DATABASE
The Journal of Biological Databases and Curation

Database, 2019, 1–13
doi: 10.1093/database/baz068
Original article

Original article

Re-curation and rational enrichment of knowledge graphs in Biological Expression Language

Charles Tapley Hoyt[1,2,*], Daniel Domingo-Fernández[1,2], Rana Aldisi[1,2], Lingling Xu[1,2], Kristian Kolpeja[1], Sandra Spalek[1], Esther Wollert[1], John Bachman[3], Benjamin M. Gyori[3], Patrick Greene[3] and Martin Hofmann-Apitius[1,2]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany, [2]Bonn-Aachen International Center for Information Technology, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany, and [3]Laboratory of Systems Pharmacology, Harvard Medical School, 200 Longwood Ave, 02115 Boston, MA, USA

*Corresponding author: Tel: +49 2241 14-2268; Fax: +49 2241 14-2656; Email: charles.hoyt@scai.fraunhofer.de

Citation details: Hoyt,C. T., Domingo-Fernàndez,D., Aldisi,R. *et al.* Re-curation and rational enrichment of knowledge graphs in Biological Expression Language. *Database* (2019) Vol. 2019: article ID baz068; doi:10.1093/database/baz068

Received 31 January 2019; Revised 3 April 2019; Accepted 29 April 2019

Abstract

The rapid accumulation of new biomedical literature not only causes curated knowledge graphs (KGs) to become outdated and incomplete, but also makes manual curation an impractical and unsustainable solution. Automated or semi-automated workflows are necessary to assist in prioritizing and curating the literature to update and enrich KGs. We have developed two workflows: one for re-curating a given KG to assure its syntactic and semantic quality and another for rationally enriching it by manually revising automatically extracted relations for nodes with low information density. We applied these workflows to the KGs encoded in Biological Expression Language from the NeuroMMSig database using content that was pre-extracted from MEDLINE abstracts and PubMed Central full-text articles using text mining output integrated by INDRA. We have made this workflow freely available at https://github.com/bel-enrichment/bel-enrichment.

Database URL: https://github.com/bel-enrichment/results

Background

The rapid accumulation of unstructured knowledge in the biomedical literature has motivated its structuring and formalization so computers can assist in large-scale reasoning and interpretation. Several standard formats have been proposed for storing newly structured knowledge, including Systems Biology Markup Language [SBML; (1)], Biological Pathways Exchange Language [BioPAX; (2)],

Biological Expression Language [BEL; (3)] and Gene Ontology Causal Activity Models (4). Accompanying these standards are public repositories containing content generated both in academic and industrial contexts such as the BioModels Database (5), Pathway Commons (6), NDEx (7), Bio2RDF (8), Open PHACTS (9) and BEL Commons (10).

Even though each standard focuses on different aspects of modeling knowledge in systems and networks biology, they all give rise to knowledge graphs (KGs) consisting of biological entities (nodes), their interrelations (edges) and their associated metadata. While KGs have been useful for qualitative modeling of biochemical networks (11, 12), cellular signaling (13–15), gene regulatory pathways and genetic interactions (16, 17), metabolic pathways (18, 19) and other systems biology applications, there are several challenges associated with their use. First, they contain noise arising from curation, from the loss of information due to representation and from normalization of different knowledge representations (20–22). Second, they are generally an incomplete representation of the current state of scientific knowledge due to the large amount of uncurated, unstructured knowledge in the literature. Third, they progressively become out of date as scientific experimentation and investigation elucidate new knowledge (23). Finally, they often lack biological contextual information such as organelle, cell, cell line, tissue, organ, phenotype or disease specificity (24, 25).

KGs also suffer from issues in the normalization and mapping of entities. Though interoperability standards and resources like the Minimal Information Required in the Annotation of Models [MIRIAM; (26)] and Identifiers.org (27) have been developed and implemented to promote the semantic interoperability of biological models (and by extension, KGs), curators often encounter concepts that are not present in high-quality, publicly available terminologies and cannot capture the incident knowledge in a semantically meaningful way. These situations require enriching previously existing terminologies or, in some cases, developing new ones. For situations when the appropriate concept/term is unclear, several tools have been developed and made freely available to the community to help curators build semantically interoperable models including the Ontology Lookup Service [OLS; (28)], the Ontology Mapping Service (OxO; https://www.ebi.ac.uk/spot/oxo), Zooma (https://www.ebi.ac.uk/spot/zooma) and CEDAR Workbench (29). Further, recent work from Domingo-Fernández *et al*. on mapping pathways between major databases (30) and a critical assessment of their overlaps and contradictions (31) has shown that the adoption of standards like MIRIAM has been slow and that while the syntax of the varying formats used by each database may be correct, their semantic interoperability is still lacking.

Motivation

Accurately structuring and formalizing the unstructured knowledge in the biomedical literature requires careful planning and manual effort from trained curators. The scope of a given project must be defined based on its scientific goals (e.g. to support the interpretation of data, to generate a disease-specific knowledgebase etc.) and limited in its literature content sources (e.g. abstracts, full text, patents etc.) based on a project-specific metric for quality and relevance—both of which are nebulous in description and difficult to generate. The scope must also be limited to certain classes of biological entities, their interrelations and the standard formats that are capable of expressing them. For instance, the entities, relations and formats used during curation are different for protein complex assemblies curated by the Complex Portal (32) and regulatory interactions curated by the Signaling Network Open Resource (33). Similarly, curation guidelines must be defined reflecting these limits. For example, the guidelines of a project designed to model Tau aggregation inhibitors from the chemistry literature might encourage the curators to include direct binding partners of those inhibitors (e.g. GSK-3β, CDK5 etc.) but explicitly exclude the biological mechanisms through which the inhibitors' targets result in Tau aggregation that would better be curated during a different project focusing on capturing molecular biology from its primary literature. While there is no alternative to proper planning, several semi-automated curation workflows such as BEL information extraction workflow (BELIEF) (34) and the sbv IMPROVER (35) provide assistance by automatically detecting entities and relations for curators to accept or fix in order to increase productivity and enforce correct syntax and semantics. However, these and similar systems are limited in their ability to capture the relevant chemistry and biology, and reversion to manual curation is often necessary. Most issues arise from the complexity of terminology used in domain-specific biology and the heterogeneity of chemical nomenclature in general. Further, compositions of entities and concepts (e.g. there is a flexible and extensible terminology for complexes of proteins, such as the variety of combinations of subunits that form various nicotinic receptors) remain a challenge as natural language used in scholarly articles is constantly evolving. Finally, the issues of insufficient resources and fixed timelines apply to most curation projects, as aptly described by (36).

In the AETIONOMY project (https://www.aetionomy.eu), we manually curated NeuroMMSig, an inventory of multiscale and multimodal KGs that capture mechanistic knowledge in the context of neurological disorders (37). We encoded it in BEL because it is appropriate for qualitative causal, correlative and associative relationships between biological entities, processes and measurements across modes and scales. However, it is currently suffering from the issues we have previously described: it has not been assessed for confidence, is becoming outdated and needs to be enriched following a rational approach that best prioritizes the flood of recent literature.

To address this, we have developed and applied two workflows, described in this paper: the first is for re-curating existing BEL documents to ensure their syntactic and semantic correctness in a scenario where there was neither prior syntax validation, curation guidelines for entity nomenclature nor a second curator for achieving inter-annotator agreement. The second is a semi-automated algorithm and reproducible workflow for updating and rationally enriching an existing KG that lessens the burden of identifying relevant literature, reduces the overhead, as defined by Rodríguez-Esteban (36), and generates more, higher quality, relevant content.

We applied these workflows to a selection of KGs in NeuroMMSig and evaluated the curation effort (time) and quality in comparison to purely manual curation and other previously reported semi-automated curation workflows. We increased the number of nodes and edges in the selected KGs respectively by approximately five and seven times while maintaining the specificity of the KGs. With an improvement to the content underlying NeuroMMSig, the mechanism enrichment algorithm on its corresponding web service can return more correct and robust results to support the analysis of neuroimaging and genomics data for clinical trials in Alzheimer's disease, Parkinson's disease and epilepsy. Finally, we have made this workflow freely available at https://github.com/bel-enrichment/bel-enrichment so others can include it in their own curation workflows.

Methods

We first present the re-curation workflow for syntactic and semantic quality assurance before presenting our proposed approach for updating and rational enrichment.

Syntactic quality assurance

We developed a workflow using git (https://git-scm.com), GitHub (https://github.com), PyBEL (44) and a novel PyBEL extension PyBEL-Git (39) in order to identify and address syntactical issues in the BEL documents generated during the AETIONOMY project [https://www.aetionomy.eu; (40–44) and exposed through the NeuroMMSig mechanism enrichment server (37).

This workflow can be implemented in other web-based version control systems such as GitLab (https://gitlab.com) and Atlassian BitBucket (https://bitbucket.org) as well as directly integrated with continuous integration systems such as GitLab CI/CD (https://docs.gitlab.com/ee/ci), Travis-CI (https://travis-ci.com) and BitBucket Pipelines (https://bitbucket.org/product/features/pipelines) using the instructions provided at https://github.com/pybel/pybel-git with minimal configuration.

Semantic quality assurance

We selected 10 signatures (and their corresponding BEL documents) from NeuroMMSig based on their druggability (number of proteins targeted by drugs that have been assessed in clinical trials), their novelty (less preference given to subgraphs corresponding to hypotheses that have repeatedly failed in the clinic, namely amyloid-beta aggregation) and their amenability to assay development (based on expert advice) as an example for the re-curation workflow outlined below. An enumeration and statistics can be found in Table 1, and the signatures can be explored through BEL Commons. Because BEL was developed by the biomarker discovery company Selventa before the wide adoption of semantic resources like Identifiers.org, the Open Biomedical Ontology Foundry and the OLS, the language used a custom format for storing the names and identifiers of entities in major biomedical databases and ontologies such as the HUGO Genome Nomenclature Consortium [HGNC; (45)], Chemical Entities of Biological Interest [ChEBI; (46)], the Gene Ontology [GO; (4)], Medical Subject Headings [MeSH; (47)], the Disease Ontology [DO; (48)], the Human Phenotype Ontology [HPO; (49)], the Cell Line Ontology (50), the Experimental Factor Ontology (51) and others. Additionally, Selventa provided several entity type-specific, manually curated terminologies for chemicals, protein families, protein complexes and diseases for entities that had not yet been included in any of the other existing resources.

Because the Selventa terminologies are no longer maintained and the publicly available terminologies have far surpassed them in coverage, the first step in re-curation was to normalize entities to high-quality, publicly available terminologies. For example, chemicals were normalized to identifiers from ChEBI, ChEMBL (52) and PubChem (53) whenever possible; protein families and complexes were normalized to FamPlex (54); and diseases were normalized to DO and HPO. Further, because the BEL documents from

Table 1. Statistics for the number of BEL nodes and BEL statements in the 10 KGs selected from the NeuroMMSig inventory before re-curation (using the version last updated on 6 December 2016), after re-curation and after enrichment

Label	Description	Before re-curation		After re-curation		After enrichment	
		Nodes	Edges	Nodes	Edges	Nodes	Edges
Tau protein subgraph	The downstream effects of the post-translational modification, aggregation and transport of the Tau protein	191	493	261	733	708	2054
DKK1 subgraph, GSK3 subgraph	The interaction partners with GSK-3β and its targets of post-translational modification. The complementary DKK1 pathway is a specific signaling cascade upstream of GSK-3β	128	254	174	377	376	1165
Inflammatory response	Processes related to inflammation in the context of Alzheimer's disease	182	373	341	743	2003	7607
Insulin signal transduction	The molecular relationships between insulin resistance and inflammation, motivated by epidemiological studies that suggested a correlation between Alzheimer's disease (AD) and type II diabetes (55).	251	739	315	881	612	1973
Amyloidogenic subgraph	The downstream effects of the amyloid precursor protein (APP), its protein modifiers and its cleavage products	493	1223	652	1751	2090	7436
Non-amyloidogenic subgraph	Chemicals and processes known to down-regulate the expression of the transcript corresponding to APP or the abundance of the APP protein	195	359	325	635	795	2238
Apoptosis and cell death	Processes relevant to AD that result in apoptosis including the Caspase subgraph, XIAP subgraph and Complement system subgraph	104	143	170	229	1065	2401
Acetylcholine subgraph	Pathways including biological entities and processes related to cholinergic neurons and acetylcholine transmission	106	197	148	337	423	1275
GABA subgraph	Pathways including biological entities and process related to GABAergic neurons and GABA transmission	21	30	91	190	305	721
Reactive oxygen species subgraph	The effects of reactive oxygen species, including the Myeloperoxidase subgraph, Hydrogen peroxide subgraph, Free radical formation subgraph and Nitric oxide subgraph	104	173	126	224	1401	6277
Total		1188	3529	1704	5391	5850	23 811

Later, we discuss these statistics in terms of INDRA statements—the discrepancies are due to the ontological reasoner applied in the conversion process from INDRA statements to BEL statements.

Table 2. Confidence annotations using the Likert scale for re-curation

Confidence	Rationale
None	If the evidence string is nonsense or contains no reasonable biological knowledge, delete it and the related statements entirely. It is okay to remove BEL statements that are not supported.
Low	If it's not clear what BEL should represent the biology, add SET Confidence = "Low" for later discussion.
Medium	If the statement is wrong, fix it and add the annotation SET Confidence = "Medium".
High	If statement can be asserted from the given evidence, add the annotation SET Confidence = "High".

AETIONOMY were all produced before 2015, the entities that were curated using their labels (instead of stable identifiers) needed to be updated. A short investigation showed that HGNC and GO were the least stable namespaces (e.g. change of preferred label, splitting of entries, merging of entries and deprecation of entries), but combined they had less than 100 entities to be addressed. We therefore concluded that manual intervention was more appropriate than developing complicated systems for updating labels. While it is not intended to be the focus of this article, we have also begun to build a custom terminology (available at https://github.com/pharmacome/terminology) to supplement the publicly available ones for a small number (less than 1000) of terms that had not been included in other resources.

After ensuring both the correctness of BEL syntax and namespace usage, a remaining major aspect of re-curation is to address the issues arising from curation lacking inter-annotator agreement. BEL statements and their corresponding annotations (metadata) were generated by several independent curators and had not undergone quality control either by comparison with the results of independent curation of the same document by a second curator or even minimally checked by a second curator. We applied the following simple guidelines:

1. Second curator*:* check and label all relevant statements with a `SET Confidence` annotation using the Likert scale as described in Table 2.
2. Third curator (curation leader): after all relevant statements had been checked for correctness, check all statements with `SET Confidence = "High"` or `SET Confidence = "Medium"`. Change the confidence to `SET Confidence = "Very High"` on agreement. Otherwise, fix the statement.

The existence of the confidence guideline can be checked with the PyBEL command line interface with the following command: `pybel compile --required-annotations "Confidence"`.

Proposed approach for updating and rational enrichment

Next, we developed and applied a procedure for enriching a given BEL document in order to cope with the mounting issues of out of dateness and incompleteness. Our approach identifies nodes with low information density and uses a large-scale corpus of biomedical literature that has been pre-processed by automated relation extraction methods to identify the most relevant literature, evidences and ultimately relations. Notably, the previously described quality assurance (i.e. re-curation) workflows for checking and addressing the syntactic and semantic correctness of a given BEL document were necessary to decrease the noise input into the procedure. Following the re-curation of the 10 NeuroMMSig subgraphs, we applied the following procedure for rational enrichment:

1. KG pre-processing: nodes corresponding to the same gene (i.e. RNA, microRNA, protein and variants thereof) are collapsed, non-causal relationships (e.g. correlative, associative, ontological etc.) are removed and several entity types (i.e. abundances, reactions, pathologies and biological processes) are removed. While non-causal relations may be useful for exploration and in some analyses of a KG, their removal results in a graph of genes and their causal interrelations that allow the following enrichment steps to prioritize genes based on the amount of causal information available, which is required for several standard algorithms for analyzing experimental data such as Reverse Causal Reasoning (56). Further, the removal of other entity types and the collapsing of genes and their corresponding products were motivated by the focus of such algorithms on interactions between genes and their products. These pre-processing steps could be modified to fit other downstream applications. For example, the entity filter could be modified to include reactions and metabolites if the downstream application were to investigate metabolic flux.

2. Application of information density metric: the remaining nodes are ranked by an information density function. We used the sum of the node in-degree and out-degree as this corresponds to the amount of causal information for a given gene that is available in the KG. In this scenario, isolated nodes correspond to genes for which there is no causal information about its interactions with other proteins, and leaves (i.e. entities with only one edge) correspond to nodes that have very limited information.
3. Automated relation extraction: the top-ranked genes are used as a query to a KG generated by large-scale automated biological relation extraction. We used the Integrated Network and Dynamical Reasoning and Assembler [INDRA; (57)] and applied several filters to find the most relevant and novel relations. First, the relations that were already curated and in the KG were excluded. Second, INDRA was used to calculate a confidence score (between 0.0 and 1.0) for each relation based on evidences from structured databases and the frequency of occurrence of similar statements. Those statements with a low confidence score (<0.80) were removed to increase the precision and therefore reduce the curation overhead. While INDRA integrates relations extracted from multiple reading systems, a corpus of relations from a single machine reading system, such as EVEX, would serve the same purpose (58).
4. Conversion to BEL: different automated relation extraction systems present various information (e.g. entity offsets, events, triggers etc.) in ways that are not amenable to curation. Because INDRA already normalizes this information for several systems to several varieties of the `indra.Statement` Python class, we developed a converter to BEL using PyBEL that can be used directly with the `indra.assemblers.PybelAssembler` Python class. Finally, this information is exported to an Excel sheet with several additional columns for tracking INDRA statement provenance, curator provenance, the correctness of BEL statements, the type of errors found and the changes made to incorrect BEL statements. Examples and links to the full results can be found in the supplementary information. This process often results in the addition of entities that were excluded during KG pre-processing, such as biological processes and pathologies, as well as the inclusion of additional namespaces based on their corresponding priorities encoded in the converter.

For each round of rational enrichment, the procedure was applied to generate several curation sheets corresponding to the lowest information genes. Each row was checked with the following procedure:

1. Place an 'x' in the 'Checked' column.
2. If the BEL statement correctly corresponds to the 'Evidence' column, place an 'x' in the 'Correct' column.
3. Else if the BEL statement can be improved (e.g. assignment of entity types, relation etc.), correct it and place an 'x' in the 'Changed' column and annotate the error type in the 'Error Type' column using a controlled vocabulary (see the supplementary data). Additional guidelines for categorizing error types can be found at https://github.com/pharmacome/curation/blob/master/indra-errors.rst.
4. Else if the BEL statement does not correspond to the 'Evidence' column and cannot be improved, then 'x' should neither be placed in the 'Correct' nor the 'Changed' column.
5. If the 'Evidence' column contains other BEL statements that were not extracted, duplicate the current row's provenance (reference, evidence etc.) and add the additional BEL statements. Place an 'x' in the 'Changed' column but not the 'Correct' column.
6. If there are other BEL statements that can be extracted, make a new line with all of the same provenance information (uuid, reference, evidence etc.) and just place an 'x' in the 'Changed' column.

This procedure was applied iteratively: as the low information density nodes from the first round gained new relations, the KG was expanded and further low information density nodes were added. There are several improvements that could be made to the information density function and prioritization of the resulting extracted statements. For example, relations found by INDRA between low information density nodes and high information density nodes could be prioritized to maintain the scope and focus of a KG.

Results and discussion

While applying the re-curation workflow outlined in Figure 1, we identified large sections of poor-quality curation that had to be removed. Additionally, some evidences in the BEL document that were previously incompletely curated were completed. Re-curation also required the updating of namespaces from the 2015 versions to the most current and necessitated some additional revisions.

To evaluate the enrichment workflow outlined in Figure 2, we defined weekly curation rounds in which each of the 5 curators were tasked to curate the enrichment template generated by INDRA for the first 30 prioritized genes. Curators worked 10 hours per round for 1 month (4 weeks; 1 round per week) to curate BEL statements from a pool of 113 genes. A database of statements

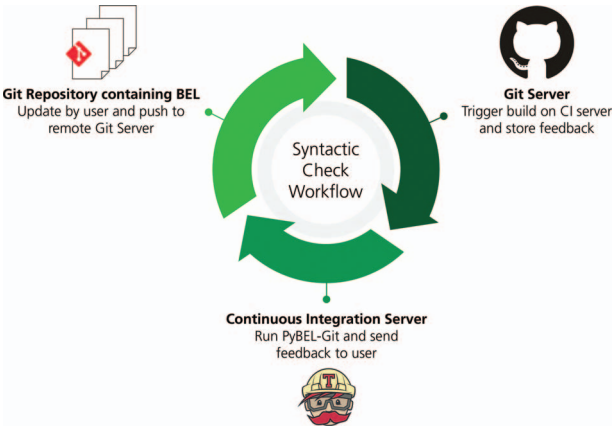

Figure 1. A workflow for syntactic quality assessment. This figure can be found on FigShare at https://doi.org/10.6084/m9.figshare.7643006.v1.

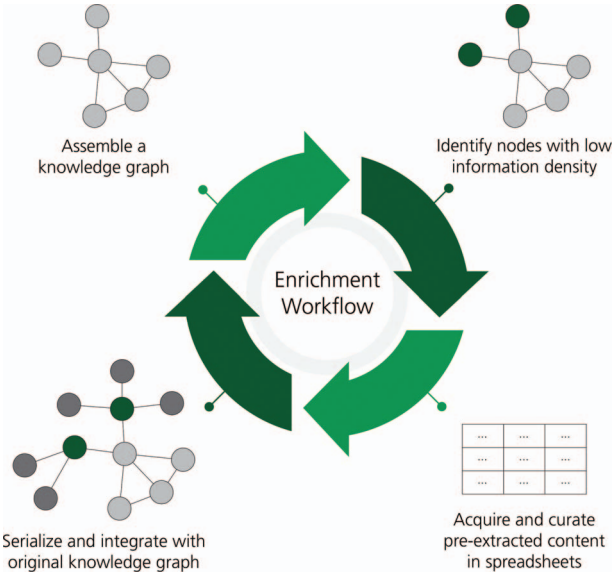

Figure 2. A workflow for the rational enrichment of knowledge graphs. This figure can be found on FigShare at https://doi.org/10.6084/m9.figshare.7642964.v1.

was generated by INDRA using the REACH (59, 60) and Sparser (61) readers to extract a total of 17096 statements containing these genes from all MEDLINE abstracts and PubMed Central full-text articles available in August 2018. Of these, 2989 were manually evaluated. A total of 917 statements (30.7%) were marked as correct by the curators, 1454 statements (48.6%) required manual corrections and the remainder (20.7%) could not be corrected. The criteria for correctness was that 'all' aspects of the statement, including the subject and object entities, relationship type, phosphorylation and other post-translational modifications, were extracted to the same extent as careful manual curation could. Ultimately, excluding the statements that could not be corrected, 79.3% of the automatically extracted, manually revised BEL statements were recovered. After curation, the recovered statements were converted into a BEL KG that contained 4228 nodes and 17002 edges complementary to the original 10 subgraphs selected from NeuroMMSig. The discrepancies in the number of INDRA statements to BEL statements are due to the ontological reasoning process that occurs during conversion. For example, INDRA statements about protein complex formation are converted to bi-directional BEL statements, INDRA statements about post-translationally modified proteins induce edges to the reference protein and INDRA statements about bound proteins create a variety of additional BEL nodes representing their constituents and membership edges connecting them.

There are two main aspects that are commonly used to formally evaluate a biocuration workflow: the time required to complete the task and quality of the curation compared with a gold standard. To evaluate whether the proposed approach for rational enrichment allows curating a larger amount of statements without compromising the quality, we calculated the average number of minutes required to curate one statement using our proposed workflow and compared it with previous estimates calculated conducting manual curation of BEL statements [(34); (62)] (Figure 3a). While the average curation effort was significantly lower than manual curation [2.19 minutes per BEL statement in our workflow vs. 3.2 minutes per BEL statement in manual curation reported by (62)], our calculations included the time used by the curators to annotate the various errors made by the reading system(s). Therefore, if the curation exercise would have exclusively focused on curating BEL statements, the average would have been even lower. Moreover, it is important to note that our proposed approach does not explicitly require the time nor expertise required for corpora generation because the reading systems (e.g. REACH and Sparser) and assembly systems (i.e. INDRA and PyBEL) are applied to all available literature. Unfortunately, it was not possible to make a direct comparison to the reported 1.7 minutes per BEL statement reported by Madan *et al.* (34) due to several confounding variables and the unavailability of further statistical information about the variability of curation time when using BELIEF.

Although the amount of time required to curate a certain amount of statements with the proposed approach is lower compared to standard manual curation, the curation effort is also highly variable depending on which gene was curated (Figure 3a). To investigate how the curation effort depends on the accuracy of the reader extracting BEL statements, we compared the average curation effort between genes whose statements were accurately and poorly extracted (Figure 3b). We observed that the curation effort required

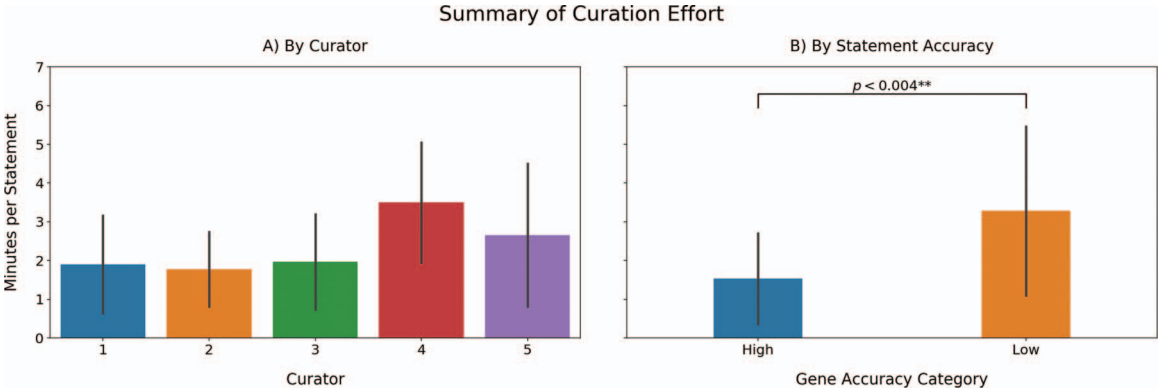

Figure 3. a) Recovered BEL statements per minute. Note that the time reported here includes the time invested in annotate the statement as well as INDRA errors. b) A comparison of the curation effort between genes for which INDRA had high accuracies (top 20) and genes presenting low accuracies (bottom 20).

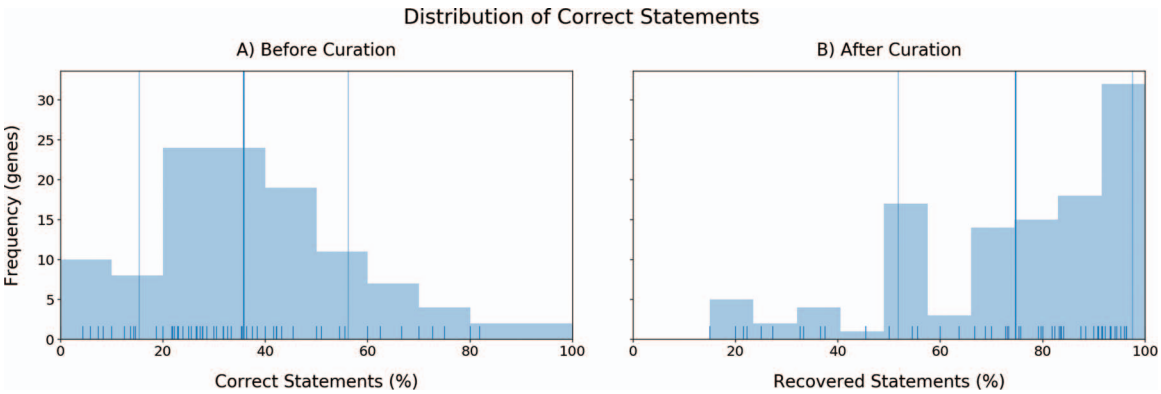

Figure 4. a) The distribution of the accuracies in triple identification by INDRA for each gene. X-axis: Correct statements (%). Y-axis: Number of genes (frequency). b) Distribution of recovered statements after curation (mean: 74.63%).

to extract statements in genes whose statements were highly accurate (top 20) was significantly less ($P < 0.004$; Student's *t*-test) than the effort required to curate low accuracy (bottom 20) genes, which effectively took as long as manual curation. We conclude that the high variability associated with the average curation times per curator can be explained by the extra invested time in the genes presenting low recall.

The second aspect we evaluated was the performance in terms of quality. To investigate the direct quality of the BEL statements coming from INDRA, we analyzed the distributions of correct statements before curation observed in each gene (accuracy investigation) (Figure 4a). Most of the genes presented accuracies close to the mean accuracy (35.75%) with only a few outliers whose limited number of extracted statements lead to their respective high or low accuracies (see Supplementary Figure 1). Furthermore, in accordance with previous research assessing the quality of automatic and manual relation extraction (63), the accuracies we observed again indicated that BEL statements must be manually curated in order to generate high-quality networks. After curation, the distribution of statements that were correct plus statements that were fixed during curation (i.e. excluding statements that were incorrect and could not be fixed) shifted completely to long-tailed distribution with an average of 74.63% BEL statements successfully extracted (Figure 4b). The remaining statements (~25%) could either not be coded in BEL or contained any relevant information about the particular gene.

While curating the BEL statements, we also annotated the errors made throughout the process of reading, assembly by INDRA and conversion to BEL by PyBEL in order to identify common mistakes and to assist in the improvement of these three systems. The results showed that the most common error is caused by the name entity recognition system that identifies the entities participating in the relation (Figure 5). Other common errors arose from the improper assignment of the subject and object entities, from evidences that did not actually include relations between the subject and object entities and statements that were semantically incorrect due to a negation word (e.g. not, no, none, neither etc.).

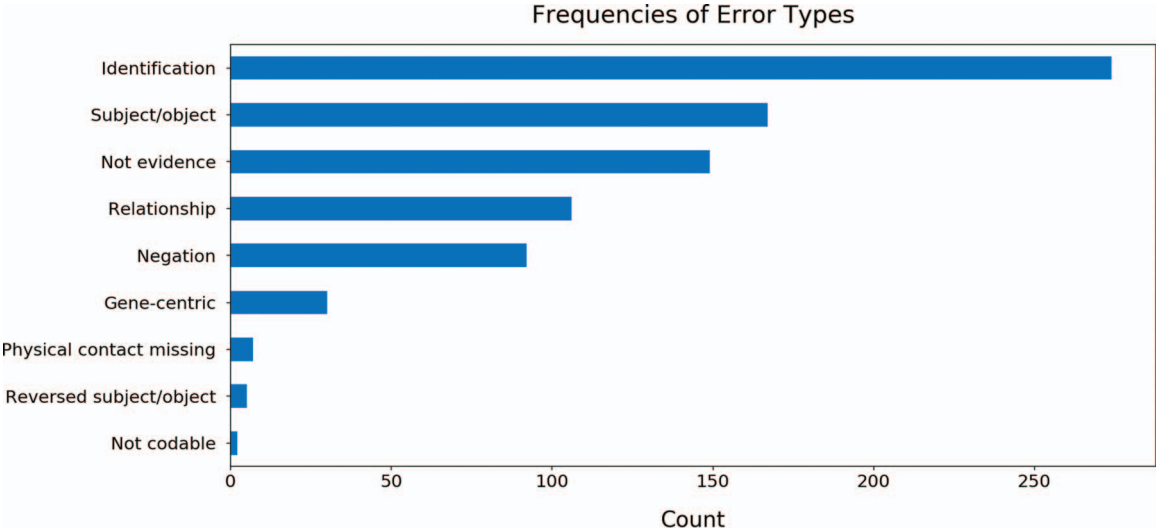

Figure 5. The frequencies of common errors found while curating BEL statements generated from 113 genes. Further details about each error type and the annotation process are available in the guidelines available at https://github.com/pharmacome/curation/blob/master/indra-errors.rst.

The five curators were tasked with tagging interesting examples of the common mistakes that could be used to inform the development of the reading systems (REACH, Sparser etc.) and the assembly systems (INDRA and PyBEL). Because the authors of this manuscript maintain the INDRA and PyBEL packages, identifying the causes of errors in assembly was relatively straightforward. For example, BEL statements containing biological processes were consistently output using invalid BEL syntax, including the *activity()* function, which is reserved for proteins and other physical entities. We addressed this by updating the previously mentioned *indra.assemblers.PybelAssembler* class. Another error type that was not addressed until after the evaluation was completed was the determination of the role of direct physical interaction in causal relations. INDRA makes use of linguistic cues from the text mining systems along with information from protein–protein interaction databases to determination if a relation involves a physical interaction between proteins, but this information was not incorporated into the *indra.assemblers.PybelAssembler* class. Instead, by default all relations were output using BEL statements implying physical contact: 'directly increases' (i.e. increases via contact) and directly decreases (i.e. decreases via contact). This issue has since been fixed. In general, the direct/indirect distinction is difficult to detect automatically in natural language, though it is very important in the generation of mechanistic and mathematical models arising from biological knowledge.

In Table 3, we present a small sampling of the errors and corresponding suggestions for improvement in the reading systems. We present a much more thorough enumeration of the errors found in statements for the 113 curated genes in the supplementary information. Besides generating new content quickly, this curation procedure includes information to allow for the evaluation of the automated relation extraction systems and for the proposition of improvements. For example, new groundings can be proposed for entities that were often mismatched. A prominent example was the misidentification of tau (a human protein) and taurine (an amino acid).

Additionally, new rules could be suggested for rule-based systems to avoid issues with the misidentification of the order of the subject and object as in the example of 'Bak expression was also induced in cells overexpressing the stress-induced transcription factor GADD153, but Bak expression was inhibited in cells expressing an antisense GADD153 construct' (64) whose use of the passive voice may have caused REACH to interpret the statement as 'Bak increased GADD153'. Ultimately, we believe we can use these examples to provide useful feedback to the developers of the reading systems and improve future extraction.

After applying the re-curation workflow to our selection of KGs in the NeuroMMSig inventory, we increased the number of nodes from 1188 to 1704 (∼1.5×) and edges from 3529 to 5391 (∼1.5×). After applying the enrichment workflow, the number of nodes increased to 5850 (∼5×) and edges to 23811 (∼7×). A more granular summary can be found in Table 1. With a 5× increase in nodes, we would expect to see a 10× increase in edges if the new nodes were completely disconnected from the pre-existing nodes in the KG, which shows that we have been able to maintain the specificity of the KGs to a reasonable degree. In total, our curators spent 80 hours

Table 3. Examples of errors that resulted in suggestions for improvements for the underlying relation extraction systems

Gene	Evidence	Issue	Suggestion
MRC1	In conclusion, these results suggest that BCR and ABL kinase abrogates MMR activity to inhibit apoptosis and induce mutator phenotype. (65)	MRC1, also known as MMR, was confused with mismatch repair (MMR)	Machine learning methods generating contextual word embeddings could be used to improve the named entity recognition component such as NeuralCoref (https://github.com/huggingface/neuralcoref).
TIMP1	In our work, the restoration of cholesterol efflux capacities from EPA-enriched human monocyte-derived macrophages (HMDM) treated with both the adenylate cyclase activator forskolin and the phosphodiesterase inhibitor IBMX strongly suggests that EPA decreased the ABCA1 mediated cholesterol efflux from HMDM through a PKA dependent pathway. (66)	TIMP1, also known as EPA, was confused with eicosapentaenoic acid (EPA)	Improve the named entity recognition (disambiguation) process, for example, by updating synonym dictionaries in rule-based systems.
TRPV1	Moreover, recently TRPV1 has been demonstrated to be either inhibited or activated by PIP 2. (67)	Only the inhibition relationship was extracted	Rule-based relation extraction systems could be appended with new rules to handle sentences with multiple objects. This and similar examples could be included in the training data for machine learning-based relation extraction.
NUMB	This interaction is mediated by the NPXY motif of LNX1 and leads to ubiquitination of Numb by the RING domain of LNX1, thereby targeting Numb to proteasomal degradation. (68)	The complex sentence structure of 'ubiquitination' and 'targeting' event were not resolved properly, and the ubiquitination was omitted.	Rule-based systems like REACH that explicitly handle ubiquitination events could be appended with new rules.
USF2	Taken together, the results shown in Figure 5A–C suggest that USF2 stimulates the transcriptional activity of NFκB by enhancing the degradation of IκBα. (69)	Relation should be treated as an indirect, rather than direct, increase	Update the INDRA PybelAssembler to make use of information about whether a relation is mediated through physical contact.

on the enrichment step to generate 17 002 new BEL statements with an average rate of 3.54 edges per minute. The resulting enriched KG can be used in reproductions of previous analyses leveraging the NeuroMMSig inventory to assess their robustness, deliver new insights and improve future analyses when the results are incorporated into a future release of the NeuroMMSig mechanism enrichment server. Additionally, the statements comprise a large training set for future machine learning approaches for text mining.

Conclusions

We have proposed and applied a generalizable workflow for enriching and updating existing biological KGs with a focus on the reduction of curation time both in literature triage and in extraction. While its realization involved spreadsheets rather than a *bona fide* curation interface, we believe that it could be adopted by both BEL-specific curation interfaces [e.g. BELIEF and BioDati Studio (https://studio.demo.biodati.com)] and more general biological relation curation interfaces [e.g. NOCTUA (http://noctua.berkeleybop.org), Factoid (https://github.com/PathwayCommons/factoid) and WikiPathways (70)]. Furthermore, INDRA is flexible enough to generate curation sheets for curators familiar with formats other than BEL, such as BioPAX or SBML.

This workflow is by no means the ultimate solution for finding relevant content. Using pre-extracted statements as a stand-in for relevance allows a given KG to be expanded, but it requires several rounds to find the limits of a given pathway or graph, during which the scope of the curation could be lost. We plan to investigate other methods for identifying relevant content by combining topic modeling with mind maps to not only identify content at the entity level, but on a higher abstraction that allows for capturing of entire areas of biology. These methods could compensate for the simplications that we made to the curation task, such as removing relations containing chemicals, biological processes and phenotypes. Additionally, they could enable

earlier-stage curation that is more focused on achieving reasonable coverage of the available knowledge rather than high granularity enrichment.

Ultimately, as automated relation extraction technologies improve, they will be used to more significantly supplement manual curation efforts. We expect to see many upcoming workflows leveraging these exciting prospects.

Authors' contributions

C.T.H. and D.D.F. conceived and designed the study and authored this manuscript. C.T.H., D.D.F., R.A., L.X., S.S., E.W. and K.K. performed curation. J.B., B.M.G. and P.G. provided data. M.H.A. supervised the project.

Availability of data and materials

The pybel-git Python package that was used to assess syntactic quality is openly available at https://github.com/pybel/pybel-git. All other code and analysis is openly available at https://github.com/bel-enrichment.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

We would like to thank Stephan Gebel for his organizational support and Alina Enns and Keerthika Lohanadan for their help in the curation tasks.

Funding

Fraunhofer Society under the MAVO Project; Human Brain Pharmacome; EU/EFPIA Innovative Medicines Initiative Joint Undertaking under AETIONOMY (115568 to D.D.F.); European Union's Seventh Framework Programme (FP7/2007-2013); EFPIA.

Conflict of interest. None declared.

References

1. Hucka,M., Finney,A., Sauro,H.M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531. https://doi.org/10.1093/bioinformatics/btg015.
2. Demir,E., Cary,M.P., Paley,S. *et al.* (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.*, **28**, 1308–1308. https://doi.org/10.1038/nbt1210-1308c.
3. Slater,T. (2014) Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discov. Today*, **19**, 193–198. https://doi.org/10.1016/j.drudis.2013.12.011.
4. Carbon,S., Dietze,H., Lewis,S.E. *et al.* (2017) Expansion of the gene ontology knowledgebase and resources: the gene ontology consortium. *Nucleic Acids Res.*, **45**, D331–D338. https://doi.org/10.1093/nar/gkw1108.
5. Glont,M., Nguyen,T.V.N., Graesslin,M. *et al.* (2018) BioModels: expanding horizons to include more modelling approaches and formats. *Nucleic Acids Res.*, **46**, D1248–D1253. https://doi.org/10.1093/nar/gkx1023.
6. Cerami,E.G., Gross,B.E., Demir,E. *et al.* (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, 685–690. https://doi.org/10.1093/nar/gkq1039.
7. Pratt,D., Chen,J., Welker,D. *et al.* (2015) NDEx, the Network Data Exchange. *Cell Systems*, **1**, 302–305. https://doi.org/10.1016/j.cels.2015.10.001.
8. Belleau,F., Nolin,M.-A., Tourigny,N. *et al.* (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.*, **41**, 706–716. https://doi.org/10.1016/j.jbi.2008.03.004.
9. Williams,A.J., Harland,L., Groth,P. *et al.* (2012) Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today*, **17**, 1188–1198. https://doi.org/10.1016/j.drudis.2012.05.016.
10. Hoyt,C.T., Domingo-Fernández,D. and Hofmann-Apitius,M. (2018) BEL Commons: an environment for exploration and analysis of networks encoded in Biological Expression Language. *Database (Oxford)*, **2018**, 1–11. https://doi.org/10.1093/database/bay126.
11. Rausanu,S. *et al.* (2015) Computational models for inferring biochemical networks. *Neural Comput. Appl.*, **26**, 299–311. https://doi.org/10.1007/s00521-014-1617-x.
12. Yugi,K., Kubota,H., Hatano,A. *et al.* (2016) Trans-omics: how to reconstruct biochemical networks across multiple 'omic' layers. *Trends Biotechnol.*, **34**, 276–290. https://doi.org/10.1016/j.tibtech.2015.12.013.
13. Pilalis,E., Koutsandreas,T., Valavanis,I. *et al.* (2015) KENeV: a web-application for the automated reconstruction and visualization of the enriched metabolic and signaling super-pathways deriving from genomic experiments. *Comput. Struct. Biotechnol. J.*, **13**, 248–255. https://doi.org/10.1016/j.csbj.2015.03.009.
14. Pon,A., Jewison,T., Su,Y. *et al.* (2015) Pathways with PathWhiz. *Nucleic Acids Res.*, **43**, W552–W559. https://doi.org/10.1093/nar/gkv399.
15. Tripathi,S., Flobak,Å., Chawla,K. *et al.* (2015) The gastrin and cholecystokinin receptors mediated signaling network: a scaffold for data analysis and new hypotheses on regulatory mechanisms. *BMC Syst. Biol.*, **9**, 1–15. https://doi.org/10.1186/s12918-015-0181-z.
16. Kandasamy,K., Mohan,S.S., Raju,R. *et al.* (2010) NetPath: a public resource of curated signal transduction pathways. *Genome Biol.*, **11**, R3. https://doi.org/10.1186/gb-2010-11-1-r3.
17. Kamburov,A., Stelzl,U., Lehrach,H. *et al.* (2013) The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.*, **41**, 793–800. https://doi.org/10.1093/nar/gks1055.
18. Caspi,R., Billington,R., Ferrer,L. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480. https://doi.org/10.1093/nar/gkv1164.

19. Wishart,D.S., Feunang,Y.D., Marcu,A. *et al.* (2018) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.*, **46**, D608–D617. https://doi.org/10.1093/nar/gkx1089.
20. Nickel,M., Murphy,K., Tresp,V. *et al.* (2016) A review of relational machine learning for knowledge graphs. *Proc. IEEE*, **104**, 11–33. https://doi.org/10.1109/jproc.2015.2483592.
21. Mihindukulasooriya,N., Hassanzadeh,O., Dash,S. *et al.* (2017) Towards comprehensive noise detection in automatically-created knowledge graphs. In: *CEUR Workshop Proceedings*, Vol. **1963**, pp. 1–4.
22. Pujara,J., Augustine,E. and Getoor,L. (2017) Sparsity and noise: where knowledge graph embeddings fall short. In: *Conference on Empirical Methods in Natural Language Processing*, pp. 1752–1757.
23. Wadi,L., Meyer,M., Weiser,J. *et al.* (2016) Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods*, **13**, 705–706. https://doi.org/10.1038/nmeth.3963.
24. Hofmann-Apitius,M., Ball,G., Gebel,S. *et al.* (2015) Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders. *Int. J. Mol. Sci.*, **16**, 29179–29206. https://doi.org/10.3390/ijms161226148.
25. Saqi,M., Lysenko,A., Guo,Y.-K. *et al.* (2018) Navigating the disease landscape: knowledge representations for contextualizing molecular signatures. *Brief. Bioinform.*, bby025. https://doi.org/10.1093/bib/bby025.
26. Laibe,C. and Le Novère,N. (2007) MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. *BMC Syst. Biol.*, **1**, 58. https://doi.org/10.1186/1752-0509-1-58.
27. Juty,N., Le Novere,N. and Laibe,C. (2012) Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res.*, **40**, 580–586. https://doi.org/10.1093/nar/gkr1097.
28. Cote,R., Jones,P., Apweiler,R. *et al.* (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**, 1–7. https://doi.org/10.1186/1471-2105-7-97.
29. Gonçalves,R.S., O'Connor,M.J., Martínez-Romero,M. *et al.* (2017) The CEDAR workbench: an ontology-assisted environment for authoring metadata that describe scientific experiments. *Lecture Notes in Computer Science*, Vol. 10588, pp. 103–110. https://doi.org/10.1007/978-3-319-68204-4_10.
30. Domingo-Fernández,D., Hoyt,C.T., Bobis Alvarez,C. *et al.* (2018) ComPath: an ecosystem for exploring, analyzing, and curating pathway databases. *NPJ Syst. Biol. Appl.*, **5**, 3. https://doi.org/10.1038/s41540-018-0078-8.
31. Domingo-Fernández,D., Mubeen,S., Marín-Llaó,J. *et al.* (2019) PathMe: merging and exploring mechanistic pathway knowledge. *BMC Bioinformatics*, **20**, 243. https://doi.org/10.1186/s12859-019-2863-9.
32. Meldal,B.H.M., Forner-Martinez,O., Costanzo,M.C. *et al.* (2015) The complex portal—an encyclopaedia of macromolecular complexes. *Nucleic Acids Res.*, **43**, D479–D484. https://doi.org/10.1093/nar/gku975.
33. Perfetto,L., Briganti,L., Calderone,A. *et al.* (2016) SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res.*, **44**, D548–D554. https://doi.org/10.1093/nar/gkv1048.
34. 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 (Oxford)*, **2016**, 1–17. https://doi.org/10.1093/database/baw136.
35. Guryanova,S. and Guryanova,A. (2017) sbv IMPROVER: modern approach to systems biology. *Methods Mol. Biol.*, **1613**, 21–29. https://doi.org/10.1007/978-1-4939-7027-8_2.
36. Rodriguez-Esteban,R. (2015) Biocuration with insufficient resources and fixed timelines. *Database (Oxford)*, **2015**, 1–9. https://doi.org/10.1093/database/bav116.
37. Domingo-Fernández,D., Kodamullil,A.T., Iyappan,A. *et al.* (2017) Multimodal mechanistic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. *Bioinformatics*, **33**, 3679–3681. https://doi.org/10.1093/bioinformatics/btx399.
38. Hoyt,C.T., Konotopez,A. and Ebeling,C. (2018) PyBEL: a computational framework for Biological Expression Language. *Bioinformatics*, **34**, 703–704. https://doi.org/10.1093/bioinformatics/btx660.
39. Hoyt,C.T. (2018) cthoyt/pybel-git v0.0.1 (Version v0.0.1). Zenodo. http://doi.org/10.5281/zenodo.1491432 (19 November 2018, date last accessed).
40. Irin,A.K., Tom Kodamullil,A., Gündel,M. *et al.* (2015) Computational modelling approaches on epigenetic factors in neurodegenerative and autoimmune diseases and their mechanistic analysis. *J. Immunol. Res.*, **2015**, 1–10. https://doi.org/10.1155/2015/737168.
41. Kodamullil,A.T., Younesi,E., Naz,M. *et al.* (2015) Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis. *Alzheimers Dement*, **11**, 1329–1339. https://doi.org/10.1016/j.jalz.2015.02.006.
42. Naz,M., Kodamullil,A.T. and Hofmann-Apitius,M. (2016) Reasoning over genetic variance information in cause-and-effect models of neurodegenerative diseases. *Brief. Bioinform.*, **17**, 505–516. https://doi.org/10.1093/bib/bbv063.
43. Emon,M.A.E.K., Kodamullil,A.T., Karki,R. *et al.* (2017) Using drugs as molecular probes: a computational chemical biology approach in neurodegenerative diseases. *J. Alzheimers Dis.*, **56**, 677–686. https://doi.org/10.3233/JAD-160222.
44. Hoyt,C.T., Domingo-Fernández,D., Balzer,N. *et al.* (2018) A systematic approach for identifying shared mechanisms in epilepsy and its comorbidities. *Database (Oxford)*, **2018**, 269860. https://doi.org/10.1093/database/bay050.
45. Yates,B., Braschi,B., Gray,K.A. *et al.* (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625. https://doi.org/10.1093/nar/gkw1033.
46. Hastings,J., De Matos,P., Dekker,A. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, 456–463. https://doi.org/10.1093/nar/gks1146.
47. Rogers,F.B. (1963) Medical subject headings. *Bull. Med. Libr. Assoc.*, **51**, 114–116.
48. Schriml,L.M., Mitraka,E., Munro,J. *et al.* (2018) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962. https://doi.org/10.1093/nar/gky1032.
49. Köhler,S., Carmody,L., Vasilevsky,N. *et al.* (2018) Expansion of the Human Phenotype Ontology (HPO) knowledge base and

resources. *Nucleic Acids Res*., 1–10. https://doi.org/10.1093/nar/gky1105.

50. Sarntivijai,S., Lin,Y., Xiang,Z. *et al.* (2014) CLO: the cell line ontology. *J. Biomed. Semantics*, **5**, 1–10. https://doi.org/10.1186/2041-1480-5-37.
51. Malone,J., Holloway,E., Adamusiak,T. *et al.* (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118. https://doi.org/10.1093/bioinformatics/btq099.
52. Gaulton,A., Hersey,A., Nowotka,M.L. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res*., **45**, D945–D954. https://doi.org/10.1093/nar/gkw1074.
53. Kim,S., Thiessen,P.A., Bolton,E.E. *et al.* (2016) PubChem substance and compound databases. *Nucleic Acids Res*., **44**, D1202–D1213. https://doi.org/10.1093/nar/gkv951.
54. Bachman,J.A., Gyori,B.M. and Sorger,P.K. (2018) FamPlex: a resource for entity recognition and relationship resolution of human protein families and complexes in biomedical text mining. *BMC Bioinformatics*, **19**, 1–14. https://doi.org/10.1186/s12859-018-2211-5.
55. Karki,R., Tom Kodamullil,A. and Hofmann-Apitius,M. (2017) Comorbidity analysis between Alzheimer's disease and type 2 diabetes mellitus (T2DM) based on shared pathways and the role of T2DM drugs. *J. Alzheimers Dis*., **60**, 721–731.
56. Catlett,N.L., Bargnesi,A.J., Ungerer,S. *et al.* (2013) Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics*, **14**, 340. https://doi.org/10.1186/1471-2105-14-340.
57. Gyori,B.M., Bachman,J.A., Subramanian,K. *et al.* (2017) From word models to executable models of signaling networks using automated assembly. *Mol. Syst. Biol*., **13**, 954. https://doi.org/10.15252/msb.20177651.
58. Van Landeghem,S., Hakala,K., Rönnqvist,S. *et al.* (2012) Exploring biomolecular literature with EVEX: connecting genes through events, homology, and indirect associations. *Adv. Bioinformatics*, **2012**, 582765. https://doi.org/10.1155/2012/582765.
59. Valenzuela-Escárcega,M.A., Hahn-Powell,G., Surdeanu,M. *et al.* (2015) A domain-independent rule-based framework for event extraction. In: *Proceedings of ACL-IJCNLP 2015 System Demonstrations*. Association for Computational Linguistics and the Asian Federation of Natural Language Processing. Stroudsburg, PA, USA pp. 127–132. https://doi.org/10.3115/v1/P15-4022.
60. Valenzuela-Escárcega,M.A., Babur,Ö., Hahn-Powell G, *et al.* (2018) Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database (Oxford)*, **2018**, 1–14. https://doi.org/10.1093/database/bay098.
61. McDonald,D.D. (2000) Issues in the representation of real texts: the design of KRISP. In: *Natural Language Processing and Knowledge Representation*, pp. 77–110. https://github.com/ddmcdonald/sparser.
62. Szostak,J., Ansari,S., Madan,S. *et al.* (2015) Construction of biological networks from unstructured information based on a semi-automated curation workflow. *Database (Oxford)*, **2015**, bav057, https://doi.org/10.1093/database/bav057.
63. 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 (Oxford)*, **2016**, baw067, https://doi.org/10.1093/database/baw067.
64. Lovat,P.E., Oliverio,S., Corazzari,M. *et al.* (2003) Bak: a downstream mediator of fenretinide-induced apoptosis of SH-SY5Y neuroblastoma cells. *Cancer Res*., **63**, 7310–7313.
65. Stoklosa,T., Poplawski,T., Koptyra,M. *et al.* (2008) BCR/ABL inhibits mismatch repair to protect from apoptosis and induce point mutations. *Cancer Res*., **68**, 2576–2580.
66. Fournier,N., Tardivel,S., Benoist,J.-F. *et al.* (2016) Eicosapentaenoic acid membrane incorporation impairs ABCA1-dependent cholesterol efflux via a protein kinase A signaling pathway in primary human macrophages. *Biochim. Biophys. Acta*, **1861**, 331–341.
67. Morelli,M.B., Amantini,C., Nabissi,M. *et al.* (2014) Cross-talk between alpha 1D-adrenoceptors and transient receptor potential vanilloid type 1 triggers prostate cancer cell proliferation. *BMC Cancer*, **14**, 921.
68. Young,P., Nie,J., Wang,X. *et al.* (2005) LNX1 is a perisynaptic Schwann cell specific E3 ubiquitin ligase that interacts with ErbB2. *Mol. Cell. Neurosci*., **30**, 238–248.
69. Wang,L., Li,H., Zhang,Y. *et al.* (2009) HINT1 inhibits β-catenin/TCF4, USF2 and NFκB activity in human hepatoma cells. *Int. J. Cancer*, **124**, 1526–1534.
70. Slenter,D.N., Kutmon,M., Hanspers,K. *et al.* (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res*., **46**, D661–D667.