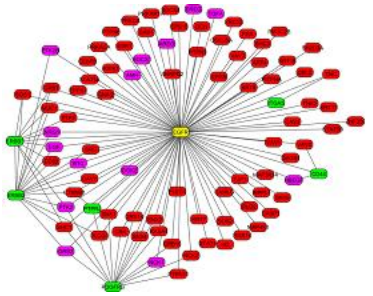
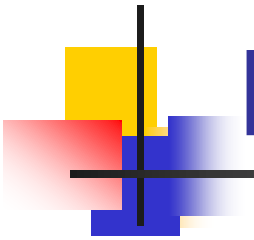
Learning of Protein Interaction Networks



Presenter: Yanjun Qi

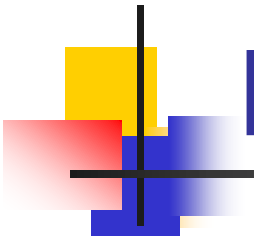
Ph.D. Thesis Defense
2008 / 05

Language Technologies Institute, School of Computer Science
Carnegie Mellon University



Road Map

- Protein-Protein Interaction (PPI) Network
- Learning of PPI Networks
  - Link prediction
  - Important group detection
- Summary
  - Thesis statement & contributions
  - Future work



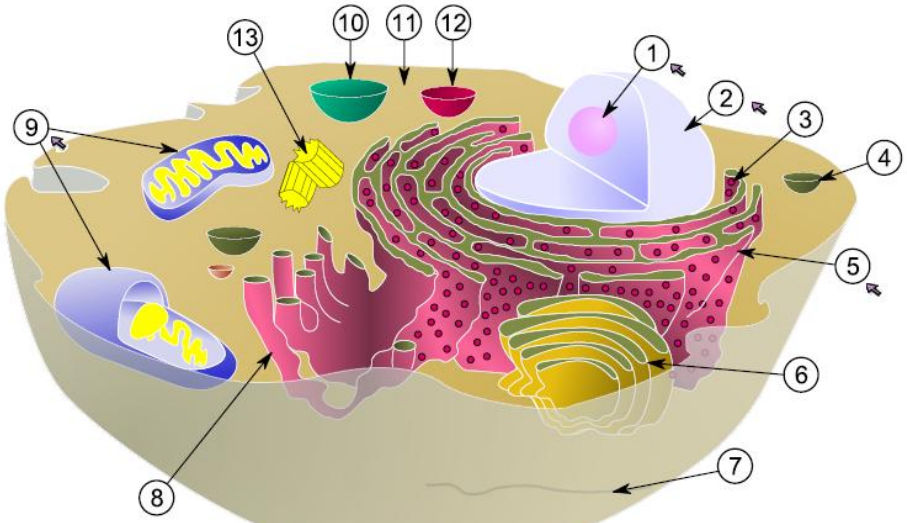
Road Map

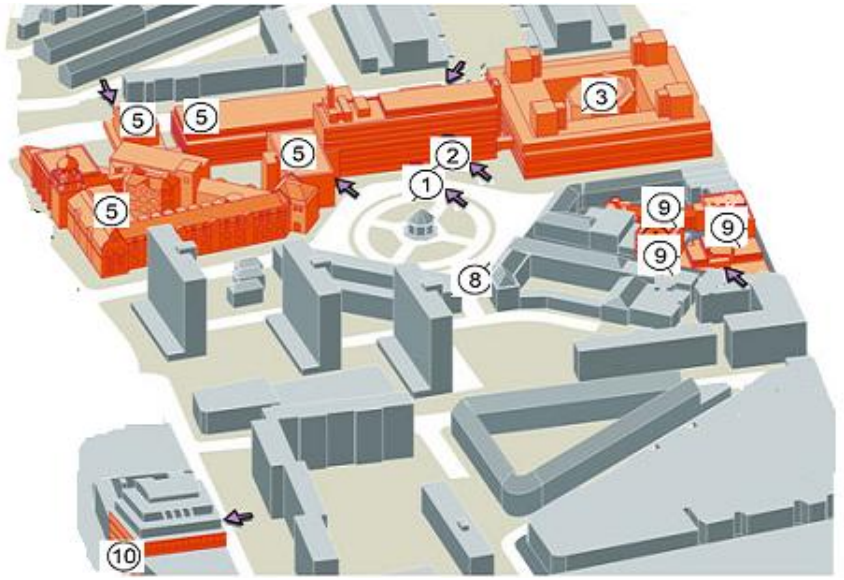
- Protein-Protein Interaction (PPI) Network
- Learning of PPI Networks
  - Link prediction
  - Important group detection
- Summary
  - Thesis statement & contributions
  - Future work

Background: Cell

- Cell
  - The basic living unit of life
- City
  - The basic unit of human society
- Protein
  - Chief actors within the cell
  - Participate in every biological process
- Human Being
  - Main actors within the city
  - Participate in every social activity

Cell Compartments

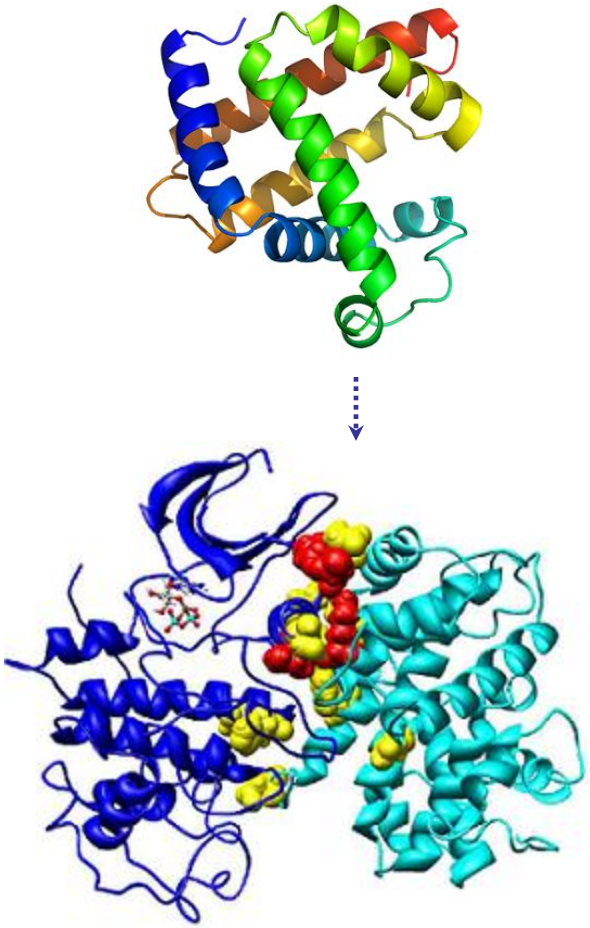




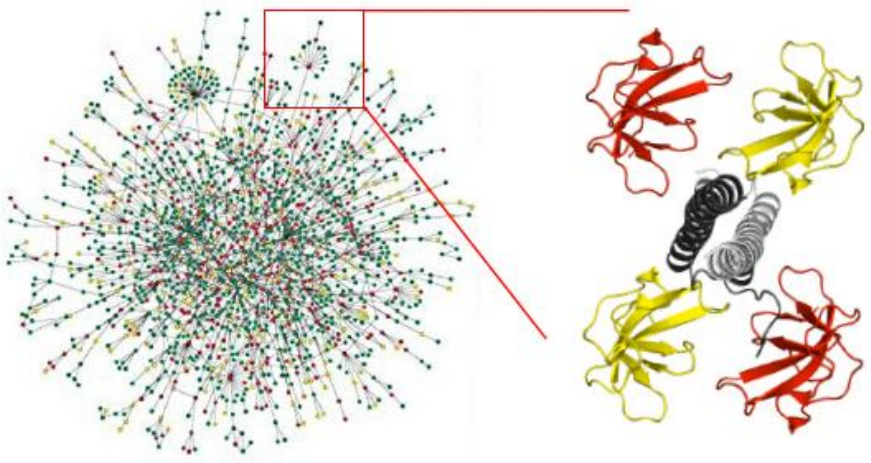
Parts	▪ Cell	▪ City
1. Center	Nucleolus	Chief executive
2. Information Center	Nucleus	City hall
5. Transport Network	ER	Subway
9. Power Generator	Mitochondria	Power plant
…	…	…

Proteins and Interactions

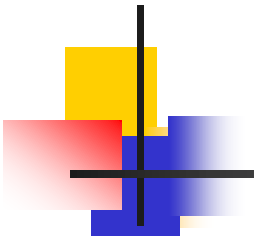
- Every function in the living cell depends on proteins
- Proteins are made of a linear sequence of amino acids and folded into unique 3D structures
- Proteins can bind to other proteins physically
  - Enables them to carry out diverse cellular functions



Protein-Protein Interaction (PPI) Network



- PPIs play key roles in many biological systems
- A complete PPI network (naturally a graph)
  - Critical for analyzing protein functions & understanding the cell
  - Essential for diseases studies & drug discoveries

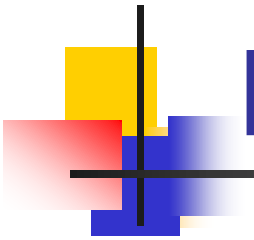


PPI Biological Experiments

- *Small-scale* PPI experiments
  - One protein or several proteins at a time
  - Small amount of available data
  - Expensive and slow lab process

- *Large-scale* PPI experiments
  - Hundreds / thousands of proteins at a time
  - Noisy and incomplete data
  - Little overlap among different sets

➔ Large portion of the PPIs still missing or noisy !

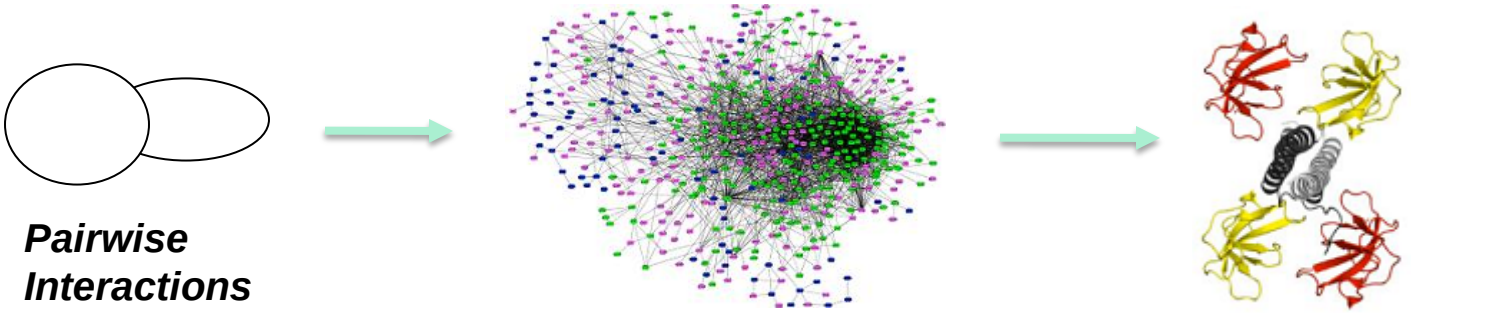


Road Map

- Protein-Protein Interaction (PPI) Network
- Learning of PPI Networks
  - Link prediction
  - Important group detection
- Summary
  - Thesis statement & Contributions
  - Future work

Learning of PPI Networks

- Goal I: Pairwise PPI (links of PPI graph)
  - Most protein-protein interactions (pairwise) have not been identified or noisy
  - ➔ Missing link prediction !
- Goal II: "Complex" (important groups)
  - Proteins often interact stably and perform functions together as one unit ("complex")
  - Most complexes have not be discovered
  - ➔ Important group detection !



Pairwise Interactions

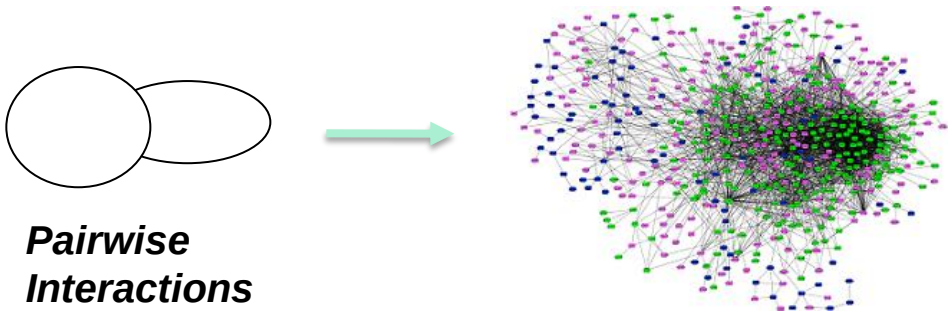
Link Prediction

PPI Network

Protein Complex

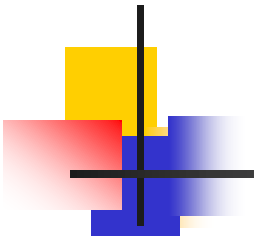
Group Detection

Goal I: Missing Link Prediction



Pairwise Interactions

PPI Network

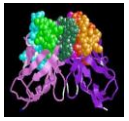
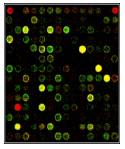


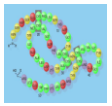
PPI Prediction through Data Fusion

- Motivation
  - Lots of other biological information available
  - Implicitly related to PPI relationship (for example, co-expressed genes)
  - Utilize this information to improve the quality of protein interaction data
- Objectives
  - To infer PPI reliably and to provide interesting biological hypotheses for validation
  - To provide useful information for the design of laboratory experiments

Related Biological Data

- Overall, four categories:
  - Direct high-throughput experimental data: Two-hybrid screens (Y2H) and mass spectrometry (MS)
  - 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.

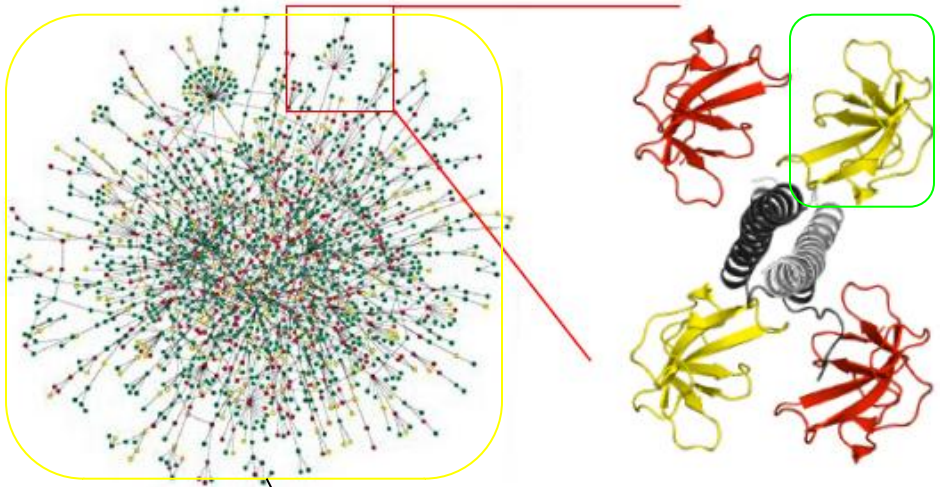



direct

Indirect

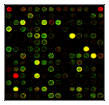
➔ Utilize implicit evidence and available direct experimental results together

Related Data Evidence

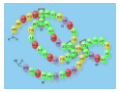


Attribute Evidence of Each Protein

Sequence



Expression

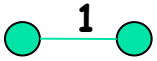


Structure

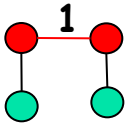
Annotation

……

Relational Evidence Between Proteins



Synthetic lethal

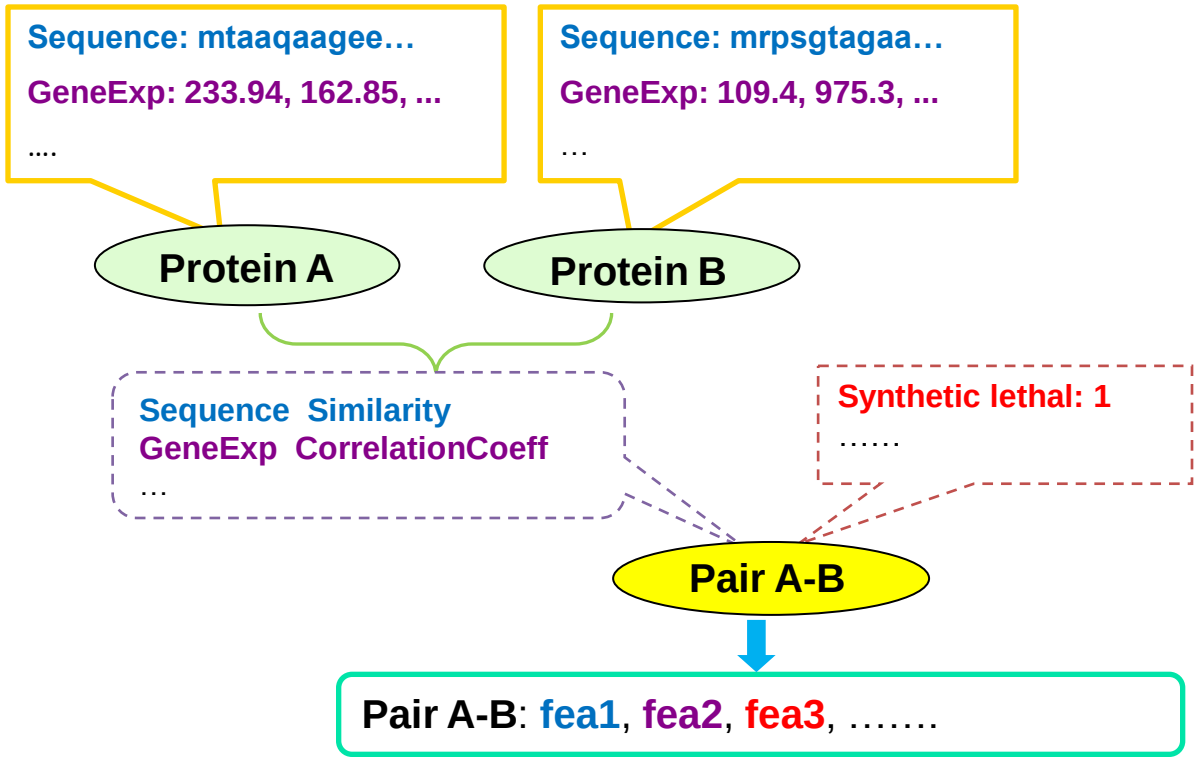


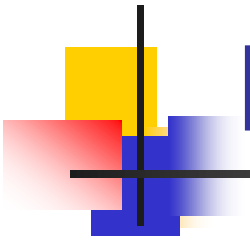
Relation expanding

……

Feature Vector for (Pairwise) Pairs

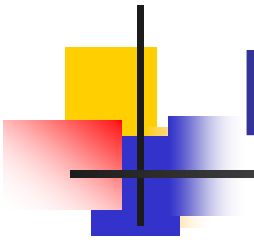
- For data representing protein-protein pairs, use directly
- For data representing single protein (gene), calculate the (biologically meaningful) similarity between two proteins for each evidence





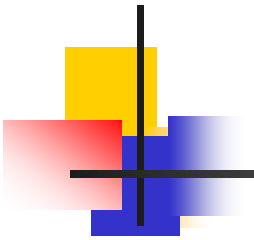
Problem Setting

- For each protein-protein pair:
  - Target function: interacts or not ?
  - Treat as a binary classification task
- Feature Set
  - Feature are heterogeneous
  - Most features are noisy
  - Most features have missing values
- Reference Set:
  - Small-scale PPI set as positive training (thousands)
  - No negative set (non-interacting pairs) available
  - Highly skewed class distribution
    - Much more non-interacting pairs than interacting pairs
    - Estimated: 1 out of ~600 yeast; 1 out of ~1000 human



Previous Work

- Jansen,R., et al., Science 2003
  - Bayes Classifier
- Lee,I., et al., Science 2004
  - Sum of Log-likelihood Ratio
- Zhang,L., et al., BMC Bioinformatics 2004
  - Decision Tree
- Bader J., et al., Nature Biotech 2004
  - Logistic Regression
- Ben-Hur,A. et al., ISMB 2005
  - Kernel Method
- Rhodes DR. et al., Nature Biotech 2005
  - Naïve Bayes



Systematic Comparison

- Previous methods differ in three aspects
  - Reference sets for training and testing;
  - Features and how they were extracted
  - Learning methods
- Thus, we collect a benchmark data set for supervised PPI prediction
  - To investigate how three aspects affect the prediction performance

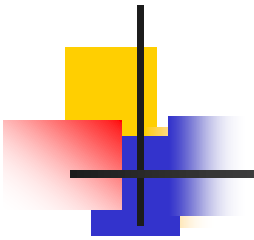
Y. Qi, Z. Bar-Joseph, J. Klein-Seetharaman, Proteins 2006

Systematic Comparison

- Key Factors
  - Prediction target (three types)
    - Not equally difficult (computationally)
    - (1) physical interaction, (2) co-complex relationship, (3) pathway co-membership task
  - Feature encoding
    - (1) "detailed" style, and (2) "summary" style
    - Feature importance varies
  - Classification method
    - Random Forest & Support Vector Machine

Details in the paper

Y. Qi, Z. Bar-Joseph, J. Klein-Seetharaman, Proteins 2006



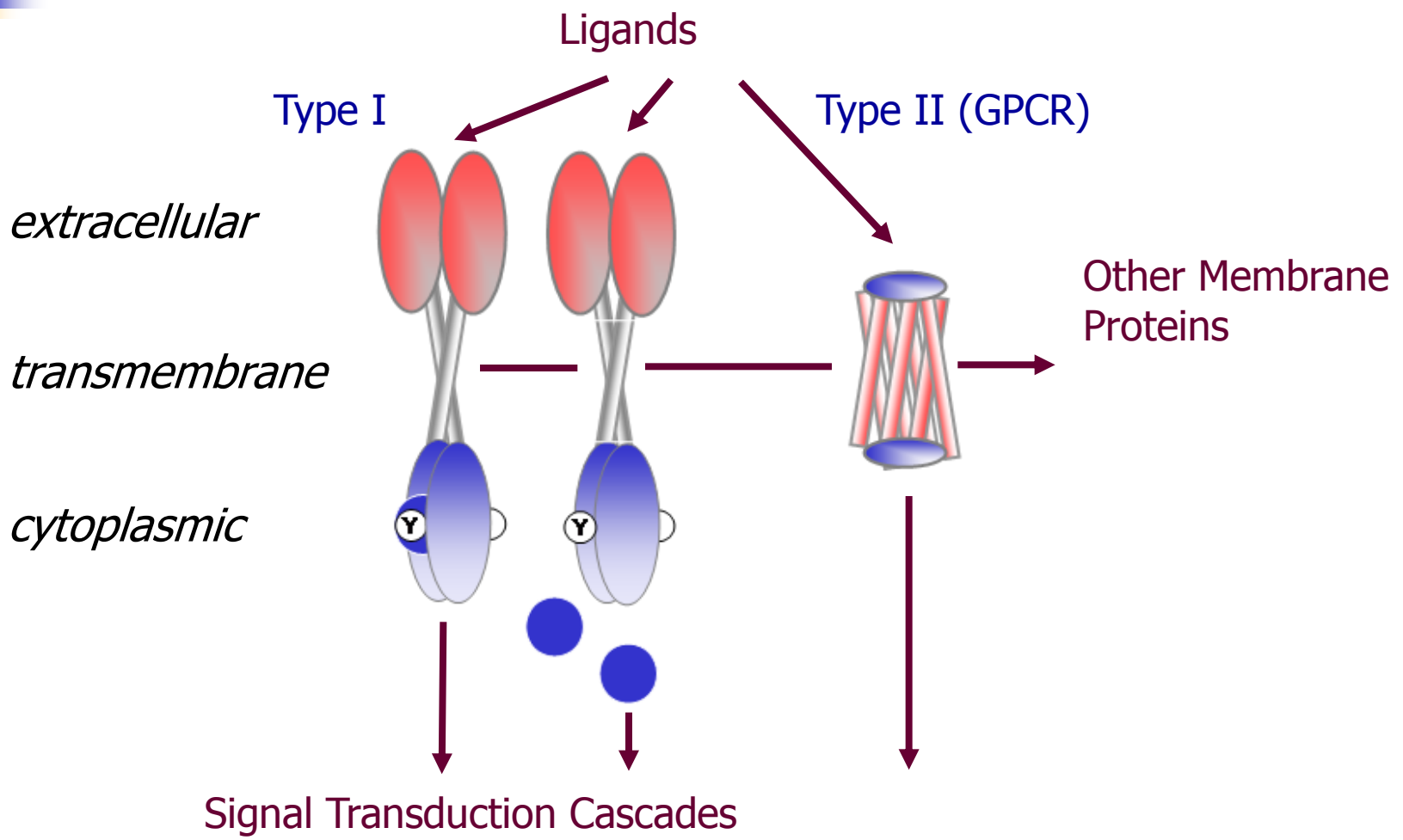
Methods Proposed

- Combined approach for sub-network PPI
  - Infer PPI reliably and validate experimentally
- PPI prediction using ranking
  - Find protein pairs that are "similar" to positive PPIs
- PPI prediction by multiple view learning
  - Infer PPI reliably and generate guidance info. to help biological experiments' design

Methods Proposed

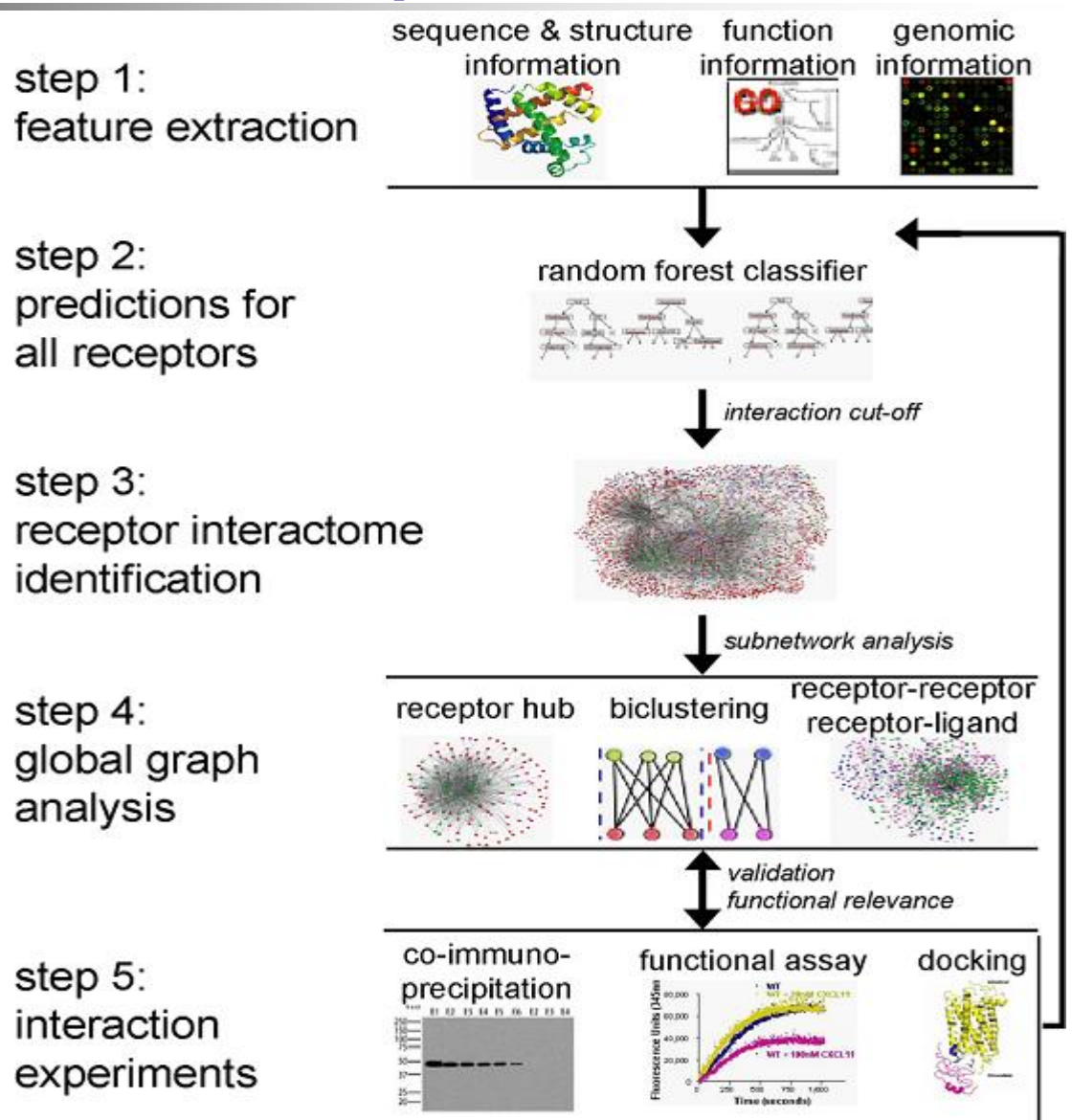
- Combined approach for sub-network PPI
  - Infer PPI reliably and validate experimentally
- PPI prediction using ranking
  - Find protein pairs that are "similar" to positive PPIs
- PPI prediction by multiple view learning
  - Infer PPI reliably and generate guidance info. to help biological experiments' design

Human Membrane Receptors



PPI Predictions for Human Membrane Receptors

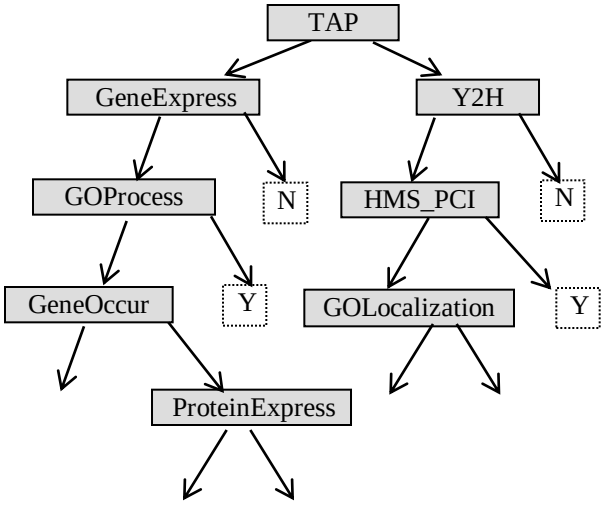
- A combined approach
  - Binary classification
  - Global graph analysis
  - Biological feedback & validation

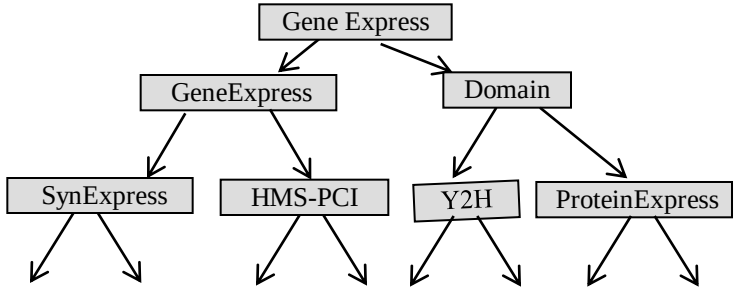


Y. Qi, et al 2008

Step 2: Binary Classification

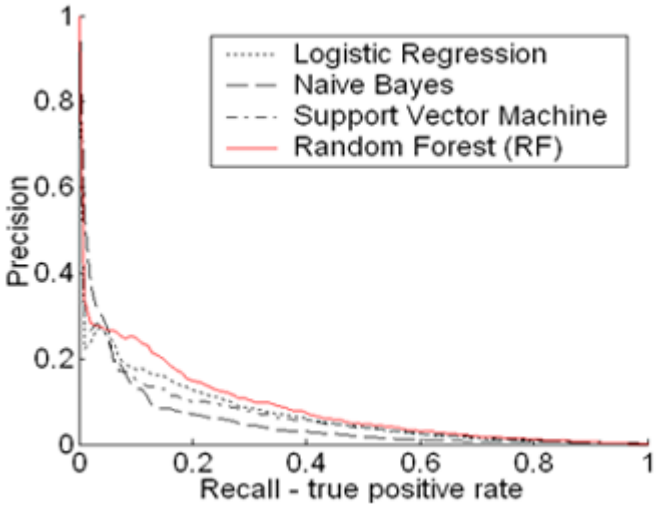
- Random Forest Classifier
  - A collection of independent decision trees (ensemble classifier)
  - Each tree is grown on a bootstrap sample of the training set
  - Within each tree's training, for each node, the split is chosen from a bootstrap sample of the attributes

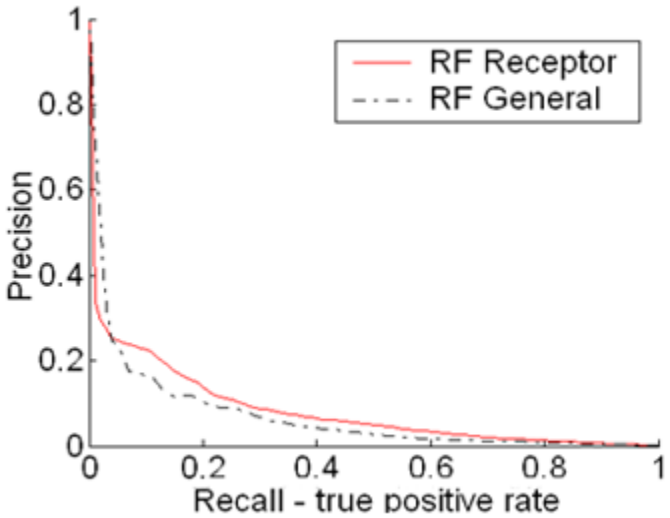




- Robust to noisy feature
- Can handle different types of features

Step 2: Binary Classification





- Compare Classifiers

(27 features extracted from 8 different data sources, modified with biological feedbacks)

- Receptor PPI (sub-network) to general human PPI prediction

Step 3-4: Global Graph Analysis

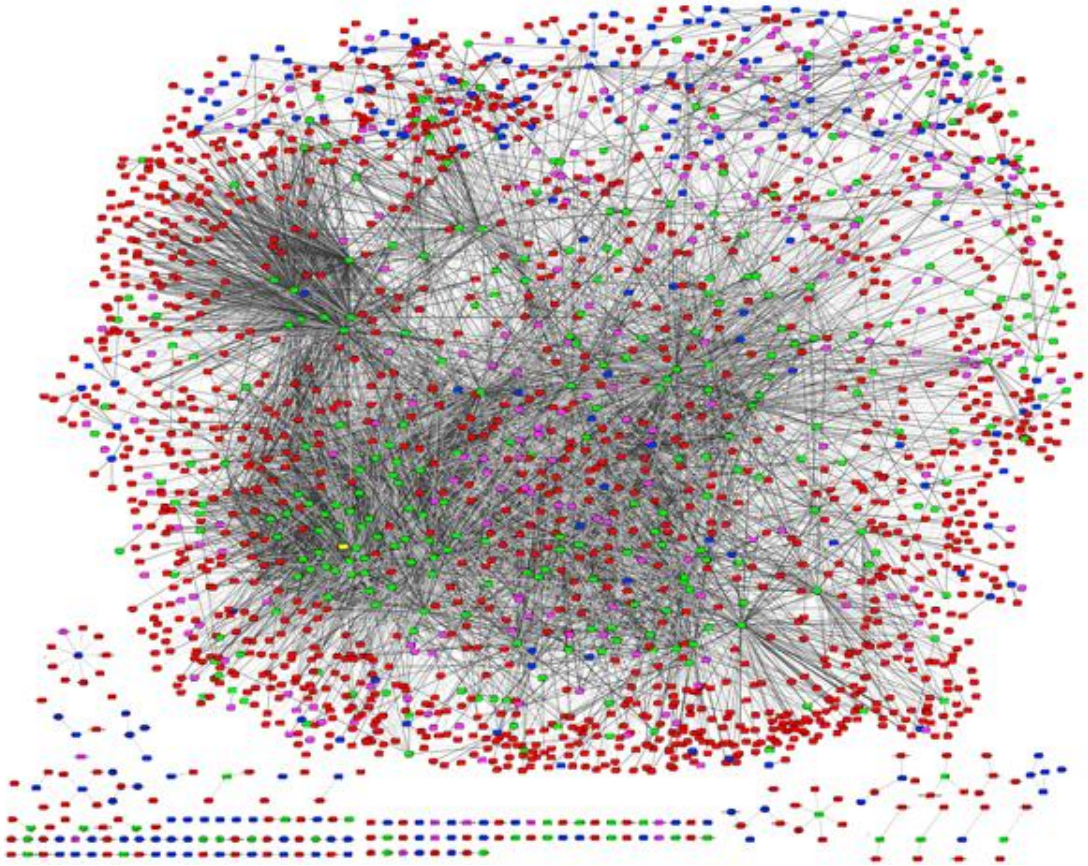
- Type I Receptor
- GPCR
- Ligand
- Other

Proteins

Receptor	551
Other	1752

Interactions

Total	9144
HPRD Known	1462



- Degree distribution / Hub analysis / Disease checking
- Graph modules analysis (from bi-clustering study)
- Protein-family based graph patterns (receptors / receptors subclasses / ligands / etc)

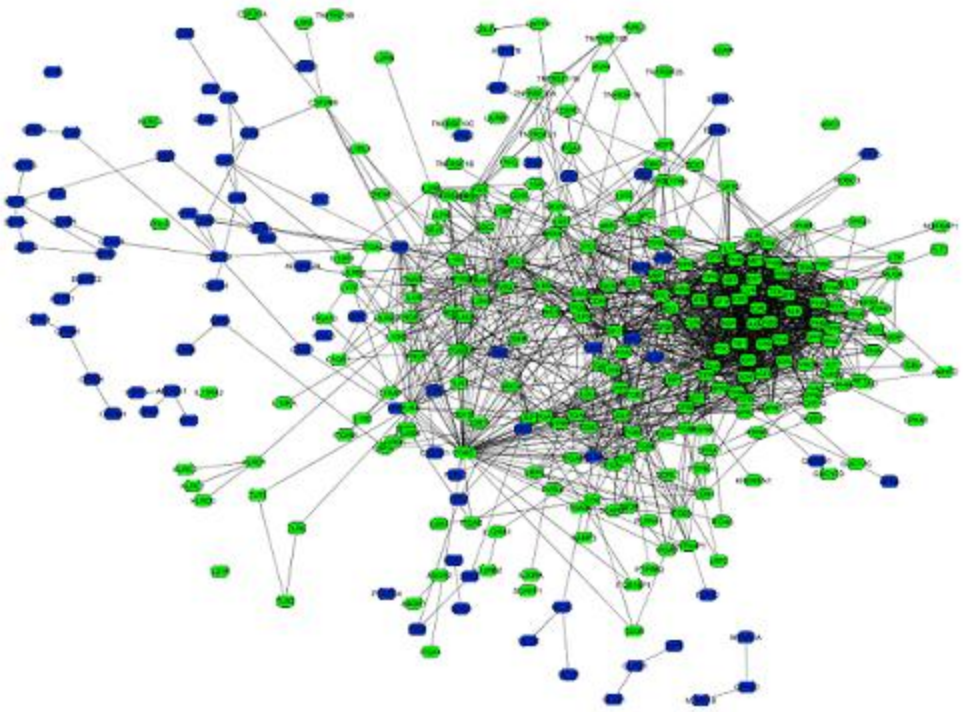
Step 4: Global Graph Analysis

- Network analysis reveals interesting features of the human membrane receptor PPI graph

For instance:

- Two types of receptors (GPCR and non-GPCR (Type I))

- GPCRs less densely connected than non-GPCRs (Green: non-GPCR receptors; blue: GPCR)

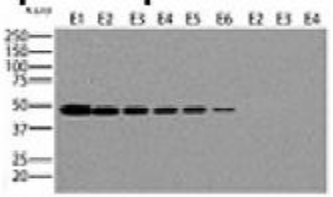


Step 5: Experimental Validation

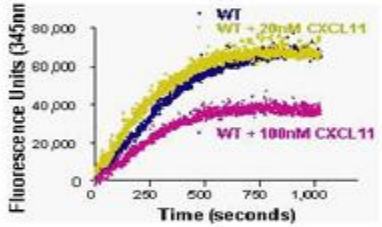
- Five of our predictions were chosen for experimentally tests and three were verified
  - EGFR with HCK (pull-down assay)
  - EGFR with Dynamin-2 (pull-down assay)
  - RHO with CXCL11 (functional assays, fluorescence spectroscopy, docking)

  - Experiments @ U.Pitt School of Medicine

co-immuno-precipitation



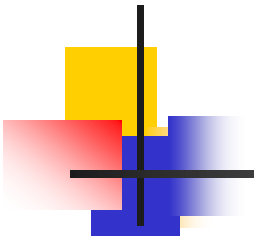
functional assay



docking

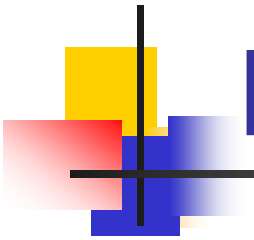
Details in the paper

Y. Qi, et al 2008



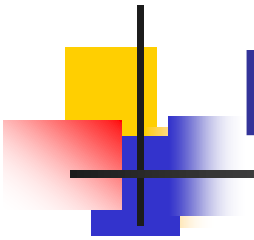
Methods Proposed

- Combined approach for sub-network PPI
  - Infer PPI reliably and validate experimentally
- PPI prediction using ranking
  - Find protein pairs that are "similar" to positive PPIs
- PPI prediction by multiple view learning
  - Infer PPI reliably and generate guidance info. to help biological experiments' design



Motivation

- Current situation of PPI task
  - Only a small positive (interacting) set available
  - No negative (not interacting) set available
  - Highly skewed class distribution
    - Much more non-interacting pairs than interacting pairs
  - The cost for misclassifying an interacting pair is higher than for a non-interacting pair
  - Accuracy measure is not appropriate here
- Try to handle this task with ranking
  - Rank the known positive pairs as high as possible
  - At the same time, have the ability to rank the unknown positive pairs as high as possible

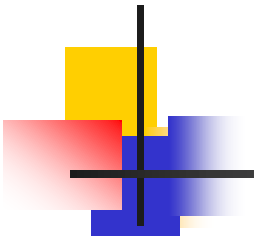


Method

- Handle this task using ranking
  - Find a distance / similarity function to measure the pairwise difference / similarity between protein pairs
  - Use kNN (or similar methods) to calculate the confidence score of a candidate pair based on the training set
  - Rank the test pairs to an ordered list by this score

Details in the paper

Y. Qi, J. Klein-Seetharaman, Z. Bar-Joseph, PSB 2005



Methods Proposed

- Combined approach for sub-network PPI
- PPI prediction using ranking
- PPI prediction by multiple view learning

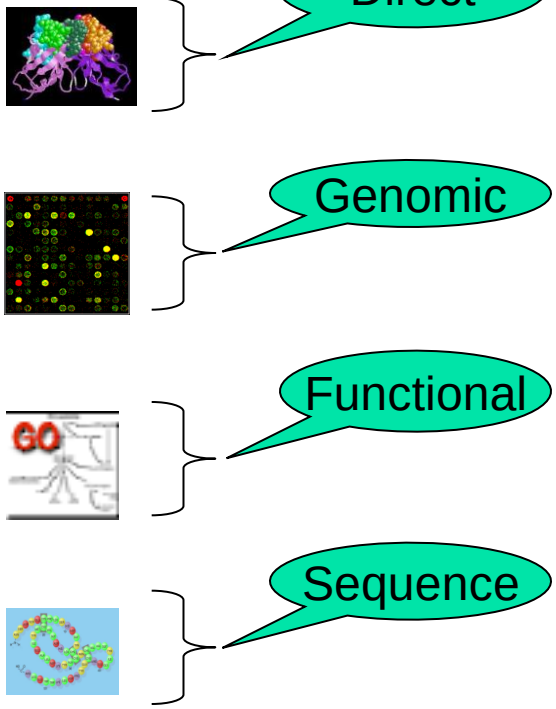
Motivation: Multiple View Learning

- Features are heterogeneous in nature
- Give guidance information for biological experimental design
  - Useful for biologists to know how features contributed to a specific prediction
  - Researchers may have various opinions regarding the liability of diverse features sources
  - Intrinsically different PPI pairs correlate differently with feature sources

Y. Qi, J. Klein-Seetharaman,Z. Bar-Joseph, BMC Bioinformatics 2007

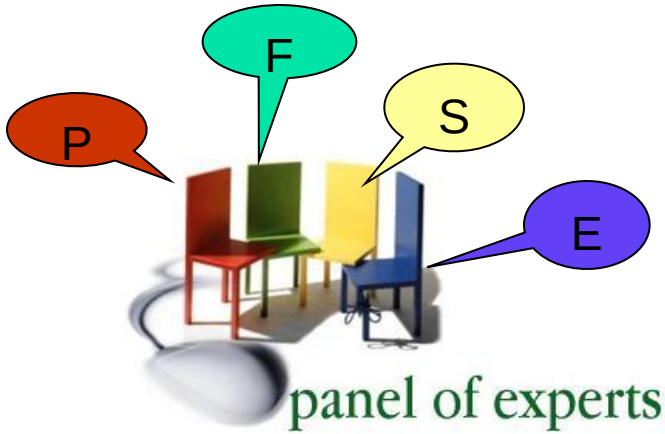
Split Features into Multi-View

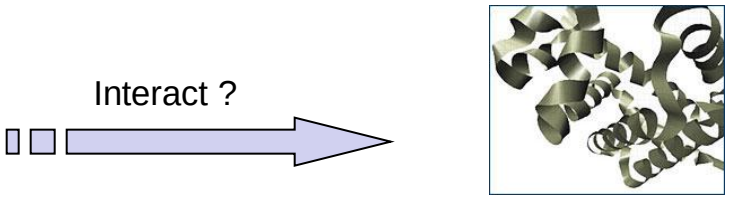
- Overall, four feature groups:
  - **P:** Direct highthroughput experimental data: Two-hybrid screens (Y2H) and mass spectrometry (MS)
  - **E:** Indirect high throughput data: Gene expression, protein-DNA binding, etc.
  - **F:** Functional annotation data: Gene ontology annotation, MIPS annotation, etc.
  - **S:** Sequence based data sources: Domain information, gene fusion, homology based PPIs, etc.



Y. Qi, J. Klein-Seetharaman, Z. Bar-Joseph, BMC Bioinformatics 2007

Mixture of Feature Experts (MFE)

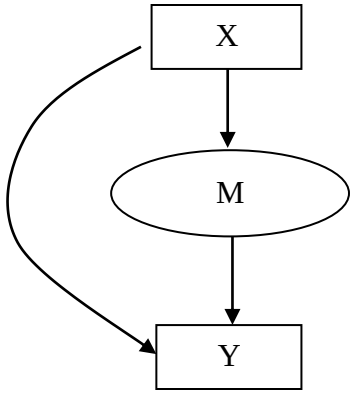




- Make protein interaction prediction by
  - Weighted voting from the four roughly homogeneous feature categories
  - Treat each feature group as a prediction expert
  - The weights are also dependent on the input example

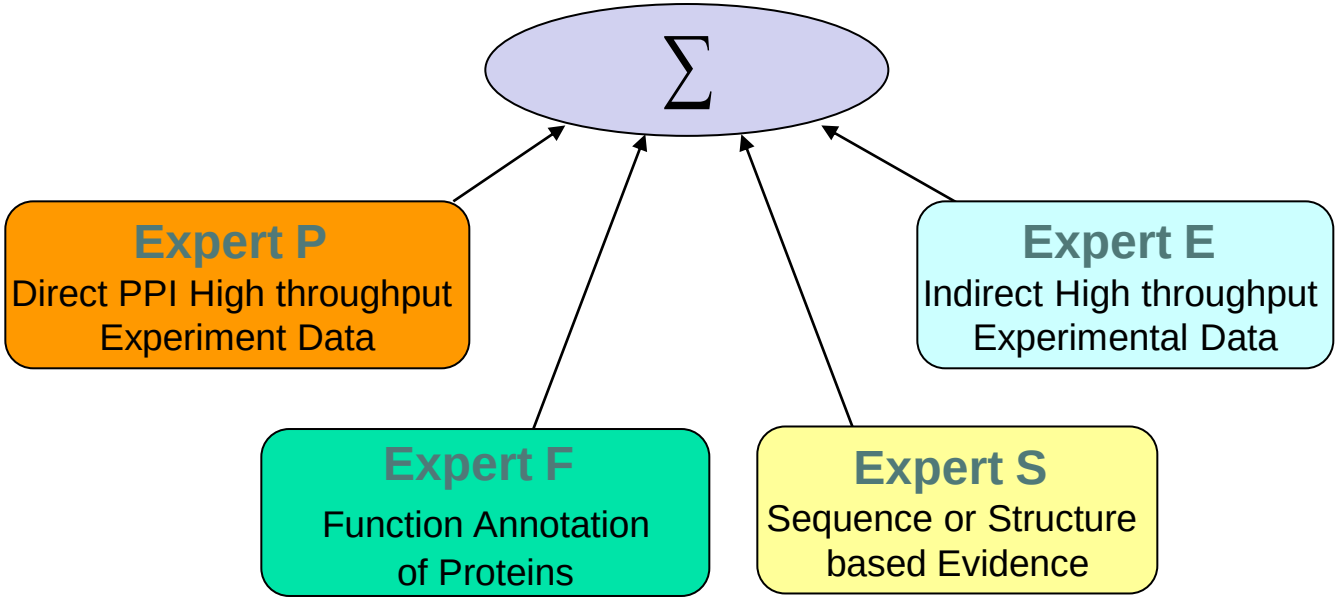
Mixture of Feature Experts (MFE)

- 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)

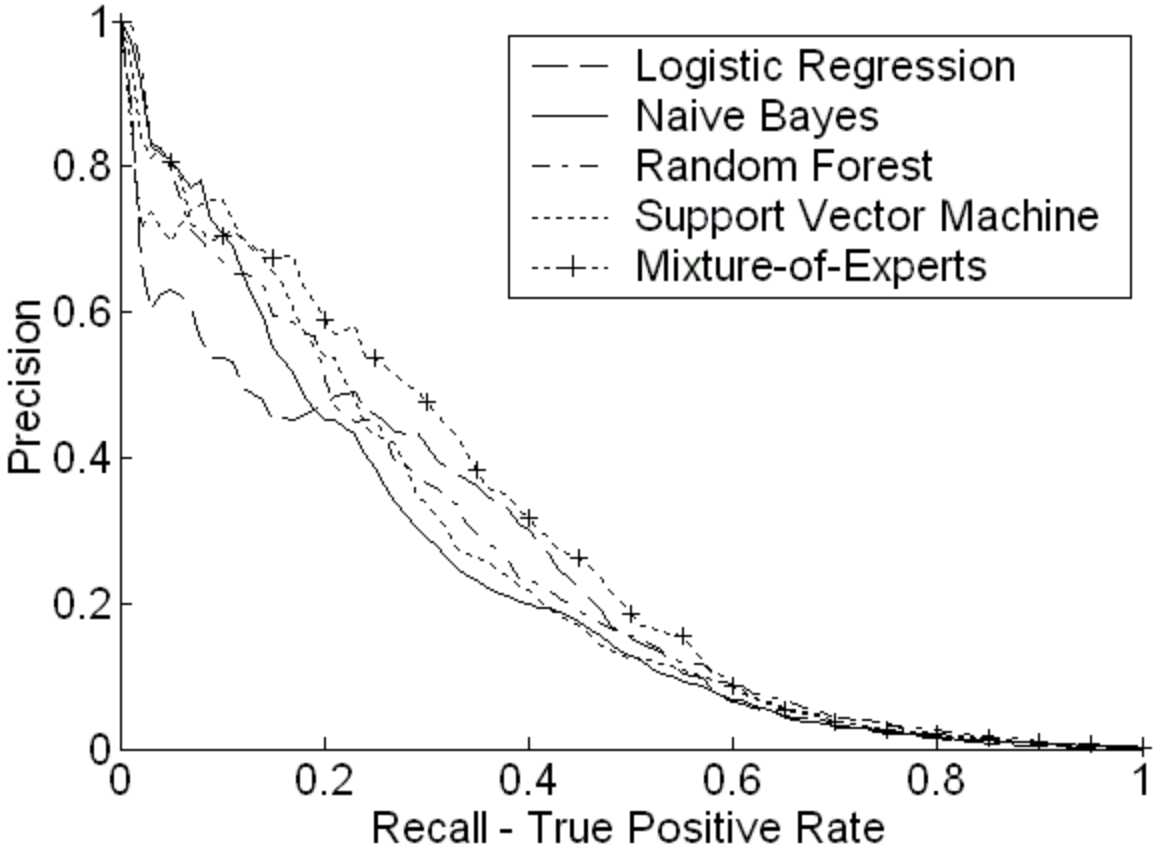
Mixture of Four Feature Experts

- Handling missing value
  - Add additional feature column for each feature having low feature coverage
  - MFE uses present / absent information when weighting different feature groups
- The posterior weight for expert *i* in predicting pair *n*
  - The weight can be used to indicate the importance of that feature view (expert) for this specific pair

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

Performance

- 162 features for yeast physical PPI prediction task
- Features extracted in "detail" encoding
- Under "detail" encoding, the ranking method is almost the same as RF (not shown)



A Simple Usage of Experts' Weights

- 300 candidate protein pairs
- 51 predicted interactions
  - 33 validated already
  - 18 newly predicted

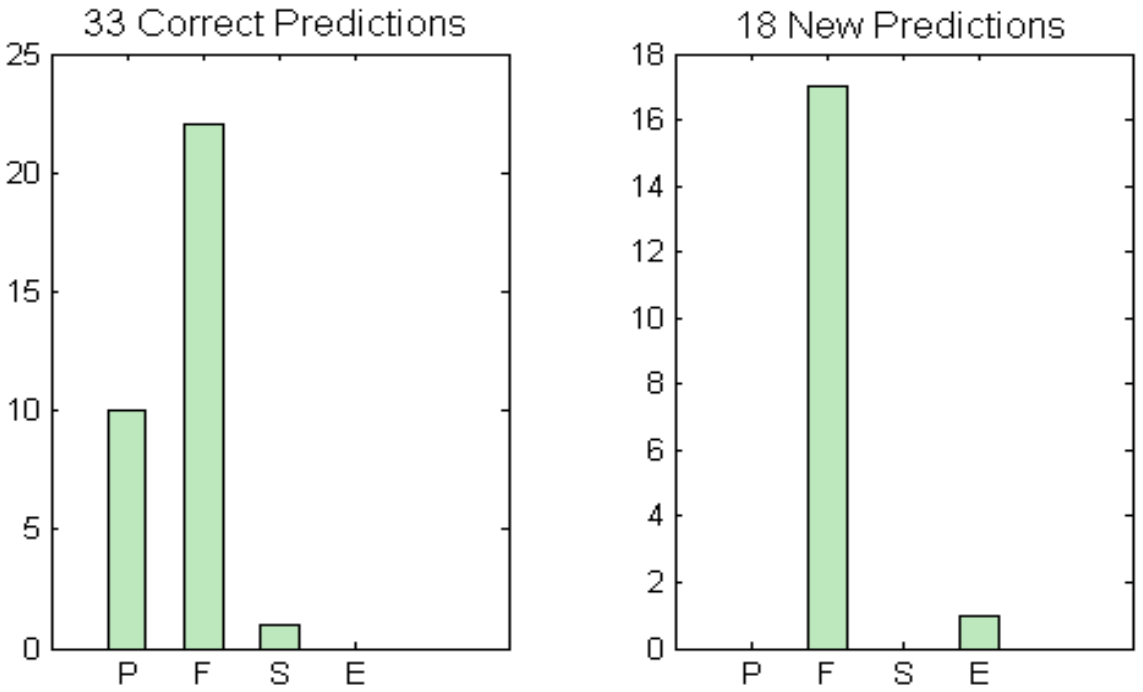
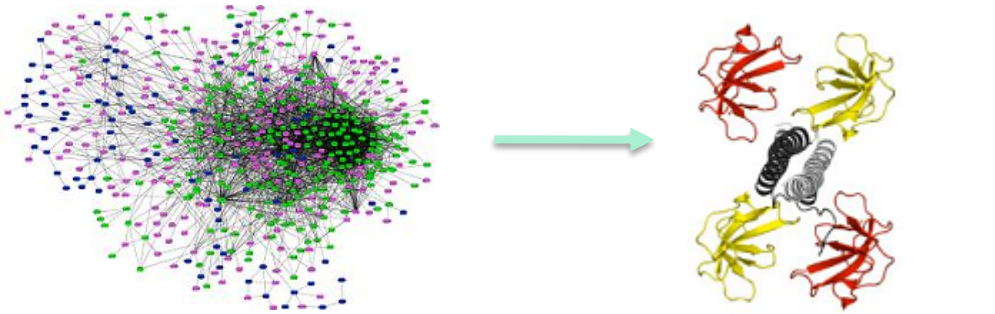


Figure: The frequency at which each of the four experts has maximum contribution among validated and predicted pairs

Goal II: Important Group Detection

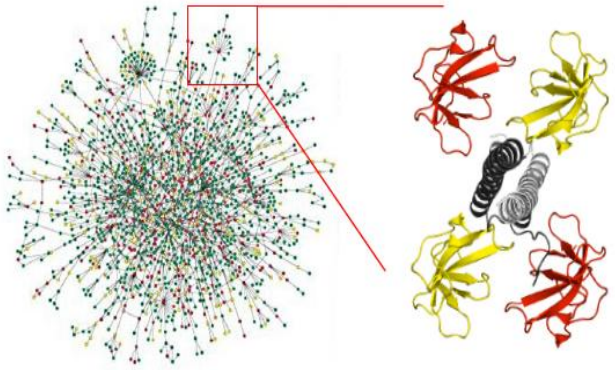


PPI Network ***Protein Complex***

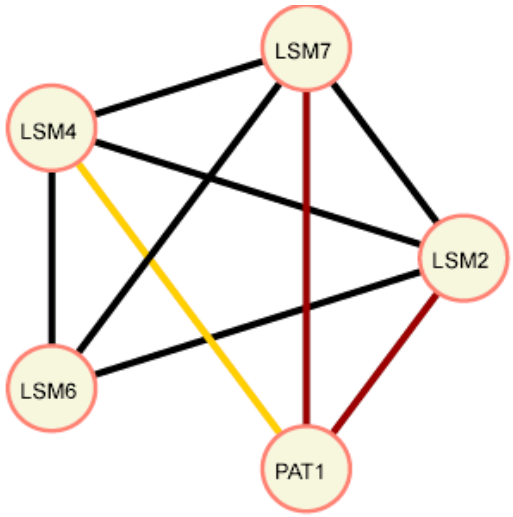
Protein Complex

➔ Group detection within the PPI network

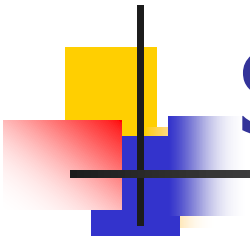
- Proteins form associations with multiple protein binding partners stably (termed “complex”)
- Complex member interacts with part of the group and work as an unit together
- Identification of these important sub-structures is essential to understand activities in the cell



Identify Complex in PPI Graph



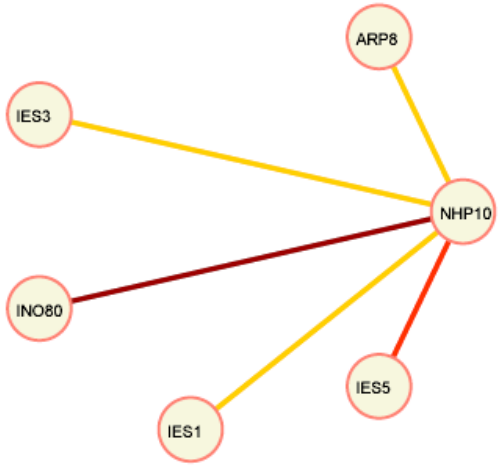
- PPI network as a weighted undirected graph
  - Edge weights derived from supervised PPI predictions: Goal I
- Previous work
  - Unsupervised graph clustering style
  - All rely on the assumption that complexes correspond to the dense regions of the network

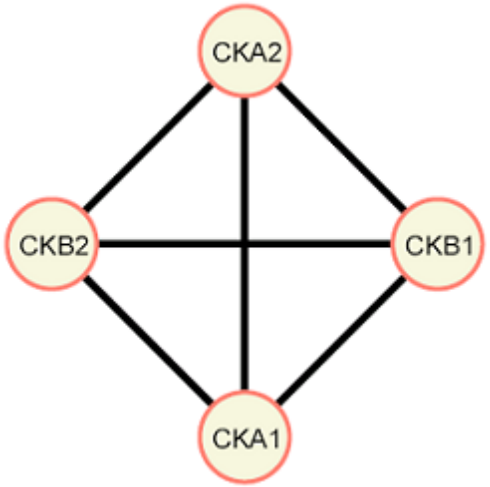


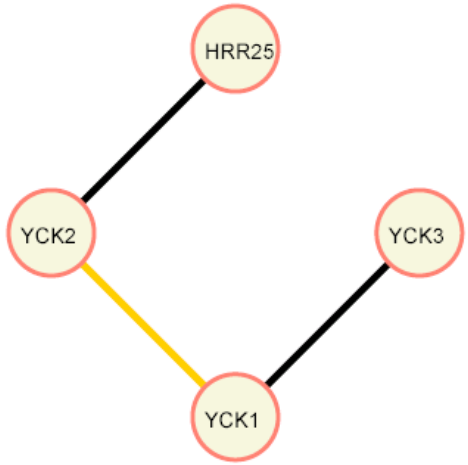
Some Facts

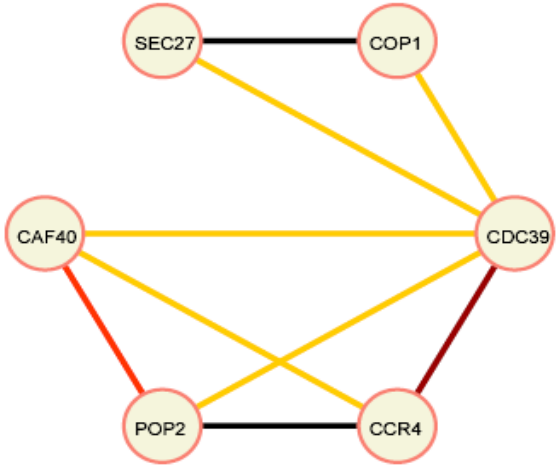
- Many other possible topological structures
- A small number of complexes available from reliable experiments
- Complexes also have functional /biological properties (like weight / size / …)

Possible topological structures

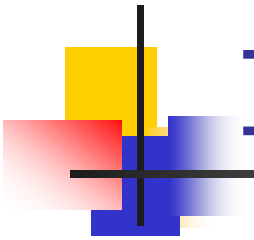








Edge weight color coded



Identify Complex in PPI

- Objectives
  - Make use of the small number of known complexes ➔ supervised
  - Model the possible topological structures ➔ subgraph statistics
  - Model the biological properties of complexes ➔ subgraph features

Properties of Subgraph

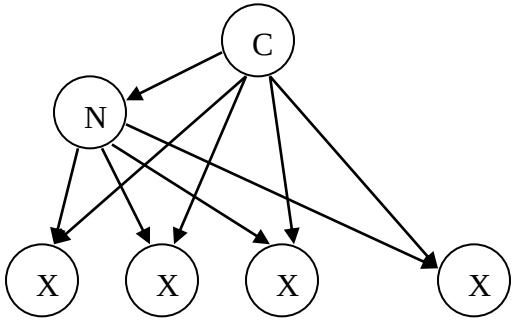
- Subgraph properties as features in BN
  - Various topological properties from graph
  - Biological attributes of complexes

No.	Sub-Graph Property
1	Vertex Size
2	Graph Density
3	Edge Weight Ave / Var
4	Node degree Ave / Max
5	Degree Correlation Ave / Max
6	Clustering Coefficient Ave / Max
7	Topological Coefficient Ave / Max
8	First Two Eigen Value
9	Fraction of Edge Weight > Certain Cutoff
10	Complex Member Protein Size Ave / Max
11	Complex Member Protein Weight Ave / Max

Model Complex Probabilistically

➔ Assume a probabilistic model (Bayesian Network) for representing complex sub-graphs

- Bayesian Network (BN)
  - *C* : If this subgraph is a complex (1) or not (0)
  - *N* : Number of nodes in subgraph
  - *Xi* : Properties of subgraph



$$L = \log \frac{p(c=1 \mid n, x_1, x_2, \ldots, x_m)}{p(c=0 \mid n, x_1, x_2, \ldots, x_m)}$$

Model Complex Probabilistically

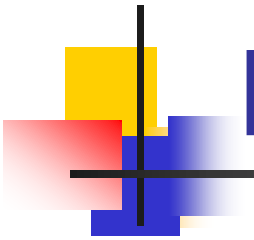
- BN parameters trained with MLE
  - Trained from known complexes and random sampled non-complexes
  - Discretize continuous features
  - Bayesian Prior to smooth the multinomial parameters

- Evaluate candidate subgraphs with the log ratio score *L*

$$L = \log \frac{p(c=1 \mid n, x_1, x_2, \ldots, x_m)}{p(c=0 \mid n, x_1, x_2, \ldots, x_m)} = \log \frac{p(c=1)\, p(n \mid c=1) \prod_{k=1}^{m} p(x_k \mid n, c=1)}{p(c=0)\, p(n \mid c=0) \prod_{k=1}^{m} p(x_k \mid n, c=0)}$$

Discover Complexes through Heuristic Local Search

- Identify Complexes ➔ Search for high scoring subgraphs
- Lemma: *Identifying the set of maximally scoring subgraphs in our PPI graph is NP-hard*
- Employ the iterated simulated annealing search on the log-ratio score

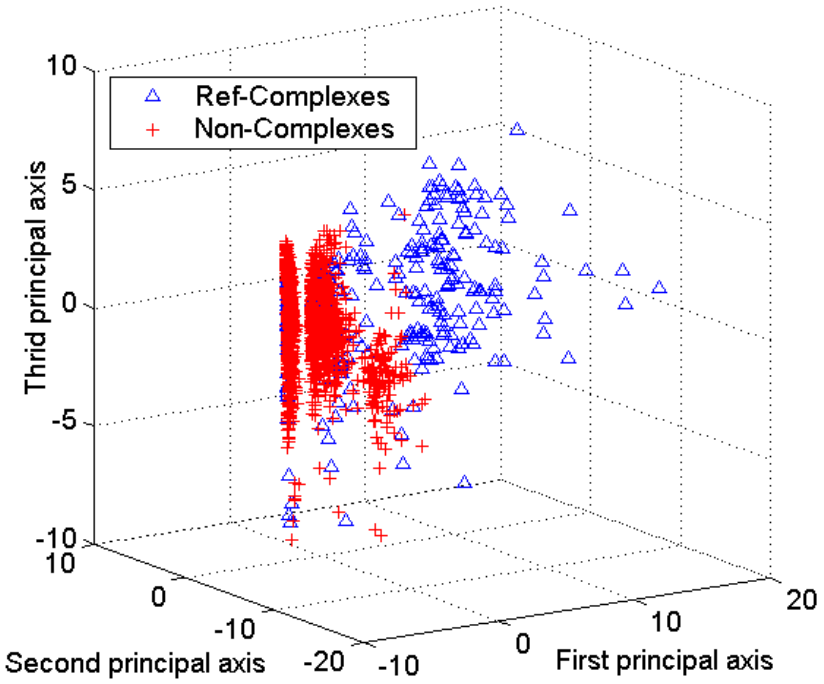


Experimental Setup

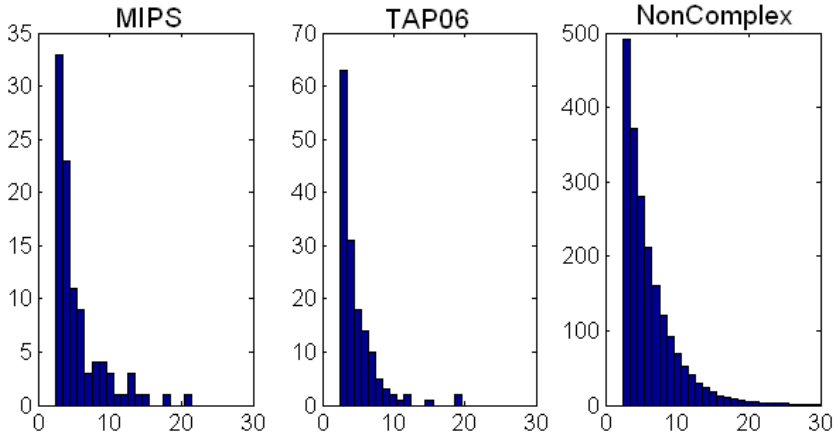
- Positive training data:
  - Set1: MIPS Yeast complex catalog: a curated set of ~100 protein complexes
  - Set2: TAP05 Yeast complex catalog: a reliable experimental set of ~130 complexes
  - Complex size (nodes' num.) follows a power law
- Negative training data
  - Generate from randomly selected nodes in the graph
  - Size distribution follows the same power law as the positive complexes

Data Distribution

Feature distribution

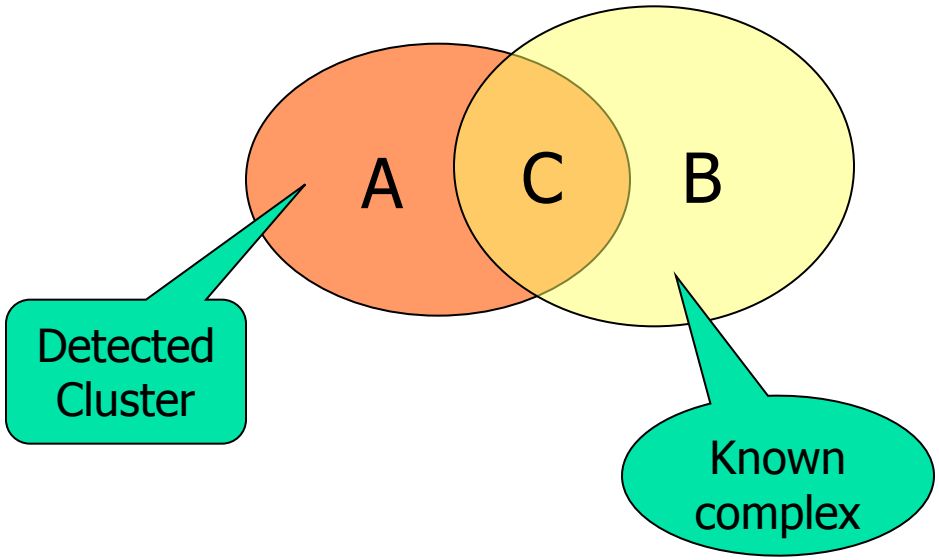


Node size distribution



Evaluation

- Train-Test style (Set1 & Set2)
- Precision / Recall / F1 measures
- A cluster "detects" a complex if



A : Number of proteins only in cluster
B : Number of proteins only in complex
C : Number of proteins shared

If overlapping threshold *p set as 50%*

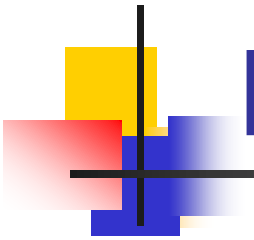
$$\frac{C}{A+C} > p \quad \& \quad \frac{C}{B+C} > p$$

Performance Comparison

- On yeast predicted PPI graph (~2000 nodes)

- Compare to a popular complex detection package: MCODE (search for highly interconnected regions)
- Compare to local search relying on density evidence only
- Compared to local search with complex score from SVM (also supervised)

Methods	Precision	Recall	F1
Density	0.180	0.462	0.253
MCODE	0.219	0.075	0.111
SVM	0.211	0.377	0.269
BN	0.266	0.513	0.346



Road Map

- Protein-Protein Interaction (PPI) Network
- Learning of PPI Networks
  - Link prediction
  - Important group detection
- Summary
  - Thesis statement & contributions
  - Future work

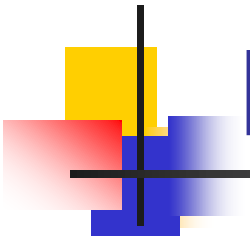
Thesis Statement

This dissertation provides a systematic computational framework for discovering protein-protein interactions (PPI) and for identifying important patterns within PPI networks.

The computational predictions yielded by this framework suggest a number of novel biological hypotheses that have been verified with subsequent laboratory experimentations.

Contributions

1. A systematic study and a benchmark dataset for supervised PPI prediction in yeast
2. Infer PPI reliably and validate experimentally ➔ A combined computational and experimental method for human receptor PPI predictions
3. Find protein pairs that are "similar" to positive PPIs ➔ PPI prediction with ranking for yeast PPI identifications
4. Infer PPI reliably and generate guidance info. to design biological experiments ➔ Mixture of feature experts method for PPI identifications in yeast and human
5. Supervised group detection for protein complexes
6. Two web services (one for yeast PPI predictions and one for human receptor PPI predictions)

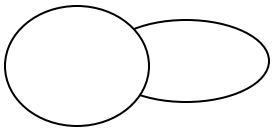


Future Work

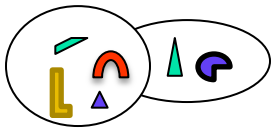
- Link prediction
  - Active learning to assist biological experiments
  - Semi-supervised learning for hard cases
  - Joint learning considering multiple links
  - Virus to host PPI predictions (bipartite graph)
- Group detection
  - better complex model
  - better search algorithm
- Pathway identification (chain structure)
- Global graph analysis of PPI network
- Protein function prediction (hierarchy labels)
- Domain/motif interaction detection (binding sites)

Learning of PPI Networks

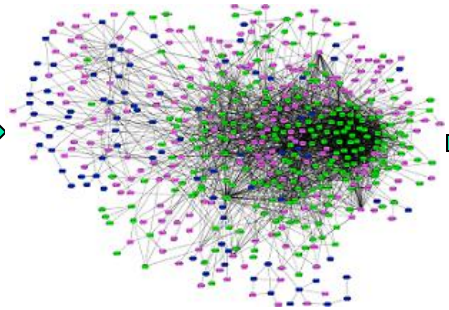
PSB 05
PROTEINS 06
BMC Bioinfo 07
CCR 08



Pairwise Interactions



Domain/Motif Interactions



PPI Network

Human-PPI (Revise 08)
HIV-Human PPI (Revise)

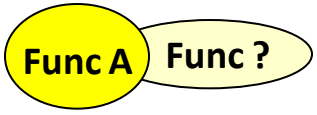
Protein Complex

ISMB 08

Pathwav

Prepare

Function Implication



Genome Biology 08

Thanks !

Questions ?