Enriching Publications with Structured Digital Abstracts: the human-machine experiment

Florian Leitner[1], Andrew Chatr-aryamontri[2], Arnaud Ceol[2], Martin Krallinger[1], Luana Licata[2], Lynette Hirschman[3],Gianni Cesareni[2], Alfonso Valencia[1]

[1]Structural Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
[2]Department of Biology, University of Rome Tor Vergata, Rome, Italy
[3]The MITRE Corporation, Bedford, MA, USA.

A majority of articles in the biological domain report relationships between biological entities, such as genes, proteins, or metabolites. Computational Biology and especially Systems Biology utilize this information to establish biological networks, providing a holistic view of cell organization and function [1, 2]. While recent progress in experimental technologies has resulted in an exponential growth of published information, the process of extracting this information and converting it into structured database entries is lagging behind because of the intrinsic difficulties and slow pace of the human-based curation process [3].

In the protein interaction domain, a group of major public interaction data providers have come together in the IMEx consortium (imex.sourceforge.net) and agreed to share the curation effort and exchange completed records on molecular interaction data. At the current funding level, the IMEx databases can only deal with about 2,000 out of the estimated 10,000 protein interaction articles published yearly. The stark contrast of these numbers stresses the need for a significant increase in the number of processed articles if we want to avoid generating increasing amounts of “stagnant” information.

One possible avenue to significantly extend the number of protein interactions captured by public repositories would include authors in the annotation effort. As suggested by Orchard et al., authors could be asked to submit the relevant structured information required by the databases during the editorial process, as defined by the minimum information requirement for reporting protein interaction experiments (MIMIx standard) [4].

Approaches for adding structured information to scientific publications have been discussed: most prominently 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. We proposed the use of automated systems to assist authors and/or curators during the annotation process as one possible solution to these problems [7].

The debate is stalling because of the absence of hard data for the fundamental questions at hand:

1. What is the error rate and the subjectivity in the process of manuscript curation (both for authors and database curators)?
2. Can authors help by providing structured information, and, are they willing to help?
3. Is this extra work useful in increasing the amount and accuracy of existing structured information, i.e., can databases make use of it?
4. Can such a procedure be realistically implemented in an editorial process?
5. How efficient are automatic information extraction strategies and can they take over or assist curators/authors in these tasks?

In February 2007, the FEBS Letters editorial board designed an experiment with the intent of addressing these questions. The pilot project on Structured Digital Abstracts (SDAs) [8] focused on protein interaction (PPI) and asked authors to provide structured information as a variant of the MIMIx standard, after the manuscript's acceptance. In addition, trained curators of the MINT [9] PPI database reviewed these annotations.

From March to December 2007, 76 authors were invited to take part in this FEBS Letters SDA experiment. Three quarters of them accepted and the articles were published with an appended SDA reporting the identifiers of the interacting proteins, the interaction type (physical interaction, co-localization, enzymatic reaction, etc.) and the experimental method used to support the interaction. Of the authors who did not accept, the main reasons were concerns about possible delays in the publication of their articles. The majority of the participants who responded to an accompanying questionnaire demonstrated interest in the experiment. However, a few authors explicitly said that they would be discouraged from publishing in FEBS Letters if the SDA procedure were made compulsory. The authors declared that this extra responsibility took, on average, one hour of their time and that the most demanding task was the identification of the protein database identifiers.

The follow-up question then is if text mining and information extraction approaches can be used to support the annotation process, maintaining quality while reducing cost and time by aiding authors and curators in the task. In the context of the BioCreative initiative (www.biocreative.org) on information extraction, we organized a community effort ("BioCreative II.5") to directly answer these questions in a setup that, with minor adjustments, could form part of an online editorial procedure (see Supplement 1). The proposed scenario envisions authors or curators sending publications to the BioCreative Meta-Server (BCMS) [10], a text-mining meta-server that in turn distributes these texts to annotation servers, a collection of specialized text-mining servers provided by more than a dozen research groups throughout the world. These annotation servers then mine for relevant data and return results in ranked order with normalized confidence values using a pre-defined communication protocol. The BCMS would then validate and aggregate the results, returning them to the authors/curators, from which they then select the relevant annotations.

The BioCreative evaluation was designed to assess the performance of automatic extraction methods when challenged with the following three tasks:

- Identification of articles reporting protein interaction information, to investigate the possibility of helping database curators in the selection of relevant PPI articles.
- Identification of the UniProt IDs of the proteins participating in an interaction supported by an experiment described in the article - a process called “protein normalization” (N.B., this task is not comparable, and much harder, than “regular” protein identification tasks, where all proteins mentioned in an article are normalized, given that the actual number of proteins with experimental interaction evidence is only a small fraction of the proteins mentioned in an article).
- Identification of all binary interaction pairs that are experimentally validated in the article.

Fifteen teams responded to the challenge that was held after the FEBS Letters experiment: The training phase began in March 2009, the test phase took part in June, and the challenge culminated in a workshop held during October 2009 in Madrid. The teams were asked to provide annotation servers, while the BCMS was used to evaluate the system results. For training, the teams were provided with the text from the manuscripts annotated during the FEBS Letters SDA experiment; the system outputs were then evaluated on an independent test set using this real-time, web-service-based setup. On average, the annotation servers needed 2 minutes to annotate an article. There will be a special journal issue scheduled for September 2010, describing the systems used in BioCreative II.5.

All annotations from both the FEBS Letters experiment and the BioCreative challenge were collected and the performance of each result type, namely **(a)** authors, **(b)** database curators, **(c)** curators using the data provided by the authors, and **(d)** automated text-mining systems, was evaluated in terms of precision, recall (coverage) and balanced F-measure (harmonic mean between precision and recall). In addition, for the automated results, the area under the interpolated precision/recall-curve (AUC iP/R) [11] was calculated (see Table 1, left, and Supplement 2). iP/R curves express the behavior of precision and recall while iterating over ranked results.

In addition to the authors, two curators from the MINT database and one from another PPI database, adopting the same curation procedures, independently annotated the manuscripts. These annotations were used to estimate an inter-annotator agreement based on two sample sets of 21 common documents. When comparing protein identifiers annotated from curators of the same database, we calculated an agreement of 93%, while comparison between curators of two different databases had an agreement of 81%.

To evaluate all these results, a consensus annotation, known as “ground truth”, or “gold standard”, had to be created. To this end, the three curator annotations were consolidated: whenever curators observed a disagreement between their annotations, they continued reviewing the manuscript until a consensus was reached. Furthermore, both MINT curators and the BioCreative organizers made final corrections and updates to the gold standards in the context of the BioCreative challenge.

The authors provided annotations for a total of 52 articles, while the test set for the annotation servers consisted of 61 FEBS Letters articles from the year before the experiment (unknown to the BioCreative participants at the time of the test phase). We received a total of 134 result sets for the three tasks (article classification, identifier and pairs extraction) from the text-mining systems. The detailed results are discussed in Supplement 2, and Supplement 3 contains the full set of results for each submission and task; selected results are shown in Table 1, right.

For any of the results provided by one of the three parties involved - a system, author, or curator - to be beneficial for another, their annotations should not be completely overlapping to help increase the final coverage. Therefore, we calculated the overlaps between the three parties’ final results (see Figure 1). However, as the SDA annotations were publicly available at the time of the BioCreative challenge, the automated systems could not be tested on the official articles annotated by the authors and were evaluated on an independent test set (as discussed above). To calculate the overlaps between the three parties, we therefore used the final training runs, as they cover the public FEBS Letters experiment articles (in total, there are 33 common articles between curator, authors, and system annotation results; the performance of the system used - Team 10, Run 5, the high-scoring *test* set result system - is about 10% better on the training set). In this figure, we only used the top 6 results from the automated system for each article (see Supplement 2 explaining the choice of this result size).

The results from systems, authors, and curators are disjoint over large parts of their annotations, making it likely that each party would be able to assist the others (Figure 1, top left). Authors and automated systems have just 1/4 of their annotations in common, and joining the high-scoring system’s result with the author annotations would increase the coverage of author annotations on the gold standard by 28% with a small investment of additional time (see Figure 1, top right and discussion in Supplement 2). When joining system and curator annotations (Figure 1, bottom right), the increase in coverage is not as large (5%), but the system reproduces nearly 2/3 of the annotations the curators made and therefore it can be argued that text-mining could help speed up the process of retrieving the protein identifiers. In comparison, joining author and curator identifier annotations has a similar picture as for curators and systems, although with many fewer false positives (15 vs. 101, Figure 1, bottom left and right). Comparing this last picture to the gain curators made when integrating author data (“authors+curator”) shows that the author data helped them to remove exactly these 7 FP annotations, and to integrate the 12 false negatives the authors did find.

These comparisons show 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 (Table 1, left side). Curators alone achieved a recall of 89%. When curators based their work on the author annotations, their results increased to 94% coverage (at the same precision) as when working alone (Table 1, left, line 4). Comparing the top-scoring system’s identifier results to the performance of authors indicates that the (organism filtered, ortholog mapped) 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 (an “ensemble” system); the high-scoring ensemble system achieved a precision of 83%, a recall of 73%, and a balanced f-measure of 0.75 - better than the authors alone.

The four main conclusions of the FEBS Letters experiment in response to the initial questions are that **i)** many authors, albeit not all, understand the importance of structured information accompanying their articles and are willing to cooperate, **ii)** because of the relatively low accuracy of authors’ submissions, the initial procedure did not result in saving of curators’ time, nevertheless, **iii)** the interplay between authors and curators increases the fidelity of the annotation process, and **iv)**, that adding author-generated structured annotations to the editorial process is possible.

The analysis of the results of the BioCreative II.5 challenge provides solid data to answer the fifth question about machine-generated annotations: 14 out of 52 result sets from three teams for the interaction protein identification had more than 0.5 AUC iP/R in the “interacting protein identification task”, and 7 of the 41 result sets from four teams performed over 0.25 AUC iP/R in the “pair task”. Based on the overall quality of the systems’ results, we predict that combining results of various systems will further improve the performance over the numbers presented here. Combining the results from all systems yields a maximally achievable recall of 87% over the whole test set (i.e., micro-averaged) – significantly higher than the (micro-averaged) author (58%) or curator (82%) coverage alone. This is consistent with BioCreative II which showed that the combination of system outputs yielded better results than any single system [12].

We conclude that machine generated annotations are a convenient starting point for establishing SDAs. Retrieving database identifiers in the human curation process is the most time consuming task and could be alleviated by providing assistance through automated systems. In addition, text mining can help to select relevant articles for curation (article classification). As anecdotal evidence, the text-mining systems managed to clearly separate two articles from the extended article classification test set (595 articles, containing the 61 relevant articles) that were originally annotated as negative articles in the ground truth but actually do contain curation-relevant interaction information. The specific benefits of having annotation systems inserted into such pipelines will be elucidated by a usability study that will form part of the BioCreative III challenge planned for September 2010.

Acknowledgements

A.V., F.L., L.H., and M.K. would like to thank Scott A. Mardis for his valuable contributions to both the BioCreative challenge and the evaluation phase. G.C. would like to thank Patricia McCabe, Felix Wieland, Giulio Superti-Furga, Felix Goni, Iain Mowbray and Adriaan Klinkenberg for their support to the FEBS Letters experiment.

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 the Spanish National Bioinformatics Institute (www.inab.org), a platform of Genoma España. The work in the MINT database is supported by FEBS, AIRC and Telethon grants. The participation of L.H. in BioCreative II.5 was funded in part by US National Science Foundation grant IIS-0844419.

References

1. Aloy, P. & Russell, R.B. Structural systems biology: modeling protein interactions. Nature Reviews Molecular Cell Biology 7, 188-197 (2006).
2. Kitano, H. Computational systems biology. Nature 420, 206-210 (2002).
3. Seringhaus, M.R. & Gerstein, M.B. Publishing perishing? Towards tomorrow's information architecture. BMC Bioinformatics 8, 17 (2007).
4. Orchard, S. et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol 25, 894-898 (2007).
5. Gerstein, M., Seringhaus, M. & Fields, S. Structured digital abstract makes text mining easy. Nature (2007).
6. Hahn, U., Wermter, J., Blasczyk, R. & Horn, P. Text mining: powering the database revolution. Nature (2007).
7. Leitner, F. & Valencia, A. A text-mining perspective on the requirements for electronically annotated abstracts. FEBS Letters 582, 1178-1181 (2008).
8. Ceol, A., Chatr-Aryamontri, A., Licata, L. & Cesareni, G. Linking entries in protein interaction database to structured text: The FEBS Letters experiment. FEBS Letters 582, 1171-1177 (2008).
9. Chatr-aryamontri, A. et al. MINT: the Molecular INTeraction database. Nucleic Acids Res 35, D572-574 (2007).
10. Leitner, F. et al. Introducing meta-services for biomedical information extraction. Genome Biol 9 Suppl 2, S6 (2008).
11. Davis, J. & Goadrich, M. The relationship between precision-recall and ROC curves. Proceedings of the 23rd international conference on Machine … (2006).
12. Krallinger, M. et al. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biol 9 Suppl 2, S1 (2008).

13 Chen, L. et al. Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics (2005) vol. 21 (2) pp. 248-56

Figureă

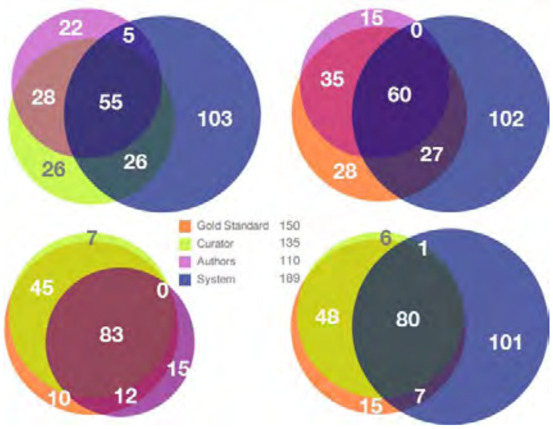


Figure 1. Annotation result overlaps shown as Venn diagrams. The system data is from Team 10, training set run 5, but only taking the top 6 results for each article. Numbers are for the protein identifiers annotated by the corresponding method on the 33 common articles used in this study. For an interpretation, see text. (Clockwise)*Top left:* Comparison of the result sets from all three sources of annotations. All others show the overlaps relative to the gold standard: *Top right:* Authors and text-mining systems. *Bottom right:* Curators and text-mining systems. *Bottom left:* Curators and authors.

Task	Class	Precision	Recall	F-Score
Protein Identifiers	systems	74%	55%	0.59
	authors	84%	66%	0.71
	curators	96%	89%	0.91
	authors+curators	96%	94%	0.95
Interaction Pairs	systems	53%	34%	0.37
	authors	72%	57%	0.59
	curators	93%	83%	0.86
	authors+curators	93%	89%	0.90

Task	Class
Protein Identifiers	best F-score (T42, S1)
	best AUC iP/R (T10, R5)
Interaction Pairs	best F-score and AUC iP/R incl. MINT data (T18, R1)

Table 1. *Left:* Recall (coverage), precision and balanced F-scores for the protein identification and the interaction pair tasks: Results for the best automated system in the BioCreative challenge after homonym ortholog filtering (see Supplement 2); the authors' and DB curator's annotation performance; and evaluation of the structured information produced by curators using the author annotations. *Right:* Detailed performance values of the top-scoring systems in the BioCreative challenge for both tasks, again after homonym ortholog filtering. The highest balanced F-score and AUC iP/R systems are shown. In the trade off between precision and recall the best F-score systems optimize precision and the best AUC iP/R systems optimize recall. Interestingly, the same system has the best F-score in the identifier and pair tasks, and also has the highest AUC iP/R score for the pairs task (Team 42, S1). The performance of a system that used information about existing interactions from MINT is shown for comparison (bottom, right).

Table 1. *Left:* Recall (coverage), precision and balanced F-scores for the protein identification and the interaction pair tasks: Results for the best automated system in the BioCreative challenge after homonym ortholog filtering (see Supplement 2); the authors' and DB curator's annotation performance; and evaluation of the structured information produced by curators using the author annotations. *Right:* Detailed performance values of the top-scoring systems in the BioCreative challenge for both tasks, again after homonym ortholog filtering. The highest balanced F-score and AUC iP/R systems are shown. In the trade off between precision and recall the best F-score systems optimize precision and the best AUC iP/R systems optimize recall. Interestingly, the same system has the best F-score in the identifier and pair tasks, and also has the highest AUC iP/R score for the pairs task (Team 42, S1). The performance of a system that used information about existing interactions from MINT is shown for comparison (bottom, right).

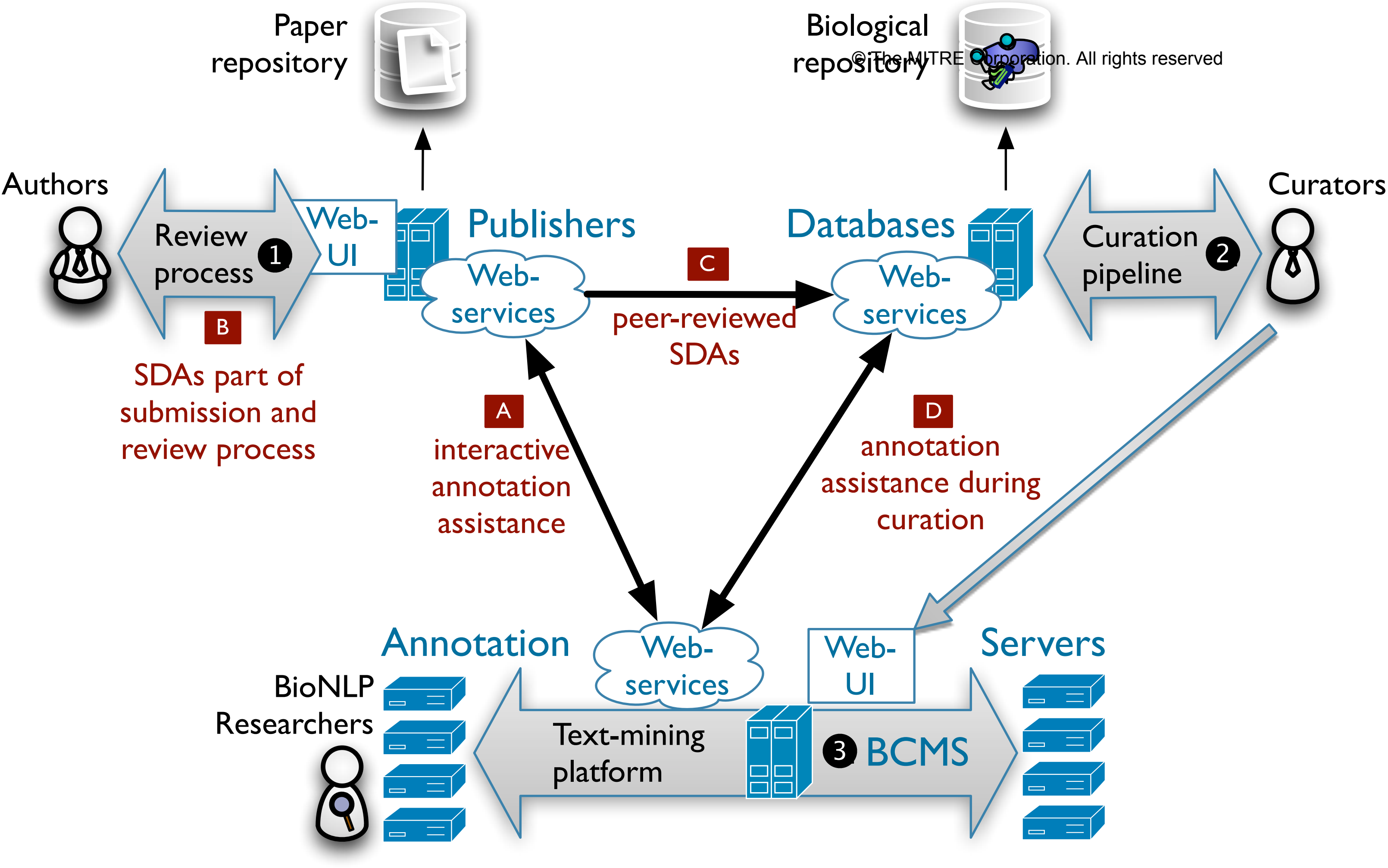
Supplement 1: SDAs in the publishing process.

Most of the workflow required to integrate structured information in scientific publications and databases is already in place (Supplemental Figure 1, 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), 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.

By setting up the necessary connections, SDAs can become an integral part of the publication and curation processes (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 BCMS when submitting papers for publication to the publisher’s server, using the author’s knowledge to reduce irrelevant results. 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. The benefits are considerable: publishers can provide practical structured access to their publications using the SDAs to catalog and index them. Making SDAs part of the review process would improve the quality of the annotations, and the SDAs can be directly integrated into biological repositories with the proper pointers to *and* from the textual sources. The curation process will become more efficient as the quality of peer-reviewed SDAs improves, and the process will benefit from text mining facilities to mine for database identifiers – currently, a very time consuming manual process. Text mining itself might improve because better systems can be developed by learning from the increased number of annotations brought into existence by the SDAs. The largest benefit for the scientific community will be the significant increase in coverage of the biological repositories.

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.


Paper repository
Biological repository
Authors
Review process
1
Web-UI
Publishers
Web-services
Databases
Web-services
Curation pipeline
2
Curators
C
peer-reviewed SDAs
B
SDAs part of submission and review process
A
interactive annotation assistance
D
annotation assistance during curation
Annotation
Web-services
Web-UI
Servers
BioNLP Researchers
Text-mining platform
3
BCMS

Supplement 2: Technical Explanations and Results

This section describes the challenges for automated systems, metrics for interactive curation, results, and definitions.

The Challenge of Identifying Interacting Proteins and Data Preparation

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.

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 the 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), 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. 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 inherent difficulty of disambiguating the article's species origins 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%. Combing the mapping and filtering step 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 16pp/37% for identifiers and 15pp/69% for pairs, or 2-3pp better than organism filtering only. The following table gives an overview of the mapping and filtering impacts for this highest-scoring system:

T42 S1 results	**Identifiers**		**Pairs**	
Filter	Macro-avrg'd F-Score	Macro-avrg'd AUC iP/R	Macro-avrg'd F-Score	Macro-avrg'd AUC iP/R
None (raw results)	0.429	0.386	0.221	0.194
Homon. ortholog map.	0.462	0.443	0.253	0.217
Organism filtering	0.560	0.493	0.349	0.287
Mapping & filtering	0.588	0.530	0.374	0.315

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-scores 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 below), then the results suggest that the systems could be used even in their present state to provide useful information to authors or curators.

Note: The BioCreative system results as provided in this publication are the data *after application of the homonym ortholog filtering process.* The full data set is not included because the size of the data is an order of magnitude larger but is available on request to FL. The F-measure is in terms of balanced F-measure on the entire filtered result set for each system.

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 an hour to annotate their articles and that the most time consuming task was finding the protein identifiers. 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 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 now 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 Table 1), and it would take approximately 15 minutes of time.

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

Results

Article Classification Task

The best performing article classification system had an AUC iP/R score of 0.7 at 92% accuracy (all results from each submission and task are listed in Supplement 3). 15/37 results achieve AUC iP/R scores over 0.5, and 20 results sets have an accuracy higher than 80%.

ACT Data Format in Provided Data Files

1. Article ID
2. Classification (true/false)
3. Rank
4. Confidence score (in the classification)

Protein Identification Task

For the protein identification task, the high-scoring AUC iP/R system, after finding 1/3 of the relevant IDs (recall), had nearly ¾ of its results correct (71% precision), and over the whole result set the system found ¾ of the relevant results at 14% precision. The protein identification system with the highest balanced F-score found about half of the IDs in the gold standard and ¾ of all reported results were correct. For the pairs task, the high-scoring balanced F-score system – which is the same system that had the highest F-score for identifiers and also the highest AUC iP/R score for pairs – found 1/3 of all pairs and half of its results were correct (see Table 1, right). Overall,14 out of 52 result sets from three teams for the interaction protein identification task had over 0.5 AUC iP/R.

INT Data Format in Provided Data Files (incl. FEBS Letters experiment data)

1. Article ID
2. UniProt accession
3. Rank (n/a for FEBS Letters experiment data [human annotation])
4. Confidence score (in the annotation)

Interaction Pairs Task

It is obvious that the pairs task is more difficult and has a lower performance than the identification of protein IDs due to the additional challenge of identifying correct combinations. The highest scoring system had an AUC iP/R score of 0.31 (see Table 1, right). Overall, 7 of the 41 result sets from four teams performed over 0.25 AUC iP/R in the pairs task, and more than half perform over 0.25 AUC iP/R.

INT Data Format in Provided Tables (incl. FEBS Letters experiment data)

1. Article ID
2. UniProt accession of partner A
3. UniProt accession of partner B
4. Rank (n/a for FEBS Letters experiment data [human annotation])
5. Confidence score (in the annotation)

Formulas and Calculation

Nomenclature

p – precision
r – recall
TP – true positive
FP – false positive
FN – false negative
TN – true negative

Accuracy

(TP + TN) / (TP + FP + FN + TN)

Area Under the Interpolated Precision/Recall Curve

The AUC **A** of the interpolated P/R function **f$_{pr}$** is defined as follows:

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

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

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'** ≥ **r**. (An algorithmic implementation would then seek for the max p's ("interpolated precision") by reversely traversing the recall/precision value pairs from highest to lowest recall.)

F-Measure

$F_\beta = (\beta^2 + 1) * p * r / (\beta^2) * p + r$

Where **β** is a measure of the relative importance of recall as compared to precision.

Balanced F or $F_1 = 2 (p * r) / (p + r)$

Matthew's Correlation Coefficient

(TP * TN – FP * FN) / sqrt((TP + FP)(TP + FN)(TN + FP)(TN + FN))

Precision

TP / (TP + FP)

Recall

TP / (TP + FN)

Sensitivity

TP / (TP + FN)

Specificity

TN / (TN + FP)