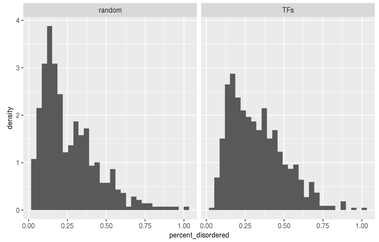
Predicting Protein Interactions of Intrinsically Disordered Protein Regions

Benjamin Yeh, Department of Computer Science, Stanford University

CS 229 Project, Autumn 2018
https://github.com/bentyeh/disprot
bentyeh@stanford.edu

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[1,2,3,4]. However, less is known about how these disordered regions interact with other proteins. Such research is important for several reasons: 1) a recent estimate[5] 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[6,7]. While there are many protein-protein interaction (PPI) prediction algorithms[8], they are largely based on knowledge from 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.



Comparison of disorder prevalence between transcription factors and control (random) sequences from the human genome. A motivating example in the study of disordered PPI's is the difference in percent disorder between transcription factors (TFs) and random sequences in the human genome. TFs are thought to recruit transcriptional complexes (such as mediator) via their disordered domains.

Data

The labeled dataset was borrowed from Perovic et al.[9], consisting of 90253 unique protein-protein pairs where at least one protein was considered "intrinsically disordered" by DisProt[10]. Within this dataset, 19796 (22%) pairs were considered to be interacting (positive) and 70457 (78%) to be non-interacting (negative) by HIPPIE[11]. 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.

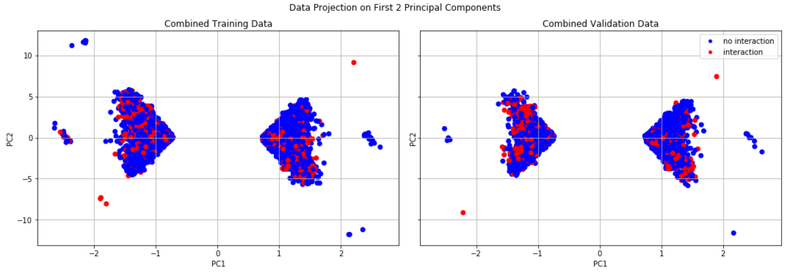
References

1. Romero, Pedro, et al. "Sequence complexity of disordered protein." *Proteins: Structure, Function, and Bioinformatics* 42.1 (2001): 38-48.
2. Peng, Kang, et al. "Length-dependent prediction of protein intrinsic disorder." *BMC Bioinformatics* 7.1 (2006): 208.
3. Ishida, Takashi, and Kengo Kinoshita. "PrDOS: prediction of disordered protein regions from amino acid sequence." *Nucleic Acids Research* 35. (2007): W460-W464.
4. Dosztanyi, Zsuzsanna, et al. "The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins." *Journal of Molecular Biology* 347.4 (2005): 827-839.
5. Ali, Muhammad, and Yiva Ivarsson. "High-throughput discovery of functional disordered regions." *Molecular Systems Biology* 14.5 (2018): e8377.
6. Wright, Peter E., and H. Jane Dyson. "Intrinsically Disordered Proteins in Cellular Signalling and Regulation." *Nature Reviews Molecular Cell Biology* 16.1 (2015): 18–29.
7. Liu, Jiangang et al. "Intrinsic Disorder in Transcription Factors." *Biochemistry* 45.22 (2006): 6873–6888.
8. Singh, Rohit, et al. "StructDNet: a web service to predict protein-protein interactions using a structure-based approach." *Nucleic Acids Research* 38. (2010): W508-W515.
9. Perovic, Vladimir et al. "IDPpi: Protein-Protein Interaction Analyses of Human Intrinsically Disordered Proteins." *Scientific Reports* 8.1 (2018): 10563.
10. Piovesan, Damiano, et al. "DisProt 7.0: a major update of the database of disordered proteins." *Nucleic Acids Research* 45.D1 (2016): D219-D227.
11. Schaefer, Martin H., et al. "HIPPIE: Integrating protein interaction networks with experiment based quality scores." *PloS One* 7.2 (2012): e31826.
12. Xiao, Nan, et al. "prot/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences." *Bioinformatics* 31.11 (2015): 1857-1859.
13. Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12.Oct (2011): 2825-2830.
14. Oates, Matt E., et al. "D2P2: database of disordered protein predictions." *Nucleic acids research* 41.D1 (2012): D508-D516.
15. Szklarczyk, Damian, et al. "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible." *Nucleic acids research* (2016): gkw937.
16. Chatr-Aryamontri, Andrew, et al. "The BioGRID interaction database: 2017 update." *Nucleic acids research* 45.D1 (2017): D369-D379.
17. Riffle, Michael, et al. "ProXL (Protein Cross-Linking database): A platform for analysis, visualization, and sharing of protein Cross-Linking mass spectrometry data." *Journal of proteome research* 15.8 (2016): 2863-2870.
18. Dosztányi, Zsuzsanna, Bálint Mészáros, and István Simon. "ANCHOR: web server for predicting protein binding regions in disordered proteins." *Bioinformatics* 25.20 (2009): 2745-2746.
19. Liu, Xiaoxia, et al. "Identifying protein complexes based on node embeddings obtained from protein-protein interaction networks." *BMC bioinformatics* 19.1 (2018): 332.

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 protr[12], 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 2449-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.

All data were normalized as z-scores (mean 0, variance 1) then visualized through PCA plots to understand how well featurization separated the binary-labelled data. To reduce the feature complexity, only the top 446 principal components (corresponding to singular values > 1) were retained. Finally, the dataset was renormalized as z-scores and split 60-20-20 into training, validation, and test sets.

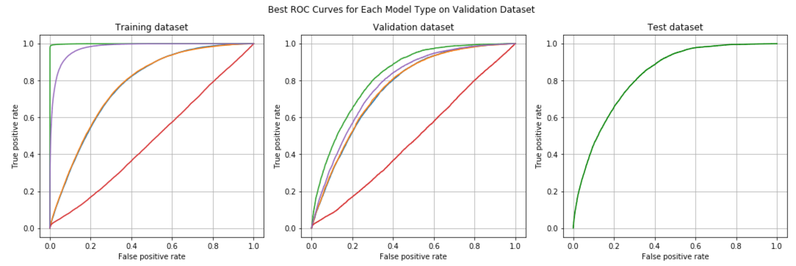


Projection of training and validation data onto first 2 principal components. PCA plots did not reveal any clear linear decision boundary, suggesting that nonlinear models may be more appropriate. The symmetry of the plots is a likely consequence of data augmentation procedures.

Models

The linear models tested included L2-regularized logistic regression and support vector machines (SVM). The non-linear models tested included random forest (RF) classifiers, Gaussian-kernel SVMs, and neural networks. Each model type was evaluated at several hyperparameters. The Python package scikit-learn[13] was used to build and train the models and evaluate their accuracy using AUROC as the primary metric.

Results



Model	Hyperparameters	AUROC (train)	AUROC (validation)	AUROC (test)
Random forest	100 trees, max depth 50	0.9995	0.8245	0.8268
Neural network	α=0.001	0.9775	0.7832	--
SVM (linear)*	C=0.1	0.7731	0.7625	--
Logistic regression	C=0.1	0.7713	0.7605	--
SVM (RBF)**	C=1	0.4794	0.4822	--

ROC curves across training and validation datasets for the best-performing (on the validation dataset) model of each model type. The ROC curve for the overall best-performing model is also shown on the test dataset. AUROC values in the legend correspond to performance on the validation dataset. *Did not converge after 1000 iterations. **Did not converge after 50 iterations.

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 turning 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[14], String[15], BioGRID[16], and proxl[17]) 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[18], and using embeddings of protein complexes derived from PPI networks[19].