Dispredict3.0: Prediction of Intrinsically Disordered Proteins with Protein Language Model

Md Wasi Ul Kabir, Md Tamjidul Hoque*

Department of Computer Science, University of New Orleans, New Orleans, LA, USA.

Abstract

Many biologically active proteins/protein regions fail to form a stable three-dimensional structure, yet they exhibit biological functions. These proteins are called Intrinsically disordered proteins (IDPs), and the regions are called Intrinsically disordered regions (IDRs). They play vital roles in various biological processes. These disordered regions have significant implications in properly annotating function and drug design for critical diseases. IDRs are structurally and functionally very different from ordered proteins and therefore require special experimental and computational tools for identification and analyses. Thus, the identification of IDRs is a time-consuming task. This research aims to develop a machine learning method to predict proteins' disordered regions (IDRs) accurately. We have developed a novel method named Dispredict3.0, the evolutionary information from a protein language model. Our method shows superior performance compared to the state-of-the-art method.

Keywords: disorder proteins, protein language model, machine learning.

Introduction

Intrinsic disorder proteins and protein regions do not have a well-defined secondary and tertiary structure. We know that the disorder region of the protein has existed for 20 years [1]. For example, we have abundant disorder protein sequences, eukaryotes, in which disorder proteins are found in over 30% of proteins [2]. Though disorder proteins lack well-defined structures, they play important roles in many cellular processes. They are also the main reason related to various diseases [3].

* Corresponding author: thoque@uno.edu

Several community-driven assessments were done to evaluate the disorder predictors. The first community-driven assessment was held in 2005 in Critical Assessment of protein Structure Prediction (CASP) [4] to evaluate the disordered predictors. Recently, the Critical Assessment of Protein Intrinsic Disorder (CAID), which is specially designed to assess disorder prediction, was held in 2018 [4]. In CAID assessment evaluates several recent tools, i.e., fIDPnn [2], AUCpreD [5], ESpritz [6], RawMSA [7], SPOT-Disorder2 [8], and SPOT-Disorder-Single [9]. The top-performing tool in CAID is the fIDPnn method (AUC= 0.814). The fIDPnn method extracts structural and functional information from different tools and encodes the features by aggregating the profile data at residue, window, and protein levels. They train a deep neural network for disorder prediction.

In this work, we developed a novel disorder prediction name Dispredict3.0. The improvement comes from the representation of a protein language model which is trained on large protein sequences. Recently we have seen several protein languages models in the literature, i.e., Evolutionary Scale Modeling (ESM) [10], DNABERT [11], RNABERT [12]. ESM trains a deep contextual language model on 86 billion amino acids across 250 million protein sequences spanning evolutionary diversity, and the trained model contains information about biological properties in its representations. The representations are learned from sequence data alone. The authors show that the learned representation space has information on the structure of proteins from the level of biochemical properties of amino acids to the remote homology of proteins [10]. It also has information about the secondary and tertiary structures [10].

The main contribution of this paper is to show that the representation from the protein’s language model has proven to improve protein disorder prediction. We have used an optimized Light Gradient Boosting Machine to train the machine learning model. We have experimentally shown that the Dispredict3.0 outperforms the state-of-the-art method.

Material and Method

Dataset

We have collected the same dataset used in the fIDPnn [2] method to have a fair comparison. The authors [2] of the fIDPnn method curated the 745 proteins dataset from the Disprot 7.0 database [13]. The Disprot database is the gold standard for disorder regions annotation. The disport database is consistently updated, and the recent release is from December 2021. The dataset has three-part- the train, test, and validation sets. The training, test, and validation set contain 445, 176, and 100 proteins. The training and test set share less than 25% similarities between them. The dataset is highly imbalanced in the number of ordered and disordered residues, as shown in Table 1.

Table 1. Statistics of ordered and disordered residues in the training, test, and validation dataset.

Disordered/Ordered	Train	Test
No. of Disordered residues	50387	17871
No. of Ordered residues	169565	48675
Total No. of Residues	219952	66546

Framework of Dispredict3.0

Dispredict3.0 extracts residue-level, window-level, and protein-level information of protein sequence using the fIDPnn tool. The fIDPnn method extracts these features from other methods, i.e., PSIPRED [14], IUPred [15], PSI-BLAST [16]. We have used all the features from fIDPnn as the method performs very well in the last CAID [4] competition. Moreover, we extract evolutionary information from the Transformer based protein language model, named Evolutionary Scale Modeling (ESM), from Facebook AI Research [10]. ESM trains on large protein sequences and can represent residues. To reduce the dimensionality of ESM features, we have used Principal Component Analysis (PCA) [17]. Table 2 shows the number of features extracted from different tools. After collecting all the features from four different tools, we train an optimized Light Gradient Boosting Machine algorithm [18]. The framework of Dispredict3.0 is shown in figure 1.

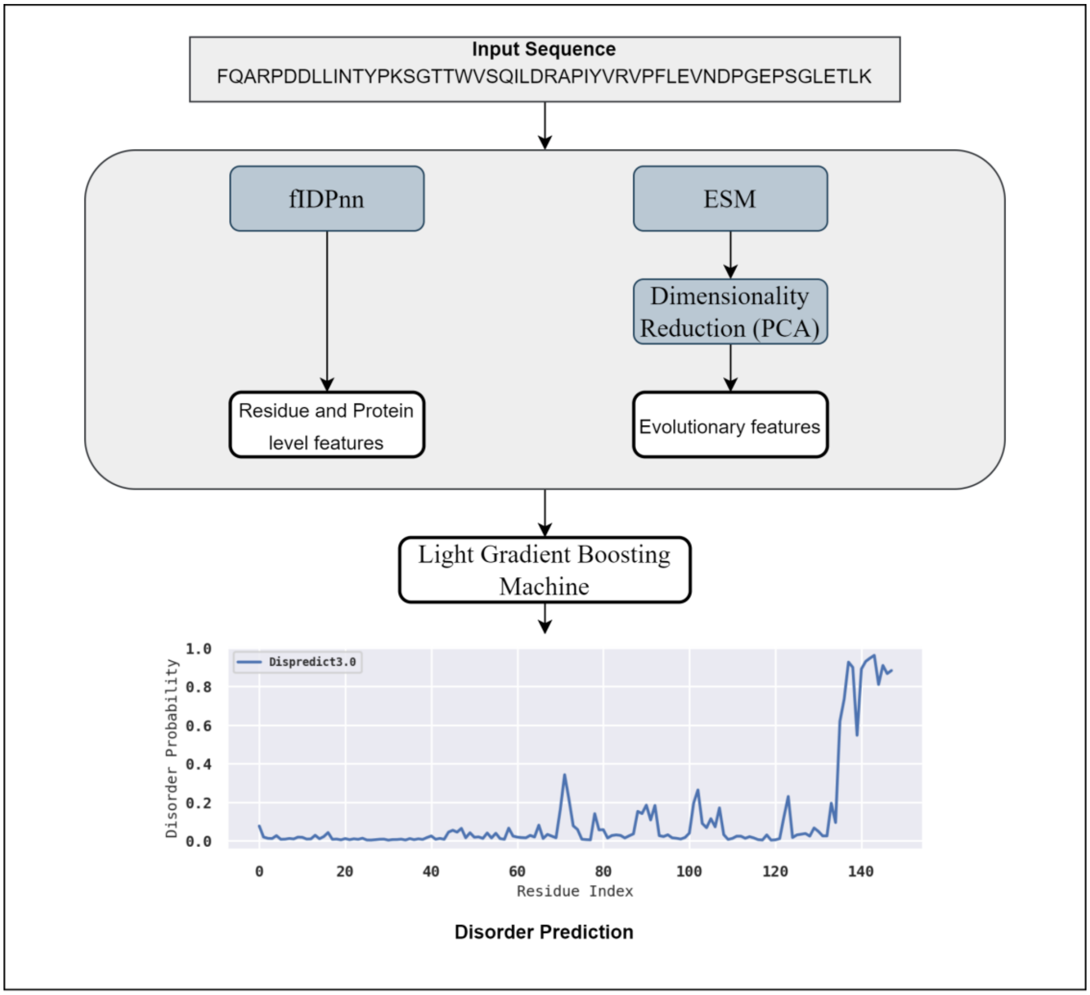
Input Sequence
FQARPDDLLINTYPKSGTTWVSQILDRAPIYVRVPFLEVNDPGEPSGLETLK
fIDPnn
ESM
Dimensionality Reduction (PCA)
Residue and Protein level features
Evolutionary features
Light Gradient Boosting Machine
Dispredict3.0
Disorder Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
Residue Index
Disorder Prediction

Figure 1. The framework of the Dispredict3.0 method for disordered prediction. Dispredict3.0 collects features from fIDPnn, and ESM and trains a Light Gradient Boosting Machine for disorder prediction.

Table 2. The number of features extracted from different tools.

Methods	No. of Features
fIDPnn	317
ESM	3843
Total Features	4160

Performance Evaluation Metrics

To evaluate the performance of Dispredict3.0, we have selected widely used metrics for the imbalance dataset, i.e., the Area under Receiver operating characteristic curve (AUC), F1-score, Mathews Correlation Coefficient (MCC), and Kappa score. ROC Curve represents the performance of a classifier at various threshold settings. AUC metric is threshold dependent, whereas F1-score, MCC, and Kappa metrics are threshold independent. Table 3 shows the definition of the selected metrics.

Table 3. Performance Metrics

Name of Metric	Definition
True Positive (TP)	Correctly predicted positive samples
True Negative (TN)	Correctly predicted negative samples
False Positive (FP)	Incorrectly predicted positive samples
False Negative (FN)	Incorrectly predicted negative samples
F1-score (Harmonic mean of precision and recall)	$\frac{2TP}{2TP+FP+FN}$
Mathews Correlation Coefficient (MCC)	$\frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FP) \times (TN+FN}}$
Kappa	$\frac{2 \times (TP \times TN - FP \times FN)}{(TP+FN) \times (TP+FP) \times (TN+FP) \times (TN+FN)}$

Results

Prediction of intrinsic disorder

To find the best method for disorder prediction, we experiment with the eight machine learning methods. Table 4 shows the cross-validation results on the training dataset in terms of the metrics AUC, F1-score, kappa, and MCC. We found that, on average, Light Gradient Boosting Machine performs better.

Table 4. Comparison of individual machine learning methods' prediction results in 10-fold cross-validation on the training dataset.

Model	AUC	F1-score	Kappa	MCC
Light Gradient Boosting Machine	0.982	0.901	0.864	0.867
Decision Tree Classifier	0.933	0.905	0.865	0.865
Extra Trees Classifier	0.967	0.840	0.785	0.798
Gradient Boosting Classifier	0.936	0.798	0.727	0.736

Linear Discriminant Analysis	0.917	0.760	0.668	0.670
Logistic Regression	0.917	0.756	0.660	0.661
Ada Boost Classifier	0.900	0.721	0.619	0.624
K Neighbors Classifier	0.801	0.619	0.455	0.455

Ablation analysis

We have also experimented with the effect of each feature set in terms of prediction. Table 5 shows the test set results with an incremental increase of the features set. We found that adding representation from different layers of ESM helps to have a better prediction. If we collect representation from all 34 layers, each residue is represented by a large matrix of size (34 × 1281). So, we have implemented an incremental PCA to reduce the dimension. We have selected the number of components as 3184 to retain 93.02% variance of the original data. We found that the reduced dimension from PCA helps achieve better AUC and F1-score. Our training dataset is highly imbalanced (Table 1), and F1-score, kappa, and MCC are threshold-dependent metrics. As discussed in the next section, these motivate us to find the optimum threshold value.

Table 5. Analysis of the effect of different feature sets.

Name	**AUC**	**F1-score**	**Kappa**	**MCC**	**Imp-AUC**	**Imp-F1-score**	**Imp-Kappa**	**Imp-MCC**	**Imp-Average**
Base Scores (fIDPnn)	0.837	0.558	0.445	0.469					
fIDPnn	0.834	0.575	0.447	0.455	-0.32	3.21	0.45	-2.94	0.10
fIDPnn + Layer 0	0.831	0.580	0.454	0.461	-0.69	4.08	1.89	-1.53	0.94
fIDPnn + Layer 0 and 33	0.851	0.610	0.487	0.491	1.67	9.38	9.22	4.86	6.28
fIDPnn + Layer 33	0.852	0.615	0.495	0.501	1.81	10.33	11.15	6.90	7.55
fIDPnn + Layer 15	0.851	0.620	0.502	0.508	1.65	11.29	12.74	8.44	8.53
fIDPnn + Layer 0 and 15	0.849	0.624	0.506	0.512	1.47	11.88	13.60	9.20	9.04
fIDPnn + Layer 15 and 33	0.854	0.632	0.517	0.523	2.11	13.32	16.00	11.52	10.74
fIDPnn + Layer 0, 15 and 33	0.857	0.631	**0.518**	**0.525**	2.37	13.25	**16.25**	**11.97**	**10.96**
fIDPnn + PCA	**0.858**	**0.633**	0.517	0.522	**2.56**	**13.62**	16.12	11.45	10.94

Best score values are **bold-faced.** Here, ‘Imp’ stands for improvement. The ‘Imp’ indicates the improvement in percentage achieved by Dispredict3.0 over the Base Scores (fIDPnn) for the corresponding evaluation metric.

Threshold Optimization

Threshold Optimization is considered an important part of machine learning, especially when the training dataset is highly imbalanced. The trained model becomes biased to one class for the imbalanced dataset. We optimize the threshold based on the ROC AUC curve to overcome the problem. Threshold optimization can be done based on the training or the validation set, or both. We found the best threshold=0.353 when we take an average of the optimum threshold from both the training and validation dataset, as shown in Table 6.

Table 6. Threshold optimization based on the training and validation set.

Name	Threshold	AUC	F1-score	Kappa	MCC	Imp-AUC	Imp-F1-score	Imp-Kappa	Imp-MCC	Imp-Average
Base Scores (fIDPnn)	0.500	0.837	0.558	0.445	0.469	-	-	-	-	-
Default threshold	0.500	0.858	0.631	0.516	0.522	2.49	13.09	15.73	11.33	10.66
Training Set	0.583	0.858	0.616	0.504	0.516	2.49	10.46	13.24	10.11	9.08
Validation Set	0.123	0.858	**0.655**	0.505	0.514	2.49	**17.47**	13.44	9.66	10.76
Training + Validation Set	0.353	0.858	0.648	**0.525**	**0.525**	2.49	16.14	**17.76**	**12.09**	**12.12**

Here, 'Imp' stands for improvement. The 'Imp' indicates the improvement in percentage achieved by Dispredict3.0 over the fIDPnn method for the corresponding evaluation metric.

Comparison with existing methods

We compare Dispredict3.0 with six recently published disorder predictors. All results are generated by running the tool on our own server. Dispredict3.0 outperforms all of them. Compared to the state-of-the-art methods, Dispredict3.0 has an improvement of 2.54%, 16.22%, 12.05%, and 17.85% in terms of AUC, F1-score, MCC, and kappa, respectively. Table 7 and Figure 2 show that Dispredict3.0 outperforms all six predictors.

Table 7. Comparison of Dispredict3.0 with the six disordered predictors.

Method	AUC	F1-score	MCC	Kappa
NetSurfP - 2.0	0.753	0.536	0.350	0.348
IUPred3	0.762	0.584	0.412	0.407
Metapredict	0.766	0.595	0.423	0.411
Espritz	0.795	0.544	0.401	0.398
SPOT-Disorder-Single	0.802	0.596	0.448	0.448
fIDPnn	0.837	0.558	0.469	0.445
DisPredict3.0	**0.858**	**0.648**	**0.525**	**0.525**
Imp (%)	2.54%	16.22%	12.05%	17.85%

Here, ‘Imp’ stands for improvement. The ‘Imp’ indicates the improvement in percentage achieved by Dispredict3.0 over the fIDPnn method for the corresponding evaluation metric.

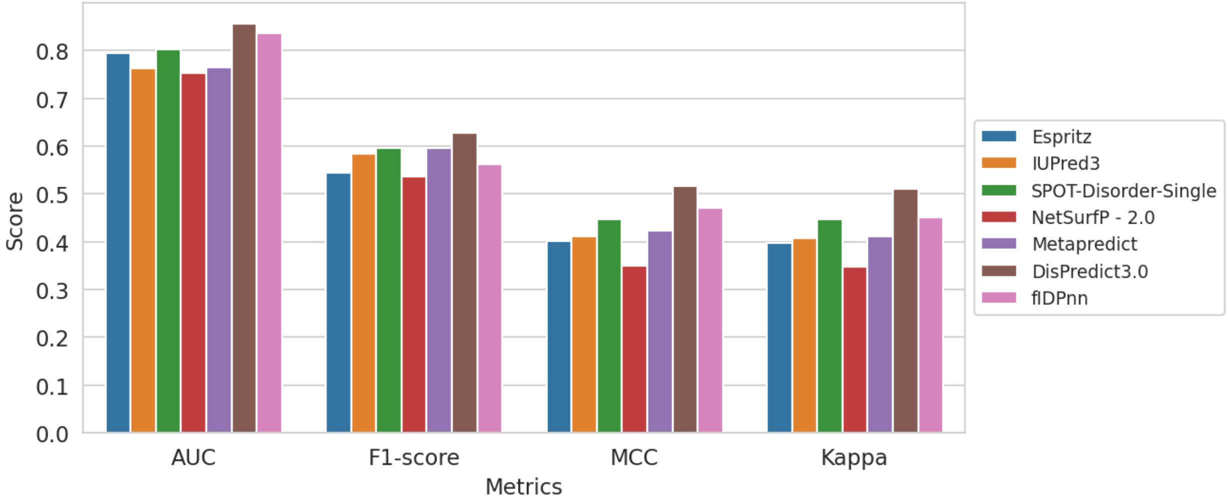

Figure 2. Comparison of Dispredict3.0 with the six disordered predictors.

We have also plotted the ROC curve and precision-recall curve compared to the existing methods, as shown in Figures 3 and 4. The curve clearly Dispredict3.0 performs better than the other methods in terms of ROC curve and precision-recall curve.

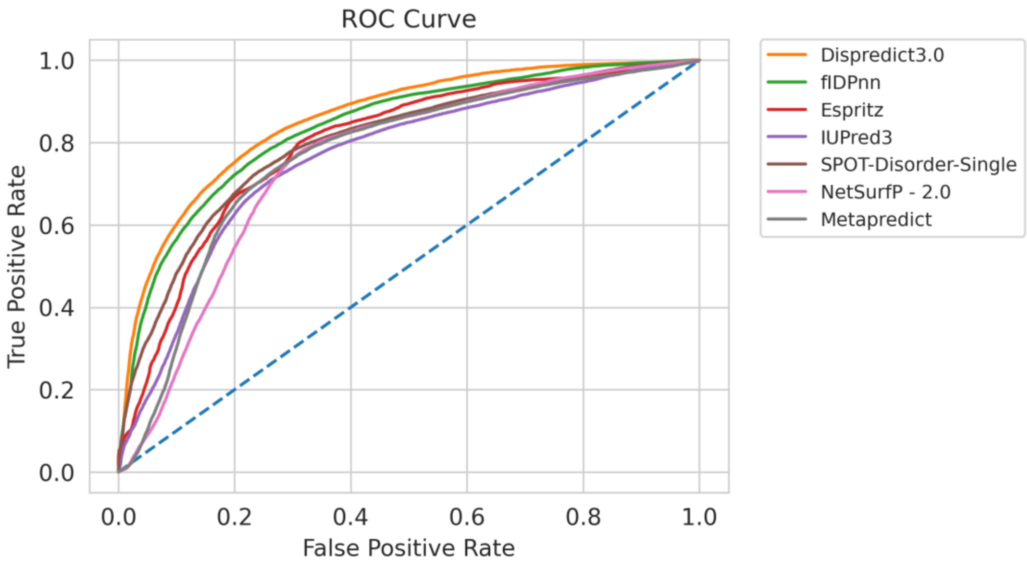

Figure 3. The ROC curve for different methods on the test dataset.

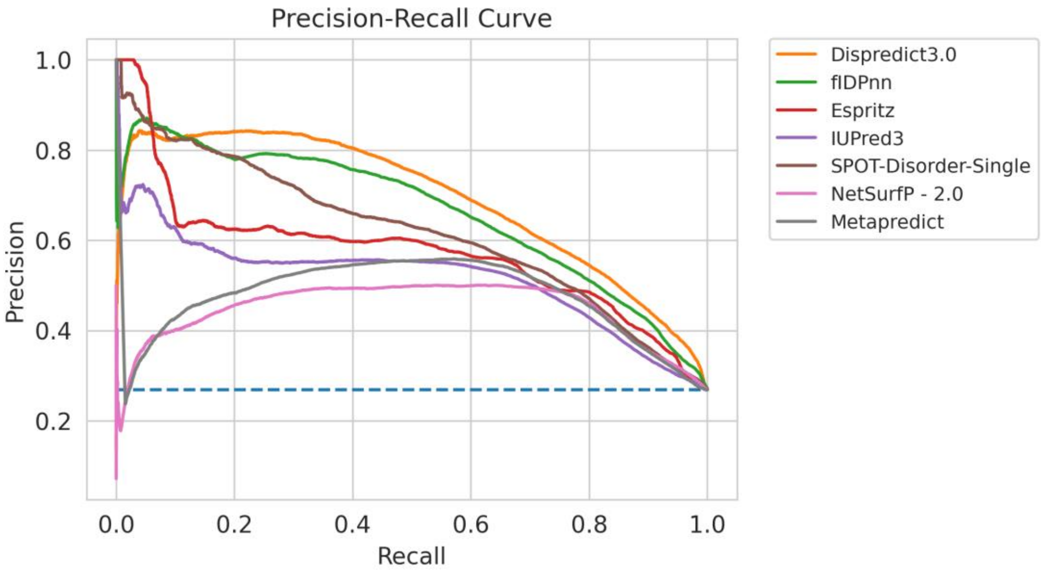

Figure 4. The precision-Recall curve for different methods on the test dataset.

Prediction of fully disordered proteins

We further investigate the Dispredict3.0 prediction capabilities for the fully disordered proteins. Figure 5 shows the performance of Dispredict3.0 compared to the existing methods. The results of Dispredict3.0 are shown with an optimized threshold. The SPOT-Disorder-Single and IUPred3 perform better in AUC and MCC, but Dispredict3.0 shows good results in terms of F1-score and Kappa.

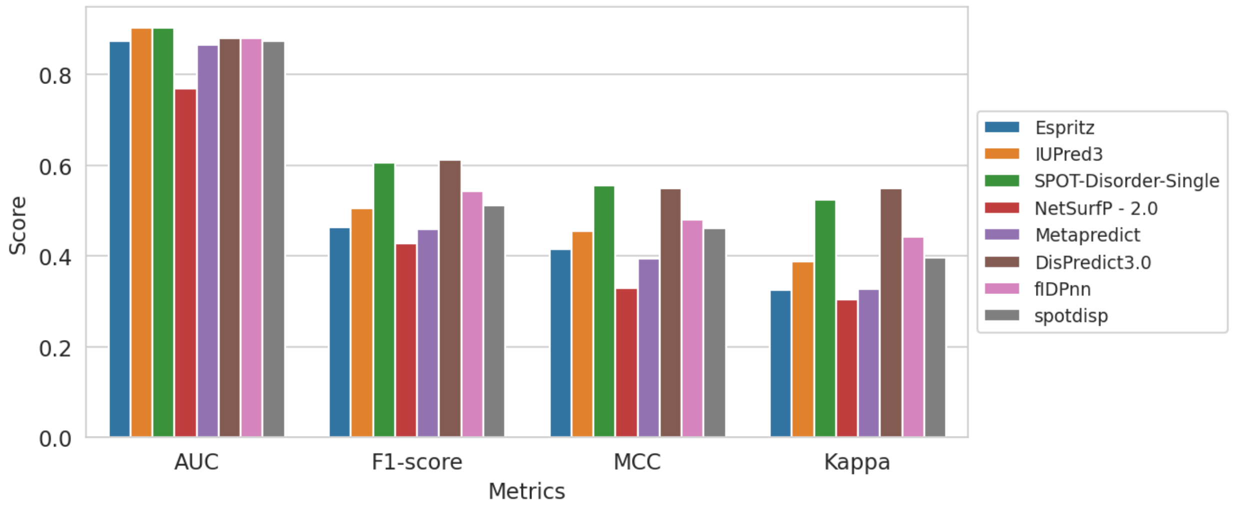

Figure 5. Performance comparison of Dispredict3.0 on fully disordered proteins.

Conclusions

Here we presented a novel disorder predictor, Dispredict3.0, that uses the representation from a protein langue model that helps to improve the performance of disorder prediction. Further, we have used PCA to reduce the dimensions of the representation and train an optimized Light Gradient Boosting Machine for disorder prediction. The experimental results show that Dispredict3.0 outperforms the existing methods for disorder prediction. Though the Dispredict3.0 method depends on a large protein language model, the method is not computationally expensive. We plan to do a computation complexity analysis with other methods in the future.

Furthermore, we will do a t-test analysis to show the statistical significance of Dispredict3.0 compared to other methods. We are also developing a new test set from the latest Disprot release to show that the performance of Dispredict3.0 is robust. We believe this tool will be helpful to the researcher to find disorder regions.

Data Availability

The code and data related to the development of Dispredict3.0 can be found here https://github.com/wasicse/Dispredict3.0

Author Contributions

Data collection and processing M.K., Conceived and designed the experiments: M.K., and M.T.H. Performed the experiments: M.K., Analysed the data: M.K., and M.T.H. Contributed reagents/materials/analysis tools: M.K., Wrote the paper: M.K., and M.T.H. All authors have read and agreed on the publication of the final version of the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Quaglia F, Mészáros B, Salladini E, Hatos A, Pancsa R, Chemes LB, Pajkos M, Lazar T, Peña-Díaz S, Santos J *et al*: **DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation**. *Nucleic Acids Research* 2022, **50**(D1):D480-D487.
2. Hu G, Katuwawala A, Wang K, Wu Z, Ghadermarzi S, Gao J, Kurgan L: **flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions**. *Nat Commun* 2021, **12**(1):4438.
3. Erdős G, Pajkos M, Dosztányi Z: **IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation**. *Nucleic Acids Research* 2021, **49**(W1):W297-W303.
4. Predictors C, DisProt C, Necci M, Piovesan D, Tosatto SCE: **Critical assessment of protein intrinsic disorder prediction**. *Nat Methods* 2021, **18**(5):472-481.
5. Wang S, Ma J, Xu J: **AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields**. *Bioinformatics* 2016, **32**(17):i672-i679.
6. Walsh I, Martin AJM, Di Domenico T, Tosatto SCE: **ESpritz: accurate and fast prediction of protein disorder**. *Bioinformatics* 2012, **28**(4):503-509.
7. Mirabello C, Wallner B: **rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments**. *PLoS ONE* 2019, **14**(8):e0220182.
8. Hanson J, Paliwal KK, Litfin T, Zhou Y: **SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning**. *Genomics, Proteomics & Bioinformatics* 2019, **17**(6):645-656.
9. Hanson J, Paliwal K, Zhou Y: **Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures**. *J Chem Inf Model* 2018, **58**(11):2369-2376.

10. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J *et al*: **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences**. *Proc Natl Acad Sci USA* 2021, **118**(15):e2016239118.
11. Ji Y, Zhou Z, Liu H, Davuluri RV: **DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome**.9.
12. Yamada K, Hamada M: **Prediction of RNA-protein interactions using a nucleotide language model**. In.: Bioinformatics; 2021.
13. Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidović R, Dosztányi Z *et al*: **DisProt 7.0: a major update of the database of disordered proteins**. *Nucleic Acids Research* 2017, **45**(D1):D219-D227.
14. Buchan DWA, Jones DT: **The PSIPRED Protein Analysis Workbench: 20 years on**. *Nucleic Acids Research* 2019, **47**(W1):W402-W407.
15. Mészáros B, Erdős G, Dosztányi Z: **IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding**. *Nucleic Acids Research* 2018, **46**(W1):W329-W337.
16. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Research* 1997, **25**(17):3389-3402.
17. Jolliffe IT, Cadima J: **Principal component analysis: a review and recent developments**. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 2016, **374**(2065):20150202.
18. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y: **LightGBM: a highly efficient gradient boosting decision tree**. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA*. Curran Associates Inc. 2017: 3149–3157.