High-throughput prediction of disordered moonlighting regions in protein sequences

Fanchi Meng[1] and Lukasz Kurgan[2,1*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, T6G 2V4, Canada.
[2]Department of Computer Science, Virginia Commonwealth University, Richmond, 23284, U.S.A.

*Corresponding author: lkurgan@vcu.edu; 804-827-3986

Abstract

Intrinsically disordered regions lack stable structure in their native conformation but are nevertheless functional and highly abundant, particularly in Eukaryotes. Disordered moonlighting regions (DMRs) are intrinsically disordered regions that carry out multiple functions. DMRs are different from moonlighting proteins that could be structured and that are annotated at the whole-protein level. DMRs cannot be identified by current predictors of functions of disorder that focus on specific functions rather than multifunctional regions. We conceptualized, designed and empirically assessed first-of-its-kind sequence-based predictor of DMRs, DMRpred. This computational tool outputs propensity for being in a DMR for each residue in an input protein sequence. We developed novel amino acid indices that quantify propensities for functions relevant to DMRs and used evolutionary conservation, putative solvent accessibility and intrinsic disorder derived from the input sequence to build a rich profile that is suitable to accurately predict DMRs. We processed this profile to derive innovative features that we input into a Random Forest model to generate the predictions. Empirical assessment shows that DMRpred generates accurate predictions with area under receiver operating characteristic curve=0.86 and accuracy=82%. These results are significantly better than the closest alternative approaches that rely on sequence alignment, evolutionary conservation and putative disorder and disorder functions. Analysis of abundance of putative DMRs in the human proteome reveals that as many as 25% of proteins may have long (>30 residues) DMRs. A webserver implementation of DMRpred is available at http://biomine.cs.vcu.edu/servers/DMRpred/

Keywords: Intrinsic disorder; moonlighting regions; prediction; human proteome; protein function.

Introduction

Intrinsically disordered regions (IDRs) lack stable structure in their native conformation[1]. Proteins with IDRs are prevalent in nature. According to some estimates around 20% of residues in Eukaryotic proteins are disordered[2,3] and about half of human proteins have at least one long (>30 consecutives residues) IDR[4,5]. Proteins with IDRs carry out numerous functions that rely on protein-protein and protein-nucleic acids interactions (e.g., translation, transcription, and chromosome condensation), are involved in a variety of signaling functions, and facilitate regulation of protein

functions via posttranslational modifications[2,6-14]. Substantial efforts have been made to predict and computationally characterize IDRs[15-17]. There are over 40 methods that predict IDRs and some of them were empirically shown to provide very accurate predictions[18-22]. Moreover, progress has been made in recent years to predict functions of intrinsically disordered regions[15]. Example predictors include Anchor[23], MoRFpred[24], fMoRFpred[25], and MoRFCHiBi[26] that predict disordered protein-protein binding regions, DFLpred[27] that outputs putative disordered linkers, and DisoRDPbind[28,29] that predicts disordered protein-protein, -RNA and -DNA binding regions. High degree of plasticity of IDRs allows them to bind more than one ligand and carry out multiple functions[30-32]. According to our estimates in Ref. [15], about 37% of the functionally annotated IDRs in the DisProt database[33] perform multiple functions.

Multifunctional (moonlighting) proteins were reviewed and discussed in Ref. [34]. A moonlighting protein is a single polypeptide chain that has multiple autonomous and unrelated functions that cannot be simply associated with separate domains[35]. Example mechanisms that lead to the moonlighting activities include interactions with multiple ligands, presence of multiple oligomerization states, and expression in different cell types and cellular locations[34,36]. Whereas the multifunctionality of DMRs stems from their high degree of plasticity that allows a single IDR to bind multiple ligands, serve as a linker and/or perform entropic functions[30-32]. The moonlighting proteins can be predicted computationally from sequences and other information about these proteins, such as protein-protein interactions and gene expression profiles [35-38]. However, these methods make predictions only at the protein level, not at the residue or sequence region level that is necessary to identify the multifunctional IDRs. Here, we use the term "moonlighting" to describe regions in the protein chain that perform more than one function. The disordered moonlighting regions (DMRs) are different from moonlighting proteins since they concern regions rather than complete protein chains and since they focus specifically on the intrinsic disorder. Recent research has revealed many examples of diverse and functionally important DMRs[39-42]. While many DMRs can be found in the DisProt database, a significantly larger number of these regions is awaiting to be discovered. Computational prediction of DMRs is necessary to find these regions among the large and growing number of IDRs.

Current predictors of functions of IDRs do not predict DMRs and current methods that predict moonlighting proteins cannot be applied at the region level. To this end, we propose first-of-its-kind methods that accurately predicts DMRs from protein sequences, DMRpred. DMRpred aims to separate DMRs from other types of regions including monofunctional IDRs and structured regions. DMRpred uses a sophisticated predictive model to generate a numeric propensity for each residue being in a DMR. The particularly innovative aspects of this work are: 1) development of a new dataset of proteins with DMRs; 2) design and use of novel scales that quantify propensities of amino acids for functions that are relevant to DMRs; 3) use of an original approach to build predictive inputs that aggregate structural and functional characteristics based on putative IDRs; and 4) inclusion of the first attempt to quantify DMRs in the human proteome.

Table I. Annotations of functions for disordered regions in DisProt 7.0.3. We exclude posttranslational modifications since these are not intrinsic functions of IDRs. They are located in IDRs and DisProt does not provide their exact positions.

Level 1	Level 2
Entropic chain	Flexible linker/spacer
	Entropic bristle
	Entropic clock
	Entropic spring
	Structural mortar
	Self-transport through channel
Molecular recognition – assembler	Assembler
	Localization (targeting)
	Localization (tethering)
	Prions (self-assembly, polymerization)
	Liquid-liquid phase separation/demixing (self-assembly)
Molecular recognition – scavenger	Neutralization of toxic molecules
	Metal binding/metal sponge
	Water storage
Molecular recognition – effectors	Inhibitor
	Disassembler
	Activator
	cis-regulatory elements (inhibitory modules)
	DNA bending
	DNA unwinding
Molecular recognition – display site	Limited proteolysis
Molecular recognition – chaperone	Protein detergent/solvate layer
	Space filling
	Entropic exclusion
	Entropy transfer

Materials and methods

Datasets and annotation of DMRs

The data comes from two sources: DisProt[43] and Protein Data Bank (PDB)[44]. We use DisProt 7.0.3 to collect disordered proteins and extract annotations of DMRs. We use PDB to collect structured proteins that are necessary to ensure that our model does not predict DMRs for them. After removing 10 proteins from DisProt that have incorrect annotations (e.g., annotations out of bounds of protein chains) we parsed the remaining 693 proteins. They include 2,108 disordered regions with length ranging between 5 and 2,400 residues.

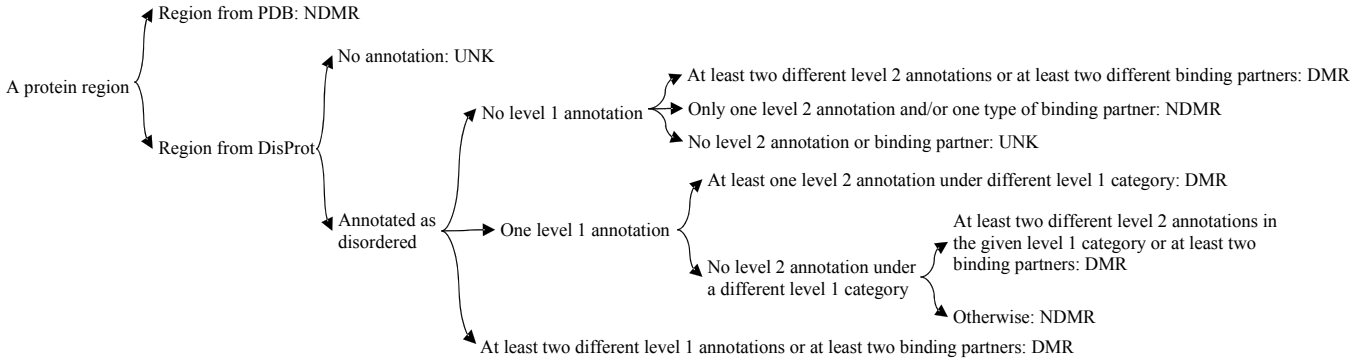

Figure 1. Flow chart to define disordered moonlighting regions based on the functional annotations given in Table I and binding partner annotations defined in DisProt 7.0.3.

DisProt 7.0.3[33] has two levels of functional annotations (Table I) which are separate from the binding partner annotations. We define DMR as a disordered region that has at least two distinct functions. We use the hierarchy of the functional annotations in DisProt to ensure that the functions used to annotate DMRs are distinct. We ensure that each DMR has at least two different level 1 annotations, at least two different level 2 annotations, or one level 1 annotation with at least one level 2 annotation under different level 1 category. Moreover, DMRs also include disordered regions that are annotated to have at least two different types of binding partners (e.g., DNA and protein). The annotations of binding partners include protein-protein, protein-DNA, protein-RNA, protein-lipid, protein-metal, protein-inorganic salt and protein-small molecule binding. Our annotation protocol does not mix the annotations of functions and binding partners in order to secure a conservative set of multifunctional regions, i.e., a combination of a functional annotation and an annotation of a binding partner is not used to annotate DMRs. Detailed annotation protocol is shown in Figure 1. For a given disordered region in DisProt, we consider this region as one of the three classes:

1 a disordered moonlighting regions (DMR),
2 a non-disordered moonlighting region (NDMR) that includes monofunctional disordered regions (that have a known function) and structured regions, and
3 a region of unknown type (UNK).

UNK regions include disordered regions without functional annotation and regions in DisProt without any annotations. Residues in the UNK regions are excluded from our analysis. We do not use them to neither build nor assess the model. We include structured proteins collected from the Protein Data Bank (PDB)[45] and residues from these proteins are annotated as the NDMR residues. Figure 1 explains how DMR residues, NDMR residues and UNK residues are annotated. For example, the protein region annotated using the path at the bottom of the Figure 1 has at least two different functional annotations that belong to different level 1 categories, and thus it is annotated as a DMR. We note that some of the monofunctional IDRs from our dataset could be re-labeled as DMRs in the future as additional and different functional annotations are collected. We accommodate for that by setting the evaluation criteria to allow for a small amount of false positives (non-DMRs predicted as DMRs).

We define DMR residues and NDMR residues as residues that are in DMRs and NDMRs, respectively. We include all proteins that have annotated DMRs and proteins with residues annotated as either DMR or NDMR (both monofunctional and structured); we exclude proteins that contain only unknown annotations. Consequently, we select 139 out of the 693 proteins from DisProt that have 12,910 DMR residues. We also collect high-resolution structured monomer proteins from PDB using the following criteria: chain length $\geq$ 30 residues; resolution $\leq$ 2.0 Å; number of chains (asymmetric unit) = 1; number

of chains (biological assembly) = 1, and number of entities = 1. We collected 2,927 such monomers in February 2017. We filter out proteins that have non-standard amino acids (AAs) or disordered residues (missing residue or marked as REMARK 465). This ensures that the selected proteins contain only standard AAs and are structured. Next, we select a representative subset of these proteins that share low sequence similarity. We run BlastClust[46] with the length coverage > 70% and the identity threshold = 25%. We pick one representative sequence from each of the 298 clusters to ensure that remaining proteins share low similarity. To balance the number of disordered and structured proteins, we randomly select 139 proteins from the set of 298 structured proteins. We combine the two sets of 139 proteins to form the dataset of 278 proteins. We divide these 278 proteins at random into two subsets of equal size, a training dataset that we use to design and parameterize the predictive model, and a test dataset to perform blind validation. We further subdivide the training dataset into four equally sized subsets (12.5% of the original dataset) to perform four-fold cross validation. We ensure that the training and test datasets as well as the four cross-validation folds share sequence identity below 25%. To do that, we run BlastClust on the 278 proteins, using the same parameters as above, and we place each of the resulting 263 clusters that include similar sequences (≥ 25% identity) into one of the five protein sets that is chosen at random. The first four subsets (12.5% of the original dataset) constitute the four folds of the training set and the remaining fifth subset (50% of the original dataset) is used as the test dataset. We ensure that each of the five subsets has similar ratio of DMR to NDMR residues by randomly resampling clusters, if needed. The training dataset (with annotated cross-validation folds) and test dataset are available at http://biomine.cs.vcu.edu/servers/DMRpred/. The training (test) dataset includes 140 (138) proteins with 6,261 (6,649) DMR residues and 16,466 (17,449) NDMR residues; the latter set of residues includes structured and monofunctional disordered residues. We did not balance the number of DMR and NDMR residues to ensure that the predictive model inferred from these data does not overpredict DMRs, i.e., it should predict a small fraction of residues as DMRs.

Evaluation criteria

The prediction is a numeric score between 0 and 1 that represents propensity for a given residue to be part of a DMR. The score can be also converted into a binary prediction using a threshold. A residue with a putative score greater than or equal to a given threshold is predicted as part of a DMR, otherwise it is predicted as part of a NDMR. We assess the predictive quality of the putative propensities with the receiver operating characteristic (ROC) curve and the area under ROC (AUC_R). To plot the ROC curves and quantify AUC_R values, we calculate true-positive rates (TPRs) and false-positive rates (FPRs) by comparing binary predictions with native annotations at different thresholds imposed on the predicted scores:

$$TPR = TP/(TP + FN) = TP / number\_of\_DMR\_residues \quad (1)$$

$$FPR = FP/(FP + TN) = FP/ number\_of\_NDMR\_residues \quad (2)$$

where TP is the number of true positives (correctly predicted residues in DMRs), FN is the number of false negatives (predicted incorrectly residues that are part of DMRs), FP is the number of false positive (incorrectly predicted residues that are part of NDMRs), and TN is the number of true negatives (correctly predicted residues in NDMRs). Given TPR and FPR values generated at different thresholds ranging from 0 and 1, we plot the ROC curve and calculate the corresponding AUC_R value. Moreover, we plot the precision-recall curves and calculate the area under the precision-recall curve (AUC_{PR}). Recall and precision are defined as equation (5) and (4), respectively.

Motivated by the fact that vast majority of the residues are located in NDMRs (they are "negatives") we perform assessment of the propensities when FPR is low, at or below 5%. The 5% value accommodates for the possibility that some of the monofunctional IDRs could be in fact multifunctional because

additional, different functions of these regions are not yet determined. This also ensures that the corresponding predictions include putative DMR residues which are likely correctly predicted, i.e., only a small fraction of these predictions are false positives. Correspondingly, we calculate $AUC_{RlowFPR}$ that covers the low range of FPR values between 0 and 0.05. Since $AUC_{RlowFPR}$ values are rather small and difficult to assess directly, we compute $AUC_{Rratio} = AUC_{RlowFPR}/AUC_{Rrandom\_lowFPR}$, where $AUC_{RlowFPR}$ is divided by the AUC_R of a random predictor (for which FPR always equals to TPR) in the same FPR range. This ratio quantifies the rate of improvement over a random predictor, i.e., ratio > 1 means that a given method is better than random and ratio = 2 means that this method is twice better than random.

We also assess the binary predictions defined using the threshold that results in FPR = 5%. For the binary predictions, we use accuracy, precision, recall and Matthews Correlation Coefficient (MCC):

$$\text{Accuracy} = (TP + TN)/\text{number\_of\_all\_residues} \quad (3)$$

$$\text{Precision} = TP/\text{number\_of\_all\_predicted\_DMR\_residues} \quad (4)$$

$$\text{Recall} = TP/\text{number\_of\_all\_native\_DMR\_residues} \quad (5)$$

$$MCC = (TP \times TN - FP \times FN)/\text{sqrt}((TP + FP)(TP + FN)(TN + FP)(TN + FN)) \quad (6)$$

The AUCR ranges between 0.5 and 1 where 0.5 denotes random prediction and 1 denotes perfect prediction. Accuracy, precision and recall range between 0 and 1 where 0 denotes that no residues were predicted correctly and 1 denotes perfect prediction. MCC ranges between -1 and 1, where -1 denotes that inverted prediction (all DMR residues are predicted as NDMR residues and vice versa), 0 denotes a random result and 1 denotes a perfect prediction.

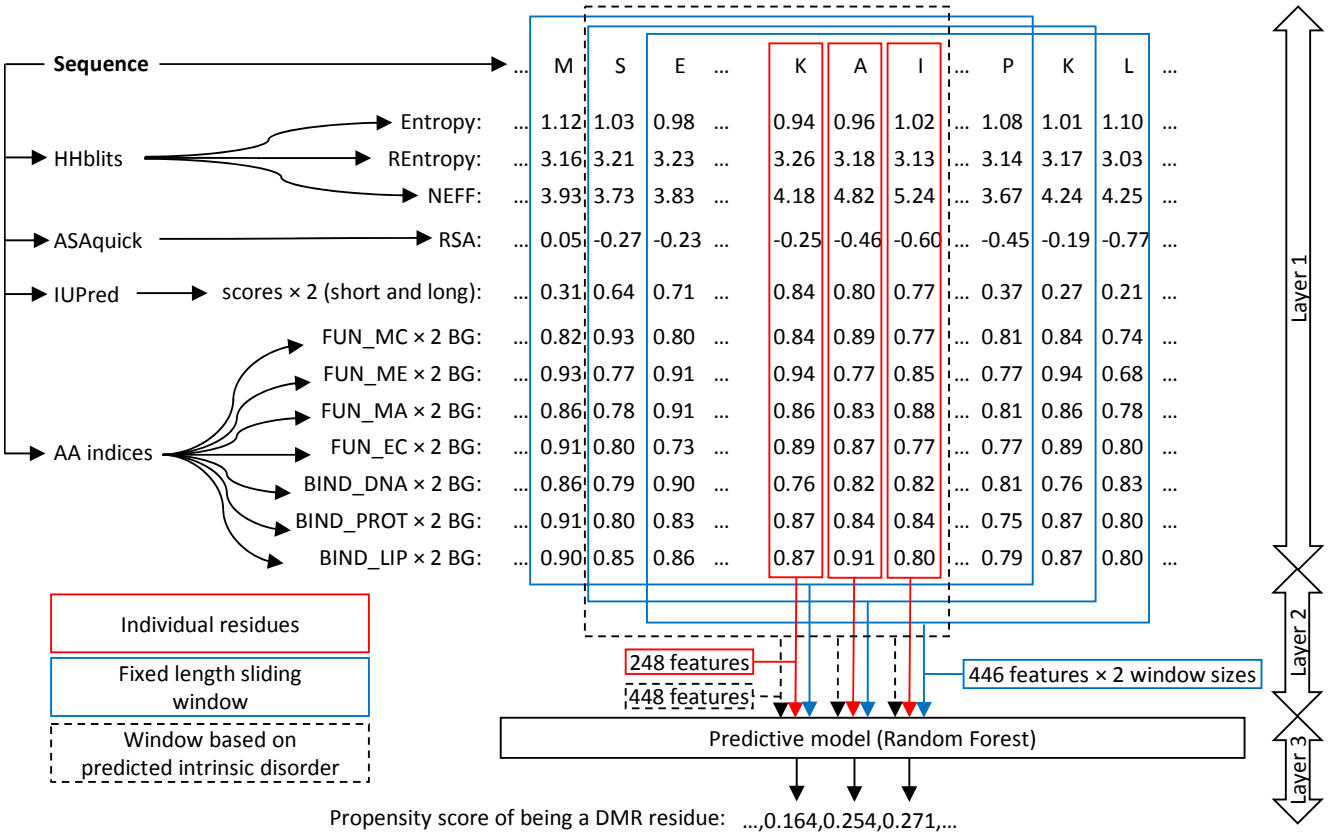

Figure 2. Architecture of DMRpred.

Architecture of DMRpred

The architecture of DMRpred (Figure 2) includes three layers:

1. *Sequence profile*: we represent the input sequence by a set of numeric values that quantify biophysical and structural properties of residues in this sequence.
2. *Feature representation*: for each residue in the input protein we convert the profile into a set of features that quantify relevant properties for this residue and its neighbors in the sequence.
3. *Prediction*: The features are input into a predictive model that generates the propensities for residues to be in DMRs.

Sequence profile

We consider several relevant biophysical and structural properties to define the sequence profile. They include sequence conservation, relative solvent accessibility, intrinsic disorder and a set of novel AA indexes (Figure 2). The indices quantify propensity of individual AA types to carry out functions that are relevant to DMRs.

We compute conservation from the multiple sequence alignment produced with HHblits[47]. HHblits is based on profile – profile alignments computed with the hidden Markov models. It was shown in Ref. [47] to be faster and more sensitive than the sequence-based alignment with PSI-BLAST[48]. To further reduce the runtime, we run HHblits against the Pfam database (as of February 2017), instead of the default UniProt20 database, and we iterate twice. Running HHblits against Pfam database is 12 times faster when compared to using UniProt20; average per protein runtime is 8 seconds vs. 102 seconds. Using the outputs of HHblits, we quantify the conservation in three ways: entropy, relative entropy (REntropy) and the local diversity (NEFF). The entropy is calculated using the 20 AA emission frequencies and relative entropy is calculated by considering the HHblits null model frequencies as the background frequency[49]. The NEFF(*i*) output from HHblits measures the diversity of sub-alignment for residue *i* that contains all sequences that have a residue at position *i* of the full alignment. A smaller entropy, larger relative entropy and smaller NEFF indicate a more conserved residue. We invert the values of entropy and NEFF by subtracting their values from the corresponding maximal value. This makes these values consistent with the other properties, where a larger number indicates a more conserved residue that has a higher chance to carry out function(s) relevant to DMRs.
We calculate the relative solvent accessibility (RSA) with ASAquick[50]. ASAquick predicts the relative accessible surface area from a single sequence (without alignment). ASAquick is orders of magnitude faster than most of the other predictors of RSA that require multiple sequence alignment. It produces prediction in less than a second for a protein that is 500 residues long. We normalize the output of ASAquick to the 0 to 1 range where a larger number means that the corresponding residue is more solvent exposed.

Intrinsic disorder is predicted with IUPred[51,52]. We selected this particular method since it is fast, was ranked as one of the top methods in several recent benchmarks[15,19,21] and was utilized in several other disorder function predictors, such as DFLpred[27] and DisoRDPbind[28]. We use both the short region and long region versions of IUPred. The output of IUPred ranges from 0 to 1 and larger numbers suggest a higher likelihood for the intrinsic disorder.

Table II. Propensities that quantify enrichment or depletion of amino acid types for specific functions/biding partners computed when using all residues in the training dataset as the background. Columns correspond to functions/binding partners where FUN_MC: molecular recognition – chaperone; FUN_ME: molecular recognition – effectors; FUN_MA: molecular recognition – assembler; FUN_EC: entropic chain; BIND_DNA: protein-DNA binding; BIND_PROT: protein-protein binding; and BIND_LIP: protein-lipid binding. For each amino acid, we list its fractional difference (FD) value defined based on the difference in composition between the query sample (residues in a given functional region) and the background sample. Positive FD values indicate enrichment and negative value indicates depletion. Statistical significance of the fractional difference is quantified with the *p*-values; *p*-value < 0.01 is considered statistically significant. Bold font shows amino acids for which the FD values are significantly different.

Amino	FUN_MC		FUN_ME		FUN_MA		FUN_EC		BIND_DNA		BIND_PROT		BIND_LIP	
Acid	FD	*p*-value	FD	*p*-value	FD	*p*-value	FD	*p*-value	FD	*p*-value	FD	*p*-value	FD	*p*-value
A	-0.119	0.205	**0.420**	**0.000**	0.126	0.115	-0.032	0.531	**0.211**	**0.003**	0.111	0.030	-0.208	0.040
C	0.119	0.561	-0.391	0.068	-0.054	0.786	**-0.669**	**0.000**	**-0.395**	**0.003**	**-0.447**	**0.000**	-0.489	0.031
D	-0.081	0.603	0.164	0.225	0.066	0.509	-0.192	0.098	**0.247**	**0.002**	0.119	0.037	**0.400**	**0.000**
E	**0.723**	**0.000**	0.017	0.920	0.048	0.647	**1.109**	**0.000**	0.073	0.284	**0.536**	**0.000**	**0.334**	**0.001**
F	-0.228	0.198	**-0.514**	**0.001**	-0.164	0.191	**-0.579**	**0.000**	**-0.409**	**0.000**	**-0.419**	**0.000**	-0.073	0.716
G	-0.202	0.048	**0.376**	**0.000**	**0.233**	**0.003**	0.042	0.610	-0.143	0.029	**0.129**	**0.007**	**0.473**	**0.000**
H	0.411	0.028	0.138	0.517	0.089	0.597	-0.304	0.072	0.007	0.941	0.164	0.061	0.584	0.011
I	-0.175	0.222	**-0.388**	**0.002**	**-0.449**	**0.000**	-0.178	0.148	**-0.307**	**0.001**	**-0.359**	**0.000**	-0.244	0.106
K	**0.605**	**0.000**	-0.076	0.386	**0.445**	**0.000**	**0.292**	**0.005**	**1.123**	**0.000**	**0.388**	**0.000**	**0.370**	**0.001**
L	-0.089	0.385	0.075	0.415	-0.195	0.026	**-0.236**	**0.004**	**-0.326**	**0.000**	**-0.238**	**0.000**	-0.250	0.021
M	0.004	0.912	-0.426	0.017	-0.168	0.225	-0.355	0.022	-0.160	0.176	**-0.372**	**0.000**	-0.324	0.066
N	-0.242	0.087	**-0.352**	**0.004**	-0.203	0.052	-0.183	0.144	**-0.269**	**0.003**	**-0.318**	**0.000**	-0.385	0.012
P	0.128	0.296	**0.293**	**0.006**	0.136	0.128	0.305	0.016	0.144	0.074	**0.372**	**0.000**	0.207	0.064
Q	-0.170	0.201	0.155	0.296	-0.184	0.095	**0.358**	**0.002**	-0.148	0.091	0.059	0.306	0.003	0.897
R	0.252	0.157	0.142	0.360	0.237	0.039	0.011	0.987	**0.316**	**0.000**	0.025	0.752	**-0.283**	**0.008**
S	**-0.338**	**0.002**	**0.334**	**0.001**	**0.284**	**0.001**	0.210	0.015	**0.246**	**0.000**	**0.218**	**0.000**	-0.013	0.816
T	0.006	0.942	**-0.310**	**0.007**	-0.123	0.188	-0.033	0.658	**-0.216**	**0.005**	-0.095	0.075	-0.133	0.212
V	-0.002	0.930	**-0.433**	**0.000**	**-0.317**	**0.001**	**-0.290**	**0.005**	**-0.218**	**0.004**	**-0.310**	**0.000**	-0.323	0.013
W	-0.318	0.189	-0.493	0.036	**-0.623**	**0.002**	-0.469	0.022	**-0.724**	**0.000**	**-0.538**	**0.000**	0.021	0.751
Y	**-0.552**	**0.002**	-0.324	0.061	-0.143	0.364	**-0.609**	**0.000**	**-0.458**	**0.000**	**-0.458**	**0.000**	-0.064	0.733

Table III. Propensities that quantify enrichment or depletion of amino acid types for specific functions/biding partners computed when using disordered residues in the training dataset as the background. Columns correspond to functions/binding partners where FUN_MC: molecular recognition – chaperone; FUN_ME: molecular recognition – effectors; FUN_MA: molecular recognition – assembler; FUN_EC: entropic chain; BIND_DNA: protein-DNA binding; BIND_PROT: protein-protein binding; and BIND_LIP: protein-lipid binding. For each amino acid, we list its fractional difference (FD) value defined based on the difference in composition between the query sample (residues in a given functional region) and the background sample. Positive FD values indicate enrichment and negative value indicates depletion. Statistical significance of the fractional difference is quantified with the *p*-values; *p*-value < 0.01 is considered statistically significant. Bold font shows amino acids for which the FD values are significantly different.

Amino	FUN_MC		FUN_ME		FUN_MA		FUN_EC		BIND_DNA		BIND_PROT		BIND_LIP	
Acid	FD	*p*-value	FD	*p*-value	FD	*p*-value	FD	*p*-value	FD	*p*-value	FD	*p*-value	FD	*p*-value
A	-0.221	0.035	0.257	0.010	-0.001	0.872	-0.147	0.096	0.068	0.361	-0.018	0.834	**-0.299**	**0.005**
C	**3.385**	**0.000**	**1.368**	**0.004**	**2.626**	**0.000**	0.279	0.523	**1.352**	**0.001**	**1.136**	**0.001**	1.013	0.031
D	-0.220	0.085	-0.012	0.744	-0.097	0.333	**-0.314**	**0.004**	0.058	0.580	-0.051	0.413	0.179	0.076
E	0.154	0.076	**-0.317**	**0.001**	**-0.292**	**0.000**	**0.417**	**0.000**	**-0.275**	**0.000**	0.036	0.585	-0.099	0.533
F	-0.067	0.783	-0.418	0.017	0.007	0.963	**-0.496**	**0.002**	-0.287	0.029	**-0.300**	**0.004**	0.105	0.514
G	**-0.405**	**0.000**	0.022	0.548	-0.081	0.477	-0.220	0.010	**-0.362**	**0.000**	**-0.161**	**0.004**	0.096	0.789
H	-0.016	0.990	-0.208	0.172	-0.243	0.089	**-0.512**	**0.000**	-0.300	0.012	-0.190	0.051	0.097	0.834
I	0.257	0.188	-0.067	0.611	-0.165	0.246	0.254	0.130	0.051	0.645	-0.026	0.817	0.152	0.347
K	-0.023	0.848	**-0.432**	**0.000**	-0.116	0.107	-0.213	0.012	**0.301**	**0.000**	**-0.148**	**0.007**	-0.163	0.116
L	0.379	0.011	**0.623**	**0.000**	0.214	0.060	0.148	0.282	0.014	0.885	0.148	0.079	0.133	0.272
M	0.078	0.782	-0.385	0.058	-0.106	0.515	-0.307	0.086	-0.092	0.544	**-0.325**	**0.005**	-0.269	0.168
N	0.466	0.045	0.243	0.342	0.526	0.011	**0.576**	**0.004**	0.403	0.011	0.308	0.023	0.188	0.260
P	0.093	0.505	0.254	0.037	0.099	0.328	0.256	0.082	0.106	0.300	**0.330**	**0.000**	0.171	0.171
Q	-0.153	0.274	0.172	0.301	-0.169	0.167	**0.386**	**0.006**	-0.133	0.202	0.075	0.382	0.022	0.982
R	0.225	0.254	0.113	0.499	0.208	0.108	-0.011	0.895	**0.288**	**0.008**	-0.001	0.999	-0.310	0.010
S	**-0.380**	**0.001**	0.255	0.014	0.209	0.029	0.141	0.132	0.171	0.036	0.146	0.031	-0.069	0.479
T	0.172	0.227	-0.197	0.164	0.019	0.886	0.127	0.369	-0.090	0.414	0.054	0.510	0.014	0.964
V	**0.478**	**0.004**	-0.153	0.241	0.010	0.980	0.052	0.673	0.161	0.191	0.025	0.797	0.008	0.786
W	0.528	0.306	0.103	0.793	-0.170	0.619	0.183	0.768	-0.388	0.144	0.020	0.997	**1.255**	**0.003**
Y	-0.283	0.244	0.071	0.647	0.356	0.075	-0.378	0.048	-0.144	0.383	-0.140	0.292	0.481	0.051

A unique to DMRpred part of the profile is the propensity of AAs to carry out functions that are relevant to DMRs. We quantify these AA indices with Composition Profiler[53]. The indices measure enrichment or depletion of specific AA types in the corresponding functional IDRs. First, we extract all functional IDRs from the training dataset. We consider the functions that we use to define DMRs and that have at least 1000 residues; the latter ensures that we have enough data for statistical analysis. We cover seven functions: molecular recognition – chaperone (FUN_MC), molecular recognition – effectors (FUN_ME), molecular recognition – assembler (FUN_MA), entropic chain (FUN_EC), protein-DNA binding (BIND_DNA), protein-protein binding (BIND_PROT) and protein-lipid binding (BIND_LIP). For each of the seven functions we use the corresponding regions as a query to run the Composition Profiler. The Profiler compares a given query to a background. We consider two types of background (BG): all residues and disordered residues from the training dataset. The former type of background results in the computation of differences between a specific set of functional residues and a generic set of all AAs. The latter type focuses on the differences between a specific set of functional residues, which are disordered, and a set of all disordered AAs. For each of the 20 AAs, the Composition Profiler outputs a fractional difference of the composition between the query and the background. Positive (negative) fractional differences indicate enriched (depleted) AAs. The Profiler also outputs *p*-values

that measure statistical significance of the fractional differences. We consider p-value < 0.01 as statistically significant. Tables II and III provide the fractional differences for the 20 AA types, the seven functions and two backgrounds.

Feature representation

Using the sequence profile, we empirically generate a rich set of features to represent every residue in the input sequence. The features quantify information about individual biophysical and structural properties and their combinations, e.g., we combine conservation and solvent accessibility.
We generate features for each residue by considering the information about the residue itself and its neighbors in the sequence. The use of the neighboring residues is inspired by the fact that the disordered moonlighting residues form regions composed of consecutive AAs that share certain functional and structural properties. We define neighbors using two types of sequence windows: a sliding window of a fixed length (defined based on size of native DMRs in the training dataset) centered on the residue that we currently predict; and the putative disordered regions (disordered window) that includes this residue (Figure 2). To the best of our knowledge, we are the first to use the latter window type. We do not pad windows for residues at the termini of the sequence and accordingly the features are normalized by the length of the window. The length of the second type of the windows varies and is determined by the length of the putative disordered regions generated with IUPred_short and IUPred_long. The use of the fixed size sliding windows is motivated by the design of related methods, such as MoRFpred[24], fMoRFpred[25], DisoRDPbind[28] and DFLpred[27]. Using the individual and combined biophysical and structural properties that are quantified for individual residues and based on the two types of windows, we compute 1588 features for each residue in the input protein chain. A detailed description of these features can be found in the Supplement.

Design of the predictive model

We use the feature vector for each of the 22,727 residues in the training dataset to generate a predictive model using a machine learning algorithm. This model outputs a propensity score that a given residue is a DMR residue. We consider three algorithms: Logistic regression[54], Naive Bayes[55] and Random Forest[56] using their implementations in the Weka platform[57].

We conduct feature selection for the logistic regression and Naive Bayes algorithms. Random Forest algorithm automatically selects features when building the trees. We use the best-first search to implement the selection. First, we calculate the AUC_R values when using individual features to make predictions on each of the four training folds, and we rank the features by their averaged (over the four training folds) AUC_R values. We run the 4-folds cross validation on the training dataset using the logistic regression and Naive Bayes with the top-ranked feature to initialize the set of selected features. Next, we add the next-ranked feature to the current feature set if this results in a higher average AUC_R than the AUC_R obtained before the feature was added. We scan the sorted feature set once.

Results

Selection of the DMRpred's predictive model

We parameterize the three considered algorithms and compare their predictive performance using the training dataset to select the model that offers favorable predictive quality. Unlike logistic regression and Naive Bayes algorithms that do not require parametrization, we perform a grid search to find optimal

parameters for the Random Forest. We select parameters that result in the highest AUC_R measured with the 4-fold cross validation on the training dataset. Based on suggestions from Ref. [58], we consider the number of trees = $\{2^7, 2^8, 2^9, 2^{10}\}$, the number of features randomly selected for each tree node = $\{\log_2(N), \text{sqrt}(N)\}$ where N is the total number of features = 1588, and % of samples for each tree node (bag percent) = {20%, 30%, 40%, 50%}. There are total of 32 combinations of parameter values.

Table IV. Results based on 4-fold cross validation on the training dataset.

Algorithm	ACC	PREC	Recall	MCC	AUC_R	AUC_{Rratio}	AUC_{PR}
Random Forest	0.837	0.803	0.536	0.560	0.868	15.314	0.742
Logistic Regression	0.813	0.772	0.452	0.488	0.867	11.275	0.618
Naive Bayes	0.769	0.603	0.414	0.358	0.795	4.140	0.305

Table IV summarizes the results that correspond the highest AUC_R based on the cross validation on the training dataset for the three algorithms. We report the average accuracy, precision, recall, MCC, AUC_R, AUC_{Rratio} and AUC_{PR} over the 4 cross validation folds. We implement DMRpred using the Random Forest model that secures the best value for all measures. The parameters that were used to generate this model are: number of trees = 512, number of features per tree node = 39, and bag percent = 30%.

Table V. Comparison of DMRpred with designs that do not use: predicted RSA (No RSA), sequence conservation (No CON), putative intrinsic disorder (No ID), AA indices (No AAI), sliding windows (No SWIN), windows based on predicted IDRs (No IDWIN) and any windows (No WIN). The results are based on bootstrapping cross validation on the training dataset and are ranked by the AUC_R value. + means that DMRpred is significantly better than a given configuration (p-value < 0.01). Results are sorted in the descending order by AUC_R.

Feature set	ACC	PREC	Recall	MCC	AUC_R	AUC_{Rratio}	AUC_{PR}
DMRpred	0.837	0.803	0.536	0.560	0.868	15.314	0.742
No RSA	0.810^+	0.770^+	0.436^+	0.476^+	0.861^+	13.387^+	0.721
No SWIN	0.823	0.784	0.493	0.521	0.856^+	14.941	0.736^+
No AAI	0.816^+	0.770^+	0.461^+	0.493^+	0.854^+	13.457^+	0.722
No IDWIN	0.793^+	0.731^+	0.383^+	0.420^+	0.854^+	10.092^+	0.688^+
No CON	0.787^+	0.707^+	0.356^+	0.391^+	0.823^+	10.375^+	0.660^+
No WIN	0.782^+	0.711^+	0.340^+	0.381^+	0.818^+	7.450^+	0.615^+
No ID	0.773^+	0.686^+	0.307^+	0.346^+	0.781^+	8.679^+	0.591^+

Analysis of the DMRpred's predictive model

DMRpred combines sequence conservation, predicted RSA, putative IDRs and AA indices that quantify propensity for functions that are relevant to DMRs to define the sequence profile. It also uses two types of windows to generate features: sliding windows and windows based on predicted IDRs. We assess contributions of different parts of the profile and different window types to the predictive performance of our model. To do that we run the 4-fold cross validation on the training dataset with the best performing Random Forest model that excludes features that utilize a given part of the profile or a given type of window. We run the grid search to parametrize the Random Forest model for each subset of features. Table V summarizes results for each of these configurations. We report the average AUC_R, accuracy, precision, recall, MCC, AUC_R, AUC_{Rratio} and AUC_{PR} computed over the 4 folds. We also

evaluate statistical significance of the differences between DMRpred and each of the other configurations. We bootstrap the cross-validation results by randomly selecting 50% proteins 100 times, and we run paired *t*-test between these 100 measurements to evaluate the significance. We validated normality of these measurements using the Anderson-Darling test at the 0.05 significance.

DMRpred significantly outperforms all other configurations in accuracy, precision, recall, MCC, AUC_R, and AUC_{Rratio}. The only exception where the decrease is not statistically significant is when the sliding windows are not used. The AUC_{PR} is also significantly smaller for the configurations without the sliding window, without the window based on the predicted IDRs, and when we do not use sequence conservation and putative intrinsic disorder. This means that all elements of the sequence profile as well as the use of the disorder region-based windows significantly contribute to the DMRpred's predictive performance. In other words, their use results in a significant increase of at least one measure of predictive quality. Based on the magnitudes of the decrease in AUC_R values, the most relevant information for the prediction of DMRs includes the putative intrinsic disorder, the use of both types of windows to compute features, and sequence conservation. These factors are well-grounded in the characteristics of DMRs that are by definition disordered and include functional residues that are typically highly conserved. The windows are needed to capture differences in the intrinsic characteristics of DMRs (which form segments in the sequence) and the residues that surround these regions.

Empirical comparison with alternative approaches to predict disordered moonlighting regions

We compare the predictive performance of DMRpred with several alternative approaches that could be used to identify DMRs. Since DMRs are a subset of IDRs, they can be perhaps identified using predictors of disordered regions. Thus, we include three predictors of disordered regions: the popular Espritz[59] and IUPred[52] methods and one of the newest methods, SPOT-disorder[60], which is a successor of another popular method SPINE-D[61]. We use the three available versions of Espritz that were designed based on the three main sources of disorder annotations: NMR structures (Espritz_NMR), X-ray structures (Espritz_X-ray) and the DisProt database (Espritz_DisProt), and two available versions of IUPred: long and short. We also include four representative methods that predict specific types of functional IDRs. They include DisoRDPbind[28] that predicts disordered protein-DNA, protein-RNA and protein-protein binding regions, Anchor[62] that generates putative disordered protein-binding regions, and two methods that predict molecular recognition features (MoRFs): MoRFpred[24] and fMoRFpred[25]. MoRFs are protein-binding IDRs that undergo disordered-to-order transition upon interaction. Inclusion of these four methods is motivated by the fact that DMRs carry out multiple functions that include binding to proteins and nucleic acids. Moreover, DMRs should include evolutionarily conserved residues. Thus, we also use sequence conservation computed from alignments produced with HHblits to identify disordered moonlighting residues[47]. Finally, we include a default/typical approach to predict functional residues/regions that relies on the sequence alignment. We run PSI-BLAST[48] with default parameters for each protein in the test dataset against all proteins in the training dataset. Using the most similar training protein we copy its annotations onto the test protein for the positions with identical residues or conservative substitutions in the alignment. Residues that are not aligned as assumed to be NDMRs.

We use the corresponding author-provided webservers or implementations to run the abovementioned tools. We use the MoRFpred and fMoRFpred webservers to collect their predictions, which we utilize as

a proxy for the propensities for DMR residues. We utilize standalone software for Anchor and Espritz and the DisoRDPbind's webserver to obtain its three propensity scores for protein-protein, protein-DNA and protein-RNA binding. Because DMRs carry out multiple functions, we combine two or three DisoRDPbind's scores to represent the propensity that a given residue binds multiple partners. We combine the scores in two ways: as average of the two highest scores among the three scores DisoRDPbind produces; and as average the three scores. We run HHblits for each sequence in the test set against the default UniProt20 database to compute the conservation scores. We calculate entropy and relative entropy[49] from the 20 AA emission frequencies and use the NEFF(i) scores for each residue i that are directly output by HHblits to produce three estimates of conservation.

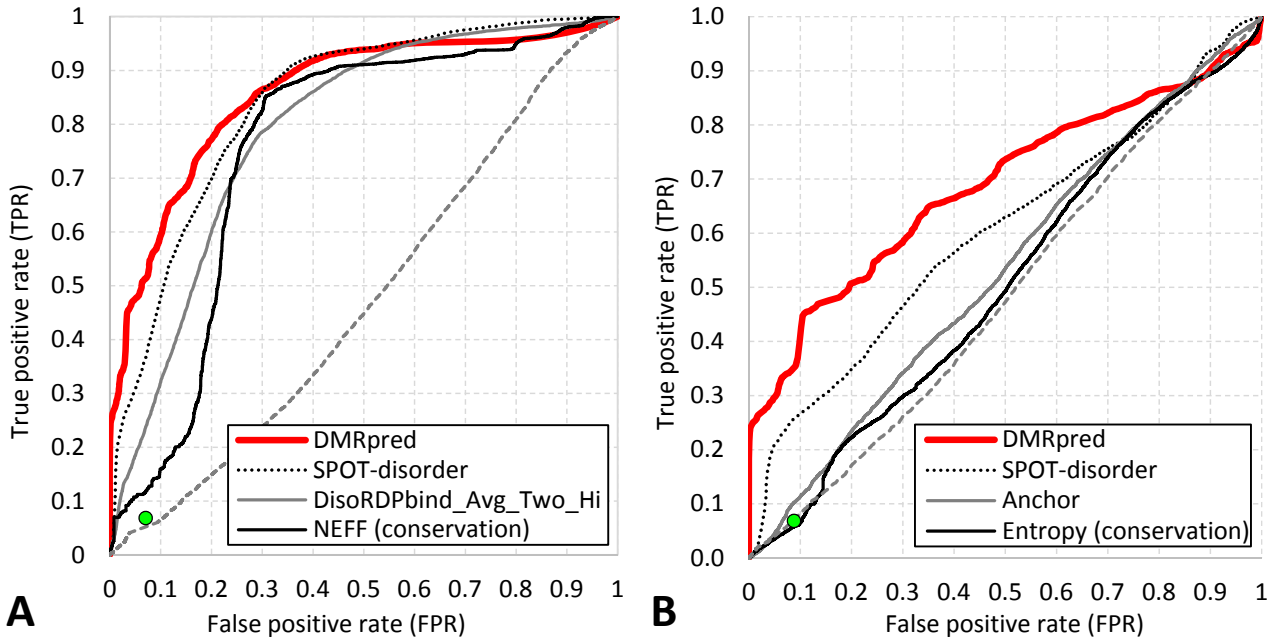

Figure 3. ROC curves for DMRpred and the other best-performing predictors (having highest AUC_R on the corresponding test dataset) from each group of methods. Panel A shows results on the complete test dataset (Table 6) while panel B is for a version of the test dataset with only monofunctional IDRs and DMR (Table 7). Alignment-based results generated with PSI-BLAST are shown with a single point since this approach generates only binary predictions.

Table VI compares the predictive performance of these approaches with DMRpred on the test dataset. We sort the methods by AUC_R within each group defined by their typical prediction target: DMRs, functional IDRs, all IDRs, MoRFs, and predictions based on sequence conservation. We rank these groups based on the highest within-group AUC_R values. We assess statistical significance of the differences between the predictive performance of DMRpred and each of the other 15 methods. We bootstrap the results by randomly selecting 50% of test proteins 100 times, and we run paired *t*-test between these 100 measurements to evaluate the significance. The measurements are normal based on the Anderson-Darling test at the 0.05 significance. We also show the ROC curves (Figure 3A) and precision-recall curves (Figure 4A) for DMRpred and one best-performing predictor having the highest AUC_R or AUC_{PR}, respectively, from each group of methods. We visualize the results generated using the alignment with the PSI-BLAST using a single point since this approach generates only binary predictions.

Table VI. Assessment of predictions on the test dataset. + means that DMRpred is significantly better than a given other method (p-value < 0.01). Accuracy, precision, recall and MCC are calculated at 5% false positive rate (see Materials and Methods for details). The best results for each measure of predictive performance are shown in bold font. NA (not available) is due to the fact that PSI-BLAST provides only the binary predictions.

Prediction target	Methods	Accuracy	Precision	Recall	MCC	AUC_R	AUC_{Rratio}	AUC_{PR}
Disordered moonlighting regions	DMRpred	**0.820**	**0.788**	**0.474**	**0.511**	**0.856**	**14.638**	**0.748**
	PSI-BLAST	0.692$^+$	0.269$^+$	0.068$^+$	-0.004$^+$	NA	NA	NA
Disordered regions	SPOT-disorder	0.773$^+$	0.701$^+$	0.308$^+$	0.353$^+$	0.840$^+$	8.294$^+$	0.632$^+$
	IUPred-long	0.756$^+$	0.654$^+$	0.247$^+$	0.288$^+$	0.792$^+$	5.697$^+$	0.552$^+$
	Espritz_NMR	0.746$^+$	0.616$^+$	0.210$^+$	0.245$^+$	0.781$^+$	4.369$^+$	0.518$^+$
	IUPred-short	0.723$^+$	0.495$^+$	0.126$^+$	0.135$^+$	0.745$^+$	2.259$^+$	0.466$^+$
	Espritz_X-ray	0.741$^+$	0.595$^+$	0.193$^+$	0.224$^+$	0.739$^+$	3.573$^+$	0.470$^+$
	Espritz_DisProt	0.694$^+$	0.154$^+$	0.024$^+$	-0.058$^+$	0.663$^+$	0.411$^+$	0.338$^+$
Functional disordered regions	DisoRDPbind_AvgTwoHigh	0.739$^+$	0.584$^+$	0.184$^+$	0.213$^+$	0.790$^+$	4.270$^+$	0.526$^+$
	DisoRDPbind_AvgThree	0.726$^+$	0.512$^+$	0.137$^+$	0.149$^+$	0.775$^+$	2.757$^+$	0.484$^+$
	Anchor	0.734$^+$	0.562$^+$	0.169$^+$	0.193$^+$	0.746$^+$	3.681$^+$	0.485$^+$
Sequence conservation	NEFF	0.719$^+$	0.461$^+$	0.107$^+$	0.108$^+$	0.754$^+$	1.552$^+$	0.466$^+$
	Entropy	0.706$^+$	0.333$^+$	0.065$^+$	0.030$^+$	0.699$^+$	2.817$^+$	0.406$^+$
	Relative entropy	0.705$^+$	0.318$^+$	0.061$^+$	0.022$^+$	0.698$^+$	1.042$^+$	0.398$^+$
MoRF regions	fMoRFpred	0.707$^+$	0.284$^+$	0.041$^+$	0.004$^+$	0.474$^+$	0.955$^+$	0.255$^+$
	MoRFpred	0.703$^+$	0.298$^+$	0.055$^+$	0.012$^+$	0.470$^+$	1.382$^+$	0.270$^+$

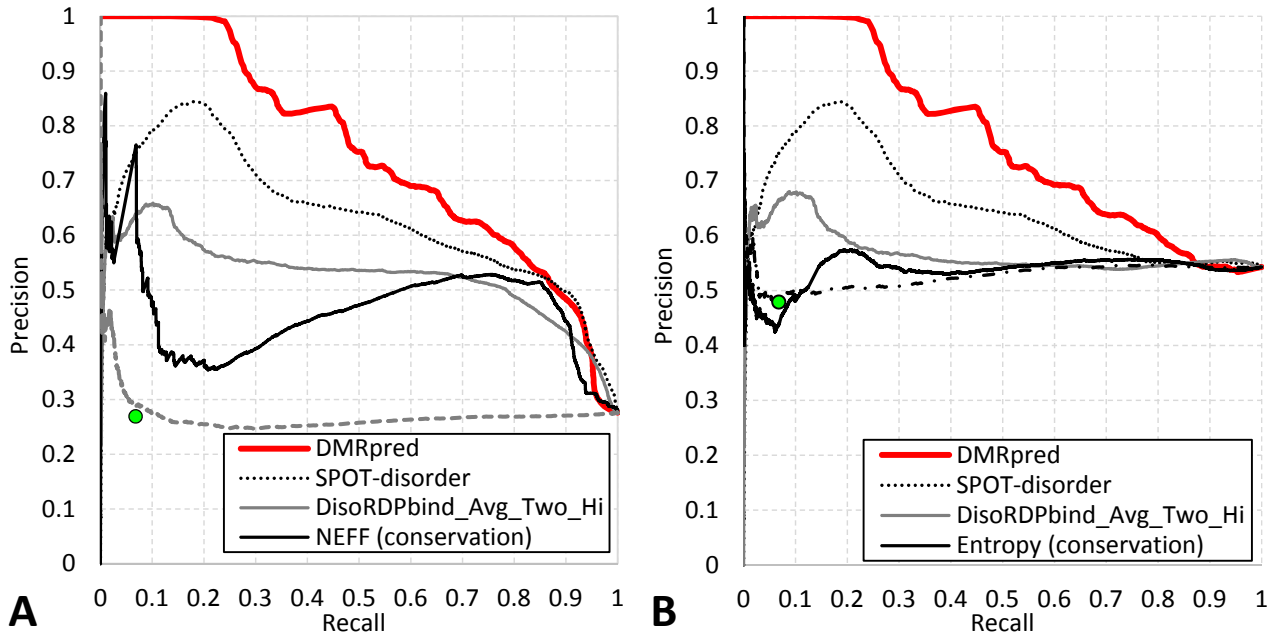

Figure 4. Precision-recall curve for DMRpred and the other best-performing predictors (with highest AUC_{PR} on the corresponding test dataset) from each group of methods. Panel A shows results on the complete test dataset (Table 6) while panel B is for a version of the test dataset with only monofunctional IDRs and DMR (Table 7). Alignment-based results by PSI-BLAST are shown with a single point since PSI-BLAST generates only binary predictions.

Table VI reveals that DMRpred offers the best predictive performance and that it significantly outperforms all other methods for all considered measures (p-value < 0.01). DMRpred's AUC_{Rratio} = 14.6 which means that its AUC_R for predictions at low FPR (< 5%) is about 14.6 times better than random. This represents about 75% improvement over the second best SPOT-disorder that secures AUC_{Rratio} = 8.3. DMRpred's AUC_R = 0.86. This high value is reflected in the ROC curve shown in Figure 3A. We note a relatively large gap between ROC for DMRpred and the other methods for FPRs ≤ 0.2. High FPRs are not practical since they result in the number of false positives that is higher than the numbers of true positive; this is because only about 27% of residues in the test dataset are DMR residues. Interestingly, DMRpred has a steep ROC curve for very low FPRs. It finds 15.1% of native DMR residues without producing any false positives. DMRpred also secures the highest AUC_{PR} value. From Figure 3A, we observe that the precision-recall curve for DMRpred is well above the corresponding curves of all other methods, especially for the values of recall < 23% where precision = 100%. Moreover, our predictor obtains precision = 90% (80%) at recall of 27% (47%). This again highlights the ability of DMRpred to provide high quality predictions for very low FPRs, which is crucial given that number of NDMR residues is larger than the number of DMRs. Using predictions calibrated to a low false positive rate (FPR) = 5%, DMRpred's accuracy = 82% and precision = 78.8%, which means it correctly predicts 82% of residues and 78.8% of the putative DMR residues. In contrast, the other approaches make correct predictions for between 70% and 77% of residues, and between 15% and 70% of the predicted DMR residues. DMRpred also secures much higher recall, MCC, and AUC_{Rratio} when compared to the other methods. Recall = 47.4% means it correctly finds 47.4% of native DMR residues when it's FPR = 5%, i.e., the fraction of NDMR residues incorrectly predicted as DMR is only 5%. As expected, predictors of disorder have on average higher recall than the predictors of functional disordered regions and MoRFs. This is because only a subset of the disordered regions are targeted by the methods that predict functional disordered regions and MoRFs. We note that DMRpred has better recall than the disorder predictors for the low values of FPR (Figure 3A), which again points to DMRpred's ability to provide substantially more accurate predictions of DMRs. However, this trend reverses for the higher values of FPR > 0.3 where the higher recall for the disorder predictors stems from the fact that all DMRs are disordered regions (Figure 3A). Finally, DMRpred's MCC = 0.51, which indicates strong correlation between the predicted DMR annotations and the native DMR annotations. This is compared to the second best MCC = 0.35 that is secured by SPOT-disorder.

The predictors of the intrinsically disordered residues (SPOT-disorder, the three versions of Espritz and the two versions of IUPred) offer significantly lower predictive performance, compared to DMRpred, because they predict all IDRs irrespective of their function(s) while majority of IDRs are not DMRs. Correspondingly, these methods over-predict DMRs. For instance, for the same predicted positive rate = 25% (number of predicted DMR residues divided by number of native DMR residues), DMRpred generates only 29 false positives while SPOT-disorder produces 268 false positives and the three Espritz versions generate between 613 and 1,254 false positives. The lower than DMRpred's recall for the disorder predictors stems from the fact that they predict a smaller number of true positives for the fixed at 5% false positive rate that is used in this tests. However, their recall is still higher than the recall of methods that predict MoRFs, functional disordered regions and methods that rely on the sequence conservation, which is because these are optimized to predict a broader category of all disordered residues.

DisoRDPbind, Anchor, MoRFpred and fMoRFpred generate putative IDRs that interact with DNA, RNA or proteins, instead of multi-functional DMRs, some of which may also implement functions that do not involve binding to nucleic acids and proteins (e.g., entropic regions and metal-binding regions). The relatively low recall of these four methods compared to DMRpred (Table III) suggest that they find

only a small subset of DMRs. The low predictive performance of the conservation scores (MCC < 0.11 and recall < 0.11) suggests that using the evolutionary conservation alone is not sufficient to separate DMRs and non-DMRs. This is because many of the NDMR residues could be conserved, including residues that interact with one ligand (residues in the monofunctional structured and disordered regions) and residues that are crucial for structural integrity of the protein fold. Moreover, the low predictive quality of PSI-BLAST (MCC = 0, precision = 26.9% and recall = 6.8%) is likely due to the fact that the test dataset shares low sequence similarity with the training proteins (< 25%) and thus alignment cannot find reliably similar chains to transfer the DMR annotations.

Table VII. Assessment of predictions on a version of the test dataset that includes only monofunctional IDRs and DMRs (structured regions are excluded). + means that DMRpred is significantly better than a given other method (*p*-value < 0.01). Accuracy, precision, recall and MCC are calculated at 5% false positive rate (see Materials and Methods for details). The best results for each measure of predictive performance are shown in bold font. NA (not available) is due to the fact that PSI-BLAST provides only the binary predictions.

Prediction target	Methods	Accuracy	Precision	Recall	MCC	AUC_R	AUC_{Rratio}	AUC_{PR}
Disordered moonlighting regions	DMRpred	**0.595**	**0.880**	**0.294**	**0.319**$^+$	**0.687**	**11.140**	**0.772**
	PSI-BLAST	0.454$^+$	0.479$^+$	0.068$^+$	-0.037$^+$	NA	NA	NA
Disordered regions	SPOT-disorder	0.548$^+$	0.833$^+$	0.208$^+$	0.231$^+$	0.601$^+$	3.348$^+$	0.650$^+$
	Espritz_NMR	0.478$^+$	0.654$^+$	0.080$^+$	0.059$^+$	0.544$^+$	0.831$^+$	0.572$^+$
	IUPred-long	0.497$^+$	0.728$^+$	0.117$^+$	0.115$^+$	0.532$^+$	1.645$^+$	0.587$^+$
	IUPred-short	0.453$^+$	0.444$^+$	0.032$^+$	-0.041$^+$	0.526$^+$	0.611$^+$	0.554$^+$
	Espritz_X-ray	0.456$^+$	0.481$^+$	0.039$^+$	-0.026$^+$	0.489$^+$	0.774$^+$	0.536$^+$
	Espritz_DisProt	0.434$^+$	0.000$^+$	0.000$^+$	-0.167$^+$	0.378$^+$	0.000$^+$	0.448$^+$
Functional disordered regions	Anchor	0.458$^+$	0.513$^+$	0.044$^+$	-0.013$^+$	0.530$^+$	0.746$^+$	0.557$^+$
	DisoRDPbind_AvgTwoHigh	0.483$^+$	0.680$^+$	0.089$^+$	0.076$^+$	0.519$^+$	1.524$^+$	0.568$^+$
	DisoRDPbind_AvgThree	0.460$^+$	0.527$^+$	0.047$^+$	-0.007$^+$	0.497$^+$	0.677$^+$	0.534$^+$
Sequence conservation	Entropy	0.449$^+$	0.395$^+$	0.028$^+$	-0.059$^+$	0.510$^+$	0.527$^+$	0.539$^+$
	Relative entropy	0.453$^+$	0.455$^+$	0.035$^+$	-0.037$^+$	0.504$^+$	0.752$^+$	0.538$^+$
	NEFF	0.472$^+$	0.626$^+$	0.070$^+$	0.042$^+$	0.472$^+$	1.773$^+$	0.526$^+$
MoRF regions	MoRFpred	0.456$^+$	0.488$^+$	0.040$^+$	-0.024$^+$	0.485$^+$	0.905$^+$	0.528$^+$
	fMoRFpred	0.447$^+$	0.358$^+$	0.023$^+$	-0.071$^+$	0.413$^+$	0.511$^+$	0.476$^+$

Empirical comparison with alternative approaches on dataset with mono-functional IDRs and DMRs

DMRpred predicts DMRs, which are a subset of IDRs that are multifunctional. While current predictors of IDRs and functions of IDRs fairly accurately differentiate between structured and disordered residues[18-21], a unique feature of DMRpred is that it is capable to distinguish between monofunctional and multifunctional IDRs (DMRs).

We test and compare this feature of DMRpred with the other 15 considered approaches. Results of an empirical comparison of predictive performance of these methods on a version of the test dataset that includes only the monofunctional IDRs and multifunctional IDRs (DMRs) (i.e., test dataset where structured residues are removed) are shown in Table VII. Figures 3B and 4B show the corresponding ROC curves and precision-recall curves, respectively. Like on the complete test

dataset, DMRpred secures the best results across the entire and comprehensive spectrum of measures of predictive performance. In particular, it obtains AUC_R close to 0.7 and AUC_{Rratio} slightly above 11. The latter means that DMRpred outperforms a random predictor by 11 folds when generating predictions characterized by low false positive rates. Table VII shows that our tool achieves 88% precision, close to 30% recall and MCC = 0.32 at the low 5% false positive rate. The differences in the predictive performance between DRMpred and each of the other 15 methods are statistically significant (p-value < 0.01). As expected we observe that the disorder predictors (SPOT-disorder, Espritz and IUPred) struggle to separate mono and multifunctional IDRs. They secure $AUC_R \leq 0.6$ and $AUC_{Rratio} < 3.4$. The other alternative predictors offer an even lower predictive quality with $AUC_R \leq 0.53$ and $AUC_{Rratio} < 1.8$. Lastly, we observe that the results on this dataset are overall worse than on the original test dataset that also includes structured residues. This trend extends over all considered methods and is expected since it is relatively easier to differentiate between DMR residues and structured residues compared to DMR vs. monofunctional IDR residues. Moreover, as we mention in Materials and Methods, some of the monofunctional IDRs could be re-labeled as DMRs as additional and different functional annotations are discovered in the future. This issue contributes to the lower predictive performance since the fraction of potentially mislabeled non-DMR residues is higher in this test dataset compared to the original dataset that includes structured residues.

Case study

We use the serine/threonine-protein phosphatase 2B (DisProt ID: DP00092) from the test dataset to illustrate predictions by DMRpred. This protein has two IDRs, one at the N-terminus (positions 1 to 13 [63]), and the other at the C-terminus (positions 373 to 521 [63]). The first IDR has no functional annotations or binding partners in DisProt and by our definition it is annotated as unknown (neither DMR nor NDMR). The second IDR includes a protein-binding region where calmodulin binds (positions 373 to 468) [63-65] and an auto-inhibitory domain (positions 371 and 511) [64,65], which define it as DMR. Figure 5 plots the outputs of DMRpred, and other methods with the highest AUC_R for a given prediction target (Table VI), except the MoRF predictors that have $AUC_R < 0.5$. DMRpred's scores (red line) at the C-terminus are high, which correctly suggests a DMR there. The scores for the structured catalytic domain (positions 14 to 373; blue horizontal line) and the N-terminus are low, suggesting that there are no DMRs there. We argue that DMRpred's prediction for the IDR at the N-terminus is possibly correct, given that this extensively studied protein does not yet have a functional annotation for this short region. To compare, SPOT-disorder (black dotted line) correctly identifies the IDR at the N-terminus and also partially predicts the IDR at the C-terminus. This prediction highlights our observation that the IDR predictors are likely to over-predict DMRs. Interestingly, the average of the two highest scores from DisoRDPbind (gray line) fails to identify the native DMR, although we observe a slight increase in the scores at the N-terminus due to higher values for its protein-binding predictions. Finally, conservation scores (black solid line) are not suitable to identify DMRs since they point to several highly conserved regions that do not line up with DMRs. Overall, we conclude that DMRpred offers reasonably accurate predictions for this protein that cannot be substituted with outputs of the other methods.

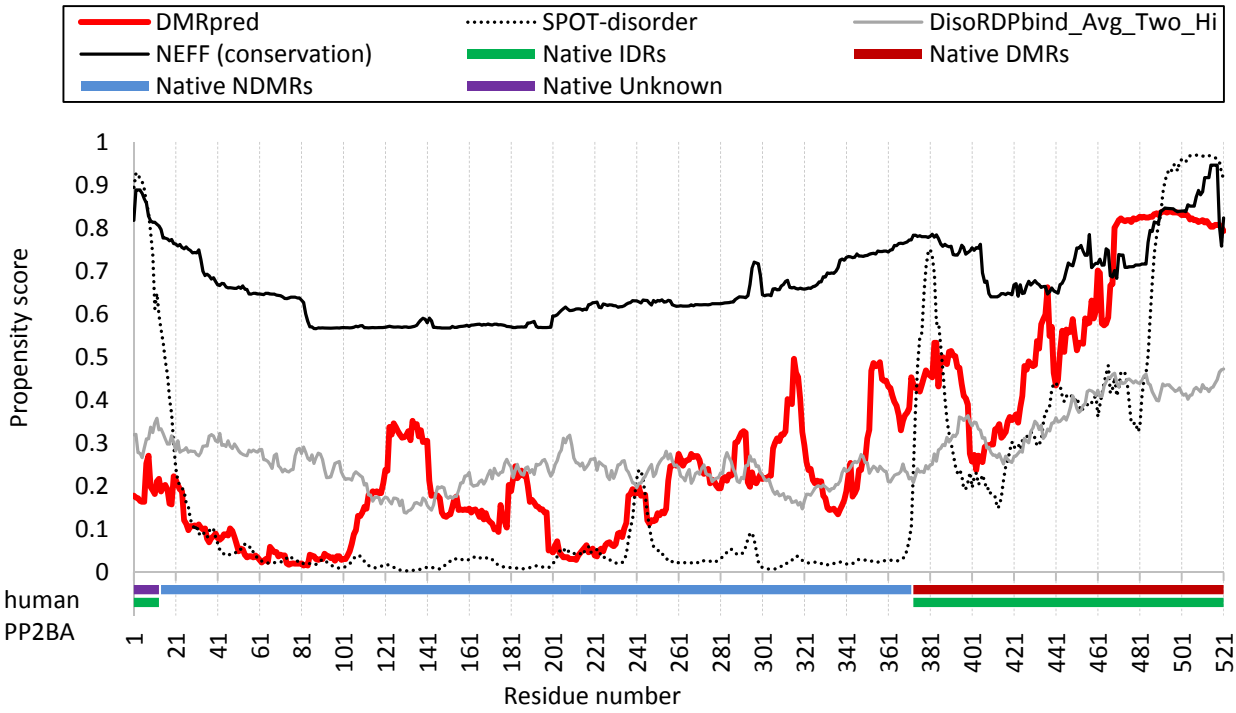

Figure 5. Predictions for human PP2BA protein (serine/threonine-protein phosphatase 2B; DisProt ID: DP00092). The horizontal lines at the bottom show native disordered regions (IDRs; light green), native DMRs (dark green), NDMRs (blue) and unknown regions (violet). We include outputs from DMRpred (thick red line), DisoRDPbind (gray), SPOT-disorder (dotted black) and conservation scores from HHblits (solid black).

Prediction and analysis of DMRs in the human proteome

We characterize putative DMRs and IDRs in the complete reviewed human proteome that we collected from UniProt[66,67]. We retrieved the annotations of intrinsic disorder for these proteins from the MobiDB database[68,69]. We use 19,917 human proteins after removing about 200 proteins that could not be mapped to MobiDB. We make predictions with DMRpred and annotate DMR residues based on the binary predictions that are calibrated to produce 5% false positive rate on the test dataset, i.e., residues with propensities ≥ 0.761 are assumed as DMR residues. We annotate putative DMRs as segments of at least four consecutive DMR residues. This is in line with the definition of all IDRs that are expected to include at least four consecutive amino acids[18,19]. We found about 32 thousand putative DMRs in the human proteome, which corresponds to around 30% of the 107 thousand putative IDRs. This is similar to the 37% rate of DMRs among the IDRs included in DisProt, which was reported in[15]. We focus our analysis on long (≥ 30 consecutive residues) putative DMRs and IDRs since they are recognized as a distinct class of biologically functional domains[4,70,71].

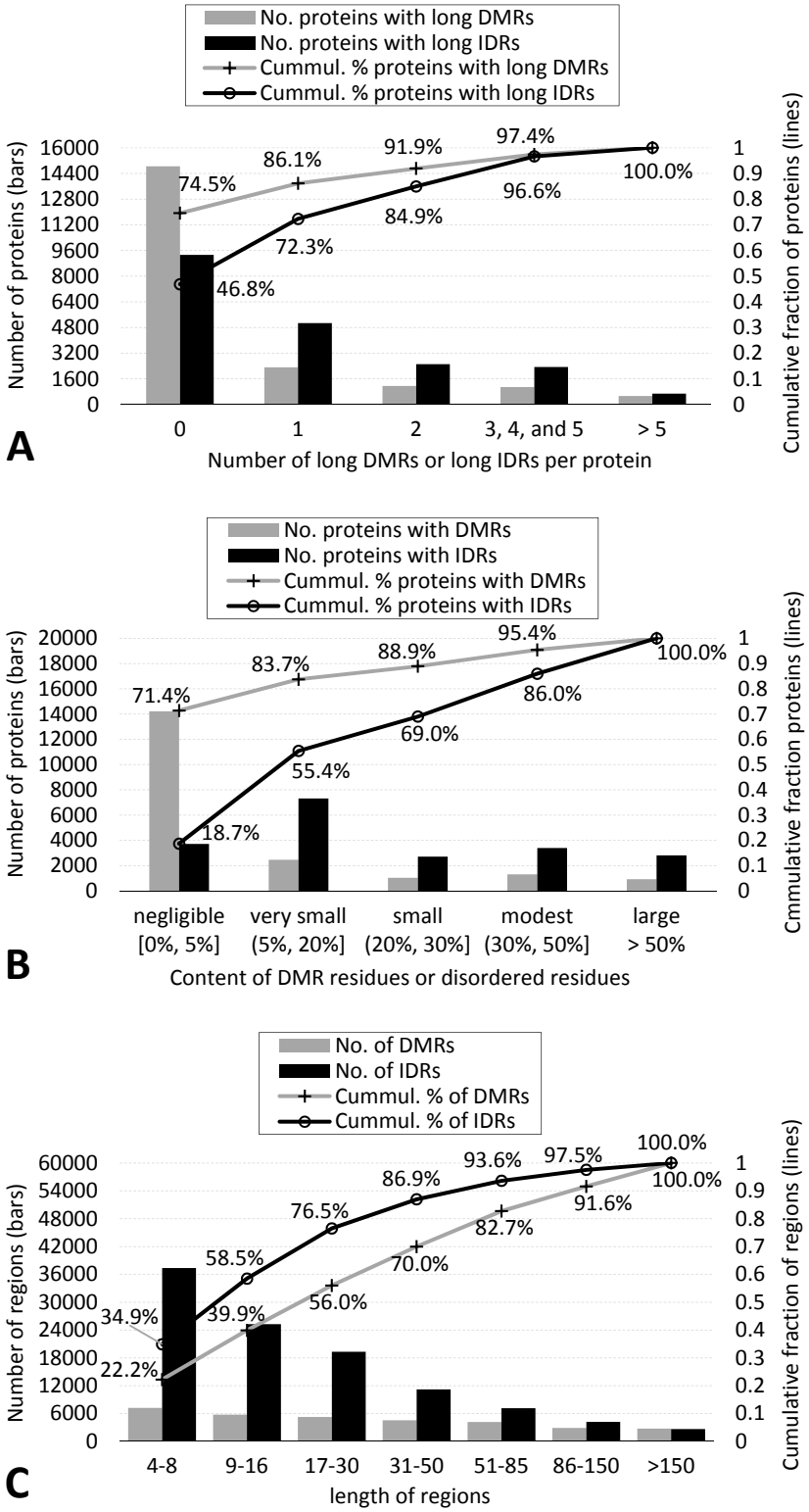

Figure 6. Analysis of putative DMRs in the complete reviewed human proteome. Panel A shows relation between the number of long (>30 consecutive residues) putative DMRs and intrinsically disordered regions (IDRs). Bars show number of proteins that have the number of long regions given on the *x*-axis. Panel B summarizes content of residues in putative DMRs and putative IDRs. Bars

show the number of proteins with content ranges given on the *x*-axis. Panel C analyzes the length of putative DMRs and IDRs. Bars show the number of regions with length ranges given on the *x*-axis. Lines show the corresponding cumulative fractions of proteins in each of the three panels. The putative DMRs were generated with DMRpred. Putative IDRs were collected from MobiDB.

Figure 6A suggests that about 53% of human proteins have at least one long IDR. This agrees with recent estimates that ranged between about 45%[4] and 50%[5]. Interestingly, we show that about 25% of human proteins may have at least one long DMR, and 8% may have three or more long DMRs. Figure 6B shows that about 29% human proteins have DMRs. The content of DMR residues among the remaining 71% proteins is below 5%. These are considered spurious predictions since that the false positive rate of DMRpred is estimated to be 5%. To compare, about 81% of human proteins have disordered residues, i.e., their disorder content > 5%. Such substantial difference in the rate of disorder vs. DMR content is reasonable given that only a small fraction of IDRs are DMRs. Further analysis shows that about 11% of human proteins are predicted to have at least modest content of DMRs (>30% of their residues are in DMRs) and 4.6% to have large content (> 50%) (Figure 6B). These results also reveal that three times as many proteins (approximately 31%) have > 30% disorder content. The latter result is in good agreement with a recent analysis in Ref. [4] where about 31.5% of proteins were predicted to have at least 30% of disordered residues. The histograms of length of DMRs and IDRs are given in the Figure 6C. Both histograms follow the same trend, where there are gradually fewer regions that are longer. The main differences are the overall number of regions that, as expected, is much lower in the case of DMR. The rate of decline that is also lower for DMRs; see black bars (for IDRs) and gray bars (for DMRs) in Figure 6A. Interestingly, our analysis reveals that most of the very long disordered regions are possibly DMRs, given that the number of regions longer than 150 residues is similar when comparing IDRs and DMRs.

DMRpred's webserver

DMRpred's webserver is available at http://biomine.cs.vcu.edu/servers/DMRpred. Users only need to provide FASTA-formatted protein sequence(s) to obtain predictions that are computed on the server side. The webserver outputs a propensity score for each residue in the input sequence(s) for being a DMR residue. It also produces binary predictions that are generated from the propensities using the cutoff = 0.761; residues with propensity ≥ 0.761 are predicted as DMR residues. This cutoff was calibrated to provide 5% FPR on the test dataset. The DMRpred's webserver allows batch submissions of up to 50 sequences at one time. The sequences should be at least 21 residues long since ASAquick that is embedded into DMRpred requires this. Users are encouraged to provide email address which is used to provide notification when the prediction is finished and a private URL where the results can be downloaded from. Whether or not the email is provided, the results are also made available in the browser window, given that the user will not close it when the results are being processed. DMRpred is relatively fast. The webserver produces prediction for a protein with length of about 500 residues in less than one minute.

Conclusions

We conceptualized, designed, tested and deployed DMRpred, the first-of-its-kind computational method for the prediction of DMRs directly from protein sequences. DMRpred uses the input sequence to derive a comprehensive profile that includes sequence conservation, putative relative

solvent accessibility and intrinsic disorder, and a novel set of residue-level propensities for functions that are relevant to DMRs. The information in this profile is aggregated using sliding windows and an innovative type of windows defined based on putative IDRs. Features extracted from this profile are input to the Random Forest model to make the predictions.

We empirically demonstrate that the various parts of the profile and the two types of windows are useful for the prediction. Results on a blind test dataset reveal that DMRpred provides accurate predictions of DMRs. The predictive quality of DMRpred is statistically significantly higher than the predictive performance of a comprehensive set of alternative approaches to make these predictions. Predictions on the complete human proteome reveal that as many as 25% of human proteins may have at least one long DMR. A webserver that implements DMRpred is freely available at http://biomine.cs.vcu.edu/servers/DMRpred/.

Conflict of Interest

Authors declare no conflict of interests.

Funding

This research was supported in part by the National Science Foundation grant 1617369 to L.K.

References

1. Habchi J, Tompa P, Longhi S, Uversky VN. Introducing protein intrinsic disorder. Chem Rev 2014;114(13):6561-6588.
2. Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, Hu G, Uversky VN, Kurgan L. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cellular and molecular life sciences : CMLS 2015;72(1):137-151.
3. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337(3):635-645.
4. Pentony MM, Jones DT. Modularity of intrinsic disorder in the human proteome. Proteins 2010;78(1):212-221.
5. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol 2008;18(6):756-764.
6. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. Intrinsic Disorder and Protein Function†. Biochemistry 2002;41(21):6573-6582.
7. Babu MM. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. Biochem Soc Trans 2016;44(5):1185-1200.
8. Hu G, Wu Z, Uversky VN, Kurgan L. Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions. Int J Mol Sci 2017;18(12).
9. Wang C, Uversky VN, Kurgan L. Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. Proteomics 2016;16(10):1486-1498.
10. Meng F, Na I, Kurgan L, Uversky VN. Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments. Int J Mol Sci 2016;17(1).

11. Fuxreiter M, Toth-Petroczy A, Kraut DA, Matouschek A, Lim RY, Xue B, Kurgan L, Uversky VN. Disordered proteinaceous machines. Chem Rev 2014;114(13):6806-6843.
12. Xue B, Blocquel D, Habchi J, Uversky AV, Kurgan L, Uversky VN, Longhi S. Structural disorder in viral proteins. Chem Rev 2014;114(13):6880-6911.
13. Peng Z, Xue B, Kurgan L, Uversky VN. Resilience of death: intrinsic disorder in proteins involved in the programmed cell death. Cell Death Differ 2013;20(9):1257-1267.
14. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM. Classification of Intrinsically Disordered Regions and Proteins. Chemical Reviews 2014;114(13):6589-6631.
15. Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cell Mol Life Sci 2017;74(17):3069-3090.
16. Lieutaud P, Ferron F, Uversky AV, Kurgan L, Uversky VN, Longhi S. How Disordered is My Protein and What is Its Disorder For? A Guide Through the "Dark Side" of the Protein Universe. Intrinsically Disordered Proteins 2016;4(1):e1259708.
17. Meng F, Uversky V, Kurgan L. Computational Prediction of Intrinsic Disorder in Proteins. Curr Protoc Protein Sci 2017;88:2 16 11-12 16 14.
18. Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K. Assessment of protein disorder region predictions in CASP10. Proteins 2014;82 Suppl 2:127-137.
19. Peng ZL, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci 2012;13(1):6-18.
20. Necci M, Piovesan D, Dosztanyi Z, Tompa P, Tosatto SCE. A comprehensive assessment of long intrinsic protein disorder from the DisProt database. Bioinformatics 2017.
21. Walsh I, Giollo M, Di Domenico T, Ferrari C, Zimmermann O, Tosatto SCE. Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics 2015;31(2):201-208.
22. Fan X, Kurgan L. Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. J Biomol Struct Dyn 2014;32(3):448-464.
23. Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics 2009;25(20):2745-2746.
24. Disfani FM, Hsu W-L, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 2012;28(12):i75-i83.
25. Yan J, Dunker AK, Uversky VN, Kurgan L. Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst 2016;12(3):697-710.
26. Malhis N, Gsponer J. Computational identification of MoRFs in protein sequences. Bioinformatics 2015;31(11):1738-1744.
27. Meng F, Kurgan L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. Bioinformatics 2016;32(12):i341-i350.
28. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Research 2015.
29. Peng Z, Wang C, Uversky VN, Kurgan L. Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. Methods Mol Biol 2017;1484:187-203.
30. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics 2008;9 Suppl 1:S1.

31. Sun X, Rikkerink EH, Jones WT, Uversky VN. Multifarious roles of intrinsic disorder in proteins illustrate its broad impact on plant biology. Plant Cell 2013;25(1):38-55.
32. Uversky VN. Intrinsic Disorder-based Protein Interactions and their Modulators. Curr Pharm Design 2013;19(23):4191-4213.
33. Piovesan D, Tabaro F, Micetic I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidovic R, Dosztanyi Z, Elofsson A, Gasparini A, Hatos A, Kajava AV, Kalmar L, Leonardi E, Lazar T, Macedo-Ribeiro S, Macossay-Castillo M, Meszaros A, Minervini G, Murvai N, Pujols J, Roche DB, Salladini E, Schad E, Schramm A, Szabo B, Tantos A, Tonello F, Tsirigos KD, Veljkovic N, Ventura S, Vranken W, Warholm P, Uversky VN, Dunker AK, Longhi S, Tompa P, Tosatto SC. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res 2016;D1:D219-D227.
34. Jeffery CJ. Moonlighting proteins. Trends in Biochemical Sciences 1999;24(1):8-11.
35. Khan IK, Kihara D. Genome-scale prediction of moonlighting proteins using diverse protein association information. Bioinformatics 2016.
36. Khan Ishita K, Kihara D. Computational characterization of moonlighting proteins. Biochemical Society Transactions 2014;42(6):1780-1785.
37. Gómez A, Domedel N, Cedano J, Piñol J, Querol E. Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins? Bioinformatics 2003;19(7):895-896.
38. Khan IK, Chitale M, Rayon C, Kihara D. Evaluation of function predictions by PFP, ESG, and PSI-BLAST for moonlighting proteins. BMC Proceedings 2012;6(7):1-5.
39. Olah J, Bertrand P, Ovadi J. Role of the microtubule-associated TPPP/p25 in Parkinson's and related diseases and its therapeutic potential. Expert Rev Proteomics 2017;14(4):301-309.
40. Redwan EM, Al-Hejin AM, Almehdar HA, Elsaway AM, Uversky VN. Prediction of Disordered Regions and Their Roles in the Anti-Pathogenic and Immunomodulatory Functions of Butyrophilins. Molecules 2018;23(2).
41. Migliaccio AR, Uversky VN. Dissecting physical structure of calreticulin, an intrinsically disordered Ca(2+)-buffering chaperone from endoplasmic reticulum. J Biomol Struct Dyn 2017:1-20.
42. Thieulin-Pardo G, Schramm A, Lignon S, Lebrun R, Kojadinovic M, Gontero B. The intriguing CP12-like tail of adenylate kinase 3 from Chlamydomonas reinhardtii. FEBS J 2016;283(18):3389-3407.
43. Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidović R, Dosztányi Z, Elofsson A, Gasparini A, Hatos A, Kajava AV, Kalmar L, Leonardi E, Lazar T, Macedo-Ribeiro S, Macossay-Castillo M, Meszaros A, Minervini G, Murvai N, Pujols J, Roche DB, Salladini E, Schad E, Schramm A, Szabo B, Tantos A, Tonello F, Tsirigos KD, Veljković N, Ventura S, Vranken W, Warholm P, Uversky VN, Dunker AK, Longhi S, Tompa P, Tosatto SCE. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Research 2017;45(D1):D219-D227.
44. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research 2000;28(1):235-242.
45. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403-410.
47. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Meth 2012;9(2):173-175.

48. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 1997;25(17):3389-3402.
49. Wang K, Samudrala R. Incorporating background frequency improves entropy-based residue conservation measures. BMC Bioinformatics 2006;7:385.
50. Faraggi E, Zhou Y, Kloczkowski A. Accurate single-sequence prediction of solvent accessible surface area using local and global features. Proteins: Structure, Function, and Bioinformatics 2014;82(11):3170-3176.
51. Dosztányi Z, Csizmók V, Tompa P, Simon I. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. Journal of Molecular Biology 2005;347(4):827-839.
52. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21(16):3433-3434.
53. Vacic V, Uversky V, Dunker AK, Lonardi S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. BMC Bioinformatics 2007;8(1):211.
54. Le Cessie S, Van Houwelingen J. Ridge estimators in logistic regression. Applied statistics 1992:191-201.
55. John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. 1995. Morgan Kaufmann Publishers Inc. p 338-345.
56. Breiman L. Random forests. Mach Learn 2001;45(1):5-32.
57. Frank E, Hall MA, Witten IH. The WEKA Workbench. Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition: Morgan Kaufmann; 2016.
58. Oshiro TM, Perez PS, Baranauskas JA. How Many Trees in a Random Forest? In: Perner P, editor. Machine Learning and Data Mining in Pattern Recognition MLDM 2012. Volume 7376. Berlin, Heidelberg: Springer; 2012. p 154-168.
59. Walsh I, Martin AJM, Di Domenico T, Tosatto SCE. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 2012;28(4):503-509.
60. Hanson J, Yang Y, Paliwal K, Zhou Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. Bioinformatics 2017;33(5):685-692.
61. Zhang T, Faraggi E, Xue B, Dunker AK, Uversky VN, Zhou Y. SPINE-D: Accurate Prediction of Short and Long Disordered Regions by a Single Neural-Network Based Method. Journal of biomolecular structure & dynamics 2012;29(4):799-813.
62. Mészáros B, Simon I, Dosztányi Z. Prediction of Protein Binding Regions in Disordered Proteins. PLoS Comput Biol 2009;5(5):e1000376.
63. Kissinger CR, Parge HE, Knighton DR, Lewis CT, Pelletier LA, Tempczyk A, Kalish VJ, Tucker KD, Showalter RE, Moomaw EW, et al. Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. Nature 1995;378(6557):641-644.
64. Ye Q, Feng Y, Yin Y, Faucher F, Currie MA, Rahman MN, Jin J, Li S, Wei Q, Jia Z. Structural basis of calcineurin activation by calmodulin. Cell Signal 2013;25(12):2661-2667.
65. Wang H, Du Y, Xiang B, Lin W, Li X, Wei Q. A renewed model of CNA regulation involving its C-terminal regulatory domain and CaM. Biochemistry 2008;47(15):4461-4468.
66. The UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017;45(D1):D158-D169.
67. UniProt C. UniProt: a hub for protein information. Nucleic Acids Res 2015;43(Database issue):D204-212.

68. Di Domenico T, Walsh I, Martin AJM, Tosatto SCE. MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics 2012;28(15):2080-2081.
69. Potenza E, Di Domenico T, Walsh I, Tosatto SC. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res 2015;43(Database issue):D315-320.
70. Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, Uversky VN. Close encounters of the third kind: disordered domains and the interactions of proteins. Bioessays 2009;31(3):328-335.
71. Peng Z, Mizianty MJ, Kurgan L. Genome-scale prediction of proteins with long intrinsically disordered regions. Proteins 2014;82(1):145-158.