

Text Mining and Protein Annotations: the Construction and Use of Protein Description Sentences

Martin Krallinger¹ Rainer Malik²
mkrallinger@cnio.es rainer@cs.uu.nl

Alfonso Valencia¹
valencia@cnio.es

¹ Dep. Struct. Comp. Biology Spanish National Cancer Centre (CNIO), Melchor Fernández Almagro, 3, E-28029 Madrid, Spain

² Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany

Abstract

Existing biological knowledge stored as structured database records has been extracted manually by database curators analyzing the scientific literature. Most of this information was derived from sentences which describe biologically relevant aspects of genes and gene products. We introduce the Protein description sentence (Prodisen) corpus, a useful resource for the automatic identification and construction of text-based protein and gene description records using information extraction and text classification techniques. Basic guidelines and criteria relevant for the construction of a text corpus of functional descriptions of genes and proteins are proposed. The steps used for the corpus construction and its features are presented. Moreover, some of the potential applications of the Prodisen corpus for biomedical text mining purposes are explored and the obtained results are presented.

Keywords: text mining, information extraction, biological annotations, gene description, corpus construction

1 Introduction

The rapid growth of the biomedical literature, as deposited in databases such as PubMed, together with the demands posed by the biology community for efficient access to gene descriptions increased the interest in using information extraction (IE) and text mining strategies to identify relevant functional descriptions from scientific literature.

Attempts to use computational tools to extract a variety of functional information from the literature, ranging from protein interactions [1], gene product annotations [10, 16] to gene-disease association [13] have been made in the past.

For the development and evaluation of such systems, high quality training and test data collections are crucial. Despite the number of described applications, only few attempts to construct such data sets have been made so far. In most cases, biomedical text mining systems were evaluated based on small data collections. These collections often lack sufficient descriptions on both data selection criteria as well as quality of the data in terms of inter-annotator agreement [8]. Although developing suitable biomedical corpora is a time consuming and labor-intensive task which requires the involvement of domain experts, some high quality biomedical corpora exist, such as the GENIA corpus [9]. Moreover Cohen *et al.* [5] explored basic aspects which influence the usage of existing biomedical corpora.

An aspect often neglected when using existing biomedical corpora is that they are often based on previous information retrieval (IR) steps using very specific query terms. In the case of the GENIA

corpus, a previous selection step for PubMed articles with the MeSH terms *human*, *blood cell* and *transcription factor* was done. Previous filtering steps using very specific query terms or selecting only certain journals may result in limiting the type of functional descriptions contained in the derived corpora.

Other available data collections were specifically tailored for very narrow text mining tasks such as the identification of gene mentions in text [6, 19] and are therefore not suitable for the discovery of functional descriptions where proteins are referred to using for instance anaphoric expressions.

The first BioCreAtIvE challenge [2] resulted in a useful text passage corpus for the extraction of human protein annotations based on controlled vocabulary terms (Gene Ontology concepts). Nevertheless, it does not consider other relevant descriptions such as protein interactions or gene-disease associations. Other resources used by text mining applications [14] were derived from data stored in existing biological databases: Mainly the Gene Ontology Annotation (GOA) database [3] and the GeneRif database [12]. Both of these collections were not constructed for use by biomedical text mining applications. Their applicability for information extraction tools show significant limitations. GOA only provides references to PubMed entries and not individual sentences or passages supporting the actual annotations. In GeneRif, many entries do not correspond to PubMed sentences or were modified by the user which entered the record. These entries often lack the experimental context which support the gene descriptions.

To complement existing resources and to help in the efforts to bridge the gap between IR approaches and Text Mining, we developed the Protein descriptions in sentences (Prodisen) corpus. It contains a set of positive sentences corresponding to protein (and gene) descriptions and a contrast set of negative sentences with no functional information associated. Prodisen explores the whole PubMed collection without previous document restriction using specific query terms. As most of the existing NLP tools take sentences as the basic processing unit the Prodisen corpus proves to be especially suitable for discovering sentence derived features. The Prodisen sentences derived from a large corpus of PubMed abstracts have been manually categorized by domain experts.

The most significant gene description types considered include aspects related to their interactions with other biomolecules, their molecular function, cellular location or associations to disease conditions.

The procedure for the construction of Prodisen is designed to be reproducible and easy to extend by distributing the corpus together with the used corpus construction script. The Prodisen gene description sentences include anaphoric cases and are not limited to genes or gene products and also include general protein family names. To assess the difficulty of identifying gene descriptions from PubMed the inter-observer agreement was measured. Also basic characteristics such as the relative sentence position of gene descriptions within abstracts were measured.

2 Corpus Construction

For the construction of a biomedical corpus the first step involves the selection of journal articles or abstracts. The most important repository for life sciences literature citations, with over 15 million entries is the PubMed database, a service provided by the National Library of Medicine (NLM) [18].

Only a fraction of the PubMed entries are related to molecular biology, containing also a considerable amount of publications related to other domains such as clinical sciences, nursing, psychology, chemistry or engineering. Moreover, many of the abstracts in the Molecular Biology domain do not necessarily contain information on gene or gene product function.

For a system which tries to extract protein (and gene) descriptions from PubMed it is therefore crucial to identify not only those publications which are related to biological function, but also the actual sentences describing relevant aspects of the association between gene products and functions. One strategy could thus be to identify first proteins mentioned in text, and then select those sentences which correspond to descriptions of their functions. This approach has two main limitations; the first

is related to the difficulty in detecting protein/gene mentions in text due to the ambiguity of their actual names, in addition to the use of indirect references these names. To illustrate this problem, consider the following example:

Anaphora example [PMID:11827972]:

Sentence 1: "*CasL/HEF1 belongs to the p130(Cas) family.*"

Sentence 2: "*It is tyrosine-phosphorylated following beta(1) integrin and/or T cell receptor stimulation and is thus considered to be important for immunological reactions.*"

In the this example (anaphora), the protein, is mentioned in sentence 1. Although sentence two is obviously describing relevant aspects of this protein it is not mentioning the protein name explicitly, but referring to it using an anaphoric expression (e.g. through the pronoun *it*).

To take into account gene descriptions which do not explicitly state the characterized genes, in case of the Prodisen construction, domain experts were presented sequentially with the sentences together with the corresponding abstract. Then they had to classify each sentence into one of three basic categories: (Y) the sentence is useful as gene description, independent of whether the gene name is mentioned or a referential expression is used (based on the abstract context); (N) the sentence does not correspond to a gene description or (D) uncertain ambiguous cases, where the expert was not sure. A sample cases of sentences belonging to the negative data set (not suitable to derive protein, gene product or protein family) descriptions is:

"*N 15337019 10 Group B-1 (54 patients) underwent lateral prophylactic sternal reinforcement before placement of peristernal wires.*"

The uncertain class (D) included sentences where it was not clear from the context and annotator background knowledge if it was referring to a gene description. This class also included cases of wrong sentence splitting or uninformative short sentence fragments.

2.1 Gene Description Classes

The sentences containing protein description information were additionally classified in those 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. These cases include descriptions where the protein names appear as well as cases where referring expressions were used or it could be inferred that the sentence contains a relevant protein description (based on the context). The classes considered include:

1. Descriptions related to molecular functions. Example: 10896954 1 "*We recently characterized a novel protein, GIT1, that interacts with G protein-coupled receptor kinases and possesses ADP-ribosylation factor (ARF) GTPase-activating protein activity.*"
2. Descriptions related to biological processes. Example: 10648622 1 "*We present here the characterization of SPB1, an essential yeast gene that is required for ribosome synthesis.*"
3. 3- Descriptions which refer to descriptions of cellular location. Example: 10777497 4 "*Localization and biochemical fractionation studies show that Psr1p is associated with the plasma membrane via a short amino-terminal sequence also present in Psr2p.*"
4. Descriptions related to associations to diseases, symptoms or treatments. Example: 10551832 3 "*The dysfunction of ALDP is responsible for X-linked adrenoleukodystrophy (X-ALD), a neurodegenerative disorder in which saturated very long-chain fatty acids accumulate because of their impaired peroxisomal beta-oxidation.*"

Table 1: Prodisen corpus in numbers. R: Prodisen random corpus, E: Prodisen enriched corpus, Sent.: total number of sentences, Abs.: total number of abstracts, Words: total number of words, Y: total number of sentences useful to derive gene descriptions, N: total number of sentences which are not useful as gene descriptions, D: total number of sentences which correspond to uncertain cases.

| Corpus | Sent. | Abs. | Words | Y | N | D |
|--------|--------|-------|---------|-------|--------|-----|
| R | 10,039 | 1,234 | 224,890 | 1,704 | 7,899 | 436 |
| E | 11,125 | 1,232 | 244,549 | 7,693 | 2,949 | 483 |
| All | 21,164 | 2,466 | 469,439 | 9,397 | 10,848 | 919 |

5. Descriptions referring to interactions, e.g. protein interactions and dimerization or protein-compound interactions. Example: 10648622 6 “*Coimmunoprecipitation experiments show that Spb1p is associated in vivo with the nucleolar proteins Nop1p and Nop5/58p.*”
6. Information related to the gene expression (e.g. in which tissues a given gene is expressed) 11294880 6 “*NCKX3 transcripts were most abundant in brain, with highest levels found in selected thalamic nuclei, in hippocampal CA1 neurons, and in layer IV of the cerebral cortex.*”
7. Descriptions related to homology information. Example: 11274158 2 “*In this study, we identified a human homolog of RVS161, termed BIN3 (bridging integrator-3), and a Schizosaccharomyces pombe homolog of RVS161, termed hob3+ (homolog of Bin3).*”
8. Descriptions of sequence and structural features (including mutations, protein family, isoforms, post-translational modifications, SNP, chromosome mapping). Example 1: 10570961 3 “*The previously characterized hsp40 gene contains only three exons and two introns.*”
9. Other useful gene descriptions such as information related to phenotypes, experimental usage (markers) and enzyme kinetics.

All these basic types of gene descriptions refer to different relevant aspects that characterize genes, proteins and protein families and represent the diversity of annotation information stored in different biological annotation databases. In practice, single gene description sentences are often not limited to one single description aspect, but provide several characterization types.

2.2 Prodisen Random Corpus

Text mining applications which claim to extract functional descriptions from PubMed have to consider that only a fraction of abstracts contained in PubMed are related to Molecular Biology and of those only some sentences are related to gene descriptions. Thus, to address the detection of gene description sentences from the whole PubMed collection and to estimate the total amount of functional description sentences in PubMed, we constructed a set of randomly selected abstracts.

We have classified each of the sentences from those abstracts into one of the three previously described categories. As shown in Table 1, the Prodisen random corpus contains over 10,000 sentences, where around 15 percent have been classified as gene description sentences. Taking into account that there are over 7 million PubMed entries which contain abstracts in English, and that the average length of these abstracts is around 9 sentences, one would expect that there are around 9.5 million sentences in PubMed which contain gene description related information.

2.3 Prodisen Enriched Corpus

Often existing training collections suffer from class-imbalance problems, which require special treatments when used as input data for text classifiers. It is thus useful to have a data collection consisting

Table 2: PubMed article usage overlap between different biological databases. For each database the number of (non-redundant) PubMed articles used was extracted and the overlap between different databases was calculated.

| DB | GOA | GeneRif | UniProt | OMIM | PDB | IntAct |
|---------|--------|---------|---------|--------|--------|--------|
| GOA | 29,248 | 3,972 | 15,409 | 9,465 | 135 | 764 |
| GeneRif | 3,972 | 84,380 | 4,890 | 6,637 | 620 | 283 |
| UniProt | 15,409 | 4,890 | 112,476 | 19,859 | 5,061 | 764 |
| OMIM | 9,465 | 6,637 | 19,859 | 88,766 | 296 | 193 |
| PDB | 135 | 620 | 5,061 | 296 | 11,790 | 35 |
| IntAct | 764 | 283 | 764 | 193 | 35 | 1184 |

in a balanced set of positive and negative training instances. As the proportion of positive cases, i.e. sentences corresponding to gene descriptions, is relatively small in the random Prodisen corpus, we propose a strategy to construct an enriched set of abstracts in terms of gene descriptions. This strategy is based on PubMed article citation overlap between different biological annotation databases, each of them with a different focus regarding the gene description type. For instance, OMIM, as the primary data source for genetic disorders, was chosen because it includes much information about diseases, symptoms or treatments (Description Type 4), whereas PDB was chosen for the information about structural features (Description Type 8). GOA includes information related to cellular localization and gene expression (Description Types 3 and 6). GeneRif and UniProt are very good sources for all types of gene annotation.

To have a collection of abstracts that covers all the previously defined relevant gene description types we extracted first for each database the number of (non-redundant) PubMed articles used for their annotations. Then we identified the articles (with PubMed abstracts) which were used as citations by several different databases. Table 2 illustrates the binary overlap between the citations extracted for each of the biological databases. The core set of the enriched Prodisen corpus consisted in the articles cited by the following biological databases: UniProt, OMIM, GeneRif and GOA. Additional articles cited in GOA and PDB, Pfam or IntAct were included. The resulting Prodisen enriched corpus contained over 11 thousand sentences, with over 7 thousand gene description sentences (see Table 1).

2.4 Prodisen Inter-Annotator Agreement

To estimate the difficulty of identifying gene description relevant sentences, the measurement of annotator agreement is useful. Reader disagreement is often based on the certain degree of subjective interpretation characteristic of natural language. Measurements of annotator agreement is useful to assess the reliability and consistency of classification techniques and compare performance of human and computer-based classification strategies. The importance of measuring inter-annotator agreement is now clear after the experience of the community-wide evaluation efforts, especially the BioCre-AtIvE, see [4]. Reader disagreement is often based on a certain degree of subjective interpretation, as a general characteristic of natural language is that often more communicated than is said.

The difficulty of interpretation is particularly obvious in domain specific literature, such as the molecular biology literature. Authors of biomedical articles continuously make presuppositions on the background knowledge of their targeted scientific community. Also the context of a given sentence is crucial in order to make inference of its underlying semantics.

Measuring the observer agreement of categorical data has been also addressed in other domains, such as clinical medicine, i.e. reproducibility of disease classifications [11]. We developed two independent basic classifications following the guidelines for what should be considered as a valid gene description. The agreement scores for both Prodisen sets regarding the three classification types are

Table 3: Agreement indices of the polytomous ratings of the two Prodisen data sets. For a more detailed description on observer agreement calculation used the agreement indices refer to [11].

| Agreement Index | Description | Random | Enriched |
|-----------------|-------------|--------|----------|
| P(o) | Overall | 0.856 | 0.705 |
| P(pos) | Positive | 0.656 | 0.799 |
| P(neg) | Negative | 0.928 | 0.306 |
| P(e) | Chance | 0.662 | 0.490 |
| Kappa | Corrected | 0.574 | 0.421 |

provided in Table 3 and is in line with other experiences obtained for similar scenarios.

3 The Disease Prodisen Collection

There is an increasing interest in using text mining to detect functional characterizations and interactions of disease-associated proteins from the literature. The Prodisen strategy was thus applied to construct a sentence corpus which contains relevant descriptions of diseases-associated genes and proteins. This was done by taking the overlapping citations between OMIM and GOA annotations of the Molecular Function class where the annotations are based on experimental evidences described in these publications. Each of the sentences within this collection was then manually classified by a domain expert according to several types, whether the sentence corresponds to:

- a) a gene description (in the sense as defined for the baseline Prodisen)
- b) a gene/protein-disease association
- c) a gene/protein-function descriptions
- d) an interaction description
- e) a experimental method is mentioned

In addition a relevance score was provided to the gene description ranking from 0 to 10. The resulting collection contains a total of 5,455 manually classified sentences. A total of 3,445 function-related sentences were labeled, as well as 201 and 921 disease and interaction-related sentences respectively. Note that although all the abstracts were cited in OMIM, and thus have been used to derived associations of human genes to diseases, only few explicit gene to disease associations are contained in the abstracts. In contrast a considerable fraction of the sentences describe interactions (both genetic as well as physical interactions). The collection of interaction sentences form the disease Prodisen has been included as part of the training data provided from the Second BioCreAtIvE Challenge protein-interaction task (<http://biocreative.sourceforge.net/>).

4 Initial Analysis of the Prodisen Corpus

The initial inspection of the corpus reveals a clear association between the relative position of the informative sentence within an abstract and the classification as a gene description. Many of the gene descriptions are contained in the final part of abstracts (Figure 1). This is in line with the general discourse structure of scientific abstracts: (1) very general introduction, (2) more specific aspects, aims or problems, (3) experimental methods and (4) obtained results and conclusions. This is actually also in line of the average relevance ranks obtained for the disease Prodisen sentence classification, where a slight increase over the relative sentence intervals was observed.

Table 4: Detection of interaction sentences from Disease Prodisen collection. Type 1: bag of word approach after stop word removal; type 2: bag of word after stop word removal, stemming and case conversion; type 3: word- POS pairs, type 4: only stemmed words previously labeled as nouns using domain specific POS tagger (MedPost); type 5: only stemmed words previously labeled as verbs using domain specific POS tagger (MedPost); type 6: only stemmed words previously labeled as either verbs or nouns using domain specific POS tagger (MedPost); type 7: like type 6 but without stemming.

| Feature Types | Accuracy | Precision | Recall |
|---------------|----------|-----------|--------|
| Type 1 | 84.38% | 82.97% | 86.50% |
| Type 2 | 85.88% | 84.58% | 87.75% |
| Type 3 | 82.5% | 81.55% | 84.00% |
| Type 4 | 72.88% | 71.23% | 76.75% |
| Type 5 | 78.75% | 81.59% | 74.25% |
| Type 6 | 80.88% | 78.92% | 84.25% |
| Type 7 | 82.5% | 80.95% | 85.00% |

To estimate the number of protein mentions which appear in description sentences of the enriched Prodisen corpus we applied automatic bio-entity recognition (ABNER system, [15]). The mean number of protein entity mention occurrences detected in this collection was of 1.91 compared to 1.05 of the sentences which do not correspond to description relevant sentence types. When considering the fraction of sentences which lack protein mentions detected by the Bio-NER system, only 0.18 % of the description sentences contained either no direct protein mention or protein mentions which were not detected by the entity recognition tool. This fraction was considerable higher in case of the sentences which do not correspond to protein descriptions (0.40%). These differences were not as outstanding for other biological entities, namely cell lines, cell types and DNA.

Finally in order to analyze verbs which are often found in protein and gene description sentences we applied POS tagging of the description sentences in the enriched Prodisen corpus. This resulted in a collection of 232 verbs often found in protein description sentences. Among the most frequent verbs which appear in the descriptions are *expressed*, *required*, *identified*, *involved*. Both these verbs as well as POS-patterns extracted from the gene description corpus are available as additional materials.

5 Detecting Protein Interaction Descriptions

The detection of protein-protein interactions, has been addressed using both, experimental as well as bioinformatics methods. Characterizations of protein interactions have been of practical relevance to understand cell signaling pathways known to be involved in cancer. Many interactions are described in the literature and a considerable fraction of the the protein descriptions within the disease Prodisen subset belong to this class. To demonstrate the use Prodisen for the detection of interaction descriptions from text we implemented a interaction sentence classifier. The retrieval of interaction-related sentences is actually the first step in efficient protein-interaction extraction. Most existing interaction extraction systems are based on the co-occurrence of protein or gene mentions within abstracts or sentences, regardless whether these are actually related to interaction information. We constructed a sentence classifier which uses the sentences of the disease Prodisen corpus classified manually as interaction relevant and not relevant to derive a suitable training and test set. This collection contains a balanced set of sentences, 1,842 in total, thereof 921 interaction relevant. Several strategies to select and process features for the classifier have been tested, including stemming, case conversion, and Part-of Speech tagging (see Table 4). A Support Vector Machine classifier was used, and several types of kernel functions tested (linear, polynomial and radial basis function). No significant differences

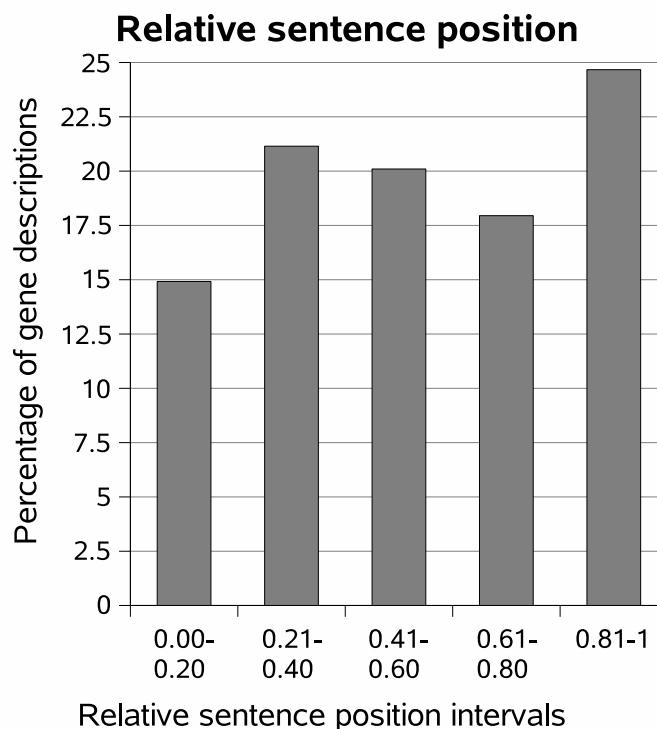


Figure 1: Relative sentence position intervals vs. percentage gene description sentences, derived from the enriched Prodisen corpus. Most of the gene description sentences are contained in the final part of abstracts. The obtained percentages of description sentences for each position are confirmed by the generally used discourse structure of abstracts, where the first sentences are very general statements, followed by more detailed aspects, experimental methods and obtained results.

between the used kernel functions were encountered. The best performing strategy was based on a bag of word feature approach, with previous stop word filtering, Porter stemming and case conversion. This approach reached an accuracy of 85.88% and a precision/recall of 84.58%/87.75% respectively on the test set. Although the recognition of functional verbs to detect interactions has been extensively explored, when used as features in the sentence classification approach its main limitation is related to a moderate recall, in part associated to the difficulty in detecting domain specific verbs which are often mislabeled using current POS-taggers. The most discriminative stemmed feature words for the classifier for the positive class (interaction relevant) included *interact*, *bind* and *domain*.

6 Prodisen Availability

The Prodisen corpus (both the random and the enriched set) is available online at: <http://www.pdg.cnb.uam.es/martink/PRODISEN/>

It is provided in an easy to use format together with the Prodisen construction script, to ensure both, the reconstruction and the possible customized extension of the Prodisen corpus. The Prodisen corpus will be revised and extended periodically.

7 Discussion and Conclusions

One of the difficulties in developing text mining applications in biology is the general lack of suitable training and test data. We explored the construction of a corpus of protein and gene description sentences, which is not limited to specific description types. It has thus a more extended use, especially for sentence and passage retrieval and automatic protein summaries generation. The protein description types considered in Prodisen are directly related to the information stored in existing biological annotation databases, and thus reflects the main information demands posed by the biology community. The presented corpus construction strategy also constitutes a novel approach in terms of selection of the actual sentences for manual curation. The randomly selected PubMed entries are especially useful to estimate the amount of functional descriptions contained in PubMed, which can be up to 9.5 million sentences. Many of those sentences are semantically redundant, in the sense that the provided description appears a number of times, e.g. the association between insulin and diabetes. Nevertheless, it points out that there is still a considerable gap between the amount of functional descriptions contained in biological databases and those stored in the scientific literature.

The Prodisen corpus provides also useful data to connect information retrieval and information extraction efforts addressing the construction of combined protein description identification strategies, as it is crucial to select first those documents and sentences which are potentially relevant.

One of the observations made during the process of building Prodisen is that a number of descriptions do not explicitly mention the gene or protein names but use referring expressions (e.g. anaphora) instead. So far, only a few attempts have been made to develop tools for anaphora resolution specifically for the biomedical domain.

We also propose a strategy to construct a gene description enriched corpus by selecting those articles which have been used by multiple different databases, each focusing on different description types. Using this approach, the selected abstracts have a higher probability of containing a large number of protein description sentences, as they have been curated by independent databases and contain a variety of different description types, often based on experimental results.

The manual description classification required often deep inference based on contextual information as well as on the background knowledge of the domain experts, which may result in more than one interpretation. Considerable reader variability is often encountered also in other domains such as in clinical sciences for agreement in screening of mammography [7]. We demonstrated the use of the Prodisen corpus in training and testing of information extraction techniques dedicated to the identification of gene/protein descriptions in text. It can be a useful resource for both, bag of words based approaches focusing on word frequencies, as well as for the discovery of description patterns and predicate argument structures (PAS) [17]. Additionally the corpus can be used to derive contextual word frequencies for protein mention discovery (disambiguation). Finally, the availability of the Prodisen construction script will facilitate the extension of the corpus with additional data, or with the creation of more specific types of protein descriptions.

References

- [1] Blaschke, C. and Valencia, A., The potential use of SUISEKI as a protein interaction discovery tool, *Genome Inform.*, 12:123–134, 2001.
- [2] Blaschke, C., Leon, E.A, Krallinger, M., and Valencia, A., Evaluation of BioCreAtIvE assessment of task 2, *BMC Bioinformatics*, 6(Suppl. 1):S16, 2005.
- [3] Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R., The Gene Ontology Annotation (GOA) Database: Sharing knowledge in UniProt with Gene Ontology, *Nucleic Acids Res.*, 32(Database issue):D262–D266, 2004.

- [4] Camon, E.B., Barrell, D.G., Dimmer, E.C., Lee, V., Magrane, M., Maslen, J., Binns, D., and Apweiler, R., An evaluation of GO annotation retrieval for BioCreAtIvE and GOA., *BMC Bioinformatics*, 6(Suppl. 1):S17, 2005.
- [5] Cohen, K.B., Fox, L., Ogren, P.V., and Hunter, L., Corpus design for biomedical natural language processing, *Proc. ACL-ISMB Workshop*, 38–45, 2005.
- [6] Franzen, K., Eriksson, G., Olsson, F., Asker, L., Liden, P., and Coster, J., Protein names and how to find them, *Int. J. Med. Inf.*, 67(1–3):49–61, 2002.
- [7] Gram, I.T., Bremnes, Y., Ursin, G., Maskarinec, G., Bjurstam, N., and Lund, E., Percentage density, Wolfe’s and Tabar’s mammographic patterns: Agreement and association with risk factors for breast cancer, *Breast Cancer Res.*, 7(5):R854–R861, 2005.
- [8] Hersh, W., Evaluation of biomedical text-mining systems: Lessons learned from information retrieval, *Brief. Bioinform.*, 6(4):344–356, 2005.
- [9] Kim, J.D., Ohta, T., Tateisi, Y., and Tsujii, J., GENIA corpus–semantically annotated corpus for bio-textmining, *Bioinformatics*, 19(Suppl. 1):i180–i182, 2003.
- [10] Krallinger, M., Padron, M., and Valencia, A., A sentence sliding window approach to extract protein annotations from biomedical articles, *BMC Bioinformatics*, 6(Suppl. 1):S19, 2005.
- [11] Kundel, H.L. and Polansky, M., Measurement of observer agreement, *Radiology*, 228(2):303–308, 2003.
- [12] Mitchell, J.A., Aronson, A.R., Mork, J.G., Folk, L.C., Humphrey, S.M., and Ward, J.M., Gene indexing: Characterization and analysis of NLM’s GeneRIFs, *AMIA Annu. Symp. Proc.*, 460–464, 2003.
- [13] Perez-Iratxeta, C., Bork, P., and Andrade, M.A., Association of genes to genetically inherited diseases using data mining, *Nat. Genet.*, 31(3):316–319, 2002.
- [14] Raychaudhuri, S., Chang, J.T., Sutphin, P.D., and Altman, R.B., Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature, *Genome Res.*, 12(1):203–214, 2002.
- [15] Settles, B., Biomedical named entity recognition using conditional random fields and rich feature sets, *Proc. NLPBA/COLING*, 2004.
- [16] Stoica, E. and Hearst, M., Predicting gene functions from text using a cross-species approach, *Pac. Symp. Biocomput.*, 11:88–99, 2006.
- [17] Wattarujeekrit, T., Shah, P.K., and Collier, N., PASBio: Predicate-argument structures for event extraction in molecular biology, *BMC Bioinformatics*, 5:155, 2004.
- [18] Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A., and Wagner, L., Database resources of the National Center for Biotechnology, *Nucleic Acids Res.*, 31(1):28–33, 2003.
- [19] Yeh, A., Morgan, A., Colosimo, M., and Hirschman, L., BioCreAtIvE Task 1A: Gene mention finding evaluation, *BMC Bioinformatics*, 6(Suppl. 1):S2, 2005.