Better Modules in Protein-Protein Interaction Networks

Saket Navlakha, Michael C. Schatz, and Carl Kingsford*
Dept. of Computer Science, Center for Bioinformatics and Computational Biology,
Institute for Advanced Computer Studies, University of Maryland, College Park
* Corresponding author: carlk@cs.umd.edu

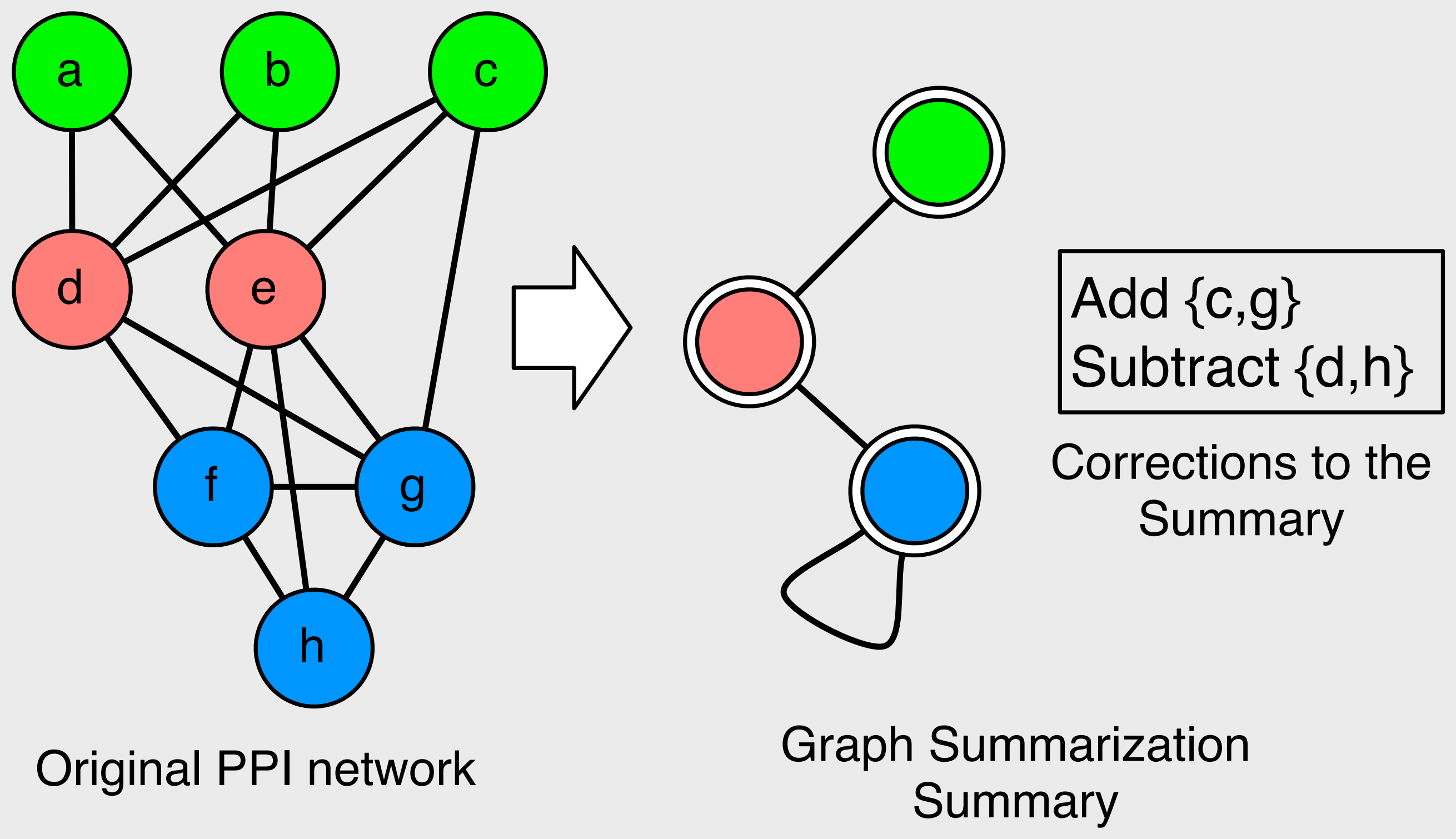
Module-assisted Prediction

A technique called Graph Summarization (GS, ref. 2) can be used to partition protein-protein interaction networks to reveal modules that are more biologically relevant than the clusters produced by other graph partitioning techniques. We apply GS to predict Gene Ontology annotations of biological process for proteins of unknown annotations. We also apply it to detecting membership in protein complexes, as annotated in the MIPS catalog. GS outperforms other approaches such MCODE, MCL and modularity.

Graph Summarization Example

GS produces a compressed ***summary graph***, with nodes combined into supernodes and superedges representing bicliques, and a ***list of corrections*** to the summary.

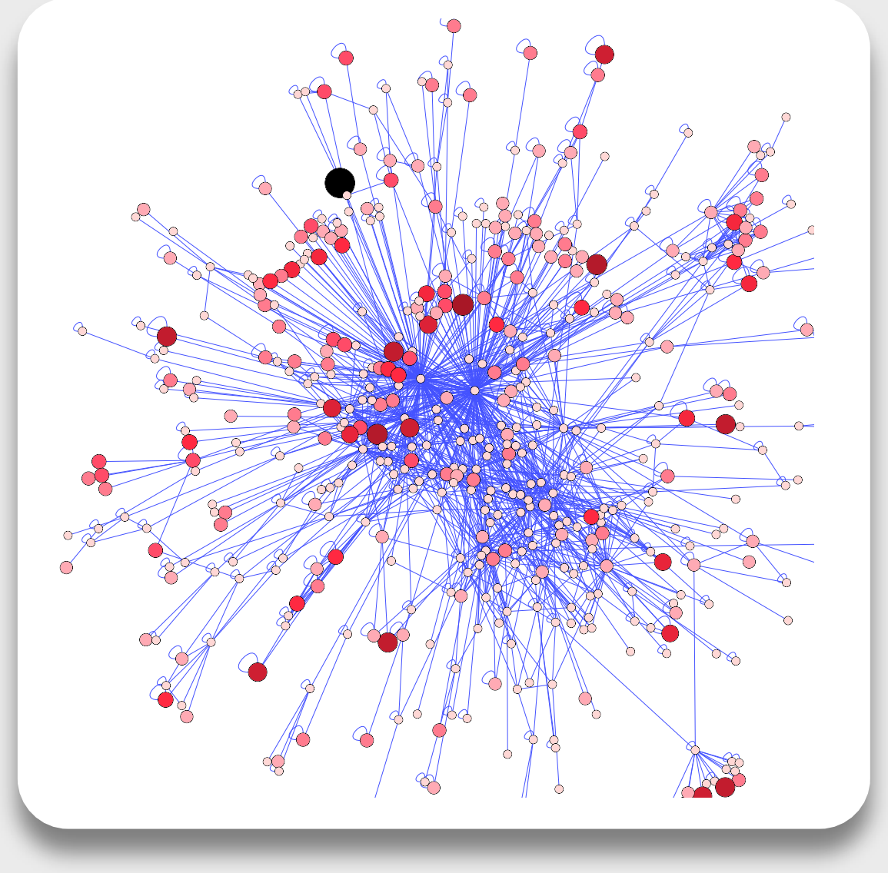
Goal: minimize # edges in the summary plus # of corrections.



Corrections to the Summary

Original PPI network

Graph Summarization Summary



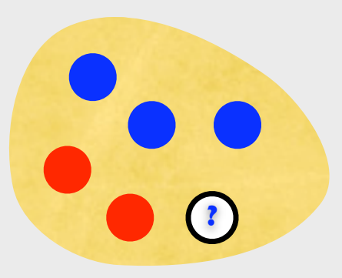
GS of yeast PPI network.
Before: 5,492 proteins
After: 1632 supernodes.

Transferring Annotations

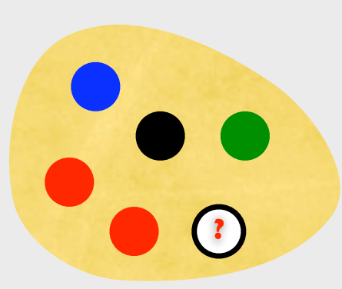
We decompose the PPI network into modules using several graph clustering approaches: MCODE [4], MCL [3], modularity [5], and Graph Summarization [1,2].

We transfer annotations within these modules using one of three standard transfer techniques.Performance is assessed with leave-one-out cross validation.

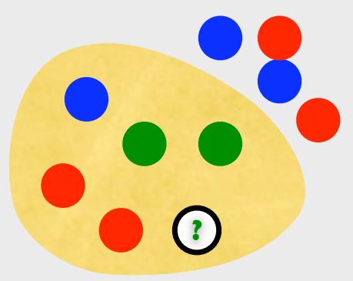
Majority: transfer if > 50% annotated proteins have the annotation.



Plurality: transfer the most common annotation(s).



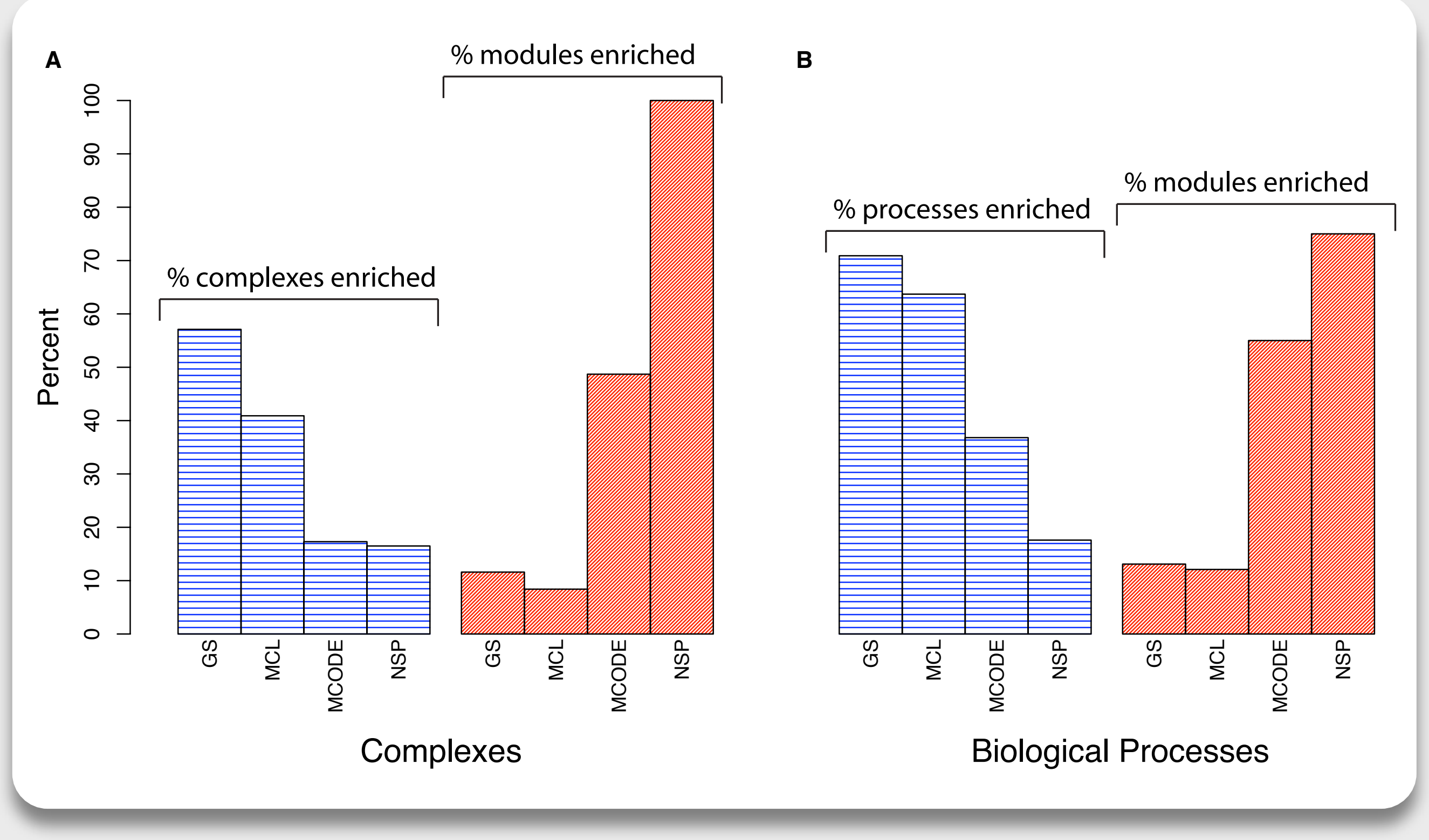
Hypergeometric: transfer the statistically enriched annotations.



Results: Module Quality

GS modules cover more
At right: the % of annotations statistically enriched in at least 1 module (**blue**) and the % of modules that are enriched for at least one annotation (**red**). A wider variety of annotations are enriched in some GS module.

Modularity creates only 8 modules and MCODE clusters only 6% of the network. A larger % of their modules are enriched for some annotation, but this is a poor indicator of predictive performance.



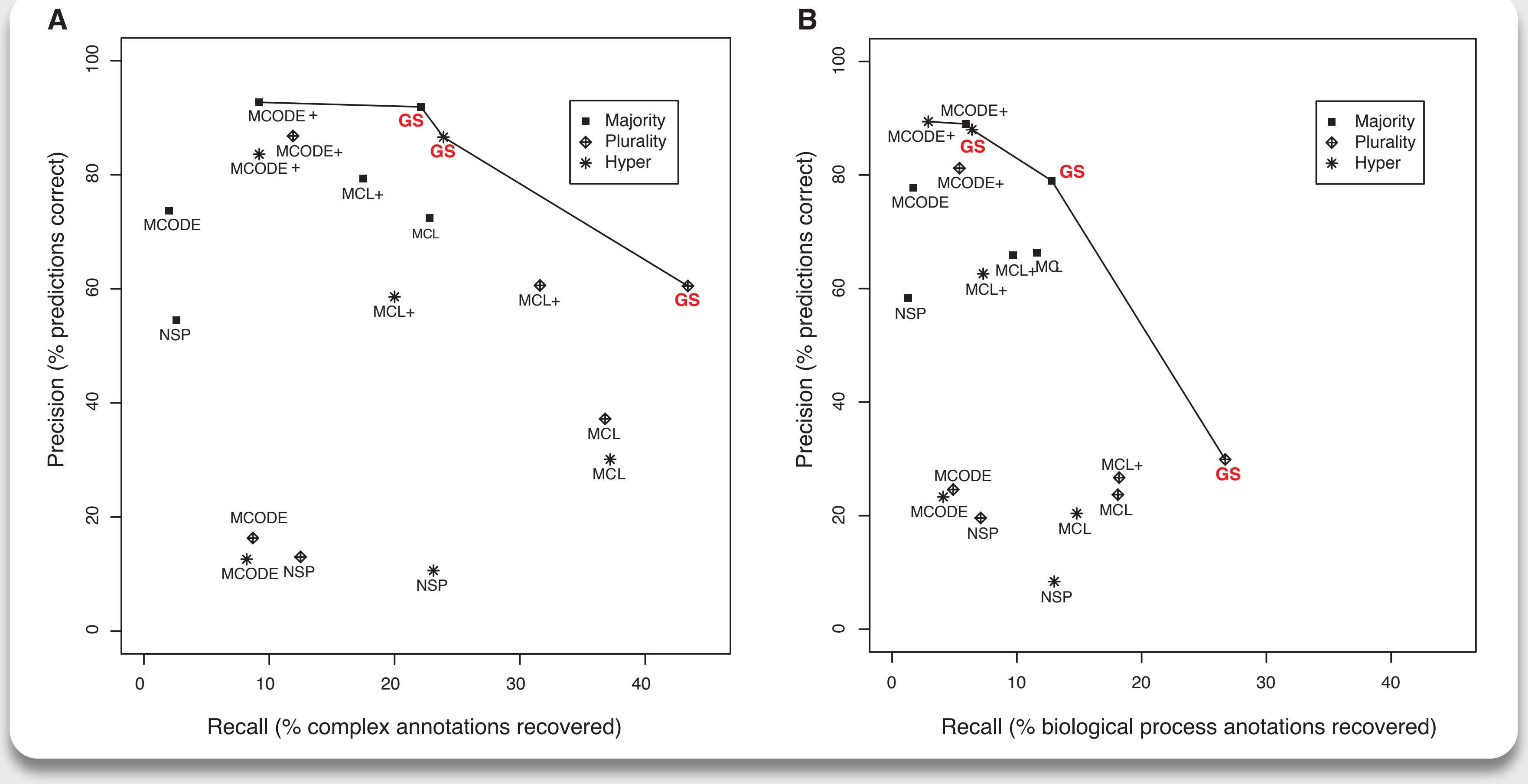
Results: Predicting New Annotations

Interaction network
The PPI network for *Saccharomyces cerevisiae* was downloaded from IntAct (5,492 proteins, 40,332 edges). Similar results are obtained if only edges with ≥ 2 supporting publications are included.

Known annotations
GO biological process annotations for the terms selected in [7] were taken from the SGD. Protein complex annotations were taken from the lowest level of the MIPS complex hierarchy.

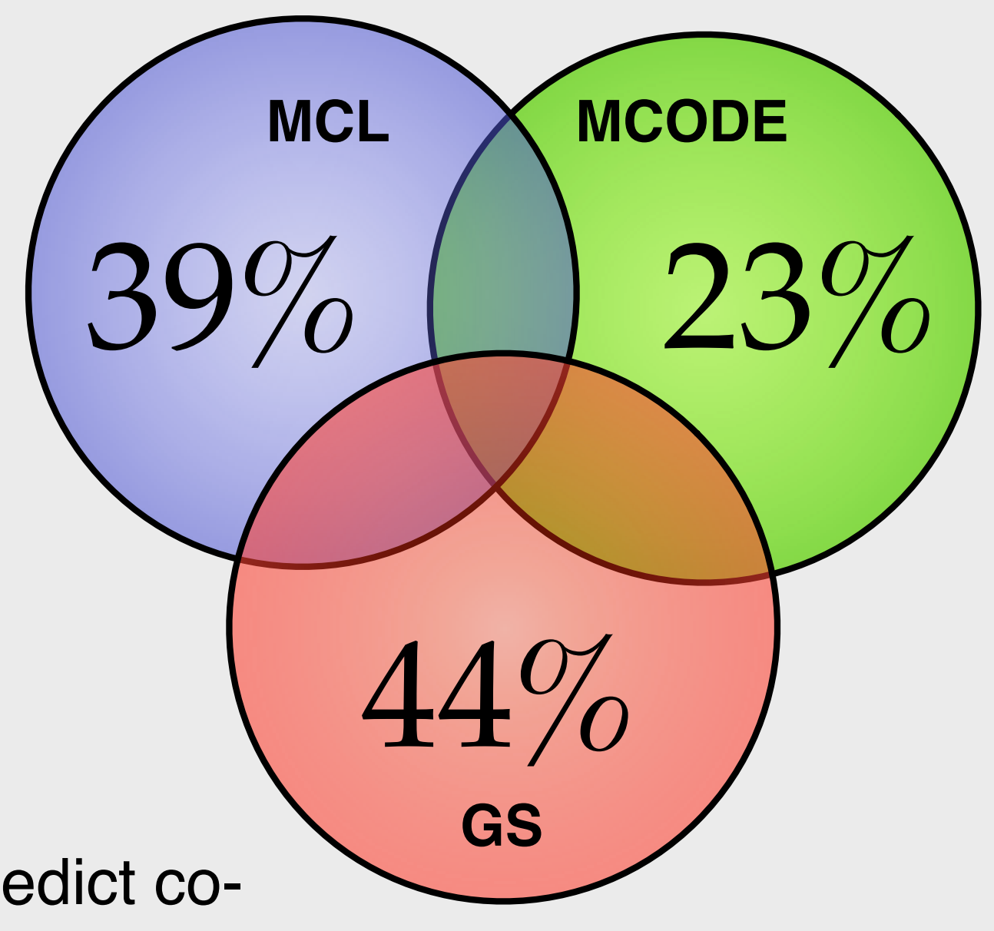
GS outperforms MCODE, modularity, MCL
Precision and recall on leave-one-out cross validation is shown using 3 different module-detection methods and 3 different annotation transfer schemes. MCL and MCODE have several parameters. Results are reported for both their default values and parameters tuned to maximize precision (marked with a +). GS has no parameters to tune.

Graph Summarization is always Pareto optimal: no other method dominates it in both precision and recall. Though MCODE can be tuned to achieve precision equal to GS, it has lower recall.



Results: Comparing Predictions

Methods complement each other
A large % of correct predictions are unique to each method using the majority transfer rule and MIPS complex annotations. Similar results hold for biological process annotations and other transfer rules.



Predicting co-complexed pairs
The GS correction list can also be used to predict co-complexed pairs of proteins, with generally higher precision than the method of completing defective cliques [6].

Conclusions

GS produces better biological modules
GS groups together proteins with similar interaction partners, leading to better complex and biological process modules than MCODE, MCL or modularity.

MDL principle
GS is mathematically well-founded and has no parameters to tune.

GS is versatile
It generalizes bipartite cores, cliques, stars. It also explicitly handles noise via a corrections list.

References

[1] S. Navlakha, M.C.Schatz, C. Kingsford. Revealing Biological Modules via Graph Summarization. *J. Comp. Biol.* **16**(2), *in press.* (Presented at RECOMB Systems Biology, 2008.)

[2] S. Navlakha, R. Rastogi, N. Shrivastava. Graph summarization with bounded error. In *Proc. of SIGMOD 2008*, pages 419–432.

[3] S. van Dongen. (2000) A cluster algorithm for graphs, Tech Report, CWI.

[4] G. D. Bader and C. W. V. Hogue. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**:2.

[5] M. E. J. Newman. (2000) Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103:8577–8582.

[6] Yu, H., Paccanaro, A., Trifonov, V., *et al*. (2006) Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* **22**:823–829.

[7] C. L. Myers, D. R. Barrett, M. A. Hibbs, C. Huttenhower, and O. G. Troyanskaya. (2006) Finding function: evaluation methods for functional genomic data. *BMC Genomics*, **7**:187.