Selective Integration of Multiple Biological Data for Supervised Network Inference

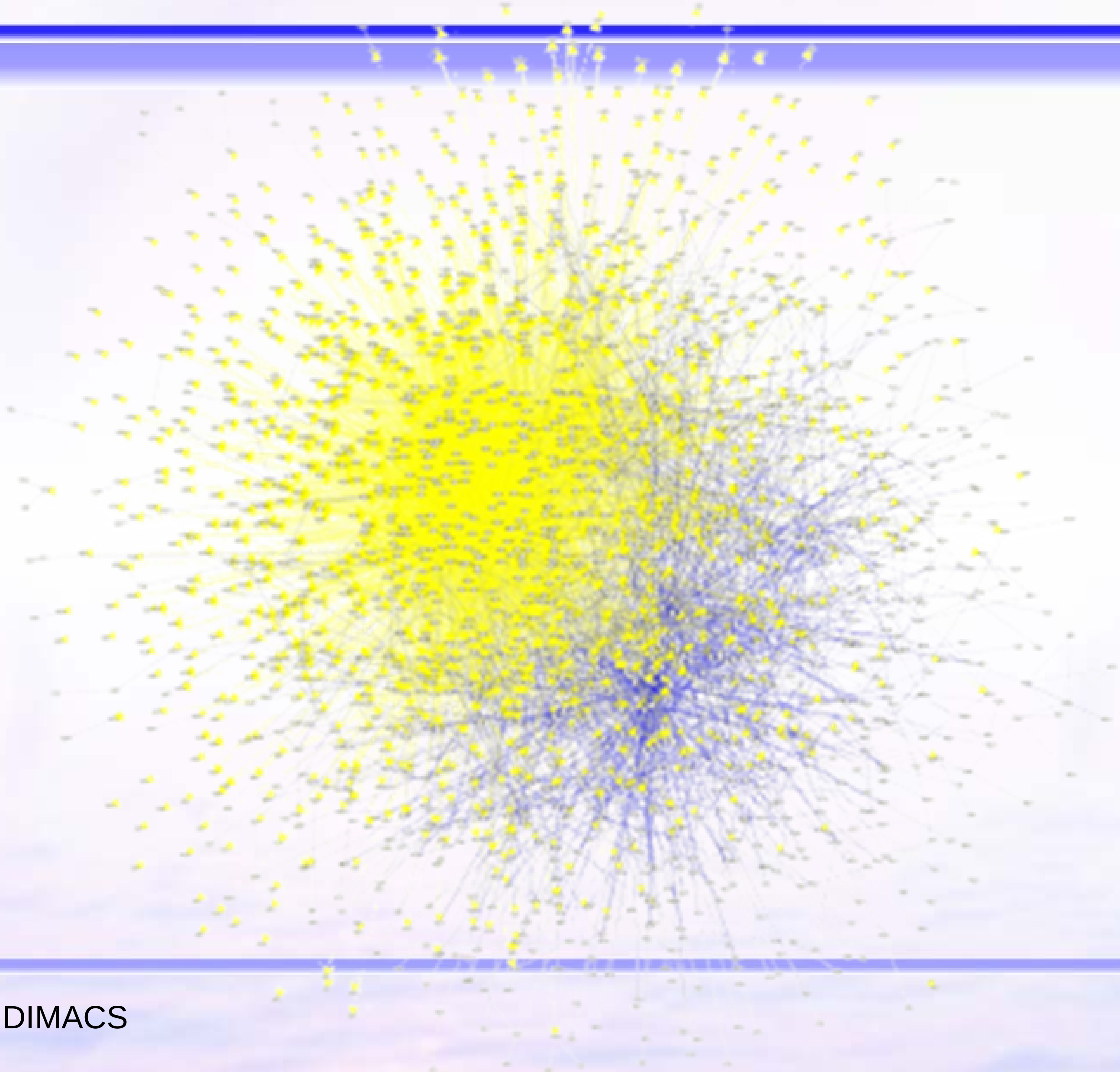
Koji Tsuda **National Institute for Advanced Industrial Science and Technology (AIST), Tokyo, Japan**

Joint work with Tsuyoshi Kato and Kiyoshi Asai

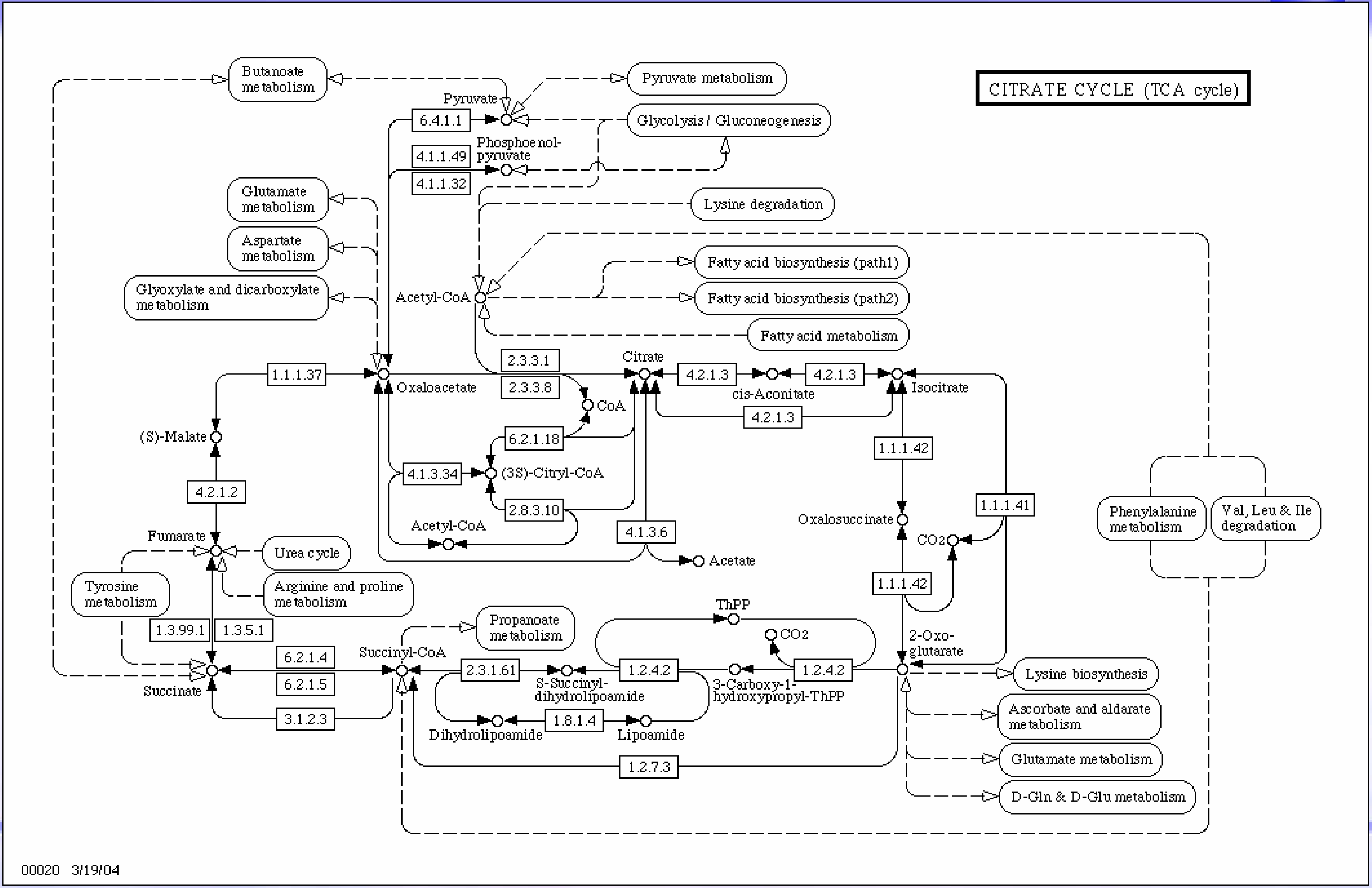
Biological Networks

- **Physical Interaction network**
  - **Edge ⇔ Two proteins physically interact (e.g. docking)**
- **Metabolic networks of enzymes**
  - **Edge ⇔ Two enzymes catalyzing successive reactions**
- **Gene regulatory networks**
- **Large graphs with sparse connections**
  - **1,000~10,000 nodes**
  - **10,000 – 100,000 edges**

Physical Interaction Network



Metabolic Network



Statistical Inference of Networks

- **Infer the network by data about proteins**
  - **Gene expressions, Phylogenetic profiles etc**

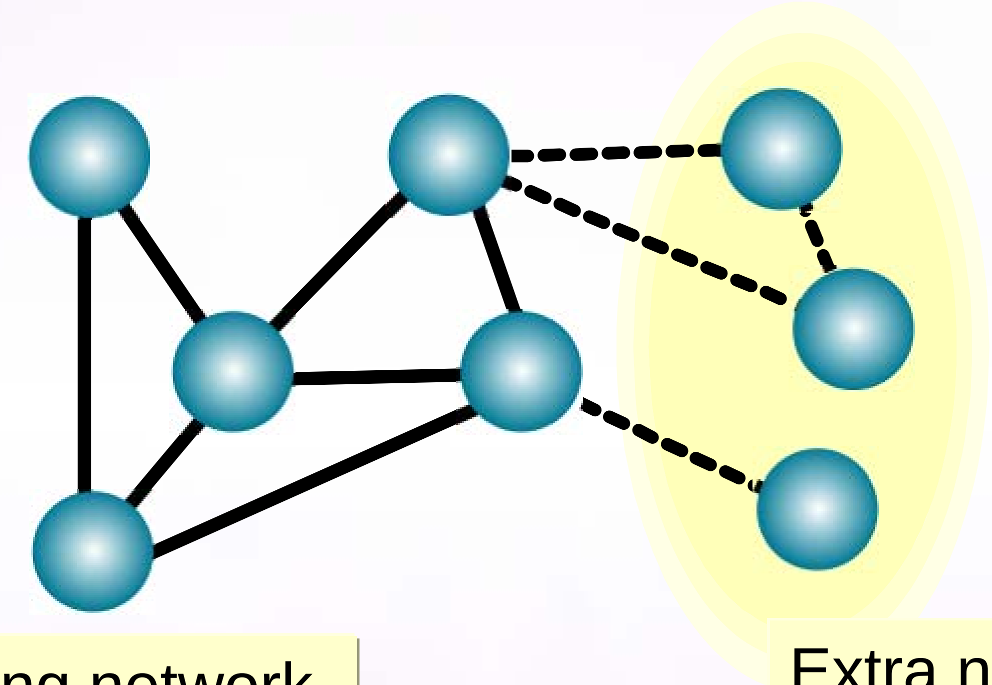
- **Propose a Kernel-based inference method**
  - **1. Supervised Inference**
    - **Learning from data and training network**
  - **2. Weighted combination of multiple data**
    - **Identify unnecessary data that do not contribute for network inference**

Unsupervised → Supervised Inference

- **Unsupervised network inference**
  - **Bayesian network (Friedman et al., 2000)**
  - **Infer every edge from scratch (no known edges)**

- **Supervised network inference**
  - **A part of the network is known (training network)**
  - **Infer the rest of the network from data and training net**
  - **Kernel CCA (Yamanishi et al., ISMB, 2004)**

Supervised Network Inference



Training network

Extra nodes

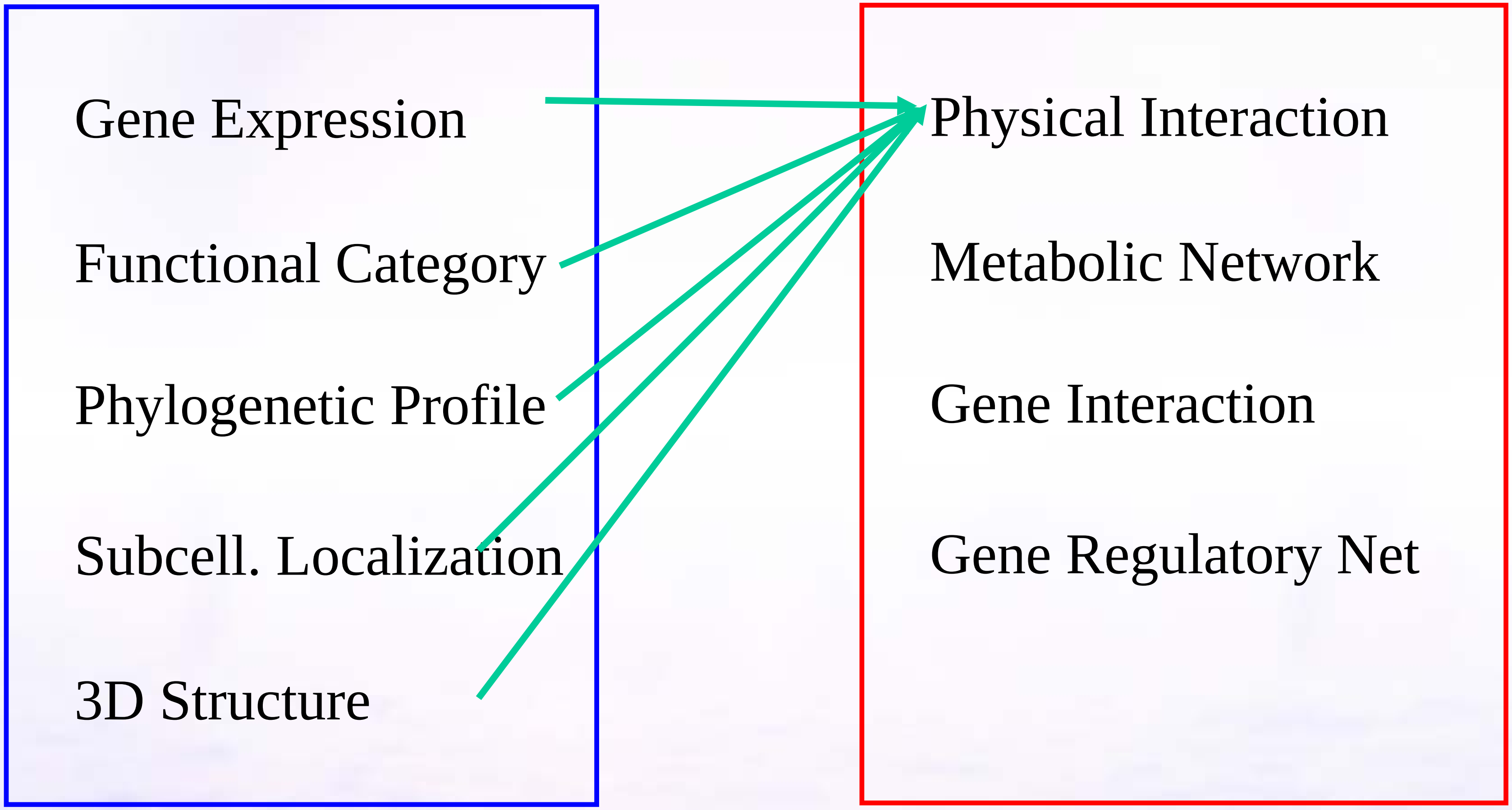
Single Data ⇨ Multiple Data

- **Multiple data for inferring networks**
  - **Gene expression profiles**
  - **Subcellular locations**
  - **Phylogenetic profiles**

- **Identify relevant data for inference**
- **Weighted integration of multiple data !**
  - **Feature selection to data selection**

- **Kernel CCA: No mechanism for data selection**

Inferring a Network from Multiple Data



Gene Expression

Functional Category

Phylogenetic Profile

Subcell. Localization

3D Structure

Physical Interaction

Metabolic Network

Gene Interaction

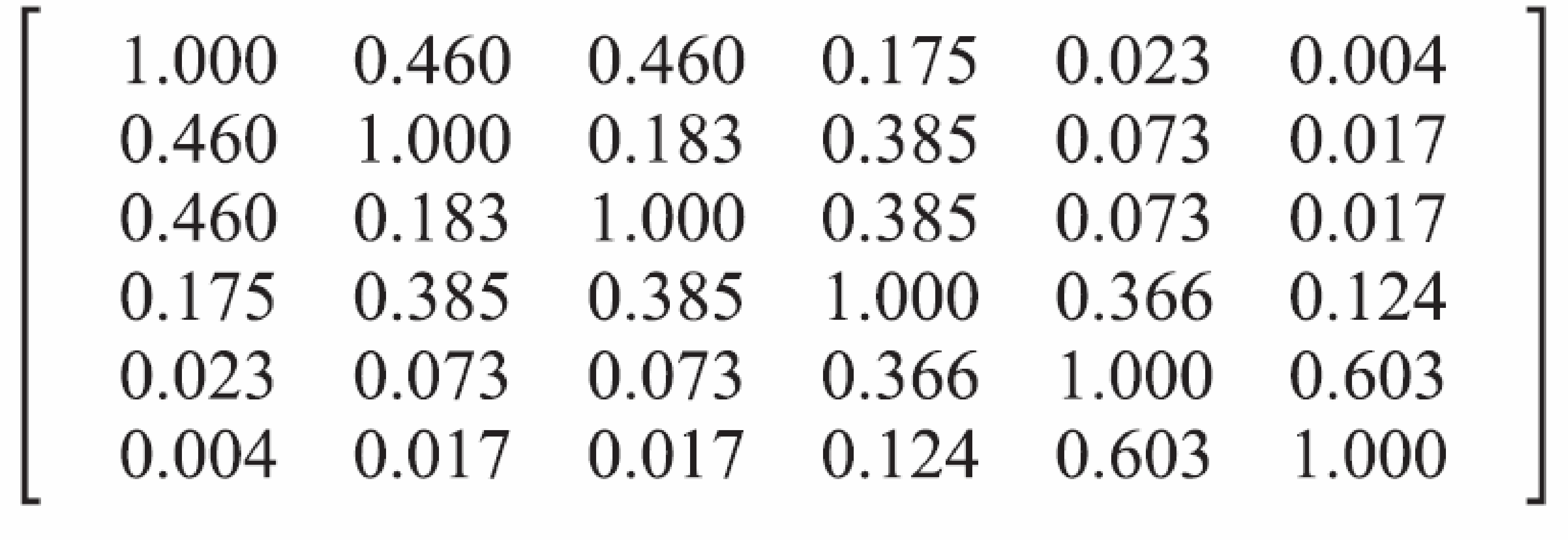
Gene Regulatory Net

Outline

- **Network Inference from a kernel matrix**
  - **Unsupervised, Single Data**
  - **Thresholding: Nearest neighbor connection**
- **Incorporating the training network**
  - **Supervised, Single Data**
  - **Kernel Matrix Completion (Tsuda et al., 2003)**
- **Weighted integration of multiple data**
  - **Supervised, Multiple Data**
  - **Weights determined by the EM algorithm**

Unsupervised, Single Data

- **Convert the data to a kernel matrix**
  - **Similarity among proteins**
  - **Gene expression: Pearson correlation**
  - **Phylogenetic profile: Tree kernel (Vert 2002)**
  - **3D structure: Graph kernel (Borgwardt et al., 2005)**

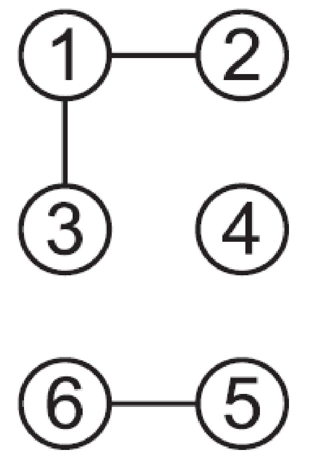


$$\begin{bmatrix} 1.000 & 0.460 & 0.460 & 0.175 & 0.023 & 0.004 \\ 0.460 & 1.000 & 0.183 & 0.385 & 0.073 & 0.017 \\ 0.460 & 0.183 & 1.000 & 0.385 & 0.073 & 0.017 \\ 0.175 & 0.385 & 0.385 & 1.000 & 0.366 & 0.124 \\ 0.023 & 0.073 & 0.073 & 0.366 & 1.000 & 0.603 \\ 0.004 & 0.017 & 0.017 & 0.124 & 0.603 & 1.000 \end{bmatrix}$$

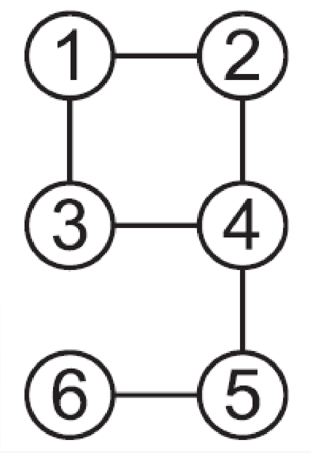
Construct the network by thresholding

- **Establish an edge where the kernel value is more than threshold**

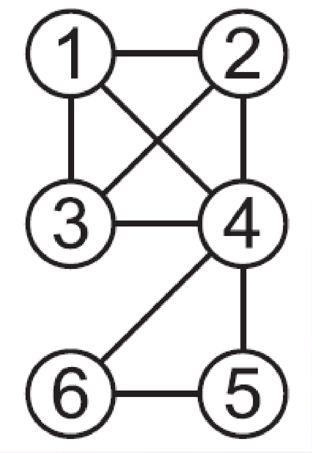
t=0.1



t=0.2

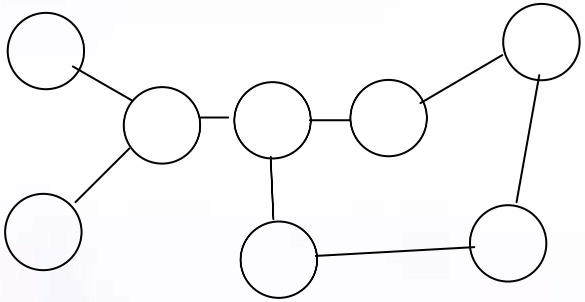


t=0.4



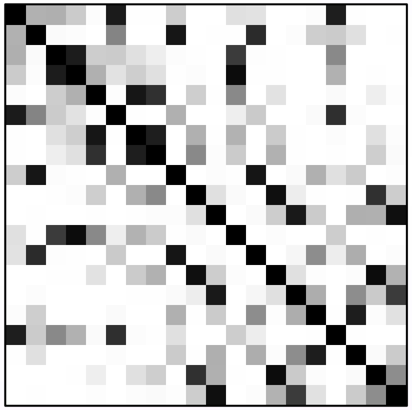
Supervised, Single Data

- **Known Training Network (Only for first n nodes)**



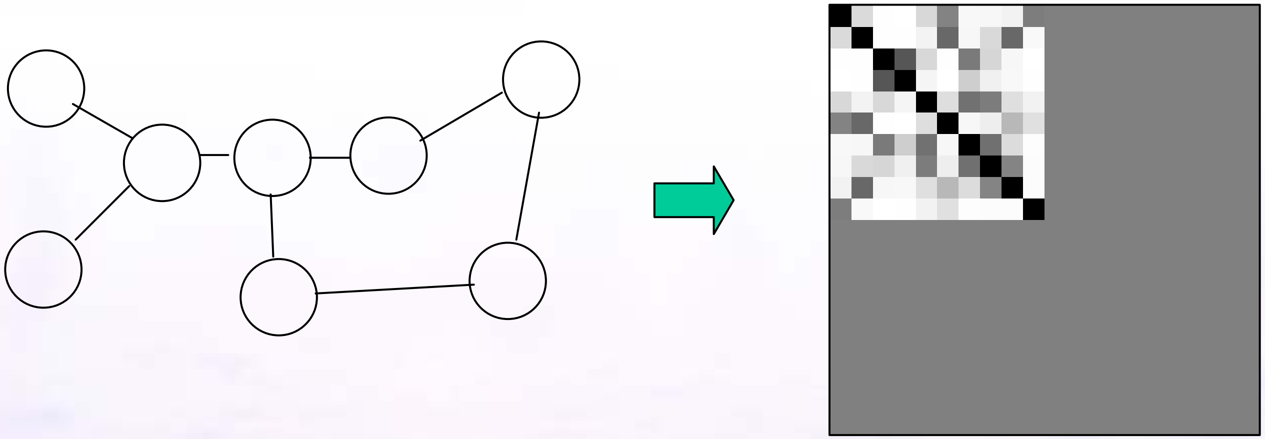
- **Data about all proteins**

Kernel Matrix



Incomplete kernel matrix from training network

- **Convert the training graph to a kernel matrix**
- **Synchronizing the representation**
- **Diffusion kernel (Kondor and Lafferty, 2002)**
  - **Measure closeness of nodes by random walking**



*Thresholding approximately recover the original network

Computation of Diffusion Kernel

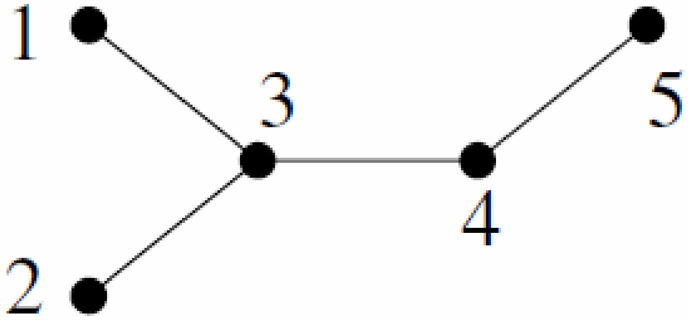
- **A: Adjacency matrix,**
- **D: Diagonal matrix of Degrees**
- **L = D-A: Graph Laplacian Matrix**
- ***Diffusion kernel matrix***

$$K = \exp(-\beta L)$$

  - β **: Diffusion paramater**

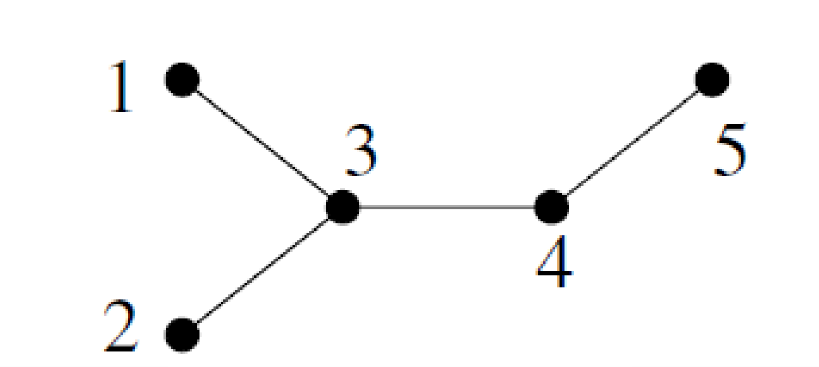
- **Characterizes closeness among nodes**
- **Often used with SVM (Lanckriet et al, PSB 2004)**

Adjacency Matrix and Degree Matrix



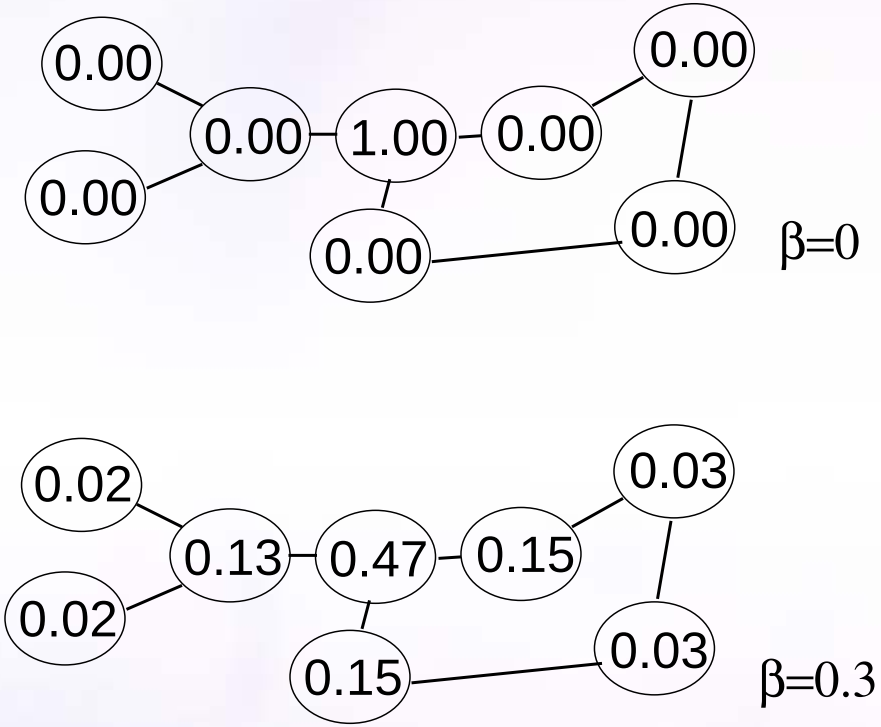
$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Graph Laplacian Matrix L



$$L = D - A = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

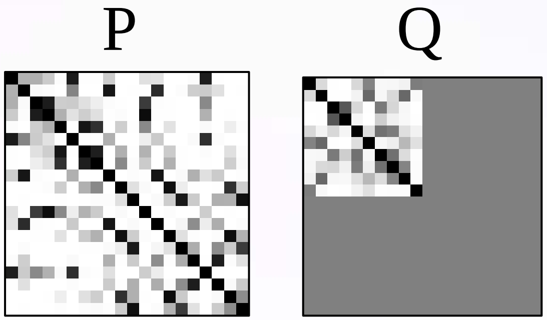
Actual Values of Diffusion Kernels



Closeness from the “central node”

Kernel Matrix Completion

- **P: Kernel matrix of the data**
- **Q: Incomplete kernel matrix**



$$Q = \begin{bmatrix} K_I & Q_{vh} \\ Q_{vh}^{\mathrm{T}} & Q_{hh} \end{bmatrix}$$

- **Missing values estimated by minimizing the KL divergence**

$$\textbf{Minimize}\ \mathrm{KL}(Q,P)\ \ \textbf{w.r.t.}\ Q_{vh}, Q_{hh}$$

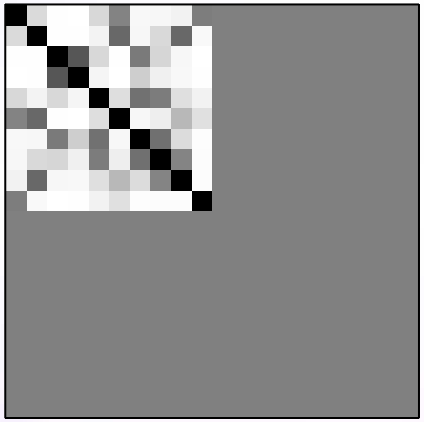
$$\mathrm{KL}(Q,P) = \tfrac{1}{2}\,\mathrm{tr}(P^{-1}Q) - \tfrac{1}{2}\,\mathrm{logdet}(P^{-1}Q) - \tfrac{1}{2}\,\ell$$

- **Closed from solution Q***
- **Threshold Q* to obtain the network**

Supervised, Multiple Data

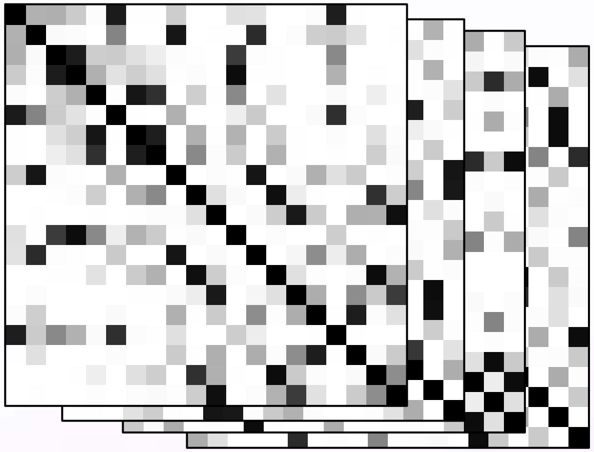
- **Known Training Network**

Diffusion Kernel Matrix

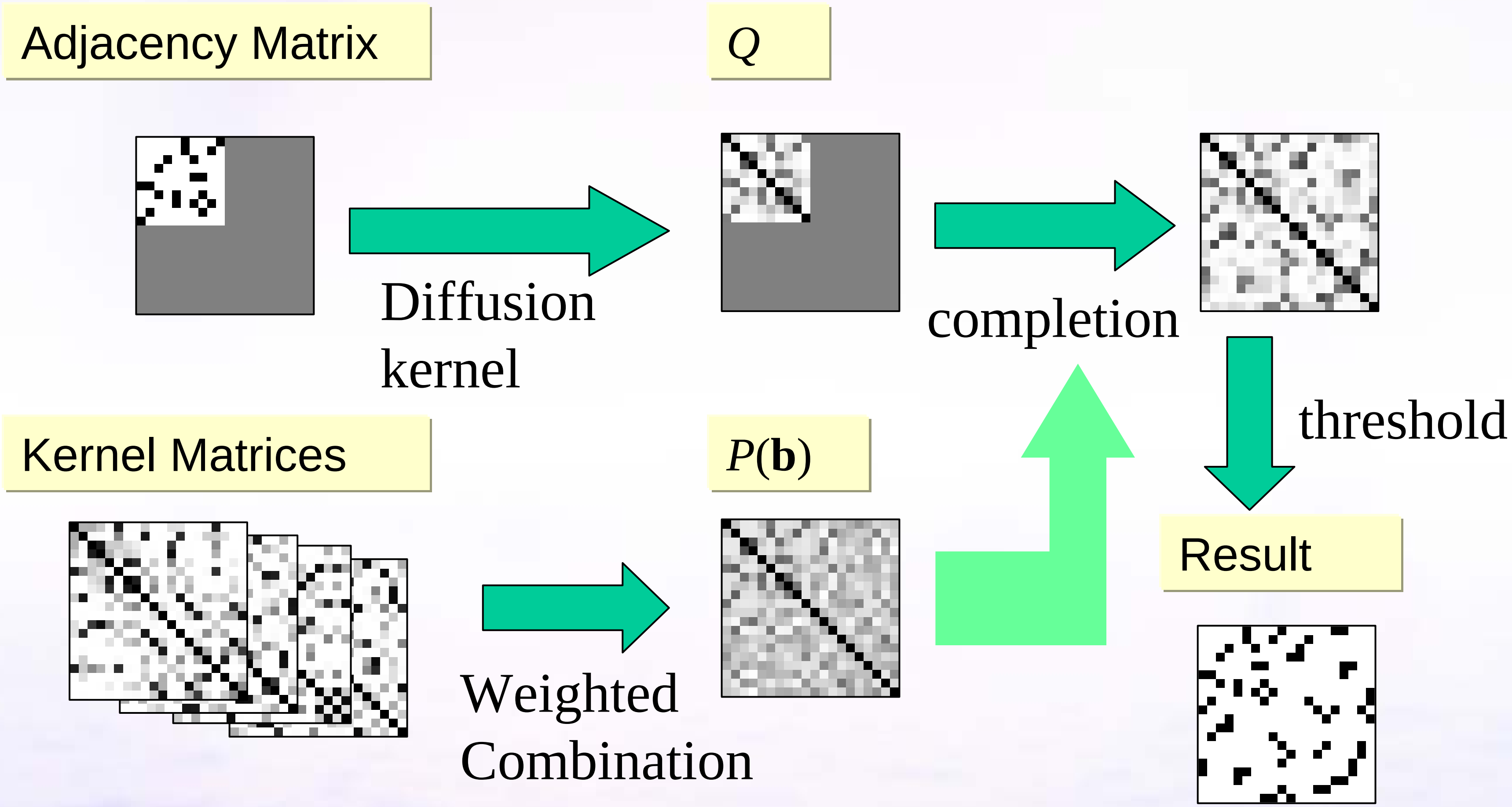


- **Multiple data about all proteins**

Kernel Matrices



Overview of Our Approach



Adjacency Matrix

Diffusion kernel

Q

completion

threshold

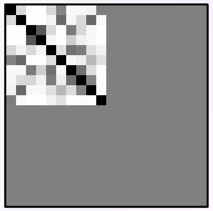
Kernel Matrices

Weighted Combination

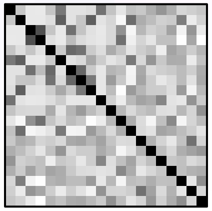
$P(\mathbf{b})$

Result

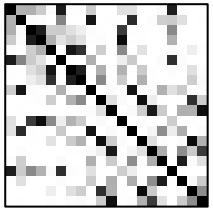
Notations



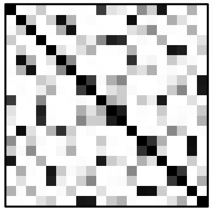
$$Q = \begin{bmatrix} K_I & Q_{vh} \\ Q_{vh}^{\mathrm{T}} & Q_{hh} \end{bmatrix}$$



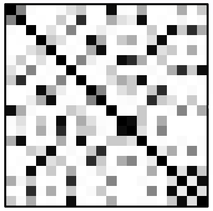
$P(\mathbf{b})$



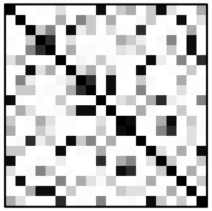
K_1



K_2



K_3



K_4

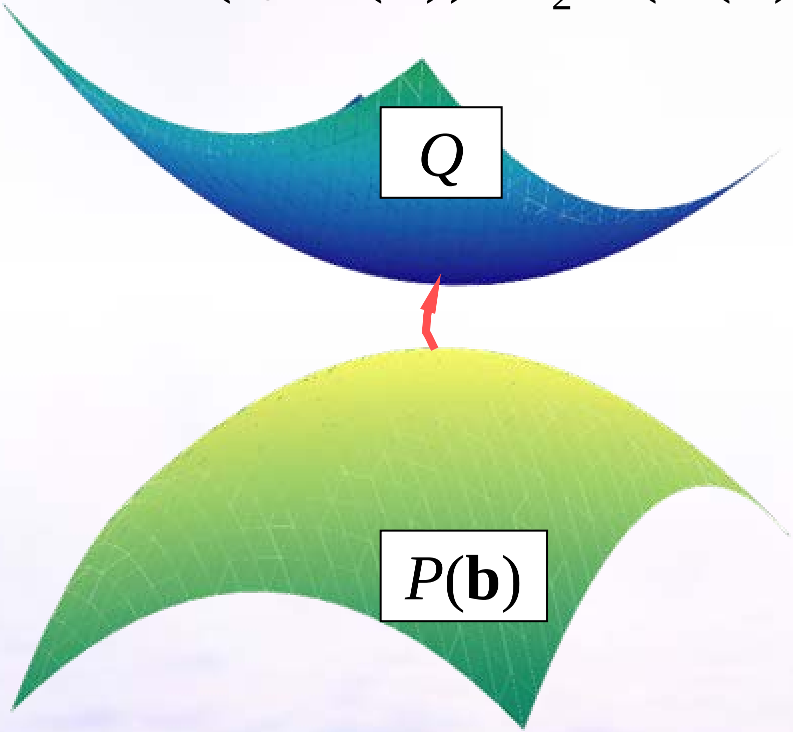
$$P(\mathbf{b}) = \sum_{i=1}^{n_k} b_i K_i + \sigma^2 I$$

Unknowns: Submatrices Q_{vh}, Q_{hh}, Weights $\mathbf{b}$

Objective Function

- **KL divergence**

$$\mathrm{KL}(Q, P(\mathbf{b})) = \tfrac{1}{2}\,\mathrm{tr}(P(\mathbf{b})^{-1}Q) - \tfrac{1}{2}\,\mathrm{logdet}(P(\mathbf{b})^{-1}Q) - \tfrac{1}{2}\,\ell$$



Minimize w.r.t.

Submatrices Q_{vh}, Q_{hh},
Weights $\mathbf{b}$

Solved by the EM algorithm

EM Algorithm

- **Repeat the following two steps**
  1. **E-step: minimize** $\mathrm{KL}(Q, P(\mathbf{b}))$ **w.r.t.** $\boldsymbol{Q_{vh}}$**,**$\boldsymbol{Q_{hh}}$
  2. **M-step: minimize** $\mathrm{KL}(Q, P(\mathbf{b}))$ **w.r.t.** b

- **E-step: Same as the single kernel case**

- **M-step: Cannot be solved in closed form**

EM Algorithm for Extended Matrices

- **Extended Kernel Matrices**

$$\tilde{Q} = \begin{bmatrix} Q & Q_{xz} \\ Q_{xz}^{\mathrm{T}} & Q_{zz} \end{bmatrix} \qquad R(\mathbf{b}) = \begin{bmatrix} P(\mathbf{b}) & \Lambda \\ \Lambda^{\mathrm{T}} & \sigma^2 I \end{bmatrix}$$

where

$$Q_{xz} \in \Re_{\ell \times n_k \ell}, \quad Q_{zz} \in \Re_{n_k \ell \times n_k \ell}$$

$$K_i = \Lambda_i \Lambda_i^{\mathrm{T}}, \qquad \Lambda = \left\lfloor \Lambda_1, \cdots, \Lambda_{n_k} \right\rfloor$$

- **The solution of the following problem is also optimal in the original problem**

$$\min_{Q_{vh}, Q_{hh,} Q_{xz}, Q_{zz}, \mathbf{b}} \mathrm{KL}(\tilde{Q}, R(\mathbf{b}))$$

Solutions of the steps

- **E-step**

$$Q_{vh} = K_I P_{vv}^{-1} P_{vh}, \qquad Q_{hh} = P_{hh} - P_{vh}^T P_{vv}^{-1} P_{vh} + P_{vh}^T P_{vv}^{-1} K_I P_{vv}^{-1} P_{vh}$$

$$V_z^{-1} = \sigma^{-2} \Lambda^T \Lambda, \qquad Q_{zz} = V_z + \sigma^{-4} V_z \Lambda^T Q \Lambda V_z$$

- **M-step**

$$b_k = \frac{1}{N} \sum_{j=(k-1)N+1}^{kN} \left[Q_{zz} \right]_{jj}$$

Edge prediction experiments

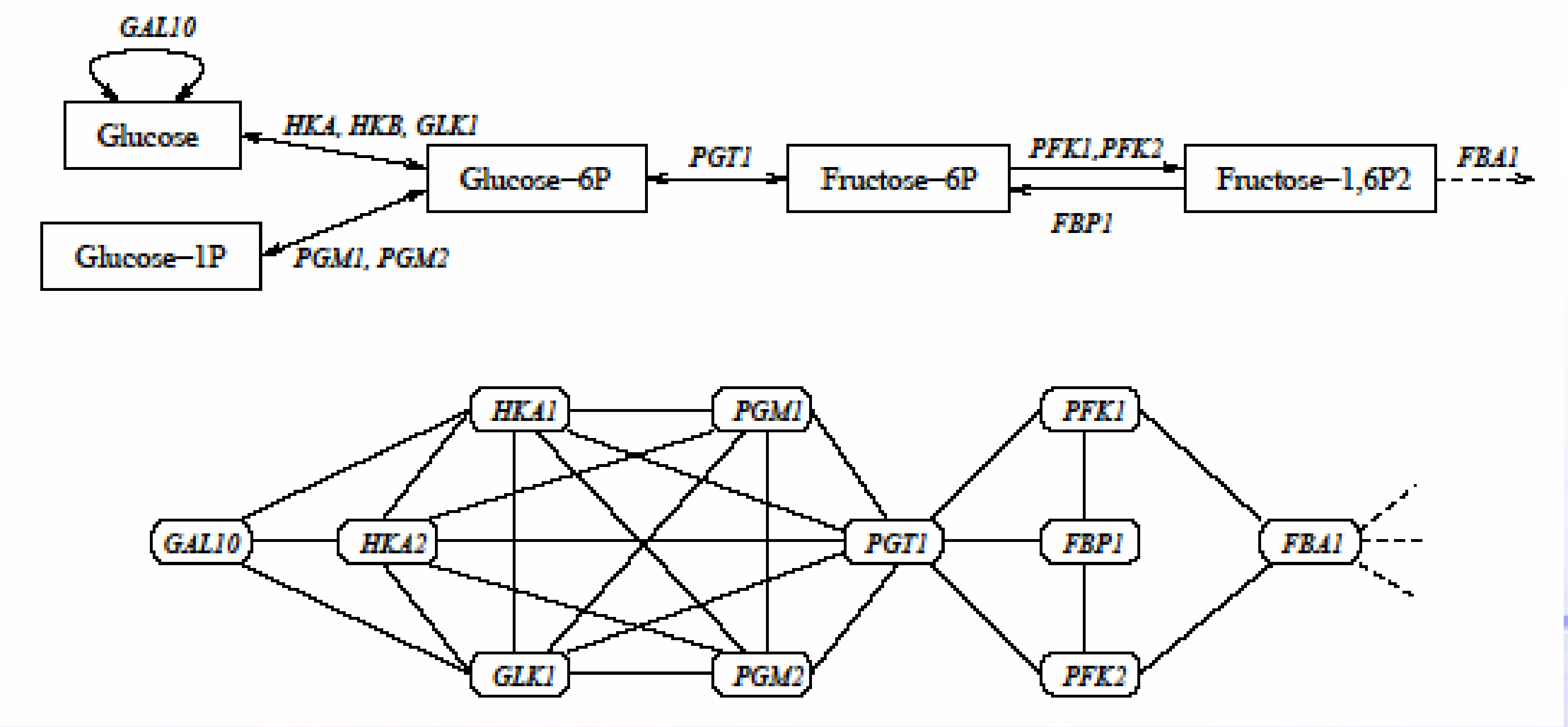
Network	**・Metabolic Network（KEGG）** **・Protein interaction network（von Mering, 2002）**
Data	**・exp: gene expression** **・y2h: Interaction net by yeast2hybrid** **・loc: subcellular location** **・phy: phylogenetic profile** **・rnd1,…,rnd4: random noise**
Methods	**・Q: Proposed method** **・P: Simple combination of kernel matrices** **・cca: kernel CCA (without noises)**
Evaluation	**ROC score of edge prediction accuracy** **(10-fold cross validation)**

Metabolic network

- Made from LIGAND Database (KEGG)

(Vert and Kanehisa, NIPS, 2003)

- Connect enzymes of two successive reactions
- 769 nodes, 3702 edges



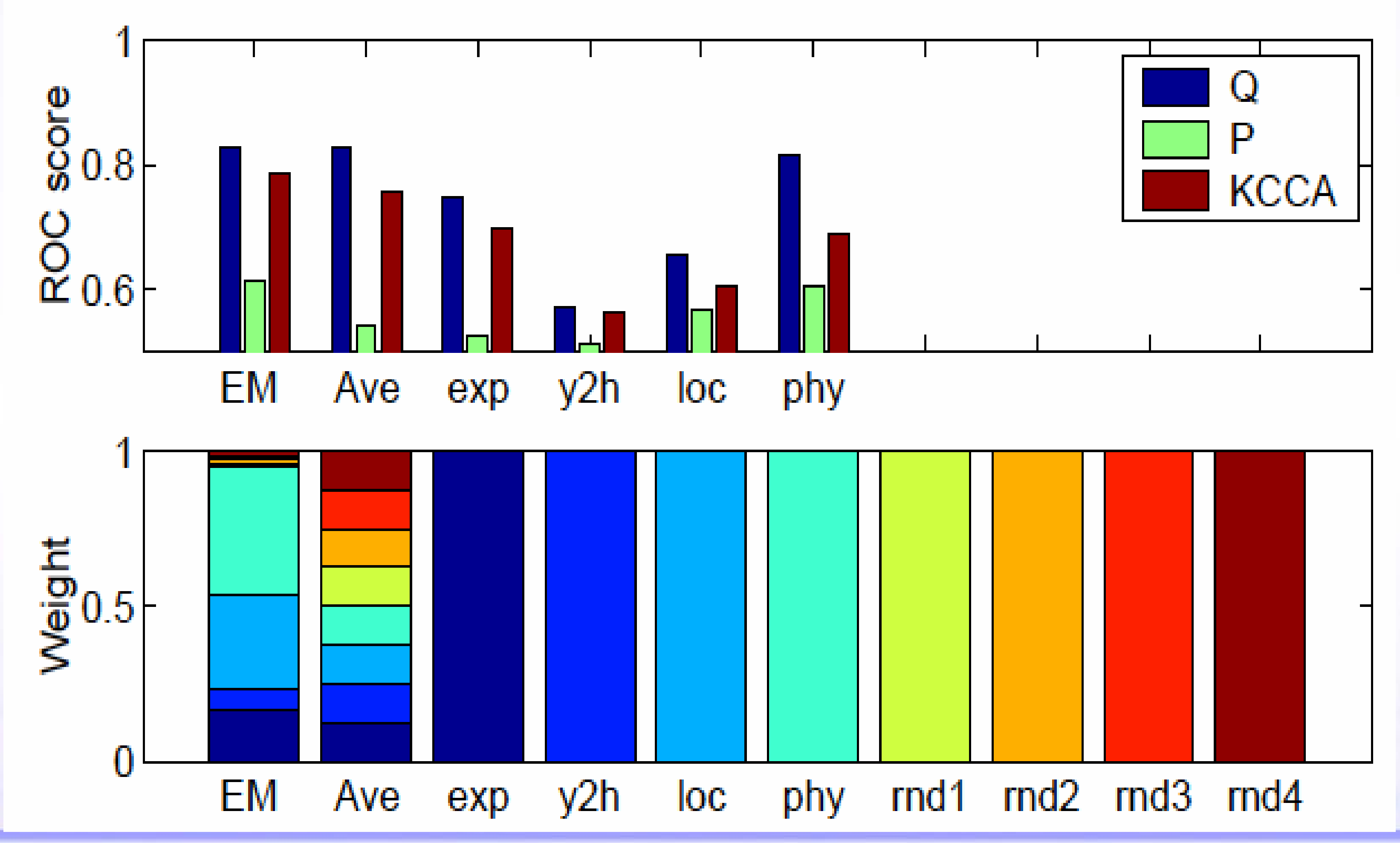
Interaction network (Von Mering et al., Nature, 2002)

- **Middle Confidence**
- **Interactions validated by multiple experiments**
  - **High-throughput yeast two hybrid**
  - **Correlated mRNA expression**
  - **Genetic interaction**
  - **Tandem affinity purification,**
  - **High-throughput mass-spectrometric protein complex identification**
- **984nodes, 2438 edges**

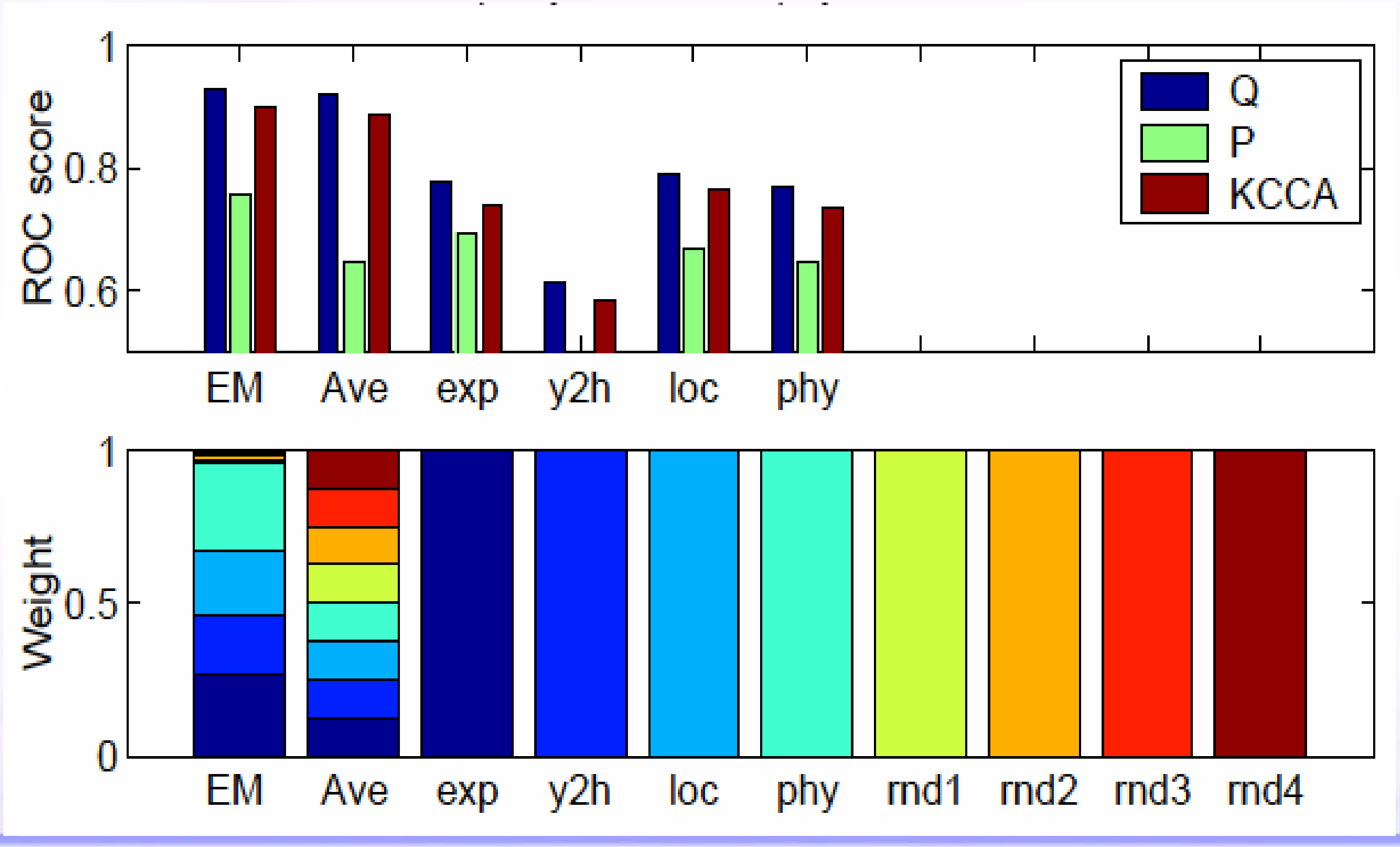
Dataset Details

Metabolic Net	**http://www.genome.jp/kegg/**
Interaction	**Von Mering et al., Nature, 417 399--403 , 2002**
Expression	**Spellman et al., MBC, 9, 3273—3297, 1998** **Eisen et al., PNAS, 95, 14863—8, 1998**
Y2H	**Ito et al., PNAS, 98, 4569—74, 2001** **Uetz et al., Nature, 10, 601—3, 2000**
Subcellular location	**Huh et al. Nature, 425, 686-91, 2003**
Phylogenetic profile	**http://www.genome.jp/kegg/**

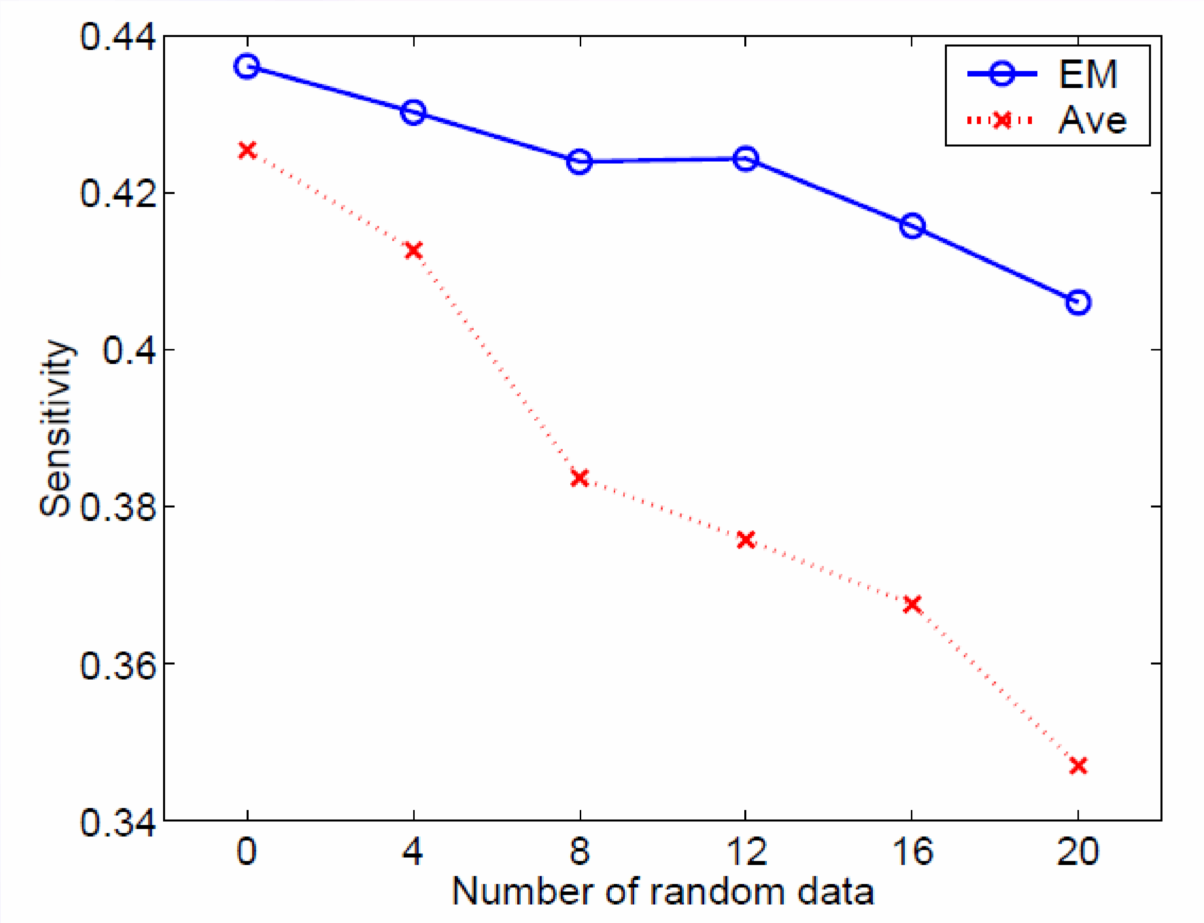
Metabolic Network



Physical Interaction Network



Introduce More Random Matrices (Metabolic network)



Sensitivity at 95% specificity

Summary of Experiments

- **Simple combination (P)　＜ Completed matrix（Q)**
  - **Training network is essential**

- **Selection did not improve accuracy**

- **Accuracy comparable to kernel CCA**

- **Automatic selection of datasets**
    - **4 noise kernel matrices removed**

Conclusion

- **Supervised Inference of Network**
  - **Part of network known**
  - **Selection from multiple data**
  - **Formulation as kernel completion problem**
  - **Validation experiments on metabolic and interaction networks**
- **Future work**
  - **Biological interpretation of selection results**
  - **Applications to non-bio data**

T. Kato, K. Tsuda, and K. Asai. Selective integration of multiple biological data for supervised network inference. ***Bioinformatics*****, 21(10):2488--2495, 2005.**

Experiments

Data

- *The functional catalogue provided by the MIPS Comprehensive Yeast Genome Database (CYGD-mips.gsf.de/proj/yeast).*

- *In a total of 6355 yeast proteins, however, Only 3588 have class labels.*

13 CYGD functional Classes

1. metabolism
2. energy
3. cell cycle and DNA processing
4. transcription
5. protein synthesis
6. protein fate
7. cellular transportation and transportation mechanism
8. cell rescue, defense and virulence
9. interaction with cell environment
10. cell fate
11. control of cell organization
12. transport facilitation
13. others

Data

Network created from Pfam domain structure. A protein is represented by a 4950-dimensional binary vector, in which each bit represents the presence or absence of one Pfam domain. An edge is created if the inner product between two vectors exceeds 0.06. The edge weight corresponds to the inner product.

Co-participation in a protein complex (determined by tandem affinity purification, TAP). An edge is created if there is a bait-prey relationship between two proteins.

Protein-protein interactions (MIPS physical interactions)

Genetic interactions (MIPS genetic interactions)

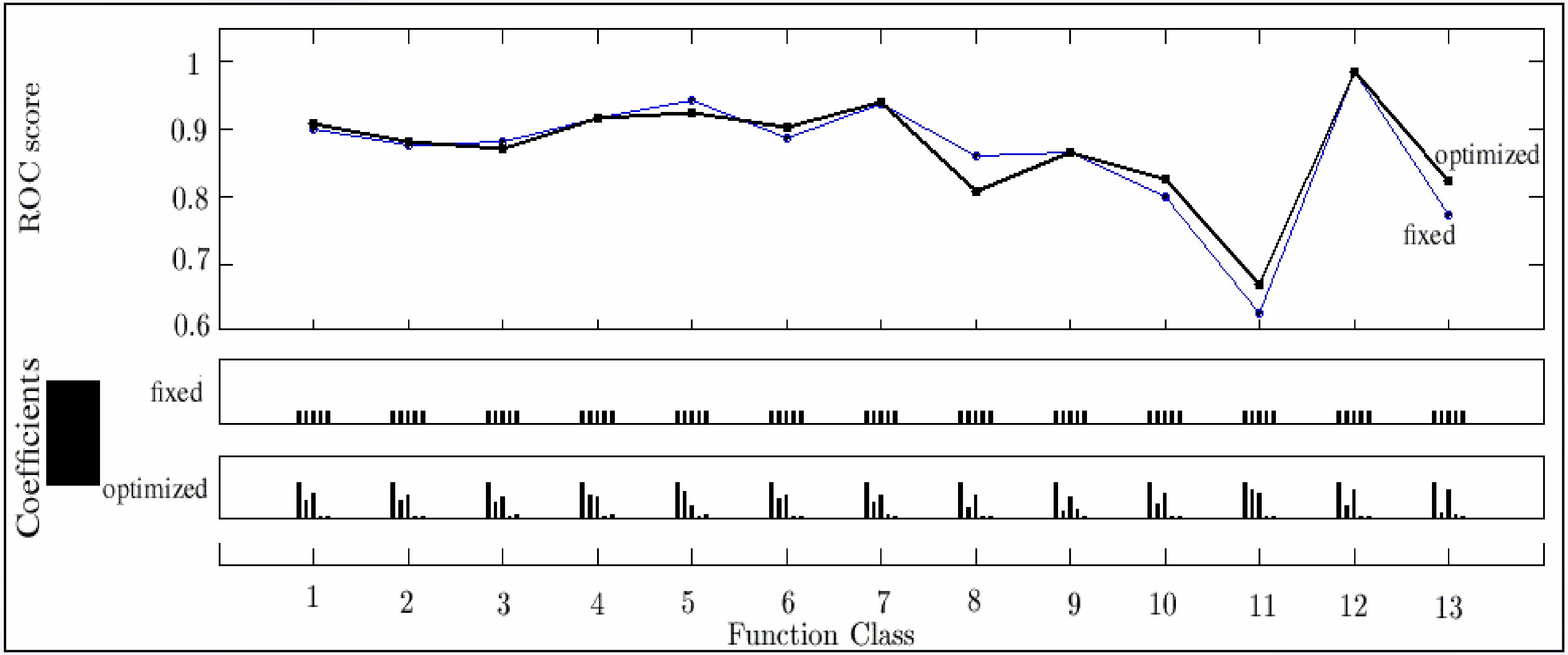
Network created from the cell cycle gene expression measurements [Spellman et al., 1998]. An edge is created if the Pearson coefficient of two profiles exceeds 0.8. The edge weight is set to 1. This is identical with the network used in [Deng et al., 2003]

Design

The Performance Comparison Between …

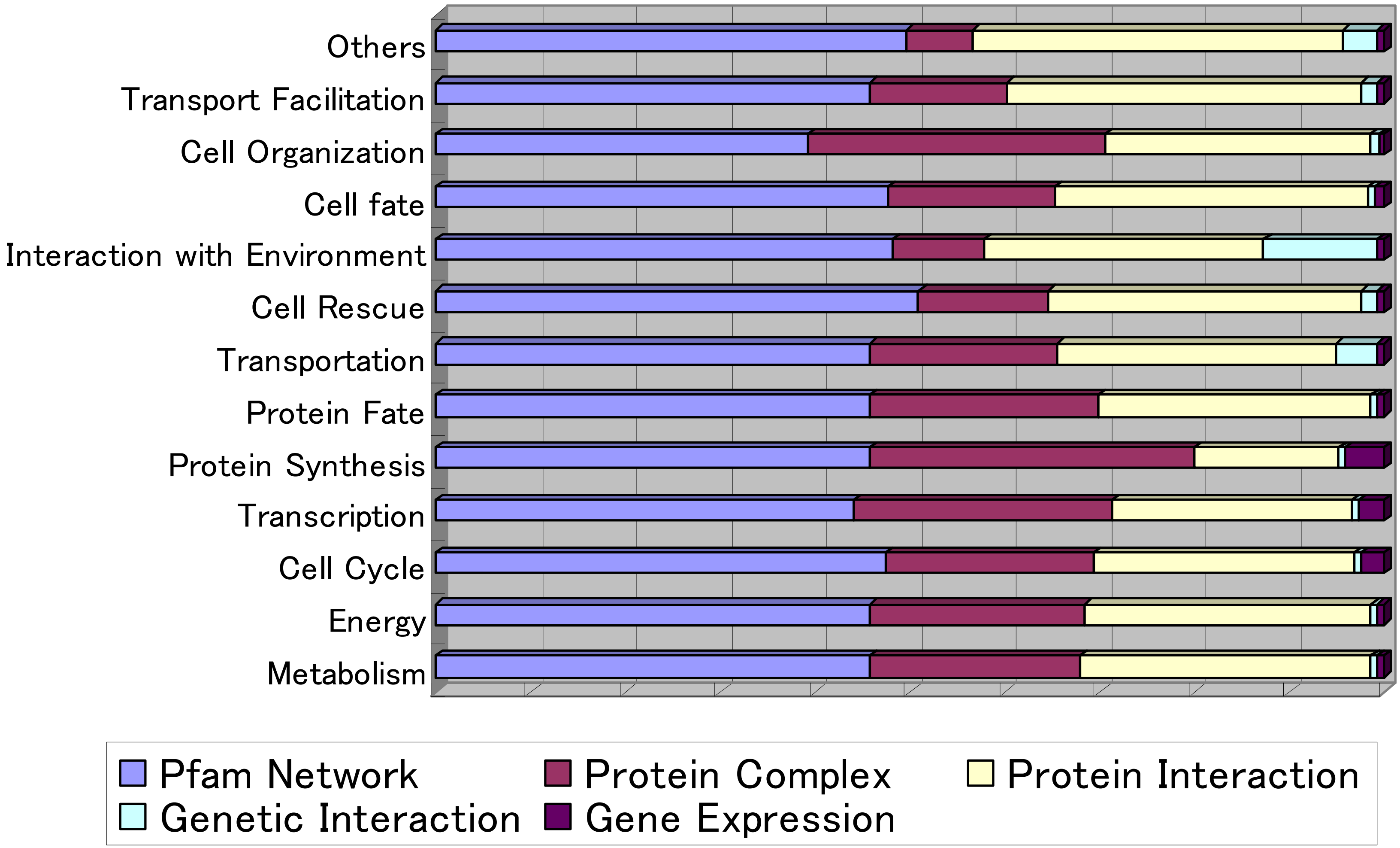
L_k	*Laplacian of Individual Graph*
L_{opt}	*Laplacian of Combined Graph with Optimized Weights*
L_{fix}	*Laplacian of Combined Graph with Fixed (equal) Weights*
MRF	*Markov Random Field, proposed by Deng et al [2003]*
SDP/SVM	*Semi-definite Programming based Support Vector Machines, proposed by Lanckriet et al [2004]*

Results : ROC score with Weights – Classwise, L_{fix} vs. L_{opt}



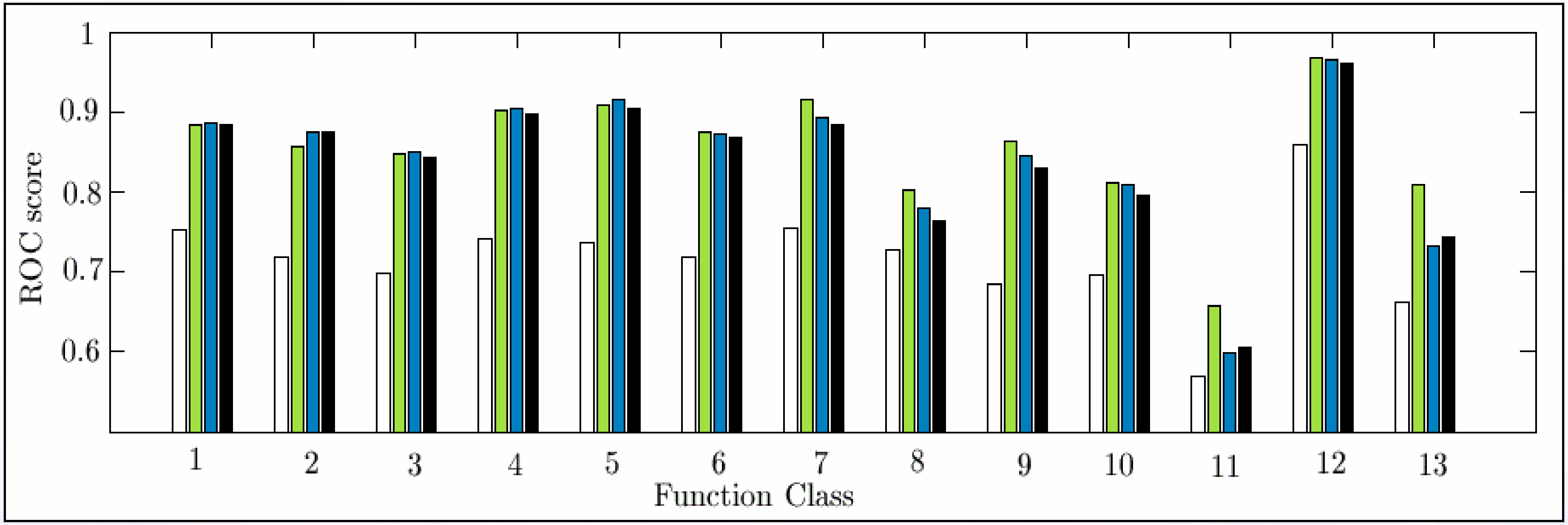
The optimization of weights did not always lead to better ROC scores (except for the classes 10, 11, 13). However, the advantage of L_{opt} is that the redundant networks are automatically identified.

Obtained Weight Parameters



Results : *ROC scores of L_{opt}, L_{fix} , MRF, and SDP/SVM*

White: MRF
Green: SDP/SVM
Blue: L_{fix}
Black: L_{opt}



For most classes, the proposed method achieves high scores, which are similarto the SDP/SVM methods.

Results : Computational Time

Average Computation Time

Combining Graphs with Fixed Weights : *1.41 seconds* (std. 0.013)*

Combining Graphs with Optimized Weights : *49.3 seconds* (std. 14.8)*

Nearly linearly proportional to the number of non-zero entries of sparse matrices

SDP/SVM : *Approx. 60 min (G. Lanckriet, personal communication)*

$O(n^3)+O((m+n)^2n^{2.5})$

* *measured in a standard 2.2Ghz PC with 1GByte memory*

Conclusion

- **Extended Label Propagation for Multiple Networks**
- **Good Prediction Accuracy in Yeast Protein Function Experiments**
- **Fast and Scalable**
- **Redundant / Irrelevant Networks Excluded**
- **Biological Implications??**