Predicting the functions of proteins in PPI networks from global information

Hossein Rahmani

Joint work with: Hendrik Blockeel, Andreas Bender

5 September 2009

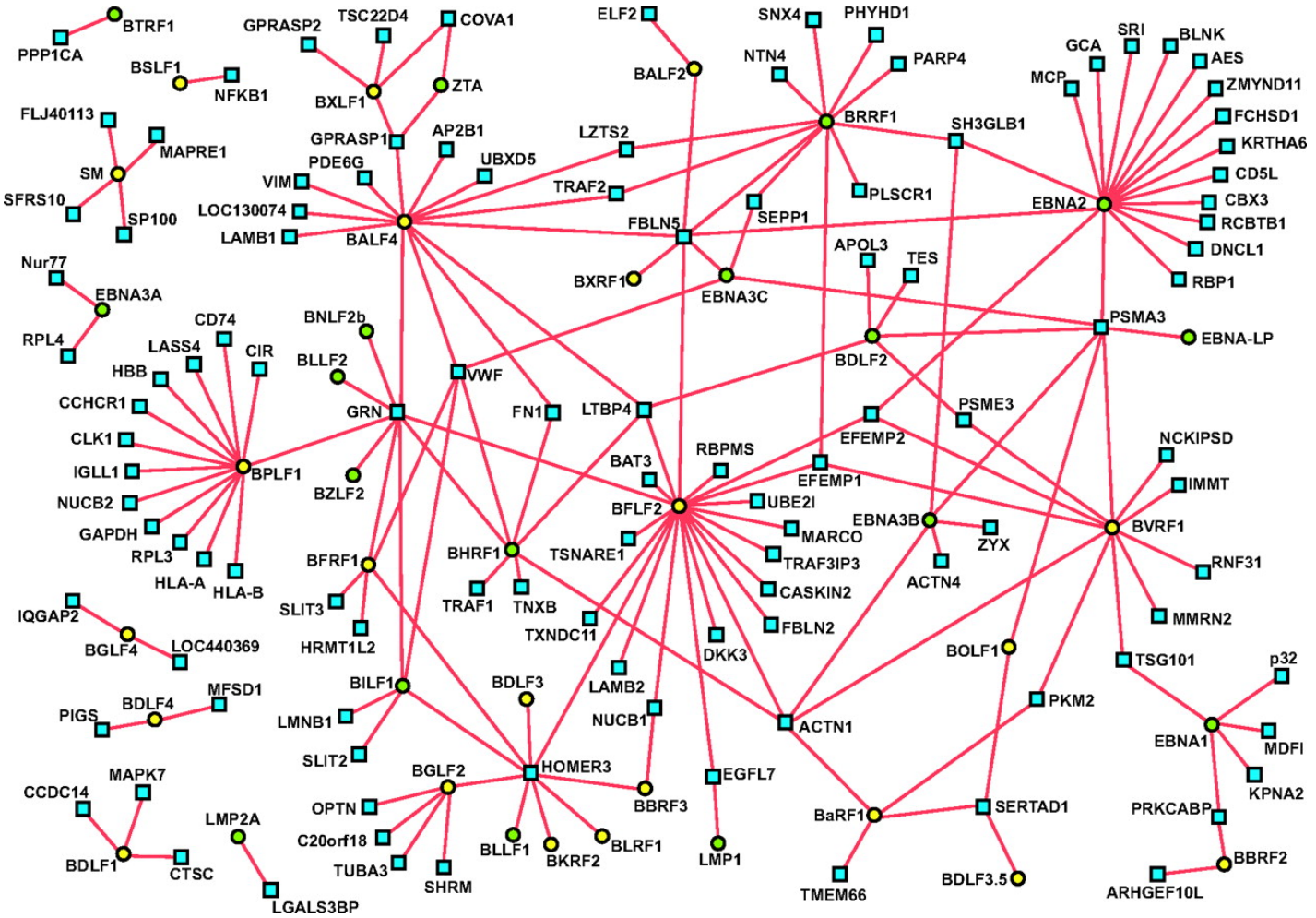
Outline

- Problem Statement
  - Formal Description of PPI Networks
  - Transductive Approaches
  - Inductive Approaches
- Related Works
- Proposed Methods
  - Global Inductive Method
- Experiment
  - Comparison of Learners
  - Comparison with a Transductive Method
- Conclusions and Future Works

PPI Network

- Protein-Protein Interaction Network
- Represented as an undirected graph
  - Node set $V \rightarrow$ Proteins
  - Edge set $E \rightarrow$ Direct interaction
- $\forall v \in V$ is described by a description $d(v) \in D$
  - $d(v)$ derived from the network structure
  - No additional information, such as the protein structure is available
- $\forall v \in V$ optionally is annotated with a label $l(v) \in L$
  - Labels $l(v)$ are sets of protein functions
  - E.g., metabolism, transcription, protein synthesis and etc
- We assume there is a true labeling function λ that is $l(v) = \lambda(v)$ where $l(v)$ is defined
- Task: Find a suitable $\lambda(v)$ where $l(v)$ is not defined

PPI Network Sample



Function Categories

Functional Category	Proteins
01 METABOLISM	1514
02 ENERGY	367
10 CELL CYCLE AND DNA PROCESSING	1012
11 TRANSCRIPTION	1077
12 PROTEIN SYNTHESIS	480
14 PROTEIN FATE (folding, modification, destination)	1154
16 PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)	1049
18 REGULATION OF METABOLISM AND PROTEIN FUNCTION	253
20 CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES	1038
30 CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM	234
32 CELL RESCUE, DEFENSE AND VIRULENCE	554
34 INTERACTION WITH THE ENVIRONMENT	463
38 TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS	120
40 CELL FATE	273
41 DEVELOPMENT (Systemic)	69
42 BIOGENESIS OF CELLULAR COMPONENTS	862
43 CELL TYPE DIFFERENTIATION	452
99 UNCLASSIFIED PROTEINS	1393

Multi-Label Proteins

- Many proteins have more than one function
- Problem:
  - If we have n possible functions
  - Goal: Predicting a subset of these functions for unclassified proteins
- Most of the off-the-shelf machine learning techniques can predict a single value, not a set of values
- Transform task of predicting set of functions to
  - n single-function prediction tasks
  - Use binary classification for each possible function
    - Class True: Protein has that function
    - Class False: Protein does not have that function

Transductive Learning

- Task: Predict the label of all the nodes
- Input: $G = (V, E, d, l)$ with l a partial function
- Output: Complete version $G' = (V, E, d, l')$ with l' a complete function that is consistent with l
- l' should approximates λ by optimization criterion o
- o expresses our assumption about λ
  - E.g., directly connected nodes tend to have similar labels
  - Number of $\{v_1, v_2\}$ edges where $l'(v_1) \neq l'(v_2)$ edges should be minimal
- Our assumption about λ is called bias of transductive learner

Inductive Learning

- Task: Learn a function $f : D \rightarrow L$ that maps a node description $d(v)$ onto its label $l(v)$
- Input: $G = (V, E, d, l)$ with l a partial function
- Output: $f : D \rightarrow L$ such that $f(d(v)) = l(v)$
- Note: f differs from l in that it maps D, not V, onto L
  - It can make prediction for node v that is not in the original network, as long as $d(v)$ is known
- Biases
  - Transductive bias: Assumption expressed by optimization criterion o
  - Description bias D
  - Inductive bias: Choice of learning algorithm that is used to learn f from $(d(v), l(v))$ couples

Related Works

1. Transductive approaches
   - Local: Majority Rule and its extensions
   - Global: Global Optimization and Functional Clustering
2. Inductive approaches
   - Local: Topological Redundancies
   - Global: ? $\rightarrow$ Our Method

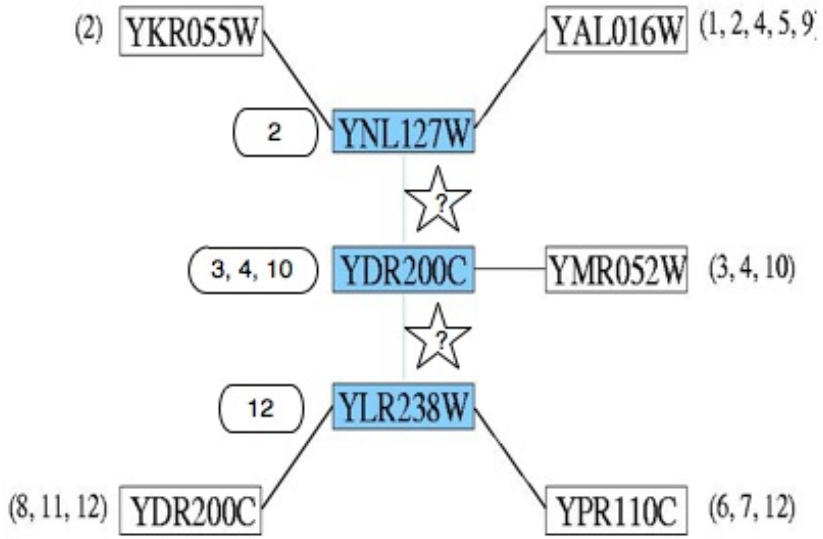
Schwikowski, B.(2000)

Milenkovic, T.(2008)

Majority Rule

- Local transductive method
- Assumption: Two Interacting proteins have something in common (e.g., same function)
- Predicted function: Most common function(s) among classified partners

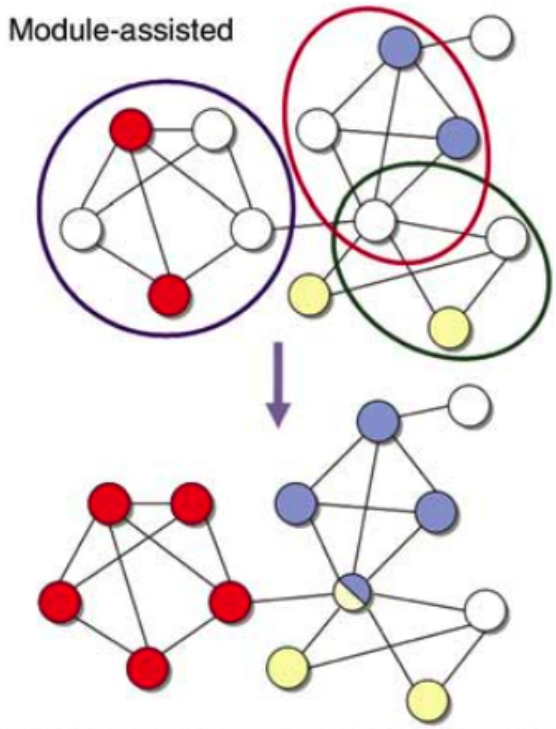
Majority Rule



- Problem: Links unclassified-unclassified proteins completely neglected
- Solution: Global optimization methods

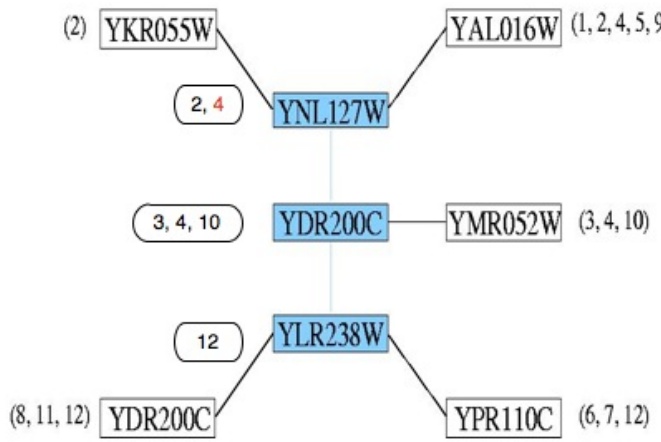
Functional Clustering

- Global transductive method
- Assumption: Dense regions are a sign of the common involvement in biological process
- Predict the function of unclassified protein based on the cluster they belong to



Global Optimization

- Global transductive method
- Links unclassified/unclassified proteins also taken into account
- Any probable function assignment to the whole set of unclassified proteins is considered
  - Counting number of interacting pairs with no common functions
  - Select the function assignment with lowest value

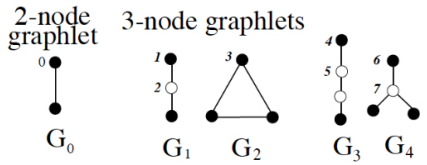


Topological Approaches

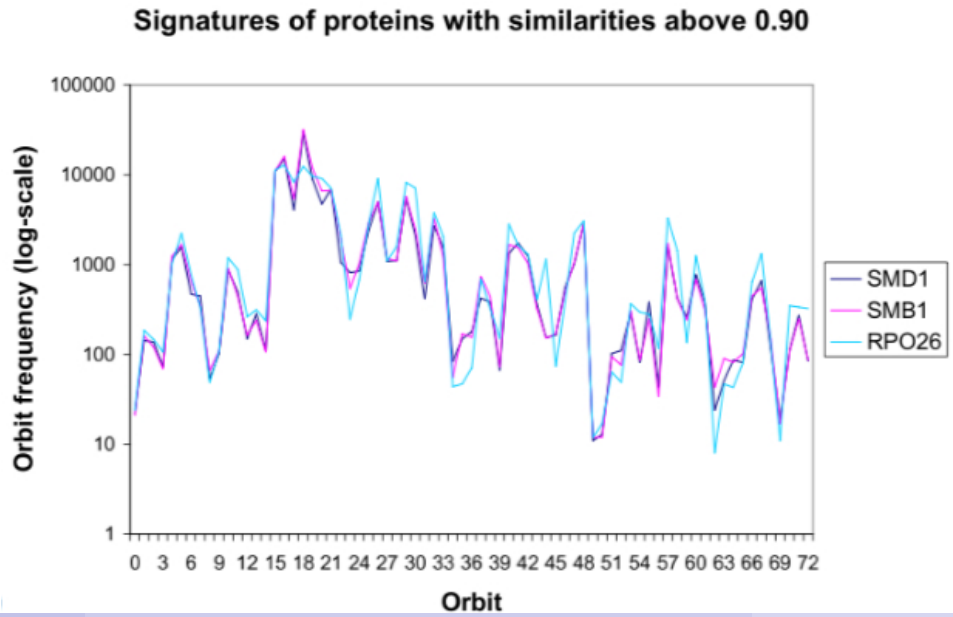
- Local inductive method
- Node description $d(v)$ is built based on the local neighborhood
- Count number of patterns (e.g., graphlet) around the proteins
- Make the signature vector for each protein
- Assumption: Proteins with high similar signature vector have same functions

Topological Approaches

- Some topological patterns (Number of considered patterns = 73)



- Orbit: One of the previous patterns
- Same orbit frequency $\longrightarrow$ same function



A Global Description of Proteins

- Global inductive approach
- Node description
  - N nodes in the network numbered from 1 to N
  - Each node is described by an n-dimensional vector
  - i'th component in the vector of node v gives the length of shortest path between v and node i
  - Probelm: Large Graph $\rightarrow$ very high dimensional descriptions
  - Solution: Reduce dimensionality by focusing on shortest-path distance to a few "important" nodes
    - Feature selection problem

Important Proteins

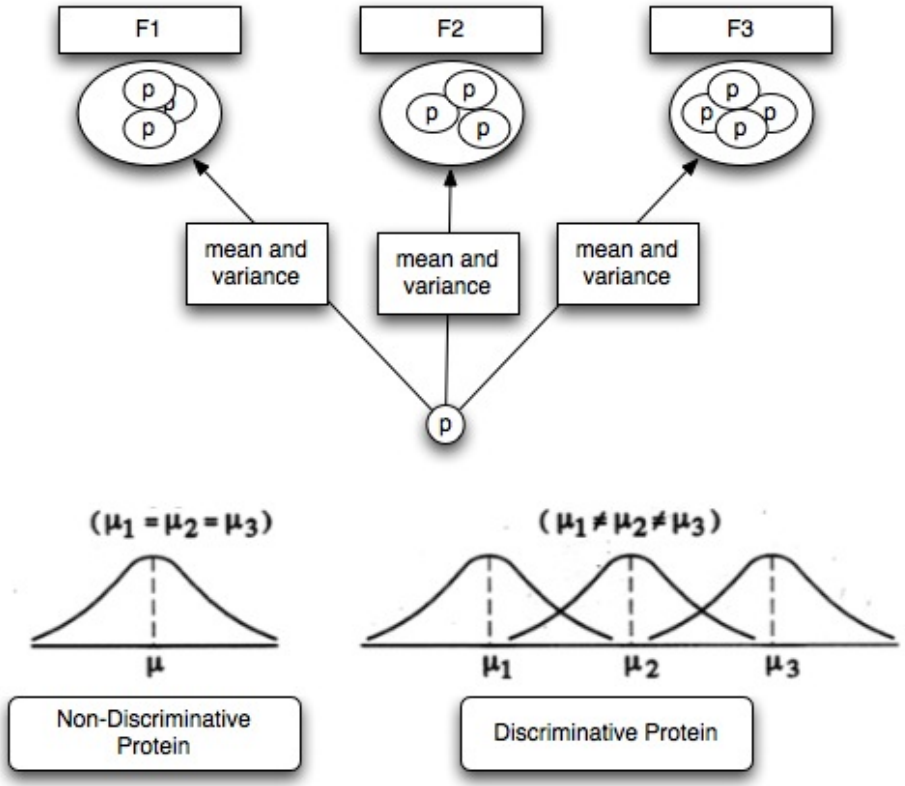
- Definition: Node i is important if the shortest-path distance of some node v to node i is likely to be relevant for v's classification
- Feature f_i denotes the shortest-path distance to node i
- Anova based feature selection
  - For each function j, let G_j be the set of all proteins that have that function j
  - Let $\bar{f}_{ij}$ be the average f_i value in G_j
  - Let $var(f_{ij})$ the variance of the f_i in G_j
  - Anova (analysis of variance) based relevancy measure:

$$A_i = \frac{Var_j[\bar{f}_{ij}]}{Mean_j[var(f_{ij})]} \quad (1)$$

  - A high A_i denotes a high relevance of feature f_i

Important Proteins

- Simple Example:
  - Three functions: $F1$, $F2$ and $F3$
  - Proteins: P's



Experiments

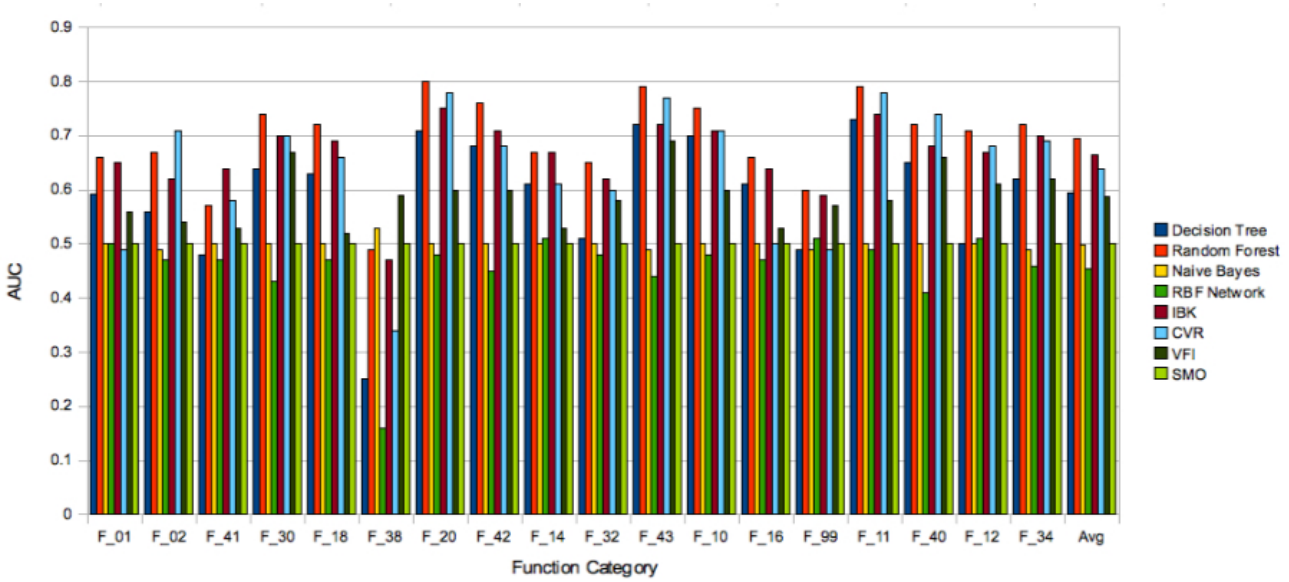
1. Find the suitable learning system
2. Compare to transductive learner (i.e., Majority Rule)
   - Metrics:
     - Area under the ROC Curve (AUC), Precision, Recall and F1
3. Datasets:
   - DIP-Core, VonMering, Krogan and MIPS

	Num of proteins	Num of interactions
DIP-Core	2388	4400
Krogan	2708	14246
MIPS	7928	44514
VonMering	2401	22000

Comparison of Learners

- Target: Choose the best inductive learning method for protein function prediction
- Analyze methods available on Weka data mining toolbox
  - Decision tree (J48), Random forest, Instance based learner (IBK), Naive Bayes, Radial basis function networks, Support vector machine (SMO), Classification Via Regression (CVR) and Voting Feature Intervals (VFI)
  - Input of the methods:
    - Select 700 important proteins based on Anova measure
    - Find the shortest path of each protein to those selected proteins
    - Use this information as the input of Weka
  - Find the AUC of each method in 10-fold cross validation
  - DIP-Core dataset

Comparison of Learners



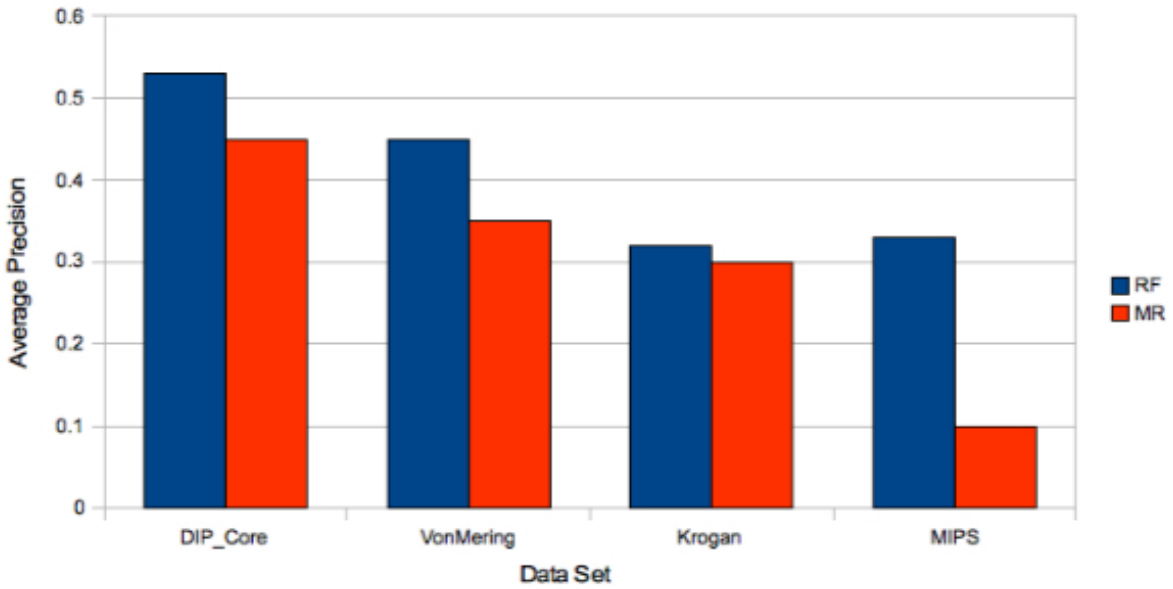
- Random Forest performs best among all the learners in 14 out of 18 cases
- Other 4 cases are all characterized by a very high class skew
- Random Forest: Best candidate for learning from this type of data

Comparison with a Transductive Method

- Use Random Forest for function prediction
  - Select 700 nodes based on Anova measure
  - Find the shortest path of each protein to those selected proteins
  - Use this information as the input of Weka
- Use Majority Rule for function prediction
  - Select three most occured functions in the neighborhood of the protein
- Compare Random Forest (RF) to Majority Rule (MR) based on comparison measures in 10-fold cross validation

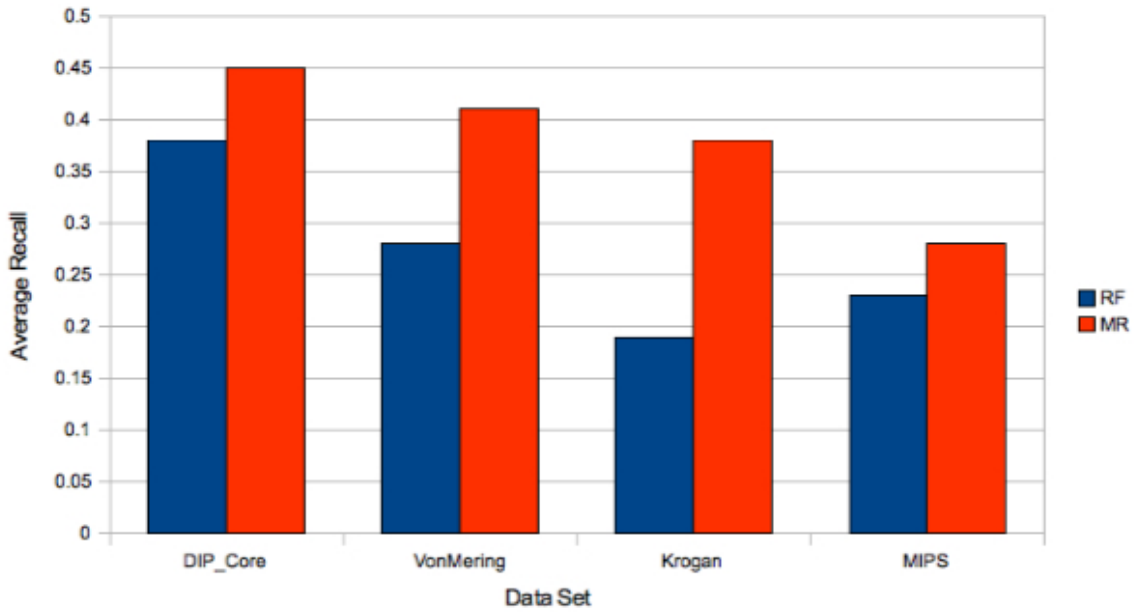
Average Precision

- Compare RF and MR based on average precision
- RF has higher precision (11 % higher in average)



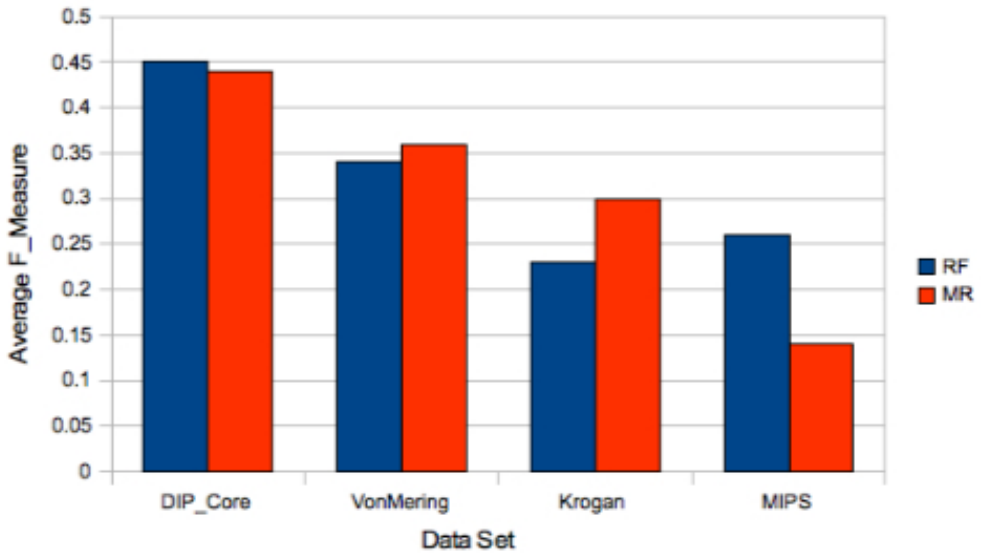
Average Recall

- Compare RF and MR based on average recall
- RF has smaller recall (10 % smaller in average)



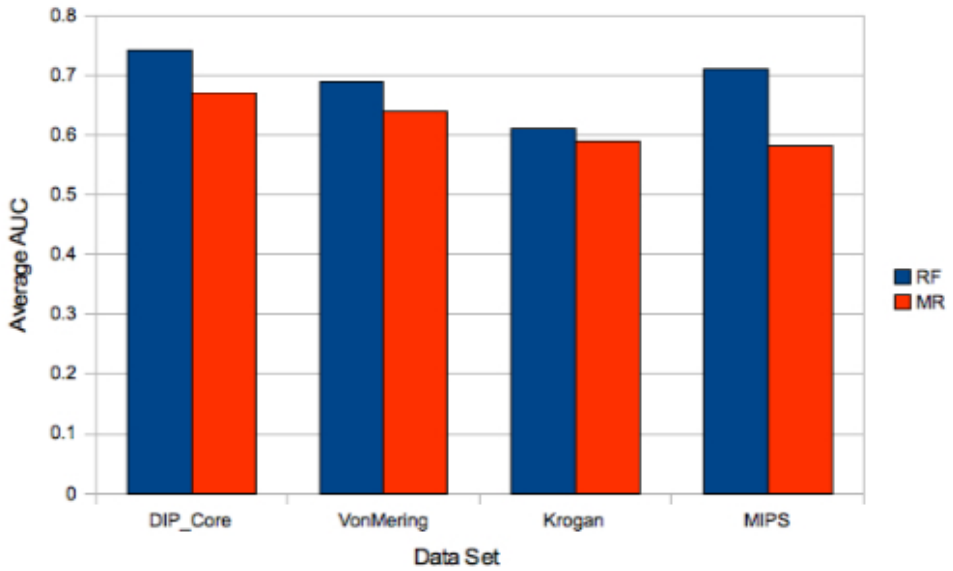
Average F1

- Compare RF and MR based on average F1-measure
- RF and MR perform almost similarly with respect to Fmeasure.



Average AUC

- Compare RF and MR based on average AUC
- RF tends to have higher scores (+6%)



Conclusions and Future Works

- Function prediction in PPI networks
- Discuss inductive and transductive methods and their biases
- Propose global node description formalism based on Anova
  - Global inductive method
  - Proved that our description (distance to important proteins) is informative
  - Outperforms the Majority Rule approach according to AUC and Precision
- Future Works:
  - Compare to more previous methods
  - Find out to what extent the global protein description is complementary to that used in other approaches
  - Multi-label classification

Thanks!