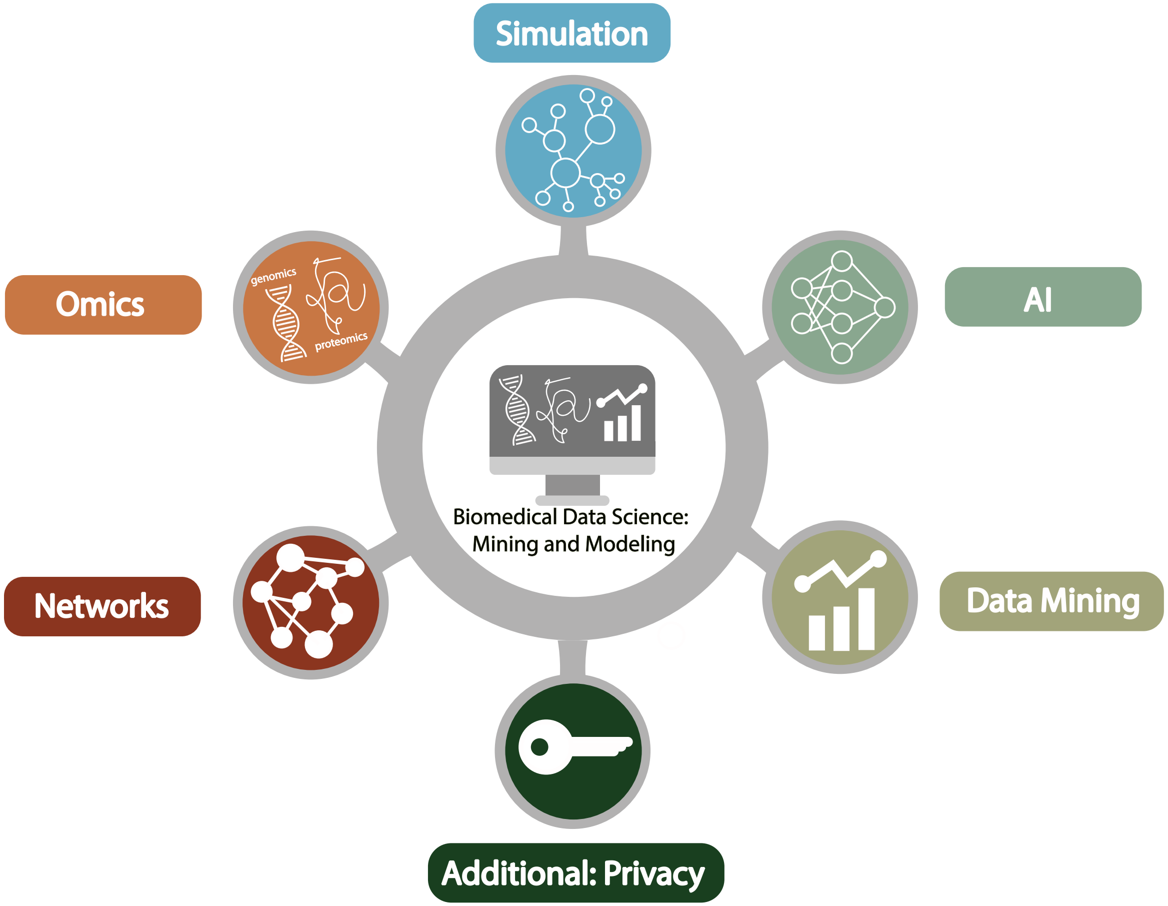
Biomedical Data Science (GersteinLab.org/courses/452)
Network Prediction (25m10e)



Mark Gerstein
Yale U.

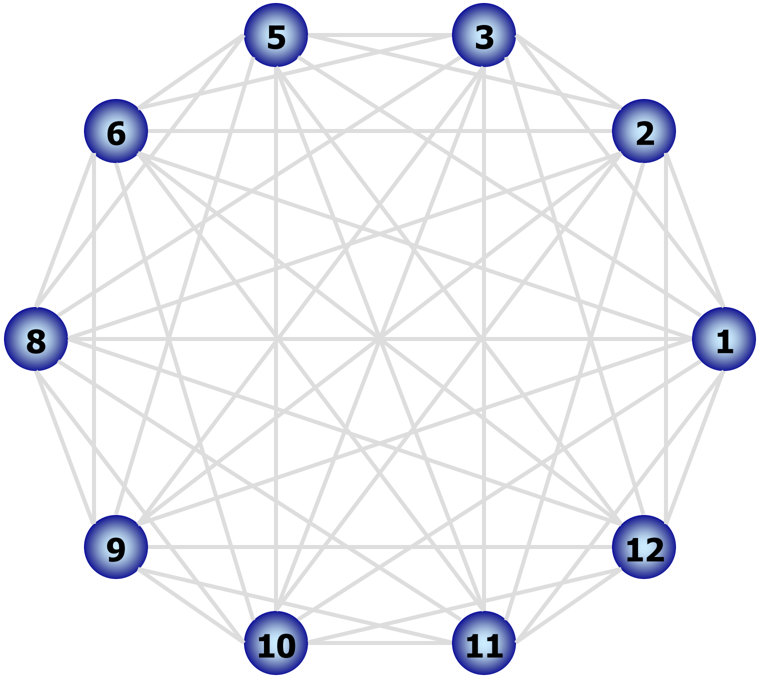
Last edit in spring '25. Substantially edited compared to '23 with a new video in '25.

Predicting Networks
via Bayesian Integration

Problem Motivation

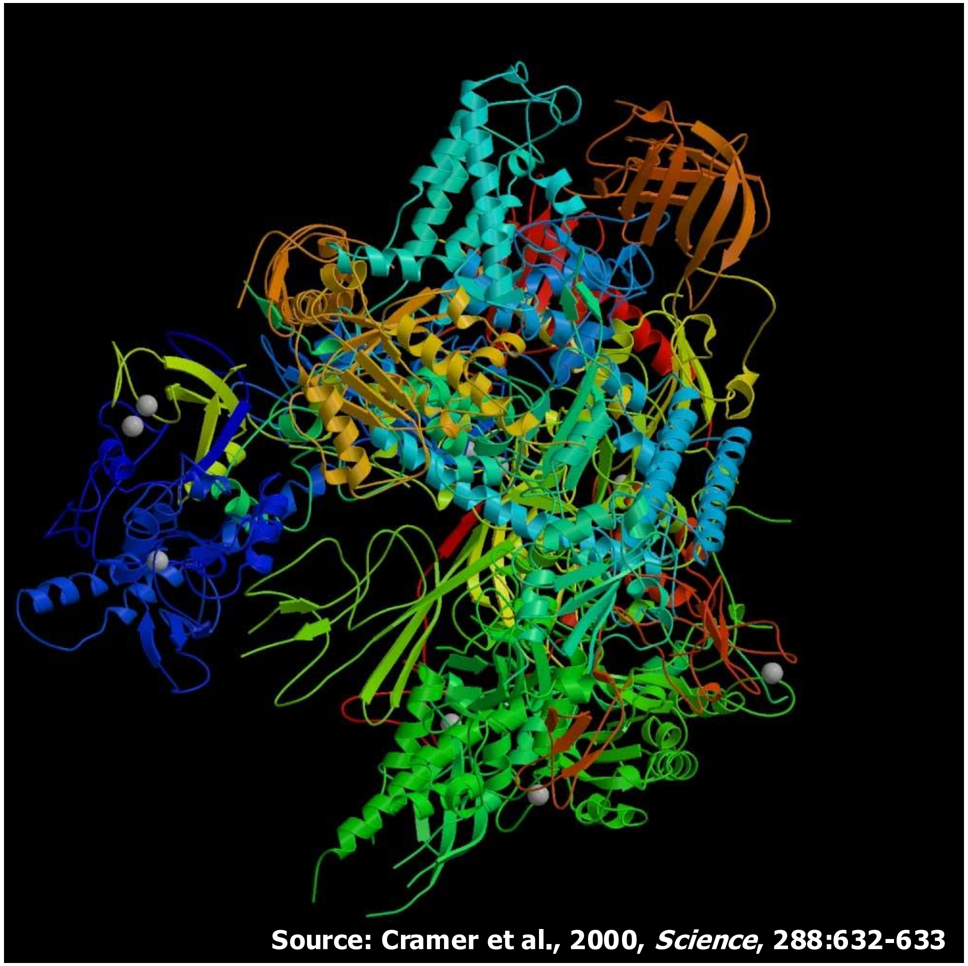
RNA polymerase II: Structure

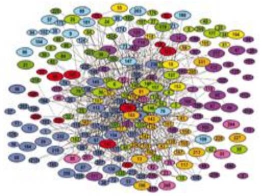
Which subunits interact? Based on Binding experiments



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

Compare with Gold Std. Structure

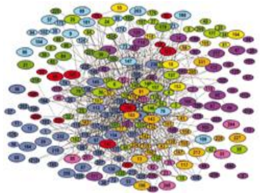

Source: Cramer et al., 2000, *Science*, 288:632-633



Binding Experiments on Subunit Pairs

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	1	0	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1		1	1	0	1	0		0	0	0	0		0	1	0		0	0		0		
Pull-down 2		1	1	1	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1		0	1	0	1	0		0	0	0	0		0	0	0		0	0		0		
Pull-down 3		1								1								1	0	1	0	0	1	0																					
Cross-linking	1	1	1	1	1		0	1	1	1	1	1	0		1	1	1					1	1		1	0					1														
Far Western 1		1	1							1	1							1	0	0		0	0	0	0	1		0	0	0															
Far Western 2				1	1		1	1	1			1	1		1	1	1		0	0		0	1	0	0	0		0	0	0	0		0	0	0		0	0	0				0	0	0
Far Western 3																		1	0	0		0	1	0																					

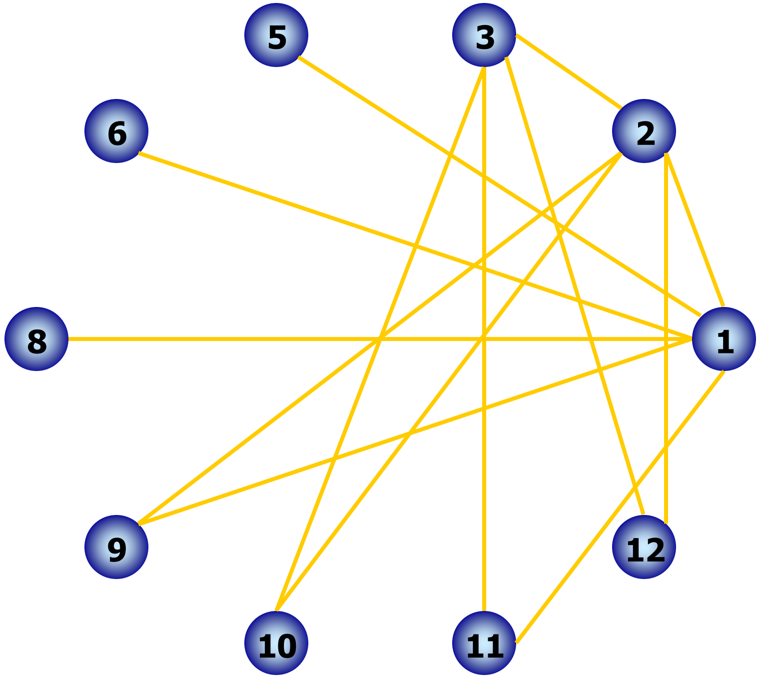
Interaction experiments
before structure was known



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

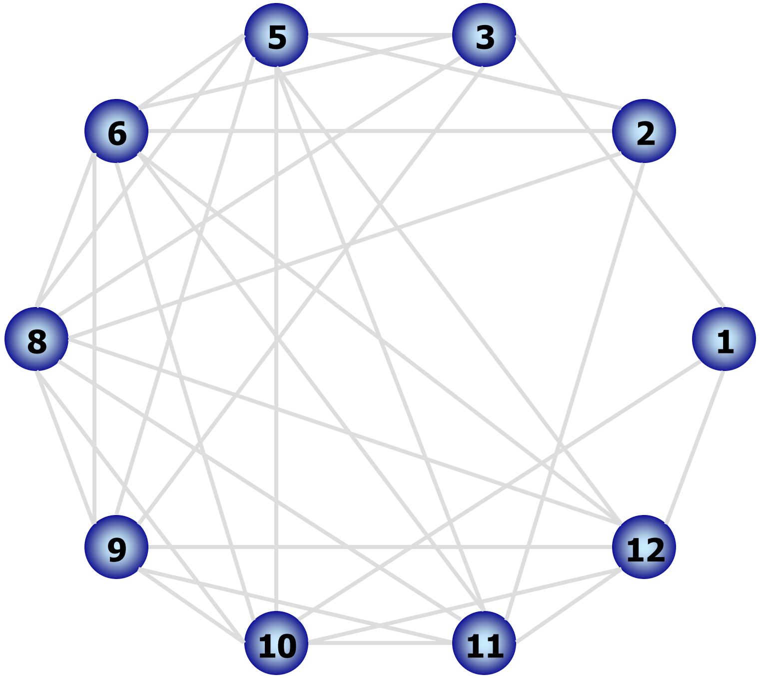
Gold-Standard Positive (GSTD+): 13



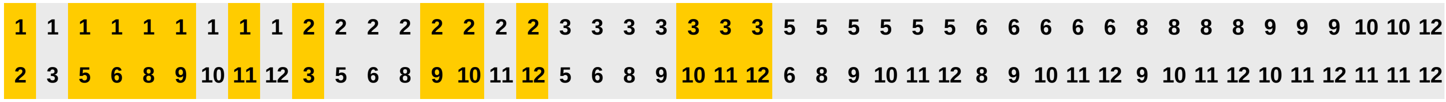
Gold-Standard Negatives

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



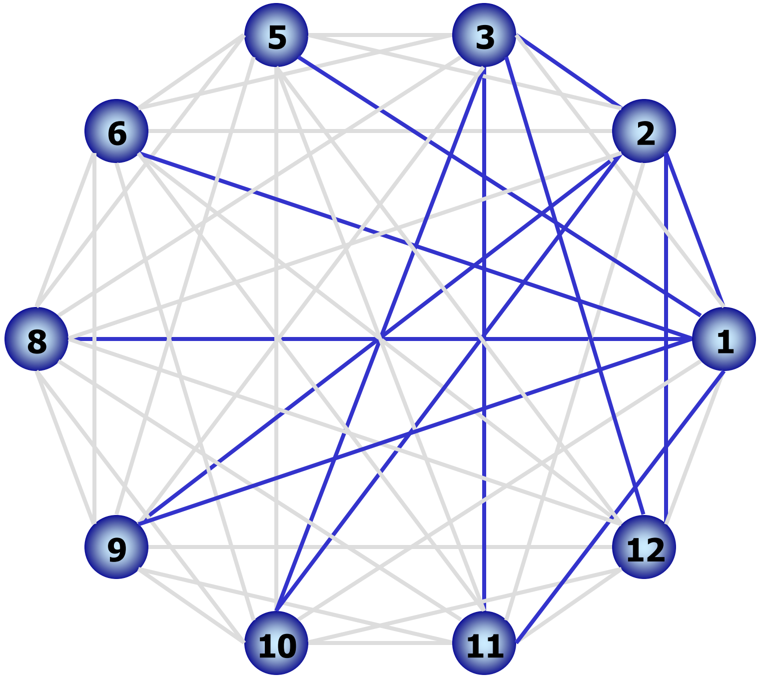
RNA Polymerase II: Gold-Standards

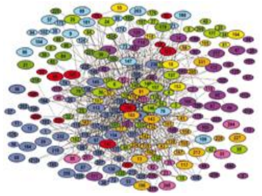


Subunits	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	5	5	5	5	5	5	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12

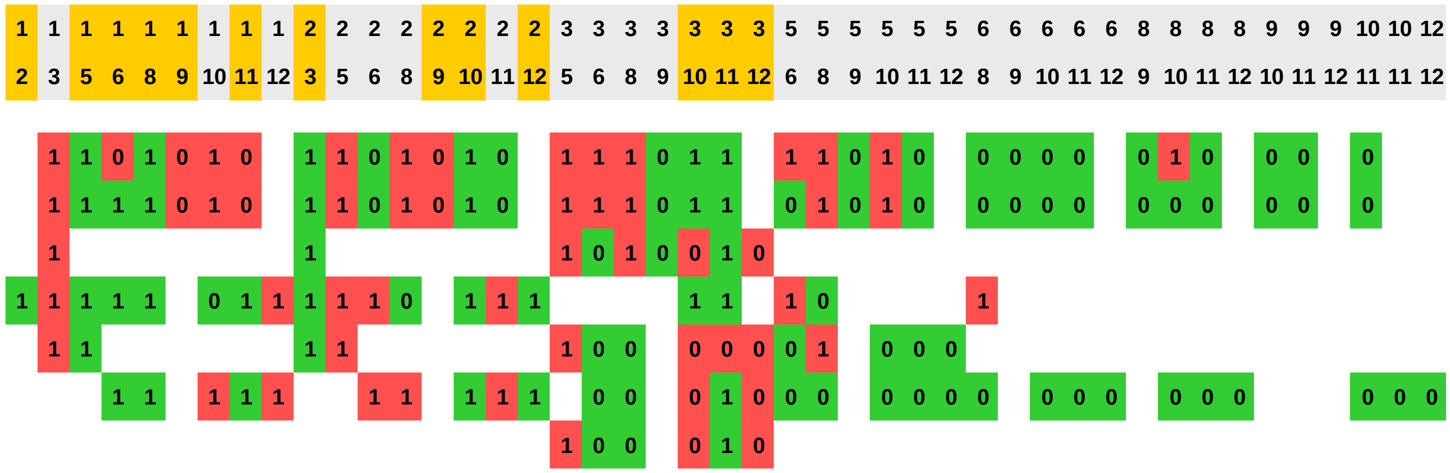
Gold-Standard Positive (GSTD+): 13

Gold-Standard Negative (GSTD-): 32





Assess Quality and Coverage of PPints



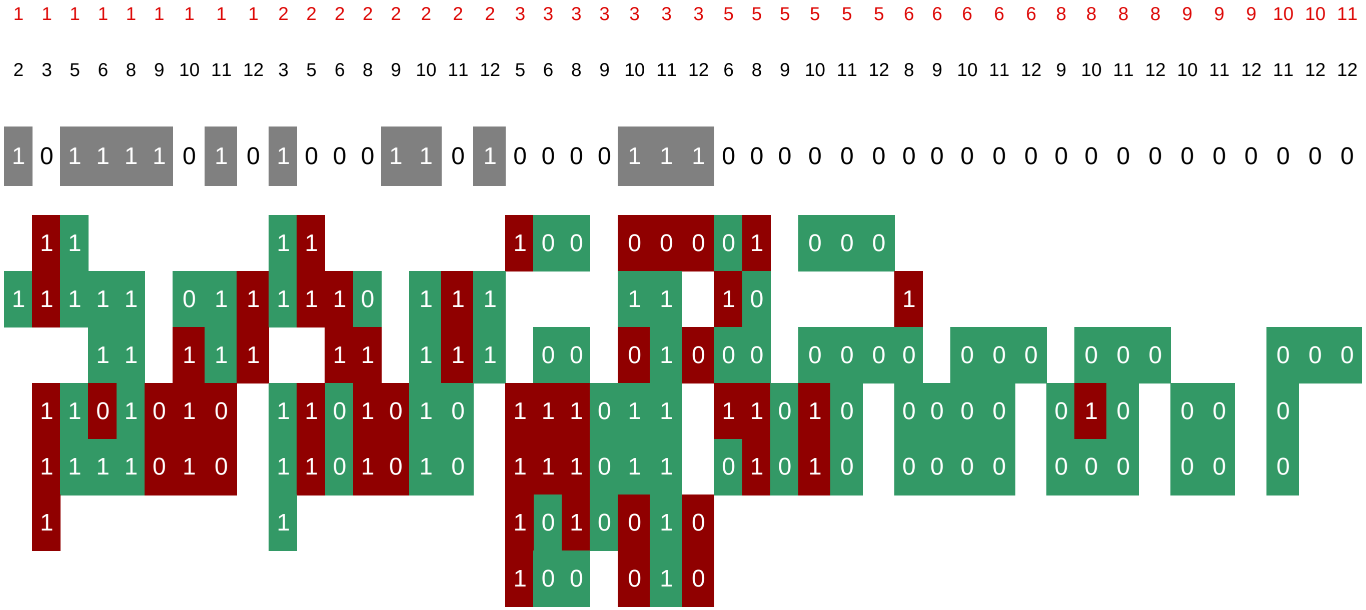
Subunits	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
Subunits	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10
Pull-down 1		1	1	0	1	0	1	0		1	1	0	1	0	1
Pull-down 2		1	1	1	1	0	1	0		1	1	0	1	0	1
Pull-down 3		1								1					
Cross-linking	1	1	1	1	1		0	1	1	1	1	1	0		1
Far Western 1		1	1							1	1				
Far Western 2				1	1		1	1	1			1	1	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

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	
Pull-down 2	0		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	1		1	0				
Far Western 1			1	0	0		0	0	0	0	1		0	0	0
Far Western 2	1	1		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															

🟦 GSTD+
⬜ GSTD-
🟩 True
🟥 False

Re-drawn in a different color scheme



Subunit A	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
Subunit B	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10
structural contact	1	0	1	1	1	1	0	1	0	1	0	0	0	1	1
Far western		1	1							1	1				
Cross-linking	1	1	1	1	1		0	1	1	1	1	1	0		1
Far western				1	1		1	1	1			1	1		1
Pull-down		1	1	0	1	0	1	0		1	1	0	1	0	1
Pull-down		1	1	1	1	0	1	0		1	1	0	1	0	1
Pull-down		1								1					
Far western															

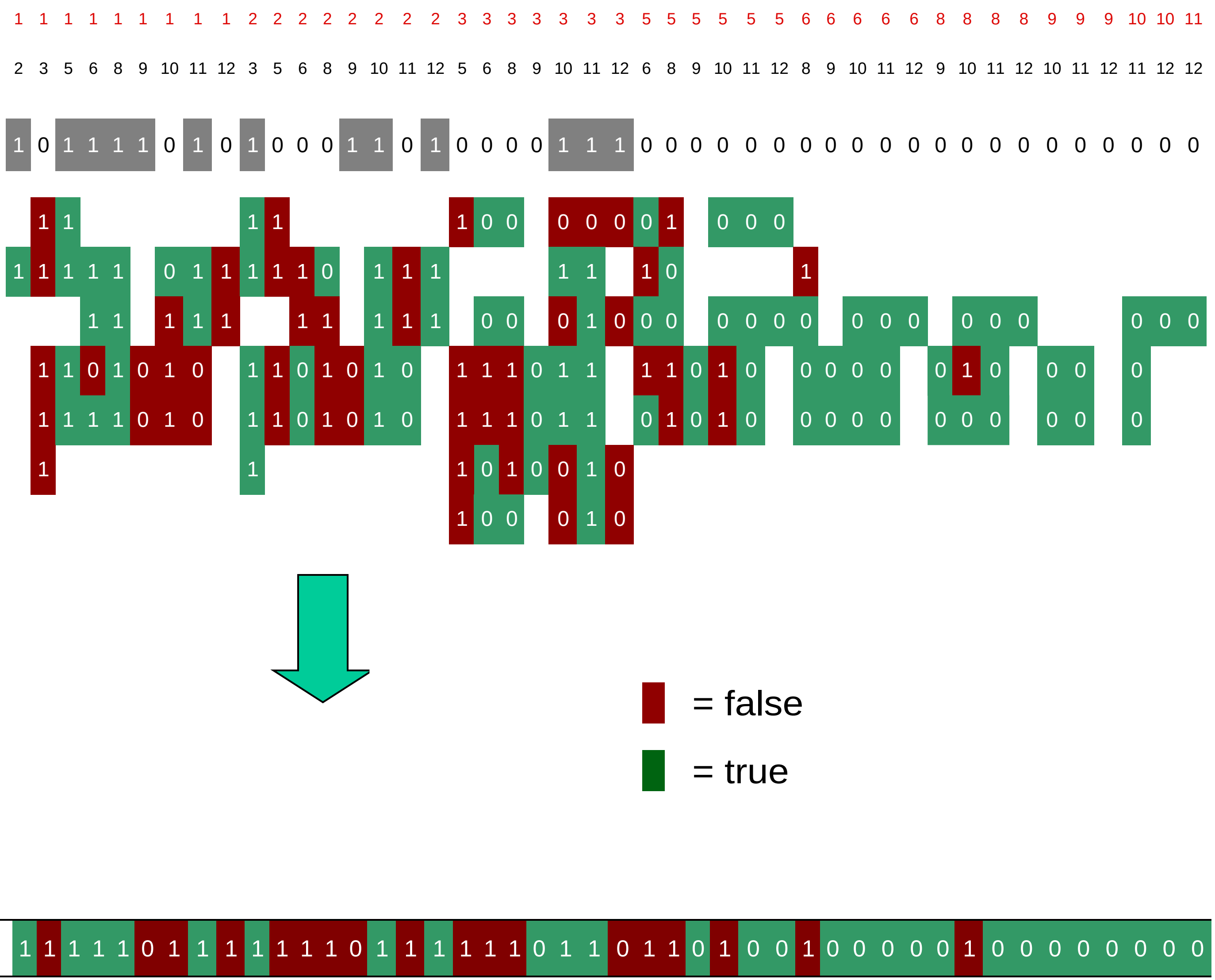
Subunit A	2	2	3	3	3	3	3	3	3	5	5	5	5	5	5
Subunit B	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12
structural contact	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0
Far western			1	0	0		0	0	0	0	1		0	0	0
Cross-linking	1	1					1	1		1	0				
Far western	1	1		0	0		0	1	0	0	0		0	0	0
Pull-down	0		1	1	1	0	1	1		1	1	0	1	0	
Pull-down	0		1	1	1	0	1	1		0	1	0	1	0	
Pull-down			1	0	1	0	0	1	0						
Far western			1	0	0		0	1	0						

Subunit A	6	6	6	6	6	8	8	8	8	9	9	9	10	10	11
Subunit B	8	9	10	11	12	9	10	11	12	10	11	12	11	12	12
structural contact	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Far western															
Cross-linking	1														
Far western	0		0	0	0		0	0	0				0	0	0
Pull-down	0	0	0	0		0	1	0		0	0		0		
Pull-down	0	0	0	0		0	0	0		0	0		0		
Pull-down															
Far western															

🟥 = false

🟩 = true

Data integration: RNA polymerase II

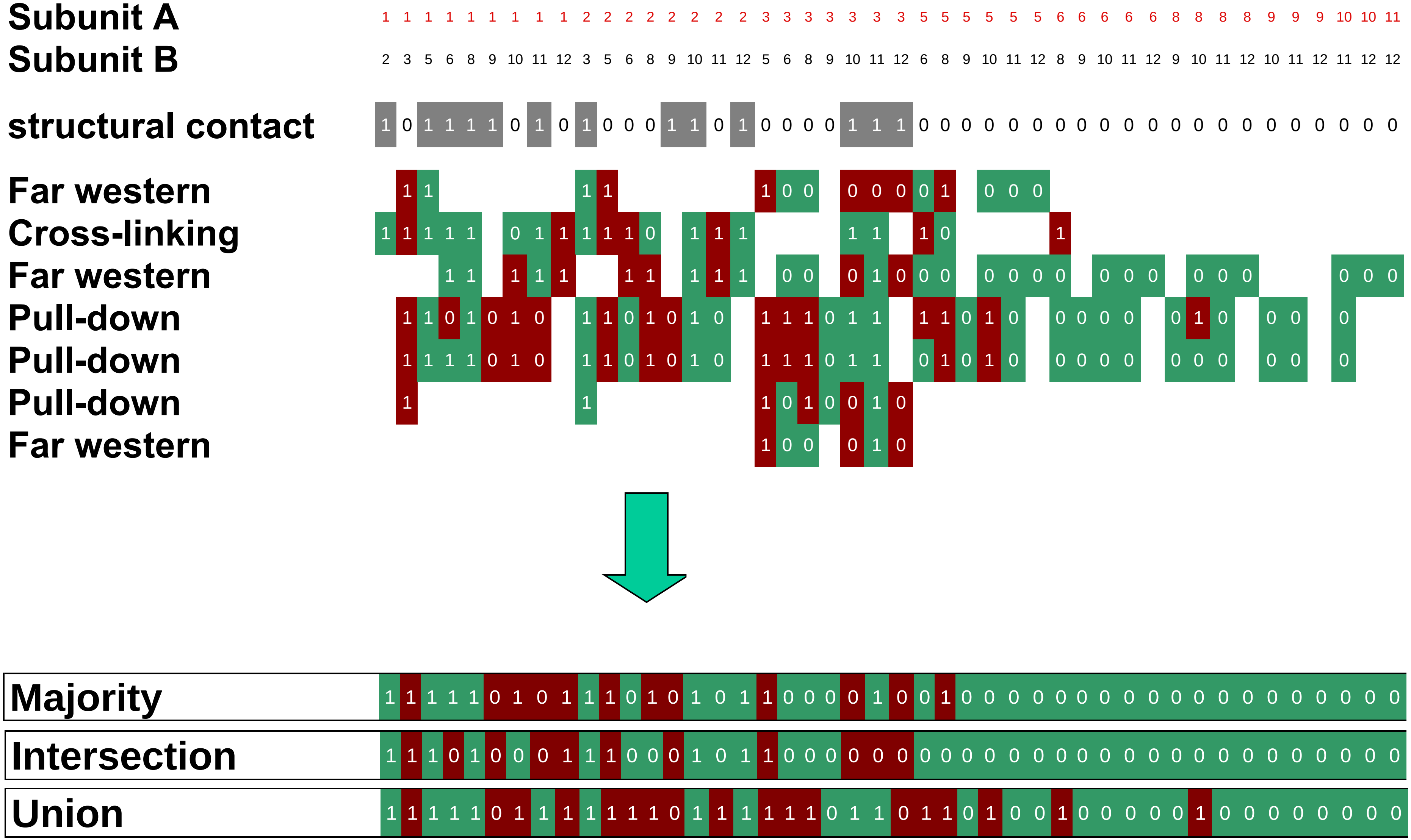


Subunit A	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	5	5	5	5	5	5	6	6	6	6	6	8	8	8	8	9	9	9	10	10	11
Subunit B	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12	8	9	10	11	12	9	10	11	12	10	11	12	11	12	12
structural contact	1	0	1	1	1	1	0	1	0	1	0	0	0	1	1	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Far western		1	1							1	1							1	0	0		0	0	0	0	1		0	0	0															
Cross-linking	1	1	1	1	1		0	1	1	1	1	1	0		1	1	1					1	1		1	0					1														
Far western				1	1		1	1	1			1	1		1	1	1		0	0		0	1	0	0	0		0	0	0	0		0	0	0		0	0	0				0	0	0
Pull-down		1	1	0	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1		1	1	0	1	0		0	0	0	0		0	1	0		0	0		0		
Pull-down		1	1	1	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1		0	1	0	1	0		0	0	0	0		0	0	0		0	0		0		
Pull-down		1								1								1	0	1	0	0	1	0																					
Far western																		1	0	0		0	1	0																					
Union	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0

🟥 = false

🟩 = true

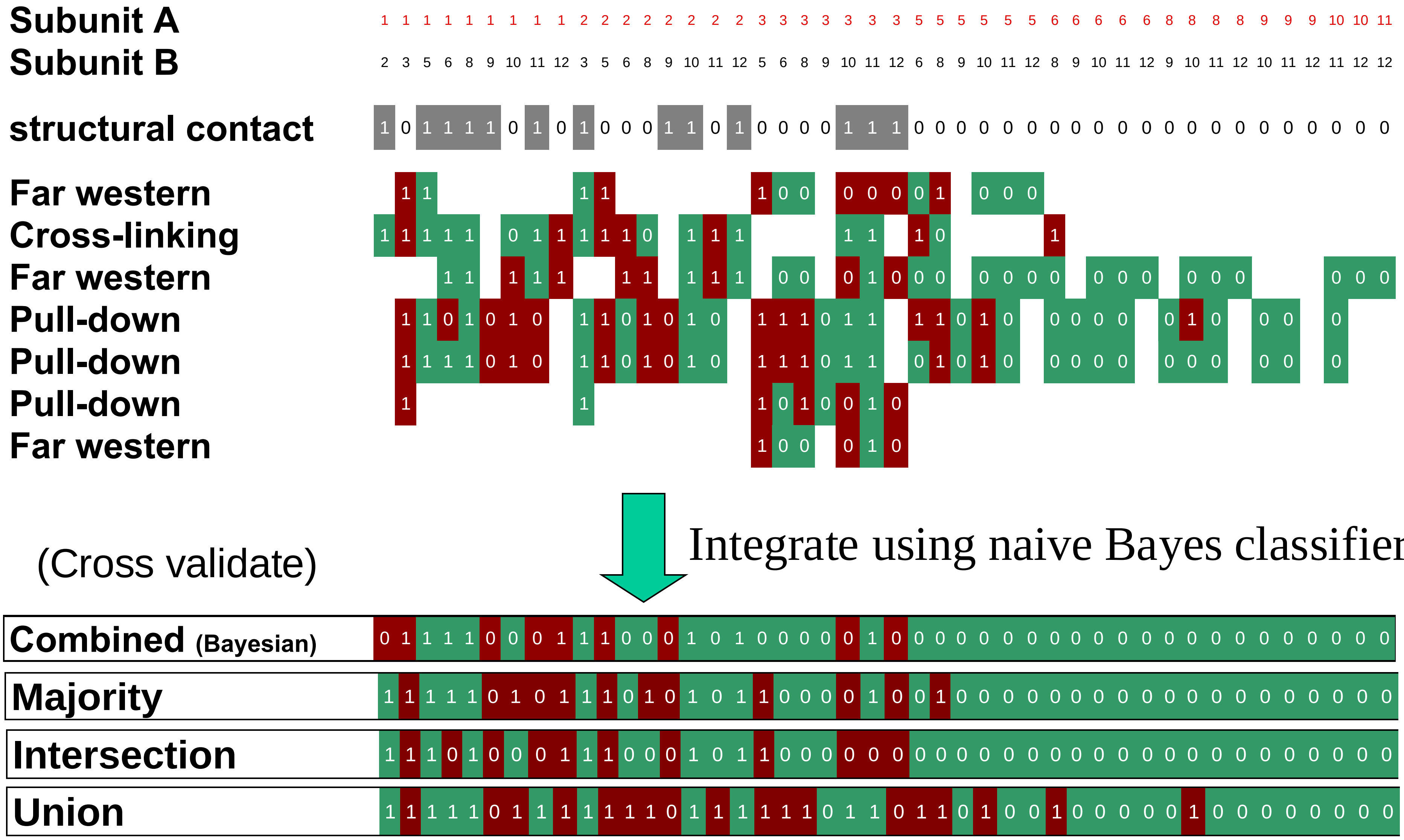
Data integration: RNA polymerase II



Subunit A	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	11
Subunit B	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	12	12
structural contact	1	0	1	1	1	1	0	1	0	1	0	0	0	1	1	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Far western		1	1							1	1							1	0	0		0	0	0	0	1		0	0	0															
Cross-linking	1	1	1	1	1		0	1	1	1	1	1	0		1	1	1					1	1		1	0					1														
Far western				1	1		1	1	1			1	1		1	1	1		0	0		0	1	0	0	0		0	0	0	0		0	0	0		0	0	0				0	0	0
Pull-down		1	1	0	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1		1	1	0	1	0		0	0	0	0		0	1	0		0	0		0		
Pull-down		1	1	1	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1		0	1	0	1	0		0	0	0	0		0	0	0		0	0		0		
Pull-down		1								1								1	0	1	0	0	1	0																					
Far western																		1	0	0		0	1	0																					

Majority	1	1	1	1	1	0	1	0	1	1	1	0	1	0	1	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Intersection	1	1	1	0	1	0	0	0	1	1	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Union	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Data integration: RNA polymerase II



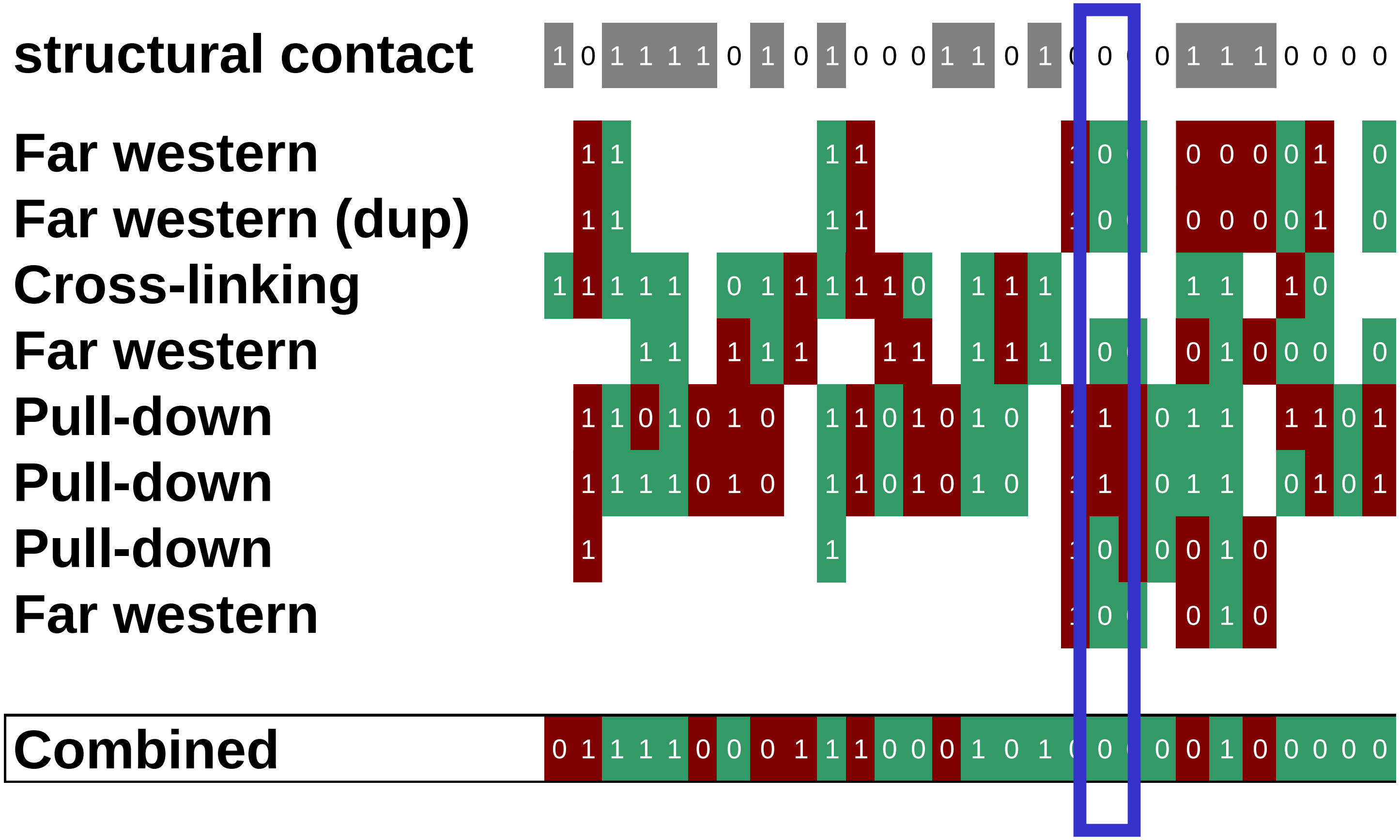
Subunit A	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	5	5	5	5	5	5	6	6	6	6	6	8	8	8	8	9	9	9	10	10	11
Subunit B	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12	8	9	10	11	12	9	10	11	12	10	11	12	11	12	12
structural contact	1	0	1	1	1	1	0	1	0	1	0	0	0	1	1	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Far western		1	1							1	1							1	0	0		0	0	0	0	1		0	0	0															
Cross-linking	1	1	1	1	1		0	1	1	1	1	1	0		1	1	1					1	1		1	0					1														
Far western				1	1		1	1	1			1	1		1	1	1		0	0		0	1	0	0	0		0	0	0	0		0	0	0		0	0	0				0	0	0
Pull-down		1	1	0	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1		1	1	0	1	0		0	0	0	0		0	1	0		0	0		0		
Pull-down		1	1	1	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1		0	1	0	1	0		0	0	0	0		0	0	0		0	0		0		
Pull-down		1								1								1	0	1	0	0	1	0																					
Far western																		1	0	0		0	1	0																					

Integrate using naive Bayes classifier

(Cross validate)

Combined (Bayesian)	0	1	1	1	1	0	0	0	1	1	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Majority	1	1	1	1	1	0	1	0	1	1	1	0	1	0	1	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Intersection	1	1	1	0	1	0	0	0	1	1	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Union	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Weighted Voting: the Likelihood Ratio



structural contact 1 0 1 1 1 1 0 1 0 1 0 0 0 1 1 0 1 0 0 1 1 1 0 0 0 0

Far western 1 1 1 1 0 0 0 0 0 1 0

Far western (dup) 1 1 1 1 0 0 0 0 0 1 0

Cross-linking 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 0

Far western 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0

Pull-down 1 1 0 1 0 1 0 1 1 0 1 0 1 0 1 0 1 1 1 1 0 1

Pull-down 1 1 1 1 0 1 0 1 1 0 1 0 1 0 1 0 1 1 0 1 0 1

Pull-down 1 1 0 0 0 1 0

Far western 0 0 1 0

Combined 0 1 1 1 1 0 0 0 1 1 1 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0

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

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

Predicting Networks via Bayesian Integration

Formalism

Similar form to many other classifiers – e.g. SVC, LDA, logistic (see ISLR for comparison)

Supervised Classification by Weighted Voting

Simple Vote: $R = f_1 + f_2 + f_3 + ... + f_n$ 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:

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

If has prior knowledge w_0

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

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]

More Bayes Rule

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

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

$$w_0 = \log \frac{P}{N}$$

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

Assume independent (Naïve Bayes)

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

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

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

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

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

Update form of Bayes Rule

Updating our prior odds (w_0) by successively adding likelihood ratio terms (w_1,w_2,w_3) for each feature to get a final posterior odds

$$\log\left(\frac{P(I \mid f_1, f_2, f_3, \ldots)}{P(\sim I \mid f_1, f_2, f_3, \ldots)}\right) = \underset{\substack{\uparrow \\ w_1}}{\log\frac{TPR_1}{FPR_1}} + \underset{\substack{\uparrow \\ w_2}}{\log\frac{TPR_2}{FPR_2}} + \underset{\substack{\uparrow \\ w_3}}{\log\frac{TPR_3}{FPR_3}} + \ldots + \underset{\substack{\uparrow \\ w_0}}{\log\frac{P}{N}}$$

This simple additive update form supposedly, secretly invented by A Turing during WWII [Spiegelhalter & Brooks, The Art of Uncertainty ('25)]

Estimating Probabilities & Dummy Counts

- 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 $X=x$

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

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

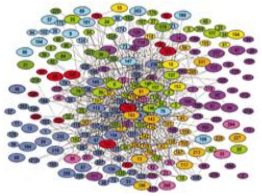
$$\frac{n_{x|y} + mp}{n_y + m}$$

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
via Bayesian Integration

Worked Example
(for reference)



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		

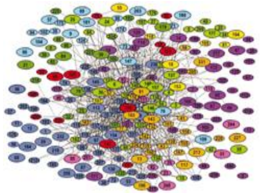
$$L_1 = \frac{p\left(x_1 \mid GSTD+\right)}{p\left(x_1 \mid GSTD-\right)}$$

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

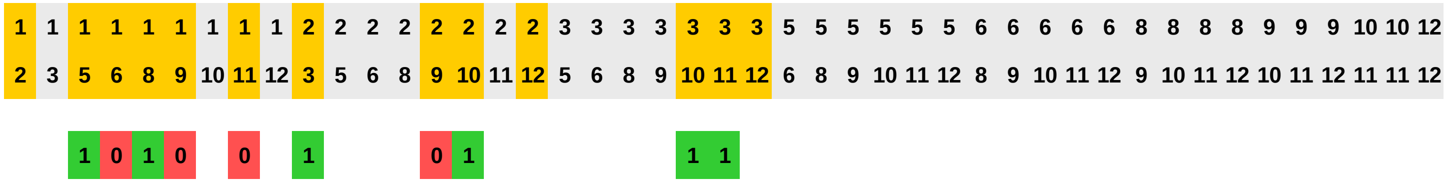
Likelihood Ratio
for Feature *f*:

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

🟦 **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																						

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

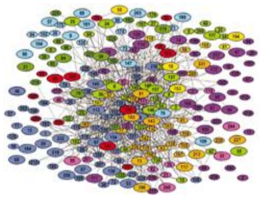
$$L_0 = \frac{p(x_0 \mid GSTD+)}{p(x_0 \mid GSTD-)} = \frac{4/13}{}$$

🟦 **GSTD+**

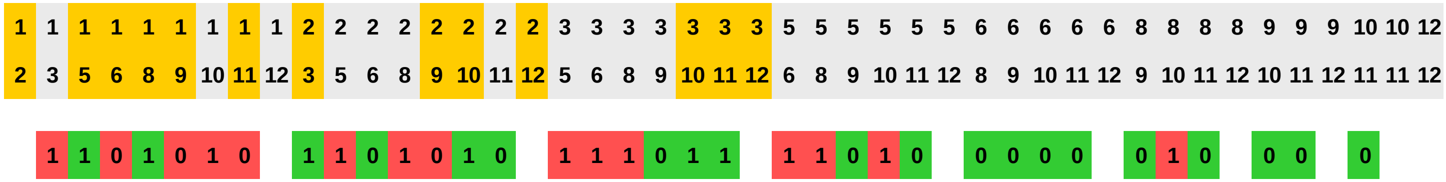
⬜ **GSTD-**

🟩 **True**

🟥 **False**



Calculating Likelihood Ratios

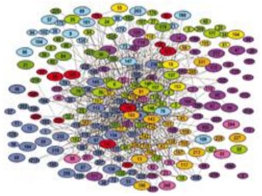


Subunits	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	5	5	5	5	5	5	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12
Pull-down 1		🟥 1	🟩 1	🟥 0	🟩 1	🟥 0	🟥 1	🟥 0		🟩 1	🟥 1	🟩 0	🟥 1	🟥 0	🟩 1	🟩 0		🟥 1	🟥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

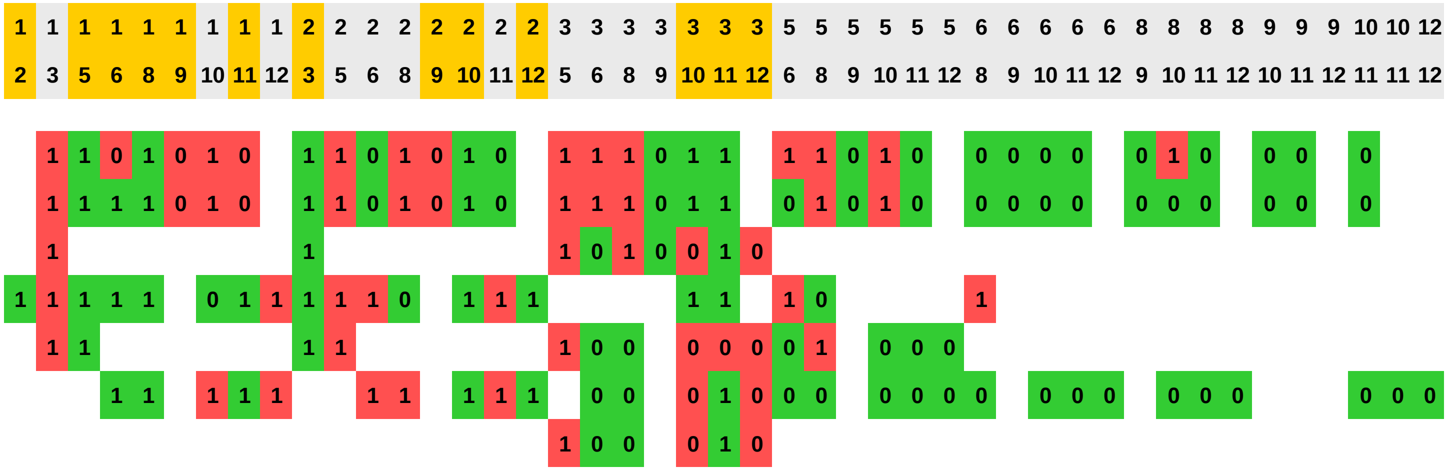
$$L_1 = \frac{p\left(x_1 \mid GSTD+\right)}{p\left(x_1 \mid GSTD-\right)} = \frac{6/13}{11/32} = 1.34$$

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

🟦 GSTD+
⬜ GSTD-
🟩 True
🟥 False



Calculating Likelihood Ratios

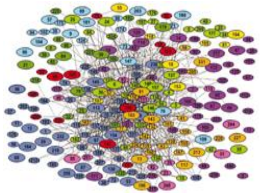


Subunits	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3
Subunits	2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12	5	6	8	9	10	11
Pull-down 1		1	1	0	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1
Pull-down 2		1	1	1	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1
Pull-down 3		1								1								1	0	1	0	0	1
Cross-linking	1	1	1	1	1		0	1	1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

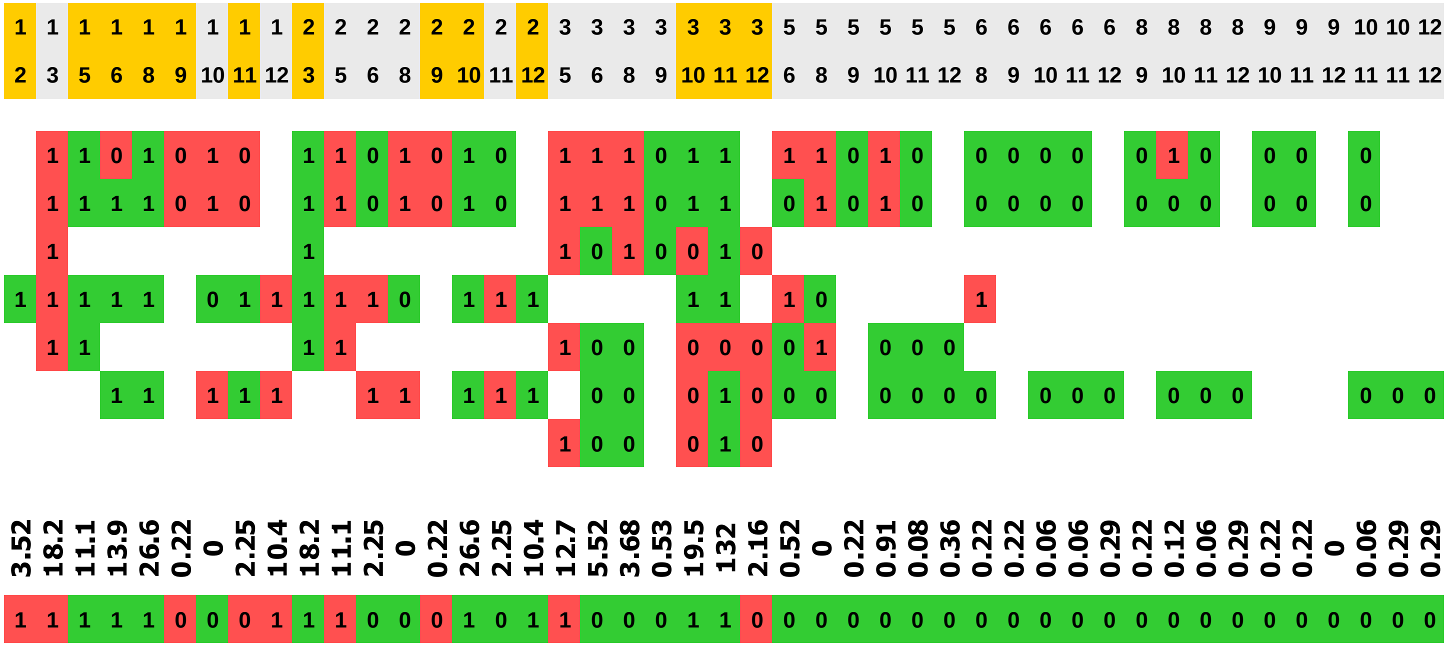
Subunits	3	5	5	5	5	5	5	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	12	6	8	9	10	11	12	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12
Pull-down 1		1	1	0	1	0		0	0	0	0		0	1	0		0	0		0		
Pull-down 2		0	1	0	1	0		0	0	0	0		0	0	0		0	0		0		
Pull-down 3	0																					
Cross-linking		1	0					1														
Far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

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



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
Pull-down 2		1	1	1	1	0	1	0		1	1	0	1	0	1
Pull-down 3		1								1					
Cross-linking	1	1	1	1	1		0	1	1	1	1	1	0		1
Far Western 1		1	1							1	1				
Far Western 2				1	1		1	1	1			1	1		1
Far Western 3															
	3.52	18.2	11.1	13.9	26.6	0.22	0	2.25	10.4	18.2	11.1	2.25	0	0.22	26.6
Combined (Bayes)	1	1	1	1	1	0	0	0	1	1	1	0	0	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	
Pull-down 2	0		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	1		1	0				
Far Western 1			1	0	0		0	0	0	0	1		0	0	0
Far Western 2	1	1		0	0		0	1	0	0	0		0	0	0
Far Western 3			1	0	0		0	1	0						
	2.25	10.4	12.7	5.52	3.68	0.53	19.5	132	2.16	0.52	0	0.22	0.91	0.08	0.36
Combined (Bayes)	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0

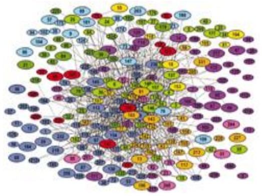
Subunits	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
Subunits	8	9	10	11	12	9	10	11	12	10	11	12	11	11	12
Pull-down 1	0	0	0	0		0	1	0		0	0		0		
Pull-down 2	0	0	0	0		0	0	0		0	0		0		
Pull-down 3															
Cross-linking	1														
Far Western 1															
Far Western 2	0		0	0	0		0	0	0				0	0	0
Far Western 3															
	0.22	0.22	0.06	0.06	0.29	0.22	0.12	0.06	0.29	0.22	0.22	0	0.06	0.29	0.29
Combined (Bayes)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

↑

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

- 🟦 **GSTD+**
- ⬜ **GSTD-**
- 🟩 **True**
- 🟥 **False**

“Weighted Voting”



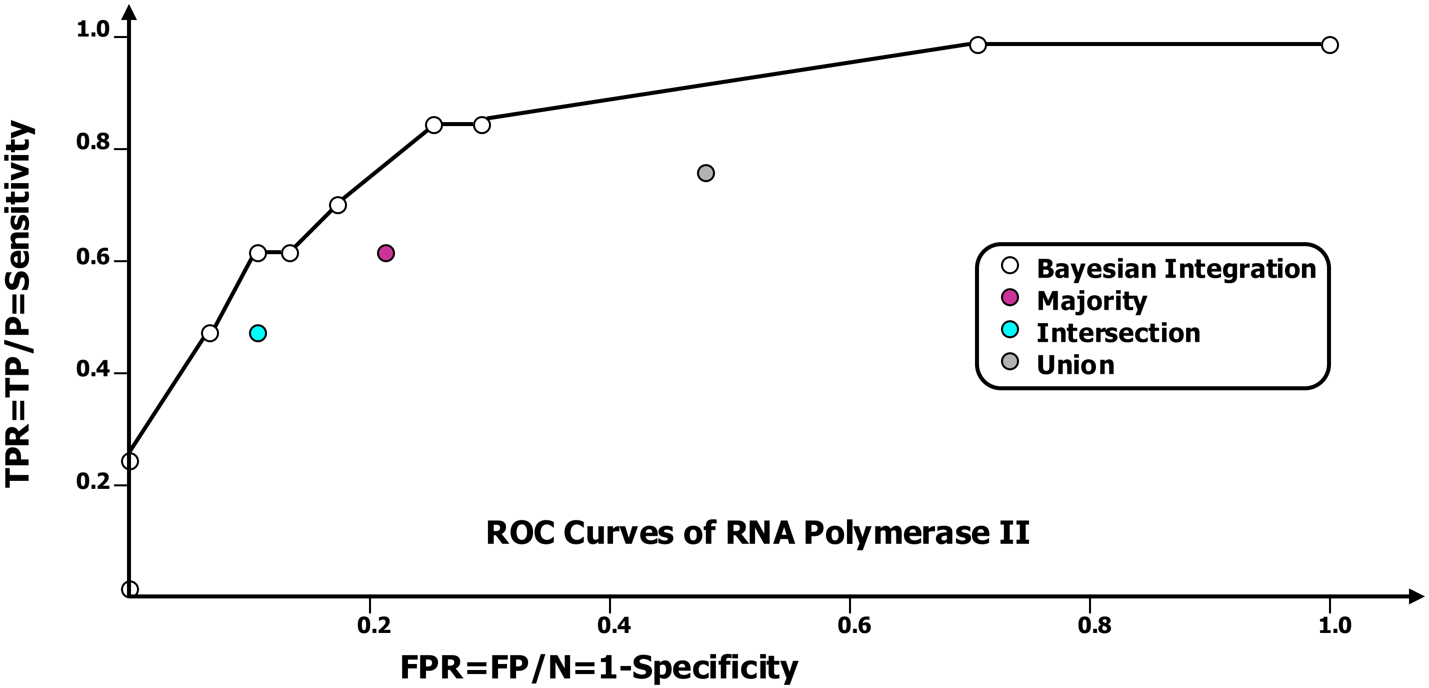
Data Integration: ROC Curve

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
	3.52	18.2	11.1	13.9	26.6	0.22	0	2.25	10.4	18.2	11.1	2.25	0	0.22	26.6
Combined (Bayes)	1	1	1	1	1	0	0	0	1	1	1	0	0	0	1
Majority	1	1	1	1	1	0	1	0	1	1	1	0	1	0	1
Intersection	1	1	1	0	1	0	0	0	1	1	1	0	0	0	1
Union	1	1	1	1	1	0	1	1	1	1	1	1	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
	2.25	10.4	12.7	5.52	3.68	0.53	19.5	132	2.16	0.52	0	0.22	0.91	0.08	0.36
Combined (Bayes)	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0
Majority	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0
Intersection	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Union	1	1	1	1	1	0	1	1	0	1	1	0	1	0	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
	0.22	0.22	0.06	0.06	0.29	0.22	0.12	0.06	0.29	0.22	0.22	0	0.06	0.29	0.29
Combined (Bayes)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Majority	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Intersection	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Union	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0

🟦 **GSTD+**
⬜ **GSTD-**
🟩 **True**
🟥 **False**

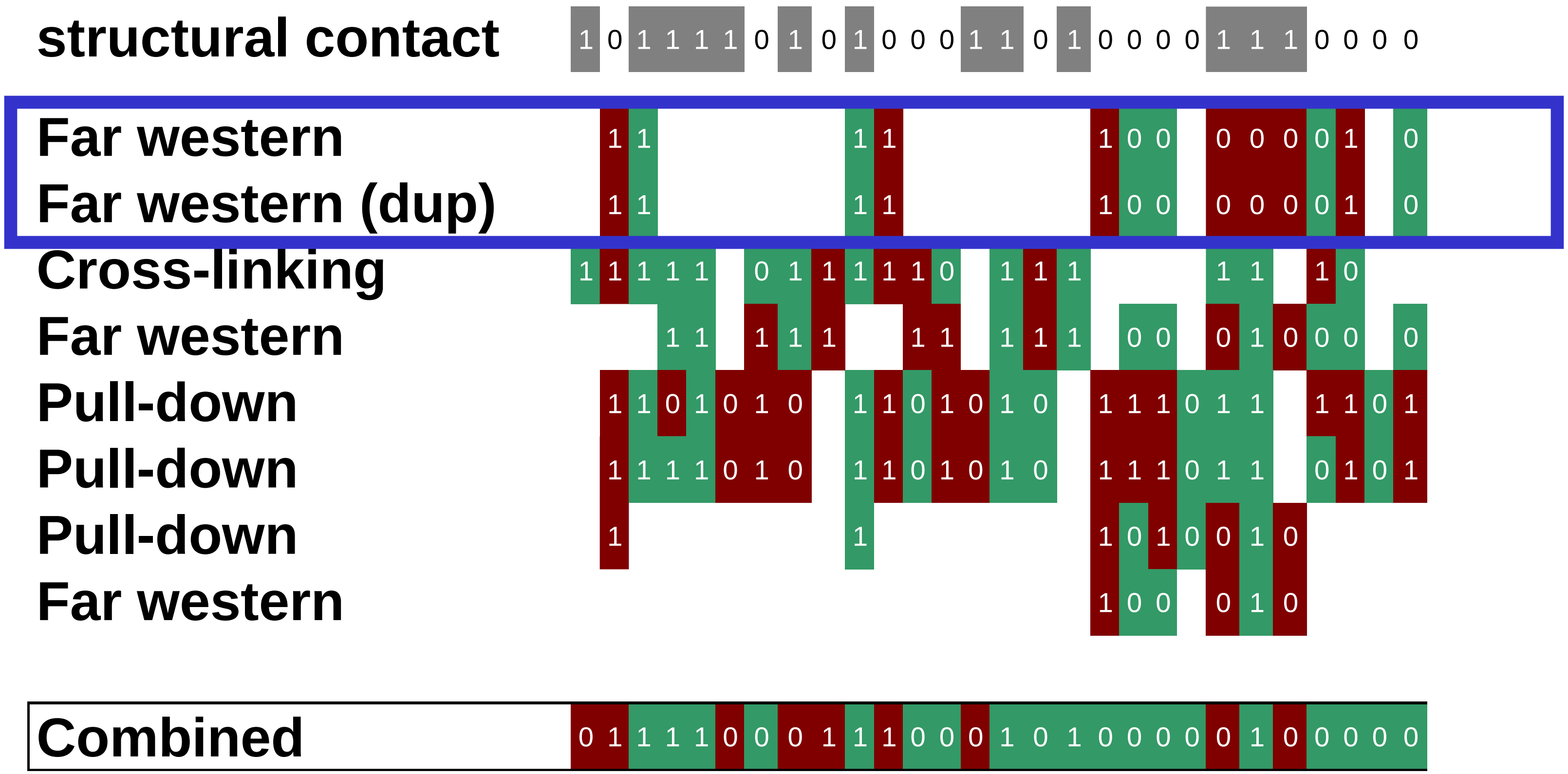


ROC Curves of RNA Polymerase II

Predicting Networks
via Bayesian Integration

Feature Correlation

Correlations between similar features



structural contact	1	0	1	1	1	1	0	1	0	1	0	0	0	1	1	0	1	0	0	0	0	1	1	1	0	0	0	0
Far western		1	1							1	1							1	0	0		0	0	0	0	1		0
Far western (dup)		1	1							1	1							1	0	0		0	0	0	0	1		0
Cross-linking	1	1	1	1	1		0	1	1	1	1	1	0		1	1	1					1	1		1	0		
Far western				1	1		1	1	1			1	1		1	1	1		0	0		0	1	0	0	0		0
Pull-down		1	1	0	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1		1	1	0	1
Pull-down		1	1	1	1	0	1	0		1	1	0	1	0	1	0		1	1	1	0	1	1		0	1	0	1
Pull-down		1								1								1	0	1	0	0	1	0				
Far western																		1	0	0		0	1	0				
Combined	0	1	1	1	1	0	0	0	1	1	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0

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

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

Predicting Networks
via Bayesian Integration

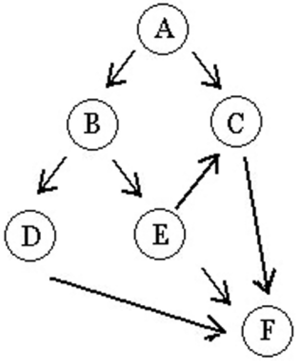
Bayesian Networks

Bayesian network is a model that represent the relationship between variables

lectures.gersteinlab.org - [Gao et al. ('24) Bioinfo. (submitted)]

A **Bayesian network** is a directed acyclic graph (DAG) that specifies a joint distribution over X as a product of local conditional distributions, one for each node:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \overset{\text{def}}{=} \prod_{i=1}^{n} p(x_i \mid x_{\text{Parents}(i)})$$

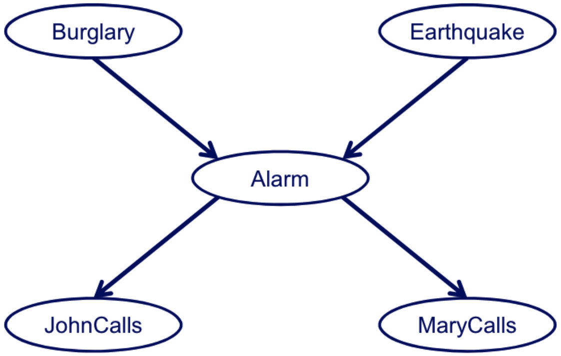


P(A,B,C,D,E,F) = P(F|C,D,E)P(A,B,C,D,E)

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

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

With the network structure (local conditional probabilities), the model describes the joint probability in a more compact manner



Number of parameters (assuming all variables are binary):

Chain rule:

$$P(B,E,A,J,M) = \underset{1}{P(B)}\underset{2}{P(E|B)}\underset{4}{P(A|B,E)}\underset{8}{P(J|B,E,A)}\underset{16}{P(M|B,E,A,J)}$$

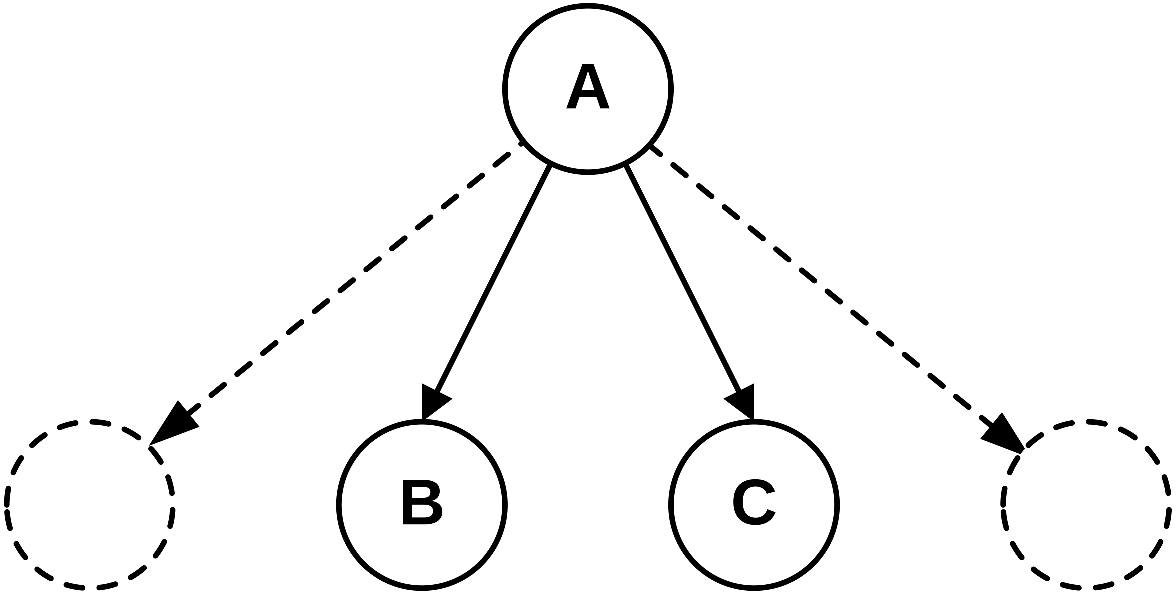
BN:

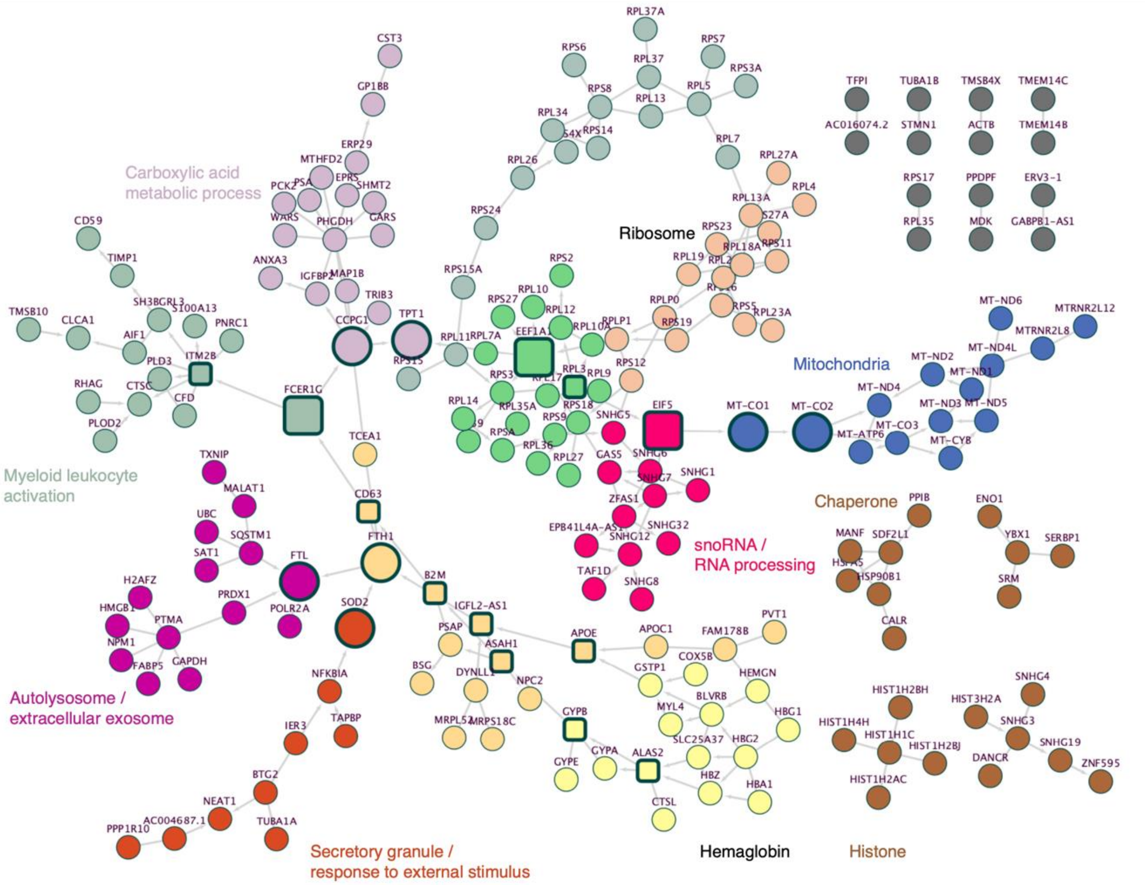
$$P(B,E,A,J,M) = \underset{1}{P(B)}\underset{1}{P(E)}\underset{4}{P(A|B,E)}\underset{2}{P(J|A)}\underset{2}{P(M|A)}$$

lectures.gersteinlab.org - [Gao et al. ('24) Bioinfo. (submitted)]

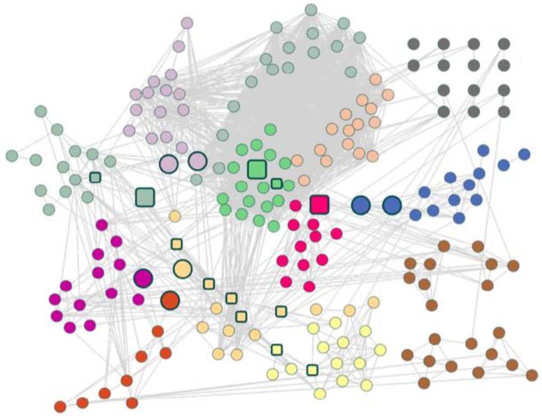
Naive Bayes

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

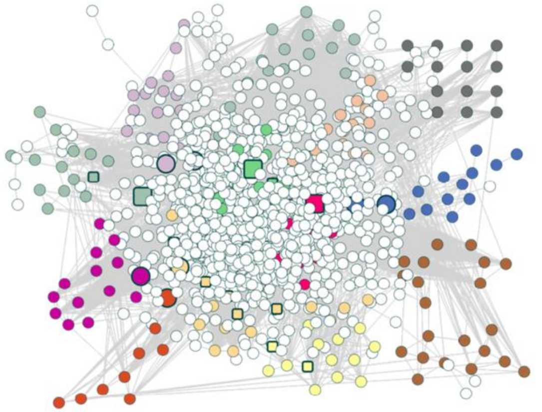




Bayesian Network



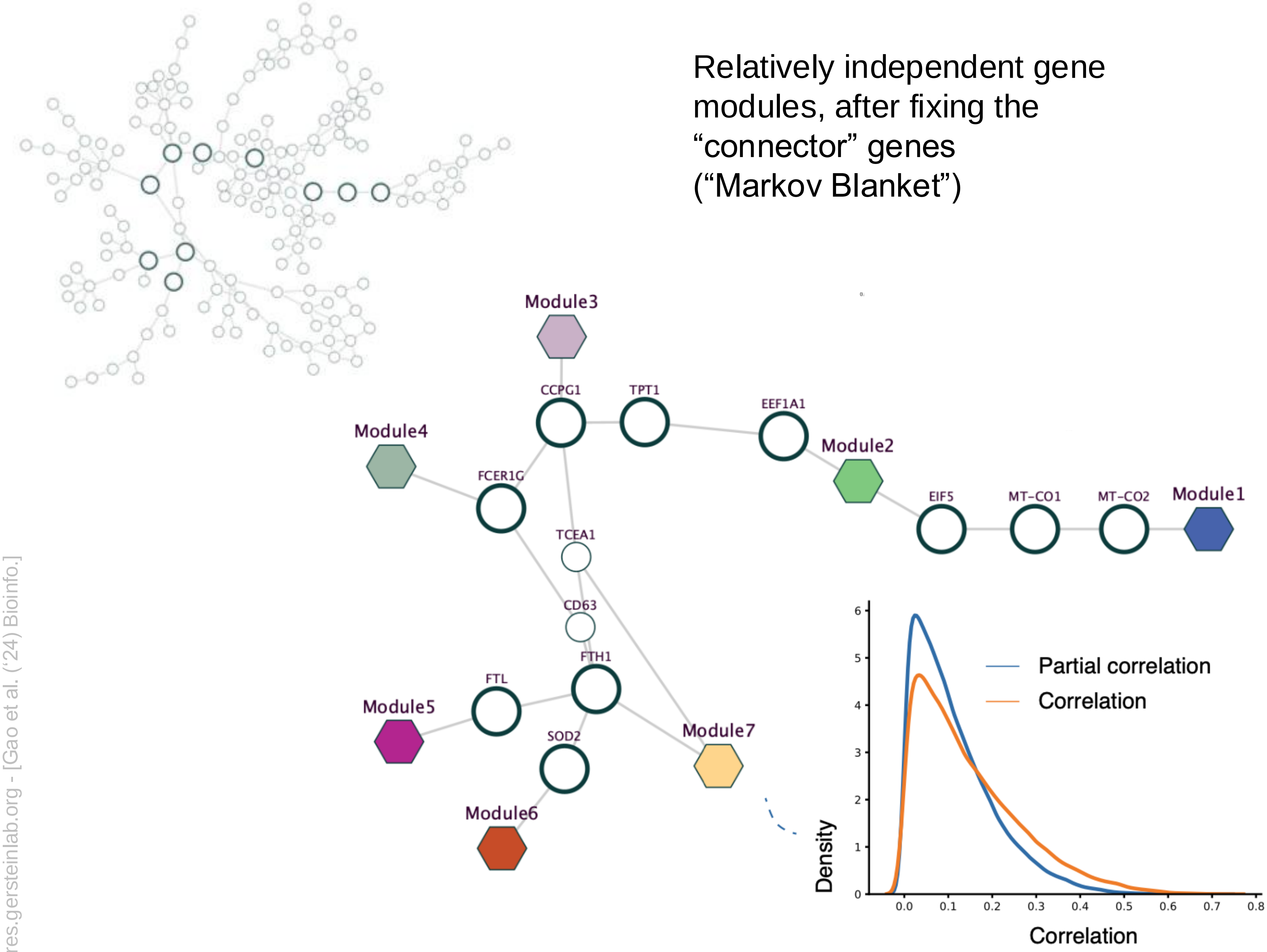
Co-expression



Co-expression + uninformative genes

lectures.gersteinlab.org - [Gao et al. ('24) Bioinfo. (submitted)]

Relatively independent gene modules, after fixing the "connector" genes ("Markov Blanket")



ures.gersteinlab.org - [Gao et al. ('24) Bioinfo.]

References

- James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert
  An Introduction to Statistical Learning: with Applications in R
  [ISLR (2nd edition)]
  https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1071614177/ + https://www.statlearning.com
  (Chapter 4.4.4 and 4.7.5 gives background on Naive Bayes.)

- Edwards et al. (2002). Trends in Genetics, 18(10), 529–536.
  Bridging structural biology and genomics: assessing protein interaction data with known complexes.
  https://doi.org/10.1016/s0168-9525(02)02763-4
  (Relates to the worked example.)