Improved Knowledge Acquisition and Creation of Structured Knowledge for Systems Toxicology

Sam Ansari, Justyna Szostak, Marja Talikka, Julia Hoeng
Research & Development
Philip Morris Products SA (part of PMI)
Neuchatel, Switzerland
sam.ansari@pmi.com, justyna.szostak@pmi.com, marja.talikka@pmi.com, julia.hoeng@pmi.com

Juliane Fluck
Fraunhofer Institute for Algorithms and Scientific Computing
Sankt Augustin, Germany
juliane.fluck@scai.fraunhofer.de

***Abstract*—Ever-increasing scientific literature enhances our understanding on how toxicants impact biological systems. In order to utilize this information in the growing field of systems toxicology, the published data must be transformed into a structured format suitable for knowledge modelling, reasoning, and ultimately high throughput data analysis and interpretation. Consequently, there is an increasing demand from systems toxicologists to access such knowledge in a computable format, here biological network models. The BEL Information Extraction workFlow (BELIEF) automatically extracts biological entities and causal relationships from any text resource and converts them into a formalized language, the Biological Expression Language (BEL). BEL is a machine- and human-readable language that represents molecular relationships and events as semantic triples: subject–relationship–object. In addition to the automatic extraction through text mining, BELIEF also features a curation interface to verify and modify the proposed triples and benefits from BEL's human-readability. The curation interface facilitates this curation task by providing relevant information to ensure high curation accuracy and fast processing. The resulting BEL triples are then assembled to biological network models that represent specific biological processes for a given context, e.g., organism, tissue type, disease state. These biological network models can then be verified in a crowd-based approach utilizing a collaborative web-based platform before finally sharing them through a publicly available and specialized repository. In this strategy paper, we summarize over various solutions to challenges in the knowledge-based systems toxicological assessment.**

Keywords-component; text mining; BEL; knowledge management; network models; curation; systems toxicology.

I. Introduction to Systems Toxicology

Systems toxicology supports the detailed understanding of the mechanisms by which biological systems respond to toxicants. This understanding can be used to assess the risk of chemicals, drugs or consumer products. In this work, we give a summary on critical developments within the biological, toxicological, as well as computational domain that has led to this knowledge-based systems toxicological assessment approach [1].

A. Omics Profiling in Systems Toxicology, from Basic Research to Risk Assessment

Towards the end of the second millennium a new field emerged in toxicological assessment when the high throughput analysis of the transcriptome called transcriptomics was first used to identify the cellular components and signaling pathways involved in toxicity response and in relationship to harm and disease [2][3]. This approach was called Omics profiling. Omics is the name of a group of biological fields, such as genomics for the discipline in genetics, proteomics for the study of proteins, metabolomics for the study of chemical processes involving metabolite (and the related field of lipidomics for lipids), transcriptomics, and others. This addition of Omics profiling to toxicological assessments opened new possibilities to better understand how different compounds cluster into similar mechanistic classes based on the molecular response profile they inflict on the test system. It has also enabled the discovery and validation of exposure-response biomarkers, as well as the classification and ranking of drug candidates [4]. Omics profiling also guides the development of new and more precise toxicological endpoints and targeted cellular assays [5] and can be valuable in the approximation of the lowest dose that results in the perturbation of the system, especially when data are sparse and when the toxicant affects more than a single pathway [6][7].

Initiatives, such as the ToxCast [8], Adverse Outcome Pathways (AOPs) [9], FutureTox [10], and the Comparative Omics profiling Database (CTD; http://ctdbase.org/) [11], share the common goal to turn the sole use of apical endpoints into predictive toxicology by gaining better understanding of toxicological mechanisms [12]. The

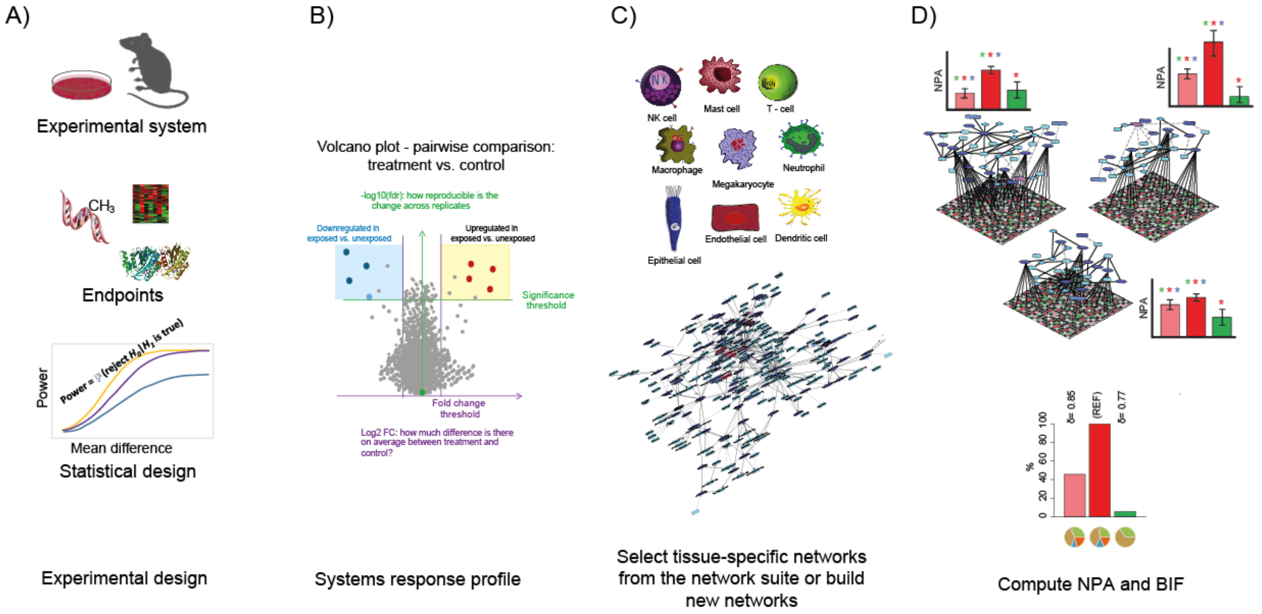

A)
Experimental system
CH3
Endpoints
Power
Power = P(reject H0|H1 is true)
Mean difference
Statistical design
Experimental design
B)
Volcano plot - pairwise comparison: treatment vs. control
-log10(fdr): how reproducible is the change across replicates
Downregulated in exposed vs. unexposed
Upregulated in exposed vs. unexposed
Significance threshold
Fold change threshold
Log2 FC: how much difference is there on average between treatment and control?
Systems response profile
C)
NK
NK cell
Mast cell
T - cell
Macrophage
Megakaryocyte
Neutrophil
Epithelial cell
Endothelial cell
Dendritic cell
Select tissue-specific networks from the network suite or build new networks
D)
NPA
***
*
%
0
20
40
60
80
100
δ= 0.85
(REF)
δ= 0.77
Compute NPA and BIF

Figure 1. Overview of the Systems Toxicology Workflow. A) The workflow starts with a careful selection of the experimental system and robust statistical design based on the choice of endpoints to be measured. B) Common endpoints in systems toxicology are the high throughput Omics measurements. Transcriptomics data can be analysed to obtain the systems response profiles triggered by the exposure. C) The transcriptomics data is often analysed in the context of causal biological network models for mechanistic interpretation and the models have to be carefully chosen to align with the biological context that the experiment has been conducted in. D) Finally, sophisticated algorithms are used to compute the Network Perturbation Amplitude (NPA) for each network and the aggregated overall Biological Impact Factor (BIF).

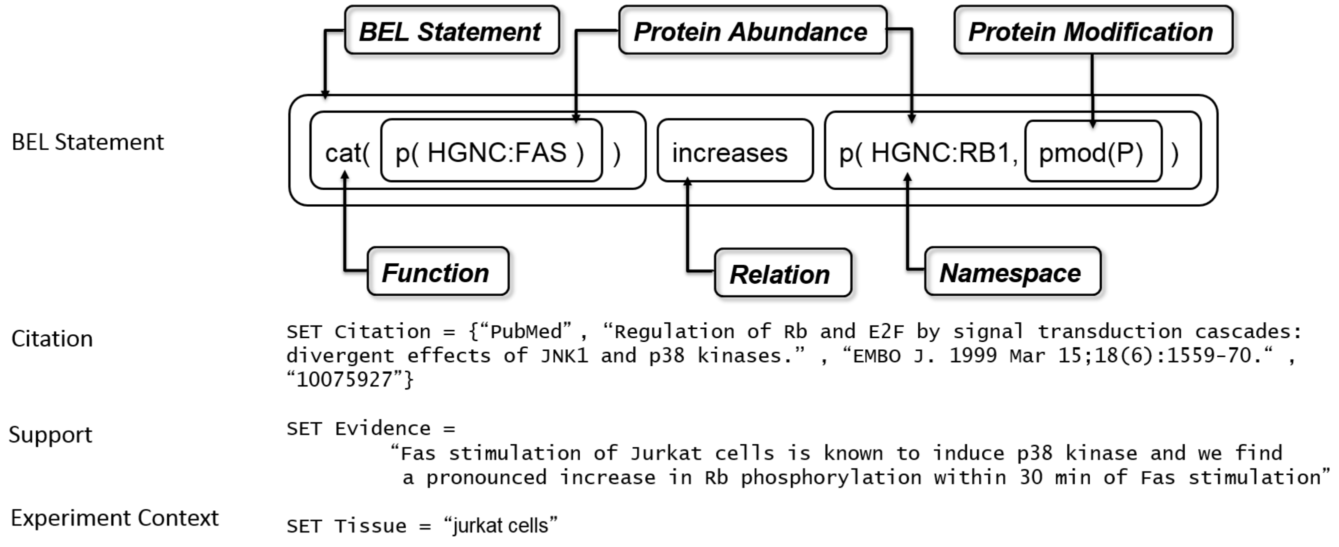
ultimate goal of the toxicity testing in the 21st century is to use the insight gained from in vitro assays to predict adverse effects observed in laboratory animals and humans [13].

The comprehensive integration of classic toxicology approaches and Omics profiling forms the basis of systems toxicology. There is also strong emphasis on the development of computational platforms to enable quantitative analysis of molecular changes in response to a stressor and to accurately model the toxicological system [14].

B. *Knowledge Modelling and Computational Analysis as Core Instruments in Systems Toxicology*

There are a number of ways to extract meaningful signals from high throughput measurements involving a variety of suited software solutions on top of sophisticated laboratory instruments. At the same time the data generated by these high throughput methods creates a challenge in the analysis and interpretation with traditional data processing applications. Different pathway tools provide ways to analyze high throughput data, mainly differentially expressed genes to highlight the biological processes, in which the genes are known to function. These partially web-based tools that include the DAVID Bioinformatics Database [15] or the Gene Set Enrichment Analysis (GSEA) method [16] employ pathway repositories, such as Reactome [17], Biocarta, [18], and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [19] that enable the mapping of the regulated genes into pathways. Fortunately, as the number of datasets increases, more literature that describes biological relationships is published. Using this knowledge, the creation of causal relationships from scientific evidences is becoming an efficient and popular methodology to analyze molecular data. Such models allow the interpretation of data in the context of directional graphs with signed interactions (edges) between biological entities (nodes) [20][21][22][23], providing the network perspective for the stressor response [24].

Causal biological network models are also the cornerstone of the workflow for impact assessment that we have developed over the past years [25][26]. The workflow starts with the design of appropriate experiments for data production. This includes the choice of exposure regimen, experimental test system, and the selection of measurements that will be made (see Figure 1). In addition to apical endpoints, transcriptomics profiling is almost always included to allow mechanistic interpretation of exposure effects. After conducting rigorous quality controls and statistical analyses, the gene expression changes are converted into a systems response profile that illustrates how the level of each molecular entity (here mRNA) is changed in response to the exposure. Finally, the transcriptomics data are analysed in the context of causal biological network models that summarize the *a priori*



Citation

```
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”}
```

Support

```
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”
```

Experiment Context

```
SET Tissue = “jurkat cells”
```

Figure 2. BEL Nanopub Overview. The three crucial elements of a BEL nanopub are the BEL Statement showing the knowledge statement in a triple and controlled terminology, as well as the citation information and actual evidence sentence. Experiment context is an optional field to simplify the triple assembly into biological network models. BEL nanopubs are coded in XML.

knowledge in a given context [25][26] (see Figure 1).

The network models, here causal network models, were built form scientific literature to reflect biological processes that are, e.g., assumed impacted in the lung and vascular system in response to cigarette smoke exposure [27][28]. The current suite of models consists of network families describing cell proliferation [29], cell fates [30], cell stress [31], pulmonary inflammation [32], tissue repair and vascular inflammation [33].

The causal network models are used in combination with transcriptomics data and computational algorithms that transform gene expression changes into Network Perturbation Amplitude (NPA) scores and the aggregated Biological Impact Factor (BIF) [34][35]. Such calculation requires the network model to contain a measurable layer reflecting gene regulations by some of the entities in the model backbone. The transcriptomics data are used to infer the activities of the backbone nodes based on the regulation of their target gene in the dataset.

These inferred changes in the backbone node activities are evaluated in the context of the overall network topology; the NPA score depicts the predicted effect that the exposure has on the biology described by the network model [26][34][35]. In some cases it is beneficial to get an overview of the overall biological impact that a stressor elicits on the test system. The BIF is an aggregation of the NPA scores stemming from perturbation of the individual biological processes included in the network model suite [35]. The advantage of computing a single holistic score is that it allows a high-level comparison of biological effects resulting from different exposures. Several use cases employing this approach have been published [36][37][38][39][40][41].

The following sections of this strategy paper will summarize the key modules in the implementation of this knowledge-based toxicological assessment.

II. Knowledge Acquisition for Knowledge Modelling

In this section, we describe the toolset and implementation strategy for knowledge acquisition and ultimately knowledge modelling in order to analyze and interpret systems toxicology data.

A. BEL and the BEL framework

Today, an overall accepted and widely spread exchange format for knowledge is through the unstructured text, natural language. Natural language contains many redundancies, uses varying vocabularies, introduces complication by using grammar and different sentence structures, as well as containing implications. All these factors make the unstructured text as such useless for knowledge management through computational tools. Our causal biological network models represent the aggregation of unstructured scientific knowledge formalized in the Biological Expression Language BEL [42]. BEL is a computer and human readable language specially designed to formalize biological relationships and allow the construction of network models to facilitate downstream computational analyses. While there are other conventions available that allow the formalization of unstructured biological knowledge, e.g., BioPAX [43] and SBML [44], BEL presents a considerable advantage that it is simple to read and edit by a biologist, because the formalization is close to natural language with simplification into a triple, here subject, predicate, and object as well as restriction of vocabulary via defined namespaces (see Figure 2). BEL conserves causal, e.g., increases and decreases, and non-causal correlative relationships, here positive correlation, negative correlation, and association. Based on Semantic-web technology, BEL uses a nanopublication model for publishing an assertion, together with attribution and provenance metadata [45] (Guidelines for Nanopublication http://nanopub.org/guidelines/working_draft/). A BEL nanopub is the smallest unit of information and represents a biological relationship with its provenance. Two elements are crucial, the BEL statement and the evidence, where evidence is the supporting text and citation information. Additionally, each BEL statement can be associated with experimental context information such as organism, organ, tissue, cell line, disease state and more. This context information is finally used to construct biological network models under specific experimental conditions.

In BEL subjects and objects are represented by a function of biological entities controlled by the BEL syntax and BEL terms that are managed in namespaces. A function can be the abundance of a particular biological entity, here chemical abundance *a()*, protein *p()*, genes *g()*, RNA *r()* or micro-RNA *mRNA()*. Biological process *bp()* or disease *path()* functions capture cellular parameters or processes in BEL. A detailed description of the BEL syntax and the use of BEL namespaces is described in detail on www.openbel.org.

BEL is accompanied by a set of tools packaged in the BEL framework (www.openbel.org). These tools allow the syntactic and semantic validation and compilation of single BEL triple into an assembled network model. The BEL framework also includes a knowledge assembly models managing software and a connector for Cytoscape network visualization software [46] in order to visualize and analyze the assembled networks in graph.

B. From Knowledge to BEL, the BEL Information Extraction workFlow (BELIEF)

An approach that is becoming more and more popular and that started back in the 80s is to either manually or automatically curate / parse and semantically annotate natural language word by word, sentence by sentence. The domain of text analytics with the help of linguistics was established and is increasingly developing tools and algorithms that better identify entities either via extended vocabularies or sophisticated statistical methods such as machine learning. At the same time, domain experts as well as curators and computational scientists focus on a way to define formats for a better and more applicable representation of knowledge. As it stands today, text mining either focuses on high recall and rather low precision, which is the case for most automated solutions, or on low recall but high precision, which is typically the case for manual curation.

The solution obviously lies in semi-automated knowledge extraction where linguistic tools identify relevant entities from natural language with a high recall and are manually curated for high precision.

Fluck et al. have addressed the challenge in efficiently extracting knowledge from the rich source of scientific articles by proposing a workflow combining both, the automated extraction method using text mining methodology and the manual curation of these results to ensure precision [47]. In the next two sections we will explain the challenges each methodology (manual versus automated) has and how in the third section both approaches can be combined into an efficient workflow, BELIEF.

1) Limitations of Manual Curation

Manual curation typically has the goal to bring unstructured text into structured data where various information sources are brought into one repository, here typically a database, and the data are annotated with controlled terminology. For a long time great effort has been made to build biological databases such as, e.g., UniProtKB/Swiss-Prot (www.uniprot.org/), MGI (www.informatics.jax.org/), HGNC (www.genenames.org/). The manual curation process is the dominant methodology for these efforts where a team of curators reviews literature and other knowledge sources for annotations given the specific context. These annotations are then stored in a structured format into databases [48][49][50]. However, one large source for variability even in these structured repositories is the variability from curator to curator defined by the experience the curator has and the effort the curator is taking. In fact, it was shown that expert curators present an accuracy of 90% for a specific task while the inter-curator agreement ranges from 77% to a minimum of 31% [51][52]. These results demonstrate the issues of this sophisticated yet very time consuming process that jeopardizes the quality and goal of these standardized repositories to some extent. Even when annotation guidelines are specified in order to create harmonization across different curators given a specific task, the personal variability is still high and impactful in the biological domain. Therefore, high-quality manual curation of the scientific literature is a very challenging and time-consuming effort and impacts the progress in the creation of these biological databases [53].

2) Limitations of Automated Curation

With the limitations of human curation, computational teams started focusing on automated curation processes to, in most cases, replace parts of the manual curation task [54]. Tools for named entity recognition (NER) for gene and protein name recognition are widely used within the database community. Typically tools such as Textpresso [55] and ProMiner [56] are employed to identify specific entity classes [57][58]. While the speed and output quantity in which automated annotations perform is impressive, the

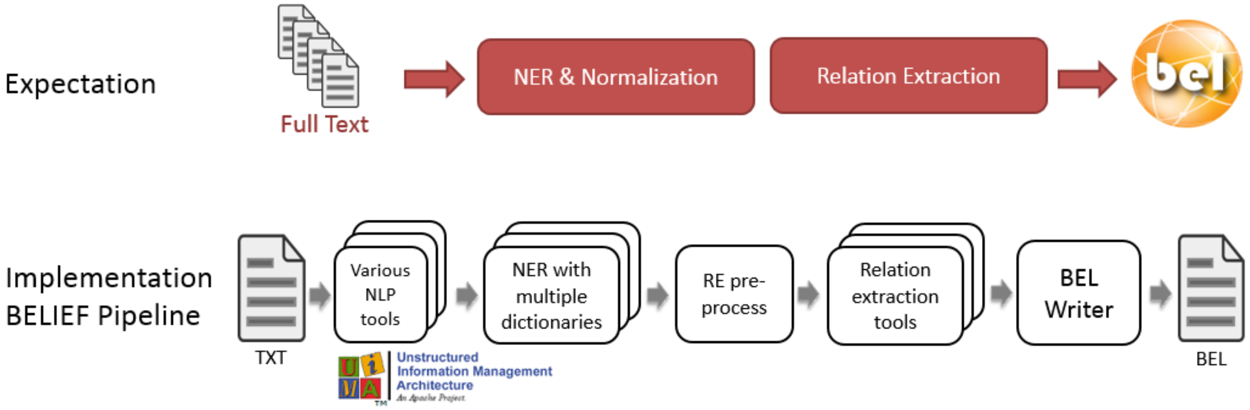


Figure 3. BELIEF Text Mining Pipeline Overview. The Expectation line shows a very high-level view on the functionality. In the row "Implementation BELIEF Pipeline" all relevant modules are shown. Various NLP tools are used for detecting and splitting sentences, identifying words etc. In the next step NER is used to detect relevant entities with given dictionaries, here namespaces. The relationships between these detected entities is captured in the next step and finally a BEL nanopub compliant output is generated.

results also have a greatly increased rate of errors compared to manual curation [53]. More specifically, it was shown that automated curation caused critical errors in assignments of particular functions that may affect as many as 30% of the proteins, and may even exceed 80% of individual protein families [48]. In a recent paper that modeled annotation errors in the Gene Ontology database, it was estimated that up to 49% of sequences functionally annotated by automatic sequence comparison methods could be mis-annotated [59]. Comparable errors were also observed in other analyses [48][60][61]. The introduction of errors in the public database could lead to severe error propagation that could make the data useless and even misleading when it comes to the interpretation of experimental data [48][62].

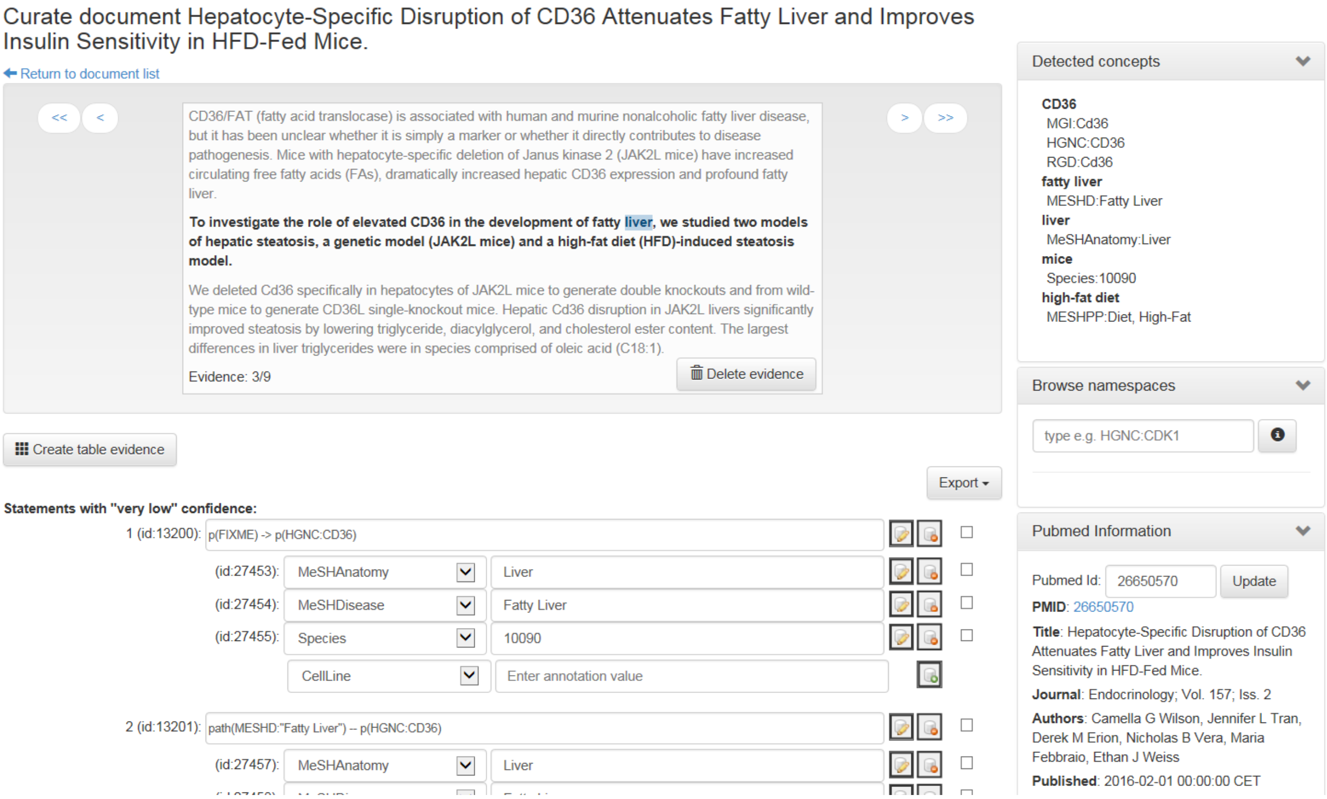
3) Solution and Performance of a Semi-automated Solution

Although text mining solutions did take over parts of the manual curation tasks, there was no approach shown before where a full knowledge statement was extracted and coded into a predefined syntax and become subject for manual curation. Looking at both approaches in annotating and extracting knowledge from unstructured text, the strength and weaknesses become obvious. The manual approach obviously results in much more reliable output at the cost of time, effort and harmonization / reproducibility. At the same time, automated annotation and extraction has a dramatically improved curation speed with full reproducibility but lacking precision. These strength and weaknesses are complementary and suggest a combined approach, here semi-automated approach. Especially in biological research and pathway modeling the identification of relationships between entities, e.g., protein-protein, drug-protein interactions or protein-disease relationships is crucial for mechanistic and network analyses. To be efficient, the automated curation process would have to be able to mimic the human ability to infer relations from the text. Text mining tools are currently not only able to detect and identify biological entities in the text, but they are also able to infer the relationships between these entities. The accuracy of text mining tools was estimated and demonstrated a high-performance of about 82-85% overall (Elsevier), 80 % for ProMiner for human and mouse gene/protein name recognition and about 50% for BioRat [63][64]. Altogether these evidences demonstrate that text mining is an efficient tool to curate unsolved amounts of data with a consistent quality for data detection and annotation.

In 2014 Fluck et al. released the BEL Information Extraction workFlow BELIEF (see Figure 3) [47]. The BELIEF infrastructure embeds an extraction information workflow combined with NER and relation extraction (RE) methods into a state of the art environment.

The combination of various linguistic tools into one workflow requires an extra effort in normalizing the results (see Figure 3).

BELIEF addresses the biological network model curation needs by identifying chemical, gene/protein, and biological process and disease terms in scientific articles. Additionally to that BELIEF identifies relationships through a combination of specialized ontologies and linguistics rules. On top of the BELIEF text mining pipeline sits the BELIEF Dashboard that provides users a manual curation interface for the automatically extracted BEL nanopubs (see Figure 4).



Curate document Hepatocyte-Specific Disruption of CD36 Attenuates Fatty Liver and Improves Insulin Sensitivity in HFD-Fed Mice.

CD36/FAT (fatty acid translocase) is associated with human and murine nonalcoholic fatty liver disease, but it has been unclear whether it is simply a marker or whether it directly contributes to disease pathogenesis. Mice with hepatocyte-specific deletion of Janus kinase 2 (JAK2L mice) have increased circulating free fatty acids (FAs), dramatically increased hepatic CD36 expression and profound fatty liver.

To investigate the role of elevated CD36 in the development of fatty liver, we studied two models of hepatic steatosis, a genetic model (JAK2L mice) and a high-fat diet (HFD)-induced steatosis model.

We deleted Cd36 specifically in hepatocytes of JAK2L mice to generate double knockouts and from wild-type mice to generate CD36L single-knockout mice. Hepatic Cd36 disruption in JAK2L livers significantly improved steatosis by lowering triglyceride, diacylglycerol, and cholesterol ester content. The largest differences in liver triglycerides were in species comprised of oleic acid (C18:1).

Evidence: 3/9

Statements with "very low" confidence:

1 (id:13200):	p(FIXME) -> p(HGNC:CD36)		□
(id:27453):	MeSHAnatomy	Liver	□
(id:27454):	MeSHDisease	Fatty Liver	□
(id:27455):	Species	10090	□
	CellLine	Enter annotation value	
2 (id:13201):	path(MESHD:"Fatty Liver") -- p(HGNC:CD36)		□
(id:27457):	MeSHAnatomy	Liver	□

Detected concepts

CD36
MGI:Cd36
HGNC:CD36
RGD:Cd36
fatty liver
MESHD:Fatty Liver
liver
MeSHAnatomy:Liver
mice
Species:10090
high-fat diet
MESHPP:Diet, High-Fat

Browse namespaces

type e.g. HGNC:CDK1

Pubmed Information

Pubmed Id: 26650570

PMID: 26650570

Title: Hepatocyte-Specific Disruption of CD36 Attenuates Fatty Liver and Improves Insulin Sensitivity in HFD-Fed Mice.

Journal: Endocrinology; Vol. 157; Iss. 2

Authors: Camella G Wilson, Jennifer L Tran, Derek M Erion, Nicholas B Vera, Maria Febbraio, Ethan J Weiss

Published: 2016-02-01 00:00:00 CET

Figure 4. Overview of the BELIEF Dashboard. The main window contains two sections: 1. the evidence text with the next sentence and browsing buttons, and 2. the automatically generated BEL nanopub with options for modifications and export. The right banner contains supporting information such as the concepts together with the namespace sources that were detected, a namespace browser to assign new concepts that were not detected, as well as the citation information that automatically retrieved all required information based on the Pubmed ID provided.

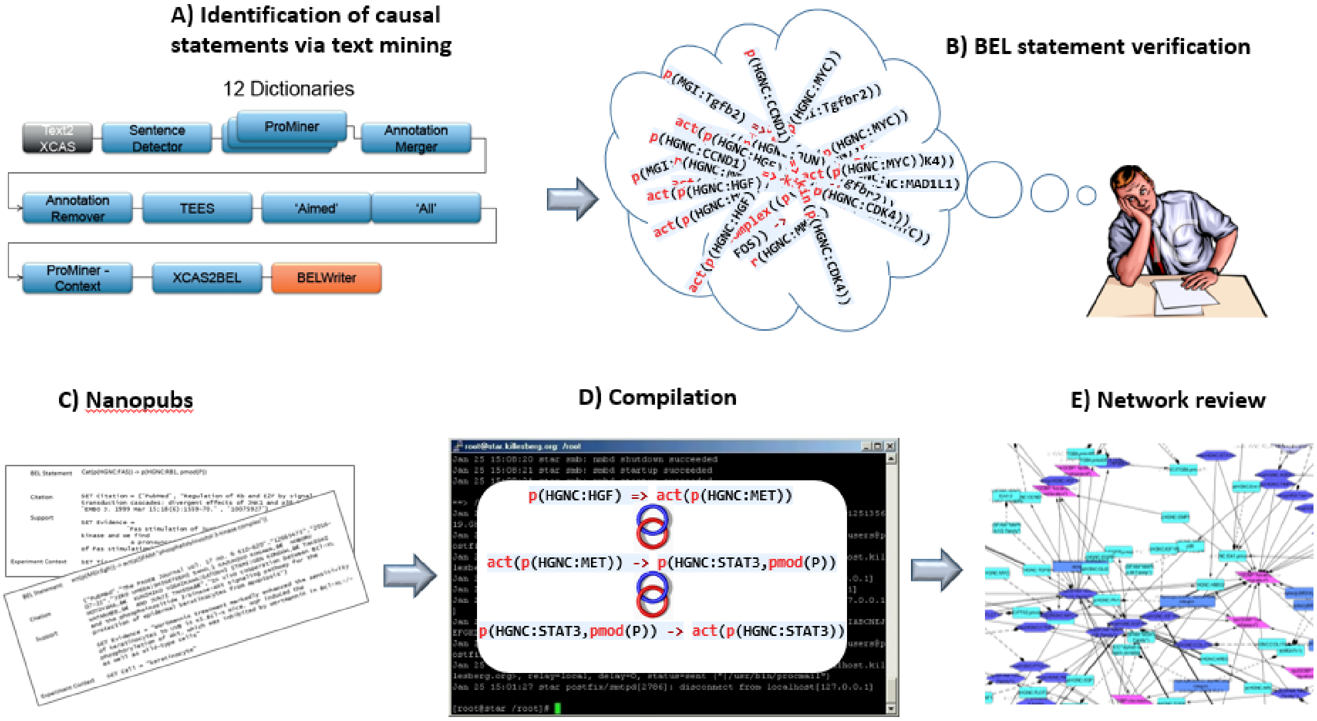
The dashboard offers the possibility to the curators to visualize, edit, correct, and delete statements to ensure precision on the high recall output from the underlying text mining pipeline. In an assessment Fluck et al. not only showed a higher detection rate of this combined curation approach in BELIEF but also a much higher user acceptance rating on the simplification of the curation effort [47].

III. Knowledge Assembly and Verification

With the creation of BEL nanopubs extracted from and curated in BELIEF, causal network models can be assembled using the BEL framework tools. These network models are typically assembled from BEL nanopubs given a specific context. After assembling, these networks are further assessed either with experimental data or additional knowledge sources, e.g., databases or other scientific articles. However, this verification can also be carried out in a crowd-based approach. Boué et al. developed a web-based platform for a collaborative verification of these causal network models [27]. In their publication the authors show the outcome of a community challenge called Network Verification Challenge (NVC) by using their platform and verifying 50 biological network models relevant to lung biology and early COPD. Each participant was given the opportunity to confirm, reject, or modify the networks on a website (https://bionet.sbvimprover.com/) and to add mechanistic detail [65]. The challenge showed that even for a group of domain experts unfamiliar with BEL, the crowd performed well at representing scientific findings in BEL. In a similar setup Fluck and Rinaldi performed a BEL task in the BioCreative V challenge. The goal of the challenge was to address curation challenges presented in BEL with text mining solutions. The outcome was that even for computerized systems, BEL did not bring a challenge in adapting algorithms and addressing the challenge tasks well [66].

IV. Knowledge Sharing

As previously stated, the collaborative approach in verifying the representation of literature-based knowledge proved most useful. To ensure continuity in this review approach, network models must be shared and available to the public domain to allow the community use, review, and further provide feedback. In fact, one of the strongest motivators for participants going through the verification process was the use of these networks for their own research projects.



```
p(HGNC:HGF) => act(p(HGNC:MET))
act(p(HGNC:MET)) -> p(HGNC:STAT3,pmod(P))
p(HGNC:STAT3,pmod(P)) -> act(p(HGNC:STAT3))
```

Figure 5. Process to build Causal Biological Network Models using BEL. A) The causal statements are identified in scientific literature and processed by text mining software within BELIEF. B) Domain expert biologists verify the statements using the curation dashboard in BELIEF. C) Semi-automated curation workflow gives rise to nanopubs that contain all essential information about the causal statements. D) The compilation process builds the isolated nanopubs into a coherent representation that can be E) visualized and verified using software such as Cytoscape.

However, the representation of network information in databases is not trivial. Boue et al. prepared an overview on network-based databases and their attributes [27]. In the same article, the authors present the Causal Biological Network database that is specialized for BEL triples and allows to query the data and find the right network model in the large number of available network models. Currently, the database contains biological network models that reflect causal signaling pathways across a wide range of biological processes, including cell fate, cell stress, cell proliferation, inflammation, tissue repair, and angiogenesis in the pulmonary and cardiovascular context. The database is openly accessed giving access to over 120 manually curated and well annotated biological network models. The database uses MongoDB that stores all network models and previous versions of each mode as JSON objects. With these objects users can query the database for genes, proteins, biological processes, small molecules, and keywords in the network descriptions in order to access the required network. On top of the database is a query and visualization layer that allows the users to browse the content and visualize the networks featuring filters for nodes and edges. A link to the supporting text in pubmed is available with each edge (http://causalbionet.com).

V. Future of Knowledge Modeling in Systems Toxicology

In the field of network toxicology, the current static models will eventually be replaced with dynamic models that can capture time and dose effects and provide better predictions on toxic outcomes [14]. This can be accomplished only by developing new and / or combining current modeling languages to handle differential equations that allow dynamic modelling in the context of a priori knowledge. There is also a need to invest in tools that can make conversions from one language to another (e.g., BEL to SBML) so that recorded knowledge is not syntax dependent and therefore limited in the toolset linked to the given syntax. There is still substantial work in enlarging the namespaces and controlled vocabulary to allow the curation of all species context. For instance, the zebrafish is a very attractive model system in modern toxicology research and it would be unfortunate if high throughput measurements from such species cannot be interpreted using causal biological network models. While the semi-automated curation workflow described here and in [67] is a major step towards efficient curation, it would further benefit from automated literature identification within a topic, and eventually the identified articles could be connected to the text mining tool, after which the exposed statements would be verified by curators. Such an approach would quickly pave the way to systems toxicologist's vision of a robust systems toxicology knowledgebase to keep up to date with the growing scientific knowledge. However, regardless how large, independent curation efforts cannot harness all the available information, leaving significant gaps in knowledge. One way to accomplish this is the education of researches about controlled vocabularies for expressing study results and enforcing a nanopub submission along with scientific manuscripts. In essence, a nanopub is the smallest unit of information and represents a biological relationship with its provenance derived from a publication. Ideally the entire content, including figures and tables with captions, could be represented in this format [68]. A community-driven approach the Concept Web Alliance has described guidelines for writing a nanopub, which has to consist of a statement, the origin of the statement, and the origin of the nanopub [45]. Even big datasets, often rather hidden in the supplemental parts, could be made more expedient when expressed in machine-readable formats with sufficient metadata on origin and context. Such approach has been tested in the assertion of differential gene expression in Huntington's disease [69], and the Open PHACTS Discovery Platform provides a guideline for precompetitive nanopub creation and outlines how nanostatements can be cited following their usage in a discovery project [70]. While nanopubs could be formalized in any modelling language, the aforementioned conversion tools would enable more efficient use of the growing toxicology knowledgebase.

VI. Conclusion

The emerging field of systems toxicology encourages new approaches in the processing of experimental data. Unlike many standard toxicological approaches, the amount of data generated for a single sample requires sophisticated computational approaches for the processing and computationally available *a priori* knowledge for the interpretation. In this work, we present a workflow that creates these computer-readable knowledge clusters, here biological network models (see Figure 1). The starting point is the identification of relevant knowledge sources. The workflow continues with computational approaches to create knowledge statements (here BEL Nanopubs) and their manual curation by experts to ensure correctness of the knowledge formalization. This leads to the creation of assembled network models that can be used with computational methods (here the network perturbation amplitude) to calculate an overall biological impact factor for toxicity assessment (see Figure 5). As knowledge changes and extends by time, a strategy must be put in place to ensure knowledge representation based on the current opinion in a specific biological field.

Acknowledgment

We acknowledge Stephanie Boué for the creation of various graphics used in the figures.

References

[1] S. Ansari, "Knowledge acquisition and application for product risk impact analyses in an industrial setup," Proc. Eighth International Conference on Information, Process, and Knowledge Management (EKNOW 2016), IARIA, 2016, Venice, Italy.

[2] E. F. Nuwaysir, M. Bittner, J. Trent, J. C. Barrett, and C. A. Afshari, "Microarrays and toxicology: The advent of toxicogenomics," Molecular Carcinogenesis, vol. 24, pp. 153-159, 1999.

[3] T. Storck, M. Von Brevern, C. Behrens, J. Scheel, and A. Bach, "Transcriptomics in predictive toxicology," Current Opinion in Drug Discovery & Development, vol. 5, pp. 90-97, 2002.

[4] C. A. Afshari, H. K. Hamadeh, and P. R. Bushel, "The evolution of bioinformatics in toxicology: advancing toxicogenomics," Toxicological Sciences, vol. 120, pp. 225-237, 2010.

[5] M. North and C. D. Vulpe, "Functional toxicogenomics: Mechanism-centered toxicology," International Journal of Molecular Sciences, vol. 11, pp. 4796-4813, 2010.

[6] I. Moffat, N. L. Chepelev, S. Labib, J. Bourdon-Lacombe, B. Kuo, J. K. Buick, et al, "Comparison of toxicogenomics and traditional approaches to inform mode of action and points of departure in human health risk assessment of benzo [a] pyrene in drinking water," Critical Reviews in Toxicology, vol. 45, pp. 1-43, 2015.

[7] R. S. Thomas, M. A. Philbert, S. S. Auerbach, B. A. Wetmore, M. J. Devito, I. Cote, et al, "Incorporating new technologies into toxicity testing and risk assessment: Moving from 21st century vision to a data-driven framework," Toxicological Sciences, vol. 136, pp. 4-18, 2013.

[8] A. M. Richard, R. S. Judson, K. A. Houck, C. M. Grulke, P. Volarath, I. Thillainadarajah, et al, "The toxcast chemical landscape: Paving the road to 21st century toxicology," Chemical Research in Toxicology, vol.. 29, pp. 1225-1251, 2016.

[9] D. L. Villeneuve, D. Crump, N. Garcia-Reyero, M. Hecker, T. H. Hutchinson, C. A. LaLone, et al, "Adverse outcome pathway development ii: Best practices," Toxicological Sciences, vol. 142, pp. 321-330, 2014.

[10] J. C. Rowlands, M. Sander, J. S. Bus, and F. O. Committee, "Futuretox: Building the road for 21st century toxicology and risk assessment practices," Toxicological Sciences, vol 137, pp. 269-277, 2013.

[11] A. P. Davis, C. J. Grondin, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, et al, "The comparative toxicogenomics database's 10th year anniversary: Update 2015," Nucleic Acids Research, vol. 43, pp. 914-920, 2015.

[12] R. Benigni, "Predictive toxicology today: The transition from biological knowledge to practicable models," Expert Opinion on Drug Metabolism & Toxicology, vol. 12, pp. 989-992, 2016.

[13] D. Krewski, D. Acosta Jr, M. Andersen, H. Anderson, J. C. Bailar III, K. Boekelheide, et al, “Toxicity testing in the 21st century: A vision and a strategy,” Journal of Toxicology and Environmental Health, Part B, vol. 13, pp. 51-138, 2010.

[14] S. J. Sturla, A. R. Boobis, R. E. FitzGerald, J. Hoeng, R. J. Kavlock, K. Schirmer, et al, “Systems toxicology: From basic research to risk assessment,” Chemical Research in Toxicology, vol. 27, pp. 314-329, 2014.

[15] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources," Nature Protocols, vol. 4, pp. 44-57, 2009.

[16] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, et al, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” Proceedings of the National Academy of Sciences of the United States of America, vol. 102, pp. 15545-15550, 2005.

[17] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, et al, “Reactome: A knowledgebase of biological pathways,” Nucleic Acids Research, vol. 33, pp. 428-432, 2005.

[18] D. Nishimura, “Biocarta,” Biotech Software & Internet Report: The Computer Software Journal for Scient, vol. 2, pp. 117-120, 2001.

[19] M. Kanehisa and S. Goto, “Kegg: Kyoto encyclopedia of genes and genomes,” Nucleic Acids Research, vol. 28, pp. 27-30, 2000.

[20] L. Perfetto, L. Briganti, A. Calderone, A. C. Perpetuini, M. Iannuccelli, F. Langone, et al, “Signor: A database of causal relationships between biological entities,” Nucleic Acids Research, vol. 44, pp. 548-554, 2015.

[21] L. Chindelevitch, D. Ziemek, A. Enayetallah, R. Randhawa, B. Sidders, C. Brockel, et al, “Causal reasoning on biological networks: Interpreting transcriptional changes,” Bioinformatics, vol. 28, pp. 1114-1121, 2012.

[22] D. Djordjevic, A. Yang, A. Zadoorian, K. Rungrugeecharoen, and J. W. Ho, “How difficult is inference of mammalian causal gene regulatory networks?,” PloS One, vol. 9, Nov. 2014, doi: 10.1371/journal.pone.0111661.

[23] A. Krämer, J. Green, J. Pollard, and S. Tugendreich, “Causal analysis approaches in ingenuity pathway analysis (ipa),” Bioinformatics, vol. 30, pp. 523-530, 2013.

[24] W. Zhang, “Network toxicology: A new science,” Computational Ecology and Software, vol. 6, pp. 31-40, 2016.

[25] J. Hoeng, R. Deehan, D. Pratt, F. Martin, A. Sewer, T. M. Thomson, et al, “A network-based approach to quantifying the impact of biologically active substances,” Drug Discovery Today, vol. 17, pp. 413-418, 2012.

[26] J. Hoeng, M. Talikka, F. Martin, S. Ansari, D. Drubin, A. Elamin, et al, “Toxicopanomics: Applications of genomics, transcriptomics, proteomics and lipidomics in predictive mechanistic toxicology,” CRC Press, Boca Raton, USA, 2014.

[27] S. Boué, M. Talikka, J. W. Westra, W. Hayes, A. Di Fabio, J. Park, et al, “Causal biological network database: A comprehensive platform of causal biological network models focused on the pulmonary and vascular systems,” Database, Apr. 2015, doi: 10.1093/database/bav030.

[28] S. Boué, M. Talikka, J. W. Westra, W. Hayes, A. Di Fabio, J. Park, et al, "Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems," Database, vol. 2015, Apr. 2015, doi: 10.1093/database/bav030.

[29] J. W. Westra, W. K. Schlage, B. P. Frushour, S. Gebel, N. L. Catlett, W. Han, et al, “Construction of a computable cell proliferation network focused on non-diseased lung cells,” BMC Systems Biology, vol. 5, Jul. 2011, doi: 10.1186/1752-0509-5-105.

[30] S. Gebel, R. B. Lichtner, B. Frushour, W. K. Schlage, V. Hoang, M. Talikka, et al, “Construction of a computable network model for DNA damage, autophagy, cell death, and senescence,” Bioinformatics and Biology Insights, vol. 7, pp. 97-117, 2013.

[31] W. K. Schlage, J. W. Westra, S. Gebel, N. L. Catlett, C. Mathis, B. P. Frushour, et al, “A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue,” BMC Systems Biology, vol. 5, Oct. 2011, doi: 10.1186/1752-0509-5-168.

[32] J. W. Westra, W. K. Schlage, A. Hengstermann, S. Gebel, C. Mathis, T. Thomson, et al, “A modular cell-type focused inflammatory process network model for non-diseased pulmonary tissue,” Bioinformatics and Biology Insights, vol. 7, pp. 167-192, 2013.

[33] H. De León, S. Boué, W. K. Schlage, N. Boukharov, J. W. Westra, S. Gebel, et al, “A vascular biology network model focused on inflammatory processes to investigate atherogenesis and plaque instability,” Journal of Translational Medicine, vol. 12, Jun. 2014, doi: 10.1186/1479-5876-12-185.

[34] F. Martin, A. Sewer, M. Talikka, Y. Xiang, J. Hoeng, and M. C. Peitsch, “Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models,” BMC Bioinformatics, vol. 15, Jul. 2014, doi: 10.1186/1471-2105-15-238.

[35] A. Sewer, F. Martin, W. K. Schlage, J. Hoeng, and M. C. Peitsch, “Quantifying the biological impact of active substances using causal network models,” Computational Systems Toxicology, vol. 2015, pp. 223-256, doi: 10.1007/978-1-4939-2778-4_10.

[36] B. Phillips, E. Veljkovic, S. Boué, W. K. Schlage, G. Vuillaume, F. Martin, et al, “An 8-month systems toxicology inhalation/cessation study in apoe−/− mice to investigate cardiovascular and respiratory exposure effects of a candidate modified risk tobacco product, ths 2.2, compared with conventional cigarettes,” Toxicological Sciences, vol. 151, pp. 426-464, 2015.

[37] A. R. Iskandar, I. Gonzalez-Suarez, S. Majeed, D. Marescotti, A. Sewer, Y. Xiang, et al, “A framework for in vitro systems toxicology assessment of e-liquids,” Toxicology Mechanisms and Methods, vol. 26, pp. 389-413, 2016.

[38] S. Ansari, K. Baumer, S. Boué, S. Dijon, R. Dulize, K. Ekroos, et al, “Comprehensive systems biology analysis of a 7-month cigarette smoke inhalation study in c57bl/6 mice,” Scientific Data, vol. 3, Jan. 2016, doi: 10.1038/sdata.2015.77.

[39] U. Kogel, I. G. Suarez, Y. Xiang, E. Dossin, P. Guy, C. Mathis, et al, “Biological impact of cigarette smoke compared to an aerosol produced from a prototypic

modified risk tobacco product on normal human bronchial epithelial cells," Toxicology In Vitro, vol. 29, pp. 2102-2115, 2015.

[40] A. R. Iskandar, Y. Xiang, S. Frentzel, M. Talikka, P. Leroy, D. Kuehn, et al, "Impact assessment of cigarette smoke exposure on organotypic bronchial epithelial tissue cultures: A comparison of mono-culture and co-culture model containing fibroblasts," Toxicological Sciences, vol. 147, pp. 207-221, 2015.

[41] B. Phillips, E. Veljkovic, M. J. Peck, A. Buettner, A. Elamin, E. Guedj, et al, "A 7-month cigarette smoke inhalation study in c57bl/6 mice demonstrates reduced lung inflammation and emphysema following smoking cessation or aerosol exposure from a prototypic modified risk tobacco product," Food and Chemical Toxicology, vol. 80, pp. 328-345, 2015.

[42] T. Slater, "Recent advances in modeling languages for pathway maps and computable biological networks," Drug Discovery Today, vol. 19, pp. 193-198, 2014.

[43] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, et al, "The biopax community standard for pathway data sharing," Nat Biotech, vol. 28, pp. 935-942, 2010.

[44] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, et al, "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," Bioinformatics, vol. 19, pp. 524-531, 2015.

[45] P. Groth, A. Gibson, and J. Velterop, "The anatomy of a nanopublication," Information Services & Use, vol. 30, pp. 51-56, 2010.

[46] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, et al, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," Genome Research, vol. 13, pp. 2498-2504, 2003.

[47] J. Fluck, S. Ansari, J. Szostak, J. Hoeng, M. Zimmermann, M. Hofmann-Apitius, and M. Peitsch, "Belief - a semiautomatic workflow for bel network creation," 6th International Symposium on Semantic Mining in Biomedicine (SMBM 2014), 2014, Aveiro, Portugal.

[48] A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt, "Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies," PLoS Computational Biology, vol. 5, Dec. 2009, doi: 10.1371/journal.pcbi.1000605.

[49] C. UniProt, "The universal protein resource (uniprot) 2009," Nucleic Acids Res, vol. 37, pp. 169-174, 2009.

[50] T. C. Wiegers, A. P. Davis, K. B. Cohen, L. Hirschman, and C. J. Mattingly, "Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (ctd)," BMC Bioinformatics, vol. 10, Oct. 2009, doi: 10.1186/1471-2105-10-326.

[51] C. N. Arighi, B. Carterette, K. B. Cohen, M. Krallinger, W. J. Wilbur, P. Fey, et al, "An overview of the biocreative 2012 workshop track iii: Interactive text mining task," Database, Jan. 2013, doi: 10.1093/database/bas056.

[52] J. L. Warner, P. Anick, and R. E. Drews, "Physician inter-annotator agreement in the quality oncology practice initiative manual abstraction task," Journal of Oncology Practice / American Society of Clinical Oncology, vol. 9, May 2013, doi: 10.1200/JOP.2013.000931.

[53] R. Winnenburg, T. Wachter, C. Plake, A. Doms, and M. Schroeder, "Facts from text: Can text mining help to scale-up high-quality manual curation of gene products with ontologies?," Briefings in Bioinformatics, vol. 9, pp. 466-478, 2008.

[54] Z. Lu and L. Hirschman, "Biocuration workflows and text mining: Overview of the biocreative 2012 workshop track ii," Database, Nov. 2012, doi: 10.1093/database/bas043.

[55] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature," PLoS Biol, vol. 2, Nov. 2004, doi: 10.1371/journal.pbio.0020309.

[56] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, "Prominer: Rule-based protein and gene entity recognition," BMC Bioinformatics, vol. 6, May 2005, doi: 10.1186/1471-2105-6-S1-S14.

[57] K. Van Auken, P. Fey, T. Z. Berardini, R. Dodson, L. Cooper, D. Li, et al, "Text mining in the biocuration workflow: Applications for literature curation at wormbase, dictybase and tair," Database, Nov. 2012, doi: 10.1093/database/bas040.

[58] H. J. Drabkin, J. A. Blake, and M. G. I. Database, "Manual gene ontology annotation workflow at the mouse genome informatics database," Database, Oct. 2012, doi: 10.1093/database/bas045.

[59] C. E. Jones, A. L. Brown, and U. Baumann, "Estimating the annotation error rate of curated go database sequence annotations," BMC Bioinformatics, vol. 8, May 2007, doi: 10.1186/1471-2105-8-170.

[60] C. Andorf, D. Dobbs, and V. Honavar, "Exploring inconsistencies in genome-wide protein function annotations: A machine learning approach," BMC Bioinformatics, vol. 8, Aug. 2007, doi: 10.1186/1471-2105-8-284.

[61] D. Devos and A. Valencia, "Intrinsic errors in genome annotation," Trends in Genetics: TIG, vol. 17, pp. 429-431, 2001.

[62] P. D. Karp, "What we do not know about sequence analysis and sequence databases," Bioinformatics, vol. 14, pp. 753-754, 1998.

[63] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, et al, "Overview of biocreative ii gene normalization," Genome Biology, vol. 9, Sep. 2008, doi: 10.1186/gb-2008-9-s2-s3.

[64] D. P. Corney, B. F. Buxton, W. B. Langdon, and D. T. Jones, "Biorat: Extracting biological information from full-length papers," Bioinformatics, vol. 20, pp. 3206-3213, 2004.

[65] A. A. Namasivayam, A. F. Morales, Á. M. F. Lacave, A. Tallam, B. Simovic, D. G. Alfaro, et al, "Community-reviewed biological network models for toxicology and drug discovery applications," Gene Regulation and Systems Biology, vol. 10, pp. 51-66, 2016.

[66] F. Rinaldi, T. R. Ellendorff, S. Madan, S. Clematide, A. van der Lek, T. Mevissen, et al, "Biocreative v track 4: A shared task for the extraction of causal network information using the biological expression language," Database, Jul. 2016, doi: 10.1093/database/baw067.

[67] J. Szostak, S. Ansari, S. Madan, J. Fluck, M. Talikka, A. Iskandar, et al, "Construction of biological networks from unstructured information based on a semi-automated

curation workflow," Database, Jun. 2015, doi: 10.1093/database/bav057.

[68] B. Mons, H. van Haagen, C. Chichester, J. T. den Dunnen, G. van Ommen, E. van Mulligen, et al, "The value of data," Nature Genetics, vol. 43, pp. 281-283, 2011.

[69] E. Mina, M. Thompson, R. Kaliyaperumal, J. Zhao, Z. Tatum, K. M. Hettne, et al, "Nanopublications for exposing experimental data in the life-sciences: A huntington's disease case study," Journal of Biomedical Semantics, vol. 6, Feb. 2015, doi: 10.1186/2041-1480-6-5.

[70] A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, et al, "Open phacts: Semantic interoperability for drug discovery," Drug Discovery Today, vol. 17, pp. 1188-1198, 2012.