Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources

Marcin J. Mizianty, Wojciech Stach, Ke Chen, Kanaka Durga Kedarisetti, Fatemeh Miri Disfani and Lukasz Kurgan∗
Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada T6G 2V4

ABSTRACT
Motivation: Intrinsically disordered proteins play a crucial role in numerous regulatory processes. Their abundance and ubiquity combined with a relatively low quantity of their annotations motivate research toward the development of computational models that predict disordered regions from protein sequences. Although the prediction quality of these methods continues to rise, novel and improved predictors are urgently needed.

Results: We propose a novel method, named MFDp (Multilayered Fusion-based Disorder predictor), that aims to improve over the current disorder predictors. MFDp is as an ensemble of 3 Support Vector Machines specialized for the prediction of short, long and generic disordered regions. It combines three complementary disorder predictors, sequence, sequence profiles, predicted secondary structure, solvent accessibility, backbone dihedral torsion angles, residue flexibility and B-factors. Our method utilizes a custom-designed set of features that are based on raw predictions and aggregated raw values and recognizes various types of disorder. The MFDp is compared at the residue level on two datasets against eight recent disorder predictors and top-performing methods from the most recent CASP8 experiment. In spite of using training chains with ≤25% similarity to the test sequences, our method consistently and significantly outperforms the other methods based on the MCC index. The MFDp outperforms modern disorder predictors for the binary disorder assignment and provides competitive real-valued predictions. The MFDp’s outputs are also shown to outperform the other methods in the identification of proteins with long disordered regions.

Availability: http://biomine.ece.ualberta.ca/MFDp.html
Supplementary information: Supplementary data are available at Bioinformatics online.
Contact: lkurgan@ece.ualberta.ca

1 INTRODUCTION
The intrinsically disordered proteins (IDPs), also referred to as unstructured proteins, lack stable tertiary structure under physiological conditions in vitro. The IDPs play a crucial role in transcriptional regulation, translation and cellular signal transduction (Dunker et al., 2008). Their prevalence was also implicated in various human disorders including the neurodegenerative diseases such as Huntington, Parkinson’s and Alzheimer’s disease (Raychaudhuri et al., 2009). However, the functional role of IDPs is not as well understood when compared with the well-packed proteins. Importantly, the annotations of the disorder lag behind the rapidly accumulating number of known protein chains. To compare, the curated DisProt database (Sickmeier et al., 2007) includes little over 500 chains, the PDB database (Berman et al., 2000) that allows finding unstructured segments in the solved tertiary structures (which are assumed to be equivalent to disordered segments) includes ∼58,000 proteins, while the overall number of known protein chains is >9 million. The disorder is frequently observed in regions with low-sequence complexity and with low content of hydrophobic amino acids, which would often form a core of a folded globular protein (Dyson and Wright, 2005; Uversky et al., 2000). These and other sequence characteristics can be used to differentiate between disordered and ordered regions, which in turn implies that disorder is predictable from the sequence. The past decade observed development of a number of computational models for the prediction of the disordered regions. These methods allow for high-throughput annotations of protein chains which provide a viable solution to close the annotation gap. These predictors could be categorized into four groups: (i) methods based on relative propensity of amino acids to form disorder/ordered regions which include GlobPlot (Linding et al., 2003), IUPred (Dosztányi et al., 2005), FoldIndex (Ptitsyn et al., 2005) and Ucon (Schlessinger et al., 2007a); (ii) methods built utilizing machine-learning classifiers, such as DIProE (Hecker et al., 2008), DISOPRED (Jones and Ward, 2003) DISOPRED2 (Ward et al., 2004), PrDOS (Ishida and Kinoshita, 2007), POODLE-S (Shimizu et al., 2007a), POODLE-L (Hirose et al., 2007), POODLE-W (Shimizu et al., 2007b), Spritz (Vullo et al., 2006), DisPSSMP (Su et al., 2006), DisPSSMP2 (Su et al., 2007), IUP (Yang and Yang, 2006), NORSnet (Schlessinger et al., 2007b) and OnD-CRFs (Wang and Sauer, 2008); (iii) methods based on a meta-approach which combines predictions from multiple base predictors including recent MD (Schlessinger et al., 2009), metaDOS (Ishida and Kinoshita, 2008), GS-metaDisorder (J.Bujnicki, unpublished data) and MULTICOM (Cheng et al., 2005); and (iv) methods based on analysis of predicted 3D structural models, such as DSSPclust (McGuffin, 2000). Since 2002, the sequence-based disorder predictors are biannually assessed and compared during the critical assessment of structure prediction (CASP) experiments. Although the prediction quality continues to rise, as shown in the most recent CASP8 (Novotný-Brik et al., 2009), improved prediction methods are still urgently needed (Schlessinger et al., 2009).

We propose a novel architecture, named MFDp (Multilayered Fusion-based Disorder predictor), that aims to improve the overall quality of the prediction when compared with modern methods. We analyze and quantify improvements provided by MFDp for a generic prediction of all disordered segments and also for the prediction of long-disordered segments. The latter is motivated by recent results that show that long-disordered regions are useful for target selection for structure determination of integral membrane protein (Punta et al., 2009).
et al., 2009), they are implicated in protein–protein recognition (Tompa et al., 2009) and were found helpful in prediction of protein crystallization propensity (Slabinski et al., 2007).

The MFDp is based on four novel ideas. First, motivated by the observation that combining orthogonal predictors is helpful (Oldfield et al., 2005; Schlessinger et al., 2009), we fuse four complementary disorder predictors. This is in contrast to earlier ensemble-based solutions that combined predictors selected in an ad hoc manner. When compared to the recently proposed orthogonal ensemble-based MD predictor (Schlessinger et al., 2009), we consider different aspects to judge dissimilarity. MD combines four predictors that tackle different ‘types’ of disorder, while we combine three methods that differ in their approach to perform predictions. We use a machine learning-based predictor DISOPRED2, residue propensity-based IUPred and a recent DISOclust that is based on tertiary structure predictions (i.e. prediction based on the sequence-derived tertiary structure). The usage of the 3D-based predictor is the novel aspect of our ensemble. We select DISOPRED2 since it was demonstrated to provide high-quality predictions and to be orthogonal to other machine learning-based methods (Schlessinger et al., 2009). The IUPred provides two models that specialize in prediction of long- and short-disordered regions, respectively. The DISOclust utilizes a premise that ordered residues should be conserved in 3D space in multiple models, whereas residues that vary or are consistently missing are correlated with the disorder. It predicts per-residue error in multiple fold recognition models which is followed by an analysis of the conservation of the per-residue errors across all models (McGuffin, 2008).

Second, we use the most comprehensive selection of the input information sources when compared with the recent methods (Table 1). Similar to MD, DISpro and MULTICOM, we use the sequence profiles (the most widely used inputs; Table 1), predicted secondary structure (SS) [disordered regions are characterized by lack of SS (Radivojac et al., 2004, 2007; Vucetic et al., 2003)] and solvent accessibility [unstructured regions have a large solvent-accessible area (Schlessinger et al., 2009)]. We also utilize sequence-based predictions of backbone dihedral torsion angles, B-factor [which are associated with disordered regions (Zhang et al., 2009)] and the sequence terminus indicator [similarly to PrDOS (Su et al., 2007)]. The usage of the backbone angles is motivated by their usefulness in building-scoring function for fold recognition and 3D structure prediction (Wu and Zhang, 2008), and by the success of the 3D-based DISOclust that ranked fourth in CASP8 (Noivirt-Brik et al., 2009). We tried predictions of signal peptides, but we did not find them useful.

Third, we aggregate the predictions, before feeding them into a classifier, using a sliding window to facilitate predictions of long-disordered regions. The aggregation utilizes neighboring predictions to perform averaging, to find maximal and minimal prediction value/probability and to analyze local SS conformations, which helps the classifier to find longer stretches of disordered residues.

Fourth, we use two-layered architecture where the first layer includes three predictors, one designed for all disordered residues, one for short (<30 residues), and one for long segments. These predictors use different inputs encoded using numerical descriptors derived from the abovementioned sources. The second layer combines their outputs to generate our disorder predictions.

MFDp is capable of recognizing various types of disorder, since it is trained using a dataset that includes residues from disordered regions of all sizes and that combines proteins from the DisProt database and X-ray structures from the PDB.

### 2 METHODS

#### 2.1 Datasets

The proposed method was designed and tested on a dataset with 514 protein sequences. These sequences were harvested from the PDB and the DisProt databases. The PDB sequences were filtered using the called PDB list (Wang and Dunbrack, 1993) to extract a high-quality and low-sequence identity subset. We selected sequences that have structures with R-factor <0.2 and resolution <2.0Å, and that are characterized by sequence identity <5%.

The above follows the selection process used by GS-meta Disorder, one of the leading methods in CASP8. Since most protein chains in PDB are completely ordered, we kept randomly selected 20% of the ordered proteins. We extracted the entire set of 1195 proteins from the current release 4.9 of the DisProt and we merged them with the PDB chains. The combined set was
2.2 Definition of disorder

Prior work shows that the assignment of the disordered regions performed using different experimental methods is not always consistent (Vucetic et al., 2003). The disorder predictors that were trained on regions identified by one experimental method could be less accurate for prediction of disordered regions that were characterized by other methods (Schlessinger et al., 2007a, b). To date, there is no golden standard for the assignment of the disordered regions. In the past CASP experiments the disordered regions were defined as residues that lack coordinates in structures solved by X-ray crystallography and as residues that exhibit high variability within ensemble or are annotated as disordered in REMARK 465 by experimentalists for the structures solved by NMR (Bordoli et al., 2007; Noivirt-Brik et al., 2009). Consequently, the above definition is used to annotate disordered residues in the CASP8 dataset. The MdX dataset utilizes two types of annotations to generalize the proposed predictive model. The chains extracted from PDB are annotated using the above definition, while the chains from DisProt use the curated annotations extracted from this database.

2.3 Evaluation criteria and test procedures

The assessment of the predictions uses the same criteria as in the CASP experiments (Bordoli et al., 2007; Noivirt-Brik et al., 2009). The evaluation was performed per-residue since per-protein predictions are more prone to overfitting (Schlessinger et al., 2009). The predictions are at two levels: (i) the binary value that defines whether a given residue is or is not disordered; and (ii) the real value that quantifies probability of disorder. The binary predictions were assessed using five measures:

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

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

Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}}

\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{N}}

\text{F1} = 2 \times \frac{\text{S} \times \text{MCC}}{\text{S} + \text{MCC}}

\text{F2} = \frac{4 \times \text{F1}}{3 \times \text{F1} + 1}

TP = \text{true positives (correctly predicted disordered residues)},

FN = \text{false negatives (disordered residues that were predicted ordered)},

TN = \text{true negatives (correctly predicted ordered residues)},

FP = \text{false positives (ordered residues that were predicted disordered)}

F1 and F2 are equal zero when all residues are predicted to disorder and one experimental method could be less accurate for prediction of disordered regions identified by other methods. The disorder predictors that were trained on regions identified by other methods (Schlessinger et al., 2009) are designated to predict all disordered residues (SVMALL), and disordered residues in long (≥30 residues) (SVMLONG) and short segments (<30 residues) (SVMSHORT). Each classifier is designed individually, i.e. uses different parameters and different set of input features (see Sections 2.5 and 2.6) computed from the IUPred, DISOclust and DISOPRED2 predictions, the sequence, the sequence profiles and the various relevant sequence-based predictions. The latter predictions are performed using methods chosen based on two criteria: (i) their quality is comparable to the leading relevant predictors; and (ii) they provide a stand-alone implementation that could be utilized to build an autonomous implementation of the proposed disorder predictor. We utilize PSIPRED (McGuffin et al., 2000) for the SS prediction, Real-SPINE3 for the prediction of the relative solvent accessibility (RSA) and backbone dihedral torsion angles (Faraggi et al., 2009), PROBval for the B-factor prediction and residue flexibility prediction (Schlessinger et al., 2006) and RPSF for the signal peptide prediction (Plewczyński et al., 2008). We also use IUPred to predict globular domains. The selection of the SVM as the classifiers was motivated by its prior extensive use in disorder prediction (Hirose et al., 2007; Ishula and Kinoshita 2007, 2008; Peng et al., 2006; Shimizu et al., 2007a, b, Vullo et al. 2006). We use a linear-kernel-based SVM introduced in Fan et al. (2008). The real-valued prediction is computed as a maximum among the probabilities generated by SVMALL, SVMSHORT and SVMLONG (we combine all predicted disordered residues) and the resulting value is binarized using the threshold that equals 0.37. The value of the threshold was found using the MdX dataset. We ’smooth out’ our binary predictions by filtering out short segments (<2 residues). On the first pass, we convert predicted ordered residues in segments <2 residues to disordered residues with probability of 0.37. On the second pass, we reassess the short, <2 residues, segments of disordered residues as ordered with probability of 0.369. The MfDp predictor is shown in Figure 1.

2.4 Architecture

The MfDp utilizes four complementary disorder predictors, IUPred LONG (IUPREDL), IUPred SHORT (IUPRED3), DISOPRED2 and DISOclust, the sequence, the PSSM profile (Altschul et al., 1997), the predicted SS, the predicted RSA, the predicted globular domains (IUPREDG), backbone dihedral torsion angles and signal peptides. The predictors were run with their default parameters. The PSSM profiles were generated using the non-redundant (nr) database from NCBI (downloaded on November 19, 2009), which was filtered using PPHIE (Jones and Swindells, 2002) to remove low-complexity regions, transmembrane regions and coiled-coil segments.

We utilize a sliding window of size 15 centered over the predicted residue to extract the features. We use the raw numerical values for each of the 15 positions, which include: the probability of the prediction of disordered residues from IUPREDL, IUPRED3, DISOPRED2 and DISOclust; the PSSM values; the probabilities of prediction of helix, coil and strand conformations from PSIPRED; the predicted B-factor values; the predicted solvent accessibility and backbone angles; and the binary values denoting whether a given position in the window is predicted as a signal peptide.
We designed three SVMs using different versions of the MxD dataset. The
when aggregating, we ignore the positions in the window that are outside
the average biserial correlation, from each category. We take all features
in a given category if their number < τ. Next, we run 5-fold cross validation
on the MxD dataset where we vary the value of τ between 1 and 15, and for
each τ we parameterize the value of constant C of the linear SVM classifiers.
We consider C in consecutive powers of 2 between 2^{-3} and 2^7. For each of
the prediction models (ALL, LONG and SHORT) τ = 1 provides the best values
of AUC in the 5-fold cross validation on the MxD dataset. Next, starting
from the least correlated feature from the selected 17 features, we removed
a given feature if it increases AUC value for the 5-fold cross validation.
By removal, we mean the removal of the feature from the current model
and the selection of the threshold, using all fully disordered
prediction models (ALL, LONG and SHORT) of the MxD dataset; these chains have similar
disorder characteristics to the chains in the CASP8. The training
chains from the MxD dataset share at most 25% identity to any
chain the CASP8 set making it challenging to perform predictions.
On the contrary, the methods we compare with large training
datasets that may share substantial similarity to the CASP8 targets
(i.e only a handful of CASP8 target were considered as free-modeling),
which could raise their predictive quality.

The cross-validated predictions of MFDp are compared to its
input predictors, including: DISOPRED2; IUPred, and JUPredS
(IUPred predictions for short and long segments); DISoclust; and
PROFBval, with and selected other recent methods including the

for each following features we calculate its average correlation (over the
current best step) we quantify predictive quality of the features selected in the first step using the linear SVMs. We construct feature sets by selecting top τ features, according to
their average biserial correlations, from each category. We take all features
in a given category if their number < τ. Next, we run 5-fold cross validation
on the MxD dataset where we vary the value of τ between 1 and 15, and for
each τ we parameterize the value of constant C of the linear SVM classifiers.
We consider C in consecutive powers of 2 between 2^{-3} and 2^7. For each of
the prediction models (ALL, LONG and SHORT) τ = 1 provides the best values
of AUC in the 5-fold cross validation on the MxD dataset. Next, starting
from the least correlated feature from the selected 17 features, we removed
a given feature if it increases AUC value for the 5-fold cross validation.
By removal, we mean the removal of the feature from the current model
and the selection of the threshold, using all fully disordered
prediction models (ALL, LONG and SHORT) of the MxD dataset; these chains have similar
disorder characteristics to the chains in the CASP8. The training
chains from the MxD dataset share at most 25% identity to any
chain the CASP8 set making it challenging to perform predictions.
On the contrary, the methods we compare with large training
datasets that may share substantial similarity to the CASP8 targets
(i.e only a handful of CASP8 target were considered as free-modeling),
which could raise their predictive quality.

The cross-validated predictions of MFDp are compared to its
input predictors, including: DISOPRED2; IUPred, and JUPredS
(IUPred predictions for short and long segments); DISoclust; and
PROFBval, with and selected other recent methods including the

3 RESULTS

3.1 Comparison with existing predictors

We compare MFDp with relevant predictors on the MxD and CASP8
datasets. For the MxD dataset, we use 5-fold cross validation to
assess our predictions; we use 75% of each of the five training
subsets to compute SVMs, the other 25% to find the threshold used
for binarization of the probabilities, and the test subsets to evaluate
predictions. In contrast, predictions from other method are generated
on the entire dataset (without cross-validation) using either web-
servers or stand-alone implementations provided by the authors.
For the test on the CASP8 dataset, the MFDp is trained, including
computation of SVMs (using the same parameters as for the MxD
dataset) and the selection of the threshold, using all fully disordered
and PDB chains in the MxD dataset; these chains have similar
disorder characteristics to the chains in the CASP8. The training
chains from the MxD dataset share at most 25% identity to any
chain the CASP8 set making it challenging to perform predictions.
On the contrary, the methods we compare with large training
datasets that may share substantial similarity to the CASP8 targets
(i.e only a handful of CASP8 target were considered as free-modeling),
which could raise their predictive quality.

The cross-validated predictions of MFDp are compared to its
input predictors, including: DISOPRED2; IUPred, and JUPredS
(IUPred predictions for short and long segments); DISoclust; and
PROFBval, with and selected other recent methods including the

where X \in \{PSSM_{AA}, SS_{SSX}, RSA, IUPREDL, IUPREDs, DISOPRED, DISoclust\} is the name of a given feature, AA with j \in \{1, 2, ..., 20\} is the amino acid type, SSX with k \in \{H, E, C\} is the type of the SS, avg_h with m \in \{3, 5, 7, 9, 11, 13, 15, 19\} is the position in the window where 0 denotes the
position of the predicted residue, and X_{avg_h} = \sum_{w_i \cap \text{cont}} X_{m}(w_i) / m
for the residue at position w_i (the actual position in the sequence equals i + \text{shift}).
When aggregating, we ignore the positions in the window that are outside
of the chain for the residues close to the sequence terminus, i.e. we sum
only the values for the positions in the chain and this sum is divided by the
total number of these positions. The features are divided into 17 categories
based on the source information used including sequence, PSSM, amino-acid
frequency, conservation, SS, RSA, torsion angles, DISOPRED2, DISoclust,
IUPREDL, IUPREDs, B-factor, signal peptides, purification tag, globular
domains, strict flexibility and non-strict flexibility. Detailed description is
provided in the appendix at http://biomine.ece.ualberta.ca/MFDp.html.

Fig. 1. Architecture of the MFDp method.

belongs to a globular domain, is part of a purification tag or is predicted
to their average biserial correlations. The set of the selected features is
ordered and disorder residues. We sort the features in each category according
to their average biserial correlations. The set of the selected features is
initialized with the feature that has the highest biserial correlation. Next,
Comparison of predictive quality measured on the MxD and CASP8 datasets when considering all disordered regions.

Following the CASP8, predictions were assessed authors), i.e. these residues are ignored when computing the quality improvements, as tested using Shapiro–Wilk test at the parameter value.

We compare predictive quality when considering all disordered regions and for the regions that are at least 30 residues long, respectively. The selection of the length threshold is consistent with values used in the recent investigation of protein–protein recognition mechanisms (Tompa et al. 2009) and in design of predictors for long segments (Han et al. 2009; Peng et al. 2006). The same as in the CASP8, we discard the disordered regions with three or fewer residues (Noivirt-Brik et al., 2009), private correspondence with authors), i.e. these residues are ignored when computing the quality measures. Similarly, for the results on the long segments we discard the shorter regions. Following the CASP8, predictions were assessed using bootstrapping in which 80% of the targets were randomly selected 1000 times and we report the corresponding averages and standard deviations see Tables 2 and 3. We recomputed the results for the CASP8 submissions and we found them consistent with the original report (Noivirt-Brik et al., 2009). We also analyze statistical significance of the differences between MFDp and the other methods. We compared the 1000 paired average results for MCC, SEN, AUC and $Sw_{01}$ obtained from using the bootstrapping with 80% of the targets on both CASP8 and MxD datasets. Since the averages do not follow normal distribution, as tested using Shapiro–Wilk test at the 0.05 significance, we use Wilcoxon rank sum test and we measure significance of the differences at 0.01 level. Tables 2 and 3 show that MFDp consistently outperforms all methods based on the MCC index. These improvements are statistically significant with $P < 0.01$ when compared with all considered competitors on both datasets and for both all and long-disordered segments, except for NORSnet and IUPREDI on the long regions and the CASP8 dataset. Results on the MxD dataset demonstrate that MFDp significantly outperforms its input methods as well as MD, NORSnet and Ucon when considering all disordered regions.

We report the averages and corresponding SDs for bootstrapping with 1000 repetitions of 80% of chains. Underlined methods are used as inputs for MFDp. The methods are sorted by the MCC values and the highest values are shown in bold. Results of tests of significance of the differences between MFDp and the other methods are given in the "significance" column. The tests compare average values from 1000 bootstrapping repetitions. The + and – mean that MFDp is statistically significantly better/worse with $P < 0.01$, and means that results are not significantly different.

| Dataset | Predictor | ACC | SEN | SPEC | AUC |
|---------|-----------|-----|-----|------|-----|
| MxD     | MFDp      | 0.44| 0.51| 0.75 | 0.81|
| MD       | MFDp      | 0.43| 0.48| 0.73 | 0.81|
| DISOPRED2 | MFDp    | 0.40| 0.44| 0.72 | 0.78|
| IUPredL  | MFDp      | 0.39| 0.42| 0.72 | 0.78|
| IUPredS  | MFDp      | 0.37| 0.38| 0.69 | 0.73|
| NORSnet  | MFDp      | 0.34| 0.37| 0.68 | 0.73|
| DISOclust | MFDp     | 0.33| 0.40| 0.72 | 0.79|
| Ucon     | MFDp      | 0.31| 0.34| 0.68 | 0.77|
| PROFcalc | MFDp      | 0.19| 0.22| 0.61 | 0.81|

| CASP8   | MFDp      | 0.61| 0.63| 0.82 | 0.89|
|---------|-----------|-----|-----|------|-----|
| MD       | MFDp      | 0.43| 0.48| 0.73 | 0.81|
| DISOPRED2 | MFDp    | 0.40| 0.44| 0.72 | 0.78|
| IUPredL  | MFDp      | 0.39| 0.42| 0.72 | 0.78|
| IUPredS  | MFDp      | 0.37| 0.38| 0.69 | 0.73|
| NORSnet  | MFDp      | 0.34| 0.37| 0.68 | 0.73|
| DISOclust | MFDp     | 0.33| 0.40| 0.72 | 0.79|
| Ucon     | MFDp      | 0.31| 0.34| 0.68 | 0.77|
| PROFcalc | MFDp      | 0.19| 0.22| 0.61 | 0.81|

We report the averages and corresponding SDs for bootstrapping with 1000 repetitions of 80% of chains. Underlined methods are used as inputs for MFDp. The methods are sorted by the MCC values and the highest values are shown in bold. Results of tests of significance of the differences between MFDp and the other methods are given in the "significance" column. The tests compare average values from 1000 bootstrapping repetitions. The + and – mean that MFDp is statistically significantly better/worse with $P < 0.01$, and means that results are not significantly different.
the binary prediction measured using MCC and $S_m$. The fact that MFDp is successful in maintaining high-quality predictions for the long segments is important given that predictors designed for long disordered regions are usually less successful in predicting short regions (Obradovic et al., 2005; Peng et al., 2005). Our method is outperformed on the $S_m$ and AUC values for the prediction of all segments by the top five CASP8 methods, but we significantly improve over these methods based on the MCC index. We also significantly improve over their binary assignments, measured with the predictors IUPred, NORSnet, DISOPRED2, Ucon and PROFbval are summarized between 0 and 0.4 on the MxD dataset and between 0.1 and 0.4 on the CASP8 dataset when compared with all predictors except the top-performing methods in the CASP8. Our method provides comparable TP rates for the FP rates $>0.2$ when compared with the top performers from the CASP8 experiment. Overall, we conclude that MFDp outperforms modern existing predictors for binary disorder predictions and provides competitive real-valued predictions.

### 3.2 Predictions of proteins with long-disordered segments

We investigate the quality of our predictions applied to find proteins with long ($\geq 30$ residues)-disordered regions. This binary prediction is motivated by the fact that this information is useful for target selection (Punta et al., 2009; Slabinski et al., 2007) and protein–protein recognition (Tompa et al., 2009). Similar evaluation was also performed for the MD (Schlessinger et al., 2009). Although $\sim$45% of disordered residues in CASP8 are in the long segments (allowing for a relatively robust per-residues evaluations), there are only 15 such segments and thus we did not use this dataset. The results on the MD dataset obtained using 5-fold cross validation for the MFDp and the predictions from the servers for MD, DISOPRED2, IUPred, NORSnet, DISOclust, Ucon and PROFbval are summarized

### Table 3. Comparison of predictive quality measured on the MxD and CASP8 datasets when considering disordered regions that are $\geq 30$-residues long

| Dataset | Predictor | MCC | $S_m$ | ACC | SENS | SPEC | AUC |
|---------|-----------|-----|-------|-----|------|------|-----|
| MxD     | MFDp      | 0.44 | ±0.02 | 0.52 | ±0.02 | 0.76 | ±0.01 |
|         | MD        | 0.44 | ±0.02 | 0.50 | ±0.02 | 0.75 | ±0.01 |
|         | IUPred    | 0.41 | ±0.02 | 0.45 | ±0.01 | 0.72 | ±0.01 |
|         | DISOPRED2 | 0.40 | ±0.02 | 0.46 | ±0.02 | 0.73 | ±0.01 |
|         | NORSnet   | 0.37 | ±0.02 | 0.41 | ±0.02 | 0.70 | ±0.01 |
|         | PROFbval  | 0.37 | ±0.01 | 0.38 | ±0.01 | 0.69 | ±0.01 |
|         | DISOclust | 0.33 | ±0.01 | 0.40 | ±0.01 | 0.70 | ±0.01 |
|         | Ucon      | 0.32 | ±0.01 | 0.37 | ±0.01 | 0.68 | ±0.01 |
|         | PROFbval  | 0.19 | ±0.01 | 0.22 | ±0.01 | 0.61 | ±0.00 |
| CASP8   | MFDp      | 0.60 | ±0.10 | 0.73 | ±0.09 | 0.86 | ±0.04 |
|         | NORSnet   | 0.59 | ±0.15 | 0.60 | ±0.15 | 0.80 | ±0.07 |
|         | DISOPRED2 | 0.58 | ±0.11 | 0.68 | ±0.10 | 0.84 | ±0.05 |
|         | 379       | 0.56 | ±0.10 | 0.71 | ±0.08 | 0.86 | ±0.04 |
|         | 297       | 0.54 | ±0.09 | 0.73 | ±0.08 | 0.87 | ±0.04 |
|         | IUPred    | 0.53 | ±0.11 | 0.59 | ±0.10 | 0.79 | ±0.05 |
|         | 79        | 0.52 | ±0.09 | 0.71 | ±0.08 | 0.85 | ±0.04 |
|         | 153       | 0.50 | ±0.09 | 0.72 | ±0.08 | 0.86 | ±0.04 |
|         | 69        | 0.44 | ±0.08 | 0.69 | ±0.07 | 0.83 | ±0.03 |
|         | DISOclust | 0.37 | ±0.07 | 0.63 | ±0.07 | 0.81 | ±0.04 |
|         | MD        | 0.36 | ±0.09 | 0.58 | ±0.11 | 0.79 | ±0.05 |
|         | Ucon      | 0.28 | ±0.10 | 0.41 | ±0.12 | 0.71 | ±0.06 |
|         | PROFbval  | 0.16 | ±0.04 | 0.31 | ±0.05 | 0.66 | ±0.03 |

We report the averages and corresponding SDs for bootstrapping with 1000 repetitions of 80% of chains. Underlined methods are used as inputs for MFDp. The methods are sorted and the significance of the differences between MFDp and the other methods are given in the significance columns. The tests compare average values from 1000 bootstrapping repetitions. The + and − mean that MFDp is statistically significantly better/worse with $P < 0.01$, and * means that results are not significantly different.
MFDp - Multilayered fusion-based disorder predictor

Fig. 2. ROCs for the predictions on the (A) MxD and (B) CASP8 datasets.

Table 4. Comparison of predictions of proteins with long (≥30 residues) disordered segments on the MxD datasets.

| Predictor | MCC  | ACC  | SENS | SPEC | AUC  |
|----------|------|------|------|------|------|
| MFDp     | 0.53 | 0.77 | 0.82 | 0.71 | 0.86 |
| Ucon     | 0.52 | 0.73 | 0.53 | 0.95 | 0.85 |
| DISOPRED2| 0.52 | 0.76 | 0.71 | 0.81 | 0.82 |
| IUPredS  | 0.52 | 0.75 | 0.63 | 0.87 | 0.83 |
| DISOclust| 0.51 | 0.74 | 0.60 | 0.89 | 0.83 |
| MD       | 0.49 | 0.74 | 0.67 | 0.81 | 0.80 |
| NORSnet  | 0.48 | 0.71 | 0.51 | 0.93 | 0.80 |
| PROFbval | 0.47 | 0.73 | 0.50 | 0.64 | 0.82 |

Underlined methods are used as inputs for MFDp. The methods are sorted by the MCC values and the highest values are shown in bold.

in Table 4. The corresponding ROC curves are shown in Figure 3 (for 0–0.5-TP-rate range) and in the appendix (for the entire range). The results demonstrate that MFDp outperforms all considered predictors as measured using MCC, ACC and AUC values. This suggests that our method is a useful resource for annotation of proteins with long disordered regions.

3.3 Case studies
We selected CASP8 target T0480 (PDBid 2K4X) which was solved with NMR, and CASP8 target T0404 (PDBid 2DFE) which was solved with X-ray crystallography, as our case studies. We compare side-by-side prediction of MFDp, its input predictors DISOPRED2, IUPred, DISOclust, the recent ensemble-based MD, and two top-performing on the CASP8 (with respect to MCC) McGuffin [379] and GeneSilicoMetaServer [297] methods (Fig. 4). For the T0480, which is a small protein with ~50% disordered residues located at sequence termini, MFDp finds both disordered segments, achieves below-average MCC of 0.59 and improves over predictions of its ensemble predictors. The MD over-predicts disorder and GeneSilicoMetaServer generates predictions comparable to that of MFDp. The T0404 includes 25% disordered residues with one longer segment away from the termini. Most of the predictors find both disordered segments, and they also incorrectly annotate another disordered segment at the N-terminus, except IUPREDL which predicts only the middle segment. The GeneSilicoMetaServer and MD slightly over-predict disorder, and the segment in the center of the chain is most accurately predicted by MFDp, DISOPRED2 and McGuffin. Although these predictions should not be assumed typical, they demonstrate that the MFDp can outperform all of its input predictors.

4 CONCLUSIONS
The MFDp predicts disordered regions of all sizes and different types. On average, it improves over the predictions of current solutions for the binary disorder assignment and provides competitive real-value predictions, as evaluated on two datasets with low-sequence similarity. The MFDp’s outputs are also shown to outperform other predictors for the identification of proteins with long-disordered regions. The main reasons for the improvements are the application of a comprehensive set of information sources including complementary disorder predictions, sequence profiles and other relevant sequence-based predictors of structural and functional protein characteristics, novel two-layer architecture and the usage of custom features that aggregate raw prediction inputs.
Fig. 4. Comparison of predictions from MFDp, DISOPRED2 (DP2), IUPRED (IUP), IUPRED2 (IUPS) and DISOclust (DISO), MD, and two CASP8 predictors with the highest MCC, McGuffin (379) and GeneSilicoMetaServer (297) for CASP8 targets T0480 (on the left) and T0404 (on the right). The ‘–’ and ‘D’ denote the ordered and disordered residues, respectively. The actual disorder annotations are shown in the first line.

Funding: NSERC Canada (to L.K.); Icore/Alberta Ingenuity Scholarships (to W.S., K.C. and K.D.K.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. et al. (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. *Bioinformatics*, 23, 3433–3434.

Ambroisi,L et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402.

Berman,H.M et al. (2000) The Protein Data Bank. *Nucleic Acids Res.*, 28, 235–242.

Bordoli,L et al. (2007) Assessment of disorder predictions in CASP7. *Proteins*, 69 (Suppl. 8), 129–136.

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

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

Dunker,A.K et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402.

Hecker,J et al. (2008) Protein disorder prediction at multiple levels of sensitivity and specificity. *BMC Genomics*, 9 (Suppl. 1), S9.

Hoos,S et al. (2007) POODLE-L: A two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics*, 23, 2046–2053.

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

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

Jones,D.T and Swindells,M.B. (2002) Getting the most from PSI-BLAST. *Trends Biochem. Sci.*, 27, 161–164.

Jones,D.T and Ward,J.L. (2003) Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, 53 (Suppl. 6), 573–578.

Kondrashov,AS et al. (2005) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, 33, 3701–3708.

Kumar,M et al. (2008) Intermolecular disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, 24, 1798–1804.

McGuffin,L.J et al. (2000) The PSSPREDD protein structure prediction server: *Bioinformatics*, 16, 404–405.

Naray-Bink,O et al. (2009) Assessment of disorder predictions in CASP8. *Proteins*, 77 (Suppl. 9), 210–216.

Obradovic,Z et al. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, 64, 176–182.

Ollier,J et al. (2005) Comparing and combining predictors of mostly disordered proteins. *Bioinformatics*, 21, 2989–2990.

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

Plewczynski,D et al. (2008) Prediction of signal peptides in protein sequences by neural networks. *Acta Biochim. Pol.*, 55, 261–267.

Pilia,G et al. (2005) Foldsfinder: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, 21, 3435–3438.

Punta,M et al. (2009) Structural genomics target selection for the New York consortium on membrane protein structure: I. Struct. Func. Genom., 10, 255–266.

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

Radivojac,P et al. (2007) Intrinsic disorder and functional proteins. *Biochim. Biophys. Acta*, 1774, 1493–1456.

Raychoudhuri,S et al. (2009) The role of intrinsically unstructured proteins in neurodegenerative disease. *PLoS One*, 4, e5656.

Schönherr,S et al. (2006) PROFUnfold: predict flexible and rigid residues in proteins. *Bioinformatics*, 22, 891–893.

Schönherr,S et al. (2007a) Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, 23, 2376–2384.

Schönherr,S et al. (2007b) Natively unstructured loops differ from other loops. *PLoS Comput. Biol.*, 3, e110.

Schönherr,S et al. (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One*, 4, e4453.

Shiina,K et al. (2007a) POODLE-L: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. *Bioinformatics*, 23, 2376–2384.

Shiina,K et al. (2007b) Predicting mostly disordered proteins by using structure-unknown protein data. *Bioinformatics*, 8, 78.

Sickmeier-M et al. (2007) UniProt: the Database of Disordered Proteins. *Nucleic Acids Res.*, 35, D768–D770.

Slijepcevic,P et al. (2007) The challenge of protein structure determination - lessons from structural genomics. *Prot. Sci.*, 16, 2472–2482.

Su.C.T et al. (2007) PDIA: integrated protein disorder analyzer. *Nucleic Acids Res.*, 35, 465–472.

Su.C.T et al. (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *Bioinformatics*, 7, 319.

Tomp,P et al. (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. *Biosew., 31, 328–335.

Uversky,VN et al. (2000) Why are “natively unfolded” proteins unstructured under physiologic conditions? *Prot. Sci.*, 41, 415–427.

Vaccari,S et al. (2003) Flavors of protein disorder. *Prot. Sci.*, 82, 573–584.

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

Wang,G and Dunbrack,R.L. Jr. (2003) PSICEN: a protein sequence calling server. *Bioinformatics*, 19, 1589–1591.

Wang,G and Strauss,U.H. (2008) OnD-CRF: predicting order and disorder in proteins using conditional random fields. *Bioinformatics*, 24, 2138–2139.

Wu,S and Zhang,Y (2008) MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, 72, 547–556.

Yang,M.Q and Yang,Y.J. (2006) IUP: intrinsically unstructured protein predictor – a software tool for analyzing polyepitope sequences. In Sixth IEEE Symposium on Bioinformatics and Bioengineering. IEEE-Computer Society, Washington DC, USA, pp. 16–18.

Zhang,M et al. (2009) On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins*, 76, 617–636.