BELIEF: a semi-automated curation tool to build mechanistic causal biological knowledgebase from unstructured scientific information

Justyna Szostak[1], Sam Ansari[1], Marja Talikka[1], Juliane Fluck[2], Sumit Madan[2], Florian Martin[1], Manuel C. Peitsch[1], Julia Hoeng[1]

[1] Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland

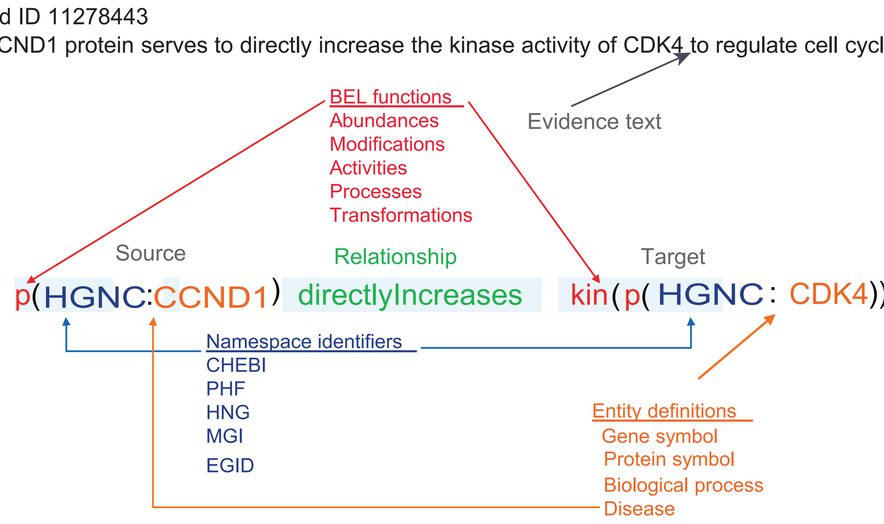
[2] Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Introduction

Ever-increasing scientific literature enhances our understanding of how toxicants impact biological systems, and there is an increasing demand from systems biologists/toxicologists to have access to the existing knowledge in a structured, and preferably, computable format (1-2).
Knowledge curation into computable format requires a well-defined structured and standardized language and entity/relationship recognition software for efficient knowledge retrieval. We recently introduced the BEL Information Extraction workFlow, (BELIEF) (3). BELIEF automatically extracts biological entities and causal relationships from text and converts them into a computable language, the biological expression language (BEL), a machine-and human-readable language that codes molecular relationships as semantic triplets: subject–relationship–object (www.openbel.org). BELIEF also allows human review and correction of the proposed statements.
Here we show how the semi-automated curation workflow employing BELIEF facilitates the construction of causal biological network models describing disease specific processes and an efficient, objective and specific interpretation of molecular data.

Biological Expression Language (BEL)

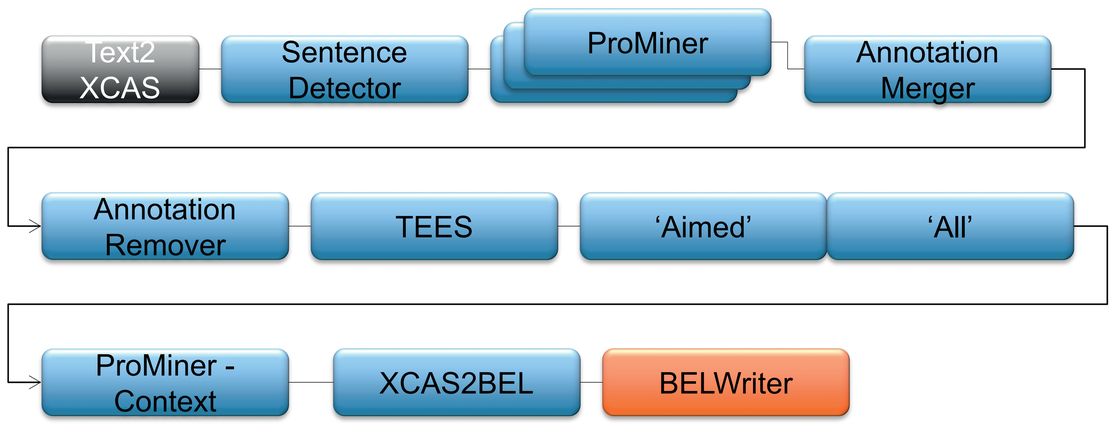
PubMed ID 11278443
"The CCND1 protein serves to directly increase the kinase activity of CDK4 to regulate cell cycle progression."



http://www.openbel.org

BEL (Biological expression Language) represents scientific findings in a computable format and captures experimental contex (species, disease, cell, cell line, tissue).
Nodes represent biological molecular mechanisms (ex."kinase activity of AKT1"), while edges encode signed causal relationships between the nodes.

Text mining pipeline

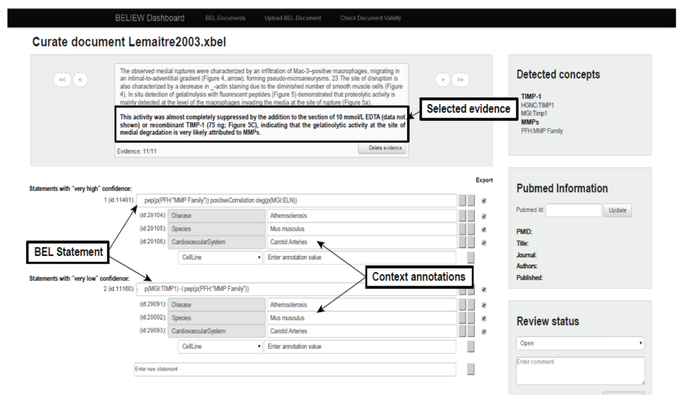


The pipeline consists of several linguistic modules that start with the conversion for the UIMA environment and includes named entity recognitions across 12 BEL compliant dictionaries as well as modules for annotation normalization and machine learning based relationship entity recognitions. The pipeline ends with a BEL writer module to create BEL compliant output.

http://belief.scai.fraunhofer.de/BeliefDashboard/

The BEL Information Extraction workFlow (BELIEF)

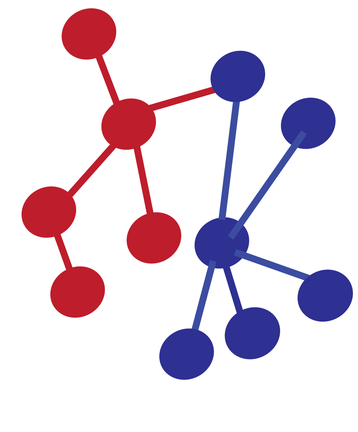
Scientific Literature | Text Mining pipeline | Curation Interface in the BELIEF dashboard | Transformation of scientific evidence into BEL statements | Computable knowledge



Causal relationships extracted from literature are written into the BEL:

"The Mouse protein with the MGI name IL18 increases the peptidase activity of the Mouse protein with the MGI name Mmp13"

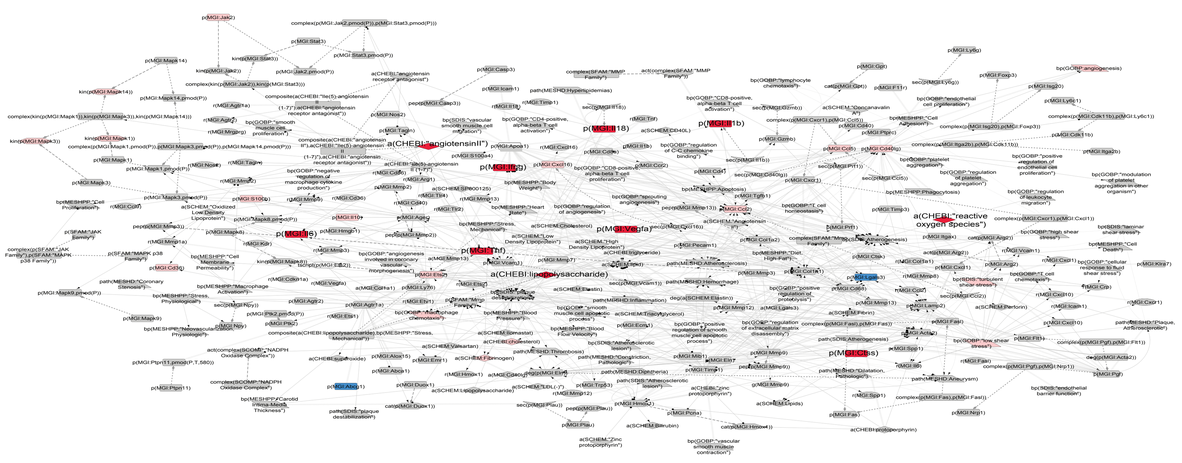
p(MGI:IL18) ---------> pep(p(MGI:Mmp13))



This schema describes the creation of computable knowledge from scientific literature to the biological mechanistic network model. The workflow is initiated with the selection of scientific articles and submission to the text mining pipeline where all text documents are processed and biological entities and relationships are recognized, extracted and assembled. The curation interface within the knowledge extraction pipeline gives access to the assembled entities that are coded in a BEL statement. The curation interface assists curators in the review of knowledge triplets (statements). The extracted causal relationships are then compiled into a mechanistic network model. The mechanistic network model represents molecular interactions accompanied with contextual information about the experiments.

Atherosclerosis Plaque Destabilization network model

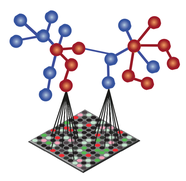
Graphical representation of the Atherosclerosis Plaque Destabilization network model built using BELIEF



The Atherosclerosis Plaque Destabilization network model contains 303 Nodes and 795 Edges.The nodes of the networks correspond to molecular biological entities (e.g., protein abundances, activities, and biological processes). Network edges connect two nodes and represent the cause-and-effect relationship between the corresponding entities. Compiled BEL statements represent knowledge in the graphical view.
33 full-text articles were selected and processed with the assisted curation pipeline. Three of the most connected nodes indicate biological processes of "plaque destabilization," "atherogenesis," and "positive regulation of smooth muscle cell apoptosis," all of which closely reveal the context that was modelled using the network.

Network scoring

NPA: Network Perturbation Amplitude



Find x that minimizes $\sum_{(i,j)}^{edges} \omega(i,j)\cdot(x_i - sign(i,j)\cdot x_j)^2$ subject to $x|_{transcript\ layer} = \beta$ where $\omega(i,j)$ gives the edge weights.

Evaluate $\sum_{(i,j)}^{edges} (x_i + sign(i,j)\cdot x_j)^2$ where the sum includes all the edges of the functional layer.

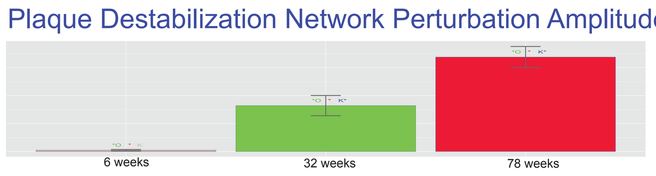
The NPA algorithm scores the transcriptomic data against the network model using backward reasoning similar to RCR. It integrates gene expression data with a two-layer cause-and-effect network by using its full topology to identify, interpret and quantify the perturbation of the network in response to any treatment (4).

Dataset description

The dataset GSE10000 (5) was used to paint and score the network. It contains gene expression measurements from the aortas of wild-type and ApoE$^{-/-}$ mice at 6, 32 and 78 weeks of age.

Data Set	Reference	Species	Title	Platform	Tissue	Regime	Perturbation
GSE 10000	PMC2626665	Mouse	Lymphotoxin beta receptor signaling promotes tertiary lymphoid organogenesis in the aorta adventitia of aged ApoE$^{-/-}$ mice.	Affymetrix Mouse Genome 430A 2.0 Array	Aorta	Standart chow	Age 6/32/78 weeks

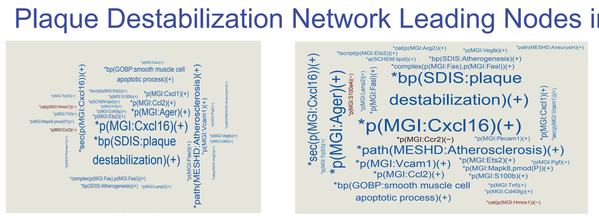
Plaque Destabilization Network Perturbation Amplitude Score in aortas of ApoE-/- vs. wt mice



The network is perturbed in a time -dependent manner: the network perturbation amplitude (NPA) scores for the Atherosclerosis Plaque Destabilization network model show an increase of the network activation in ApoE-/- mice at 32 and 78 weeks of age when compared with wt.

Symbol legend:*: O and K statistic p-values below 0.05 (in color), .O and K. p-values between 0.05 and 0.1(in grey).

Plaque Destabilization Network Leading Nodes in aortas of ApoE-/- vs. wt mice



The NPA allowed identification of the key contributors to the perturbation, referred to as leading nodes. Leading nodes for the Atherosclerosis Plaque Destabilization network model are shown with a "*", for contrasts where the associated NPA is significant (Cxcl16, Ager and plaque destabilization, smooth muscle cell apoptosis).

Conclusions

BELIEF is a semi-automated curation workflow that combines text mining with a user friendly curation interphase allowing the extraction of causal molecular relationships from scientific literature.

The standardized language used by BELIEF enables the conversion of unstructured knowledge into computable biological network models.

Combined with backward reasoning algorithms, the network models built using BELIEF can be used to extract biological significance from noise for quantitative impact assessment.

References

1. Hoeng J, Deehan R, Pratt D, Martin F, Sewer A, Thomson TM et al. A network-based approach to quantifying the impact of biologically active substances. Drug discovery today. 2012;17(9-10):413-8. doi:10.1016/j.drudis.2011.11.008.
2. Hoeng J, Talikka M, Martin F, Sewer A, Yang X, Iskandar A et al. Case study: the role of mechanistic network models in systems toxicology. Drug discovery today. 2014;19(2):183-92. doi:10.1016/j.drudis.2013.07.023.
3. Szostak J, Ansari S, Madan S, Fluck J, Talikka M, Iskandar A et al. Construction of biological networks from unstructured information based on a semi-automated curation workflow. Database : the journal of biological databases and curation. 2015;2015:bav057. doi:10.1093/database/bav057.
4. Martin F, Sewer A, Talikka M, Xiang Y, Hoeng J, Peitsch MC. Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. BMC bioinformatics. 2014;15:238. doi:10.1186/1471-2105-15-238.
5. Grabner R, Lotzer K, Dopping S, Hildner M, Radke D, et al. (2009) Lymphotoxin beta receptor signaling promotes tertiary lymphoid organogenesis in the aorta adventitia of aged ApoE-/- mice. J Exp Med 206: 233-248.
2014;15:238. doi:10.1186/1471-2105-15-238.

PMI RESEARCH & DEVELOPMENT

This work was funded by Philip Morris International

Systems Toxicology 2016
Real World Applications and Opportunities
January 27-29, 2016