Collecting a Large Corpus from all of MEDLINE

Jörg Hakenberg* **Ulf Leser**
Knowledge Management
in Bioinformatics
Humboldt-Universität zu Berlin
Berlin, Germany
{hakenberg,leser}@informatik.
hu-berlin.de

Harald Kirsch **Dietrich Rebholz-Schuhmann**
Rebholz group
European Bioinformatics
Institute (EBI)
Hinxton/Cambridge, UK
{kirsch,rebholz}@ebi.ac.uk

Abstract

We present our ideas and first results for a system to extract interactions between proteins from scientific publications. This system consists of three main stages. First, we extract a large sample of sentences from unannotated text. Second, we generate language patterns using multiple sentence alignment to identify consensus phrases. Last, we apply these patterns to arbitrary text, again using sentence alignment. In this paper, we concentrate on the first step, were we extract a training sample from MEDLINE. We search for occurrences of both partners of a known protein-protein interaction in a single sentence and further refine the resulting set to exclude false positives. We are able to extract almost 68,000 examples for sentences that discuss protein-protein interactions.

1 Introduction

The extraction of protein-protein interactions from scientific literature has become one of the best studied applications for text mining during the last years. Many systems have been proposed that aim to solve this and related problems, for instance protein-gene and gene-gene relations. Most approaches presented are based on statistics, Markov modeling (Xiao et al., 2005), sentence parsing (Daraselia et al., 2004), rules (Saric et al., 2005), or patterns (Blaschke and Valencia, 2002; Hao et al., 2005; Hakenberg et al., 2005). Many machine learning techniques strongly depend on training samples, and even some pattern matching techniques learn a pattern set from an annotated sample. Manually annotating such samples is very time-consuming and tedious, and requires domain experts, for whom there is often no immediate benefit. For these and other reasons, there exist only a few and small annotated samples for the task of protein-protein interaction extraction. SPIES provides a corpus of 891 sentences annotated for protein-protein interactions (Hao et al., 2005), as well as IEPA (Ding et al., 2002), and others[1]. Combining those different corpora to one large sample typically leads to inconsistent, contradictory, and maybe overlapping data. Some results from cross-training experiments (named entity recognition) have been published, for example in Zhou et al. (2004). For other tasks, there are even less or none publicly available sources. One of the main ideas we pursue in this project deals with automatically gathering a training corpus for subsequent learning methods.

The approach of matching language patterns against arbitrary text for the task of protein-protein interaction extraction has been studied well during the last years (see above). Depending on the system, either these patterns (also referred to as rules or frames) contain generalizations, or the matching techniques tolerate changes. For instance, the insertion of an adjective or determiner never changes the overall meaning. This way, patterns capture not only exact matches, but also slight variations of language. Previous experiments revealed that searching for language patterns can be a very precise method (around 90%). The achieved recall values are not convincing, at least at high levels of precision. Keeping a certain level of precision does not allow for too broad generalizations, but on the other hand, even marginally less strict patterns would boost the recall. An ob-

[1] See http://compbio.uchsc.edu/corpora/ for a collection.

* To whom correspondence should be addressed.

vious idea is to keep patterns as strict as possible, and just increase their overall number to ensure high recall. (Blaschke and Valencia, 2002) mention a set of 31, (Hao et al., 2005) start with 192 patterns and reduce them subsequently. Most times, there is much manual work involved in the creation and curation of patterns, or the initial pattern set gets learned from manually annotated sample sentences.

Our idea to overcome the shortage of samples is to automatically extract a large number of examples from all of MEDLINE, using a fact database to look up possible relations. This should be applicable to any types of relations, such as protein-protein/protein-gene interactions, enzyme kinetics, drug-target relations, etc., as long as corresponding fact database is available. The system simply looks for every encountered pair of entities, if this is also contained in the database. In addition, we filter for the occurrence of certain words near the mentioning.

2 System and methods

We parse all of MEDLINE for a context that contains at least one pair of the entity class needed. In our case, a context always is a single sentence, but this can easily be adjusted to two or more consecutive sentences or a paragraph, for instance. We then apply named entity recognition to the sentence. For this step, is is mandatory to not only recognize names, but map them to their respective identifiers in a database. We use the tool described in Kirsch et al. (2005), that recognizes protein names and maps them to the UniProt database (Bairoch et al., 2005). The dictionary for this tool so far includes 195,908 terms, and was parsed from the database entries (names and synonyms). One term might fit multiple UniProt entries (IDs), and multiple entries might map to the same ID. We extracted 41,748 pairs of interacting proteins from the IntAct database (Hermjakob et al., 2004) (31,471 proteins, May 2005). Whenever we find two or more proteins in a sentence, we search for every possible combination of these in the IntAct pairs. This way, a single sentence might mention more than one interaction at once. With an additional filter we try to exclude simple enumerations and other 'random' co-occurrences of IntAct pairs. We skip all sentences that do not contain a word referring to an interaction between proteins. Such words are, for instance, 'phosphorylates', 'inhibitor', etc.. All in all, we use 160 nouns, 520 verbs, and seven adjectives (including number and conjugations), based on the collection provided by Temkin and Gilder (2003).

For our own pattern matching technique that searches for patterns in arbitrary text and deduces the semantics of the new text, we use a method that resembles sequence alignment. We reduce all sentences collected from MEDLINE to their 'core', that is where the interaction is mentioned, plus a certain boundary. These cores are much shorter, and do not contain the large variety to the left and right of them. The alignment is much faster with shorter patterns, there are less patterns, and not for every possible full sentence there has to be a pattern. The alignment uses an end-space-free strategy, so that the shortest subsentence is matched against the pattern.

3 Results

We collected almost 68,000 sentences from MEDLINE that contain at least one pair from IntAct (see Table 1). 31,000 of these contain an interaction verb; almost as many contain an interaction noun; and there are some that contain both.

Manual inspection of the collection revealed that both the quality and variety are surprisingly good. Table 2 lists a few examples for patterns containing interaction verbs.

We stored all patterns in the form '`ANY/PTN inhibit/v:G ANY/PTN`', each pair depicting the token and part-of-speech tag observed. '`ANY/PTN`' refers to a protein with an arbitrary name, and '`inhibit/v:G`' refers to a verb in gerund form, in this case the token 'inhibit'. For all patterns and part-of-speech tags, please consult the supplementary information.

4 Discussion and Conclusions

In this paper, we have presented a system to extract an annotated corpus from large text collections using a fact database. In our case, we collected sentences describing protein-protein interactions occurring in IntAct from all of MEDLINE. Ideally, this method provides a large sample, not only useful for our subsequent steps, but as a corpus for training systems proposed by others as well. In addition, it could be easily adapted to collect corpora for other tasks as well, such as enzyme kinetics (taking BRENDA (Schomburg et al., 2004) or Kinetikon (Menz et al., 2005) as databases), pro-

proteins in IntAct	31,471
protein pairs from IntAct	41,748
sentences with at least one IntAct pair	67,870
IntAct pairs in the sentences	117,460
sentences with IntAct pair + interaction verb	3,498
reduced to zero-word boundary, uniques	21,191
reduced to one-word boundary, uniques	24,403

Table 1: Statistics for MEDLINE sentences and interaction pairs from IntAct. Uniques include the reduction of protein names to their entity class.

PTN) *binds* to PTN
PTN) *binds* to its receptor (PTN
PTN) *binds* to the cytoplasmic tail of PTN
PTN is *associated* with PTN
PTN (+) T cells *expressing* PTN
PTN) , which *encodes* the Drosophila PTN
PTN) *recruits* PTN
PTN) *recruits* the adapter molecule PTN
PTN site near the promoter *bound* c- PTN
PTN site was specifically *recognized* by c- PTN

Table 2: Some examples for language patterns collected from MEDLINE. Sentences were reduced to their core, omitting boundaries. 'PTN' indicates proteins, while interaction words are in *italics*.

tein annotations from Gene Ontology (GOA, Camon et al. (2004)), targets of drugs (TTD, Chen et al. (2002)), and many others. Prerequisites are, however, sufficiently good named entity recognition, and further filters like words indicating associations.

In addition to simple co-occurrences of entities in a sentence that also contains a word referring to a relation, we try to ensure quality by taking only entities known to interact with each other. Named entity recognition itself guarantees precision/recall levels of about 80% only. A false positively recognized protein is unlikely to take part in an interaction known from IntAct. Predicting two false positives, where this pair also occurs in IntAct, is even less likely. By scanning for words indicating interactions, we reduce the probability of extracting a 'false' sentence even further. At the moment, we exctract interactions from single sentences only. It is clear, however, that many evidences for relations spread across multiple sentences or even whole paragraphs. We thus study how our approach can be altered to find these occurrences as well, though it is much harder to remove false positives.

Other than providing a simple and fast method to collect an annotated corpus, we pursue several other ideas. From a computer linguistic point of view, it might be interesting to study the existence and usage of certain patterns of speech. This means, for example a particular verb or group of verbs would determine the basic structure of the sentence it is used in, or is even used only in a defined manner. Such analyses can be performed very easily on a large sample corpus.

Future research directions

Our next steps will focus on the application of the collected language patterns to the extraction of protein-protein interactions from arbitrary text. Even after reduction the pattern set is still too large to apply every single one to every new sentence. In addition, the pattern set misses a lot of interactions, because it is in no way generalized, and even its large size surely cannot grasp all varieties of human language. We would like to generalize the patterns by subsequently transforming the most similar patterns into *consensus patterns*, as provided by *multiple sentence alignment*. These consensus patterns describe motifs commonly observed in scientific text, using information such as tokens, lemmata, part-of-speech tags, etc. They form word sequences containing not only fixed positions, but also free positions and weighted lists of

possible word occurrences, even interdependent.

The support of each (consensus) pattern is given with a data set (the collected language patterns or a benchmark corpus). It determines the pattern's usefulness and reveals possible containment in other patterns, when calculated separately and jointly. Precision and recall of each pattern can be determined as well. For fast applications, only the best n patterns could be used.

5 Supplementary information

For supplementary information, please see http://www.informatik.hu-berlin.de/~hakenber/publ/suppl/.

6 Acknowledgements

This work is supported by the German Federal Ministry of Education and Research (BMBF) under grant contract 0312705B. The Knowledge Management in Bioinformatics Group is a member of the Berlin Center for Genome Based Bioinformatics (BCB). Funding for the Rebholz group is provided by the Network of Excellence "Semantic Interoperability and Data Mining in Biomedicine" (NoE 507505). JH was additionally supported by the German Foreign Exchange Service (DAAD), reference number D/05/26768.

References

Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, *et al.* 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, 33(Database issue):D154–9.

Christian Blaschke and Alfonso Valencia. 2002. The Frame-Based Module of the SUISEKI Information Extraction System. *IEEE Intelligent Systems*, 17:14–20.

Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, *et al.* 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32(Database issu):D262–D266.

Xin Chen, Zhiliang Ji, and Yuzong Z. Chen. 2002. TTD: Therapeutic Target Database. *Nucleic Acids Research*, 30(1):412–415.

Nikolai Daraselia, Anton Yuryev, Sergej Egorov, Svetlana Novichkova, Alexander Nikitin, and Ilya Mazo. 2004. Extracting human protein interactions from medline using a full-sentence parser. *Bioinformatics*, 20(5):604–611.

Jing Ding, Daniel Berleant, Dan Nettleton, and Eve S. Wurtele. 2002. Mining MEDLINE: Abstracts, Sentences, or Phrases? In *Pacific Symposium on Biocomputing (PSB)*, pages 326–337, Kaua'i, Hawaii, USA, Jan. 3-7.

Jörg Hakenberg, Conrad Plake, Ulf Leser, Harald Kirsch, and Dietrich Rebholz-Schuhmann. 2005. LLL'05 Challenge: Genic Interaction Extraction with Alignments and Finite State Automata. In *Proc Learning Language in Logic Workshop (LLL05) at the 22nd Int Conf on Machine Learning*, Bonn, Germany.

Yu Hao, Xiaoyan Zhu, Minlie Huang, and Ming Li. 2005. Discovering patterns to extract protein-protein interactions from the literature: Part II. *Bioinformatics*, 21(15):3294–3300.

Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, *et al.* 2004. IntAct: an open source molecular interaction database. *Nucl Acid Res*, 32(Database issue):D452–D455.

Harald Kirsch, Sylvain Gaudan, and Dietrich Rebholz-Schuhmann. 2005. Distributed modules for text annotation and IE applied to the biomedical domain. *Int Journal of Medical Informatics*. Advance access online.

Stephan Menz, Christof Dehmel, Axel Kowald, and Edda Klipp. 2005. Kinetikon - database for kinetic parameters. In *Workshop Datenbanken im NGFN*, Cologne, Germany, March 9.

Jasmin Saric, Lars Juhl Jensen, Rossitza Ouzounova, Isabel Rojas, and Peer Bork. 2005. Large-scale Extraction of Protein/Gene Relations for Model Organisms. In *Proc Symp on Semantic Mining in Biomedicine*, page 50, Hinxton, UK, April 10-13.

Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, *et al.* 2004. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*, 32:D431–D433, Jan 1.

Joshua M. Temkin and Mark R. Gilder. 2003. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 19(16):2046–2053.

Juan Xiao, Jian Su, GuoDong Zhou, and ChewLim Tan. 2005. Protein-Protein Interaction Extraction: A Supervised Learning Approach. In *Proc Symp on Semantic Mining in Biomedicine*, pages 51–59, Hinxton, UK, April 10-13.

GuoDong Zhou, Dan Shen, Jie Zhang, Jian Su, Soon-Heng Tan, and ChewLim Tan. 2004. Recognition of protein/gene names from text using an ensemble of classifi ers and effective abbreviation detection. In *BioCreAtIvE Workshop*, Granada, Spain, March 28-31.