Accurate Identification of Protein Disorder from Sequence

Sumaiya Iqbal, Denson Smith, Md Tamjidul Hoque
Department of Computer Science, University of New Orleans, New Orleans, LA 70148
{siqbal1,dsmith8,thoque}@uno.edu

Bioinformatics and Machine Learning Lab
Computer Science Department
University of New Orleans
http://cs.uno.edu/~tamjid/

Abstract

Motivation. Association of intrinsically disordered proteins (IDPs) or regions (IDRs) of proteins with critical human diseases has urged their identification to be a crucial research in the area of bioinformatics and computational biology. About 80% of the human disordered proteins contain at least one amyloidogenic region that are directly linked with Parkinson's and Alzheimer's diseases or, type II diabetes.

Contributions. The aims of this research are to improve the current prediction accuracy of disordered protein residues from the protein sequence alone, and to investigate the intriguing connections among intrinsic disorder and human diseases that can facilitate IDPs based drug discovery. With a view to this, we developed DisPredict3, to classify ordered versus disordered residues from protein sequence.

Conclusions. DisPredict3 showed competitive performance. It is possible to identify the critical binding regions within disordered regions that undergo disorder to structure transitions using disorder prediction tool. Thus, it will be interesting to further investigate the tool for binding region or, induced folding region prediction.

This work is supported by the Louisiana Board of Regents (BoR) through the Board of Regents Support Fund, LEQSF (2013-16)-RD-A-19. We gratefully acknowledge BoR.

What is Protein Disorder?

- Proteins may misfold and remain unstructured, either in full sequence or, in a region, known as IDPs or, IDRs
- Being unstructured, IDPs have higher accessible surface for interaction with partners. Thus, they are biologically active
- IDPs (or, IDRs) are characterized by lack of secondary and tertiary structures.

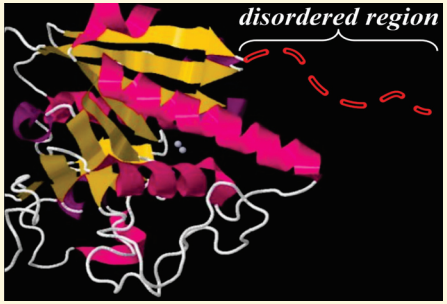
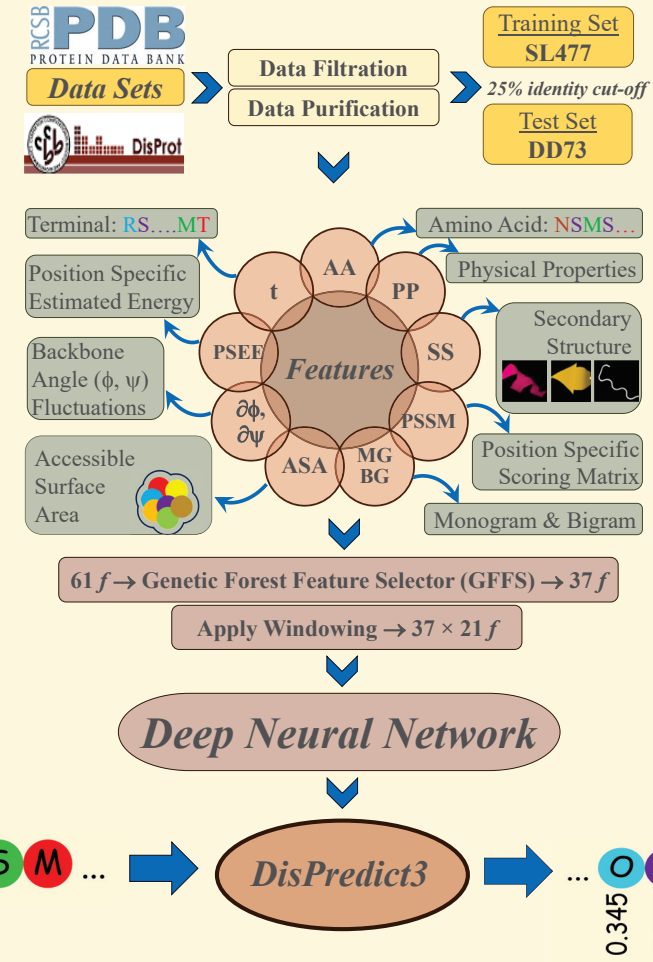


Figure 1: Crystal structure of human GTA complexed with lactose (ID: 1ZI1, Length: 293, IDR: 176 – 293).

Why Computational Method?

- Experimental annotation of disordered proteins is costly, both in terms of money and time.
- Computational tools play an alternative and vital role in understanding functions of disordered proteins.

Illustration of Workflow



Structural Properties of Disorder

- Disordered proteins or, regions are are characterized by
  - High solvent exposure
  - Marginal secondary structure
  - Energetically unfavorable condition

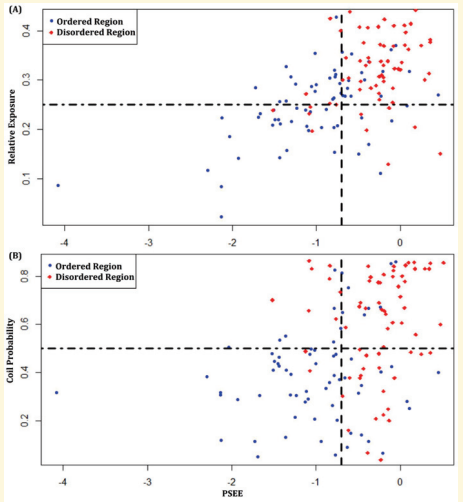


Figure 2: Correlation between (A) PSEE vs. relative exposure and (B) PSEE vs. coil probability of ordered regions (*blue circle*) and disordered regions (*red diamond*) of DisProt database. The vertical dashed line separates the average PSEE of all ordered and disordered region and the horizontal dash-dotted line separates the ordered and disordered regions with more and less than (A) 25% exposure and (B) 50% coil probability.

Classification Performance

Table 1: Performance comparison among five disordered Residue predictor on DD73 dataset.

	ACC	PPV	MCC	AUC
DisPredict3	**0.858**	0.818	**0.698**	**0.914**
DisPredict2	0.832	**0.857**	0.680	0.902
DisPredict	0.829	0.806	0.663	0.890
SPINE-D	0.822	0.765	0.639	0.890
MFDp	0.828	0.796	0.658	0.880

Probability Output Comparison

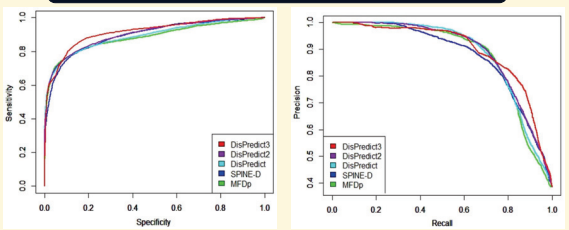


Figure 3: Comparison in terms of ROC and Precision-Recall curve .

Significance of Disorder Prediction

- Disorder predictor can identify the disorder (without amyloid formation) to order (with amyloid formation) transition of amyloidogenic regions (ARs) from AMYPdb.

Figure 4: Probability plot for β-2-micro-globulin