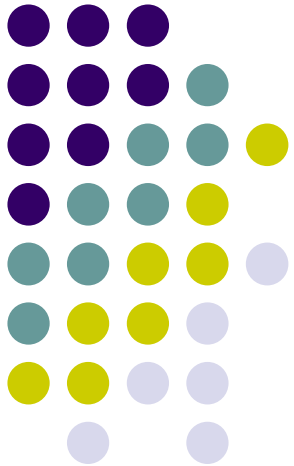
Carnegie Mellon

A Mixture of Feature Experts Approach for Protein-Protein Interaction Prediction

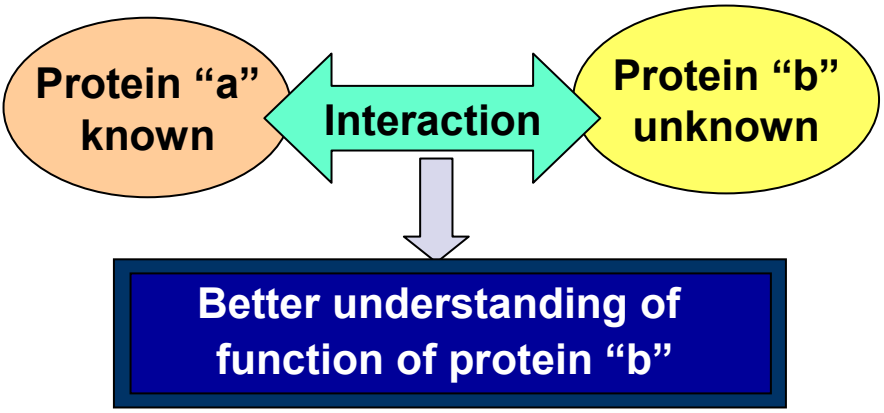


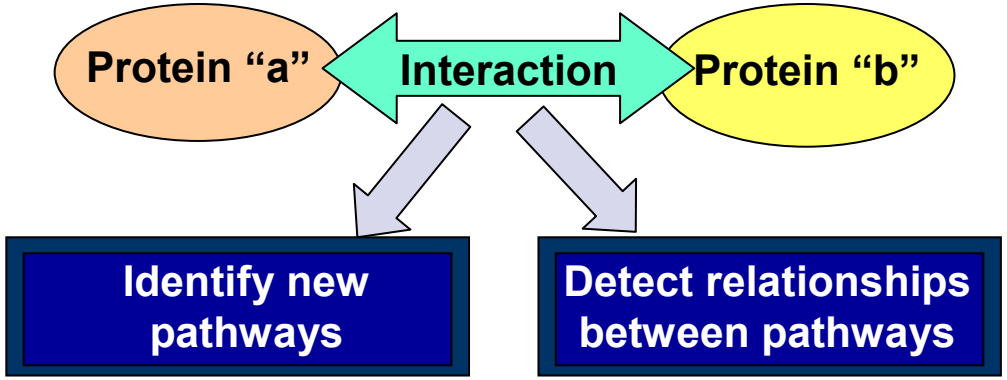
Yanjun Qi[1], Judith Klein-Seetharaman[1,2] and Ziv Bar-Joseph[1]

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh
[2]Department of Pharmacology, University of Pittsburgh Medical School, Pittsburgh

Email: {qyj, judithks, zivbj}@cs.cmu.edu

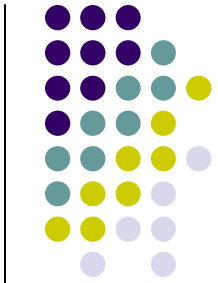
Importance of Protein Interactions





- Need: comprehensive identification of Protein-Protein Interactions (PPI)
  - To systematically define proteins' functions
  - To decipher the molecular mechanisms underlying given biological functions

Approaches



- Experimental:
  - direct high throughput data: mass spectrometry and yeast-two-hybrid, Y2H
    - High false-positive and false-negative rate, especially Y2H
    - Incomplete, with majority remains to be discovered, especially for human
    - Surprisingly small overlap among different sets
- Computational:
  - treat direct data as features and use in combination with other implicitly related biological information
    - Example: If two proteins are co-expressed, they may interact.

Our Goal :

- Integrate multiple biological data sources to :
  - Predict protein interacting pairs in yeast more accurately and completely
    - Different example may benefit from using different feature sets
  - Give guidance /help for biological lab experiments
    - Useful for biologists to know which features contributed to specific predictions
      - (Researchers may have various opinions regarding the liability of diverse features)
      - (Different features also have diverse reliability)

Related Works

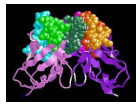
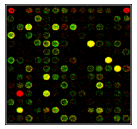
- Jansen,R. et al., Science 2003
  - Use Bayes classifier to classify candidate protein pairs interact or not
- Zhang,L. et al., BMC Bioinformatics 2004
  - Decision tree to classify a candidate protein pair in same complex or not
- Ben-Hur,A. et al., ISMB 2005
  - kernel method in conjunction with a support vector machine classifier
- Qi,Y., et al, PSB 2005
  - Random Forest Similarity based weighted k-NN classifier

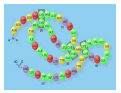
supervised classification

Above methods either estimate feature importance globally or implicitly for a specific interaction prediction!

Features Used

- Overall, four categories: (roughly homogeneous within category)
  - Direct high-throughput experimental data
    - (two-hybrid screens and mass spectrometry)
  - Indirect high throughput data
    - (gene expression, protein-DNA binding etc.)
  - Functional annotation data
    - (gene ontology annotation, mips annotation, etc.)
  - Sequence based data sources
    - (domain information, gene fusion, homology based PPIs, etc.)

direct

Indirect

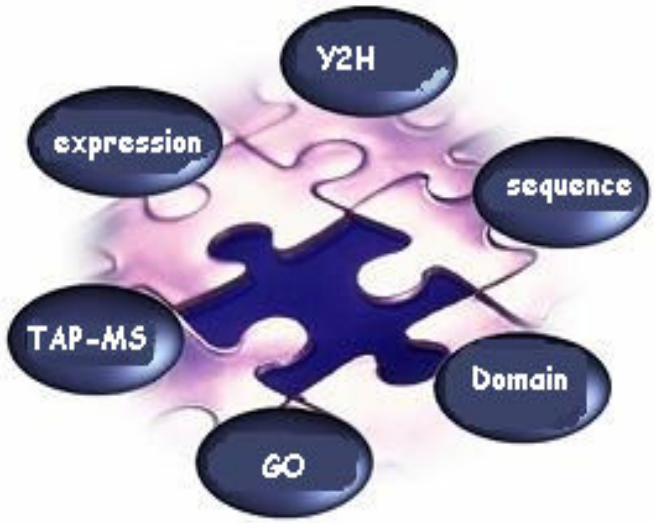
Data Properties

Challenges:

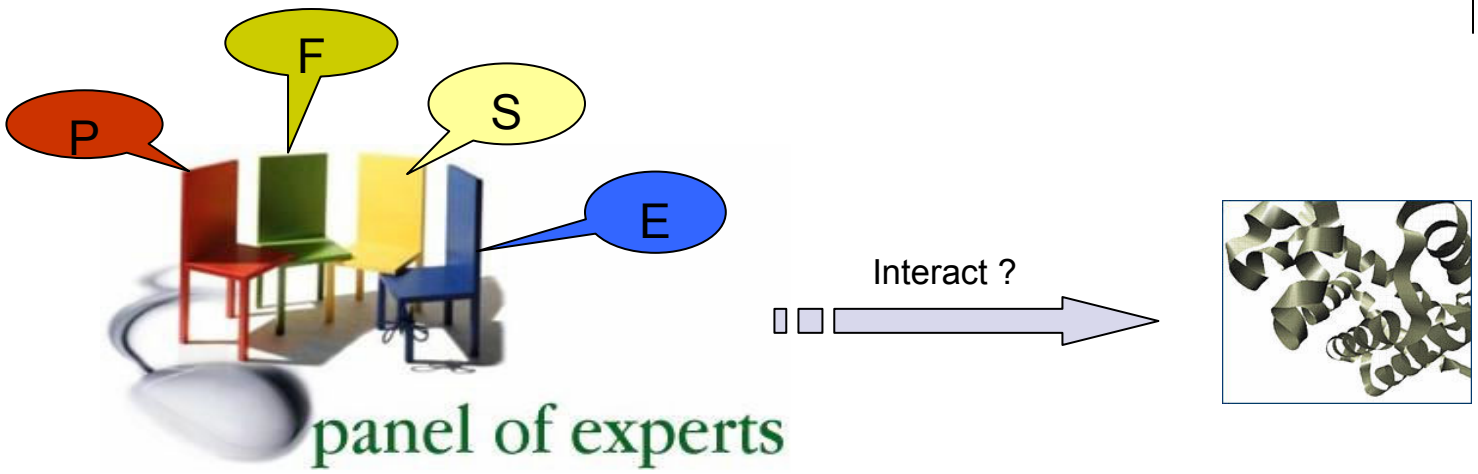
- Most data are noisy
- Many missing values
- Data is often correlated

Potential advantages:

- Data from heterogeneous sources
- Redundant features are also important and can provide complementary information



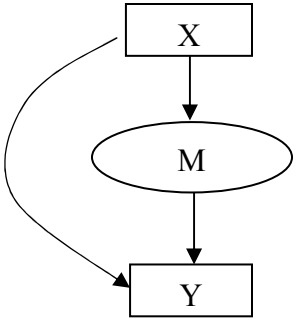
Method – Mixture of Feature Experts



- Make protein interaction prediction by
  - Weighted voting from the four roughly homogeneous feature categories
  - Treat each feature category as a prediction expert
  - The weights are also dependent on the input feature

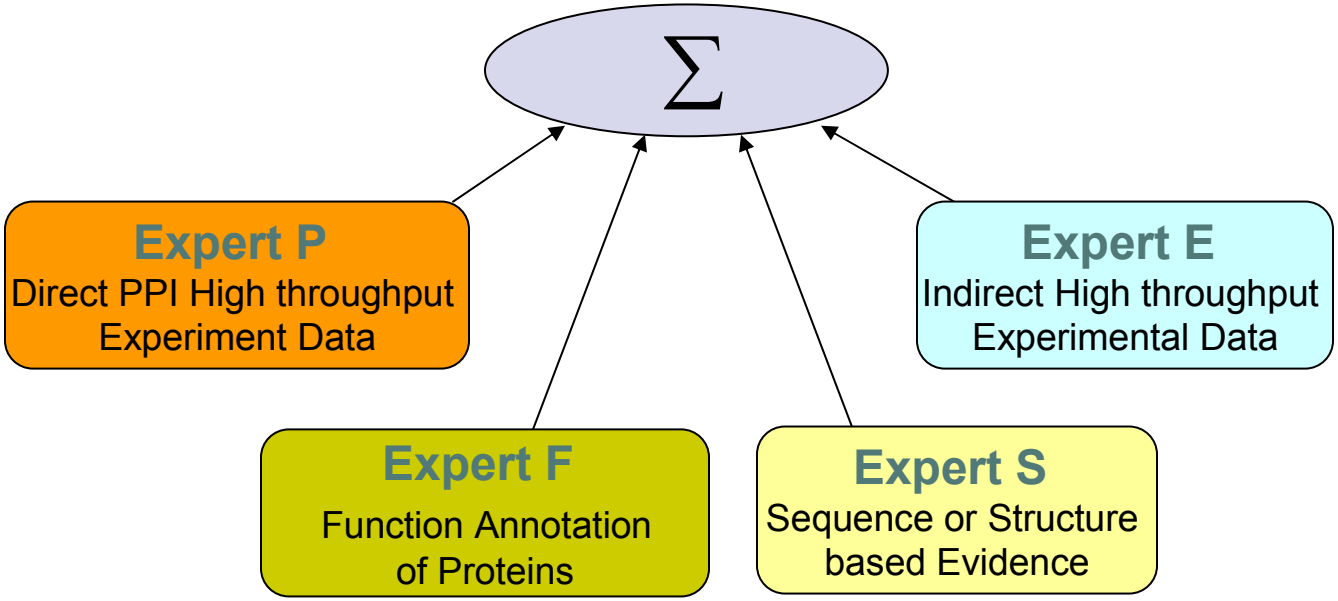
Mixture of Experts (ME)

- A single layer tree with experts at the leaves
- A root gate is used to integrate experts
- Weights assigned on each expert by the root gate
  - Depends on the input set for a given pair
- Hidden variable “M” represents the choice of expert



$$p(Y \mid X) = \sum_{M} p(Y \mid X, M)\, p(M \mid X)$$

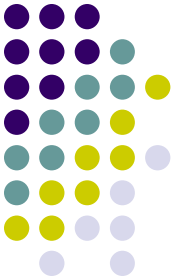
Mixture of Four Feature Experts



$$p(y^{(n)} \mid x^{(n)}) = \sum_{i=1}^{4} p(m_i^{(n)} = 1 \mid x^{(n)}, v) * p(y^{(n)} \mid x^{(n)}, m_i^{(n)} = 1, w_i)$$

- Parameters (w_i, v) are trained using EM
- Experts and root gate use logistic regression (ridge estimator)

Mixture of Feature Experts

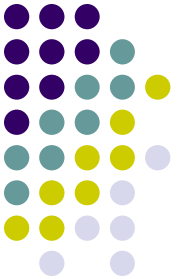


- Handling missing value
  - Add additional feature column for each feature having low feature coverage
  - ME could then also use present / absent information when weighting different features

- The posterior weight for expert *i* in predicting pair *n*
  - The weight can be used to indicate the importance of that feature category (expert) for this specific pair

$$h_i^{(n)} = P(m_i^{(n)} = 1 \mid y^{(n)}, x^{(n)}, v^t, w^t) = \frac{P(m_i^{(n)} = 1 \mid x^{(n)}, v^t) * p(y^{(n)} \mid x^{(n)}, m_i^{(n)} = 1, w_i^t)}{\sum_{j=1}^{4} P(m_j^{(n)} = 1 \mid x^{(n)}, v^t) * p(y^{(n)} \mid x^{(n)}, m_j^{(n)} = 1, w_j^t)}$$

Experiments



- Measurements
  - AUC score: The area under the ROC curve
  - Partial AUC score: measures the area under the ROC curve within a specific region
    - We are interested with the performance where the false positive rate is low
  - Tradeoff between accurateness / completeness

- Reference Set
  - Only a small positive (interacting) set available (small scale interaction experimental result)
  - Highly skewed class distribution
    - Much more non-interacting pairs than interacting pairs
  - The ratio of positive pairs to negative (random) pairs is roughly 1 : 600 in yeast

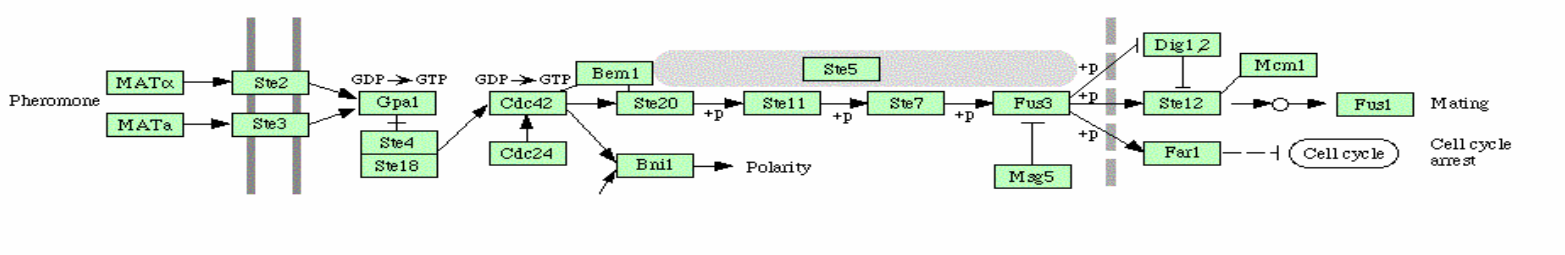
Performance Comparison

- Compare with four other classifiers
  - Random Forest (RF)
  - Logistic regression (LR)
  - Support Vector Machine (SVM)
  - Naïve Bayes (NB)

- Used randomly train & test style to evaluate the performance

Methods	**AUC**	AUC STD	**R50**	R50 STD
LR	0.8823	0.0330	0.2866	0.0707
NB	0.9349	0.0158	0.2486	0.0472
RF	0.9321	0.0142	0.2688	0.0482
SVM	0.9159	0.0247	0.2585	0.0638
ME	**0.9463**	**0.0137**	**0.3080**	**0.0780**

Validate on Yeast Pheromone Pathway



- 25 proteins involved in this pathway
- Test all possible 300 protein pairs
- 51 predicted interactions
  - 33 validated already
  - 18 newly predicted

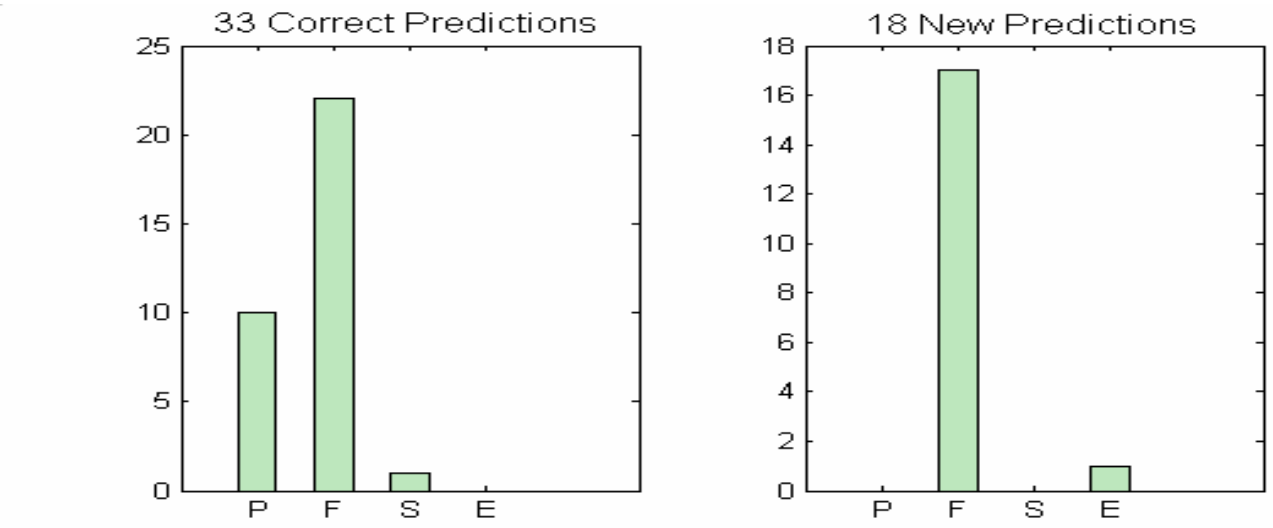
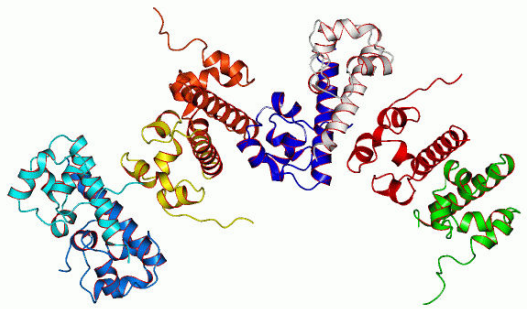


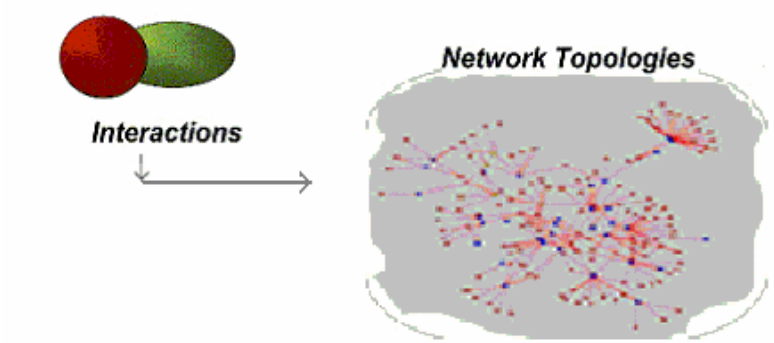
Figure: The frequency at which each of the four experts has maximum contribution among validated and predicted pairs

Future Work

- Extend to other species
  - (for example, Human)

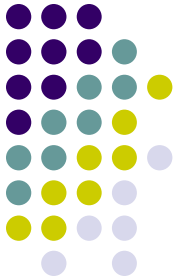


- Graph mining on the full predicted protein interaction network



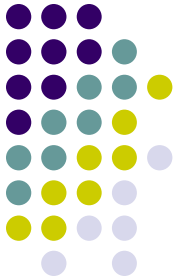
Thank You

Questions ?



Extra Slides

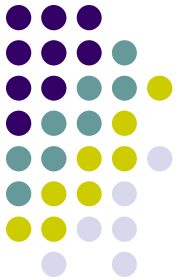
Features



Expert	Feature Category	Num. Features	Coverage (Percentage)
P expert	HMS_PCI Mass	1	8.3
	TAP Mass	1	8.8
	Yeast-2-Hybrid	1	3.9
F expert	GO Molecular Function	21	80.7
	GO Biological Process	33	76.1
	GO Component	23	81.5
	Essentiality	1	100
	MIPS Protein Class	25	4.6
	MIPS Mutant Phenotype	11	9.4
S expert	Gene Neighborhood / Gene Fusion / Gene Co-occur	1	100
	Sequence Similarity	1	100
	Homology based PPI	4	100
	Domain-Domain Interaction	1	100
E expert	Gene Expression	20	88.9
	Protein Expression	1	42.8
	Protein-DNA TF group binding	16	98.0
	Synthetic Lethal	1	7.6

Reference: Y. Qi, Z. Bar-Joseph, J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction", PROTEINS: Structure, Function, and Bioinformatics. 2005

Reference Set Situation



- Existing PPI Set:
  - Only a small positive (interacting) set available (small scale interaction experimental result)
  - No negative (not interacting) set available
  - Highly skewed class distribution
    - Much more non-interacting pairs than interacting pairs

- Reference set we use:
  - The ratio of positive pairs to negative (random) pairs is roughly 1 : 600 in yeast

	SET	#PAIRS	NOTE
Reference Set	Positive Set	~ 3000	From [DIP]
	Random Set		Random Generated (excluded above POS set)