Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 2 April 2015
Applications Note

Genome analysis

Predicting candidate genes from phenotypes, functions, and anatomical site of expression

Jun Chen [1,†] **Azza Althagafi** [1,2,†] **and Robert Hoehndorf**[1,*]

[1] King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical & Mathematical Sciences and Engineering (CEMSE) Division, Thuwal 23955, Saudi Arabia.
[2] Taif University, Faculty of Computing and Information Technology, Taif, 26571, Saudi Arabia

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Over the past years, many computational methods have been developed to incorporate information about phenotypes for disease gene prioritization task. These methods generally compute the similarity between a patient's phenotypes and a database of gene-phenotype to find the most phenotypically similar match. The main limitation in these methods is their reliance on knowledge about phenotypes associated with particular genes, which is not complete in humans as well as in many model organisms such as the mouse and fish. Information about functions of gene products and anatomical site of gene expression is available for more genes and can also be related to phenotypes through ontologies and machine learning models.

Results: We developed a novel graph-based machine learning method for biomedical ontologies which is able to exploit axioms in ontologies and other graph-structured data. Using our machine learning method, we embed genes based on their associated phenotypes, functions of the gene products, and anatomical location of gene expression. We then develop a machine learning model to predict gene–disease associations based on the associations between genes and multiple biomedical ontologies, and this model significantly improves over state of the art methods. Furthermore, we extend phenotype-based gene prioritization methods significantly to all genes which are associated with phenotypes, functions, or site of expression.

Availability: Software and data are available at `https://github.com/bio-ontology-research-group/DL2Vec`.

Contact: robert.hoehndorf@kaust.edu.sa

1 Introduction

Understanding the molecular mechanisms underlying a set of abnormal phenotypes is important for diagnosis, prevention, and development of therapies. Methods to identify and study these mechanisms include observational, experimental, and computational approaches. In particular, in rare diseases, deciphering the mechanisms underlying a set of phenotypes is often limited due to small sample sizes. Computational methods that can reveal or model mechanisms in these diseases often rely on biological background knowledge.

Several computational methods have been developed to prioritize candidate genes for a particular disease or set of abnormal phenotypes (Tranchevent *et al.*, 2016; Tomar *et al.*, 2019; Guala and Sonnhammer, 2017; Zhang *et al.*, 2018; Feng, 2017). Many such methods rely on identifying similarities between genes and suggest new candidates based on such a similarity (Gillis and Pavlidis, 2012). This similarity can be computed on several known features about a gene, including phenotype associations (Greene *et al.*, 2016), distance within an interaction network (Peng *et al.*, 2018), or functional similarity (Liu *et al.*, 2018; Schlicker and Albrecht, 2010)

Phenotype-based methods have been particularly successful in finding candidate genes causing Mendelian diseases (Hoehndorf *et al.*, 2011; Washington *et al.*, 2009; Shefchek *et al.*, 2020). Phenotype-based methods compare disease phenotypes to known genotype–phenotype associations and suggest candidate genes based on phenotype similarity measures (Köhler *et al.*, 2009). While these methods are successful, their main limitation is the incomplete knowledge of phenotypes that are associated with particular genotypes. One approach to overcome this limitation is the use of phenotype associations from model organism experiments together with ontologies that integrate phenotypes across different species (Washington *et al.*, 2009; Hoehndorf *et al.*, 2011; Smedley *et al.*, 2013;

[†] These authors contributed equally.

Bone *et al.*, 2016; Wang *et al.*, 2017a). Although the use of model organisms expanded the scope of prioritizing candidate genes, there is only a limited amount of information about phenotype associations available for genotypes in model organism; furthermore, genes for which there are no orthologs in model organisms cannot benefit from cross-species phenotype-based gene prioritization approaches.

One possibility to overcome the limited information on genotype–phenotype associations is the use of prediction models that predict phenotypes, and efforts such as the Computational Assessment of Function Annotation (CAFA) (Zhou *et al.*, 2019) challenge regularly evaluate function and phenotype prediction models; while function prediction methods have increased significantly in performance and provide accurate predictions at least for some types of functions (Zhou *et al.*, 2019), phenotype predictions still perform worse than function prediction methods (Jiang *et al.*, 2016). Phenotypes arise from a genotype and interactions with the environment (Johannsen, 1911), and predicting the endophenotypes resulting from molecular aberrations requires the use of knowledge about molecular interactions as well as physiological interactions within and between cells, tissues, and organs.

Logical axioms, as used in many phenotype ontologies to formally characterize and standardize phenotype descriptions (Mungall *et al.*, 2010; Köhler *et al.*, 2019; Gkoutos *et al.*, 2018), relate phenotypes systematically to biological functions and anatomical locations, and thereby integrate physiology, anatomy, and abnormal phenotypes within a unifying formal framework (Mungall *et al.*, 2010; Gkoutos *et al.*, 2018; Shefchek *et al.*, 2020). The axioms in phenotype ontologies rely on ontologies that can be applied across different species. In particular, biological processes, functions, and cellular anatomy are described using the Gene Ontology (GO) (Ashburner *et al.*, 2000), and anatomical sites are described using the UBERON anatomy ontology (Mungall *et al.*, 2012); both ontologies are designed to integrate information across different species, and multiple large databases contain information that relate biological entities with classes in these ontologies. Phenotype ontologies therefore not only integrate background knowledge but can also be used to integrate data associated with different ontologies; in particular, they can be used to integrate functions of gene products, anatomical site or tissue of gene expression, and phenotypes resulting from a gene's loss of function.

Using ontologies and the background knowledge they contain in machine learning models can significantly improve their performance (Smaili *et al.*, 2020). Here, we developed an ontology-based machine learning method to prioritize candidate genes based on abnormal phenotypes observed in mouse models, the normal functions of gene products, and anatomical location of gene expression. Our method combines axioms in ontologies and annotations to ontology classes. We evaluate several machine learning methods that utilize ontology axioms, and develop a novel graph-based method that overcomes several limitations of existing methods, in particular when applying machine learning to different ontologies in which classes are not related mainly through subclass axioms but rather through other types of axioms. We demonstrate that our approach improves significantly compared to established phenotype-based gene prioritization methods, and further extends the application of these methods to all genes for which either their functions or their anatomical location of expression is known.

2 Result

2.1 Phenotype-based prioritization of candidate genes

We developed a method based on deep learning to rank candidate causative genes given a set of abnormal phenotypes that characterize a genetically-based disease. We prioritize, or rank, genes based on three distinct types of features that can be associated with a gene: phenotypes associated with the gene's orthologs in the mouse; the functions and cellular locations of the gene products for which the gene encodes; and the anatomical locations at which the gene is expressed. Each of these features is expressed using biomedical ontologies and we use the ontology as part of the learning problem. For this purpose, we first embed the information about genes and diseases together with the ontologies used to characterize them in a vector space and then use a supervised machine learning model to predict whether a gene is causative of a set of phenotypes or disease.

Specifically, we obtain the annotations of human genes with functions and cellular locations encoded by the Gene Ontology (GO) (Ashburner *et al.*, 2000) from the GO Annotation database (Huntley *et al.*, 2014), their anatomical site of expression in functional genomics experiments (GTEx Consortium, 2015) encoded using the UBERON anatomy ontology (Mungall *et al.*, 2012), and the phenotypes of their mouse orthologs from the Mouse Genome Informatics (MGI) database (Smith *et al.*, 2017) and characterized using the Mammalian Phenotype Ontology (MP) (Smith and Eppig, 2009). Furthermore, we obtain phenotype annotations of human diseases with the Human Phenotype Ontology (HPO) (Robinson *et al.*, 2008) from the HPO database (Köhler *et al.*, 2019). To combine the annotations using the different ontologies, we use the integrated PhenomeNET ontology (Rodríguez-García *et al.*, 2017).

We "embed" each gene and disease, their ontology-based annotations, and the ontologies used in the annotations, in a vector space. An embedding is a function from gene or disease identifiers, and from entities in ontologies, into a real-valued vector $\Re^n$ of size n (with n being a parameter of the embedding) such that some properties of the ontologies are preserved in $\Re^n$. Initially, we use the Onto2Vec (Smaili *et al.*, 2018), OPA2Vec (Smaili *et al.*, 2019), methods to generate the embeddings as they have performed well in similar tasks before. We also use SmuDGE (Alshahrani and Hoehndorf, 2018), which generates feature vectors for entities represented in a knowledge graph and encodes for (parts of) the knowledge contained in ontologies. We generate embeddings individually using phenotype, GO, and UBERON annotations; because these annotations are available for different numbers of genes, we also generate a set of embeddings based on the union of all genes and their annotations (i.e., for genes that have annotations from one, two, or all three datasets) as well as another set of embeddings only for genes that have annotations from all three sources.

We use a pointwise learning-to-rank model (see Materials and Methods, and Supplementary Figure 1), to prioritize gene–disease pairs based on gene–disease associations in the Online Mendelian Inheritance in Men (OMIM) database (Amberger *et al.*, 2011). Our model is based on neural networks; given a pair of embedding vectors G and D as input, the model independently transforms the embeddings into a lower-dimensional representations using two fully-connected hidden layers, and then computes the inner product followed by a sigmoid function that outputs a value between 0 and 1, and which we use as the prediction score for an association between G and D.

We train and test our model based on 10-fold cross-validation; in each fold we split our data by the disease (and not by the gene–disease association pair) to ensure that the diseases on which we test have not been seen during training. Within each split, we use 10% of the data as the testing data used to report the final results of our model, and we use the other 90% data to train the model and tune its parameters; within these 90% of training data in each fold, we use a randomly chosen set of 90% for training and 10% for validation. We use sub-sampling of "unknown" associations between genes and diseases to generate negative associations for each disease; we sample 20 negatives for each positive association. We then use binary cross-entropy as the loss function to optimize the ranking model and use the Adam optimizer (Kingma and Ba, 2014) to train our model.

For the evaluation of our learning-to-rank model, we rank all genes for each disease based on their prediction score (within the testing set). We then use the receiver operating characteristic (ROC) curve (Fawcett, 2006) and the area under the ROC curve (ROCAUC) to evaluate how the known positive gene–disease pairs rank among all the possible pairs. Supplementary Figure 2 shows the ROC curves for our prediction model when using different embedding methods, and Table 1 summarizes the results of the cross-validation. We find that we can identify causative genes best when using phenotypes, while the predictive performance decreases when using features derived from gene functions and anatomical site of gene expression.

We hypothesize that one of the reasons for the observed difference in predictive performance between the different data types is the inability of Onto2Vec and OPA2Vec to capture longer distance dependencies through which phenotypes, functions, and anatomical locations are connected within the PhenomeNET ontology. In particular, Word2Vec is equivalent to factorizing a matrix which contains the pointwise mutual information (PMI) (Church and Hanks, 1990) of words within a context window (Levy and Goldberg, 2014), and this measure is only based on directly co-occurring tokens (within the context window considered by Word2Vec). When using Onto2Vec or OPA2Vec, genes and diseases will only directly co-occur with the ontology classes used to characterize them (i.e., phenotypes, GO functions, and UBERON anatomical classes for genes, and phenotypes for diseases), as well as all their superclasses (because Onto2Vec and OPA2Vec compute the transitive closure over the subclass hierarchy and add them to the set of asserted axioms). Consequently, even if a phenotype class is defined based on an anatomical location or a function, this anatomical location or function class will not co-occur with a gene or disease that is associated with this phenotype. For example, the class *Ventricular septal defect* (`HP:0001629`) is defined as an incomplete closure of the *Interventricular septum* (`UBERON:0002094`), which in turn is constrained to be a part of the *Heart* (`UBERON:0000948`) in UBERON. When embedding genes based on their anatomical site of expression (i.e., using the UBERON ontology) and diseases based on their phenotypes, Onto2Vec and OPA2Vec will only add subclass relations as directly co-occurring tokens to use in the embedding but not the classes that are linked indirectly through axioms.

We hypothesize that by incorporating these indirect associations will allow us to better utilize the background knowledge contained in the ontologies and further improve predictive performance, and we develop a novel embedding approach for ontologies that aims to improve the embedding of ontologies with many complex axioms, as well as embeddings of entities which are annotated with classes that do not stand in a subclass relation but are related through more complex axioms.

2.2 Embedding graph-based representations of ontologies

Our novel embedding approach is inspired by the OWL2Vec (Holter *et al.*, 2019) as well as the Walking RDF & OWL (Alshahrani *et al.*, 2017) and SmuDGE (Alshahrani and Hoehndorf, 2018) methods which first convert ontologies into a graph based on syntactic patterns within the ontology axioms, and then apply a knowledge graph embedding (Wang *et al.*, 2017b) on the resulting graph. However, our method extends these approaches to incorporate more complex forms of axioms into the generated graph so that the complexity of the axioms in a cross-species phenotype ontology such as PhenomeNET (Rodríguez-García *et al.*, 2017) can be utilized.

We have defined a transformation function (shown in Table 2 that is used to convert ontology axioms in the Web Ontology Language (OWL) (Grau *et al.*, 2008) format into a graph. The transformation function considers logical operators as well as quantifiers, and converts them into edges (or subject–predicate–object triples) of a graph. The function is applied to all logical axioms in an ontology, determines whether the precondition or preconditions of the function are satisfied for the axiom, and if they are satisfied it adds one or more edges to the graph.

Our transformation considers axioms pertaining to classes (the ontology, or TBox). Associations between a gene G and an associated ontology class C can be modelled in OWL as an axiom `G SubClassOf: has-function some C` (or using some other relation instead of `has-function`) and, as a consequence, a direct edge between G and C will be created through our algorithm. We convert all axioms from the PhenomeNET ontology, and the annotations of gene and disease entities with their ontology classes, into a graph representation using the transformation function in Table 2. After generating the graph, we apply iterated random walks starting at nodes of the graph to generate a corpus, and use Word2Vec (Mikolov *et al.*, 2013) to generate embeddings for nodes and edge labels based on this corpus.

We repeat our supervised training process using our novel embedding method. The results are summarized in Table 1 (left-hand side, “without PPI”) and ROC curves for this task shown in Supplementary Figure 2. While the performance in predicting gene–disease associations using only phenotype annotations is comparable to the predictive performance observed when using Onto2Vec and OPA2Vec, we observe a significant improvement in ROCAUC when using features encoded using GO ($p = 2.02 \times 10^{-125}$, Mann-Whitney U test) and UBERON ($p = 2.77 \times 10^{-152}$, Mann-Whitney U test), indicating that our approach can better capture relations between classes that are related through complex axioms instead of only subclass axioms. The SmuDGE and OWL2Vec methods are more similar to our approach and their performances are closer to our method; however, we still improve over both SmuDGE and OWL2Vec when using only GO and UBERON as features. We also reports recall (hits) at ranks 1, 10, and 100.

2.3 Adding network information

Since our embedding approach is based on graphs and random walks, it can naturally accommodate other graph-structured information in addition to the graph generated from the ontology axioms. There are many biological networks that also relevant to understanding gene–disease associations (Alanis-Lobato *et al.*, 2016; van Dam *et al.*, 2018; Al-Harazi *et al.*, 2016), in particular interaction networks. To determine whether our method is able to utilize this information, we conduct another experiment in which we add functional interactions between proteins obtained from the STRING database (Szklarczyk *et al.*, 2019) to the knowledge base. We add the interactions to our graph as direct `interacts-with` edges between genes (or, equivalently, for an interaction between proteins P_1 and P_2, we add the axiom $P_1 \sqsubseteq \exists$interacts-with.P_2 to the knowledge base and convert them according to the conversion rules in Table 2), and then we repeat our workflow and predict associations between genes and diseases based on the new embeddings (which now also contain information about interactions between genes/proteins in addition to the associations with ontology classes as in the previous experiment). The results of this experiment are shown in Table 1 (right-hand side, “with PPI”) and Supplementary Figure 3, which shows the overall performance obtained from our method using network information and its comparison with embeddings based on other methods.

We find that our workflow results in the best-performing model with regard to ROCAUC, in particular when comparing the embeddings generated using ontologies of different domains, such as when comparing diseases (characterized with phenotypes) and genes characterized by their function or anatomical site of expression.

		Gene Disease Associations without PPI				Gene Disease Associations with PPI			
Methods	**Features**	**ROCAUC**	**Hits@1**	**Hits@10**	**Hits@100**	**ROCAUC**	**Hits@1**	**Hits@10**	**Hits@100**
Onto2Vec	MP	0.903 [0.835-0.931]	0.012	0.207	0.355	0.929 [0.889-0.956]	0.019	0.274	0.440
	GO	0.829 [0.789-0.910]	0.004	0.131	0.315	0.856 [0.832-0.917]	0.008	0.226	0.364
	UBERON	0.700 [0.638-0.739]	0.003	0.080	0.145	0.698 [0.654-0.740]	0.006	0.063	0.118
	Intersection	0.918 [0.885-0.935]	0.012	0.256	0.442	0.928 [0.800-0.959]	0.016	0.300	0.490
	Union	0.958 [0.903-0.975]	0.009	0.229	0.406	0.957 [0.929-0.968]	0.013	0.258	0.445
OPA2Vec	MP	0.907 [0.864-0.923]	0.019	0.204	0.372	0.910 [0.814-0.947]	0.015	0.250	0.427
	GO	0.826 [0.802-0.862]	0.015	0.194	0.313	0.841 [0.821-0.880]	0.009	0.219	0.354
	UBERON	0.717 [0.650-0.784]	0.017	0.127	0.208	0.727 [0.702-0.759]	0.006	0.114	0.191
	Intersection	0.920 [0.880-0.936]	0.011	0.256	0.445	0.928 [0.812-0.948]	0.020	**0.304**	0.468
	Union	0.954 [0.937-0.967]	0.008	0.197	0.378	0.959 [0.943-0.965]	0.013	0.248	0.457
OWL2Vec	MP	0.955 [0.920-0.968]	0.025	0.282	0.602	0.958 [0.943-0.970]	0.030	0.250	0.627
	GO	0.899 [0.805-0.934]	0.021	0.216	0.564	0.903 [0.843-0.937]	0.027	0.258	0.574
	UBERON	0.800 [0.703-0.875]	0.025	0.162	0.454	0.843 [0.809-0.928]	0.023	0.215	0.501
	Intersection	0.948 [0.898-0.975]	0.031	0.225	0.618	0.953 [0.901-0.986]	0.032	0.275	0.629
	Union	0.968 [0.953-0.978]	0.029	0.248	0.617	0.971 [0.963-0.985]	0.025	0.245	0.628
SmuDGE	MP	0.957 [0.941-0.974]	**0.042**	0.225	0.614	0.956 [0.930-0.969]	0.051	0.246	0.632
	GO	0.894 [0.851-0.925]	0.025	0.243	0.551	0.904 [0.835-0.953]	0.029	0.253	0.556
	UBERON	0.815 [0.696-0.879]	0.038	0.194	0.460	0.847 [0.804-0.881]	0.032	0.200	0.507
	Intersection	0.956 [0.931-0.980]	0.027	0.248	0.641	0.951 [0.935-0.960]	0.024	0.233	0.611
	Union	0.977 [0.969-0.985]	0.024	0.244	0.615	0.973 [0.952-0.986]	0.028	0.214	0.582
DL2Vec (our model)	MP	0.957 [0.931-0.971]	0.026	**0.284**	0.638	0.959 [0.941-0.994]	**0.052**	0.258	**0.654**
	GO	0.910 [0.892-0.947]	0.023	0.235	0.579	0.921 [0.896-0.942]	0.044	0.273	0.585
	UBERON	0.842 [0.685-0.898]	0.020	0.155	0.470	0.854 [0.821-0.889]	0.019	0.190	0.515
	Intersection	0.954 [0.928-0.977]	0.032	0.252	**0.642**	0.956 [0.931-0.978]	0.036	0.295	0.650
	Union	**0.978 [0.967-0.998]**	0.024	0.244	0.635	**0.976 [0.948-0.989]**	0.037	0.255	0.637

Table 1. Evaluation results for predicting gene–disease associations using embeddings generated from the Mammalian Phenotype (MP) ontology, Gene Ontology (GO), and UBERON anatomy ontology. The intersection represents embeddings generated jointly from all three types of ontologies and associations, limited to genes that have associations to all three ontologies, while union represents embeddings generated jointly from all three types of ontologies and associations, limited to genes that have associations in at least one of the three ontologies. For ROCAUC, we report the intervals obtained from cross-validation. The results "without PPI" use a graph based on ontology axioms and associations between genes/diseases and ontology classes; the results "with PPI" also include functional interactions between genes/proteins as pat of the graph.

Condition 1	Condition 2	Triple(s)
$A \sqsubseteq QR_0 \ldots QR_m D$ $A \equiv QR_0 \ldots QR_m D$	$D := B_1 \sqcup \ldots \sqcup B_n \mid B_1 \sqcap \ldots \sqcap B_n$	$\langle A, (R_0 ... R_m), B_i \rangle$ for $i \in 1 \ldots n$
$A \sqsubseteq B$		$\langle A, SubClassOf, B \rangle$
$A \equiv B$		$\langle A, EquivalentTo, B \rangle$

Table 2. The transformation rules to convert axioms in ontology O into a graph using DL2Vec. Q represents an arbitrary quantifier or cardinality restriction, A and B_i represent arbitrary class names, and R_i represent arbitrary relation names. Our algorithm iterates through all axioms in O and determines whether the conditions are satisfied; if the condition or conditions are satisfied, the corresponding triple or triples are added to the resulting graph. For example, 'feeding behavior' $\sqsubseteq$ behavior will result in the triple ⟨'feeding behavior', $SubClassOf$, behavior⟩ being added. The first rule captures more complex axioms where multiple relations could be used as part of the axiom and D is either a class name or a complex class description (union or intersection); in the latter case, multiple triples are added to the resulting graph.

3 Discussion

We designed a novel method to prioritize candidate genes given a set of abnormal phenotypes associated with a genetically-based disease; our method uses information about genes obtained from animal model phenotypes, the functions of gene products, the anatomical location of gene expression, and interaction networks, as well as a large amount of background knowledge contained in biomedical ontologies. Our method improves over other phenotype-based methods in several ways.

First, we use a pointwise learning-to-rank machine learning model which improves the predictive performance when evaluated using gene–disease associations from the Online Mendelian Inheritance in Men (OMIM) (Amberger *et al.*, 2011) database; our model is designed to directly learn the similarities between two embeddings and results in improved predictive performance when compared to other models (Smaili *et al.*, 2019, 2018) used to predict gene–disease associations based on embeddings.

Second, we developed a novel method to exploit complex axioms by converting them into a graph and relying on graph embeddings; we show that this approach improves performance significantly when embedding multiple ontologies that are only linked through complex axioms. This advance is particularly important in ontologies that are heavily formalized using OWL and that are interlinked, such as the ontologies in the collaborative OBO Foundry effort (Smith *et al.*, 2007). For example, using DL2Vec, we are able to prioritize the association between a Mendelian form of cataract (`OMIM:604307`) and the gene CRYGC within the first two ranks when incorporating the GO, whereas OPA2Vec and Onto2Vec rank this gene below rank 1,000. One of the key phenotypes of cataract is *visual impairment* (`HP:0000505`) which is defined, in the HPO, as a decreased *visual perception* (`GO:0007601`); based on this formal definition, DL2Vec creates an edge between *visual impairment* and *visual perception*. The gene CRYGC is associated with the GO class *visual perception*. When performing the iterated random walks from either the disease node or the gene node, we find that multiple walks use this edge and therefore lead to a direct co-occurrence of both the disease and the gene with the nodes representing *visual impairment* as well as *visual perception*; applying Word2Vec on these walks results in the gene embedding and disease embedding to become more similar to each other and allows DL2Vec to prioritize the association at one of the top ranks.

Third, our method prioritizes candidate genes for a set of abnormal phenotypes using a combination of gene expression, function, network, phenotype data, and ontologies. In contrast to methods that rely on knowledge about disease-associated genes in order to prioritize new candidates, the input to our method are only the phenotypes observed in a patient. In our approach, prioritization of candidate genes does not rely on knowledge (or existence) of other genes associated with the same phenotypes. We achieve this by combining the different annotations on two distinct levels: first, the different annotations (phenotype, function, expression) are combined on the level of a gene or gene product (which we do not distinguish), so that a single entity (the gene and its products) is associated with all three types of information; second, we also utilize the links between ontologies directly. The links between the classes in ontologies allow us to establish new relations between the different features associated with genes, and these features are not accessible without utilizing the ontology axioms. This makes our approach applicable to Mendelian disease for which no genes may be known to be associated (or where only a single gene is associated), and where features of known disease-associated genes could not be used to identify novel causative genes. While approaches based on the guilt-by-association principle generally perform well on diseases or phenotypes with several known associated genes (Chen *et al.*, 2009; Tranchevent *et al.*, 2016; Gillis and Pavlidis, 2012; Singleton *et al.*, 2014; Schlicker and Albrecht, 2010), our method has a broader range of application.

Fourth, while there are several phenotype-based methods that are applied widely for prioritizing candidate genes (Smedley *et al.*, 2013; Cornish *et al.*, 2018; Köhler *et al.*, 2009), they are limited to genes with associated phenotypes. As there are only a limited number of human genes with associated phenotypes, this set of genes can be expanded significantly by incorporating phenotypes of human orthologs in animal models (Smedley *et al.*, 2013); however, even using animal model phenotypes will leave about half of human genes without any phenotype associations, either due to lack of phenotype associations in animal models or due to the absence of orthologs for a human gene (Shefchek *et al.*, 2020). We significantly expand phenotype-based gene prioritization methods to genes that have either phenotype associations, are associated with GO functions, or have known sites of expression. While the predictive performance of our method is lower for genes that do not have phenotype associations than for genes with associated phenotypes, we show that we can nevertheless identify disease-associated genes by comparing phenotypes to gene functions or to anatomical locations.

Additionally, our model is extensible and can include additional features if they can be encoded using ontologies. For example, we can expand our model using gene expression in individual celltypes, using the Celltype Ontology (CL) (Bakken *et al.*, 2017). We experimented using single-cell RNAseq data from the Tabula Muris project (The Tabula Muris Consortium *et al.*, 2018) in which genes are annotated with the CL. From this dataset, we obtain 17,149 associations between genes and one or more classes from CL. We added the CL annotation of genes as well as the disease phenotype annotations and performed the same experiments as for the other three ontologies. Without including the functional interactions between genes, we obtain a ROCAUC of 0.906 $(0.883 - 00.949)$ for predicting gene–disease associations (Hits@1, Hits@10, and Hits@100 are 0.037, 0.299, and 0.634, respectively). These results show that single cell gene expression can provide more information for predicting gene–disease associations than tissue-level gene expression encoded using Uberon. One key limitation in using celltype-specific gene expression is that CL is used in fewer axioms within phenotype ontologies (compared to Uberon or GO), and therefore our method will not exploit relations between phenotypes and celltypes as well as relations between the other ontologies.

Our method still has several limitations. Our conversion from OWL into a graph does not consider all OWL axioms, and the conversion also treats different types of restrictions and axiom types identically although their semantics is different. In the future, we plan to extend the method to convert any OWL axioms into a graph representation, relying, for example, on relational patterns defined in the OBO Relation Ontology (Smith *et al.*, 2007), and also rely on inferred axioms for generating the graph such as implemented in the Onto2Graph method (Rodríguez-García and Hoehndorf, 2018).

Another major limitation of our approach is that it is inherently transductive and not inductive. In particular, the diseases with their phenotype associations must be known in our workflow before generating embeddings and training our prediction model, and it is not straightforward to apply the approach to a new set of phenotypes (such as the phenotypes observed in an individual). This limitation is shared by many graph embedding and knowledge graph embedding approaches (Wang *et al.*, 2017b). However, this limitation can be overcome either through the use of inductive methods for learning on knowledge graphs, such as graph neural networks (Scarselli *et al.*, 2008; Kipf and Welling, 2016), or by including patients with their phenotypes as part of the original data (or graph), training the model on gene–disease associations and applying it to predict candidate genes for the patient nodes. However, extending our approach to an inductive setting will allow for an easier combination of our approach with methods to find pathogenic causative variants based on observed phenotypes and next generation sequencing data (Robinson *et al.*, 2014; Boudellioua *et al.*, 2017).

Finally, we treat all genes that are not known to be associated with the disease as negatives and consequently have many more negative than positive associations. This has two consequences; first, we may incorrectly classify an association as negative when a gene is associated with the disease but this association is not yet known. Second, while the overall predictive performance of our method improves over the state of the art and ROCAUC is usually above 0.9 in our evaluation, the recall at the first ranks is still low and rarely exceeds 5% at the first rank. The reason for this difference between the evaluation measures is the imbalanced dataset we use for evaluation, where all genes not known to be associated with a disease are considered negative for that disease. Our evaluation therefore does not consider any additional knowledge about potential associations between a gene and disease. However, in a realistic scenario in which new genes are evaluated for their association with a Mendelian disease, more information is usually available, either from evaluating the pathogenicity of variants found in affected individuals, filtering by pedigree and mode of inheritance, or filtering by variants found in unrelated individuals with the same phenotypes; after such a workflow, usually less than 100 genes remain as potential candidates (Alfares *et al.*, 2020) (in contrast to 9,886 in our evaluation), and recall at top ranks will improve.

4 Conclusions

We developed a method for prioritizing candidate genes given a set of phenotypes associated with a disease. Our method can utilize different types of features characterized through ontologies, and significantly improves the phenotype–based prediction of disease genes. While previous phenotype–based gene prioritization methods are only applicable when phenotype associations are known for genes, our method can be applied to a much larger number of genes for which either functions, sites of expression, phenotypes, or interactions with other genes are known.

5 Materials and Methods

5.1 Ontology and annotation resources

We downloaded Gene Ontology (GO) (Ashburner *et al.*, 2000) annotations of 18,495 human gene products (495,719 annotations in total) from the Gene Ontology website on 2020-03-20. We excluded the GO annotations where the evidence code indicated that the annotation was inferred from electronic annotation (IEA) or for which no biological data is available (ND).

We obtained phenotype annotations for 13,529 mouse genes, including 228,214 associations between genes and Mammalian Phenotype Ontology (Smith and Eppig, 2009) classes, from the file `MGI_GenePheno.rpt` available at the Mouse Genome Informatics (MGI) database (Smith *et al.*, 2017). Phenotype associat ions were downloaded on 2020-03-20. We map each mouse gene to their human ortholog using the file `HMD_HumanPhenotype.rpt` available at the MGI database, resulting in 9,879 human genes where the mouse ortholog has phenotype associations.

We further downloaded the Tissue Expression Profiles (GTEx) dataset (GTEx Consortium, 2015) from the Gene Expression Atlas (Papatheodorou *et al.*, 2020) on 2020-03-20. GTEx characterizes gene expression across 53 tissues. We map the Ensembl protein identifiers to Entrez gene identifiers using the mapping provided by the Entrez database (Maglott *et al.*, 2010). We set a threshold for whether a gene is expressed or not in a tissue by setting a cutoff of 4.0 transcripts per million (TPM); this threshold is determined experimentally (see Supplementary Figure 4). Finally we obtained 20,538 Entrez genes which have expression above this threshold in one or more tissue. We map each tissue to the Uberon Anatomy Ontology (Mungall *et al.*, 2012), downloaded from the AberOWL ontology repository on 2020-03-20. We exclude the expression in *EBV-transformed lymphocyte* and *transformed skin fibroblast*, since these two tissues are not available in the Uberon ontology.

The PhenomeNET Ontology (Rodríguez-García *et al.*, 2017) is a cross-species ontology which integrates multiple species-specific phenotype ontologies as well as related ontologies such as the Gene Ontology and the Uberon Anatomy Ontology. We downloaded the PhenomeNET ontology from the AberOWL ontology repository on March 20, 2020.

5.2 Evaluation datasets

We obtain associations between 2,542 human diseases and 2,885 genes from the file `MGI_DO.rpt` available at the MGI database, downloaded on 2020-03-20; the dataset contains 4,051 gene–disease associations in total, where diseases are represented using their OMIM identifier (Amberger *et al.*, 2011).

As our gene–phenotype associations are to mouse genes (resulting from a loss of function of that gene) while our evaluation set uses human gene identifiers, we need to identify human orthologs of the mouse genes. We identify the mouse orthologs of human genes, and human orthologs of mouse genes, using the file `HMD_HumanPhenotype.rpt` at the MGI database, downloaded on 2020-03-20. Supplementary Table 1 summarizes our training and evaluation data. For each type of feature, there is a different number of associated genes, and consequently a different number of gene–disease associations we can identify; most disease-associated genes have features in all three datasets.

We use functional interactions between proteins obtained from the STRING database (Szklarczyk *et al.*, 2019) on March 01, 2020. The interaction dataset contains 19,355 proteins and 11,759,455 edges between them. We mapped the proteins to the UniProt database and filter out those entries that did not map to the UniProt database. Further, STRING provides a confidence score for an interaction and we only keep interactions with a confidence of at least 700. The remaining interaction network consists of 17,178 proteins with 840,672 interactions.

5.3 Embedding methods

Onto2Vec (Smaili *et al.*, 2018) is a method to learn the semantic embedding representations of biological entities and by extracting features from ontology–based annotations, axioms and ontology structures. It directly utilizes the axiom features and also indirectly infer new logical axiom features by applying the HermiT OWL reasoner (Motik *et al.*, 2009). Onto2Vec collects data and axioms as "sentences" and uses a skip–gram model to learn the vector representation for each word. OPA2Vec (Smaili *et al.*, 2019) is an extension of Onto2Vec which includes the annotation axioms in ontologies and uses transfer learning to assign them a semantics.

A random walk of length k on a graph $G = (V, E)$ is a sequence of nodes and edges $n_1, e_1, \ldots, e_{k-1}, n_k$ such that for all i, $1 \leq i < k$, $e_i \equiv (n_i, n_{i+1}) \in E$ and n_{i+1} was chosen randomly from all neighbors of n_i. Here, we also include edge labels in the walk. For the purpose of selecting neighboring nodes, we treat the graph as undirected. We generate 80 walks from each node, and stop the walks after 20 steps. When including functional interactions between proteins in our knowledge base, we increased the number of walks to 200 and stop the walks after 30 steps to account for the higher node degree.

We implement our embedding algorithm in a software called DL2Vec and make the source code, together with our experiments, freely available under the GNU General Public License version 3.

5.4 Word2Vec model

Word2Vec (Mikolov *et al.*, 2013) is a language model for learning vector representations of words based on co-occurrence within a context window. We use the skipgram model of Word2Vec (Mikolov *et al.*, 2013). Given a sentence with N words, the skipgram model reads the sentence with a window kernel size c and maximizes the co-occurrence probability of words that appear in the same window.

We apply the SkipGram algorithm on our node and edge sequence corpus, which is generated by a random walk on the heterogeneous graph. We set the skipgram parameters to a window size of 10, and min_count value to 1. The training process iterates 20 times, and it outputs a 200 dimensional embedding for each entity.

5.5 Pointwise learning-to-rank prediction model

We use a pointwise learning-to-rank model to predict associations between genes and diseases. The model takes two vectors V_1 and V_2 as input for two independent neural networks ν_1 and ν_2. We then calculate the inner product of $\nu_1(V_1)$ and $\nu_2(V_2)$ and use a sigmoid function to obtain a similarity score between V_1 and V_2. We train this model using binary cross entropy as loss function. Each neural network ν_1 and ν_2 consists of two hidden layers, the first with 256 neurons and the second with 50. We use 20% dropout (Srivastava *et al.*, 2014) after each layer, followed by a LeakyReLU (Xu *et al.*, 2015) as the activation function. The model parameters are optimized using the Adam (Kingma and Ba, 2014) optimizer.

5.6 Evaluation

We use the Receiving Operating Characteristic (ROC) curve (Fawcett, 2006) to assess the performance of our classification model. The ROC curve is a plot of the true positive rate as a function of the false positive rate. To compute true positive and false positive rate, we rank all genes for each disease, and compute the average true and false positive rates at each rank. We then generate the ROC curve, and compute the area under the ROC curve, as the averages across all diseases. We also report the recall at rank n (Hits@n).

We compute differences in the area under the ROC curve using the non-parametric Mann Whitney U test (Nachar *et al.*, 2008). For the test, we test the significance of ranking true positive associations differently between two prediction models. We consider differences as significant if $p < 0.05$. In order to compare the performance of the embeddings generated from phenotypes (MP), gene expression (UBERON), and biological functions (GO) directly, we focus on genes which have annotations to all three ontologies as evaluation set; the number of genes that have annotations in all three ontologies is 9,886.

Funding

This work was supported by funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3454-01-01, URF/1/3790-01-01, FCC/1/1976-08-01, and FCC/1/1976-08-08.

References

Al-Harazi, O. *et al.* (2016). Integrated genomic and network-based analyses of complex diseases and human disease network. *Journal of Genetics and Genomics*, **43**(6), 349–367.

Alanis-Lobato, G. *et al.* (2016). Hippie v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic acids research*, page gkw985.

Alfares, A. *et al.* (2020). What is the right sequencing approach? Solo VS extended family analysis in consanguineous populations. *BMC Medical Genomics*, **13**(1).

Alshahrani, M. and Hoehndorf, R. (2018). Semantic disease gene embeddings (smudge): phenotype-based disease gene prioritization without phenotypes. *Bioinformatics*, **34**(17), i901–i907.

Alshahrani, M. *et al.* (2017). Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics*, **33**(17), 2723–2730.

Amberger, J. *et al.* (2011). A new face and new challenges for online mendelian inheritance in man (OMIM). *Human mutation*, **32**(5), 564–567.

Ashburner, M. *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**(1), 25–29.

Bakken, T. *et al.* (2017). Cell type discovery and representation in the era of high-content single cell phenotyping. *BMC bioinformatics*, **18**(17), 7–16.

Bone, W. P. *et al.* (2016). Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genetics in Medicine*, **18**(6), 608–617.

Boudellioua, I. *et al.* (2017). Semantic prioritization of novel causative genomic variants. *PLOS Computational Biology*, **13**(4), 1–21.

Chen, J. *et al.* (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, **37**(suppl_2), W305–W311.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, **16**(1), 22–29.

Cornish, A. J. *et al.* (2018). PhenoRank: reducing study bias in gene prioritization through simulation. *Bioinformatics*, **34**(12), 2087–2095.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(8), 861–874.

Feng, B.-J. (2017). Perch: a unified framework for disease gene prioritization. *Human mutation*, **38**(3), 243–251.

Gillis, J. and Pavlidis, P. (2012). "Guilt by Association" is the exception rather than the rule in gene networks. *PLOS Computational Biology*, **8**(3), 1–13.

Gkoutos, G. V. *et al.* (2018). The anatomy of phenotype ontologies: principles, properties and applications. *Briefings in Bioinformatics*, **19**(5), 1008–1021.

Grau, B. C. *et al.* (2008). OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, **6**(4), 309–322.

Greene, D. *et al.* (2016). Phenotype similarity regression for identifying the genetic determinants of rare diseases. *The American Journal of Human Genetics*, **98**(3), 490–499.

GTEx Consortium (2015). The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**(6235), 648–660.

Guala, D. and Sonnhammer, E. L. (2017). A large-scale benchmark of gene prioritization methods. *Scientific reports*, **7**, 46598.

Hoehndorf, R. *et al.* (2011). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, **39**(18), e119–e119.

Holter, O. M. *et al.* (2019). Embedding OWL ontologies with OWL2Vec. In *CEUR Workshop Proceedings*, volume 2456, pages 33–36.

Huntley, R. P. *et al.* (2014). The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Research*, **43**(D1), D1057–D1063.

Jiang, Y. *et al.* (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, **17**(1), 1–19.

Johannsen, W. (1911). The genotype conception of heredity. *The American Naturalist*, **45**(531), 129–159.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Köhler, S. *et al.* (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, **85**(4), 457–464.

Köhler, S. *et al.* (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic acids research*, **47**(D1), D1018–D1027.

Köhler, S. *et al.* (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, **85**(4), 457 – 464.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Liu, W. *et al.* (2018). Gene ontology enrichment improves performances of functional similarity of genes. *Scientific reports*, **8**(1), 1–12.

Maglott, D. *et al.* (2010). Entrez gene: gene-centered information at NCBI. *Nucleic acids research*, **39**(suppl_1), D52–D57.

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

Motik, B. *et al.* (2009). Hypertableau Reasoning for Description Logics. *Journal of Artificial Intelligence Research*, **36**, 165–228.

Mungall, C. J. *et al.* (2010). Integrating phenotype ontologies across multiple species. *Genome biology*, **11**(1), R2.

Mungall, C. J. *et al.* (2012). Uberon, an integrative multi-species anatomy ontology. *Genome biology*, **13**(1), R5.

Nachar, N. *et al.* (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, **4**(1), 13–20.

Papatheodorou, I. *et al.* (2020). Expression atlas update: from tissues to single cells. *Nucleic Acids Research*, **48**(D1), D77–D83.

Peng, J. *et al.* (2018). Measuring phenotype-phenotype similarity through the interactome. *BMC bioinformatics*, **19**(5), 114.

Robinson, P. N. *et al.* (2008). The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, **83**(5), 610–615.

Robinson, P. N. *et al.* (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, **24**(2), 340–348.

Rodríguez-García, M. Á. and Hoehndorf, R. (2018). Inferring ontology graph structures using OWL reasoning. *BMC bioinformatics*, **19**(1), 7.

Rodríguez-García, M. Á. *et al.* (2017). Integrating phenotype ontologies with PhenomeNET. *Journal of biomedical semantics*, **8**(1), 58.

Scarselli, F. *et al.* (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, **20**(1), 61–80.

Schlicker, A. and Albrecht, M. (2010). FunSimMat update: new features for exploring functional similarity. *Nucleic acids research*, **38**(suppl_1), D244–D248.

Shefchek, K. A. *et al.* (2020). The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research*, **48**(D1), D704–D715.

Singleton, M. V. *et al.* (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *The American Journal of Human Genetics*, **94**(4), 599–610.

Smaili, F. Z. *et al.* (2018). Onto2vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics*, **34**(13), i52–i60.

Smaili, F. Z. *et al.* (2019). Opa2vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics*, **35**(12), 2133–2140.

Smaili, F. Z. *et al.* (2020). Formal axioms in biomedical ontologies improve analysis and interpretation of associated data. *Bioinformatics*, **36**(7), 2229–2236.

Smedley, D. *et al.* (2013). Phenodigm: analyzing curated annotations to associate animal models with human diseases. *Database*, **2013**.

Smith, B. *et al.* (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, **25**(11), 1251.

Smith, C. L. and Eppig, J. T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, **1**(3), 390–399.

Smith, C. L. *et al.* (2017). Mouse genome database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Research*, **46**(D1), D836–D842.

Srivastava, N. *et al.* (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, **15**(1), 1929–1958.

Szklarczyk, D. *et al.* (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, **47**(D1), D607–D613.

The Tabula Muris Consortium *et al.* (2018). Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, **562**(7727), 367.

Tomar, S. *et al.* (2019). Specific phenotype semantics facilitate gene prioritization in clinical exome sequencing. *European Journal of Human Genetics*, **27**(9), 1389–1397.

Tranchevent, L.-C. *et al.* (2016). Candidate gene prioritization with Endeavour. *Nucleic acids research*, **44**(W1), W117–W121.

van Dam, S. *et al.* (2018). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, **19**(4), 575–592.

Wang, J. *et al.* (2017a). MARRVEL: integration of human and model organism genetic resources to facilitate functional annotation of the human genome. *The American Journal of Human Genetics*, **100**(6), 843–853.

Wang, Q. *et al.* (2017b). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, **29**(12), 2724–2743.

Washington, N. L. *et al.* (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology*, **7**(11).

Xu, B. *et al.* (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Zhang, Y. *et al.* (2018). Prioritizing disease genes with an improved dual label propagation framework. *BMC bioinformatics*, **19**(1), 47.

Zhou, N. *et al.* (2019). The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, **20**(1), 1–23.