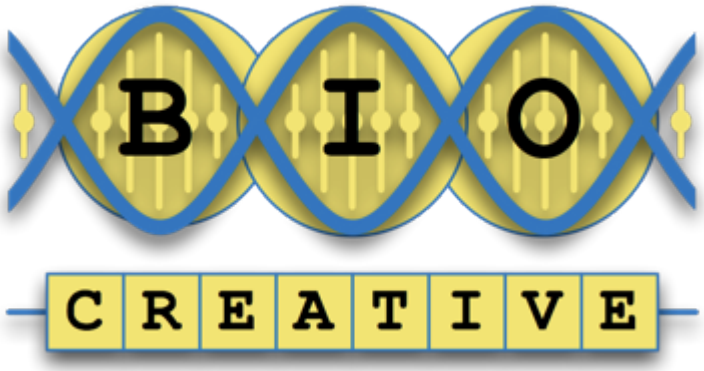
The BioCreative III challenge :

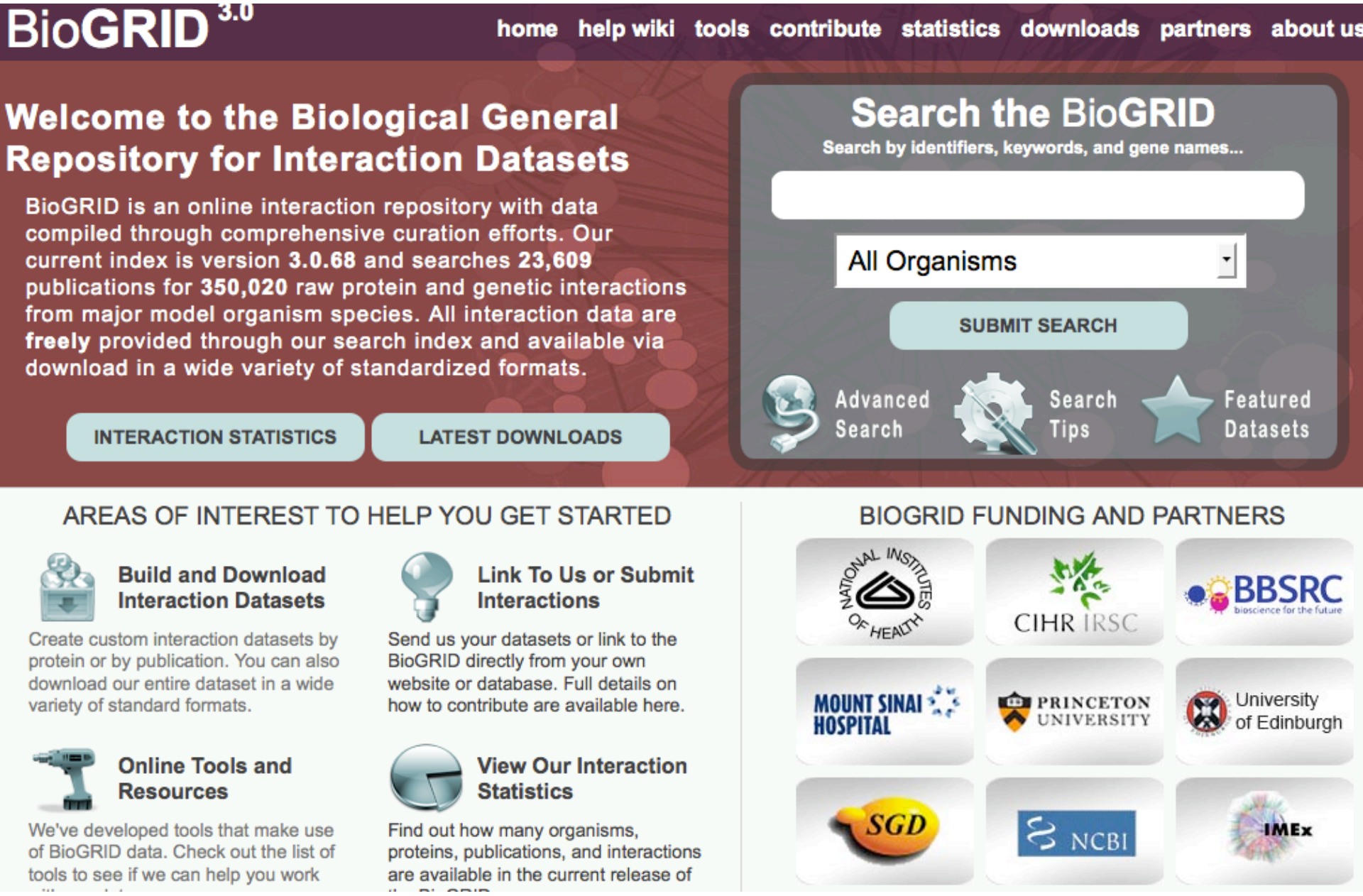
Protein-Protein Interaction Task



Washington, 14th of September 2010

Martin Krallinger,
Spanish National Cancer Centre
(CNIO)

Martin Krallinger, Florian Leitner, Miguel Vazquez, Alfonso Valencia (2010)


BioGRID 3.0
home help wiki tools contribute statistics downloads partners about us
Welcome to the Biological General Repository for Interaction Datasets
BioGRID is an online interaction repository with data compiled through comprehensive curation efforts. Our current index is version 3.0.68 and searches 23,609 publications for 350,020 raw protein and genetic interactions from major model organism species. All interaction data are freely provided through our search index and available via download in a wide variety of standardized formats.
INTERACTION STATISTICS
LATEST DOWNLOADS
Search the BioGRID
Search by identifiers, keywords, and gene names...
All Organisms
SUBMIT SEARCH
Advanced Search
Search Tips
Featured Datasets
AREAS OF INTEREST TO HELP YOU GET STARTED
Build and Download Interaction Datasets
Create custom interaction datasets by protein or by publication. You can also download our entire dataset in a wide variety of standard formats.
Link To Us or Submit Interactions
Send us your datasets or link to the BioGRID directly from your own website or database. Full details on how to contribute are available here.
Online Tools and Resources
We've developed tools that make use of BioGRID data. Check out the list of tools to see if we can help you work
View Our Interaction Statistics
Find out how many organisms, proteins, publications, and interactions are available in the current release of
BIOGRID FUNDING AND PARTNERS
NATIONAL INSTITUTES OF HEALTH
CIHR IRSC
BBSRC bioscience for the future
MOUNT SINAI HOSPITAL
PRINCETON UNIVERSITY
University of Edinburgh
SGD
NCBI
IMEx

1. Background and motivation

go to: HomoMINT: an inferred human network Domino: a domain peptide interactions database VirusMINT: a v

MINT | Home | Search | Curation | Statistics | Download

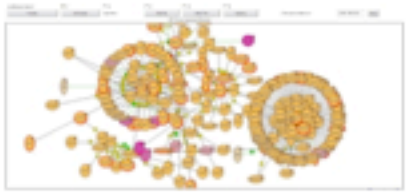
Statistics:
- 86565 interactions
- 30947 proteins
- 3624 pmids

FEBS Letters special issue: **the Digital, Democratic Age of Scientific Abstracts**

The spreadsheet for data submission to the FEBS Letters experiment: is available here

Welcome to MINT, the Molecular INTeraction database. MINT focuses on **experimentally verified protein-protein interactions** mined from the scientific literature by expert curators. The full MINT dataset can be freely downloaded.

The curated data can be analyzed in the context of the high throughput data and viewed graphically with the **'MINT Viewer'**.



MINT has signed the **IMEx agreement** (http://www.imexconsortium.org/) to share curation efforts and supports the Protein Standard Initiative (PSI) recommendation.

FEBS Letters and the FEBS Journal in collaboration with MINT enhance the content of their articles with the addition of Structured Digital Abstracts

Please, in any articles making use of the data extracted from MINT, refer to *MINT, the molecular interaction database: 2009 update. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G. Nucleic Acids Res. 2010 Jan;38(Database issue):D532-9. Epub 2009 Nov 6.*[Abstract]

Overview

1. General Introduction, background and motivation

2. Interaction Article Task (ACT)
   - 2.1. Introduction, task definition, participation
   - 2.2. Data preparation: training, development, test sets
   - 2.3. Results
   - 2.4. Participating systems methods
   - 2.5. ACT discussion & conclusions

3. Interaction Method Task (IMT)
   - 3.1. Introduction, task definition, participation
   - 3.2. Data preparation: training, development, test sets
   - 3.3. Results
   - 3.4. Participating systems methods
   - 3.5. IMT discussion & conclusions

4. Conclusions & outlook

5. Acknowledgements

Protein-Protein Interaction (PPI)

Specific physical contacts with molecular binding between proteins, both transient as well we stable contacts.

PPI information: literature, large scale experiments, bioinformatics predictions

Public repositories integrate information from large- and small-scale PPI experiments reported in the scientific literature

Pathguide contains information about 325 biological pathway related resources and molecular interaction related resources (pathguide.org)

Annotation effort shared by various interaction databases: BioGRID, MINT, BIND, CORUM, DIP,HAPPI,HPRD,I2D,InnateDB,IntAct,InteroPorc, iRefIndex, iRefWeb, MatrixDB,MIPS, PC, PIMRider

Common vocabulary and standards to improve consistency and Efficiency of PPI annotations: PSI-MI

Martin Krallinger, Florian Leitner, Miguel Vazquez, Alfonso Valencia (2010)

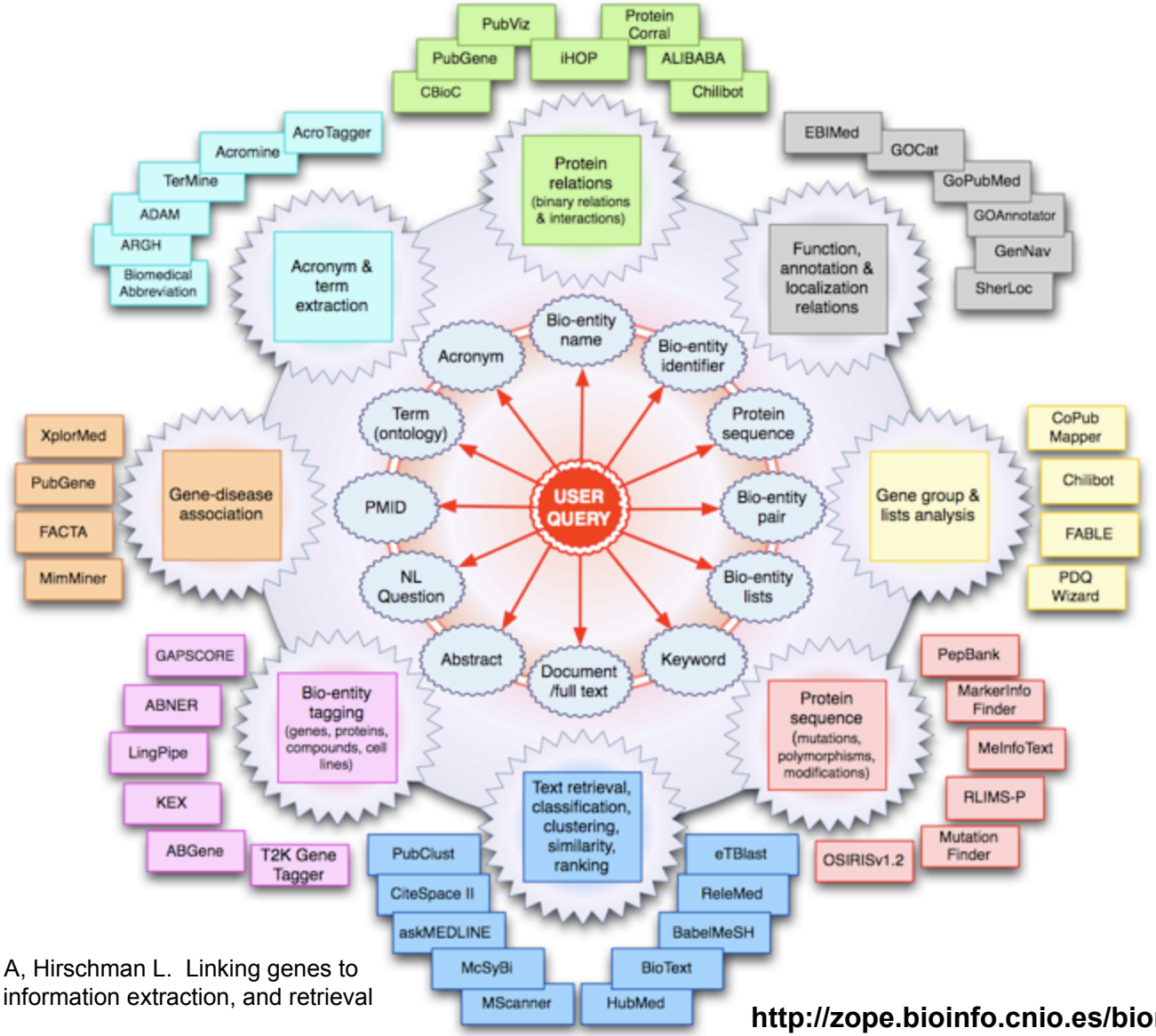
PPI Databases

Acronym	Database Full Name and URL	PPI Sources	Type of MI	Species	*n* Proteins (Dec. 2009)	*n* Interactions (Dec. 2009)
Primary Databases: PPI experimental data (curated from specific SSc & LSc published studies)						
BIND	Biomolecular Interaction Network Database, http://bond.unleashedinformatics.com/	Ssc & Lsc published studies (literature-curated)	PPIs & others	All	[31,972]	[58,266]
BioGRID	Biological General Repository for Interaction Datasets, http://www.thebiogrid.org/	Ssc & Lsc published studies (literature-curated)	PPIs & others	All	[28,717]	[108,691]
DIP	Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu/dip/	Ssc & Lsc published studies (literature-curated)	Only PPIs	All	20,728	57,683
HPRD	Human Protein Reference Database, http://www.hprd.org/	Ssc & Lsc published studies (literature-curated)	Only PPIs	Human	27,081	38,806
IntAct	IntAct Molecular Interaction Database, http://www.ebi.ac.uk/intact/	Ssc & Lsc published studies (literature-curated)	PPIs & others	All	[60,504]	[202,826]
MINT	Molecular INTeraction database, http://mint.bio.uniroma2.it/mint/	Ssc & Lsc published studies (literature-curated)	Only PPIs	All	30,089	83,744
MIPS-MPact	MIPS protein interaction resource on yeast, http://mips.gsf.de/genre/proj/mpact/	Derived from CYGD	Only PPIs	Yeast	1,500	4,300
MIPS-MPPI	MIPS Mammalian Protein-Protein Interaction Database, http://mips.gsf.de/proj/ppi	Ssc published studies (literature-curated)	Only PPIs	Mammalian	982	937
Meta-Databases: PPI experimental data (integrated and unified from different public repositories)						
APID	Agile Protein Interaction DataAnalyzer, http://bioinfow.dep.usal.es/apid/	BIND, BioGRID, DIP, HPRD, IntAct, MINT	Only PPIs	All	56,460	322,579
MPIDB	The Microbial Protein Interaction Database, http://www.jcvi.org/mpidb/	BIND, DIP, IntAct, MINT, other sets (exp & lit.-curated)	Only PPIs	Microbial	7,810	24,295
PINA	Protein Interaction Network Analysis platform, http://csbi.ltdk.helsinki.fi/pina/	BioGRID, DIP, HPRD, IntAct, MINT, MPact	Only PPIs	All	[?]	188,823
Prediction Databases: PPI experimental and predicted data ("functional interactions", i.e., interactions *lato sensu* derived from different types of data)						
MiMI	Michigan Molecular Interactions, http://mimi.ncibi.org/MimiWeb/	BIND, BioGRID, DIP, HPRD, IntAct, & nonPPI data	PPIs & others	All	[45,452]	[391,386]
PIPs	Human PPI Prediction database, http://www.compbio.dundee.ac.uk/www-pips/	BIND, DIP, HPRD, OPHID, & nonPPI data	PPIs & others	Human	[?]	[37,606]
OPHID	Online Predicted Human Interaction Database, http://ophid.utoronto.ca/	BIND, BioGRID, HPRD, IntAct, MINT, MPact, & nonPPI data	PPIs & others	Human	[?]	[424,066]
STRING	Known and Predicted Protein-Protein Interactions, http://string.embl.de/	BIND, BioGRID, DIP, HPRD, IntAct, MINT, & nonPPI data	PPIs & others	All	[2,590,259]	[88,633,860]
UniHI	Unified Human Interactome, http://www.mdc-berlin.de/unihi/	BIND, BioGRID, DIP, HPRD, IntAct, MINT, & nonPPI data	PPIs & others	Human	[22,307]	[200,473]

The table divided in three sections: **primary databases**, which include PPIs from large- and small-scale (Lsc & Ssc) experimental data that are usually obtained from curation of research articles (8 resources included: BIND, BioGRID, DIP, HPRD, IntAct, MINT, MIPS-MPACT, MIPS-MPPI); **meta-databases**, which include PPIs derived from integration and unification of several primary repositories (3 resources: APID, MPIDB, PINA); **prediction databases**, which include PPIs from experimental analyses together with predicted PPIs obtained from the analyses of heterogenous biological data (5 resources: MiMI, PIPs, OPHID, STRING, UniHI). The table shows the total number of proteins and interactions that were reported by each repository in December 2009 (as far as we could see in the respective Web site). The numbers are in brackets [ ] when the repository includes PPIs and other types of interactions (e.g., protein-ligand interactions or for the case of prediction databases nonPPI data). The question mark [?] indicates that the number of distinct proteins included is such repository could not be found in the Web.
doi:10.1371/journal.pcbi.1000807.t001

De Las Rivas J, Fontanillo C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput Biol. 2010 Jun 24;6(6):e1000807.

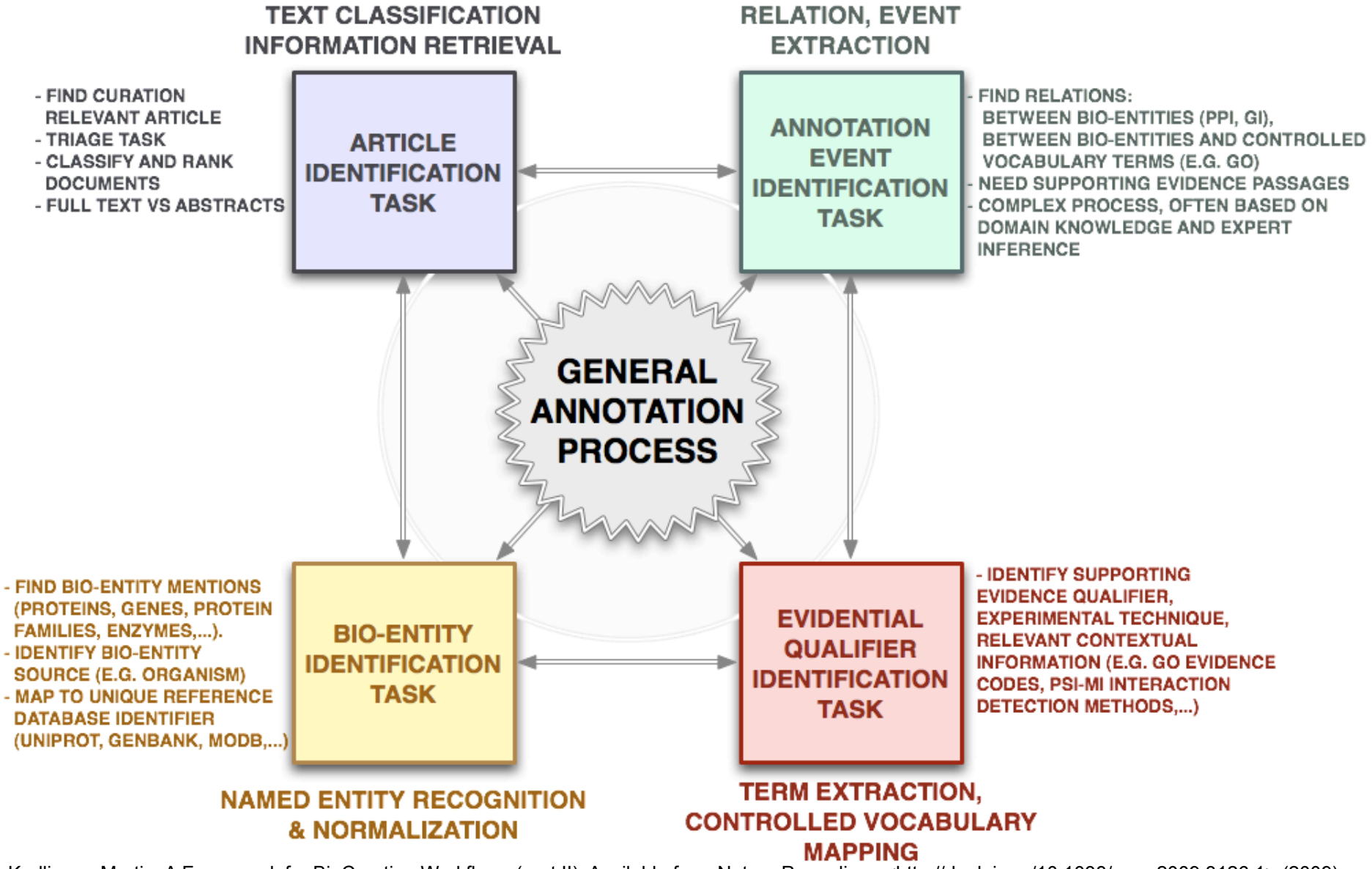
BioNLP applications



Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol. 2008;9 Suppl 2:S8.

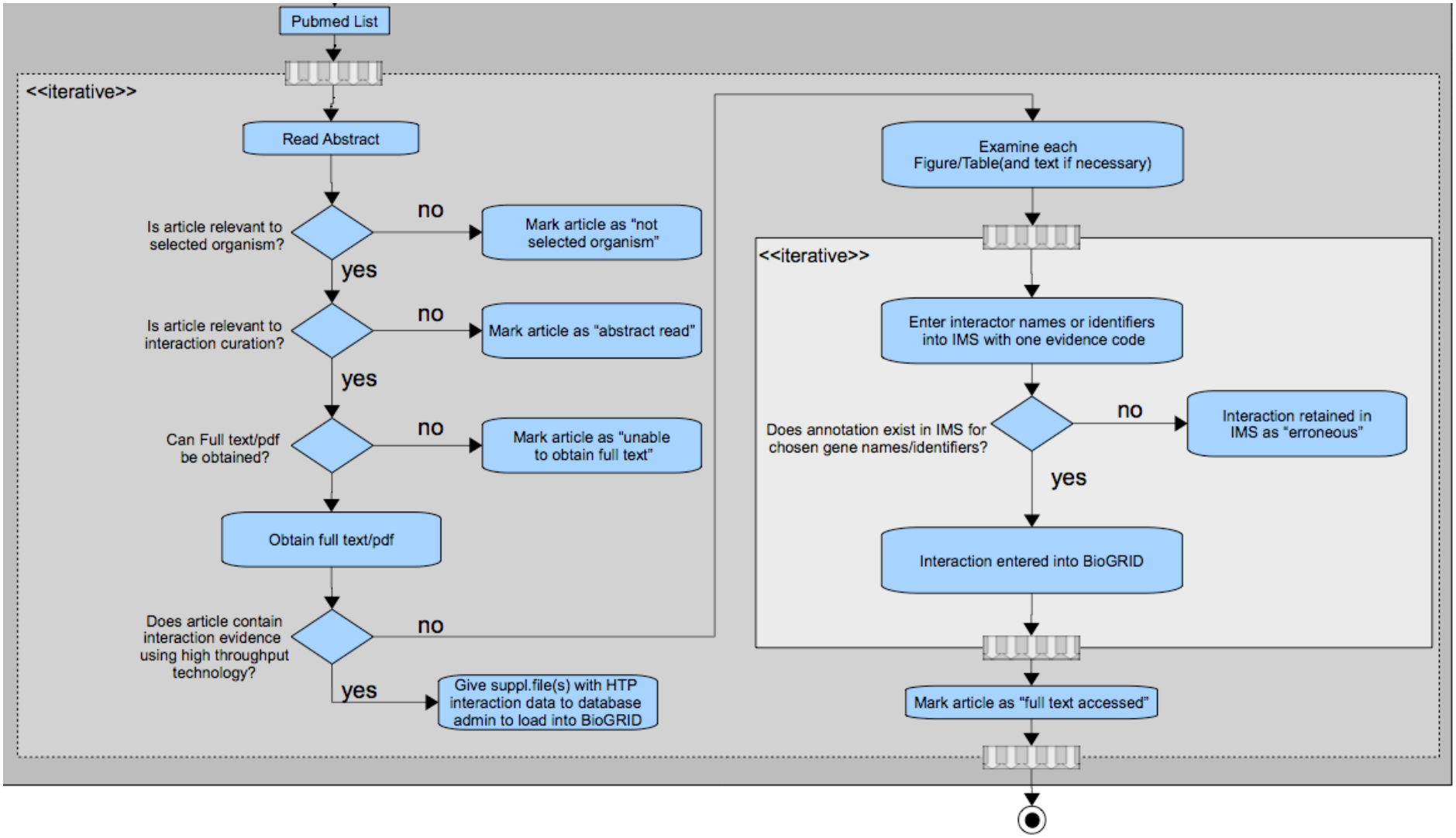
http://zope.bioinfo.cnio.es/bionlp_tools

Biocuration workflows : Tasks & curation pipeline



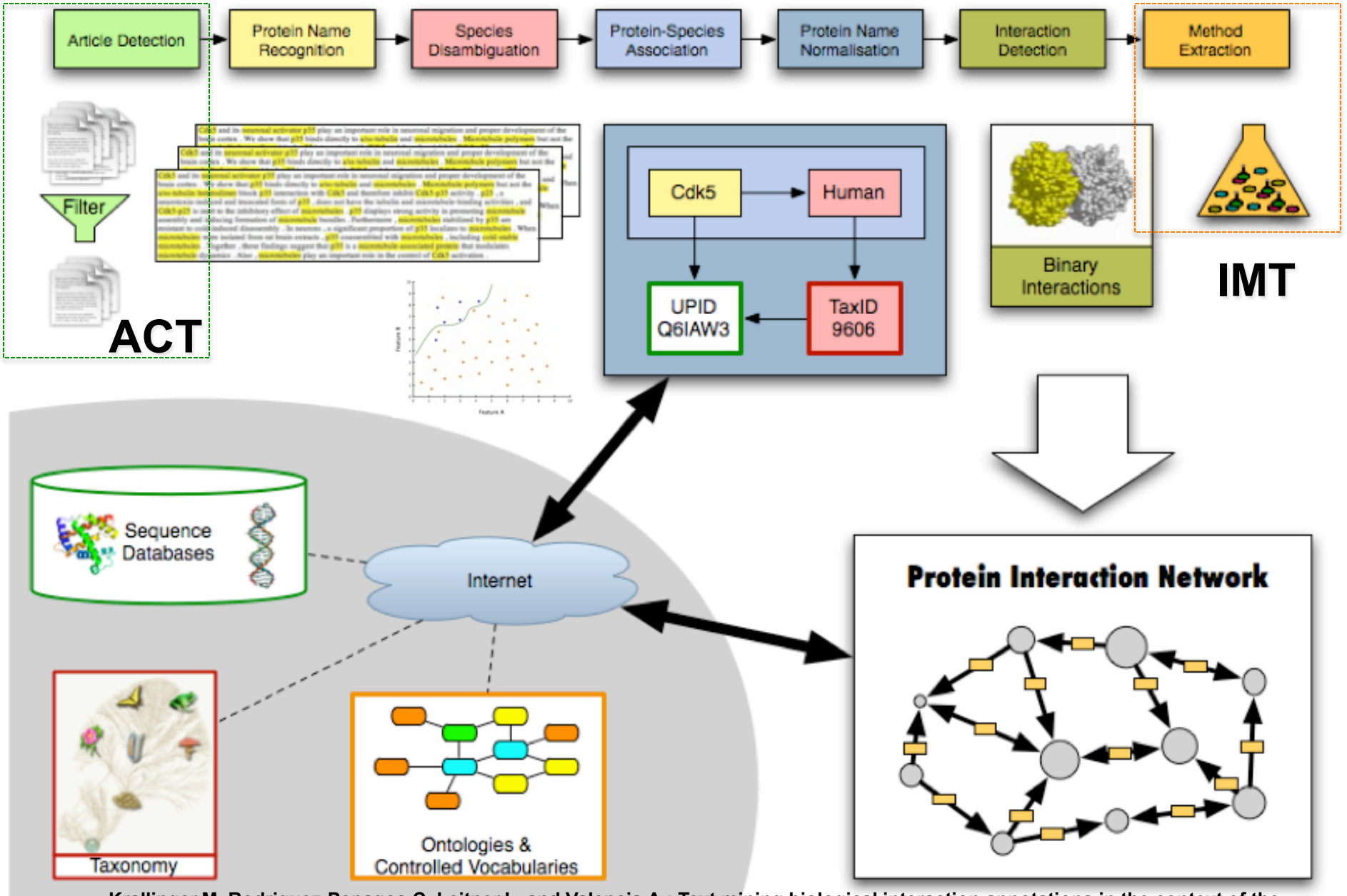
Krallinger, Martin. A Framework for BioCuration Workflows (part II). Available from Nature Precedings <http://dx.doi.org/10.1038/npre.2009.3126.1> (2009)

BioGRID Biocuration workflow



Provided by Andrew Winter (http://wiki.thebiogrid.org/doku.php/curation_description)

1. Background and motivation



Krallinger,M. Rodriguez-Penagos,C. Leitner,L. and Valencia,A.: Text mining biological interaction annotations in the context of the Semantic Web Proceedings of the Network Tools and Applications in Biology (NETTAB) workshop (2007) pp. 47-53

ACT: Article categorization task

- Binary classification of recent PubMed abstracts as PPI relevant
- Predictions provided together with a confidence score in the]0..1] range
- Evaluation based on AUC iP/R (also additional analysis, f-score, accuracy)
- NOT balanced set, abstracts, journals of biocuration interest
- Exhaustive manual revision by three domain experts and refinement based on database curators of BioGRID and MINT
- IAA pairwise percentage agreement between MINT & BioGRID 95%.
- Article ID ➠ Class ➠ [Rank ➠] Confidence

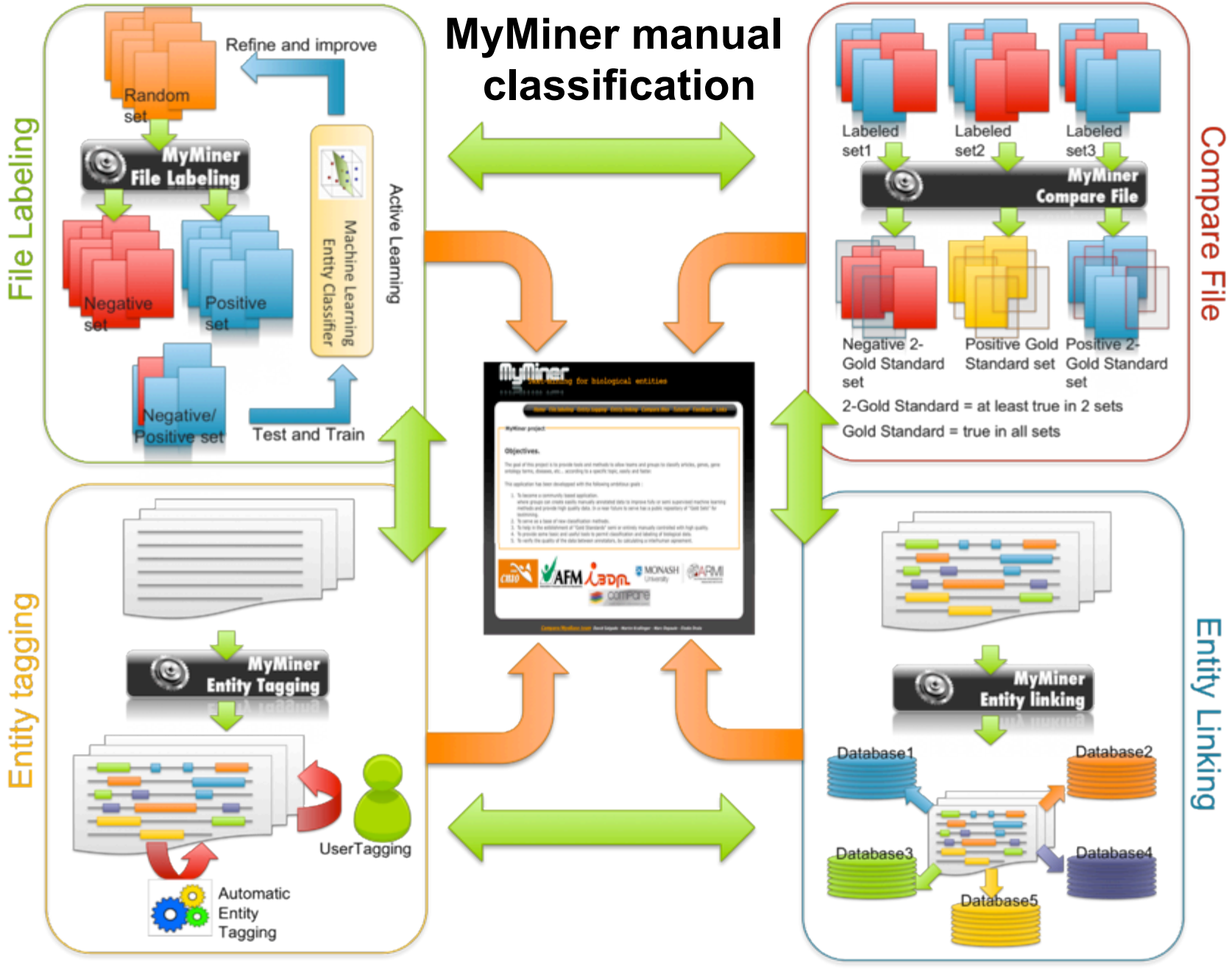
TRAINING SET (Balanced)	DEVELOMENT SET (Unbalanced)	TEST SET (Unbalanced)
total size: 2280	total size: 4000	total size: 6000
+ PPI: 1140	+ PPI: 682	+ PPI: 910
Not PPI: 1140	Not PPI: 3318	Not PPI: 5090
proportion: 50%	proportion: 17.05%	proportion: 15.17%

ACT participating teams

TEAM	LEADER	INSTITUTION	# RUNS	ONLINE
65	Fabio Rinaldi	University of Zurich	5	N
70	Sérgio Matos	Universidade de Aveiro, IEETA	5	N
73	W John Wilbur	NCBI	5	N
81	Luis Rocha	Indiana University	10	Y
89	Shashank Agarwal	University of Wisconsin-Milwaukee	10	Y
90	Xinglong Wang	National Centre for Text Mining	5	N
92	Keith Noto	Tufts University	1	N
100	Zhiyong Lu	NCBI\NLM\NIH	4	N
104	Jean-Fred Fontaine	Max Delbrück Center	5	N
88	Ashish Tendulkar	IIT Madras	2	N

- 10 Teams, 52 runs, two teams also submitted online runs

3. Interaction Article Task



AUC iP/R

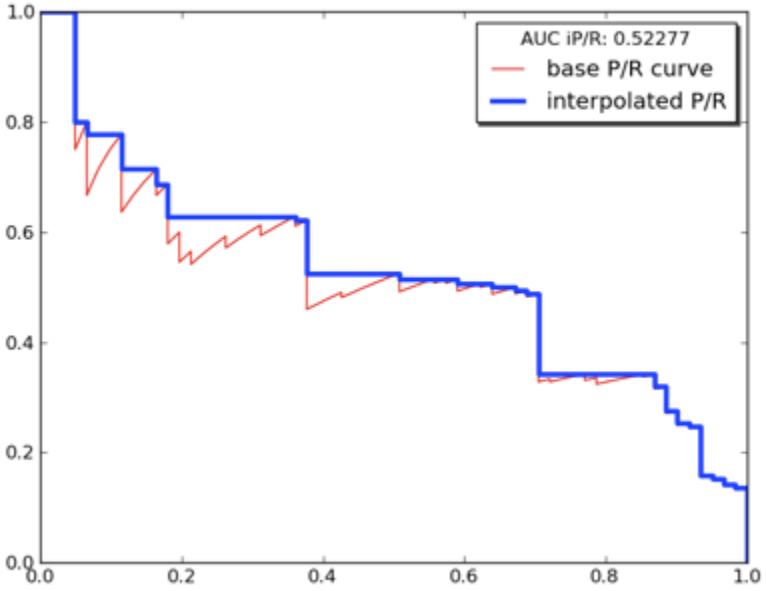
$$A(f_{pr}) := \sum_{i=1}^{n} (p_{i_j} * (r_j - r_{j-1}))$$

$$p_i(r) = max_{r' \geq r} p(r')$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$



3. Interaction Article Task

Team	Run/Srvr	Accuracy	Specificity	Sensitivity	F-Score	MCC	AUC iP/R
T65	RUN_1	88.68%	97.64%	38.57%	50.83%	0.48297	63.85%
T65	RUN_2	87.93%	93.07%	59.23%	59.82%	0.52727	63.89%
T65	RUN_3	67.05%	64.19%	83.08%	43.34%	0.34244	41.74%
T65	RUN_4	73.68%	74.13%	71.21%	45.08%	0.34650	41.74%
T65	RUN_5	88.00%	94.40%	52.20%	56.89%	0.50255	62.39%
T70	RUN_1	*56.45%*	*49.70%*	94.18%	39.62%	0.31789	56.76%
T70	RUN_2	87.41%	96.11%	38.79%	48.32%	0.43346	56.76%
T70	RUN_3	81.92%	83.61%	72.53%	54.91%	0.46563	56.76%
T70	RUN_4	47.77%	39.04%	**96.59%**	35.95%	0.27060	56.76%
T70	RUN_5	86.84%	98.62%	20.99%	32.62%	0.34488	56.76%
T73	RUN_1	87.55%	91.81%	63.74%	60.83%	0.53524	65.91%
T73	RUN_2	**89.15%**	94.95%	56.70%	61.32%	**0.55306**	67.96%
T73	RUN_3	87.78%	92.61%	60.77%	60.14%	0.52932	65.89%
T73	**RUN_4**	88.88%	94.34%	58.35%	**61.42%**	0.55054	**67.98%**
T73	RUN_5	87.62%	92.18%	62.09%	60.33%	0.53031	65.37%
T81	RUN_1	59.03%	58.76%	60.55%	30.96%	0.13949	19.93%
T81	RUN_2	58.47%	57.86%	61.87%	31.12%	0.14219	19.69%
T81	RUN_3	25.37%	14.72%	84.95%	25.66%	-0.00344	15.66%
T81	RUN_4	63.45%	69.16%	31.54%	20.74%	0.00538	16.20%
T81	RUN_5	69.17%	77.35%	23.41%	18.72%	0.00645	*15.63%*
T81	SRVR_10	85.38%	99.61%	5.82%	10.78%	0.17771	50.25%
T81	SRVR_11	84.73%	99.86%	0.11%	*0.22%*	*-0.00272*	46.02%
T81	SRVR_12	84.30%	98.86%	*2.86%*	5.23%	0.05244	32.11%
T81	SRVR_13	84.88%	99.92%	0.77%	1.52%	0.05791	18.59%
T81	SRVR_9	84.88%	**99.98%**	0.44%	0.88%	0.05220	44.19%
T88	RUN_1	42.63%	35.11%	84.73%	30.94%	0.15238	21.97%
T88	RUN_2	56.92%	53.73%	74.73%	34.47%	0.20417	26.04%
T89	RUN_1	80.02%	80.90%	75.06%	53.26%	0.44911	61.29%
T89	RUN_2	81.00%	81.75%	76.81%	55.08%	0.47242	62.13%
T89	RUN_3	82.40%	83.85%	74.29%	56.15%	0.48180	60.48%
T89	RUN_4	87.73%	94.79%	48.24%	54.40%	0.47967	43.76%
T89	RUN_5	87.27%	91.81%	61.87%	59.58%	0.52082	48.47%
T89	SRVR_4	77.80%	77.84%	77.58%	51.46%	0.43152	57.44%
T89	**SRVR_5**	78.05%	78.15%	77.47%	51.71%	0.43424	57.56%
T89	SRVR_6	79.90%	81.00%	73.74%	52.67%	0.44073	54.97%
T89	SRVR_7	86.25%	92.06%	53.74%	54.24%	0.46156	41.58%
T89	SRVR_8	86.87%	90.39%	67.14%	60.80%	0.53336	47.40%
T90	RUN_1	88.73%	95.15%	52.86%	58.73%	0.52736	51.14%
T90	RUN_2	88.70%	94.97%	53.63%	59.01%	0.52890	51.65%
T90	RUN_3	88.32%	93.93%	56.92%	59.64%	0.52914	65.24%
T90	RUN_4	88.93%	96.03%	49.23%	57.44%	0.52237	49.26%
T90	RUN_5	88.60%	95.05%	52.53%	58.29%	0.52204	50.83%
T92	RUN_1	86.22%	90.77%	60.77%	57.22%	0.49155	50.99%
T100	RUN_1	88.77%	96.82%	43.74%	54.15%	0.50005	61.62%
T100	RUN_2	88.27%	93.89%	56.81%	59.49%	0.52732	61.86%
T100	RUN_3	81.13%	82.69%	72.42%	53.80%	0.45256	60.25%
T100	RUN_4	81.85%	82.85%	76.26%	56.04%	0.48270	63.75%
T104	RUN_1	80.12%	80.69%	76.92%	53.99%	0.45999	53.67%
T104	RUN_2	80.07%	80.47%	77.80%	54.21%	0.46370	53.67%

3. Interaction Article Task

Team	Run/Srvr	Accuracy	Specificity	Sensitivity	F-Score	MCC	AUC iP/R
T65	RUN_1	88.68%	97.64%	38.57%	50.83%	0.48297	63.85%
T65	RUN_2	87.93%	93.07%	59.23%	59.82%	0.52727	63.89%
T65	RUN_3	67.05%	64.19%	83.08%	43.34%	0.34244	41.74%
T65	RUN_4	73.68%	74.13%	71.21%	45.08%	0.34650	41.74%
T65	RUN_5	88.00%	94.40%	52.20%	56.89%	0.50255	62.39%
T70	RUN_1	*56.45%*	*49.70%*	94.18%	39.62%	0.31789	56.76%
T70	RUN_2	87.41%	96.11%	38.79%	48.32%	0.43346	56.76%
T70	RUN_3	81.92%	83.61%	72.53%	54.91%	0.46563	56.76%
T70	RUN_4	47.77%	39.04%	**96.59%**	35.95%	0.27060	56.76%
T70	RUN_5	86.84%	98.62%	20.99%	32.62%	0.34488	56.76%
T73	RUN_1	87.55%	91.81%	63.74%	60.83%	0.53524	65.91%
T73	RUN_2	**89.15%**	94.95%	56.70%	61.32%	**0.55306**	67.96%
T73	RUN_3	87.78%	92.61%	60.77%	60.14%	0.52932	65.89%
T73	**RUN_4**	88.88%	94.34%	58.35%	**61.42%**	0.55054	**67.98%**
T73	RUN_5	87.62%	92.18%	62.09%	60.33%	0.53031	65.37%
T81	RUN_1	59.03%	58.76%	60.55%	30.96%	0.13949	19.93%
T81	RUN_2	58.47%	57.86%	61.87%	31.12%	0.14219	19.69%
T81	RUN_3	25.37%	14.72%	84.95%	25.66%	-0.00344	15.66%
T81	RUN_4	63.45%	69.16%	31.54%	20.74%	0.00538	16.20%
T81	RUN_5	69.17%	77.35%	23.41%	18.72%	0.00645	*15.63%*
T81	SRVR_10	85.38%	99.61%	5.82%	10.78%	0.17771	50.25%
T81	SRVR_11	84.73%	99.86%	0.11%	*0.22%*	*-0.00272*	46.02%
T81	SRVR_12	84.30%	98.86%	*2.86%*	5.23%	0.05244	32.11%
T81	SRVR_13	84.88%	99.92%	0.77%	1.52%	0.05791	18.59%
T81	SRVR_9	84.88%	**99.98%**	0.44%	0.88%	0.05220	44.19%

3. Interaction Article Task

Team	Run/Srvr	Accuracy	Specificity	Sensitivity	F-Score	MCC	AUC iP/R
T88	RUN_1	42.63%	35.11%	84.73%	30.94%	0.15238	21.97%
T88	RUN_2	56.92%	53.73%	74.73%	34.47%	0.20417	26.04%
T89	RUN_1	80.02%	80.90%	75.06%	53.26%	0.44911	61.29%
T89	RUN_2	81.00%	81.75%	76.81%	55.08%	0.47242	62.13%
T89	RUN_3	82.40%	83.85%	74.29%	56.15%	0.48180	60.48%
T89	RUN_4	87.73%	94.79%	48.24%	54.40%	0.47967	43.76%
T89	RUN_5	87.27%	91.81%	61.87%	59.58%	0.52082	48.47%
T89	SRVR_4	77.80%	77.84%	77.58%	51.46%	0.43152	57.44%
T89	**SRVR_5**	78.05%	78.15%	77.47%	51.71%	0.43424	57.56%
T89	SRVR_6	79.90%	81.00%	73.74%	52.67%	0.44073	54.97%
T89	SRVR_7	86.25%	92.06%	53.74%	54.24%	0.46156	41.58%
T89	SRVR_8	86.87%	90.39%	67.14%	60.80%	0.53336	47.40%
T90	RUN_1	88.73%	95.15%	52.86%	58.73%	0.52736	51.14%
T90	RUN_2	88.70%	94.97%	53.63%	59.01%	0.52890	51.65%
T90	RUN_3	88.32%	93.93%	56.92%	59.64%	0.52914	65.24%
T90	RUN_4	88.93%	96.03%	49.23%	57.44%	0.52237	49.26%
T90	RUN_5	88.60%	95.05%	52.53%	58.29%	0.52204	50.83%
T92	RUN_1	86.22%	90.77%	60.77%	57.22%	0.49155	50.99%
T100	RUN_1	88.77%	96.82%	43.74%	54.15%	0.50005	61.62%
T100	RUN_2	88.27%	93.89%	56.81%	59.49%	0.52732	61.86%
T100	RUN_3	81.13%	82.69%	72.42%	53.80%	0.45256	60.25%
T100	RUN_4	81.85%	82.85%	76.26%	56.04%	0.48270	63.75%
T104	RUN_1	80.12%	80.69%	76.92%	53.99%	0.45999	53.67%
T104	RUN_2	80.07%	80.47%	77.80%	54.21%	0.46370	53.67%
T104	RUN_3	64.93%	59.86%	93.30%	44.66%	0.38161	53.67%
T104	RUN_4	69.78%	66.25%	89.56%	47.34%	0.40530	53.67%
T104	RUN_5	86.27%	98.47%	18.02%	28.47%	0.30064	53.67%

3. Interaction Article Task

TEAM & RUN	AUC iP/R
T73_RUN_4	0.6798
T73_RUN_2	0.6796
T73_RUN_1	0.6591
T73_RUN_3	0.6589
T73_RUN_5	0.6537
T90_RUN_3	0.6524
T65_RUN_2	0.6389
T65_RUN_1	0.6385
T100_RUN_4	0.6375
T65_RUN_5	0.6239
T89_RUN_2	0.6213
T100_RUN_2	0.6186
T100_RUN_1	0.6162
T89_RUN_1	0.6129
T89_RUN_3	0.6048

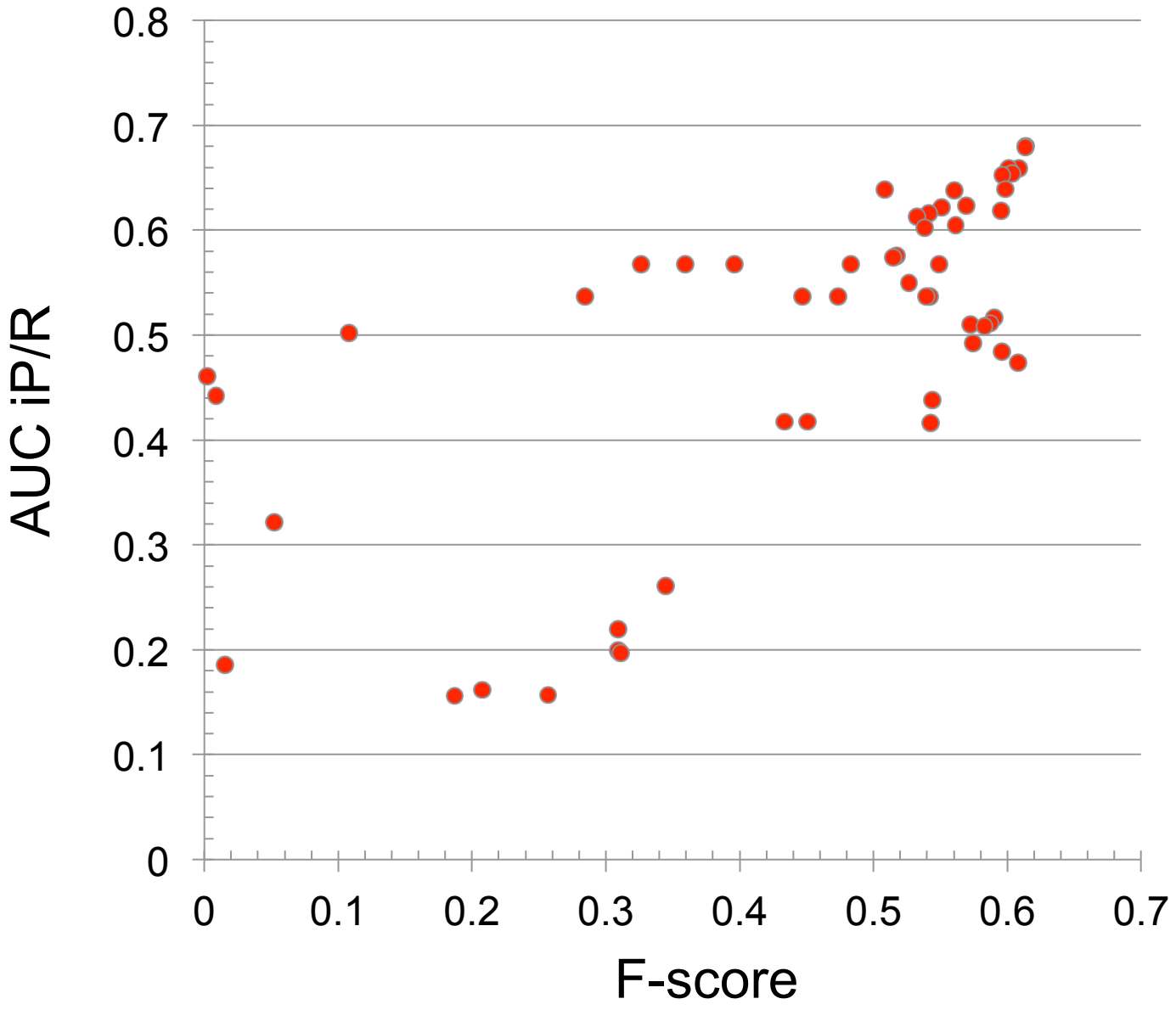
TEAM & RUN	F-SCORE
T73_RUN_4	0.6142
T73_RUN_2	0.6132
T73_RUN_1	0.6083
T89_SRVR_8	0.608
T73_RUN_5	0.6033
T73_RUN_3	0.6014
T65_RUN_2	0.5982
T90_RUN_3	0.5964
T89_RUN_5	0.5958
T100_RUN_2	0.5949
T90_RUN_2	0.5901
T90_RUN_1	0.5873
T90_RUN_5	0.5829
T90_RUN_4	0.5744
T92_RUN_1	0.5722

3. Interaction Article Task

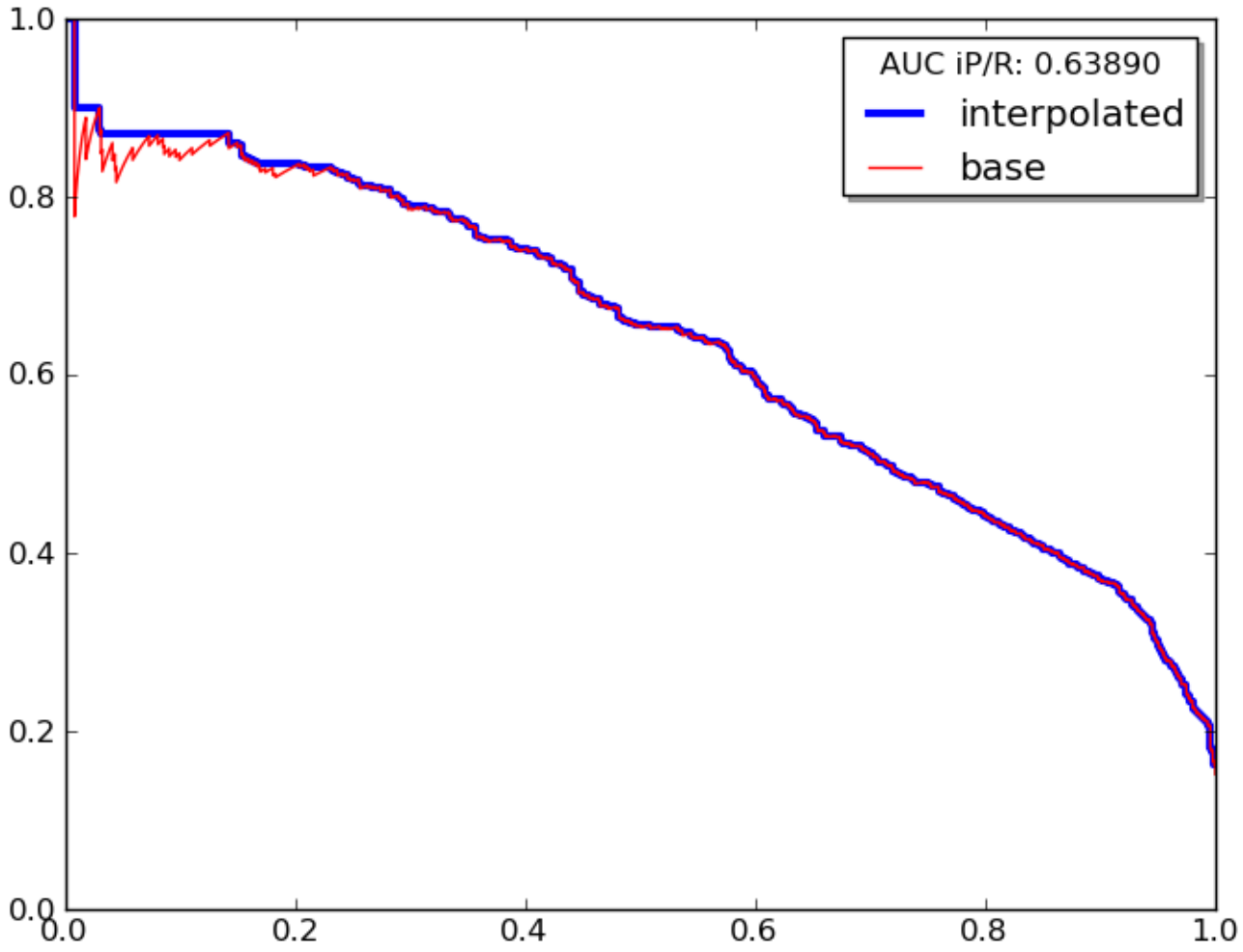
TEAM & RUN	ACCURACY
T73_RUN_2	0.8915
T90_RUN_4	0.8893
T73_RUN_4	0.8888
T100_RUN_1	0.8877
T90_RUN_1	0.8873
T90_RUN_2	0.887
T65_RUN_1	0.8868
T90_RUN_5	0.886
T90_RUN_3	0.8832
T100_RUN_2	0.8827
T65_RUN_5	0.88
T65_RUN_2	0.8793
T73_RUN_3	0.8778
T89_RUN_4	0.8773
T73_RUN_5	0.8762

TEAM & RUN	MCC
T73_RUN_2	0.55306
T73_RUN_4	0.55054
T73_RUN_1	0.53524
T89_SRVR_8	0.53336
T73_RUN_5	0.53031
T73_RUN_3	0.52932
T90_RUN_3	0.52914
T90_RUN_2	0.5289
T90_RUN_1	0.52736
T100_RUN_2	0.52732
T65_RUN_2	0.52727
T90_RUN_4	0.52237
T90_RUN_5	0.52204
T89_RUN_5	0.52082
T65_RUN_5	0.50255

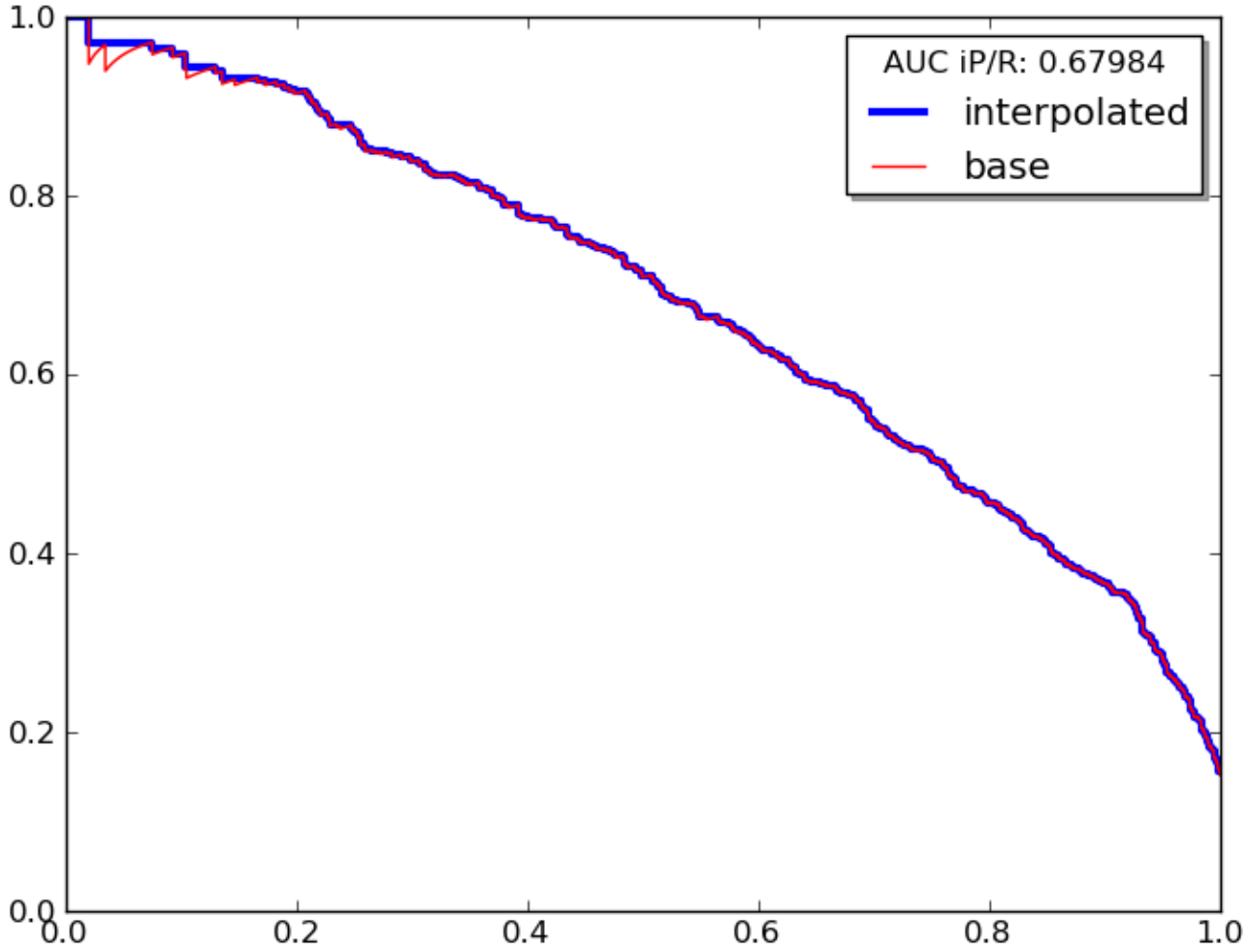
3. Interaction Article Task



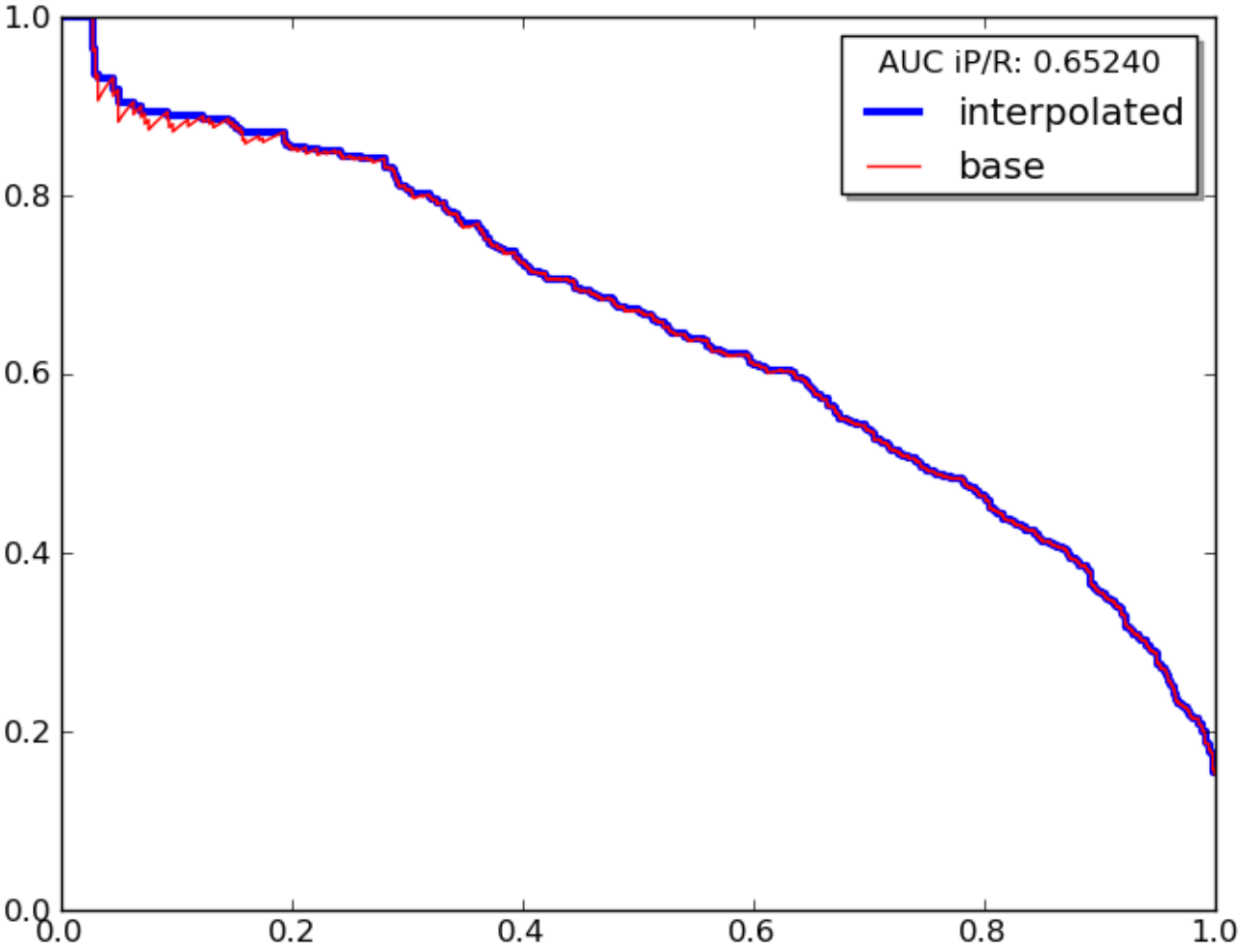
Team 65 Run 2


AUC iP/R: 0.63890
interpolated
base
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Team 73 Run 4


AUC iP/R: 0.67984
interpolated
base
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Team 90 Run 3


AUC iP/R: 0.65240
interpolated
base
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

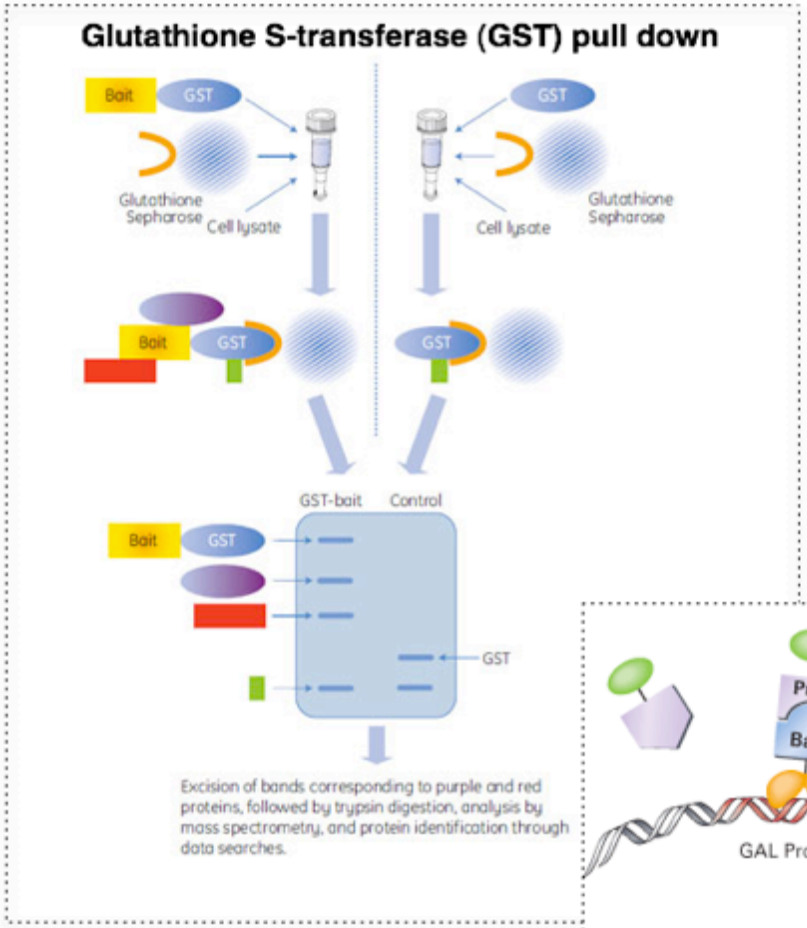
Overview of participating ACT methods

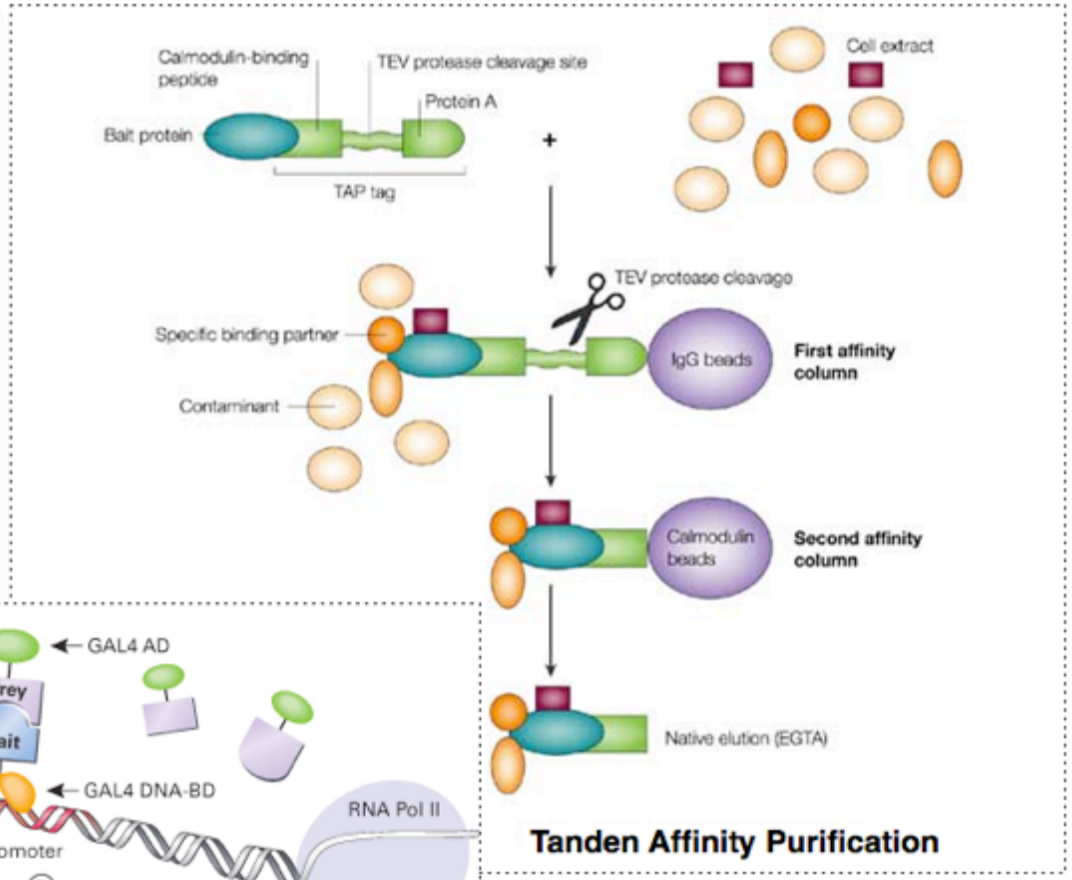
- A considerable number used supervised learning methods (also one applied semi-supervised learning).
- Methods: SVM, naïve Bayes, logistic regression, max. entropy.
- Not general correction for class imbalance (some added negative examples from closely related articles in PubMed).
- Explored features: PSI-MI, MeSH, BioLexicon, authors, journal, institutions, bigrams, POS tagging, NER (genes, proteins, organisms), interaction terms

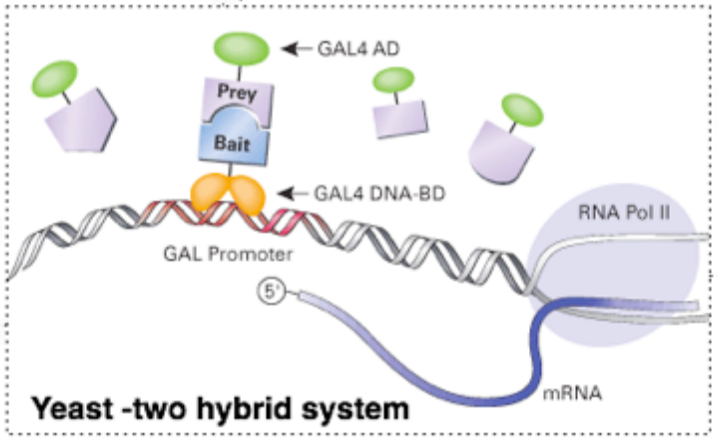
Conclusions & Outlook

- Participants could generate competitive enough results to make their systems useful for improving the PPI curation pipeline
- Could be used to score the abstracts of the most relevant journals for biocuration
- Also true ambiguity in some cases for humans
- Evaluation of participating submissions against each of the three curators individually as well as against BioGRID/MINT classified subset of test set
- Analysis of efficiency in terms of curation time saved by using these systems
- Need of online availability annotation servers
- Combined system seems to increase performance

IMT: Interaction Method Task







IMT: Interaction Method Task

- Interaction detection Methods are important as evidential qualifier for PPIs
- Standardized vocabulary and ontology for formalizing the concepts relevant for experimental methods used to characterize PPI methods (PSI-MI).
- Return ranked list of PSI-MI identifiers: interaction detection method subset.
- Comparison between the automatically generated results and the manual annotations generated by BioGRID and MINT database curators

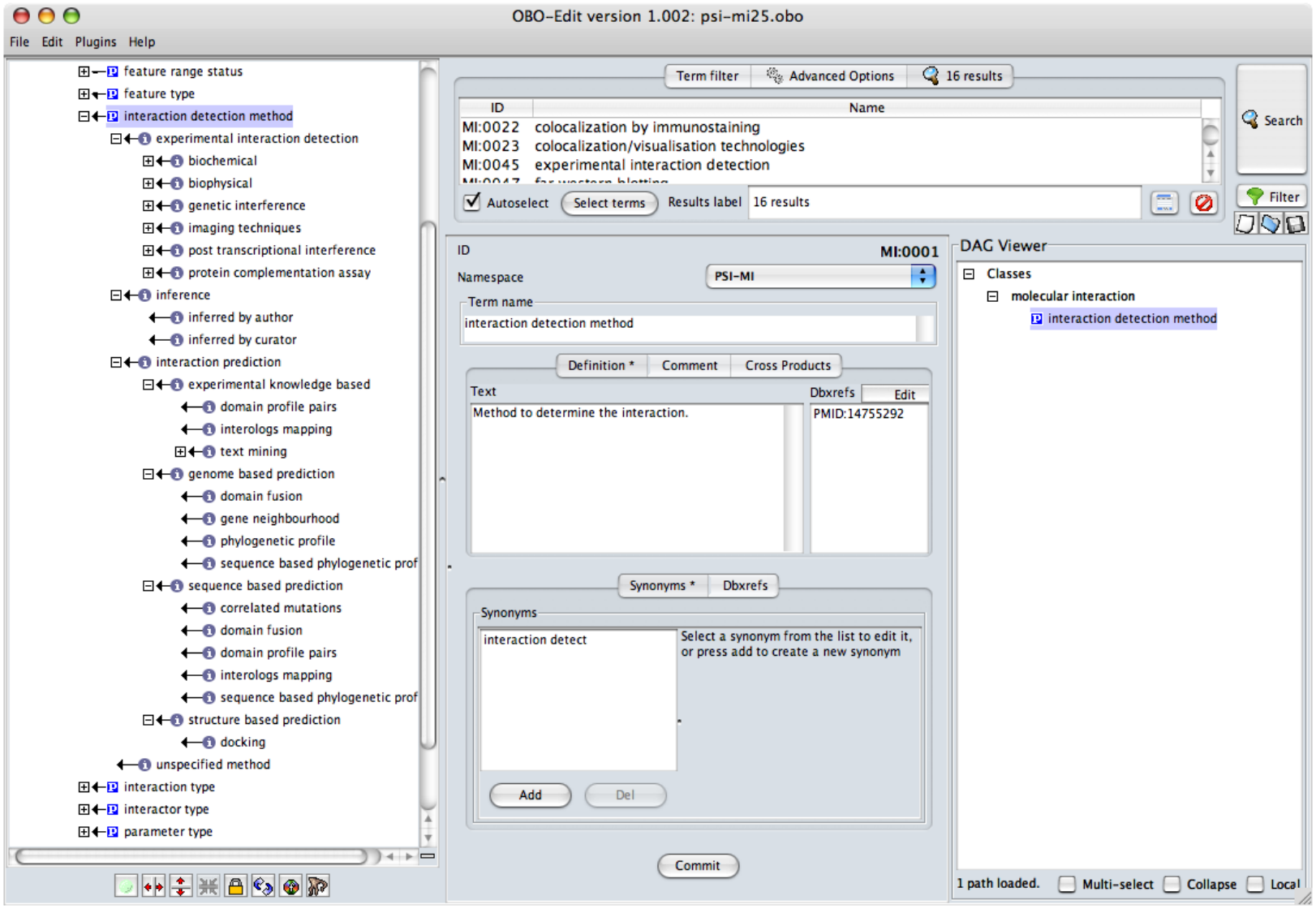
Article ID ➠ PSI-MI Id ➠ [Rank ➠] Confidence ➠ Evidence Text

Provide textual evidence passage for human interpretation

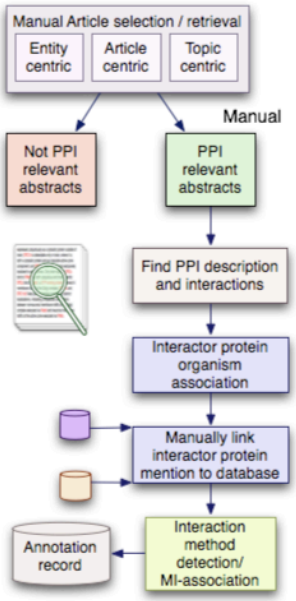
IMT results are to be returned in six tab-separated columns, consisting of:

1. Article identifier
2. Interaction Detection Method MI identifier
3. Unique rank in the range [1..N], where N is the total number of hits for that article.
4. Confidence for that concept in the range]0..1], i.e., excluding zero-confidence.
5. Evidence string (max 500 characters) derived from the full text paper

IMT: Annotation granularity

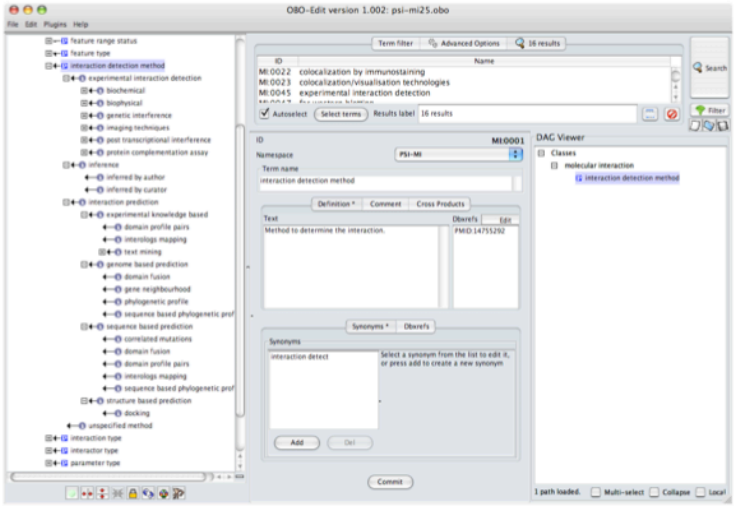


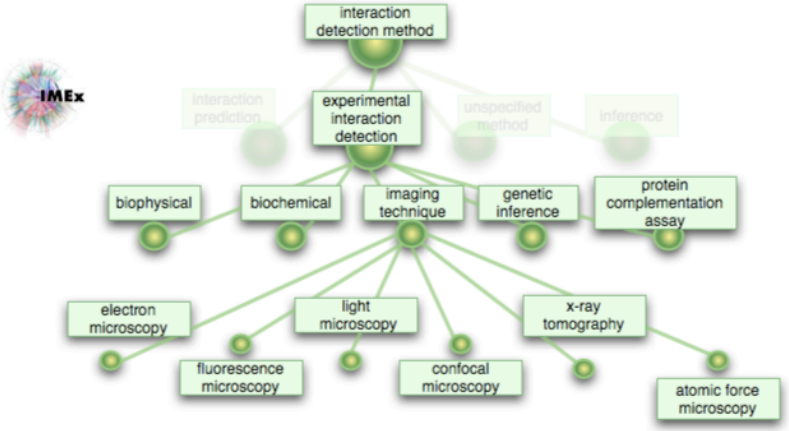
3. PPI ANNOTATION & PSI-MI ONTOLOGY



Sentences for text-mining - textual evidence of manually curated interactions.
(*IntAct (Hermjakob et al.,* http://www.ebi.ac.uk/intact)

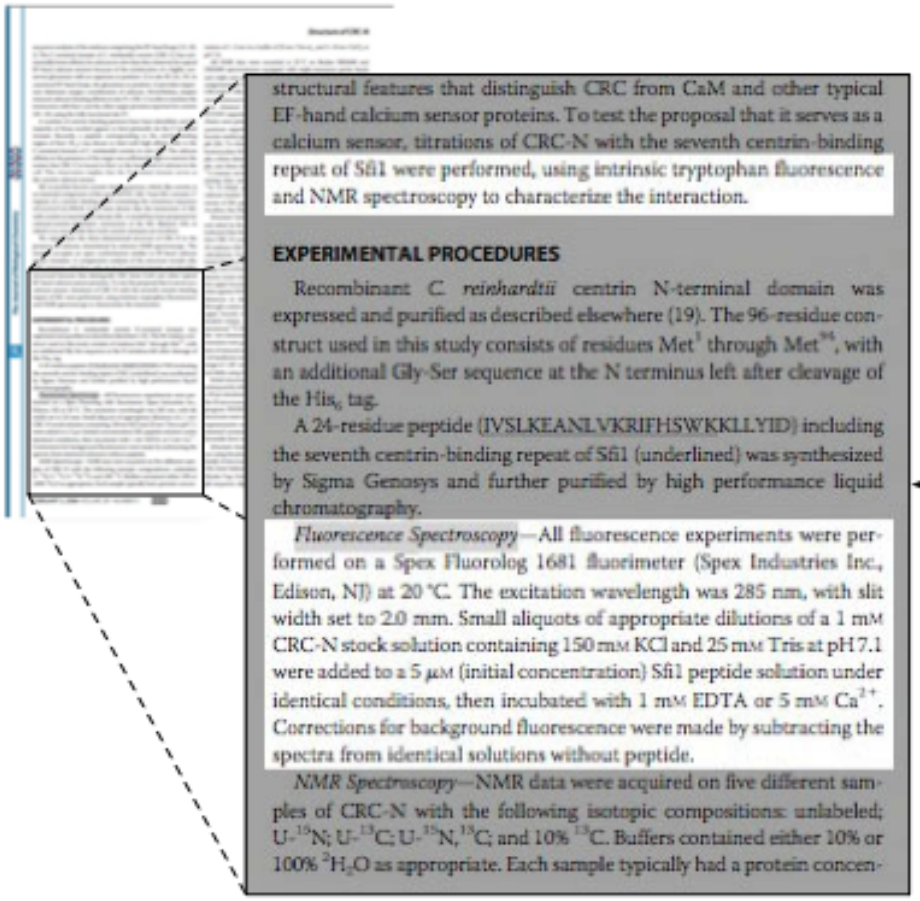
10068665,EBI-1555892,"An immunoblot of the GST-SH2 domain pull-down eluates, incubated with the anti-SHIP MoAb P1D7, showed a clear phosphorylation-dependent interaction of the p85 C-terminal SH2 domain with 145 kD SHIP (Fig 8, lanes 5 and 6). Identical immunoblots incubated with the monoclonal anti-SHIP antibody P2C6 also indicated an interaction of the p85 C-terminal SH2 domain with 145 kD SHIP"
10215621,EBI-1392955,we carried out GST pull-down experiments with 35S-labeled Exd proteins (Fig. 6). GST-Hth bound Exd
10409688,EBI-935417,binding of recombinant eIF4E to immobilized GST-eIF4G1 fusion proteins.
10564262,EBI-1560775,GST-Apl2p (1) bound clathrin heavy chain





A non-redundant collection of interaction evidence passages From IntAct database linked to interaction detection methods (MI-Identifiers) as defined by the PSI-MI was constructed to derive interactor protein affixes (and interaction patterns).

Articles



structural features that distinguish CRC from CaM and other typical EF-hand calcium sensor proteins. To test the proposal that it serves as a calcium sensor, titrations of CRC-N with the seventh centrin-binding repeat of Sfi1 were performed, using intrinsic tryptophan fluorescence and NMR spectroscopy to characterize the interaction.

EXPERIMENTAL PROCEDURES

Recombinant *C. reinhardtii* centrin N-terminal domain was expressed and purified as described elsewhere (19). The 96-residue construct used in this study consists of residues Met^{1} through Met^{94}, with an additional Gly-Ser sequence at the N terminus left after cleavage of the His_6 tag.

A 24-residue peptide (IVSLKEANLVKRIFHSWKKLLYID) including the seventh centrin-binding repeat of Sfi1 (underlined) was synthesized by Sigma Genosys and further purified by high performance liquid chromatography.

Fluorescence Spectroscopy—All fluorescence experiments were performed on a Spex Fluorolog 1681 fluorimeter (Spex Industries Inc., Edison, NJ) at 20 °C. The excitation wavelength was 285 nm, with slit width set to 2.0 mm. Small aliquots of appropriate dilutions of a 1 mM CRC-N stock solution containing 150 mM KCl and 25 mM Tris at pH 7.1 were added to a 5 μM (initial concentration) Sfi1 peptide solution under identical conditions, then incubated with 1 mM EDTA or 5 mM Ca^{2+}. Corrections for background fluorescence were made by subtracting the spectra from identical solutions without peptide.

NMR Spectroscopy—NMR data were acquired on five different samples of CRC-N with the following isotopic compositions: unlabeled; U-^{15}N; U-^{13}C; U-^{15}N,^{13}C; and 10% ^{13}C. Buffers contained either 10% or 100% $^{2}H_2O$ as appropriate. Each sample typically had a protein concen-

PSI-MI 2.5 Ontology

- interaction detection method
  - experimental interaction detection
    - biophysical
      - circular dichroism
      - electron resonance
      - fluorescence technology
        - bioluminescence resonance energy transfer
        - classical fluorescence spectroscopy
        - fluorescence correlation spectroscopy
        - fluorescence polarization spectroscopy
        - fluorescence-activated cell sorting
        - fluorescent resonance energy transfer
        - homogeneous time resolved fluorescence
      - isothermal titration calorimetry
      - light scattering
      - mass spectrometry studies of complexes
      - molecular sieving
      - nuclear magnetic resonance
      - scintillation proximity assay
      - surface plasmon resonance
      - x-ray crystallography
    - protein complementation assay
    - genetic interference
    - post transcriptional interference
    - biochemical
    - imaging techniques

IMT data sets

TRAINING SET

Total nr. articles: 2003
Unique PSI-MI IDs: 86
Total PSI-ID-article links: 4348
Avrg IDs/article: 2.17

DEVELOPMENT SET

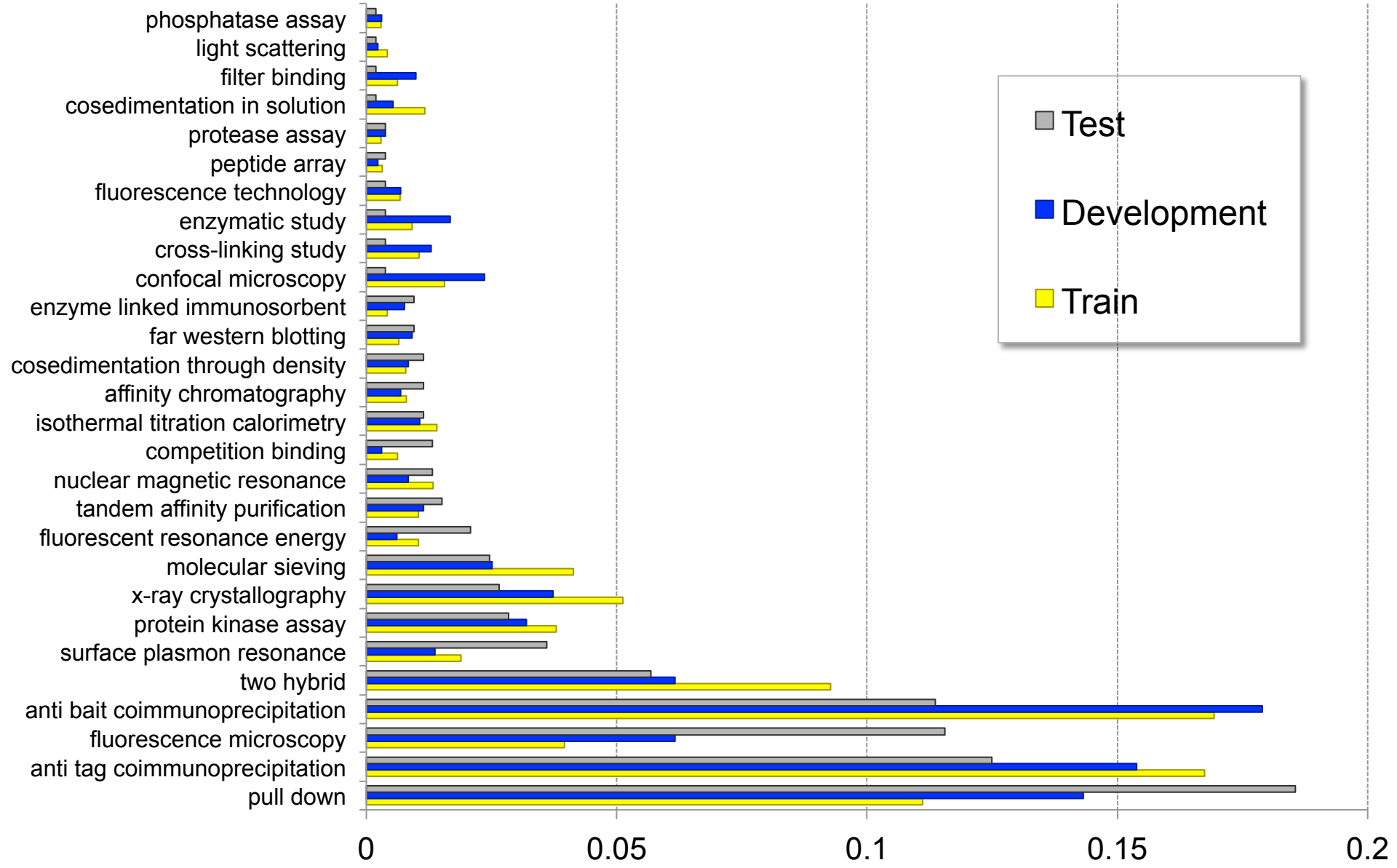
Total nr. articles: 587
Unique PSI-MI IDs: 71
Total PSI-ID-article links: 1316
Avrg IDs/article: 2.24

TEST SET

Total nr. articles: 223
Unique PSI-MI IDs: 46
Total PSI-ID-article links: 528
Avrg IDs/article: 2.36

Total subset of 115 PSI-MI terms

4. Interaction Method Task



IMT participating teams

TEAM	LEADER	INSTITUTION
65	Fabio Rinaldi	University of Zurich
69	Robert Leaman	Arizona State University
70	Sérgio Matos	Universidade de Aveiro, IEETA
81	Luis Rocha	Indiana University
88	Ashish Tendulkar	IIT Madras
89	Shashank Agarwal	University of Wisconsin-Milwaukee
90	Xinglong Wang	National Centre for Text Mining
100	Zhiyong Lu	NCBI\NLM\NIH

- 8 Teams, 42 runs, two teams also submitted online runs

4. Interaction Method Task

Team	Run/Srvr	Docs	Precision	Recall	F1 Score	AUC iP/R
T65	RUN_1	**222**	9.35%	**83.21%**	0.16322	0.47884
T65	RUN_2	**222**	2.45%	**100.00%**	0.04750	0.44034
T65	RUN_3	**222**	9.99%	79.38%	0.17163	0.47650
T65	RUN_4	**222**	33.48%	42.88%	0.35403	0.30927
T65	RUN_5	**222**	*2.44%*	**100.00%**	*0.04735*	0.50111
T69	RUN_1	214	54.87%	57.91%	0.52392	0.52112
T69	RUN_2	211	57.01%	57.35%	0.53415	0.51844
T69	RUN_3	203	60.24%	56.41%	0.54454	0.51470
T69	RUN_4	199	62.46%	55.17%	**0.55060**	0.51013
T69	RUN_5	190	64.24%	52.44%	0.54354	0.49390
T70	RUN_1	143	51.78%	35.01%	0.37838	0.31402
T70	RUN_2	72	71.76%	36.81%	0.45608	0.36215
T70	RUN_3	*30*	**80.00%**	41.50%	0.51508	0.41500
T70	RUN_4	205	31.65%	38.72%	0.31747	0.32295
T70	RUN_5	159	36.36%	*21.26%*	0.24754	0.18976
T81	RUN_1	**222**	4.44%	63.91%	0.08191	0.22022
T81	RUN_2	221	9.39%	41.92%	0.14117	0.19766
T81	RUN_3	**222**	13.51%	28.35%	0.17414	*0.17010*
T81	RUN_4	**222**	13.21%	29.57%	0.17341	0.20388
T81	RUN_5	209	21.93%	24.64%	0.21339	0.18733
T88	RUN_1	219	29.10%	45.04%	0.33601	0.38590
T88	RUN_2	220	28.67%	45.53%	0.33353	0.38373
T89	RUN_1	200	54.78%	53.37%	0.50905	0.46061
T89	RUN_2	200	54.95%	53.23%	0.50760	0.46423
T89	RUN_3	201	54.05%	53.25%	0.50234	0.45330
T89	RUN_4	199	54.48%	54.18%	0.51254	0.47211
T89	RUN_5	201	55.30%	56.12%	0.52377	0.47807
T89	SRVR_4	200	55.33%	55.61%	0.52112	0.47636
T89	SRVR_5	199	54.09%	54.00%	0.50962	0.47650
T89	SRVR_6	201	55.14%	56.12%	0.52350	0.48047
T89	SRVR_7	203	50.46%	55.66%	0.50064	0.47392
T89	SRVR_8	199	54.04%	54.05%	0.50840	0.47534
T90	RUN_1	200	56.11%	51.59%	0.50720	0.44687
T90	RUN_2	203	56.37%	53.19%	0.51203	0.47159
T90	RUN_3	217	55.29%	59.90%	0.54616	**0.52974**
T90	RUN_4	177	63.98%	46.89%	0.51355	0.44118
T90	RUN_5	164	66.26%	46.78%	0.52021	0.44458
T100	RUN_1	213	47.26%	54.97%	0.47062	0.43312
T100	RUN_2	**222**	41.19%	54.61%	0.44178	0.43238
T100	RUN_3	**222**	35.29%	45.53%	0.37496	0.32459
T100	RUN_4	**222**	35.29%	45.53%	0.37496	0.32459
T100	RUN_5	125	56.40%	30.65%	0.37011	0.29387
Team	**Run/Srvr**	**Docs**	**Precision**	**Recall**	**F1 Score**	**AUC**

AUC iP/R: Area under the interpolated precision/recall

Macro-averaged

TEAM & RUN	AUC iP/R
T90_RUN_3	0.52974
T69_RUN_1	0.52112
T69_RUN_2	0.51844
T69_RUN_3	0.5147
T69_RUN_4	0.51013
T65_RUN_5	0.50111
T69_RUN_5	0.4939
T89_SRVR_6	0.48047
T65_RUN_1	0.47884
T89_RUN_5	0.47807
T65_RUN_3	0.4765
T89_SRVR_5	0.4765
T89_SRVR_4	0.47636
T89_SRVR_8	0.47534
T89_SRVR_7	0.47392

Micro-averaged

TEAM & RUN	AUC iP/R
T90_RUN_3	0.35423
T69_RUN_1	0.34302
T69_RUN_2	0.33824
T69_RUN_3	0.32539
T69_RUN_4	0.31711
T89_SRVR_6	0.30049
T89_SRVR_5	0.30046
T89_RUN_5	0.2998
T89_SRVR_4	0.29926
T89_SRVR_8	0.29766
T69_RUN_5	0.29373
T89_SRVR_7	0.29303
T89_RUN_4	0.2922
T65_RUN_5	0.29016
T89_RUN_2	0.28589

F1 Score

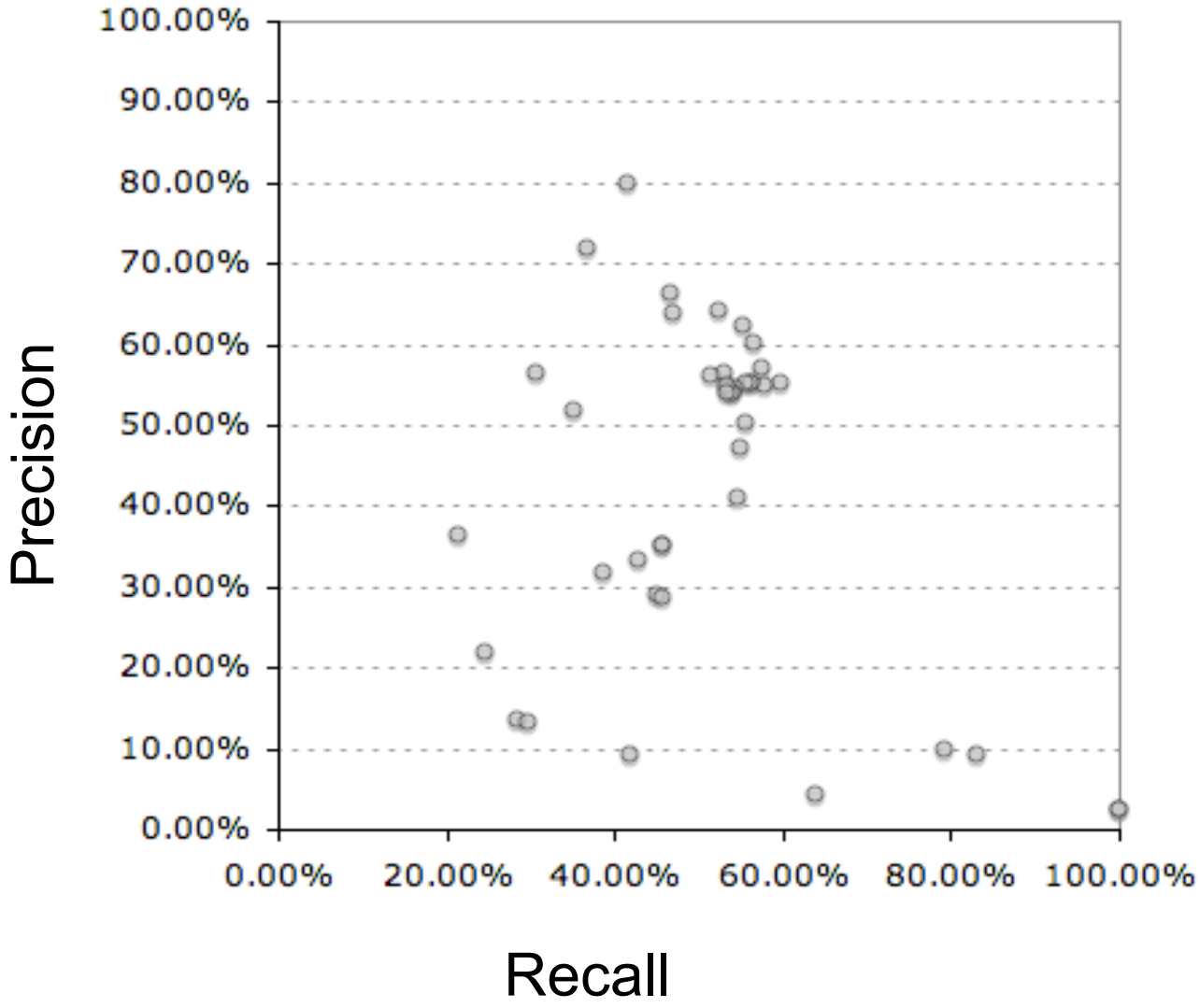
Macro-averaged

TEAM & RUN	F1 score
T69_RUN_4	0.5506
T90_RUN_3	0.54616
T69_RUN_3	0.54454
T69_RUN_5	0.54354
T69_RUN_2	0.53415
T69_RUN_1	0.52392
T89_RUN_5	0.52377
T89_SRVR_6	0.5235
T89_SRVR_4	0.52112
T90_RUN_5	0.52021
T70_RUN_3	0.51508
T90_RUN_4	0.51355
T89_RUN_4	0.51254
T90_RUN_2	0.51203
T89_SRVR_5	0.50962

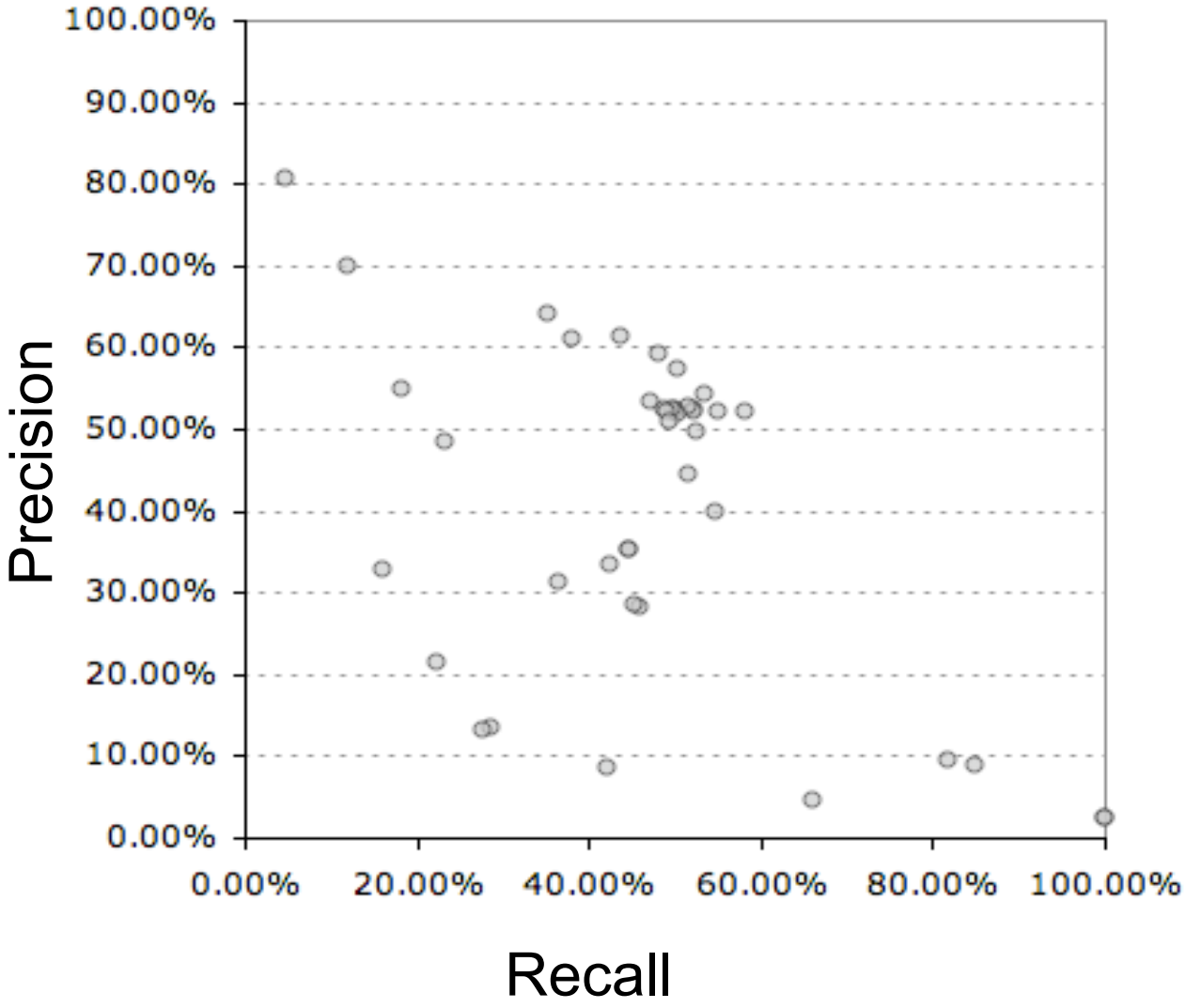
Micro-averaged

TEAM & RUN	F1 score
T90_RUN_3	0.55117
T69_RUN_2	0.5392
T69_RUN_3	0.53589
T69_RUN_1	0.53506
T69_RUN_4	0.5304
T89_RUN_5	0.52381
T89_SRVR_6	0.52232
T89_SRVR_4	0.52157
T89_RUN_4	0.51167
T89_SRVR_5	0.51163
T89_SRVR_7	0.51013
T89_SRVR_8	0.51011
T69_RUN_5	0.50998
T89_RUN_1	0.50977
T90_RUN_2	0.50591

Precision vs recall (macro-averaged)



Precision vs recall (micro-averaged)



‘Easier’ methods

Method	Nr. Articles	average TP	average FN	average FP
pull down	98	***44.07***	53.93	30.10
anti tag coimmunoprecipitation	66	***33.60***	32.40	63.17
anti bait coimmunoprecipitation	60	***29.83***	30.17	69.45
two hybrid	30	***18.33***	11.67	29.57
surface plasmon resonance	19	***12.55***	6.45	14.29
coimmunoprecipitation	51	***8.95***	42.05	23.12
fluorescence microscopy	61	***7.69***	53.31	16.26
fluorescent resonance energy transfer	11	***6.93***	4.07	17.90
x-ray crystallography	14	***6.90***	7.10	25.19
protein kinase assay	15	***4.55***	10.45	18.81

‘Easier’ methods

Method	Nr. Articles	average TP	average FN	average FP
pull down	98	44.07	53.93	30.10
anti tag coimmunoprecipitation	66	33.60	32.40	63.17
anti bait coimmunoprecipitation	60	29.83	30.17	69.45
two hybrid	30	18.33	11.67	29.57
surface plasmon resonance	19	12.55	6.45	14.29
coimmunoprecipitation	51	8.95	42.05	23.12
fluorescence microscopy	61	7.69	53.31	16.26
fluorescent resonance energy transfer	11	6.93	4.07	17.90
x-ray crystallography	14	6.90	7.10	25.19
protein kinase assay	15	4.55	10.45	18.81

Example TP prediction

['19056683', 'MI:0018', '2', '0.9985588388335134', 'TP53INP2 Interacts with GABARAP and GABARAP-like2 Proteins Proteins interacting with TP53INP2 were identi\xef\xac\x81ed by yeast two-hybrid screening of a HeLa cDNA library.\n'] ['two hybrid', '2 hybrid', '2-hybrid', '2H', '2h', 'classical two hybrid', 'Gal4 transcription regeneration', 'two-hybrid', 'yeast two hybrid']

‘Intermediate’ cases

['19741093', 'MI:0096', '2', '0.327747', 'loaded onto an SDS - PAGE gel for Western blot analysis . Figure 1 . GST - Apl5 - ear binds HOPS subunits . Pulldowns on GSTApl5 - ear resin were performed as described (see Materials and Methods) with 150 OD600 nm ml of] ['pull down']

['19218236', 'MI:0096', '1', '0.668406', 'due to the bridging effect of SirT1 (see below) . In an in vitro binding assay , GST - DBC1 efficiently pulled down in vitro - translated SUV39H1 (Fig. 2b) , suggesting that the binding is a direct interaction . These '] ['pull down']

['18625238', 'MI:0114', '3', '0.557822', 'helix in the EF loop (Leu63 - Ala65) . CHIR - AB1 forms homodimers Although the CHIR - AB1 protein used for crystallization was purified from the monomeric peak , crystal packing created a symmetric CHIR - AB1 dimer in which residues \n'] ['x-ray crystallography', 'X-ray', 'x-ray diffraction']

‘Difficult’ cases

['19481529', 'MI:0424', '1', '0.630389', 'phosphorylated Ser437Ala mutant , suggesting phosphorylation of PACS-2 Ser437 was required for binding 14-3-3 proteins . We then conducted a fluorescence polarization assay to determine quantitatively whether phosphorylated'] ['protein kinase assay']

['18922473', 'MI:0006', '2', '0.472072315860236', 'Interaction between the endogenous TRAF6 and TAK1 in AML12 cells as determined by immunoprecipitation with anti - TAK1 antibody , followed by anti - TRAF6 Western blot . The TGF - \xce\xb2 treatment was for 30 minutes and the total rabbit IgG \n'] ['anti bait coimmunoprecipitation', 'anti bait coip']

Overview of participating systems: BioNLP methods

- Most teams used the provided PDF to text conversions
- A considerable fraction carried out some sort of preprocessing and sentence splitting.
- Several different supervised models used: 2 SVM, 2 Logistic regressions, naïve Bayes, random forest, decision tree, KNN.
- Most teams able to provide proper scores/ranks.
- Few teams made use of ontological relationships
- Few carried out NER for genes or organisms
- A couple of teams expanded training set of MINT and IntAct database content
- Most expanded the dictionary with additional synonyms

IMT Discussion & conclusions

Main difficulties relate to the range of different expressions that may refer to a given experimental method, handling PDF articles, heterogeneous journal composition

Some methods can be used in other context that are not PPI relevant

Some methods terms/acronyms are ambiguous (e.g. 2H or CD)

Complexity to mapping to the right granularity of terms in the ontology

Importance of evidence passages for human interpretation

Use of method task for quick filtering of relevant articles and to improve retrieval of experimental qualifiers for PPI

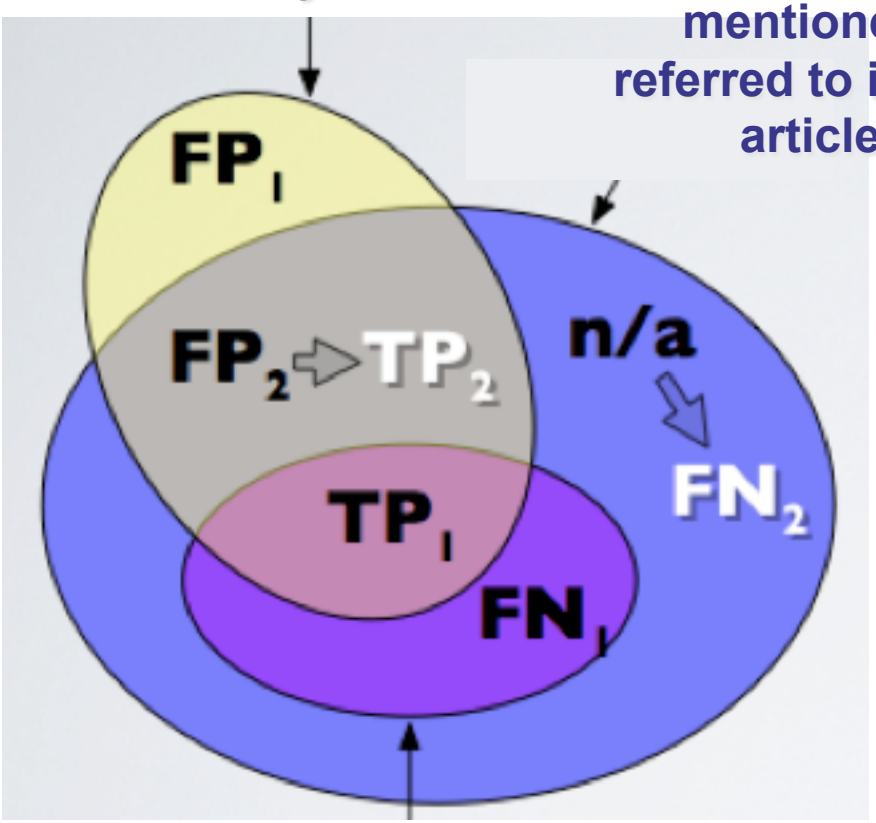
Assist PPI databases in the method annotation

Tools need to be available

IMT: Interaction Method Task

Results from Automated systems

All methods mentioned/ referred to in the article



Methods experimentally supporting PPI

PPI Task Conclusions

BCIII tasks addressed relevant aspects for both database curators as well as general biologists

Provided a large training, development and test set collection

The classification of PPI relevant abstracts using participating systems is useful to improve the selection or relevant articles for Database curators and biologists.

Need of systems to be accessible online

ACT systems can decrease considerable the manual selection time Of relevant documents

Additional text-based annotations needed for improving the systems

Acknowledgements

Databases: BioGRID (Andrew Chatr-aryamontri, Andrew Winter) and MINT (Livia Perfetto, Luana Licata, Marta Iannuccelli, Leonardo Briganti, Gianni Cesareni) for preparing the data annotations

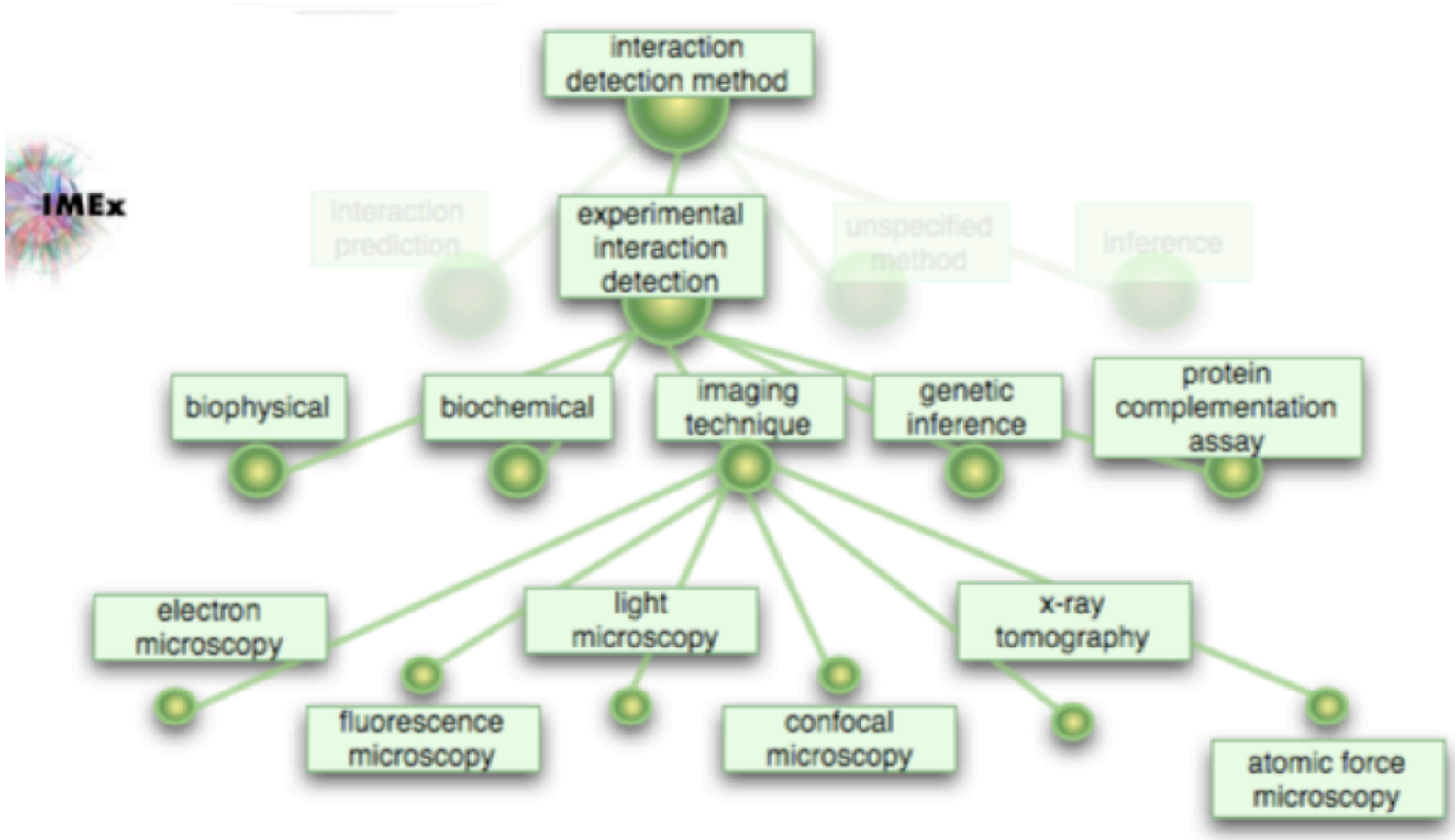
Participants: for implementing their systems and submitting their predictions

CNIO (Florian Leitner, Miguel Vazquez, Alfonso Valencia)

Publishers: Elsevier, Wiley, NPG, Rockefeller University Press, American Society for Biochemistry and Molecular Biology, American Society of Plant Biologists

BCIII Organizers: for organizing this event, feedback and coordination.

Thanks!


interaction detection method
IMEx
experimental interaction detection
biophysical
biochemical
imaging technique
genetic inference
protein complementation assay
electron microscopy
light microscopy
x-ray tomography
fluorescence microscopy
confocal microscopy
atomic force microscopy