Bioinformatics, 33(5), 2017, 685–692
doi: 10.1093/bioinformatics/btw678
Advance Access Publication Date: 5 December 2016
Original Paper

Structural bioinformatics

Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks

Jack Hanson[1,]*, Yuedong Yang[2,]*, Kuldip Paliwal[1] and Yaoqi Zhou[2,]*

[1]Signal Processing Laboratory, Griffith University, Brisbane 4122, Australia and [2]Institute for Glycomics, Griffith University, Gold Coast 4215, Australia

*To whom correspondence should be addressed.

Associate editor: Anna Tramontano

Received on June 7, 2016; revised on September 29, 2016; editorial decision on October 22, 2016; accepted on October 26, 2016

Abstract

Motivation: Capturing long-range interactions between structural but not sequence neighbors of proteins is a long-standing challenging problem in bioinformatics. Recently, long short-term memory (LSTM) networks have significantly improved the accuracy of speech and image classification problems by remembering useful past information in long sequential events. Here, we have implemented deep bidirectional LSTM recurrent neural networks in the problem of protein intrinsic disorder prediction.
Results: The new method, named SPOT-Disorder, has steadily improved over a similar method using a traditional, window-based neural network (SPINE-D) in all datasets tested without separate training on short and long disordered regions. Independent tests on four other datasets including the datasets from critical assessment of structure prediction (CASP) techniques and >10 000 annotated proteins from MobiDB, confirmed SPOT-Disorder as one of the best methods in disorder prediction. Moreover, initial studies indicate that the method is more accurate in predicting functional sites in disordered regions. These results highlight the usefulness combining LSTM with deep bidirectional recurrent neural networks in capturing non-local, long-range interactions for bioinformatics applications.
Availability and Implementation: SPOT-disorder is available as a web server and as a standalone program at: http://sparks-lab.org/server/SPOT-disorder/index.php.
Contact: j.hanson@griffith.edu.au or yuedong.yang@griffith.edu.au or yaoqi.zhou@griffith.edu.au
Supplementary information: Supplementary data is available at *Bioinformatics* online.

1 Introduction

One of the major discoveries in recent years is that protein sequences not only encode for regions that possess well-defined unique three-dimensional structures, but also for regions that lack the tendency to form a structure. These Intrinsically Disordered Proteins (IDPs), or Regions in proteins (IDRs), have been found to fulfill a wide variety of crucial biological roles (commonly involving regulatory and signaling functions) (Dyson and Wright, 2005; Rigden, 2009; Uversky *et al.*, 2005), with a particular prevalence in eukaryotes (Dunker *et al.*, 2000). The flexibility and multiple structural states of IDPs offer unique advantages over ordered proteins (Receveur-Bréchot *et al.*, 2006; Rigden, 2009), thus vindicating the evolutionary development and propagation of intrinsic disorder (Rigden, 2009). In fact, natural protein sequences are more disordered than random sequences (Yu *et al.*, 2016). IDPs have been implicated in many human diseases, including cancer, cardiovascular and neurodegenerative diseases and genetic diseases (Raychaudhuri *et al.*, 2009; Uversky *et al.*, 2005, 2008). Thus, it is important to identify IDPs or IDRs as a means to better understand the functional mechanisms of proteins.

Intrinsic disorder in proteins has been studied by both experimental and theoretical methods. Common experimental methods for determining protein structure, such as X-ray spectroscopy, Nuclear Magnetic Resonance (NMR) spectroscopy and Circular Dichroism (CD) spectroscopy, have been the main tools for characterization of disorder in proteins (Dunker *et al.*, 2001; Dyson and Wright, 2000; Rigden, 2009). These methods make up for the bulk of annotated IDPs and IDRs in the disordered protein database DisProt (Sickmeier *et al.*, 2007), which at its current version (v6.02) consists of 694 IDPs and 1539 proteins containing IDRs taken from the Protein DataBank (PDB), an archive of protein structures (Berman *et al.*, 2000). The largest single collection of disordered proteins has 25 833 annotated proteins (Potenza *et al.*, 2015). Meanwhile, the UniProt/TrEMBL protein sequence database (Bairoch *et al.*, 2005) currently consists of 60 971 489 protein entries, most of which do not yet have their structures nor disordered regions determined. Genome-scale computational studies suggest that 30–50% proteins contain IDRs, and 15% are IDPs (Rigden, 2009; Tompa *et al.*, 2006). Due to the arduous and costly process of determining the flexible structural states of IDPs and IDRs experimentally (Radivojac *et al.*, 2004), the disparity between experimentally annotated and unannotated proteins grows rapidly.

To bridge this growing gap between sequenced and disorder-annotated proteins, many computational methods have been established to discriminate intrinsically disordered regions from structured regions at a fraction of the time and cost of the experimental methods. A majority of these methods employ machine-learning techniques such as Artificial Neural Networks (ANNs) (Rumelhart *et al.*, 1985) and Support Vector Machines (SVMs) (Vapnik, 1998). For example, SVMs have been utilized in MFDp (Mizianty *et al.*, 2010), POODLE L and S (Hirose *et al.*, 2007; Shimizu *et al.*, 2007), PrDOS (Ishida and Kinoshita, 2007), Spritz (Vullo *et al.*, 2006) and with an RBF kernel in the most recently released model, Dispredict (Iqbal and Hoque, 2014). ANNs have also enjoyed widespread success in many accurate predictors, such as DisEMBL (Linding *et al.*, 2003*a*), DISOPRED (Jones and Ward, 2003), DISpro (Cheng *et al.*, 2005), NORSnet (Schlessinger *et al.*, 2007*a*), the PONDR series (Romero *et al.*, 2001), PROFbval (Schlessinger *et al.*, 2006), RONN (Yang *et al.*, 2005), SPINE-D (Zhang *et al.*, 2012) and Espritz (Walsh *et al.*, 2012), which is the only classifier to utilize a Bidirectional Recurrent ANN (BRNN) (Baldi *et al.*, 1999; Schuster and Paliwal, 1997). Other computational methods are based on the analysis of amino acid propensities, physicochemical properties and statistical potential, such as FoldIndex (Prilusky *et al.*, 2005), GlobPlot (Linding *et al.*, 2003*b*), IUPred (Dosztányi *et al.*, 2005) and UCON (Schlessinger *et al.*, 2007*b*). Complementary methods can be combined to form meta-predictors, which employ the outputs from other classifiers to form their own enhanced predictions. Examples are CSpritz (Walsh *et al.*, 2011), MD (Schlessinger *et al.*, 2009), metaPrDOS (Ishida and Kinoshita, 2008), MFDp2 (Mizianty *et al.*, 2013) and PONDR-FIT (Xue *et al.*, 2010).

Recently, the application of deep neural networks with more than two hidden layers to proteins has permitted a better learning of deep and complex relationships between sequences, structures and functions of proteins, and advanced the accuracy of pairwise contact prediction (Di Lena *et al.*, 2012; Eickholt and Cheng, 2012), secondary structure and solvent accessible surface-area prediction (Heffernan *et al.*, 2015, 2016; Paliwal *et al.*, 2015; Qi *et al.*, 2012) and protein disorder prediction (Eickholt and Cheng, 2013; Paliwal *et al.*, 2015). However, common deep learning techniques, such as recurrent neural networks and window-based artificial neural networks, while effective at propagating local errors within sequence neighbors, are ineffective at modeling long-range (non-local) interactions between amino acid residues that are structural but not sequence neighbors (Hochreiter *et al.*, 2001). Because residue–residue interactions are dominated by structural neighbors, how to account for them is the key for improving sequence-based prediction of protein structural and functional properties.

The long-range dependence between a series of time-resolved events can be better captured by enforcing the constant error flow so that useful long-range interactions can be memorized (Hochreiter and Schmidhuber, 1997). In this Long Short-Term Memory (LSTM) network, hidden layers are made of memory blocks containing one or more LSTM cells. Each LSTM cell has the discretion to either forget, input to or output the Constant Error Carousel (CEC, i.e. a fixed weight of 1 in the absence of outside signal). The CEC passes through every LSTM cell in the entire sequential event, acting as a memory backbone effectively connecting the whole sequence (Hochreiter and Schmidhuber, 1997), in either the forwards or backwards directions. LSTM-based neural networks have successfully applied to speech and image-related problems in which long-range memory is the key for accurate interpretation and prediction (Graves and Schmidhuber, 2005; Vinyals *et al.*, 2015).

In this paper, we hope to capture nonlocal interactions that are essential for determining whether a protein will fold (structured) or will not fold (intrinsically disordered) into a unique three-dimensional structure, by employing deep bidirectional LSTM cells. The bidirectional network will allow us to capture both forward and backward information contained in protein sequences. The new method, called SPOT-disorder (Sequence-based Prediction Online Tools for disorder), is found to be highly effective in predicting both short and long disordered regions without separated training (Zhang *et al.*, 2012), despite disordered regions of different sizes having different compositions of amino acids (Radivojac et al., 2004; Rigden, 2009). Independent tests and applications to the targets from Critical Assessment of Structure Prediction (CASP9 and 10) (Monastyrskyy *et al.*, 2011, 2014) have confirmed that SPOT-disorder is comparable to or more accurate than all the methods compared, regardless of which datasets were employed for comparison.

2 The machine learning approach

2.1 Neural network

As shown in Figure 1, the proposed method comprises of a three hidden-layer BRNN (Schuster and Paliwal, 1997), utilizing a recurrent feed-forward layer with a Rectified Linear Unit (ReLU) activation function in the first layer, succeeded by LSTM cell layers in the second and third layer (Hochreiter and Schmidhuber, 1997). In this architecture, the recurrent layer consists of 200 nodes and a bias node in each direction, and the LSTM layers contain 200 one-cell memory blocks of one cell each in each direction. The size of memory block (200 nodes) was chosen after proving to be the best compromise between memory usage and performance accuracy in training. We trained another network with an increased 1000 nodes per layer and found it did not improve on the performance of our method.

The model was trained utilizing the BackPropagation Through Time (BPTT) algorithm (Pineda, 1987; Werbos, 1988; Williams and Zipser, 1989) in the Torch framework (Collobert *et al.*, 2002). This model has compensated for possible overtraining in the training phase by the use of the dropout algorithm, with a dropout percentage of 50% in each of the hidden layers (Srivastava *et al.*, 2014).

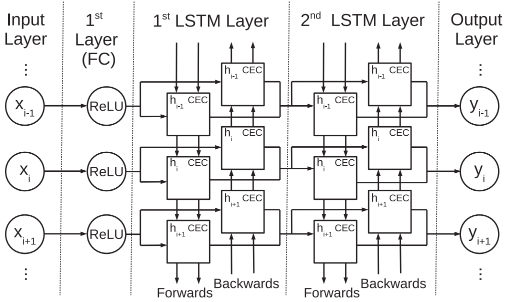

Fig. 1. The layout of the entire network (focusing on sequence position *i*). Here *x* and *y* denote input and output, respectively, while CEC denotes the Constant Error Carousel, FC denotes a fully connected layer and h the immediately recurrent connection in an LSTM.

Torch accounts for dropout during the training phase, meaning that the weights in each layer do not have to be scaled in testing. The training parameters also incorporated a decaying learning rate initialized at 0.001, and a momentum term of 0.99 (Rumelhart *et al.*, 1985).

The learning rate was determined in initial training, where larger learning rates of >0.001 did not enable the network to converge. We employed the step learning rate decay technique to reduce the learning rate by 1% per epoch (Senior *et al.*, 2013). That is, the learning rate was initialized at 0.001, and then systematically annealed to $\approx 6 \times 10^{-4}$ within 50 epochs. This allowed the model to learn finer detail as it progressed through training. Finally, the two outputs of this network are squeezed into a probability distribution through the use of the softmax function (Bishop, 2006).

2.2 Input features

Similar to our previous method (Zhang *et al.*, 2012), input features for disorder prediction included evolutionary, predicted structural properties and physicochemical properties for each amino acid at its position in the polypeptide chain.

The evolutionary content is established through the use of a Position-Specific Scoring Matrix (PSSM), generated by three iterations of the PSI-BLAST algorithm (Altschul *et al.*, 1997) against the NCBI's Non-Redundant (NR) sequence database for each protein. The Shannon entropy is also calculated to represent the information content in these probabilities per residue (Shannon, 1948). The average Shannon entropy over the entire protein is also utilized as an input feature of the general conservation of the whole protein. This leads to a total of 22 evolutionary features used for prediction.

We also obtained 17 predicted structural features from the SPIDER2 predictor (1 ASA, 1 CN, 4 HSE based on Cα and Cβ atoms, respectively and the 11 predicted SS and sine/cosine of the backbone angle values) (Heffernan *et al.*, 2015, 2016). Finally, seven commonly-used physicochemical properties, including hydrophobicity and polarizability, are used as features as provided by Meiler *et al.* (2001).

Thus, using these features resulted in a 46-length feature vector for each amino acid. These parameters were scaled within the range of [0,1] before being passed through the model, based on the minima and maxima of the training data.

2.3 Datasets

Two datasets were used to train and independently test the network. We employed the same dataset DM4229 from the SPINE-D classifier (Zhang *et al.*, 2012) in which 3000 chains (DM3000) were selected for training and cross validation and 1229 chains (DM1229) for independent testing. In addition, we employed two independent test sets (SL and MxD datasets) that were used to test SPINE-D and MFDp (Mizianty *et al.*, 2010), respectively, as well as another independent test set, MobiDB (Potenza *et al.*, 2015; Walsh *et al.*, 2014). The initial DM4229 dataset comprised of the original Disprot dataset, obtained prior to the 5th of August, 2003 (Vucetic *et al.*, 2005). It contains 72 fully disordered proteins from the Disprot database v5.0 (Sickmeier *et al.*, 2007) and 4125 high-resolution, non-redundant structures determined by X-ray crystallography after removing sequences with similarity >25% by the Blastclust algorithm (Altschul *et al.*, 1997).

The SL dataset of 477 proteins (Sirota *et al.*, 2010) was created by re-annotating Disprot to more accurately reflect disordered and ordered regions. The MxD dataset was originally obtained by Mizianty *et al.* (2010) and further reduced by Iqbal and Hoque (2014) to 444 proteins after removing proteins with unknown amino acid residues. Removing the overlap between SL477 and DM4229 using 25% sequence identity cutoff led to SL329 as an independent test set. As some methods employ MxD444 as training set, we also removed the overlap between SL329 and MxD444 and obtained SL117. Furthermore, we obtained the datasets from CASP9 and CASP10 (Monastyrskyy *et al.*, 2011, 2014) as additional test sets.

The MobiDB dataset consists of 25 833 proteins labeled through several methods of curation: data directly taken from Disprot labels; indirectly inferred labels from PDB structures; and predicted labels through a consensus of disorder predictors (Potenza *et al.*, 2015). We trimmed this dataset down to 12 019 proteins after removing the overlap between MobiDB and the SL477 and DM4229 datasets at 20% sequence similarity. Also, proteins with unknown amino acids were discarded, as well as proteins under 30 residues long. This resulted in our final and largest test set, Mobi11925.

2.4 Performance evaluation

Performance is assessed in this paper through the analysis of binary labels and raw prediction values. The raw prediction probabilities are obtained at the output of the network through the use of the softmax function. The discrete labels are generated by the comparison of these probabilities with a pre-calculated threshold *T*. For protein disorder prediction, we assume disorder labels to represent positive samples and order labels to represent negative samples. Each output from the classifier can be sorted into one of four outcomes depending on the label of the sample: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Sensitivity ($S_e = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$) and specificity ($S_p = \frac{\mathrm{TN}}{\mathrm{TN+FP}}$) are two metrics which measure the performance of each class in binary classification. Sensitivity illustrates the classifier's ability to correctly allocate samples into the disordered (or positive) class, whereas specificity does the same for the ordered (or negative) class. These two measures are often combined into a single metric to form the balanced accuracy measurement ($\mathrm{Acc} = \frac{S_e + S_p}{2}$). Another balanced metric is the commonly-used Matthews' Correlation Coefficient (MCC) (Matthews, 1975). This metric calculates the correlation between the predicted and obtained binary classifications, and can be calculated from a confusion matrix as follows:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \quad (1)$$

Even though this measure combines information from both columns in the confusion matrix, this metric presents a balanced metric independent of class skew.

Another single-valued metric is the Youden index S_w. This metric is often used in conjunction with the area under the Receiver Operating Characteristic (ROC) curve (AUC) as it is an indication of the margin between the random predictor (where $S_e = 1 - S_p$) and the obtained ROC for the chosen threshold (Schisterman *et al.*, 2005):

$$S_w = \frac{W_d\mathrm{TP} - W_o\mathrm{FP} + W_o\mathrm{TN} - W_d\mathrm{FN}}{W_dN_d + W_oN_o}, \quad (2)$$

where W_o and W_d are the ratio of disordered and ordered residues, respectively, and N_d and N_o are the total number of disordered and ordered residues, respectively. It was shown by Lobanov *et al.* (2010) that there is a linear relationship between the weighted score and the accuracy metrics ($S_w = 2 \cdot \mathrm{Acc} - 1$). Because these two metrics are tautological, the accuracy metric will be eschewed in this paper for the Youden index. For further AUC analysis, the *P*-value metric can be calculated, which calculates the significance of the difference between two AUC values on a common dataset (Hanley and McNeil, 1982). All of these measurements based on the discrete labels of the predictor have a maximum value of 1, indicating a perfect level of prediction. Therefore, the performance between predictors can be evaluated and compared using the AUC, MCC or S_w metrics after being tested on a common dataset.

2.5 Method comparison

To compare against other methods, we downloaded the standalone version for DisEMBL 1.4 (Download: http://dis.embl.de/html/download.html), DISOPRED 3.16 (Download: http://bioinf.cs.ucl.ac.uk/software_downloads/), Dispredict 1.0 (Download: https://github.com/tamjidul/DisPredict_v1.0), Espritz 1.1 (Download: http://protein.bio.unipd.it/download/), SPINE-D (Download: http://sparks-lab.org/index.php/Main/Downloads) and MetaDisorder (Download: https://github.com/Rostlab/MetaDisorder), which included the predictors NORSnet, Profbval and UCON. We also submitted our sequences to the online servers of Dispro (Server URL: http://scratch.proteomics.ics.uci.edu/), IUP-short and IUP-long (Server URL: http://iupred.enzim.hu/), MFDp and MFDp2 (Server URL: http://biomine-ws.ece.ualberta.ca/MFDp2/), PONDR-fit (Server URL: http://www.disprot.org/pondr-fit.php) and PONDR-VLXT (Server URL: http://www.pondr.com/cgi-bin/PONDR/pondr.cgi). These methods will also provide a comparison between sequence-based and profile-based predictors. Profile-based predictors (such as Dispredict, DISOPRED, the MetaDisorder ensemble, MFDp 1 & 2, SPINE-D and the proposed method) utilize an evolutionary profile in their predictions. This generally provides a more accurate prediction than solely sequence-based methods, but greatly increases the time taken for prediction from the sequence, limiting their practicality for large-scale predictions. Note that we are using the single-sequence model for Espritz due to time constraints, which is known to provide inferior results to the multi-sequence model (Walsh *et al.*, 2012).

3 Results

3.1 Training and testing

SPOT-disorder was trained and cross-validated by the DM3000 set (ten-fold cross validation) and independently tested by DM1229 and SL329. As shown in Table 1, the performance for the two randomly divided large datasets (DM3000 and DM1229) is essentially the same, where the obtained S_w's and AUC's for both datasets are virtually equal, indicating the robustness of the training. For the smaller test set SL329, SPOT-disorder has an even better performance with MCC = 0.67. SPOT-disorder is more accurate than SPINE-D, which employed traditional window-based ANNs for the same cross validations and test sets. The AUC values for DM1229 increased from 0.877 by SPINE-D to 0.894 by SPOT-disorder and S_w from 0.60 to 0.63.

Table 1. Performance of SPOT-disorder on the 10-fold cross-validation subset of DM3000, and test sets DM1229, SL329 and Mobi11925

Dataset	AUC	S_w[c]	MCC[d]	S_e[d]	S_p[d]	#O[a]	#D(long)[b]
DM3000	0.892	0.63	0.60	0.54	0.98	656634	74170 (33778)
DM1229	0.894	0.63	0.57	0.51	0.98	276748	29082 (11496)
SL329	0.905	0.66	0.67	0.67	0.96	51292	39544 (34470)
Mobi11925	0.891	0.63	0.50	0.49	0.98	2907992	191633 (63751)

[a] #O: the number of ordered residues.

[b] #D: the number of disordered residues and the number in long disordered regions in parentheses.

[c] The S_w was obtained using the threshold that maximized the S_w score in DM3000 cross-validation (0.16).

[d] MCC, S_e and S_p were obtained using the threshold which maximized MCC during DM3000 cross-validation (0.49).

Table 2. Performance of SPOT-disorder on the SL329 dataset with individual feature groups removed

Omitted feature group	AUC	S_w	MCC
SPOT-Disorder	0.905	0.661	0.660
PSSM & Entropy	0.903	0.662	0.663
Physicochemical Properties	0.902	0.668	0.672
Backbone Structure	0.902	0.647	0.643
Contacts/Solvent Exposure	0.896	0.642	0.638

Table 2 examines the contributions of individual feature groups to the overall performance of SPOT-disorder in the SL329 test set. We divided features into four groups: evolutionary (PSSM + entropy), physiochemical properties, predicted backbone structure (secondary structure and backbone angles) and predicted contacts (solvent exposure and contact numbers). The table shows that removing contacts predicted by SPIDER2 has the largest impact, followed by secondary structure on all three measures (AUC, Sw and MCC). The importance of predicted contacts for disorder prediction confirms the capability of the LSTM technique to pick up long-range, nonlocal interactions. Interestingly, the evolutionary features (PSSM and entropy) are not as important as one might think, likely due to other features such as secondary structures and contacts being predicted with the PSSM information.

Supplementary Table S1 compares SPOT-disorder against a number of 12 other predictors in addition to SPINE-D in SL329. The ROC curves for these methods are shown in Supplementary Figure S1. In the table, we also present the results from SL290 after excluding 39 chains >1000 residues in length, as some predictors' servers do not accept proteins over this length. We note that the SPINE-D results for SL329 are slightly better than previously reported by Zhang *et al.* (2012), likely because the updated, larger protein sequence library is employed in generating PSSM by PSIBLAST. Table S1 shows that SPOT-disorder has the best performance across all performance metrics excluding Dispredict, which employed the MxD444 dataset as the training set, which was found to have a high sequence similarity to SL329. In fact if we excluding chains in SL329 overlapped with MxD444 (SL117), the performance of Dispredict is significantly worse than SPOT-Disorder (AUC = 0.671, S_w = 0.38, MCC = 0.42 for Dispredict, compared to AUC = 0.930, S_w = 0.71, MCC = 0.70 for SPOT-Disorder). Importantly, the AUC obtained by SPOT-disorder is significantly better than all of the other methods (P-value $< 10^{-5}$ for all methods) across the datasets in Supplementary Table S1.

3.2 CASP9 and CASP10 predictions

To further compare our methods, SPOT-disorder is applied to CASP9 targets and CASP10 targets. Notice that SPINE-D and SPOT-disorder were trained by the same training sets built prior to 2010 and SPINE-D participated in the blind prediction of CASP9. Thus both CASP datasets can be considered as independent test sets for SPOT-disorder. CASP9 has 117 targets with 23656 ordered residues and 2427 disordered residues, with only 560 disordered residues in long disordered regions (≥30 residues). Similarly, CASP10 has 94 targets with 22 688 ordered residues and 1502 disordered residues. There are only 260 disordered residues in long disordered regions (≥30 residues). Thus both datasets mainly test the ability of a computational method for detecting short disordered regions.

The performance of SPOT-disorder in CASP9 and CASP10 is compared to other methods in Supplementary Tables S2 and S3 and Figures S3 and S4 for ROC curves, respectively. The performance of these other methods is obtained from Monastyrskyy *et al.* (2011, 2014), respectively. As CASP predictions contain predictors which are trained on different objectives (i.e. maximizing S_w or MCC), both of the thresholds from Table 1 are used in calculating SPOT-disorder's performance. For CASP9, SPOT-disorder has the best performance in AUC, S_w and MCC values. For CASP10, SPOT-disorder has an insignificantly lower AUC than PrDOS-CNF (0.903 versus 0.907, a one-tailed P-value = 0.28), yet achieves a higher MCC value (0.55 versus 0.53) and the highest S_w (0.63 versus 0.56 for POODLE and metaprdos2). SPOT-disorder scores marginally higher than DISOPRED in AUC (0.903 versus 0.897, a one-tailed P-value = 0.22), but, more significantly, scores higher in MCC (0.55 versus 0.53). The comparison of these results to SPINE-D's performance in the CASP datasets shows that SPOT-disorder significantly outperforms SPINE-D, with one-tailed P-values of 5×10^{-5} and 5×10^{-3} for CASP9 and CASP10, respectively. The relatively higher P-value could be caused by the small set in CASP10, which contains very few long (>30 residues) disordered regions.

3.3 MobiDB

To further confirm the accuracy of SPOT-disorder we applied our method to MobiDB, our largest database consisting of >10 000 annotated proteins. Table 3 and Figure 2 compared SPOT-disorder with 11 other methods. Several methods listed in Supplementary Table S1 (MFDp 1 & 2, and Dispro) are not compared here because the size of the database prohibits its submission for online prediction. We have listed the result of DISOPRED for convenience although it was directly trained by a significant portion of MobiDB. SPOT-disorder continues to consistently improve over other methods compared. In particular, the difference between the ROC curves given by SPOT-disorder and the next best SPINE-D (not including DISOPRED) is statistically significant with a one-tailed p-value of $< 10^{-6}$. Our values of AUC, S_w and MCC are 0.891, 0.630 and 0.401, respectively, all of which are the highest among all the methods compared, excluding DISOPRED. We noted that faster single-sequence-based models are generally less accurate than slower profile-based methods due to their lack of evolutionary information on this dataset.

Table 3. Performance of various methods on Mobi11925, grouped by use of evolutionary features

Predictor		Mobi11925		
		AUC	S_w	MCC
Profile	SPOT-disorder[a]	0.891	0.630	0.401
	DISOPRED[b,c]	0.887	0.466	0.494
	SPINE-D	0.882	0.610	0.397
	MD[b]	0.813	0.405	0.305
	NORS[b]	0.738	0.227	0.234
	PROFbval[b]	0.734	0.335	0.165
	UCON[b]	0.734	0.312	0.206
Sequence	Espritz	0.824	0.164	0.333
	PONDR-fit	0.822	0.441	0.341
	DisEMBL	0.789	0.343	0.327
	IUP-short	0.784	0.368	0.341
	IUP-long	0.740	0.289	0.270
	PONDR	0.733	0.336	0.192

[a] SPOT-disorder's MCC value is slightly different from that reported in Table 1 as the threshold used here is based on S_w.

[b] Omits a single protein of length >10000 residues long.

[c] DISOPRED employed a significant portion of proteins in Mobi11925 as the training set. The result is listed here for convenience.

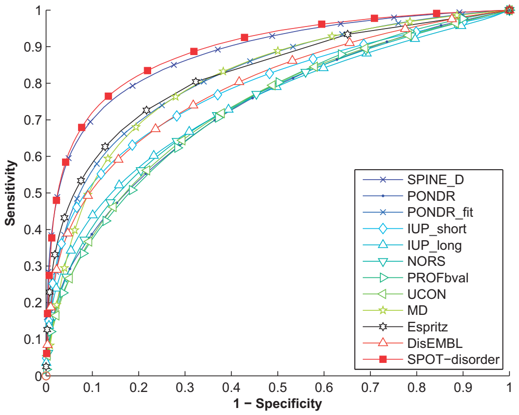

Fig. 2. Receiver Operating Characteristic curves by various methods as labeled for the Mobi11925 dataset.

Also of note is that AUC and S_w values are essentially the same among DM3000, DM1229 and Mobi11925 as shown in Table 1. Mobil11925 has the lowest MCC value because it has the lowest ratio of disordered residues (disorder:order = 1:15).

3.4 Length dependence

In our previous method SPINE-D, residues in long disordered (≥30 residues) regions are predicted separately from those in short disordered (<30 residues) regions. It is of interest to know if SPOT-Disorder would be biased toward disordered regions of certain length without length-separated training. Supplementary Figure S2 compares the performance of SPOT-disorder with other methods for the SL290 dataset in disordered regions of different sizes. It is clear that SPOT-disorder either matches or exceeds the performance of SPINE-D in each length segment and improves over other methods in comparison, suggesting a small but systematic overall improvement. SPOT-disorder also follows the general curve of most of the predictors, illustrating regions in the dataset which are more easily predicted than others, such as regions of length (45,60] and (120,180].

To remove the uncertainty due to the small sample size, we further examined length dependence using the large MobiDB dataset. As shown in Figure 3, the larger amount of data makes a smoother dependence of performance on the size of disordered regions. Interestingly, SPOT-Disorder improves over SPINE-D consistently for all sizes of disordered regions except the last two bins: (120, 180] and >180. The improvement in shorter disordered regions is consistent with the fact that these disordered regions tend to associated with more nonlocal interactions as they are embedded within structured regions.

4 Discussion

This paper represents the first application of a long short-term memory network in protein disorder prediction. The new technique has the best performance or has matched the best performance among all methods compared for all independent test sets (Table 3; Supplementary Tables S1–S3). This was achieved despite SPOT-Disorder being trained on an unbalanced dataset (DM3000), and independently tested on both a more balanced dataset (SL329), datasets dominated by short disordered regions (CASP9 and CASP10) and the large MobiDB dataset of >10 000 proteins.

One strength of SPOT-disorder is its ability to handle disordered residues in disordered regions of different lengths. Some previous methods relied on separate training of short or long disordered regions in a single method (Zhang *et al.*, 2012) or two different methods such as short and long versions in IUPRED (Dosztányi *et al.*, 2005) and PONDR (Peng *et al.*, 2005, 2006; Romero *et al.*, 2001). Other methods such as DISpro (Cheng *et al.*, 2005) and Predisorder (Deng *et al.*, 2009) trained on disordered regions without separating long and short disordered regions. SPOT-disorder confirmed that it is possible to provide the best prediction in short and long disordered regions without specific training. Interestingly, according to the largest test set (Figure 3), most methods, including this study, have the best performance for the disordered regions of (30,45] amino acids long. The performance goes down quickly for the long disordered regions with more than 180 residues, indicating that prediction of very long disordered regions remains a challenge.

While the main objective of this study is to highlight the importance of long short-term memory for further improving on the current disorder predictors, it is of interest to examine its ability to identify potential functional regions in disordered proteins. Early studies have shown that a dip in disorder probability is associated with induced folding (Mohan et al., 2006; Oldfield *et al.*, 2005). Several machine-learning tools were developed specifically for prediction of binding sites in disorder (Garner *et al.*, 1999), such as alpha-MoRF (Oldfield *et al.*, 2005), MoRFpred (Disfani *et al.*, 2012), ANCHOR (Mészáros *et al.*, 2009) and DISOPRED3 (Jones and Cozzetto, 2015). We previously showed that semi-disorder (defined as predicted disorder probability around 0.5) is implicated in induced folding and protein aggregation (Zhang *et al.*, 2013) and plays a significant role in temperature adaption of disordered proteins (Wang *et al.*, 2013). As an initial test for SPOT-disorder in binding site prediction, we employed a test set of 9 proteins with annotated disordered binding sites, collected by Jones and Cozzetto (Jones and Cozzetto, 2015). Following our previous work (Zhang *et al.*, 2013), we defined semi-disorder as a region enclosing a predicted disorder probability of 0.5. Table 4 shows that using only two parameters for defining semi-disorder, SPOT-disorder yields substantially more accurate prediction of binding sites (MCC = 0.31 and $S_w = 0.41$) than other methods specifically trained for predicting binding sites (MCC<0.13 and S_w<0.12). The optimized semi-disorder region is found to be between 0.28 and 0.69. This result highlights the potential of SPOT-disorder in predicting functional regions of protein disorder and strengthens the existence of a third state between disordered and structured states to account for induced folding and aggregation in intrinsically disordered regions (Zhang *et al.*, 2013).

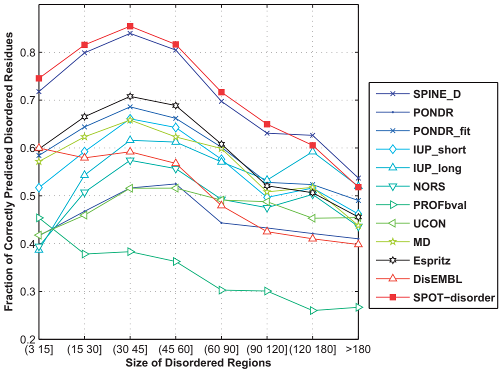

Fig. 3. The classification accuracy of the predictors per the size of the disordered regions for the Mobi11925 dataset.

These successes can be attributed to the ability of LSTMs to recognize non-local interactions, as the limit of focusing on local interactions has seemingly been reached by the use of traditional neural networks. Steady further improvement in all of the datasets tested suggests the usefulness of LSTMs in other bioinformatics areas where non-local interactions are important.

In this study, to greatly improve the speed of training this network, we trained this network using a Graphics Processing Unit (GPU) for accelerated training. The use of GPUs in training neural networks has been shown to decrease the training time by a factor of ≈20 (Oh and Jung, 2004). Training the network through 50 epochs took about 1 week on a single Quadro K5200 graphics card. The testing phase of a neural network involves doing a forward pass through the network, which is computationally trivial and can easily be performed on a CPU. For example, a forward pass only takes 0.8 seconds for a protein with 110 residues and 6 seconds for a protein with 1194 residues. The most time-consuming portion is calculation of the PSSM by PSI-BLAST. It takes about 10 and 50 minutes/iteration for those same two proteins, respectively, on a single CPU of Intel Xeon E5-1650 v2 @3.50GHz. The computing cost of PSSM is the main reason for slow calculation for profile-based methods.

Table 4. Prediction of protein binding sites from several predictors

Predictor	S_w	MCC	Sens	Spec
SPOT-disorder (dual)	0.407	0.309	0.52	0.89
SPOT-disorder (single)[a]	0.239	0.142	0.96	0.28
DISOPRED3[b]	0.105	0.126	0.15	0.96
MoRFpred[b]	0.112	0.104	0.19	0.92
MFSPSSMpred[b]	0.106	0.092	0.21	0.90
ANCHOR[b]	−0.175	−0.092	0.29	0.54

[a] Results obtained by considering disordered regions as binding sites and using only one threshold.

[b] Results obtained from Jones and Cozzetto (2015).

Acknowledgements

We gratefully acknowledge the support of the Griffith University eResearch Services Team and the use of the High Performance Computing Cluster "Gowonda" to complete this research. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

Funding

This work is supported by National Health and Medical Research Council of Australia (Contract grant numbers: 1059775 and 1083450); Australian Research Council s Linkage Infrastructure, Equipment and Facilities funding scheme (Contract grant number: LE150100161) to Y.Z.

References

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

Bairoch,A. *et al*. (2005) The universal protein resource (uniprot). *Nucleic Acids Res*., **33** (**suppl 1**), D154–D159.

Baldi,P. *et al*. (1999) Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**, 937–946.

Berman,H.M. *et al*. (2000) The protein data bank. *Nucleic Acids Res*., **28**, 235–242.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer, Berlin.

Cheng,J. *et al*. (2005) Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining Knowl. Discov*., **11**, 213–222.

Collobert,R. *et al*. (2002). Torch: a modular machine learning software library. Technical report IDIAP.

Deng,X. *et al*. (2009) Predisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics*, **10**, 436.

Di Lena,P. *et al*. (2012) Deep architectures for protein contact map prediction. *Bioinformatics*, **28**, 2449–2457.

Disfani,F.M. *et al*. (2012) Morfpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*, **28**, i75–i83.

Dosztányi,Z. *et al*. (2005) Iupred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.

Dunker,A.K. *et al*. (2001) Intrinsically disordered protein. *J. Mol. Graph. Model*., **19**, 26–59.

Dunker,A.K. *et al*. (2000) Intrinsic protein disorder in complete genomes. *Genome Informatics*, **11**, 161–171.

Dyson,H.J. and Wright,P.E. (2000) Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. *Methods Enzymol*., **339**, 258–270.

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

Eickholt,J. and Cheng,J. (2012) Predicting protein residue–residue contacts using deep networks and boosting. *Bioinformatics*, **28**, 3066–3072.

Eickholt,J. and Cheng,J. (2013) Dndisorder: predicting protein disorder using boosting and deep networks. *BMC Bioinformatics*, **14**, 1.

Garner,E. *et al*. (1999) Predicting binding regions within disordered proteins. *Genome Informatics*, **10**, 41–50.

Graves,A. and Schmidhuber,J. (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw*., **18**, 602–610.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, **143**, 29–36.

Heffernan,R. *et al*. (2016) Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics*, **32**, 843–849.

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

Hirose,S. *et al*. (2007) Poodle-l: a two-level svm prediction system for reliably predicting long disordered regions. *Bioinformatics*, **23**, 2046–2053.

Hochreiter,S. *et al*. (2001) *Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies*. IEEE Press, Piscataway, NJ.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput*., **9**, 1735–1780.

Iqbal,S. and Hoque,M.T. (2014). Dispredict: a predictor of disordered protein from sequence using rbf kernel. Technical report Tech. Report TR-2014/1.

Ishida,T. and Kinoshita,K. (2007) Prdos: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res*., **35** (**suppl 2**), W460–W464.

Ishida,T. and Kinoshita,K. (2008) Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, **24**, 1344–1348.

Jones,D.T. and Cozzetto,D. (2015) Disopred3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*, **31**, 857–863.

Jones,D.T. and Ward,J.J. (2003) Prediction of disordered regions in proteins from position specific score matrices. *Proteins Struct. Funct. Bioinformatics*, **53**, 573–578.

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

Linding,R. *et al*. (2003b) Globplot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res*., **31**, 3701–3708.

Lobanov,M.Y. *et al*. (2010) Library of disordered patterns in 3d protein structures. *PLoS Comput. Biol*., **6**, e1000958.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct*., **405**, 442–451.

Meiler,J. *et al*. (2001) Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Mol. Model. Annu*., **7**, 360–369.

Mészáros,B. *et al*. (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol*., **5**, e1000376.

Mizianty,M.J. *et al*. (2013) Mfdp2: accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. *Intrinsically Disord. Proteins*, **1**, e24428.

Mizianty,M.J. *et al*. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, **26**, i489–i496.

Mohan,A. *et al*. (2006) Analysis of molecular recognition features (morfs). *J. Mol. Biol*., **362**, 1043–1059.

Monastyrskyy,B. *et al*. (2011) Evaluation of disorder predictions in casp9. *Proteins: Struct. Funct. and Bioinformatics*, **79**, 107–118.

Monastyrskyy,B. *et al*. (2014) Assessment of protein disorder region predictions in casp10. *Proteins Struct. Funct. and Bioinformatics*, **82**, 127–137.

Oh,K.S. and Jung,K. (2004) Gpu implementation of neural networks. *Pattern Recogn*., **37**, 1311–1314.

Oldfield,C.J. *et al*. (2005) Coupled folding and binding with α-helix-forming molecular recognition elements. *Biochemistry*, **44**, 12454–12470.

Paliwal,K. *et al*. (2015) A short review of deep learning neural networks in protein structure prediction problems. *Adv. Tech. Biol. Med*., **3**, 139.

Peng,K. *et al*. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.

Peng,K. *et al*. (2005) Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinformatics Comput. Biol*., **3**, 35–60.

Pineda,F.J. (1987) Generalization of back-propagation to recurrent neural networks. *Phys. Rev. Lett*., **59**, 2229.

Potenza,E. *et al*. (2015) Mobidb 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res*., **43**, D315–D320.

Prilusky,J. *et al*. (2005) Foldindex[copyright]: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.

Qi,Y. *et al*. (2012) A unified multitask architecture for predicting local protein properties. *PLoS One*, **7**, e32235.

Radivojac,P. *et al*. (2004) Protein flexibility and intrinsic disorder. *Protein Sci*., **13**, 71–80.

Raychaudhuri,S. *et al*. (2009) The role of intrinsically unstructured proteins in neurodegenerative diseases. *PLoS One*, **4**, e5566.

Receveur-Bréchot,V. *et al*. (2006) Assessing protein disorder and induced folding. *Proteins Struct. Funct. Bioinformatics*, **62**, 24–45.

Rigden,D.J. (2009) *From Protein Structure to Function with Bioinformatics*. Springer, Berlin.

Romero,P. *et al*. (2001) Sequence complexity of disordered protein. *Proteins Struct., Funct. Bioinformatics*, **42**, 38–48.

Rumelhart,D.E. *et al*. (1985). Learning internal representations by error propagation. Technical report DTIC Document.

Schisterman,E.F. *et al*. (2005) Optimal cut-point and its corresponding youden index to discriminate individuals using pooled blood samples. *Epidemiology*, **16**, 73–81.

Schlessinger,A. *et al*. (2007a) Natively unstructured loops differ from other loops. *PLoS Comput. Biol*., **3**, e140.

Schlessinger,A. *et al*. (2007b) Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, **23**, 2376–2384.

Schlessinger,A. *et al*. (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One*, **4**, e4433.

Schlessinger,A. *et al*. (2006) Profbval: predict flexible and rigid residues in proteins. *Bioinformatics*, **22**, 891–893.

Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* , **45**, 2673–2681.

Senior,A. *et al*. (2013) An empirical study of learning rates in deep neural networks for speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6724–6728.

Shannon,C.E. (1948) A note on the concept of entropy. *Bell Syst. Tech. J*., **27**, 379–423.

Shimizu,K. *et al*. (2007) Poodle-s: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics*, **23**, 2337–2338.

Sickmeier,M. *et al*. (2007) Disprot: the database of disordered proteins. *Nucleic Acids Res*., **35 (suppl 1)**, D786–D793.

Sirota,F.L. *et al*. (2010) Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. *BMC Genomics*, **11 (Suppl 1)**, S15.

Srivastava,N. *et al*. (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res*., **15**, 1929–1958.

Tompa,P. *et al*. (2006) Prevalent structural disorder in e. coli and s. cerevisiae proteomes. *J. Proteome Res*., **5**, 1996–2000.

Uversky,V.N. *et al*. (2005) Showing your id: intrinsic disorder as an id for recognition, regulation and cell signaling. *J. Mol. Recogn*., **18**, 343–384.

Uversky,V.N. *et al*. (2008) Intrinsically disordered proteins in human diseases: introducing the d2 concept. *Annu. Rev. Biophys*., **37**, 215–246.

Vapnik,V.N. (1998) *Statistical Learning Theory*, vol. **1**. Wiley, New York.

Vinyals,O. *et al*. (2015) Show and tell: a neural image caption generator. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164.

Vucetic,S. *et al*. (2005) Disprot: a database of protein disorder. *Bioinformatics*, **21**, 137–140.

Vullo,A. *et al*. (2006) Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res*., **34 (suppl 2)**, W164–W168.

Walsh,I. *et al*. (2014) Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics*, **31**, 201–208.

Walsh,I. *et al*. (2012) Espritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.

Walsh,I. *et al*. (2011) Cspritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res*., **39 (suppl 2)**, W190–W196.

Wang,J. *et al*. (2013) The role of semidisorder in temperature adaptation of bacterial flgm proteins. *Biophys. J*., **105**, 2598–2605.

Werbos,P.J. (1988) Generalization of backpropagation with application to a recurrent gas market model. *Neural Netw*., **1**, 339–356.

Williams,R.J. and Zipser,D. (1989) A learning algorithm for continually running fully recurrent neural networks. *Neural Comput*., **1**, 270–280.

Xue,B. *et al*. (2010) Pondr-fit: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta Proteins Proteomics*, **1804**, 996–1010.

Yang,Z.R. *et al*. (2005) Ronn: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.

Yu,J.F. *et al*. (2016) Natural protein sequences are more intrinsically disordered than random sequences. *Cell. Mol. Life Sci*., **73**, 2949–2957.

Zhang,T. *et al*. (2013) Intrinsically semi-disordered state and its role in induced folding and protein aggregation. *Cell Biochem. Biophys*., **67**, 1193–1205.

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