MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins

Marco Necci[1], Damiano Piovesan[1], Zsuzsanna Dosztányi[2,3] and Silvio C.E. Tosatto[1,4,*]

[1]Department of Biomedical Sciences, University of Padua, *Viale G. Colombo 3, 35121 Padova, Italy*. [2]MTA-ELTE Lendület Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/c , Budapest, Hungary. [3]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, PO Box 7,H-1518 Budapest, Hungary. [4]CNR Institute of Neuroscience, *Viale G. Colombo 3,* 35121 *Padova, Italy.*

Supplementary Material

Performance Measures

Performances of single prediction methods against the reference are provided by a series of metrics commonly used to evaluate classifiers, namely balanced accuracy, precision, sensitivity, specificity, F1-score, Matthews correlation coefficient (MCC), false positive regions (FPreg), true positive regions (TPreg) and predicted regions (PredRegs). Almost all metrics are calculated from the confusion matrix, which can be obtained from the comparison of predictions and reference. All predictions and all reference protein states, Positive (P) for ID and Negative (N) for structure, are concatenated and a single confusion matrix is calculated from these single sequences, resulting in per-residue statistics. Notice that those residues positions that in the reference correspond to unknown annotation (see (Walsh *et al.*, 2015)), are filtered out from both reference and prediction sequences.

Precision or Positive Predictive Value (PPV) is calculated as:

$$\frac{TP}{TP+FP}$$

Sensitivity, also called recall or True Positive Rate (TPR) is calculated as:

$$\frac{TP}{TP+FN}$$

Specificity, also known as True Negative Rate (TNR) is calculated as:

$$\frac{TN}{TN+FP}$$

Balanced accuracy is the arithmetic mean of recall and specificity, while F1-score is the harmonic mean of precision and sensitivity

$$Balanced\ accuracy = \frac{Specificity + Sensitivity}{2} \qquad F1 = \frac{2 \cdot Precision \cdot Sensitivity}{Precision + Sensitivity}$$

Both F1-score and balanced accuracy are metrics for classifier evaluation, that, to some extent, handle class imbalance. Depending of which of the two classes (N or P) outnumbers the other, each metric outperforms the other: if N >> P → F1 is better; if P >> N → balanced accuracy is better. The MCC is a correlation coefficient between the observed and predicted binary classifications. It is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. It is calculated as:

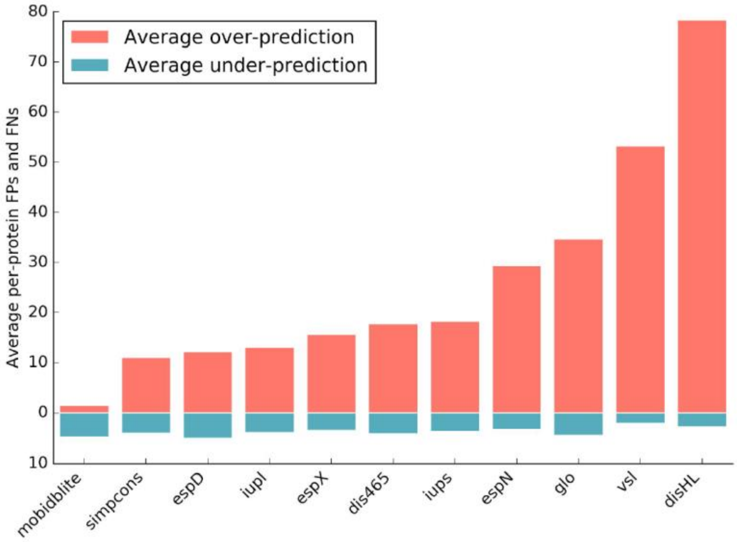
$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The FPreg and TPreg metrics are different from those explained above since their computation is not based on the confusion matrix. To compute these values, all predicted IDRs of at least 3 residues are aligned to all observed IDRs of at least 3 residues. (NB: This threshold is raised to 20 when considering only long disordered regions) TPreg is the number of predicted regions whose overlap with a reference region is at least 50% and FPreg is the number of predicted regions whose overlap with a reference region is less than 50%. PredRegs is the total number of regions predicted by a method.

Supplementary Table 1. Performance comparison for short and long disorder.

Method	Balanced Accuracy	Precision	Recall	Specificity	F1	MCC	FPreg	TPreg	PredRegs
DisEmbl-465	0.56	0.17	0.17	0.96	0.17	0.12	2,951	1,502	7,245
DisEmbl-HL	0.72	0.23	0.54	0.90	0.33	0.30	40,836	14,771	80,788
ESpritz Disprot	0.73	0.38	0.51	0.95	**0.43**	**0.40**	16,108	13,728	46,754
ESpritz NMR	0.70	0.32	0.46	0.95	0.38	0.34	30,286	15,181	70,230
ESpritz Xray	0.69	0.11	0.67	0.71	0.19	0.18	100,367	15,225	176,411
Globplot	0.60	0.13	0.33	0.87	0.18	0.13	44,127	5,764	75,750
IUPred long	0.63	0.28	0.31	0.96	0.29	0.25	19,835	5,460	34,574
IUPred short	0.71	0.32	0.48	0.94	0.38	0.35	24,408	14,251	59,710
VSL2b	**0.76**	0.17	**0.70**	0.81	0.27	0.28	59,648	**15,987**	117,302
Consensus	0.70	0.44	0.42	0.97	**0.43**	**0.40**	14,546	12,535	39,617
MobiDB-lite	0.57	**0.63**	0.14	**1.00**	0.23	0.28	**568**	1,468	**2,338**

All values are shown as percentages. The top performing method in each category is shown in bold and the second best underlined. All ID residues are considered regardless of the disorder region length (i.e. length cutoff = 0).



Supplementary Figure S1. Distribution of over- and under-predicted ID residues in the dataset. A histogram is shown for each method. The average number of over-predicted ID residues (FP, upper part, in red) is compared with under-predicted ID residues (FN, lower part, in blue). Both are calculated as the arithmetic mean of the per-protein results.

References

Walsh,I. *et al.* (2015) Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics*, **31**, 201–208.