Predicting Protein Functions and Domain Interactions from Protein Interactions

Fengzhu Sun, PhD

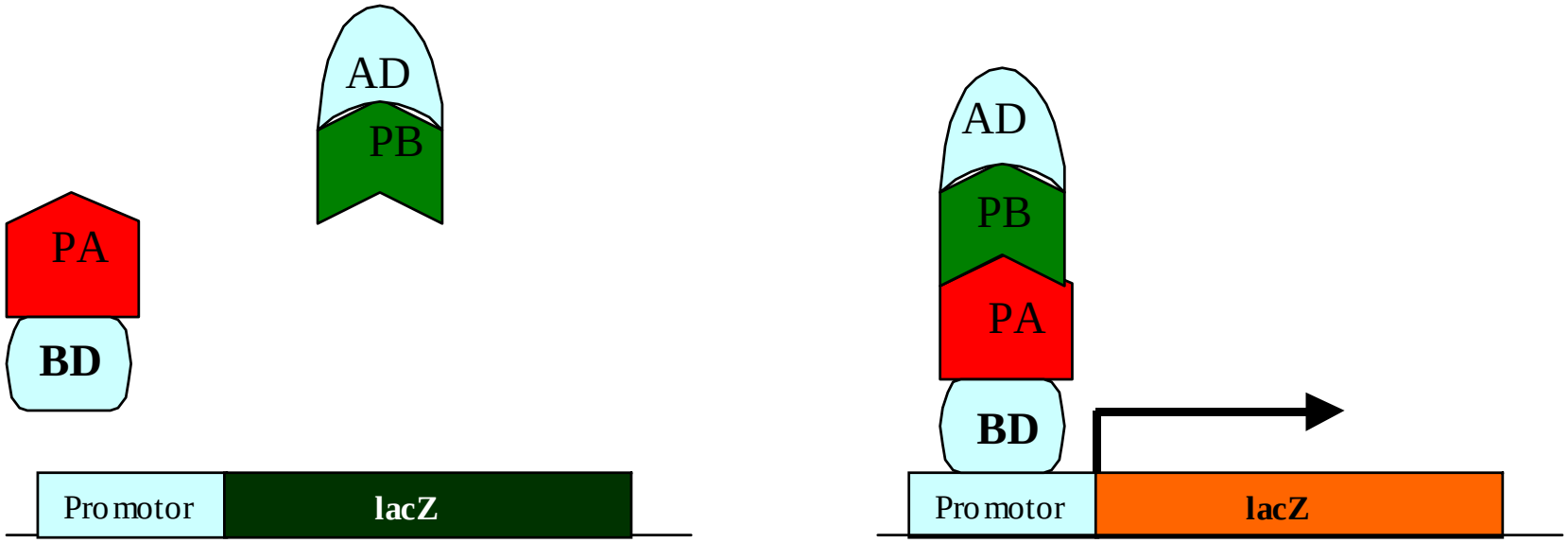
Center for Computational and Experimental Genomics
University of Southern California

Outline

- High-throughput genome technologies
- Genome Databases
- Problems we have studied
- Protein essentiality and genomic information
- Diffusion Kernel based Markov random field (DK-MRF) model
- Results on protein essentiality and GO functions

High-throughput Genome Technologies

Protein physical interactions yeast two-hybrid analysis



BD: Gal4 DNA-binding domain;
AD: Gal4 transcription-activation domain;
lacZ expression detected ➔ PA-PB interact.

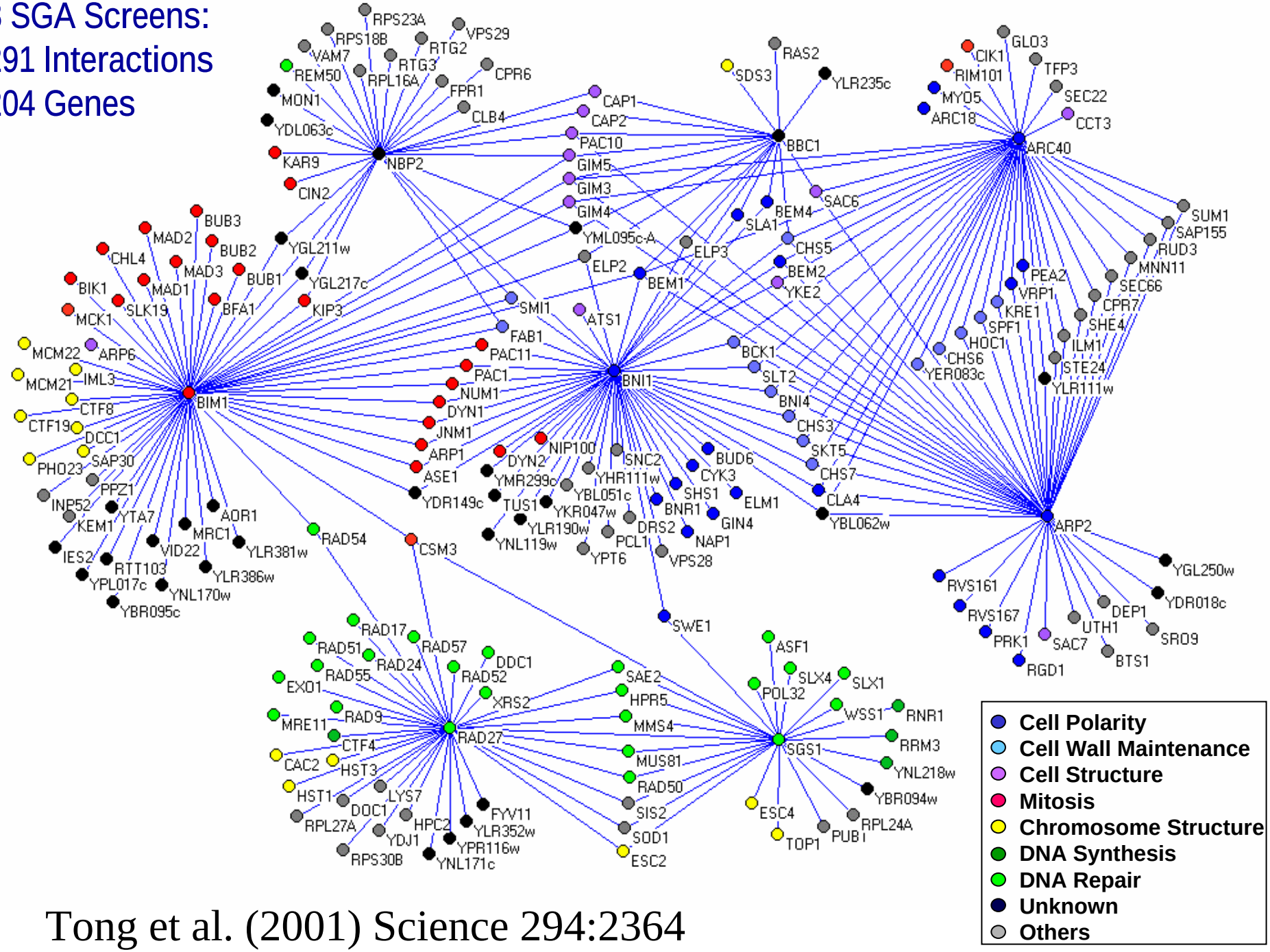
Genetic Interactions

Synthetic lethal mutations

Two mutations are synthetically lethal if cells with either of the single mutations are viable but cells with both mutations are in viable. As with suppressor analysis, synthetic lethal mutations often indicate that the two mutations affect a single function or pathway.

Tong et al. (2001) Science 294:2364

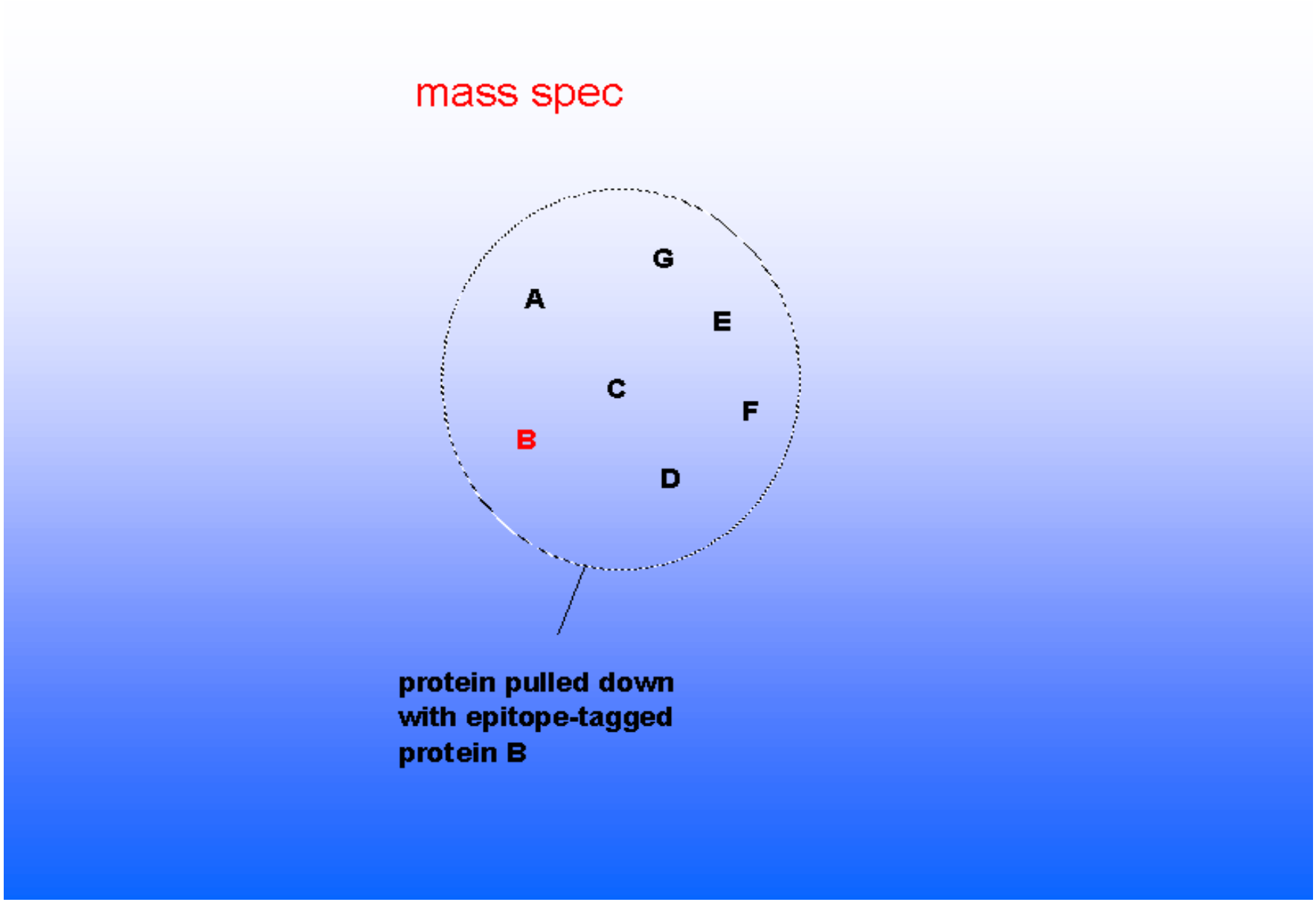
8 SGA Screens:
291 Interactions
204 Genes



Tong et al. (2001) Science 294:2364

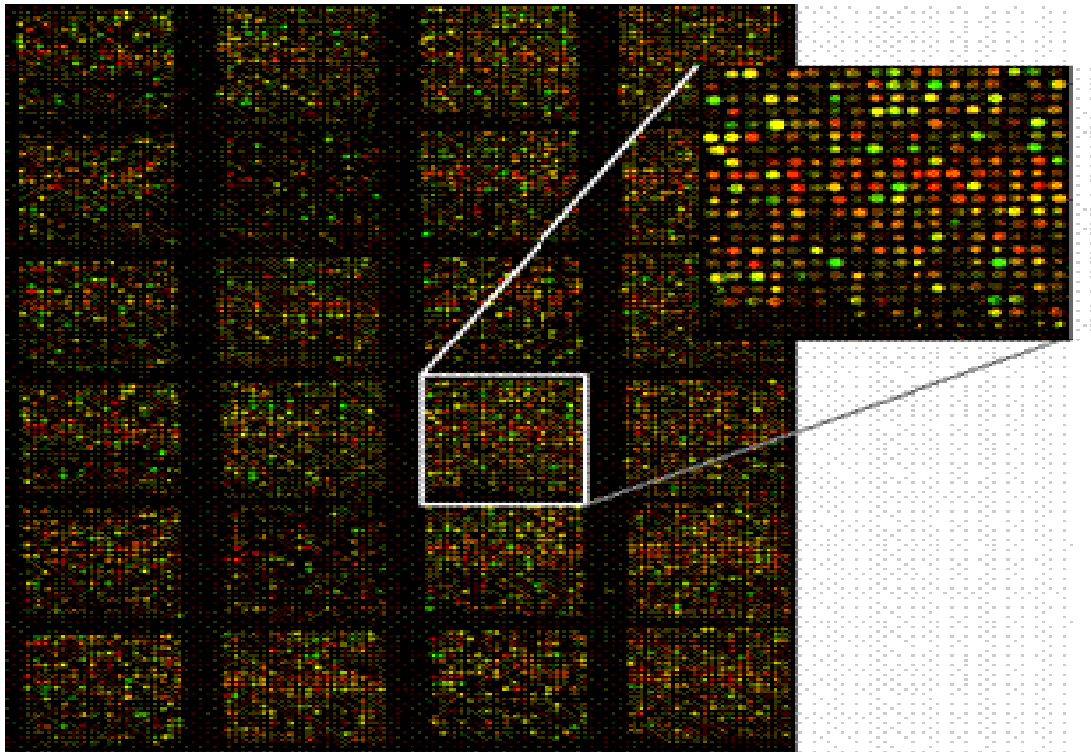
Protein complex identification using Mass Spectrometry

- Attach an affinity tag to many target “bait” proteins
- Introduce the DNA encoding these baits into cells and allowing them to be expressed and form complexes with other proteins
- Proteins extracted with the tagged bait are identified using MS methods



Gavin et al. (2002) Nature 415:141

Microarrays



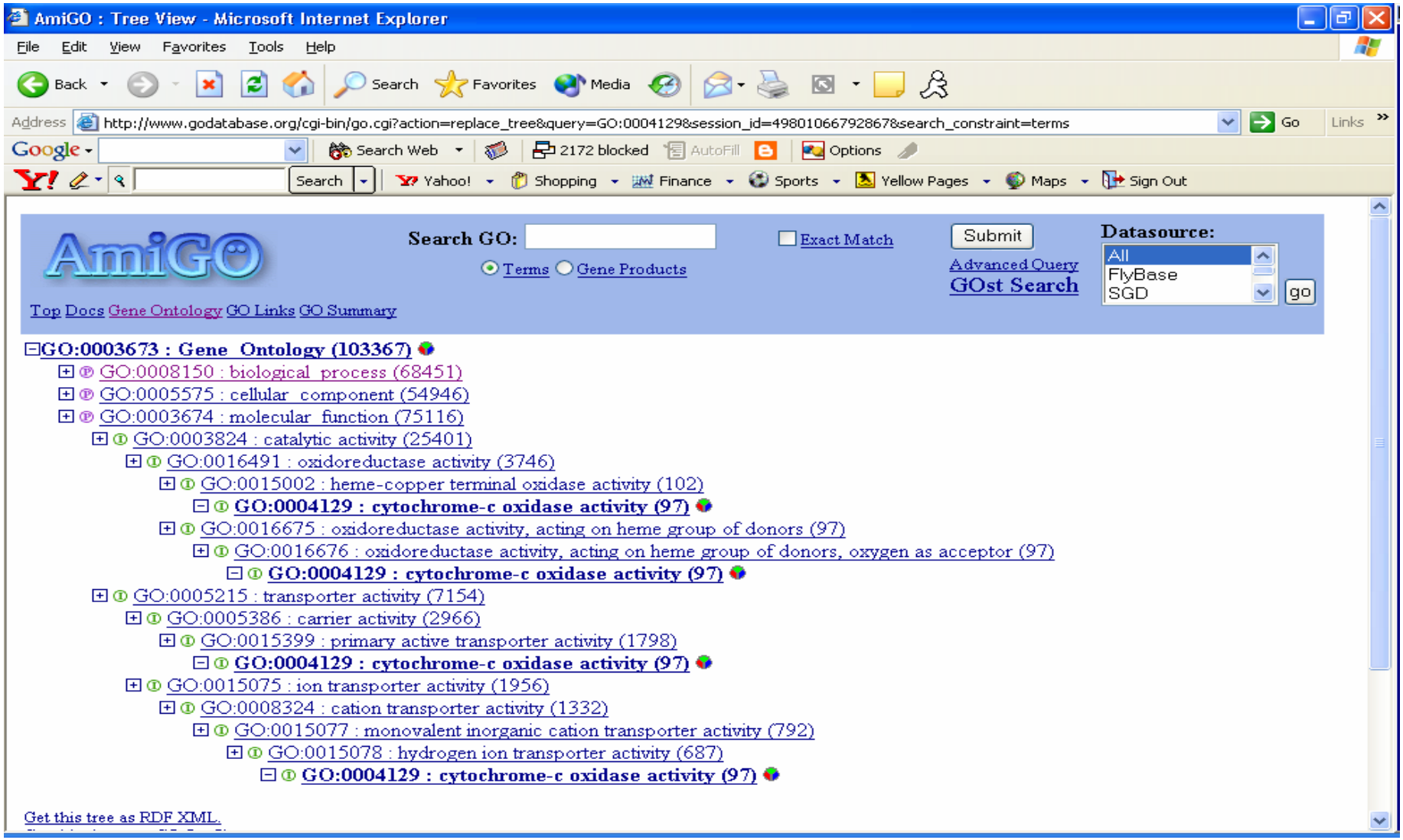
Protein Interaction Databases

- MIPS (mips.gsf.de)
- DIP: Database of Interacting Proteins (dip.doe-mbi.ucla.edu)
- BIND: Biomolecular Interaction Network Database (www.bind.ca)
- GRID: General Repository for Interaction Datasets (biodata.mshri.on.ca/grid)

Gene Ontology

- Includes many databases, including several of the world's major repositories for plant, animal and microbial genomes
- Consists of three structured ontologies that describe gene products in terms of
  - Biological processes
  - Cellular location
  - Molecular functions

GO Functions



Problems we have studied

- Reliability of observed interactions for different technologies (Deng et al. PSB 2003)
- Predicting protein domain interactions from protein interactions from different organisms, phylogenetic profiles, geneontology (Deng et al. Genome Research 2002, Lee et al. BMC Bioinformatics, 2006)
- Mixture models for stochastic network motifs in stochastic protein interaction networks and gene regulation networks (Jiang et al. PNAS 2006)

Problems we have studied

- Gene prioritization by integrating gene expression and protein interactions (Ma et al. Bioinformatics 2006)
- Integrative approaches for identifying candidate genes and gene regulation pathways (Tu et al. ISMB 2006)
- Protein function prediction by integrating protein interactions, gene expressions, domains, etc (Deng et al. JCB, 2003, 2004, Lee et al. OMICS 2006)

The relationship Between Protein Essentiality and Genomic Information

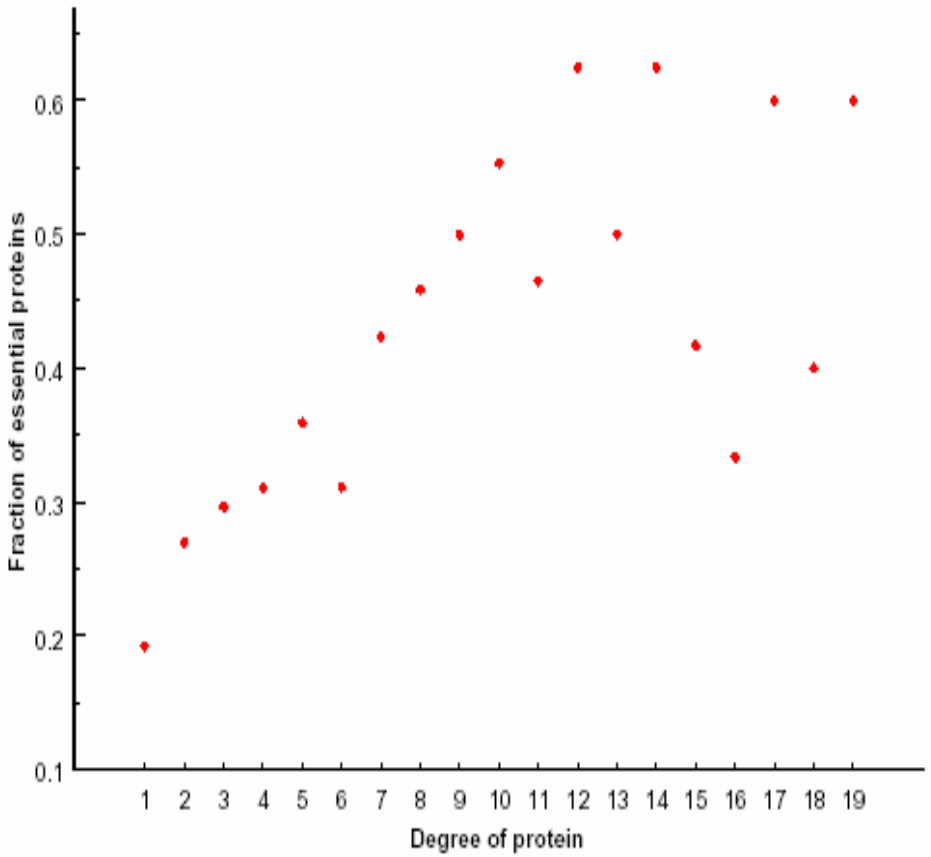
Protein Essentiality

Essential Genes are those when mutated or knocked out under given conditions, the cell will not survive. Other genes are referred as viable or non-essential

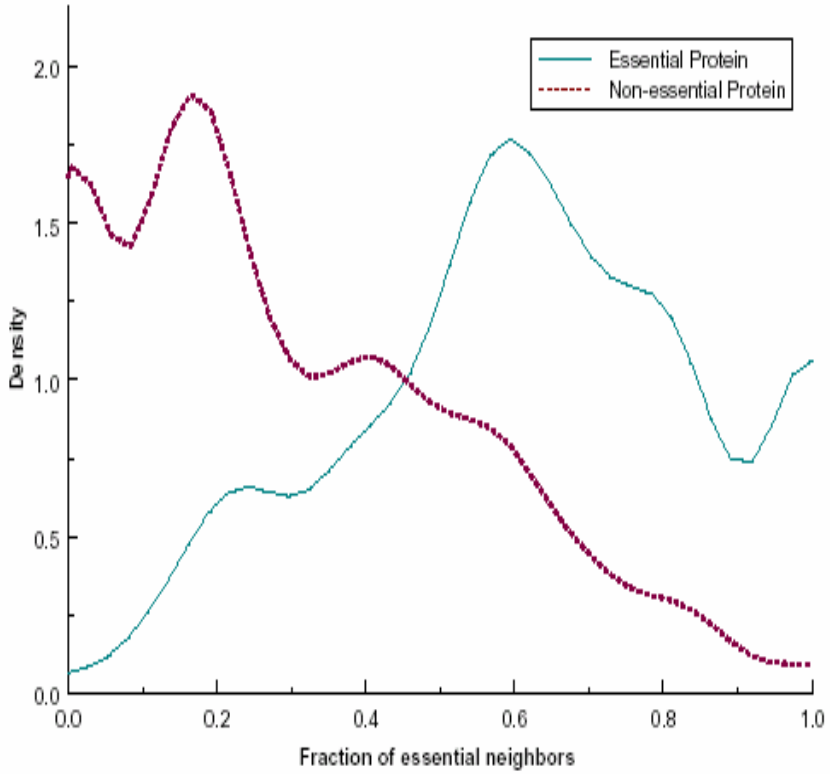
Question: What are the relationships between protein essentiality and genomic information?

Protein essentiality vs interaction

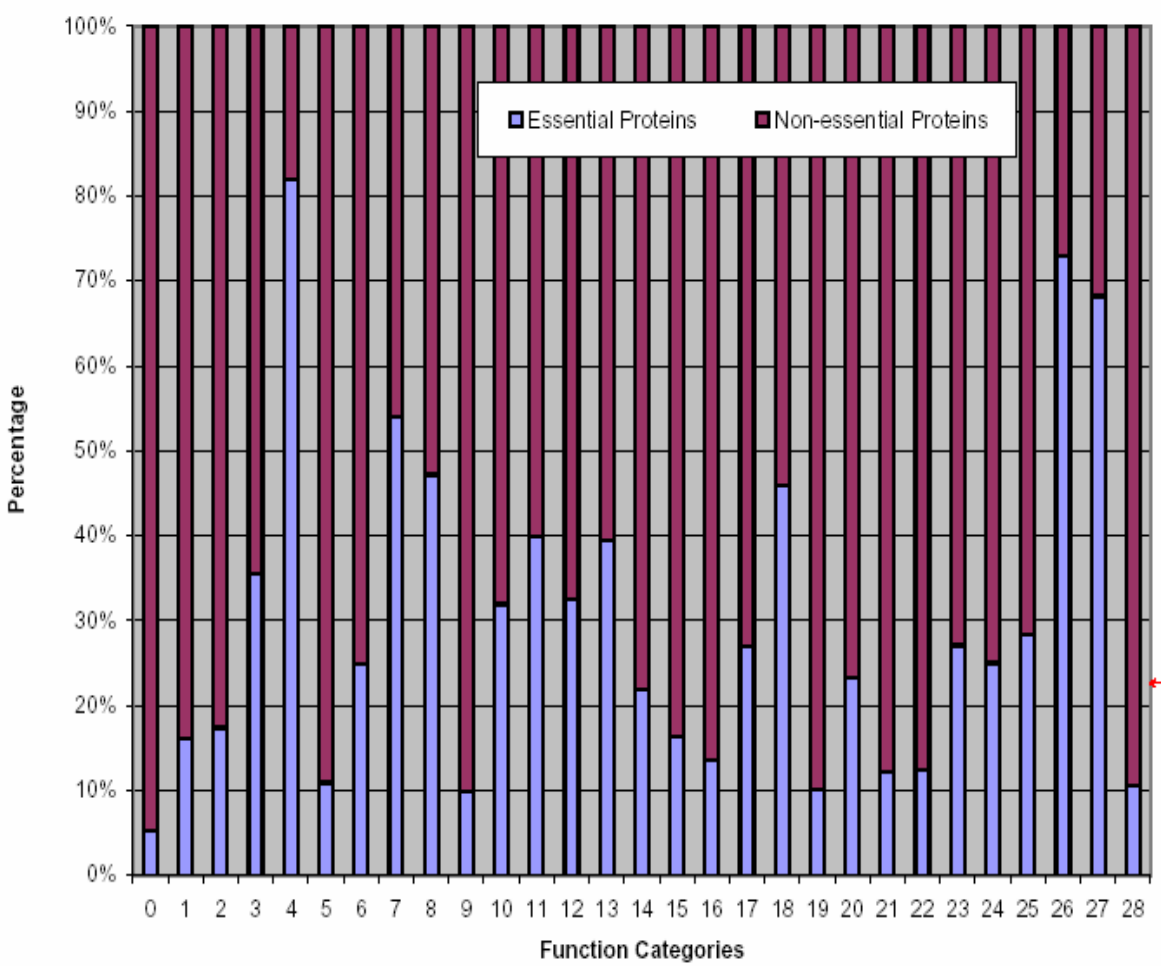
Essentiality vs degree



Essentiality vs fraction of essential neighbors

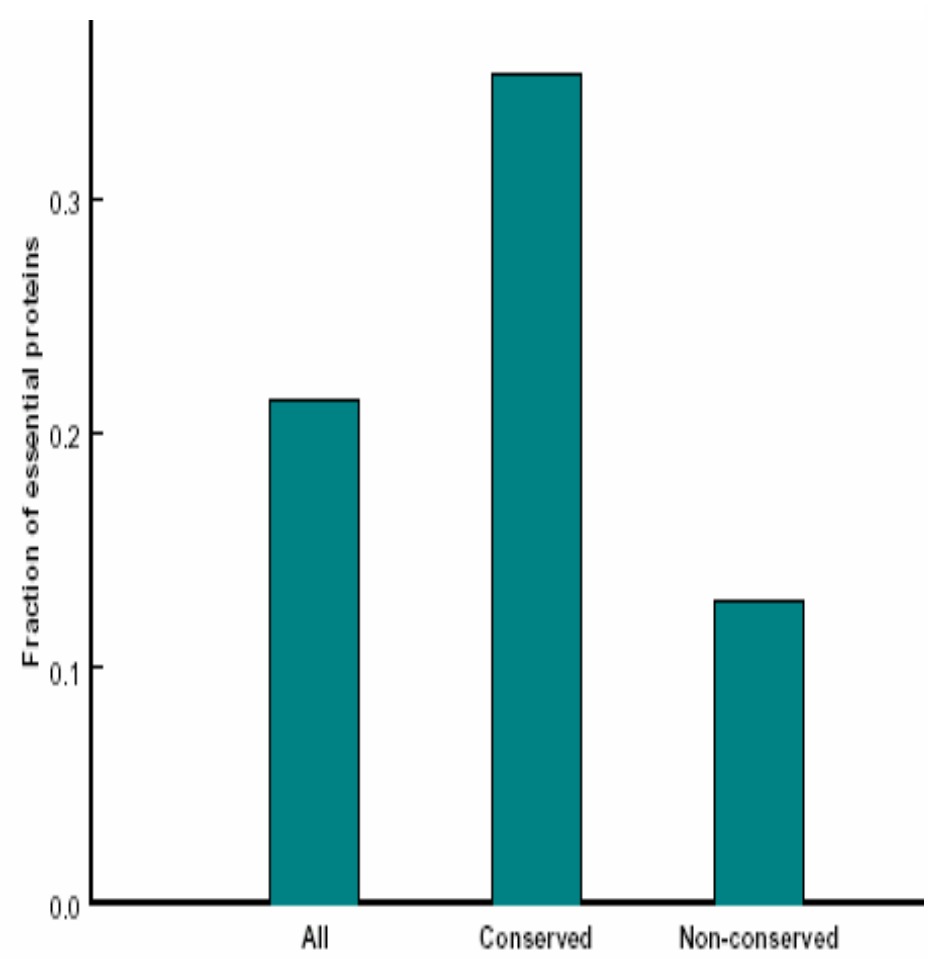


Essentiality vs GO Biological Processes

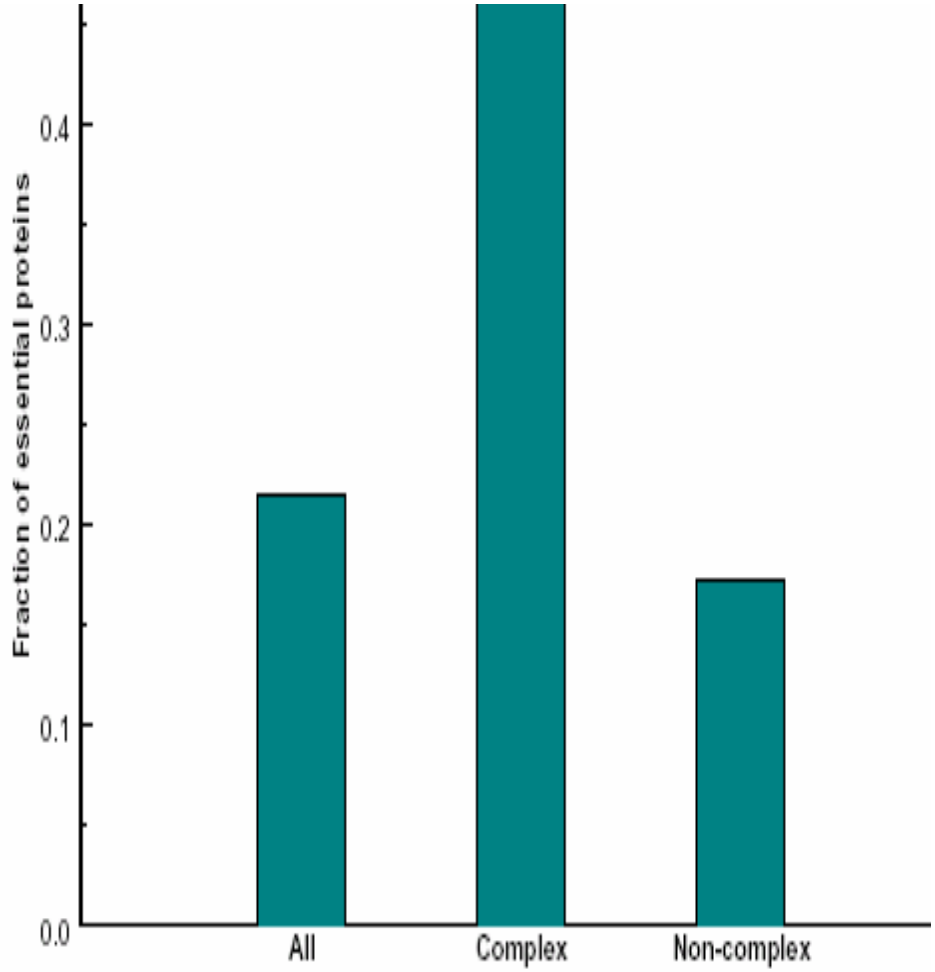


Essential Proteins are Significantly enriched in

- 4(rRNA processing)
- 26(mRNA processing)
- 27(RNA splicing)
- 7(DNA replication)
- 8(mitosis)



Conserved proteins are more likely to be essential



Proteins in complexes are more likely to be essential

Understanding Protein Essentiality

Two Powerful Methods

MRF: Build a MRF model for functional assignment and calculate the posterior probability assignment for unknown proteins

Pros: probability assignment, study the contribution of each factor

Cons: prediction accuracy lower than SVM

SVM: Define protein similarity based on diffusion kernel for a network. Then use SVM to define a classifier

Pros: higher prediction accuracy

Cons: not easy to evaluate the contribution of each factor

Basic Ideas

- Constructing a network with proteins as nodes and interactions as edges
- Define a diffusion kernel on the network
- A Markov Random Field using the diffusion kernel
- MCMC to assign protein essentiality for unknown proteins

Diffusion Kernel on a Network
(Kondor and Lafferty 2002)

$$K(i,j) = \{e^{\tau H}\}_{i,j}$$

$$H = \begin{cases} 1 & \text{if protein } i \text{ interacts with protein } j; \\ -d_i & \text{if protein } i \text{ is the same as protein } j; \\ 0 & \text{otherwise.} \end{cases}$$

d_i is the degree of protein i

Advantages of diffusion kernels

- Define similarity between all protein pairs, not just immediate neighbors
- Take degrees of nodes into consideration
- The parameter τ can be adjusted to achieve high prediction accuracy

Diffusion Kernel Based Markov Random Field (DK-MRF)

- Motivation: The kernel matrix defines similarity between proteins. Similar proteins tend to have similar functions.
- Without interaction data, the probability of functional labeling

$$\Pr(X) \propto \pi^{N_1}(1-\pi)^{N_0}$$

N_1: # proteins with the function

N_0: # proteins without the function

π : fraction of proteins with the function

X_i = 1 if the ith protein is essential and o otherwise

- Given the interaction data, our believe for the network is proportional to

$$\exp(\alpha N_1 + \beta_{10} D_{10} + \beta_{11} D_{11} + \beta_{00} D_{00}),$$

$$N_1 = \sum_i I\{x_i = 1\},$$

$$D_{11} = \sum_{i<j} K(i,j) I\{x_i = 1, x_j = 1\},$$

$$D_{10} = \sum_{i<j} K(i,j) I\{(x_i = 1, x_j = 0) \text{ or } (x_i = 0, x_j = 1)\},$$

$$D_{00} = \sum_{i<j} K(i,j) I\{x_i = 0, x_j = 0\}.$$

DK-MRF continued

The prior distribution for X is

$$\Pr(X \mid \theta) = \frac{1}{Z(\theta)} \exp(-U(x))$$

The potential function is

$$U(X) = -(\alpha N_1 + \beta_{10} D_{10} + \beta_{11} D_{11} + \beta_{00} D_{00})$$

Conditional Marginal Distribution

$$\log \frac{\Pr(X_i = 1 \,|X_{[-i]}, \theta)}{1 - \Pr(X_i = 1 \,|X_{[-i]}, \theta)}$$
$$= \alpha + (\beta_{10} - \beta_{00}) K_0(i) + (\beta_{11} - \beta_{10}) K_1(i).$$

$$K_0(i) = \sum_{j \neq i} K(i,j) I\{x_j = 0\},$$
$$K_1(i) = \sum_{j \neq i} K(i,j) I\{x_j = 1\}.$$

MRF vs DK-MRF

If

$$K(i,j) = \begin{cases} 1, & \text{i, j are neighbors} \\ 0, & \text{otherwise} \end{cases}$$

then the DK-MRF model reduces to the MRF model

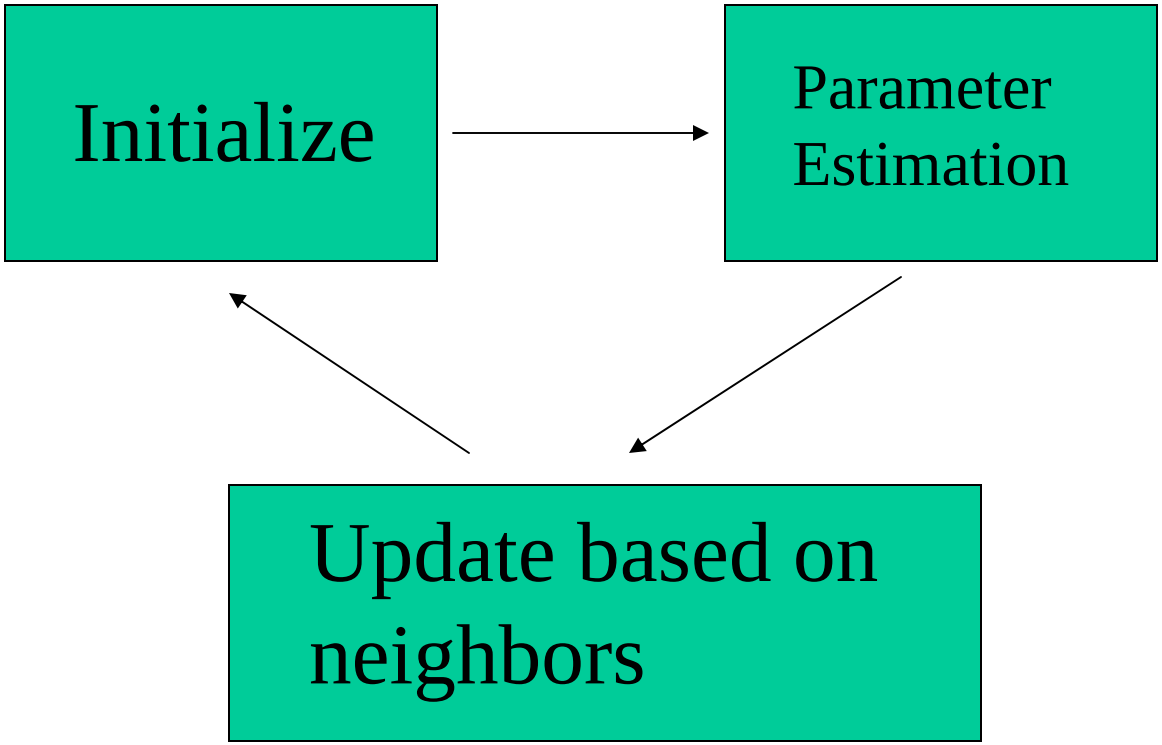
Problems

- Estimate the parameters in the model

$$\theta = (\alpha, \beta_{10} - \beta_{00}, \beta_{11} - \beta_{10})$$

- Estimate the posterior probability of the essentiality for the unknown proteins conditional on the functions of the annotated proteins and the network

Markov Chain Monte Carlo



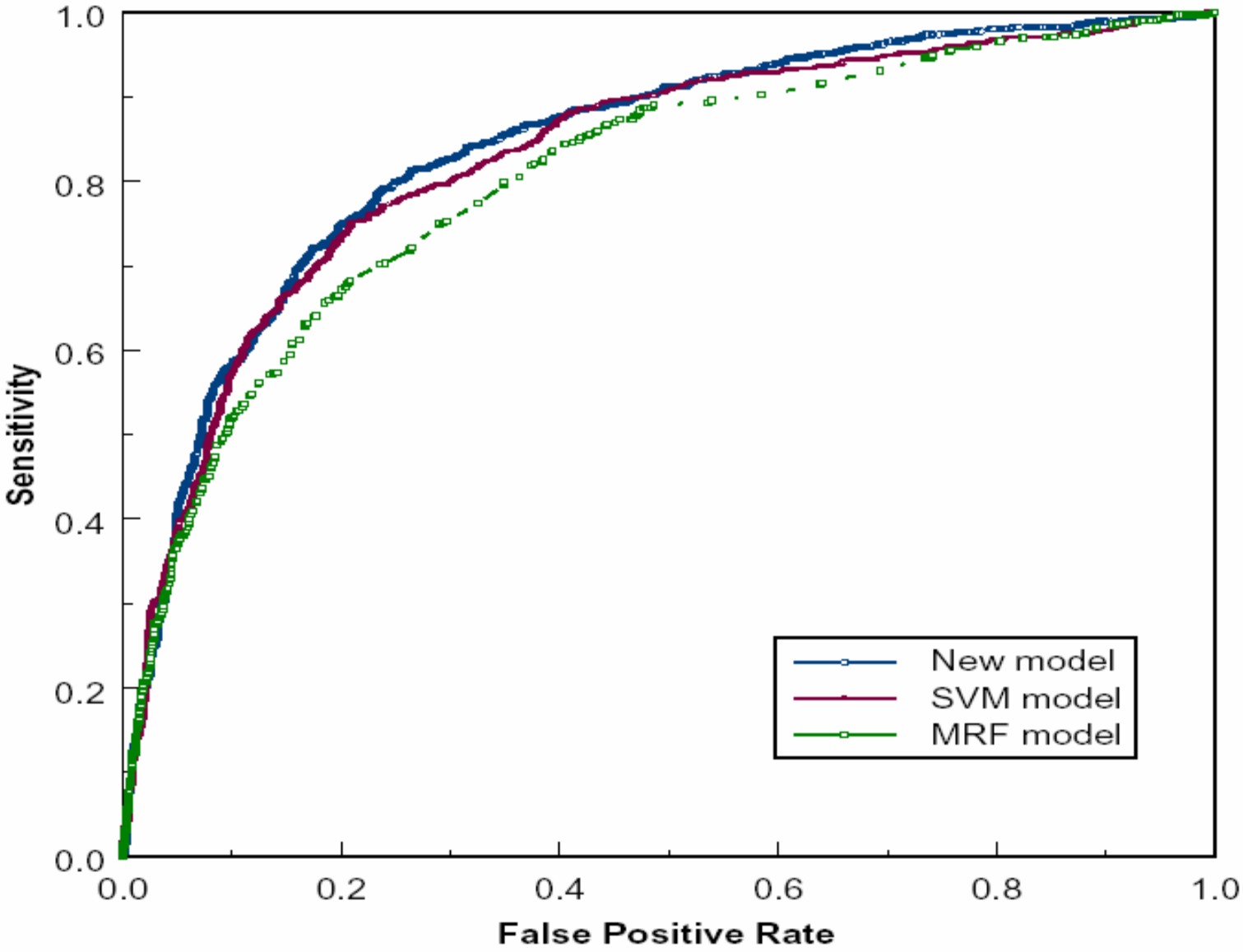
Cross-validation of known proteins

Leave-one-out test: for every known protein, assume it as unknown, predict its functions using the DK-MRF model, and compare the predictions with the annotations.

Annotation	Prediction	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

false positive rate = FP/(TN+FP)

sensitivity = TP/(TP+FN)


1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
New model
SVM model
MRF model

Extending MRF to Incorporate Other Data Sources

- Interactions - pairwise
  - Physical Interaction (MIPS)
  - Highly correlated Gene Pairs (SGD)
- Complex/Cluster - groups
  - Protein Complex (TAP)
  - Gene Cluster
- Sequence Features
  - Protein domains (Pfam)

- Networks are incorporated through diffusion kernels, K

- Simple features, e.g conservation, complex, are incorporated using linear terms

- Complex features, e.g GO functions, domains, are first used to define kernels and then incorporated into the model

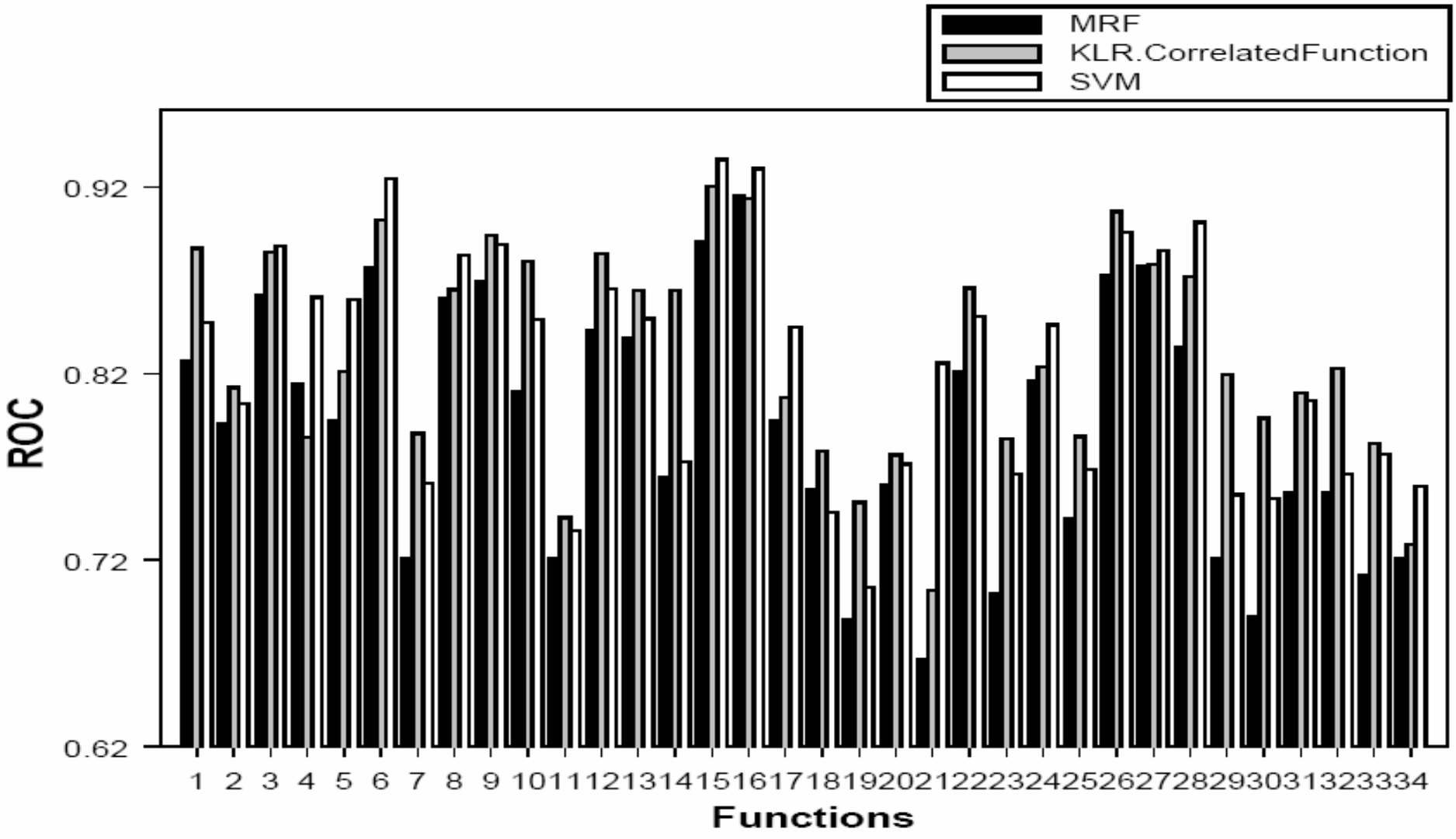
Diffusion kernel based logistic regression model

$$\log \frac{\Pr(X_i = 1 | \text{ network, features})}{1 - \Pr(X_i = 1 | \text{ network, features})} = \alpha + \beta_+ \sum_{j:X_j=1, j\neq i} K(i,j) + \beta_- \sum_{j:X_j=0, j\neq i} K(i,j)$$
$$+ \sum_{k=1}^{u} \lambda_k \mathbf{I}_i^{f_k} + \sum_{s=1}^{v} \left(\theta_{s+} \sum_{j:X_j=1, j\neq i} \mathbf{f}_i^s \cdot \mathbf{f}_j^s + \theta_{s-} \sum_{j:X_j=0, j\neq i} \mathbf{f}_i^s \cdot \mathbf{f}_j^s \right)$$

Feature selection

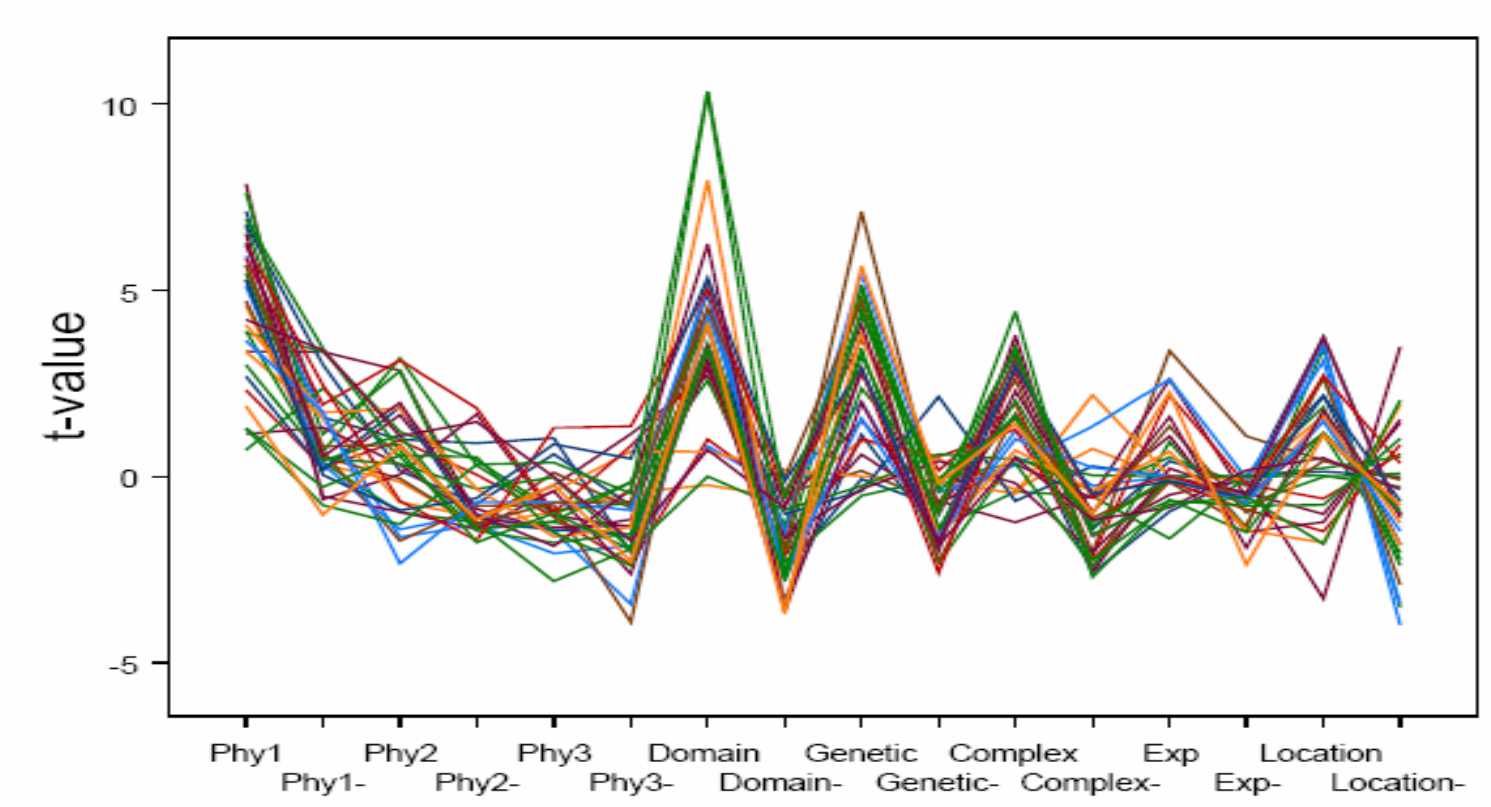
1st Selection		**2nd Selection**		**3rd Selection**	
feature	ROC score	feature	ROC score	feature	ROC score
Interaction	0.842	Domain	0.846	Domain	0.859
Domain	0.639	Function	0.857	Location	0.860
Function	0.732	Location	0.851	Complex	0.858
Location	0.689	Complex	0.842	Conservation	0.867
Complex	0.602	Conservation	0.855		
Conservation	0.651				
Selected feature	interaction	Selected feature	function	Selected feature	conservation
4th Selection		**5th Selection**		**with all features**	
feature	ROC score	feature	ROC score		ROC score
Domain	0.867	Domain	0.853		0.855
Location	0.869	Complex	0.869		
Complex	0.867				
Selected feature	location	Selected feature	-		

ROC scores for 34 GO functions



Contribution of data sources

- A KLR approach explore the contribution of neighbors to the functions of proteins of interest.
  - The t-value based on the t-test statistic for the hypothesis that the parameter of interest is zero



A collaborative blind study on mouse functions

- Protein interactions, domains, expression profiles, phylogenetic profiles, disease relation, phenotypes, are provided for all the 21K mouse proteins
- Training: GO functions for part of the known proteins
- Test: Some known proteins are retained for validations
- Novel: new discoveries since Feb. 2006

Objective: The usefulness and limitations of current methods applied to a high organism. Nine groups participated in this exercise

Results

	Biological Process		Cellular component		Molecular function	
	31-100 (239)	>100 (100)	31-100 (48)	>100 (30)	31-100 (111)	>100 (35)
Validation	85	84	86	83	94	93
Test	85	85	85	85	89	90
Novel	71	71	73	68	80	83

The average ROC score (*100) for GO functional categories with at least 31 genes based on cross-validation, the blind test set, and novel gene set experimentally verified since Feb. 2006.

Surprisingly the prediction accuracy for Mouse is similar to that for the Yeast.

Conclusions

- Developed a diffusion kernel based MRF model for protein function prediction
- The KLR model has similar performance as SVM approach, but the model is much simpler with only three parameters
- Integrating other data sources can increase the performance the protein function prediction
- As a model based approach, KLR can be used to study the contribution of each data source

Related Work

- Letovsky and Kasif, Bioinformatics 19: suppl 1197-1204

  The same as MRF, details differ (MIPS)

- Karaoz U et al. PNAS, March 2, 2004

  Application to GO

- Vazquez et al. Nature Biotechnology, 2003

  Considered multiple functions using ideas similar to MRF

- Lanckriet et al. PSB 2004

  Used SVM for function prediction

Predicting Protein Domain Interactions

Why study domain-domain interaction?

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

Notation

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.

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

Assumptions

- 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.

Combine possible errors of data into the model

False positive

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

False negative

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

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).

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

Parameters re-estimation

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.

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

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

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

Observed interactions: T=5719;

Proteins: N=6359 (SGD).

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

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

We choose fp=1.0E-5, fn= 0.8-0.95.

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).

Acknowledgements

- Co-advised with Tim Ting Chen
- Supported by NSF, NIH and USC
- Zhidong Tu, Minghua Deng, Hyun-ju Lee
- The Mouse function project is organized by F. Roth and T. Hughes

References

- Hyunju Lee, et al. (2006) *BMC Bioinformatics 7:269*
- Hyunju Lee, et al. (2006) *OMICS: A Journal of Integrative Biology*, 10:40-55
- Deng MH, et al. (2004) *Bioinformatics* 20:895-902
- Deng MH, et al. (2004) *Journal of Computational Biology*, 11:463-476
- Deng MH, et al. (2002) *CSB2002*, 117-126. Also on *Journal of Computational Biology*, 10:947-960
- Deng MH, et al. (2002) *RECOMB2002:*117-126. Also on *Genome Research* **12:**1540-1548.