Prediction and Characterization of Disorder-Order Transition Regions in Proteins by Deep Learning

Ziang Yan[1], Satoshi Omori[1], Kazunori D Yamada[1], Hafumi Nishi[1] and Kengo Kinoshita[1,2,3] *

[1]Graduate School of Information Sciences, Tohoku University, Sendai, 980-8579, Japan

[2]Tohoku Medical Megabank Organization, Tohoku University, Sendai, 980-8573 Japan.

[3]Institute of Development, Aging, and Cancer, Tohoku University, Sendai, 980-8575 Japan.

*To whom correspondence should be addressed.

Abstract

The biological functions of proteins are traditionally thought to depend on well-defined three-dimensional structures, but many experimental studies have shown that disordered regions lacking fixed three-dimensional structures also have crucial biological roles. In some of these regions, disorder–order transitions are also involved in various biological processes, such as protein-protein interaction and ligand binding. Therefore, it is crucial to study disordered regions and structural transitions for further understanding of protein functions and folding. Owing to the costs and time requirements of experimental identification of natively disordered or transitional regions, the development of effective computational methods is a key research goal. In this study, we used overall residue dependencies and deep representation learning for prediction and reused the obtained disordered regions for the prediction of disorder–order transitions. Two similar and related prediction tasks were combined. Firstly, we developed a novel deep learning method, Res-BiLstm, for residue-wise disordered region prediction. Our method outperformed other predictors with respect to almost all criteria, as evaluated using an independent test set. For disorder-order transition prediction, we proposed a transfer learning method, Res-BiLstm-NN, with an acceptable but unbalanced performance, yielding reasonable results. To grasp underlining biophysical principles of disorder-order transitions, we performed qualitative analyses on the obtained results and discovered that most transitions have strong disordered or ordered preferences, and more transitions are consistent with the ordered state than the disordered state, different from conventional wisdom. To the best of our knowledge, this is the first sizable-scale study of transition prediction.

Availability: https://github.com/Yanzziang/Transition_Disorder_Prediction

Contact: kengo@ecei.tohoku.ac.jp

Introduction

Intrinsically disordered proteins or regions do not form stable structures under physiological conditions and yet have crucial roles in biological processes, including the recognition of proteins, nucleic acids, and other types of molecules, post-translational modifications, and alternative splicing (Dyson and Wright, 2005; Nishi, et al., 2013; Tompa, 2005; Uversky and Dunker, 2010; Uversky, et al., 2008; Wright and Dyson, 1999). Some of these disordered proteins or regions undergo a transition from the disordered to ordered state (or vice versa) upon interactions with other molecules. In particular, segments that transition from disorder to order via protein–protein interactions are recognized as molecular recognition features (MoRFs) and have been explored extensively (Kotta-Loizou, et al., 2013; Mishra, et al., 2018; Mohan, et al., 2006; van der Lee, et al., 2014; Yan, et al., 2016).

In response to the increasing importance of intrinsically disordered regions and disorder–order transitions, numerous disordered region predictors have been developed in the last two decades (Liu, et al., 2019). These predictors are roughly divided into physicochemical property-based approaches (such as IUPred (Dosztanyi, et al., 2005) and FoldIndex (Prilusky, et al., 2005)) and machine learning-based approaches (such as DisEMBL (Linding, et al., 2003), SPINE-D (Zhang, et al., 2012), SPOT-Disorder (Hanson, et al., 2017),and SPOT-Disorder-single (Hanson, et al., 2018)). These software and servers, particularly with machine-learning based approaches, have achieved fairly good accuracy, which allows us to explore potentially disordered regions without waiting for experimental verification. Similarly, predictors of functional sites in disordered regions have also been developed and the majority of these employ protein-interacting MoRFs as training datasets (Disfani, et al., 2012; Hanson, et al., 2019).

Even though it is widely accepted that disorder-order transitions can occur not only by protein interactions but also by small ligand binding, allosteric effects, and other environmental changes, not many prediction studies have been performed on non-MoRF transitions. Furthermore, disordered protein regions have recently been recognized as druggable regions (Ruan, et al., 2019), implying that drugs should be designed to cause disorder–order transitions to work effectively. To fully understand or control the functions and dynamic behaviors of protein disordered regions, a more general understanding of disorder–order transitions and transition regions is essential.

In this study, we developed a disorder–order predictor, Res-BiLstm, by employing state-of-the-art machine learning techniques, such as the bidirectional long short-term memory (LSTM) with residual connections. LSTM, a variant of recurrent neural networks (RNNs), can deal with contextual information for input data and thus are widely utilized for biological sequence analyses (Jiang, et al., 2018; Liu and Gong, 2019; Yamada and Kinoshita, 2018). Our predictor exhibited equivalent or superior performance to those of existing methods. We then expanded our method to predict transition sites recorded in the Protein Data Bank (PDB). This method, Res-BiLstm-NN, inspired by the effectiveness of transfer learning in MoRFs, enabled the application of disorder information to the prediction of transitions and exhibited acceptable but unbalanced performance. An exploratory analysis using our predictors revealed that transition sites have unique features compared to those of non-transition (order or disorder) regions, different from the conventional view. To the best of our knowledge, this is the first attempt to capture the general characteristics of transition sites without assumptions regarding molecular interactions.

Methods

Datasets

Disordered protein dataset

A dataset consisting of 4229 disordered proteins from SPINE-D (Zhang, et al., 2012) was used. A total of 3000 proteins were selected for training, 400 proteins were selected for validation in the learning process, and the remaining 829 proteins were used as an independent test set. The disordered protein dataset was extremely unbalanced; disordered regions only occupied approximately 10% of all disordered proteins. Table S1 enumerates the unbalanced disordered and ordered residues in training, validation, and test datasets.

Transition site dataset

To build the transition site dataset, all human enzymes stored in the Universal Protein Resource (UniProt) (UniProt Consortium, 2018) were retrieved. All identified protein structures of human enzymes (Du, et al., 2014) were extracted from PDB (Berman, et al., 2002), and the ss_dis.txt file (Berman, et al., 2002) was used to obtain notations about the order/disorder states. Then, summary statistics about structural information at the residue level was obtained and residues were classified as transition, non-transitional ordered, or disordered sites. First, the residues of all extracted X-ray structures from PDB were aligned to the corresponding UniProt amino acid sequences with reference to the SIFTS file (Dana, et al., 2019; Velankar, et al., 2013). Sequence segments whose structural information were not recorded in any structure files were removed. Order/disorder labels indicated that the residues were always observed as ordered/disordered among protein structures for the protein. The transition label indicated that the residues were identified as both ordered and disordered at least once among available protein structures. Proteins without transition labels were excluded, resulting in 2995 proteins in the final dataset. Figure S1 summarizes the process of building the dataset. The human protein dataset was split into 2200 proteins for training, 300 proteins for validation, and 495 proteins for the independent test set. The dataset was unbalanced. Residues labeled as transition, non-transitional order, and disorder occupied 75,793, 828,617, and 59,560 residues in the human protein dataset respectively (Table S2), where the residues located within three residues of the N-terminus or C-terminus were excluded from statistical analyses.

Model features

There are 20 kinds of amino acids constituting variable-length protein sequences. To obtain the amino acid characters, similar to SPINE-D, evolutionary information, physicochemical properties, and predicted structural information were used as features for each amino acid within a protein, instead of sequence-based information. The 20 evolutionary features were extracted from a position-specific scoring matrix (PSSM), which was generated by three iterations of a PSI-BLAST search (Altschul, et al., 1997) against the non-redundant protein sequence dataset UniRef50 (Suzek, et al., 2015).

Seven physicochemical properties defined by Meiler et al. (Meiler, et al., 2001), including van der Waals volume, hydrophobicity, isoelectric point, and so on, were also used to characterize amino acids. In addition, 17 predicted structural features, such as probabilities of secondary structures, torsional angles, contact number, and so on, were generated using SPIDER2 (Heffernan, et al., 2015). These parameters were used to generate a 44-length feature vector for each amino acid.

As a distinction, in three-state classification, these features were particularly scaled from 0 to 1 by the min-max normalization method. Figure S2 shows the feature matrices for different protein sequences.

Neural network architectures

Architecture of disordered region prediction

Unlike standard feedforward neural networks, like multilayer perceptron (MLP), RNNs can deal with variable-length input sequences by memory functions for previous input information. In particular, LSTM (Gers, et al., 2000; Hochreiter and Schmidhuber, 1997) and bidirectional LSTM (Graves and Schmidhuber, 2005) have excellent learning performance. The bidirectional LSTM improves performance by reading input sequences in both forward and backward directions simultaneously. Owing to its mechanism, the bidirectional LSTM is essentially a deep network architecture and could be difficult to be optimized due to the gradient varnishing or exploding problem. To resolve this issue, residual learning, which adopts shortcut connections to perform identity mapping (He, et al., 2016), was introduced. In this study, the bidirectional LSTM layers and residual connections were combined to obtain disordered region predictions, in an approach termed stacked Residual Bidirectional LSTM network, or Res-BiLstm. In brief, Res-BiLstm consists of five bidirectional LSTM layers with concatenated output vectors from forward and backward LSTM blocks. Residual connections were added from one layer to the next. In the network, the sigmoid unit is adopted at the last discriminated layer for structural binary classification of each amino acid. The layout of the Res-BiLstm network is shown in Fig. S3. In the architecture, 100 neural units are set inside the recurrent hidden layers of LSTM blocks. An input protein sequence is $\mathbf{x} \in \mathbb{R}^{L\times 44}$, where L and 44 indicate the length of the protein sequence and number of input features for an amino acid, respectively. Input biological features for position t within the protein sequence, $\mathbf{x}_t \in \mathbb{R}^{44}$, are represented layer-by-layer and transformed to a fixed vector $\mathbf{h}_t^{l=1} \in \mathbb{R}^{200}$-by performing the following forward transformations on $\mathbf{x}_t$:

$$\vec{\mathbf{h}}_t^{l=1} = \vec{F}^{l=1}\left(\vec{\mathbf{h}}_{t-1}^{l=1}, \vec{\mathbf{s}}_{t-1}^{l=1}, \mathbf{x}_t\right) \quad ,$$

$$\overleftarrow{\mathbf{h}}_t^{l=1} = \overleftarrow{F}^{l=1}\left(\overleftarrow{\mathbf{h}}_{t+1}^{l=1}, \overleftarrow{\mathbf{s}}_{t+1}^{l=1}, \mathbf{x}_t\right),$$

where arrow, $\mathbf{h}$, $\mathbf{s}$, $l = a$, and F are forward or backward direction, the output vector, internal memory of the LSTM cell, a-th layer, and gating functions of LSTM blocks-in the forward and backward direction. The concatenated output vector $h_t^{l=1} \in \mathbb{R}^{200}$is calculated by the following equation:

$$\mathbf{h}_t^{l=1} = [\vec{\mathbf{h}}_t^{l=1}; \overleftarrow{\mathbf{h}}_t^{l=1}].$$

Thus, the output vector from fifth LSTM layer, $\mathbf{h}_t^{l=5}$ is calculated as follows:

$$\mathbf{h}_t^{l=5} = F^{l=5}(\ldots(F^{l=3}(F^{l=2}(\mathbf{h}_t^{l=1}) + \mathbf{h}_t^{l=1}) + \mathbf{h}_t^{l=2})\ldots) + \mathbf{h}_t^{l=4}.$$

Finally, the representation vector $\mathbf{x}_t$ and $\mathbf{h}_t^{l=5}$ are passed into the sigmoid function for structural binary classification. The Res-BiLstm model is trained by a weighted binary cross entropy function to assign each amino acid to one of two classes, disordered or ordered, as follows:

$$E = \frac{1}{L}\sum_{t=1}^{L}(y_t^d \log P(y_t = 1|\mathbf{x}_t) \times \xi + y_t^c \log(1 - P(y_t = 0|\mathbf{x}_t))),$$

where $y_t^d$and y_t^c represent the corresponding class label of disordered and order regions and the class weighting ξ is added to account for unbalanced classes. In our problem, the weight, ξ is set to 8.853 for the disordered class based on the ratio of the two classes among all amino acids of training set.

Prediction of the architecture of transitions

Apart from natively disordered and ordered regions, proteins can also contain regions undergoing a spontaneous disorder–order transition. MoRFs are disordered regions undergoing a unidirectional disorder-to-order transition while binding to partners. Previous computational prediction methods for MoRFs show relatively good performance using biological features extracted from protein sequences (Sharma, et al., 2018). A transfer learning method based on predictions of disordered regions showed improved MoRF identification (Hanson, et al., 2019). These results suggest that disorder–order transition regions can be predicted based on sequence information, and knowledge of disorder is transferable for the prediction of transition regions. In this study, a transfer learning method was implemented to bridge two similar tasks, residue-wise disordered region prediction and transition prediction. Our aim was to improve transition identification learning based on the transition dataset we generated by transferring learned knowledge for a sequence-to-structure mapping based on the disordered protein dataset.

Similar to the architecture of previous MoRF prediction, we extracted latent representations from the pre-trained disordered region predictor Res-BiLstm as input for the transition prediction task. As theoretically demonstrated in the previous subsection, the Res-BiLstm network with a constant error carousel in each LSTM block can effectively overcome gradient vanishing and exploding, which adequately maintains the actual dependencies among amino acids within a sequence. Deep stacking with residual connections between each LSTM layer enables more powerful characterization through layer-by-layer disorder encoding. The last disorder representation containing a high level of disorder–order structure information is reused for further transition encoding. The architecture of the transition prediction is shown in Fig. S4. Our method, Res-BiLstm-NN, involves the last representation layer of the pre-trained predictor Res-BiLstm and an MLP including three fully connected hidden layers with 100 nodes and dropout (ratio of 0.2) in each layer. As in Res-BiLstm, the sigmoid function is used at the last discriminant layer for the binary classification.

In the model, $\mathbf{h}_t^{l=5}$ is used to execute transition encoding, which can be described by the following equations:

$$\boldsymbol{u}_t = \sigma(\mathbf{W}^{l=3}\sigma(\mathbf{W}^{l=2}\sigma(\mathbf{W}^{l=1}\mathbf{h}_t^{l=5} + \mathbf{b}^{l=1}) + \mathbf{b}^{l=2}) + \mathbf{b}^{l=3}),$$

where, σ represents ReLU, $\boldsymbol{u}_t \in \mathbb{R}^{100}$ is the encoded vector from the last fully connected layer, and **W** and **b** are parameters to be learned.

Finally, in the encoded vector $\mathbf{x}_t$, $\mathbf{u}_t$ is passed to a sigmoid decision unit for binary transition classification. Similar to Res-BiLstm, the error is calculated by a weighted binary cross entropy by the following equation:

$$E = \frac{1}{L}\sum_{t=1}^{L}\left(y_t^p \log P(y_t = 1|x_t) \times \xi + y_t^q \log\left(1 - P(y_t = 0|x_t)\right)\right),$$

where $y_t^p$and y_t^q represent the corresponding class labels for structural transitions and non-transitions. The class weigh ξ is set to 11.3 for unbalanced classes.

Optimizer and regularization

The two models Res-BiLstm and Res-BiLstm-NN were trained by the Adam optimizer with a learning rate of 0.001. Early stopping was executed when the validation error stops decreasing in the learning process to reduce the risk of overfitting.

Computational environment

The training processes were implemented using the NVIDIA V100 Graphics Processing Unit for rapid training on the NIG supercomputer at the ROIS National Institute of Genetics. As a framework for neural network implementation, Keras version 2.2.4 and Tensorflow version 1.13.1 were utilized as back-end software.

Results

Disordered region prediction

Performance evaluation

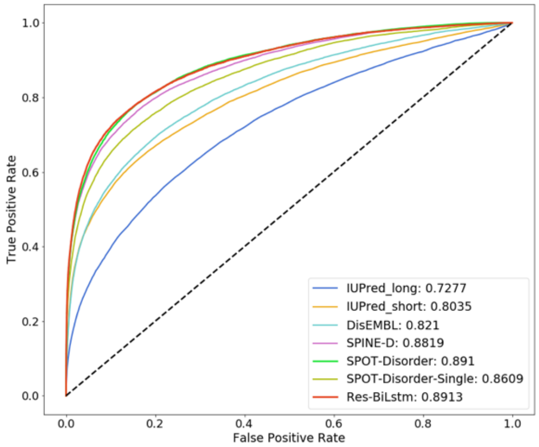

Fig. 1 Comparison of sensitivity, specificity, and AUC values for several disordered region predictors

The performance of Res-BiLstm was evaluated using the test dataset. For comparison, we also evaluated the performance of IUPred (long) and IUPred (short), DisEMBL, SPINE-D, and SPOT-Disorder. As a criterion, we utilized the area under the receiver operating characteristic (ROC) curve (AUC). As shown in Fig. 1, Res-BiLstm achieved the highest AUC for the independent test dataset, indicating the effectiveness and superiority of stacked residual LSTM networks. AUC values were similar for Res-BiLstm and SPOT-Disorder (differing by 0.0003), implying that stacked residual representation learning does not provide a significant benefit compared with a shallow-layers BiLstm method. SPINE-D was comparable to our method, and this can likely be explained by the common dataset used for network training. In addition to the AUC value, other metrics were employed for the balanced and comprehensive measurement of model performance. We used sensitivity and specificity as metrics of the ability to correctly identify disordered and ordered residues, respectively. Owing to unbalanced classes in the disordered protein dataset, traditional accuracy metrics are not directly applicable. Instead, we used a balanced accuracy measurement (ACC) by the calculating mean of the sensitivity and specificity values. We applied various metrics to compare the performance of several disordered region predictors, as shown in Table 1. Res-BiLstm and SPOT-Disorder outperformed other predictors with respect to almost all metrics, including SPOT-Disorder-Single, which utilized a common dataset for network training. This superior performance is likely due to the additional effective evolutionary, structural, and physicochemical features compared with pure sequence information. The performance of SPINE-D, utilizing nearly identical features to those of Res-BiLstm, was relatively weaker than that of our methods, implying that a sliding window approach restricts the discriminant analysis of sequences to structures. In addition to AUC, Res-BiLstm exhibited a much higher sensitivity than that of SPOT-Disorder, thereby yielding the best ACC, indicating that this method exhibits the most balanced performance for correctly identifying both disordered and ordered regions. MCC and F1 values for our method were second only to SPOT-Disorder; it is likely that the higher specificity of SPOT-Disorder plays an important role in unbalanced datasets containing large samples of ordered regions and small samples of disordered regions.

Table 1. Comparison of major metrics among several predictors

	AUC	ACC	Se	Sp	MCC	F1
IUPred (long)	0.728	0.619	0.283	0.955	0.267	0.318
IUPred (short)	0.804	0.706	0.466	0.946	0.403	0.454
DisEMBL	0.821	0.706	0.463	0.950	0.411	0.461
SPINE-D	0.882	0.801	**0.748**	0.854	0.420	0.447
SPOT-Disorder	**0.891**	0.736	0.493	0.978	**0.541**	**0.567**
SPOT-Disorder-Single	0.861	0.676	0.365	**0.987**	0.486	0.485
Res-BiLstm	**0.891**	**0.808**	0.702	0.913	0.493	0.529

The maximum value in each column is shown in bold. Here, ACC, Se, Sp, MCC and F1 stands for accuracy, sensitivity, specificity, Mathew's correlation coefficient and F1 score.

Example of disorder prediction outputs

To determine actual output probabilities for an entire protein sequence, we selected an experimentally annotated protein from the UniProt database (ID: P03176), thymidine kinase. We obtained the prediction probability curves for thymidine kinase from several disordered region predictors, as shown in the Fig. S5.

Res-BiLstm obtained a more stable output than those of predictors based on energy estimation methods. Moreover, the output probabilities obtained using Res-BiLstm included more extreme values (extremely high for the actual disorder class or low for the actual order class) than those of predictors that learn discriminant patterns using other computational methods. Based on these results, our method can more effectively identify both disordered and ordered regions.

Transition prediction

Performance evaluation

The performance of the transition region predictor Res-BiLstm-NN was evaluated on the test dataset. The confusion matrix for the binary classification problem and evaluation metrics are shown in Fig. S6 and Table 2, respectively. As shown in Table S3, our method identified disordered and ordered regions with high accuracy. However, more than half of the transitions were incorrectly classified as structurally non-transitional states, and the ratio of correctly predicted transitions and incorrectly predicted non-transitional regions made it difficult to correctly identify transitions using Res-BiLstm-NN. As presented in Table 2, based on AUC and ACC values, our method exhibited acceptable performance in general. The low MCC and F1 values indicated an unbalanced performance for detecting transition regions. These results reflect the difficulty in predicting transition regions.

Table 2. The performance of Res-BiLstm-NN

	AUC	ACC	Se	Sp	MCC	F1
Res-BiLstm-NN	0.741	0.641	0.432	0.849	0.194	0.259

Example of predicted transition sites

We selected DNA polymerase kappa (UniProt ID: Q9UBT6) to analyze the actual output probabilities of Res-BiLstm-NN,. As presented in Fig. S7, almost all of the transition sites were correctly predicted. However, a number of structurally unchangeable regions were also incorrectly classified, suggesting that some may undergo a potential transformation in specific environmental conditions. Interestingly, the output probability curve exhibits relatively drastic fluctuations even among residues in close proximity, implying that transitions are likely to be determined by or influenced by specific environmental conditions than by their original inherent structures.

The three-dimensional structure of the DNA polymerase kappa protein is presented in Fig. 2. It was found that one of the correctly predicted regions (shown in blue) is closely located to thymidine triphosphate, suggesting that our predictor successfully identified the disorder-order transition region which may associated with ligand binding and then protein function.

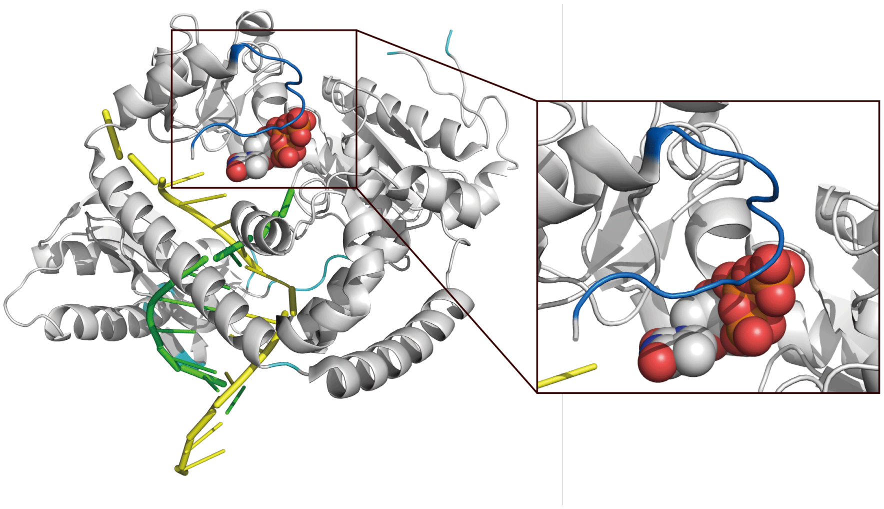

Fig. 2. Three dimensional structure of DNA polymerase kappa (UniProt ID: Q9UBT6, PDB ID: 2OH2 chain B). DNA strands are shown in yellow and green. The ligand (thymidine triphosphate) is shown as sphere. The correctly predicted transition regions are shown in cyan and blue. It is suggested that the disorder-order transition of the blue region might associate with ligand binding.

Discussion

We developed an accurate predictor of disordered protein regions, Res-BiLstm, and applied it to the first large-scale analysis of structural transitions based on a transfer learning method, Res-BiLstm-NN, which uses the internal representation of protein disorder from Res-BiLstm. The AUC values for an independent test dataset indicated an acceptable overall performance but values for sensitivity, MCC, and F1 highlight the difficulty in identifying the transition class. This can potentially be explained by the strong disorder–order preference, potential unknown transitions, and inherently inseparable sequence information distribution.

We applied Res-BiLstm, which produces disordered or ordered predictions for each residue, to the transition site test dataset. This process allowed us to explore the similarity between transitions and natively disorder/order sites. We also performed disordered region prediction using the architecture of Res-BiLstm trained on the order-disorder-transition dataset without transition sites. The results are shown in Fig. 3. The probability distribution predicted by the established predictor is presented in the left-hand panel. A relatively small fraction of transitions was in an intermediate state, and the majority showed a strong disorder or order preference. The output distribution from direct learning (right-hand panel) shows that a large proportion of transitions were classified as ordered regions. Compared to an intermediate state between disorder and order, most transition sites had strong disorder or order preference. Thus, the inseparable similarity due to the strong disorder-order preference can negatively impact model performance. More structural transition sites are likely to be similar to ordered sites than to disordered sites, whereas substantial evidence shows that transition regions prefer the disordered states for entropic advantage when they are in ligand free states, and the number of known disorder-to-order transitions associated with ligand binding far exceeds the opposite cases in common datasets, such as DisProt (Uversky, 2002).

To further explore transition preferences, we performed a three-state (transition, non-transitional order, and non-transitional disorder) classification based on the transition dataset. We employed a similar architecture to Res-BiLstm, but used the softmax function in the last discriminant layer instead of the sigmoid function. As presented in Fig. S8, the low sensitivity of transitions indicates that transition site prediction from sequence information alone is difficult. In other words, evolutionary, structural, and physicochemical information extracted solely from protein sequences might not be sufficient to identify whether a region can undergo a structural transformation or not. Furthermore, transition sites tend to be misclassified to non-transitional ordered sites more frequently than to non-transitional disordered sites. This indicates that transition sites are similar to ordered sites than to disordered sites, to a certain degree. Table S4 shows various metrics for individual binary-class evaluations. Disorder-Others performed better than Order-Others with respect to most metrics, such

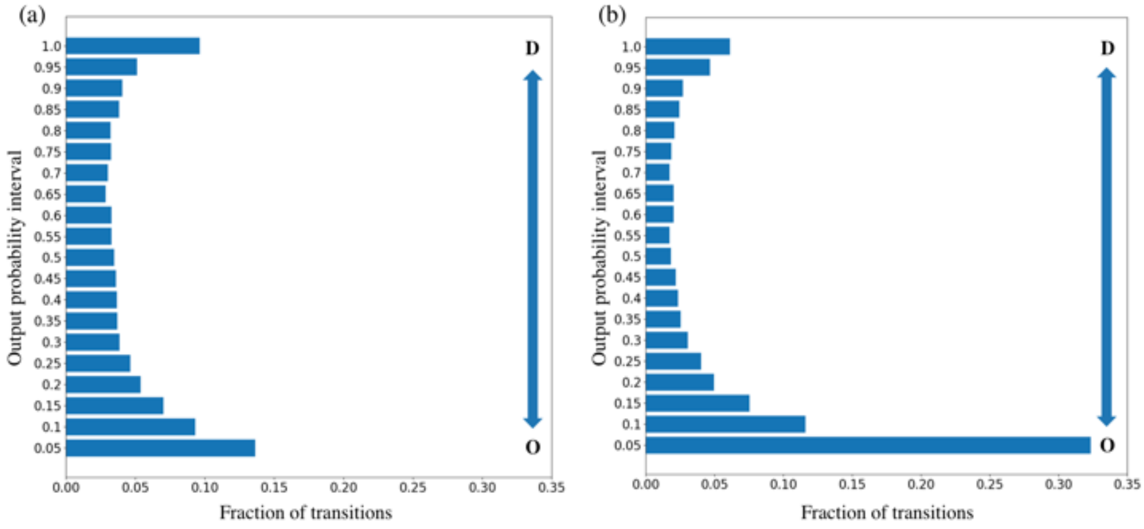

Fig. 3. Predicted output distributions of transitions. Disordered region prediction by (a) Res-BiLstm and (b) same architecture in direct learning.

as AUC, ACC, and MCC, supporting the similarity between transition sites and ordered sites, different from the conventional view.

Conclusion

We developed two accurate protein disordered region predictors, Res-BiLstm and Res-BiLstm-NN. Our predictors showed decent performance for predicting ordered/disordered and transition regions. Exploratory analyses using our predictors revealed that transition sites have unique features compared to non-transition (ordered or disordered) regions. To the best of our knowledge, this is the first attempt to capture the general characteristics of transition sites without assumptions regarding molecular interactions.

Acknowledgements

Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics and on the ToMMo supercomputer at Tohoku Medical Megabank Organization.

Funding

This work was supported in part by Platform Project for Supporting in Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from AMED under Grant Number JP19am0101067.

Conflict of Interest: none declared.

References

Altschul, S.F., *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25(17):3389-3402.

Berman, H.M., *et al*. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography* 2002;58(6):899-907.

Dana, J.M., *et al*. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res* 2019;47(D1):D482-D489.

Disfani, F.M., *et al*. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* 2012;28(12):i75-83.

Dosztanyi, Z., *et al*. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005;21(16):3433-3434.

Du, J., *et al*. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Mol Biosyst* 2014;10(9):2441-2447.

Dyson, H.J. and Wright, P.E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 2005;6(3):197-208.

Gers, F.A., Schmidhuber, J. and Cummins, F. Learning to forget: Continual prediction with LSTM. *Neural Computation* 2000;12(10):2451-2471.

Graves, A. and Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 2005;18(5-6):602-610.

Hanson, J*., et al*. Identifying Molecular Recognition Features in Intrinsically Disordered Regions of Proteins by Transfer Learning. *Bioinformatics* 2019.

Hanson, J., Paliwal, K. and 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.

Hanson, J*., et al*. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* 2017;33(5):685-692.

He, K.M*., et al*. Deep Residual Learning for Image Recognition. *2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr)* 2016:770-778.

Heffernan, R*., et al*. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* 2015;5:11476.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation* 1997;9(8):1735-1780.

Jiang, K.Y*., et al*. Identifying tweets of personal health experience through word embedding and LSTM neural network. *Bmc Bioinformatics* 2018;19.

Kotta-Loizou, I., Tsaousis, G.N. and Hamodrakas, S.J. Analysis of Molecular Recognition Features (MoRFs) in membrane proteins. *Biochim Biophys Acta* 2013;1834(4):798-807.

Linding, R*., et al*. Protein disorder prediction: implications for structural proteomics. *Structure* 2003;11(11):1453-1459.

Liu, J. and Gong, X. Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. *BMC Bioinformatics* 2019;20(1):609.

Liu, Y., Wang, X. and Liu, B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief Bioinform* 2019;20(1):330-346.

Meiler, J*., et al*. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Journal of Molecular Modeling* 2001;7(9):360-369.

Mishra, P.M., Uversky, V.N. and Giri, R. Molecular Recognition Features in Zika Virus Proteome. *J Mol Biol* 2018;430(16):2372-2388.

Mohan, A*., et al*. Analysis of molecular recognition features (MoRFs). *J Mol Biol* 2006;362(5):1043-1059.

Nishi, H*., et al*. Regulation of protein-protein binding by coupling between phosphorylation and intrinsic disorder: analysis of human protein complexes. *Mol Biosyst* 2013;9(7):1620-1626.

Prilusky, J*., et al*. FoldIndex((c)): a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 2005;21(16):3435-3438.

Ruan, H*., et al*. Targeting intrinsically disordered proteins at the edge of chaos. *Drug Discovery Today* 2019;24(1):217-227.

Sharma, R*., et al*. OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. *Bioinformatics* 2018;34(11):1850-1858.

Suzek, B.E*., et al*. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;31(6):926-932.

Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 2005;579(15):3346-3354.

UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2018;46(5):2699.

Uversky, V.N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 2002;11(4):739-756.

Uversky, V.N. and Dunker, A.K. Understanding protein non-folding. *Biochim Biophys Acta* 2010;1804(6):1231-1264.

Uversky, V.N., Oldfield, C.J. and Dunker, A.K. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 2008;37:215-246.

van der Lee, R., *et al*. Classification of intrinsically disordered regions and proteins. *Chem Rev* 2014;114(13):6589-6631.

Velankar, S., *et al*. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* 2013;41(Database issue):D483-489.

Wright, P.E. and Dyson, H.J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 1999;293(2):321-331.

Yamada, K.D. and Kinoshita, K. De novo profile generation based on sequence context specificity with the long short-term memory network. *BMC Bioinformatics* 2018;19(1):272.

Yan, J., *et al*. Molecular recognition features (MoRFs) in three domains of life. *Mol Biosyst* 2016;12(3):697-710.

Zhang, T., *et al*. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *J Biomol Struct Dyn* 2012;29(4):799-813.