The literature curated microbial binary protein-protein interactions

Seesandra V Rajagopala[1], Johannes Goll[1], ND Deve Gowda[2], Kumar C Sunil[2], Björn Titz[3], Arnab Mukherjee[1], Sharmilla S Mary[2], Naresh Raviswaran[2], Chetan S Poojari[2], Svetlana Shtivelband[1], Stephen M Blazie[1], Julia Hoffman, Peter Uetz[1,§]

1 The J. Craig Venter Institute, Rockville, 20850 MD, USA
2 Indgen Life Technologies, Bangalore - 560 004 (Karnataka), India
3 Institute for Genetics, Forschungszentrum Karlsruhe, Karlsruhe, Germany

§ corresponding author:
9712 Medical Center Drive
Rockville, Maryland, USA

Phone: +1-301-795-7589
Fax: +1-301-294-3142

e-mail: uetz@jcvi.org

ABSTRACT

Prokaryotic protein-protein interactions are underrepresented in currently available databases. Here we describe a reference dataset (MPI-LIT) focusing on microbial binary protein-protein interactions and associated experimental evidence that we have manually curated from 36,852 scientific abstracts and full texts. The MPI-LIT dataset comprises 1,240 experimental descriptions that describe a non-redundant set of 748 interactions of which 659 (88%) are not reported in public databases. To estimate the curation quality, we compared our dataset with a union of manually curated microbial data from IntAct, DIP, BIND and MINT. Among common abstracts, we achieve a sensitivity of up to 66% for interactions and 77% for PSI-MI annotations of experimental methods. For both, we estimate the false positive rate to be less then 4%. Compared to the other datasets, MPI-LIT has the lowest fraction of interaction experiments per abstract (1.5) and the highest coverage of strains (92) and scientific articles (814). As a case study, we evaluated interaction confidence estimation methods which are implemented in STRING and show that most of them descriminate well between high-throughput data from pull down studies and our high-quality reference set. The microbial reference set is accessible at http://www.jcvi.org/mpidb/?dbsource=MPI-LIT

1. INTRODUCTION

Microbes represent the vast majority of completely sequenced genomes (1). Recent metagenomics projects have fortified this dominance even more with bacteria representing about 90% of all sequences in the global ocean sampling dataset (2). Clearly, an understanding of protein function and physiology requires a detailed understanding of their interactions with both other proteins and small molecules.

Surprisingly, compared to eukaryotes the protein interactions of microbial species are largely unexplored: while the microbial interaction database (3) reports ~20,000 microbial interactions, in contrast general interaction databases, such as IntAct report on the order of 100,000 eukaryotic interactions. Although some interactions, especially from high-throughput studies, are reported in public databases, the majority remain hidden in the primary scientific literature.

Due to the ambiguity of free text, especially of protein names, species/strains and experimental methods, natural language processing algorithms are unable to reliably extract all the interacting proteins and associated data automatically. Thus manual curation remains key. Manual curation of protein interactions poses a number of problems, including curation inconsistency, myriad possible levels of annotation details, and a large volume of text to be analysed. Such a manual curation of protein-protein and genetic interactions have been carried out for *Saccharomyces cerevisiae* (4) and human (5).

Here we report a manual curated dataset (MPI-LIT) that we have extracted from 814 publications focusing on microbial species. This dataset comprises 1,240 experimental descriptions (PubMed IDs and Experimental Methods) that link 940 proteins by 748 binary protein-protein interactions (Table 1). While our dataset does not appear to be large, it is the largest manually curated dataset published so far for microbial protein interactions and thus will serve as a true positive reference set, It can be used to evaluate interaction confidence estimation methods, to estimate the confidence of high-thorughput microbial interaction datasets (e.g. generated by yeast-two-hybrid or co-affinity purification screens) and to train automatic literature mining algorithms

2. Method

2.1 Literature Curation Strategy

Phase I: PubMed Search

We started by manually curating interactions from the primary literature, similar to previous efforts for yeast (4) and human (5). A PubMed search for "Bacteria" OR "*Escherichia coli*" OR "*Salmonella*" OR "*Bacillus subtilis*" OR "*Pseudomonas*" AND (interaction OR interact OR interacts OR bind OR binds), yielded 36,852 articles as of August 14, 2006 that potentially contain microbial protein interaction data.

Phase II: Text Analysis

From these articles, interacting protein pairs, the respective microbial species, and the experimental method were extracted from the abstract. During this phase, we were able to identify 4,046 protein names and 2,303 method descriptions (Table 1)

Phase III: Protein ID and ontology mapping

As proteins were usually represented by their common names in the publications, we set up an automated protein identification pipeline. We systematically screened microbial versions of the UniProtKB/Swiss-Prot (6) and Biothesaurus (7) databases to match proteins to the latest stable UniProt accessions based on their common names and species. Using this pipeline, we were able to uniquely identify 1,254 proteins (30%).

For 2,796 proteins, we could not identify a unique UniProt accession: most commonly the strain could not be uniquely identified. For example, a UniProt search for *Escherichia coli* and protein/gene name RecA results in 11 different UniProt entries comprising 10 sequenced strains. Thus, we manually mapped 1,790 (44%) common names species/strain pairs to the latest stable UniProt accessions.

1,006 (25%) protein names could not be matched at all and these proteins were removed from the final curated dataset. Such deleted entries include non-protein entities that were initially identified as proteins such as small-molecules or protein complexes (“RNA polymerase”), non-microbial proteins, and misspelled common names. In a related effort, we manually mapped the free-text experimental methods descriptions onto experimental methods defined by the PSI-MI ontology.

Interaction vs. Experiment

Interactions are defined as unique pairs of latest stable UniProt accessions. An interaction experiment is defined by an interaction, an experimental method (PSI-MI), and a publication (PubMed ID). An interaction can be part of more than one experiment whenever such an interaction is reported by a different method and/or different study.

2.2 Datasets

The MPI-UNION dataset has been downloaded from the MPIDB database (3) and is a microbial subset of IntAct, DIP, BIND and MINT interactions filtered for binary interactions. The *E. coli* K12 STRING scores were downloaded from the STRING database (version 7.1). *E. coli* K12 GO annotations were collected from the Gene Ontology Annotation (GOA) Database (04/01/08).

2.3 GO Term enrichments

We used the topGO R package to detect significantly enriched GO terms in MPI-LIT “Improved scoring of functional group from gene expression data by decorrelating GO graph structures. The classic algorithm based on gene count using the elim

methods was applied to minimize the false-positive rate). The test statistic is based on Fisher's exact test.

3. Results

3.1.1 Abstract overlap

Our PubMed search retrieved 36,852 abstracts of which only a tiny fraction (203, 0.6 %) are reported in the MPI-UNION dataset. Our PubMed search missed 299 abstracts that are reported to contain microbial interactions in MPI-UNION. This indicates that our query missed a fraction of relevant abstracts (Fig.1A). This might be due to the fact that the search was limited to abstracts. Articles which describe interactions in the full text, which is true for most of the medium and high-throughput PPI studies, would have been missed. During the Phase II and Phase III curation process we were able to remove non relevant abstracts and those for which we could not uniquely identify the interacting proteins (Table 1). After Phase III, the fraction of known abstracts in our literature set increased 18 fold, and we were able to collect 738 new abstracts (Fig. 1B) describing microbial protein-protein interactions.

3.1.2 Interaction curation

To estimate the curation fidelity, we compared the 76 MPI-LIT articles overlapping with MPI-UNION (Fig. 1B). We estimate the sensitivity of interaction detection by comparing the number of interactions overlapping between both datasets. The MPI-LIT curation efforts achieve a sensitivity of 66% when using MPI-UNION as a reference set (note that strains were merged into species and the common names were compared. Vice versa, the MPI-UNION obtained the same sensitivity when using MPI-LIT as a reference. This indicates that independent literature curation efforts, MPI-LIT and MPI-UNION, miss an estimated 34% of interactions (false negatives) (Fig. 1C). This is in line with a previous manual curation study conducted by Reguly et al. that reported a false negative rate of 20%, and up 50% depending on the reference set and databases (4). The possible reason for false negatives in MPI-LIT is the Phase II curation was limited to abstracts. If interactions are described in the full text and not mentioned in the abstracts, curators failed to report the interactions.

Interestingly, MPI-LIT and MPI-UNION curated 36 unique interactions for the common set of 76 articles (Fig. 1C). We re-examined these interactions in the primary articles - Out of the 36 unique MPI-LIT interactions which are not reported in MPI-UNION, 4 interactions turned out to be false positives. The estimated false positive rate for MPI-LIT curtains is 4% (i.e., 4 false positive out of 105 interactions), similar rate (4%) of manual curtain errors was also reported in the privies study (4). Vice versa, out of the 31 unique MPI-UNION interactions, one interaction turned out to be false. The estimated false positive rate for MPI-UNION curtains is 1% (i.e., 1 false positive out of 105 interactions). The nature of false positives in the MPI-LIT dataset, include curator typo errors (2%) and wrong protein ID mapping (2%). These wrong entries were removed in the final dataset.

3.1.3 Method curation

Based on 76 articles that are common to both MPI-LIT and MPI-UNION, 69 common interactions have been curated from these articles (Fig. 1C), ignoring strain variations. Of these, 52 interactions were identical, including the strains. Each protein interaction can have more then one experimental method if the same interaction is reported from more then one study or from a different experiment. We estimate the sensitivity of experimental method annotation by comparing the PSI-MI terms of these 52 interactions

(Fig. 1D). Each pair of interacting proteins can have more then one experimental method, if the interaction is verified by more then one experiment or publication. In total 78 experimental methods were curated for the 52 interactions by MPI-LIT and MPI-UNION. We merged all terms at level 1 of the PSI-MI ontology (*biophysical*, *protein complementation assay*, *genetic interference*, *post transcriptional interference*, *biochemical*, *imaging techniques*). For such a merged set, we estimate the average sensitivity to be on the order of 77% for MPI-LIT (see methods) using MPI-UNION as a reference set. Although we assume that we have missed a considerable fraction of experimental descriptions, the curated descriptions are very likely to be true positives. Based on the MPI-UNION reference set, we estimate the false-positive rate to be lower than 0.5%.

3.2 The MPI-LIT dataset

The MPI-LIT dataset covers 1,240 experimental descriptions comprising 784 non-redundant bacterial PPIs of 92 species/strains extracted from 814 abstracts. The coverage of abstracts and species/strains is significantly higher than those compiled by curation efforts represented in the MPI-UNION dataset (Table 2). Notably, the 784 PPIs in MPI-LIT are supported by 1,240 experiments. This indicates that on average an interaction is either confirmed by more than one experimental method and/or by multiple publications (Table 2). Most of the interactions in MPI-LIT are reported for *E. coli* (54%), *Bacillus subtils* (11%) and *Salomella typhimurium* (3%) (Supplementary Table S1).This is in line with our PubMed query which emphasizes these three species.

On average we identified 0.92 interactions per article, indicating these interactions are culled from small-scale studies. Within MPI-UNION, MINT curated an average of 1.25 interactions per article, whereas DIP reported 13, BIND reported 15, and IntAct reported 67. This indicates enrichment of the interactions derived from medium and high-throughput experiments (Fig. 2A).

We wondered whether certain molecular functions, cellular components and biological processes are enriched in the literature curated dataset. . To investigate this, we looked for significantly enriched GO terms in each of the three subontologies. We did this by comparing the frequency of GO terms of genes that are present in a MPI-LIT subset of *E. coli* K12 interactions with those present in the whole *E. coli* K12 genome (see methods). Table 3 lists the top ten enriched GO terms for the Biological Process subontology. A broad range of processes have been enriched, among them metabolic processes such as “GO:0006457 protein folding”, “GO:0006260 DNA replication”, „GO:0006281 DNA repair”; cellular processes such as “GO:0007049 cell cycle” and “GO:0051301 cell division”; as well as processes involved in response to stimuli like “GO:0006950 response to stress” , ”GO:0006935 chemotaxis” and “GO:0009432 SOS response”. Finally, localization processes such as “GO:0065002 intracellular protein transport across a membrane” were enriched as well. This broad variety of enriched processes indicates that there is no bias towards certain groups of biological processes. Notably, the molecular function “GO:0005515 protein binding” and cellular component “GO:0043234 protein complex” were found to be highly enriched in the other subontologies, reflecting the protein interaction nature of the dataset. A list of all enriched terms and highlighted GO graphs can be found in the supplement (Supplement table S2).

Gene Ontology (GO) terms represented within MPI-LIT dataset. We used an *E. coli* K12 subset of MPI-LIT and the GO terms of *E. coli* K12 genome for the analysis.

Interestingly MPI-LIT dataset represents wide range of functional profiles, and does not show bias towards specific functional class. However, as expected the GO term "protein binding" is significantly overrepresented in the "Molecular Function" GO terms, which is in line with the dataset nature (Table 3). Even though there is no bias towards specific functional groups or pathways, in the "Biological Process" the GO terms mainly stress response, cell division, cells cycle, DNA replication, DNA repair and chemotaxis. These overrepresentations indicate the MPI-LIT dataset has higher coverage of proteins involved in these pathways.

3.3 Application (benchmark the confidence estimation methods)

Several methods have been proposed to estimate the biological relevance or confidence of protein interaction data..To give an example of how MPI-LIT can be used o benchmark such methods, we compared methods that evaluate the degree of co-annotation, co-expression, co-citation, gene-fusion, co-pathway membership, co-occurrence, and gene-neighborhood using our manually curated interaction data and data we obtained from high-throughput experiments.

Most of these measures are pre-computed for a variety of microbial genomes in the STRING database. We used an *E. coli* K12 subset of MPI-LIT (355 interactions, 45%) and compared their STRING values with those of binary interactions from two *E. coli* pull-down experiments (SPOKE model) (8,9).

One sided two sample wilcoxon tests revealed that STRING's gene neighborhood, co-occurrence, experiments and text mining methods scored MPI-LIT interactions significantly better then either of the high-throughput pull-down datasets (Fig. 3, Mann-Whitney U test, $p<0.05$), indicating that such methods are well suited for data quality estimation. In contrast, the gene fusion, gene expression, and pathway neighbor methods were not able to clearly separate the datasets. For example, it is known that co-expressed and pathway neighboring proteins do not necessarily interact and vice versa. Overall, STRING's probalistic combined score discriminates very well between MPI-LIT and inferred interactions from high-thorughput pull-down experiments ($p<0.01$ for either pull-down datasets) underlining STRING's usefulness for interaction confidence estimation.

[We should also provide an estimate of how many interactions remain in the literature un-curated; this may include an estimate of how many interactions do we get from full text as opposed to abstracts. If this number cannot be estimated we should discuss reasons why not and how this problem could be solved. It would be also informative to have a simple diagram with the numbers of interactions curated per year and the corresponding abstracts that contained these PPIs. Maybe there is a trend that helps us to estimate future efforts necessary to curate the literature in a more targetd fashion and more efficiently.]

4. Conclusions

The extreme diversity of microbes and their extraordinary rate of evolution necessitate particular care of rapidly evolving protein interactions. Thus we believe that this reference dataset will be a valuable source of information for the microbiologist. It comprises 1,240 experimental descriptions (PubMed IDs and

Methods) that describe a unique set of 748 interactions of which 659 interactions were not previously reported in public databases (Fig. 2C). All data can be downloaded at http://www.jcvi.org/mpidb/?dbsource=MPI-LIT. For all MPI-LIT interactions we provide a pair of UniProt accession numbers, the experimental method (PSI-MI controlled vocabulary) and the PubMed ID.
The false negative rate for MPI-LIT and also MPI-UNION dataset was 34%, which is much higher then expected for the manual curation. Thus several independent curation efforts are needed to reach full coverage – a fact already noted by Reguly et al. we also conclude that around 30% of interactions are missed when we curated interactions from the abstracts.

Considering the fact that small-scale protein interaction studies are usually believed to be of higher quality than high-throughput data, the MPI-LIT dataset can be used as training set for PPI literature mining algorithms, as a true positive reference set for PPI confidence estimations, to predict interlogs for other species, and for integrative bioinformatics analysis. Based on our results, we plan to focus our curation efforts on the Journal of Bacteriology and Molecular Microbiology to increase the coverage of this reference dataset. We are planning to coordinate our literature curation efforts under the umbrella of the IMEx consortium. .

ACKNOWLEDGEMENTS
We thank Harihara Dixit, Srikanta Dixit, Lakshmi C Madhusudana and Guruprasad B Rajegowda from Indgen Life Technologies Mysore, for their support with literature curation.

Figure legends

Fig. 1
Literature curation fidelity. (A) Abstract overlap between MPI-LIT PubMed query retrieved abstracts and the “MPI-UNION” datasets represented abstracts. (B) The abstract overlap between MPI-LIT curated abstracts containing PPIs and MPI-UNION dataset. (C) Assessment of the sensitivity and false negative rate of interaction curation. Here we compare the curation fidelity of MPI-LIT and MPI-UNION overlapping abstracts. (D) Assessment of the sensitivity and false negative rate of method curation by comparing MPI-LIT and MPI-UNION overlapping interactions.

Fig. 2
Characterization of the MPI-LIT dataset. (A) Number of interactions per publication in MPI-LIT, MINT, DIP, BIND and IntAct datasets. (B) Graph showing the addition of the microbial protein interaction data over time. (C). Overlapping interactins between different dataset, MPI-UNION dataset is is a microbial subset of IntAct, DIP, BIND and MINT interactions filtered for binary interactions.
Fig. 3
(A) Assessment of interaction confidence methods implemented in the STRING database. For all methods, except for gene fusion and KEGG pathway neighbors, confidence estimates for MPI-LIT were significantly greater then those of the pull down datasets (Mann-Whitney U test, $p<0.05$).
(B) Distribution of GO terms for protein involved in physical interactions. Here we compared GO biological process terms for *E. coli* proteins in MPI-LIT and high-

throughput complex purification datasets of *E. coli* (8,9). The mean shared annotation is significantly higher for MPI-LIT than for the Butland and Arifuzzaman datasets.

REFERENCES

1. Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K. and White, O. (2001) The Comprehensive Microbial Resource. *Nucleic Acids Res*, **29**, 123-125.
2. Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W. *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biol*, **5**, e16.
3. Johannes Goll, S.V.R., Shen C. Shiau, Hank Wu, Peter Uetz. (2008) MPIDB: The microbial protein interaction database. *Bioinfomatics (in review process)*.
4. Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hon, G.C., Myers, C.L., Parsons, A., Friesen, H., Oughtred, R., Tong, A. *et al.* (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol*, **5**, 11.
5. Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K., Gronborg, M. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, **13**, 2363-2371.
6. Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic acids research*, **34**, D187-191.
7. Liu, H., Hu, Z.Z., Zhang, J. and Wu, C. (2006) BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**, 103-105.
8. Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., Ara, T., Nakahigashi, K., Huang, H.C., Hirai, A. *et al.* (2006) Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res*, **16**, 686-691.
9. Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N. *et al.* (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, **433**, 531-537.

Tables

Table 1: Literature-curation strategy

Phase	Method	Publications	Experiments	Interactions	Proteins
I. Literature search	PubMed	36,852	-	-	-
II. Text analysis	Abstracts/Full text read & curated	1,732	2,303	2,289	4,046
III. Protein ID and ontology mapping	Proteins mapped onto UniProt ID's Methods mapped onto PSI-MI	814	1,240	748	940

Table 2: Microbial binary protein-protein interaction datasets

Data set	Species and strains	Experiments[1]	Non-redundant interactions[2]	Abstracts	Interaction per abstract
MPI-LIT	92	1,240	748	814	0.9
MINT	63	234	170	136	1.25
DIP	32	1,404	1,403	109	12.8
BIND	58	1,576	1,564	102	15.3
IntAct	73	13,887	13,242	196	67.5
MPI-UNION	142	15,848	15,077	501	30

[1]experiment is a unique experimental method or unique PubMed ID describing a protein-protein interaction

Table 3

Rank	GO ID	Term	ECO K12 Genes	MPI-LIT K12 Genes Observed	MPI-LIT K12 Genes Expected
Molecular Function					
1	GO:0005515	protein binding	839	142	90.09
2	GO:0003677	DNA binding	472	88	50.68
3	GO:0005524	ATP binding	354	69	38.01
4	GO:0016987	sigma factor activity	9	8	0.97
5	GO:0051082	unfolded protein binding	13	9	1.4
Biological Process					
1	GO:0009432	SOS response	17	14	1.99
2	GO:0006935	chemotaxis	22	15	2.58
3	GO:0065002	intracellular protein transport across a membrane	11	10	1.29
4	GO:0007049	cell cycle	57	24	6.68
5	GO:0051301	cell division	52	24	6.09
6	GO:0006457	protein folding	28	14	3.28
7	GO:0006950	response to stress	153	53	17.93
8	GO:0006281	DNA repair	68	28	7.97
9	GO:0006260	DNA replication	69	33	8.09
10	GO:0006352	transcription initiation	7	6	0.82

Fig 1

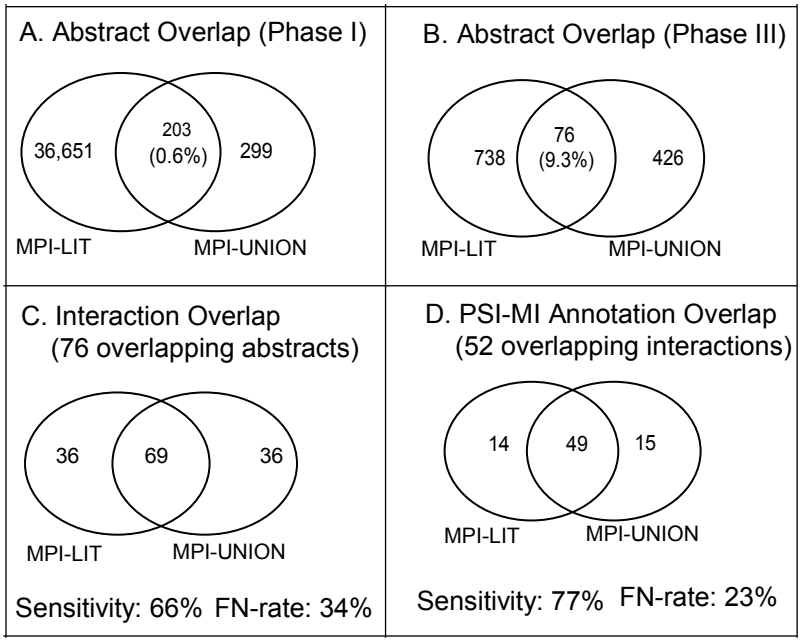
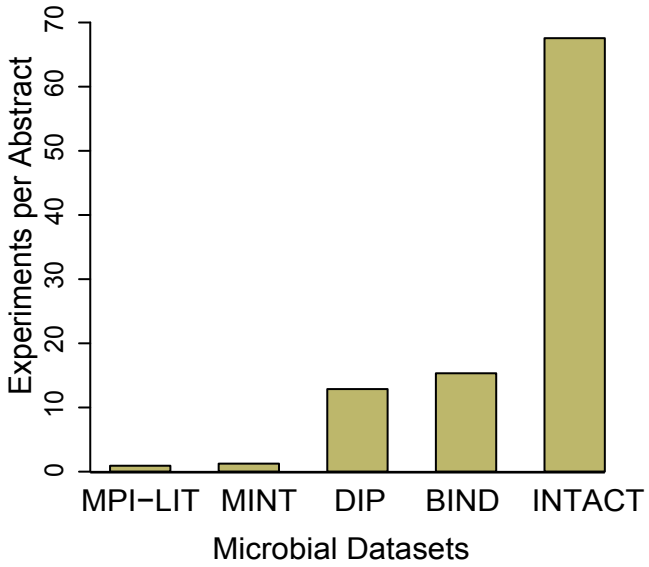
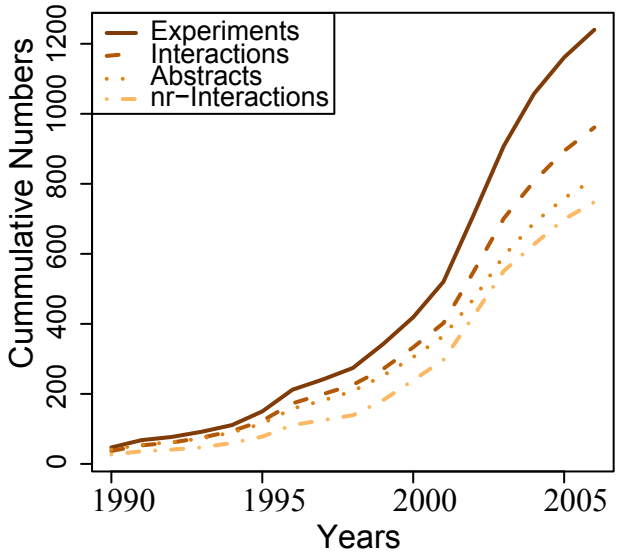

A. Abstract Overlap (Phase I)
36,651
203
(0.6%)
299
MPI-LIT
MPI-UNION
B. Abstract Overlap (Phase III)
738
76
(9.3%)
426
MPI-LIT
MPI-UNION
C. Interaction Overlap
(76 overlapping abstracts)
36
69
36
MPI-LIT
MPI-UNION
Sensitivity: 66% FN-rate: 34%
D. PSI-MI Annotation Overlap
(52 overlapping interactions)
14
49
15
MPI-LIT
MPI-UNION
Sensitivity: 77% FN-rate: 23%

Fig 2

A. Average Interaction per Abstract



B. Interaction Data Over Time



C. Interaction Overlap

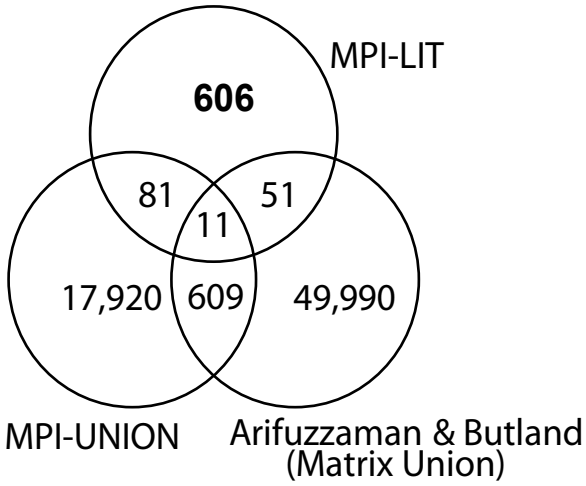
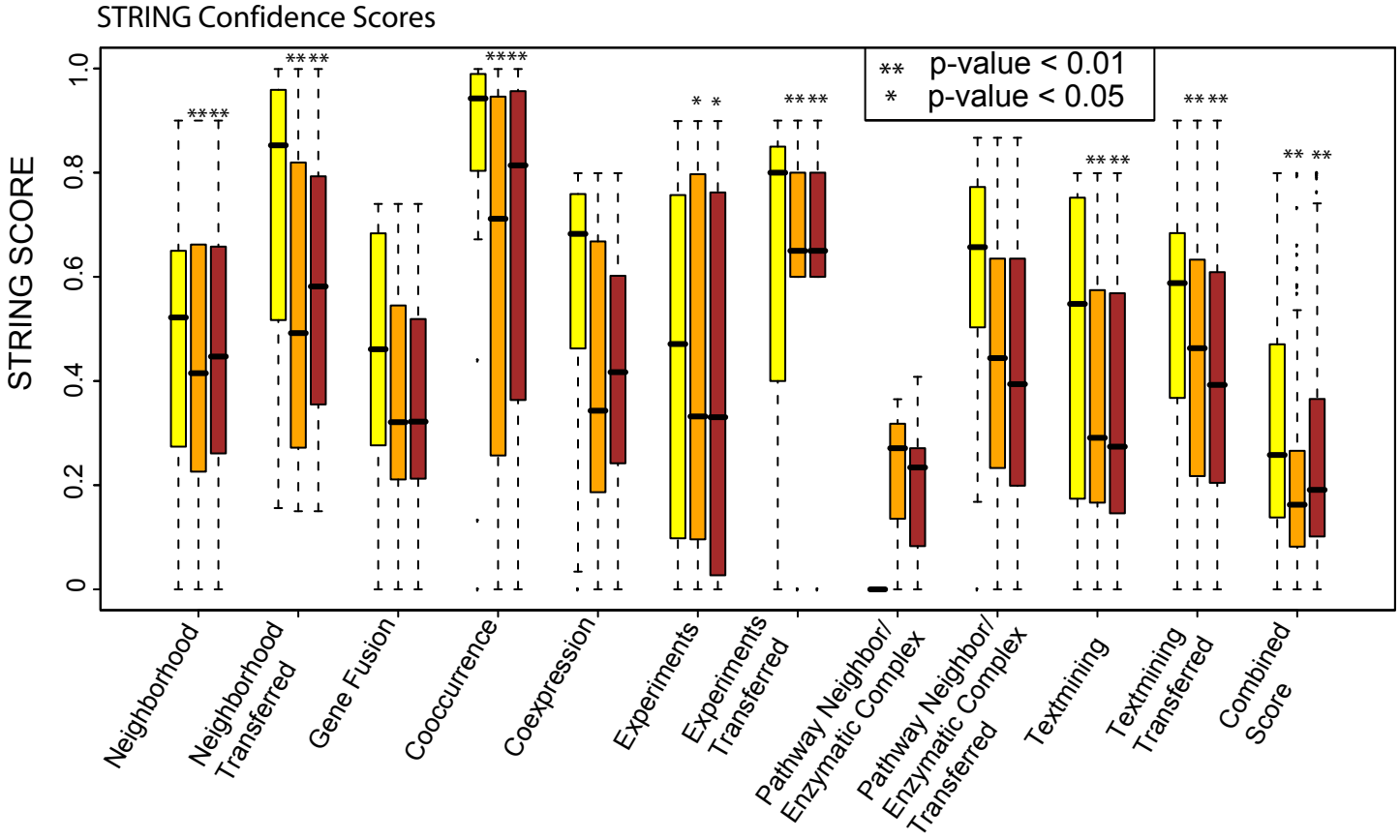


Fig 3


STRING Confidence Scores
STRING SCORE
0
0.2
0.4
0.6
0.8
1.0
** p-value < 0.01
* p-value < 0.05
Neighborhood
Neighborhood Transferred
Gene Fusion
Cooccurrence
Coexpression
Experiments
Experiments Transferred
Pathway Neighbor/ Enzymatic Complex
Pathway Neighbor/ Enzymatic Complex Transferred
Textmining
Textmining Transferred
Combined Score