

# Flexible Text Mining Strategies for Drug Discovery

**David Milward**  
Linguamatics Ltd.  
St. John's Innovation Centre  
Cowley Rd., Cambridge, UK  
david.milward  
@linguamatics.com

**Christian Blaschke**  
bioalma  
Ronda de Poniente 4 2° C-D,  
28760 Tres Cantos,  
Madrid, Spain  
blashcke@bioalma.com

**Jean-Marc Neefs**  
Johnson & Johnson  
Janssen-Pharmaceutica, NV  
Turnhoutseweg 30,  
2340 Beerse, Belgium

**Mark-Christoph Ott**  
Johnson & Johnson  
Janssen-Pharmaceutica, NV  
Turnhoutseweg 30,  
2340 Beerse, Belgium

**Rudi Verbeeck**  
Johnson & Johnson  
Janssen-Pharmaceutica, NV  
Turnhoutseweg 30,  
2340 Beerse, Belgium

**Andrew Stubbs**  
Johnson & Johnson  
Janssen-Pharmaceutica, NV  
Turnhoutseweg 30,  
2340 Beerse, Belgium

## Abstract

The two text mining strategies: finding co-occurrences of biological entities within documents, and finding relationships using Natural Language Processing, are often seen as competitors. Here we adopt a flexible approach where the techniques are adapted and combined to suit the nature of the document corpus, and the specific task. The approach was tested on three tasks relevant to cancer treatment: finding kinases associated with cancer, finding gene mutations, and finding interactions between proteins associated with cancer. The paper describes the use of entity disambiguation, co-occurrence and linguistic processing in these tasks, and provides an overview of the methodology and results.

## 1 Introduction

As the amount of scientific literature continues to increase, extracting information from the literature is becoming both more important, and harder to achieve by traditional methods.

Text mining offers the opportunity to ask new questions, and to so immediately. Even in areas where there are existing hand-curated databases answering particular questions, text mining allows users to keep up-to-date with the latest

literature, examine a wider range of publications, and look for contextual factors that may have become of interest after the databases were constructed.

In the drug discovery process, experimentation followed by data analysis provides candidate biological targets. However, a critical step is then to use the literature to assess the applicability of the targets for therapeutic intervention (Fickett and Hayes, 2004). Although some therapeutic target knowledge is available from curated databases, text mining is important to ensure that knowledge is as current as possible.

In this paper we look at three case studies where different text mining strategies were employed. This suggests the need to use flexible tools that can embody more than one strategy.

## 2 Methodology

For each of the three case studies, the documents were prepared and searched using the same process.

The first step was to process Medline abstracts using bioalma's Text Detective software in order to recognize proteins/genes, diseases and compounds. Disambiguation of gene names is particularly challenging (Chen & Friedman, 2005). bioalma's Text Detective uses carefully designed rules and several lexicons of biological concepts to disambiguate and normalize mentions of biological entities to appropriate database references (Tamames, 2005). The abstracts were marked-up

using XML tags, and then passed to Linguamatics' I2E software for indexing.

The I2E software is an interactive information extraction engine which combines search with information extraction (Milward et al., 2005). In I2E, the user can start with a standard search for words within a document, then refine this to require the words to be in the same sentence, or in a particular linguistic pattern. The results are output as HTML tables, as Excel spreadsheets or in a format suitable for database entry. The system pre-indexes documents to allow fast querying. All linguistic processing is done prior to indexing, including "tokenisation" to split a string of characters into individual words, "sentence splitting" to recognise sentence ends, "tagging" to recognise parts of speech such as nouns or verbs, and "chunking" to group words into meaningful units according to their parts of speech

For each of the tasks there was no existing "gold standard" to evaluate against, so we employed several techniques:

1. For comparison of recall using different strategies we used existing hand-curated databases, even if incomplete, to get relative performance.
2. For estimating recall, we hand annotated a subset of documents related to specific subclasses of the task, e.g. a specific class of the disease, or a subset of the kinases.
3. For precision we used hand curation of results.

As well as extracting tables of relationships, the I2E system highlights the part of a document where the relationship (and its components) are found. This was found to be helpful for both hand curation and hand annotation.

In all cases we were primarily interested in "fact-based" precision and recall i.e. how many facts were extracted, and with what accuracy. This contrasts with "instance-based" precision and recall that count multiple instances of a single fact.

### 3 The Case Studies and Results

#### 3.1 Kinases associated with Cancer

The aim of this task was to find kinases associated with a particular cancer.

The first strategy we employed was to look for kinases and a specific cancer co-occurring within the same abstract (see e.g. Erhardt et al., 2006, for discussion of co-occurrence strategies). The fact-based precision score was impressive: 99% of the kinases suggested to be associated with the cancer were determined to be correct after hand curation. The instance-based precision (measuring how many kinases were correctly associated with the cancer in each document) was similarly high at 98%.

The use of co-occurrence within a sentence rather than within an abstract ruled out the one bad result, as did the use of a frequency cut-off (more than one co-occurrence). However, in both cases recall dropped by a third.

The use of further linguistic constraints, on top of co-occurrence within a sentence made no sense in this case, since this would potentially reduce recall further, and there was no possibility in gaining in precision over the 100% figure for co-occurrence within a sentence. Could this result have been predicted prior to evaluation? Some possible indicators were:

1. The relationship of *association* is bi-directional so there is no need for linguistics to determine if the kinase associates with the disease, or the disease with the kinase (contrast with e.g. *phosphorylation* where we do need to distinguish between "A phosphorylates B" and "B phosphorylates A")
2. *Associates* is a very general relationship, so there is no need to distinguish between different kinds of relationships between the two entities (note that this does mean that kinases involved in preventative vs. causative roles are not distinguished).
3. Most abstracts only talk about a single disease.

In future work we intend to explore this last point further, since it suggests we may want to use different strategies for abstracts and full-text articles, even for the same task. This might involve constraining co-occurrence within certain

sections of a document (or ruling out e.g. the Discussion section), or monitoring the number of disease occurrences within the co-occurrence window and adjusting the window accordingly.

### 3.2 Mutations of Kinases associated with Cancer

The aim of this study was to find information concerning mutations of kinases associated with cancer, and the kind of mutation. To pick out terms indicating a mutation, we used keywords that are good indicators of mutation, such as the following:

Missense, Deletion, Insertion, Heterozygosity, LOH, Homozygosity ...

These were augmented by phrases or patterns where the single word by itself was too noisy e.g.:

Chromosome\* abnormality  
Delete\* ... chromosome\* (within 5 words)

The second pattern above looks for delete and its morphological variants such as “deletion”, “deleted” etc. followed within five words by “chromosome” or “chromosomes”.

Three strategies were investigated for finding associations of the mutations with a particular kinase. The first looked for any mutation term in the same abstract as a kinase. The second looked for the mutation and the kinase within the same sentence. The third looked for a variety of constructions involving a mutation and a gene, including the constructions in bold font such as:

1. these studies showed a 31% (5 of 16) **LOH of MKK4** that is not associated with coding region mutations
2. overall, our data suggest that **mutations in CHEK2** may contribute to prostate cancer risk
3. we have previously reported **a truncating mutation (1100delC) in CHK2**

For all three strategies, we restricted the search to documents containing terms for a particular kind of cancer. In order to establish recall for finding mutations (but not for finding kinases), we took all documents containing a kinase associated with a particular cancer, and curated these

to find mutations. We used the system to highlight potential mutations terms to speed up this process for positive hits. The Precision and Recall figures for the three strategies were as follows:

Mutation and Kinase	Precision (%)	Recall (%)
in same abstract	70	92
in same sentence	80	60
in syntactic pattern	100	32

These results show a classic recall/precision tradeoff, so the choice of strategy is dependent upon how much hand curation the user is prepared to do.

### 3.3 Protein Interactions associated with Cancer

The final case study applied the tools to the problem of finding protein-protein interactions. NLP-based information extraction tools have been applied to this task for several years (e.g. Thomas et al., 2000, Humphreys et al., 2000, Friedman et al. 2001). NLP is generally used because:

1. the relationships are not symmetric: we need to distinguish between “Raf phosphorylates Mek” and “Raf was phosphorylated by Mek”.
2. There tend to be a large number of proteins in each abstract, and even in each sentence. Co-occurrence tends to pick up “sisterhood relationships” e.g. A-Raf and B-Raf both being Raf proteins.

In our case, the protein-protein relationships were constrained to appear in abstracts containing terms for the specific cancer of interest, so we were mixing precise linguistic patterns for the relationship, with co-occurrence in the abstract for the disease term.

Precise linguistic patterns were used to ensure high accuracy, and the correct direction of relationship. A set of patterns were used to cover constructions including verbal relationships, nominals, relative clauses and combinations of nominal and verbal relationships. Example constructions include:

1. **PACAP** transiently **increased c-fos gene expression**
2. **IRS-1** , an **activator** of **PI3-kinase**
3. ... of **uPA** which interacts with **uPAR**
4. the ability of **IL-6** to **induce** tyrosine **phosphorylation** of **STAT3**

Negative constructions such as “no phosphorylation”, “did not phosphorylate” or “failed to phosphorylate” were excluded.

This combination of precise patterns and the use of gene/protein achieved a precision of over 90%. Results were independently checked according to whether the relationship was correctly picked up from the text, and was biologically plausible. Regarding recall, initial comparison with a curated pathway database suggests that many new relationships were found, but also many relationships are missed. This needs further investigation, but is at least partly due to the abstracts (as opposed to the full text articles) not containing the relationships.

#### 4 Conclusion

This study has shown that text mining can provide very high quality results from large numbers of abstracts. The case studies also suggest that the best results are obtained by being flexible in the text mining strategy adopted, mixing language processing and co-occurrence according to the task, the document set and the required balance between recall and precision. The first study used co-occurrence within an abstract. The second showed how co-occurrence within an abstract, within a sentence, and using linguistic patterns are all appropriate strategies depending on the recall and precision required. The third showed the use of precise linguistic patterns.

In future work we would like to provide more concrete measures to help predict which strategy will be best in particular circumstances, and to investigate more automatic adaptation of strategies on a per document basis according to these measures.

#### Acknowledgements

We would like to thank Luis Gomez at bioalma and James Thomas at Linguamatics for their help

in running the experiments. We would also like to thank colleagues at Johnson & Johnson who helped with curation, and Paul Milligan at Linguamatics for his comments on previous drafts.

#### References

- Chen L, Liu H, Friedman C. 2005. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2):248-56.
- Fickett, J. and Hayes, W. 2004. Text Mining for Drug Discovery. *European Pharmaceutical Contractor*. Autumn 2004.
- Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17 Suppl. 1, S74-S82.
- Humphreys, K., Demetriou, G. and Geizauskas, R. 2000. Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structure. *Pac. Symp. Biocomp.*, pp. 502-513.
- Milward, D., Bjärelund, M., Hayes, W., Maxwell, M., Öberg, L., Tilford, N., Thomas, J., Hale, R., Knight, S. and Barnes, J. 2005. Ontology-based Interactive Information Extraction from Scientific Abstracts. *Comparative and Functional Genomics*, 6:67-71.
- Ramón A-A Erhardt, Reinhard Schneider, Christian Blaschke. Introduction and status of text-mining techniques applied to biomedical text. *Drug Discovery Today*, 2006
- Tamames, J. 2005. Text detective: a rule-based system for gene annotation in biomedical texts. *BMC Bioinformatics* 6 (Suppl 1:S10). Epub, May 24.
- Thomas, J., Milward, D., Ouzounis, C., Pulman, S. and Carroll, M. 2000. Automatic extraction of protein interaction from scientific abstracts. *Pac Symp Biocomp*, 541-552.