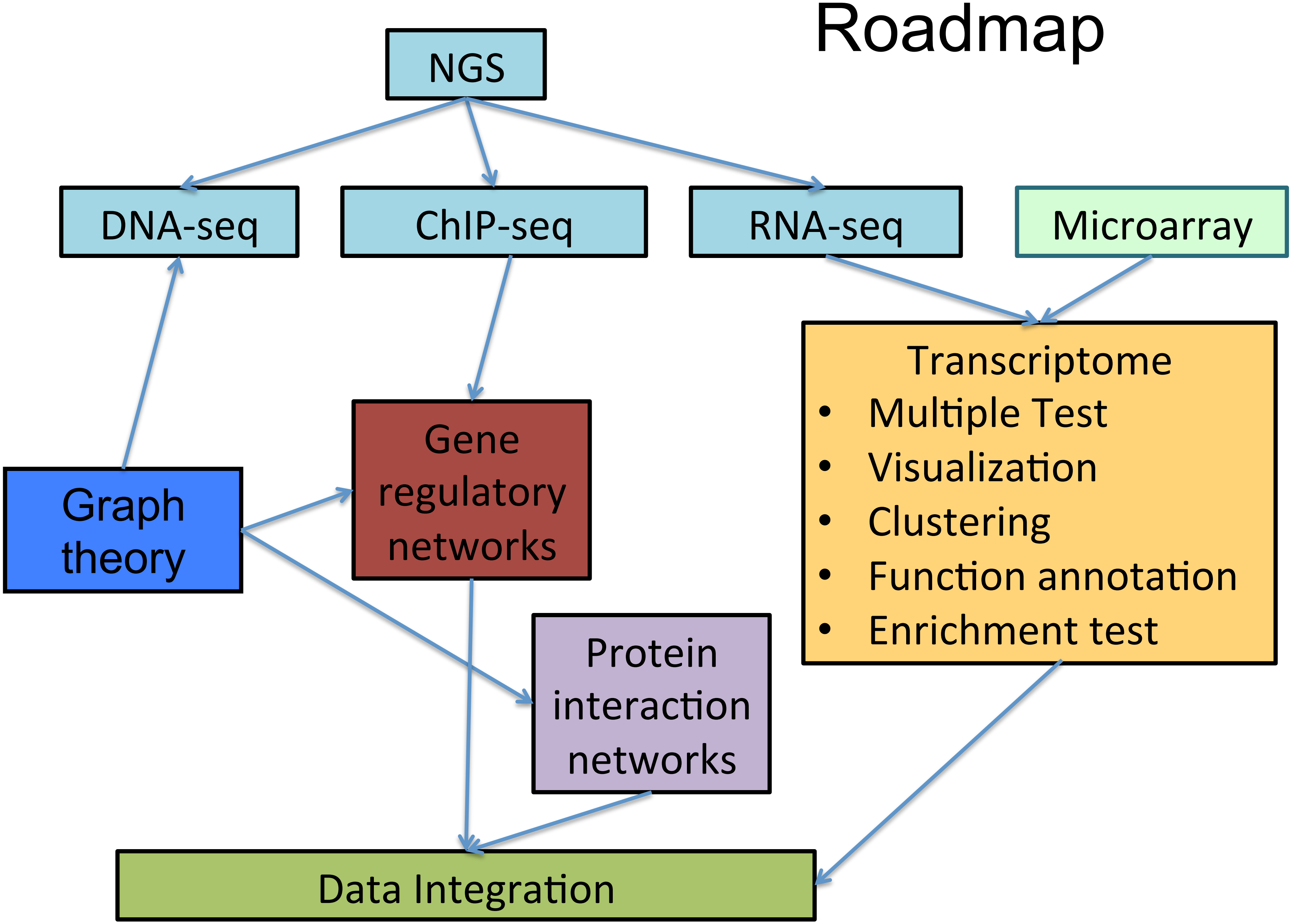
Protein-Protein Interaction Network

Lecture 1

Roadmap



NGS

DNA-seq

ChIP-seq

RNA-seq

Microarray

Graph theory

Gene regulatory networks

Transcriptome

- Multiple Test
- Visualization
- Clustering
- Function annotation
- Enrichment test

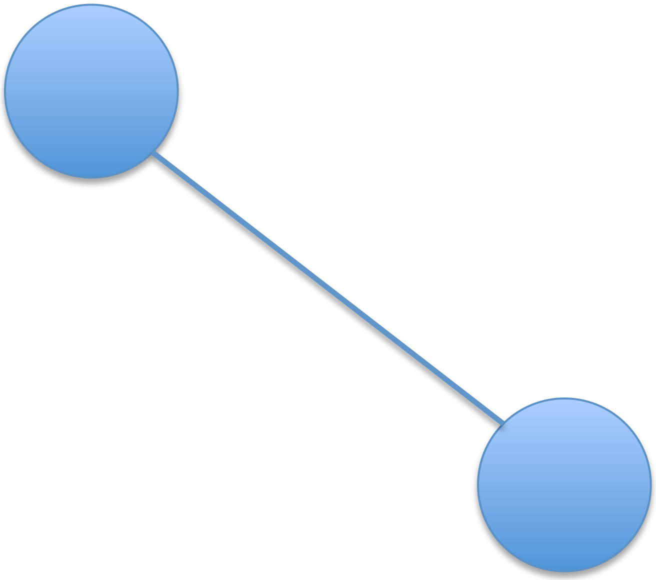
Protein interaction networks

Data Integration

Outline

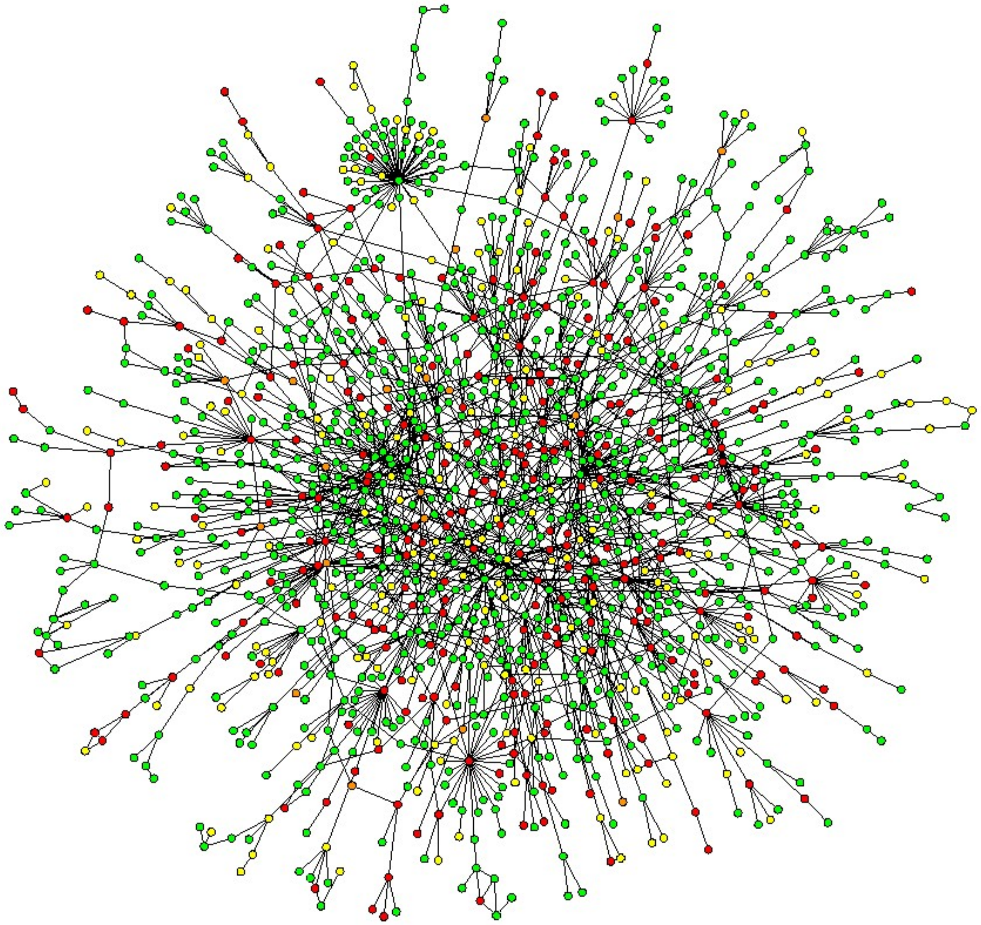
- Protein-Protein Interaction Model
- How to get PPI
  - Experimental methods (methods, results, assessing and filtering)
  - Bioinformatic methods
- PPI databases
- network properties
- Analysis method
- Integration with other omic data

Graph Model



Vertex
Edge

Yeast protein interaction network

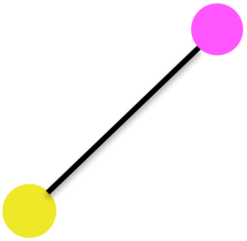


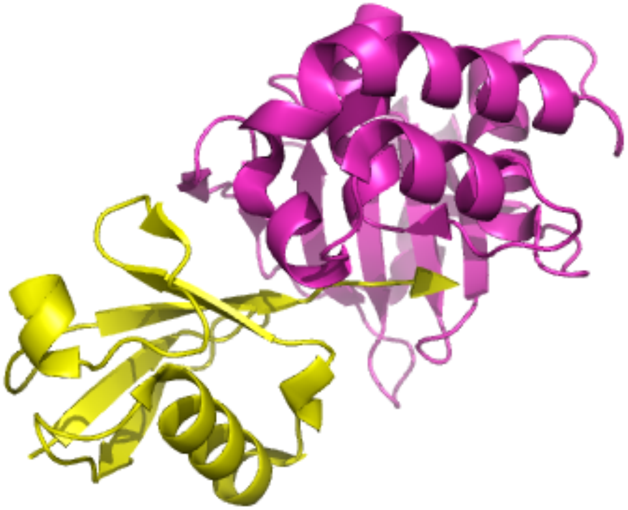
What kind of interactions?

- Protein Physical Interactions
  - Protein-protein binding
  - Enzyme and its substrates
  - Enzyme and its inhibitor
  - Protein Chaperon
  - Protein complexes

Protein Binding

- L-protein and ubiquitin

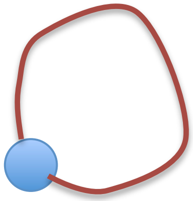


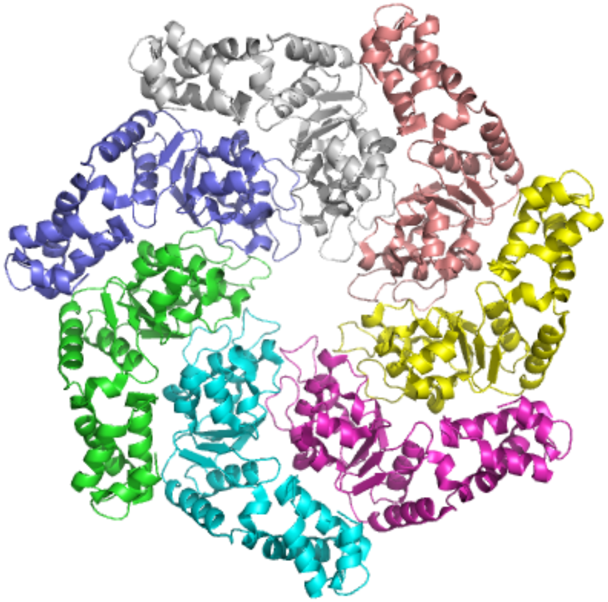


PDB: 3PRP

Protein Binding

- NtrC1 ATPase domains form a Heptamer

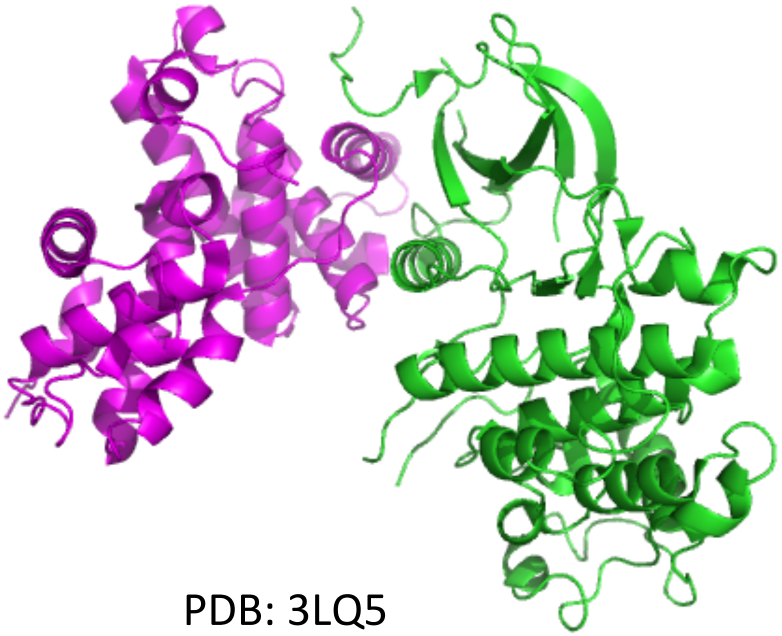




3M0E

Enzyme and its substrate

- Cell division protein kinase 9 and Cyclin-T1
- Trigger Mcl-1 Down-Regulation and Apoptotic Cell Death in Neuroblastoma Cells



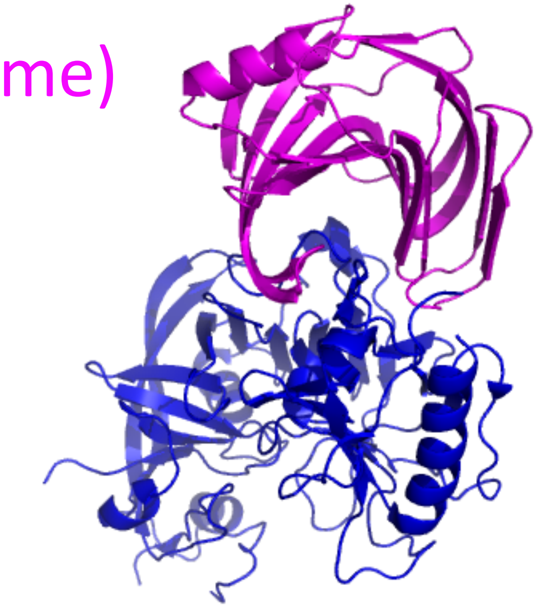
PDB: 3LQ5

Enzyme and its inhibitor

- Xylanase is a class of enzymes which degrade the linear polysaccharide beta-1,4-xylan into xylose, thus breaking down hemicellulose, one of the major components of plant cell walls.

Xylanase (Enzyme)

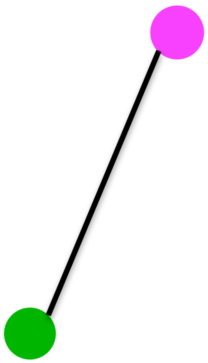
TAXI (Inhibitor)

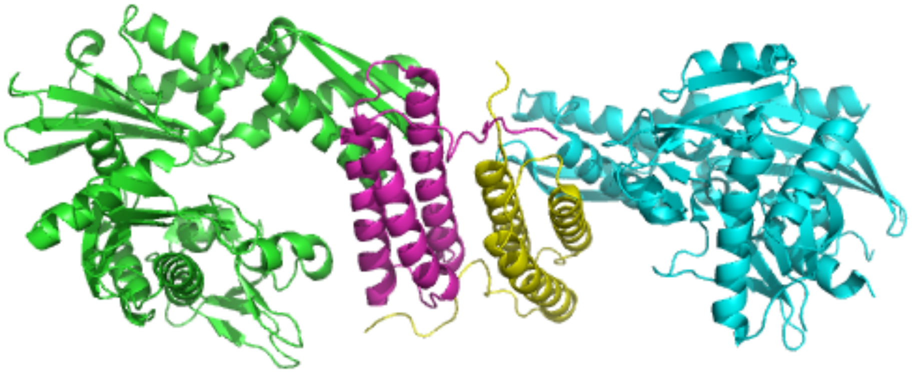


PDB: 2B42

Protein Chaperone

- Complex between the BAG5 BD5 and Hsp70 NBD

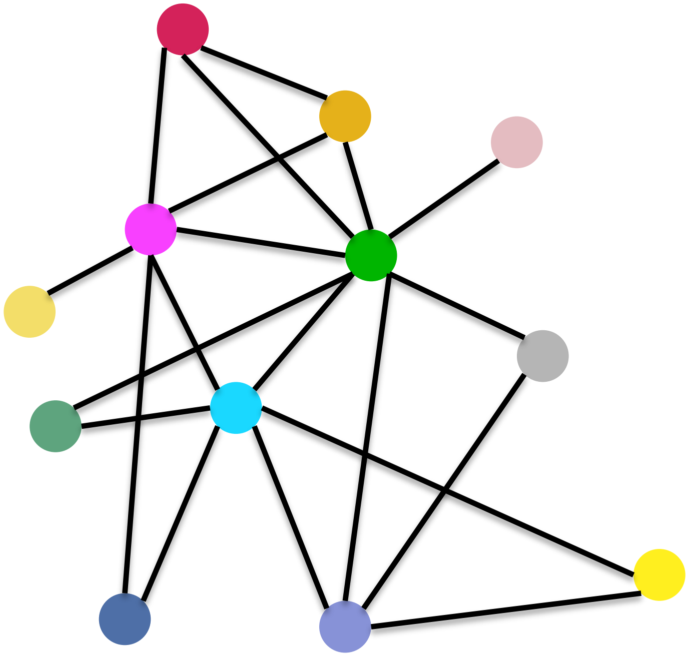


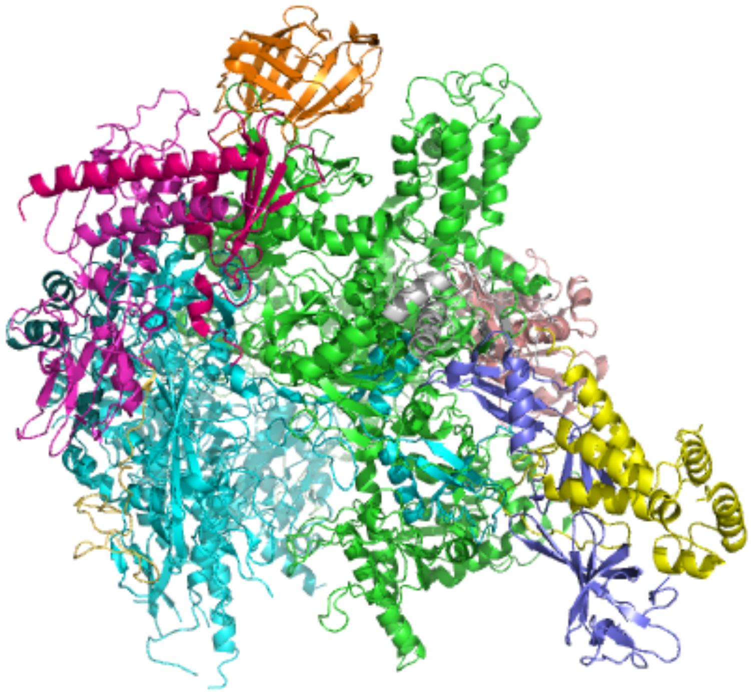


PDB: 3A8Y

Protein Complex

- 12-subunit RNA Polymerase II

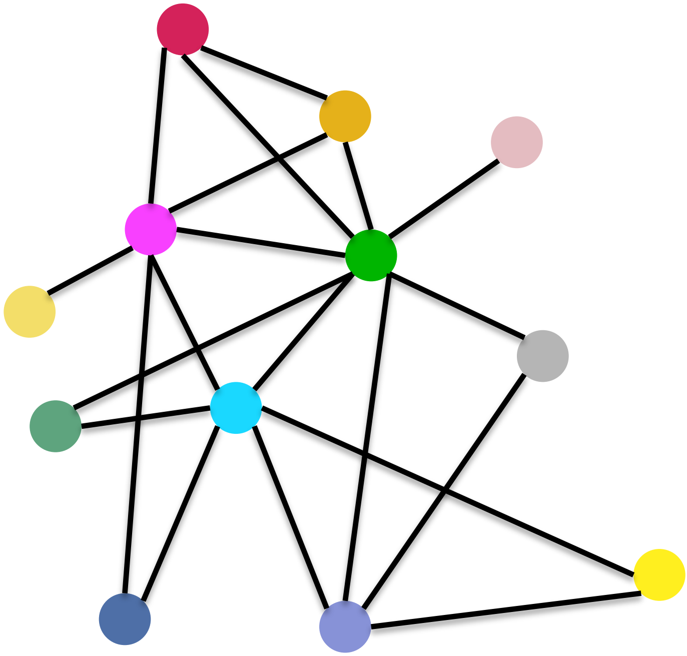




PDB: 2B8K

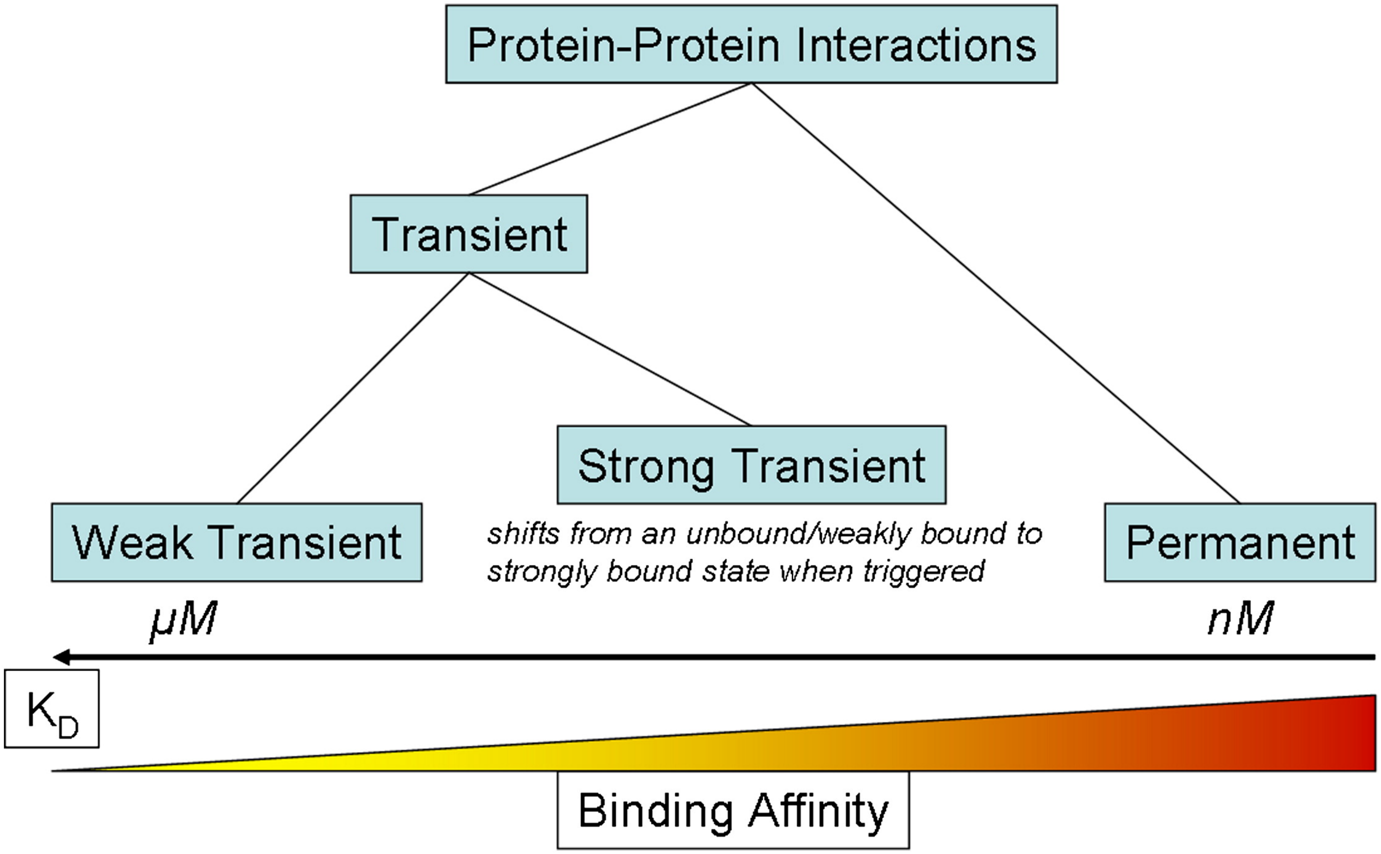
Protein Complex

- What is the connection density for this graph?



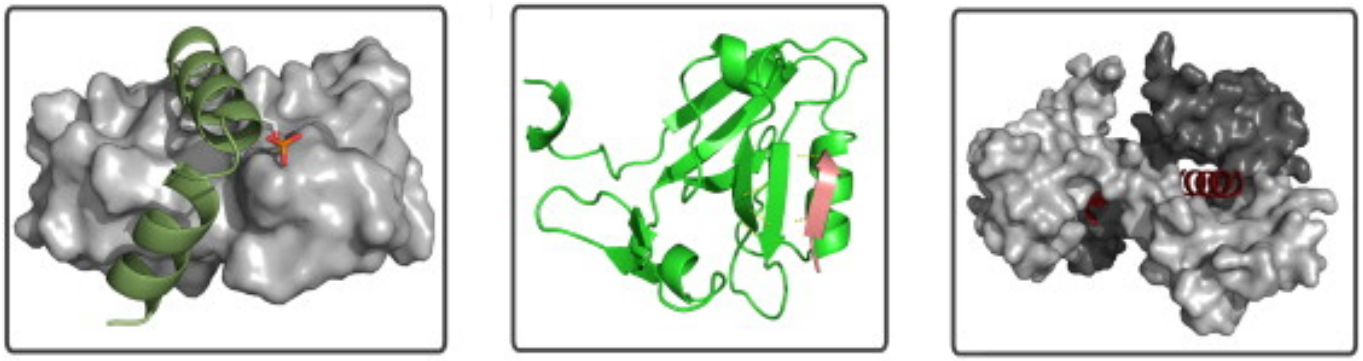
$$Q = \frac{|E|}{V(V-1)/2}$$

Permanent or Transient interactions


Protein-Protein Interactions
Transient
Strong Transient
Weak Transient
shifts from an unbound/weakly bound to strongly bound state when triggered
Permanent
μM
nM
K_D
Binding Affinity

Perkins *et al.* Structure (2010)

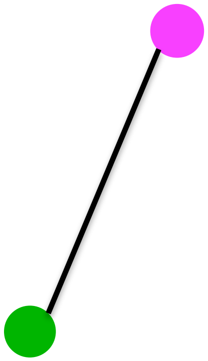
Permanent or Transient interactions



A KIX domain of CBP with KID peptide of CREB

B PSD-95 PDZ domain with its peptide

C Calcineurin - Calmodulin complex



- Difficult to measure the transient interactions.
- How to distinguish permanent and transient interactions in PPI network?

Perkins *et al.* Structure (2010)

What kind of information PPI network cannot provide?

- Protein binding affinity? No
- Network topology? Yes
- Protein binding interface? No
- Protein function? We will try

PPI networks for entire genomes

- The potential number of interactions is huge, and the number of real interactions is probably very large.
  - ~16 000–26 000 different interaction pairs in the yeast. Grigoriev Nucleic acid Research (2003)
  - ~600,000-250,000,000 interaction pairs in human genome.
- However, the current status to the knowledge of those interactions is still poor; only a small portion of those protein interaction pairs have been discovered.
- The large amount of interaction pairs is also a challenge to study them. The "network" is a suitable tool to study on the PPI data.

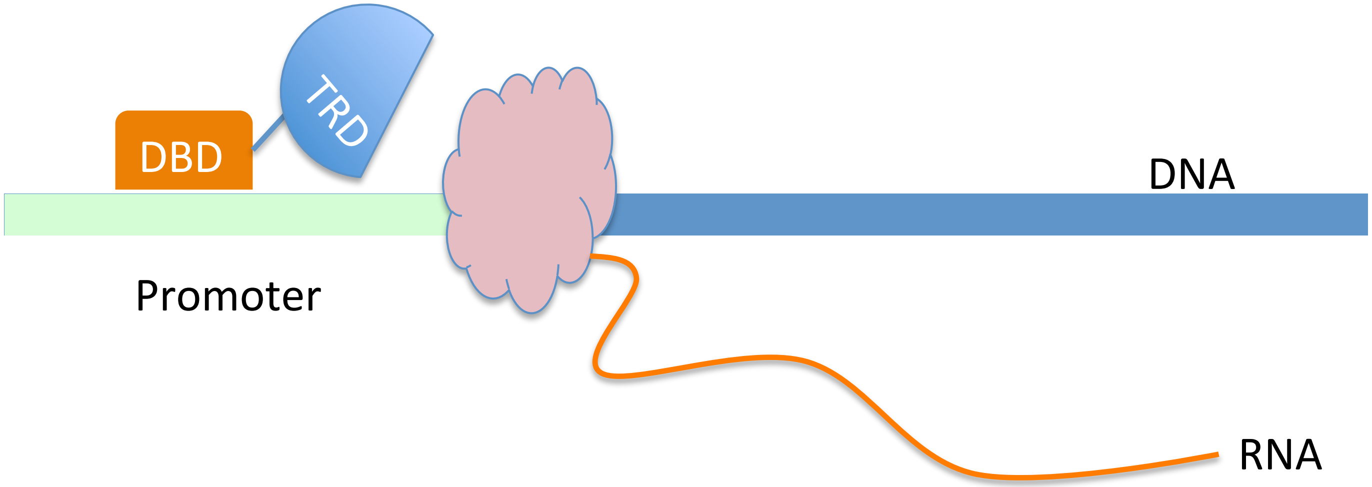
Outline

- Protein-Protein Interaction Model
- How to get a PPI network
  - Experimental methods: Y2H, MS etc.
  - Bioinformatic methods
- PPI databases and network properties
- Analysis method
- Integration with other omic data

Experimental methods

- **Co-immunoprecipitation** is considered to be the gold standard assay for protein–protein interactions, especially when it is performed with endogenous (not overexpressed and not tagged) proteins.
- **Pull-down assays** are a common variation of immunoprecipitation and are used identically, although this approach is more amenable to an initial screen for interacting proteins.
- **Chemical cross-linking** is often used to "fix" protein interactions in place before trying to isolate/identify interacting proteins.
- **Yeast two-hybrid assay**
- **Tandem Affinity purification**
- **Protein microarray**
- **Phage display**

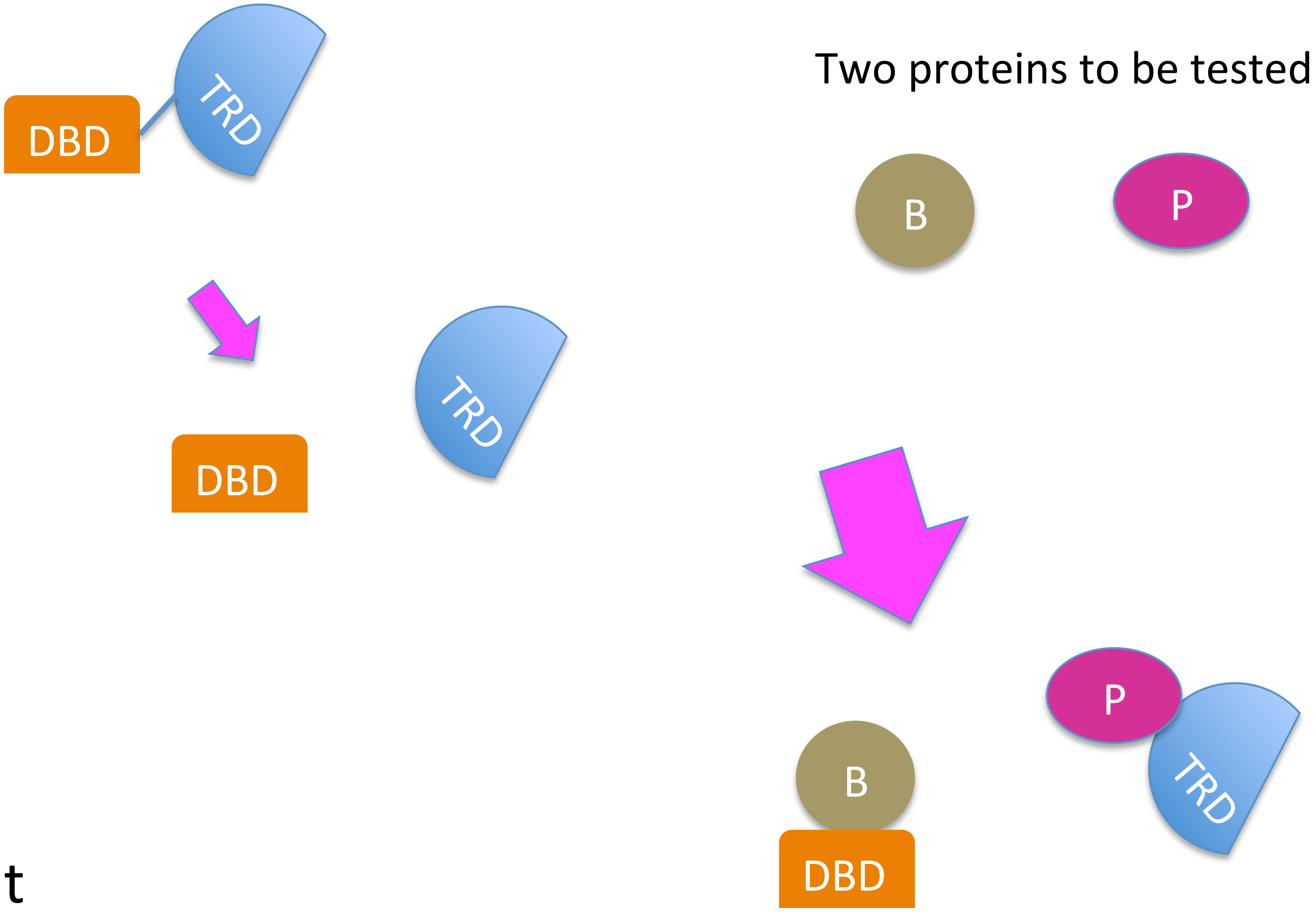
Yeast Two-hybrid Assay



DBD :DNA binding domain

TAD: Transcriptional Activation domain

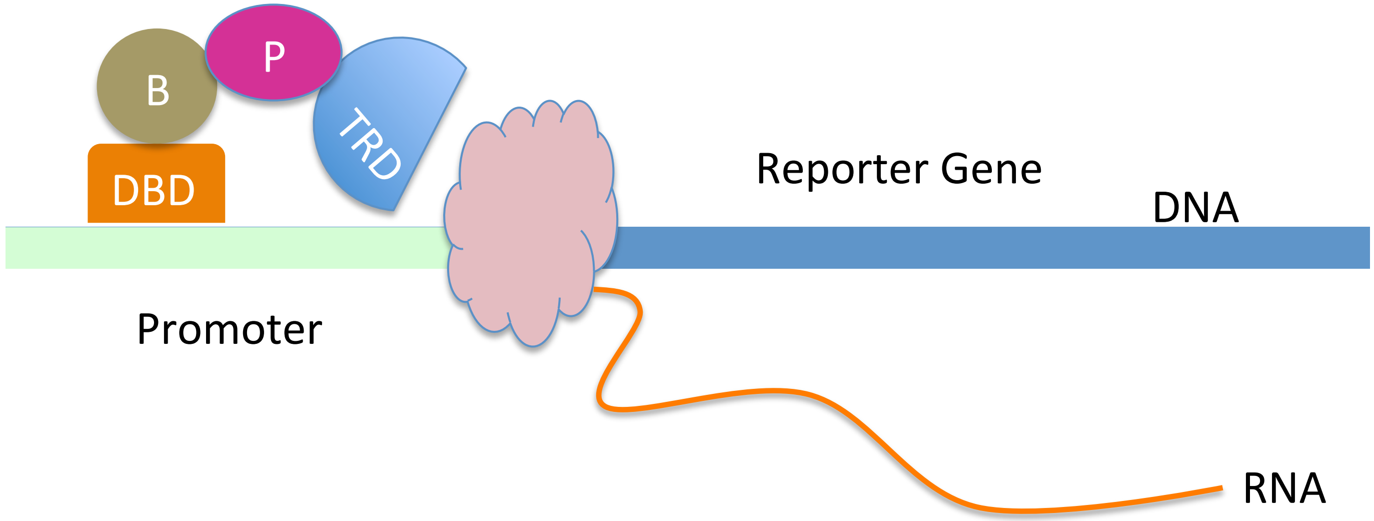
Yeast Two-hybrid Assay



Two proteins to be tested

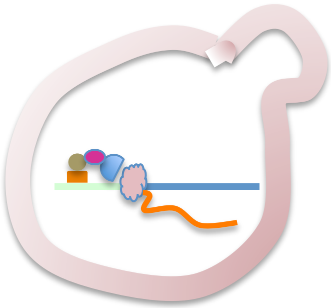
B :Bait
P: Prey

Yeast Two-hybrid Assay

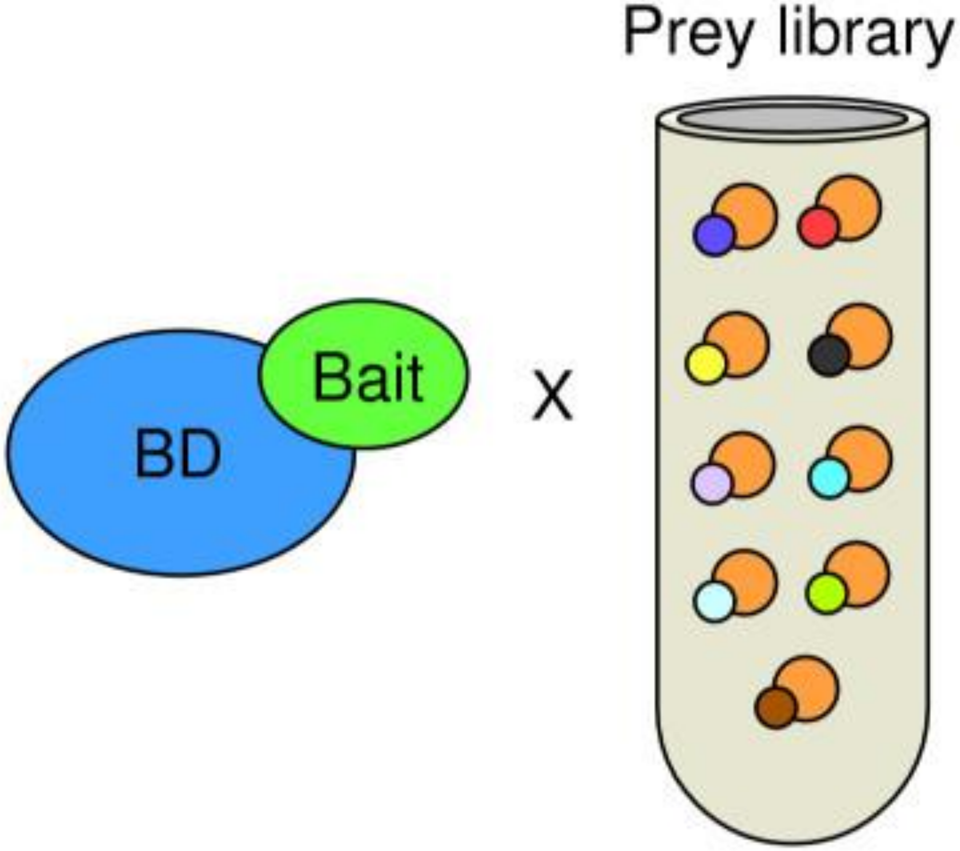


Transcription factor: Gal4

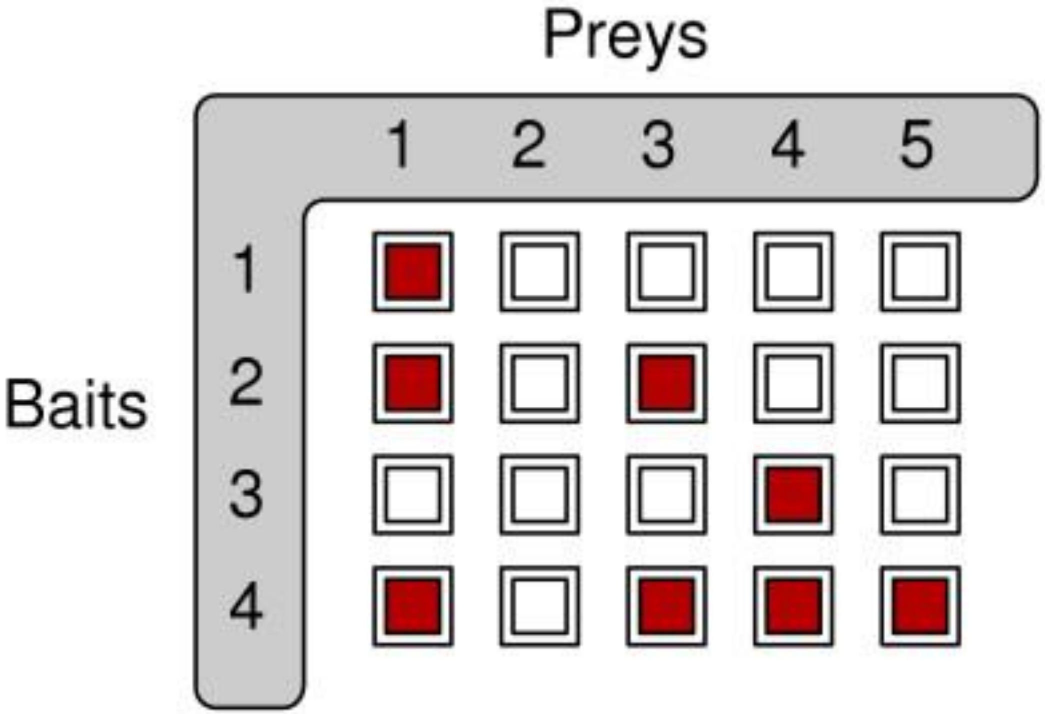
Reporter gene: LacZ



Yeast Two-hybrid Assay



Yeast Two-hybrid Assay



Baits \ Preys	1	2	3	4	5
1	🟥	□	□	□	□
2	🟥	□	🟥	□	□
3	□	□	□	🟥	□
4	🟥	□	🟥	🟥	🟥

What does this matrix is?

Yeast 2-hybrid Assay

- Pros
  - Easy/fast
  - No purification required
  - *In vivo* conditions
  - Can be adapted for high throughput screens
  - Can detect transient interactions

Yeast 2-hybrid Assay

- Cons
  - prone to false negatives because
    - protein doesn’t fold,
    - protein doesn’t localize to nucleus,
    - interference from endogenous protein,
    - fusion protein doesn’t interact like native protein,
    - fusion may be toxic to cell
  - prone to false positives
    - auto-activation
    - indirect interactions
  - not quantitative
  - no control over post-translational modifications
  - only test binary interactions

Yeast 2-hybrid assay for an entire genome

Uetz et al. Nature (2000) 403, 623-627

Two strategies:

1. "array" approach: ~6,000 activation domain hybrid transformants mated to 192 DNA binding domain fusion transformants only 20% of interactions (281) reproducible (many auto-activate), and 3.3 positives per interaction-competent protein
2. "high-throughput screen" approach: 5,345 ORFs cloned separately into DNA-binding and activation domain plasmids (2 reporter genes); DBD fusions pooled and mated to AD fusions; 12 clones per pool sequenced, gave 692 unique interactions (472 seen more than once) 1.8 positives per interaction-competent protein.

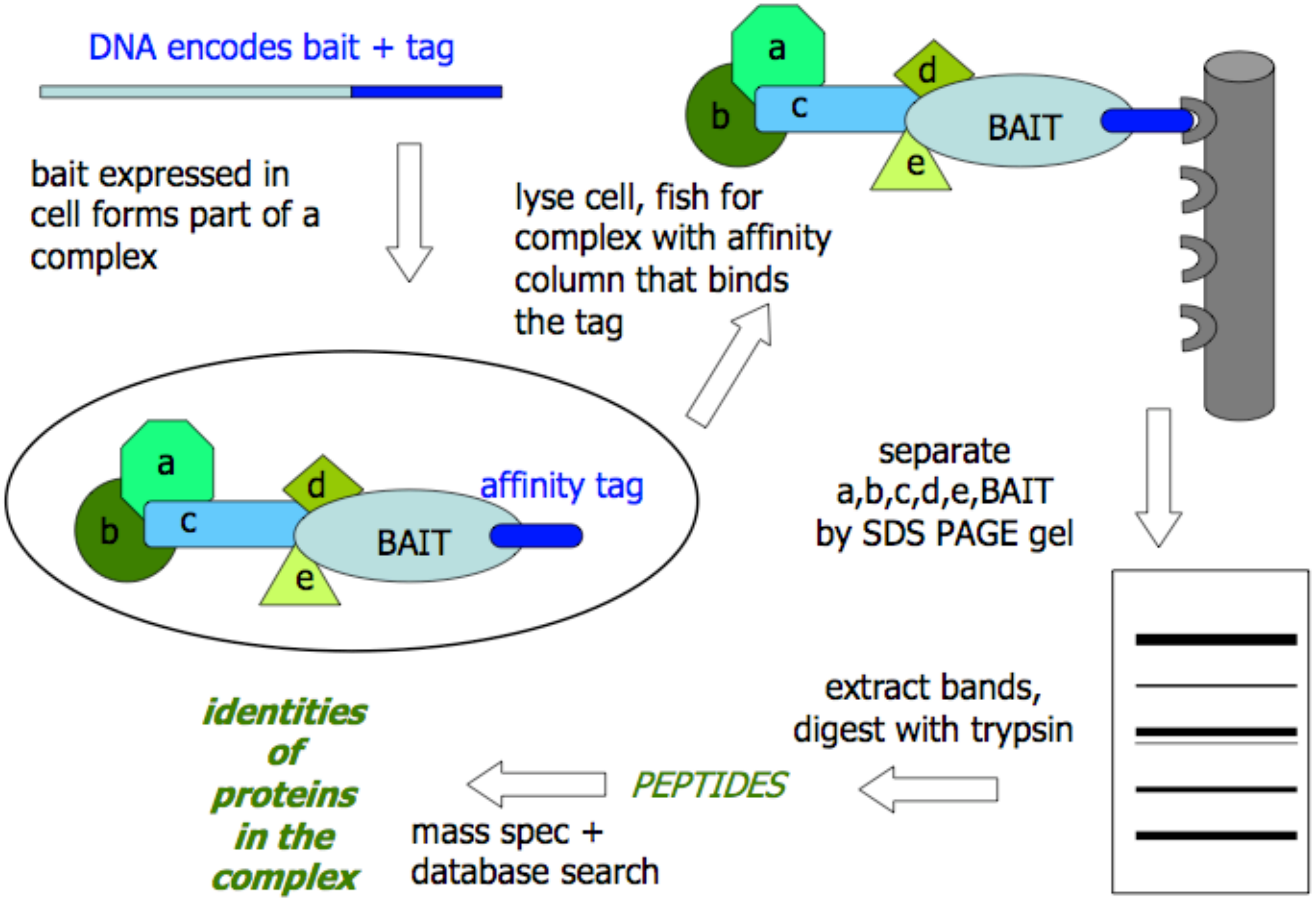
Experimental methods

- **Co-immunoprecipitation** is considered to be the gold standard assay for protein–protein interactions, especially when it is performed with endogenous (not overexpressed and not tagged) proteins.
- **Pull-down assays** are a common variation of immunoprecipitation and are used identically, although this approach is more amenable to an initial screen for interacting proteins.
- **Chemical cross-linking** is often used to "fix" protein interactions in place before trying to isolate/identify interacting proteins.
- **Yeast two-hybrid assay**
- **Tandem Affinity purification (TAP)**
- **Protein microarray**
- **Phage display**

Tandem Affinity Purification (TAP)

- Most proteins interact with several other proteins (estimate 2-10).
- Many proteins in the cell are found in complexes. For some purposes, knowing the identities of the members of the clusters is as useful, or more useful, than knowing the directly interacting partners.
- Tandem Affinity purification (TAP) is a method for characterizing the clusters directly, rather than one interaction at a time.

TAP/MS spectrometry



DNA encodes bait + tag

bait expressed in cell forms part of a complex

lyse cell, fish for complex with affinity column that binds the tag

separate a,b,c,d,e,BAIT by SDS PAGE gel

extract bands, digest with trypsin

PEPTIDES

mass spec + database search

identities of proteins in the complex

TAP/MS spectrometry for an entire genome

- Gavin et al. Nature(2002) 415, 141-147;
  - Cellzome 1,167 bait proteins in Yeast genome
  - TAP tag inserted at 3' end of gene; proteins under endogenous promoter 2 rounds of purification
  - 232 distinct complexes with 2 to 83 proteins per complex new cellular role proposed for 344 proteins
  - To assess confidence:

    Repeat the experiment -only 70% reproducible using the same bait
    Use different proteins in the complex as the bait, see if we can recover the same proteins in the complex.
- Ho et al. Nature(2002) 415, 180-183;
  - 725 bait proteins in yeast; 1,578 interacting proteins FLAG tag, proteins transiently overexpressed
  - To assess confidence: 74% of interactions reproducible in small scale co-IP/blot

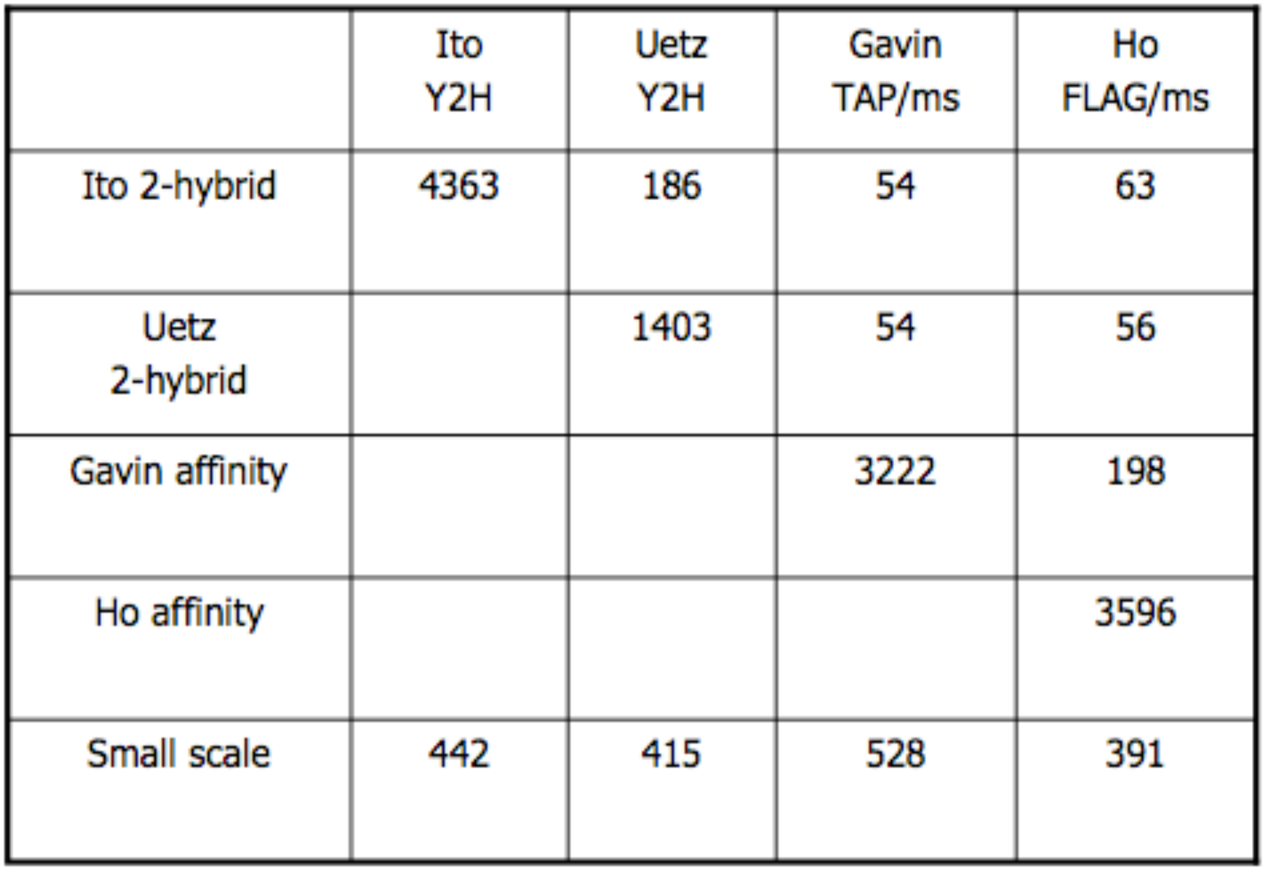
TAP/MS assay

- Pros
  - get the whole complex
  - Proteins are likely to share a function
  - very sensitive -can detect ~15 copies per cell
  - *in vivo* conditions
  - can be adapted for high-throughput screens

TAP/MS assay

- Cons
  - doesn't determine direct or indirect interactions
  - not reliable for small proteins (< 15 kD)
  - affinity tag may interfere with interactions or with the function of essential proteins
  - prone to false positives, e.g. "sticky" proteins
  - prone to false negatives
    - won't get every protein every time
    - complex must survive purification
  - not quantitative

Overlap of high-throughput interaction studies is LOW



	Ito Y2H	Uetz Y2H	Gavin TAP/ms	Ho FLAG/ms
Ito 2-hybrid	4363	186	54	63
Uetz 2-hybrid		1403	54	56
Gavin affinity			3222	198
Ho affinity				3596
Small scale	442	415	528	391

data from Salwinski & Eisenberg, Current Opinion in Structural Biology (2003) 13, 377-382

Conclusions

- Lots of protein-protein interaction data are now available for yeast, but it is not very reliable and not comprehensive.
- Need additional accessing and filtering steps.
- Nevertheless, these data have inspired the development of many computational methods.
- To facilitate computational analysis, need to disseminate the data in a usable form! This is often a rate limiting step in systems biology.

High throughput interaction data

- Not reliable
- Noisy

- Computational methods for improving the quality of interaction data
  – Assessment and validation

Assessing and filtering Criteria

- Promiscuity criteria
- Overlap criteria
- Topology criteria

Assessing and filtering Criteria

- Promiscuity criteria
  - In most high-throughput interaction studies, a few proteins are observed to interact promiscuously. Generally these are removed from the analysis.
  - Problem: some interactions may be real!
- Examples:
  - Using TAP/MS even without a bait, 17 proteins were found in pull-downs by Gavin et al. 49 other proteins found to have a similar frequency of interaction to these false positives were thrown out.
  - Using Yeast 2-hybrid, proteins were observed to make many interactions in many screens usually discarded as probably false positives.

Assessing and filtering Criteria

- Promiscuity criteria
- Overlap criteria
- Topology criteria

Assessing and filtering Criteria

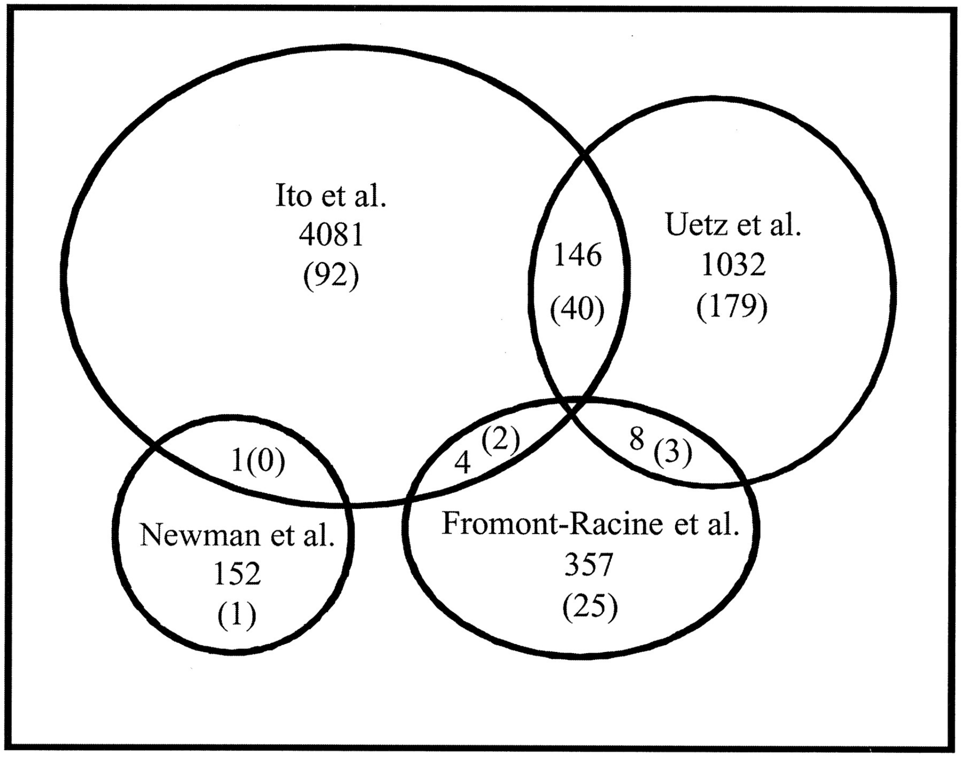
- Overlap criteria
  - An interaction has higher possibility to be real if two different types of methods discover it.
- Methods:
  - With interaction data.
  - With non-interaction data.

Assessing and filtering Criteria

With interaction data:

intersection is low!

E.g. compare Y2H and TAP/MS. Unfortunately, overlap is low.



Assessing and filtering Criteria

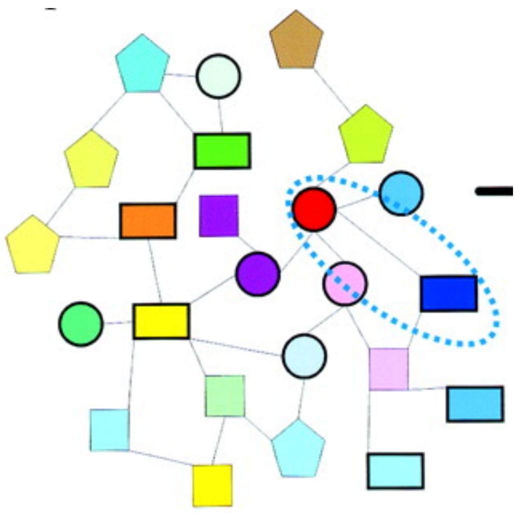
- Overlap criteria
- Methods:
  - With non-interaction data.
    - Expression Profile Reliability (EPR)
    - Homology methods -Paralogous Verification (PVM)
    - Domain Pair Verification (DPV)

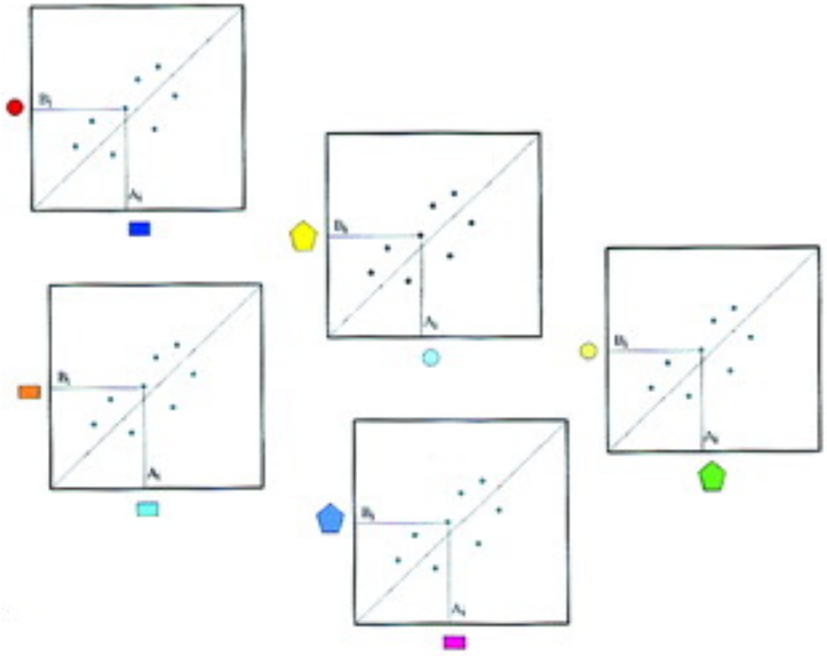
Deane et al. (2002) *Mol. Cell. Proteomics*

Expression Profile Reliability (EPR)

- **Expression Profile Reliability Index** (*EPR Index*) evaluates the quality of a large-scale protein-protein interaction data sets by comparing the expression profile.
- Two proteins have high possibility to interact with each other, if they **co-express**.

EPR

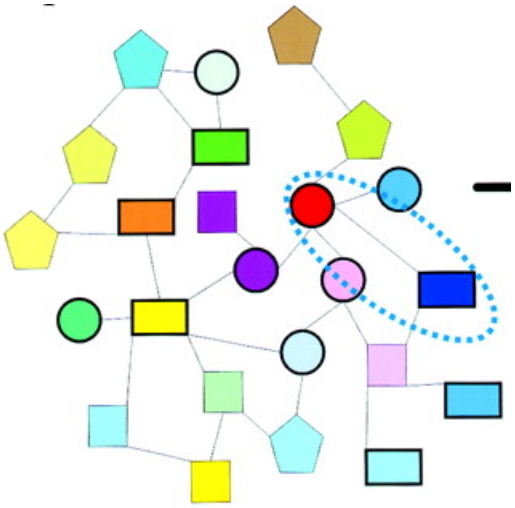


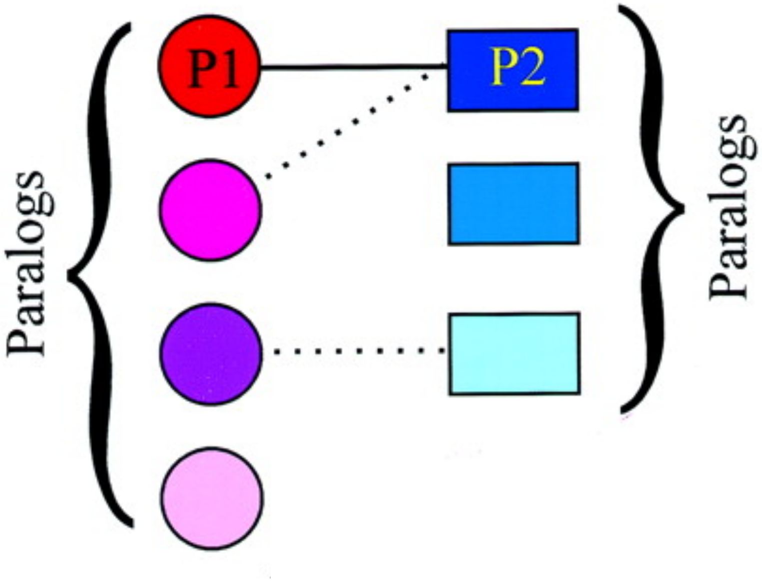


Collect the mRNA expression levels of the interaction pairs under several conditions, and calculate their expression correlations.

Deane et al. (2002) *Mol. Cell. Proteomics*

Paralogous Verification Method (PVM)





Count the number of paralogous interactions,
If the PVM score =2, they have a interaction.

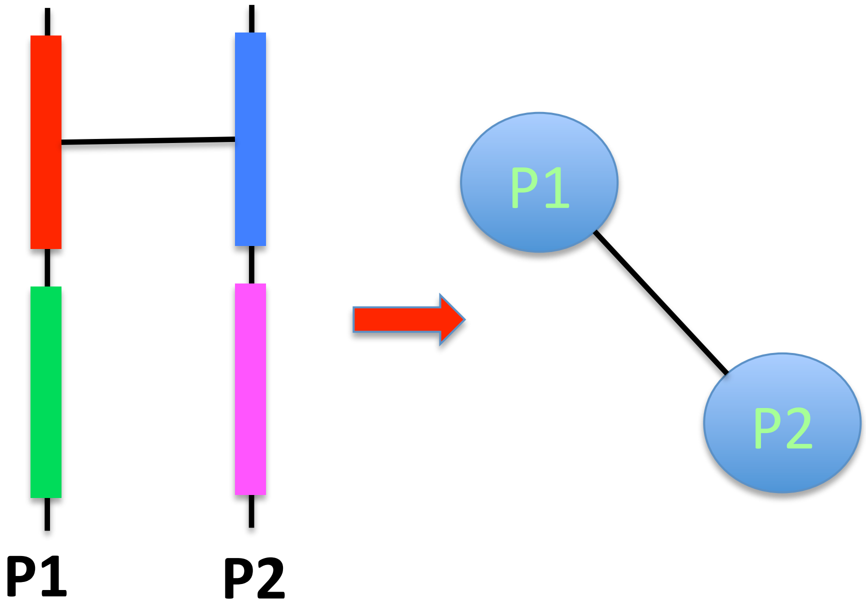
Homologous sequences are **paralogous** if they were separated by a gene duplication event: if a gene in an organism is duplicated to occupy two different positions in the same genome, then the two copies are paralogous.

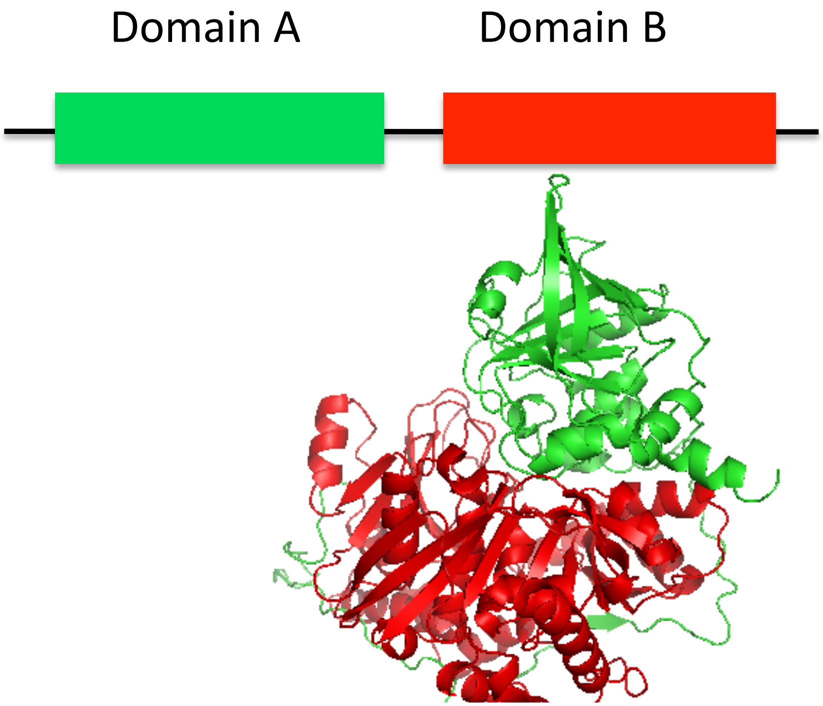
Paralogous Verification Method (PVM)

- PVM is very accurate; if a pair scores by PVM, it is almost certainly a true interaction.
- PVM does not have good coverage; it is not sensitive. PVM only confirms around 50% high-confidence samples. This is because many examples of paralogous complexes are sparse.

Domain Pair Verification (*DPV*)

- If two domains have an interaction, any two proteins that have those two domains also have interactions.
- Protein 3D structures can provide the atomic detains for protein interactions.
- The solved structures most are a single domain instead of a full length protein.



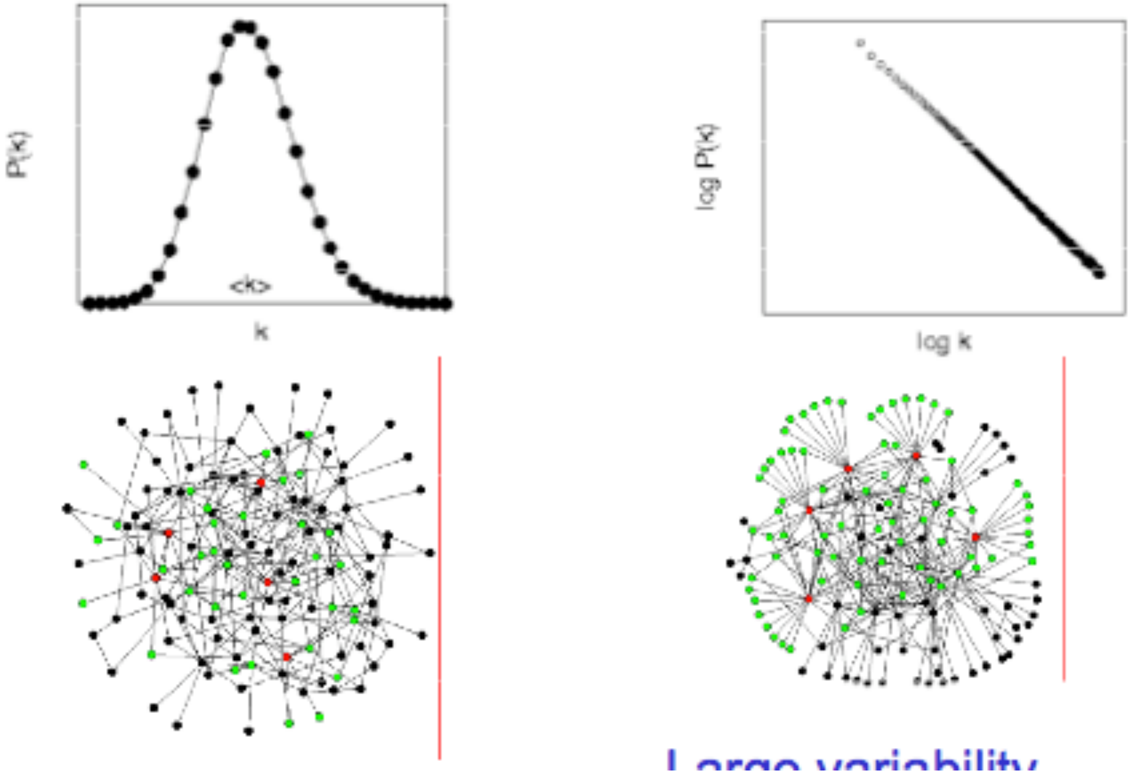


Assessing and filtering Criteria

- Promiscuity criteria
- Overlap criteria
- Topology criteria

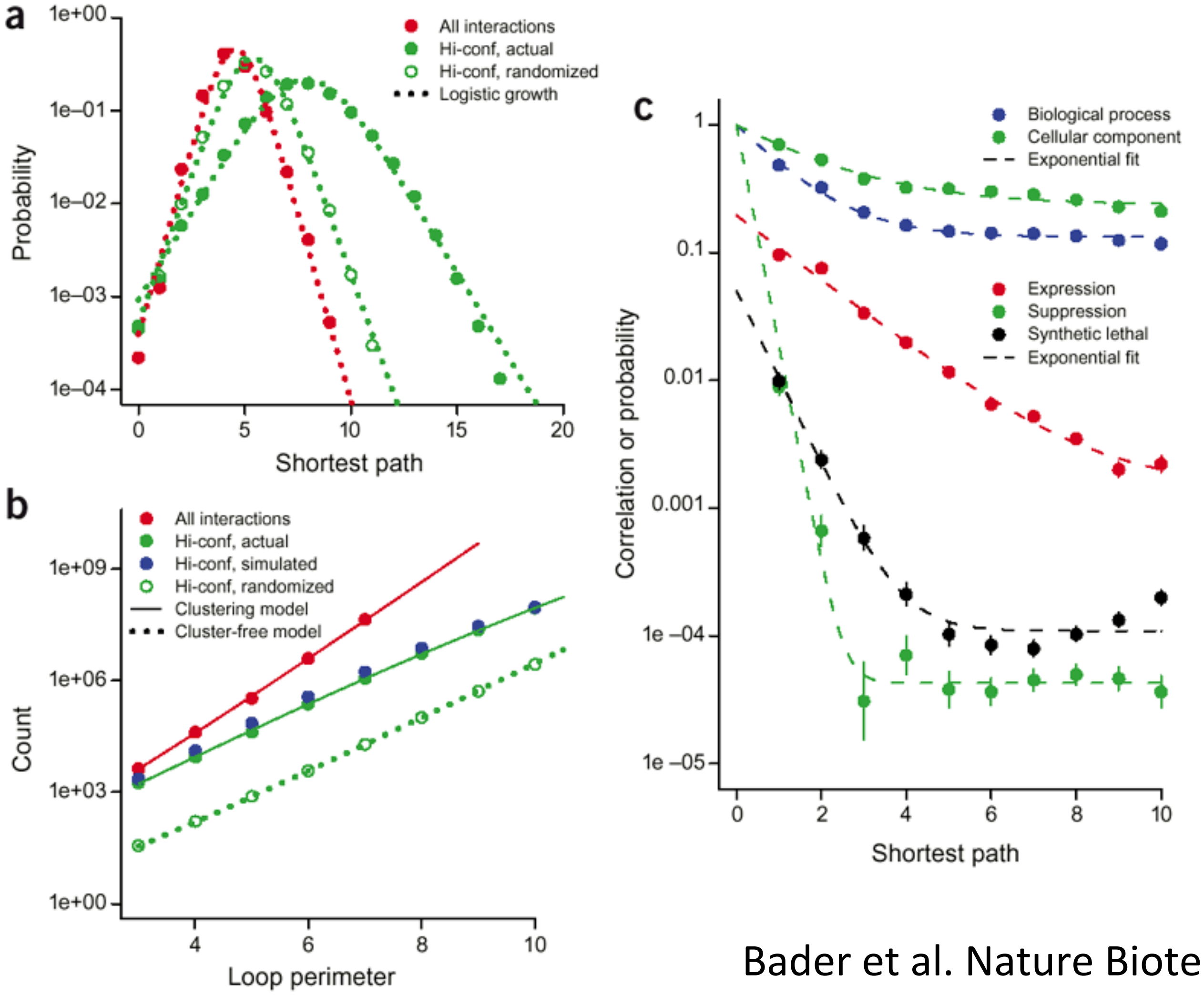
A scale free network

- Power-law degree distributions were found in diverse networks



Topology criteria

- Use information about the observed vs. expected interaction network.



Bader et al. Nature Biotechnology (2003) 22, 78-85

Outline

- Protein-Protein Interaction Model
- How to get PPI
  - Experiments: Y2H, MS, etc.
  - Bioinformatics
- PPI databases and network properties
- Analysis method
- Integration with other omic data

Why do we need bioinformatics way to generate PPI networks?

- Only model organisms have high throughput PPI data. For example, yeast and human. How about maize?
- High throughput method is expensive and time consuming.

Bioinformatics methods

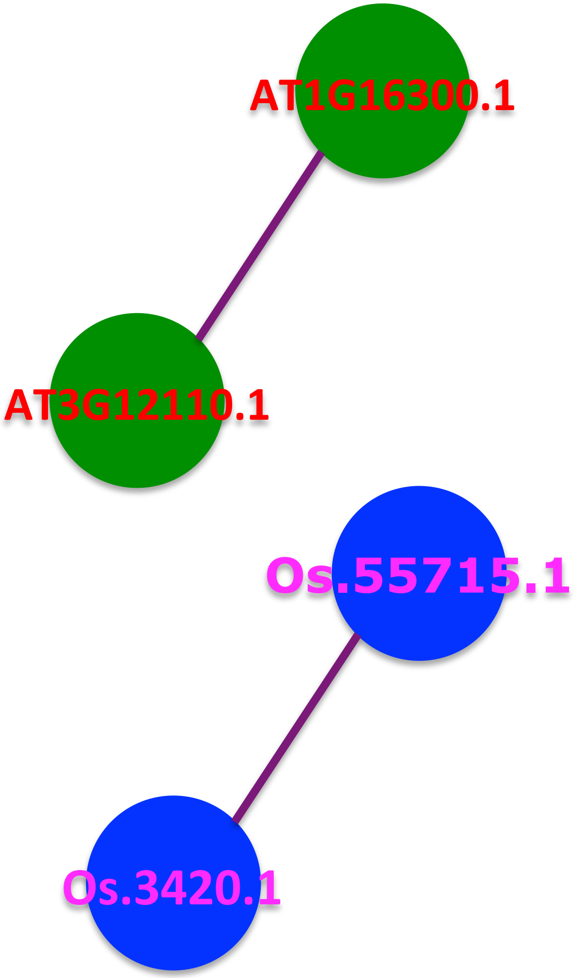
- Homologous method to find Orthology
- Combination with other information, such as expression profile, GO annotations.
- Prediction
  - Sequence method
  - Structural based method
- Text mining

An example: Rice PPI

- http://www.harvest-web.org/

Rice	ATH
Os.3420.1	AT3G12110.1
…	…
Os.52771.1	AT5G60390.3
Os.55715.1	AT1G16300.1
Os.5492.1	AT3G56070.2
…	…

7000 15000



Bioinformatics methods

- Homologous method to find Orthology
- Prediction
  - Sequence method
  - Structural based method
- Text mining
- Infer from other networks, such as expression profile, GO annotations.

Predicting protein-protein interactions

- Sequence methods
- How can you predict that an interaction might occur between two proteins based purely on sequence data?

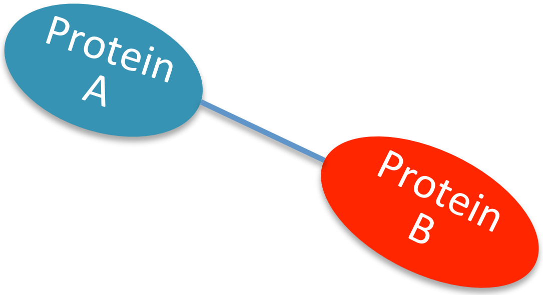
Valencia & Paz o s, (2002) Current Opin ion in Structural Biolog y 12, 368-373
Skrabanek et al. (2008) Mol Biotechnol. 38(1):1-17.

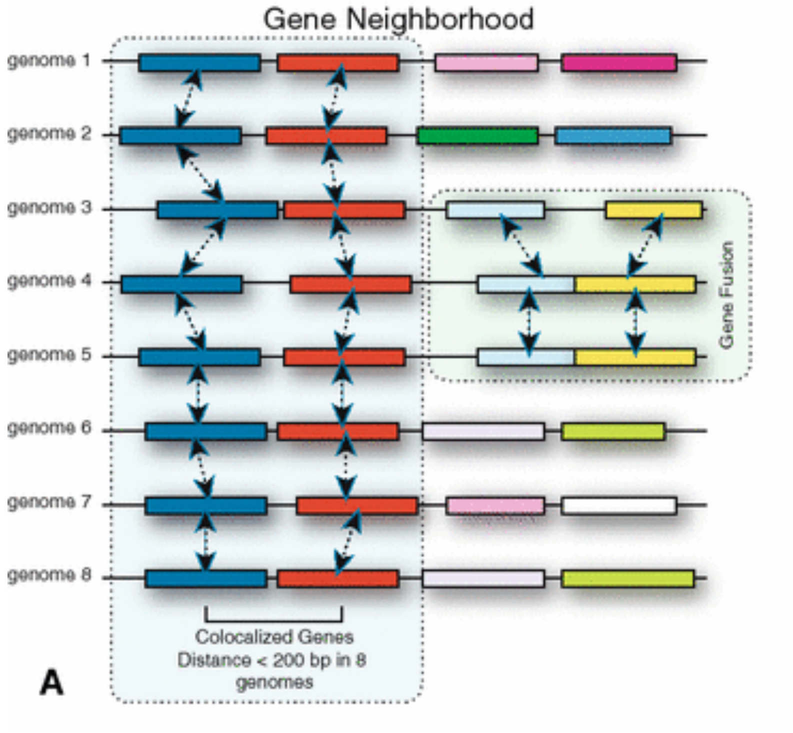
Prediction PPI with sequences

- Gene neighborhood
- Gene fusions
- Phylogenetic profiles
- Co-evolution
- Correlated Mutation
- Domain interaction

Prediction PPI with sequences

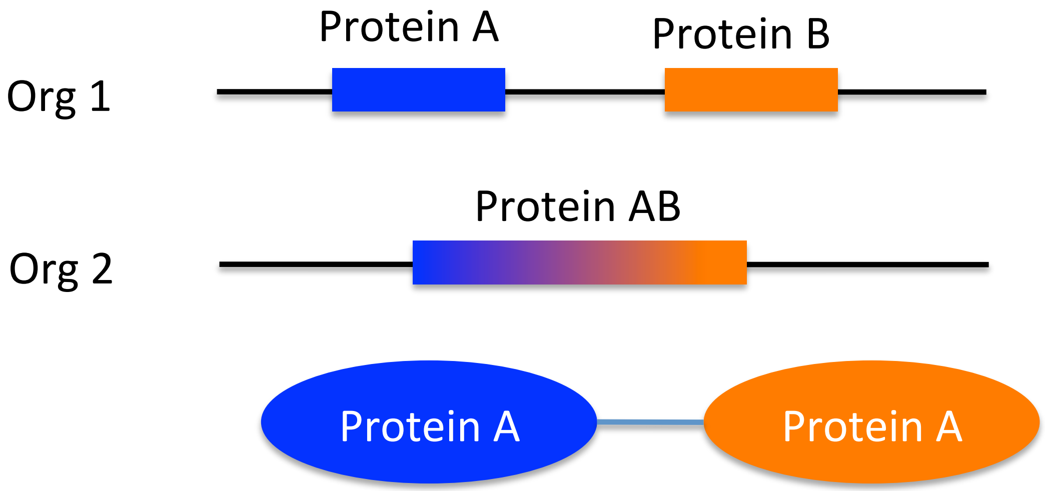
- Gene neighborhood
  - for bacteria, the arrangement of genes in operons means that interacting proteins are often encoded in adjacent sites in the genome





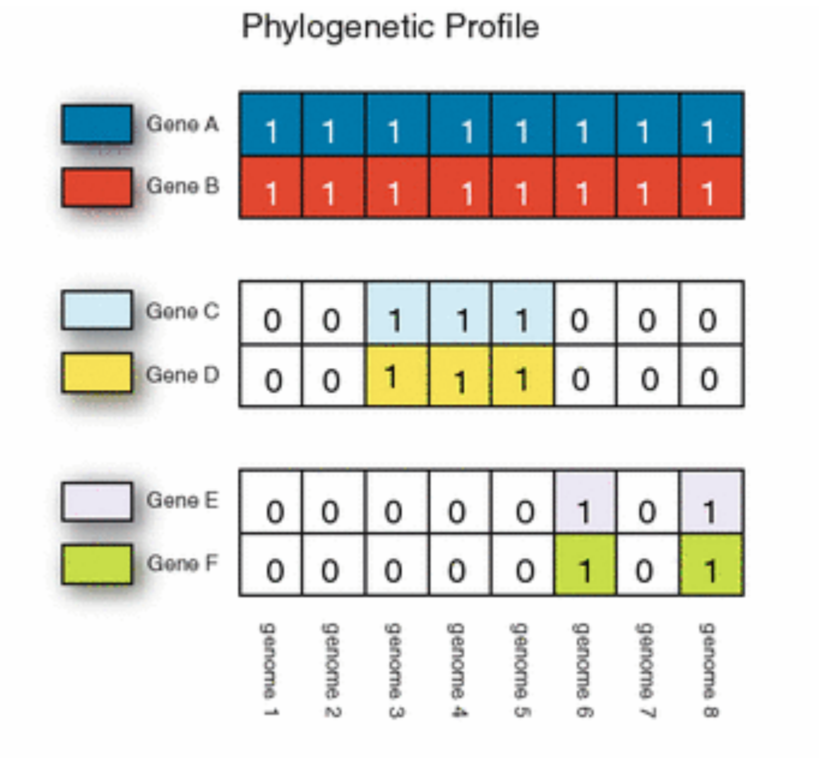
Prediction PPI with sequences

- Gene fusions
  - genes encoding interacting proteins in one organism are sometimes fused into a single gene in another. Look for these occurrences.



Prediction PPI with sequences

- Phylogenetic profiles
  - based on the joint presence/absence of a pair of proteins in a large number of genomes.

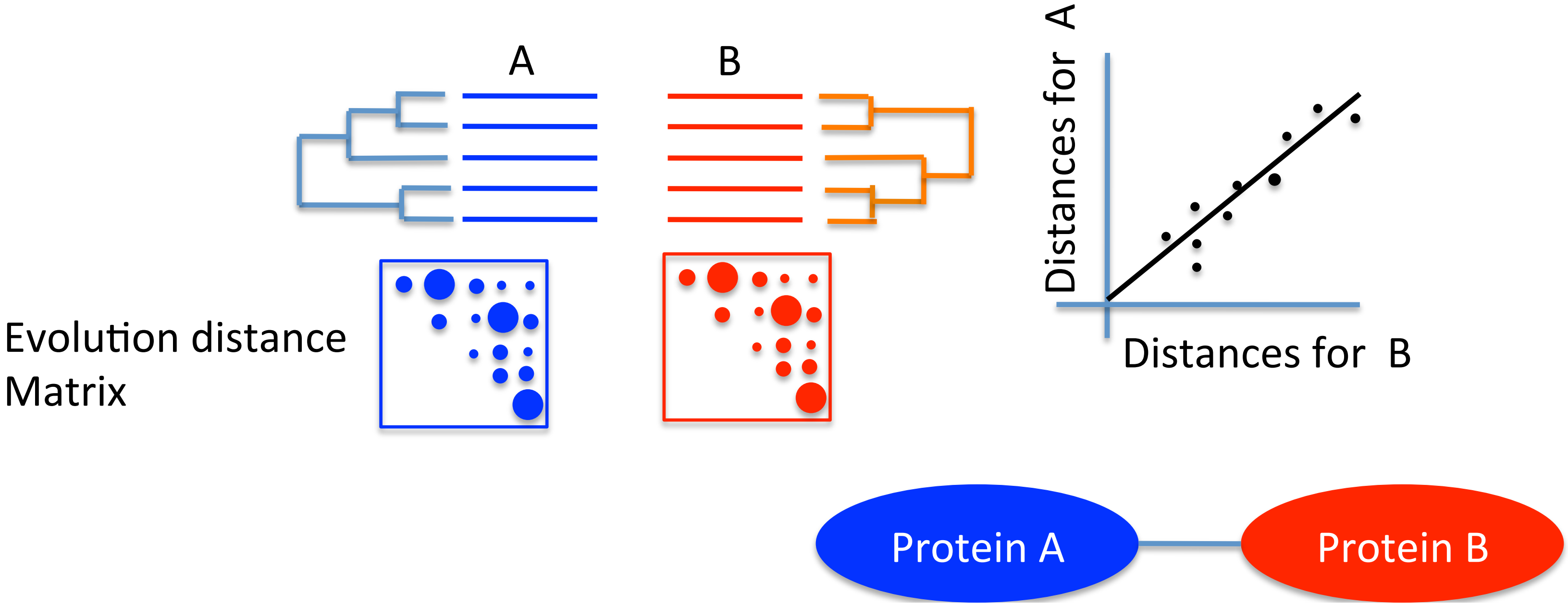


Phylogenetic Profile

	genome 1	genome 2	genome 3	genome 4	genome 5	genome 6	genome 7	genome 8
Gene A	1	1	1	1	1	1	1	1
Gene B	1	1	1	1	1	1	1	1
Gene C	0	0	1	1	1	0	0	0
Gene D	0	0	1	1	1	0	0	0
Gene E	0	0	0	0	0	1	0	1
Gene F	0	0	0	0	0	1	0	1

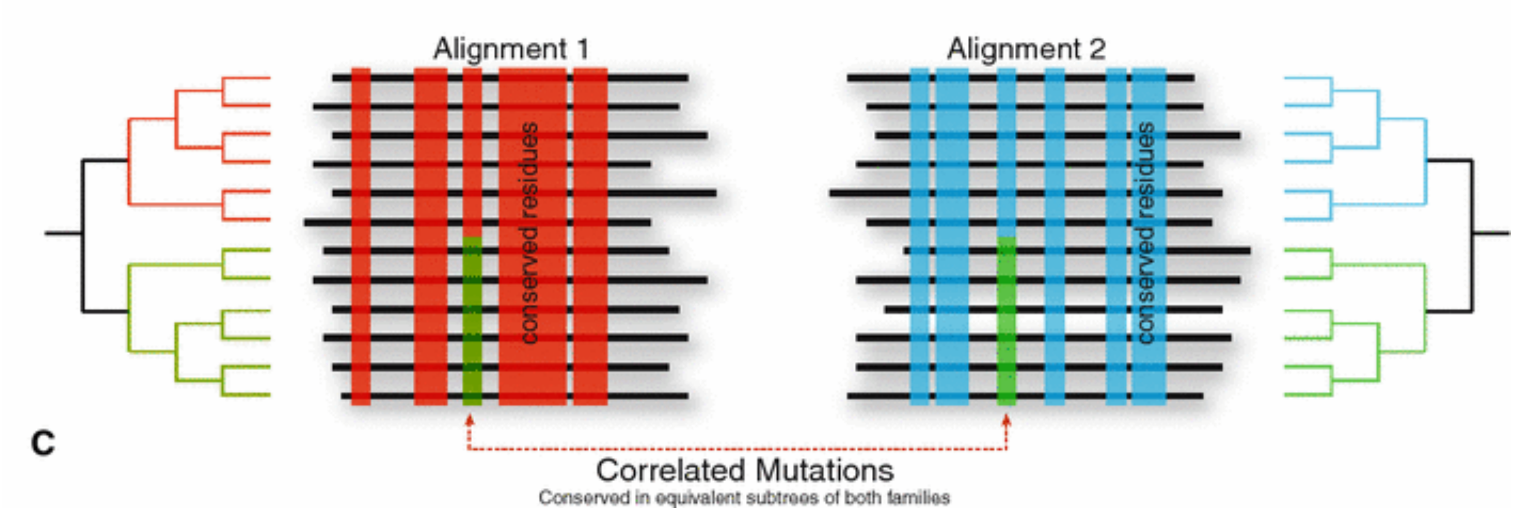
Prediction PPI with sequences

- Co-evolution
  - – as assessed by similarity of phylogenetic trees. “mirrortree” method compares the distance matrices for generating trees;



Prediction PPI with sequences

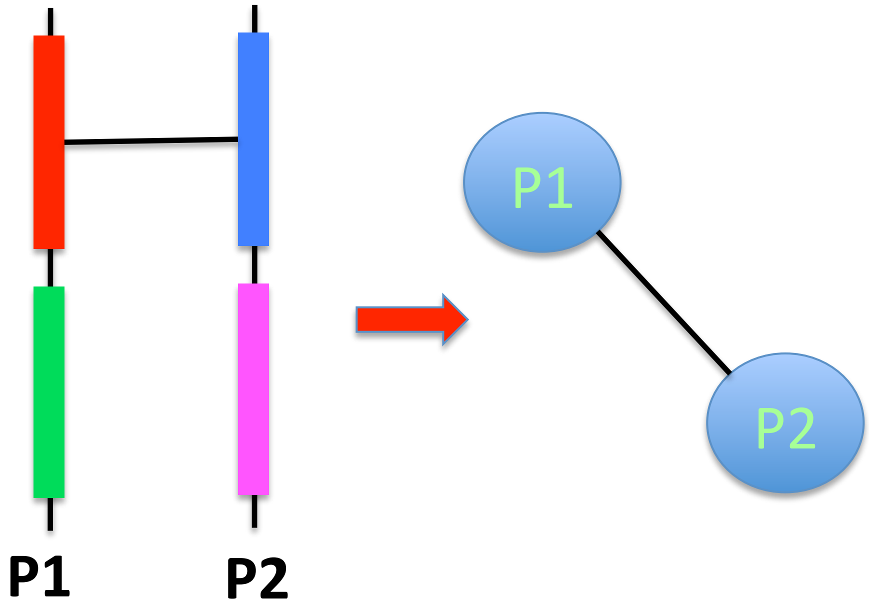
- Correlated mutations
  - the idea is that interacting positions on different proteins should co- evolve so as to maintain the interface. Look for correlation between sequence changes at one position and those at another position in a multiple sequence alignment.

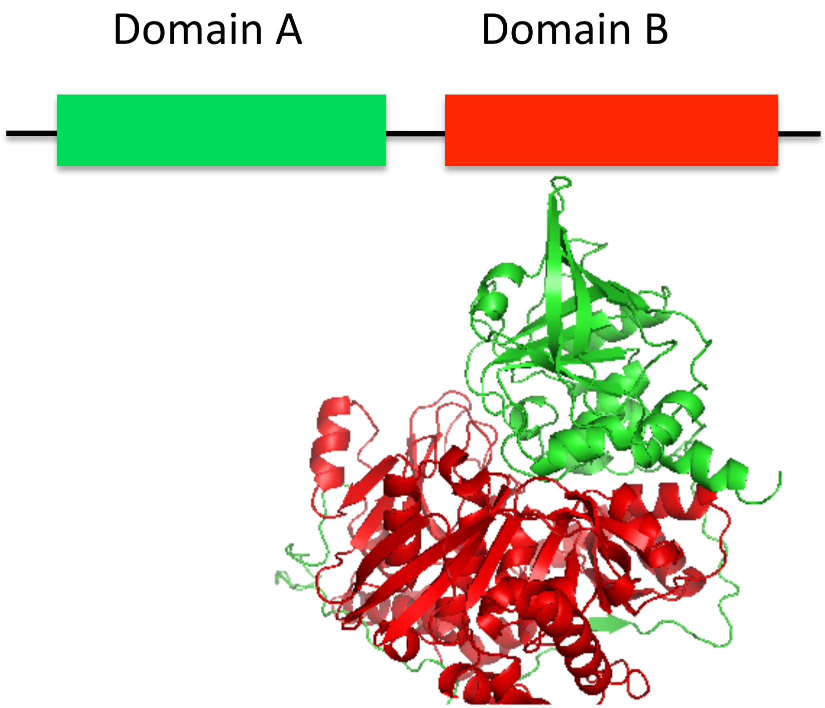


Süel et al. (2002) Nature Strut. Bio.
Pazos & Valencia (2002) Proteins

Prediction PPI with Sequence

- Domain interaction, similar to Domain Pair Verification (*DPV*)
- If two domains have an interaction, any two proteins that have those two domains also have interactions.
- Protein 3D structures can provide the atomic detains for protein interactions.
- The solved structures most are a single domain instead of a full length protein.



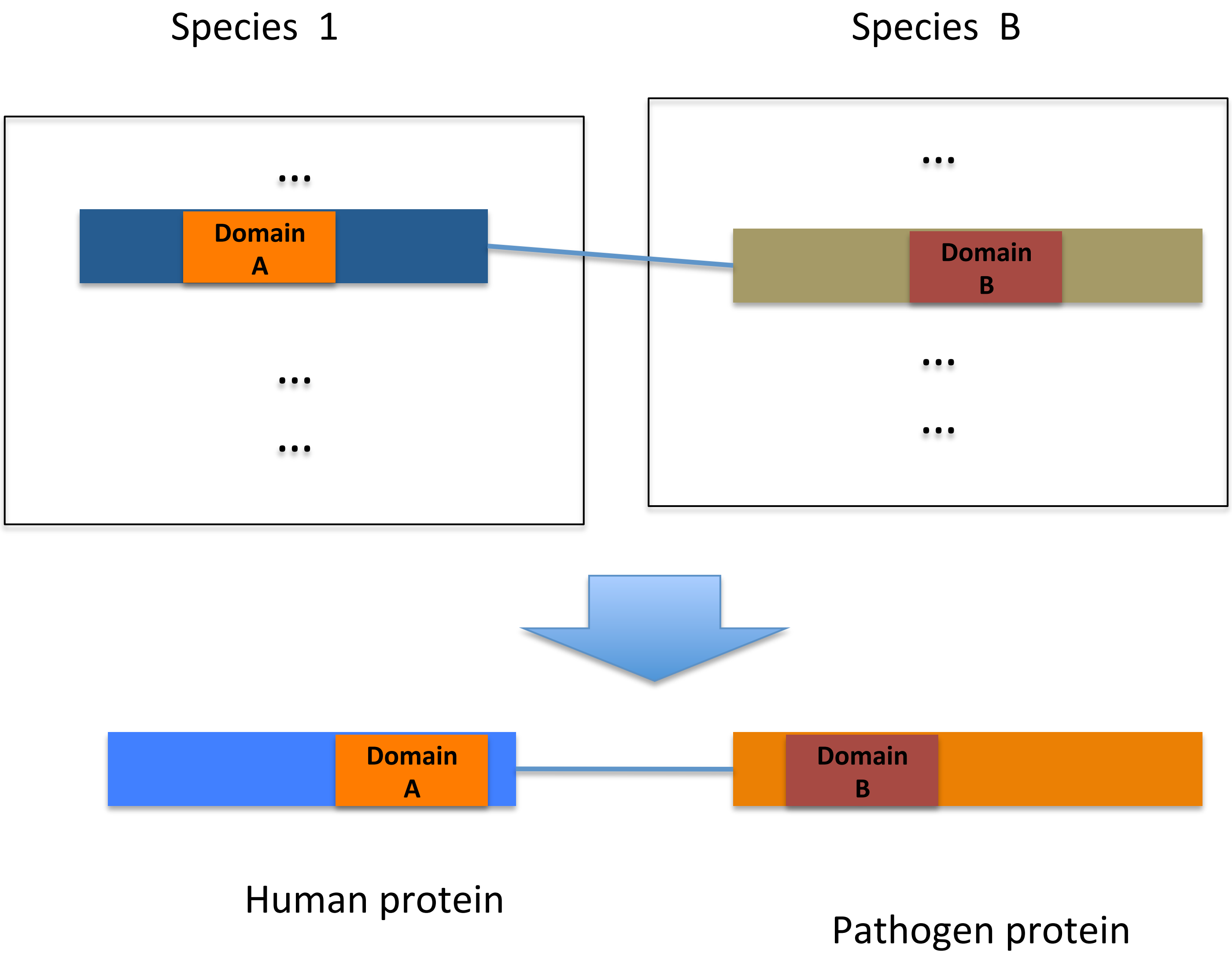


prediction of host-pathogen PPI

- *Plasmodium falciparum* is responsible for the most severe form of malaria.
- Host-pathogen PPs play a vital role in initiating infection.
- Integrate intra-species PPI datasets with protein–domain profiles to predict host-pathogen PPI networks

Dyer et al. (2007) Bioinformatics 12(13) i159

Prediction of Pathogen-Host PPI



Species 1

Species B

…

Domain A

…

…

…

Domain B

…

…

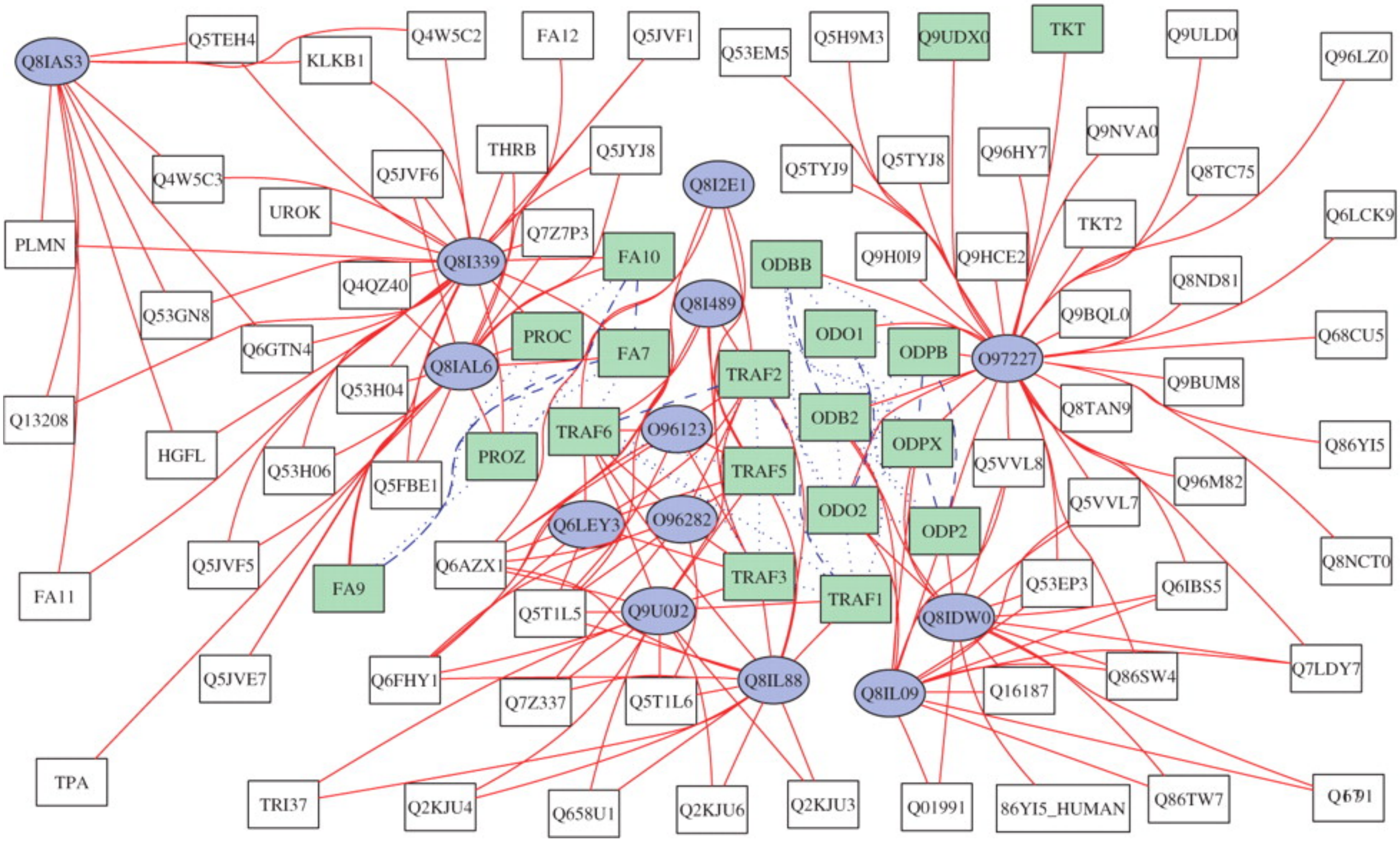
Domain A

Domain B

Human protein

Pathogen protein

prediction of host-pathogen PPI



Dyer et al. (2007) Bioinformatics 12(13) i159

Prediction PPI with sequences

- Problems: they need lots of sequences, and the methods are very sensitive to the alignment method we used.

Web tools for PPI prediction with sequences

- AllFUSE (Enright *et al. 2001*, Gene fusions, http://www.ebi.ac.uk/research/cgg/allfuse/)
- STRING (Snel *et al.* 2000, Gene Co-Localization, gene-fusion, phylogenetic profiles, http://www.bork.embl-heidelberg.de/STRING/)
- WIT (Overbeek *et al.* 2000, Orthology/phylogenetic profiles/gene co-localization, http://wit.mcs.anl.gov/WIT2/)
- Predictome (Mellor *et al.* 2002, Gene Co-Localization, gene-fusion, phylogenetic profiles, http://predictome.bu.edu/)
- COGs (Tatusov *et al.* 1997, Orthology/phylogenetic profiles, http://www.ncbi.nlm.nih.gov/COG/)