

POSTGENOMIC ANALYSIS  
PROTEIN-PROTEIN INTERACTIONS  
FUNCTIONAL PREDICTIONS

Ioannis Xenarios, Lukasz Salwinski,  
Joyce Duan, Charlotte Deane

# OUTLINE

Protein-Protein interactions what is known experimentally

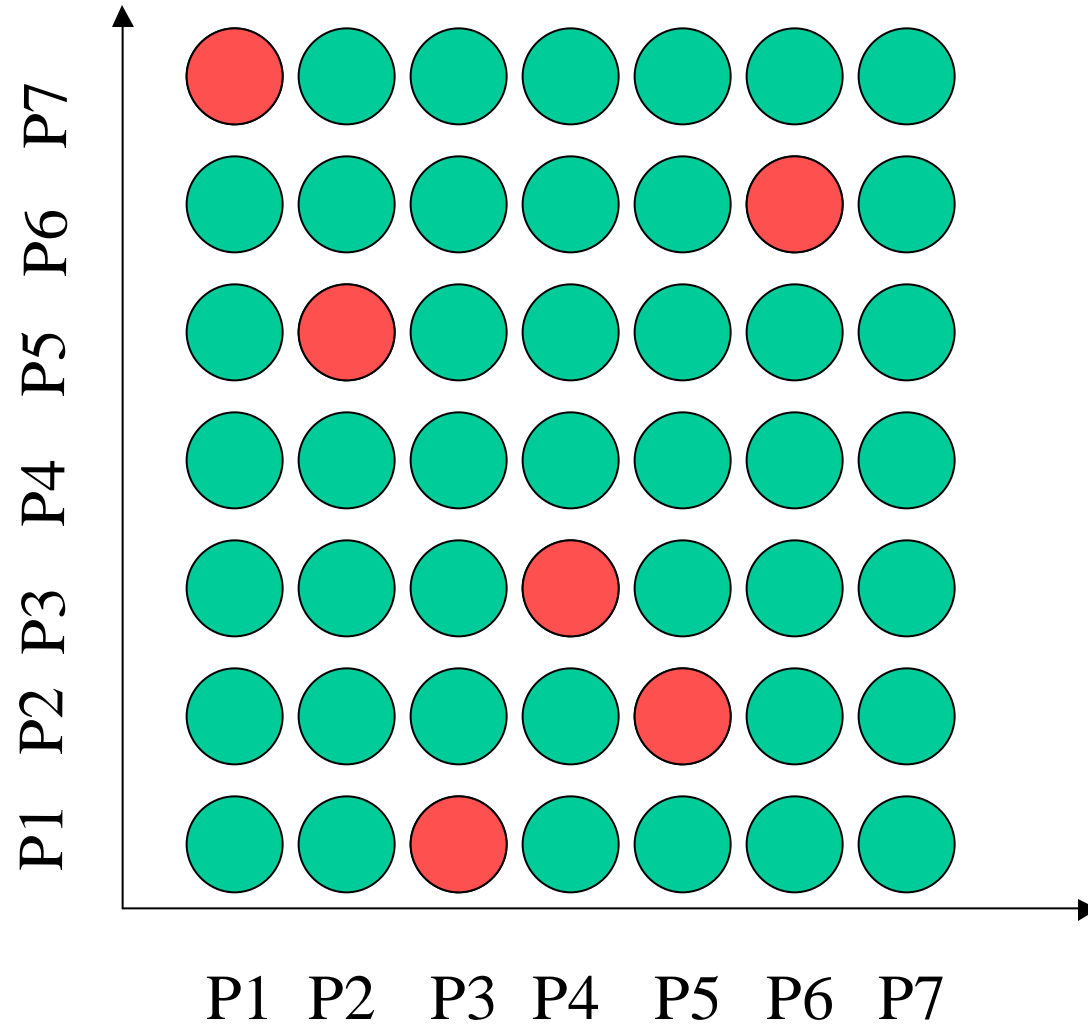
Methods to analyze large scale protein interaction dataset  
(and assess their validity)

Inferring function using genomic information  
(more by Mike Thompson Protein Pathways)

Protein interactions dataset and graph theory

# PROTEIN INTERACTIONS IN A SIMPLE GENOME

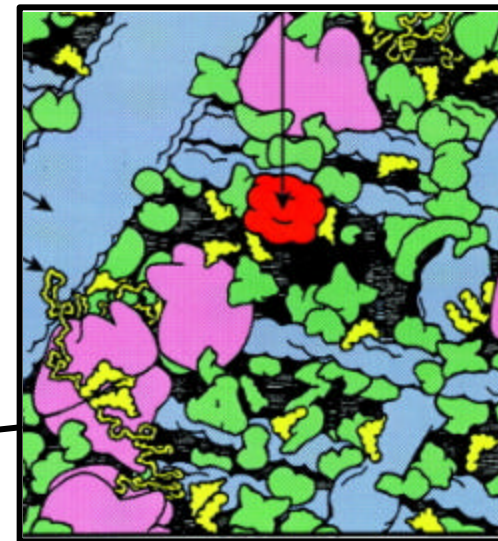
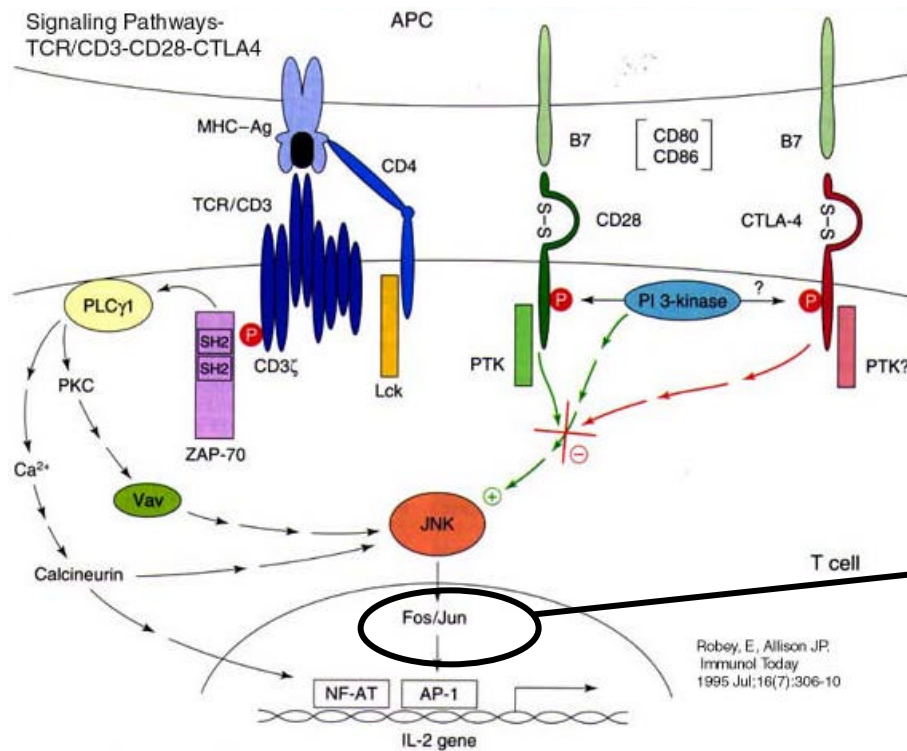
## A NOT SIMPLE SPACE OF INTERACTION TO ANALYZE



49 Potential interactions

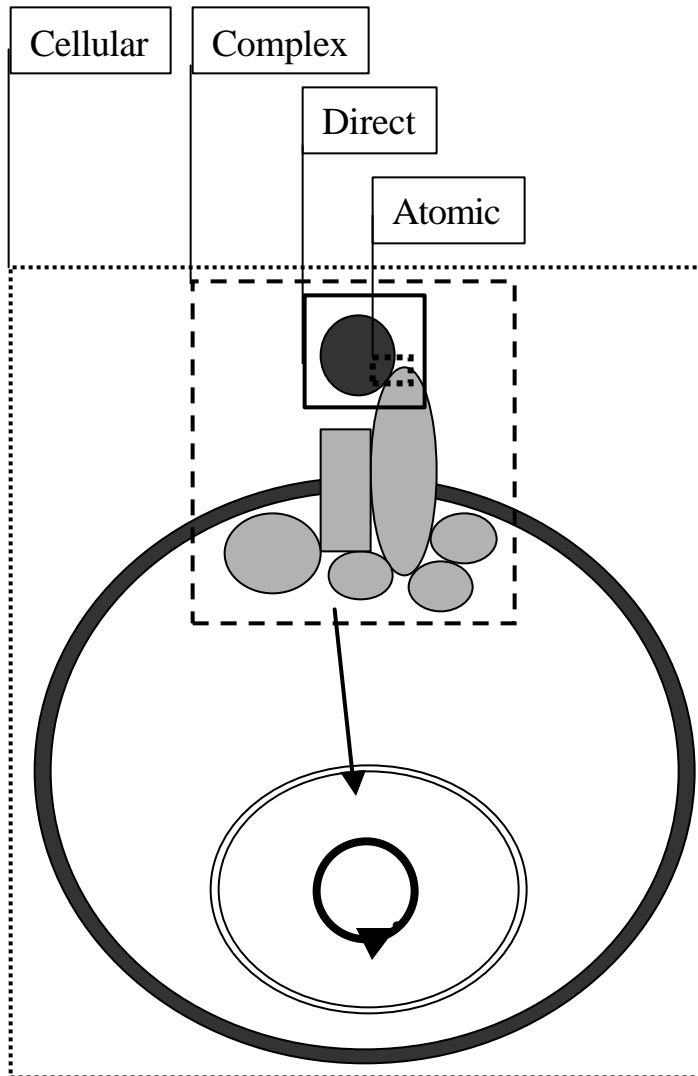
# PROTEIN-PROTEIN INTERACTIONS

Protein-protein interactions are essential to every aspect of life:  
From biochemical pathways, cell cycle,  
cell signaling, cell maintenance



Minton JBC 2001 Apr; (276) 10557-80  
adapted from David Goodsell (Machinery of life)

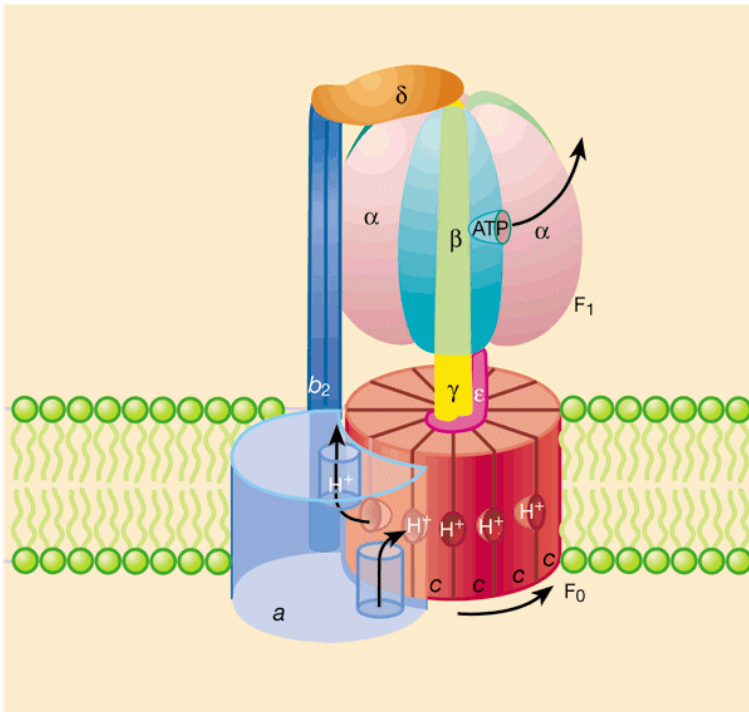
# METHODS TO DETECT PROTEIN-PROTEIN INTERACTIONS



	Level of observation			
	Atomic interactions	Direct interactions	Complex interactions	Cellular interactions
X-ray				
Competition binding				
Gel retardation assays				
ELISA				
Gel retardation assays				
Two hybrid test				
Affinity Column				
BIAcore sensor chip/plasmon resonance				
Electron Paramagnetic Resonance				
Gel filtration chromatography				
Mass spectrometric screening				
Cross-linking				
Co-immunoprecipitation				
Co-sedimentation				
Sizing Column				
Sucrose gradient sedimentation				
Copurification				
Electron microscopy				
Native Gel				
Immunoprecipitation				
Immunofluorescence				
Immunolocalization				
Immunostaining				
FRET Analysis				
Monoclonal antibody blockade				
Interaction adhesion assay				
Knock-out				
Antisense				
Transient coexpression				

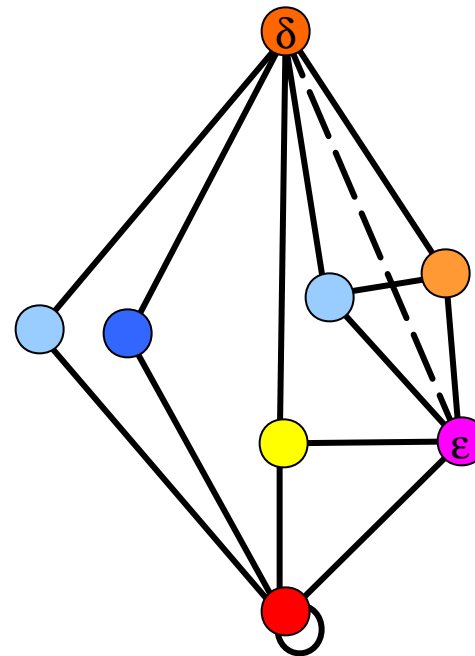
# HOW TO REPRESENT PROTEIN-PROTEIN INFORMATION IN DIP AND HOW MUCH SHOULD WE TRUST THE DATA?

ATP Synthase



Paul D. Boyer Nature (1999)

ATP Synthase in DIP



———— XRay structure Stock et al. Science (1999)

- - - - - Two-hybrid assay

Moritani C et al. Biochim Biophys (1996)

# PROTEIN-PROTEIN INTERACTION QUERYING THE DIP

<http://dip.doe-mbi.ucla.edu>

The screenshot displays the DIP database interface in a Mozilla browser window. The main page shows search results for the query 'actin'. A detailed view of a DIP node (DIP:310N) is shown, along with a detailed view of a DIP link (DIP:1143E).

**Database of Interacting Proteins**

**TEXT SEARCH RESULTS**

Interaction	DIP Node	Links	PIR	SWISSPROT	GENBANK
DIP:1175E	DIP:310N	→	ATRY	ACT_YEAST	gi:71632
DIP:973N	DIP:310N	→	S63211	SLA2_YEAST	gi:21312
DIP:310N	DIP:310N	→	ATRY	ACT_YEAST	gi:71632
DIP:1179E	DIP:225N	→	A46584		gi:53994
DIP:2454E	DIP:310N	→	ATBY	ACT_YEAST	gi:71632
DIP:809N	DIP:310N	→	S47005	GCS1_YEAST	gi:10709
DIP:310N	DIP:310N	→	ATBY	ACT_YEAST	gi:71632
DIP:2455E	DIP:1144N	→	S61023		gi:21322
DIP:310N	DIP:310N	→	ATRY	ACT_YEAST	gi:71632
DIP:10677E	DIP:2250N	→	EFBY1A	EPIA_YEAST	gi:12300
DIP:310N	DIP:310N	→	ATRY	ACT_YEAST	gi:71632
DIP:2456E	DIP:712N	→	S28394	ST26_YEAST	gi:32003
DIP:310N	DIP:310N	→	ATBY	ACT_YEAST	gi:71632
DIP:2201N	DIP:310N	→	S48385	TPM2_YEAST	gi:62695
DIP:310N	DIP:310N	→	ATRY	ACT_YEAST	gi:71632
DIP:2458E	DIP:2202N	→	S64375	TRF1_YEAST	gi:21319
DIP:2692N	DIP:310N	→	A32183	TPM1_YEAST	gi:13610
DIP:310N	DIP:310N	→	ATRY	ACT_YEAST	gi:71632
DIP:5121N	DIP:310N	→	S50445	VACS_YEAST	gi:1077594
DIP:310N	DIP:310N	→	ATBY	ACT_YEAST	gi:71632
DIP:1140E	DIP:310N	→	ATBY	ACT_YEAST	gi:71632
DIP:890N	DIP:310N	→	S54468	AIP1_YEAST	gi:1079591
DIP:4115N	DIP:310N	→	S67137	ESAL_YEAST	gi:2132098
DIP:10543E	DIP:310N	→	ATBY	ACT_YEAST	gi:71632

**DIP NODE**

**DIP:310N** PIR [ATRY](#) SwissProt [ACT\\_YEAST](#) GenBank [gi:71632](#)

Name/Description actin

CrossRef YPD: [ACT1](#) SGD: [BND1](#) MIPS: [YEL282](#)

Organism *Saccharomyces cerevisiae* Localization EC Function

Keywords Methylated amino acid.

**DIP LINK**

**DIP:1143E**

**Protein A** DIP:310N PIR: ATRY SwissProt: ACT\_YEAST GenBank: gi:71632  
Name/Description actin

**Protein B** DIP:46N PIR: A31360 SwissProt: FROP\_YEAST GenBank: gi:83404  
Name/Description profilin

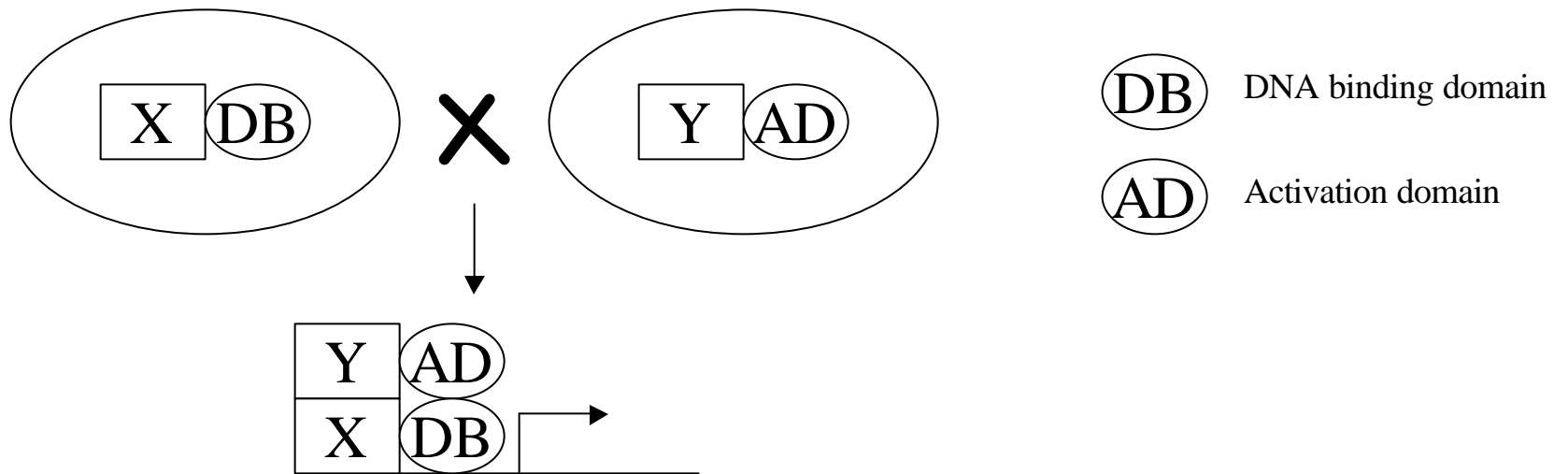
#	Method	Class Details	Source
1	Two hybrid test	---	PMID:7719650
2	Copurification	---	PMID:1606975
3	In vitro binding	---	PMID:9002982
4	Two hybrid test	---	PMID:10681120
5	Two hybrid test	---	PMID:11482916

Document Done (0.665 sec)

Document Done (0.052 sec)

LARGE SCALE  
PROTEIN-PROTEIN  
INTERACTIONS SCREENS

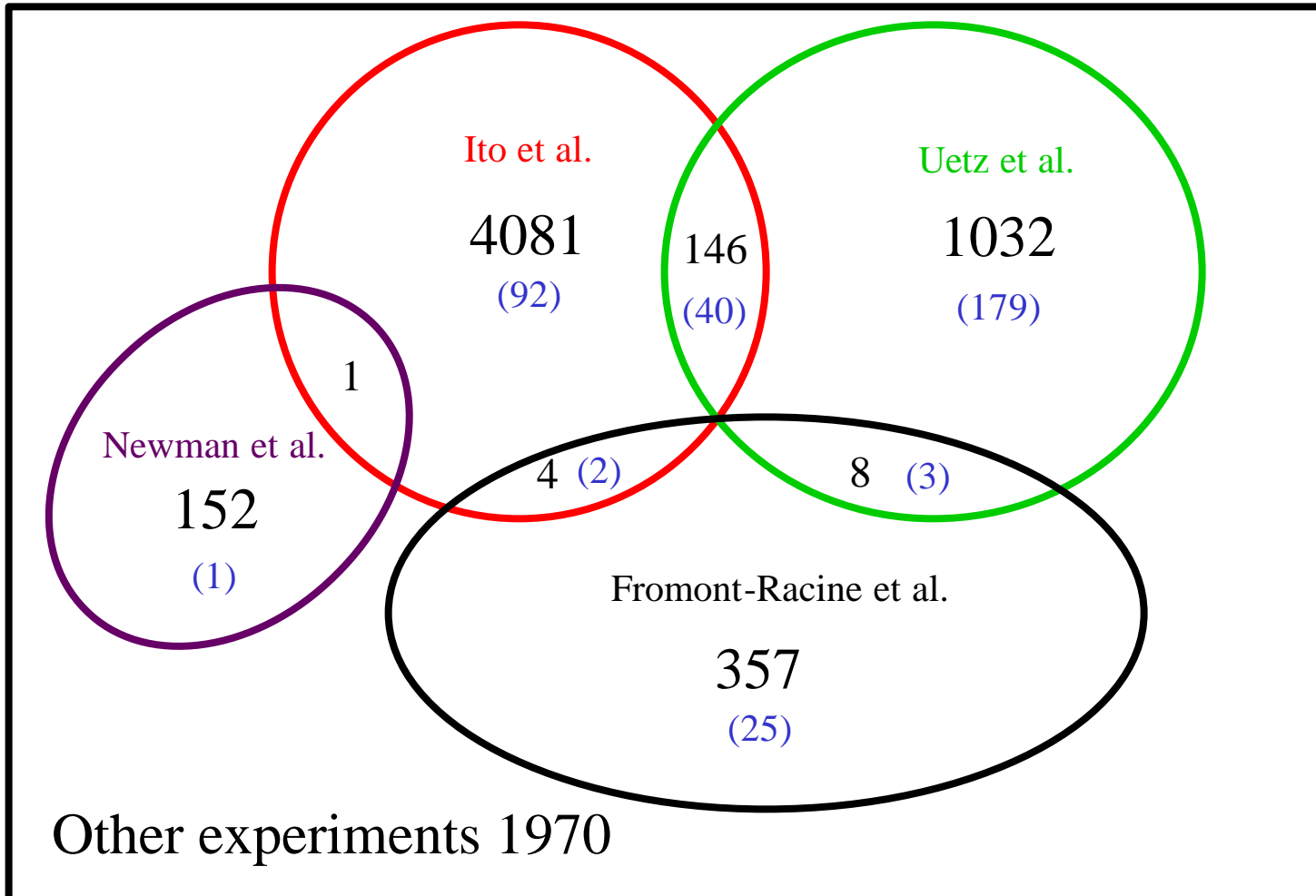
# LARGE SCALE TWO HYBRID APPROACH TO DETECT PROTEIN-PROTEIN INTERACTIONS



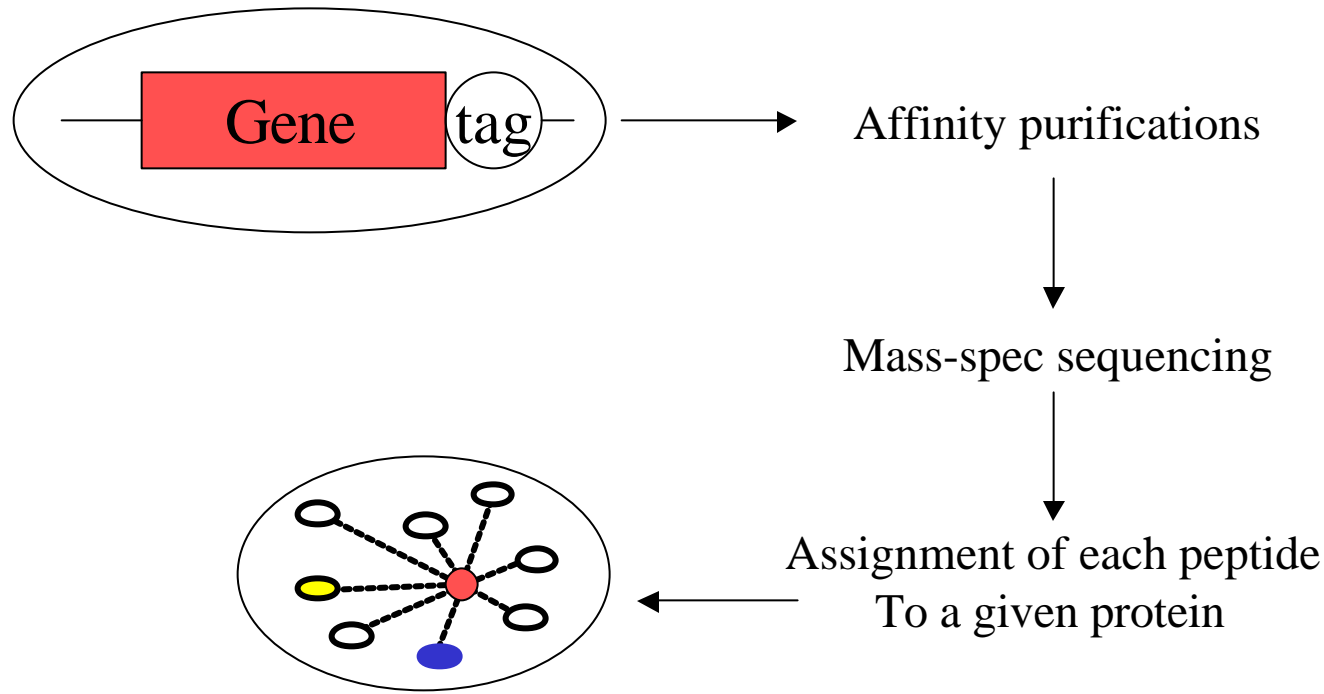
## Problems:

- Over-expression not natural amount of each proteins
- Transcription factor cannot be used (give false positives signals)
- False localization (driven by the nuclear localization signal)
- Wrong orientations or instability of the fused proteins
- Membrane proteins are more complicated to detect

**Lack of OVERLAP**  
**in large scale**  
**genomic protein interaction analysis in yeast**



# LARGE SCALE PURIFICATION AND MASS-SPECTROMETRIC OF YEAST PROTEIN COMPLEXES



## Problems:

Number of purifications steps (as well as fractionation)

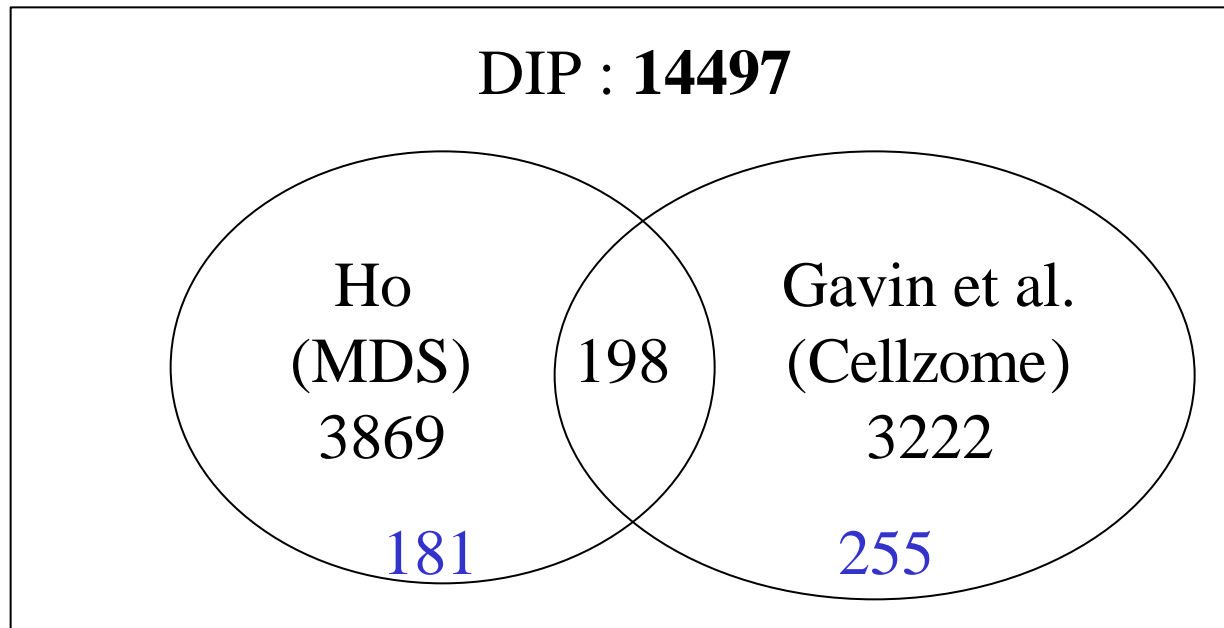
Protein with high  $k_{\text{off}}$  cannot be detected

For some genes the level of homology is so high that the exact protein can elude the analysis

Proteins can be found in different complexes (results will be an aggregation)

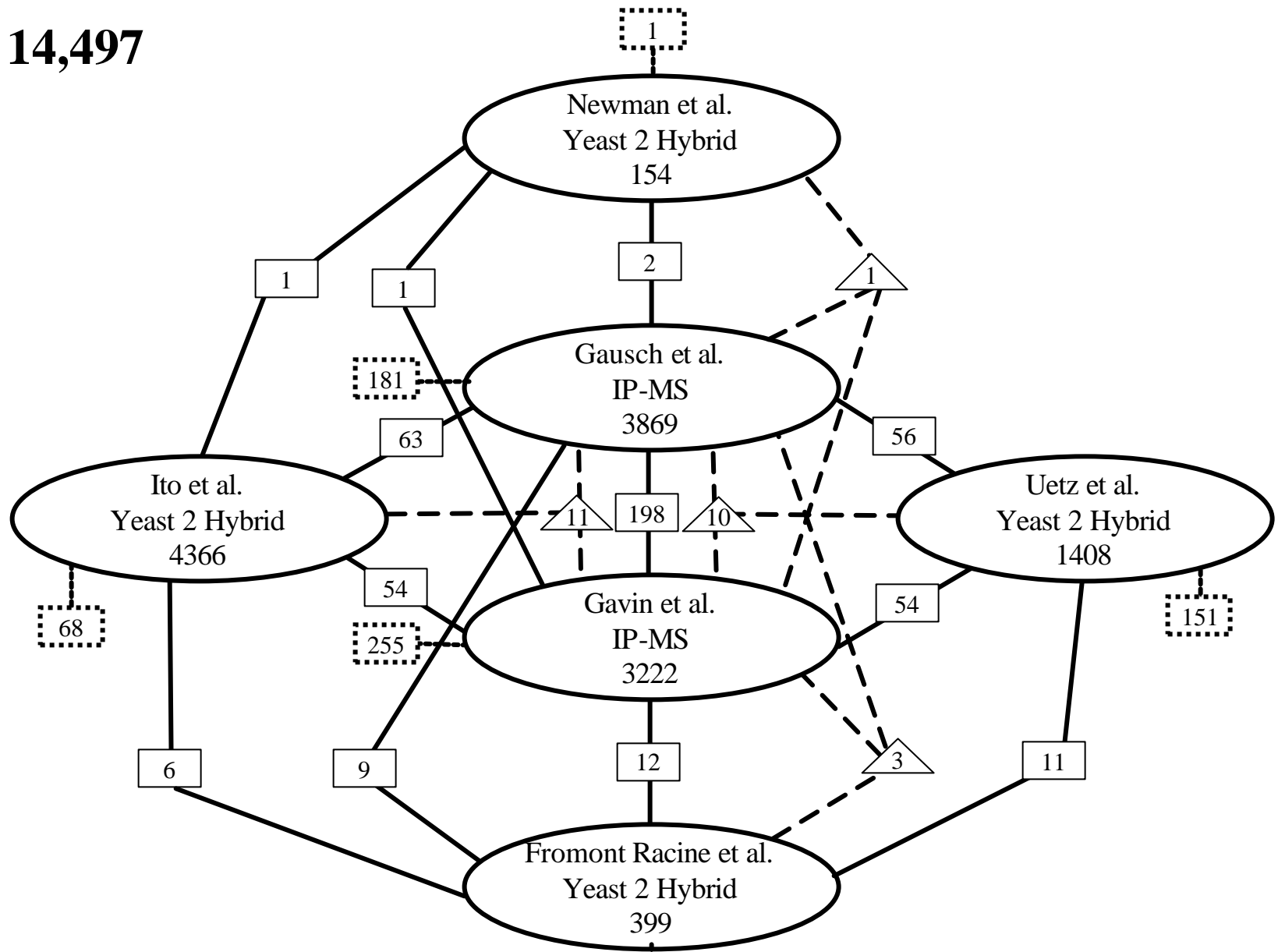
Interactions complexes are not always reciprocal

# LARGE SCALE PURIFICATION AND MASS-SPECTROMETRIC OF YEAST PROTEIN COMPLEXES



NOT A LOT OF OVERLAP AGAIN  
Not better than yeast two hybrid!

**DIP: 14,497**



——— Overlap between any two genome-wide interactions screens  
 - - - Overlap among three genome-wide interactions screens  
 ..... Overlap between any genome-wide screen with non genome-wide interaction screen

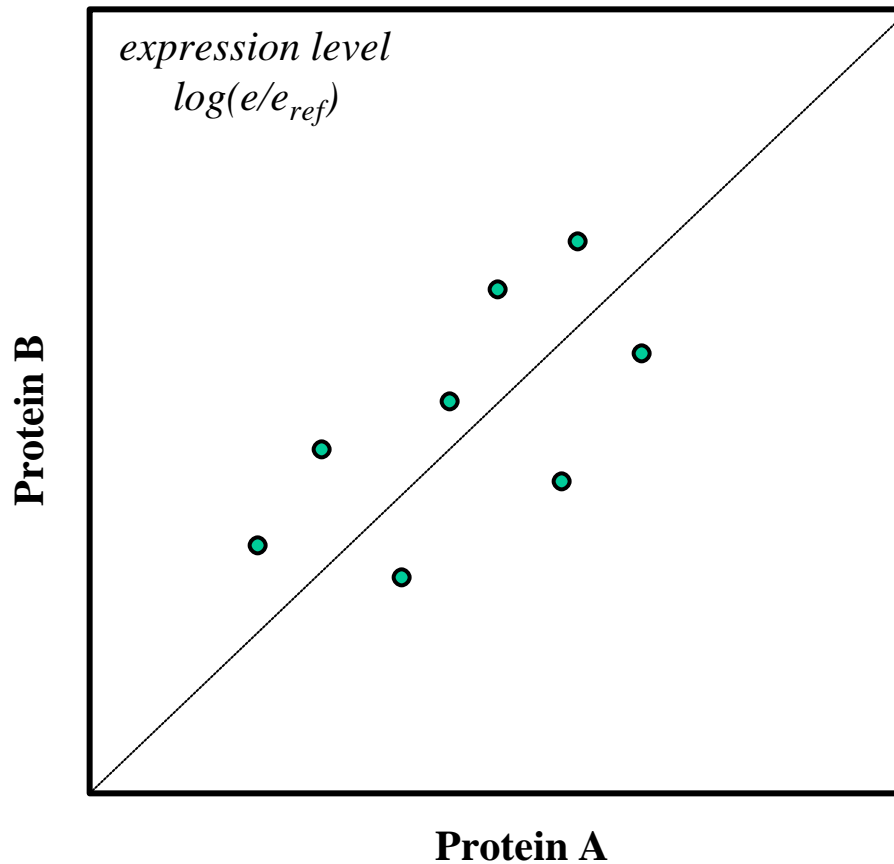
LARGE SCALE PROTEIN  
INTERACTIONS SCREENS  
AND  
VALIDATION METHODS

# Relating gene expression data and protein interaction data

Assumption protein that functions together have same  
Patterns of expression throughout various conditions  
e.g. heat-shock, low carbon, etc (for yeasts)

Can we use the microarrays analysis to determine the  
Fraction of protein-protein interactions that are “correct”

# USING EXPRESSION DATA TO ESTIMATE THE ERROR RATE OF PROTEIN PROTEIN INTERACTIONS DATASET(S)

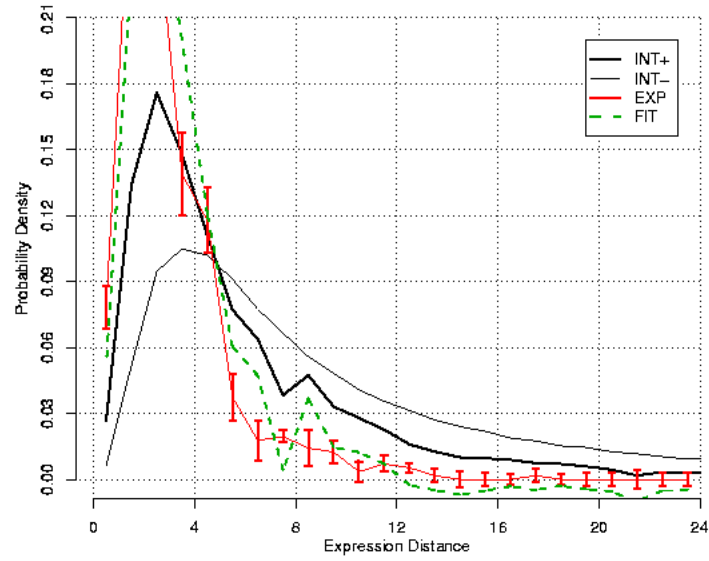


**Euclidean distance**

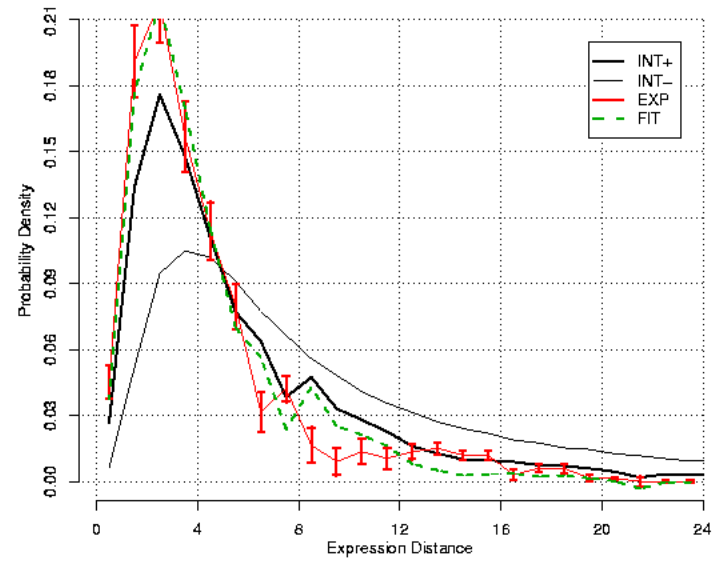
$$d = \sqrt{\sum_i (A_i - B_i)^2}$$

# STABLE COMPLEXES

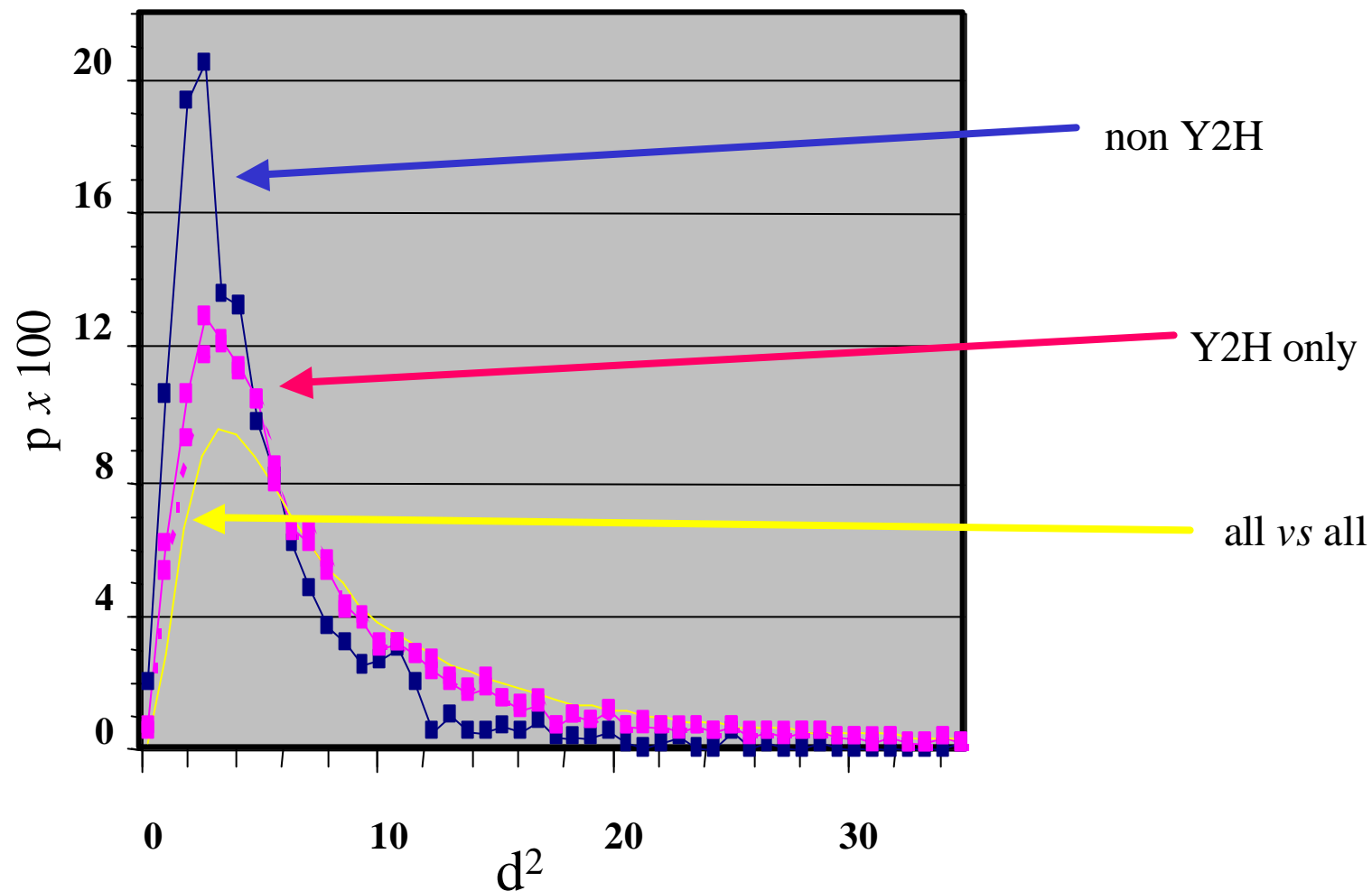
Proteasome



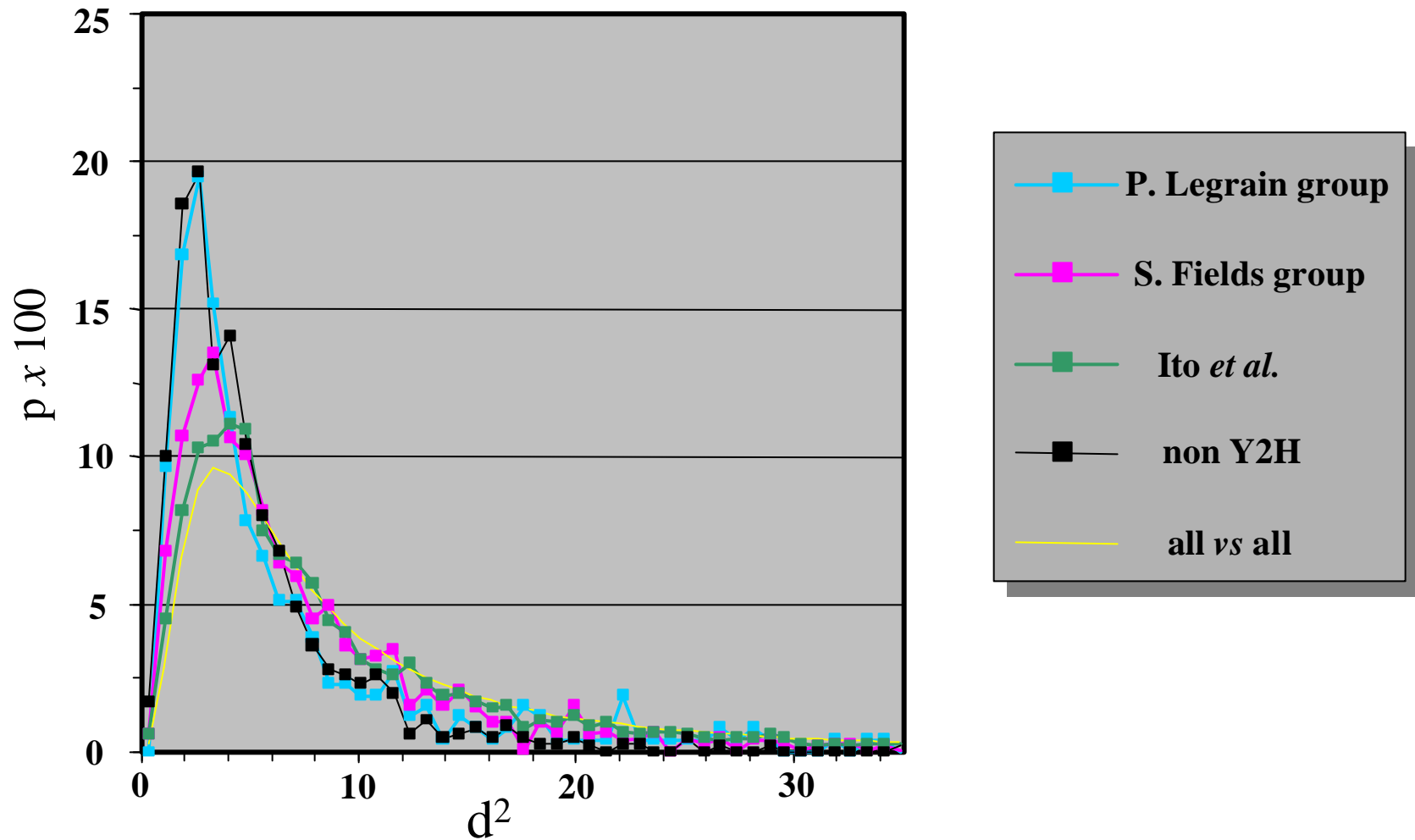
Ribosome (cytoplasmic)



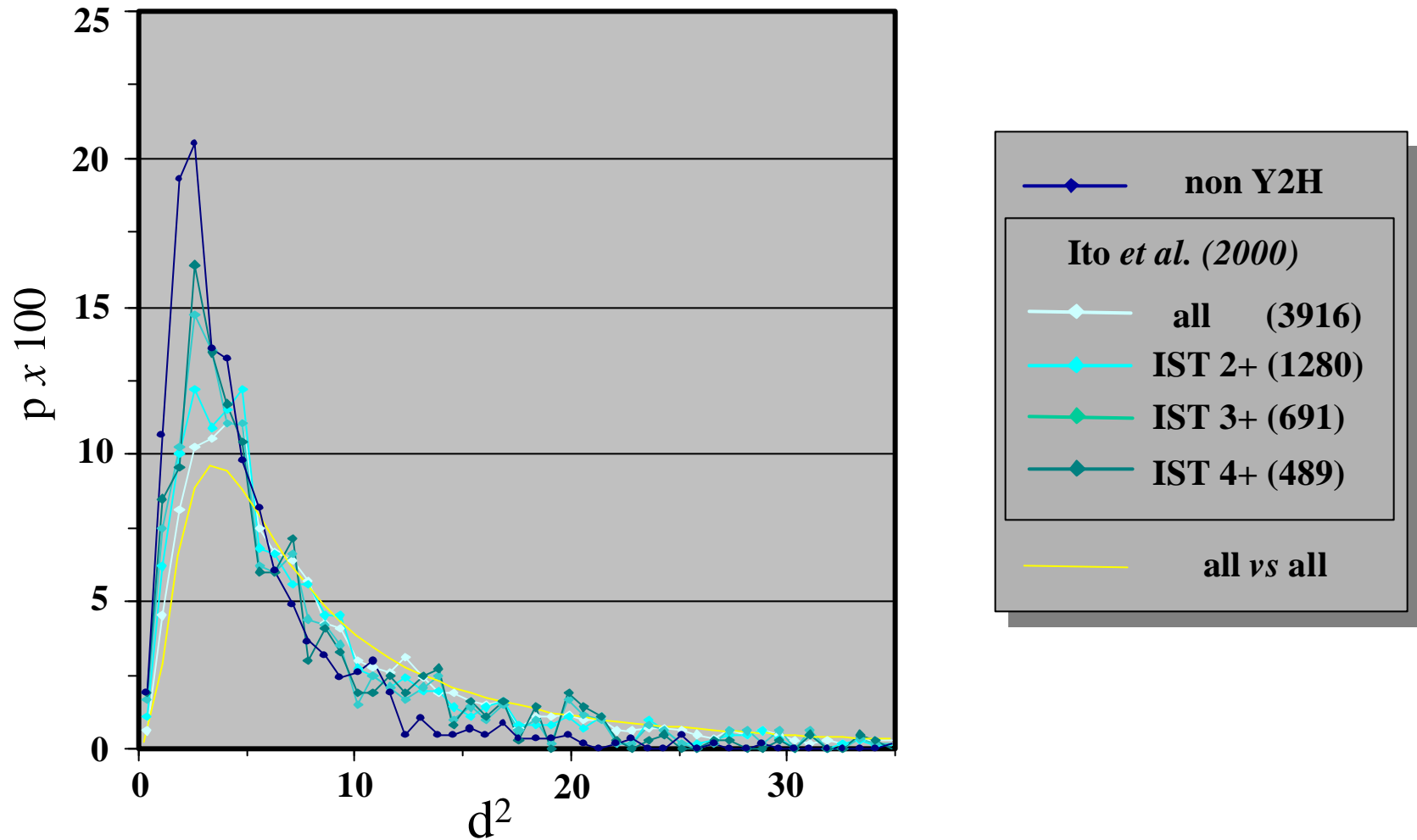
# DIFFERENCE IN DISTANCE MEASURE BETWEEN PROTEIN PROTEIN INTERACTION DETERMINED BY STANDARD METHOD AS COMPARED TO YEAST TWO HYBRID METHOD



# SPLITTING THE YEAST TWO HYBRID DATASET

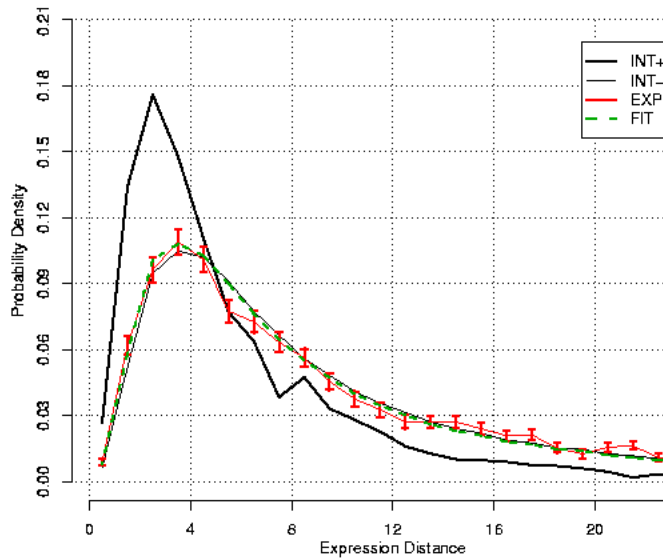


# DEPENDANCE ON THE NUMBER OF IST AND THE QUALITY OF THE DATA ?

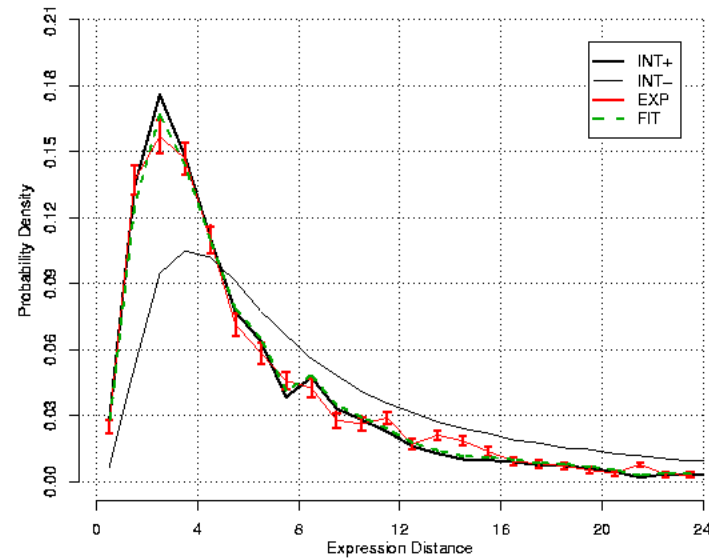


# LARGE SCALE PURIFICATION AND MASS-SPEC ANALYSIS OF YEAST PROTEIN COMPLEXES TALE OF A GOOD - A BAD DATASET

**Ho (MDS) 3869**



**Gausch (Cellzome) 3222**



Fraction of  
'correct interaction'

$7.81 \pm 3.7$

$89.5 \pm 6.4$

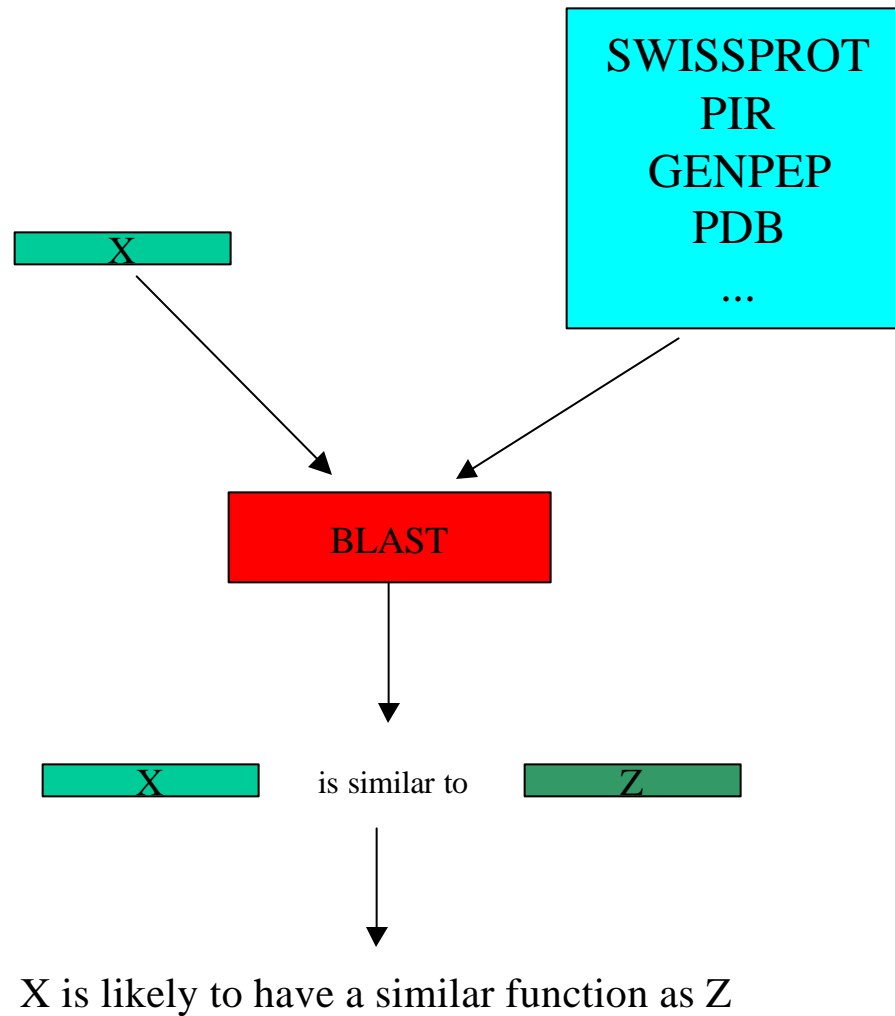
## Fraction of true positives in the different large scale screen

Experimental setup	Number of interactions observed	EPR index
Genome-wide yeast two hybrid	6114	32.3% $\pm$ 3.4%
Ito dataset	4366	19.5% $\pm$ 3.6%
Uetz dataset	1406	48.7% $\pm$ 6.9%
Mass spec MDS	3595	7.8% $\pm$ 3.7%
Mass spec Cellzome	3222	89.5% $\pm$ 6.6%

# PREDICTING PROTEINS THAT FUNCTION TOGETHER

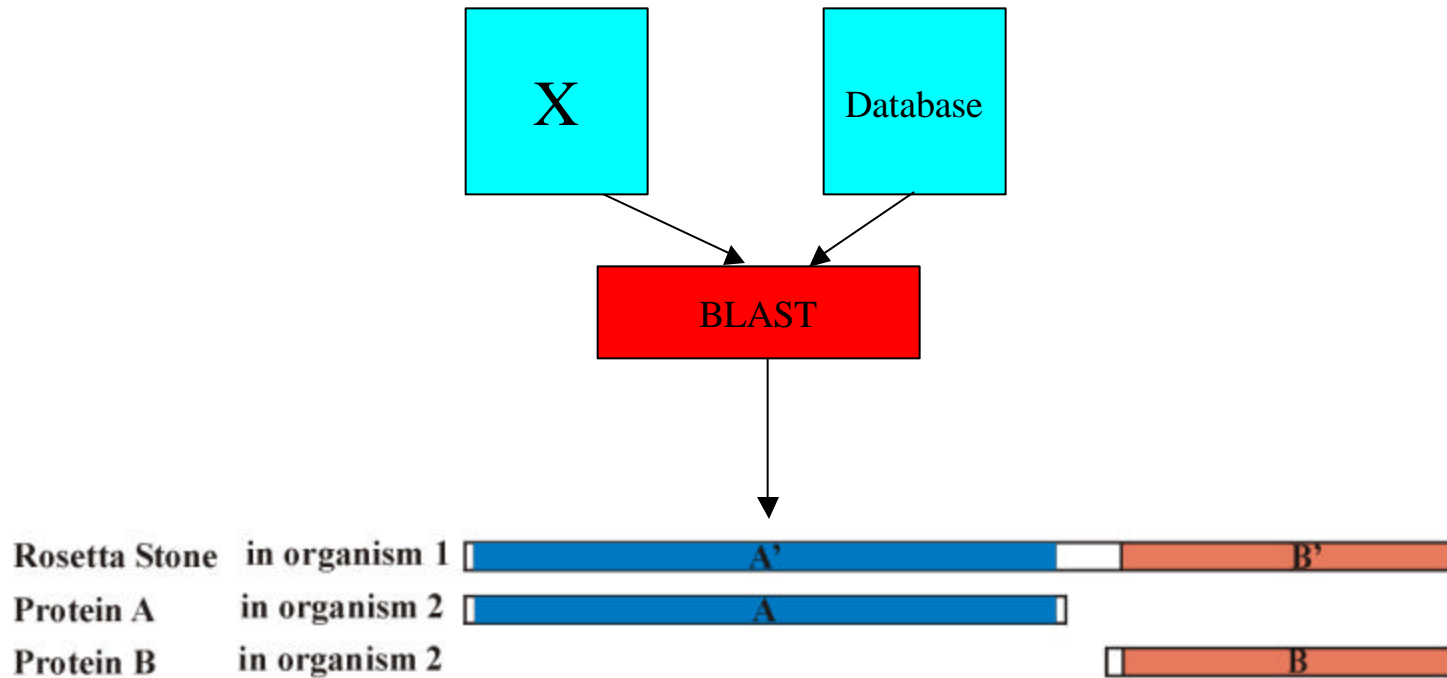
Eisenberg *et al.* 2000  
Nature

# STANDARD HOMOMOLOGY BASED METHOD



# NON-HOMOLOGY BASED METHOD

## (i) Rosetta Stone



Marcotte *et al.* Nature 1999  
Enright *et al.* Nature 1999

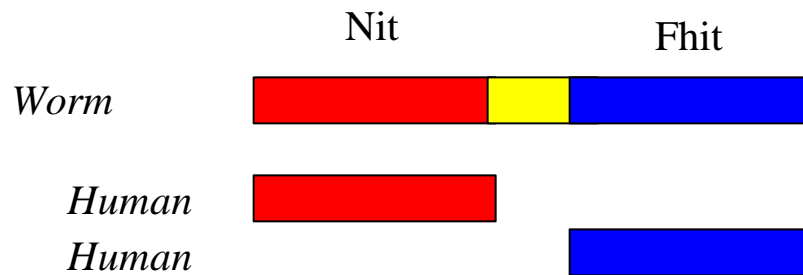
# EXAMPLE OF A ROSETTA STONE PROTEIN

## worm NitFhit

Fhit: fragile Histidine triad associated  
with a wide variety of cancer

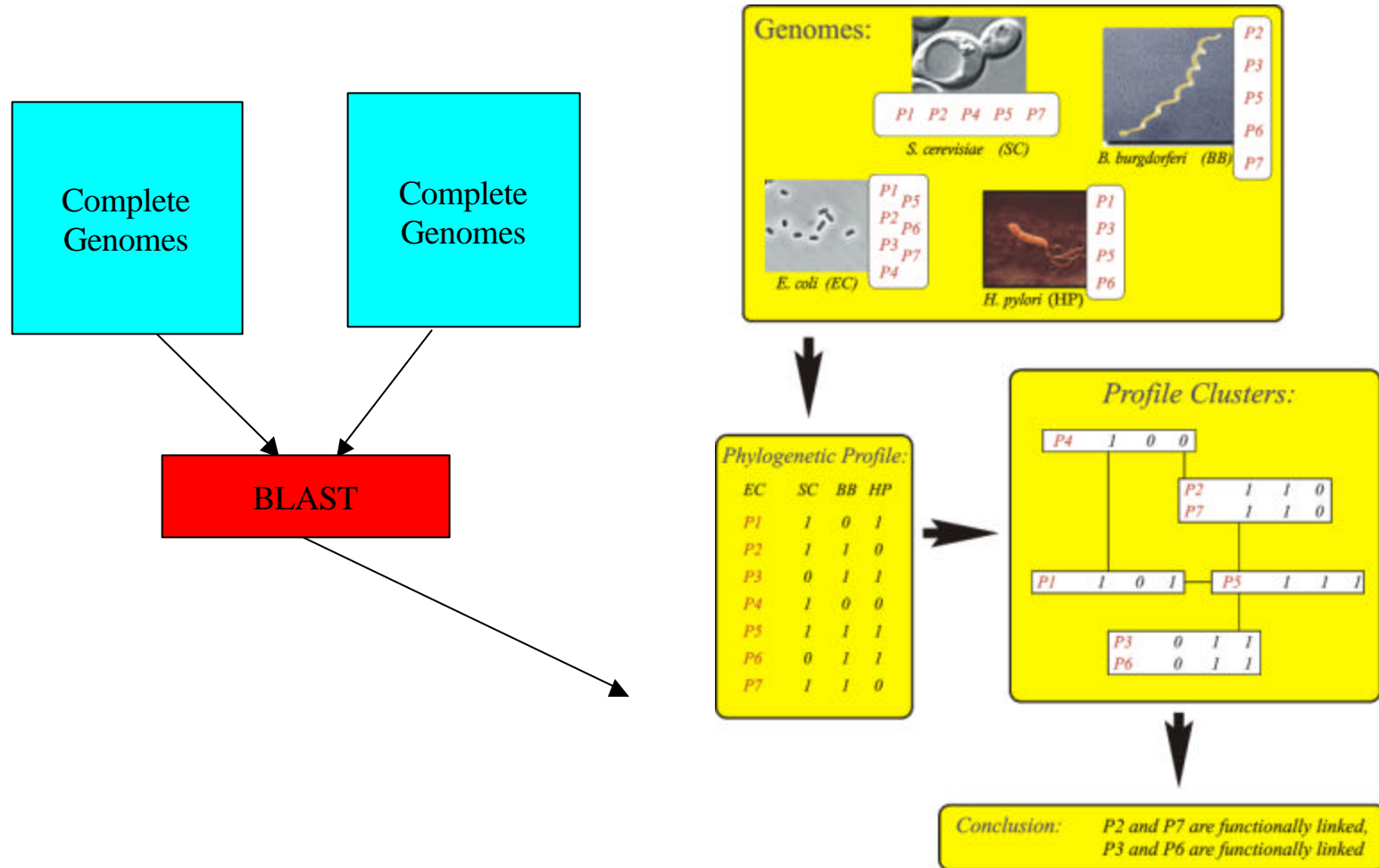


No interacting partners known

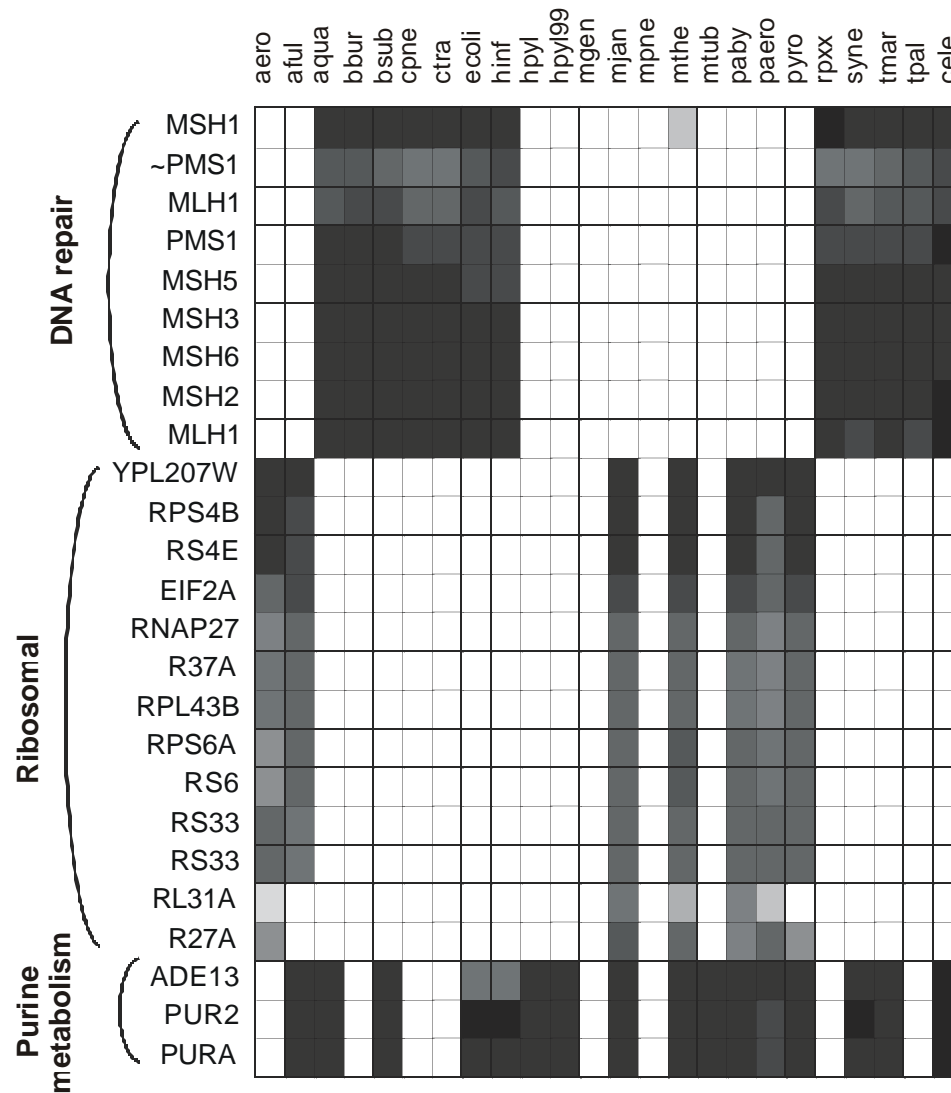


# NON-HOMOLOGY BASED METHOD

## (ii) Phylogenetic profiles

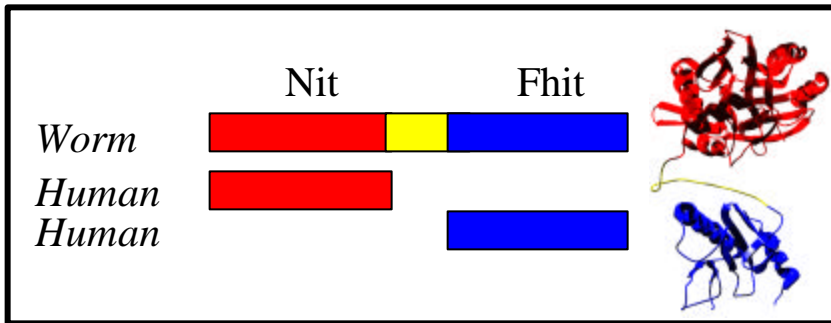


# EXAMPLES OF PHYLOGENETIC PROFILES

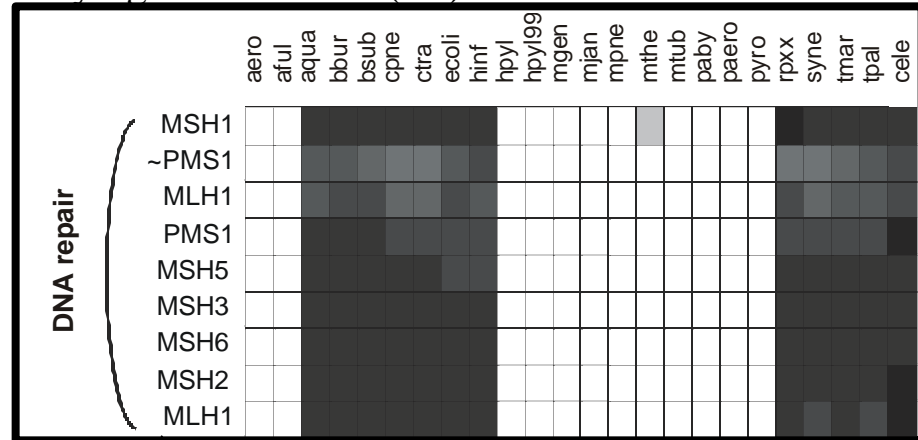


# INTEGRATING FUNCTIONAL LINKAGES AND EXPERIMENTAL PROTEIN INTERACTION DATA

Rosetta Stone (RS)

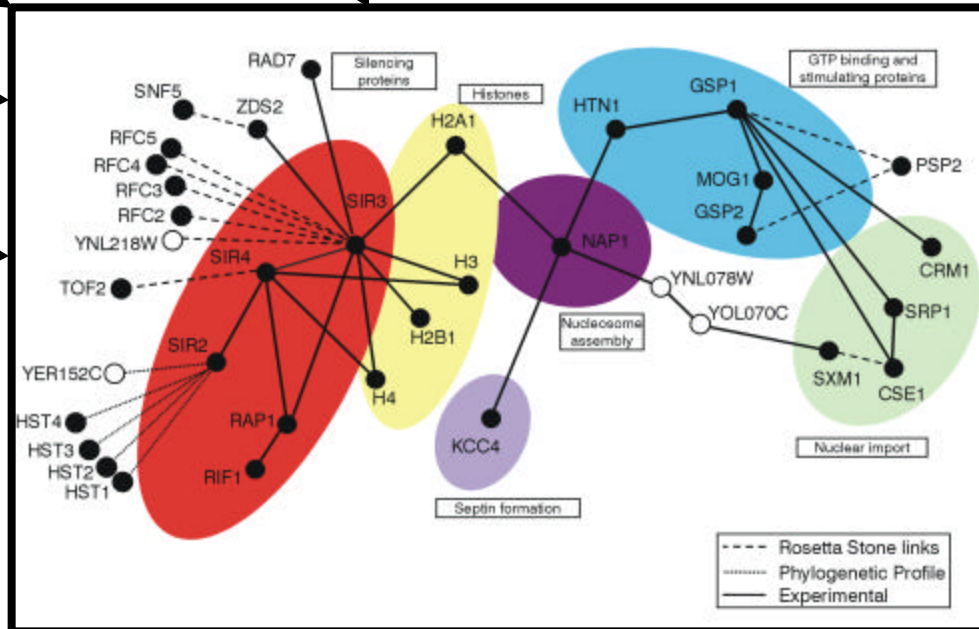


Phylogenetic Profile (PP)



Gene Neighbours (GN)

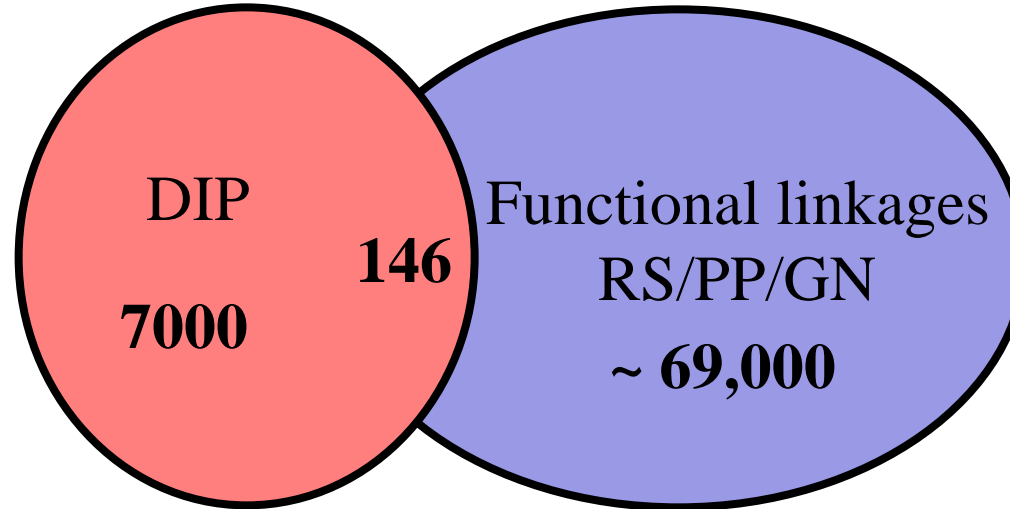
mRNA correlation (CC)



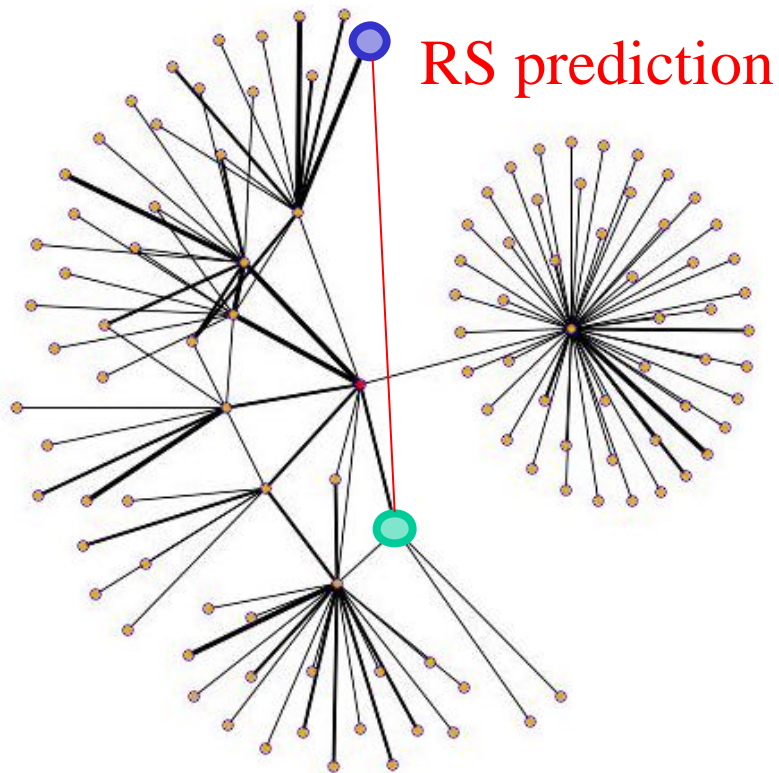
Reviewed in Eisenberg *et al.* Nature 2000

HOW GOOD ARE  
THOSE METHODS  
compared to  
KNOWN  
PROTEIN-PROTEIN  
INTERACTIONS ?

# COMPARING PROTEIN INTERACTIONS AND FUNCTIONAL LINKAGES OF THE YEAST PROTEOME



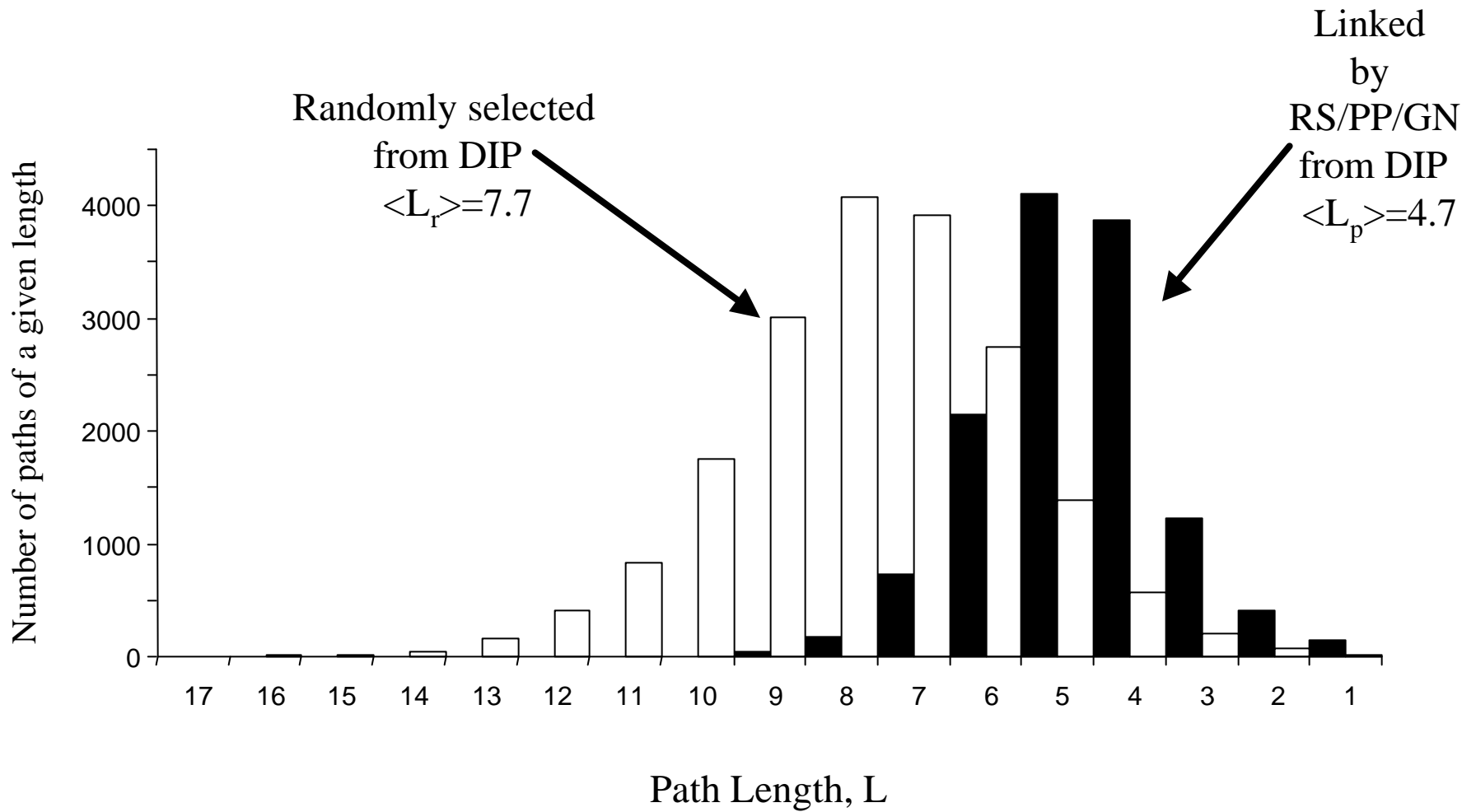
# COMPARING PATH LENGTHS BETWEEN TWO FUNCTIONALLY LINKED PROTEINS USING THE EXPERIMENTAL INTERACTION NETWORK



How many interaction steps  
between  
the blue and the green proteins

# PATH LENGTH

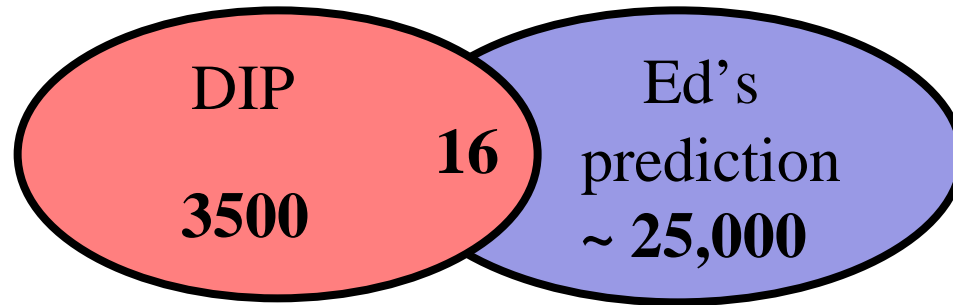
## FUNCTIONALLY LINKED PROTEINS vs RANDOM PAIRS



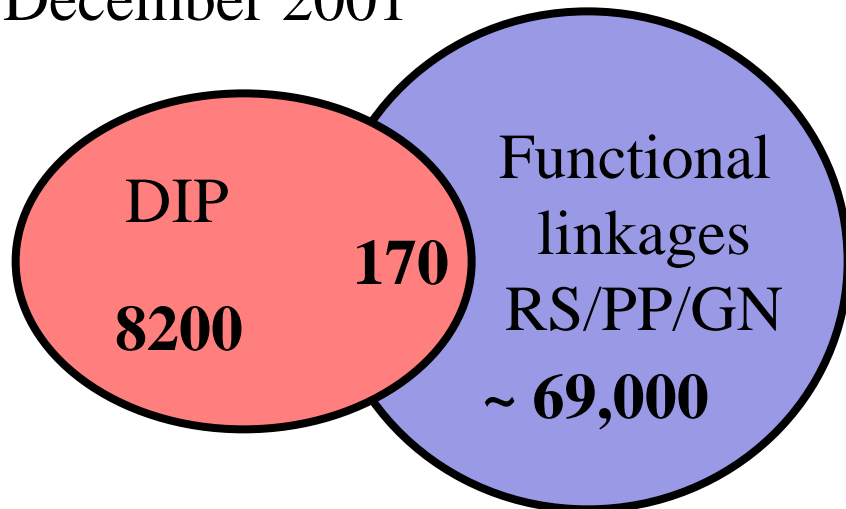


# OVER THE TIME THE NUMBERS OF OVERLAPING INTERACTIONS FOUND INCREASE

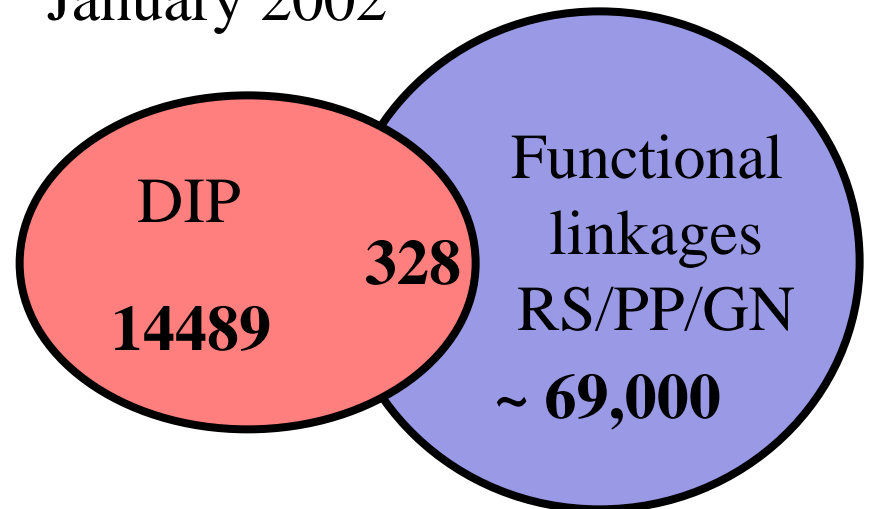
December 2000



December 2001



January 2002



Still the path distributions between predicted linkages  
And randomly selected proteins shows a clear differences

## Prediction methods and EPR index

Prediction method	Number of functionally linked proteins	EPR index
<b>Marcotte 1999</b>		
Rosetta Stone (RS)	1900	29.1% + 7.2%
Phylogenetic profiles (PP)	1963	0% + 5.9%
RS + PP	680	51.8% + 11.7%
<b>De Lisi 2002</b>		
Fusion pairs De Lisi	11840	31.0% + 3.1%
Phylogenetic Profiles	1306	0% + 7.0%
<b>Protein Pathways</b>		
<b>(&gt;0.5 pred accuracy)</b>		
Rosetta Stone, Phylogenetic Profiles, Gene neighbors	63140	29.0% ± 1.8%
Rosetta Stone	36883	35.5% ± 2.2%
Phylogenetic profiles	17116	16.8% + 2.2%
<i>Protein Pathways (PP,RS,GN)</i> <i>(0.1-0.2 prediction)</i>	56443	14.8% + 1.4%

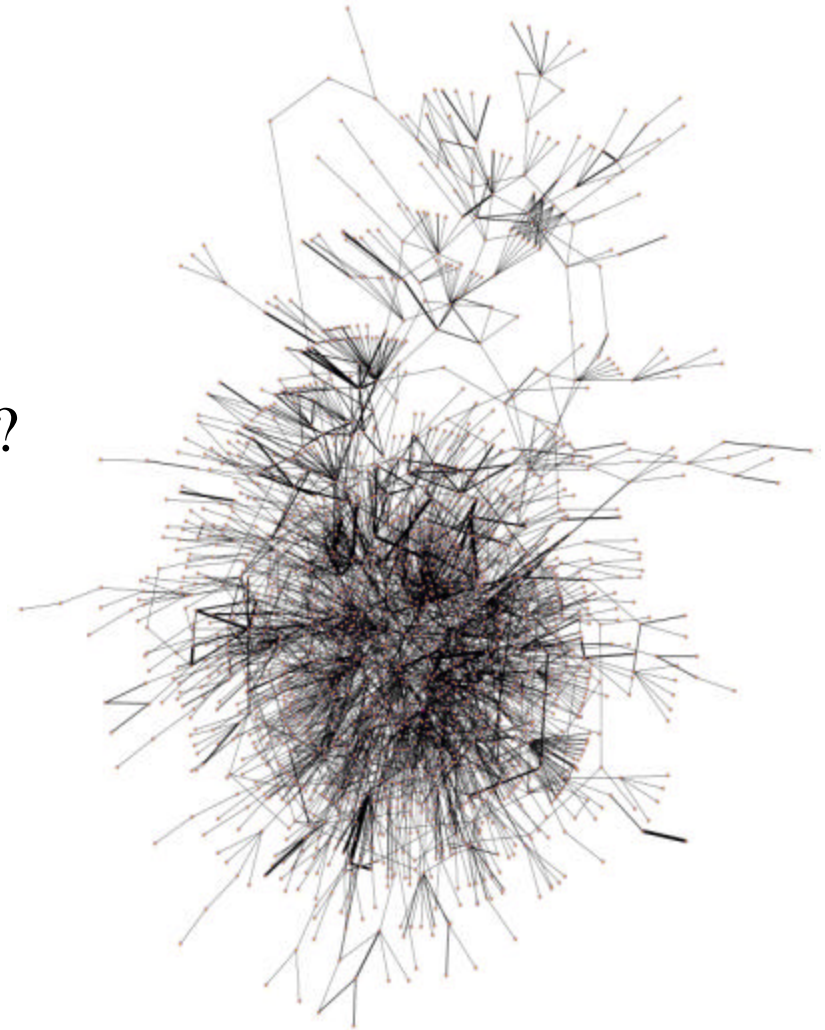
PROTEIN-PROTEIN  
INTERACTIONS  
AND  
GRAPH(s)

...

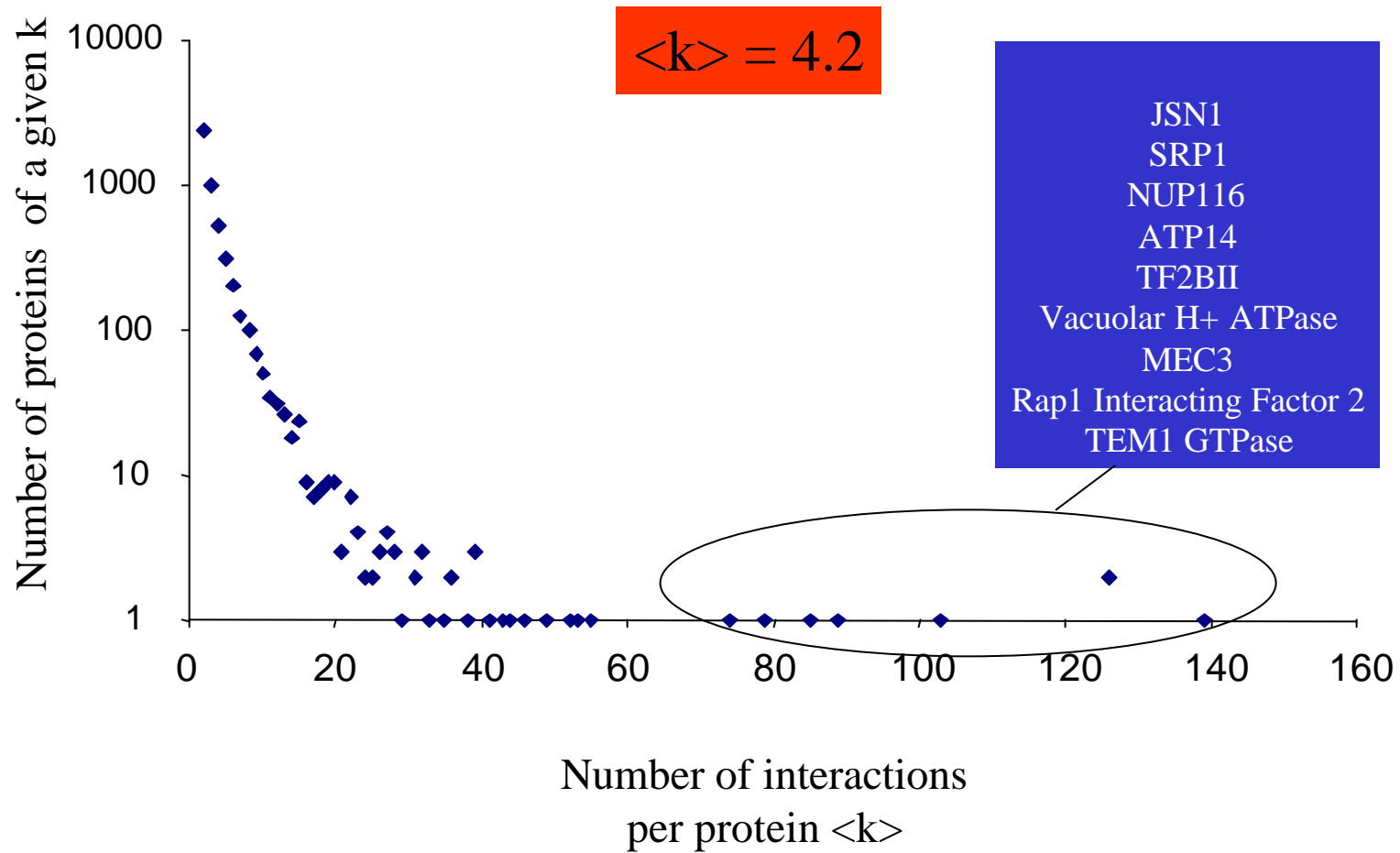
# WHAT KIND OF PROPERTIES DOES THE CONNECTED DIP NETWORK POSSESS?

Are some of these properties  
quantifiable  
and can they answer some  
biologically relevant questions ?

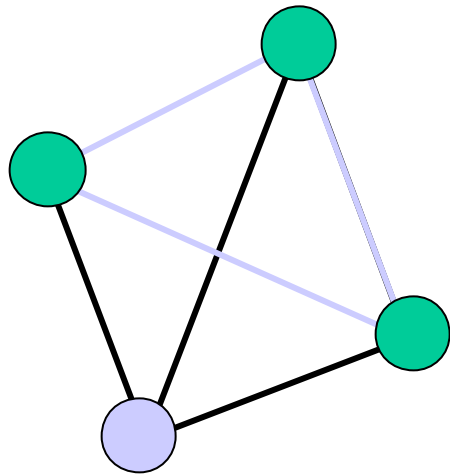
**Connectivity**  
**Path length**  
**Robustness**



# HOW MANY INTERACTIONS PER PROTEINS ?



## CLUSTERING COEFFICIENT ( $C_v$ ) SHOWS IF A PROTEIN IS PART OF A COMPLEX



$k$  = number of edges

$k_{\max}$  = max number of possible edges

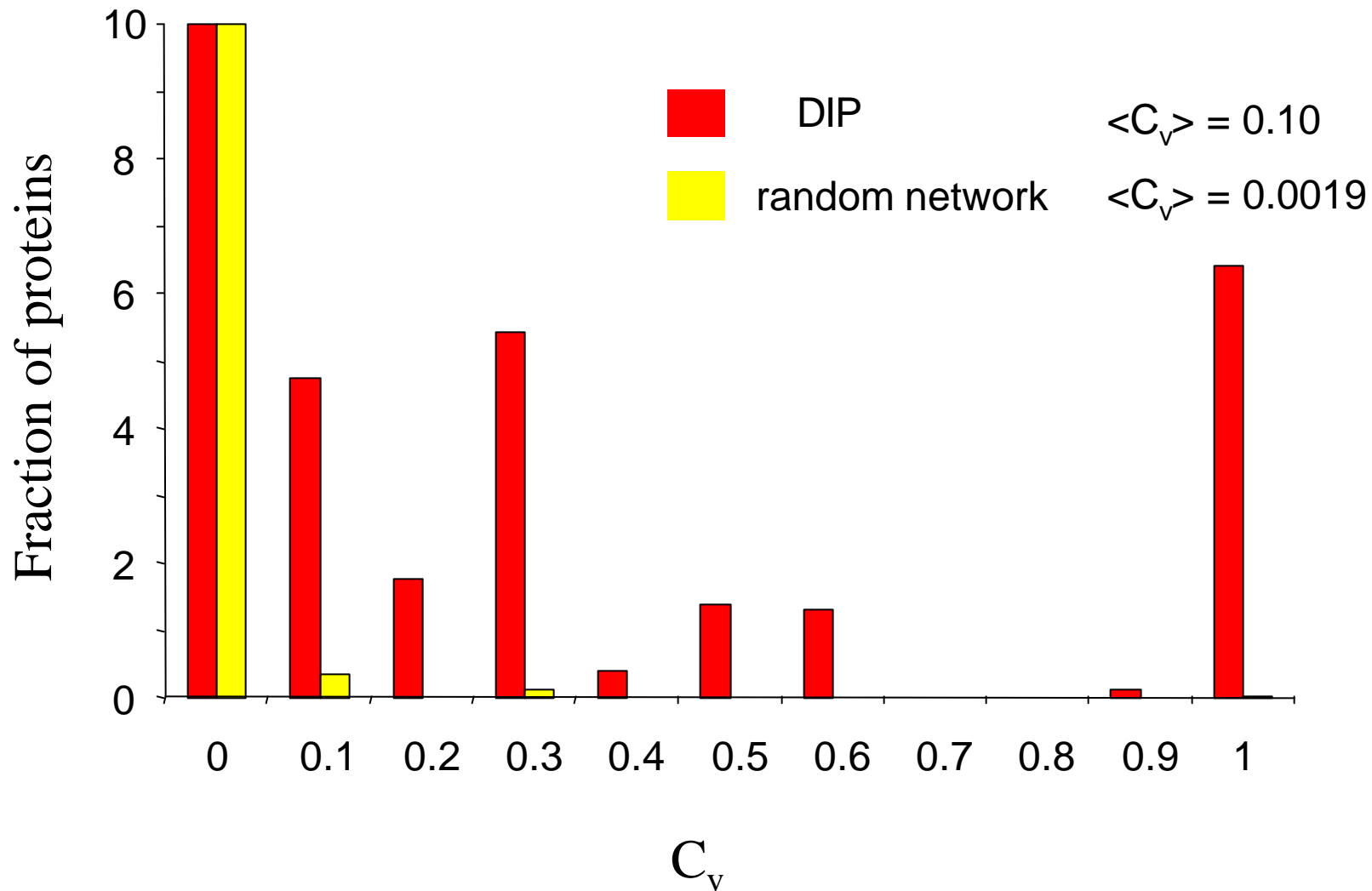
$k_{\text{obs}}$  = number of edges in the 1st shell

$$k=3 \quad k_{\max}=3(3-1)/2=3$$

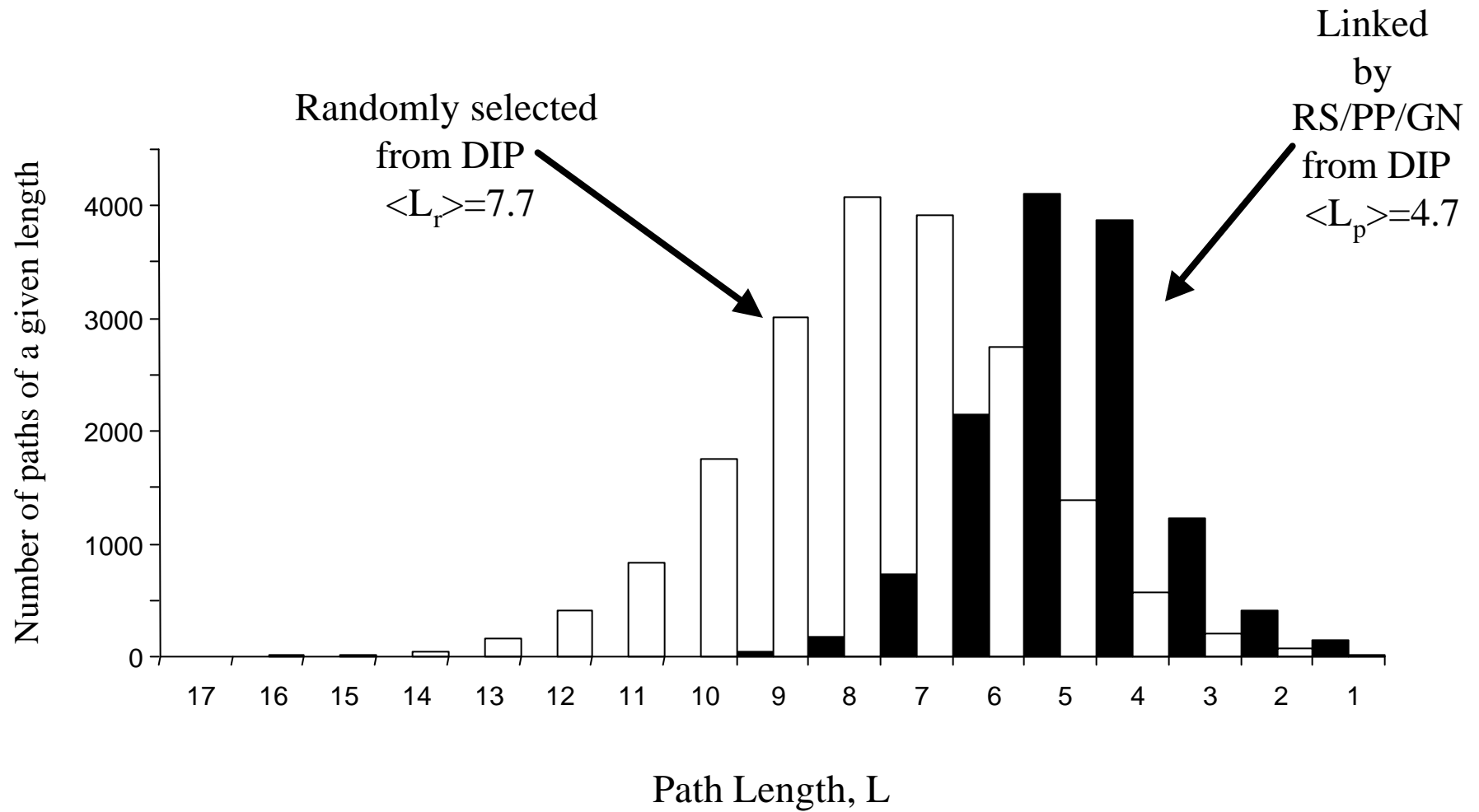
$$k_{\text{obs}}=1 \Rightarrow C_v=k_{\text{obs}}/k_{\max}=1/3$$

$$k_{\text{obs}}=3 \Rightarrow C_v=k_{\text{obs}}/k_{\max}=1$$

DISTRIBUTION OF  $C_v$  VALUES  
IN DIP COMPARED TO A RANDOM NETWORK  
WITH THE SAME NUMBER OF EDGES AND NODES

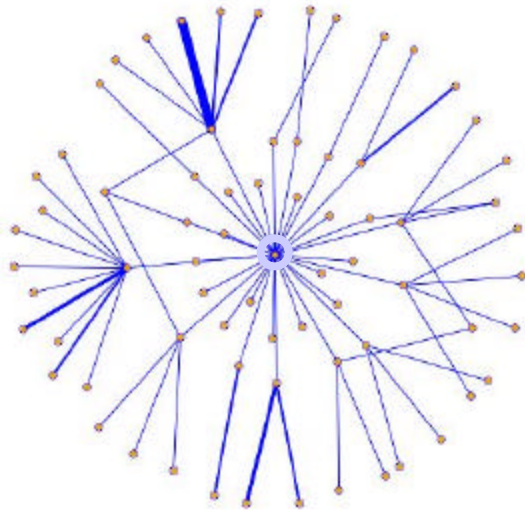


# PATH LENGTH FUNCTIONALLY LINKED PROTEINS vs RANDOM PAIRS



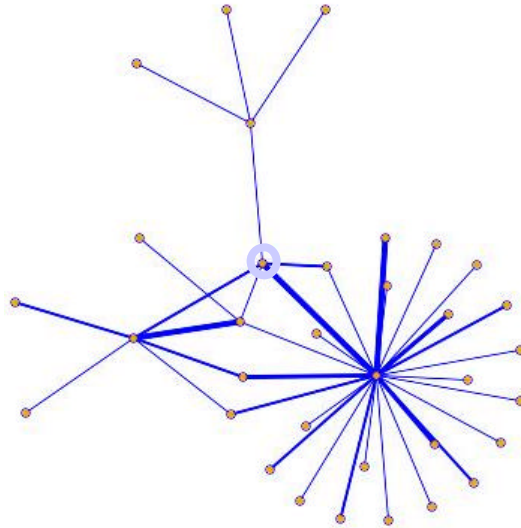
# ARE CONNECTIVITY AND FUNCTION RELATED ?

$C_v = 0 - 0.1$



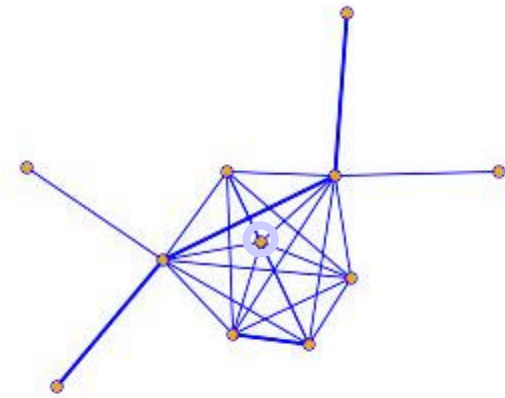
SNP1  
Actin  
Karyopherin  
MSL1

$C_v = 0.2 - 0.7$



CAP  
Ste5  
Kss1  
Fus3

$C_v = 0.8-1$



ORC1 complex  
ATP synthase