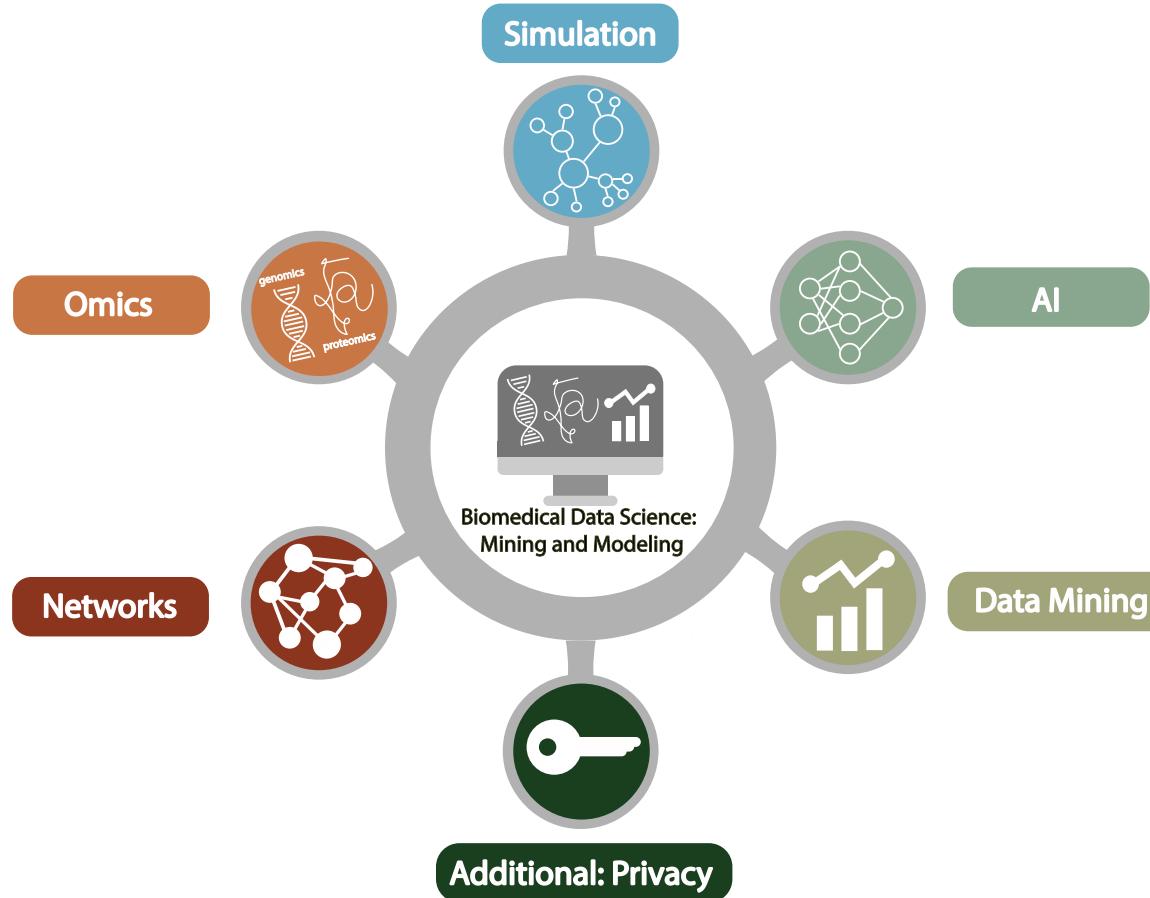


# Biomedical Data Science

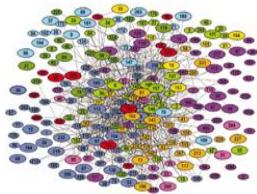
(GersteinLab.org/courses/452)

## Network Prediction

(23m10e)

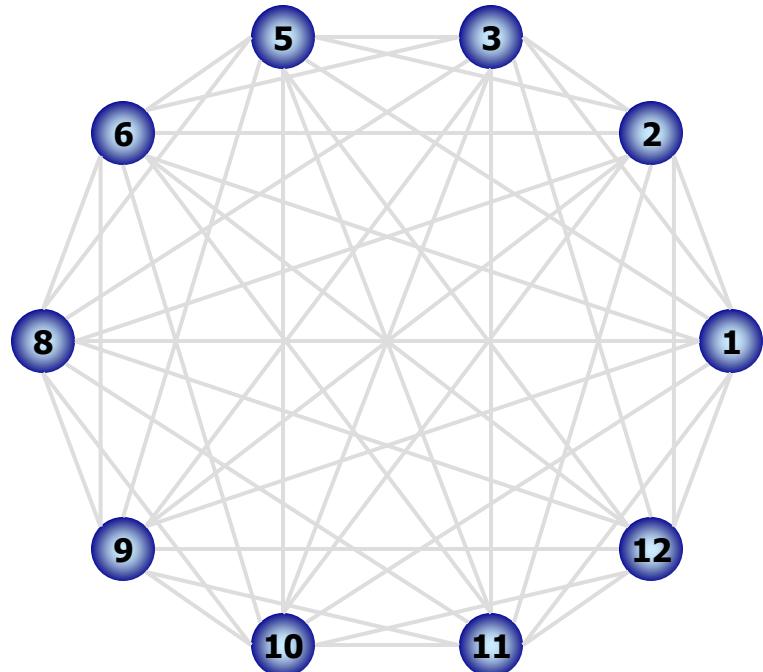


# Predicting Networks via Bayesian Integration: Problem Motivation



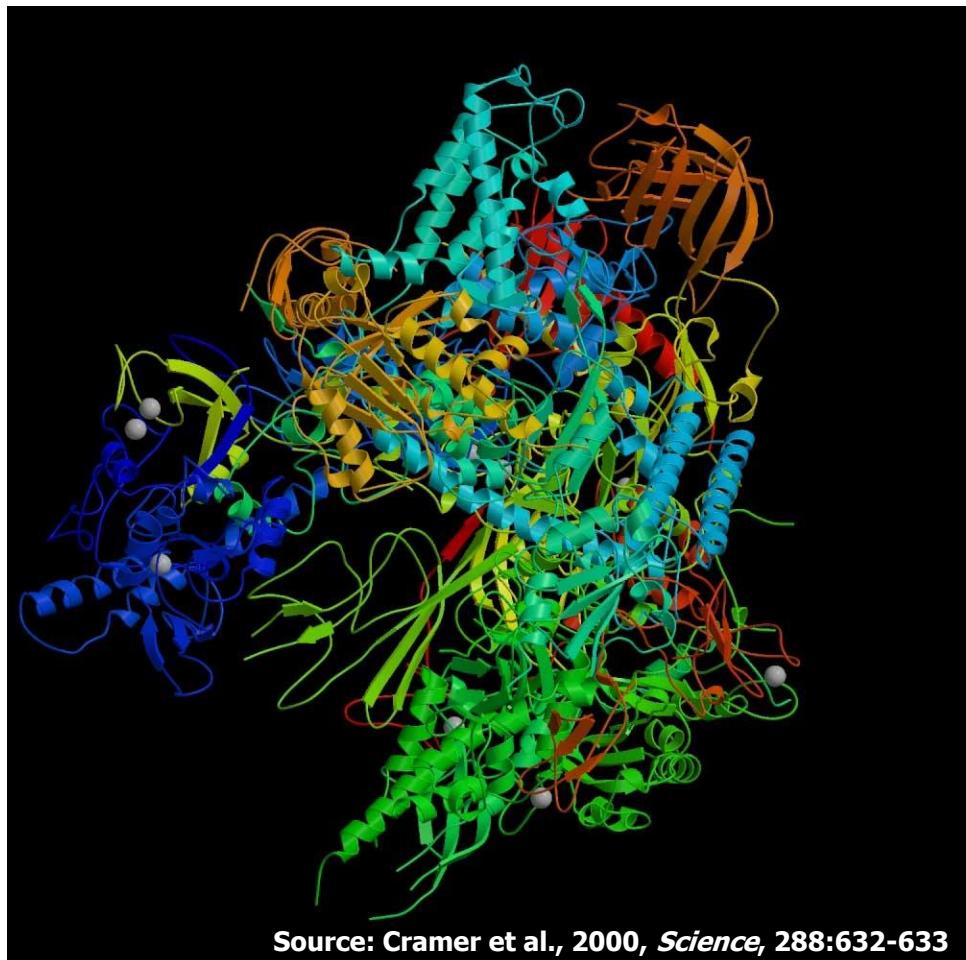
# RNA polymerase II: Structure

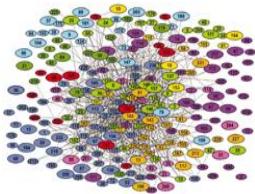
Which subunits interact?  
Based on Binding  
experiments



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

Compare with Gold Std. Structure





## Binding Experiments on Subunit Pairs

**Subunits**

**Subunits**

**Pull-down 1**

**Pull-down 2**

**Pull-down 3**

**Cross-linking**

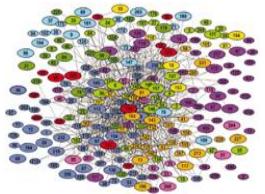
**Far Western 1**

**Far Western 2**

**Far Western 3**

	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	5	5	5	5	5	6	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12													
	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										
<b>Pull-down 1</b>	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	0														
<b>Pull-down 2</b>	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	0														
<b>Pull-down 3</b>	1									1								1	0	1	0	0	1	0																															
<b>Cross-linking</b>	1	1	1	1	1		0	1	1	1	1	0		1	1	1			1	1		1	0						1																										
<b>Far Western 1</b>	1	1								1	1							1	0	0		0	0	0	0	1		0	0	0		0	0	0	0	0	0	0	0	0	0	0													
<b>Far Western 2</b>		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	0	0														
<b>Far Western 3</b>																		1	0	0		0	1	0																															

Interaction experiments  
***before*** structure was known

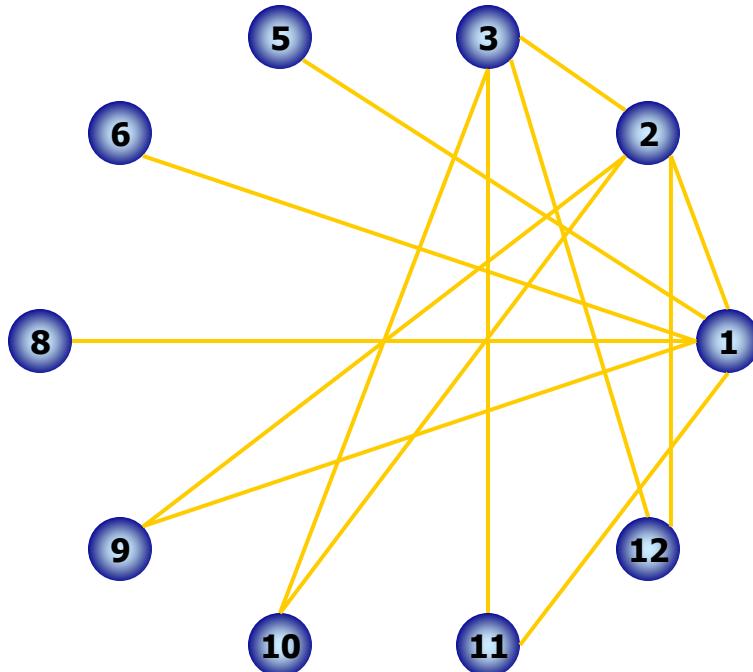


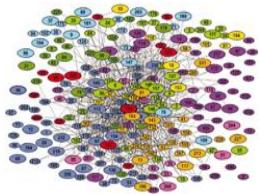
# Gold-Standard Positives

Subunits  
Subunits

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



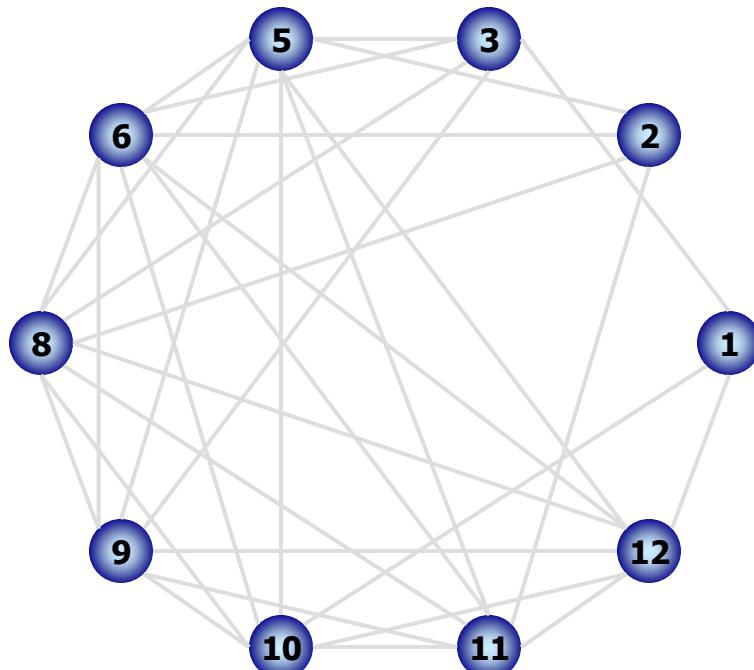


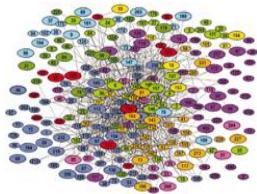
# Gold-Standard Negatives

**Subunits**  
**Subunits**

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

Gold-Standard Negative (GSTD-): 32





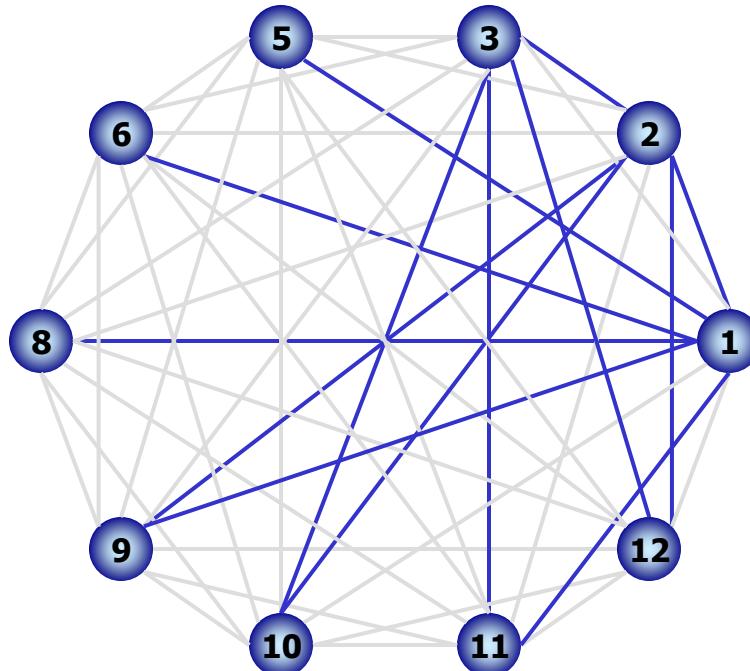
## RNA Polymerase II: Gold-Standards

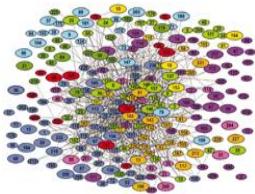
**Subunits**  
**Subunits**

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

Gold-Standard Positive (GSTD+): 13

Gold-Standard Negative (GSTD-): 32





## Assess Quality and Coverage of PPints

**Subunits**  
**Subunits**

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

**Pull-down 1**

1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

**Pull-down 2**

1	1	1	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	1	0	1	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	1	0	0	1	0	0	0	0	0	0	0	0	0	0
---	--	--	--	--	--	--	--	--	--	---	--	--	--	--	--	--	--	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

**Cross-linking**

1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

**Far Western 1**

1	1						1	1	1	1	0	1	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 2**

		1	1				1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
--	--	---	---	--	--	--	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

**Far Western 3**

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

GSTD+  
GSTD-  
True  
False

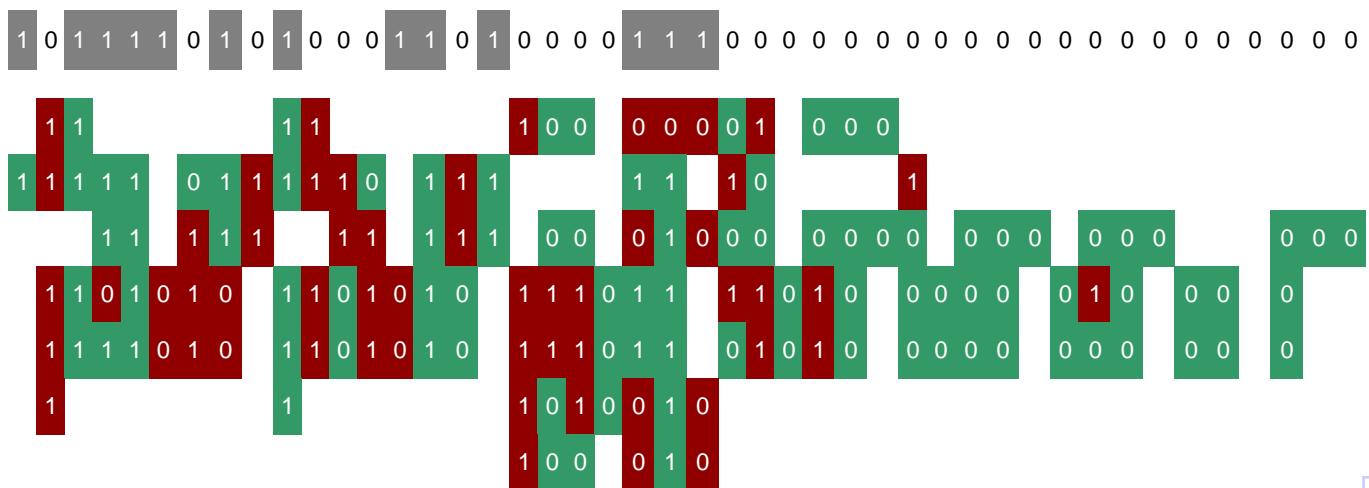
# Re-drawn in a different color scheme

**Subunit A**

**Subunit B**

**structural contact**

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



■ = false

■ = true

# Data integration: RNA polymerase II

**Subunit A**

**Subunit B**

**structural contact**

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

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 12

**Far western**

**Cross-linking**

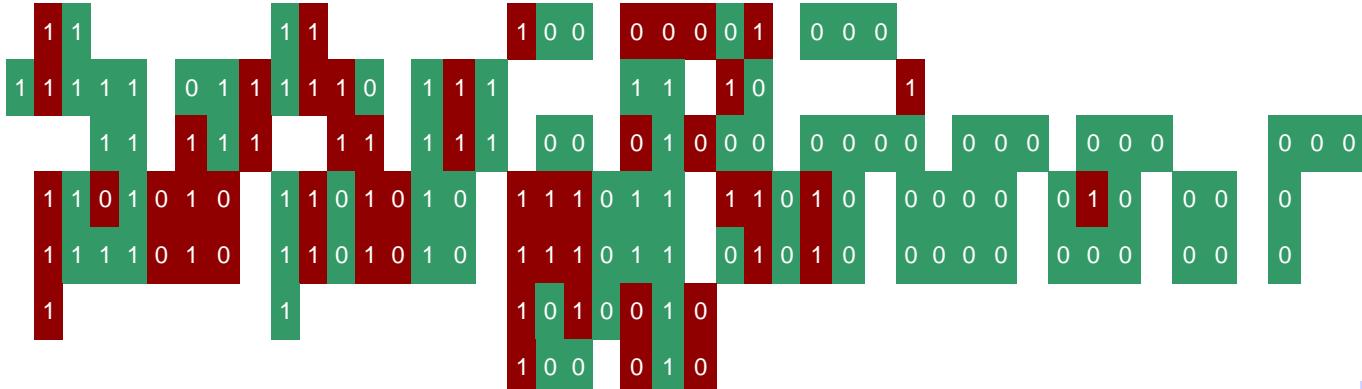
**Far western**

**Pull-down**

**Pull-down**

**Pull-down**

**Far western**



■ = false

■ = true

**Union**



# Data integration: RNA polymerase II

**Subunit A**

**Subunit B**

**structural contact**

**Far western**

**Cross-linking**

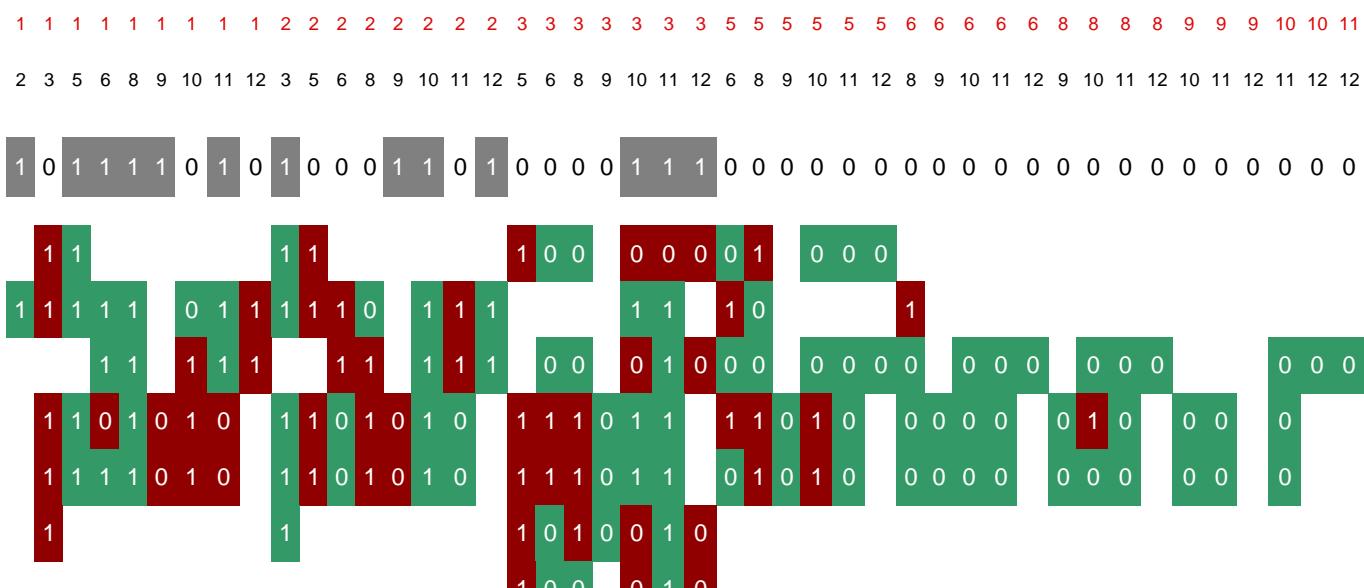
**Far western**

**Pull-down**

**Pull-down**

**Pull-down**

**Far western**



**Majority**



**Intersection**



**Union**



# Data integration: RNA polymerase II

**Subunit A**

**Subunit B**

**structural contact**

**Far western**

**Cross-linking**

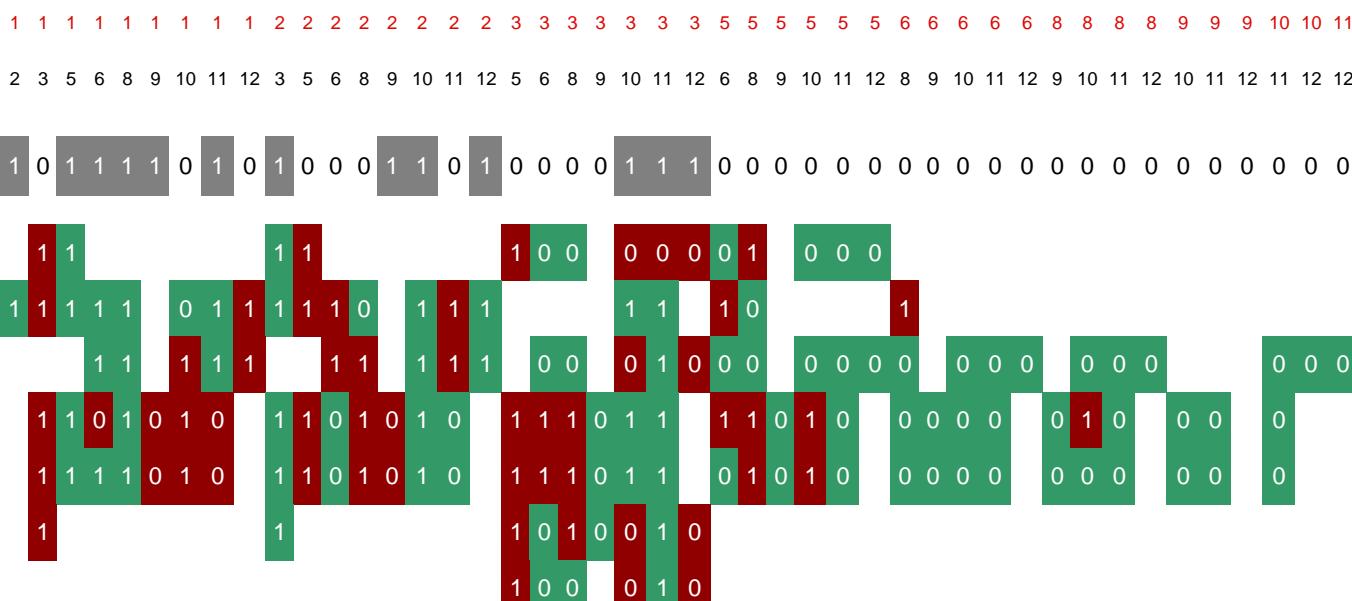
**Far western**

**Pull-down**

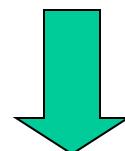
**Pull-down**

**Pull-down**

**Far western**



(Cross validate)

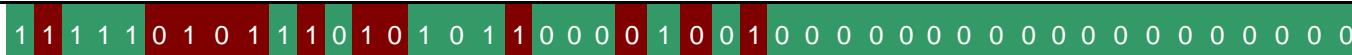


Integrate using naive Bayes classifier

**Combined (Bayesian)**



**Majority**



**Intersection**



**Union**



# Weighted Voting: the Likelihood Ratio

**structural contact**



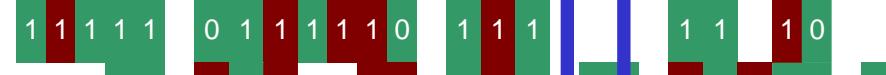
**Far western**



**Far western (dup)**



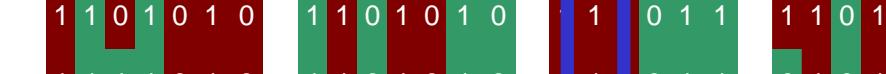
**Cross-linking**



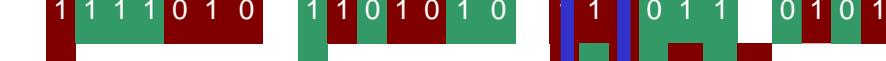
**Far western**



**Pull-down**



**Pull-down**



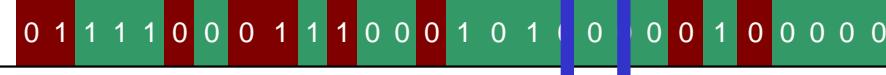
**Pull-down**



**Far western**



**Combined**



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

With weights: **likelihood ratio L** = L<sub>1</sub> + L<sub>2</sub> + L<sub>3</sub> ...

# Predicting Networks via Bayesian Integration: Intuition & Formalism

Derived from  
"perceptron model"  
 $R = \langle w, f \rangle + b$

# Supervised Classification by Weighted Voting

**Simple Vote:**  $R = f_1 + f_2 + f_3 + \dots + f_n$       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:**

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

If has prior knowledge  $w_0$

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

# Classification by Voting

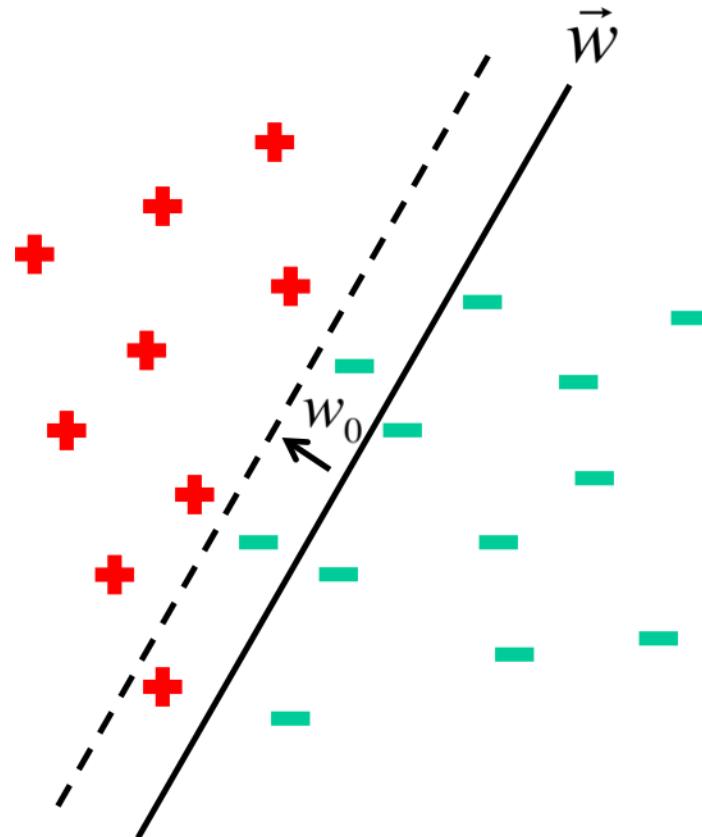
On Training Set

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

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

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

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



# 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**

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

Assume Naïve Bayes =  $\frac{P(f_1 \mid I)P(f_2 \mid I)P(f_3 \mid I)\dots P(I)}{P(f_1, f_2, f_3, \dots)}$   
(independent)

$$\begin{aligned} P(\sim I \mid f_1, f_2, f_3, \dots) &= \frac{P(f_1, f_2, f_3, \dots \mid \sim I)P(\sim I)}{P(f_1, f_2, f_3, \dots)} \\ &= \frac{P(f_1 \mid \sim I)P(f_2 \mid \sim I)P(f_3 \mid \sim I)\dots P(\sim I)}{P(f_1, f_2, f_3, \dots)} \end{aligned}$$

$$\begin{aligned} \log\left(\frac{P(I \mid f_1, f_2, f_3, \dots)}{P(\sim I \mid f_1, f_2, f_3, \dots)}\right) &= \log\left(\frac{P(f_1 \mid I)}{P(f_1 \mid \sim I)} \frac{P(f_2 \mid I)}{P(f_2 \mid \sim I)} \frac{P(f_3 \mid I)}{P(f_3 \mid \sim I)} \dots \frac{P(I)}{P(\sim I)}\right) \\ &= \log \frac{TPR_1}{FPR_1} + \log \frac{TPR_2}{FPR_2} + \log \frac{TPR_3}{FPR_3} + \dots + \log \frac{P}{N} \end{aligned}$$

More Bayes Rule

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

The equation shows the log-likelihood ratio of the joint probability of the evidence given the hypothesis versus the joint probability of the evidence given the hypothesis being false. This is decomposed into the sum of individual log-likelihood ratios for each feature  $f_i$ . Below the equation, four upward arrows point from the weights  $w_1, w_2, w_3, w_0$  to the corresponding terms in the sum.

## More Bayes Rule

# Estimating Probabilities

- We have so far estimated  $P(X=x | Y=y)$  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  $P(X=x | Y=y)=0.05$  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

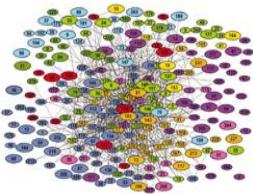
- Replace  $n_{x|y}/n_y$  by:

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

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 via Bayesian Integration: Worked Examples

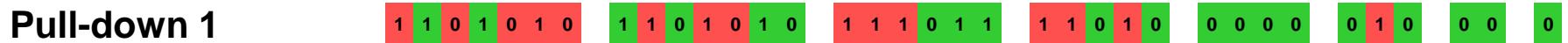


# Likelihood Ratios

**Subunits**



**Subunits**



**Pull-down 1**

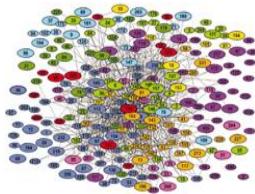
$$L_1 = \frac{p(x_1 | GSTD +)}{p(x_1 | GSTD -)}$$

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

Likelihood Ratio  
for Feature  $f$ :

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

- █ GSTD+
- █ GSTD-
- █ True
- █ False

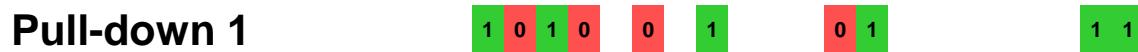


# Calculating Likelihood Ratios

**Subunits**



**Subunits**



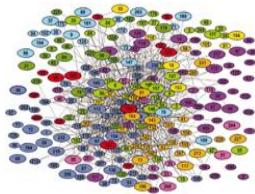
**Pull-down 1**



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

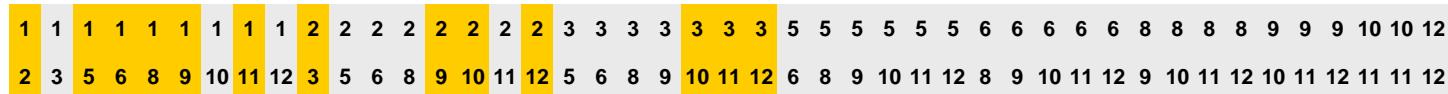
$$L_0 = \frac{p(x_0 | GSTD+)}{p(x_0 | GSTD-)} = \frac{4/13}{}$$

- █ GSTD+
- █ GSTD-
- █ True
- █ False



# Calculating Likelihood Ratios

**Subunits**



**Subunits**

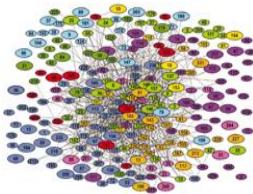
**Pull-down 1**



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

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

- █ GSTD+
- █ GSTD-
- █ True
- █ False



# Calculating Likelihood Ratios

**Subunits**  
**Subunits**

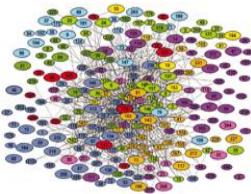
1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	5	5	5	5	5	6	6	6	6	6	8	8	8	8	9	9	9	10	10	12
2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12	5	6	8	9	10	11	12	6	8	9	10	11	12	8	9	10	11	12	10	11	12			

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

1 1 0 1 0 1 0	1 1 0 1 0 1 0	1 1 1 0 1 1 1	1 1 0 1 0 1 0	0 0 0 0	0 1 0	0 0	0
1 1 1 1 0 1 0	1 1 0 1 0 1 0	1 1 1 0 1 1 1	1 1 0 1 0 1 0	0 0 0 0	0 0 0	0 0	0
1	1	1	1 0 1 0 1 0	1 1	1 0 1 0 1 0	1	1
1 1 1 1 1	0 1 1 1 1 0	1 1 1 1 1 0	1 1 1 1 1 0	1 1 1 1 1 0	1 1 1 1 1 0	1	1
1 1	1 1	1 1 1	1 1	1 1 1	1 0 0	0 0 0	0 0 0
1 1	1 1 1	1 1	1 1	1 1 1	0 0	0 0 0	0 0 0
1 1	1 1 1	1 1	1 1	1 1 1	0 0	0 0 0	0 0 0
1 1	1 1 1	1 1	1 1	1 1 1	0 0	0 0 0	0 0 0

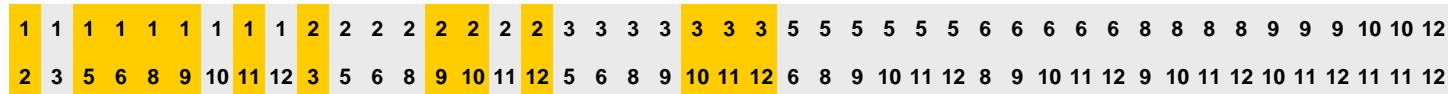
<b>Pull-down 1</b>	$L1 = (6/13) / (11/32) = 1.34$	$L0 = (4/13) / (14/32) = 0.70$
<b>Pull-down 2</b>	$L1 = (7/13) / (9/32) = 1.91$	$L0 = (2/13) / (16/32) = 0.31$
<b>Pull-down 3</b>	$L1 = (2/13) / (3/32) = 1.64$	$L0 = (2/13) / (2/32) = 2.46$
<b>Cross-linking</b>	$L1 = (10/13) / (7/32) = 3.52$	$L0 = (0/13)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

- █ GSTD+
- █ GSTD-
- █ True
- █ False

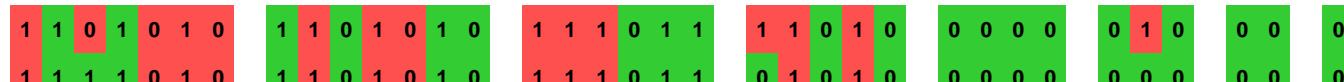


# Data Integration: ROC-Curve

**Subunits**  
**Subunits**



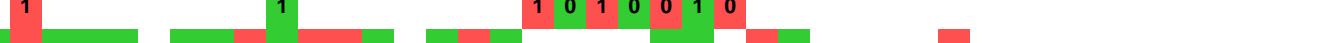
**Pull-down 1**



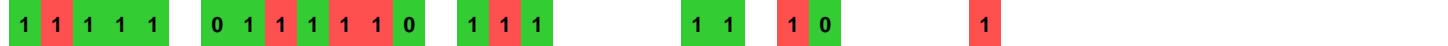
**Pull-down 2**



**Pull-down 3**



**Cross-linking**



**Far Western 1**



**Far Western 2**



**Far Western 3**



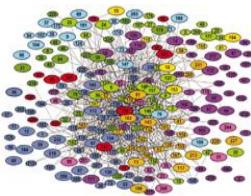
**Combined (Bayes)**



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

- █ GSTD+
- █ GSTD-
- █ True
- █ False

**“Weighted Voting”**



# Data Integration: ROC Curve

**Subunits**  
**Subunits**

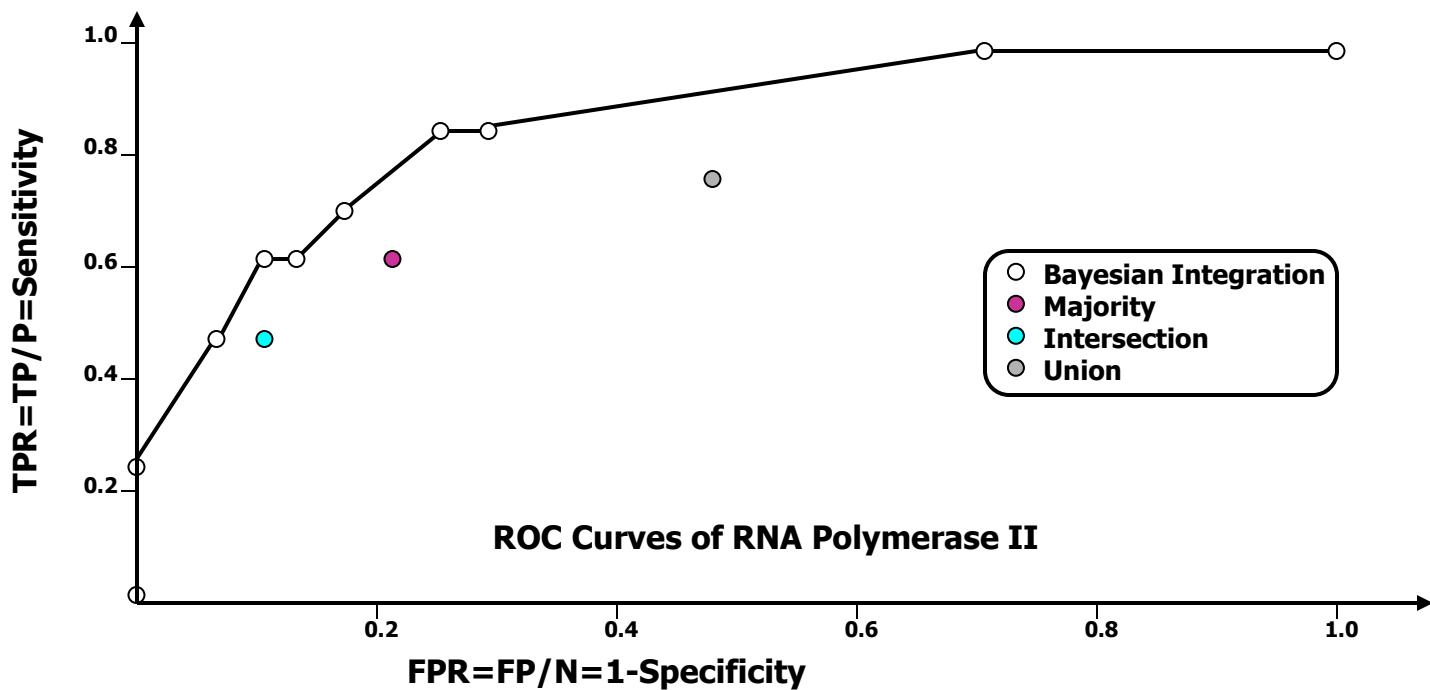
1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	5	5	5	5	6	6	6	6	8	8	8	8	9	9	9	10	10	12
2	3	5	6	8	9	10	11	12	3	5	6	8	9	10	11	12	5	6	8	9	10	11	12	8	9	10	11	12	9	10	11	12	10	11	12			

**Combined (Bayes)**

1	1	1	1	1	1	0	0	0	1	1	0	0	0	1	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

**Majority  
Intersection  
Union**

1	1	1	1	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	1	0	0	1	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
1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	0	1	0	1	0	1	0	1	0	0	0	0	0



# Predicting Networks via Bayesian Integration: Features Correlation

# Correlations between similar features

**structural contact**



**Far western**



**Far western (dup)**



**Cross-linking**



**Far western**



**Pull-down**



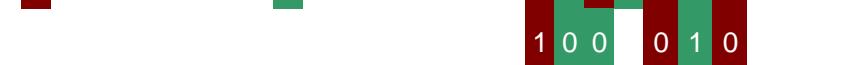
**Pull-down**



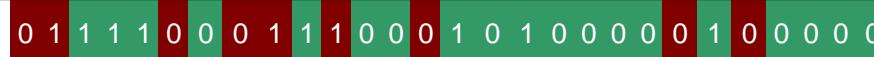
**Pull-down**



**Far western**



**Combined**



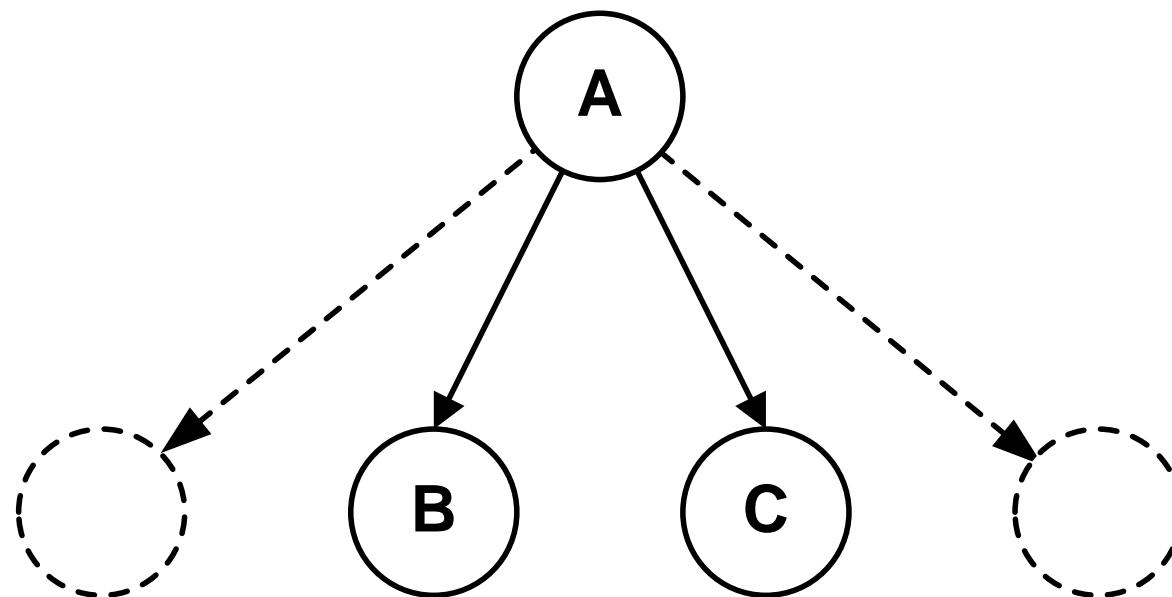
<b>GS</b>	1	0	0	1	1	1	...
<b>F1</b>	1	0	0	1	0	0	...
<b>F2</b>	1	0	0	1	0	0	...
<b>F3</b>	0	0	1	0	1	0	...
<b>F4</b>	0	1	0	0	1	0	...
<b>F5</b>	0	1	0	0	1	0	...
<b>F6</b>	1	0	1	1	0	0	...

# Feature Correlation and Fully Connected Bayes

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

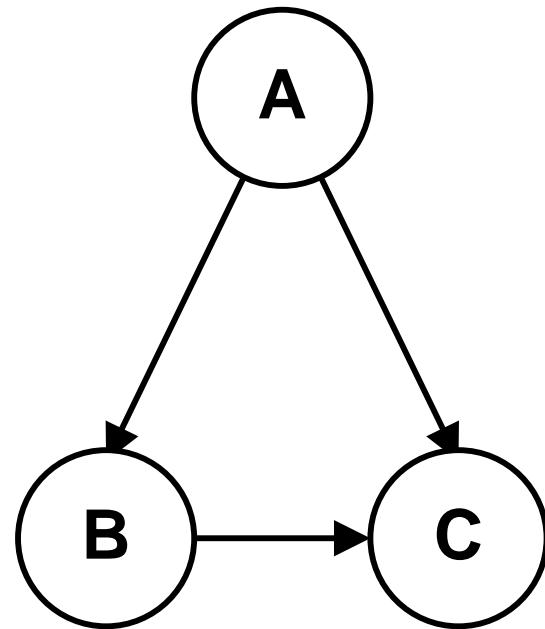
# Naive Bayes

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

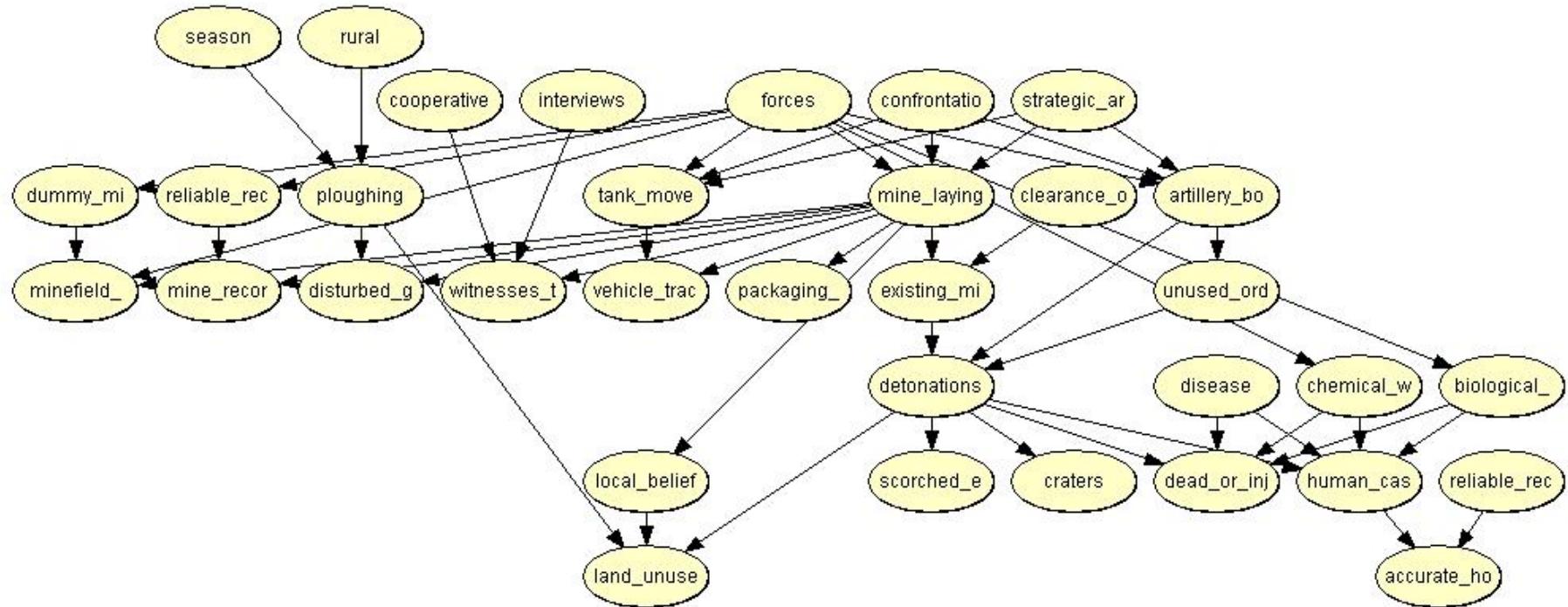


## A ‘correct’ factorisation

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



# 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