Analyzing Incomplete Biological Pathways Using Network Motifs

Maya Eldayeh

Michael Hahsler

DBI Retreat, May 6 and 12, 2011
UT Southwestern Medical Center

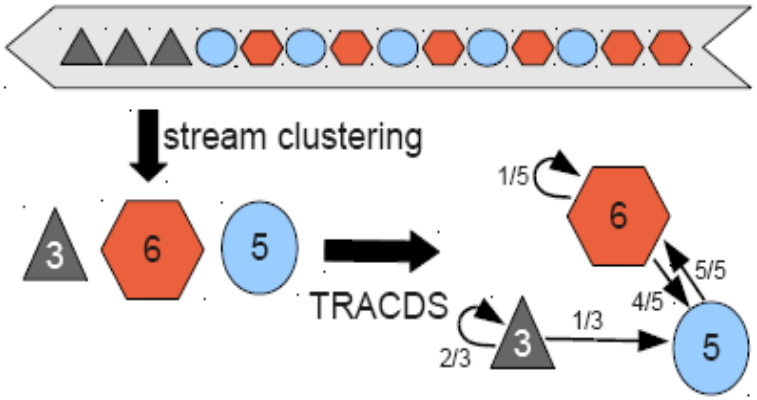
Intelligent Data Analysis Lab
Department of Computer Science and Engineering

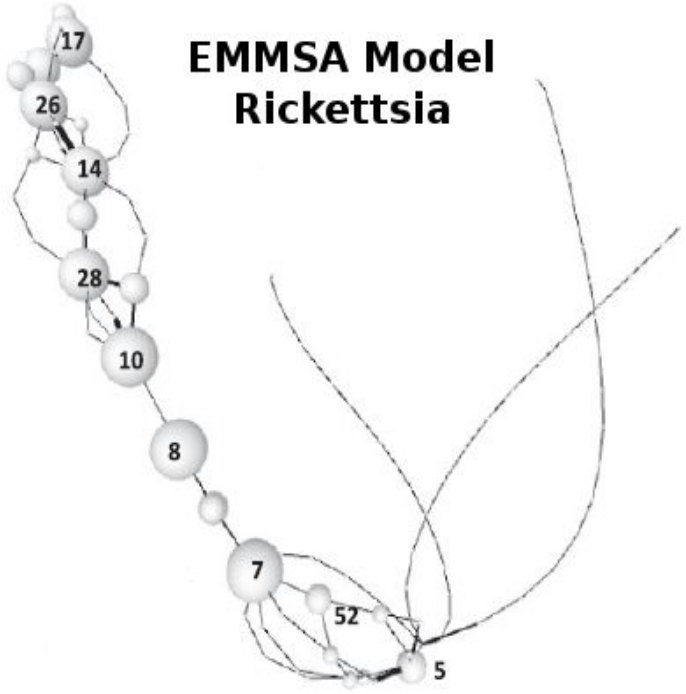
SMU | Bobby B. Lyle School of Engineering

IDA@SMU

http://lyle.smu.edu/IDA/

EMM Sequence Analysis
Sequence information preserving data stream clustering





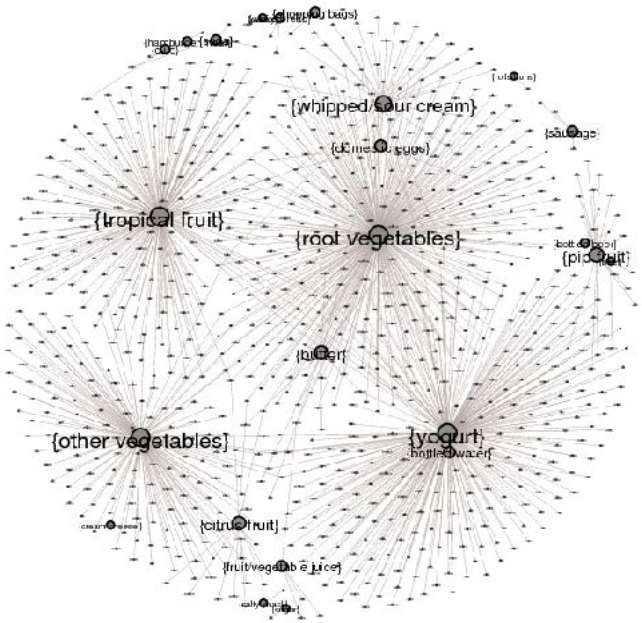
http://cran.r-project.org/package=rEMM

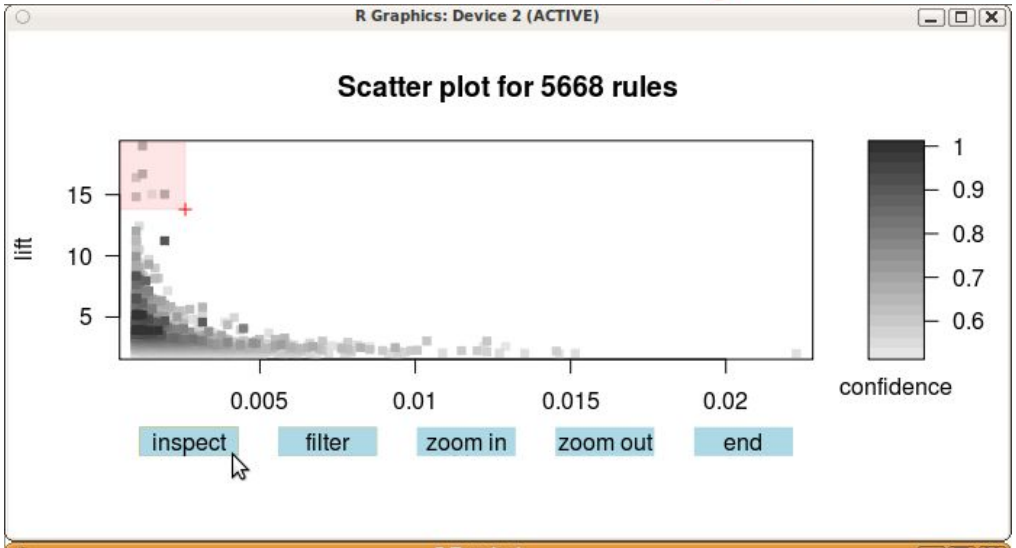
Intelligent Data Analysis Lab

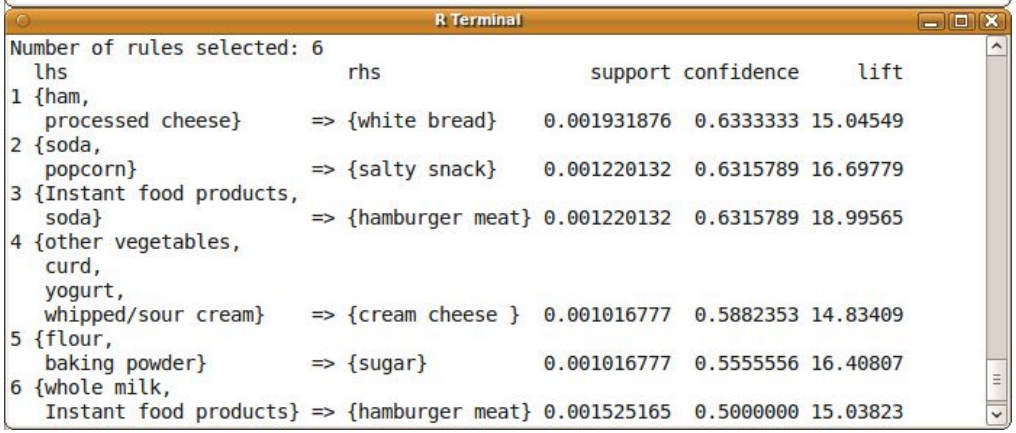
SMU | Bobby B. Lyle School of Engineering

Association Rules

- Interest measures
- Classification
- Visualization

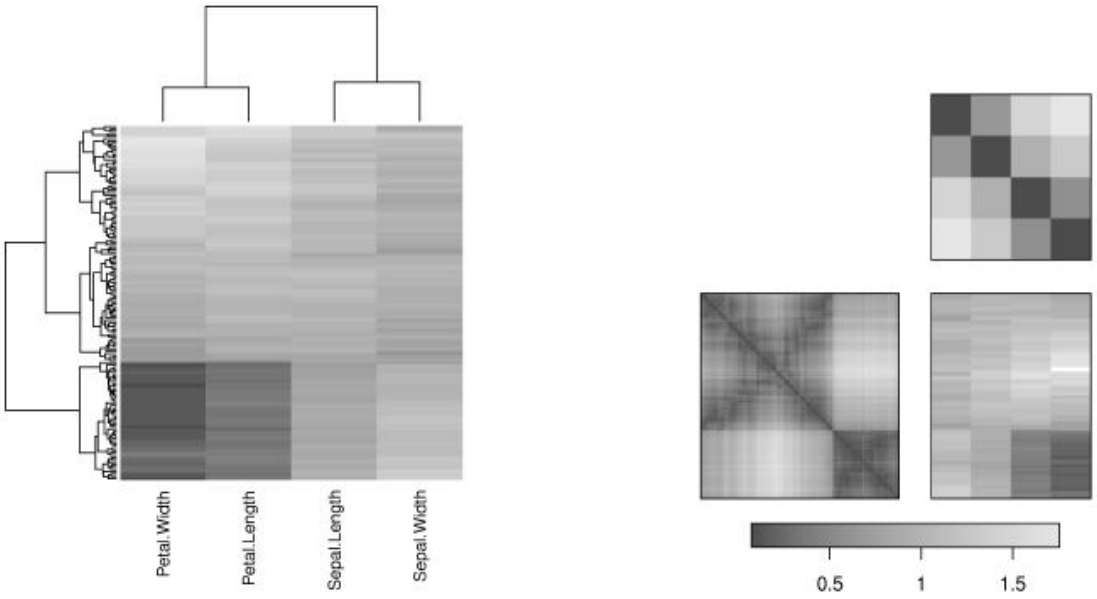




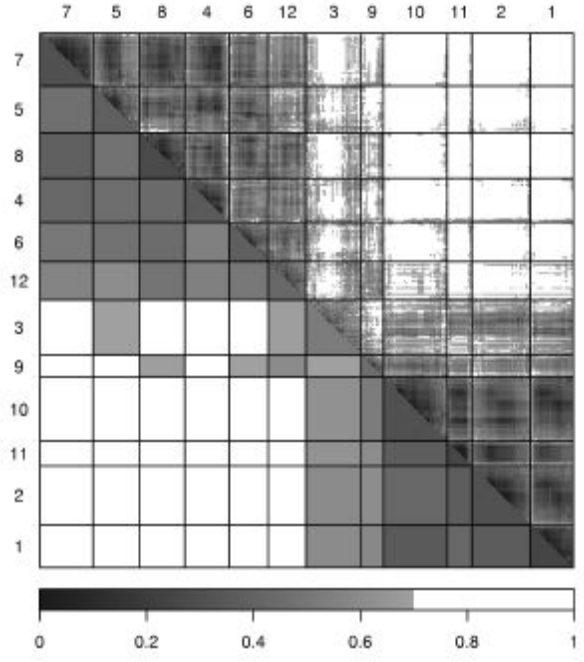


http://cran.r-project.org/package=arules
http://cran.r-project.org/package=arulesViz
http://cran.r-project.org/package=arulesSequences

Visualization using seriation



Iris data set



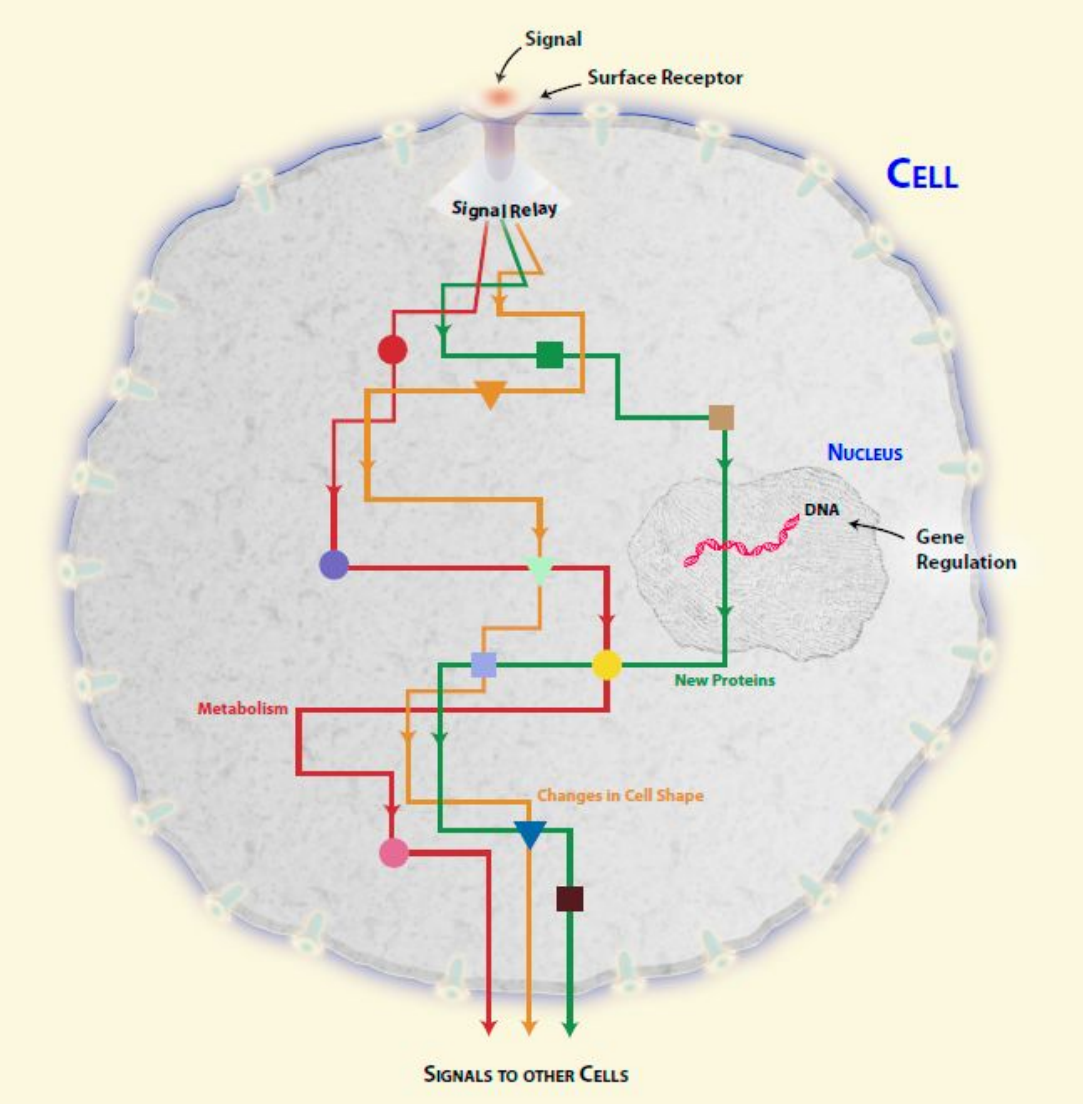
Votes 1984
64 binary variables
PAM with Jaccard k=12

http://cran.r-project.org/package=seriation

Talk Outline

1. Pathway Completion Problem
2. Pathway Motifs
3. Fit & Complete Algorithm
4. Experimental Results
5. Future Work

Biological Pathways

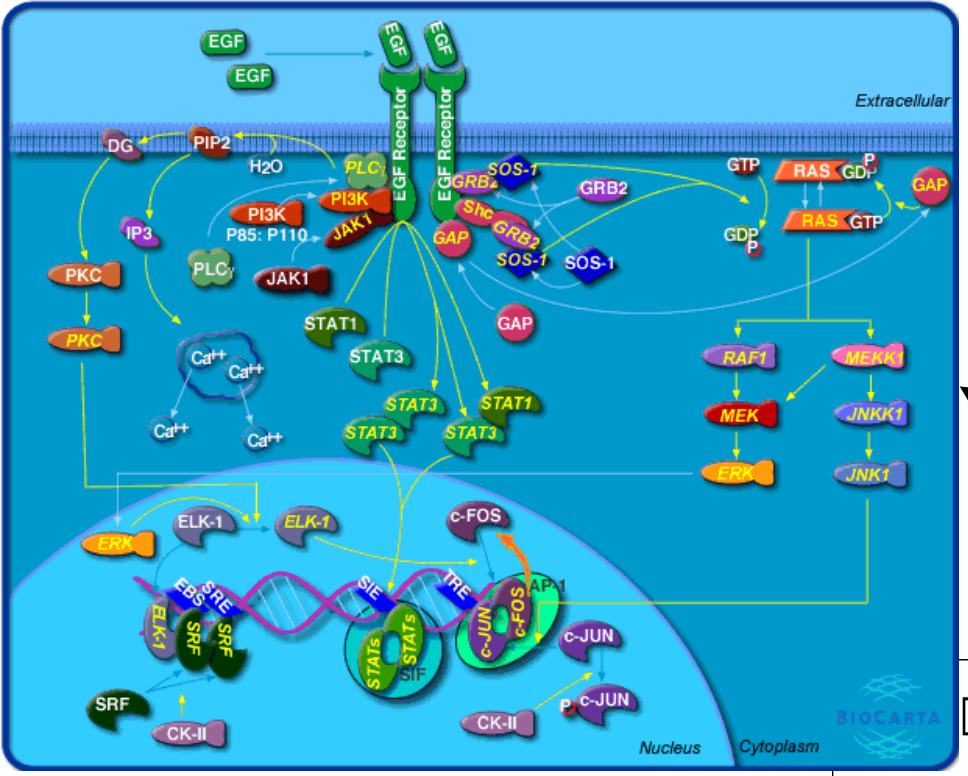


"Biological pathways are distinct, experimentally-validated subnetworks of proteins within the larger PPI network that interact with each other by well defined mechanisms to regulate a specific biologic phenotype." [8]

http://www.genome.gov/multimedia/illustrations/Biological_Pathways.pdf

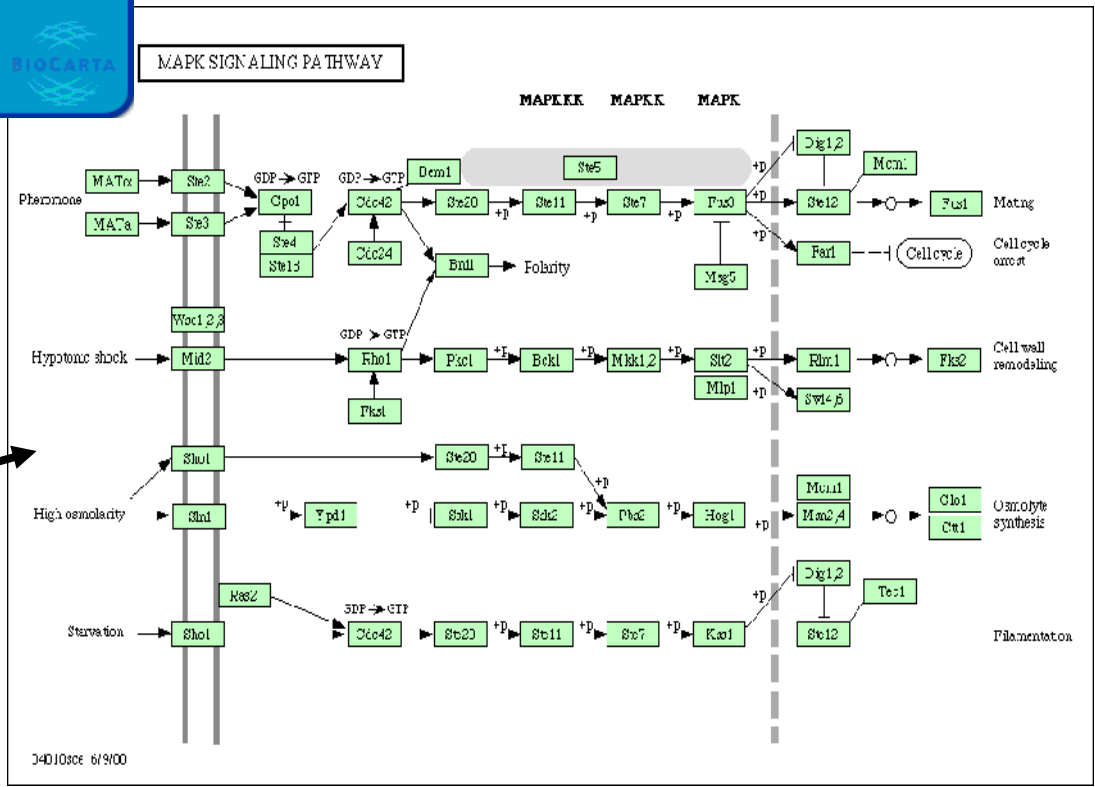
- Represented as diagrams, manually created, stored as digital pictures (e.g. *.gif)

BioCarta



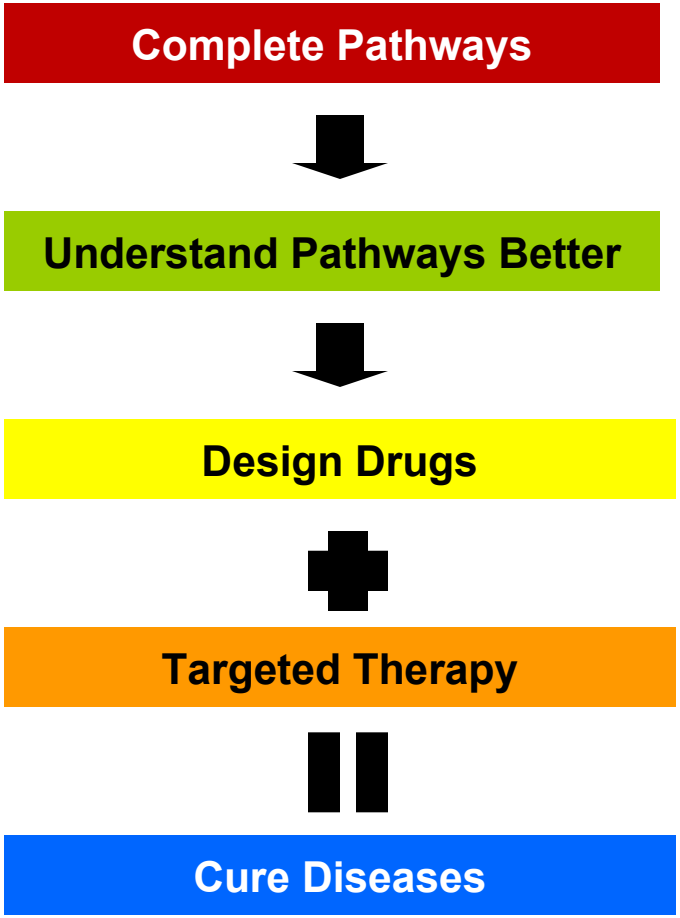
- KEGG: xml files available for download

KEGG



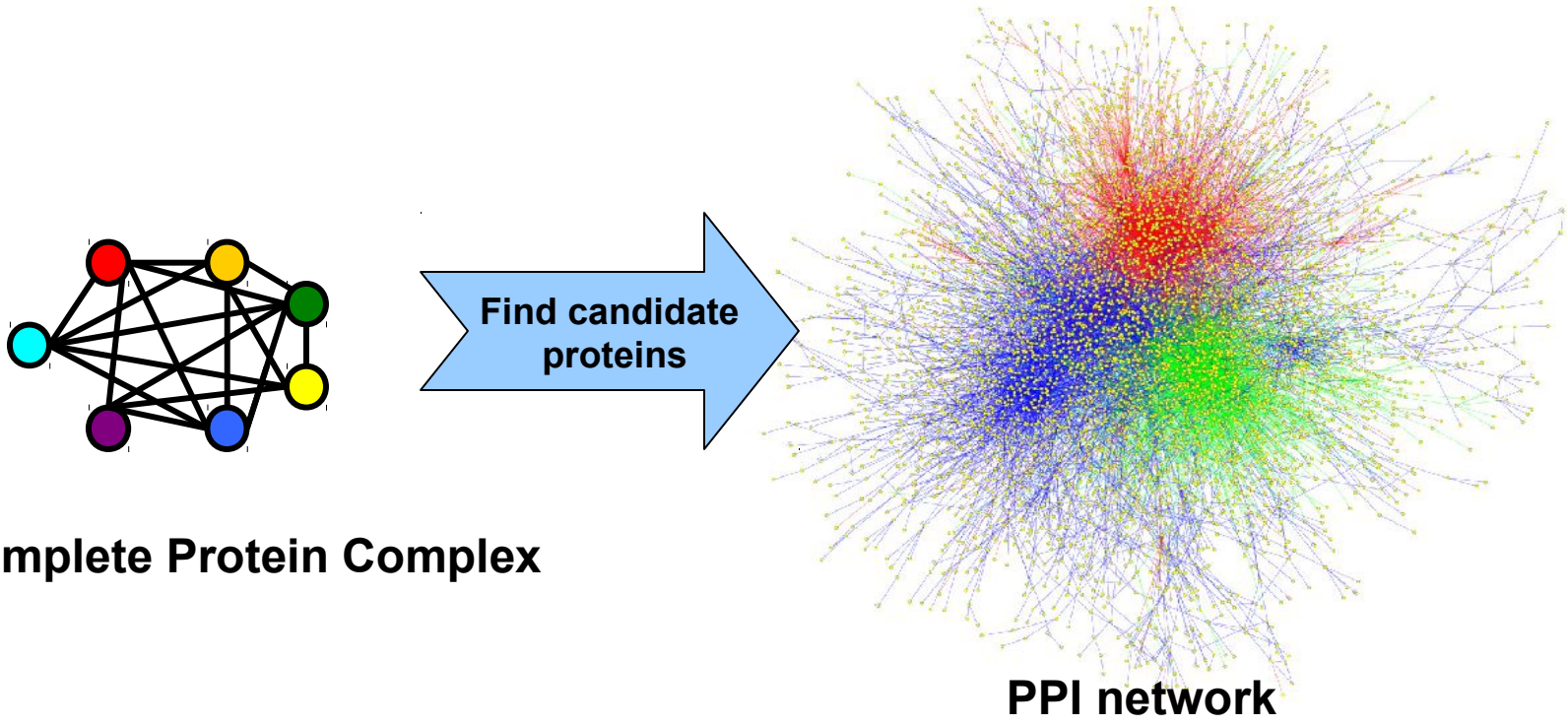
Pathway Completion

Providing candidate **proteins and their location in the pathway** helps to plan and execute targeted experiments.



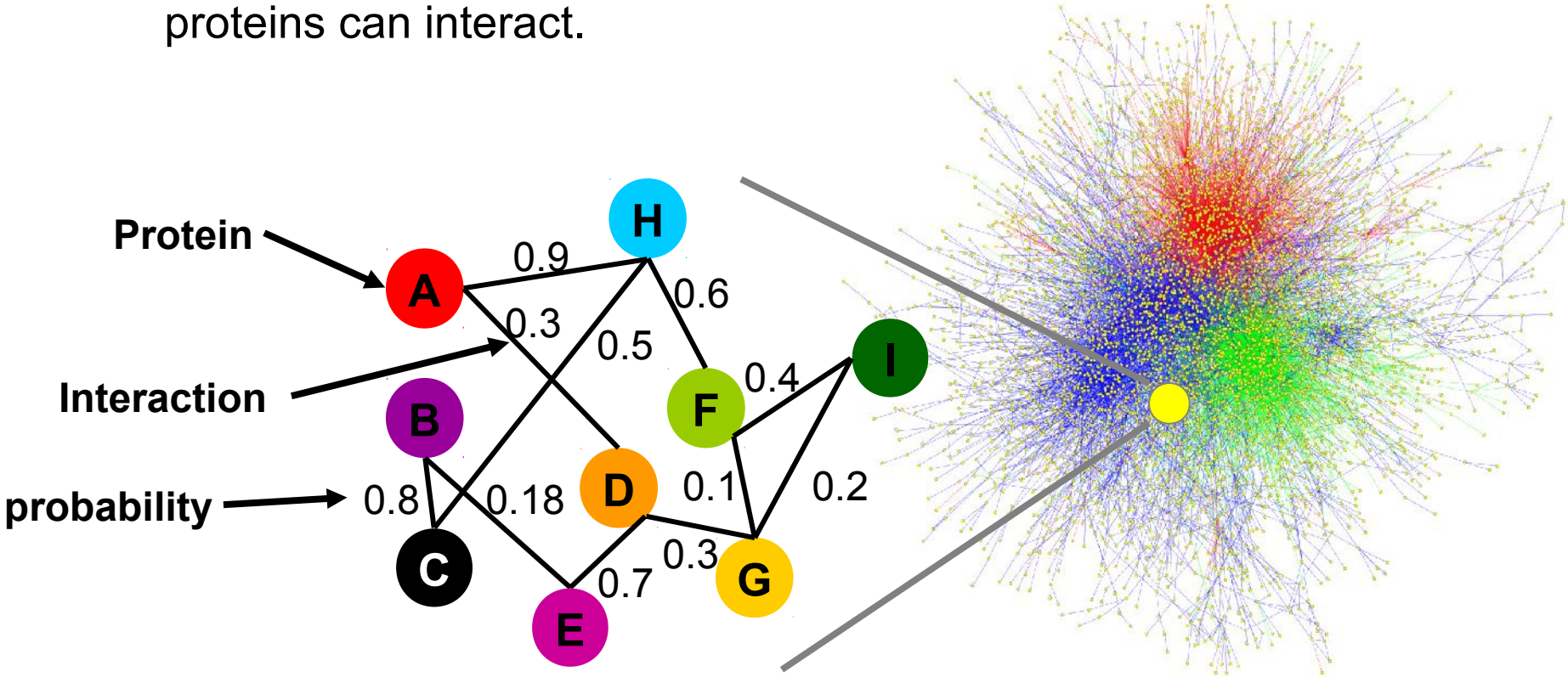
Similar Problem: Protein Complex Membership

A related studied problem is the protein complex membership problem.



Probabilistic PPI Networks

- Computational techniques to create probabilistic PPI network from the results of various experiments
- These networks have weights = probability that the proteins can interact.



Current Approaches: Membership Problem

Given a probabilistic PPI network and a protein complex:

- Uncover a list of candidate proteins from the network ranked according to degree of membership = **PROXIMITY**

Graph theory techniques and computational Method:

- **Network Reliability Using Monte Carlo Simulation (MCS):** Asthana et al. – Genome Research 2004 [4]
- **Random Walk (RW):** Can et al. – BIOKDD 2005 [1]
- **Net-Flow (NF):** Camoglu et al. – Advances in Bioinformatics 2009 [11]

All have:

- Used probabilistic PPI network
- Used 27 benchmark protein complexes
- Performed leave-one-out cross validation to test accuracy

RW applied technique on 10 benchmark pathways.

Current Approaches: Membership Problem

SMU | Bobby B. Lyle School of Engineering

Given a probabilistic PPI network and a protein complex:

- Uncover a list of candidate proteins from the network ranked according to degree of membership = **PROXIMITY**

Graph theory techniques and computational Method:

- **Network Reliability Using Monte Carlo Simulation (MCS):** Asthana et al. – Genome Research 2004 [4]
- **Random Walk (RW):** Can et al. – BIOKDD 2005 [1]
- **Net-Flow (NF):** Camoglu et al. – Advances in Bioinformatics 2009 [11]

All have:

- Used probabilistic PPI network
- Used 27 benchmark protein complexes
- Performed leave-one-out cross validation to test accuracy

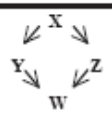
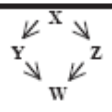
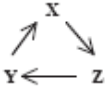
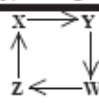
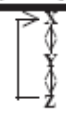
RW applied technique on 10 benchmark pathways.

Pathway Completion Problem

- Extracts candidate member proteins for a pathway from the network and indicates their location in the pathway via using motifs
- Different from complex/pathway membership problem

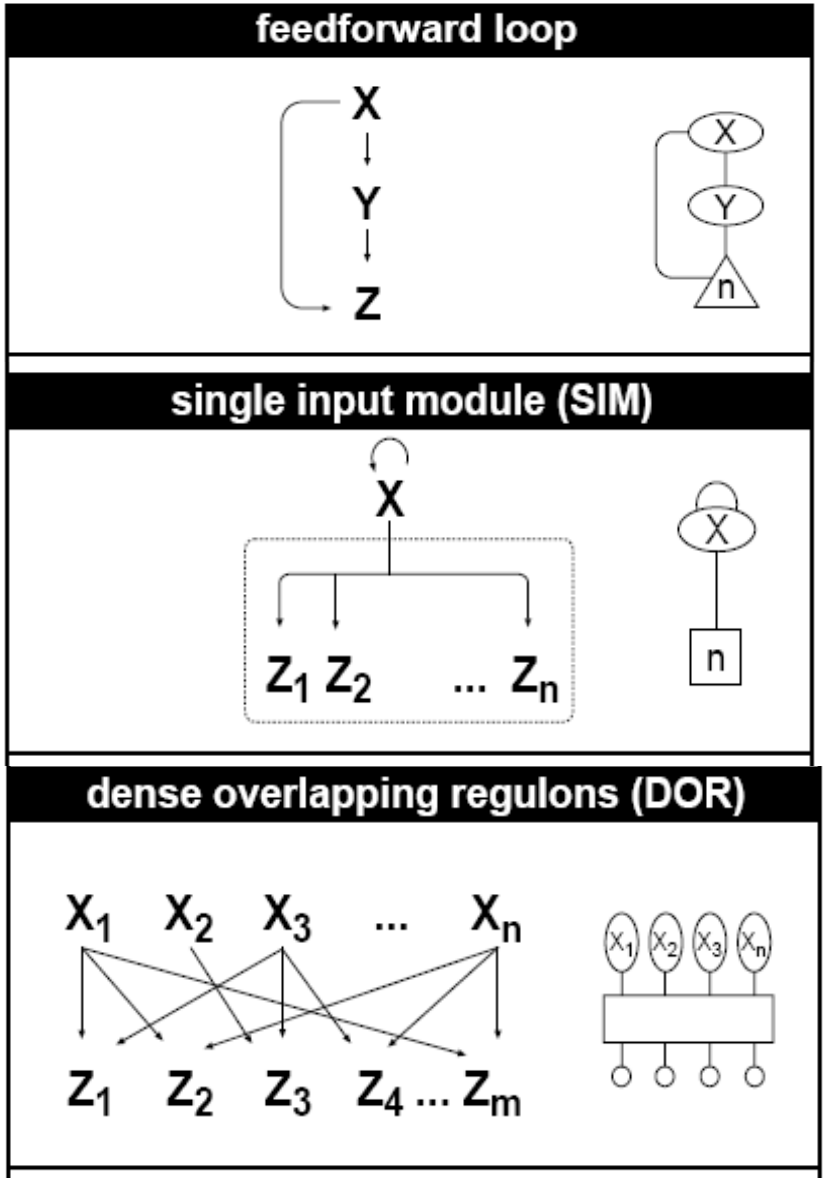
Method / Features	Membership	Completion
Complex/Pathway	Mainly Complex	Pathway
Probabilistic PPI network	**Yes**	**Yes**
Use Structure	**No**	**Yes**
Rank member protein	**Yes**	**Yes**
Provide location information	**No**	**Yes**

Motifs in Different Systems

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			
Electronic circuits (forward logic chips)					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

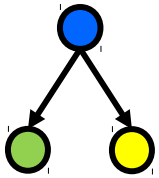
Network motifs are, “patterns of interconnections occurring in complex networks at numbers that are significantly high…” [9]

Motifs in Pathways

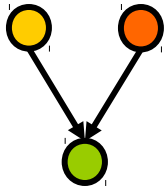


~ Linear Motif (LM)

~ Single Input Motif (SIM)



~ Multiple Input Motif (MIM)

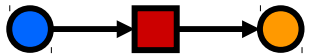
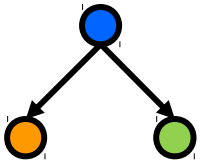
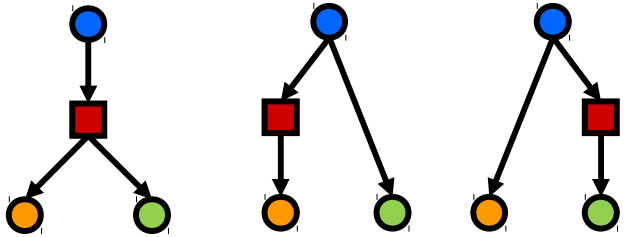
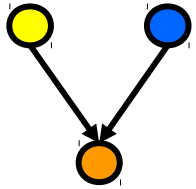
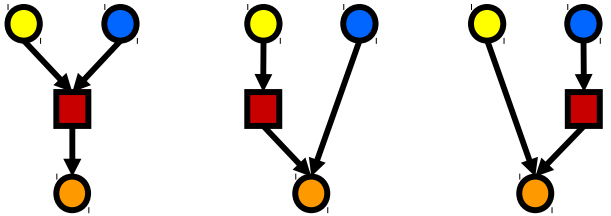


[10]

Proposed Complete Motif Structures

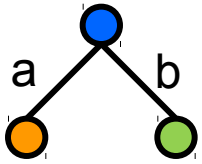
SMU | Bobby B. Lyle School of Engineering

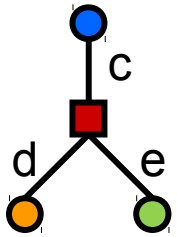
Use the idea of network motifs

Motif Name	Original Structure	Proposed Complete Structure
Linear		
Single Input		
Multiple Input		

Scoring

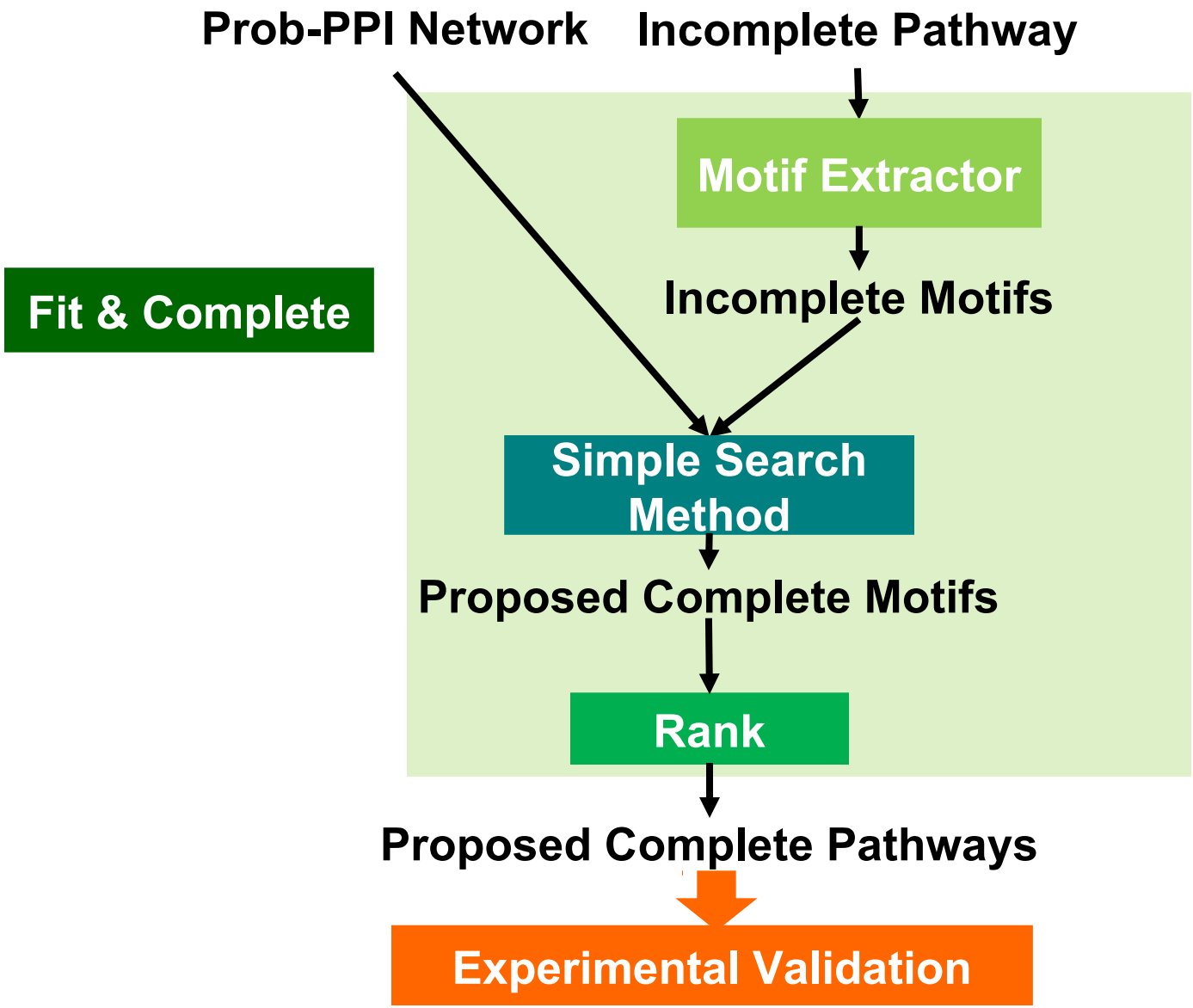
- Determine if proposed complete motif structures are “better” than the original motif
- Simple scoring scheme as a starting point
- Use minimum, maximum, and average of the weights on the edges from the probabilistic PPI network
- Want to incorporate biological knowledge





Original Motif Score	**Proposed Complete Motif Score**
Min(a, b)	Min(c, d, e)
Max(a, b)	Max(c, d, e)
Avg(a, b)	Avg(c, d, e)

The Fit and Complete Algorithm



The Fit & Complete Algorithm

Input: weighted graph, $G = (V,E)$ // *probabilistic PPI network*

directed graph, $G' = (V', E')$ // *incomplete pathway*

(1) Generate L, the set of linear motif sub-graphs of G'

(2) For each l in L do // for each motif in L

(3) For each edge e in l

// compute the score of the original motif

(4) Find w_e in G, the weight of the corresponding edge in G to e

(5) Add w_e to l, the edge weight of e

(6) S_L = scoreMotif(l) // returns s_l the score for motif l

(7) (L^c, S^c) = Search(G, l) // finds L^c the ordered set of possible complete motifs *and* S^c the ordered set of their scores

(8) Return (S_L, S^c, L^c) // S_L the set of original motif scores

The Fit and Complete Algorithm

Based on the unrealistic assumption that probabilistic PPI networks contains all interactions

The Search Method

Input: weighted graph, $G = (V,E)$ // *probabilistic PPI network*

motif l *in* L

(1) Let $L^c = \Phi$, the empty ordered set of possible complete motifs

(2) Let $S^c = \Phi$, the empty ordered set of their scores

(3) Let v_1 and v_2 be the two vertcies of l

(4) Find N the set of all common first neighbors of v_1 and v_2 of l in G

(5) For each n in N

(6) Find w_n, the weight on the edge incident in G

(7) Create motif l^c with edges $\{(v_1, n), (n, v_2)\}$ and find the edge weights in G

(8) Add l^c to L^c // l^c is the complete motif

(9) S^c = scoreMotif(l^c) // returns s_l^c the score of l^c

(10) Return (L^c, S^c)

Preliminary Experiments

SMU | Bobby B. Lyle School of Engineering

Yeast Network	Description	Number of Proteins	Number of Probabilistic interactions
Naïve Bayes [4]	• Probabilistic network Derived from results of 4 large-scale experiments	3,112	12,594
ConfidentNet [6]	• Probabilistic functional network connections among proteins are predicted using a Bayesian approach by integrating 5 different datasets	5,552	235,222
PIT-Network [7]	• Combination of predicted and experimental using Naïve Bayes	5,240	91,768

Example Pathways

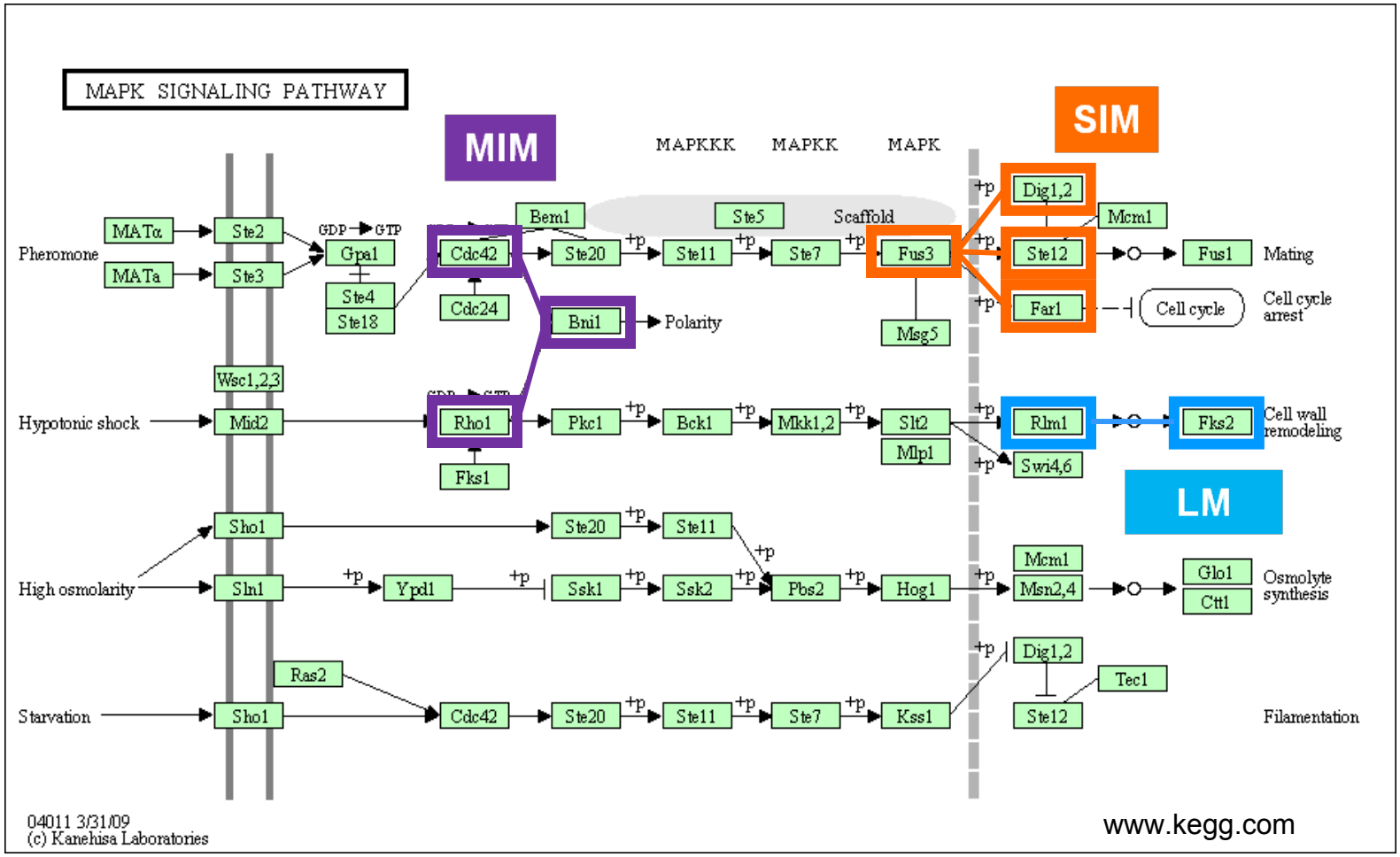
SMU | Bobby B. Lyle School of Engineering

- Applied our algorithm to ConfidentNet and five KEGG pathways.

- ConfidentNet has **5,552 proteins** and **235,222 probabilistic interactions.**

Pathways	Proteins	Interactions	LM	MIM	SIM	Motifs
Endocytosis	17	17	19	5	4	28
Cell Cycle	57	89	63	12	13	88
Regulation of Autophagy	17	14	14	1	2	17
Meiosis	59	145	108	27	20	155
MAPK Signaling Pathway	51	61	65	14	11	90

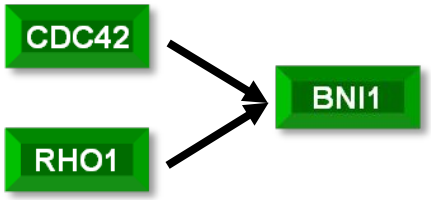
Experiments and Results



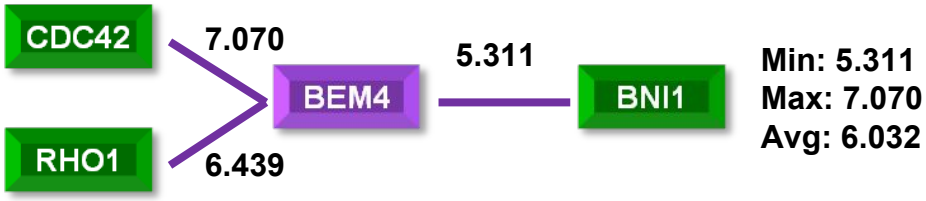
Results: MIM – Examples from first run

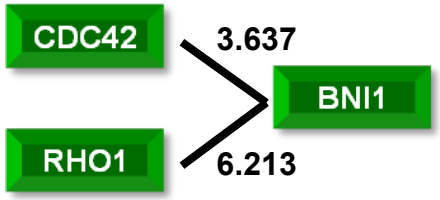
SMU | Bobby B. Lyle School of Engineering

MAPK Pathway

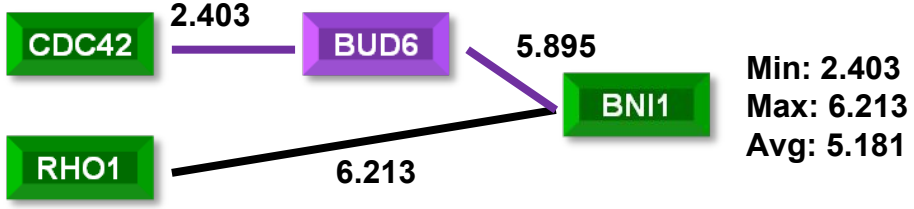


Probabilistic PPI Network





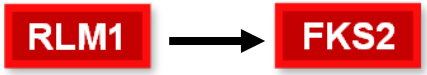
Min: 3.637
Max: 6.213
Avg: 4.925



Results: LM – Examples from first run

SMU | Bobby B. Lyle School of Engineering

MAPK Pathway



Probabilistic PPI Network

Min: 3.920
Max: 5.404
Avg: 4.662

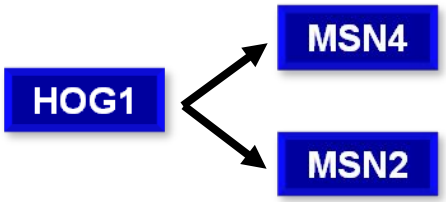
Min: 1.319
Max: 1.319
Avg: 1.319

Min: 3.908
Max: 5.296
Avg: 4.602

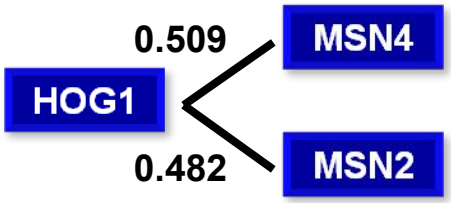
Results: SIM – Examples from first run

SMU | Bobby B. Lyle School of Engineering

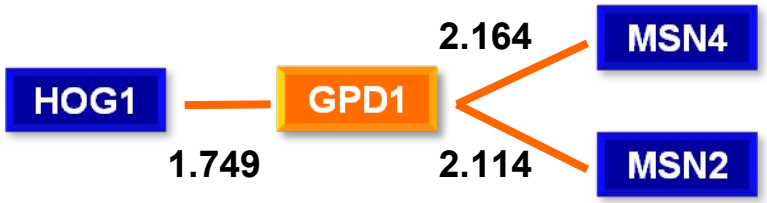
MAPK Pathway



Probabilistic PPI Network

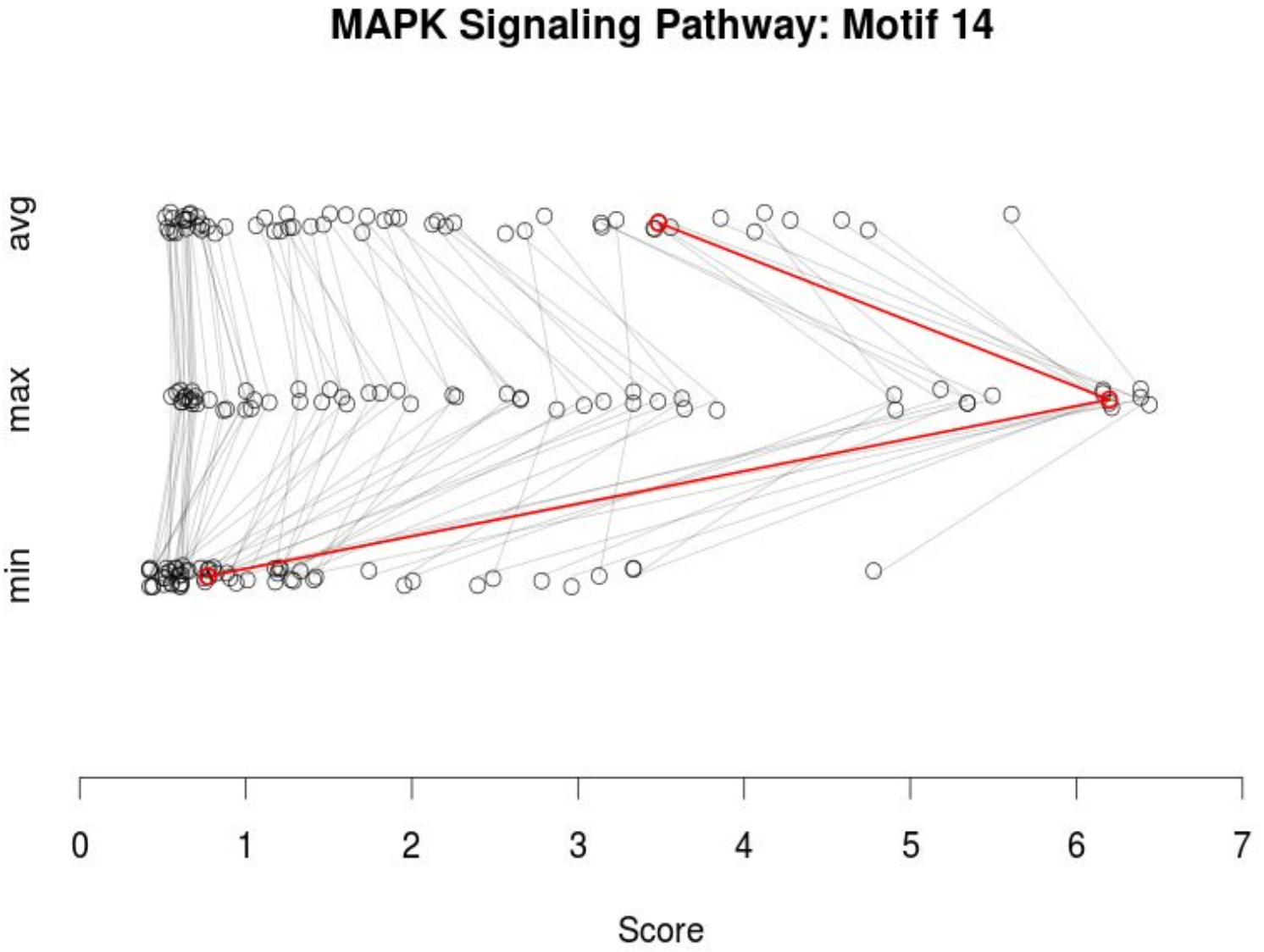


Min: 0.482
Max: 0.509
Avg: 0.495

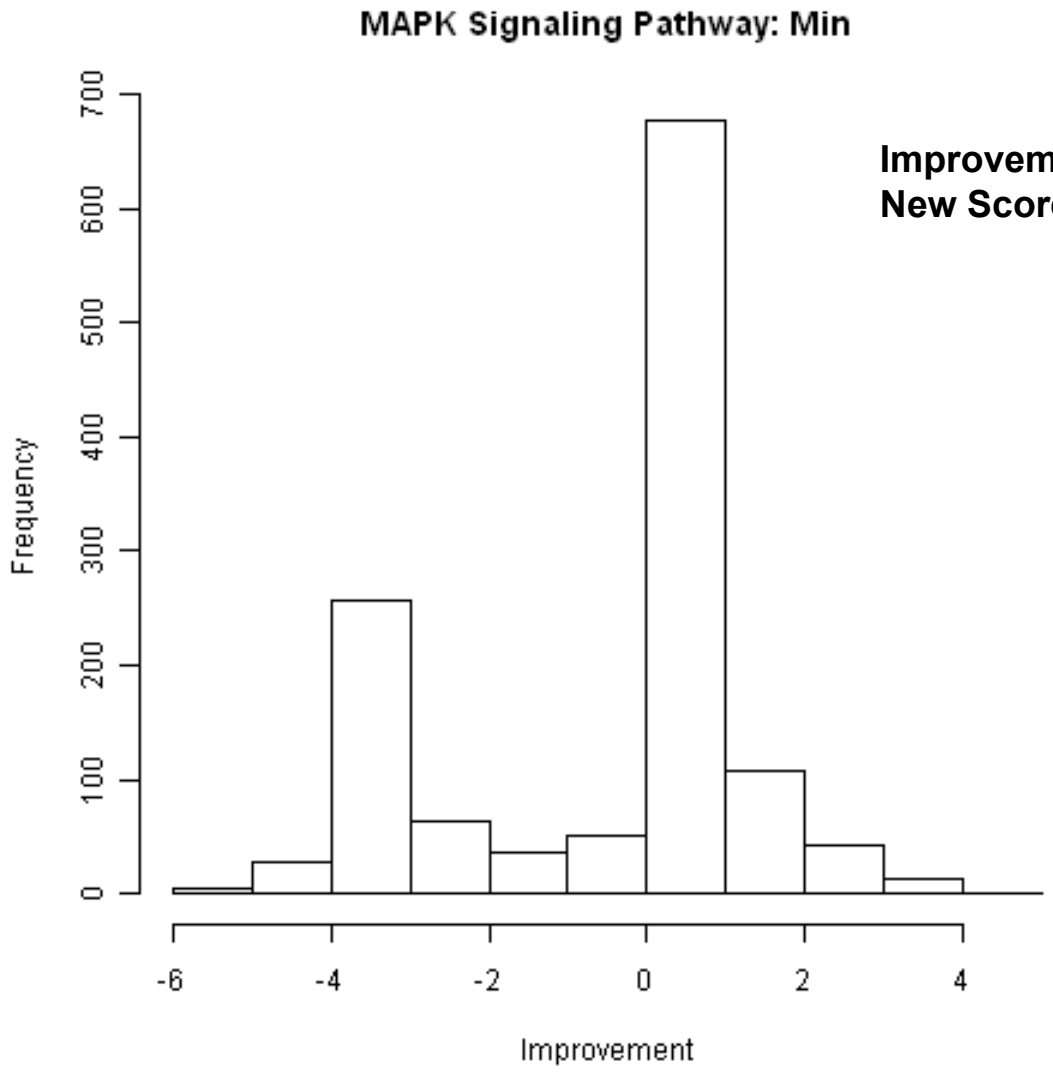


Min: 1.749
Max: 2.164
Avg: 1.944

Results for one Motif: Parallel Coordinates

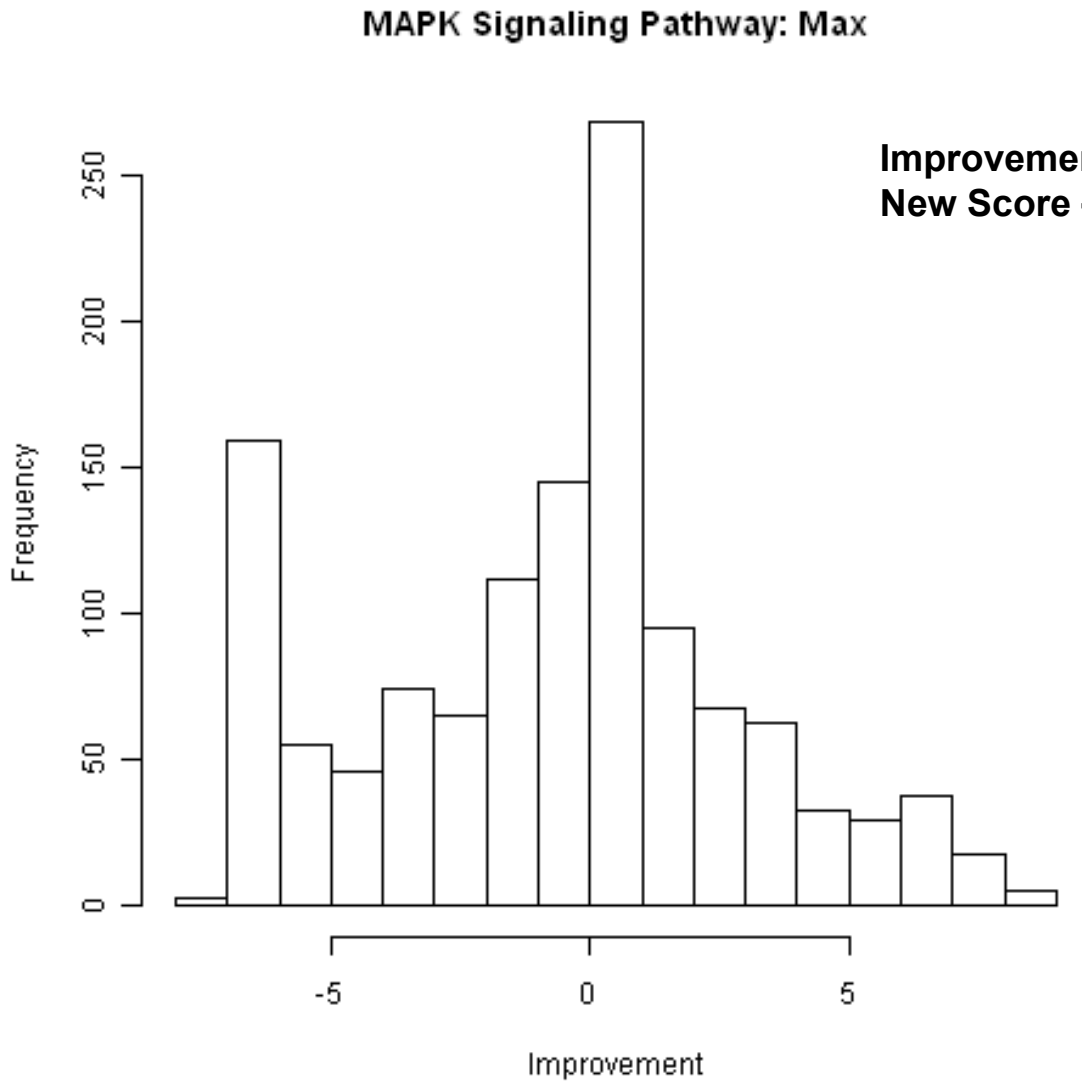


Results: Improvement over all Motifs - Min



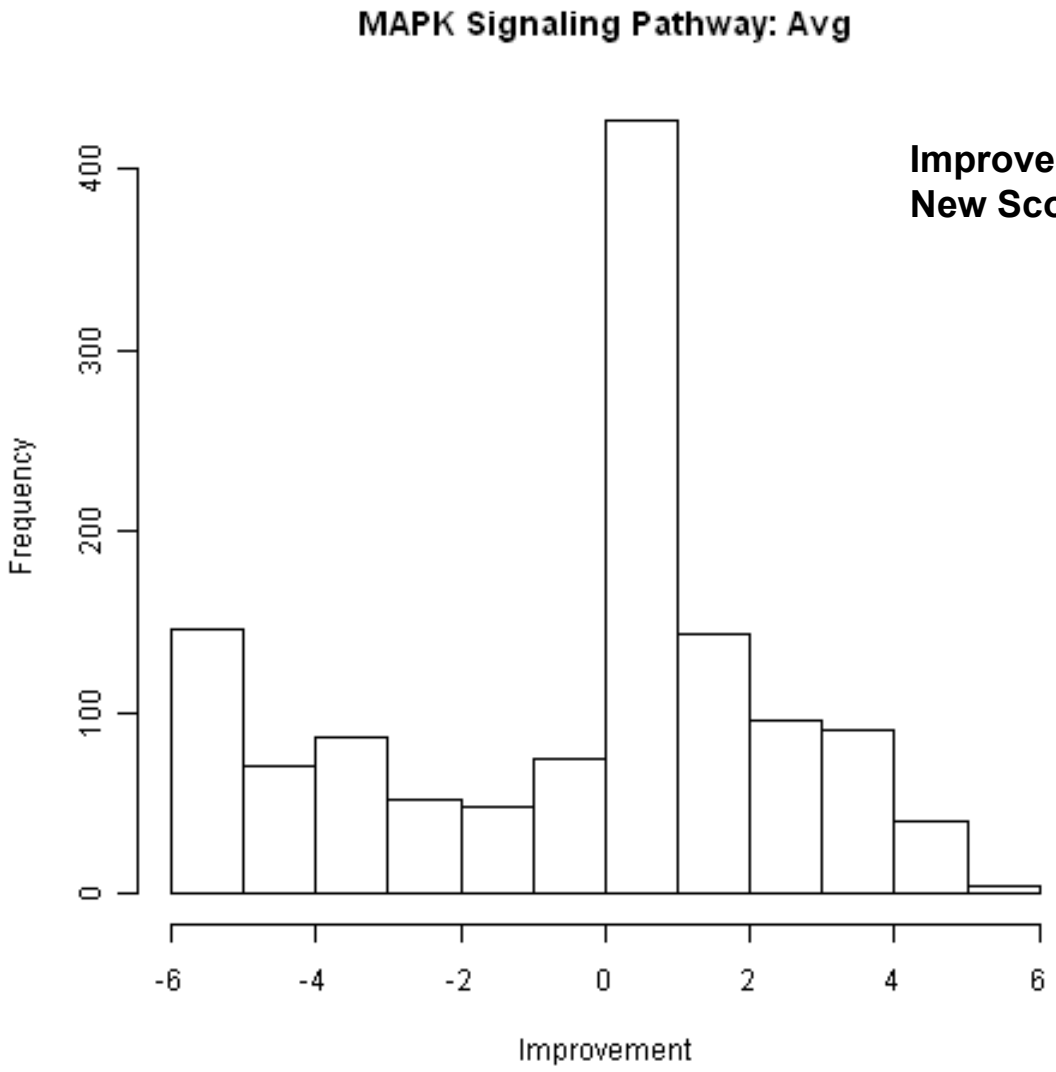
Improvement =
New Score – Original Score

Results: Improvement over all Motifs - Max



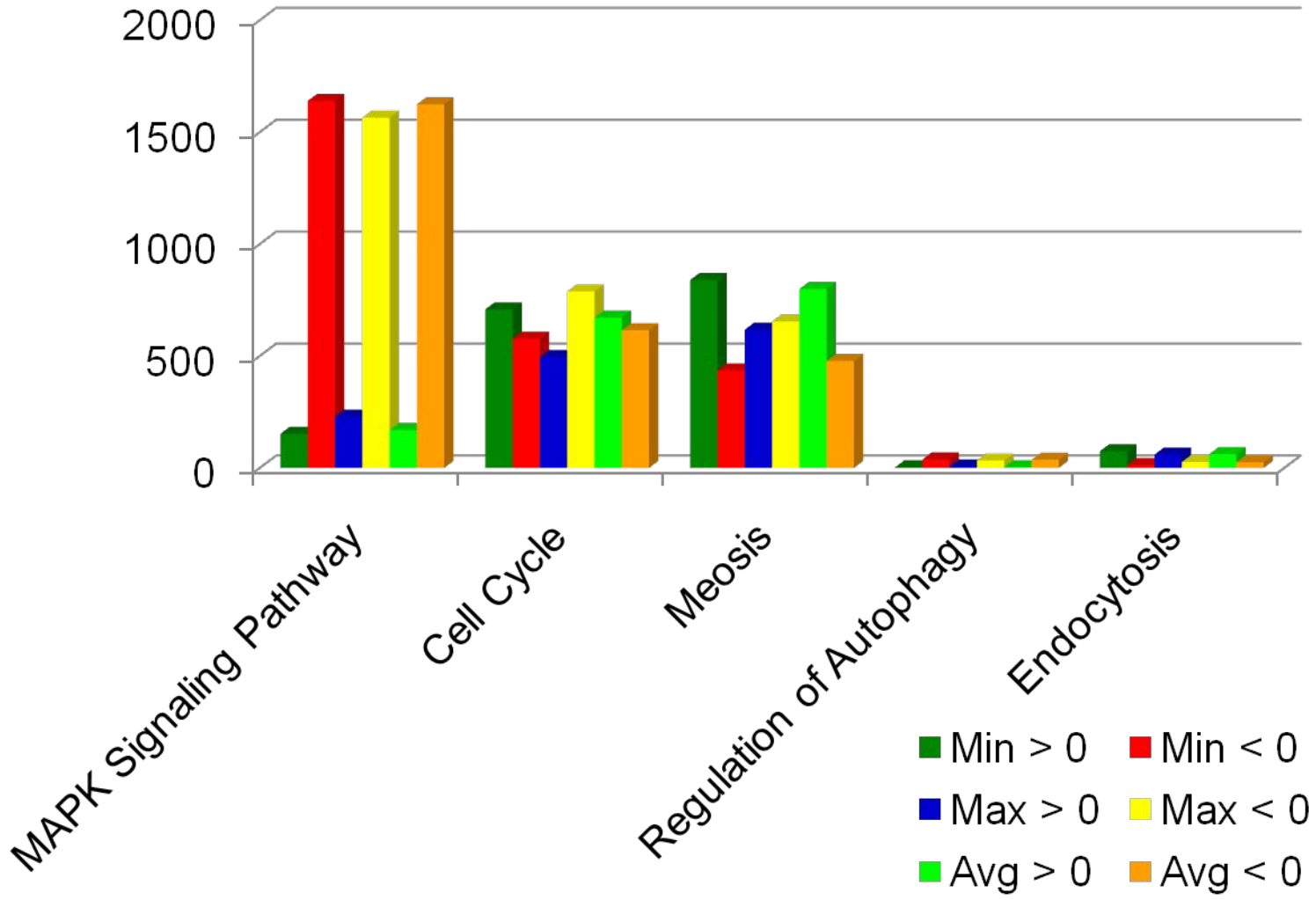
Improvement =
New Score – Original Score

Results: Improvement over all Motifs - Avg



Improvement =
New Score – Original Score

Results: Summary Statistics



Future Work

- Modify our search method to handle noise; implemented RW
- Biologically motivated scores
- Explore edge deletion as well (not only protein addition)
- Handle MIMs and SIMs
- Algorithm Accuracy (leave-one-out cross validation)