Subject Section

Prediction of novel virus–host interactions by integrating clinical symptoms and protein sequences

Wang Liu-Wei1, Şenay Kafkas2, Jun Chen1, Jesper Tegnér1,3 and Robert Hoehndorf1,2,*

1 Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia,
2 Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia,
3 Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia.

* To whom correspondence should be addressed.

Abstract

Motivation: Infectious diseases from novel viruses are becoming a major public health concern. Fast identification of virus–host interactions can reveal mechanistic insights of infectious diseases and shed light on potential treatments and drug discoveries. Current computational prediction methods for novel viruses are based only on protein sequences. Yet, it is not clear to what extent other important features, such as the symptoms caused by the viruses, could contribute to a predictor. Disease phenotypes (i.e., symptoms) are readily accessible from clinical diagnosis and we hypothesize that they may act as a potential proxy and an additional source of information for the underlying molecular interactions between the pathogens and hosts.

Results: We developed DeepViral, a deep learning method that predicts potential protein–protein interactions between human and viruses. First, human proteins and viruses were embedded in a shared space using their associated phenotypes, functions, taxonomic classification, as well as formalized background knowledge from biomedical ontologies. By extending a sequence learning model with phenotype features, our model can not only significantly improve over previous sequence-based approaches for inter-species interaction prediction, but also identify pathways of viral targets under a realistic experimental setup for novel viruses.

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

Contact: robert.hoehndorf@kaust.edu.sa

1 Introduction

Infectious diseases emerging unexpectedly from novel pathogens have been a major public health concern around the globe (Jones et al., 2008). Pathogens disrupt host cell functions (Finlay and Cossart, 1997) and target immune pathways (Dyer et al., 2010) through complex inter-species interactions of proteins (Dyer et al., 2008), RNA (Fajardo et al., 2015) and DNA (Weitzman et al., 2004). The study of pathogen–host interactions (PHI) can therefore provide insights into the molecular mechanisms underlying infectious diseases and guide the discoveries of novel therapeutics or provide a basis for repurposing of available drugs. For example, a previous study of many PHIs showed that pathogens typically interact with the protein hubs (those with many interaction partners) and bottlenecks (those of central location to important pathways) in human protein–protein interaction (PPI) networks (Dyer et al., 2008). However, due to cost and time constraints, experimentally validated pairs of interacting pathogen–host proteins are limited in number. Moreover, there exists a time delay for a validated PHI to be included in a database of PHIs, often requiring manual curation of the literature or text mining efforts (Thieu et al., 2012). Therefore, the computational prediction of PHIs is a useful complementary approach in suggesting candidate interaction partners out of all the human proteins.
Existing PHI prediction methods typically utilize features of the interacting proteins, such as PPI network topology, protein structures and sequences, or functional profiling such as Gene Ontology similarity and KEGG pathway analysis (Nourani et al., 2015). While protein functions have been shown to predict intra-species (e.g., human) PPIs (Guzzi et al., 2011; Jain and Bader, 2010; Pesquita et al., 2009) and such protein specific features exist for some extensively studied pathogens, such as Mycobacterium tuberculosis (Huo et al., 2015) and HIV (Mukhopadhyay et al., 2014), for most of the novel pathogens, these features are rare and expensive to obtain. As new virus species are being discovered each year, with potentially many more to come (Woolhouse et al., 2012), a method is needed to rapidly identify candidate interactions from information that can be obtained quickly—for example, the signs and symptoms of the host, which may be utilized as a proxy for the underlying molecular interactions between host and pathogen proteins.

The phenotypes elicited by pathogens, i.e., the signs and symptoms observed in a patient, may provide information about molecular mechanisms (Gkountou et al., 2018). The information that phenotypes provide about molecular mechanisms is commonly exploited in computational studies of Mendelian disease mechanisms (Oellrich et al., 2016; Schöpfel et al., 2012), for example, to suggest candidate genes (Hoehndorf et al., 2011; Meehan et al., 2017) or diagnose patients (Köhler et al., 2009), but the information can also be used to identify drug targets (Hoehndorf et al., 2013a) or gene functions (Hoehndorf et al., 2013b). To the best of our knowledge, phenotypes and phenotype similarity have not yet been utilized for the prediction of PHIs.

We hypothesize that the phenotypes elicited by an infection with a pathogen are, among others, the result of molecular interactions, and that knowledge of the phenotypes in the host can be used to suggest the protein perturbations, from which these phenotypes arise. While a large number of phenotypes resulting from infections are a consequence of immune system processes that are shared across a wide range of different types of pathogens, certain hallmark phenotypes, such as decreased CD4 cell-count in infections with HIV (Ford et al., 2017) or microcephaly resulting from Zika virus infections (Mikako et al., 2016), can be used to suggest interacting host proteins, through which these symptoms are elicited.

One common limitation of the PH prediction problem is the lack of ground truth negative data. A recent method DeNovo (Eid et al., 2015) adopted a "dissimilarity-based negative sampling" for each virus protein, the negatives are sampled from human proteins that do not have known interactions with other similar virus proteins (above a certain sequence similarity threshold). Another method based on protein sequences (Zhou et al., 2018) samples negatives from only the set of host proteins that are less than 80% similar (in terms of sequence similarity) from the host proteins in the positive training data. By construction, these sampling schemes make the human proteins in the negative set different from the positive set; when used not only for training a model but also for evaluating the model's performance, this sampling scheme has the potential to overestimate the actual performance for finding novel PHIs. In a more realistic evaluation for a novel virus species, a model would be evaluated on all the host proteins that it could potentially interact with, regardless of sequence similarity.

We developed a machine learning method, DeepVirus, to predict potential interactions between viruses and all human proteins for which we can generate the relevant features. Firstly, the features of phenotypes, functions and taxonomic classifications are embedded in a shared space for human proteins and viruses. We then extend a sequence model by incorporating the phenotype features of viruses into the model. We show that the joint model trained on both the sequences and phenotypes can significantly improve over the state-of-the-art method and predict potential PHIs in a realistic experimental setup for novel viruses and predict human protein targets that are enriched for relevant pathways.

# 2 Materials and methods

## 2.1 Data sources of interactions, phenotypes, functions and ontologies

Interactions between hosts and pathogens were downloaded from the Host Pathogen Interaction Database (HPIDB) (Ammari et al., 2016). The database contains 32,758 distinct pairs of protein-protein interaction between human and viruses, equipped with a corresponding MI-score (see Section 2.3) and the virus has a family taxon present in the NCBI taxonomy (Sayers et al., 2009). The phenotypes associated with pathogens were collected from the PathoPhenoDB (Karkas et al., 2018), a database of manually curated and text-mined associations of pathogen, diseases and phenotypes. We downloaded the PathoPhenoDB database version 1.2.1 (http://patho.phenomemrowser.net/).

The phenotypes associated with human genes were collected from the Human Phenotype Ontology (HPO) database (Köhler et al., 2018), and the phenotypes associated with mouse genes and the orthologous gene mappings from mouse genes to human genes, originated from the Mouse Genome Informatics (MGI) database (Smith et al., 2018). The Entrez gene IDs in HPO and MGI were mapped to reviewed Uniprot protein IDs using the Uniprot Retrieve/ID mapping tool (https://www.uniprot.org/uploadlists/). The Gene Ontology annotations of human proteins (release date 2020-02-22) were downloaded from the Gene Ontology Consortium (Ashburner et al., 2000; The Gene Ontology Consortium, 2017).

To add background knowledge from biomedical ontologies of phenotypes and GO classes, we downloaded the cross-species PhenomeNET Ontology (Hoehndorf et al., 2011; Rodríguez-García et al., 2017), which is built upon and includes the Gene Ontology (Ashburner et al., 2000; The Gene Ontology Consortium, 2017), from the AberOWL ontology repository (Hoehndorf et al., 2015a) on September 13, 2018. We obtained the NCBI Taxonomy classification (Sayers et al., 2009) as an ontology in OWL format (version 2018-07-27) from EMBL-EBI ontology repository (https://www.ebi.ac.uk/ols/ontologies/ncbitaxon).

## 2.2 Learning feature embeddings

To generate feature embeddings, we use DL2Vec (Chen et al., 2020), a recent method for learning features for entities (in our case, the human proteins and viruses) from their associations to ontology classes. DL2Vec first converts the ontologies and entity associations into a graph, with the classes and entities as the nodes and the associations and ontology axioms as the edges. Then several random walks are performed, starting from the entities over to the ontology graph and thereby generating a corpus of walks in the form of sentences capturing the graph neighborhoods and thereby the ontology axioms. Following the construction of such sentences, a Word2Vec skipgram model (Mikolov et al., 2013) is used to learn an embedding for each entity by learning from the corpus. The resulting embedding is a vector representation of an entity capturing its co-occurrence relations with other entities within the graph generated by DL2Vec. For an example, the embedding of a virus contains the feature information from its neighborhood on the graph, i.e., its phenotypes and its taxonomic relatives.

## 2.3 Positive and negative sampling

For training DeepVirus, curated virus–host interactions were obtained from HPIDB (Ammari et al., 2016), a database of host–pathogen protein protein interactions. Next two positive sets were constructed by filtering using different confidence levels. HPIDB (Ammari et al., 2016) provides MI-scores (Villaveces et al., 2015), a confidence score for molecular
interactions, for the curated PHIs from two sources, IntAct (Kerrien et al., 2011) and VirHostNet (Guirimand et al., 2015). We filter HPIDB with a MIscore threshold at 0.4 to construct a high confidence positive set of PHIs, reducing the number of distinct interaction pairs to 3,600. Our threshold of 0.4 is chosen to ensure high confidence as it filters out the peak of the data between 0.3 and 0.4, as shown in Figure 1.

2.5 Experimental setup and evaluation metrics

A method to predict PHIs for a novel virus should have the capacity to predict for all human proteins that the novel virus could potentially interact with, and realistically simulate a scenario where a novel virus emerges, for which we have no known interactions and no knowledge about the molecular functions of the viral proteins.

To evaluate the model realistically for a novel virus, the predictive performance is evaluated in a leave-one-family-out (LOFO) cross validation manner, in which we leave out one virus family in our positive set for testing, 20% of the remaining families for validation, and the rest 80% for training. The objective of the LOFO cross-validation is to evaluate the model under a scenario where the novel virus emerges from a novel virus family in the situation where we have no knowledge about its protein interactions.

For each viral protein in the test family, we rank all the human proteins by the predicted likelihood of interaction and evaluate the model by aggregating the normalized ranks of the true positive interacting human proteins to compute the area under the receiver operating characteristic (ROC) curve (Fawcett, 2006). Due to the large number of negatives in relation to the positives, normalized ranks approximate the true negative rate (TNR) of the ROC curve. A high ROC AUC indicates the ability of the model to prioritize the true positive proteins among all the human proteins. We also evaluate by the hit rates at rank 10 and rank 100, denoted Hit@10/100, which are defined as the proportion of true positive human proteins ranked within the top 10 or 100 across all the human proteins.

3 Results

DeepViral is a model that predicts potential protein interactions between viruses and human from the protein sequences and feature embeddings of phenotypes, functions and taxonomies. To enable predictions based on such different types of features we embed them in a shared representation space. Then we incorporate these feature embeddings with a protein sequence model to predict for potential PHIs of novel viruses. The workflow of DeepViral is illustrated in Figure 2.

3.1 Embedding features of viruses and human proteins from phenotypes, functions and taxonomies

We start with the biological hypothesis that phenotypes (i.e., symptoms) elicited by viruses in their hosts can act as a proxy for the underlying molecular mechanisms of the infection, and therefore may provide additional information to the prediction of potential PHIs for novel viruses.

To generate feature embeddings for human proteins and virus taxa, we apply a recent representation learning method DL2Vec (Chen et al., 2020), which learns feature embeddings for entities based on their annotations to ontology classes (see Section 2.2). DL2Vec takes two types of inputs: the associations of the entities with ontology classes (e.g., human proteins and their functions), and the ontologies themselves. DL2Vec exploits the underlying semantic information to provide formalized background knowledge through connecting different ontologies of phenotypes, functions and virus taxonomy.

For representing virus taxa through the phenotypes they elicit in their hosts, we use the phenotype associations for viruses from PathoPhenoDB (Kathos et al., 2018), a database of pathogen to host phenotype (signs and symptoms) associations. To increase the coverage of phenotypes beyond PathoPhenoDB, the taxonomic relations of the viruses were added from the NCBI Taxonomy (Sayers et al., 2009). By adding these taxonomic relations (as annotations of viruses to DL2Vec), we propagate the known phenotypes along the taxonomic hierarchies and learn a generalized embedding for viruses that do not have any phenotype annotations in PathoPhenoDB but have close relatives that do.

![Fig. 1: The distribution of MIscores in HPIDB.](image-url)
Similarly, for representing human proteins, we use the annotations of
their associated phenotypes from the Human Phenotype Ontology (HPO)
database (Köhler et al., 2018), the phenotypes associated with their mouse
orthologs from the Mouse Genome Informatics (MGI) database (Smith et al.,
2018), and their protein functions from the Gene Ontology (GO)
database (Ashburner et al., 2000; The Gene Ontology Consortium, 2017).

To provide DL2Vec with structured background knowledge through
ontologies, we use the cross-species phenotype ontology PhenomeNET
(Hoehndorf et al., 2011; Rodríguez-García et al., 2017) to associate human
and mouse phenotypes, the Gene Ontology (Ashburner et al., 2000; The
Gene Ontology Consortium, 2017) to incorporate knowledge of protein
functions and the NCBI Taxonomy ontology (Sayers et al., 2009) for the
taxonomic relations between viruses. These ontologies contain formalized
biological background knowledge (Hoehndorf et al., 2015b), which has
the potential to significantly improve the performance of these features in
machine learning and predictive analyses (Smaili et al., 2019).

3.2 The joint prediction model from phenotypes and
sequences

DeepViral consists of a phenotype model trained on phenotypes caused
by a viral infection and a sequence model trained on protein sequences, as
shown in Figure 2 (b). The models take a pair of virus and human proteins
as input and predicts the likelihood of their interaction. The inputs for a
human protein are its feature embedding and sequence, and the features
for a viral protein is its sequence and the feature embedding of the virus
species it belongs to. The sequence model projects the protein sequence
into a low dimension vector representation, which is concatenated with the
vector projected from the embedding by the phenotype model to form a
joint representation of the proteins. A dot product is performed over the two
vector representations of the pair of proteins to compute their similarity,
which then is used as input to a sigmoid activation function to compute
their predicted probability of interaction.

To compare with the state-of-the-art method for PHI prediction of
novel viruses, we train our model on the four datasets provided by a recent
machine learning method (Zhou et al., 2018) for predicting PHI of Ebola
and H1N1. DeepViral is able to improve over the previous method in all
the evaluation metrics across these datasets, as summarized in Table 1.
Notably, integrating the embeddings of protein features (based on GO)
with the sequences performs better in almost all the metrics as compared
to the sequence only model.

3.3 Joint prediction for novel viruses from novel families

We then apply DeepViral following the experimental setup described in
Section 2.5, to evaluate the prediction performance of the model under
the scenario where a novel virus (from a novel family) emerges and no
previous knowledge (except about its protein sequences and phenotypes)
is known. Under this setting, we evaluated the performances of DeepViral
with different combinations of features of virus and human proteins, as
summarized in Table 2. The evaluation was performed on the full and
the high confidence dataset respectively, and only on the viral families
that have more positives than a threshold (set to be 1% of the total
number of positives) in order to reduce the amount of runtime for model
evaluation. The models are run 5 times for each set of features to compute
the confidence interval of the ROCAUC. For the models using only the
feature embeddings of viruses as input (i.e. the first six models in
Table 2), a potential interaction is predicted between a human protein
and a virus species, instead of a protein–protein interaction. In an
evaluation where the inputs are not symmetric, e.g., only using the sequences
of human proteins but not viruses, an additional dense layer is added to
project the longer representation to the same dimension as the other so
that the dot product can be performed.

Overall, among the features of human proteins, the protein function
annotations based on GO perform better than phenotypes. Sequences
from human and viral proteins almost always improve the performances,
with the combination of GO, virus phenotypes and protein sequences
performing the best. The dataset without filtering the positives tend to
give higher ROCAUCs across most features, likely due to the larger set of
training data. However, the models trained on the high confidence dataset
perform consistently higher in the hit rates at rank 10 and 100, potentially
a result of smaller but more distinctive set of proteins.

Remarkably, the predicted human proteins for the viruses are enriched
in pathways of viral targets. We aggregate the top 100 predicted
Table 1. Comparison with the state-of-the-art methods on the datasets of Ebola and H1N1 (Zhou et al., 2018) (the performances of the previous method are from Table 5 of the original paper). All evaluation metrics are computed following the original paper: ACC - accuracy, PPV - positive prediction value (precision), SN - sensitivity, AUC - area under the ROC curve. Different combinations of the features are used for training DeepViral: Seq - the protein sequences, Pheno - the virus phenotype embeddings, HP, MP and GO - the embeddings of human proteins from HPO, MGI and GO, respectively. The bold numbers represent the better metric for a dataset. The dash lines (–) mean that the datasets are not applicable: the training sets TR3 and TR4 contain host and viral proteins of other species (for transfer learning), but currently our set of feature embeddings only contain host proteins and viral proteins of other species.

Table 2. Evaluation results of DeepViral for predicting PHIs under a realistic setting for novel viruses. HP, MP and GO denote the source of the human phenotype embeddings, i.e., HPO, MGI and GO, respectively. For the models using only the phenotypes of the viruses, i.e., without sequences, the predicted interaction is between a human protein and a virus species (the dash line indicates the absence of viral proteins). The mean and confidence interval of ROCAUCs are provided as well as the mean of the hit rates at rank 10 and 100, i.e., Hit@10/100. The best performing combination of features across all datasets are bolded.

4 Discussion

We developed DeepViral, a machine learning method for predicting PHIs between viruses and human. DeepViral is, to the best of our knowledge, the first predictor using clinical phenotypes as an additional feature in PHI prediction and it turned out to provide a significant improvement over purely sequence-based methods. Phenotype-based approaches have been successful in predicting disease-gene associations for Mendelian diseases (Hoehndorf et al., 2011) and intra-species PHIs (Alshahrani et al., 2017), but have not yet been used for the prediction of (inter-species) PHIs in infectious diseases. Our model avoids the bottle-neck of identifying the molecular functions of pathogen proteins, by instead introducing a novel and - in the context of infectious diseases – rarely explored type of feature, the phenotypes elicited by pathogens in their hosts, as a “proxy” for the molecular mechanisms, which in turn eventually produce the observed clinical phenotypes.

By challenging DeepViral with novel viruses, we could extract specific pathways being attacked by the viruses, as indicated by the predicted interactions in the human proteome. The focus of our method on utilizing features generated based on endo-phenotypes observed in humans and mice (Schofeld et al., 2016) has therefore the crucial advantage that we can identify host-pathogen interactions that may contribute to particular signs and symptoms. For example, our model consistently prioritizes the interaction between the proteins of Zika virus (NCBITaxon:64320) and DDX11 (UniProt:CO00573) in humans. Infections with Zika virus have the potential to result in abnormal embryogenesis and, specifically, microcephaly (Wang et al., 2017). Phenotypes associated with DDX11 in the mouse ortholog include abnormal embryogenesis, microcephaly, and
abnormal neural tube closure (Chen et al., 2016). DDX3X mutations in humans have been found to result in intellectual disability, specifically in females and affecting individuals in a dose-dependent manner (Blok et al., 2015). While DDX3X has previously been linked to the infectivity of Zika virus (Doñate-Macián et al., 2018), our model further suggests a role of DDX3X in the development of the embryogenesis phenotypes resulting from Zika virus infections.

While improving over a previous model on Ebola and H1N1 (Zhou et al., 2018), we argue that the performance of DeepViral on these datasets may have been over-estimated due to the negative sampling scheme based on sequence similarity that is used not only for training but also for evaluation of the model. Under a more realistic evaluation procedure that considers all host proteins as potential interaction partners for novel viruses, the achieved predictive performances are considerably lower. This calls for future efforts in the direction of PHI prediction of novel viruses, an issue today of increasing relevance to global public health. Accurate predictions of potential PHIs for novel pathogens with rapidly obtainable features would be an important aid for the understanding of infectious disease mechanisms and the repurposing of existing drugs.

An example of such a novel virus is the novel coronavirus SARS-CoV-2, which as of 21st April 2020 reached more than 2.5 million infected cases and 170 thousand fatalities globally (Dong et al., 2020) in a timespan of 5 months. Based on a recently released dataset of 332 PHIs from 26 viral proteins of SARS-CoV-2 (Gordon et al., 2020), we applied DeepViral by treating it as a novel family (with no other Coronaviridae viruses in the dataset) and achieved a ROCAUC of 0.738 (0.730–0.747), which is within the observed variability in predicting for different virus families, as shown in Figure 4. This family-wise variability suggests that the learned features to predict for PHIs may have different generalization power across families, possibly a result of varying degrees of (dis)similarity between the virus families. Nevertheless, optimizing the predictive power for a single virus, e.g., SARS-CoV-2, requires a case-by-case experimental setup. Specifically in the case of SARS-CoV-2, one can potentially relax the leave-one-family-out evaluation, as we have prior knowledge about other species in its family, e.g., SARS and MERS, as well as their interactions with hosts and protein functions (Thiel et al., 2003). This is indeed a topic for further investigation.

There are several limitations that can be addressed by future efforts. One is the scarcity of training data for inter-species PPIs and this may be leveraged by transfer learning on the much larger intra-species PPI data available for humans and other model organisms. We also ignored other types of PHIs outside virus–human interactions in our current study, such as those of other hosts, e.g., plants and fishes, and other pathogens, e.g., bacteria and fungi, which have been shown to improve the prediction performance for Ebola (but not H1N1) in a previous method (Zhou et al., 2018) (see Table 1; the training sets TR3 and TR4 contain host proteins from species other than human). Additionally, predicting tissue-specific PHIs would also provide additional insights, as proteins of both human (Eagerberg et al., 2014) and viruses (Jarosinski et al., 2012) often have tissue-specific expressions and functions.
Acknowledgements
We would like to thank Maxat Kalmanov and Mona Alshahrani for their advice on earlier versions of this work. We also thank Jeffery Law for making public the mappings of the SARS-CoV-2 proteins.

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/1/3790-01-01.

References
Alshahrani, M. et al. (2017). Neuro-symbolic representation learning on biological knowledge graphs. Bioinformatics, 33(17), 2723–2730.
Ammari, M. G. et al. (2016). Hupid 2.0: a curated database for host-pathogen interactions. Database, 2016, haw103.
Ashburner, M. et al. (2000). Gene ontology: tool for the unification of biology. Nature genetics, 25(1), 25.
Bloks, L. S. et al. (2015). Mutations in ddx3x are a common cause of unexplained intellectual disability with gender-specific effects on wnt signaling. The American Journal of Human Genetics, 97(2), 343 – 352.
Chen, C.-Y. et al. (2016). Targeted inactivation of murine ddx3x: essential roles of ddx3x in placentaion and embryogenesis. Human Molecular Genetics, 25(14), 2985–2992.
Chen, J. et al. (2020). Predicting candidate genes from phenotypes, functions, and anatomical site of expression. bioRxiv.
Chollet, F. et al. (2015). Keras. https://keras.io.
Dohate-Macián, P. et al. (2018). The tpv4p channel links calcium influx to ddx3x activity and viral infectivity. Nature Communications, 9, 2307.
Dong, E. et al. (2020). An interactive web-based dashboard to track covid-19 in real time. The Lancet Infectious Diseases.
Dyer, M. D. et al. (2008). The landscape of human proteins interacting with viruses and other pathogens. PLOS Pathogens, 4(2), 1–14.
Dyer, M. D. et al. (2010). The human-bacterial pathogen protein interaction networks of bacillus anthracis, francisella tularensis, and yersinia pestis. PLOS ONE, 5(8), 1–12.
Eid, F.-E. et al. (2015). DeNovo: virus-host sequence-based protein–protein interaction prediction. Bioinformatics, 31(8), 1144–1150.
Fagerberg, L. et al. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Molecular & Cellular Proteomics, 13(2), 397–406.
Fajardo, J., et al. (2015). Disruption of specific rna-rna interactions in a double-stranded rna virus inhibits genome packaging and virus infectivity. PLOS Pathogens, 11(12), 1–22.
Farwlett, T. (2006). An introduction to ROC analysis. Pattern Recogn Lett, 27(8), 861 – 874.
Finlay, B. B. and Cossart, P. (1997). Exploitation of mammalian host cell functions by bacterial pathogens. Science, 276(5313), 718–725.
Foord, N. et al. (2017). The evolving role of cdr cell counts in hiv care. Current Opinion in HIV and AIDS, 12(2), 123–128.
Gkoutos, G. V. et al. (2018). The anatomy of phenotype ontologies: principles, properties and applications. Briefings in Bioinformatics, 19(5), 1008–1021.
Gordon, D. E. et al. (2020). A sars-cov-2-human protein-protein interaction map reveals drug targets and potential drug-repurposing. bioRxiv.
Guirimand, T. et al. (2015). Virhostnet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic acids research, 43(D1), D583–D587.
Guzzi, P. H. et al. (2011). Semantic similarity analysis of protein data: assessment with biological features and issues. Briefings in Bioinformatics, 13(5), 569–585.
Hoehndorf, R. et al. (2011). Phenomenet: a whole-phenome approach to disease gene discovery. Nucleic acids research, 39(18), e119–e119.
Hoehndorf, R. et al. (2013a). Mouse model phenotypes provide information about human drug targets. Bioinformatics.
Hoehndorf, R. et al. (2013b). Systematic analysis of experimental phenotype data reveals gene functions. PLoS ONE, 8(4), e60847.
Hoehndorf, R. et al. (2015a). Aber-owl: a framework for ontology-based data access in biology. BMC bioinformatics, 16, 1(1), 26.
Hoehndorf, R. et al. (2015b). The role of ontologies in biological and biomedical research: a functional perspective. Briefings in Bioinformatics.
Hua, T. et al. (2015). Prediction of host - pathogen protein interactions between mycobacterium tuberculosis and homo sapiens using sequence motifs. BMC bioinformatics, 16(1), 100.
Jain, S. and Bader, G. D. (2010). An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. BMC bioinformatics, 11(1), 562.
Jarsaul, B. et al. (2012). Fluorescently tagged pal47 of marrk’s disease virus reveals differential tissue expression of the tegument protein in vivo. Journal of Virology, 86(5), 2428–2436.
Jassal, B. et al. (2020). The reactome pathway knowledgebase. Nucleic acids research, 48(D1), D498–D503.
Jones, K. E. et al. (2008). Global trends in emerging infectious diseases. Nature, 451(7181), 990–993.
Kafkas, S. et al. (2018). PathophenoDB: linking human pathogens to their disease phenotypes in support of infectious disease research. bioRxiv.
Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27–30.
Kerrien, S. et al. (2011). The IntAct molecular interaction database in 2012. Nucleic Acids Research, 40(D1), D841–D846.
Köhler, S. et al. (2018). Expansion of the human phenotype ontology (hpo) knowledge base and resources. Nucleic Acids Research, page gky1105.
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.
Meenan, T. F. et al. (2017). Disease model discovery from 3,328 gene knockouts by the international mouse phenotyping consortium. Nature genetics, 49(8), 1231–1238.
Mikolajczyk, T. et al. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc.
Mukar, J. et al. (2016). Zika virus associated with microcephaly. New England Journal of Medicine, 374(10), 951–958.
Mukhopadhyay, A. et al. (2014). Incorporating the type and direction information in predicting novel regulatory interactions between hiv-1 and human proteins using a biclustering approach. BMC bioinformatics, 15(1), 26.
Nourani, E. et al. (2015). Computational approaches for prediction of pathogen-host protein-protein interactions. Frontiers in Microbiology, 6, 94.
Oellrich, A. et al. (2016). The digital revolution in phenotyping. Briefings in Bioinformatics, 17(5), 819–830.
Pesquita, C. et al. (2009). Semantic similarity in biomedical ontologies. PLoS computational biology, 5(7).

“output” — 2020/4/22 — 12:29 — page 7 — #7
Raudvere, U. et al. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1), W181–W198.

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

Sayers, E. W. et al. (2009). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 37(suppl_1), D5–D15.

Schofield, P. N. et al. (2012). Mouse genetic and phenotypic resources for human genetics. *Human Mutation*.

Schofield, P. N. et al. (2016). 25 - the informatics of developmental phenotypes. In R. B. R. D. Morris-Kay, editor, *Kauffman's Atlas of Mouse Development Supplement*, pages 307 – 318. Academic Press, Boston.

Slenter, D. N. et al. (2018). Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research*, 46(D1), D661–D667.

Smalli, F. Z. et al. (2019). Formal axioms in biomedical ontologies improve analysis and interpretation of associated data. *Bioinformatics*. btz920.

Smith, C. L. et al. (2018). 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. *The journal of machine learning research*, 15(1), 1929–1958.

The Gene Ontology Consortium (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1), D331–D338.

Thiel, V. et al. (2003). Mechanisms and enzymes involved in sars coronavirus genome expression. *Journal of General Virology*, 84(9), 2305–2315.

Thiæs, T. et al. (2012). Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics*, 28(6), 867–875.

Villaveces, J. M. et al. (2015). Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database*, 2015.

Wang, A. et al. (2017). Zika virus genome biology and molecular pathogenesis. *Emerging Microbes & Infections*, 6(3), e13.

Weitzman, M. D. et al. (2004). Interactions of viruses with the cellular dna repair machinery. *DNA Repair*, 3(8), 1165 – 1173. BRIDGE OVER BROKEN ENDS - The Cellular Response to DNA Breaks in Health and Disease.

Woolhouse, M. et al. (2012). Human viruses: discovery and emergence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1604), 2864.

Zhou, X. et al. (2018). A generalized approach to predicting protein-protein interactions between virus and host. *BMC genomics*, 19(6), 568.