The BEL Information Extraction Workflow (BELIEF): Updates and Evaluation

Sam Ansari[1], Sumit Madan[2], Justyna Szostak[1], Philipp Senger[2], Marja Talikka[1], Juliane Fluck[2], and Julia Hoeng[1]
[1]Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland (part of Philip Morris International group of companies)
[2]Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany

Introduction

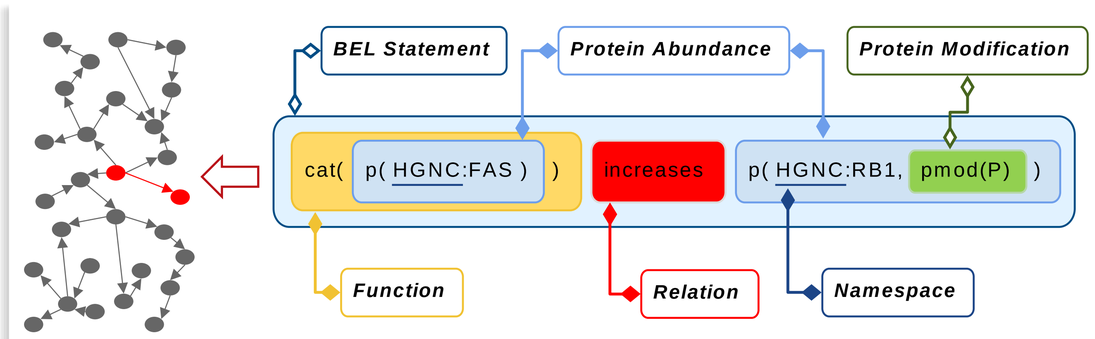
The pace in which knowledge is published in the scientific domain is much higher than its application in the interpretation of biological data. In order to reduce this gap, methods are required to convert literature knowledge into a more applicable format, here network models in the biological domain [1].

In 2014, we had introduced the BEL Information Extraction workFlow (BELIEF) [2], a semi-automated workflow featuring a text mining pipeline as well as a curation interface. Based on natural language processing (NLP) BELIEF automatically extracts biological entities as well as detects the relationship they have with eachother. These triples are coded in BEL and used for the interpretation of mainly high-throughput data such as transcriptomic data [3].

In this poster, we present the new version with an improved text mining pipeline as well as a new curation interface and show performance indicators that were collected from the BioCreative V Track 4 setup [4,5] and IAT.

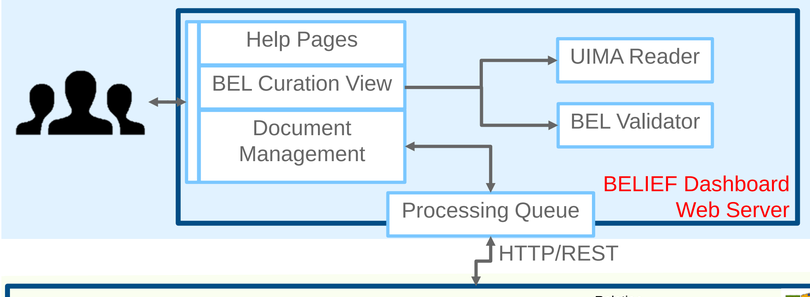
BEL (Biological Expression Language)

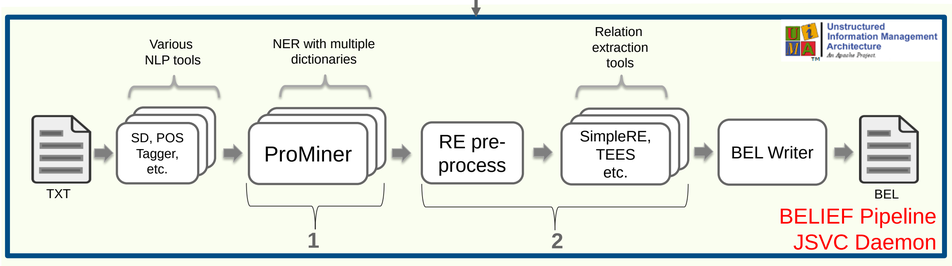
BEL [6] is a machine and human-readable language that represents molecular relationships and events as semantic triples where context can include information about the biological and experimental system in which the relationships were observed as well as the supporting publications cited.
Unlike other knowledge representation standards such as BioPAX and SBML, BEL comes very close to natural language and proved suitable as the exchange format between text mining and human curation.



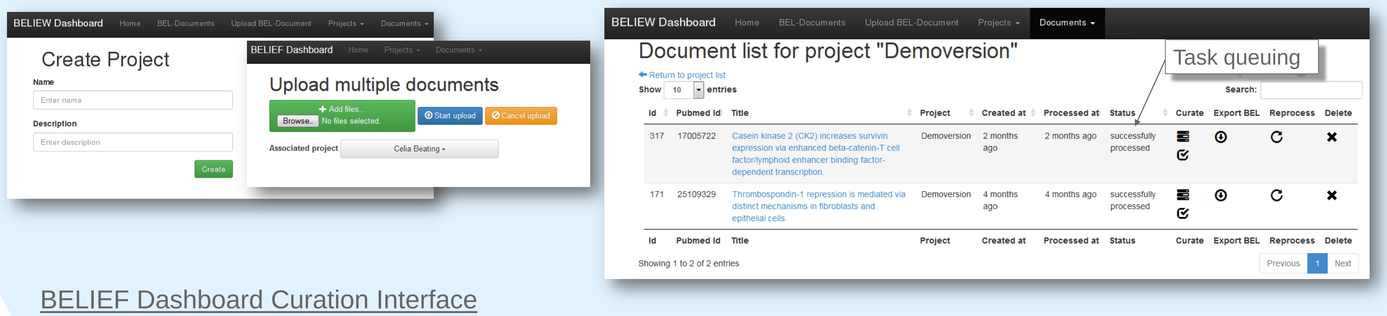

```
SET Citation = {"PubMed", "Regulation of
Rb and E2F by signal transduction
cascades: divergent effects of JNK1 and
p38 kinases." , "EMBO J. 1999 Mar
15;18(6):1559-70." , "10075927"}

SET Evidence =
        "Fas stimulation of Jurkat cells is
         known to induce p38 kinase and we
         find a pronounced increase in Rb
         phosphorylation within 30 min of
         Fas stimulation"

SET Tissue = "jurkat cells"
```

BELIEF Improvements



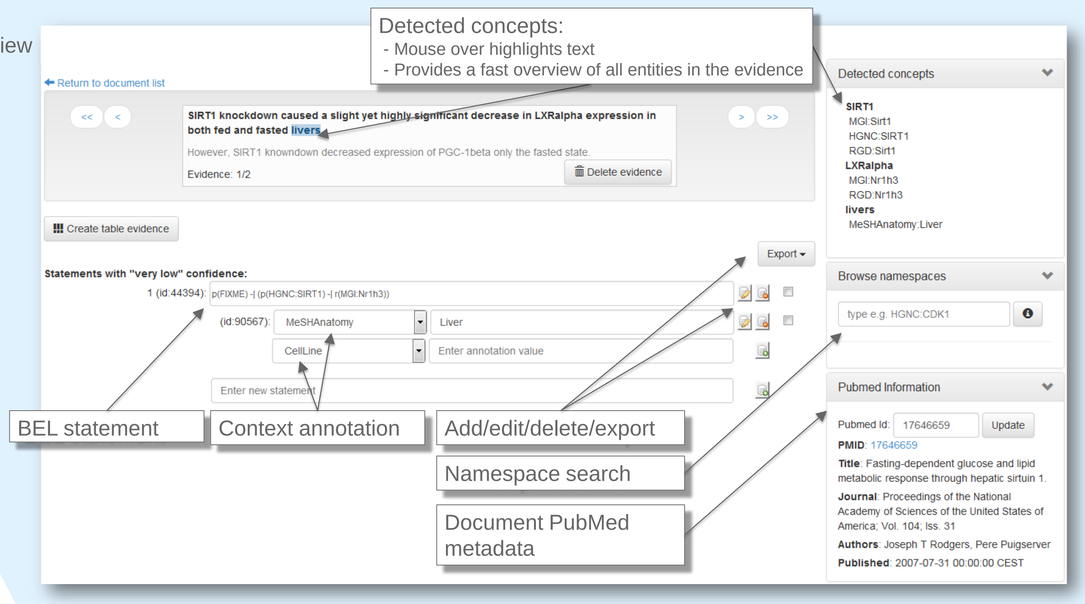


Project and Task Management

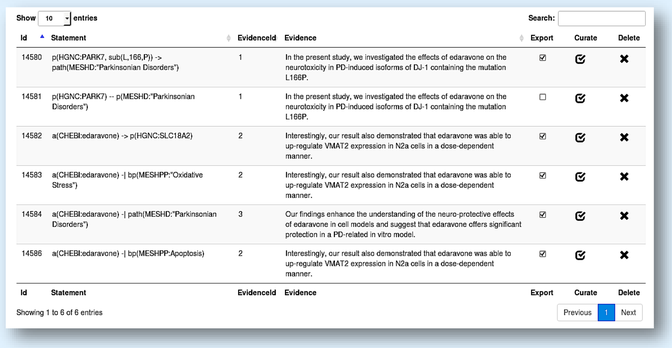


BELIEF Dashboard Curation Interface

Evidence centric curation view



Statement centric curation view

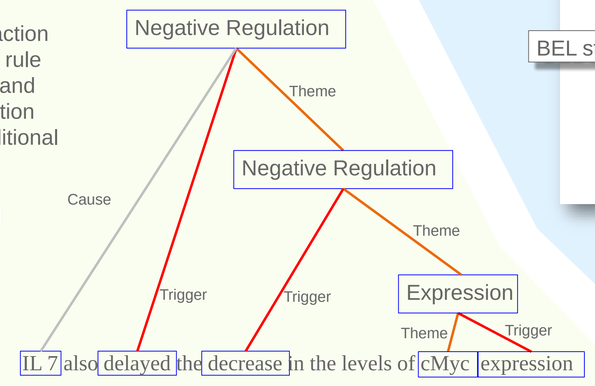


1. Integrated dictionaries for different classes with normalization

Entity class	Resources	BEL namespace
Human genes/proteins	EntrezGene/Uniprot	HGNC
Mouse genes/proteins	EntrezGene/Uniprot	MGI
Rat genes/proteins	EntrezGene/Uniprot	RGD
Protein family names	OpenBEL	SFAM
Protein complex names	OpenBEL	SCOMP
Protein complex names	Gene Ontology	GOCC
Biological processes	Gene Ontology	GOBP
Chemical names	OpenBEL	SCHEM
Chemical names	ChEBI	CHEBI
Chemical names	ChEMBL	CHEMBL
Disease names	MeSH	MESHD
Anatomical names	MeSH	MeSHAnatomy
Cell lines	Cell Line Ontology	CellLine
Cell structures	MeSH	CellStructure

2. The Relation Extraction (RE) preprocess selects unified entity annotations for relation extraction. The relation extraction tools only receive their own unified annotation, all other matches are preserved for the human curator.

The relation extraction is performed with rule based extraction and relation classification as well as an additional deep parsing via BioNLP tools:
- LibLINEAR [7]
- TEES2.1 [8]



Performance

Dictionary	Recall rate initial version	Recall rate application adapted
Genes/Protein: (HGNC)	80 %	93 %
Chemical compounds: ChEBI	15 %	66 %
Chemical compounds: SCHEM	30 %	75 %
Chemical compounds: ChEBI + SCHEM+ ChEMBL	not determined	91 %
Selventa-human-complex	40 %	46 %
GO-Complex	not determined	64 %
Selventa-human-complex + Complex	not determined	82 %
GO-Function	22 %	not determined
Selventa-human-families	8 %	77 %

Use case: relation between small molecules (mainly protein inhibitors) and their targets

Learnings for higher recall (BELIEF 2014 vs current:
- Use external and internal (OpenBEL) resources for named entity recognition and ensure mapping and normalization
- Combine various resources

NLP Sentence Detection	~6% error
NLP Tokenization	~8% error
NER Different Classes	~15% error
RE Multi-step	~25% error

complexity

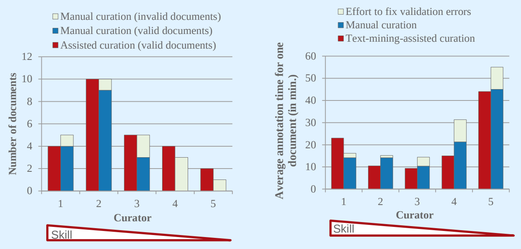
Propagated error

Class	Precision	Recall	F-score
Term	81.34	72.67	76.76
Function-Secondary	66.67	39.29	49.44
Function	51.16	33.33	40.37
Relationship-Secondary	56.65	73.76	64.09
Relationship	67.37	31.68	43.10
Statement	59.15	20.79	30.77

Test set prediction results for several classes of BioCreative V BEL track task 1 (100 sentences).
Compared with the outcome of the BEL track task 1, BELIEF generated the highest F-score for entirely correct BEL statements (30.8% versus 20.2% for the best BioCreative evaluated system)

Knowledge Type detected	Manual Curation Network Knowledge	Assisted Curation Network Knowledge
Number of nodes	63	76
Number of edges	94	128
Number of evidences	21	26
Chemical abundance	5	7
Protein abundance	30	32
RNA abundance	2	3
Complex abundance	3	5
Biological process	10	14

First Application: Manual Curation vs. a semi-automated curation process for causal knowledge extraction



At the BioCreative V IAT track, 5 curators that did not know the system before tested the BELIEF Dashboard. The testers went through a tutorial and training before they received an annotation guideline to perform the actual testing. Although none of the testers had BELIEF experience, they curated faster (except tester 1) with more extracted statements (20% more on average [data not shown]).

A System Usability Scale based on a questionnaire was filled and resulted in a score of 67 which is an average usability score for a very specialized tool. Below some comments:
"*The complexity was in the BEL language itself; the BELIEF system actually made it easier to start understanding how interactions were encoded.*"
"*The system is very easy to learn for a user who is already familiar with BEL.*"
"*In particular, the preselected protein identifiers were immensely useful (which I only found out when I tried to find them by hand).*"

Summary

BELIEF in its current version better supports domain experts in different stages of knowledge acquisition and network model creation. The results certify that BELIEF shows an improved performance in both, accuracy and recall, as well as a good system usability. The overall impression of all untrained testers was that BELIEF speeds up and further simplifies the creation of BEL statements.

The new and impactful features are:
- Single point of entry including document and task management
- Reduced BEL coding effort due to full and partial BEL statement generation and validation on modifications
- Automatic citation from the Pubmed ID
- Two curation views to facilitate curation (evidence and statement centric view)
- Possibility to use custom dictionaries and re-running the text mining pipeline with these
- Show adjacent sentences to support curation

The key learnings from the user acceptance testing are:
- The success of curation tools lies in providing all relevant information to the curator and limit the curation task to the actual limitations of the automated system
- The preparation of annotation guidelines is critical for consistent annotations across several users
- Comprehensive supporting material is required to facilitate BEL coding
- Collaborative curation is becoming more and more common and should be supported

http://www.scaiview.com/belief/

References

[1] Hoeng, J., Deehan, R., Pratt, D., et al. (2012) Drug Discov. Today, 17, 413–8, A network-based approach to quantifying the impact of biologically active substances.
[2] Fluck, J., Madan, S., Ansari, S., et al. (2014) Proc. 6th Int. Symp. Semant. Min. Biomed. pp 109–113, BELIEF - A semiautomatic workflow for BEL network creation.
[3] Martin, F., Sewer, A., Talikka, M., et al. (2014) BMC Bioinformatics, 15, 238, Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models.
[4] Fluck, J., Madan, S., Ansari, S., et al. (2016) Submitt. to Database J. Biol. Databases curation, 2016, Training corpora for the extraction of causal relationships coded in Biological Expression Language (BEL).
[5] Rinaldi, F., Effendorf, T., Madan, S., et al. (2016) Database (Oxford)., 2016, submitted, BioCreative V Track 4: A Shared Task for the Extraction of Causal Network Information in Biological Expression Language.
[6] Slater, T. (2014) Drug Discov. Today, 19, 193–198, Recent advances in modeling languages for pathway maps and computable biological networks.
[7] Fan, R. E., Chang, K. W. and Hsieh, C. J. (2008) J. Mach. Learn., 9, 1871–1874, LIBLINEAR: A library for large linear classification.
[8] Björne, J., Ginter, F. and Salakoski, T. (2012) BMC Bioinformatics, 13 Suppl 1, S4, University of Turku in the BioNLP'11 Shared Task.

The research described in this poster was funded by Philip Morris International

The 9th International Biocuration Conference
Geneva, Switzerland
April 10-14, 2016

Philip Morris International Research & Development, Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland
T: +41 58 242 21 11, **F**: +41 58 242 28 11, **W**: www.pmi.com