# An Overview of BioCreative II.5

Florian Leitner, Scott A. Mardis, Martin Krallinger, Gianni Cesareni, Lynette A. Hirschman, and Alfonso Valencia

**Abstract**—We present the results of the BioCreative II.5 evaluation in association with the FEBS Letters experiment, where authors created Structured Digital Abstracts to capture information about protein-protein interactions. The BioCreative II.5 challenge evaluated automatic annotations from 15 text mining teams based on a gold standard created by reconciling annotations from curators, authors, and automated systems. The tasks were to rank articles for curation based on curatable protein-protein interactions; to identify the interacting proteins (using UniProt identifiers) in the positive articles (61); and to identify interacting protein pairs. There were 595 full-text articles in the evaluation test set, including those both with and without curatable protein interactions. The principal evaluation metrics were the interpolated area under the precision/recall curve (AUC iP/R), and (balanced) F-measure. For article classification, the best AUC iP/R was 0.70; for interacting proteins, the best system achieved good macroaveraged recall (0.73) and interpolated area under the precision/recall curve (0.58), after filtering incorrect species and mapping homonymous orthologs; for interacting protein pairs, the top (filtered, mapped) recall was 0.42 and AUC iP/R was 0.29. Ensemble systems improved performance for the interacting protein task.

**Index Terms**—Text mining, text analysis, natural language processing, molecular biology, biological curation.

## 1 Introduction

Biologists, authors, and database curators all face difficulties when trying to make use of published text mining and information extraction systems implemented for the biomedical literature [1]. Several aspects make it cumbersome to determine which approaches are competitive for a particular task; these include heterogeneity of result formatting, different prior assumptions and data selection criteria underlying each system design, and the variability of evaluation settings [2]. The lack of independent evaluation data collections as well as limitations related to system accessibility have motivated a series of community challenges, carried out with the aim of gaining a better understanding of the performance and methods used to solve biologically relevant text mining tasks that could scale to real-world applications [3]. The BioCreative (Critical Assessment of Information Extraction Systems in Biology) challenges have attracted considerable interest, as reflected by the number of participating systems (over 40 for BioCreative II), and citations to the results [2], [3], as well as reuse of the resulting annotated corpora and evaluation software. The first two BioCreative challenges, BioCreative I and II, posed several tasks covering different levels of granularity and complexity, from the identification of biological entities within sentences to the extraction of complex biological annotations according to predefined literature curation guidelines followed by database annotators.

An important initial step for most biotext mining systems is the correct recognition of mentions of biological entities of interest, especially genes and proteins; this was evaluated via the Gene Mention task of BioCreative I and II [4], [5]. It is also of practical importance to provide links between each article and the unique database identifiers of the biological entities mentioned in these articles. Biological annotation databases (e.g., UniProt [6], or model organism databases) typically associate a unique identifer for each biological entity in the database. The focus of the Gene Normalization task in BioCreative I and II was to return such lists of gene/protein identifiers, given a collection of articles (abstracts). For BioCreative I [7], the task was to return unique gene identifiers associated with gene products from a set of abstracts for model organisms (fly, mouse, and yeast); for BioCreative II, the task focused on human gene products [8] and used EntrezGene as the source of gene identifiers. The Gene Normalization task addresses a common step carried out by most database curators, who typically provide their functional annotations through associations of normalized bioentities to controlled vocabulary or ontology terms. The extraction of such functional associations, namely, of human gene products to Gene Ontology terms, was pursued in the advanced task of BioCreative I [9], where protein-GO term relations as well as the corresponding evidence passages had to be extracted from full-text articles.

Several databases are engaged in manual annotation of protein-protein interactions (PPIs) from the literature, including MINT [10], IntAct [11], and BioGRID [12]. The automatic detection of PPIs from the literature has been the focus of multiple text mining systems. During BioCreative II, the PPI task evaluated the performance of automated systems for several tasks designed to follow the manual literature curation workflow [13]. This task covered

1. the detection and ranking of abstracts according to the relevance for deriving PPI annotations,

- *F. Leitner, M. Krallinger, and A. Valencia are with the Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. E-mail: valencia@cnio.es.*
- *S.A. Mardis and L.A. Hirschman are with MITRE Corporation, Bedford, MA 01730. E-mail: {mardis, lynette}@mitre.org.*
- *G. Cesareni is with the Department of Biology, University of Rome Tor Vergata, Rome, Italy, and IRCSS Santa Lucia Rome, Italy.*

*For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org.*
*Digital Object Identifier no. 10.1109/TCBB.2010.61.*

1545-5963/10/$26.00 © 2010 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

2. the extraction of the normalized protein interaction pairs,
3. the retrieval of suitable protein interaction evidence passages from full-text articles as well as
4. the automatic detection of the interaction detection experimental methods mentioned in the papers.

To ensure that the PPI annotations followed commonly used standards adopted by the biocuration community, the evaluation data was prepared by experienced curators from two different databases, MINT and IntAct [11].

The results obtained in the Gene Mention and Normalization tasks of BioCreative II suggested that a combined or ensemble system could provide improved predictive power by integrating the results from multiple systems. This motivated the implementation of the BioCreative Meta-Server platform (BCMS) [14], which is able to display and integrate multiple predictions from various annotation servers based on Web services and accessible to end users via a Web interface.

## 2 Background on SDAs and the *FEBS Letters* Experiment

The vast majority of scientific results in the biological domain are disseminated in a format that is optimized for human consumption while little effort is made by authors and publishers to make the published information suitable for automatic retrieval and processing.

The *FEBS Letters* editorial board, which convened in Vienna in July 2007, discussed the issue and resolved to start rectifying this trend by asking authors to provide structured information to be added to standard manuscript text. The structured information was meant to be appended to traditional abstracts as a structured summary. The goal was to maintain human readability while at the same time using database cross-references to precisely define the biological entities, and controlled vocabularies to specify their relationships. For the experimental phase, starting in March 2008, authors, after acceptance of their manuscripts, were asked to submit information related to protein interactions for which they were reporting new experimental evidence [15]. The author-submitted information was monitored and, whenever necessary, edited by professional curators of the MINT protein interaction database [16], [2] and finalized via a reciprocal exchange of information with the authors. After this experimental phase, the editors realized that this procedure entailing author involvement after manuscript acceptance was unduly slowing down publication times. The alternative of urging authors to include structured information before submission, as originally suggested by the MIMIx proposal [17], was not considered because of the prospect of discouraging authors with submission requests that were not yet enforced by competitor journals. Presently, *FEBS Letters* articles, and more recently, the *FEBS Journal*, offer structured digital abstracts but these are now produced by MINT curators who, after composing them, ask for authors' approval. This corpus of articles represents a unique, albeit limited, collection of scientific articles annotated by professional curators through a clearly defined procedure and as such represents an ideal benchmark for comparison of performance of natural language processing designed to extract biological information.

## 3 BioCreative II.5 Tasks

Following the needs of biocuration pipelines [18], such as for the MINT database, and in the light of the *FEBS Letters* experiment, three tasks were announced that were similar to BioCreative II. These tasks were chosen because they seemed feasible for text mining and were thought to be the most likely candidates for reducing manual (human) curation workload:

1. **Article categorization task (ACT):** Binary classification of articles (document classification) as relevant for curation, i.e., for extracting PPI annotations.
2. **Interactor normalization task (INT):** Lists of identifiers of proteins for which the article reports evidence for an interaction.
3. **Interaction pair task (IPT):** Lists of binary interaction pairs as protein identifier pairs per article.

For both the normalization (INT) and pair (IPT) tasks, the interacting proteins were required to have experimental evidence for their corresponding interaction in the article; it was not sufficient to include proteins described as interacting with each other in the text if the article did not present experimental evidence. The protein identifier space the systems had to select from was defined as the complete set of UniProt [6] major release 15.0 primary accessions. UniProt includes both manually curated SwissProt entries, as well as automatically generated entries in TrEMBL. For the PPI task, a great proportion of the identifiers in both the training and test sets come from the manually curated SwissProt [19] entries, with a much smaller subset from TrEMBL.

### 3.1 Classification: ACT

The binary classification (true/false) of full-text articles as relevant for curation—i.e., containing protein-protein interaction annotations—had to be reported together with a confidence score for the classification in the (0-1] range (i.e., excluding zero). Participants were provided with 595 full-text articles for both the training and test sets, out of which 61 articles were curation-relevant (after the challenge, two additional articles in the test set had to be reclassified as relevant, for a total of 63 articles in this set—see below). The articles were provided in two formats: raw XML with the corresponding DTD, and as extracted UTF-8 "plain-text" files in a format that preserved order, and section headers and titles.

### 3.2 Normalization: INT

For the normalization task (INT), participating teams were asked to return a ranked list of proteins that were detected as being used in an interaction with experimental evidence for a given interaction in the article ("interacting proteins"). These lists consisted of: the primary UniProt accessions ("protein identifiers"); a normalized confidence score (again, in the range (0-1]); a unique, positive integer rank for each protein identifier; and an optional evidence passage that triggered the extraction of the given normalization—although this optional evidence passage ultimately

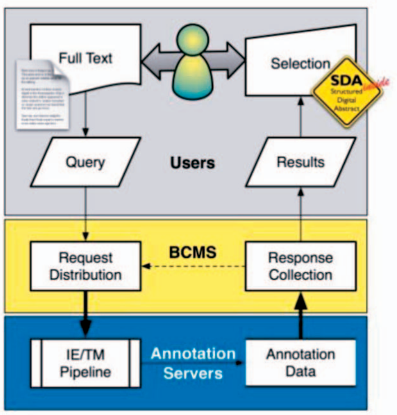


Fig. 1. The online scenario for the BioCreative II.5 challenge. This diagram depicts a possible scenario for the automated generation of annotations by online users (authors, curators, readers, etc.). Users would submit an article (full text) via a query to the BioCreative Meta-Server (BCMS). The BCMS then distributes the article to all known annotation servers (bold arrow down), which run the article through their text mining pipelines, and return the extracted data to the BCMS (bold arrow up). The BCMS then returns the valid data (existing UniProt accessions, correct data formats, etc.). If any problem occurred with an annotation server, the request is reissued to that server (dotted arrow). Otherwise, after collecting the responses, a consensus prediction result is delivered to the user. A user interface would allow the user to find relevant results for the Structured Digital Abstract. The upper, gray part is a hypothetical setup, while the colored lower parts are already in place and form the online scenario used in this challenge.

was not used by any of the teams. A total of 61 articles with annotations existed for both the training and test sets. However, all teams received the full 595 articles during the test phase and were not told which articles in the set were relevant until the end of the challenge (as this was part of the ACT task). All teams provided results for at least 21/61 articles in the test set, with an average of 55 annotated articles for the INT.

### 3.3 Interaction Pairs: IPT

For the pairs task, participants were asked to return a list of nonredundant, binary interaction pairs that have experimental evidence for the interaction in the article, again with a normalized confidence score (range: (0-1]) and a unique, positive integer rank. Directionality of the interactions was disregarded. Again, reporting an evidence passage for the classification was offered optionally, but was not used by any of the participants. As for the INT tasks, participants were given all 595 articles with no prior information about their relevance and were only evaluated on the articles for which they reported results out of the 61 relevant articles. The smallest result (from Team 42, see results) covered 21/61 articles, with the average result set reporting annotations on 47 articles.

### 3.4 The BioCreative Meta-Server

To make the challenge more realistic and to support a scenario of generating full-text annotations for Structured Digital Abstracts "on-the-fly" by an online mechanism, we asked participants to take part in a novel online experiment (see Fig. 1). To this end, we implemented a specialized version of the BioCreative Meta-Server [14]. Through this platform, participants could register their classification servers ("annotation servers") via a Web interface (up to five annotation servers per participating group), instruct the BCMS to send them annotation requests together with the next article in the queue, and review the current state of the process, including error and technical problem reports. This architecture uses a Web-service protocol (XML-RPC) to communicate between the BCMS and an annotation server. The Meta-Server sent the articles as UTF-8 formatted plain text to the annotation servers, which responded with a predefined data structure reporting results after having analyzed the text. The annotation servers had 10 min time to respond/produce annotations per article, although the average response time during the test phase was just slightly over 2 min/article. In total, 10 of the 15 teams agreed to participate in this advanced online setting, simulating a real-world annotation scenario. This makes BioCreative II.5, to the best of our knowledge the first online, live challenge carried out for text mining.

### 3.5 Challenge Metrics

The background of this challenge was to provide a human annotator (e.g., DB curator or the article's author) with the best possible set of machine-generated annotations, from which the annotator then would choose the relevant items. The most common utility measures for information extraction systems are *recall* and *precision*. Recall measures the "coverage" the result set has over the true results (commonly called "ground truth," or "gold standard"), while precision measures the percentage of correct answers in the result set relative to its complete size. Therefore, BioCreative II.5 used two common ways of measuring the quality of the results [20]: 1) Balanced *F-measure*, a metric that informs us about the overall quality of recall and precision as harmonic mean on the complete result set,

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

where p is precision, r is recall, and $\beta$ is a variable to balance the trade-off between precision and recall—for the balanced F-measure, $\beta$ is 1. Alternatively, 2) the area under the interpolated precision/recall curve (AUC iP/R),

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

where $p_i$ (*interpolated precision*) is defined as

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

and A the area under the curve, is a measure of (decreasing) precision at increasing levels of recall while iterating over a ranked list of results. In this context, AUC iP/R can be understood as a measure of quality for result ranking and favoring recall, while F-score in this context might be understood as measuring the quality of the overall result set (without taking ranking or confidence into account). High-recall systems might produce orders of magnitude larger results than high-precision systems, thereby gaining a large coverage over the true set (gold standard), but at an ultimately low precision for their complete result set, and achieving higher AUC iP/R scores if they use a good ranking scheme. High-precision systems, on the other hand, trade the coverage on the gold standard for significantly fewer, but very precise results, therefore potentially scoring higher in the F-measure.

For the article classification, F-measure would not be a very adequate estimate of overall set performance, as it does not take true negatives (TN) into account, which are known in the case of this task (as opposed to the INT and IPT tasks). Instead, we use two standard measures from Information Retrieval, *accuracy* and *Matthew's Correlation Coefficient* (MCC) [21]. Accuracy can be decomposed into two elements: 1) *Specificity*, a measure of how well the classifier was able to distinguish irrelevant results, where 100 percent specificity means no false positive (FP) was picked up by the system: $\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$) and 2) *Sensitivity*, a measure of how often the classifier misses out on relevant results ($\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$), where TP is true positive and FN is false negative), similar to the above Recall measure.

These two measures are then combined to establish the Accuracy of an approach:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}.$$

This measure therefore quantifies the degree of closeness of the system to the actual true value. It is most easily understood by comparison to Precision, through the "target analogy" on a dart board: accuracy is a measure of how well the darts' hits gravitate around the board's center (the "bulls eye," TP+TN), while precision measures how close together the darts were placed on the board (TP only), i.e., it is not possible to achieve high accuracy without also having a high precision.

In addition to accuracy, we employed the more reliable, yet similar MCC score, which produces the correlation coefficient between the observed (true) and the predicted (annotated) binary classifications in the range $[-1, 1]$. A coefficient of 1 would represent a perfect classification result, 0 an average, "random" result, and $-1$ an inverse classification:

$$\begin{aligned} MCC &= \sqrt{\frac{\chi^2}{n}} \\ &= \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \end{aligned}$$

MCC is also known as the $\Phi$ (phi) coefficient. The major advantage over accuracy is that the MCC is unbiased by sets where the two classes have very different sizes (as is the case in this challenge, with 61 true versus 534 false articles).

### 3.6 BioCreative Data Sets

Given the publicly known background of the *FEBS Letters* SDA experiment, the existing SDA-containing articles generated by the authors could not be used as a test set for the participants, since reproducing the results from available results would have been an all-too-easy way to circumvent this challenge. Instead, these SDA-containing articles, all from *FEBS Letters*' 2008 volume, were used as training set. For the test set, articles from *FEBS Letters*' 2007 volume were chosen because of their close temporal proximity; this also avoided the problem of introducing articles from different journals into the challenge that might create an additional bias in the sets due to the different interests of each journal. The MINT staff curated all articles and the classifications are available through the MINT database.

In addition to the online setting, we were interested in establishing the quality of text mining systems on *de novo* data. Therefore, participants were asked to not use the MINT database as a resource in their information extraction system, or report the use to us if they had. The use of existing annotation resources, contrary to *de novo* extraction, can be understood as "evidence mining," a task not directly relevant to the challenge's objectives. To ensure that all participants were honestly reporting the use of MINT, nine of the 61 annotated true articles in the test set were taken from the years 2002-2006 and are articles describing PPIs that had not yet been curated by the MINT team and did not form part of their database (or, as a matter of fact, of any of the IMEx consortium PPI DBs, such as IntAct) during the time of the challenge, a fact unknown to the participants during the test phase. Only one team reported the use of MINT for some of their result sets, and we observed—just for these runs—a significant performance difference between this "secret" set of nine articles and the "official" 52 test set articles. Thus, we can safely assume that all participants honestly attempted to generate novel results, unless they indicated otherwise.

To generate the gold standards for both the training and test sets, there was a combined curation effort between the MINT and BioCreative teams. Each article was annotated by at least two or more MINT curators. Additionally, annotations that the systems consistently were not able to reproduce (false positives made by all systems) were reexamined by the BioCreative team and the MINT curators, leading to two cases where the gold standard was corrected after the test phase. As a result, these efforts have produced a very high-quality ground truth data set available from the organizers (included in the "BioCreative II.5 Elsevier Corpus," available through the BioCreative Website at http://www.biocreative.org/).

In addition, as we already detailed elsewhere [22], we also measured the agreement scores between curators to establish an "upper bound" for the automated system performance. We measured the performance differences (agreement) of two MINT curators and an MINT curator compared to the curator of another database. Due to the different curation protocols between databases and their slightly different targets, agreement was lower in interdatabase measurements. We established an interdatabase agreement of 81 percent, while the intradatabase agreement increases this score to 93 percent (see [22]).

## 4 Results

In total, 15 teams registered for the challenge, and each team was allowed to participate in any number of the announced tasks (see Supplement 1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/2010.61, for a mapping between team IDs and the official participants). We received 37 test set results from eight teams for the ACT, 10 teams submitted 52 normalization runs (INT) on the test set, and another nine of the 15 teams participated in the interacting pairs task (IPT), sending us 45 result sets. All data and result sets are available online at http://www.biocreative.org/events/biocreative-ii5/results-and-data/.

Each participating system received all 595 articles from the test set without any indication which articles were relevant for extracting PPIs; however, we only evaluated relevant articles for which a team actually produced results. For the smallest result set, this means that the server actually produced results for 83 articles, out of which 21 were relevant and evaluated. There are several reasons for this evaluation strategy: first of all, as participants were not informed which articles would be relevant for INT and IPT, a system might decide autonomously if an article is relevant to limit workload. This is even more reasonable given that there was a very strict time limit for the online (but not the offline) part of the challenge, set to 10 minutes/article. Furthermore, since we believed that time would matter for the long-term intended applications—users would undoubtedly grow very unsatisfied if it took too long to generate results—we encouraged the teams to optimize the performance of their online servers. Finally, all results are evaluated using macroaveraging (i.e., the final score is calculated by averaging over the individual results per article). This means, the result reflects the average performance of a system on articles for which it actually does report results, and not the score of the system on the set, which would make the score relevant for the chosen test set, but would not inform us about the performance of the system on articles for which it produces results: It is obvious that a system has zero performance if it does not produce results for an article, but including these empty results in the final score would not let us estimate the performance of the system where it did produce results. Therefore, we stayed in the tradition of former BioCreative challenges and only evaluated articles for which the participants did submit results. However, this means that the recall scores are not comparable across systems that provided answers for a different number of documents.

In addition to the raw results presented here, we applied several postprocessing procedures to the annotations to explore how the raw results could be improved with minimal human intervention. For example, systems were not informed about the relevant (correct) species that the article is treating and had to disambiguate the most plausible species themselves. However, the human expert (e.g., the paper's author or a database curator) would likely know the species the article is describing, while for automated systems this is difficult to distinguish, because authors tend to not mention species at all, or list them in the methods section or list many different species, so that for an automated system, it is difficult to identify the relevant species for a given protein. Furthermore, the protein might come from several species, not only one—including cross-species interactions. Therefore, we used the gold standard UniProt identifier to establish the relevant (correct) species for the protein, and provided this as the hypothetical input from a human expert. This information can be used to run two postprocessing procedures on the results:

**1.** *Mapping homonym orthologs* to their correct results. For this process, we created clusters of all homonymous proteins in UniProt (protein and gene names, including synonymous—but excluding numeric and one-letter names), and used the UniRef50 clusters (sequences with at least 50 percent sequence identity, which is commonly used as an estimate to establish if a protein is homolog to another or not; see the Box on Homology for explanations of the concept of orthology). Establishing intersecting subsets from these two cluster sets produces homonym homolog sets. Then, for the mapping process, we extracted all homonym orthologs to gold standard annotations by only taking the proteins in a set where a gold standard protein occurs that are from different organisms than the gold standard protein (i.e., reducing the homologs to orthologs only, see Box 1). This correlation is then used to map any homonymous ortholog result to the correct gold standard entry. Therefore, it allows us to estimate the performance regarding the error introduced by choosing the right protein, but from the wrong species. This postprocessing step converted some false positives into true positives, but did not make large changes to the result size.

Box 1 – Homology: orthologs and paralogs

In genetics and proteomics, homology refers to similar (not necessarily identical) gene/protein sequences to a specified degree (expressed as a percentage value, e.g., UniRef50 are clusters of homologous protein sequences that share 50% sequence identity). Homologous sequences can occur in the same species' gene pool, or can be found when two sequences in distinct species share a similar sequence. The first case are called paralogous sequences and are thought to have arisen by a gene duplication event, after which each gene might assume the same or, more commonly, different functions in the cell. The latter case, when two distinct species share a similar sequence, is called orthology. This is thought to occur after a speciation event (i.e., the point at which one species gives rise to another or the point at which two species' common ancestor is replaced by two distinct lines). From there on, the gene pool of each species undergoes different evolutionary and therefore mutation patterns. The result is that the sequences start to diverge and may no longer be (nearly) exact copies. This might even result in functional paralogs, where these two genes start to carry out different tasks in each organism, but more commonly, the genes tend to maintain the same function with diverging sequences.

**2.** *Filtering wrong species* from the results. Instead of attempting to map wrong organism results to their correct entry, we can also just remove all organisms which the expert user does not want to see, again simulated by establishing the organism via the gold standard. In contrast to the mapping process in (1) above that increased recall, this filtering step removed many false positives from the result set, benefitting high recall systems that sacrificed precision by reporting a large quantity of homonymous proteins from many different organisms.

TABLE 1
ACT Results (Raw)

| Team | Run | TP | FP | FN | TN | Specificity | Sensitivity | Accuracy | MCC | P at full R | AUC iP/R |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | S04 | 60 | 273 | 3 | 259 | 0.487 | 0.952 | 0.536 | 0.272 | 0.138 | 0.464 |
| 9 | R01 | 42 | 107 | 21 | 425 | 0.799 | 0.667 | 0.785 | 0.331 | 0.110 | 0.291 |
| 9 | R02 | 44 | 118 | 19 | 414 | 0.778 | 0.698 | 0.770 | 0.329 | 0.107 | 0.298 |
| 9 | R03 | 42 | 114 | 21 | 418 | 0.786 | 0.667 | 0.773 | 0.316 | 0.106 | 0.281 |
| **9** | **R04** | 42 | 73 | 21 | 459 | 0.863 | 0.667 | 0.842 | 0.413 | **0.265** | 0.550 |
| 9 | R05 | 42 | 79 | 21 | 453 | 0.852 | 0.667 | 0.832 | 0.396 | 0.255 | 0.560 |
| 9 | S14 | 33 | 20 | 30 | 512 | 0.962 | 0.524 | 0.916 | 0.525 | 0.133 | 0.648 |
| 9 | S26 | 44 | 49 | 19 | 483 | 0.908 | 0.698 | 0.886 | 0.514 | 0.107 | 0.615 |
| **9** | **S27** | 20 | 5 | 43 | 527 | 0.991 | 0.317 | **0.919** | 0.472 | 0.176 | 0.568 |
| 9 | S28 | 26 | 10 | 37 | 522 | 0.981 | 0.413 | 0.921 | 0.508 | 0.113 | 0.675 |
| **9** | **S29** | 44 | 33 | 19 | 499 | 0.938 | 0.698 | 0.913 | **0.583** | 0.117 | 0.672 |
| 13 | R01 | 63 | 532 | 0 | 0 | 0.000 | 1.000 | 0.106 | 0.000 | 0.120 | 0.600 |
| 13 | R02 | 63 | 532 | 0 | 0 | 0.000 | 1.000 | 0.106 | 0.000 | 0.154 | 0.447 |
| 13 | R03 | 63 | 532 | 0 | 0 | 0.000 | 1.000 | 0.106 | 0.000 | 0.149 | 0.577 |
| 13 | R04 | 63 | 532 | 0 | 0 | 0.000 | 1.000 | 0.106 | 0.000 | 0.158 | 0.594 |
| 13 | R05 | 63 | 532 | 0 | 0 | 0.000 | 1.000 | 0.106 | 0.000 | 0.164 | 0.647 |
| 14 | R01 | 57 | 227 | 6 | 305 | 0.573 | 0.905 | 0.608 | 0.295 | 0.106 | 0.223 |
| 14 | R02 | 46 | 182 | 17 | 350 | 0.658 | 0.730 | 0.666 | 0.246 | 0.106 | 0.228 |
| 14 | R03 | 60 | 327 | 3 | 205 | 0.385 | 0.952 | 0.445 | 0.218 | 0.106 | 0.173 |
| 14 | R04 | 47 | 193 | 16 | 339 | 0.637 | 0.746 | 0.649 | 0.240 | 0.106 | 0.226 |
| **14** | **R05** | 61 | 518 | 2 | 14 | 0.026 | **0.968** | 0.126 | -0.010 | 0.106 | 0.127 |
| 14 | S08 | 56 | 350 | 7 | 182 | 0.342 | 0.889 | 0.400 | 0.153 | 0.115 | 0.208 |
| 14 | S25 | 56 | 350 | 7 | 181 | 0.341 | 0.889 | 0.399 | 0.152 | 0.115 | 0.208 |
| 16 | R01 | 22 | 37 | 41 | 495 | 0.930 | 0.349 | 0.869 | 0.288 | 0.108 | 0.395 |
| 16 | R02 | 17 | 49 | 46 | 483 | 0.908 | 0.270 | 0.840 | 0.174 | 0.122 | 0.278 |
| 16 | R03 | 21 | 38 | 42 | 494 | 0.929 | 0.333 | 0.866 | 0.270 | 0.120 | 0.394 |
| 16 | R04 | 33 | 30 | 30 | 502 | 0.944 | 0.524 | 0.899 | 0.467 | 0.141 | 0.550 |
| 16 | R05 | 27 | 36 | 36 | 496 | 0.932 | 0.429 | 0.879 | 0.361 | 0.180 | 0.454 |
| **20** | **S32** | 37 | 23 | 26 | 509 | 0.957 | 0.587 | 0.918 | 0.556 | 0.196 | **0.699** |
| 20 | S33 | 54 | 84 | 9 | 448 | 0.842 | 0.857 | 0.844 | 0.510 | 0.210 | 0.657 |
| 31 | S18 | 36 | 30 | 27 | 502 | 0.944 | 0.571 | 0.904 | 0.505 | 0.109 | 0.435 |
| 31 | R01 | 34 | 22 | 29 | 510 | 0.959 | 0.540 | 0.914 | 0.525 | 0.107 | 0.413 |
| 31 | R02 | 34 | 22 | 29 | 510 | 0.959 | 0.540 | 0.914 | 0.525 | 0.107 | 0.413 |
| 31 | R03 | 34 | 22 | 29 | 510 | 0.959 | 0.540 | 0.914 | 0.525 | 0.106 | 0.505 |
| **31** | **R04** | 13 | 2 | 50 | 530 | **0.996** | 0.206 | 0.913 | 0.398 | 0.107 | 0.315 |
| **31** | **R05** | 13 | 2 | 50 | 530 | **0.996** | 0.206 | 0.913 | 0.398 | 0.107 | 0.315 |
| 32 | S07 | 30 | 92 | 0 | 0 | 0.000 | 1.000 | 0.246 | 0.000 | n/a | 0.149 |

*All results from the ACT. Online submissions are denoted by "S," offline submissions by "R." The highest score for each category is marked in bold. "P at full R" refers to the precision measured at the point where all true articles are found in the ranked list (full recall). The best classification system was S29 from Team 9 (MCC 0.583), the best ranking system S32 from Team 20 (AUC iP/R 0.699). Team 32 only submitted one incomplete classification run where data was sent unintentional and shown as reference only. Team 13 classified all articles as positive, and therefore, has an MCC score of zero.*

**3.** Finally, the *combination* of both approaches can be used to estimate the "optimal result": mapping all gold standard orthologs and filtering any leftover wrong species. This is an optimistic score, because it does not include all possible homonym ortholog mappings, and assumes that the human user would have already chosen which mappings are correct, filtering irrelevant mappings. It also can be understood as an estimate of the best possible performance if the systems would have known the species a priori.

Table 1 shows the raw results on the test set for each system in the article classification task, Table 2 for the normalization (identifier) task, and Table 3 for the interaction pair task. In Table 4, the various effects of postprocessing the results with the described approaches is shown for the highest recall (AUC iP/R) and highest precision (F-measure) systems after applying the corresponding procedure in the normalization and pair tasks (for the ACT, no such postprocessing can be made, obviously).

Another way of investigating the relative performance of systems is by plotting the results in a 2D scatter plot using the F-measure and AUC iP/R results, where high precision and high recall systems occupy the highest values on their respective axis. In addition, systems that balance the two approaches become apparent, as they are close to the top right corner (e.g., Team 18 for INT and Team 42 for IPT). For the raw INT results, this is shown in Fig. 2, for the raw IPT scores, the plot is shown in Fig. 3.

## 5 Analysis of Results

### 5.1 Article Classification

The first task in a curation pipeline is the identification of relevant articles. In the context of PPIs, this would relate to the identification of articles that report interactions with experimental evidence. As explained, participants were asked to return a Boolean classification for each article, together with a confidence score for this classification. For offline results (not submitted via the BCMS), participants were also allowed to rank the results from the most to least relevant article. In the case of the online results, this ranking was established by ordering the articles from the largest confidence score applied to a true classification to the largest confidence score for a false classification. Therefore, Accuracy and MCC measure how well the binary classification was done, while the main evaluation function, AUC iP/R, measures how well the confidence score/ranking scheme

TABLE 2
INT Results (Raw)

| Team | Run | EvDoc | TP | FP | FN | Precision | Recall | F-Score | AUC iP/R |
|---|---|---|---|---|---|---|---|---|---|
| **10** | **R01** | 61 | 158 | 20730 | 94 | 0.012 | 0.652 | 0.024 | **0.435** |
| 10 | R02 | 61 | 158 | 20688 | 94 | 0.012 | 0.652 | 0.024 | 0.432 |
| 10 | R03 | 61 | 157 | 20715 | 95 | 0.012 | 0.646 | 0.024 | 0.429 |
| 10 | R04 | 61 | 158 | 20730 | 94 | 0.012 | 0.652 | 0.024 | 0.433 |
| 10 | R05 | 61 | 162 | 20980 | 90 | 0.013 | 0.676 | 0.024 | 0.431 |
| 10 | S09 | 61 | 146 | 1638 | 106 | 0.087 | 0.591 | 0.145 | 0.428 |
| 10 | S21 | 61 | 157 | 12798 | 95 | 0.015 | 0.646 | 0.028 | 0.432 |
| 10 | S23 | 61 | 157 | 12713 | 95 | 0.015 | 0.646 | 0.028 | 0.428 |
| 10 | S24 | 61 | 157 | 12798 | 95 | 0.015 | 0.646 | 0.028 | 0.432 |
| 14 | R01 | 56 | 60 | 139 | 173 | 0.304 | 0.291 | 0.277 | 0.247 |
| 14 | R02 | 45 | 44 | 110 | 127 | 0.277 | 0.283 | 0.263 | 0.258 |
| 14 | R03 | 60 | 83 | 394 | 165 | 0.192 | 0.359 | 0.224 | 0.276 |
| 14 | R04 | 46 | 47 | 132 | 127 | 0.251 | 0.291 | 0.259 | 0.260 |
| 14 | S08 | 48 | 28 | 144 | 153 | 0.165 | 0.161 | 0.149 | 0.133 |
| 14 | S25 | 57 | 43 | 198 | 200 | 0.190 | 0.188 | 0.172 | 0.149 |
| 18 | R01 | 60 | 133 | 771 | 116 | 0.206 | 0.561 | 0.275 | 0.373 |
| 18 | R02 | 59 | 82 | 238 | 165 | 0.253 | 0.355 | 0.278 | 0.306 |
| 18 | R03 | 60 | 104 | 415 | 145 | 0.234 | 0.439 | 0.284 | 0.338 |
| ***18*** | ***R04*** | 60 | 105 | 414 | 144 | 0.236 | 0.441 | ***0.286*** | 0.330 |
| 18 | R05 | 59 | 101 | 377 | 146 | 0.230 | 0.415 | 0.276 | 0.309 |
| 22 | R01 | 61 | 140 | 4993 | 112 | 0.029 | 0.596 | 0.055 | 0.373 |
| 22 | R02 | 61 | 157 | 32346 | 95 | 0.006 | 0.679 | 0.012 | 0.357 |
| 22 | R03 | 61 | 140 | 5003 | 112 | 0.029 | 0.596 | 0.054 | 0.381 |
| **22** | **R04** | 61 | 158 | 31361 | 94 | 0.006 | **0.683** | 0.013 | 0.379 |
| 22 | R05 | 61 | 131 | 6738 | 121 | 0.020 | 0.544 | 0.038 | 0.310 |
| 22 | S05 | 61 | 134 | 4796 | 118 | 0.028 | 0.584 | 0.052 | 0.301 |
| 22 | S10 | 60 | 131 | 6746 | 107 | 0.019 | 0.593 | 0.037 | 0.237 |
| 22 | S11 | 61 | 134 | 4792 | 118 | 0.028 | 0.576 | 0.052 | 0.303 |
| 22 | S12 | 61 | 131 | 4197 | 121 | 0.031 | 0.579 | 0.058 | 0.292 |
| 22 | S13 | 61 | 120 | 5492 | 132 | 0.022 | 0.513 | 0.041 | 0.242 |
| 26 | S16 | 50 | 37 | 1965 | 179 | 0.028 | 0.181 | 0.045 | 0.079 |
| 31 | S18 | 34 | 55 | 285 | 95 | 0.162 | 0.397 | 0.218 | 0.181 |
| 31 | R01 | 32 | 46 | 274 | 90 | 0.144 | 0.397 | 0.201 | 0.160 |
| 31 | R02 | 32 | 46 | 274 | 90 | 0.144 | 0.397 | 0.201 | 0.164 |
| 31 | R03 | 32 | 48 | 272 | 88 | 0.150 | 0.406 | 0.209 | 0.178 |
| 31 | R04 | 61 | 136 | 15002 | 116 | 0.009 | 0.582 | 0.018 | 0.019 |
| 31 | R05 | 32 | 63 | 2588 | 73 | 0.028 | 0.511 | 0.052 | 0.043 |
| 32 | R01 | 61 | 105 | 1592 | 147 | 0.068 | 0.444 | 0.113 | 0.178 |
| 32 | S07 | 61 | 101 | 1575 | 151 | 0.067 | 0.420 | 0.110 | 0.166 |
| 37 | R01 | 52 | 89 | 582 | 127 | 0.212 | 0.485 | 0.235 | 0.242 |
| 37 | R02 | 61 | 149 | 6187 | 103 | 0.025 | 0.604 | 0.048 | 0.048 |
| 37 | R03 | 54 | 98 | 486 | 128 | 0.221 | 0.503 | 0.257 | 0.304 |
| 37 | R04 | 61 | 157 | 5327 | 95 | 0.031 | 0.636 | 0.058 | 0.055 |
| **42** | **S01** | 21 | 38 | 62 | 54 | **0.434** | 0.482 | **0.429** | 0.386 |
| 42 | S02 | 56 | 100 | 603 | 141 | 0.167 | 0.475 | 0.219 | 0.303 |
| 42 | S03 | 58 | 96 | 681 | 150 | 0.144 | 0.451 | 0.200 | 0.258 |
| 42 | S19 | 39 | 46 | 223 | 111 | 0.195 | 0.333 | 0.226 | 0.227 |
| 42 | S20 | 55 | 119 | 1243 | 116 | 0.105 | 0.535 | 0.167 | 0.333 |
| 51 | R01 | 61 | 22 | 837 | 230 | 0.006 | 0.091 | 0.012 | 0.048 |
| 51 | R02 | 61 | 42 | 1451 | 210 | 0.013 | 0.177 | 0.023 | 0.085 |
| 51 | R03 | 61 | 130 | 9902 | 122 | 0.018 | 0.530 | 0.034 | 0.138 |
| 51 | R04 | 61 | 127 | 9561 | 125 | 0.019 | 0.519 | 0.035 | 0.137 |

*All results from the INT. Online submissions are denoted by "S," offline submissions by "R." The highest score for each category is marked in bold. "EvDoc" refers to the number of relevant documents for which the system returned results and on which it was evaluated. TP: true positives, FP: false positives, FN: false negatives. Team 42 submitted the highest F-score run (S01, 0.429), while Team 10 achieved the highest AUC iP/R (R01, 0.435). However, Team 42's run S01 only accounts for 1/3 of the documents in the set (21/61), while the next best F-score run from Team 18 (R04, bold italic)—at a significantly lower F-score of 0.286—submitted results for all but one article.*

fits the ground truth. Some participants only assigned one class to all articles, resulting in possibly good AUC iP/R scores at low accuracy and a correlation coefficient of zero.

The article classification task in this challenge seems to have been harder than most former challenges, which most likely can be attributed to the fact that former BioCreatives used abstracts only, while having to classify an article based on the complete text is possibly a harder task. However, with both a sensitivity and sensitivity score of about 85 percent measured in one of the runs submitted by the two teams with exceptionally high MCC and AUC iP/R scores (Teams 9 and 20), these systems seem more than fit even when working on full text.

We analyzed the results by dividing the true and false articles into two separate sets and plotting each article against its average rank (x-axis) and the standard deviation of the rank (y-axis) assigned by each result set we received. When we do this, several clusters of articles become apparent (see Fig. 4). The upper half shows the distribution for positive (true) articles, the lower half for the negative (false) articles. On the (horizontal) x-axis, the average rank assigned by the best runs from each team is used; ranks are weighted by the AUC iP/R score of the run. On the (vertical) y-axis, the standard deviation of each article's rank distribution within the used runs is plotted. This means, the closer an article is to the left, the more systems, on average, evaluated

TABLE 3
IPT Results (Raw)

| Team | Run | EvDoc | TP | FP | FN | Precision | Recall | F-Score | AUC iP/R |
|---|---|---|---|---|---|---|---|---|---|
| 14 | R01 | 56 | 22 | 170 | 174 | 0.128 | 0.146 | 0.116 | 0.128 |
| 14 | R02 | 45 | 15 | 101 | 126 | 0.116 | 0.129 | 0.109 | 0.123 |
| 14 | R03 | 60 | 33 | 561 | 180 | 0.074 | 0.193 | 0.086 | 0.155 |
| 14 | R04 | 46 | 18 | 122 | 125 | 0.120 | 0.145 | 0.121 | 0.134 |
| 14 | S08 | 48 | 7 | 116 | 141 | 0.037 | 0.038 | 0.036 | 0.035 |
| 14 | S25 | 57 | 15 | 268 | 196 | 0.068 | 0.083 | 0.063 | 0.052 |
| **18** | **R05** | 39 | 34 | 76 | 126 | **0.290** | 0.236 | **0.222** | 0.208 |
| 22 | R01 | 61 | 66 | 36251 | 150 | 0.002 | 0.349 | 0.004 | 0.168 |
| **22** | **R02** | 61 | 74 | 151621 | 142 | 0.000 | **0.435** | 0.001 | 0.153 |
| 22 | R03 | 61 | 65 | 36348 | 151 | 0.002 | 0.345 | 0.004 | 0.174 |
| 22 | R04 | 61 | 78 | 151617 | 138 | 0.001 | 0.447 | 0.001 | 0.171 |
| 22 | R05 | 61 | 52 | 60016 | 164 | 0.001 | 0.253 | 0.002 | 0.088 |
| 22 | S05 | 61 | 54 | 15034 | 162 | 0.004 | 0.299 | 0.007 | 0.079 |
| 22 | S10 | 60 | 48 | 14847 | 155 | 0.003 | 0.303 | 0.006 | 0.072 |
| 22 | S11 | 61 | 53 | 15035 | 163 | 0.004 | 0.294 | 0.007 | 0.097 |
| 22 | S12 | 61 | 55 | 13575 | 161 | 0.004 | 0.333 | 0.008 | 0.069 |
| 22 | S13 | 61 | 40 | 15114 | 176 | 0.003 | 0.194 | 0.005 | 0.036 |
| 26 | S16 | 50 | 2 | 5185 | 183 | 0.000 | 0.008 | 0.000 | 0.000 |
| 31 | S18 | 34 | 7 | 48 | 130 | 0.146 | 0.066 | 0.065 | 0.057 |
| 31 | R01 | 32 | 2 | 49 | 123 | 0.063 | 0.034 | 0.036 | 0.034 |
| 31 | R02 | 32 | 3 | 47 | 122 | 0.094 | 0.044 | 0.051 | 0.044 |
| 31 | R03 | 32 | 9 | 40 | 116 | 0.180 | 0.108 | 0.116 | 0.096 |
| 31 | R04 | 61 | 10 | 6835 | 206 | 0.003 | 0.067 | 0.005 | 0.006 |
| 31 | R05 | 32 | 1 | 333 | 124 | 0.001 | 0.010 | 0.002 | 0.000 |
| 32 | R01 | 29 | 6 | 57 | 82 | 0.123 | 0.101 | 0.103 | 0.086 |
| 32 | S07 | 29 | 6 | 57 | 82 | 0.123 | 0.101 | 0.103 | 0.091 |
| 37 | R03 | 52 | 35 | 820 | 148 | 0.080 | 0.278 | 0.090 | 0.150 |
| 37 | R04 | 52 | 36 | 811 | 147 | 0.084 | 0.287 | 0.094 | 0.150 |
| 37 | R05 | 46 | 37 | 783 | 139 | 0.090 | 0.307 | 0.094 | 0.177 |
| 37 | R06 | 54 | 43 | 605 | 147 | 0.103 | 0.323 | 0.116 | 0.187 |
| **37** | **R07** | 45 | 41 | 463 | 134 | 0.115 | 0.347 | 0.123 | **0.223** |
| 42 | S01 | 21 | 15 | 113 | 60 | 0.213 | 0.296 | 0.221 | 0.194 |
| 42 | S07 | 56 | 35 | 1880 | 175 | 0.022 | 0.226 | 0.031 | 0.061 |
| 42 | S03 | 58 | 25 | 1428 | 188 | 0.019 | 0.173 | 0.030 | 0.053 |
| 42 | S19 | 39 | 14 | 418 | 113 | 0.027 | 0.098 | 0.037 | 0.033 |
| 42 | S20 | 55 | 54 | 3498 | 149 | 0.027 | 0.325 | 0.046 | 0.079 |
| 51 | R01 | 27 | 4 | 1954 | 111 | 0.002 | 0.059 | 0.005 | 0.020 |
| 51 | R02 | 36 | 11 | 3222 | 130 | 0.018 | 0.126 | 0.027 | 0.034 |
| 51 | R03 | 61 | 24 | 19001 | 192 | 0.008 | 0.160 | 0.013 | 0.046 |

*Official results from the IPT. Online submissions are denoted by "S," offline submissions by "R." The highest score for each category is marked in bold. "EvDoc" refers to the number of relevant documents for which the system returned results and on which it was evaluated. TP: true positives, FP: false positives, FN: false negatives. Team 18 submitted the highest F-score run (R05, 0.222), while Team 37 achieved the highest AUC iP/R (R07, 0.223). Four runs from Team 18 were removed because they used MINT data (R05 also produces [correct] results on the "secret set"—see text); these runs achieve significantly higher scores. And two runs from Team 37 were removed, as the limit was five runs for online and offline submissions each.*

TABLE 4
INT and IPT Results after Postprocessing for the Highest Scoring Teams

| Task | Class | Team Run | F-Score | Team Run | AUC iP/R |
|---|---|---|---|---|---|
| INT | raw | T18 R04 | 0.286 | T10 R01 | 0.435 |
| | raw mapped | | 0.300 | | 0.478 |
| | raw filtered | | 0.379 | | 0.527 |
| | raw map. & filt'd. | | 0.396 | | 0.573 |
| | best mapped | T42 S01 | 0.462 | T10 R05 | 0.478 |
| | best filtered | | 0.560 | | 0.528 |
| | best map. & filt'd. | | 0.588 | | 0.575 |
| IPT | raw | T18 R05 | 0.222 | T37 R07 | 0.223 |
| | raw mapped | | 0.222 | | 0.252 |
| | raw filtered | | 0.291 | | 0.288 |
| | raw map. & filt'd. | | 0.291 | | 0.312 |
| | best mapped | T42 S01 | 0.253 | T37 R07 | n/a |
| | best filtered | | 0.349 | | n/a |
| | best map. & filt'd. | | 0.374 | T42 S01 | 0.315 |

*This table shows the improvements of the final scores when postprocessing the results with the described steps (see text). For each task, the best system from the raw results is shown with its initial score and the improvement from homonym ortholog mapping (only), organism filtering (only), and the combined step of mapping and filtering. Additionally, where applicable, the system that would achieve the highest result after the postprocessing step is shown (as it is possible that a formerly lower ranked system achieves a better postprocessing result than the best raw result system). In total, four outstanding teams are shown here: Teams 18 and 42, which dominate both tasks in the F-score rankings with high-precision systems, Team 10 with the best normalization systems, and Team 37 with the best interaction pair classification system.*

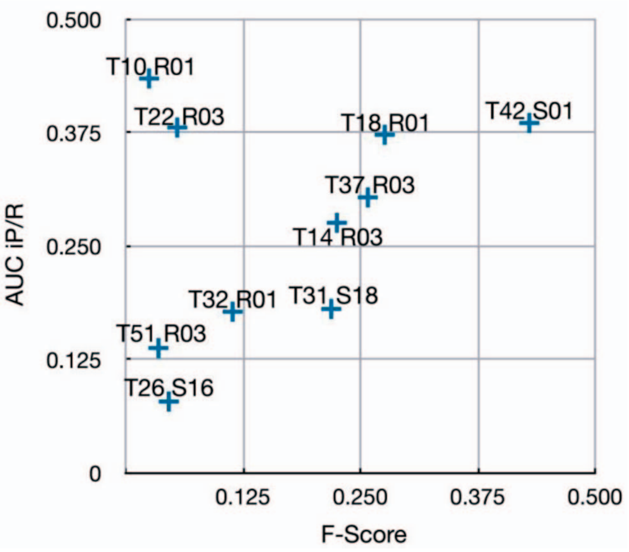


Fig. 2. INT plot of F-score versus AUC iP/R. The scatter plot shows the best AUC iP/R runs (y-axis, "ranking & recall") of each team against their corresponding F-score results (x-axis, "overall set score & precision").

the article as relevant, the closer it is to the bottom, the higher the agreement was between the systems about its placement. At an StdDev of 225 ranks—meaning ranks were spread out nearly over all the possible ranks—it becomes apparent that systems agreed much less on the placement of positive than negative articles (1/3rd versus 1/30th of the articles have a rank StdDev over 225). There are 17 articles in the positive set that were identified as relevant by virtually all systems ("easy articles," see figure), as well as two more articles in the negative set that fall in this same category but were falsely identified as negative in the initial gold standard (two errors in GS, see figure and following paragraph). About 13 articles in the positive set were consistently hard to identify, while over 36 articles in the negative set were clearly classified as highly irrelevant by most systems. Finally, articles at the lower right front of the negative articles were shown to contain nonrelevant genetic interactions.

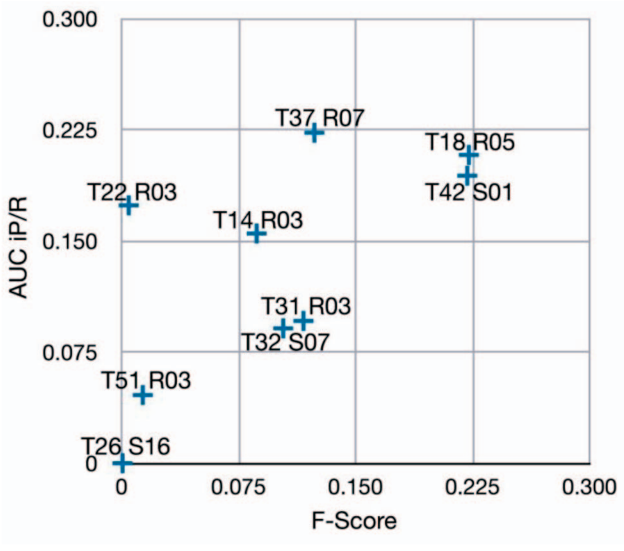


Fig. 3. IPT plot of F-score versus AUC iP/R. The scatter plot shows the best AUC iP/R runs (y-axis, "ranking & recall") of each team against their corresponding F-score results (x-axis, "overall set score & precision").

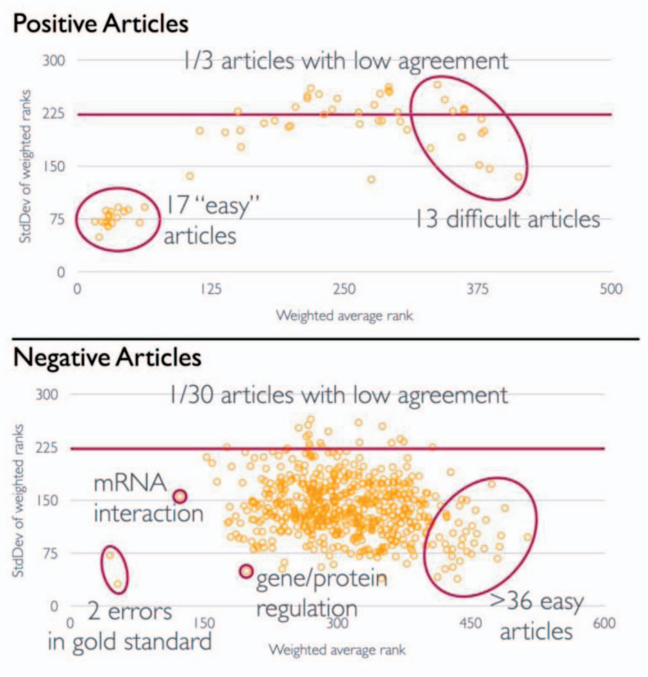


Fig. 4. Plotting true and false articles by average rank and standard deviation. See text for explanations.

We analyzed some of the properties of the articles in these clusters, which helped to establish the possible nature of the difficulties encountered. As the task is to identify curation-relevant articles describing PPIs, we hypothesized that the word "interaction" might be especially common in the true set. As a matter of fact, the noun "interaction" is the second most frequent noun in the positive set (after "cell," the most frequent noun in both sets), and occurs, on average, three times more often per article in the positive than the negative set. Furthermore, for the cluster of articles most systems clearly identified as relevant ("easy articles"), this frequency increases to nearly eight times higher than in the negative set. By contrast, hard to identify articles have an average frequency of this term lower than even the average negative set frequency, and commonly only contain indirect interaction descriptions, such as protein phosphorylation descriptions. Furthermore, when looking into the false positives, they mostly contain non-PPIs, such as genetic interactions (e.g., promotor binding). Finally, this clustering prominently distinguished two articles in the negative test set that all systems rather consistently classified as relevant. Investigating these two cases showed that these two samples were actually true articles that were missed by the curators. They was a posteriori reclassified as positive in the gold standard and led to the slight imbalance between true and negative articles in the training and test sets (61/534 in the training set, 63/532 in the test set).

## 5.2 Protein Normalization (INT)

The normalization results might seem surprisingly low when compared to results calculated from our gene/protein normalization corpora from BioCreative I or II. However, there are several aspects that are unique to BioCreative II.5 (and the PPI task in BioCreative II) and make it several

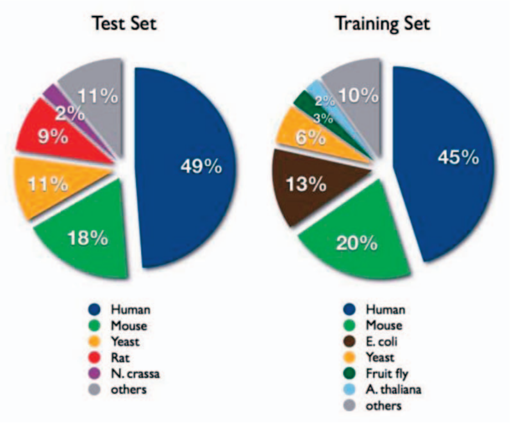


Fig. 5. Species differences between training and test set. The two pie charts show the distribution of species among the proteins in the training and test set. Although the differences look large here, the overall “agreement” between the two sets (number of proteins with matching species in the other set/total number of proteins in both sets) is 82 percent. The main difference is accounted for by E. coli proteins in the training set (13 percent of training set annotations), for which no annotation exists in the test set.

degrees more challenging over the data found in currently existing normalization corpora. First of all, the mappings had to be found in the full text of the articles and not just in a limited text space such as sentences or paragraphs, and participants did not receive high-quality training data with exact annotations of protein mentions that give rise to the corresponding normalizations. Second, there was no limitation in species, meaning that the systems had to disambiguate species themselves and were never told which species they should map to. This problem is amplified by taking into account that about one-quarter of the protein annotations in the test set belong to species that do not occur in the training set and vice versa (see Fig. 5). Finally, the goal was to return normalizations for proteins that actually have an interaction description in the article and are backed by experimental evidence. This meant that although systems were correctly reporting identifiers for proteins mentioned in the article, these were counted as false positives because of this limitation. This factor creates a decreased precision score for the systems (while recall is not commonly much affected, see Fig. 6). Therefore, the results of this challenge underestimate the ability of the automated systems to correctly map protein mentions to database identifiers, and these results should not be compared to regular normalization tasks. However, by modeling the challenge on a real curation task, we believe that we have created a setup that simulates the problems that automated systems would need to deal with when applied to large-scale biological data mining scenarios that commonly impose very specific requirements about the information they need to extract. Fig. 6 shows how the true annotation to proteins with experimental evidence influences the result evaluation. If all proteins mentioned in the article counted as correct results, a portion of the false positives that systems report would change to correct annotations $(\mathrm{FP}_2 \rightarrow \mathrm{TP}_2)$, although it would also increase the number of false negatives $(\mathrm{n/a} \rightarrow \mathrm{FN}_2)$. The net effect, however, is an average increase in precision when all mentions can be annotated, while recall stays roughly equal. A common counterargument to this assumption is the “Cooperative Effect”: For example, given a system reporting on an article annotating mouse proteins: if that system had missed the mouse association in the first place (e.g., falsely reporting only human proteins), it would actually only decrease the system’s scores if all mouse protein mentions were included.

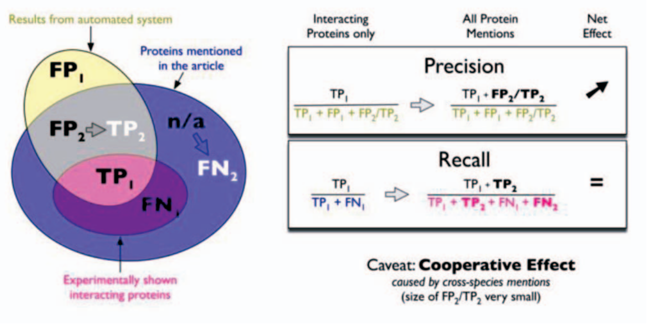

Results from automated system
Proteins mentioned in the article
$\mathrm{FP}_1$
$\mathrm{FP}_2 \Rightarrow \mathrm{TP}_2$
n/a
$\mathrm{FN}_2$
$\mathrm{TP}_1$
$\mathrm{FN}_1$
Experimentally shown interacting proteins
Interacting Proteins only
All Protein Mentions
Net Effect
Precision
$\frac{\mathrm{TP}_1}{\mathrm{TP}_1 + \mathrm{FP}_1 + \mathrm{FP}_2/\mathrm{TP}_2} \Rightarrow \frac{\mathrm{TP}_1 + \mathrm{FP}_2/\mathrm{TP}_2}{\mathrm{TP}_1 + \mathrm{FP}_1 + \mathrm{FP}_2/\mathrm{TP}_2}$ ↗
Recall
$\frac{\mathrm{TP}_1}{\mathrm{TP}_1 + \mathrm{FN}_1} \Rightarrow \frac{\mathrm{TP}_1 + \mathrm{TP}_2}{\mathrm{TP}_1 + \mathrm{TP}_2 + \mathrm{FN}_1 + \mathrm{FN}_2}$ =
Caveat: Cooperative Effect
caused by cross-species mentions
(size of $\mathrm{FP}_2/\mathrm{TP}_2$ very small)

Fig. 6. The effect of the experimental evidence annotation restriction on precision.

## 5.3 Interaction Pairs (IPT)

The same logic applies to the interaction pair results, where the difficulty is increased because of the squared number of possibilities of combining binary groupings after normalization, and also because, in quite a few cases, *in vitro* cross-species interactions are reported. It is nontrivial to train a system that can report cross-species interactions without incurring a large negative cost caused by introducing many false positive interactions. The effect of this factor can be seen in the difference between the raw results and the homonym ortholog evaluation of the IPT.

However, for many scientific annotation needs, identifying the proteins that actually have experimental evidence is important; this “limitation” is required to create scientifically sound protein interaction maps in Systems Biology that are backed up by empirical evidence. On a side note, given that for the biological use of this data, the experimental evidence is of utmost importance, it is rather discouraging that no team reported the textual evidence from which it had drawn the classification (“evidence passage”).

Interesting aspects are the differences between techniques applied by the teams, including differences between each of the runs they submitted. For detailed descriptions of each system, see the corresponding publication from each team in this special issue. However, we asked the participants to fill in a short questionnaire after the test phase, which all but one team completed.

From the three strongest teams in the ACT (9, 20, and 31), Team 31 reported using support vector machines (SVM) for classification (the other two did not disclose their supervised learning technique), and none of them used part-of-speech (POS) tagging or any specialized parsing techniques. In general, only Team 31, which also participated in the two other tasks (teams 9 and 20 only participated in the ACT), used a specialized natural language processing in biology (BioNLP) pipeline that could handle problems such as species mentions or protein normalization. We therefore can conclude that this aspect does not seem critical to this task, and we know from the earlier BioCreative II that SVMs are very powerful at document classification.

The highest scoring teams in the normalization task (10 and 42) both used conditional random fields and

Team 10 used a broad selection of NLP techniques, including conjunction handling, pattern matching, POS tagging, and shallow parsing. However, both teams had methods in place to handle the necessary processing for this task: recognition of species mentions, normalization of species (to their taxonomic ID), and disambiguation of protein mentions and of protein-organism associations. The only difference here was that Team 10 additionally employed biosyllable handling (that is, for example, special handling for words with biologically significant endings, such as -ase in terms like "protein kinase").

For the most successful teams in the IPT (teams 18, 37, and 42), the approaches seem to be technically quite variable: Team 42 only used pattern matching, Team 18 employed conjunction handling including shallow parsing, while Team 37 employed a very wide range of NLP techniques, even deep parsing of the grammatical structures of the text. Regarding biologically relevant language processing techniques, again all three teams made use of a wide range of the commonly used approaches mentioned for INT, and only Team 18 had a way of handling biosyllables. In general, although the use of these BioNLP techniques is no guarantee for excellent performance, it seems impossible to achieve significant results without them.

Regarding NLP approaches, all but one team (Team 9) had specialized approaches for special characters (e.g., greek letters, roman numerals, etc.). Teams 9, 10, and 22 used additional (nonbiological) NLP resources for training their classifiers, and teams 9, 22, and 42 used additional training material on top of the provided training set. Given that all these four teams achieved noteworthy results, this is another factor that might have contributed to their high performance. Most of these mentioned high-scoring teams (except 20, 31, and 37) used additional biological resources (e.g., protein databases, MeSH terms, etc.) in their systems. Common NLP resources used by a few teams seem to be the LingPipe [23] framework, and protein taggers—specifically the GeniaTagger [24] and ABNER [25].

### 5.4 Assessing Support for the Human Analyst

It is of special interest to investigate how the systems might support a range of human users in their work. Some results of this analysis are reported in [22], where we were able to show that each source (automated systems, authors, curators) provided novel annotations that were missed by the others, making it likely that each could assist the others in this task. However, from the above analysis of results, it is clear that the particular metrics chosen for BioCreative II.5 give only limited insight into the potential of the automated systems to help the user.

This raises two related questions: can we define one or more scenarios where an automated system could help a class of end users; and if so, what are the appropriate performance metrics? In choosing metrics, we need to keep in mind two distinct audiences: metrics needed by developers to optimize their system; and metrics for potential end users (e.g., authors and curators) to assess the impact of an automated system on their work. These measures do not necessarily have to be the same, but it is important to synchronize them so that the metrics used to optimize system performance can lead to improved performance for the end user(s).

In BioCreative II.5, the use of the AUC based on interpolated precision and recall rewarded high recall on the INT and IPT tasks, at the expense of precision. As noted above, systems had a hard time with several aspects of the problem, leading to overgeneration (low precision) of results. The low precision scores dominated recall scores in the balanced F-measure, resulting in low scores; the highest "raw" F-measure of 0.40 was for a high-precision run (precision of 0.38 and recall of 0.41), but this system provided annotations for only 21 of the 61 articles. Authors had a hard time with different aspects of the problem when preparing their structured digital abstracts. In their feedback, the authors noted that one of the most cumbersome aspects was the need to associate each protein with the correct UniProt identifier. Finding the correct identifier requires an understanding of UniProt, including selecting the right identifier from different proteins with similar names and from orthologous proteins from different species with the same name. Authors reported spending around an hour to generate a structured digital abstract for their article—a fair part of that time was spent looking up the UniProt identifiers. We can make a conservative estimate of that time by assuming that the authors spent 15 minutes of their time on UniProt code lookup—approximately 5 minutes per interacting protein, since authors annotated, on average, three to four proteins per document. In contrast, the automated systems were very fast ($\sim$2 minutes processing time per document), and achieved recall comparable to or better than the recall of the authors—close to 70 percent for the best single system, which provided annotations for all 61 documents (precision 0.006, recall 0.68, balanced F-measure of 0.013).

These results suggest a specific scenario where the user (an author or curator who knows the interacting proteins) could quickly skim a list of candidate proteins to select the correct identifiers. This should be possible, provided that the automated system returns a ranked list of interacting proteins plus contextual information, e.g., the protein name as mentioned in the paper, its UniProt identifier, its "standard" name, species, and symbol, plus a linkage to the place(s) in the paper where it is mentioned. This would provide sufficient information for a curator or author to run through a list of 30 candidates per document in 5 minutes, assuming 10 seconds to accept or reject an entry in the list. This could provide the user with comparable performance in less time: from our estimates above, authors achieved a 66 percent recall in about 15 minutes; the best automated system achieved a slightly better recall of 69 percent at a cutoff of 30 candidates per document, using the filtered and homolog-mapped data (as shown in Table 4). This suggests that an automated system can help the user to create a comparably rich answer set (2/3 of the interacting proteins) in less than half the time (2 minutes to run the system plus 5 minutes to screen 30 answers, compared to 15 minutes to generate). And we can achieve even better scores by combining inputs from multiple systems (an ensemble system—discussed below); the trick is to select the right cutoffs to support the intended use, and to select the metrics that will encourage—in this case—high recall with good ranking (right answers in the top 30).

To support this use case, a false negative is far more costly than a false positive, so we want to weight recall far more heavily than precision. This is because it is relatively cheap (in terms of time) to reject false positives from a list;

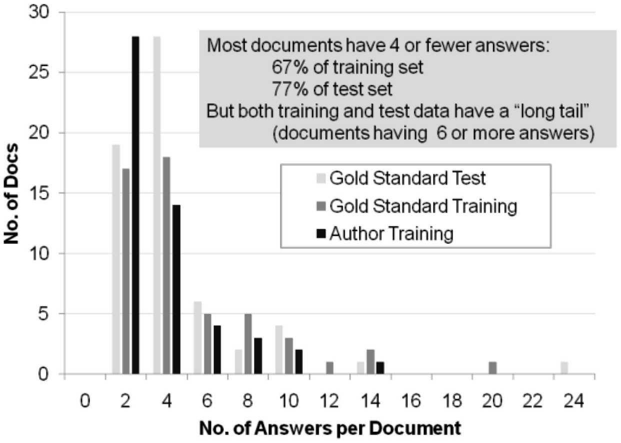


Fig. 7. Interacting proteins per document. The x-axis shows the number of interacting proteins per document; the y-axis shows the number of documents with that number of proteins for both the test collection and the training collection (based on the gold standard). The third set of bars shows the number of annotations provided by the authors on the training set. Most documents (77 percent of test set and 67 percent of training set) have four or fewer answers; however, both sets have a "long tail" of documents with six or more answers.

however, if the correct identifier is not on the list, then it is likely to be missed (a false negative) or it will require significant user effort to look it up (costly in terms of time). The choice of $\beta$ (the weighting of recall) depends on how long the list is (the cutoff) and also on the expected ratio of true positives to false positives in a document.

To estimate this, we first looked at the distribution of true positives per document in the gold standard in both the training and test sets (Fig. 7). We assumed we could run an ideal system with optimal answer ranking, such that all the gold standard true positives were ranked higher than any false positives. In this scenario, both recall and precision depend on the choice of cutoff: for example, if the answer list is cut off at 10, this ideal system would give perfect recall for any document with less than 10 answers—but somewhat lower precision. However, the (few) documents that had more than 10 proteins in the gold standard would have perfect precision, but lowered recall.

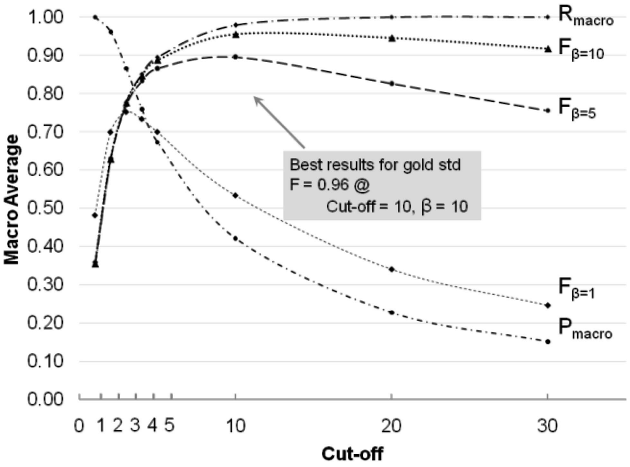


Fig. 8. Training data. Optimal system answers at different cutoffs and betas. Mixed dot-dash lines are the recall and precision at different cutoffs (given on the x-axis); the pure dashed lines represent F-measure for different values of $\beta$. The maximum F-measure (0.96) is achieved for a cutoff of 10 and a $\beta$ of 10.

TABLE 5
Representations Prepared for the Classifier

| j.febslet.2008.01.065<br>P52194 | | j.febslet.2008.01.065<br>P69203 | |
|---|---|---|---|
| class | correct | class | incorrect |
| t22/run4guessed | 1 | t10/RUN_5guessed | 1 |
| t22/run4rank | 0.166667 | t10/RUN_5rank | 0.058824 |
| t22/run4conf | 0.856742 | t10/RUN_5conf | 0.001721 |
| t32/1guessed | 1 | | |
| t32/1rank | 0.166667 | | |
| t32/1conf | 1 | | |

*Two representations are shown: on the left, for a protein that is correctly identified for the given document, and on the right, an incorrectly identified protein. Author data, when used, is treated identically as any system run, although no confidence is available to be used.*

To explore the relation between cutoff, answer distribution, and weighting factor, we varied both the answer cutoff value and the weighting factor $\beta$. These results are shown in Fig. 8. The best results for this ideal system on the training data were achieved at a cutoff of 10 and $\beta = 10$ (a macroaveraged weighted F-measure of 0.96).

### 5.5 Ensemble System—Combine Systems to Get Better Results

So far, we have focused on individual "best system" results. However, in BioCreative II, we demonstrated that a combination of systems (an "ensemble" system) generally outperforms even the best single system. We ran experiments to create an ensemble system for two different scenarios: the first was an ensemble system made up of the best test submission from each group, run at a cutoff of 30 and a $\beta = 10$. In the second experiment, we built an ensemble system by combining results from author curations plus system results; for this, we used the number of answers provided by the author as a cutoff that varied by document.

The ensemble runs were produced using a trained classifier that combines the output of multiple sources 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.

The output of the ensemble is produced by reordering all the candidates for a given document according to the probability of correctness assigned by the classifier. Results are then scored by the BCII.5 evaluation script. Because author annotations were available only for a portion of the training set, it was not possible to train and test on independent data sets as in BCII. Therefore, all the author ensemble results reported here are fivefold cross-validation averages utilizing the 48 available documents. When creating the ensemble (without author annotations) to run on the blind test data, we train on the BCII.5 training corpus and test on the test corpus. This is made possible because all participants were asked to submit the output from their runs on both the test and training data sets.

During training and testing of the ensemble system, the classifier was passed instances such as shown in Table 5. For each participant run (*sysrun*) used as input, the classifier is given an indication whether that run reported the answer anywhere in its result (*sysrun*guessed), the reciprocal of the rank of the answer (*sysrun*rank), and the numerical confidence supplied by the run (*sysrun*conf). Using only the *sysrun*-guessed feature reduces the ensemble to a simple weighted voting scheme. Experiments indicated both rank and confidence contributed positively to the final performance.

Only one run from each team is used in creating a particular ensemble system, so as to limit bias that might accrue from overrepresentation of any one team. Furthermore, the various runs from a given team are less likely to contain statistically independent errors that can be exploited by the classifier to obtain an improved result.

We chose which runs to include by selecting, for each team, the run that maximized a particular metric. In general, choosing runs that maximized recall or AUC tended to produce ensemble runs that did the same; similarly, runs that maximized precision or F-measure tended to produce ensemble runs that were better in those metrics. In this paper, we report ensemble runs that were based on system runs maximizing the microaveraged recall.

The experimental setup we chose has two opportunities to impose a cutoff in the number of answers utilized:

1. a training cutoff—the number of answers per system per document used when training the classifier and
2. a candidate selection cutoff—the number of answers per system per document gathered as candidates for reranking by the ensemble classifier.

Through experimentation, we found that performance was degraded when we trained on fewer answers than we used during candidate selection. Furthermore, training with many more answers than we used during selection provided limited improvements, and in some experiments, caused a drop in performance. Consequently, most runs are conducted with the training cutoff equal to the selection cutoff.

The selection cutoff, because it bears on the size of the candidate pool for the ensemble, has a definite impact on the ensemble system's recall. Choosing an optimal value for the selection cutoff depends on the evaluation criteria, in particular, on how many answers per document will be evaluated. When we evaluated performance on the top N results only, setting the selection cutoff near to N produced best results.

We experimented with two ensemble systems. The first was trained using the runs provided on the training data, and tested against the test data gold standard. We computed the F-measure at $\beta = 10$ at a cutoff of 30. The resulting system (using the "raw" data, with training and candidate selection cutoffs of 5) achieved a recall of 0.67, a precision of 0.10, and a weighted F-measure of 0.62, evaluated on all 61 test documents. This ensemble system did better than the high-scoring comparable single run—for example, at a cutoff of 30, team 10 R5 had a recall of 0.60 and a precision of 0.08.

We also created an ensemble system that combined the author's list of interacting proteins with the lists generated by the automated systems (using the highest scoring microaveraged AUC run from each group). The ensemble used the top 10 candidates per document from each system to train the classifier, and also 10 candidates per system per document to create the ranked list for evaluation. In order to provide results that were comparable to those produced by the authors, the cutoff was chosen to be the number of answers provided by the author for each document. This produced a hybrid author/system ensemble that achieved a macroaveraged balanced F-measure of 0.75 (recall of 0.83, precision 0.73)—distinctly better than the authors alone: 0.71 F-measure, with 0.66 recall and 0.84 precision.

## 6 Conclusions

We have carried out a successful evaluation of automated systems creating protein-protein interaction annotations. With 15 teams participating and 134 results sets in total, the size of this challenge was definitely large enough to draw a number of interesting conclusions. In addition, the challenge used a corpus of 1,190 articles for which Elsevier, the publisher of FEBS Letters, granted us permission for continued distribution as the "BioCreative II.5 Elsevier corpus" via the BioCreative Web site. From these 1,190 articles, 122 have high-quality protein normalization and protein interaction pair annotations that can be used in the future for training and evaluating text mining systems. This forms an openly available collection of continuous full text (both as raw XML and processed to UTF-8 format) that is free to use for the scientific community.

The metrics reported on the raw results appear quite low; however, these results do not reflect the potential utility of the automated systems. By applying postprocessing techniques and using more task-centric metrics, we can get a better sense of the potential utility of the results provided by the automated systems and how these results can provide valuable data to increase the performance and throughput of human annotators. It is probable that such postprocessing steps would form part of a real pipeline. Furthermore, some of these improvements could be achieved with very minimal human interaction. We estimate that use of an appropriately configured automated system could speed up by at least a factor of 2 the task of finding UniProt identifiers for interacting proteins, with no loss in overall accuracy.

Once again, we found that an ensemble system provided better results than any single system; we also developed an ensemble system that combined both author annotations and annotations from automated systems. This author-system ensemble improved the author-only results from 0.71 to 0.75 balanced F-measure.

The upcoming BioCreative III evaluation (September 2010) will include a task focused on interactive use of automated systems for the protein normalization task and the selection of appropriate user-centric metrics. This will provide an ideal setting to determine whether we see the expected time savings from the use of automated systems applied to the gene/protein normalization task; it will also provide developers with direct feedback from curators of multiple databases.

## 7 Contributions

Scott A. Mardis was responsible for the running of the ensemble experiments and the overall coordination of the article. Florian Leitner was responsible for the running of the BioCreative tasks, analysis of the results, and the write-up of those sections of the paper. Martin Krallinger was responsible for the background and comparisons to the previous BioCreative evaluations. Gianni Cesareni was responsible for the contributions from the MINT database team and the *FEBS Letters* experiment. Lynette A. Hirschman was responsible for the analysis of alternative metrics to assess utility of the text mining results for authors and curators. Alfonso Valencia and Gianni Cesareni were jointly responsible for the conception of the BioCreative/FEBS Letters experiment (BioCreative II.5).

## ACKNOWLEDGMENTS

The participation of Scott A. Mardis and Lynette A. Hirschman in BioCreative II.5 was funded by the US National Science Foundation (NSF) grant IIS-0844419. The work of Florian Leitner, Martin Krallinger, and Alfonso Valencia 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 of Gianni Cesareni on the MINT database is supported by FEBS, AIRC, and Telethon grants.

## REFERENCES

[1] R.B. Altman, C.M. Bergman, J. Blake, C. Blaschke, A. Cohen, F. Gannon, L. Grivell, U. Hahn, W. Hersh, L. Hirschman, L.J. Jensen, M. Krallinger, B. Mons, S.I. O'Donoghue, M.C. Peitsch, D. Rebholz-Schuhmann, H. Shatkay, and A. Valencia, "Text Mining for Biology—the Way Forward: Opinions from Leading Scientists," *Genome Biology,* vol. 9, suppl. 2, p. S7, 2008.

[2] C. Blaschke, L. Hirschman, A. Yeh, and A. Valencia, "Critical Assessment of Information Extraction Systems in Biology," *Comparative and Functional Genomics,* vol. 4, pp. 674-677, 2003.

[3] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman, and A. Valencia, "Evaluation of Text-Mining Systems for Biology: Overview of the Second BioCreative Community Challenge," *Genome Biology,* vol. 9, suppl. 2, p. S1, 2008.

[4] L. Smith, L.K. Tanabe, R.J. Ando, C.J. Kuo, I.F. Chung, C.N. Hsu, Y.S. Lin, R. Klinger, C.M. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C.A. Struble, R.J. Povinelli, A. Vlachos, W.A. Baumgartner, Jr., L. Hunter, B. Carpenter, R.T. Tsai, H.J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. Mana-Lopez, J. Mata, and W.J. Wilbur, "Overview of BioCreative II Gene Mention Recognition," *Genome Biology,* vol. 9, suppl. 2, p. S2, 2008.

[5] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman, "BioCreAtIvE Task 1A: Gene Mention Finding Evaluation," *BMC Bioinformatics,* vol. 6, suppl. 1, p. S2, 2005.

[6] "The Universal Protein Resource (UniProt) 2009," *Nucleic Acids Research,* vol. 37, pp. D169-D174, Jan. 2009.

[7] L. Hirschman, M. Colosimo, A. Morgan, and A. Yeh, "Overview of BioCreAtIvE Task 1B: Normalized Gene Lists," *BMC Bioinformatics,* vol. 6, suppl. 1, p. S11, 2005.

[8] A.A. Morgan, Z. Lu, X. Wang, A.M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H.H. Liu, R. Torres, M. Krauthammer, W.W. Lau, H. Liu, C.N. Hsu, M. Schuemie, K.B. Cohen, and L. Hirschman, "Overview of BioCreative II Gene Normalization," *Genome Biology,* vol. 9, suppl. 2, p. S3, 2008.

[9] C. Blaschke, E.A. Leon, M. Krallinger, and A. Valencia, "Evaluation of BioCreAtIvE Assessment of Task 2," *BMC Bioinformatics,* vol. 6, suppl. 1, p. S16, 2005.

[10] A. Ceol, A. Chatr-Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni, "MINT, the Molecular Interaction Database: 2009 Update," *Nucleic Acids Research,* vol. 38, pp. D532-D539, Jan. 2010.

[11] A. Chatr-Aryamontri, S. Kerrien, J. Khadake, S. Orchard, A. Ceol, L. Licata, L. Castagnoli, S. Costa, C. Derow, R. Huntley, B. Aranda, C. Leroy, D. Thorneycroft, R. Apweiler, G. Cesareni, and H. Hermjakob, "MINT and IntAct Contribute to the Second BioCreative Challenge: Serving the Text-Mining Community with High Quality Molecular Interaction Data," *Genome Biology,* vol. 9, suppl. 2, p. S5, 2008.

[12] C. Stark, B.J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: A General Repository for Interaction Datasets," *Nucleic Acids Research,* vol. 34, pp. D535-D539, Jan. 2006.

[13] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia, "Overview of the Protein-Protein Interaction Annotation Extraction Task of BioCreative II," *Genome Biology,* vol. 9, suppl. 2, p. S4, 2008.

[14] F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J. Hakenberg, C. Plake, C.J. Kuo, C.N. Hsu, R.T. Tsai, H.C. Hung, W.W. Lau, C.A. Johnson, R. Saetre, K. Yoshida, Y.H. Chen, S. Kim, S.Y. Shin, B.T. Zhang, W.A. Baumgartner, Jr., L. Hunter, B. Haddow, M. Matthews, X. Wang, P. Ruch, F. Ehrler, A. Ozgur, G. Erkan, D.R. Radev, M. Krauthammer, T. Luong, R. Hoffmann, C. Sander, and A. Valencia, "Introducing Meta-Services for Biomedical Information Extraction," *Genome Biology,* vol. 9, suppl. 2, p. S6, 2008.

[15] A. Ceol, A. Chatr-Aryamontri, L. Licata, and G. Cesareni, "Linking Entries in Protein Interaction Database to Structured Text: The FEBS Letters Experiment," *FEBS Letters,* vol. 582, pp. 1171-1177, Apr. 2008.

[16] A. Chatr-Aryamontri, A. Ceol, L.M. Palazzi, G. Nardelli, M.V. Schneider, L. Castagnoli, and G. Cesareni, "MINT: The Molecular INTeraction Database," *Nucleic Acids Research,* vol. 35, pp. D572-D574, Jan. 2007.

[17] S. Orchard, L. Salwinski, S. Kerrien, L. Montecchi-Palazzi, M. Oesterheld, V. Stumpflen, A. Ceol, A. Chatr-Aryamontri, J. Armstrong, P. Woollard, J.J. Salama, S. Moore, J. Wojcik, G.D. Bader, M. Vidal, M.E. Cusick, M. Gerstein, A.C. Gavin, G. Superti-Furga, J. Greenblatt, J. Bader, P. Uetz, M. Tyers, P. Legrain, S. Fields, N. Mulder, M. Gilson, M. Niepmann, L. Burgoon, J. De Las Rivas, C. Prieto, V.M. Perreau, C. Hogue, H.W. Mewes, R. Apweiler, I. Xenarios, D. Eisenberg, G. Cesareni, and H. Hermjakob, "The Minimum Information Required for Reporting a Molecular Interaction Experiment (MIMIx)," *Nature Biotechnology,* vol. 25, pp. 894-898, Aug. 2007.

[18] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D.P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White, and S.Y. Rhee, "Big Data: The Future of Biocuration," *Nature,* vol. 455, pp. 47-50, Sept. 2008.

[19] A. Bairoch, B. Boeckmann, S. Ferro, and E. Gasteiger, "Swiss-Prot: Juggling between Evolution and Stability," *Briefings in Bioinformatics,* vol. 5, pp. 39-55, Mar. 2004.

[20] C.D. Manning, D.R. Prabhakar, and S. Hinrich, *Introduction to Information Retrieval.* Cambridge Univ. Press, 2008.

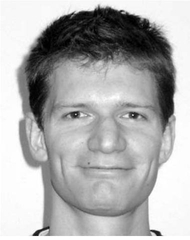
[21] B.W. Matthews, "Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme," *Biochimica et Biophysics Acta,* vol. 405, pp. 442-451, Oct. 1975.

[22] F. Leitner, A. Chatr-Aryamontri, A. Ceol, M. Krallinger, L. Licata, S. Mardis, L. Hirschman, G. Cesareni, and A. Valencia, "Enriching Publications with Structured Digital Abstracts: The Human-Machine Experiment," accepted for publication in *Nature Biotechnology,* 2010.

[23] B. Carpenter, "LingPipe," http://www.alias-i.com/, 2010.

[24] Y. Tsuruoka, Y. Tateishi, J.D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, "GENIA Tagger: Developing a Robust Part-of-Speech Tagger for Biomedical Text," *Proc. 10th Panhellenic Conf. Informatics,* pp. 382-392, 2005.

[25] B. Settles, "ABNER: An Open Source Tool for Automatically Tagging Genes, Proteins and Other Entity Names in Text," *Bioinformatics,* vol. 21, pp. 3191-3192, July 2005.



**Florian Leitner** completed the master's degree in molecular biology, specializing in computational biology, with Frank Eisenhaber at the IMP in Vienna, Austria. Currently, he is working toward the PhD degree at the Spanish National Cancer Research (CNIO) in Spain with Alfonso Valencia, specializing in text mining. He previously worked for Rebecca Wade at the EML in Heidelberg, Germany, and for Markus Jaritz at the Novartis Research Campus in Vienna.

**Scott A. Mardis** received the BS degree in computer engineering and the MS degree in electrical engineering from the University of Illinois at Urbana-Champaign and the MS and PhD degrees in computer science from Cornell University. He is a lead research scientist in the Information Technology Center of MITRE Corporation in Bedford, Massachusetts. From 1988 to 1993, he worked at Bell Laboratories in Naperville, Illinois, on telecommunications hardware and software. Since 2000, he has worked at MITRE in the areas of human language understanding, bioinformatics, and software analysis.

**Martin Krallinger** is working toward the PhD degree under Alfonso Valencia specializing in biomedical text mining. He is currently with the Structural Biology and Biocomputing Group of the Spanish National Cancer Research Center (CNIO). He has served as a co-organizer of the Second BioCreAtIvE Challenge Workshop and also for the Workshop on Text Mining for the BioCuration Workflow at the Third International Biocuration Conference.

**Gianni Cesareni** received a degree in physics from the University of Rome La Sapienza. Thereafter, he spent three years in Cambridge in the laboratory of Sidney Brenner. He then moved to EMBL in Heidelberg, Germany, where he led a group working on the mechanisms controlling plasmid DNA replication. He is currently a professor of genetics at the University of Rome Tor Vergata, Italy. Since 1989, he has taught and worked in Rome. He is interested in the interplay between specificity and promiscuity in the protein interaction network mediated by protein recognition modules. He is the founder of the MINT protein interaction database.

**Lynette A. Hirschman** received the BA degree in chemistry from Oberlin College, the MA degree in German literature from the University of California, Santa Barbara, and the PhD degree in mathematical linguistics from the University of Pennsylvania in 1972. She is director of biomedical informatics in the Information Technology Center at MITRE Corporation in Bedford, Massachusetts. She is a founding organizer of BioCreative (Critical Assessment of Information Extraction for Biology) and of the BioLINK text mining SIG at ISMB. She is also on the board of the Genomic Standards Consortium, working on metadata capture for metagenomics.

**Alfonso Valencia** received formal training in population genetics and biophysics from the Universidad Complutense de Madrid and was awarded the PhD degree in 1988 from the Universidad Autónoma de Madrid. From 1989 to 1994, he was a postdoctoral fellow at the laboratory of C. Sander at the European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. He is director of the program in structural biology and biocomputation at CNIO (the Spanish National Cancer Research Center). In 1994, he set up the Protein Design Group at the Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC) in Madrid, where he was appointed as research professor in 2005. He is a member of the European Molecular Biology Organization, and founder and former vice president of the International Society for Computational Biology. He is a founding organizer of the annual European Computational Biology Conferences and serves on the Scientific Advisory Board of the Swiss Institute for Bioinformatics, Biozentrum, Basel, as well as the steering committee of the European Science Foundation Programme on Functional Genomics (2006-2011). He is a co-organizer of the BioCreative challenges, co-executive editor of *Bionformatics*, serves on the editorial board of the *EMBO Journal* and the *EMBO Reports*, and is director of the Spanish National Bioinformatics Institute (INB), a platform of Genoma España.