Predicting Co-Complexed Protein Pairs Using Genomic and Proteomic Data Integration

Lan V. Zhang[1], Sharyl L. Wong[1], Oliver D. King[1], Frederick P. Roth[1]

Keywords: protein-protein interaction, protein complex, decision tree, data integration, machine learning

1 Introduction.

Identifying all protein-protein interactions in an organism is a major objective of proteomics. A related goal is to know which protein pairs are present in the same protein complex. High-throughput methods such as yeast two-hybrid (Y2H) and affinity purification coupled with mass spectrometry (APMS) have been used to detect interacting proteins on a genomic scale [1-4]. However, both Y2H and APMS methods have substantial false-positive rates. Aside from high-throughput interaction screens, other gene- or protein-pair characteristics may also be informative of physical interaction. Therefore it is desirable to integrate multiple datasets and utilize their different predictive value for more accurate prediction of co-complexed relationship.

2 Results.

Using a probabilistic decision tree approach, we integrated high-throughput protein interaction data with other gene- and protein-pair characteristics to predict co-complexed protein (CCP) pairs. Our predictions proved more sensitive and specific than predictions based on Y2H or APMS methods alone or in combination (Figure 1). Among the top predictions not annotated as CCPs in our reference set of protein complexes (obtained from the MIPS complex catalog), a significant fraction were found to physically interact according to a separate database (YPD, Yeast Proteome Database), and the remaining predictions may potentially represent unknown CCPs (Table 1).

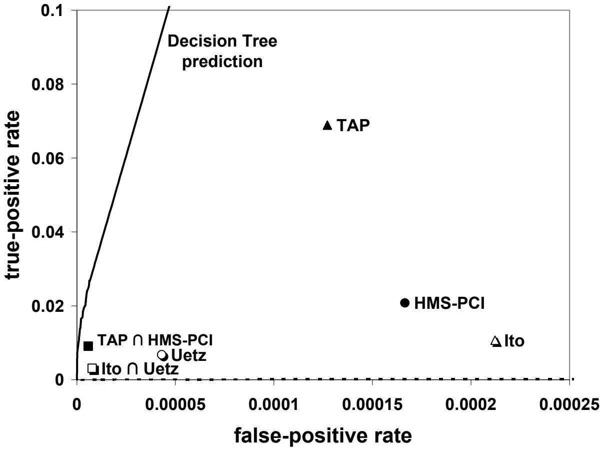

Figure 1: ROC (Receiver Operating Characteristic) curve of decision tree predictions in comparison with four high-throughput datasets (two YPD studies: Ito *et al.* and Uetz *et al.*, and two APMS studies: Gavin *et al.* and Ho *et al.*), as well as their simple combinations (intersection of the two Y2H studies and intersection of the two APMS studies). Solid line represents decision tree predictions, while dotted line represents random predictions.

[1] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, MA 02115, USA. E-mail: `fritz_roth@hms.harvard.edu`

Rank	Protein 1	Protein 2	Score	YPD Complex Annotation
1	Rpl40Bp	Rps31p	0.943	
2	Rps31p	Rpl40Ap	0.938	
3	Smc1p	Smc3p	0.864	Cohesin
4	Gpt2p	Sec28p	0.857	
5	Pwp2p	Utp13p	0.844	Small subunit processome
5	Sgn1p	Pub1p	0.844	
7	Rdh54p	Rad5p	0.833	
7	Arp3p	Rvs167p	0.833	
7	Arp3p	Srv2p	0.833	
10	Spt5p	Rpb3p	0.800	Paf1p complex
10	Spt5p	Rpo21p	0.800	Paf1p complex
12	Pwp2p	Dip2p	0.776	Small subunit processome
12	Pwp2p	Ylr409C	0.776	
12	Sap190p	Sap155p	0.776	
12	Sap190p	Sap185p	0.776	
12	Pph21p	Pph22p	0.776	
12	Nop7p	Fpr4p	0.776	
12	Sap185p	Sap155p	0.776	
12	Sik1p	Cbf5p	0.776	
12	Nop2p	Ebp2p	0.776	Pre-60S ribosomal particle
12	Rpa135p	Ret1p	0.776	
22	Pwp2p	Asc1p	0.750	
22	Drs1p	Spb4p	0.750	
24	Rsm10p	Mrps5p	0.744	Mrp4p-associated complex (mitochondrial ribosome)
24	Mtr3p	Rrp45p	0.744	Exosome 3'-5' exoribonuclease complex
24	Rrp40p	Rrp46p	0.744	Exosome 3'-5' exoribonuclease complex
24	Rrp40p	Ski6p	0.744	Exosome 3'-5' exoribonuclease complex

Table 1: Top 27 predictions that are not annotated as CCPs in the reference set

3 Conclusion.

We demonstrated that the probabilistic decision tree approach can be successfully used to predict co-complexed protein (CCP) pairs from other gene- or protein-pair characteristics. Our top-scoring CCP predictions provide testable hypotheses for experimental validation.

References

[1] Ho, Y., Gruhler A., Heilbut A., Bader G.D., *et al.*, 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature,* 415:180-183.

[2] Ito, T., Chiba T., Ozawa R., Yoshida M., *et al.*, 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences USA,* 98:4569-4574.

[3] Uetz, P., Giot L., Cagney G., Mansfield T.A., *et al.*, 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature,* 403:623-627.

[4] Gavin, A.C., Bosche M., Krause R., Grandi P., *et al.*, 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,* 415:141-147.