

# Consensus pattern alignment to find protein-protein interactions in text

Jörg Hakenberg<sup>1</sup>Michael Schroeder<sup>1</sup>Ulf Leser<sup>2</sup>

hakenbergj@biotec.tu-dresden.de ms@biotec.tu-dresden.de leser@informatik.hu-berlin.de

<sup>1</sup> Biotechnological Centre, Technische Universität Dresden, 01307 Dresden, Germany<sup>2</sup> Computer Science Department, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

“Don’t I know you from somewhere?” – comparing new to known texts plays a key role in the system we propose for searching protein–protein interactions (PPIs). Our system builds on an inexact pattern matching strategy, where patterns (linguistic frames) reflect the compositional structure of known occurrences of PPIs in text. To describe this structure, part-of-speech tags (verbs etc.) and entity classes (proteins), words, and word stems are used. Consider the sentences “Sky1p phosphorylates Npl3p” and “Akt phosphorylates beta-catenin”. Both have a structure in common that connects two proteins with a single verb. From comparable systems proposed before [1, 2], it became clear that collecting a suitable set of patterns is of major importance, and this step forms the main component of our system. From the IntAct database [5], we extract all pairs of proteins known to interact. We scan PubMed for textual evidences for each such interaction, and retain all single sentences that describe them. Using pairwise sentence alignment as a similarity scoring function, we perform a clustering on the resulting set of sentences. Within each cluster, multiple sentence alignment (MSA) identifies commonalities and variable positions across all sentences, expressed in a consensus pattern. Figure 1 shows an example MSA with four sentences that define one consensus pattern. We can now align such consensus patterns against arbitrary text to extract new PPIs.

Our system yields a maximum recall of 69% –which was the best reported among all participating systems–, a maximum precision of 45% and maximum F1-measure of 41% on the BioCreative test set. Our method works completely independent from the training corpus, which we did not use at any stage. Thus, we intrinsically exclude any risk of overfitting, and believe that our approach should work equally well for related extraction problems, such as finding protein–disease associations.

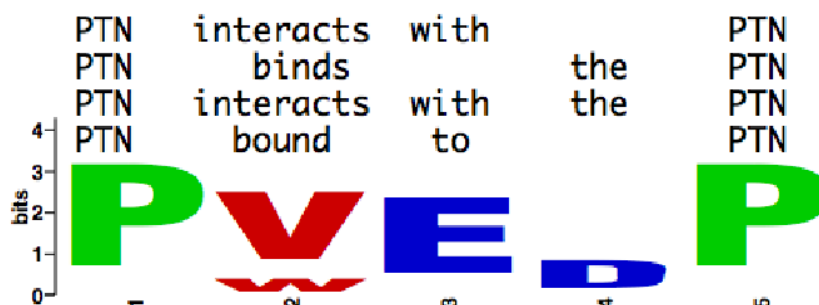


Figure 1: Sequence logo –the consensus pattern– for four short sentences. Height of a character corresponds to the information content (entropy) at this position; the larger a character, the more conserved it is across multiple sentences. PTN/P are wildcards for any protein name; V, verb, present tense; W, verb, simple past; E, preposition; D, determiner. Logo created with WebLogo ([weblogo.berkeley.edu](http://weblogo.berkeley.edu)).

## Methods

The system we propose falls into two components: searching sentences that contain two identified proteins and searching for PPIs described in these sentences. The initial recognition of protein names is based on a dictionary derived from UniProt (protein names, gene names and respective synonyms). The identification is a variation of the system we presented for the GN task (see elsewhere in this proceedings). The extraction of PPIs builds on ideas presented with the Ali Baba tool [3].

### *Named entity recognition*

For the initial recognition of protein names, we built a dictionary using synonyms provided by UniProt/TrEmbl (description and gene name fields) for the approximately 200,000 proteins listed in the IPS data set. Each synonym mapped to all UniProt identifiers that share this synonym. Multiple IDs appeared mainly for abbreviations, which often have different expansions, and proteins shared across multiple organisms. We added term variations (plural/singular forms, changes in capitalization, structural variations) to this dictionary. It was also very important to expand the list of candidate IDs by cross-checking for proteins sharing similar synonyms in UniProt. For example, UniProt contains the name “Hoxb6” only for a protein in the mouse, and uses the name “HOXB6” for human and others. From the training data it became clear that authors often would use “Hoxb6” to refer to the human ortholog, however. Thus, we iteratively expanded the list of IDs for each name variant based on case-insensitive comparisons. We finally compiled a finite state automaton from all entries for fast spotting of candidate names in text.

### *Named entity normalization*

Named entity normalization (NEN) was a very important step in the IPS task, and a proper protein name disambiguation was necessary. Our disambiguation builds on a subsequent reduction of candidate UniProt IDs for each recognized name (see our GN task paper.) The highest impact on performance came with the reduction to organisms. We used the Ali Baba tool to recognize organism names in the corresponding abstracts. We compared these identified organisms to the annotations of each potential UniProt entry. Comparison was based on the controlled vocabulary provided by UniProt [6], which we enriched using the NCBI Taxonomy to include other common names, as well as manual curation (so that “patients” would trigger “human.”) Sometimes it was not possible to restrict the IDs to only one candidate. In such cases, we would report the first standard name (for higher precision) or all remaining (higher recall.) We noticed that in most cases, at least one standard name (out of a predicted PPI pair) was correctly found by the disambiguation. When the second was not correctly found, however, this still accounted for an overall false positive and a missing annotation.

### *Interaction extraction*

We applied a sentence alignment against a pre-compiled set of patterns on every sentence that contained at least two proteins. Such patterns describe typical occurrences of evidences that mention PPIs. Very simple examples for these would be [ *protein* binds to *protein* ] or [ *protein* bound to the *word* domain of *protein* ]. Here, *protein* and *word* are wildcards for every protein recognized by the first component, or arbitrary words, respectively. To find such patterns, we applied the following strategy. First, we collected a large set of sentences from PubMed that most likely describe PPIs. To find such sentences, we used the IntAct database [5] and searched for sentences that contain an interaction pair in PubMed. Each protein in IntAct can be mapped to a UniProt ID and, using the above recognition, we scanned the full PubMed database for any occurrence of a pair of proteins known to interact. We reduced each sentence to the core phrase (potentially) describing the interaction and searched for typical words (“binds”, “associated”, “complex”, etc.) For more details, please refer to [4], examples are shown in Table 1. Starting with a set of more than 200,000 such sentences, we computed a pairwise similarity using sentence alignment. The input for these alignments were tokens, token stems, and part-of-speech tags for each position in a sentence. A distance matrix containing pairwise alignment scores for all pairs of core phrases was used to construct a guide tree for clustering (comparable to ClustalW.) On each cluster, we then performed a multiple sentence alignment to compute a consensus pattern that best describes the sentences. Figure 1 shows an example for an MSA to compute such a consensus pattern (POS tags only.) We found ca. 10,000 such consensus patterns, many of which were as simple as the aforementioned examples, but with many rather complex patterns as well.

Sentence alignment provides an inexact matching strategy for sequences of words; this allows for (often observed) deletions or insertions of words with minor influence on the overall statement (for instance, adjectives and determiners.) Consensus patterns bring two main advantages: i) they consist

of word sequences actually observed in evidence texts and are thus very specific; ii) they combine observations made across multiple evidences into one pattern and thus generalize well.

---

<i>protein binds to protein</i>
( <i>protein</i> ) <b>binds</b> to its <b>receptor</b> ( <i>protein</i>
<i>protein binds to the cytoplasmic tail of protein</i>
<i>protein recruits the adapter molecule protein</i>
<i>protein site was specifically recognized by c- protein</i>
<i>protein and protein compete for binding to protein</i>
( <i>protein</i> ) results in <b>decreased</b> <i>protein</i> synthesis
Arabidopsis <i>protein</i> ( <i>protein</i> ) <b>associates</b> with both <i>protein</i> and <i>protein</i>
cytosolic <i>protein</i> is <b>associated</b> with a <b>complex</b> of <i>protein</i> ( <i>protein</i> )
<i>protein</i> , a modular <b>adapter</b> which in muscle cells <b>interacts</b> with members of the <i>protein</i> family including <i>protein</i>
<i>protein induces activation</i> of coagulation and fibrinolysis through an exclusive <b>effect</b> on the <i>protein</i>
<i>protein</i> was previously found to <b>interact</b> with the KRAB silencing domain of <i>protein</i> and with the <i>protein</i>

---

Table 1: Examples for phrases collected from PubMed. Sentences were reduced to their core. *protein* indicates proteins of arbitrary name, while all other words and symbols appeared as such; interaction words are **bold**.

## Analysis

Short description	Precision	Recall	F1 (in%)
<i>min=2; ids=2; organism=a,h,m,y,l</i>	7.7	69.4	13.2
<i>min=3; ids=1; organism=a,l</i>	15.0	65.1	22.5
<i>min=1; ids=1; organism=a,l</i>	44.5	41.7	40.5

Table 2: Results for different strategies on the IPS test set. *min*, minimum of identified interactions per pair and article required for a prediction; *ids*, number of submitted IDs per protein in case more than one was left after NEN; *organism*, order of assignment to organisms for unresolved proteins: take organism found in abstract, take human, mouse, yeast, or highest ranked gene (1).

Table 2 shows the results of our method depending on different settings. First, we see that proper NEN was crucial regarding the overall outcome. We found that associating a protein with an organism was quite easy, and our paper for the GN task discusses how intra-organism ambiguities could be solved. We encountered most NEN-related problems as a result from erroneous PDF to text conversion, an issue that has been discussed elsewhere. For example, in many of the plain texts, Greek letters, which were crucial to identify the right member of a family, were missing. Some false positive predictions were found as discussed in dangling text or not annotated in the gold standard for various reasons (different understanding of an interaction; not main thrust of publication.) Thus, tuning towards IPS-task-specific annotations on the training corpus might help. Regarding the “main thrust”, we found that many of the PPIs were discussed quite often within a single publication, so even requiring at least three evidences did not influence the recall much, but increased the precision. PPIs mentioned only once in the Introduction, for instance, could be filtered out. Evaluations of our approach on other corpora revealed quite different results. On the SPIES corpus [2], the method showed a precision around 80% at 50% recall. There are two differences compared to the results on IPS: (1) the figures are lower in general and (2), the order of precision and recall have changed. NER/NEN is not necessary for SPIES, which consists of 1000 single sentences that all contain at least one PPI.

## References

- [1] Blaschke, C. and Valencia, A., The Frame-Based Module of the SUISEKI Information Extraction System, *IEEE Intelligent Systems*, 17(2):14–20, 2002.
- [2] Huang, M., Zhu, X., Hao, Y., Payan, D.G., Qu, K., and Li, M., Discovering patterns to extract protein-protein interactions from full texts, *Bioinformatics*, 20(18):3604–3612, 2004.
- [3] Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., and Leser, U., AliBaba: PubMed as a graph, *Bioinformatics*, 22(19):2444–2445, 2006.
- [4] Hakenberg, J., Leser, U., Kirsch, H., Rebholz-Schuhmann, D., Collecting a large corpus from all of Medline, *Proc. Symposium on Semantic Mining in Biomedicine*, 2006.
- [5] See <http://www.ebi.ac.uk/intact/>
- [6] See <http://www.expasy.org/cgi-bin/speclist>