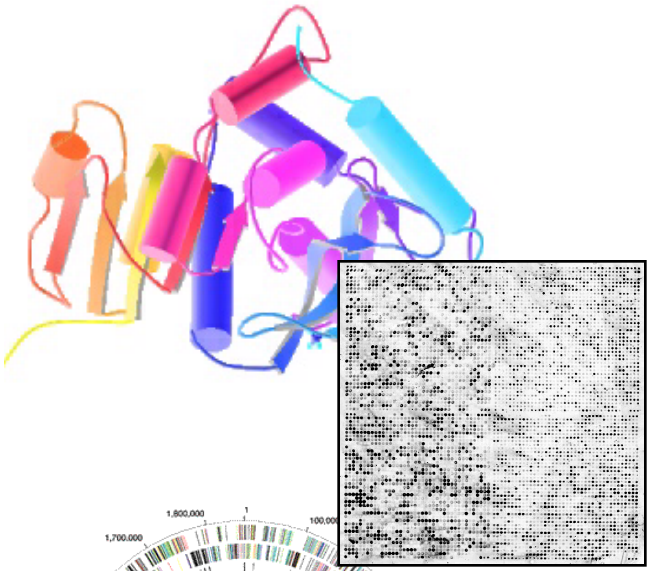
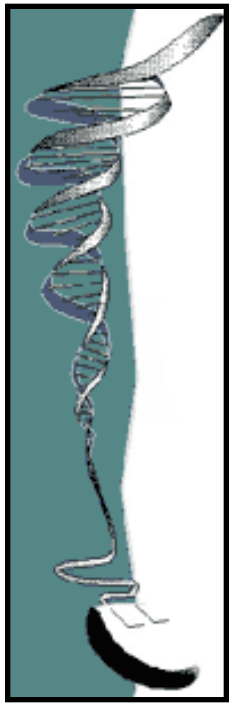
Bioinformatics: Predicting Networks





Mark Gerstein, Yale University
gersteinlab.org/courses/452
(last edit in Spring '17)

Origin of Biological Networks

Origin of Networks

- Protein-protein interactions
  - ◊ Phosphorylation networks
- Metabolic Networks
- Regulatory networks
  - ◊ from Chip-Seq (see next slide)

- "Squared" scale
  - ◊ 6K genes in yeast but ~18M potential interactions (6000 chose 2 pairs of interactions)

Data Flow: Chip-seq expts. to co-associating peaks

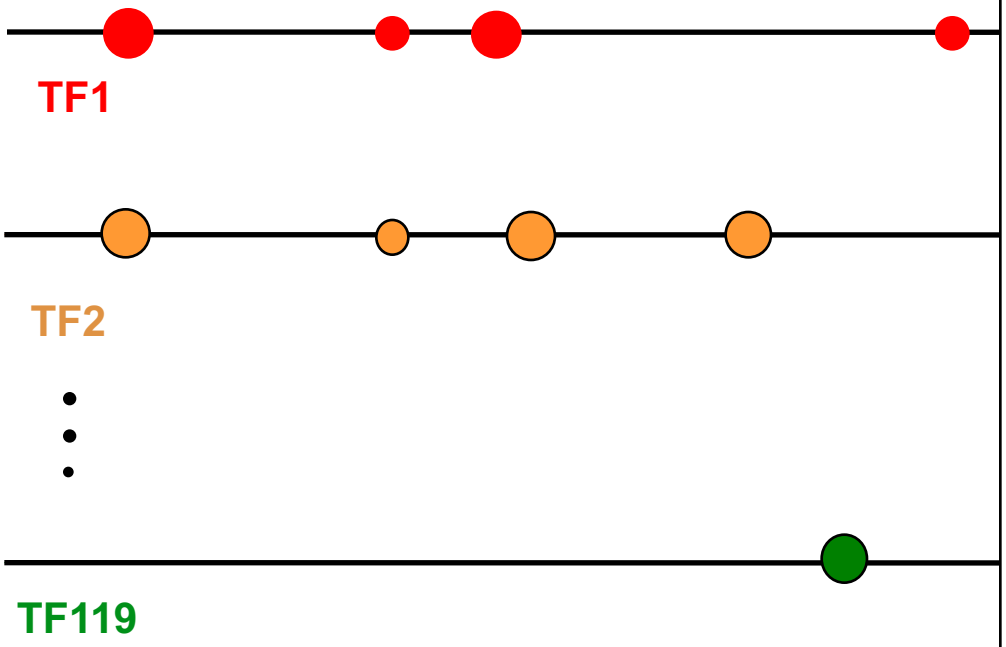
119 TFs from 458 ChIP-Seq experiments (2 Tb tot.)

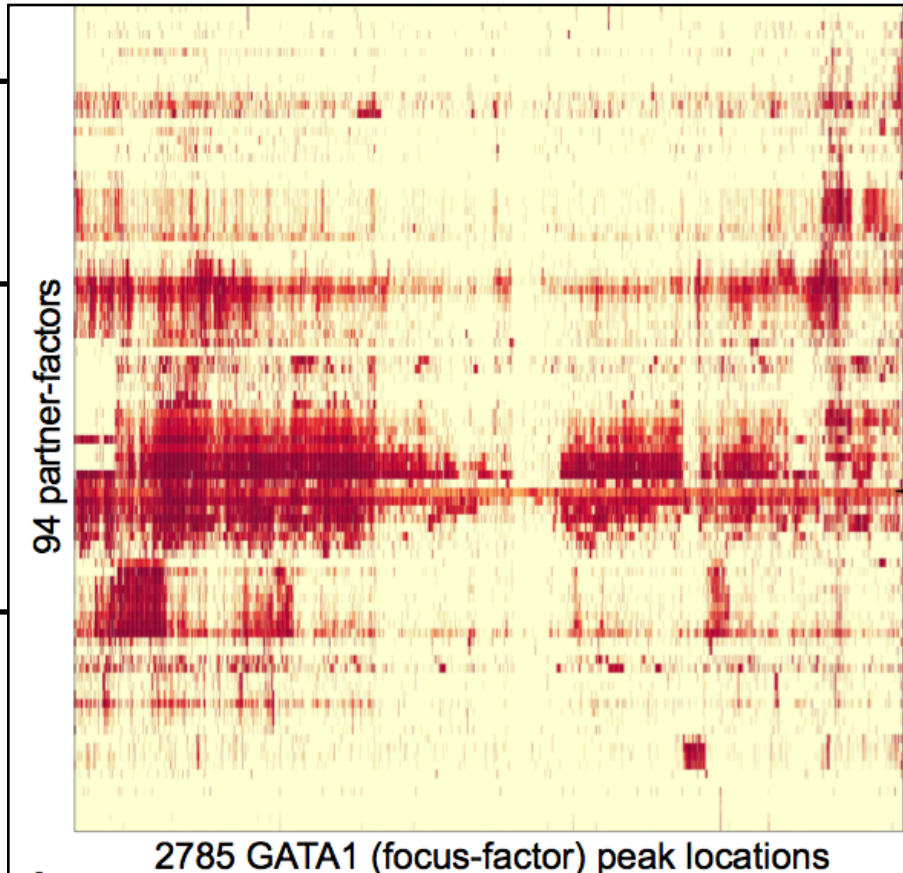
↓

Signal Tracks

↓

7M Peaks from Uniform Peak Calling





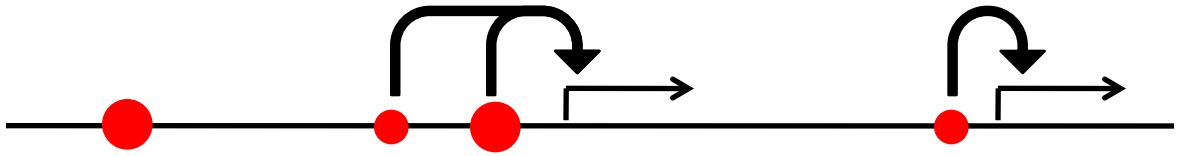
[Gerstein et al. Nature (in press, '12)]

Data Flow: peaks to proximal & distal networks

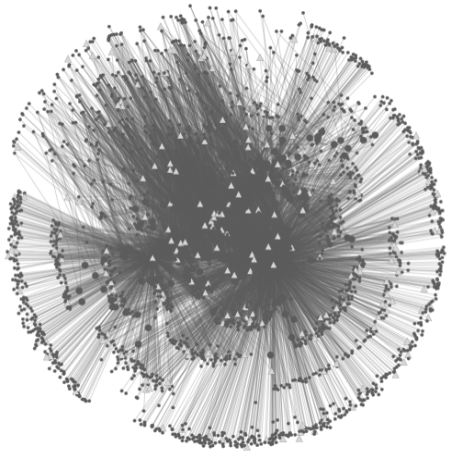
[Cheng et al., *Bioinfo.* ('11);
Gerstein et al. *Nature* (in press, '12) ;
Yip et al., *GenomeBiology* (in press, '12)]

Peak Calling

Assigning TF binding sites to targets

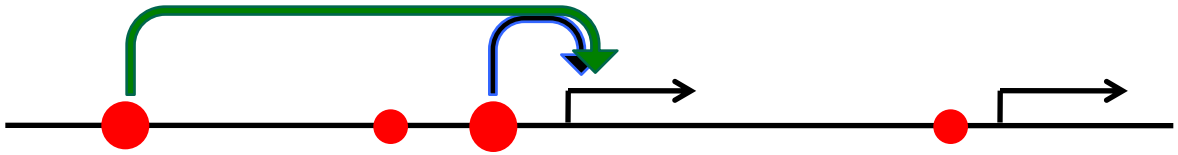


~500K Edges

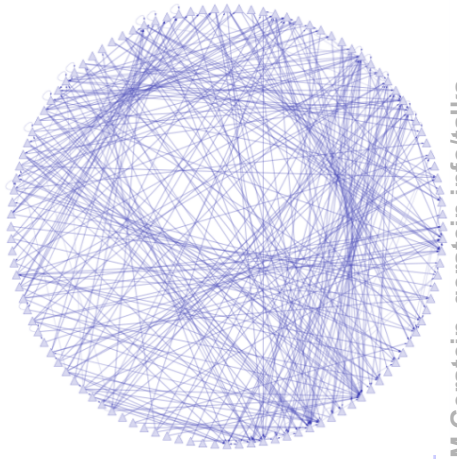


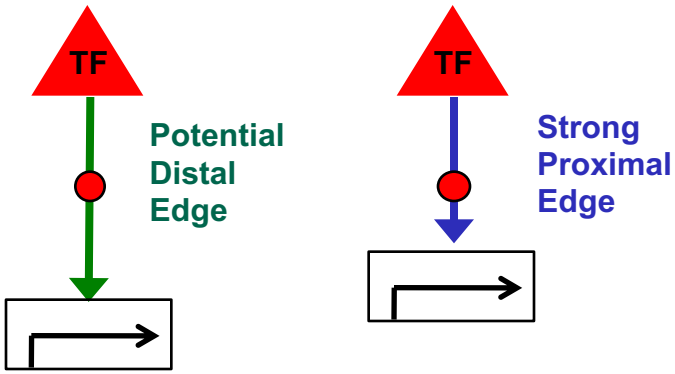
Filtering high confidence edges & distal regulation

Based on stat. model combining
signal strength & location relative to typical binding

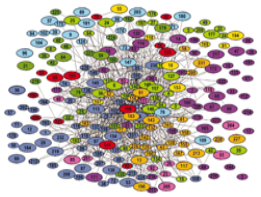


~26K Edges



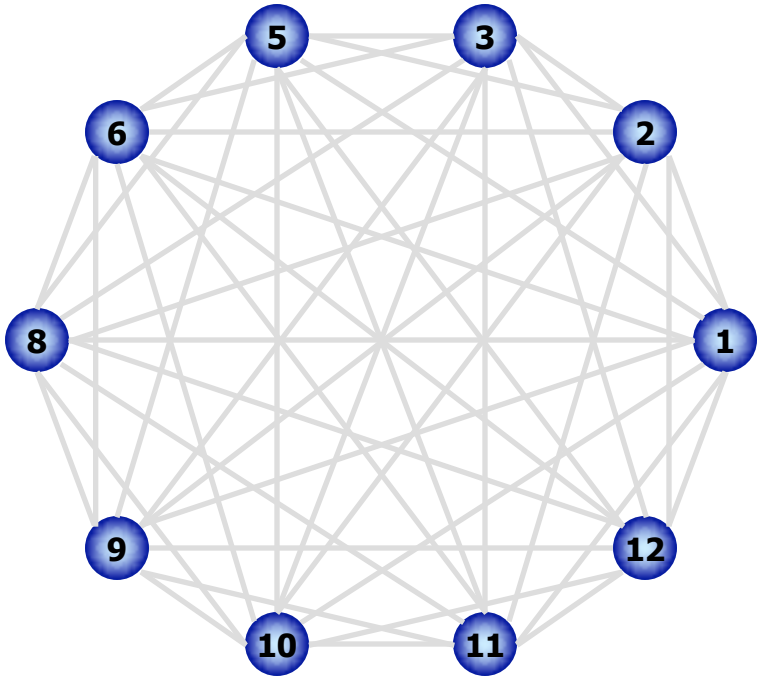


Predicting Networks via Bayesian Integration: Problem Motivation



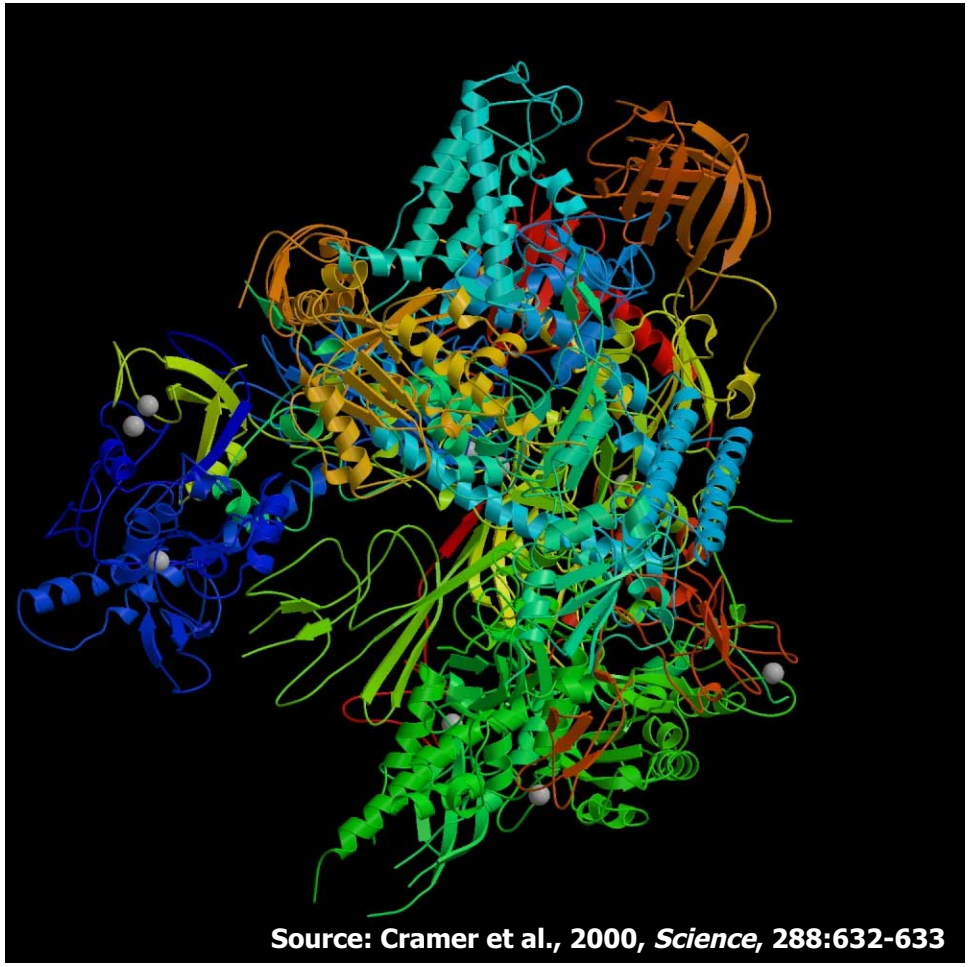
RNA polymerase II: Structure

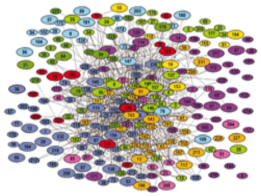
Which subunits interact? Based on Binding experiments



Source: Edwards et al., 2002, *Trends in Genetics*

Compare with Gold Std. Structure





Binding Experiments on Subunit Pairs

Subunits	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
Subunits	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12
Pull-down 1		1	1	0	1	0	1	0		1	1	0	1	0	1	0	
Pull-down 2		1	1	1	1	0	1	0		1	1	0	1	0	1	0	
Pull-down 3		1								1							
Cross-linking	1	1	1	1	1		0	1	1	1	1	1	0		1	1	1
Far Western 1		1	1							1	1						
Far Western 2				1	1		1	1	1			1	1		1	1	1
Far Western 3																	

Subunits	3	3	3	3	3	3	3	5	5	5	5	5	5
Subunits	5	6	8	9	10	11	12	6	8	9	10	11	12
Pull-down 1	1	1	1	0	1	1		1	1	0	1	0	
Pull-down 2	1	1	1	0	1	1		0	1	0	1	0	
Pull-down 3	1	0	1	0	0	1	0						
Cross-linking					1	1		1	0				
Far Western 1	1	0	0		0	0	0	0	1		0	0	0
Far Western 2		0	0		0	1	0	0	0		0	0	0
Far Western 3	1	0	0		0	1	0						

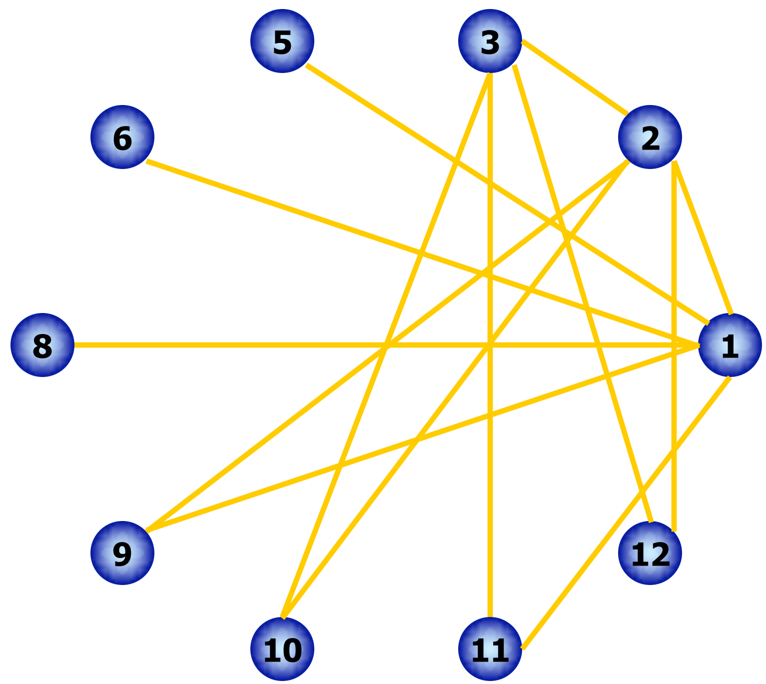
Subunits	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12
Pull-down 1	0	0	0	0		0	1	0		0	0		0		
Pull-down 2	0	0	0	0		0	0	0		0	0		0		
Pull-down 3															
Cross-linking	1														
Far Western 1															
Far Western 2	0		0	0	0		0	0	0				0	0	0
Far Western 3															

Interaction experiments
before structure was known

Gold-Standard Positives

Subunits	**1**	1	**1**	**1**	**1**	**1**	1	**1**	1	**2**	2	2	2	**2**	**2**	2	**2**	3	3	3	3	**3**	**3**	**3**	5	5	5	5	5	5	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	**2**	3	**5**	**6**	**8**	**9**	10	**11**	12	**3**	5	6	8	**9**	**10**	11	**12**	5	6	8	9	**10**	**11**	**12**	6	8	9	10	11	12	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12

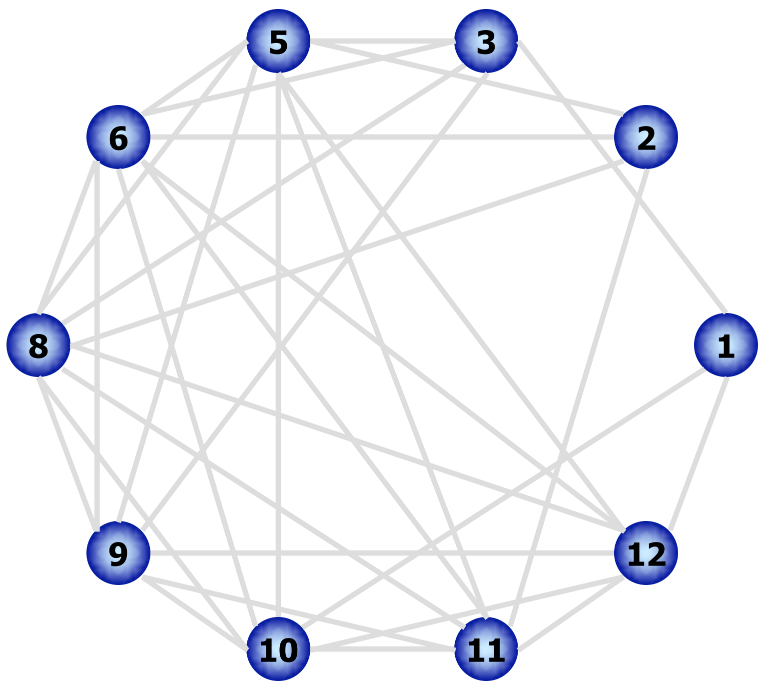
Gold-Standard Positive (GSTD+): 13



Gold-Standard Negatives

Subunits	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	5	5	5	5	5	5	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12

Gold-Standard Negative (GSTD-): 32

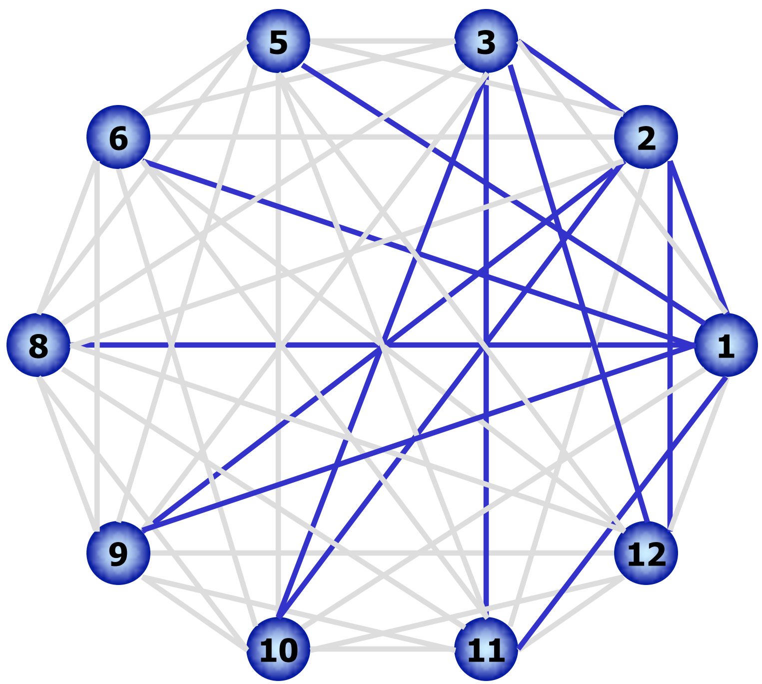


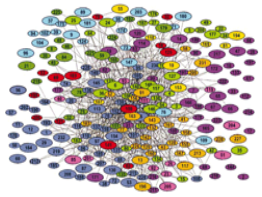
RNA Polymerase II: Gold-Standards

Subunits	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	5	5	5	5	5	5	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12

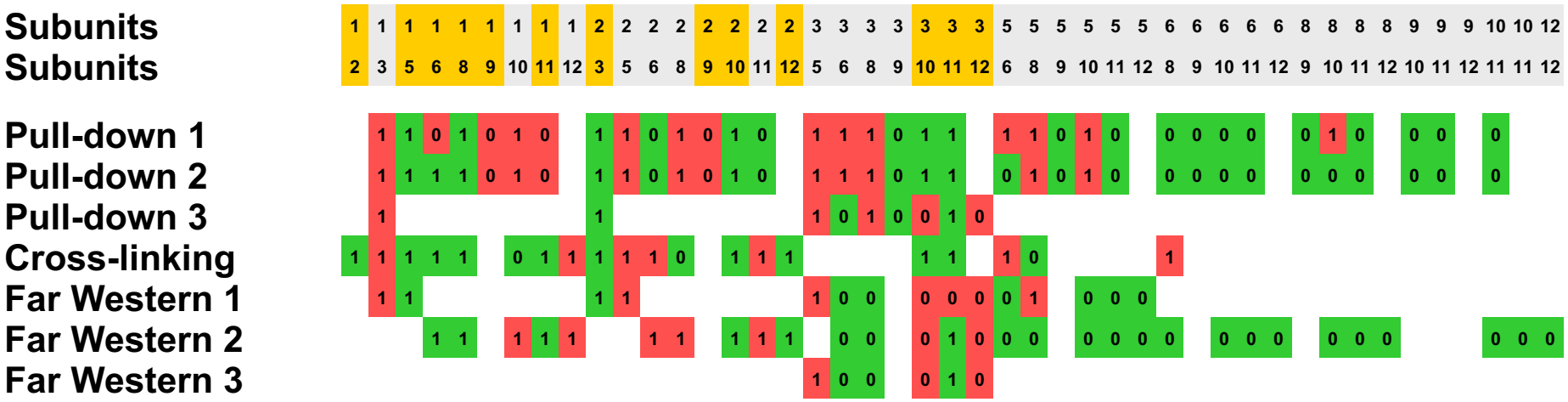
Gold-Standard Positive (GSTD+): 13

Gold-Standard Negative (GSTD-): 32



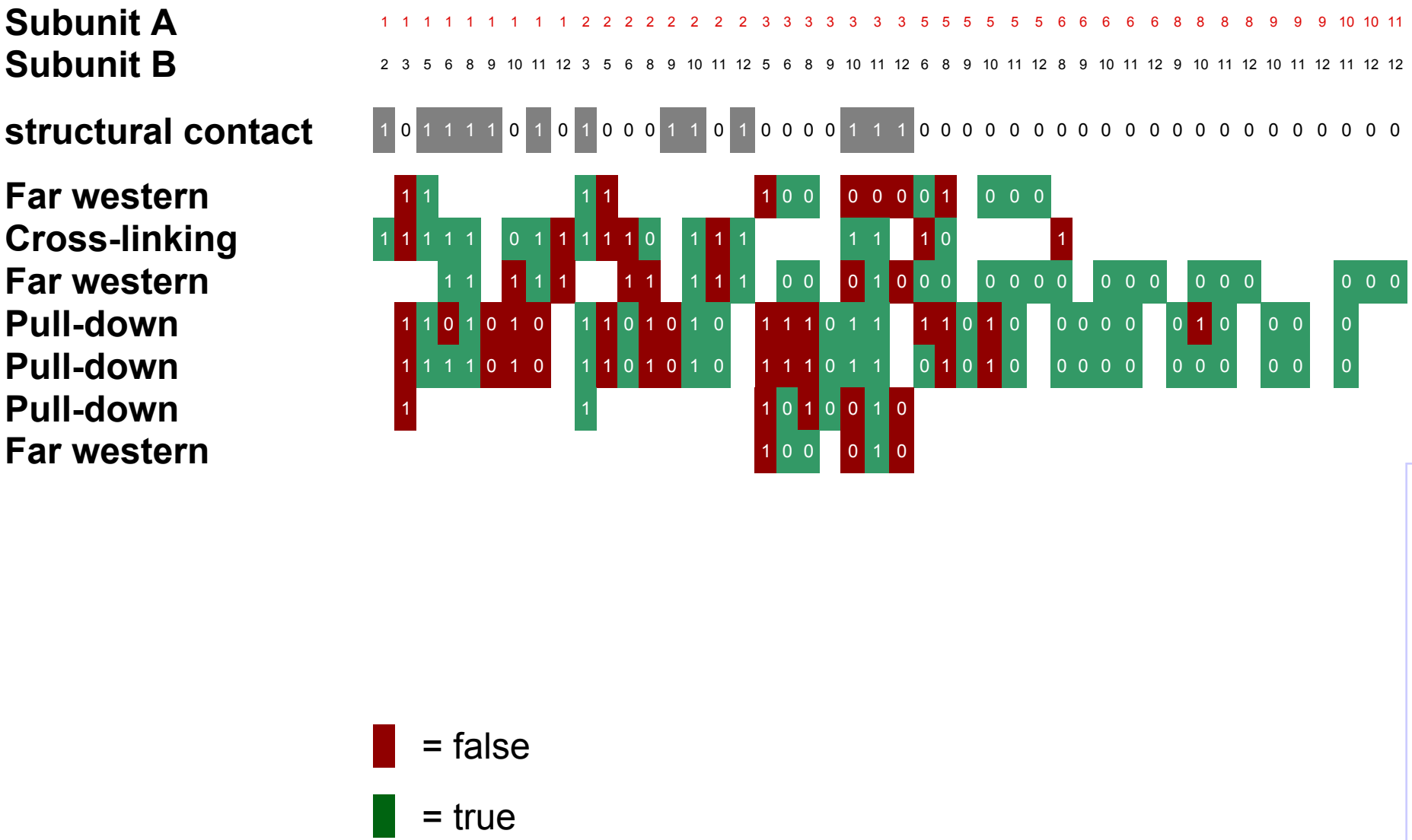


Assess Quality and Coverage of PPints

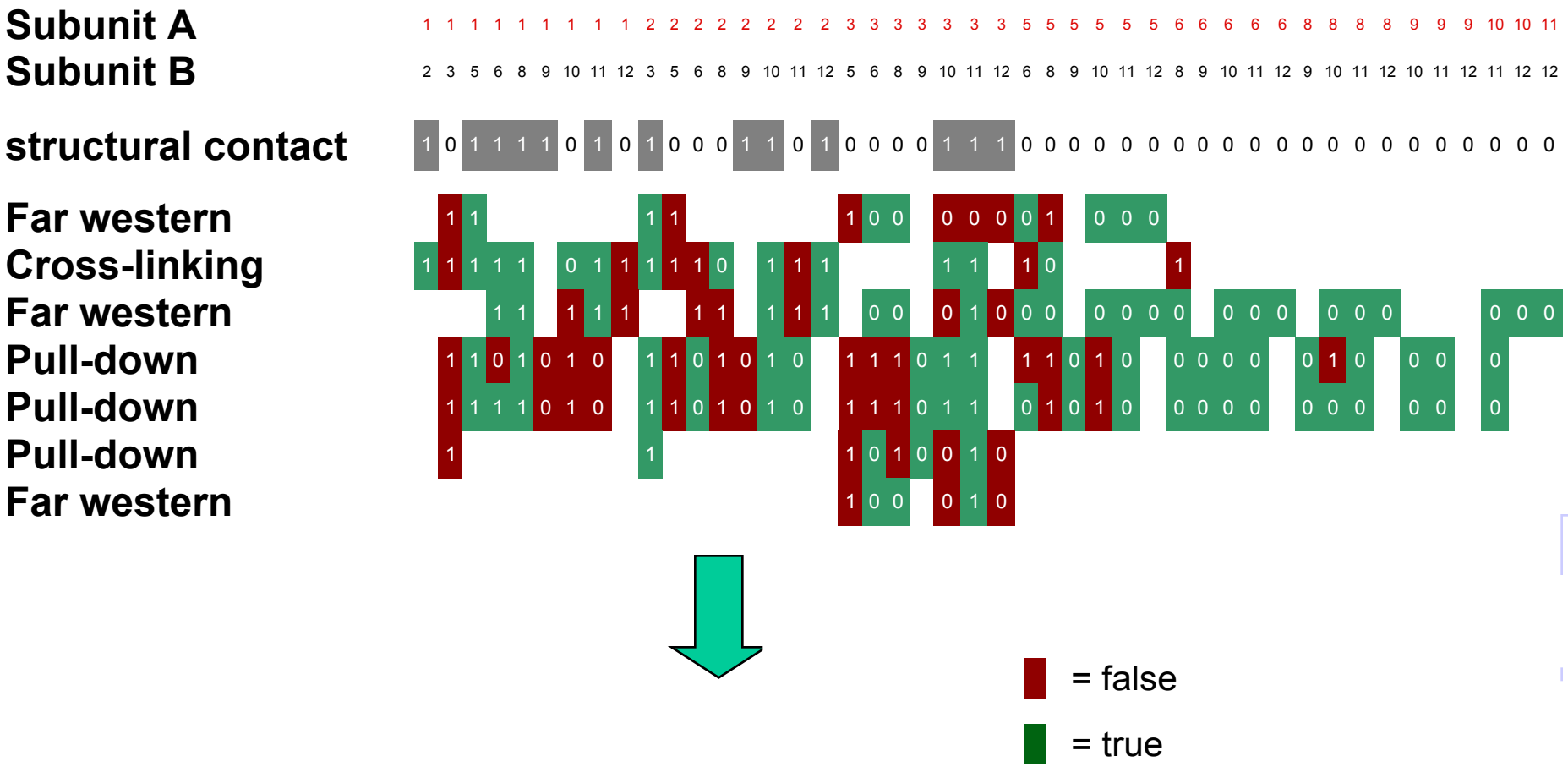


GSTD+
GSTD-
True
False

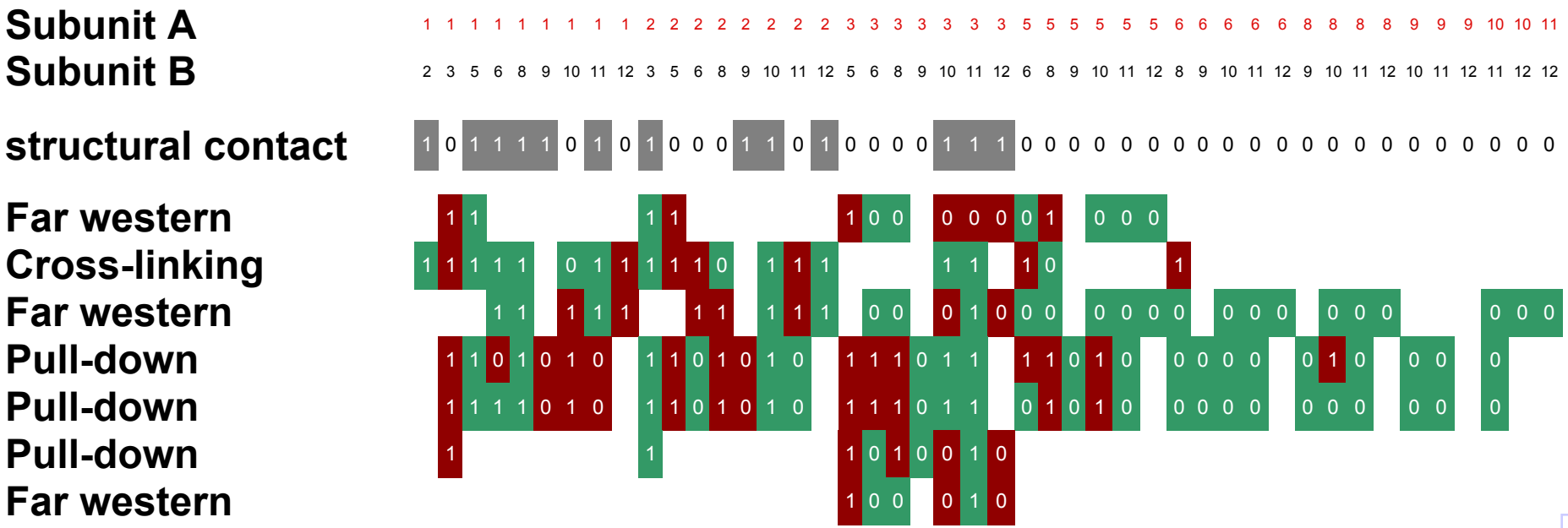
Data integration: RNA polymerase II

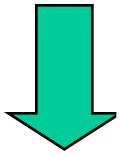


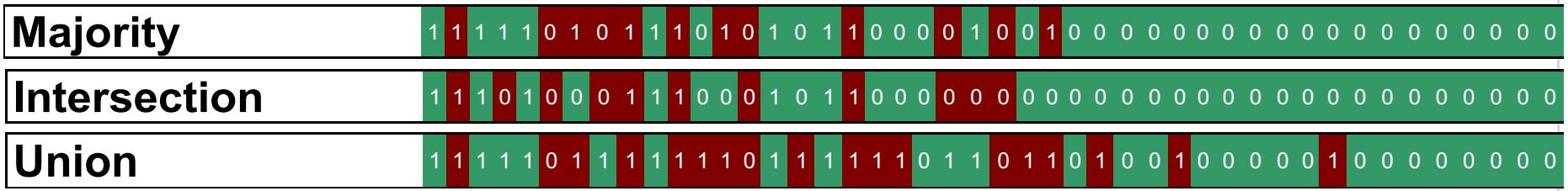
Data integration: RNA polymerase II



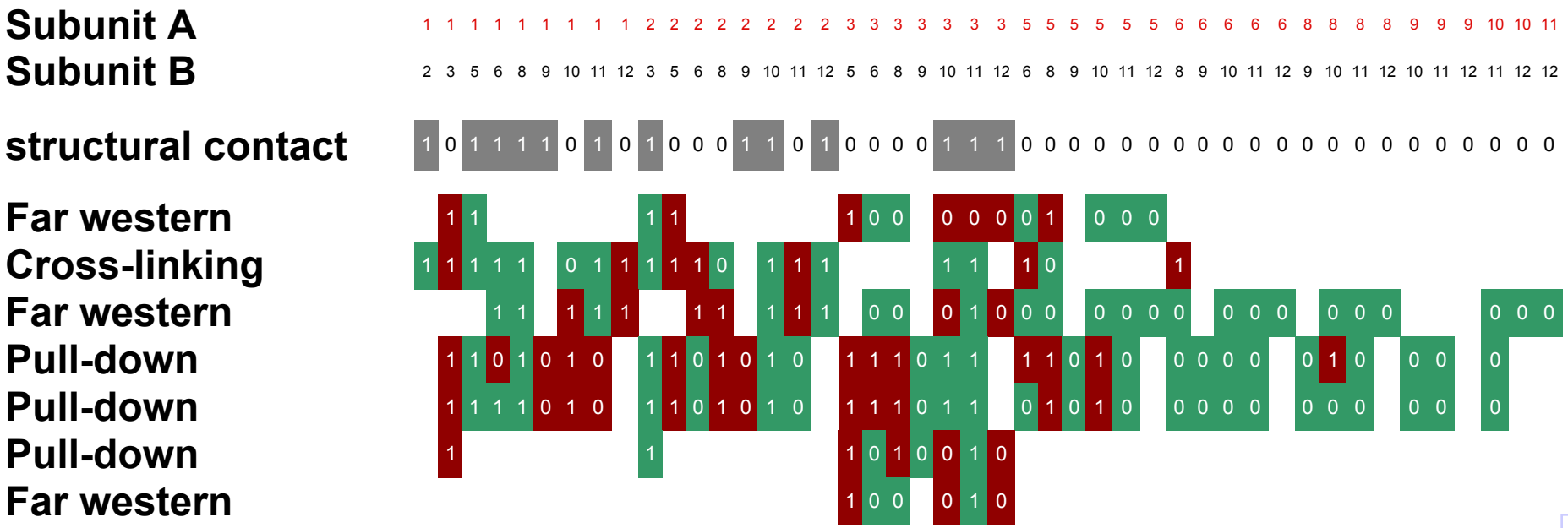
Data integration: RNA polymerase II





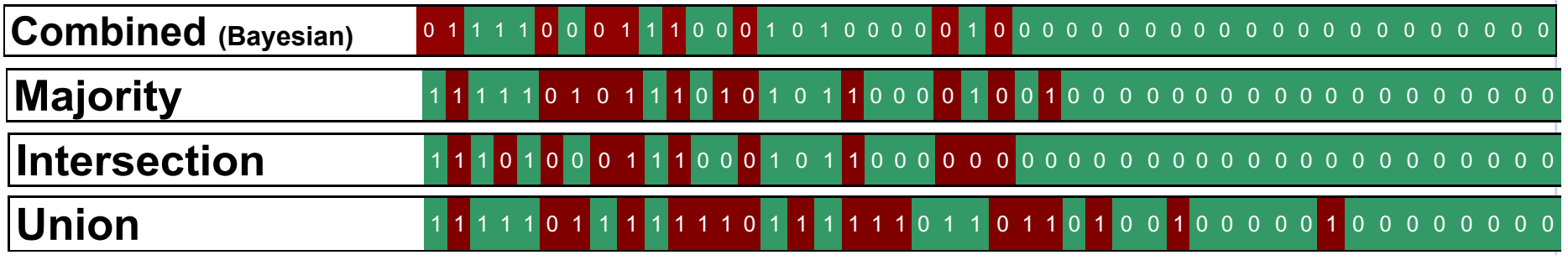


Data integration: RNA polymerase II

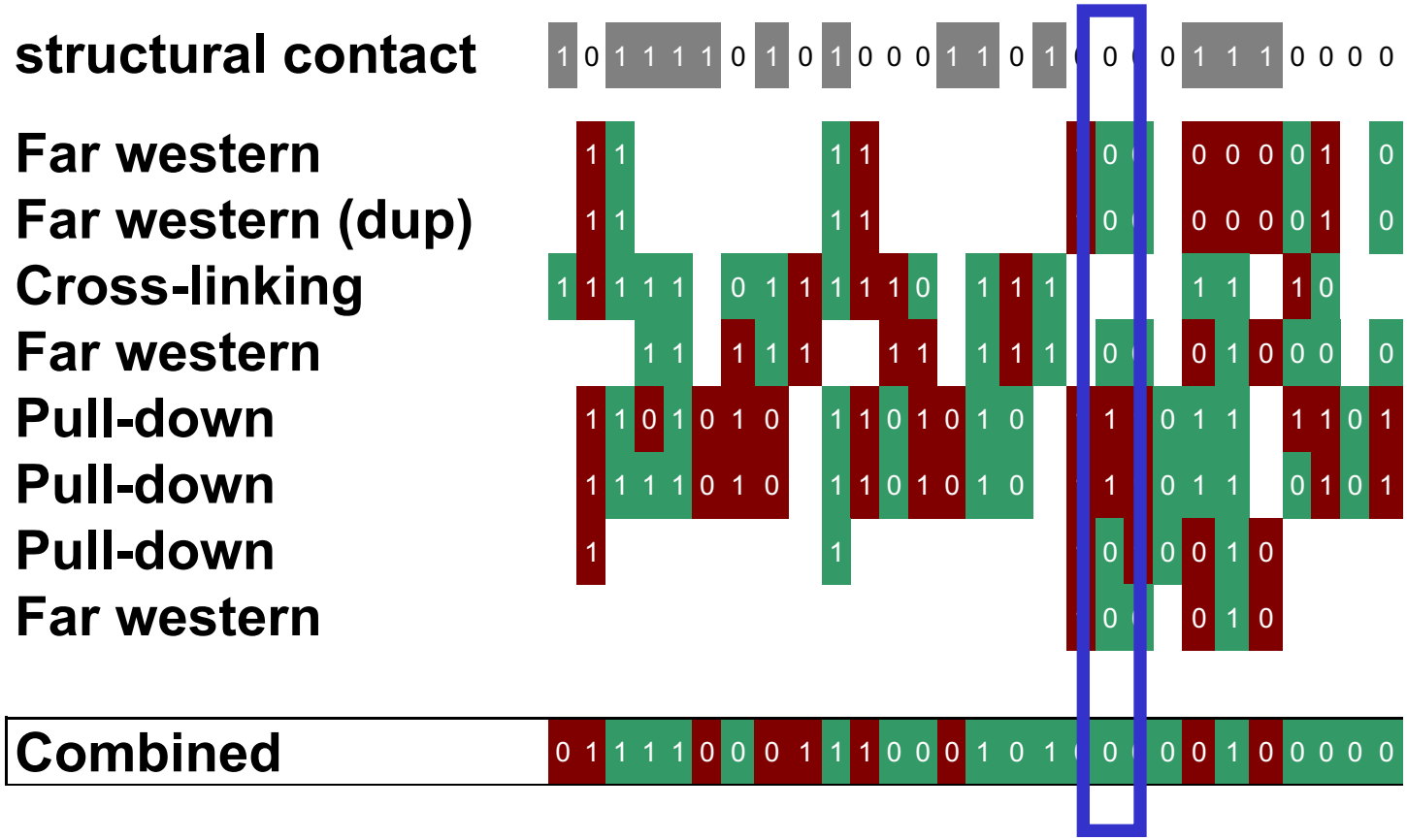


(Cross validate)

Integrate using naive Bayes classifier



Weighted Voting: the Likelihood Ratio



Maj. Vote: 0 = round(avg(0 + 0 + 0 + 1 + 1 + 0 + 0)

With weights: **likelihood ratio L** = $L_1 + L_2 + L_3 \ldots$

Predicting Networks via Bayesian Integration: Intuition & Formalism

Derived from
"perceptron model"
R = <w,f> + b

Supervised Classification by Weighted Voting

Simple Vote: $R = f_1 + f_2 + f_3 + \ldots + f_n$ With f = 1 or -1

If $\begin{cases} R>0; & I \text{ Interact} \\ R<0; & \sim I \text{ No interaction} \end{cases}$

Modify with feature weight:

$$R = w_1 f_1 + w_2 f_2 + w_3 f_3 + \ldots + w_n f_n = \vec{w} \cdot \vec{f}$$

If has prior knowledge w_0

$$R = \vec{w} \cdot \vec{f} + w_0$$

Classification by Voting

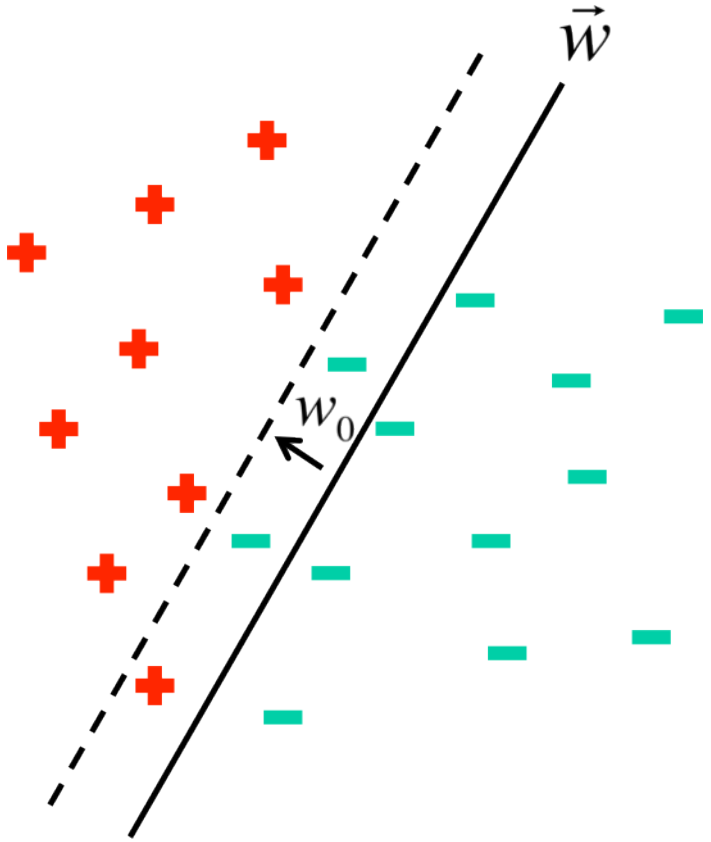
$$R = \vec{w} \cdot \vec{f} + w_0$$

$$w_1 = \log \frac{P(f_1 = 1 \mid I)}{P(f_1 = 1 \mid \sim I)}$$

$$= \log \frac{TP / P}{FR / N}$$

$$w_0 = \log \frac{P}{N}$$ (Estimated from Golden Standard)

On Training Set



Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$P(Y = y_i|X = x_j) = \frac{P(X = x_j|Y = y_i)P(Y = y_i)}{P(X = x_j)}$$

[From Mitchell, Machine Learning]

Thus

$$P(I \mid f_1, f_2, f_3, \ldots) = \frac{P(f_1, f_2, f_3, \ldots \mid I)P(I)}{P(f_1, f_2, f_3, \ldots)}$$

Assume Naïve Bayes (independent)

$$= \frac{P(f_1 \mid I)P(f_2 \mid I)P(f_3 \mid I)\ldots P(I)}{P(f_1, f_2, f_3, \ldots)}$$

$$P(\sim I \mid f_1, f_2, f_3, \ldots) = \frac{P(f_1, f_2, f_3, \ldots \mid \sim I)P(\sim I)}{P(f_1, f_2, f_3, \ldots)}$$

$$= \frac{P(f_1 \mid \sim I)P(f_2 \mid \sim I)P(f_3 \mid \sim I)\ldots P(\sim I)}{P(f_1, f_2, f_3, \ldots)}$$

$$\log\left(\frac{P(I \mid f_1, f_2, f_3, \ldots)}{P(\sim I \mid f_1, f_2, f_3, \ldots)}\right) = \log\left(\frac{P(f_1 \mid I)}{P(f_1 \mid \sim I)} \frac{P(f_2 \mid I)}{P(f_2 \mid \sim I)} \frac{P(f_3 \mid I)}{P(f_3 \mid \sim I)} \ldots \frac{P(I)}{P(\sim I)}\right)$$

$$= \log \frac{TPR_1}{FPR_1} + \log \frac{TPR_2}{FPR_2} + \log \frac{TPR_3}{FPR_3} + .. + \log \frac{P}{N}$$

More Bayes Rule

$$\log\left(\frac{P(I \mid f_1, f_2, f_3, \ldots)}{P(\sim I \mid f_1, f_2, f_3, \ldots)}\right) = \log\frac{TPR_1}{FPR_1} + \log\frac{TPR_2}{FPR_2} + \log\frac{TPR_3}{FPR_3} + .. + \log\frac{P}{N}$$

$$\uparrow \qquad \uparrow \qquad \uparrow \qquad \uparrow$$

$$w_1 \qquad w_2 \qquad w_3 \qquad w_0$$

More Bayes Rule

Estimating Probabilities

- We have so far estimated $P(X=x \mid Y=y)$ by the fraction $n_{x|y}/n_y$, where n_y is the number of instances for which $Y=y$ and $n_{x|y}$ is the number of these for which $X=x$
- This is a problem when n_x is small
  - ◊ E.g., assume $P(X=x \mid Y=y)=0.05$ and the training set is s.t. that $n_y=5$. Then it is highly probable that $n_{x|y}=0$
  - ◊ The fraction is thus an underestimate of the actual probability
  - ◊ It will dominate the Bayes classifier for all new queries with $X=x$

$$\frac{\#\text{ count with feature i in GS+}}{\#\text{ count with feature i in GS-}} = \frac{TPR_i}{FPR_i}$$

m-estimate

- Replace $n_{x|y}/n_y$ by:

$$\frac{n_{x|y} + mp}{n_y + m}$$

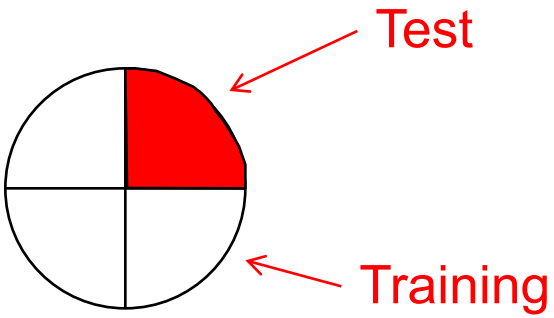
Dummy Counts

- Where *p* is our prior estimate of the probability we wish to determine and *m* is a constant
  - ◊ Typically, $p = 1/k$ (where *k* is the number of possible values of *X*)
  - ◊ *m* acts as a weight (similar to adding *m* virtual instances distributed according to *p*)

Predicting Networks: Assessment via Cross-Validation

Training and Testing Set

- Cross Validation: Leave one out, seven-fold



4-fold

Cross Validation

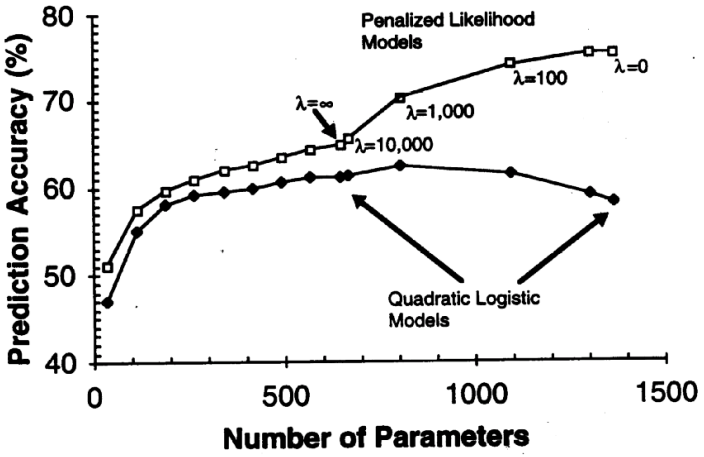
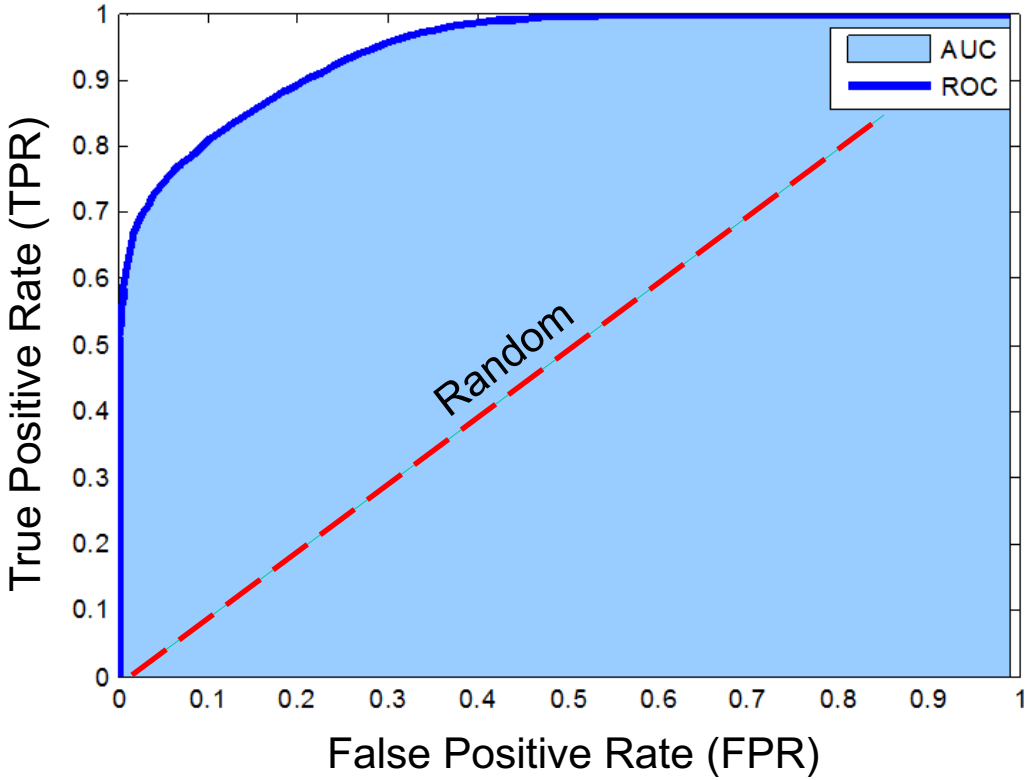


Figure 2. Comparison of prediction accuracy (correctly predicted residues as a proportion of total residues) versus effective number of parameters for linear-logistic models (number of parameters ≤640) and penalized likelihood models for crossvalidated (◆) and uncrossvalidated (❑) results. The values of the penalty parameter λ are shown.

Credits: Munson, 1995;
Garnier et al., 1996

Intuition :
ROC Curve & the prior

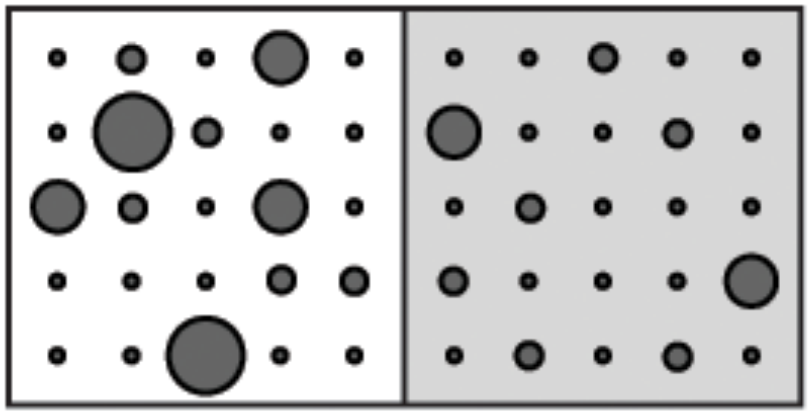


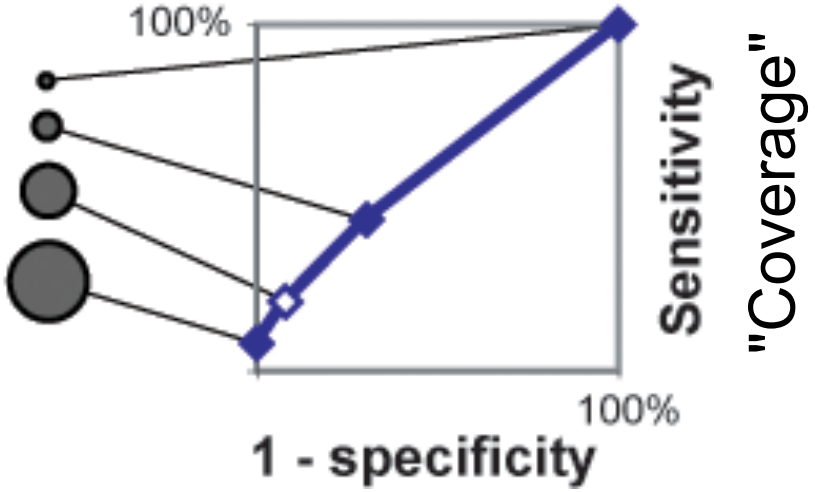
$$TPR = TP / P = TP / (TP + FN)$$

$$FPR = FP / N = FP / (FP + TN)$$

[From Biometrical Fusion - input statistical distribution]

Comparison of Predictions against a Positive and Negative Gold Standard

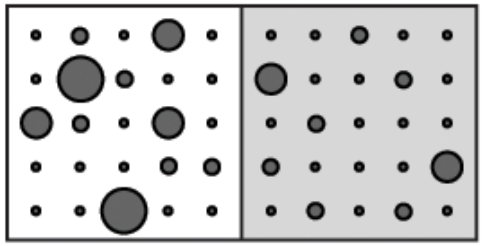


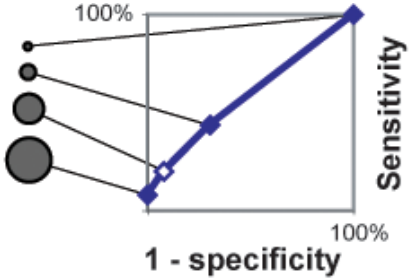


Threshold "predictions" at different levels and compare to + and - gold standards

"Error Rate"

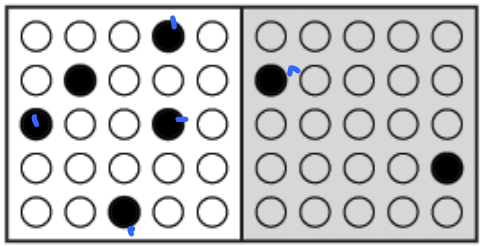
ROC plot
(cross validated)



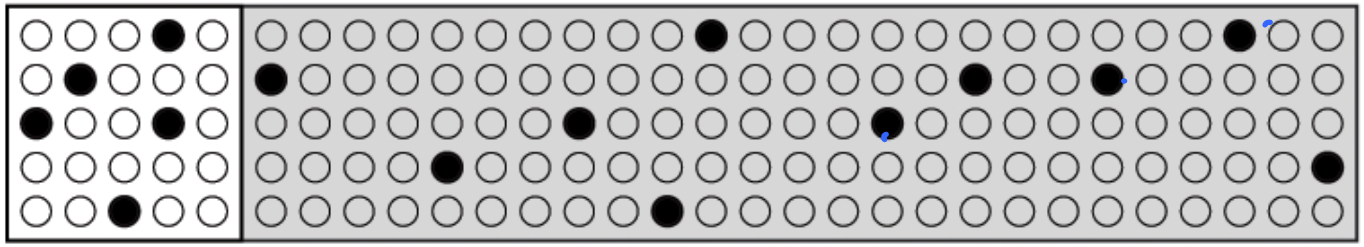


Effect on Predictions of Large Number of Negatives

Sensitivity **1- specificity** **Positive predictive value**

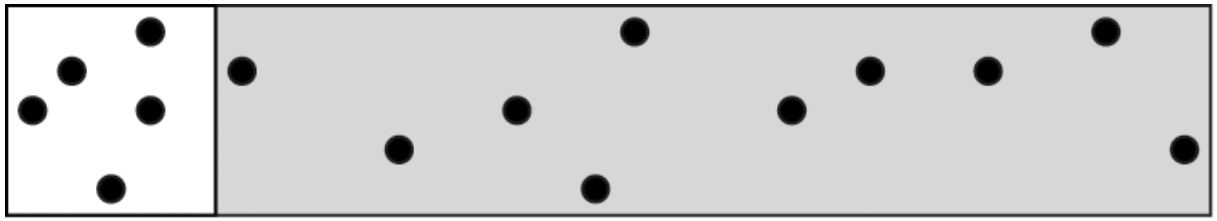


$\frac{5}{25} = 20\%$ $\frac{2}{25} = 8\%$ $\frac{5}{5+2} \approx 71\%$

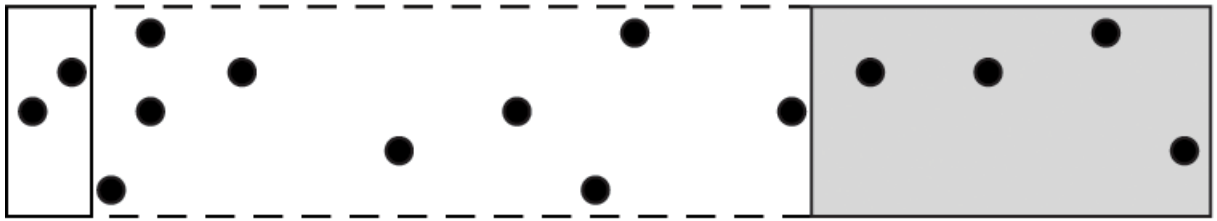


$\frac{5}{25} = 20\%$ $\frac{10}{125} = 8\%$ $\frac{5}{5+10} \approx 33\%$

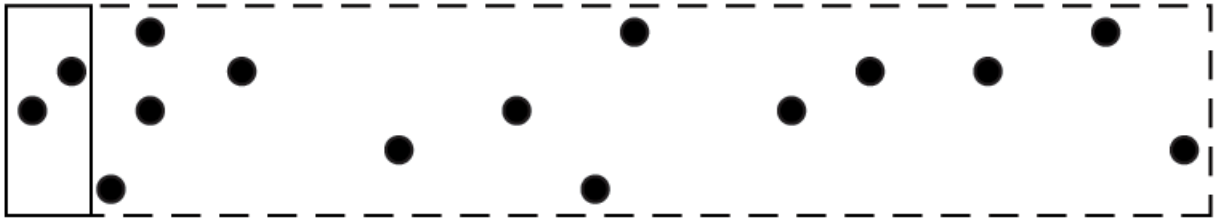
Importance of Balanced Positive and Negative Examples



$\frac{5}{?} = ?$ $\frac{10}{?} = ?$ $\frac{5}{5+10} \approx \mathbf{33\%}$

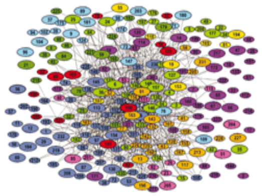


$\frac{2}{?} = ?$ $\frac{4}{?} = ?$ $\frac{2}{2+4} \approx \mathbf{33\%}$ (estimate)



$\frac{2}{?} = ?$ $\frac{?}{?} = ?$ $\frac{2}{2+?} = ?$

Predicting Networks via Bayesian Integration: Worked Examples



Likelihood Ratios

Subunits	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
Subunits	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10
Pull-down 1		1	1	0	1	0	1	0		1	1	0	1	0	1

Subunits	2	2	3	3	3	3	3	3	3	5	5	5	5	5	5
Subunits	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12
Pull-down 1	0		1	1	1	0	1	1		1	1	0	1	0	

Subunits	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12
Pull-down 1	0	0	0	0		0	1	0		0	0		0		

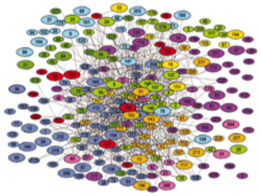
$$L_1 = \frac{p(x_1 \mid GSTD+)}{p(x_1 \mid GSTD-)}$$

$$L_0 = \frac{p(x_0 \mid GSTD+)}{p(x_0 \mid GSTD-)}$$

Likelihood Ratio
for Feature *f*:

$$L_f \equiv \frac{p(x_f \mid GSTD+)}{p(x_f \mid GSTD-)}$$

GSTD+
GSTD-
True
False



Calculating Likelihood Ratios

Subunits	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	5	5	5	5	5	5	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12
Pull-down 1			1	0	1	0		0		1				0	1							1	1																						

$$L_1 = \frac{p(x_1 \mid GSTD+)}{p(x_1 \mid GSTD-)} = \frac{6/13}{}$$

$$L_0 = \frac{p(x_0 \mid GSTD+)}{p(x_0 \mid GSTD-)} = \frac{4/13}{}$$

- GSTD+
- GSTD-
- True
- False

Calculating Likelihood Ratios

Subunits	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
Subunits	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10
Pull-down 1		1	1	0	1	0	1	0		1	1	0	1	0	1

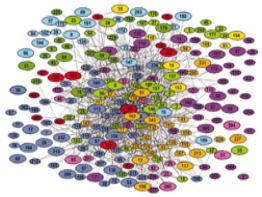
Subunits	2	2	3	3	3	3	3	3	3	5	5	5	5	5	5
Subunits	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12
Pull-down 1	0		1	1	1	0	1	1		1	1	0	1	0	

Subunits	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12
Pull-down 1	0	0	0	0		0	1	0		0	0		0		

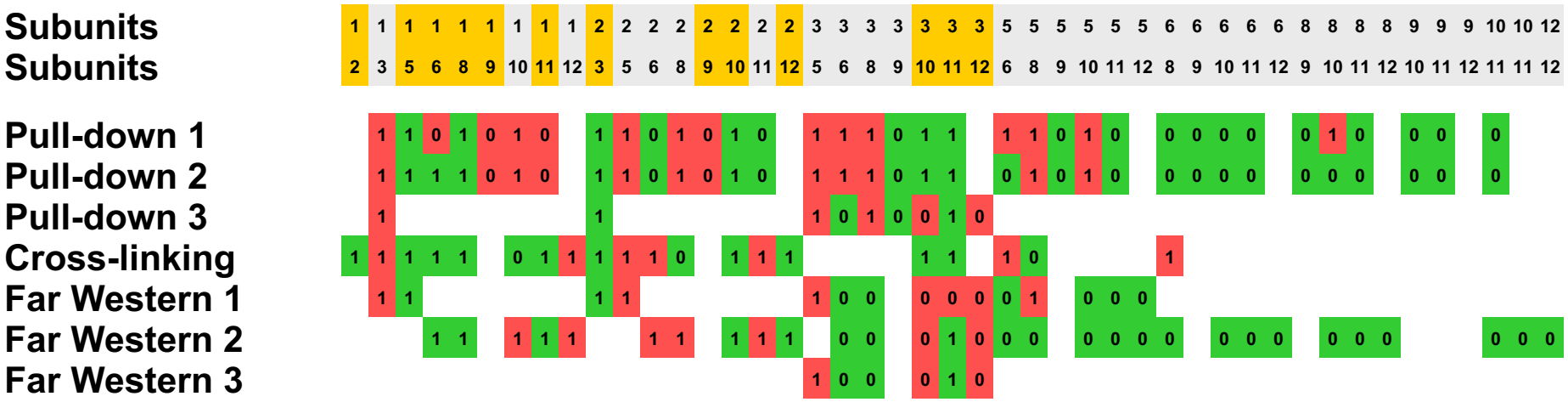
$$L_1 = \frac{p(x_1 \mid GSTD+)}{p(x_1 \mid GSTD-)} = \frac{6/13}{11/32} = 1.34$$

$$L_0 = \frac{p(x_0 \mid GSTD+)}{p(x_0 \mid GSTD-)} = \frac{4/13}{14/32} = 0.70$$

GSTD+
GSTD-
True
False

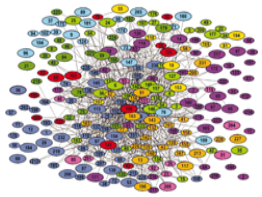


Calculating Likelihood Ratios

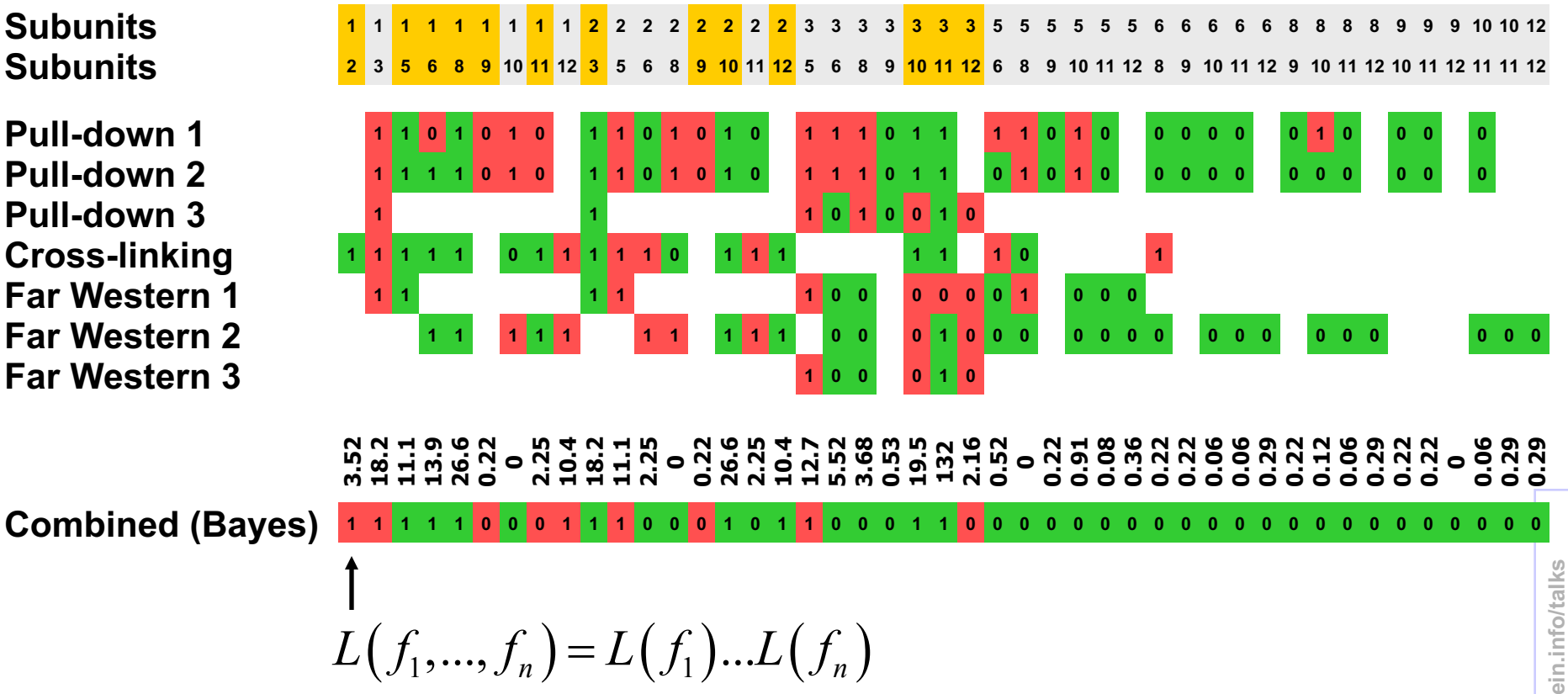


Pull-down 1	*L1* = (6/13) / (11/32) = 1.34	*L0* = (4/13) / (14/32) = 0.70
Pull-down 2	*L1* = (7/13) / (9/32) = 1.91	*L0* = (2/13) / (16/32) = 0.31
Pull-down 3	*L1* = (2/13) / (3/32) = 1.64	*L0* = (2/13) / (2/32) = 2.46
Cross-linking	*L1* = (10/13) / (7/32) = 3.52	*L0* = (0/13) / (3/32) = 0
Far Western 1	*L1* = (2/13) / (4/32) = 1.23	*L0* = (3/13) / (6/32) = 1.23
Far Western 2	*L1* = (6/13) / (5/32) = 2.95	*L0* = (2/13) / (17/32) = 0.29
Far Western 3	*L1* = (1/13) / (1/32) = 2.46	*L0* = (2/13) / (2/32) = 2.46

GSTD+
GSTD-
True
False



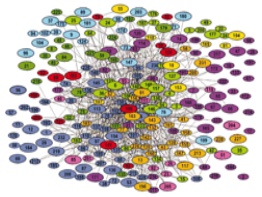
Data Integration: ROC-Curve



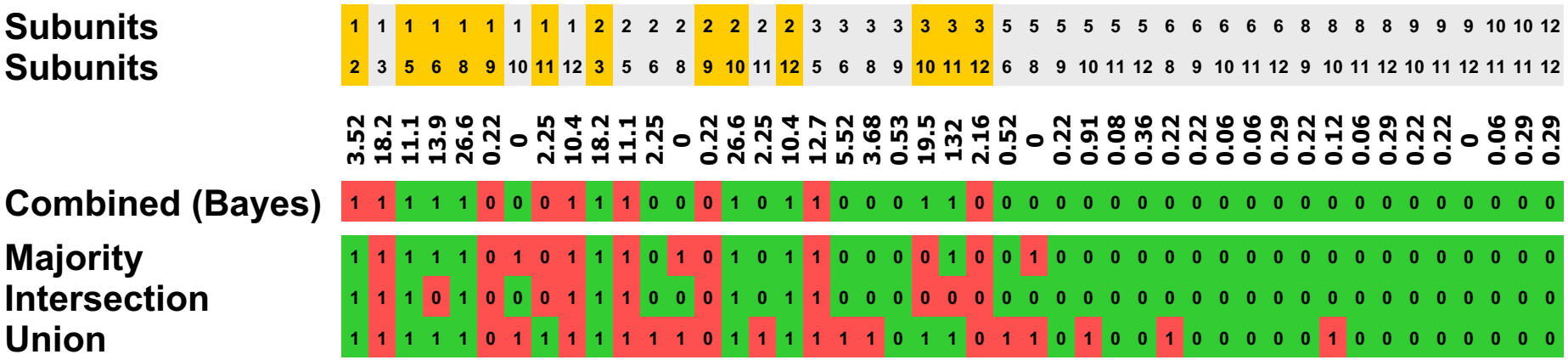
$$L(f_1,\ldots,f_n) = L(f_1)\ldots L(f_n)$$

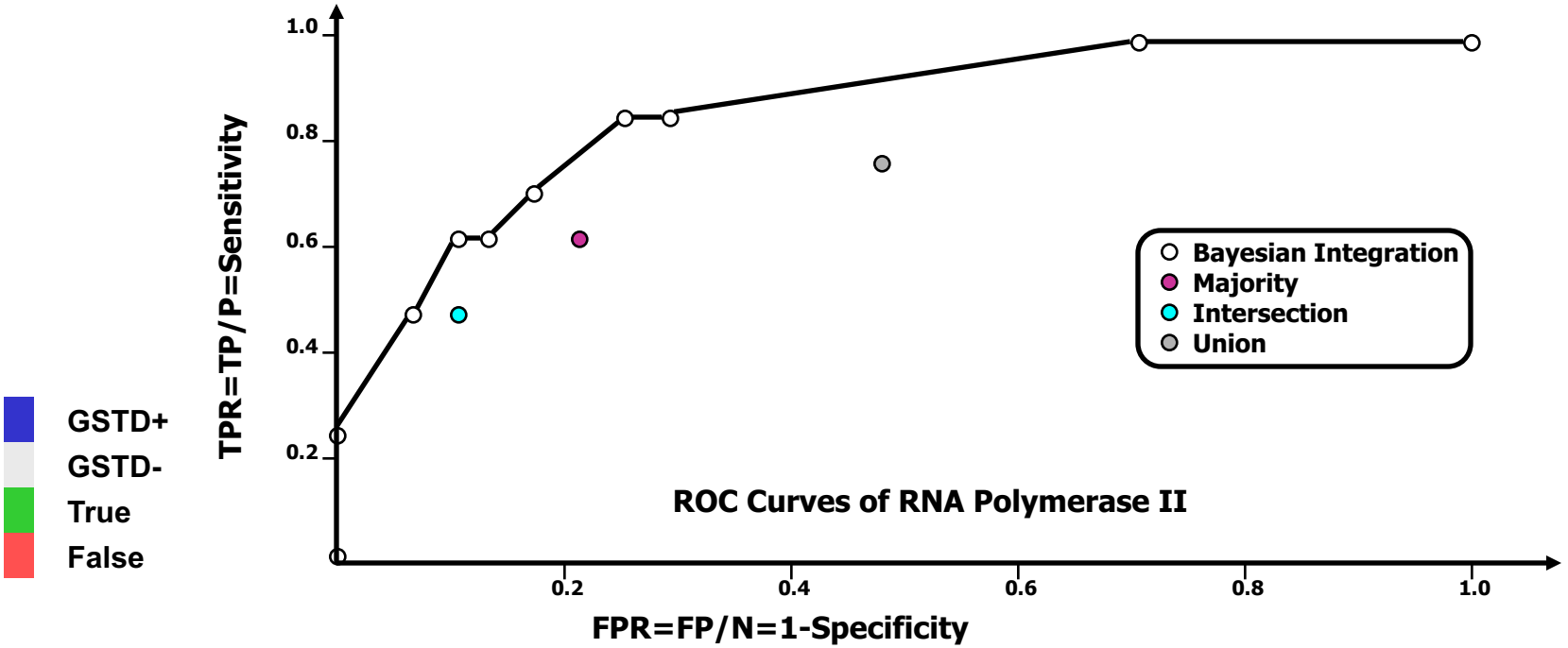
GSTD+
GSTD-
True
False

"Weighted Voting"



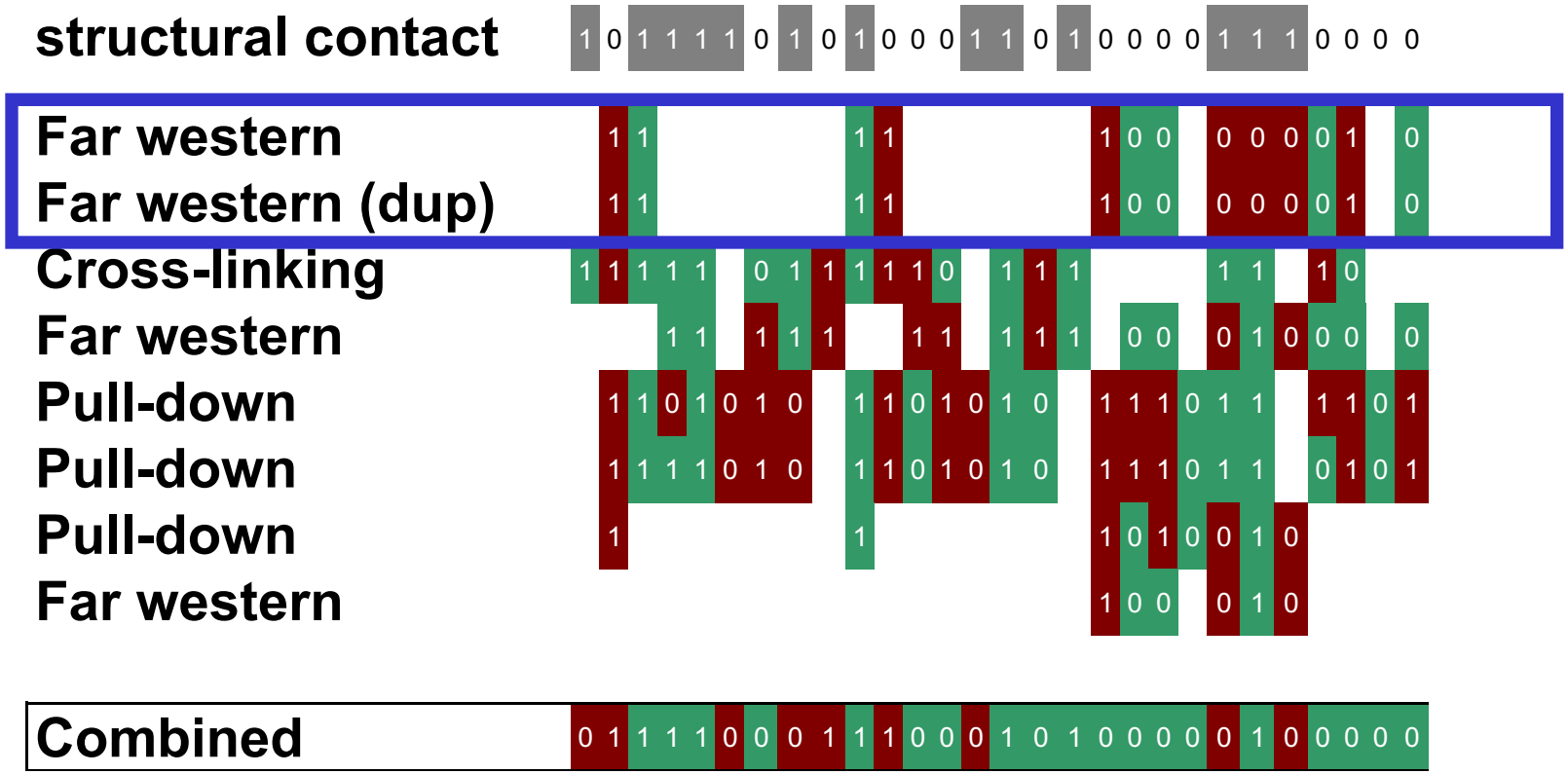
Data Integration: ROC Curve





Predicting Networks via Bayesian Integration: Features Correlation

Correlations between similar features



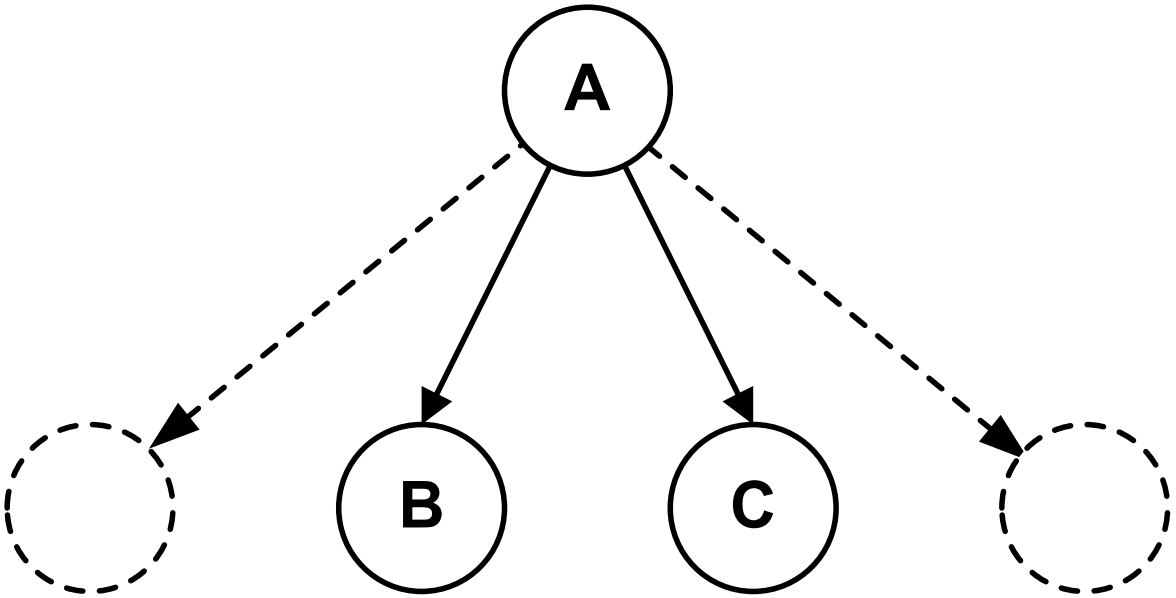
GS	**1**	**0**	**0**	**1**	**1**	**1**	**…**
F1	**1**	**0**	**0**	**1**	**0**	**0**	**…**
F2	**1**	**0**	**0**	**1**	**0**	**0**	**…**
F3	**0**	**0**	**1**	**0**	**1**	**0**	**…**
F4	**0**	**1**	**0**	**0**	**1**	**0**	**…**
F5	**0**	**1**	**0**	**0**	**1**	**0**	**…**
F6	**1**	**0**	**1**	**1**	**0**	**0**	**…**

Feature Correlation and Fully Connected Bayes

$$w_{4,5} = \log \frac{P(f_4 = 1, f_5 = 1 \mid I)}{P(f_4 = 1, f_5 = 1 \mid \sim I)}$$

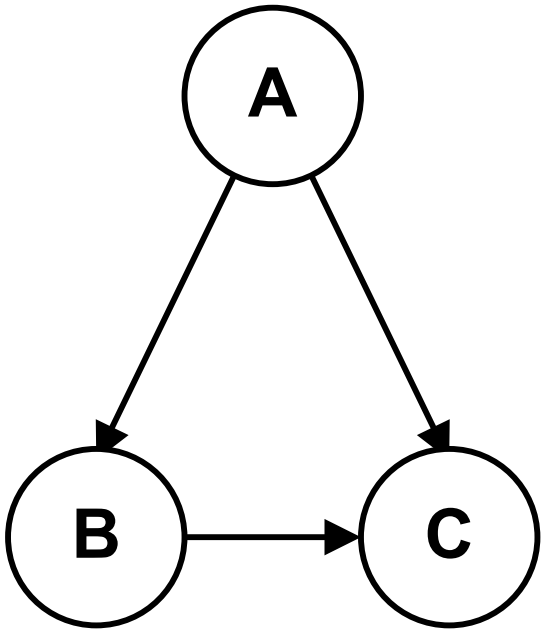
Naive Bayes

$$P(A,B,C) = P(C|A)P(B|A)P(A)$$

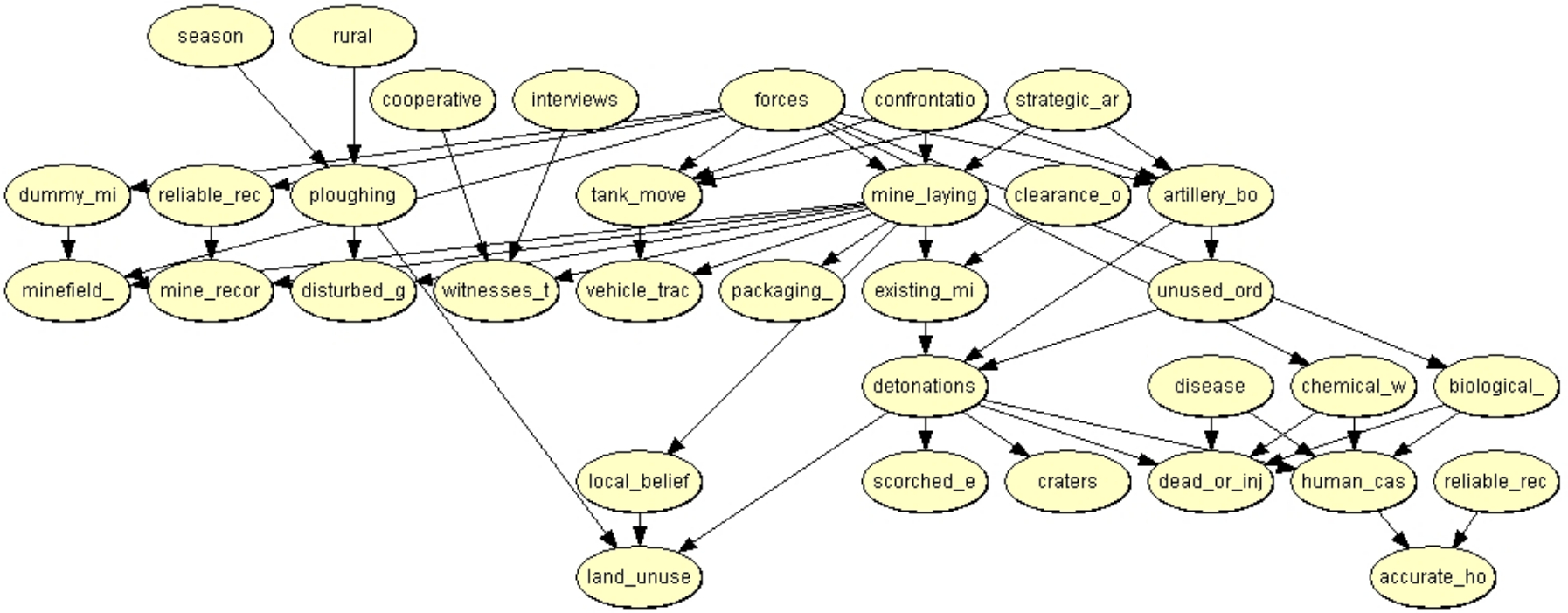


A ‘correct’ factorisation

$$P(A,B,C) = P(C|A,B)P(B|A)P(A)$$



A Typical BBN



Predicting Networks via Bayesian Integration: Real Thing but with a few features

Papers on Predicting Protein Interactions

- A Enright et al. (**1999**) "Protein interaction maps for complete genomes based on gene fusion events." Nature. 402(6757):86-90.
- E Marcotte et al. (1999) "A Combined Algorithm for Genome-Wide Prediction of Protein Function." Nature 402, 83-86 (1999).
- E Marcotte et al. (1999) "Detecting Protein Function & Protein-Protein Interactions from Genome Sequences." Science 285, 751-753
- M Pellegrini et al. (1999) "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc.Natl. Acad. Sci. 96, 4285-4288.
- R Jansen et al. (2003). "A Bayesian networks approach for predicting protein-protein interactions from genomic data." Science 302: 449-53.
- **I Lee et al. (2004) "A Probabilistic Functional Network of Yeast Genes". *Science* 206: 1555-1558**
- H Yu et al. (2004) "Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs." Genome Res 14: 1107-18.
- **L Lu et al. (2005) "Assessing the limits of genomic data integration for predicting protein networks." Genome Res 15: 945-53.**
- A Ramani et al. (2005) "Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome." Genome Biology 6:r40.
- Xia et al. (**2006**) "Integrated prediction of the helical membrane protein interactome in yeast." J Mol Biol. 357:339-49