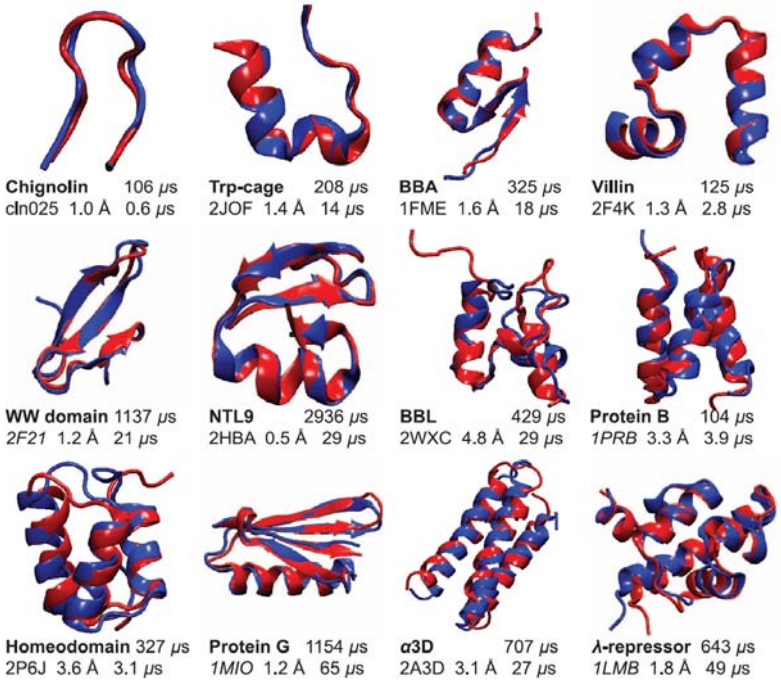
- L12 - Introduction to Protein Structure; Structure Comparison & Classification
- L13 - Predicting protein structure
- L14 - Predicting protein interactions
- L15 - Gene Regulatory Networks
- L16 - Protein Interaction Networks
- L17 - Computable Network Models

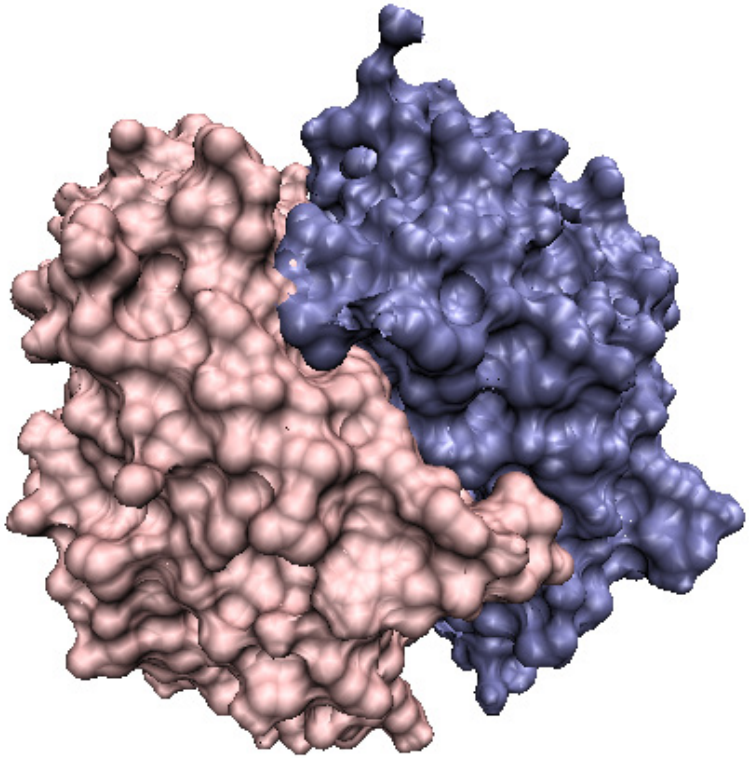
Predictions

Last time: protein structure



© American Association for the Advancement of Science. All rights reserved. This content is excludedfrom our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Lindorff-Larsen, Kresten, Stefano Piana, et al. "How Fast-folding Proteins Fold." *Science* 334, no. 6055 (2011): 517-20.

Now: protein interactions



© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

Prediction Challenges

- Predict effect of point mutations
- Predict structure of complexes
- Predict all interacting proteins

Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions

•DOI: 10.1002/prot.24356

"Simple" challenge: Starting with known structure of a complex: predict how much a mutation changes binding affinity.

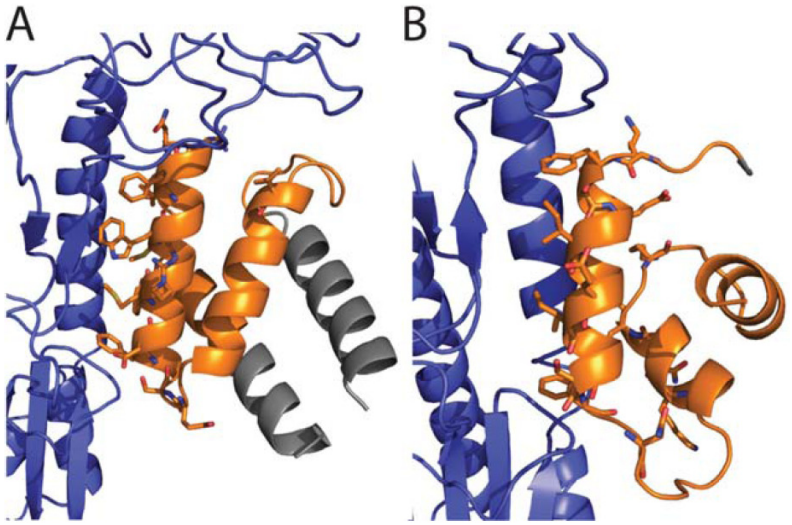
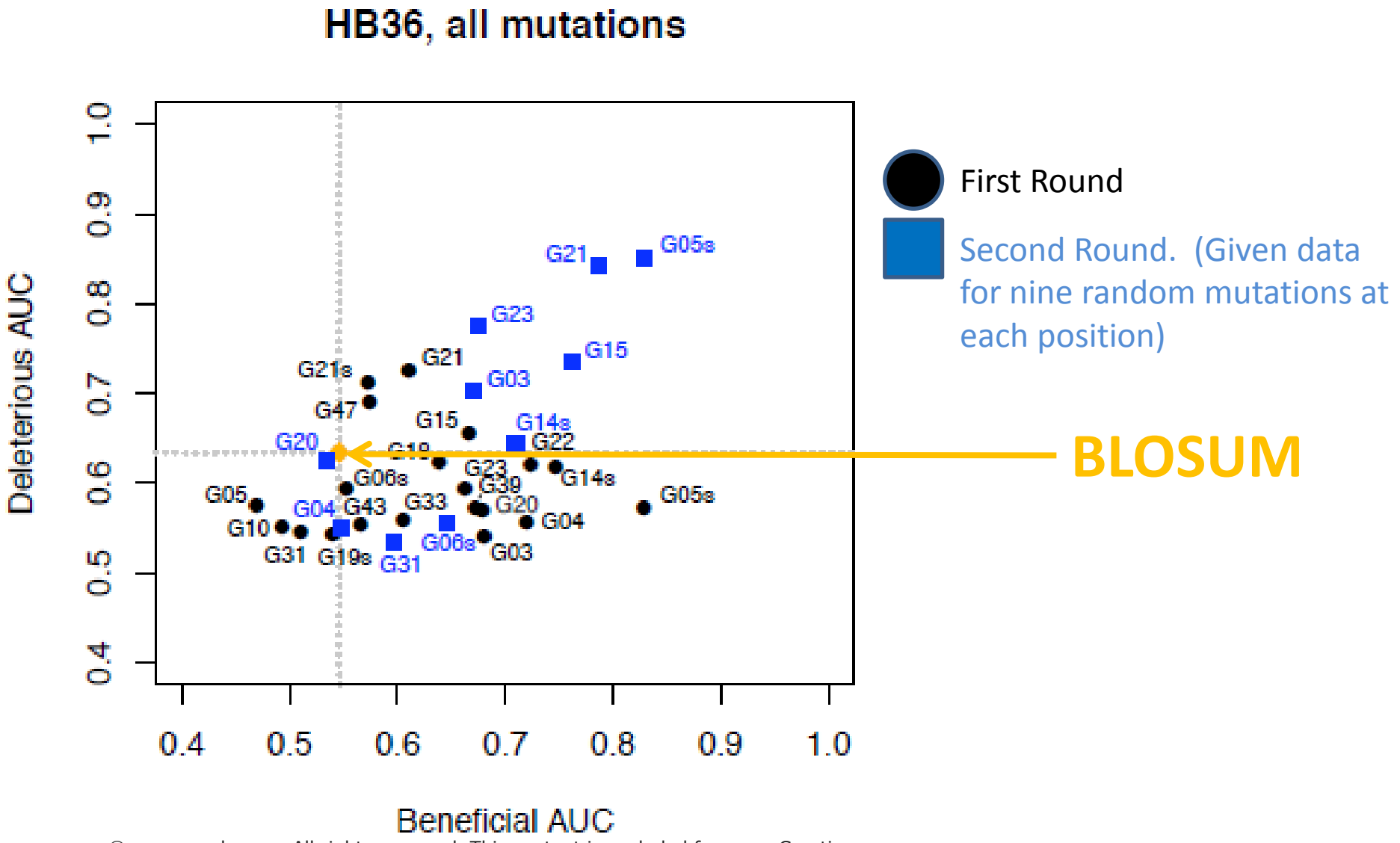


Figure 1

The structures of (**A**) HB36 (**B**) HB80 in complex with HA (blue) which were provided to participants. Residues probed in the deep sequencing enrichment experiment are in orange; the remainder are in grey. Residues at the interface are represented as sticks.

© Wiley Periodicals, Inc. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Moretti, Rocco, Sarel J. Fleishman, et al. "Community-wide Evaluation of Methods for Predicting the Effect of Mutations on Protein–protein Interactions." Proteins: Structure, Function, and Bioinformatics 81, no. 11 (2013): 1980-7.

Area under curve for predictions (varying cutoff in ranking)

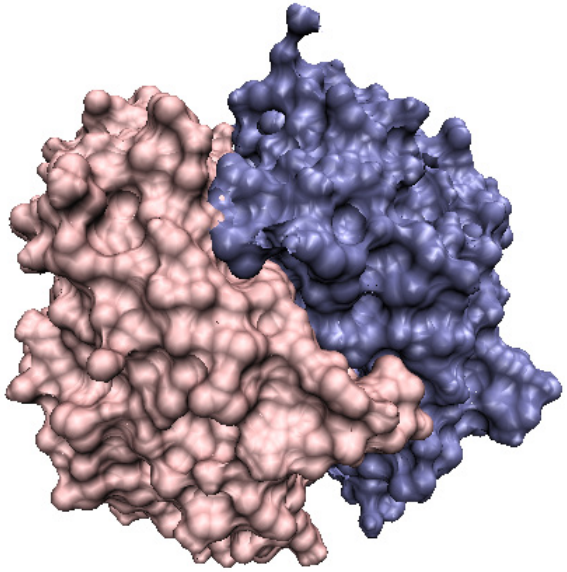


© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

•DOI: 10.1002/prot.24356

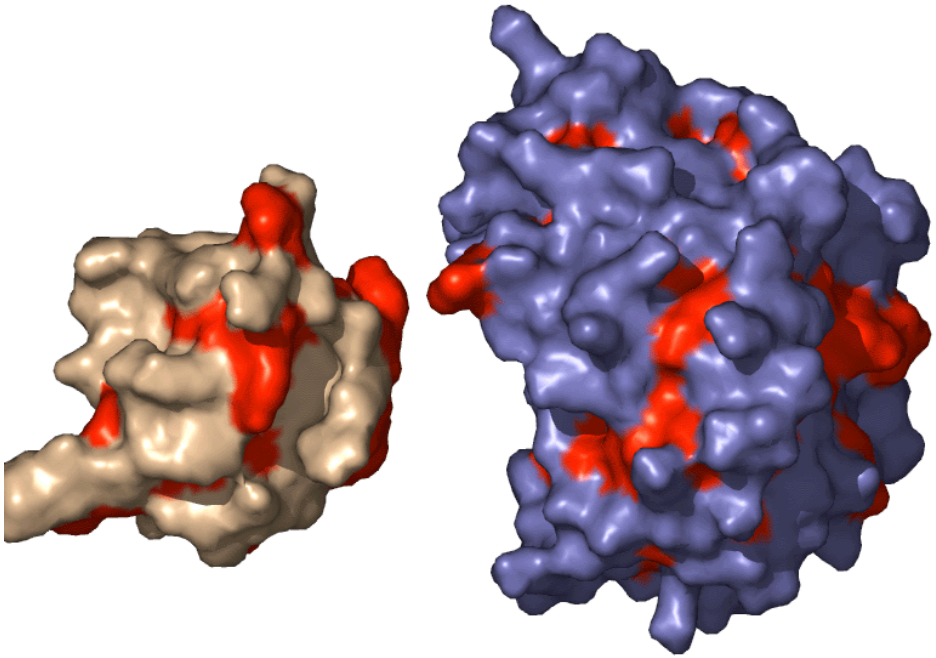
Predicting Structures of Complexes

- Can we use structural data to predict complexes?
- This might be easier than quantitative predictions for site mutants.
- But it requires us to solve a docking problem



© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

Docking

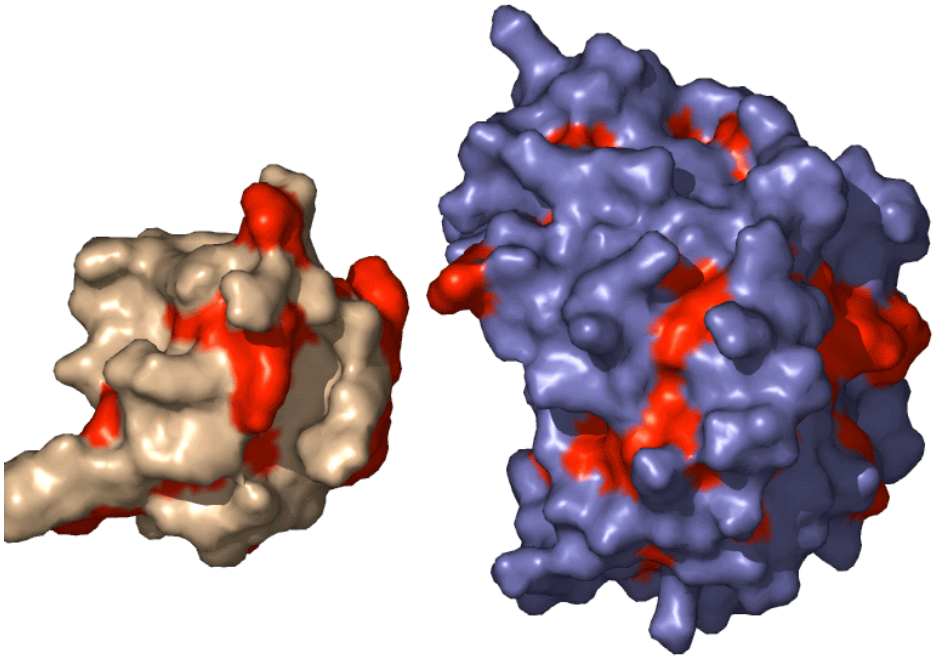


Courtesy of Nurcan Tuncbag. Used with permission.

Which surface(s) of protein A interactions with which surface of protein B?

N. Tuncbag

Docking



Courtesy of Nurcan Tuncbag. Used with permission.

Imagine we wanted to predict which proteins interact with our favorite molecule.

For each potential partner:

- Evaluate all possible relative positions and orientations
  - allow for structural rearrangements
    - measure energy of interaction

This approach would be extremely slow!

It's also prone to false positives.

Why?

Reducing the search space

- Efficiently choose potential partners before structural comparisons
- Use prior knowledge of interfaces to focus analysis on particular residues

Next

PRISM

Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement.

Tuncbag N, Keskin O, Nussinov R, Gursoy A.

http://www.ncbi.nlm.nih.gov/pubmed/22275112

PrePPI

Structure-based prediction of protein–protein interactions on a genome-wide scale

Zhang, et al.

http://www.nature.com/nature/journal/v490/n7421/full/nature11503.html

PRISM's Rationale

There are limited number of protein "architectures".

Protein structures can interact via similar architectural motifs even if the overall structures differ

Find particular surface regions of proteins that are spatially similar to the complementary partners of a known interface

N. Tuncbag

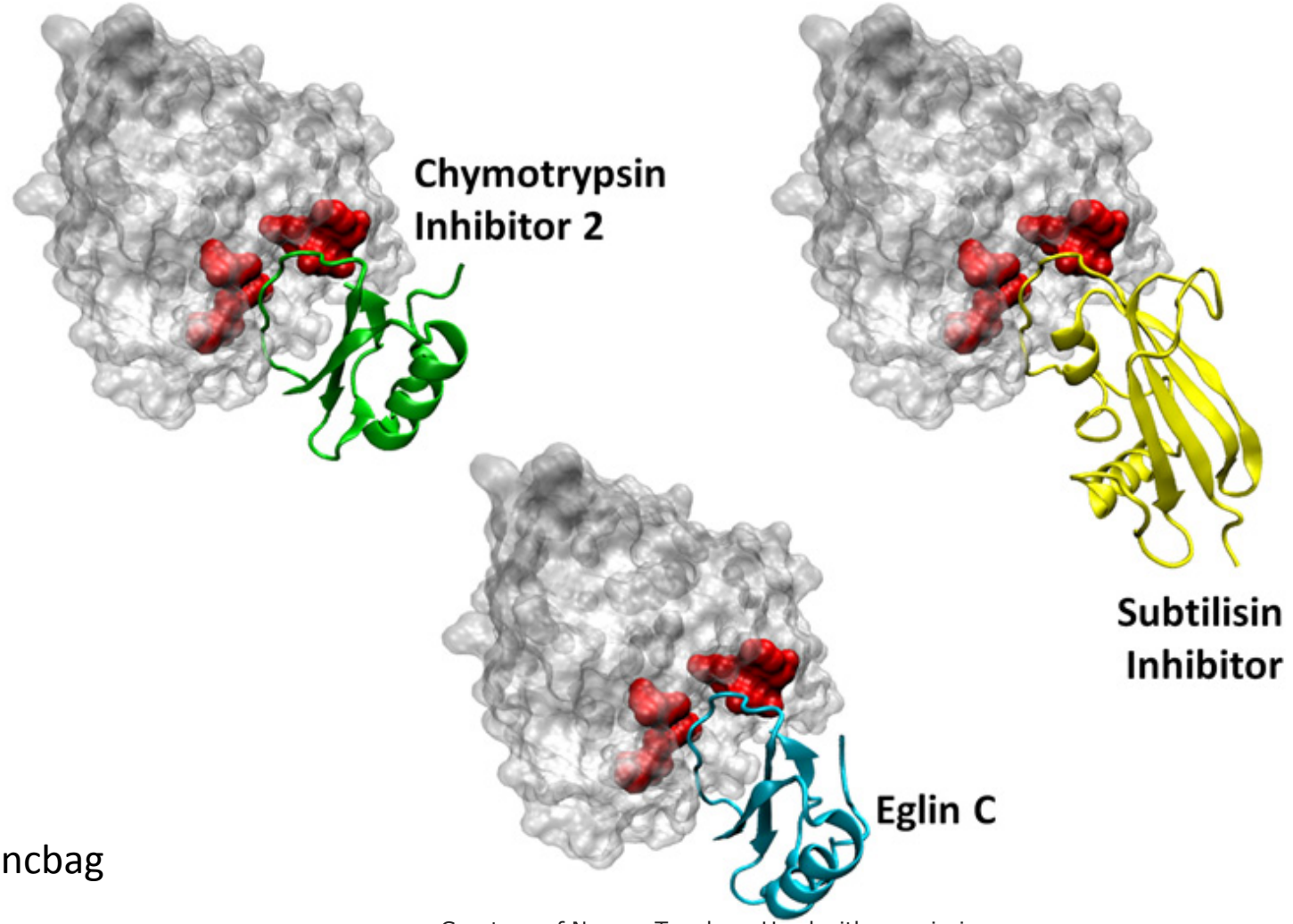
PRISM's Rationale

- Two components:
  - **rigid-body structural comparisons** of target proteins to known template protein-protein interfaces
  - **flexible refinement using a docking energy function**.
- Evaluate using structural similarity and evolutionary conservation of putative binding residue **'hot spots'**.

N. Tuncbag

Subtilisin and its inhibitors

Although global folds of Subtilisin's partners are very different, binding regions are structurally very conserved.



N. Tuncbag

Courtesy of Nurcan Tuncbag. Used with permission.

Hotspots

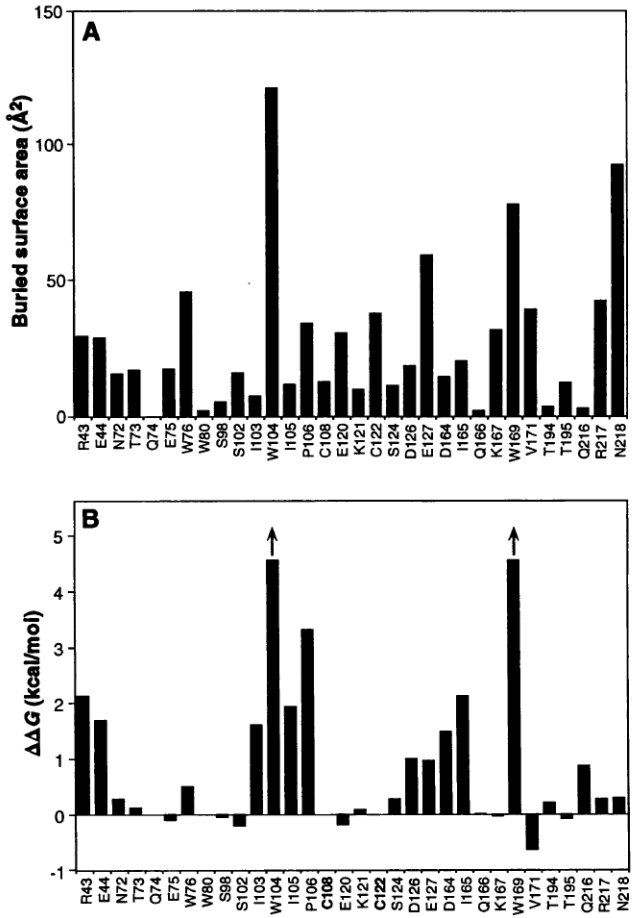
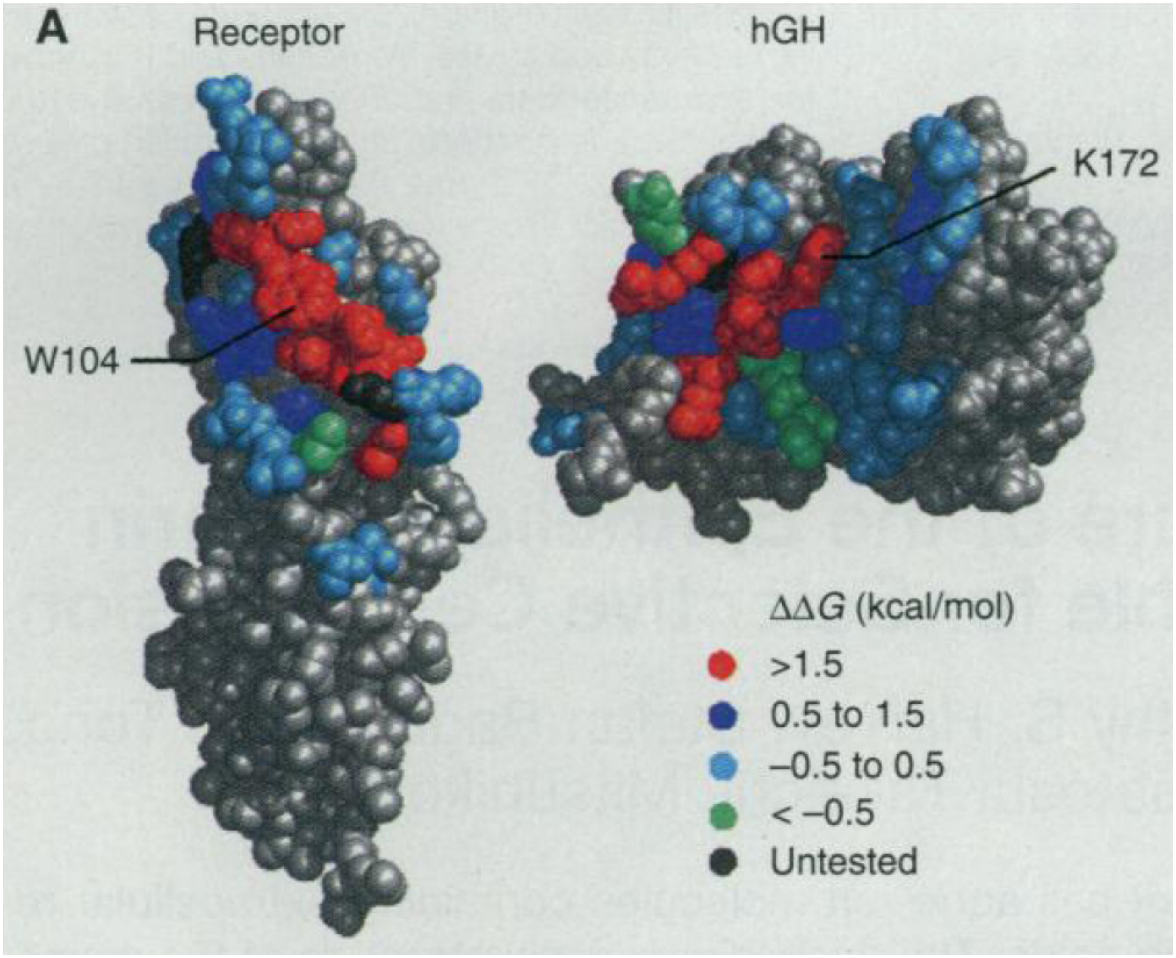


Fig. 1. Contribution of only a subset of contact residues to net binding energy. (**A**) Loss of solvent-accessible area (*7*) of the side chain portion of each residue in the hGHbp on forming a complex with hGH. (**B**) Difference in binding free energy between alanine-substituted and wild-type hGHbp $(\Delta\Delta G)_{mut\text{-}wt}$ at contact residues (*5*). Negative values indicate that affinity increased when the side chain was substituted by alanine.

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Clackson, Tim and James A. Wells. "A Hot Spot of Binding Energy in a Hormone-Receptor Interface." *Science* 267, no. 5196 (1995): 383-6.

Figure from Clackson & Wells (1995).

Hotspots

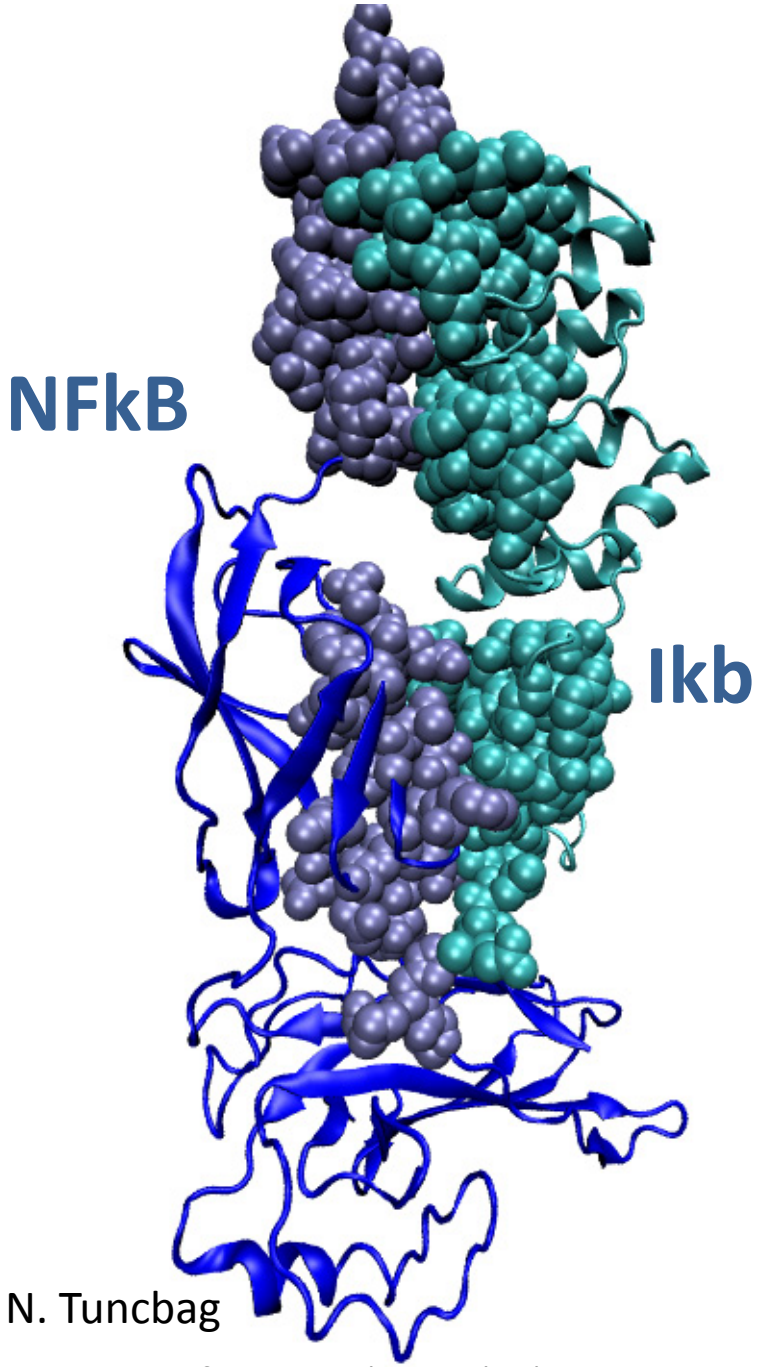


© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Clackson, Tim and James A. Wells. "A Hot Spot of Binding Energy in a Hormone-Receptor Interface." *Science* 267, no. 5196 (1995): 383-6.

Figure from Clackson & Wells (1995).

- Fewer than 10% of the residues at an interface contribute more than 2 kcal/mol to binding.
- Hot spots
  - rich in Trp, Arg and Tyr
  - occur on pockets on the two proteins that have complementary shapes and distributions of charged and hydrophobic residues.
  - can include buried charge residues far from solvent
  - O-ring structure excludes solvent from interface

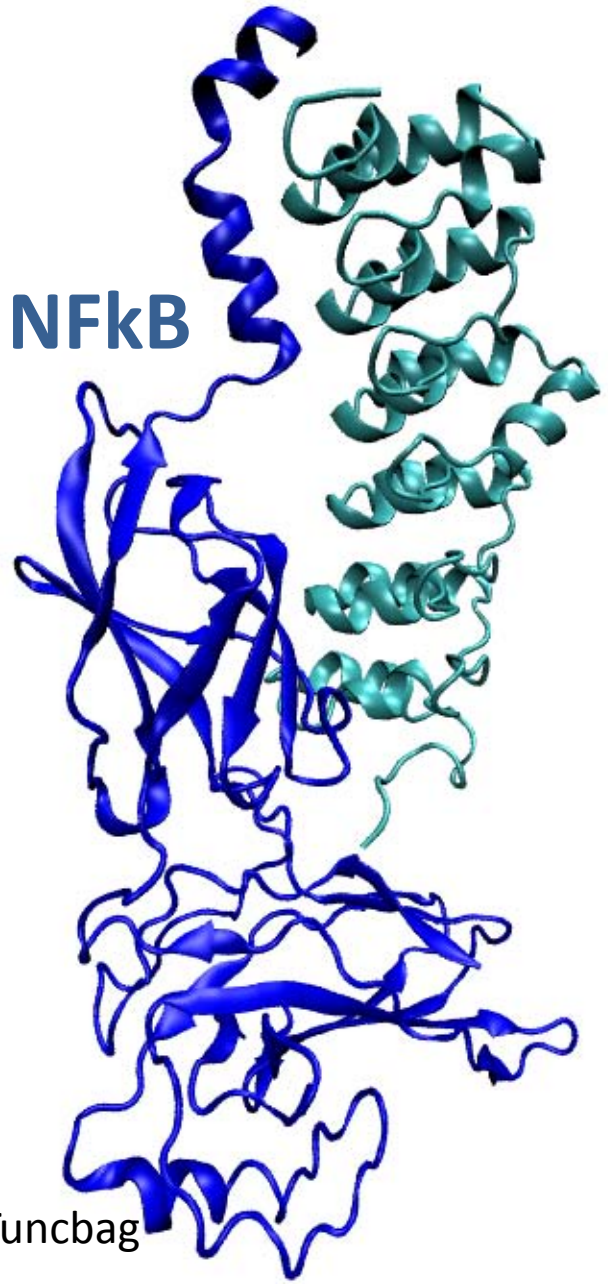
http://onlinelibrary.wiley.com/doi/10.1002/prot.21396/full



N. Tuncbag

Courtesy of Nurcan Tuncbag. Used with permission.

1. Identify interface of template (distance cutoff)



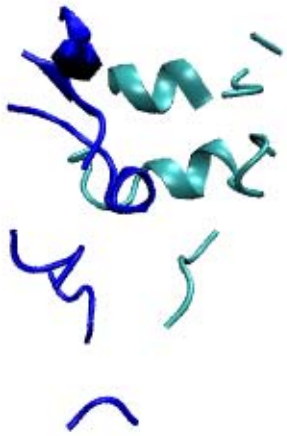
N. Tuncbag

Courtesy of Nurcan Tuncbag. Used with permission.

1. Identify interface of template (distance cutoff)

1. Identify interface of template (distance cutoff)

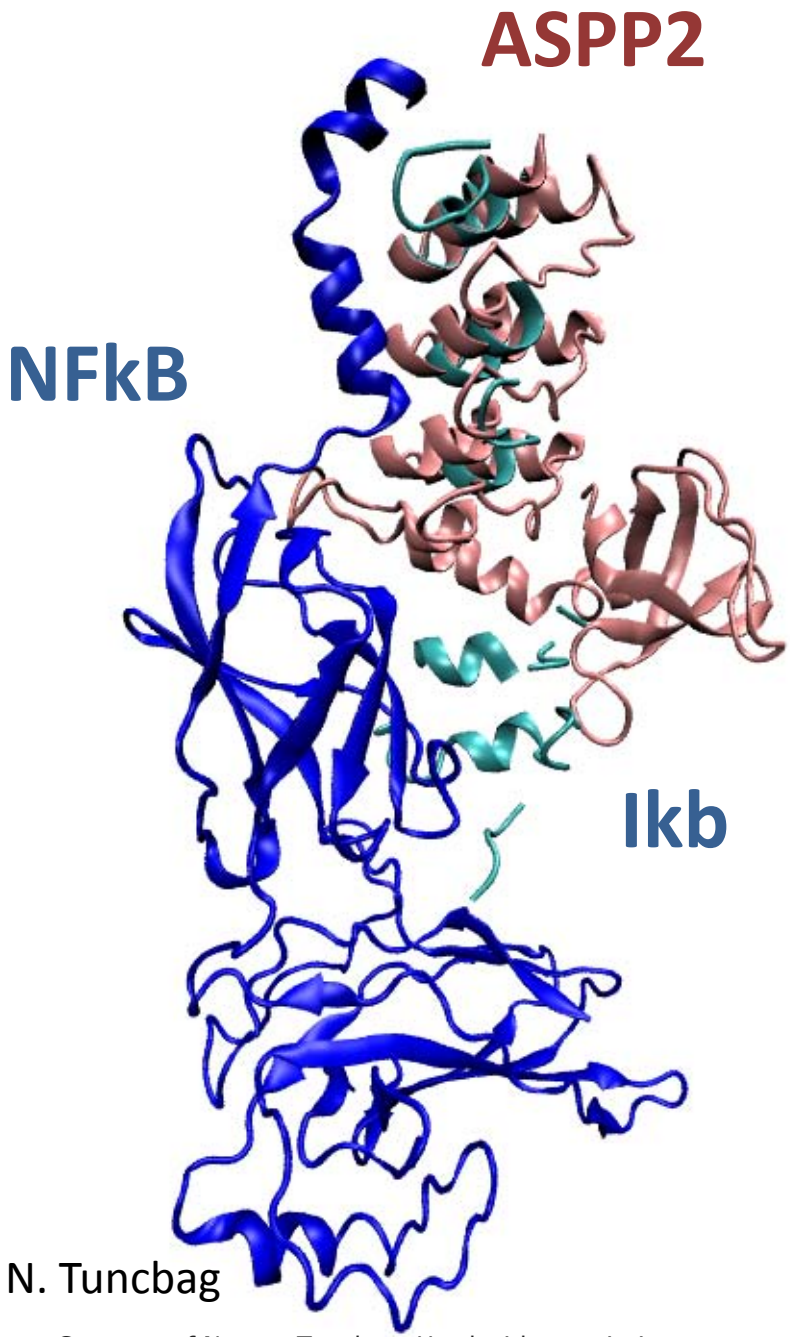
NFkB



Ikb

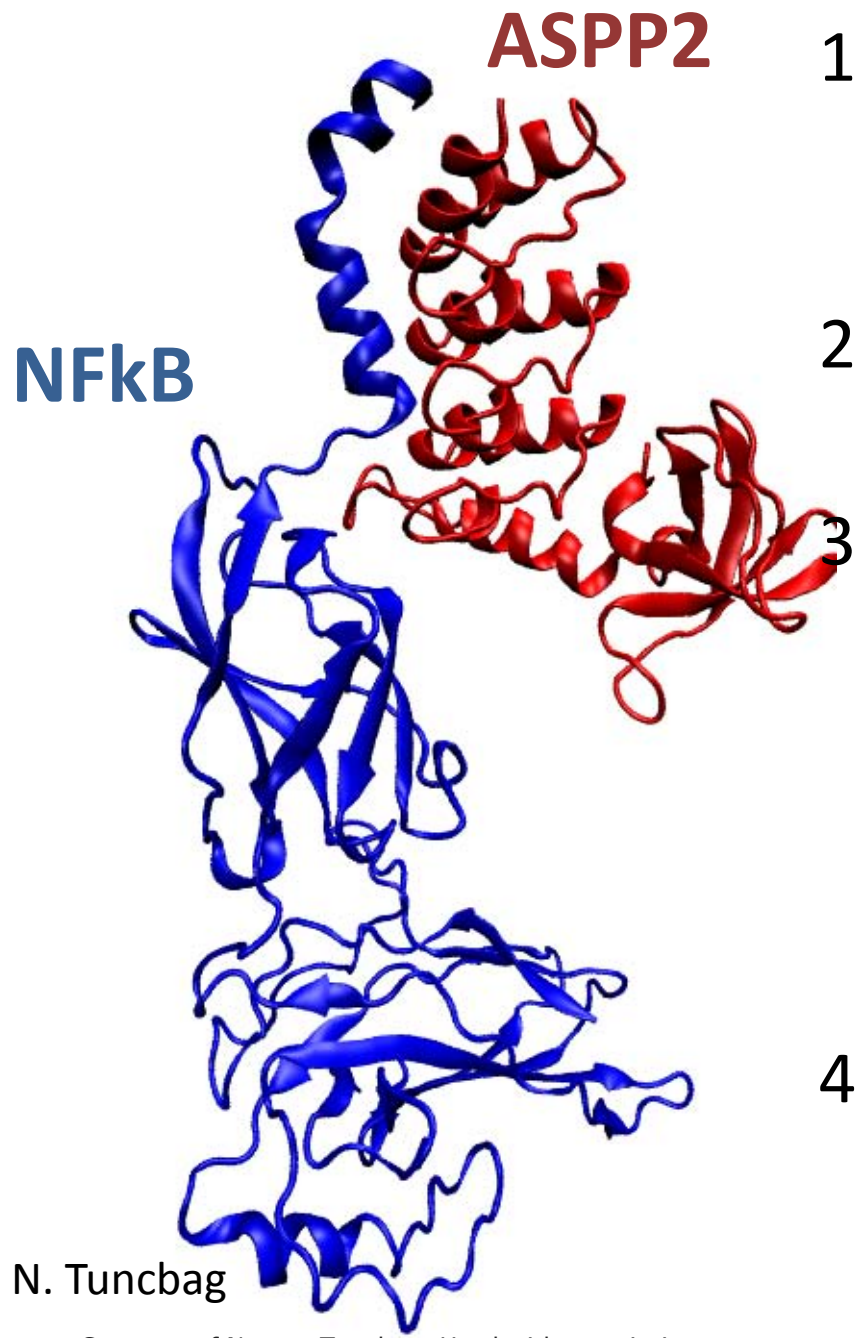
Courtesy of Nurcan Tuncbag. Used with permission.

N. Tuncbag



1. Identify interface of template (distance cutoff)
2. Align entire surface of query to half-interfaces
3. Test
    1. Overall structural match
    2. Structural match of at hotspots
    3. Sequence match at hotspots

Courtesy of Nurcan Tuncbag. Used with permission.

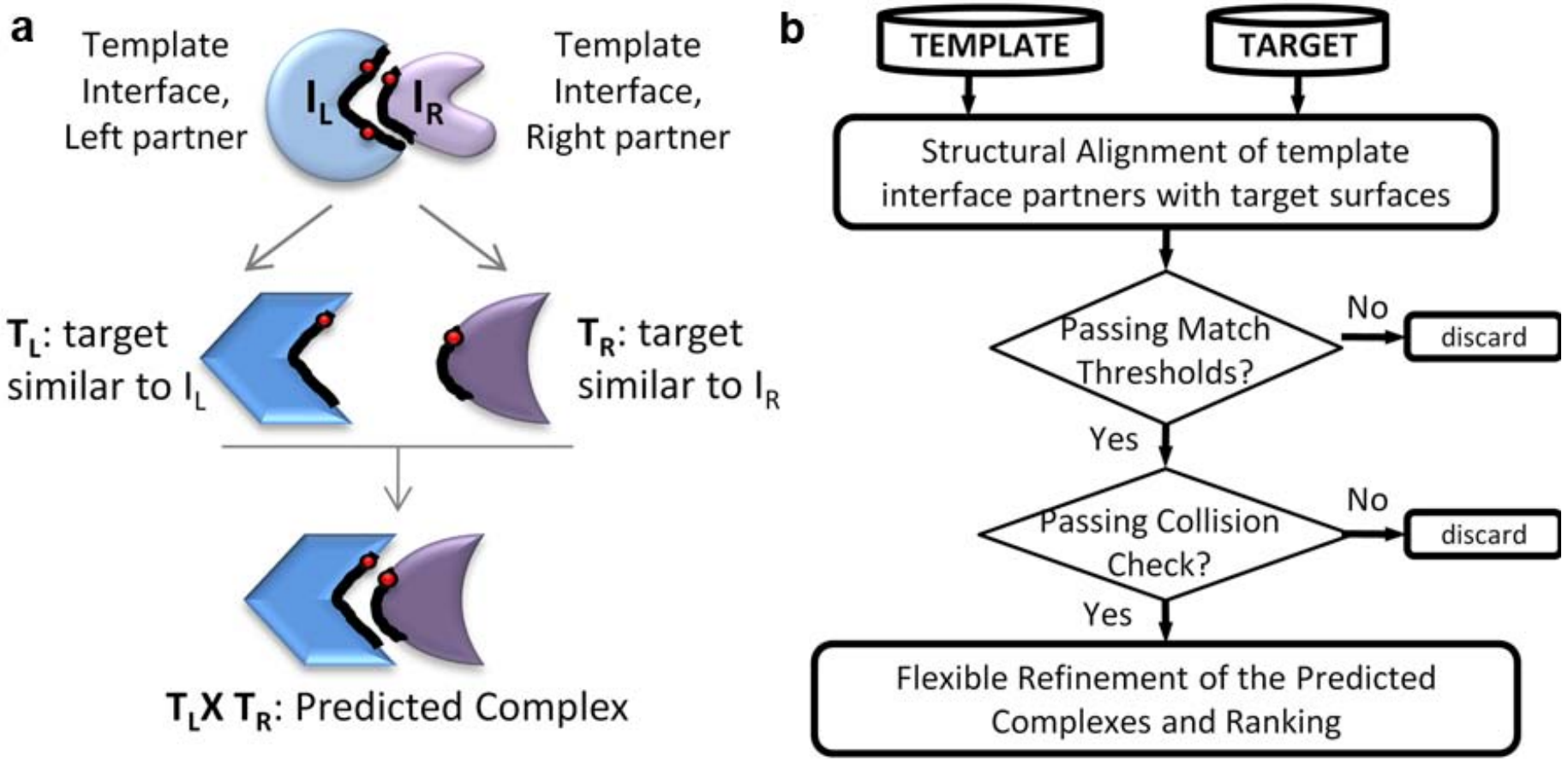


N. Tuncbag

Courtesy of Nurcan Tuncbag. Used with permission.

1. Identify interface of template (distance cutoff)
2. Align entire surface of query to half-interfaces
3. Test
    1. Overall structural match
    2. Structural match of at hotspots
    3. Sequence match at hotspot
4. Flexible refinement

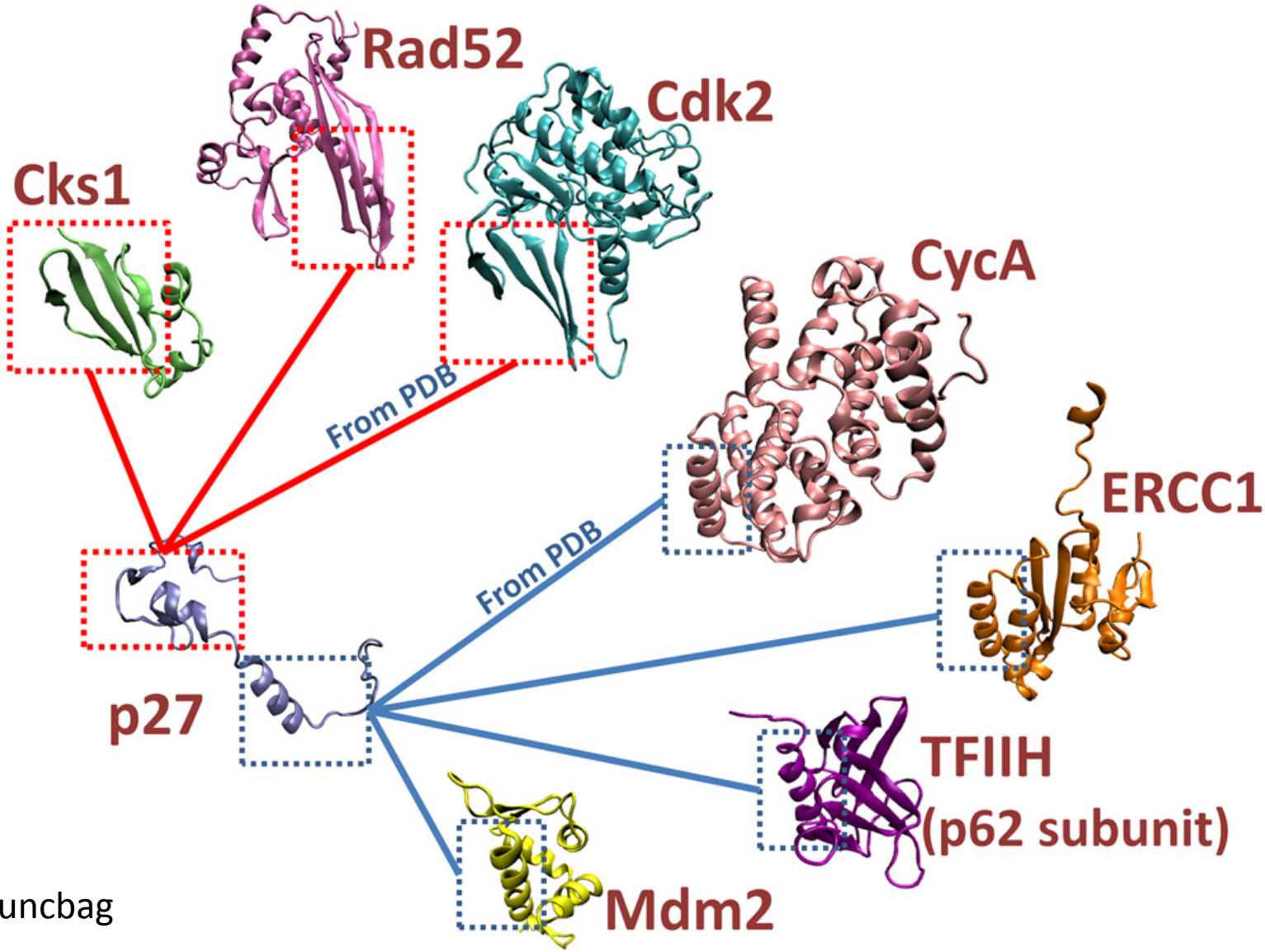
Flowchart



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Tuncbag, Nurcan, Attila Gursoy, et al. "Predicting Protein-protein Interactions on a Proteome Scale by Matching Evolutionary and Structural Similarities at Interfaces using PRISM." *Nature Protocols* 6, no. 9 (2011): 1341-54.

Structural match of template and target does not depend on order of residues

Predicted p27 Protein Partners



N. Tuncbag

Courtesy of Nurcan Tuncbag. Used with permission.

Next

PRISM

Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement.

Tuncbag N, Keskin O, Nussinov R, Gursoy A.

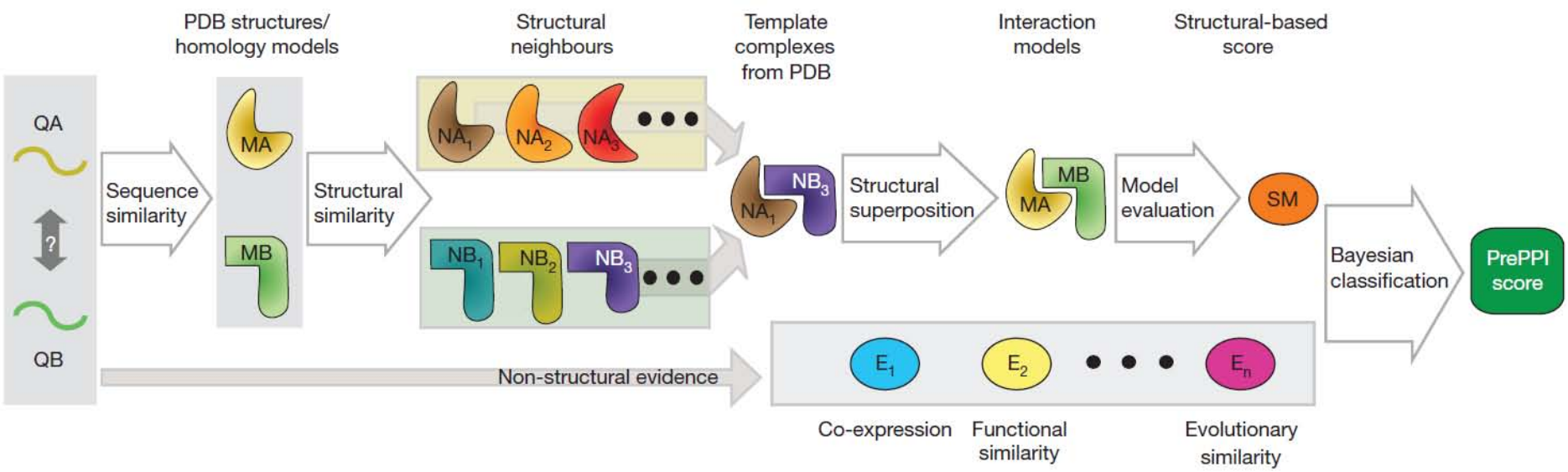
http://www.ncbi.nlm.nih.gov/pubmed/22275112

PrePPI

Structure-based prediction of protein–protein interactions on a genome-wide scale

Zhang, et al.

http://www.nature.com/nature/journal/v490/n7421/full/nature11503.html



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Zhang, Qiangfeng Cliff, Donald Petrey, et al. "Structure-based Prediction of Protein-protein Interactions on a Genome-wide Scale." *Nature* 490, no. 7421 (2012): 556-60.

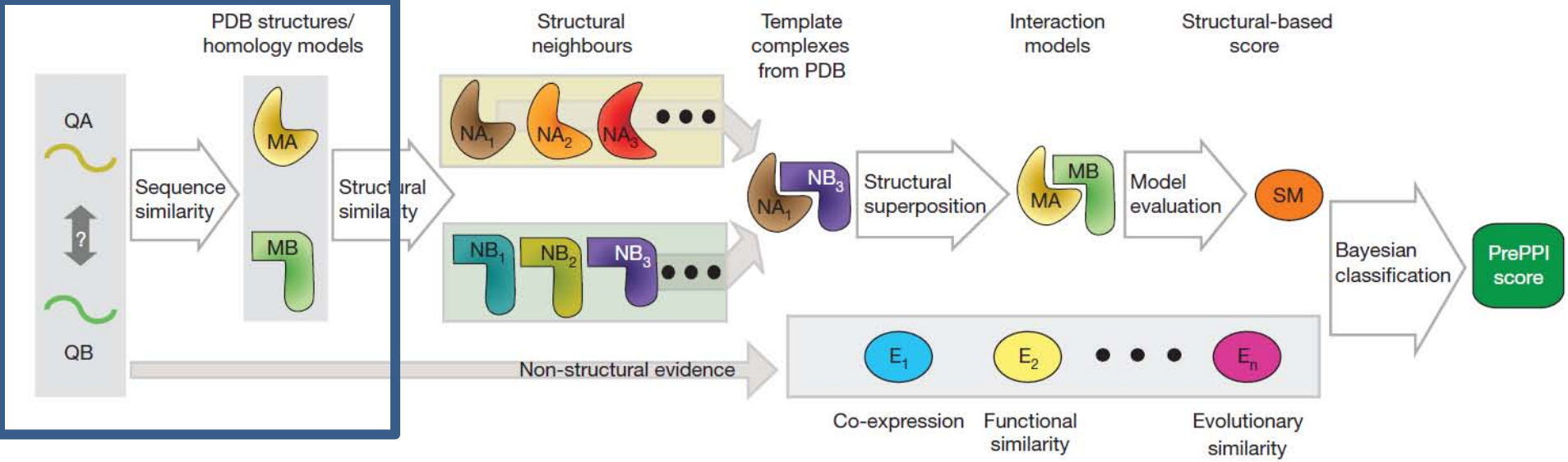
PrePPI

Scores potential templates without building a homology model

Criteria

- Geometric similarity between the protomer and template
- Statistics based on preservation of contact residues

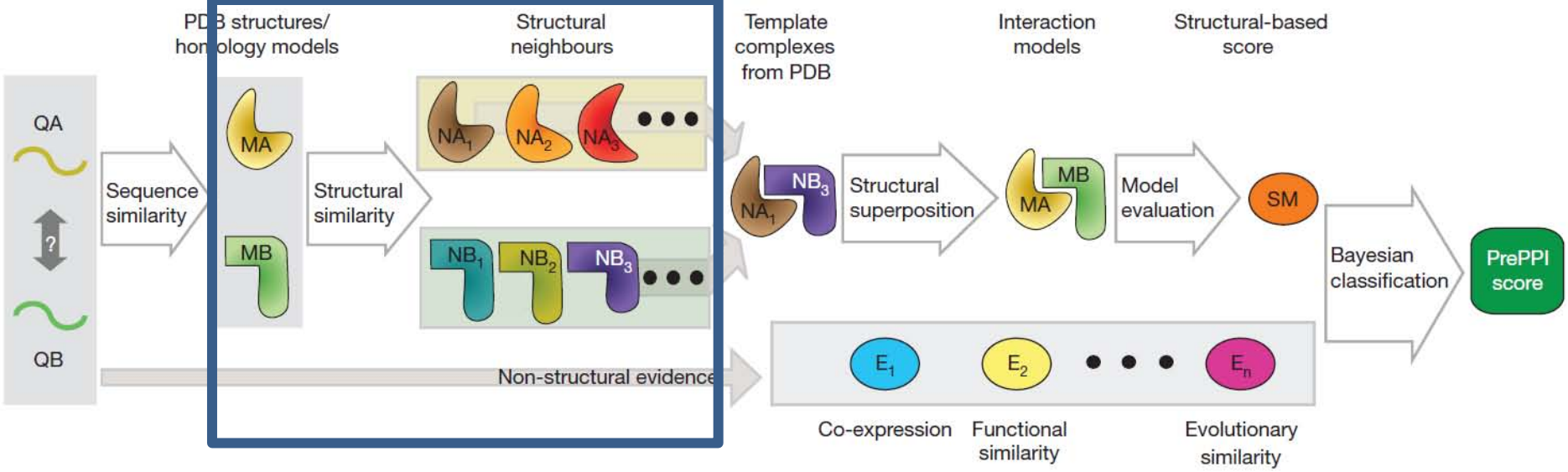
Structure-based prediction of protein–protein interactions on a genome-wide scale
Nature 490, 556–560 (25 October 2012) doi:10.1038/nature11503



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Zhang, Qiangfeng Cliff, Donald Petrey, et al. "Structure-based Prediction of Protein-protein Interactions on a Genome-wide Scale." *Nature* 490, no. 7421 (2012): 556-60.

1. Find homologous proteins of known structure (MA,MB)
3. Look for structure of a complex containing structural neighbors
2. Find structural neighbors (NA,NB) (avg. 1,500 neighbors/structure)
4. Align sequences of MA,MB to NA,NB based on structure

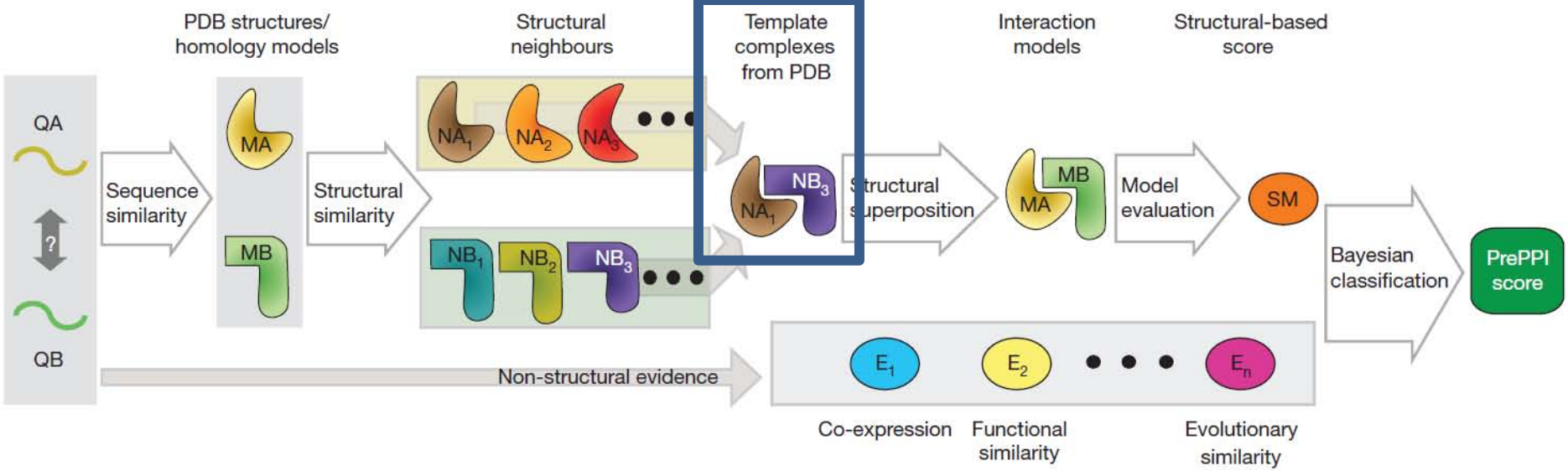
Structure-based prediction of protein–protein interactions on a genome-wide scale
Nature 490, 556–560 (25 October 2012) doi:10.1038/nature11503



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Zhang, Qiangfeng Cliff, Donald Petrey, et al. "Structure-based Prediction of Protein-protein Interactions on a Genome-wide Scale." *Nature* 490, no. 7421 (2012): 556-60.

1. Find homologous proteins of known structure (MA,MB)
2. Find structural neighbors (NA_i,NB_i)(avg:1,500 neighbors/structure)
3. look for structure of a complex containing structural neighbors
4. Align sequences of MA,MB to NA,NB based on structure

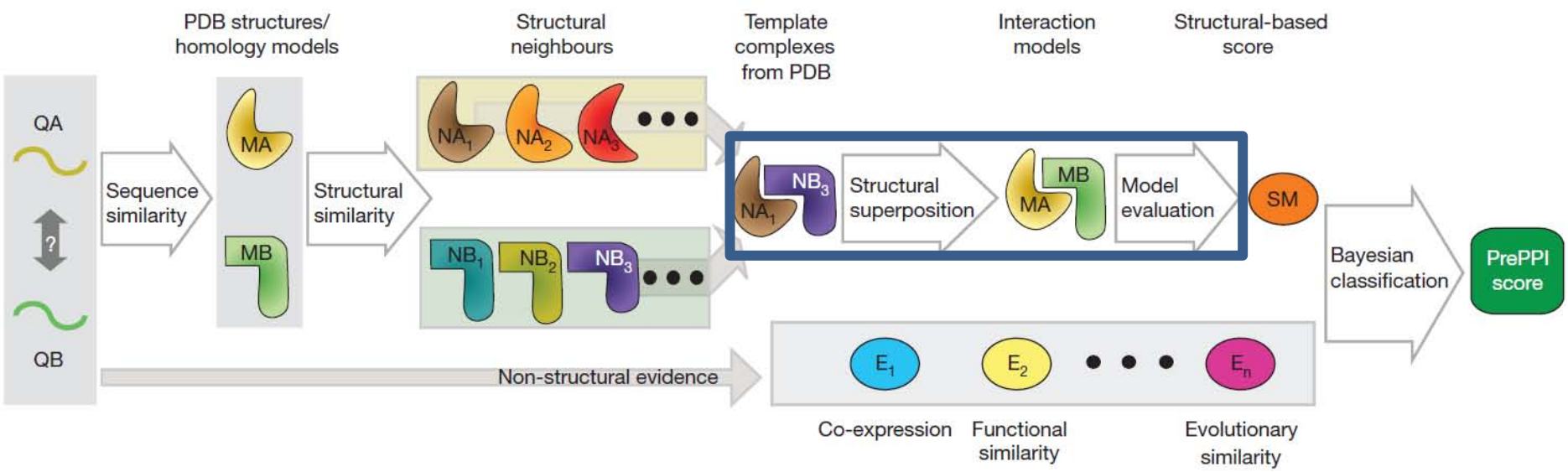
Structure-based prediction of protein–protein interactions on a genome-wide scale
Nature 490, 556–560 (25 October 2012) doi:10.1038/nature11503



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Zhang, Qiangfeng Cliff, Donald Petrey, et al. "Structure-based Prediction of Protein-protein Interactions on a Genome-wide Scale." *Nature* 490, no. 7421 (2012): 556-60.

1. Find homologous proteins of known structure (MA,MB)
2. Find structural neighbors (NA_i,NB_i)(avg:1,500 neighbors/structure)
3. Look for structure of a complex containing structural neighbors
4. Align sequences of MA,MB to NA,NB based on structure

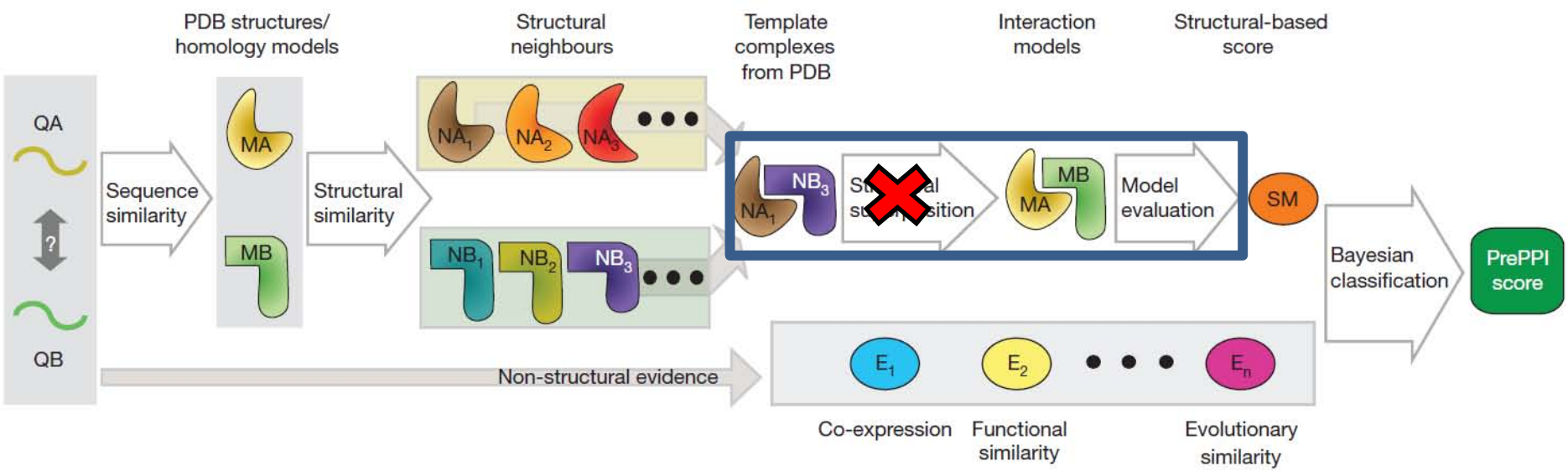
Structure-based prediction of protein–protein interactions on a genome-wide scale
Nature 490, 556–560 (25 October 2012) doi:10.1038/nature11503



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Zhang, Qiangfeng Cliff, Donald Petrey, et al. "Structure-based Prediction of Protein-protein Interactions on a Genome-wide Scale." *Nature* 490, no. 7421 (2012): 556-60.

1. Find homologous proteins of known structure (MA,MB)
2. Find structural neighbors (NA_i,NB_i)(avg:1,500 neighbors/structure)
3. Look for structure of a complex containing structural neighbors
4. Align sequences of MA,MB to NA,NB based on structure

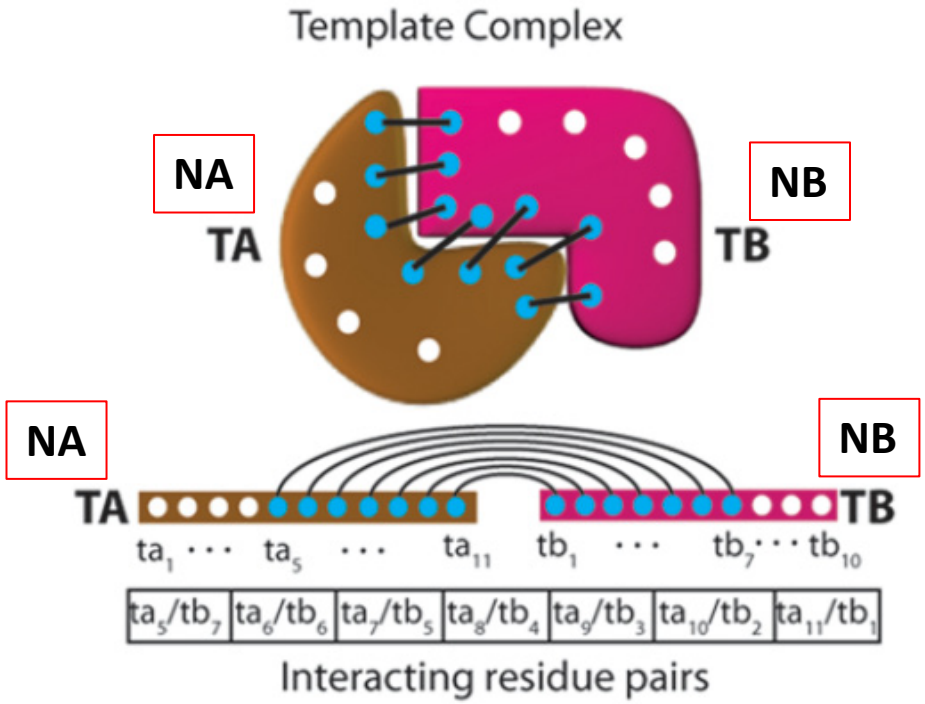
Structure-based prediction of protein–protein interactions on a genome-wide scale
Nature 490, 556–560 (25 October 2012) doi:10.1038/nature11503



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Zhang, Qiangfeng Cliff, Donald Petrey, et al. "Structure-based Prediction of Protein-protein Interactions on a Genome-wide Scale." *Nature* 490, no. 7421 (2012): 556-60.

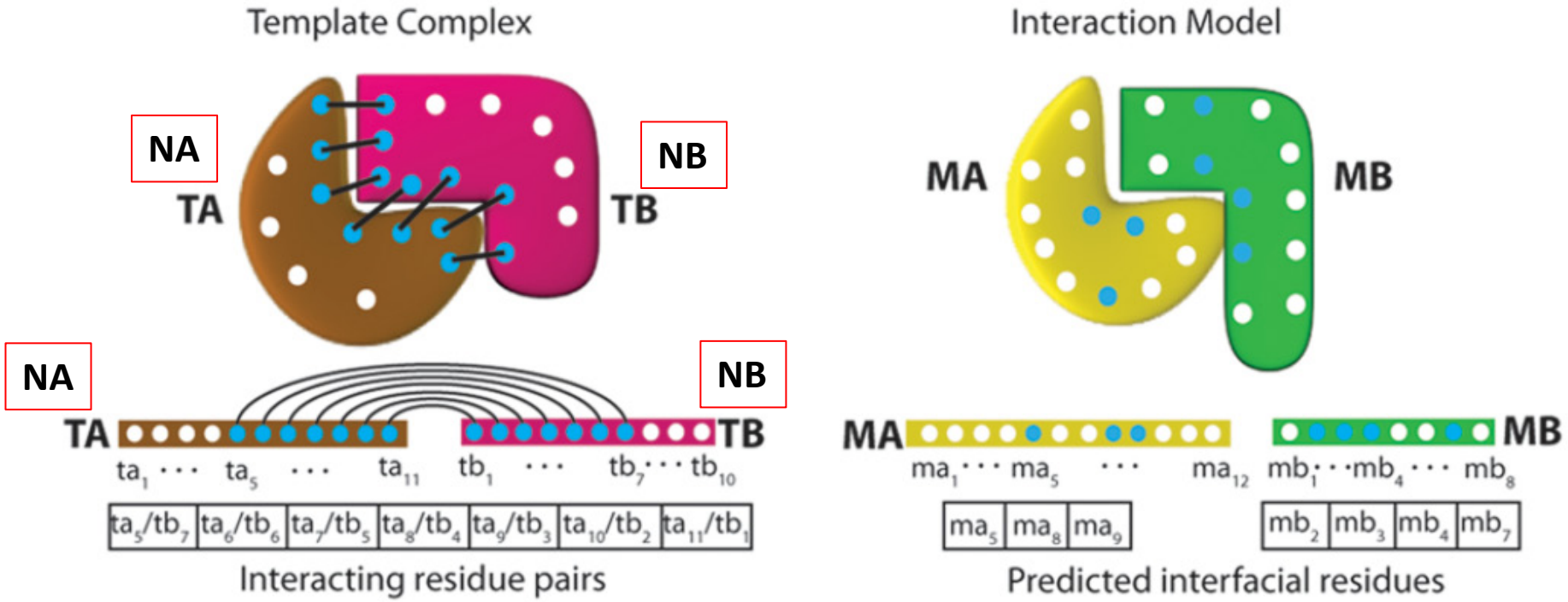
1. Find homologous proteins of known structure (MA,MB)
2. Find structural neighbors (NA_i,NB_i)(avg:1,500 neighbors/structure)
3. Look for structure of a complex containing structural neighbors
4. Align sequences of MA,MB to NA,NB based on structure

Structure-based prediction of protein–protein interactions on a genome-wide scale
Nature 490, 556–560 (25 October 2012) doi:10.1038/nature11503



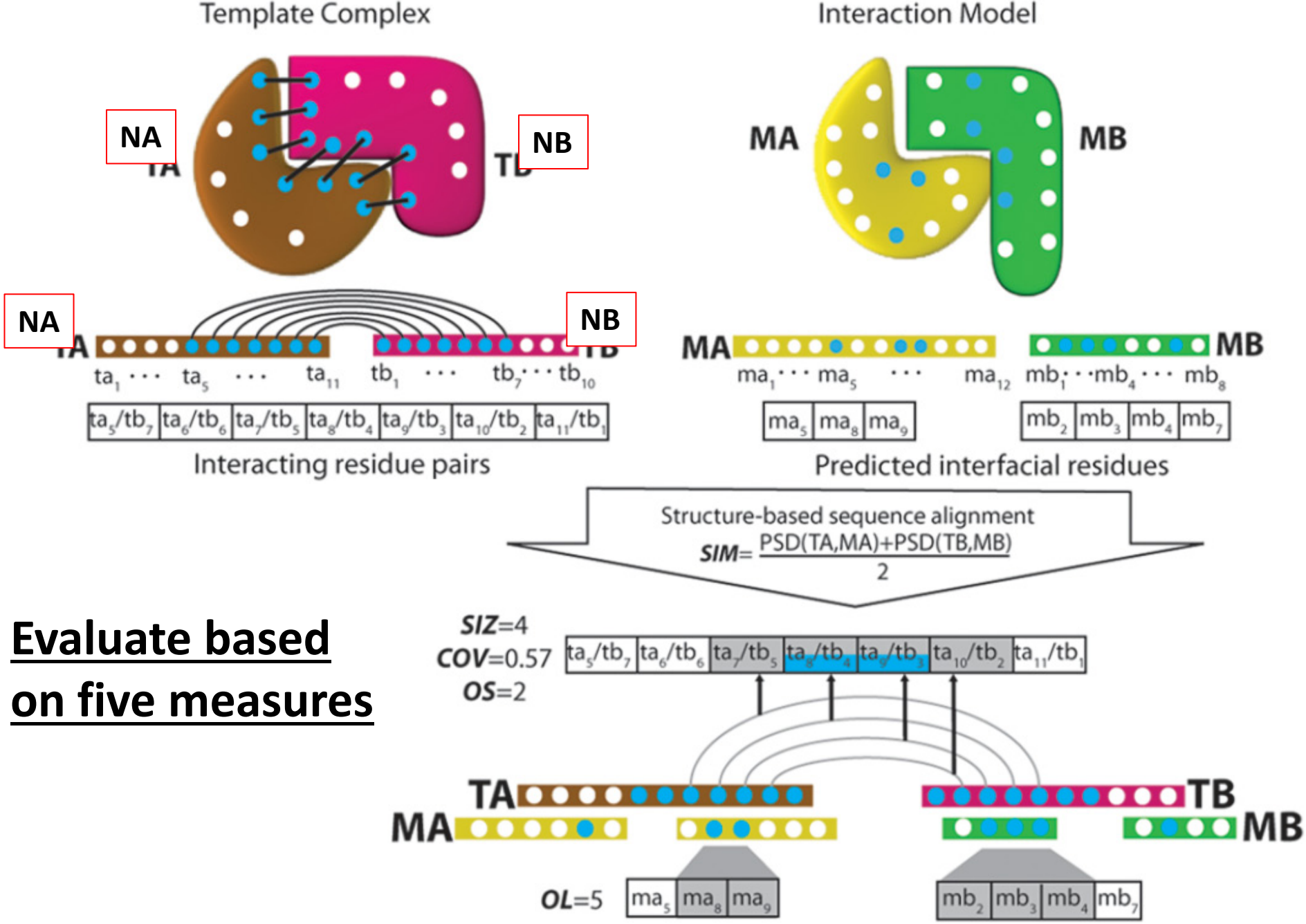
© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

1. Identify interacting residues in template complex (Called NA1 NB3 in rest of paper)



© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

1. Identify interacting residues in template complex (Called NA1 NB3 in rest of paper)
2. Predict interacting residues for the homology models

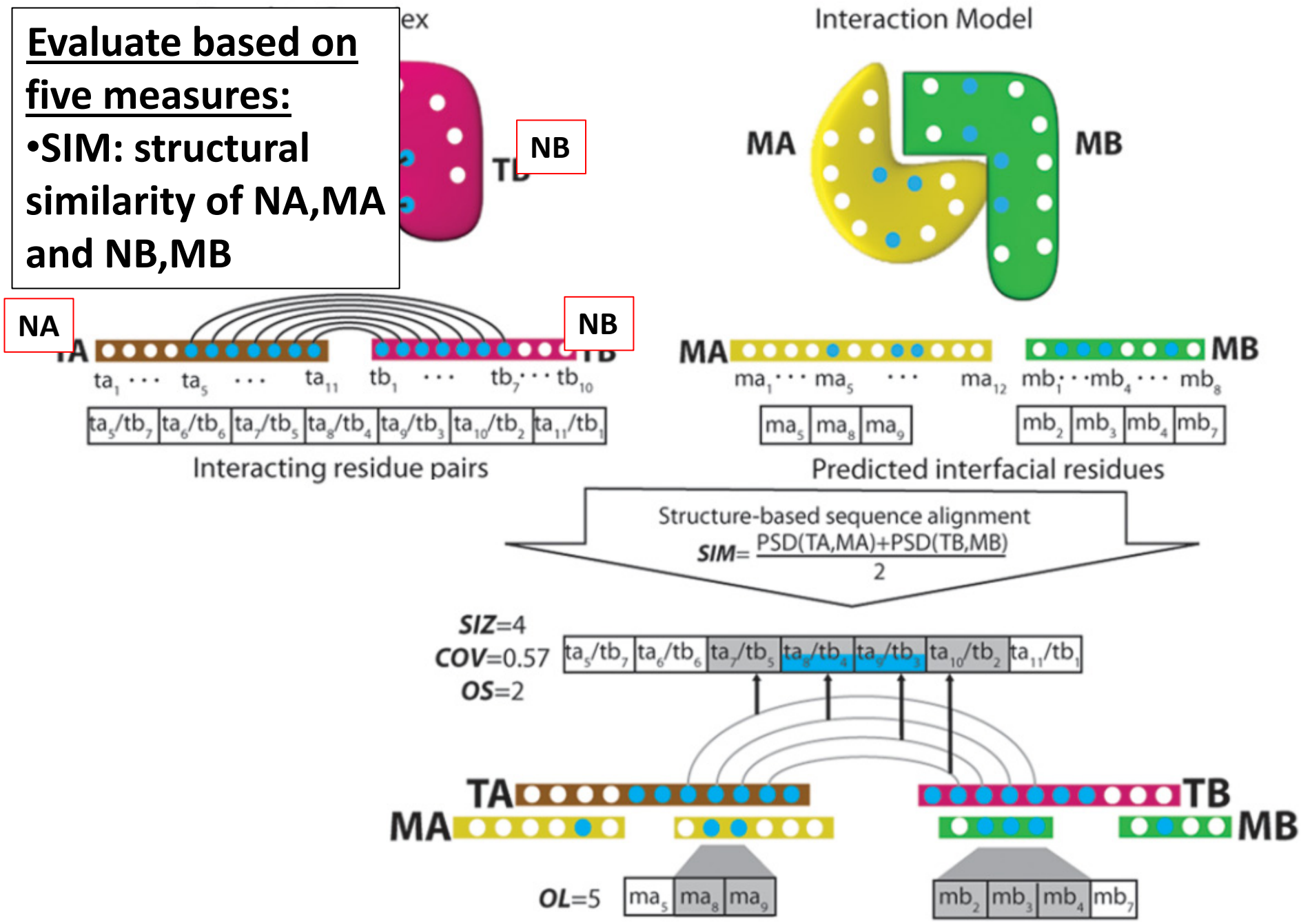


Evaluate based on five measures

© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

Evaluate based on five measures:

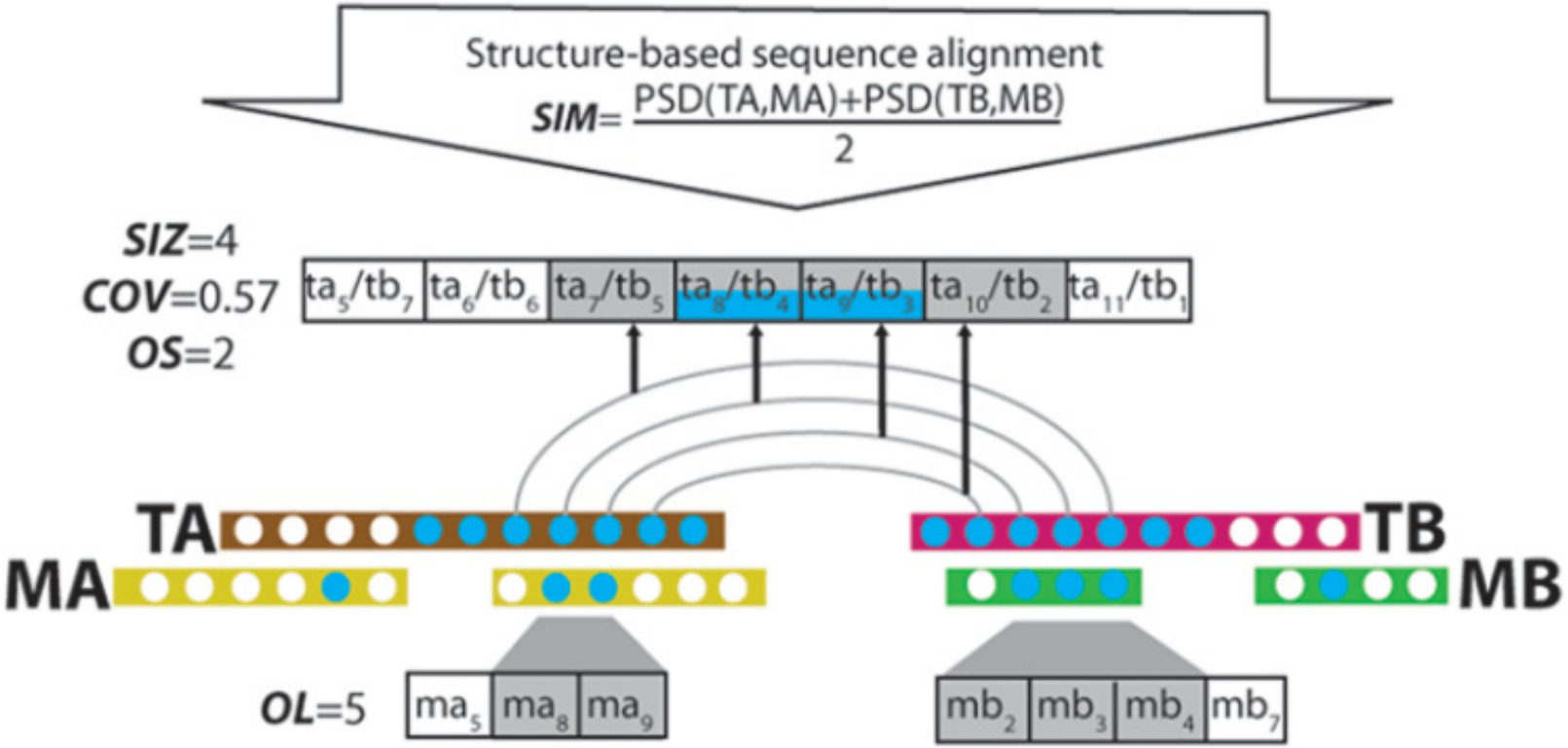
- **SIM: structural similarity of NA,MA and NB,MB**



© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

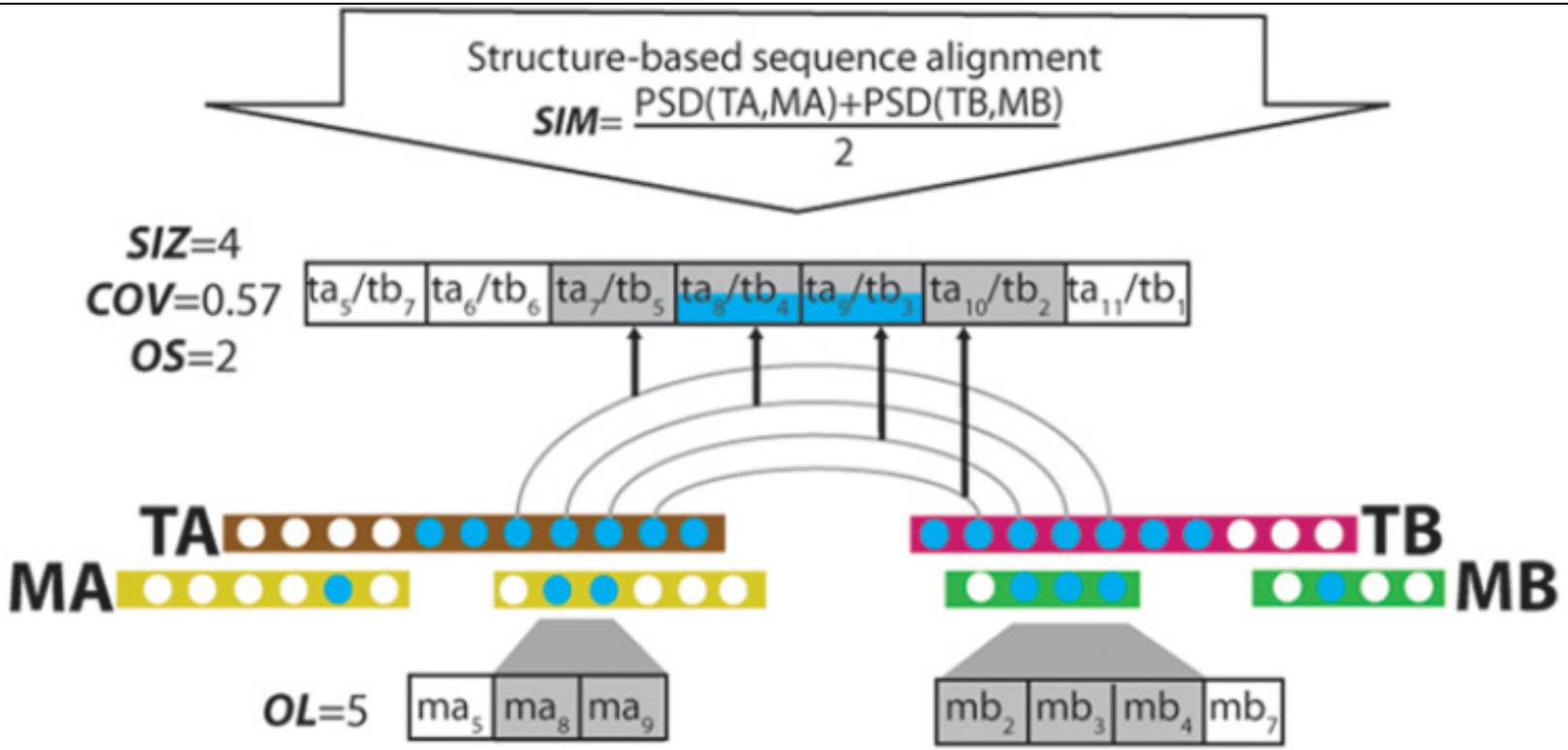
Evaluate based on five measures:

- **SIM: structural similarity of NA,MA and NB,MB**
- **SIZ (number) COV (fraction) of interaction pairs can be aligned anywhere**
- **OS subset of SIZ at interface**
- **OL number of aligned pairs at interface**

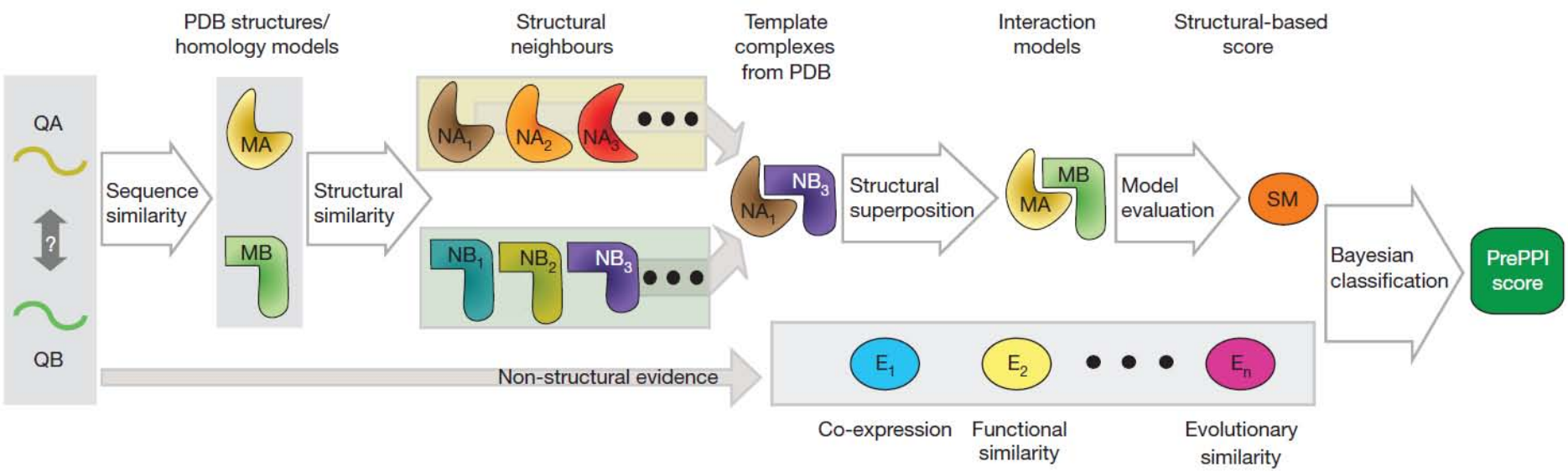


© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

"The final two scores reflect whether the residues that appear in the model interface have properties consistent with those that mediate known PPIs (for example, residue type, evolutionary conservation, or statistical propensity to be in protein–protein interfaces)." ????



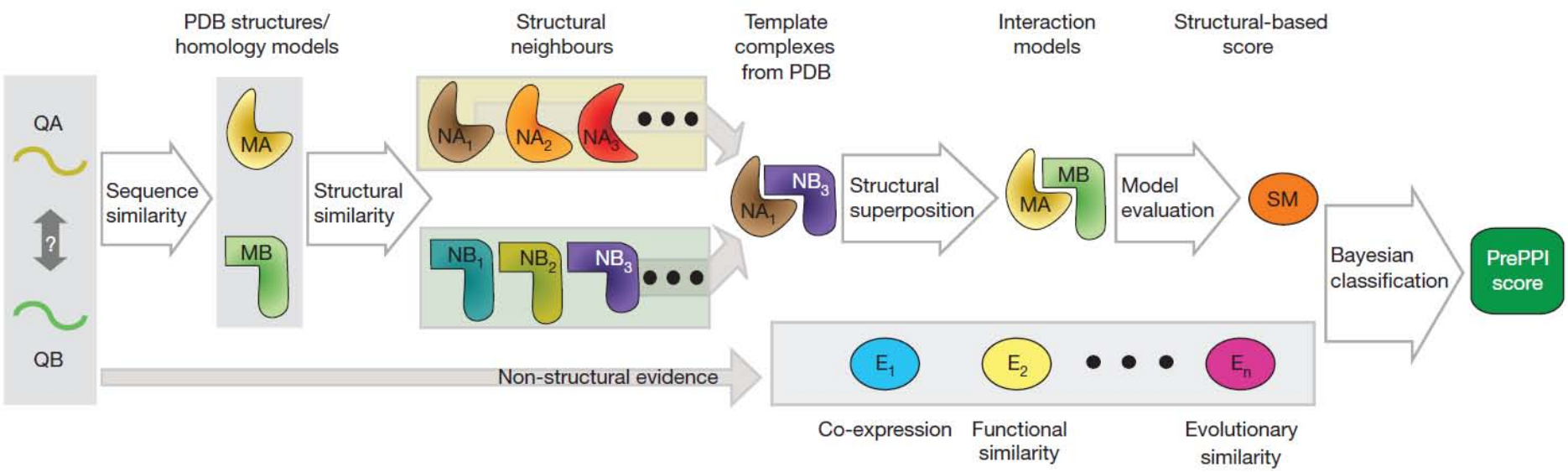
© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Zhang, Qiangfeng Cliff, Donald Petrey, et al. "Structure-based Prediction of Protein-protein Interactions on a Genome-wide Scale." *Nature* 490, no. 7421 (2012): 556-60.

1. Find homologous proteins of known structure (MA,MB)
2. Find structural neighbors (NA_i,NB_i)(avg:1,500 neighbors/structure)
3. Look for structure of a complex containing structural neighbors
4. Align sequences of MA,MB to NA,NB based on structure
5. Compute five scores
6. Train Bayesian classifier using "gold standard" interactions

Structure-based prediction of protein–protein interactions on a genome-wide scale
Nature 490, 556–560 (25 October 2012) doi:10.1038/nature11503



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Zhang, Qiangfeng Cliff, Donald Petrey, et al. "Structure-based Prediction of Protein-protein Interactions on a Genome-wide Scale." *Nature* 490, no. 7421 (2012): 556-60.

1. Find homologous proteins of known structure (MA,MB)
2. Find structural neighbors (NA_i,NB_i)(avg:1,500 neighbors/structure)
3. Look for structure of a complex containing structural neighbors
4. Align sequences of MA,MB to NA,NB based on structure
5. Compute five scores
6. Train Bayesian classifier using "gold standard" interactions

We will examine Bayesian classifiers soon

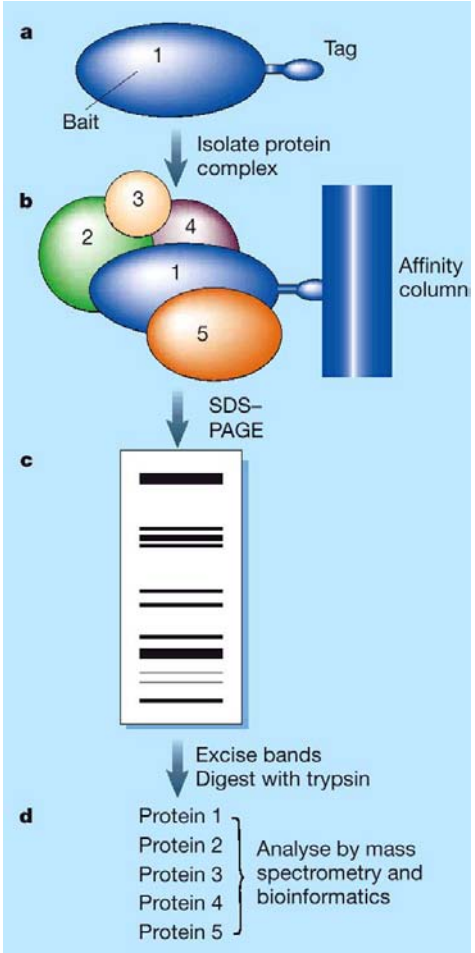
Outline

- Structural prediction of protein-protein interactions
- High-throughput measurement of protein-protein interactions
- Estimating interaction probabilities
- Bayes Net predictions of protein-protein interactions

Detecting protein-protein interactions

What are the likely false positives?

What are the likely false negatives?



Gavin, A.-C. *et al*. *Nature* **415**, 141-147 (2002).

Ho, Y. *et al*. *Nature* **415**, 180-183 (2002).

Proteomics: Protein complexes take the bait

Anuj Kumar and Michael Snyder

Nature 415, 123-124(10 January 2002)

doi:10.1038/415123a

Courtesy of Macmillan Publishers Limited.
Used with permission.

Source: Kumar, Anuj, and Michael Snyder. "Proteomics: Protein Complexes take the Bait." *Nature* 415, no. 6868 (2002): 123-4.

40

Mass-spec for protein-protein interactions

- Extremely efficient method for detecting interactions
- Proteins are in their correct subcellular location.

Limitations?

Mass-spec for protein-protein interactions

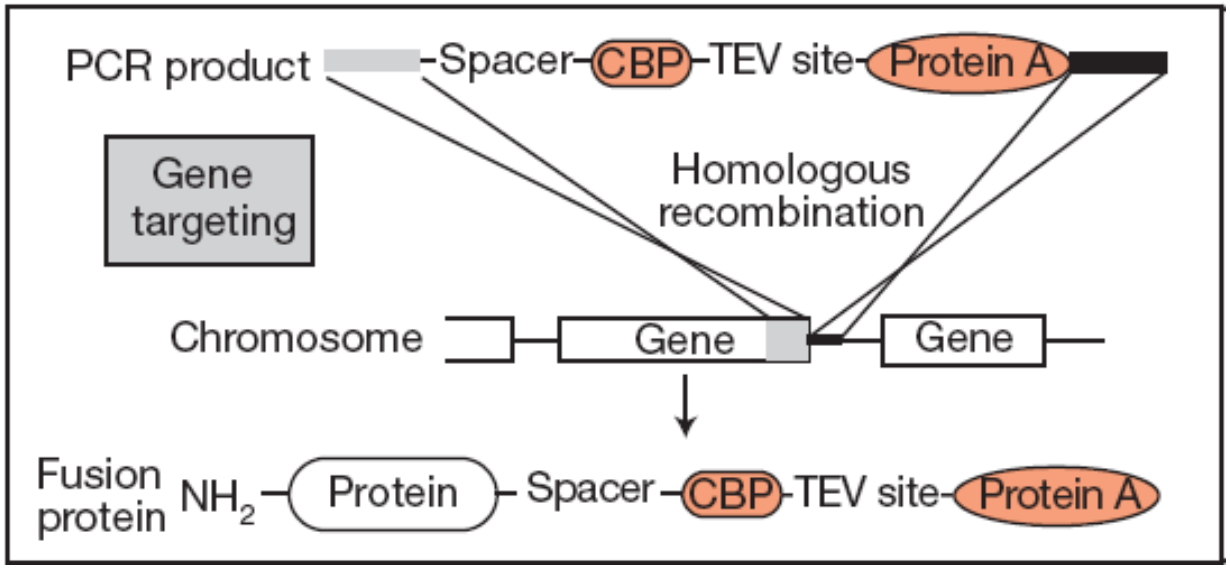
- Extremely efficient method for detecting interactions
- Proteins are in their correct subcellular location.

Limitations?

- overexpression/tagging can influence results
- only long-lived complexes will be detected

Tagging strategies

Gavin et al. (2002) Nature.



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Gavin, Anne-Claude, Markus Bösche, et al. "Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes." *Nature* 415, no. 6868 (2002): 141-7.

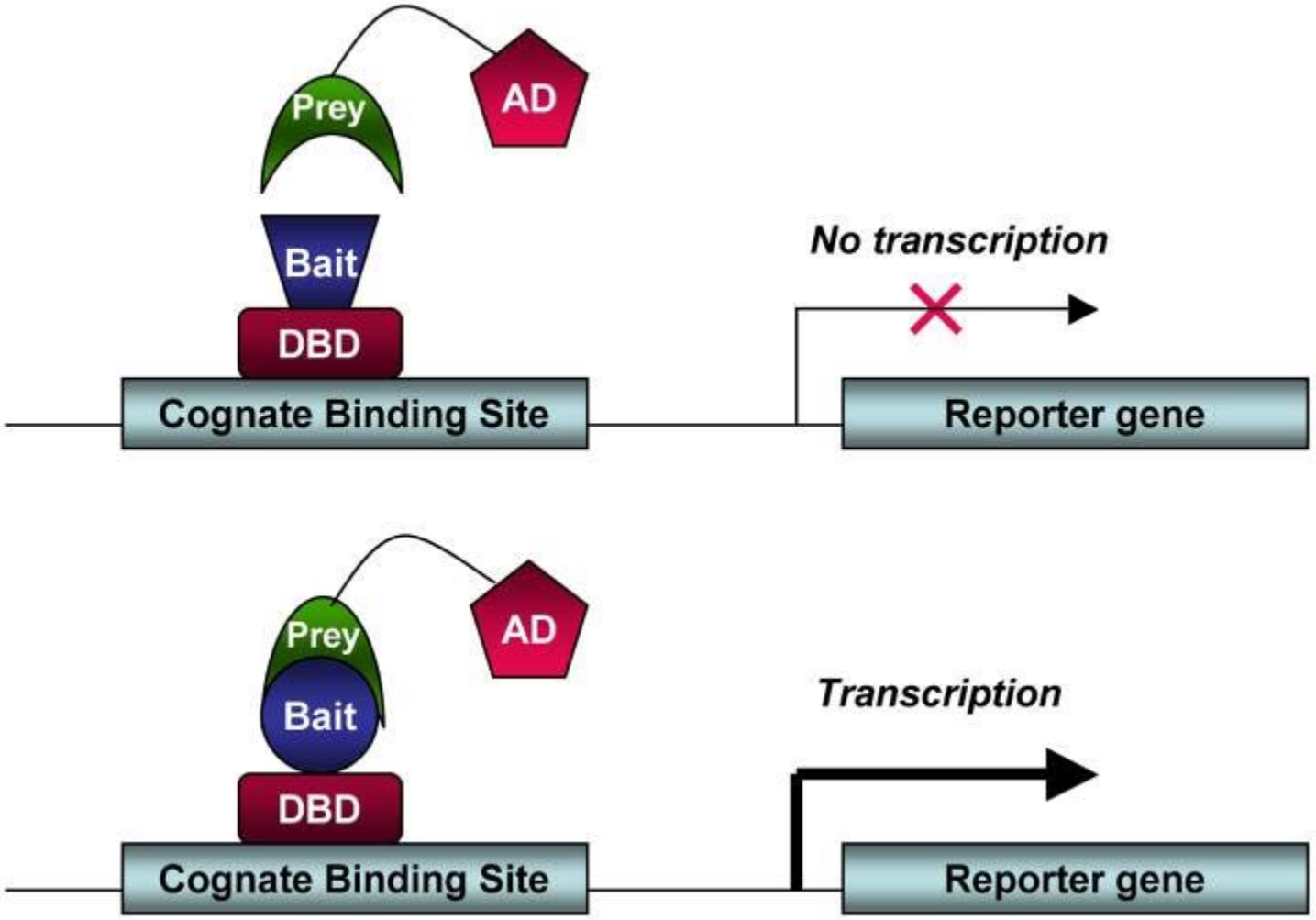
TAP-tag (Endogenous protein levels)

Tandem purification

1. Protein A-IgG purification
2. Cleave TEV site to elute
3. CBP-Calmodulin purification
4. EGTA to elute

Ho et al. (2002) Nature over-expressed proteins and used only one tag.

Yeast two-hybrid

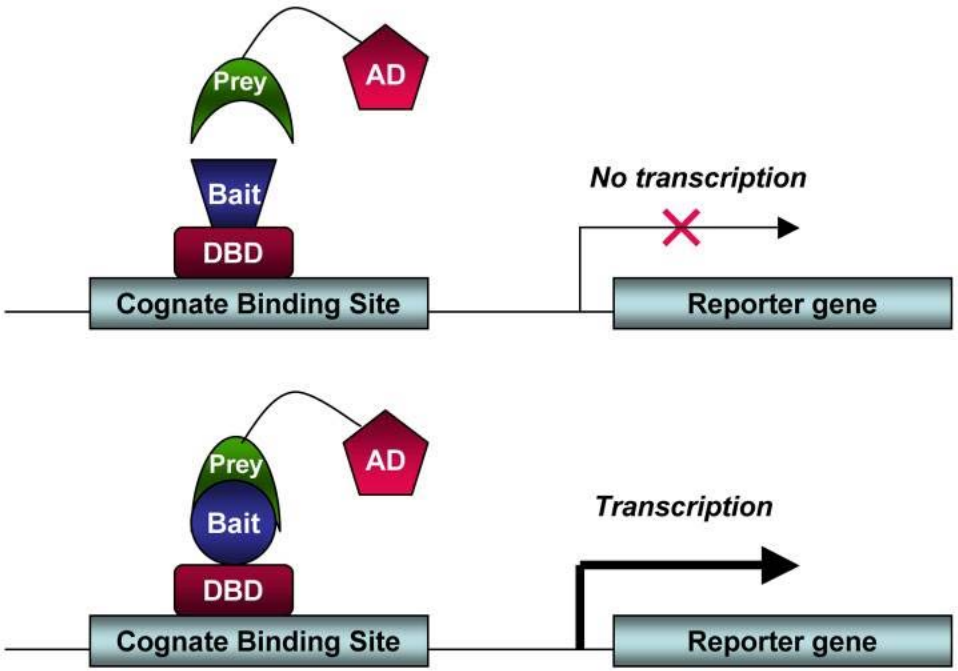


Courtesy of BioTechniques. Used with permission.

Source: Ratushny, Vladimi, and Erica A. Golemis. "Resolving the Network of Cell Signaling Pathways using the Evolving Yeast Two-hybrid System." *Biotechniques* 44, no. 5 (2008): 655.

Biotechniques. 2008 Apr;44(5):655-62.
Ratushny V, Golemis E.

How does this compare to mass-spec based approaches



Courtesy of BioTechniques. Used with permission.
Source: Ratushny, Vladimi, and Erica A. Golemis. "Resolving the Network of Cell Signaling Pathways using the Evolving Yeast Two-hybrid System." *Biotechniques* 44, no. 5 (2008): 655.

- **Does not require purification – will pick up more transient interactions.**
- **Biased against proteins that do not express well, or are incompatible with the nucleus**

Biotechniques. 2008 Apr;44(5):655-62.

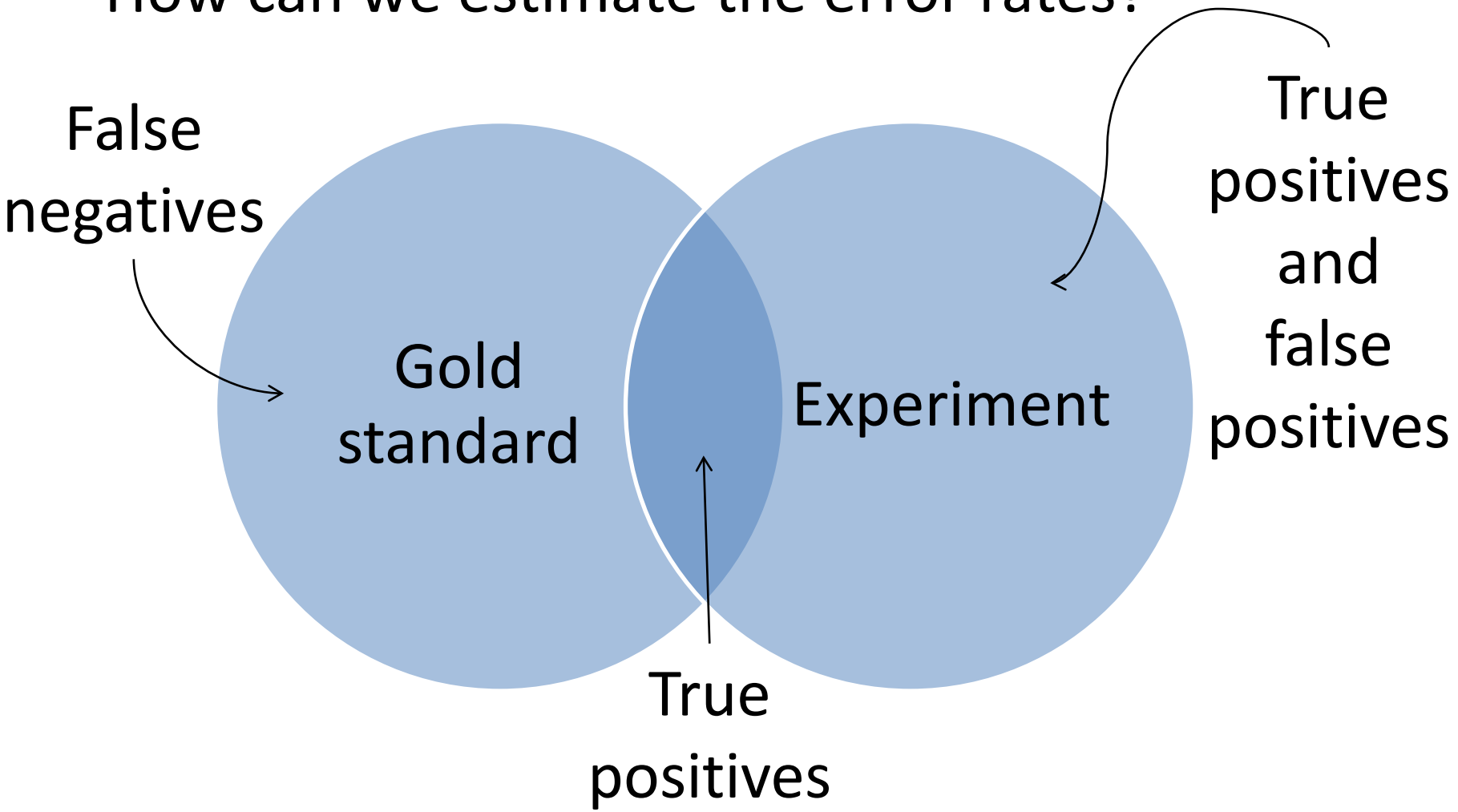
RatushnyV , Golemis E.

Outline

- Structural prediction of protein-protein interactions
- High-throughput measurement of protein-protein interactions
- Estimating interaction probabilities
- Bayes Net predictions of protein-protein interactions

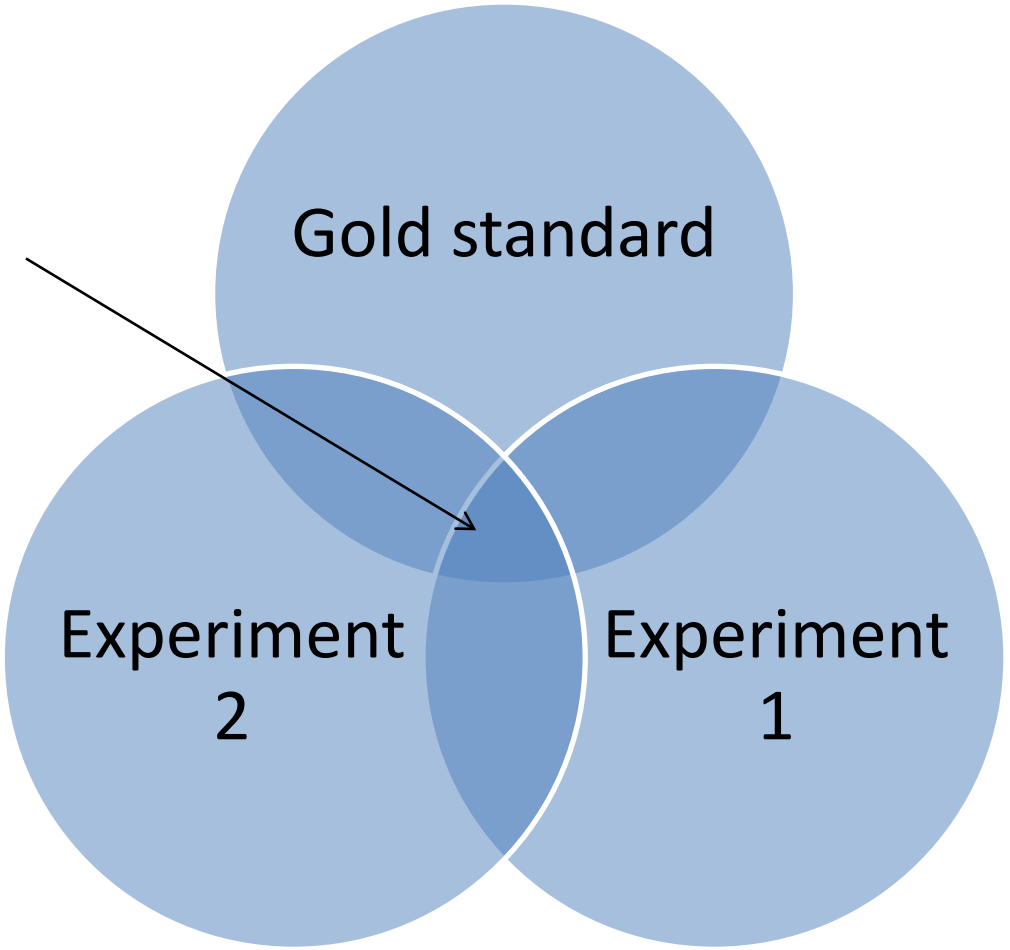
Error Rates

- How can we estimate the error rates?



Error Rates

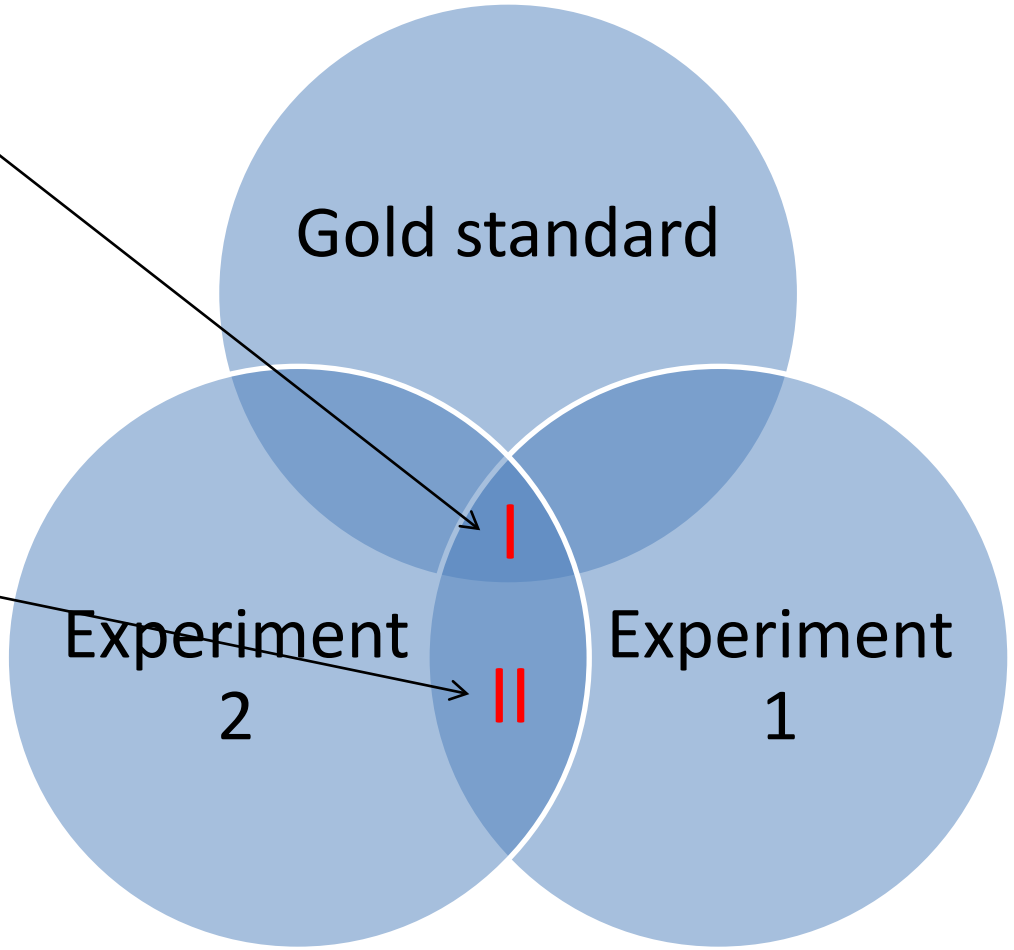
True positives from gold standard



Data Integration

I =True positives from gold standard

II=Consensus true positives



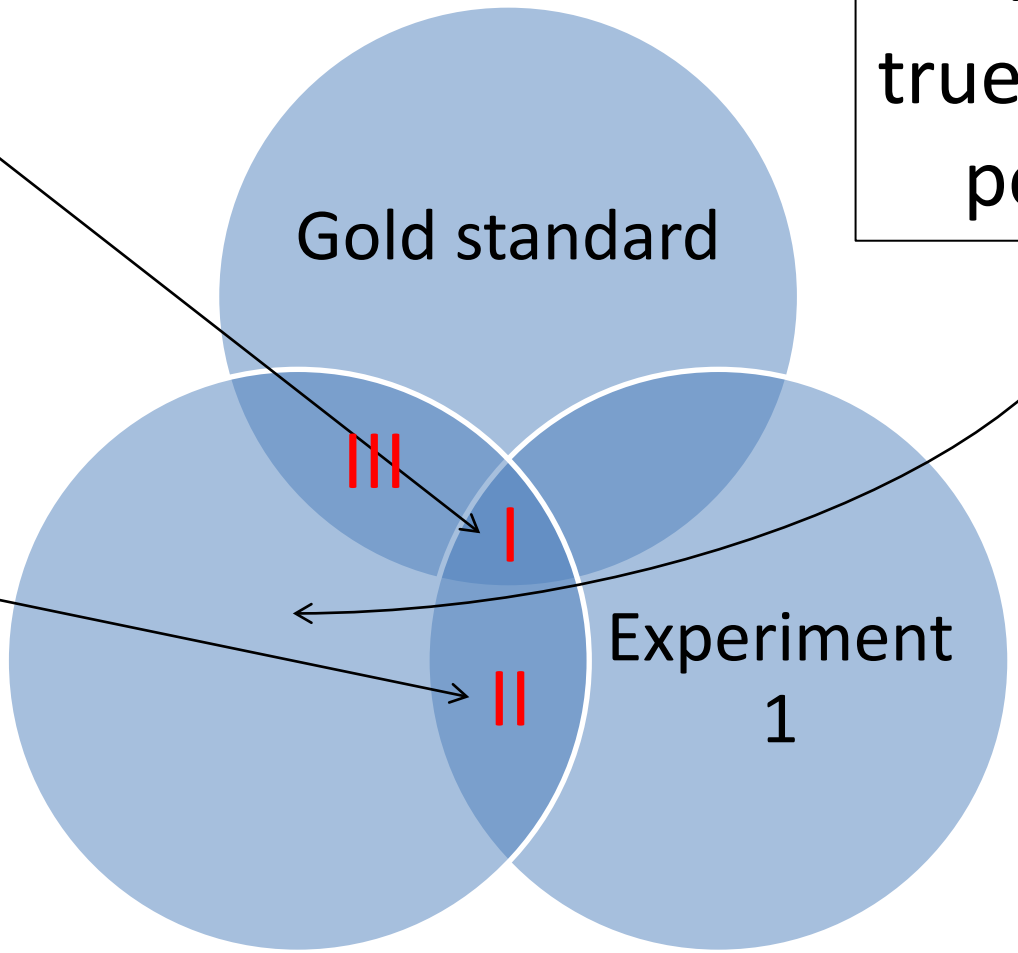
Fraction of consensus present in gold standard=I/II

Data Integration

I =True positives from gold standard

Mix of true and false positives

II=Consensus true positives



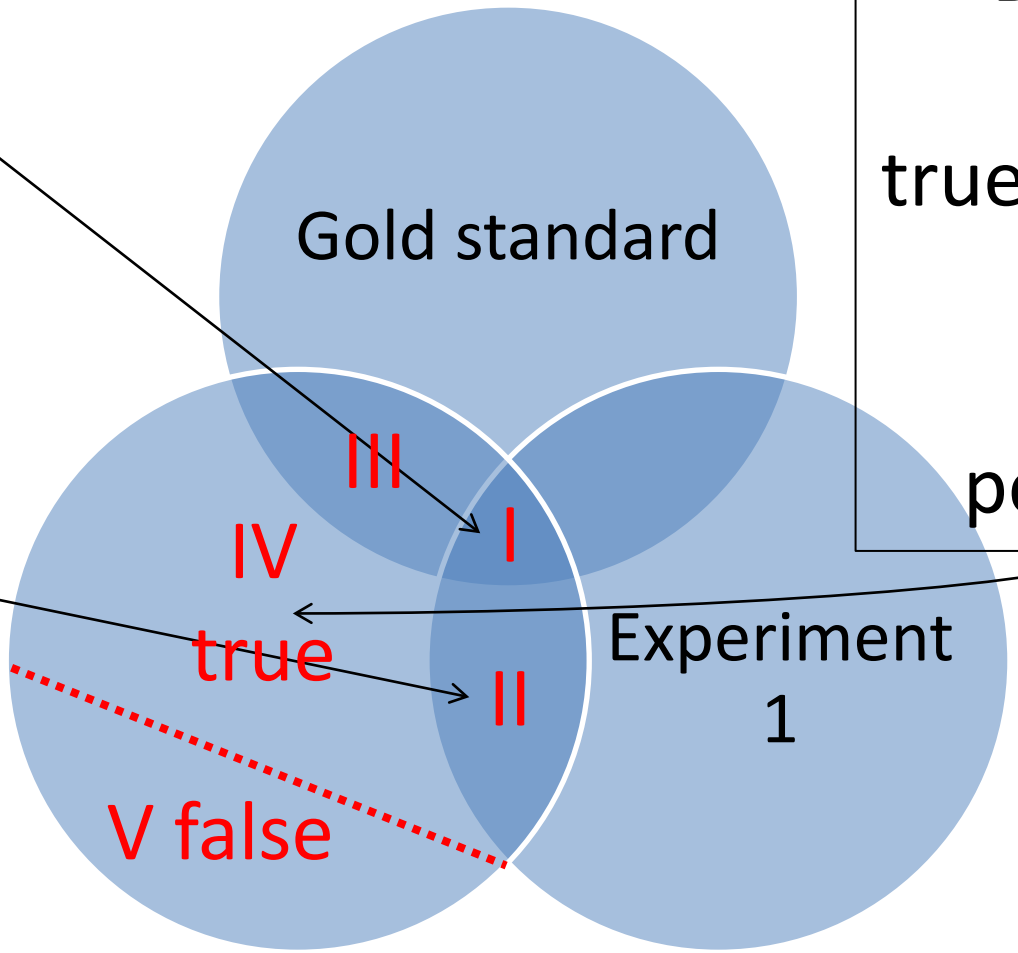
Fraction of consensus present in gold standard=I/II

Data Integration

I =True positives from gold standard

II=Consensus true positives

Define:
IV = true positives
V = false positives



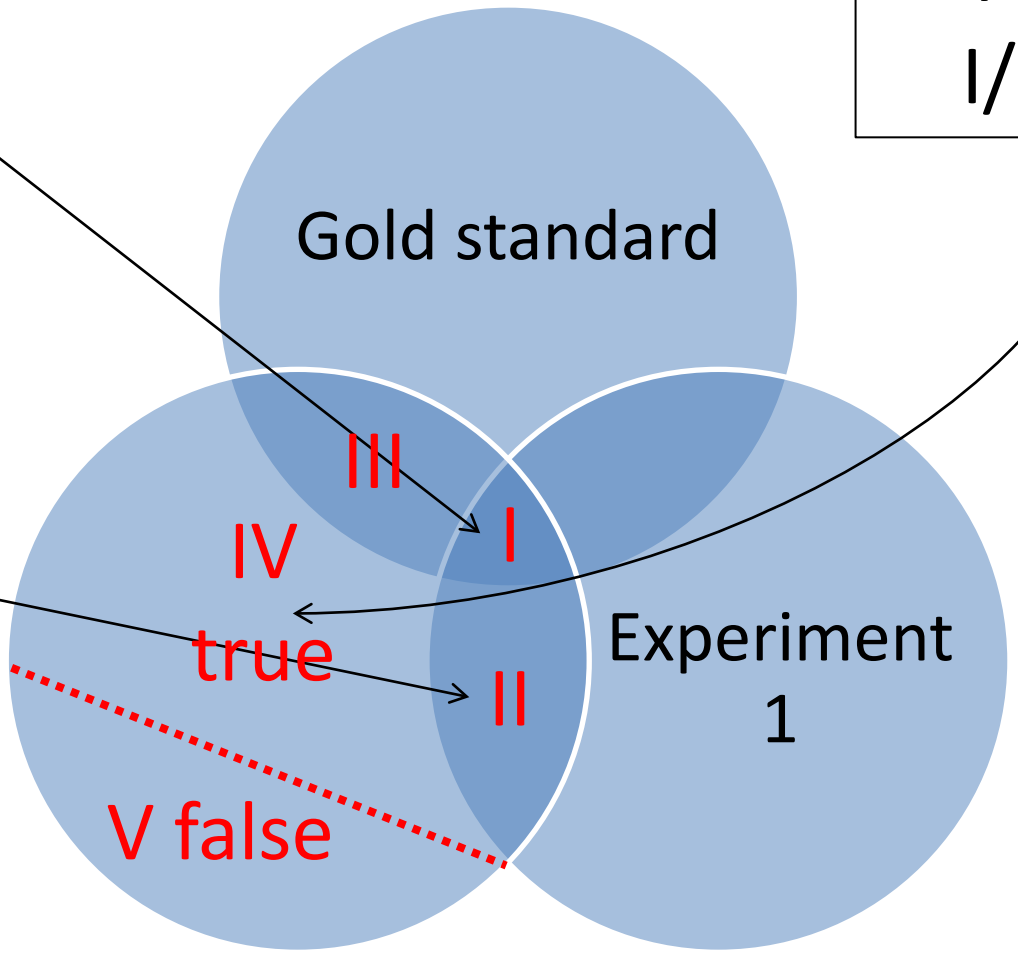
Fraction of consensus present in gold standard=I/II

Data Integration

I =True positives from gold standard

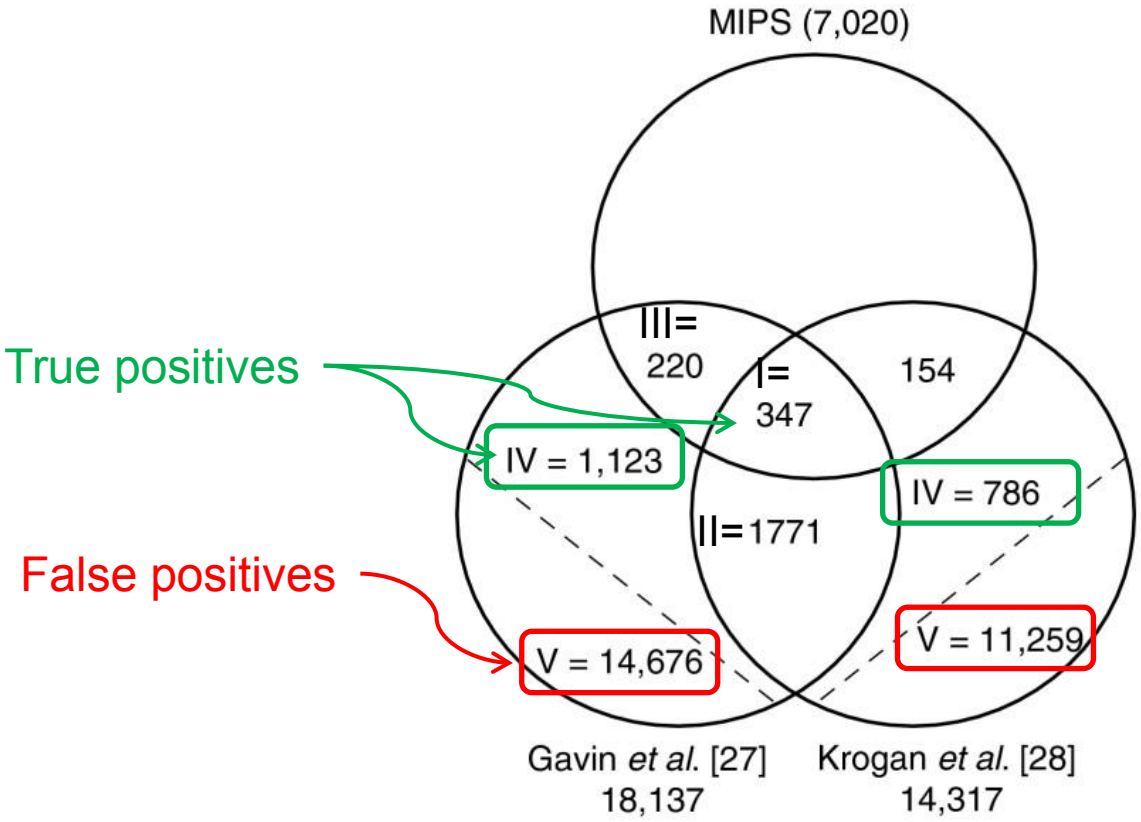
II=Consensus true positives

Assume I/II=III/IV



Fraction of consensus present in gold standard=I/II

Estimated Error Rates



Courtesy of BioMed Central Ltd. Used with permission.
Source: Hart, G. Traver, Arun K. Ramani, et al. "How Complete are Current Yeast and Human Protein-interaction Networks." *Genome Biology* 7, no. 11 (2006): 120.

Assume that all of regions I and II are true positives.

If MIPS has no bias toward either Krogan or Gavin,

then the fraction of TP in MIPS will be the same in the common data (I/II) and the unique data (III/IV)

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

How complete are current yeast and human protein-interaction networks?
G Traver Hart, **Arun K Ramani** and **Edward M Marcotte**
Genome Biology 2006, **7:**120doi:10.1186/gb-2006-7-11-120

Table 1

Yeast protein-interaction assay false-positive rates: yeast datasets

Dataset	Number of interactions	Derived false-positive rate* (%)	Published false-positive rate (%)	Average false-positive rate (%)
Uetz *et al*. [35]	854	46 [32]	32 [24]†,47 [44], 50 [37], 51 [42]	45
Ito [36]	4,393	89 [32]	71 [24]†, 78 [41], 85 [37], 91 [44]	83
Gavin *et al*. [16]	3,180	68 [32]	14 [24]†, 22 [4], <72 (upper bound [20])	35
Ho *et al*. [17]	3,618	83 [32], 81, 82, 80	55 [24]†, <97 (upper bound [20])	76
Jansen *et al*. [22]	15,922	81, 79	-	80
Gavin *et al*. [27]	18,137	78, 82, 86‡	-	82
Krogan *et al*. [28]	14,317 (7,123 core)	75, 79, 66‡ (59, 65, 37 core)	-	73 (54 core)
Overall	51,419			72

*This interaction assay false-positive rate is taken from D'haeseleer and Church [32] or derived using the method therein. Multiple values derive from choosing either the GRID [2] or MIPS [33] reference sets. ‡This interaction assay false-positive rate is calculated with the EPR server of Deane *et al*. [42]. †The mean of four values estimated from Table S3 of Lee *et al*. [24] by fitting the interaction set as a linear combination of true-positive (small scale interactions) and false-positive (random pairs) interactions.

Hart *et al*. *Genome Biology* 2006 **7**:120 doi:10.1186/gb-2006-7-11-120

Table 3

Human protein-interaction assay false-positive rates: human datasets

Dataset	Number of unique interactions	Derived false-positive rates* (%)	Published false-positive rates (%)	Average false-positive rates (%)
Lehner and Fraser [40]	58,700 (9,396 core)	96, 94, 93 (86, 81, 69 core)	-	94 (79 core)
Rhodes *et al*. [23]	38,379	87, 86, 83	-	85
Stelzl *et al*. [15]	3,150 (902 core)	98, 98 (94,95 core)	70 [45]	98 (86 core)
Rual *et al*. [14]	2,611	87, 93	8-66 [14]†, 54 [45]	58
Overall	100,242			90

*This interaction assay false-positive rate is derived using the method of D'haeseleer and Church [32] and a reference set of 20,296 unique interactions from HPRD [54], BIND [55], Reactome [56], and Ramani *et al*. [49]. Multiple values derive from different choices of comparison sets. †A range of six values (mean 48%) estimated from Table 1 of Rual *et al*. [14] by fitting the interaction set CCSB-HI1 as a linear combination of true positive (LCI core) and false positive (all possible) interactions.

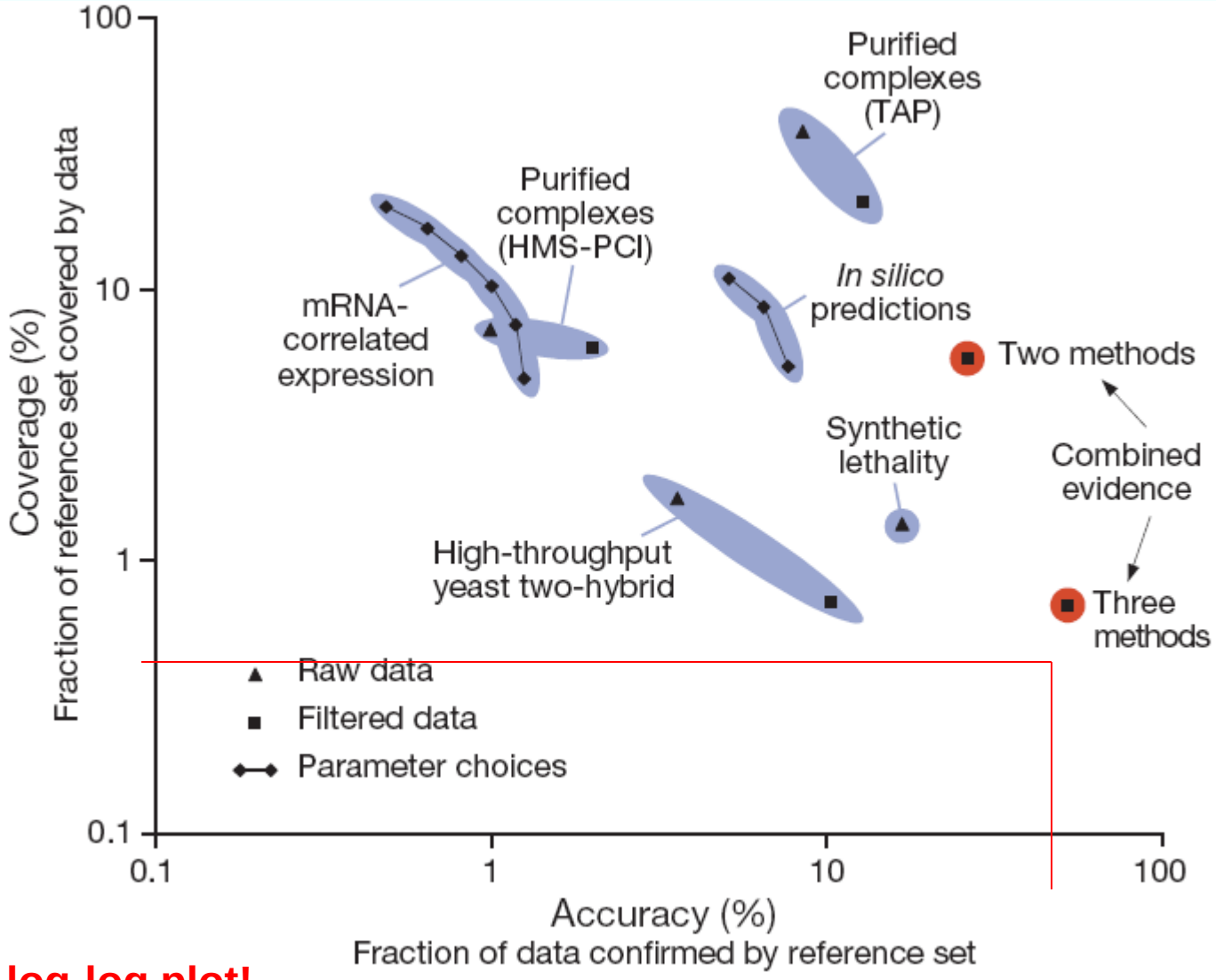
Hart *et al*. *Genome Biology* 2006 **7**:120 doi:10.1186/gb-2006-7-11-120

Courtesy of BioMed Central Ltd. Used with permission.

Source: Hart, G. Traver, Arun K. Ramani, et al. "How Complete are Current Yeast and Human Protein interaction Networks." *Genome Biology* 7, no. 11 (2006): 120.

Finding real interactions

- Take only those that are reported by >1 method?
- Filter out “sticky” proteins?
- Estimate probability of each interaction based on data.
- Use external data to predict



Note: log-log plot!

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Von Mering, Christian, Roland Krause, et al. "Comparative Assessment of Large-scale Data Sets of Protein–protein Interactions." *Nature* 417, no. 6887 (2002): 399-403.

Comparative assessment of large-scale data sets of protein–protein interactions
von Mering, *et al. Nature* **417**, 399-403 (23 May 2002) | doi:10.1038/nature750

Finding real interactions

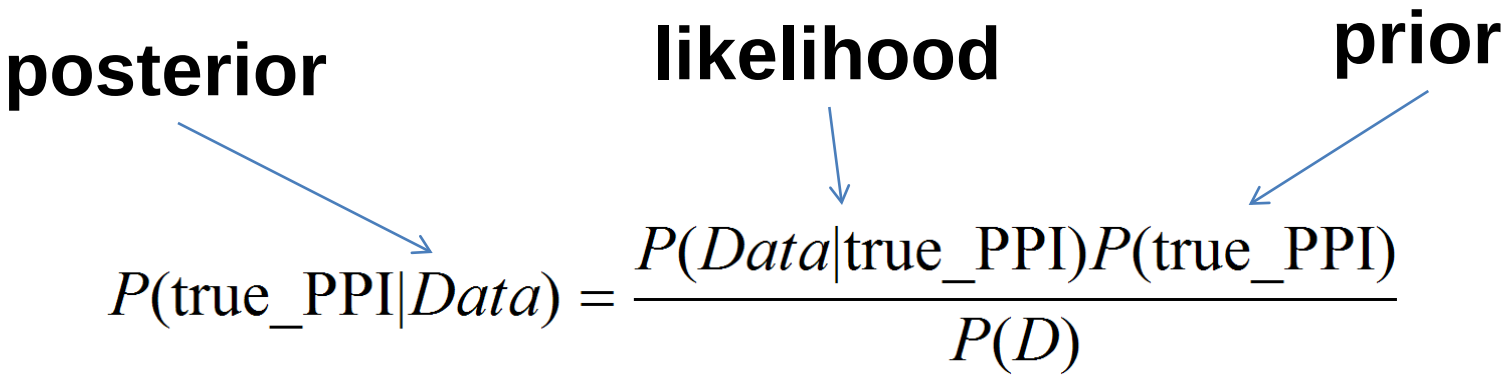
- Estimate probability of each interaction based on data.
  - How can we compute $P(real\_PPI \mid Data)$?

$$Data = (two\_hybrid, mass\_spec, co\_evolution, co\_expression, \ldots)$$

or

$$Data = (mass\_spec\_expt\_1, mass\_spec\_expt\_2, \ldots)$$

Bayes Rule


posterior
likelihood
prior
$P(\text{true\_PPI}|Data) = \frac{P(Data|\text{true\_PPI})P(\text{true\_PPI})}{P(D)}$

Naïve Bayes Classification

posterior **likelihood** **prior**

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

likelihood ratio = ratio of posterior probabilities

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

if > 1 classify as true

if < 1 classify as false

How do we compute this ?

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 =

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

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

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 =

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

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

Ranking function = $\log\left[\frac{P(Data\,|\,true\_PPI)}{P(Data\,|\,false\_PPI)}\right]$

Ranking function =

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

We assume the observations are independent
(we'll see how to handle dependence soon)

Ranking function =

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

We assume the observations are independent
(we'll see how to handle dependence soon)

Ranking function =

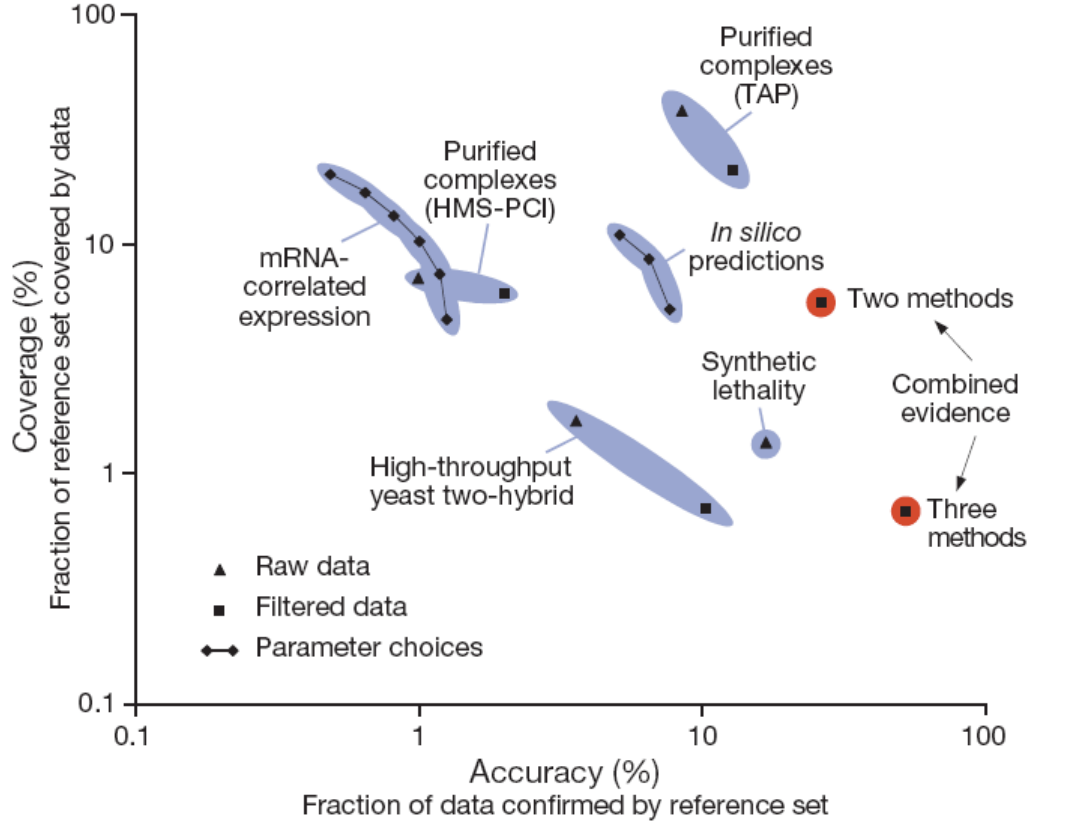
$$\log\left[\frac{P(Data \mid true\_PPI)}{P(Data \mid false\_PPI)}\right] = \prod_{i}^{M} \frac{P(Observation_i \mid true\_PPI)}{P(Observation_i \mid false\_PPI)}$$

We assume the observations are independent
(we'll see how to handle dependence soon)

We can compute these terms if we have a set of high-confidence positive and negative interactions .

Exactly how we compute the terms depends on the type of data.

For affinity purification/mass spec. see
Collins et al. Mol. Cell. Proteomics 2007
http://www.mcponline.org/content/6/3/439.long



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Von Mering, Christian, Roland Krause, et al. "Comparative Assessment of Large-scale Data Sets of Protein–protein Interactions." *Nature* 417, no. 6887 (2002): 399-403.

Instead of requiring an interaction to be detected in all assays, we can rank by

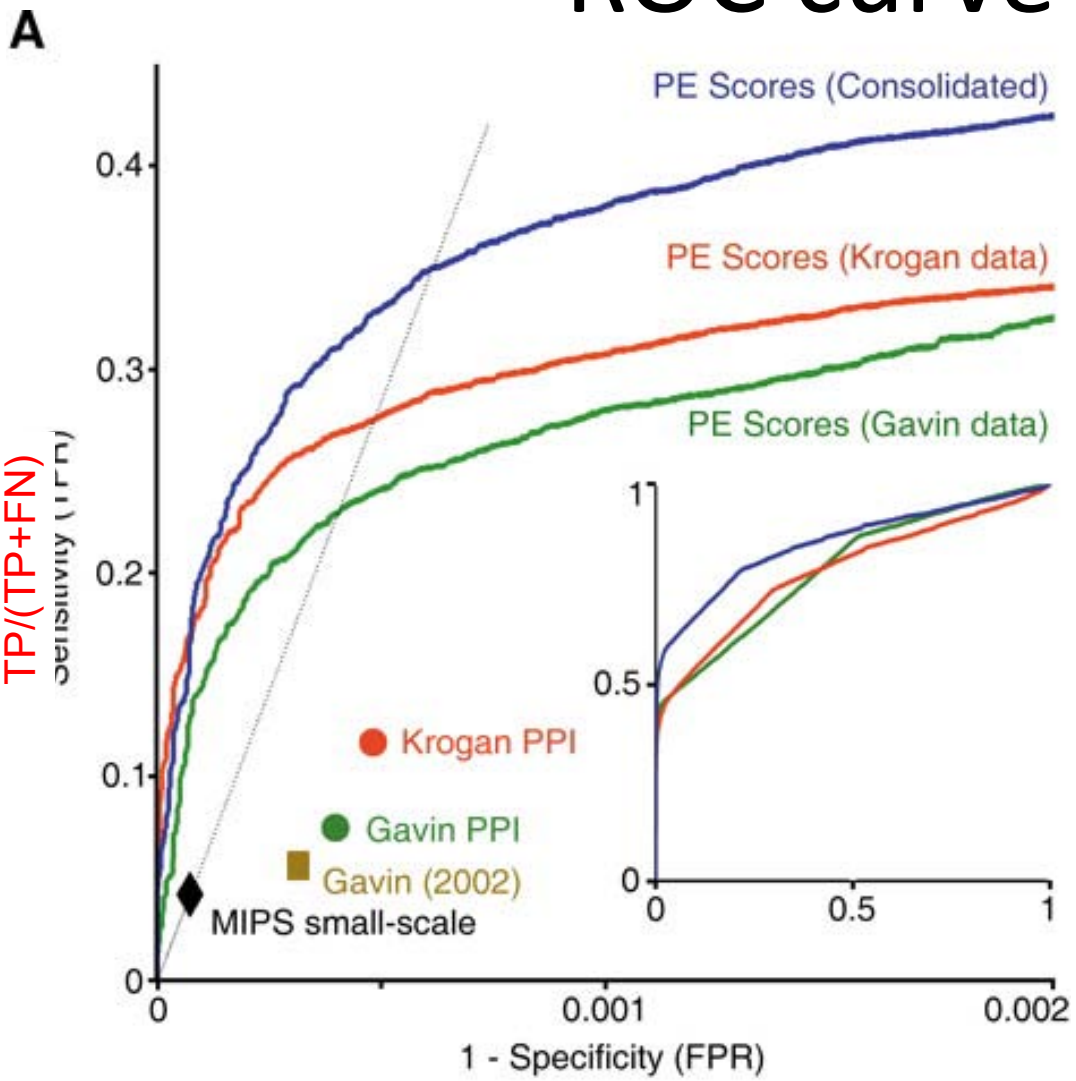
$$P(\text{true\_PPI}|Data)$$

Comparative assessment of large-scale data sets of protein–protein interactions
von Mering, *et al. Nature* **417**, 399-403 (23 May 2002) | doi:10.1038/nature750

ROC curve

Compute using high-confidence positives (from complexes)

TP/(TP+FN)



© American Society for Biochemistry and Molecular Biology. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Collins, Sean R., Patrick Kemmeren, et al. "Toward a Comprehensive Atlas of the Physical Interactome of Saccharomyces Cerevisiae." *Molecular & Cellular Proteomics* 6, no. 3 (2007): 439-50.

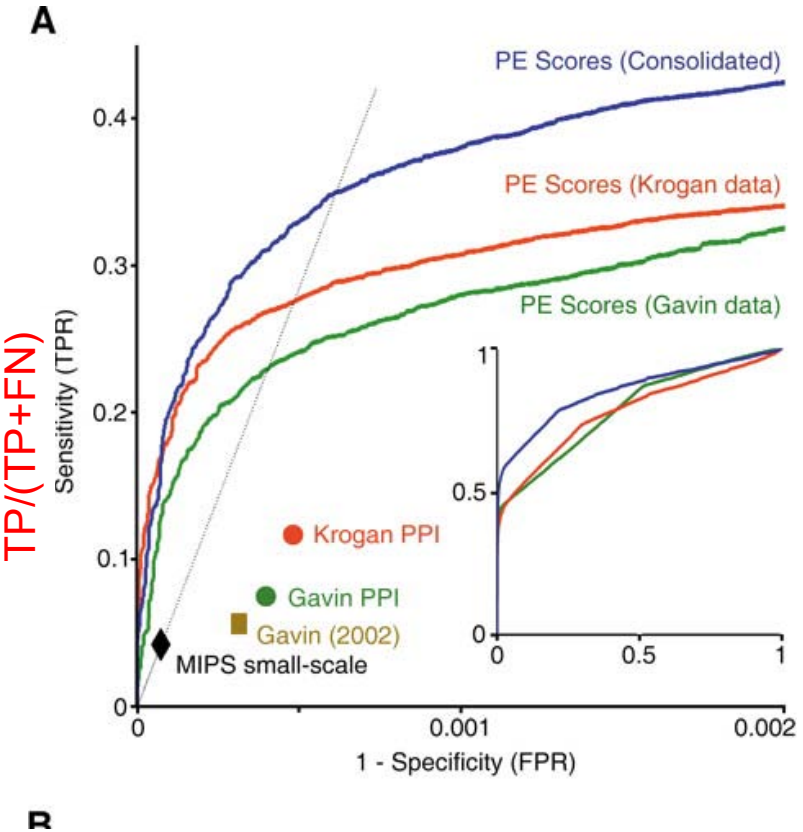
True positives = ?

True negatives = ?

Compute using high-confidence negatives FP/(TN+FP)

Collins et al. Mol. Cell. Proteomics 2007

ROC curve



© American Society for Biochemistry and Molecular Biology. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Collins, Sean R., Patrick Kemmeren, et al. "Toward a Comprehensive Atlas of the Physical Interactome of Saccharomyces Cerevisiae." *Molecular & Cellular Proteomics* 6, no. 3 (2007): 439-50.

Compute using high-confidence positives (from complexes)
TP/(TP+FN)

Compute using high-confidence negatives
FP/(TN+FP)

ROC curve

True positives = interactions between proteins that occur in a complex annotated in a human-curated database (MIPS or SGD).

True negatives = proteins pairs that

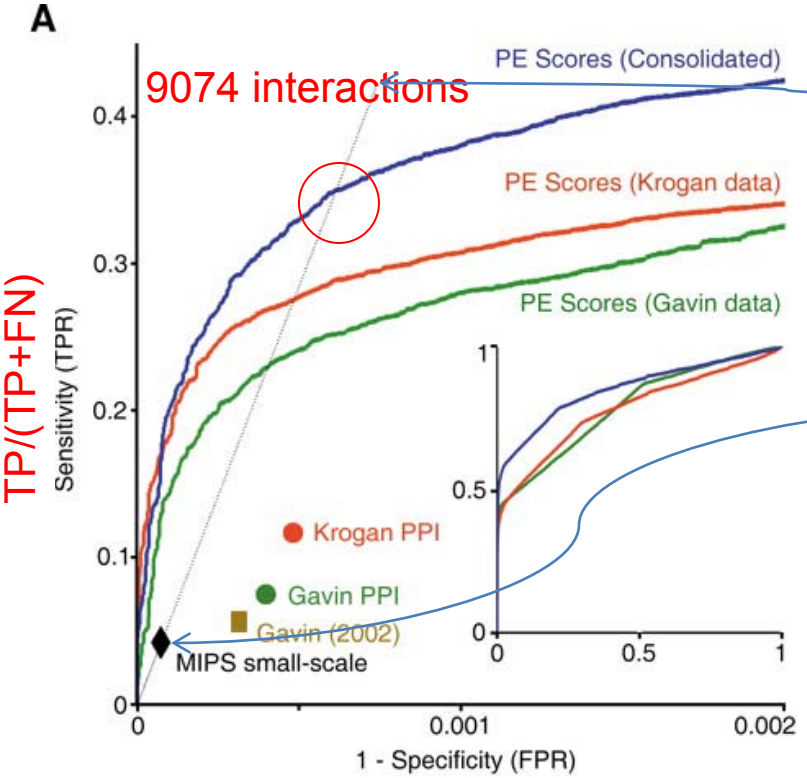
1. are annotated to belong to distinct complexes
2. have different sub-cellular locations OR anticorrelated mRNA expression

Collins et al. Mol. Cell. Proteomics 2007

ROC curve

Compute using high-confidence positives (from complexes)

TP/(TP+FN)



Error rate equivalent to literature curated

Literature curated (excluding 2-hybrid)

© American Society for Biochemistry and Molecular Biology. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Collins, Sean R., Patrick Kemmeren, et al. "Toward a Comprehensive Atlas of the Physical Interactome of Saccharomyces Cerevisiae." *Molecular & Cellular Proteomics* 6, no. 3 (2007): 439-50.

Compute using high-confidence negatives

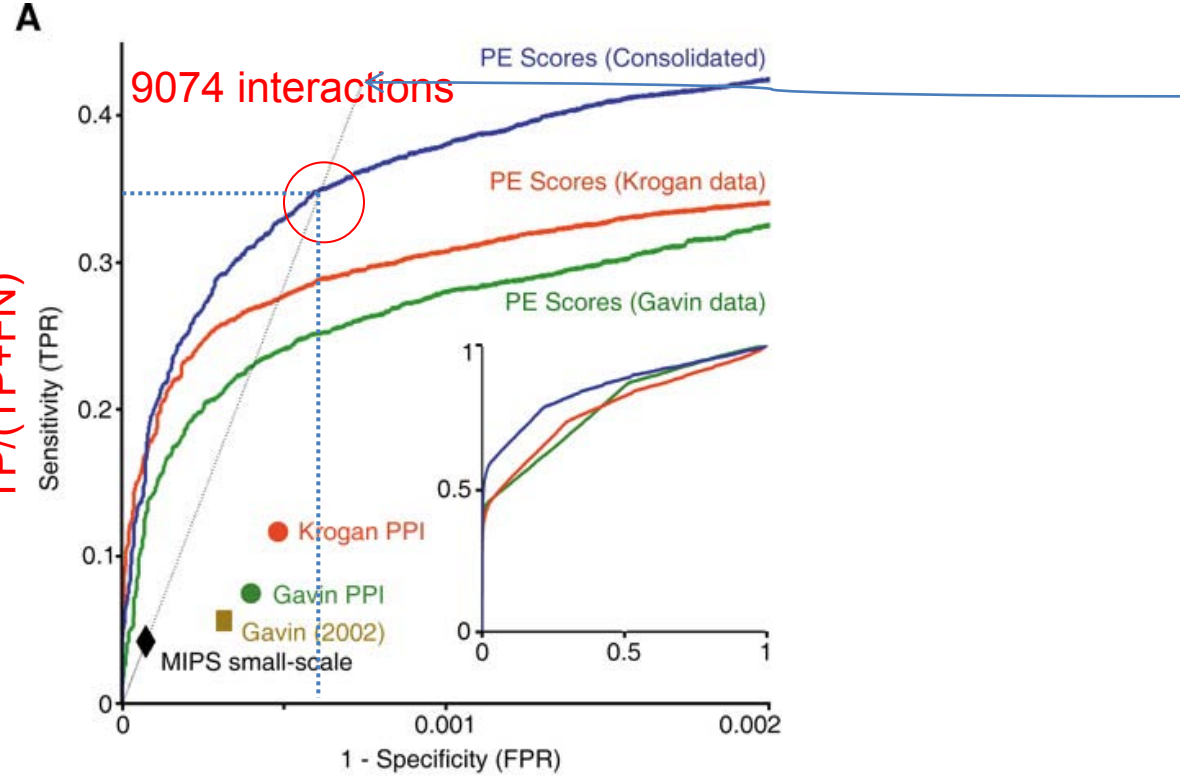
FP/(TN+FP)

ROC curve

Collins et al. Mol. Cell. Proteomics 2007

ROC curve

Compute using high-confidence positives (from complexes)
TP/(TP+FN)



© American Society for Biochemistry and Molecular Biology. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Collins, Sean R., Patrick Kemmeren, et al. "Toward a Comprehensive Atlas of the Physical Interactome of Saccharomyces Cerevisiae." *Molecular & Cellular Proteomics* 6, no. 3 (2007): 439-50.

Compute using high-confidence negatives

FP/(TN+FP)

ROC curve

Collins et al. Mol. Cell. Proteomics 2007

Outline

- Structural prediction of protein-protein interactions
- High-throughput measurement of protein-protein interactions
- Estimating interaction probabilities
- Bayes Net predictions of protein-protein interactions

Bayesian Networks

A method for using probabilities to reason

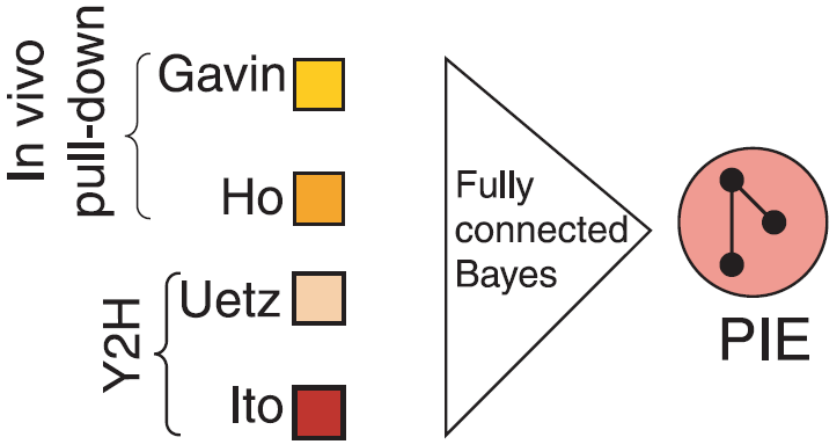
In Biology

- Gene regulation
- Signaling
- Prediction

Bayesian Networks

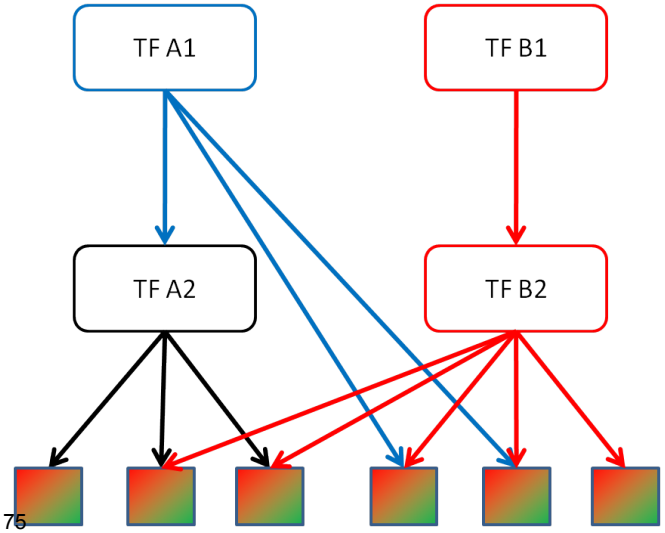
- Bayesian Networks are a tool for reasoning with probabilities
- Consist of a graph (network) and a set of probabilities
- These can be “learned” from the data

Bayesian Networks



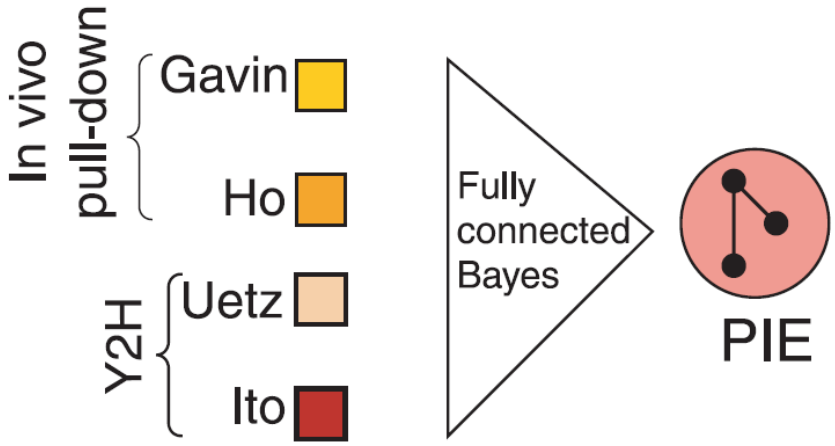
© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Jansen, Ronald, Haiyuan Yu, et al. "A Bayesian Networks Approach for Predicting Protein-protein 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



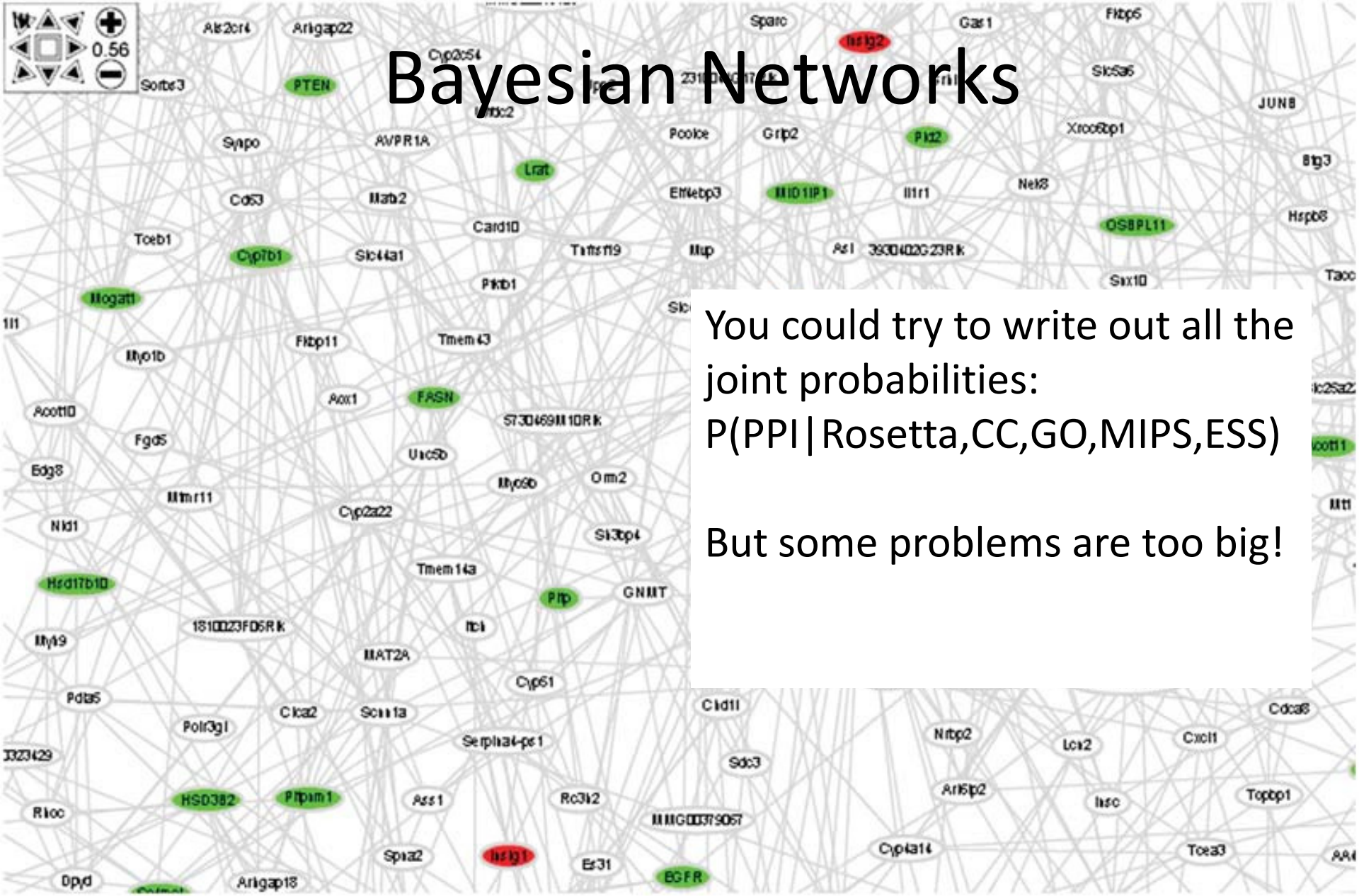
You could try to write out all the joint probabilities:

$P(PPI|Y2H_{Uetz}, Y2H_{Ito}, IP_{Gavin}, IP_{Ho})$

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

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.

Bayesian Networks



You could try to write out all the joint probabilities:
P(PPI|Rosetta,CC,GO,MIPS,ESS)

But some problems are too big!

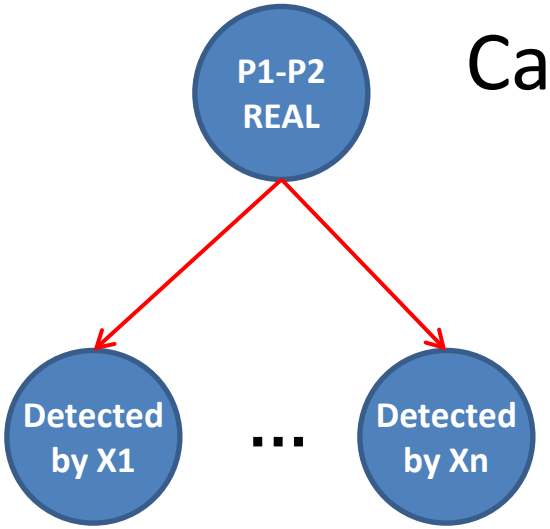
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Yang, Xia, Joshua L. Deignan, et al. "Validation of Candidate Causal Genes for Obesity that Affect Shared Metabolic Pathways and Networks." *Nature genetics* 41, no. 4 (2009): 415-23.

Bayesian Networks

- Complete joint probability tables are large and often unknown
- N binary variables = 2^N states
  - only one constraint (sum of all probabilities =1)

  => 2^N - 1 parameters

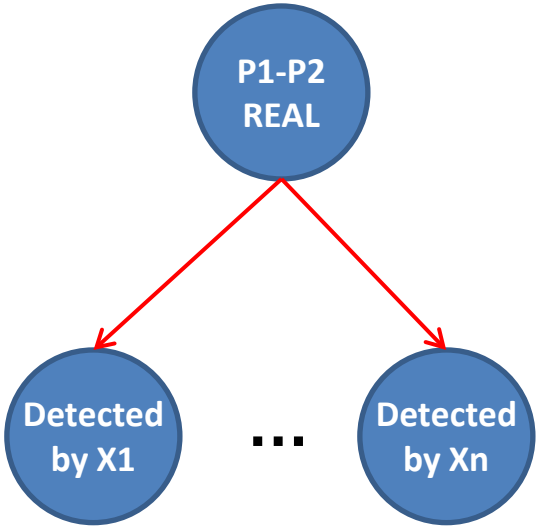
Graphical Structure Expresses our Beliefs

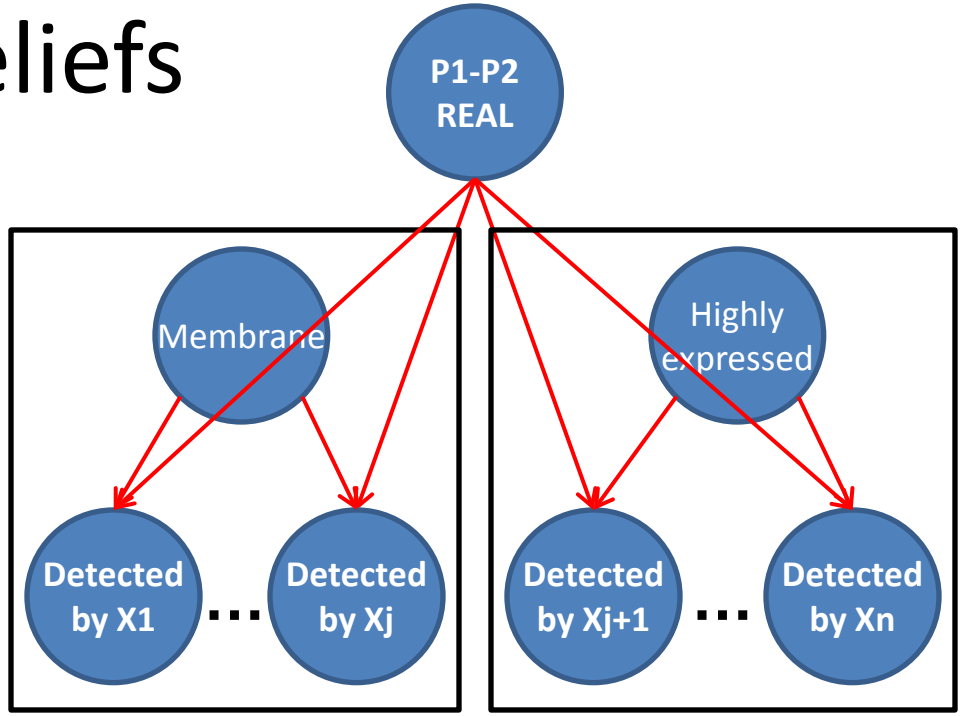


Cause (often “hidden”)

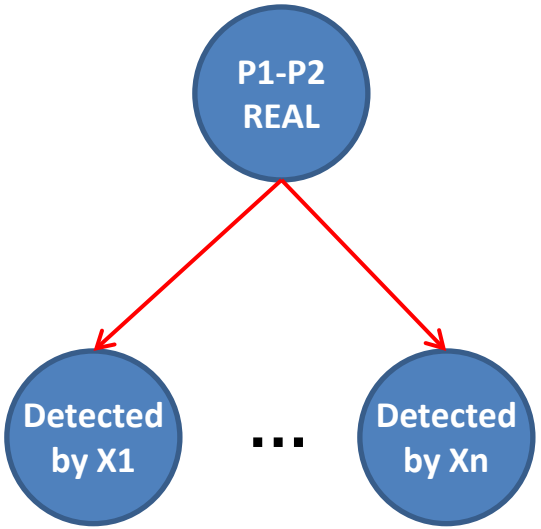
Effects (observed)

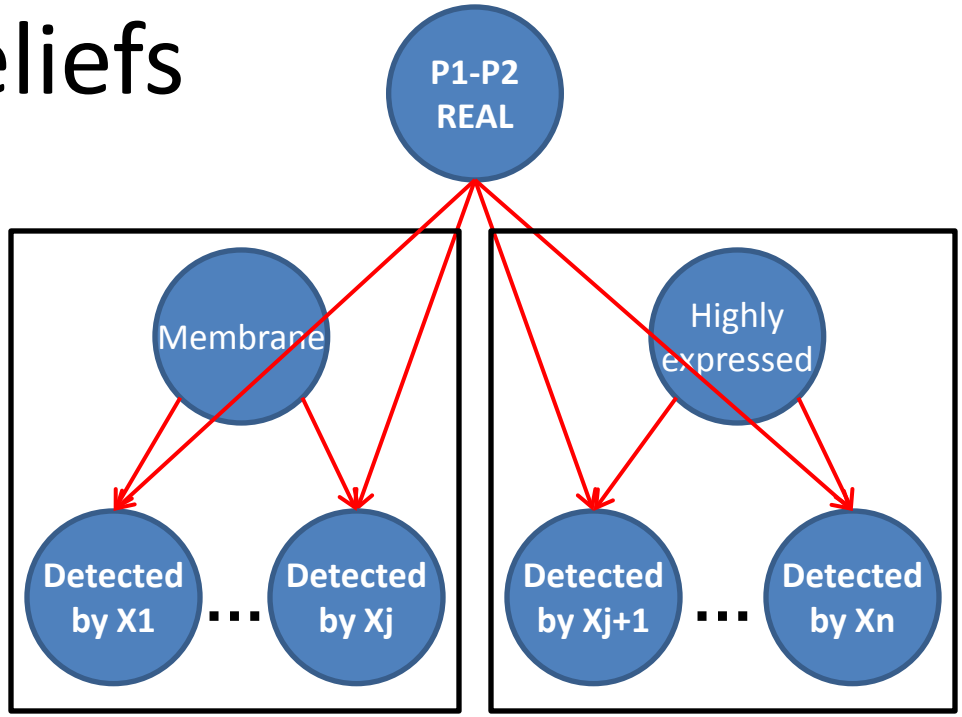
Graphical Structure Expresses our Beliefs





Graphical Structure Expresses our Beliefs





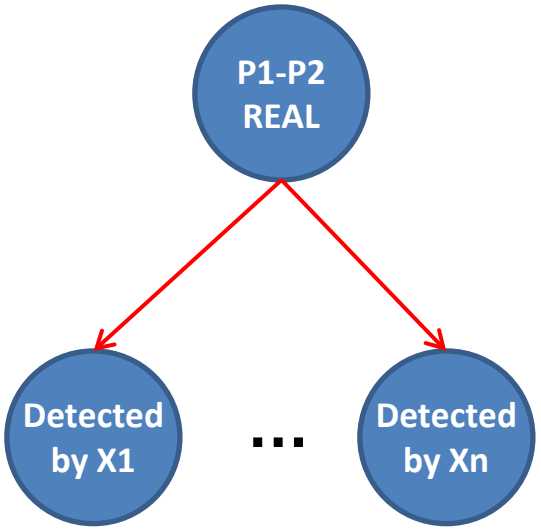
Naïve Bayes assumes all observations are independent

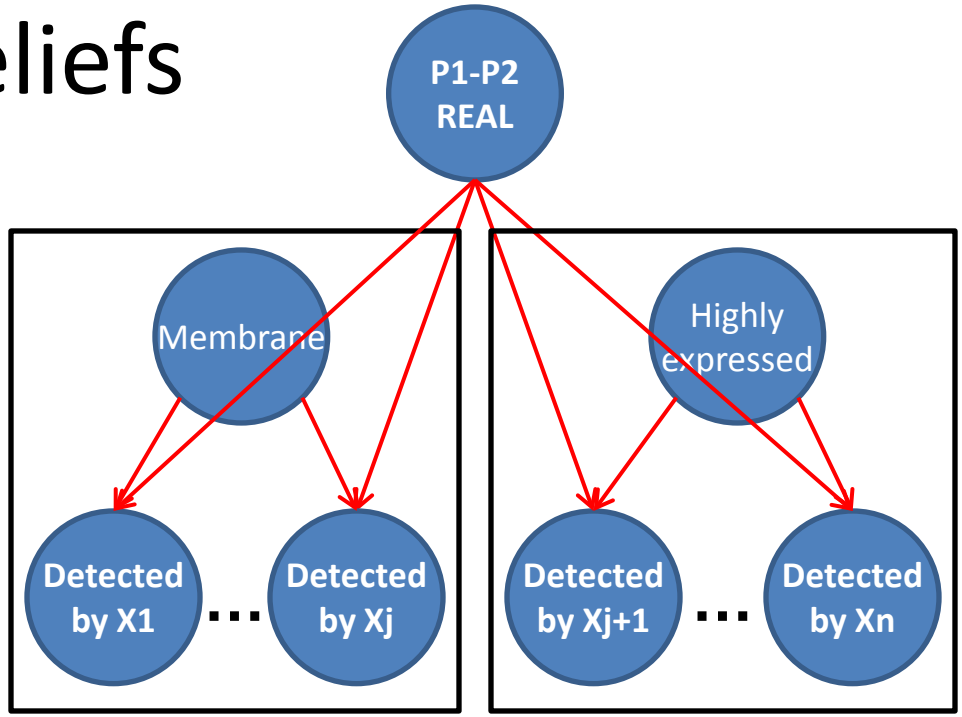
$$P(X_1 \ldots X_n|\text{PPI}) = \prod_i [P(Xi|\text{PPI})]$$

But some observations may be coupled.

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

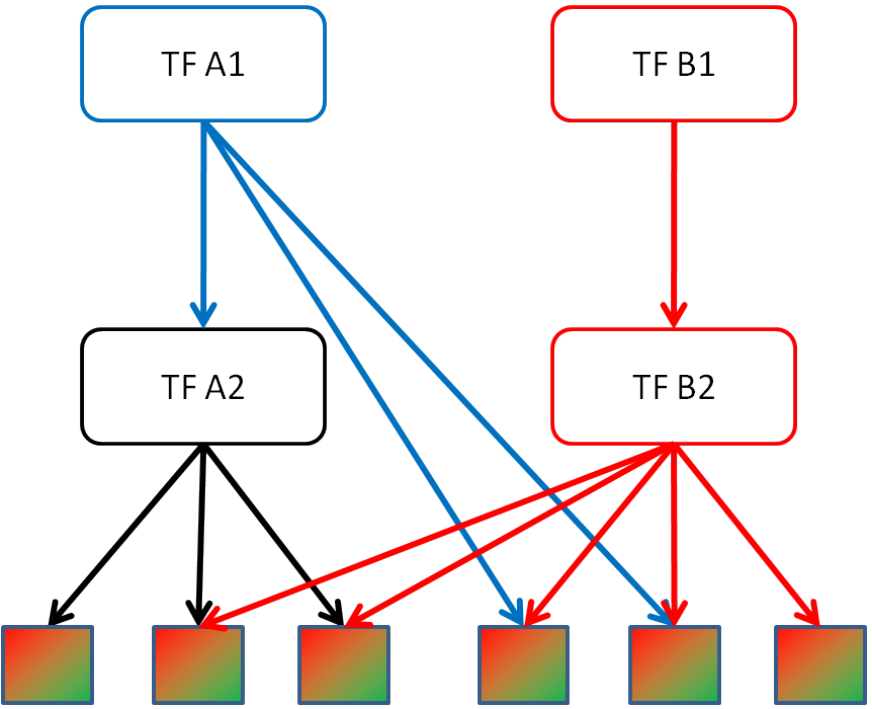
Graphical Structure Expresses our Beliefs





- The graphical structure can be decided in advance based on knowledge of the system or learned from the data.

Graphical Structure



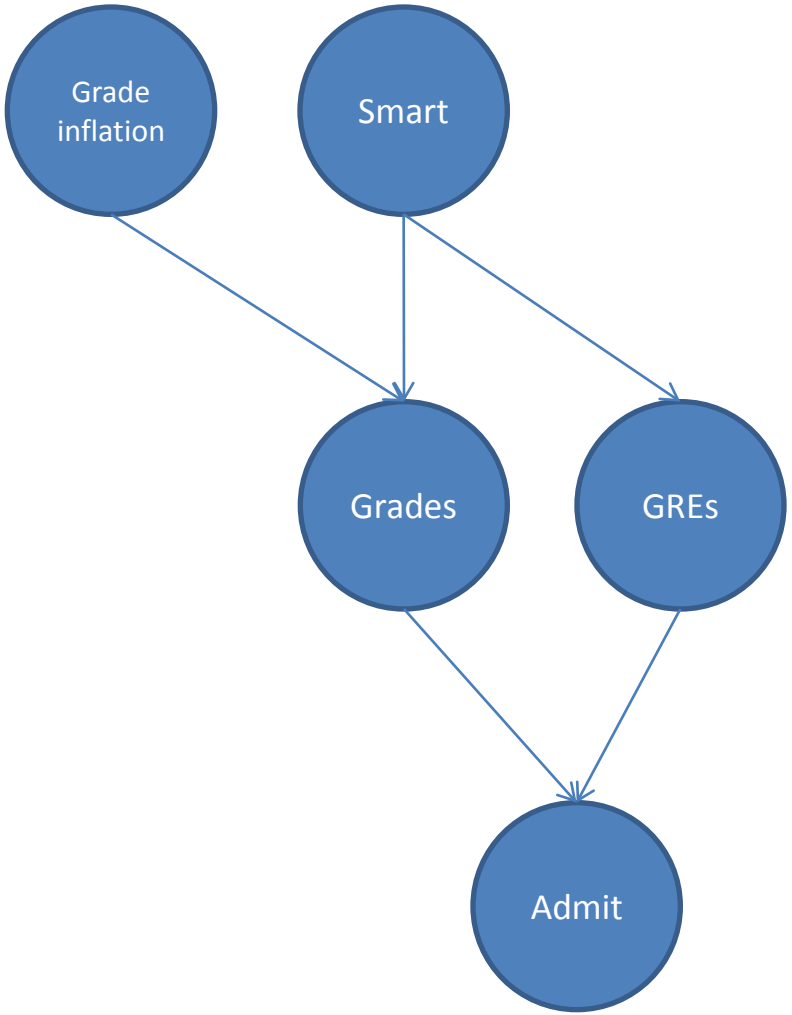
In a Bayesian Network, we don't need the full probability distribution.

A node is independent of its ancestors given its parents.

For example:

The activity of a gene does not depend on the activity of TF A1 once I know TF B2.

Automated Admissions Decisions



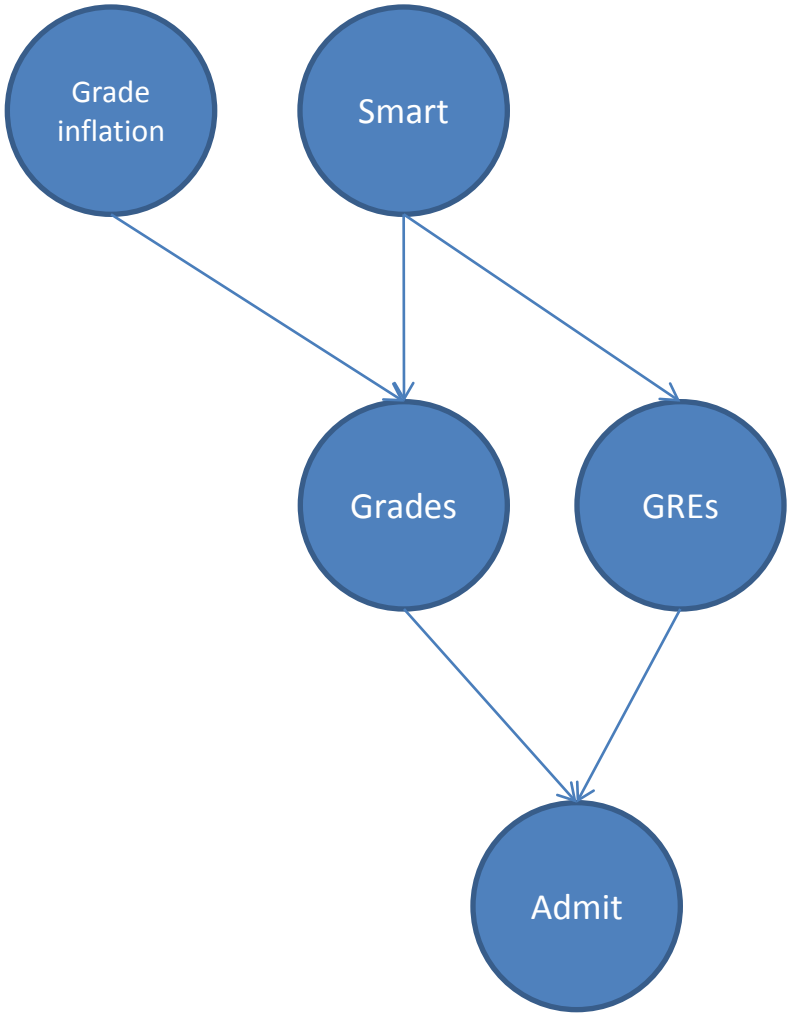
Prediction: we observe the "causes" (roots/ parents) and want to predict the "effects" (leaves/children).

Given Grades, GREs will we admit?

Inference: we observe the "effects" (leaves/ children) and want to infer the hidden values of the "causes" (roots/parents)

You meet an admitted the student. Is s/he smart?

Automated Admissions Decisions

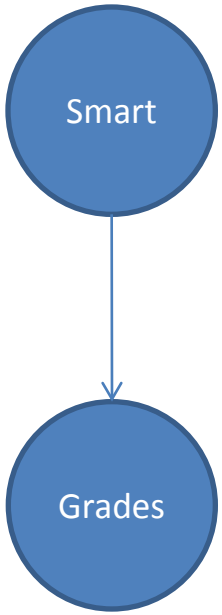


Making predictions/inferences requires knowing the joint probabilities:

P(admit, grade inflation, smart, grades, GREs)

We will find conditional probabilities to be very useful

Automated Admissions Decisions



Joint Probability

S	G (above threshold)	P(S,G)
F	F	0.665
F	T	0.035
T	F	0.06
T	T	0.24

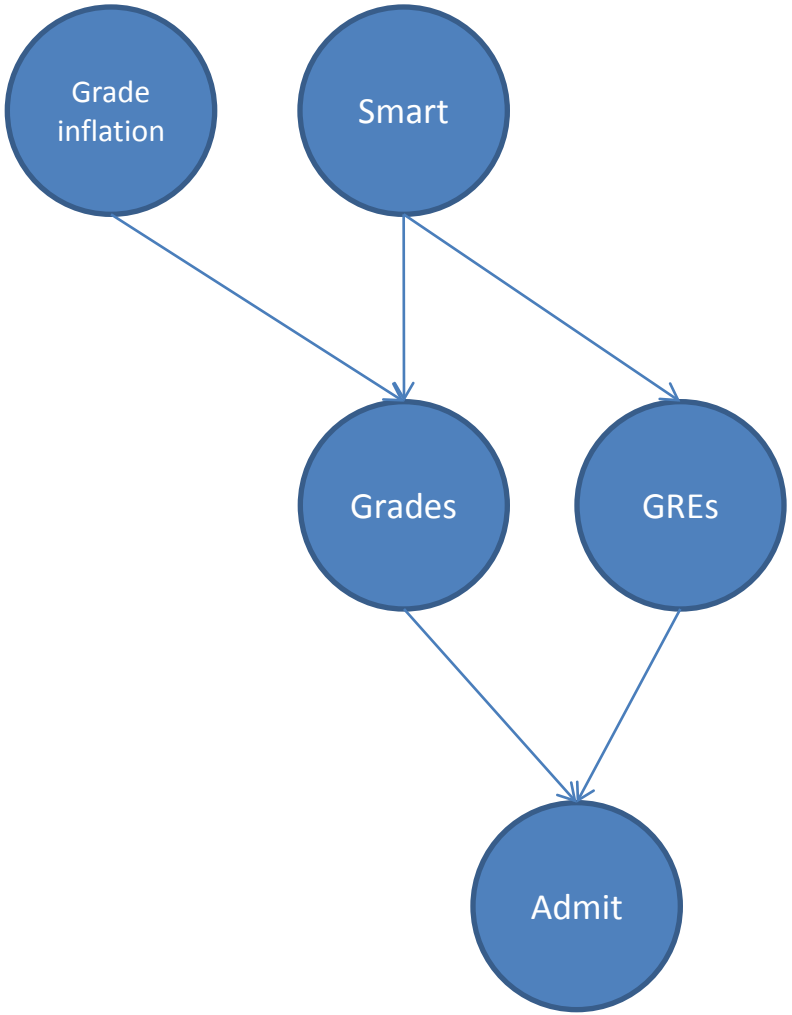
Conditional Probability

S	P(G=F|S)	P(G=T|S)
F	0.95	0.05
T	0.2	0.8

P(S=F)=0.7	P(S=T)=0.3

Formulations are equivalent and both require same number of constraints.

Automated Admissions Decisions



In a Bayesian Network, we don't need the full probability distribution.

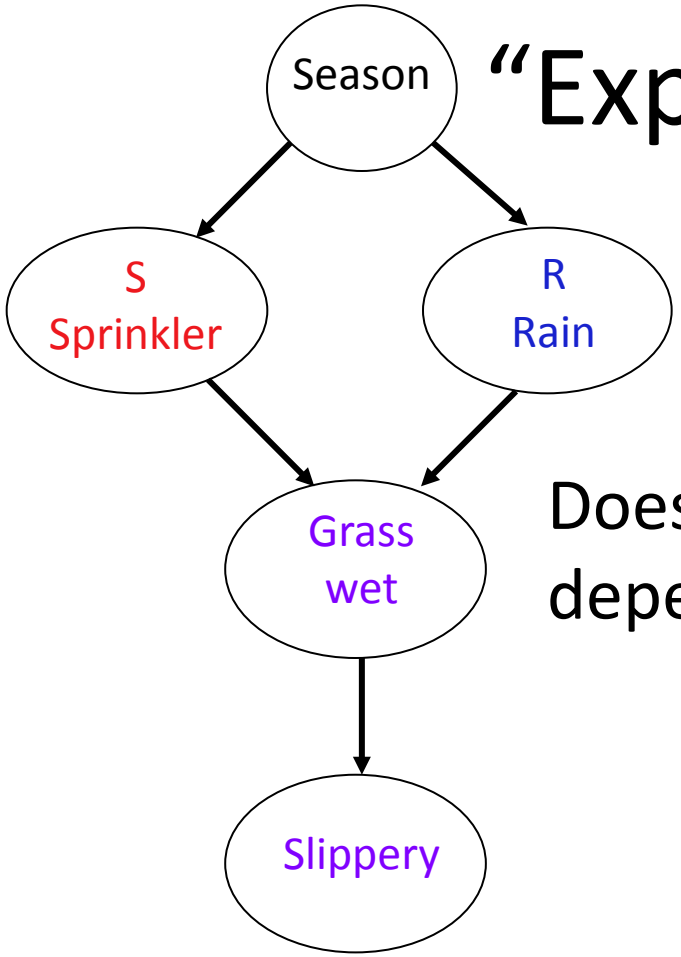
The joint probability depends only on "parents."

For example:

GRE scores do not depend on the level of grade inflation at the school, but the grades do.

$$P(X_1 \ldots X_n) = \prod_i [P(Xi|Parents(X_i))]$$

“Explaining Away”



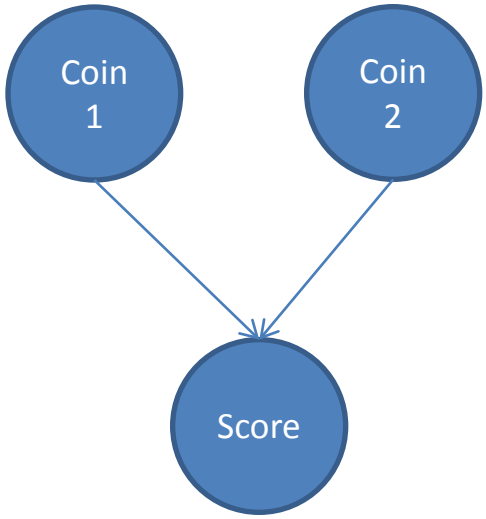
Does the probability that it’s raining depend on whether the sprinkler is on?

In a causal sense, clearly not.

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

"Explaining Away"

$p(C_1=H)=p(C_1=T) =. 5$



$p(C_2=H)=p(C_2=T) =. 5$

C_1	C_2	Score
H	H	1
T	T	1
H	T	0
T	H	0

Does the prob. that C_1=H on depend on whether C_2 =T?

If we know the score, then our belief in the state of C1 is influenced by our belief in the state C2.

$$p(C_2 = H|S = 1, C_1 = T) = \frac{p(C_2 = H, S = 1, C_1 = T)}{p(S = 1, C_1 = T)} = 0$$

How do we obtain a BN?

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

Learning Models from Data

- Assume we know the structure, how do we find the parameters?
- Define an objective function and search for parameters that optimize this function.

Learning Models from Data

- Two common objective functions
  - Maximum likelihood:
    - Define the likelihood over training data $\{X_i\}$:

$$L(\theta) = P(Data|\theta) = \sum_{i}^{N} P(Xi|\theta)$$

$$\theta_{ML} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} P(Data|\theta)$$

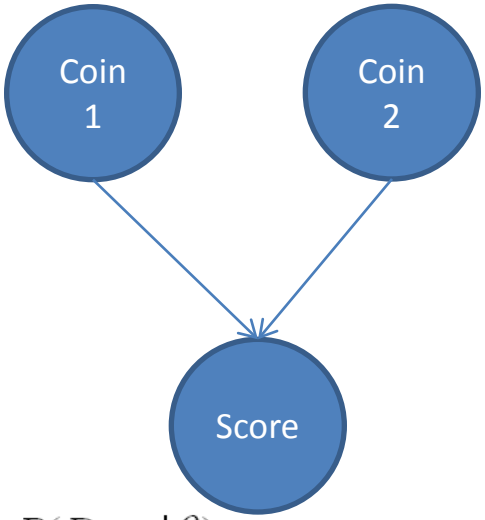
  - Maximum posterior:

$$\theta_{MAP} = \arg\max_{\theta} P(\theta|Data) = \arg\max_{\theta} \frac{P(Data|\theta)P(\theta)}{P(D)}$$

- Good search algorithms exist:
  - Gradient descent, EM, Gibbs Sampling, …

Learning Models from Data

$p(C_1=H)=?$
$p(C_1=T)=?$



$p(C_2=H)=?$
$p(C_2=T)=?$

C_1	C_2	Score
H	H	?
T	T	?
H	T	?
T	H	?

$$\theta_{ML} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} P(Data|\theta)$$

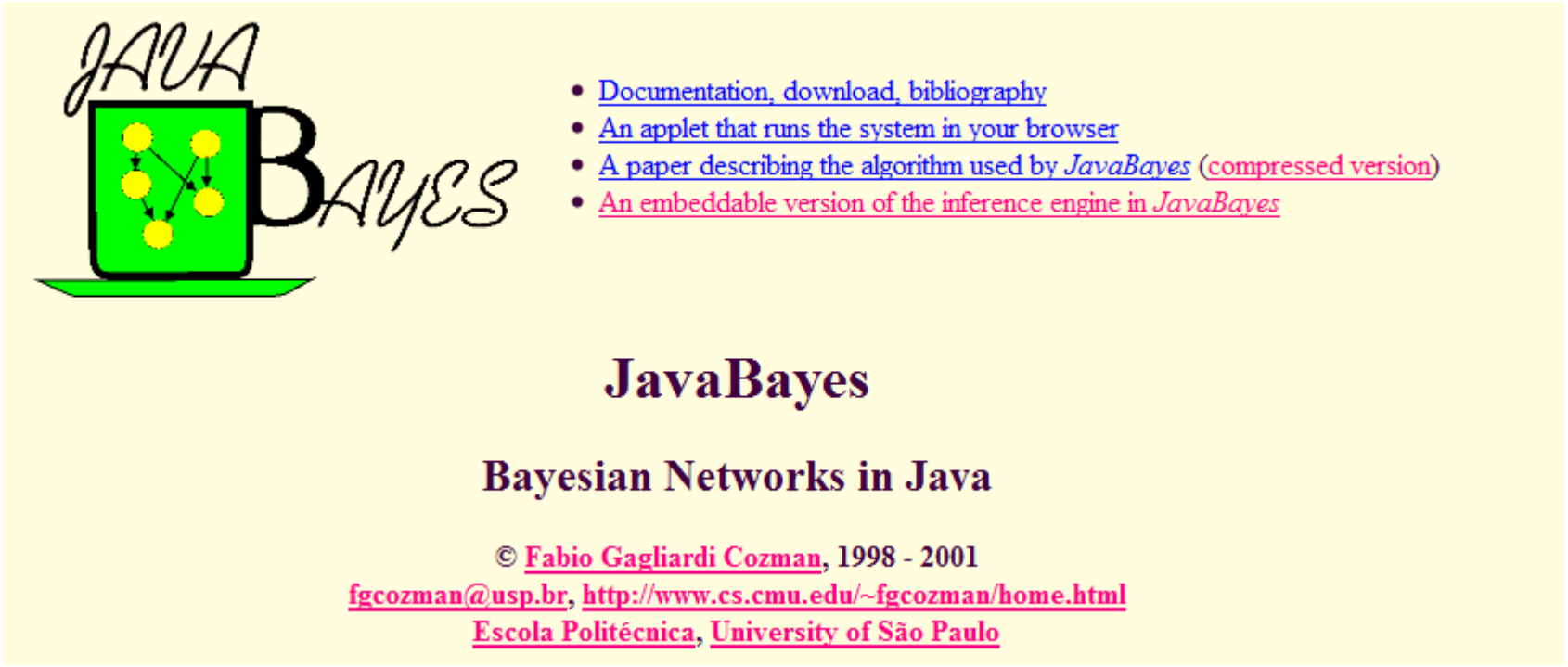
$$\theta_{MAP} = \arg\max_{\theta} P(\theta|Data) = \arg\max_{\theta} \frac{P(Data|\theta)P(\theta)}{P(D)}$$

D= {(C1,C2,S)} = {(H,T,0), (H,H,1) …}

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 measurements are likely independent?
    - Which nodes should regulate transcription?
    - Which should cause changes in phosphorylation?

Resources to learn more



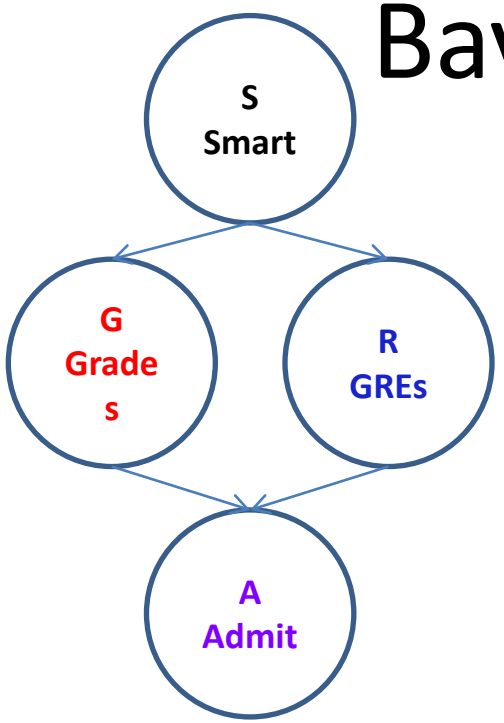
© Fabio Gagliardi Cozman. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

Kevin Murphy's tutorial: http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html

A worked “toy” example

Best to work through on your own

Chain Rule of Probability for Bayes Nets



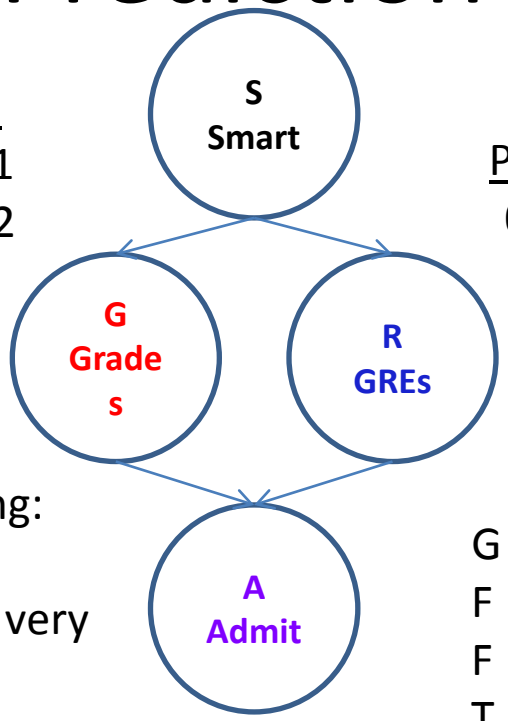
Recall:
$P(X,Y) = P(X|Y)P(Y)$

$$P(S,G,R,D) = P(S)P(G|S)P(R|G,S)P(A|S,G,R)$$

$$= P(S)P(G|S)P(R|S)P(A|G,R) \quad \text{(why?)}$$

(because of conditional independence assumption)

Prediction with Bayes Nets



G		
S	P(F)	P(T)
F	0.9	0.1
T	0.8	0.2

S	
P(F)	P(T)
0.5	0.5

R		
S	P(F)	P(T)
F	0.8	0.2
T	0.2	0.8

A			
G	R	P(F)	P(T)
F	F	0.9	0.1
F	T	0.8	0.2
T	F	0.5	0.5
T	T	0.2	0.8

Rough grading: being smart doesn't help very much!

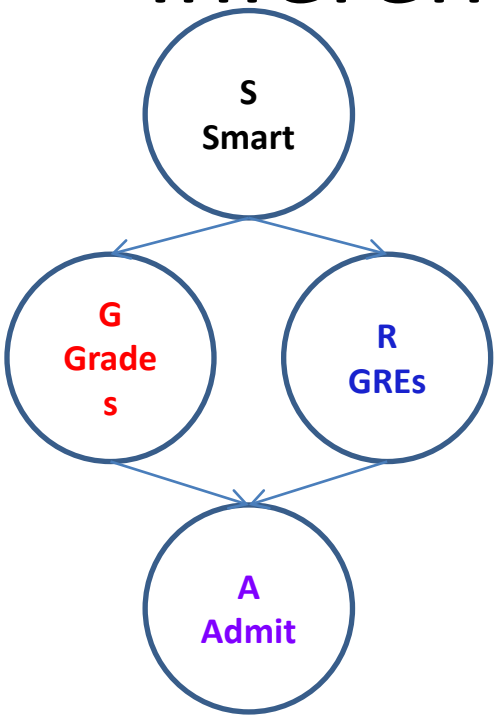
GREs are better correlated with intelligence than the grades

$$P(A=T|S=T) = \sum_{G=F,T}\sum_{R=F,T} P(G|S=T)P(R|S=T)P(A=T|G,R)$$

$$= (0.8)(0.2)(0.1) + (0.8)(0.8)(0.2) + (0.2)(0.2)(0.5) + (0.2)(0.8)(0.8) = .29$$

(F F) (F T) (T F) (T T)

Inference with Bayes Nets



$P(S=T|A=T) = P(S=T,A=T)/P(A=T)$

Or, using Bayes Rule:

$= P(S=T)P(A=T|S=T)/P(A=T)$

$P(S=T) = 0.5$

$P(A=T|S=T)$ calculated on previous slide $=.29$

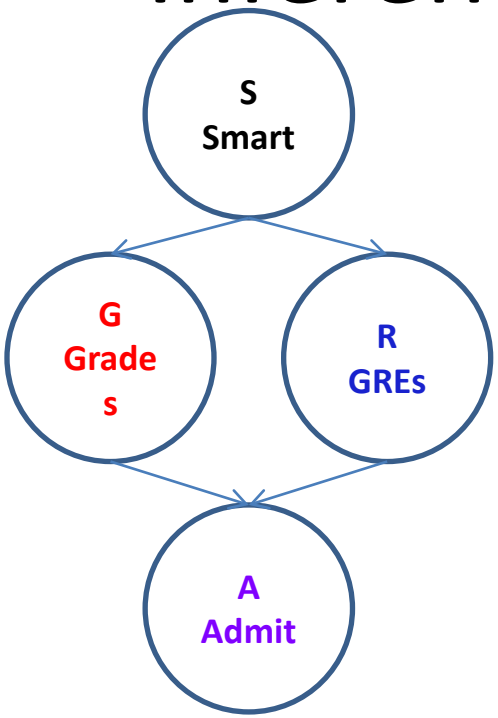
$P(A=T|S=F) = .14$

$$P(A=T) = \sum_{S=F,T} \sum_{G=F,T} \sum_{R=F,T} P(S)P(G|S)P(R|S)P(A=T|G,R)$$

$= P(S=T)P(A=T|S=T) + P(S=F)P(A=T|S=F) = 0.21$

$P(S=T) = 0.5$, $P(S=F) = 0.5$, $P(A=T|S=F)$ calculated analogously to $P(A=T|S=T)$

Inference with Bayes Nets



If a student is not admitted, is it more likely they had bad GREs or bad grades?

Compute P(R=F|A=F) and P(G=F|A=T)

Tedious but straightforward to compute

$$P(R=F|A=F) = P(R=F,A=F) / P(A=F) = [\sum_{G=F,T} \sum_{R=F,T} P(S)P(G=F)P(R)P(A=F)]/P(A=F)$$

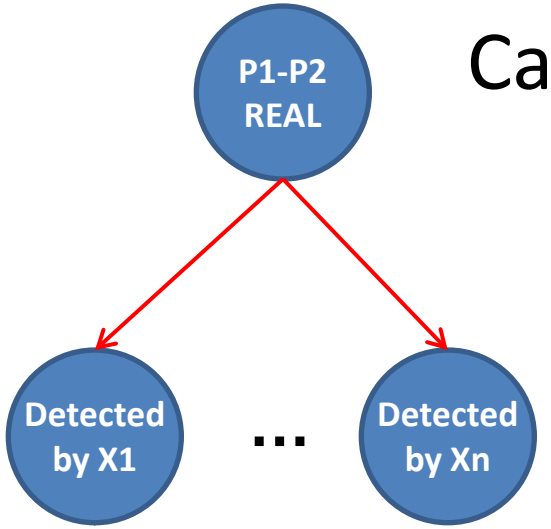
$$P(A=F) = \sum \sum \sum P(S)P(G|S)P(R|S)P(A=T|G,R) \text{ (as before)}$$

$$\frac{P(G=F|A=T)}{P(R=F|A=F)} = \frac{.92}{.56} = 1.6$$

End of worked example

Goal

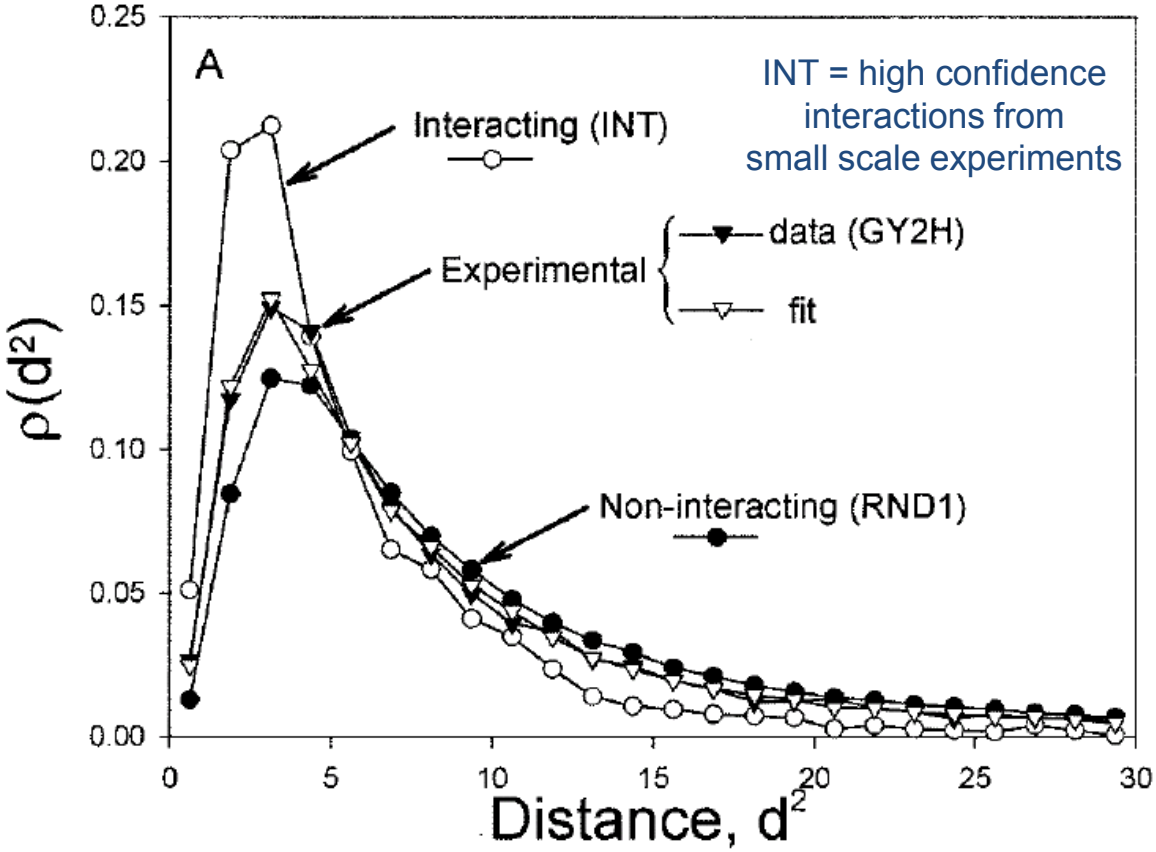
- Estimate interaction probability using
  - Affinity capture
  - Two-hybrid
  - Less physical data



Cause (often “hidden”)

Effects (observed)

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



© American Society for Biochemistry and Molecular Biology. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Deane, Charlotte M., Łukasz Salwiński, et al. "Protein Interactions Two Methods for Assessment of the Reliability of High Throughput Observations." *Molecular & Cellular Proteomics* 1, no. 5 (2002): 349-56.

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.

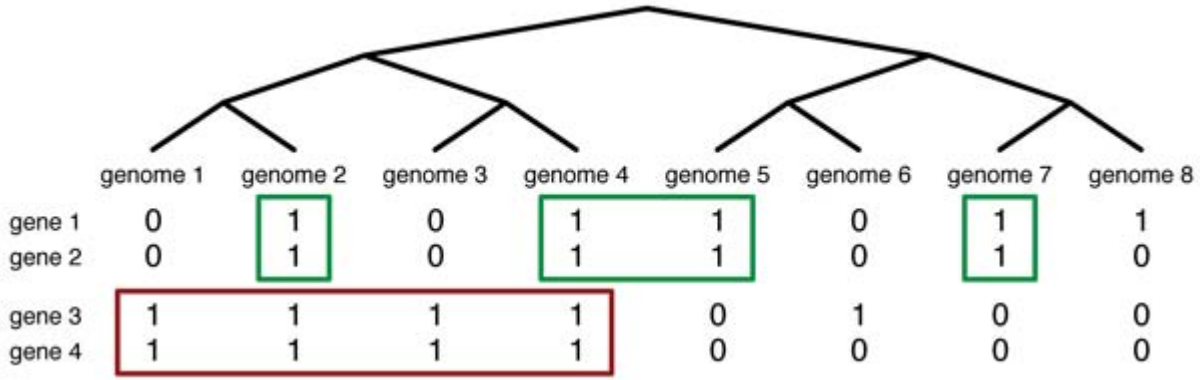
d = "distance" that measures the difference between two mRNA expression profiles

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 ⇨

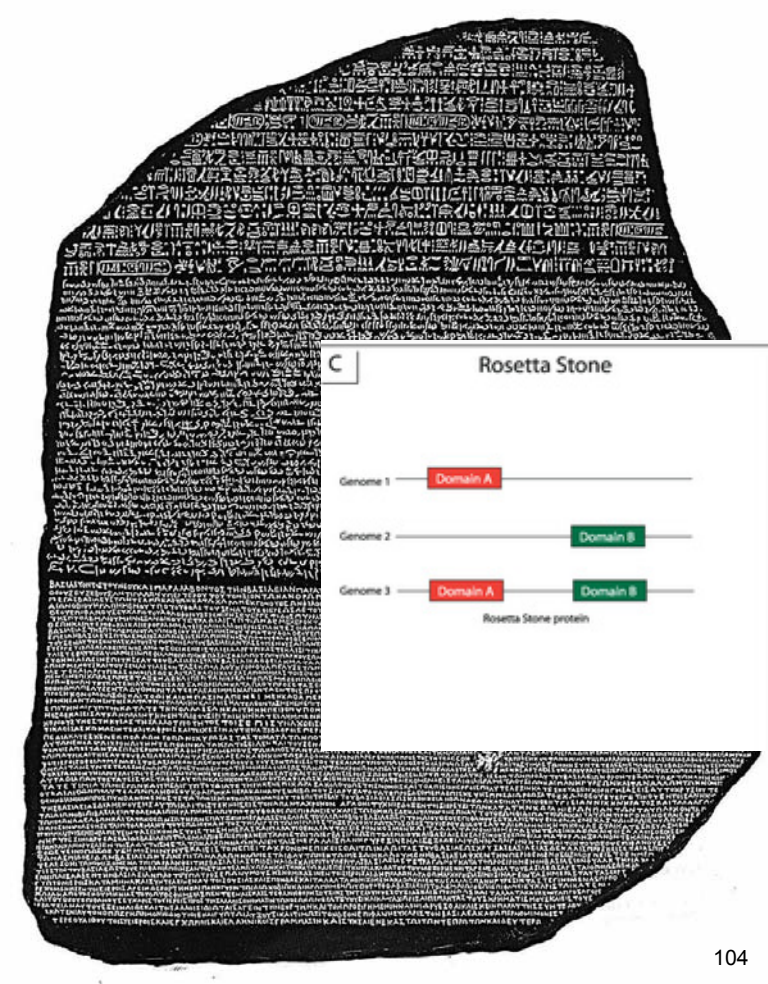


Courtesy of Cokus et al. License: CC-BY.
Source: Cokus, Shawn, Sayaka Mizutani, et al. "An Improved Method for Identifying Functionally Linked Proteins using Phylogenetic Profiles." *BMC Bioinformatics* 8, no. Suppl 4 (2007): S7.

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

Advantage of Bayesian Networks

- Data can be a mix of types: numerical and categorical
- Accommodates missing data
- Give appropriate weights to different sources
- Results can be interpreted easily

Requirement of Bayesian Classification

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

Positive training data: MIPS

- Hand-curated from literature

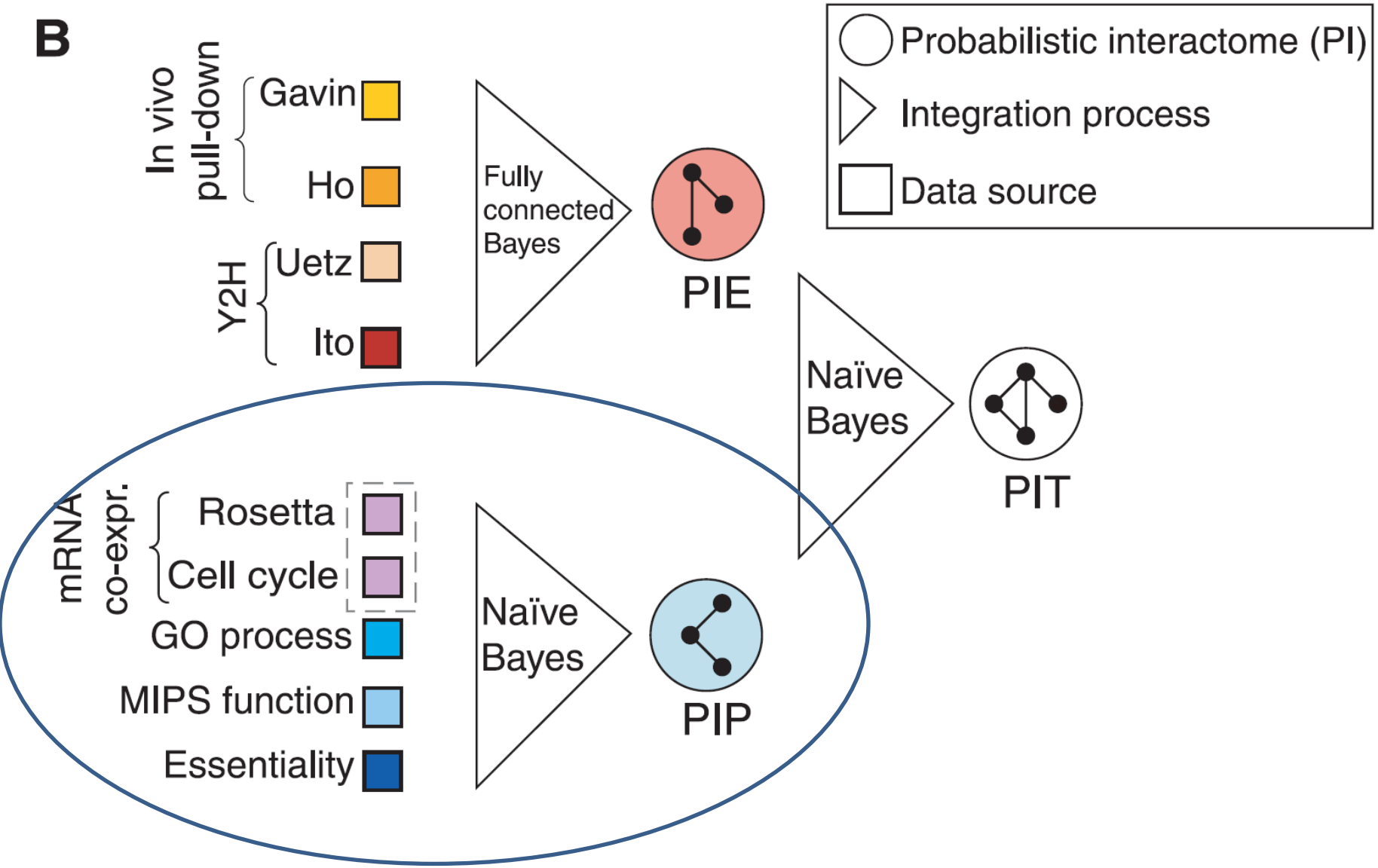
Negative training data:

- 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	

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
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.



© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
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 =

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

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= $\frac{P(f \mid pos)}{P(f \mid neg)}$

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

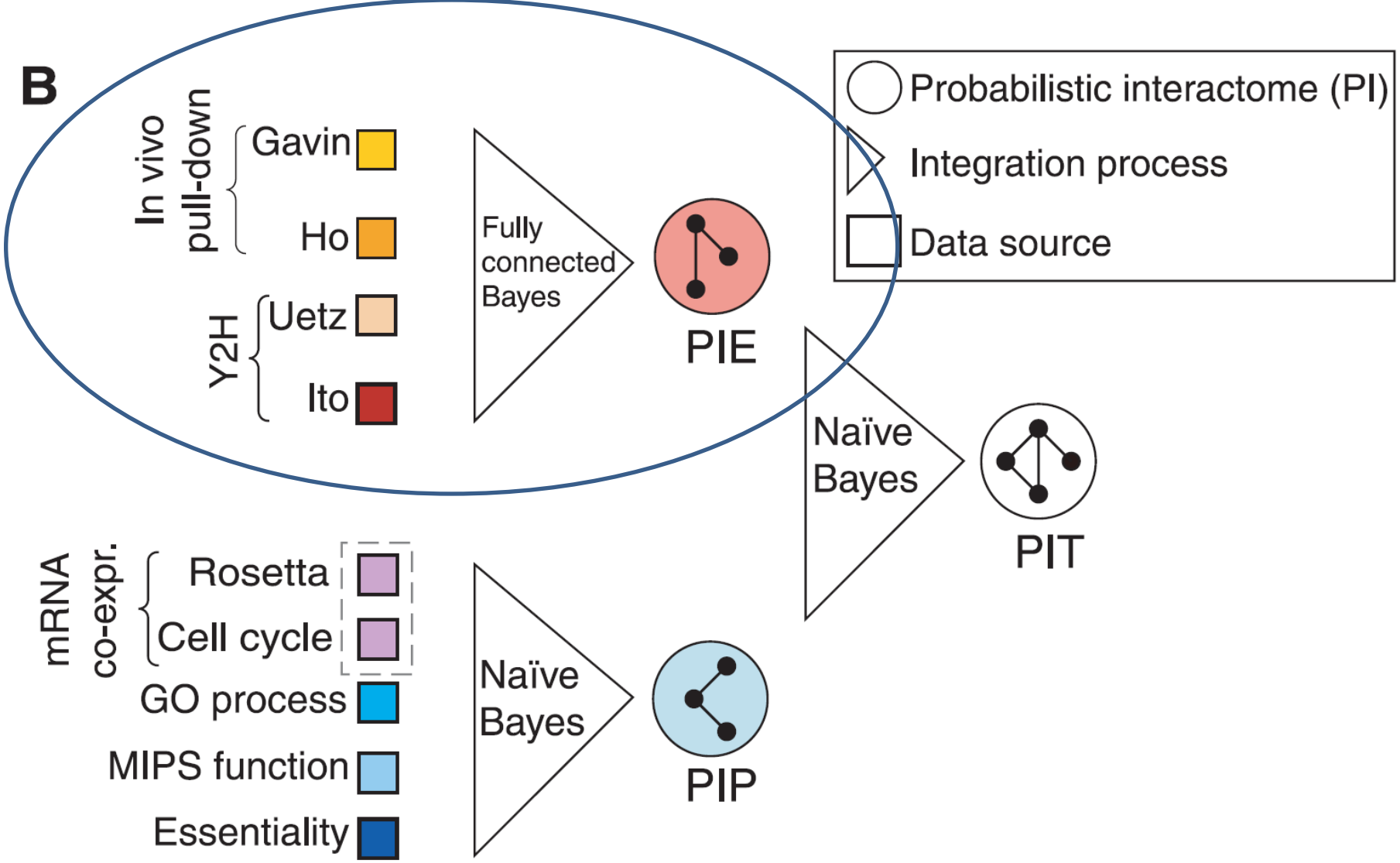
1,114/2150

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

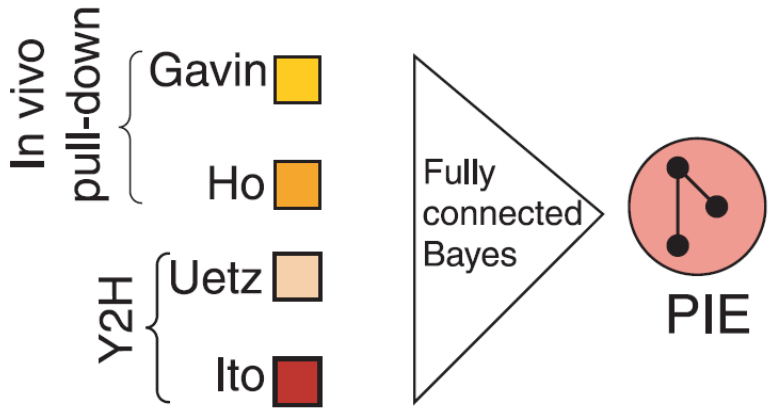
Expression correlation		# protein pairs	Gold standard overlap					P(exp|pos)	P(exp|neg)	L
			pos	neg	sum(pos)	sum(neg)	sum(pos)/ sum(neg)			
Values	0.9	678	16	45	16	45	0.36	2.10E-03	1.68E-05	124.9
	0.8	4,827	137	563	153	608	0.25	1.80E-02	2.10E-04	85.5
	0.7	17,626	530	2,117	683	2,725	0.25	6.96E-02	7.91E-04	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



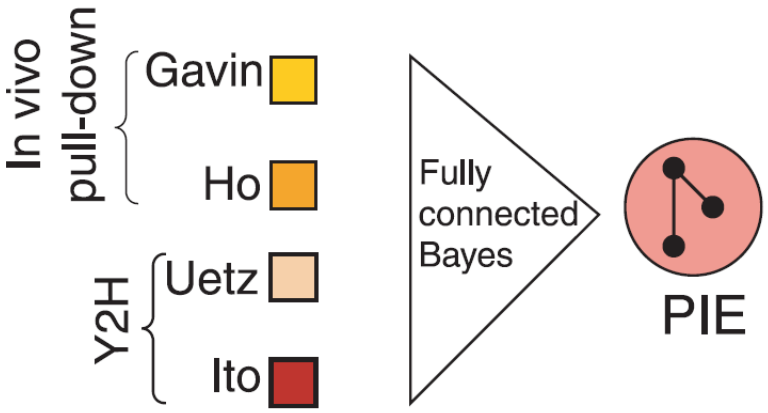
© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Jansen, Ronald, Haiyuan Yu, et al. "A Bayesian Networks Approach for Predicting Protein-protein Interactions from Genomic Data." *Science* 302, no. 5644 (2003): 449-53.



Fully connected →
Compute probabilities for all 16 possible combinations

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

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
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.

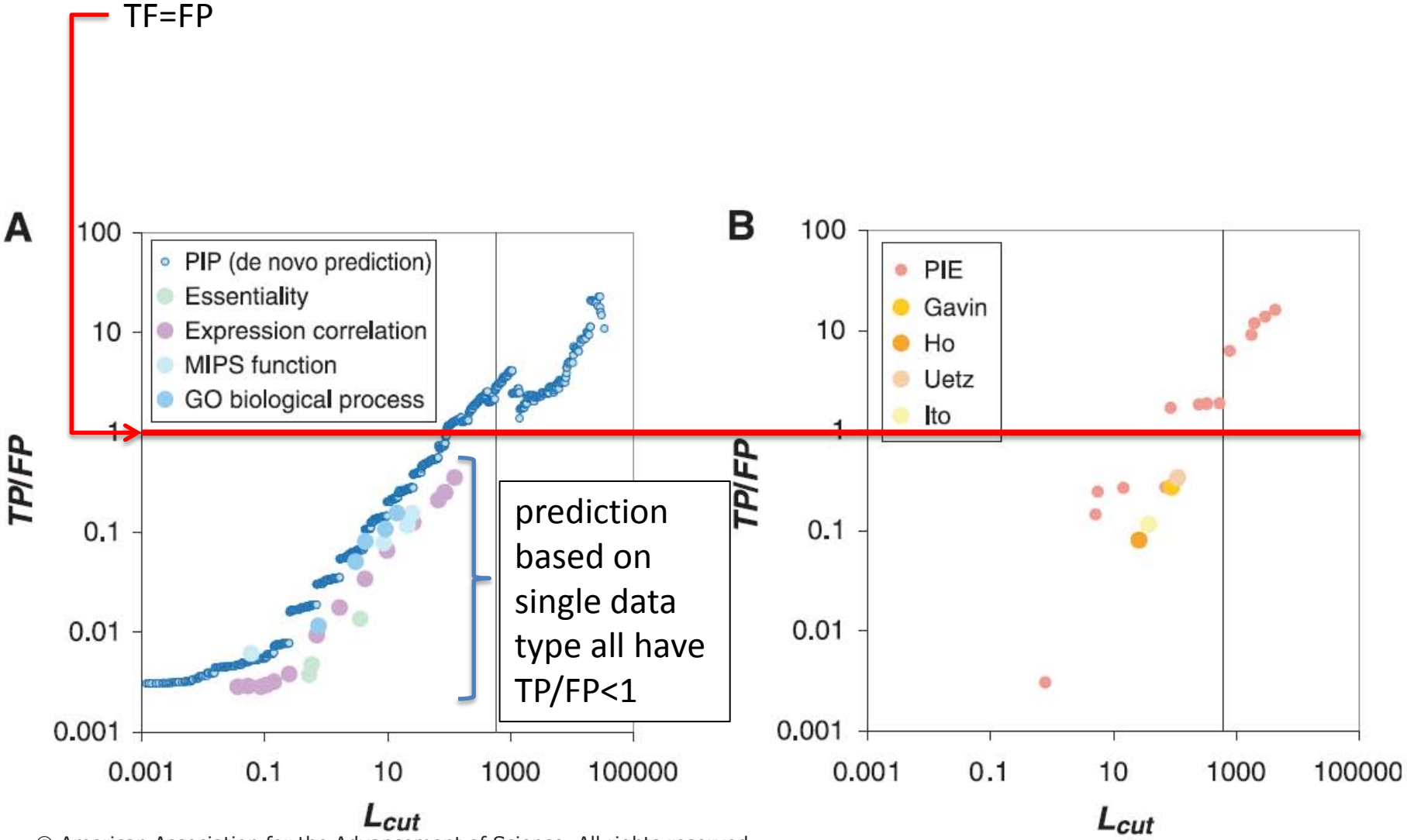


Interpret with caution, as numbers are small

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

© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

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.



© American Association for the Advancement of Science. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

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.

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

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

Summary

- Structural prediction of protein-protein interactions
- High-throughput measurement of protein-protein interactions
- Estimating interaction probabilities
- Bayes Net predictions of protein-protein interactions

MIT OpenCourseWare
http://ocw.mit.edu

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

Spring 2014

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.