DeNovo: Pilot and Generalization Studies

This document explains the pilot studies accompanying DeNovo. The codes and data needed to reproduce these studies are available on the paper web page.

1 Study ST1: Test DeNovo on Bacteria-Human

Here we test how well DeNovo, which is designed with virus properties in mind, generalizes to predicting PPIs for novel bacteria targeting human.

1.1 Data

A new data set of virus-human PPIs was collected from HPIDB (accessed September 21st, 2015). The data were originally generated in a separate study [1]. We used PPIs from three bacteria with human: *Bacillus anthracis*, *Yersinia pestis*, and *Francisella tularensis*; call them B1, B2, and B3 respectively. B1 belongs to a bacterial phylum different from that of B2 and B3, while B2 and B3 share the same class but differ in their taxonomic order.

B1 has 3057 PPIs, B2 has 402, and B3 has 1346. Approximately speaking, each Bacteria shares half of its human interaction partners with at least one of the other two bacteria: B1, B2, and B3 interact uniquely with 426, 1197, and 861 human proteins, and share 572, 911, and 887 partners with the other two bacteria, respectively.

1.2 Methods

Features were extracted from the positive examples of the three sets. Our dissimilarity-based negative sampling scheme was used to generate the negative examples at T^* with positive to negative ratio of 1:10 for each bacterial protein. An SVM model in each training-testing round was trained and tested with C^* and γ^*.

We followed the same testing scheme as in the family partitioning, where each of these different bacteria was set aside for testing, while the interactions from the other two bacteria were used for training.

1.3 Results

In average, prediction accuracy was about 97%. Sensitivity, specificity, accuracy were 94%, 97.2%, and 96.42% for B1; 94.8%, 98.3%, and 97.47% for B2; and

94.9%, 98.3%, and 97.32% for B3.

1.4 Discussion

Prediction for novel bacteria using DeNovo trained only on PPIs of the other two bacteria with human results in near optimal accuracy. Approximately half of the human proteins in testing appeared in training. This result supports our intuition that predicting for bacteria is an easier task than for viruses, for viral families have the constraint of dissimilarity in sequence.

2 Study ST2: Inter-Pathogen Prediction

How well DeNovo generalizes when used to predict PPIs of the host with pathogens other than the one(s) in training?

2.1 Data

The host in this study is *Arabidopsis thaliana*. In the training set, the interactions are collected from the PPIRA database of *A. thaliana* and the bacterium *Ralstonia solanacearum* (Accessed September 26, 2015). There are 3074 PPIs between 119 of the bacteria proteins and 1442 of the host proteins.

The testing data set is PPIs between the same host and other pathogens including some bacteria different from *R. solanacearum*. A set of 543 PPIs were collected from HPIDB (accessed September 24, 2015) of 149 unique pathogen proteins. The testing set contain 127 PPIs that belong to bacteria, 384 to fungi, and 30 to viruses.

2.2 Methods

Features were extracted from the positive interactions of the training set. All-versus-all sequence dissimilarity distances were calculated for the bacterium proteins to generate the negative examples with a 1:1 ratio of positives to negatives. An SVM model was trained on these examples.

Features were extracted for the testing set as normal. The dissimilarity distances were measured for each protein of the pathogens with the bacterial proteins in training to generate the negative examples for testing.

2.3 Results and Discussion

Sensitivity achieved was 84.6% while specificity was 65%. This result suggests that different pathogens may be interacting differently with the host proteins, through different interaction interfaces. This can be due to the nature of each pathogen and its proteins.

3 Study ST3: Virus-Bacterium Inter-Pathogen Prediction

How do viruses and bacteria differ in their interactions with human proteins?

3.1 Data

All VirusMentha PPIs used in the main study of DeNovo are used here for training. For testing, 8857 bacteria-human PPIs from HPIDB were employed. 2765 of these interactions have human partners that appear in VirusMentha.

3.2 Methods

Features of the VirusMentha PPIs were extracted, and negative examples generated at T^* using a ratio of 1:1 of positives to negatives. An SVM model was trained on all of the PPIs of VirusMentha at once with C^* and γ^* and tested on the bacteria-human PPIs positive examples after their features were extracted. No negative examples were created, for our concern here is how much of the true interactions can be captured from different pathogen interactions.

3.3 Results and Discussion

DeNovo could recognize 67.7% of the interactions with known human proteins, and 69.9% of the interactions with foreign human proteins. The total sensitivity was 69.2%. Although the model was trained on similar human proteins for the first set, it does not perform better than in the second set, suggesting different interaction interfaces employed by human proteins in the two cases or by proteins of the two types of pathogens.

4 Studies ST4 and ST5: Scenarios of Testing for a Novel Virus

Suppose a novel virus emerges, and we want to give a list of potential interacting and non-interacting human proteins, or we have a list of putative interactions of that novel virus with human proteins, and we want to assess how probable they are. Here we test both cases. In ST4, foreign viral proteins are tested against foreign human proteins to imitate the second situation. In ST5, only the viral proteins are foreign to imitate the first situation.

4.1 Data

All of VirusMentha PPIs serve as positive training examples, and the negative examples are generated with T^*. The testing set was obtained from an external data set, HPIDB, with 1612 PPIs of foreign human proteins for ST4, and 1830

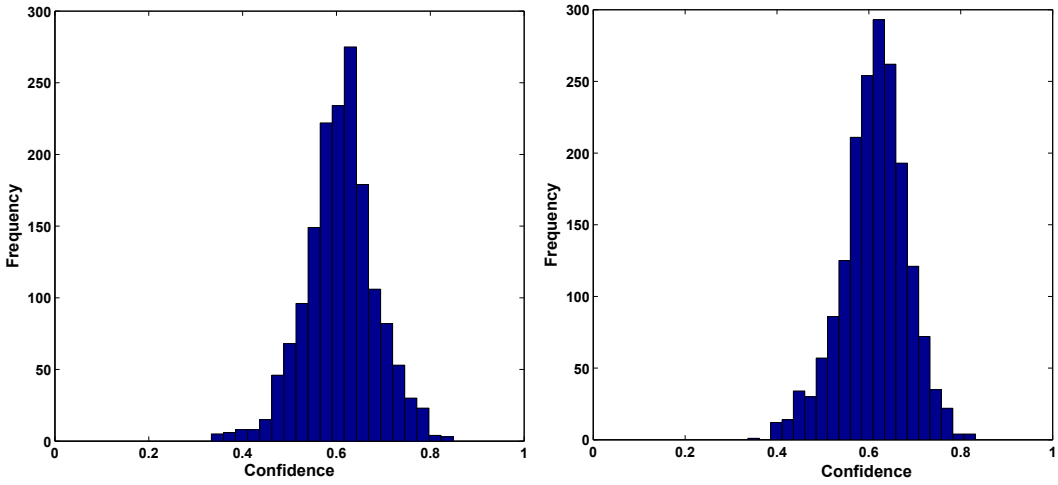

Figure 1: Confidence of positive PPI prediction in studies ST4 (Left) and ST5 (Right). Sensitivity in both cases was about 94%.

PPIs of the known human proteins for ST5. The viral proteins in both cases are foreign to the training set.

4.2 Method

A model is then trained on the training set with the optimal parameters (C^* and γ^*). For ST4, the features were extracted with no negative sampling, and the sensitivity was tested on the trained model. In ST5, the dissimilarity distance between the novel viral proteins and those of VirusMentha were calculated to generate negative examples for the testing set with T^* and a 1:1 ratio. The trained model was tested on the positive and negative interactions.

4.3 Results and Discussion

Recognition of the positive interactions in both studies was roughly the same, 94%. The specificity in study ST5 was 81%.

In real scenarios, where a list of putative virus-human PPIs are given to estimate their probabilities, confidence scores will help the researchers make informative decisions. Figure 1 shows the distribution of the confidence of prediction in both studies.

If a researcher interest is to generate a list of likely and/or unlikely interactions for a set of novel viral proteins, he/she should pair each viral protein with all human proteins in the training set, extract their features vectors, and use the trained model to assess the probability (confidence) of each interaction to be a positive one. The researcher then can select the interactions above (or below) a certain confidence threshold.

It seems that the model is sufficient to recognize the positive PPIs even if the human proteins are foreign, suggesting that the same interaction discriminating features are maintained in the foreign human proteins and across the foreign viral proteins, unlike the case in Study ST3 and ST2 where the recognition was relatively low where prediction was for PPIs of different pathogens than used in training the models.

5 Study ST6: Sequence Motifs

Sequence motifs can be a candidate prediction tool for virus-host PPI prediction, however, it does not fit novel virus for few are already known for different hosts. In this study, we demonstrate that DeNovo is sufficient to capture interaction-related features other than SLiM sequence information.

5.1 Data and Methods

We downloaded the SLiMs known in viruses from the ELM database, accessed Oct. 5, 2015. Using protein UniProt IDs, we intersected viral proteins carrying these ELMs with the viral proteins in VirusMentha (the data set used in the main study).

We then divided VirusMentha into two sets of interactions: the ELM-set that contains PPIs of viral proteins that have any of the ELMs, and the non-ELM set that contains the remaining PPIs. The non-ELM set works as a training set. We then generated the negative examples for the training set and for the ELM-set using the dissimilarity method of DeNovo with a 1:1 ratio of positive to negatives.

A masked set was generated from the ELM-set by masking the known SLiMs in each viral protein so that these SLiMs are not considered in feature extraction. Features were extracted from the training, masked testing, and non-masked testing sets. A SVM model with $C^* = 10$ and $\gamma = 10^{-3}$ was trained on the training set. The trained model was tested against the masked and non-masked testing sets. Confidence of prediction in both cases was retrieved, and then used to calculate standard error.

5.2 Results

The SLiM set consists of 219 instances in 316 unique viral protein identifiers. VirusMentha originally has 445 viral proteins. The two sets intersect in 39 proteins in 66 ELM instances. There were a total of 2015 positive PPIs and 1940 negative PPIs for the ELM-set, and 3430 positives and 3219 negatives in the non-ELM set. Accuracy, sensitivity, and specificity for the non-Masked set were 81.90%, 80.71%, and 83.06%, respectively; and for the masked set were 82.59%, 81.65%, and 83.53%, respectively. Standard error was 0.82%.

5.3 Discussion

The change in accuracy observed when the known viral ELMs are masked is insignificant, however, it was an increase in all evaluation measures, not a decrease. This suggests that DeNovo captures other features in viral proteins associated with interacting or non-interacting pairs that are different from viral SLiMs. The studied SLiMs may be too few to make a significant difference in accuracy.

5.4 Conclusion

Considering SLiMs as additional features in DeNovo is not likely to bring an increase in accuracy of prediction.

6 Study ST7: Data Mining Approaches

To assess how data mining techniques that learn from positive PPIs alone can be affected by the viral dissimilarity property among families, we used a one-class classifier to model the true PPIs in the two grouping criteria and test the accuracy of prediction as in the main study.

6.1 Methods

A One-class SVM model was trained on the positive interactions of the viral groups in Criteria 1 and 2 in a leave-one-out setting. The trained model in each case was then used to classify the positive examples in the remaining group into positive and negatives. The accuracy was measured by how many of the examples where classified as positives.

6.2 Results

In assessing how mining the true interactions alone may affect the prediction for novel viruses, with a confidence threshold of 0.5, the accuracy of prediction in the grouping criterion 1 was 38% and for criterion 2 was 25%. This demonstrates that using true interactions is not enough to capture discriminating features and model the true interactions, especially in the case of viruses with large dissimilarity in sequence with those in the training, and in the case when human proteins are also foreign to the training set.

7 Study ST8: One-class Negative Sampling

7.1 Motivation

Data mining approaches are used in PPI prediction to bypass the problem of generating negative examples needed for classic machine learning techniques. In a recent study [4], the authors combine these two approaches by generating

negative examples from the true interactions using one-class SVMs to label the highly similar examples as positives, and the remaining as negatives. They then used two-class SVMs to train and make prediction on these two classes.

The above method seems to compete with our proposed negative sampling method, for it was able to break the accuracy barrier of predicting in HIV-human PPIs domain that was set at about 85%, and moved it to higher than 89%. However, our intuition is that this method will be biased in the context of virus-host PPIs. We thus expect it to give high accuracy (in multiple virus-host PPIs case) that does not reflect real prediction power reinforced by the proposed negative sampling method, but reflects the method bias in this case. Thus, we designed and conducted three studies to assess this bias claim.

7.2 Methods

We used the family grouping (Criterion 2 in the main study) to assess how exceptional accuracy this negative sampling method can reach. In each one of the 10 families of PPI sets, we train an SVM model on the PPIs of that family, and then used the same model to give a score for each PPI. The 0.5 threshold was used as a balancing point. A PPI with a higher score is considered a positive example, and one below is considered a negative example.

We then used the leave-one-out testing strategy on the families, by training a two-class SVM on the examples from 9 families at a time and test on the examples from the remaining family. The accuracy measures were averaged over the 10 testing sets.

In the second study, we tested the negative examples that our dissimilarity-based method generates for each family on the models generated above. Each negative set from a family was tested on the model trained on the other families as usual.

In the third study, we retrieved the scores of confidence in positive PPIs from the VirusMentha data set for the total of 5445 PPIs used in the main study. Only 301 PPIs have score above 0.5. We used a one-class SVM to learn the structure or model of this data set. We then used the same model to generate scores of how likely each PPI example from the training set is to belong to the major component of similar interactions as modeled by the one-class SVM. We used again the threshold of 0.5 for comparison to the real score in the data set.

7.3 Results

Grouping true PPIs from the different 10 families into positive and negative examples on each family independently, and then testing using two-class SVMs resulted in exceptionally high accuracy. The weighted average accuracy was 91.6%, the weighted average sensitivity was 88.6% and the weighted average specificity was 94.5%. The support vector ratio was 19% in average. Using the same models to test DeNovo negative examples in each case, there was no single example correctly classified.

In the study that compares the original confidence score to the assigned score by one-class SVM models, 100 out of 301 examples originally of score above 0.5 were correctly identified, and 2722 examples of score below 0.5 were assigned score above 0.5. If we name the class of examples above 0.5 as positive, the results correspond to 33% sensitivity, 53% specificity, and 52% total accuracy. The support vector ratio in this study is 50%, reflecting the ease of discrimination between the two clusters.

7.4 Discussion

The low support vector ratio (in the models trained on examples generated by the one-class SVMs on viral family groups) indicates the easiness of the classification task. This observation does not align with the fact that virus families have almost no sequence similarities among themselves, which makes the prediction task very complicated. This reported increase in accuracy corresponds to about 80-fold of standard error in the investigated case of family prediction.

However, testing the generated models on the negative examples our dissimilarity-based negative sampling method produced resulted in flat zero specificity. Taking into consideration that our negative examples at dissimilarity threshold of $T = 0.8$ are highly likely to be true negatives as reasoned in the main study, the flat zero indicates the clustering performed using the one-class classifier does not align with the true positive and negative PPI classes.

From the bias in the scoring study, it is evident that scoring using a one-class SVM model does not correspond to positive and negative classes, and that the features used by the classifier to model the highly similar positive examples do not align with discriminating features that can separate true PPIs from negative ones. This further demonstrates the bias in this one-class negative sampling method presented in [4] when applied to virus-host PPIs.

The results support our intuition of the one-class SVM negative sampling bias (in the case of PPIs from multiple virus-host pairs). Our understanding of this problem is that the non-linear complicated problem of discriminating between positive and negative PPIs was transferred to a linearly separable problem using a clustering-like technique. There was no control on the features that guaranteed the two linearly separable classes were indeed corresponding to positive and negative PPIs. The two-class classifier was then used to train and test on these two classes; this is like clustering some data and then use a classifier to assess how well the two clusters are separated, thus exceptional high accuracy is expected. But these clusters do not necessarily correspond to the two PPI classes we are interested in, as demonstrated by our results above.

References

[1] Matthew D Dyer, Chris Neff, Max Dufford, Corban G Rivera, Donna Shattuck, Josep Bassaganya-Riera, TM Murali, and Bruno W Sobral. The

human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. *PLoS One*, 5(8):e12089, 2010.

[2] Meghana Kshirsagar, Jaime Carbonell, and Judith Klein-Seetharaman. Multisource transfer learning for host-pathogen protein interaction prediction in unlabeled tasks. In *NIPS Workshop on Machine Learning for Computational Biology*, 2013.

[3] Suyu Mei. Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. *PLoS ONE*, 8(11):e79606, 2013.

[4] Suyu Mei and Hao Zhu. A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. *Scientific Reports*, 5, 2015.