Using existing biomedical resources to detect and ground terms in biomedical literature

Kaljurand, K ; Rinaldi, Fabio ; Kappeler, T ; Schneider, G

Abstract: We present an approach towards the automatic detection of names of proteins, genes, species, etc. in biomedical literature and their grounding to widely accepted identifiers. The annotation is based on a large term list that contains the common expression of the terms, a normalization step that matches the terms with their actual representation in the texts, and a disambiguation step that resolves the ambiguity of matched terms. We describe various characteristics of the terms found in existing term resources and of the terms that are used in biomedical texts. We evaluate our results against a corpus of manually annotated protein mentions and achieve a precision of 57% and recall of 72%.

DOI: https://doi.org/10.1007/978-3-642-02976-9_32

Using existing biomedical resources to detect and ground terms in biomedical literature

Kaarel Kaljurand, Fabio Rinaldi, Thomas Kappeler, Gerold Schneider

Institute of Computational Linguistics, University of Zurich
kalju@ifi.uzh.ch, rinaldi@ifi.uzh.ch, kappeler@bluewin.ch,
gschneid@ifi.uzh.ch

Abstract. We present an approach towards the automatic detection of names of proteins, genes, species, etc. in biomedical literature and their grounding to widely accepted identifiers. The annotation is based on a large term list that contains the common expression of the terms, a normalization step that matches the terms with their actual representation in the texts, and a disambiguation step that resolves the ambiguity of matched terms. We describe various characteristics of the terms found in existing term resources and of the terms that are used in biomedical texts. We evaluate our results against a corpus of manually annotated protein mentions and achieve a precision of 57% and recall of 72%.

1 Introduction

The complexity of biological organisms and the success of biological research in describing them, have resulted in a large body of biological entities (genes, proteins, species, etc.) to be indexed, named and analyzed. Proteins are among the most important entities. They are an essential part of an organism and participate in every process within cells. Most proteins function in collaboration with other proteins, and one of the research goals in molecular biology is to identify which proteins interact.

While the number of different proteins is large, the amount of their possible interactions and combinations is even larger. In order to record such interactions and represent them in a structured way, human curators who work for knowledge base projects, e.g. MINT[1] and IntAct[2] (see [5] for a detailed overview), carefully analyze published biomedical articles. As the body of articles is growing rapidly, there is a need for effective automatic tools to help curators in their work. Such tools must be able to detect mentions of biological entities in the text and tag them with identifiers that have been assigned by existing knowledge bases. As the names that are used to reference the proteins can be very ambiguous, there is a need for an effective ambiguity resolution.

[1] http://mint.bio.uniroma2.it
[2] http://www.ebi.ac.uk/intact

In this paper, we describe the task of automatically detecting names of proteins, genes, species, experimental methods, and cell lines in biomedical literature and grounding them to widely accepted identifiers assigned by three different knowledge bases — UniProt Knowledgebase (UniProtKB)[3], National Center for Biotechnology Information (NCBI) Taxonomy[4], and Proteomics Standards Initiative (PSI) Molecular Interactions (MI) Ontology[5].

The term annotation uses a large term list that is compiled on the basis of the entity names extracted from the mentioned knowledge bases and from a list of cell line names. This resulting list covers the common expression of the terms. A term normalization step is used to match the terms with their actual representation in the texts. Finally, a disambiguation step resolves the ambiguity (i.e. multiple IDs proposed by the annotator) of the matched terms.

The work presented here is part of a larger effort undertaken in the OntoGene project[6] aimed at improving biomedical text mining through the usage of advanced natural language processing techniques. The results of the entity detection feed directly into the process of identification of protein interactions. Our approach relies upon information delivered by a pipeline of NLP tools, including sentence splitting, tokenization, part of speech tagging, noun and verb phrase chunking, and a dependency-based syntactic analysis of input sentences [7]. The syntactic parser takes into account constituent boundaries defined by previously identified multi-word entities. Therefore the richness of the entity annotation has a direct beneficial impact on the performance of the parser, and thus leads to better recognition of interactions.

2 Term resources

As a result of the rapidly growing information in the field of biology, the research community has realized the need for consistently organizing the discovered information — assign identifiers to biological entities, enumerate the names by which the entities are referred to, interlink different resources (e.g. existing knowledge bases and literature), etc. This has resulted in large and ever-growing knowledge bases (lists, ontologies, taxonomies) of various biological entities (genes, proteins, species, etc.). Fortunately, many of these resources are also freely available and machine processable. These resources can be treated as linguistic resources and used as an input for the creation of large term lists. Such lists can be used to annotate existing biomedical publications in order to identify the entities mentioned in these publications. In the following we describe four resources: UniProtKB, NCBI Taxonomy, PSI-MI Ontology, and CLKB.

[3] http://www.uniprot.org
[4] http://www.ncbi.nlm.nih.gov/Taxonomy/
[5] http://psidev.sourceforge.net/mi/psi-mi.obo
[6] http://www.ontogene.org

2.1 UniProtKB

The UniProt Knowledgebase (UniProtKB)[7] assigns identifiers to 397,539 proteins and describes their amino-acid sequences. The identifiers come in two forms: numeric accession numbers (e.g. `P04637`), and mnemonic identifiers that make visible the species that the protein originates from (e.g. `P53_HUMAN`). In the following we always use the mnemonic identifiers for better readability.

In addition to enumerating proteins, UniProtKB lists their names that are commonly used in the literature. The set of names covers names with large lexical difference (e.g. both 'Orexin' and 'Hypocretin' can refer to protein `OREX_HUMAN`), but usually not names with minor spelling variations (e.g. using a space instead of a hyphen).

We extracted 626,180 (different) names from the UniProtKB XML file. The ambiguity of a name can be defined as the number of different UniProtKB entries that contain the name. UniProtKB names can be very ambiguous. This follows already from the naming guideline which states that "a recommended name should be, as far as possible, unique and attributed to all orthologs"[8]. Thus, a protein that is found in several species has one name but each of the species contributes a different ID. In UniProtKB, the average ambiguity is 2.61 IDs per name. If we discard the species labels, then the average ambiguity is 1.05 IDs. Ambiguous names (because the respective protein occurs in multiple species) are e.g. 'Cytochrome b' (1770 IDs), 'Ubiquinol-cytochrome-c reductase complex cytochrome b subunit' (1757), 'Cytochrome b-c1 complex subunit 3' (1757). Ambiguous names (without species labels) are e.g. 'Capsid protein' (103), 'ORF1' (97), 'CA' (88).

Table 1 shows the orthographic/morphological properties of the names in UniProtKB in terms of how much certain types of characters influence the ambiguity. Non alphanumeric characters or change of case, while increasing ambiguity, influence the ambiguity relatively little. But as seen from the last column, digits matter a lot semantically. These findings motivate the normalization that we describe in section 3.1. Table 1 also shows the main cause for ambiguity of the names — the same name can refer to proteins in multiple species. While these proteins are identical in some sense (similar function or structure), the UniProtKB identifies them as different proteins.

2.2 NCBI Taxonomy

The National Center for Biotechnology Information provides a resource called NCBI Taxonomy[9], describing all known species and listing the various forms of species names (e.g. scientific and common names). As explained in section 2.1, knowledge of these names is essential for disambiguation of protein names.

[7] We use the manually annotated and reviewed Swiss-Prot section of UniProtKB version 14, in its XML representation

[8] `http://www.uniprot.org/docs/nameprot`

[9] `http://www.ncbi.nlm.nih.gov/Taxonomy/`

Table 1. Ambiguity of UniProtKB terms. `ID_ORG` stands for the actual identifiers, which include the species ID. `ID` stands for artificially created identifiers where the qualification to the species has been dropped. "Unchanged" = no change done to the terms; "No whitespace" = all whitespace is removed; "Alphanumeric" = only alphanumeric characters are preserved; "Lowercase" = all characters are preserved but lowercased; "Alpha" = only letters are preserved.

	Unchanged	No whitespace	Alphanumeric	Lowercase	Alpha
`ID_ORG`	2.609	2.611	2.624	2.753	10.616
`ID`	1.049	1.050	1.053	1.058	4.145

We compiled a term list on the basis of the taxonomy names list[10], but kept only names whose ID mapped to a UniProtKB species "mnemonic code" (such as `ARATH`)[11]. The final list contains 31,733 entries where the species name is mapped to the UniProtKB mnemonic code. To this list, 8877 entries were added where the genus name is abbreviated to its initial (e.g. 'C. elegans') as names in such form were not included in the source data. These entries can be ambiguous in general (e.g. 'C. elegans' can refer to four different species), but are needed to account for such frequently occurring abbreviation in biomedical texts. Furthermore, six frequently occurring names that consist only of the genus name were added. In these cases, the name was mapped to a unique identifier (e.g. 'Arabidopsis' was mapped to `ARATH`), as it is expected that e.g. 'Arabidopsis' alone is always used to refer to *Arabidopsis thaliana*, and never to e.g. *Arabidopsis lyrata*.

2.3 PSI-MI Ontology

The Proteomics Standards Initiative (PSI) Molecular Interactions (MI) Ontology[12] contains 2207 terms (referring to 2163 PSI-MI IDs) related to molecular interaction and methods of detecting such interactions (e.g. 'western blot', 'pull down'). There is almost no ambiguity in these names in the ontology itself. Several reasons motivate including the PSI-MI names in our term list. First, names of experimental methods are very frequent in biomedical texts. It is thus important to annotate such names as single units in order to make the syntactic analysis of the text more accurate. Second, in some cases a PSI-MI name contains a substring which happens to be a protein name (e.g. 'western blot' contains a UniProtKB term 'blot'). If the annotation program is not aware of this, then some tokens would be mistagged as protein names. Third, some PSI-MI terms overlap with UniProt terms, meaning that the corresponding proteins play an important function in protein interaction detection, but are not the subject of the actual interaction. An example of this is 'GFP' (PSI-MI ID `0367`, UniProtKB

[10] `ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz` (file `names.dmp`)
[11] `http://www.uniprot.org/help/taxonomy`
[12] `http://psidev.sourceforge.net/mi/psi-mi.obo`

ID GFP_AEQVI), which occurs in sentences like "interaction between Pop2p and GFP-Cdc18p was detected" where the reported interaction is between POP2 and CDC18, and GFP only "highlights" this interaction.

2.4 Cell line names

Cell line names occur frequently in biomedical articles, and one has to be aware of these names in order to avoid tagging them as e.g. protein names. Secondly, almost every cell line comes from one species (although also "chimeric" cell lines are sometimes used), thus the mention of a cell line in a sentence can give a hint of which species the given sentence is about.

We extracted 8741 cell line names from the Cell Line Knowledgebase (CLKB)[13] which is a compilation of data (names, identifiers, cell line organisms, etc.) from various cell line resources (HyperCLDB, ATCC, MeSH) [8]. The data is provided in the standard RDF format. The cell line names in CLKB contain very little ambiguity and synonymy.

CLKB does not map the cell line organism labels to NCBI IDs. This is not directly possible because the organism label often points to a strain, breed, or race of a particular organism (e.g. 'human, Caucasian', 'mouse, BALB/c'), but NCBI does not assign IDs with such granularity. In total, there are 257 organism labels, the most frequent of which we map to the UniProtKB species mnemonic codes (e.g. HUMAN, MOUSE) and the rest to a dummy identifier.

2.5 Compiled term list

We compiled a term list of 1,688,224 terms based on the terms extracted from UniProtKB, NCBI, PSI-MI, and CLKB, listing the term name, the term ID, and the term type in each entry. The type corresponds roughly to the resource the term originates from. For UniProtKB, there are two types, PROT and GEN. For NCBI, there are six types, distinguishing between common and scientific names, and the rank of the name in the taxonomy. For the PSI-MI Ontology terms and CLKB cell line names there is one type — MI or CLKB, respectively. The frequency distribution of types is listed in table 2. There is relatively little type ambiguity — three terms ('P22', 'LI', 'D2') can belong to three different types, 300 terms to two different types. In the latter case, the ambiguity is between PROT/GEN and CLKB in 209 cases, and between PROT/GEN and MI in 69 cases.

3 Automatic annotation of terms

Using the described term list, we can annotate biomedical texts in a straightforward way. First, the sentences and tokens are detected in the input text. We use the LingPipe[14] tokenizer and sentence splitter which have been trained on

[13] http://stateslab.org/data/CellLineOntology/
[14] http://alias-i.com/lingpipe/

Table 2. Frequency distribution of types in the compiled term list, together with the source of the IDs that are assigned to the terms.

Frequency	Type ID		Description
884,641	`PROT`	UniProt	UniProtKB protein name
752,019	`GEN`	UniProt	UniProtKB gene name
16,979	`ocs`	NCBI	NCBI common name, species or below
8877	`oss`	NCBI	NCBI scientific name, species or below
8877	`ogs2`	NCBI	`oss` name, genus abbreviated (e.g. 'A. thaliana')
8741	`CLKB`	NCBI	CLKB cell line name
3316	`oca`	NCBI	NCBI common name, above species
2561	`osa`	NCBI	NCBI scientific name, above species
2207	`MI`	PSI-MI	PSI-MI term
6	`ogs1`	NCBI	NCBI selected genus name (e.g. 'Arabidopsis')

biomedical corpora. The tokenizer produces a granular set of tokens, e.g. words that contain a hyphen (such as 'Pop2p-Cdc18p') are split into several tokens, revealing the inner structure of such constructs which would e.g. allow to discover the interaction mention in "Pop2p-Cdc18p interaction". The processing then annotates the longest possible and non-overlapping sequences of tokens in each sentence, and in the case of success, assigns all the possible IDs (as found in the term list) to the annotated sequence. The annotator ignores certain common English function words (we use a list of ∼50 stop words), as well as figure and table references (e.g. 'Fig. 3a' and 'Table IV').

3.1 Normalization

In order to account for possible orthographic differences between the terms in the term list and the token sequences in the text, a normalization step is included in the annotation procedure. The same normalization is applied to the term list terms in the beginning of the annotation when the term list is read into memory, and to the tokens in the input text. In case the normalized strings match exactly, the input sequence is annotated with the IDs of the term list term. Our normalization rules are similar to the rules reported in [1,10], e.g.

- Remove all characters that are not alphanumeric or space
- Remove lowercase-uppercase distinction
- Normalize Greek letters and Roman numerals, e.g. 'alpha' → 'a', 'IV' → '4'
- Remove the final 'p' if it follows a number, e.g. 'Pan1p' → 'Pan1'
- Remove certain species-indicating prefixes (e.g. 'h' for human, 'At' for *Arabidopsis thaliana*), but in this case, admit only IDs of the given species

In general, these rules increase the recall of term detection, but can lower the precision. For example, sometimes case distinction is used to denote the same protein in different species (e.g. according to UniProtKB, the gene name 'HOXB4' refers to `HXB4_HUMAN`, 'Hoxb4' to `HXB4_MOUSE`, and 'hoxb4' to `HXB4_XENLA`). The gain in recall, however, seems to outweigh the loss of precision.

3.2 Disambiguation

A marked up term can be ambiguous for two reasons. First, the term can be assigned an ID from different term types, e.g. a UniProtKB ID and a PSI-MI Ontology ID. This situation does not occur often and usually happens with terms that are probably not interesting as protein mentions (such as 'GFP' discussed in section 2.3). We disambiguate such terms by removing all the UniProtKB IDs. (Similar filtering is performed in [9].) Second, the term can be assigned several IDs from a single type. This usually happens with UniProtKB terms and is typically due to the fact that the same protein occurs in many different species. Such protein names can be disambiguated in various ways. We have experimented with two different methods: (1) remove all the IDs that do not reference a species ID specified in a given list of species IDs; (2) remove all IDs that do not "agree" with the IDs of the other protein names in the same textual span (e.g. sentence, or paragraph) with respect to the species IDs.

For the first method, the required species ID list can be constructed in various ways, either automatically, on the basis of the text, e.g. by including species mentioned in the context of the protein mention, or by reusing external annotations of the article. We present in [2] an approach to the detection of species names mentioned in the article. The species mentions are used to create a ranked list, which is then used to disambiguate other entities (e.g. protein mentions) in the text.

The second method is motivated by the fact that according to the IntAct database, interacting proteins are usually from the same species: less than 2% of the listed interactions have different interacting species. Assuming that proteins that are mentioned in close proximity often constitute a mention of interaction, we can implement a simple disambiguation method: for every protein mention, the disambiguator removes every UniProtKB ID that references a species that is not among the species referenced by the IDs of the neighboring protein mentions.

In general, the disambiguation result is not a single ID, but a reduced set of IDs which must be further reduced by a possible subsequent processing step.

4 Evaluation

We evaluated the accuracy of our automatic protein name detection and grounding method on a corpus provided by the IntAct project[15]. This corpus contains a set of 6198 short textual snippets (of 1 to about 3 sentences), where each snippet is mapped to a PubMed identifier (of the article the snippet originates from), and an IntAct interaction identifier (of the interaction that the snippet describes). In other words, each snippet is a "textual evidence" that has allowed the curator to record a new interaction in the IntAct knowledge base. By resolving an interaction ID, we can generate a set of IDs of interacting proteins and a set of species involved in the interaction, for the given snippet. Using the PubMed identifiers, we can generate the same information for each mentioned article. By

[15] `ftp://ftp.ebi.ac.uk/pub/databases/intact/current/various/data-mining/`

Table 3. Results obtained on the IntAct snippets, with various forms of disambiguation, measured against PubMed IDs. The evaluation was performed on the complete IntAct data (*all*), and on a 5 times smaller fragment of IntAct (*subset*) for which we automatically extracted the species information. Three forms of disambiguation were applied: IntAct = species lists from IntAct data; TX = species lists from our automatic species detection; span = the species of neighboring protein mentions must match. Additionally, combinations of these were tested: e.g. IntAct & span = IntAct disambiguation followed by span disambiguation. The best result in each category is in boldface.

Disamb. method	Corpus	Precision	Recall	F-Score	True pos.	False pos.	False neg.
No disamb.	all	0.03	0.73	0.05	2237	81,662	848
IntAct	all	0.56	0.73	0.63	2183	1713	804
span	all	0.03	0.71	0.06	2186	68,026	899
IntAct & span	all	0.57	0.72	**0.64**	2147	1599	840
span & IntAct	all	0.57	0.72	0.64	2142	1631	821
No disamb.	subset	0.02	0.69	0.04	424	20,344	188
IntAct	subset	0.51	0.71	0.59	414	397	170
span	subset	0.02	0.67	0.05	407	16,319	205
IntAct & span	subset	0.53	0.69	**0.60**	404	363	180
span & IntAct	subset	0.52	0.69	0.59	399	369	177
TX	subset	0.42	0.59	0.49	340	478	241
TX & span	subset	0.43	0.57	**0.49**	332	445	249
span & TX	subset	0.42	0.57	0.48	329	457	244

comparing the sets of protein IDs reported by the IntAct corpus providers, and the sets of protein IDs proposed by our tool, we can calculate the precision and recall values.

We annotated the complete IntAct corpus by marking up with an entry in the term list the token sequences that the normalization step matched. Each resulting annotation includes a set of IDs which was further reduced by the two disambiguation methods described in 3.2, i.e. some or all of the IDs were removed. Results before and after disambiguation are presented in table 3. The results show a relatively high recall which decreases after the disambiguation. This change is small however, compared to the gain in precision. False negatives are typically caused by missing names in UniProtKB, or sometimes because the normalization step fails to detect a spelling variation. A certain amount of false positives cannot be avoided due to the setup of task — the tool is designed to annotate all proteins contained in the sentences, but not all of them necessarily participate in interactions, and thus are not reported in the IntAct corpus.

5 Related work

There is a large body of work in named entity recognition in biomedical texts. Mostly this work does not cover grounding the detected named entities to exist-

ing knowledge base identifiers. Recently, however, as a result of the BioCreative workshop, more approaches are extending from just detecting entity mentions to "normalizing" of the terms. In general, such normalization handles gene names (by grounding them to EntrezGene[16] identifiers). [6] gives an overview of the BioCreative II gene normalization task.

A method of protein name grounding is described in [10]. It uses a rule-based approach that integrates a machine-learning based species tagger to disambiguate protein IDs. The reported results are similar to ours. In the BioCreative Meta Server (BCMS)[17] [3], 2 out of 13 gene/protein taggers annotate using UniProtKB protein identifiers. The Whatizit[18] webservice annotates input texts with UniProtKB, Gene Ontology[19], and NCBI terms. A preliminary comparison showed that our approach gives results of similar quality.

Several linguistic resources have been compiled from existing biomedical databases. BioThesaurus[20] is a thesaurus of gene and protein names (and their synonyms and textual variants) where each name is mapped to a UniProtKB identifier [4]. The latest version 5.0 of BioThesaurus contains more than 9 million names, extracted from 35 different databases. The biggest contributer, however, is UniProtKB, mainly its TrEMBL section.

ProMiner[21] is a closed source dictionary-based named entity tagger that uses an entity name database compiled from a wide variety of sources for gene, protein, disease, tissue, drug, cell line, and other names. Detailed information about this resource has not been published.

6 Conclusions and future work

The main goal of the work described in this paper is to reliably identify protein mentions in order to identify protein-protein interactions in a subsequent processing step. We use a large term list compiled from various sources, and a set of normalization rules that match the token sequences in the input sentences against the term list. Each matched term is assigned all the IDs that are possible for this term. The following disambiguation step removes most of the IDs on the basis of the term context and knowledge about the species that the article discusses. For the evaluation, we have used the freely available IntAct corpus of snippets of textual evidence for protein-protein interactions. To our knowledge, this corpus has not been used in a similar evaluation before.

In the future, we would like to include more terminological resources in the annotation process. While the described four resources (UniProtKB, NCBI Taxonomy, PSI-MI Ontology, CLKB cell line names) contain the most important names used in biomedical texts, there exist other names that are frequently

[16] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
[17] http://bcms.bioinfo.cnio.es
[18] http://www.ebi.ac.uk/webservices/whatizit/
[19] http://www.geneontology.org
[20] http://pir.georgetown.edu/iprolink/biothesaurus/
[21] http://www.scai.fraunhofer.de/prominer.html

used but that are not covered by these resources, e.g. names of certain chemical compounds, diseases, drugs, tissues, etc.

Acknowledgments

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel, Switzerland. The authors would like to thank the three anonymous reviewers of AIME'09 for their valuable feedback.

References

1. Jörg Hakenberg. What's in a gene name? Automated refinement of gene name dictionaries. In *Proceedings of BioNLP 2007: Biological, Translational, and Clinical Language Processing; Prague, Czech Republic*, 2007.
2. Thomas Kappeler, Kaarel Kaljurand, and Fabio Rinaldi. TX Task: Automatic Detection of Focus Organisms in Biomedical Publications. In *BioNLP 2009, NAACL/HLT*, Boulder, Colorado, 4–5 June 2009.
3. F Leitner, M Krallinger, C Rodriguez-Penagos, J Hakenberg, C Plake, C-J Kuo, C-N Hsu, RT-H Tsai, H-C Hung, WW Lau, CA Johnson, R Saetre, K Yoshida, YH Chen, S Kim, S-Y Shin, B-T Zhang, WA Baumgartner, L Hunter, B Haddow, M Matthews, X Wang, P Ruch, F Ehrler, A Ozgur, G Erkan, DR Radev, M Krauthammer, T Luong, and R Hoffmann. Introducing meta-services for biomedical information extraction. *Genome Biology*, 9(Suppl 2):S6, 2008.
4. Hongfang Liu, Zhang-Zhi Hu, Jian Zhang, and Cathy Wu. BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22(1):103–105, 2006.
5. Suresh Mathivanan, Balamurugan Periaswamy, TKB Gandhi, Kumaran Kandasamy, Shubha Suresh, Riaz Mohmood, YL Ramachandra, and Akhilesh Pandey. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, 7(Suppl 5):S19, 2006.
6. AA Morgan, Z Lu, X Wang, AM Cohen, J Fluck, P Ruch, A Divoli, K Fundel, R Leaman, J Hakenberg, C Sun, H-h Liu, R Torres, M Krauthammer, WW Lau, H Liu, C-N Hsu, M Schuemie, KB Cohen, and L Hirschman. Overview of BioCreative II gene normalization. *Genome Biology*, 9(Suppl 2):S3, 2008.
7. Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13, 2008.
8. Sirarat Sarntivijai, Alexander S. Ade, Brian D. Athey, and David J. States. A bioinformatics analysis of the cell line nomenclature. *Bioinformatics*, 24(23):2760–2766, 2008.
9. Lorraine Tanabe and W. John Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124–1132, 2002.
10. Xinglong Wang and Michael Matthews. Distinguishing the species of biomedical named entities for term identification. *BMC Bioinformatics*, 9(Suppl 11):S6, 2008.