Overview of BEL Track: Extraction of Complex Relationships and their Conversion to BEL

Sumit Madan[1], Justyna Szostak[2], Jens Dörpinghaus[1], Julia Hoeng[2] and Juliane Fluck[1]
[1]Fraunhofer Institute SCAI, Schloss Birlinghoven, 53757 Sankt Augustin, Germany
[2]Philipp Morris International R&D, Philip Morris Products S.A, Quai Jeanrenaud 5, 2000 Neuchatel, Switzerland

Abstract— **Biological signaling is complex and our knowledge about it is often only available in literature. Signaling can involve small molecules as well as proteins that can be activated or deactivated by various regulations such as ligand binding, complex forming, modification status or miRNA binding. Changes in signaling influence biological processes and/or are involved in disease etiology. The Biological Expression Language (BEL) has been created to store this kind of information in a structured form that can be used for network generation and visualization as well as interpretation of experimentally generated data. The BioCreative VI BEL track provides training data and an evaluation environment to encourage the text mining community to tackle the automatic extraction of such complex relationships as well as converting it to BEL. Although only a few groups participated in this track, the groups participating the second time could drastically increase their performance. The best system reached 32% F-score for extraction of complete BEL statements (task 1) and, when given the named entities, above 49%. Beside rule-based systems, methods involving hierarchical sequence labeling and neural networks are adapted to this task. For the second task in the BEL track, finding evidence text snippets for a given statement, despite the provided training data, only one team took part.**

Keywords—text mining; relation extraction; named entity recognition; entity normalization; evaluation; BEL

I. Introduction

Biological Expression Language (BEL) has been introduced to allow the formalization of causal biomedical relationships (1). The resulting BEL statements can be assembled into causal networks, which can be easily queried or used for data mining. The manual extraction of this information from literature is time consuming. Consequently, there is a high demand for automatic support. Currently, the automatic extraction has not yet reached the performance to allow for fully automatic extraction of this information. Even for humans, depending on training and application area, formalizing complex knowledge is demanding and not unambiguous. Nevertheless, due to the increase of large scale experiments and the requirements of molecular knowledge in precision medicine, the demand for structured biomedical knowledge is increasing. The BEL track has been introduced in BioCreative V 2015 to provide training data for the extraction and BEL translation of complex relationships over a set of different named entity classes (2). In addition to the relationship information, BEL has the advantage to provide provenance information in form of text evidences that are well suited to serve as training set. During the first BEL track, we appraised a large corpus of BEL documents resulting in a training corpus of BEL statement–sentences pairs (3). A development and the test set were analyzed by annotators and curated to assure that all possible BEL statements are associated with the sentences. In addition, information resources to explain BEL and an evaluation framework has been set up (2). For the current task, novel and yet unpublished test data from the disease context of Ulcerative Colitis (4,5) has been created. In the next chapter, the task, relevant resources, the created test set and result summaries of the participating systems are presented.

II. Task Overview

A. Biological Expression Language and used namespaces

BEL statements encode semantic triples with subject, relationship and object. An example BEL statement with the corresponding sentence is shown in Fig. 1. For the BEL track, we focus on two causal relationship types: increase and decrease. Subject and object contain entities that are normalized to so-called namespaces. Those namespaces are generated from database entries (e.g. human genes from HGNC[1], mouse genes from MGI[2] and chemical entities from ChEBI [3] database). Other namespace origin either from ontologies, such as the biological processes subtree of the Gene Ontology for GOBP[4], or from terminologies, such as the diseases namespace MESHD [5] from the Medical Subject Heading terminology disease subtree. By using the normalized entities from such namespaces, the resulting statements can be integrated and merged to networks as well as aligned to other data.

For the different entities, different class abundances are assigned: the abundance function *a()* is assigned to chemicals, *bp()* for biological processes and *path()* (pathology) for diseases. For genes, different abundances *g()* (gene), *r()*

We acknowledge support of our research from Philip Morris International.

[1] HGNC stands for HUGO Gene Nomenclature Committee (http://www.genenames.org)

[2] MGI stands for Mouse Genome Informatics (http://www.informatics.jax.org/)

[3] ChEBI stands for Chemical Entities of Biological Interest (https://www.ebi.ac.uk/chebi/)

[4] GOBP stands for Gene Ontology Biological Process

[5] MESH disease subtree is available at https://meshb.nlm.nih.gov/treeView.

(mRNA) or *p()* (protein) are possible, but only *p()* is used in the BEL track to reduce the complexity. In addition to those class abundances, different functions can also be assigned to the biological entities. For the BEL track, we focus on the protein phosphorylation function *pmod()*, the activity function *act()* to describe a protein activation, the *tloc()* function to describe a translocation, *deg()* for degradation and finally *complex()* to describe protein complexes. For a more detailed description we refer to Rinaldi et al. (2).

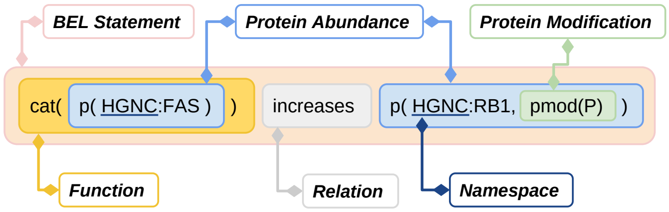


Fig. 1 Example of a BEL statement extracted manually for the sentence "Fas stimulation of Jurkat cells is known to induce p38 kinase and we find a pronounced increase in Rb phosphorylation within 30 min of Fas stimulation" (PMID:10075927). Reprinted with permission from Rinaldi et al. (2).

B. *Task description and evaluation*

The BEL track challenge is organized into two tasks evaluating the complementary aspects of the problem:

- ***Task 1: Given textual evidence for a BEL statement, generate the corresponding BEL statement.***
- ***Task 2: Given a BEL statement, provide at most 10 additional evidence sentences.***

Extraction of relationships and their coding in BEL is a complex task due to the different entity, relationship and function types. Therefore, we simplified the challenge further and provided a cascade model for evaluation of task 1. A detailed overview of all simplifications is provided online[6]. In short, HGNC or EntrezGene identifiers are accepted for the same statement and mouse orthologous identifiers are accepted as well. For the abundance function of those namespaces, all correct abundance are accepted. Furthermore, for the modification function *pmod()* and the translocation function *tloc()*, the number of arguments are reduced.

The cascade evaluation model is described in detail in Rinaldi et al. (2). Only syntactically correct statements in a correct format are accepted, but different levels of performance are evaluated. A submitted full BEL statement is automatically cut into its fragments to ensure evaluations on lower levels. On term level, only the correctness of BEL terms is assessed. On this level, the correctness of the discovered entities, the correctness of associated namespaces and their format as well as the correctness of the associated abundance/process function is measured.

On function level the correctness of discovered function is evaluated. Functions are only accepted together with their argument, the BEL term. As simplification, a complex function is only valid if at least one of its arguments is correct. On the secondary function level, the correctness of a function alone was measured, regardless of the correctness of their term-arguments.

In the relationship-level evaluation, only the entities and the relationships are considered, functions that are part of a BEL statement are not taken into account on this level. Yet again, two levels of evaluation are considered. For a full score relationship, subject, object and the relationship type must be correct. For the secondary relationship level, partial relationships, containing two correct units out of three (subject, object and relationship type), are considered fulfilled. Finally, we evaluated how many BEL statements are entirely correct.

For task 2, up to 10 sentences for each BEL statements were accepted from the participating systems. Those statements were evaluated on two levels: In the 'fully supportive level', the sentence must contain all necessary information for a biologist to create the BEL statement. In the 'partially supportive level', the sentence is correct, when context information from the paper is taken into account. For more detailed information on the evaluation criteria, we refer to Rinaldi et al. (2).

III. Training Data and Preparation of New Test Set

The training data and test data from BioCreative V (2015) is available at the datasets page[7]. The description of the training set selection and annotation are described in detail in Fluck et al. (3). For the generation of the 2017 BEL track task 1 test set, we decided to use a real-world use case and extracted new data in the disease context of Ulcerative Colitis. For the test set, we restricted the named entity classes to those that can be normalized to the gene/protein namespaces HGNC and MGI, CHEBI for chemical names, MESH for disease names and GO for biological processes and a restricted set of relationship types and functions defined above. For the extraction, we used automatic support in form of the BELIEF workflow (6). This workflow pre-annotates and normalizes the named entities and suggests BEL statements in a user-friendly curation environment. In the frontend, the curator can browse through the text, search for unrecognized named entities and edit or add statements. Only user-selected statements are exported by the system.

In a first step, two curators independently extracted the information from the same full texts. In the comparison of results, it became clear that it is very tedious to extract all possible statements for a document and, in addition, we would create very similar statements only with different experimental settings, which are irrelevant for the task 1. It was also not feasible that independent curators select the same sentences for curation in full text. Therefore, we decided towards a more straightforward approach. One person selected the relevant sentences and extracted all BEL statements from this sentence. The second curator analyzed and edited this set. Finally, differences were discussed in an annotation jamboree. In Table I, an overview about the number of different statements, articles, entities and relationship types and functions of the task

[6] http://wiki.openbel.org/display/BIOC/All+Functions+Evaluation+Overview

[7] https://wiki.openbel.org/display/BIOC/Datasets

1 test set are given. For task 2 BEL statements were curated from abstracts to make sure that at least one sentence could be found for every BEL statement.

TABLE I. DISTRIBUTION OF TERM, FUNCTION AND RELATIONSHIP TYPES IN THE TRAINING AND TEST SETS

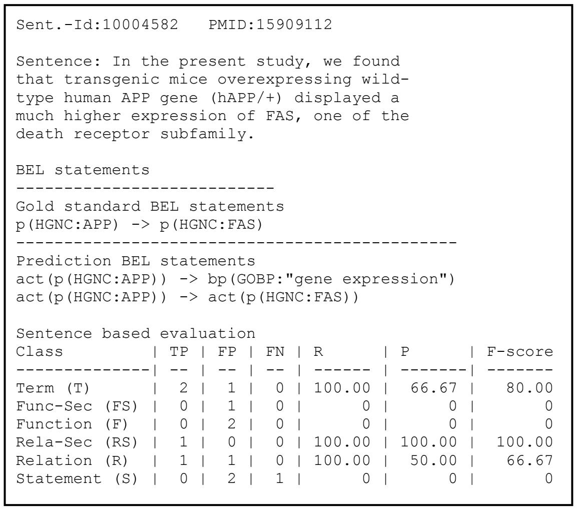
Type	*Training*	*Test 2015*	*Test 2017*
Terms			
p()	19.918	346	328
a()	1.927	37	52
bp()	877	31	23
path()	244	15	2
Functions			
act()	6.332	36	79
pmod()	1.411	9	36
complex()	750	15	5
tloc()	406	13	10
deg()	205	6	4
sub()	23	0	0
trunc()	6	0	0
Relations			
increases	8.112	155	130
decreases	2.956	53	68

IV. SUPPORTING RESOURCES

The participants were provided with a range of supporting resources and a comprehensive documentation[8], containing a description of the format and detailed explanation of the evaluation process. The evaluation on the different levels of a single BEL statement was illustrated using a set of concrete example submissions as reference. Additionally, a validation and an evaluation interface[9] was provided for the participants to validate and test their generated statements during the development phase. The BEL statement validator checks the user provided BEL statements with respect to formal correctness. It provides specific error messages for invalid BEL statements. For sample, training and 2015 test set, the evaluation interface evaluates the input BEL statements based on the evaluation criteria such as term, function, relationship and full statement level. An example of a candidate evaluation is shown in Figure 2.

Further supporting resources included the BEL statements from the training, sample and 2015 test set in BioC format, which we generated automatically using a converter based on the official ruby-based BEL parser[10] and an open-source BioC ruby module[11]. A tab-separated format that contains all fragments of the BEL statements (terms, functions and relations) was automatically generated from the sample, training and 2015 test set, using the same BEL parser mentioned above. These were provided to the participants as supporting material (c.f. (2)).

For task 1 stage 1, we also provided the normalized names of all biological processes that occur in the test set as extracting such concepts is still a non-trivial task. Finally, for task 1 stage 2, we provided a file with entity information with offsets and the associated normalized concept with the namespace.



```
Sent.-Id:10004582   PMID:15909112

Sentence: In the present study, we found
that transgenic mice overexpressing wild-
type human APP gene (hAPP/+) displayed a
much higher expression of FAS, one of the
death receptor subfamily.

BEL statements
---------------------------
Gold standard BEL statements
p(HGNC:APP) -> p(HGNC:FAS)
----------------------------------------------
Prediction BEL statements
act(p(HGNC:APP)) -> bp(GOBP:"gene expression")
act(p(HGNC:APP)) -> act(p(HGNC:FAS))

Sentence based evaluation
Class          | TP | FP | FN | R      | P      | F-score
--------------| -- | -- | -- | ------ | -------| -------
Term (T)       |  2 |  1 |  0 | 100.00 |  66.67 |   80.00
Func-Sec (FS)  |  0 |  1 |  0 |      0 |      0 |       0
Function (F)   |  0 |  2 |  0 |      0 |      0 |       0
Rela-Sec (RS)  |  1 |  0 |  0 | 100.00 | 100.00 |  100.00
Relation (R)   |  1 |  1 |  0 | 100.00 |  50.00 |   66.67
Statement (S)  |  0 |  2 |  1 |      0 |      0 |       0
```

Fig. 2 An example of a candidate evaluation. The example shows the candidate sentence, the gold standard and predicted statements. For all primary and secondary levels the scores are provided (2). Abbreviations: PMID (PubMed identifier), true positive (TP), false positive (FP), false negative (FN), recall (R), precision (P). Adapted and reprinted with permission from Fluck et al. (7).

V. RESULTS

A. *Task 1: Given textual evidence for a BEL statement, generate the corresponding BEL statement.*

Four teams contributed results of their information extraction systems for task 1. A maximum number of three submissions was permitted. Table II shows the results for the task 1 in stage 1. Here the teams had to provide their own term recognition. The results are color-coded in shades of green according to the values of F-score (F), the main evaluation criterion, and supplemented by the values for precision (P) and recall (R). The best results for each evaluation metrics are marked up in bold script. In general, all teams took part on all structural levels except team 390, which excluded the function level. Team 379 and 411 already took part in the BioCreative V BEL track task 1 (2015).

For the full statement level, the best system with the team id 379 achieved an F-measure of 32%. This illustrates the difficulty of this highly structured prediction task. The teams 411 and 390 had a similar performance, although their results were quite different on other evaluation levels, e.g. the term level.

[8] https://wiki.openbel.org/display/BIOC/BioCreative+BEL+Task+Challenges

[9] http://bio-eval.scai.fraunhofer.de/cgi-bin/General_server.rc

[10] https://github.com/OpenBEL/bel.rb

[11] https://github.com/dongseop/simple_bioc

TABLE II. EVALUATION OF STAGE 1 OF TASK 1 (PREDICTION OF BEL STATEMENTS WITHOUT GOLD STANDARD ENTITIES)

Team Id	Run	Terms			Function			Function Second.			Relation			Relation Second.			Statement		
		F	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
411	r1	63.24	84.62	50.49	33.99	44.83	27.37	51.24	67.39	41.33	40.22	55.38	31.58	62.92	88.19	48.91	22.99	33.33	17.54
	r2	57.75	81.93	44.59	31.08	43.4	24.21	38.67	69.05	38.67	36.78	53.33	28.07	57.73	86.84	43.23	20.71	31.82	15.35
	r3	61.24	88.27	46.89	32.88	47.06	25.26	46.15	64.29	36	37.43	56.14	28.07	62.03	92.24	46.72	21.15	33.98	15.35
380	r1	50.88	76.82	38.03	6	60	3.16	7.5	60	4	16.77	31.71	11.4	45.14	80	31.44	7.38	15.71	4.82
	r2	55.29	81.01	41.97	6.06	75	3.16	7.59	75	4	21.52	38.64	14.91	51.06	84	36.68	10.67	22.22	7.02
	r3	67.83	72.22	63.93	20.17	50	12.63	31.25	71.43	20	24.69	28.25	21.93	62.25	70.95	55.46	10.44	12.9	8.77
379	r1	74.14	78.18	70.49	**40.54**	56.6	31.58	**55.28**	70.83	45.33	43.65	51.81	37.72	86.17	89.62	82.97	32.28	40.67	26.75
	r2	72.89	78.71	67.87	40.29	63.64	29.47	54.39	79.49	41.33	**43.77**	52.12	37.72	**86.71**	93	81.22	**32.45**	41.22	26.75
390	r1	**76.39**	81.18	72.13	0	0	0	0	0	0	29.87	25.55	35.96	65.19	60.45	70.74	18.08	16.1	20.61
	r2	**76.39**	81.18	72.13	0	0	0	0	0	0	28.92	24.19	35.96	65.23	59.29	72.49	17.88	15.53	21.05

TABLE III. EVALUATION OF STAGE 2 OF TASK 1 (PREDICTION OF BEL STATEMENTS WITH GOLD STANDARD ENTITIES)

Team Id	Run	Terms			Function			Function Second.			Relation			Relation Second.			Statement		
		F	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
411	r1	83.93	99.11	72.79	36.36	47.46	29.47	46.77	59.18	38.67	57.22	73.29	46.93	83.33	98.8	72.05	31.30	46.15	23.68
	r2	86.09	99.15	76.07	40.51	50.79	33.68	51.16	61.11	44	56.08	70.67	46.49	83.92	98.82	72.93	30.95	44.63	23.68
	r3	85.45	99.13	75.08	39.24	49.21	32.63	50	60.38	42.67	57.6	73.47	47.37	83.63	98.81	72.49	31.79	46.61	24.12
380	r1	90.20	98.83	82.95	12.39	38.89	7.37	21.74	58.82	13.33	42.52	52.94	35.53	84.24	96.61	74.67	22.66	32	17.54
379	r1	87.65	90.56	84.92	51.75	77.08	38.95	57.63	79.07	45.33	**66.83**	76.7	59.21	**92.06**	95.75	88.65	49.2	63.01	40.35
	r2	86.4	90.94	82.3	**52.55**	85.71	37.89	**58.93**	89.19	44	**66.83**	76.7	59.21	91.92	97.55	86.9	**49.6**	64.34	40.35
	r3	87.41	90.81	84.26	51.75	77.08	38.95	57.63	79.07	45.33	**66.83**	76.7	59.21	91.99	96.63	87.77	49.2	63.01	40.35
390	r1	**91.33**	99.23	84.59	0	0	0	0	0	0	43.51	41.6	45.61	86.36	90.05	82.97	23.61	25	22.37
	r2	88.36	99.18	79.67	0	0	0	0	0	0	42.47	44.29	40.79	83.41	91.19	76.86	24.06	28.07	21.05
	r3	76.71	98.96	62.62	0	0	0	0	0	0	25.68	29.38	22.81	71.76	85.98	61.57	15.06	18.47	12.72

On the secondary relation level, it is sufficient that only two out of the three components of relationship (subject, object, relation type) are correct for the statement to count as a positive match. This level is introduced in order to give credit to results, which although partially correct, could still be useful in the context of a semi-automated approach as suggestions to a human curator. Here, up to 87% F-score were achieved, which is quite impressive. Whereas on the relation level, the highest F-measure was around 43%. This shows a drop of around 40% in comparison to the secondary level.

Furthermore, the comparison of all teams shows that most of the teams (except team 390) built systems that focus more on precision rather than recall. High scores on the relation level do not necessarily correlate with high scores on the full statement level. This is due to the fact that full statement level combines all structural levels.

The results for task 1 stage 2 are shown in Table III. In this stage, the gold standard concepts together with their specific text spans were made available to the teams. All teams could significantly benefit and improve on the level of full statements, which shows the importance of high-quality term recognition for further higher-level recognition tasks. Team 379 reached the highest F-score of 49.6% with the provided terms. In comparison to stage1, the score was increased up to 17% on the same test set. Similar performance increases can be seen on the function and relation level, too.

B. *Task 2: Given a BEL statement, provide at most 10 additional evidence sentences.*

For this task only one team participated. The team asked the organizers whether they can choose a different setting for the submission. In agreement with the organizers, two runs with two different configurations and only 5 ranked sentences for each run were submitted. The correctness of the provided evidence sentences was evaluated manually and rated on two different levels of strictness:

1. Fully supportive: Relationship is fully expressed in the sentence.
2. Partially supportive: Relationship can be extracted from the sentence if context sentences or biological background knowledge are taken into account.

To evaluate the quality of the curation results, we calculated an inter-annotator agreement. To tackle this task part of the manual curation was carried out by two different curators. For 150 entries, we observed a high agreement of 93% (kappa statistic: 0.75) and 91% (kappa statistic: 0.79) for the categories fully and partially supportive, respectively.

TABLE IV. EVALUATION RESULTS OF TASK 2 INCLUDING MEAN AVERAGE PRECISION (MAP)

Runs	Criterion	TP	FP	Precision	MAP
Run 1	Full	117	265	30.6%	59.6%
	Partial	175	207	45.8%	77.5%
Run 2	Full	121	261	31.7%	50.2%
	Partial	192	190	50.3%	76.7%

As shown in Table IV the system provided 382 evidence sentences for 98 BEL statements in each run (mean 3.9 sentences per statement). In run 1 for 55 BEL statements, there was at least one entirely correct evidence sentence, for 71 statements at least one sentence meeting the partially supportive evaluation condition, and in run 2, 58 and 70 BEL statements satisfied the fully and partially supportive evaluation condition, respectively. Table IV also shows the detailed numbers for TP, FP and the resulting precision at the micro level. Around one third of all sentences fully expressed the desired relationship. In order to assess the ranking quality of the system, we computed the mean average precision (MAP). Although the first run has a slightly lower precision compared to the second run, the MAP is considerably higher, especially for full supportive sentences. Overall, based on the results and the low number of participants, task 2 seems to be as difficult as task 1.

ACKNOWLEDGMENT

We would like to thank Lisa Langnickel for her support in the generation of the test data and the evaluation of task 2 data.

REFERENCES

1. Slater, T. and Song, D. Saved by the BEL: ringing in a common language for the life sciences. 2012.
2. Rinaldi, F., Ellendorff, T. R., Madan, S., et al. (2016) BioCreative V track 4: a shared task for the extraction of causal network information using the Biological Expression Language, *Database*, 2016.
3. Fluck, J., Madan, S., Ansari, S., et al. (2016) Training and evaluation corpora for the extraction of causal relationships encoded in biological expression language (BEL)., *Database (Oxford).*, 2016.
4. Sartor, R. B. (2006) Mechanisms of Disease: pathogenesis of Crohn’s disease and ulcerative colitis, *Nat. Clin. Pract. Gastroenterol. Hepatol.*, 3, 390–407.
5. Kaistha, A. and Levine, J. (2014) Inflammatory bowel disease: the classic gastrointestinal autoimmune disease., *Curr. Probl. Pediatr. Adolesc. Health Care*, 44, 328–34.
6. Madan, S., Hodapp, S., Senger, P., et al. (2016) The BEL information extraction workflow (BELIEF): evaluation in the BioCreative V BEL and IAT track, *Database*, 2016, baw136.
7. Fluck, J., Madan, S., Ellendorff, T. R., et al. (2015) Track 4 Overview: Extraction of Causal Network Information in Biological Expression Language (BEL), *Proc. Fifth BioCreative Chall. Eval. Work.*