Predicting Protein Interactions of Intrinsically Disordered Protein Regions
Benjamin Yeh, Department of Computer Science, Stanford University

Motivation
Over the last two decades, many algorithms have been developed to predict regions of disorder (where there is no stable secondary or tertiary structure) within protein sequences. However, less is known about how these disordered regions interact with other proteins. Such research is important for several reasons: 1) a recent estimate suggests that over a third of human proteins are intrinsically disordered; and 2) these intrinsically disordered proteins (IDPs) have widespread roles in cellular processes, such as cell signaling and regulation. While there are many protein-protein interaction (PPI) prediction algorithms, they are largely based on knowledge of curated databases or models of energetically favorable interactions, both of which tend to rely on known protein structures. IDPs thus pose a unique challenge for PPI prediction.

Features
Each protein-protein interaction pair was represented by concatenating the feature vectors of its constituent proteins. The features of individual proteins, calculated with the R package prot4R, can be broadly classified into length-independent features (amino acid and dipeptide composition, and transition frequencies) and length-dependent features (pseudo-amino acid composition (PAAC) descriptors and autocorrelative measures). In total, this yields a 2494-dimensional vector for each protein and a 4898-dimensional vector for each protein-protein pair. The dataset was also readily augmented: since whether two proteins interact should not depend on the order of the proteins, both orderings of concatenation of the individual protein feature vectors were included. Therefore, the fully-featurized augmented dataset was a 176548-samples by 4898-features matrix.

Data
The labeled dataset was borrowed from Perovic et al consisting of 90253 unique protein-protein pairs where at least one protein was considered “intrinsically disordered” by DisProt. Within this dataset, 19796 (22%) pairs were considered to be interacting (positive) and 70457 (78%) to be non-interacting (negative) by HIPPIE. This dataset was then filtered for proteins with length greater than 50 amino acids to avoid trivial length-dependent auto-correlative feature descriptors, leaving 88274 pairs.

Results
The linear models generally demonstrated less variance (overfitting) but higher bias than the nonlinear models. The results were not surprising given that the PCA plots failed to show strong evidence of linear decision boundaries. The RF models fit the data very well and had the best generalized performance on the validation dataset, despite significant overfitting. The best AUROC score achieved here (0.8268) surpassed that reported by Perovic et al. (0.745), which may be due to our data augmentation method. Unfortunately, interpreting the performance of the RF models is difficult due to their ensemble nature and the PCA dimensionality-reduction step prior to training. It is therefore almost impossible to concretely explain what protein pair characteristics are favorable for interactions versus non-interactions.

Future Work
Simple extensions of current work include considering all 4898 features instead of the PCA-reduced 446 features; finer hyperparameter tuning to reduce overfitting; and trying more advanced nonlinear models, such as larger neural networks. Broader datasets can be collected by incorporating diverse data sources (e.g., D2P2, STRING, BioGRID, and proxi) with unique experimental and computational descriptions of PPIs. New featurization strategies that may improve separation of labelled data include using co-evolution information and energy models to account for stabilization of disordered domains upon interactions with other proteins, and using embeddings of protein complexes derived from PPI networks.