BELIEF Dashboard – a Web-based Curation Interface to Support Generation of BEL Networks

Sumit Madan[1,*], Sven Hodapp[1], and Juliane Fluck[1]

[1]Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany
{sumit.madan,sven.hodapp,juliane.fluck}@scai.fraunhofer.de

Abstract. The relevance of network-based approaches in systems biology to achieve a better understanding of biological mechanisms has increased enormously. The Biological Expression Language (BEL) is well designed to collate findings from scientific literature into biological network models. To facilitate encoding and biocuration of such findings in BEL, a free and user-friendly web-based curation interface called BELIEF Dashboard has been developed. The interface incorporates an information extraction workflow to support the biocurator with text mining techniques.
BELIEF Dashboard allows easy curation of generated BEL statements and their context annotations. It also visualizes the evidence sentences and highlights detected named entities with their disambiguated normalized names. Resulting BEL statements and their context annotations can be syntactically and semantically verified which ensures consistency in the BEL network. Additional system functionalities provide a document management system, the retrieval of citation information from PubMed, a search interface for OpenBEL namespaces and annotation definitions, and also the export of valid BEL documents. In summary, the curation tool supports experts in different stages of systems biology network building. We show that BEL coding with BELIEF Dashboard reduces the curation effort and improves the efficiency.
Availability: BELIEF Dashboard is available at the following link: http://www.scaiview.com/belief

Keywords: Network generation, Biological Expression Language (BEL), OpenBEL, Text Mining, Relation Extraction.

1 Introduction

To study the complex mechanisms of biological systems, network-based approaches such as protein-protein interaction, metabolic, signaling, regulatory, or co-expression networks are emerging in the field of systems biology. A computational representation of knowledge in a well-defined structured and standardized language is required for systematic network analysis. Biological PAthway eXchange language (BioPAX) [1], an ideal format for database exchange, and Systems Biology Markup Language (SBML) [2], a structured XML-based language, are the two most popular network-

modeling languages in systems biology. Despite both have been used to model biochemical reaction networks, they have the drawback of being not human readable. They also have restricted possibilities to represent the various knowledge layers published in scientific literature.

BEL (Biological Expression Language) is designed and engineered for systems biology experts to drive biological network-based analytics [3]. BEL conserves the biological causal and correlative relationships gathered from scientific literature with contextual and provenance information in a computable form. The relationships, also referred to as BEL statements, are triples decomposable into a subject, a predicate, and an object (c.f. Fig**.** 1). An assembly of many BEL statements produces a causal network knowledgebase which can be used to understand and analyze the underlying biological mechanisms. For example, several network models have been used to understand the cause-and-effect mechanisms of pulmonary and vascular systems [4, 5].

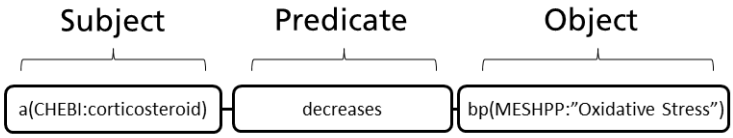

Fig. 1. Structure of a BEL statement [6]. The example BEL statement describes that an abundance of the chemical *corticosteroid reduces* the biological process *Oxidative Stress*. For identification and disambiguation of domain specific terms and concepts several predefined BEL namespaces (e.g. CHEBI) are used.

The extraction of BEL statements and thus the construction of network models have been done manually by experienced users through biocuration. Biocuration is a process to translate biological findings and information to a formal and structured representation of biological data [7]. The extraction of such findings from scientific literature is an important and time-consuming task [7]. To handle the huge volume of emerging and published literature and accelerate the curation, biocurators need support through automated mining systems [7]. Text mining assisted biocuration systems which extract current knowledge from the scientific literature, represent one of the several examples of such a system. In this paper, we present the BELIEF Dashboard which uses a text mining workflow to support knowledge extraction from literature and allows biocuration of causal and correlative BEL statements with their context annotations.

2 System Description and Functionalities

The BELIEF workflow contains two main components – the text mining pipeline (BELIEF Pipeline) and the web-based curation tool (BELIEF Dashboard) (c.f. Fig. 2). Both components communicate through a HTTP/REST API.

2.1 BELIEF Pipeline

The component BELIEF Pipeline includes an UIMA[1]-based text mining workflow where several state-of-the-art natural language processing (NLP), named entity recognition (NER) and relationship extraction (RE) tools are implemented and combined to facilitate the information extraction.

The dictionary- and rule-based NER software ProMiner which achieved an f-measure of 0.8 in the BioCreative II NER assessments has been used in the workflow [8, 9]. To detect several classes of named entities (like human genes, chemicals, biological processes, diseases etc.), various dictionaries have been integrated into the workflow. These dictionaries are used and optimized for systems biology use cases [10].

For relationship classification a LibLinear-based approach [11] as well as the BioNLP-based software TEES (Turku Event Extraction System) [12] are integrated. Unlike most other approaches, the workflow converts the automatically extracted results into full and valid BEL documents. A comparison of BioNLP-shared task annotation with BEL and its conversion to BEL is described in Fluck et al. 2013 [13]. Further details of the text-mining pipeline can be found in Fluck et al. 2014 [10].

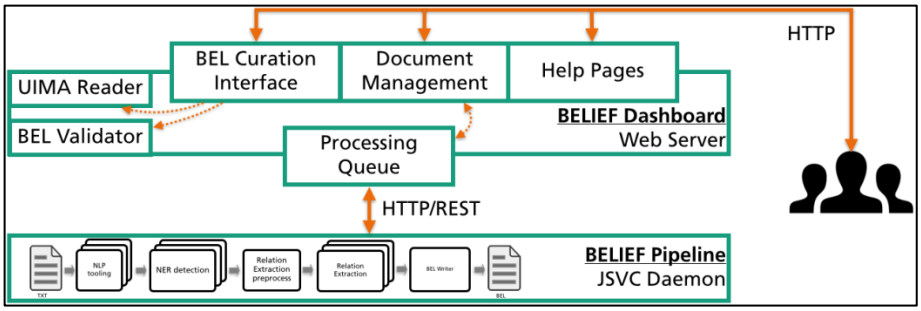

Fig. 2. Architecture of semi-automated information extraction workflow BELIEF [10]. The workflow consists of a text mining pipeline BELIEF Pipeline and a web-based biocuration tool BELIEF Dashboard.

2.2 BELIEF Dashboard

The interactive web application BELIEF Dashboard supports the curation of BEL statements and context annotations which are automatically extracted by the pipeline. During the implementation of the application, we applied the concepts of user-friendliness and user-interactivity to achieve high biocuration efficiency. The web application framework Grails is the basis of the implementation. Grails allows a seamless integration of both Java-based frameworks UIMA and OpenBEL. The UIMA framework is used to parse the results of the pipeline. The OpenBEL framework is used to parse and validate the BEL statements with their context annotations

[1] http://uima.apache.org

against the various OpenBEL namespaces and annotation definitions. The application uses the H2 database to save the persistent data.

Help Pages. The web interface offers help pages to introduce several functionalities of the system. As BEL is not widely used within the biomedical community, a short introduction page is offered to get familiar with the main concepts of BEL. Also the process of biocuration with the curation interface is explained in a detailed step-by-step tutorial. Various features of the user interface are described in detail as well.

Project and Document Management System. A project in BELIEF Dashboard represents a curation process project which can contain several documents. The manual curation is performed on a single document. For this purpose a document management is available which allows adding, updating and deleting any text documents. After curation a user can download or export a BELScript document containing the curation results.

Document Processing Queue. Every added text document is queued into a processing queue. As soon as the text mining pipeline BELIEF is idle again, it automatically starts retrieving the next unprocessed document from the queue and starts processing it. The document status changes to 'processing started'. After the processing has finished the results are pushed to the BELIEF Dashboard. When the processing is successfully completed the system will change the document status to 'successfully processed'. In case of processing failure the status changes to 'processing failed'.

BEL Curation Interface. The successfully processed documents are available for manual curation to the users. For this purpose the user can choose between the statement-centric or the evidence-centric curation view. The statement-centric view lists all detected statements with their evidences in tabular form. The evidence-centric curation view (c.f. Fig. 3) visualizes the extracted BEL statements for an evidence. To enable a better understanding of the context, sentences surrounding the evidence text are shown as well. The detected concepts found in the evidence are listed in the right area. The annotated text in the evidence is highlighted when the mouse is held over the detected concepts. The concepts consist of BEL namespaces and their normalized names detected by the NER software ProMiner. The identified BEL statements are shown in the lower area of the curation view. It allows adding, updating and deleting statements. The context annotations such as organism, anatomy, or disease can be edited as well. The statements and context annotations are automatically validated for correct syntax, valid semantics, namespaces and reference citation. The errors and warnings are provided to the user by a notification box on the lower right side of the view.

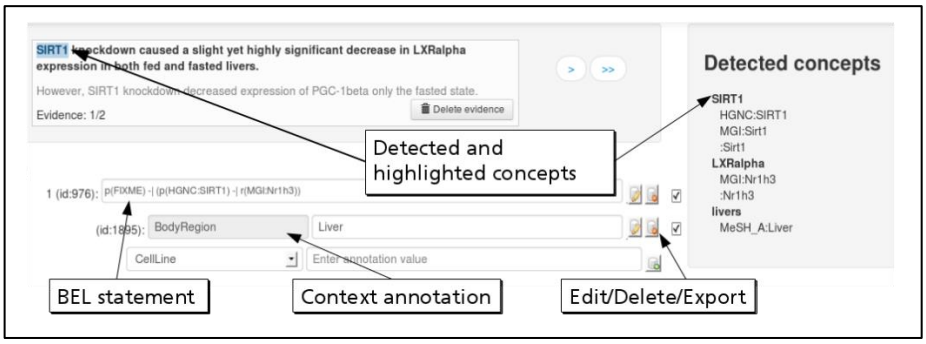

Fig. 3. Screenshot of the evidence-centric curation view. The top left side visualizes the evidence text. Detected concepts in the current evidence text are shown in the top right area. The bottom left area allows the curation of BEL statements and their context annotations.

BEL Validator. Only valid statements and context annotations can be used for the network creation. Thus, it is essential that the BEL data is syntactically and semantically valid. The BEL statements and context annotations are validated as soon as they are added or changed in the system. The input data is validated in two different ways – syntactically and semantically. If the data is invalid, a message will be triggered and visualized to the curator.

3 Evaluation results

3.1 Text mining results

The performance of the text mining tool has been reported in Fluck et al. [10]. ProMiner software has been used for NER which reached recall and precision value of approximately 80% for human and mouse gene/protein name recognition in the past BioCreative assessments [8, 9]. The evaluation of relation extraction was done on sentence level where only sentences with correctly annotated proteins were considered. The LibLinear-based classification method and TEES extracted 60% and 42% correct protein pairs respectively. An overall recall rate of 74% was achieved with the combination of both methods. For results of the LibLinear-based method, more manual curation effort is necessary to generate complete and valid BEL statements.

3.2 IAT task results

The *BioCreative V User InterActive Task* (IAT) conducts a formal evaluation of the text mining systems for a specific biocuration task. The systems are evaluated upon performance (time-on-task and accuracy of the text mining assisted curation compared to manual curation) and a subjective measure via a user survey [14].

Participants. A total of seven curators were invited by the task organizers and were provided access to the BELIEF Dashboard out of which only five participated fully in this task. In the beginning of the task a survey was conducted for the participants. Before the assignment of the tasks, the participants were first introduced to BEL and subjected to a three-document-based training session to demonstrate the curation process and important characteristics of the system. The curators were randomly divided into two groups.

Document corpus. A total of 20 PubMed abstracts were included in the document corpus. The documents were selected from different domain areas but considering the context for which BELIEF was created. These documents were divided into two sets (Set1 and Set2) containing 10 documents each. Table 1 describes the usage of the document sets by the two participant groups. Set1 was assigned to Group1 for the text mining assisted curation and to Group2 for manual curation, whereas the curation type was transposed for the dataset Set2.

	Assisted curation	**Manual curation**
Group1 (2 Users)	Set1	Set2
Group2 (3 Users)	Set2	Set1

Table 1. Usage of the document sets by the two participant groups

Annotation Guidelines. To guide the biocuration and create consistent high quality annotations, well defined annotation guidelines play a crucial role. The prepared guidelines are designed to be easy to follow. The guidelines for curation with BELIEF Dashboard are available on the help page.

Biocuration results. In total 25 documents were curated in 392 minutes (6.53 hours) with the BELIEF Dashboard. The manual curation produced 24 documents in 374 minutes (6.23 hours) out of which 8 documents (33.3 %) were invalid. These documents include several syntax errors which obstructs the BEL compilation to networks. The documents curated with the BELIEF Dashboard contained 243 BEL statements and those with the manual curation contained 202 statements. The manually curated invalid documents contained 64 statements (31.68 %). Fig. 4. shows the number of curated documents for each annotator for both curation types. All curators produced invalid documents but only during the manual curation.

The time required by the biocurators for curation with BELIEF Dashboard and for manual curation is shown in Fig. 5. Additional effort has to be considered regarding the several invalid documents from manual curation. The following tasks are necessary to clean up these invalid documents: (1) compile document with OpenBEL Framework, (2) localize the syntax error in the BEL document, (3) fix the error, and (4) repeat until all errors are fixed. To fulfill these tasks we are assuming additional 10 minutes effort for a single invalid document. The additional time has also been added to the average annotation time (c.f. Fig. 5.). In comparison to manual curation

the curators used approximately 22% less time for assisted curation. This calculation also considers additional effort which is essential to generate consistent networks.

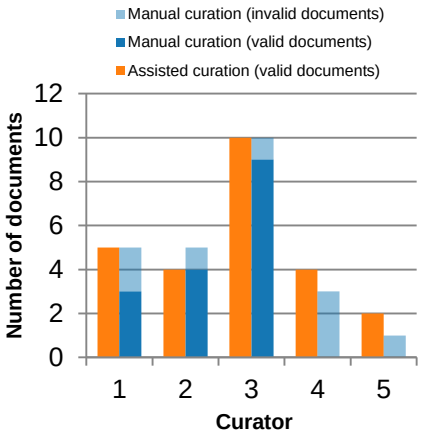

Fig. 4. Number of documents curated by all curators

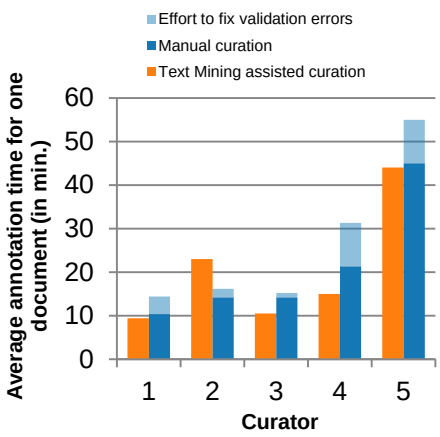

Fig. 5. Time usage by annotators and curation type

Survey results. The second survey consisted of a 10-item questionnaire. These questions are derived from the *System* Usability *Scale*[2] (SUS) method which measures the perception of usability and learnability of a software product. From the 10 questions, eight measure the usability. The remaining two questions quantify the learnability. The score ranges from 0 to 100 and a score of 68 (not a percentage) is considered average. BELIEF Dashboard achieved a calculated average SUS score of 66.67. The system has reached an average SUS score of 67.185 and 64.58 for usability and learnability respectively.

Within the survey, the participants were also asked to comment the questions if needed. Some of the comments are listed below:

1. *Complexity of BEL*: "The complexity was in the BEL language itself; the BELIEF system actually made it easier to start understanding how interactions were encoded." and "The system is very easy to learn for a user who is already familiar with BEL."
2. *Concept search*: "More tools could be added to help the curator such as a search by term not only by NameSpace+Term. It would also be useful to know when the available NameSpaces that are available were updated in the case that there is a missing term one can know if it is a bug or an update issue."
3. *Named entity detection*: "In particular, the preselected protein identifiers were immensely useful (which I only found out when I tried to find them by hand)."
4. *Relation extraction*: "It was cumbersome to sort through the less relevant results, but overall the system was easy to use and the disambiguation was helpful."

[2] http://www.measuringu.com/sus.php

5. *BEL validator*: "No way of suggesting how to correct mistakes or any correct examples."
6. *Team support*: "I found it really interesting to participate in this task, people were extremely helpful and fast in answering the questions I had."

3.3 Conclusion

We presented a user-friendly web-based curation interface which incorporates a semi-automatic knowledge extraction workflow to support network building for systems biology. It allows biocurators to extract knowledge from biomedical literature and curate causal and correlative relationships encoded into BEL. During the IAT task the curation interface was evaluated based on the performance and user survey. The task helped to identify various aspects of the interface which were helpful for the curation and revealed important issues for future improvements. We showed that BELIEF Dashboard increases the curation efficiency in comparison to the manual curation.

4 Acknowledgements

We acknowledge the help of Selventa Inc. supporting the integration of the OpenBEL framework in our system. We thank Heinz-Theodor Mevissen to help us integrating the ProMinerNER system into the pipeline.

We are very grateful for the support of BioCreative IAT team who provided us the opportunity to take part in the IAT task and organized the whole task. Various discussions and feedback rounds helped us to create a usable curation environment for the biocurators. We also want to thank the participants who tested our system and helped us identifying several improvement points.

Funding. We acknowledge support of our research from Philipp Morris International.

Conflict of Interest. None declared.

5 References

1. Demir E, Cary MP, Paley S, et al. (2010) The BioPAX community standard for pathway data sharing. Nat Biotechnol 28:935–942. doi: 10.1038/nbt.1666
2. Finney a M, Hucka M, Sauro HM, et al. (2001) The Systems Biology Markup Language. Mol Biol Cell 12:708.
3. Slater T (2014) Recent advances in modeling languages for pathway maps and computable biological networks. Drug Discov Today 19:193–198.
4. Schlage WK, Westra JW, Gebel S, et al. (2011) A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue. BMC Syst Biol 5:168. doi: 10.1186/1752-0509-5-168
5. Boué S, Talikka M, Westra JW, et al. (2015) Causal biological network database: a comprehensive platform of causal biological network models focused on the

pulmonary and vascular systems. Database (Oxford) 2015:bav030–. doi: 10.1093/database/bav030

6. Fluck J (2015) BELIEF – A semiautomatic workflow for BEL network creation. In: Conf. From Big Data to Smart Knowl. – Text Data Min. Sci. Econ. http://textmining.congressbuero.de/sites/default/files/textmining/Fluck_Talk.pdf.

7. Howe D, Costanzo M, Fey P, et al. (2008) Big data: The future of biocuration. Nature 455:47–50. doi: 10.1038/455047a

8. Hanisch D, Fundel K, Mevissen H-T, et al. (2005) ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics 6 Suppl 1:S14.

9. Fluck J, Mevissen H-T, Dach H, et al. (2007) ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries. In: Hirschmann L, Krallinger M, Valencia A (eds) Proc. Second BioCreative Chall. Eval. Work. pp 149–151

10. Fluck J, Madan S, Ansari S, et al. (2014) BELIEF - A semiautomatic workflow for BEL network creation. In: Proc. 6th Int. Symp. Semant. Min. Biomed. pp 109–113

11. Bobic T, Klinger R, Thomas P, et Al. (2012) Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactions. In: Proc. ROBUS-UNSUP 2012 Jt. Work. Unsupervised Semi-Supervised Learn. NLP. pp 35–43

12. Björne J, Salakoski T (2013) TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task. In: Proc. BioNLP Shar. Task 2013 Work. Association for Computational Linguistics, pp 16–25

13. Fluck J, Klenner A, Madan S, et al. (2013) BEL networks derived from qualitative translations of BioNLP Shared Task annotations. In: Proc. BioNLP Shar. Task 2013 Work. Association for Computational Linguistics (ACL), pp 80–88

14. Arighi C (2015) BioCreative V - User Interactive Task (IAT). http://www.biocreative.org/tasks/biocreative-v/track-5-IAT/. Accessed 21 Aug 2015