Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

DeepPheno: Predicting single gene knockout phenotypes

Maxat Kulmanov [1], and Robert Hoehndorf[1,*]

[1]Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences & Engineering Division, King Abdullah University of Science and Technology, 4700 King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Predicting the phenotypes resulting from molecular perturbations is one of the key challenges in genetics. Both forward and reverse genetic screen are employed to identify the molecular mechanisms underlying phenotypes and disease, and these resulted in a large number of genotype–phenotype association being available for humans and model organisms. Combined with recent advances in machine learning, it may now be possible to predict human phenotypes resulting from particular molecular aberrations.

Results: We developed DeepPheno, a method for predicting the phenotypes resulting from complete loss-of-function in single genes. DeepPheno uses the functional annotations with gene products to predict the phenotypes resulting from a loss-of-function; additionally, we employ a two-step procedure in which we predict these functions first and then predict phenotypes. This allows us to predict phenotypes associated with any known protein-coding gene. We evaluate our approach using evaluation metrics established by the CAFA challenge and compare with top performing CAFA 2 methods. Our method achieves an $F_{\max}$ of 0.46 which is a significant improvement over state-of-the-art $F_{\max}$ of 0.36. Furthermore, we show that predictions generated by DeepPheno are applicable to predicting gene–disease associations based on comparing phenotypes, and 60% of predictions made by DeepPheno interact with a gene that is already associated with the predicted phenotype.

Availability: `https://github.com/bio-ontology-research-group/deeppheno`

Contact: robert.hoehndorf@kaust.edu.sa

1 Introduction

Many human diseases have a genetic basis and are caused by abnormalities in the genome. Due to their high heterogeneity, many disorders are still undiagnosed and despite significant research their genetic basis has not yet been established (Tifft and Adams, 2014). Understanding how disease phenotypes evolve from an organism's genotype is a significant challenge.

Reverse genetic screens can be used to investigate the causality of perturbing molecular mechanisms on a genetic level and observing the resulting phenotypes (Manis, 2007). For example, the International Mouse Phenotyping Consortium (IMPC) (Collins *et al.*, 2007) aims to associate phenotypes with loss of function mutations using gene knockout experiments, and similar knockout experiments have been performed in several model organisms. Further genotype–phenotype associations for the laboratory mouse and other model organisms are also systematically extracted from literature and recorded in model organism databases (Smith and Eppig, 2015; Bult *et al.*, 2018; Cherry *et al.*, 2011)

Similarity between observed phenotypes can also be used to infer similarity between molecular mechanisms, even across different species (Washington *et al.*, 2009). In humans, the Human Phenotype Ontology (HPO) (Köhler *et al.*, 2017) provides an ontology for characterizing phenotypes and a wide range of genotype–phenotype associations have been created based on the human phenotype ontology. Further information about genotype–phenotype associations is collected in databases such as Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.*, 2005), Orphanet (Hoehndorf and Gkoutos, 2012), ClinVar (Landrum *et al.*, 2016), and DECIPHER (Firth *et al.*, 2009).

With the number of genotype–phenotype associations available now, it may be possible to predict the phenotypic consequences resulting from some changes on the level of the genotype using machine learning. Several methods have been developed to automatically predict or generate genotype–phenotype associations. To predict phenotype associations,

these methods use different sources such as literature (Kahanda *et al.*, 2015; Collier *et al.*, 2015; Singhal *et al.*, 2016), functional annotations (Kulmanov *et al.*, 2018; Kahanda *et al.*, 2015; Dogan, 2018), protein–protein interactions (Kahanda *et al.*, 2015), expression profiles (Xu *et al.*, 2009; Labuzzetta *et al.*, 2016), genetic variations (Chen *et al.*, 2014; Kahanda *et al.*, 2015), or their combinations (Kahanda *et al.*, 2015). The general idea behind most of these methods is to find genetic similarities, or interactions, and transfer phenotypes between genes based on the assumption that similar or interacting genes are involved in similar or related phenotypes (Gillis and Pavlidis, 2012).

Phenotypes arise from complex biological processes which include genetic interactions, protein–protein interactions, physiological interactions, and interactions of an organism with environmental factors as well as lifestyle and response to chemicals. A large number of these interactions can be described using the Gene Ontology (GO) (Ashburner *et al.*, 2000), and GO annotations are available for a large number of proteins from thousands of species (Huntley *et al.*, 2014).

We developed a novel approach of predicting gene–phenotype associations from function annotations of gene products. We use the Gene Ontology (GO) (Ashburner *et al.*, 2000; The Gene Ontology Consortium, 2018) function annotations as our main feature and predict HPO classes. We propagate both functional and phenotype annotations using the hierarchical structure of GO and HPO and train a deep neural network model which learns to map sets of GO annotations to sets of HPO annotations. One limitation of predicting phenotypes from functions is that not all genes have experimental functional annotations. We overcome this limitation by using the function prediction method DeepGOPlus (Kulmanov and Hoehndorf, 2019) which has been trained on a large number of proteins and can generate accurate functional annotations using only protein sequence. We evaluate our method using the latest phenotype annotations from the HPO (Köhler *et al.*, 2017) and using the evaluation dataset from the Computational Assessment of Function Annotation (CAFA) challenge (Radivojac *et al.*, 2013), and we compare our results with the top performing methods in CAFA 2 (Jiang *et al.*, 2016) and one recent phenotype prediction method, HPO2GO (Dogan, 2018). We demonstrate a significant improvements over the state of the art in each evaluation.

To further validate the usefulness of our phenotype predictions in computational biology, we test whether we can predict gene-disease associations from the predicted phenotype annotations. We compute semantic (phenotypic) similarity between gene–phenotype annotations and disease–phenotype annotations, and our results show that the phenotype annotations generated by DeepPheno are predictive of gene–disease associations. We further analyzed the predictions generated by our method and found that, in average, more than 60% of the predicted genes for a phenotype interact with genes that are already associated with the phenotype, suggesting that some of our false positive predictions might actually be truly associated within a phenotype module.

2 Materials and methods

2.1 Evaluation and training data

2.1.1 Training and testing dataset

We downloaded Human Phenotype Ontology (HPO) version released on 15th of April, 2019 and phenotype annotations released on 3rd of June, 2019 from `https://hpo.jax.org`. HPO Team provides annotations from OMIM and ORPHANET. We use gene–phenotype annotations from both sources and build a dataset of 4,073 genes annotated with 7,465 different phenotypes. In total, the dataset has 150,653 annotations with average 36 annotations per gene. 3,944 of the genes map to manually reviewed and annotated proteins from UniProtKB/SwissProt (Consortium, 2018). We refer to this dataset as June2019. We use reviewed genes/proteins and split the dataset by genes into 90% training and 10% testing sets and use 10% of the training set as a validation set to tune the parameters of our prediction model.

Our main features from which we predict phenotypes are gene functions. We use the Gene Ontology (GO) (Ashburner *et al.*, 2000; The Gene Ontology Consortium, 2018) released on 17th of April, 2019 and GO annotations (uniprot_sprot.dat.gz) from UniProtKB/SwissProt released in April, 2019. We construct three different datasets with different types of functional annotations for the same genes. In the first dataset we use all annotations from the file which includes electronically inferred ones. The second dataset has only experimental functional annotations. We filter experimental annotations using the evidence codes `EXP`, `IDA`, `IPI`, `IMP`, `IGI`, `IEP`, `TAS`, `IC`, `HTP`, `HDA`, `HMP`, `HGI`, `HEP`. For the third dataset, we use GO functions predicted by DeepGOPlus (Kulmanov and Hoehndorf, 2019).

2.1.2 CAFA2 evaluation dataset

To compare our method with state-of-the-art phenotype prediction methods we follow the CAFA (Radivojac *et al.*, 2013) challenge rules and generate a dataset using a time based split. CAFA2 (Jiang *et al.*, 2016) challenge benchmark data was collected from January 2014 until September 2014. We train our model on phenotype annotations that were available before the challenge started and evaluate the model on annotations that appeared during the challenge period. Similarly, we use function annotations that were available before a challenge starts and train DeepGOPlus on a UniProt/Swissprot version released in January 2014. We use the same versions of HPO and GO that were used in the challenge.

2.1.3 Protein–protein interactions data

To evaluate false positive predictions generated by our approach we use protein–protein interactions (PPI) data. We download StringDB PPI networks (Szklarczyk *et al.*, 2018) version 11 published in January, 2019. StringDB is a database of protein's functional associations which includes direct physical and indirect functional interactions. StringDB combines interactions from several primary PPI databases and adds PPI interactions that are predicted with computational methods. We use interactions with a score of 700 or above in order to filter high confidence interactions.

2.2 Baseline and comparison methods

2.2.1 Naive approach

We use several approaches to benchmark and compare our prediction results. The "naive" approach was proposed by the CAFA (Radivojac *et al.*, 2013) challenge as one of the basic methods to assign GO and HPO annotations. Here, each query gene/protein g is annotated with the HPO classes with a prediction score computed as:

$$S(g, p) = \frac{N_p}{N_{total}} \tag{1}$$

where p is a HPO class, N_p is a number of training genes annotated by HPO class p, and N_{total} is a total number of training genes. It represents a prediction based only on the total number of genes associated with a class during training.

2.2.2 CAFA2 Methods

The CAFA2 (Jiang *et al.*, 2016) challenge evaluates several phenotype prediction methods which present state-of-the-art performance for this task. We train and test our model on the same data and compare our results with top performing methods. The CAFA3 (Zhou *et al.*, 2019) challenge

also evaluated HPO predictions but did not release the evaluation results yet.

2.2.3 HPO2GO

HPO2GO predicts HPO classes by learning association rules between HPO and GO classes based on their co-occurrence in annotations (Dogan, 2018). The idea is to map every HPO class p to a GO class f and score the mapping with the following formula:

$$S(p, f) = \frac{2 * N_{p\&f}}{N_p + N_f} \tag{2}$$

where $N_{p\&f}$ is the number of genes annotated with both p and f, N_p is the number of genes annotated with p and N_f is the number of genes annotated with f. In the prediction phase the mappings are used to assign HPO classes to genes with available GO classes.

2.3 Evaluation metrics

In order to evaluate our phenotype predictions and compare our method with other competing methods we use the CAFA (Radivojac *et al.*, 2013) evaluation metrics $F_{\max}$ and $S_{\min}$ (Radivojac and Clark, 2013). Additionally, we report the area under the precision-recall curve (AUPR) (Davis and Goadrich, 2006).

$F_{\max}$ is a maximum gene/protein-centric F-measure computed over all prediction thresholds. First, we compute average precision and recall using the following formulas:

$$pr_i(t) = \frac{\sum_p I(f \in P_i(t) \wedge p \in T_i)}{\sum_p I(p \in P_i(t))} \tag{3}$$

$$rc_i(t) = \frac{\sum_p I(f \in P_i(t) \wedge p \in T_i)}{\sum_p I(p \in T_i)} \tag{4}$$

$$AvgPr(t) = \frac{1}{m(t)} \cdot \sum_{i=1}^{m(t)} pr_i(t) \tag{5}$$

$$AvgRc(t) = \frac{1}{n} \cdot \sum_{i=1}^{n} rc_i(t) \tag{6}$$

where p is an HPO class, T_i is a set of true annotations, $P_i(t)$ is a set of predicted annotations for a gene i and threshold t, $m(t)$ is a number of proteins for which we predict at least one class, n is a total number of proteins and I is an identity function which returns 1 if the condition is true and 0 otherwise. Then, we compute the $F_{\max}$ for prediction thresholds $t \in [0, 1]$ with a step size of 0.01. We count a class as a prediction if its prediction score is higher than t:

$$F_{\max} = \max_t \left\{ \frac{2 \cdot AvgPr(t) \cdot AvgRc(t)}{AvgPr(t) + AvgRc(t)} \right\} \tag{7}$$

$S_{\min}$ computes the semantic distance between real and predicted annotations based on information content of the classes. The information content $IC(c)$ is computed based on the annotation probability of the class c:

$$IC(c) = -\log(Pr(c|P(c)) \tag{8}$$

where $P(c)$ is a set of parent classes of the class c. The $S_{\min}$ is computed using the following formulas:

$$S_{\min} = \min_t \sqrt{ru(t)^2 + mi(t)^2} \tag{9}$$

where $ru(t)$ is the average remaining uncertainty and $mi(t)$ is average misinformation:

$$ru(t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{c \in T_i - P_i(t)} IC(c) \tag{10}$$

$$mi(t) = \frac{1}{n} \sum_{i=1}^{n} \sum_{c \in P_i(t) - T_i} IC(c) \tag{11}$$

2.4 Gene–Disease association prediction

To evaluate predicted phenotype annotations we downloaded gene-disease association data from OMIM (Hamosh *et al.*, 2005). The OMIM database provides associations for around 6,000 diseases and 14,000 genes. We filter these associations with the genes from our randomly split test set and their associated diseases. In total, the dataset has 561 associations of 395 genes with 548 diseases.

We predict an association between gene and disease by comparing their phenotypes. Our hypothesis is that if a gene and disease are annotated with similar phenotypes then there could be an association between them (Hoehndorf *et al.*, 2011). We compute Resnik's (Resnik, 1999) semantic similarity measure for pairs of phenotype classes and use the Best-Match-Average (BMA) (Schlicker *et al.*, 2006) strategy to combine similarites for two sets of annotations. We use the similarity score to rank diseases for each gene and report recall at top 10 rank, recall at top 100, mean rank and the area under the receiver operating characteristic curve (ROC AUC) (Fawcett, 2006) for each prediction method.

Resnik's similarity measure is defined as the most informative common anchestor (MICA) of the compared classes in the ontology. First, we compute information content (IC) for every class with following formula:

$$IC(c) = -\log(p(c))$$

Then, we find Resnik's similarity by:

$$Sim_{Resnik}(c_1, c_2) = IC(MICA(c_1, c_2))$$

We compute all possible pairwise similarities of two annotation sets and combine them with:

$$Sim_{BMA}(A, B) = \frac{\underset{c_1 \in A}{avg} \left(\max_{c_2 \in B}(s(c_1, c_2))\right) + \underset{c_1 \in B}{avg} \left(\max_{c_2 \in A}(s(c_1, c_2))\right)}{2}$$

where $s(x, y) = Sim_{Resnik}(x, y)$.

2.5 Training and tuning of model parameters

We evaluated several models with two, three and four fully connected layer models. We selected the number of units for each layer from $\{250, 500, ..., 2000\}$ with dropout rate from $\{0.2, 0.5\}$ and learning rate from $\{0.01, 0.001, 0.0001\}$ for the Adam optimizer (Kingma and Ba, 2014). We performed 50 trials of random search for best parameters for each type of the models. Our model achieved best validation loss with a fully connected layer with 750 units, dropout rate of 0.5 and learning rate of 0.001. We use the TensorFlow 2.0 (Abadi *et al.*, 2016) machine learning system with Keras API and tune our parameters with Keras Tuner.

3 Results

3.1 DeepPheno Model

Predicting phenotypes from genotypes is one of the biggest challenges in genetics. It requires understanding of biological mechanisms from the molecular over the physiological to the behavioral level. Phenotypes arise from complex interactions between genotype and environment. Our aim is to predict phenotypes which result from the loss of function of a single gene. In order to achieve this we need different types of genetic features such as functions, molecular and multi-cellular interactions, pathways and physiological interactions. Our approach is to use existing experimental

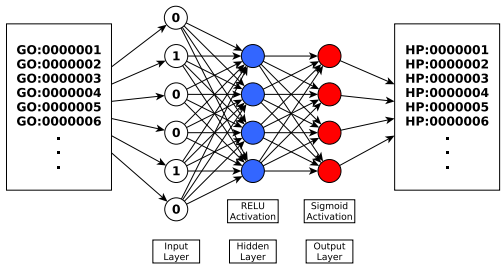

Fig. 1. Neural network model architecture.

and predicted function annotations, and learn the associations between combinations of functions and phenotype annotations. We use OMIM (Hamosh *et al.*, 2005) diseases and their phenotype annotations from the HPO (Köhler *et al.*, 2017) to learn associations between sets of GO functional annotations and sets of HPO phenotype annotations. Since experimental annotations are not available for all genes, we also utilize the sequence-based function prediction method DeepGOPlus (Kulmanov and Hoehndorf, 2019), which uses information of function annotations in many organisms, to fill this gap and predict the functions of gene products. Consequently, DeepPheno can predict phenotypes for all protein-coding genes with an available sequence.

Our phenotype prediction model is a two layer fully-connected neural network which takes a sparse binary vector of functional annotation features as input and outputs phenotype annotation scores for selected terms for prediction. Figure 1 describes the model architecture. The first layer is responsible for reducing the dimensionality of our sparse input and the second layer is a multi-class multi-label classification layer with sigmoid activation functions for each neuron. We use a dropout layer after the first fully-connected layer in training phase to avoid overfitting.

3.2 Evaluation and comparison

We evaluate our method on the phenotype annotations from the HPO database (Köhler *et al.*, 2017) released in June 2019. We randomly split the dataset into training, validation and testing sets; we split the data by genes so that if a gene is included in one of the three sets, the gene is present with all its annotations. We tune all parameters of our models using the validation set and report evaluation results on the unseen testing set. We train four neural network models using the same phenotype annotations and same training/testing split, but with different functional annotations as features. The first model is called DeepPhenoGO and it uses only experimental function annotations. The second model is called DeepPhenoIEA and is trained on all annotations from UniProtKB/SwissProt including predicted ones (i.e., including annotations with an evidence code indicated the annotation was electronically inferred). The predicted annotations are usually based on sequence or structural similarity and multiple sequence alignments. The third model is our main model called DeepPheno. We train DeepPheno on functions predicted by our DeepGOPlus method. Finally, we train DeepPhenoAll model on combined annotations from UniProtKB/Swissprot and DeepGOPlus. We propagate all annotations using the structure of the HPO ontology.

We first compare our results with phenotype annotations generated by the "naive" method (see Section 2.2.1 for details). The naive method predicts the most frequently annotated phenotype classes in the training set for all genes in the testing set. It achieves an $F_{\max}$ of 0.402 which is close to our results and higher than the state-of-the-art methods in the CAFA2 challenge. Despite such performance, naive annotations do not have any practical use as they are identical for every gene. Our

Method	$F_{\max}$	$S_{\min}$	**AUPR**
Naive	0.402	151.992	0.357
DeepPhenoGO	0.452	144.177	0.429
DeepPhenoIEA	0.463	143.088	0.441
DeepPheno	0.463	**141.124**	0.440
DeepPhenoAll	**0.470**	142.343	**0.444**

Table 1. The comparison of performance on the June2019 dataset.

Method	$F_{\max}$	$S_{\min}$	**AUPR**
Naive	0.367	89.326	0.299
HPO2GO	0.309	97.337	0.177
DeepPhenoGO	0.388	88.859	0.330
DeepPhenoIEA	0.398	89.374	**0.339**
DeepPheno	**0.402**	89.053	**0.339**
DeepPhenoAll	0.396	**88.104**	0.338

Table 2. The comparison of performance on the first CAFA2 challenge dataset.

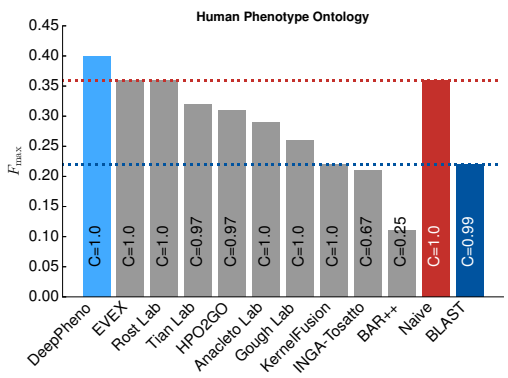

Fig. 2. Comparison of DeepPheno with CAFA2 top 10 methods and HPO2GO.

neural network model achieves an $F_{\max}$ of 0.452 when we train it with only experimental GO annotations and, as expected, it improves to 0.463 when we add predicted annotations in UniProtKB. We expected results to improve because experimental GO annotations are not complete and the electronically inferred function annotations can add missing information. The model trained with DeepGOPlus annotations achieves the same $F_{\max}$ of 0.463, but results in the best performance among all models in our $S_{\min}$ evaluation (which also considers the specificity of the predicted phenotype classes, see Section 2.3). Merging all GO annotations results in a slight improvement of phenotype predictions suggesting that DeepGOPlus annotations add more information to the predicted annotations in UniProtKB. The results show that DeepPheno would be the best choice for annotating protein sequences without experimental GO annotations since it requires only DeepGOPlus annotations which can be predicted from protein amino acid sequence information alone.

To compare our method with other methods, we trained and tested our model using the CAFA2 challenge data, i.e., using the training and testing data as well as the ontologies provided in CAFA2 (see Section 2.1.2). We further evaluated annotations for CAFA2 targets provided by the HPO2GO method (Dogan, 2018). The top performing methods in CAFA2 achieve an $F_{\max}$ of around 0.36 (Radivojac *et al.*, 2013). Our DeepPhenoGO model trained using only experimental GO annotations achieve $F_{\max}$ of 0.388. Models which use predicted annotations improve $F_{\max}$ score and we achieve the best $F_{\max}$ of 0.402 with the DeepPheno model trained with DeepGOPlus annotations.

Method	Hits10 (%)	Hits100 (%)	Mean Rank	AUC ROC
Naive	2	18	275.24	0.50
DeepPhenoGO	12	46	167.12	0.70
DeepPhenoIEA	14	51	147.59	0.73
DeepPheno	15	50	149.58	0.73
DeepPhenoAll	15	51	147.85	0.73
RealHPO	**88**	**96**	**13.86**	**0.98**

Table 3. The comparison of performance on gene–disease association prediction.

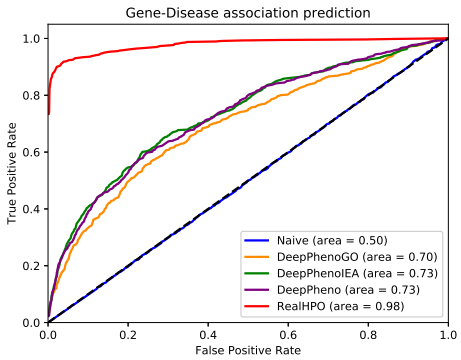

Fig. 3. Semantic similarity based prediction of gene-disease associations using predicted and manually asserted phenotype annotations.

3.3 Application to predicting gene–disease associations

Phenotype annotations have many applications, including prediction of candidate genes that are causally involved in diseases (Köhler *et al.*, 2009; Jagadeesh *et al.*, 2019; Hoehndorf *et al.*, 2011). For example, we can compare gene–phenotype associations to disease phenotypes and prioritize candidate diseases for diagnosis (Jagadeesh *et al.*, 2019) or prioritize genetic variants that are causative for a disease (Smedley *et al.*, 2015; Boudellioua *et al.*, 2019). These methods can suggest candidate genes for rare diseases and improve identification of causative variants in a clinical setting (Jagadeesh *et al.*, 2019). Here, our aim is to test whether predicted phenotype annotations generated by DeepPheno are applicable for gene–disease association predictions.

One of the widely used methods for comparing ontology class annotations is semantic similarity (Harispe *et al.*, 2014). We compute semantic similarity scores between gene–phenotype annotations and disease–phenotype annotations and use it to rank diseases for each gene. Such an approach is widely used to predict gene–disease associations based on phenotypes (Köhler *et al.*, 2009; Hoehndorf *et al.*, 2011) and allows us to determine how our predicted annotations can contribute to such an application.

As expected, phenotype annotations generated by the naive approach are not predictive of gene–disease associations and resulted in a performance with ROCAUC of 0.50. On the other hand, similarity scores from existing annotations from HPO performed nearly perfectly with a ROCAUC of 0.98. The reason for this nearly perfect prediction is that most of the OMIM diseases are associated with only one gene and share almost same phenotype annotations as their associated genes because of how the gene–phenotype associations have been generated (Köhler *et al.*, 2017). Predicting gene–disease associations using the DeepPheno models that rely on electronically inferred GO function annotations resulted in an ROCAUC of 0.73, which is slightly higher than predicting gene–disease associations using the model trained with experimental GO annotations which resulted in ROCAUC of 0.70. Table 3 summarizes the results and Figure 3 depicts the ROC curves. Overall, this evaluation shows that our model is predicting phenotype associations which can be used for predicting gene–disease associations.

3.4 Evaluation of false positives

In our experiments, we consider the absence of knowledge about an association as a negative association. When we predict such an association, we consider this a false positive prediction. However, current gene–phenotype associations in the database are not complete and some of the false positive predictions generated by our method may actually be a correct association. To test this hypothesis, we further evaluate if our predicted genes are interacting with a phenotype-associated gene using the underlying assumption that phenotypes are determined by network modules of interacting genes and gene products (Wang *et al.*, 2019; Han *et al.*, 2015).

To evaluate our false positives, we generated DeepPheno predictions for 18,860 genes with corresponding protein entries in Swissprot/UniProtKB. Then, for every specific phenotype, we compared false positive genes and genes that are interacting with the phenotype-associated gene using the STRING database (Szklarczyk *et al.*, 2018). We then computed the percentage of overlap. We found that on average, 62% of our false positives interact with a phenotype-associated gene and may contribute to a phenotype-module within the interaction network. We tested whether this finding is significant by performing a random simulation experiment. For every phenotype, we generated the same number of random gene associations as our predictions and computed the average overlap of false positives and interacting genes. We repeated this procedure 1,000 times and tested the significance of having an average of 62% overlap. We found that random predictions give an average of 24% overlap and the probability of observing 62% overlap under a random assignment of genes to phenotypes is less than 0.001, demonstrating that the genes we predict for a phenotype are significantly more likely to directly interact with a known phenotype-associated gene.

4 Discussion

DeepPheno can predict sets of gene–phenotype associations from gene functional annotations. Specifically, it is designed to predict phenotypes which arise from a complete loss of function of a gene. Together with function prediction methods such as DeepGOPlus (Kulmanov and Hoehndorf, 2019), DeepPheno can, in principle, predict phenotype associations for protein-coding genes using only the protein's amino acid sequence. However, DeepGOPlus was trained on experimentally annotated sequences of many organisms, including several animal model organisms. It further combines global sequence similarity and a deep learning model which learns to recognize sequence motifs as well as some elements of protein structure. The combination of this information is implicitly used in DeepGOPlus and its predictions, and is therefore able to predict physiological functions that are closely related to the abnormal phenotypes predicted by DeepPheno.

We evaluated DeepPheno on two datasets and compared its predictions with the top performing methods in the CAFA2 challenge. DeepPheno showed overall the best performance in these evaluations; however, the improvement in terms of the $F_{\max}$ measure was not far from the naive classifier. The naive classifier annotates all genes with the same set of annotations based on their annotation frequency in training set. We hypothesize that the naive classifier achieves an $F_{\max}$ close to DeepPheno and other phenotype prediction models because of the propagation of

annotations using the hierarchical structure of HPO. Most of the top level classes will be associated with almost all of the genes after this propagation process. Therefore, we further evaluated the predicted annotations and demonstrated that DeepPheno annotations can be used to prioritize gene–disease associations whereas the naive annotations do not perform better than a random classifier in this task.

In addition, we analyzed our false positive predictions for several diseases and found that in more than 60% of the false positive gene-phenotype association, the gene is interacting with a truly associated gene. We further investigated in some cases whether the false positive gene–phenotype associations were reported to be significant in GWAS studies. We used the GWAS Catalog (Buniello *et al.*, 2018) to determine if some of our false predictions were reported in GWAS studies, and tested the phenotypes Type II Diabetes mellitus (`HP:0005978`) and Hypothyroidism (`HP:0000821`) since they are specific phenotypes in HPO and are very well studied with many GWAS analyses. Type II Diabetes is associated with 151 genes in the HPO database and we predict 227 genes with DeepPheno. 59 genes from our predictions were already included in the HPO database and we consequently generated 168 false positives. We found that our false positive associations with genes NPHP4, BEST3, and TXNL4B were reported to be associated with Type II Diabetes in GWAS Catalog while the others were not. NPHP4 was also associated by a GWAS study using the UK Biobank dataset (`https://www.nealelab.is/uk-biobank`). For the Hypothyroidism (`HP:0000821`) phenotype, we predict 152 genes which are not in our training set, of which the CTLA4, ARID5B, and TMEM131 genes are also reported in the GWAS Catalog; a GWAS study using the UK Biobank dataset reported an association with FANCA gene which is in our false positive set. As we predict genes that interact with phenotype-associated genes, and while the predicted genes do mostly not reach genome-wide significance in GWAS studies, a possible explanation may be that some of the predicted genes have only low effect sizes, preventing them from being detected in a GWAS study. In future research, it may be possible to explore this hypothesis further by directly evaluating interaction models and DeepPheno's predictions on GWAS datasets.

Currently, DeepPheno suffers from several limitations. Firstly, our model uses an ordinary two-layer fully-connected neural network which does not incorporate ontological axioms and the hierarchical structure of the classes in HPO. Since there are no hard constraints on class dependencies, the model might produce inconsistent predictions which have to be resolved after generating the predictions. Secondly, we use only functional annotations as features. This gives our model the ability to predict phenotypes for many genes; however, phenotypes are very complex and do not only depend on functions, and including other information may further improve our model. Specifically, we plan to include different types of interactions between genes as well as gene expression information. Finally, we can only predict a limited number of phenotypes for which we find at least 50 annotated genes. This limitation is caused by the need to train our neural network model. One way to overcome this limitation is to include phenotype associations with different evidence, such as those derived from GWAS study instead of using only phenotypes resulting from Mendelian disease as included in the HPO database.

Acknowledgements

We acknowledge the use of computational resources from the KAUST Supercomputing Core Laboratory.

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.

References

Abadi, M. *et al.* (2016). Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA. USENIX Association.

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

Boudellioua, I. *et al.* (2019). Deeppvp: phenotype-based prioritization of causative variants using deep learning. *BMC Bioinformatics*, **20**(1), 65.

Bult, C. J. *et al.* (2018). Mouse Genome Database (MGD) 2019. *Nucleic Acids Research*, **47**(D1), D801–D806.

Buniello, A. *et al.* (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, **47**(D1), D1005–D1012.

Chen, Y.-C. *et al.* (2014). A probabilistic model to predict clinical phenotypic traits from genome sequencing. *PLOS Computational Biology*, **10**(9), 1–11.

Cherry, J. M. *et al.* (2011). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, **40**(D1), D700–D705.

Collier, N. *et al.* (2015). Phenominer: from text to a database of phenotypes associated with omim diseases. *Database*, **2015**, bav104.

Collins, F. S. *et al.* (2007). A new partner for the international knockout mouse consortium. *Cell*, **129**(2), 235.

Consortium, T. U. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**(D1), D506–D515.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA. ACM.

Dogan, T. (2018). Hpo2go: prediction of human phenotype ontology term associations for proteins using cross ontology annotation co-occurrences. *PeerJ*, **6**, e5298–e5298. 30083448[pmid].

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

Firth, H. V. *et al.* (2009). DECIPHER: Database of chromosomal imbalance and phenotype in humans using ensembl resources. *American journal of human genetics*, **84**(4), 524–533.

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.

Hamosh, A. *et al.* (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **33**(suppl_1), D514–D517.

Han, S. K. *et al.* (2015). Network modules of the cross-species genotype-phenotype map reflect the clinical severity of human diseases. *PLOS ONE*, **10**(8), 1–13.

Harispe, S. *et al.* (2014). The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, **30**(5), 740–742.

Hoehndorf, R. and Gkoutos, G. V. (2012). A translational medicine approach to orphan diseases. In *Proceedings of the Virtual Physiological Human conference*.

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

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

Jagadeesh, K. A. *et al.* (2019). Phrank measures phenotype sets similarity to greatly improve mendelian diagnostic disease prioritization. *Genetics in Medicine*, **21**(2), 464–470.

Jiang, Y. *et al.* (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, **17**(1), 184.

Kahanda, I. *et al.* (2015). Phenostruct: Prediction of human phenotype ontology terms using heterogeneous data sources [version 1; referees: 2 approved]. *F1000Research*, **4**(259).

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Kulmanov, M. and Hoehndorf, R. (2019). DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*.

Kulmanov, M. *et al.* (2018). Ontology-based validation and identification of regulatory phenotypes. *Bioinformatics*, **34**(17), i857–i865.

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.* (2017). The human phenotype ontology in 2017. *Nucleic Acids Research*, **45**(D1), D865.

Labuzzetta, C. J. *et al.* (2016). Complementary feature selection from alternative splicing events and gene expression for phenotype prediction. *Bioinformatics*, **32**(17), i421–i429.

Landrum, M. J. *et al.* (2016). Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, **44**(D1), D862–D868.

Manis, J. P. (2007). Knock out, knock in, knock down — genetically manipulated mice and the nobel prize. *New England Journal of Medicine*, **357**(24), 2426–2429.

Radivojac, P. and Clark, W. T. (2013). Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, **29**(13), i53–i61.

Radivojac, P. *et al.* (2013). A large-scale evaluation of computational protein function prediction. *Nat Meth*, **10**(3), 221–227.

Resnik, P. (1999). Semantic similarity in a taxonomy: An Information-Based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, **11**, 95–130.

Schlicker, A. *et al.* (2006). A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, **7**(1), 302.

Singhal, A. *et al.* (2016). Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. *PLOS Computational Biology*, **12**(11), 1–19.

Smedley, D. *et al.* (2015). Next-generation diagnostics and disease-gene discovery with the exomiser. *Nature Protocols*, **10**, 2004 EP –.

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 Semantics*, **6**, 11.

Szklarczyk, D. *et al.* (2018). 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 Gene Ontology Consortium (2018). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, **47**(D1), D330–D338.

Tifft, C. J. and Adams, D. R. (2014). The national institutes of health undiagnosed diseases program. *Curr Opin Pediatr*, **26**(6), 626–633. 25313974[pmid].

Wang, Q. *et al.* (2019). Co-expression network modeling identifies key long non-coding rna and mrna modules in altering molecular phenotype to develop stress-induced depression in rats. *Translational Psychiatry*, **9**, 125.

Washington, N. L. *et al.* (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLOS Biology*, **7**(11), 1–20.

Xu, M. *et al.* (2009). Automated multidimensional phenotypic profiling using large public microarray repositories. *Proceedings of the National Academy of Sciences*, **106**(30), 12323–12328.

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