Misprediction of structural disorder in extremophiles

Rita Pancsa[1], Denes Kovacs[1], Pallab Bhowmick[1], Peter Tompa[1,2]

[1]VIB Department of Structural Biology, Brussels, Belgium
[2]Institute of Enzymology, Hungarian Academy of Sciences, Budapest, Hungary

Introduction: Whereas the concept of intrinsic disorder derives from biophysical observations of the lack of structure of proteins under native conditions, many of our respective concepts rest on genome scale bioinformatics predictions based on amino acid sequence. It is established that most predictors work reliably on proteins commonly encountered, but it is often neglected that we know very little about their performance on proteins derived from microorganism that thrive in environments of extreme temperature, pH or salt concentration. To address the accuracy of disorder prediction in these extremophiles, we collected such proteins from the Protein Data Bank[1] (PDB) and predicted their level of structural disorder by various algorithms; similar calculations were also performed on complete proteomes. Looking at the species composition of the DisProt Database[2] – which usually serves as an essential training set for disorder prediction methods – also helped to clarify our view on this problem.

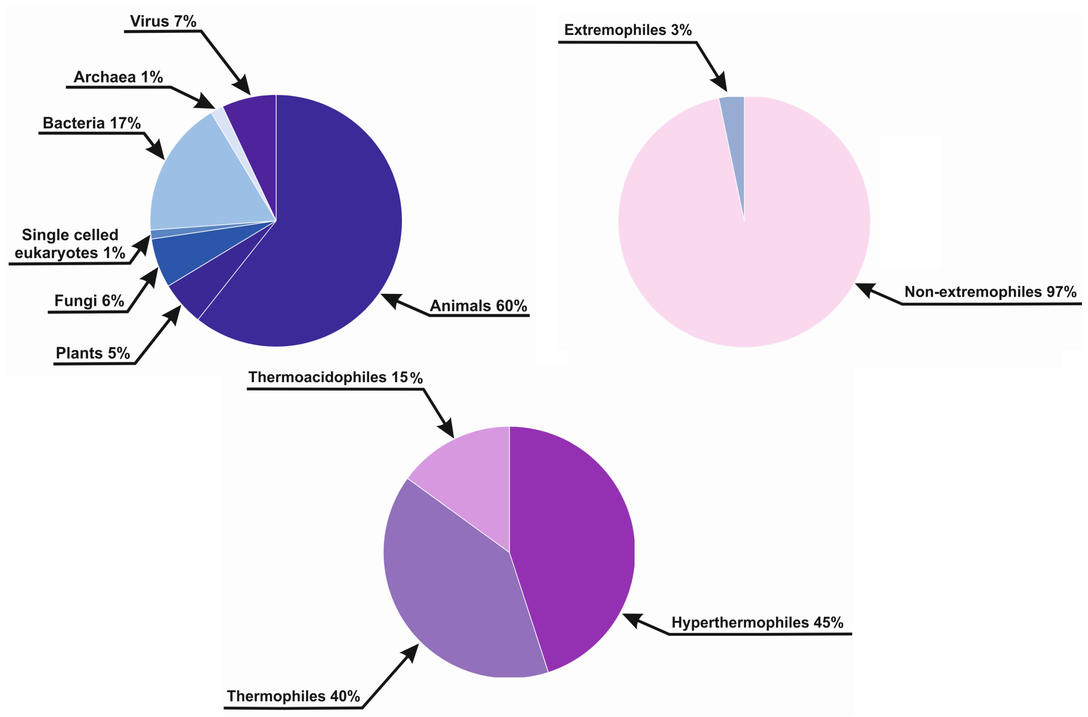

Figure 1: Species distribution of the sequences in the DisProt database.
The majority of sequences stored in the DisProt database belong to multicellular organisms, particularly animals. Only 3% of them came from extremophile organisms, mainly from hyperthermophiles and thermophiles, and some of the extremophile groups, like halophiles, alkaliphiles, psychrophiles or radiotolerants are not represented at all.

Results: We found evidence supporting the idea of protein disorder misprediction in case of extremophile organisms. The species distribution of the DisProt database (Figure 1) shows very low ratio of extremophile proteins, many of the major groups not being represented at all (like halophiles, alkaliphiles, psychrophiles and radiotolerants). However, most of the existing disorder prediction methods were trained to distinguish between the properties of proteins in this database and those in the PDB.
When extracting bacterial and archaeal sequences from the PDB database and predicting their structural disorder with various methods (methods based on different approaches showed very similar tendencies) we found huge differences between the disorder content of proteins in some extremohpile groups and the reference mesophile set (Figure 2). The group of halophiles and radiotolerants showed significantly higher disorder content than that of mesophiles, while other groups (acidophiles, hyperthermophiles, thermophiles and psychrophiles) were found to be significantly less disordered.
The mean disorder contents for similarly grouped prokaryotic complete proteomes reflected similar tendencies (Figure 3) with the exception of the acidophile group.
Finally, we found structures for such highly similar enzyme pairs in the PDB, where one of the proteins belongs to an exremophile and the other to a mesophile organism. These also clearly showed that despite the almost identical structural properties, disorder prediction methods give surprisingly different results because of the adaptive changes affecting extremophile sequences (Figure 4).

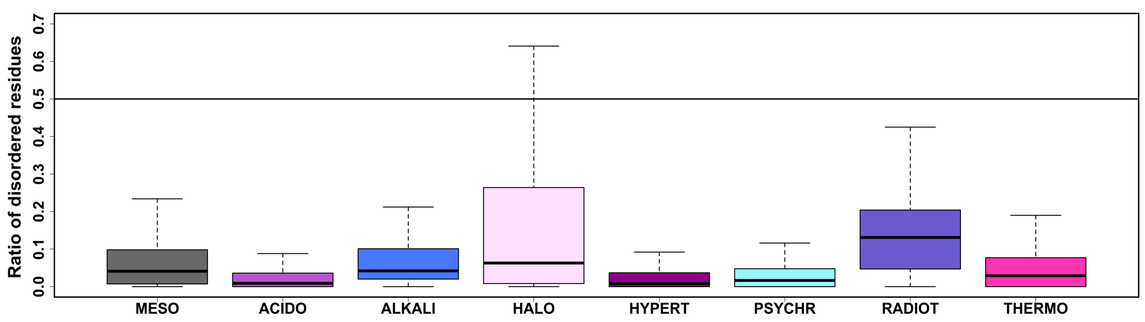

Figure 2: Structural disorder of PDB sequences for different exremophile groups.
All the bacterial and archaeal entries were downloaded from the PDB, in which only proteins are present. Sequence identity filter of 90% and a sequence length filter of at least 31 residues was applied. The sequences were sorted to a reference mesophile group and to different extremophile groups according to the information found for the species. The ratio of disordered residues was calculated for every protein based on the IUPred method[3] and these were plotted for every group. Abbreviations stand for: MESO – mesophiles, ACIDO – acidophiles, ALKALI – alkaliphiles, HALO – halophiles, HYPERT – hyperthermophiles, PSYCHR – psychrophiles, RADIOT – rediaotolerants and THERMO – thermophiles.

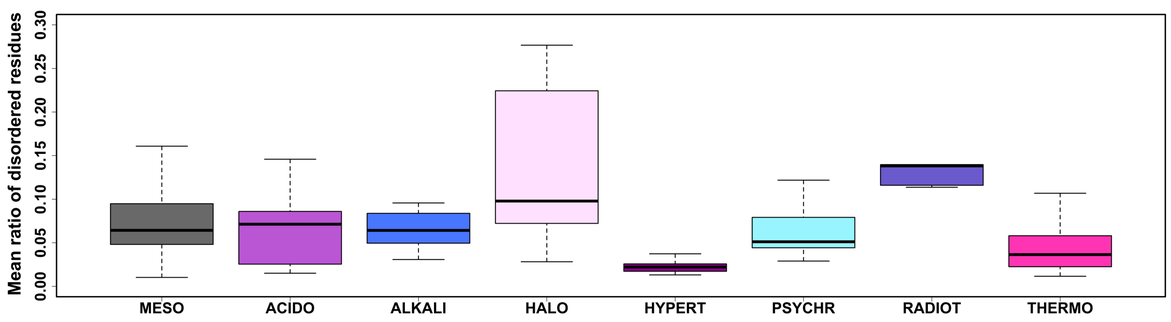

Figure 3: Mean disorder content of complete proteomes for different exremophile groups.
We collected all the prokaryotic representative proteomes with 75% of co-membership treshold from the PIR databese[4]. The species were grouped the same way as in case of PDB sequences. For all proteins in the 898 bacterial and 96 archaeal proteomes the ratio of disordered residues (disorder content) was calculated based on the predictions of the IUPred method. The mean value was gained for every proteome, and plotted for the different groups. The abbreviations are described by Figure 2.

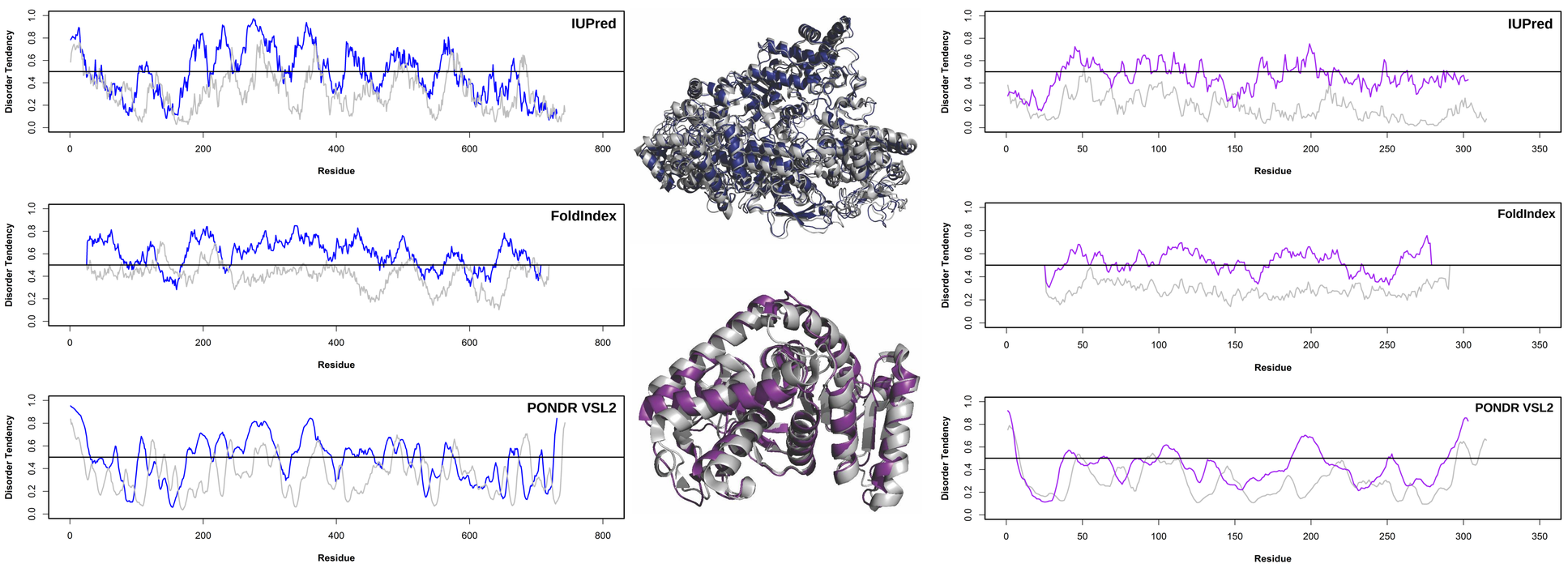

Figure 4: Examples for highly similar enzyme pairs coming from different lifestyle groups with the comparison of their predicted disorder pattern.
The upper structural alignment compares two catalase-peroxidase enzymes, a mesophile (grey; PDB: 1SJ2) and a halophile (dark blue; PDB: 1ITK). The comparison of their disorder prediction results are shown on the left with three different methods (IUPred, FoldIndex[5], PONDR VSL2[6]) using the same color code. On the bottom, the structural alignment of two malate dehydrogenase enzymes is presented, a mesophile (grey; PDB: 3TL2) and a halophile (purple; PDB: 1D3A). Predictions with the same three methods can be found on the right.

Conclusion: We found systematic overprediction of disorder in the case of bacteria living at very high salt concentrations (halophiles) and those surviving gamma radiation (radiotolerants). Underprediction is more difficult to prove, but it seems to be the case of bacteria living at extremely low- (psychrophiles) and extremely high- (hyperthermophiles and thermophiles) temperatures. Acidophiles and alkaliphiles did not show big differences from mesophiles on the whole proteome level, which is probably due to their general ability of using H^+ pumps to keep their cytosol at near neutral pH. This mechanism protects their proteome from the pH extremities outside the cells. Although sparseness of data on ordered – and particularly on disordered – proteins from extremophiles precludes the development of dedicated predictors at present, we do suggest to use those disorder prediction methods, which take into consideration evolutionary relationships when investigating structural disorder in extremophiles.

Acknoledgements:
This work was supported by the Research Foundation Flanders (FWO) Odysseus grant G.0029.12, and the Marie Curie Initial Training Network project 264257 (IDPbyNMR) from the European Commission.

References:
1. Berman HM et al. (2000) Nucleic Acids Res 28: 235-242.
2. Sickmeier M et al. (2007) Nucleic Acids Res 35:D786-93.
3. Dosztanyi Z et al. (2005) J Mol Biol 347: 827-839.
4. Chen_ C et al. (2011) PLoS One 6(4):e18910.
5. Prilusky J et al. (2005) Bioinformatics 21(16):3435-8.
6. Obradovic Z et al. (2005) Proteins 61(S7):176-182.