DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2013, Article ID bat020, doi:10.1093/database/bat020

Original article

Preliminary evaluation of the CellFinder literature curation pipeline for gene expression in kidney cells and anatomical parts

Mariana Neves[1,*], Alexander Damaschun[2], Nancy Mah[3], Fritz Lekschas[2], Stefanie Seltmann[2], Harald Stachelscheid[2], Jean-Fred Fontaine[3], Andreas Kurtz[2,4] and Ulf Leser[1]

[1]Humboldt-Universität zu Berlin, Knowledge Management in Bioinformatics, Berlin, 10099, Germany, [2]Berlin Brandenburg Center for Regenerative Therapies, Charité, Berlin, 13353, Germany, [3]Max Delbrück Center for Molecular Medicine, Berlin, 13092, Germany and [4]Adult Stem Cell Research Center, Seoul National University, College of Veterinary Medicine, Seoul, 151-742, Republic of Korea

***Corresponding author:** Tel: +49 (0) 30 2093 3902; Fax: +49 (0) 30 2093 5484; Email: neves@informatik.hu-berlin.de

Submitted 30 November 2012; Revised 28 February 2013; Accepted 12 March 2013

Citation details: Neves,M., Damaschun,A., Mah,N., *et al*. Preliminary evaluation of the CellFinder literature curation pipeline for gene expression in kidney cells and anatomical parts. *Database* (2013) Vol. 2013: article ID bat020; doi: 10.1093/database/bat020.

Biomedical literature curation is the process of automatically and/or manually deriving knowledge from scientific publications and recording it into specialized databases for structured delivery to users. It is a slow, error-prone, complex, costly and, yet, highly important task. Previous experiences have proven that text mining can assist in its many phases, especially, in triage of relevant documents and extraction of named entities and biological events. Here, we present the curation pipeline of the CellFinder database, a repository of cell research, which includes data derived from literature curation and microarrays to identify cell types, cell lines, organs and so forth, and especially patterns in gene expression. The curation pipeline is based on freely available tools in all text mining steps, as well as the manual validation of extracted data. Preliminary results are presented for a data set of 2376 full texts from which >4500 gene expression events in cell or anatomical part have been extracted. Validation of half of this data resulted in a precision of ~50% of the extracted data, which indicates that we are on the right track with our pipeline for the proposed task. However, evaluation of the methods shows that there is still room for improvement in the named-entity recognition and that a larger and more robust corpus is needed to achieve a better performance for event extraction.
Database URL: http://www.cellfinder.org/

Introduction

Biomedical literature curation is the process of automatically and/or manually compiling biological data from scientific publications and making it available in a structured and comprehensive way. Databases that integrate information derived in some way from scientific publications include, for instance, model organism databases (1), protein–protein interactions (2) and gene–chemical–disease relationships (3). Typical literature curation workflows include the following steps (4): triage (selection of relevant publications), biological entities identification (e.g. genes/proteins, diseases, etc.), extraction of relationships (e.g. protein–protein interactions, gene expression, etc.), association of biological processes with experimental evidence, data validation and recoding into the database. Therefore, literature curation requires a careful reading of publications by domain experts, which is known to be a time-consuming task. Additionally, the increasing growth of available publications prevents a comprehensive manual

Database, Vol. 2013, Article ID bat020, doi:10.1093/database/bat020

curation of intended facts and previous studies show that it is not feasible (5).

Recent advances in text mining methods have facilitated its application in most of the literature curation stages. Challenges have contributed to the improvement and availability of a variety of methods for named-entity prediction (6), and more specifically for gene/protein prediction and normalization (7, 8). Also binary relationships (9) and event extraction (10) have been improved, and its current performance allows its use on large scale projects (11). Finally, integrated ready-to-use workbenches have also been available, such as @Note (12), Argo (13), MyMiner (14) and Textpresso (15), although the performance and scalability to larger projects is still dubious for some of them. A comparison between some of them is found in this survey on annotation tools for the biomedical domain (16).

Previous reports (17, 18) and experiments (19) have confirmed the feasibility of text mining to assist literature curation and recent surveys (4, 20) show that, indeed, it is already part of many biological databases workflows. For instance, text mining support is being explored for the triage stage in FlyBase (21), for curation of regulatory annotation in (22) and also in the AgBase (23), Biomolecular Interaction Network Database (BIND) (24), Immune Epitope Database (IEDB) (25) and The Comparative Toxicogenomics Database (CTD) (26) databases. Additionally, many solutions have been proposed for the CTD database during a recent collaborative task (27). Further, Textpresso has been widely used to prioritize document and for Gene Ontology (GO) terms (28) annotation in WormBase and The Arabidopsis Information Resource (TAIR) (29). Named-entity recognition has also been included in the curation workflow of Mouse Genome Informatics (MGI) (30) for gene/protein extraction, and in Xenbase (31) for gene and anatomy terms, for instance. Finally, few databases have tried automatic relationships extraction methods: protein phosphorylation information has been extracted using rule-based pattern templates (32), recreation of events has been carried out for the Human Protein Interaction Database (HHPID) database (33) and revalidation of relationships for the PharmGKB database (34).

We present the first description of the curation pipeline for the CellFinder database (http://www.cellfinder.org/), a repository of cell research, which aims to integrate data derived from many sources, such as literature curation and microarray data. It is based on a novel ontology [Cell: Expression, Localization, Development, Anatomy (CELDA) (http://cellfinder.org/about/ontology)], which allows standardization and integration to other available ontologies on the cell and anatomy domains. Hence, the CellFinder platform provides a framework for comprehensive descriptions of human tissues, cells and commonly used model organisms on molecular and functional levels, *in vivo* and *in vitro*.

The CellFinder pipeline for literature curation integrates state-of-art freely available tools for the document triage, recognition of a variety of entity types and extraction of biological processes. Curation is carried out for full text documents available at the PubMed Central Open Access (PMC OA) subset (http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/), and manual intervention from curators is currently only necessary for querying new documents for curation and validation of the derived biological processes. In both cases, web-based tools are being used, which allow their integration into the CellFinder web site. We are not aware of prior usage of available systems for the automatic extraction of biological events. For instance, Xenbase manually annotates gene expression events (31), whereas others databases use proprietary systems (34) or tools, which do not allow re-use for other domains (33).

Our literature curation pipeline has been evaluated using a dataset on the kidney cell research. The kidney consists of >26 cell types, which arise and organize into several anatomical structures during a conserved developmental process (35). Kidney disease culminates from a common sclerotic pathway involving epithelial-mesenchymal transition, extracellular matrix remodeling and vascular changes (36). Multiple renal and non-renal (e.g. inflammatory) cell types are involved in these processes, with dynamic gene expression patterns and functions (37). Therefore, to identify relevant research describing cells and their interactions in normal and diseased kidney, we decided to include species-independent experimental and clinical data of renal disease and of kidney development in CellFinder. For the kidney cell use case, information is compiled about characterization of gene expression profiles in cells and other anatomical locations, such as tissues and organs. Hence, named-entity extraction is performed for genes, proteins, cell lines, cell types, tissues and organs. Gene expression events are then extracted between a gene/protein and a certain cell or anatomical part. The sentence below illustrates one such example (PMID 18989465):

> *On the other hand, the podoplanin expression occurs in the differentiating odontoblasts and the expression is sustained in differentiated odontoblasts, indicating that odontoblasts have the strong ability to express podoplanin.*

We are aware of only two previous publications, which report extraction of gene expression in anatomical locations from biomedical texts. OpenDMAP (38) uses Protégé and UIMA-based components, and it has been evaluated for three applications: protein transport, protein interactions and cell type-specific gene expression. OpenDMAP extract genes/proteins and cells using A Biomedical Named

Entity Recognizer (ABNER) (39) and a short list of trigger words. Relationships between the triple gene-cell-trigger are identified based on manual pattern templates. It reports precision of 64% and recall of 16% from an evaluation of 324 NCBI's GeneRIFs, which consists of short descriptions of gene functions.

A more comprehensive study on the expression of genes in anatomical location was carried out in (40) with the Gene Expression Text Miner system. The work included extending 150 abstracts from the BioNLP corpus (41) with annotations for anatomical parts and cell lines, as well as relationships to the existing gene expression events. Genes/proteins were extracted using GNAT (42), anatomical part and cell line recognition was performed by Linnaeus (43) using 13 anatomical ontologies and one for cell lines. A list of expression triggers was manually built, and association between the entities is also rule-based. Evaluation on the extended 150 abstracts resulted in a precision of almost 60% and a recall of 24%.

The next section will describe the CellFinder curation pipeline and the methods that are used in each stage. Results for the experiments performed for most of the steps are shown in the section 'Results' followed by discussion on the more important aspects of the pipeline in the section 'Discussion and future work'.

Methods and materials

The curation pipeline for the CellFinder database includes the following steps (cf. Figure 1): triage of potential relevant documents, retrieval of full text, linguistic preprocessing, named-entity recognition, post-processing, relationship extraction, manual validation of the results and integration of gene expression events into the database. This section describes details on the methods used in each phase.

Triage

Document triage is usually the first step in any literature curation workflow and consists of retrieving potential relevant publications for manual curation or for further processing by a text mining pipeline. In the CellFinder project,

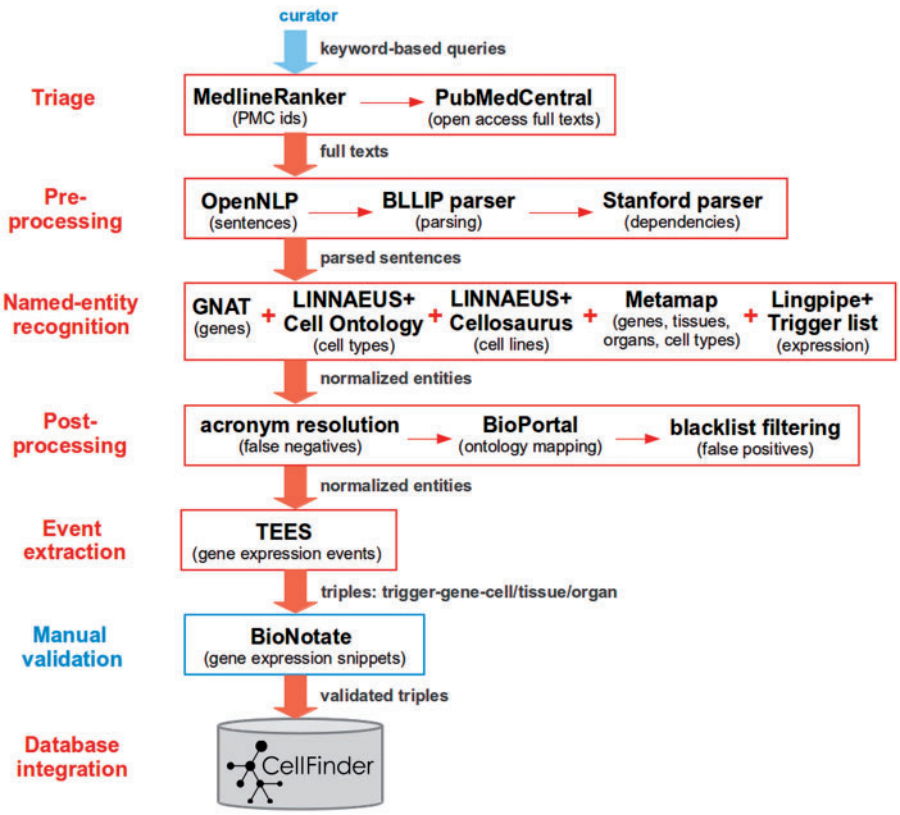

Figure 1. Overview of the literature curation pipeline for the CellFinder database. It includes the following steps: triage of potential relevant documents, retrieval of full text, preprocessing (sentence splitting, tokenization and parsing), named-entity recognition (genes, proteins, cell lines, cell types, organs, tissues, expression triggers), gene expression events extraction, manual validation of the results and integration into the database. Automatic procedures are shown in red, whereas the manual ones are shown in blue.

Database, Vol. 2013, Article ID bat020, doi:10.1093/database/bat020

we aim to curate only full texts documents, which are available for text mining purposes, i.e. the ones included in the PMC OA subset. Although it is a much smaller collection than the whole Medline, this subset currently contains >200 000 documents.

In our pipeline, document triage was performed by querying MedlineRanker (44), a machine learning based text categorization system. We have performed eight queries to MedlineRanker as follows: 'kidney tubular epithelial EMT', 'kidney vascular endothelial interstitium', 'kidney glomerular basement membrane', 'kidney mesangial space podocyte', 'kidney development differentiation pronephros', 'kidney extra cellular matrix, 'kidney regeneration mesenchymal precursor' and 'corticomedullary junction'. The search terms were aimed to identify cells, genes and structures that relate to cells contained in nephrons and tubules, such as epithelial cells, endothelial cells and podocytes, as well as cell changes associated with mesenchymal–epithelial transition (EMT) and fibrosis, changes in extracellular matrix and relevant proteins and in cells during kidney development, such as mesenchymal precursor cells.

Each query retrieved a list of 10 000 (MedlineRanker's cut-off) potential PMC relevant documents, including many repeated documents found across lists. After a post-processing step, which included verification on whether documents were part of the PMC OA subset and exclusion of repeated entries, a list of 2376 documents was derived. Documents were retrieved from PMC and were processed through our text mining pipeline.

Pre-processing

Full texts documents were first split by sentences using the OpenNLP toolkit (http://opennlp.apache.org/) and then parsed by the Brown Laboratory for Linguistic Information Processing (BLLIP) parser (https://github.com/dmcc/bllip-parserV) (45) (also known as McClosky-Charniak parser). Part-of-speech tags, tokenization and full parsing were derived from the BLLIP parser output. Dependency trees were built using the Stanford parser (http://nlp.stanford.edu/software/lex-parser.shtml). Part-of-speech, tokenization and parsing information are only necessary for the gene expression extraction (cf. 'Event Extraction' below).

Named-entity recognition

Named-entity recognition has been performed for five entity types: genes/proteins, cell lines, cell types, anatomical parts and gene expression triggers. Extraction is based on available state-of-art systems and dictionary or ontology-based approaches, without any adaption nor retraining. Methods are similar to the ones investigated in previous experiments performed with the CellFinder corpus (46). To enable data integration into the CellFinder database, all extracted mentions must be normalized to any of the ontologies or terminologies currently supported by our database: Cell Ontology (CL) (47), Cell Line Ontology (CLO) (48), EHDAA2 (49), Experimental Factor Ontology (EFO) (50), Foundational Model of Anatomy (FMA) (51), GO (52), Adult Mouse Anatomy (MA) (53) and Uberon (54).

We identify genes using GNAT (42), a system for extraction and normalization of gene and protein mentions. GNAT assigns confidence scores (up to 1.0) to the gene/protein candidates. Based on previous experiments (46), we have decided for a threshold score of 0.25 for filtering out potentially wrong gene/protein predictions. GNAT provides identifiers for all gene mentions with respect to the EntrezGene database (55).

Cell lines are recognized based on the version 4.0 of Cellosaurus (ftp://ftp.nextprot.org/pub/current_release/controlled_vocabularies/ cellosaurus.txt), a manually curated vocabulary of cell lines provided by the Swiss Institute of Bioinformatics. Synonyms from Cellosaurus were automatically expanded according to space and hyphens, such as 'BSF-1', 'BSF 1' and 'BSF1', resulting in a list of >41 000 synonyms for 15 245 registered cell lines. Matching of the derived list of synonyms and the full texts is performed by Linnaeus (43).

For the recognition of cell types and anatomical parts, we use Metamap (56), a system for Unified Medical Language System (UMLS) concept extraction. We configured Metamap to generate acronym variants and restricted results by the following semantic types: 'Cell' for cell types and 'Anatomical Structure', 'Body Location or Region', 'Body Part, Organ or Organ Component', 'Body Space or Junction', 'Body Substance', 'Body System', 'Embryonic Structure', 'Fully Formed Anatomical Structure' and 'Tissue' for anatomical parts. Metamap uses natural language processing techniques for breaking the text into phrases and further match them to UMLS concepts. From the potential matches returned by Metamap, we record not only the ones with highest score but also those that have the longest matching with the respective phrase.

Cell types have also been extracted using an ontology-based approach in which synonyms from the CL are matched against the full texts. It consists on a list of 2786 cell types from 1491 terms and matching is again performed by Linnaeus (43). Finally, triggers are extracted based on a list of 509 expression triggers, which was built manually. Terms from the list are matched against the full text using Lingpipe (http://alias-i.com/lingpipe/).

Post-processing

Acronym resolution. Metamap includes a step for acronym resolution, which returns a list of the pairs of abbreviations and long forms found as equivalent. However, Metamap sometimes recognizes the plural of some abbreviations but not the singular form or it does not return some abbreviations as a mention, but only the long

forms. For instance, for cell types, Metamap recognizes 'hESCs' as an acronym for 'human embryonic stem cells', but not its singular form 'hESC'. Further, although it lists the pair 'hESCs' and 'human embryonic stem cells' as being equivalent, only the long form is returned as a mention. Based on the list of pairs of abbreviations and long forms returned by Metamap, we try to match missed abbreviations and singular forms using Lingpipe.

Ontology mapping. Metamap returns annotations with regard to Concept Unique Identifier (CUI) terms, the original UMLS identifiers. Whenever available, we map them to FMA and GO terms using mappings available at the UMLS database. CUI terms are also mapped to other ontologies and terminologies supported by UMLS, but not by CellFinder, such as the CRISP Thesaurus (http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CSP/).To increase the recall of anatomical terms, we mapped UMLS CUI terms to CRISP terms [using mappings available at BioPortal (57)], and then further to other ontologies supported by CellFinder (e.g. CL, CLO, EHDAA2, MA, Uberon). Annotations returned by Metamap, which could not be automatically mapped to any supported ontology, are not removed, as identifiers could still be provided manually before integration of the data into the CellFinder database (not yet supported in the current curation workflow).

Blacklist filtering. Blacklists of manually curated mentions and identifiers are used for filtering out potential false predictions for all four entity types. This list was manually built based on the analysis of wrongly extracted annotations from the two corpora used for evaluation (cf. section 'Results'). The list of mentions contains only one entry for cell line ('FL'), 39 for anatomical parts (e.g. 'organism', 'tissue' and 'analysis'), 31 entries for cell types (e.g. 'cell' and 'stem cell') and 79 entries for genes/proteins (e.g. 'anti', 'repair', 'or in'). The list of identifiers include those which refer to broad concepts such as 'cell' (FMA:68646) or 'tissue' (FMA:9637). We filter out extracted mentions associated to any of the identifiers in this list.

Event extraction

Results from sentence splitting, tokenization, part-of-speech tagging, parsing, dependency tags and named entities are integrated into the so-called 'Interaction XML' file format (https://github.com/jbjorne/TEES/wiki/TEES-Overview) (58) used by the Turku Event Extraction System (TEES) (59). TEES is an event extraction system, which uses multiclass Support Vector Machine on a rich graph-based feature set for trigger, edge and negation detection. Despite recent improvement of relation extraction methods (10), TEES seems to be the only available system suitable to be re-trained with novel corpora from any domain without the need of performing changes in its source code.

We trained TEES in a gold-standard set of 20 full text annotated documents, 10 on human embryonic stem cell research (hereafter called CF-hESC), whose entities annotations have been previously published (46) and a new set of 10 full texts documents on kidney stem cell research (hereafter called CF-Kidney). Both corpora have been manually annotated with the five entity types (gene/proteins, cell lines, cell types, anatomical parts, expression triggers) and gene expression events (cf. example in Figure 2). These events are composed of a trigger, which is always linked to two arguments, a gene/protein (hereafter called 'Gene' argument) and a cell line, cell type or anatomical part (hereafter called 'Cell' argument). We split both corpora into three parts (training, development and test) and perform experiments using one corpus or a combination of both for training. Details on the corpora are shown in Table 1.

TEES receives the Interaction XML file as input and returns a new XML file, which includes predictions for the 'Cell' and 'Gene' relationships. The later are subsequently combined to compose complete gene expression events by

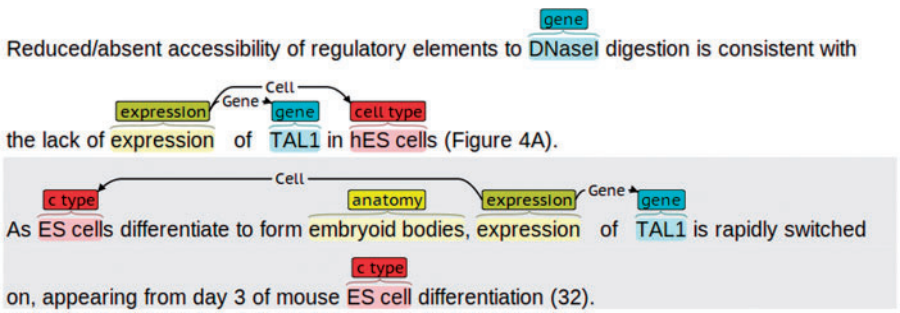

Figure 2. Examples of gene expression events for the kidney stem cell corpus (PMID 17389645, PMCID PMC1885650). Each expression trigger (dark yellow) is always related with only one gene/protein (in blue) and only one cell (in yellow) or anatomical part (in red). However, the corpus was also annotated with entities, which do not take part in any event. Visualization of the corpus was provided by Brat annotation tool (60).

Database, Vol. 2013, Article ID bat020, doi:10.1093/database/bat020

Table 1. Statistics on the corpora

Features	CF-hESC			CF-Kidney		
	Training	Development	Test	Training	Development	Test
Documents	6	2	2	6	2	2
Sentences	1379	259	539	1578	618	383
Sentences with entities	944	163	302	1344	527	314
Sentences with events	147	26	40	240	210	122
Entities	4158	583	1260	4834	3443	1748
Genes/proteins	1264	163	355	1440	1338	782
Cell lines	198	72	141	11	8	1
Cell types	1556	179	524	917	259	72
Anatomical parts	921	137	173	2116	1380	617
Expression triggers	219	32	67	350	458	276
Relationships	944	160	390	1144	1404	1320
Expression-Gene/protein	472	84	195	572	702	660
Expression-CellLine	13	6	36	14	5	
Expression-CellType	435	56	122	411	398	86
Expression-anatomy	24	18	37	147	299	574

Information is shown for the training, development and test data sets of the CF-hESC and CF-Kidney data sets. It includes number of documents, sentences, sentences with entities and sentences with events. Number of annotations is presented by entity type, and the number of events also shown according to the entities participating in the relationships.

checking the presence of both a 'Gene' and a 'Cell' relationship linked to the same trigger. TEES relationships are restricted to entities present in the same sentence; therefore, the same restriction is valid for all derived events.

Manual validation

We applied TEES-trained models on the kidney cell data set of 2376 full texts. Results were manually validated using Bionotate (61), a collaborative open-source text annotation tool. Bionotate presents a snippet of text along with annotated entities, a question, and a list of possible answers. Curators were instructed to give one answer per snippet, and although Bionotate allows changing the span of the named entities, for this experiment, curators were asked only to answer the question. Bionotate selects snippets randomly among all those included in its repository. A snippet is no longer presented to the user when a certain number of agreements (equal answers) have been reached. For this experiment, one answer from any of our expert curators suffices.

We have converted the output from TEES event extractor system to the XML format of the Bionotate. Snippets are composed of the sentence in which the event occurs and the two previous and subsequent sentences, for a better understanding of the context (cf. Figure 3). Additionally, a link to the respective PubMed entry is provided, in case those curators needed to check the abstract or full text of the publication before answering the question. The questions assessed whether there was a gene expression event taking place in the snippet, including its negation, whether the named entities were correctly recognized or if the publication was relevant for the kidney cell research. This resulted in the following possible answers: [1] Yes, an event is taking place and all entities are correct. [2] Yes, but the text says the gene expression is NOT taking place. [3] No, no event is taking place although all entities are correct. [4] No, this is not a gene expression trigger. [5] No, this is not a gene. [6] No, this is not a cell or anatomical part. [7] No, both gene and cell or anatomical part are incorrect. [8] No, the snippet (publication) does not seem to be relevant for CellFinder.

Results

In this section, we describe the evaluation performed for the methods used in the various stages of the text mining pipeline. We also present an overview of the data, which have been extracted by our curators with the help of the pipeline. The triage phase has not been directly evaluated, except for the answer number 8 during the manual validation of results (cf. 'Manual validation' in this section).

Evaluation of the named-entity recognition and event extraction will be shown in terms of precision (P), recall (R) and f-score (F). Precision represents the ratio of the correct predictions of a particular system among all the returned ones. On the other hand, recall corresponds to

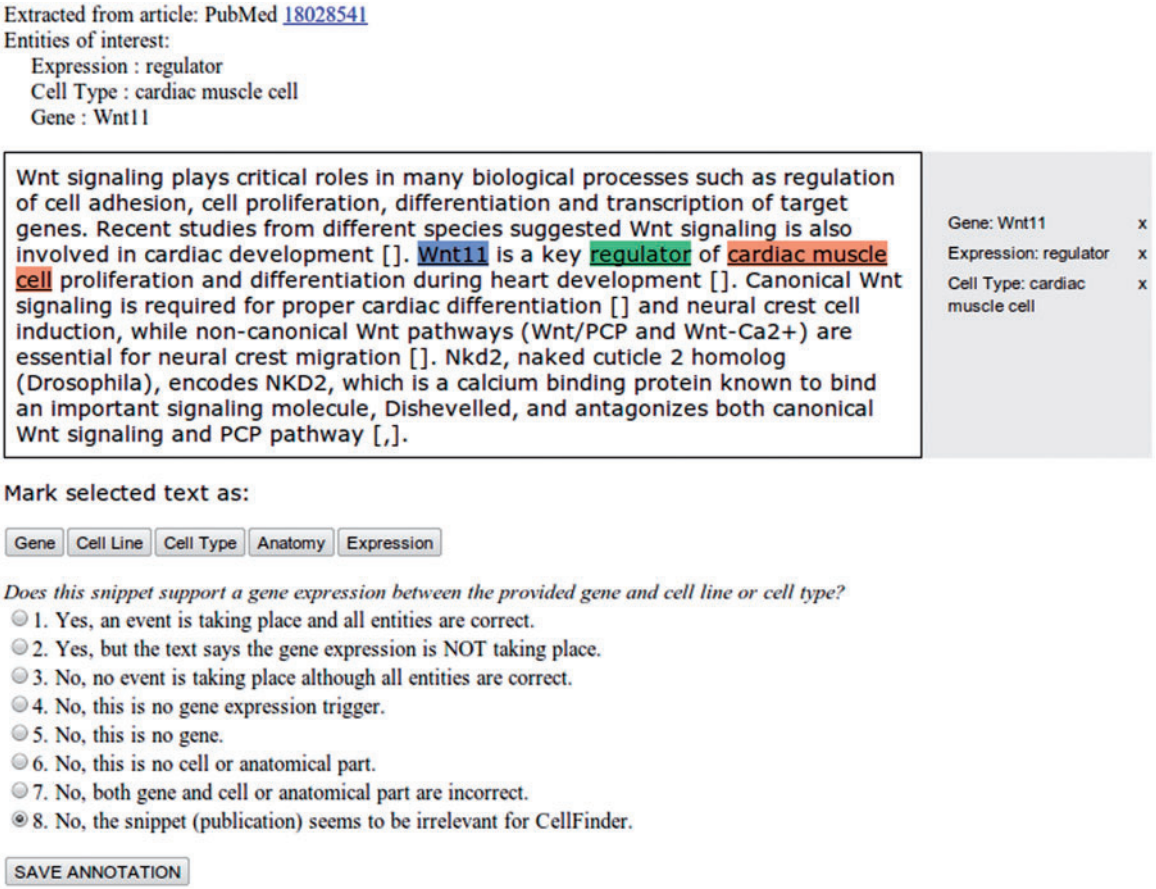

Figure 3. Screen-shot of Bionotate configured for the validation of the gene expression events. Three named-entities are always pre-annotated: a trigger (in green), a gene (in blue) and a cell line, cell type or anatomical part (in red). The answers assess whether the biological event is taking place, its negation, the accuracy of the named-entity recognition and the relevancy of the publication from where the snippet was derived.

the ratio of gold-standard annotations, which were actually returned by the system. Finally, the f-score is a harmonic average of both measures and shows the overall performance of a system.

Pre-processing

During the pre-processing step, sentence splitting in all 2376 full text documents resulted in a total of 581 350 sentences. Parsing and dependency tags conversion was successfully for 578 572 of them. The parsing information is only used by the TEES system (cf. 'Event extraction' in section 'Methods and materials'), which means that although named-entity recognition was carried out in all sentences, only those correctly parsed ones were analyzed by TEES.

Named-entity recognition

Named-entity extraction was evaluated on the development and test gold-standard documents belonging to the human embryonic and kidney stem cell research (cf. Table 1), but only the development data sets were used for further improvements of methods, such as trigger list or blacklist construction and error analysis (cf. section 'Discussion and future work'). Table 2 shows the evaluation of each entity type for both corpora. The 'Exact' evaluation assesses annotations, which matched regarding span and entity type, whereas 'Overlap+Type' allowed overlapping spans for annotations of the same type and 'Overlap' let annotations to have different types. The latter is particularly helpful regarding overlapping annotations between cell lines, cell types and anatomical parts, as any of these entity types corresponds to the same argument 'Cell' in the gene expression event (cf. Figure 2).

Recall is particularly low for genes/proteins in the development data set of the CF-Kidney corpus owing to a high number of annotations from a few genes/proteins, which have been missed by the system: 'Gata3' (155), 'Ret' (97) and 'EpCAM' (83). Some of these were found by GNAT but with a recall lower than the threshold we have considered. Cell lines are very rare in the CF-Kidney corpus, and the eight

Database, Vol. 2013, Article ID bat020, doi:10.1093/database/bat020

Table 2. Evaluation of the automatic named-entity recognition on the CF- hESC and CF-Kidney corpora

Corpora		Match	Entity types (recall/F-score)				
			Genes	C. lines	C. types	Anatomy	Expression
CF-hESC	Development	Ex.	0.61/0.54	0.68/0.61	0.14/0.15	0.34/0.34	0.72/0.15
		OT	0.75/0.65	0.94/0.85	0.62/0.66	0.48/0.45	0.91/0.19
		Ov.	0.82/0.69	0.94/0.81	0.70/0.73	0.72/0.62	0.97/0.20
	Test	Ex.	0.68/0.65	0.40/0.49	0.25/0.28	0.30/0.25	0.45/0.08
		OT	0.76/0.72	0.58/0.65	0.58/0.65	0.43/0.35	0.54/0.09
		Ov.	0.77/0.71	0.61/0.69	0.77/0.82	0.81/0.71	0.55/0.10
CF-Kidney	Development	Ex.	0.34/0.45	1.00/0.33	0.17/0.26	0.69/0.75	0.68/0.43
		OT	0.35/0.46	1.00/0.33	0.18/0.27	0.88/0.87	0.69/0.43
		Ov.	0.46/0.56	1.00/0.34	0.77/0.80	0.90/0.89	0.76/0.47
	Test	Ex.	0.69/0.76	1.00/0.33	0.89/0.86	0.67/0.74	0.80/0.42
		OT	0.70/0.77	1.00/0.33	0.93/0.89	0.69/0.76	0.80/0.42
		Ov.	0.70/0.77	1.00/0.33	0.94/0.91	0.72/0.77	0.81/0.42

Results are shown for the development and test data sets in the format recall/F-score. Matching is evaluated regarding same span and entity type (Ex.), overlapping span and same type (OT) and overlapping span of any entity type (Ov.).

identical cell lines of the development data set and the only one of the test data set were correctly extracted (thus recall 1.0). Finally, recall is also particularly low for cell types in the development data set, even when allowing overlaps. Indeed, there is a great variety of cell types (>100), which could not be recognized, especially cell types, which in fact represent gene expressions events, such as 'NCAM + NTRK2 + cells' or 'Gata3−/Ret− cells'.

The ontology mapping post-processing step could automatically map a total of 171 (CF-hESC corpus) and 121 (CF-Kidney corpus) additional annotations to an identifier from any of the ontologies supported in CellFinder. They had been previously extracted by Metamap, but they were associated only to the UMLS CUI identifier. However, 1342 (33%) and 961 (16%) of the extracted annotations, respectively, remain assigned only to the UMLS CUI identifier, with respect to the total number of cell types and anatomical parts.

The acronym resolution procedure has resulted in a slight increase in recall for cell types and anatomy, without loss of f-score (result not shown). For instance, recall for cell types in the CF-hESC corpus increased from 64 to 70% (result not shown) owing to the extraction of acronyms such as 'MEF' (mouse embryonic fibroblast) or 'EB' (embryoid body), which have not been previously returned by Metamap.

Finally, blacklist filtering of terms also allowed a modest improvement of precision for both corpora (result not shown). For instance, precision for genes/proteins in the CF-hESC corpus increased from 43 to 50% (result not shown) owing to filtering out annotations such as 'or in' or 'membrane', which had been recognized by GNAT and genes or proteins.

The named-entity extraction methods were run on the 2376 full texts and resulted in a total of >2 200 000 mentions for all five entity types. Details on the extracted annotations are presented in Table 3, such as the number of mentions for each entity type, distinct text spans and distinct identifiers.

Table 3. Statistics on the extracted named entities

Annotations	Genes	C. lines	C. types	Anatomy	Expression
Distinct mentions	702 829	81 074	183 820	565 860	681 370
Distinct spans	34 222	1825	9142	14 874	892
Distinct ids	34 353	11 875	1150	4300	

For each entity type, the number of annotations, distinct spans and identifiers is shown. Sometimes more than one identifier is assigned to a mention, therefore their high number. Trigger words (Expression) are not normalized to any ontology.

Event extraction

To extract gene expression events, we investigated training TEES on three models: CF-hESC corpus (6 full text documents), CF-Kidney corpus (6 full text documents) and a mix of both (12 full text documents) (hereafter called CF-Both). Input to TEES should include three data sets: training, development and test. During the training step, TEES automatically configures its parameters using the development data set and presents an evaluation of its own for the test set. Details on the performance of the relationship extraction is shown in Table 4 for the three training models, as well as for the complete events further performed by the authors. This is the performance of TEES without the influence of the named-entity recognition predictions of our text mining pipeline, as only gold-standard documents are used during the training step. Recall of the relationships range from 60 to 70% while precision is also

Table 4. Evaluation of TEES during training

Data sets	Relationship	Development			Test		
		P	R	F	P	R	F
CF-hESC	Cell	0.86	0.56	0.68	0.77	0.45	0.57
	Gene	0.91	0.68	0.78	0.82	0.90	0.86
	Event	0.60	0.35	0.44	0.38	0.53	0.44
CF-Kidney	Cell	0.71	0.50	0.59	0.62	0.68	0.65
	Gene	0.60	0.82	0.69	0.73	0.75	0.74
	Event	0.17	0.49	0.25	0.12	0.56	0.20
CF-Both	Cell	0.77	0.55	0.65	0.69	0.64	0.67
	Gene	0.67	0.81	0.73	0.69	0.84	0.76
	Event	0.55	0.48	0.51	0.50	0.56	0.53

Evaluation is shown for the 'Cell' and 'Gene' relationships and for the development and test data sets, as described in Table 1. The complete events derived from a 'Cell' and a 'Gene' argument associated to the same trigger are also shown. For each training run, evaluation is carried out on the corresponding development and test data sets, i.e. two documents for each single corpus (CF-hESC and CF-Kidney) and four documents when training on the joined corpus (CF-Both). Predictions were performed over the gold-standard named-entity annotations. 'P' refers to 'Precision', 'R' to 'Recall' and 'F' to 'F-score'.

good, from 60 to almost 90%. Both the recall and precision drop when considering the complete events, and recall is not always as high as the argument with the lower recall. This is due to the fact that TEES predicts the 'Cell' and 'Gene' relationships independently, and many of them are not associated to the same trigger.

In Table 5, we show the performance of TEES relationship extraction when using the predictions obtained in the named-entity recognition step, as well as gene expression events derived from the binary relationships. This is the final performance of our text mining pipeline for the extraction of gene expression events on cell and anatomical locations. Additionally, we include the performance for the prediction of the triplets gene-cell-trigger, which represent every possible combination of annotations from these three arguments in the same sentence. Therefore, it represents the higher possible recall for the event extraction provided the predicted named entities.

Results are shown using the approximate span matching, i.e. for each argument, overlapping matches are allowed, but entities should have the same type as well as equality of the argument type ('Cell' or 'Gene'). For the development data set and when using the CF-Kidney corpus for training TEES, whether alone or together with the CF-hESC corpus, no complete event was extracted. This is due to two reasons: (i) the low recall of genes/proteins and cell types for the CF-Kidney corpus (cf. Table 2, evaluation OT) and (ii) the inability of the CF-Kidney model to extract events from documents from other domains, i.e with different cell type nomenclature. Indeed, no gene expression events have been extracted from the two development documents of the CF-hESC corpus included in the development data set of the CF-Both corpus. This probably due to the high complexity and variability of the cell types in the CF-Kidney corpus, with examples such as 'NCAM− cell' or 'EpCAM−NCAM−NTRK2+ cells'.

We have run TEES using the three models (CF-hESC, CF-Kidney and CF- Both) on the 2376 documents and the named-entities previously extracted (cf. Table 3). We have obtained only 115 and 178 gene expression events for the CF-Kidney and CF-Both models, respectively, whereas the CF-hESC model retrieved 4280 events. The latter were derived from almost 127 000 binary relationships, i.e. the complete events correspond to only 14% of the original extracted relationships. The last column of Table 5 summarizes the number of relationships and derived events, which have been obtained using each training model.

Manual validation

The gene expression events obtained with the three models were converted to the Bionotate XML format, and snippets were loaded into its repository. Curators have manually validated 2741 snippets, which contained events predicted by the three distinct models. Results are summarized in Table 6. The validated data, one file per snippet in the Bionotate's XML format, is available for download at the CellFinder web site (http://cellfinder.org/about/annotation/).

Validation for the events extracted using the CF-hESC model, the best performing one according to the evaluation and the number of predictions, can be summarized as follows. About 51% (answers 1 and 2) of the gene expression events have been extracted correctly, as well as the participating entities. This includes both positive and negative statements of gene expression in cell in anatomical parts. Exactly 17% (answers 3 and 4) of the snippets described processes not related to gene expression, although the gene, cell or anatomy were correctly recognized. Almost 25% (answers 5, 6 and 7) of the extracted events contained a wrong identified gene/protein, cell/

Database, Vol. 2013, Article ID bat020, doi:10.1093/database/bat020

Table 5. Evaluation of gene expression extraction

Data sets	Relationship/Event	Development			Test			Predictions
		P	R	F	P	R	F	
CF-hESC	Cell	0.43	0.06	0.10	0.76	0.33	0.46	14 551
	Gene	0.35	0.22	0.27	0.76	0.79	0.77	112 372
	Events	0.50	0.08	0.14	0.27	0.05	0.08	4280
	Triplets	0.06	0.51	0.10	0.05	0.35	0.09	
CF-Kidney	Cell	0.44	0.02	0.05	0.52	0.57	0.55	109 934
	Gene	0.62	0.06	0.10	0.77	0.69	0.73	5520
	Event							115
	Triplets	0.02	0.19	0.04	0.02	0.28	0.05	
CF-Both	Cell	1.0	0.01	0.02	0.70	0.64	0.67	69 079
	Gene	0.33	0.01	0.01	0.69	0.84	0.76	3792
	Event							178
	Triplets	0.02	0.22	0.04	0.03	0.30	0.05	

We have trained the TEES system on three data sets: CF-hESC, CF-Kidney and CF-Both. Results for the 'Cell' and 'Gene' relationships were provided by TEES during processing of the documents. Performance for complete events is evaluated allowing overlapping matches for entity spans, but with equality of entity types and argument types. The triplets correspond to every possible combination of the triggers, genes/proteins, cells or anatomical parts in the same sentence, i.e. the highest possible recall for any relationship extraction system provided the predictions for the entities. The 'Pred.' column presents the number of relationships or complete events, which have been extracted from the 2376 full texts on kidney research when using each of the training models. 'P' refers to 'Precision', 'R' to 'Recall' and 'F' to 'F-score'.

Table 6. Evaluation of the gene expression snippets in Bionotate

Answers	CF-hESC		CF-Kidney		CF-Both		Total	
	No. snippets	%	No. snippets	%	No. snippets	%	No. snippets	%
1. Yes	1204	49.1	34	29.5	6	3.3	1244	45.4
2. Yes (negation)	47	1.9	3	2.6	0	0	50	1.8
3. No (but entities correct)	218	9.0	8	7.0	1	0.6	227	8.3
4. No (trigger wrong)	194	8.0	28	24.3	78	43.8	300	11.0
5. No (gene wrong)	346	14.1	11	9.6	6	3.4	363	13.2
6. No (cell/anatomy wrong)	207	8.5	26	22.6	9	5.1	242	8.8
7. No (gene/cell/anatomy wrong)	55	2.2	4	3.5	1	0.6	60	2.2
8. No (irrelevant document)	177	7.2	1	0.9	77	43.2	255	9.3
Total	2448	100	115	100	178	100	2741	100

A total of 2741 snippets (gene expression events) were validated. These events were predicted by the three models used for training TEES event extraction system. Percentages for each answer are also shown.

anatomy or both of them, which means that precision was higher than the average for the named-entity recognition (cf. Table 2). Finally, 7.2% of the snippets turned out to belong to documents, which are irrelevant to the kidney cell domain, which gives a hint on the performance of the triage step.

Discussion and future work

We have described our preliminary text mining pipeline for the extraction of five entity types and gene expression events. In this section, we discuss the most important results derived from this first experiment with our text mining curation pipeline.

Named-entity recognition

In the named-entity recognition step, we have considered only state-of-art and freely available tools, and we did not train specific systems with the gold-standard corpora discussed here. Results for entity extraction are in-line with previous published ones (46), although data sets are

different and, therefore, results are not directly comparable. A high recall is preferable over a high precision, as events cannot be predicted if the participating entities have not been previously extracted. On the other hand, a high number of wrong predictions slow down the validation process, and therefore, a balance between precision and recall (given by the f-score) is also desirable. Provided the still low recall for some entities, and the consequent low recall of the event extraction, future work should still focus on the improvement of the named-entity prediction.

Regarding genes/proteins extraction, most of the missing annotations could have been recognized by GNAT if we had used a lower threshold. Other tools could also be combined with GNAT, such as GeneTUKit (62) or BANNER (63). Additionally, use of domain-specific post-processing, such as 'whitelists' of genes/proteins, would certainly help, and future work will concentrate on these two approaches. Recall for genes/proteins increases considerably for both development data sets when allowing overlaps and an improvement is also perceived when type equality is relieved, which shows that some genes overlap with some cells names or anatomical parts, such as 'C34' (a gene) and 'C34 cell' (a cell type).

Cell lines are not as common as cell types in our corpora, specially in the CF-Kidney corpus where this entity type is almost non-existent (cf Table 1). However, it plays an important role in the cell research, and scientific literature reports many gene expression events, which take place in cell cultures. Restricting our evaluation to the CF-hESC corpus, recall varies from 60 to >90% when allowing overlapping spans (cf. Table 2), but it is still not satisfactory, and dictionary-based methods might not be sufficient. Missing annotations for cell lines are mostly due to the absence of the synonym in any of the available thesaurus or ontologies, such as 'SD56', which is not included in Cellosaurus. Thus, future work will include training a machine learning system for cell line recognition, including annotation of additional gold-standard documents.

Improvement of the event extraction starts with the improvement of the recall for the named entities. Performance of cell types and anatomical parts are rather variable. A good recall is usually obtained when releasing equality of types, and further experiments should consider unifying the cell types and anatomical parts in our corpora. If fact, previous studies on the CF-hESC corpus show that inter-annotator agreement for these entity types was low (46). Overlaps between cell types and anatomical parts should not be a problem for the gene expression event extraction, as both entity types takes part in the 'Cell' argument.

Cell types were sometimes poorly recognized for the CF-Kidney data set, owing to the high variability of the nomenclature and the presence of gene expression in its contents, such as 'NCAM+NTRK2+ cells' or 'Gata3−/Ret− cells'. Thus, improvements on cell type extraction should also focus on training machine learning algorithms. Mapping cell types with such a pattern to an identifier is also a challenge, as these terms are not included in any available ontology. The prior identification of the original cell type in which the gene is being expressed can help in the normalization of these cells, an information that is usually present in the text, although not always in the same sentence.

Expression triggers are extracted based on a manually curated list, which assures a high recall. Low recall, such as the ones for the development data set of the CF-Kidney corpus are due to unusual trigger words, such as '-' (negative expression), 'dim' and 'bright'.

Event extraction

We obtained the gene expression events using the TEES edge detection module, which extracted relationships between expression triggers and a gene/protein, cell or anatomy. TEES allows training the system with novel corpora, and during the training step, examples are generated for all combinations of entities provided in the training corpus. Therefore, a few relationships returned by TEES are related to the wrong entity type. For instance, it extracts some 'Gene' arguments associated to cells or anatomical parts and some 'Cell' arguments related to genes, although no such examples can be found in any of our gold-standard corpora. TEES extracts the relationships independently. Therefore, the recall of the binary relationships does not correspond to the recall of the complete gene expression event. Future work on event extraction will also include trying additional event extraction systems, such as (64, 65).

Use of more annotated documents might also improve the event extraction. Further experiments can also be performed using available corpora, such as the set of annotated abstracts of the Gene Expression Text Miner corpus (40). Additionally, a careful analysis of the wrongly extracted events returned by TEES when using gold-standard annotations (cf. low precision for CF-Kidney corpus in Table 4) could reveal inconsistencies in the manual annotations in our corpora. To avoid huge differences between development and test results, a cross-validation could have been investigated. In summary, a cross-validation in a larger and more robust corpus could provide more stable results.

Nevertheless, these preliminaries results on extraction of gene expression in cells and anatomical parts are certainly interesting for the many groups working on event extraction, as this is one of the first curation experiment to use a event extraction system, which had not been developed by the authors. Additionally, it is probably the first external evaluation of TEES on a new corpus, one of the very few event extraction systems available to the public. Finally, the use of corpora from two distinct cell research domains

Database, Vol. 2013, Article ID bat020, doi:10.1093/database/bat020

shows how large differences in results are dependent on the corpus and the corresponding learned model.

Processing of the data set of 2376 full text documents for kidney cell research resulted in a high number of entities but apparently a low number of extracted events. However, recall is unknown, as well as the number of publications, which described expression of genes in cells and anatomical parts for the kidney cell research. The number of correct gene expression events is certainly low compared with the number of processed documents, but number of irrelevant publications in our collection is also unknown and could be higher than 6%, as reported by answer number 8 of the validation (cf. Table 6).

Next event extraction tasks will involve recognition of additional relationships, such as identifying the cell type or tissue from which a certain cell line was derived. Future work will also include additional biological processes, such as cell differentiation. These relationships have already been annotated in the two gold-standard corpora discussed here and involve the same entities whose recognition is already included in our pipeline.

Manual validation

Manual validation of 2741 snippets reported that half of them contained correctly recognized entities and gene expression events, which is in line with the precision of TEES shown in Table 5. Curators reported that most mistakes concentrated on incomplete extraction of genes/proteins and cell types, such as the recognition of 'TGF' instead of 'TGF-beta'. Feedback from the validation will help to improve both recall and precision for the named-entity recognition by adding more terms to the blacklists (potential wrong predictions) and by creating 'whitelists' (potential missing annotations).

Curators reported a positive first experience with Bionotate, although changes in visual interface, short-cuts and functional features have been suggested as future work. Next experiments will also focus on the validation of the identifiers, which were automatically assigned during the named-entity recognition, as well as allowing curators to change the span of the pre-annotated entities, a feature already supported by Bionotate. Validation of the normalized identifiers is an important step before final integration of the results into the CellFinder database. Version 2.0 of Bionotate (66) supports this functionality and will certainly be considered for integration in our pipeline.

Conclusions

We presented here our preliminary results for the text mining pipeline for curation of gene expression events in cells in anatomical parts for the CellFinder database. Our pipeline relies only on open-source or freely available tools, and evaluation for each stage has been carried out based on gold-standard corpora. We are not aware of previous database curation pipelines where text mining methods have been used in all of the following stages: triage, named-entity recognition and event extraction.

We performed named-entity extraction extraction for genes/proteins, cell lines, cell types, tissues, organs and gene expression triggers. Gene expression events were extracted using machine learning algorithms trained on manually annotated corpora from two domains, human embryonic stem cells and kidney cell research. Results for both the name-entity recognition and event extraction steps are promising, although improvements are still necessary to achieve a higher recall and precision.

The text mining pipeline has been used to process 2376 full texts documents on kidney cell research and resulted in a total of >60 000 distinct entities and >4500 gene expression events. Half of the events have been manually validated by experts, and about half of them were classified as describing a gene expression taking place in a cell or anatomical part.

Acknowledgements

The authors are thankful for the valuable support of Jari Björne (University of Turku) on the TEES system and Thomas Stoltmann (Humboldt-Universität zu Berlin) for technical support. They also thank Hanno Loring, Krithika Hariharan, Bella Roßbach and Isidora Paredes for curation of documents.

Funding

Deutsche Forschungsgemeinschaft (DFG) (LE 1428/3-1 and KU 851/3-1). Funding for open access charge: LE 1428/3-1.

Conflict of interest. None declared.

References

1. Hirschman,J., Berardini,T.Z., Drabkin,H.J. *et al*. (2010) A MOD(ern) perspective on literature curation. *Mol. Genet. Genomics*, **283**, 415–425.
2. Turinsky,A.L., Razick,S., Turner,B. *et al*. (2010) Literature curation of protein interactions: measuring agreement across major public databases. *Database*, article ID baq026; doi: 10.1093/database/baq026.
3. Wiegers,T., Davis,A., Cohen,K.B. *et al*. (2009) Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (ctd). *BMC Bioinformatics*, **10**, 326.
4. Hirschman,L., Burns,G.A., Krallinger,M. *et al*. (2012) Text mining for the biocuration workflow. *Database*, article ID bas020; doi: 10.1093/database/bas020.
5. Baumgartner,W.A., Cohen,K.B., Fox,L.M. *et al*. (2007) Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, **23**, i41–i48.

6. Kim,J.-D., Ohta,T., Tsuruoka,Y. *et al*. (2004) Introduction to the bio-entity recognition task at jnlpba. In: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, JNLPBA '04*. Association for Computational Linguistics, Stroudsburg, PA, pp. 70–75.
7. Smith,L., Tanabe,L., Ando,R. *et al*. (2008) Overview of BioCreative II gene mention recognition. *Genome Biol.*, **9**, S2.
8. Morgan,A.A., Lu,Z., Wang,X. *et al*. (2008) Overview of BioCreative II gene normalization. *Genome Biol.*, **9 (**Suppl. 2), S3.
9. Tikk,D., Thomas,P., Palaga,P. *et al*. (2010) A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS Comput. Biol.*, **6**, e1000837.
10. Kim,J.-D., Nguyen,N., Wang,Y. *et al*. (2012) The genia event and protein coreference tasks of the bionlp shared task 2011. *BMC Bioinformatics*, **13 (**Suppl. 11), S1.
11. Gerner,M., Sarafraz,F., Bergman,C.M. *et al*. (2012) Biocontext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*, **28**, 2154–2161.
12. Lourenco,A., Carreira,R., Carneiro,S. *et al*. (2009) @Note: a workbench for biomedical text mining. *J. Biomed. Inform.*, **42**, 710–720.
13. Rak,R., Rowley,A., Black,W. *et al*. (2012) Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database*, article ID bas010; doi: 10.1093/database/bas010.
14. Salgado,D., Krallinger,M., Depaule,M. *et al*. (2012) Myminer: a web application for computer-assisted biocuration and text annotation. *Bioinformatics.*, **28**, 2285–2287.
15. Müller,H.-M., Kenny,E. E. and Sternberg,P.W. (2004) Textpresso: an ontology- based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.
16. Neves,M. and Leser,U. (2012) A survey on annotation tools for the biomedical literature. *Brief. Bioinform*, 1–14.
17. Rebholz-Schuhmann,D., Kirsch,H. and Couto,F. (2005) Facts from text—is text mining ready to deliver? *PLoS Biol.*, **3**, e65.
18. Winnenburg,R., Wächter,T., Plake,C. *et al*. (2008) Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief. Bioinform.*, **9**, 466–478.
19. Alex,B., Grover,C. and Haddow,B. (2008) Assisted curation: does text mining really help. *Pac. Symp. Biocomput.*, **2008**, 556–567.
20. Lu,Z. and Hirschman,L. (2012) Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database*, article ID bas043; doi: 10.1093/database/bas043.
21. McQuilton,P. (2012) Opportunities for text mining in the flybase genetic literature curation workflow. *Database*, article ID bas039; doi: 10.1093/database/bas039.
22. Aerts,S., Haeussler,M., van Vooren,S. *et al*. (2008) Text-mining assisted regulatory annotation. *Genome Biol.*, **9**, R31.
23. Pillai,L., Chouvarine,P., Tudor,C.O. *et al*. (2012) Developing a biocuration workflow for AgBase, a non-model organism database. *Database*, article ID bas038; doi: 10.1093/database/bas038.
24. Donaldson,I., Martin,J., de Bruijn,B. *et al*. (2003) Prebind and textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11.
25. Wang,P., Morgan,A., Zhang,Q. *et al*. (2007) Automating document classification for the immune epitope database. *BMC Bioinformatics*, **8**, 269.
26. Kim,S., Kim,W., Wei,C.-H. *et al*. (2012) Prioritizing pubmed articles for the comparative toxicogenomic database utilizing semantic information. *Database*, article ID bas042; doi: 10.1093/database/bas042.
27. Wiegers,T.C., Davis,A.P. and Mattingly,C.J. (2012) Collaborative biocuration—text-mining development task for document prioritization for curation. *Database*, article ID bas037; doi: 10.1093/database/bas037.
28. Harris,M.A., Clark,J., Ireland,A. *et al*. (2004) The gene ontology (go) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
29. Van Auken,K., Fey,P., Berardini,T.Z. *et al*. (2012) Text mining in the biocuration workflow: applications for literature curation at wormbase, dictybase and tair. *Database*, article ID bas040. doi: 10.1093/database/bas040.
30. Dowell,K., McAndrews-Hill,M., Hill,D. *et al*. (2009) Integrating text mining into the mgi biocuration workflow. *Database*, article ID bap019.
31. Bowes,J.B., Snyder,K.A., Segerdell,E. *et al*. (2010) Xenbase: gene expression and improved integration. *Nucleic Acids Res.*, **38**, D607–D612.
32. Hu,Z.Z., Narayanaswamy,M., Ravikumar,K.E. *et al*. (2005) Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, **21**, 2759–2765.
33. Jamieson,D.G., Gerner,M., Sarafraz,F. *et al*. (2012) Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database. *Database*, article ID bas023; doi: 10.1093/database/bas023.
34. Rinaldi,F., Clematide,S., Garten,Y. *et al*. (2012) Using ODIN for a PharmGKB revalidation experiment. *Database*, article ID bas021; doi: 10.1093/database/bas021.
35. Raciti,D., Reggiani,L., Geffers,L. *et al*. (2008) Organization of the pronephric kidney revealed by large- scale gene expression mapping. *Genome Biol.*, **9**, R84.
36. Zeng,R., Han,M., Luo,Y. *et al*. (2011) Role of Sema4C in TGF-beta1-induced mitogen-activated protein kinase activation and epithelial–mesenchymal transition in renal tubular epithelial cells. *Nephrol. Dial. Transplant.*, **26**, 1149–1156.
37. Tarabra,E., Giunti,S., Barutta,F. *et al*. (2009) Effect of the monocyte chemoattractant protein-1/CC chemokine receptor 2 system on nephrin expression in streptozotocin- treated mice and human cultured podocytes. *Diabetes*, **58**, 2109–2118.
38. Hunter,L., Lu,Z., Firby,J. *et al*. (2008) OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*, **9**, 78.
39. Settles,B. (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, **21**, 3191–3192.
40. Gerner,M., Nenadic,G. and Bergman,C.M. (2010) An exploration of mining gene expression mentions and their anatomical locations from biomedical text. In: *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, BioNLP '10*. Association for Computational Linguistics, Stroudsburg, PA, pp. 72–80.
41. Ohta,T., Kim,J.-D., Pyysalo,S. *et al*. (2009) Incorporating genetag-style annotation to genia corpus. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, BioNLP '09*. Association for Computational Linguistics, Stroudsburg, PA, pp. 106–107.
42. Hakenberg,J., Plake,C., Leaman,R. *et al*. (2008) Inter-species normalization of gene mentions with gnat. *Bioinformatics*, **24**, i126–i132.

43. Gerner,M., Nenadic,G. and Bergman,C. (2010) Linnaeus: a species name identification system for biomedical literature. *BMC Bioinformatics*, **11**, 85.
44. Fontaine,J.-F., Barbosa-Silva,A., Schaefer,M. *et al*. (2009) MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res.*, **37**, W141–W146.
45. Charniak,E. and Johnson,M. (2005) Coarse-to-fine n-best parsing and maxent discriminative reranking. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05*. Association for Computational Linguistics, Stroudsburg, PA, pp. 173–180.
46. Neves,M., Damaschun,A., Kurtz,A. *et al*. (2012) Annotating and evaluating text for stem cell research. In: *Proceedings of the Third Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2012) at Language Resources and Evaluation (LREC)*. Istanbul, Turkey, pp. 16–23.
47. Bard,J., Rhee,S. and Ashburner,M. (2005) An ontology for cell types. *Genome Biol.*, **6**, R21.
48. Sarntivijai,S., Xiang,Z., Meehan,T.F. *et al*. (2011) Cell line ontology: redesigning the cell line knowledgebase to aid integrative translational informatics. In: Bodenreider,O., Martone,M.E. and Ruttenberg,A. (eds), *ICBO*, Vol. 833: CEUR Workshop Proceedings. CEUR-WS.org.
49. Bard,J. (2012) A new ontology (structured hierarchy) of human developmental anatomy for the first 7 weeks (carnegie stages 1–20). *J. Anat.*, **221**, 406–416.
50. Malone,J., Holloway,E., Adamusiak,T. *et al*. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
51. Rosse,C. and Mejino,J.L.V. (2008) The foundational model of anatomy ontology. In: Burger,A., Davidson,D. and Baldock,R. (eds), *Anatomy Ontologies for Bioinformatics*, Vol. 6. Computational Biology, Springer, London, pp. 59–117.
52. Consortium,T.G.O. (2010) The gene ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
53. Hayamizu,T.F., de Coronado,S., Fragoso,G. *et al*. (2012) The mouse-human anatomy ontology mapping project. *Database*, article ID bar066; doi: 10.1093/database/bar066.
54. Mungall,C., Torniai,C., Gkoutos,G. *et al*. (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5.
55. Maglott,D., Ostell,J., Pruitt,K.D. *et al*. (2005) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33** (Database issue), D54–D58.
56. Aronson,A.R. and Lang,F.-M. (2010) An overview of metamap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.*, **17**, 229–236.
57. Whetzel,P.L., Noy,N.F., Shah,N.H. *et al*. (2011) BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Res.*, **39**, 541–545.
58. Pyysalo,S., Airola,A., Heimonen,J. *et al*. (2008) Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, **9 (**Suppl. 3), S6.
59. Bjorne,J., Ginter,F. and Salakoski,T. (2012) University of Turku in the BioNLP'11 Shared Task. *BMC Bioinformatics*, **13**, S4.
60. Stenetorp,P., Pyysalo,S., Topic,G. *et al*. (2012) BRAT: a web-based tool for nlp-assisted text annotation. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon, France, pp. 102–107.
61. Cano,C., Monaghan,T., Blanco,A. *et al*. (2009) Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. *J. Biomed. Inform.*, **42**, 967–977.
62. Huang,M., Liu,J. and Zhu,X. (2011) Genetukit: a software for document-level gene normalization. *Bioinformatics*, **27**, 1032–1033.
63. Leaman,R. and Gonzalez,G. (2008) Banner: an executable survey of advances in biomedical named entity recognition. *Pac. Symp. Biocomput.*, **13**, 652–663.
64. Bui,Q.-C. and Sloot,P.M. (2012) A robust approach to extract biomedical events from literature. *Bioinformatics*, **28**, 2654–2661.
65. Neves,M., Carazo,J.-M. and Pascual-Montano,A. (2009) Extraction of biomedical events using case-based reasoning. In: *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*. Boulder, CO, pp. 68–76.
66. Cano,C., Labarga,A., Blanco,A. *et al*. (2011) Social and semantic web technologies for the text-to-knowledge translation process in Biomedicine, doi:10.5772/13560.