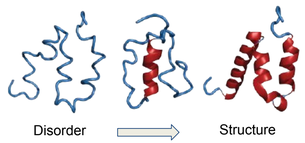
Ensemble Prediction of Intrinsically Disordered Regions in Proteins

Ahmed Attia
aattia@stanford.edu

Introduction

Intrinsically Disordered Regions (IDRs) are regions of proteins which lack a stable structure. Defying the structure-function paradigm, IDRs have functional roles in foundational biological processes including transcriptional regulation, translation, and cellular signal transduction [1].



Because of this importance, a diversity of predictors has emerged, each utilizing various protein characteristics, such as low hydrophobicity and high net charge, to predict disorder. Here we test whether an ensemble method can combine approaches to provide better predictions than its primary predictors.

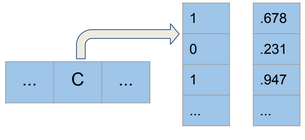
Dataset Construction

There are two main classes of reliable data on IDRs. The first is annotation of IDRs from literature collected in the Disprot database. The second is IDRs inferred through experimental sources such as X-ray crystallography and NMR spectroscopy.

We construct a dataset of 438 proteins utilizing disorder and order annotations from scientific literature, X-ray crystallography, and NMR spectroscopy. The data is pulled from MobiDB with literature annotations coming from Disprot.

Ensemble Inputs

For every protein, we obtain order/disorder predictions for each of its amino acids using 11 different IDR predictors.



Then we construct two feature vectors for every amino acid, one with the primary predictors' binary predictions and another with the primary predictors' continuous values reflecting confidence of an amino acid being ordered/disordered.

We compare the performance of ensemble methods with datasets containing each type of feature vectors to the performance of primary predictors.

Number of Examples	102007
Number of Features	11

Methodology

I. Majority Rule Voting, Averaging

- Baseline ensemble methods were majority rule voting (used for binary features) and averaging (used for continuous features).
- Majority rule voting predicts based upon the most popular binary prediction of the component IDR predictors.
- For averaging, if the probability of an amino acid being disordered averaged over the 11 primary IDR predictors was more than half then the amino acid was classified as disordered.

II. Logistic Regression

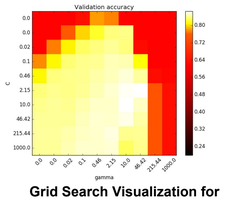
- Logistic regression with L2 regularization was used for classification (with both binary and continuous features)

III. SVM

- For the SVM estimator we used a gaussian (rbf) kernel
- Hyperparameter tuning of γ and C: γ is a parameter of the rbf below, C trades off a simpler decision surface with misclassification error on the training data.
- Grid search over logarithmic range used for hyperparameter tuning.

$$K(x, x') = -\gamma \|x - x'\|^2$$

Radial Basis Function



Grid Search Visualization for Continuous Features

IV. Random Forest

- A random forest classifier with 10 trees was used for classification for both the binary and continuous features.

Primary Predictor Results

Individual Predictor	Sensitivity	Specificity	Accuracy	Individual Predictor	Sensitivity	Specificity	Accuracy
DisEMBL465	.4191	.9503	.7414	GlobPlot	.4073	.9100	.6952
disHL	.5301	.7205	.6456	IUpredL	.6442	.9167	.8056
ESpritzD	.4197	.9574	.7459	IUpredS	.5496	.9409	.7870
ESpritzN	.6862	.8681	.7966	JRONN	.7258	.8125	.7789
EspritzX	.5714	.9553	.8043	VSL2b	.8067	.7750	.7875
foldIndex	.6223	.7593	.7061				

Table 1: Sensitivity, Specificity, and Accuracy for each of the primary DR Predictors.

Ensemble Results

- For logistic regression, SVM, and random forest 3-fold cross validation is used with the validation metrics reported.

Binary Feature Ensemble	Sensitivity	Specificity	Accuracy	Continuous Feature Ensemble	Sensitivity	Specificity	Accuracy
Majority Rule	..9257	.6520	.8181	Averaging	.9749	.4690	.7785
Logistic Reg.	.6905	.9158	.8259	Logistic Reg.	.7290	.9182	.8446
SVM	.7204	.9169	.8385	SVM	.8953	.9658	.9384
Random Forest	.7163	.9183	.8377	Random Forest	.8432	.9674	.9191

Table 2: Sensitivity, Specificity, and Accuracy for each of the Ensemble Predictors.

Analysis

- Best performing method is SVM using continuous features as ensemble inputs, achieving a ~13% improvement in accuracy over the most accurate primary predictor while topping both the maximum sensitivity and specificity.
- Continuous feature ensemble methods are the most promising: the next best performing method is random forest also on continuous features
- Nonlinear methods are the best performers for ensemble learning, but even the simple majority rule has slightly better accuracy than the best individual primary classifier.
- The improvement in performance with ensemble method is likely a result of eliminating the weaknesses of individual predictors. These weaknesses can stem from different sources of training data [2] and result in highly variable predicted levels of disorder across individual predictors [3].

Future Work

- Incorporate additional primary IDR predictors
- Dimensionality reduction of constructed feature vectors (similar primary predictors)
- Further hyperparameter tuning

Acknowledgements

- BiocomputingUP Lab, University of Padua: Providing data on IDR annotations and predictions.

[1]H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," Nature Reviews Molecular Cell Biology, vol. 6, no. 3, pp. 197–208, 2005.
[2]L. P. Kozlowski and J. M. Bujnicki, "MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins," BMC Bioinformatics, vol. 13, no. 1, p. 111, 2012.
[3]J. Atkins, S. Boateng, T. Sorensen, and L. Mcguffin, "Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies," International Journal of Molecular Sciences, vol. 16, no. 8, pp. 19040–19054, 2015.