Improvements in protein function prediction using confidence in protein interactions

Kathryn Doroschak[1], Thomas Schaffner[2], Benjamin Hescott[2], Lenore Cowen[2]

1 Department of Computer Science, University of Minnesota, Minneapolis, MN, U.S.
2 Department of Computer Science, Tufts University, Medford, MA, U.S.

Introduction

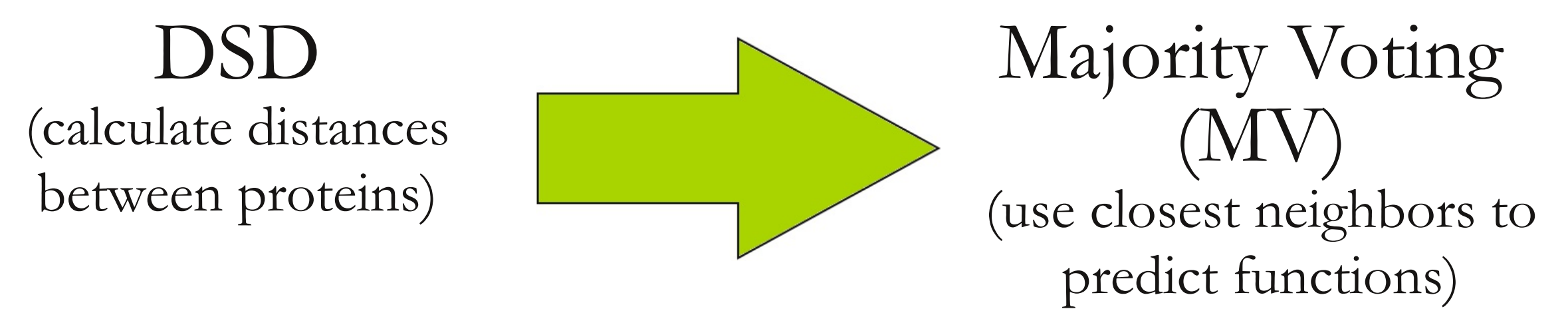
Characterizing protein function is a crucial part of understanding biological systems. Most computational methods for function prediction are limited by:

a) Indistinct functional neighborhoods from using shortest paths in "small world" protein-protein interaction (PPI) networks, and
b) Data quality issues inherent in PPI databases.

The first problem has been addressed by combining majority voting and diffusion state distance (DSD). DSD is a fairly new metric that leverages graph diffusion to better capture distances within a network.

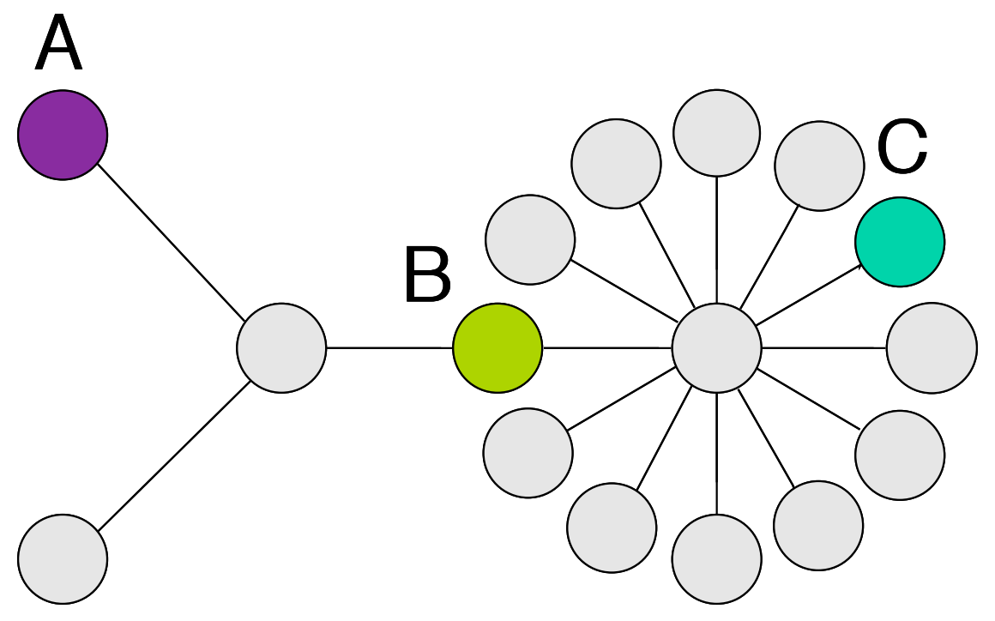
Here we focus on the second limitation, addressing data quality. We do so by introducing confidence to function prediction, assigning scores to each edge in the PPI network.

Function Prediction using DSD + MV



DSD
(calculate distances between proteins)

Majority Voting (MV)
(use closest neighbors to predict functions)



With regular distance:
d (A,B) = d (B,C)

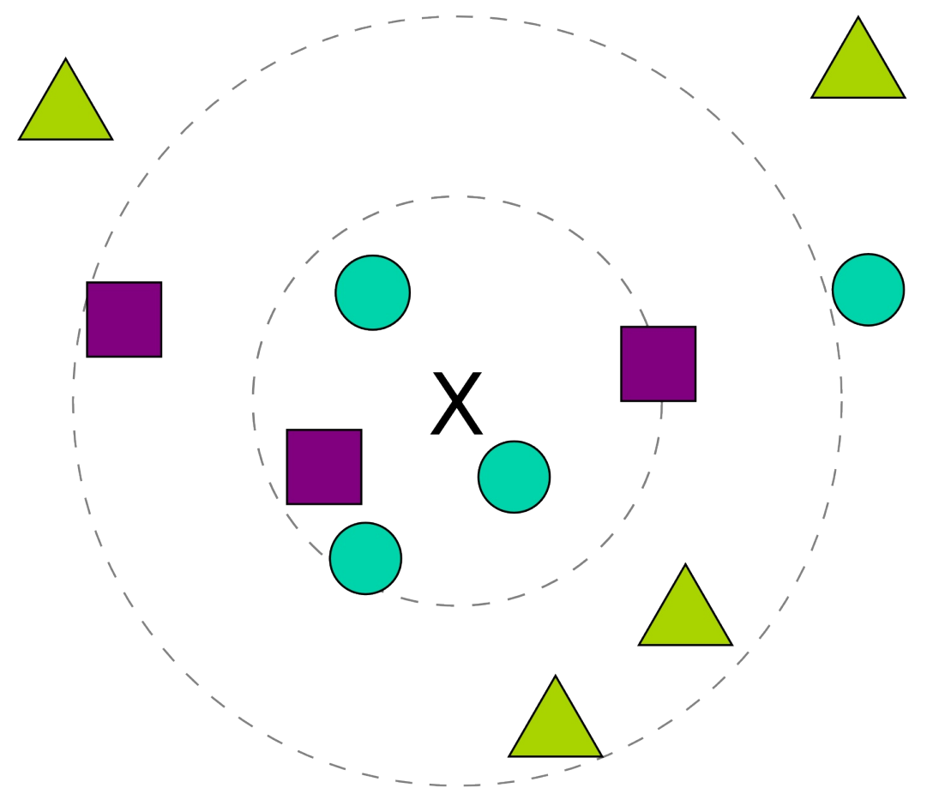
With DSD:
d (A,B) < d (B,C)

How DSD works

Many PPIs contain hubs (high degree nodes), which create a small world property in the network. This property hurts algorithms that rely on shortest paths, e.g. Majority Vote, because most paths are similarly short. To offset this, DSD uses graph diffusion to distinguish paths via hubs from paths via lower degree nodes [1].

How Majority Vote works

Majority Vote [2] can be thought of as "guilt by association" -- protein functions are predicted based on neighbors' functions. To do this, we allocate votes to each neighbor, either unweighted (equal vote for each neighbor) or weighted. Each neighbor votes for its own function. The node is assigned the function with the most votes.



X's closest 5 neighbors are:
(close) ● ■ ● ● ■ (far)

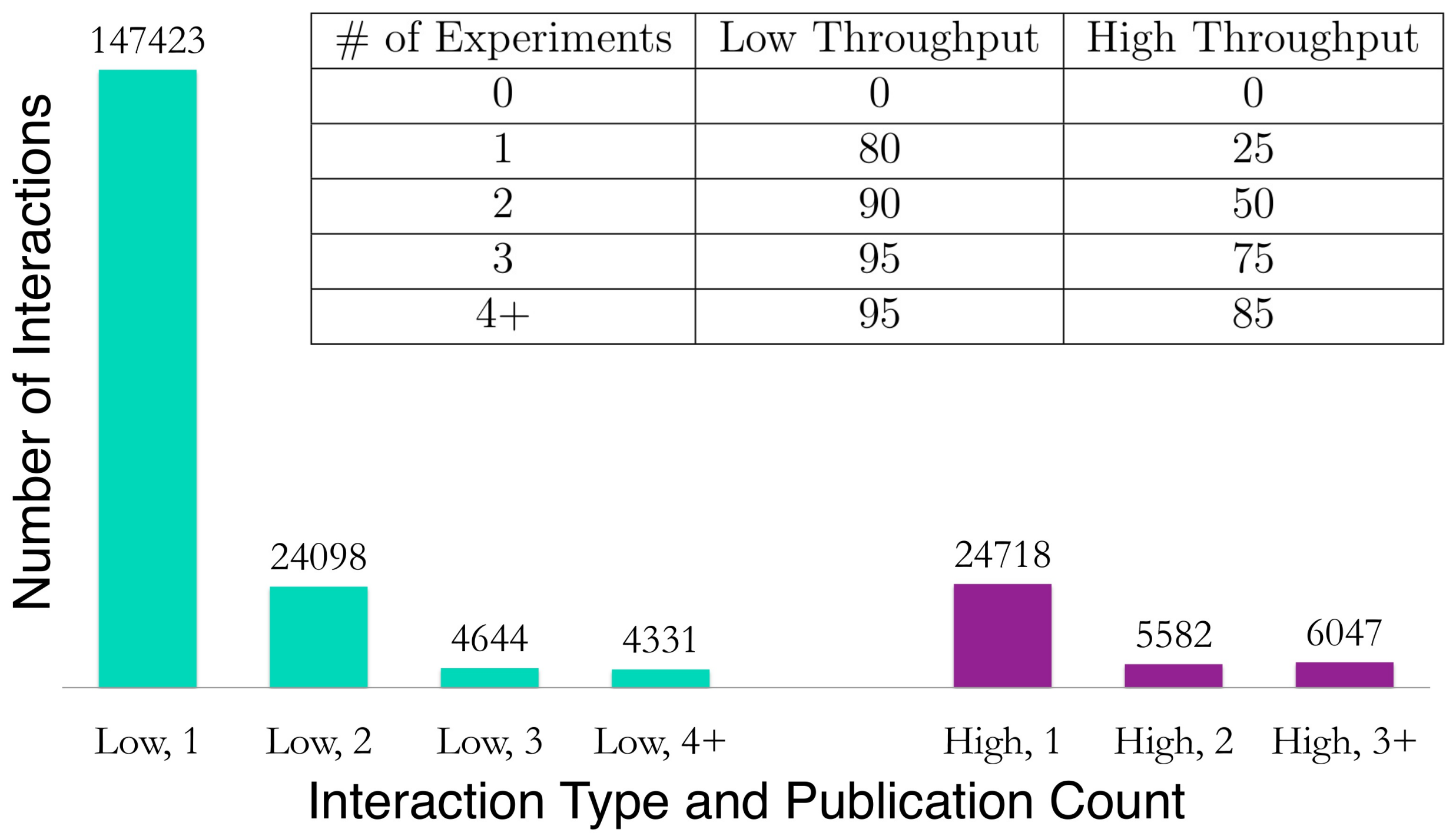
X is probably ●
based on closeness & number

Data Sources

Protein-Protein Interactions:	BioGRID S. cerevisiae v3.2.101 [3]
Functional Annotations:	MIPS FunCat v2.1 [4]
Confidence Scores:	Literature-based (see below)

Generating Confidence Scores

Confidence scores are derived from the volume and type of experiments conducted for each edge. Multiple publications for an edge serve as verification for that edge. Additionally, high-throughput experiments tend to be less reliable due to their tendency to produce false positives. The cutoff for high and low throughput is 100 interactions [5].



# of Experiments	Low Throughput	High Throughput
0	0	0
1	80	25
2	90	50
3	95	75
4+	95	85

Publication counts are derived from S. cerevisiae protein interactions found in the BioGRID database v3.2.101 [3]. There were originally 324,743 interactions, filtered down to 216,842 (shown here) by removing non-ORF proteins and duplicates.

Applying Confidence Scores

Two steps in the pipeline, two ways to add confidence:

1. Majority voting with confidence as weights
2. DSD with confidence as probability in random walks

MV + Confidence

Simply allocate votes by confidence. The more confident we are that two nodes are connected, the more votes the pair gets.

DSD + Confidence

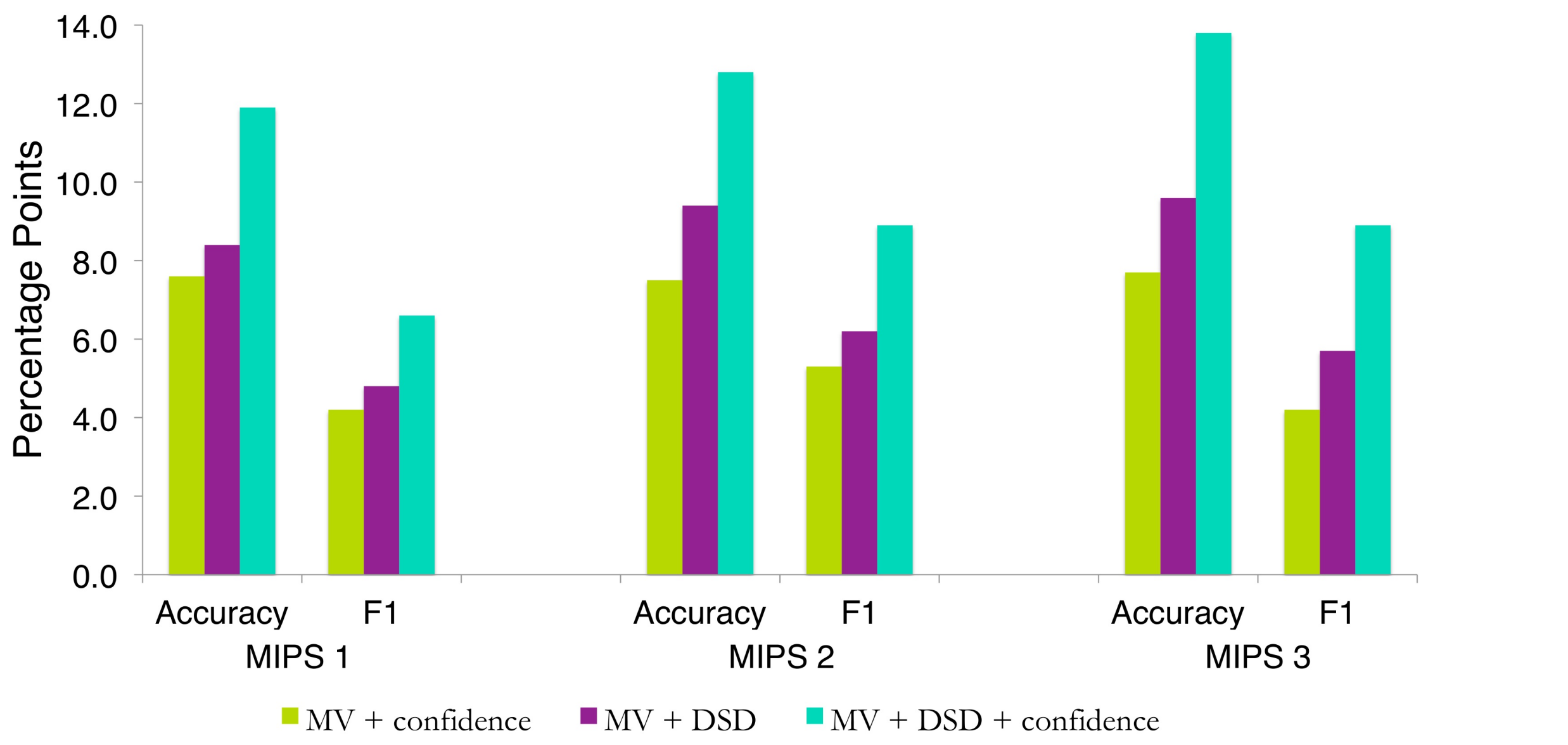
Weight the random walk within DSD so that higher confidence edges are taken more frequently. We normalize the confidence with respect to the other edges coming out of a given node.

Results

Adding confidence to majority voting helps significantly, but DSD still performs even better, both with and without confidence. DSD with confidence performs best by far, likely because it addresses the issues of both shortest path distribution and data quality.

	MIPS 1		MIPS 2		MIPS 3	
	Accuracy	F1 score	Accuracy	F1	Accuracy	F1
Majority Vote (MV)	58.0	45.0	45.3	32.7	40.7	30.4
MV, confidence as weights	65.6	49.2	52.8	38.0	48.4	34.6
MV, weighted original DSD	66.4	49.8	54.7	38.9	50.3	36.1
MV, weighted DSD with conf	69.9	51.6	58.1	41.6	54.5	39.3

Summary of MV performance improvements using various confidence techniques, 2-fold cross-validation, and 10 voting neighbors.



MV performance increases relative to the baseline (MV with no DSD or confidence).

Summary

Confidence significantly improves both DSD and ordinary-distance majority voting (MV) by accounting for data unreliability.

- Accuracy improved 11.9 pp from original MV to MV using DSD with confidence.
- Accuracy improved 3.5 pp from MV using original DSD to MV using DSD with confidence.

Our confidence techniques can be easily integrated with other function prediction methods; DSD with confidence can be used with any shortest-path function prediction method by replacing ordinary distance.

References

[1] Cao M, Zhang H, Daniels N, Park J, Crovella M, Cowen L, Hescott B (2013) Going the distance for protein function prediction. In review.
[2] Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. Nature Biotechnology 18: 1257-1261
[3] Stark C, Breitkreutz B, Reguly T, Boucher L, Breitkreutz A, et al. (2006) Biogrid: a general repository for interaction datasets. Nucleic Acids Res 34: D535-D539
[4] Ruepp A, Zollner A, Maier D, Albermann K, Hani J, et al. (2004) The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic acids research 32: 5539-55
[5] Chen Y, Rajagopala S, Stellberger T, Uetz P (2010) Exhaustive benchmarking of the yeast two-hybrid system. Nature Methods 7:667-668