Bioinformatics, 32(8), 2016, 1144–1150
doi: 10.1093/bioinformatics/btv737
Advance Access Publication Date: 16 December 2015
Original Paper

Sequence analysis

DeNovo: virus-host sequence-based protein–protein interaction prediction

Fatma-Elzahraa Eid[1,2,*], Mahmoud ElHefnawi[3] and Lenwood S. Heath[1]

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA, [2]Department of Systems and Computer Engineering, Faculty of Engineering, Al-Azhar University, Cairo, Egypt and [3]Biomedical Informatics and Chemoinformatics Research Group, Department of Informatics and Systems, Center of Excellence for Advanced Sciences, National Research Center, Giza, Egypt

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on 29 July 2015; revised on 8 December 2015; accepted on 12 December 2015

Abstract

Motivation Can we predict protein–protein interactions (PPIs) of a novel virus with its host? Three major problems arise: the lack of known PPIs for that virus to learn from, the cost of learning about its proteins and the sequence dissimilarity among viral families that makes most methods inapplicable or inefficient. We develop DeNovo, a sequence-based negative sampling and machine learning framework that learns from PPIs of different viruses to predict for a novel one, exploiting the shared host proteins. We tested DeNovo on PPIs from different domains to assess generalization.
Results: By solving the challenge of generating less noisy negative interactions, DeNovo achieved accuracy up to 81 and 86% when predicting PPIs of viral proteins that have no and distant sequence similarity to the ones used for training, receptively. This result is comparable to the best achieved in single virus-host and intra-species PPI prediction cases. Thus, we can now predict PPIs for virtually any virus infecting human. DeNovo generalizes well; it achieved near optimal accuracy when tested on bacteria–human interactions.
Availability and implementation: Code, data and additional supplementary materials needed to reproduce this study are available at: https://bioinformatics.cs.vt.edu/~alzahraa/denovo.
Contact: alzahraa@vt.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Predicting protein interactions of a novel virus with its host is a 3-fold data problem. First, few or no protein–protein interactions (PPIs) may be available for a novel virus to learn from, prohibiting the application of the typical single virus-host PPI prediction methods. Second, learning advanced protein features (like structure) for a novel virus protein can be time-consuming, a challenge to obtain (Perdigão *et al.*, 2015), and a generalization barrier. Third, sequence dissimilarity among viral families (King *et al.*, 2012) makes transferring knowledge from PPIs of some families to another family insufficient to model its PPIs. Predicting inter-species PPIs (which involve proteins from two different organisms) imposes additional difficulty over intra-species PPI prediction.

There are four interleaving components to address for the case of predicting PPIs of novel viruses with their shared host: PPI samples to learn from; features to represent the PPIs in a way that allows generalization; a learning approach that deals with the difficulties imposed by this type of prediction; and the consequent problem of generating negative PPI samples to learn from.

Features. Prediction methods need features extracted from PPI examples to learn. Four major feature extraction approaches are typical in pathogen-host PPI prediction. Sequence homology-based approaches (Zhou *et al.*, 2014), structure-based approaches (Dyer *et al.*, 2007; Zhou *et al.*, 2013b) and domain-based approaches (Evans *et al.*, 2009; Sarmady, 2010; Zhou *et al.*, 2013b) transfer known interactions from one protein to another based on sequence homology, structure similarity, or sharing of interaction interfaces, respectively (Nourani *et al.*, 2015; Zhou *et al.*, 2013a). Additionally, sequence dissimilarities among viral families hamper the generalization of these knowledge-based approaches to all novel viruses. The fourth category extracts features from the amino acid sequences of the protein pairs. It is successfully used on single virus-host PPI prediction (Cui *et al.*, 2012; Dyer *et al.*, 2011) and adopted in DeNovo.

Learning approaches. Three major learning approaches dominate inter-species PPI prediction (Nourani *et al.*, 2015): classic machine learning, where a model learns to discriminate between positive and negative PPIs (Cui *et al.*, 2012; Dyer *et al.*, 2011; Tastan *et al.*, 2009); data mining approaches that model true PPIs and measure a putative PPI similarity to them (Nouretdinov *et al.*, 2012); and transfer learning, where knowledge of PPIs is selectively transferred from a rich domain of PPIs to the target one for which prediction is needed using commonalities between the two domains (Kshirsagar *et al.*, 2013; Mei, 2013). We anticipate that data mining and transfer learning approaches are not adequate for novel virus PPI prediction, mainly because virus families carry no sequence similarities among themselves sufficient to make data mining or transfer learning from available PPIs of other viruses possible.

Data problem. In predicting PPIs for a virus with its host, either there are a sufficient number of known PPIs, few, or none. In the first case, usually a model is trained on the available PPIs to predict additional ones (Dyer *et al.*, 2011; Tastan *et al.*, 2009). As most viruses do not have many known PPIs, these single virus-host prediction approaches are not universally applicable. Recently, transfer learning is applied to learn from PPIs between a pathogen and its host and make prediction for another pathogen through a transfer function that maps knowledge between the domains (Kshirsagar *et al.*, 2013; Mei, 2013). However, transfer function design can be a challenge to build between viral families where members share no sequence similarities and may require prior knowledge of the two domains.

Negative sampling. A machine learning model needs to learn the difference between the positive and negative classes, yet there is no gold standard of non-interacting proteins nor is there enough information to build one (Shen *et al.*, 2007). Typically, negative examples are generated randomly, for other knowledge-based methods are mostly biased (Ben-Hur and Noble, 2005). Data mining approaches bypass negative sampling by modeling true interactions. For example, conformal prediction is used in a recent study (Nouretdinov *et al.*, 2012) where the conformance of a putative PPI is evaluated against the known PPIs and classified as positive if its prediction confidence is above a predefined threshold.

Solution overview. This study addresses the generalized problem of predicting PPIs between a virus and its host without the need to know any of their interactions in advance. The key idea is to learn interaction features for the host proteins rather than for specific viral proteins. The model depends only on protein sequences to extract features and generate unbiased negative examples to ensure generalization. Data are partitioned in a way that simulates the situation of a new virus to reasonably evaluate the performance. We further make three hypotheses on what attributes potentially guide the learning process and use our framework to test them to link biological and computational aspects of the prediction process. We name the proposed solution DeNovo as an analogy to *de novo* protein structure prediction, for it makes prediction without prior knowledge of any interactions between the two organisms to learn from. See Supplementary Material S8 for the solution workflow.

DeNovo demonstrates the feasibility of predicting interactions of a virus protein with human proteins by learning from the available interactions of other viruses that also infect human, even with little or no sequence similarity to the viral proteins in training. The framework achieves accuracy comparable to the best reported for intra-species PPI prediction (using sequence information) despite the challenging nature of the generalized problem. We additionally demonstrate DeNovo generalization to other pathogens.

2 Methods

DeNovo divides the virus-human PPIs data set into training and testing pairs in a way that simulates a novel virus prediction situation. DeNovo learns only from the amino acid sequence of the interacting pairs. To overcome the low prediction accuracy when random negative examples are used in training, a sequence-based negative sampling method is proposed. Its performance is assessed using other data sets and compared with other methods to demonstrate generalization and robustness.

2.1 Data set and preparation

Virus-human PPIs were collected from VirusMentha (accessed June 2014) (Calderone *et al.*, 2014). Each viral protein is given by its UniProtKB identifier and NCBI taxon identifiers of its corresponding virus and viral subfamily (Wheeler *et al.*, 2007). The data set contains 5753 unique interactions between 2357 human proteins and 453 viral proteins, covering 173 different virus species in 25 subfamilies. Sequences of the human and viral proteins were retrieved from UniProt (UniProt Consortium and Others, 2014).

Interactions were filtered down to 5445 PPIs between 2340 human proteins and 445 viral proteins, covering 172 viral species in 28 subfamilies (Supplementary Material S6). A PPI was eliminated if one of its proteins has no corresponding reviewed sequence in UniProt (UniProt Consortium and Others, 2014), if the viral protein belongs to a virus whose taxon identifier is not annotated as hosted by human in UniProt, or if the viral protein belongs to a virus that lies on a branch (on the taxonomy tree) with limited representation in the data set (a threshold of 10 PPIs was used).

2.2 Problem definitions

Novel virus PPI prediction problem. To the heart of DeNovo is the usage of virus-host PPIs available for a set of viruses and a shared host to predict PPIs of the host proteins with another virus that is foreign to the available virus set. The problem can be formalized as:

Given: a list of viruses $v_1, v_2, \ldots, v_n, v_{n+1}$ infecting a host h, a set of proteins H for h, a set of proteins V for $v_1, v_2, \ldots, v_n$, a set of positive interactions between members of H and V, and a set of proteins V_{n+1} for v_{n+1};

Find: all likely interactions between members of H and V_{n+1}.

Negative sampling problem. We call the process of generating negative examples from the positive ones 'negative sampling'. In the context of PPI prediction, a negative interaction (example) resulting from this process is a pair of proteins unlikely to interact. For the special case of virus-host PPIs between multiple viruses and a shared host, the negative sampling problem can be defined as:

Given: a set of host proteins H, a set of viral proteins V, and a list of positive interactions between members of H and V;

Generate: a list of negative (unlikely) interactions between members of H and V.

Data partitioning problem. Machine learning applications typically split the available data into training and testing sets. Here, we use ‘data partitioning’ to refer to dividing the data set based on some criterion into multiple subsets, each of which can serve as a testing set while the remaining subsets are grouped into a corresponding training set. This problem can be defined in the context of virus-host PPI prediction as:

Given: a set of host proteins H, a set of viral proteins V, a list of positive and negative interactions between members of H and V, and a criterion Cr on sets of interactions;

Find: a partitioning of the interactions into subsets, each of which satisfies Cr.

2.3 Hypotheses

Hypothesis 1 (features learning). We hypothesize that if we use a large number of interactions between different viruses and host proteins to train a classification model, it can learn to classify interactions with respect to the host proteins, rather than specific viral proteins. We anticipate that the classifier will extract features associated with the host proteins being interacting or not interacting. In such a case, PPI prediction would be possible for a foreign viral protein as the model will evaluate whether it is likely or unlikely to interact with each of the human proteins it was trained on.

Hypothesis 2 (shared interaction partner). We hypothesize that viral proteins with high sequence similarity can theoretically interact with a large number of similar host proteins. This hypothesis is based on two observations. First, viral protein interaction interfaces are likely to be different and more flexible than those of human proteins (Tokuriki *et al.*, 2009). Second, their interactions with the host proteins are extensively mediated with viral short linear motifs (SLiMs), which give them the flexibility to interact with a larger number of host proteins (Davey *et al.*, 2011). Using negative random sampling for the virus-host PPIs case will possibly lead to a large number of false negative examples.

Hypothesis 3 (sequence similarity). We hypothesize that strong sequence similarity between a tested viral protein and some viral proteins in the training set enhances the classification decision accuracy. Some interaction-related features may be conserved across the related viral proteins and consequently reflected in their sequences and the corresponding feature vectors. A novel virus not only has no known interactions, but it also may have no strong sequence similarities with any viral protein in the training set. We thus anticipate that the accuracy in the case of predicting PPIs of a novel virus can be much lower than when there are some related viral proteins in the training set.

2.4 Dissimilarity-based negative sampling

Random sampling. In the typical random negative sampling method, for a viral protein x, the set of human proteins that serve as partners in the negative intersections with x are picked randomly from the set of all human proteins (in the data set) that x does not interact with. We expect random sampling to produce many more incorrect negative examples than expected in the intra-species PPI case (according to Hypotheses 2 and 3 in Section 2.3), misleading the learning process and lowering the prediction sensitivity.

Proposed sampling. Our negative sampling method aims at reducing the expected large number of false negative examples by the random sampling. According to our hypotheses, if two viral proteins are similar in sequence, a human protein that interacts with one of them cannot be paired with the other as a negative example. The Dissimilarity-Random-Sampling algorithm (Supplementary Material S9) calculates all-versus-all global alignment bit-scores of the viral proteins. Bit-scores are then normalized, and their complements are used as the dissimilarity distances between the pairs of viral proteins. After excluding the unlikely negative examples based on a dissimilarity threshold T, random sampling is performed over the remaining negative interactions.

Dissimilarity threshold. Dissimilarity threshold T is used to pick less likely interactions as negative examples. We need to define an optimal value for the dissimilarity threshold (T^*) that best serve this purpose. According to the twilight zone concept of protein sequence alignments, when the sequence identity score between a pair of proteins falls below 20%, structure similarity between them is minimal (Rost, 1999). Considering the central dogma of genomics (sequence determines structure determines function), by selecting T^* at 0.8 (which corresponds to 20% sequence similarity), we increase the probability that the two proteins have no similar interaction interfaces, and consequently they are less likely to share an interaction partner. Thus, if viral proteins x and y interact with human proteins h_x and h_y, respectively, and the dissimilarity distance between x and y is >0.8, then x is less likely to interact with h_y, and thus the pair (x, hy) can serve as a negative (unlikely) interaction.

2.5 Data partitioning

Partitioning requirements. We need to divide the data set into training and testing subsets such that the testing subset contains interactions of viruses taxonomically far from those in the training set in a way that imitates predicting for a novel virus (or for a known virus with no or few known interactions). Additionally, we are interested in dividing the entire data set into multiple subsets, each of which can serve as a testing subset, while the remaining subsets are grouped into a corresponding training subset (Section 2.2). A random split cannot satisfy these requirements.

Taxonomy rank as partitioning criterion. When a novel virus emerges, it is more likely to be of a low taxonomy rank (an isolate or strain) than to be of higher rank (a species or genus), because a new virus typically evolves from known ones. However, we cannot specify a fixed taxonomy rank (species for example) as a partitioning criterion because the branches of the viral taxonomy tree significantly vary in length (Supplementary Material S4). Some subsets would carry interactions of viral proteins nearly identical or entirely different from the other viral proteins in the remaining subsets, making performance evaluation inconsistent.

Criterion 1 (partial blindness). In this scheme, interactions associated with leaf nodes under the same immediate ancestor in the viral taxonomy tree are grouped together into a single subset. This partitioning makes interactions from very related viruses (and consequently related viral proteins) fall into the same testing subset, but still allows some more distantly related viral proteins in the same virus family to be utilized in training (as they fall in some other subsets), hence the name partial blindness. See Supplementary Material S10 for an example. The testing, in this case, will be equivalent to predicting for a novel virus of the same rank as the ancestor of the testing subset.

Criterion 2 (complete blindness). We want to examine how the framework performs in the case of a novel virus that not only has no known interactions, but also its proteins have no strong sequence similarities with any viral protein in the training set (Hypothesis 3, Section 2.3). To test this case, we need to make the viral proteins in the testing set S completely dissimilar in sequence to all viral proteins in the training set R. We thus developed the complete blindness criterion to describe this requirement: interactions of viral proteins within the same viral family are grouped together into a single subset. Members of a single viral family share some similarity in sequence among themselves and no similarities to the other viral families (according to the ICTV taxonomy).

2.6 Learning model

Features. We used a feature extraction scheme developed for intra-species PPI prediction and successfully adapted to the inter-species case (Cui *et al.*, 2012; Dyer *et al.*, 2011; Shen *et al.*, 2007). It first clusters the 20 amino acids based on similarities of physiochemical properties known to drive most PPIs (dipoles and volumes of side chains). Residues of each protein are mapped to the corresponding cluster numbers. The frequency of each possible 3-mer is calculated in each mapped protein, generating a feature vector that is then normalized over [0,1] for each protein independently. The two normalized vectors of an interacting (or negative) pair are concatenated into a single feature vector representing the interaction. See Supplementary Material S11 for more details.

Classification model. We used support vector machines (SVMs) for their classification power and ability to tolerate high noise (Ben-Hur and Weston, 2010). We used the radial basis function kernel $K(u,v) = \exp(-\gamma * |u-v|^2)$ and the SVM implementation from LIBSVM version 3.18 (Chang and Lin, 2011).

Model parameters. To pick optimal values for the model parameters C (error penalty) and γ (the kernel parameter), a grid search over C and γ was performed with nested 5-fold cross-validation. Five exponentially-spaced values for each parameter were used; C in the range $[10^{-3},10^{1}]$ and γ in $[10^{-4},10^{0}]$, as suggested by best practice (Chang and Lin, 2011). The search was conducted at each value of the dissimilarity threshold T in the range [0,1] with a 0.1 step by training and testing SVM models on each training-testing pair for each (C, γ, T) combination. The resultant optimal parameter values are $C^* = 10$ and $\gamma^* = 10^{-3}$ (see Supplementary Material S1).

Partitioning. The data set was divided into 49 and 10 subsets according to Criteria 1 and 2, respectively. The i^{th} subset was used for testing at a time against a model trained on the interactions from all the remaining subsets, forming the i^{th} (R, S) training-testing pair. A negative to positive examples ratio of 1:1 is maintained throughout the study. Rounds of training and testing were performed over the different pairs with different dissimilarity thresholds T. The performance of the system was averaged over the entire data set for each of the partitioning schemes. Thus, no data are lost in testing while the performance evaluation is consistent with the goal of the study.

Accuracy measures. The performance was evaluated using standard accuracy, sensitivity and specificity for each (R, S) pair (Nourani *et al.*, 2015). We also report the support vector ratio for each model and prediction confidence. To quantify the overall performance across all the different (R, S) pairs, averages of the above measures were calculated (for each parameter value set), weighted by the sample size of the testing unit S in each pair.

2.7 Pilot and generalization studies

We conducted a set of pilot studies to assess DeNovo generalization and robustness. Each study is denoted STx, where x is the study number, for ease of reference. See Supplementary Material S12 for details of these studies.

Generalization to bacteria. ST1 examines how well DeNovo generalizes to predict bacteria-human PPIs for a novel bacterium. It also compares DeNovo performance to a recent transfer learning approach (Kshirsagar *et al.*, 2013). ST2-trained DeNovo on PPIs of *Arabidopsis thaliana* with a specific bacterium and tested the model on PPIs from other pathogens with *A.thaliana*. DeNovo was further trained on all VirusMentha virus-human PPIs and tested on bacteria-human PPIs in ST3.

Generalization within viruses. DeNovo was tested on virus-human PPIs for foreign viral and human proteins (not in the training set) in ST4, and on PPIs of foreign viral proteins with known human proteins in ST5. ST4 and ST5 work as application scenarios on how to use DeNovo to estimate the probability of a list of putative PPIs of a novel virus (ST4) and to generate a list of interacting and/or non-interacting human partners for that virus (ST5).

Assessing sequence-based features. Assessing how adding SLiM features may improve DeNovo accuracy, in ST6 we masked the SLiMs known in the viral proteins, grouped the PPIs of these proteins into a testing data set, trained DeNovo on the remaining PPIs, and compared the prediction accuracy when the SLiMs were masked and not masked.

Comparing with data mining-based learning and sampling. To compare DeNovo to data mining techniques, ST7 used a one-class classifier to model the true PPIs in both grouping schemes by training and testing on positive PPIs of different sets to assess how much true interactions these techniques may capture. In ST8, we compared our negative sampling method to one recently proposed (Mei and Zhu, 2015), which uses one-class SVMs to separate true PPIs into positives and negatives. We used that method on the ten viral family sets, and further tested the models on our negative samples of these families. We additionally compared the result of splitting the PPIs into positive and negatives against their original confidence score.

3 Results and discussion

DeNovo makes it possible to predict PPIs of a novel virus with human accurately by using PPIs from different viruses, sequence-based features and our negative sampling method. The results demonstrate the generalization and robustness of DeNovo.

3.1 Negative sampling

Random sampling. Random sampling is prone to produce false negative examples, but it is widely accepted in intra-species PPI prediction, for it is proven to be unbiased (Ben-Hur and Noble, 2005). When tested on the partial and full simulation of novel virus prediction (Criteria 1 and 2 at $T = 0$), random sampling accuracy achieved 60 and 48%, respectively. With the same features and SVM models, random sampling in intra-species and single virus-host PPI prediction achieved a maximum of 86 and 81.6%, respectively (Cui *et al.*, 2012; Shen *et al.*, 2007). This significant difference in accuracy demonstrates how difficult predicting for novel viruses is, and the high noise introduced by random sampling in this case.

Performance in partial blindness. The proposed negative sampling method outperforms the typical random sampling ($T = 0$) in both cases of partial and full simulation of novel virus prediction; see Figure 1 and Supplementary Material S2. Testing with the 49 (R, S) pairs resulting from the partially blind partitioning, the accuracy raised to 76.4% at $T^* = 0.8$, and eventually increased to 86% with $T = 0.9$. Accuracy, sensitivity and specificity were increased when T^* was used with 16, 16.8 and 17.6%, respectively, over those achieved with random sampling. The accuracy gain is about 60-fold of the standard error (0.27%). Additionally, a 12% drop was observed in support vectors ratio memorized by the trained model, which reflects less over-fitting and an easier learning task.

Performance in complete blindness and testing hypothesis 1. The accuracy improvement is maintained even when the sequences of the novel virus proteins are completely dissimilar to all other viral proteins used in training the model (Criterion 2); see Figure 1 and Supplementary Material S3. An increase in accuracy of 23.12% was achieved, which corresponds to a 96-fold standard error improvement over random sampling, supporting our learning hypothesis (Section 2.3).

Stability. Comparative high accuracies were also achieved when the model was trained with parameter values other than the optimal ones (C^* and γ^*), reflecting the stability of this method; see Supplementary Material S1. Additionally, the monotonic increase in accuracy is maintained with the increase in T, regardless of the partitioning criterion or the data set used for training; see Figure 1 and Supplementary Materials S1 and S2.

Testing hypothesis 2. The results from this negative sampling method also support our shared interaction partner hypothesis (Section 2.3), for when more dissimilarity was enforced in pairing the negative example partners (by increasing T), the accuracy and the sensitivity improved. Such improvement suggests that less noisy negative examples were generated, facilitating the classification learning process.

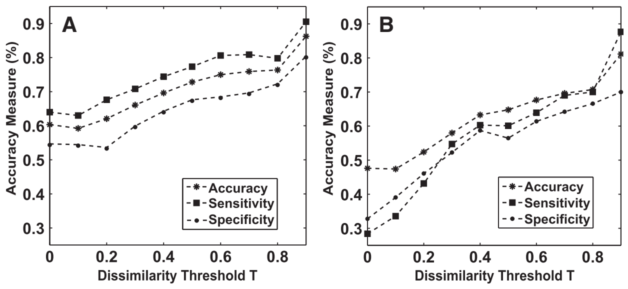

Fig. 1. Accuracy, sensitivity and specificity measured at different sequence dissimilarity threshold values, testing with **(A)** partial blindness, and **(B)** full blindness grouping criteria. Dotted lines are for visualization purposes

Comparison with data mining-based sampling. A recent negative sampling method clusters the true PPIs into two subsets: a large cluster of highly similar PPIs and a smaller distant cluster (Mei and Zhu, 2015). The study hypothesizes that most of the PPIs reported should be true ones and will cluster together, and the fewer PPIs falsely reported as positives will cluster away from the true ones. A model is then trained and tested on subsets of the two clusters. When we replicated the study for the case of viral families scenario, exceptionally high accuracy was achieved (91.6%). However, when we examined the members of each cluster, we found that (at least in the case of viral proteins) the two clusters do not necessarily correspond to real positive and negative PPI classes, suggesting that this method is not suitable for viral-host PPIs. See Supplementary Material S12 for details.

Advantages of the proposed sampling. Our negative sampling method picks the least likely interactions as negative examples to ease the classifier task for better discrimination between the true and unlikely interactions. Meanwhile, the dissimilarity distance constraint maintains the unbiased characteristic of the typical random sampling. First, it does not exclude human proteins similar to those interacting with the viral protein in question, x, but it excludes those interacting with viral proteins similar in sequence to x. Second, the dissimilarity distances are normalized independently for each viral protein (Supplementary Material S5).

With the demonstrated robustness and bias-free of the proposed dissimilarity-based negative sampling method, it can safely and efficiently replace the typical random negative sampling, solving a central challenge in PPI prediction problems (Nourani *et al.*, 2015).

3.2 Data partitioning

Subsets. The partial and full simulation of predicting PPIs of a novel virus (Criteria 1 and 2) give 49 and ten viral subsets respectively. See the taxonomy tree in Supplementary Material S4 with the 49 subsets color-coded, and Supplementary Material S7 for the 10 family names. Influenza A virus PPIs are grouped into three subsets (H1N1, H3N2 and H5N1), solely representing Orthomyxovirus in family grouping. Available PPIs from Hepatitis C virus genotypes fall into three subsets in Criterion 1, which contribute along with two other viruses into the viral family group of *Flaviviridae*. PPIs of *Herpesviridae* fall into seven subsets in Criterion 1, each of which

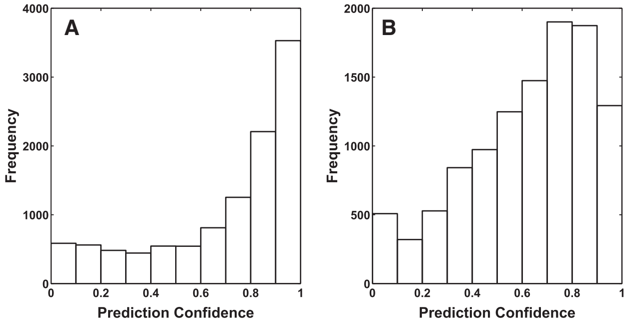

Fig. 2. Confidence of prediction with **(A)** partial blindness grouping and **(B)** full blindness grouping, using the model optimal parameters (C^*, γ^* and T^*)

roughly corresponds to one of the human herpesvirus types. Similarly, human viral types of *Adenoviridae* and *Papillomaviridae* fall into five and seven subsets, respectively.

Partitioning and negative sampling. Prediction in the case of partial blindness is more accurate and more confident than with full blindness partitioning; see Figure 2. However, partitioning does not affect and is not affected by the negative sampling at $T^* = 0.8$. At this threshold (in both schemes), negative human partners are picked from the partners of the other viral proteins that are placed in the midnight zone with respect to the viral protein in question. The purpose of partitioning is to simulate the situation of a novel virus, not to aid the negative sampling.

Learning. The accuracy difference between the two schemes is explained by and supports Hypothesis 3. In the partial blindness case, PPIs from related viral proteins (to the ones in testing) contribute to the training set, making the learning task easier. Testing on all interactions from a single family means that the model is not trained on any viral proteins with sequence similarity to the ones in testing. Thus, the model learns most of the discriminating features with respect to human proteins and other viral proteins and may miss some features that distinguish the interactions of the foreign viral proteins. This result confirms the generalization characteristic of DeNovo and further supports our learning hypothesis (Hypothesis 1).

3.3 Features

Sequence-based features. The main reason we utilize sequence-based features is to maintain generalization of DeNovo to any other PPI domain and to maintain prediction ability to any novel virus once its proteins are sequenced. Advanced features [structure, gene ontology (GO) or interaction domains] require prior knowledge that can be difficult and time-consuming to obtain (Guo *et al*., 2008; Shen *et al*., 2007). Sequence dissimilarity among viral families hampers using homology-based and structure-based features in novel virus-host PPI prediction. Amino acid sequence information has been proven to be sufficient for PPI prediction (Guo *et al*., 2008) (we also demonstrated this in Section 3.4).

Domain-based features. Although domain-based prediction is applied (on a limited scale) in the pathogen-host area (Zhou *et al*., 2013b), application for novel virus prediction needs prior knowledge of virus domains and related information (Nourani *et al*., 2015). As such knowledge is not available for many viruses and is difficult to obtain, one solution is to predict viral domains computationally for new viral proteins and apply domain-based methods on them. However, viral proteins are mostly loosely packed, enriched in disordered regions (Tokuriki *et al*., 2009), or of unknown and unmapped structure (Perdigão *et al*., 2015), making prediction of viral domains questionable.

SLiM-based features. Similarly, domain-based methods that use SLiMs in predicting viral PPIs face the same objections: the generalization problem for lack of data, and difficulty of SLiM prediction (for SLiMs are short and degenerate in nature). Some studies assign viral proteins whose sequences are predicted or known to have SLiMs similar to ones known in human as interacting candidates with human proteins carrying the corresponding counter domains (Evans *et al*., 2009; Sarmady, 2010). These methods are limited by the 239 known human SLiM classes and the 219 known viral SLiM instances (Dinkel *et al*., 2013).

We assessed how sequence motif information may help to improve the accuracy of DeNovo prediction by masking the viral SLiMs in testing PPIs (Study ST6 in Section 2.7). No change in accuracy was observed (Supplementary Material S12), suggesting that SVMs capture other interaction associated features that are sufficient to make the discrimination, or the available number of SLiMs is insufficient to make a difference.

3.4 Prediction approach

We reasoned that our negative sampling method is robust and not biased (Section 3.1). We thus find classic machine learning (which requires addressing the challenge of negative sampling) more suitable for the domain of predicting PPI of a novel virus than data mining and transfer learning approaches, as discussed below.

Data mining. In data mining-based learning, a putative PPI is classified as likely or unlikely based on its similarity to the known PPIs, without the need of negative examples. Our main concern is that dissimilarities among viral families are likely to cause a model representing true PPIs to be insufficient to describe interactions of foreign viruses. We tested this hypothesis in Study ST7 (Section 2.7). With a similarity threshold of 0.5, accuracies of 38% and 25% were obtained for partitioning schemes 1 and 2, respectively. These low accuracies suggest that modeling true PPIs of some viruses is insufficient to capture interaction-discriminating features for other viruses.

Transfer learning. Our concern of using transfer learning in virus-host PPI prediction for novel viruses is the need of prior knowledge to design a transfer function specific for each novel virus. We compared the performance of DeNovo, which uses sequence-based features, with transfer learning that uses more advanced features (Kshirsagar *et al*., 2013) on PPIs from 3 bacterial pathogens of different taxonomic orders (Study ST1 in Section 2.7). With near optimal accuracy of 97% and low support vector ratio, DeNovo demonstrates that its use of SVMs is sufficient (and probably a better fit) to capture discriminative features that otherwise need to be explicitly modeled in transfer learning. Another study uses homology knowledge transfer of GO features in HIV proteins (Mei, 2013). Such advanced features hamper generalization to most viruses where GO data can be time-consuming and a challenge to obtain.

3.5 Generalization

DeNovo generalizes well. We tested our model on five data sets assessing how it behaves in different pathogen domains, across pathogens with a shared host and on PPIs whose protein partners are foreign to the model. These tests demonstrate how to apply DeNovo and its generalization ability and robustness. However, we did not examine generalization to non-human host PPIs with viral proteins due to lack of data.

We used DeNovo on three PPI sets from different bacteria with human (Study ST1). Each set was tested once against a model trained on the other two, similar to the Criterion 2 scenario. DeNovo achieved average accuracy, sensitivity, specificity and support vector ratio of 97%, 94.5, 97.5 and 28.8, 94.5%, 97.5%, and 28.8%, respectively. This near-optimal result demonstrates the prediction ability of DeNovo. It also shows how predicting for bacteria is easier than for novel viruses, for bacteria do not have the dissimilarity constraint among families.

To further assess the difference between viral and bacterial proteins in prediction, DeNovo was trained on all VirusMentha virus-human PPIs and tested on bacteria-human PPIs (Study ST3). The model could recognize up to 69% of the tested PPIs. DeNovo was further trained on PPIs from a plant host with a bacterial pathogen and tested on proteins of different pathogens (viruses, fungi, and bacteria) that target the same host (Study ST2). Sensitivity reached 84.6%, while specificity was 65%, reflecting some differences of these pathogen proteins and/or their interaction interfaces.

When DeNovo was tested on PPIs from viral proteins foreign to VirusMentha, it correctly classified 94% of them (Study ST5). When both human and viral proteins were foreign (Study ST4), the prediction accuracy was approximately the same. This result supports our learning hypothesis (Hypothesis 1) that DeNovo can learn with respect to host proteins rather than specific viral proteins.

4 Conclusion

We introduced the problem of predicting virus-human PPIs for a novel virus, for which we do not have any known interactions. Our proposed solution, DeNovo, solved this problem efficiently by (i) exploiting the known PPIs between a large number of viruses and their shared host (human) to learn from; (ii) introducing a method to reduce the expected noise in generating the negative PPI examples from the true ones; and (iii) partitioning the data set in a way that simulates the situation of a novel virus, hence makes performance evaluation realistic. The model achieved significantly high accuracy while only using protein sequences to learn and predict, making the prediction possible for virtually any virus infecting human. We demonstrated that DeNovo generalizes well, and is robust and biologically sound.

Acknowledgements

Thanks to Dr Liqing Zhang, Prof. Clifford Shaffer and Dr Bert Huang for their helpful discussions. Thanks also to the anonymous reviewers for their insightful suggestions and comments.

Funding

This work was supported by the program of Egypt and the missions sector of the Egyptian ministry of higher education and the associated funding program is (VT-MENA program of Egypt). We express appreciation for the support of NSF [grant DBI-1062472].

Conflict of Interest: none declared.

References

Ben-Hur,A. and Noble,W.S. (2005) Kernel methods for predicting protein–protein interactions. *Bioinformatics*, **21**, i38–i46.

Ben-Hur,A. and Weston,J. (2010). A users guide to support vector machines. In: Carugo,O., Eisenhaber,F. (eds.) *Data Mining Techniques for the Life Sciences*. Springer, New York, pp. 223–239.

Calderone,A. *et al.* (2014) VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Res.*, **43**, D588–D592.

Chang,C.C. and Lin,C.J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27:1–27:27.

Cui,G. *et al.* (2012) Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics*, **13**, S5.

Davey,N.E. *et al.* (2011) How viruses hijack cell regulation. *Trends Biochem. Sci.*, **36**, 159–169.

Dinkel,H. *et al.* (2013). The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.*, **42,** D259–D266.

Dyer,M.D. *et al.* (2007) Computational prediction of host-pathogen protein–protein interactions. *Bioinformatics*, **23**, i159–i166.

Dyer,M.D. *et al.* (2011) Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect., Genet. Evol.*, **11**, 917–923.

Evans,P. *et al.* (2009) Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med. Genomics*, **2**, 27.

Guo,Y. *et al.* (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Res.*, **36**, 3025–3030.

King,A.M. *et al.* (2012). Virus Taxonomy: Classification and Nomenclature of Viruses. *Ninth Report of the International Committee on Taxonomy of Viruses, volume 9*. Elsevier, San Diego, CA.

Kshirsagar,M. *et al.* (2013). Multisource transfer learning for host-pathogen protein interaction prediction in unlabeled tasks. In *NIPS Workshop on Machine Learning for Computational Biology*, Lake Tahoe, NV, pp. 3–6.

Mei,S. (2013) Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. *PLoS One*, **8**, e79606.

Mei,S. and Zhu,H. (2015) A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. *Sci. Rep.*, **5**, 8034.

Nourani,E. *et al.* (2015) Computational approaches for prediction of pathogen-host protein-protein interactions. *Front. Microbiol.*, **6**, Published online.

Nouretdinov,I. *et al.* (2012). Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method. In *Pacific Symposium on Biocomputing*, vol. **311**, pp. 311–322, World Scientific.

Perdigão,N. *et al.* (2015). Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci.*, **112**, 15898–15903.

Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.

Sarmady,M. (2010). HIV protein sequence signatures for crosstalk with host proteins. PhD Thesis, Drexel University.

Shen,J. *et al.* (2007) Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad Sci. USA*, **104**, 4337–4341.

Tastan,O. *et al.* (2009). Prediction of interactions between HIV-1 and human proteins by information integration. In *Pacific Symposium on Biocomputing*. NIH Public Access, pp. 516–527.

Tokuriki,N. *et al.* (2009) Do viral proteins possess unique biophysical features? *Trends Biochem. Sci.*, **34**, 53–59.

UniProt Consortium and Others. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.

Wheeler,D.L. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.

Zhou,H. *et al.* (2013a) Progress in computational studies of host–pathogen interactions. *J. Bioinform. Comput. Biol.*, **11**, 1230001.

Zhou,H. *et al.* (2013b) Stringent DDI-based prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. *BMC Syst. Biol.*, **7**, S6.

Zhou,H. *et al.* (2014) Stringent homology-based prediction of H. sapiens-M. tuberculosis h37rv protein-protein interactions. *Biol. Direct*, **9**, 5.