Methods that are optimized to predict natively unstructured regions in proteins use different methodologies. While many of these methods performances are comparable, they are rather complementary to each other and identify different proteins to have unstructured regions. Here, we show an example in which two methods that use statistical potentials (IUPred [3] and FoldIndex [4]) accurately identify a protein with unstructured region that undergo disorder-order transition coupled with DNA binding. In this example the interface is large and contains many charged residues, thus, the signal of disorder from the amino acid composition can be captured (Fig. App_1). Conversely, methods that are trained on missing residues from Xray structures and use predicted secondary structure as an input can be fooled by this information because the secondary structure elements are predicted as they appear in the structured complex form (differently from the unstructured monomer form). Here, we show an example for such case; the
secondary structure prediction by PSIPRED [5] is highly correlated with DISOPRED2 [6] output (Fig. App_2) (DISOPRED2 prediction depends on PSIPRED output among other properties); note that the secondary structure prediction by PSIPRED is similar to the secondary structure elements of MAX protein as they appear in the complex structure (Fig 7). While it has been shown that the interfaces of unstructured regions that are involved in protein-protein interactions are rather large and are abundant with charged and polar residues, the potentials based method can be accurate in identifying this type of unstructured regions as well. However, in some cases, the interfaces of the unstructured regions with their partners are hydrophobic. Due to the fact that most statistical potential based methods associate hydrophobicity with stability they may misclassify such proteins as structured (example in Fig. App_3 and Fig. 8).

As an example, we compared the predicted contacts for the cold climate plant protein BN28a (DisProt identifier DP00216), which was found to be unstructured by both far-UV CD and NMR experiments [1], and the well-structured L23 ribosomal protein (PDB id 1jj2_R (Klein, et al., 2001)). “White spaces” (Fig. App_2A, top left triangle, Supporting Online Material) dominated the BN28a map, indicating low predicted contact probabilities for this protein (Fig. App_2B, red curve, Supporting Online Material), whereas contacts for the well-structured L23 ribosomal protein (Fig. App_2A, lower right triangle, Supporting Online Material) were enriched in high probability pairwise interactions (Fig. App_2B, blue curve). Darker areas in the L23 map might correspond to actual interactions between secondary structure segments, as is the case for the region in the blue box in Fig. App_2A, where PROFcon correctly identified interactions between beta strands 3 (residues 49-56 as defined by DSSP [2] secondary structure assignments) and 4 (residues 62-68).
Figures for Supporting Online Material

Fig. App_1

Fig. App_1 (was fig. 2): Different facets of unstructured regions. There are several approaches to the definition of unstructured regions. (A) Residues that are missing in electron density maps (green arrow) may appear unstructured in one protein structure (PDB chain 1TC1_A (Focia, et al., 1998)) and ordered in another structure of the same protein (PDB chain 1TC1_B (Focia, et al., 1998)). (B) Natively unstructured regions are usually long (>30 residues) and may have helices and strands. They may become structured upon complex formation and usually have a functional role. (C) In contrast to unstructured regions, residues with high B-values (in red) appear on the electron density map but have higher thermal energy. For instance, the switch II region in RAS protein (PDB identifier 5P21 (Pai, et al., 1990)). (D) An extreme variant of loopy regions are NORS regions that are depleted of helices and strands for stretches extending over 70 residues and more. When such regions form regular 3D structures they usually remain abundant in loops (PDB identifier 1KI0 (Abad, et al., 2002)).
Fig. App_2

Fig. App_2 (was fig. 3): PROFcon captures low contact propensity of unstructured regions. PROFcon predictions can be translated into 2D residue-residue interaction maps. (A) The 2D-map of the unstructured cold climate plant protein BN28a (in red, DisProt entry DP00216; upper left triangle) is significantly less dark than the map of the structured L23 ribosomal protein (in blue, PDB code 1jj2_R; lower right triangle). Each point on the 2D-contact map represents a predicted interaction between two residues; the darkness of the point is correlated with the likelihood of this interaction to occur. (B) The corresponding two distributions of the values predicted by PROFcon appear very different. PROFcon reliably predicts the protein of regular structure to have many more pairwise interactions than the protein with unstructured regions.
Fig. App_3: Distribution of $e_i$ values (was fig. 4). The predicted $e_i$ values (Eqn. 3) are given for three different data sets. The sets correspond to residues from well-ordered proteins (blue), proteins with unstructured regions (red) and the subset of the latter that contains only the unstructured residues (according to DisProt database).
Fig. App_4

A.

B.

Fig. App_4: IUPred and FoldIndex predictions for MAX protein. The output of the disorder predictors IUPred and FoldIndex are presented. Both methods use the amino acid composition in a given window to calculate the tendency of a residue to be in unstructured region. IUPred predicted all residues to have high probability of disorder (A) and FoldIndex calculated all residues to be in unstructured region (B).
Fig. App_2: DISOPRED2 and PSIPRED predictions for MAX protein. The output of the disorder predictor DISOPRED2 and the secondary structure prediction method PSIPRED are presented. DISOPRED2 utilizes secondary structure predictions information by PSIPRED to predict unstructured regions. (A) DISOPRED2 accurately identifies the termini of the protein (residues 1-21 and 107-160) to be unstructured. (B) This output is in agreement with PSIPRED secondary structure prediction that identifies the termini regions to be mostly in the coil state. The bHLH-LZ DNA binding domain of this protein (residues 22-106) is correctly predicted to be mostly helical leading DISOPRED2 to give a string signal of order for that region.
**Fig. App_6**

**Fig. App_3: IUPred prediction for Myosin heavy chain.** The output of the disorder predictors IUPred for myosin heavy chain is presented. The lever arm (residues 765-832) has a very strong signal of order that is resulted from hydrophobic residues in that region.
References for Supporting Online Material


