molecules

MDPI

Article

Misprediction of Structural Disorder in Halophiles

Rita Pancsa [1,*], Denes Kovacs [2] and Peter Tompa [1,2,3,*]

[1] Institute of Enzymology, Research Centre for Natural Sciences of the Hungarian Academy of Sciences, 1117 Budapest, Hungary
[2] VIB Center for Structural Biology (CSB), 1050 Brussels, Belgium; dkovacs@vub.ac.be
[3] Structural Biology Brussels (SBB), Vrije Universiteit Brussel (VUB), 1050 Brussels, Belgium
* Correspondences: pancsa.rita@ttk.mta.hu (R.P.); peter.tompa@vub.be (P.T.); Tel.: +36-1-382-6705 (R.P.); +32-2-629-1924 (P.T.)

Academic Editor: Quan Zou
Received: 15 January 2019; Accepted: 26 January 2019; Published: 29 January 2019

Abstract: Whereas the concept of intrinsic disorder derives from biophysical observations of the lack of structure of proteins or protein regions under native conditions, many of our respective concepts rest on proteome-scale bioinformatics predictions. 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 of microorganisms that thrive in environments of extreme temperature, pH, or salt concentration, which may cause adaptive sequence composition bias. To address this issue, we predicted structural disorder for the complete proteomes of different extremophile groups by popular prediction methods and compared them to those of the reference mesophilic group. While significant deviations from mesophiles could be explained by a lack or gain of disordered regions in hyperthermophiles and radiotolerants, respectively, we found systematic overprediction in the case of halophiles. Additionally, examples were collected from the Protein Data Bank (PDB) to demonstrate misprediction and to help understand the underlying biophysical principles, i.e., halophilic proteins maintain a highly acidic and hydrophilic surface to avoid aggregation in high salt conditions. Although sparseness of data on disordered proteins from extremophiles precludes the development of dedicated general predictors, we do formulate recommendations for how to address their disorder with current bioinformatics tools.

Keywords: extremophile; halophile; structurally disordered; intrinsically disordered; disorder prediction; sequence composition bias; adaptation; charge distribution

1. Introduction

The recent recognition of intrinsically disordered proteins and regions (IDPs/IDRs) that lack a well-defined structure under physiological conditions [1–4] revolutionized structural biology research. Such proteins perform their functions as conformational ensembles and are prevalent in most organisms, with eukaryotes typically having more than prokaryotes [5,6]. Based on diverse biophysical techniques, we have conclusive evidence for the structural disorder of 1539 protein regions within 694 proteins, which are deposited in the DisProt database (release 6.02) [7]. This database provides accurate residue-level assessment of the order/disorder status of proteins and served as a training set for several dedicated predictors [8,9].

Since experimental data on disorder in DisProt cover only a tiny fraction of (even the best covered human) proteome, many of our basic concepts and generalizations of protein disorder rely on bioinformatics predictions, including its phylogenetic distribution and evolution [5,6]. However, it has never been directly addressed how much we can extrapolate our conclusions to species outside those covered in DisProt. That is, a potentially serious and thus far unappreciated limitation of applying predictors is that they have all been trained on disordered proteins from organisms of a rather limited phylogenetic circle and range of habitats. Due to these reasons, any type of evolutionary adaptation causing systematic compositional bias in protein sequences could potentially mislead the prediction methods. For instance, proteins of (micro)organisms living under extreme environmental conditions (extremophiles) are under severe challenge not to lose their structural integrity [10–18]. Extremophile sequences are practically absent from DisProt (Figure 1), which raises the question of whether they can be accurately handled by disorder prediction algorithms, i.e., how much we should trust such prediction results.

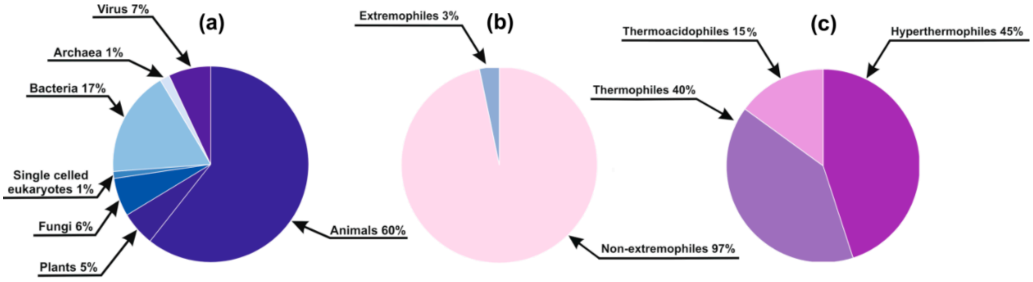

Figure 1. Species distribution of the DisProt Database: (**a**) Distribution of DisProt entries among large phylogenetic groups; (**b**) Distribution of the represented species among extremophiles and non-extremophiles; (**c**) Distribution of the 20 extremophilic proteins among different lifestyle extremities.

Although several prior observations caution that there might be systematic prediction errors due to these reasons [19–22], those have never been directly addressed. For example, when structural disorder was addressed in Archaea and in extremophiles, very high levels of disorder were reported for halophiles and alkaliphiles [20,22], and interpreted as a sign of increased need for complex regulation in them [22]. However, from the point of view of evolutionary adaptation, the saline lakes where most halophiles live represent relatively static environmental conditions coupled with a low number of possible competitors, which do not imply the increased need for complex regulatory mechanisms and consequently structural disorder. Additionally, Archaea are consistently predicted to have less disorder than Bacteria, but it has never been noted that this could well be due to their significant bias for hyperthermophilic species [16] and not due to their phylogenetic identity *per se*. Although biased amino acid composition is the primary reason for structural disorder [23–25], and proteins of extremophiles do possess unusual amino acid composition [10,26,27], these two are based on different underlying reasons and therefore should be carefully separated and their relationship should be precisely analyzed before interpreting any prediction results gained for such sequences.

Here we investigate this phenomenon by applying disorder predictors to the proteomes and individual proteins of extremophiles and comparing their predicted disorder to those of mesophiles. We report systematic prediction errors in a particular extremophile group, explain how the observed tendencies make biological sense, and formulate some general recommendations to arrive at a more realistic picture when using the available methods on their sequences.

2. Results

2.1. Species Distribution of the DisProt Database

In the DisProt Database (release 6.02) [7], which served as a training set for most disorder prediction methods, only 20 sequences (<3% of the database) belong to species living among extreme conditions (10 Archaea and 10 Bacteria; nine hyperthermophiles and eight thermophiles, with three of them also displaying acidophilic characteristics (thermoacidophiles); Figure 1). Thus, we can conclude that extremophile groups are hardly or not at all represented in DisProt, which precluded the training of any methods that could distinguish them from other DisProt sequences.

2.2. Assessing Structural Disorder in Extremophile Groups

We calculated the mean fractions of disordered residues for the proteomes belonging to each extremophile lifestyle group and compared their distributions to that of the mesophilic reference group for Bacteria and Archaea, separately, with three prediction methods (see further details on these predictors in Materials and Methods) (Figure 2). We found similar tendencies as Vicedo et al. [20]. Average protein length was also compared between the different groups and mesophiles (as reference); moreover, the comparisons were repeated using only protein subsets with annotated Enzyme Commission (EC) numbers (by UniProt [28]; Figure S1), hereafter referred to as enzymes, to see if the observed deviations were due to the addition or loss of disordered proteins/regions or if they were rather due to an adaptive change generally affecting the compositions of all proteins (including enzymes), without necessarily affecting their structural status. All calculated data are presented in Tables S1 and S2 for Bacteria and Archaea, respectively.

The three prediction methods provided similar tendencies among the lifestyle groups, even though they showed remarkable differences in the absolute values of predicted disorder. Additionally, full proteomes and enzymes showed very similar tendencies with each method, with enzymes having lower predicted disorder in absolute terms (Figure 2 and Figure S1).

Acidophilic and alkaliphilic bacteria did not show significant disorder deviations from mesophiles, which is in line with the fact that they maintain circumneutral intracellular pH through diverse pH homeostasis mechanisms [29,30], and therefore did not need to extensively adapt the compositions of their intracellular proteins to their environment.

Thermophiles and hyperthermophiles had significantly lower predicted disorder with almost all prediction methods. Among Archaea there are many hyperthermophiles but few thermophiles, so for the former a statistically stronger comparison could be made, while for the latter the comparison was too weak to show significant difference. Among Bacteria, both the many thermophiles and few hyperthermophiles showed significantly lower disorder content than mesophiles with two of the three methods. Therefore, if there are sufficient numbers of proteomes to compare, both thermophiles and hyperthermophiles show a significant reduction in disorder content, with hyperthermophiles having more extreme values. Actually, the physical basis of protein stability in (hyper)thermophiles has been analyzed in great detail in the case of folded proteins, leading to the general picture that extended H-bond and electrostatic networks on the surface, better atomic packing, and reduced surface dynamics all contribute to increased thermal stability [29,30]. Whereas this effect can be rationalized and predicted, evolutionary strategies against other destabilizing physical and chemical factors have been studied in much less detail and are thus poorly understood. The involvement of structural disorder in heat resistance has been addressed in one study only [31], leaving the question, if predicted reduced disorder is also a structural fact, open. Other extreme conditions have not yet been generally approached by either experiment or bioinformatics, i.e., our study is the first to address this question in a systematic manner.

Whether thermophilic and hyperthermophilic proteins are affected by a misprediction of disorder is very difficult to answer. The reason for this is that they hardly have any disordered regions whose predictions can be fairly compared to the corresponding mesophilic homologs because all of them have been lost or at least heavily shortened through evolution, probably for the sake of increasing thermal stability [31]. We have selected the five most closely related mesophile–hyperthermophile species pairs from our dataset (four bacterial and one archaeal, see Table 1), downloaded their 1:1 orthologous protein pairs from the OMA Orthology database, searched for all the $\geq$40 residues IUPred-predicted consecutive long disordered regions (LDRs) in the mesophilic species, aligned the orthologous pairs by ClustalW 2.0.12, identified the corresponding regions within the hyperthermophile orthologs using the alignments, and finally calculated how big percent of the length of the mesophilic LDRs have been preserved in them (Table 1). We found that on average only half of the identified LDRs preserved at least 75% of their length in the hyperthermophiles, with many of the well-preserved regions belonging to ribosomal proteins or matching the regions that we have previously identified as the most essential

LDRs of prokaryotic housekeeping proteins, which are so strongly preserved in evolution that even minimal genomes (endosymbiotic bacteria that have undergone extreme genome reduction) have them [32].

Table 1. Orthologous IUPred-predicted consecutive long disordered regions (LDRs) between mesophilic and hyperthermophilic species.

Mesophile Species	Hyperthermophile Species	Number of 1:1 Protein Orthologs	Number of LDRs (≥40 res.) in Mesophile	Number of LDRs with ≥75% of Length Preserved (of which Ribosomal)
Clostridium perfringens (ATCC 13124)	*Caldicellulosiruptor saccharolyticus (ATCC 43494)*	717	15	6 (3)
Clostridium perfringens (ATCC 13124)	*Thermaerobacter marianensis (ATCC 700841)*	539	9	7 (3)
Clostridium perfringens (ATCC 13124)	*Thermoanaerobacter tengcongensis (DSM 15242)*	782	15	7 (3)
Desulfitobacterium hafniense (DSM 10664)	*Carboxydothermus hydrogenoformans (DSM 6008)*	1109	36	15 (3)
Methanococcus maripaludis (S2)	*Methanocaldococcus jannaschii (DSM 2661)*	1078	8	6 (4)

We checked the presence of the previously independently obtained non-ribosomal obligate LDRs in some well-known hyperthermophilic organisms and compared them to reference mesophilic organisms. Such examples include the *C*-terminal tails of GroEL, DnaK, and the single-stranded DNA-binding protein, SSB. It turns out that in hyperthermophiles even these absolutely essential IDRs have been largely shortened, i.e., when aligning these segments with their mesophilic homologs, the hyperthermophilic sequences are heavily gapped (Figure 3), even though the folded fractions of the same proteins are well-preserved. These examples illustrate well that the very low predicted disorder content of these extremophile proteomes is definitely mainly due to the loss of IDRs. Since hyperthermophiles live at not only extreme, but very static habitats with hardly any competition, their gene sets have undergone a severe reduction to only 2000 ± 500 protein-coding genes in both Bacteria and Archaea, which is significantly less than those of mesophilic species (Bacteria $p = 0.0222$; Archaea $p = 0.0160$). This process went along a radical erosion of their remaining proteins (Figure 2 and Figure S1), i.e., the average protein length in hyperthermophilic Archaea is significantly shorter than in mesophiles ($p < 0.0001$; Figure 2). This erosion mostly affected their IDRs, which were probably thermodynamically unfavorable and provided complex regulatory possibilities, which may be superfluous in such static environments.

Psychrophile bacteria did not show significant disorder deviations from mesophiles. This might be due to the fact that the few organisms belonging here represent a mixture of real psychrophiles and psychrotolerants, which can grow in cold environments but do not require them.

There are two groups for, which we observed significantly higher predicted disorder content than for mesophiles: radiotolerants (among Bacteria) and halophiles (among both Bacteria and Archaea). The group of radiotolerants was largely made up by the *Deinococcus* genus, which is an ancient group with no clear relationship to any of the other known bacterial lineages [29]. *Deinococcus* species have several unique protein families as well as orphans and *Deinococcus*-specific proteins of unknown function [30,31]. Their extraordinary resistance to ionizing radiation, oxidative stress, desiccation, and other damaging conditions has never been attributed to an amino acid composition bias of their proteins, but rather to the action of specific DNA repair [29] and antioxidation systems, and transcription regulation during stress response [30].

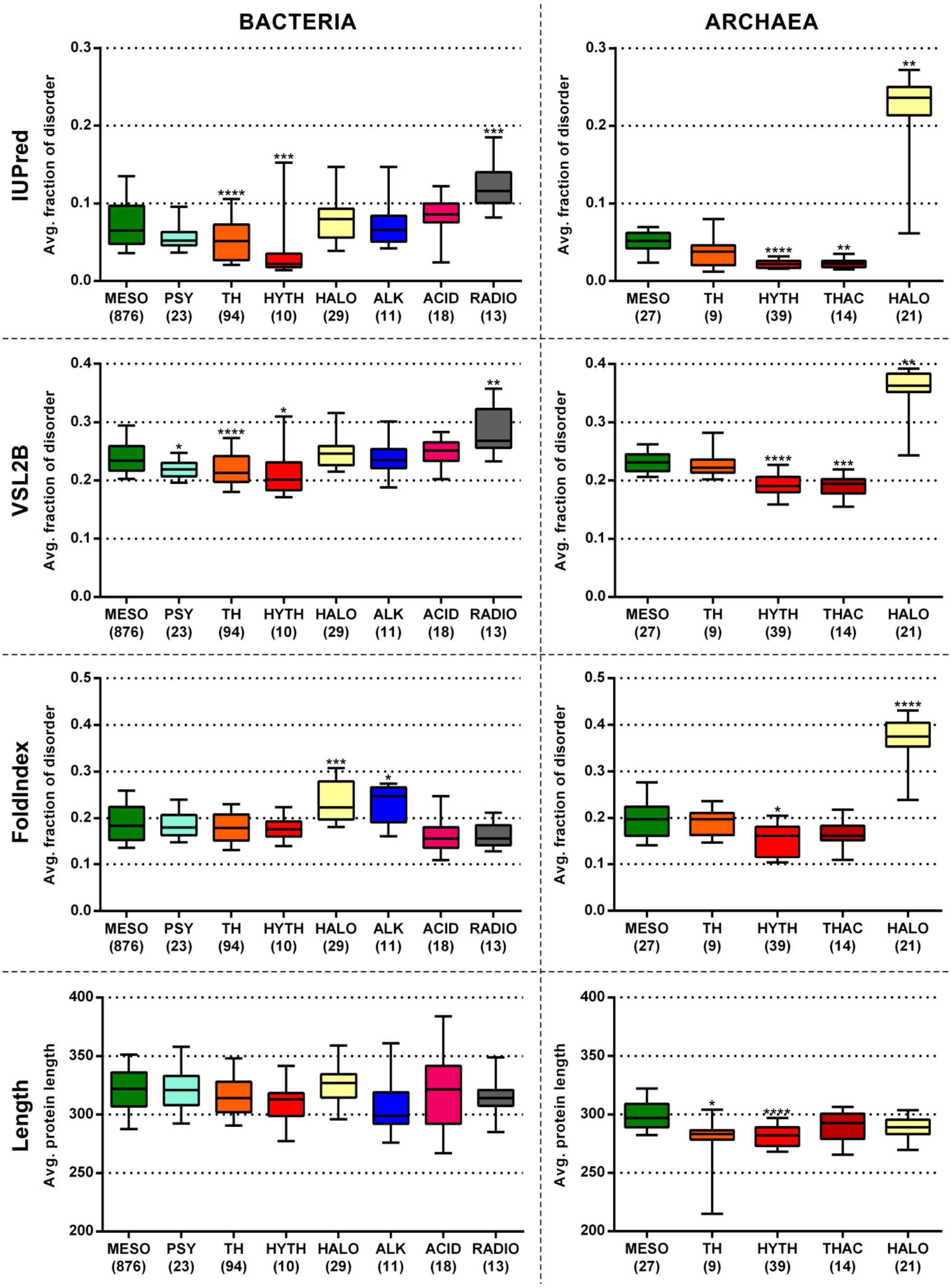

Figure 2. Predicted disorder content of Bacteria and Archaea with different lifestyles. The average fraction of disordered residues was calculated for each proteome by three prediction methods. The distribution of the calculated values of each lifestyle group was compared to that of mesophiles by Dunn's multiple comparison test for Bacteria and Archaea separately. The significance levels of the distribution differences are indicated as stars above the boxes (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$). The number of proteomes in the different lifestyle groups are indicated below the labels. MESO—mesophiles, PSY—psychrophiles, TH—thermophiles, HYTH—hyperthermophiles, HALO—halophiles, ALK—alkaliphiles, ACID—acidophiles, RADIO—radiotolerants, THAC—thermoacidophiles.

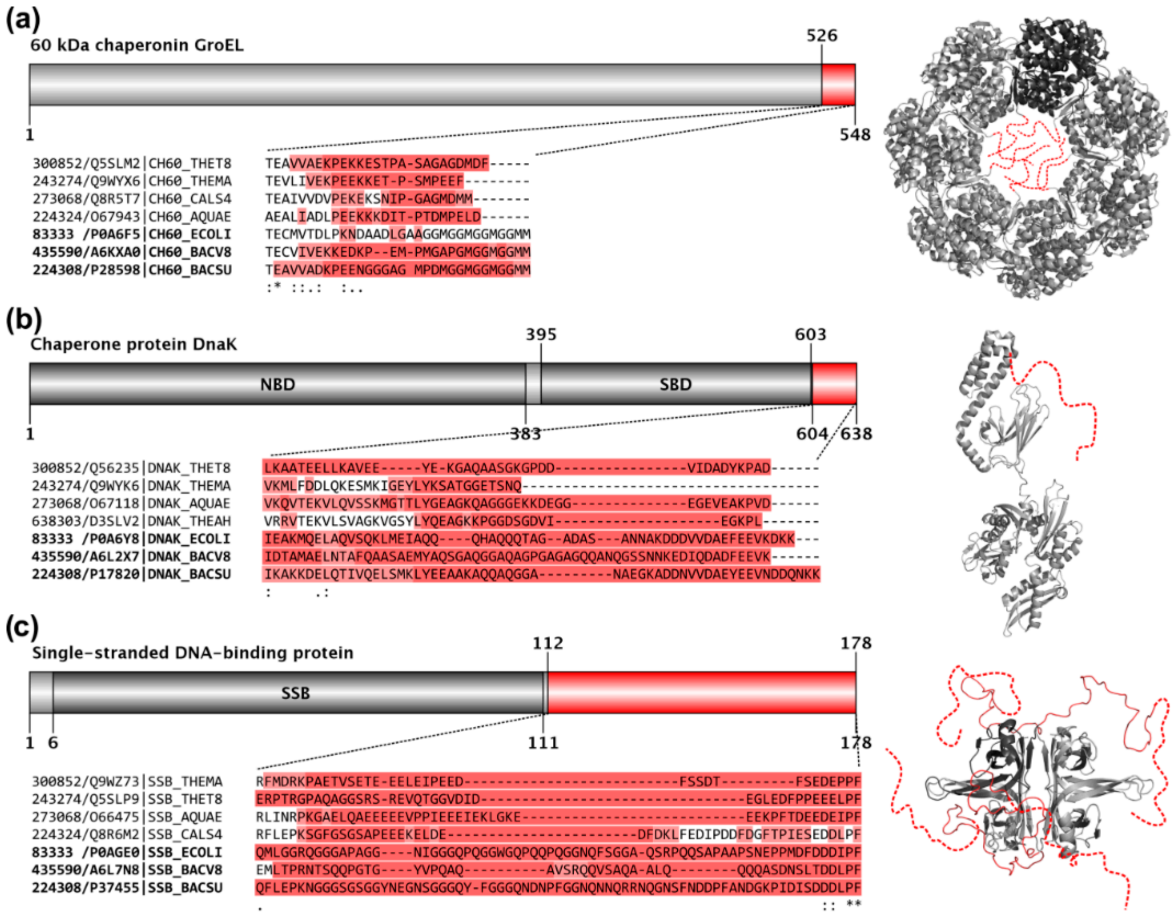
(a)
60 kDa chaperonin GroEL
526
1
548
300852/Q5SLM2|CH60_THET8 TEAVVAEKPEKKESTPA-SAGAGDMDF-----
243274/Q9WYX6|CH60_THEMA TEVLIVEKPEEKKET-P-SMPEEF--------
273068/Q8R5T7|CH60_CALS4 TEAIVVDVPEKEKSNIP-GAGMDMM-------
224324/O67943|CH60_AQUAE AEALIADLPEEKKKDIT-PTDMPELD------
83333 /P0A6F5|CH60_ECOLI TECMVTDLPKNDAADLGAAGGMGGMGGMGGMM
435590/A6KXA0|CH60_BACV8 TECVIVEKKEDKP--EM-PMGAPGMGGMGGMM
224308/P28598|CH60_BACSU TEAVVADKPEENGGGAG MPDMGGMGGMGGMM
:* ::.: :..
(b)
Chaperone protein DnaK
395
603
NBD
SBD
1
383
604
638
300852/Q56235|DNAK_THET8 LKAATEELLKAVEE-----YE-KGAQAASGKGPDD---------------VIDADYKPAD------
243274/Q9WYK6|DNAK_THEMA VKMLFDDLQKESMKIGEYLYKSATGGETSNQ-----------------------------------
273068/O67118|DNAK_AQUAE VKQVTEKVLQVSSKMGTTLYGEAGKQAGGGEKKDEGG-------------EGEVEAKPVD------
638303/D3SLV2|DNAK_THEAH VRRVTEKVLSVAGKVGSYLYQEAGKKPGGDSGDVI-------------------EGKPL-------
83333 /P0A6Y8|DNAK_ECOLI IEAKMQELAQVSQKLMEIAQQ----QHAQQQTAG--ADAS---ANNAKDDDVVDAEFEEVKDKK---
435590/A6L2X7|DNAK_BACV8 IDTAMAELNTAFQAASAEMYAQSGAQGGAQAGPGAGAGQQANQGSSNNKEDIQDADFEEVK------
224308/P17820|DNAK_BACSU IKAKKDELQTIVQELSMKLYEEAAKAQQAQGGA---------NAEGKADDNVVDAEYEEVNDDQNKK
: .:
(c)
Single-stranded DNA-binding protein
112
178
SSB
1 6
111
178
300852/Q9WZ73|SSB_THEMA RFMDRKPAETVSETE-EELEIPEED-------------------------FSSDT---------FSEDEPPF
243274/Q5SLP9|SSB_THET8 ERPTRGPAQAGGSRS-REVQTGGVDID-------------------------------EGLEDFPPEEELPF
273068/O66475|SSB_AQUAE RLINRPKGAELQAEEEEEVPPIEEEIEKLGKE--------------------------EEKPFTDEEDEIPF
224324/Q8R6M2|SSB_CALS4 RFLEPKSGFGSGSAPEEEKELDE----------------------DFDKLFEDIPDDFDGFTPIESEDDLPF
83333 /P0AGE0|SSB_ECOLI QMLGGRQGGGAPAGG----NIGGGQPQGGWGQPQQPQGGNQFSGGA-QSRPQQSAPAAPSNEPPMDFDDDIPF
435590/A6L7N8|SSB_BACV8 EMLTPRNTSQQPGTG----YVPQAQ----------AVSRQQVSAQA-ALQ--------QQQASDNSLTDDLPF
224308/P37455|SSB_BACSU QFLEPKNGGGSGSGGYNEGNSGGGQY-FGGGQNDNPFGGNQNNQRRNQGNSFNDDPFANDGKPIDISDDDLPF
. :: **

Figure 3. In hyperthermophiles, even essential IDRs are largely shortened. Domain maps, structures, aligned sequences, and IUPred-predicted disordered segments are depicted for three essential proteins, GroEL (**a**), DnaK (**b**), and SSB (**c**), with conserved disordered *C*-terminal regions that likely exist and function as disordered ensembles in all species [32]. On the grey domain maps, the residue boundaries of conserved disordered segments (in red) and known domains (in darker grey) are provided for the *E. coli* K12 protein. The red regions of the domain maps complemented by a few residue positions are highlighted as Clustal Omega 1.2.2. multiple sequence alignments below the domain maps. The sequences of the organisms are identified by their Taxonomy/UniProt identifiers. In each case there are four hyperthermophilic sequences shown on the top and three mesophilic reference sequences bellow (identifiers in bold) as a reference. In the alignments the background of the residues are colored according to the corresponding IUPred predictions; residues with a score >0.5 have darker red background, residues with a score between 0.5 and 0.4 have lighter red background, while residues with a score <0.4 have no background. The *E. coli* structures of the corresponding proteins, GroEL (Protein Data Bank(PDB) ID:2NWC), DnaK (PDB ID:2KHO), and SSB (PDB ID:1QVC), are also depicted in light grey with the conserved disordered segments marked by red or added as red dashed lines (or as the combination of the two). For a better understanding of the correspondences, in both the heptameric GroEL and tetrameric SSB structures, one chain is depicted by darker grey than the others.

Therefore, the high level of predicted disorder content in these proteomes is probably not a result of misprediction, but (1) the fact that many of their proteins contain long intrinsically disordered, often low complexity regions that are not present in non-extremophile homologues [31], and likely contribute to desiccation resistance [32]; (2) the increased presence of orphan proteins in each species [29] (orphan proteins have a high propensity for disorder in general [33,34]); and (3) the increased numbers of DNA repair and transcription regulation proteins that have relatively high levels of structural disorder in general [35,36].

Halophilic Bacteria grow over an extended range of salt concentrations (3–20% NaCl, w/v and above), with many being only moderately halophilic, unlike the obligate halophilic Archaea whose growth is restricted to high saline environments [33]. This could be the reason why archaeal halophiles had uniformly very high disorder content with all three prediction methods (higher than any other lifestyle group), while bacterial halophiles were only found to be significantly more disordered than mesophiles with FoldIndex. To see if there is a relationship between the level of salt requirement/tolerance of bacterial halophiles and their predicted proteome-level disorder content, we classified the respective 29 species into three groups, moderately halophilic, halophilic, and extremely halophilic (see further details on this classification in Materials and Methods and in Table S3), and compared their predicted disorder content with all three prediction methods to that of mesophiles (Figure 4). While moderately halophilic species were not different from any of the three prediction methods, halophilic species were more disordered than mesophiles with FoldIndex and extremely halophilic species with both IUPred and PONDR VSL2B. The largely elevated disorder levels seen for extreme halophiles among both Bacteria and Archaea are due to a completely different reason than in the case of radiotolerants. Since halophiles usually live in static environments, it is very unlikely that they would benefit from maintaining an increased complexity of regulatory mechanisms that would require the excessive involvement of disordered proteins/regions. Halophilic Bacteria and Archaea were also reported to have some large unstructured hydrophilic low complexity protein regions that may provide them with resistance to dehydration [34]. This is in line with the suggested chaperone effect of IDPs, such as late embryogenesis abundant (LEA) proteins and dehydrins [35,36], which appears as a widespread stress-related protective mechanism in diverse organisms such as bacteria [34], plants [37], or animals [38]. Still, this is probably not the primary reason for the high predicted disorder in halophiles; rather, it is due to their unusually high net negative surface charge, i.e., an increase in Asp and Glu residues and a decrease in nonpolar and basic surface residues, which may help them avoid aggregation in high salt conditions [26,27]. The proteomes as well as the IUPred-predicted disordered regions of halophiles have significantly more acidic residues than those of mesophiles for both Bacteria and Archaea (Figure S2). We propose that the high net negative surface charge misleads disorder prediction methods, especially the ones that are purely based on sequence, so that they predict even well-folded halophilic proteins/enzymes to be disordered.

2.3. Halophile–Mesophile Protein Homologs with Available Structures Clearly Support Overprediction of Disorder in Halophiles

To investigate the role of high net negative surface charge of halophilic proteins, we decided to select a few enzymes that have both an extreme halophile and a mesophile version in the Protein Data Bank (PDB) with high sequence- and structure-identity and similar length. Since there are not many of such pairs and it is difficult to perform a targeted search, we show here three examples that we could identify in a relatively short time. These include a pair of malate-dehydrogenase enzymes from *Haloarcula marismortui* (halophile; PDB:1D3A) and *Bacillus anthracis* (mesophile; PDB:3TL2) (Figure 5), a pair of nucleoside diphosphate kinase enzymes from *Halobacterium salinarum* (halophile; PDB:2AZ3) and *Staphylococcus aureus* (mesophile; PDB:3Q86) (Figure S3), and a pair of catalase peroxidase enzymes from *H. marismortui* (halophile; PDB:1ITK) and *Mycobacterium tuberculosis* (mesophile; PDB:1SJ2) (Figure S4). Disorder predictions were carried out on these pairs by a wide variety of prediction methods, namely IUPred [39], FoldIndex [40], RONN [41], PONDR VSL2B [42], PrDOS [43], PONDR VSL2P [42], DisPro [44], Disopred2 [6], Globplot2 [45], and CSpritz [46]. To help understand the biophysical principles underlying the markedly elevated disorder profiles of the halophilic proteins we also showed their surface charge patterns. The marked differences between the surface charge patterns of homologous halophile–mesophile protein pairs have already been previously demonstrated for dihydrofolate reductase [47,48], malate-dehydrogenase [49], 2Fe-2S ferredoxin [50], proliferating cell nuclear antigen (PCNA) [51], manganese superoxide dismutase [52], elongation factor EF-Tu [53], and glutamate dehydrogenase [54].

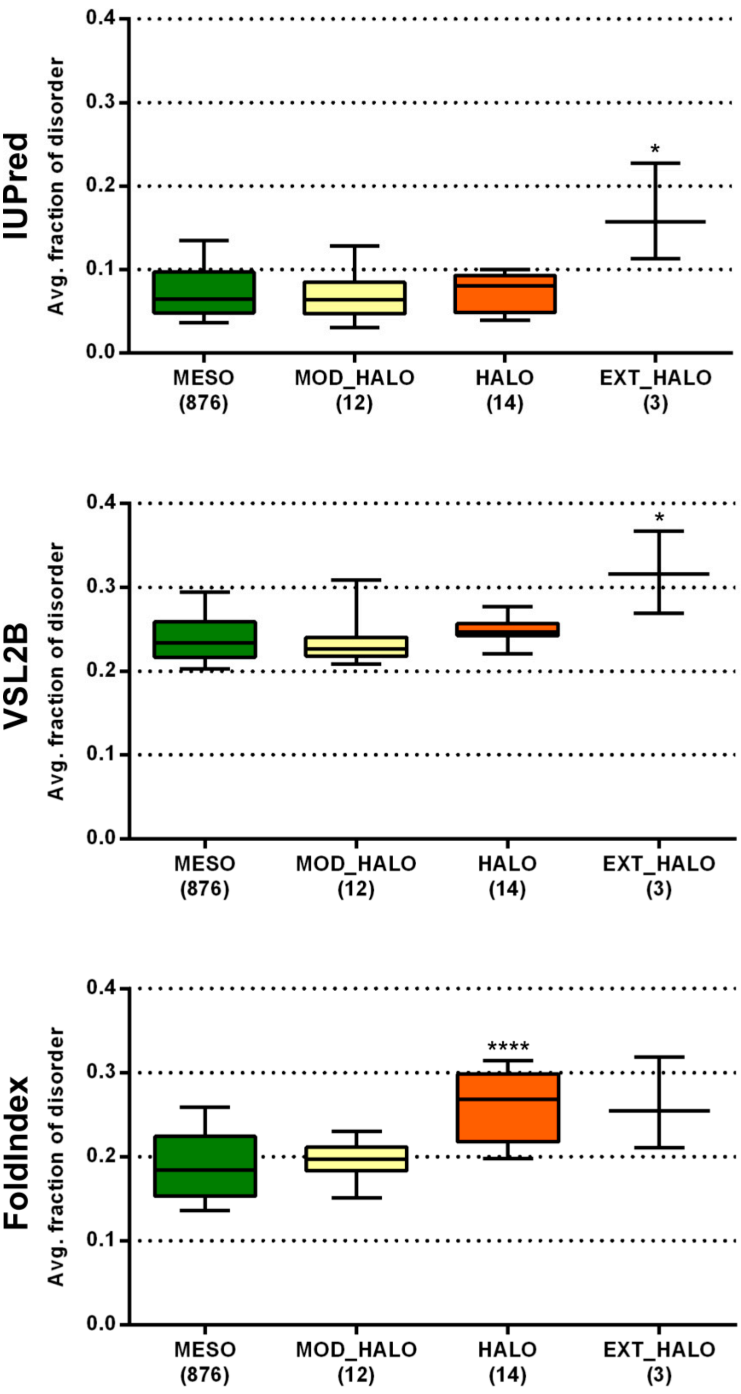

Figure 4. Predicted disorder content of different extremity classes of bacterial halophiles. The average fraction of disordered residues was calculated for each proteome by three prediction methods. The distribution of the calculated values of each class was compared to that of mesophiles by Dunn's multiple comparison test, with the significance levels indicated as stars above the boxes (* $p < 0.05$; **** $p < 0.0001$). MESO—mesophiles, MOD_HALO—moderate halophiles, HALO—halophiles, EXT_HALO—extreme halophiles.

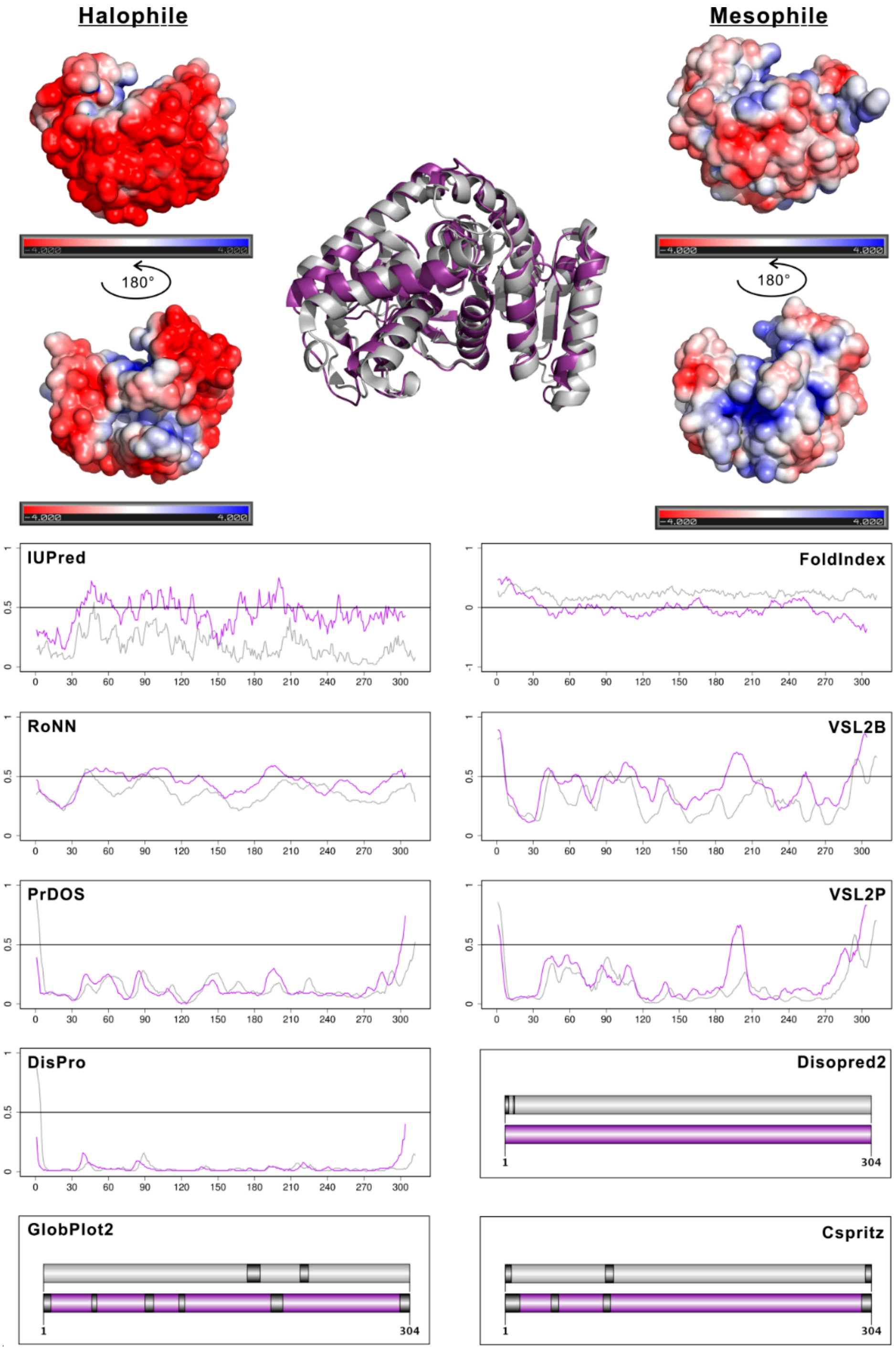

Figure 5. Disorder predictions with ten methods for a halophilic and a mesophilic malate–dehydrogenase enzyme. The halophilic enzyme from *Haloarcula marismortui* (PDB:1D3A) is shown on the left and with purple color in the structural alignment and within the disorder prediction plots. The mesophilic enzyme from *Bacillus anthracis* (PDB:3TL2) is shown on the right and with grey color in the structural alignment and within the disorder prediction plots. For the different disorder prediction methods, order–disorder thresholds were defined as suggested in their respective original publications. In each case values are above the probability threshold mean disorder, except for FoldIndex where it is the opposite. For the seven methods that provide continuous values for disorder probability we show the predicted values as curves. The curves are continuous and just placed on each other, not fitted/aligned. Purple curves show the predictions for the halophilic, grey ones for the mesophilic enzyme. For the three methods that only provide a binary classification of residues as ordered/disordered we show the proteins as colored bars with the predicted disordered regions depicted as dark grey stripes and with the length of the halophilic protein indicated.

In each presented and published case, along with high sequence- and structure-identity, a vast excess of negative charges was found on the surface of the halophilic protein. Proteome-wide computational analyses also supported the finding that this feature does not depend on the individual protein structures, but is applicable to all proteins encoded by halophilic organisms [55] and thus can be regarded as a general adaptive strategy to survive in high salt concentrations. For the three enzyme pairs depicted in Figure 5, Figure S3, and Figure S4 we saw markedly elevated predicted disorder values for the halophilic counterparts with IUPred, FoldIndex (here smaller values mean a larger propensity for disorder), VSL2B, and RONN, and somewhat elevated values with GlobPlot2. Thus, these purely sequence-based predictors are more prone to overestimate (mispredict) disorder due to the adaptive sequence composition bias typical for halophiles. However, the remaining methods that relied not only on sequence but also on phylogenetic information [55] mostly by generating PSI-BLAST profiles (PrDOS, VSL2P, DisPro, DISOPRED2, CSpritz) provided better results. When using PrDOS, DisPro DISOPRED2, or CSpritz, the differences between the predicted values for mesophiles and halophiles were negligible for all three protein pairs.

2.4. Archaea have Lower Disorder Levels than Bacteria due to Their Bias Towards Hyperthermophilic Species

The disorder levels of the Kingdoms of living things have been compared many times by different research groups using different prediction methods [5,6,21]. The conclusions were always similar. Archaea had the lowest disorder level, significantly lower than Bacteria or Eukaryotes. What we observed here, however, by two of the three applied prediction methods, was that mesophilic archaea were not significantly less disordered than mesophilic bacteria (Figure 6). These results imply that Archaea (39/112, ~35% of hyperthermophilic species) are less disordered than Bacteria (10/1052, ~1%) due to their bias towards species living in extreme environments that necessitate such adaptive sequence composition changes, and not due to their phylogenetic identity *per se*.

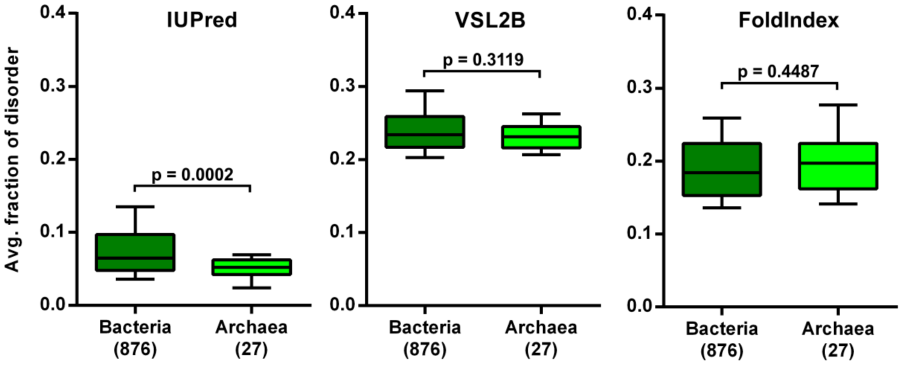

Figure 6. Comparison of the disorder levels of mesophilic Bacteria and Archaea. The average fraction of disordered residues was calculated for each proteome by three prediction methods and the resulting distributions were compared between Bacteria and Archaea by Mann–Whitney U tests, with p-values and the number of proteomes in the two groups indicated.

3. Discussion

In this work we show that the evolutionary adaptation strategies of extremophile proteins to maintain their functions under demanding environmental conditions can mislead disorder prediction methods, especially the ones purely based on sequence, which do not usually face such sequence composition deviations during their training process.

For the three extremophile groups with significant deviations in their proteome-level disorder contents from those of mesophiles, we tried to determine the reasons underlying the observed tendencies. While hyperthermophiles and radiotolerants seem to lack [31] or amass disordered proteins and regions, respectively, as part of their adaptation strategies, in the case of halophiles the very high predicted disorder content is most probably a misprediction issue. Not just the whole proteomes but also the enzymes of halophiles were predicted with very high disorder content. Thus, the observed deviation could not be due to a general trend of employing additional disordered proteins or proteins being extended by disordered regions, but rather to all proteins being predicted with higher values due to their adaptive compositional biases. In halophiles there is an excess of acidic residues on the surface of proteins [48,56], because their interaction with the solvent can maintain a hydration layer that allows halophilic proteins to retain their functional state at molar salt concentrations (the residue–ion interaction model of halophilic protein stabilization) [57]. However, this adaptive compositional bias makes them look similar to intrinsically disordered proteins [56] in the eye of disorder prediction methods. This assumption was confirmed by analyzing halophile–mesophile protein pairs with available structures. Although halophile structures seemed essentially identical to their mesophile counterparts, they had a distinctive pattern of acidic surface charges, which led to significantly elevated predicted disorder tendencies. We thus confidently claim that the repeatedly reported high disorder content in halophiles [20,22] is a result of misprediction by multiple prediction methods, since their well-folded proteins were also predicted to be disordered. Therefore, halophilic organisms do not employ an increased complexity of regulatory mechanisms as proposed previously based on their high predicted disorder levels [22].

Based on our results, it is the lifestyle extremities of the majority of Archaea that lead to very low levels of predicted disorder. Almost all archaeal species live unusual lifestyles. Even the ones categorized as mesophilic are obligate or facultative anaerobe methanogen species living in sediments or wastewater. Importantly, there are many extremely hyperthermophilic (~35%), thermophilic (8%), and thermoacidophilic (12.5%) species among them, which lower their predicted proteome disorder levels. Sequenced extremely halophilic species that could counterbalance this effect by the observed overprediction make up only a smaller fraction of archaeal proteomes (~20%). Based on the observed data we suppose that if one could carefully select archaeal and bacterial species from the same environments and compare them, there would not be a significant difference between their disorder levels.

The structural state of proteins in general, and therefore also protein disorder, is largely dependent on environmental conditions (pH, salt, metals, cofactors, pressure, crowding, etc.). However, disorder predictors work based on the sequence, or other sequence-related more complex features (evolutionary information, predicted secondary structure, etc.), and thus are absolutely not able to relate to the above conditions. On top of this issue, the proteins of organisms facing extreme environmental conditions might have adaptive sequence compositional changes that make their folded proteins look similar to (mesophilic) IDPs. Since the amount of proven IDPs is insufficient to develop dedicated predictors for any group of extremophiles, we can only call for the cautious use of the current methods and formulate some general rules that can be applied to arrive at a more realistic picture when using them. Based on our experience with halophiles, second and third generation prediction methods taking into account evolutionary information, for instance, PrDOS, DisPro DISOPRED2, or CSpritz, are more likely to perform well on sequences with specific adaptive changes, so we would recommend using such slower, albeit more reliable tools on extremophile sequences.

4. Materials and Methods

The collections of representative bacterial and archaeal proteomes with 75% of co-membership threshold were downloaded from the PIR database (release 5.7) [58]. The resulting 1052 bacterial and 112 archaeal proteomes were grouped according to the lifestyle extremities of the corresponding species, which were assigned mainly by relying on the annotations in GOLD (Genomes OnLine Database) [59]. For Bacteria, 8 main lifestyle groups were created: mesophiles, thermophiles, hyperthermophiles, halophiles, acidophiles, alkaliphiles, psychrophiles, and radiotolerants. Species with an optimal growth temperature above 75 °C were considered to be hyperthermophiles, while those with 45 to 75 °C were grouped as thermophiles. Mixed groups (like thermoacidophiles) were not created because the increased fragmentation of the data could obscure important trends. If a given species showed more than one lifestyle extremity, it was put into each of the corresponding groups. For Archaea, 5 groups were created: mesophiles, halophiles, thermophiles, hyperthermophiles, and (hyper)thermoacidophiles, since the other pure extremity groups were not represented by a sufficient number of species (<5). Among Archaea there were many (hyper)thermoacidophiles, which is why we created a separate group for them. Except for those, if a given species showed more than one lifestyle extremity, it was put into each of the corresponding groups.

We searched for the optimal and maximally tolerated NaCl concentrations for the 29 halophilic bacterial species and classified them into three groups (Table S3). We accepted bacteria as extremely halophilic/halotolerant if they could grow in >30 w/v% NaCl. We classified them as halophilic if they required at least 3% NaCl for growth but could also tolerate 15–25%, and moderately halophilic if they required no or <10% NaCl and could not tolerate >15%.

Disorder was calculated for all the sequences in each proteome by three popular prediction methods: IUPred [39], PONDR® VSL2B [42], and FoldIndex [40]. These methods were chosen based on their representation of different theoretical approaches, and because they allowed for the prediction of whole proteomes in a reasonable time. IUPred predicts structural disorder using sequence information alone, based on the possible pairwise interaction energies (calculated from precomputed statistic potentials for each possible amino acid pair) between a given residue and the residues of the surrounding sequence windows. If enough possible favorable interactions prevail with the given protein segment, the amino acid is predicted to be ordered, but if not, it is predicted to be disordered. PONDR® VSL2B is a support vector machine that uses sequence composition features as inputs. FoldIndex takes into consideration the net charge and the sum of normalized hydrophobicity scale values (like charge-hydrophobicity plots) for a sequence window to distinguish between ordered and disordered residues.

The FoldIndex method was reprogrammed based on the simple formula provided in the original article and then the predicted scores were inverted on the order/disorder threshold value, so that the higher values meant more disorder, like in case of the other methods. This reinvented FoldIndex was used for the proteome-level analysis, while for the individual proteins in Figure 5, Figure S3, and Figure S4 the original values provided by the FoldIndex server were used.

Dunn's multiple comparison tests were applied to evaluate the statistical significance of distribution differences when more than two datasets were compared (comparisons of multiple lifestyle groups to mesophiles), while Mann–Whitney U tests were used for pairwise comparisons (e.g., between the bacterial and archaeal mesophiles).

All data processing in this study was performed using custom Perl scripts (v5.10.1, Perl Foundation, Holland, MI, USA). All statistical analyses were implemented in GraphPad Prism 6 (GraphPad Software, San Diego, CA, USA). The Pymol molecular graphics tool (v1.2r2, Schrödinger, LLC, New York, NY, USA), the statistical analysis programming language R (R Foundation, Vienna, Austria), and Microsoft Excel 2010 (Microsoft Corporation, Redmond, WA, USA) were also used for figure preparation.

Supplementary Materials: The following are available online.

Author Contributions: Conceptualization, P.T. and R.P.; methodology, R.P., P.T., and D.K.; investigation, R.P. and D.K.; data curation, R.P.; writing—original draft preparation, R.P. and P.T.; writing—review and editing, P.T. and R.P.; visualization, R.P. and D.K.; funding acquisition, P.T and R.P.

Funding: This research was funded by the Research Foundation Flanders (FWO), grant number G.0029.12; the National Research, Development, and Innovation Office (NKFIH), grant numbers K124670 and FK128133; and the Hungarian Academy of Sciences, grant number PREMIUM-2017-48.

Acknowledgments: The authors kindly acknowledge the technical support of Pallab Bhowmick.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V.N. Introducing protein intrinsic disorder. *Chem. Rev.* **2014**, *114*, 6561–6588. [CrossRef] [PubMed]
2. Van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, A.K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T.; et al. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **2014**, *114*, 6589–6631. [CrossRef] [PubMed]
3. Wright, P.E.; Dyson, H.J. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **1999**, *293*, 321–331. [CrossRef] [PubMed]
4. Wright, P.E.; Dyson, H.J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18–29. [CrossRef] [PubMed]
5. Pancsa, R.; Tompa, P. Structural disorder in eukaryotes. *PloS ONE* **2012**, *7*, e34687. [CrossRef] [PubMed]
6. Ward, J.J.; Sodhi, J.S.; McGuffin, L.J.; Buxton, B.F.; Jones, D.T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **2004**, *337*, 635–645. [CrossRef]
7. Sickmeier, M.; Hamilton, J.A.; LeGall, T.; Vacic, V.; Cortese, M.S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V.N.; et al. DisProt: The Database of Disordered Proteins. *Nucleic Acids Res.* **2007**, *35*, D786–D793. [CrossRef]
8. Dosztanyi, Z.; Meszaros, B.; Simon, I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Briefings Bioinf.* **2010**, *11*, 225–243. [CrossRef]
9. Ferron, F.; Longhi, S.; Canard, B.; Karlin, D. A practical overview of protein disorder prediction methods. *Proteins* **2006**, *65*, 1–14. [CrossRef]
10. Cambillau, C.; Claverie, J.M. Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.* **2000**, *275*, 32383–32386. [CrossRef]
11. Imanaka, T. Molecular bases of thermophily in hyperthermophiles. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* **2011**, *87*, 587–602. [CrossRef]
12. Kiraga, J.; Mackiewicz, P.; Mackiewicz, D.; Kowalczuk, M.; Biecek, P.; Polak, N.; Smolarczyk, K.; Dudek, M.R.; Cebrat, S. The relationships between the isoelectric point and: Length of proteins, taxonomy and ecology of organisms. *BMC Genom.* **2007**, *8*, 163. [CrossRef] [PubMed]
13. Knight, C.G.; Kassen, R.; Hebestreit, H.; Rainey, P.B. Global analysis of predicted proteomes: Functional adaptation of physical properties. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 8390–8395. [CrossRef] [PubMed]
14. Li, W.F.; Zhou, X.X.; Lu, P. Structural features of thermozymes. *Biotechnol. Adv.* **2005**, *23*, 271–281. [CrossRef]
15. Pollo, S.M.; Zhaxybayeva, O.; Nesbo, C.L. Insights into thermoadaptation and the evolution of mesophily from the bacterial phylum Thermotogae. *Can. J. Microbiol.* **2015**, *61*, 655–670. [CrossRef]
16. Robb, F.T.; Clark, D.S. Adaptation of proteins from hyperthermophiles to high pressure and high temperature. *J. Mol. Microbiol. Biotechnol.* **1999**, *1*, 101–105.
17. Szilagyi, A.; Zavodszky, P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: Results of a comprehensive survey. *Structure* **2000**, *8*, 493–504. [CrossRef]
18. Taylor, T.J.; Vaisman, II. Discrimination of thermophilic and mesophilic proteins. *BMC Struct. Biol.* **2010**, *10* (Suppl. 1), S5. [CrossRef]
19. Meszaros, B.; Simon, I.; Dosztanyi, Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.* **2009**, *5*, e1000376. [CrossRef]
20. Vicedo, E.; Schlessinger, A.; Rost, B. Environmental Pressure May Change the Composition Protein Disorder in Prokaryotes. *PloS ONE* **2015**, *10*, e0133990. [CrossRef] [PubMed]
21. Xue, B.; Dunker, A.K.; Uversky, V.N. Orderly order in protein intrinsic disorder distribution: Disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* **2012**, *30*, 137–149. [CrossRef] [PubMed]
22. Xue, B.; Williams, R.W.; Oldfield, C.J.; Dunker, A.K.; Uversky, V.N. Archaic chaos: Intrinsically disordered proteins in Archaea. *BMC Syst. Biol.* **2010**, *4* (Suppl. 1), S1. [CrossRef]
23. Dunker, A.K.; Lawson, J.D.; Brown, C.J.; Williams, R.M.; Romero, P.; Oh, J.S.; Oldfield, C.J.; Campen, A.M.; Ratliff, C.M.; Hipps, K.W.; et al. Intrinsically disordered protein. *J. Mol. Graphics Modell.* **2001**, *19*, 26–59. [CrossRef]
24. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **2002**, *27*, 527–533. [CrossRef]

25. Uversky, V.N.; Gillespie, J.R.; Fink, A.L. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* **2000**, *41*, 415–427. [CrossRef]
26. Elcock, A.H.; McCammon, J.A. Electrostatic contributions to the stability of halophilic proteins. *J. Mol. Biol.* **1998**, *280*, 731–748. [CrossRef] [PubMed]
27. Graziano, G.; Merlino, A. Molecular bases of protein halotolerance. *Biochim. Biophys. Acta* **2014**, *1844*, 850–858. [CrossRef] [PubMed]
28. The UniProt, C. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169. [CrossRef]
29. Calligari, P.A.; Calandrini, V.; Ollivier, J.; Artero, J.B.; Hartlein, M.; Johnson, M.; Kneller, G.R. Adaptation of Extremophilic Proteins with Temperature and Pressure: Evidence from Initiation Factor 6. *J. Phys. Chem. B* **2015**, *119*, 7860–7873. [CrossRef] [PubMed]
30. Katava, M.; Kalimeri, M.; Stirnemann, G.; Sterpone, F. Stability and Function at High Temperature. What Makes a Thermophilic GTPase Different from Its Mesophilic Homologue. *J. Phys. Chem. B* **2016**, *120*, 2721–2730. [CrossRef] [PubMed]
31. Burra, P.V.; Kalmar, L.; Tompa, P. Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes. *PloS ONE* **2010**, *5*, e12069. [CrossRef] [PubMed]
32. Pancsa, R.; Tompa, P. Essential functions linked with structural disorder in organisms of minimal genome. *Biol. Direct* **2016**, *11*, 45. [CrossRef] [PubMed]
33. Litchfield, C.D. Halophiles. *J. Ind. Microbiol. Biotechnol.* **2002**, *28*, 21–22. [CrossRef]
34. Krisko, A.; Smole, Z.; Debret, G.; Nikolic, N.; Radman, M. Unstructured hydrophilic sequences in prokaryotic proteomes correlate with dehydration tolerance and host association. *J. Mol. Biol.* **2010**, *402*, 775–782. [CrossRef] [PubMed]
35. Boothby, T.C.; Pielak, G.J. Intrinsically Disordered Proteins and Desiccation Tolerance: Elucidating Functional and Mechanistic Underpinnings of Anhydrobiosis. *BioEssays* **2017**, *39*. [CrossRef] [PubMed]
36. Tompa, P.; Csermely, P. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J.* **2004**, *18*, 1169–1175. [CrossRef] [PubMed]
37. Hundertmark, M.; Hincha, D.K. LEA (late embryogenesis abundant) proteins and their encoding genes in Arabidopsis thaliana. *BMC Genom.* **2008**, *9*, 118. [CrossRef]
38. Boothby, T.C.; Tapia, H.; Brozena, A.H.; Piszkiewicz, S.; Smith, A.E.; Giovannini, I.; Rebecchi, L.; Pielak, G.J.; Koshland, D.; Goldstein, B. Tardigrades Use Intrinsically Disordered Proteins to Survive Desiccation. *Mol. Cell.* **2017**, *65*, 975–984. [CrossRef]
39. Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **2005**, *347*, 827–839. [CrossRef]
40. Prilusky, J.; Felder, C.E.; Zeev-Ben-Mordehai, T.; Rydberg, E.H.; Man, O.; Beckmann, J.S.; Silman, I.; Sussman, J.L. FoldIndex: A simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **2005**, *21*, 3435–3438. [CrossRef]
41. Yang, Z.R.; Thomson, R.; McNeil, P.; Esnouf, R.M. RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **2005**, *21*, 3369–3376. [CrossRef] [PubMed]
42. Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Dunker, A.K. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **2005**, *61* (Suppl. 7), 176–182. [CrossRef]
43. Ishida, T.; Kinoshita, K. PrDOS: Prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* **2007**, *35*, W460–W464. [CrossRef] [PubMed]
44. Hecker, J.; Yang, J.Y.; Cheng, J. Protein disorder prediction at multiple levels of sensitivity and specificity. *BMC Genom.* **2008**, *9* (Suppl. 1), S9. [CrossRef]
45. Linding, R.; Russell, R.B.; Neduva, V.; Gibson, T.J. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **2003**, *31*, 3701–3708. [CrossRef] [PubMed]
46. Walsh, I.; Martin, A.J.; Di Domenico, T.; Vullo, A.; Pollastri, G.; Tosatto, S.C. CSpritz: Accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res.* **2011**, *39*, W190–W196. [CrossRef]
47. Bohm, G.; Jaenicke, R. A structure-based model for the halophilic adaptation of dihydrofolate reductase from Halobacterium volcanii. *Protein Eng.* **1994**, *7*, 213–220. [CrossRef] [PubMed]

48. Pieper, U.; Kapadia, G.; Mevarech, M.; Herzberg, O. Structural features of halophilicity derived from the crystal structure of dihydrofolate reductase from the Dead Sea halophilic archaeon, Haloferax volcanii. *Structure* **1998**, *6*, 75–88. [CrossRef]
49. Dym, O.; Mevarech, M.; Sussman, J.L. Structural features that stabilize halophilic malate dehydrogenase from an archaebacterium. *Science* **1995**, *267*, 1344–1346. [CrossRef]
50. Frolow, F.; Harel, M.; Sussman, J.L.; Mevarech, M.; Shoham, M. Insights into protein adaptation to a saturated salt environment from the crystal structure of a halophilic 2Fe-2S ferredoxin. *Nat. Struct. Biol.* **1996**, *3*, 452–458. [CrossRef]
51. Winter, J.A.; Christofi, P.; Morroll, S.; Bunting, K.A. The crystal structure of Haloferax volcanii proliferating cell nuclear antigen reveals unique surface charge characteristics due to halophilic adaptation. *BMC Struct. Biol.* **2009**, *9*, 55. [CrossRef] [PubMed]
52. Liao, J.; Liu, M.Y.; Chang, T.; Li, M.; Le Gall, J.; Gui, L.L.; Zhang, J.P.; Jiang, T.; Liang, D.C.; Chang, W.R. Three-dimensional structure of manganese superoxide dismutase from Bacillus halodenitrificans, a component of the so-called "green protein". *J. Struct. Biol.* **2002**, *139*, 171–180. [CrossRef]
53. Baldacci, G.; Guinet, F.; Tillit, J.; Zaccai, G.; de Recondo, A.M. Functional implications related to the gene structure of the elongation factor EF-Tu from Halobacterium marismortui. *Nucleic Acids Res.* **1990**, *18*, 507–511. [CrossRef] [PubMed]
54. Britton, K.L.; Stillman, T.J.; Yip, K.S.; Forterre, P.; Engel, P.C.; Rice, D.W. Insights into the molecular basis of salt tolerance from the study of glutamate dehydrogenase from Halobacterium salinarum. *J. Biol. Chem.* **1998**, *273*, 9023–9030. [CrossRef] [PubMed]
55. Meng, F.; Uversky, V.N.; Kurgan, L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell. Mol. Life Sci.* **2017**, *74*, 3069–3090. [CrossRef] [PubMed]
56. Arakawa, T.; Yamaguchi, R.; Tokunaga, H.; Tokunaga, M. Unique Features of Halophilic Proteins. *Curr. Protein Pept. Sci.* **2017**, *18*, 65–71. [CrossRef]
57. Lenton, S.; Walsh, D.L.; Rhys, N.H.; Soper, A.K.; Dougan, L. Structural evidence for solvent-stabilisation by aspartic acid as a mechanism for halophilic protein stability in high salt concentrations. *Phys. Chem. Chem. Phys.* **2016**, *18*, 18054–18062. [CrossRef]
58. Chen, C.; Natale, D.A.; Finn, R.D.; Huang, H.; Zhang, J.; Wu, C.H.; Mazumder, R. Representative proteomes: A stable, scalable and unbiased proteome set for sequence analysis and functional annotation. *PloS ONE* **2011**, *6*, e18910. [CrossRef]
59. Pagani, I.; Liolios, K.; Jansson, J.; Chen, I.M.; Smirnova, T.; Nosrat, B.; Markowitz, V.M.; Kyrpides, N.C. The Genomes OnLine Database (GOLD) v.4: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **2012**, *40*, D571–D579. [CrossRef]