Terminological resources for Text Mining over Biomedical Scientific Literature

Fabio Rinaldi [a,*] , Kaarel Kaljurand [a] , Rune Sætre [b,c] .

[a] *Institute of Computational Linguistics, IFI, University of Zurich, Switzerland*

[b] *Tsujii Laboratory, University of Tokyo, Japan*

[c] *Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU)*

Abstract

We present a combined terminological resource for text mining over biomedical literature. The purpose of the resource is to allow the detection of mentions of specific biological entities in scientific publications, and their grounding to widely accepted identifiers. This is an essential process, useful in itself, and necessary as an intermediate step for almost every type of more complex text mining application.

We discuss some of the properties of the terminology for this domain, in particular the degree of ambiguity, which constitutes a peculiar problem for text mining applications. Without a correct recognition and disambiguation of the domain entities, no reliable results can be produced.

Finally, we discuss an application that makes use of the resulting terminological knowledge base. We annotate an existing corpus of sentences about protein interactions. The annotation consists of a normalization step that matches the terms in our resource with their actual representation in the corpus, and a disambiguation step that resolves the ambiguity of matched terms. The evaluation shows a precision of 57% and recall of 72%.

Key words:
Information Extraction, Text Mining, Terminological Resources, Biomedical Literature

* *Contact author*: Tel: +41 44 635 7132; Fax: +41 44 635 6809;
Email addresses: rinaldi@ifi.uzh.ch (Fabio Rinaldi), kaljurand@gmail.com (Kaarel Kaljurand), satre@idi.ntnu.no (Rune Sætre).

1 Introduction

The complexity of biological organisms and the recent progress 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 biological databases, e.g. MINT [1] [1] and IntAct [2] [2] (see [3] 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 databases. As the names that are used to reference the proteins can be very ambiguous, there is a need for an effective ambiguity resolution.

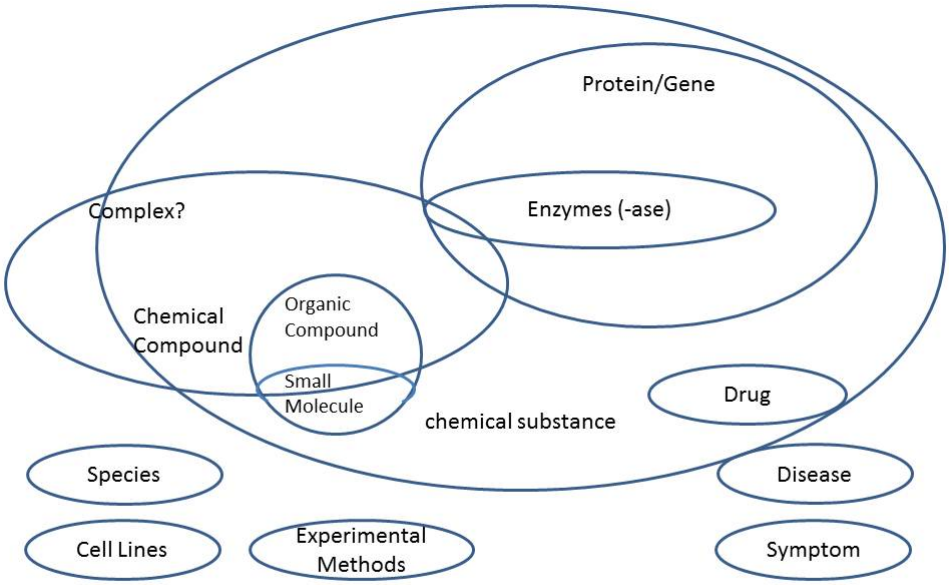
In this paper, we describe the task of automatically detecting names of various entities of relevance (e.g. proteins, genes, species, experimental methods, cell lines) in biomedical literature and grounding them to widely accepted identifiers assigned by standard Knowledge Bases (KB), such as the UniProt Knowledgebase (UniProtKB) [4], [3] the National Center for Biotechnology Information (NCBI) Taxonomy, [4] or the Proteomics Standards Initiative (PSI) Molecular Interactions (MI) Ontology [5]. [5]

The term annotation process is based upon a large term list that is compiled using the entity names extracted from the mentioned knowledge bases and from a list of cell line names. This resulting list covers the most common expressions for each term. A term normalization step is used to match the terms with their actual representation in the texts. Finally, a disambiguation step resolves the ambiguities (i.e. multiple IDs proposed by the annotator) among 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 module feed directly into the process of protein interaction detection. 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 [6]. The syntactic parser takes into account constituent boundaries defined by previously identified multi-word entities. Therefore the richness of the entity annotation has

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

Figure 1 Diagram showing the organization of the entities under consideration.



a direct beneficial impact on the performance of the parser, and thus leads to better recognition of interactions.

Additional input has been provided by the AGRA project,[7] which aims at ... (please complete). Resources such as Enzymes, Medical Terms, etc.. are relevant for/because... (please complete). These entities are related as described in Figure1.

This paper is structured in the following way. In section Section 2 we describe the terminological resources that we have used, in section Section 3 we describe the process of automatic annotation of biomedical texts using these resources, in section Section 4 we describe the evaluation method and results, in section Section 5 we review related work, and finally, in section Section 6 we draw conclusions and describe future work.

2 Term resources

As a result of the rapidly growing amount of available information in the field of biology, the research community has realized the need for consistently organizing the discovered knowledge - e.g. 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.

[7] <http://agra.fzv.uni-mb.si/>

Such lists can be used to annotate existing biomedical publications in order to identify the entities mentioned in these publications. In the rest of this section we describe some of these resources and how they can be used as a source of terminology for text mining purposes.

2.1 *Proteins and Genes*

There are several public protein/gene name knowledge bases available. Here we will look at four of them in more detail. They contain roughly the same information, but because they are made for slightly different purposes, there are small variations in the information registered for a "single protein". UniProt tries to capture all protein names from several species, while Entrez Gene focuses on the genes that these proteins are translated from. Affymetrix [8] is a company making Micro-Array chips for physical experiments, so they have to deal with another type of ambiguity between (fragments of) proteins that have similar physical properties. Among the resources hosted by the ExPASy (Expert Protein Analysis System) server at the Swiss Institute of Bioinformatics, one of particular interest for our work is the ENZYME database, [9] which describes special proteins that have an enzymatic effect inside our cells. These enzymes are typically named after the substances that they act on.

Since there are several common identifier systems in use, there are also several independent services available for mapping between the different identifiers. One example of such a system is the IdConverter [10] .

2.1.1 *UniProt KB*

The UniProt Knowledgebase (UniProtKB) assigns identifiers to a vast number of proteins and describes their amino-acid sequences. UniProt is organized in two section: SwissProt (manually curated) and TrEMBL (automatically derived). The experiments described in this paper are based on the XML version of the SwissProt section of UniProtKB version 14. 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). UniProt sees as one of its functions to help with the standardization of protein nomenclature and thus tries to cover all the common ways of referring to a protein [11] , while at the same time specifying a single name as "recommended name", following certain naming guidelines [12] . In addition, the

[8] <http://www.affymetrix.com/>
[9] <http://au.expasy.org/enzyme/>
[10] <http://idconverter.bioinfo.cnio.es/>
[11] <http://www.uniprot.org/faq/9>
[12] <http://www.uniprot.org/docs/nameprot>

Table 1 Frequency ranking of paths to XML elements that contain terms in UniProtKB.

Frequency	XPath
752,019	/uniprot/entry/gene/name
397,539	/uniprot/entry/protein/recommendedName/fullName
284,782	/uniprot/entry/protein/alternativeName/fullName
90,397	/uniprot/entry/protein/recommendedName/shortName
65,500	/uniprot/entry/protein/alternativeName/shortName
16,400	/uniprot/entry/protein/component/recommendedName/fullName
8913	/uniprot/entry/protein/domain/recommendedName/fullName
6339	/uniprot/entry/protein/component/alternativeName/fullName
5269	/uniprot/entry/protein/domain/alternativeName/fullName
5023	/uniprot/entry/protein/component/recommendedName/shortName
1416	/uniprot/entry/protein/CdAntigenName
1207	/uniprot/entry/protein/domain/recommendedName/shortName
1069	/uniprot/entry/protein/component/alternativeName/shortName
787	/uniprot/entry/protein/domain/alternativeName/shortName

names of functional domains and components of proteins, and also names of genes that encode the proteins are provided. UniProtKB attempts to cover proteins of all species. The top five species ranked by the number of their different proteins are *Homo sapiens* (Human) with 20,325 proteins, *Mus musculus* (Mouse) with 15,915, *Rattus norvegicus* (Rat) with 7170, *Arabidopsis thaliana* (Mouse-ear cress) with 6970, and *Saccharomyces cerevisiae* (Baker's yeast) with 6553. [13]

We extracted 626,180 (different) names from the UniProtKB XML file, using the XPath expressions listed in Table 1. 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"* [14] . 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 2 shows the orthographic/morphological properties of the names in UniProtKB in terms of how much certain types of characters influence the ambiguity. Non alphanumeric

[13] The amount of proteins for a species reflects the amount of research done on the given species, rather than the amount of proteins that the species has.
[14] <http://www.uniprot.org/docs/nameprot>

Table 2 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

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, i.e. they are very discriminative among different proteins. These findings motivate the normalization that we describe in section Section 3.2. Table 2 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.

2.1.2 Entrez Gene

Entrez Gene [15] is a gene-based resource supplying connections for map, sequence, expression, structure, functional and homology data. Entrez Gene assigns identifiers to 6.3 million genes, from 6440 different species, ranging from humans to viruses. The database is updated daily, and we have automatic processing in place in order to ensure that our dictionary is up to date.REMOVE THIS CLAIM?

Every Entrez Gene identifier is linked to the terms commonly used to describe the gene, including official name, nomenclature name, aliases and other designations. Each entry also has a short description which we use to extract more possible synonyms for a given term. Most of the entries contain 2 names for each gene, but the most studied proteins usually have around 5-10 synonyms. Each gene has an official reference term (ideally unique). However, because occasionally very simple reference terms are used, a degree of ambiguity remains. For example, *A* is the name of 27 genes (e.g. GeneID: 396713, *blood group A transferase-like*), *9* corresponds to 71 genes (e.g. GeneID: 920967. *9 tail spike protein*). On average, each of the extracted official gene terms corresponds to 1.2 identifiers.

2.1.3 Affymetrix Identifiers for Micro Array probes

Affymetrix produces Micro Array chips for all the proteins in several different species. Annotation information is stored in a single file for each array type, and that information is updated quarterly. In June 2010, there were 58 annotation files with references to SwissProt and Entrez Gene available. These files cover 1.5 million probe-sets, which are used to detect the presence of a protein in a biological experiment. Among these probe-set ids, 800,000 are

[15] <http://www.ncbi.nlm.nih.gov/gene>

mapped to SwissProt identifiers, and 860,000 are mapped to Entrez Gene identifiers. In the AGRA project, [16] Affymetrix probes have to be mapped to protein identifiers in order to use existing BioNLP systems like Facta+ [8]. [17]

As already mentioned, there is some ambiguity when mapping a probe (which usually contains just a part of a protein sequence) to a protein identifier. In the current version of Affymetrix, 456,000 (57%) of the SwissProt entries are ambiguous, but only 37,000 (4%) of the Entrez Gene entries. This reflects the fact that genes are closer to the probes used on the chip, than proteins are. The Affymetrix annotation tables also contains references into other common protein knowledge bases such as UniGene, Ensembl, EC, OMIM, RefSeq, FlyBase, AGI, WormBase, MGI, RGD, SGD, Gene Ontology, Pathway, InterPro, Trans Membrane and QTL.

2.1.4 Enzymes

Our list of enzymes was extracted from the ENZYME database hosted by the ExPASy proteomics server of the Swiss Institute of Bioinformatics (SIB). It contains 4188 identifiers with a total of 27,000 synonyms (6.5 pr. id). All identifiers have at least two synonyms, e.g. the enzyme commission (EC) number, and an "English" name. The highest synonym count is EC:2.7.10.1 (Drosophila Eph kinase) with more than 200 synonyms. Enzymes are usually named after their substrate (the chemical they catalyze), with the word ending -ase added. Because of this, a single protein may be associated with multiple EC numbers and multiple proteins may be associated with the same EC number.

2.2 Species

The National Center for Biotechnology Information provides a resource called NCBI Taxonomy [18] , describing all known species and listing the various forms of species names (e.g. scientific and common names). As explained in section Section 2.1.1, knowledge of these names is essential for disambiguation of protein names.

We compiled a term list on the basis of the taxonomy names list, [19] but kept only names whose ID mapped to a UniProtKB species "mnemonic code" (such as *ARATH* for *Arabidopsis Thaliana*). [20] 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

[16] <http://agra.fzv.uni-mb.si/>
[17] <http://refine1-nactem.mc.man.ac.uk/facta_events/>
[18] <http://www.ncbi.nlm.nih.gov/Taxonomy/>
[19] <ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz>
[20] <http://www.uniprot.org/help/taxonomy>

(e.g. 'Arabidopsis' was mapped to *ARATH*), as it is expected that e.g. 'Arabidopsis' alone is generally used to refer to *Arabidopsis thaliana*, and not to e.g. *Arabidopsis lyrata*. [21]

2.3 *Experimental Methods*

The Proteomics Standards Initiative (PSI) Molecular Interactions (MI) Ontology [22] 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 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 it is necessary to be aware of these names in order to avoid tagging them as e.g. protein names. Besides, almost every cell line comes from one species (although also "chimeric" cell lines are sometimes used), thus the mention of a cell line can give a useful hint of which species are central in a given document or document fragment.

We extracted 8741 cell line names from the Cell Line Knowledgebase (CLKB) [23] which is a compilation of data (names, identifiers, cell line organisms, etc.) from various cell line resources (HyperCLDB, ATCC, MeSH) [7]. 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

[21] We maintain a way to distinguish between terms that are sourced unambiguously from a single DB entry, and simplified terms introduced by us. Ultimately it is up to the application to decide in a given context which interpretation of a term is the most reliable.
[22] <http://psidev.sourceforge.net/mi/psi-mi.obo>
[23] <http://stateslab.org/data/CellLineOntology/>

(_CLKB). When processing CLKB, we ignored (did not follow) the MeSH ID cross references, which would have maybe provided additional synonyms (and possibly ambiguity).

Another interesting resource for cell line names is HyperCLDB [24] which covers 9 cell line catalogues with the total of about 5600 different cell line names. However, because it is more difficult to link this resource to NCBI, and because most if it appears to be included in CLKB anyway, we have so far restricted our source of cell line names to CLKB.

2.5 Small molecules and chemical compounds

This section of our dictionary (and the enzyme section as well) was largely derived from resources developed for the FACTA project [8]. The naming conventions for small molecules and compounds are in general more complex than those for protein and thus pose additional challenges to a text mining system. However, small molecules and chemical compounds are usually named after their structure, and there are also more guidelines available for proper naming than for proteins in general. The main source for chemical compounds is the ChEBI [25] and the CAS [26] databases. ChEBI, also known as Chemical Entities of Biological Interest, focuses on 'small' chemical compounds, and is a part of the Open Biomedical Ontologies effort. Unlike CAS, ChEBI does not contain many entities that are encoded by the genome.

The CAS registry contains a wide variety of substances, including the world's largest collection of organic and inorganic compounds, metals, alloys, minerals, organometallics, elements, isotopes, nuclear particles, proteins and nucleic acids, polymers, nonstructurable materials (UVCBs). The registry contains term identifiers, like "CAS:100-33-4", which is linked to synonyms for that substance, ranging from "Pentamide", via "C19H24N4O2" to " Benzenecarboximidamide, 4, {4'-[1,5-pentanediylbis(oxy)]bis-} ".

Our current list of compounds contains 129,000 identifiers, with a total number of 624,000 synonyms (4.8 pr. id). More than thousand of the identifiers have only one synonym (like CAS:991-56-0 is C31H16BrN3O7, or CAS:55508-42-4 is Neurotensin), while one identifier has 746 synonyms (CAS:95422-24-5, containing both ethyl and methyl in many different forms).

2.6 Medical terms

The domain of articles in PubMed is not only molecular interactions, but also the effects of these interactions on human health, including diseases, their symptoms, and medicines. Therefore we also extracted disease and symptom names from the Unified Medical Language System, [27] and medicine names from the DrugBank [28].

[24] <http://bioinformatics.istge.it/hypercldb/>
[25] Chemical Entities of Biological Interest: <http://www.ebi.ac.uk/chebi/>
[26] Chemical Abstract Service <http://cas.org/>
[27] <http://www.nlm.nih.gov/research/umls/>
[28] <http://www.drugbank.ca/>

2.6.1 Disease (UMLS)

The diseases part of UMLS covers 112,000 unique identifiers, with a total of 425,000 synonyms (3.8 pr. id). 28,000 identifiers have only one synonym (like UMLS:C1876206, LADD SYNDROME), while seven identifiers have more than 100 synonyms (like UMLS:C0851140, CIN 3).

2.6.2 Symptom (UMLS)

The disease symptoms part of the UMLS covers 6163 identifiers, and 23,000 synonyms (3.7 pr. identifier). 1800 identifiers have just one synonym (like UMLS:C1868968, *"Regurgitation of medication"*), while the most ambigous identifier has 63 synonym terms (UMLS:C0039070, *"Faint"*, *"syncope"*, *"swoon"* etc.)

2.6.3 Drug (DrugBank)

The DrugBank database [9] is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target information (i.e. sequence, structure, and pathway). The database contains nearly 4800 drug entries. Each DrugCard entry contains more than 100 data fields with half of the information being devoted to drug/chemical data. We extracted 9505 unique identifiers from the DrugBank. They are covered by 33,000 synonyms (3.5 pr. id). 3900 of the identifiers have one associated synonym, while one identifier can be represented by 320 terms (DrugBank:APRD00552 is B50, *"Biamine"*, *"Big Friends"*, etc.).

2.7 Compiled term list

We compiled a term list of 11,387,557 terms based on the terms extracted from all the mentioned resources, including for each entry the term name, the term ID, and the term type. 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 3. 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.

In the term list, 746,226 of the terms are multi-word units (e.g. 258,835 contain two tokens, 189,948 three tokens, and about 1000 terms even more than 10 tokens). We did not normalize the names to any canonical representation nor generate all possible spelling variations of the names. Our text mining system includes a 'normalization' module which is used once when the terms are stored into an internal data structure, and again when candidate terms from the document are processed. This approach is capable of recognizing as equivalent terms that differ in a number of surface details like spacing, hyphenation, etc.

Table 3 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')

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 [29] tokenizer and sentence splitter which have been trained on 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 non-overlapping sequences of tokens *"starting at any given point"*, and 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 about 50 stop words), as well as figure and table references (e.g. 'Fig. 3a' and 'Table IV').

3.1 Preprocessing

In the preprocessing step the input text is transformed into XML (if needed) and a set of linguistic annotations are applied using LingPipe [30], namely: sentence splitting, tokenization, and part-of-speech tagging.

The LingPipe part-of-speech tagger has been trained on biomedical texts and also the sentence splitter and tokenizer are aware of the nature of biomedical texts. E.g. the tokenizer provides a very granular tokenization as often two protein names are hyphenated together and as such they should be split up. We have modified the LingPipe sentence splitter by adding a list of abbreviations commonly found in species names (e.g. 'sp.', 'subsp.'). Note

[29] <http://alias-i.com/lingpipe/>
[30] <http://alias-i.com/lingpipe/>

that at the moment the subsequent steps in our term annotation pipeline ignore the part-of-speech information. This information, however, is used by the syntactic parser to detect syntactic dependencies between tokens.

3.2 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 [10,11], e.g.

- Remove all characters that are not alphanumeric or space
- Remove lowercase-uppercase distinction
- Normalize Greek letters and Roman numerals, e.g. 'alpha' $\rightarrow$ 'a', 'IV' $\rightarrow$ '4'
- Remove hyphens if between alphanumeric strings
- Remove the final 'p' if it follows a number, e.g. 'Pan1p' $\rightarrow$ '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. This is probably due to the fact that authors are not very good at respecting such conventions, and often inadvertently introduce minor variants.

3.3 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 Section 2.3). We disambiguate such terms by removing all the UniProtKB IDs. (Similar filtering is performed in [12].) 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 [13] 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. [31] 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 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 Section 3.3, i.e. some or all of the IDs were removed. Results before and after disambiguation are presented in Table 4. 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.

[31] <ftp://ftp.ebi.ac.uk/pub/databases/intact/current/various/data-mining/>

Table 4 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

5 Related work

There is a large body of work in named entity recognition in biomedical texts. Mostly this work focuses on the detection of the entity mentions in the text and does not cover the second aspect of the problem, namely grounding the detected named entities to existing knowledge base identifiers. Recently, however, as a result of the BioCreative workshop, more approaches are extending from just detecting entity mentions to their normalization to a standardized version of the term, or to a database identifier. The most common application is the normalization of gene names, as practiced in the Gene Normalization tasks of the recent BioCreative competition [14], using Entrez Gene as the reference database.

Various web services have become recently available which provide term detection and grounding in arbitrary text provided by the user. One of the outcomes of the BioCreative II

competition was a web-based tool called the BioCreative MetaServer (BCMS) [15], [32] which combines the results from different remote servers to provide an 'harmonized' entity annotation service, including gene and protein normalization. Another well-known annotation service is Whatizit,[16] [33] a webservice which annotates input texts with UniProtKB, Gene Ontology, [34] and NCBI terms. A preliminary comparison showed that our approach gives results of similar quality. Harvester [17] [35] is a service that crosslinks many bioinformatic sites with protein information. They make their results available on their homepage, but unlike this project, it is not possible to download the resources in the form of dictionaries for use in other programs.

[18] provides a thorough overview of terminological resources for chemical compounds (PubChem [36] , KEGG [37]), and describes a machine learning approach (using conditional random fields) to detect names of chemical compounds in biomedical literature. [19] describes a computationally intensive approach towards the detection of terminological variants for the purpose of correct term matching and disambiguation. A method of protein name grounding is described in [11]. 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.

JoChem [20] is an example of a dictionary that focus on a narrow subset of the classes that we cover in this paper. Their dictionary is available in two different formats, including an XML file in the Simple Knowledge Organization System format on their web site. [38] In order to standardize the dictionary efforts, we also make our dictionary available in the same format as the JoChem dictionary. DO WE WANT TO KEEP THIS CLAIM?

Several linguistic resources have been compiled from existing biomedical databases. The BioThesaurus [21] [39] is a thesaurus of gene and protein names (and their synonyms and textual variants). The latest version (6.0) of BioThesaurus contains more than 9 million names, extracted from 35 different databases. The biggest contributor, however, is UniProtKB, mainly its TrEMBL section. Another extensive compilation of biomedical terms is the BioLexicon [22]. The BioLexicon aims at being a linguistically-rich resource, containing not only information about terminology, but also verbs, adjectives and adverbs which are of relevance for biomedical text processing. Additionally, it includes semantic links among the terms, such as derivations relations. The domain relevant verbs (658) are further specified by subcategorization frames (1710), as well as semantic event frames (850) which increase the range of possible usages of this resource. Originally compiled within the scope of the BOOTStrep European Project [23], the BioLexicon is now available through the European Language Resources Association (ELRA).

PathText [24] is an interesting example of an advanced application which combines text mining technologies with a sophisticated visualization approach for the graphical rendering of

[32] <http://bcms.bioinfo.cnio.es>
[33] <http://www.ebi.ac.uk/webservices/whatizit/>
[34] <http://www.geneontology.org>
[35] <http://harvester.fzk.de/harvester/>
[36] <http://pubchem.ncbi.nlm.nih.gov>
[37] <http://www.genome.jp/kegg/>
[38] <http://www.biosemantics.org/chemlist>
[39] <http://pir.georgetown.edu/iprolink/biothesaurus/>

biological pathways, linking each relation with its supporting information in the biomedical literature.

6 Conclusions and future work

In this paper we presented a large terminological resource, compiled through the aggregation of a number of different manually-curated sources. We also presented results related to the lexical properties of such resources, specifically the degree of ambiguity of the terms, and we inspected the causes of such ambiguity, in particular for protein names. This information is of vital importance for the implementation of an efficient term normalization and grounding algorithm.

We believe that our harmonized terminological knowledge base constitutes a valuable resources for all research groups involved in biomedical text mining.The current version is available at <http://www.ontogene.org/resources/>. We are constantly monitoring and updating the terminological resources included in our system. Additional resources are added when judged of sufficient quality.

Additionally, we presented the usage of the terminology in an application focusing on the detection of protein-protein interactions. For the evaluation, we have used the freely available IntAct corpus of snippets of textual evidence for protein-protein interactions. We show results which are certainly competitive. The same approach has been recently used within the BioCreative II.5 task of protein-protein detection, where it has obtained the best results [25].

7 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 activity of R.S. is partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan) and Genome Network Project (MEXT, Japan).

References

[1] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, G. Cesareni, MINT: a Molecular INTeraction database, FEBS Letters 513 (1) (2002) 135–140. 1

[2] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, R. Apweiler, IntAct: an open source molecular interaction database, Nucl. Acids Res. 32 (suppl 1) (2004) D452–455. 1

[3] S. Mathivanan, B. Periaswamy, T. Gandhi, K. Kandasamy, S. Suresh, R. Mohmood, Y. Ramachandra, A. Pandey, An evaluation of human protein-protein interaction data in the public domain, BMC Bioinformatics 7 (Suppl 5) (2006) S19. 1

[4] UniProt Consortium, The universal protein resource (uniprot), Nucleic Acids Research 35 (2007) D193–7. 1

[5] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, C. Grant SG, Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, I. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, A. R., The hupo psi's molecular interaction format - a community standard for the representation of protein interaction data, Nat. Biotechnol 22 (2004) 177–183. 1

[6] F. Rinaldi, T. Kappeler, K. Kaljurand, G. Schneider, M. Klenner, S. Clematide, M. Hess, J.-M. von Allmen, P. Parisot, M. Romacker, T. Vachon, OntoGene in BioCreative II, Genome Biology 9 (Suppl 2) (2008) S13. 1

[7] S. Sarntivijai, A. S. Ade, B. D. Athey, D. J. States, A bioinformatics analysis of the cell line nomenclature, Bioinformatics 24 (23) (2008) 2760–2766. 2.4

[8] Y. Tsuruoka, J. Tsujii, S. Ananiadou, FACTA: a text search engine for finding associated biomedical concepts, Bioinformatics 24 (21) (2008) 2559–2560. 2.1.3, 2.5

[9] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets., Nucleic Acids Res 36 (Database issue) (2008) D901–6. 2.6.3

[10] J. 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. 3.2

[11] X. Wang, M. Matthews, Distinguishing the species of biomedical named entities for term identification, BMC Bioinformatics 9 (Suppl 11) (2008) S6. 3.2, 5

[12] L. Tanabe, W. J. Wilbur, Tagging gene and protein names in biomedical text, Bioinformatics 18 (8) (2002) 1124–1132. 3.3

[13] T. Kappeler, K. Kaljurand, F. Rinaldi, TX Task: Automatic Detection of Focus Organisms in Biomedical Publications, in: BioNLP 2009, NAACL/HLT, Boulder, Colorado, 2009. 3.3

[14] A. Morgan, Z. Lu, X. Wang, A. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H.-h. Liu, R. Torres, M. Krauthammer, W. Lau, H. Liu, C.-N. Hsu, M. Schuemie, K. Cohen, L. Hirschman, Overview of BioCreative II gene normalization, Genome Biology 9 (Suppl 2) (2008) S3. 5

[15] F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J. Hakenberg, C. Plake, C.-J. Kuo, C.-N. Hsu, R.-H. Tsai, H.-C. Hung, W. Lau, C. Johnson, R. Saetre, K. Yoshida, Y. Chen, S. Kim, S.-Y. Shin, B.-T. Zhang, W. Baumgartner, L. Hunter, B. Haddow, M. Matthews, X. Wang, P. Ruch, F. Ehrler, A. Ozgur, G. Erkan, D. Radev, M. Krauthammer, T. Luong, R. Hoffmann, Introducing meta-services for biomedical information extraction, Genome Biology 9 (Suppl 2) (2008) S6. 5

[16] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, A. Jimeno, Text processing through Web services: calling Whatizit, Bioinformatics 24 (2) (2008) 296–298. 5

[17] U. Liebel, B. Kindler, R. Pepperkok, 'Harvester': a fast meta search engine of human protein resources, Bioinformatics 20 (12) (2004) 1962–1963. 5

[18] C. Kolarik, R. Klinger, C. M. Friedrich, M. Hofmann-Apitius, J. Fluck, Chemical Names: Terminological Resources and Corpora Annotation, in: Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference), Marrakech, Morocco, 2008. 5

[19] Y. Tsuruoka, J. McNaught, J. Tsujii, S. Ananiadou, Learning string similarity measures for gene/protein name dictionary look-up using logistic regression, Bioinformatics 23 (20) (2007) 2768–2774. 5

[20] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. M. Hendriksen, B. J. A. Schijvenaars, E. M. v. Mulligen, J. Kleinjans, J. A. Kors, A dictionary to identify small molecules and drugs in free text, Bioinformatics 25 (22) (2009) 2983–2991. 5

[21] H. Liu, Z.-Z. Hu, J. Zhang, C. Wu, BioThesaurus: a web-based thesaurus of protein and gene names, Bioinformatics 22 (1) (2006) 103–105. 5

[22] Y. Sasaki, S. Montemagni, P. Pezik, D. Rebholz-Schuhmann, J. McNaught, S. Ananiadou, Biolexicon: A lexical resource for the biology domain, in: Proc. of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), 2008. 5

[23] E. Buyko, S. Piao, Y. Tsuruoka, K. Tomanek, J.-D. Kim, J. McNaught, U. Hahn, J. Su, S. Ananiadou., Bootstrep annotation scheme: Encoding information for text mining, in: Corpus Linguistics 2007 - Proceedings of the 4th Corpus Linguistics Conference, 2007. 5

[24] B. Kemper, T. Matsuzaki, Y. Matsuoka, Y. Tsuruoka, H. Kitano, S. Ananiadou, J. Tsujii, PathText: a text mining integrator for biological pathway visualizations, Bioinformatics 26 (12) (2010) i374–381. 5

[25] F. Rinaldi, G. Schneider, K. Kaljurand, S. Clematide, T. Vachon, M. Romacker, OntoGene in BioCreative II.5, IEEE/ACM Transactions on Computational Biology and Bioinformatics 7 (3) (2010) 472–480. 6