

# Supporting online material

## TOC for Supporting Online Material

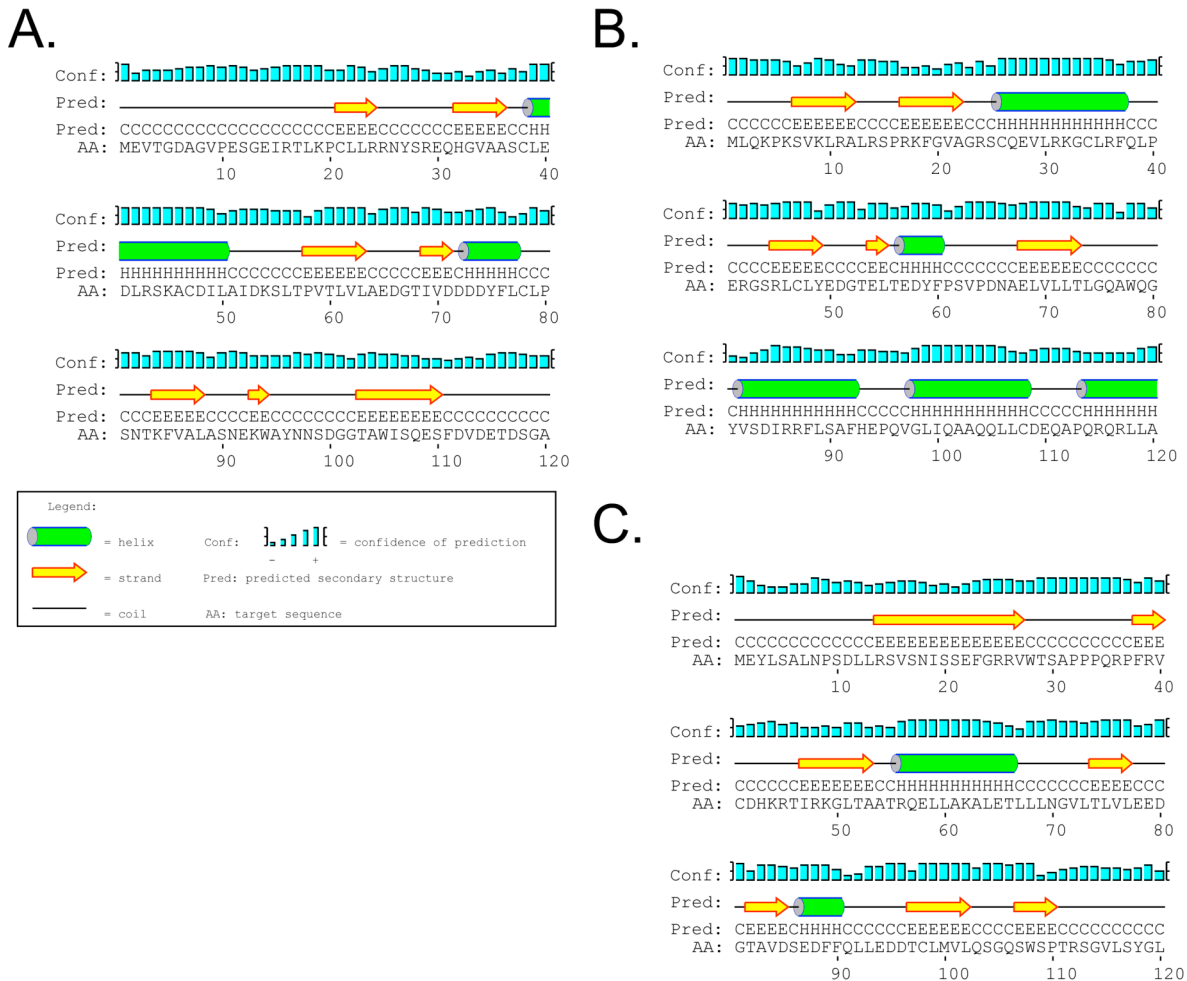
<b>TOC for Supporting Online Material.....</b>	<b>1</b>
<b>Synopsis for Supporting Online Material .....</b>	<b>2</b>
<b>Figures for Supporting Online Material .....</b>	<b>3</b>
Fig. App_1.....	3
Fig. App_2.....	4
Fig. App_3.....	5
<b>Tables .....</b>	<b>7</b>
Table App_1 .....	7
<b>References for Supporting Online Material.....</b>	<b>9</b>

## Synopsis for Supporting Online Material

Many methods that are optimized to predict natively unstructured regions in proteins are trained and tested on residues that are missing from X-ray structures. Although this definition has some advantages (for instance, the definition is more or less unified among different researchers and can be easily interpreted), missing residues do not necessarily represent a biological phenomenon. In fact, it has been shown that residues in these regions are similar in amino acid composition and secondary structure content to flexible structure loops (1). Therefore, these methods cannot always distinguish between loops that are structured and loops that aren't. Here, we show one example in which the secondary structure prediction by PSIPRED (2) (Fig. App\_1) is highly correlated with DISOPRED2 (3) output (Fig. 5A) (DISOPRED2 prediction depends on PSIPRED output among other properties); the locations of the peaks of the prediction are correlated with loops location. NORSnet, however, is optimized to make the distinction between natively unstructured loops and structured loops. Since both NORSnet and DISOPRED2 use predicted secondary structure information (which can reach about 80% accuracy), they can distinguish whether two homologues proteins are 'loopy' or not and overcome the high similarity of the amino acid compositions (Fig. App\_2). Conversely, methods that use only amino acid composition cannot always make that distinction.

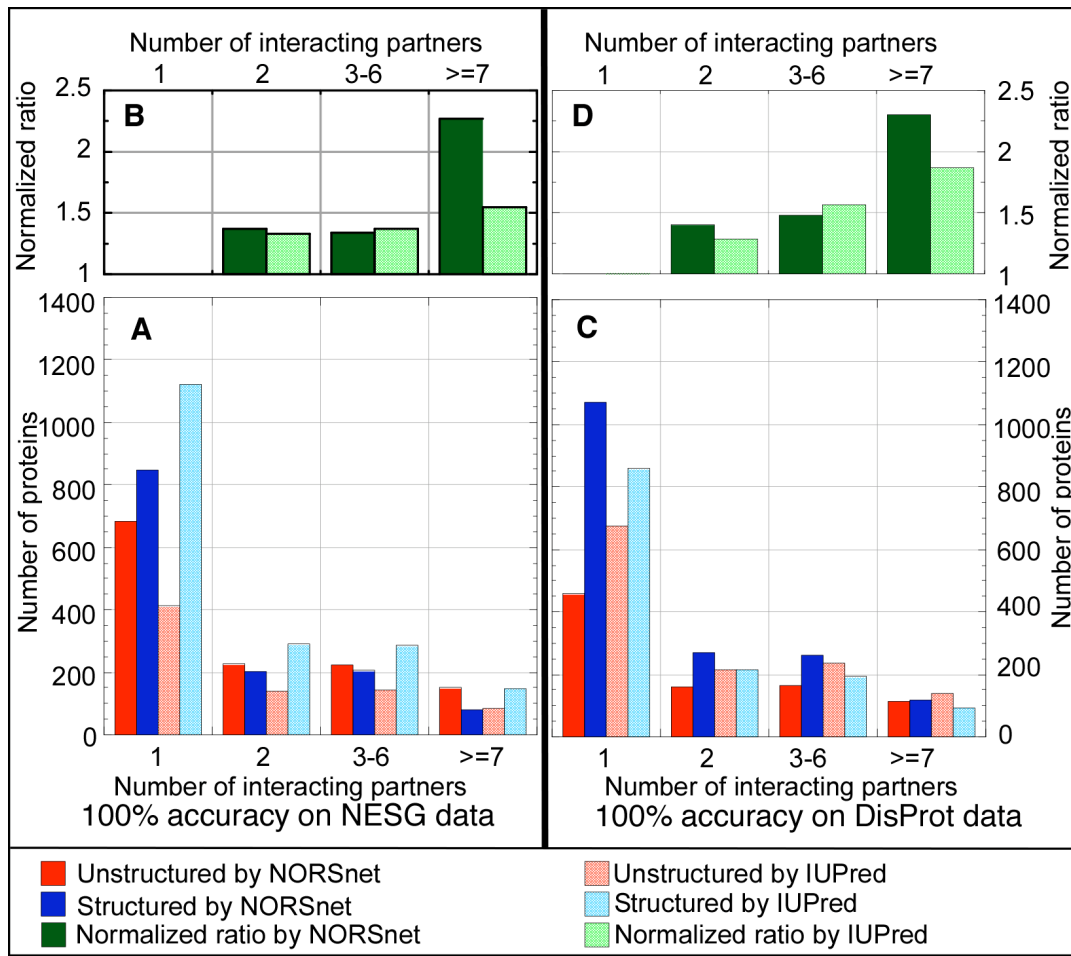


Fig. App\_2



**Fig. App\_2: Secondary structure predictions of the N-termini domains of DFF45, DFF40 and CIDE-B.** Although the N-term domain of DFF45 is unstructured PSIPRED predicts secondary structure elements within that region (A). DFF40 and CIDE-B, homologues of DFF45, however, do not have natively unstructured N-termini (B-C). Interestingly, to an extent this is reflected in the predicted secondary structure prediction of these proteins as DFF45 is predicted more residues to be in coil.

**Fig. App\_3**



**Fig. App\_3: Unstructured regions over-represented in protein-protein hubs of worm.** Similarly to Fig. 7, we ran IUPred on worm proteins that are involved in protein-protein interactions. NORSnet data is identical to the one presented in Fig. 7. The number of proteins that are predicted to be either unstructured or well-structured is plotted against the number of interacting partners for two different thresholds of reliability of the two methods: **A+B** were compiled for thresholds at which both methods maintained 100% accuracy for the NESG data (Fig. 4), while graphs **C+D** were compiled for 100% accuracy on DisProt (Fig. 3). Since the number of observed interaction partners falls off dramatically, we had to group the data into bins of roughly equal sizes (x-axes). **A+C** show the results for the number of proteins predicted in each bin of interaction partners, while **B+D** show the normalized ratios to zoom into the difference between unstructured and structured proteins in each bin. These ratios were compiled as  $\text{Ratio}(\text{bin}) = \frac{\{\#\text{unstructured}(\text{bin})/\#\text{structured}(\text{bin})\}}{\{\#\text{unstructured}(1)/\#\text{structured}(1)\}}$ . As all ratios are above 1, proteins with more than one interaction partners have more

unstructured regions than proteins with one partner. For the thresholds at which both methods achieved 100% accuracy on the DisProt dataset, both IUPred and NORSnet identified unstructured regions in 98 proteins that interact with seven partners or more. IUPred predicted 37 proteins with unstructured regions that NORSnet did not identify and NORSnet predicted 17 proteins with unstructured regions that IUPred had missed.

## Tables

Table App\_1

Number	NESG ID <sup>a</sup>	Sequence length	Disorder signal <sup>b</sup>
1	AR2242	107	Largely
2	BhR21	117	Partly
3	CvR16	205	Partly
4	FR254	163	Largely
5	HR1506	79	Largely
6	HR1538	62	Largely
7	HR1821	157	Partly
8	HR1974	120	Largely
9	HR2078	170	Largely
10	HR2130	173	Largely
11	HR224	87	Largely
12	HR2299	113	Largely
13	HR36	115	Partly
14	HR8	76	Largely
15	HR919	208	Largely
16	HR922	154	Largely
17	HR997	189	Largely
18	KR12	231	Largely
19	LmR11	103	Partly
20	MaR51	125	Partly
21	MhR22	75	Largely

22	MhR41	206	Partly
23	MrR47	128	Partly
24	PsR51	76	Largely
25	SR128	193	Partly
26	SmR3	62	Largely
27	SpR5	62	Largely
28	WR46	193	Partly
29	XR5	50	Largely
30	YR8	155	Largely

**Table S1: Dataset of unstructured proteins from NorthEast Structural Genomics Consortium**

- a NESG id referred to identifiers given by the NESG consortium.
- b Disorder signal referred to different levels of signal of a protein to be unstructured from NMR experiments. *Largely* marked largely unstructured proteins, e.g., (i) their HSQC has high signal to noise and very low dispersion and (ii) their HetNOE data is clear negative; *partly* marked partly unstructured proteins, which have some local structure but overall obey the same criteria; 20 proteins were identified as largely unstructured and 10 proteins were identified as partly unstructured.

## References for Supporting Online Material

1. Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D. and Dunker, A.K. (2004) Protein flexibility and intrinsic disorder. *Protein Science*, **13**, 71-80.
2. McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404-405.
3. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology*, **337**, 635-645.