Protein Name Tagging in the Immunological Domain

Renata Kabiljo
School of Crystallography Birkbeck
University of London
Malet Street, London WC1E 7HX
r.kabiljo@mail.cryst.bbk.ac.uk

Adrian J Shepherd
School of Crystallography Birkbeck
University of London
Malet Street, London WC1E 7HX
a.shepherd@mail.cryst.bbk.ac.uk

Abstract

The research described in this paper addresses the following question: How well do generic protein/gene name taggers perform when they are applied to full-text articles from the sub-domain of immunology (a subdomain with its own distinctive protein nomenclature)? To answer this question we have created a new corpus – ImmunoTome – consisting of ten full-text immunological articles in which the names of proteins have been manually annotated. Our results show that a single tagger – ABNER trained on the BioCreAtivE corpus – performs significantly better than the other taggers we evaluated when applied to ImmunoTome. ImmunoTome is available from immunominer.cryst.bbk.ac.uk/tome.html.

1 Introduction

Large amounts of useful immunological data are to be found exclusively in full-text journal articles. Much of these data concern the key proteins involved in the immune response, notably antibodies, antigenic proteins, and cytokines. For our research as members of the ImmunoGrid Consortium[1], our ultimate aim is to devise methods capable of automatically extracting this information from the literature. As a crucial first step, we need to identify the protein entities themselves.

Some of the key protein entities of the immune system (or the genes that encode them) have their own, distinctive nomenclature, notably the CD nomenclature for leukocyte surface molecules (e.g. *CD4*, *CD8*) and the HLA nomenclature for the human leukocyte antigen system (e.g. *B40*, *DR14*, *Dw25*). Other important classes of immune system proteins have names that typically start with three upper-case letters, e.g. *TCR* (T cell receptors). Names containing a mixture of upper-case letters and digits are likely to be particularly easy for taggers to identify, as relatively few non-protein words have this form.

There are a number of freely-available named-entity taggers for proteins and other biomedical entities. These taggers have typically been trained on one or more of three widely-used biomedical corpora – GENIA, BioCreAtivE and Yapex. All three corpora consist of manually annotated abstracts (or sentences from abstracts) taken almost exclusively from non-immunological papers.

This raises an important question: How well do taggers trained on such corpora perform when they are applied to a specific sub-domain such as immunology characterized by its own distinctive protein nomenclature? This question is the starting point for the research described in this paper. Here we compare the performance of four freely-available taggers designed to annotate the names of proteins and other biomedical entities in natural language texts. The four taggers are: LingPipe[2], trained on GENIA (Kim *et al.*, 2003); NLProt (Mika & Rost, 2004), trained on Yapex (Franzén *et al.*, 2002); Gapscore (Chang *et al.*, 2004), a rule-based tagger; and ABNER (Settles, 2005). ABNER can be run in two modes: one trained on a simplified version of GENIA known as the JNLPBA corpus (Kim *et al.*, 2004), and the other trained on BioCreAtivE (Yeh *et al.*, 2005).

2 Methods

2.1 The ImmunoTome corpus

In order to assess the performance of generic protein taggers on full-text immunological articles, we created a new corpus – ImmunoTome. ImmunoTome consists of ten full-text articles from the *Journal of Immunology*, each containing at least one reference to the proteins *CD4* or *CD8*. (The latter criterion was adopted because we are particularly interested in the molecular aspects of the adaptive immune system.)

[1] www.immunogrid.eu

[2] www.alias-i.com/lingpipe/index.html

In ImmunoTome, protein names are annotated regardless of their context. For example, in the phrase "CD4+CD8- cells" both "CD4" and "CD8" are annotated as proteins. We have annotated both the names of proteins and the names of the genes that code for those proteins, but not non-coding entities such as promoters and enhancers. We believe this approach is a reasonable compromise for many biomedical text-mining applications, as a clear distinction between protein and gene names is often impossible.

In designing ImmunoTome, we have aimed to adopt good practices relevant to the development of biomedical corpora, including the provision of explicit annotation guidelines. ImmunoTome was created by two annotators with prior experience of developing the ProSpecTome corpus (Kabiljo *et al.*, 2007). Inter-annotator agreement was calculated using a single article after an iterative process of guideline and annotation refinement using the other nine. When the annotations of the second annotator were scored against the annotations of the first, an F-score of 82% was achieved. When credit was given for overlapping annotations, this rose to 96%.

Note that, although of sufficient size for evaluation purposes, ImmunoTome is much smaller than standard training corpora. It is therefore not large enough to facilitate the retraining of existing tagging tools.

2.2 Tagger evaluation

To calculate approximate upper and lower bounds on the performance of different taggers, we assessed their performance using both "strict" and "sloppy" matching criteria. When strict criteria are applied, a tagger is required to match a given protein name exactly to score a "hit". When sloppy criteria are applied, the tagger scores a "hit" provided part of the protein name is matched.

The extent to which exact matching is required in practice is application-dependent. However, in terms of the fair evaluation of tagger performance, the use of strict matching criteria has a clear disadvantage; the performance of a tool will vary significantly depending on essentially arbitrary choices made by the annotators of the evaluation corpus (e.g. is the word "mouse" part of the protein name in the phrase "mouse oxytocin"?). With sloppy matching criteria, on the other hand, there is a risk that a tool will gain credit even when it has missed the core part of a protein name (e.g. if it exclusively annotated either the word "activated" or "protein" in the phrase "activated ras-1 protein").

We investigated a random set of 100 tagged protein names that count as hits with sloppy criteria, but as misses with strict criteria. In every case the core of the protein name was contained within the annotation. In 18 cases an erroneous word had been incorporated (e.g. the word "namely" in "namely IFN-gamma"), in 21 cases the discrepancies were associated with the conjunction "and" (e.g. "CD4 and CD8 coreceptors" is annotated as a single name in ImmunoTome, but as two proteins by the tagger), and the remainder involved more-or-less legitimate extensions to, or contractions of, the name as annotated in ImmunoTome (e.g. "Ag antivenin" instead of "antivenin").

3 Results

3.1 Comparative performance of taggers

The performance of our chosen taggers on four corpora is given in table 1. These results show that ABNER is the best-performing tagger on all corpora, with the BioCreAtivE version of ABNER registering the best scores on ImmunoTome using both strict and sloppy matching criteria. All the other taggers show a significant drop in performance when evaluated on ImmunoTome.

	Y	J	P	I
Sloppy matching criteria				
ABNER (B)	80.3	76.0	85.3	78.3
ABNER (G)	73.9	84.1	80.1	69.5
NLProt	N/A	70.8	81.2	66.1
LingPipe	65.3	79.1	67.4	53.7
Gapscore	80.5	68.6	81.3	56.6
Strict matching criteria				
ABNER (B)	54.2	60.8	59.4	54.0
ABNER (G)	48.4	67.9	62.0	47.8
NLProt	N/A	45.8	59.7	43.8
LingPipe	43.4	62.8	47.0	33.6
Gapscore	57.4	38.3	52.9	30.7

Table 1. The F-scores produced by five taggers when applied to four corpora. Abbreviations are as follows: Y = Yapex; J = JNLPBA evaluation corpus; P = ProSpecTome (Kabiljo *et al.*, 2007); I = ImmunoTome; B = BioCreAtivE; G = GENIA. As NLProt was trained on the Yapex corpus, no fair test score can be provided for this combination.

ImmunoTome differs from the other corpora in two important respects: it contains texts from a distinctive sub-domain; and it comprises full-text articles rather than abstracts. In order to shed light on the relative impact of these differences, we independently evaluated the taggers that were not trained on GENIA using the subset of GENIA abstracts containing the annotations *CD4* or *CD8* (86 abstracts from a total of 2,000). The results are shown in table 2.

	Full GENIA corpus	Subset of GENIA
Sloppy matching criteria		
ABNER (B)	79.7	83.5
NLProt	74.2	77.1
Gapscore	73.8	77.0
Strict matching criteria		
ABNER (B)	64.3	66.8
NLProt	49.7	54.8
Gapscore	40.8	48.1

Table 2. The F-scores of three taggers – none of which were trained using the GENIA corpus – applied to GENIA and to an immunological subset of GENIA.

These results suggest that taggers find it easier to correctly identify protein names from the immunological sub-domain than from general biomedical texts. To explore possible reasons for this, we analyzed the most frequently-occurring protein names in three corpora (table 3). Note that the top ten protein names in ImmunoTome account for 43% of the total annotations in that corpus – much higher than the equivalent figures for Yapex and GENIA (6% and 9% respectively). This is to be expected given the repetitious use of protein names in full-text articles.

Yapex	GENIA	ImmunoTome
NF-kappa B (28)	NF-kappa B (862)	CD4 (518)
Tat (27)	NF-kappaB (542)	CD8 (348)
CD4 (26)	IL-2 (535)	TCR (156)
p53 (26)	transcription factors (332)	TRX1 (143)
NF-kappaB (23)	AP-1 (322)	TCR-αß (74)
GM-CSF (22)	IL-4 (314)	CD40 (59)
IL-2 (22)	transcription factor (283)	TNF (58)
SMN (22)	TNF-α (245)	IFN-γ (53)
IL-6 (22)	IFN- γ (227)	CD40L (52)
SUMO-1 (21)	cytokines (200)	2C TCR (51)

Table 3. The top ten occurring protein names in three corpora. The number of occurrences of each name is recorded in parentheses. Different forms of the same name (e.g. “NF-kappa B” and “NF-kappaB”) are recorded separately.

From table 3 it appears that names of forms that are likely to prove comparatively easy for taggers to identify are more prevalent in ImmunoTome. In particular, names made from a combination of upper-case letters and digits account for six out of ten names on the ImmunoTome list, compared with four for Yapex and three for GENIA. On the other hand, names in lower case (easily confused with general vocabulary) or title case (easily confused with generic proper names) are more prevalent in Yapex and GENIA. (Note that the appearance of general references to proteins – e.g. “cytokines” – exclusively on the GENIA list is attributable to the annotation guidelines associated with that corpus.)

3.2 Information content of ImmunoTome

The distribution of protein names across the different sections of the full-text articles in ImmunoTome are summarized in table 4. Of the total number of protein names, less than 5% are found in the abstracts. When the same calculation is performed for distinct protein names, less than 10% are found in the abstracts. Unsurprisingly, the number of protein names uniquely found in the abstracts of ImmunoTome is very low (though, perhaps surprisingly, non-zero).

	TA	I	M	RD
total words	2262	6331	7418	33786
total annotated	171	484	397	2437
total distinct	84	212	275	518
total distinct & exclusive	14	111	189	359
annotated / words (%)	7.6	7.6	5.4	7.2
distinct / words (%)	3.7	3.4	3.7	1.5
exclusive / words (%)	0.6	1.8	2.5	1.1

Table 4. The information content of the ImmunoTome corpus broken down by section. Abbreviations are as follows: TA = Title + Abstract; I = Introduction; M = Materials and Methods; RD = Results + Discussion. The “distinct & exclusive” total records the number of distinct protein names that are exclusively found within a given section.

With respect to the density of information, the results are less clear-cut. Ultimately it makes sense to select the most relevant sections for a given application, and relevance is not something that can be assessed by a simple analysis of name density (Shah *et al.*, 2003).

Whatever the application, it is certainly worth taking into account the variable performance of protein taggers on the different sections of full-text articles. This is summarized for ImmunoTome in table 5.

There are two notable features of these results. Firstly, all taggers except Gapscore perform best on the Introduction section, in spite of the fact that all the taggers except Gapscore were trained on abstracts. This is a surprising result and one that warrants further investigation.

	TA	I	M	RD
ABNER (B)	79.4	83.1	72.4	78.2
ABNER (G)	74.2	75.3	63.2	69.2
NLProt	67.4	75.9	48.0	67.4
LingPipe	60.4	61.1	52.1	52.1
Gapscore	63.2	54.6	46.4	46.4

Table 5. The F-scores of five taggers on different sections within the ImmunoTome corpus evaluated using sloppy matching criteria. Abbreviations are the same as for table 4.

Secondly, all tools perform worst on the Materials and Methods section. A common problem here is the relatively high density of proper names such as *Pharmingen* and *Sweden.*

4 Conclusion

That ABNER (BioCreAtivE) proves to be the best single tagger when applied to ImmunoTome reinforces the conclusion we reached elsewhere (Kabiljo *et al.*, 2007). It may be significant that this version of ABNER did much better than the version trained on GENIA. Further investigation is needed to decide whether this is attributable to the content of the BioCreAtivE corpus, to its annotation guidelines, or to other factors.

This is, we believe, an important finding. The construction of new training corpora is very time consuming, hence it is highly unlikely that multiple training corpora focusing on specific biomedical sub-domains will become available in the foreseeable future. In these circumstances, researchers wishing to perform named entity recognition in a biomedical sub–domain have little option but to use one or more existing taggers. Our results show that, at least for the sub-domain of immunology, this does not lead to a large drop in performance – provided that the chosen tagger is ABNER (BioCreAtivE).

Our results also show that taggers have particular problems when annotating the Materials and Methods sections in ImmunoTome. This is likely to be true more generally, and suggests that for some applications it is sensible to exclude Materials and Methods sections altogether.

Finally, using multiple corpora to evaluate the performance of different protein taggers potentially gives us deeper insights into their relative performance. From this perspective, ImmunoTome complements existing corpora, and will offer a new dimension to future analyses. In this role, the usefulness of ImmunoTome is enhanced by the provision of explicit annotation guidelines, and the assessment of inter-annotator agreement reported above.

Acknowledgements

The research of R.K. is supported by European Commission under FP6-2004-IST-4 contract no. 028069 (the ImmunoGrid project) and the UK ORSAS scheme. We would like to thank Diana Stoycheva, University of Heidelberg, for her help in creating the ImmunoTome corpus.

References

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

Kristofer Franzén, Gunnar Eriksson, Fredrik Olsson, Lars Asker, Per Lidén and Joakim Cöster. 2002. Protein names and how to find them. *International Journal of Medical Informatics*, 67(1-3):49-61.

Renata Kabiljo, Diana Stoycheva and Adrian J. Shepherd. 2007. ProSpecTome: a new tagged corpus for protein named entity recognition. *Proceedings of the Annual Meeting of the ISMB BioLINK Special Interest Group on Text Data Mining*, Vienna, 19 July 2007, 24-27.

Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi and Jun'-ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl. 1):i180-i182.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Teteisi and Nigel Collier. 2004. Introduction to the Bio-Entity Recognition Task at JNLPBA. *Proceedings of JNLPBA 2004*, 70-75.

Sven Mika and Burkhard Rost. 2004. Protein names precisely peeled off free text. *Nucleic Acids Research*, 32(Web server issue): W634-W637

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

Parantu K. Shah, Carolina Perez-Iratxeta, Peer Bork and Miguel A. Andrade. 2003. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics*, 4:20.

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