Published in final edited form as:

Entropy (Basel). 2019 August ; 21(8): . doi:10.3390/e21080764.

Entropy, Fluctuations, and Disordered Proteins

Eshel Faraggi[1,2,*], **A. Keith Dunker**[3], **Robert L. Jernigan**[4], **Andrzej Kloczkowski**[5,6]

[1]Department of Physics, Indiana University Purdue University Indianapolis, Indianapolis, IN 46202, USA

[2]Research and Information Systems, LLC, 1620 E. 72nd ST., Indianapolis, IN 46240, USA

[3]Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202, USA

[4]Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USA

[5]Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, Columbus, OH 43205, USA

[6]Department of Pediatrics, The Ohio State University, Columbus, OH 43205, USA

Abstract

Entropy should directly reflect the extent of disorder in proteins. By clustering structurally related proteins and studying the multiple-sequence-alignment of the sequences of these clusters, we were able to link between sequence, structure, and disorder information. We introduced several parameters as measures of fluctuations at a given MSA site and used these as representative of the sequence and structure entropy at that site. In general, we found a tendency for negative correlations between disorder and structure, and significant positive correlations between disorder and the fluctuations in the system. We also found evidence for residue-type conservation for those residues proximate to potentially disordered sites. Mutation at the disorder site itself appear to be allowed. In addition, we found positive correlation for disorder and accessible surface area, validating that disordered residues occur in exposed regions of proteins. Finally, we also found that fluctuations in the dihedral angles at the original mutated residue and disorder are positively correlated while dihedral angle fluctuations in spatially proximal residues are negatively correlated with disorder. Our results seem to indicate permissible variability in the disordered site, but greater rigidity in the parts of the protein with which the disordered site interacts. This is another indication that disordered residues are involved in protein function.

*Correspondence: efaraggi@gmail.com; Tel.: +1-317-565-3783.

Author Contributions: Conceptualization, E.F., A.K.D., R.L.J., and A.K.; methodology, E.F.; software, E.F.; validation, E.F.; formal analysis, E.F.; investigation, E.F., A.K.D., R.L.J., and A.K.; resources, E.F.; data curation, E.F.; writing—original draft preparation, E.F.; writing—review and editing, E.F., A.K.D., R.L.J., and A.K.; visualization, E.F.; supervision, A.K.; project administration, E.F., A.K.D., R.L.J., and A.K.; and funding acquisition, A.K. and R.J.

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

Keywords

protein disorder; protein structure; entropy; fluctuations; mutations

1. Introduction

Protein disorder, where whole proteins or protein segments are either unstable or metastable, has proven to be a critical property to understand function in biological systems [1–33]. Such disordered proteins and regions have been demonstrated to become more abundant as organism complexity increases [23,29,34–37]. This increase in disorder with organism complexity likely results from the key roles played by disorder in the signaling and regulatory processes underlying cellular differentiation, cell cycle control, gene regulation, and protein–protein interactions, especially enabling the existence of hubs [10,13,38–44]. The origin of these effects arises because disordered proteins enable more diverse function, yet are still able to maintain a high degree of specialization in their specific interactions.

Understanding entropic effects in protein systems is usually difficult, and understanding molecular stability is arguably one of the most important problems in molecular biology, particularly of interest for proteins. This problem is clearly important for understanding the relationship between protein function, structure, and sequence. The full knowledge of protein stabilities requires the reliable evaluation of energies and entropies. This would also aid in the evaluation of structural models of the large number of genes with unknown protein structures. Even a partial understanding of the relationship between entropic sequence effects and structure has already yielded significant success [45–47], and will most likely lead to further success in the future.

Entropy and disorder are intimately related. The aim here was to explore this relationship for proteins. In this work, we analyzed datasets of related proteins with known sequences and structures. We split each cluster of related proteins into two sets, one having a greater sequence similarity than the other. We do this as a way to further explore the relationship to sequence variability. We analyzed the variations in sequence and structure among these sets and quantify their entropies.

Entropy is a global variable. It is the logarithm of the number of phase space states accessible to a system. In very large systems, such that they appear continuous, the number of states is estimated from the phase space volume. Since velocity coordinates in a protein system are related to thermal degrees of freedom and we assumed a constant temperature, the momentum distribution of different conformational states will be similar and, to a good approximation, the difference in the entropy of two states will depend only on the difference between their configuration space volumes.

A previous study of entropy in protein systems by Franzosa and Xia [48] investigated the constraints that structure imposes upon protein evolution. They found that solvent exposure is the most significant structural determinant of residue evolution and also identified a weak effect from the packing density. The relationship between solvent exposure and entropy they found was “strong, positive, and linear”. We investigated these relationships.

2. Materials and Methods

To obtain a dataset of both sequence and structure of related proteins, we clustered PDB [49–51] structures with resolution better than 3 Angstrom, at 25% or greater sequence identity (SID). This was done by first clustering the PDB at 99% SID, to remove redundancies, and then clustering at 25% SID. We used BLASTClust [52] with default parameters to do the clustering. We selected the largest clusters and divided each cluster into two separate sets having SID values in the range of 30–50% SID for one set (A types) and 60–80% for the second set (B types). These collections represent more diverse A sets and less diverse B sets, respectively.

The most abundant cluster of related proteins we found is that of antibodies. We use the notation of L1 to refer to this set. The PDB ID for the structural seed of this set is 5U68E, where the fifth character gives the chain ID. The second most abundant cluster is for kinases. We use the notation of L2 to refer to this set. The PDB ID for the structural seed of this set is 3TTIA. Per review request, we included the third largest cluster as an additional test for some of our results as described in the text. The PDB seed for this cluster is 5F1OB and it is labeled L3. Cartoon representations of the seed structures for L1, L2, and L3, are given in Figure 1. In total, we have 6 sets of proteins: L1, L2, and L3, each split into two sets as described above. The number of proteins in each set and the corresponding descriptors are given in Table 1.

For each of these sets, we collected their sequences and executed a multiple sequence alignment (MSA) using CLUSTALW [53,54] with default parameters. The length of the alignment obtained for each set is also given in Table 1. We observed that, the more varied is the set, the longer is the length of the alignment, as expected. In addition, we found that the more distant are the proteins (lower SID), the longer is the alignment. That is also to be expected since aligning more distant proteins would require inserting more gaps to accommodate the larger sequence variations. We refer to a given column in the MSA as an MSA site. To avoid cases of sparseness in the data, we only used MSA sites having a count of at least 20 amino acids. The number of MSA sites obeying this condition is also given in Table 1. Finally, we give the means, medians, and standard deviations, for the TM-scores [55] between the seed structure and all the rest of the structures in the corresponding set. The Template Modeling score (TM-score) is a parameter to measure protein structure similarity. It is calculated from the distances between the residues of two aligned proteins:

$$TM - score = max[\frac{1}{L_t}\sum_{i=1}^{L_a}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}], \quad (1)$$

where the maximum is taken over all possible alignments. L_t is the length of the target protein, L_a is the length of the protein that is aligned to it, d_i is the distance between the residues at alignment location i, and d_0 is a scaling distance, optimized to $d_0 = 1.24\sqrt[3]{L_t - 15} - 1.8$. With such calibration, the TM-score does not depend on the protein length, and varies between 0 and 1, with 1 indicating a perfect match. A TM-score below 0.2

indicates structurally unrelated proteins, while a TM-score greater than 0.5 indicates the two proteins belong to the same fold.

We observed that the largest set L1 is distributed around relatively distinct structures with some similarity, having a TM-score of just below 0.5. All three sets have TM-scores with means ranging between 0.37 and 0.67. We also observed an appropriate increase in mean and median TM-score for the more similar structures (the B sets - SID 60–80%). We note that the set L2B with SID 60–80% is composed of two overlapping structure clusters. This is exhibited in the range of TM-score and also somewhat skews the results for this set.

To evaluate the amount of disorder at a given MSA site, we introduce the parameter δ that counts the excess number of disordered residues:

$$\delta = \frac{N_{dis} - N_{ord}}{N_{res}}. \tag{2}$$

Here, N_{dis} is the number of residues at an MSA site classified as disordered, i.e., their coordinates appear in the missing coordinates section of the corresponding PDB file (remark 465), and N_{ord} is the number of residues classified as ordered, i.e., with coordinates in the corresponding PDB file. N_{res} is the total number of residues per MSA site, i.e., the total number of sequences that have a residue at this site in the MSA. Note that in a few instances there is a discrepancy between the residue type as it appears in the Uniprot [56] sequence information and that in the PDB file. We discarded these cases and did not classify them as either ordered or disordered, however, they appear in N_{res}. Hence, in general $N_{res} \geq N_{dis} + N_{ord}$, and $-1 \leq \delta \leq 1$. At a given MSA site, $\delta = 1$ indicates all residues at that site are classified as disordered, while $\delta = -1$ indicates that all residues at that site are classified as ordered. Using this approach, for any specific mutation site, we established linked measures of both disorder and structural order. Note that, for a given MSA site, with a number of proteins having a residue at that location, those residues with missing coordinates in the PDB file are labeled disordered. For those residues with coordinates in the PDB, we calculated the structural features. Hence, for a given MSA site, we have both structural and disorder information and we investigated the relationship between them. To ensure that we have some sampling points per MSA site we restricted our attention only to those MSA sites that have at least 20 proteins contributing a residue to the alignment. Note that, since the protein sequences were obtained from Uniprot, residues without PDB coordinates can still contribute to the alignment.

To characterize the structure and fluctuations at a given MSA site, we proceeded as follows. We started by finding the closest long-range contact (CLRC), that is, the closest residue in space which has a sequence separation of more than 5 residues from the MSA site. The distances between the C^{β} (C^{α} for GLY) atoms was used to measure the distance between residues. For all MSA sites, we calculated the average of these distances, d_{av}, and their standard deviations, d_{sd}. We quantified the rotational relationship by calculating the cosines of the angles between the N-C^{α}, C^{α}-C, and C^{α}-C^{β} (0.0 for GLY) bonds for the residue pair identified by a given MSA site and its CLRC site. The cosine of the angle was obtained by taking the dot product of the bond vectors and normalizing by their lengths. The average

values of these were identified as $c1_{av}$, $c2_{av}$, $c3_{av}$, respectively, and the standard deviations of the cosines of these angles are denoted as $c1_{sd}$, $c2_{sd}$, and $c3_{sd}$, respectively.

We also calculated the Shannon entropy for the original MSA site with respect to residue type fluctuations, i.e., from those aligned sequences at a given MSA site we created a probability distribution for residue type and then used $\sum p\ ln(p)$ to calculate the sequence entropy. We performed a similar procedure on the probability distribution for residue types of the CLRC. The Shannon entropy parameter for the original MSA site is denoted s_1, and that for the CLRC s_2. In Figure 2, we give example plots of δ (Figure 2A), the two entropies s_1 and s_2 (Figure 2B), d_{av} (Figure 2C), and d_{sd} (Figure 2D), for all MSA sites having at least 20 residues contributing to the alignment for the set L1A. The relationships even between this limited set seem complex. In what follows, we try to further characterize the features of a given MSA site, and as a first approach determine the correlations between these different features.

In addition, we calculated the average and standard-deviations for the accessible surface area, relative accessible surface area (RSA), and the φ and ψ dihedral angles for both the MSA site and the CLRC site. For the MSA site, we identify these parameters as $a1_{av}$, $a1_{sd}$, $ra1_{av}$, $ra1_{sd}$, $\varphi 1_{av}$, $\varphi 1_{sd}$, $\psi 1_{av}$, and $\psi 1_{sd}$, respectively. For the CLRC, we identify these as $a2_{av}$, $a2_{sd}$, $ra2_{av}$, $ra2_{sd}$, $\varphi 2_{av}$, $\varphi 2_{sd}$, $\psi 2_{av}$, and $\psi 2_{sd}$, respectively. We also calculated the propensity of secondary structure types at the MSA and CLRC sites. In general, we use the index 1 for the MSA site and the index 2 for the CLRC site. The letters *h, c, and e* refer to the helix, coil, and, sheet secondary structure types, respectively. We use the same secondary structure assignment scheme as in SPINE-X [57–59]. From these propensities, we calculate a probability distribution for secondary structure states and from that we calculate the Shannon entropy for secondary structure. We use the notation s_{ss1} for this entropy for the MSA site and s_{ss2} for this entropy at the CLRC.

3. Results and Discussion

To estimate the relationships between the various parameters and the disorder propensity of an alignment site, δ, we calculated Pearson correlation coefficients (correlations) between them. In Tables 2–7, we give the correlations between the various sequence and structure parameters we calculated. Correlations to entropic parameters are given in Table 2. We found that there is a consistent negative correlation between s_2 and δ. The only exception is a marginal positive correlation for L3B. Since L3B is a smaller set, this result may be due to not enough statistics. The positive correlation between s_2 and δ seems to indicate a significant average residue conservation for residues proximate to disordered sites. This can be an indication of their importance for function, and this will be further explored in future work. The correlation values between δ and the entropy of the disordered site are weaker. This indicates that the residue type substitution rate is not significantly different between ordered and disordered residues in this case. This behavior of the correlation could come about because disordered sites form structure with protein or nucleic acid partners, with the resulting structure imparting increased conservation for the amino acids involved in the formation of the complex. Some disordered regions have high conservation throughout [60],

possibly because they form multiple partnerships such that nearly all of their residues are important for at least one critical structure.

In Table 3, we give the correlations between the different entropic parameters. Overall, there is a positive correlation between them. In addition, as expected, correlations within the more similar set L2B appear larger than those of the less similar sets L2A. For set L1, the situation is less clear and may be an indication that set L1 carries more noise.

Correlations between δ and spatial parameters are given in Table 4. We found significant positive correlations between δ and d_{sd}. This is to be expected from the definition of disorder. It is interesting to note that the correlations to the fluctuations of the rotational degrees of freedom ($c1_{sd}$, $c2_{sd}$, $c3_{sd}$) are all negative. This is a puzzle since it seems to associate larger rotational fluctuations with less disorder. It may be an indication that disorder fluctuations are more abundant in the radial directions than in the rotational ones.

To further test the correlation between δ and d_{av} and d_{sd}, we performed a similar analysis on the third most abundant cluster of related sequences of structures deposited in the PDB. The general properties for this cluster are labeled under L3 in Table 1. The seed PDB structure for this protein chain is 5F1OB; a representative cartoon of it is given in Figure 1. For the more diverse subgroup of L3 (30–50% SID), we found correlations of 0.219 and 0.195 for d_{av} and d_{sd}, respectively. For the less diverse subgroup of L3 (60–80% SID), we found correlations of 0.361 and 0.225 for d_{av} and d_{sd}, respectively. These trends are in line with the results for clusters L1 and L2. To estimate the statistical significance of the observed positive correlations between δ and d_{av} and d_{sd}, we performed the following analysis for the set L1A. We started by selecting a random subset of 200 points and calculated the correlation for this subset. We then repeated this process 15 times and use a majority vote to determine the sign of the correlation. Hence, we can consider such a round a flip of a coin, with two possible outcomes. We conducted 10 such rounds and obtained a positive correlation for all of them. In analogy with coins, this would correspond to a p-value of 2^{-10}, indicating confident rejection of the null hypothesis that the correlations are random.

Correlations between δ and accessible surface area parameters are given in Table 5. They are mostly significantly positive, and similarly for the RSA. This is in agreement with the general observation that disordered residues occur in exposed regions of proteins. The negative correlations with fluctuations in accessible surface area may be due to the same observation, as disordered residues would tend to remain exposed, and hence have reduced fluctuations. We could not identify any consistent trend from the correlations with the dihedral angles values (Table 6). However, we did find one for the fluctuations in the dihedral angles. Fluctuations of the dihedral angles are positively correlated with disorder propensity for the original residues, and negatively correlated for the CLRC residues. This is in agreement with the results for the Shannon entropy at these sites, indicating allowed variability at the disordered site, and increased rigidity in the parts of the protein where this disordered site interacts. This may be another indication that disordered residues are involved in protein function.

Correlations between δ and probabilities of secondary structure types are given in Table 7. As expected, we found a significant negative correlation. We found the strongest negative correlation with the propensity of β-sheets. This is also expected. We also found a negative correlation between the entropy of secondary structure and δ for both the original site and its CLRC. This indicates that as disorder is increased at a given MSA site, it becomes more probable for the secondary structure to be of a particular type. This results is consistent with our previous observations for the CLRC. It is difficult to evaluate their significance for the original site since, we have a mixture of disorder and order information.

In Figure 3, we plot the entropy of secondary structure at a given MSA site (s_{ss1}) versus the value of δ at that site. In Figure 4, we plot the same at the CLRC site (s_{ss2} versus δ). In both cases, we see a scatter of points on the x-axis. This indicates a strong effect that is due to conserved sites with zero entropy. If we remove these sites from the calculations of the entropy, we get a reversal in the sign of the correlation, going from −0.213 and −0.141 to 0.139 and 0.247 for Figures 3 and 4, respectively.

We also calculated the correlations between secondary structure type probabilities of the original site and the CLRC. Our aim here was to study the relationship between the structure at the MSA and CLRC sites. Specifically, if there is a difference in that relationship between sites that tend to be more or less disordered. In Table 8, we give these values. We also calculated these correlation values separately for MSA sites with $\delta \geq 0$ (more disordered) and with $\delta < 0$ (more ordered). There is a clear positive correlation for secondary structure types regardless of the state of disorder. One should note that for set L1 there is very little helix conformation, as observed in Figure 1A. This is the reason for the low correlation for this case in Table 8 as there are not enough data.

Finally, part of our aim in the research was to find differences in behavior between two sets of proteins, one with more related proteins, and the other a more diverse set of proteins. Unfortunately, we do not feel confident in drawing conclusions from the data regarding that question. However, future studies may find the data presented here useful.

4. Conclusions

We investigated the relationship between entropy and disorder using native protein structures found in the PDB. By finding clusters of related proteins and studying the MSA of the sequences of these clusters, we were able to establish a link between sequence, structure, and disorder information. We introduced several parameters as measures of fluctuations at a given MSA site and used these as plausible representative of the sequence and structure entropy at that site. We then defined a disorder propensity of an MSA site, δ, and calculated the Pearson correlations between it and our fluctuation parameters. Overall, we found a tendency for negative correlations between disorder and structure. We also found evidence for residue-type conservation for those residues in close proximity to potentially disordered sites. Mutations at the disordered site itself appear to be allowed.

We found significant positive correlations between δ and the fluctuations in the system. This is to be expected from the definition of disorder. It is interesting to note that the correlations

to the fluctuations of the rotational degrees of freedom ($c1_{sd}$, $c2_{sd}$, and $c3_{sd}$) are all negative. This may be an indication that disorder fluctuations are more abundant in the radial direction than in rotational directions but this result will be investigated in future studies.

As expected, we found positive correlations for disorder and accessible surface area, indicating that disordered residues occur in exposed regions of proteins. We found a negative correlations for disorder with fluctuations in accessible surface area. This seems to indicate that disordered residues would tend to remain exposed, and hence with reduced RSA fluctuations. We also found that fluctuations in the dihedral angles at the original mutated residue and disorder are positively correlated while dihedral angle fluctuations in the CLRC residue are negatively correlated with disorder. This agrees with the results for the Shannon entropy at these sites, indicating permissible variability in the disordered site, but greater rigidity in the parts of the protein with which the disordered site interacts. This is another indication that disordered residues are involved in protein function. We also found indications that, as disorder is increased at a given MSA site, it becomes more probable for the secondary structure to be of a particular type.

Acknowledgments

This research was supported in part by NSF grant CNS-0521433, the Lilly Endowment, Inc., and the Indiana METACyt Initiative, through their support for the Indiana University Pervasive Technology Institute. A.K. acknowledges the Visiting Professorship Award to visit the Future Value Creation Research Center at Graduate School of Informatics, Nagoya University, Japan.

Funding: This research was funded by NSF grant DBI 1661391, and NIH grants R01 GM127701 and R01 GM127701-01S1.

Abbreviations

The following abbreviations are used in this manuscript

SID	Sequence Identity
MSA	Multiple Sequence Alignment
CLRC	Closest Long-Range Contact
RSA	Relative Accessible Surface Area

References

1. Dunker AK; Garner E; Guilliot S; Romero P; Albrecht K; Hart J; Obradovic Z; Kissinger C; Villafranca JE Protein disorder and the evolution of molecular recognition: Theory, predictions and observations. Pac. Symp. Biocomput 1998, 3, 473–484.
2. Dunker AK; Obradovic Z The protein trinity: Linking function and disorder. Nat. Biotechnol 2001, 19, 805–806. [PubMed: 11533628]
3. Dunker AK; Brown CJ; Lawson JD; Iakoucheva LM; Obradovic Z Intrinsic disorder and protein function. Biochemistry 2002, 41, 6573–6582. [PubMed: 12022860]
4. Iakoucheva LM; Brown CJ; Lawson JD; Obradović Z; Dunker AK Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol 2002, 323, 573–584. [PubMed: 12381310]
5. Tompa P Intrinsically unstructured proteins. Trends Biochem. Sci 2002, 27, 527–533. [PubMed: 12368089]

6. Tompa P The functional benefits of protein disorder. J. Mol. Struct. Theochem 2003, 666, 361–371.
7. Iakoucheva LM; Radivojac P; Brown CJ; O'Connor TR; Sikes JG; Obradovic Z; Dunker AK The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 2004, 32, 1037–1049. [PubMed: 14960716]
8. Tompa P; Csermely P The role of structural disorder in the function of RNA and protein chaperones. FASEB J. 2004, 18, 1169–1175. [PubMed: 15284216]
9. Ward JJ; Mcguffin LJ; Bryson K; Buxton BF; Jones DT The DISOPRED server for the prediction of protein disorder. Bioinformatics 2004, 20, 2138–2139. [PubMed: 15044227]
10. Dunker AK; Cortese MS; Romero P; Iakoucheva LM; Uversky VN Flexible nets: The roles of intrinsic disorder in protein interaction networks. FEBS J. 2005, 272, 5129–5148. [PubMed: 16218947]
11. Fink AL Natively unfolded proteins. Curr. Opin. Struct. Biol 2005, 15, 35–41. [PubMed: 15718131]
12. Bustos DM; Iglesias AA Intrinsic disorder is a key characteristic in partners that bind 14–3-3 proteins. Proteins Struct. Funct. Bioinform 2006, 63, 35–42.
13. Dosztanyi Z; Chen J; Dunker AK; Simon I; Tompa P Disorder and sequence repeats in hub proteins and their implications for network evolution. J. Proteome Res 2006, 5, 2985–2995. [PubMed: 17081050]
14. Sickmeier M; Hamilton JA; LeGall T; Vacic V; Cortese MS; Tantos A; Szabo B; Tompa P; Chen J; Uversky VN; et al. DisProt: The database of disordered proteins. Nucleic Acids Res. 2006, 35, D786–D793. [PubMed: 17145717]
15. Hilser VJ; Thompson EB Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. Proc. Natl. Acad. Sci. USA 2007, 104, 8311–8315. [PubMed: 17494761]
16. Radivojac P; Iakoucheva LM; Oldfield CJ; Obradovic Z; Uversky VN; Dunker AK Intrinsic disorder and functional proteomics. Biophys. J 2007, 92, 1439–1456. [PubMed: 17158572]
17. Tompa P; Prilusky J; Silman I; Sussman J Structural disorder serves as a weak signal for intracellular protein degradation. Proteins Struct. Funct. Bioinform 2008, 71, 903–909.
18. Tompa P; Fuxreiter M Fuzzy complexes: Polymorphism and structural disorder in protein-protein interactions. Trends Biochem. Sci 2008, 33, 2–8. [PubMed: 18054235]
19. Fong JH; Shoemaker BA; Garbuzynskiy SO; Lobanov MY; Galzitskaya OV; Panchenko AR Intrinsic disorder in protein interactions: Insights from a comprehensive structural analysis. PLoS Comput. Biol 2009, 5, e1000316. [PubMed: 19282967]
20. Vavouri T; Semple JI; Garcia-Verdugo R; Lehner B Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell 2009, 138, 198–208. [PubMed: 19596244]
21. Mittag T; Kay LE; Forman-Kay JD Protein dynamics and conformational disorder in molecular recognition. J. Mol. Recognit 2010, 23, 105–116. [PubMed: 19585546]
22. Hagai T; Azia A; Tóth-Petróczy Á; Levy Y Intrinsic disorder in ubiquitination substrates. J. Mol. Biol 2011, 412, 319–324. [PubMed: 21802429]
23. Schlessinger A; Schaefer C; Vicedo E; Schmidberger M; Punta M; Rost B Protein disorder: A breakthrough invention of evolution? Curr. Opin. Struct. Biol 2011, 21, 412–418. [PubMed: 21514145]
24. Bustos DM The role of protein disorder in the 14–3-3 interaction network. Mol. Biosyst 2012, 8, 178–184. [PubMed: 21947246]
25. Vacic V; Markwick PR; Oldfield CJ; Zhao X; Haynes C; Uversky VN; Iakoucheva LM Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. PLoS Comput. Biol 2012, 8, e1002709. [PubMed: 23055912]
26. Zhang T; Faraggi E; Xue B; Dunker AK; Uversky VN; Zhou Y SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method. J. Biomol. Struct. Dyn 2012, 29, 799–813. [PubMed: 22208280]
27. Zhang T; Faraggi E; Li Z; Zhou Y Intrinsically semi-disordered state and its role in induced folding and protein aggregation. Cell Biochem. Biophys 2013, 67, 1193–1205. [PubMed: 23723000]

28. Hughes SL; Schart V; Malcolmson J; Hogarth KA; Martynowicz DM; Tralman-Baker E; Patel SN; Graether SP The importance of size and disorder in the cryoprotective effects of dehydrins. Plant Physiol. 2013, 163, 1376–1386. [PubMed: 24047864]
29. Uversky VN A decade and a half of protein intrinsic disorder: Biology still waits for physics. Protein Sci 2013, 22, 693–724. [PubMed: 23553817]
30. Habchi J; Tompa P; Longhi S; Uversky VN Introducing protein intrinsic disorder. Chem. Rev 2014, 114, 6561–6588. [PubMed: 24739139]
31. Berlow RB; Dyson HJ; Wright PE Functional advantages of dynamic protein disorder. FEBS Lett. 2015, 589, 2433–2440. [PubMed: 26073260]
32. Varadi M; Zsolyomi F; Guharoy M; Tompa P Functional advantages of conserved intrinsic disorder in RNA-binding proteins. PLoS ONE 2015, 10, e0139731. [PubMed: 26439842]
33. Zhang T; Faraggi E; Li Z; Zhou Y Intrinsic disorder and Semi-disorder prediction by SPINE-D In Prediction of Protein Secondary Structure; Humana Press: New York, NY, USA, 2017; pp. 159–174.
34. Brown CJ; Johnson AK; Dunker AK; Daughdrill GW Evolution and disorder. Curr. Opin. Struct. Biol 2011, 21, 441–446. [PubMed: 21482101]
35. Mosca R; Pache RA; Aloy P The role of structural disorder in the rewiring of protein interactions through evolution. Mol. Cell. Proteom 2012, 11.
36. Niklas KJ; Cobb ED; Dunker AK The number of cell types, information content, and the evolution of complex multicellularity. Acta Soc. Bot. Pol 2014, 83, 337–347.
37. Yruela I; Oldfield CJ; Niklas KJ; Dunker AK Evidence for a strong correlation between transcription factor protein disorder and organismic complexity. Genome Biol. Evol 2017, 9, 1248–1265. [PubMed: 28430951]
38. Haynes C; Oldfield CJ; Ji F; Klitgord N; Cusick ME; Radivojac P; Uversky VN; Vidal M; Iakoucheva LM Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput. Biol 2006, 2, e100. [PubMed: 16884331]
39. Singh GP; Ganapathi M; Dash D Role of intrinsic disorder in transient interactions of hub proteins. Proteins Struct. Funct. Bioinform 2007, 66, 761–765.
40. Xie H; Vucetic S; Iakoucheva LM; Oldfield CJ; Dunker AK; Uversky VN; Obradovic Z Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J. Proteome Res 2007, 6, 1882–1898. [PubMed: 17391014]
41. Dunker AK; Bondos SE; Huang F; Oldfield CJ Intrinsically disordered proteins and multicellular organisms In Seminars in Cell & Developmental Biology; Elsevier: Amsterdam, The Netherlands, 2015; Volume 37, pp. 44–55.
42. Niklas KJ; Bondos SE; Dunker AK; Newman SA Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications. Front. Cell Dev. Biol 2015, 3, 8. [PubMed: 25767796]
43. Niklas KJ; Dunker AK; Yruela I The evolutionary origins of cell type diversification and the role of intrinsically disordered proteins. J. Exp. Bot 2018, 69, 1437–1446. [PubMed: 29394379]
44. Zhou J; Zhao S; Dunker AK Intrinsically disordered proteins link alternative splicing and post-translational modifications to complex cell signaling and regulation. J. Mol. Biol 2018, 430, 2342–2359. [PubMed: 29626537]
45. Marks DS; Colwell LJ; Sheridan R; Hopf TA; Pagnani A; Zecchina R; Sander C Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 2011, 6, e28766. [PubMed: 22163331]
46. Schmiedel JM; Lehner B Determining protein structures using deep mutagenesis. Nat. Genet 2019, 51, 1177–1186. [PubMed: 31209395]
47. Rollins NJ; Brock KP; Poelwijk FJ; Stiffler MA; Gauthier NP; Sander C; Marks DS Inferring protein 3D structure from deep mutation scans. Nat. Genet 2019, 51, 1170–1176. [PubMed: 31209393]
48. Franzosa EA; Xia Y Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol 2009, 26, 2387–2395. [PubMed: 19597162]

49. Bernstein FC; Koetzle TF; Williams GJ; Meyer EF Jr.; Brice MD; Rodgers JR; Kennard O; Shimanouchi T; Tasumi M The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol 1977, 112, 535–542. [PubMed: 875032]
50. Berman HM; Westbrook J; Feng Z; Gilliland G; Bhat TN; Weissig H; Shindyalov IN; Bourne PE The protein data bank. Nucleic Acids Res. 2000, 28, 235–242. [PubMed: 10592235]
51. Berman H; Henrick K; Nakamura H Announcing the worldwide protein data bank. Nat. Struct. Mol. Biol 2003, 10, 980.
52. Altschul SF; Gish W; Miller W; Myers EW; Lipman DJ Basic local alignment search tool. J. Mol. Biol 1990, 215, 403–410. [PubMed: 2231712]
53. Thompson JD; Higgins DG; Gibson TJ CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994, 22, 4673–4680. [PubMed: 7984417]
54. Larkin MA; Blackshields G; Brown N; Chenna R; McGettigan PA; McWilliam H; Valentin F; Wallace IM; Wilm A; Lopez R; et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007, 23, 2947–2948. [PubMed: 17846036]
55. Zhang Y; Skolnick J TM-align: A protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005, 33, 2302–2309. [PubMed: 15849316]
56. Consortium U UniProt: The universal protein knowledgebase. Nucleic Acids Res. 2018, 46, 2699. [PubMed: 29425356]
57. Faraggi E; Yang Y; Zhang S; Zhou Y Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure 2009, 17, 1515–1527. [PubMed: 19913486]
58. Faraggi E; Xue B; Zhou Y Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins Struct. Funct. Bioinform 2009, 74, 847–856.
59. Faraggi E; Zhang T; Yang Y; Kurgan L; Zhou Y SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J. Comput. Chem 2012, 33, 259–267. [PubMed: 22045506]
60. Zhou J; Oldfield CJ; Yan W; Shen B; Dunker AK Intrinsically disordered domains: Sequence→ disorder→ function relationships. Protein Sci. 2019.

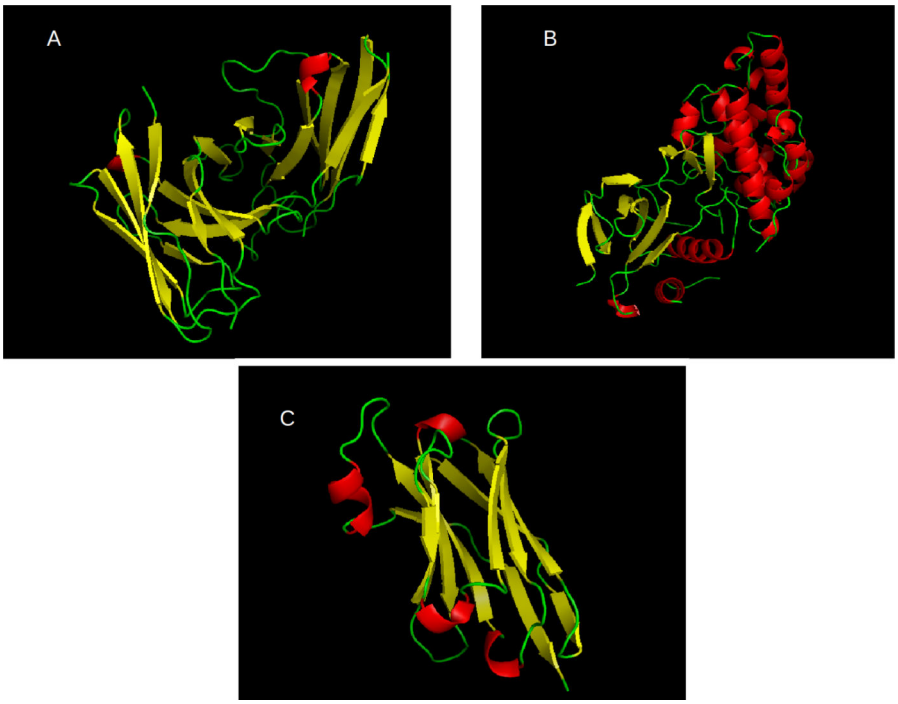

Figure 1.

Cartoons of the structures for the seeds for the sets L1, L2, and L3. (**A**) Seed for the L1 sets L1A and L1B, an antibody fragment, PDBID: 5U68E. (**B**) Seed for the L2 sets L2A and L2B, JNK3 mitogen-activated protein kinase 10, PDBID: 3TTIA. (**C**) Seed for the L3 sets L3A and L3B, nanobody MU551, PDBID: 5F1OB. The color scheme is according to the secondary structure types, with beta strands yellow, helix red and coil green. Note that we keep a dark background to aid in viewing loops and especially loops with missing residues.

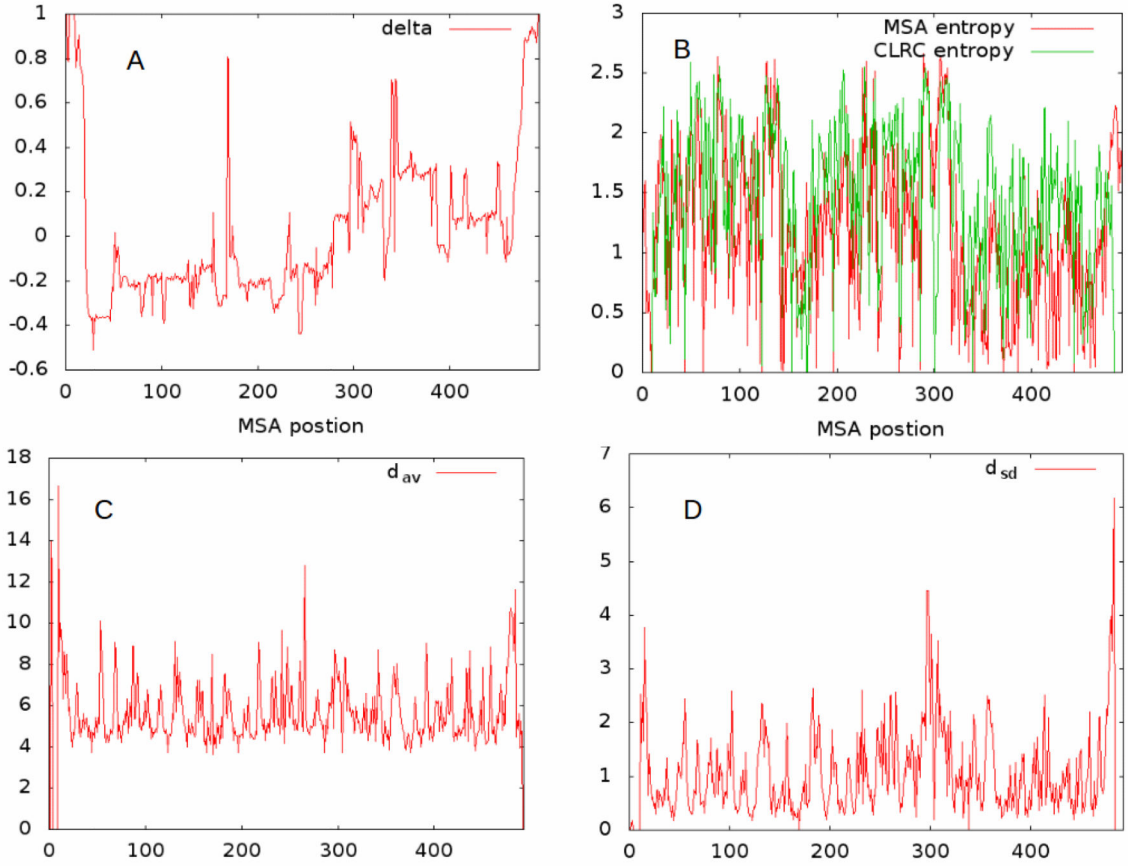

Figure 2.

Example plots of: δ (**A**); the two entropies s_1 and s_2 (**B**); d_{av} (**C**); and d_{sd} (**D**), for all MSA sites having at least 20 residues contributing to the alignment for the set L1A.

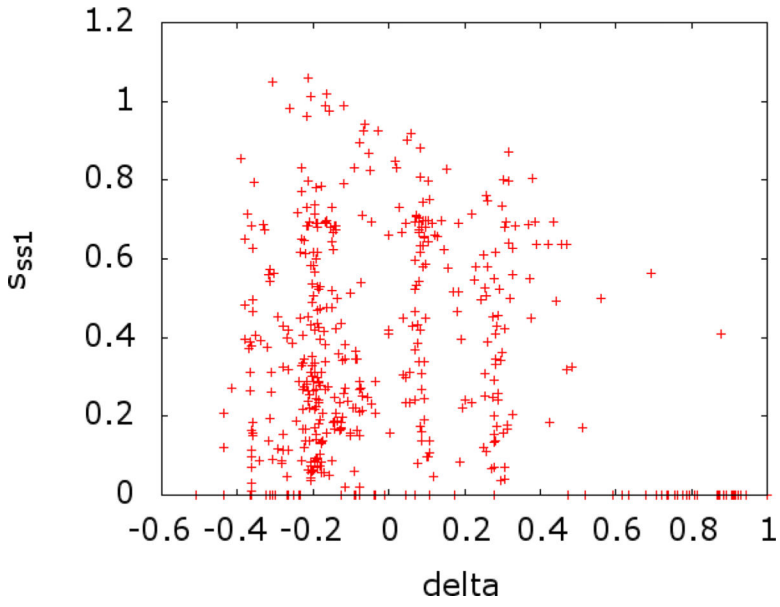

Figure 3.
The entropy of secondary structure at a given MSA site versus the value of δ at that site.

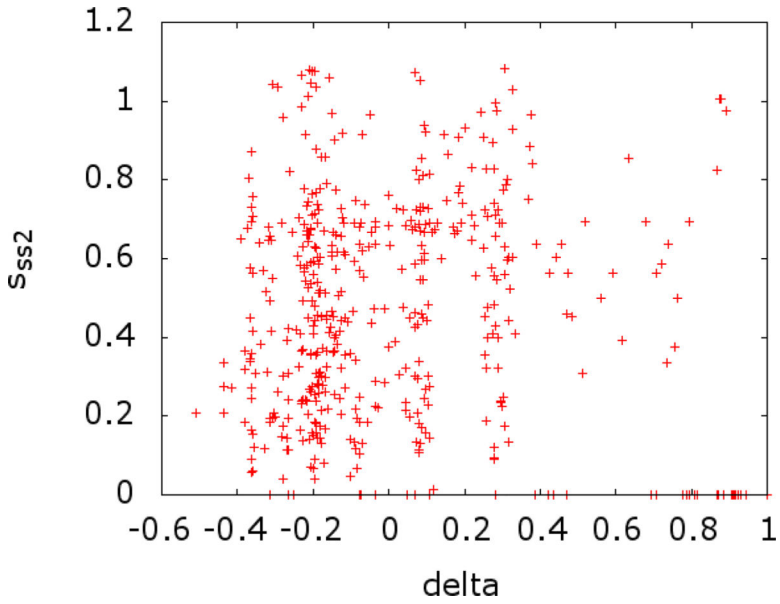

Figure 4.
The entropy of secondary structure at the CLRC to the MSA site versus the value of δ at that site.

Table 1.

Properties of the protein sets used.

Set:	**L1**		**L2**		**L3**	
SID to seed:	**30–50%**	**60–80%a**	**30–50%**	**60–80%**	**30–50%**	**60–80%**
Number of proteins	1759	586	398	261	378	393
Length of MSA	666	565	930	498	535	228
>20 MSA sites	494	397	569	157	355	182
TMS mean	0.45	0.49	0.51	0.37	0.53	0.67
TMS median	0.43	0.49	0.24	0.36	0.23	0.87
TMS STDEV	0.06	0.06	0.26	0.05	0.32	0.30
Number of residues in seed protein	294		464		163	
Seed PDBID	5U68E		3TTIA		5F1OB	
Function title	Antibody fragment		JNK3 mitogen-activated kinase		Nanobody MU551	

Properties of the most abundant clusters of related proteins in the PDB.

Table 2.

Correlations between the disorder propensity δ and entropic parameters.

Parameter	L1		L2		L3	
	30–50%	**60–80%**	**30–50%**	**60–80%**	**30–50%**	**60–80%**
s_1	−0.074	0.092	0.030	0.240	0.020	0.250
s_2	−0.402	−0.316	−0.346	−0.286	−0.145	0.042
s_{ss1}	−0.213	−0.138	0.050	−0.077	−0.205	0.080
s_{ss2}	−0.141	−0.216	−0.039	−0.101	−0.128	0.022

s_1 and s_2 are entropies with respect to fluctuations in residue type for the MSA and CLRC sites, respectively, and s_{ss1} and s_{ss2} are entropies with respect to fluctuations in secondary structure assignment for the MSA and CLRC sites, respectively.

Table 3.

Correlations between the entropic parameters.

Parameter	L1		L2	
	30–50%	**60–80%**	**30–50%**	**60–80%**
s_1, s_2	0.467	0.319	0.526	0.681
s_1, s_{ss1}	0.267	0.179	0.280	0.403
s_1, s_{ss2}	0.207	0.206	0.280	0.544
s_2, s_{ss1}	0.442	0.424	0.284	0.483
s_2, s_{ss2}	0.562	0.601	0.458	0.745
s_1, s_{ss2}	0.375	0.370	0.374	0.549

Correlation between the different entropic parameters calculated for the different sets of aligned proteins.

Table 4.

Correlations between the disorder propensity δ and spatial parameters.

Parameter	L1		L2	
	30–50%	**60–80%**	**30–50%**	**60–80%**
d_{av}	0.082	−0.097	0.354	0.296
d_{sd}	0.191	0.114	0.425	0.248
$c1_{av}$	0.259	0.181	0.026	0.147
$c1_{sd}$	−0.131	−0.163	−0.013	−0.182
$c2_{av}$	0.247	0.142	0.052	0.236
$c2_{sd}$	−0.137	−0.148	−0.042	−0.143
$c3_{av}$	−0.050	−0.097	0.061	−0.237
$c3_{sd}$	−0.243	−0.159	−0.002	−0.161

Correlations between spatial characteristics and the disorder propensity at a given MSA site.

Table 5.

Correlations between the disorder propensity δ and ASA parameters.

Parameter	L1		L2	
	30–50%	**60–80%**	**30–50%**	**60–80%**
$a1_{av}$	0.167	0.093	0.388	0.450
$a1_{sd}$	−0.077	−0.051	0.340	0.120
$a2_{av}$	0.129	−0.040	0.317	0.253
$a2_{sd}$	−0.153	−0.235	0.215	−0.078
$ra1_{av}$	0.180	0.090	0.409	0.409
$ra1_{sd}$	−0.040	−0.042	0.393	0.133
$ra2_{av}$	0.163	0.006	0.320	0.214
$ra2_{sd}$	−0.163	−0.206	0.214	−0.186

Correlations between ASA characteristics and the disorder propensity at a given MSA site.

Table 6.

Correlations between the disorder propensity δ and dihedral angle parameters.

Parameter	L1		L2	
	30–50%	**60–80%**	**30–50%**	**60–80%**
$\varphi 1_{av}$	0.269	0.216	0.410	0.326
$\varphi 1_{sd}$	0.106	0.029	0.480	−0.026
$\psi 1_{av}$	−0.006	−0.041	0.213	0.054
$\psi 1_{sd}$	0.035	0.028	0.262	0.235
$\varphi 2_{av}$	0.184	0.272	−0.003	0.269
$\varphi 2_{sd}$	−0.204	−0.173	−0.015	−0.314
$\psi 2_{av}$	−0.264	−0.243	−0.079	−0.292
$\psi 2_{sd}$	−0.156	−0.125	−0.082	−0.192

Correlations between dihedral angles characteristics and the disorder propensity at a given MSA site.

Table 7.

Correlations between the disorder propensity δ and secondary structure probabilities.

Parameter	L1		L2	
	30–50%	**60–80%**	**30–50%**	**60–80%**
ss1h	−0.093	−0.067	−0.378	−0.120
ss1c	−0.333	−0.183	−0.303	−0.354
ss1e	−0.339	−0.189	−0.207	−0.393
ss2h	−0.104	−0.084	−0.385	−0.150
ss2c	−0.316	−0.144	−0.400	−0.319
ss2e	−0.424	−0.264	−0.294	−0.575

Correlations between secondary structure probabilities and the disorder propensity at a given MSA site.

Table 8.

Correlations between secondary structure type probabilities.

Type	L1		L2	
	30–50%	**60–80%**	**30–50%**	**60–80%**
Helix	0.198	0.179	0.509	0.289
Coil	0.611	0.480	0.414	0.442
Sheet	0.741	0.716	0.643	0.620
Helix ($\delta \geq 0$)	−0.033	−0.007	0.551	0.175
Coil ($\delta \geq 0$)	0.566	0.332	0.893	0.993
Sheet ($\delta \geq 0$)	0.768	0.585	0.721	0.658
Helix ($\delta < 0$)	0.277	0.307	0.450	0.278
Coil ($\delta < 0$)	0.588	0.559	0.341	0.354
Sheet ($\delta < 0$)	0.707	0.770	0.632	0.523

Correlations between the secondary structure probabilities of the original MSA site and its CLRC.