

Medea elements, are being developed with the intent of driving genes refractory to malaria and dengue fever into mosquito populations. Gene drive was not an issue that was considered when the terms of the Cartagena Protocol were first negotiated and, as noted in the guidance document, the fact that mosquitoes are a vector of human disease poses “new considerations and challenges during the risk assessment process.” Questions arise as to whether the risks of this technology should be weighed against the potential to control disease on a global scale. These issues must be addressed in a clear and open way, making further discussion essential.

A related issue is the exemption of GM mosquitoes in transit or destined for contained use from the AIA procedure. The AIA procedure was written, in part, with the intent of protecting developing countries against threats to biosafety due to a lack of resources to conduct their own risk assessment. Even so, during negotiations of the protocol, countries with strong biotech industries successfully argued that GM organisms in transit or containment pose negligible risks and thus the AIA procedure would restrict trade unnecessarily if applied to them. For GM mosquitoes with invasive gene drive systems, the risks are non-negligible because breaches of containment are impossible to rule out and, once released, just a few escapees could be capable of spreading transgenes on a global scale. The exemption must therefore be re-examined in these cases.

The scenario of containment is particularly relevant to GM mosquitoes because, before an open release, trials are being discussed that would take place in field cages exposed to the ambient environment in a location that the species naturally inhabits. This is an important step in a phased assessment of risks and efficiency; however, before these trials, developing countries are not entitled to request that the importing country pay for a preliminary risk assessment because the AIA procedure does not apply. This issue is not mentioned in the guidance document and was likely considered to be beyond its scope; however, the Cartagena Protocol clearly provides inadequate protection in this scenario, and further discussion is essential before field trials become a reality.

Another pressing issue hinted at in the guidance document is the inapplicability of the Cartagena Protocol to modified mosquitoes that do not fit the definition of “modern biotechnology.” The protocol applies to living modified (LM) organisms

developed using *in vitro* nucleic acid techniques; however, it does not apply to mosquitoes modified by other means having similar implications for biodiversity and human health.

The most noteworthy variety of non-LM mosquitoes is an *A. aegypti* line infected with the wMelPop strain of *Wolbachia*, an inherited bacterium capable of manipulating its host's reproductive biology in a manner that promotes its spread through a population. As it turns out, this *Wolbachia* infection is associated with several physiological changes beneficial for disease control, including reduced mosquito lifespan, reduced dengue viral load and reduced ability to obtain blood meals with age⁵. However, the existence of physiological changes in conjunction with invasiveness draws into question the wider implications these changes have on biosafety and highlights the fact that biosafety issues are not limited to genetic modification.

To address this issue, the guidance document states that “although the focus of this guidance is on LM mosquitoes, in principle, it may also be useful for the risk assessment of similar non-LM mosquito strategies.” This is an important point; however, it is in no way legally binding. Non-LM mosquitoes are beyond the scope of the Cartagena Protocol. However, given that their biosafety implications are as serious as those for LM mosquitoes, further discussion is needed on how they should be regulated. Even-handed regulation will ensure that one strategy is not chosen

over another purely for its immunity to onerous requirements.

In conclusion, the guidance document of the sub-working group represents an important first step towards incorporating the biosafety issues posed by GM mosquitoes into the Cartagena Protocol. It raises a number of important considerations regarding risk assessment that may be largely adequate for releases of sterile and self-limiting GM mosquitoes. However, for strategies involving mosquitoes capable of replacing entire populations with disease-incompetent varieties, several issues still need to be resolved. For these strategies, a balance must be sought between the precautionary principle, respect for the sovereignty of states and the ethical mandate to prevent disease on a global scale. Further discussion is needed to address the international regulatory challenges posed by GM mosquitoes in working towards the goal of global vector-borne disease control.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

John M Marshall

Department of Infectious Disease Epidemiology,
Imperial College London, London, UK.
e-mail: john.marshall@imperial.ac.uk

1. <http://www.cbd.int/biosafety/protocol.shtml>
2. <http://www.cbd.int/doc/meetings/bs/bsrarm-02/official/bsrarm-02-05-en.pdf>
3. Fu, G., et al. *Proc. Natl. Acad. Sci. USA* **107**, 4550–4554 (2010).
4. Marshall, J.M. & Taylor C.E. *PLoS Med.* **6**, e1000020 (2009).
5. Moreira, L.A. et al. *PLoS NTDs* **3**, e568 (2009).

The *FEBS Letters*/BioCreative II.5 experiment: making biological information accessible

To the Editor:

Current publications lack structured representations of the entities and relationships they report on. As a consequence, information retrieval is hampered and much of the scientific literature is poorly accessible unless it is organized in domain-specific databases by expert curation¹. However, manual curation is a slow process and databases lag behind, failing to cover much of the published information. The combined effort of the IMEx group deals with

only ~20% of the estimated 10,000 protein interaction articles published yearly (Supplementary Methods). To explore new publication strategies, the *FEBS Letters* experiment asked authors to supply structured annotations for their publications that were linked to databases with the intervention of professional bio-curators². The BioCreative II.5 challenge then compared these annotations provided by authors and curators to automated systems³. Combining these two efforts has generated the first quantitative

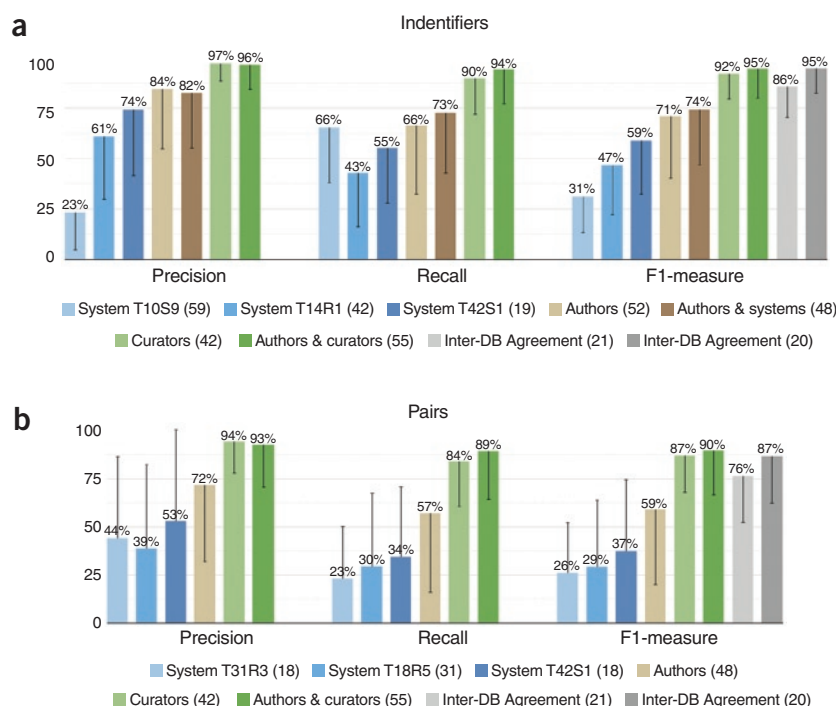


Figure 1 Evaluation results. The performance of annotation sources when (a) identifying an article's interacting proteins and (b) identifying binary protein interaction pairs (both by their (UniProt) database identifiers). Data obtained from five data sources: automatic extraction methodologies ('System T10 S9', 'System T14 R1' and 'System T42 S1'; where T10, T14 and T42 indicate team runs and S and R indicate online and offline submission number, respectively; curators ('Curators'), authors ('Authors'), authors and curators combined ('Authors & Curators'); or authors and extraction methodologies combined ('Authors & Systems'). Figures in parentheses next to each source indicate the number of articles that source annotated. Results are displayed as average precision, recall and the balanced F measure (F_1 -measure) per article (F_1 weighs precision and recall equally; error bars: s.d.) for both protein identifier and interaction pair annotations from all three sources. For reference, inter-annotator agreement (IAA) is shown: 'Inter-DB Agreement' indicates IAA between curators of different databases (light gray bars); 'Intra-DB Agreement' DBs indicates IAA between curators of the same database. Light brown bars show the performance of the authors alone, whereas the dark brown bars are results from an ensemble classifier using both automatic extraction methodologies' and authors' annotations to evaluate the impact of joining both outputs. The test set articles used by the automatic extraction methodologies (Systems; blue bars) do not overlap with the articles annotated by the other sources. The Authors & Systems combination therefore is based on results from the automatic extraction methodologies' training set. The training set, Authors, Curators, Authors & Curators, and Authors & Systems have 33 articles in common. For more details on the distribution of articles, see **Supplementary Methods**.

data to support the debate on ways to supplement publications with structured information.

One way to extend the number of interactions captured by public repositories is to enlist authors in the annotation effort. As suggested by Orchard *et al.*⁴, authors could be asked to submit the relevant information during the editorial process, as defined by the minimum information requirement for reporting protein interaction experiments (MIMIx). The pros and cons of possible approaches for adding structured information to scientific publications have been discussed. Gerstein

*et al.*⁵ and Hahn *et al.*⁶ argued over the extent to which quality, consistency and stable support could be expected from authors, and to what extent automatic systems can help in this process. However, the debate is stalling because of the absence of comparable data about these approaches.

In February 2008, the FEBS editorial board designed an experiment asking authors to provide structured information on protein interactions along the lines of the MIMIx recommendations. Trained curators from the Molecular Interaction (MINT) protein-protein interaction (PPI) database⁷ then reviewed these annotations

and included them in their repository. A structured, human- and machine-readable summary was appended to the traditional abstract as a structured digital abstract (SDA). From March to December 2008, 76 authors were invited to take part in this experiment; 57 accepted, and the published articles contained these SDAs, reporting the identifiers of the interacting proteins, the interaction type (physical interaction, co-localization, enzymatic reaction and so forth) and the experimental method used to verify the interaction.

To complement this study, the BioCreative (<http://www.biocreative.org/>) organizers challenged text-mining researchers to reproduce a subset of the annotations provided by these SDAs. As data for this challenge ('BioCreative II.5'), Elsevier (Amsterdam) provided 122 PPI-describing articles with their SDAs, and over 1,000 negative articles from *FEBS Letters*, which did not describe PPIs with experimental evidence.

The evaluation presented here was designed to assess the performance of automatic extraction methodologies as compared with the results of curators and authors on (i) identification of the binary interaction pairs and (ii) the corresponding identification of the interacting proteins described in the article (by their (UniProt) database identifiers). The results were restricted to those proteins participating in an interaction supported by experimental evidence. Additionally, BioCreative II.5 assessed the capacity of automatic systems to retrieve articles reporting experimentally validated protein interaction information. This third task is of direct interest to database curators in the process of selecting relevant articles.

Curators also generated annotations based on author data (authors & curators). Finally, we created an ensemble system that produced annotations by combining author and system data (systems & authors). To evaluate the results from all five sources (Fig. 1), a 'consensus annotation' was created. Three independent curators consolidated their annotations until a consensus was reached. The performance of each source was evaluated in terms of precision, recall (coverage) and balanced F-measure (the harmonic mean between precision and recall) against this consensus. The results, including the evaluation of an inter-annotator agreement (IAA) study as reference, are shown in Figure 1. Detailed background information can be found in the **Supplementary Methods** and

Supplementary Results. From this analysis, we can draw four major conclusions.

First, the majority of authors publishing PPI articles in *FEBS Letters* during the period of the experiment (76 in total) were willing and able to provide structured information: 57 of the authors agreed to participate in the experiment, 56 provided curation-relevant PPI data and 17 out of 22 authors who responded to a questionnaire accompanying the experiment expressed their interest in SDAs.

Second, because of the relatively low accuracy of authors' submissions, the use of authors' annotations did not result in saving of curators' time; however, by combining the annotations of authors with other sources, we are able to show that the data quality increases relative to their individual results (Fig. 1), 'Curator' versus 'Authors & Curator', and 'Authors' versus 'Authors & Systems'. This claim is also explained by an investigation of the overlap of annotations between the three sources (Supplementary Results).

Third, adding SDAs to the editorial process was not only desirable for the *FEBS Letters* editorial processes, but also technically possible. These SDAs continue to form part of the two FEBS (Federation of European Biochemical Societies) periodicals *FEBS Letters* (published by Elsevier) publications and *FEBS Journal* (published by Blackwell). Furthermore, journals incorporating SDAs are likely to increase the impact and visibility of their articles: unambiguously labeled articles will be more relevant in database search results.

Finally, the (online) systems used to assist in the annotation process required on average 2.3 min (s.d. = 1.4 min) to provide annotations on the test set articles, whereas 22 of the authors reported spending 68 min (s.d. = 72 min) on average. Our measurements of MINT curators over 36 PPI articles show an average of 50 min (s.d. 26 min) per article. According to both authors and curators, retrieving the correct identifiers is the most time-consuming task. Therefore, having automated systems provide a list of relevant identifiers is likely to save authors and curators a significant amount of time. As we show in the **Supplementary Results**, systems are able to return identifiers that both authors and curators miss. And, combining the identifiers reported by authors with results from multiple

automated systems ('Systems & Authors' in Fig. 1) achieved an *F*-measure of 0.75—better than the authors alone.

In addition, from the BioCreative II.5 article classification results, we established that text mining could help to select PPI-reporting articles for database curators (highest accuracy = 92%). However, given the results of the systems on interaction pairs, it is also clear that complex curation needs cannot be transferred to purely automated approaches. On the other hand, with the relatively good results of systems on identifiers, integrating text mining in the process is likely to generate a significant reduction of time. We need to acknowledge the problem of the relatively small sample of articles used in this first quantitative study, a problem that would be solved by sustained integration of author annotations in the publication and/or curation process, as in the ongoing *FEBS Letters* and *FEBS Journal* experiments. An investigation of integrating annotation systems into the human curation process will form part of the next BioCreative challenge (BC III). In the **Supplementary Discussion**, we propose a simple infrastructure for assembling text mining, authors, and curators in this process based on Leitner and Valencia⁸.

Articles labeled with SDAs simplify retrieval and identification of relevant articles for readers. Furthermore, author-provided SDAs would directly benefit areas of research requiring large amounts of high-quality biological data, such as systems biology⁹, as more structured data would be produced. In general, SDAs are likely to generate a fundamental impact on the relation between scientists and publications. Our results show that, by combining text mining, authors and curators, it is possible to increase the quality of the annotations and it is plausible that the curation process will be more efficient.

All participants in the BioCreative II.5 project are listed in the **Supplementary List of Participants**.

Note: Supplementary information is available on the Nature Biotechnology website.

AUTHOR CONTRIBUTIONS

F.L. and A.C.-a. contributed equally to this work. G.C. and A.V. contributed equally to this work. F.L. was in charge of the BioCreative challenge and analysis of the results, and the comparisons to the

FEBS data; A.C.-a. was in charge of the *FEBS Letters* experiment, data curation, author studies and inter-annotator agreement studies; S.A.M. performed the ensemble experiments; A.C. and L.L. contributed to the inter-annotator agreement studies and data curation; M.K. contributed to the BioCreative II.5 challenge; L.H. was responsible for analysis of alternative metrics to assess the BioCreative results; G.C. was responsible for the contributions from the MINT database and the *FEBS Letters* experiment; A.V. was responsible for the BioCreative challenge.

ACKNOWLEDGMENTS

The authors acknowledge the BioCreative II.5 participants. G.C. thanks P. McCabe, F. Wieland, G. Superti-Furga, F. Goni, I. Mowbray and A. Klinkenberg for their support. L. Salwinski participated in the inter-annotator agreement study. The work on BioCreative was supported by the European Commission FP6 NoEs ENFIN LSHG-CT-2005-518254 and EMBRACE LSG-CT-2004-512092, and by grants BIO2007-66855 from the Spanish Ministerio de Ciencia e Innovación and by the Spanish National Bioinformatics Institute (<http://www.inab.org/>). The work on the MINT database is supported by FEBS, AIRC and Telethon grants. The participation of L.H. and S.A.M. in BioCreative II.5 was funded by US National Science Foundation grant IIS-0844419.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Florian Leitner¹, Andrew Chatr-aryamontri², Scott A Mardis³, Arnaud Ceol², Martin Krallinger¹, Luana Licata², Lynette Hirschman³, Gianni Cesareni^{2,4} & Alfonso Valencia¹

¹Structural and Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. ²Department of Biology, University of Rome Tor Vergata, Rome, Italy. ³The MITRE Corporation, Bedford, Massachusetts, USA. ⁴Research Institute 'Fondazione Santa Lucia', Rome, Italy, and IRCSS Santa Lucia, Rome, Italy. e-mail: cesareni@uniroma2.it; valencia@cnio.es

1. Seringhaus, M.R. & Gerstein, M.B. *BMC Bioinformatics* **8**, 17 (2007).
2. Ceol, A., Chatr-Aryamontri, A., Licata, L. & Cesareni, G. *FEBS Lett.* **582**, 1171–1177 (2008).
3. Leitner, F. et al. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **7**, 385–399 (2010).
4. Orchard, S. et al. *Nat. Biotechnol.* **25**, 894–898 (2007).
5. Gerstein, M., Seringhaus, M. & Fields, S. *Nature* **447**, 142 (2007).
6. Hahn, U., Wermter, J., Blasczyk, R. & Horn, P. *Nature* **448**, 130 (2007).
7. Chatr-aryamontri, A. et al. *Nucleic Acids Res.* **35**, D572–D574 (2007).
8. Leitner, F. & Valencia, A. *FEBS Lett.* **582**, 1178–1181 (2008).
9. Aloy, P. & Russell, R.B. *Nat. Rev. Mol. Cell Biol.* **7**, 188–197 (2006).

Supplement - The BioCreative II.5 Participants

Team	First Name	Last Name	Email	Institution	Country
7	Cheng-Ju	Kuo	cju.kuo@gmail.com	Institute of Information Science, Academia Sinica	TW
7	Maurice Han Tong	Ling		Singapore Polytechnic	SG
7	Chun-Nan	Hsu		Institute of Information Science, Academia Sinica	TW
9	Luis	Rocha	rocha@indiana.edu	Indiana University	US
9	Jasleen	Kaur		Indiana University	US
9	Ahmed	Hamed		Indiana University	US
9	Alaa	Abi-Haidar		Indiana University	US
9	Azadeh	Nematzadeh		Indiana University	US
9	Artemy	Kolchinsky		Indiana University	US
10	Richard Tzong-Han	Tsai	BIOSMILE.Web.Search@gmail.com	IASL/IISR	TW
10	Hong-Jie	Dai		Academia Sinica/YZU	TW
10	yuehyang	bow		Academia Sinica/YZU	TW
10	Tsung-Kai	Lin		Academia Sinica/YZU	TW
10	ChiHsin	Huang		Academia Sinica/YZU	TW
13	Hongfang	Liu	hl224@georgetown.edu	Georgetown	US
13	Manabu	Torii		Georgetown	US
13	Zhangzhi	Hu		Georgetown	US
14	Frederic	Ehrler	ehrlar@unige.ch	University of Geneva	CH
14	Patrick	Ruch		Haute école de gestion de Genève	CH
14	Julien	Gobeill		University of Geneva	CH
14	Emilie	Pasche		University of Geneva	CH
16	Man	Lan	lanman.sg@gmail.com	Institute for Infocomm Research	SG
16	Jian	Su		I2R	SG
18	Yifei	Chen	yifechen@vub.ac.be	Computational Modeling Lab,Vrije Universiteit Brussel	BE
18	Feng	Liu		Computational Modeling Lab,Vrije Universiteit Brussel	BE
18	Bernard	Manderick		Computational Modeling Lab,Vrije Universiteit Brussel	BE
20	Kyle	Ambert	ambertk@ohsu.edu	Oregon Health & Science University	US
20	Aaron	Cohen		Oregon Health & Science University	US
22	Rune	Sætre	satre@idi.ntnu.no	University of Tokyo	JP
22	Takuya	Matsuzaki		University of Tokyo	JP

22 Kazuhiro	Yoshida		University of Tokyo	JP
22 Makoto	Miwa		University of Tokyo	JP
22 Yoshinobu	Kano		University of Tokyo	JP
26 Wei	Chih-Hsuan	p7896116@mail.ncku.edu.tw	NCKU KMLab	TW
26 Yi-Tsung	Tang		NCKU-csie	TW
31 Yong-gang	Cao	caomoexpert@hotmail.com	University of Wisconsin-Milwaukee NLP group	US
31 Qing	Zhang		University of Wisconsin-Milwaukee NLP group	US
31 Shashank	Agarwal		University of Wisconsin-Milwaukee NLP group	US
31 Balaji Polepalli	Ramesh		University of Wisconsin-Milwaukee NLP group	US
31 Zuofeng	Li		University of Wisconsin-Milwaukee NLP group	US
32 Karin	Verspoor	karin.verspoor@ucdenver.edu	UC Denver	US
32 Kevin	Cohen		UC Denver	US
32 Helen	Johnson		UC Denver	US
32 Chris	Roeder		UC Denver	US
32 William	Baumgartner		UC Denver	US
37 Fabio	Rinaldi	rinaldi@cl.uzh.ch	University of Zurich	CH
37 Kaarel	Kaljurand		University of Zurich	CH
37 Gerold	Schneider		University of Zurich	CH
37 Martin	Romacker		Novartis, NITAS	CH
37 Simon	Clematide		University of Zurich	CH
42 Jörg	Hakenberg	joerg.hakenberg@asu.edu	Arizona State University	US
42 Nguyen	Ha Vo		Arizona State University	US
42 Luis	Tari		Arizona State University	US
42 Chitta	Baral		Arizona State University	US
42 Robert J.	Leaman		Dept Biomedical Informatics, Arizona State University	US
42 Siddhartha	Jonnalagadda		Dept Biomedical Informatics, Arizona State University	US
42 Ryan	Sullivan		Dept Biomedical Informatics, Arizona State University	US
42 Christopher	Miller		Dept Biomedical Informatics, Arizona State University	US
42 Graciela	Gonzalez		Dept Biomedical Informatics, Arizona State University	US
51 Jay	Urbain	jay.urbain@gmail.com	Milwaukee School of Engineering	US

Supplementary Methods for “The FEBS Letters/BioCreative II.5 experiment: making biological information accessible”

Florian Leitner^a, Andrew Chatr-aryamontri^b, Scott A. Mardis^c, Arnaud Ceol^b, Martin Krallinger^a, Luana Licata^b, Lynette Hirschman^c, Gianni Cesareni^{b,d} and Alfonso Valencia^a

- a) Structural and Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
- b) Department of Biology, University of Rome Tor Vergata, Rome, Italy.
- c) The MITRE Corporation, Bedford, MA, USA.
- d) Research Institute “Fondazione Santa Lucia”, Rome, Italy, and IRCSS Santa Lucia, Rome, Italy.

This supplement describes the corpora (“document collections”) used for the various methods presented, data preparation strategies, our approach to estimate annotation time requirements, the setup for the ensemble system (systems + authors) and the combined system, as well as mathematical definitions of the evaluation functions.

Estimating the number of PPI articles p.a.

The combined number of articles indexed in Google Scholar (advanced search) for the year 2009 in the subject areas “Biology, Life Sciences, and Environmental Science” (Bio) and “Medicine, Pharmacology, and Veterinary Science” (Med) is currently 209,500: 122,000 in Bio + 145,000 in Med - 57,500 articles in Bio AND Med. This is a number similar to Highwire's 220,000 articles available for the same year. On the other hand, 20 to 30 % of the articles of FEBS Letters, Journal EMBO J, or EMBO reports contain protein interaction information according to our curators. Using some query logic, it is possible to estimate the number of articles reporting PPI informations per year (p.a.). Terms describing methods to detect PPIs that are highly specific to this topic can be used to create the estimate. All queries are done for all articles in these two subject areas but excluding patents, and only for the year 2009. Results are required to have “at least summaries”. To enhance the specificity of the results, the most distinct noun in the BC II.5 corpus (see below) of PPI articles, namely “interaction”, as established during the challenge, is combined with a method term using ANDing. The five most important PPI experimental method description terms (defined by the PSI:MI ontology) using this logic were determined to be: “coimmunoprecipitation”, “two hybrid”, “pull down”, “tandem affinity purification”, and “complementation assay”. The final query is then constructed by requiring the noun “interaction” and any of these five terms ([Google Scholar query link](#)) and is run against all articles in the two subject areas Bio and Med. This produces a result set of approximately 11,400 articles. Note that a similar query on PubMed would produce significantly less results, as NLM does not index the full text, only abstracts. As Google Scholar has full text access to most of the

relevant journals, it is safe to assume that a conservative low estimate of PPI articles p.a. is about 10,000.

Corpora usage

All articles in the corpora are from the FEBS Letters journal. Due to the fact that authors had made annotations that were public during the time of the challenge, we could not responsibly use these articles (from 2008) in the test set. Instead, author-annotated articles were used in the training set and a set of articles from the year before (2007) without public SDAs was used as test set. This forced us to apply the official training set results from the systems for the following two methods: building an ensemble system (systems + authors), and overlap evaluation (see Supplementary Results).

Authors corpus: The author corpus consisted of the 52 articles from 2008 for which participating authors (57 in total) returned evaluation-relevant PPI data (i.e., 56/57 authors returned annotations, while one author did not, and 4 of the 56 authors did not return annotations for protein identifiers or pairs). 33 of these articles form the “overlap corpus” mentioned below and 38 are part of the curator corpus. 51 of these articles are part of the author + curator corpus (i.e., the 55 “authors + curators” articles miss one article from this set, as we only have author annotations for it).

Curator corpus: The curator corpus consisted of 42 articles for which the curators generated annotations not basing their work on the author data, 38 of which overlap with the author corpus, and 33 of which in turn form the “overlap corpus”.

Authors + curator corpus: For the authors + curator evaluation we had 55 articles, covering 51 of the above author articles plus the four “missing articles” in the author corpus for which authors had submitted annotations without pair/identifier information. These four articles none the less were used by the curator for generating author-based annotations (while one of the 52 author submissions is not part of this corpus).

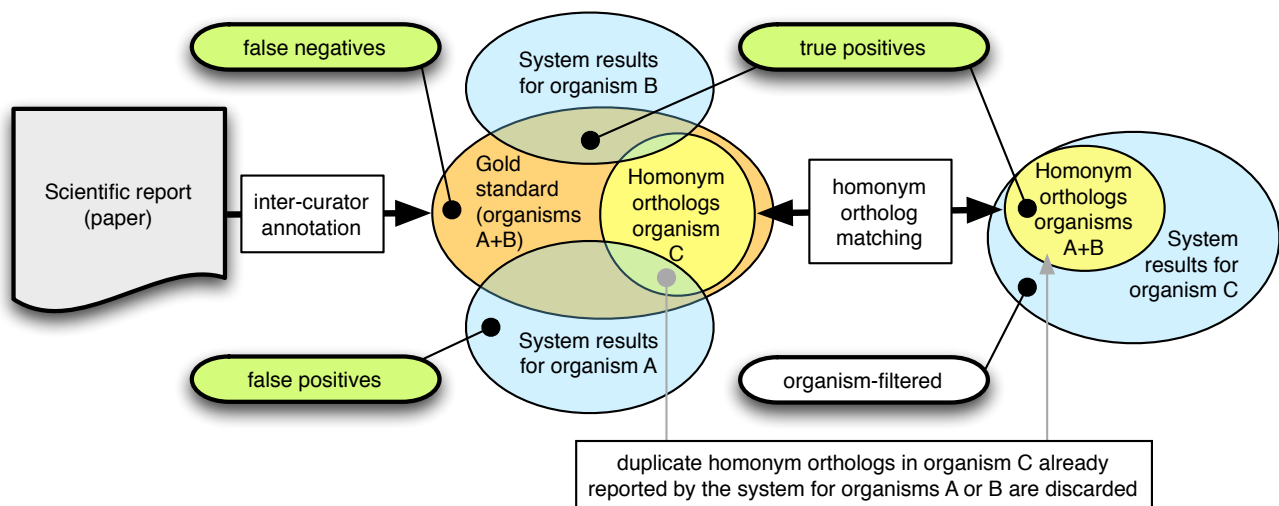


Figure 1. Combined homonym ortholog matching and organism filtering. After curating the gold standard (gold, center) from the paper (gray, left), that in this diagram contains proteins for the hypothetical organisms A and B, a system produces automated results for that paper (blue ovals) that contains proteins for organisms A, B and C. In principal, these annotations either count as true positives, or false positives, while missed results are false negatives (annotated with green ovals). Proteins in the system's result for organism C that have homonymous orthologs in the gold standard for organisms A or B are mapped and count as true positives (yellow circles). Results for organism C that have no mapping are filtered and do not count as false positives, while wrong (false positive) results for organisms A or B still are counted. Duplicates produced by the homonym ortholog mapping step are counted once only (white box at bottom).

Training set: For the BioCreative challenge, the training set consisted of 61 articles from 2008, mostly containing the publicly available SDAs created by authors and/or curators (45 of these articles overlap with the author corpus and were used in the ensemble system described later, 39 overlap with the curator corpus, and 48 with the author + curator corpus). For the overlap calculations presented in the Supplementary Results, we used a set of 33 articles from the original 61 training set articles for which we had annotations from all three sources (i.e., this training set, and the authors and curator corpora).

Test set: The test set consisted of 61 articles, mostly from 2007, but also included nine articles from 2002-2006 that had no annotations in any PPI repository at the time of the challenge. These nine articles were added to ensure the challenge participants (15 research groups in total) were not mining for evidence in the PPI databases. This corpus has no overlaps with any other corpus, as the SDA annotations were publicly available at the time of the BioCreative challenge, so the automated systems could not be tested on the articles annotated by authors and instead had to be evaluated on an independent test set

Overlap corpus: For the overlap study presented in the Supplementary Results we used the 33 articles that overlap between the training set, the author corpus, and the curator corpus.

Homonym ortholog matching and organism filtering (data preparation)

Species identification is a hard task for automatic annotation because authors often do not mention the species and because experiments carried out with proteins from different species are common in the PPI literature. As a practical solution, in addition to the “raw” system results, we applied two filters to estimate how well the automated systems could perform if augmented by information readily provided by an author or curator.

In a two-step process, we map homonymous orthologs to their correct gold standard entries and then filter out irrelevant species (organism filtering, see Figure 1), since in a real life application, it will be relatively easy for authors and curators to disambiguate the species and select correct orthologs if necessary. See the Supplementary Results for an investigation on the impact of these two post-processing steps.

The map of homonymous orthologs is created by extracting (i) all UniRef50 (release 15.0) sequence sets and (ii) grouping all homonymous UniProt records (r 15.0) by their gene and protein names (case-insensitive matching, filtering single character or pure numerals). Each set in one list is then intersected with all sets in the other to generate a list of homonymous *homolog* sets. Finally, for each result in the gold standard, we extract all identifiers in these homonymous *homolog* sets that contain a gold standard identifier,

but remove any protein identifiers for the same organism as the gold standard identifier from each extracted set to produce sets of homonymous *orthologs* (meaning, to avoid creating mappings to the same organism (*homologs*) as the gold standard identifier, we only extracted mappings to other organisms (potential *orthologs*)).

Organism filtering is based on looking up the TaxIDs of the gold standard proteins in their UniProt records and then removing any protein identifiers the systems reported that do not belong to one of those TaxIDs.

Metrics for system-aided author curation

We also made estimates of whether it would be possible to use the automated system results to facilitate the author's work in generating a structured digital abstract. The authors reported that they needed, on average, an hour to annotate their articles and that the most time consuming task was finding the protein identifiers (see Supplementary Results). Given that an average article had about 4 proteins and 3 interacting protein pairs, we made a conservative estimate that it would take the author, on average, 5 minutes to find a UniProt ID for an interacting protein. We further estimated that when presented with a proper list of possible interacting proteins and their identifiers, it would take an author 10 sec. to make a yes/no decision about a candidate protein (This estimate, of course, depends on the quality of the user interface and will be investigated in BioCreative III.) Using these estimates, we can compare expected time and quality of author curation with and without aid from an automated system.

- With no system, we would expect the author to find 2-3 of the 4 interacting proteins in an article (recall 66% from

Figure 1 in the main article), and it would take approximately 10-15 minutes of time.

- Using an automated system, we estimate that the author would spend 1 minute inspecting the 6 top ranked results of the best or a combined/ensemble system, which would yield a comparable recall (2-3 interacting proteins) – but in far less time (1 minute vs. 10-15 minutes).

Note that at a depth of the 6 top-ranked results, more than half of the automated systems (27/52) achieve a recall above 50% (the highest being 68%). This means that those systems would report in their top 6 candidates at least two of the four identifiers found in an average article. In the future, if automated systems can improve their ranking of answers or by using an ensemble system, this should bring most of the correct answers into the top 30, allowing an author to quickly annotate 80-90% of the correct interacting proteins in 5 minutes (30*10 seconds).

Creating an ensemble system and combining system results

For the *ensemble system* (systems + authors) we used the number of answers provided by the author as a cut-off that varied by document. The ensemble runs were produced using a trained classifier that combines the output of multiple raw system results (i.e., before the homonym ortholog mapping and organisms filtering described above) into a single ranked list. Using a set of documents for training, we created a maximum entropy classifier that we used to predict whether a particular protein is properly in the gold standard based on 5-fold cross-validation.

We also created a *combined system* (independent of the ensemble system mentioned above) joining the top 30 answers per document from each group's best run.

Formulas and calculation

NOMENCLATURE

p – precision

r – recall

TP – true positive [count]

FP – false positive [count]

FN – false negative [count]

TN – true negative [count]

AVERAGING

Note that IAA, F-measure, Recall, and Precision can be calculated over the complete results set (called “micro-averaging”), or for each article separately and then averaging over the number of articles (called “macro-averaging”). This second approach is more independent of the underlying set and therefore the approach used for all numbers in this publication.

IAA - INTER-ANNOTATOR AGREEMENT

The IAA is defined as the percentage of common annotations from two annotation sets A and B over the number of total annotations contained in those sets:

$$IAA = \frac{n \cdot \left| \bigcap_{i=1}^n A_i \right|}{\sum_{i=1}^n |A_i|}$$

This measure has the advantage over JPA (see below) of being directly comparable to the F_1 -measure (see below), as it can be shown that the above formula is the same as an F_1 -score that can be calculated between two annotation sets if one is used as reference for the other:

$$\begin{aligned} IAA &= \frac{2 \cdot |A \cap B|}{|A| + |B|} = \\ &= \frac{TP + TP}{(TP + FP) + (TP + FN)} = \\ &= \frac{2 \cdot TP}{2 \cdot TP + FN + FP} = \\ &= 2 \cdot \frac{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}{\frac{TP}{TP + FP} + \frac{TP}{TP + FN}} = \\ &= 2 \cdot \frac{p \cdot r}{p + r} = F_1 \end{aligned}$$

JPA - JOINT PROBABILITY OF AGREEMENT

The JPA is defined as the percentage of annotations the annotators have in common over the total number of annotations they made:

$$JPA = \frac{\left| \bigcap_{i=1}^n A_i \right|}{\left| \bigcup_{i=1}^n A_i \right|}$$

Where A_i are the annotation sets generated by n annotators. This measure is more strict than IAA, and we use micro-averaging for the numbers shown in this publication; however, its results are not comparable to the F-measure.

AUC iP/R - AREA UNDER THE INTERPOLATED PRECISION/RECALL CURVE

This is the official evaluation function of the challenge results (AUC iP/R is part of the online results available from the www.biocreative.org site). The AUC A of the interpolated P/R function f_{pr} is defined as follows:

$$\begin{aligned} A(f_{pr}) &= \sum_{j=1}^n p_{ij} \cdot (r_j - r_{j-1}) \\ p_i(r) &= \max_{r' \geq r} p(r') \end{aligned}$$

Where n is the total number of correct hits and p_i is the highest interpolated precision for the correct hit j at r_j , the recall for that hit. Interpolated precision p_i is calculated for each recall r by taking the highest precision at r or any $r' \geq r$.

F-MEASURE

$$F_\beta = \frac{(1 + \beta^2) \cdot (p \cdot r)}{\beta^2 \cdot (p + r)}$$

Where β is a measure of the relative importance of recall as compared to precision. The balanced F-measure (F_1) weighs precision and recall equally:

$$F_1 = 2 \cdot (p \cdot r) / (p + r)$$

PRECISION

$$TP / (TP + FP)$$

RECALL

$$TP / (TP + FN)$$

Supplementary Results for “The FEBS Letters/BioCreative II.5 experiment: making biological information accessible”

Florian Leitner^a, Andrew Chatr-aryamontri^b, Scott A. Mardis^c, Arnaud Ceol^b, Martin Krallinger^a, Luana Licata^b, Lynette Hirschman^c, Gianni Cesareni^{b,d} and Alfonso Valencia^a

- a) Structural and Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
- b) Department of Biology, University of Rome Tor Vergata, Rome, Italy.
- c) The MITRE Corporation, Bedford, MA, USA.
- d) Research Institute “Fondazione Santa Lucia”, Rome, Italy, and IRCSS Santa Lucia, Rome, Italy.

This supplement describes in greater detail the results of the FEBS Letters SDA experiment and the inter-annotator agreement, explains the challenges for automated systems, illustrates a detailed, comparative analysis of results from all three main sources, and reports on the ensemble and combined system results.

FEBS Letters SDA experiment - detailed author results and responses

76 authors were asked to participate in the FEBS Letters SDA experiment, and 57 of them accepted the invitation. Of these 57, 52 returned PPI curation-relevant annotations. Of the 19 authors who did not join the FEBS experiment, the main reasons were concerns about possible delays in the publication of their articles. 22 of the participating authors also responded to an accompanying questionnaire, and 17 demonstrated interest in the experiment. Seven authors complained about insufficient guidelines for creating the annotations (authors received an Excel form containing the explanations for creating the annotations.) However, two authors explicitly said that they would be discouraged from publishing in FEBS Letters if the SDA procedure were made compulsory. The 22 authors that returned the questionnaire declared that generating the SDAs took them, on average, 68 minutes (i.e., about one hour) of their time (with a Standard Deviation of 72 min.) and that the most demanding task was the identification of the protein database identifiers.

Inter-annotator agreement (IAA) studies

Two curators from the MINT database and one from another PPI database, adopting the same curation procedures, independently annotated sample sets of PPI manuscripts. These annotations were used to estimate an inter-annotator agreement based on sets of 21 common documents between each curator pairing.

When comparing protein identifiers annotated from curators of the same database, we calculated an (macro-averaged) inter-annotator agreement (IAA) of 95% (with a [micro-averaged] joint probability of agreement, JPA, of 87%), while comparison between curators of two different databases had an

agreement of 86% (JPA: 70%). When comparing interaction pairs, curators of the same database had an agreement of 87% (JPA: 69%), while curators from different databases only had an agreement of 76% (JPA: 60%).

The intra-database IAA scores reflect the approximately expected values for curator performance and indeed match well with the F_1 scores of the curator evaluation (Figure 1, curator results, main article). The formulas for calculating IAA and JPA are shown in the Supplementary Material.

The challenge of identifying interacting proteins

There are two problems that make it difficult for automated systems to identify interacting proteins: species identification and distinguishing interacting proteins from other proteins mentioned in an article. See the Supplementary Methods for details on how homonym ortholog maps and organism filtering were implemented.

Homonym ortholog mapping had the effect of raising the scores on the protein pair task considerably (3pp/14% for the highest-scoring system, T42 S1), and to a lesser extent on the protein identification task (3pp/8%, for T42 S1). Organism filtering, because of the removal of large quantities of false positives from the result set, has a much larger impact on the results than homonym ortholog mapping. Systems mainly report these large numbers of false positives due to the mentioned difficulty of disambiguating the article's species origins (e.g., authors might not even mention the source species of the reported proteins) and due to the protein name ambiguity (proteins from different organisms with the same names). Organism filtering increases the protein identification score of T42 S1 by 13pp/30% and the pairs result by 13pp/58%. Combining the mapping and filtering steps produces an optimized, highest achievable result from the automatically generated data after human processing (i.e., the result set now only contains correct orthologs and correct species results). For system T42 S1 the overall increase resulting from mapping and filtering is

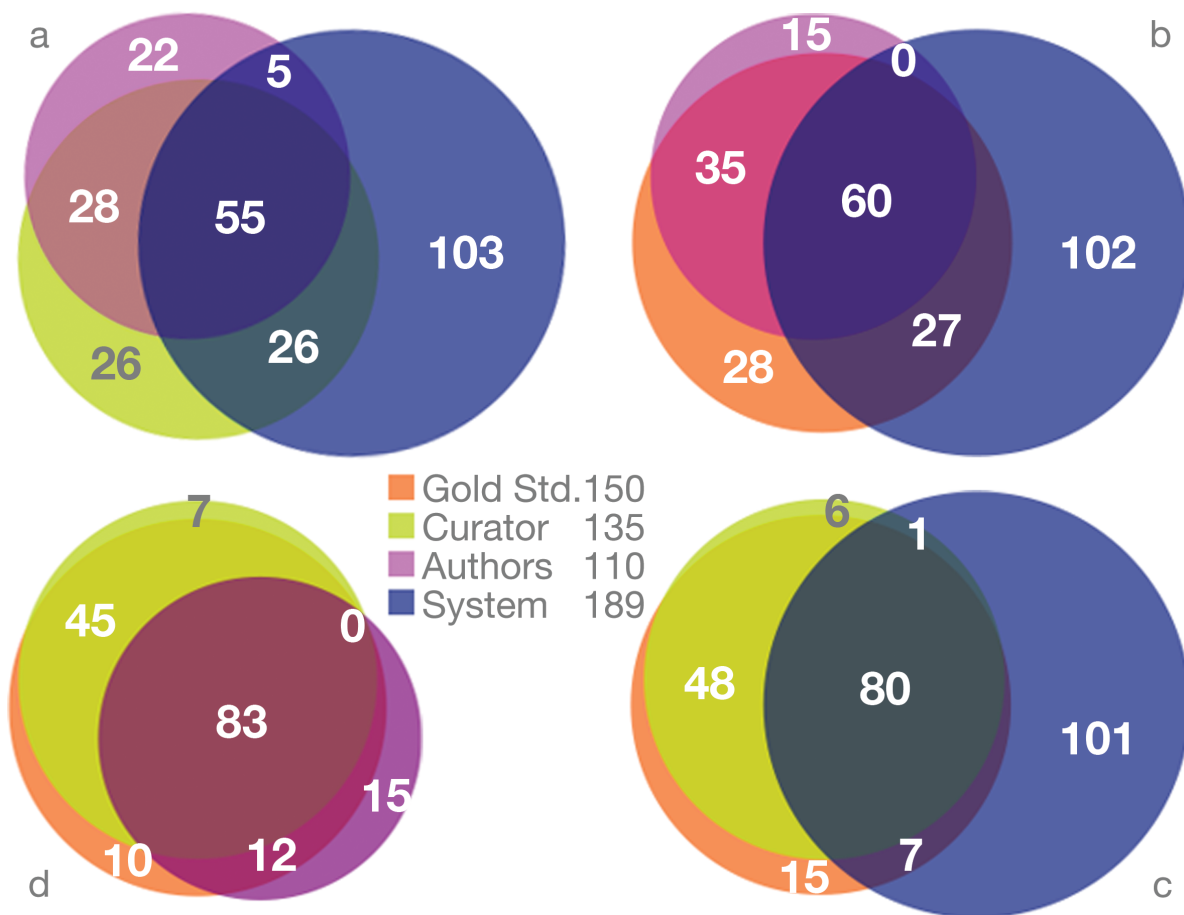


Figure 1. Overlap of annotations by authors, curators and automatic systems, shown as Venn diagrams. The number of annotations are for the protein identification task and refer to 33 (training set) articles that all three sources had in common. The automatic systems, for each article, produce a relatively long list of proteins ranked according to relevance. For a fair comparison with the other annotation methods we have only considered the system's top six proteins for each article. The system data in this figure are from Team 10, training set run 5 (also see Supplementary Methods). For an interpretation, see text. (Clockwise) *Top left, a*: Comparison of the result sets from all three sources of annotations. All others show the overlaps relative to the gold standard: *Top right, b*: Authors and text-mining systems. *Bottom right, c*: Curators and text-mining systems. *Bottom left, d*: Curators and authors.

16pp/37% for identifiers and 15pp/69% for pairs, or 2-3pp better than organism filtering only.

It is also difficult for automated systems to distinguish interacting proteins from other proteins mentioned in an article. This is another source of the large numbers of false positives - even after homonym ortholog filtering - resulting in low precision. Low precision, in turn, drives down the balanced F-measure when looking at the entire ranked result sets (Note that F-measure does not take into account ranking in any way.) However, if we apply a cut-off (e.g., 6 top-ranked results per document), or if we use an F-measure weighted to emphasize recall (see definition of F_β in the Supplementary Methods), then the results suggest that the systems could be used even in their present state to provide useful information to authors or curators. In summary, the identification task is not comparable, and much harder, than "regular" protein identification tasks where all

proteins mentioned in an article are to be identified, given that the actual number of proteins with experimental interaction evidence is only a small fraction of the proteins mentioned in an article and their identification requires additional analysis steps.

The BioCreative system results as quoted in this publication are the results *after application of the homonym ortholog filtering process*. The individual system results before and after each post-processing step can be downloaded here as an Excel file: [BC II.5 results, on www.biocreative.org](http://www.biocreative.org/BC-II.5-results).

Comparing system, authors and curator result sets

Due to the estimates presented in the Supplementary Methods (*Metrics for system-aided author curation*), we used the best system's top 6 results for each article when calculating the annotation overlaps (Figure 1). In total, there are 33

articles the three sources have in common; the performance of the system that was used in this analysis - T10 R5, the highest scoring AUC iP/R system on the test set - is about 10% better on the training than on the test set; As the SDA annotations were publicly available at the time of the BioCreative challenge, the automated systems could not be tested on the articles annotated by the authors and were evaluated on an independent test set (see Supplementary Material). To calculate the overlaps between the three sources the training set result used was 18%/10pp higher performing than its test set result (AUC iP/R of 0.575 in the test set vs. 0.679 in the training set).

The Venn diagrams in Figure 1 show that the results from systems, authors, and curators are disjoint over large parts of their annotations, making it likely that each source might be able to assist the others (Fig. 1a). Authors and automated systems have just 1/4 of their annotations in common, and joining the results of the best-scoring system with authors' annotations would increase the gold standard coverage of author annotations by 28%/18pp (see Fig. 1b). When joining system and curator annotations (Fig. 1c), the increase in coverage is not as large (5%), but the system reproduces nearly 2/3 of the annotations the curators made. Thus, it can be argued that text mining could assist both authors and curators and help speed up the process of retrieving correct protein identifiers. In comparison, using authors' annotations, curators obtain a similar benefit in coverage with the authors' results and the added advantage of a much smaller number of false positives (15 vs. 101, Fig. 1d vs. 1c). Comparing this last Venn Diagram (1d) to the gain curators had in coverage when integrating author data (Figure 1,

curators vs. authors + curators, main article) shows that the author data helped them to remove FP annotations shown in the diagram above, and to integrate the false negatives the authors did find. It is also worth noting that the automated systems produced more annotations than the other sources; some of these additional annotations will refer to proteins mentioned in the article that are not relevant to the task. However, these extra results negatively affect calculations of precision and balanced F-measure.

Ensemble and combined system results

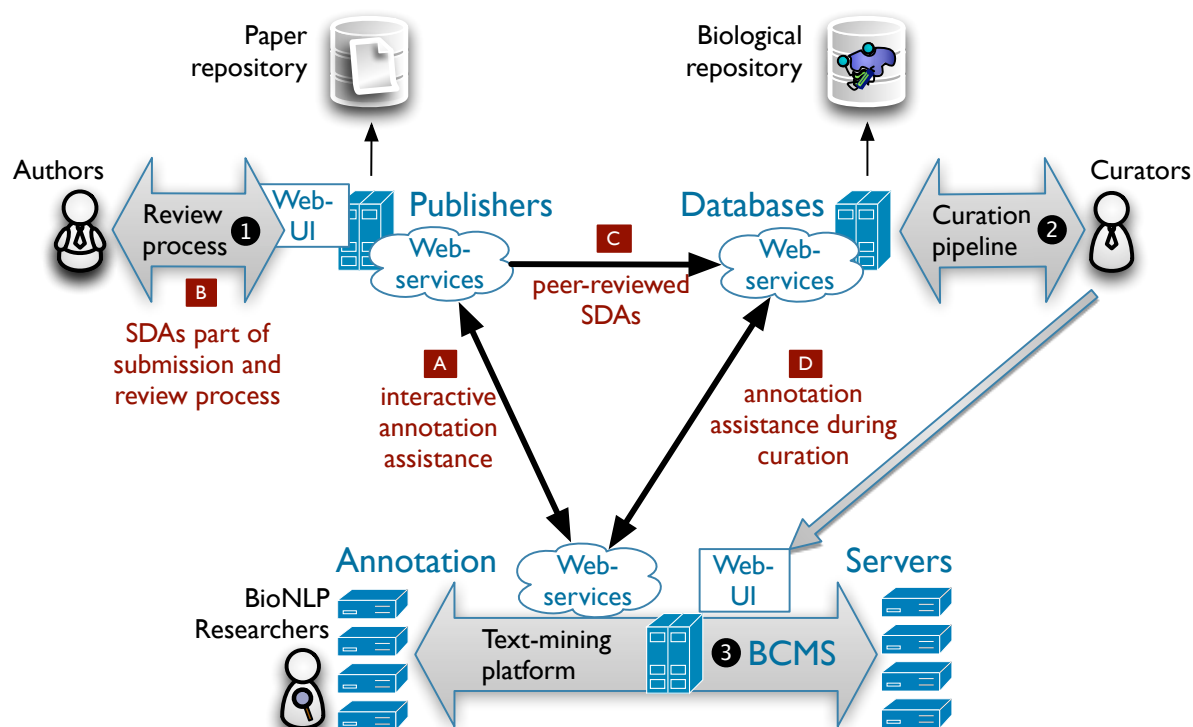
The overlap comparisons showed that the three approaches could assist each other: systems could help the authors, who in turn could aid curators, improving the overall performance. For example, authors working alone found about 2/3 of the relevant identifiers, while 84% of their annotations were correct (Figure 1, main article, authors). Curators alone achieved a recall of 89%. Comparing the top-scoring system's identifier results to the performance of authors indicates that the (*post-processed*) results of the automated systems lag behind in balanced F-measure; however, we also ran an experiment to combine author input with inputs from multiple automated systems using the *raw (unprocessed)* results (an "ensemble system"); This ensemble system achieved a precision of 83%, a recall of 73%, and a balanced f-measure of 0.75 - better than the authors alone (see Figure 1, systems + authors, main article). Furthermore, the "combined system" (using the top 30 answers per document from each group's best run) achieved a macro-averaged recall of 84% - higher than the (macro-averaged) author recall (55%) and not far below the curator recall (89%).

A service model for integrating Structured Digital Abstracts in the publishing process.

Supplementary Discussion for “The FEBS Letters/BioCreative II.5 experiment: making biological information accessible”

Florian Leitner^a, Andrew Chatr-aryamontri^b, Scott A. Mardis^c, Arnaud Ceol^b, Martin Krallinger^a, Luana Licata^b, Lynette Hirschman^c, Gianni Cesareni^{b,d} and Alfonso Valencia^a

- a) Structural and Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
- b) Department of Biology, University of Rome Tor Vergata, Rome, Italy.
- c) The MITRE Corporation, Bedford, MA, USA.



- d) Research Institute “Fondazione Santa Lucia”, Rome, Italy, and IRCSS Santa Lucia, Rome, Italy.

We present a possible model to programmatically integrate SDAs into the publishing process and for enriching existing biological repositories. Our intent is to spark a discussion about ways to include author-provided data for public manuscripts.

Most of the workflow required to integrate structured information in scientific publications and databases is already in place (Figure, 1-3): (1) Publishers provide interfaces for authors connecting them with the review process, leading to new publications that are stored in web-accessible repositories of papers. (2) Databases have professional bio-curators filling biological repositories with structured records that can be accessed by web services. (3) Text miners (BioNLP, Natural Language Processing in Biology) have joined forces to provide the BioCreative Meta-Server (BCMS)¹, a platform connected to a growing number of Annotation Servers with a unified web-protocol with the ability to extract biologically

relevant information from text. Concerning the text-mining service (3), there are other alternatives available, such as the web-services offered by Reflect⁴ or Whatizit⁵, or, for example the paper annotation service offered by PaperMaker⁶. However, the BCMS is quite different from those approaches, as it does not provide its “own” annotation service, but rather integrates many different providers under one point-of-access (and, indeed, Reflect and Whatizit are planned to be part of text-mining services offered by the BCMS platform in the future, just as iHOP^{1,2} already is). All the services registered on the BCMS platform get queried in parallel for any incoming request and the results are pooled. Therefore, the BCMS acts as a “meta-service” to access multiple providers under one URI (unique resource identifier) with a unified data format and constitutes a federated repository of text-mining-based annotations.

By setting up the necessary connections, SDAs can become an integral part of the publication and curation processes (Figure, A-D, red): Publishers would integrate the text mining resources (A) with their paper submission user interfaces (Web-UI) for authors (B) to facilitate the initial creation of the SDAs: This process would be formulated as an interactive protocol with the text-mining service. When papers are submitted for publication to the publisher's server, they could be forwarded to the text-mining service, possibly including annotations from authors to reduce irrelevant results (especially, specifying the relevant species for the article). After reviewers and authors agree on the annotations, the peer-reviewed SDA, now included as part of the paper, gets sent to the databases (C), and curators can decide to accept or revise annotations before adding them to their repository. Text-mining services might be used to process old publications without SDAs (D), which can be done directly via a web-UI or integrated into the curation pipeline using web services.

After the results from the FEBS experiment and BC challenge we are convinced of the feasibility of facilitating author-mediated SDA generation by including text-mining tools into the review process.

By using a multi-step approach - automated systems, authors, reviewers and curators - the quality of annotations should increase, and articles with SDAs that have not (yet) been curated by a database would include some essential annotations. The benefits are considerable, as we argue in the main paper. The largest benefit for the scientific community will be the significant increase in coverage of the biological repositories. An open question is the storage of these annotations. Currently, the FEBS Letters and Journal store this information internally. Alternatively, a public repository could be a more open model that links annotations back to their article by using the article's DOI, storing annotations for any article and all journals.

Given that the scientific and technical elements seem to be in place, future progress is in the hands of the stakeholders. Databases and publishers, with the collaboration of authors, annotators and editors, can provide the essential connections between editorial houses and biological databases in which text mining will play an essential role.

References

1. Fernández et al. **iHOP web services**. *Nucleic Acids Res* (2007) vol. 35 (2) pp. W21--W26
2. Hoffmann and Valencia. **A gene network for navigating the literature**. *Nat Genet* (2004) vol. 36 (7) pp. 664
3. Leitner et al. **Introducing meta-services for biomedical information extraction**. *Genome Biol* (2008) vol. 9 Suppl 2 pp. S6
4. Pafilis et al. **Reflect: augmented browsing for the life scientist**. *Nat Biotechnol* (2009) vol. 27 (6) pp. 508--510
5. Rebholz-Schuhmann et al. **Text processing through Web services: calling Whatizit**. *Bioinformatics (Oxford, England)* (2008) vol. 24 (2) pp. 296-8
6. Rebholz-Schuhmann et al. **PaperMaker: validation of biomedical scientific publications**. *Bioinformatics* (2010) vol. 26 (7) pp. 982