

# Transcriptional Regulation of Protein Complexes In and Across Species

K. Tan, T. Shlomi, H. Feizi, T. Ideker, and R. Sharan

*Proc. Natl. Acad. Sci. USA.*, 104 (4), 1283-1288, 2007

Presented by:

D. Beck

03/29/2007

# Introduction

- Objectives
- Data
- Methods
- Results
- Issues and Concerns

# Objectives

- Experimental noise and network structure make identification of functional modules in protein-protein interaction networks (PPINs) difficult.
- Evidence exists demonstrating that proteins within a complex are often encoded by genes which are controlled by the same transcription factor (TF).
- Identification of modules by finding coregulated protein complexes.

# Objectives (cont'd)

- Validation of the modules by functional enrichment, correlation of expression, and phylogenetic conservation.
- Extension of technique across species to identify conserved protein complexes and unknown transcriptional interactions.

# Data: Protein-Protein Interactions

- Yeast and fly PPI data was obtained from the Database of Interacting Proteins (DIP) in November 2005.
- Confidence estimates were computed by a logistic regression model with inputs  $X = \{\text{number of experimental observations, Pearson coefficient, small-world clustering coefficient}\}$  [1]
- Positive training data was taken from Munich Information Center for Protein Sequences (MIPS); negative training data was obtained by random pairs.

# Data: Protein-Protein Interactions (cont'd)

- $w, v$ : proteins
- $N$ : number of proteins in the network
- $\Pr(T_{uv} | X)$ : Probability of a true interaction between  $u$  and  $v$  with input  $X$ .
- $B_i$ : Training parameters

$$C_{vw} = -\log \sum_{i=|N(v) \cap N(w)|}^{\min\{|N(v)|, |N(w)|\}} \frac{\binom{|N(v)|}{i} \binom{N - |N(v)|}{|N(w)| - i}}{\binom{N}{|N(w)|}}$$

$$\Pr(T_{uv} | X) = \frac{1}{1 + \exp(-\beta_0 - \sum_{i=1}^3 \beta_i X_i)}$$

# Data: Protein-DNA Interactions

- Transcriptional interactions for yeast were obtained from Harbison, et. al. [2].
- The confidence of each interaction was set at 0.96, based on the 4% false positive rate cited in the paper.
- Experimentally derived and manually curated coregulated protein complexes were obtained from the MIPS database.

# Data: Gene Expression Profiles and Sequences

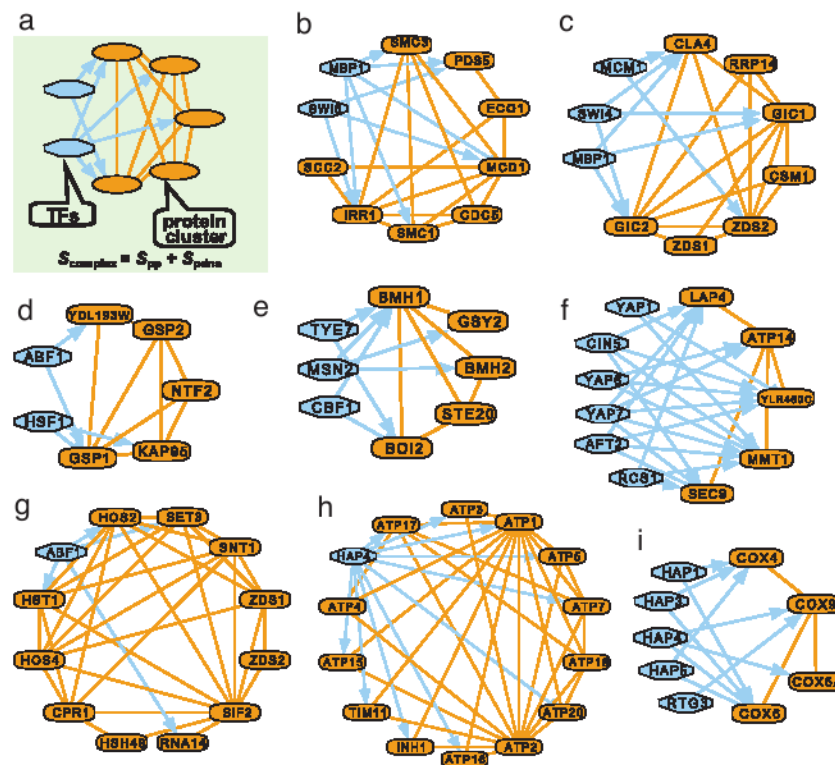
- Yeast gene expression measurements were obtained from the Stanford Microarray Database and covered 973 conditions across a wide variety of cellular states.
- Expression profiles used in validation were performed for the wild-type haploid BY4147 and *rpn4* deletion strains under heat-shock induced stress.
- Normalized log expression values were obtained for the latter using the VERA package.

# Data: Gene Expression Profiles and Sequences (cont'd)

- Yeast sequence data was obtained from the Washington University Genome Sequencing Center.
- Fly sequence data was obtained from the UCSC Genome Browser website.

# Methods: Network Construction

- Orange nodes are proteins; blue nodes are TFs.
- Orange edges indicate PPIs; blue edges indicate TF regulation of a protein.



# Methods: Protein Complex Scoring and Discovery

- A greedy approach that starts from high-scoring seeds consisting of a protein and two interacting partners plus at least one transcription factor is used to identify clusters.
- These seeds are refined using local search to obtain a log likelihood ratio score for a cluster.
- A significance score is computed by comparing clusters to clusters obtained by application of the algorithm to a random network.
- Retained clusters are filtered so that the overlap between any two clusters is less than 80%.

# Methods: Protein Complex Scoring and Discovery (cont'd)

- $a, b$ : interaction probabilities
- $P(u,v)$ : probability of interaction in experimentally derived network.
- $R(u,v)$ : probability of interaction in random network.
- $S$ : protein cluster.
- $T$ : transcription factor.

$$L(S, T) = \sum_{u,v \in S} \log \frac{\alpha P(u, v) + (1 - \alpha)(1 - P(u, v))}{R(u, v)P(u, v) + (1 - R(u, v))(1 - P(u, v))} + \sum_{u \in S, t \in T} \log \frac{\beta P(u, t) + (1 - \beta)(1 - P(u, t))}{R(u, t)P(u, t) + (1 - R(u, t))(1 - P(u, t))}$$

# Results: Identified Clusters

- 72 protein complexes were identified in yeast.
- Association between MIPS complexes and transcription factors was computed using a hypergeometric distribution and transcription factor data from Harbison, et. al. [2].
- Only 9 of the manually curated MIPS complexes were strongly associated with a transcription factor; 50 of the identified complexes had no overlap with these 9.

# Results: Validation of Identified Complexes

- Complexes were validated using functional enrichment, expression coherency, and phylogenetic coherency.
- Functional annotations were obtained from yeast GO annotations.
- Expression correlations were obtained from the mean pairwise Pearson correlation of gene expression profiles among complex members.

# Results: Validation of Identified Complexes (cont'd)

- Phylogenetic coherency was obtained from mean similarity of sequence alignment computed by a Jaccard measure.

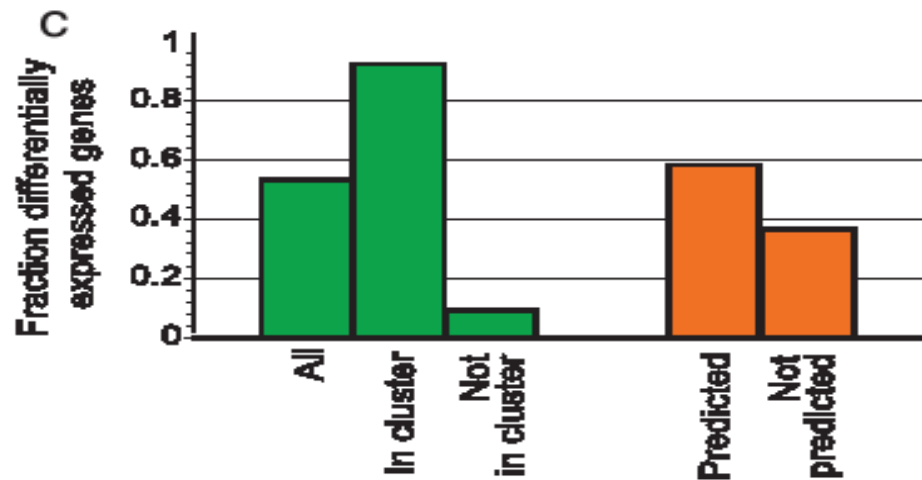
**Table 1. Validation of yeast clusters by functional enrichment, expression coherency, and conservation coherency of their members\***

	Complex source	GO enrichment, %	Expression coherency, %	Conservation coherency, %
MS-derived complexes	Ho <i>et al.</i> (26)	61	8	24
	Gavin <i>et al.</i> (27)	77	9	36
Protein clusters	Current study	99	26	22
Coregulated clusters	Current study	100	45	59

\*All analyses were restricted to clusters of size at least 7, although the same trends were observed over a wide range of cluster size cutoffs.

# Results: Experimental Validation of Novel Clusters

- Expression profiles for wild-type and *rpn4* deletion strains were used to validate predicted complexes.



**Fig. 2.** Transcriptional interaction prediction in yeast. (a) Receiver operating characteristics curve of the logistic regression classifier. AUC, area under the curve;  $S_n$ , sensitivity;  $S_p$ , specificity. (b) An example of a predicted cluster regulated by Rpn4. Orange arrows, known Rpn4 TIs from Harbison *et al.* (28); purple, newly predicted Rpn4 TIs. Shades of red represent  $P$  values ( $\leq 0.05$ ) for differential gene expression. (c) Fraction of differentially expressed genes in various gene sets. Green, genes bound by Rpn4 from the Harbison data; orange, genes in cluster models but not bound by Rpn4 based on the Harbison data.

# Results: Extension Across Species

- The algorithm was applied to fly protein networks using yeast transcriptional interaction data.
- This was validated by computing the probability of overlap between these complexes and clusters computed from fly PPIN only.
- It is claimed that 24 highly functionally enriched complexes were found.
- Conservation of transcription factors was obtained by sequence alignment of fly promoter regions and known fly transcription factors.

# Results: Extension Across Species (cont'd)

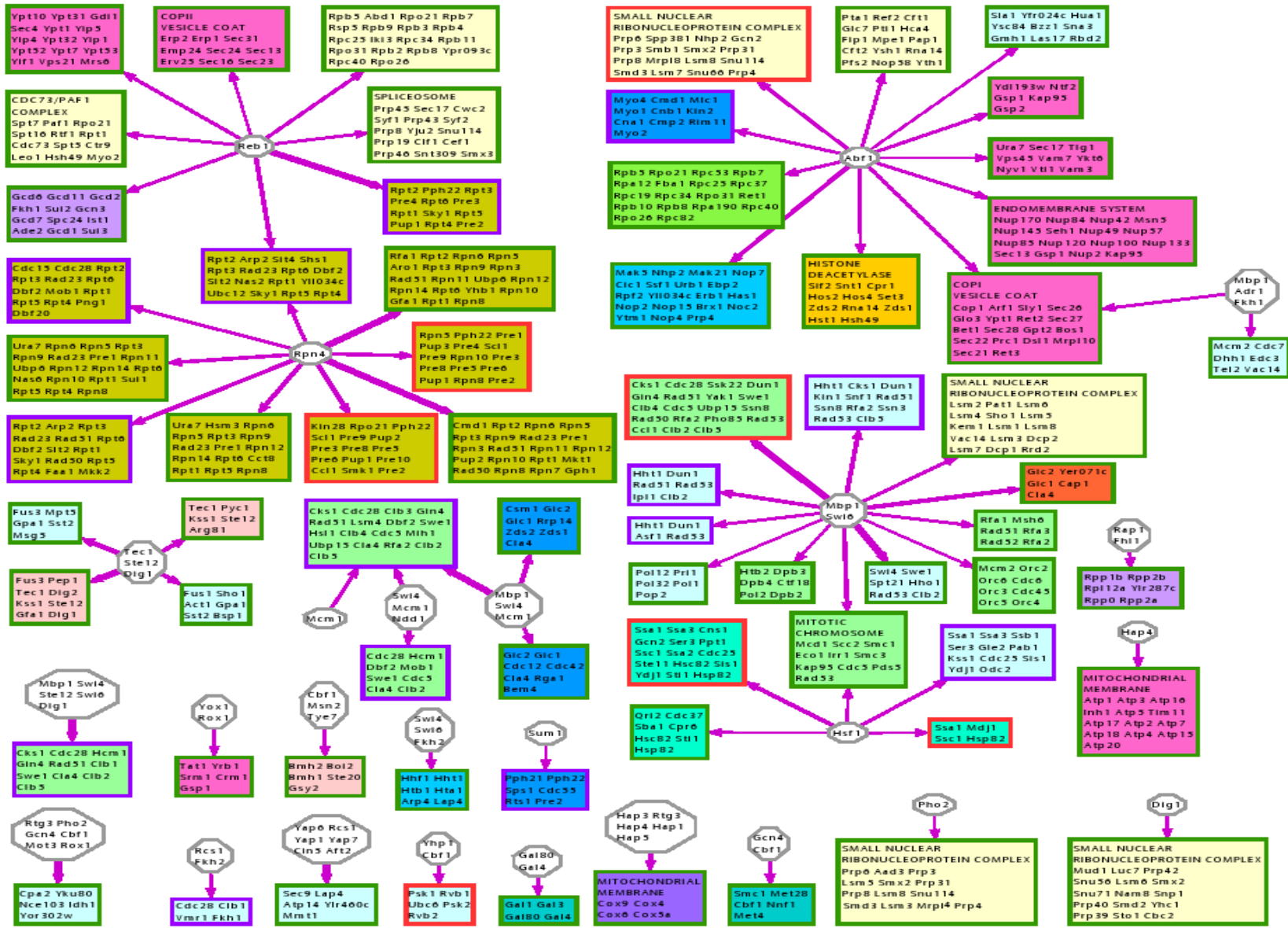
- It is claimed that three levels of conservation in transcription factors were found.
- 1) Both the transcription factor and DNA binding motif were conserved.
- 2) The transcription factor was conserved but the DNA binding motif was not.
- 3) Neither were conserved.
- It is hypothesized that transcriptional regulatory networks diverge more quickly than protein networks.

# Results: Cellular Machinery and Regulation in Yeast

## GO Cellular Processes

- amino acid deacetylation
- biosynthesis
- budding
- catabolism
- cell biogenesis
- cell communication
- cell cycle
- energy pathways
- nucleic acid metabolism
- protein folding
- regulation of cell shape & size
- small molecule metabolism
- transcription
- transport
- not enriched

- found using yeast data
- found using yeast and fly data
- found in both searches
- TF
- protein-DNA interaction



# Issues and Concerns

- Lack of detail in the paper and supplementary information causes difficulty in evaluation of the methods and results.
- Cannot evaluate algorithm as it was not included in either paper or supplementary information.
- Detailed information on the methods used to extend the technique across species is suspiciously absent.

# References

- [1] *Conserved Patterns of Protein Interaction in Multiple Species.* R. Sharan, S. Suthram, R. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. Karp and T. Ideker. *Journal: Proc. Natl. Acad. Sci. USA* 102, pp. 1974-1979, 2005.
- [2] *Transcriptional regulatory code of a eukaryotic genome.* C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, and R.A. Young. *Nature* 431, 99 - 104 (2004).