Efficient extraction of protein-protein interactions from full-text articles

Jörg Hakenberg, Robert Leaman, Nguyen Ha Vo, Siddhartha Jonnalagadda, Ryan Sullivan, Christopher Miller, Luis Tari, Chitta Baral, Graciela Gonzalez

Abstract—Proteins and their interactions govern virtually all cellular processes, such as regulation, signaling, metabolism, and structure. Most experimental findings pertaining to such interactions are discussed in research articles, which in turn get curated by protein interaction databases. Authors, editors, and publishers benefit from efforts to alleviate the tasks of searching for relevant articles, evidence for physical interactions, and proper identifiers for each protein involved. The BioCreative II.5 community challenge addressed these tasks in a competition-style assessment, to evaluate and compare different methodologies, to make aware of the increasing accuracy of automated methods, and to guide future implementations.

In this paper, we present our approaches for protein named entity recognition including normalization, and for extraction of protein-protein interactions from full text. Our overall goal is to identify efficient individual components, and we compare various compositions to handle a single full-text article in between ten seconds and two minutes. We propose strategies to transfer document-level annotations to the sentence-level, which allows for the creation of a more fine-grained training corpus; we use this corpus to automatically derive around 5000 patterns. We rank sentences by relevance to the task of finding novel interactions with physical evidence, using a sentence classifier built from this training corpus. Heuristics for paraphrasing sentences help to further remove unnecessary information that might interfere with patterns, such as additional adjectives, clauses, or bracketed expressions.

In BioCreative II.5, we achieved an f-score of 22% for finding protein interactions, and 43% for mapping proteins to UniProt IDs; disregarding species, f-scores are 30% and 55%, respectively. On average, our best-performing setup required around two minutes per full text. All data and pattern sets as well as Java classes that extend third-party software are available as supplementary information.

1 Introduction

Proteins and their interactions govern virtually all cellular processes, such as regulation, signaling, metabolism, and structure. Maps of protein-protein interactions (PPIs) are crucial to understanding biological processes on a systems level. A variety of factors influence the formation of protein complexes [1]. Laboratory techniques to detect, analyze, and quantify protein interactions are reviewed in [2] (biochemical approaches) and [3] (molecular and cellular approaches). Results from high-throughput methods such as yeast two-hybrid assays (Y2H) and affinity purification/mass spectroscopy (AP/MS) are the most abundant found in PPI databases. For instance, 64% of all PPIs in IntAct [4] currently stem from two-hybrid screens[1]. Deane et al. [5] assessed the accuracy of high-throughput methods and estimated an error rate of around 50% for *Saccharomyces cerevisiae* data (about 8000 distinct interactions, 6000 derived from four independent high-throughput Y2H screens, all from DIP [6]). Other studies suggest that Y2H performs better in terms of false positive rates than AP/MS; predictions can be validated further with protein complementation assays (PCAs) [7].

- *J. Hakenberg, N. Vo, and C. Baral are with the Department Computer Science and Engineering, Arizona State University, Tempe, AZ, 85283, USA.*
- *R. Leaman, S. Jonnalagadda, R. Sullivan, C. Miller, and G. Gonzalez are with the Department of Biomedical Informatics, Arizona State University, Phoenix, AZ, 85004, USA.*
- *L. Tari is with Hoffmann-La Roche Inc., Nutley, NJ, 07110, USA.*

Databases follow different ways of curating protein-protein interactions. The first is direct submission by authors, after a publication has been accepted (but not necessarily been published yet), sometimes required per the respective journal's policy. The second is journal "shadowing", in which database curators follow new issues of a fixed set of particular journals and curate reported interactions; some major databases such as IntAct and MINT cooperate such to minimize the overlap in curated articles and thus workload. The third way of curating protein interactions is a "topical curation", in which database maintainers pick a subject (such as a Gene Ontology term) of interest to collaborators and try to find and curate each article published referring to this subject (such as an experiment that involves a protein assigned to the GO term). This last way in particular deals with the large amount of legacy data, decades of publications on PPIs, often retrieved using the PubMed citation index. In an effort to quantify these legacy data, [8] found around 150,000 distinct protein-protein interactions of human genes/proteins within one million PubMed abstracts from 1990 to 2007.

Database curation of PPIs comes at a high cost, as curators are either Ph.D. level scientists specifically hired for this task, or researchers working closely with the database and providing topical or in-house data. Thus, any of these efforts would largely benefit from automated systems that handle one or multiple of the following steps with sufficient accuracy:

1) identify abstracts/full text articles that contain relevant data,
2) spot relevant passages in a given article (as opposed to established background information),
3) recognize mentions of relevant biomedical entities (proteins, organisms),
4) map each entity mention to a database identifier (such as a UniProt ID for each protein),
5) extract relationships between entities (such as mapping a protein to an organism or finding protein-protein interactions), and also
6) extract additional information on the experiments described (interaction detection method, clone library, antibodies, etc.).

To alleviate the task of curation and provide quick overviews of experimental findings, some journals started to offer Structured Digital Abstracts (SDAs), which in addition to an author-provided abstract summarize found interactions and map proteins to UniProt IDs etc. Currently, ways are sought to alleviate the task of creating SDAs for either publishers or authors. One of the most time-consuming steps for authors in writing SDAs is to find the correct UniProt identifier for the proteins used in the experiments. While authors know about the reported protein-protein interactions, articles have to be searched meticulously if SDAs are created on the publisher's side. Thus, authors and curators both need tools to simplify writing concise SDAs; tools are also needed to deal with the vast amount of legacy data that was not published together with SDAs.

In recent years, the field of biomedical text mining has seen a number of community challenges to address certain aspects of the ongoing research. Most of these are held in a manner similar to the "Critical assessment of methods of protein structure prediction" (CASP; see [9] for the latest edition). Among the earliest, the "Text Retrieval Conference" (TREC) included a specialized genomics track each year from 2003 to 2007 [10], focusing on information retrieval (IR). BioNLP/JNLPBA addressed the problem of named entity recognition for five types of entities [11]. The BioNLP Shared Task'09 dealt with the extraction of various kinds of molecular events involving genes and proteins (such as regulation, binding, protein catabolism) [12]. The BioCreative community challenge, so far held in 2003, 2006, and 2009, focused on recognition of gene/protein names in text; gene mention normalization, mapping gene names to EntrezGene IDs; extracting protein-protein interactions, mapping proteins to UniProt identifiers; as well as relevance ranking of articles; see the respective overviews in [13], [14], [15]. BioCreative I in 2003 also addressed functional annotation of proteins with Gene Ontology terms, based on textual evidence. Overall, the extraction of relationships among biomedical entities such as genes, proteins, diseases, and mutations from scientific text has been an active research topic for most of the past decade. Most attentions clearly has been paid to the recognition of genes and the extraction of protein-protein interactions from PubMed abstracts; more and more approaches tackle associations between genes and diseases, genes and mutations, drugs and enzymes, proteins and Gene Ontology annotations, and so on, as well as they are shifting to the use of full text articles. Currently, about 2.5 million full text articles are available via PubMed, many of them have open access through archives such as NIH's PubMed Central and publishers such as BioMed Central and the Public Library of Science (PLoS).

BioCreative II.5 in 2009 dealt with the task of helping database curators to find and identify protein-protein interactions in full-text articles (IPT task); to map proteins to UniProt identifiers (INT task); and to rank texts according to relevance for curation (ACT task). All these tasks are potentially beneficial also to authors submitting aforementioned Structured Digital Abstracts, which necessitate at least finding appropriate UniProt IDs for all proteins studied. As BioCreative II.5 addressed system performance for tasks related to curation, the definition of a protein-protein interaction went along the lines of "any association between two proteins that is worth curating for a given publication." From the SDAs provided with the training data, we see that this included physical interactions as well as co-localizations; in any case, all PPIs in BioCreative II.5 are binary and undirected relations. The exact definition of an interaction depends on the targeted database (here: MINT; BioCreative II obtained data from IntAct, another IMEx consortium member [16], [4], [17]). The second half of the definition concerns curation; this can be interpreted as "would a database curator consider a publication as a reference for the given interaction?" Some published experiments yield the first indication that two proteins interact; some experiments try to confirm a given interaction, possibly by using a different detection method; on the other hand, many publications would cite interactions found previously, for instance, to provide background on the proteins studied. While in the first two cases, a curator could decide to include the publication as reference for first/further evidence, the last kind of publication would mostly certainly not be considered (for the given interaction; but possibly for another one). Overall, a protein-protein interaction should be reported if the underlying publication provides physical evidence, by providing results (images!) of a pull-down, immunoblot, fluorescence analysis, etc. All these aspects lead to a relevance ranking step of automatically extracted protein interactions, which we will describe in Section 2.4.

For BioCreative II.5, fifteen participating research groups returned 134 submissions in *online* and *offline* scenarios: the online scenario required each team to set up CASP-like annotations servers [9], [18], which would take a previously unseen query document and return predictions within ten minutes per document; the offline scenario allowed for bulk download and annotation of a set of texts, with a one week time constraint for

returning predictions. In this paper, we describe our approaches to extract protein-protein interactions from full-text articles (IPT task) and to map individual proteins to their respective UniProt identifiers (INT). Note that we did not submit results to the aforementioned article relevance ranking task (ACT), although this could be emulated according to how many proteins/interactions our methods are able to extract, together with their respective predicted ranks.

In the following, we start with a detailed explanation of our methods, which include named entity recognition, entity mention normalization, relevance ranking of sentences, sentence simplification, generating patterns for relationship extraction, and sentence-level annotation of the originally document-level training data. We then describe our data sets and evaluation on BioCreative II.5 and external data, both of the whole system and its individual components. We conclude with a discussion of our findings. Non-trivial data used in our experiments, such as pattern sets or extensions of third-party software packages, are provided as supplementary information. For related work, we refer the reader to the other articles published in this special issue, together with the overview paper by Hirschman et al. [15]. Some of the approaches we discuss build on earlier systems described in [19] (gene mention normalization; pattern generation) and [20] (inter-species gene mention normalization). We discuss related work inline with the methods, data sets, and discussion where appropriate.

2 METHODS

Our system consists of a sequence of building blocks, which we will describe in separate sections in the following. The approach for protein-protein interaction extraction is based on linguistic patterns that we extract from training data, after re-annotating it on the sentence level (original annotations were given on the document level). This requires sufficient named entity recognition and normalization (also referred to as identification). We use our patterns together with the OpenDMAP parser [21] for efficient analysis of a large amount of texts. Sentences are analyzed and ranked for relevance, regarding containment of protein-protein interactions with physical evidence and referral to novelties. We also experiment with sentence simplification, that is, heuristics to paraphrase sentences by reducing noun-phrases to head nouns etc. We evaluated each individual method on task-external data; these results can be found in the part on evaluation, Section 3.

As one of the tracks BioCreative II.5 targeted online web-services for processing full text articles, our goal was to find an efficient and still effective composition of modules to solve the various sub-tasks involved. Thus, we backed off from some approaches presented previously, such as GNAT for entity mention normalization and an alignment-based pattern matching algorithm (see [20], [19]). The focus of all building blocks in our systems is on annotation time, and our goal was to provide a service that can handle a full-text article in 10 seconds and still have reasonable accuracy; our most accurate composition/tuning parameters should still be able to analyze a full text in about two minutes.

2.1 Named entity recognition and identification

One important motivation for studying named entity recognition (NER) in biomedical documents is as a building-block step in a larger information extraction pipeline. NER systems in recent years have tended towards primarily employing machine learning techniques, including conditional random fields, due to the consistently high performance these techniques provide when trained on a high-quality corpora such as the BioCreative II gene mention corpus [22]. In contrast, however, systems for identifying proteins (also called normalization and grounding, EMN) have largely focused on dictionary-based techniques; some notable recent systems include GNAT [20] and GeNo [23]. For BioCreative II.5, we have identified several attributes which are useful for supporting the identification of proteins found, including:

1) the ability to associate a *confidence* with each mention found;
2) improving *consistency* via enforcing a one-sense-per-document assumption;
3) generating a list of *candidate proteins* (identifications) to which each mention could refer.

Our protein name recognition and identification approach thus roughly falls into these steps: *1)* machine learning-based NER of protein names, *2)* dictionary-based recognition of species names, *3)* generating candidate IDs for each recognized mention based on dictionary matching, *4)* filtering of IDs by species, *5)* ranking of candidate IDs.

2.1.1 Recognition of protein names

Our method for locating proteins in text centered on the use of the BANNER named entity recognition system. BANNER is based machine learning utilizing conditional random fields, and utilizes a rich feature set widely surveyed from the literature [24]. For more information on conditional random fields, the reader is referred to the technical report by [25]. For BioCreative II.5, we used a model created using the BioCreative II gene mention training data (15,000 sentences, 18,285 gene mentions).

Confidence scores

We extended the implementation of tagging in the BANNER named entity recognition system to provide the n-best labelings for each sentence, that is, the n labelings with the highest probability according to the conditional random fields model. We then normalize the probabilities by dividing each by the sum of all n. The mention confidence is then taken to be the sum of the sentence

probabilities in which that mention appears. We empirically noted that the probabilities output for each labeling become negligible after $n \approx 10$. Consider the following short sentence fragment having two alternative labelings (predicted protein mention underlined):

"splicings of the human serotonin transporter gene"

"splicings of the human serotonin transporter gene"

In these cases, the normalized sentence probabilities compute to 0.53 and 0.47 in our model, respectively.

Consistency

One of the strengths of conditional random field based taggers is the ability to locate mentions based on the context surrounding the mention. This provides the beneficial effect of allowing mentions to be tagged even though they contain unfamiliar vocabulary. Not every mention will appear in an unambiguous context, however, implying that there will be mentions which are correctly found and tagged in one context but will not be correctly tagged in another context within the same document. To handle such inconsistencies, we employed a variation on the one-sense-per-document assumption utilizing confidence values calculated for each mention. We create a set out of all mentions containing the same text, and then find the mention with the highest mention confidence. We then assign that confidence to every mention within the set. Since most proteins are mentioned several times in the text, this has the effect of forcing all mentions with a non-zero likelihood to appear with the same context as the mention with the strongest context.

We note that there are common exceptions to the notion that biomedical text should be consistent under the tentative definition that consistency refers to the degree that identical token sequences receive the same labels within a document. These exceptions include ambiguous names such as abbreviations, which may be used to refer to entities of many different types. Our methodology is still reasonable in light of names which are ambiguous with protein names since they are not likely to occur in contexts strongly indicative of a protein name. A second exception includes embedded names, such as the "HD" in the name "HD gene", which refers to Huntington disease. These were not common in our dataset.

2.1.2 Recognition of species names

Identification of proteins first requires to map each protein to the correct organism. To recognize species names, we created a dictionary from all entries in the NCBI Taxonomy [26]. These included the NCBI Taxonomy's preferred name, common names, and synonyms, resulting in 405,279 different species, with a total of 557,915 unique names. However, we restricted the number of proteins from UniProt and TrEMBL to species associated with proteins in the training data, see Section 2.2.1 on the next page. We extended this dictionary by including cell lines that can be traced to a single origin species. For example, "HeLa" cells imply human, whereas "NIH-3T3" cells imply mouse. Our collection covers 1390 names of cell lines from eight different species; see supplementary data. For some model organisms, we added additional commonly used names that were not contained in the NCBI entries, such as "patients" referring to human. Some generic names (mostly in the rank of genus) had to be mapped to a specific organism. For instance, we map all occurrences of "mice" (genus Mus) to *M. musculus*, "rats" to *R. norvegicus*, "flies" and "Drosophila" to *D. melanogaster*, and "yeast" to *S. cerevisiae*.

It showed that some species names are ambiguous, for instance, with common English words and other biomedical types. Examples include "laser", "beta", "bears", "This", and "cancer". For BioCreative II.5, we performed word sense disambiguation using the heuristic rules: we removed some of these occurrences in any case (such as "bears", "laser", "codon"), and required another synonym or hyponym of the species to appear in the same text for some other occurrences (a mention of "cancer" required any additional reference to crustaceans, for example).

2.2 Entity mention normalization

Each individually recognized protein mention needs to be mapped to a UniProt identifier. Our strategy was to first assign candidate identifiers based on dictionary-matches. We then narrow down each list of IDs per protein mention by species. We have shown previously [19], [20] how genes/proteins can be disambiguated by using context profiles, that is, information available on each protein that can be compared to the current text and measure the overlap in GO terms, disease associations, tissue specificity, chromosomal location of genes, protein length and mass, and so on. For efficiency required in the online tasks in BioCreative II.5, however, we do not perform actual disambiguation, but rank IDs by dictionary match and species only. We describe these two steps in this section.

2.2.1 Assigning candidate identifiers to proteins

Dictionary-based NER approaches create a short list of potential or candidate identifications that can then be disambiguated. Our named entity recognition system, however, only specifies the location and entity type of each mention. We thus extended BANNER to also perform candidate generation by using approximate string comparison techniques to compare the text of the protein mentions found to the text of known protein names in UniProt. The string comparison employed was the Jaccard index over the position–independent set of tokens in the mention and in the protein name, and was implemented using a trie data structure for efficiency. Each string comparison (recognized mention versus dictionary entry) results in a string similarity score, which we can use to rank individual identifiers, in addition to the aforementioned probability assigned to each mention.

We decided to use all of UniProt/SwissProt, but included data from TrEMBL only for species that were associated with more than one protein in the training set gold standard; we found ten such species. The distribution of species differed slightly between training and test set, in that the test set consisted to 7% each of rat and bovine proteins (none in the training set), and the training set contained about 14% E. coli proteins (none in the test set); while distributions for human, mouse, and yeast proteins were roughly the same. Thus, due to the aforementioned reduction of TrEMBL data, we would not find any rat/bovine protein exclusive to TrEBML (but still all rat/bovine SwissProt proteins).

2.2.2 Mapping protein mentions to species

After candidate identifiers have been assigned to each protein mention, we narrow down each list by potential species. We experimented with four different methods:

- scan the text to the left of the protein mention and pick the closest species; if no species was found, scan to the right;
- assign species by occurrence frequency per document; for example, if 70% of the species mentioned in a document refer to human, pick the first best human protein; if a candidate list for a protein mention does not contain human proteins, pick the 2^{nd} most frequent species, and so on;
- assign species by occurrence frequency in title and abstract alone;
- pick the first protein ID (which corresponds to a random choice).

After experiments on the training data, we decided to use the first method (data not shown). If no species was found in the entire text, we picked a species that was valid for at least one candidate protein at random.

2.3 Relationship extraction

In our experiments for protein-protein interaction extraction, we concentrate on approaches that employ linguistic patterns and compare them to co-occurrence methods as a baseline. In order to obtain a suitable set of patterns, we pursue an automatic strategy that generates patterns from a training corpus. We search for groups of similar (partial) sentences describing PPIs and try to express their commonalities using patterns. In this paper, we base such commonalities on the flat structure of a sentence, that is, sequences expressed in terms of word categories, words, and lemmata, as opposed to parse trees etc. as seen in other approaches. Owing to the web-service scenario in BioCreative II.5, we sought for relation extraction approaches that are fast at application time, just like for all other components discussed here. We compared several techniques of generating and applying linguistic patterns in [19]. We found that patterns can be generated from automatically created, sentence-level training data, based on PubMed abstracts and interaction databases such as IntAct. Sentence patterns can also be generated from document-level annotations, as was the case in BioCreative II.5 IPS. While the first strategy yields high-recall patterns that result in good precision for the extraction of any kind of protein-protein interactions, we found that the second strategy yields better results on the BioCreative 2/II.5 data, albeit being very subtask-specific: BioCreative II.5 seeks for novel, physical interactions; it includes (as opposed to some other benchmarks) co-localizations and co-purifications, excludes protein-gene interactions, and so on. This improvement in performance can certainly be attributed to the close ties between training and test data, which is not given in the first strategy that extracts patterns from arbitrary PubMed abstracts.

At application time, we found that the best performing strategy is an alignment of patterns against new text. However, the processing time is very long, in particular when the pattern sets become large; alignment algorithms are quadratic to compute the alignment score, and cubic when the traceback is needed (e.g., to identify exact pairs plus relation-indicating keywords). We introduced some drastic improvements concerning time performance in [27], which included indexing and filtering of patterns to reduce the amount of alignments to be computed, and which led to an only small decrease in recall.

In this paper, we decided on a method that generates task-specific patterns by clustering annotated training data; we then introduce *conceptualizations of tokens*: that is, lists of words referring to similar concepts (here: protein interactions) that are interchangeable without changing the overall meaning of a sentence and also keeping the sentence grammatical. In the remainder of this section, we first explain how we derived sentence-level annotations from the document-level gold standard. These are used to generate linguistic patterns from all positive sentences in the training data. In the last part of this section, we explain how we used the OpenDMAP framework [21] for maintaining and matching patterns.

2.3.1 Generating training data

Our pattern generating techniques require sentence-level annotation of entities and relationships, whereas the BioCreative II.5 IPS data is annotated on a document level. That means that per document, the list of relevant interactions is known in terms of pairs of UniProt IDs, but no positions in the text, evidence sentences or paragraphs, kinds of interactions, and so on, are given. To annotate the full text articles on a sentence level, we apply NER and assignment of candidate IDs as described in the previous sections. Thus, the outcome is a text annotated with protein mentions and a list of candidate identifiers per mention. We narrow down these candidate IDs according to the IDs given in the gold standard per document. We then skim through the list of interacting pairs per document and pick all sentences that contain any one (or more) of the pairs. A further filtering step removes sentences that do not

contain any word from a pre-defined list of keywords referring to protein-protein interactions. This list is based on the one used in [19], and extended with words found in the structured digital abstracts provided with the training data (keywords such as 'co-localization' that are task-specific). For a recent study of the predictive power of individual such predicates we refer the interested reader to [28]. We were able to recover evidences for 93 of the 223 pairs in the training data; see Section 3.3 for details. We provide the automatically annotated training data, consisting of 784 positive out of 14,844 sentences in total, as supplementary information; thus, on average, we have almost eight evidence sentences per pair found.

2.3.2 Generating patterns

To generate patterns, we use all sentences that contain a protein interaction mentioned in the gold standard. We reduce each sentence to the shortest, continuous snippet that contains both proteins from the gold standard pair as well as a keyword that indicates an interaction. This keyword then becomes a placeholder for similar keywords, so that similar sentences/snippets can be found in new text. As an example, consider the snippet "protein A interacts with protein B", where 'interacts' might be replaced with 'co-localizes' or 'binds'. We distinguish between interaction-indicating verbs, nouns, adverbs, and adjectives, further sub-dividing each category by interaction types and/or tense/conjugation/declension. Examples for such *concept groups* are shown in Figure 1 for regulators, regulations, activities (all three are represented in groups of singular nouns), and verbs (here: past tense). Note that these groupings –in particular for verbs– aim at identifying predicates that can be exchanged with each other so that the resulting sentence will still be grammatically correct; the groupings usually are not restricted to a single biological concepts.

We used such groupings not only for interaction-indicating keywords, but also for tokens in domain-independent word categories, for example to replace 'is', 'are', 'was', 'were' with each other. Note that we explicitly do not use part-of-speech tagging, but rather have pre-defined grouped lists. All such snippets are converted into OpenDMAP patterns, see next section, and we provide these patterns as supplementary information; Figure 2 shows some examples.

2.3.3 Pattern matching with OpenDMAP

For curation and matching of linguistic patterns, we used the OpenDMAP framework [21], which was previously used for BioCreative 2 and related tasks [29], [21]. OpenDMAP patterns are backed by an ontology, in the simplest case a concept hierarchy containing genes, proteins, species and other entities on on hand, and tokens arranged by word category (for instance, verbs by tempus) and meaning (such as our interaction-indicating keywords) on the other hand. This ontology is easily maintained using the Protègè editor. Our implementation requires few additional Java classes to run OpenDMAP with our patterns; these can be found in the supplementary information.

2.3.4 Sentence-level and figure caption co-occurrence

As a fast baseline, we also set up a service that skips the pattern matching and simply extracts protein pairs based on sentence-level co-occurrence. Thus, the only filter to remove false positives is relevance ranking, which we describe in the next section. We treat each figure and table caption as a single sentence.

2.4 Relevance ranking of sentences

The particular task of BioCreative II.5 IPT required to extract novel protein-protein interactions for which physical evidence is provided in a given article. We thus sought a module that ranks sentences according to both criteria (novel and physical evidence). We implemented this module as a classifier learned from the aforementioned, pre-processed training data, which has PPI annotations on a sentence- instead of document-level. We used LibSVM as machine learner, trained on the entire training data (9750 sentences).

Only sentences predicted as positive are fed into subsequent modules; the input are sentences annotated during the protein NER step; a threshold may be applied to influence precision/recall. This ranking module also helps reducing the number of sentences that have to be fed to subsequent, possibly time-consuming modules, thus reducing the overall processing time, which is beneficial for an online scenario.

The features we used for ranking characterize each sentence in using the following aspects:

1) section: in which major section does the sentence occur — Abstract, Introduction, Results, ...
2) position: where in a paragraph can a sentence be found — mapped to a value between 0 (⇒ section heading) and 1 (⇒ last sentence)
3) is the sentence a (sub-) section heading?
4) does the sentence occur in a) a figure caption or b) a table caption?
5) does the sentence contain a reference to a) a figure, b) a table, c) a cited paper, or d) supplementary information?
6) bag-of-words, lemmatized; drawn from current and previous sentence (treated as separate features)
7) number of proteins mentioned in the sentence.

The rationales behind some of these features are as follows. We are trying to find sentences that express a novel, major finding of the current publication, relating to a physical protein interaction. In such publications, protein interactions are often mentioned as sub-headings in the Results or Discussion section, and often quite literally so ("A co-localizes with B" as a heading). Major findings are often mentioned at the very beginning (as an 'appetizer') or very end (as a summary) of a section. A reference to another publication points away from a finding being novel; while a reference to a figure or

regulator (noun, singular)	accelerator, acceptor, activator, agent, anchor, catalysator, catalyzator, deconjugator, down-regulator, downregulator, effector, immunoreactant, inhibitor, interactor, ligand, ligase, modifier, promoter, promotor, reactant, receptor, regulator, repressor, stimulator, suppressor, target, up-regulator, upregulator
regulation (noun, singular)	acetylation, activation, degradation, demethylation, dephosphorylation, depletion, destabilization, destruction, disruption, down-regulation, downregulation, elevation, expression, hyperexpression, induction, inhibition, methylation, modification, modulation, mono-ubiquitination, monoubiquitination, multi-ubiquitination, multiubiquitination, mutation, obstruction, over-expression, overexpression, phosphorylation, poly-ubiquitination, polyubiquitination, stabilization, stimulation, transacetylation, transactivation, transcription, up-regulation, upregulation
activity (noun, singular)	abolishment, abrogation, acceleration, accumulation, activity, addition, affection, amplification, augmentation, augmention, conjugation, control, conversion, cross-reactivity, deconjugation, expansion, exposion, immunoreactivity, inactivation, infection, ligation, mediation, participation, precipitation, prevention, production, promotion, reduction, regulation, sequestration, substitution, sumoylation, suppression, transduction, ubiquitination
interaction (verb, past tense)	abolished, abrogated, absorbed, accelerated, accepted, accumulated, acetylated, activated, added, affected, amplified, anchored, antagonized, arrested, assembled, associated, attached, augmented, blocked, bound, broke, catalysed, catalyzed, cleaved, co-eluted, coeluted, co-immunoprecipitated, coimmunoprecipitated, coinfected, complexed, conjugated, contacted, controlled, controled, co-purified, copurified, cross-linked, crosslinked, deacetylated, deconjugated, decreased, degraded, demethylated, depended, dephosphorylated, depleted, derived, destabilized, destructed, detached, dimerized, disassembled, disassociated, discharged, disrupted, down-regulated, downregulated, elevated, encoded, encompassed, enhanced, evoked, exhibited, formed, fused, hastened, immunoblotted, immunoprecipitated, immunoreacted, impaired, inactivated, incited, increased, induced, infected, influenced, inhibited, initiated, injected, interacted, interfered, interplayed, ligated, linked, methylated, modified, modulated, mono-ubiquitinated, multi-ubiquitinated, oxidised, oxidized, participated, phosphorylated, poly-ubiquitinated, polyubiquitinated, potentiated, precipitated, prevented, produced, reacted, recognised, recognized, recruited, reduced, regulated, related, removed, repaired, replaced, repressed, required, responded, restricted, severed, stabilized, stained, stimulated, substituted, sumoylated, suppressed, synthesised, synthesized, targeted, tethered, transacetylated, transactivated, transcribed, transduced, transfered, transformed, treated, ubiquitinated, up-regulated, upregulated

Fig. 1. Examples of four conceptualized groups of tokens used in our experiments. Any such token in a training/test sentence can be replaced with any of the others in the same group. See supplementary information for the entire set.

... association between MAP1S and SOCS3 ...
⇒ {w-interact-attachment} {w-preposition-of} [interactor1] and [interactor2]
... binding activity between TSC-22 and fortilin ...
⇒ {w-interact-activity-noun-s} {w-preposition-of} [interactor1] and [interactor2]
... a novel role for MafG in HIF-1alpha accumulation ...
⇒ [interactor1] {w-in} [interactor2] {w-interact-activity-noun-s}
... localization of dysferlin, a binding partner of affixin, in ...
⇒ [interactor1] , {w-determiner}? {w-interact-attachment} partner {w-preposition-of} [interactor2]
... modification of Prox1 by SUMO-1 ...
⇒ {w-interact-regulation-noun-s} {w-preposition-of} [interactor1] {w-preposition-by} [interactor2]

Fig. 2. Example source sentences and resulting patterns; see supplementary information for the whole set. Terms in square brackets refer to slot fillers, here: interacting proteins. Curly brackets represent concepts used as placeholders for word lists, see Figure 1 for examples. '?' indicates optional terms. Tokens without brackets ('partner') are fixed.

table points to additional information that the authors apparently thought useful to underline their findings. Facts mentioned in the introduction of a paper are most often repetitions of known facts (background information involving the entities discussed in the paper); mentions in abstracts, results, and conclusions are more likely to point to novel findings. Such sentences often contain phrases such as "Here we show that ..." or "We conclude that ...", indicating novel findings.

2.5 Paraphrasing sentences

The idea behind our paraphrasing approach is to simplify sentences to retain only information necessary for relationship extraction. Our strategy follows four steps, explained in detail in [30]:

1) deleting uninformative words,
2) replacing entity names with a single-word tag,
3) replacing noun phrases with the head noun, and
4) simplification of syntax.

All these steps serve the purpose of arriving at a sentence that is simpler to parse for a dependency parser, creating less parsing errors. In the framework presented for BioCreative II.5, simplification serves mostly to increase recall by removing unnecessary "fillers" from both patterns and new sentences.

Deletion of uninformative words — Each sentence is first preprocessed to remove phrases that are not essential to the sentence. This includes removal of section indicators, which are phrases that specify the name of the section at the beginning of the sentence and are followed by a colon. These section indicators typically do not contain a verb. Another type of removal is the removal of phrases in parentheses, which include citations and numbering in sentences that represent lists. Other than removal, the preprocessing step involves a simple transformation of partial hyphenated words, which are words that begin or end with a hyphen. Such words are typically parts of the nearby hyphenated words. A partial hyphenated word is transformed by combining it with the nearest hyphenated word that follows or precedes the partial hyphenated word, depending if the partial hyphenated word begins or ends with a hyphen. For instance, the phrase "alpha- and beta-catenin" is transformed into "alpha-catenin and beta-catenin". In addition, we re-

move introductory phrases that frequently occur at the beginning of a sentence; examples are the underlined parts in "These results suggest that affixin is involved in reorganization of subsarcolemmal cytoskeletal actin [...]" and "As reported previously, alphaPIX was specifically coimmunoprecipitated by [...]", which we remove.

Replacement of entity names — Named entities occur frequently in biomedical text, and due to their inherent complex structure they are one of the main reasons for natural language parsers to perform poorly on biomedical text. Our approach is to replace each entity name (typically noun phrases) with a single element. For the system described in this paper, we focus on named genes, as recognized by BANNER and described in Section 2.1. Each named entity will also be numbered, so we would replace each such name with tokens like 'GENE0', 'GENE1', and so on. To satisfy the linking requirements when using a deep parser (Link Grammar in our case), we also have to consider the grammatical category of each name (that is, singular or plural). To address this issue, single elements are concatenated with an 's' if the following verb is not third-person singular.

Replacement of noun phrases — The occurrences of multiword technical terms involved in biomedical text imply that such terms introduce inaccuracy while calculating the syntactic information available in the sentence; for instance, many parsers would join adjectives with their corresponding nouns [31]. Our approach uses LingPipe [32] for shallow parsing to identify noun phrases and replace them with single elements. As in the replacement of gene names, grammatical category has to be taken into account. A single element is considered singular when the following verb is third-person singular or the determiner preceded by the element is either 'a' or 'an'. The single element is otherwise considered as plural and an 's' is attached to the end of the element.

Replacement of gene names with place holders like 'GENE0' does not generally lead to loss of context for the task of PPI extraction, as we maintain a list of place holders cross-referenced to the corresponding original gene names. However, replacing noun phrases with place holders like 'REPNP0' (for the first such replaced noun phrase) can cause loss of context because of skipping the words that indicate association. Hence, we replace the noun phrases with the head noun of the phrase as identified using the Stanford Parser.

Simplification of syntax — In [30], we discuss the necessity of building a ratioed metric for determining the grammatical correctness of a sentence. Every sentence can be uniquely associated with the two-tuple of null count and disjunct cost obtained from the cost vector of Link Grammar output. The null count (which represents words left out in the Link Grammar parse) needs more attention than the disjunct cost (which represents linkages marked as less likely by Link Grammar). Since null counts and disjunct costs are typically less than ten (that is, single digit numbers), for the purpose of easy comparison and for capturing the two-tuples in one dimension, we define a GRAM value of a sentence to be ten times the null count plus its disjunct cost. It is an easy proof that a GRAM value is equivalent to the two-tuple of null count and disjunct cost, under the assumption that the disjunct cost of the corresponding collection of sentences is not more than ten. Any syntactic simplification will be approved only if the resulting sentences are collectively at least as grammatically correct as the original sentence alone, that is, the sum of GRAM values of the parts should be less than or equal to the GRAM value of the original sentence. We implemented rules for prefix subordination, infix subordination and if-then coordination (for details see [33]). These rules were also adapted recently by SimText [34], a text simplification system for improving the readability of medical literature.

2.5.1 Limitations and future work

Since the process of determining grammatical correctness requires processing the sentence and its components with Link Grammar multiple times, the last step is computationally expensive and is not suitable.for an online competition like BC 2.5 that has time constraints. So, we didn't use syntactic simplification in our pipeline of BC 2.5. We also noted that Link Grammar is not as efficient as statistical parsers like [35] and hence we wish to change the GRAM metric to use notions based on normalized probability and transpose the algorithm by replacing Link Grammar dependencies with Stanford dependencies.

2.6 Ranking of extracted pairs and proteins

Ultimately, we rank all pairs and individual proteins per article. This ranking is influenced by the confidence scores we obtained for each protein mention, the disambiguation to find a UniProt ID per protein, the relevance score of the ranked sentence, as well as the number of evidences found for each pair in the overall article. We can also use the later value for filtering, requiring a minimum number of occurrences of each pair throughout the article. We set a threshold for each individual of the aforementioned scores as well as the combined score to filter out likely false positives; the combined score was the product of all individual scores, normalized to [0..1] by dividing by the maximum score observed in the corresponding article.

3 EVALUATION AND RESULTS

We first describe the evaluation results relevant to the BioCreative II.5 INT and IPT tasks. The second half of this section presents results obtained for individual modules, as established in intrinsic evaluations. Submissions in BioCreative II.5 were evaluated in two ways for the interaction normalization (INT) and the interaction pair extraction (IPT) tasks: 1) using the raw submitted data

and 2) using the submitted data mapped to orthologous proteins. For the first evaluation, submissions predict the exact protein in terms of its UniProt ID to achieve a true positive; for the second, prediction of a protein that was homologous to the true protein was also considered correct. The reason for introducing "homonym ortholog mapping and organism filtering" (HOF) was that it is –for automated systems– often impossible to disambiguate species: authors do not always mention the species, or multiple species are plausible for a given protein mention; also, some interactions include proteins from different species. In BioCreative II.5, putative organism(s) were selected per protein by curators, and a predicted protein was mapped to these target organism(s) if an orthologuous protein with the same name existed for the later; see [15] for details. Thus, this second assessment essentially made up for errors in mapping a protein to an organism, alleviating the task of protein normalization.

In addition to precision and recall, two metrics were used in BioCreative II.5 for ranking submissions: F-score and AUC interpolated precision/recall (iP/R) curves. As per the initial task description, submissions were assessed using the macro-averaged F-score for INT and IPT. They were later also ranked according to AUC iP/R. To obtain AUC iP/R curves, the highest precision at each recall point is calculated. The interested reader can find a detailed discussion regarding f-measure versus AUC scores (for PPI extraction systems) in [36].

Our systems for INT and IPT were tuned towards high F-score, in particular, trying to balance precision and recall (this does not hold for all of our co-occurrence-based submissions, which concentrate on high recall). From the overall data (see overview article, [15]), it can be seen that high AUC iP/R can practically be guaranteed by tuning systems towards high recall. For both INT and IPT this can simply be achieved by submitting multiple identifiers per protein, thus leading to hundreds of UniProt IDs per article. The highest AUC iP/R submission of 43.5% contained an average of 83 IDs for each of the 252 relevant proteins (20,888 IDs in total for the 61 relevant documents, 342 IDs on average per document), with a resulting precision of 1.2%. For a database curation scenario, narrowing down the number of IDs per protein/article might prove suitable, for curators as well as authors writing an SDA; thus, we focus on reporting F-scores in the remainder.

For the INT task, mapping proteins that take part in an interaction to UniProt identifiers, we achieve a maximum F-score of 42.9% on the raw data set; see Table 1. The second best f-score was obtained by team 18, achieving 28.6%. Our system yields the highest precision among the approaches presented by all teams (43.4%, +13% over the next best approach), and a maximum recall of 53.5% (best team: 59.1%), both again on the raw data.

In the IPT task, extracting protein interactions from full-text articles, we achieve an F-score of 22.1%, slightly behind the best system (team 18), which achieves 22.2% on the raw data; see Table 2. Our system obviously benefits from the homonym-ortholog mapping (see overview paper [15]), after which we achieve an F-score of 30.1%, outperforming all other approaches.

3.1 Server configurations

All our five submissions were implemented as online servers with different settings, referred to as *s01*, *s02*, *s03*, *s19*, and *s20* in the following. In essence, we varied the methods for interaction extraction (pattern-based; sentence-level co-occurrence; sentence and figure caption co-occurrence), applied or skipped relevance ranking, and applied or skipped sentence simplification; for relevance ranking, we varied the threshold above which an interaction (IPT) or individual protein (INT) was reported. For protein named entity recognition and identification, we varied the thresholds considering mention and identification probabilities based on BANNER and dictionary matching. We also varied the minimum support for a predicted pair, that is, the minimum number of extracted evidences discussing a pair in the given document. For all five servers, we excluded self-interactions.

- *s01* used pattern-based extraction, relevance ranking, and sentence simplification. Thresholds were set to low confidence, but we required three occurrences per interaction in an article before reporting it. Overall, this setup was thought to yield the best F-score performance, with more or less balanced precision and recall, but also requiring the longest processing time per full text.
- *s02* extracted interactions based on sentence-level co-occurrence, requiring high confidence predictions for protein mentions, normalizations, and pairs. This setup employed relevance ranking to filter out likely false positives in terms of "novel interactions with physical evidence". At least hree occurrences of an interaction had to be found, otherwise we did not report it.
- *s03* was set up similar to *s02*, with the addition of figure caption co-occurrences; two proteins had to co-occur within the same figure caption, not sentences of a caption. The minimum number of evidences per interaction was set to one, but we considered only sentences that had exactly two proteins.
- *s19* used essentially the same configuration as *s01*, but without relevance ranking of sentences.
- *s20* was set up as a "high recall" run: interaction pair extraction was based on simple sentence-level and figure caption co-occurrence; the thresholds for protein mentions, protein identification, and relevance ranking were set to low values; minimum support was one.

The high recall run, *s20*, yielded almost 52% recall at 11% precision for the IPT task; the best pattern-based recall was 30% at 38% precision (*s01*). For individual proteins, *s20* achieved 64% recall, compared to 55% for the best pattern-based configuration (again *s01*).

TABLE 1
Interaction pair normalization (INT) raw results (left half) and ortholog-mapped results (right half). The table shows results for our five experiments (grey); for other participants, only their respective best result is mentioned. Each row starts with a team number plus a key ("run4") to further identify the experiment, *cf.* overview paper [15]. o indicates an online run, f offline. Results are macro-averaged, in percent, sorted by F-score in raw results.

Result set	Results	P	R	F1	AUC iP/R	Result set	Results	P	R	F1	AUC iP/R
42 s01 o	100	**43.4**	48.2	**42.9**	38.6	42 s01 o	100	**67.3**	52.2	**55.1**	49.1
18 run4 f	519	23.6	44.1	28.6	25.1	31 s18 o	340	46.7	52.6	44.5	37.5
14 run1 f	199	30.4	29.1	27.7	24.7	37 run3 f	584	44.1	54.7	41.8	44.0
37 run3 f	584	22.1	50.3	25.7	30.8	18 run3 f	519	41.1	51.9	40.9	38.0
42 s19 o	269	19.5	33.3	22.6	22.7	14 run1 f	199	55.6	36.4	40.5	33.3
42 s02 o	703	16.7	47.5	21.9	30.3	42 s19 o	269	47.5	42.5	39.6	37.8
31 s18 o	340	16.2	39.7	21.8	18.1	42 s02 o	703	35.3	52.5	35.5	44.5
42 s03 o	777	14.4	45.1	20.0	25.8	42 s03 o	777	33.7	50.7	34.7	39.6
42 s20 o	1362	10.5	53.5	16.7	33.3	10 s09 o	1784	25.5	**71.8**	34.0	**60.1**
10 s09 o	1784	8.7	**59.1**	14.5	**42.8**	32 1 f	1697	28.3	55.9	32.5	39.3
32 1 f	1697	6.8	44.4	11.3	17.8	42 s20 o	1362	27.4	63.8	32.1	52.6
22 s12 o	4328	3.1	57.9	5.8	29.2	22 run1 f	5133	20.4	69.7	26.6	56.3

TABLE 2
Interaction pair extraction (IPT) raw results (left half) and ortholog-mapped results (right half).

Result set	Results	P	R	F1	AUC iP/R		Results	P	R	F1	AUC iP/R
18 run5 f	612	**29.0**	23.6	**22.2**	17.5		110	30.9	23.6	23.2	19.1
42 s01 o	407	21.3	29.6	22.1	19.4		128	**38.0**	29.6	**30.1**	25.7
37 run7s f	2068	11.5	**34.7**	12.3	**22.2**		504	18.2	39.2	19.1	32.4
14 run4 f	572	12.0	14.5	12.1	13.4	(run1, f)	192	20.9	17.4	16.2	16.5
31 UWMFull f	86	18.0	10.8	11.6	9.6		49	37.2	21.0	23.6	19.4
32 s07 o	219	12.3	10.1	10.3	9.1		63	23.6	18.1	19.6	17.5
42 s20 o	27106	2.7	32.5	4.6	7.9		3552	11.1	51.5	14.6	26.0
42 s19 o	1306	2.7	9.8	3.7	3.3		432	13.5	16.6	11.9	10.9
42 s02 o	10488	2.2	22.6	3.1	6.1		1915	14.8	39.6	16.1	27.0
42 s03 o	10082	1.9	17.3	3.0	5.3		1453	14.5	32.5	15.9	23.6
51 precision2 f	25692	1.8	12.6	2.7	3.4		3233	7.5	19.6	8.4	9.1
22 s12 o	131243	0.4	33.3	0.8	6.9	(run1, f)	36317	10.6	**64.2**	14.3	**35.2**

TABLE 3
Composition and performance of individual servers. Average processing time in minutes per article, based on the full set of 595 test documents, 61 of which contained interactions; F1-score in % for raw and ortholog-mapped data; for detailed results see Tables 1 and 2. Note the large processing time in the co-occurence based configuration s02, which was due to a server outage and should be similar to s03.

ID	Setup	Time	F, IPT, raw	orth.	F, INT, raw	orth.
s01	pattern–based	2m 11s	22.1	30.1	42.9	55.1
s19	pattern–based, w/o relevance ranking	45s	3.7	11.9	22.6	39.6
s02	sentence–level co-occurrence	2m 12s	3.1	16.1	21.9	35.5
s03	sentence & figure caption co-occurrence	1m 16s	3.0	15.9	20.0	34.7
s20	sentence & figure caption co-occurrence, low confidence	11s	4.6	14.6	16.7	32.1

3.2 Pattern set from BC II and BC II.5 training data

We identified 1409 snippets in the BC II.5 training data, leading to 971 unique patterns; 200 of these patterns occurred twice or more often in the training snippets. From the BioCreative II training data, consisting of 700 full text articles, we extracted a total of 18,206 snippets, yielding 11,062 unique patterns; 4608 of these patterns occurred two or more times. We merged the two sets of patterns with support two or more (200+4608) into the final set we used for our experiments; there were surprisingly few redundancies (34; only 17% out of the 200), so the overall number of unique patterns used was 4774. The supplementary information contains each pattern together with its support, including the ones that have a single supporting evidence; note again that we did not use those in our experiments, they are provided for completeness only. Table 4 lists all patterns generated from the re-annotated BC II.5 training set that have a support of four or more in this set (see supplementary data for whole set). These 24 patterns cover 15.3% of the training examples. The patterns with a support of two or more together cover 50.3% of the training data.

3.3 Where to find interactions

Annotation of the training data on a sentence level led to an analysis as to where protein interactions are typically discussed in full text articles. The basis for this analyses were 3,794 sentences from the training set that contain two or more proteins (but not necessarily interactions). We were able to recover 93 out of the 223 protein pairs

TABLE 4
Patterns with highest support (≥4) in the training data. See Section 2.3.2 for explanations of the markup. *DASH* and *COMMA* are placeholders for a single dash/hyphen and a comma, respectively.

Pattern	Support
{w-interact-attachment} {w-prep-of} [interactor1] and [interactor2]	75
[interactor1] DASH [interactor2] {w-interact-attachment}	25
[interactor1] {w-interact-verb-s} {w-prep-with} [interactor2]	21
[interactor1] {w-interact-verb-s} [interactor2]	20
[interactor1] {w-is} {w-interact-verb-d} {w-prep-for} [interactor2]	18
{w-interact-attachment} {w-prep-of} [interactor1] {w-prep-with} [interactor2]	14
[interactor1] {w-is} {w-interact-verb-d} {w-prep-with} [interactor2]	14
[interactor1] {w-interact-verb-d} {w-prep-with} [interactor2]	12
[interactor1] and [interactor2] {w-interact-verb-i}	10
[interactor1] {w-interact-attachment} {w-prep-with} [interactor2]	8
{w-interact-attachment} {w-prep-of} [interactor1] and {w-determiner}? [interactor2]	8
[interactor1] {w-interact-verb-s} [interactor2] {w-interact-adjective-dashed}	6
{w-interact-verb-i} {w-prep-of} [interactor1] and [interactor2]	6
[interactor1] deficient {c-cell} COMMA {w-interact-regulation-noun-s} {w-prep-of} [interactor2]	6
{w-interact-verb-d} [interactor1] DASH [interactor2]	6
[interactor1] as {w-determiner}? {w-adjective}? [interactor2] {w-interact-attachment}	5
[interactor1] {w-interact-verb-s} {w-prep-with} {w-determiner}? [interactor2]	5
[interactor1] {w-in} [interactor2] {w-interact-adjective-dashed}	5
[interactor1] {w-aux} {w-interact-verb-i} [interactor2]	4
[interactor1] {w-is} {w-interact-verb-d} {w-prep-for} [interactor2] {w-interact-adjective-dashed}	4
[interactor1] {w-in} [interactor2] {w-interact-activity-noun-s}	4
[interactor1] and [interactor2] {w-interact-verb-s}	4
[interactor1] {w-interact-verb-s} {w-prep-with} {w-determiner}? {w-adjective}? domain {w-prep-of} [interactor2]	4
[interactor1] mediates {w-interact-attachment} {w-prep-with} [interactor2]	4

in the training set. This means that per each of these pairs, we found at least one evidence sentence that contains both proteins. Note that these data refer to auto–annotated sentences, as described in Section 2.3.1; we skipped the disambiguation step by reducing the sets of candidate UniProt identifiers to the ones known in the gold standard for a particular document. Missing protein pairs resulted from false negatives in the NER step, missed assignment of candidate IDs, proteins of a pair never occurring together in a single sentence, and pairs discussed in tables, figures, or supplementary data (where we analyzed captions only).

These 93 interaction pairs cover 120 out of the 239 different proteins in the gold standard. Counting individual proteins, without the constraint that they have to be in the same sentence as an interaction partner from the gold standard, we found 134 of the 239 proteins somewhere in the corresponding document. 28 of the pairs not found contained an identifier that was not from UniProt/SwissProt, concerning 41 individual proteins. 27 (>10%) of the training pairs occurred in a single table (document ID: `2008.02.82`), but were never mentioned in the full text (tables are not contained in the training data except for captions), and thus eluded our method. Other pairs were mentioned only in figures (see training document `2008.01.65`, Figure 5, for an example) and never in the full text; in some cases, even the figure caption did not mention one or both proteins. The remaining pairs were missed by NER or candidate assignment.

The BioCreative II.5 IPS training data consists of 61 full text documents, with 9750 sentences in total. Our automated sentence-level annotation of the training data (see Section 2.3.1) revealed that 2291 sentences contain exactly one protein, and 3794 sentences contain two or more proteins (23 and 39%, respectively). Among the 3794 sentences with at least two proteins, we found 784 sentences that contain one or more protein pairs known to interact from the gold standard. We provide these sentences as supplementary information, with markup for proteins (including UniProt ID), pairs, and source of the evidence regarding document and position with the document. As mentioned above, we could not find all gold standard pairs in single sentences of the training data. Some pairs would be mentioned only in tables (which are not part of the training data except for table captions), only in figures (such as in a blot), the evidence is spread across multiple sentences (maybe using anaphora), our NER step missed one or both protein mentions, or our EMN step failed to assign the correct UniProt ID(s).

We analyzed the 784 sentences regarding their position within each document to get an idea of where interactions can typically be found in a full text publication (also cmp. Ding *et al.* [37]). Figure 3 shows these data, distinguishing between the two aspects "section" (*Title*, *Abstract*, *Introduction*, *Methods*, *Results*, *Results and Discussion*, *Discussion*, *Conclusions*, *Figure*, *Table*, and three types of *Supplementary data*) and position within a subsection/paragraph (heading, first sentence of a paragraph, sentence in the middle, and last sentence). Note that data from figures and tables stem from their captions only, not the actual figure or table. Note also that *Results and Discussion* refers to a frequently occurring main section heading, and not data joined from two different sections. Values used in Figure 3 include duplicates, that is, sentences that contain more than one interacting

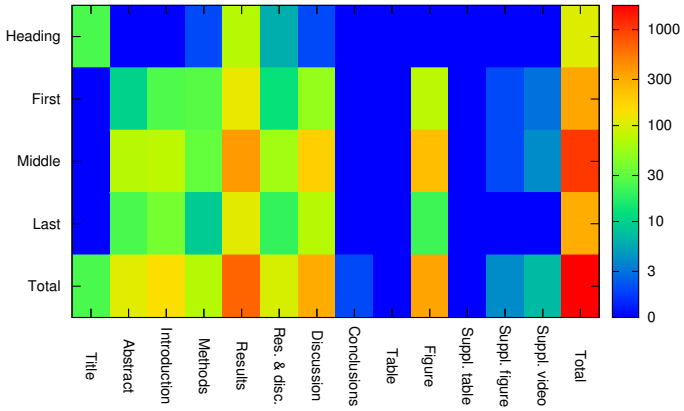

Fig. 3. Positions within full text articles where interacting proteins are discussed. Horizontal axis: number of sentences per main section (Title, Introduction, Methods, etc.); vertical axis: number of sentences grouped by position within a paragraph (heading of a section or subsection, first sentence of a paragraph, sentence in the middle, and last sentence). Figures and Tables are not distinguished by the section they occur in. The last column and row indicate totals per section and position, respectively.

pair are counted multiple times. Thus, the total number of evidences we obtained from the 784 sentences was 1784. 38% (682 out of 1784) interaction pairs occurred in *Results* sections, 19% in *Figure* captions, and 17% (308) in *Discussion* sections. *Abstracts* contained 6% of the sentences that discuss interactions. Grouping sentences by their position within a subsection/paragraph, it showed that 45% of all headings that contain two or more proteins discuss one or more interactions; for first, middle, and last sentences the percentages are 28, 28, and 32%, respectively.

On average, each full text article in the training set contained 3.65 interactions, possibly with multiple mentions anywhere in the text. Looking into the coverage of the interaction pairs by each section, we found that 77% of the interaction pairs were mentioned (once or multiple times) in *Results*, 57% were mentioned in *Introductions* and *Figures*, 56% in *Discussions*, 50% in *Abstracts*, and 27% in *Titles*. *Results and Discussions* contained 5%, and *Methods* only 2%. Though we were able to recover evidences for only 42% of the gold standard interactions from the training set, these numbers can be seen as rough estimates for occurrences of interactions in full text articles; this holds in particular because a large number of the missing pairs might be discussed in tables and figures only and thus not be accessible by our methods (see beginning of this section).

TABLE 5
Evaluation of sentence simplification on AIMed. *"And"* means we count a PPI only if it was predicted by both systems, with and without simplification; for *"or"* combination, the PPI was predicted by at least one. P/R/F1 in %.

Experiment	Precision	Recall	F-score
without simplification	49	53	51
with simplification	**53**	55	53
with *and* without	52	46	50
with *or* without (incl.)	48	**61**	**54**

3.4 INT- and IPT-independent analysis of modules

3.4.1 Sentence simplification

We used PIE [38], a machine learning PPI extraction tool that uses parse trees generated from Stanford parses, for evaluation on the AIMed corpus [39]. The results are shown in Table 5. The results of evaluation and error analysis allow us to conclude that sentence simplification, although still needing improvements, leads to improved PPI extraction results using PIE. This indicates sentence simplification used as a preprocessing step for NLP-based systems could improve PPI extraction. In particular, we observed that simplification can largely increase the recall, while maintaining or also increasing precision, depending on the setup. We tested the combination of predictions from sentence before simplification with predictions extracted after simplification, using AND or OR semantics, as detailed in Table 5. Not surprisingly, the highest recall thus results from an OR combination, but highest precision was obtained after simplification.

3.4.2 Protein named entity recognition

To evaluate the performance of our NER component, we previously experimented with BANNER on the BioCreative 2 Gene Mention (GM) data set. In its latest version, we obtained an f-score of 86.4% (88.7% precision at 84.3% recall) on the BC2 GM test set; the model was computed from the entire BC 2 GM training set. BC2 GM consists of 15,000/5,000 sentences from PubMed abstracts (fixed split for training/testing), with 18,285/6,331 mentions of genes and gene products.

3.4.3 Protein mention normalization

To normalize proteins names to UniProt identifiers, we use an adapted version of GNAT [20]. GNAT was evaluated on data derived from the BioCreative 1 and 2 test sets, thus mapping genes to their respective EntrezGene entry. The derived data consists of 100 annotated abstracts and contains 320 genes from 13 species. Due to the origin of the data set, 295 genes refer to either mouse, human, yeast, or fruit fly. In a cross-species evaluation, GNAT achieves an f-score of 81.4%; for human, mouse, fly, and yeast, individual f-scores are 85.4, 81.0, 75.3, and 89.6%, respectively.

3.4.4 Sentence ranking by relevance

We performed a five-fold cross-validation of the sentence relevance ranking module on the entire sentence-level training set, independent of actual pairs and their UniProt identifiers; averaged accuracy for this module was 94%.

4 Discussion and Conclusions

We presented here a system for extracting protein-protein interactions from full-text and mapping each protein to a UniProt identifier. We showed how to derive a sentence-level training set from a document-level gold standard to generate patterns, which we use in the OpenDMAP framework to analyze new text. We rank sentences by relevance to containing novel interactions and evidence for physical interactions. Sentence simplification helps to rid of unnecessary information (filler words) in both patterns and new text. Using this approach, we achieve an f-score of 22.1% for finding protein interactions, and 42.9% for mapping proteins to UniProt IDs, as evaluated during the BioCreative II.5 community challenge. It shows that our method outperforms other approaches for the relation extraction task and is on par (-0.1%) with the best system for protein mention normalization. Disregarding the species in the normalization task, thus finding the matching group of orthologuous proteins only, we can outperform all other systems, obtaining an F-score of 30.1%.

As our system largely benefits from the homonym-ortholog mapping (HOF; +8% in F-score). The jump results from an increase in precision (+16.3%), while recall remains the same (29.6%), due to the nature of the mapping. This leads to the conclusion that while our system often can identify the group of orthologuous proteins correctly, it lacks in mapping proteins to the correct species, as shown by the large gain in precision.

In future work, we also aim to identify "negative patterns" to filter false positives. In the same manner, we experiment with ranking patterns and/or sorting them by largest spans covered whenever multiple patterns match the same sentence. As an example, consider the partial sentence "protein A binds to B and C to D"; a pattern "P1 binds P2 and P3", which can easily be found in all training sets, would lead to the wrong conclusion, namely A binding to C, while in reality C binds to D.

Another interesting perspective would be to encode parse information (shallow parse, constituents, and/or dependencies) in our patterns; see, for instance, [40], who assess the usage of several deep parsers in extracting PPIs. One drawback with respect to the online scenario we are focusing on in this article would be the added overhead in parsing individual sentences (possibly filtered by a prior ranking step). Depending on sentence, parser, and parameters, parsing can take between 10 and 30 seconds to obtain useful parse trees, by far exceeding the time constraints we envision for our system.

Notes

[1] See http://www.ebi.ac.uk/intact/statisticView, Oct 2009.

Appendix

Supplementary Information

Annotated training data

BC II.5 training data annotated for individual interactions on the sentence level.

OpenDMAP patterns

Auto-generated patterns for protein interactions, based on BC II.5 training data, to be used with OpenDMAP [21].

Additional classes for OpenDMAP

Class files that demonstrate the extension of OpenDMAP [21] to use our pattern files.

List of cell lines mapped to species

In addition to species' names from the NCBI Taxonomy, we compiled a list of 1390 names of cells and cell lines that can each be mapped to an individual origin species.

Acknowledgments

GG, RL, CM, and RS acknowledge support from Science Foundation Arizona grant CAA 0277-08, The Arizona Alzheimer's Disease Data Management Core under NIH Grant NIA P30 AG-19610, and the State of Arizona Alzheimer's Disease Research Consortium. Parts of this research (CB, NV, LT, JH) were funded by the grants NSF 0412000, SFAZ CAA 0289-08, and NSF OCI 0950440. JH thanks the Fulton School of Engineering for support.

References

[1] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc Natl Acad Sci*, vol. 93, no. 1, pp. 13–20, January 9 1996.

[2] J. A. Miernyk and J. J. Thelen, "Biochemical approaches for discovering protein-protein interactions," *Plant J*, vol. 53, no. 4, pp. 597–609, Feb 2008.

[3] S. Lalonde, D. W. Ehrhardt, D. Loqué, J. Chen, S. Y. Rhee, and W. B. Frommer, "Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations," *Plant J*, vol. 53, no. 4, pp. 610–35, Feb 2008.

[4] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk, and H. Hermjakob, "The intact molecular interaction database in 2010," *Nucl. Acids Res.*, vol. 38, pp. D525–D531, 2010.

[5] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations," *Molecular & Cellular Proteomics*, vol. 1, no. 5, pp. 349–356, May 2002.

[6] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D449–51, Jan 2004.

[7] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, and M. Vidal, "High-quality binary protein interaction map of the yeast interactome network," *Science*, vol. 322, no. 5898, pp. 104–10, Oct 2008.

[8] K. Fundel, R. Küffner, and R. Zimmer, "Relex—relation extraction using dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365–371, 2007.

[9] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, and A. Tramontano, "Critical assessment of methods of protein structure prediction - round viii," *Proteins*, vol. 77 Suppl 9, pp. 1–4, 2009.

[10] W. Hersh and E. Voorhees, "Trec genomics special issue overview," *Information Retrieval*, vol. 12, no. 1, Feb 2009.

[11] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity task at jnlpba," in *Proc. Joint Workshop on Natural Language Processing in Biomedicine and its Applications, JNLPBA*, 2004, pp. 70–75.

[12] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii, "Overview of bionlp'09 shared task on event extraction," in *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 1–9.

[13] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, "Overview of biocreative: critical assessment of information extraction for biology," *BMC Bioinformatics*, vol. 6, no. Suppl 1, p. S1, 2005.

[14] 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, no. Suppl 2, p. S1, 2008.

[15] L. A. Hirschman, S. A. Mardis, G. Cesareni, M. Krallinger, F. Leitner, and A. Valencia, "An Overview of BioCreative II.5," *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 2010, Special Issue on BioCreative II.5, to appear.

[16] 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 Res*, vol. 38, no. Database issue, pp. D532–9, Jan 2010.

[17] S. Orchard, S. Kerrien, P. Jones, A. Ceol, A. Chatr-Aryamontri, L. Salwinski, J. Nerothin, and H. Hermjakob, "Submit your interaction data the imex way: a step by step guide to trouble-free deposition," *Proteomics*, vol. 7 Suppl 1, pp. 28–34, Sep 2007.

[18] F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J. Hakenberg, C. Plake, C.-J. Kuo, C.-N. Hsu, R. T.-H. Tasi, H.-C. Hung, W. W. Lau, C. A. Johnson, R. Sætre, K. Yoshida, Y. H. Chen, S. Kim, S.-Y. Shin, B.-T. Zhang, W. A. Baumgartner, L. Hunter, B. Haddow, M. Matthew, X. Wang, P. Ruch, F. Ehrler, A. Özgür, 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, no. S2, p. S6, 2008.

[19] J. Hakenberg, C. Plake, L. Royer, H. Strobelt, U. Leser, and M. Schroeder, "Gene mention normalization and interaction extraction with context models and sentence motifs," *Genome Biology*, vol. 9, no. S2, p. S14, 2008.

[20] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder, and G. Gonzalez, "Inter–species normalization of gene mentions with GNAT," *Bioinformatics*, vol. 24, no. 16, pp. i126–i132, September 2008.

[21] L. Hunter, Z. Lu, J. Firby, W. A. B. Jr, H. L. Johnson, P. V. Ogren, and K. B. Cohen, "OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression," *BMC Bioinformatics*, vol. 9, p. 78, 2008.

[22] L. Smith, L. K. Tanabe, R. J. nee 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, L. Hunter, B. Carpenter, R. T.-H. 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-Vazquez, and W. J. Wilbur, "Overview of biocreative ii gene mention recognition," *Genome Biology*, vol. 8, no. Suppl. 2, p. S2, 2008.

[23] J. Wermter, K. Tomanek, and U. Hahn, "High-performance gene name normalization with geno," *Bioinformatics*, vol. 25, no. 6, pp. 815–21, Mar 2009.

[24] R. Leaman and G. Gonzalez, "Banner: An executable survey of advances in biomedical named entity recognition," in *Proc. Pacific Symposium on Biocomputing*, Kona, Hawaii, USA, 2008, pp. 652–663.

[25] R. Klinger and K. Tomanek, "Classical probabilistic models and conditional random fields," Department of Computer Science, Dortmund University of Technology, Tech. Rep., 2007.

[26] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. John Wilbur, E. Yaschenko, and J. Ye, "Database resources of the national center for biotechnology information," *Nucleic Acids Res*, Nov 2009.

[27] P. Palaga, L. Nguyen, U. Leser, and J. Hakenberg, "High-performance information extraction with alibaba," in *Proceedings of EDBT 2009, Demo*, St. Petersburg, March 23-26 2009.

[28] D. Rebholz-Schuhmann, A. Jimeno-Yepes, M. Arregui, and H. Kirsch, "Measuring prediction capacity of individual verbs for the identification of protein interactions," *Journal of Biomedical Informatics*, vol. 43, no. 2, pp. 200–207, April 2010.

[29] W. A. Baumgartner, Z. Lu, H. L. Johnson, J. G. Caporaso, J. Paquette, A. Lindemann, E. K. White, O. Medvedeva, K. B. Cohen, and L. Hunter, "An integrated approach to concept recognition in biomedical text," in *Proc 2nd BioCreative Challenge Evaluation Workshop*, 2007, pp. 257–271.

[30] S. Jonnalagadda and G. Gonzalez, "Sentence Simplification Aids Protein-Protein Interaction Extraction," in *Proc Int Symp Languages in Biology and Medicine (LBM)*, Seogwipo-si, Jeju Island, South Korea, Nov 8-10 2009.

[31] S. Clark and J. R. Curran, "Wide-coverage efficient statistical parsing with ccg and log-linear models," *Computational Linguistics*, vol. 33, no. 4, pp. 493–552, 2007.

[32] "LingPipe." [Online]. Available: http://alias-i.com/lingpipe/

[33] A. Siddharthan, "Syntactic simplification and text cohesion," Ph.D. dissertation, University of Cambridge, UK, 2003.

[34] E. Ong, J. Damay, G. Lojico, K. Lu, and D. Tarantan, "Simplifying text in medical literature," *Journal of Research in Science, Computing and Engineering*, vol. 4, no. 1, pp. 37–47, 2007.

[35] D. McClosky and E. Charniak, "Self-training for biomedical parsing," in *Proceedings of ACL-08: HLT, Short Papers*. Columbus, Ohio, USA: Association for Computational Linguistics, June 2008, pp. 101–104.

[36] A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski, "A graph kernel for protein-protein interaction extraction," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Columbus, Ohio: Association for Computational Linguistics, June 2008, pp. 1–9.

[37] J. Ding, D. Berleant, D. Nettleton, and E. S. Wurtele, "Mining medline: Abstracts, sentences, or phrases?" in *Proc. Pacific Symposium on Biocomputing*, Kaua'i, Hawaii, USA, Jan. 3-7 2002, pp. 326–337.

[38] S. Kim, S.-Y. Shin, I.-H. Lee, S.-J. Kim, R. Sriram, and B.-T. Zhang, "PIE: an online prediction system for protein-protein interactions from text," *Nucleic Acids Res*, 2008.

[39] R. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. W. Wong, "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, pp. 139–155, 2005. [Online]. Available: ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/

[40] Y. Miyao, K. Sagae, R. Sætre, T. Matsuzaki, and J. Tsujii, "Evaluating contributions of natural language parsers to protein–protein interaction extraction," *Bioinformatics*, vol. 25, no. 3, pp. 394–400, 2009.

Jörg Hakenberg received a doctoral degree (Dr. rer. nat.) from Humboldt-Universität zu Berlin, Germany. He is currently a research associate with the Computer Science Department at Arizona State University. His research interests are bioinformatics, text and data mining, and biological network analysis.

Robert Leaman received his Bachelors of Science degree in Computer Science from Brigham Young University. After spending several years in industry, he is now a Ph.D. student in Computer Science at Arizona State University. His research interests include computational linguistics, text mining, and named entity recognition.

Nguyen Ha Vo obtained a Master's degree in Computer Engineering from Chonnam National University, Korea. He is currently pursuing his Ph.D. at Arizona State University. His research interests are information extraction and biology knowledge representation.

Siddartha Jonnalagadda obtained a Bachelors of Technology (Honors) in Computer Science and Engineering from the Indian Institute of Technology, Kharagpur. He is currently pursuing his Ph.D. in the Department of Biomedical Informatics, Arizona State University, with biomedical natural language processing, computational linguistics and distributional semantics as the main areas of research.

Ryan Sullivan

Christopher Miller is currently pursuing a Medical Doctorate at Temple University School of Medicine, having completed a Master in Science degree in Biomedical Informatics at Arizona State University.

Luis Tari received a Ph.D. degree in computer science from Arizona State University. He is currently a postdoc research fellow in text mining at Hoffmann-La Roche. His research interests include information extraction, information retrieval and bioinformatics.

Chitta Baral is a professor in computer science at Arizona State University. His research interests are in the areas of artificial intelligence, knowledge representation, natural language processing, and bioinformatics.

Graciela Gonzalez is an Assistant Professor in Biomedical Informatics at Arizona State University. Her research interests are in the areas of natural language processing, knowledge representation, and translational bioinformatics.