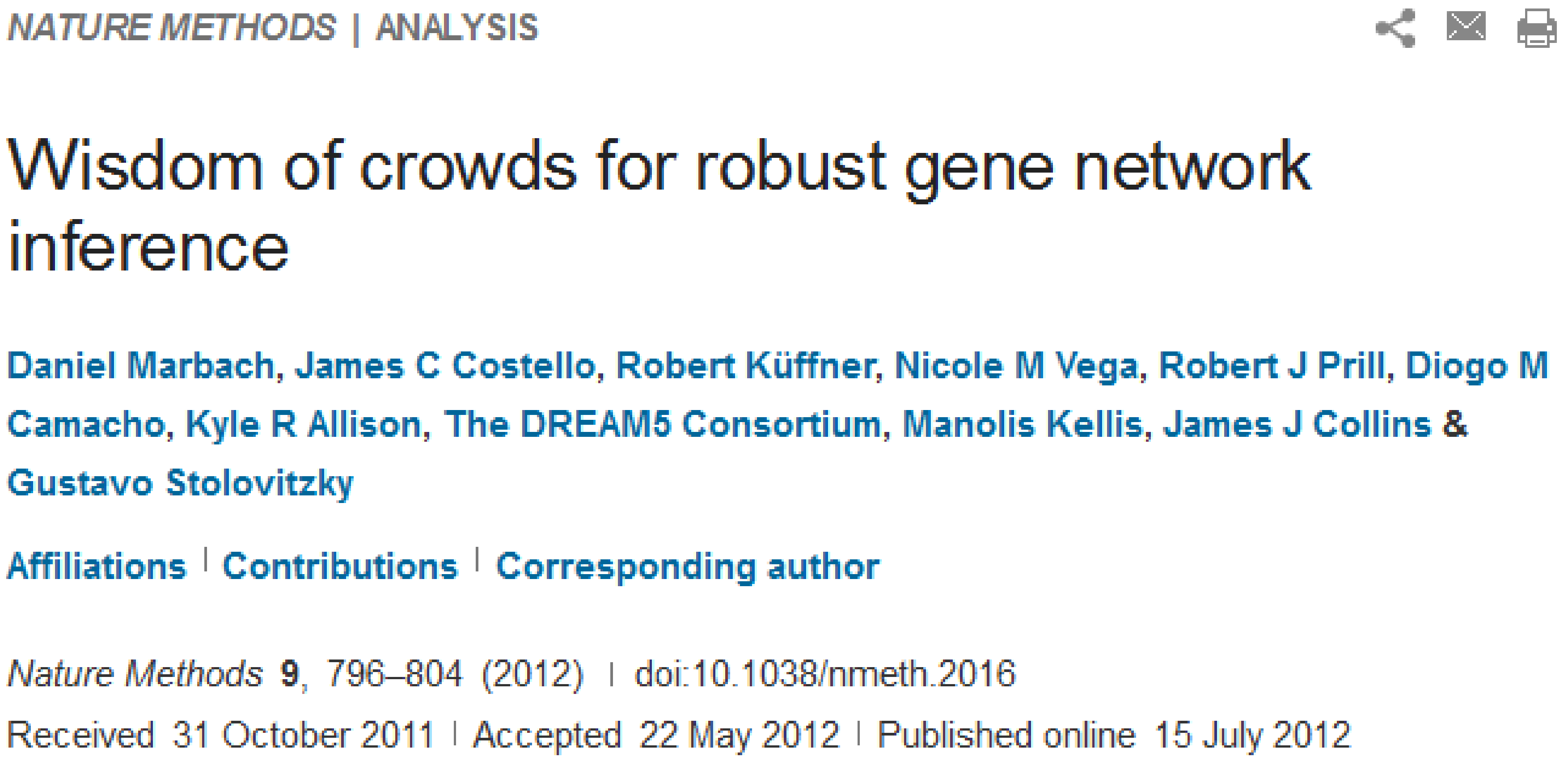
MINDy modulators

	Potential Modulators		
Source of targets	Signaling (542)	TFs (598)	Any (3,131)
Database	91	99	
ARACNe	80	85	
ALL	[25/296]	[32/296]	296

MINDy selects between 10-20% of candidates!

Outline

- Bayesian Networks for PPI prediction
- Gene expression
  - Distance metrics
  - Clustering
  - Signatures
  - Modules
    - Bayesian networks
    - Regression
    - Mutual Information
    - Evaluation on real and simulated data



NATURE METHODS | ANALYSIS

Wisdom of crowds for robust gene network inference

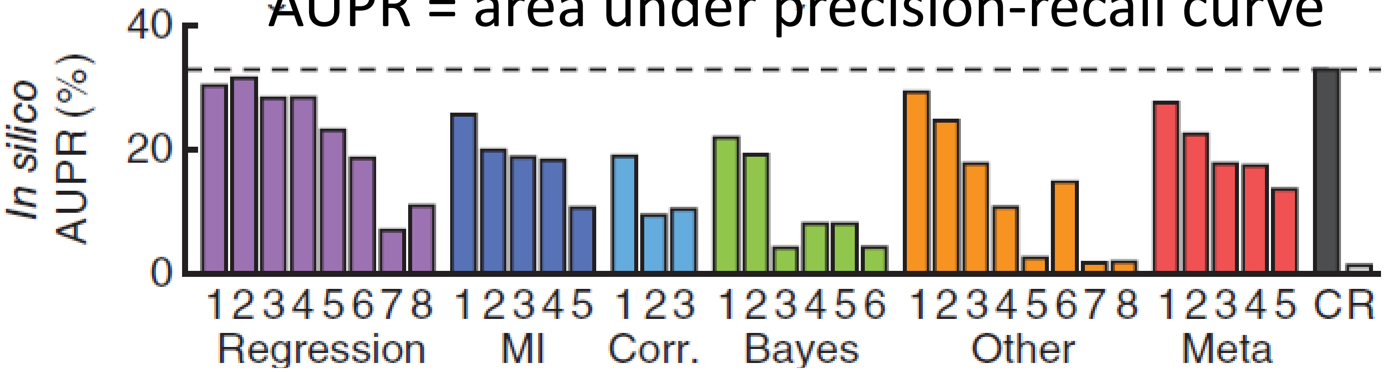
Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, The DREAM5 Consortium, Manolis Kellis, James J Collins & Gustavo Stolovitzky

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

AUPR = area under precision-recall curve



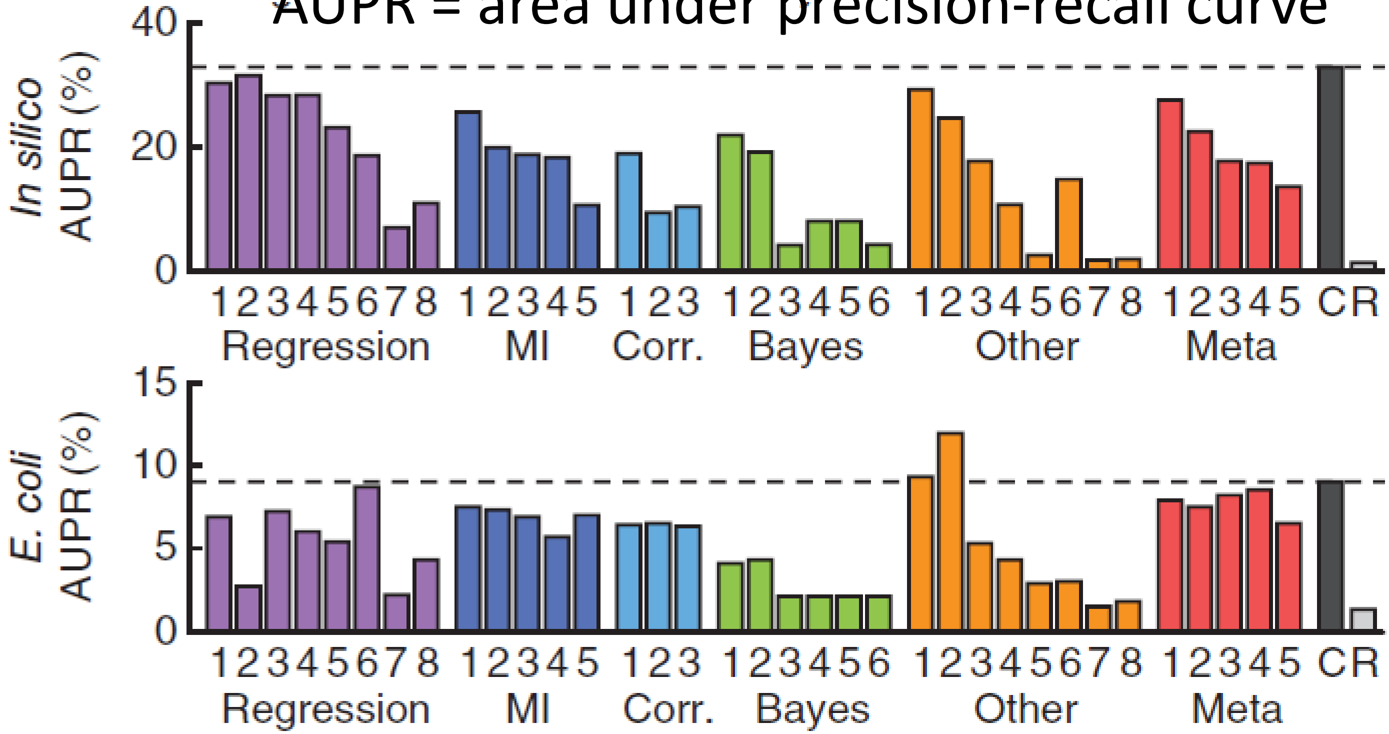
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

Area under precision-recall curve

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

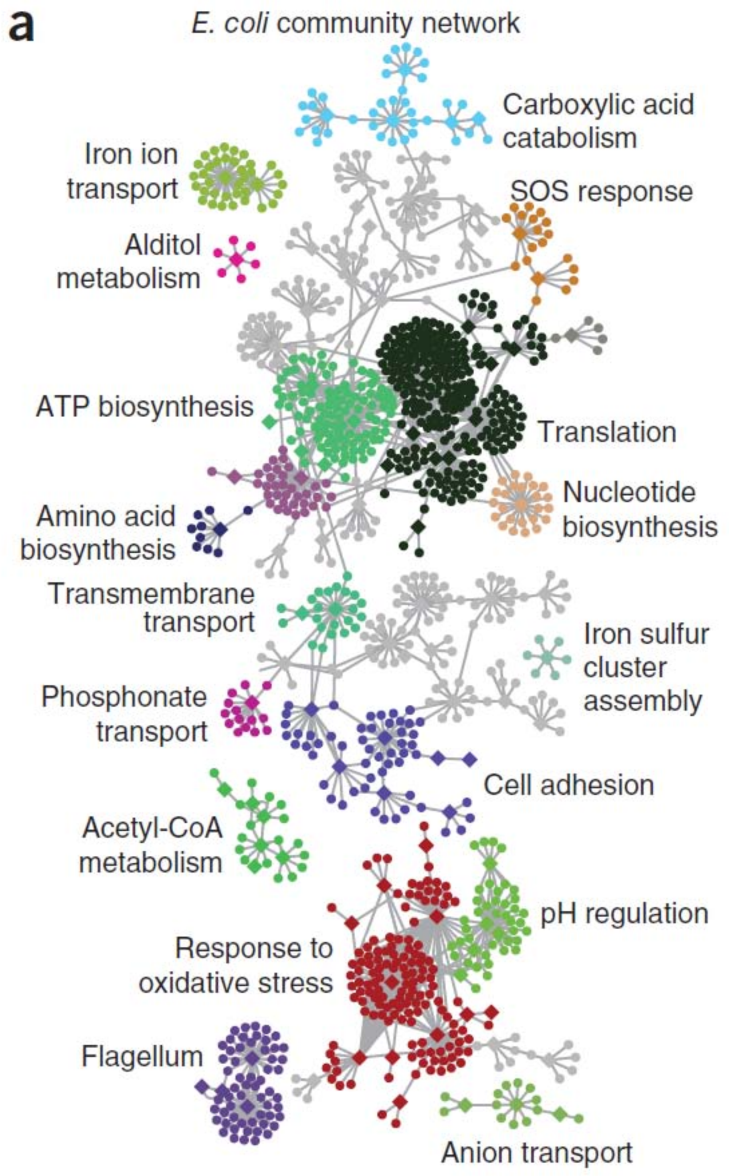
AUPR = area under precision-recall curve

Area under precision-recall curve



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.

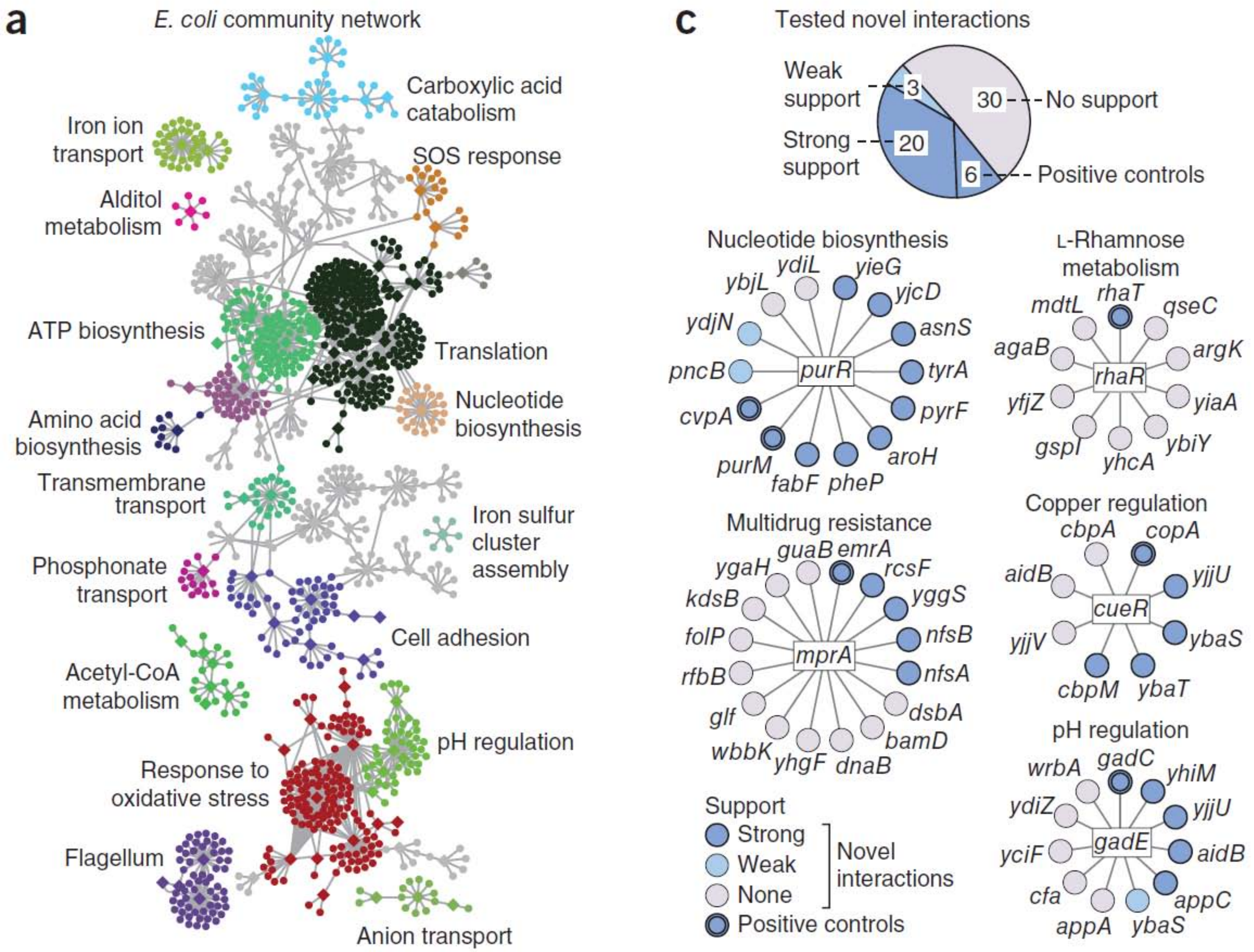
Wisdom of crowds for robust gene network inference
Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016



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



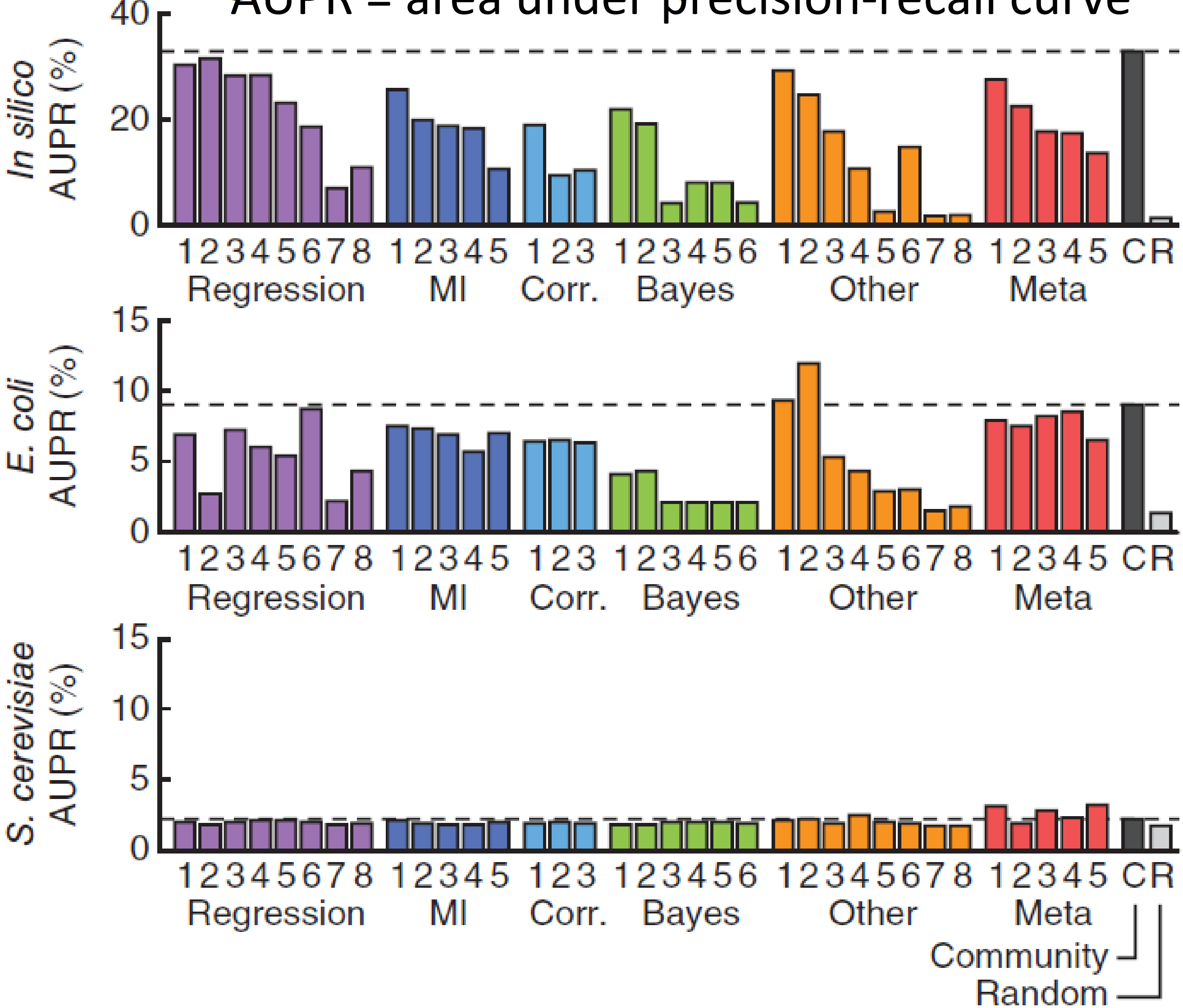
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Marbach, Daniel, James C. Costello, et al. "Wisdom of Crowds for Robust Gene Network Inference." *Nature Methods* 9, no. 8 (2012): 796-804.

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

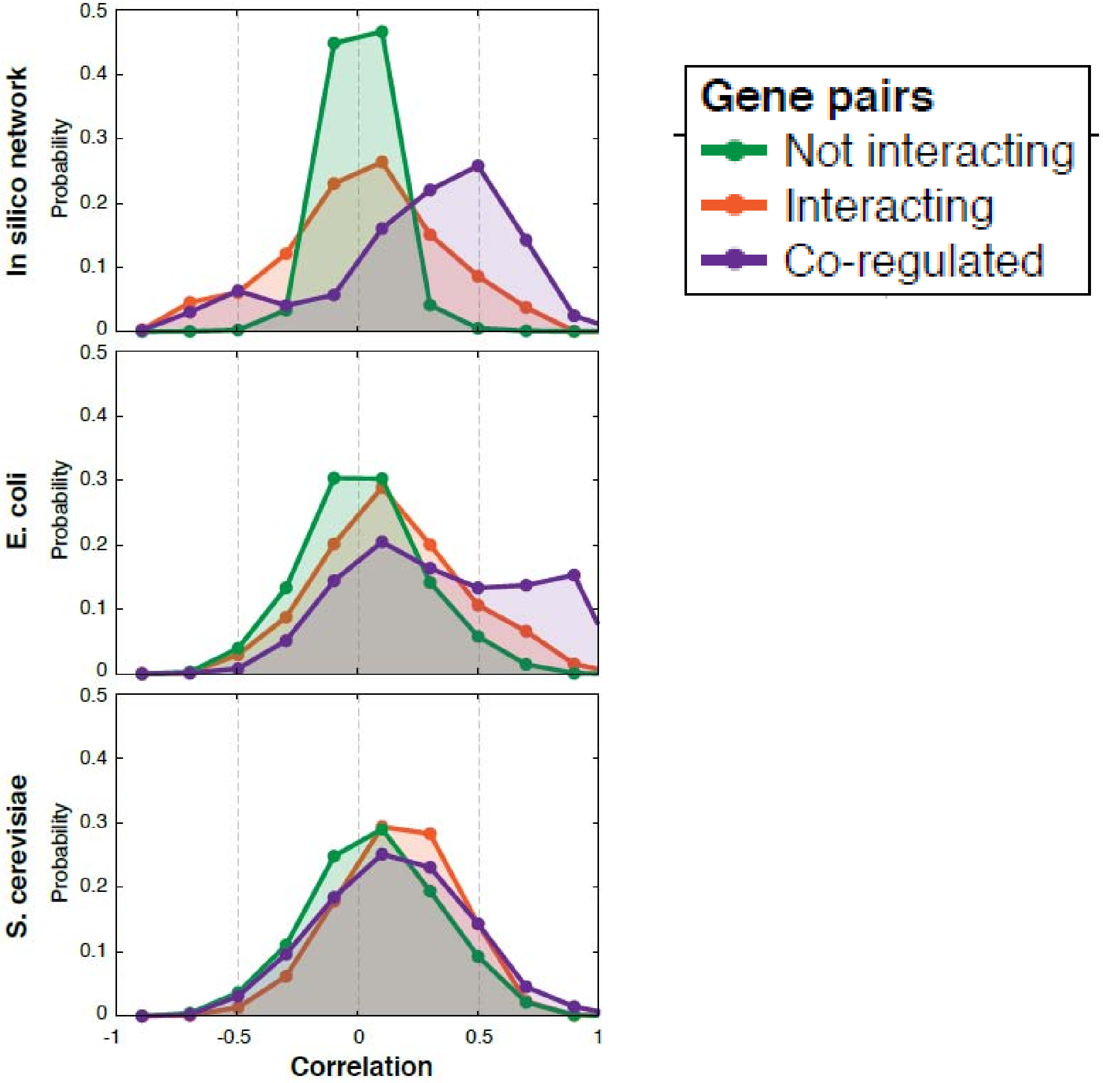
Area under precision-recall curve



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

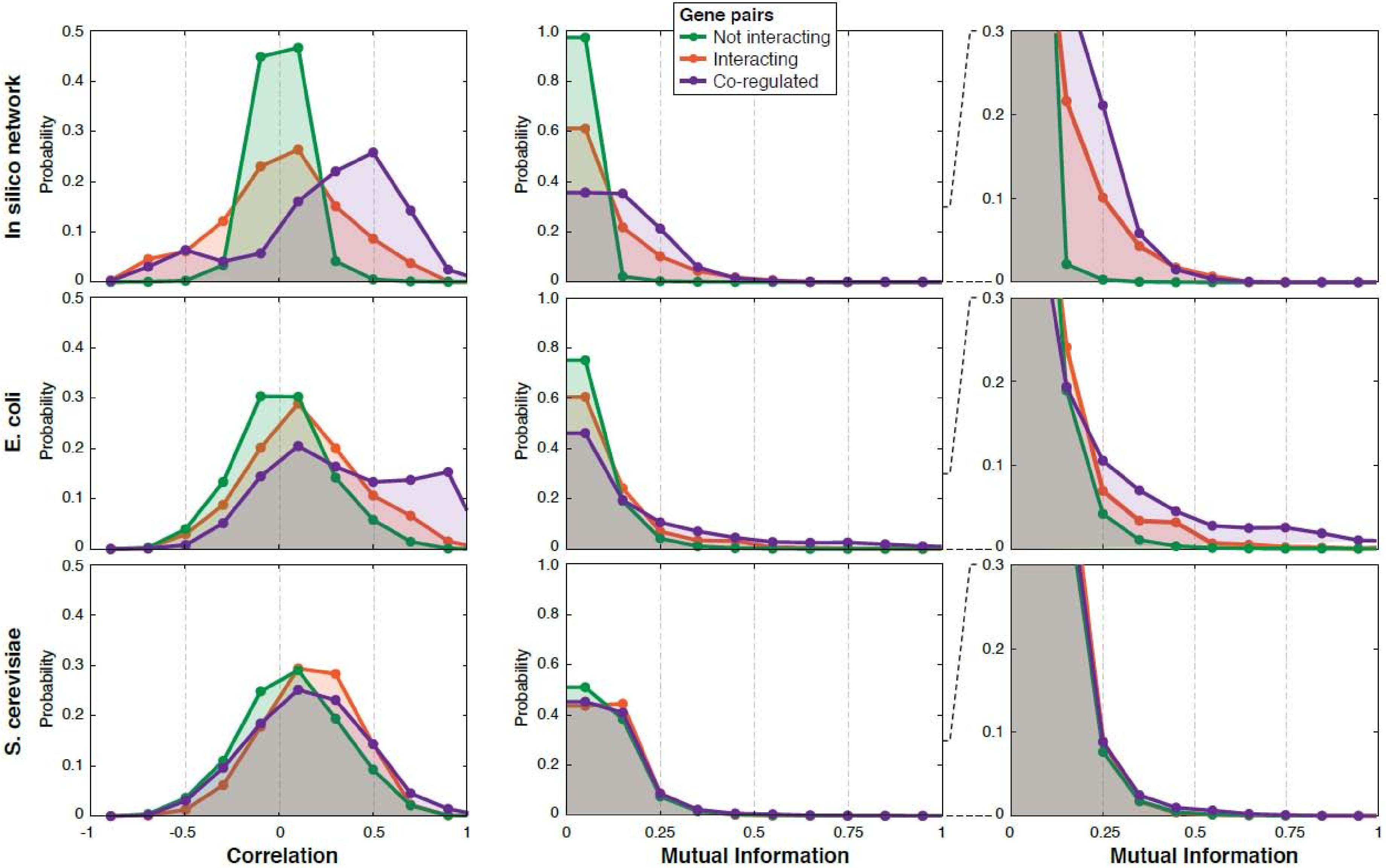
Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016



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

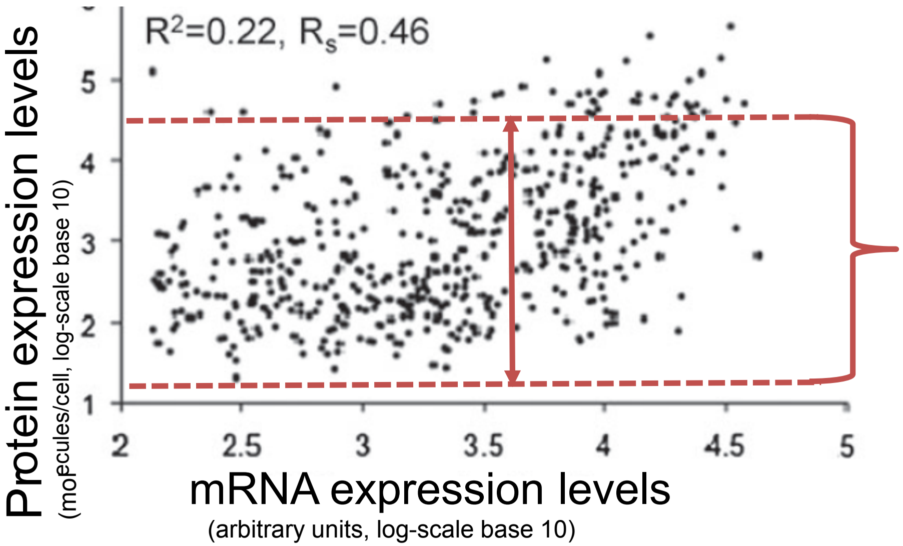
Nature Methods 9, 796–804 (2012) doi:10.1038/nmeth.2016

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

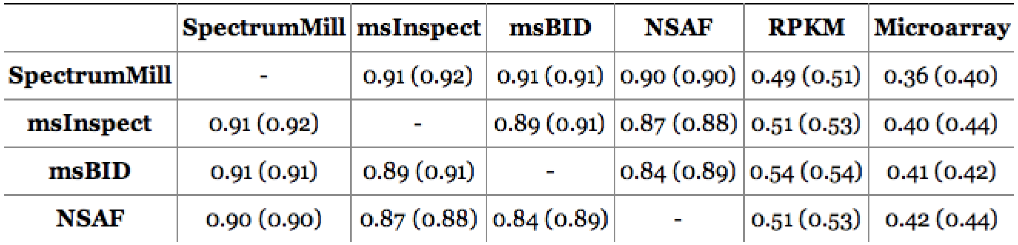


1,
of
c

000 fold range
protein
oncentrations

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.

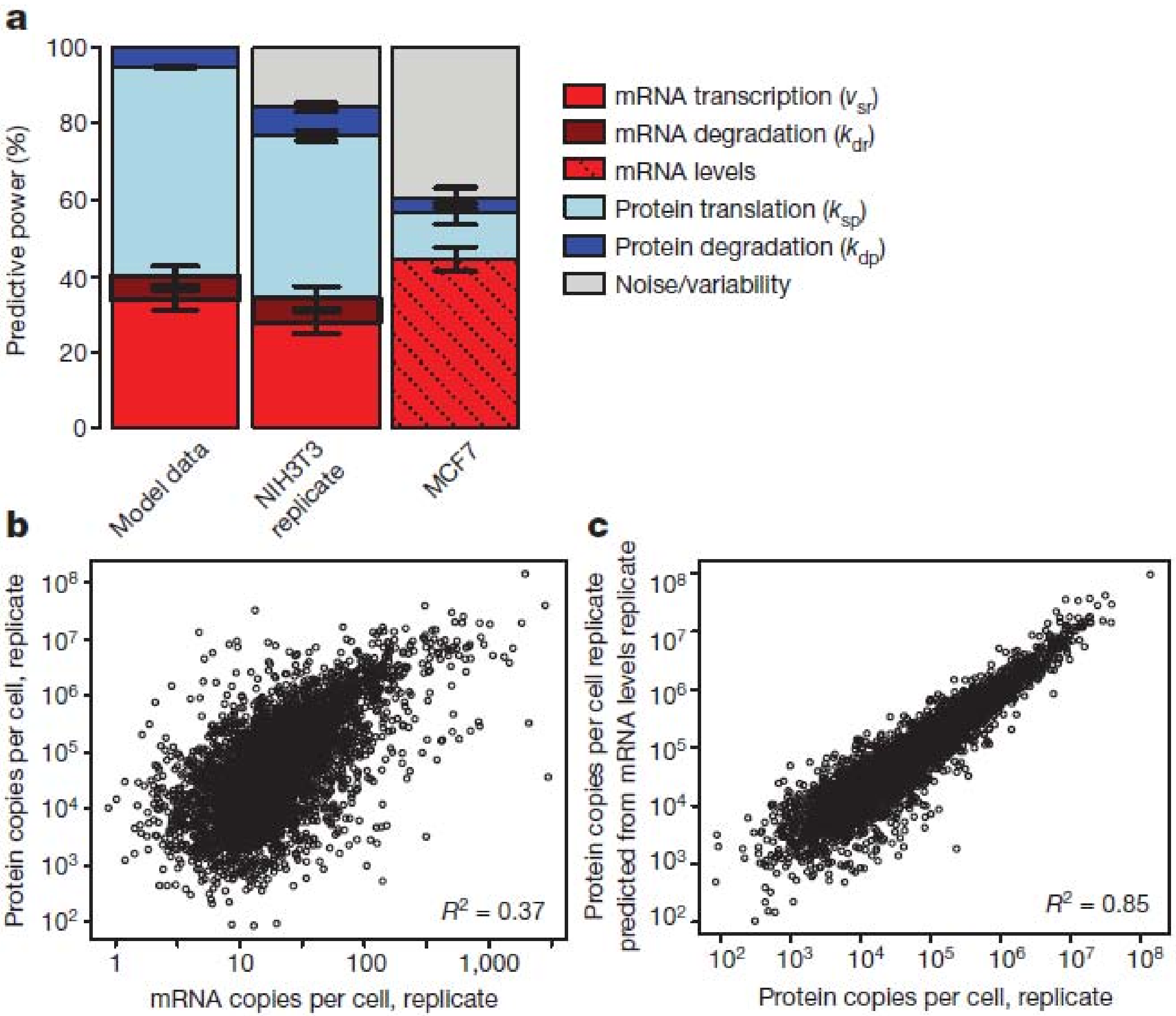
Raquel de Sousa Abreu, Luiz Penalva, Edward Marcotte and Christine Vogel, *Mol. BioSyst.*, 2009 DOI: 10.1039/b908315d



	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.

- 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

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