Journal of Biomolecular Structure and Dynamics, 2013
http://dx.doi.org/10.1080/07391102.2013.775969

Taylor & Francis
Taylor & Francis Group

Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus

Xiao Fan and Lukasz Kurgan*

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada

Communicated by Ramaswamy H. Sarma

(*Received 23 November 2012; final version received 11 February 2013*)

Availability of computational methods that predict disorder from protein sequences fuels rapid advancements in the protein disorder field. The most accurate predictions are usually obtained with consensus-based approaches. However, their design is performed in an ad hoc manner. We perform first-of-its-kind rational design where we empirically search for an optimal mixture of base methods, selected out of a comprehensive set of 20 modern predictors, and we explore several novel ways to build the consensus. Our method for the prediction of disorder based on Consensus of Predictors (disCoP) combines seven base methods, utilizes custom-designed set of selected 11 features that aggregate base predictions over a sequence window and uses binomial deviance loss-based regression to implement the consensus. Empirical tests performed on an independent benchmark set (with low-sequence similarity compared with proteins used to design disCoP), shows that disCoP provides statistically significant improvements with at least moderate magnitude of differences. disCoP outperforms 28 predictors, including other state-of-the-art consensuses, and achieves Area Under the ROC Curve of .85 and Matthews Correlation Coefficient of .5 compared with .83 and .48 of the best considered approach, respectively. Our consensus provides high rate of correct disorder predictions, especially when low rate of incorrect disorder predictions is desired. We are first to comprehensively assess predictions in the context of several functional types of disorder and we demonstrate that disCoP generates accurate predictions of disorder located at the post-translational modification sites (in particular phosphorylation sites) and in autoregulatory and flexible linker regions. disCoP is available at http://biomine.ece.ualberta.ca/disCoP/.

Keywords: intrinsic disorder; high-throughput prediction; consensus; meta-prediction; linear regression

1. Introduction

Intrinsic disorder, which is defined as a lack of stable structure in a native conformation, is prevalent in proteins. According to some estimates, about 10% of proteins are fully disordered (Tompa, 2002), about 33% of eukaryotic proteins have at least one long (>30 consecutive amino acids) intrinsically disordered region (Ward, Sodhi, McGuffin, Buxton, & Jones, 2004), and a recent study of 67 eukaryotic proteomes demonstrates that on average over 30% of residues are disordered (Xue, Dunker, & Uversky, 2012). Intrinsically disordered regions were shown to implement various important cellular functions, including signal transduction, cell division, transcription and translation. (Tompa, 2002). Recent works also show that disorder is implicated in various human diseases including cancer, cardiovascular, neurodegenerative, and genetic diseases (Midic, Oldfield, Dunker, Obradovic, & Uversky, 2009; Raychaudhuri, Dey, Bhattacharyya, & Mukhopadhyay, 2009; Uversky, Oldfield, & Dunker, 2008)). In spite of these advances, our knowledge and understanding of the functional roles of intrinsic disorder is limited, particularly compared with the structured proteins. This combined with the above-mentioned importance of the disorder strongly motivates further research in this area. To date, only a relatively small number of about 700 proteins is annotated with the intrinsic disorder (Sickmeier et al., 2007), and this limits further studies. Several experimental methods can be used to characterize disorders, such as NMR spectroscopy, circular dichroism spectroscopy, protease digestion, X-ray crystallography, etc. However, they are characterized by a relatively low throughput that cannot accommodate for the rapidly accumulating number of known protein chains. This spurred development of high-throughput computational methods that predict intrinsic disorders from protein sequences.

The computational predictors are categorized into four groups (Peng & Kurgan, 2012a): (1) methods based on relative propensity of amino acids including GlobPlot

*Corresponding author. Email: lkurgan@ece.ualberta.ca

(Linding, Russell, Neduva, & Gibson, 2003), IUPred (Dosztányi, Csizmok, Tompa, & Simon, 2005), FoldIndex (Prilusky et al., 2005), Ucon (Schlessinger, Punta, & Rost, 2007); (2) methods that utilize machine learning, such as DisEMBL (Linding et al., 2003), DISOPRED (Jones & Ward, 2003), DISOPRED2 (Ward, McGuffin, Bryson, Buxton, & Jones, 2004), DISpro (Cheng, Sweredoski, & Baldi, 2005), RONN (Yang, Thomson, McNeil, & Esnouf, 2005), PONDR family of predictors (Obradovic, Peng, Vucetic, Radivojac, & Dunker, 2005; Obradovic et al., 2003; Peng, Radivojac, Vucetic, Dunker, & Obradovic, 2006; Peng et al., 2005; Romero et al., 2001; Vucetic, Brown, Dunker, & Obradovic, 2003), IUP (Yang & Yang, 2006), Spritz (Vullo, Bortolami, Pollastri, & Tosatto, 2006), ProfBval (Schlessinger, Yachdav, & Rost, 2006), DisPSSMP (Su, Chen, & Ou, 2006), IDPA (Su, Chen, & Hsu, 2007), POODLE family of predictors (Hirose, Shimizu, Kanai, Kuroda, & Noguchi, 2007; Shimizu, Muraoka, Hirose, Tomii, & Noguchi, 2007), NORSnet (Schlessinger, Liu, & Rost, 2007), OnD-CRFs (Wang & Sauer, 2008), and most recently ESpritz (Walsh, Martin, Di Domenico, & Tosatto, 2012) and SPINE-D (Zhang et al., 2012); (3) methods based on a consensus-based approach that combine multiple disorder predictors including PreDisorder (Deng, Eickholt, & Cheng, 2009), metaPrDOS (Ishida & Kinoshita, 2008), MD (Schlessinger, Punta, Yachdav, Kajan, & Rost, 2009), PONDR-FIT (Xue, Dunbrack, Williams, Dunker, & Uversky, 2010), MFDp (Mizianty et al., 2010), CSpritz (Walsh et al., 2011), and MetaDisorder (Kozlowski & Bujnicki, 2012); and (4) methods based on analysis of predicted 3D structural models, such as PrDOS (Ishida & Kinoshita, 2007) and DISOclust (McGuffin, 2008). We note that consensus-based predictors are also called ensembles and meta-predictors, and we use these terms interchangeably. The interest in computational disorder prediction continues (Monastyrskyy, Fidelis, Moult, Tramontano, & Kryshtafovych, 2011; Noivirt-Brik, Prilusky, & Sussman, 2009) and new approaches that would provide improved predictive quality are needed.

An interesting characteristic of the disorder predictors is the fact that they are quite diverse in their design and objectives. Firstly, the disorder is annotated using a few different approaches and some predictors aim to find certain corresponding types/flavors of the disorder (Vucetic et al., 2003). For example, DisEMBL (Linding et al., 2003) includes three versions, one that aims to predict loops/coils, as defined by DSSP (Kabsch & Sander, 1983), another to predict "hot loops" defined based on B-factors, and the third that focuses on disorder defined based on missing coordinates in X-ray structures. Similarly, ESpritz (Walsh et al., 2012) is composed of three versions, one for the disorder defined using the missing coordinates, another using the disorder annotations for the DisProt database (Sickmeier et al., 2007), and the third based on the annotations extracted from NMR structures (Martin, Walsh, & Tosatto, 2010). A few studies have shown that predictors trained using a particular type of annotations are usually less accurate in predicting other types of disordered regions (Oldfield et al., 2005; Schlessinger, Liu, et al., 2007; Vucetic et al., 2003). This is expected and it suggests that a complete prediction of all types of disorder requires combining these methods together. Secondly, individual predictors use different information generated from the input sequences. This information includes position-specific scoring matrix (PSSM) profiles, sequence complexity, hydrophobicity, net-charge, interaction energy, predicted secondary structure, intrachain contacts, solvent accessibility, flexibility, etc. (Dosztanyi, Csizmok, Tompa, & Simon, 2005; Dosztányi, Mészáros, & Simon, 2010; Linding, Russell, et al., 2003; Uversky, Gillespie, & Fink, 2000). Thirdly, these methods utilize a wide range of models that calculate the predictions, ranging from simple scoring function to more sophisticated models, like Support Vector Machines (SVMs), conditional random fields, and various types of Neural Networks (NNs) (He et al., 2009; Peng & Kurgan, 2012a). These three observations suggest the results generated by these methods also likely vary; certain methods may perform particularly well for certain type of disorder or specific types of input chains. This diversity motivates development of consensus-based predictors that combine multiple methods to improve overall, across all types of disorder and all types of proteins, predictive quality. These ensembles should improve predictive performance given that the input methods are characterized by orthogonality (Brown, Wyatt, Harris, & Yao, 2005). Earlier research in disorder prediction indeed demonstrates that a mixture of orthogonal methods provides improvements (Oldfield et al., 2005), and a more recent study further supports this finding (Peng & Kurgan, 2012b).

A few meta-predictors of intrinsic disorder were already developed. PreDisorder averages the outputs of the methods that were evaluated in CASP8, except for a few inaccurate predictors (Deng et al., 2009). MD employs four predictors and utilizes a NN to combine them (Schlessinger et al., 2009). PONDR-FIT also adopts a NN to combine predictions generated by six methods (Xue et al., 2010). metaPrDOS (Ishida & Kinoshita, 2008) and MFDp (Mizianty et al., 2010) use SVM to implement the ensemble composed on seven and three disorder predictors, respectively. Finally, the two newest meta-predictors perform relatively simple averaging of predictions provided by three methods in the case of CSpritz (Walsh et al., 2011) and weighted averaging of thirteen methods in the case of MetaDisorder (Kozlowski & Bujnicki, 2012). We note that the methods that compose the consensus are selected in a relatively ad hoc fashion, based on an arbitrary inclusion of the

largest possible number of methods (Deng et al., 2009; Kozlowski & Bujnicki, 2012), their availability to the authors and predictive quality measured per individual method (Ishida & Kinoshita, 2008), and their availability to the authors combined with an argument that individual methods differ in their aims or design (Mizianty et al., 2010; Schlessinger et al., 2009; Walsh et al., 2011; Xue et al., 2010). Moreover, the authors usually consider a relatively narrow set of predictors (Ishida & Kinoshita, 2008; Mizianty et al., 2010; Schlessinger et al., 2009; Walsh et al., 2011; Xue et al., 2010), or otherwise they do not use a rational approach to select a subset of methods that positively contribute to the final result (Deng et al., 2009; Kozlowski & Bujnicki, 2012); the latter may lead to the inclusion of methods that negatively affect the predictive performance.

To this end, we propose a new consensus-based disorder predictor that implements three novel ideas. First, we utilize a first-of-its-kind empirical/rational approach to select the best performing set of predictors out of a comprehensive set of 20 input methods. Second, motivated by the recent success of the averaging-based consensuses (Kozlowski & Bujnicki, 2012; Walsh et al., 2011), we explore more advanced averaging-based approaches to combine the predictions. We empirically compare simple averaging and two more sophisticated solutions that use different types of regressions. Third, instead of using raw values predicted by the input methods, we designed a set of features that aggregate the raw values utilizing a sliding window. We perform empirical feature selection to find the best performing features, and the resulting set of 11 aggregation-based features that are generated based on predictions from the selected seven methods are inputted into our regression. Moreover, our consensus is optimized using a new measure that aims to maximize true positive rate, especially when low false-positive rate is desired; this means that our method can be used to generate a conservative set of high-quality disorder predictions. Empirical tests on an independent benchmark set of proteins (that shares low sequence similarity to the proteins used to design our approach) show that our method for prediction of disorder based on Consensus of Predictors (disCoP) offers improved predictive performance. We are also the first to perform comprehensive evaluation of disorder predictions in the context of different functional types of disorder.

2. Materials and methods

2.1. Evaluation criteria

The assessment of the predictors is consistent with the CASP9 experiment (Monastyrskyy et al., 2011) and we also evaluate predicted disorder content. We evaluate predictions at three levels: (1) the binary values that predict whether a given residue is disordered or is not disordered; (1) the real values that quantify probability of disorder for each residue; and (3) the disorder content computed for the entire protein sequence. The binary predictions are assessed by the following three measures:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specifity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

where TP (true positives) and TN (true negatives) are the counts of correctly predicted disordered and structured residues, respectively, and FP (false positives) and FN (false negatives) are the counts of incorrectly predicted disordered and structured residues, respectively. Matthews correlation coefficient (MCC) was recommended in CASP9 (Monastyrskyy et al., 2011); its values range between -1 and 1 with 0 denoting random prediction and higher absolute values denoting more accurate predictions. We do not compute Q2 accuracy that was deemed unsuitable in CASP9 (Monastyrskyy et al., 2011). Moreover, the balanced accuracy, which is defined as average of sensitivity and specificity, can be easily inferred from the two individual measures that we report, and we do not report S_w that was shown to be linearly correlated with the balanced accuracy.

The receiver operating characteristic (ROC) curve, which represents relation between true positive rates (TPR) = TP/(TP + FN) and false-positive rates (FPR) = FP/(FP + TN), is used to evaluate the predicted probability. We compute the Area Under ROC curve (AUC) that ranges between .5 (for a random predictor) and 1 (for a perfect predictor). Since we aim to produce a consensus that focuses on generating high TPRs for the low FPRs (left-hand-side part of the ROC curve), we introduce another measure called *constrained AUC*. Given that the maximum number of false positives (structured residues predicted as disordered) equals to the number of all positives (natively disordered residues), we calculate the constrained AUC using the FPR values with the upper bound = FP/(FP + TN) = P/N, where P is the number of positives and N is the number of negatives (natively ordered residues). In other words, the constrained AUC is the area under the ROC curve constrained to the FPR range between 0 and the upper bound, that is, area where number of false positives will not exceed the number of native positives. The upper

bound is the ratio of native disordered and ordered residues in the TRAINING dataset and equals .299. Since the maximal possible value of the constrained AUC is equal to the upper bound, we normalize it (using min-max normalization) to the [0, 1] range. Higher values of the constrained AUC correspond to better predictive quality for the low FPR range. We use the constrained AUC to optimize the design of our consensus.

The quality of the disorder content prediction is assessed using difference between the predicted and native disorder content (DDC) = (average_predicted_content − average_native_content). The disorder content is defined as a ratio between the number of disordered and all residues in a given dataset of proteins. The DDC measure shows whether a given predictor over-predicts (DDC value is positive) or under-predicts (DDC value is negative) the disorder at the chain level.

2.2. Datasets and evaluation protocols

The proposed disCoP is designed and tested on a large benchmark dataset that was originally developed to validate the MFDp predictor (Mizianty et al., 2010). The sequences were harvested from the PDB and the DisProt databases. After removing sequences that share > 25% identify and four sequences for which some considered methods (NORSnet, Profbval, Ucon and MD) failed to produce predictions, 305 proteins from DisProt and 205 proteins from PDB are included. The chains collected from PDB were annotated following the protocol from CASP (Monastyrskyy et al., 2011). The annotations for chains from the Disprot database were enriched with PDB REMARK 465 annotations following (Sirota et al., 2010). Specifically, the chains taken from Disprot were searched using PSI-BLAST (Altschul et al., 1997) against PDB (3 iterations, *e*-value < .001) and we selected the best hit with sequence identity $\geqslant$ 98% (over aligned region) and alignment coverage $\geqslant$ 90%. Using the alignment, the PDB-based disorder/order annotations were mapped for the aligned region and we kept the original disorder annotations from Disprot. The resulting 510 proteins were divided into equally sized TRAINING (with 255 proteins) and TEST (255 proteins) datasets; we note that the chains in the test set share at most 25% similarity to the chains from the training dataset.

The TRAINING dataset was used to empirically design the disCoP consensus. This was performed using threefold cross-validation to assure that we do not overfit this dataset; we divided the chains such that each of the three folds has similar number of amino acids and fraction of disordered residues. The resulting model was tested and compared with other disorder predictors on the TEST dataset. We generated two additional subsets of the TEST dataset. The first subset called TEST_SHORT includes 234 chains that are shorter that 1000 residues; this was motivated by the fact that the most recent meta-predictor MetaDisorder (Kozlowski & Bujnicki, 2012) cannot predict longer chains. We used this dataset to compare with MetaDisorder, its input predictors, and the other predictors considered in this work. The second subset, called TEST_FUNCTION, is utilized to evaluate disorder predictions when considering different functional types of disorder. The annotation of the types is based on the DisProt database that defines 38 functional subclasses (Dunker, Brown, Lawson, Iakoucheva, & Obradovic, 2002; Sickmeier et al., 2007). The TEST dataset covers 16 functions and we further removed the subclasses with less than five proteins since they would not have enough data for a statistically sound evaluation. Key characteristics of the remaining 6 subclasses in the TEST dataset are shown in Table 1. Using TEST_FUNCTION, we assess the predictions of disorder for the subset of disordered residues that correspond to a given functional subclass. We combine the disordered residues annotated with a given functional subclass with a subset of structured residues from the same chains to maintain the ratio between disordered and ordered residues from the TRAINING dataset. We repeat the selection of the ordered residues 10 times and we report the average (over the 10 repetitions) AUC values.

The TRAINING and TEST datasets are available at http://biomine.ece.ualberta.ca/disCoP/.

Table 1. Summary of the annotations of the six subclasses of disorder in the TEST dataset.

Functional subclass	# Proteins	# DRs	# Disordered residues	Average length of DRs	Notes
Protein–protein binding	60	79	8724	110.43	Binds to protein partner(s)
Substrate or ligand binding	28	43	2906	67.58	Binds to substrate(s) and/or ligand(s)
Flexible linkers or spacers	16	24	576	24.00	Provides separation and permits movement between adjacent domains
Protein-DNA binding	11	15	917	61.13	Binds to DNA
Phosphorylation	10	10	1857	185.70	Guides the addition of a phosphate
Autoregulatory	6	9	631	72.5	Involved in the regulation of protein function/activity

Note: DR stands for disordered region. The subclasses are sorted by the number of proteins that include them.

We also assess significance of the improvements offered by the disCoP consensus. We compare disCoP with any of the other considered predictors using ten repetitions of a randomly select subset of 50% of proteins in the TEST (or TEST_SHORT) datasets. The results for a given pair of predictors are compared using the Student's *t*-test, if distributions were normal, or with the Mann–Whitney U-test, if not. Distribution type was verified using the Anderson–Darling test using the *p*-value of .05.

2.3. Considered disorder predictors

We consider a comprehensive set of methods that are accessible to the end users as either standalone software or a web server, which are published on a reputable peer-reviewed scientific venue or were evaluated in a CASP experiment, and which are not restricted to short chains (e.g. like MetaDisorder can only predict chains with less than 1000 residues). These methods cover the four types of disorder predictors and include:

- two relative propensity-based methods: IUPred (Dosztanyi et al., 2005) and Ucon (Schlessinger, Punta, et al., 2007)
- eight machine learning-based predictors: DISOPRED2 (Ward, McGuffin, et al., 2004), DRIPPRED (MacCallum, 2004), RONN (Yang et al., 2005), VSL2B (Peng et al., 2006), ProfBval (Schlessinger et al., 2006), NORSnet (Schlessinger, Liu, et al., 2007), ESpritz (Walsh et al., 2012), and SPINE-D (Zhang et al., 2012)
- four consensus-based methods: MD (Schlessinger et al., 2009), MFDp (Mizianty et al., 2010), PONDR-FIT (Xue et al., 2010), and CSpritz (Walsh et al., 2011)
- two 3D prediction–based predictors: PrDos (Ishida & Kinoshita, 2007) and DISOclust (McGuffin, 2008).

We used IUPred and CSpritz in both of their versions, one for prediction of short and the other for long disordered segment. The ESpritz method includes three versions, which address three types of disorder annotations: based on missing coordinates in X-ray structures, using annotations from DisProt and based on NMR structures (Martin et al., 2010). Including the different versions of IUPred, CSpritz, and ESpritz, we consider total of 20 methods; see Supplementary Table S1 and Supplementary Figure S1.

2.4. Selection of consensus model

We consider three models to implement the consensus, including a simple average (SA) and two linear regression-based models.

2.4.1. Simple average

The output of the consensus is computed as an arithmetic average of the probabilities generated by the input methods. Similar approach was utilized in PreDisorder (Deng et al., 2009), in one of the designs considered in MD (Schlessinger et al., 2009), and in CSpritz (Walsh et al., 2011). The main advantages of this approach are simplicity and speed. However, it does not consider predictive quality and orthogonality of the input methods, as it treats all of them equally. An improved strategy based on a weighted average was recently used in MetaDisorder (Kozlowski & Bujnicki, 2012), where weight values were optimized by a genetic algorithm.

2.4.2. Regression-based consensuses

We use an empirical approach to generate a weighted average. For a given set of predictors, we use regression that fits the consensus-based predictions into the native disorder annotations (based on a training dataset) to calculate the weights. This way, the resulting weights should accommodate for the orthogonality of the input methods. Given a set of real-valued and binary-valued outputs (features) obtained from the input (base) predictors for each residue $X \in R^{t \times n}$ and the target native binary annotation of disorder $y \in \{-1, 1\}^{t \times 1}$, a general form of regression is

$$w = arg\min_{w}(L(f(X, w)), y) + \alpha R(w))$$

where t is the number of residues, n is the number of features, f is a mapping function that transforms X and w to $\hat{y} = f(X, w)$ which approximates y, L is a loss function that is used to penalize the difference between y and $\hat{y}$, R is a regularizer that smoothes the output values, and α is a scalar (set to a default value of .01) to adjust trade-off between the loss and regularizer. The output generated by the corresponding consensus is $\hat{y} = \text{sign}(x^{\mathrm{T}}w - b)$, where the sign corresponds to the two different outcomes (disordered vs. structured residues).

Since the number of input data points (residues) is relatively large, complex, and nonlinear mapping functions that are computationally intensive cannot be utilized. We evaluate two mapping functions: squared error, which is one of the most commonly used loss function, and binomial deviance. In contrast to the squared error, the binomial deviance assumes larger penalty when the difference between y and $\hat{y}$ is larger. These two optimization criteria are defined as follows:

$$\text{Squared error} \quad \min_{w}(\|Xw - y\|_2^2 + \|\alpha w\|_2^2)$$

Binomial deviance

$$\min_{w,b}(\| \log_2(1+2)^{-y(Xw-b)}\|_1 + \|\alpha w\|_2^2)$$

where $L2$ norm is applied as the regularizer in order to solve a convex problem.

The outputs generated by the input predictors are combined using a weighted average where the weights are estimated with the least square and minimum binomial deviance loss regressions (BDR).

The overall procedure to find empirically best performing consensus is as follows. First, we try all combinations of consensuses with k input/base predictors selected out of the 20 considered method that are combined using each of three models, including SA, least square regression (LSR), and BDR. For each k and model type, we select the consensus that generates the highest value of constrained AUC based on threefold cross-validation on the TRAINING dataset. The value of k is initiated with 2 and incremented by 1 as long as the corresponding best consensus improves value of the constrained AUC. The constrained AUCs for the three models and different values of k are shown in Table 2.

Table 2 demonstrates that a SA improves the predictive quality when compared with the best individual method, ESpritz-Disprot, which confirms the observations in (Deng et al., 2009; Walsh et al., 2012). However, the improvements have relatively small magnitude and they decrease/disappear when combining more

Table 2. Constrained AUC obtained for the three consensus models: SA, LSR, and BDR, for different values of k. The results are based on threefold cross-validation on the TRAINING dataset and we select consensuses with the largest value of the constrained AUC. k denotes the number of predictors selected from the considered 20 methods. The first row shows the constrained AUC value of the best performing individual predictor, ESpritz-Disprot. The best combination for each model is shown in bold font. The best overall consensus that uses BDR model and $k=8$ is compared with two other consensuses of 8 predictors: one with the top 8 performing (based on the cross-validated constrained AUC on the TRAINING dataset) methods and the other with the most recently published 8 methods.

k # predictors	k selected predictors with the highest value of constrained AUC based on threefold cross-validation on the TRAINING dataset	Constrained AUC
1	ESpritz-Disprot	.5454
SA model		
2	**ESpritz-Disprot, SPINE-D**	.5729
3	ESpritz-Disprot, SPINE-D, CSpritz-long	.5709
4	ESpritz-Disprot, SPINE-D, CSpritz-long, CSpritz-short	.5648
5	ESpritz-Disprot, SPINE-D, CSpritz-long, CSpritz-short, IUPRED-short	.5554
LSR model		
2	ESpritz-Disprot, CSpritz-long	.5774
3	ESpritz-Disprot, CSpritz-long, Profbval	.5818
4	ESpritz-Disprot, CSpritz-long, MD, SPINE-D	.5891
5	ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2	.5957
6	ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2, ESpritz-Xray	.5992
7	ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2, ESpritz-Xray, ESpritz-NMR	.6028
8	**ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2, ESpritz-Xray, ESpritz-NMR, DISOclust**	.6053
9	ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2, ESpritz-Xray, ESpritz-NMR, DISOclust, DRIP-PRED	.5938
BDR model		
2	ESpritz-Disprot, CSpritz-long	.5780
3	ESpritz-Disprot, CSpritz-long, MD	.5825
4	ESpritz-Disprot, CSpritz-long, MD, SPINE-D	.5924
5	ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2	.5992
6	ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2, DISOclust	.6036
7	ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2, DISOclust, ESpritz-Xray	.6068
8	**ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2, DISOclust, ESpritz-Xray, ESpritz-NMR**	.6099
8 top constrained AUC	ESpritz-Disprot, CSpritz-long, MD, SPINE-D, MFDp, CSpritz-short, DISOPRED2, VSL2B	.6009
8 latest predictors	ESpritz-Disprot, ESpritz-Xray, ESpritz-NMR, SPINE-D, CSpritz-long, CSpritz-short, MFDp, PONDR-FIT	.5941
9	ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2, DISOclust, ESpritz-Xray, ESpritz-NMR, VSL2B	.5983

predictors. As expected, use of the regression leads to larger improvements. Our empirical results indicate that BDR provides better results when compared with the LSR, although the differences are relatively small. Both best-performing, over all values of k, regression-based consensuses include the same eight base predictors. We note that these eight selected methods are not the top eight best-performing individual predictors. Table 2 compares the BDR-based consensus of the selected eight methods with the BDR-based consensuses of two other heuristically selected sets of eight methods: one with the top eight performing (based on the cross-validated constrained AUC on the TRAINING dataset) methods and the other with the most recently published eight methods. Although these two consensuses also offer improvements, their magnitude is smaller than the magnitude of the improvements provided by our, empirically selected set of the eight predictors. This demonstrates the advantage of the empirical design.

2.5. Feature generation and selection

Motivated by the fact that disordered residues are usually grouped into segments and the successful use of segment-based features generated from predicted disorder in the MFDp method (Mizianty et al., 2010), we further improve our consensus by considering aggregation of the outputs generated by the selected eight input predictors. The aggregated input predictions, using both binary and real-valued predictions, constitute a set of features that are inputted into the BDR model. Moreover, based on the success of methods that separate predictions of short- and long-disordered segments (Hirose et al., 2007; Mizianty et al., 2010; Peng et al., 2005; Zhang et al., 2012), we derive ternary predictions from the binary predictions, where the predicted ordered residues are denoted by 0, predicted disordered residues in short segments (below 30 consecutive amino acids) as .5, and predicted disordered residues in long segments (30 or more consecutive amino acids) as 1. As a result, we aggregate three outputs for each of the eight selected input predictors: binary predictions, ternary predictions, and real-valued probabilities. The final set of considered features includes binary, ternary, and real-valued predictions for a given predicted residue, the means and medians of the real-valued predictions using windows of different sizes that are centered on the predicted residue, and the disorder content (fraction of disordered residues) and average value of the ternary predictions using the same windows. The corresponding 123 features are summarized in Table 3.

Given the large number of considered features (8 predictors $\times$ 123 features = 984 features) and the fact that they are redundant/correlated (e.g. features generated from the same type of the output and different window sizes), we perform a simple empirical, two-step feature selection to find a well-performing subset of the features. In the first step, the features are sorted by their constrained AUCs, which are calculated using the TRAINING dataset. Next, the sorted features are divided into groups, where each groups corresponds to a given predictor and a given type of features, including probability-derived features, binary and ternary predictions, and features aggregated from the binary and ternary predictions. In the second step, the sorted features are scanned (from best to worst) once and the BDR models are built based on the threefold cross-validation on the TRAINING dataset. We start with the model that uses the top performing feature and we add another feature to the current feature set if it improves the constrained AUC by at least .005 and if another feature from the same type/predictor is not already in the current feature set. We use the cross-validation, limit the selection to one scan and assume that improvements must be larger than .005 to assure that the resulting feature set is relatively small and that it does not overfit the training dataset. As a result, 11 of 984 features are selected and these features utilize seven of the eight predictors: ESpritz-Disprot, CSpritz-long, MD, SPINE-D, DISOPRED2, DISOclust, and ESpritz-Xray. This corresponding consensus obtains the averaged cross-validated constrained AUC of .6411 over the three folds in the TRAINING dataset, compared with the constrained AUC of .6099 (see Table 2) obtained by the consensus that uses all eight methods and the raw (without aggregation) predictions. Supplementary Figure S2 shows the values of the constrained AUC along the second step of the feature selection. The use of 6 features results in a consensus that matches the constrained AUC of the best consensus

Table 3. Description of features generated from the three types of outputs of each input predictor.

Type of the output	Description	Number of features for each predictor
Real-valued probability	Probability for predicted residue	1
	Mean probability in windows with sizes 3, 5, …, 61	30
	Median probability in windows with sizes 3, 5, …, 61	30
Binary prediction	Binary prediction for predicted residue	1
	Disordered content in windows with sizes 3, 5, …, 61	30
Ternary prediction	Ternary prediction for predicted residue	1
	Mean values in windows with sizes 3, 5, …, 61	30

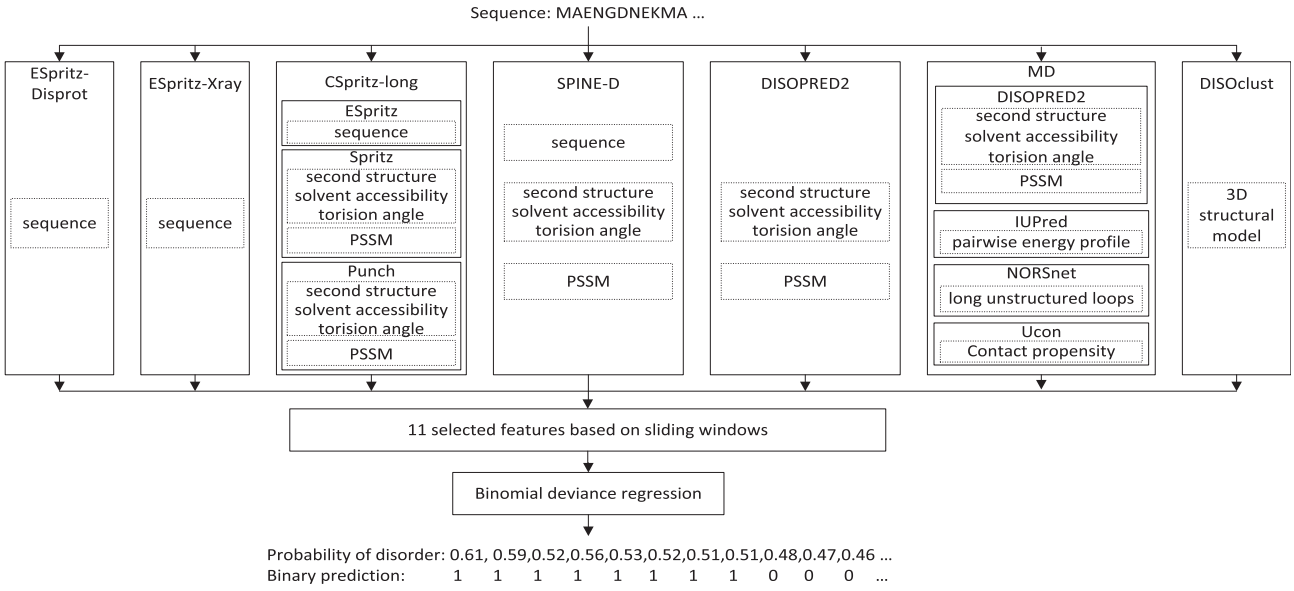

Figure 1. Overall architecture of the disCoP predictor.

that utilizes the raw values, and the 5 additional features improve the constrained AUC by .03.

2.6. disCoP predictor

The outline of the final design of our disCoP predictor is shown in Figure 1. The input sequence is first predicted by the selected seven methods (ESpritz-Disprot, ESpritz-Xray, CSpritz-long, SPINE-D, DISOPRED2, MD, and DISOclust), two of which (CSpritz-long and MD) are also consensuses. Some of these predictors (CSpritz-long, SPINE-D, and DISOPRED2) internally use additional predictions such as secondary structure, solvent accessibility, and torsion angles. The outputs of these predictors are aggregated using the 11 selected features that are inputted into the BDR model; these features are discussed in detail in the Results section. The model outputs the real-valued probabilities that can be binarized using threshold of .5.

2.7. Evaluation of disCoP model on the TRAINING dataset

The disCoP model is designed using 3-fold cross-validation where the corresponding folds are selected such that they have similar size (i.e. similar number of amino acids) and similar fraction of disordered residues. To demonstrate that disCoP preserves the improved predictive quality over less uniformly distributed datasets, we evaluate our method using ten repetitions of the 3-fold cross-validations (where folds are likely uneven in the disorder amount) and jackknife test (also called leave-one-out test) on the TRAINING dataset. These results are compared against the results obtained by the 20 disorder predictors listed in Section 2.3, see Table 4. The disCoP predictions obtained with the repeated cross–validations, and jackknife tests have similar predictive quality. The standard deviations that quantify differences between repetitions of the cross-validation test are relatively small, which means that disCoP generates similar results, irrespective of the selection of the training datasets. The quality of the disCoP predictions measured with the constrained AUC, AUC, and MCC is better than the predictive quality of each of the other 20 predictors. These improvements are statistically significant at p-value $< .01$ (based on a statistical test that considers 10 repetitions of randomly chosen 50% of the proteins from the TRAINING dataset). For instance, based on the jackknife test disCoP improves over the best other method ESpritz-Disprot by .0365 in the constrained AUC (which corresponds to a relative improvement by $100\% \times (.5822-.5457)/.5457 = 6.7\%$), by .0184 in AUC (relative improvement by 2.2%), and by .0384 in MCC (relative improvement by 8.8%). The new consensus predictor also provides the smallest value of DDC (Difference between the predicted and native Disorder Content), which means that it does not over- or under-predict the amount of disorder. The design of the disCoP, which maximizes the constrained AUC, leads to relatively high values of TPR (true positive rate) for a given value of FPR (false positive rate), especially for a low range of FPR values; see the ROC curves in the Supplementary Figure S1.

We also developed a web server that implements disCoP, called disCoP_WS. Since the authors of ESpritz do not allow using their methods in a publicly available consensus (private communication) and CSpritz does not offer a standalone implementation that can be run locally,

Table 4. Comparison of predictive quality on the TRAINING dataset. The disCoP predictor is evaluated based on 10 repetitions of the 3-fold cross-validation (CV) test and jackknife (JK) test (also called leave-one-out test). For the CV test, we provide the average results over the 10 repetitions ± the corresponding standard deviations. Methods that are used as inputs to disCoP are shown in bold font. The highest value for each measure is given in bold font and predictors are sorted in the descending order by their constrained AUC values. The last column provides the DDC (Difference between the predicted and native Disorder Content) values where a positive/negative value means that a given predictor over-/under-predicts the overall amount of disorder in a chain. "Sig" columns show statistical significance of differences measured based on 10 repetitions on randomly chosen 50% of the proteins from TRAINING dataset; we select one subset of randomly chosen 50% of proteins for each of the 10 repetitions of the CV test; +/=/− indicate that disCoP is significantly better/not significantly different/significantly worse than another method at p-value < .01. Two "Sig" columns compare the 20 disorder predictors listed in Section 2.3 against the results of disCoP based on the 10 repetitions of CV and based on the JK test, respectively.

Methods	Constrained AUC	Sig CV	Sig JK	AUC	Sig CV	Sig JK	MCC	Sig CV	Sig JK	Sensitivity	Specificity	DDC
disCoP (CV)	**.5823** ± .0212		=	.8384 ± .0067		=	.4747 ± .0126		=	**.6455** ± .0490	.8498 ± .0306	.03 ± .03
disCoP (JK)	.5822	=		**.8401**	=		**.4757**	=		.603	**.876**	**.00**
ESpritz-Disprot	.5457	+	+	.8217	+	+	.4373	+	+	.721	.772	.11
CSpritz-long	.5401	+	+	.8155	+	+	.4182	+	+	.744	.736	.14
SPINE-D	.5041	+	+	.7917	+	+	.4173	+	+	.783	.704	.18
CSpritz-short	.4993	+	+	.7842	+	+	.3991	+	+	.761	.704	.17
MFDp	.4896	+	+	.7984	+	+	.4215	+	+	.744	.740	.14
MD	.4756	+	+	.7948	+	+	.4006	+	+	.665	.780	.09
IUPred-short	.4650	+	+	.7715	+	+	.3666	+	+	.526	.846	.01
ESpritz-NMR	.4590	+	+	.7584	+	+	.3234	+	+	.779	.605	.25
IUPred-long	.4584	+	+	.7745	+	+	.3833	+	+	.605	.807	.06
ESpritz-Xray	.4569	+	+	.7706	+	+	.3758	+	+	.640	.777	.09
DISOclust	.4558	+	+	.7534	+	+	.3113	+	+	.770	.600	.26
VSL2B	.4530	+	+	.7724	+	+	.3800	+	+	.774	.673	.20
DISOPRED2	.4522	+	+	.7651	+	+	.3779	+	+	.650	.771	.10
PONDR-FIT	.4471	+	+	.7705	+	+	.3848	+	+	.628	.792	.07
PrDos	.4388	+	+	.7644	+	+	.3725	+	+	.598	.804	.06
RONN	.4197	+	+	.7489	+	+	.3394	+	+	.661	.727	.13
Norsnet	.3881	+	+	.7332	+	+	.3208	+	+	.533	.806	.04
DRIP-PRED	.3773	+	+	.7147	+	+	.2963	+	+	.701	.648	.20
Ucon	.3668	+	+	.7219	+	+	.2776	+	+	.550	.756	.08
ProfBval	.3340	+	+	.6792	+	+	.1821	+	+	**.829**	.374	.44

disCoP_WS excludes ESpritz-Disprot, ESpritz-Xray, and CSpritz and uses the remaining four predictors. We repeated the feature selection using the four methods and the abovementioned protocol. As a result, disCoP_WS utilizes nine features generated from the outputs of the four methods, which are inputted into the BDR model; these features are discussed in the Results section. disCoP_WS is freely available at http://biomine.ece.ualberta.ca/disCoP/.

3. Results

3.1. Comparison with other predictors on the TEST datasets

The disCoP and disCoP_WS methods are comprehensively compared with the other considered predictors using the independent (with chains that share < 25% sequence similarity with the chains in the TRAINING dataset) TEST dataset. We collected results from IUPred-long and IUPred-short (Dosztanyi et al., 2005), DRIP-PRED (MacCallum, 2004), DISOPRED2 (Ward, McGuffin, et al., 2004), RONN (Yang et al., 2005), VSL2B (Peng et al., 2006), ProfBval (Schlessinger et al., 2006), Ucon (Schlessinger, Punta, et al., 2007), NORSnet (Schlessinger, Liu, et al., 2007), PrDos (Ishida & Kinoshita, 2007), DISOclust (McGuffin, 2008), MD (Schlessinger et al., 2009), MFDp (Mizianty et al., 2010), PONDR-FIT (Xue et al., 2010), CSpritz-long and CSpritz-short (Walsh et al., 2011), ESpritz-Disprot, ESpritz-Xray, and ESpritz-NMR (Walsh et al., 2012), and SPINE-D (Zhang et al., 2012); see Table 5. These methods include publicly available versions of the top predictors from CASP9 (Monastyrskyy et al., 2011), such as PrDos, DISOPRED, SPINE-D, and MFDp.

Table 5 shows that disCoP outperforms each of the other 20 predictors based on AUC, constrained AUC, and MCC. The improvements are statistically significant, although in some cases the corresponding magnitude is modest, relative to the overall range of values of AUC and MCC. For instance, disCoP improves over the

Table 5. Comparison of predictive quality on the TEST dataset. Methods that are used as inputs to disCoP are shown in bold font. The highest value for each measure is given in bold font and predictors are sorted in the descending order by their constrained AUC values. The last column provides the DDC (Difference between the predicted and native Disorder Content) values where a positive/negative value means that a given predictor over/under-predicts the overall amount of disorder in a chain. "Sig" columns show statistical significance of differences measured based on 10 repetitions on randomly chosen 50% of the proteins from TEST dataset; +/=/− indicate that disCoP is significantly better/not significantly different/significantly worse than another method at p-value < .01. disCoP_WS is a web server version of disCoP that excludes ESpritz and CSpritz predictors (see "disCoP predictor" section for details).

Methods	Constrained AUC	Sig	AUC	Sig	MCC	Sig	Sensitivity	Specifity	DDC
disCoP	**.6249**		**.8498**		**.5037**		.673	.856	.04
ESpritz-Disprot	.6063	+	.8325	+	.4835	+	.709	.819	.07
disCoP_WS	.5819	+	.8254	+	.4759	+	.669	.838	.04
CSpritz-long	.5667	+	.8324	+	.4508	+	.733	.775	.11
MD	.5666	+	.8181	+	.4501	+	.656	.826	.05
SPINE-D	.5427	+	.8093	+	.4229	+	.761	.728	.15
MFDp	.5370	+	.8159	+	.4472	+	.734	.771	.12
CSpritz-short	.5250	+	.8062	+	.4369	+	.767	.737	.15
DISOPRED2	.5077	+	.7877	+	.4230	+	.653	.807	.07
VSL2B	.4995	+	.7909	+	.4068	+	.774	.701	.18
DISOclust	.4971	+	.7784	+	.3638	+	.783	.647	.22
PrDos	.4967	+	.7908	+	.4280	+	.607	.840	.03
ESpritz-Xray	.4952	+	.7851	+	.4047	+	.621	.814	.06
IUPred-short	.4726	+	.7707	+	.3691	+	.498	**.864**	**−.01**
PONDR-FIT	.4693	+	.7841	+	.4089	+	.612	.823	.05
IUPred-long	.4663	+	.7700	+	.3864	+	.558	.840	.02
ESpritz-NMR	.4509	+	.7484	+	.3183	+	.737	.639	.22
Norsnet	.4366	+	.7363	+	.3609	+	.551	.825	.03
RONN	.4353	+	.7599	+	.3644	+	.659	.752	.11
Ucon	.4104	+	.7446	+	.3294	+	.565	.790	.06
DRIP-PRED	.4034	+	.7224	+	.3024	+	.693	.662	.19
ProfBval	.3516	+	.7063	+	.2096	+	**.846**	.389	.44

second best ESpritz-Disprot by .015 in AUC (which corresponds to a relative improvement by 100% × (.8498−.8325)/.8325 = 2.1%), .02 in constrained AUC (relative improvement by 3.1%), and .02 in MCC (relative improvement by 4.2%). The improvements over the third-best CSpritz-long have larger magnitude and equal .06 in constrained AUC (relative improvement by 10.2%) and .05 in MCC (relative improvement by 11.7%), with a modest improvement by .015 in AUC (relative improvement by 2.1%). disCoP also provides one of the smallest values of DDC (Difference between the predicted and native Disorder Content), which means that it provides balanced predictions. In contrast, some methods over-predict the disorder, such as DISOclust and ESpritz-NMR that predict 22% more disorder than the native annotations suggest. Importantly, the results also demonstrate that disCoP can be utilized to provide a conservative subset of high quality disorder predictions. Our predictor has high specificity (generates relatively few false positives) coupled with a high value of constrained AUC. This can be observed using the ROC curves for the top 6 performing methods according to AUC and constrained AUC (including disCoP) given in Figure 2; Supplementary Figure S3 shows the ROC curves for all considered methods. disCoP provides higher values of TPR (true positive rate) given the same value of FPR (false positive rate), especially for the low values of FPR; see inset in Figure 2. This means that residues predicted by disCoP as disordered with high

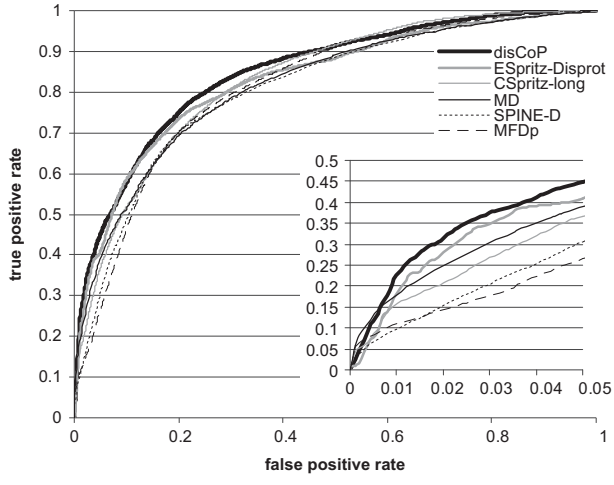

Figure 2. ROC curves of the disCoP and 5 other top-performing (based on both AUC and constrained AUC; see Table 5) predictors on the TEST dataset. The inset in the bottom-right corner shows the ROC curve constrained to the false positive rate (FPR) between 0 and .05.

probabilities are more likely to be in fact disordered, i.e. fewer of them are structured.

As expected, disCoP_WS (the web server version of disCoP) provides lower AUC and MCC when compared to disCoP. This is due to the fact that disCoP_WS does not include the ESpritz (which is banned from inclusion into publicly available meta-predictors) and CSpritz (which has no standalone version) predictors, both of which provide accurate results on the TEST dataset. However, we emphasize that disCoP_WS improves over each of the four methods that are its inputs, including SPINE-D, DISOPRED2, MD, and DISOclust. Compared with the best performing of these four methods MD, disCoP_WS improves AUC and MCC by .01 (relative improvement by 1%) and .025 (relative improvement by 5.7%), respectively. Overall, the web server version provides strong predictive quality with particularly high specificity and low DDC.

We also compare with the most recent meta-predictors, MetaDisorder (Kozlowski & Bujnicki, 2012), on the TEST_SHORT dataset; see Table 6. MetaDisorder allows predictions for chains of up to 1000 residues and TEST_SHORT is a subset of the TEST dataset with 234 of these shorter chains. The comparison covers disCoP, disCoP_WS and 28 other methods. These methods include the same 20 predictors as in Table 5; four versions of MetaDisorder: MetaDisorder, MetaDisorderMD, MetaDisorderMD2, and MetaDisorder3d (Kozlowski & Bujnicki, 2012); and four other predictors that are inputs to MetaDisorder: iPDA (Su et al., 2007), Poodle-S (Shimizu et al., 2007), Poodle-L (Hirose et al., 2007), and DISpro (Cheng et al., 2005). Predictions of the eight

Table 6. Comparison of predictive quality on the TEST_SHORT dataset. TEST_SHORT dataset includes 234 chains from the TEST dataset that are shorter that 1000 residues since the MetaDisorder server can be applied only to these shorter proteins. MetaDisorder and its input predictors, for which predictions were collected using the MetaDisorder server, are denoted using underline. Methods that are used as inputs to disCoP are shown in bold font. The highest value for each measure is given in bold font and predictors are sorted in the descending order by their constrained AUC values. The last column provides the DDC (Difference between the predicted and native Disorder Content) values where a positive/negative value means that a given predictor over/under-predicts the overall amount of disorder in a chain. "Sig" columns show statistical significance of differences measured based on 10 repetitions on randomly chosen 50% of the proteins from TEST dataset; +/=/− indicate that disCoP is significantly better/not significantly different/significantly worse than another method at p-value < .01. disCoP_WS is a web server version of disCoP that excludes ESpritz and CSpritz predictors (see "disCoP predictor" section for details).

Method	Constrained AUC	Sig	AUC	Sig	MCC	Sig	Sensitivity	Specifity	DDC
disCoP	.6822		.8819		.5707		.681	.890	.00
MFDp	.6671	+	.8759	+	.5530	+	.751	.838	.07
disCoP_WS	.6516	+	.8612	+	.5522	=	.682	.878	.01
MetaDisorder	.6406	+	.8614	+	.5485	+	.737	.843	.05
ESpritz-Disprot	.6364	+	.8548	+	.5247	+	.725	.832	.06
CSpritz-long	.6358	+	.8529	+	.5200	+	.784	.789	.06
MetaDisorderMD	.6348	+	.8605	+	.5409	+	.676	.874	.01
MetaDisorderMD2	.6260	+	.8582	+	.5386	+	.711	.852	.04
SPINE-D	.6229	+	.8486	+	.5005	+	.774	.779	.11
MD	.5952	+	.8472	+	.5135	+	.675	.855	.03
iPDA	.5898	+	.8411	+	.4651	+	.504	.916	−.06
DISOCLUST	.5840	+	.8206	+	.4297	+	.797	.692	.18
VSL2B	.5836	+	.8228	+	.4615	+	.770	.747	.13
DISOPRED2	.5774	+	.8281	+	.4975	+	.655	.855	.02
Poodle-L	.5736	+	.8110	+	.4679	+	.644	.840	.03
ESpritz-Xray	.5728	+	.8165	+	.4735	+	.650	.841	.03
DISpro	.5672	+	.8160	+	.4145	+	.329	.967	−.15
Poodle-S	.5602	+	.8280	+	.4780	+	.578	.886	−.02
PONDR-FIT	.5585	+	.8183	+	.4848	+	.632	.860	.01
PrDos	.5571	+	.8325	+	.4949	+	.623	.872	.00
IUPRED-short	.5508	+	.8018	+	.4533	+	.520	.901	−.05
IUPRED-long	.5460	+	.8045	+	.4642	+	.555	.889	−.03
Espritz-NMR	.5292	+	.7785	+	.3654	+	.732	.683	.17
RONN	.5183	+	.7909	+	.4336	+	.670	.796	.07
Norsnet	.4944	+	.7613	+	.4264	+	.530	.879	−.03
CSpritz-short	.4794	+	.7454	+	.3739	+	.610	.788	.10
DRIP-PRED	.4677	+	.7454	+	.3547	+	.680	.717	.13
Ucon	.4655	+	.7670	+	.3716	+	.560	.822	.02
MetaDisorder3d	.4481	+	.7783	+	.3065	+	.330	.918	−.11
ProfBval	.4034	+	.7274	+	.2452	+	.846	.424	.39

latter methods were collected from the MetaDisorder server and they are underlined in Table 6.

The results demonstrate that disCoP outperforms the MetaDisorder consensus by a statistically significant margin. The improvements over the best-performing version called MetaDisorder equal .02 in AUC (relative improvement by 2.4%), .04 in constrained AUC (relative improvement by 6.5%), and .02 in MCC (relative improvement by 4.1%). We note that MetaDisorder includes 13 predictors, compared with seven that are used in disCoP. The likely reasons for the improved predictive quality are the empirical selection of the input predictors and use of aggregated features. We also note that disCoP predicts chains longer than 1000 residues.

3.2. *Evaluation for functional types of disorder*

We perform first-of-its kind evaluation of disorder predictions for different functional types of disorder, which are annotated based on the DisProt database (Sickmeier et al., 2007). We collected annotations for six functional types: disorder-mediated protein–protein binding, protein-DNA binding, substrate or ligand binding, flexible linkers or spacers, disordered phosphorylation sites, and disordered autoregulatory regions; see Table 1. We assess whether residues of a given type are correctly predicted as being disordered; see details in the "Datasets and evaluation protocols" section. The corresponding AUC values for disCoP, disCoP_WS and each of the 20 considered predictors and each of the functional types on the TEST_FUNCTION dataset are shown in Table 7. The Table ranks all methods based on an average rank across the six functional types. The corresponding Table that reports the values of the constrained AUC is shown in the Supplementary Table S2. Table 7 shows that disCoP obtains the highest average rank of AUC, which means that its predictions are on average ranked the highest across the six functional types of disorder. The ESpritz-Disprot method provides the most accurate disorder predictions for residues involved in disorder-mediated binding, including binding with proteins, DNA, and other ligands. On the other hand, disCoP provides accurate results for the prediction of disorder that implements other considered functions, such as linker and autoregulatory regions. Both of these methods perform well for disordered post-translational modification sites, in particular phosphorylation sites.

Table 7. AUC values measured on the TEST_FUNCTION dataset for disCoP, disCoP and 20 other predictors for the six functional types of disorder. The AUC values are averages over the 10 repetitions with different randomly selected sets of structured residues (see "Datasets and evaluation protocols" section for details). Methods that are used as inputs to disCoP are shown in bold font. The highest value for each functional type is given in bold font. The methods are sorted by average rank of AUC, which is the average over the ranks for individual functional types. disCoP_WS is a web server version of disCoP that excludes ESpritz and CSpritz predictors (see "disCoP predictor" section for details).

	Functional types related to binding			Other functional types				
Method	Protein-protein binding	Substrate or ligand binding	Protein-DNA binding	Flexible linkers or spacers	Phosphorylation	Autoregulatory	Average AUC	Average rank of AUC
disCoP	.869	.755	.793	**.794**	**.898**	**.879**	.831	**1.7**
ESpritz-Disprot	**.881**	**.783**	**.834**	.767	**.898**	.871	**.839**	2.2
MD	.838	.725	.816	.774	.893	.843	.815	3.7
CSpritz-long	.831	.728	.760	.792	.881	.873	.811	4.0
disCoP_WS	.835	.712	.787	.773	.888	.848	.807	4.2
MFDp	.815	.725	.751	.752	.864	.828	.789	7.5
CSpritz-short	.786	.721	.753	.782	.826	.827	.783	8.3
SPINE-D	.804	.688	.730	.768	.868	.840	.783	8.5
ESpritz-Xray	.789	.722	.743	.746	.838	.796	.772	10.0
VSL2B	.788	.714	.741	.716	.840	.807	.768	11.2
PrDos	.784	.645	.715	.778	.838	.835	.766	11.8
DISOPRED2	.786	.654	.697	.761	.850	.835	.764	12.2
PONDR-FIT	.781	.720	.729	.726	.823	.789	.761	13.2
DISOCLUST	.782	.664	.696	.748	.858	.823	.762	13.2
IUPRED-long	.782	.703	.703	.711	.806	.774	.747	15.2
Norsnet	.751	.626	.746	.710	.839	.773	.741	16.0
IUPRED-short	.770	.710	.697	.729	.798	.765	.745	16.2
RONN	.760	.676	.686	.719	.809	.783	.739	16.5
Ucon	.757	.692	.716	.705	.759	.762	.732	17.7
ProfBval	.706	.674	.704	.705	.772	.744	.718	19.3
ESpritz-NMR	.746	.665	.674	.670	.807	.765	.721	19.5
DRIP-PRED	.708	.643	.638	.707	.805	.755	.709	20.5

Importantly, our results provide useful clues to potential users of disorder predictors when they aim to investigate certain functional aspects. We observe that disordered sites that are involved in binding to smaller ligands (excluding proteins and DNA) and linker regions are more difficult to predict when compared to the other functional types. Moreover, the users could match their study with a particular method that allows them to maximize predictive accuracy.

3.3. Predictive model

The disCoP uses a set of selected 11 features that aggregate disorder predictions generated by seven methods and which are combined together to predict disorder using binomial deviance loss-derived regression; the design of this model was performed empirically based on threefold cross-validation on the TRAINING dataset. We analyze contributions of these features and their relation with the native disorder annotations. Table 8 lists the selected features together with a few measures that evaluate their individual predictive performance, their contribution to the consensus based on the results of feature selection, and their corresponding weights in the regression; the same information for the disCoP_WS predictor is summarized in the Supplementary Table S3. The feature names define the input predictor (the part of the name before underscore) and type of output and aggregation (the part after underscore) where median *i* and mean *i* correspond to median and mean probability in window of size $2 \times i + 1$, respectively, and content *i* and Lcontent *i* correspond to content of binary and ternary predictions in window of size $2 \times i + 1$, respectively (see "Feature generation and selection" for details).

We observe that all selected features are calculated based on aggregate values; none of the raw predictions are used. This demonstrates that our design of features that implements aggregation is beneficial to the prediction. Five features are based on predicted real-valued probabilities (mean and median-based features) and six utilize the binary and ternary predictions (content and Lcontent features). Probability-based aggregation uses smaller window sizes compared with the aggregated binary/ternary predictions; this justifies our feature selection that chooses the preferred window sizes separately for each feature type.

The individual performance is assessed with constrained AUC when each feature is used individually to predict disorder on the TRAINING dataset and their biserial correlation with the native annotation of disorder. As expected, these two measures are strongly correlated with each other, with the Pearson's correlation coefficient (PCC) of .87, and with the AUC of the corresponding predictors on the TRAINING dataset, with PCC of .80 for the biserial correlation and .54 for the constrained AUC. This suggests that predictive performance of aggregated predictions is correlated with the overall performance of the corresponding predictor, which is expected. The performance of the selected features when

Table 8. Summary of 11 features used in the disCoP consensus. The features are sorted by their constrained AUC when used individually to predict the disorder based on threefold cross-validation on the TRAINING dataset. The biserial correlation was computed against the native disorder annotation in the TRAINING dataset. The "Constrained AUC then added to consensus" gives the value of the constrained AUC when a given feature was added into the consensus during the feature selection. The last column lists weights in the regression including a bias (free weight), which is listed in the last row. The features with negative weights are given in bold font. The first part of the feature name (before underscore) identifies the input predictor; the second part shows the particular type of output and aggregation where median*i* and mean*i* correspond to median and mean probability in window of size $2 \times i+1$, respectively, and content *i* and Lcontent *i* correspond to content of binary and ternary predictions in window of size $2 \times i+1$, respectively (see "Feature generation and selection" for details).

	Predictive performance of individual features			
Features	Constrained AUC of individual features	Biserial correlation with native disorder	Constrained AUC when added to consensus	Regression weights
ESpritz-Disprot_median6	.546	.487	.546	.252
CSpritz-Long_median7	.542	.475	.579	.317
MD_mean8	.490	.456	.584	.005
ESpritz-Disprot_Lcontent28	.481	.450	.589	−.026
SPINE-D_content24	.483	.453	.602	.106
DISOPRED_mean15	.471	.416	.612	.042
DISOclust_median17	.471	.393	.619	.009
ESpritz-Xray_mean26	.462	.420	.624	−.079
MD_content29	.459	.427	.630	−.030
CSpritz-Long_content30	.444	.423	.635	−.082
DISOPRED_Lcontent30	.431	.385	.641	−.051
Bias				.268

used together in the consensus is measured using constrained AUC values that were obtained when adding these features during the feature selection (see also Supplementary Figure S2) and their weights in the regression model. Most of the initial gain in performance when adding the features comes from the use of probability-based aggregates that utilize predictions of the well-performing ESpritz, CSpritz and MD predictors. However, later on, we add features that utilize predictors with lower overall predictive performance (such as DISOclust and ESpritz-Xray). Also, the features added at the end of the feature selection (that are listed at the bottom of Table 8) apply larger window sizes. This means that the larger sequence context that they utilize can be effectively used to adjust contributions from the more locally computed (using smaller windows) features. Interestingly, features with negative weights in the regression (shown in bold font in Table 8), which reduce the probability of the disorder prediction by the consensus, are also associated with aggregations based on larger window sizes. These features do not utilize the DISOclust and SPINE-D predictors, which are characterized by relatively large values of DDC (e.g. see Table 5), that is which over-predict the disorder content. This means that predictions from more conservative methods are used to balance the over-predictions from these two methods.

We also provide additional, empirical motivation for combining multiple features together. Figure 3 shows a scatter plot of values of two features, DISOclust_median17 and CSpritz-Long_median7, together with the annotation of the disorder/order for a randomly selected subset of 10% of residues in the TRAINING dataset. These features are based on two predictors with substantially different predictive profiles, as measured with their overall performance, DDC, and balance between sensitivity and specificity; see Table 5. The scatter plot demonstrates that combining these two features leads to a better discriminatory power compared to using them separately. The feature based on the DISOclust predictions (*y*-axis) shows that majority of residues with values above .5 are disordered. However, with the help of the second feature based on the CSpritz predictions (*x*-axis), these residues can be more accurately classified as disordered if the second feature has values at above .5. In other words, some of the residues incorrectly predicted as disordered by DISOclust-based feature can be corrected with the CSpritz-based feature and vice versa.

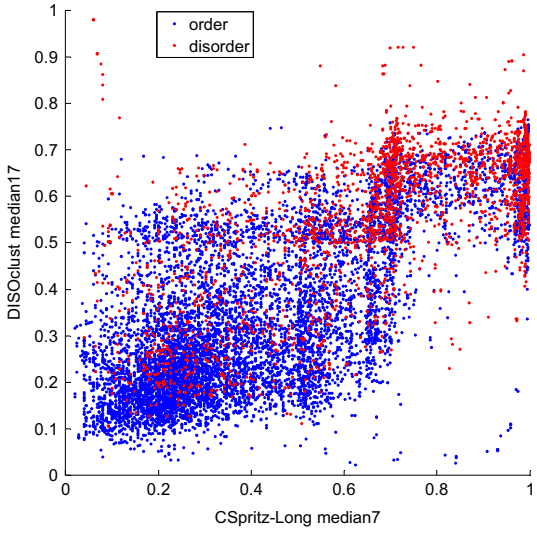

Figure 3. Scatter plot of values of two features, DISOclust_median17 (*y*-axis) and CSpritz_Long_median7 (*x*-axis), for randomly selected subset of 10% of residues from the TRAINING dataset. Colors denote outcomes, with red for disordered residues and blue for ordered/structured residues.

3.4. *Case studies*

We visualize predictions from our disCoP and its base/input predictors using two case studies that concern proteins from the TEST set with different disorder characteristics. We plot the native annotations of disorder and predicted probabilities and binary annotations of disorders, which for the considered predictors are based on a cut-off of .5; see Figure 4. While these cases should not be considered as typical, they attempt to visualize the results at the protein level and highlight differences between predictions generated by the disCoP and its input methods.

The first case is the potassium voltage-gated channel protein Shaker (DisProt ID DP00267) (Hoshi, Zagotta, & Aldrich, 1990), which is a relatively long chain with one native relatively short disordered region (in the vicinity of the N-terminus) that is annotated as a flexible linker; see Figure 4(a). The Figure reveals that all base/input predictors correctly find disorder at the N-terminus. However, they also predict a long disordered segment at the C-terminus, and a few disordered residues or short segments in the middle of the chain. Our consensus effectively reduces the over-prediction of the disorder while preserving the prediction of the disordered region at the N-terminus.

The second case is the DNA repair protein (DisProt ID DP00091) (Iakoucheva et al., 2001), which is a short chain with two fairly long native disordered regions that implement protein–protein binding; see Figure 4(b). The native disorder is located at both termini. The predictions of different base predictors vary in their probability profiles. The ESpritz-Xray and DISOPRED2 methods under-predict the disorder, while the ESpritz-Disprot, DISOclust and MD over-predict the amount of disorder. SPINE-D predicts fairly accurate amount of disorder per chain, but misaligns the predicted disordered segments with respect to the position of the native segments. The CSpritz-Long method provides the highest quality predictions among the seven input meth-

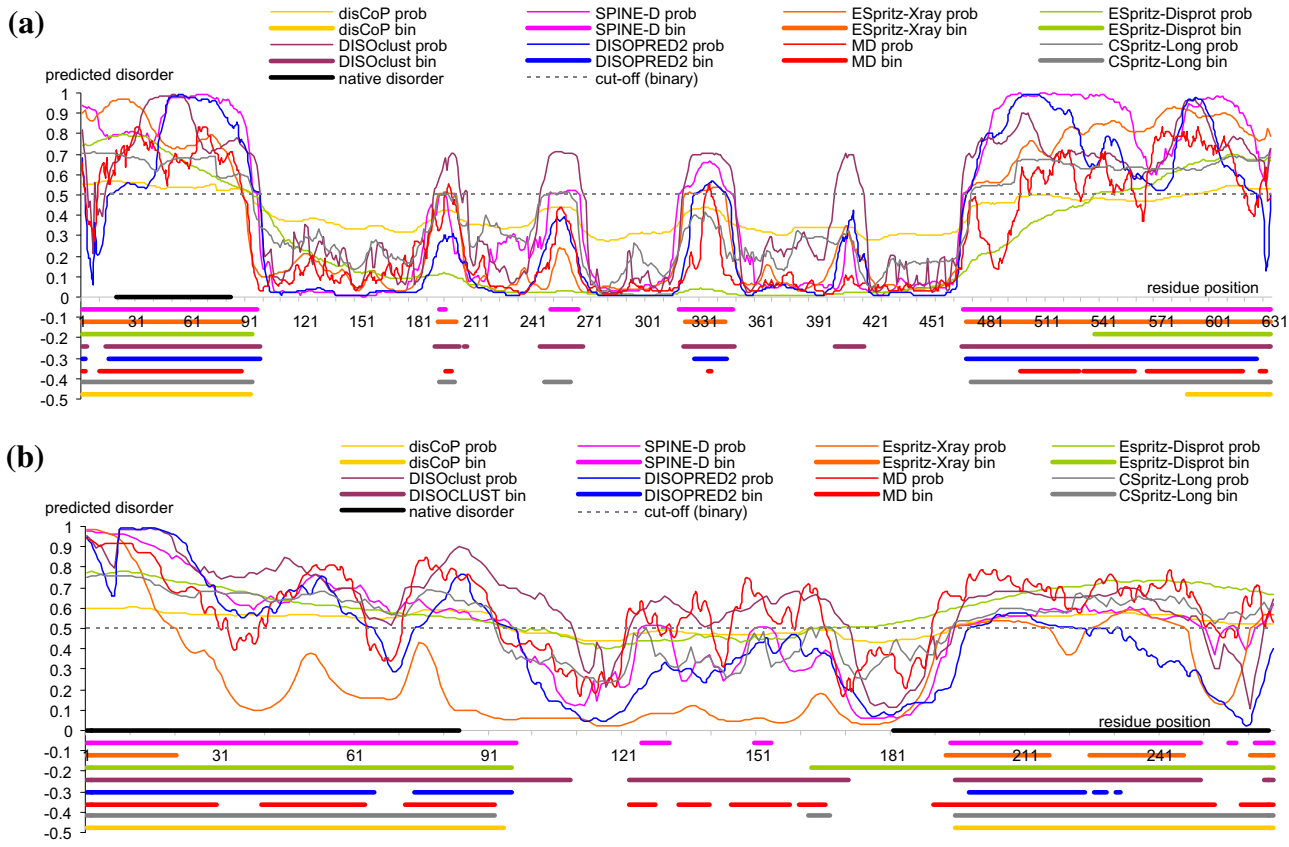

Figure 4. Prediction of disordered residues for (panel a) potassium voltage-gated channel protein Shaker (DisProt ID DP00267) and (panel b) DNA repair protein (DisProt ID DP00091), which were taken from TEST dataset, by disCoP and its seven input predictors. The *x*-axis shows positions in the protein sequence. Probability values are shown by thin lines at the top of the figure. The cut-off of .5 to convert probabilities into binary predictions is shown using a dashed horizontal line. The native MoRF regions are annotated using black horizontal line. The binary predictions are denoted using thick horizontal lines below 0 on the *y*-axis.

ods. Our disCoP improves slightly over the CSpritz-Long by removing a short disordered segment in the middle of the chain.

Overall, disCoP offers predictions with a relatively low false-positive rate (rate of incorrectly predicted disordered residues) by removing some of the "less reliable" disorder predictions generated by its input predictors. We also observe that our method provides a smoother profile of predicted probabilities, which results from the averaging done by the regression.

4. Conclusions

Our study builds upon the diversity of the existing disorder prediction to propose a consensus that provides improved predictive quality. We address the shortage of meta-methods which go beyond the current designs that are performed in an ad hoc manner. Our approach, named disCoP, takes advantage of three novel aspects: empirical selection of a well performing/complementary set of predictors (using a comprehensive list of state-of-the-art methods); use of a more sophisticated averaging via binomial deviance-based regression to implement the consensus; and utilization of a custom designed features that aggregate input binary and real-valued predictions. Our consensus uses 11 features that are calculated based on outputs of seven disorder predictors. Empirical evaluation on an independent test dataset demonstrates that our consensus provides improved predictive quality, when compared with its input predictors and a wide-range of other predictors, including a state-of-the-art consensus that applies twice as many input predictors. Moreover, our evaluation that considers different functional types of disorder shows that disCoP provides strong disorder predictions on several types of disorder, including disordered phosphorylation sites and autoregulatory and flexible linker regions. The latter evaluation provides useful clues for users who can now tailor selection of a particular predictor for a given disorder-mediated function that they investigate.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2013.775969.

Acknowledgement

The authors thank Mr. Marcin Mizianty for help with the implementation and testing of the disCoP_WS method.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*, 3389–3402.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Journal of Information Fusion, 6*, 5–20.

Cheng, J., Sweredoski, M. J., & Baldi, P. (2005). Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery, 11*, 213–222. doi:10.1007/s10618-005-0001-y

Deng, X., Eickholt, J., & Cheng, J. (2009). PreDisorder: *Ab initio* sequence-based prediction of protein disordered regions. *BMC Bioinformatics, 10*, 436. doi:10.1186/1471-2105-10-436

Dosztanyi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of Molecular Biology, 347*, 827–839. doi:10.1016/j.jmb.2005.01.071

Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics, 21*, 3433–3434. doi:10.1093/bioinformatics/bti541

Dosztányi, Z., Mészáros, B., & Simon, I. (2010). Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Briefings in Bioinformatics, 11*, 225–243. doi:10.1093/bib/bbp061

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., & Obradovic, Z. (2002). Intrinsic disorder and protein function. *Biochemistry, 41*, 6573–6582.

He, B., Wang, K., Liu, Y., Xue, B., Uversky, V. N., & Dunker, A. K. (2009). Predicting intrinsic disorder in proteins: An overview. *Cell Research, 19*, 929–949.

Hirose, S., Shimizu, K., Kanai, S., Kuroda, Y., & Noguchi, T. (2007). POODLE-L: A two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics, 23*, 2046–2053. doi:10.1093/bioinformatics/btm302

Hoshi, T., Zagotta, W. N., & Aldrich, R. W. (1990). Biophysical and molecular mechanisms of Shaker potassium channel inactivation. *Science, 250*, 533–538.

Iakoucheva, L. M., Kimzey, A. L., Masselon, C. D., Bruce, J. E., Garner, E. C., Brown, C. J., … Ackerman, E. J. (2001). Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Science, 10*, 560–571. doi:10.1110/ps.29401

Ishida, T., & Kinoshita, K. (2007). PrDOS: Prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Research, 35*(Web Server issue), W460–W464. doi:10.1093/nar/gkm363

Ishida, T., & Kinoshita, K. (2008). Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics, 24*, 1344–1348. doi:10.1093/bioinformatics/btn195

Jones, D. T., & Ward, J. J. (2003). Prediction of disordered regions in proteins from position specific score matrices. *Proteins, 53*(Suppl 6), 573–578. doi:10.1002/prot.10528

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22*, 2577–2637. doi:10.1002/bip. 360221211

Kozlowski, L. P., & Bujnicki, J. M. (2012). MetaDisorder: A meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics, 13*, 111-2105-13-111. doi:10.1186/1471-2105-13-111; 10.1186/1471-2105-13-111

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., & Russell, R. B. (2003). Protein disorder prediction: Implications for structural proteomics. *Structure, 11*, 1453–1459.

Linding, R., Russell, R. B., Neduva, V., & Gibson, T. J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Research, 31*, 3701–3708.

MacCallum, B. (2004). Order/disorder prediction with self organizing maps. [Map]. Retrieved from http://www.webcitation.org/query.php?url=http://www.forcasp.org/paper2127.html doi:10.1186/1471-2164-9-s1-s9

Martin, A. J., Walsh, I., & Tosatto, S. C. (2010). MOBI: A web server to define and visualize structural mobility in NMR protein ensembles. *Bioinformatics, 26*, 2916–2917. doi:10.1093/bioinformatics/btq537

McGuffin, L. J. (2008). Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics, 24*, 1798–1804. doi:10.1093/bioinformatics/btn326

Midic, U., Oldfield, C. J., Dunker, A. K., Obradovic, Z., & Uversky, V. N. (2009). Protein disorder in the human diseasome: Unfoldomics of human genetic diseases. *BMC Genomics, 10*(Suppl 1), S12. doi:10.1186/1471-2164-10-S1-S12

Mizianty, M. J., Stach, W., Chen, K., Kedarisetti, K. D., Disfani, F. M., & Kurgan, L. (2010). Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics, 26*, i489–i496. doi:10.1093/bioinformatics/btq373

Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A., & Kryshtafovych, A. (2011). Evaluation of disorder predictions in CASP9. *Proteins, 79*, 107–118. doi:10.1002/prot.23161

Noivirt-Brik, O., Prilusky, J., & Sussman, J. L. (2009). Assessment of disorder predictions in CASP8. *Proteins, 77*(Suppl 9), 210–216. doi:10.1002/prot.22586

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., & Dunker, A. K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins, 53*(Suppl 6), 566–572 doi:10.1002/prot.10532

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., & Dunker, A. K. (2005). Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins, 61*(Suppl 7), 176–182. doi:10.1002/prot.20735

Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., & Dunker, A. K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry, 44*, 1989–2000. doi:10.1021/bi047993o

Peng, Z. L., & Kurgan, L. (2012a). Comprehensive comparative assessment of in-silico predictors of disordered regions. *Current Protein and Peptide Science, 13*, 6–18.

Peng, Z. L. & Kurgan, L. (2012b). On the complementarity of the consensus-based disorder prediction. *Pacific Symposium on Biocomputing, 17*, 176–187.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., & Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics, 7*, 208. doi:10.1186/1471-2105-7-208

Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., Dunker, A. K., & Obradovic, Z. (2005). Optimizing long intrinsic disorder predictors with protein evolutionary information. *Journal of Bioinformatics and Computational Biology, 3*, 35–60.

Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., & Sussman, J. L. (2005). FoldIndex: A simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics, 21*, 3435–3438. doi:10.1093/bioinformatics/bti537

Raychaudhuri, S., Dey, S., Bhattacharyya, N. P., & Mukhopadhyay, D. (2009). The role of intrinsically unstructured proteins in neurodegenerative diseases. *PLoS ONE, 4*, e5566.

Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., & Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins, 42*, 38–48.

Schlessinger, A., Liu, J., & Rost, B. (2007). Natively unstructured loops differ from other loops. *PLoS Comput Biology, 3*, e140.

Schlessinger, A., Punta, M., & Rost, B. (2007). Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics, 23*, 2376–2384. doi:10.1093/bioinformatics/btm349

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., & Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE, 4*, e4433.

Schlessinger, A., Yachdav, G., & Rost, B. (2006). PROFbval: Predict flexible and rigid residues in proteins. *Bioinformatics, 22*, 891–893. doi:10.1093/bioinformatics/btl032

Shimizu, K., Muraoka, Y., Hirose, S., Tomii, K., & Noguchi, T. (2007). Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics, 8*, 78. doi:10.1186/1471-2105-8-78

Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., … Dunker, A. K. (2007). DisProt: The database of disordered proteins. *Nucleic Acids Research, 35*(Database issue), D786–D793. doi:10.1093/nar/gkl893

Sirota, F., Ooi, H., Gattermayer, T., Schneider, G., Eisenhaber, F., & Maurer-Stroh, S. (2010). Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. *BMC Genomics, 11*, 1–17. doi:10.1186/1471-2164-11-S1-S15

Su, C. T., Chen, C. Y., & Hsu, C. M. (2007). iPDA: Integrated protein disorder analyzer. *Nucleic Acids Research, 35* (Web Server issue), W465–W472. doi:10.1093/nar/gkm353

Su, C. T., Chen, C. Y., & Ou, Y. Y. (2006). Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics, 7*, 319. doi:10.1186/1471-2105-7-319

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends in Biochemical Sciences, 27*, 527–533.

Uversky, V. N., Gillespie, J. R., & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins, 41*, 415–427.

Uversky, V. N., Oldfield, C. J., & Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: Introducing the D2 concept. *Annual Review of Biophysics, 37*, 215–246. doi:10.1146/annurev.biophys.37.032807.125924

Vucetic, S., Brown, C. J., Dunker, A. K., & Obradovic, Z. (2003). Flavors of protein disorder. *Proteins, 52*, 573–584. doi:10.1002/prot.10437

Vullo, A., Bortolami, O., Pollastri, G., & Tosatto, S. C. (2006). Spritz: A server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Research, 34* (Web Server issue), W164–W168. doi:10.1093/nar/gkl166

Walsh, I., Martin, A. J., Di Domenico, T., & Tosatto, S. C. (2012). ESpritz: Accurate and fast prediction of protein disorder. *Bioinformatics, 28*, 503–509. doi:10.1093/bioinformatics/btr682

Walsh, I., Martin, A. J., Di Domenico, T., Vullo, A., Pollastri, G., & Tosatto, S. C. (2011). CSpritz: Accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Research, 39* (Web Server issue), W190–W196. doi:10.1093/nar/gkr411

Wang, L., & Sauer, U. H. (2008). OnD-CRF: Predicting order and disorder in proteins using [corrected] conditional random fields. *Bioinformatics, 24*, 1401–1402. doi:10.1093/bioinformatics/btn132

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., & Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics, 20*, 2138–2139. doi:10.1093/bioinformatics/bth195

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology, 337*, 635–645. doi:10.1016/j.jmb.2004.02.002

Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K., & Uversky, V. N. (2010). PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochimica Et Biophysica Acta (BBA) – Proteins & Proteomics, 1804*, 996–1010. doi:10.1016/j.bbapap.2010.01.011

Xue, B., Dunker, A. K., & Uversky, V. N. (2012). Orderly order in protein intrinsic disorder distribution: Disorder in 3500 proteomes from viruses and the three domains of life. *Journal of Biomolecular Structure and Dynamics, 30*, 137–149. doi:10.1080/07391102.2012.675145;10(1080/07391102), 2012, 675145

Yang, Z. R., Thomson, R., McNeil, P., & Esnouf, R. M. (2005). RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics, 21*, 3369–3376. doi:10.1093/bioinformatics/bti534

Yang, M. Q., & Yang, J. Y. (2006). IUP: Intrinsically unstructured protein predictor – a software tool for analyzing polypeptide sequences. *Sixth IEEE Symposium on BioInformatics and BioEngineering, BIBE 2006*, 3–11.

Zhang, T., Faraggi, E., Xue, B., Dunker, A. K., Uversky, V. N., & Zhou, Y. (2012). SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method. *Journal of Biomolecular Structure and Dynamics, 29*, 799–813.