AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields

Sheng Wang¹,²,*, Jianzhu Ma¹ and Jinbo Xu¹,*

¹Toyota Technological Institute at Chicago, Chicago, IL, USA and ²Department of Human Genetics, University of Chicago, Chicago, IL, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Protein intrinsically disordered regions (IDRs) play an important role in many biological processes. Two key properties of IDRs are (i) the occurrence is proteome-wide and (ii) the ratio of disordered residues is about 6%, which makes it challenging to accurately predict IDRs. Most IDR prediction methods use sequence profile to improve accuracy, which prevents its application to proteome-wide prediction since it is time-consuming to generate sequence profiles. On the other hand, the methods without using sequence profile fare much worse than using sequence profile.

Method: This article formulates IDR prediction as a sequence labeling problem and employs a new machine learning method called Deep Convolutional Neural Fields (DeepCNF) to solve it. DeepCNF is an integration of deep convolutional neural networks (DCNN) and conditional random fields (CRF); it can model not only complex sequence–structure relationship in a hierarchical manner, but also correlation among adjacent residues. To deal with highly imbalanced order/disorder ratio, instead of training DeepCNF by widely used maximum-likelihood, we develop a novel approach to train it by maximizing area under the ROC curve (AUC), which is an unbiased measure for class-imbalanced data.

Results: Our experimental results show that our IDR prediction method AUCpreD outperforms existing popular disorder predictors. More importantly, AUCpreD works very well even without sequence profile, comparing favorably to or even outperforming many methods using sequence profile. Therefore, our method works for proteome-wide disorder prediction while yielding similar or better accuracy than the others.

Availability and Implementation: http://raptorx2.uchicago.edu/StructurePropertyPred/predict/

Contact: wangsheng@uchicago.edu, jinboxu@gmail.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1. Introduction

It has been known that proteins without stable structure could still perform biological functions (Jirgensons, 1958). With more proteins (or protein regions) of no unique structure being discovered (Jensen et al., 2013), the intrinsically disordered proteins (IDP) or intrinsically disordered regions (IDR) are found to be involved in many important biological processes (Oldfield and Dunker, 2014), and their occurrence is proteome-wide (Nguyen Ba et al., 2012), where at least 6% of the residues in the annotated sequences from SwissProt (Boeckmann et al., 2003) were considered as disordered (Di Domenico et al., 2012). The discovered IDP/IDRs are curated and stored in the DisProt database (Sickmeier et al., 2007). Since the experimental structure determination methods are laborious and expensive, IDRs are accumulating slowly in DisProt (Sickmeier et al., 2007). Therefore, it is important to develop accurate and high-throughput computational methods to predict disordered regions.

Roughly speaking, computational methods for IDR prediction belong to three categories: sequence-based, template-based and consensus (Deng et al., 2012; He et al., 2009). Sequence-based methods predict IDRs using sequence (and/or sequence profile) information, but not experimentally solved structures (as templates) (Dosztányi et al., 2005; Eickholt and Cheng, 2013; Walsh et al., 2012; Ward et al., 2004; Zhang et al., 2012). Template-based methods (Jones
and Cozzetto, 2015) use the solved protein structures as templates to help with IDR prediction. Consensus methods combine individual predictors (Ishida and Kinoshita, 2007; Hirose et al., 2007; Kozlowski and Bujnicki, 2012; Xue et al., 2010), either sequence-based or template-based, to yield better accuracy. For sequence-based methods, some predictors make use of sequence profile (Jones and Cozzetto, 2015; Walsh et al., 2012; Zhang et al., 2012), while the others do not (Dosztányi et al., 2005; Peng et al., 2006; Frilusky et al., 2005; Walsh et al., 2012). Since sequence profiles contain evolutionary information, they usually result in better prediction accuracy (Walsh et al., 2012). However, it is time-consuming to generate sequence profiles, so profile-based methods may not be suitable for large-scale proteome-wide prediction. Further, there are also a large number of proteins with very sparse sequence profiles, so it is still important to study IDR prediction without sequence profiles (Deng et al., 2015; Walsh et al., 2012).

Many methods employ machine learning to integrate different protein features for IDR prediction (Eickholt and Cheng, 2013; Hirose et al., 2007; Yang et al., 2005). Methods such as Neural Networks (Yang et al., 2005) and Support Vector Machines (Hirose et al., 2007) ignore interdependency among residues and predict the disorder state of one residue independent of the others. Nevertheless, the disorder states of sequentially adjacent residues are correlated (Romero et al., 1998). To exploit this correlation, (Wang and Sauer, 2008) describes a linear-chain Conditional Random Fields (CRF) (Lafferty et al., 2001) for IDR prediction.

This article presents a new, in-house deep learning model called Deep Convolutional Neural Fields (DeepCNF) (Wang et al., 2016a) for proteome-wide IDR prediction. DeepCNF is an integration of Conditional Random Fields (CRF) and Deep Convolutional Neural Network (DCNN) (Lee et al., 2009). DeepCNF can also be viewed as CRF with DCNN as its non-linear feature generating function. DeepCNF embraces the merits of both CRF and DCNN, so it can model not only complex sequence-structure relationship, but also correlation among adjacent residues. Few deep learning methods have been developed for disorder prediction, such as DNdisorder (Eickholt and Cheng, 2013) by Cheng’s group. However, our new method AUCpred differs from Cheng’s work as follows. AUCpred is designed specifically for efficient proteome-wide IDR prediction since it works well without using sequence profile, while Cheng’s method needs sequence profile. In algorithmic level, (a) we use DCNN while DNdisorder uses a deep belief network (DBN) constructed from restricted Boltzmann machine (RBM). DCNN is better than DBN in capturing longer-range sequential information in a hierarchical manner; (b) our method considers the correlation of the ‘ordered/disordered’ states of sequentially adjacent residues while DNdisorder does not; and (c) we use a better and novel training algorithm.

Since the distribution of ordered and disordered residues is highly imbalanced, we develop a new method instead of the widely used maximum-likelihood method to train DeepCNF. That is, we train DeepCNF by maximizing area under the ROC curve (AUC), which is an unbiased measure for class-imbalanced data. It is important to deal with the issue of imbalanced class distribution for IDR prediction since by predicting all residues to be ordered we may have ~94% per-residue accuracy, but such a prediction has AUC ~ 0.5 and is completely useless. AUC is insensitive to changes in class distribution (Cortes and Mohri, 2004) because that the ROC curve specifies the relationship between false positive (FP) rate and true positive (TP) rate, which are independent of class distribution (Fawcett, 2004). However, it is very challenging to directly optimize AUC. A few algorithms have been developed to maximize AUC on imbalanced unstructured data (Herschtal and Raskutti, 2004; Joachims, 2005; Narasimhan and Agarwal, 2013), but to the best of our knowledge, there is no such an algorithm for imbalanced sequence data addressed here. To maximize AUC of our DeepCNF model, we formulate the AUC function in a ranking framework, approximate it by a polynomial Chebyshev function (Calders and Jaroszewicz, 2007) and then use L-BFGS (Liu and Nocedal, 1989) to optimize it.

Our experimental results on several benchmarks show that our disorder prediction method AUCpred greatly outperforms existing popular disorder predictors of the same category (i.e. sequence-based methods), including the recent deep learning methods (Eickholt and Cheng, 2013; Wang et al., 2015). Even without sequence profile, our method compares favorably to or outperforms other methods using sequence profile. Since it is time-consuming to generate sequence profiles, this makes our method an ideal tool for high-quality proteome-wide disorder prediction.

2 Results

2.1 Datasets

We use four publically available datasets to train, validate, and evaluate our method AUCpred. In particular, we use the UniProt90 dataset (Di Domenico et al., 2012) to train the model parameters, and employ the CASP9 (Monastyrskyy et al., 2011), CASP10 (Monastyrskyy et al., 2014) targets and the recent CAMEO (http://www.cameo3d.org/sp/1-year/) targets to evaluate prediction accuracy (Haas et al., 2013). To remove redundancy with the evaluation data, we filter UniProt90 by the following two criteria: (i) proteins released after May 01, 2010 (i.e. the start date of CASP9) are removed; and (ii) proteins sharing >25% sequence identity with the CASP and CAMEO targets are removed. In total, there are 13 800 proteins in the final UniProt90 dataset. A 10-fold cross-validation was performed to train the model parameters. For more details, see Supplemental Material.

CASP9 and CASP10 have 117 and 94 test proteins, respectively. We merge them into a single dataset. CASP11 was held in 2014, but did not release any disorder prediction test targets. We also tested the competing methods on the 229 hard CAMEO targets released from September 16, 2014 to September 16, 2015.

2.2 Programs to compare

Since our method is sequence-based, we do not compare it with some past CASP-participating programs such as DISOclust (McGuffin, 2008), biomine_dr_mixed (Monastyrskyy et al., 2014), POODLE (Monastyrskyy et al., 2014), MULTICOM-construct (Monastyrskyy et al., 2014), IntFoldD2-DR (Roche et al., 2011) and GSMetadisorder (Kozlowski and Bujnicki, 2012) because they are mainly consensus-based. However, we do compare with consensus-based PrDOS-CNF (Ishida and Kinoshita, 2007, 2008) since it is the winner of CASP10. Instead, we mainly compare our method with the following sequence- and/or template-based predictors: DNdisorder (Eickholt and Cheng, 2013), IUpred (Dosztányi et al., 2005), Espritz (Walsh et al., 2012) and DisoPred3 (Jones and Cozzetto, 2015).

IUpred does not use sequence profile and has two prediction modes: IUpred-long (or IUpred+) for long disordered regions and IUpred-short (or IUpred+) for short ones. Espritz also has two prediction modes: one uses sequence profile (denoted as Espritz+) and the other (denoted as Espritz-) does not. Espritz has three different versions trained by X-ray structures, NMR structures and long
disordered regions derived from DisoProt, respectively. We test the version trained by X-ray structures, which performs the best among the three (Walsh et al., 2012). DisoPred3 makes use of sequence profile and also supports template-based prediction. We use DisoPred3T and DisoPred3A to denote the template-based and profile-based modes, respectively. All the programs are run with their default parameters.

To show the performance of maximized-AUC training, we compare AUCpreD with the same deep architecture trained by maximal likelihood, which is denoted as DeepCNF-D (Wang et al., 2015). To do a fair comparison, we also create two prediction modes: one uses sequence profile (denoted as DeepCNF-Dp) and the other (denoted as DeepCNF-Da) does not.

### 2.3 Evaluation criteria

Five performance metrics are used: (i) threshold-dependent measures including balanced accuracy (Acc), sensitivity (Sens), specificity (Spec) and the Matthews correlation coefficient (Mcc); and (ii) threshold-independent measures including area under the Precision-Recall curve (AUCpreD), or typically denoted as Average Precision) and area under the ROC curve (AUC). Each method predicts a likelihood score (probability) of one residue being in a disordered state and then classifies a residue with a predicted score larger than a given threshold as disordered. Each competing method has its own default threshold value. For our method AUCpreD, we set this default threshold to 0.5 according to CASP (Monastyrskyy et al., 2014).

True positives (TP) and true negatives (TN) are the numbers of correctly predicted disordered and ordered residues, respectively; whereas false positives (FP) and false negatives (FN) are the numbers of misclassified disordered and ordered residues, respectively. Sensitivity and specificity are defined as TP/(TP + FN) and TN/(TN + FP), respectively. Recall is the same as sensitivity, and precision is defined as TP/ (TP + FP). Acc is the average of sensitivity and specificity. Mcc is defined as (TP × TN − FP × FN)/√[(TP + FP)(TN + FP)(TP + FN)(TN + FN)]. Given a certain threshold, ROC curve plots 1-specificity and sensitivity, whereas Precision-Recall curve plots recall and precision.

Since only ~6% of residues are disordered, Sens, Spec and Acc may not be as good measures (Fawcett, 2004) as Mcc, AUC and AUCp, which are generally regarded as balanced measures that can be used on class-imbalanced data. MCC ranges from −1 to +1, with +1 representing a perfect prediction, 0 random prediction and −1 total disagreement between prediction and ground truth. AUC has a minimum value 0.0, a random value 0.5, and the best value 1.0. AUCp has a minimum value 0.0 and the best value 1.0. The ROC curves are known to overestimate performance of predictors on the imbalanced data (Davis and Goadrich, 2006). To address this potential issue, we complemented the AUC analysis with AUCp.

### 2.4 Overall performance

Tables 1 and 2 show the per-residue and per-protein performance of our method AUCpreD and the others on the CASP and CAMEO data, respectively. In terms of Mcc, AUCp and AUC on CASP, our profile-based AUCpreDp greatly outperforms the others including the consensus-based predictor PrDOS-CNF, and the template-based predictor DisoPred3T. DisoPred3A is even worse than AUCpreDp (our profile-free version) in terms of AUC, Acc and (average) per-protein Mcc.

Evaluated by Mcc, AUCp and AUC on CAMEO, AUCpreDp outperforms all the others including DisoPred3T. The P-values listed in both tables show that in terms of per-protein Mcc (i.e. Mcc calculated on each protein and then averaged), the difference between AUCpreDp and the others is statistically significant. Compared to our method AUCpreDp, DisoPred3T performs slightly better on CASP than CAMEO maybe because DisoPred3 uses a template database built in 2014, which contains proteins similar to many CASP targets, but not to the recent CAMEO targets. Note that Espritz has the best sensitivity among all the methods, but at the cost of specificity.

Although DeepCNF-D and DNdisorder also employ deep learning to disorder prediction, especially DeepCNF-D that applies the same model architecture as AUCpreD, their performance is not as good as AUCpreD in terms of Mcc, AUCp and AUC. In general, the performance of the three methods is in the order of DNdisorder < DeepCNF-D < AUCpreD, which illustrates the benefit of our method from the convolutional deep architecture, as well as the maximal-AUC training algorithm.

**Performance on long, internal and external disorder regions.** We evaluated the performance of all the predictors on long, internal,

### Table 1. Per-residue and per-protein performance on the CASP targets

| Methods          | Acc | Sens | Spec | Mcc | AUCp | AUC | Per-protein |
|------------------|-----|------|------|-----|------|-----|-------------|
|                  |     |      |      |     |      |     |             |
| AUCpreDp         | 0.76| 0.53 | 0.98 | 0.55| 0.61 | 0.89|             |
| DeepCNF-Dp       | 0.74| 0.51 | 0.97 | 0.48| 0.49 | 0.86|             |
| Espritzp         | 0.72| 0.52 | 0.92 | 0.38| 0.40 | 0.81|             |
| DNdisorder       | 0.74| 0.59 | 0.89 | 0.37| 0.44 | 0.83|             |
| DisoPred3A       | 0.67| 0.36 | 0.98 | 0.47| 0.50 | 0.84|             |
| DisoPred3T       | 0.68| 0.39 | 0.99 | 0.54| 0.58 | 0.85|             |
| PrDOS-CNF        | 0.71| 0.44 | 0.98 | 0.53| 0.58 | 0.86|             |
| AUCpreD          | 0.72| 0.47 | 0.97 | 0.51| 0.56 | 0.86|             |
| DeepCNF-D        | 0.71| 0.46 | 0.96 | 0.42| 0.45 | 0.82|             |
| Espritzp         | 0.75| 0.60 | 0.89 | 0.38| 0.43 | 0.83|             |
| IUpredS,L        | 0.66| 0.36 | 0.95 | 0.33| 0.35 | 0.68|             |
| IUpredS,L        | 0.57| 0.18 | 0.95 | 0.16| 0.20 | 0.59|             |

P-value is calculated between AUCpreDp and the other methods in terms of per-protein Mcc. The best values are shown in bold. See text for more explanation.
Table 2. Per-residue and per-protein performance on the CAMEO targets

| Methods            | Per-residue |             |             |             |             | Per-protein |             |             |             |
|-------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                   | Acc         | Sens        | Spec        | Mcc         | AUCpr       | AUC         | Acc         | Sens        | Spec        | Mcc         |
| Predictors using sequence profile or template |             |             |             |             |             |             |             |             |             |             |
| AUCpreDp          | 0.75        | 0.53        | 0.97        | 0.49        | 0.55        | 0.88        | 0.76        | 0.58        | 0.95        | 0.48        | N/A         |
| DeepCNF-Dp        | 0.73        | 0.51        | 0.95        | 0.43        | 0.47        | 0.84        | 0.74        | 0.53        | 0.95        | 0.44        | 2.9E–04     |
| Espritzβ          | 0.74        | 0.58        | 0.89        | 0.38        | 0.44        | 0.80        | 0.74        | 0.66        | 0.82        | 0.37        | 3.0E–11     |
| DNdisorder        | 0.73        | 0.59        | 0.88        | 0.36        | 0.43        | 0.79        | 0.73        | 0.67        | 0.78        | 0.34        | 6.8E–14     |
| DisoPred3A        | 0.70        | 0.46        | 0.95        | 0.42        | 0.45        | 0.83        | 0.72        | 0.49        | 0.95        | 0.42        | 1.2E–05     |
| DisoPred3T        | 0.70        | 0.43        | 0.96        | 0.44        | 0.47        | 0.83        | 0.72        | 0.47        | 0.96        | 0.43        | 7.3E–05     |
| PrDOS-CNF         | 0.71        | 0.45        | 0.97        | 0.45        | 0.46        | 0.84        | 0.73        | 0.51        | 0.95        | 0.44        | 2.4E–04     |
| IUpredL           | 0.65        | 0.38        | 0.92        | 0.29        | 0.30        | 0.72        | 0.60        | 0.32        | 0.88        | 0.21        | 6.4E–22     |
| IUpredS           | 0.69        | 0.36        | 0.93        | 0.37        | 0.39        | 0.79        | 0.72        | 0.56        | 0.88        | 0.38        | 7.8E–08     |
| Predictors without using sequence profile |             |             |             |             |             |             |             |             |             |             |             |
| AUCpreDp          | 0.72        | 0.48        | 0.95        | 0.45        | 0.48        | 0.84        | 0.74        | 0.55        | 0.93        | 0.45        | 0.45        |
| DeepCNF-Dp        | 0.70        | 0.46        | 0.94        | 0.38        | 0.41        | 0.79        | 0.71        | 0.48        | 0.92        | 0.39        | 3.6E–07     |
| Espritzβ          | 0.74        | 0.61        | 0.87        | 0.37        | 0.43        | 0.80        | 0.74        | 0.68        | 0.79        | 0.35        | 8.1E–11     |
| IUpredL           | 0.69        | 0.46        | 0.93        | 0.37        | 0.39        | 0.79        | 0.72        | 0.56        | 0.88        | 0.38        | 7.8E–08     |
| IUpredS           | 0.65        | 0.38        | 0.92        | 0.29        | 0.30        | 0.72        | 0.60        | 0.32        | 0.88        | 0.21        | 6.4E–22     |

and external disorder regions in Supplemental Material. As shown in Supplemental Tables 1 and 2, on CAMEO targets, our method AUCpreDp greatly exceeds the others in terms of Mcc, AUCpr and AUC. Even without using sequence profile, AUCpreDp is comparable to or outperforms the others in terms of Mcc, AUCpr and AUC.

Performance without sequence profile. Without using sequence profile, our method AUCpreDp also performs very well. In terms of AUC and Mcc on both datasets, AUCpreDp performs even better than the template-based DisoPred3A and DNdisorder. AUCpreDp also has a comparable performance with the template-based DisoPred3T. AUCpreDp performs much better than the predictors without using sequence profile such as IUpred and Espritz in terms of both AUC and Mcc.

Performance on proteins with sparse sequence profile. We further examine the performance of AUCpreDp on proteins with little homologous information. We use Neff (Wang et al., 2011) to measure the amount of homologous information available for a protein sequence, which ranges from 1 to 20. A protein with a small Neff has a sparse sequence profile, i.e. little homologous information; while a protein with a large Neff may have a large amount of homologous information and thus, high-quality sequence profile. Tested on the CASP and CAMEO targets with Neff ≤ 2, the AUC values obtained by AUCpreDp (AUCpreDp) are 0.89 (0.90) and 0.86 (0.86), respectively. The average per-residue Mcc values obtained by AUCpreDp (AUCpreDp) are 0.59 (0.62) and 0.47 (0.48), respectively. In contrast, the AUC values obtained by DisoPred3A (DisoPred3A) are only 0.84 (0.84) and 0.80 (0.79), respectively. The average per-residue Mcc values obtained by DisoPred3A (DisoPred3A) are 0.54 (0.55) and 0.41 (0.41), respectively. These results imply that (1) AUCpreDp (i.e. sequence profile not used) is comparable to the profile-based AUCpreDp on a protein with a sparse sequence profile, and (2) AUCpreDp outperforms the profile-based DisoPred3A at a large margin on proteins with a sparse sequence profile.

2.5 Large-scale prediction on the human proteome
The large-scale prediction on the entire human proteome is an important application of disorder prediction (Walsh et al., 2012). We extract the human protein sequences from UniProt. Since it takes time to generate sequence profile for a very long protein, we use only proteins with <1700 residues, which results in 19 385 human proteins. We use manually curated and/or experimental determined disorder annotations from the MobiDB database (http://mobidb.bio.unipd.it/) as ground truth (i.e. excludes all prediction and consensus from different predictors). Specifically, MobiDB annotates ground truth disorder regions using two information sources: (i) manual curation from DisProt (Sickmeier et al., 2007), and (ii) experimental PDB information (Di Domenico et al., 2012) (see one example in Supplementary Figure S4). In total there are 1 066 878 and 97 815 residues in order and disorder states, respectively.

On this test set, the average running times of predictors without using sequence profile information, i.e. AUCpreDp, Espritzp, and IUpred (long and short), are 5, 5 and 1 s, respectively. For predictors that use sequence profile or template information, say AUCpreDp, Espritzp, DisoPred3, and PrDOS-CNF, the average running times are 38, 32, 45 and 30 min, respectively. We did not run DNdisorder and PrDOS-CNF since they do not provide downloadable version.

When analyzing the whole proteome, it is important to limit the number of false positives to avoid drawing false conclusions on the prevalence of disorder (Walsh et al., 2012). That is, the sensitivity level at a high specificity level is more important than at a low specificity level. As shown in Figure 1, AUCpreDp has the best ROC curve when the specificity level is larger than 0.95. At the specificity level 0.99, AUCpreDp and AUCpreDp have sensitivity around 0.4 and 0.35, respectively, while the other predictors have sensitivity less than 0.25. As shown in Table 3, AUCpreDp exceeds all the others in terms of Mcc, AUCpr, and AUC, whereas AUCpreDp exceeds the others in terms of Mcc and AUC. These results imply that AUCpreDp is a good tool for high-throughput disorder prediction.

When only the human protein sequences sharing less than 25% identity to the training proteins are evaluated, the (AUC, Mcc) values obtained by AUCpreDp, AUCpreDp, DeepCNF-Dp, DeepCNF-Dp, Espritzp, Espritzp, DisoPred3T, DisoPred3T, IUpred-long and IUpred-long are (0.94, 0.72, 0.65), (0.90, 0.63, 0.57), (0.89, 0.59, 0.54), (0.86, 0.55, 0.51), (0.87, 0.57, 0.48), (0.87, 0.56, 0.46), (0.89, 0.50, 0.50), (0.89, 0.41, 0.49), (0.85, 0.54, 0.50) and (0.83, 0.51, 0.45), respectively.

3 Method
3.1 Disorder prediction by deep convolutional neural fields
Definition of disordered residues. Following the definition in (Monastersky et al., 2011), we label a residue as disordered (label 1) if it is in a segment of more than three residues missing atomic
coordinates in the X-ray structure. The other residues are labeled as ordered (label 0).

Deep Convolutional Neural Fields (DeepCNF) for disorder prediction. As shown in Supplemental Figure 1, DeepCNF has two modules: (i) the Conditional Random Fields (CRF) module consisting of the top layer and the label layer, and (ii) the deep convolutional neural network (DCNN) for the b-th neuron at position i of the K-th layer (i.e. the top layer in Supplemental Figure S1) for label a, and W, U, and T are the model parameters to be trained. Specifically, W is the parameter for the neural network, U is the parameter connecting the top layer to the label layer and T is for label correlation.

3.2 Training DeepCNF by maximizing AUC

3.2.1 The AUC function

In a binary classification problem with unbalanced data, the performance of a predictor is usually measured by AUC (Fawcett, 2004). AUC can be derived from ranking of positive/negative examples. With a perfect ranking, all positive examples are ranked higher than negatives and thus, the AUC value is 1. Even a single error in the ranking results in an AUC value less than 1. A random ranking has an AUC value at 0.5, regardless of class distribution. Each pairwise ranking involves one positive sample and one negative sample, so AUC maintains a perfect 50–50% class distribution (Cortes and Mohri, 2004).

Formally, the AUC of a predictor function \(P_o\) on label \(\tau\) is defined as:

\[
AUC(P_o, \tau) = \left( P_o(y_i^+) > P_o(y_i^-) \mid i \in D^+, j \in D^- \right).
\] (4)

Where \(P\) is the probability over all pairs of positive and negative examples, \(P_o(y_i^+)\) is the predicted probability of the label at position i being \(\tau\), \(D^+\) is a set of positive examples with true label \(\tau\) and \(D^-\) is a set of negative examples with true label not being \(\tau\).

We can use the marginal probability at one position to predict its label. The marginal probability is defined as follows.

\[
P_i^+(y_i^+) = \frac{1}{Z(X)} \sum_{y_i^+} [\delta(y_i = \tau) \cdot \exp(P_{1:1}(X, y, \theta))].
\] (5)
Where \( F_{i,1}(X, y, \theta) = \sum_{j} y_j \Phi(X, y, j, \theta) \). The following Wilcoxon–Mann–Whitney statistic (Hanley and McNeil, 1982) is an unbiased estimator of AUC(\( P_{\theta}, \tau \)):

\[
AUC_{\text{MW}}(P_{\theta}, \tau) = \sum_{i \in D} \sum_{j \in D \setminus \{i\}} \delta(P_{\theta}(y'_j | X) > P_{\theta}(y'_i | X)) / |D|^2 - |D|.
\] (6)

Where \( \delta(A > B) \) is an indicator function equaling to 1 if and only if \( A > B \). Summing over all the labels, the overall objective function is \( \sum_{\tau} AUC_{\text{MW}}(P_{\theta}, \tau) \).

### 3.2.2 Polynomial approximation of the AUC function

For a large dataset, the computational cost of AUC by Equation (6) is high. Recently, (Calders and Jaroszewicz, 2007) proposed a polynomial approximation of AUC which can be computed in linear time. The key idea is to approximate the indicator function \( \delta(x > 0) \), where \( x \) represents \( P_{\theta}(y'_j | X) \), by a polynomial Chebyshev approximation. That is, we approximate \( \delta(x > 0) \) by \( \sum_{d=0}^{D} c_d x^d \) where \( d \) is the degree and \( c_d \) is the coefficient of the polynomial (Calders and Jaroszewicz, 2007).

Let \( n_1 = |D|^2 \) and \( n_0 = |D| \). Using the polynomial Chebyshev approximation, we can approximate Equation (6) as follows:

\[
AUC_{\text{MW}}(P_{\theta}, \tau) = \frac{1}{n_0 n_1} \sum_{x=0}^{D} \sum_{y=0}^{D} \delta(P_{\theta}(y'_j | D) > P_{\theta}(y'_i | D)) \eta(P_{\theta}(y'_j | D), \tau). \]
\] (7)

where \( \eta_{ij} = c_d \left( \frac{\mu}{\lambda} \right)^{(i-1)d} \) and \( \mu(P_{\theta}(y'_j | D), \tau) = \sum_{i \in D} P(\tau) P(y'_i) \). For a linear-chain structure with length \( L \), we have \( s(P_{\theta}(y'_j | D), \tau) = \sum_{i=1}^{L-1} \delta_i^+ P(y'_i) \) where \( \delta_i^+ \) is an indicator function. If the true label of the \( i \)-th position equals to \( \tau \), \( \delta_i^+ \) is equal to 1; otherwise 0.

### 3.2.3 Complexity analysis

We derive the gradient of the polynomial approximation of AUC [see Equation (7)] for our DeepCNF model in the Supplemental Material. The time and space complexity of gradient calculation is \( O(d^2 \cdot |D|^2 + L) \) and \( O(|D| \cdot |L|) \), respectively, which is linear when protein length is much larger than the number of labels and the degree of the polynomial Chebyshev approximation.

Since a deep neural network is applied, we may not be able to solve the training problem to global optimum. Instead we use the L-BFGS (Liu and Nocedal, 1989) algorithm to find a suboptimal solution.

### 3.3 Protein features

Our method has two prediction modes depending on if sequence profile is used or not. When sequence profile is used, each residue has residue-, evolution- and structure-related features. Otherwise, only residue-related features are used.

**Residue-related features.** (a) amino acid identity represented as a binary vector of 20 elements; (b) amino acid physico-chemical properties (7 values from Table 1 in (Meiler et al., 2001)); (c) propensity of being at endpoints of a secondary structure segment (11 values from Table 1 in (Duan et al., 2008)); (d) correlated contact potential (40 values from Table 3 in (Tan et al., 2006)); and (e) reduced AAIndex (5 values from Table 2 in (Archleby et al., 2005)). These features may allow for a richer representation of amino acids (Ma and Wang, 2015; Walsh et al., 2012).

**Evolution-related features.** Residues in ordered and disordered state are found to have different substitution patterns due to different selection pressures (Dunker et al., 1998). We use position specific scoring matrix (PSSM) generated by PSI-BLAST (Altschul et al., 1997) to encode the evolutionary information of the sequence under prediction. PSSM has been widely used for disorder prediction (Walsh et al., 2012) and other applications. We also use the HHM profile generated by HHpred (Söding, 2005), which is complementary to PSSM to some degree.

**Structure-level features.** A strong tendency has been observed that disordered regions tend to exist in the coiled state (Becker et al., 2013) and expose to solvent (Oldfield and Dunker, 2014). Therefore, we use local structural features, such as predicted secondary structure and solvent accessibility. Specifically, we predict secondary structure by RaptorX-SS8 (Wang et al., 2011, 2016c) and solvent accessibility by AcconPred (Ma and Wang, 2015) from sequence profile.

### 4 Conclusions

This article has presented a new deep learning method AUCpreD for protein disorder prediction. AUCpreD distinguishes itself from the others in that it applies a deep probabilistic graphical model DeepCNF to model complex sequence–structure relationship and directly optimizes the AUC measure to deal with the imbalanced distribution of disordered and ordered residues. DeepCNF allows us to model complex sequence–disorder relationship by a deep hierarchical architecture, and exploit interdependency between adjacent order/disorder states. Experimental results show that AUCpreD performs much better than the state-of-the-art methods of the same category in terms of AUC, AUC\(^\text{MW}\) and Mcc. On long disordered regions and terminal/internal regions, AUCpreD also performs the best. Even without using sequence profile, AUCpreD still compares favorably to or outperforms the methods that use sequence profile or even protein templates.

Our prediction method can also be integrated into a consensus method to further improve disorder prediction (Deng et al., 2012). It is also possible for AUCpreD to incorporate template information to further improve prediction accuracy. Instead of using linear-chain CRF, we may model a protein by Markov Random Fields (MRF) which can capture long-range residue interactions (Xu et al., 2015). As suggested in (Schlessinger et al., 2007), the predicted residue-residue contact information (Ma et al., 2015; Wang et al., 2016b) could further contribute to disorder prediction under the MRF model.

In addition to disorder prediction, our maximum-AUC training algorithm could be applied to many sequence labelling problems with imbalanced label distributions (He and Garcia, 2009). For example, in post-translation modification (PTM) site prediction, the phosphorylation and methylation sites occur much less often than normal residues (Blom et al., 2004). In eight-state secondary structure prediction, curvature region (state S), beta loop (state T) and irregular loop (state L) are also very rare (Wang et al., 2011, 2016c). We believe that our AUC maximization training method can help with these problems (Wang et al., 2016c).

### Acknowledgements

The authors are also grateful to the computing power provided by the UChicago Beagle and RCC allocations.
Funding
This study was supported by the National Institutes of Health (R01GM0897532 to J.X.) and National Science Foundation (DBI-0960390 to J.X.).

Conflict of Interest: none declared.

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.

Atchley,W.R. et al. (2003) Solving the protein sequence metric problem. Proc. Natl. Acad. Sci. USA, 102, 6395–6400.

Becker,J. et al. (2013) On the encoding of proteins for disordered regions prediction. PLoS One, 8, e82252.

Blok,N. et al. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics, 4, 1633–1649.

Boeckmann,B. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TREMBL in 2003. Nucleic Acids Res., 31, 365–370.

Calder,S. and Jaroszewicz,S. (2007) Efficient AUC optimization for classification. Knowl. Discov. Data.: PKDD 2007., Springer, Berlin, pp. 42–53.

Cortes,C. and Mohri,M. (2004) AUC optimization vs. error rate minimization. Adv. Neural Inform. Process. Syst., 16, 313–320.

Davis,J. and Goadrich,M. (2006) The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, ACM, New York, pp. 233–240.

Deng,X. et al. (2012) A comprehensive overview of computational protein disorder prediction methods. Mol. Biostyst., 8, 114–121.

Deng,X. et al. (2015) An overview of practical applications of protein disorder prediction and drive for faster, more accurate predictions. Int. J. Mol. Sci., 16, 15384–15404.

Di Domenico,T. et al. (2012) MobIDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics, 28, 2080–2081.

Dostányi,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.

Duan,M. et al. (2008) Position-specific residue preference features around the ends of helices and strands and a novel strategy for the prediction of secondary structures. Protein. Sci., 17, 1505–1512.

Dunker,A. et al. (1998) Protein disorder and the evolution of molecular recognition: theory, predictions and observations. Pac. Symp. Biocomput., 473–484.

Eickholt,J. and Cheng,J. (2013) DNdisorder: predicting protein disorder using coupling analysis and supervised learning. Bioinformatics, 31, 3506–3513.

McCullagh,L.-J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics, 24, 1798–1804.

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

Monastyrskyy,B. et al. (2011) Evaluation of disorder predictions in CASP9. Prot. Struct. Funct. Bioinform., 79, 107–118.

Monastyrskyy,B. et al. (2014) Assessment of protein disorder region predictions in CASP10. Prot. Struct. Funct. Bioinform., 82, 127–137.

Narasimhan,H. and Agarwal,S. (2013) A structural [SVM] based approach for optimizing partial AUC. In: Proceedings of the 30th International Conference on Machine Learning, pp. 516–524.

Nguyen Ba,A.N. et al. (2012) Proteome-wide discovery of evolutionary conserved sequences in disordered regions. Sci Signal., 5, res1.

Oldfield,C.J. and Dunker,A.K. (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. Annu. Rev. Biochem., 83, 553–584.

Peng,K. et al. (2006) Length-dependent prediction of protein intrinsic disorder. Bioinformatics, 22, 208–210.

Prlić,A. et al. (2005) FoldIndex®: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics, 21, 3435–3438.

Roche,D.B. et al. (2011) The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. Nucleic Acids Res., 39, W171–W176.

Romero,P. et al. (1998) Thousands of proteins likely to have long disordered regions. Pac. Symp. Biocomput., 437–448.

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

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

Söding,J. (2005) Protein homology detection by HMM–HMM comparison. Bioinformatics, 21, 951–960.

Tan,Y.H. et al. (2006) Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. Proteins Struct. Funct. Bioinform., 64, 587–600.

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

Wang,L. and Sauer,U.H. (2008) OnD-CRF: predicting order and disorder in proteins conditional random fields. Bioinformatics, 24, 1401–1402.

Wang,S. et al. (2015) DeepCIF-D: predicting protein ordered/disordered regions by weighted deep convolutional neural fields. Int. J. Mol. Sci., 16, 17315–17330.
Wang, S. et al. (2016a) Protein secondary structure prediction using deep convolutional neural fields. *Sci. Rep.*, 6, 18962.

Wang, S. et al. (2016b) CoinFold: a web server for protein contact prediction and contact-assisted protein folding. *Nucleic Acids Res.*, 44, W361–W366.

Wang, S. et al. (2016c) RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res.*, 44, W430–W435.

Wang, Z. et al. (2011) Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics*, 11, 3786–3792.

Ward, J. J. et al. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, 20, 2138–2139.

Xu, J. et al. (2013) Protein Homology Detection through Alignment of Markov Random Fields Using MRFalign. Springer, Berlin.

Xue, B. et al. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta (BBA) Proteins Proteom.*, 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.

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.