Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 438–448

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Design and Development of Hydrophobicity and Net charge Based Artificial Neural Network Model for IDP/IDPR Prediction

Subrata Sinha[a*], Sanchita Sharma[a], Surabhi Johari[b], Ashwani Sharma[c], Sanchaita Rajkhowa[a]

[a]*Centre for Biotechnology and Bioinformatics, Dibrugarh University, Dibrugarh -786004, Assam, India*
[b]*School of Biosciences, IMS University Courses Campus,Ghaziabad-201002, Uttarpradesh*
[c] *Insight Biosolutions, Pépinière d'entreprises Biopôle, 6 Rue Pierre Joseph Colin,35000 Rennes, France*

Abstract

Intrinsically disordered proteins (IDPs) and proteins containing intrinsically disordered protein regions (IDPRs) often affect protein structure determination. IDP/IDPRs cause failures in the protein structure determination pipeline leading to enhanced experimental cost and time. Hence the primary sequence of proteins is often analyzed using disorder prediction servers so that manipulations can be made to the protein sequence to aid its expression, purification, and crystallization. Assessment of existing IDP/IDPR prediction methods with CASP 10 targets shows the scope of improvement of prediction. In this paper, an attempt has been made to develop a hydrophobicity and net charge based artificial neural network model for protein disorder prediction. Subsequently, a prediction tool has been developed using Java using the Neural Network Schema. The prediction tool has been tested with CASP 10 targets and the developed predictor is found to outperform other predictors i.e RONN, PONDR VLXT, DisEMBL, FOLDINDEX, and GLOBPLOT in Sensitivity (0.983), Precision (0.854), Accuracy (0.937), MCC (0.840), and AUC (0.937). DisEMBL and GLOBPLOT are found to have better specificity i.e 0.909 and 0.979 respectively than the proposed predictor i.e 0.890.

*Keywords:*ANN; IDP/IDPR Prediction;Bioinformatics;Protein Disorder Prediction.

10.1016/j.procs.2023.01.026

1. Introduction

Some proteins or particular regions of proteins lack a well-defined tertiary structure in their native state or ordered three-dimensional structure. Such proteins are called intrinsically disordered proteins (IDPs) and likewise, the unstructured regions are called intrinsically disordered protein regions (IDPRs) [1]. IDP/IDPRs generally lack bulky hydrophobic amino acids; hence, they are unable to form the well-organized hydrophobic core that makes up a structured domain [2]. These disordered regions often cause difficulties for experimental studies related to structure determination [3], as these regions are inherently flexible, which can make proteins extremely difficult to crystallize, and hence X-ray diffraction analysis may be unfeasible. Experimental data such as those generated via nuclear-magnetic resonance imaging (NMR) or X-ray crystallography (if crystals can be obtained), may be hard to interpret due to random or missing coordinates obtained for the disordered regions [4, 5]. Therefore, proteins are often analyzed using protein prediction servers before experimental analyses to identify disordered regions. If such regions are predicted to exist then manipulations can be made to the protein sequence to aid its expression, purification, and crystallization [6]. Knowledge of protein disorder is not only important in the structure determination pipeline, but also it plays an essential role in various fields of bioinformatics like evolutionary studies [7], drug discovery [8-12], and disease-related studies [13-15]. As a result, protein disorder prediction tools have been widely adopted in various applications, while researchers have been continuously making efforts in improving and developing protein disorder prediction methods [16]. Given the time and cost associated with identifying protein disorder from experimental methods, several computational approaches have been developed to predict protein disorder from a protein's primary sequence [1, 6, 12, 14, and 17]. Hence there is a need to select certain physicochemical properties of protein which can be calculated from the primary sequence of protein for the prediction of protein disorder. There are two such major forces viz. hydrophobicity and net charge which determine the folding and unfolding of protein and can be calculated from the primary sequence.

Hydrophobicity is the tendency of the nonpolar molecules to cluster in a polar solvent to minimize contact with the polar solvent and such nonpolar molecules are called hydrophobic molecules. Hydrophobic residues (viz. Ile, Leu, Val, Trp, Tyr, Phe) form the core in stable proteins around which the protein sequence folds and forms the tertiary structure and this phenomenon is related to the relative insolubility of nonpolar residues in a polar solvent. The IDPs/IDPRs have comparatively low hydrophobicity than the stable, ordered proteins. The proteins or protein regions that are experimentally proved as disordered contain a fewer number of hydrophobic residues in comparison to the ordered proteins, thus they fail to benefit from the hydrophobic interactions that drive the formation of the core in stable proteins.

The net charge of a protein molecule is the arithmetic average of all charges. Proteins contain both positively charged (Arg, Lys, and His) and negatively charged amino acids (Asp and Glu), and these amino acids can either stabilize or destabilize the proteins by charge-charge interactions [21]. When the positive and negative charges on protein are equal, the net charge is zero. A high net charge (either positive or negative) can destabilize the proteins utilizing the repulsive force between the charges of equal sign [18]. IDPs/IDPRs are rich in polar and charged residues (Arg, Gln, Ser, Glu, and Lys) and thus the net charge in IDPs/IDPRs is comparatively high than the stable proteins[19]. A high net charge means that there is a huge amount of un-neutralized charge. Hence, IDPs/IDPRs are unable to form a stable tertiary structure due to the repulsive force generated by the charges of equal sign [20].

The existing disorder predictors do not show satisfactory accuracy. A few of the predictors are listed below in the table.

Table 1.Existing Protein Disorder Predictors

Predictors	Accuracy predicted by RONN	Parameter used	Based on
RONN	0.849	Bio-basis function	Neural network trained on disordered proteins
PONDR VLXT [1]	0.787	Local aa composition, flexibility, hydropathy, etc	Neural network
DisEMBL[Linding 2003b]	0.911	Coils with high temperature factors.	Neural networks trained on X-ray structure data
FoldIndex	0.776	Charge/hydrophaty analyzed locally	Uversky algorithm
GlobPlot	0.582	Secondary Structure propensity	Russell/Linding scale of disorder

Among the given predictors only FoldIndex uses hydrophobicity and net charge implementing the algorithm of Uversky and co-workers[22]. The accuracy of FoldIndex has been studied with protein sequences derived from DisProtdatabase[23] and an accuracy of 0.625 has been found. Yang, Z.R. et al., 2005 had studied the nine disorder prediction tools against 80 protein sequences derived from the Protein Data Bank and found that FoldIndex has an accuracy of 0.771[24]. Jaime Prilusky, et al.,2005, made a study on the accuracy of FoldIndex and they found it to be 0.825. So a attempt has been made to re-examine the accuracy of few predictors and propose a new predictor with better accuracy.

The methods adopted by the existing predictors has also been studied and found several experimental techniques such as X-ray crystallography, NMR spectroscopy, circular dichroism, protease digestion, and Stokes radius determination, along with several computational techniques such as artificial neural networks (ANNs), support vector machines (SVMs), logistic regression, and discriminate analysis have so far been used to detect disordered proteins. Past research has shown that ANNs using amino acid properties are an effective tool for predicting protein disorder. ANN with Backpropagation (BP) learning algorithm is widely used in solving various classifications and forecasting problems. Even though BP convergence is slow but it is guaranteed. According to Hajek M (2005), a neural network derives its computing power through, firstly, its massively parallel distributed structure and, secondly, its ability to learn and, therefore, generalize [25]. Generalization refers to the neural network producing reasonable outputs for inputs not encountered during training (learning). These two information processing capabilities make it possible for neural networks to solve complex (large-scale) problems that are currently intractable. Therefore, a vast number of Researchers design the disorder prediction tool based on this Neural Network model. RONN, PONDR and DisEMBLetc, are some of the existing disorder prediction tools which are based on ANN and show better accuracy than non-neural network models like FoldIndex and GlobPlot as shown in Table 1. In addition to that, it can be observed that FoldIndex shows better accuracy than GlobPlot. That is why an attempt has been made to develop an ANN predictor with parameters of FoldIndex (hydrophobicity, net charge) instead of GlobPlot(Secondary structure propensity).

2. Industrial Applicability

In 1978, the same year that functional disorder was indicated by X-ray crystallography, NMR revealed the highly charged, functional tail of histone H5 to be disordered. Since then, NMR 3D structural determination has led to the characterization of several proteins containing functional, yet disordered regions. In 1996 Steven W et al. worked on X-ray and NMR structure of human Bcl-xL and found no electron density was observed for residues Ser 28 to Val 80[26]. In 2004, Bandaru attempted to crystallize the NEIL-1 protein and he was failed. Then he analyzed the sequence with the help of PONDR and found that in the C-termini of the protein 106th residues are disordered. Then he made a construct excluding the c terminal 106th residues and successfully crystallize the protein [27]. Derewenda (2010) also found difficulties in experimental studies due to IDPs[3].

During the target selection process in structural genomics/biology, intrinsic protein disorder is important to consider since disordered regions at the N- and C- termini (or even within domains) often lead to difficulties in protein expression, purification, and crystallization. Failure in any single step of the structure determination leads to a great financial loss as the entire protein structure determination pipeline fails. It is therefore essential to be able to predict if the target protein is potentially disordered/unstructured.

Summarizing the above it has been found that if the protein contains a long disordered chain one cannot find its 3D structure by X-ray crystallography while almost 90% of structures in PDB are solved by this method. For such protein solution-based methods such as circular dichroism, NMR spectroscopy is applicable. So the knowledge of the presence of disordered regions in a protein can guide the experimental biologist in choosing the correct method as well as designing expressible constructs for X-ray crystallography.

Every year a huge amount of sequences are targeted to determine the structure but very few of the attempts seem to be successful.

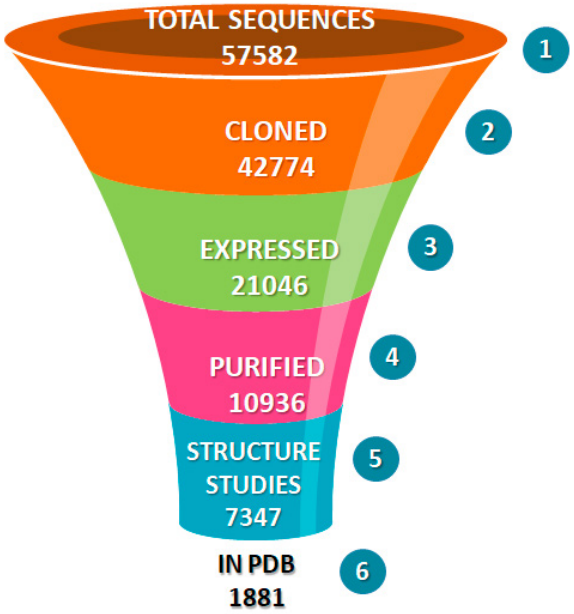

Fig. 1. A graphical representation of the biological targets being studied by the High-throughput and Membrane Protein Centers, as determined from depositions to TargetTrack

From the above data, it can be observed that from a total of 57582 sequences only 1881(i.e 3%) are successfully converted to structure format in PDB. Among the remaining 97% of failed attempts, some could be because of the presence of disordered regions in the sequences.

3. Scope

The inaccuracies of existing disorder predictors in predicting the sequence to be ordered or disordered have kept the path of development of new disorder predictors open. Inaccurate disorder prediction can lead to a wrong biological conclusion in evolutionary biology studies, and drug discovery and will misguide the protein structure determination pipeline. Till now several studies have been done to determine the sensitivity, specificity, and accuracy of the existing disorder prediction tools. Among all only one disorder predictor has been developed based on the two physiochemical properties of the protein (net charge and hydrophobicity) i.eFoldIndex using a sliding window. Our study is also based on the results of a neural network trained on a programmatically derived set of protein sequences which takes Net charge and Hydrophobicity as inputs and output is predicted either as 0(for disordered regions/parts) or as 1(for ordered regions/parts).

The results of blind-testing a panel of nine disorder prediction tools against 80 protein sequences derived from the Protein Data Bank done by RONN shows that FoldIndex has an accuracy of 0.771. Jaime Prilusky et al., 2005, made a study on the accuracy of FoldIndex and they found it to be 0.825. The accuracy of the same with protein sequences derived from the DisProt database has been studied and found to be 0.625. From the above-mentioned information, a question has been raised on the accuracy of the existing disorder prediction tool. Therefore an attempt has been made to develop an *in-house* protein disorder prediction tool.

4. Materials and Method

4.1 Materials

JDK (Java Development Kit) 1.8.0, Bio Java Library, DisProt Database, Neuroph Studio2.92.

4.2Method

4.2(a) Experimental Setup

i) Primary sequence for 800 proteins has been downloaded from the DisProt database and then the ordered and disordered regions of the sequence are fragmented by an *in-house* java program. Hydrophobicity and Net charge of each fragmented sequence have been calculated using a Bio Java program developed *in-house*.
ii) A neural network model has been designed in which hydrophobicity and net-charge were used as input neurons, the neural network has 4 neurons in the hidden layer. The neural network model has one output neuron which will have a value of 1 for ordered sequences and 0 for disordered sequences.

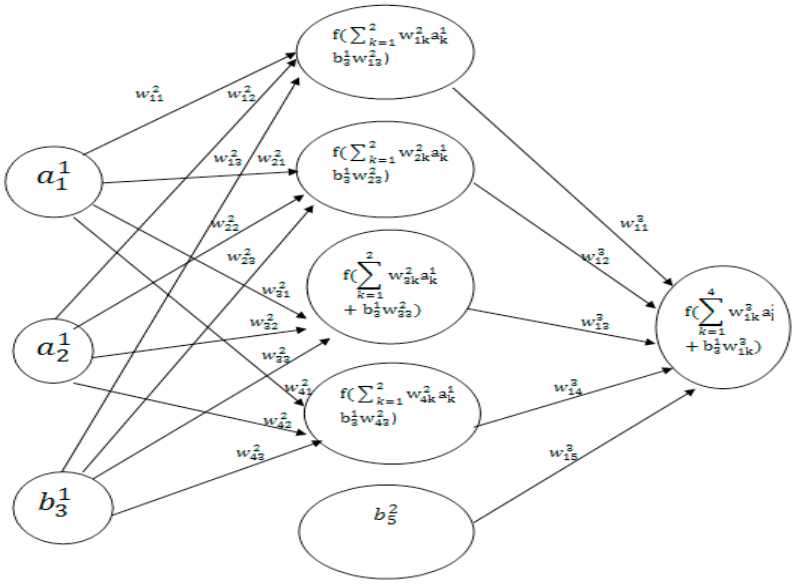

Fig. 2. A general neural network schema showing the nodes, weight values and formula to calculate final node value

iii) The neural network has been designed with above specification in Neuroph Studio 2.92 as shown in figure 3. The training of the neural network was started with learning rate 0.01, momentum 0.7, maximum error 0.01 and default iterations.

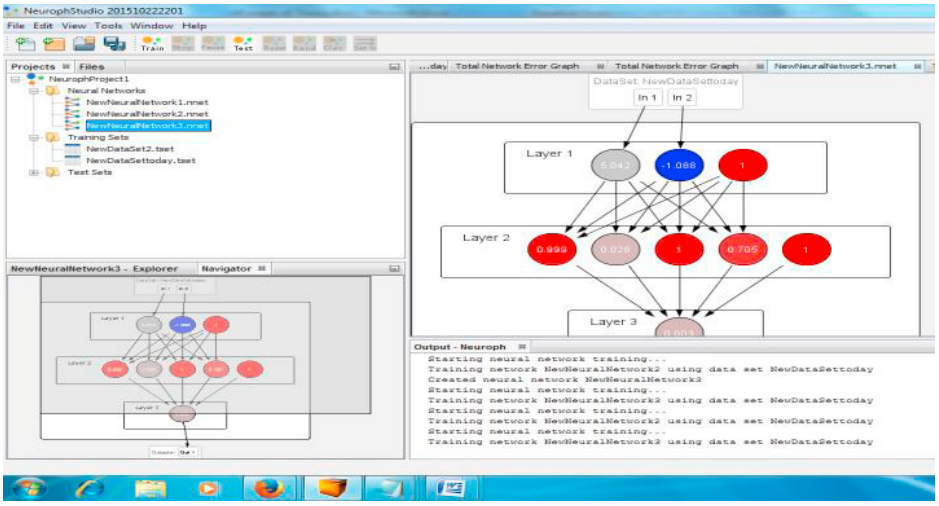

Fig. 3Neuroph studio 2.92 with neural network models and training set and test sets.

4.2(b) Evaluation

The neural network model generated from neuroph studio has been evaluated by RMSE ,MAE ,MSE and comparison of classification error with testing error which is calculated

$$ME = \sum_{i=1}^{n}(P_i - E_i) \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i-1}^{n}|P_i - E_i| \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - E_i)^2 \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - E_i)^2} \quad (4)$$

Where Pi is predicted disorder percentage and Ei is experimental disorder percentage

The predictor has been assessed with the targets released by CASP 10 experiments [28]. The true positives, false positives, true negatives, and false negatives have been calculated. The disordered regions predicted as disordered are considered as True Positive (TP), the disordered regions predicted as the ordered region are considered as False Negative (FN), the ordered regions predicted as the disordered region is considered as False Positive(FP), and the ordered regions predicted as ordered is considered as True Negative (TN). The evaluation was done using binary metrics such as Sensitivity (Sens.), Specificity (Spec.), Precision (Prec.), Mathew Correlation Coefficient (MCC), and Balanced Accuracy (Acc.) which can be calculated as follows:

$$\text{Sens.} = TP/(TP + FN) \quad (5)$$

$$\text{Spec.} = TN/(TN + FP) \quad (6)$$

$$\text{Prec.} = TP/(TP + FP) \quad (7)$$

$$\text{MCC} = ((TP * TN) - (FP * FN))/\sqrt{\{(TP + FP)(TP + FN)(TN + FP)(TN + FN)\}} \quad (8)$$

$$\text{Acc.} = 0.5 * \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right). \quad (9)$$

4.2(c) Workflow of Experiment

The overall workflow of the experiment methodology has been shown below

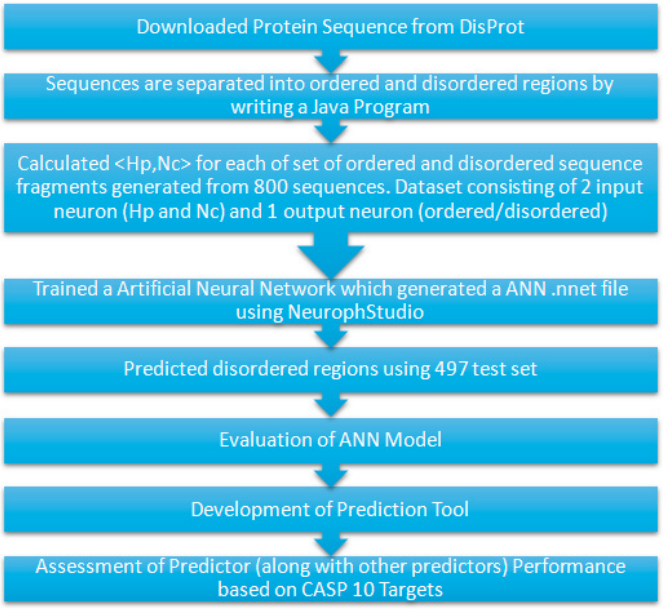

Fig. 4.Flowchart of overall Methodology

5. Results

The data generated by the java program on DisProt sequences has resulted in an ordered set of sequences and a disordered set of sequences. Net-charge and hydrophobicity have been calculated for each set of ordered and disordered sequence fragments. The entire data produced has been divided into Training Set (80%) and Test Set (20%) for 5-fold cross-validation and trained on a Neuroph IDE. The generated 5 Artificial Neural Network models (.nnet file) were used to predict whether the sequence is ordered or disordered on a 20% test set for each fold. The training error and testing errors were analyzed and the results are given below:

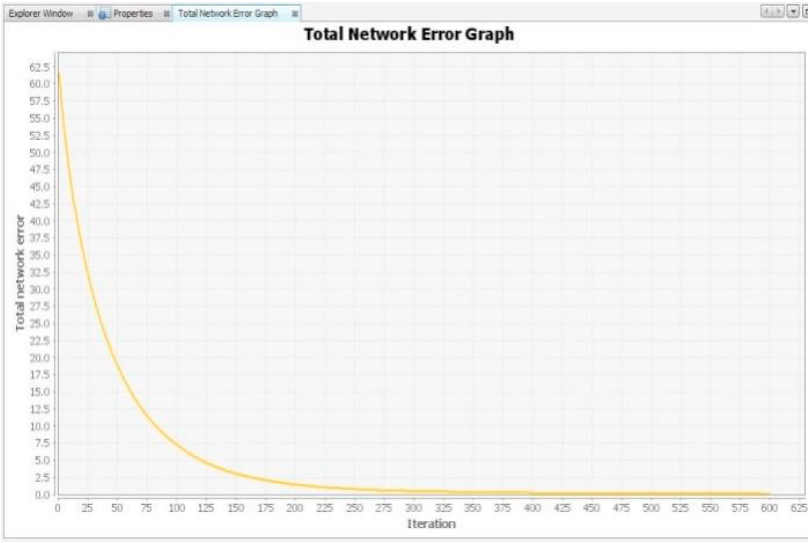

Fig. 5.Graphical Representation of Total Neural Error for 2381 sequences

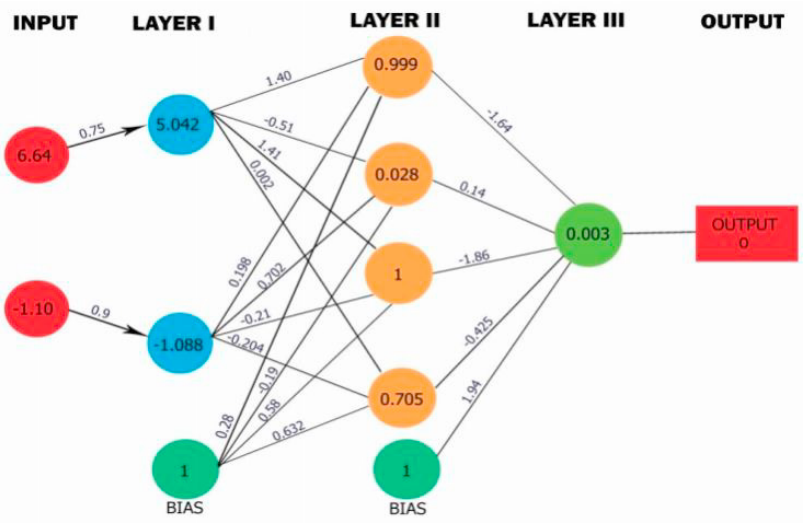

Fig. 6. Neural network schema with two nodes and one biased node in the input layer, four nodes and one biased node in the hidden layer, and one node in the output layer. The schema shown above takes one instance of net charge and hydrophobicity as input and predicts the output which is near to 1, hence the schema is for the prediction of ordered sequence/regions.

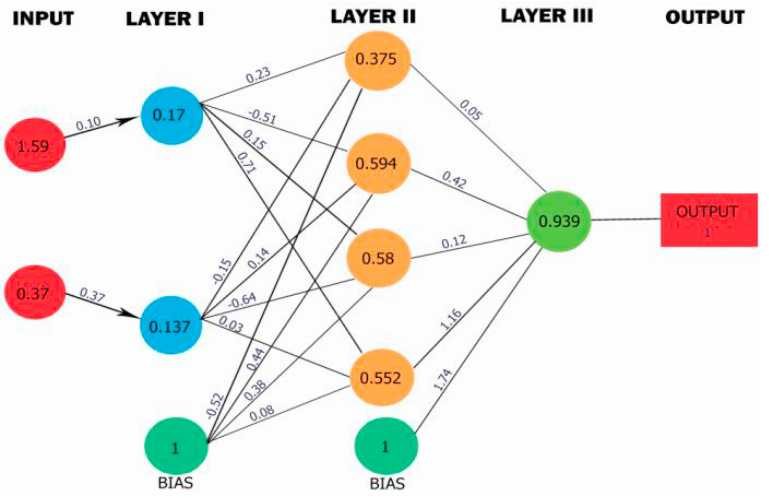

Fig. 7. Neural network schema with two nodes and one biased node in the input layer, four nodes and one biased node in the hidden layer and one node in the output layer. The schema shown above takes one instance of net charge and hydrophobicity as input and predicts the output which is near 0, hence the schema is for the prediction of disordered sequence/regions.

Table 2.RMSE and MAE of the prediction model for each Test Set.

NN Model	ME	MAE	MSE	RMSE
1	0.04	0.04	0.006	0.08
2	0.04	0.05	0.008	0.08951
3	0.05	0.05	0.008	0.09013
4	0.05	0.05	0.010	0.097663
5	0.04	0.05	0.007	0.080686

Table 3.Comparison of Training and Test MSE/RMSE

Serial no.	Training MSE	Training RMSE	Testing MSE	Testing RMSE
1	0.058	0.240	0.222	0.471
2	0.014	0.118	0.006	0.077
3	0.001	0.031	0.008	0.089
4	0.019	0.137	0.005	0.070
5	0.057	0.238	0.007	0.083

The Training error (classification error) of each training was ≈ 0.001 whereas the Testing error (Generalised error) was found to be > 0.001 hence more training of the network will be needed to make the prediction more accurate. Once the training is complete a neural network schema is generated, using which a Java-based disorder prediction tool has been developed. The Schema of the proposed predictor is given below:

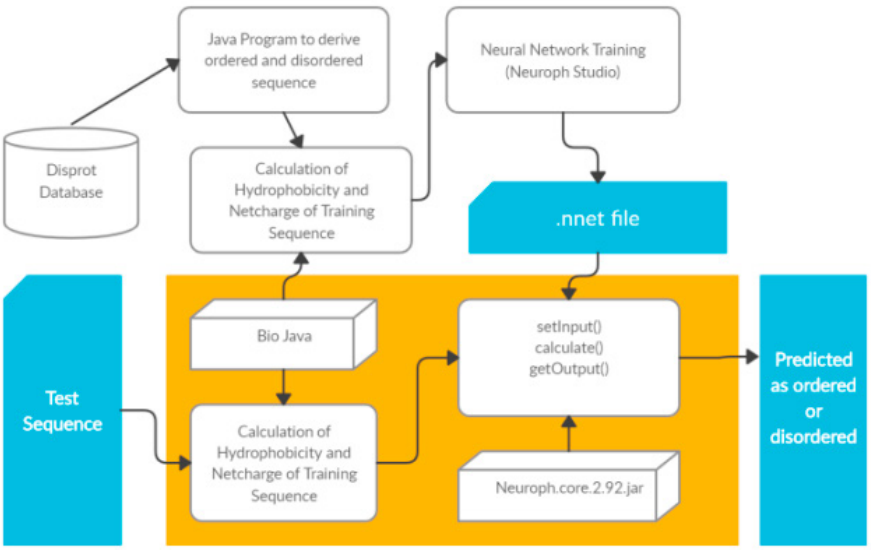

Fig. 8. Schema of the prediction tool

The tool has been tested on CASP 10 targets to calculate the sensitivity, Specificity, Precision, Accuracy, and MCC of each prediction as given in Table 4.

Table 4.Performance of Predictors

Predictor	Sens.	Spec.	Prec.	Acc.	MCC	AUC
RONN	0.486	0.163	0.037	0.324	0.232	0.736
PONDR	0.477	0.833	0.147	0.655	0.174	0.654
DisEMBL	0.568	0.909	0.258	0.738	0.321	0.738
FOLDINDEX	0.265	0.816	0.082	0.541	0.041	0.541
GLOBPLOT	0.525	0.979	0.455	0.753	0.453	0.773
DisPredHN	0.983	0.890	0.854	0.937	0.840	0.937

RONN which is based on Bio-basis neural network shows an accuracy of 0.324 while testing on 94 CASP 10 targets, whereas it shows an average AUC value, however, it shows the unacceptable precision value which was found to be 0.037 while tested on CASP 10 targets. Neural network based predictor PONDR which is based on Local amino acid composition, flexibility and hydropathy of protein sequence shows satisfactory specificity of 0.833 and average accuracy of 0.655 and AUC of 0.654 on CASP 10 targets, but again precision was found to be very low which is 0.174. Neural network based predictor DisEMBL which is based on temperature factor shows satisfactory specificity of 0.909, Accuracy of 0.738 and AUC of 0.738 whereas it shows average sensitivity of 0.568 but DisEMBL did not show good precision and MCC value which is 0.258 and 0.321 respectively. FoldIndex is the only predictor based on net charge and hydrophobicity and based on the Uversky algorithm. FoldIndex shows good specificity of 0.816 and average accuracy of 0.541 but it did not show good sensitivity, and precision while tested on CASP 10 targets. However, GLOBPLOT which is based on secondary structure propensity shows the highest specificity of 0.979 and outperformed all predictors in terms of specificity whereas it shows the average result for all binary metrics. The result shows DisPredHN outperforms other predictors in terms of AUC (0.937), MCC (0.840), Accuracy (0.937), Precision (0.854), and Sensitivity (0.983), the only exception is DisEMBL and GLOBPLOT which are found to have better specificity i.e 0.909 and 0.979 respectively than the proposed predictor i.e 0.890.

6. Conclusion

The demand for accurate prediction of natively disordered regions in proteins is being driven by structural genomics initiatives worldwide to increase the success rate, particularly for studies on eukaryotic proteins. The disorder prediction into the construct design process can reduce the amount of time and resources devoted to non-crystallizable or otherwise badly behaved proteins, allowing researchers in the wet lab to tackle the real technological challenges posed by high-throughput studies. No disorder prediction technique is so accurate that it can be entirely trusted, and a strategy of using several well-performing methods and looking for common prediction features may be the best way to attempt the reliable identification of regions of disorder in unknown protein sequences. The proposed artificial neural network-based approach, called DisPredHN combines information derived from hydrophobicity and the net charge of the sequence shows higher accuracy. DisPredHN have outperformed in terms of sensitivity, precision, accuracy, MCC and AUC with respect to other predictors. The proposed predictor outperformed other predictors, because it has considered two major factors which decides stability of protein which are hydrophobicity and net-charge there by giving near accurate predictions. Among the other predictors with which prediction result has been compared only FoldIndex considers these parameters, however FoldIndex do not employ artificial neural network model leading to poor prediction on all binary metrics. It is anticipated that DisPredHN will be of great use to experimental biologists wishing to optimize constructs for expression and crystallization or who wish to identify features within a studied protein sequence. It is further expected that progress to come from a deeper understanding of disordered proteins, which will lead to more systematic definitions of the phenomenon.

Acknowledgment

The authors wish to thank Centre for Biotechnology and Bioinformatics Studies, Dibrugarh University for the infrastructure support and Indian Science and Technology Foundation, Delhi for technical support.

References

[1] Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) "Sequence complexity of disordered protein", *Proteins*, **42(1)**: 38–48

[2] Uversky VN, Gillespie JR, Fink AL (2000) "Why Are Natively Unfolded Proteins Unstructured Under Physiologic Conditions?" *Proteins,* **41(3)**:415-427

[3] Derewenda ZS (2010) "Application of protein engineering to enhance crystallizability and improve crystal properties", *Actacrystallographica. Section D, Biological crystallography.* **66(5)**:604-15

[4] Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. (2004) Protein flexibility and intrinsic disorder. Protein Science, **13(1)**:71-80

[5] Romero P, Obradovic Z, Dunker AK (1997) "Sequence Data Analysis for Long Disordered Regions Prediction in the CalcineurinFamily",*Genome Informatics.*, **8**: 110–124

[6] Linding R, Russell RB, Neduva V, Gibson TJ (2003) "GlobPlot: Exploring protein sequences for globularity and disorder", *Nucleic Acids Research*, **31(13)**: 3701-3708

[7] Huang H, Sarai A (2012) "Analysis of the relationships between evolvability, thermodynamics, and the functions of intrinsically disordered proteins/regions", *Computational Biology and Chemistry*, **41**: 51-57

[8] Cheng Y, LeGall T, Oldfield CJ, Mueller JP, Van YY, Romero P, Cortese MS, Uversky VN, Dunker AK (2006) ,"Rational drug design via intrinsically disordered protein", *Trends Biotechnology* , **24(10)**:435-442

[9] Dunker AK, Uversky VN (2010) "Drugs for protein clouds: Targeting intrinsically disordered transcription factors", *Current Opinion in Pharmacology*, **10 (6)**: 782–788

[10] Uversky VN (2012) "Intrinsically disordered proteins and novel strategies for drug discovery". *Expert Opinion on Drug Discovery,* **7(6)**: 475-488

[11] Chen CY, Tou WI (2013) "How to design a drug for the disordered proteins?", *Drug Discovery Today,* 18(19-20): 910–915

[12] Marasco D, Scognamiglio PL (2015) "Identification of inhibitors of biological interactions involving intrinsically disordered proteins", *International Journal of Molecular Sciences*, **16 (4)**:7394–7412

[13] Uversky VN, Davé V, Iakoucheva LM, Malaney P, Metallo SJ, Pathak RR, Joerger AC (2014) "Pathological Unfoldomics of Uncontrolled Chaos: Intrinsically Disordered Proteins and Human Diseases",*Chemical Reviews,* **114(13)**:6844–6879

[14] Uversky VN, Oldfield CJ, Dunker AK (2008) "Intrinsically disordered proteins in human diseases: introducing the D2 concept", *Annual Review of Biophysics*, **37**: 215–246

[15] Cheng Y, LeGall T, Oldfield CJ, Dunker AK, Uversky VN (2006) "Abundance of intrinsic disorder in protein associated with cardiovascular disease" *Biochemistry*, **45(35)**:10448–10460

[16] Deng X, Gumm J, Karki S, Eickholt J, Cheng J (2015) "An Overview of Practical Applications of Protein Disorder Prediction and Drive for Faster, More Accurate Predictions", *International Journal of Molecular Sciences*, **16(7)**:15384-15404

[17] Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB(2003), "Protein disorder prediction: implications for structural proteomics". *Structure*, **11(11)**:1453–1459.

[18] Pace CN, Alston RW, Shaw KL (2000) "Charge–charge interactions influence the denatured state ensemble and contribute to protein stability", *Protein Science*, **9(7)**: 1395–1398

[19] Dong F, Zhou HX (2002) "Electrostatic Contributions to T4 Lysozyme Stability: Solvent-Exposed Charges versus Semi-Buried Salt Bridges",*Biophysical Journal*, **83(3)**:1341–1347

[20] Uversky VN, Dunker AK (2013) "The case for intrinsically disordered proteins playing contributory roles in molecular recognition without a stable 3D structure", *F1000 Biology Reports*, **5 (2013)**: 1. doi:10.3410/B5-1

[21] Matthew, J B, and F R Gurd (1986) "Stabilization and destabilization of protein structure by charge interactions", *Methods in Enzymology*, **130**: 437-53

[22] Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL (2005) "FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics, Sequence analysis, *Bioinformatics*, **21(16)**:3435-3438

[23] Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK (2007) "DisProt: the database of disordered proteins", *Nucleic Acids Research*, **35(Database issue)**:D786-93

[24] Yang ZR, Thomson R, McNeil P, Esnouf RM (2005) "RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins", *Bioinformatics*, **21 (16)**:3369–3376.

[25] M.Hajek (2005), "Neural Networks", University of Kwazala. Available at http://www.cs.ukzn.ac.za/notes/Neural Networks.

[26] Muchmore SW, Sattler M, Liang H, Meadows RP, Harlan JE, Yoon HS, Nettesheim D, Chang BS, Thompson CB, Wong SL, Ng SL, Fesik SW (1996) "X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death", *Nature*, **381(6580)**:335-341

[27] Bandaru V, Cooper W, Wallace SS, Doublié S (2004) "Overproduction, crystallization and preliminary crystallographic analysis of a novel human DNA-repair enzyme that recognizes oxidative DNA damage", *ActaCrystallogr D BiolCrystallogr.* **60(6)**:1142-1144.

[28] Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K (2014) "Assessment of protein disorder region predictions in CASP10", *Proteins* **82(2)**:127-137