Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4585–4596
Marseille, 11–16 May 2020

PROGENE — A Large-scale, High-Quality Protein-Gene Annotated Benchmark Corpus

Erik Faessler, Luise Modersohn, Christina Lohr and Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab,
Friedrich-Schiller-Universität Jena, Jena, Germany
SMITH Consortium of the German Medical Informatics Initiative
{erik.faessler | luise.modersohn | christina.lohr | udo.hahn}@uni-jena.de

Abstract

Genes and proteins are the fundamental entities of molecular genetics and deeper knowledge about their interactions constitutes a cornerstone for advancing precision medicine. We here introduce PROGENE (formerly called FSU-PRGE), a corpus that reflects our efforts to cope with this important class of named entities within the framework of a long-lasting large-scale annotation campaign at the Jena University Language & Information Engineering (JULIE) Lab. We partitioned the entire corpus into 11 subcorpora covering various biological domains to achieve an overall subdomain-independent corpus. It consists of 3,308 MEDLINE abstracts with over 36k sentences and more than 960k tokens annotated with nearly 60k named entity mentions. Two annotators strove for carefully assigning entity mentions to classes of genes/proteins as well as families/groups, complexes, variants and enumerations of those where genes and proteins are represented by a single class. The main purpose of the corpus is to provide a large body of consistent and reliable annotations for supervised training and evaluation of machine learning algorithms in this relevant domain. Furthermore, we provide an evaluation of two state-of-the-art baseline systems—BIOBERT and FLAIR—on the PROGENE corpus. We make the evaluation datasets and the trained models available as a benchmark to encourage comparable evaluations of new methods in the future.

Keywords: Genes and proteins, text corpus, annotation, biomedical corpus, evaluation

1. Introduction

Genes and proteins are the fundamental entities of molecular genetics and deeper knowledge about their interactions constitutes a cornerstone for advancing precision medicine. Hence, they attracted the interest of the biomedical NLP community for a long time already, with focus on two information extraction tasks. The first one, *named entity recognition and grounding*, aims at locating gene and protein mentions in a document and, in a second step, mapping such mentions onto some well-established identifier system or name space, such as provided by Gene Ontology (GO) (Gene Ontology Consortium, 2001), UNIPROT (UniProt Consortium, 2017), NCBI GENE (Brown et al., 2015) or KEGG (Kanehisa et al., 2012). Both tasks were particularly featured in several iterations of the BIOCREATIVE Challenge (Hirschman et al., 2005; Krallinger et al., 2008; Arighi et al., 2011; Mao et al., 2014). The second task aims at *extracting relations* between genes or proteins in terms gene/protein interactions from documents (as investigated in several iterations within the BioCreAtIvE (Krallinger et al., 2008; Leitner et al., 2010; Arighi et al., 2011; Doğan et al., 2017) and BioNLP relation extraction challenges (Kim et al., 2009; Pyysalo et al., 2011; Kim et al., 2011; Pyysalo et al., 2012; Kim et al., 2013; Nédellec et al., 2013; Bossy et al., 2015)). Also of interest to the biomedical NLP community are other types of gene-induced relations (as witnessed by corresponding challenge tasks) like gene–disease relations (Pyysalo et al., 2015; Wang et al., 2019) or gene-chemical interactions (Krallinger et al., 2017).

For both kinds of tasks, annotated corpora are needed—either as repositories from which *training* data for classifiers can be drawn in a supervised learning mode, or as gold standards for *evaluating* the performance of named entity or relation taggers (Rebholz-Schuhmann et al., 2013). Yet, existing corpora differ greatly in size, thematic focus, annotation quality, granularity of the underlying conceptual entity representation and the way individual entity classes are defined—even when different corpora cover the same entity classes. Thus, merging all available annotations into one large corpus and regard it as a coherent source of gene/protein annotations (Wang et al., 2009; Wang et al., 2010; Galea et al., 2018) might not be advisable. This stresses the need for large-scaled, consistently annotated and quality-checked corpora for specific entity classes.

We here describe such a large-scale protein annotation campaign of MEDLINE abstracts and introduce the PROGENE corpus, with special emphasis on biologically reliable and annotation-wise consistent metadata. An earlier version of that corpus is described by Hahn et al. (2010) and has community-widely been referred to as the FSU-PRGE corpus (cf., e.g., Habibi et al. (2017; Dang et al. (2018)). However, the FSU-PRGE version did not offer fully detailed annotation levels but collapsed the largest part of the available annotations in the single *protein/gene (PrGe)* class for reasons of simplicity. In PROGENE, we now enrich the whole corpus with annotation levels for genes/proteins, families/groups, complexes, variants and enumerations in their original annotation format. The new corpus version is available at DOI 10.5281/zenodo.3698568.

2. A Survey of Gene/Protein Annotation

In this section, we give a chronological overview of existing corpora with focus on gene or protein annotations summarized in Table 1. Unless stated otherwise, the listed corpora use a single annotation level to manually[1] mark occurrences

[1]Hence, we here exclude an alternative stream of work on so-called silver standards (Rebholz-Schuhmann et al., 2010; Sousa et al., 2019) where annotations are derived using automatic taggers.

Name	Text Type	# Texts	# Sentences	# Tokens	# Genes	# Relations
Named Entity Focused (Genes/Proteins)						
GENIA v3.02	abstracts	2,000	20,546	472,006	30,269	n/a
JNLPBA	abstracts	2,404	24,806	568,786	35,366	n/a
GENETAG	sentences	20,000	20,000	547,801	23,996	n/a
PRODISEN	abstracts	2,466	21,000	469,000		n/a
OSIRIS	abstracts	105	1,043	28,697	768	n/a
AIMED PROTEINS	abstracts	748	7,785	195,396	5,287	n/a
PENNBIOIE	abstracts	2,514	14,305	357,313	17,427	n/a
CRAFT v4.0	full texts	97	30,830	793,651	23,578	n/a
CELLFINDER	full texts	10	2,177	65,031	1,621	n/a
IGN	abstracts	543	4,807	126,368	5,948	n/a
GNORMPLUS	abstracts	694	6,583	168,853	10,639	n/a
PHARMACONER	clinical reports	1,000		396,988	3,009	n/a
PROGENE	abstracts	3,308	36,223	960,757	59,514	n/a
Relation Focused (with Genes/Proteins as Arguments)						
IEPA	various	200	243	151,74	1109	335
LLL	sentences	77	77	1,496	117	165
ITI/TXM	full texts	455	137,400	3,900,000	16,1448	44,686
BIOINFER	sentences	1,100	1,100	33,858	6,349	2,662
AIMED INTERACTIONS	abstracts	225	2,202	59,700	4,227	1,069
GENEREG	abstracts	314	3,125	82,747	1,971	1,770
VARIOME	full texts	33	6,051	172,987	4,613	12,885
BIONLP 2009	abstracts	1,210	11,346	267,450	14,969	Task 1: 13,588 Task 2: 13,623
BIONLP 2011 GE TASK	abstracts	1,210	11,346	267,450	14,969	13,603
	full texts	14		80,962	6,580	4,444
BIONLP '13/'16 GE TASK	full texts	34		187,989	12,068	9,364
AGAC	abstracts	500	5,080		1,154	1,514
BIOCREATIVE VI	abstracts	597	5,724	149,469	8,833	760
PGXCORPUS	sentences	945	945	29,016	1,762	2,875

Table 1: Text corpora with gene/protein annotations ('#' stands for 'number of', 'n/a' for 'not applicable') in the order of publication year. The number of texts relates to the document sort making up the corpus, i.e. abstracts, full texts, sentences or other. The number of relations refers to all relations in the corpus, not just those that incorporate genes or proteins.

of genes and gene products, such as proteins or mRNA, in texts. It is common practice in biomedical annotation to create a more abstract annotation class as an umbrella for all entities that relate to specific parts of a DNA sequence due to the difficulties that arise in the attempt of a more fine-grained approach (Ohta et al., 2009). Especially the difference between the mentions of actual gene sequences as opposed to the expressed proteins of the same gene is sometimes difficult to distinguish, even for human experts (Hatzivassiloglou et al., 2001). The statistics on token, sentence, gene/protein and annotation counts given for each corpus in this section are taken from the respective official publication, the overview table of Habibi et al. (2017)[2] or computed by ourselves, in this order of availability.

The structure of this section is as follows. We define two groups of corpora. The first are those that are annotated for genes and/or proteins – possibly amongst other entity classes – without primarily disclosing relation information about the annotated entities. The second group consists of corpora that ultimately aim for the annotation of relations between gene/protein entities. Both groups are internally ordered chronologically by first release of the corpus or their publication year. In this manner, this section serves as a historical overview of gene/protein corpus creation (for alternative surveys of text corpora annotated with gene/protein metadata, cf. Habibi et al. (2017); Pyysalo et al. (2008) compare five early protein relation corpora—AIMED, BIOINFER, HPRD50, IEPA, and LLL).

Proteins or Genes.

The **GENIA** corpus (Ohta et al., 2002; Kim et al., 2003) was the first large corpus annotated for biomedical entities containing about 472k tokens. Its documents were gathered from a MEDLINE search for "*human*, *blood cells* and *transcription factors*," thus defining the thematic focus of the corpus. GENIA is a notable exception of the otherwise common practice to collapse genes, gene products and other related concepts into a single class. It uses an ontology of 47 classes including `proteins` and `DNA` which both exhibit a number of subclasses like `complex` or `subunit` (proteins) and `family_or_group` or `domain_or_region`

[2]Even though they applied their own segmentation, the numbers they found are close to the official ones when available.

(DNA) the latter of which is roughly used to denote genes in the annotated text, as long as the physical instance of DNA is referred to. However, this fine annotation granularity comes at the cost of lowered annotation consistency (Ohta et al., 2009). To alleviate this issue and to make GENIA comparable to annotation schemes using higher-level concepts, such as the GENETAG corpus (see below), GENIA was reannotated with the `GGP (gene or gene product)` class (Ohta et al., 2009). This follow-up version contains 12k GGP annotations and 15,5k annotations of the original protein class. The GENIA corpus also includes a subset annotated for protein-protein-interactions, the GENIA interaction corpus that was used for the BIONLP Shared Tasks in 2009 (Kim et al., 2009) and 2011 (Kim et al., 2011) (see below).

The **JNLPBA** challenge corpus (Kim et al., 2004) comprises all of GENIA version 3.02 as its training set. The corpus added another 404 documents with nearly another 97k tokens annotated for a wide variety of entities of importance in molecular biology. Taken together, the corpus features 569k tokens and 35k annotations for genes or proteins. There is a revised, cleaned version of this corpus available (Huang et al., 2019) that (according to the authors) is more consistently annotated than the original one.

Another early corpus containing gene and protein annotations, also of a considerable size, is **GENETAG** (Tanabe et al., 2005). It comprises nearly 550k tokens and 24k annotations of genes or proteins. The documents constituting the corpus were selected by automatically estimating the probability of MEDLINE abstracts containing gene mentions or not. By applying the same method to the sentences of those abstracts exactly 20,000 sentences, half of them chosen from the highest-ranking sentences and the other half from the lowest-ranking sentences, were ultimately sampled for the corpus. In effect, no specific biological domain was targeted. The corpus was used for the BIOCREATIVE I (Yeh et al., 2005) (15,000 sentences) and II (Smith et al., 2008) (all 20,000 sentences) challenges on gene mention recognition. GENETAG took another route to gene/protein annotation than GENIA and created an umbrella class for genes and gene products as was described at the beginning of the section and later also adopted for the GENIA corpus.

Whereas the corpora discussed so far assigned an annotation span to gene/protein information, **PRODISEN** is agnostic to such span information. Recognizing the issue of thematic disbalance in existing annotated biomedical corpora, PRODISEN is an approach to sample PUBMED on a large scale with as little topic bias as possible. For the Protein Description Sentence (PRODISEN) corpus (Krallinger et al., 2006), randomly selected PUBMED articles were screened by biological experts who had to classify sentences into three categories: whether they explicitly or implicitly contained information about proteins (and genes), or did not contain such descriptions, or whether the experts were unsure about making such a decision. The first class was additionally classified in sentences containing information on relevant aspects of a gene, gene product, gene group, protein family or protein domain based on the analysis of the contextual information. Altogether, PRODISEN incorporated 21k sentences extracted from 2,466 abstracts, with 469k tokens. The corpus was split into two halves, one containing the randomly selected abstracts, the other one enriched by citation overlap.

The **OSIRIS** (Furlong et al., 2008) system finds mentions of Single Nucleotide Polymorphisms (SNPs), i.e., gene variations, in the scientific literature. For its evaluation, a corpus of 29k tokens with 768 gene/protein annotations (Habibi et al., 2017) was created.

The **AIMED** corpus (Pyysalo et al., 2008) contains two parts, one annotated purely for genes/proteins, AIMED PROTEINS, the other for protein-protein-interactions, AIMED INTERACTIONS (for details, cf. Table 1). The documents are focused on the human genome. Taking both parts together, AIMED is composed of 9987 sentences, 255k tokens and contains 9514 gene/protein annotations.

The **PENNBIOIE** corpus,[3], version 1.0, consists of two thematically disparate parts, one dealing with oncology, the other with the inhibition of the CYP450 enzyme. The whole corpus comes with 357k tokens and about 17k annotations (Habibi et al., 2017) of genes and gene-related entities in the oncology subset and `substance` annotations in the CYP450 subset that mostly refer to proteins.

In contrast to using abstracts from MEDLINE, the **CRAFT** corpus takes full texts from PUBMED CENTRAL. The original release that was described in Bada et al. (2012) contained 67 full text articles annotated for a range of entities including genes/proteins. For the BIONLP OST 2019 CRAFT Task (Baumgartner et al., 2019), another 30 articles were added as test data, amounting to 97 full texts. For the statistics given in this paper, especially in Table 1, we downloaded release 4.0.0 from GITHUB[4] and extracted the *PoS tags and sentences* and the PR(OTEIN) levels with the script offered in the download. The corpus contains nearly 800k token and 24k gene/protein annotations.

CELLFINDER (Neves et al., 2012) is also based on full-texts (10 documents, 2,177 sentences), with a focus on stem cell research. It offers annotations for anatomy, cell components, cell lines, cell types, genes and species mentions, including 65,031 tokens with 1,621 gene/protein annotations.

Two corpora originated from the BIOCREATIVE II Challenge on Gene Normalization, the Instance-Level Gene Normalization corpus **IGN** (Dai et al., 2013) and the **GNORMPLUS** corpus (Wei et al., 2015), mainly composed of the MEDLINE abstracts used in that challenge. Both extend the original annotation data, basically lists of NCBI Gene identifiers for the human genes mentioned in the abstracts, with mention-level annotations rooted in the abstract text by providing the respective character offsets. Both corpora also provide the NCBI GENE identifiers for the mentions with a varying level of completeness; the GNORMPLUS data also contain annotations for gene families or groups and domain motifs. To this point, both corpora consist of 543 abstracts, with 4,8k sentences and 126k tokens. The GNORMPLUS corpus contains additional 151 documents from the Citation GIA test collection.[5]

[3] `https://catalog.ldc.upenn.edu/LDC2008T21` `https://catalog.ldc.upenn.edu/LDC2008T20`

[4] `https://github.com/UCDenver-ccp/CRAFT/tree/v4.0.0`

[5] `https://ii.nlm.nih.gov/TestCollections`

The **PHARMACONER** corpus evolved from the BIONLP OST 2019 PHARMACONER Task (Gonzalez-Agirre et al., 2019) and is unique for several reasons. Unlike all the other English-language corpora discussed here, it features Spanish language data and primarily deals with chemical compounds and drugs, but it also carries 3,009 protein annotations. It consists of a manually classified collection of 1,000 clinical case report sections (397k tokens) derived from open access Spanish medical publications, named the Spanish Clinical Case Corpus (SPACCC).[6]

Relations Involving Proteins or Genes as Arguments.

The **IEPA** corpus (Ding et al., 2002) was created to represent a diverse set of interactions between chemicals, mostly proteins. The main goal was to compare relation extraction efforts on different sizes of textual units. The corpus contains 243 sentences, with 15k tokens. There are 1,109 annotations of genes or proteins whose names occur in a list of 16 gene/protein names (Pyysalo et al., 2008). While the original documents were drawn from MEDLINE abstracts, the corpus documents themselves represent diverse text spans, from single sentences to whole abstract text bodies which describe the interaction between two entities.

The **LLL** corpus was created for the Learning Language in Logic – Genic Interaction Extraction Challenge (Nédellec, 2005) and comes with 77 sentences involving protein-gene interactions in *Bacillus subtilis*. The annotations consist of agent-target pairs where agents are proteins and targets are genes resulting in 117 annotations of genes/proteins belonging to a list of 116 names (Pyysalo et al., 2008).

The **ITI TXM** corpus (Alex et al., 2008) annotation effort resulted in two large biomedical corpora. The first corpus copes with protein-protein-interactions (PPI), the second focuses on tissue expressions (TE). Both corpora are larger than all other corpora we discuss here, including our own, with 2M and 1.9M tokens, respectively. They feature a `protein` annotation level due to its importance for the relation types in focus, and also contain a number of other entity types that may also play a role in those relations like `complex`, `mRNAcDNA`, `disease` and others (altogether 15 entity classes). For the TE corpus only, a `gene` annotation level was added. The PPI corpus was selected from full texts containing keywords pointing at protein-protein-interactions such as `bind`, `interact` etc. and contains 89k protein annotations, whereas the TE corpus holds 61k protein and 12k gene annotations. However, The combined corpus was not publicly available as of March 2020.

The **BIOINFER** corpus (Pyysalo et al., 2007) was created by searching PUBMED for pairs of proteins known for their interactions. From the found articles, the abstracts were searched for occurrences of these pairs which resulted in 1,100 sentences (about 34k tokens) in the original version. For the annotation process, BIOINFER builds on the GENIA physical type and relation ontology and each sentence is separately assigned annotations at the entity, relationship, and (syntactic) dependency level. Entity types include physical entities, such as individual genes, proteins, protein families and compexes, or RNA, processes (e.g., phosphorylation) and properties associated with entity states, e.g., amount, location, function, dynamics, and physical state. The relationship ontology covers four major classes, namely partonomic `part-of` and taxonomic `is-a`, (experimental) observations, and causal relations. BIOINFER contains 6,349 entity and 2,662 relationship annotations.

The **GENEREG** corpus (Buyko et al., 2010) consists of 314 PUBMED abstracts dealing with the regulation of gene expression in the model organism *E. coli*. Annotation is based on nine categories referring to the Gene Regulation Ontology (GRO) (Beisswanger et al., 2008), viz. `Gene Expression`, `Transcription`, `Regulation of Gene Expression (ROGE)`, `Positive ROGE`, `Negative ROGE`, and `Experimental Intervention`, with subtypes `Genetic Modification`, `Artificial Increase`, and `Artificial Decrease`. GENEREG comes with 1,770 relation annotations that are linked to the GENIA corpus, as well as to biomedical and general language domain lexicons. There is also an overlap with PROGENE for 149 documents, yet gene annotations in GENEREG focus more on additional entity types (such as transcription factors) than on finer granularity, the goal of PROGENE.

The **VARIOME** corpus focuses on "human genetic variation and its relationship to disease" (Verspoor et al., 2013). Its annotations of genes, body parts, patient cohorts and other clinically relevant types culminate in the annotation of relationships between those entities to express that patients have a particular genetic variation and how this relates to diseases the patients have. The corpus is built from 33 PUBMED CENTRAL full texts with 173k tokens, 6,051 sentences and 4,613 gene/protein annotations.

The **BioNLP Shared Task** on Event Extraction issues a series of continuously updated and enhanced relation-centered corpora aiming at the construction of an NFκb knowledge base. Starting with the first challenge in 2009, the complete reference corpus (including training, development and test data) (Kim et al., 2009) is composed of 1,210 PUBMED abstracts, with roughly 11k sentences or 267k tokens. It is based on the GENIA event corpus (Kim et al., 2008) and contains 13.6k annotations of nine different event types taken from the GENIA event ontology (`Gene expression`, `Transcription`, `Protein catabolism`, `Phosphorylation`, `Localization`, `Binding`, `Regulation`, `Positive` and `Negative regulation`). Gene/protein annotations were taken from the modified GENIA corpus (Ohta et al., 2009).

The follow-up event featured the GENIA Event (GE) Task 2011(Kim et al., 2011) using the same nine relation types (and annotations) as in 2009, but added to the abstract portion from 2009 a new full-text segment composed of 14 articles (with 82k tokens) , thus adding 6,580 annotations for genes and proteins and 4,444 relation annotations.

The third edition of this shared task in 2013 came up with a new, more recent full-text-only corpus extracted from the Open Access subset of PUBMED CENTRAL (Kim et al., 2013). Four new event types were added; `Protein modification` and its three sub-types, `Ubiquitination`, `Acetylation` and `Deacetylation`. Furthermore, the `Protein modification` types were modified such that they were

[6] `https://github.com/PlanTL-SANIDAD/SPACCC`

directly linked to causal entities, which was only possible through `Regulation` events in previous editions. This corpus comprises 34 full texts (14 were taken over from the 2011 campaign), with 188k tokens, 12k protein and 9,4k event annotations.

For the fourth edition of the BIONLP Shared Task in 2016 the original corpus from the third round was cleaned and further augmented by assignments of UNIPROT IDs for named entity grounding (Kim et al., 2016).

In the 2019 edition of the BIONLP Shared Task, Wang et al. (2019) introduced the **Active Gene Annotation Corpus (AGAC)** for the task of drug repurposing. They collected 500 MEDLINE abstracts, with slightly more than 5k sentences, using the Mesh terms *"Mutation/physiopathology"* and *"Genetic Disease"* and annotated AGAC with four annotators for eleven types of named entities, which were categorized into bio-concepts (e.g., Variation or Pathway), regulation types, and other entities (among them Disease, Enzyme, and 1,1k mentions of Gene/Protein), as well as two types of thematic relations between them.

For the latest Precision Medicine Track in **BioCreative VI**, Doğan et al. (2017) created two corpora, one with relevance judgments for abstracts in the precision medicine domain, the second annotated with PPIs that are affected by a mutation. Other interaction relations have not been annotated for this corpus. The gene/protein arguments of the annotated PPIs are also marked in the corpus and labeled with their NCBI GENE ID. We determined the statistics about the relation corpus reported in Table 1 from the download of the relation training dataset from the BIOCREATIVE website[7] in XML format. We extracted the titles and abstracts from the XML documents and counted 5,724 sentences with nearly 150k tokens. From the XML annotations, we gathered 1,762 gene annotations and 2,875 mentions of relations.

The **PGXCORPUS** (Legrand et al., 2020) focuses on the pharmacogenomics domain. The corpus download contains 945 tokenized sentences taken from 911 PUBMED abstracts (roughly 30k tokens). 1708 (plus 54 genomic variation) annotations relate to the level of `Gene_or_protein`, besides the `Chemical`, `Phenotype` and more general `Genomic_factor` annotation classes (overall, 6,761 PGx entities and 2,875 relationships between them were annotated).

3. The PROGENE Corpus

3.1. Data and Annotation Setup

The development of the PROGENE corpus started at the JULIE Lab Jena in 2008 (at that time, informally referred to as FSU-PRGE). There were two annotators, one of them an expert biologist, the other an NLP researcher with strong biomedical background. The main goals of the annotation project were

- to construct a consistent and (as far as possible) subdomain-independent and comprehensive protein-annotated corpus
- to differentiate between protein families and groups, protein complexes, protein molecules, protein variants (e.g. alleles) and elliptic enumerations of proteins—much needed distinctions for professional biologists.

To achieve a large coverage of biological subdomains, documents from multiple existing protein/gene corpora were reannotated. To increase coverage, new document sets were created. All documents are abstracts from PUBMED/MEDLINE. The final corpus consists of the union of all the documents in the different subcorpora. The annotation guidelines were primarily created by the expert biologist with support from the other annotator. The active learning-based Jena Annotation Environment (JANE) (Tomanek et al., 2007) was chosen to manage the annotation project. JANE leverages the MMAX2 tool (Müller and Strube, 2006) for the annotation process which is why the primary annotation format of the corpus is the MMAX2 format.

An overview of the subcorpora is given in Table 2. The subcorpus designations are of a technical nature which is the result of the original document selection process to reach a large domain coverage. We keep the names for reference to the original data.

Entity Type	# of Entity Annotations	Percentage
protein	43,070	0.72
protein family or group	12,304	0.21
protein complex	2,858	0.05
protein variant	665	0.01
protein enum	617	0.01
Total	59,514	1.00

Table 2: Number of occurrences of named entities within the PROGENE corpus

In the following, we provide an overview of the annotation levels in PROGENE. Despite the convention that level names carry a *protein* prefix, all levels also include annotations for gene and mRNA mentions. Thus, the annotations make no difference between proteins, genes, or mRNAs.

protein. Mentions of genes or proteins are regarded as textual mentions referring to an entity that can be found in a relevant database, most importantly UNIPROT.[8] This class of entities also includes mentions with promotor designations or organism indicators. Note that mentions that actually denote a group of exactly two elements also belong to this class, leaving the *protein enumeration* class for larger groups and families. Consider as an example: $[STAT5]_{protein}$, even though it is a group for *STAT5a* and *STAT5b*, but for only exactly those two.

protein_family_or_group. This class bundles families or groups of genes/proteins, e.g. $[transcription\ factors]_{protein\_family\_or\_group}$ or $[aquaporins]_{protein\_family\_or\_group}$. However, it does not incorporate very general terms (*lipoprotein*), locations (*mitochondrial genes*), functions (*RNA-binding proteins*)

[7] `https://biocreative.bioinformatics.udel.edu/`

[8] `https://www.uniprot.org/`

or similarity-descriptions (*Caspase-like proteins*). The identification of such groups and families is especially important for gene/protein grounding tasks which commonly assign database IDs to mentions of concrete genes/proteins but not for groups or families of them, frequently leading to false positives for taggers that do not differentiate between the two.

protein_complex. Complexes of at least two different proteins, e.g., *[IL-2 receptor]*$_{protein\_complex}$.

protein_variant. Annotations of allelic variants of a gene or protein isoforms, e.g., *[apoE2]*$_{protein\_variant}$.

protein_enum. Elliptic enumerations of two or more proteins, e.g., *[STAT11 and 12]*$_{protein\_enum}$. This annotation level is not used for enumerations of separate, yet complete protein/gene names, such as with *[STAT4]*$_{protein}$ *and [STAT11]*$_{protein}$.

The corpus does not contain nested annotations. For example, the elements of an enumeration are not annotated as proteins. In general, the annotation guidelines strive to avoid high complexity in the structure of the annotations in order to achieve a higher annotation consistency.

3.2. Corpus Characteristics

Putting all pieces together, the PROGENE corpus consists of 3,308 documents with 36,223 sentences and 960,757 tokens. For a more detailed description of the sub-corpora, cf. Table 3 which depicts the distribution of documents and sentences as well as a short description of the corresponding sub-corpora. Each sub-corpus is contained in a directory of its own in the download package so that either specific sub-corpora can directly be addressed for specific in-depth analysis or the entire corpus for more general purposes.

As can be seen from Table 2 and Figure 1 the number of entities is not equally distributed within the corpus. The label `protein` dominates the annotated entities with around 43k occurrences summing up to about 70% of all entities.

Figure 1 contains the overall count of entities as well as their number of extensions—the tokens other than the first—if an entity consists of more than one token. The

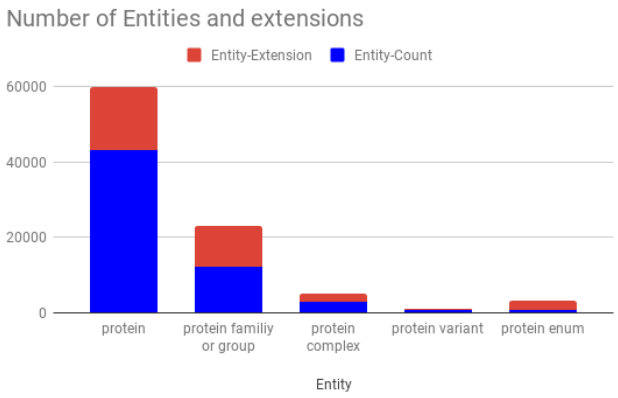

Figure 1: Number of entities (blue) and the number of their extensions or additional tokens (red)

`protein family or group` and `protein enum` are frequently comprised of multiple tokens which is especially true for the `protein family or group` class. Yet the annotated spans of `protein variant` and `protein enum` are typically two or four tokens long, respectively. On second sight, this is not surprising as the `protein family or group` describes gene/protein families which often have the token *factor* or *protein* attached as second token. Similarly, the `protein enum` class describes enumerations of more than one protein (see Section 3.1. for detailed descriptions).

The download release of the PROGENE corpus contains the annotations in *MMAX2* and *IOB* format.

4. Baseline Classifiers

When we want to evaluate different methods on a given dataset, a baseline is needed for better comparison. Here, we provide a realistic multi-class classification baseline for this corpus. We decided to use two state-of-the-art methods: FLAIR (Akbik et al., 2019) and BIOBERT (Lee et al., 2020). For statistical evaluation, we split the PROGENE corpus into a fixed set of 10 train-test partitions (with disjoint test sets) and performed a 10-fold cross-validation. The respective code is contained in the download release as well.

4.1. FLAIR

ELMO (Peters et al., 2018) marked a breakthrough in contextualized embedding techniques. In this approach, word embeddings are created which depend on their particular lexical surroundings in a text rather than representing each word with a single, static embedding vector. Akbik et al. (2018) extended the approach by introducing a purely character-based technique that does not use a fixed vocabulary of words any more. This method was implemented in FLAIR (Akbik et al., 2019), an NLP framework mainly for sequence tagging and text classification using PYTORCH (Paszke et al., 2017). As with ELMO, BILSTM-based language models are trained that, at test/prediction time, create vector representations for each (character) position in a given text. This is done in a forward and backward manner based on the head or the tail of the text, respectively, relative to the specific position in the text.

In our experiments, we used custom-trained FLAIR (forward and backward) and FASTTEXT embeddings. For the training of all three embedding models, we created a set of nearly 750M PUBMED abstracts with 1,226B words according to FASTTEXT without pre-processing for text normalization. The documents were selected by tagging a late 2018 snapshot of MEDLINE for genes with BANNER (Leaman and Gonzalez, 2008) that had been trained on the complete train and test set of the BIOCREATIVE II (Smith et al., 2008) Gene Mention data. Documents in which BANNER found at least one gene mention were added to the corpus dedicated for embedding learning.

For the **FASTTEXT embeddings**, we set the number of dimensions to 300. To use them in FLAIR, it was necessary to convert them into the GENSIM format (Řehůřek and Sojka, 2010). The text representation was used to create the final vector representation employed in our FLAIR tagger.

Subcorpus name	# Documents	# Sentences	Description
genes cytorec	253	2,433	MEDLINE abstracts focusing on cellular receptors
genes genetag 1	594	7,155	Part 1 of the reannotated GENETAG corpus (Tanabe et al., 2005), disjoint ffom genetag 2
genes genetag 2	541	6,317	Part 2 of the reannotated GENETAG corpus (Tanabe et al., 2005), disjoint from genetag 1
genes LLL / AiMed	296	2,824	all documents from which sentences were drawn for the LLL corpus (57 documents), 116 documents from the AIMED corpus plus 123 additional documents from MEDLINE
genes PIR	282	2,778	MEDLINE abstracts selected to cover proteins in the PIR database (`http://pir.georgetown.edu/`)
genes x45 shuffled	317	3,441	The 'shuffled' suffix refers to a random selection of documents from a larger base set for annotation. It is only kept for consistency reasons internal to the JULIE LAB
proteins 0	201	2,102	despite the naming similarity, these documents are disjoint from those in 'proteins 0 shuffled'
proteins 0 shuffled	236	2,829	despite the naming similarity, these documents are disjoint from those in 'proteins 0'
proteins 5	324	3,662	MEDLINE abstracts focusing on proteins
proteins ecoli	148	1,519	MEDLINE abstracts focusing on papers using *E. coli*
proteins KIR	114	1,146	MEDLINE abstract selected to cover proteins from the KIR database (`https://www.ebi.ac.uk/ipd/kir/`)
total	**3308**	**36,223**	

Table 3: Detailed information and important hints for using the PROGENE corpus and its various sub-corpora.

FLAIR embeddings are BILSTM language models. In order to train them, we split the embedding corpus into train, dev and test subsets. The dev and test sets were approximately 1% of the size of the train set or 150k lines each, where the lines were the title or the abstract text of a MEDLINE citation and for each citation its title and its text body was included as separate lines. The forward and backward models were trained with the same set of parameters (with the exception of the direction) and were learned using FLAIR in version 0.4.2. We employed the FLAIR-provided *chars* dictionary since we mainly dealt with English text. We used a single hidden layer for each model with a size of 2048, respectively. We set the sequence length to 250, the mini batch size to 100, the patience to 25, the anneal factor to 0.25 and the maximum of epochs to 10. We chose those parameters similar to the recommendations of the FLAIR team which they reported to work well for them. We trained both models for 40 days on a GeForce GTX 1080 graphics card. The forward model achieved its best and final performance on the dev set after 23 days. The dev loss had reached 0.71, the perplexity was 2.04. The test set for the final model showed a loss of 0.72 and a perplexity of 2.05. The backward model took 13 days to reach the minimal dev loss of 0.72, perplexity 2.06, the test performance was also a loss of 0.72 and a perplexity of 2.06.

The three embedding models were then used to build one single *stacked embedding*, i.e., the concatenation of the embedding vectors created for a specific word occurrence in the text (the FLAIR word representations encode the state of the language model after or before the word, for the forward and backward model, respectively). We created one FLAIR model for each of the 10 cross-validation splits where from each train partition 5% was used as the dev set by sampling each 20th sentence from the original train set.

FLAIR leverages a CRF-BILSTM architecture for sequence tagging. We set the hidden size of the BILSTM to 256, 1 hidden layer, used no dropout, set the learning rate to 0.12, the anneal factor to 0.5, the patience to 3, the mini batch size to 32 and the maximum number to 200. Those settings were chosen in accordance with the ones recommended by the FLAIR team, as well. All training runs terminated after 60 to 80 epochs due to no further improvements on the development set and the following vanishing of the learning rate by annealing. Each model training took 4 to 5h on an NVIDIA GeForce RTX 2080.

4.2. BioBERT

In contrast to FLAIR, bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019) constitute an unsupervised method to train language models on unlabeled texts. To fine-tune the model for a specific domain or to solve a specific task, only one additional output layer is needed and no further adjustments in the model itself are required.

Lee et al. (2020) recently proposed a variant of BERT, BIOBERT, especially adapted to the biomedical domain,

with both code and models accessible on GitHub.[9] As vanilla BIOBERT is only able to service binary classification tasks, we modified the code to enable *multi*-class classification.

We used the BIOBERT PUBMED v1.0 model trained on PUBMED abstracts and PUBMED CENTRAL full articles as the basis for fine-tuning on the training dataset using an NVIDIA GeForce RTX 2080 (8 GB) GPU with 100 epochs and a batch size of 16 for training. The other parameters were left as suggested by Lee et al. (2020). It is important to mention that BIOBERT uses its own tokenization when enabling the `do_predict` flag which splits all non-word characters. Thus, the number of generated tokens and the ones in the test set may differ, unless the tokenization is adapted, which leads the program to crash. Fine-tuning on one of the train-test splits usually took about 7 hours, whereas prediction consumed only about 5 minutes. The largest amount of time for prediction is needed to load the language model and start the process, not the prediction itself. The training time can be significantly reduced by using the default number of epochs (3) which, in turn, lowers the amount of training time to less than one hour per run.

4.3. Results

Although FLAIR works fine on the original tokenization of PROGENE, for better comparison this evaluation has been carried out on exactly the same tokenization as created for BERT. As can be seen in Table 4, FLAIR performed slightly better than BIOBERT.

Entity	Metric	FLAIR	BioBERT
proteins	Precision	86.42	82.93
	Recall	90.64	86.31
	F1	**0.885**	**0.846**
protein complex	Precision	78.87	71.4
	Recall	66.35	58.25
	F1	**0.719**	**0.638**
protein enum	Precision	68.91	55.08
	Recall	55.73	60.91
	F1	**0.610**	**0.572**
protein family or group	Precision	80.01	72.03
	Recall	75.57	72.68
	F1	**0.777**	**0.723**
protein variant	Precision	72.78	50.61
	Recall	36.48	28.12
	F1	**0.480**	**0.356**
All	Accuracy	97.59	97.04
	Precision	84.66	79.77
	Recall	85.39	81.21
	F1	**0.850**	**0.805**

Table 4: Classification results of FLAIR and BIOBERT on the 10-fold cross-validation.

The single classification results and the confusion matrix (see Figure 2) show that FLAIR consistently produces higher scores than BIOBERT except for the recall value on the `protein_enum` class. The most prominent mismatched categories are `protein` and `protein_variant`. Taking the definitions of both labels into consideration, this is not surprising. A `protein_variant` is defined as an allelic variant of a gene or protein isoform, thus its name is in many cases only slightly different from the corresponding proteins name and thus hard to distinguish. In a similar way, it appears that `protein_complex` and `protein_family_or_group` are hard classes to recognize. Entities of those categories often do not give away their nature just by their name or context since they are often used interchangeably with individual proteins in the literature and thus require resources like family lists.

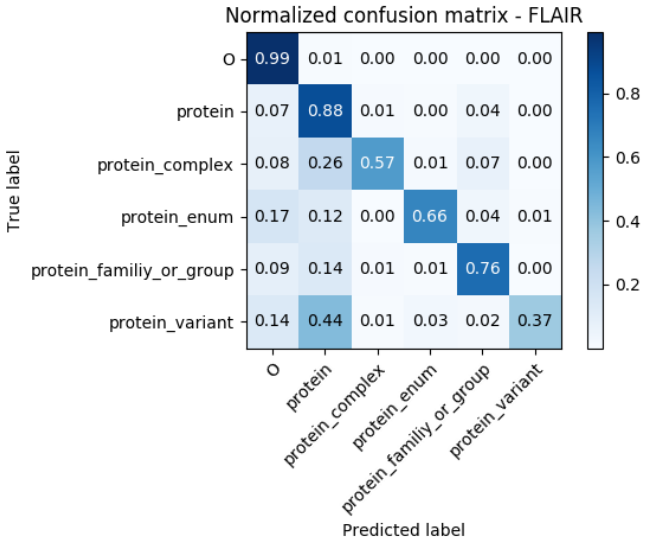

Figure 2: Confusion matrix of the FLAIR prediction results over all 10 test-splits. The rows show the true label. The columns depict how often entities of the true label were predicted as the label represented by the column. The matrix for BIOBERT has a similar appearance.

5. Conclusion

In this paper, we described the PROGENE corpus with annotations for five named entity types for genes/proteins, namely `protein`, `protein family or group`, `protein complex`, `protein variant` and `protein enum`. PROGENE consists of 3,308 documents with about 36k sentences or more than 960k tokens.

We built a consistent and as far as possible subdomain-independent and comprehensive protein-annotated corpus with a metadata set of nearly 60k manually added, fine-grained entity annotations from different protein types. Although the creation of the initial corpus started 10 years ago, it is still the largest publicly available abstract-based protein/gene corpus world-wide, resulting from a single-site annotation campaign.

As a baseline for comparison, we tested two state-of-the-art classifiers: FLAIR and BIOBERT. The results show higher scores on our specific cross-validation splits for FLAIR in terms of accuracy, precision, recall and f1-score.

[9] `https://github.com/dmis-lab/biobert`

Acknowledgements This work was supported by BMBF within the SMITH project under grant 01ZZ1803G and DFG under grant HA 2079/8-1 within the STAKI2B2 project. We thank all annotators, as well as André Scherag, Danny Ammon, and all members of the Data Integration Center of the Jena University Hospital.

6. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In Pierre Isabelle, et al., editors, *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics: Main Conference. Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1638–1649.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR : an easy-to-use framework for state-of-the-art NLP. In Waleed Ammar, et al., editors, *NAACL-HLT 2019 — Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Demonstrations Session. Minneapolis, Minnesota, USA, June 3-4, 2019*, pages 54–59. Association for Computational Linguistics (ACL).

Alex, B., Grover, C., Haddow, B., Kabadjov, M. A., Klein, E., Matthews, M., Roebuck, S., Tobin, R., and Wang, X. (2008). The ITI TXM corpora: tissue expressions and protein-protein interactions. In Sophia Ananiadou, et al., editors, *BioTxtM 2008 — Proceedings of the 1st Workshop on Building and Evaluating Resources for Biomedical Text Mining @ LREC 2008. Marrakech, Morocco, May 26, 2008*, pages 11–18, Paris. European Language Resources Association (ELRA).

Arighi, C. N., Lu, Z., Krallinger, M., Cohen, K. B., Wilbur, W. J., Valencia, A., Hirschman, L., and Wu, C. H. (2011). Overview of the BIOCREATIVE III workshop. *BMC Bioinformatics*, 12(Suppl 8):S1.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K. M., Blake, J. A., and Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13:#161.

Baumgartner, W. A., Bada, M., Pyysalo, S., Ciosici, M. R., Hailu, N. D., Pielke-Lombardo, H., Regan, M., and Hunter, L. E. (2019). CRAFT Shared Tasks 2019 overview: integrated structure, semantics, and coreference. In Jin-Dong Kim, et al., editors, *BioNLP-OST 2019 — Proceedings of the 5th Workshop on BioNLP Open Shared Tasks @ EMNLP-IJCNLP 2019. Hong Kong, China, November 4, 2019*, pages 174–184. Association for Computational Linguistics (ACL).

Beisswanger, E., Lee, V., Kim, J.-j., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., and Hahn, U. (2008). Gene Regulation Ontology (GRO): design principles and use cases. In Stig Kjær Andersen, et al., editors, *MIE 2008 — Proceedings of the 21st International Congress of the European Federation for Medical Informatics. eHealth Beyond the Horizon: Get IT There. Gothenburg, Sweden, 25-28 May 2008*, volume 136 of *Studies in Health Technology and Informatics*, pages 9–14. IOS Press.

Bossy, R., Golik, W.-t., Ratkovic, Z., Valsamou, D., Bessières, P., and Nédellec, C. (2015). Overview of the Gene Regulation Network and the Bacteria Biotope Tasks in BIONLP '13 Shared Task. *BMC Bioinformatics*, 16(Suppl 10):S1.

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K. D., Maglott, D. R., and Murphy, T. D. (2015). GENE : a gene-centered information resource at NCBI. *Nucleic Acids Research*, 43(D1 (Database issue)):D36–D42.

Buyko, E., Beisswanger, E., and Hahn, U. (2010). The GENEREG corpus for gene expression regulation events. An overview of the corpus and its in-domain and out-of-domain interoperability. In Nicoletta Calzolari, et al., editors, *LREC 2010 — Proceedings of the 7th International Conference on Language Resources and Evaluation. La Valletta, Malta, May 17-23, 2010*, pages 2662–2666. European Language Resources Association (ELRA).

Dai, H.-J., Wu, J. C.-Y., and Tsai, R. T.-H. (2013). Collective instance-level gene normalization on the IGN corpus. *PLoS One*, 8(11):e79517.

Dang, T. H., Le, H.-Q., Nguyen, T. M., and Vu, S. T. (2018). D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. *Bioinformatics*, 34(20):3539–3546.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. N. (2019). BERT : pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, *NAACL-HLT 2019 — Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, USA, June 2-7, 2019*, volume 1: Long and Short Papers, pages 4171–4186. Association for Computational Linguistics (ACL).

Ding, J., Berleant, D., Nettleton, D., and Wurtele, E. S. (2002). Mining MEDLINE: abstracts, sentences, or phrases? In Russ B. Altman, et al., editors, *PSB 2002 — Proceedings of the Pacific Symposium on Biocomputing. Kauai, Hawaii, USA, January 3-7, 2002*, pages 326–337, Singapore. World Scientific Publishing.

Doğan, R. I., Chatr-aryamontri, A., Kim, S., Wei, C.-H., Peng, Y., Comeau, D. C., and Lu, Z. (2017). BIOCREATIVE VI Precision Medicine Track: creating a training corpus forming protein-protein interactions affected by mutations. In Kevin Bretonnel Cohen, et al., editors, *BioNLP 2017 — Proceedings of the 16th SIGBioMed Workshop on Biomedical Natural Language Processing @ ACL 2017. Vancouver, British Columbia, Canada, August 4, 2017*, pages 171–175. Association for Computational Linguistics (ACL).

Furlong, L. I., Dach, H., Hofmann-Apitius, M., and Sanz, F. (2008). OSIRIS v1.2: a named entity recognition system for sequence variants of genes in biomedical literature. *BMC Bioinformatics*, 9(1):#84.

Galea, D., Laponogov, I., and Veselkov, K. (2018). Exploiting and assessing multi-source data for supervised biomedical named entity recognition. *Bioinformatics*, 34(14):2474–2482, July.

Gene Ontology Consortium. (2001). Creating the GENE ONTOLOGY resource: design and implementation. *Genome Research*, 11(8):1425–1433.

Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., and Krallinger, M. (2019). PHARMACONER: Pharmacological Substances, Compounds and Proteins Named Entity Recognition Track. In Jin-Dong Kim, et al., editors, *BioNLP–OST 2019 — Proceedings of the 5th Workshop on BioNLP Open Shared Tasks @ EMNLP-IJCNLP 2019. Hong Kong, China, November 4, 2019*, pages 1–10, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Habibi, M., Weber, L., Neves, M. L., Wiegandt, D. L., and Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Hahn, U., Tomanek, K., Beisswanger, E., and Faessler, E. (2010). A proposal for a configurable silver standard. In Nianwen Xue, et al., editors, *LAW IV — Proceedings of the 4th Linguistic Annotation Workshop @ ACL 2010. Uppsala, Sweden, 15-16 July 2010*, pages 235–242, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Hatzivassiloglou, V., Duboué, P. A., and Rzhetsky, A. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*, 17(Supplement 1):S97–S106.

Hirschman, L., Yeh, A. S., Blaschke, C., and Valencia, A. (2005). Overview of BIOCREATIVE : critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Supplement 1):S1.

Huang, M.-S., Lai, P.-T., Tsai, R. T.-H., and Hsu, W.-L. (2019). Revised JNLPBA corpus: a revised version of biomedical NER corpus for relation extraction task. *arXiv.org: arXiv:1901.10219*.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1 (Database issue)):D109–D114.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus: a semantically annotated corpus for biotextmining. *Bioinformatics*, 19(Suppl 1):i180–i182.

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. H. (2004). Introduction to the Bio-Entity Recognition Task at JNLPBA. In Nigel Henry Collier, et al., editors, *JNLPBA 2004 — Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications @ COLING 2004. Geneva, Switzerland, August 28-29, 2004*, pages 70–75.

Kim, J.-D., Ohta, T., and Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:#10.

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J. (2009). Overview of BIONLP '09 Shared Task on Event Extraction. In Jun'ichi Tsujii, editor, *BioNLP 2009 — Proceedings of the BioNLP 2009 Shared Task on Event Extraction @ NAACL-HLT 2009. Boulder, Colorado, USA, June 5, 2009*, volume Companion Volume, pages 1–9. Association for Computational Linguistics (ACL).

Kim, J.-D., Wang, Y., Takagi, T., and Yonezawa, A. (2011). Overview of Genia Event Task in BIONLP Shared Task 2011. In Jun'ichi Tsujii, et al., editors, *BioNLP 2011 — Proceedings of the BioNLP Shared Task 2011 Workshop on Biomedical Natural Language Processing @ ACL-HLT 2011. Portland, Oregon, USA, 24 June 2011*, pages 7–15. Association for Computational Linguistics (ACL).

Kim, J.-D., Wang, Y., and Yasunori, Y. (2013). The GENIA Event Extraction Shared Task, 2013 edition: overview. In Claire Nédellec, et al., editors, *BioNLP 2013 — Proceedings of the BioNLP Shared Task 2013 Workshop @ ACL 2013. Sofia, Bulgaria, August 9, 2013*, pages 8–15. Association for Computational Linguistics (ACL).

Kim, J.-D., Wang, Y., Colic, N., Baek, S. H., Kim, Y. H., and Song, M. (2016). Refactoring the GENIA event extraction shared task toward a general framework for ie-driven kb development. In Claire Nédellec, et al., editors, *BioNLP 2016 — Proceedings of the 4th BioNLP Shared Task Workshop @ ACL 2016. Berlin, Germany, 13 August 2016*, pages 23–31, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Krallinger, M., Malik, R., and Valencia, A. (2006). Text mining and protein annotations: the construction and use of protein description sentences. *Genome Informatics*, 17(2):121–130.

Krallinger, M., Morgan, A. A., Smith, L. H., Leitner, F., Tanabe, L. K., Wilbur, W. J., Hirschman, L., and Valencia, A. (2008). Evaluation of text mining systems for biology: overview of the second BIOCREATIVE community challenge. *Genome Biology*, 9(Suppl 2):S1.

Krallinger, M., Rabal, O., Akhondi, S. A., Pérez, M. P., Santamaría, J., Rodríguez, G. P., Tsatsaronis, G., Intxaurrondo, A., López, J. A., Nandal, U., Van Buel, E., Chandrasekhar, A., Rodenburg, M., Laegreid, A., Doornenbal, M. A., Oyarzabal, J., Lourenço, A., and Valencia, A. (2017). Overview of the BIOCREATIVE VI Chemical-Protein Interaction Track. In *Proceedings of the 6th* BIOCREATIVE VI *Challenge Evaluation Workshop*, volume 1, pages 141–146.

Leaman, R. and Gonzalez, G. H. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In Russ B. Altman, et al., editors, *PSB 2008 — Proceedings of the 13th Pacific Symposium on Biocomputing 2008. Kona, Hawaii, USA, January 2008*, pages 652–663.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BIOBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February.

Legrand, J., Gogdemir, R., Bousquet, C., Dalleau, K., Devignes, M.-D., Digan, W., Lee, C.-J., Ndiaye, N.-C., Petitpain, N., Ringot, P., Smaïl-Tabbone, M., Toussaint, Y., and Coulet, A. (2020). PGXCORPUS, a manually annotated corpus for pharmacogenomics. *Scientific Data*, 7:#3.

Leitner, F., Mardis, S. A., Krallinger, M., Cesareni, G., Hirschman, L., and Valencia, A. (2010). An overview of BIOCREATIVE II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):385–399.

Mao, Y., Van Auken, K., Li, D., Arighi, C. N., McQuilton, P., Hayman, G. T., Tweedie, S., Schaeffer, M. L., Laulederkind, S. J. F., Wang, S.-J., Gobeill, J., Ruch, P., Luu, A. T., Kim, J.-j., Chiang, J.-H., Chen, Y.-D., Yang, C.-J., Liu, H., Zhu, D., Li, Y., Yu, H., Emadzadeh, E., Gonzalez, G. H., Chen, J.-M., Dai, H.-J. J., and Lu, Z. (2014). Overview of the Gene Ontology Task at BIOCREATIVE IV. *Database: The Journal of Biological Databases and Curation*, 2014:#bau086.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, et al., editors, *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, pages 197–214. Peter Lang.

Nédellec, C., Bossy, R., Kim, J.-D., Kim, J.-j., Ohta, T., Pyysalo, S., and Zweigenbaum, P. (2013). Overview of BIONLP Shared Task 2013. In *BioNLP 2013 — Proceedings of the BioNLP Shared Task 2013 Workshop @ ACL 2013. Sofia, Bulgaria, August 9, 2013*, pages 1–7. Association for Computational Linguistics (ACL).

Nédellec, C. (2005). Learning Language in Logic – Genic Interaction Extraction Challenge. In J. Cussens et al., editors, *LLL '05 — Proceedings of the 4th Learning Language in Logic Workshop @ ICML 2005. Bonn, Germany, 7 August 2005*, pages 31–37.

Neves, M. L., Damaschun, A., Kurtz, A., and Leser, U. (2012). Annotating and evaluating text for stem cell research. In Sophia Ananiadou, et al., editors, *BioTxtM 2012 — Proceedings of the 3rd Workshop on Building and Evaluating Resources for Biomedical Text Mining @ LREC 2012. Istanbul, Turkey, May 26, 2012*, pages 16–23. European Language Resources Association (ELRA).

Ohta, T., Tateisi, Y., and Kim, J.-D. (2002). The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In Mitchell P. Marcus, editor, *HLT 2002 — Human Language Technology Conference. Proceedings of the 2nd International Conference on Human Language Technology Research. San Diego, California, USA, March 24-27, 2002*, pages 82–86, San Francisco/CA. Morgan Kaufmann.

Ohta, T., Kim, J.-D., Pyysalo, S., Wang, Y., and Tsujii, J. (2009). Incorporating GENETAG-style annotation to GENIA corpus. In Kevin Bretonnel Cohen, et al., editors, *BioNLP 2009 — Proceedings of the Workshop on Biomedical Natural Language Processing @ NAACL-HLT 2009. Boulder, Colorado, USA, June 4-5, 2009*, pages 106–107, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PYTORCH. In *Proceedings of the Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques @ NIPS 2017. Long Beach, California, USA, December 9, 2017*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C. T., Lee, K., and Zettlemoyer, L. S. (2018). Deep contextualized word representations. In Marilyn A. Walker, et al., editors, *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA, June 1-6, 2018*, volume 1: Long Papers, pages 2227–2237. Association for Computational Linguistics (ACL).

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., and Salakoski, T. (2007). BIOINFER : a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:#50.

Pyysalo, S., Airola, A., Heimonen, J., Björne, J., Ginter, F., and Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl 3):S6.

Pyysalo, S., Ohta, T., and Tsujii, J. (2011). Overview of the Entity Relations (REL) Supporting Task of BIONLP Shared Task 2011. In Jun'ichi Tsujii, et al., editors, *BioNLP 2011 — Proceedings of the BioNLP Shared Task 2011 Workshop on Biomedical Natural Language Processing @ ACL-HLT 2011. Portland, Oregon, USA, 24 June 2011*, pages 83–88, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Pyysalo, S., Ohta, T., Rak, R., Sullivan, D., Mao, C., Wang, C., Sobral, B., Tsujii, J., and Ananiadou, S. (2012). Overview of the ID, EPI and REL tasks of BIONLP Shared Task 2011. *BMC Bioinformatics*, 13(Suppl 11):S2.

Pyysalo, S., Ohta, T., Rak, R., Rowley, A. D., Chun, H.-W., Jung, S.-J., Choi, S.-P., Tsujii, J., and Ananiadou, S. (2015). Overview of the Cancer Genetics and Pathway Curation Tasks of BIONLP Shared Task 2013. *BMC Bioinformatics*, 16(Suppl 10):S2.

Rebholz-Schuhmann, D., Jimeno-Yepes, A., van Mulligen, E. M., Kang, N., Kors, J. A., Milward, D., Corbett, P., Buyko, E., Tomanek, K., Beisswanger, E., and Hahn, U. (2010). The CALBC Silver Standard Corpus for biomedical named entities: a study in harmonizing the contributions from four independent named entity taggers. In *LREC 2010 — Proceedings of the 7th International Conference on Language Resources and Evaluation. La Valletta, Malta, May 17-23, 2010*, pages 568–573, Paris. European Language Resources Association (ELRA).

Rebholz-Schuhmann, D., Kafkas, Ş., Kim, J.-H., Li, C., Jimeno-Yepes, A., Hoehndorf, R., Backofen, R., and Lewin, I. (2013). Evaluating gold standard corpora against gene/protein tagging solutions and lexical resources. *Journal of Biomedical Semantics*, 4:#28.

Smith, L. H., Tanabe, L. K., Johnson [nee Ando], R., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., Ling, Y.-S., Klinger, R., Friedrich, C. M., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C. A., Povinelli, R. J., Vlachos, A., Baumgartner, W. A., Hunter, L. E., Carpenter, B., Tsai, R. T.-H., Dai, H.-J. J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P. W., Blaschke, C., Torres, R., Neves, M. L., Nakov, P. I., Divoli, A., Maña López, M. J., Mata, J., and Wilbur, W. J. (2008). Overview of BIOCREATIVE II Gene Mention Recognition. *Genome Biology*, 9(Suppl 2):S2.

Sousa, D., Lamurias, A., and Couto, F. M. (2019). A silver standard corpus of human phenotype-gene relations. In Jill Burstein, et al., editors, *NAACL-HLT 2019 — Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, USA, June 2-7, 2019, pages 1487–1492. Association for Computational Linguistics (ACL).

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

Tomanek, K., Wermter, J., and Hahn, U. (2007). Efficient annotation with the Jena ANnotation Environment (JANE). In Branimir K. Boguraev, et al., editors, *LAW [I] — Proceedings of the [1st] Linguistic Annotation Workshop @ ACL 2007. Prague, Czech Republic, June 28-29, 2007*, pages 9–16, Stroudsburg/PA. Association for Computational Linguistics (ACL).

UniProt Consortium. (2017). UNIPROT: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1 (Database issue)):D158–D169.

Verspoor, K. M., Jimeno-Yepes, A., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., and Plazzer, J.-P. (2013). Annotating the biomedical literature for the human variome. *Database: The Journal of Biological Databases and Curation*, 2013:#bat019, April.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In René Witte, et al., editors, *Proceedings of the Workshop on New Challenges for NLP Frameworks @ LREC 2010. La Valletta, Malta, May 22, 2010*, pages 45–50.

Wang, Y., Kim, J.-D., Sætre, R., Pyysalo, S., and Tsujii, J. (2009). Investigating heterogeneous protein annotations toward cross-corpora utilization. *BMC Bioinformatics*, 10:#403.

Wang, Y., Kim, J.-D., Sætre, R., Pyysalo, S., Ohta, T., and Tsujii, J. (2010). Improving the inter-corpora compatibility for protein annotations. *Journal of Bioinformatics and Computational Biology*, 8(5):901–916.

Wang, Y., Zhou, K., Gachloo, M., and Xia, J. (2019). An overview of the Active Gene Annotation Corpus and the BIONLP OST 2019 AGAC Track Tasks. In Jin-Dong Kim, et al., editors, *BioNLP-OST 2019 — Proceedings of the 5th Workshop on BioNLP Open Shared Tasks @ EMNLP-IJCNLP 2019. Hong Kong, China, Nov. 4, 2019*, pages 62–71. Association for Computational Linguistics (ACL).

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2015). GNORMPLUS : an integrative approach for tagging genes, gene families, and protein domains. *BioMed Research International*, 2015:#918710.

Yeh, A. S., Morgan, A. A., Colosimo, M., and Hirschman, L. (2005). BIOCREATIVE Task 1A : Gene Mention finding evaluation. *BMC Bioinformatics*, 6(Suppl 1):S2.