14 PREDICTING PROTEIN INTERACTIONS

* Prediction challenges
  + Effect of point mutations
  + Structure of complexes
  + All interacting proteins.
* Use structural data to predict complexes.
* Simply want to see if two proteins will interact or not (Docking problem)

1. Docking
   which surface of A interacts with which surface of B?
   * ∀ partners,
     + evaluate all possible relative position + orientation
     + Allow for structural changes
     + Measure energy of interaction

2. Problems: Slow, False Positives.
   * no taking into account relative strength of interaction w/ all other proteins.
   (may not be the best interaction.

3. Limit Search Space
   * filter BEFORE structural comparisons,
   * avoid expensive commute until confident

4. PRISM, PrePPI

4.1 PRISM idea:
* limited # of protein "architectures."
* architectural motifs define interactions
* find spatially similar regions of protein complements of known interface.

[Predict "hot spots" Deep Learning]

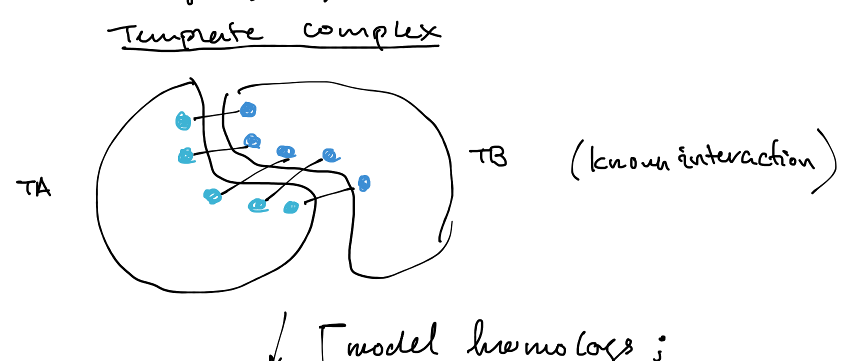
4.1.2 Two Parts
* rigid body structural comparisons of target proteins to known PPInterface
* flexible refinement using docking energy
* evaluation metric: structural similarity & conservation of binding "hot spots"

4.1.3 Algorithm
* identify interface of template (dist. cutoff)
* align surface of query to half interfaces
* ignore rest of structure, test
  structure match
  " ——— " at hotspots
  sequence match " ——— "
* flexible refinement

4.1.4 Use PDB for template interaction surfaces
(PDB = Protein Data Bank)

4.2 PrePPI
* find homologous proteins of known structure M_A, M_B to the Query Q_A, Q_B
* find structural neighbors $\{NA_i, NB_i\}$ [~1500/structure]
* find known interaction among neighbors.
* align sequences of M_A, M_B to N_A, N_B based on structure

4.2.1 Aligning Sequences.
Template complex

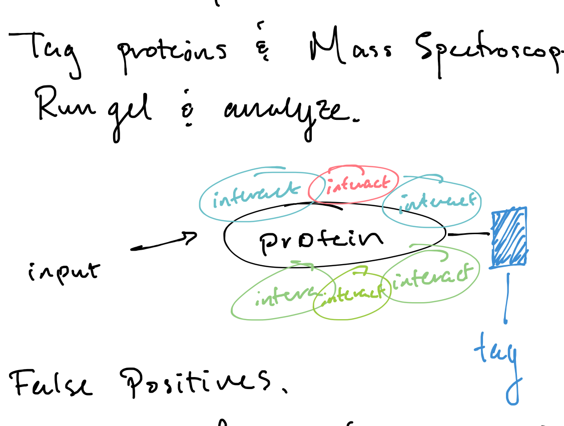


[model homologs; are they strong interaction?]
evaluate on 5 metrics
1. SIM (structural similarity)
2. SIZ (#) COV (fraction) of interaction pairs that can be aligned.
3. PredUs, PINUP, cons-PPISP

* Drawback: doesn't perform flexible structure analysis.

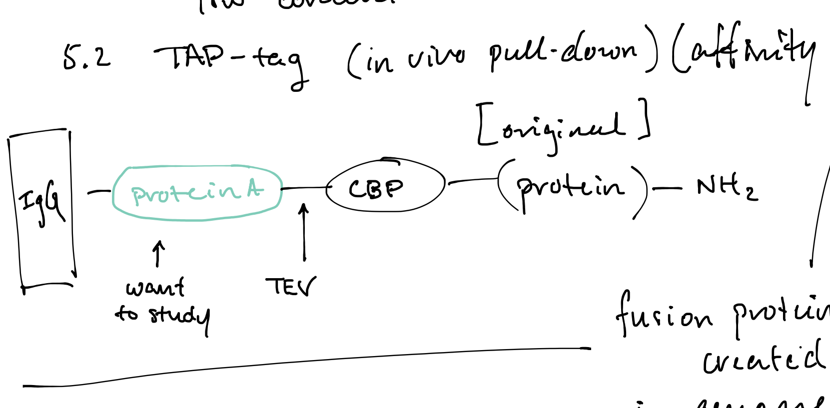
5. Outline
* High thruput measuring PPI
* Estimate interaction probabilities.
* BayesNet prediction of PPI

5.1 "Protein Complexes take the bait"
* Tag proteins & Mass Spectroscopy
* Run gel & analyze.



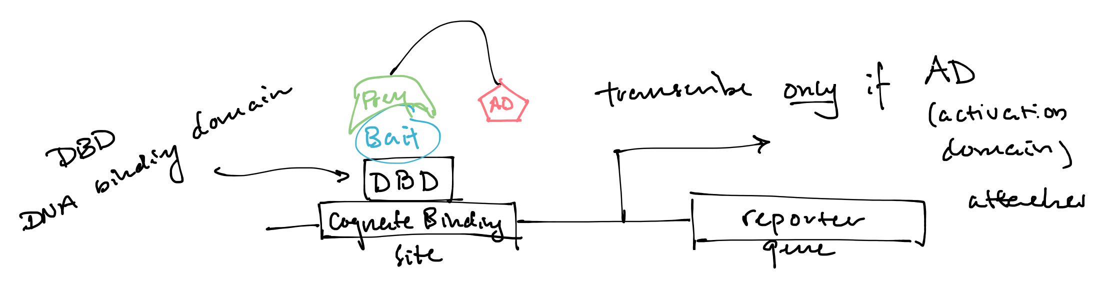
* False Positives.
  - non-specific binding proteins
  - proteins binding to tag
* False Negatives
  - near tag interactions.
  - low concentration levels

5.2 TAP-tag (in vivo pull-down) (affinity capture)



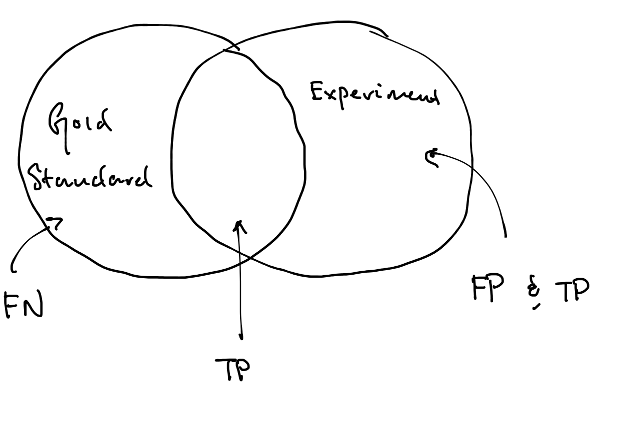
fusion protein created in genome.

5.3 Yeast Two Hybrid

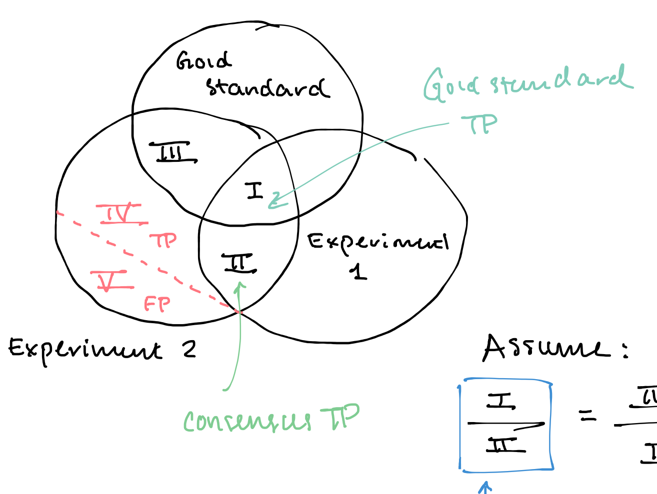


transcribe only if AD (activation domain) attached

6. Error Rates



6.1 Data Integration



Assume:

$$\frac{I}{II} = \frac{III}{IV}$$

($\frac{I}{II}$ = fraction of TP captured by gold standard)

6.2 Lots of False Positives!
* want to build a probability of each interaction based on external data

6.3 Bayesian Model

$$Pr[\text{real PPI} \mid \text{Data}] = \frac{Pr[\text{data} \mid \text{true PPI}]\, Pr[\text{true PPI}]}{Pr[\text{data}]}$$

$$\text{likelihood ratio} = \frac{Pr[\text{true PPI} / \text{data}]}{Pr[\text{false PPI} \mid \text{data}]}$$

(cancel out $Pr[\text{data}]$)

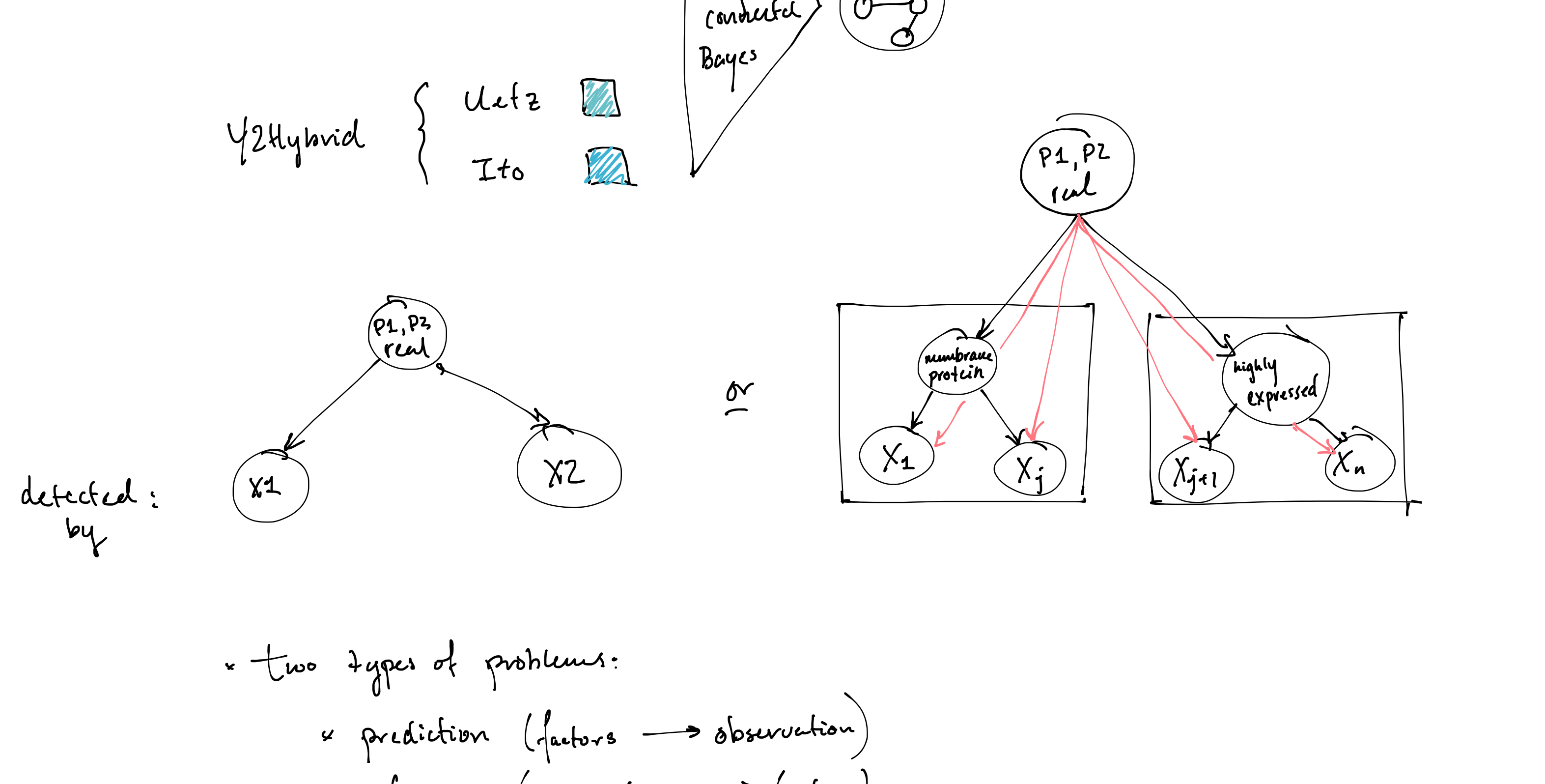
$$RF = \text{ranking function} = \log\left[\frac{Pr[\text{data} \mid \text{true PPI}]}{Pr[\text{data} \mid \text{false PPI}]}\right]$$

(i.e. ignore priors)

∀ (A, B) pairs.
Protein A, Protein B
↪ compute RF_{AB} based on experiments $1, 2, \ldots, n$

6.4 Bayesian Networks.
* in bio: generegulation, signaling, prediction
* consists of graph & set of probabilities. ("learned" from data)

6.5 Predicting interactions.



* two types of problems:
  * prediction (factors → observation)
  * inference (observation → factors)

7. Estimate INT P using other kinds of data

7.1 Gene expression
* Correlated expression of genes.

7.2 Co-evolution
* multiple instances of appearance/disappearance of same gene pairs