Disordered Function Conjunction:

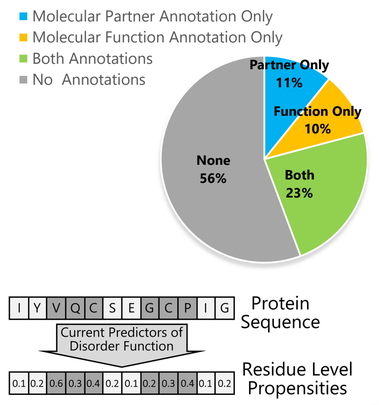
On the in-silico function annotation of intrinsically disordered regions

COMPUTER SCIENCE

Authors: Sina Ghadermarzi, Akila Katuwawala, Christopher J Oldfield, Amita Barik, Lukasz Kurgan
Presenter: Sina Ghadermarzi – PhD Student – ghadermarzis@vcu.edu

Motivation

- Intrinsically disordered proteins carry out important biological functions despite lacking stable 3D structure
- Many experimentally determined disordered regions lack functional annotation
- Current computational method predict disorder functions at the residue level
- Novel computational methods are needed to predict functions of disordered regions

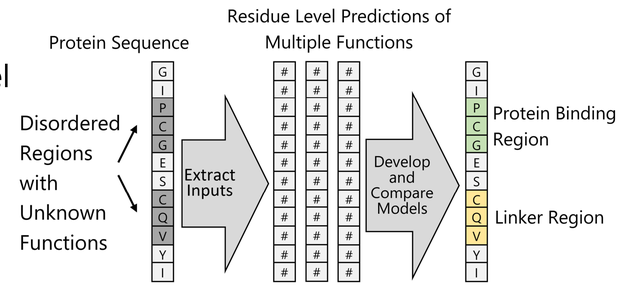


Goal

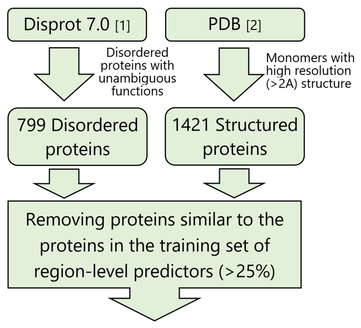
Investigate feasibility of developing an accurate predictor of functions of disordered regions

Methods

- Compare use of different inputs (residue level function predictions and sequence itself)
- Compare different computational predictive models
- Evaluate quality of different setups for prediction of functions of disordered regions



Dataset



Features

Protein sequence with a specified region

I Y V Q C S E G C P I G

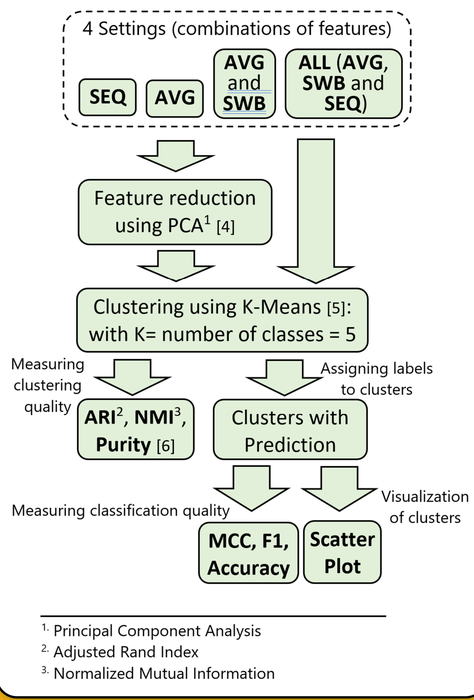
Residue level predictors: ANCHOR, fMoRFpred, DisoRDPbind (DNA, RNA, Protein), DFLpred [3]

Extracting region level from residue level features

SEQ (21 features)	**SWB** (6 features)	**AVG** (6 features)
Amino acid compositions and sequence length	Sliding window based features	Average based features

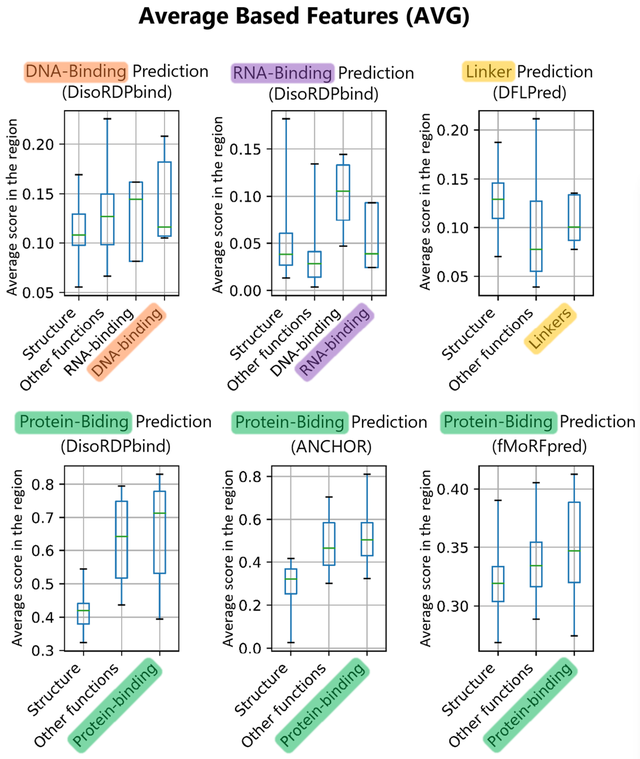
3 Groups of region level features

Method- Evaluation



1. Principal Component Analysis
2. Adjusted Rand Index
3. Normalized Mutual Information

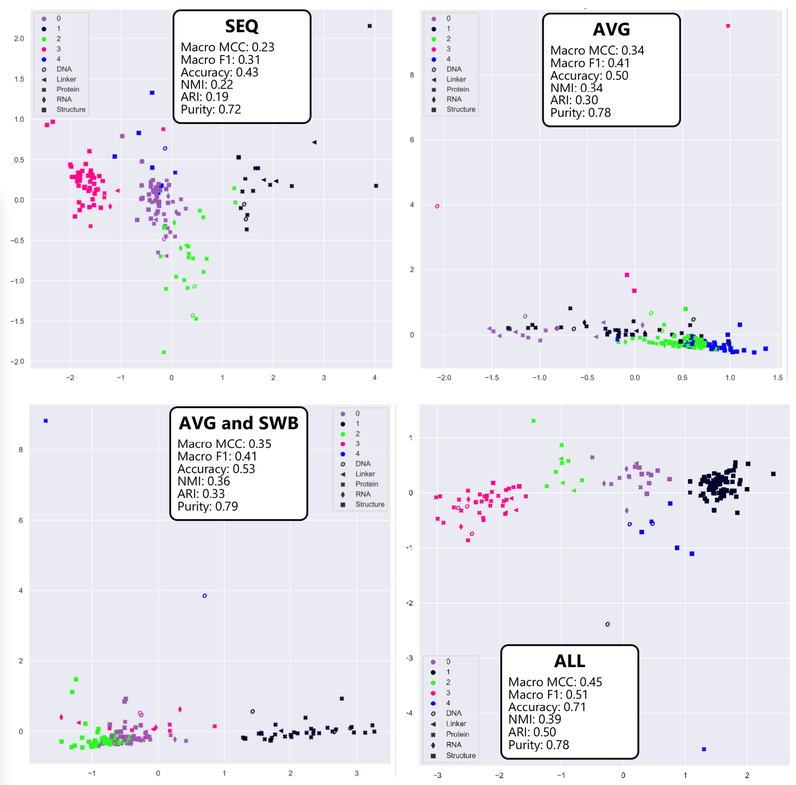
Results



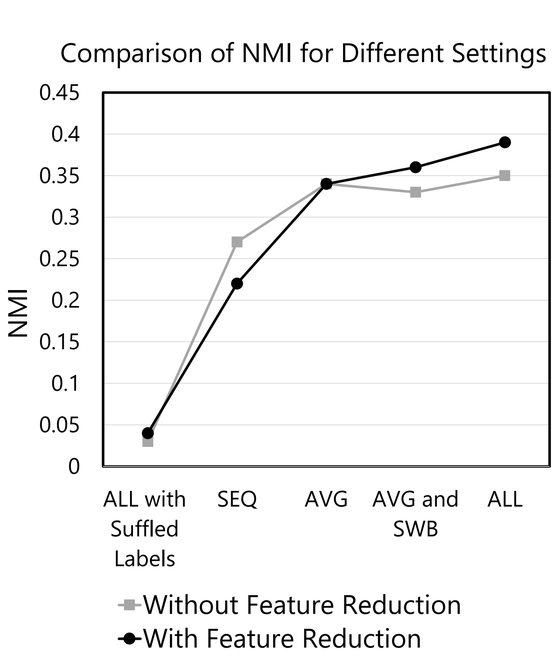
Box plots of averages of residue level prediction. Each panel compares the average-based features between three sets of regions: structured, disordered regions with other functions, and the disordered regions with the predicted function.

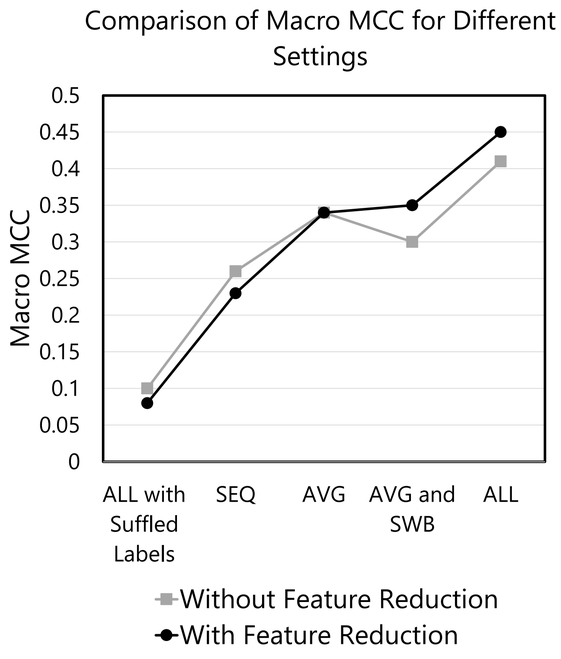
Insights

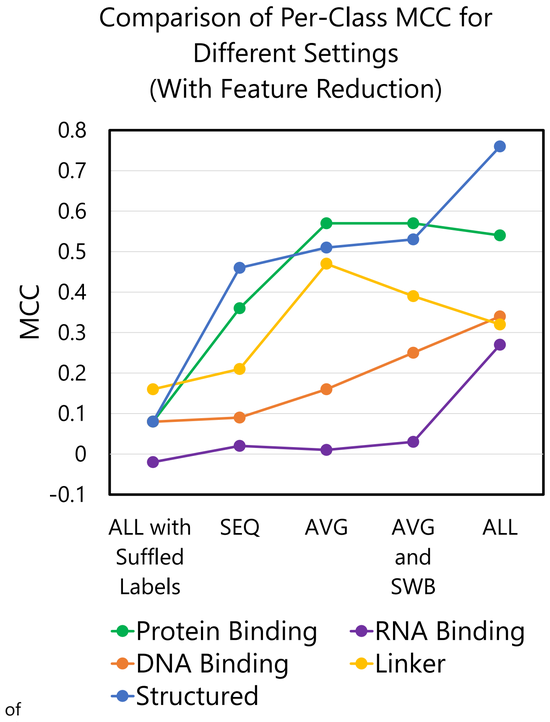
- AVG is modestly predictive
- Use of clustering and feature reduction helps to better differentiate functions for the regions
- Both clustering and classification quality are improved with addition of each group of features
- Different prediction quality is observed for different labels (functions) which is expected due to different number of labeled data available for them



Visualization of clusters and labels. The clusters for four settings of features (with dimensionality reduction) are visualized in scatter plots. The respective clustering and classification scores are shown in boxes. The colors show the clusters and shape of the markers show the label of the proteins. Reduction to two dimensions were done using Kernel-PCA method with polynomial kernel.







Clustering and classification quality. Clustering quality measured by NMI and overall classification quality measured by macro MCC are shown for both of with and without dimensionality reduction cases. The "ALL with shuffled labels" shows the value of the measure when calculated for randomly shuffled labels.

Summary and Conclusion

- Functions of disordered regions can be predicted accurately using residue level predictions
- Combining the predictions of multiple residue level functions leads to more accurate region level predictions
- Addition of sequence information provides further improvements

References

[1] Piovesan, D., et al., *DisProt 7.0: a major update of the database of disordered proteins.* Nucleic Acids Research, 2016. 45(D1):D219-27.
[2] Burley, S.K., et al., *RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Research, 2018. 47(D1):D464-74.*
[3] Katuwawala, A., S. Ghadermarzi and L. Kurgan, *Chapter Nine - Computational prediction of functions of intrinsically disordered regions*, in *Progress in Molecular Biology and Translational Science*, V.N. Uversky, Editor. 2019, Academic Press. p. 341-369
[4] Jolliffe, I.T. and J. Cadima, *Principal component analysis: a review and recent developments.* Philos Trans A Math Phys Eng Sci, 2016. 374(2065):20150202.
[5] Jain, A.K., *Data clustering: 50 years beyond K-means.* Pattern Recognition Letters, 2010. 31(8):651-66
[6] Pfitzner, D., R. Leibbrandt, and D. Powers, *Characterization and evaluation of similarity measures for pairs of clusterings.* Knowledge and Information Systems, 2009. 19:361-94