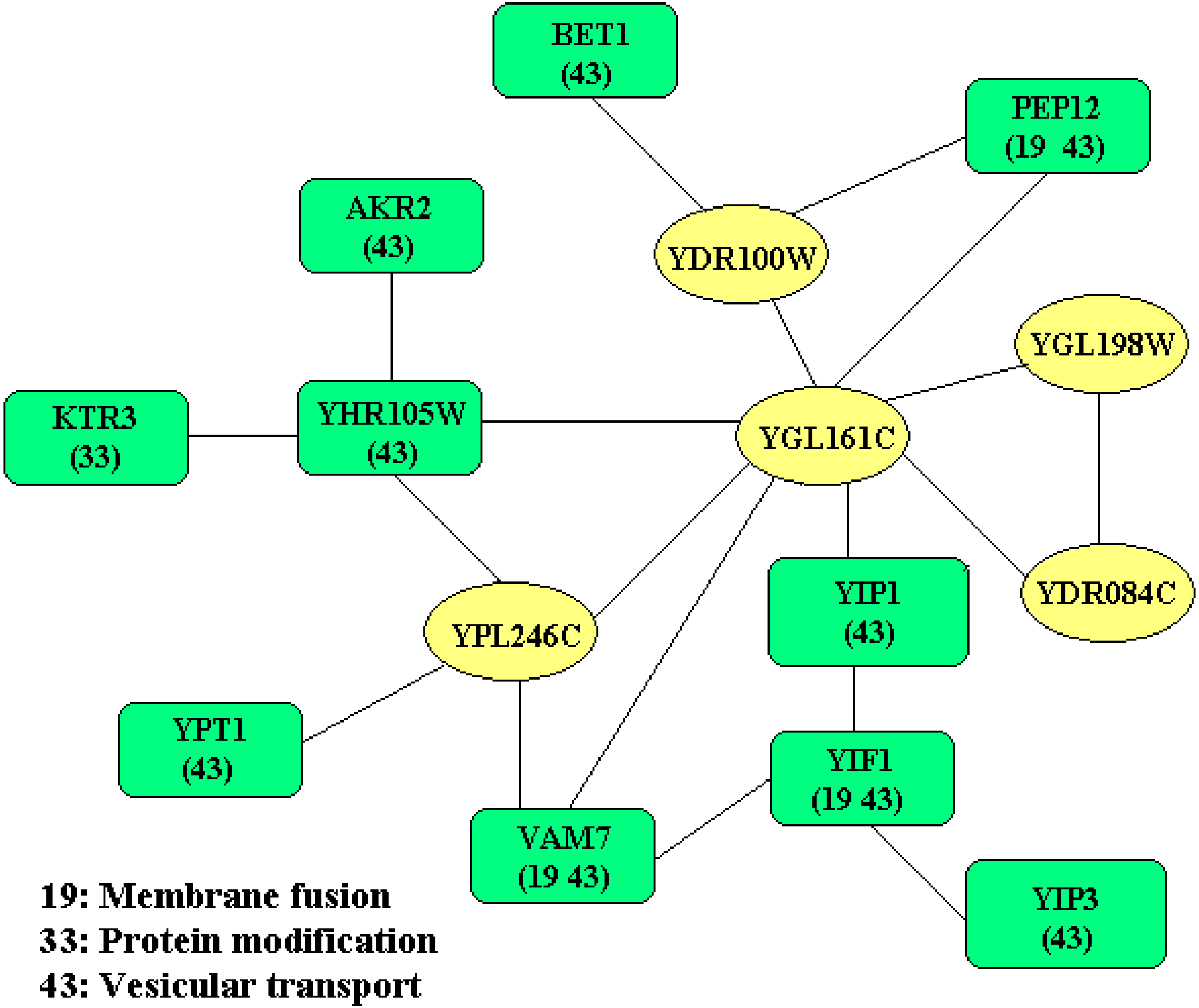
第8-2章: Network Based Inference

- Network based function prediction
  - Markov random field model
  - Kernel logistic regression method
- Network based gene expression study
  - Hidden Markov random field



BET1 (43)

PEP12 (19 43)

AKR2 (43)

YDR100W

YGL198W

KTR3 (33)

YHR105W (43)

YGL161C

YIP1 (43)

YDR084C

YPL246C

YPT1 (43)

YIF1 (19 43)

VAM7 (19 43)

YIP3 (43)

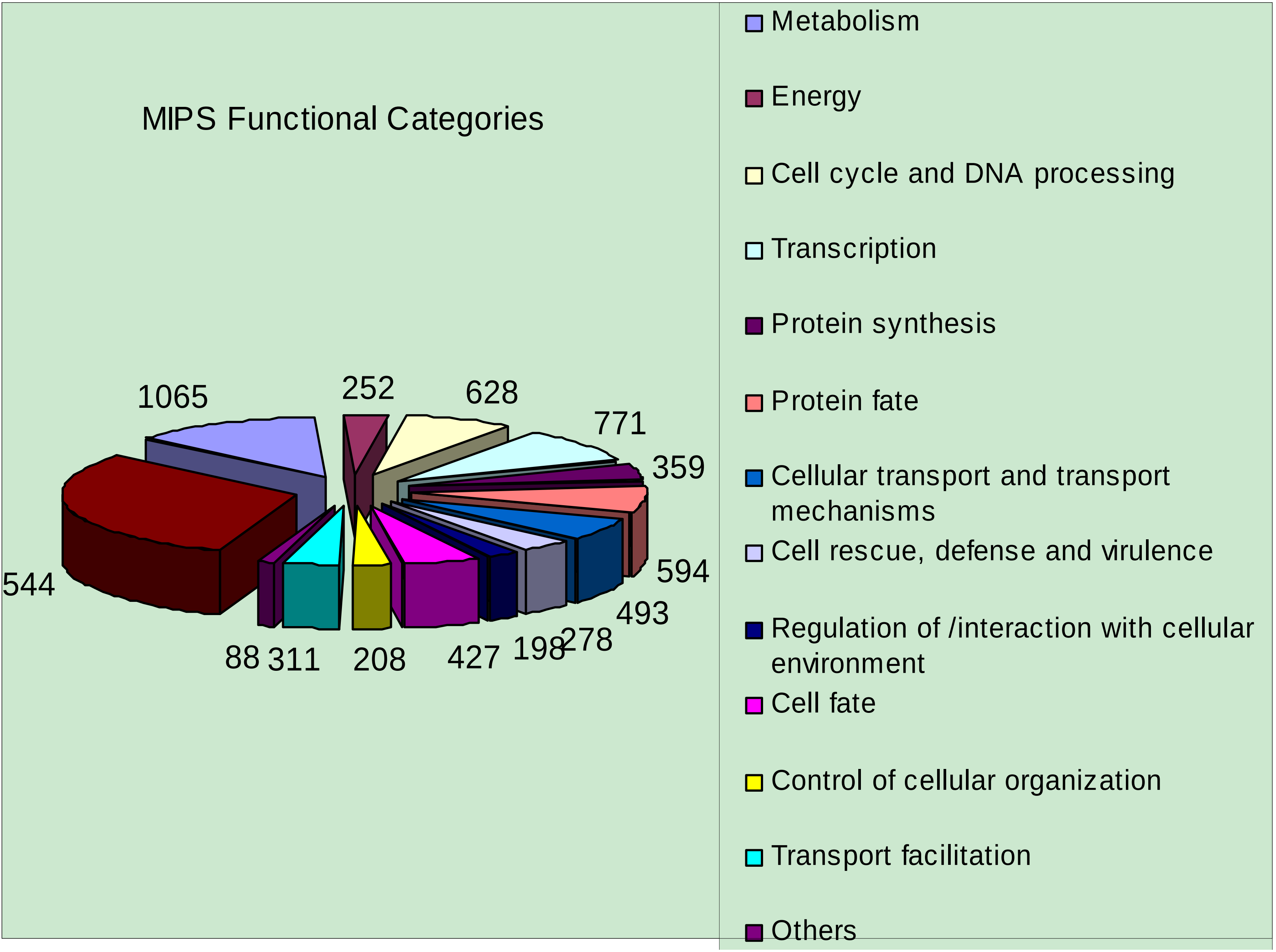
19: Membrane fusion
33: Protein modification
43: Vesicular transport

Protein Functions

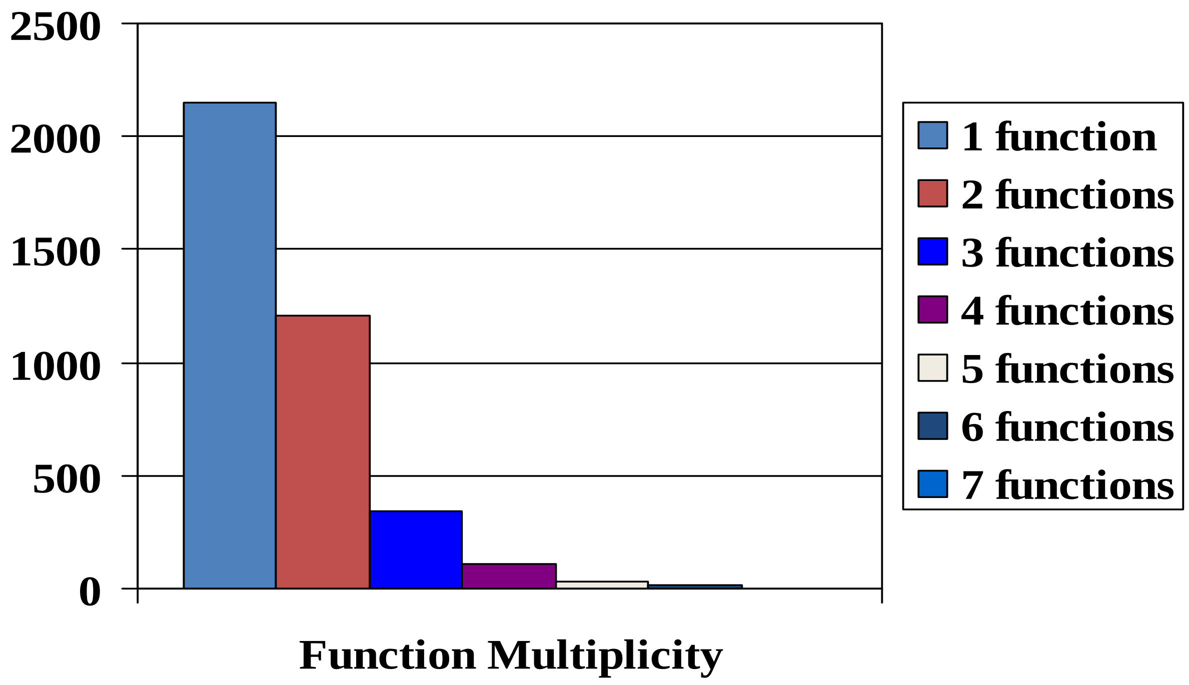
- YPD (Yeast Proteome Database).
- MIPS (Munich Information Center for Protein Sequences).
- GO (Gene Ontology).

Hierarchical: MIPS, GO.

Non-hierarchical: YPD.


MIPS Functional Categories
1065
252
628
771
359
594
493
278
198
427
208
311
88
2544
Metabolism
Energy
Cell cycle and DNA processing
Transcription
Protein synthesis
Protein fate
Cellular transport and transport mechanisms
Cell rescue, defense and virulence
Regulation of /interaction with cellular environment
Cell fate
Control of cellular organization
Transport facilitation
Others

Multiple Function Problem



We consider each function separately. 1-0 problem

Functional Clues

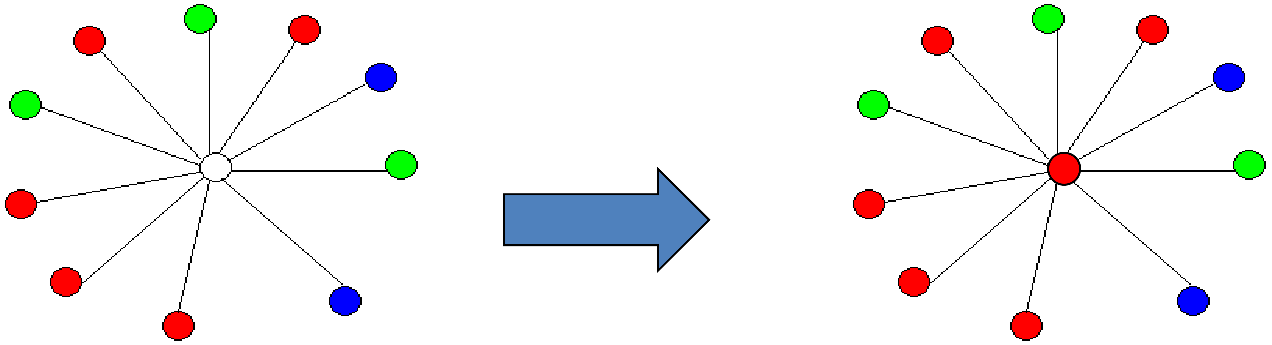
- Sequence similarity (BLAST).
- Domain, motif, transcriptional factor binding, as well as some signals in sequence.
- Phenotype data.
- Gene expression (clustering)
- Protein-protein interactions

Protein-protein Interactions

- DIP (Database of Interacting Protein).
- BIND (The Biomolecular Interaction Network Database).
- GRID (The General Repository for Interaction Datasets).
- MIPS physical interaction.

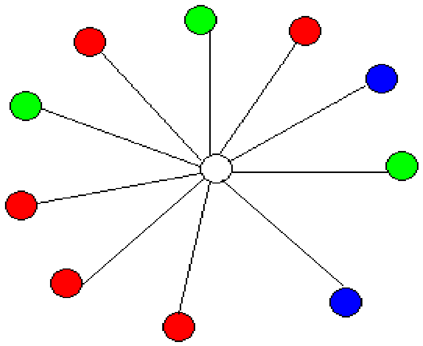
Basic Idea

- Interacted proteins are more likely to be in the similar function group, so we can infer the function of the unknown proteins by their interaction partners.



Two Basic Methods

- Neighborhood-counting method (Schwikowski 2000)
- Chi-square method (Hishigaki 2001)

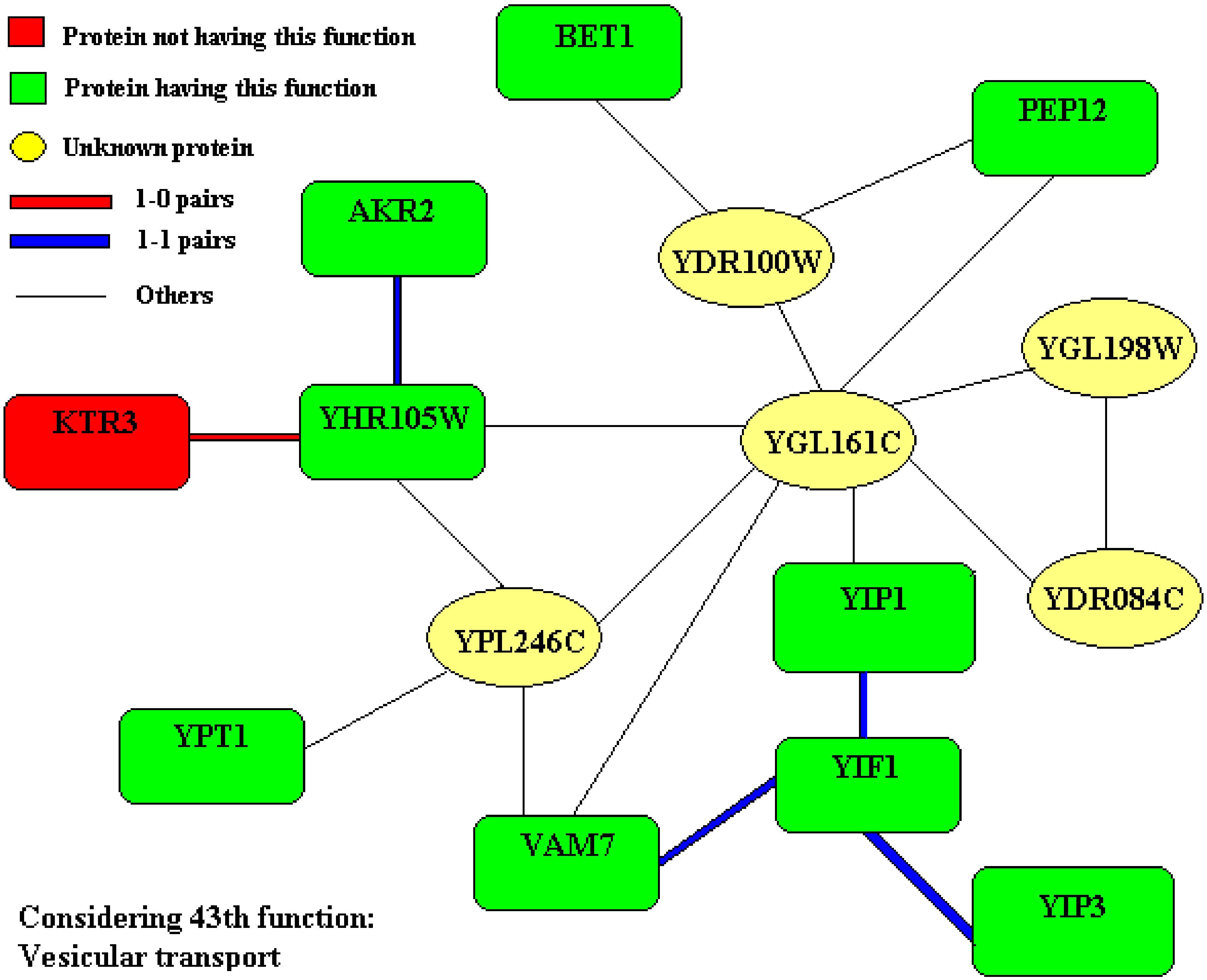


$$S_i = \frac{(n_i - e_i)^2}{e_i}$$

$$e_i = p_i \sum_j n_j$$

Where p_i is the fraction of protein having the ith function in the whole map.

Local → Global



Considering 43th function:
Vesicular transport

Markov Random Field (MRF)

- Motivation: proteins involved in same functional process are likely to interact.
- Without interaction data, the probability of functional labeling

N_1: # proteins with the function

N_0: # proteins without the function

π: fraction of proteins with the function

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

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

$$a^{N_{11}} b^{N_{10}} c^{N_{00}}$$

N$_{ij}$: # interacting protein pairs one with annotation *i* and the other *j*

- Combining the above two equations, the Gibbs distribution is

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

- The potential function

$$U(x) = -\alpha N_1 - \beta N_{10} - \gamma N_{11} - N_{00}$$

Mathematical Problems

- Estimate the parameters in the model

$$\theta = (\alpha, \beta, \gamma)$$

- Estimate the posterior probability of the annotations for the unannotated proteins conditional on the functions of the annotated proteins

Estimation of Parameters

- Using sub-network containing annotated proteins only.
- A Logistic model can be derived (maximum psudo-likelihood estimation).

$$\log \frac{Pr(X_i = 1|X_{[-i]}, \theta)}{1 - Pr(X_i = 1|X_{[-i]}, \theta)}$$
$$= \alpha + (\beta - 1)M_0^{(i)} + (\gamma - \beta)M_1^{(i)}$$

Logistic Regression

Observed data:

Response	X_1	X_2	…	X_n
y_1	x_{11}	x_{12}	…	x_{1n}
y_2	x_{21}	x_{22}	…	x_{2n}
…	…	…	…	…
y_N	x_{N1}	x_{N2}	…	x_{Nn}

Maximum likelihood estimation, the likelihood function is

$$L(\alpha_0, \cdots, \alpha_n) = \prod_{i=1}^{n} Pr(Y_i = y_i)$$

$$= \prod_{y_i=1} \left(\frac{e^{\alpha_0 + \sum_{j=1}^{n} \alpha_j x_{ij}}}{1 + e^{\alpha_0 + \sum_{j=1}^{n} \alpha_j x_{ij}}} \right) \prod_{y_i=0} \left(\frac{1}{1 + e^{\alpha_0 + \sum_{j=1}^{n} \alpha_j x_{ij}}} \right)$$

Gibbs Sampler

Sampling the assignment with probability

$$
\begin{aligned}
& Pr(X_i = 1 | X_{[-i]}, \theta) \\
& = \frac{Pr(X_i = 1 | X_{[-i]}, \theta)}{Pr(X_i = 1 | X_{[-i]}, \theta) + Pr(X_i = 0 | X_{[-i]}, \theta)} \\
& = \frac{\exp(\alpha + (\beta - 1) M_0^{(i)} + (\gamma - \beta) M_1^{(i)})}{1 + \exp(\alpha + (\beta - 1) M_0^{(i)} + (\gamma - \beta) M_1^{(i)})}
\end{aligned}
$$

Algorithm

- *Initialization*. Assign function (1 or 0) for unannotated proteins with prior probability
- *Update* the function assignment for each unknown protein until convergence
  - Compute the posterior probability for each protein based on other proteins (interaction partners).
  - Assign function (1 or 0) for it based on this posterior probability
- *Output*

Iterative Conditional Mode

- Greedy assign function according to the conditional probability

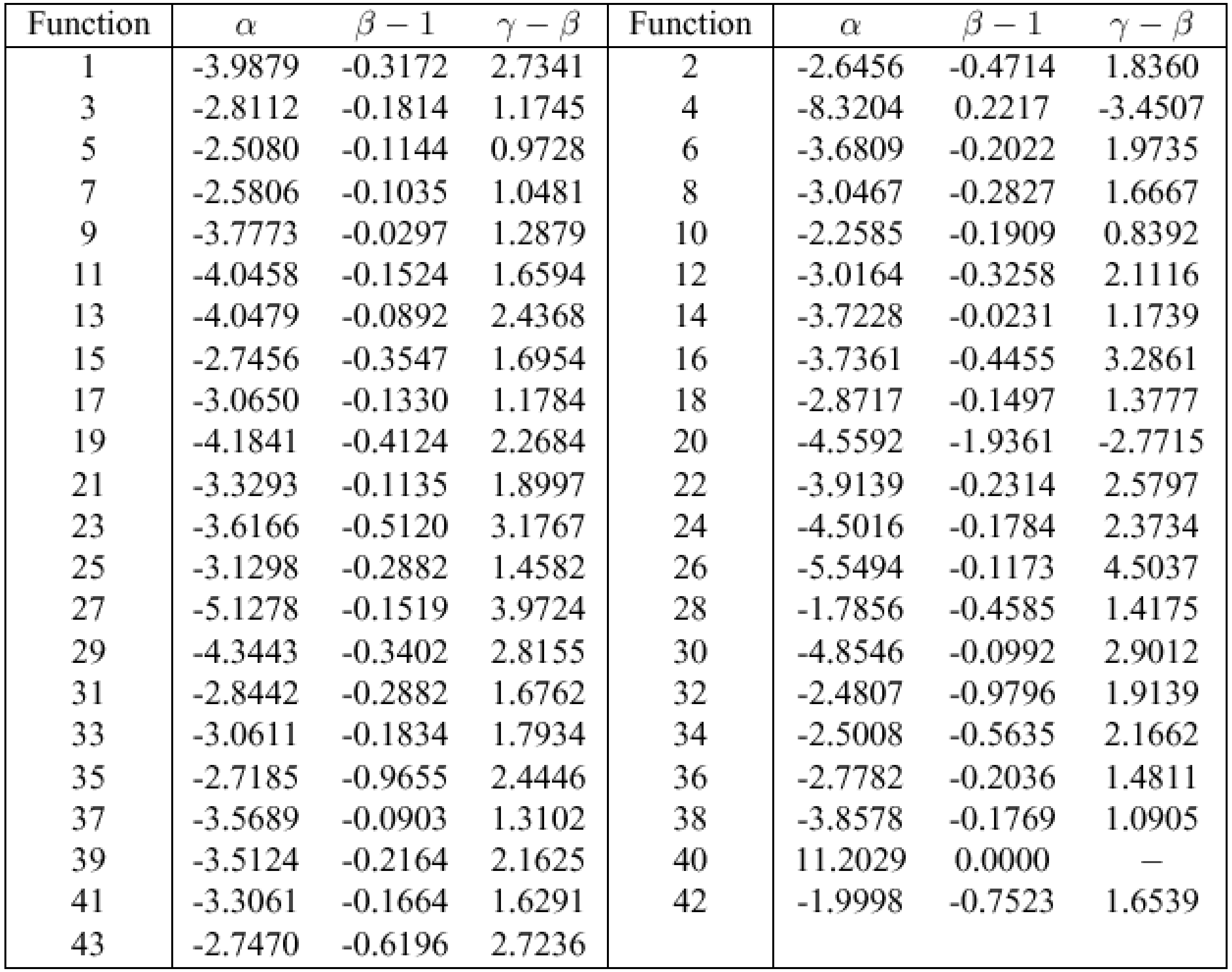
$$
\begin{aligned}
&Pr(X_i = 1|X_{[-i]}, \theta) \\
&= \frac{Pr(X_i = 1|X_{[-i]}, \theta)}{Pr(X_i = 1|X_{[-i]}, \theta) + Pr(X_i = 0|X_{[-i]}, \theta)} \\
&= \frac{\exp(\alpha + (\beta - 1)M_0^{(i)} + (\gamma - \beta)M_1^{(i)})}{1 + \exp(\alpha + (\beta - 1)M_0^{(i)} + (\gamma - \beta)M_1^{(i)})}
\end{aligned}
$$

Data

- YPD (Feb.15, 2002), 6281 proteins, 43 classses of “cellular role.”
- MIPS physical interactions: 2559 interaction pairs, 120 self-interaction

		Number of interaction partners						
	All	0	1	2	3	4	5	>=6
Annotated	3854	2399	1455	785	514	346	252	186
Unannotated	2427	2005	422	131	45	15	12	7
Total	6281	4404	1877	916	559	361	264	193

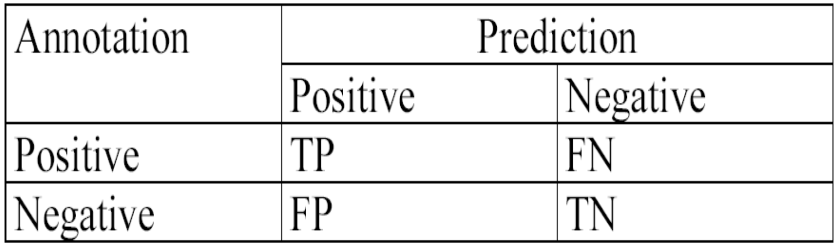
Estimated Parameters



Function	α	$\beta - 1$	$\gamma - \beta$	Function	α	$\beta - 1$	$\gamma - \beta$
1	-3.9879	-0.3172	2.7341	2	-2.6456	-0.4714	1.8360
3	-2.8112	-0.1814	1.1745	4	-8.3204	0.2217	-3.4507
5	-2.5080	-0.1144	0.9728	6	-3.6809	-0.2022	1.9735
7	-2.5806	-0.1035	1.0481	8	-3.0467	-0.2827	1.6667
9	-3.7773	-0.0297	1.2879	10	-2.2585	-0.1909	0.8392
11	-4.0458	-0.1524	1.6594	12	-3.0164	-0.3258	2.1116
13	-4.0479	-0.0892	2.4368	14	-3.7228	-0.0231	1.1739
15	-2.7456	-0.3547	1.6954	16	-3.7361	-0.4455	3.2861
17	-3.0650	-0.1330	1.1784	18	-2.8717	-0.1497	1.3777
19	-4.1841	-0.4124	2.2684	20	-4.5592	-1.9361	-2.7715
21	-3.3293	-0.1135	1.8997	22	-3.9139	-0.2314	2.5797
23	-3.6166	-0.5120	3.1767	24	-4.5016	-0.1784	2.3734
25	-3.1298	-0.2882	1.4582	26	-5.5494	-0.1173	4.5037
27	-5.1278	-0.1519	3.9724	28	-1.7856	-0.4585	1.4175
29	-4.3443	-0.3402	2.8155	30	-4.8546	-0.0992	2.9012
31	-2.8442	-0.2882	1.6762	32	-2.4807	-0.9796	1.9139
33	-3.0611	-0.1834	1.7934	34	-2.5008	-0.5635	2.1662
35	-2.7185	-0.9655	2.4446	36	-2.7782	-0.2036	1.4811
37	-3.5689	-0.0903	1.3102	38	-3.8578	-0.1769	1.0905
39	-3.5124	-0.2164	2.1625	40	11.2029	0.0000	–
41	-3.3061	-0.1664	1.6291	42	-1.9998	-0.7523	1.6539
43	-2.7470	-0.6196	2.7236				

Cross Validation

- Leave-one-out test.
- *Precision*
  - fraction of the correct predictions within all predictions.
- *Recall*
  - Fraction of the correct predictions within all the test samples

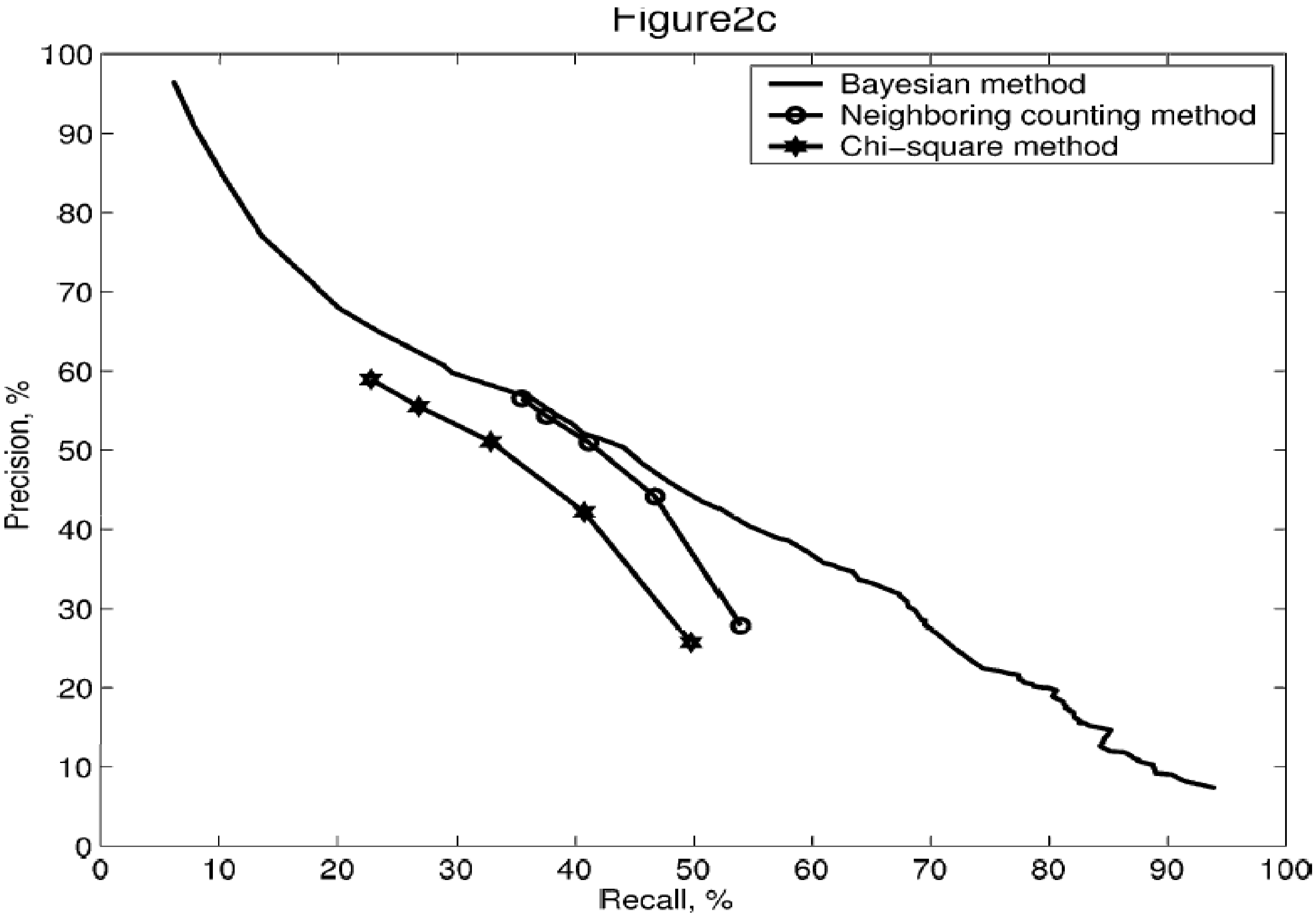


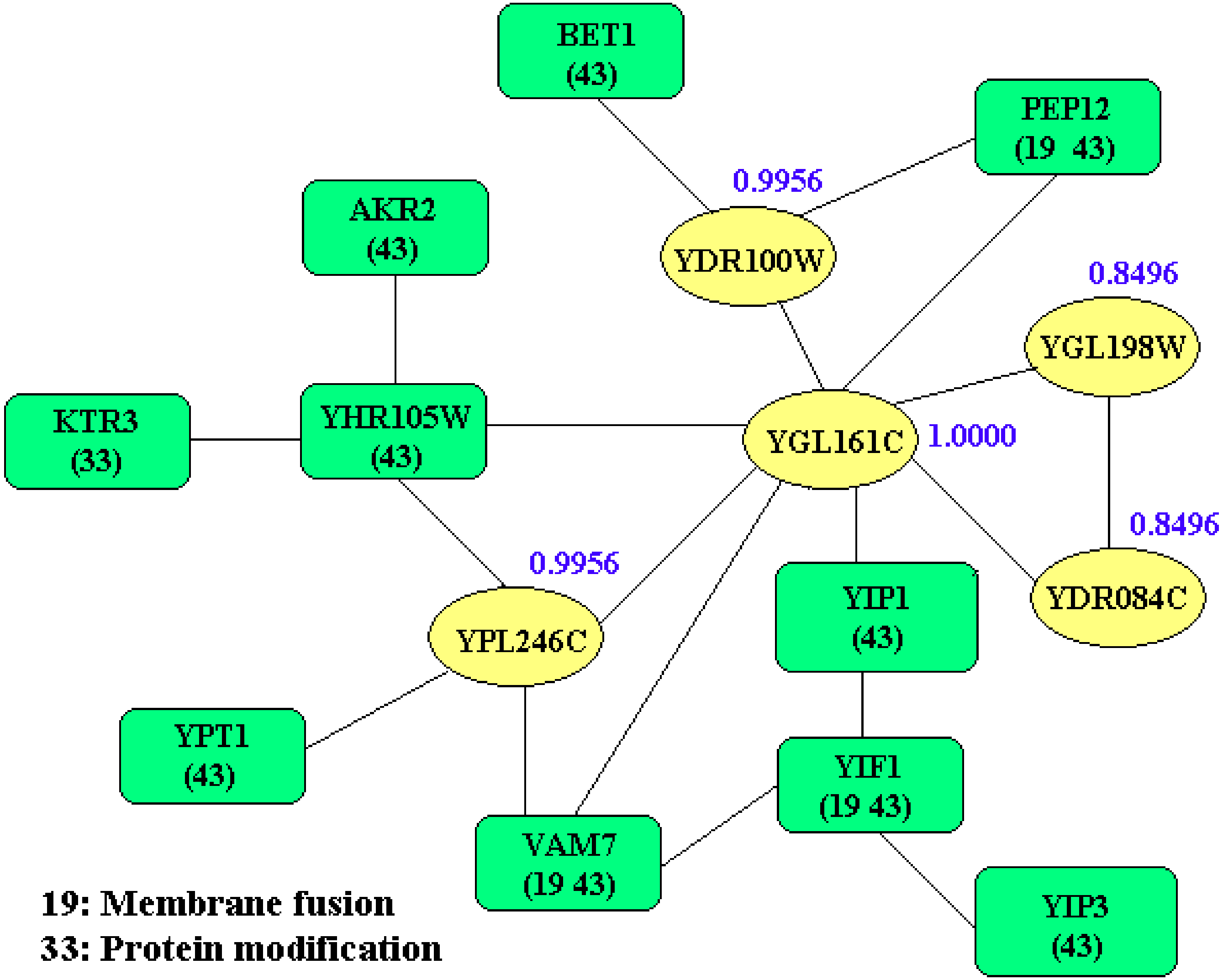
Annotation	Prediction	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Comparison of Different Methods



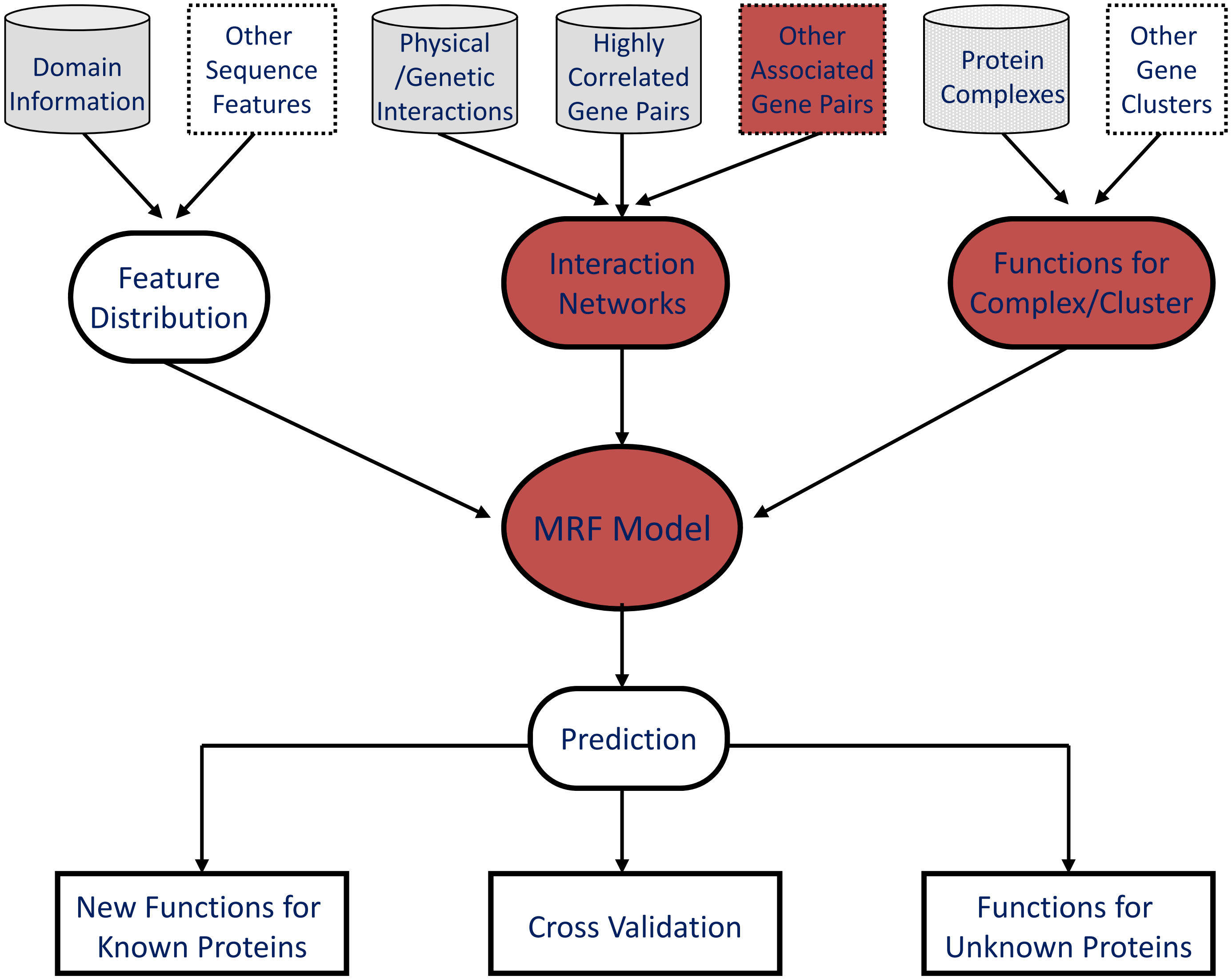


19: Membrane fusion
33: Protein modification
43: Vesicular transport

Extending MRF to Incorporate Other Data Sources

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

A General MRF Model

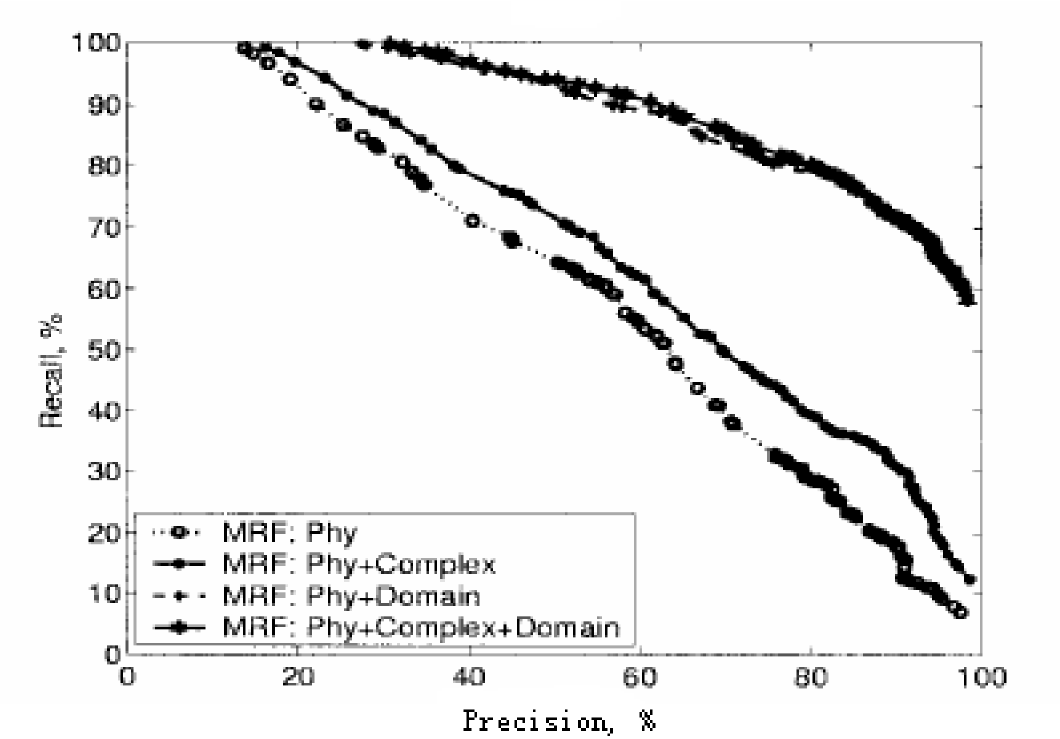

Domain Information
Other Sequence Features
Physical /Genetic Interactions
Highly Correlated Gene Pairs
Other Associated Gene Pairs
Protein Complexes
Other Gene Clusters
Feature Distribution
Interaction Networks
Functions for Complex/Cluster
MRF Model
Prediction
New Functions for Known Proteins
Cross Validation
Functions for Unknown Proteins

Key Points in MRF

- The *prior probability* of an unknown protein, X_i, is obtained from protein complex information.
- Given the functional assignment of a protein, the probability of sequence features and domain information
- Given the functional assignment of all the proteins, the believe of all the networks

$$
\begin{aligned}
&\Pr(X_i \mid \mathbf{X}^{\text{known}}) \\
&= \sum_{\mathbf{X}^{\text{other\_unknown}}} \Pr(X_i \mid \mathbf{X}^{\text{other\_unknown}}, \mathbf{X}^{\text{known}}) \Pr(\mathbf{X}^{\text{other\_unknown}} \mid \mathbf{X}^{\text{known}})
\end{aligned}
$$

Leave-one-out Test for Known Proteins



Deng et al. 2004, *J. Comput. Biol.,11:463-475*

Related Works

- Letovsky and Kasif, ISMB2003, also Bioinformatics 19-Sl: i197-i204 (MRF)
- Vazquez et al. Nature Biotech., 2003 (Ising model, simulated annealing)
- Karaoz et al. PNAS, March 2, 2004. (MRF)
- Lanckriet et al. PSB 2004 (SVM)
- Leone and Pagnai. Bioinformatics 21: 239-247, 2005. (MRF, loopy message passing)

Other Works

- Samanta and Liang, PNAS 100:12579-12583, 2003.

  Criteria: if two proteins share significantly larger number of common interaction partners than random, they have close functional associations.

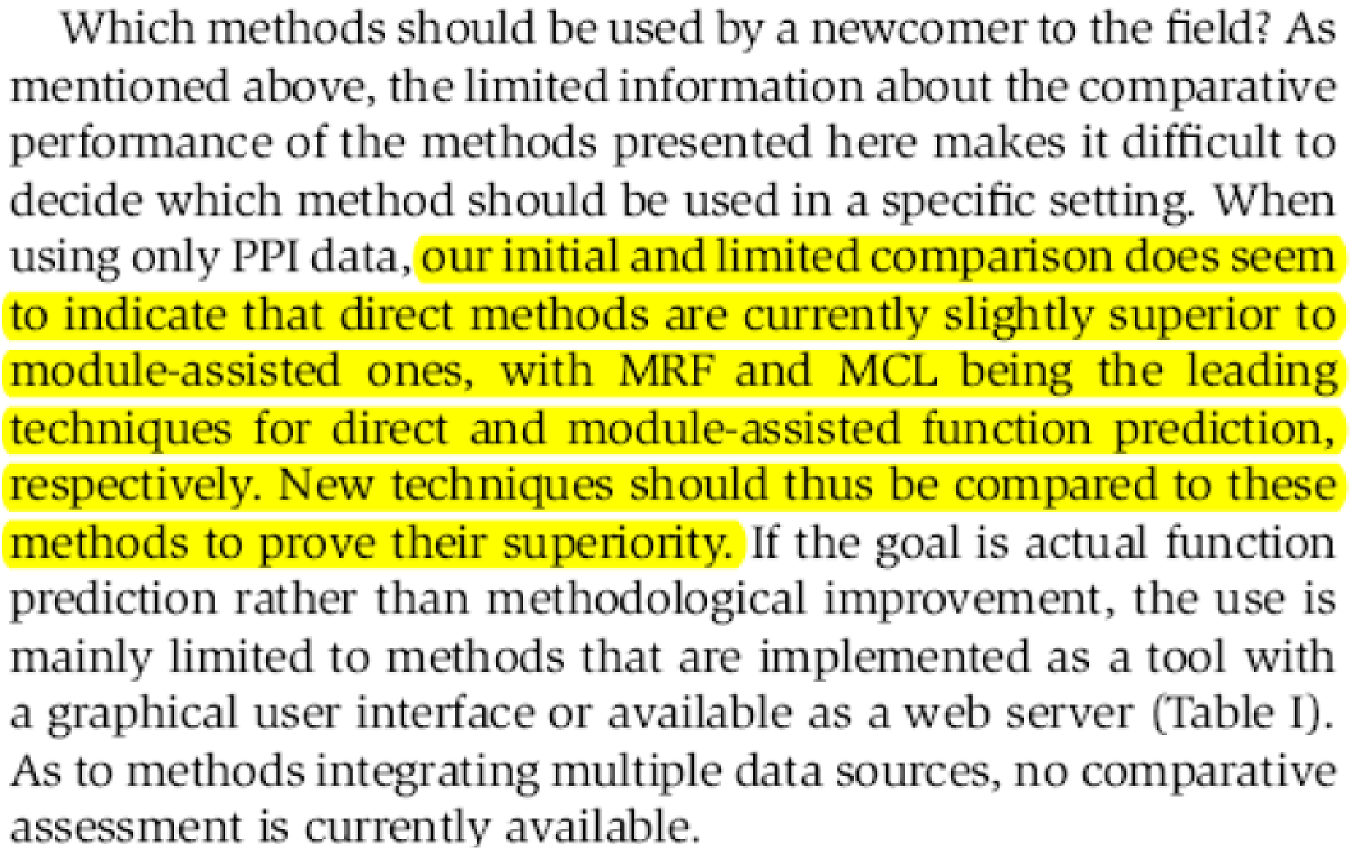
- Zhou et al. PNAS 99: 12783-12788, 2002.

  Transitive Functional Annotation by shortest path analysis of co-expressed gene network.

- Module based methods

A Good Review Paper

- Sharan et al. Mol. System Biol. 2007



Which methods should be used by a newcomer to the field? As mentioned above, the limited information about the comparative performance of the methods presented here makes it difficult to decide which method should be used in a specific setting. When using only PPI data, our initial and limited comparison does seem to indicate that direct methods are currently slightly superior to module-assisted ones, with MRF and MCL being the leading techniques for direct and module-assisted function prediction, respectively. New techniques should thus be compared to these methods to prove their superiority. If the goal is actual function prediction rather than methodological improvement, the use is mainly limited to methods that are implemented as a tool with a graphical user interface or available as a web server (Table I). As to methods integrating multiple data sources, no comparative assessment is currently available.

References

- Schwikoski et al. 2000, *Nature Biotech.18:1257-1261*
- Hishigaki et al. 2001, *Yeast, 18:523-531*
- Zhou et al. 2002, *PNAS, 99: 12783-12788*
- Deng et al. 2003, *J. Comput. Biol.10: 947-960*
- Letovskyet al.2003, *Bioinformatics, 19-Sl: i197-i204*
- Samanta et al. 2003, *PNAS, 100:12579-12583*
- Vazquez et al. 2003, *Nature Biotech., 697-700*
- Deng et al. 2004, *Bioinformatics,20:895-902*
- Deng et al. 2004, *J. Comput. Biol.,11:463-475*
- Lanckriet et al. 2004, *PSB, 300-311*
- Karaoz et al. 2004, *PNAS, 101:2888-2893*
- Leone et al. 2005, *Bioinformatics, 21: 239-247*
- Sharanet al. 2007, *Mol. System Biol.3: 88*

Kernel Logistic Regression Model

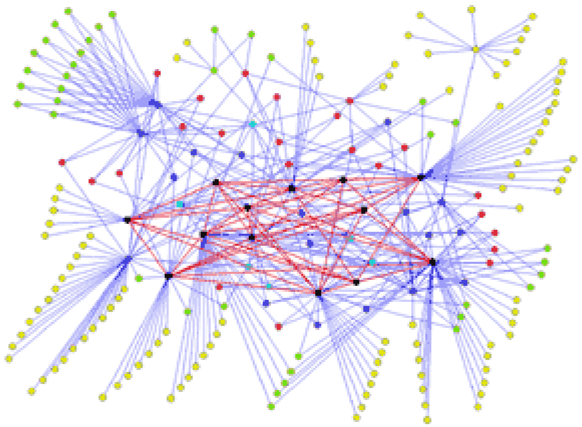
- Kernel logistic regression method
- Yeast function prediction
- Mouse function prediction

- References:

1. Lee et al. *Omics: integrative biology*, 2006.
2. Pena-Castillo *et al., Genome biology,* 9:S2, 2008.

Kernels for Protein Interaction

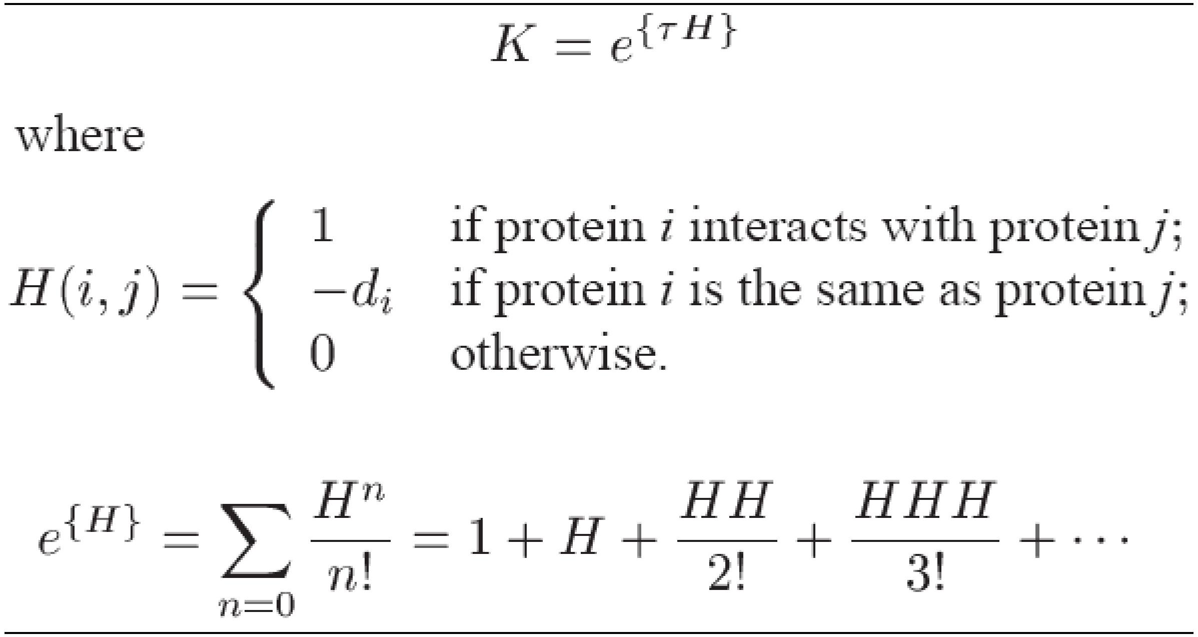
- Motivation: Two interacting proteins are more likely to share similar functions
- Adjacent kernel



$$K(i,j) = \begin{cases} 1 & \text{if protein } i \text{ interacts with protein } j; \\ 0 & \text{otherwise.} \end{cases}$$

Diffusion Kernel for Protein Interaction

- Diffusion kernel K calculates the similarity distance between any two nodes in the network



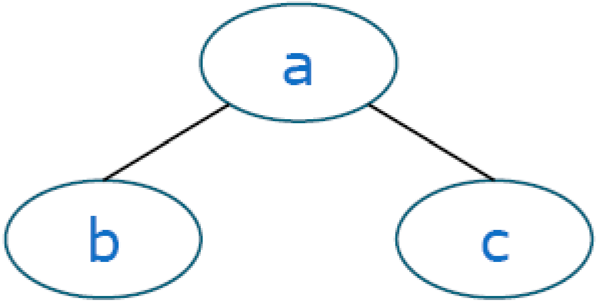
$$K = e^{\{\tau H\}}$$

where

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

$$e^{\{H\}} = \sum_{n=0} \frac{H^n}{n!} = 1 + H + \frac{HH}{2!} + \frac{HHH}{3!} + \cdots$$

Diffusion Kernel for Protein Interaction



H =

$$\begin{array}{cccc} & a & b & c \\ a & 0 & 1 & 1 \\ b & 1 & 0 & 0 \\ c & 1 & 0 & 0 \end{array}$$

K=e(H, tau=0.5) =

$$\begin{array}{cccc} & a & b & c \\ a & 0.48 & 0.26 & 0.26 \\ b & 0.26 & 0.67 & 0.07 \\ c & 0.26 & 0.07 & 0.67 \end{array}$$

Kernels for Other Data

- Gene expression

$$K_n(i,j) = (1 + \mathrm{PCC}(i,j))^n$$

- Domain: suppose v_i is the domain membership of protein i

$$K_n(i,j) = (1 + v_i v_j)^n$$
$$K_n(i,j) = (1 + \frac{v_i v_j}{|v_i| + |v_j| - |v_i v_j|})^n$$

Kernel Logistic Regression Model (I)

- Assume: Probability for X

$$\begin{cases} Pr(X|\theta) = \dfrac{1}{Z(\theta)} \exp(-U(x)) \\ U(x) = -\alpha N_1 - \beta_{10} D_{10} - \beta_{11} D_{11} - \beta_{00} D_{00} \end{cases}$$

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

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

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

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

Kernel Logistic Regression Model (II)

$$\log \frac{Pr(X_i = 1|X_{[-i]}, \theta)}{1 - Pr(X_i = 1|X_{[-i]}, \theta)}$$

$$= \alpha + (\beta_{10} - \beta_{00})K_0(i) + (\beta_{11} - \beta_{10})K_1(i)$$

$$K_0(i) = \sum_{i \neq j} K(i, j)I(X_j = 0)$$

$$K_1(i) = \sum_{i \neq j} K(i, j)I(X_j = 1)$$

Kernel Logistic Regression Model (III)

- Multiple data integration

$$\log \frac{Pr(X_i = 1|X_{[-i]}, \theta)}{1 - Pr(X_i = 1|X_{[-i]}, \theta)}$$

$$= \gamma + \sum_{d=1}^{D} \delta_0^{(d)} M_0^{(d)}(i) + \sum_{d=1}^{D} \delta_1^{(d)} M_1^{(d)}(i)$$

$$M_0^{(d)}(i) = \sum_{i \neq j} M^{(d)}(i,j) I(X_j = 0)$$

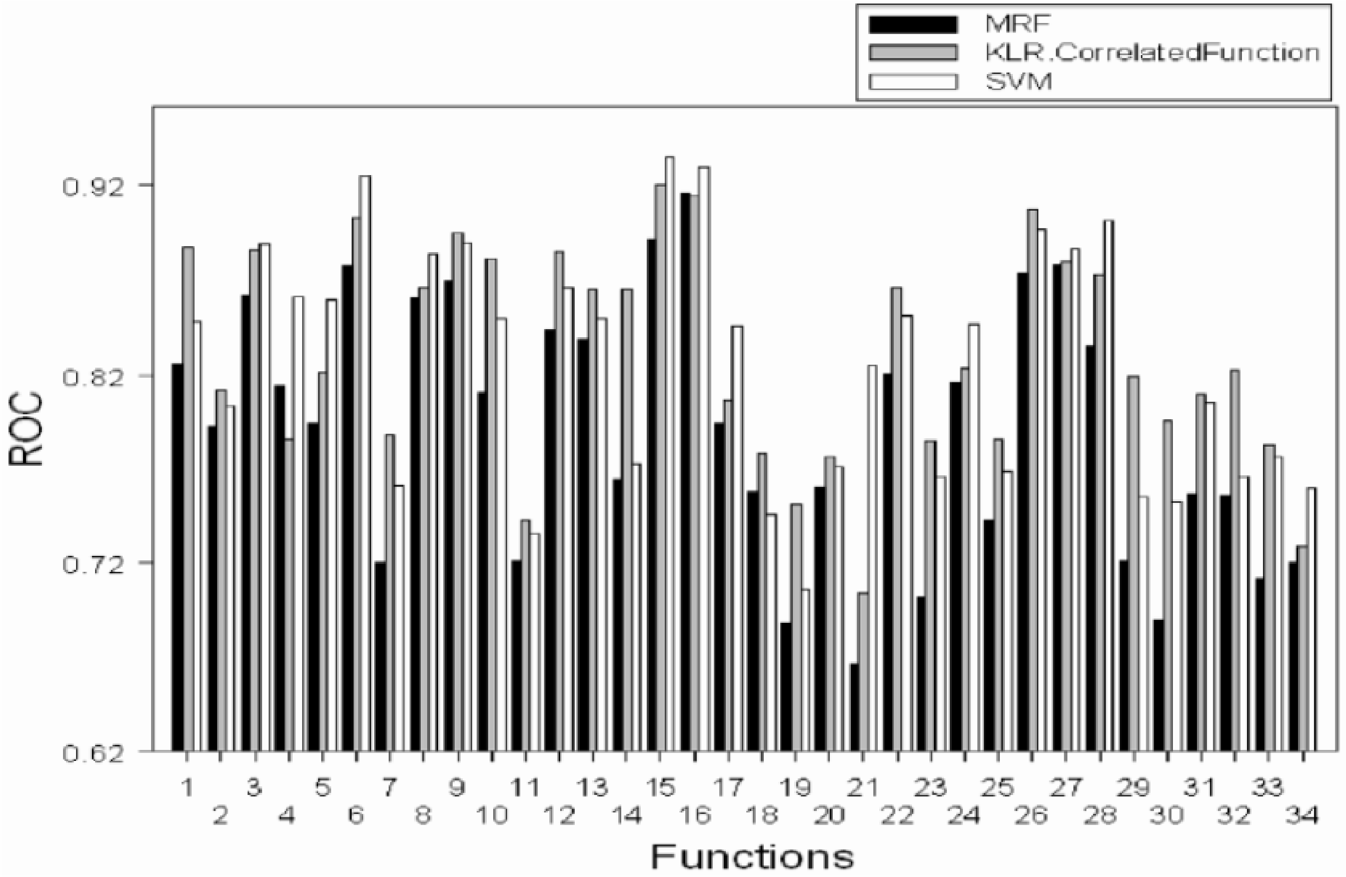
$$M_1^{(d)}(i) = \sum_{i \neq j} M^{(d)}(i,j) I(X_j = 1)$$

- D: number of data sources
- $M^{(d)}(i,j)$ the kernel matrix of d-th data source

Yeast Data Sets

- Yeast protein functions
  - Protein : gene list from the SGD database
  - Function : 34 functions from Gene Ontology (GO) biological process
- Physical protein interactions
  - 2,566 interactions from MIPS (http://mips.gsf.de)
- Protein-domaininteractions
  - Pfam(http://www.sanger.ac.uk/Software/Pfam/iPfam/)
- Protein localization
  - Huh *et al.(2003)*
- Protein complex
  - Ho *et al. (2002). high-throughput mass spectrometric protein complex identification (HMS-PCI)*

Comparison of Different Methods



Lee et al. Omics: integrative biology 2006

Validation of Predictions

- Ten newly added annotated proteins (probability > 0.2) between May 2003 and December 2005, and their function annotations

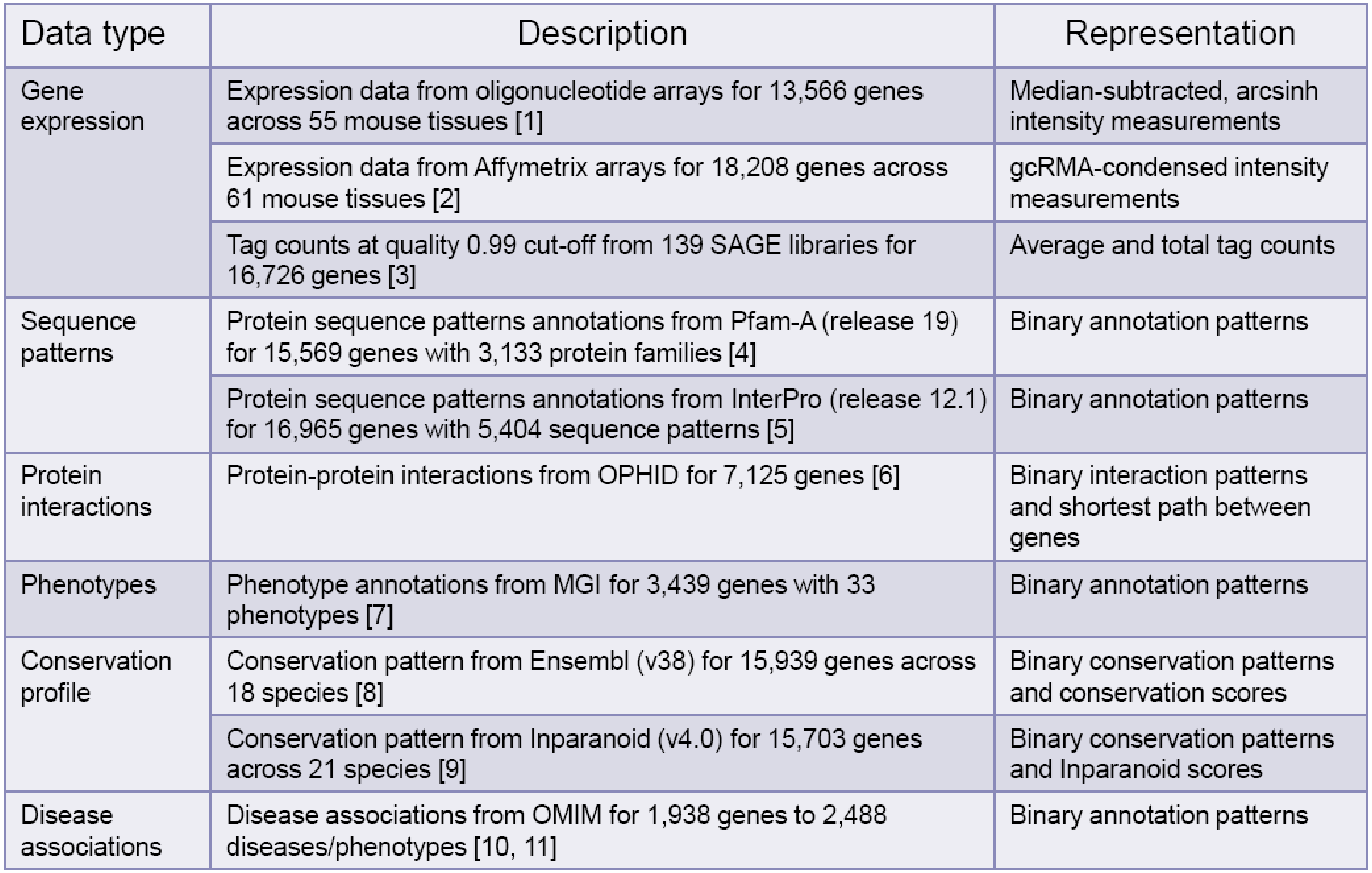
Protein	*New annotation*	*MRIN**	*Probability*
FRQ1	Regulation of signal transduction	1	0.404
YLR254C	Nuclear migration, microtubule-mediated	3	0.906
YMR009W	Methionine salvage	20	1.000
YPR118W	Methionine salvage	20	1.000
PIG2	Regulation of glycogen biosynthesis	23	0.268
HRT3	Ubiquitin-dependent protein catabolism	24	0.477
PSY3	Error-free DNA repair	26	0.334
YJU2	Nuclear mRNA splicing, via spliceosome	27	0.916
YLR424W	Nuclear mRNA splicing, via spliceosome	27	0.299
YDR140W	Peptidyl-glutamine methylation	31	0.320

Lee et al. Omics: integrative biology 2006

Mouse Function Prediction

- Motivation: Evaluation of computational function prediction for mammals
- Nine bioinformatics teams participated.
- A standardized collection of mouse functional genomic data;
- GO functional categories;
- 21603 mouse genes (76%);
- Randomized training data.
- Randomized held-out test data.
- Novel test data.

Data Sets (I)

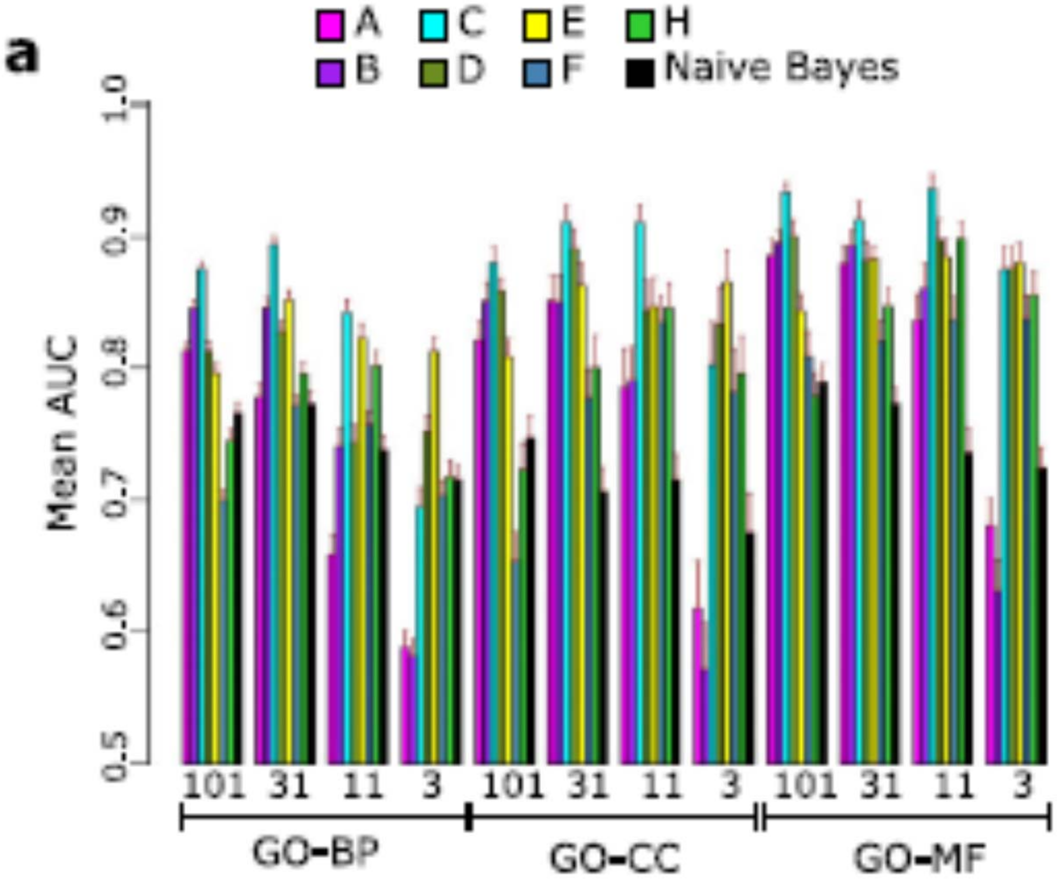


Data type	Description	Representation
Gene expression	Expression data from oligonucleotide arrays for 13,566 genes across 55 mouse tissues [1]	Median-subtracted, arcsinh intensity measurements
	Expression data from Affymetrix arrays for 18,208 genes across 61 mouse tissues [2]	gcRMA-condensed intensity measurements
	Tag counts at quality 0.99 cut-off from 139 SAGE libraries for 16,726 genes [3]	Average and total tag counts
Sequence patterns	Protein sequence patterns annotations from Pfam-A (release 19) for 15,569 genes with 3,133 protein families [4]	Binary annotation patterns
	Protein sequence patterns annotations from InterPro (release 12.1) for 16,965 genes with 5,404 sequence patterns [5]	Binary annotation patterns
Protein interactions	Protein-protein interactions from OPHID for 7,125 genes [6]	Binary interaction patterns and shortest path between genes
Phenotypes	Phenotype annotations from MGI for 3,439 genes with 33 phenotypes [7]	Binary annotation patterns
Conservation profile	Conservation pattern from Ensembl (v38) for 15,939 genes across 18 species [8]	Binary conservation patterns and conservation scores
	Conservation pattern from Inparanoid (v4.0) for 15,703 genes across 21 species [9]	Binary conservation patterns and Inparanoid scores
Disease associations	Disease associations from OMIM for 1,938 genes to 2,488 diseases/phenotypes [10, 11]	Binary annotation patterns

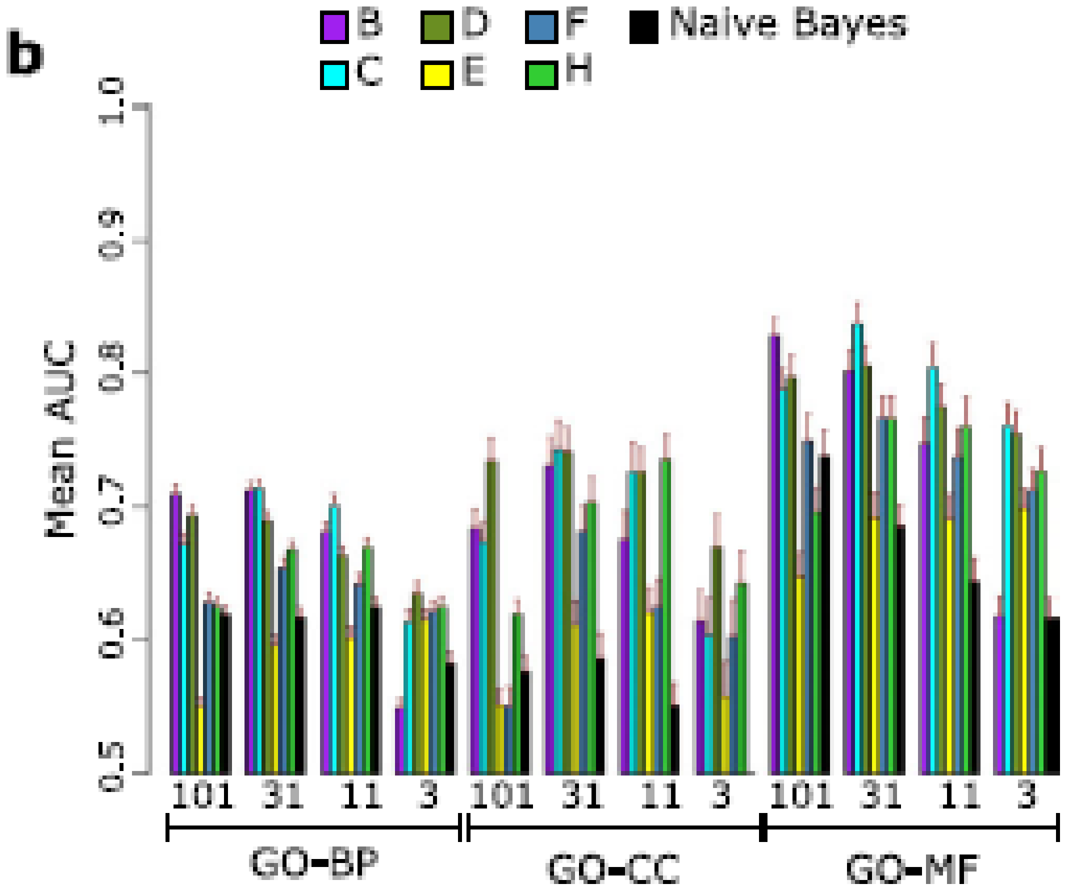
Data Sets (II)

- Zhang W et al. J Biol2004, 3(5):21.
- Su AI et al. PNAS USA 2004, 101(16):6062–7.
- SiddiquiAS et al. PNAS USA 2005, 102(51):18485–90.
- Finn RD et al. Nucleic Acids Res 2006, 34:D247–51.
- MulderNJ et al. Nucleic Acids Res 2005, 33:D201–5.
- Brown KR, JurisicaI: Bioinformatics 2005, 21(9):2076–82.
- EppigJT, Nucleic Acids Res 2007, 35:D630–7.
- KasprzykA et al. Genome Res 2004, 14:160–9.
- O’Brien KP et al. Nucleic Acids Res 2005, 33:D476–80.
- Wheeler DL et al. Nucleic Acids Res 2007, 35:D5–12.
- Hamosh A et al. Nucleic Acids Res 2005, 33:D514–7.

On Held-out Test Data



On Novel Test Data



Network-based Genomics Analysis

References

1. Zhi Wei and Hongzhe Li. A Markov random field model for network-based analysis of genomic data. Bioinformatics 23(12): 1537–1544,2007.
2. Hongzhe Li, Zhi Wei and John Maris. A hidden Markov random field model for genome-wide association studies. Biostatistics 11(1):139–150, 2010.
3. Zhi Wei and Hongzhe Li. A hidden spatial-temporal markov random field model for network-based analysis of time course gene expression data. Annals of Applied Statistics 2(1): 408-429, 2008.
4. Min Chen, Judy Cho and Hongyu Zhao. Incorporating Biological Pathways via a Markov Random Field Model in Genome-Wide Association Studies. PloS Genetics 7(4): e1001353, 2011.

Integrating Network Structure to Gene Expression Analysis

- The gene network gives a prior: interacted genes are more likely to have same differential expression pattern.

- A Markov random field model is used to model this network constraint.

Hidden Markov Random Field (HMRF)

- X: status of genes, differential expression (DE) or equal expression (EE), which is hidden
- Y: observed gene expression level of each gene.
- Data: Protein-protein interaction network, gene expression profiles with m replicates under one condition, and n replicates under another condition

Gamma-gamma Model for Gene Expression Data

- For ith gene, Gamma distribution with latent mean

$$f(y|\mu_i) = \frac{\lambda_i^\alpha y^{\alpha-1} \exp(-\lambda_i y)}{\Gamma(\alpha)}, \lambda_i = \frac{\alpha}{\mu_i}$$

- The latent mean has a inverse Gamma distribution $\pi(\mu)$ over all genes with shape parameter α_0 and scale parameter v.

Gamma-gamma Model for Gene Expression Data

- So the joint density is

$$f(y_i) = \int f(y_i|\mu_i)\pi(\mu_i)d\mu_i$$

- For genes under the first condition

$$f(y_{i1}, \cdots, y_{im}) = K_1 \frac{(\prod_{j=1}^m y_{ij})^{\alpha-1}}{(v + \sum_{j=1}^m y_{ij})^{m\alpha+\alpha_0}}$$

$$K_1 = \frac{v^{\alpha_0}\Gamma(m\alpha+\alpha_0)}{\Gamma^m(\alpha)\Gamma(\alpha_0)}$$

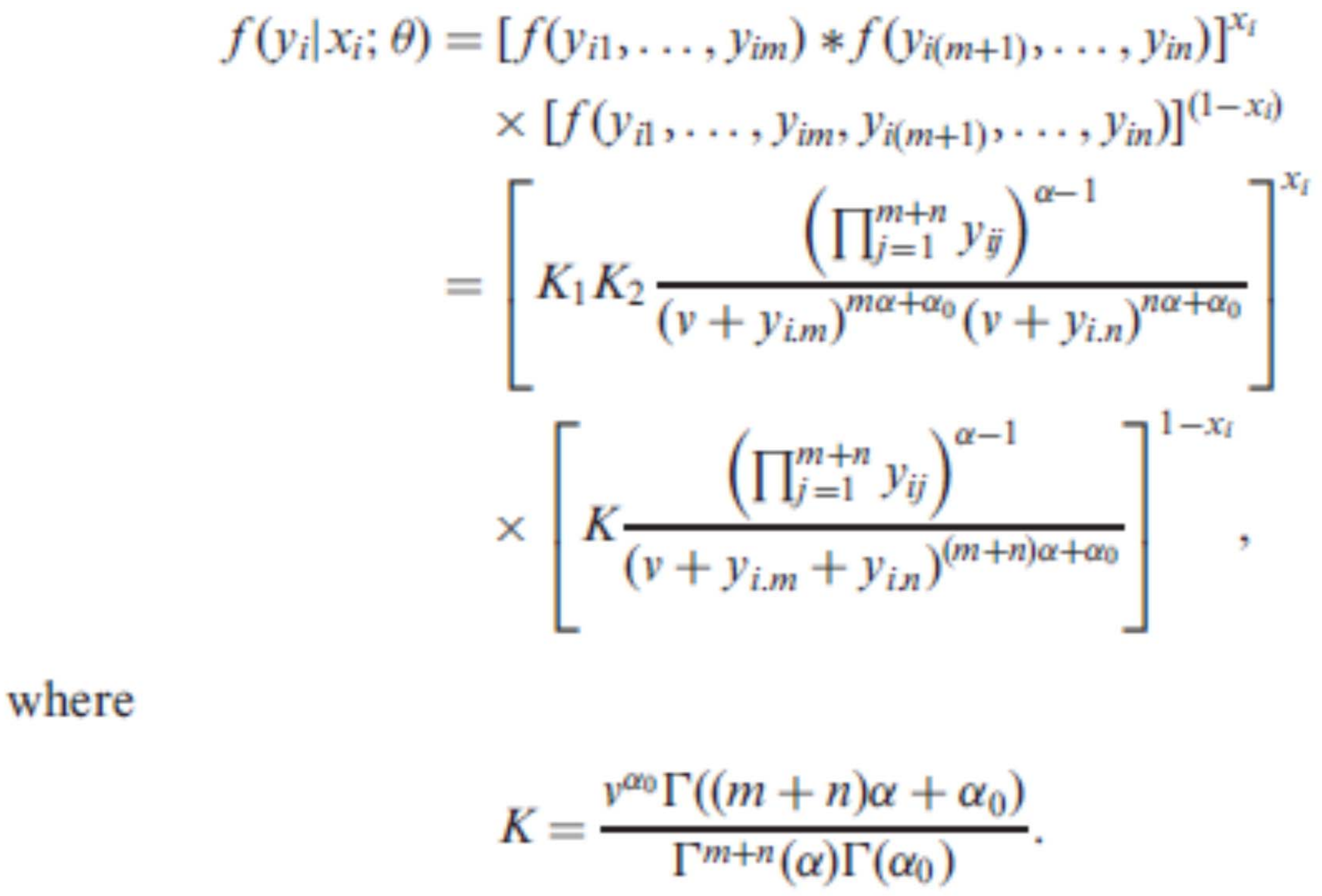
Gamma-gamma Model for Gene Expression Data

- For genes under the second condition

$$f(y_{i(m+1)}, \cdots, y_{i(m+n)}) = K_2 \frac{(\prod_{j=m+1}^{m+n} y_{ij})^{\alpha-1}}{(v + \sum_{j=m+1}^{m+n} y_{ij})^{n\alpha+\alpha_0}}$$

$$K_2 = \frac{v^{\alpha_0}\Gamma(n\alpha + \alpha_0)}{\Gamma^n(\alpha)\Gamma(\alpha_0)}$$

Emission Probability



$$\begin{aligned} f(y_i|x_i;\theta) &= [f(y_{i1},\ldots,y_{im}) * f(y_{i(m+1)},\ldots,y_{in})]^{x_i} \\ &\quad \times [f(y_{i1},\ldots,y_{im},y_{i(m+1)},\ldots,y_{in})]^{(1-x_i)} \\ &= \left[K_1 K_2 \frac{\left(\prod_{j=1}^{m+n} y_{ij}\right)^{\alpha-1}}{(v+y_{i.m})^{m\alpha+\alpha_0}(v+y_{i.n})^{n\alpha+\alpha_0}} \right]^{x_i} \\ &\quad \times \left[K \frac{\left(\prod_{j=1}^{m+n} y_{ij}\right)^{\alpha-1}}{(v+y_{i.m}+y_{i.n})^{(m+n)\alpha+\alpha_0}} \right]^{1-x_i}, \end{aligned}$$

where

$$K = \frac{v^{\alpha_0}\Gamma((m+n)\alpha+\alpha_0)}{\Gamma^{m+n}(\alpha)\Gamma(\alpha_0)}.$$

Emission Probability

- Assumption: Given any particular realization x, the random variables Y=(Y_1,Y_2, . . . ,Y_p) are conditionally independent.
- Assumption: Each Y_i has the same unknown conditional density function f(y_i | x_i), dependent only on X_i.

$$l(y|x) = \prod_{i=1}^{p} f(y_i|x_i, \theta) \qquad (1)$$

Markov Random Field (MRF)

- Using a MRF to model the dependency of differential expression status

$$P(x|\Phi) \propto \exp(\gamma_0 n_0 + \gamma_1 n_1 - \beta n_{01})$$

$$n_0 = \sum_{i=1}^{p} (1 - x_i), \quad n_1 = \sum_{i=1}^{p} x_i$$

$$n_{01} = \sum_{(i,j) \in E} (1 - x_i) x_j$$

Conditional Probability

- By considering any two realizations which differ only at gene i,

$$P(X_i = k | X_{[-i]}, \Phi) = P(X_i = k | X_{\partial i}, \Phi) \propto \exp(\gamma_k - \beta u_{i,1-k})$$

where $u_{i,1-k}$ is the number of neighbors of gene i having state 1-k (k=1 or 0)

Psudo-likelihood Estimation of the Parameters

- Psudo-likelihood

$$\begin{aligned} l(x;\Phi) &= \prod_{i=1}^{p} p_i(x_i | x_{\partial i}; \Phi) \qquad (2) \\ &= \prod_{i=1}^{p} \frac{\exp[(1-x_i)(\gamma_0 - \beta u_{i1}) + x_i(\gamma_1 - \beta u_{i0})]}{\exp[\gamma_0 - \beta u_{i1}] + \exp[\gamma_1 - \beta u_{i0}]} \end{aligned}$$

Algorithm (I)

1. Obtain an initial estimation $\hat{x}$ of the true state x*, using a simple two sample t-test.

2. Estimate θ by the value $\hat{\theta}$ which maximizes the likelihood $l(y|\hat{x}, \theta)$ [see equation (1)].

3. Estimate Φ by the value $\hat{\Phi}$ which maximizes the conditional likelihood $l(x; \Phi)$ [see Equation (2)] based on current $\hat{x}$.

Algorithm (II)

4. Carry out a single cycle of ICM based on the current $\hat{x}, \hat{\theta}, \hat{\Phi}$, to obtain a new $\hat{x}$. Specifically, for i=1 to p, update x_i which maximizes

$$P(x_i|y, \hat{x}_{[-i]}) \propto f(y_i|x_i;\theta)p_i(x_i|\hat{x}_{\partial i};\hat{\Phi})$$

5. Go to step 2 for a fixed number of cycles or until approximate convergence of $\hat{x}$.

Simulation Study

- Network:1668 nodes and 8011 edges; totally 33 pathways
- Genes in the K pathways to be DE and the rest of genes to be EE (k=5,9, 13,17)
- r0= 1, r1=1, b=2.
- GG model for the expression, 3 replicates in each condition with a=10, a0=0.9, v=0.5
- Simulations are repeated 100 times.

% of DE in simulated data	Model	Sensitivity	Specificity	FDR
	MRFGG	0.682 (0.064)	0.999 (0.001)	0.013 (0.011)
$P = 0.115$ (0.005)	GG	0.640 (0.035)	0.998 (0.001)	0.023 (0.015)
	tTEST1	0.495 (0.033)	0.966 (0.005)	0.347 (0.037)
	tTEST2	0.007 (0.009)	1.000 (0.000)	0.014 (0.075)
	MRFGG	0.743 (0.067)	0.997 (0.003)	0.018 (0.014)
$P = 0.189$ (0.008)	GG	0.664 (0.027)	0.996 (0.002)	0.023 (0.012)
	tTEST1	0.495 (0.029)	0.966 (0.005)	0.229 (0.029)
	tTEST2	0.010 (0.010)	1.000 (0.000)	0.009 (0.041)
	MRFGG	0.793 (0.037)	0.991 (0.006)	0.020 (0.011)
$P = 0.357$ (0.009)	GG	0.698 (0.020)	0.990 (0.004)	0.024 (0.008)
	tTEST1	0.497 (0.020)	0.966 (0.005)	0.110 (0.017)
	tTEST2	0.019 (0.012)	1.000 (0.000)	0.008 (0.023)
	MRFGG	0.835 (0.036)	0.975 (0.011)	0.030 (0.012)
$P = 0.486$ (0.008)	GG	0.718 (0.018)	0.982 (0.006)	0.025 (0.008)
	tTEST1	0.496 (0.017)	0.966 (0.006)	0.068 (0.012)
	tTEST2	0.026 (0.014)	1.000 (0.001)	0.011 (0.022)

Summaries are averaged over 100 simulations; standard errors are shown in parentheses. tTest1: two-sample *t*-test using *P*-value of 0.05 as cutoff point; tTEST2: two-sample *t*-test for FDR = 0.05 using the procedure of Benjamini and Hochberg.

Gamma Distribution on Wiki

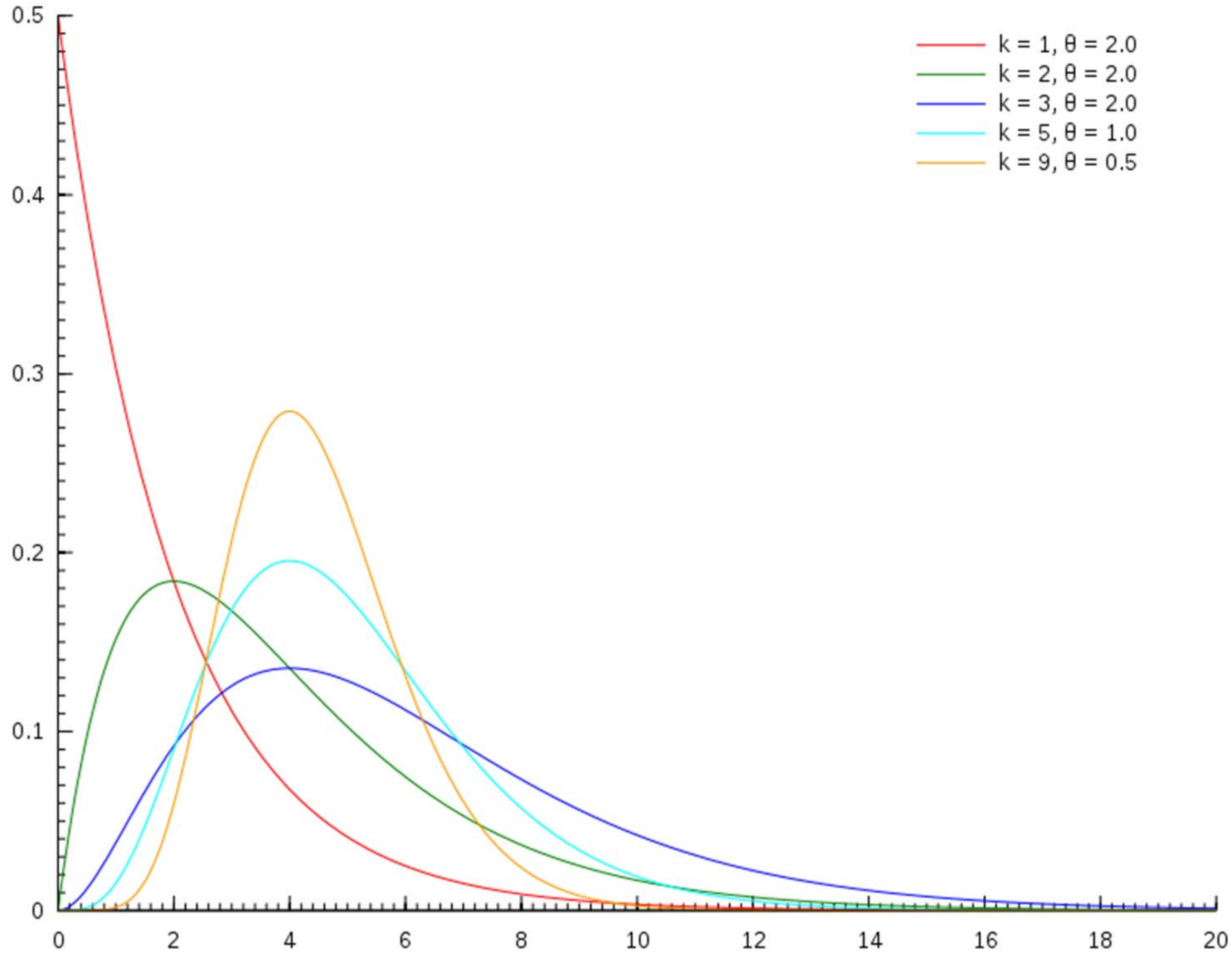
- In life test, the waiting time until death is generally model as Gamma distribution
- Density function (shape α , scale β)

$$\mathrm{Gamma}(x;\alpha,\beta)=\frac{x^{\alpha-1}\beta^{\alpha}e^{-x\beta}}{\Gamma(\alpha)}$$

- Expectation: $E[\ln(X)]=\psi(\alpha)-\ln(\beta)$
- Gamma distribution is the conjugate prior for many distributions

http://en.wikipedia.org/wiki/Gamma_distribution

Gamma Distribution on Wiki



http://en.wikipedia.org/wiki/Gamma_distribution

Inverse-Gamma Distribution on Wiki

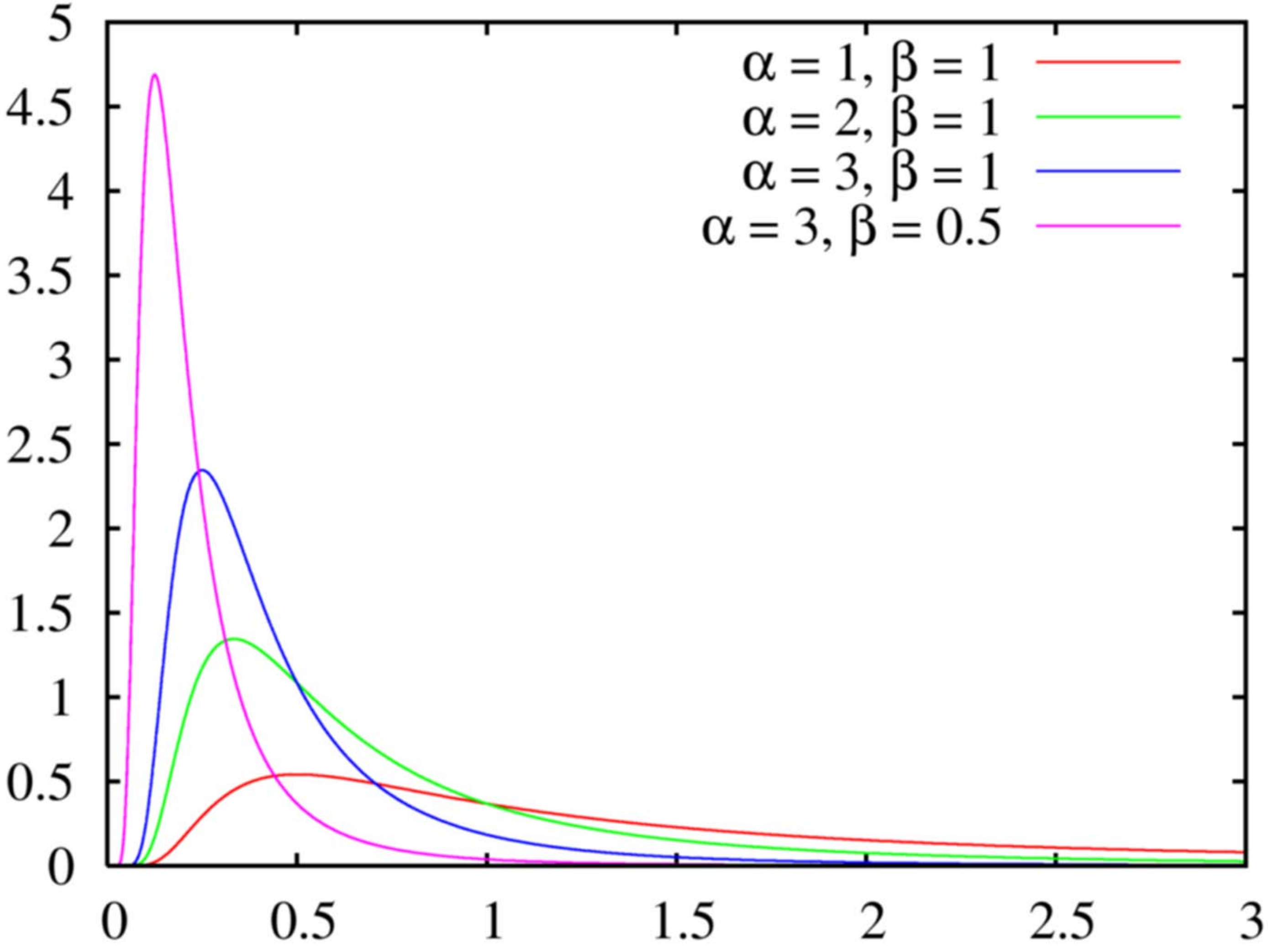
- Density function

$$f(x;\alpha,\beta) = \frac{\beta^\alpha x^{-\alpha-1}\exp(-\frac{\beta}{x})}{\Gamma(\alpha)}$$

- E(1/X)=a/b
- If X~Inv-Gamma(a,b), kX~Inv-Gamma(a,kb)
- If X~Gamma(a,b), 1/X~Inv-Gamma(a,b)
- If X~Gamma(k,t), 1/X~Inv-Gamma(k,1/t)

http://en.wikipedia.org/wiki/Inverse-gamma_distribution

Inverse-Gamma Distribution on Wiki



http://en.wikipedia.org/wiki/Inverse-gamma_distribution