Bioinformatics, 35(12), 2019, 2133–2140
doi: 10.1093/bioinformatics/bty933
Advance Access Publication Date: 8 November 2018
Original Paper

Databases and ontologies

OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction

Fatima Zohra Smaili, Xin Gao* and Robert Hoehndorf *

Computer, Electrical & Mathematical Sciences and Engineering (CEMSE) Division, Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on August 5, 2018; revised on November 2, 2018; editorial decision on November 5, 2018; accepted on November 7, 2018

Abstract

Motivation: Ontologies are widely used in biology for data annotation, integration and analysis. In addition to formally structured axioms, ontologies contain meta-data in the form of annotation axioms which provide valuable pieces of information that characterize ontology classes. Annotation axioms commonly used in ontologies include class labels, descriptions or synonyms. Despite being a rich source of semantic information, the ontology meta-data are generally unexploited by ontology-based analysis methods such as semantic similarity measures.

Results: We propose a novel method, OPA2Vec, to generate vector representations of biological entities in ontologies by combining formal ontology axioms and annotation axioms from the ontology meta-data. We apply a Word2Vec model that has been pre-trained on either a corpus or abstracts or full-text articles to produce feature vectors from our collected data. We validate our method in two different ways: first, we use the obtained vector representations of proteins in a similarity measure to predict protein–protein interaction on two different datasets. Second, we evaluate our method on predicting gene–disease associations based on phenotype similarity by generating vector representations of genes and diseases using a phenotype ontology, and applying the obtained vectors to predict gene–disease associations using mouse model phenotypes. We demonstrate that OPA2Vec significantly outperforms existing methods for predicting gene–disease associations. Using evidence from mouse models, we apply OPA2Vec to identify candidate genes for several thousand rare and orphan diseases. OPA2Vec can be used to produce vector representations of any biomedical entity given any type of biomedical ontology.

Availability and implementation: `https://github.com/bio-ontology-research-group/opa2vec`

Contact: robert.hoehndorf@kaust.edu.sa or xin.gao@kaust.edu.sa

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biological knowledge is widely spread across different types of resources. Biomedical ontologies have been highly successful in providing the means to integrate data across multiple disparate sources by providing an explicit and shared specification of a conceptualization of a domain (Gruber, 1995). Notably, ontologies provide a structured and formal representation of biological knowledge through logical axioms (Hoehndorf *et al.*, 2015b), and ontologies are therefore widely used to capture information that is extracted from literature by biocurators (Bodenreider, 2008). However, ontologies do not only include a formal, logic-based structure but also include many pieces of meta-data that are primarily intended for

human use, such as labels, descriptions or synonyms (Smith *et al.*, 2007).

Due to the pervasiveness of ontologies in the life sciences, many applications have been built that exploit various aspects of ontologies for data analysis and to construct predictive models. For example, a wide selection of semantic similarity measures have been developed to exploit information in ontologies (Al-Mubaid and Nguyen, 2006; Jiang and Conrath, 1997; Leacock and Chodorow, 1998; Lin *et al.*, 1998; Li *et al.*, 2003; Resnik *et al.*, 1999; Wu and Palmer, 1994), and semantic similarity measures have successfully been applied to the prediction of protein–protein interactions (Pesquita *et al.*, 2009), gene–disease associations (Köhler *et al.*, 2009) or drug targets (Hoehndorf *et al.*, 2014).

Recently, a set of methods have been developed that can characterize nodes and edges in knowledge graphs through 'embeddings'. A knowledge graph is a directed graph which consists of nodes that represent entities within a domain of knowledge, labeled edges which represent relations between these entities, and an inference mechanisms that enables the generation of new relations between entities in the graph. A knowledge graph embedding is a function that maps entities (nodes and edges) in a knowledge graph to vectors within an n-dimensional vector space subject to constraints that aim to preserve certain structural features of the graph within the vector space; several methods to generate knowledge graph embeddings have been developed that primarily differ in the constraints they employ on the mapping function (Bordes *et al.*, 2013; Nickel *et al.*, 2016a; Ristoski and Paulheim, 2016). These methods are used to produce feature vectors for entities represented in a knowledge graph and encode for (parts of) the knowledge about the entity that is represented in a knowledge graph. Knowledge graph embeddings have already been applied successfully in the biological domain to predict relations between biological entities (Alshahrani *et al.*, 2017; Alshahrani and Hoehndorf, 2018). However, ontologies, in particular those in the biomedical domain, cannot easily be represented as graphs (Rodríguez-García and Hoehndorf, 2018); rather, they constitute logical theories that are best represented as sets of axioms (Baader *et al.*, 2003).

Recently, we developed Onto2Vec, a method that generates feature vectors from the formal logical content of ontologies (Smaili *et al.*, 2018), and we could demonstrate that Onto2Vec can outperform existing semantic similarity measures. Here, we extend Onto2Vec to OPA2Vec (Ontologies Plus Annotations to Vectors) to jointly produce vector representations of entities in biomedical ontologies based on both the semantic content of ontologies (i.e. the logical axioms) and the meta-data contained in ontologies as Web Ontology Language (OWL) (Grau *et al.*, 2008; W3C OWL Working Group, 2009) annotation axioms. We combine multiple types of information contained in biomedical ontologies, including asserted and inferred logical axioms, datatype properties and annotation axioms to generate a corpus that consists of both formal statements, natural language statements and other annotation axiom values that relate entities to literals. We then apply a Word2Vec model to generate vector representations for any entity named in the ontology. We further extend our method by incorporating information from biomedical literature. Using transfer learning, we apply a pre-trained Word2Vec model in OPA2Vec to significantly improve the performance in encoding natural language phrases and statements.

We evaluate OPA2Vec using two different ontologies and applications: first, we use the Gene Ontology (GO) (Ashburner *et al.*, 2000) to produce vector representations of yeast and human proteins and determine their functional similarity and predict interactions between them; second, we evaluate our method on the PhenomeNET ontology (Hoehndorf *et al.*, 2011; Rodríguez-García *et al.*, 2017) to infer vector representations of genes and diseases and use them to predict gene–disease associations. We demonstrate that OPA2Vec can produce task-specific and trainable representations of biological entities that significantly outperform both Onto2Vec and traditional semantic similarity measures in predicting protein–protein interactions and gene–disease associations. OPA2Vec is a generic method which can be applied to any ontology formalized in OWL, and OPA2Vec is freely available from `https://github.com/bio-ontology-research-group/opa2vec/`.

2 Materials and methods

2.1 Encoding ontologies plus annotations as vectors

Ontologies formalized in the Web Ontology Language (OWL) (Grau *et al.*, 2008) are based on a Description Logic (Baader *et al.*, 2003). In Description Logics, an ontology is described as the combination of a TBox and an ABox (Horrocks *et al.*, 2006). The TBox is a set of axioms that formally characterize classes (e.g. `behavior SubClassOf: 'biological process'`), while the ABox contains a set of axioms that characterize instances (e.g. `SAMN01832237 instanceOf: Biosample`). The TBox and ABox together are used by the Onto2Vec method (Smaili *et al.*, 2018) to generate dense vector representations; to achieve this goal, Onto2Vec treats asserted or inferred axioms as sentences which form a corpus, and vectors are generated using Word2Vec (Mikolov *et al.*, 2013a,b).

In addition to the TBox and ABox (i.e. to the formal, logical axioms characterizing the domain), ontologies contain a large amount of meta-data in the form of annotation axioms (Hoehndorf *et al.*, 2015b; Smith *et al.*, 2007), and while the axioms are important for automated processing of ontologies, the annotation axioms provide crucial information for humans. OWL annotation axioms relate OWL entities (classes, instances, properties or axioms) to a literal using an OWL annotation property; we call the literal the 'value' of the annotation property (W3C OWL Working Group, 2009). Ontology meta-data consist of the set of non-logical annotation axioms that describe different aspects of ontology classes, relations or instances. For example, most ontologies associate entities with a label, a natural language description, several synonyms, etc. While such meta-data are distinct from the formal content of an ontology and therefore not exploited by methods such as Onto2Vec, they nevertheless provide valuable information about ontology classes, relations and instances.

OPA2Vec (Ontologies Plus Annotations to Vectors) is a novel machine learning method that combines both the formal content of ontologies and the meta-data expressed as OWL annotation axioms to generate feature vectors for any named entity in an ontology; the vectors encode for both the formal and informal content that characterize and constrain the entities in an ontology. OPA2Vec further uses an OWL reasoner, with a choice of either the Elk (Kazakov *et al.*, 2014) or HermiT reasoner (Shearer *et al.*, 2008), to access the deductive closure of an ontology. While HermiT supports the complete OWL 2 DL standard (W3C OWL Working Group, 2009), its worst case complexity is exponential (Horrocks *et al.*, 2006); Elk only supports the OWL 2 EL subset of OWL 2 (W3C OWL Working Group, 2009) but has polynomial complexity and can therefore be applied to larger or more complex ontologies (while losing some of the possible inferences).

Our algorithm generates sentences from OWL annotation axioms to form a corpus. From the assertion that an OWL class C has a

label L (using the rdfs: label annotation property in the OWL annotation axiom) we generate the sentence C rdfs: label L [using the complete Internationalized Resource Identifier (IRI) for C and rdfs: label], and expressing L as string literal. For example, the relation between class *Nuclear periphery* (GO: 0034399) and its label is expressed as the sentence '<http://purl.obolibrary.org/obo/GO_0034399> <http://www.w3.org/2000/01/rdf-schema#label> nuclear periphery'. If C has an annotation axioms relating it to multiple words or sentences, we generate a single sentence in which we ignore sentence or paragraph delimiters. Some annotation properties do not relate entities to strings, but, for example, to dates, numbers or other literals. An ontology may contain information about the creation date of a class or axiom; we also generate sentences from these OWL annotation axioms and render the value of the annotation property as a string. For example, the class *Transcription initiation from RNA polymerase I promoter* (GO: 0006361) has an annotation axiom that relates it to the date it was created within the GO ontology, and we generate the sentence '<http://purl.obolibrary.org/obo/GO_0006361> <http://www.geneontology.org/formats/oboInOwl#: creation_date> 2011-08-15T03'.

In OPA2Vec, we combine the corpus generated from the meta-data (i.e. OWL annotation axioms) and the inferred and asserted logical axioms (using the Onto2Vec algorithm). We then apply a Word2Vec skipgram model on the combined corpus to generate vector representations of all entities in the ontology (for technical details, see Section 2.4).

Natural language words that are used in annotation axioms have a real world linguistic meaning which cannot easily be derived from their use within an ontology alone. Therefore, we use transfer learning in OPA2Vec to assign a semantics to natural language words based on their use in a large corpus of biomedical text. In particular, we pre-train a Word2Vec model on all Medline abstracts, and another model on all open-access fulltext articles available on PubMed Central (PMC), so that natural language words are assigned a semantics (and vector representation) based on their use in biomedical literature (see Section 2.3). The vocabulary in biomedical literature overlaps with the values of annotation properties (e.g. the natural language words used to describe entities in ontologies, or the labels of the entities) but is disjoint with the vocabulary used to refer to the classes, relations and instances in an ontology (which consists of IRIs). In OPA2Vec, we therefore update the pre-trained Word2Vec model to generate vectors for the entities in the ontology, and we update the representations of words that overlap between literature and the ontology annotations.

Supplementary Figure S1 illustrates the OPA2Vec algorithm. The input of the algorithm is an ontology O in OWL format as well as a set A of instances and their associations with classes in the ontology. The output of OPA2Vec is a vector representation for each entity in O and A that encodes for the logical axioms and meta-data in O and A.

2.2 Ontology and annotation resources

We downloaded the Gene Ontology (GO) (Ashburner *et al.*, 2000) in OWL format from http://www.geneontology.org/ontology/ on September 13, 2017. We downloaded the GO protein annotations from the UniProt-GOA website (http://www.ebi.ac.uk/GOA) on September 26, 2017. We removed all annotations with evidence code IEA as well as ND. For validation, we used the STRING database (Szklarczyk *et al.*, 2017) to obtain protein–protein interaction (PPI) data for human (*Homo sapiens*) and yeast (*Saccharomyces cerevisiae*), downloaded on September 16, 2017. The yeast PPI network contains 2 007 135 interactions with 6392 unique proteins, while the human PPI network contains 11 353 057 interactions for 19 577 unique proteins.

We downloaded the PhenomeNET ontology (Hoehndorf *et al.*, 2011; Rodríguez-García *et al.*, 2017) in owl format from the AberOWL repository http://aber-owl.net (Hoehndorf *et al.*, 2015a) on February 21, 2018. We downloaded the mouse phenotype annotations from the Mouse Genome Informatics (MGI) database http://www.informatics.jax.org/ (Smith and Eppig, 2015) on February 21, 2018. We obtained a total of 302 013 unique mouse phenotype annotations. We obtained the disease to human phenotype annotations on February 21, 2018 from the Human Phenotype Ontology (HPO) database http://human-phenotype-ontology.github.io/ (Robinson *et al.*, 2008). We downloaded only the OMIM disease to human phenotype annotations which resulted in a total of 78 208 unique disease–phenotype associations. For gene–disease association prediction validation, we used the MGI_DO.rpt file from the MGI database. This file contains 9506 mouse gene-OMIM disease associations and 13 854 human gene-OMIM disease associations. To map mouse genes to human genes we used the HMD_HumanPhenotype.rpt file from the MGI database; the mapping between mouse and human genes is necessary because gene–disease associations are reported for human genes (in one of our evaluation sets) while the phenotypes and phenotype-based predictions are made for mouse genes.

To process our ontologies (GO and PhenomeNET), we used the OWL API 4.2.6 (Horridge and Bechhofer, 2011) and the Elk OWL reasoner (Kazakov *et al.*, 2014).

2.3 Text corpora

We retrieved the entire collection of article abstracts in the MEDLINE format from the PubMed database https://www.ncbi.nlm.nih.gov/pubmed/ on February 6, 2018. The total number of abstracts collected is 28 189 045. For each abstract, we removed the meta-data (publication date, journal, authors, PMID, etc.), and only kept the title of the article and the text of the abstract for training a Word2Vec model.

PubMed Central (PMC) is a repository provided by the NCBI containing full texts of peer-reviewed journal articles in the life sciences. We have downloaded all the open-access PMC articles on June 10, 2018 which resulted in a total of 4 985 333 full-text articles. We used these articles to pre-train a Word2Vec model that can be compared to the model trained on Medline.

2.4 Word2Vec

We used the ontologies, the entity annotations as well as the Medline abstracts and PMC full-text articles as the text corpora. To process this text data we used Word2Vec (Mikolov *et al.*, 2013a,b). Word2Vec is a machine learning model based on neural networks that can be used to generate vector representations of words in a text. Word2Vec is optimized in such a way that the vector representations of words with a similar context tend to be similar. Word2Vec is available in two different models: the continuous bag of word (CBOW) model and the skip-gram model. In this work, we opted for the skip-gram model which has the advantage over the CBOW model of creating better quality vector representations of words which are infrequent in the corpus. This advantage is quite useful in our case since the biological entities we want to get representations for do not necessarily occur frequently in our text corpora. In this work, we pre-trained the Word2Vec model on the set of PubMed abstracts and save the obtained model which we

eventually retrained on the ontology studied (the GO ontology and the PhenomeNET ontology). We used gridsearch to optimize the set of parameters of the skip-gram model used in this work. We used the same parameters to train Word2Vec on Medline and PMC and the ontologies dataset, except for the *min_count* which has a value of 25 for the pre-trained model on both Medline and PMC, but which we changed to 1 before training on the ontology corpus. The parameters we chose are shown in Supplementary Table S3.

2.5 Similarity

2.5.1 Cosine similarity

To calculate similarity between the vectors produced by Word2Vec, we used the cosine similarity which measures the cosine angle between the two vectors. Cosine similarity cos_{sim} between two vectors A and B is calculated as $cos_{sim}(A,B) = \frac{A \cdot B}{||A||||B||}$ where $A \cdot B$ is the dot product of A and B.

2.5.2 Semantic similarity

Resnik semantic similarity measure (Resnik *et al.*, 1999) is one of the most widely used semantic similarity measures for ontologies. Resnik's measure is based on the notion of information content which quantifies the specificity of a given class in the ontology. The information content of a class c is defined as the negative log likelihood, $-\log p(c)$, where $p(c)$ is the probability of encountering an instance of class c. Resnik similarity is formally defined as $sim(c_1, c_2) = -\log p(c_{MICA})$ where c_{MICA} is the most informative common ancestor of c_1 and c_2 in the ontology taxonomy, defined as the common ancestor with the highest information content. Biological entities can have several concept annotations within an ontology. For instance, as a protein can be involved in different biological processes and can carry several molecular functions, it can be annotated by more than one GO class. Therefore, to calculate semantic similarity between a pair of proteins, we use the best match average strategy (Pesquita *et al.*, 2009).

2.6 Supervised learning and evaluation dataset

To improve our PPI prediction and gene–disease association prediction performance, we used a neural network algorithm to train our prediction model. For our PPI prediction, we used 1015 proteins from the yeast dataset for training and 677 randomly selected proteins for testing while we used 2263 proteins from the human dataset for training and 1509 for testing. We considered as positives the pairs in the STRING database and we randomly sub-sampled negatives among all the pairs not occurring in STRING; we ensure that the cardinality of the positives and negatives are equal for the testing and the training datasets.

When predicting gene–disease associations observed in mouse models, we used 6710 gene–disease associations for training (2030 diseases and all their associations) and 2876 for testing (870 diseases and all their associations); for gene–disease associations observed in humans, we used 9698 associations for training (2978 diseases) and 4196 for testing (1276 diseases). We used the gene–disease associations from the `MGI_DO.rpt` available at MGI; we consider all other associations as negatives.

We evaluate and compare all methods on the same testing data that we obtained through the random selection: 667 yeast proteins (and all their interactions), 1509 human proteins (and all their interactions), 870 diseases from gene–disease associations in mice and 1276 diseases gene–disease associations in human.

We chose our neural network to be a feed-forward network with four layers: the first layer contains 400 input units; the second and third layers are hidden layers which contain 800 and 200 neurons, respectively; and the fourth layer contains one output neuron. We optimized parameters using a limited manual search based on best practice guidelines (Hunter *et al.*, 2012). We optimized the ANN using binary cross entropy as the loss function.

2.7 Text-mining based prediction method

We compare OPA2Vec to the text-mining based prediction method BeFree (Bravo *et al.*, 2014, 2015). BeFree extracts sets of biological associations from scientific articles. We downloaded the BeFree gene–disease prediction from the DisGeNet database (Piñero *et al.*, 2015, 2016) on September 30, 2018. The BeFree gene–disease associations are represented using UMLS concept identifiers. We use the Disease Ontology (DO) (Kibbe *et al.*, 2015; Schriml *et al.*, 2012) to map the UMLS identifiers to OMIM identifiers. Since not all diseases have both an OMIM and a UMLS identifier, we use a limited evaluation set consisting of 1194 diseases shared between BeFree and our evaluation set.

2.8 Evaluation using ROC curve and AUC

To evaluate our PPI and gene–disease prediction, we used the receiver operating characteristic (ROC) curve (Yin and Vogel, 2017) which is a widely used evaluation method to assess the performance of prediction and classification models. It plots the true positive rate (TPR or sensitivity) as a function of the false-positive rate (FPR or 1−specificity). As we do not have genuine negative instances (i.e. pairs of proteins that do not interact, or gene–disease pairs that are not associated) available, we treat all unknown associations as negatives. To compute the (FPR, TPR) pairs, we use the similarity values (or the output of the sigmoid classification layer). When predicting protein–protein interactions, for each protein we rank all other proteins based on their similarity value; when predicting gene–disease associations, we rank, for all diseases, all genes based on their similarity value. We then compute the FPR and TPR for each rank. The similarity value itself is used to rank entities while FPR and TPR is only based on the rank obtained. We report the area under the ROC curve as quantitative measure of the performance of the different methods (Yin and Vogel, 2017).

3 Results

3.1 OPA2Vec performance in predicting interactions between proteins

OPA2Vec is an algorithm that uses asserted and inferred logical axioms in ontologies, combines them with annotation axioms (i.e. meta-data associated with entities or axioms in ontologies) and produces dense vector representations of all entities named in an ontology, or entities associated with classes in an ontology (see Section 2.1 and Supplementary Fig. S1). One of the main applications of ontologies is the computation of semantic similarity (Pesquita *et al.*, 2009). As OPA2Vec combines logical axioms and annotation axioms into single vector representations, we expect that we can obtain more accurate feature vectors for biological entities than using the ontology structure alone, and that we can use this to improve the computation of semantic similarity.

To evaluate our hypothesis and demonstrate the potential of using OPA2Vec, we used the Gene Ontology (GO) as a case study (see Section 2.2). We generated a knowledge base using GO, and added either human proteins or yeast proteins as instances. We related each protein to its functions by asserting that a protein P with function F is an instance of the class `has-function some F`. We applied OPA2Vec on these two knowledge bases (one including

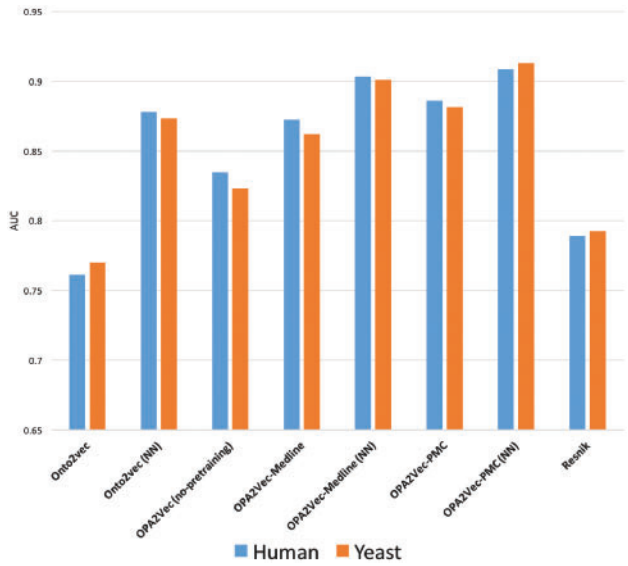

Fig. 1. AUC values of different methods for PPI prediction for yeast and human. `Onto2Vec` uses formal ontology axioms and compares vectors through cosine similarity; `Onto2Vec(NN)` uses a neural network to compare vectors; `OPA2Vec-Medline` is our method and uses formal ontology axioms, entity-class associations and annotation properties from the ontology meta-data (labels, description, synonyms, created_by) with a Word2Vec model pre-trained on Medline, and compares vectors through cosine similarity; `OPA2Vec-Medline(NN)` is `OPA2Vec-Medline` and uses a neural network to determine similarity between two protein vectors; `OPA2Vec-PMC` is similar to `OPA2Vec-Medline` but uses a Word2Vec model pre-trained on fulltext articles in PMC, and compares vectors through cosine similarity; `OPA2Vec-PMC(NN)` is `OPA2Vec-PMC` and uses a neural network to determine similarity between two protein vectors; `OPA2Vec(No pre-training)` uses same strategy as `OPA2Vec` but without a pre-trained Word2Vec model; `Resnik` is a semantic similarity measure

human proteins and the other yeast proteins) and generated vector representations for each protein and ontology class. We then used these vector representations to predict interactions between proteins as characterized in the STRING database (Szklarczyk *et al.*, 2017) by calculating the cosine similarity between each pair of protein vectors and using the obtained value as a prediction score for whether two proteins interact or not. To further improve our prediction performance, we used a neural network model to learn a similarity measure between two feature vectors that is predictive of protein–protein interactions (Smaili *et al.*, 2018). The steps we followed to predict protein–protein interactions using OPA2Vec are illustrated in Supplementary Figure S2. Figure 1 shows the AUC values obtained for OPA2Vec [Supplementary Fig. S3 shows the ROC curves and Supplementary Fig. S4 the precision–recall (PR) curves], and the comparison results against Onto2Vec and Resnik's semantic similarity measure (Resnik *et al.*, 1999) with the Best Match Average strategy (Pesquita *et al.*, 2009) for human and yeast. We found that OPA2Vec significantly improves the performance in predicting interactions between proteins in comparison to both Resnik's semantic similarity measure and Onto2Vec (e.g. improvement between Onto2Vec and OPA2Vec using cosine similarity is significant with $P = 0.031$ and $P = 0.041$ for human and yeast, respectively; one-sided Mann-Whitney U test).

To determine the contribution of each annotation property to the performance of OPA2Vec, we restricted the inclusion of annotation properties to each of the following annotation properties which are most frequently used in GO: label (`rdfs: label`), description (`obo: IAO_0000115`), synonym (`oboInOwl: hasExactSynonym`, `oboInOwl: hasRelatedSynonym`, `oboInOwl: hasBroadpi show $160#Synonym`, `oboInOwl: hasNarrowSynonym`), created by (`oboInOwl: created_by`), creation date (`oboInOwl: creation_date`) and OBO-namespace (`oboInOwl: hasOBOpi show $160#Namespace`). Supplementary Table S1 shows the relative contribution of each of the annotation properties for prediction of protein–protein interactions for human and yeast. We found that the inclusion of the natural language descriptions (`obo: IAO_0000115`) and the class labels (`rdfs: label`) results in the highest improvement of performance, while some annotation properties, such as creation date or the namespace, do not improve the prediction. Interestingly, the `created_by` annotation property adds some minor improvement to the performance. The `created_by` annotation property is used to keep track of the person who is the creator, or original editor, of a class within an ontology. Classes are often created, edited and defined by experts within a particular domain, and the same expert will add similar or related classes to the GO. Therefore, proteins with associations to classes created by the same person may have higher probability to interact due to having more similar functions.

Our analysis shows that annotation properties which describe biological entities in natural language contribute the most to the performance improvements of OPA2Vec. In particular the label and description, synonyms and created-by properties result in better, more predictive feature vector representations. Therefore, we limited our analysis to the labels, descriptions, synonyms and creator name from the ontology meta-data in further analysis (see Supplementary Fig. S5 for comparative results).

Supervised training can significantly improve the predictive performance when using two vector representations for prediction of biological associations as it has the potential to 'learn' custom, task- and dataset-specific similarity measures (Smaili *et al.*, 2018). Therefore, we trained a deep neural network (see Section 2.6) to predict whether two proteins interact given two protein vector representations as inputs. We found that this supervised approach further improves the performance of OPA2Vec (see Fig. 1 and Supplementary Fig. S3).

Furthermore, we performed all experiments twice, comparing OPA2Vec with the pre-trained models from Medline and PMC. We find that, in general, the model that has been trained on the fulltext articles in PMC performs somewhat better than the model that has been trained on Medline abstracts alone.

3.2 Evaluating performance in predicting gene–disease associations

As a second use case to evaluate OPA2Vec and demonstrate its utility, we applied our approach on the PhenomeNET ontology (Rodríguez-García *et al.*, 2017) (see Section 2.2). PhenomeNET is a system for prioritizing candidate disease genes based on the phenotype similarity (Hoehndorf *et al.*, 2011) between a disease and a database of genotype–phenotype associations. Phenotypes refer here to concrete developmental, morphological, physiological or behavioral abnormalities observed in an organism, such as signs and symptoms which make up a disease (Gkoutos *et al.*, 2005, 2017). PhenomeNET includes the PhenomeNET ontology which integrates several species-specific phenotype ontologies; it can therefore be used to compare, for example, phenotypes observed in mouse models and phenotypes associated with human disease (Hoehndorf *et al.*, 2013). We used the PhenomeNET ontology and added mouse genes and human diseases to the knowledge base as instances; we then associated each instance with a set of phenotypes. We used the

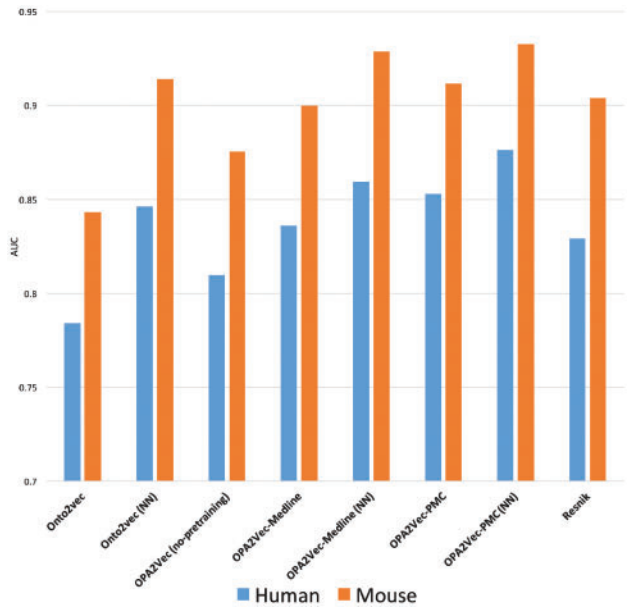

Fig. 2. AUC values for gene–disease association prediction for different methods, using human gene–disease associations (Human) and identified mouse models of human disease (Mouse) as evaluation sets

phenotypes associated with unconditional, single gene knockouts (i.e. complete loss of function mutations) available from the MGI database (Blake *et al.*, 2017) and associated them with their phenotypes, and we used the disease-to-phenotype file from the HPO database (Köhler *et al.*, 2017) to associate diseases from the Online Mendelian Inheritance in Men (OMIM) (Amberger *et al.*, 2011) database to their phenotypes. In total, our knowledge base consists of 18 920 genes and 7154 OMIM diseases.

We applied our OPA2Vec algorithm to the combined knowledge base to generate vector representations of genes and diseases, and use cosine similarity as well as a neural network to predict gene–disease associations. The corpus generated by OPA2Vec therefore consists of the set of asserted and inferred axioms from the PhenomeNET ontology, the set of annotation axioms involving labels, descriptions, synonyms and creators, and the gene and disease phenotype associations.

We then computed the pairwise cosine similarity between gene vectors and disease vectors, and we also trained a neural network in a supervised manner to predict gene–disease associations. We evaluated our results using two datasets of gene–disease associations provided by the MGI database, one containing human disease genes and another containing mouse models of human diseases. Figure 2 shows the AUC values for gene–disease prediction performance of each approach on the human disease genes and mouse models of human disease (see Supplementary Fig. S6 for the ROC curves and Supplementary Fig. S7 for the PR curves). The results utilize only labels, descriptions, synonyms and the `created_by` annotation properties as they contribute positively to prediction of gene–disease associations (see Supplementary Table S2 and Supplementary Fig. S8 for evaluation results using each annotation property). We compared the results to Resnik similarity and Onto2Vec, and found that OPA2Vec outperforms Resnik similarity and Onto2Vec in both evaluation sets. The improvement of OPA2Vec over Onto2Vec using cosine similarity is significant ($P = 0.024$ and $P = 0.026$ for human and mouse; one-sided Mann-Whitney U test), and the improvement of OPA2Vec using cosine similarity over Resnik is significant ($P = 0.0412$ and $P = 0.0307$ for human and mouse; one-sided Mann-Whitney U test). Similar to our results on predicting interactions based on GO functions associated with gene products, we find that using the Word2Vec model trained on PMC performs better than the model trained on Medline abstracts.

4 Discussion

4.1 Related work

OPA2Vec is a method that combined formal and informal content from biomedical ontologies to produce vector representations of biomedical entities. Several methods are emerging that use different types of semantically represented biological data as well as literature to produce similar kinds of vector representations and use them for prediction tasks Beam *et al.* (2018); Newman-Griffis *et al.* (2018); Alshahrani and Hoehndorf (2018). Most of these approaches, including the majority of semantic similarity algorithms (Harispe *et al.*, 2015; Pesquita *et al.*, 2009), are applied to graph-structured data and ignore the axioms that make up many biomedical ontologies. Similarly, ontology-based classification methods that have been developed to predict functions (Kulmanov *et al.*, 2018) or phenotypes (Kahanda *et al.*, 2015) utilize mainly the taxonomy of ontology, ignore other logical axioms as well as all annotation axioms. OPA2Vec is able to utilize the rich metadata, including textual definitions and labels, that are included in ontologies and in which the scientific community has invested significant resources (Smith *et al.*, 2007).

There are several text-mining systems that extract or predict gene–disease or gene–phenotype associations from literature and which rely on ontologies as background knowledge (Kahanda *et al.*, 2015). We have compared OPA2Vec to the BeFree text-mining system (Bravo *et al.*, 2014, 2015) that also identifies gene–disease associations. We limit our evaluation set to those diseases for which both OPA2Vec and BeFree can make predictions and obtain an AUC of 0.7961 for OPA2Vec and 0.7543 for BeFree, demonstrating that OPA2Vec performs better than BeFree ($P = 0.0365$, one-sided Mann-Whitney U test).

4.2 Potential for discovery of novel disease-associated genes

We apply OPA2Vec to re-analyze data obtained from high-throughput (Meehan *et al.*, 2017) and literature-curated (Blake *et al.*, 2017) mouse phenotyping experiments in order to discover new mouse models of human disease as well as candidate genes for human genetically based diseases. The main advantage of OPA2Vec in comparing human and mouse phenotypes is the ability to 'discover' orthologous phenotypes whereas previously applied methods Meehan *et al.* (2017); Hoehndorf *et al.* (2011) generally rely on explicitly encoded background knowledge to determine how phenotypes in mouse and human are related.

We predict candidate genes for over 3000 orphan diseases in OMIM (all predictions are available from our project website), many of which have not received a prediction previously (Meehan *et al.*, 2017). We manually analyzed some of the prediction results for candidate genes of orphan disease predicted by OPA2Vec but none of the competing methods. One of our predictions is E2F transcription factor 5 (E2f5, `MGI: 105091`) for an autosomal dominant variant of hydrocephalus (`OMIM: 123155`). The disease is related to a larger deletion on chromosome 8 (8q12.2-q21.2) where it has been hypothesized that a gene associated with an autosomal dominant form of hydrocephalus can be found (Vincent *et al.*, 1994).

Homozygous E2f5 knockout mice develop nonobstructive hydrocephalus (Lindeman *et al.*, 1998) as well as several other related abnormalities (Danielian *et al.*, 2016); the human ortholog of E2f5 in mice is located in the predicted region at 8q21.2, suggesting a possible involvement of E2F5 in hydrocephalus.

Similarly, we predict an involvement of DiGeorge syndrome critical region gene 8 (Dgcr8, `MGI: 2151114`) in Cayler cardiofacial syndrome (`OMIM: 125520`). Cayler cardiofacial syndrome is associated with deletions in 22q11 (Pasick *et al.*, 2013) and associated with abnormalities in facial features and the cardiovascular system (Cayler, 1969). Cardiomyocyte-specific deletion of Dgcr8 in mice leads to left ventricular malfunction, dilated cardiomyopathy and consequently premature lethality (Rao *et al.*, 2009), and the human ortholog of Dgcr8 is also located specifically at 22q11.21, i.e. in the region to which Cayler cardiofacial syndrome maps, making Dgcr8 a likely candidate for Cayler cardiofacial syndrome. While determining whether these associations are genuine causal relations will require further functional validation, we believe that our predictions are likely candidates as they are further supported by evidence that is not used in OPA2Vec.

4.3 Limitations and future work

Our approach has several limitations, some of which we intend to address as future work. While OPA2Vec can utilize OWL axioms for feature learning and prediction, it does not capture information beyond direct associations well; for these purposes, knowledge graph embeddings are more suitable as they can capture information that is more 'distant' (Nickel *et al.*, 2016b). One interesting approach in the future may be to combine the axiom- and annotation-based methods such as OPA2Vec with knowledge graph embeddings that also rely on Word2Vec such as those using random walks to explore graphs (Alshahrani *et al.*, 2017).

A related topic is the representation patterns that associate biological entities with classes in ontologies. If the entities and their relations to ontology classes are included in an OWL ontology, OPA2Vec will generate a representation for them. However, there are multiple design patterns to express particular types of information, including associations between proteins and their functions or genes and their phenotypes (Hoehndorf *et al.*, 2016; Santana da Silva *et al.*, 2017), and OPA2Vec and similar methods will enable their evaluation not only with regard to quality metrics used to evaluate ontologies (Duque-Ramos *et al.*, 2014) but also with respect to their potential to be used in machine learning and prediction.

Finally, a major limitation in OPA2Vec is the reliance on Word2Vec which is agnostic to the semantics of operators which have a well-defined meaning in OWL. In the future, we expect to find better approaches to utilize the semantics of operators and quantifiers in OWL ontologies.

5 Conclusions

We developed the OPA2Vec method to produce vector representations for biological entities in ontologies based on the formal logical content in ontologies combined with the meta-data and natural language descriptions of entities in ontologies. We applied OPA2Vec to two ontologies, the GO and PhenomeNET, and we demonstrated that OPA2Vec can significantly improve predictive performance in applications that rely on the computation of semantic similarity. We also evaluated the individual contributions of each ontology annotation property to the performance of OPA2Vec-generated vectors. Our results illustrate that the annotation properties that are used to describe details about an ontology class in natural language, in particular the labels and descriptions, contribute most to the feature vectors. We could show that transfer learning, i.e. assigning 'meaning' to words by pre-training a Word2Vec model on a large corpus of biomedical literature abstracts, could further significantly improve OPA2Vec performance in our two applications (prediction of protein–protein interactions and prediction of gene–disease associations). OPA2Vec can comprehensively encode for information in ontologies. Our method is based on accepted standards for encoding ontologies, in particular the Web Ontology Language (OWL), and has the potential to include or exclude any kind of annotation property in the generation of its features. OPA2Vec also exploits major developments in the biomedical ontologies community: the use of ontologies as community standards, and inclusion of both human- and machine-readable information in ontologies as standard requirements for publishing ontologies (Matentzoglu *et al.*, 2018; Smith *et al.*, 2007). We therefore believe that OPA2Vec has the potential to become a highly useful, standard analysis tool in the biomedical domain, supporting any application in which ontologies are being used.

Funding

The research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. FCC/1/1976-04, FCC/1/1976-06, URF/1/2602-01, URF/1/3007-01, URF/1/3412-01, URF/1/3450-01 and URF/1/3454-01.

Conflict of Interest: none declared.

References

Al-Mubaid,H. and Nguyen,H.A. (2006) A cluster-based approach for semantic similarity in the biomedical domain. In: *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, New York, USA, pp. 2713–2717.

Alshahrani,M. and Hoehndorf,R. (2018) Semantic disease gene embeddings (smudge): phenotype-based disease gene prioritization without phenotypes. *Bioinformatics*, **34**, i901–i907.

Alshahrani,M. *et al.* (2017) Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics*, **33**, 2723–2730.

Amberger,J. *et al.* (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM). *Hum Mutat.*, **32**, 564–567.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Baader,F. *et al.* (2003) *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge, UK.

Beam,A.L. *et al.* (2018) Clinical concept embeddings learned from massive sources of medical data. *CoRR.*, abs/1804.01486.

Blake,J.A. *et al.* (2017) Mouse genome database (mgd)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.*, **45**, D723–D729.

Bodenreider,O. (2008) Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb. Med. Inf.*, **2008**, 67.

Bordes,A. *et al.* (2013) Translating embeddings for modeling multi-relational data. In: Burges,C.J.C. *et al.* (eds) *Advances in Neural Information Processing Systems*, Vol. 26. Curran Associates, Inc., Red Hook, NY, USA, pp. 2787–2795.

Bravo,A. *et al.* (2014) A knowledge-driven approach to extract disease-related biomarkers from the literature. *BioMed Res. Int.*, **2014**, 1.

Bravo,À. *et al.* (2015) Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, **16**, 55.

Cayler,G. (1969) Cardiofacial syndrome. congenital heart disease and facial weakness, a hitherto unrecognized association. *Arch. Dis. Child*, **44**, 69–75.

Danielian,P.S. *et al.* (2016) E2f4 and e2f5 are essential for the development of the male reproductive system. *Cell Cycle*, **15**, 250–260.

Duque-Ramos,A. *et al.* (2014) Evaluating the good ontology design guideline (goodod) with the ontology quality requirements and evaluation method and metrics (oquare). *PLoS One*, **9**, 1–14.

Gkoutos,G.V. *et al.* (2005) Using ontologies to describe mouse phenotypes. *Genome Biol.*, **6**, R5.

Gkoutos,G.V. *et al.* (2017) The anatomy of phenotype ontologies: principles, properties and applications. *Briefings in Bioinf.*, **19**, 1008–1021.

Grau,B. *et al.* (2008) Owl 2: the next step for owl. *Web Semant. Sci. Serv. Agents World Wide Web*, **6**, 309–322.

Gruber,T.R. (1995) Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum. Comput. Stud.*, **43**, 5–6.

Harispe,S. *et al.* (2015) *Semantic Similarity from Natural Language and Ontology Analysis*. Morgan & Claypool Publishers, London, UK.

Hoehndorf,R. *et al.* (2013) An integrative, translational approach to understanding rare and orphan genetically based diseases. *Interface Focus*, **3**, 20120055.

Hoehndorf,R. *et al.* (2011) Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res.*, **39**, e119.

Hoehndorf,R. *et al.* (2014) Mouse model phenotypes provide information about human drug targets. *Bioinformatics*, **30**, 719–725.

Hoehndorf,R. *et al.* (2015a) Aber-owl: a framework for ontology-based data access in biology. *BMC Bioinformatics*, **16**, 26.

Hoehndorf,R. *et al.* (2015b) The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinf.*, **16**, 1069–1080.

Hoehndorf,R. *et al.* (2016) Large-scale reasoning over functions in biomedical ontologies. In: Roberta,F. and Werner,K. (eds) *Formal Ontology in Information Systems, Volume 283 of Frontiers in Artificial Intelligence and Applications*. IOS Press, Amsterdam, NL, pp. 299–312.

Horridge,M. and Bechhofer,S. (2011) The owl api: a java api for owl ontologies. *Semant. Web*, **2**, 11–21.

Horrocks,I. *et al.* (2006) The even more irresistible sroiq. In: Doherty,P. *et al.* (eds) *KR*. AAAI Press, Palo Alto, California, USA, pp. 57–67.

Hunter,D. *et al.* (2012) Selection of proper neural network sizes and architectures – a comparative study. *IEEE Trans. Ind. Inf.*, **8**, 228–240.

Jiang,J.J. and Conrath,D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: *Proceeding of the Int'l. Conference on Research in Computational Linguistics*. Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taipei, Taiwan, pp. 19–33.

Kahanda,I. *et al.* (2015) Phenostruct: prediction of human phenotype ontology terms using heterogeneous data sources. *F1000Research*, **4**, 259.

Kazakov,Y. *et al.* (2014) The incredible elk. *J. Autom. Reason.*, **53**, 1–61.

Kibbe,W.A. *et al.* (2015) Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.*, **43**, D1071–D1078.

Köhler,S. *et al.* (2009) Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.*, **85**, 457–464.

Köhler,S. *et al.* (2017) The human phenotype ontology in 2017. *Nucleic Acids Res.*, **45**, D865–D876.

Kulmanov,M. *et al.* (2018) Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, **34**, 660–668.

Leacock,C. and Chodorow,M. (1998) Combining local context and wordnet similarity for word sense identification. *WordNet Electron. Lexical Datab.*, **49**, 265–283.

Li,Y. *et al.* (2003) An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. Knowl. Data Eng.*, **15**, 871–882.

Lin,D. *et al.* (1998) An information-theoretic definition of similarity. In: Jude,W.S. (ed.) *ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 296–304.

Lindeman,G.J. *et al.* (1998) A specific, nonproliferative role for E2F-5 in choroid plexus function revealed by gene targeting. *Genes Dev.*, **12**, 1092–1098.

Matentzoglu,N. *et al.* (2018) Miro: guidelines for minimum information for the reporting of an ontology. *J. Biomed. Semant.*, **9**, 6.

Meehan,T. *et al.* (2017) Disease model discovery from 3, 328 gene knockouts by the international mouse phenotyping consortium. *Nat. Genet.*, **49**, 1231–1238.

Mikolov,T. *et al.* (2013a) Distributed representations of words and phrases and their compositionality. *CoRR.*, abs/1310.4546.

Mikolov,T. *et al.* (2013b) Efficient estimation of word representations in vector space. *CoRR.*, abs/1301.3781.

Newman-Griffis,D. *et al.* (2018) Jointly embedding entities and text with distant supervision. *CoRR.*, abs/1807.03399.

Nickel,M. *et al.* (2016a) Holographic embeddings of knowledge graphs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16*. AAAI Press, Palo Alto, California, USA, pp. 1955–1961.

Nickel,M. *et al.* (2016b) A review of relational machine learning for knowledge graphs. *Proc. IEEE*, **104**, 11–33.

Pasick,C. *et al.* (2013) Asymmetric crying facies in the 22q11.2 deletion syndrome: implications for future screening. *Clin. Pediatr.*, **52**, 1144–1148.

Pesquita,C. *et al.* (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, e1000443.

Piñero,J. *et al.* (2015) Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**, D833–D839.

Piñero,J. *et al.* (2016) Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.

Rao,P.K. *et al.* (2009) Loss of cardiac microrna-mediated regulation leads to dilated cardiomyopathy and heart failure. *Circulation Res.*, **105**, 585–594.

Resnik,P. *et al.* (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res. (JAIR)*, **11**, 95–130.

Ristoski,P. and Paulheim,H. (2016) Rdf2vec: rdf graph embeddings for data mining. In: Paul,G. *et al.* (eds) *International Semantic Web Conference*. pp. 498–514. Springer, Berlin, Heidelberg. Germany.

Robinson,P.N. *et al.* (2008) The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.*, **83**, 610–615.

Rodríguez-García,M.Á. and Hoehndorf,R. (2018) Inferring ontology graph structures using owl reasoning. *BMC Bioinformatics*, **19**, 7.

Rodríguez-García,M.Á. *et al.* (2017) Integrating phenotype ontologies with phenomenet. *J. Biomed. Semant.*, **8**, 58.

Santana da Silva,F. *et al.* (2017) Ontological interpretation of biomedical database content. *J. Biomed. Semant.*, **8**, 24.

Schriml,L.M. *et al.* (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.

Shearer,R. *et al.* (2008) Hermit: a highly-efficient OWL reasoner. In: Catherine,D. *et al.* (eds) *Proceedings of the Fifth OWLED Workshop on OWL: Experiences and Directions*, Ceur-ws.org, Aachen, Germany, p. 91.

Smaili,F.Z. *et al.* (2018) Onto2vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics*, **34**, i52–i60.

Smith,B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.

Smith,C.L. and Eppig,J.T. (2015) Expanding the mammalian phenotype ontology to support automated exchange of high throughput mouse phenotyping data generated by large-scale mouse knockout screens. *J. Biomed. Semant.*, **6**, 11.

Szklarczyk,D. *et al.* (2017) The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.

Vincent,C. *et al.* (1994) A proposed new contiguous gene syndrome on 8q consists of branchio-oto-renal (bor) syndrome, duane syndrome, a dominant form of hydrocephalus and trapeze aplasia; implications for the mapping of the bor gene. *Hum. Mol. Genet.*, **3**, 1859–1866.

W3C OWL Working Group. (2009) Owl 2 web ontology language: Document overview. Technical Report, W3C.

Wu,Z. and Palmer,M. (1994) Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 133–138.

Yin,J. and Vogel,R.L. (2017) Using the roc curve to measure association and evaluate prediction accuracy for a binary outcome. *Biometr. Biostatist. Int. J.*, **5**, 1.