EVEX: A PubMed-Scale Resource for Homology-Based Generalization of Text Mining Predictions

Sofie Van Landeghem[1,2], **Filip Ginter**[3], **Yves Van de Peer**[1,2] **and Tapio Salakoski**[3,4]
1. Dept. of Plant Systems Biology, VIB, Belgium
2. Dept. of Plant Biotechnology and Genetics, Ghent University, Belgium
3. Dept. of Information Technology, University of Turku, Finland
4. Turku Centre for Computer Science (TUCS), Finland
solan@psb.ugent.be, ginter@cs.utu.fi
yvpee@psb.ugent.be, tapio.salakoski@utu.fi

Abstract

In comparative genomics, functional annotations are transferred from one organism to another relying on sequence similarity. With more than 20 million citations in PubMed, text mining provides the ideal tool for generating additional large-scale homology-based predictions. To this end, we have refined a recent dataset of biomolecular events extracted from text, and integrated these predictions with records from public gene databases. Accounting for lexical variation of gene symbols, we have implemented a disambiguation algorithm that uniquely links the arguments of 11.2 million biomolecular events to well-defined gene families, providing interesting opportunities for query expansion and hypothesis generation. The resulting MySQL database, including all 19.2 million original events as well as their homology-based variants, is publicly available at http://bionlp.utu.fi/.

1 Introduction

Owing to recent advances in high-throughput sequencing technologies, whole genomes are being sequenced at an ever increasing rate (Metzker, 2010). However, for the DNA sequence to truly unravel its secrets, structural annotation is necessary to identify important elements on the genome, such as coding regions and regulatory motifs. Subsequently, functional annotation is crucial to link these structural elements to their biological function.

Functional annotation of genomes often requires extensive *in vivo* experiments. This time-consuming procedure can be expedited by integrating knowledge from closely related species (Fulton et al., 2002; Proost et al., 2009). Over the past few years, homology-based functional annotation has become a widely used technique in the bioinformatics field (Loewenstein et al., 2009).

Unfortunately, many known genotype-phenotype links are still buried in research articles: The largest biomolecular literature database, PubMed, consists of more than 20 million citations[1]. Due to its exponential growth, automated tools have become a necessity to uncover all relevant information.

There exist several text mining efforts focusing on pairwise interactions and co-occurrence links of genes and proteins (Hoffmann and Valencia, 2004; Ohta et al., 2006; Szklarczyk et al., 2011). In this paper, we present the first large-scale text mining resource which both utilizes a detailed event-based representation of biological statements and provides homology-based generalization of the text mining predictions. This resource results from the integration of text mining predictions from nearly 18M PubMed citations with records from public gene databases (Section 2). To enable such integration, it is crucial to first produce canonical forms of the automatically tagged biological entities (Section 3.1). A gene symbol disambiguation algorithm then links these canonical forms to gene families and gene identifiers (Section 3.2). Finally, a MySQL-driven framework aggregates the text-bound event occurrences into generalized events, creating a rich resource of homology-based predictions extracted from text (Section 3.3).

[1] http://www.ncbi.nlm.nih.gov/pubmed/

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 28–37,
Portland, Oregon, USA, June 23-24, 2011.

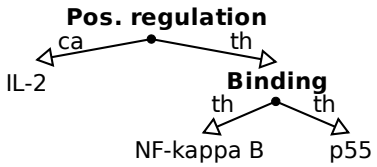

Figure 1: Event representation of the statement *IL-2 acts by enhancing binding activity of NF-kappa B to p55*, illustrating recursive nesting of events where the (th)eme of the *Positive regulation* event is the *Binding* event. The (ca)use argument is the gene symbol *IL-2*.

2 Data

Our integrative approach is based on two types of data: text mining predictions generated for the whole of PubMed (Section 2.1) and publicly available gene database records (Section 2.2).

2.1 Text mining predictions

Björne et al. (2010) have applied to all PubMed abstracts an event extraction pipeline comprising of the BANNER named entity recognizer (Leaman and Gonzalez, 2008) and the Turku Event Extraction System (Björne et al., 2009). The resulting dataset contains 36.5M occurrences of gene / gene product (GGP) entities and 19.2M occurrences of events pertaining to these entities.

The file format and information scheme of the resource correspond to the definition of the BioNLP'09 Shared Task on Event Extraction (Kim et al., 2009). Events are defined as typed relations between arguments that are either entity occurrences or, recursively, other events. There are nine possible event types: *Localization, Binding, Gene expression, Transcription, Protein catabolism, Phosphorylation, Regulation, Positive regulation,* and *Negative regulation*. Further, arguments are assigned a role: *Theme* or *Cause* for the core arguments and *AtLoc, ToLoc, Site,* and *CSite* for auxiliary arguments that define additional information such as cellular location of the event. In addition, each event occurrence may be marked as negative and/or speculative. Figure 1 depicts an example event.

2.2 Database records

During the last few decades, several large-scale databases have been designed to deal with the abundance of data in the field of life sciences. In this study, we are particularly interested in databases of gene symbols and homologous gene groups or gene families. These families are composed by clustering pairwise orthologs, which are genes sharing common ancestry evolved through speciation, often having a similar biological function.

Entrez Gene[2] is the default cross-species gene nomenclature authority, hosted by NCBI (Sayers et al., 2009). It bundles information from species-specific resources as well as from RefSeq records[3]. More than 8M Entrez Gene identifiers were collected from over 8,000 different taxa, all together referring to more than 10M distinct gene symbols, descriptions, abbreviations and synonyms. While Entrez Gene IDs are unique across taxa, gene symbols are highly ambiguous. Section 3 describes how we tackle gene symbol ambiguity across and within species.

The HomoloGene[4] database is also hosted at NCBI and provides the results of automated detection of orthologs in 20 completely sequenced eukaryotic genomes. From this resource, around 43,700 HomoloGene families were extracted, containing about 242,000 distinct genes. A second set of gene families was retrieved from Ensembl (Flicek et al., 2011). More than 13,000 Ensembl clusters were assembled comprising about 220,000 genes.

As a general rule, the functional similarity scores per homologous pair in a gene family are higher when more stringent criteria are used to define the families (Hulsen et al., 2006). While HomoloGene consists of many strict clusters containing true orthologs, bigger Ensembl clusters were obtained by assembling all pairwise orthologous mappings between genes. Ultimately, such clusters may also include paralogs, genes originated by duplication. As an example, consider the *nhr-35* gene from *C. elegans*, which has both *Esr-1* and *Esr-2* as known orthologs, resulting in the two paralogs being assigned to the same final Ensembl cluster. The Ensembl clustering algorithm can thus be seen as a more coarse-grained method while the HomoloGene mapping results in more strictly defined gene families. The implications are discussed on a specific use-case in Section 4.3.1.

[2] http://www.ncbi.nlm.nih.gov/gene
[3] http://www.ncbi.nlm.nih.gov/refseq
[4] http://www.ncbi.nlm.nih.gov/homologene

3 Methods

Widely known biomolecular events occur in many different articles, often mentioning a different gene synonym or lexical variant. Canonicalization of the entity occurrences deals with these lexical variants (Section 3.1), while the disambiguation algorithm then uniquely links canonical forms to a gene families (Section 3.2). In a final step, these links can be used to generalize the text mining events to their homology-based variants (Section 3.3).

3.1 Canonicalization of the entity occurrences

The entity occurrences predicted by BANNER (Section 2.1) follow the guidelines of GENETAG (Tanabe et al., 2005), the corpus it was trained on. These guidelines allow not only gene and gene products, but also related entities such as protein complexes and gene promoters. Furthermore, BANNER frequently tags noun phrases such as *human Esr-1 gene* rather than only the minimal symbol *Esr-1*.

To enable integration of text mining predictions with external databases, it is necessary to refine the entity occurrences to canonical forms that can be linked to gene records such as those in Entrez Gene. To this end, common prefixes and suffixes such as *gene* and *wild-type* should be removed.

In a first step towards canonicalization of the entities, a mapping table was assembled containing common contexts in which a gene symbol appears and where the full noun phrase can be reduced to that embedded symbol for the sake of information retrieval (Table 1). This mapping table was created by matching[5] a list of candidate minimal gene symbols to the extracted BANNER entities.

To define the list of candidate minimal gene symbols, two approaches have been combined. First, a set of around 15,000 likely gene symbols is extracted by looking for single token strings that were tagged by BANNER at least 50% of the times they occur in a PubMed abstract. Secondly, all official gene names are extracted from Entrez Gene. As this latter list also contains common English words such as *was* and *protein*, we have only selected those that were likely to be standalone gene symbols. We calculate this likelihood by $C_s/(C_s + C_n)$ where C_s is the number of times a string is tagged standalone and C_n is the number of times the string occurs in PubMed but is not tagged (neither as standalone, nor as part of a larger entity). This likelihood represents the proportion of standalone occurrences of the string that are tagged. We experimentally set a threshold on this value to be higher than 0.01, excluding a list of 2,865 common English words.

GGP contexts
−ORG− −GGP− gene
−GGP− sequences
mutant −GGP− proteins
−GGP− homologs
cytoplasmic wild-type −GGP−

Table 1: This table lists a few examples of entity occurrences extracted with BANNER that are resolved to the embedded minimal gene symbol (marked as −GGP−).

Subsequently, all BANNER entity occurrences are screened and likely minimal gene symbols substituted with −GGP−, resulting in generalized contexts. Then, we have matched these contexts with an extensive list of organism names from the Linneaus distribution (Gerner et al., 2010) and a small collection of miscellaneous non-formal organism terms (e.g. *monkey*), replacing all known organisms with an −ORG− placeholder. Finally, we have excluded all contexts where the embedded GGP is likely to be functionally too far removed from the embedding noun phrase (e.g. "−GGP− inhibitor"), relying on a corpus defining and categorizing such relationships (Ohta et al., 2009). Some of the contexts that were retained after this step, such as "−GGP− mutant" or "−GGP− promoter" still refer to entities that are distinctly different from the embedded GGP. These results are considered valid, as the goal of the affix stripping algorithm is to increase recall and offer explorative results involving various types of information on gene symbols.

The final list of contexts, generalized with −GGP− and −ORG− placeholders, is split into two separate lists of prefixes and suffixes, ranked by frequency. Also, numerical affixes as well as those shorter than 3 characters are discarded from these lists.

[5] All string matching steps have been implemented using the *SimString* string retrieval library (Okazaki and Tsujii, 2010).

Each text-bound entity occurrence can then be canonicalized by applying the following algorithm:

1. Replace all organism names with the placeholder `-ORG-`
2. If the string can be matched[6] to a known symbol in Entrez Gene, stop the algorithm
3. Find all occurring affixes and strip the one associated with the highest count
4. Repeat (2-3) until no more affixes match
5. Strip remaining `-ORG-` placeholders and all whitespace and non-alphanumeric characters

For example, the canonicalization of *human anti-inflammatory il-10 gene* proceeds as *-ORG- anti-inflamatory il-10 gene* → *anti-inflammatory il-10 gene* → *anti-inflammatory il-10* → *il-10*, at which point the string *il10* is matched in Entrez Gene, becoming the final canonical form. In the following section, we describe how these canonical forms are assigned to unique gene families.

3.2 Disambiguation of gene symbols

Gene name ambiguity is caused by the lack of community-wide approved standards for assigning gene symbols (Chen et al., 2005). Furthermore, authors often introduce their own lexical variants or abbreviations for specific genes.

From Entrez Gene, we have retrieved 8,034,512 gene identifiers that link to 10,177,542 unique symbols. Some of these symbols are highly ambiguous and uninformative, such as *NEWENTRY*. Others are ambiguous because they are abbreviations. Finally, many symbols can not be linked to one unique gene, but do represent a homologous group of genes sharing a similar function. Often, orthologs with similar functions are assigned similar official gene names.

The first step towards gene symbol disambiguation involves collecting all possible synonyms for each gene family from either Ensembl or HomoloGene. We strip these symbols of all whitespace and non-alphanumeric characters to match the final step in the canonicalization algorithm.

The disambiguation pipeline then synthesizes the ambiguity for all gene symbols by counting their occurrences in the gene families. Each such relation records whether the symbol is registered as an official or default gene symbol, as the gene description, an abbreviation, or a synonym. As an example, Table 2 depicts the intrinsic ambiguity of *esr1*.

Family	Type of symbol	Count
HG:47906	Default symbol	7
HG:99739	Synonym	1
HG:3740	Synonym	1
ECL:10415	Default symbol	12
ECL:8731	Synonym	1
ECL:8226	Synonym	1

Table 2: Intrinsic ambiguity of *esr1*, analysed in both HomoloGene (HG) and Ensembl clusters (ECL).

In a subsequent step, the ambiguity is reduced by applying the following set of rules, relying on a priority list imposed on the type of the symbol, ensuring we choose an official or default symbol over a description or synonym.

1. If one family has the most (or all) hits for a certain symbol and these hits refer to a symbol type having priority over other possibilities, this family is uniquely assigned to that symbol.
2. If a conflict exists between one family having the highest linkage count for a certain symbol, and another family linking that symbol to a higher priority type, the latter is chosen.
3. If two families have equal counts and type priorities for a certain symbol, this symbol can not be unambiguously resolved and is removed from further processing.
4. If the ambiguity is still not resolved, all families with only one hit for a certain symbol are removed, and steps 1-3 repeated.

The above disambiguation rules were applied to the 458,505 gene symbols in HomoloGene. In the third step, 6,891 symbols were deleted, and when the algorithm ends, 555 symbols remained ambiguous. In total, 451,059 gene symbols could thus be uniquely linked to a HomoloGene family (98%). In the *esr1* example depicted in Table 2, only the link to HG:47906 will be retained. The results for Ensembl were very similar, with 342,252 out of 345,906 symbols uniquely resolved (99%).

[6]The comparison is done ignoring whitespace and non-alphanumeric characters.

	All	Ensembl	HomoloGene
No stripping	39.9 / 67.5 / 50.2	62.8 / 70.0 / 66.2	64.2 / 69.2 / 66.6
Affix stripping	48.7 / 82.3 / 61.1	61.7 / 88.0 / 72.5	62.8 / 87.9 / 73.3

Table 3: Influence on precision, recall and F-measure (given as P/R/F) of the affix stripping algorithm on the entity recognition module, as measured across all BioNLP'09 ST entity occurrences and also separately on the subsets which can be uniquely mapped to Ensembl and HomoloGene (77.3% and 75.5% of all occurrences, respectively).

3.3 Homology-based generalization of the text mining events

In order to gain a broader insight into the 19.2M event occurrences obtained by Björne et al. (2010), it is necessary to identify and aggregate multiple occurrences of the same underlying event. This generalization also notably simplifies working with the data, as the number of generalized events is an order of magnitude smaller than the number of event occurrences.

To aggregate event occurrences into generalized events, it is necessary to first define equivalence of two event occurrences: Two event occurrences are equivalent, if they have the same event type, and their core arguments are equivalent and have the same roles. For arguments that are themselves events, the equivalence is applied recursively. The equivalence of arguments that are entities can be established in a number of different ways, affecting the granularity of the event generalization. One approach is to use the string canonicalization described in Section 3.1; two entities are then equivalent if their canonical forms are equal. This, however, does not take symbol synonymy into account. A different approach which we believe to be more powerful, is to disambiguate gene symbols to gene families, as described in Section 3.2. In this latter approach, two entity occurrences are deemed equivalent if their canonical forms can be uniquely resolved to the same gene family. Consequently, two event occurrences are considered equivalent if they pertain to the same gene families.

As both approaches have their merits, three distinct generalization procedures have been implemented: one on top of the canonical gene symbols, and one on top of the gene families defined by HomoloGene and Ensembl, respectively.

4 Results and discussion

4.1 Evaluation of entity canonicalization

The affix stripping step of the canonicalization algorithm described in Section 3.1 often substantially shortens the entity strings and an evaluation of its impact is thus necessary. One of the primary objectives of the canonicalization is to increase the proportion of entity occurrences that can be matched to Entrez Gene identifiers. We evaluate its impact using manually tagged entities from the publicly available BioNLP'09 Shared Task (ST) training set, which specifically aims at identifying entities that are likely to match gene and protein symbol databases (Kim et al., 2009). Further, the ST set comprises of PubMed abstracts and its underlying text is thus covered in our data. Consequently, the ST training set forms a very suitable gold standard for the evaluation.

First, we compare[7] the precision and recall of the BANNER output before and after affix stripping (Table 3, first column). The affix stripping results in a notable gain in both precision and recall. In particular, the nearly 15pp gain on recall clearly demonstrates that the affix stripping results in entity strings more likely to match existing resources.

Second, the effect of affix stripping is evaluated on the subset of entity strings that can be uniquely mapped into Ensembl and HomoloGene (77.3% and 75.5% of the ST entity strings, respectively). This subset is of particular interest, since the generalized events are built on top of the entities that can be found in these resources and any gain on this particular subset is thus likely to be beneficial for the overall quality of the generalized events. Here, affix stripping leads to a substantial increase in recall when compared to no stripping being applied

[7] The comparison is performed on the level of bags of strings from each PubMed abstract, avoiding the complexity of aligning character offsets across different resources.

	Entities	Ent. occ.
Canonical	1.6M (100%)	36.4M (100%)
HomoloGene	64.0K (3.9%)	18.8M (51.7%)
Ensembl	54.6K (3.3%)	18.7M (51.2%)

Table 4: Entity coverage comparison. The *entities* column gives the number of canonical entities, also shown as a percentage of all unique, canonical BANNER entities (1.6M). The *entity occurrences* column shows the number of occurrences for which the generalization could be established, out of the total number of 36.4M extracted BANNER entities.

	Events	Ev. occ.
Canonical	3223K	19.2M (100%)
HomoloGene	614K	10.2M (53%)
Ensembl	505K	10.2M (52.9%)

Table 5: Comparison of the three event generalization methods. The *events* column gives the number of generalized events and the *event occurrences* column shows the number of occurrences for which the generalization could be established, out of the total number of 19.2M text-bound event occurrences.

(around 18pp), which is offset by a comparatively smaller drop in precision (less than 2pp). Global performance increases with about 6.5pp in F-score for both the Ensembl and HomoloGene subsets.

Björne et al. (2010) used a simpler, domain-restricted affix stripping algorithm whereby candidate affixes were extracted only from NP-internal relations in the GENIA corpus (Ohta et al., 2009). This original algorithm affects 11.5% unique entity strings and results in 3.5M unique canonical forms and 4.5M unique events. In comparison, our current affix stripping algorithm results in 1.6M unique canonical forms and 3.2M unique events, thus demonstrating the improved generalization capability of the current affix stripping algorithm.

4.2 Evaluation of homology-based disambiguation

The symbol to gene family disambiguation algorithm succesfully resolves almost all gene symbols in HomoloGene or Ensembl (Section 3.2). However, not all genes are a member of a known gene family, and the event generalization on top of the gene families will thus inevitably discard a significant portion of the text mining predictions.

Table 4 shows that only a small fraction of all unique canonical entities matches the gene families from HomoloGene or Ensembl (3.9% and 3.3%, respectively). However, this small fraction of symbols accounts for approximately half of all entity occurrences in the text mining data (51.7% and 51.2%). The algorithm thus discards a long tail of very infrequent entities. Table 5 shows a similar result for the events and event occurrences. We find that mapping to HomoloGene and Ensembl results in a considerably smaller number of generalized events, yet accounts for more than half of all event occurrences (53% and 52.9%, respectively).

Finally, merging the canonical entities and the corresponding generalized events for both HomoloGene and Ensembl, we can assess the percentage of all text mining predictions that can be linked to at least one homology-based variant: 21.8M (59.8%) of all entity occurrences and 11.2M (58.4%) of all event occurrences can be resolved. Nearly 60% of entity and event occurrences in the original text mining data could thus be uniquely linked to well defined gene families. Also, as shown in Section 4.1, the 60% entities retained are expected to contain proportionally more true positives, compared to the 40% entities that could not be mapped. One might speculate that a similar effect will be seen also among events.

4.3 MySQL database and Use-cases

As the PubMed events extracted by Björne et al. (2010) are purely text-bound and distributed as text files, they can not easily be searched. One important contribution of this paper is the release of all text mining predictions as a MySQL database. During the conversion, all original information is kept, including links to the PubMed IDs and the offsets in text for all entities and triggers, referring to the original strings as they were obtained by BANNER and the event extraction system. This allows for fast retrieval of text mining data on a PubMed-scale.

As described in Section 3.3, three distinct generalization methods have been applied to the original events. On the database level, each generalization is represented by a separate set of tables for the generalized events and their arguments, aggregating important event statistics such as occurrence count and

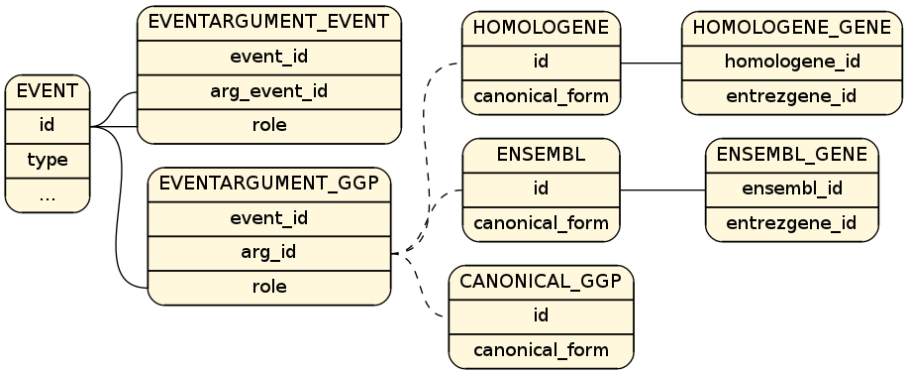

Figure 2: Database scheme of the generalized events. Three instantiations of the general scheme (i.e. the three leftmost tables) exist in the database. Following the dotted lines, each instance links to a different table in which the canonical forms and the gene identifiers can be looked up.

negation/speculation information (Figure 2). Table 5 states general statistics for the three different sets. Finally, a mapping table is provided that links the generalized events to the event occurrences from which they were abstracted. More technical details on the MySQL scheme and example queries can be found at `http://bionlp.utu.fi/`.

4.3.1 Use case: Query expansion

The MySQL database is the ideal resource to retrieve information on a PubMed-scale for a certain gene or set of genes. Suppose there would be an interest in *Esr-1*, then all abstract events on top of the canonical form *esr1* can be retrieved. However, results will display events for both the *Estrogen receptor* as well as for the much less common *Enhancer of shoot regeneration*. Furthermore, it makes no sense to add known synonyms of both genes to the query, as this will generate an incoherent list of synonyms and even more false positive hits.

In such a case, it is to be recommended to use the homology-based generalization of the events. For example, *esr1* hits the HomoloGene family HG:47906, which contains all *Estrogen receptor-alpha* genes across eukaryotic species. Canonical symbols linked to this family include *era*, *estra*, *nr3a1* and *estrogenreceptor1alpha*.

A similar analysis can be done for the Ensembl clustering, where *esr1* links to ECL:10415. However, this more coarse-grained Ensembl family contains all genes from the two closely related subgroups *Estrogen receptor* and *Estrogen related receptor*, both belonging to the *Estrogen Receptor-like* group of the superfamily of nuclear receptors (Zhang et al., 2004). On top of the synonyms mentioned previously, this family thus also includes *erb*, *esr2b*, *errbetagamma* and *similartoesrrbprotein*. By using this list for query expansion, more general text mining predictions can be retrieved.

It is to be noted that both homology-based approaches will also include events mentioning *Esr-1* as the abbreviation for *Enhancer of shoot regeneration*. While this usage is much less common, it will result in a few false positive hits. These false positives may be prevented by taking into account local context such as organism mentions, as the *Enhancer of shoot regeneration* gene is only present in *A. thaliana*. We believe our current homology-based approach could be integrated with existing or future normalization algorithms (Krallinger and Hirschman, 2007; Wermter et al., 2009) to provide such fine-grained resolution. This is regarded as interesting future work.

4.3.2 Use case: Homology-based hypotheses

Consider a newly annotated, protein-coding gene for which no database information currently exists. To generate homology-based text mining hypotheses, the orthologs of this gene first have to be defined by assessing sequence similarity through BLAST (Altschul et al., 1997).

Imagine for example a newly sequenced genome X for which a gene similar to the mouse gene *Esr-1* is identified. This gene will soon be known as "genome X *Esr-1*" and thus related to the *Esr-1* gene family. As described in Section 4.3.1, homology-

based query expansion can then be used to retrieve all events involving lexical variants and synonyms of the canonical string *esr1*.

5 Conclusions

We present a large-scale resource for research and application of text mining from biomedical literature. The resource is obtained by integrating text mining predictions in the dataset of Björne et al. (2010) with public databases of gene symbols and gene families: Entrez Gene, Ensembl, and HomoloGene. The integration is performed on the level of gene families, allowing for a number of novel use cases for both text mining and exploratory analysis of the biological statements in PubMed literature. To achieve the linking between text-based event predictions and gene databases, several algorithms are introduced to solve the problems involved.

First, we propose an algorithm for stripping affixes in entity occurrences tagged by the BANNER named entity recognizer, addressing the problem of such entities often including wider context which prevents direct matching against gene symbol databases. Using the BioNLP'09 Shared Task data as gold standard, we show that the algorithm substantially increases both precision and recall of the resulting canonical entities, the gain in recall being particularly pronounced.

Second, we propose an algorithm which assigns to the vast majority of gene symbols found in HomoloGene and Ensembl a single unique gene family, resolving the present intra-organism ambiguity based on symbol occurrence statistics and symbol type information. Matching these disambiguated symbols with the affix-stripped canonical forms of entity occurrences, we were able to assign a unique gene family from either HomoloGene or Ensembl to nearly 60% of all entities in the text, thus linking the text-bound predictions with gene databases.

Finally, we use the resolution of entity occurrences to unique gene families to generalize the events in the text mining data, aggregating together event occurrences whose arguments are equivalent with respect to their gene family. Depending on whether HomoloGene or Ensembl is used for the gene family definition, this generalization process results in 500K-600K generalized events, which together aggregate over 11.2M (58.4%) of all event occurrences in the text mining data. Being able to link the literature-based events with well-defined gene families opens a number of interesting new use-cases for biomedical text mining, such as the ability to use the homology information to search for events relevant to newly discovered sequences. The remaining 41.6% of event occurrences not generalizable to gene families can still be retrieved through an additional generalization on the level of entity canonical forms.

All relevant data, namely all original events and entities together with their canonical forms, the generalizations of events based on canonical entity forms and gene families, as well as the gene symbol to unique family mapping are made publicly available as records in a MySQL database. We also provide detailed online documentation of the database scheme and example queries. Finally, we release the affix lists used in the canonicalization algorithm.

We believe this resource to be very valuable for explorative analysis of text mining results and homology-based hypothesis generation, as well as for supporting future research on data integration and biomedical text mining.

One important future work direction is a further disambiguation of canonical gene symbols to unique gene identifiers rather than entire gene families, which would allow for more fine-grained event generalization. There is an ongoing active, community-wide research focusing on this challenge and the current resource could be integrated as an additional source of information. Another future work direction is to create a visualization method and a web interface which would allow simple, user-friendly access to the data for researchers outside of the BioNLP research community itself.

Acknowledgments

The authors would like to thank Sampo Pyysalo (University of Tokyo) and Jari Björne (University of Turku) for their contribution. SVL would like to thank the Research Foundation Flanders (FWO) for funding her research and a travel grant to Turku. This work was partly funded by the Academy of Finland and the computational resources were provided by CSC IT Center for Science Ltd., Espoo, Finland.

References

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, September.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 10–18, Morristown, NJ, USA. Association for Computational Linguistics.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Scaling up biomedical event extraction to the entire PubMed. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lifeng Chen, Hongfang Liu, and Carol Friedman. 2005. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21:248–256, January.

Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Eugene Kulesha, Pontus Larsson, Ian Longden, William McLaren, Bert Overduin, Bethan Pritchard, Harpreet Singh S. Riat, Daniel Rios, Graham R. Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Giulietta Spudich, Y. Amy Tang, Stephen Trevanion, Jana Vandrovcova, Albert J. Vilella, Simon White, Steven P. Wilder, Amonida Zadissa, Jorge Zamora, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M. Fernández-Suarez, Javier Herrero, Tim J. Hubbard, Anne Parker, Glenn Proctor, Jan Vogel, and Stephen M. Searle. 2011. Ensembl 2011. *Nucleic acids research*, 39(Database issue), January.

Theresa M. Fulton, Rutger Van der Hoeven, Nancy T. Eannetta, and Steven D. Tanksley. 2002. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell*, 14(5):1457–1467.

Martin Gerner, Goran Nenadic, and Casey M. Bergman. 2010. LINNAEUS: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85+, February.

Robert Hoffmann and Alfonso Valencia. 2004. A gene network for navigating the literature. *Nat Genet*, 36(7):664, Jul.

Tim Hulsen, Martijn Huynen, Jacob de Vlieg, and Peter Groenen. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*, 7(4):R31+, April.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Martin Krallinger and Lynette Hirschman, editors. 2007. *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, Madrid, April.

Robert Leaman and Graciela Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 652–663.

Yaniv Loewenstein, Domenico Raimondo, Oliver C. Redfern, James Watson, Dmitrij Frishman, Michal Linial, Christine Orengo, Janet Thornton, and Anna Tramontano. 2009. Protein function annotation by homology-based inference. *Genome biology*, 10(2):207, February.

Michael L. Metzker. 2010. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):31–46, January.

Tomoko Ohta, Yusuke Miyao, Takashi Ninomiya, Yoshimasa Tsuruoka, Akane Yakushiji, Katsuya Masuda, Jumpei Takeuchi, Kazuhiro Yoshida, Tadayoshi Hara, Jin-Dong Kim, Yuka Tateisi, and Jun'ichi Tsujii. 2006. An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 17–20, Sydney, Australia, July. Association for Computational Linguistics.

Tomoko Ohta, Sampo Pyysalo, Kim Jin-Dong, and Jun'ichi Tsujii. 2009. A re-evaluation of biomedical named entity - term relations. In *Proceedings of LBM'09*.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 851–859, Beijing, China, August.

Sebastian Proost, Michiel Van Bel, Lieven Sterck, Kenny Billiau, Thomas Van Parys, Yves Van de Peer, and Klaas Vandepoele. 2009. PLAZA: A comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, 21(12):3718–3731, December.

Eric W. W. Sayers, Tanya Barrett, Dennis A. A. Benson, Stephen H. H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. M. Church, Michael Dicuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y. Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. J. Lipman, Thomas L. L. Madden, Donna R. R. Maglott, Vadim Miller, Ilene Mizrachi, James Ostell, Kim D. D. Pruitt, Gregory D. D. Schuler, Edwin Sequeira, Stephen T. T. Sherry, Martin Shumway, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tatiana A. A. Tatusova, Lukas Wagner, Eugene Yaschenko, and Jian Ye. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 37(Database issue):D5–15, January.

Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars J. Jensen, and Christian von Mering. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(Database issue):D561–D568, January.

Lorraine Tanabe, Natalie Xie, Lynne H. Thom, Wayne Matten, and W. John Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6 Suppl 1.

Joachim Wermter, Katrin Tomanek, and Udo Hahn. 2009. High-performance gene name normalization with GENO. *Bioinformatics*, 25(6):815–821.

Zhengdong Zhang, Paula E. Burch, Austin J. Cooney, Rainer B. Lanz, Fred A. Pereira, Jiaqian Wu, Richard A. Gibbs, George Weinstock, and David A. Wheeler. 2004. Genomic analysis of the nuclear receptor family: New insights into structure, regulation, and evolution from the rat genome. *Genome Research*, 14(4):580–590, April.