Pacific Symposium on Biocomputing 25:171-182(2020)

# Disordered Function Conjunction: On the *in-silico* function annotation of intrinsically disordered regions

Sina Ghadermarzi[#], Akila Katuwawala[#], Christopher J. Oldfield[#], Amita Barik, and Lukasz Kurgan
*Department of Computer Science, Virginia Commonwealth University,*
*401 West Main Street, Richmond, VA 23284, U.S.A.*
*Email: lkurgan@vcu.edu*
[#] These authors contributed equally

Intrinsically disorder regions (IDRs) lack a stable structure, yet perform biological functions. The functions of IDRs include mediating interactions with other molecules, including proteins, DNA, or RNA and entropic functions, including domain linkers. Computational predictors provide residue-level indications of function for disordered proteins, which contrasts with the need to functionally annotate the thousands of experimentally and computationally discovered IDRs. In this work, we investigate the feasibility of using residue-level prediction methods for region-level function predictions. For an initial examination of the multiple function region-level prediction problem, we constructed a dataset of (likely) single function IDRs in proteins that are dissimilar to the training datasets of the residue-level function predictors. We find that available residue-level prediction methods are only modestly useful in predicting multiple region-level functions. Classification is enhanced by simultaneous use of multiple residue-level function predictions and is further improved by inclusion of amino acids content extracted from the protein sequence. We conclude that multifunction prediction for IDRs is feasible and benefits from the results produced by current residue-level function predictors, however, it has to accommodate inaccuracy in functional annotations.

*Keywords:* Intrinsically disordered proteins; Protein function; Protein-protein interactions; Protein-DNA interactions; Protein-RNA interactions; Linker regions.

## 1. Introduction

Intrinsically disordered regions (IDRs) in proteins lack a stable three-dimension structure under physiological conditions, instead existing as ensembles of conformations [1-3]. Despite this lack of structure, IDRs perform many and varied biological functions using mechanisms distinct from the mechanisms of structured proteins [4]. Further, estimates suggest that IDRs are extremely common in nature, with 25 to 40% of eukaryotic proteins containing IDRs [5-7]. While IDRs are common, only a small number are annotated with biological functions. For example, the DisProt database [8], which curates known IDRs and their functions, contains only 803 proteins. This lack of data is compounded by the difficulty of using structure-based homology techniques to transfer annotations to uncharacterized proteins; extremely high identity is required for IDRs homology (>80% identity) [9] relative to that used for structured proteins (e.g. >30% identity). This calls for computational methods for both locating IDRs and, just as importantly, determining their functions.

Prediction of disorder-associated functions is still in its infancy [10]. The first disorder function to gain popularity as a prediction target was protein-binding functions and several predictors for this have been developed to date [11-18]. Other functions have been slow to follow; we are only aware of one method for the prediction of DNA-binding and RNA-binding and another for predicting linkers [19]. The need for disorder-specific function prediction has been demonstrated; they are complementary to the predictions for the same type of function for structured regions [13].

Disorder function predictors are universally designed to predict function at the residue level [11-17, 19]. While this provides a good resolution to function prediction, this approach creates some difficulties. Primary among these is the mismatch to conventional annotation; IDR databases typically annotate functions of IDRs rather than attributing functions to individual residues [8, 9, 20]. Predicting function at the level of regions, rather than residues, matches available experimental data and annotation practices, and provides valuable functional context for the millions of putative IDRs that are available in the MobiDB [9, 21] and $D^2P^2$ [22] resources.

In this work, we investigated the feasibility of using residue level predictors of disorder function to classify IDRs according to their functions. To our knowledge, this work is the first of its kind. For this investigation, a dataset of IDRs for several function types (protein-, DNA-, and RNA-binding and linker regions) was built. This selection of functions is dictated by the availability of residue-level predictors of these functions. Our goal is to devise an approach to adapt residue-level function predictions to region-level function predictions and to demonstrate the feasibility of prediction-based annotation of several functions using multiple residue-level function predictions.

## 2. Materials and Methods

### 2.1. *Data collection*

The regions used in this work were collected from proteins deposited in the DisProt database [8] and the Protein Data Bank (PDB) [23], respectively. To create this dataset, an initial set of 799 disorder-containing proteins from DisProt and 1421 protein chains from PDB was considered. We use the entire DisProt, except for proteins with the disorder annotations flagged as ambiguous. The sequences from PDB must have high resolution structure (< 2Å), cover a complete protein chain, and exclude structures in complex with other molecules and peptides (chains < 30 residues long). We check the sequence coverage by mapping PDB chains into the corresponding UniProt protein using SIFTS [24]. The resulting 1421 structured proteins are unlikely to contain IDRs.

We ensure that that these proteins share low (<25%) sequence similarity with the training datasets of the residue-level function predictors. The 799 DisProt proteins and 1421 PDB proteins were clustered with the training sets of the considered in this article disorder function predictors: ANCHOR [17], DisoRDPbind [13, 25], fMoRFpred [12], DFLpred [19] and DMRpred [26] using CD-HIT [27] at 30% similarity threshold. The proteins that are in the clusters that include any of the proteins from these training sets were removed, leaving 229 proteins from DisProt and 417 PDB proteins that are dissimilar (<30% similarity) to the training proteins. We annotate functions of IDRs in these 229 proteins using DisProt and exclude IDRs that have multiple functions to simplify the analysis and the assessment of predictive performance. Analysis of the multi-functional IDRs will be a subject of a future study. We extract 70 functionally annotated IDRs including 51 protein-binding, 7 DNA-binding, 5 RNA-binding and 7 linker regions. We add a matching number of non-overlapping structured regions with same length as the above IDRs, which we extract at random from the sequences of the 417 PDB proteins.

### 2.2. *Computational workflow*

The computational pipeline for the functional prediction of IDRs, which also differentiates them from structured regions, has two main steps. First, the regions are represented by a fixed set of numerical features that are used to predict the functions. Second, these features are used to categorize each region to the corresponding type (function or a structured region).

#### 2.2.1. *Feature-based representation of protein regions*

We extract two categories of the region-level features from: (1) the residue-level function predictions, and (2) the sequence of the region. Features in the first category are extracted from the residue-level predictions of the disordered linkers produced by DFLpred [19], the disordered protein-binding residues generated with ANCHOR [17], DisoRDPbind [13, 25] and fMoRFpred [12], and the DNA- and RNA-binding residues by DisoRDPbind [13, 25]. We extract residue-level predictions for the disordered and structured regions from the predictions generated using the corresponding protein sequences. Next, we aggregate these residue-level predictions at the region level using an average over the residues in the region and the maximal value of the average in the 5-residues long sliding window. The latter features quantify a local (using a short sequence segment) putative propensity for each predicted function. The second category of features is extracted from the sequence of the considered regions. These features include 20-dimensional amino acid composition and the region length. We investigate the predictive value of each feature category and compare it to the use of both categories.

#### 2.2.2. *Prediction of protein region functions*

We use the region-level features to predict the region types (disorder function or structured) using two alternative approaches. Both approaches use only the information extracted from the regions without the knowledge of their types (labels), which ensures that they do not overfit the dataset. First, we utilize the most direct approach in which we use the features based on the average of the residue-level predictions of a given function to predict regions for the same function. In other words, we predict each function individually by using the features extracted from the corresponding residue-level function prediction, and we predict a given region as structured if none of the functions is predicted. Second, we use the features extracted from the residue-level predictions of all functions together and we combine them with the sequence-based features to concurrently predict multiple region types. We compare these two approaches, and investigate whether inclusion of the sequence-based features in the second approach results in improvements.

The first approach is straightforward. We represent each region using the average-based features for the given function and model these values using z-scores. Z-scores are obtained by transforming each attribute to zero mean and unit variance over the entire dataset. We predict a given function if the corresponding z-scores is the maximum among all function prediction methods. A single z-score threshold for function classification was set based on maximized MCC, above which disorder function classifications were retained and below which regions were classified as structured. The select threshold was a z-score of 1.18. A non-parametric transformation was explored but performed poorly relative to z-scores (data not shown).

The second approach relies on an unsupervised learning where we first cluster the regions using their feature-based representation and then we label the resulting clusters with the region types. Clustering finds natural groupings of regions in which the regions within a given cluster are similar with each other when compared to the regions outside the cluster. To prepare the features for clustering, we scale them to mean zero and the standard deviation of one. We compare results obtained using four feature sets: (1) all features that include features extracted from the residue-level predictions and from the sequences; (2) features extracted from the sequence; (3) features extracted from the residue level predictions; (4) a subset of (1) that includes only average region scores. The latter set is the same as the feature set utilized by the z-score-based approach. Each of these feature sets is used either directly or after it is transformed using the Principal Component

Analysis (PCA), resulting in eight setups. PCA is an unsupervised method that extracts linearly uncorrelated features using linear combination of the (possibly correlated) input features [28]. We use PCA as a pre-processing step to reduce correlations between the inputs for the subsequent clustering since this is likely to improve the results of the clustering [29]. We cluster each of the eight feature sets using the most popular clustering method that relies on the pre-define knowledge of the number of clusters, the *K*-means clustering [30]. We use the *K*-means++ version, which utilizes an improved way to initialize clusters, and the Euclidian distance that is suitable for our features. We set $k = 5$, which corresponds to the number region types that include protein-binding, RNA-binding, DNA-binding, linker and structured regions. We label the resulting clusters with the region types to maximize the micro MCC measure (which we define in Section 2.2.3). We compare these methods with an approach that uses all features (with and without PCA) for regions with randomly shuffled region types (labels). This is to verify that our methods in fact do not overfit the dataset, in which case the model based on the shuffled region types would obtain similarly good results.

### 2.2.3. *Assessment of the function prediction and clustering*

We quantify the quality of the clustering results using three popular measures: purity [31], Normalized Mutual Information (NMI) [32] and Adjusted Rand Index (ARI) [33]. These external measures quantify the relationship between cluster membership and region types. Purity is equivalent to the accuracy of prediction when we label each cluster with the most common region type within the cluster. Mutual Information quantifies the amount of information that we obtain about the region types by knowing which cluster they belong to. NMI is a variation of the Mutual Information that is adjusted to account for chance [32]. ARI is an adjusted version of the Rand Index that measures similarity of two data groupings (clustering vs. region types).

We quantify quality of the region type predictions for both considered approaches: z-score-based and clustering-based. We measure predictive quality with a comprehensive set of measures for each region type: sensitivity, specificity, F1 and Mathew's Correlation Coefficient (MCC):

$$\text{Sensitivity} = \frac{TP}{TP+FN}, \quad \text{Specificity} = \frac{TN}{TN+FP}, \quad \text{F1-Score} = \frac{2TP}{2TP+FP+FN}, \quad \text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TP is the number of regions of a given type that are correctly predicted as this type, TN is the number of regions of other types that are correctly predicted as the other types, FP is the number of the regions of the other types incorrectly predicted as the given type, and FN is the number of regions of a given type predicted as any of the other types. F1 is a harmonic average of precision and recall. MCC values around zero show lack of correlation between the predicted and actual region types while higher MCC values correspond to a stronger correlation.

Besides the per-region type measures, we calculate three measures of the overall quality that considers all region types. The accuracy is the number of correctly predicted region types divided by the number of all regions. We compute the overall F1 and the overall MCC values using micro and macro averaging. The macro averaging is an average of the five corresponding per-region-type values and by definition, it assumes that prediction for each region type has equal weight. The micro averaging quantifies predictive quality proportionally to the number of samples in each region type, i.e., regions types with more sample, like the protein-binding and structured, have proportionally larger influence on the micro averaged score. The micro averaging is computed using the same formula as the per-region type case where TP, FP, TN and FN are the sum of the five corresponding per-region type values.

We use the scikit-learn library to compute PCA, *K*-means and NMI and ARI cluster quality measures. We use in-house scripts to compute the purity measure and the prediction scores.

## 3. Results and Discussion

### 3.1. *Prediction of individual functions of IDRs*

The dataset includes IDRs annotated with protein-binding, DNA-binding, RNA-binding or linker function. Selection of these function types was based on the availability of prediction methods for each. A control set of structured regions provides a set of hard negatives that should not be predicted to have any disorder-based function. For the exploratory analysis here, this is preferable to unannotated disordered region, which are likely to have an undiscovered function. This is in contrast to IDRs with annotated functions, which have been the subjects of experimental study and less likely to have unannotated functions. For these proteins, we made disorder-based function predictions for protein-binding, DNA-binding, RNA-binding and linkers. We did not consider function prediction methods developed for structured proteins; these methods were previously determined to perform poorly for disordered regions [13]. Further, inclusion of general methods would complicate our analysis since our structured, hard-negative set was not designed to exclude any particular structure-based function.

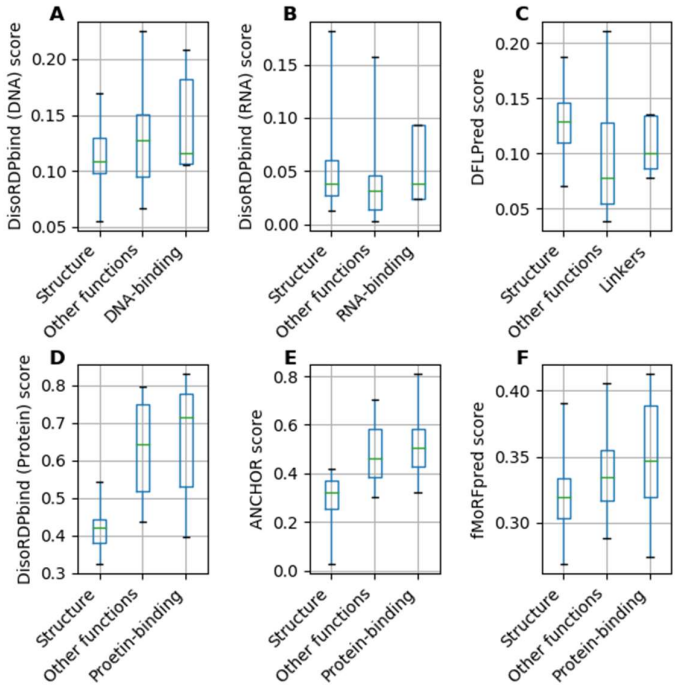


Fig. 1. Distributions of the per-region scores generated by the disorder function predictors. Each set of box plots show the distribution of the per-residue scores averaged over each region from three subsets of regions: regions with relevant disorder function for that predictor (right), disordered regions with other functions (center), and structured regions (left). Whiskers indicate the 95% and 5% extent of the data. The six prediction sets shown here are: (A) DNA-binding by DisoRDPbind, (B) RNA-binding by DisoRDPbind, (C) linkers by DFLPred, (D) protein-binding by DisoRDPbind, (E) protein-binding by ANCHOR, and (F) protein-binding by fMoRFpred.

For a simple classification of regions, we use the average of the residue-level predictions over each region as predictions of region function propensity. The distributions of these averages for each region type generated by the corresponding predictors are contrasted against the values for the other disordered regions and the structured regions in Fig. 1. Ideally, the average prediction scores would be greater for the predicted function than for other functions, and still greater than structured regions. This expectation is based on structured regions being hard negatives, while disordered regions with other types are soft negatives since they may be multifunctional. The distributions of the putative

protein-binding scores (Fig. 1D, E, and F) reflect this expectation, though protein-binding scores are not well separated from these scores for regions with other functions. The putative DNA-binding and RNA-binding scores (Fig. 1A and 1B, respectively) show mixed results. The DNA-binding regions have greater scores than structured regions, but overlap with a lower median when compared to other functions. The putative RNA-binding regions have slightly higher medians than other functions and structured regions, but these distributions largely overlap. The linker scores (Fig. 1C) have the poorest performance on this dataset, with structured regions having higher scores than other functions. However, the median predictions for the linkers are higher than other functions.

The performance when predicting each function individually using the corresponding average varies between region types. The protein-binding predictors have the highest accuracy; ANCHOR, DisoRDPbind, and fMoRFpred have MCC values of 0.66, 0.63, and 0.37, respectively. The poorer performance of fMoRFpred is not unexpected, since MoRFs are a subset of short (5 to 25 residues in length) protein-binding disordered regions [12]. Linker predictions by DFLPred show moderate performance with an MCC of 0.37. The poorest performing predictors, on this dataset, are DNA- and RNA-binding predictions by DisoRDPbind, with MCC values of 0.25 and 0.16, respectively. Further, two features of this dataset should be considered when evaluating performance. First, the dataset has been designed to exclude all proteins similar to the training sets of all predictors, so predictors are working in the difficult, extrapolatory regime. Second, only a small number of linker, DNA-binding, and RNA-binding regions are included in this dataset, so this evaluation should not be viewed as a general characterization of performance, only the difficulty of the current dataset.

In contrast to binary (a single function) predictions provided by each predictor, regions of the current dataset have mutually exclusive functions, or no function. A simple approach to combining individual binary predictors to create a five-state predictor is to use scores to resolve conflicting predictions, with the highest score being the predicted function. We use z-score based approach described in Section 2.2.2 for that purpose. Table 1 shows that these predictions perform well for the largest region types, structure (MCC = 0.66) and protein binding (MCC = 0.49), and poorly for the other types (MCC $\leq$ 0.06). This is due to poor mean and variance estimation for the regions types with smaller number of samples and the lower quality of the underlying averages (Fig. 1). One feature of these data is that predictors generally do well in distinguishing disordered functional regions from structured regions, but not as well in distinguishing between functions. This suggests that function predictions should be considered simultaneously. It should also be noted that the z-score method does not distinguish structured and disordered regions as well as a dedicated disorder prediction. We tested this using predictions produced by a popular disorder predictor, IUPred [34]. The average region-level IUPred prediction scores have a higher MCC for the structured regions (MCC = 0.76). Note that we chose not to include disorder prediction as structured regions are used as a hard-negative set for disorder-based function; the current problem is not one of structure vs. disorder.

### 3.2. *IDRs described in multidimensional space form function-related clusters*

An unsupervised approach to combining function predictions was explored by clustering regions based on the residue-level predictions and sequence features. Section 2.2.2 provides details. The resulting eight setups (using four different sets of features, each processed directly or after PCA) are compared with each other and with a control experiment where we cluster regions with randomly shuffled region types (Table 2).

Clustering metrics are the strongest for clustering with all PCA-transformed features for NMI and ARI metrics, with purity being also near the maximum observed across all clusterings (Table

2). When comparing NMI across all clusterings (Fig 1A), they have high quality compared to the control, but inclusion of the residue-level prediction-based features results in better clustering than the sequence-based features (amino acid composition) alone. Clustering quality is nearly identical for PCA-transformed and non-transformed residue-level averaged scores, which suggest little redundant information in these features. This contrasts with the features that include the maximum-windowed prediction scores ("prediction-scores only" and "all features") that show modest improvement in clustering quality following PCA transformation, which reveals some redundancy in these features.

Table 2. Clustering and classification results. Region types include structured (S), protein-binding IDRs (P), linker IDRs (L), DNA-binding IDRs (D), and RNA-binding IDRs (R). The best results for each row are in bold font.

| | | | **Z-score** | **Multidimensional unsupervised clustering without PCA** | | | | | **Multidimensional unsupervised clustering with PCA** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Metric | Region type | Function prediction average features | Control: shuffled region types and all features | Sequence features | Function prediction features | Function prediction average features | All features | Control: shuffled region types and all features | Sequence features | Function prediction features | Function prediction average features | All features |
| Clustering Scores | NMI | | | 0.03 | 0.27 | 0.33 | 0.34 | 0.35 | 0.04 | 0.22 | 0.36 | 0.34 | **0.39** |
| | ARI | | | 0.01 | 0.27 | 0.26 | 0.30 | 0.45 | -0.01 | 0.19 | 0.33 | 0.30 | **0.50** |
| | Purity | | | 0.54 | 0.72 | 0.78 | 0.78 | 0.75 | 0.54 | 0.72 | **0.79** | 0.78 | 0.78 |
| Classification Scores | Macro F1 | | 0.34 | 0.25 | 0.33 | 0.36 | 0.41 | 0.46 | 0.22 | 0.31 | 0.41 | 0.41 | **0.51** |
| | Macro MCC | | 0.24 | 0.10 | 0.26 | 0.30 | 0.34 | 0.41 | 0.08 | 0.23 | 0.35 | 0.34 | **0.45** |
| | Micro F1 | | 0.65 | 0.34 | 0.51 | 0.46 | 0.51 | 0.64 | 0.34 | 0.43 | 0.53 | 0.50 | **0.71** |
| | Micro MCC | | 0.56 | 0.18 | 0.39 | 0.33 | 0.39 | 0.55 | 0.18 | 0.29 | 0.41 | 0.38 | **0.64** |
| | Accuracy | | 0.65 | 0.34 | 0.51 | 0.46 | 0.51 | 0.64 | 0.34 | 0.43 | 0.53 | 0.50 | **0.71** |
| | Sensitivity | S | 0.82 | 0.35 | 0.67 | 0.48 | 0.44 | **0.83** | 0.19 | 0.52 | 0.53 | 0.46 | **0.83** |
| | | P | 0.59 | 0.37 | 0.29 | 0.43 | **0.67** | 0.39 | 0.65 | 0.33 | 0.57 | 0.61 | 0.61 |
| | | L | 0.14 | 0.29 | 0.57 | **0.71** | 0.57 | 0.57 | 0.29 | 0.43 | 0.71 | **0.71** | 0.43 |
| | | D | 0.14 | 0.29 | **0.57** | 0.14 | 0.14 | 0.43 | 0.29 | 0.14 | 0.14 | 0.14 | 0.43 |
| | | R | 0.00 | 0.20 | 0.00 | 0.60 | 0.40 | 0.60 | 0.00 | 0.40 | 0.40 | 0.40 | **0.60** |
| | Specificity | S | 0.84 | 0.77 | 0.89 | 0.96 | **0.99** | 0.90 | 0.87 | 0.91 | 0.96 | **0.99** | 0.93 |
| | | P | 0.88 | 0.76 | 0.95 | **0.95** | 0.92 | **0.95** | 0.44 | 0.94 | 0.94 | 0.92 | 0.90 |
| | | L | 0.90 | 0.91 | 0.90 | 0.87 | **0.96** | 0.90 | 0.92 | 0.90 | 0.90 | 0.94 | 0.95 |
| | | D | 0.93 | 0.75 | 0.70 | 0.97 | 0.97 | 0.97 | 0.85 | 0.95 | **0.99** | 0.98 | 0.96 |
| | | R | 0.97 | 0.93 | 0.97 | 0.66 | 0.62 | 0.84 | **0.99** | 0.64 | 0.68 | 0.63 | 0.89 |
| | F1 | S | 0.84 | 0.45 | 0.76 | 0.63 | 0.61 | 0.87 | 0.29 | 0.65 | 0.68 | 0.62 | **0.88** |
| | | P | 0.65 | 0.40 | 0.42 | 0.56 | **0.73** | 0.53 | 0.47 | 0.46 | 0.67 | 0.69 | 0.67 |
| | | L | 0.09 | 0.18 | 0.32 | 0.32 | 0.47 | 0.32 | 0.20 | 0.24 | 0.39 | **0.48** | 0.35 |
| | | D | 0.11 | 0.09 | 0.15 | 0.15 | 0.17 | **0.40** | 0.13 | 0.13 | 0.22 | 0.18 | 0.38 |
| | | R | 0.00 | 0.13 | 0.00 | 0.11 | 0.07 | 0.19 | 0.00 | 0.07 | 0.07 | 0.07 | **0.25** |
| | MCC | S | 0.66 | 0.13 | 0.56 | 0.49 | 0.50 | 0.73 | 0.08 | 0.46 | 0.53 | 0.51 | **0.76** |
| | | P | 0.49 | 0.14 | 0.34 | 0.47 | **0.62** | 0.44 | 0.08 | 0.36 | 0.57 | 0.57 | 0.54 |
| | | L | 0.03 | 0.14 | 0.31 | 0.34 | 0.45 | 0.31 | 0.16 | 0.21 | 0.39 | **0.47** | 0.32 |
| | | D | 0.06 | 0.02 | 0.12 | 0.12 | 0.14 | **0.37** | 0.08 | 0.09 | 0.25 | 0.16 | 0.34 |
| | | R | -0.03 | 0.09 | -0.03 | 0.10 | 0.01 | 0.21 | -0.02 | 0.02 | 0.03 | 0.01 | **0.27** |

Clusters were used directly for classification of regions into mutually exclusive function classes by assigning regions types that maximize the overall MCC. For overall classification performance – macro and micro F1 and MCC, and accuracy (Table 2) – clustering on all PCA transformed features shows the best performance. Note that these performance metrics include both unbalanced – micro F1, micro MCC, and accuracy – and class balanced – macro F1 and MCC – metrics. Examining the macro MCC across all clusterings and z-score normalization-based classification (Fig. 2B) shows the relative contribution of each feature set. Macro MCC values of each clustering are consistent with cluster quality measured by NMI (Fig. 2B), with the residue-level prediction-

based features outperforming sequence-based features. Importantly, the results for the control experiments (where we shuffle the region types) are poor (macro MCC ≤ 0.1) which suggest that our models do not overfit the dataset. The best results are secured when combining both feature categories, which suggests that they provide complementary predictive input. Moreover, as expected [29], use of PCA boosts the predictive performance. The macro and micro MCCs of the best solution are 0.45 and 0.64, respectively, which corresponds to high correlation between the predicted and the native region types.

We quantify significance of differences between the best solution (clustering with PCA and all features) and each of the other 10 results (including controls and z-score approach) with *t*-test for normal data and Wilcoxon test otherwise; we assess normality with the Anderson-Darling test at 0.05 significance. We draw 50% of regions at random 10 times and compare the corresponding 10 sets of results; this reveals whether the improvements are robust to different datasets. The micro F1, micro MCC and accuracy of the best solution are significantly higher than the 10 alternatives ($p$-value < 0.05). The macro F1 and macro MCC of the best method are higher but the difference is not significant when compared to the clustering with all features and no PCA ($p$-value = 0.21 and 0.20, respectively), while the difference is significant for the other 9 alternatives ($p$-value < 0.05). Overall, this reveals that the approach that uses all features together offers significantly better results.

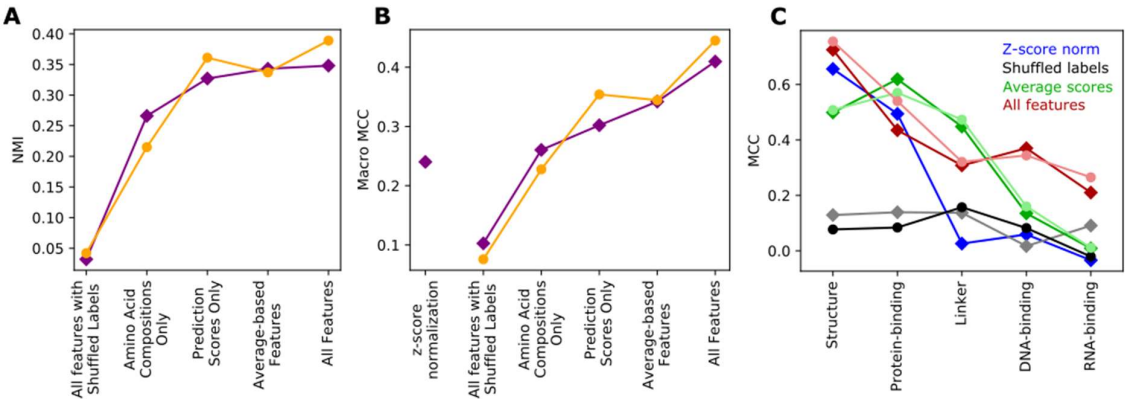


Figure 2. Clustering and classification quality. Clustering quality measured by NMI (A) and overall classification quality measured by macro MCC (B) are shown for both PCA-transformed (circles and orange lines) and non-transformed (diamond and purple lines) results. (C) Per-class MCC are shown for z-score (blue) and clustering results – both PCA-transformed (circles) and non-transformed (diamond) data for all features with shuffled labels (black and grey), function prediction average features (pink and red), and all features (light green and green).

Classification performance was also evaluated per-region type, where each region type in turn is taken as the positive prediction and the others negative, and evaluated by sensitivity, specificity, F1, and MCC (Table 2). The all PCA-transformed features clustering (the best solution) has good per-class performance across all classes (MCC ≥ 0.27), with the best performance for structured and RNA-binding regions by all metrics except specificity. Comparing the per-class MCC values for a subset of clusterings (Fig. 2C) shows that, generally, prediction performance is better for the larger classes – structure and protein-binding regions – than for the smaller classes – linker, DNA-binding, and RNA-binding regions.

To visualize clusters in the feature space, dimensions of the full feature set were reduced to two dimensions using second degree polynomial kernel-based PCA [35] (Fig. 3). Note that in this representation distances between points are not meaningful, due to the arbitrary non-linear

transformation. We observe an excellent separation between the cluster of structured regions (in gray) and the four clusters with the disordered regions (color-coded in red, yellow, green and blue). Though protein-binding regions are the most common contaminant in other function clusters, this is likely attributable to the relatively large size of this set relative to the other functional sets. The RNA-binding (green) and DNA-binding (blue) clusters are close to each other, which possibly reflects the relatively similar nature of these interactions. Comparing DNA- and RNA-prediction scores for DNA- and RNA-binding regions shows that these two predictions are modestly correlated for these region types (r=0.54), but these prediction scores are not correlated over the entire dataset (r=0.08). However, examining region cluster assignments, while 4 of the DNA-binding regions and 2 of the RNA-binding regions are incorrectly assigned, there is no DNA/RNA cluster cross assignment. This indicates that, while DNA- and RNA- binding disordered regions share sequence features, simultaneous consideration of function predictions provides sufficient information to distinguish these functions.

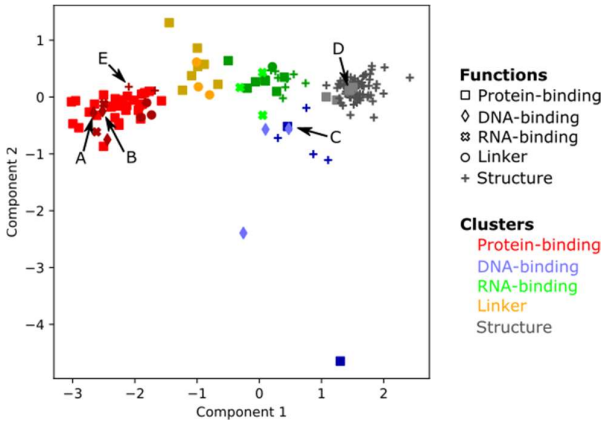


Fig. 3. Visualization of the clustering based on the all PCA-transformed features. Cluster membership is indicated by color and the region type is represented by shape. Correctly (incorrectly) classified regions are represented with lighter (darker) colors. Indicated regions (A, B, C, D, and E) are discussed in the case studies section. Axes are non-linear combinations of features derived from polynomial kernel-based PCA.

### 3.3. *Case studies*

Clustering results can be used to gain potentially novel functional insights and include misclassifications. We examined several examples of each (Fig. 3, indicated points).

We show three cases of function insights that stem from disagreement between cluster labels and annotated region types. The homeobox protein NKX3.1 (DisProt ID DP00683) has been characterized as largely disordered [36], including a homeodomain that interacts with DNA. The region N-terminal to the homeodomain contains an acidic region that interacts transiently with the homeodomain [36], where transient interactions are one mode of IDR protein-protein interactions. This IDR (Fig. 3, point A) is annotated only as DNA-binding, but its placement in a protein-binding cluster could be due to the above interaction. Another IDR misclassified this way is a single-stranded DNA-binding (SSB) protein from *T. maritima* (DisProt ID DP00996), which contains several disordered regions interspersed with a structured domain [37]. There is no direct functional evidence associated with these regions; annotations have apparently been made based on association, i.e. IDRs in an SSB. The C-terminal IDR in this protein (Fig. 3, point B) is an extension to the canonical SSB, where such an extension in *E. coli* has been found to transiently interact with the DNA binding surface and may recruit other proteins [38]. In both of these cases, these regions are likely multi-functional DNA- and protein-binding IDRs, where annotations do not reflect the protein-binding functions. Another example is a protein-binding IDR classified as DNA-binding IDR. The Antitoxin CcdA (DP00928) contains an C-terminal IDR (Fig. 3 point C) that interacts with the toxin

CccdB [39] and an N-terminal DNA binding domain [40]. The definition of the IDR region used by De Jonge *et al.* [39] overlaps with the DNA binding domain. A different definition of the C-terminal domain is shorter by 9 residues [40] eliminating the overlap. Removing these residues increases the distance to the DNA-binding cluster center, suggesting this misclassification may be due to inaccurate definition of this domain.

One of the reasons for incorrect classification of region types could be a lack protein-level context in the region level prediction. The structure of C5a peptidase (PDB ID 3EIF) contains five domains, catalytic, protease-associated, and three fibronectin type III (FN) domains. A 36 residue region from the third of the FN domains is included in our structured region set (Fig. 3, point E). This region is highly negatively charged, contributing to a large electronegative surface that plays a role in protein substrate recruitment [41]. The highly charged nature of this sequence, its role in protein interactions, and the lack of consideration of the larger sequence context in our approach likely contributes to misclassification of this region as a protein-binding IDR. An example of a protein-binding IDR classified as structured region is the N-terminal IDR from the photosystem I subunit psaH (DisProt ID DP00803, Fig. 3, point D), which is responsible for binding to other proteins. The C-terminus of psaH contains several transmembrane helices with the IDR proximal to the membrane [42]. The IDRs of transmembrane proteins often differ significantly from the IDRs of soluble proteins [43]. It is possible that the context of this IDR may be influencing its classification.

## 4. Conclusions

Our goal with this work was to investigate the feasibility of simultaneous classification of multiple IDR functions. For this goal, we adapted residue-level predictors to the region-level task. We carefully selected a difficult dataset, which had no protein similar to training proteins of all the prediction methods. This ensured that results would not be deterministic due to any possible predictor bias. The results presented here demonstrate that existing predictors are useful for the region level-predictions and that simultaneous multi-region-type prediction is feasible. We also show that this prediction benefits from the inclusion of the information extracted directly from the sequence of the regions. Development of an accurate tool for prediction of multiple functions of IDRs would require combining results of multiple residue-level predictors that address different functions, would likely benefit from a supervised approach, and needs a larger dataset to derive sufficiently large training and test sets.

There are several practical issues to consider in the development of such a tool. Many disordered regions are likely to be multi-function, which contrasts to the mutually exclusive approach taken here. While we attempted to use single function IDRs, in fact some of them are multifunctional, which we describe in section 3.3. An IDR annotation tool and the associated evaluation protocol must be able to contend with inaccuracy and incompleteness of functional annotations of IDRs. Also, inaccurate IDR definitions can introduce extraneous residues or leave out crucial residues that can complicate function prediction. Moreover, consideration of the sequence context of an IDR may be helpful for accurate function prediction. Finally, practices for annotating IDRs is to reflect the regions determined by experiments. There is little theory and no methods for dividing large IDRs in to functional domains. Since IDRs can reach lengths of a thousand residues [44], it seems likely that functional domains exist. Multifunctional IDRs can also be due to pleiotropy, e.g. acting as both a linker and binding to other proteins [45]. These considerations suggest that a multistate approach may be more appropriate for IDR function prediction.

## Acknowledgments

This research was supported in part by the National Science Foundation (grant 1617369) and the Robert J. Mattauch Endowment funds.

## References

1. Habchi, J., et al., *Introducing protein intrinsic disorder.* Chem Rev, 2014. **114**(13): p. 6561-88.
2. Lieutaud, P., et al., *How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe.* Intrinsically Disord Proteins, 2016. **4**(1): p. e1259708.
3. Oldfield, C.J., et al., *Chapter 1 - Introduction to intrinsically disordered proteins and regions*, in *Intrinsically Disordered Proteins*, N. Salvi, Editor. 2019, Academic Press. p. 1-34.
4. Dunker, A.K., et al., *Intrinsic disorder and protein function.* Biochemistry, 2002. **41**(21): p. 6573-82.
5. Peng, Z., M.J. Mizianty, and L. Kurgan, *Genome-scale prediction of proteins with long intrinsically disordered regions.* Proteins, 2014. **82**(1): p. 145-58.
6. Peng, Z., et al., *Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life.* Cell Mol Life Sci, 2015. **72**(1): p. 137-51.
7. Yan, J., et al., *RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale.* Biochim Biophys Acta, 2013. **1834**(8): p. 1671-80.
8. Piovesan, D., et al., *DisProt 7.0: a major update of the database of disordered proteins.* Nucleic acids research, 2017. **45**(D1): p. D219-D227.
9. Piovesan, D., et al., *MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins.* Nucleic acids research, 2018. **46**(D1): p. D471-D476.
10. Meng, F., V.N. Uversky, and L. Kurgan, *Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions.* Cell Mol Life Sci, 2017. **74**(17): p. 3069-3090.
11. Meszaros, B., I. Simon, and Z. Dosztanyi, *Prediction of protein binding regions in disordered proteins.* PLoS Comput Biol, 2009. **5**(5): p. e1000376.
12. Yan, J., et al., *Molecular recognition features (MoRFs) in three domains of life.* Molecular BioSystems, 2016. **12**(3): p. 697-710.
13. Peng, Z. and L. Kurgan, *High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder.* Nucleic Acids Research, 2015. **43**(18): p. e121-e121.
14. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity.* Bioinformatics (Oxford, England), 2015. **31**(6): p. 857-863.
15. Katuwawala, A., et al., *Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions.* Comput Struct Biotechnol J, 2019. **17**: p. 454-462.
16. Mizianty, M.J., et al., *Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources.* Bioinformatics, 2010. **26**(18): p. i489-96.
17. Dosztányi, Z., B. Mészáros, and I. Simon, *ANCHOR: web server for predicting protein binding regions in disordered proteins.* Bioinformatics (Oxford, England), 2009. **25**(20): p. 2745-2746.
18. Oldfield, C.J., V.N. Uversky, and L. Kurgan, *Predicting Functions of Disordered Proteins with MoRFpred.* Methods Mol Biol, 2019. **1851**: p. 337-352.
19. Meng, F. and L. Kurgan, *DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences.* Bioinformatics, 2016. **32**(12): p. i341-i350.
20. Fukuchi, S., et al., *IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature.* Nucleic acids research, 2012. **40**(Database issue): p. D507-D511.
21. Piovesan, D., et al., *MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins.* Nucleic Acids Res, 2018. **46**(D1): p. D471-D476.
22. Oates, M.E., et al., *D(2)P(2): database of disordered protein predictions.* Nucleic Acids Res, 2013. **41**(Database issue): p. D508-16.
23. Burley, S.K., et al., *Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive.* Methods in molecular biology (Clifton, N.J.), 2017. **1607**: p. 627-641.

24. Dana, J.M., et al., *SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins.* Nucleic Acids Res, 2019. **47**(D1): p. D482-D489.
25. Peng, Z., et al., *Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind.* Methods Mol Biol, 2017. **1484**: p. 187-203.
26. Meng, F. and L. Kurgan, *High-throughput prediction of disordered moonlighting regions in protein sequences.* Proteins, 2018. **86**(10): p. 1097-1110.
27. Huang, Y., et al., *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 2010. **26**(5): p. 680-682.
28. Jolliffe, I.T. and J. Cadima, *Principal component analysis: a review and recent developments.* Philos Trans A Math Phys Eng Sci, 2016. **374**(2065): p. 20150202.
29. Ding, C. and X. He, *K-means clustering via principal component analysis*, in *Proceedings of the twenty-first international conference on Machine learning*. 2004, ACM: Banff, Alberta, Canada. p. 29.
30. Jain, A.K., *Data clustering: 50 years beyond K-means.* Pattern Recognition Letters, 2010. **31**(8): p. 651-666.
31. Manning, C., P. Raghavan, and H. Schütze, *Introduction to information retrieval.* 2010. **16**: p. 356-360.
32. Pfitzner, D., R. Leibbrandt, and D. Powers, *Characterization and evaluation of similarity measures for pairs of clusterings.* Knowledge and Information Systems, 2009. **19**: p. 361-394.
33. Morey, L.C. and A. Agresti, *The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement.* Educational and Psychological Measurement, 1984. **44**(1): p. 33-37.
34. Meszaros, B., G. Erdos, and Z. Dosztanyi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding.* Nucleic Acids Res, 2018. **46**(W1): p. W329-W337.
35. Schölkopf, B., A. Smola, and K.-R. Müller, *Nonlinear Component Analysis as a Kernel Eigenvalue Problem.* Neural Computation, 1998. **10**: p. 1299-1319.
36. Ju, J.H., et al., *Interactions of the acidic domain and SRF interacting motifs with the NKX3.1 homeodomain.* Biochemistry, 2009. **48**(44): p. 10601-7.
37. DiDonato, M., et al., *Crystal structure of a single-stranded DNA-binding protein (TM0604) from Thermotoga maritima at 2.60 A resolution.* Proteins, 2006. **63**(1): p. 256-60.
38. Kozlov, A.G., M.M. Cox, and T.M. Lohman, *Regulation of single-stranded DNA binding by the C termini of Escherichia coli single-stranded DNA-binding (SSB) protein.* J Biol Chem, 2010. **285**(22): p. 17246-52.
39. De Jonge, N., et al., *Rejuvenation of CcdB-poisoned gyrase by an intrinsically disordered protein domain.* Mol Cell, 2009. **35**(2): p. 154-63.
40. Madl, T., et al., *Structural basis for nucleic acid and toxin recognition of the bacterial antitoxin CcdA.* J Mol Biol, 2006. **364**(2): p. 170-85.
41. Kagawa, T.F., et al., *Model for Substrate Interactions in C5a Peptidase from Streptococcus pyogenes: A 1.9 Å Crystal Structure of the Active Form of ScpA.* Journal of Molecular Biology, 2009. **386**(3): p. 754-772.
42. Amunts, A., et al., *Structure determination and improved model of plant photosystem I.* J Biol Chem, 2010. **285**(5): p. 3478-86.
43. Xue, B., et al., *Analysis of structured and intrinsically disordered regions of transmembrane proteins.* Molecular bioSystems, 2009. **5**(12): p. 1688-1702.
44. Mark, W.Y., et al., *Characterization of segments from the central region of BRCA1: an intrinsically disordered scaffold for multiple protein-protein and protein-DNA interactions?* J Mol Biol, 2005. **345**(2): p. 275-87.
45. Rumi-Masante, J., et al., *Structural basis for activation of calcineurin by calmodulin.* J Mol Biol, 2012. **415**(2): p. 307-17.