# Appendix for "Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources"

by Marcin J. Mizianty, Wojciech Stach, Ke Chen, Kanaka Durga Kedarisetti, Fatemeh Miri Disfani and Lukasz Kurgan

## Feature Space Representation

The proposed MFDp method utilizes four complementary disorder predictors, IUPred LONG (IUPREDL), IUPred SHORT (IUPREDS), DISOPRED2 and DISOclust; the sequence, the PSSM profile, the predicted secondary structure (SS), the predicted relative solvent accessibility (RSA), the predicted globular domains (IUPREDG), backbone dihedral torsion angles, and signal peptides. The predictors were run with their default parameters. The PSSM profiles were generated using the non-redundant (nr) database from NCBI (downloaded on Nov 19th 2009), which was filtered using PFILT (Jones & Swindells, 2002) to remove low-complexity regions, transmembrane regions, and coiled-coil segments.

We utilize a sliding window of size 15 centered over the predicted residue to extract the features. We use the raw numerical values for each of the 15 positions, which include the probability of the prediction of disordered residues from IUPREDL, IUPREDS, DISOPRED2 and DISOclust, the PSSM values, the probabilities of prediction of helix, coil and strand con-formations from PSIPRED, the predicted B-factor values, the predicted solvent accessibility and backbone angles, and the binary values denoting whether a given position in the window is predicted as a signal peptide, belongs to a globular domain, is part of a sequence tag, or is predicted as flexible by PROFbval, to encode the features. We also aggregate the raw numerical values (except for the predicted backbone angles) by computing averages in the sliding widows of sizes between 3 and 41, and by taking maximal and minimal value in the sliding widows of sizes between 3 and 15. The averages for the residue at the $i^{th}$ position in the sequence are computed as follows:

$$X\_avg\{ws_m\} = \sum_{pos=\max(i-\frac{ws_m}{2},0)}^{\min(i+\frac{ws_m}{2},seqLen)} \frac{X\_w\{pos\}}{m}$$

where $X \in \{PSSM\_\{AA_j\}, SS\_\{SS_k\}, RSA$, IUPREDL, IUPREDS, *DISOPRED2*, *DISOCLUST*$\}$ is the name of a given feature, $AA_j$ with $j \in \{1, 2, \ldots, 20\}$ is the amino acid type, $SS_k$ with $k \in \{$H, E, C$\}$ is the type of the secondary structure, $ws_m$ with $m \in \{3, 5, 7, .., 39, 41\}$ is the size of the window, $w_n$ with $n \in \{-20, -19, \ldots, 0, 1, \ldots, 19, 20\}$ is the position in the window where 0 denotes the position of the predicted residue, and $X\_w\{w_i\}$ is the value of feature $X$ for the residue at position $w_n$ (the actual position in the sequence equals $pos = i+w_n$). When aggregating, we ignore the positions in the window that are outside of the chain for the residues close to the sequence terminus, i.e., we sum only the values for the positions in the chain and this sum is dived by the total number of these positions.

We have two categories of features, which are based on the numerical and binary predictions, respectively:

1. *Features derived from numerical predictions*:
   - $XXX\_w\{w_n\}$ – *XXX value a* at $w_n$ position in the sliding window (15 features).
   - $XXX\_avg\{ws_m\}$ – Average *XXX* value for the window of size $ws_m$ (20 features).
   - $XXX\_min\{ws_m\}$ – Minimal *XXX* value for the window of size $ws_m$ (7 features).
   - $XXX\_max\{ws_m\}$ – Maximal *XXX* value for the window of size $ws_m$ (7 features).

   where *XXX* stands for normalized PSSM value for $AA_j$ ($PSSM\_\{AA_j\}$), weighted observed percentages of the actual AA (FREQ), information per position (INFO) generated by PSI-BLAST(Altschul et al., 1997); predicted

disorder probability by one of four methods (DISOCLUST, DISOPRED2, IUPREDL and IUPREDS), predicted relative solvent accessibility (RSA), or predicted B-factor (BVAL).

2. *Features derived from binary predictions*:
   *YYY_w{$w_n$}* – Binary features that encode whether the residue at $w_n$ position in the sliding window is a part of the *YYY*. Values of 0 / 1 denote that the residue is not / is a part of the *YYY*, respectively. For window positions outside of a protein we use 0 (15 features).
   *YYY_3res_{111, 100* or 001 or 110 or 011*, 000}* – Binary feature that encodes the information concerning *YYY* for a tripeptide centered on the predicted residue. Two conformations (111 and 000) denote that the residue is inside of a *YYY* or inside a non-*YYY* segment, and the remaining conformations correspond to the termini of the *YYY* (3 features).
   *YYY_dist* – Linear distance to the closest predicted *YYY*. The value is set to 1 if there is no non-*YYY* within 10 residues, -1 if there is no *YYY* within 10 residues, and it equals SIGN**dist*/10 where *dist* < 10 is the number of residues to the closest *YYY*/non-*YYY* residue, SIGN = 1 if a given residue is *YYY* and -1 if a given residue is non-XXX (1 feature).
   *YYY_content_w{$ws_m$}* – number of predicted *YYY* in window of size $ws_m$ (20 features).
   Where *YYY* stands for signal peptide (SIG), tag (TAG), globular domain (IUPREDG) or flexible residue predicted by PROFbval in Strict (BVAL_S) or non-strict (BVAL_NS) modes.

The abovementioned features are divided into seventeen categories, as described below.

***Sequence-based features*** (5 features).
*SEQ_length – length of the sequence.*
*SEQ_*abs*DistFromNTerm* – Linear distance from N termini (1 feature).
*SEQ_absDistFromTerm* – Linear distance from the closest termini (1 feature).
*SEQ_*r*elDistFromNTerm – SEQ_*abs*DistFromNTerm* divided by a sequence's length (1 feature).
*SEQ_relDistFromTerm – SEQ_absDistFromTerm* divided by the half of a sequence's length (1 feature).

***PSSM features*** (49*20 = 980 features). This category groups features derived from the PSSM profile generated by the PSI BLAST (Altschul et al., 1997). It contains of features derived for numerical predictions, where *XXX* is replaced by *PSSM_{$AA_j$} which stands for* normalized PSSM value for $AA_j$. The PSSM values were normalized using $1/(1+e^{-cs})$ where *cs* stands for the conservation score. For window positions outside of a protein we use 0.

***AA frequency*** (15 + 20 + 7 +7 = 49 features). This set includes 49 features derived for numerical predictions derived from the matrix of weighted observed percentages generated by the PSI BLAST, where *XXX* was replaced by weighted observed percentages of the actual AA.

***Conservation information*** (49 features). These features are based on the conservation information per position generated by the PSI BLAST. It contains of features derived for numerical predictions where *XXX* is replaced by information per position.

***Secondary structure*** (49*3 +9 + 6 + 6 + 13= 181 features). This category covers features derived from the secondary structure predicted by the PSIPRED. These features are computed for numerical predictions where *XXX* is replaced by *SS_{$SS_k$}* - predicted probability of $SS_k$. For window positions outside of the protein we use 0 for H and E, and 1 for C. For this category we alse computed following features:
*SS_3res_{CCC, CCH* or *CHH, HHH, CEE* or *CCE, EEE, CEC, ECE, HCH, HCE}* – Binary feature that encodes the predicted secondary structure of a tripeptide centered on the predicted residue. Three conformations (CCC, EEE, and HHH) denote that the residue is inside of a secondary structure segment and the remaining six that it is on the interface between two segments (9 features).
*SS_{$SS_k$}_seg_size_{nAvg, n80%}* – Size of the $SS_k$ segment that includes the predicted residue. If $SS_k$ is different than the predicted secondary structure of the predicted residue then the feature is set to 0, otherwise the size of the segment is normalized either by the average size of the $SS_k$ segments in the training dataset (10, 11, 4 for C, H, and E segments, respectively) or by the size at 80% of segments sizes in the training dataset

(16, 20, and 8, respectively). If a given segment is longer than the normalization value, then the value is set to 1 (2*3 = 6 features).

*SS_{$SS_k$}_seg _pos_{nAvg, n80%}* – Position inside of the $SS_k$ segment that includes the predicted residue. Value of 0 means that $SS_k$ is different than the predicted secondary structure of the predicted residue or that the residue is at the termini of the segment. Value of 1 means that the residue is in the center of the segment. Values for positions between the center and the termini are scaled linearly between 1 and 0 (2*3 = 6 features).

*SS_seg_{HCH, HCE, ECH, ECE, HCT, TCH, TCE, CHC, CEC, THC, CHT, TEC, CET}* – Binary feature that encodes the configuration of secondary structure segments adjacent to the segment that includes the predicted residue, e.g., HCH means that the predicted residue is in a coil segment adjacent to two helices on both sides. The segment types include H (helix), C (coil), E (strand), and T (terminus) (13 features).

***Relative solvent accessibility*** (22 features). These features are computed for numerical predictions were *XXX* is the RSA value predicted by the Real-SPINE3. We normalize the ASA values predicted by Real-SPINE3 using the maximal ASA values provided in (Faraggi et al., 2009).

***Torsion angles PHI*** (30 features). This category concerns PHI and PSI torsion angles predicted by the Real-SPINE3. We normalize the predicted angles to (-1; 1) interval using the following formula (angle + 180)/360. For this category we computed following features:

*{PHI,PSI}_w{$w_n$}* – Normalized predicted PHI/PSI torsion angle for residue at $w_n$ position. For positions outside of a protein we use 0 (15 features).

***Disorder predictions* DISOPRED2** (49 features). These features were computed for numerical predictions where *XXX* stands for probability of disorder predicted by the DISOPRED2.

***Disorder predictions DISOclust*** (49 features). These features were computed for numerical predictions where *XXX* stands for probability of disorder predicted by the *DISOclust*.

***Disorder predictions* IUPREDL** (49 features). These features were computed for numerical predictions where *XXX* stands for probability of disorder predicted by the IUPREDL.

***Disorder predictions* IUPREDS** (49 features). These features were computed for numerical predictions where *XXX* stands for probability of disorder predicted by the IUPREDS.

***B-factor predictions*** (49 features). These features were computed for numerical predictions where *XXX* stands for predicted B factor by PROFbval.

***Signal peptides*** (39 features). These features were computed for binary predictions were *XXX* concern signal peptides predicted by the RPSP.

***Tag*** (39 features). These features were computed for binary predictions were *XXX* are N- or C- termini tags which are introduced to ease the protein purification.

***Globular domain*** (39 features). These features were computed for binary predictions were *XXX* are globular domains predicted by IUPRED.

***Flexible residue in strict mode*** (39 features). These features were computed for binary predictions were *XXX* are flexible residues predicted by PROFbval in strict mode.

***Flexible residue in non-strict mode*** (39 features). These features were computed for binary predictions were *XXX* are flexible residues predicted by PROFbval in non-strict mode.

**Table 1.** Summary of selected features for the ALL, SHORT, and LONG models. The features are sorted by their average (over the five training subsets of the MxD dataset) biserial correlations with the with the outcomes, i.e., annotation of ordered and disorder residues

Features selected for the ALL model

| Feature name | Biserial correlation | Category | Aggregated features |
|---|---|---|---|
| IUPREDS_avg_w41 | 0.426 | iupreds | x |
| IUPREDL_avg_w29 | 0.420 | iupredl | x |
| DISOCLUST_avg_w25 | 0.418 | disoclust | x |
| DISOPRED_max_w15 | 0.404 | disopred | |
| SPINE_RSA_avg_w41 | 0.393 | spine_rsa | x |
| SPINE_RSA_min_w15 | 0.389 | spine_rsa | |
| SPINE_RSA_avg_w23 | 0.377 | spine_rsa | x |
| SPINE_RSA_min_w9 | 0.370 | spine_rsa | |
| SS_C_min_w15 | 0.368 | ss | |
| BVAL_NS_content_w41 | 0.365 | bval_ns | x |
| SPINE_RSA_avg_w13 | 0.350 | spine_rsa | x |
| SPINE_RSA_min_w5 | 0.330 | spine_rsa | |
| PSSM_avg12_G | 0.310 | pssm | x |
| BVAL_avg_w41 | 0.223 | bval | x |
| TAG_dist | 0.105 | tag | |
| SEQ_RelDistFromNTerm | -0.096 | seq | |

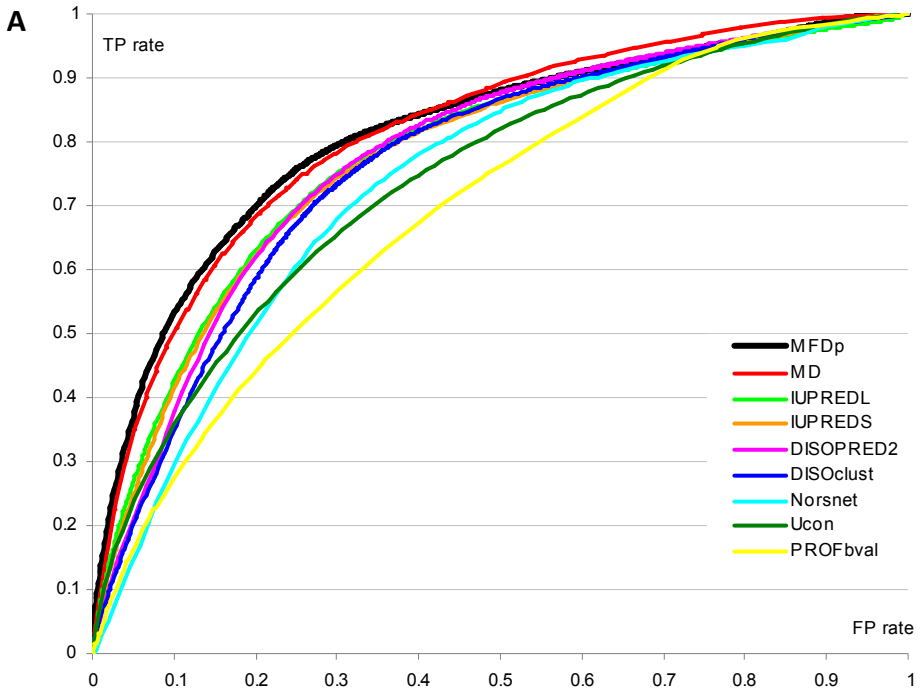
Features selected for the SHORT model

| Feature name | Biserial correlation | Category | Aggregated features |
|---|---|---|---|
| TAG_dist | 0.314 | tag | |
| TAG_content_w11 | 0.306 | tag | x |
| BVAL_max_w15 | 0.245 | bval | |
| IUPREDS_max_w15 | 0.216 | iupreds | |
| BVAL_max_w7 | 0.214 | bval | |
| BVAL_w-4 | 0.188 | bval | |
| BVAL_avg_w17 | 0.180 | bval | x |
| SS_seg_TCE | 0.173 | ss | |
| DISOCLUST_max_w9 | 0.170 | disoclust | |
| BVAL_w1 | 0.166 | bval | |
| BVAL_avg_w31 | 0.164 | bval | x |
| SPINE_RSA_min_w7 | 0.161 | spine_rsa | |
| SPINE_RSA_min_w5 | 0.157 | spine_rsa | |
| SPINE_RSA_avg_w11 | 0.150 | spine_rsa | x |
| SPINE_RSA_min_w11 | 0.150 | spine_rsa | |
| DISOPRED_max_w15 | 0.144 | disopred | |
| SPINE_RSA_min_w3 | 0.143 | spine_rsa | |
| BVAL_w4 | 0.141 | bval | |
| SEQ_RelDistFromNTerm | -0.140 | seq | |
| SPINE_RSA_avg_w19 | 0.138 | spine_rsa | x |
| SPINE_RSA_avg_w5 | 0.136 | spine_rsa | x |
| BVAL_S_content_w11 | 0.136 | bval_s | x |
| BVAL_S_content_w19 | 0.130 | bval_s | x |
| SPINE_RSA_avg_w3 | 0.125 | spine_rsa | x |
| DISOCLUST_min_w13 | 0.107 | disoclust | |
| SPINE_RSA_avg_w33 | 0.105 | spine_rsa | x |
| IUPREDL_max_w13 | 0.099 | iupredl | |
| PSSM_avg40_C | -0.091 | pssm | |
| SPINE_RSA_w0 | 0.089 | spine_rsa | |
| IUPREDG_111 | -0.056 | iupredg | |

Features selected for the LONG model

| Feature name | Biserial correlation | Category | Aggregated features |
|---|---|---|---|
| IUPREDL_avg_w41 | 0.448 | iupredl | x |
| DISOCLUST_avg_w41 | 0.435 | disoclust | x |
| IUPREDS_avg_w41 | 0.435 | iupreds | x |
| IUPREDG_content_w41 | -0.435 | iupredg | x |
| IUPREDG_000 | 0.425 | iupredg | |
| IUPREDG_111 | -0.425 | iupredg | |
| DISOPRED_max_w15 | 0.410 | disopred | |
| BVAL_NS_content_w41 | 0.383 | bval_ns | x |
| SS_C_min_w15 | 0.380 | ss | |
| BVAL_NS_content_w25 | 0.345 | bval_ns | x |
| PSSM_avg12_G | 0.338 | pssm | x |
| BVAL_NS_content_w15 | 0.304 | bval_ns | x |
| IPP_max_w15 | -0.212 | ipp | |
| FREQ_min_w13 | 0.155 | freq | |
| FREQ_min_w9 | 0.154 | freq | |
| PSI_w0 | 0.104 | spine_ang | |
| SEQ_RelDistFromNTerm | -0.060 | seq | |

**Figure 1.** ROCs for the predictions on the (A) MxD and (B) CASP8 datasets.



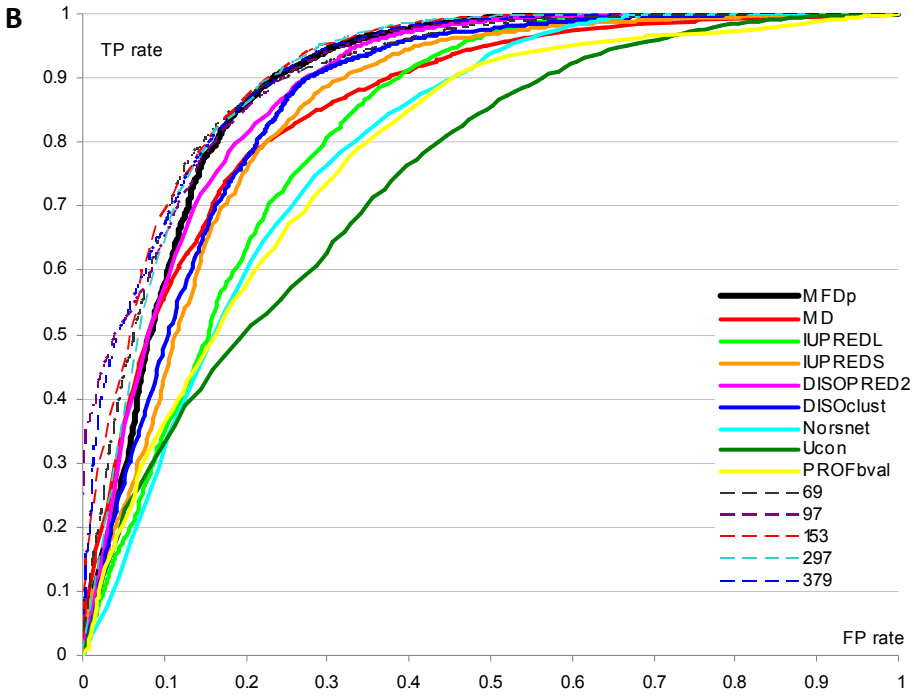



**Figure 2.** ROCs for the predictions of proteins with long disordered segments on the MxD dataset.