# Learning the Species of Biomedical Named Entities from Annotated Corpora

**Xinglong Wang and Claire Grover**

School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW, UK
{xwang,grover}@inf.ed.ac.uk

**Abstract**

In biomedical articles, terms with the same surface forms are often used to refer to different entities across a number of model organisms, in which case determining the species becomes crucial to term identification systems that ground terms to specific database identifiers. This paper describes a rule-based system that extracts 'species indicating words', such as *human* or *murine*, which can be used to decide the species of the nearby entity terms, and a machine-learning species disambiguation system that was developed on manually species-annotated corpora. Performance of both systems were evaluated on gold-standard datasets, where the machine-learning system yielded better overall results.

## 1. Introduction

Information Extraction (IE) technologies such as Named Entity Recognition (NER), Term Identification (TI) and Relation Extraction (RE) have been shown to help reduce the laborious work involved in curating the vast amount of biomedical research papers (Karamanis et al., 2007; Alex et al., 2008; Wang and Matthews, 2008). In a typical curation process for protein-protein interactions (PPI), an IE system would first recognise the protein mentions (i.e., NER) and then assign unique database identifiers to them (i.e., TI), and finally input the pairs of identifiers of the interacting proteins into a database of PPIs. As an intermediate module that disambiguates the mentions and normalises them to database identifiers, TI is essential because strings of text with the same surface form can often be used to refer to different entities. As noted in our previous work and elsewhere, determining the correct species for the protein mentions is one of the most important steps towards TI (Krauthammer and Nenadic, 2004; Chen et al., 2005; Krallinger et al., 2007; Wang, 2007).

> We found that *Plk1* can phosphorylate *Nek2* in vitro and interacts with *Nek2* in vivo.

For example, searching for the string *plk1* in the above sentence in RefSeq[1] resulted in 98 hits, whereas when a species (e.g., *mouse (Mus musculus)* was added to the query, we were able to narrow down the number of choices to two.

This paper reports on our efforts in building species-annotated corpora, in which species tags were manually assigned to occurrences of several types of biomedical entities, including proteins, genes and mRNAs, and in developing an automatic species tagger using rule-based and machine-learning approaches based on this resource.

The paper is organised as follows: Section 2 describes some of the related work. Section 3 reports on the process of annotating the species corpora and on calculating the inter-annotator-agreement. Section 4 describes a rule-based approach to detect species words (e.g., *mouse*), utilising various specialised lexicons. A number of rule-based and machine-learning based species tagging systems that utilise species words as one of the most important type of features are presented in Section 5. This section also discusses the experimental results and findings. We finally conclude and propose future work in Section 6.

## 2. Related Work

The BioCreAtIvE I & II evaluation workshops (Hirschman et al., 2005; Hirschman et al., 2007) provided forums and gold-standard datasets for the community on evaluating biomedical IE systems such as NER (Yeh et al., 2005; Wilbur et al., 2007), TI (Hirschman et al., 2004; Morgan and Hirschman, 2007), and RE (Blaschke et al., 2005; Krallinger et al., 2007). A number of tasks in the recent BioCreAtIvE II workshop have addressed the importance of species disambiguation. For example, the protein interaction pairs subtask (IPS) (Krallinger et al., 2007) resembled the work-flow of manual curation of PPIs,[2] and required identification of interacting proteins across many model organisms. The best result for this task was fairly low at 28.85% $F1$, and a number of participants have reported (e.g., Grover et al., 2007) that species ambiguity posed one of the biggest challenges. An analysis of the training dataset of IPS revealed that the interacting proteins in this corpus belong to over 60 species, and only $56.27\%$ of them are *human*.

Also, Chen et al. (2005) collected gene information from 21 organisms and quantified naming ambiguities within species, cross species, with English words and with medical terms. Their study showed that the intra-species ambiguity in gene names was negligible at $0.02\%$, whereas cross-species ambiguity was high at $14.2\%$. It suggests that resolving species ambiguity would be an effective step towards gene name identification. On the other hand, as Ananiadou et al. (2004) suggested, existing text processing resources typically lack information that can support disambiguation of terms, and such resources do not address ambiguities related to finer biological classification, such as species information.

[1] `http://www.ncbi.nlm.nih.gov/RefSeq/`. The searches were carried out on November 5, 2007.

[2] The curation task that we refer to here requires curators to identify examples of protein-protein interactions in biomedical literature, which is a laborious task requiring considerable expertise.

Our previous work (Wang, 2007) reported initial results of a species disambiguation system and the performance of TI with the system integrated. The accuracy of species tagging was $56.0\%$ as tested by 10-fold cross validation on the training data and was $75.0\%$ on the development test data. This species tagging component also improved the performance of a rule-based TI system by $10\%$. Note that those experiments were conducted on a different dataset using a different species ontology from the ones reported in this paper, and therefore the results are not comparable to those presented in this paper.

## 3. Data and Ontology

The species annotated datasets were built as part of a larger project, the TXM project (Alex et al., 2008), a three-year project which aims to produce NLP-based tools to aid curation of biomedical papers. We created two corpora in slightly different domains, EPPI (enriched PPI) and TE (Tissue Expression). The EPPI corpus consists of 217 full-text papers selected from PubMed and PubMed Central and domain experts annotated all documents for both protein entities and PPIs, as well as extra (enriched) information associated with the PPIs and normalisations of the proteins to publicly available ontologies. The TE corpus consists of 230 full-text papers, in which entities such as protein, tissue, gene and mRNA were marked up and identified, and a new tissue expression relation was marked up in addition to protein-protein interactions. We split both corpora into training ($64\%$), development test (devtest) ($16\%$) and blind test ($20\%$) datasets. In this paper, we limit our discussion to species annotation.

Proteins, protein complexes, genes and mRNAs in both EPPI and TE datasets were annotated with NCBI taxonomy IDs[3] (TaxID) denoting the model organisms that they belong to. The NCBI taxonomy is a species ontology in a tree structure, containing $267,718$ nodes, where a node can represent various levels in a hierarchy of species including genus, subgenus and species, etc. In general, a genus consists of a number of subgenus which comprise of many species.

For example, Table 1 lists three nodes in different levels of the hierarchy with regarding to *Xenopus*.

| TaxID | Name | Rank |
|---|---|---|
| 8353 | Xenopus | genus |
| 262014 | Xenopus | subgenus |
| 8364 | Xenopus tropicalis | species |

Table 1: Taxonomy records for Xenopus in the NCBI taxonomy. 'Rank' refers to the hierarchy level of the node in the ontology.

During the species annotation, entity mentions were manually assigned with IDs of the corresponding nodes (i.e., TaxIDs) in the taxonomy hierarchy. For example, given the context in the article, an annotator may assign 8364 (*Xenopus tropicalis*) to an entity mention if she thinks it belongs to that species. On the other hand, if she was not sure that it was an *Xenopus tropicalis* species, but was certain that it belonged to the genus of *Xenopus*, she would assign it with the genus TaxID instead. Sometimes authors talk about an entity without referring to any specific genus, subgenus, or species, in which case the annotator would assign a "gen" tag to the entity mention, meaning that it is used in a general sense. For example, in the following sentence, the term *nNOS* is used in a general sense, and therefore this occurrence of *nNOS* should be tagged with "gen".

> The *nNOS* mRNAs that predominate in neurons and muscle arise through activation of promoters clustered in genomic regions considerably upstream of those that contain the translation initiation codon within exon 2.

In our experiments, however, we skipped all the "gen" cases so that every entity mention is associated with an NCBI species ID.

The EPPI and TE datasets have different distributions of species. The entities in the EPPI data belong to 112 species with *human (9606)* the most frequent at $51.98\%$. For all other species the number of entities is less than $10\%$ of the total, which makes for a very long tail in the distribution. On the other hand, in the TE data, the entities belong to 61 species with *mouse (10090)* the most frequent at $44.67\%$.[4] *Human* entities are also fairly frequent and come second at $34.40\%$. See Table 2 for a more detailed breakdown of the species distributions in the datasets.

To calculate the inter-annotator-agreement, about 40% of documents were doubly annotated by different annotators. The averaged $F1$ scores of species annotation on the doubly annotated EPPI and TE datasets are $86.45\%$ and $95.11\%$, respectively, which indicates that human annotators agree well when assigning species to biomedical entities.

## 4. Detecting Species Words

Words referring to species, such as *human*, are important indicators of the species of the nearby entities. We have developed a rule-based program that detects such words, which we refer to as *species words* in this paper. The heuristic rules were compiled by inspecting various species ontology, such as the NCBI taxonomy, and the training portion of the EPPI dataset. The species words are used as features to help the rule-based and the machine-learning based species identification modules described in Section 5.

### 4.1. Overview

The species word tagger is a lexical look-up component which applies to tokenised text and which marks words that refer to a species. The input tokenised text are in XML format and attributes denoting the species detected by the program are added to the corresponding word elements. For example, content words such as *human*, *mouse*, *murine* and *D. melanogaster* would be processed by the program and tagged with TaxIDs as attributes. For the most frequent species, the Latin name is also added as an attribute.

[3] `http://www.ncbi.nlm.nih.gov/sites/entrez?db=Taxonomy`

[4] These figures were obtained from the training splits of the EPPI and TE datasets.

| | in EPPI train | | in EPPI devtest | | in TE train | | in TE devtest | |
|---|---|---|---|---|---|---|---|---|
| | ncbitax ID | inst. (%) | ncbitax ID | inst. (%) | ncbitax ID | inst. (%) | ncbitax ID | inst. (%) |
| 1 | 9606 | 20948 (52.0) | 9606 | 7650 (60.6) | 10090 | 17688 (44.7) | 9606 | 2637 (33.3) |
| 2 | 10090 | 3916 (9.7) | 8364 | 1466 (11.6) | 9606 | 13605 (34.4) | 10090 | 2390 (30.2) |
| 3 | 4932 | 3047 (7.7) | 4932 | 872 (6.9) | 10116 | 2413 (6.1) | 7227 | 597 (7.5) |
| 4 | 8364 | 2483 (6.2) | 7227 | 694 (5.5) | 7227 | 1208 (3.1) | 6239 | 414 (5.2) |
| 5 | 10116 | 1860 (4.6) | 10090 | 550 (4.4) | 3702 | 632 (1.6) | 10116 | 390 (4.9) |
| 6 | 7227 | 1452 (3.6) | 10360 | 301 (2.4) | 3888 | 586 (1.5) | 4564 | 338 (4.3) |
| 7 | 4896 | 1226 (3.0) | 11963 | 238 (1.9) | 5270 | 518 (1.3) | 3527 | 309 (3.9) |
| 8 | 6239 | 386 (1.0) | 10116 | 215 (1.7) | 6239 | 361 (0.9) | 11292 | 188 (2.4) |
| 9 | 11676 | 343 (0.9) | 4564 | 195 (1.5) | 8364 | 347 (0.9) | 7668 | 162 (2.0) |
| 10 | 10376 | 318 (0.8) | 3527 | 154 (1.2) | 8030 | 344 (0.9) | 10335 | 139 (1.8) |
| .. | .. | .. | .. | .. | .. | .. | .. | .. |
| Total | 112 | 40300 | 38 | 12632 | 61 | 39591 | 38 | 7909 |

Table 2: Distributions of species in EPPI and TE devtest and train datasets. Only the top 10 species are shown. The 'ncbitax IDs' are species IDs drawn from the NCBI taxonomy and 'inst.' are the number of instances available for the species in the corresponding dataset. The row titled 'Total' denotes the total number of species which occurred and the total number of instances in the corresponding dataset.

| | PrevWd | | | PrevWd in Sent | | | PrevWd Spread | | | PrevWd in Sent Spread | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| PPI | 81.88 | 1.87 | 3.65 | 60.79 | 5.16 | 9.52 | 63.85 | 14.17 | 23.19 | 39.74 | 50.54 | 44.49 |
| TE | 91.49 | 1.63 | 3.21 | 56.16 | 7.76 | 13.64 | 77.84 | 17.97 | 29.20 | 31.71 | 46.68 | 37.36 |

Table 3: Results (%) of the rule-based species tagger.

In more detail, the first step in the component deals with species prefixes. Protein names frequently contain prefixes indicating the species of the protein, e.g. *mSos-1*. Rules written in an *lxtransduce* grammar[5] are used to identify species prefixes for the frequently occurring cases of *human*, *mouse*, *rat*, *fly* and *yeast*. For example, the term *mSos-1* would be tagged with attributes `sprefix="mus_musculus"` and `spid="10090"`, where the name of the species is encoded in the `sprefix` attribute and the TaxID in the `spid` attribute. The second step in the component does lexical look-up to identify species words. Note that a species "word" may contain several words, for example, "E. coli".

### 4.2. Lexicons

The species word tagger uses four lexicons. The lexicon *species.lex* contains hand-compiled entries for the twenty or so most frequent and most important species. The 188 entries cover Latin and English forms for each species and allow for pluralisation (e.g. *mice*) and different tokenisations (i.e. separated and non-separated fullstops are included so that the look-up does not rely on a particular tokenisation.) The lexicon *taxonomy.lex* is derived from the NCBI taxonomy and contains $400,179$ entries. This lexicon provides a TaxID for less frequent species terms which are not covered by *species.lex*. Its entries reflect both scientific names and common names and contain a rank attribute derived from the RANK field of the taxonomy (e.g. genus, species, subspecies etc.). In the course of conversion from taxonomy to lexicon, certain entries were discarded. For example, entries with excessive punctuation and/or dates were rejected:

- *"Pseudomonas tumefaciens" (Smith and Townsend 1907) Duggar 1909*

In addition, entries which are homonyms of common English words were also rejected, e.g. *This, bear, sole* and unhelpful entries were removed, e.g. *other, unknown, unclassified*.

The lexicon *latinother.lex* is derived from a list on the UniProt web site[6] and contains $32,764$ entries. This lexicon is used to identify Latin species names which may not have been found in the previous lexicon because, for example, the NCBI taxonomy does not include short forms of Latin names. This lexicon contains both full names and short forms as well as alternative tokenisations, such as *Hyaena hyaena; H. hyaena; H . hyaena*.

The final lexicon, *extras.lex* is a hand-built lexicon containing 83 entries which provides TaxIDs for words that are not found in the taxonomy lexicon. It includes species adjectives such as *ovine* as well as common species words such as *frog* and *hamster*.

## 5. Assigning Species to Entities

### 5.1. Rule-based Approach

It is intuitive that a species word that occurs right before an entity mention (e.g., *mouse p53*) should be a strong indicator of its species. To assess how well this intuition works, we developed a rule-based system using the heuristic and species words detected by the species word tagger. We devised four rule-based systems:

[5] See `http://www.ltg.ed.ac.uk/software/ltxml2` for details of the LT-XML 2 tools developed at the LTG group at Edinburgh University.

[6] `http://www.ebi.uniprot.org/index.shtml`

- *PrevWd*: If the word preceding an entity mention is a species word, assign the species indicated by that word to the mention.
- *PrevWd in Sent*: If a word that occurs to the left of an entity mention and in the same sentence is a species word, assign the species indicated by that word to the mention.
- *PrevWd Spread*: First carry out rule *PrevWd*, and then spread the species to all the entity mentions whose surface forms are identical to the mentions tagged by rule *PrevWd*.
- *PrevWd in Sent Spread*: First carry out rule *PrevWd in Sent*, and then spread the species to all the entity mentions whose surface forms are identical to the mentions tagged by rule *PrevWd in Sent*.

Table 3 shows the evaluation results. As we can see, the precision of the system 'PrevWd' that solely relies on the previous species word to an entity mention was good but the recall was very low. The system 'PrevWd in Sent' that looks at the previous species word in the same sentence did better as measured by $F1$, but compared to 'PrevWd' its precision went down. We carried out some error analysis on the species tags predicted by the rule 'PrevWd in Sent' as follows:

- *"Expression of* CYP2B6*, a human relative of* CYP2B10 *..."* In this example, *CYP2B10* is in fact a *mouse* protein but the rule-based system tagged it as a *human* one, indicating that the rule is not always reliable.
- *"The Drosophila methyl-DNA binding protein* MBD2/3 *..."* The occurrence of *MBD2/3* was tagged as *Drosophila melanogaster (species, 7227)* by the rule but the annotator's answer was *Drosophila (genus, 7215)*. Arguably either answer should be correct as "Drosophila melanogaster" is a species under genus "Drosophila". We used a strict scoring in our evaluation without collapsing levels of the NCBI taxonomy hierarchy which made the task even more difficult.
- *"Identification of the 15FRFG domain in HIV-1* Gag *..."* This occurrence of *Gag* was tagged as *HIV-1 (11676)* by the rule but the gold standard was *HIV (12721)*. This was actually a mistake in the manual annotation and the species tagger was correct.

Spreading the species improved the $F1$ scores of both systems but the overall results were still not satisfactory, which implies that occurrences of entity mentions with the same surface forms and in the same document do not necessarily share the same model organism.

### 5.2. Machine Learning Approach

We also conducted research on machine-learning approaches to species tagging. First, we paired up a vector of contextual features with every entity mention in the training splits of the EPPI and TE data. Then, a number of Maximum Entropy models[7] were trained on such instances. The contextual features used include:

- *leftContext* The $n$ word lemmas to the left of the entity mention, without position ($n = 200$).[8]
- *rightContext* The $n$ word lemmas to the right of the entity mention, without position ($n = 200$).
- *leftSpeciesIDs* The $n$ species IDs, located to the left of the entity mention and assigned by the species word tagger described above ($n = 5$).
- *rightSpeciesIDs* The $n$ species IDs, located to the right of the entity mention and assigned by the species word tagger described above ($n = 5$).
- *leftNouns* The $n$ nouns to the left of the entity mention (with order and $n = 2$). This feature attempts to capture cases where a noun preceding a mention indicates species, e.g., *mouse protein p53*.
- *leftAdjs* The $n$ adjectives to the left of the entity mention (with order and $n = 2$). This feature is intended to capture cases where an adjective preceding a mention indicates species, e.g., *murine protein p53*.
- *leftSpeciesWords* The $n$ species word forms, identified by the species word tagger, located to the left of the entity mention ($n = 5$).
- *rightSpeciesWords* The $n$ species word forms, identified by the species word tagger, located to the right of the entity mention ($n = 5$).
- *firstLetter* The first character of the entity mention itself. Sometimes the first letters of entities indicate their species, e.g., *hP53*.
- *documentSpeciesIDs* All species IDs occurring in the article in question.
- *useStopWords* If this feature is switched on then filter out the words that appear in a pre-compiled stopword list from the above features. The list consists of frequent common English words such as prepositions (e.g., *in*) and conjunctions, etc.
- *useStopPattern* If this feature is switched on then filter out the words consisting only of digits and punctuation characters.

The 5-fold cross-validation test results are shown in Table 4 and results on devtest datasets in Table 5. We use accuracy instead of $F1$ because the machine-learning based tagger assigns a species tag to *every* entity occurrence, and therefore precision is equal to recall and to $F1$. In addition, in

[7] This software program was developed by Le Zhang at Edinburgh University. See `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`

[8] 'Without position' means that we ignored the position of the contextual words with respect to the position of the entity in question. In other words, the contextual words were treated as a bag of words.

| fold | 1 | 2 | 3 | 4 | 5 | avg. |
|---|---|---|---|---|---|---|
| EPPI | 64.67 | 56.05 | 56.29 | 67.26 | 58.63 | 60.58 |
| TE | 80.72 | 77.26 | 64.29 | 62.03 | 70.95 | 71.05 |

Table 4: Accuracy (%) of the machine-learning based species tagger tested by 5-fold cross validation on EPPI and TE *training* datasets.

| | BL | EPPI Model | TE Model | Combined Model |
|---|---|---|---|---|
| EPPI | 60.56 | 73.04 | 66.42 | **71.28** |
| TE | 33.28 | 63.91 | **70.73** | 68.18 |

Table 5: Accuracy (%) of the machine-learning based species tagger tested on EPPI and TE *devtest* datasets.

all the tests, we excluded general use of species (i.e.,"gen") and only evaluated those entities that can be assigned to specific species.

As shown in Table 5, we tested four models on the devtest portions of EPPI and TE corpora: BL, EPPI Model, TE Model, and Combined Model. BL is a baseline system, which tags the entity mentions in the devtest datasets using the most frequent species occurring in the corresponding training datasets. For example, *human* was the most frequent species in the EPPI training data, and therefore all entity occurrences in the EPPI devtest dataset were tagged with *human*. The EPPI Model was obtained by training the Maxent classifier on the EPPI training data, the TE Model by training on TE training data, and the Combined Model was trained on a combined dataset consisting of both the EPPI and TE training corpora. The Combined Model outperformed other models on the EPPI devtest dataset while the TE model yielded the best result for the TE dataset.

### 5.3. Discussion

The main problem with the machine learning approach lies in the fact that a trained model is biased toward the distribution of species that the training dataset possesses. For example, our model trained on the EPPI training data achieved 73.04% accuracy as tested on the EPPI devtest data. However, it only yielded 63.91% when tested on the TE devtest data. As we split the training and devtest portions after the annotation stage, it is reasonable to assume that the species distribution in training and devtest datasets of the same type of data are comparable. This raises the question of whether a model trained on one dataset can be ported to other datasets and can still accurately identify species of the entities.

Looking again at Table 2, which shows that the species distributions of the PPI and TE corpora are very different. This table only lists the most frequent 10 species in each corpus. In fact, the distributions have long "tails" where many species occur only a few times, which makes it nearly impossible for the learned model to detect when they occur in unseen text.

In addition, it can also happen that, a species that has never occurred in the training dataset occurs in the devtest dataset, in which case a machine learned model would have no chance to pick it up. More research into a hybrid system that makes better use of rules may be able to shed some light on this problem.

## 6. Conclusions and Future Work

We adopted the species annotated corpora developed in the TXM project and investigated various techniques for assigning species tags to biomedical entities. We found that the common heuristic of tagging an entity with the species indicated by its previous species word was not reliable: as tested on our EPPI and TE datasets, this heuristic achieved good precision of $81.88\%$ and $91.49\%$, but very low recall. Subsequently, we experimented with a machine-learning based approach and with a large set of features and with different parameter settings. Our best results were much higher than the rule-based system, with $F1$ scores over $71\%$. The problem with the machine-learning based approach, however, is that distribution of species in the training data has a big impact on the model trained on it. In other words, a species tagger trained on a corpus dominated by *human* would have little chance of achieving good results on a test corpus full of *zebra fish*. Increasing the size and coverage of the training data is an obvious solution and our experiment showed that a model trained on a combined set of data from both the EPPI and the TE domains achieved very good results on devtest dataset from either domain.

In the future we would like to explore how we can seek help from specific rules in situations when machine-learning models would not work. For example, in articles that are not talking about the common species such as *human*, *fly* and *mouse*, specific rules making use of species words might work better for detecting the species of the biomedical entities.

In addition, we would like to integrate the species tagger into term identification and relation extraction systems, making them capable of dealing with biomedical entities across multiple species.

## Acknowledgements

This work was supported by the Text Mining programme, ITI Life Sciences, Scotland.[9]

## 7. References

B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, and X. Wang. 2008. Assisted curation: does text mining really help? In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*.

S. Ananiadou, C. Friedman, and J. Tsujii. 2004. Introduction: Named entity recognition in biomedicine. *Journal of Biomedical Informatics*, 37(6):393–395.

C. Blaschke, E. A. Leon, M. Krallinger, and A. Valencia. 2005. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, 6(Suppl 1:S16).

L. Chen, H. Liu, and C. Friedman. 2005. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2):248–256.

[9] http://www.itilifesciences.com

C. Grover, B. Haddow, E. Klein, M. Matthews, L. A. Nielsen, R. Tobin, and X. Wang. 2007. Adapting a relation extraction pipeline for the BioCreAtIvE II task. In *Proceedings of the BioCreAtIvE II Workshop 2007*, Madrid.

L. Hirschman, M. Colosimo, A. Morgan, J. Columbe, and A. Yeh. 2004. Task 1B: Gene list task BioCreAtIve workshop. In *BioCreative: Critical Assessment for Information Extraction in Biology*.

L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. 2005. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl1):S1.

L. Hirschman, M. Krallinger, and A. Valencia, editors. 2007. *Second BioCreative Challenge Evaluation Workshop*. Fundación CNIO Carlos III, Madrid, Spain.

N. Karamanis, I. Lewin, R. Seal, R. Drysdale, and E. Briscoe. 2007. Integrating natural language processing with FlyBase curation. In *Proceedings of PSB*, pages 245–256, Maui, Hawaii.

M. Krallinger, F. Leitner, and A. Valencia. 2007. Assessment of the second BioCreative PPI task: Automatic extraction of protein-protein interactions. In *Proceedings of the BioCreAtIvE II Workshop 2007*, pages 41–54, Madrid, Spain.

M. Krauthammer and G. Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics (Special Issue on Named Entity Recogntion in Biomedicine)*, 37(6):512–526.

A. A. Morgan and L. Hirschman. 2007. Overview of BioCreative II gene normalisation. In *Proceedings of the BioCreAtIvE II Workshop*, Madrid.

X. Wang and M. Matthews. 2008. Comparing usability of matching techniques for normalising biomedical named entities. In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*.

X. Wang. 2007. Rule-based protein term identification with help from automatic species tagging. In *Proceedings of CICLING 2007*, pages 288–298, Mexico City.

J. Wilbur, L. Smith, and L. Tanabe. 2007. Biocreative 2 gene mention task. In *Proceedings of the BioCreAtIvE II Workshop 2007*, pages 7–16, Madrid, Spain.

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