

SUBCLASSIFYING DISORDERED PROTEINS BY THE CH-CDF PLOT METHOD[†]

FEI HUANG, CHRISTOPHER OLDFIELD, JINGWEI MENG, WEI-LUN HSU, BIN XUE[#], VLADIMIR N. UVERSKY[#], PEDRO ROMERO AND A. KEITH DUNKER^{*}

Department of Biochemistry and Molecular Biology, Indiana University School of Medicine

*410 W 10th street, Suite 5000,
Indianapolis, IN 46202*

*Emails: {huangfei, cjoldfie, menj, hsu20, or kedunker} @iupui.edu
promero@compbio.iupui.edu
{binxue or vuversk} @health@usf.edu*

Intrinsically disordered proteins (IDPs) are associated with a wide range of functions. We suggest that sequence-based subtypes, which we call flavors, may provide the basis for different biological functions. The problem is to find a method that separates IDPs into different flavor / function groups. Here we discuss one approach, the (Charge-Hydrophathy) versus (Cumulative Distribution Function) plot or CH-CDF plot, which is based the combined use of the CH and CDF disorder predictors. These two predictors are based on significantly different inputs and methods. This CH-CDF plot partitions all proteins into 4 groups: structured, mixed, disordered, and rare. Studies of the Protein Data Bank (PDB) entries and homologous show different structural biases for each group classified by the CH-CDF plot. The mixed class has more order-promoting residues and more ordered regions than the disordered class. To test whether this partition accomplishes any functional separation, we performed gene ontology (GO) term analysis on each class. Some functions are indeed found to be related to subtypes of disorder: the disordered class is highly active in mitosis-related processes among others. Meanwhile, the mixed class is highly associated with signaling pathways, where having both ordered and disordered regions could possibly be important.

Keywords: Intrinsically disordered protein, classification, CH plot, CDF prediction

1. Introduction

Unlike structured proteins folded into compact structures, intrinsically disordered proteins (IDPs) exist as flexible ensembles [1]. IDPs are very common. They comprise approximately 25% to 30% of eukaryotic proteomes [2]. Over 50% of eukaryotic proteins and 70% of signaling proteins have long disordered regions [3]. A wide range of biological activities are associated with IDPs, such as providing sites for post-translational modifications, providing sites for binding to partners via short linear motifs, acting as scaffolds by binding to multiple partners, etc. [4-6].

Studies of ordered proteins indicate that homologous proteins typically have conserved 3D structures [7-9]. Thus, structure similarity is used as an important criterion when examining related proteins. Most proteins with similar structure have a common evolutionary origin, and as a consequence their functions are typically closely related [7-9]. Databases such as SCOP [7] and CATH [8] have been constructed using this line of reasoning. These databases serve as a great

[†] This work is supported by the grants R01 LM007688-01A1, GM071714-01A2 from the National Institute of Health and EF 0849803 from the National Science Foundation

^{*} Corresponding author

[#] Current Address: Department of Molecular Medicine, University of South Florida, Tampa, FL 33620

resource for understanding the nature of the various relationships between protein structure and function, and they are widely used in various molecular and biological areas of science [7,8].

Since IDPs lack 3D structure, unique structure can't be used to partition IDPs into subtypes. We previously tried an approach based on disorder prediction to cluster IDP regions into different subtypes, which we called flavors, and some functions showed a weak partitioning among the different flavors [10]. Here our goal is to re-explore the overall idea of partitioning disordered proteins into subtypes, but using a different predictive approach. The previous approach used residue-by-residue order / disorder predictions over IDP regions of proteins [10], but a weakness of that approach was that the disordered regions varied markedly in length, which greatly complicated the interpretation.

Using *mus. musculus* protein data, here we will test an approach in which the order / disorder predictions are binary for the whole protein, indicating that a given protein is more ordered or more disordered overall. The two binary prediction tools are the charge-hydrophathy (CH) plot [12,13] and the cumulative distribution function (CDF) [14]. Applying both methods to a protein could have four possible outcomes: both methods predict order, both methods predict disorder, the CH predicts disorder while the CDF predicts order, and vice versa. When both methods predict order, the protein is likely to be predominantly structured and to be found in the Protein Data Bank (PDB) [15]. When both methods predict disorder, the proteins are likely to be IDPs with high net charge and very little structure, and thus are likely to be more extended. If CDF predicts disorder and CH predicts order or vice versa, then these two sets of proteins have both order and disorder tendencies, but with different characteristics for each tendency. Thus, overall, the CH-CDF plot separates proteins into 4 groups with differing order and disorder characteristics.

The CH-CDF plot was previously used to compare the structure-disorder tendencies of the proteomes from several species within the phylum Apicomplexa, which include Plasmodia, Trypanosomes, and Giardia [16]. The CH-CDF plot has also been used to classify the transcription factors associated with the induction of pluripotent stem cells [17] and a collection of plant-specific developmental proteins as well as their distinctive domains [18]. In all three cases, the distributions of the various proteins and domains among the four outcomes provide overviews of similarities and differences between the different sets of proteins [16-18]. According to these prior studies, the CH-CDF plot appears to be useful for identifying overall structure-disorder trends for collections of proteins. Here we apply the CH-CDF plot to the mouse proteome and then investigate whether the four outcomes are associated with differences in structure and function for these proteins.

2. Results

2.1 CH-CDF plot

First, let's illustrate the overall development of the CH-CDF plot. Figure 1A shows the placement of a disordered protein (red) and an ordered protein (blue) onto a CH plot, where the indicated linear discriminant (i.e., a linear classification boundary) was developed from a large training set of proteins [12, 13]. Note that disordered proteins have a higher net charge and lower hydrophathy compared to ordered proteins. We use the vertical distances from each protein-representing point

to the separation line as the Y-coordinate of that protein in the CH-CDF plot, so when Y is positive, the protein is indicated to be disordered. Figure 1B shows the PONDR VSL2 [19] plots for the same pair of disordered (red) and ordered (blue) proteins. In Figure 1C, the data in 1B are plotted to produce the CDF plot, where the X-axis is the prediction score and the Y axis is the total fraction of sequence loci having that score or lower. Note the different shapes for the CDFs for the ordered (blue) and disordered (red) proteins. An ordered protein's CDF curve occupies the upper part of the graph, while an IDP's CDF curve resides in the lower part of the graph. The optimal separation line, represented as a collection of 7 discrete points, was previously estimated for a large number of structured and disordered proteins [14]. The X-axis for the CH-CDF plot is calculated as the average of the vertical distances from the CDF curve to the seven boundary points. Thus, the ordered proteins are given positive values and disordered proteins are given negative values with respect to the X-axis in the CH-CDF plot.

The entire mouse (*Mus.musculus*) proteome is put onto the CH-CDF plot in Figure 1D. The descriptions of the prediction characteristics for the proteins in each quadrant are included in this plot.

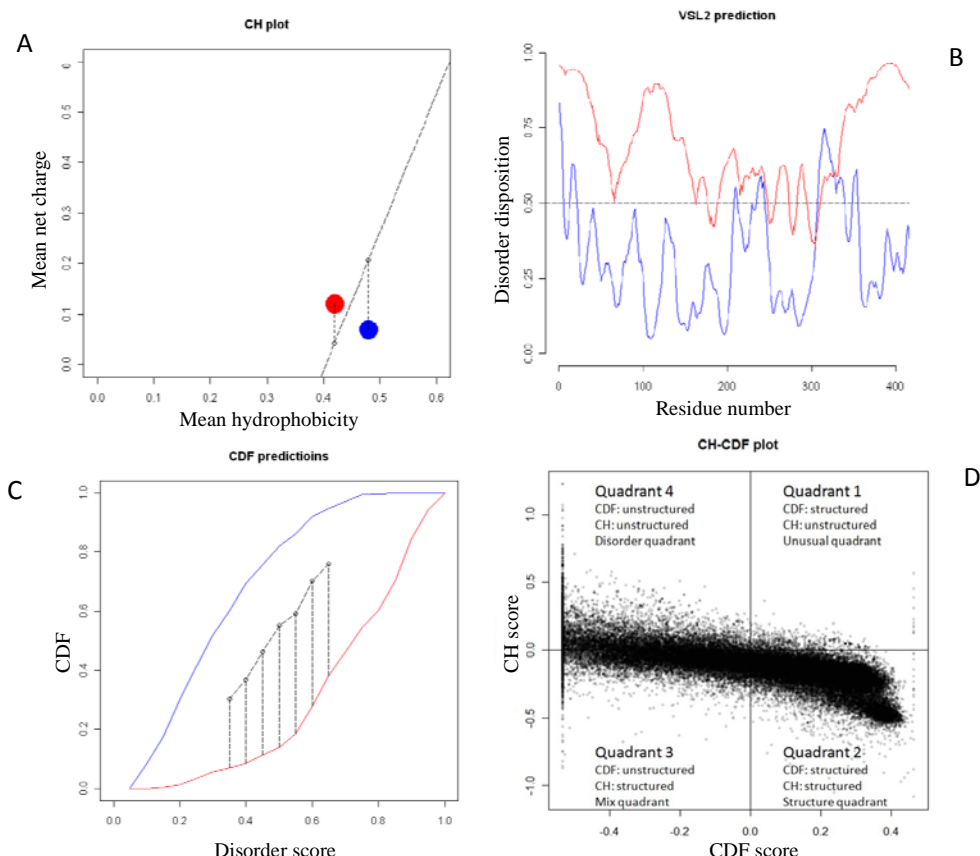


Fig.1. The CH-CDF Plot. A. An example of a CH graph, with a linear classification boundary ($y=2.743x-1.109$) and a hypothetical IDP and hypothetical structured protein. B. VSL2 prediction curve for an IDP (red) and a structured protein (blue). C. CDF curve of the two proteins in B. Vertical lines are the distance of to calculate CDF score. D. The entire mouse proteome is put onto a CH-CDF plot.

One rationale behind using CH-CDF plot in subclassification of disordered proteins is that CDF examines many more protein attributes than a CH plot, which only uses charge and hydrophathy for prediction. Consequently, the CDF curve is more sensitive to disorder than the CH plot [11]. Proteins predicted to be ordered by the CH plot but disordered by CDF (as in Q3) are low in net charge and hydrophobic, but with other features resembling an unstructured protein. Therefore, we propose that such proteins could have both disordered and ordered regions, and we refer them as mixed proteins. Meanwhile, proteins predicted to be unstructured by both methods are referred as disordered (Q4) and proteins predicted to be ordered by both predictors are likely to be structured proteins (Q2). As for proteins in Q1, their number is very small compared to other three quadrants. We do have some assumptions about them, but we have not reached any refined conclusion yet. So here, we refer them as rare proteins.

2.2 PDB coverage

PDB contains protein structures, and thus PDB is biased more towards ordered proteins than disordered. Fig. 2 shows PDB coverage percentages of various proteins vs. their length for each quadrant. By coverage percentage, we mean the percent of a given sequence that forms structure and is observed in PDB. As expected, more of the proteins in Q2 have higher coverage.

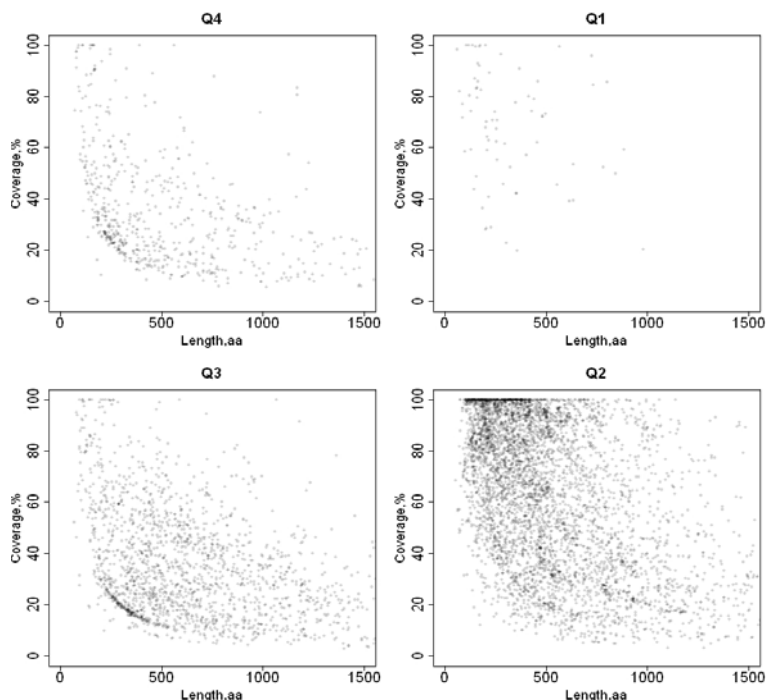


Fig.2. PDB coverage percentages of proteins classified into 4 quadrants.

To quantitate the coverage data of Figure 2, histogram summaries for each quadrant were constructed (Figure 3). When proteins are indicated to be disordered by the CDF (Q3 and Q4), the coverage summaries are similar and mostly show a small fraction of coverage. When proteins are

indicated to be structured by CDF (Q1 and Q2), the coverage summaries are similarly biased towards structure. There are other factors to consider as shown below.

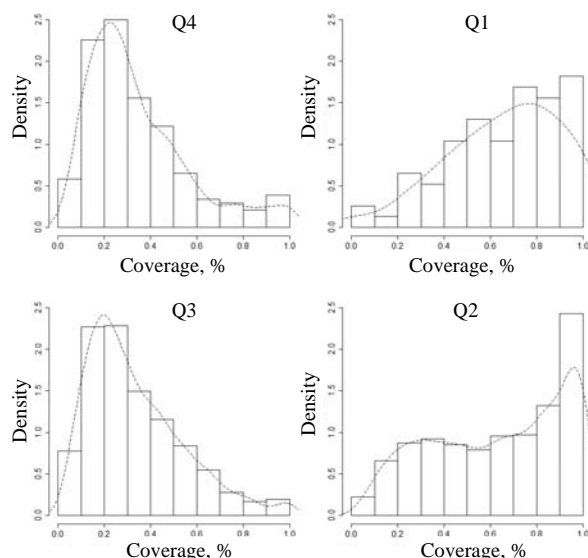


Fig.3 PDB coverage percentage histogram for all four quadrants

Another important consideration is whether a protein has any structure at all in PDB. The structure quadrant (Q2) has the highest fraction of proteins identified with at least one PDB hit, whereas the disorder quadrant (Q4) has the lowest fraction (Figure 4). Note that the mixed quadrant (Q3) actually is the second highest. Its fraction is close to the structure quadrant (Q2), and much higher than the disorder quadrant (Q4). These data suggest that mixed proteins have more structured regions than disordered proteins. Recall Fig. 2 and Fig. 3, which have shown that the coverage percentages for proteins in Q3 are very low, around 20-30% only. Taken together, these mixed proteins are more likely to have structured local regions compared to the disorder quadrant (Q4), so that they have a higher fraction of PDB hits.

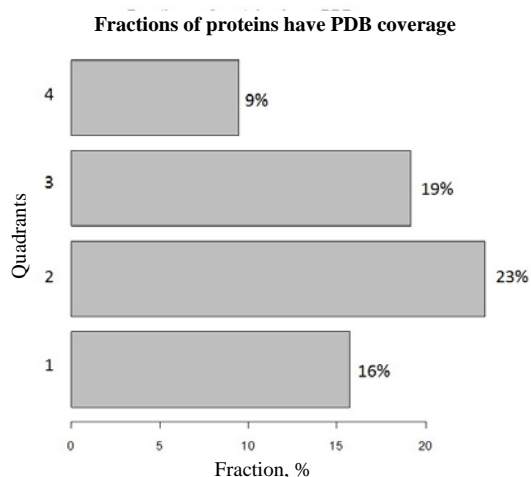


Fig.4. Fraction of protein identified with at least one PDB hit

2.3 Sequence window CH-CDF analysis

To learn more about the Q3 proteins, we dissected each protein sequence into a series of windows of 30 residues and carried out CH-CDF analysis on each segment. Table 1 summarizes results of this analysis. Proteins from the structure quadrant (Q2) have most of the windows in Q2, and the extended disorder quadrant (Q4) protein windows mostly localized in Q4. Interestingly, windows from mixed proteins (Q3) distribute mostly between quadrants Q2 Q3 and Q4, with the most hits in Q4 and slightly less hits in Q2 and Q3, suggesting that mixed proteins very likely contain a balanced distribution of ordered and disordered regions. Proteins from Q1 distribute equally in Q1 and Q2 with slightly less in Q4, again suggesting the presence of disordered regions.

Table 1. Sequence window CH-CDF analysis results

	Window quadrant localization			
	Q1	Q2	Q3	Q4
Q1 sequence windows	35%	35%	4%	26%
Q2 sequence windows	13%	68%	7%	11%
Q3 sequence windows	7%	28%	28%	37%
Q4 sequence windows	7%	13%	16%	64%

2.4 Match PDB coverage to disorder prediction

Since our analysis show that mixed proteins (in Q3) are predicted to have both disordered and ordered regions, here we attempt to verify that these predicted ordered regions are correlated with experimentally determined structures. To this end, we calculated the disorder content of the PDB covered and uncovered regions, respectively. Fig. 5 represents the results of this analysis of disorder content in all 4 quadrants.

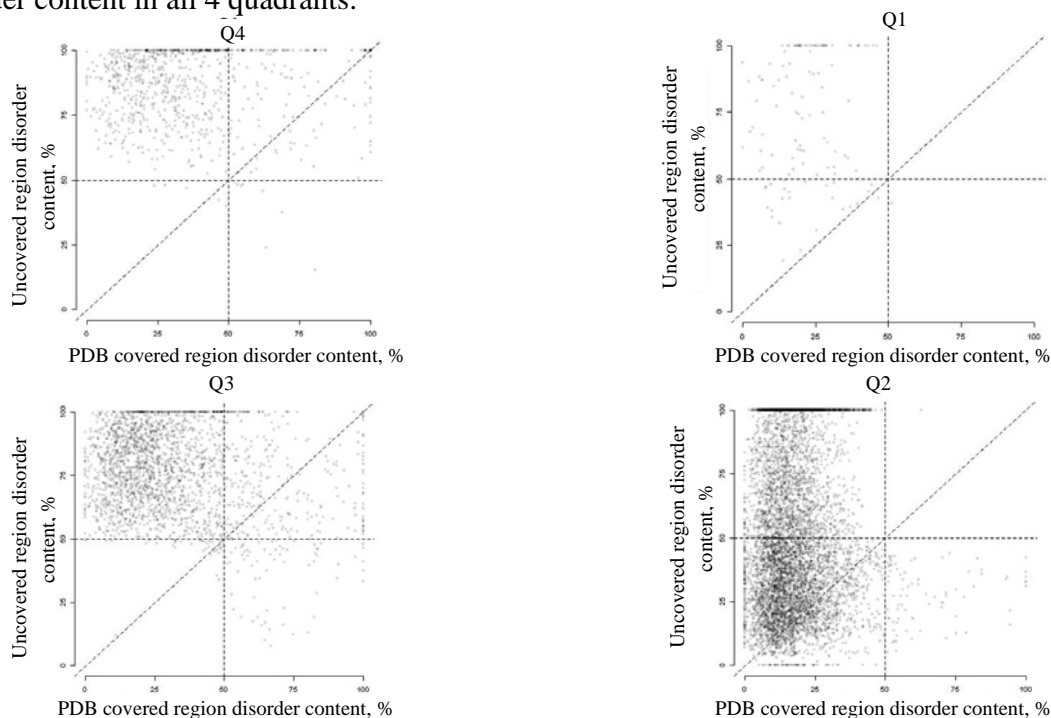


Fig.5. Percentage of disorder in PDB covered and uncovered regions

The X-axis is the disorder content of the PDB covered regions, and the Y-axis is the disorder content on the non-covered region. Disorder higher than 50% means that the corresponding region is largely predicted as disordered, whereas the score of less than 50% suggests that the region is predicted to be structured. In all 4 plots, the majority of the points clustered above the 45 degree diagonal line. We interpret this as that the regions not covered by PDB have more disorder than the covered regions.

The plot for the structure quadrant (Q2) has proteins clustered mainly to the left, both in the upper-left and lower-left corners. Those in the upper-left area could be disordered tails in these structured proteins. Those in the lower left correspond to structured regions that have not yet been crystalized. These segments are expected to be very common because many mouse proteins have multiple structured domains, and given the low percentage of PDB hits (Figure 4), it is likely that many of the structured domains of a given protein fail to make it into PDB.

In contrast to the observations for (Q2), the mixed proteins (Q3) and disordered proteins (Q4) are clustered in the upper-left corner in this plot, meaning that the regions not covered by PDB are predicted to be mostly unstructured, and those regions covered by PDB are predicted to be ordered. This indicates that disorder prediction and PDB coverage are in good agreement. Since we also showed above that mixed proteins are predicted to have both disordered and ordered regions (Table 1), it is likely that the predictions represent the true status of the protein as partially disordered and partially ordered. If this is true, it explains the mixed proteins' somewhat high fraction of PDB hits but low coverage percentages. The ordered regions are aligned to PDB sequences, but they are only a small fraction of the protein.

2.5 Function analysis for each quadrant

Previously we used a complicated prediction scheme to subdivide disordered protein regions into subtypes that we called flavors, and these different disordered flavors showed weak correlations with particular protein functions [10]. Given that the proteins in the four quadrants of the CH-CDF plots have different characteristics, it seemed reasonable to test whether these different structural subtypes exhibit functional separation. Therefore, we analyzed the proteins in each quadrant for their associations with various Gene Ontology (GO) terms.

Table 2 lists those Biological Processes GO terms found to be distinctive for each quadrant. For Q1, four of the five distinctive GO terms deal with RNA. By the CH analysis, these proteins are highly charged, and this feature may be associated with RNA association. For the Q2 structured proteins (Table 2B), most of their GO terms are related to metabolic processes and transporters. These functions are typical for structured proteins. For proteins in Q3, most of these GO terms are related to regulation or developmental pathways, including the Notch and Wnt pathways. As shown above, proteins in Q3 are likely to have both disordered and structured domains. Evidently these functions require both structured and disordered regions in the same proteins. Proteins in disorder quadrant (Q4 and table 1D) are mostly mitosis related.

Table 2. GO term analysis for four quadrants. Number of protein examples found for each GO term is listed on the right side.

Table 2A

Q1 Biological Process
tRNA methylation
tRNA wobble uridine modification
Translational termination
Positive regulation of nitric oxide
rRNA export from nucleus

Table 2B

Q2 Biological Process
Homophilic cell adhesion
Glutamine metabolic process
Phosphorylation
Sterol biosynthetic process
Peptide transport
Isoprenoid biosynthetic process
Calcium ion transport
Nucleotide metabolic process
Proteolysis involved in cellular protein catabolic process

Table 2C

Q3 Biological Process	
Regulation of transcription	Negative regulation of signal transduction
Notch signaling pathway	Pancreas development
Response to heat	Defense response to bacterium
Osteoblast differentiation	Endocytosis
Negative regulation of cell differentiation	Somitogenesis
Regulation of cell proliferation	Actin filament organization
Pituitary gland development	Wnt receptor signaling pathway
Positive regulation of neuron differentiation	Intracellular signaling cascade
Endoderm development	Epithelial cell differentiation
Organ morphogenesis	Transforming growth factor beta receptor signaling pathway

Table 2D

Q4 Biological Process
G1/S transition of mitotic cell cycle
chromosome organization
establishment of cell polarity
response to salt stress
mRNA export from nucleus

3. Discussion

3.1 Overview

IDPs may have subtypes and the different subtypes may have different functions. One previous study indicated that such disordered subtypes may exist [10]. However, the effort to subclassify IDPs such that each class has its own functional features still remains a difficult task.

Instead of relying on the training of existing data to build specific classifiers by machine learning methods, we took an alternative approach based on the hypothesis that different subtypes of IDPs should exhibit different biophysical features. Such features can be readily captured by applying two different prediction tools, CH and CDF, which use different biophysical

characteristics for their evaluation of disorder content. We therefore developed a CH-CDF plot for IDP partition.

3.2 Structural Partitioning by the CH-CDF plot

Proteins partitioned by the CH-CDF plot show a very different PDB coverage rate. The structure quadrant (Q2) has many more proteins identified with at least one PDB entry than the disorder quadrant (Q4). The mixed quadrant (Q3) has a fraction of proteins with PDB hits almost comparable to those in the structure quadrant (Q2). However, their coverage rate percentages are typically in the 20%-30% range, while the coverage in the ordered quadrant (Q2) is often as high as 90%-100%. These data suggest that mixed proteins (in Q3) have more crystal-forming ordered regions than those in the disorder quadrant (Q4).

Even though predicted to be structured, the proteins in the structure quadrant (Q2) have a significant fraction of cases with only 20% coverage (Figure 3, Q2). As indicated by the data in Table 1 and Figure 5, this result likely reflects the fact that many mouse proteins have modular structure and contain multiple structured domains. Thus, the entire protein is, overall, predicted to be structured by both the CH and CDF predictors, but if only one of the domains makes it into PDB, then such a protein could have a low coverage.

Some proteins in the mixed quadrant (Q3) and those in the disorder quadrant (Q4) have coverage percentage almost as high as 100%. After examining them individually, some of them are found to bind to ions, DNAs, RNAs, small molecules, etc. Formation of specific complexes could potentially stabilize disordered proteins, and lead to the formation of a crystallizable structure. However, there are indeed cases where such predicted disordered proteins are monomers by themselves. These need further study.

One of our early hypotheses was that proteins with relatively low net charge and high hydrophathy, e.g. predicted to be structured by CH, and yet predicted to be disordered by CDF, e.g. located in (Q3), might undergo hydrophobic collapse yet remain lacking stable structure. Such proteins would likely be native molten globules. An alternative hypothesis is that proteins in (Q3) simply contain mixtures of structured and disordered regions.

We first found that Q3 proteins have many more locally ordered sequence windows, indeed far more than the disordered quadrant (Q4), but less than the structure quadrant (Q2) (Table 1). We then showed that the amino acid sequences in proteins are predicted as mostly ordered if a PDB hit is identified for this region (Figure 5). When the sequence region is not matched with a PDB hit, it is most likely predicted to be disordered. Together these observations suggest that the quadrant (Q3) is likely to contain proteins containing relatively balanced contributions of structured and disordered regions. For this reason here we have named the proteins in this quadrant mixed rather than collapsed disorder, a description that may have appeared in previous publications [16-18]. These observations don't rule out the possibility that some of the proteins in (Q3) or even in (Q4) might be native molten globules. Further analysis and experiments are needed to identify native molten globules and determine where they fall on the CH-CDF Plot.

3.2 The rare protein quadrant (Q1)

Proteins in this quadrant are predicted to be unstructured by CH plot, but ordered by CDF. The disordered prediction from CH plot implies that a protein has high charge and is hydrophilic. Such proteins should not be predicted to be structured by CDF, so it is no wonder that the proteins in this quadrant are rare.

The density plot of PDB coverage percentage distribution for the proteins in (Q1) showed a similar pattern when compared to the structure quadrant (Q2) (fig.3). The proteins in (Q1) also have many more proteins identified with a PDB hit than those in the disorder quadrant (Q4) (fig.4). Therefore, one possibility is that these proteins are overall structured, with some high charged or hydrophilic residues, which is just the opposite of proteins in quadrant (Q3). The GO term analysis showed that 4 out of 5 of the significant GO terms are related to nucleotide processing. Further analysis shows that many of the proteins in all four quadrants including (Q1) have net positive charges rather than net negative ones. We are testing whether the positively charged proteins in (Q1) are associated with RNA binding.

3.3 Disorder subtypes and IDP functions

We tested whether the protein compartmentalization by subtypes resulted in function partition as well. For this test, we did an analysis of GO terms to determine if some terms are biased relative to others in the various quadrants (see Methods for details).

Structured proteins exhibited significant biases towards enzymatic processes and transporters. Both of these processes are well known to be associated with structured proteins [4-6]. Meanwhile, the disorder quadrant (Q4) is mainly biased towards GO terms with mitosis-related functions, which again agrees with previous observations [4-6]. On the other hand, the mixed quadrant (Q3) is highly involved in regulation pathways, which are important in development and differentiation. The recent publication on pluripotent stem cell-inducing proteins, which must be heavily involved in gene regulation, showed that these proteins are mostly localized in the mixed quadrant (Q3) [17]. Interestingly, the plant developmental proteins called GRAS straddled the structure and mixed quadrant (Q2, Q3). The N-domains of these proteins localized to the mixed and disordered quadrants (Q4, Q3), while almost all of the C-domains of these proteins localized to the structured quadrant (Q2) [18].

The flexibility provided by disordered regions could be important in such signaling events. The disordered regions could act as linkers connecting function domains. These regions could also directly bind to partners, functioning as Molecular Recognition Features (MoRFs). Such binding is usually accompanied by a disorder → order transition. Because of their flexibility, they might be able to bind to multiple partners, acting as hub proteins in signaling networks. Their flexibility is also capable of fast but short time-span binding, which also may be crucial in signaling events.

4. Methods

4.1 Protein data

The *Mus. musculus* proteome were gathered from Uniprot 15.0 [20]. A total number of 58881 sequences were obtained. Blastclust (<http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html>) with default settings were used to reduce redundancy.

4.2 PDB Coverage

PDB monomers data is downloaded from PDBe PISA. Gapped-BLAST algorithm was used to compare query sequences to PDB monomers, with default scoring matrix (BLOSUM 62). A hit was identified only when the hit region is larger than 85% of the PDB monomer sequence, and with more than 30% identity.

4.2 GO term analysis

We downloaded GO terms associated with each protein from GO Database. To reduce protein redundancy, proteins were clustered into protein families by the Blastclust program. If sequence s_i was assigned to cluster $c(s_i)$, and n_i is the total number of proteins assigned to this family, we

define a weight $w(s_i)$ for this sequence as $w(s_i) = \frac{1}{n_i}$.

Our 509,214 proteins are in association with 10,703 GO annotations. Protein sequence s_i grouped into a quadrant $k, k = 1, 2, 3, 4$ as group g_k . And for $GO_j, j = 1, 2, \dots, 10703$, there is a cluster of proteins related to GO_j , as $C_j, j = 1, 2, \dots, 10703$. Therefore, we calculate $n_{j,k}$, the number of proteins related to GO_j in quadrant k by $n_{j,k} = \sum w(s_i), s_i \in g_k \cap C_j$.

In the next step, we compare $n_{j,1}, n_{j,2}, n_{j,3}$ and $n_{j,4}$ for every specific GO term GO_j to examine if GO_j has any bias towards certain quadrant. The expected value of protein frequency in quadrant

k is calculated as $E_{j,k} = p_k \cdot \sum_{k=1}^4 n_{j,k}$, with p_k being the proportion of numbers of proteins in quadrant k . Then we compute X_j^2 , sum of expectancy, as $X_j^2 = \sum_{k=1}^4 (O_{j,k} - E_{j,k})^2 / E_{j,k}$. X_j^2 follows a chi-square distribution with 3 degrees of freedom, $X_j^2 \sim \chi^2(3)$.

Under the null hypothesis that GO_j distributed in 4 quadrants according to expectancy, we can derive p_j as a p-value for GO_j . Since multiple statistic tests are applied, we use the Bonferroni correction to adjust obtained p-value. It reduces the scale of significant results as well. A threshold of 0.05 is chosen, and GO terms with p-values lower than 0.05 are collected.

References

1. Dunker, A. K., C. J. Brown and Z. Obradovic (2002). "Identification and functions of usefully disordered proteins." *Adv Protein Chem* 62: 25-49.
2. Oldfield, C. J., Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky and A. K. Dunker (2005). "Comparing and combining predictors of mostly disordered proteins." *Biochemistry* 44(6): 1989-2000.
3. Iakoucheva, L. M., C. J. Brown, J. D. Lawson, Z. Obradovic and A. K. Dunker (2002). "Intrinsic disorder in cell-signaling and cancer-associated proteins." *J Mol Biol* 323(3): 573-84.
4. Xie, H., S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky and Z. Obradovic (2007). "Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions." *J Proteome Res* 6(5): 1882-98.
5. Vucetic, S., H. Xie, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky (2007). "Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions." *J Proteome Res* 6(5): 1899-916.
6. Xie, H., S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky (2007). "Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins." *J Proteome Res* 6(5): 1917-32.
7. Murzin, A. G., S. E. Brenner, T. Hubbard and C. Chothia (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." *J Mol Biol* 247(4): 536-40.
8. Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton (1997). "CATH--a hierarchic classification of protein domain structures." *Structure* 5(8): 1093-108.
9. Reeck, G. R., C. de Haen, D. C. Teller, R. F. Doolittle, W. M. Fitch, R. E. Dickerson, P. Chambon, A. D. McLachlan, E. Margoliash, T. H. Jukes and et al. (1987). "'Homology" in proteins and nucleic acids: a terminology muddle and a way out of it." *Cell* 50(5): 667.
10. Vucetic, S., C. J. Brown, A. K. Dunker and Z. Obradovic (2003). "Flavors of protein disorder." *Proteins* 52(4): 573-84.
11. Uversky, V. N. and A. K. Dunker "Understanding protein non-folding." *Biochim Biophys Acta* 1804(6): 1231-64.
12. Uversky, V. N., J. R. Gillespie and A. L. Fink (2000). "Why are "natively unfolded" proteins unstructured under physiologic conditions?" *Proteins* 41(3): 415-27.
13. Oldfield, C.J., Cheng, Y., Cortese, M.S., Brown, C.J., Uversky, V.N., and Dunker, A.K. Comparing and combining predictors of mostly disordered proteins. *Biochemistry*. 44:1989-2000 (2005).
14. Xue, B., C. J. Oldfield, A. K. Dunker and V. N. Uversky (2009). "CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions." *FEBS Lett* 583(9): 1469-74.
15. LeGall, T., P.R. Romero, M.S. Cortese, V.N. Uversky, and A.K. Dunker. "Intrinsic disorder in the Protein Data Bank." *J Biomol Struct Dyn* 24(4): 325-42 (2007).
16. Mohan, A., Sullivan, Jr., W.J., Radivojac, P., Dunker, A.K., and Uversky, V.N. "Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldome of early-branching eukaryotes." *Molecular Biosystems* 4: 328-340 (2008).
17. Xue, B. C.J. Oldfield, Y.Y. Van, A.K. Dunker, and V.N. Uversky. "Protein intrinsic disorder and induced pluripotent stem cells." *Mol Biosystems* (In Press)
18. Sun, Z., B. Xue, W.T. Jones, E. Rikkerink, A.K. Dunker, and V.N. Uversky. "A functionally required unfoldome from the plant kingdom: intrinsically disordered N-terminal domains of GRAS proteins are involved in molecular recognition during plant development." *Plant Mol. Biol.* (In Press).
19. Obradovic, Z., K. Peng, S. Vucetic, P. Radivojac and A. K. Dunker (2005). "Exploiting heterogeneous sequence properties improves prediction of protein disorder." *Proteins* 61 Suppl 7: 176-82.
20. Magrane, M. and U. Consortium "UniProt Knowledgebase: a hub of integrated protein data." *Database (Oxford)* 2011: bar009.