第8-1章:蛋白质相互作用的实验方法和预测

- Experimental methods
- Prediction of protein-protein interactions

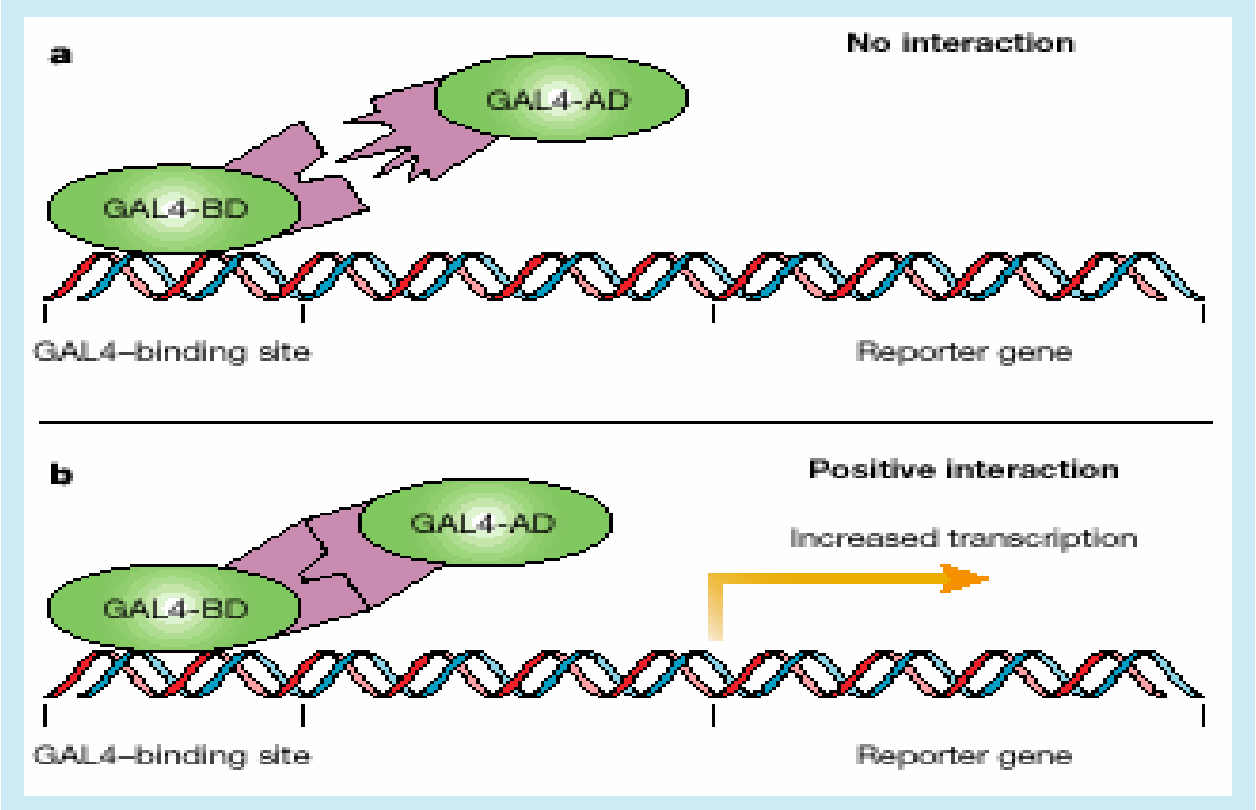
Part I: Experimental Methods

- Physical interaction
  - Yeast two hybrid system
  - TAP-mass spectrometry
- Genetic interaction
  - SGA
  - EMAP

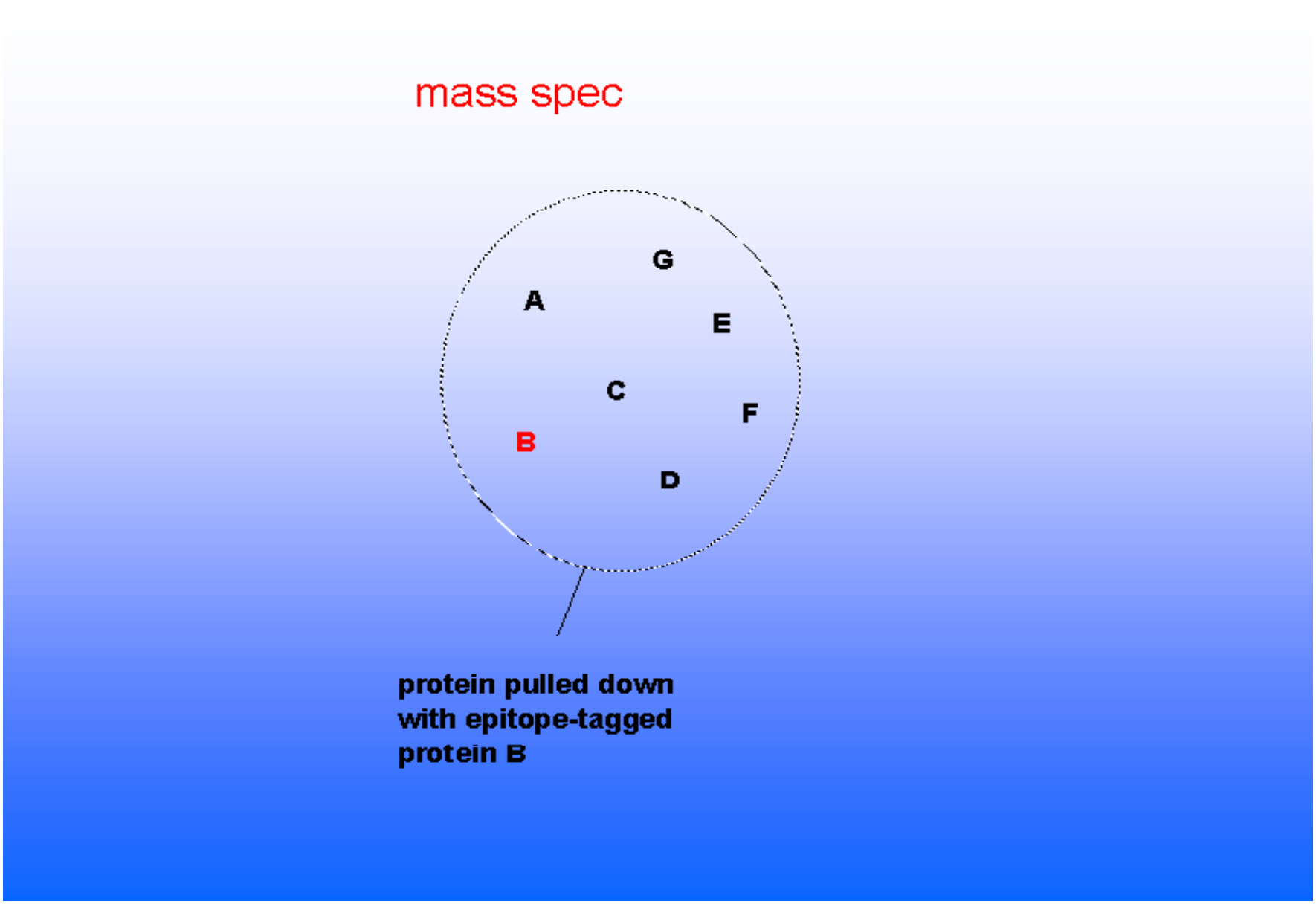
Protein-protein interactions (Experimental methods)

- Co-immunoprecipitation.
- Two-hybrid system (Uetz et al. 2000, Ito et al. 2000, 2001).
- Purified Complex by mass spectrometry
  - TAP: Tandem affinity purification (Gavin et al. 2002).
  - HMS-PCI: high-throughput mass spectrometric protein complex identification (Ho et al. 2002).

Mechanism of two-hybrid system

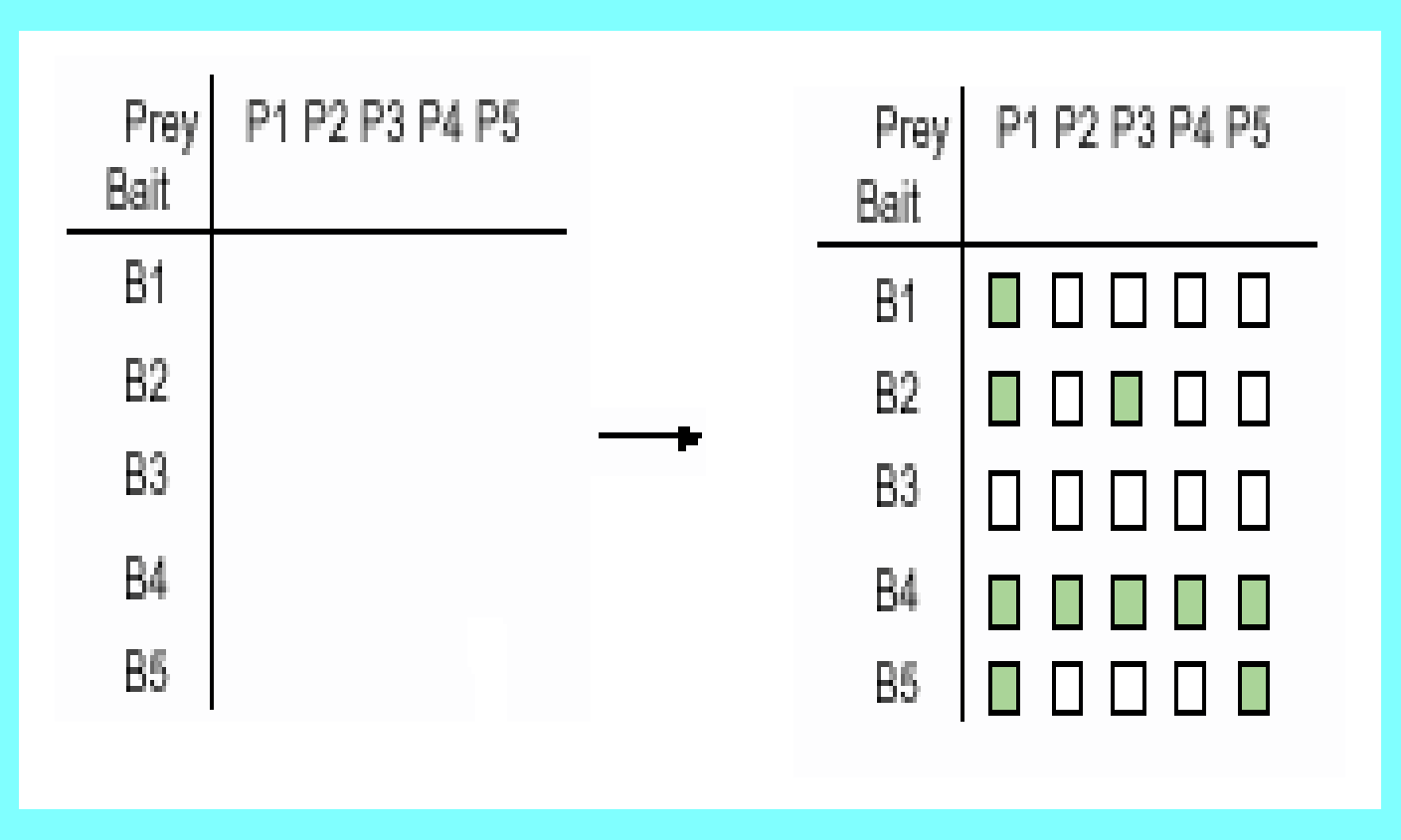


From: Nature 405, June 15, 2000, 837-846.



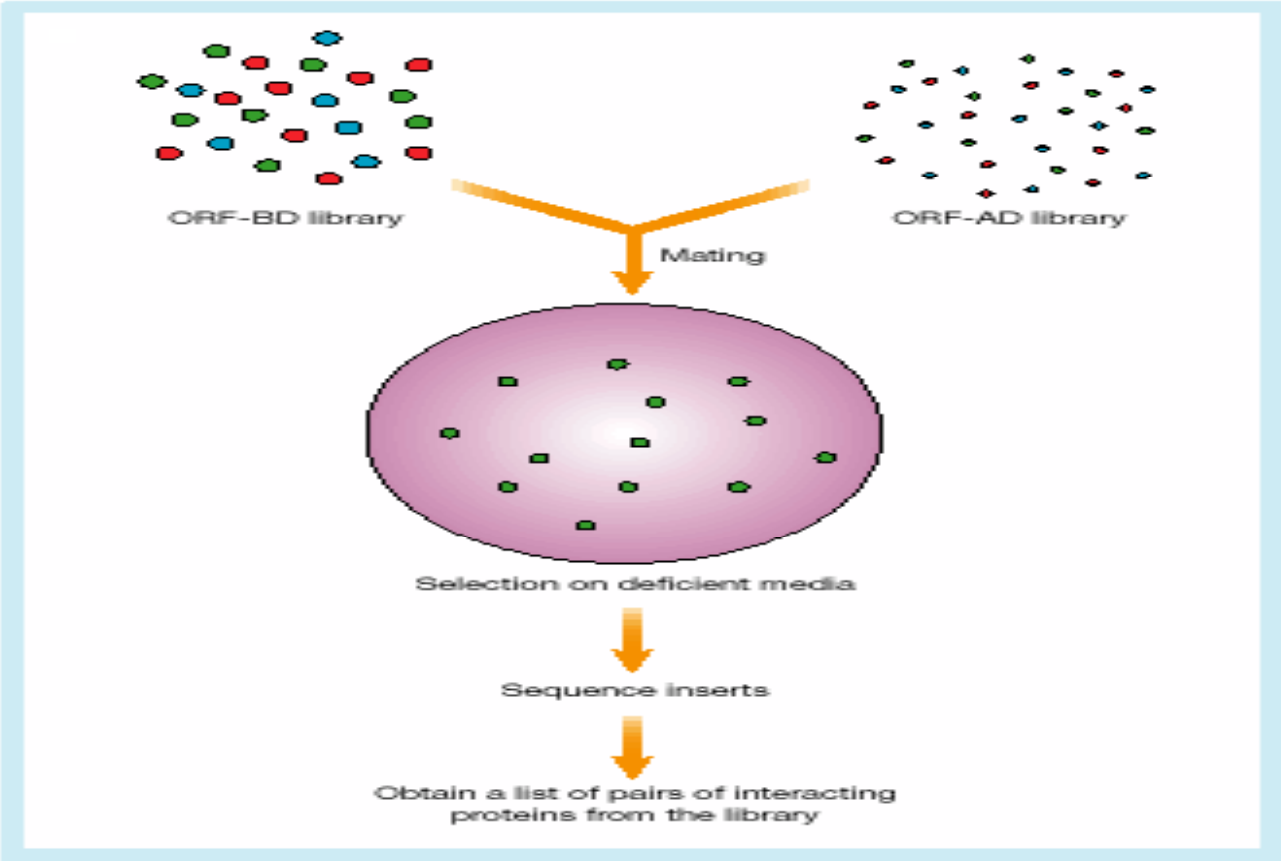
Gavin et al. (2002) Nature 415:141

Matrix method (two hybrid)



From: TRENDS in Genetics Vol.17, No.6, June 2001.

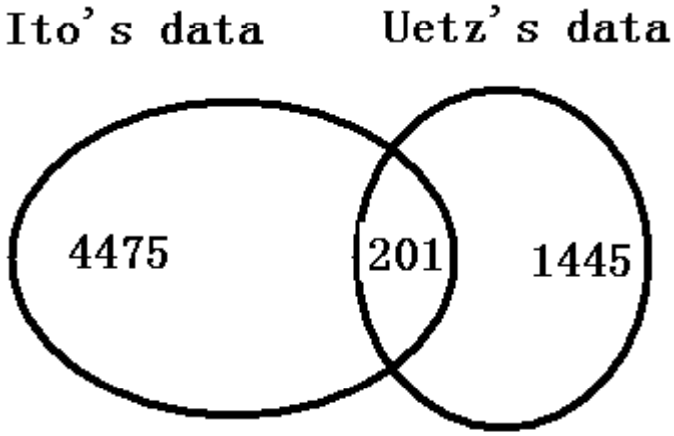
Interaction Sequence Tags (ISTs)



From: Nature 405, June 15, 2000, 837-846.

Two data sets from yeast two hybrid system

- Uetz's data (Uetz et al. 2000).
- Ito's data (Ito et al. 2000, 2001).



Possible Errors in 2-hybrid system

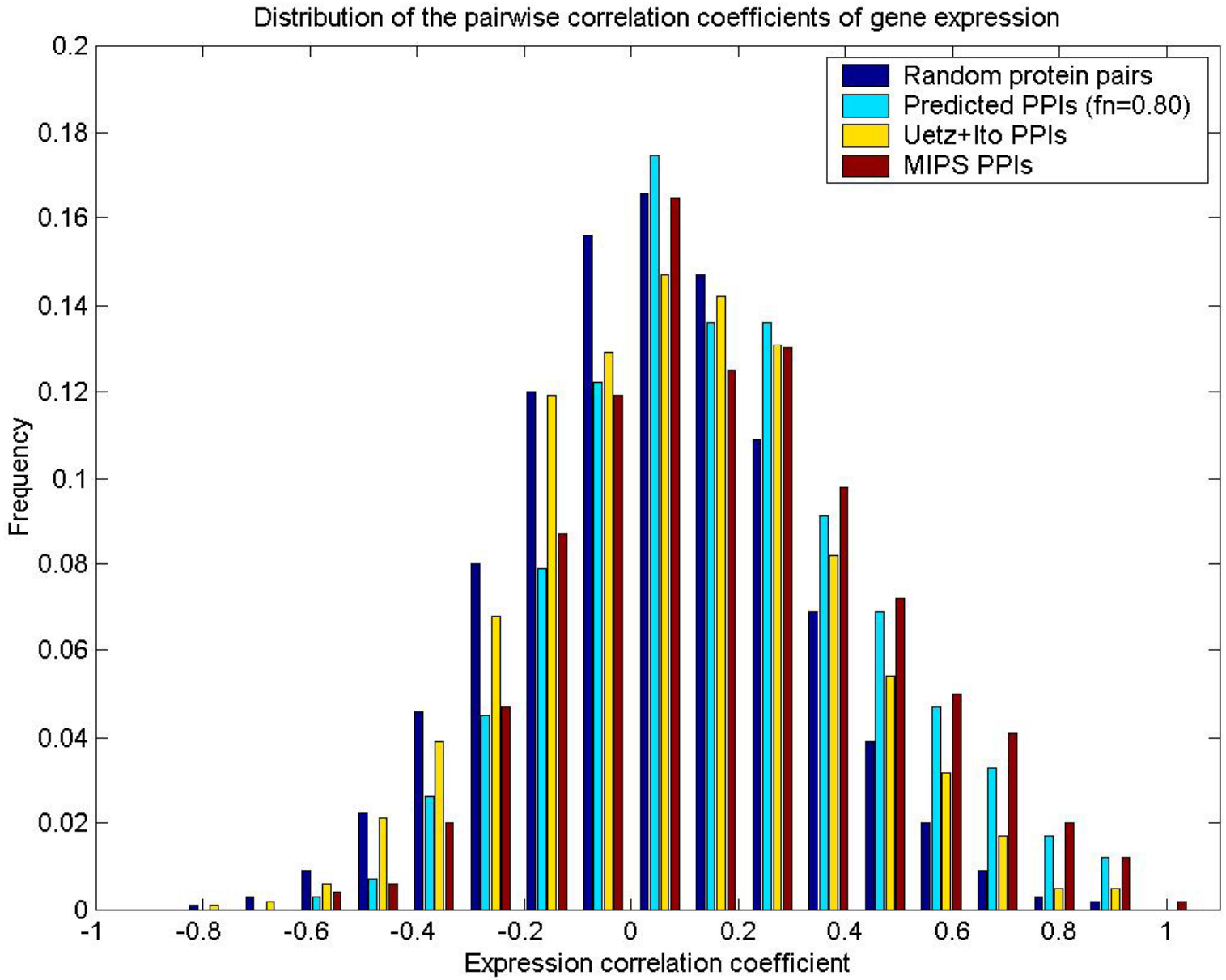
- False positive.
  - Possible mutation during PCR-amplifying.
  - Stochastic activation of reporter gene.
- False negative.
  - Membrane protein, post-translational modification protein, those self-activating reporter genes (Removed in experiment).
  - Weak interactions.

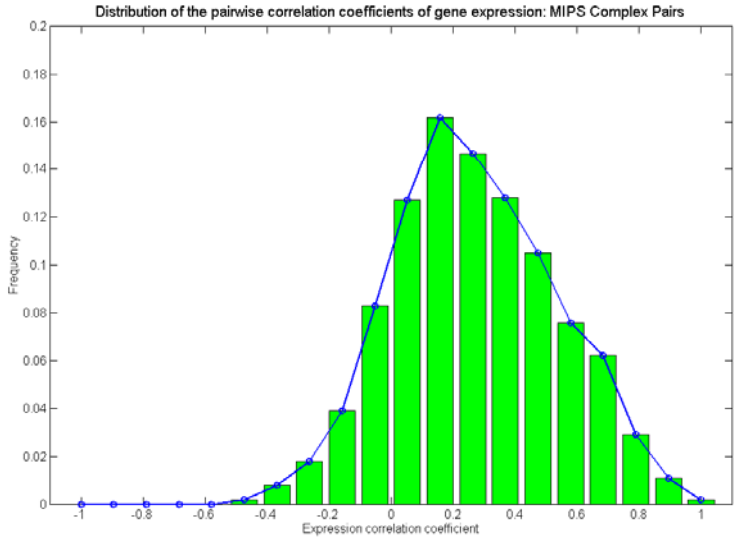
The size of interactome for yeast (5-50/protein)

Estimate the reliability of the experimental data

- Reference:

- Minghua Deng, Fengzhu Sun and Ting Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. Pacific Symposium on Biocomputing 2003: 140-151.


Distribution of the pairwise correlation coefficients of gene expression
Random protein pairs
Predicted PPIs (fn=0.80)
Uetz+Ito PPIs
MIPS PPIs
Frequency
Expression correlation coefficient
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

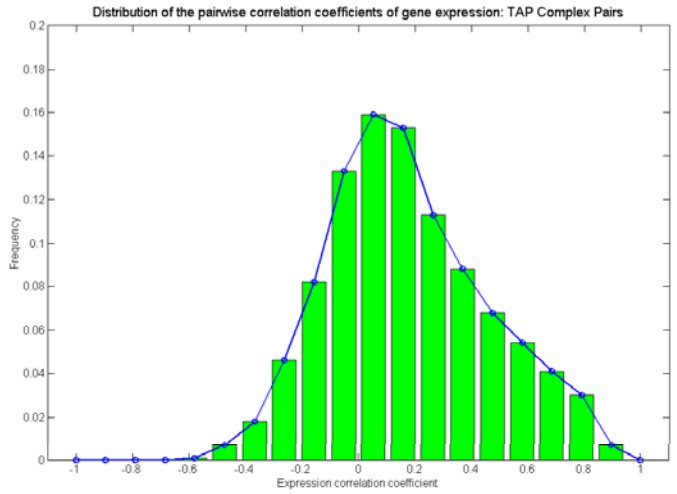


Non-interaction

$1 - \alpha$

α

Real interaction



Observed interaction data

MLE of the reliability

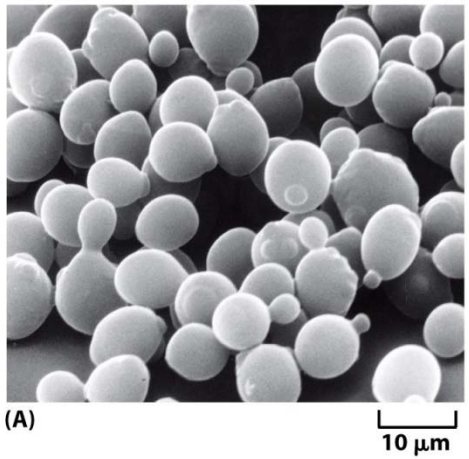
- Likelihood function

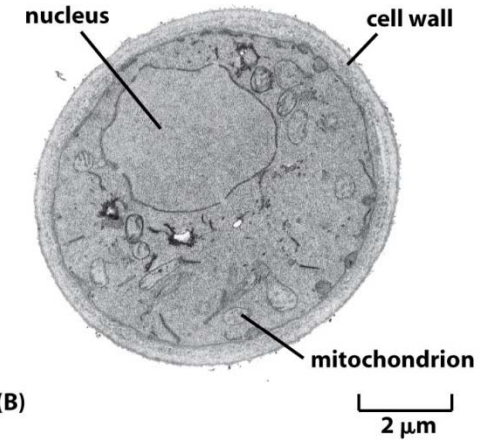
$$L(\alpha) = \prod_{k=1}^{K} (\alpha p_k + (1-\alpha) q_k)^{n_k}$$

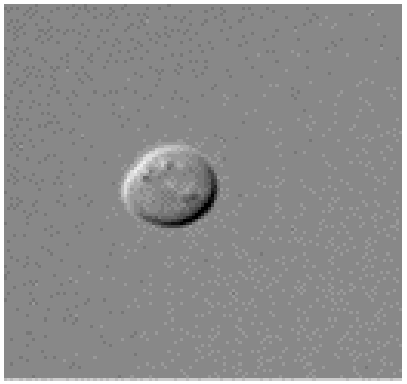
- Precision of the estimation

$$Var(\widehat{\alpha}) = \frac{1}{\sum_{k=1}^{K} n_k \frac{(p_k - q_k)^2}{(\widehat{\alpha} p_k + (1-\widehat{\alpha}) q_k)^2}}$$

Budding Yeast
Saccharomyces Cerevisiae

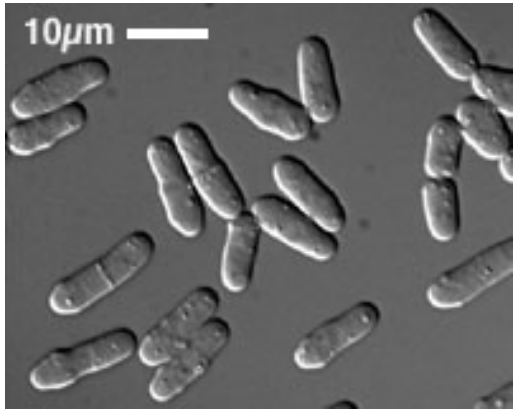


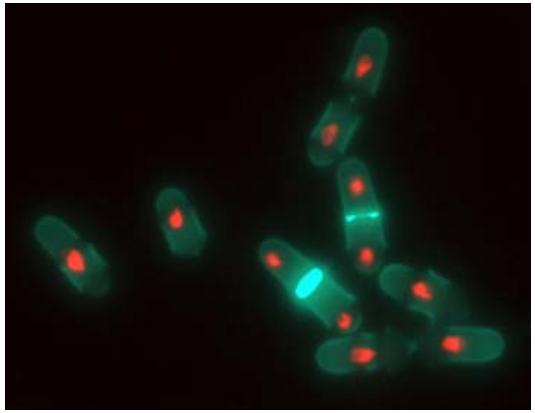
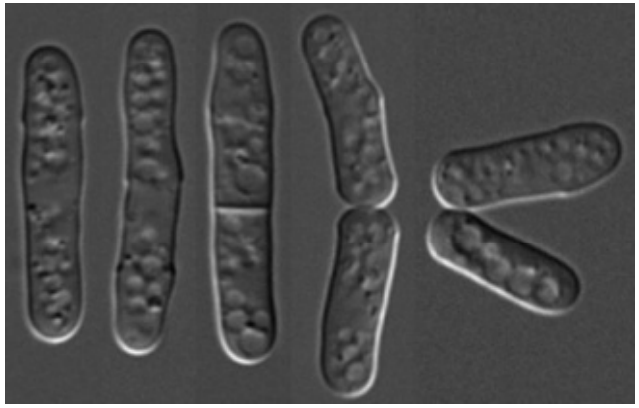




- a and α mating type, cell cycle
- 6300 genes (1997)
- Genome-wide single mutants analysis (2000~)

Fission yeast
Schizosaccharomyces Pombe



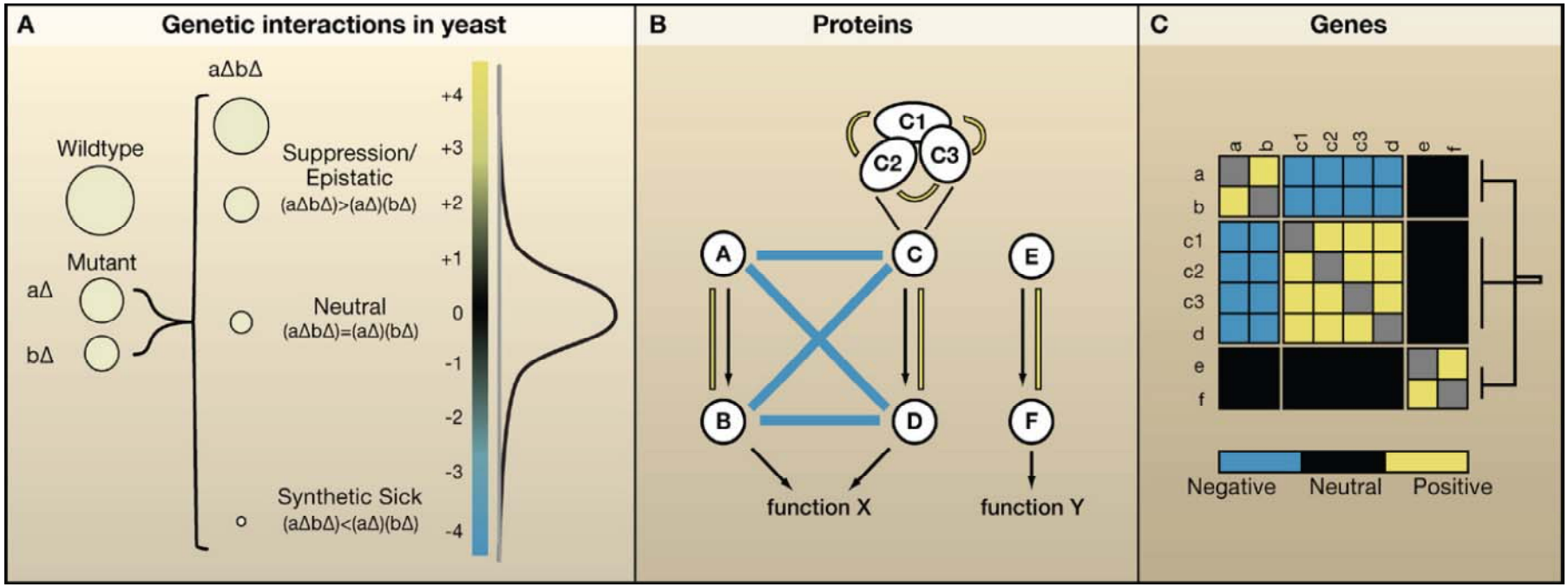
- 1000 million years separation from budding yeast;
- 13.8 Mb genome size, 4824 genes (open reading frames, OPF);
- 3 chromosomes, no genome-wide duplications; h+ and h- mating types;
- Cell cycle: 10% G1, 10% S, 70% G2 and 10% M phases.
- Genome-wide single mutants analysis (2010~)

more similar to metazoans than *S. cerevisiae*

- *cell cycle* regulation in G2/ M phase,
- gene regulation by the RNAi pathway
- the widespread presence of introns in genes

What's Genetic Interaction

- Genetic interactions between two loci can be mapped by measuring how the phenotype of an organism lacking both genes (double mutant) differs from that expected when the phenotypes of the single mutations are combined
- Null model: F(ΔAB)= F(ΔA)* F(ΔB)



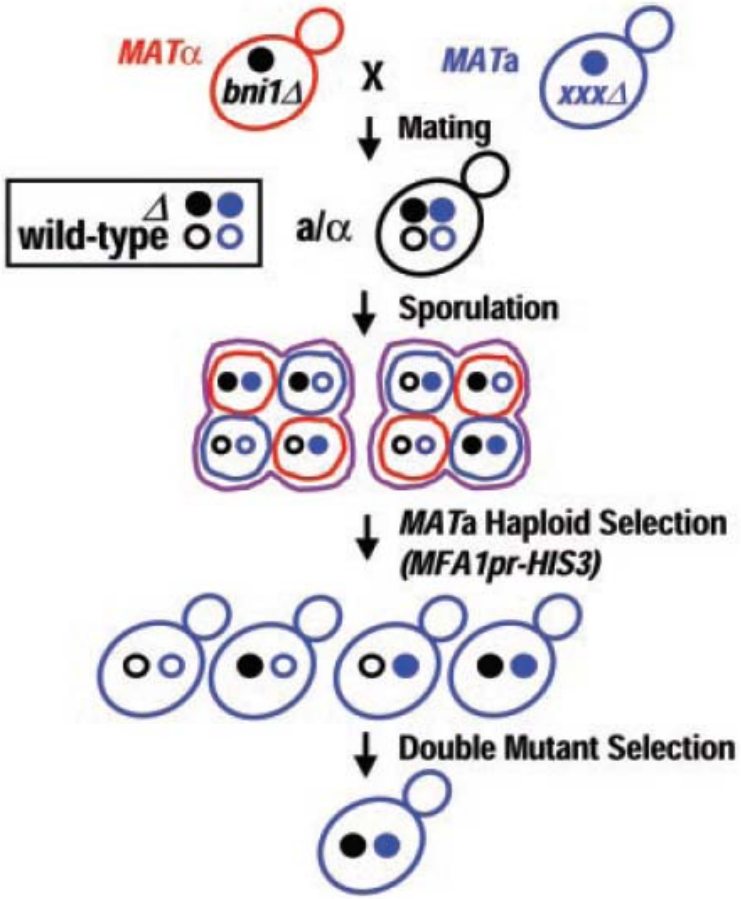
Beltrao et al. Cell 141: 739-745, 2010.

Identification of Genetic Interactions

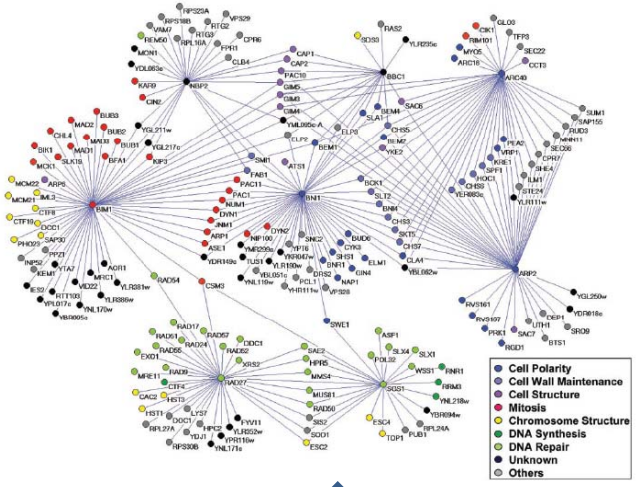
- Synthetic Gene Array (SGA) (Tong, et al. 2001)
- Diploid based Synthetic Lethality Analysis on Microarrays (dSLAM) (Pan, X., et al. 2004)

Synthetic Gene Array (SGA)

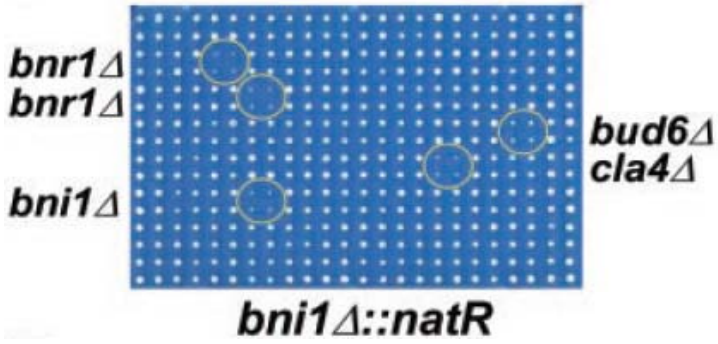
Synthetic genetic array methodology



Genetic Interaction Network



Final Screen



Amy Hin Yan Tong, et al. *Science,* 2001.

EMAP is the Extension of SGA

- EMAP: Epistatic Miniarray Profiles (Maya Schuldiner, et al. 2005. *Cell*)
- Quantitative measurement of phenotype (colony size)
  - Measure both positive and negative interactions.
  - Genome-wide (DAMP for lethal genes).

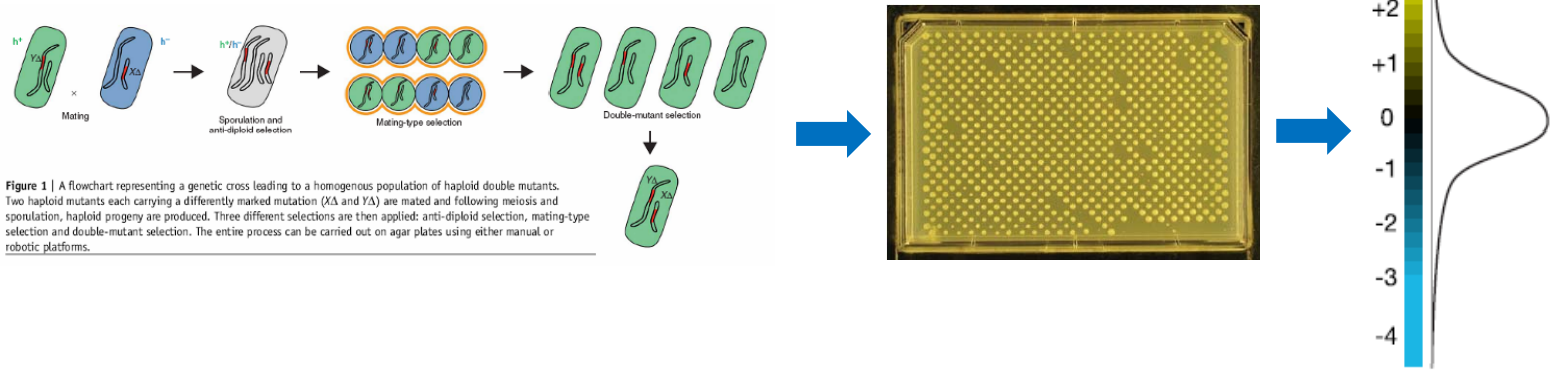


Figure 1 | A flowchart representing a genetic cross leading to a homogenous population of haploid double mutants. Two haploid mutants each carrying a differently marked mutation (XΔ and YΔ) are mated and following meiosis and sporulation, haploid progeny are produced. Three different selections are then applied: anti-diploid selection, mating-type selection and double-mutant selection. The entire process can be carried out on agar plates using either manual or robotic platforms.

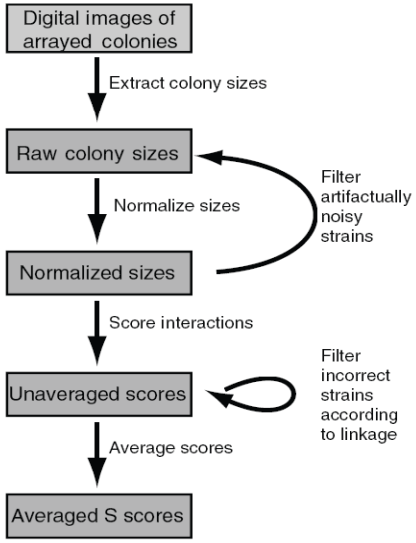
Dorothea Fiedler, et al. *Cell, 2009*

EMAP S-score

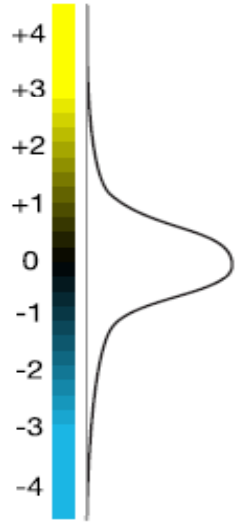
- Quantitative measure: $\epsilon = W_{ab} - W_a W_b,\ W_a = w/w_{wild}.$

No interaction	Synthetic sick/Lethality	Synthetic alleviating
$\epsilon = 0$	$\epsilon < 0$	$\epsilon > 0$

  - T-Test with null hypothesis $\epsilon = 0$



$$S_0 = \frac{\mu_{ab} - w_a \mu_b}{\sqrt{n_1 S_{ab}^2 + n_2 w_a^2 S_b^2}}$$



Sean R Collins, et al. *Genome Biology 7: R63, 2006.*

PPI databases

- MIPS: Munich Information center for Protein Sequences (http://mips.gsf.de)
- DIP: Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu)
- BIND: Biomolecular Interaction Network Database (http://www.bind.ca)
- GRID: General Repository for Interaction Datasets (http://biodata.mshri.on.ca/grid)
- MINT: Molecular Interaction Database (http://cbm.bio.uniroma2.it/mint/)

Further Reading

- For more experimental methods and databases, please read the following review paper
  - Shoemaker BA, Panchenko AR (2007) Deciphering Protein–Protein Interactions. Part I. Experimental Techniques and Databases. PLoS Comput Biol 3(3): e42. doi:10.1371/journal.pcbi.0030042.

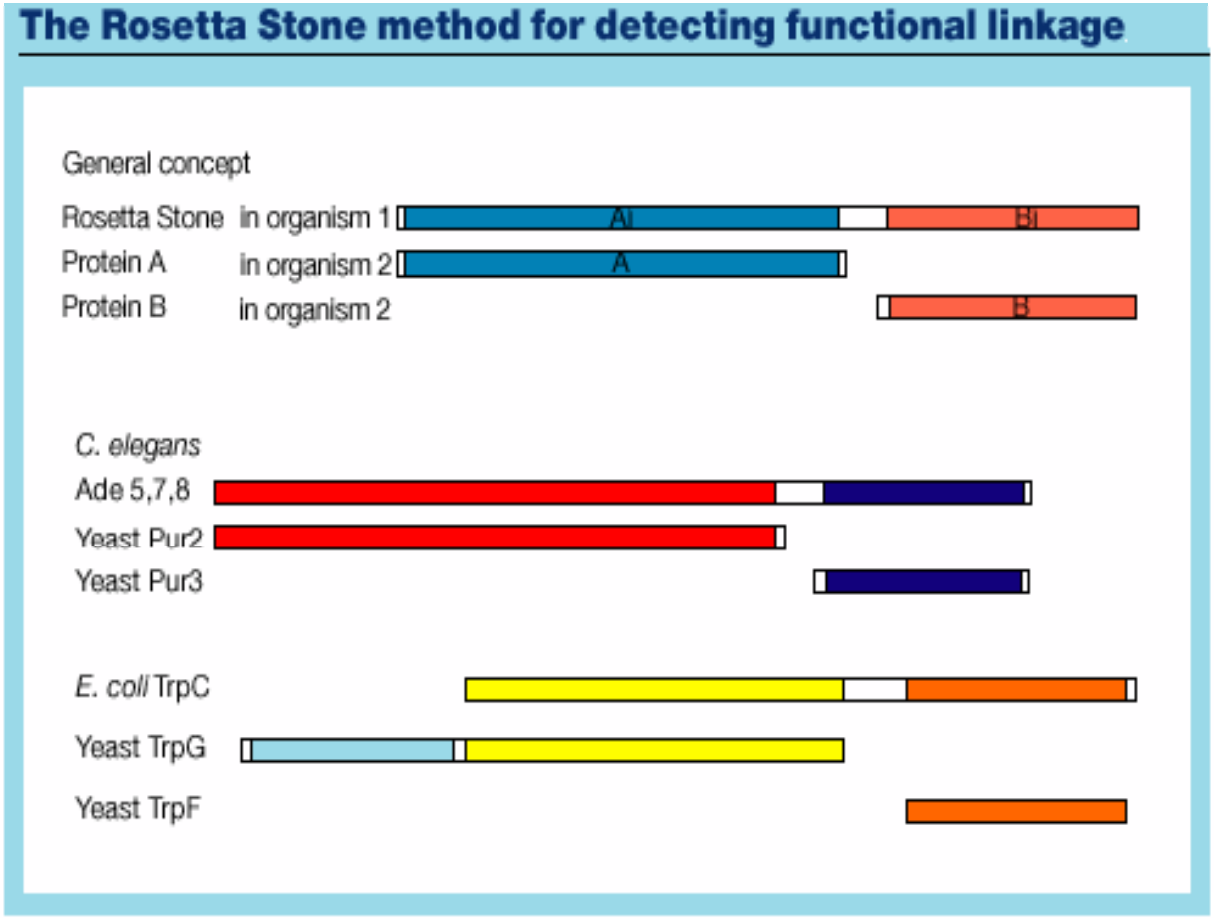
Protein-protein interactions (Computational Methods)

- Gene fusion method (A.Enright 1999. E.Maccote 1999)
- Phylogenetic profile method (M.Pellegrini 1999, D.Eisenberg, 1999).
- Gene cluster method (R.Overbeek, 1999).
- Highly co-expressed gene pairs.

Part II: Predicting Protein-protein Interactions

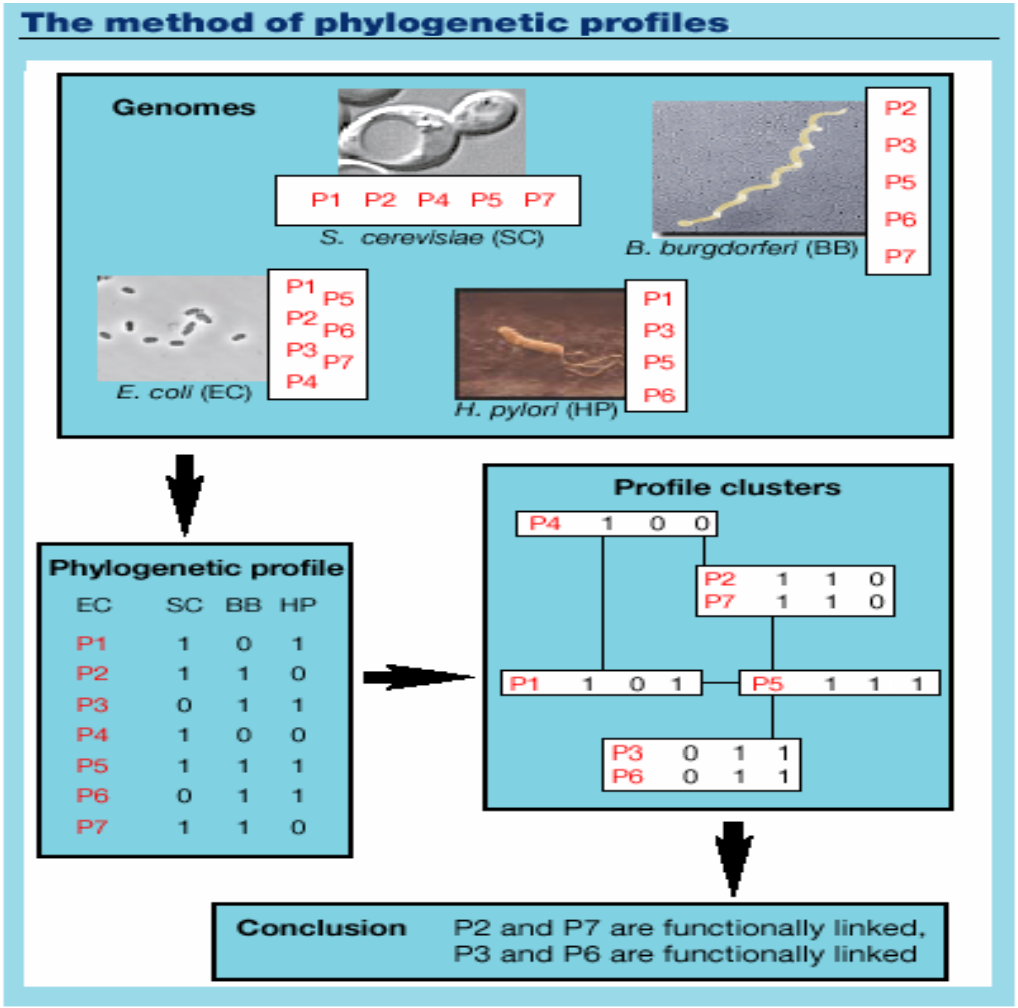
- Some computational methods
- Predicting protein-protein interaction from domains
  - Association method
  - MLE method

Rosetta Stone Method



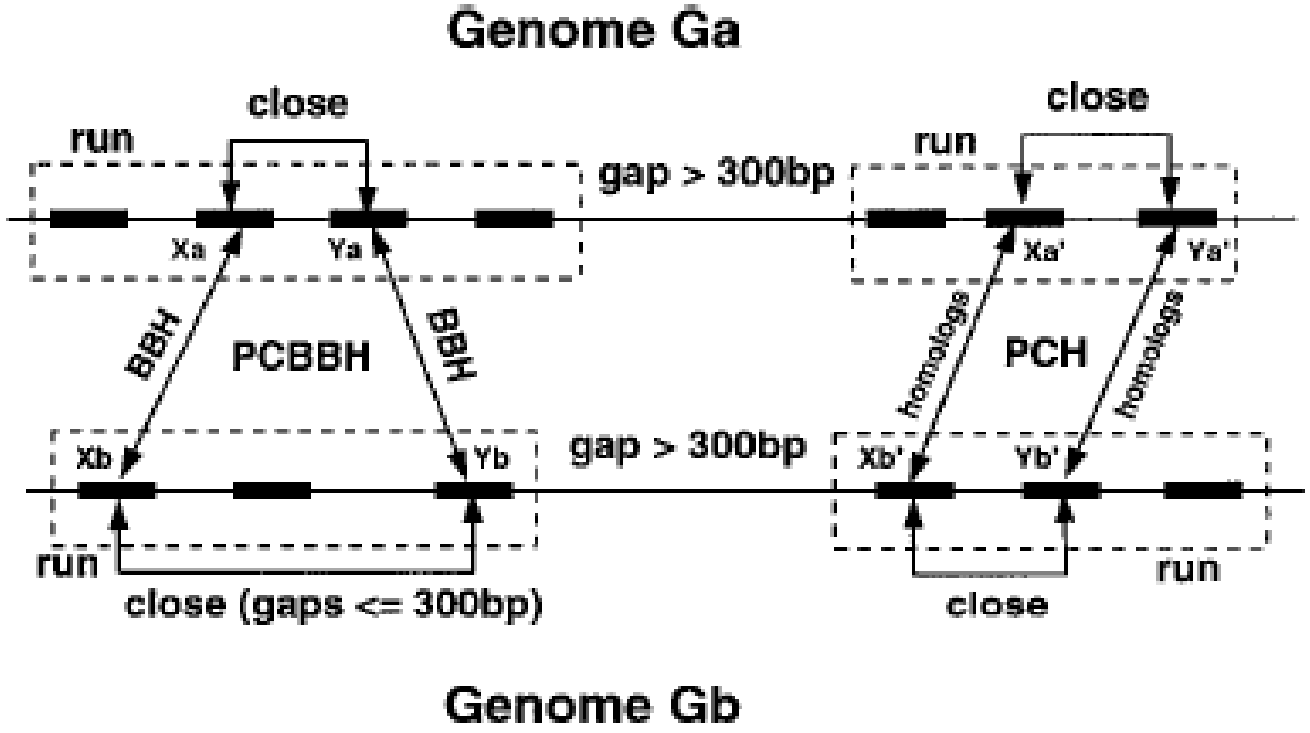
From: Nature Vol. 405, June.15, 2000, 823-826

Phylogenetic Profiles Method



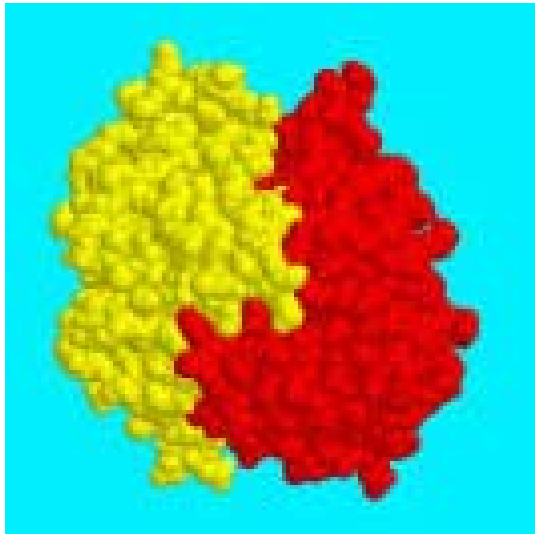
From: Nature Vol. 405, June, 15, 2000, 823-826

Using Gene Clusters to Infer Functional Coupling



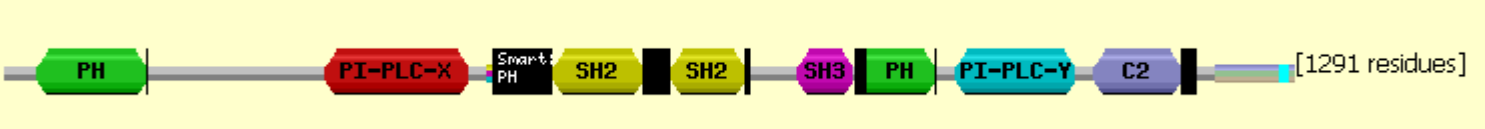
From: R.Overbeek, PNAS 96, 2896-2901, 1999.

Structure of Proteins



Predicting PPIs from Domains

- Domains are treated as elementary unit of function.
- Domains are responsible for the generation of interactions.
- Understanding protein-protein interaction at the domain level.



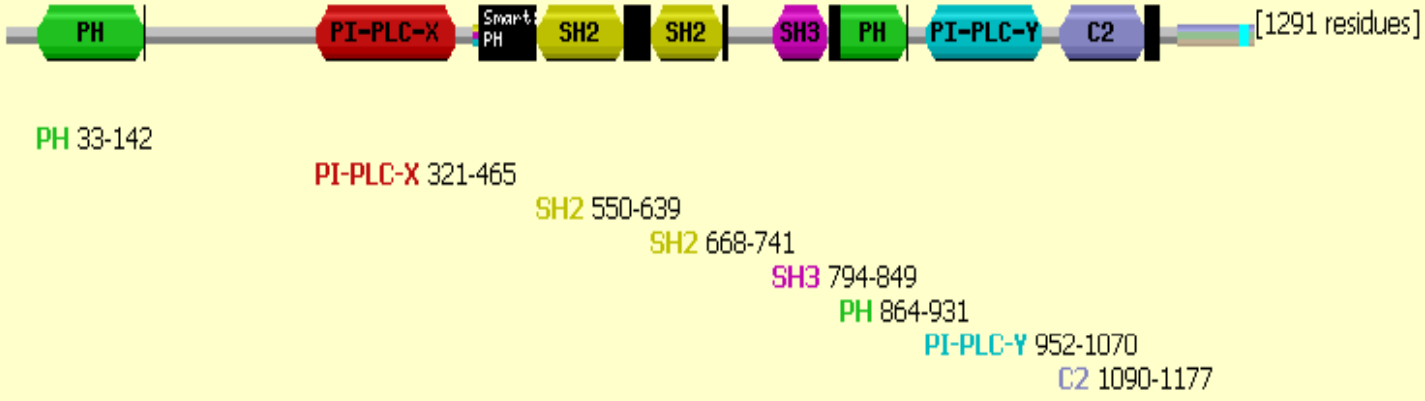
Domain Databases

- Pfam, domain classification by HMM.
- Prodom.
- PRINTS, fingerprint information of protein sequences.
- SMART, mobile domain.
- BLOCKs, multiple alignment blocks.
- Interpro.

SwissPfam entry for PIG1_BOVIN

Description from Swissprot for PIG1_BOVIN :

1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase gamma 1(ec 3.1.4.11) (plc-gamma-1) (phospholipase c-gamma-1) (plc-ii)(plc-148)



Key

signal peptide: > pfamA: > Context: > smart: > transmembrane: > low complexity: > coiled coil: > pfamB:

Source	Domain	Start	End
Pfam	PH	33	142
Pfam	PI-PLC-X	321	465

Overlapping Domains: Change the domain order using the ^ and v buttons. View the changes by clicking the 'Change order' button.

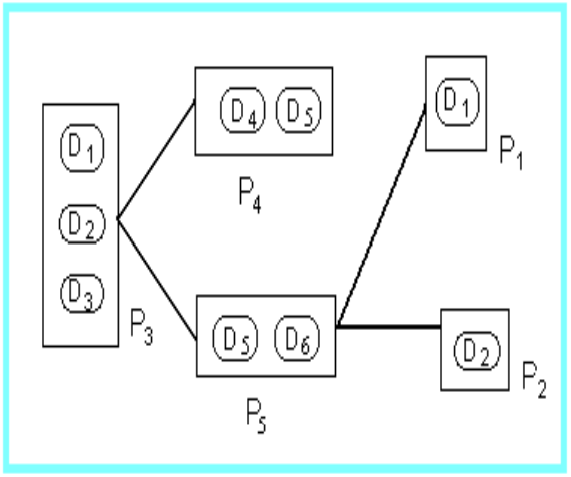
high priority

Association-A simple method

$$V(D_{ij}) = \frac{\#\{\text{Interacted protein pairs contain } D_{ij}\}}{\#\{\text{All protein pairs contain } D_{ij}\}}$$

More observed PPIs for one domain pair will give higher probability of interaction for that domain pair.

Simple Example



By assoiation method:

$$D_{34} = D_{35} = D_{36} = D_{26} = D_{16} = 1.0$$
$$D_{15} = D_{25} = 0.75, D_{14} = D_{24} = 0.5$$

Others are 0.0.

D_{15}: $\{P_{34}, P_{35}, P_{15}, P_{14}\}$.

Limitation of Association Method

- For multiple-domain proteins, this method computes the value for a certain domain pair ignoring the value of other domain-domain pairs. So it's a local one.
- This method cannot deal with possible error of the data.

Probabilistic Model

- Domain-domain interactions are independent, which means that the event that two domains interact or not does not depend on other domains.
- Two proteins interact if and only if at least one pair of domains from the two proteins interact.

Some Notations

P_{ij}: Random variable for REAL interaction of protein P_i and P_j.

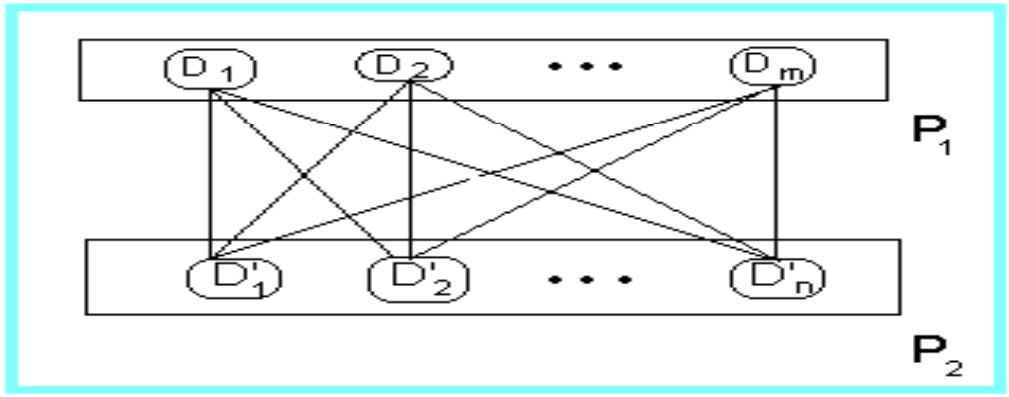
O_{ij}: Random variable for OBSERVED interaction of protein P_i and P_j.

o_{ij}: Realization of O_{ij}.

D_{ij}: Random variable for interaction of domain D_i and D_j.

$\lambda_{ij} = \Pr(\ D_i \text{ and } D_j \text{ interact})$.

Probabilistic Model (Cont.)



$$\Pr(P_{ij} \text{ interact }) = 1.0 - \prod_{D_{mn} \text{ in } P_{ij}} (1.0 - \Pr(D_{mn} \text{ interact }))$$

$$= 1.0 - \prod_{D_{mn} \text{ in } P_{ij}} (1.0 - \lambda_{mn})$$

Where $\lambda_{mn} = \Pr(D_{mn} \text{ interact })$

Modeling the Possible Errors

False positive

$$fp = Pr(O_{ij} = 1 | P_{ij} = 0)$$

False Negative

$$fn = Pr(O_{ij} = 0 | P_{ij} = 1)$$

Estimation of fp and fn

Real interactions: 5-50/protein, t1=5, t2=50;

Observed interactions: T=5719 (Uetz+Ito)

Proteins: N=6359 (SGD).

$$
\begin{aligned}
fn &= \Pr(O_{ij} = 0 \mid P_{ij} = 1) \\
&= 1.0 - \frac{\Pr(O_{ij} = 1, P_{ij} = 1)}{\Pr(P_{ij} = 1)} \\
&\geq 1.0 - \frac{\Pr(O_{ij} = 1)}{\Pr(P_{ij} = 1)} \\
&\geq 1.0 - \frac{T}{N \times t_1/2} \\
&\geq 0.64.
\end{aligned}
$$

$$
\begin{aligned}
fp &= \Pr(O_{ij} = 1 \mid P_{ij} = 0) \\
&= \frac{\Pr(O_{ij} = 1, P_{ij} = 0)}{\Pr(P_{ij} = 0)} \\
&\leq \frac{\Pr(O_{ij} = 1)}{\Pr(P_{ij} = 0)} \\
&\leq \frac{T}{N \times (N+1)/2 - N \times t_2/2} \\
&\leq 2.85E-4.
\end{aligned}
$$

Likelihood Function

$$\Pr(O_{ij} = 1)$$

$$= \Pr(O_{ij} = 1, P_{ij} = 1) + \Pr(O_{ij} = 1, P_{ij} = 0)$$

$$= \Pr(P_{ij} = 1)(1 - fn) + (1 - \Pr(P_{ij} = 1))fp$$

$$L = \prod (\Pr(O_{ij} = 1))^{o_{ij}} (1 - \Pr(O_{ij} = 1))^{1 - o_{ij}}$$

$$o_{ij} = \begin{cases} 1 & \text{if the interaction of } P_i \text{ and } P_j \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

Missing Data Problem

Parameters:

$$\theta = \{\lambda_{mn}, \forall\, m, n, fp, fn\}$$

Missing data:

$$D_{mn}^{(ij)}, P_{ij}.$$

$$D_{mn}^{(ij)} = \begin{cases} 1 & \text{if } D_m,\ D_n \text{ interact in protein pair } P_i \text{ and } P_j, \\ 0 & \text{otherwise.} \end{cases}$$

General EM Algorithm

- Observed data Y
- Missing data X
- Complete data Z=(Y, X).
- E-step (expectation).

$$\hat{Z} = E(Z|Y, \theta^{(t-1)})$$

- M-step (maximization).

$$\theta^{(t)} = \underset{\theta}{\text{Argmax}}\, L(\theta|\hat{Z}, \theta^{(t-1)})$$

Parameters Re-estimation (1)

A_m be the set of proteins containing domain D_m .
N_{mn} be the total number of protein pairs between A_m and A_n.

$$\lambda_{mn}^{(t)} = \frac{1}{N_{mn}} \sum_{i \in A_m, j \in A_n} E(D_{mn}^{(ij)} \mid O_{kl} = o_{kl}, \forall\, k, l,\ \theta^{(t-1)})$$

$$= \frac{\lambda_{mn}^{(t-1)}}{N_{mn}} \sum_{i \in A_m, j \in A_n} \frac{(1 - fn)^{o_{ij}} fn^{1-o_{ij}}}{\Pr(O_{ij} = o_{ij} \mid \theta^{(t-1)})}.$$

Parameters Re-estimation (2)

$$fp^{(t)} = \frac{\sum_{ij} \Pr(P_{ij} = 0, O_{ij} = 1 \,|\, O_{kl} = o_{kl}, \forall\, k, l, \theta^{(t-1)})}{\sum_{ij} \Pr(P_{ij} = 0 \,|\, O_{kl} = o_{kl}, \forall\, k, l, \theta^{(t-1)})}$$

$$= \frac{\sum_{o_{ij}=1} \frac{\Pr(P_{ij} = 0 \,|\, \theta^{(t-1)})}{\Pr(O_{ij} = 1 \,|\, \theta^{(t-1)})} fp^{(t-1)}}{\sum_{o_{ij}=1} \frac{\Pr(P_{ij} = 0 \,|\, \theta^{(t-1)})}{\Pr(O_{ij} = 1 \,|\, \theta^{(t-1)})} fp^{(t-1)} + \sum_{o_{ij}=0} \frac{\Pr(P_{ij} = 0 \,|\, \theta^{(t-1)})}{\Pr(O_{ij} = 0 \,|\, \theta^{(t-1)})} (1 - fp^{(t-1)})}$$

Parameters Re-estimation (3)

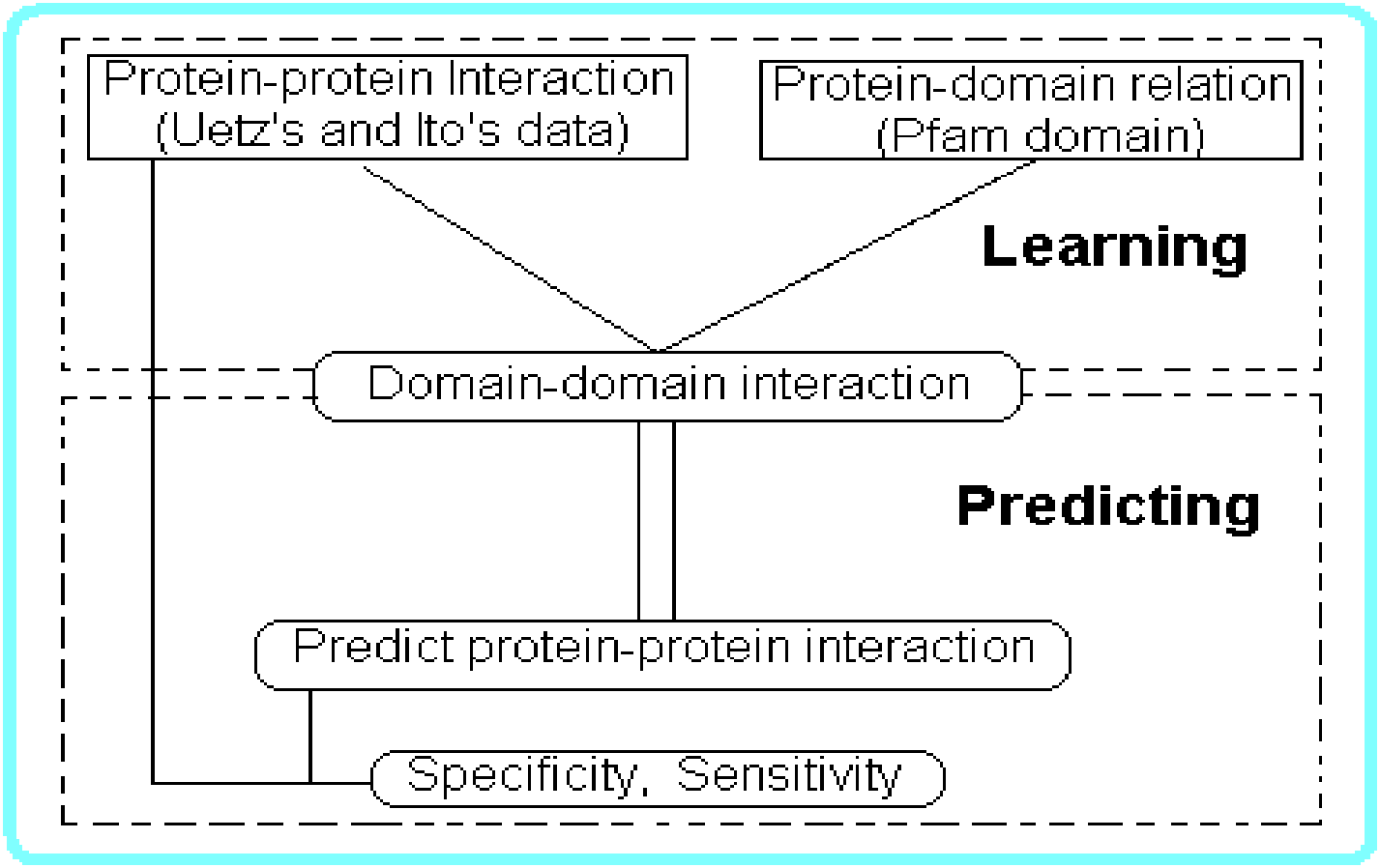
$$fn^{(t)} = \frac{\sum_{ij} \Pr(P_{ij}=1, O_{ij}=0 \,|\, O_{kl}=o_{kl}, \forall\, k,l, \theta^{(t-1)})}{\sum_{ij} \Pr(P_{ij}=1 \,|\, O_{kl}=o_{kl}, \forall\, k,l, \theta^{(t-1)})}$$

$$= \frac{\sum_{o_{ij}=0} \frac{\Pr(P_{ij}=1 \,|\, \theta^{(t-1)})}{\Pr(O_{ij}=0 \,|\, \theta^{(t-1)})} fn^{(t-1)}}{\sum_{o_{ij}=0} \frac{\Pr(P_{ij}=1 \,|\, \theta^{(t-1)})}{\Pr(O_{ij}=0 \,|\, \theta^{(t-1)})} fn^{(t-1)} + \sum_{o_{ij}=1} \frac{\Pr(P_{ij}=1 \,|\, \theta^{(t-1)})}{\Pr(O_{ij}=1 \,|\, \theta^{(t-1)})} (1 - fn^{(t-1)})}$$

Algorithm

1. Initialization: set fp and fn, choose initial values for parameters $\{\lambda_{mn}, \forall\, m, n\}$, and compute real interaction probability and observed interaction probability;
2. Update parameters $\{\lambda_{mn}, \forall\, m, n\}$ and compute the likelihood function ;
3. Go to step 2, repeat until the value of the likelihood function is unchanged (within certain error).

Flow Chart


Protein-protein Interaction
(Uetz's and Ito's data)
Protein-domain relation
(Pfam domain)
Learning
Domain-domain interaction
Predicting
Predict protein-protein interaction
Specificity, Sensitivity

Yeast Data

- Interactions (Uetz's and Ito's interaction data).
- Domain: Pfam (Pfam-A, Pfam-B).
- Proteins: SGD, N=6359.

Protein Interaction Data Sources

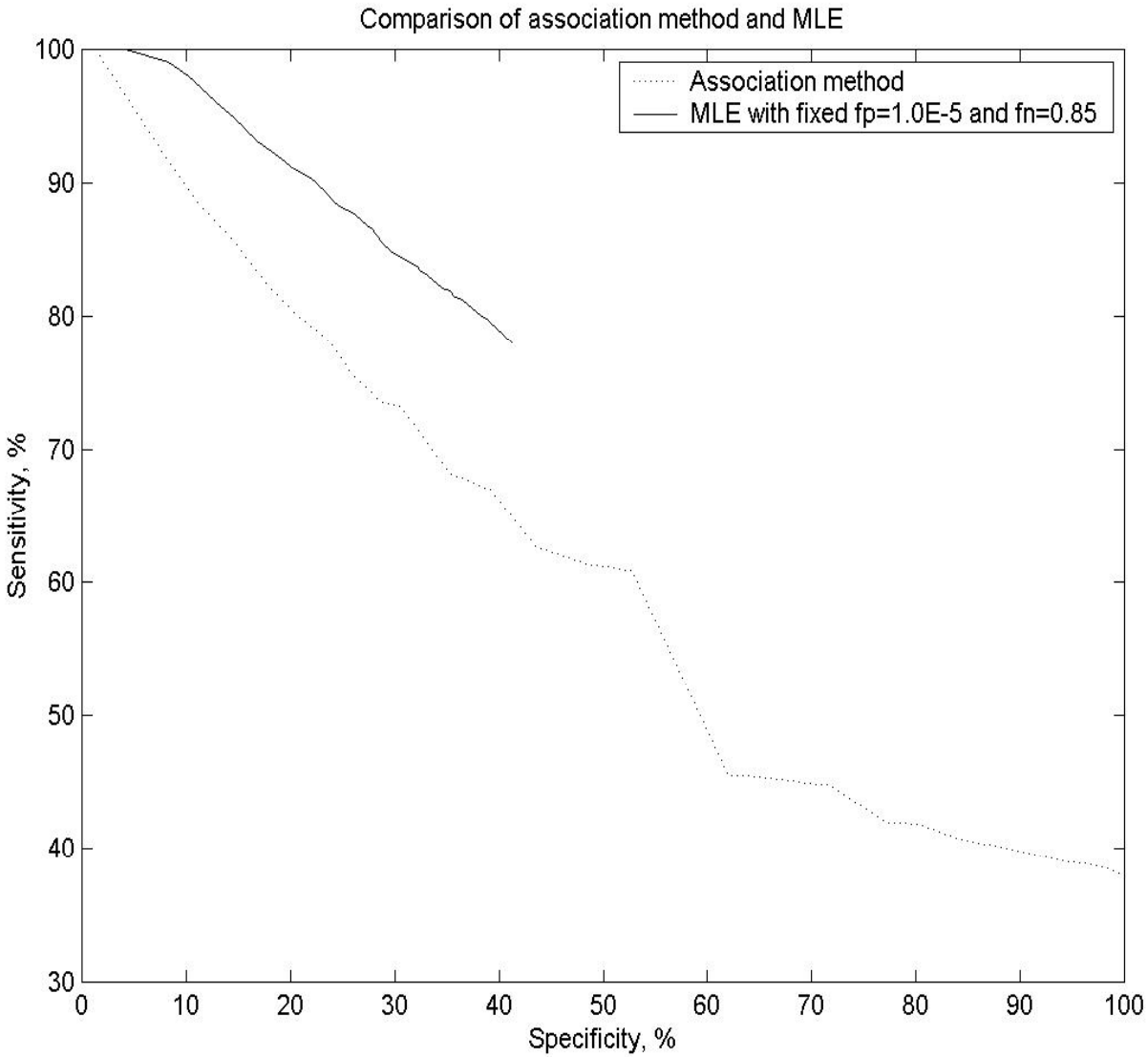
	Proteins	Pfam domains	Super - domains	PPI
Uetz	1337	1330	313	1445
Ito	3277	2776	909	4475
Uetz+Ito	3729	3124	1007	5719
Overlap	855	964	215	201

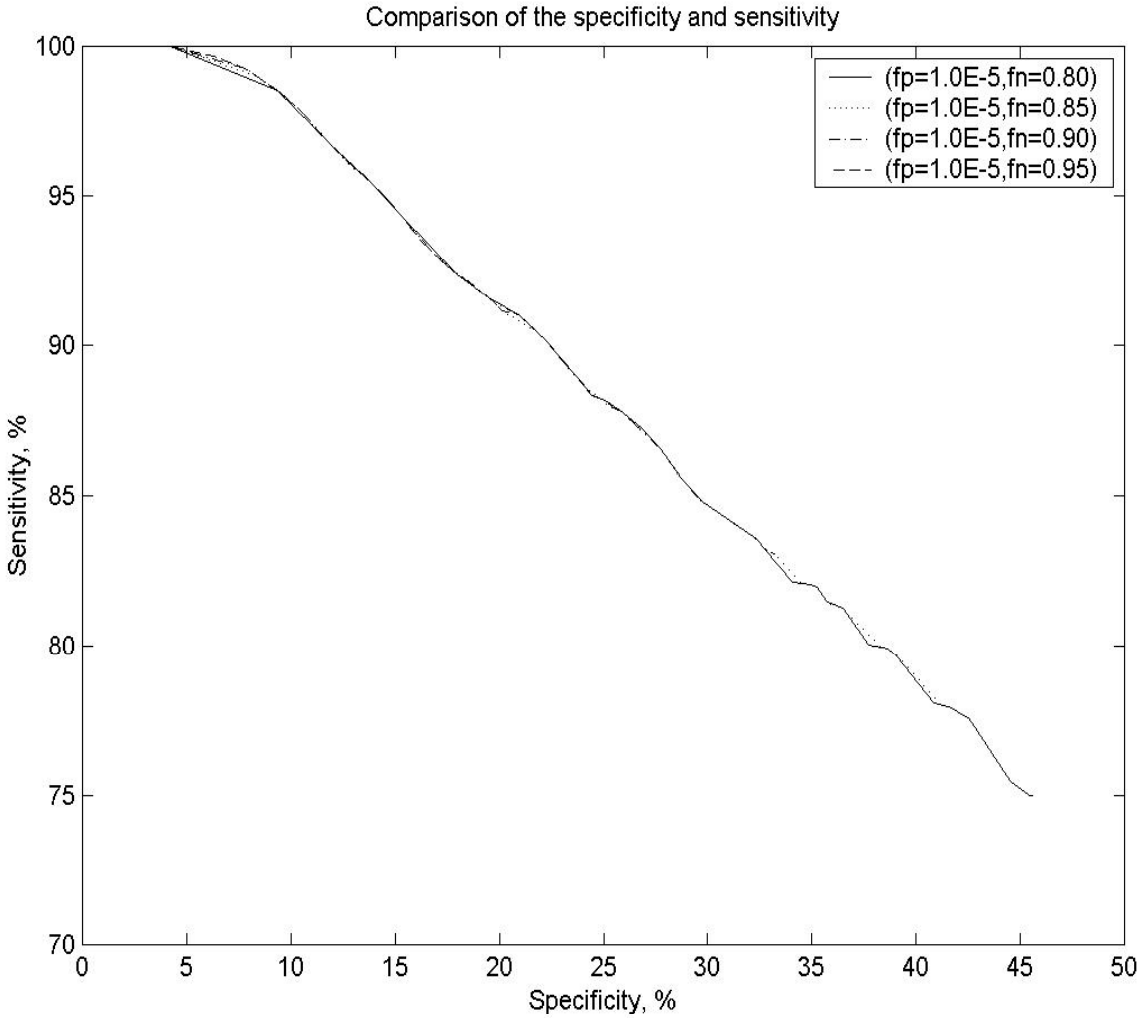
Measure the Accuracy

- Specificity and sensitivity.
- Verification by MIPS physical interactions (as TRUE interactions).
- Relationship between protein-protein interactions and expression data.

$$SP = \frac{\text{number of matches with observation}}{\text{number of prediction}}$$

$$SN = \frac{\text{number of match with observation}}{\text{number of observation}}$$


Comparison of association method and MLE
Association method
MLE with fixed fp=1.0E-5 and fn=0.85
Sensitivity, %
Specificity, %
100
90
80
70
60
50
40
30
0
10
20
30
40
50
60
70
80
90
100


Comparison of the specificity and sensitivity
(fp=1.0E-5,fn=0.80)
(fp=1.0E-5,fn=0.85)
(fp=1.0E-5,fn=0.90)
(fp=1.0E-5,fn=0.95)
Sensitivity, %
Specificity, %
100
95
90
85
80
75
70
0
5
10
15
20
25
30
35
40
45
50

Verification by Known PPIs

- MIPS physical interaction. (Totally 2570 PPIs, 1414 PPIs not overlapping with our training set).
- Compare with random matching.
  - Fold number
  - Larger fold number imply more reliable prediction

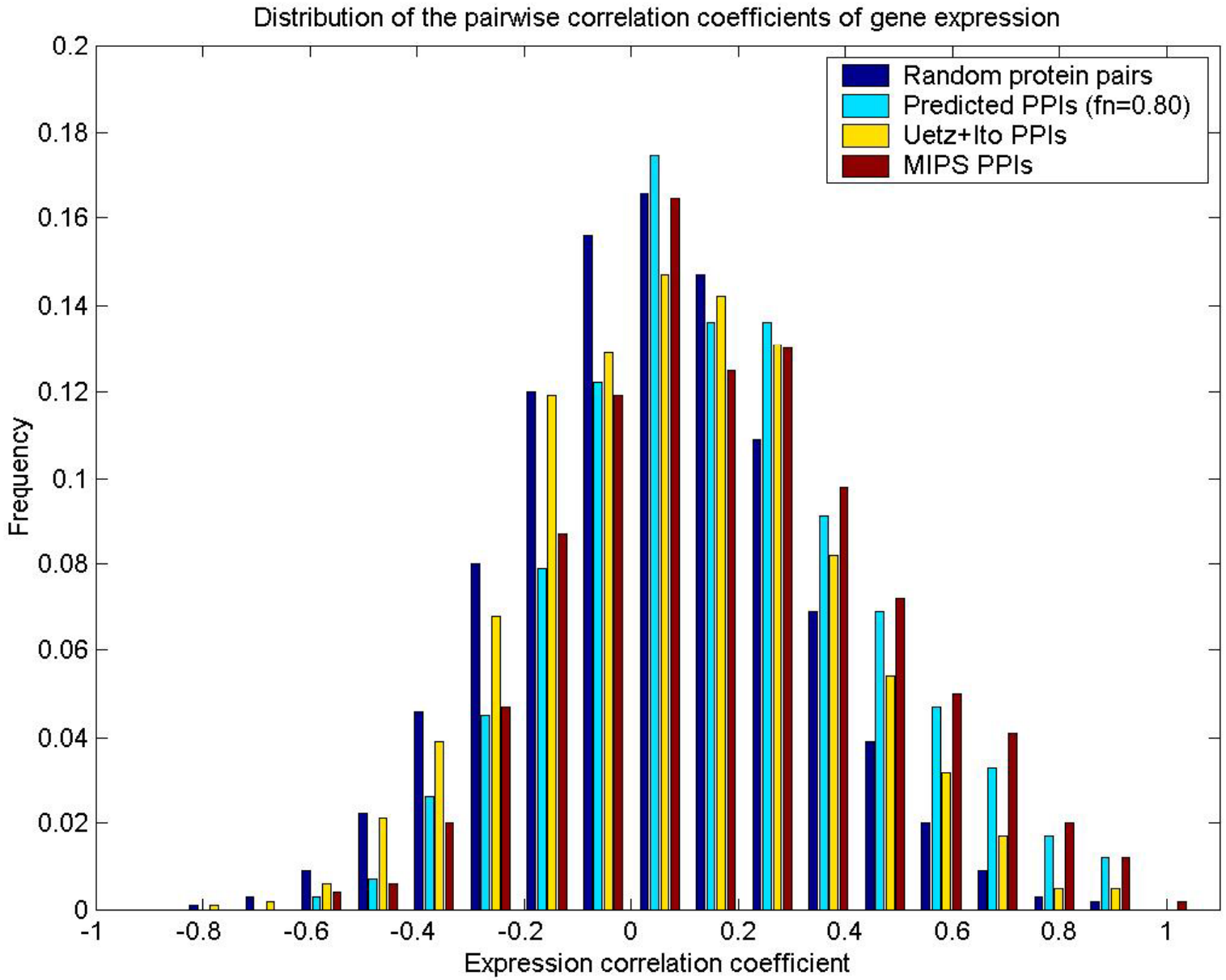
$$\#\,\mathrm{Fold} = \frac{\#\{\text{Our prediction matched to MIPS}\}}{\#\{\text{Expectation of random pairs matched to MIPS}\}}$$

Matching with MIPS PPIs

Prob	#Predict	#Train	#MIPS		#Fold
All	20221620	5719	2570	1414	1.00
>0.00	136463	5719	1265	109	11.92
>=0.20	26908	5238	1093	53	34.97
>=0.40	19360	5018	1035	48	47.85
>=0.60	14725	4775	971	47	67.53
>=0.80	12734	4647	932	43	76.02
>=0.975	10824	4461	886	40	89.88

Interaction Data Correlated With Gene Expression Data

- Interacted proteins seems to have high expression correlation
  - A.Grigorieve *Nucleic Acid Res.* 29, 2001;
  - H. Ge et al. *Nature Genetics* 29, 2001;
  - R. Jansen et al. *Genome Res*.12, 2002.
- Expression data (M.Eisen, 1998); 2465 Yeast ORFs with 79 data points/ORF.
- Pearson correlation coefficient.


Distribution of the pairwise correlation coefficients of gene expression
Random protein pairs
Predicted PPIs (fn=0.80)
Uetz+Ito PPIs
MIPS PPIs
Frequency
Expression correlation coefficient
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

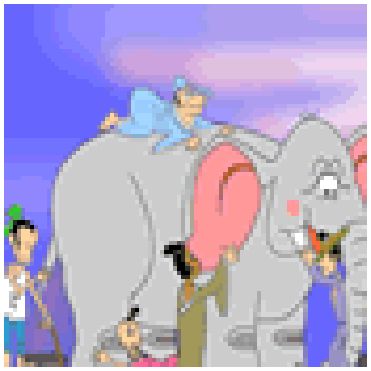
Statistics of Pairwise Correlation of Gene Expression

pairs	# pairs	mean	std	T-score	p-value	$R* > 0.5$
All ORFs	3036880	0.0428	0.2473	0.0000	5.000e-01	3.84%
$\geq$0.20	6392	0.0514	0.2550	2.7984	2.575e-03	4.79%
$\geq$0.40	4433	0.0510	0.2538	2.2232	1.311e-02	4.96%
$\geq$0.60	3318	0.0598	0.2579	3.9644	3.715e-05	5.42%
$\geq$0.80	2756	0.0626	0.2622	4.2196	1.238e-05	5.88%
$\geq$0.975	2266	0.0628	0.2637	3.8482	6.002e-05	5.87%
Uetz+Ito	1307	0.0586	0.2587	2.3213	1.015e-02	5.20%
MIPS	1106	0.1109	0.2767	9.1619	2.706e-20	8.23%

$$T = \frac{\mu_1 - \mu_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}}$$

Discussion

- Independent assumption, high order domain interaction is possible, need more parameters.
- Pfam-A and Pfam-B are in different level.
- Insufficient experimental data.



Related Works

- Inferring domain-domain interactions
  - Riley et al. Genome Biol. 2005.
  - Lee et al. BMC Bioinformatics, 2006

References

- M. Deng, S. Mehta, F. Sun, T. Chen (2002) Inferring domain–domain interactions from protein–protein interactions, Genome Res. 12: 1540–1548.
- H. Lee, M. Deng, F. Sun, T. Chen (2006) An integrated approach to the prediction of domain–domain interactions, BMC Bioinform. 7:269.
- R. Riley, C. Lee, C. Sabatti, D. Eisenberg (2005) Inferring protein domain interactions from databases of interacting proteins, Genome Biol. 6:R89.
- Shoemaker BA, Panchenko AR (2007) Deciphering Protein–Protein Interactions. Part I. Experimental Techniques and Databases. PLoS Comput Biol 3(3): e42. doi:10.1371/journal.pcbi.0030042.
- Shoemaker BA, Panchenko AR (2007) Deciphering Protein–Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. PLoS Comput Biol 3(4): e43. doi:10.1371/journal.pcbi.0030043.
- Hung Xuan Ta, Liisa Holm(2009) Evaluation of different domain-based methods in protein interaction prediction. Biochemical and Biophysical Research Communications 390:357–362.