Finding small molecule and protein pairs in scientific literature using a bootstrapping method

Ying Yan, Jee-Hyub Kim, Samuel Croset, Dietrich Rebholz-Schuhmann
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton
Cambridge
UK
{yan, jhkim, croset, rebholz}@ebi.ac.uk

Abstract

The relationship between small molecules and proteins has attracted attention from the biomedical research community. In this paper a text mining method of extracting small-molecule and protein pairs from natural text is presented, based on a semi-supervised machine learning approach. The technique has been applied to the complete collection of MEDLINE abstracts and pairs were extracted and evaluated. The results show the feasibility of the bootstrapping system, which will subsequently be further investigated and improved.

1 Introduction

Information extraction has become a major task in text-mining. A large number of studies have been carried out with the objective of developing techniques to overcome the highly ambiguous and variable nature of natural language for the extraction of information from scientific text (Song et al., 2006). Natural language processing (NLP) of biomedical text has been initiated and used for different knowledge discovery tasks such as the extraction of relationships between different types of biological objects.

Relationships between proteins and small molecules are of particular concern in the biomedical research domain. The importance of target specific small molecule research is vital in the scientific community's understanding of numerous biological processes with potential discoveries yielding various translational benefits and outcomes to public health and industry. While there has been a great number of traditional studies already completed in this field, the underlying difficulty with this type of research has been trying to understand how one molecule interacts with a target protein. Given the biological background, many researchers in Cheminformatics and Metabolomics are attempting to find the connections between small molecules and other biological entities in order to bridge the chemical and biological domains.

Of the few reported text mining approaches to this problem, Temkin and Gilder (2003) was concerned with the extraction of protein and small molecule interaction, and used a rule-based approach utilising a lexical analyser and context free grammar. Jiao and Wild (2009) presented a technique for detecting protein and small molecule interaction using a maximum entropy based learning method; this work also uses corpus-based machine learning. The main drawback of both of these studies is that they require a fully annotated corpus which is difficult to generate.

1.1 The bootstrapping method

At present a gold standard annotated corpus is not available, and constructing a reasonable annotated corpus would require an infeasible amount of manual work. Our proposed solution to this problem is to develop a semi-supervised machine learning method. In this paper a bootstrapping algorithm is presented which requires only unannotated training texts and a handful of protein small molecule pairs, known as seeds. The basic work of a bootstrapping system can be presented as an expansion engine which uses the initial seed pairs fed into the

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 172–175,
Montréal, Canada, June 8, 2012. ©2012 Association for Computational Linguistics

system to generate patterns that are used, in turn, to find more pairs. The operation of the algorithm is controlled by certain criteria that are delivered from a measurement of the quality or selectivity of patterns and discovered pairs.

Bootstrapping systems have been maturely used for information extraction purposes in other research domains, and it has been empirically shown to be a powerful method in learning lexico-syntactic patterns for extracting specific relations (Riloff and Jones, 1999). Bootstrapping systems can operate with a greatly reduced number of training examples. A bootstrapping system seems promising for the purpose of relation extraction, making it a suitable candidate method for protein and small molecule pair extraction.

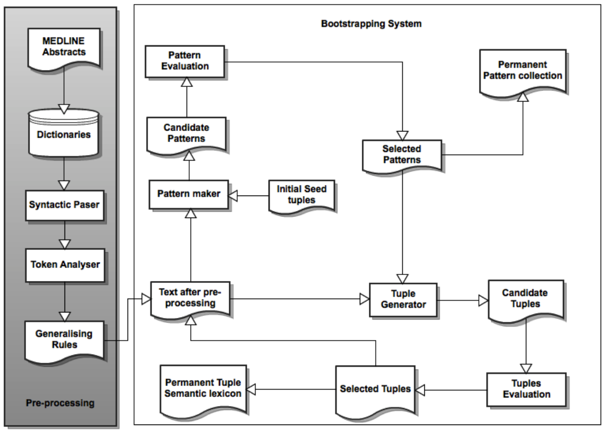

Figure 1: Extraction system architecture

2 Implementation

The typical bootstrapping method was tailored in order to improve its suitability for our extraction task, operating in the biomedical literature resource MEDLINE. The bootstrapping architecture is presented in Figure 1. The whole collection of MEDLINE was filtered using a co-occurrence approach and a named entity recogniser. In this way the sentences which contained both a protein and a small molecule were selected. The structure of patterns which are suitable to extract protein and small molecule pairs from MEDLINE was defined. Each sentence is tokenized and then normalised based on the results of syntactic parsing in order to obtain a more generalised view of the pattern. In the following sections, we describe in more detail these aspects.

2.1 Protein and small molecule recognition

Two dictionary-based named entity recognisers were used to detect the names of proteins and small molecules in the full collection of MEDLINE abstracts, with the two source dictionaries constructed using the resources UniProt (Apweiler et al., 2004) and ChEBI (De Matos et al., 2006) respectively. The following example shows the two recognisers identify a chemical object and a protein object in a sentence from a MEDLINE extract:

<chebi>Paracetamol</chebi>, 100 mg/kg, inhibited <uniprot>COX-1</uniprot> in stomach mucosa ex vivo much less effectively than in other tissues.

2.2 Sentence analysis for normalisation

It was anticipated that variations in tense and other language characteristics would cause problems in pattern generation. We therefore applied a list of normalisation steps for pattern generation. The surrounding context in the biomedical text is not normally useful and makes it difficult to identify the text and observe a clear sentence structure. The parsing result normalises patterns by eliminating non-useful components in a sentence. The step of normalisation hence increases the quality of the pattern.

The complete list of normalisation steps is as follows:

1. Replaced the representation of measurement units, such as mg/L and ml/day.
2. Employed the part-of-speech (POS) tagger GENIA (Tsuruoka et al., 2005) to analyse each token, and the tokens which are weakly related to the sentence structure were removed. So that, the only remaining tokens are the head noun of a noun phrase (NP), the verb phrase, and prepositional phrase chunks.
3. Finally a simple rule to identify the head noun was defined. In a general case, for a NP sequence, the last token is considered as the head noun. When the last token is a single character, the second last token is considered as the head noun.

Table 1: An example of a generated pattern

Seed tuple:	Paracetamol, COX-1
Found string:	*"CHEBI, UNIT, inhibit UNIPROT in mucosa than in tissue."*
Pattern:	NP_List1, UNIT, inhibit NP_List2

Constraints:	NP_List1="CHEBI*"
	NP_List2="UNIPROT*"

Keywords:	",UNIT,inhibit"

The above example after these normalisation steps becomes:

CHEBI, UNIT, inhibit UNIPROT* in mucosa than in tissue.*

where *CHEBI** and *UNIPROT** are the seeds in context.

2.3 Bootstrapping

The bootstrapping system is applied to the normalised sentences. The process starts with 100 high precision protein small molecule pairs collected from the ChEBI ontology. These pairs were retrieved by querying the ChEBI sub-ontology for the relation "has role". From the resulting data we extracted small molecules that are enzyme inhibitors together with the name of the enzyme.

2.3.1 Pattern generation and pair extraction

The concept of a bootstrapping system is that using a high precision seed pair to start the extraction engine, the system can effectively learn the pattern construction rule and the pattern constraints. Searching for the seed pairs in the corpus returns strings which are candidate extraction patterns for other pairs. The candidate patterns are made up of 'slots' and 'context strings', where the slots are either of type small-molecule or protein, and context is the text connecting the slots and the words immediately before and after the pair. By analysing the surrounding context of the slots new elements of the pattern are discovered, which can subsequently be used to search for new small-molecule protein pairs. The process of deriving a pattern from the above example is shown in Table 1.

The generated pattern can then be used to search the corpus and find other matching contexts. New pairs are retrieved from the matching context by simply locating the protein and small molecule names from the same positions as they are in the pattern.

For instance, the pattern produced in Table 1 is matched against a normalised sentence *"data suggest CHEBI, UNIT, inhibit UNIPROT"*, extracting the new pair *<trifluoperazine, CaMKII>*.

2.3.2 Evaluating seeds and patterns

The quality of the pattern is critical since patterns that generate a bad pair can introduce more false positive seeds. Therefore, within a bootstrapping system it is necessary to have a stage of pattern evaluation. Estimations of the confidence score of a pattern can be used as one of the stopping criteria. We implemented an evaluation step for both patterns and pairs based on an evaluation method developed by Agichtein and Gravano (2000). Adapting the approach to this work, if $pattern_i$ predicts tuple $t = <chemical, protein>$, and there is already a tuple $t' = <chemical, protein'>$ with high confidence, and $chemical$ from t is same as $chemical$ from t', then we could define this as a positive match of pattern ($P_{positive}$), otherwise the pattern is considered as a negative match ($P_{negative}$). So that the confidence score of pattern (P) is estimated as:

$$Conf(P) = \frac{P_{positive}}{P_{positive} + P_{negative}} \qquad (1)$$

To evaluate the pairs we again employ the method described by Agichtein and Gravano (2000). The confidence of a particular pair is a function of the number of patterns that generate it. Equation 2 shows how to calculate a confidence score for tuple T, where P is the set of patterns that derive T. C_i is the context that also contains T, $Match(C_i, P_i)$ is the degree of match of C_i and P_i.

$$Conf(T) = 1 - \prod_{I=0}^{|P|} \left(1 - \left(Conf\left(P_i\right) \cdot Match\left(C_i, P_i\right)\right)\right) \qquad (2)$$

3 Results and discussion

Table 2 shows the top 10 generated patterns ranked by the frequency that they appear in MEDLINE. As can be seen the patterns all have very simple structures. Simple patterns are more likely to be productive, i.e the simpler the structure of the pattern, the more pairs it generates. However, simple structures are also likely to generate more false negative pairs.

The pairs produced by these top 10 patterns were collected, and the confidence score then calculated using equation 1. The result implies that the confidence score of a pattern, and in turn the selectivity and productivity of the pattern, are strongly associated with the pattern's structure.

Table 2: The top 10 comment patterns

Frequency	Pattern	Confidence
68	UNIPROT* CHEBI* CHEBI	0.16
61	CHEBI* UNIPROT* UNIPROT	0.15
51	CHEBI* UNIPROT* be	0.10
49	CHEBI* UNIPROT* CHEBI	0.10
41	UNIPROT* CHEBI* be	0.21
40	CHEBI* UNIPROT*	0.08
38	UNIPROT* CHEBI* UNIPROT	0.16
37	UNIPROT* CHEBI*	0.30
26	be CHEBI* UNIPROT*	0.26
24	UNIPROT* CHEBI CHEBI* CHEBI	0.17

3.1 Quality of the extracted pairs

One hundred pairs extracted by first and second generation patterns were randomly selected for manual inspection by a domain expert curator. It was found that over 60% were valid pairs. From further examination of the cases together with their extraction patterns, it can be seen that the patterns have a high confidence score, ensuring the quality of the extracted pair. For instance, from the original text *Paracetamol, 100 mg/kg, inhibited COX-1 in stomach mucosa ex vivo much less effectively than in other tissues*, the pattern "CHEBI*, UNIT, inhibit UNIPROT*" with 0.62 confidence score derives a correct pair <*Paracetamol, COX-1*>.

Generally speaking, simple patterns are more likely to have lower confidence scores. However it was also found that the pattern quality heavily depends on the quality and reliability of the name entity recognition (NE) system.

4 Conclusions and future work

We have presented a method of detecting small molecule and protein pairs in MEDLINE abstracts. It employs semi-supervised machine learning methods to enable patterns to be automatically generated, rather than requiring human input. The approach can be used for high throughput text mining applications where manual curation is unrealistic.

The first and second iteration of results are promising and show that the approach enables many useful small molecule protein pairs to be extracted from MEDLINE using just a small number of seed pairs as input. The approach makes use of a rigorous method of evaluating the quality of generated patterns and extracted pairs. Manual inspection has been used to validate these preliminary results and has shown that approximately half of the discovered pairs represent valid small molecule protein relationships, and we expect to improve this significantly.

In future we will develop the method further and analyse the results after further algorithm iterations, enabling discovery of new patterns and consequently new pairs of proteins and small molecules that are currently undetected.

References

E. Agichtein and L. Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM.

R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, et al. 2004. UniProt: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl 1):D115–D119.

P. De Matos, M. Ennis, M. Darsow, M. Guedj, K. Degtyarenko, and R. Apweiler. 2006. ChEBI-chemical entities of biological interest. *Nucleic Acids Research*, Database Summary: 646.

D. Jiao and D.J. Wild. 2009. Extraction of CYP chemical interactions from biomedical literature using natural language processing methods. *Journal of chemical information and modeling*, 49(2):263–269.

E. Riloff and R. Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the National Conference on Artificial Intelligence*, pages 474–479. John Wiley & Sons Ltd.

M. Song, I.Y. Song, X. Hu, and H. Han. 2006. Information extraction in biomedical literature. In J. Wang, editor, *Encyclopedia of Data Warehousing and Data Mining*, pages 615–620. Information Science Reference.

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

Y. Tsuruoka, Y. Tateishi, J.D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. *Advances in informatics*, pages 382–392.