*Short Paper*

# ProSpecTome: a new tagged corpus for protein named entity recognition

Renata Kabiljo[1,*], Diana Stoycheva[2] and Adrian J Shepherd[1]

[1]School of Crystallography, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK [2]University of Heidelberg, Institute of Pharmacy and Molecular Biotechnology, Im Neuenheimer Feld 264, D-69120 Heidelberg, Germany

**ABSTRACT**

**Motivation:** This work grew out of our ongoing research on protein-protein interactions, in particular our desire to use one or more existing protein taggers for highlighting putative proteins in free text (as part of a larger interaction-mining system). Our primary motivation for developing a new evaluation corpus (i.e. a corpus designed *not* to be used for training purposes) was that we were unable to find an existing evaluation corpus that would enable us to carry out an independent comparative analysis of tool performance aligned to our application.

**Results:** We have produced a new protein-specific corpus – ProSpecTome – that is designed to facilitate the fair evaluation of protein taggers. It has been compiled by re-annotating a subset of the MEDLINE abstracts from the widely-used JNLPBA evaluation corpus. ProSpecTome combines a number of desirable features that are not shared by any other single corpus: it explicitly annotates names of proteins, but not non-coding genes; it incorporates two levels of specificity with regard to the category protein (with general references to proteins annotated separately from the names of individual proteins and protein families); the annotation guidelines used to produce the corpus together with the degree of inter-annotator agreement associated with its production are explicitly documented; and it is provided in a convenient XML format (with accompanying stylesheet so that the corpus can be easily displayed in a web browser).

**Availability:** The ProSpecTome corpus and associated annotation guidelines are freely available and can be downloaded from http://textmining.cryst.bbk.ac.uk/ProSpecTome/.

**Contact:** r.kabiljo@mail.cryst.bbk.ac.uk

## 1 INTRODUCTION

Protein Named Entity Recognition (NER) is of vital importance to a number of biomedical text-mining tasks such as the extraction of functional annotations and information about protein-protein interactions from the literature. There are a number of freely available protein taggers (i.e. tools that aim to automatically mark up protein names in natural language texts), and it is clearly desirable that we should be able to independently and reliably evaluate the performance of these tools.

Biomedical corpora in which protein names have been manually annotated play a vital role in the development and subsequent evaluation of protein taggers. Most protein taggers have been trained and/or tested using one or more of the following corpora: GENIA (Kim et al, 2003), GENETAG (Tanabe et al., 2005) and Yapex (Franzén et al., 2002). However, it is clear from a comparison of these corpora that there is considerable disagreement about what entities should be annotated as proteins. This diversity partly reflects the inherent complexity of the domain, but also the range of possible applications these corpora are designed to support. Moreover, even when two corpora agree that a given entity is a protein, there is often disagreement about where the boundaries (i.e. start-point and end-point) of the protein name are located within the text. Consequently, the performance of given tool is likely to vary considerably depending on which corpus it was trained on and, crucially, which corpus is used in its evaluation.

Here we introduce a new corpus – the ProSpecTome corpus – annotated exclusively with protein names. In designing ProSpecTome, we have re-annotated a subset of the JNLPBA (Joint Workshop on Natural Language Processing in Biomedicine and its Applications) evaluation corpus (Kim et al., 2004) described below in section 2.1. We have chosen to re-annotate part of an existing corpus, rather than start afresh with a new set of texts, for two main reasons.

Firstly, as its name suggests, the JNLPBA evaluation corpus has been deliberately "reserved" for evaluation purposes. Assuming the developers of protein taggers respect this intention, it is reasonable to assume that taggers will not have been trained on the data in this corpus. Consequently this set of texts is a natural choice when the aim is to develop a corpus for performing a fair evaluation of multiple taggers, since clearly we cannot perform a fair evaluation by testing a tool on the same data that was used to train it.

Secondly, we believe that having two contrasting sets of annotations for the same set of texts is valuable in its own right. By comparing the performance of a protein tagger on both the JNLPBA evaluation corpus and ProSpecTome, we can quantify effects that are attributable to the choice of annotation conventions in isolation from those attributable to differences in the use of language within the texts themselves.

In designing ProSpecTome, we have aimed to adopt good practices relevant to the development of biomedical corpora. Both Mani et al. (2005) and Cohen et al. (2005) stress the desirability of providing explicit annotation guidelines and assessments of inter-annotator agreement. These topics are addressed below in sections 3.2 and 3.3 respectively.

## 2 BACKGROUND

### 2.1 Existing corpora

According to a 2005 survey of corpus usage rates (Cohen et al., 2005), GENIA, GENETAG and Yapex are the three most widely-used biomedical corpora. Most of the best performing, freely available protein taggers have been trained and/or tested on one or more of these corpora.

By far the most widely-used corpus identified in the survey of Cohen et al. is the GENIA corpus. GENIA version 3.0x is a corpus of 2,000 abstracts restricted to the sub-domain of human blood cell transcription factors. In its original version, abstracts are annotated using 38 classes of biomedical entity defined in the GENIA ontology, and both nested and overlapping entities are permitted. However, for the JNLPBA, a simpler version of GENIA was provided as training data. This version has annotations for only five classes of named entity (DNA, RNA, protein, cell line and cell type) and contains neither nested nor overlapping entities. An additional GENIA test corpus comprising 404 abstracts annotated using the same, simplified approach was released for evaluation purposes in conjunction with the bio-entity recognition tasks at JNLPBA 2004. It is this latter corpus – the JNLPBA evaluation corpus – that has been partially re-annotated in the ProSpecTome corpus presented here.

Lagging some way behind GENIA in terms of popularity amongst the developers of biomedical text-mining systems are the GENETAG and Yapex corpora. The GENETAG corpus consists of 20,000 MEDLINE sentences annotated with genes and proteins. 15,000 GENETAG sentences were used for the BioCreAtIvE Task 1A Competition (Yeh et al., 2005). The Yapex corpus of protein names contains 200 abstracts, a subset of which was selected at random from the GENIA corpus. Yapex combines data from the same sub-domain as GENIA (i.e. human blood cell transcription factors) with abstracts on protein binding in humans.

Of these three corpora, only GENETAG provides detailed documentation of its annotation guidelines (Cohen et al., 2005).

### 2.2 Comparison of GENIA and Yapex annotations

The overlapping set of abstracts common to GENIA v3.0x and Yapex provides a convenient basis for comparing the degree of agreement between these two corpora. For 50 abstracts common to the simplified version of GENIA and Yapex, we have calculated the number of entities that have been annotated in the same way (identical boundaries) or similar way (overlapping boundaries) in both corpora. Taking only entities from the GENIA protein class, there are 695 protein annotations in GENIA compared with 769 in Yapex. Only 64% of the GENIA entities are identical to a Yapex entity, and only 58% of Yapex entities are identical to a GENIA entity. 81% of the GENIA entities overlap a Yapex entity, whereas 76% of Yapex entities overlap a GENIA entity.

When entities from the GENIA DNA and RNA classes are added to those from the protein class, the number of entities rises to 939. Only 50% of these GENIA entities are identical to a Yapex entity, and only 60% of Yapex entities are identical to a GENIA entity. 74% of the GENIA entities overlap a Yapex entity, whereas 95% of Yapex entities overlap a GENIA entity.

These differences in the annotations between the two corpora are attributable to a number of factors. Firstly, there is the question as to which entities get annotated as proteins. Many entities annotated as proteins in Yapex are annotated as non-protein entities in GENIA. This situation is somewhat ameliorated by treating entities from the GENIA DNA and RNA classes as protein annotations, but this has the effect of adding entities that do not code for proteins (e.g. introns and exons) to the set of GENIA "protein" annotations. Understandably, these latter entities are not annotated as proteins in Yapex.

Another factor contributing to the lack of agreement between GENIA and Yapex is the specificity with which entities are defined. The Yapex annotators "define a protein name semantically as something that denotes a single biological entity composed of one or more amino acid chains" (Franzén et al., 2002), whereas protein fragments, protein families and even very general references to proteins (e.g. the words "protein", "hormone" and "enzyme") are additionally annotated in GENIA.

Even when the two corpora agree that the same entity should be annotated, the boundaries of the annotation often differ. This reflects a different approach to handling cases where, for example, the protein name is preceded by a modifier such as "human" or followed by a modifier such as "homodimer". Examples of the contrasting approaches to protein annotation in GENIA (here incorporating entities from the DNA and RNA classes) and Yapex are given in Table 1.

**Table 1.** Examples of the same text that has been annotated differently in GENIA and Yapex. For clarity, the annotated entities have been underlined and bolded.

| GENIA annotation | Corresponding Yapex annotation |
|---|---|
| …transcription of the **nuclear proto-oncogenes c-fos** and **c-jun** | …transcription of the nuclear proto-oncogenes **c-fos** and **c-jun** |
| We compared the effects of the **deactivating cytokine interleukin 10** … | We compared the effects of the deactivating cytokine **interleukin 10** … |
| …release of chemotactic and **inflammatory cytokines** | …release of chemotactic and inflammatory cytokines |
| **c-Rel homodimer** | **c-Rel** homodimer |

## 3 RESULTS

### 3.1 Corpus overview

The ProSpecTome corpus consists of 243 articles randomly chosen from the JNLPBA evaluation corpus and re-annotated according to our own guidelines. The annotations

in ProSpecTome differ from those in the JNLPBA corpus in two key respects.

Firstly, the annotations in ProSpecTome are exclusively for protein names. Many protein names that are tagged as other entities in the JNLPBA corpus (notably as DNA or RNA entities) are tagged as proteins in ProSpecTome.

We believe this approach is a reasonable compromise for many biomedical text-mining applications. A clear distinction between protein and gene names is often impossible; a gene and the protein for which it codes frequently have the same name, and authors often use such names in an ambiguous manner that makes it impossible to assign them to separate categories. Even when it is clear from the context in which the name is used that an author is referring to a gene rather than its product, it is not obvious that the person using a protein tagger would wish such occurrences of a gene/protein name to remain untagged. On the other hand, we believe it is appropriate to exclude non-coding entities such as promoters and enhancers from our protein category.

A second important difference between the annotations in ProSpecTome and the JNLPBA corpus is that each ProSpecTome annotation has a specificity assigned to it in the form of an XML attribute ("specific"). Annotations assigned a specificity of 0 represent very general references to proteins, whereas annotations with specificity 1 refer to specific protein entities (individual proteins, protein families, complexes, etc.). Of the 4,770 protein annotations in ProSpectTome., 936 have specificity 0 and 3,834 have specificity 1. The advantage of this approach is that the corpus-user can decide whether a broad or narrow definition of a protein entity is most appropriate for the intended application.

We believe the approach to annotating protein names outlined above provides a potentially useful alternative to that provided by the JNLPBA evaluation corpus. Taking the two corpora together, several different perspectives on the naming of genes and proteins are possible.

### 3.2 ProSpecTome annotations

The guidelines used in the annotation of the ProSpecTome corpus are in the form of a set of explicit rules that specify: what names should be annotated as proteins; what specificity level (0 or 1) should be assigned to a given protein name; and how the boundaries of a given protein name should be determined. The complete set of annotation guidelines is available from the ProSpecTome website; here we briefly summarize some of the key issues.

In ProSpecTome, protein names are annotated irrespective of the context in which they appear, for example "polymerase" is annotated in the phrase "polymerase chain reaction". "IL-12 p40" is annotated in the phrase "IL-12 p40 promoter" because the gene IL-12 p40 codes for a protein, whereas the promoter itself does not. Similarly, other types of non-coding DNA and RNA (e.g. introns and exons) are not annotated. Where a protein name is followed by its abbreviation, both are annotated separately.

With respect to specificity, the names of individual proteins, protein families, parts of proteins and protein complexes are assigned a specificity of 1. Other, more general references to proteins are assigned a specificity of 0. Examples of protein names with different specificities are given in Table 2.

The majority of the ProSpecTome annotation rules are designed to ensure that the boundaries of names are annotated consistently. For example, there are rules for deciding whether relevant words that come before (e.g. "human") or after (e.g. "factor", or "complex") a protein identifier are incorporated within the annotation, and rules for handling phrases containing a conjunction (e.g. "NF-kappaB factor p50 or p52").

**Table 2.** Examples of ProSpecTome annotations with different specificities

| Specificity | Examples of terms |
|---|---|
| 0 | "Immunoglobulin", "cytokine", "52-kDa protein", "hormone receptor" |
| 1 | "Stress-activated protein kinases", "ras family of proteins", "Bcl-2 family members", "Calcineurin", "IL-10", "STAT-5" |

Examples of how the annotations in ProSpecTome differ from those in the JNLPBA corpus are given in Table 3.

**Table 3.** Examples of the same text that has been annotated differently in the JNLPBA evaluation corpus and ProSpecTome. For clarity, the annotated entities have been underlined and bolded.

| JNLPBA corpus annotation | Corresponding ProSpecTome annotation |
|---|---|
| <RNA>**estrogen receptor (ER) transcripts**</RNA> | <PROTEIN specific="1">**estrogen receptor** </PROTEIN> (<PROTEIN specific="1">**ER** </PROTEIN>) transcripts |
| Partial sequences from <DNA>**exons 1-8**</DNA> were nearly identical to the published sequence of the <RNA>**human ER mRNA.**</RNA> | Partial sequences from exons 1-8 were nearly identical to the published sequence of the <PROTEIN specific="1">**human ER**</PROTEIN> mRNA. |
| <cell_line>**Tax-expressing JPX-9 cells**</cell_line> | <PROTEIN specific="1">**Tax** </PROTEIN>-expressing JPX-9 cells |

### 3.3 Inter-annotator agreement

Given the complexity of biomedical concepts and the language used to describe them, it is unreasonable to expect different annotators to produce identical annotations using the same set of guidelines even when the annotators are highly experienced and the guidelines well designed. For the performance of a machine tagger on a given corpus to be placed in a proper perspective, it is essential that we know the extent to which the human annotators of that corpus

agreed about the annotations they made. Unfortunately, this information is not available for most biomedical corpora.

Inter-annotator agreement for biomedical NER is typically assessed using the well-known F-measure, with credit given only for names that are annotated identically by both the annotators being assessed. Mani et al. (2005) report F-scores in the range 0.65 to 0.89 for protein NER inter-annotator agreement. Similar upper levels of performance for other biomedical NER tasks are reported elsewhere (see, for example, Dingare et al., 2005).

Inter-annotator agreement for the ProSpecTome corpus was calculated on a subset of 43 abstracts. Prior to this, 200 abstracts were annotated by one annotator and these annotations were then inspected and corrected by a second annotator. The two annotators have a background in biology at both undergraduate and post-graduate levels. Both annotators then jointly refined the annotation guidelines and modified the annotations in the 200 abstracts accordingly. Finally, the remaining 43 abstracts were annotated independently by both annotators.

The inter-annotator agreement was calculated by "scoring" the annotations of the second annotator against the "gold standard" of the first annotator. The obtained F-score was 0.89, when annotations of both specificities are taken. For annotations assigned specificity 1 only, the F-score was slightly higher at 0.91. For comparison, when credit is given for overlapping annotations that do not have boundaries that match perfectly, the F-score rises to 0.98 (both specificities).

### 3.4 Performance of tools on ProSpecTome

We have performed a comparative evaluation of three protein taggers – the version of ABNER (Settles, 2005) trained on the BioCreAtIvE corpus, Gapscore (Chang et al., 2004), and NLProt (Mika and Rost, 2004) – using the ProSpecTome corpus and the same set of 243 abstracts from the JNLPBA evaluation corpus. With JNLPBA, annotations from classes DNA and RNA were merged with those from class protein on the grounds that tools performed better with this combination than with protein annotations alone. Tools were evaluated in both "strict mode" (where perfect boundary matches are required) and "sloppy mode" (where credit is given for matching part of an annotated protein). The results are summarized in Table 4.

**Table 4.** Performance of tools on ProSpecTome and on the same subset of 243 abstracts from the JNLPBA evaluation corpus. The numbers represent F-score in sloppy / strict mode.

| Tool | ProSpecTome specificity 1 | ProSpecTome specificity 0 | Subset of JNLPBA |
|---|---|---|---|
| ABNER | 0.85 / 0.62 | 0.80 / 0.56 | 0.75 / 0.63 |
| Gapscore | 0.81 / 0.53 | 0.75 / 0.48 | 0.68 / 0.37 |
| NLProt | 0.81 / 0.60 | 0.73 / 0.54 | 0.70 / 0.45 |

Two points stand out from the results in Table 4: all tools get higher scores on ProSpecTome than on the subset of JNLPBA; and all tools score better on specificity level 1 annotations than on those with specificity 0. A full explanation of these trends awaits further analysis, but – given that ProSpecTome annotations at specificity level 1 were designed to map closely to our intended application in the domain of protein-protein interactions – it is encouraging that all three tools perform best on this set of annotations.

## 4 CONCLUSIONS

In this paper, we have presented the ProSpecTome corpus, a new tagged corpus for protein named entity recognition. ProSpecTome has been specifically designed to facilitate the fair cross-evaluation of protein taggers. ProSpecTome provides a re-annotation of a subset of the widely-used JNLPBA evaluation corpus using significantly different annotation criteria. Using both corpora to evaluate the performance of a protein tagger, it is possible to undertake a more detailed analysis of its performance.

Finally, the usefulness of ProSpecTome is enhanced by its supporting documentation, notably the published set of explicit annotation guidelines available on the ProSpecTome website, and the assessment of inter-annotator agreement reported above.

## ACKNOWLEDGEMENTS

We would like to thank the UK ORSAS scheme for providing the financial support that has made this work possible.

## REFERENCES

Chang,J.T., Schutze,H. and Altman, R.B. (2004) GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*, **20** (2):216-225.

Cohen,K.B., Fox,L., Ogren,P.V. and Hunter,L. (2005) Corpus design for biomedical natural language processing. In: *Proc. ACL-ISMB Workshop*, 38-45.

Dingare,S., Nissim,M., Finkel,J., Manning,C. and Grover,C. (2005) A system for identifying named entities in biomedical text: how results from two evaluations reflect on both the system and the evaluations. *Comp. Funct. Genomics*, **6**, 77-85.

Franzén,K., Eriksson,G., Olsson,F., Asker,L., Liden,P. and Coster,J. (2002) Protein names and how to find them. *Int J Med Inf*, **67**, 49-61.

Kim,J.-D., Ohta,T., Teteisi,Y. and Tsujii,J. (2003) GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**(suppl. 1). i180-i182.

Kim,J.-D., Ohta,T., Tsuruoka,Y., Tateisi,.Y and Collier,N. (2004) Introduction to the Bio-Entity Recognition Task at JNLPBA. In: *Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA 2004)*, 70-75.

Mani,I., Hu,Z., Jang,S.B., Samuel,K., Krause,M., Phillips,J. and Wu,C.H. (2005) Protein name tagging guidelines: lessons learned, *Comp. Funct. Genomics*, **6**, 72-76.

Mika,S. and Rost,B. (2004) Protein names precisely peeled off free text. *Bioinformatics*, **20** (suppl 1):I241-247.

Settles,B. (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, **21** (14):3191-3192.

Tanabe,L., Xie,N., Thom,L.H., Matten,W. Wilbur,W.J. (2005) GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, **6**(Suppl 1):S3.

Yeh,A., Morgan,A., Colosimo,M. and Hirschman,L. (2005) BioCreAtIvE Task 1A: gene mention finding evaluation. *BMC Bioinformatics* **6**(Suppl 1):S2