SCIENCE EXPLORE
An International Journal on Science and Technology
Volume I, Issue 1
October - December 2020

ASSESSMENT OF PROTEIN DISORDER REGION PREDICTION OF PONDR BASED ON CASP10 TARGETS

Subrata Sinha
Centre for Biotechnology and Bioinformatics,
Dibrugarh University, Dibrugarh-786004,
Assam

Abstract

A great challenge in structural proteomics is to predict disordered region or disordered residues, which has significant implications in experimental studies. An extended disordered regions in protein is often difficult as they can be challenging to express, purify and crystallize the protein. Commendable works on development of protein disorder prediction has taken place since last few decades. Predictor of Natural Disordered Regions (PONDR) is once such widely used reliable protein disorder predictor. PONDR has several disorder prediction algorithms (VLXT, XL1_XT, VL3, VSL2 and CAN-XT). The article presents the assessment of PONDR disorder region prediction algorithms with CASP10 targets. The evaluation was based on the six measures i.e Sensitivity, Specificity, Precision, Accuracy, Mathew Correlation Coefficient (MCC)and Area under the ROC Curve. The result shows VSL2 algorithm delivers significantly better or moderate performance than other PONDR algorithms.

Index Terms—Proteins,Prediction algorithms, Software tools, Sensitivity and specificity, Accuracy

Introduction

some proteins or particular regions of proteins lacks a well-defined tertiary structure in their native state or ordered three-dimensional structure. Such proteins are called as intrinsically disordered proteins and likewise the unstructured regions are called as intrinsically disordered protein regions[1].

A systematic analysis of intrinsic disorder in proteins started at the turn of the century and still remains a hot research topic. PubMed search with the keywords "intrinsically disordered protein" returned continuously growing number of publications from 2009 to 2019 (as of 30th December 2019). The number of experimentally verified intrinsically disordered proteins and regions are also gradually increasing.

The reason behind increase in the studies of intrinsic disorder in proteins is because Intrinsically disordered regions have been shown to be involved in a variety of functions including the following: DNA/RNA/protein recognition, Modulation of specificity/affinity of protein binding, Molecular threading, Activation by cleavage[2], apart from the involvement in various functions, intrinsic disorder in proteins has significance in

Evolutionary and Adaptation Studies [3-7], in Disease Related Studies [8-12], in Drug Discovery [13-16], in Protein Structure Determination [17,18].

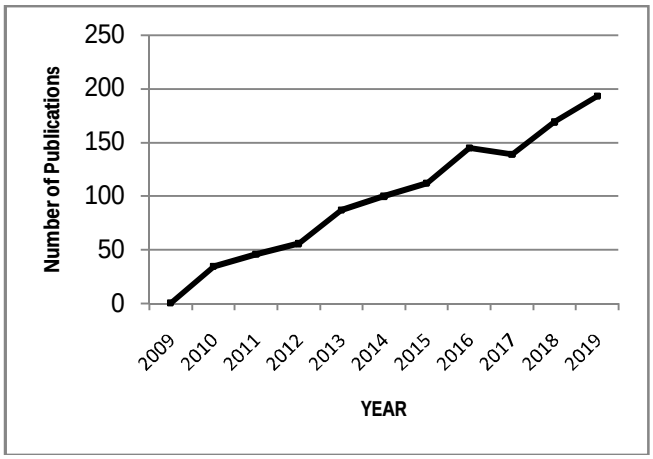

Fig. 1. Number of publications relating to "intrinsically disordered protein" on PubMed from January 2009 to December 2019.

Looking into the significance of disordered regions of proteins in various domains, various disorder region prediction methods have been developed. The first disorder predictor was published in 1997 [19], by 2009 more than 50 predictors of disorder have been developed [20] and the last review on disorder predictors shows it has reached 70 [21].
Among all the predictors, PONDR family of predictors [22-26] are found to produce consistent and reliable prediction and has evidence of supporting experimental studies [17].
We have evaluated the performance of PONDR VLXT, XL1_XT, VL3, VSL2, and CAN-XT with CASP 10 targets in predicting the disordered residues.

2 MATERIALS AND METHOD

2.1 Test Set

We have chosen the test data for evaluation of our disorder prediction model very carefully. The disorder predictors need to evaluate their accuracy based on the CASP targets released after every two years [27]. Hence the PONDR predictors has been tested with 94 targets as released in CASP 10 experiment.

2.2 Evaluation Criteria

2.2.1Binary metrics

For evaluation of disorder predictors as binary classifiers we used the (a) Sensitivity =TP/(TP+FN), (b) Specificity = TN/(TN+FP) (c) Precision= TP/(TP+FP) , (d) Balanced Accuracy (Acc) = (Sensitivity+Specificity)/2, (e) MCC = (TP.TN-FP.FN)/√((TP+FP)(TP+FN)(TN+FP)(TN+FN))

Here, TP (True Positives) is disorder residue predicted to be disordered and TN (True Negatives) is ordered residue predicted to be ordered, FP (False Positives) is ordered residue predicted to be disordered and FN (False Negatives) is disordered residue which are predicted as ordered.

2.2.2 Probability-based metrics

The accuracy of identifying disorder by assigning per-residue disorder confidence scores can be evaluated by the Receiver Operating Characteristic(ROC). A classical ROC curve

represents a monotonic function describing the balance between the true positive and false positive rates of a predictor. For a set of probability thresholds (from 0 to 1), a residue is considered as a positive example (disordered) if its predicted probability is equal to or greater than the threshold value. The area under the curve (AUC, or AUC_ROC) is used as an aggregate measure of the overall quality of a prediction method. A value of 1 corresponds to a perfect classifier, while 0.5 indicates a random prediction. Note that the ROC curve analysis works best for the probability estimates that are evenly distributed throughout the range of the allowed values. (AUC_ROC) is the only one use to describe the probability-based evaluation results .

3 RESULTS

We have analyzed 24168 residues of 94 targets released in CASP 10 experiment with PONDR predictors for Per-residue predictions and we found true positives, true negatives, false positives and false negatives for each PONDR predictors. The result is given in Table I.

TABLE I

PER-RESIDUE PREDICTIONS OF PONDR PREDICTORS ON CASP 10 TARGETS.

PREDICTORS	TP	TN	FP	FN
VLXT	614	18209	4446	899
XL1_XT	521	15527	7127	992
VL3	432	21250	1405	1081
VSL2	884	18615	4040	629
CAN-XT	301	18574	4081	1212

The binary and probability based metrics has been computed as per section B in materials and method section and the result is given in Table II.

TABLE II

COMPARISON OF PONDR DISORDER PREDICTORS BASED ON BINARY AND PROBABILITY BASED METRICS

PONDR Predictors	Sens	Spec	Prec	Acc	MCC	AUC
VLXT	0.406	0.804	0.121	0.605	0.125	0.605
XL1_XT	0.344	0.685	0.068	0.515	0.015	0.515
VL3	0.286	0.938	0.235	0.612	0.204	0.611
VSL2	0.584	0.822	0.180	0.703	0.244	0.703
CAN-XT	0.199	0.820	0.069	0.509	0.012	0.51

TABLE III
RANKING OF PONDR PREDICTORS

PONDR Predictors	Sens	Spec	Prec	Acc	MCC	AUC
VLXT	2	4	3	3	3	3
XL1_XT	3	5	5	4	4	4
VL3	4	1	1	2	2	2
VSL2	1	2	2	1	1	1
CAN-XT	5	3	4	5	5	5

For each group, Table II and III reportsthe assessm
ent scores (Sensitivity, Specificity, Precision, Accuracy, MCC and AUC_ROC) and the rank of the PONDR predictors.

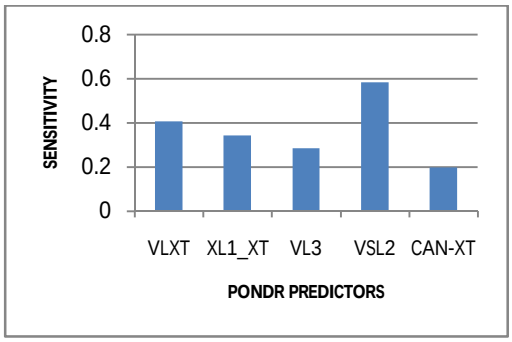

Fig.2. Sensitivity Comparison of PONDR Predictors

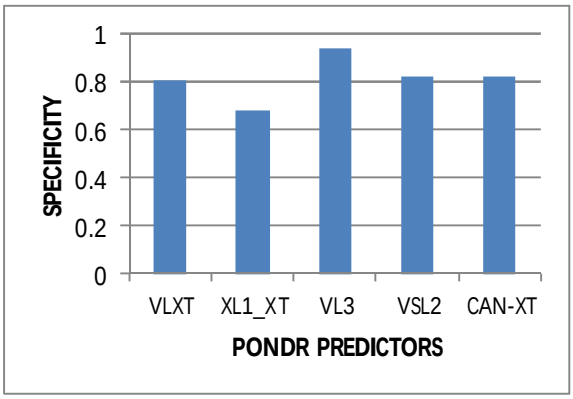

Fig. 3. Sensitivity Comparison of PONDR Predictors

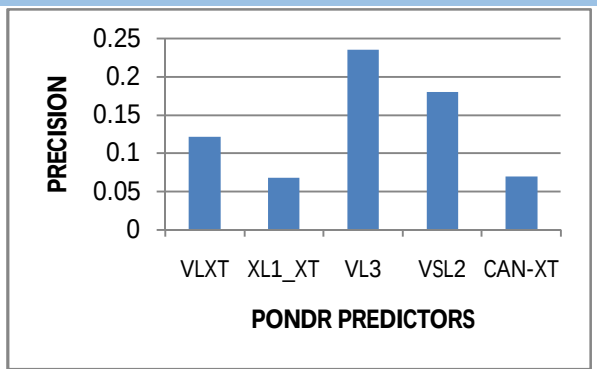

Fig. 4.Precision Comparison of PONDR Predictors

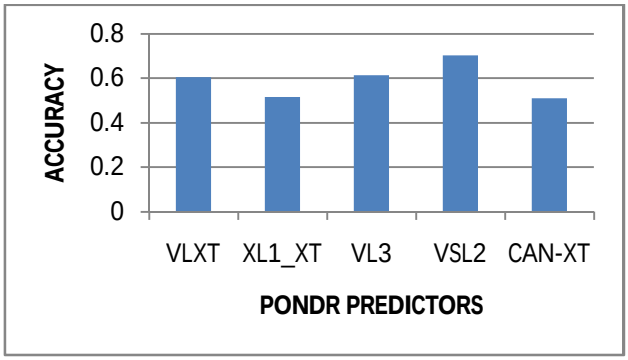

Fig. 5.Accuracy Comparison of PONDR Predictors

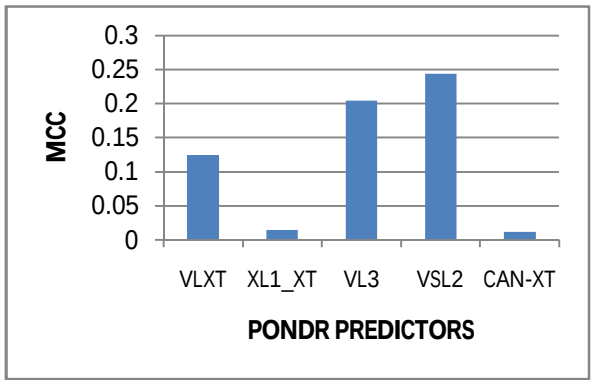

Fig. 6. MCC Comparison of PONDR Predictors

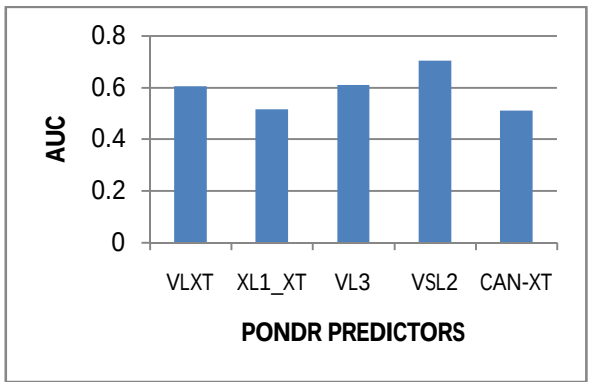

Fig. 7. AUC Comparison of PONDR Predictors

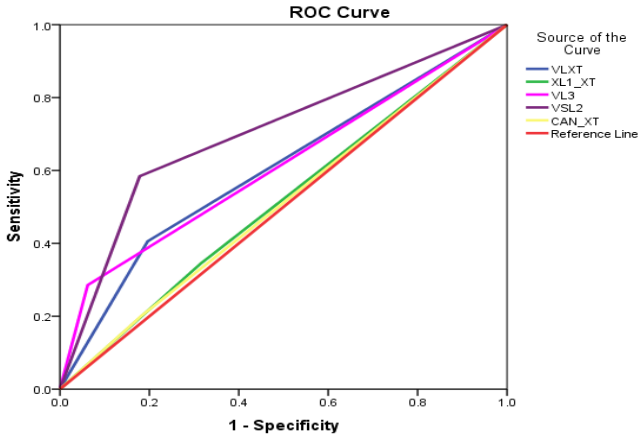

Fig. 8. Comparative ROC Curve Analysis of PONDR predictors.

Table II shows that VSL2 is the best performing algorithm with sensitivity 0.584 followed by VLXT with sensitivity 0.406, whereas VL3 found to have highest specificity i.e 0.938 followed by VSL2 i.e 0.822 and VL3 also found to have highest precision of 0.235 followed by VSL2 i.e 0.180. In terms of accuracy VSL2 once again outperformed other algoritms with accuracy 0.703 and has significantly higher MCC value i.e 0.244. The ROC Analysis shows VSL2 has highest AUC value 0.703 which assures orverall prediction capability of VSL2 in comparison to other PONDR predictors.

4 Conclusion

Significant application of disorder prediction has increased in last decade, hence the necessity to make quick and accurate predictions has also increased. There are many disorder prediction tools and PONDR is one of the widely used disorder prediction by experimental biologists. Therefore frequent accuracy assessment of PONDR is required with various data set. Our assessment of PONDR disorder prediction algorithm shows VSL2 is reliable disorder prediction algorithm among all other PONDR algorithms.

Acknowledgment

The authors wish to thank Centre for Bioinformatics Studies, Dibrugarh University for the infrastructure support.

References

[1]. A.K.Dunker*et al.* "Intrinsically disordered protein",*Journal of Molecular Graphics and Modelling*, vol. 19, pp26–59, 2001.

[2]. E.Garner*et al.*,"Predicting disordered regions from amino acid sequence: Common themes despite differing structural characterization" in *Proc* WGI, Tokyo, Japan.Dec 10-11, 1998, pp. 201-213.

[3]. A.K.Dunker e*t al.*, "Intrinsic protein disorder in complete genomes". In *Proc* WGI, Tokyo, Japan., Dec 18-19, 2000, p. 161–171.

[4]. J.J. Ward *et al.*"Prediction and functional analysis of native disorder in proteins from the three kingdoms of life". *Journal of Molecular Biology* ,vol 337, pp 635–645, 2004.

[5]. Z. Peng*et al.*, "Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life",*Cellular and Molecular Life Sciences*, vol. 72, pp.137-51, 2015.

[6]. B. Xue*et al.*, "PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino

Acids",*BiochimicaetBiophysicaActa.*vol. 1804, no. 4, pp. 996-1010, 2010.

[7]. B. Xue, A.K. Dunker, V.N. Uversky, "Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life",*Journal of Biomolecular Structure and Dynamics,* vol.30, no. 2, pp.137-49, 2012.

[8]. Y. Cheng*et al.*,"Abundance of intrinsic disorder in protein associated with cardiovascular disease", *Biochemistry*, vol. 45, no. 35, pp. 10448–10460, 2006.

[9]. V.N.Uversky, C.J. Oldfield, A.K.Dunker,"Intrinsically disordered proteins in human diseases: introducing the D2 concept",*Annual Review of Biophysics*,vol. 37, pp. 215–246, 2008.

[10]. V.N. Uversky*et al.*,"Pathological Unfoldomics of Uncontrolled Chaos: Intrinsically Disordered Proteins and Human Diseases",*Chemical Reviews*,vol. 114, no. 13, pp. 6844–6879, 2014.

[11]. J. Gsponer*et al.*, "Tight regulation of unstructured proteins: From transcript synthesis to protein degradation",*Science*,vol. 322, no. 5906, pp. 1365-1368, 2008.

[12]. V. Vacic and L.M. Iakoucheva,"Disease mutations in disordered regions—exception to the rule?",*Molecular BioSystems*, vol. 8, no. 1, pp. 27–32, 2012.

[13]. J.M. de Pereda and J.M.Andreu,"Mapping surface sequences of the tubulin dimer and taxol induced microtubules with limited proteolysis",*Biochemistry*, vol. 35, no. 45, pp. 14184-202, 1996.

[14]. J. Wang*et al.*,"Novel Strategies for Drug Discovery Based on Intrinsically Disordered Proteins (IDPs)",*International Journal of Molecular Sciences*, vol. 12, no. 5, pp. 3205–3219, 2011.

[15]. Y.Cheng *et al.*, "Rational drug design via intrinsically disordered protein.", *Trends Biotechnology*, vol. 24, no. 10, pp.435-42, 2006.

[16]. D. Marasco and P.L.Scognamiglio,"Identification of inhibitors of biological interactions involving intrinsically disordered proteins", *International Journal of Molecular Sciences*, vol. 16, pp. 7394–7412, 2015.

[17]. V. Bandaru et al. "Overproduction, crystallization and preliminary crystallographic analysis of a novel human DNA-repair enzyme that recognizes oxidative DNA damage.",*ActaCrystallographica Section D* , vol. 60, no. 6. pp.1142–1144, 2004.

[18]. J.D. Atkins*et al.*,"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,

[19]. P. Romero*et al.*, "Identifying disordered regions in proteins from amino acid sequence",in *Proc*ICNN, Houston, TX, USA, Jun 12, 1997, pp. 90–95.

[20]. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. Cell Research 2009; 19(8):929–949.

[21]. J.Li et al."An Overview of Predictors for Intrinsically Disordered Proteins over 2010–2014", *International Journal of Molecular Sciences*, vol. 16, pp. 23446-23462, 2015.

[22]. P. Romero et al., "Sequence complexity of disordered protein", *Proteins*, vol. 42, no. 1, pp. 38–48, 2001.

[23]. S.Vucetic et al., "Flavors of protein disorder",*Proteins*, vol. 52, no. 4, pp. 573–584, 2003.

[24]. Z. Obradovic et al., "Predicting intrinsic disorder from amino acid sequence",*Proteins StructFunctBioinform*, vol. 53, no. S6, pp. 566–572, 2003.

[25]. Z. Obradovic*et al.*, "Exploiting heterogeneous sequence properties improves prediction of protein disorder",*Proteins*, vol. 61, no. S7, pp. 176–182, 2005.

[26]. K.Peng*et al.*, *"Length-dependent prediction of protein intrinsic disorder"*, *BMC Bioinformatics*, vol. 7, no. 1, pp. 208, 2006.

[27]. B. Monastyrskyy et al., "Assessment of protein disorder region predictions in CASP10", *Proteins*, vol. 82, no. 2, pp. 127-137, 2014.