Towards Automatic Detection of Experimental Methods from Biomedical Literature

Thomas Kappeler, Simon Clematide, Kaarel Kaljurand, Gerold Schneider, Fabio Rinaldi*
Institute of Computational Linguistics, University of Zurich, Switzerland
kappeler@bluewin.ch,
{siclemat, kalju, gschneid, rinaldi}@ifi.uzh.ch

Abstract

In this paper we present techniques aimed at detecting, within scientific papers which describe newly discovered protein interactions, the methods used by the authors of the research to experimentally verify the interaction(s).

We compare previous results over the BioCreAtIvE data set with more recent results over a larger data set, using INTACT annotations as gold standard. This comparison shows the generality of the proposed approach and suggests that practical application of these techniques within a curation environment might not be that far away.

1 Introduction

Protein interactions play fundamental roles in biological processes (e.g. signal transduction). Biologists routinely perform experiments in order to detect or confirm protein interactions. In doing so, they use a variety of experimental methods.

Databases such as INTACT (Kerrien et al., 2006) or MINT (Zanzoni et al., 2002) aim at collecting the known interactions from the literature. The process of extracting selected items of information from the published literature in order to store such items in databases is known as "curation". This is a costly and time-consuming process, which still requires a significant amount of human resources to be performed effectively.

Tools that can support the process of curation would be extremely welcome by the community. Such tools should be capable of detecting within the papers, with high reliability, all the information that the curators need to create database records.

Repositories of protein interactions, such as INTACT and MINT, store, together with each interaction, a reference to the experimental method that was used to detect it, because this information is highly relevant to researchers. Therefore, not only the protein interactions, but also the experimental methods, need to be identified.

There is a limited number of available experimental methods for the detection of protein interactions, which are all described within the PSI-MI taxonomy (Hermjakob et al., 2004), in particular under node MI:0001 (interaction detection method). Each method is provided with a unique numerical identifier, a standard name, a definition and a list of synonyms.[1]

In order to stimulate research aimed at developing tools that support the extraction of critical information from the literature, the recent BioCreAtIvE text mining competition set up a number of tasks which partially simulate the process of curation. In particular the Protein-Protein Interaction task (PPI) was organized in four subtasks (Krallinger et al., 2008): PPI-IAS (identification of abstracts which contains curatable protein-protein interactions), PPI-IPS (identification of protein-protein interactions in abstracts), PPI-ISS (identification of sentences which provide evidence for protein-protein interactions), PPI-IMS (identification of the experimental method by means of which the interaction was verified). Our own participation to BioCreAtIvE focused on the IPS and IMS subtasks (Rinaldi et al., 2008).

In this paper we describe recent experiments aimed at testing the coverage of the IMS detection approach across a larger set of articles.

*Corresponding author

[1] For the experiments described in this paper we used version 2.5 of PSI-MI.

MI:0096 (pull down)	20.6%
MI:0007 (anti tag coip)	13.1%
MI:0018 (two hybrid)	12.7%
MI:0006 (anti bait coip)	12.1%
MI:0019 (coip)	8.8%
total	67.3 %

Table 1: The 'Big5': most frequently occurring methods in the BioCreAtIvE training data

Run	R	P	F
run 1	29.4%	65.4%	40.6%
run 2	56.8%	43.5%	49.3%
run 3	53.9%	51.3%	52.6%
run 1	20.02%	66.79%	30.81%
run 2	43.02%	40.34%	41.64%
run 3	40.96%	49.65%	44.89%

Table 2: Above: our best results over BioCreAtIvE training data. Below: our official results over BioCreAtIvE test data[3]

2 Detection of Experimental Methods

In the original BioCreAtIvE setting, the organizers asked the participants to deliver the methods coupled with the interactions to which they apply. Due to the intrinsic difficulty of the problem, coupled with the difficulty of finding the interactions, the task was later relaxed, and the participants were asked to deliver a set of experimental methods employed in the article.

The approach we used in BioCreAtIvE for the detection of the experimental methods is based on pattern matching supplemented by simple statistics. As it would have been impossible to manually develop search patterns for all 155 methods in PSI-MI, we first observed the distribution of methods in the training data. The 5 most frequently used methods alone form 67.3% of the unique pairs of methods and articles (see table 1). So we decided to focus on these methods for handcrafted patterns, informed by biological insights, and derive the rest of the patterns automatically from PSI-MI by the following process: (A) extraction of names and synonyms from PSI-MI, (B) derivation of patterns by automatic generation of variants by inclusion/deletion of spaces, tabs, newlines, returns, hyphens, etc. and allowing free variation of uppercase and lowercase.

As expected, the results of these automatically generated patterns were bad, especially for precision. Therefore, handcrafted patterns for the most frequent methods[2] were developed by our team's computational linguist and biologist in an iterative process of identifying undetected articles (false negatives), manually finding hints for methods, constructing patterns, and testing them. This process was most successful for MI:0007 (anti tag coimmunoprecipitation), MI:0006 (anti bait coimmunoprecipitation), and MI:0019 (coimmunoprecipitation). As the automatically generated patterns for MI:0096 (pull down) and MI:0018 (two hybrid) were already quite good, the handcrafted patterns did not perform much better. The approach leads to good recall but low precision (R=73.4%, P=24.3%, F=36.5%), over all file-method-pairs in the training data.

As an example of a handcrafted pattern, consider the method MI:0428, which is named "imaging techniques" in PSI-MI 2.5. This name is not actually used by authors, however strings beginning with "colocaliz" or "colocalis" (allowing hyphens and spaces within the string) are a very good indicator for this method.[4]

At this point, rather than focusing on improving the patterns, it was decided to consider the results obtained (methods for a given file) as a set of candidates, which could be filtered with statistical means.[5] A reduction from about 6.8 candidate methods (per file) to about 2.2 (as in the training data) seemed most promising. For this reduction, an empirically derived formula connecting the frequency of the method in the data and the quality of our patterns for this method was

[2]The five methods in table 1 plus MI:0428 (imaging techniques), because of low recall of the automatically generated patterns, and MI:0401 (biochemical), because of low precision.

[3]The results were evaluated by the organizers according to different criteria. We have chosen here the evaluation which corresponds to the approach used to compute the results presented in this paper (aiming at maximizing the F-score)

[4]This pattern could actually be derived from the names of several obsolete precursors (MI:0021, MI:0022, MI:0023) for MI:0428.

[5]Actually the main reason for this is the conceptual difference between "finding every mention of a method" (which our patterns already did with good precision) and finding all interaction detection methods in a file i. e. identifying the methods used by the authors to detect protein-protein interactions. The statistics are a simple way to give more importance to methods which are unlikely to be just mentioned without a connection to the detected interactions.

INTACT	BCMS	OWN	Journal
615	5958	5513	The Journal of biological chem.
280	583	0	Cell
170	1142	910	PNAS
147	1290	931	Molecular and cellular biology
143	1048	804	The EMBO journal
143	572	0	Nature
88	437	0	Science
87	626	0	Biochem. and biophys. res.com.
86	298	0	Molecular cell
75	359	0	Genes & development
58	432	102	Biochemistry
56	527	375	Oncogene
55	261	0	Journal of molecular biology
54	526	445	The Journal of cell biology
...	...	0	...
3260	22804	9080	Total

Table 3: Journal frequencies in INTACT, BCMS and our own dataset

used. For each method M we compute the following weight:

$w_M = f_M * \frac{p_M^2}{r_M^2}$

where f_M is the relative frequency of method M, while p_M and r_M are precision and recall of all patterns for method M.

The candidate methods were ranked according to their weights. We submitted 3 official runs (where the results of IMS were coupled with the results of IPS) and 3 non-official runs (where the results of IMS were not coupled with the results of IPS). Of these runs, **run 1** was maximizing precision (by giving only the best candidate and so hurting recall for all papers containing more than one method), **run 2** was maximizing recall (giving the three best candidates, so hurting precision for all papers containing one or two methods) and **run 3** was maximizing F-score (additional condition that candidates 2 and 3 reached a minimum in frequency and precision). Our best results for the training data and the official runs for the test data of BioCreAtIvE are shown in table 2.

One of the possible criticism to our approach is that the usage of methods might be time-dependent. In other words, it is reasonable to assume that some methods might be frequently used in some periods and then might go 'out of fashion', perhaps because newer and better methods take their place.

3 Evaluation

After the end of BioCreAtIvE the organizers decided to set up a publicly accessible service to give access to some of the systems which performed best in the competition. This work resulted in a meta-server (Leitner et al., 2008), which receives a request from a remote user (either via web interface or via XML-RPC) and forwards the request (via XML-RPC) to specific servers maintained by the participants. The services currently offered by the meta-server are Gene Mention, Gene Normalization, Interaction Article and Taxon Classification. The organizers defined a list of 22804 PubMed papers to be analyzed by each server (which we will call the BCMS dataset).

Interactions	Methods	%
38220	MI:0018 (two hybrid)	25.5
29268	MI:0676 (tap)	19.8
21205	MI:0096 (pull down)	14.4
20509	MI:0397 (two hybrid array)	13.5
12998	MI:0398 (two hybrid pooling)	8.8
11332	MI:0006 (anti bait coip)	7.7
9473	MI:0007 (anti tag coip)	6.4
6331	MI:0399 (2h fragment pooling)	4.3
6089	MI:0363 (inferred by author)	4.1
1842	MI:0004 (affinity chrom)	1.2
...	...	...
147584	total	100%

Table 4: Distribution of methods per interaction in INTACT

Papers	Methods	%
1121	MI:0018 (two hybrid)	34.4
1066	MI:0096 (pull down)	32.7
840	MI:0007 (anti tag coip)	25.8
761	MI:0006 (anti bait coip)	23.4
574	MI:0114 (x-ray diffraction)	17.6
287	MI:0019 (coip)	8.8
251	MI:0416 (fluorescence imaging)	7.7
123	MI:0663 (confocal microscopy)	3.8
120	MI:0424 (protein kinase assay)	3.7
115	MI:0071 (molecular sieving)	3.5
111	MI:0004 (affinity chrom)	3.4
82	MI:0676 (tap)	2.5
...	...	...
3259	total	-

Table 5: Distribution of methods per paper in INTACT[6]

Our initial aim was to offer our IMS tools as an additional service to be integrated in the meta-server, so we started from the BCMS list of articles.We also wanted to be able to test our results against already annotated articles at INTACT (which we will call the INTACT dataset).

The first problem to deal with is that of the format of the input data. Our approach requires the availability of a full-document plain text version of the original article. Initially, we considered using only articles available in PubMed Central, given the standardized XML format which

[6]Notice that one paper can contain multiple methods, so the sum of all values in this table is larger than 100%.

Year	INTACT	INTACT/BCMS	%
1978	1	0	0%
1980	1	0	0%
1987	1	0	0%
1988	4	0	0%
1989	1	0	0%
1990	2	0	0%
1991	2	0	0%
1992	2	0	0%
1993	14	0	0%
1994	21	0	0%
1995	38	4	10.5%
1996	61	11	18.0%
1997	96	25	26.0%
1998	144	33	22.9%
1999	182	54	29.7%
2000	242	58	24.0%
2001	268	67	25.0%
2002	320	80	25.0%
2003	360	64	17.7%
2004	461	66	14.3%
2005	304	50	16.4%
2006	418	131	31.3%
2007	255	6	2.4%
2008	55	0	0%

Table 6: Distribution of INTACT-curated papers per year, and their proportion in the INTACT/BCMS dataset

would definitely simplify conversion to plain text. Unfortunately BCMS has a low overlap with PubMed Central (only 35 articles).

Therefore we decided to implement our own dedicated HTML to text converters for the most frequent journals in BCMS.[7] We focused on journals which appear to have a reasonably standard HTML structure for the articles, and which were easily obtainable from our library service, obtaining a total of 9080 converted articles. Table 3 shows the most frequent journals in INTACT and BCMS, and for each of them the number of articles that we converted ('OWN'). Among the converted articles, 649 are also present in the INTACT set (as of May 31st, 2008). This is the dataset upon which we base our experiments (which in the rest of this paper will be referred to as the INTACT/BCMS dataset).

In INTACT every protein interaction is associated with the papers where it is discussed and with the experimental method that was used to detect it. Table 4 reports the most frequently used methods based on the number of interactions that they are associated with. However, there are some methods which, although used very rarely, can deliver a large number of interactions. One example is MI:0676 (tap), which is used in only 82 papers. In one of them alone (pubmed:16429126) it is associated with 21574 interactions!

[7] Although a generic HTML to text converter could have been used for the application that we describe in this paper, our aim is not only to extract the experimental methods, but also the protein interactions, using a full NLP approach, for which we need a much better conversion.

Year	P (%)	R (%)	F (%)	Big5 (%)
1995	50	66.7	57.2	71
1996	46.9	71.4	56.6	70
1997	47.9	55.7	51.5	68
1998	41.1	55.7	47.3	68
1999	44.9	58.3	50.7	65
2000	47.6	59.6	52.9	69
2001	42.5	58.6	49.3	65
2002	44.2	53.1	48.2	64
2003	44.9	51.6	48.0	62
2004	39.4	48.1	43.3	59
2005	35.7	45.9	40.2	63
2006	33.2	41.6	36.9	60
2007	44.4	61.5	51.6	60
Total	41.2	51.7	45.9	64

Table 7: Performance over INTACT/BCMS data distributed per year of publication. The last column shows the frequency of the 'Big5' experimental methods per year.

Table 5 shows the methods most frequently used, counting only once a method occurring multiple times in the same paper. As our approach delivers the methods per paper (rather than per interaction), these numbers are a more useful guideline to the relative importance of each method.

Using metadata from the corresponding PubMed entry, we get the year of publication of each INTACT paper. Using that information, we can verify how much the methods depend on the year of publication of the paper. Table 6 shows the distribution of INTACT papers by year of publication, and their proportion in our INTACT/BCMS dataset. Despite the relatively recent start of INTACT (2003), the coverage is reasonably good for the years 1997-2007.

Table 7 shows the results of applying the IMS system, as described in the previous section, to the INTACT/BCMS dataset. All tests have been performed using the modality 'max F-score' of the IMS tools, and the results apply to the association article/method (we do not consider yet the association of methods with specific protein interactions). The data provides a sufficiently large time-window, with good distribution for most of the years of observations (with the exceptions of 1995, 2007 and possibly 1996). The results are comparable to those obtained in BioCreAtIvE (both training and test), which are shown in Table 2.

Surprisingly, the value of precision is always

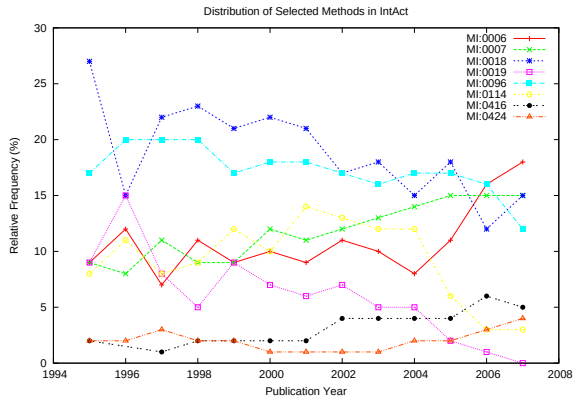

Figure 1: Trends for methods in IntAct

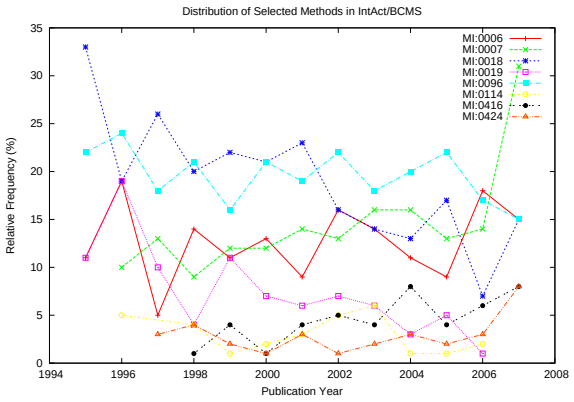

Figure 2: Trends for methods in IntAct/BCMS

lower than recall, which was not expected (in order to maximize the F-score, P and R should be almost equal). This might be caused by the fact that the articles with only one method are more frequent in IntAct (43.2%) than in BioCreAtIvE (34%). There appears to be a decreasing trend in the years 2004 to 2006 (2007 is too small to be representative), which could be caused by the emergence of new experimental methods and reduced usage of methods that were popular in previous years. However, whether this effect is due to a genuine 'aging' of experimental methods, or it is simply due to the selection of articles by IntAct curators, cannot be said on the basis of the available data.

The last column of table 7 shows that the frequency of the Big5 methods declines only slowly, and figure 1 demonstrates that emerging methods, such as MI:0114 (x-ray diffraction), MI:0416 (fluorescence imaging) and MI:0424 (protein kinase assay), take more importance even more hesitantly. Table 7 on the whole confirms that the approach as such seems not endangered by sudden 'revolutions' in the use of experimental methods and a gradual erosion of our results can be contrasted by a periodic reassessment of methods for which handcrafted patterns have to be developed.

4 Discussion

Given the limited set of documents used in our experiments, it is important to ask the question whether the results are sufficiently representative. Since our approach is based upon patterns, each of which is designed to recognize lexical hints to a given experimental method, it is obvious that the approach can be successful only as long as there is no large variation in the relative frequency of methods used in a given set of papers.

4.1 Trends

We have therefore observed the distribution and historical trends of methods in the whole IntAct dataset and compared it with the distribution and historical trends in our own dataset (see figures 1 and 2). Among the 10 most frequently used methods, the number of them shared in both sets is between 6 and 8 for each year. The most frequent methods are the same as in the BioCreAtIvE training data (see table 1). The proportion of these five methods in the IntAct data, distributed per year, is shown in the last column of table 7.

Tables 10 and 11 illustrate the performance of our search patterns for specific methods over the IntAct/BCMS dataset. Of these, the first 5 methods are searched for by handcrafted patterns, the following methods by automatically derived patterns. The disappearance of MI:0019 (coimmunoprecipitation) over time can at least partially be explained by the increase in the use of MI:0007 (anti tag coip) and MI:0006 (anti bait coip). As these are hyponyms of MI:0019, the process we observe may not be an evolution of different scientific practices but actually a semantic process: an increasing preference for the use of a more specific term, be it by the authors of the papers themselves or by curators of IntAct. The identification of 'challengers', i.e. new methods increasing in use (per paper), and so probably deserving handcrafted search patterns, is rather difficult. The most obvious candidate is method MI:0114 (x-ray crystallography), for which the automatically derived pattern does already perform very well, but MI:0416 (fluorescence microscopy) and MI:0424 (protein kinase assay) for

Methods	Papers	%
1	1408	43.2%
2	928	28.5%
3	567	17.4%
4	255	7.8%
5	81	2.5%
6	17	0.5%
7	4	0.1%

Table 8: Number of distinct methods per paper in INTACT

	real			pseudo			oracle
	1	2	3	1	2	3	-
TP	427	705	773	3260	5112	6036	715
FP	222	584	1124	0	1408	3744	642
FN	1069	791	723	3260	1408	484	781
P	65.8	54.7	40.7	100	78.4	61.7	52.7
R	28.5	47.1	51.7	50	78.4	92.6	47.8
F	39.8	56.0	45.6	66.7	78.4	74.1	50.1

Table 9: Comparison of real and simulated experiments

which the performance of the automatically derived patterns is very weak are very promising candidates (see table 11). On a lower level, methods such as MI:0004 (affinity chromatography technology), MI:0047 (far westernblotting) and MI:0071 (molecular sieving) seem to be the most interesting candidates, and the very weak performance for MI:0004 could certainly profit very much from a handcrafted search pattern.

4.2 Independence INTACT/ BioCreative

Since the original program was developed using the BioCreAtIvE training data, it is important to verify that the data on which we are testing do not have a major overlap with BioCreAtIvE training data. Among the whole 'INTACT/BCMS' articles, 453 are not in the BioCreAtIvE training data, 196 are (30.2%). Additionally, 521 of the files BioCreAtIvE training data are not in 'INTACT/BCMS'. The overlap with the BioCreAtIvE test set is less relevant, since it was not used for development of the program, however we report it here to show the independence of the two tests. 522 INTACT/BCMS files are not in BioCreAtIvE-Test, 127 are (19.6%). 231 BioCreAtIvE test files are not INTACT files.

4.3 Choosing the number of methods

The selection of the number of methods for each paper does have an impact on the final results. If the program is set to deliver always only the best ranked method, precision will be relatively good, but recall will be poor. Conversely, if always the 3 best methods are delivered, the opposite will happen. Table 9 shows the results obtained by the system if only the 1st best method is delivered, the 2 best methods, or the 3 best methods ('real').

Another way to observe the impact of the selection of the number of methods on the results is to conduct the following pseudo experiment: suppose we have the perfect ranking algorithm which delivers for each paper a list of all methods correctly ranked for relevance. If, for all papers, we always output only the best method, we are never damaging precision, but we are reducing recall of all but the 1-method-files. If, instead, we take always the two best methods, precision will be lowered by taking many unnecessary 'second best' methods, but we increase recall. Finally, if we decide to assign to all papers the three best methods, precision will be much lower and recall will keep improving. Using the data gathered directly from INTACT we can compute the results, which are also presented in table 9 ('pseudo').

Finally, we can consider the following experiment. Suppose we have an "oracle" which tells us reliably how many methods we should deliver for each paper, how good would be our results? This is a rather realistic scenario, since ideally the method detection program would be coupled with an interaction detection program,[6] therefore knowing how many methods are needed. Although we do not have at the moment a program capable of predicting how many methods should be associated to each paper, we can simulate it with data taken out of INTACT: this will be our "oracle". With such an help, we can filter the results of the method selection and ranking program, obtaining the results that are show in the last column of table 9.

These results show that, although usually our approach delivers the correct ranking for methods, there must be some cases where a correct method is ranked lower than a wrong method. A detailed inspection of these results will provide useful hints for further development.

4.4 Future Directions

The work described in this paper proves that it is possible, with reasonably simple techniques, to capture the most relevant methods with high reliability. Additional improvements to the system

[6] We are developing such a program separately, based on our BioCreAtIvE submission for the PPI-IPS task.

are likely to require complex fine tuning.

As it is impossible to handcraft rules for all the 155 methods, it would be meaningful to investigate how to improve the existing approach via machine learning. To this aim, we performed an experiment with a standard text classifier, using the methods as categories. Although the results were rather disappointing, this might have been due to the poor preparation of the input data. We intend to further investigate if better preprocessing or the usage of more sophisticated classifiers might help overcome these limitations.

The usage of other terminologies/ontologies for the extraction of synonyms (e.g. Mesh) is hampered by the unclear mapping of the relevant entries into PSI-MI entries. Without such a mapping, any attempt at using other dictionary sources would simply increase the level of noise.

Any further evaluation of the results would need to take into account the limitations of the gold standard. If the program finds a method, which has been used by the authors, and it is prominently mentioned in the paper, but it is not included in the gold standard (maybe because it is not directly related to any of the interactions annotated by INTACT curators), then it gets penalized (one FP). As an example, in PubMed 16293613 there are several mentions of "x-ray crystal structure(s)" in connection with the author's experiments, one of these mentions is in the experimental procedures section, which seems to show that method MI:0114 (x-ray crystallography) was used, but this was rated as an FP by comparison with the INTACT gold standard.

As a service to the community, we plan to make available the functionality of method identification as a web service, possibly integrated into the BioCreAtIvE meta-server described in (Leitner et al., 2008). We aim at offering coverage of all PubMed articles for which the full text is freely available, focusing in particular on PubMed Central.

5 Conclusion

We described a system capable of automatically extracting experimental methods for detection of protein interactions from biomedical scientific literature. Participation to the BioCreAtIvE II evaluation has proven the competitiveness of the approach. In this paper we have proven that the range of applicability of the system goes well beyond the scope of the BioCreAtIvE dataset. Reasonable results have been shown over literature spanning the last ten years.

Acknowledgments

We thank the anonymous reviewer for their insightful comments and helpful suggestions. This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, Basel.

References

H Hermjakob, L Montecchi-Palazzi, G Bader, J Wojcik, L Salwinski, A Ceol, S Moore, S Orchard, U Sarkans, C von Mering, B Roechert, S Poux, E Jung, H Mersch, P Kersey, M Lappe, Y Li, R Zeng, D Rana, M Nikolski, H Husi, C Brun, K Shanker, C Grant SG, Sander, P Bork, W Zhu, A Pandey, A Brazma, B Jacq, M Vidal, D Sherman, P Legrain, G Cesareni, I Xenarios, D Eisenberg, B Steipe, C Hogue, and Apweiler R. 2004. The hupo psi's molecular interaction format - a community standard for the representation of protein interaction data. *Nat. Biotechnol*, 22:177–183.

S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. 2006. IntAct - Open Source Resource for Molecular Interaction Data. *Nucleic Acids Research*.

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*. (to appear).

Florian Leitner, Martin Krallinger, Carlos Rodriguez-Penagos, Jörg Hakenberg, Conrad Plake, Cheng-Ju Kuo, Chun-Nan Hsu, Richard Tzong-Han Tsai, Hsi-Chuan Hung, William W. Lau, Calvin A. Johnson, Rune Saetre, Kazuhiro Yoshida, Yan Hua Chen, Sun Kim, Soo-Yong Shin, Byoung-Tak Zhang, William A. Baumgartner, Lawrence Hunter, Barry Haddow, Michael Matthews, Xinglong Wang, Patrick Ruch, Frédéric Ehrler, Arzucan Özgür, Günes Erkan, Dragomir R. Radev, Michael Krauthammer, ThaiBinh Luong, Robert Hoffmann, Chris Sander, and Alfonso Valencia. 2008. Introducing meta-services for biomedical information extraction. *Genome Biology*.

Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. 2008. OntoGene in BioCreative II. *Genome Biology*, 9, Suppl 2:S13.

A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. 2002. MINT: a Molecular INTeraction database. *FEBS Letters*, 513(1):135–140.

Year	tp	fn	fp	P	R	F
MI:0006 (anti bait coimmunoprecipitation)						
1995	1	0	0	100%	100%	100%
1996	2	2	1	66.7%	50%	57.2%
1997	1	2	5	16.7%	33.3%	22.2%
1998	5	5	5	50%	50%	50%
1999	4	9	12	75%	30.8%	43.7%
2000	7	11	12	36.8%	38.9%	37.8%
2001	4	9	9	30.8%	30.8%	30.8%
2002	10	20	8	55.6%	33.3%	41.7%
2003	5	18	7	41.7%	21.7%	28.5%
2004	7	11	13	35%	38.9%	36.8%
2005	2	8	11	15.4%	20%	17.4%
2006	23	33	27	46.0%	41.0%	43.4%
2007	0	2	0	0	0	0
total	71	130	110	39.2%	35.3%	37.1%
MI:0007 (anti tag coimmunoprecipitation)						
1995	0	0	1	0	0	0
1996	1	1	2	33.3%	50%	40%
1997	2	6	3	40%	25%	30.8%
1998	3	3	8	27.3%	50%	35.3%
1999	10	5	10	50%	66.7%	57.2%
2000	10	7	5	66.7%	58.8%	62.5%
2001	10	10	7	58.8%	50%	54.0%
2002	12	14	11	52.2%	46.2%	49.0%
2003	13	12	9	59.1%	52.0%	55.3%
2004	13	12	9	59.1%	52.0%	55.3%
2005	8	6	6	57.1%	57.1%	57.1%
2006	27	14	19	58.7%	65.9%	62.1%
2007	3	1	1	75%	75%	75%
total	112	91	91	55.2%	55.2%	55.2%
MI:0018 (two hybrid)						
1995	3	0	0	100%	100%	100%
1996	3	1	2	60%	75%	66.7%
1997	15	1	4	78.9%	93.8%	85.7%
1998	13	1	5	72.2%	92.9%	81.3%
1999	27	0	8	77.1%	100%	87.1%
2000	28	0	5	84.8%	100%	91.8%
2001	30	2	9	76.9%	93.8%	84.5%
2002	31	0	13	70.5%	100%	82.7%
2003	23	0	10	69.7%	100%	82.1%
2004	20	0	12	62.5%	100%	76.9%
2005	18	1	7	72%	94.7%	81.8%
2006	20	0	12	62.5%	100%	76.9%
2007	2	0	3	40%	100%	57.1%
total	233	6	90	72.1%	97.5%	82.9%
MI:0019 (coimmunoprecipitation)						
1995	2	0	1	66.7%	100%	80.0%
1996	4	0	5	44.4%	100%	61.5%
1997	4	2	12	25%	66.7%	36.4%
1998	2	1	14	12.5%	66.7%	21.1%
1999	8	5	21	27.6%	61.5%	38.1%
2000	6	3	26	18.8%	66.7%	29.3%
2001	6	3	27	18.2%	66.7%	28.6%
2002	6	7	36	14.3%	46.2%	21.8%
2003	7	2	23	23.3%	77.8%	35.9%
2004	2	3	27	6.9%	40%	11.8%
2005	2	3	21	8.7%	40%	14.3%
2006	0	2	69	0%	0%	0%
2007	0	0	4	0%	0%	0%
total	49	31	286	14.6%	61.3%	23.6%
MI:0096 (pull down)						
1995	1	1	1	50%	50%	50%
1996	4	1	1	80%	80%	80%
1997	9	2	3	75%	81.8%	78.3%
1998	12	3	3	80%	80%	80%
1999	16	3	7	69.6%	84.2%	76.2%
2000	24	4	5	82.8%	85.7%	83.7%
2001	23	3	12	65.7%	88.5%	75.4%
2002	33	9	6	84.6%	78.6%	81.5%
2003	27	2	5	84.4%	93.1%	88.5%
2004	28	4	6	82.4%	87.5%	84.9%
2005	18	6	2	90%	75%	81.8%
2006	43	9	18	70.5%	82.7%	76.1%
2007	2	0	1	66.7%	100%	80.0%
total	240	47	70	77.4%	83.6%	80.4%

Table 10: Most frequent methods (Big5): distribution per year

Year	tp	fn	fp	P	R	F
MI:0114 (x-ray crystallography)						
1995	0	0	2	0	0	0
1996	1	0	0	100%	100%	100%
1997	0	0	1	0	0	0
1998	3	0	1	75%	100%	85.7%
1999	1	0	1	50%	100%	66.7%
2000	2	1	4	33.3%	66.7%	44.4%
2001	3	1	4	42.9%	75%	54.6%
2002	8	2	4	66.7%	80%	72.7%
2003	6	4	1	85.7%	60%	70.6%
2004	0	1	2	0	0	0
2005	0	1	4	0	0	0
2006	2	3	12	14.3%	40%	21.1%
2007	0	0	0	0	0	0
total	26	13	36	41.9%	66.7%	51.5%
MI:0424 (protein kinase assay)						
1995	0	0	0	0	0	0
1996	0	0	0	0	0	0
1997	0	2	0	0	0	0
1998	0	3	0	0	0	0
1999	0	2	1	0	0	0
2000	0	1	1	0	0	0
2001	2	2	0	100%	50%	66.7%
2002	0	2	2	0	0	0
2003	0	4	2	0	0	0
2004	1	4	0	100%	20%	33.3%
2005	0	2	0	0	0	0
2006	0	10	2	0	0	0
2007	1	0	0	100%	100%	100%
total	4	32	8	33.3%	11.1%	16.6%
MI:0004 (affinity chromatography technology)						
1995	0	1	0	0	0	0
1996	0	1	0	0	0	0
1997	0	0	0	0	0	0
1998	0	1	1	0	0	0
1999	0	2	0	0	0	0
2000	0	2	0	0	0	0
2001	0	4	1	0	0	0
2002	0	4	1	0	0	0
2003	0	1	0	0	0	0
2004	1	1	2	33.3%	50%	40%
2005	0	4	1	0	0	0
2006	0	4	1	0	0	0
2007	0	0	0	0	0	0
total	1	25	7	12.5%	3.8%	5.8%
MI:0047 (far western blotting)						
1995	0	1	0	0	0	0
1996	0	0	0	0	0	0
1997	2	2	0	100%	50%	66.7%
1998	0	4	0	0	0	0
1999	1	2	0	100%	33.3%	50%
2000	0	1	0	0	0	0
2001	1	2	1	50%	33.3%	40%
2002	0	0	0	0	0	0
2003	0	4	0	0	0	0
2004	0	3	0	0	0	0
2005	0	2	0	0	0	0
2006	1	2	0	100%	33.3%	50%
2007	0	0	0	0	0	0
total	5	23	1	83.3%	17.9%	29.5%
MI:0071 (molecular sieving)						
1995	0	0	0	0	0	0
1996	0	0	0	0	0	0
1997	0	0	0	0	0	0
1998	0	0	2	0	0	0
1999	0	1	1	0	0	0
2000	0	2	2	0	0	0
2001	1	1	2	33.3%	50%	40%
2002	0	2	2	0	0	0
2003	2	4	3	40%	33.3%	36.3%
2004	2	2	3	40%	50%	44.4%
2005	0	2	4	0	0	0
2006	2	2	7	22.2%	50%	30.7%
2007	0	0	0	0	0	0
total	7	16	26	21.2%	30.4%	25.0%

Table 11: Other important methods: distribution per year