OPTIMIZING HYDROPATHY SCALE TO IMPROVE IDP PREDICTION AND CHARACTERIZING IDPS' FUNCTIONS

Fei Huang

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biochemistry and Molecular Biology,
Indiana University

January 2014

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

A. Keith Dunker, Ph.D., Chair

Jake Chen, Ph.D.

Doctoral Committee

November 19, 2013

Thomas D. Hurley, Ph.D.

Li Shen, Ph.D.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to Dr. A. Keith Dunker, for his mentorship and support through my graduate study. Dr. Dunker is a great scientist with knowledge and character. What I have learnt from Dr. Dunker will have a great influence on my future career.

I also would like to thank Dr. Vladimir Uversky. Dr. Uversky gave me valuable advices and guidance for all my research. I am indebted to Dr. Uversky for all the great ideas he came up for my research.

Thank my committee members, Dr. Jake Chen, Dr. Thomas D. Hurley, and Dr. Li Shen. Also thank my former committee members, Dr. Yaoqi Zhou and Dr. Pedro Romero, who have relocated to other Universities. Thank you for the comments and critics, many of which have turned into valuable data for publications.

Thank Chris Oldfield, for all the assistance and discussions on the research, and about life outside.

Thank all the lab members, Bin, Bo, Caron, Eshel, Jingwei, Maya, and Wei-lun. I appreciate your help and support inside and outside of the lab.

Fei Huang

Optimizing Hydropathy Scale to Improve IDP Prediction and Characterizing IDPs' Functions in PDB Dimers

Intrinsically disordered proteins (IDPs) are flexible proteins without defined 3D structures. Studies show that IDPs are abundant in nature and actively involved in numerous biological processes. Two crucial subjects in the study of IDPs lie in analyzing IDPs' functions and identifying them. We thus carried out three projects to better understand IDPs.

In the 1st project, we propose a method that separates IDPs into different function groups. We used the approach of CH-CDF plot, which is based the combined use of two predictors and subclassifies proteins into 4 groups: structured, mixed, disordered, and rare. Studies show different structural biases for each group. The mixed class has more order-promoting residues and more ordered regions than the disordered class. In addition, 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.

The 2nd project is about identifying if an unknown protein is entirely disordered. One of the earliest predictors for this purpose, the charge-hydropathy plot (C-H plot), exploited the charge and hydropathy features of the protein. Not only is this algorithm simple yet powerful, its input parameters, charge and hydropathy, are informative and readily interpretable. We found that using different hydropathy scales significantly

affects the prediction accuracy. Therefore, we sought to identify a new hydropathy scale that optimizes the prediction. This new scale achieves an accuracy of 91%, a significant improvement over the original 79%.

In our 3rd project, we developed a per-residue C-H IDP predictor, in which three hydropathy scales are optimized individually. This is to account for the amino acid composition differences in three regions of a protein sequence (N, C terminus and internal). We then combined them into a single per-residue predictor that achieves an accuracy of 74% for per-residue predictions for proteins containing long IDP regions.

A. Keith Dunker, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

IDP	Intrinsically Disordered Protein
PDB	Protein Data Bank
NMR	Nuclear magnetic resonance
CD	Circular Dichroism
PONDR	Predictors of Natural Disordered Regions
VSL2	Variously Characterized Short and Long Regions, Version 2
PSSM	Position-Specific Scoring Matrix
CASP	Critical Assessment of Protein Structure Prediction
PPI	Protein-Protein Interactions
MoRF	Molecular Recognition Feature
HOX	Homeotic
Exd	Extradentical
MCC	Matthews Correlation Coefficient
NPV	Negative Predictive Value
ROC	Receiver Operating Characteristic
CH	Charge-Hydropathy
CDF	Cumulative Distribution Function
AAindex	Amino Acid Index Database
SVM	Support Vector Machine
Trx	Thioredoxin
PPV	Positive Predictive Value

AUC	Area Under Curve

I. INTRODUCTION

1. Definition of Intrinsically Disordered Proteins (IDPs)

Intrinsically **D**isordered **P**roteins (**IDP**s) are proteins without defined 3D structures[1–5]. 'Disorder' describes the lack of structure of such proteins. The term 'intrinsically' means that their failure to fold into a specific 3D structure is inherent in the amino acid sequence[6,7]; i.e., encoded by their amino acid sequences. It is important to note that the amino acid composition of IDPs is biased towards certain amino acids, compared to the compositions of ordered proteins[4].

Many other expressions are used to describe these proteins as well. Commonly known expressions are, floppy, pliable, rheomorphic, flexible, mobile, partially folded, natively denatured, natively unfolded, natively disordered, intrinsically unfolded, vulnerable, chameleon, malleable, 4D, protein clouds, dancing proteins, proteins waiting for partners, and others[6]. These terms were proposed by many researchers when they encountered IDPs during their research, and found that these proteins exhibit features distinguishable from those of globular proteins.

Despite the differences in the use of words, almost all of these names suggest a common feature, flexibility. However, flexibility is not stringent enough to be an appropriate descriptor for these proteins. Many structured proteins, such as the binding pocket of an enzyme, may have some degrees of flexibility but they are not IDPs[6]. Protein ensembles with no preferred lowest energy conformation that adopt many different forms, is a more accurate description[8]. Thus, we use the term 'disorder' to define this state of these proteins.

Even though not illustrated in all of the terms, some researchers used names that included words such as ‘natively’ or ‘naturally’. The word ‘intrinsically’ was chosen in the end because it illustrates the two most important features of these proteins equally well: 1. disorder is inherently encoded in their amino acid sequences; 2. the state of disorder is manifested under generic physiological conditions[6].

In contrast, ‘ordered’ or ‘structured’ proteins indicate proteins that have a preferred lowest energy conformation, i.e., a well-defined 3D structure. Therefore, the term ‘disordered’ and ‘ordered or structured’ are adopted throughout this manuscript to indicate proteins that are not disordered. It is worth noting that many proteins possess both states, with each state in different regions of the sequence. We refer to these proteins as ‘partially disordered or partially ordered proteins’, and we name these regions as ‘disordered or ordered regions’.

2. Challenging the protein structure-function dogma

2.1 protein structure-function dogma

The study of IDPs has been of growing interest in the recent years[9,10]. However, it is not always this case from the beginning. For a long time, the idea that a protein can carry out function without folding into 3D structure was not recognized by the scientific community[9].

The proteins structure-function dogma has been dominant for a long time. In 1894, Emil Fischer made the astonishing discovery of enzyme specificity. He therefore proposed the ‘lock and key’ model to explain such strong specificity[11]. In 1931, Hsien Wu proposed that loss of function via protein denaturation occurred when weak

interactions were disrupted leading to loss of structure[12]. This paradigm, that structure is necessary for function, became dominant following the very large numbers of protein structures determined by X-ray crystallography[13]. To date, 94813 protein structures haven been determined and deposited in the Protein Data Bank (PDB)[14].

2.2 Different voices

During that time, however, there were examples suggesting that this dogma was not perfect. In 1953, scientists made the observation that milk protein casein is likely to be unstructured and this might help infants' digestion[15,16]. As reviewed by Sigler[17], in the 1980s several researchers found that eukaryotic transcription factors have large regions of highly unusual sequences with high content of acidic residues, termed as 'acid blobs' or 'negative noodles'. These regions fail to fold into 3D structures, yet they carry out gene regulation.

However, these examples are sparse and failed to draw wide attention. They were merely considered to be rare exceptions to the dogma. The traditional view of protein structure-function dogma still held its place until the late 1990s, when nuclear magnetic resonance (NMR) and bioinformatics study of IDPs blossomed[18–20].

2.3 NMR study reveals large disordered regions

Unlike X-ray crystallization, NMR method to study protein conformation does not require crystallization. NMR thus is more suitable to study IDPs[20]. Because of IDPs' lack of stable conformation, their NMR spectra yield multiple very different structural possibilities. Human cell cycle control protein p21 was shown by NMR to be entirely

disordered, and its regulatory functions were shown to depend on its disordered state[21]. So far, many proteins are confirmed to be disordered or having a long disordered region, where "long" is generally taken to be ≥ 30 residues [20,22–32]. Besides NMR, missing sequences in X-ray structure[19,33–35], Raman optical activity[36], Circular Dichroism (CD)[37–40], and protease sensitivity experiments[41] can also identify IDPs. Since each method has uncertainties with regard to IDP characterization, it would be an advantage to use multiple methods for IDP characterization in which case the different methods complement each other[42].

2.4 Computational method reveals amino acid composition bias for IDPs

When more and more experimental data about disordered protein accumulated, people started to ask the question why IDPs do not fold. Bioinformatics study of IDPs' composition reveals that IDPs have their own preference of amino acid residues compared to ordered proteins[4,43–46]. In general, hydrophilic or polar residues are disorder promoting. However, note that additional factors can also influence the disorder/order state of a protein or region as well, such as net charge, aromatic content, side-chain bulkiness, etc.

The folding of an ordered protein is usually driven by the hydrophobic amino acid residues, such as valine, isoleucine, leucine, phenylalanine, methionine, and tryptophan, to form a hydrophobic core. In contrast, a disordered protein usually maintains its disordered state as a result of its high content of polar residues – i.e., arginine, glutamine, serine, glutamate, and lysine – all of which readily interact with water molecules. Note that cysteine, despite being a polar residue, is missing from the above list. This is because

when oxidized, cysteines greatly promote protein conformation stability by forming disulfide bonds[47,48]. Cysteine also readily binds prosthetic groups, and thus stabilizes protein structure. Note that even though lysine could also bind to prosthetic groups, it is still a disorder promoting residue because of the charge it carries. Another interesting case here is proline. Although non-polar, proline is a very potent disorder promoting residue. With its unusual secondary amine or imine structure in which the N-H group is replaced with an N-C bond, proline lacks the N-H group critical to alpha-helix or beta-sheet formation and thus is a 'structure breaker' that often flanks alpha-helices and beta-sheets[49]. Overall, compared to strcutred proteins, IDPs are significantly enriched in P, E, S, Q and K.

3. IDP predictors

Many experimental methods are available nowadays to characterize IDPs. However, the number of determined protein sequences is so large that applying experimental method to every one of them is unrealistic. Thanks to the advances in computational biology methods, we can predict the disorder state of a protein or a protein region with fair accuracy by means of supervised learning algorithms.

In particular, given a set of labeled training data, supervised learning algorithm identifies a 'pattern' and applies this 'pattern' to new data to infer its label[50]. Given the amino acid compositional differences between structured proteins and IDPs, we can develop algorithms that predict the disordered or ordered state of a protein or region by using amino acid sequences as inputs[4,43,51].

3.1 Early predictors and C-H plot method

Currently, many IDP prediction algorithms have been developed[4,44,51–55]. The first IDP predictor was developed by R.J.P Williams in 1979[56]. Observing that IDPs having abnormally high ratio of number of charged residues divided by the number of hydrophobic residues, his algorithm separated two IDPs from a set of ordered proteins. The two proteins he studied were from the ribosome and so had a very high net charge. In this case, the high number of charged residues provides more repulsion among amino acids, and thus makes the protein less likely to fold in the absence of RNA. Furthermore, the low number of hydrophobic residues renders less driving force to fold into a hydrophobic core. However, this method developed in 1978 does not work well for proteins that have a large number of charged groups having a nearly equal balance of positive and negative charges. In this case, the large Williams ratio predicts disorder for many well folded proteins and thus has a very poor acccuracy[9].

In 2000, without knowledge of the prior work by Williams, Uversky et al used normalized net charge, not the total charge, and normalized hydrophobicity calculated from Kyte-Doolittle scale[4,57] to classify proteins as structured or as natively unfolded. Applying this method to a large number of proteins, including 91 IDPs and 275 ordered proteins gave excellent results[4]. In 2004, Sussman et al transformed this method to FoldIndex, a per-residue predictor that can be applied to predict disorder on a local region of protein[58]. However, FoldIndex did not re-train the data to obtain a new separating function, and the per-residue accuracy of FoldIndex was not evaluated. In fact, subsequent study shows that the amino acid composition of local disordered regions from partially disordered proteins different significantly from entirely disordered proteins[54,55].

As we will show herein, a simple adoption of the original linear function trained with fully disordered proteins performs poorly on partially disordered proteins.

3.2 IUPred

Another biophysical feature based predictor is IUPred, developed by Dosztányi et al. IUPred partitions disordered/ordered proteins by estimating their folding energies[52]. Without the knowledge of a protein's 3D structure, IUPred estimates the per-residue energy by assuming that the energy contribution of that residue depends on its amino acid type and potential partners in that sequence. They tried various pairwise sidechain interaction energy matrices to find the one that gave the clearest difference between the folding energies of structured and disordered proteins estimated by this method. Since disordered proteins and ordered proteins gives different folding energy estimates by this method, IUPred can predict IDPs from the approximated per-residue folding energy.

3.3 PONDR predictors

As discussed above, charge and hydropathy are not the only features having an impact on the disorder/order state of a protein. Other features, including aromatic content, sequence complexity and many others also determine if a protein/region folds or not[4,53–55,59]. In addition, a per-residue predictor that predicts local region disorder is more informative to infer the function of a protein.

PONDRs (predictors of natural disordered regions) are designed as per-residue predictors with a combination of amino acid sequence features[54,59,60]. The second generation of PONDR, namely VLXT, was a merged neural network of three sub-

predictors[61]. One predictor was trained on **v**ariously characterized, **l**ong regions of internal disorder (VL), and the other was trained on **X**-ray characterized, disordered N or C **t**ermini (XT). This predictor achieved a balanced accuracy above 70%, a value significantly better than the 50% expected by chance for two state prediction on balanced data[61].

Later, amino acid composition study on short (<30 residues) and long (>30 residues) disordered regions shows that short and long disordered regions have significantly different amino acid biases[55]. VSL2 (variously characterized short and long regions, version 2) was developed to better cope with this difference[54]. Basically, VSL2 is a meta-predictor of two sub-predictors, one trained with short and the other trained with long disordered regions data. This predictor used evolutionary information obtained by generating the position-specific scoring matrix (PSSM) of the query sequence by PSI-BLAST (Position-Specific Iterated BLAST)[62]. However, PSI-BLAST is time consuming. As a result, VSL2B, a much faster 'light' version without employing PSSM, is commonly used and achieves an accuracy that is only slightly diminished[54].

In 2010, a consensus predictor, PONDER-FIT, was developed[59]. PONDR-FIT is a neural network meta-predictor based on the outputs of 6 commonly used disorder predictors, PONDR-VSL2[54], PONDR-VL3[63], FoldIndex[58], IUPred[52], and TopIDP[49]. The outputs of these 6 predictors are then combined into a single predictor using a neural network. Compared to the best of its 6 individual predictors, PONDR-FIT shows an overall 11% improvement.

3.4 CASP (Critical Assessment of Protein Structure Prediction)

In 2002 CASP5 included disorder predictions with published evaluations, and links to several of these predictors are now available in the DisProt Database (http://www.DisProt.org)[64–67]. Note that CASP targets are results of up-to-date experiments and selected so that no prior information about them is revealed to any participants. The biannual CASP experiments serves as excellent, unbiased benchmarking for various IDP predictors.

However, the CASP disorder prediction experiment provides very unbalanced datasets for which, the number of ordered residues overwhelms disordered residues[66,68]. Also, the disorder predictions accuracies from recent CASP fluctuate, showing no significant improvement. These fluctuations are likely due to changes in the difficulties presented by the different targets, rather than fluctuations in predictor accuracies[66–68].

4. IDP abundance in nature

One application of IDP prediction is to evaluate the abundance of IDP in nature. IDP predictors haven been applied to proteomics dataset of various species[35,69–71]. In general, disorder is estimated to be much more prevalent in sequence databases and in various proteomes as compared to PDB. Eukaryotes protein sequences contain much more disorder (~33% - 50%) than prokaryotes (~15% - ~30%) and archaea (~12% - ~24%). Such predictions suggest that the human proteome, in particular, contains ~35% - 50% disordered residues[70,71]. An important, open question is whether the predicted regions of disorder remain disordered inside the cell or become structured upon association with partners or ligands.

5. IDP functions

Given such abundance, one cannot help but wonder what are the functions of IDPs? Also, why do eukaryotes have much more predicted disorder than common bacteria or archaea bacteria?

5.1 PPI (Protein-Protein Interactions)

One of the signature function of disordered protein is carried out through Molecular Recognition Features (MoRFs)[72–76]. These MoRFs are typically hydrophobic patches within a disordered region. Upon binding, the hydrophobic groups become buried and the IDP segment typically undergoes a disorder-order transitions. In some cases, significant sized parts of the original IDP region remain unstructured yet contribute to the binding affinity even while remaining unstructured. Interactions involving IDP regions that contribute to binding affinity have been called 'fuzzy complexes'[77,78].

Eukaryotic PPI networks are scale-free, a term that means a plot of the log of the number of nodes versus the log of the number of partners per node gives a straight line. Such a plot means that a few proteins in the network interact with many partners while most proteins interact with only a few partners[79–81]. Such proteins with high number of binding partners are referred to as 'hub proteins'. An often-cited analogous network is that provided by the set of airline routes, in which some big cities such as New York or Los Angles contain connecting airports for smaller cities.

It is intriguing to imagine how single proteins can bind to many partners. We suggested that the flexibility of IDPs might provide the key feature that enables one protein to bind to many partners, and we found that a number of hub proteins indeed used

disordered regions to bind to many partners[2,21,82,83]. Two IDP-based mechanisms were observed for the multipartner binding of hub proteins. In the first mechanism, one disordered region binds to many partners, or 'one-to-many binding'. In the second mechanism, many disordered regions bind to one partner or 'many-to-one binding'[2,21,82,83].

The one-to-many binding IDP regions of a hub protein are either in very close vicinity or at the exact same site[83]. When the same IDP region binds to multiple partners, this IDP region often changes its shape to fit with its different partners. For example, the N-terminal and C-terminal IDP regions of p53 each bind to more than 40 different partners by this mechanism.

Alternatively, many different disordered proteins/regions of different amino acid sequences can bind to a single ordered partner[84]. Many such examples are found in PDB, in which a single IDP contains two or even three separate MoRF regions binding to the same structured partner.

5.2 IDPs in alternative splicing

Confirmed by both experimental data and computational predicted data, alternatively spliced protein isoforms are enriched of disordered residues[85–87]. During the process of many tissue-specific alternative splicing, many MoRF containing IDP region could be either kept or spliced out, resulting in tissue-specific protein-protein interactions based on differential retention of IDP binding regions or MoRFs.

5.3 IDPs in transcription factors

Eukaryotic transcription factors have a high content of disordered residues[88,89]. The activation domains of transcription factors are often composed of IDPs. Also, structured functional domains in transcription factors are frequently flanked by disordered residues that regulate their binding to DNA.

As an example, HOX (homeotic) domain of Drosphila regulatory transcription factor Ubx is flanked by evolutionarily conserved disordered regions[90,91]. Alone without the IDP region, HOX domain binds DNA with much higher affinity. In addition, an IDP linked YPWM motif can further weakens HOX binding. To enhance its HOX domain binding affinity, Ubx interacts with Extradentical (Exd). Upon interaction, the IDP linked YPWM motif binds into a pocket of the HOX domain on Exd. Furthermore, the IDP linker between HOX domain and the YPWM motif can be alternatively spliced for different length. Therefore, we speculate that, by regulating the length of the IDP linker and the availability of surrounding Exd's binding sites, the binding affinity of Ubx's HOX domain can be either promoted or repressed.

Besides the examples discussed above, IDPs are involved in many more biological processes, such as protein-RNA interactions[92–96], allosteric regulations[97–103], chaperone functions[104–107], protein evolution[108,109], and so on. Through their dynamic features, IDPs participate in numerous biological functions through all kinds of mechanisms. It is by the dual existence of disorder and ordered state in proteins, proteins can perform wide diversity of functions and become the building blocks of life.

6. Study goals

IDPs are abundant in nature, and they play important roles in many biological processes. Understanding their functions and identifying them are the two major tasks in the study of IDPs. Here, we present two approaches, CH-CDF plot and improved C-H plot method to address these two tasks, respectively.

The CH-CDF plot method is a clustering tool that partitions IDPs into three subgroups. This partition method clusters IDPs according to their biophysical characters. Indeed, each IDP subgroup is shown to have distinguishable function features related to the biophysical features of that group.

To address the 2nd task, we improved the original C-H plot disorder predictor developed by Uversky et al. We optimized the hydropathy scale used to calculate the hydropathy in the original method. The prediction power of C-H plot is significantly improved by the newly developed scales. In addition, this scale is highly correlated with popular hydropathy scales, and thus can be used to calculate hydropathy for better understanding of protein functions.

II. MATERIALS AND METHODS

1. Subclassifying disordered protein by the CH-CDF plot method

1.1 Protein data

The *Mus. musculus* proteome were gathered from Uniprot 15.0[55]. A total number of 58881 sequences were obtained. Blastclust (http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/ blast lab.html) with default settings was used to reduce redundancy.

1.2 PDB Coverage

PDB monomers data is downloaded from PDBe PISA. The gapped-BLAST algorithm was used to compare query sequences to PDB monomers, with the 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.

1.3 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 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, $j = 1, 2, \ldots, 10723$ there is a cluster of proteins related to GO_j, as C_j, $j = 1, 2, \ldots, 10723$. 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 a 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 statistical tests are applied, we use Bonferroni correction to adjust obtained p-value. This correction reduces the scale of significant results as well. A threshold of 0.05 is chosen, and GO terms with p-values less than 0.05 are collected.

2. Optimizing IDP-Hydropathy Scale for Disorder Prediction

2.1 Dataset

Two sets of proteins were used in this study[59,60]: experimentally verified entirely disordered proteins and experimentally verified completely structured or ordered proteins. Entirely disordered proteins were taken from Disprot 6.0[64,42]. These proteins were filtered such that only those proteins with their entire sequences being disordered were retained. Our fully disordered protein dataset contains 109 disordered sequences with 22,614 amino acid residues. The set of fully structured (ordered) proteins consisting only of

single-chain and non-membrane proteins was assembled from the Protein Data Bank (PDB)[14] (http://www.rcsb.org/pdb/). Only structures determined by X-ray crystallography and characterized by unit cells with primitive space groups were kept in our dataset. Structures with ligands, disulfide bonds, or missing residues were also removed. Then a BLASTCLUST[62] analysis was performed to cluster proteins into subsets, with all members of each subset having at least 25% sequence identity with another subset member and having less than 25% sequence identity with any member of any other subset. The longest sequence in each cluster was selected to construct the fully ordered protein set. This set of experimentally determined structured proteins contains 563 fully structured protein sequences with 113,895 amino acid residues.

2.2 Benchmarking scales

We obtained 19 hydropathy scales[57,110–127] from ExPASy-ProtScale to compare with the hydropathy scale of Kyte & Doolittle (1982)[57]. A more thorough benchmarking against various amino acid indices was carried out later with 535 amino acid scales. Among these, 531 amino acid scales were obtained from the Amino Acid index database (AAindex: www.genome.jp/aaindex)[128–130]. Another 4 disorder propensity scales were also examined, including TOP-IDP[49], FoldUnfold[131], B-value[132], and DisProt[49,64,42,133].

2.3 Dealing with unbalanced data

2.3.1 Assessment metrics

Our dataset of disordered/structured proteins is highly imbalanced with 16.2% disordered and 83.8% structured based on numbers of chains or 17% disordered and 83%

structured based on numbers of amino acid residues. Accuracy, defined as the proportion of correctly classified samples in the population (Eq. 1), is not a good measurement when the number of one class dominates[134]. In fact, simply predicting every case as structured would yield an apparent accuracy close to 0.84. A better approach is to average the correct prediction of order and the correct prediction of disorder, called the balanced accuracy and calculated as follows: first, estimate the value for the correct prediction of disorder, called sensitivity (Eq. 2), and the value for the correct prediction of structure, called specificity (Eq. 3), then average the values for sensitivity and specificity[134] (Eq. 4):

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}$$ (Equation 1),

where Acc = accuracy, TP = true positive predictions, TN = true negative predictions, FP = false positive predictions, and FN = false negative predictions,

$$Sensitivity\ (Recall) = \frac{TP}{TP+FN}$$ (Equation 2),

$$Specificity = \frac{TN}{TN+FP}$$ (Equation 3),

$$Balanced\ Acc = \frac{Sensitivity+Specficity}{2}$$ (Equation 4).

The usefulness of the balanced accuracy metric is undermined by the high fraction of structured residues in the training set. That is, predicting more disordered residues

rewards sensitivity much more than the penalty in specificity, so this imbalance encourages overpredicting disorder[66,68,134]. To further help with the analysis of prediction on imbalanced data, the positive predictive value (PPV) metric was introduced[135–137]. PPV, also called "precision", is calculated as the fraction of correctly predicted disorder versus all the predicted disorder (Eq. 5):

$$PPV\ (Precision) = \frac{TP}{TP+FP} \quad \text{(Equation 5).}$$

Overpredicting disorder will result in low PPV, whereas a high PPV value indicates that a high proportion of the predicted disorder is indeed actual disorder. Combing PPV with sensitivity (also known as recall) as indicated (Eq.6) yields the F-score, which is an effective representation of the predictive power in imbalanced dataset[138]:

$$F = 2 \cdot \frac{precision \cdot recall}{precision+recall} \quad \text{(Equation 6).}$$

The F-score values range from 0 to 1, and because of the product of precision and sensitivity in the numerator, a high F-score usually means a high score for both PPV and sensitivity, or recall.

The Matthews correlation coefficient (MCC) is another very commonly used and effective metric for imbalanced datasets[66,139] (Eq. 7):

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{(Equation 7).}$$

The MCC has been observed to be highly correlated with the F-score for disorder prediction in Critical Assessment of protein Structure Prediction 9 (CASP9)[66].

In contrast to PPV, a negative predictive value (NPV) measures the correctly predicted structured proteins over all of the predicted structured proteins[135] (Eq. 8):

$$NPV = \frac{TN}{TN+FN} \quad \text{(Equation 8).}$$

A Receiver Operating Characteristic (ROC) curve is a plot of sensitivity versus specificity[140]. The area under the curve (AUC) is another often used metric for judging predictive power of an algorithm.

Given all of the above, we estimated F-score, MCC, sensitivity, specificity, AUC, PPV, and NPV as the metrics to assess the quality of the predictions that were made on the unbalanced dataset used herein. Sensitivity, specificity and AUC are informative about the correctly predicted disorder and structure of one class. PPV and NPV reveal whether the algorithm is overpredicting disorder or structure. In the end, the F-score and MCC give an overall estimate of the quality of the predictions.

2.3.2 Training method

In the current dataset, disordered proteins are outnumbered and under-represented. To develop a good predictor in the scenario of unbalanced dataset, we tried several popular methods[134]. Both under-sampling structured proteins, and oversampling disordered proteins[141–143] were implemented separately to achieve a balanced

disorder/order dataset. Synthesizing new data for the disordered class were also carried out to obtain more disordered samples[144,145]. We found that in this study, all of these methods gave similar results. The approach of adding weights to the SVM cost function[134,146,147] so that a greater penalty occurs when a disordered protein is misclassified, achieves results similar to the sampling methods above while being much simpler to implement compared to under- or oversampling. Therefore, for simplicity, here we only used the approach of using a weighted cost function.

2.4 Correlation study

The absolute value of Pearson product-moment correlation coefficient[148], r, was calculated between IDP-Hydropathy scale and each of the 513 scales from AAIndex. For each *scale* from AAIndex, the correlation of it with IDP-Hydropathy scale is calculated as in Equation 9, where IDP_i is the score for *i*th amino acid in IDP-Hydropathy scale, $Scale_i$ is the score for *i*th amino acid in that AAIndex. $\overline{IDP}$ and $\overline{Scale}$ stands for the mean value of the two scales:

$$r = \frac{\sum_{i=1}^{20}(IDP_i - \overline{IDP})(Scale_i - \overline{Scale})}{\sqrt{\sum_{i=1}^{20}(IDP_i - \overline{IDP})^2} \cdot \sqrt{\sum_{i=1}^{20}(Scale_i - \overline{Scale})^2}} \text{ (Equation 9).}$$

2.5 Benchmarking

The IDP-Hydropathy scale was derived from windows of proteins. Since entire protein sequences are applied to the original C-H plot by Uversky et al, for consistency, the benchmarking of IDP-Hydropathy scale and other 554 scales was carried out over the entire protein sequences. The normalized composition and net charge were calculated as

before. Then we obtained the 'hydropathy score' for each protein by multiplying the composition matrix and the column vector of the scale. Therefore, 2 attributes are calculated for each amino acid sequences, the 'hydropathy score' and the net charge. A linear SVM classifier was then applied to predict disorder/structure proteins.

2.6 Charge-Hydropathy plots

C-H plots were generated using our dataset with the following scales: IDP-Hydropathy, the Guy scale[110], and the Kyte-Doolitte (1982) scale[57]. The normalized net charge was calculated as previously: the absolute value of [(Arginine + Lysine) – (Glutamate + Aspartate)]/Protein Length. Then the normalized hydropathy was calculated using the indicated scales. Note that to be consistent with the original C-H plot[4], the various hydropathy scales were renormalized so as to cover the range between 0 and +1 rather than –1 to +1 as we use elsewhere herein. The linear SVM method implemented by LIBLINEAR library[81] was then applied to calculate the boundary in MATLAB (MATLAB 2012a. Natick, Massachusetts: The MathWorks Inc., 2012).

2.7 Heat map

To provide a visual comparison of variations in the different scales, a heat map was used. The heat map of IDP-Hydropathy and 9 other benchmark scales are drawn by MATLAB HeatMap function (MATLAB 2012a). The scales visualized within the heat map are all normalized to be within –1 and +1. Some of the scales are negatively correlated, so we multiplied them by –1.

3. Per-residue Charge-Hydropathy Disorder Prediction

3.1 Dataset

The data used to train the predictor is downloaded from Protein Data Bank (PDB) and Disprot. PDB data are filtered to discard structures determined by methods other than X-ray structure. PDB data with resolutions lower than 2.5, or structure sequence less than 40 amino acid residues are also discarded. Then, the PDB data and Disprot data are integrated and clustered by BLASTCLUST with 30% identity. The residues in PDB dataset with missing coordinates and residues in Disprot annotated as disordered are considered as disordered amino acid residues. All others are considered as ordered residues. Note that all sequences are extracted as they are in each database. Expression tags such as His-tags and initiating methionine(s) are included.

To pick the sequence from each cluster family, we used the following prioritized criteria as Zhang et al 2012: 1) the sequence with the largest number of disordered residues; 2) sequence with the fewest number of disordered regions (to obtain more contiguous regions of disorder); and 3) protein sequence with the largest length. In the end, we obtained 1619 protein sequences.

As shown in Peng et al 2006, the amino acid composition of long disordered regions (>30 consecutive disordered amino acids) is significantly different from the composition of short disordered regions (<=30). Therefore, our dataset is filtered for long regions.

A blind dataset was then constructed for benchmarking with other predictors. It includes sequences in PDB from 01/01/2012 till now, and a random 100 proteins from

Disprot. The training dataset excludes these sequences. However, the final predictor is constructed with all protein sequences to have better coverage of known examples.

3.2 Disorder distribution along the sequence

Number of disordered residues is calculated at the N/C terminus and internal regions. Sequences with the internal region shorter than its N/C terminus are discarded. The number of disordered residues at different regions is then normalized by the length of each region. ANOVA analysis is performed by the Matlab ANOVA function.

3.3 Correlation of amino acid composition at different regions

The disordered residue composition at each N/C terminus and internal regions are calculated. Then we used the Matlab function to calculate correlation coefficient. The plot for the differences in composition of these three regions is also generated from Matlab. The internal region composition was used as the baseline to subtract from the composition at N/C terminus.

3.4 Training via linear SVM

The N, C and internal regions are optimized individually with the similar procedures. We used LIBLINEAR to optimize the hydropathy scale to calculate local hydropathy for best disorder prediction along with local charge. Specifically, we used a window of 21 amino acid residues within the target amino acid to calculate the amino acid composition of that local region. 21 residues were chosen because it is long enough to contain a typical protein domain. For target amino acid on the N and C terminus, we

used a spacer character to if the window extends outside the sequence. Then they are applied to linear SVM with 10 fold cross-validation to obtain three coefficient vectors that maximizes the prediction power for N, C and internal regions. The 20 coefficients in the vector corresponding to 20 amino acid compositions are thus normalized to serve as the optimized hydropathy scale. The coefficients for the charge, slope and spacer character for N/C terminus are normalized accordingly to comprise the final predictor of hydropathy and charge.

III. RESULTS

1. Subclassifying disordered protein by the CH-CDF plot method

Unlike structured proteins folding into compact structures, intrinsically disordered proteins (IDPs) exist as flexible ensembles under normal physiological condition[1]. As indicated by bioinformatics studies, IDPs are very common in nature. They comprise approximately 25% to 30% of eukaryotic proteomes[149]. Over 50% of eukaryotic proteins and 70% of signaling proteins have long disordered regions[150]. 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[151–153].

Studies of ordered proteins indicate that homologues have a conserved 3D structure[154–156]. Thus, structure similarity is used as important criterion while examining a protein cluster. Most proteins with similar structure have a common evolutionary origin, and as a consequence their functions are typically closely related[154–156]. Databases such as SCOP[154] and CATH[155] have been constructed using this line of reasoning. These databases serve as a great 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[154,155].

Since IDPs lack 3D structure, 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[157]. Here our goal is to re-explore the overall idea of partitioning disordered proteins into subtypes, but using a different

predictive approach than the one we used previously. The previous approach used residue-by-residue order / disorder predictions over IDP regions of proteins[157], but a weakness of that approach was that the disordered regions varied markedly in length, which greatly complicated the interpretation.

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-hydropathy (CH) plot[4,149] and the cumulative distribution function (CDF)[60]. 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)[158]. 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 differing characteristics for each tendency. Thus, overall, the CH-CDF plot separates proteins into 4 groups with differing order and disorder tendencies.

The CH-CDF plot was previously used to compare the structure-disorder tendencies of the proteomes for several species within the phylum Apicomplexa, which include Plasmodia, Trypanosomes, and Giardia[159]. The CH-CDF plot has also been used to classify the transcription factors associated with the induction of pluripotent stem cells[160]. In both cases, the distributions of the various proteins among the four outcomes provide overviews of similarities and differences between the different sets of

proteins[159,160]. 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 differences in function for these proteins.

1.1 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 line of separation was developed from a large number of training set proteins[4,149]. Note that disordered proteins have a higher net charge and lower hydropathy compared to ordered proteins. We use the vertical distances to the separation line as the Y-coordinate of the CH-CDF plot, so when Y is positive, the protein is indicated to be disordered. Figure 1B shows the PONDR VSL2[161] plots for the same pair of disordered (red) and ordered (blue) proteins. In Figure 1C, the data in 1B are plotted to give the CDF, 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[60]. 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. Included in this plot are the descriptions of the prediction characteristics for the proteins in each quadrant.

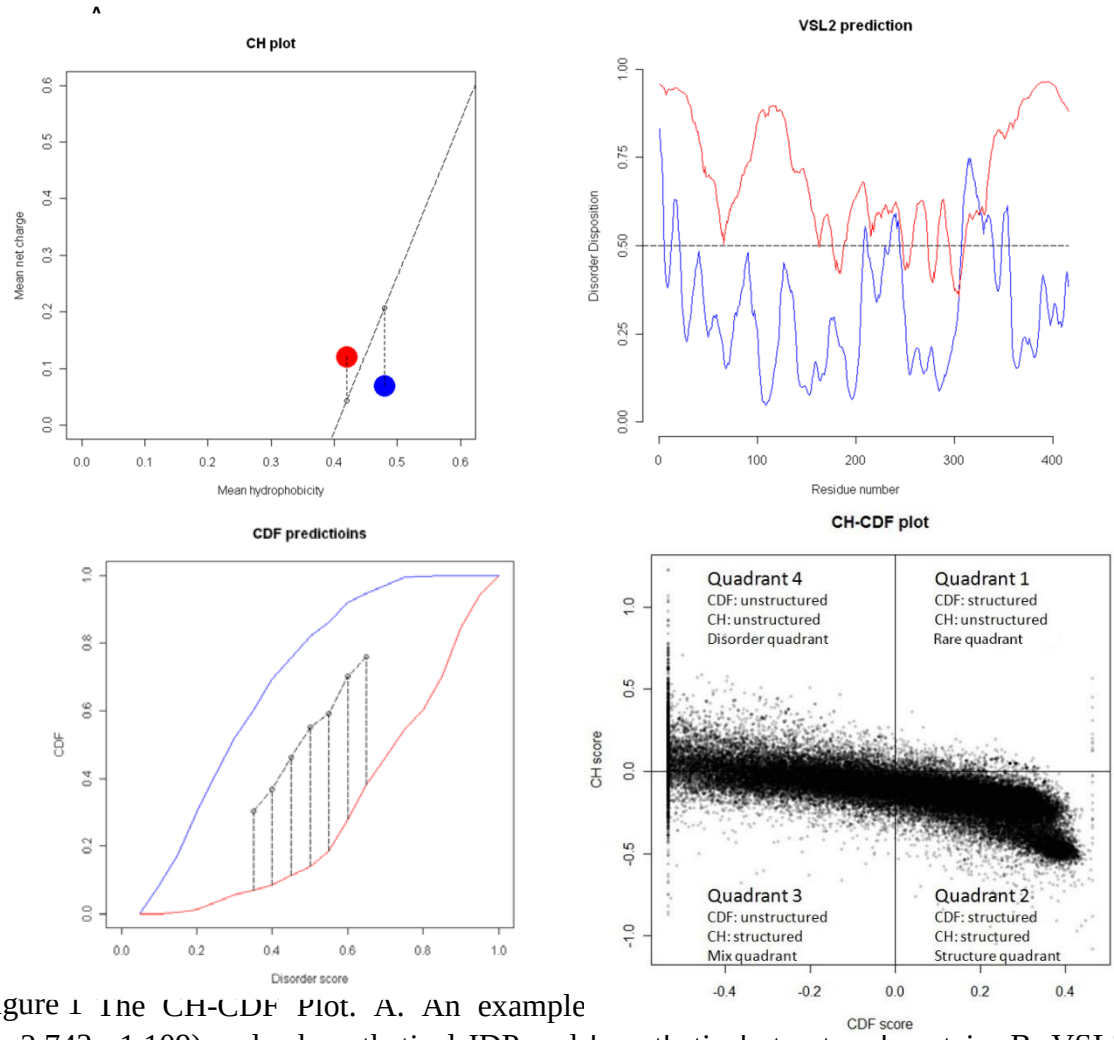


Figure 1 The CH-CDF Plot. A. An example (y=2.743x-1.109) and a hypothetical IDP and hypothetical structured protein. B. VSL2 prediction curve for an IDP (red) and a structured protein (blue). C. CDF curve of the two proteins in B. Vertical lines are the distance of to calculate CDF score. D. The entire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

One rationale behind using CH-CDF plot to subclassify disordered proteins is that CDF examines many more protein attributes than a CH plot, which only uses charge and hydropathy for prediction. Consequently, the CDF curve is more sensitive to disorder than the CH plot[9]. Proteins predicted to be ordered by the CH plot but disordered by CDF (as in Q3) are low in net charge and are 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. CH plot predicts them to be disordered, so they are typically highly charged, or hydrophilic, or both, all of which are strong indications for disorder. Since CDF method also weighs heavily on the hydropathy and charge features, the number of proteins predicted disordered by CH plot but ordered by CDF is very small. So here, we refer them as rare proteins.

1.2 PDB coverage

PDB contains protein structures, and thus PDB is biased more towards ordered proteins than disordered. Figure 2 shows PDB coverage percentages of various proteins verses 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.

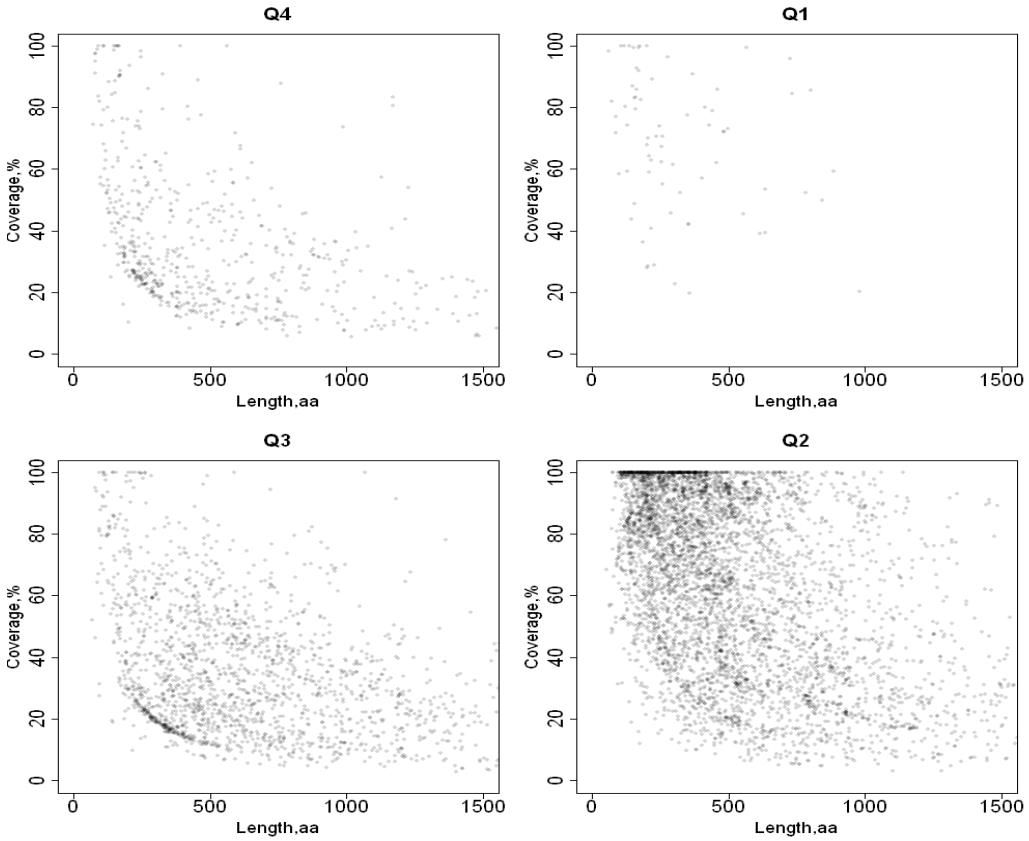


Figure 2 PDB coverage percentages of proteins classified into 4 quadrants. The PDB coverage is the combined coverage of all covered domains.

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.

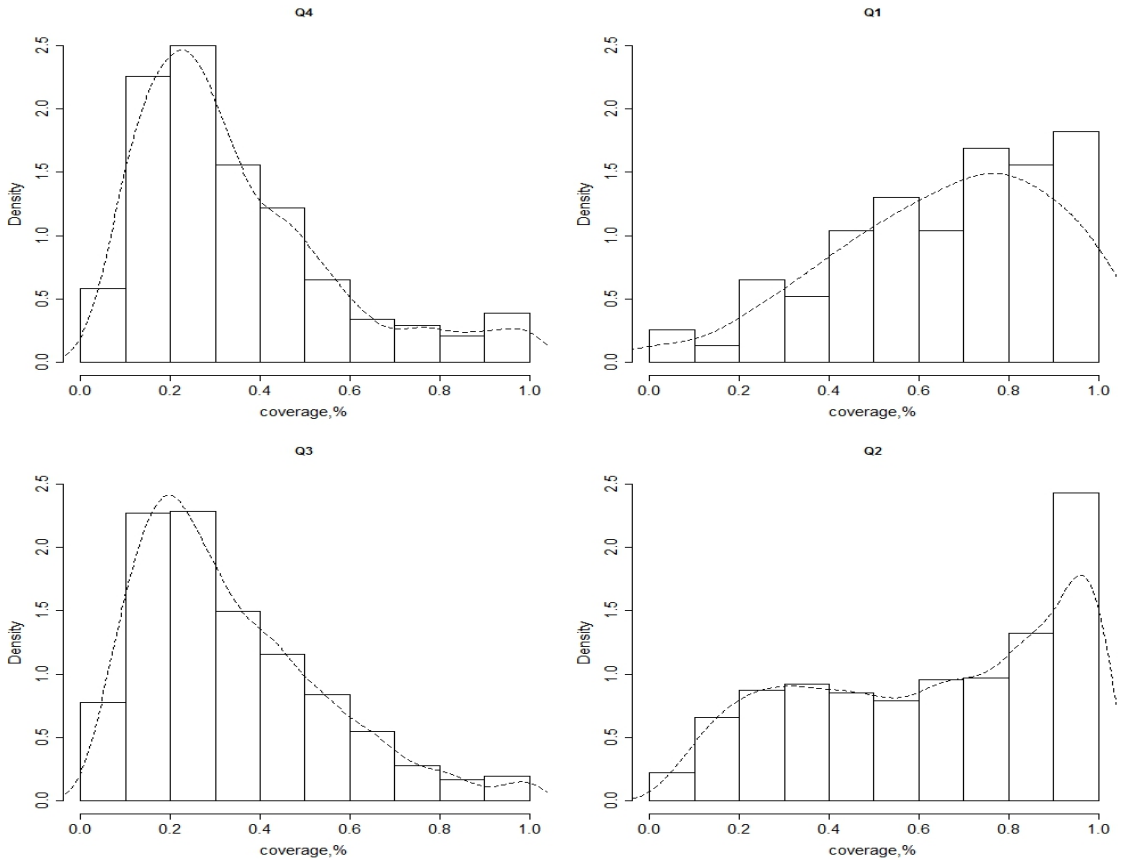


Figure 3 PDB coverage percentage histogram for all four quadrants

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, while 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 Figure 2 and Figure 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.

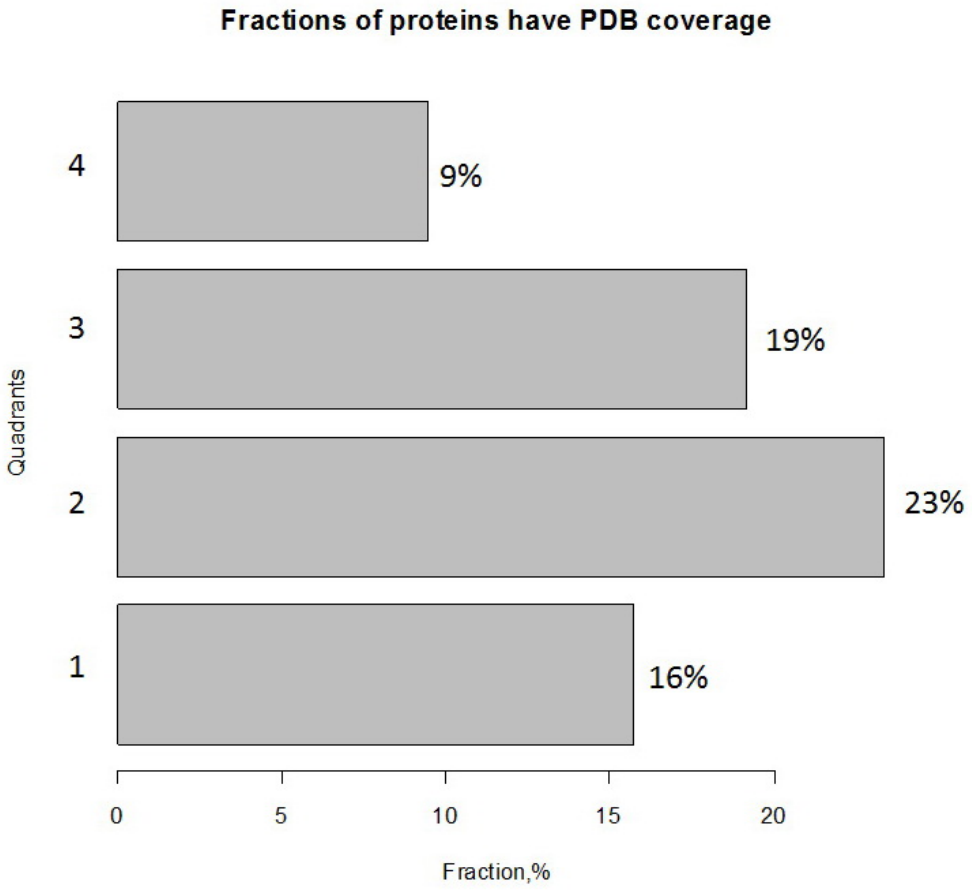


Figure 4 Fraction of protein identified with at least one PDB hit

1.3 Sequence window CH-CDF analysis

Our previous studies suggest that mixed proteins (in Q3) are likely to have both disordered regions and ordered regions. To learn more information about these proteins, we decided to dissect each protein sequence into a series of windows for our CH-CDF analysis. This is a more accurate presentation of the disorder status of a protein, as it contains segmental disorder scores. Table 1 summarizes our analysis result. Proteins from each quadrant are chopped into windows of 30 residues. Windows are analyzed by the CH-CDF method, and the fractions of windows falling into a quadrant are recorded. 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 with the most in (Q4) and slightly less in (Q2) and (Q3), suggesting that mixed proteins very likely contain a balance 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%

1.4 Match PDB coverage to disorder prediction

Since our previous studies 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. We calculated the disorder content of the PDB covered and uncovered regions, respectively. Figure 5 is the disorder content on all 4 quadrants. 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 this region is largely predicted as disordered, while less than 50% means predicted to be structured.

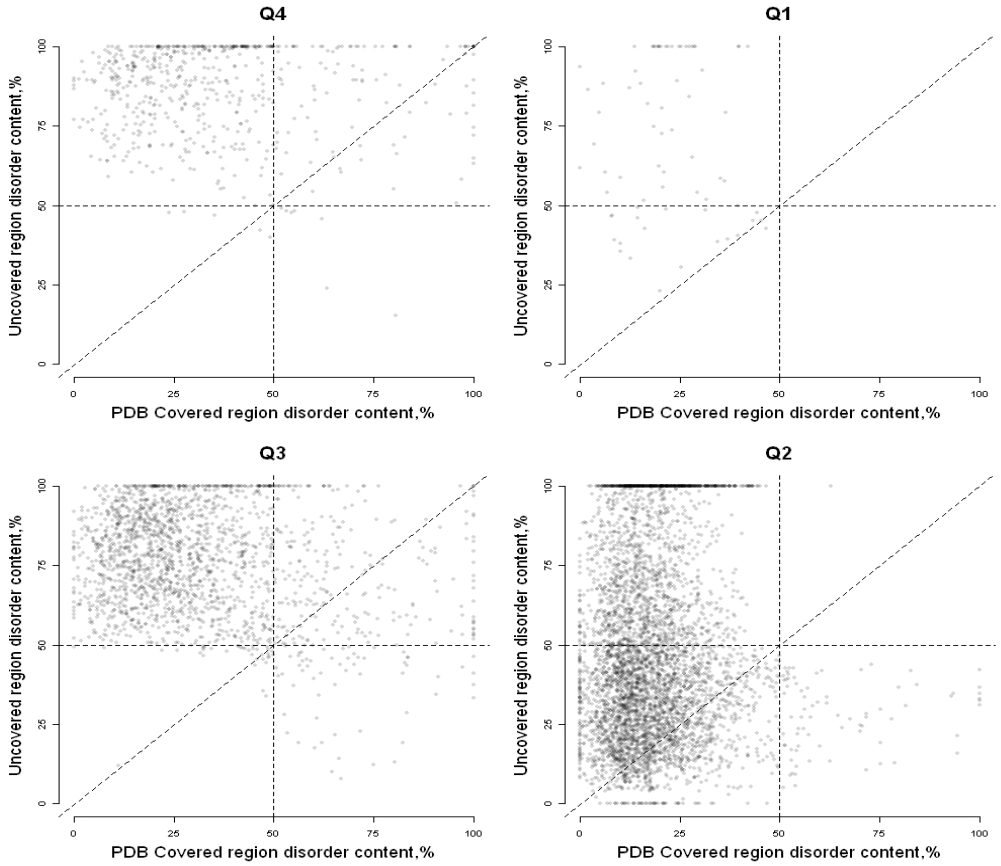


Figure 5 Percentage of disorder in PDB covered and uncovered regions

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 have residues with more than 50% predicted to be 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.

1.5 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 functions[157]. Given that the proteins in the four quadrants of the CH-CDF plots have different characteristics, it seemed reasonable to test whether these different 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 biosynthetic process
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
Notch signaling pathway
Response to heat
Osteoblast differentiation
Negative regulation of cell differentiation
Regulation of cell proliferation
Pituitary gland development
Positive regulation of neuron differentiation
Endoderm development
Organ morphogenesis
Negative regulation of signal transduction
Pancreas development
Defense response to bacterium
Endocytosis
Somitogenesis
Actin filament organization
Wnt receptor signaling pathway
Intracellular signaling cascade
Epithelial cell differentiation
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

1.6 Discussion

Since ordered proteins have different types of substructures, it is reasonable to expect that IDPs may also have subtypes and that the different subtypes may have different functions. One previous study indicated that such disordered subtypes may exist[157]. 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 common methods such as machine learning, we took an alternative approach. 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 calculations. We therefore developed a CH-CDF plot for IDP partition.

1.7 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 protein 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 among 20%-30% range, while the ordered quadrant (Q2) are as high as 90%-100%. This suggests that mixed proteins (in Q3) have more 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 examples with only 20% coverage (Figure 3, Q2). As indicated by the data in Table 1 and Figure 5, this result likely occurs because many mouse proteins have 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. Such binding could potentially stabilize them, and lead to the formation of a crystal structure. However, there are indeed cases where they are monomers by themselves. We suspect that these proteins are close to the boundary of disorder and order, which we are in the process of testing.

One of our early hypotheses was that proteins with relatively low net charge and high hydropathy, 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 without stable structure. Such proteins would likely be native molten globules. An alternative hypothesis is that proteins in (Q3) simply contain mixtures of structured regions and disordered regions.

Our following experiments attempted to test the second hypothesis that (Q3) contains mostly proteins with both structured and disordered regions. We first showed that proteins in (Q3) 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 protein 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. So it seems that the quadrant (Q3) is likely to contain proteins containing relatively balanced contributions from structured and disordered regions. For this reason here we have named the proteins in this quadrant mixed rather than collapsed disorder, which may have appeared in previous publications[159,160]. 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 such proteins and determine where they fall on the CH-CDF Plot.

1.8 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. Thus, is should be rare for such a protein to be predicted to be structured by the CDF predictor; however, this is just what happens despite the high charge and hydrophilicity. 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) (Figure 3). The proteins in (Q1) also have many more proteins identified with a PDB hit than those in the disorder quadrant (Q4) (Figure 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 collapsed 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 in the process of determining whether the positively charged proteins in (Q1) are associated with RNA binding. One approach will be testing if this is the preferred quadrant for ribosomal proteins.

1.9 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[151–153].

Meanwhile, the disorder quadrant (Q4) is mainly biased towards GO terms with mitosis-related functions, which again agrees with previous observations[151–153]. 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)[160]. 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[74,75,81] in the signaling network. Their flexibility is also capable of fast but short time-span binding, which also may be crucial in signaling events.

1.10 Summary

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-hydropathy) 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 subclassifies all proteins into 4 groups: structured, mixed, disordered, and rare. Studies of the Protein Data Bank (PDB) entries and homologues 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.

2. IDP hydropathy: A New Scale That Optimizes Disorder Prediction

2.1 Background and motivations

Intrinsically disordered proteins (IDPs) exist as flexible ensembles under normal physiological conditions, thus lacking stable tertiary structures, and yet carrying out various biological functions[2–5]. These IDPs challenge the universality of the sequence → structure → function paradigm, with biological functions associated instead with flexible ensembles rather than with structured ensembles. IDPs are involved in numerous biological activities, such as providing sites for post-translational modifications, entropic spring-based restoring forces, flexible linkers, specific binding to multiple partners, multiple binding to a specific partner, and many others[1,150–153,161–166]. Of course "structured" proteins are not lacking in motion. They too are conformational ensembles, but these ensembles involve atom fluctuation centered on equilibrium positions. Indeed, the distinction between structured ensembles and IDP ensembles is that the atom of the former has specific equilibrium positions, whereas the later do not.

Given their importance and abundance in nature, many computational tools have been developed for predicting IDPs and IDP regions from amino acid sequence, including several [1]PONDR®s[51,54,59,61], IUPred[46,52], DisoPred[163,167], SPINE-D[53], FoldIndex[58] and more than 50 others[66,68]. For the various sequence-based approaches using machine learning methodologies, hydrophobicity is widely if not universally used as one of the inputs [46,51–53,58,66,167–170].

[*] PONDR stands for Predictor of Natural Disordered Regions, and PONDR® is a registered trademark of Molecular Kinetics, Inc.

2.1.1 Protein folding, the hydrophobic effect, and disorder prediction

The hydrophobic effect makes large contribution to the folding of globular proteins[171,172]. In particular, water molecules form a dynamic 3D network by donating and accepting almost two hydrogen bonds on average from each molecule. The introduction of a nonpolar molecular that is incapable of forming hydrogen bonds with water disrupts such network. Water molecules lose reorientation and translational motion, and this process is unfavorable in terms of free energy due mainly to the loss of motional entropy[173]. As a result, hydrophobic residues are more likely to cluster on the inside of the protein to avoid contact with water molecules, whereas hydrophilic and charged residues are more likely to be exposed to surrounding solution.

Based on this theory, Uversky et al developed the Charge-Hydropathy (C-H) model to predict protein disorder[4]. In this approach, normalized net charge is plotted against normalized hydropathy, calculated from the hydropathy scale given in Kyte & Doolittle (1982)[57], giving the charge-hydropathy (C-H) plot. Remarkably, this simple C-H plot largely separates IDPs from structured proteins[56]. This model has been used both for whole protein disorder prediction via the C-H plot[56] and for residue-by-residue disorder prediction via the FoldIndex algorithm[57].

2.1.2 Various hydropathy scales

The values for the original hydrophobicity scale were estimated experimentally as the side chain free energies of transfer from selected organic solvents to water[174]. The selected organic solvents, dioxane and aqueous ethanol, were chosen because their dielectric constants are similar to the values estimated for protein interiors. Measurements

using these two solvents gave similar transfer free energy values for each of the various hydrophobic amino acids. Such free energy values for transfer from organic solvent to water are negative (e.g. spontaneous) for hydrophilic amino acids and positive (e.g. spontaneous in the opposite direction) for hydrophobic amino acids. While the original work[32] focused on the hydrophobic amino acids, later scales (reviewed in[31]) provided values for both hydrophobic and hydrophilic amino acids. To reflect the balanced importance of both hydrophobic and hydrophilic amino acids as well as to indicate a scale with both types of amino acids, Kyte and Doolittle[57] changed the name of the scale from "hydrophobic" to "hydropathic." They explained their revised name as follows: "Since hydrophilicity and hydrophobicity are no more than two extremes of a spectrum, a term that defines that spectrum would be as useful as either, just as the term light is as useful as violet light or red light. Hydropathy (strong feeling about water) has been chosen for this purpose." [31]

Since the original work of Nozaki and Tanford[32], many hydropathy scales or indices have been developed using a variety of experimental methods or using computational methods to estimate the transfer free energy values[57,110–130]. Comparison of these scales is facilitated if the various scales are normalized to a common range of values. For this normalization, we have chosen the range from –1 to +1, with positive values for the hydrophobic amino acids, thus keeping the plus-minus sign convention for hydrophobic-hydrophilic residues as used in the original Nozaki-Tanford publication[174].

The ExPASy server[175] alone provides 19 different hydrophopathy scales in ProtScale[176]. Even after normalization, the hydrophobicity value for each amino acid fluctuates by a large amount in the different scales. This raises the possibility that the

prediction accuracy of the C-H plot could be improved by using a different hydropathy scale.

Here we used the C-H plot formalism to compare the structure-disorder prediction accuracy when combined with net charge for the 19 hydropathy scales from ExPASy along with the prediction accuracies for all 535 amino acid indices obtained from the Amino Acid index database (AAindex)[128–130], TOP-IDP[49], FoldUnfold[131], B-value[132], and DisProt[49,64,42,133]. Next we used the formalism underlying the linear support vector machine[146,177] to develop a new hydropathy scale that further improves prediction of IDPs. As we show by several measures, our new scale, which we named IDP-Hydropathy, gives substantially improved predictions as compared to the originally used Kyte-Doolittle scale and also as compared to the best of the tested hydropathy scales. Here we report these comparisons of the various hydropathy scales as well our analysis of their predictions and prediction errors on our set of fully structured and fully disordered proteins. In addition to improved predictions using the C-H plot, we speculate that, given the strong negative correlation between crystallographic disorder and hydropathy[178], our new scale would likely improve disorder prediction for any algorithm that uses hydropathy as one of the inputs and that is based on training sets dominated by crystallographic disorder.

2.2 Comparing Hydropathy scale of Kyte & Doolittle (1982) with 18 other hydropathy scales

The C-H plot developed by Uversky et al[3] is a straightforward, simple, fast, yet effective whole protein disorder versus order predictor. FoldIndex is a per residue

predictor adapted from the C-H plot, using the same features of charge and hydropathy as the C-H plot. Because of their dependence on intuitive biophysical features and their simplicity, both methods are still heavily used today. However, unlike net charge, which is unambiguous, a variety of hydropathy scales have developed using quite different methods and assumptions. Thus, the various scales have the potential of being more or less useful, depending on the application.

The hydropathy scale of Kyte & Doolittle (1982) [57] has been used in both the whole protein predictor based on the CH-plot and in the FoldIndex per residue predictor. Therefore, one natural question to ask is, how well do other hydropathy scales perform compared to this particular hydropathy scale? To compare the performances of various hydropathy scales, the 19 different hydropathy scales from ExPASy were tested via C-H plots to predict the structure – disorder status of the proteins in our dataset. The results of this experiment are given in Table 3.

Table 3 The Order versus Disorder Prediction Performances of 19 Hydropathy Scales. For each scale, the citation is given by the superscript number.

Hydropathy Scales	F-Score	MCC	AUC	Sensitivity	Specificity	PPV	NPV
*Guy (1985)[110]	0.75	0.71	0.90	0.69	0.97	0.83	0.94
Miyazawa & Jernigan (1985)[111]	0.74	0.71	0.90	0.71	0.97	0.81	0.95
Manavalan & Ponnuswamy (1978)[112]	0.74	0.71	0.89	0.71	0.97	0.80	0.95
Fauchere & Pliska (1983)[113]	0.74	0.71	0.88	0.67	0.97	0.85	0.94
Rose et al. (1985)[114]	0.73	0.70	0.91	0.67	0.97	0.83	0.94
Sweet & Eisenberg (1983)[115]	0.73	0.70	0.90	0.68	0.97	0.82	0.94
Black & Mould (1991)[116]	0.70	0.67	0.87	0.64	0.97	0.82	0.93
Hopp & Woods (1981)[117]	0.69	0.66	0.88	0.62	0.97	0.81	0.93
^Kyte & Doolittle (1982)[57]	0.68	0.64	0.87	0.60	0.97	0.80	0.93
Bull & Breese (1974)[118]	0.67	0.62	0.88	0.62	0.96	0.75	0.93
Abraham & Leo (1987)[119]	0.66	0.62	0.86	0.59	0.96	0.78	0.93
Chothia (1976)[120]	0.64	0.60	0.87	0.54	0.97	0.8	0.92
Roseman (1988)[121]	0.64	0.60	0.86	0.56	0.97	0.78	0.92
Rao & Argos (1986)[122]	0.61	0.58	0.85	0.53	0.97	0.77	0.91
Janin (1979)[123]	0.58	0.54	0.86	0.50	0.96	0.74	0.91
Eisenberg et al. (1984)[124]	0.56	0.52	0.85	0.47	0.96	0.74	0.90
Tanford (1962)[125]	0.55	0.51	0.86	0.46	0.96	0.72	0.90
Welling et al. (1985)[126]	0.51	0.49	0.78	0.40	0.98	0.77	0.89
Wolfenden et al. (1981)[127]	0.45	0.42	0.79	0.35	0.97	0.69	0.89

* Hydropathy scale Guy (1985) gives the top performance.
^ Hydropathy scale Kyte & Doolittle (1982) only achieves average performance.
MCC: Matthew Correlation Coefficient
AUC: Area Under the Curve
PPV: Positive Predictive Values
NPV: Negative Predictive Values

Our data show that, in terms of disorder prediction, the hydropathy scale of Kyte & Doolittle (1982) is only average, giving the following values for the various performance metrics: 0.68 F-score, 0.60 sensitivity, 0.97 specificity, and 0.80 PPV, and ranking in the middle of the 19 hydropathy scales. The Guy (1985) hydropathy scale gives the highest F-score, a value of 0.75, which is a 10% improvement compared to the hydropathy scale of Kyte & Doolittle (1982). The hydropathy scale of Guy (1985) also reaches 0.69 sensitivity, which is a 15% improvement compared to the hydropathy scale of Kyte & Doolittle (1982). Meanwhile, use of the Guy (1985) scale maintains a PPV score of 0.83. Clearly the Guy (1985) hydropathy scale gives improved performance compared to that of Kyte & Doolittle (1982) when used with net charge to classify structured and disordered proteins via the C-H plot.

2.3 Finding the optimal hydropathy scale for IDP prediction

Since disorder prediction based on C-H plot can be significantly improved by simply adapting another hydropathy scale, it seems reasonable to ask whether another hydropathy scale can be developed that further improves the performance of the C-H plot.

2.3.1 Use of Linear SVMs to find hydropathy scales

To find a hydropathy scale that gives an improved order-disorder classification via the C-H plot methodology, we adopted a linear support vector machine (SVM)[179] for this purpose. SVMs represent a new generation of learning systems based on recent advances in statistical learning theory[146,177]. The aim in training a linear SVM is to find the separating hyperplane with the largest margin; the expectation is that the larger the

margin, the better the generalization of the classifier. Typically, the weights that are found as giving the best performance are viewed as meaningless, arbitrary parameters. However, in this particular instance, the SVM weight given to each amino acid corresponds to its hydropathy value.

Given the above, we rephrase the question of finding the optimal scale by viewing sets of protein sequences/windows as an *n* by 21 matrix (Eq. 10). The *n* rows represent n protein sequences/windows, and 21 columns are comprised of 20 normalized amino acid compositions and normalized net charge. For sequence/window *i*, $Comp_{ij}$ is its *j*'s amino acid composition, and C_i is its normalized net charge, calculated as (Eq. 11). We represent the disorder/order status of *i*th protein sequence/window as Y_i (1 or -1), thus giving:

$$\begin{bmatrix} Y1 \\ Y2 \\ \vdots \\ \vdots \\ Yn \end{bmatrix} = \begin{bmatrix} Comp_{11} & Comp_{12} & \dots & Comp_{20} & C1 \\ Comp_{21} & Comp_{22} & \dots & Comp_{20} & C2 \\ Comp_{31} & Comp_{32} & \cdots & Comp_{20} & C3 \\ \dots & \dots & \ddots & \dots & \dots \\ Comp_{n1} & Comp_{n2} & \cdots & Comp_{20} & Cn \end{bmatrix} * \begin{bmatrix} w1 \\ w2 \\ \vdots \\ \vdots \\ w20 \\ w21 \end{bmatrix} + b \text{ (Equation 10),}$$

$$where\ C_i = Comp_{iArg} + Comp_{iLys} - Comp_{iGlu} - Comp_{iAsp} \text{ (Equation 11).}$$

Then, the linear SVM is employed here to find a 21 by 1 weight vector *w*, such that $wM + b$ (bias) is closest to Y (Eq. 10). We then can transform the w1 to w20 values into our IDP-hydropathy scale by scaling them. For comparison with the other normalized hydropathy scales, these 20 weights were normalized such that they lie on the interval of -1 to +1. However, since the first published C-H plot by Uversky normalized

the Kyte-Doolittle scale to the interval of 0 to +1, we renormalize the hydropathy scale to the same interval when we draw the C-H plot.

A 10-fold cross validation was used here, and iterated for 5 times in this method. We also tested a genetic algorithm[180] and an elastic net[147] (i.e., a penalized logistic regression classifier) as alternatives for the generation of the best hydropathy scale for order / disorder classification via the C-H plot. Both of these approaches give scales with prediction performance values similar to those obtained by the SVM methodology. We chose to present the SVM approach because of its greater simplicity and elegance compared to the other methods.

2.3.2 Choosing window size for training

We previously showed that amino acid compositions associated with disordered segments exhibit changes that depend on segment length[181] and that construction of length-dependent predictors gives improved performance[54]. To minimize such length-dependent variation, we tested whether use of uniform-sized segments of protein during training would improve the subsequent classifiers based on the C-H plot. We found this to be the case. Thus, for improved training, the goal became one of finding a segment size that used as much of the data as possible while giving high quality predictions.

Different window sizes were tested as shown in Figure 6. Since entirely disordered or entirely ordered proteins are used here, larger window sizes mean more information and less noise. As a result, the general trend is that, as window size increases, the F-scores also increase, indicating higher quality predictions. However, due to the limitation of protein length, longer windows mean that more sequences must be discarded.

These opposing effects were estimated by trying various segment sizes for which a protein's sequence was divided into a collection of segments for which each successive segment is obtained by shifting the position of a given segment by half the length of the segment until there is insufficient protein remaining to cover the shifted segment.

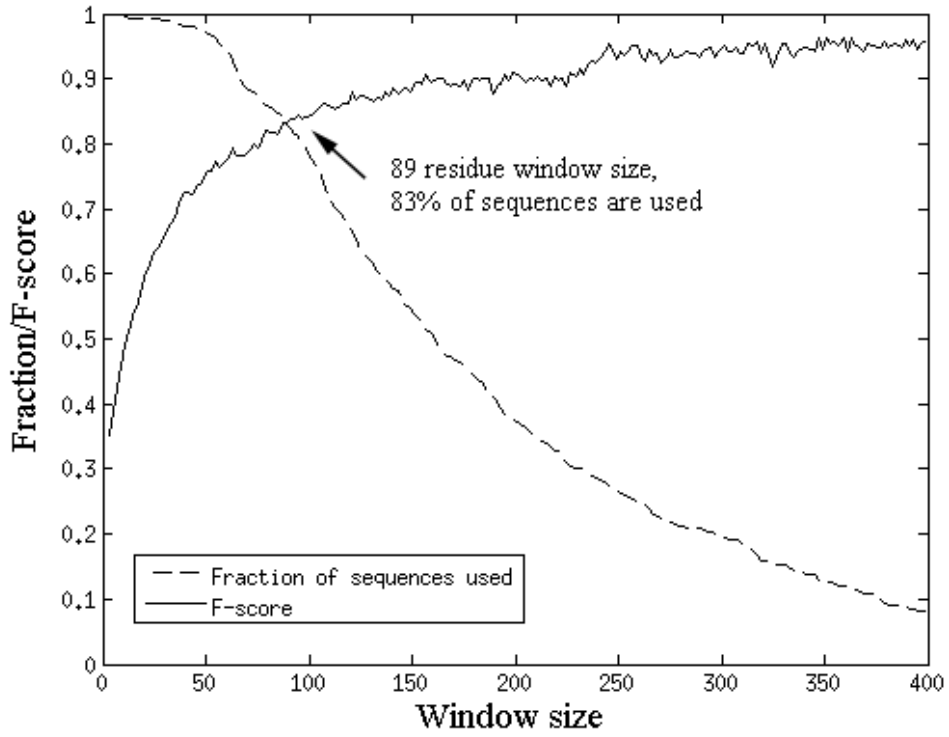


Figure 6 F-score and fraction of retained sequences versus window size.

Both F-score and fraction of retained sequences range from 0 to 1.

Using the method just described, Figure 6 compares the F-score and the fraction of sequences that are not discarded, versus the segment length used. A segment length of 89 residues was chosen as a good compromise, yielding both a high F-score and a low loss of amino acid sequences. This window size renders a relatively high performance, 0.84 F-score, and high sequence retention, 83%.

The hydropathy scale over a window size of 89 amino acids was constructed from the weight vector found by the SVM. To be consistent with the original C-H plot paper, and with previous hydropathy scale test results, this scale is applied and tested over the entire protein sequences. This new scale shows an improved performance compared to the tested 19 scales, namely: 0.86 F-score, 0.83 sensitivity, 0.98 specificity, 0.95 AUC and 0.91 PPV. We named this scale "IDP-Hydropathy" (Table 4). Note that in the original C-H plot method developed by Uversky et al, the Kyte-Doolittle scale was normalized to 0 and 1. In many hydropathy scales, however, negative values are usually used to indicate a hydrophilic residue. To accommodate this tradition and to compare different hydropathy scales, we normalized all scales to the range of -1 to 1.

Table 4 IDP-Hydropathy Scale.

Residue	W	Y	I	F	L	C	V	N	T	A
Hydropathy Score	1.00	0.54	0.34	0.25	0.23	0.22	0.19	-0.14	-0.20	-0.27
Residue	G	R	M	Q	D	S	H	K	E	P
Hydropathy Score	-0.41	-0.43	-0.46	-0.52	-0.56	-0.65	-0.71	-0.71	-0.72	-1.00

This scale is normalized to span the interval from -1 to +1.

2.4 Disorder is harder to predict

One interesting observation here is that across all tested hydropathy scales, including the IDP-Hydropathy, the specificity is high (>0.96) for all predictors, while the sensitivity is quite low compared to specificity. The sensitivity when using the scale of Guy et al is only 0.69, and the highest sensitivity is only 0.71 in the preliminary study. IDP-Hydropathy also has a relatively large gap between its sensitivity (0.83) and specificity (0.98). Disorder seems to be harder to predict than structure. We hypothesize that this results from the existence of many structure-forming segments being present within most experimentally characterized disordered proteins.

This hypothesis is supported by running per residue predictors, PONDR® VLXT[51] and VSL2[54] on our whole disordered/structured protein dataset. Fractions of predicted disorder/order over the entire disordered/ordered dataset by each predictor are displayed in Table 5. PONDR® VLXT predictor predicts residue disorder tendencies within a narrow window, and is built to be very sensitive to protein sequence local features. PONDR® VSL2, on the other hand, uses a longer window, and so its prediction is smoother with less focus on local changes. In Table 5, on average, PONDR® VLXT predicts only 58% disordered residues within an entirely disordered protein, while it predicts 78% structured residues for the sequence of a wholly structured protein. The PONDR® VSL2 prediction results are quite different. VSL2 has a comparable amount of predicted disorder residues within disordered protein as predicted structure in a structured protein. This suggests that indeed, there are many short segments with potential for structure-formation within regions within a disordered protein. This provides one possible explanation for the phenomenon in our previous study of choosing window size

as well. As the window size increases, such structure-forming small regions are smoothed out, and we observe an increase in the performance of disorder prediction.

Table 5 VLXT and VSL2 per residue prediction over our entirely disordered/structured dataset.

		Predicted			
		VLXT		***VSL2***	
		Disorder	Structure	Disorder	Structure
Dataset	Disordered	58%	~	78%	~
	Structured	~	78%	~	74%

The entries are fraction of residues that are predicted disordered/structured over the whole disordered/structured dataset. For simplicity, only diagonal entries in each predictor is shown.

2.5 Benchmark

In order to benchmark the IDP-Hydropathy scale, we compared it with 4 currently known disorder propensity scales, and with additional 531 amino acid scales retrieved from the AAIndex. As shown in Table 6, IDP-Hydropathy performed better than previously published disorder propensity scales. Its F-score of 0.86 outperformed DisProt (0.80), TOPIDP (0.79), FoldUnfold (0.77) and B-value (0.75).

Table 6 IDP-Hydropathy scale performance compared to 4 disorder propensity scales.

Hydro Scale	F-Score	MCC	AUC	Mean Acc	Sensitivity	Specificity	PPV	NPV
IDP-Hydro	0.86	0.84	0.95	0.91	0.83	0.98	0.91	0.97
DisProt	0.80	0.77	0.94	0.87	0.77	0.97	0.85	0.96
TOPIDP	0.79	0.76	0.93	0.87	0.76	0.97	0.84	0.96
FoldUnfold	0.77	0.73	0.91	0.85	0.72	0.97	0.83	0.95
Bvalue	0.75	0.72	0.92	0.83	0.67	0.98	0.87	0.94

The IDP-Hydropathy outperformed all 535 amino acids scales as well (Table 7). Compared to Linker index[182], which showed the best performance among all 535 AAIndex scales, IDP-Hydropathy scale had higher ranking for every performance metric that was measured. In particular, IDP-Hydropathy performs 5% and 6% better in terms of F-score and sensitivity than the Linker index did. Interestingly, Linker index was developed to predict 'linkers' between ordered protein domains. These linkers are rich in disordered regions. Therefore, it can also be viewed as a disorder index.

Table 7 IDP-Hydropathy scale performance compared to top 10 scales from AAIndex of top performance

Hydro Scale	F-Score	MCC	AUC	Mean Acc	Sensitivity	Specificity	PPV	NPV
IDP-Hydro	0.86	0.84	0.95	0.91	0.83	0.98	0.91	0.97
Linker index	0.82	0.80	0.93	0.88	0.78	0.98	0.89	0.96
beta-sheet 1	0.80	0.78	0.92	0.86	0.74	0.98	0.90	0.95
beta-strands	0.80	0.77	0.90	0.86	0.74	0.98	0.88	0.95
beta-sheet 2	0.80	0.77	0.93	0.87	0.75	0.98	0.87	0.95
beta-sheet 3	0.79	0.76	0.90	0.86	0.74	0.98	0.86	0.95
Contact number	0.79	0.76	0.91	0.86	0.74	0.97	0.85	0.95
Buriability	0.78	0.75	0.91	0.85	0.73	0.97	0.86	0.95
Hydropathy	0.78	0.75	0.92	0.85	0.71	0.98	0.87	0.95
Partition energies	0.77	0.74	0.92	0.85	0.73	0.97	0.84	0.95
Interactivity	0.77	0.74	0.90	0.85	0.73	0.97	0.83	0.95

The ROC curve of IDP-Hydropathy, along with Linker Index, which is the top performance scale in AAIndex, top correlated scales in AAIndex, hydropathy scale of Guy (1985) from ExPAsy, and hydropathy scale of Kyte & Doolittle (1982) are represented in Figure 7, which clearly shows that the ROC for IDP-Hydropathy outperforms all others.

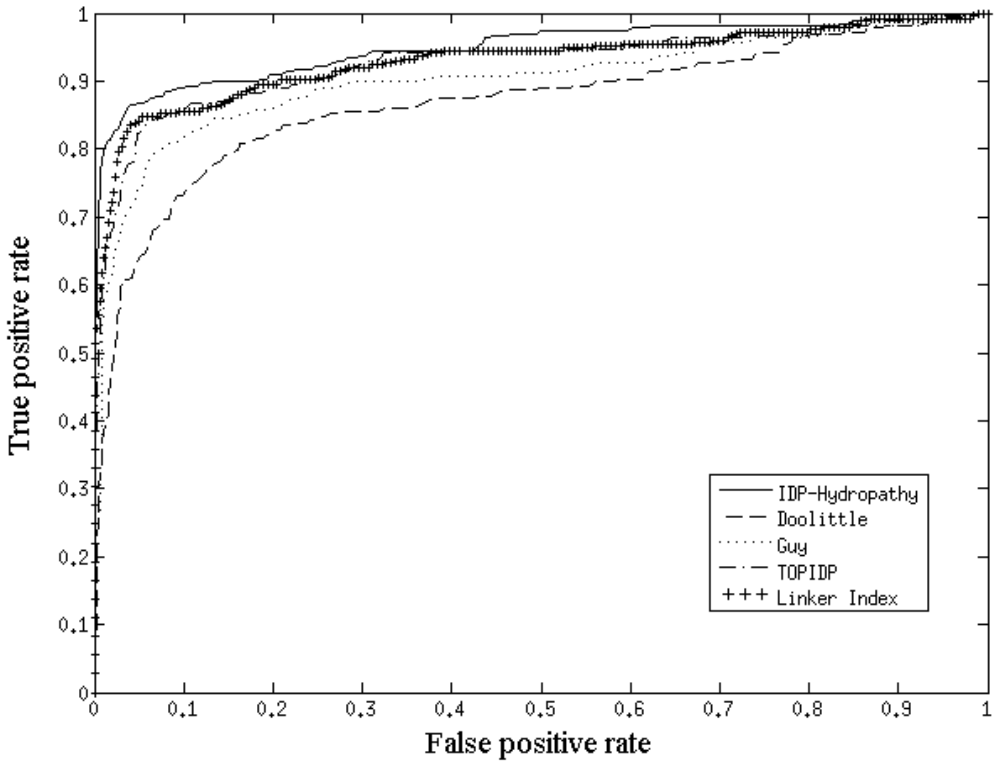


Figure 7 ROC curves for order-disorder classification

The True Positive Rate is plotted versus the False Positive Rate. The IDP-hydropathy is compared with the next two highest ranking scales, Linker Index and TOP IDP, as well as with the highest ranking hydropathy scale, Guy (1985), and the originally used hydropathy scale, Kyte & Doolittle (1981).

2.6 Correlation study

Since IDP-Hydropathy scale is derived via computation, and focused on maximizing prediction accuracy rather than being based on real physical attributes, another question to ask is if this scale is truly a hydropathy scale or if it contains input from other amino acid properties. One way to test this possibility is to study how this scale correlates with the 531 scales given in the AAIndex set. As shown in Table 8, the most correlated scale is the Interactivity scale, which is basically a hydrophobicity scale obtained through residue-residue interaction. The 2nd and 3rd most correlated scales are both from the work of Zhou and co-workers in which the amino acid stability and buriability were estimated[183]. These scales are highly correlated with hydropathy scales as well. The 4th scale, Linker index, is actually a form of disorder propensity index as discussed earlier. The next, Tanford hydrophobicity scale[125], is established by calculating the free energy transfer of amino acid side chains and backbone peptides from water to ethanol and dioxane solutions. This study shows that the IDP-Hydropathy scale shows the highest correlations with other hydrophobicity or hydropathy scales, or scales closely associated with hydrophobicity or hydropathy.

Table 8 Top 10 IDP-Hydropathy correlated AAIndex amino acid scales

AAIndex Name	**|Correlation|**	**F-score**
Interactivity scale,	0.87	0.76
The stability scale	0.87	0.72
Buriability	0.85	0.78
Linker index	0.83	0.82
Transfer energy, organic solvent	0.83	0.72
Normalized frequency of beta-sheet from CF	0.81	0.76
Normalized frequency of beta-sheet	0.81	0.77
Bitterness (Hydrophobicity parameters)	0.80	0.73
Conformational preference for antiparallel beta-strands	0.80	0.80
Flexibility parameters	0.80	0.73

For an easier representation of different AAIndex scales, we grouped AAIndex scales to the four categories according to their experimental procedures as “Interactivity”, “Buriability”, “Transfer energy” and “Others”. We use “Interactivity” to indicate that this scale is obtained by measuring the interactions between the amino acids within a structured protein. “Buriability” means that the authors compared the surface residues and inner buried residues to calculate their scales. The 3rd category, “Transfer energy” measures the amount of energy associated with the transfer of residues from polar to non-ploar phase. “Others” are methods other than these three, including vapor pressure measurements, chromatography, and so on. These 4 groups are adapted because IDP-Hydropathy is mostly related to AAIndex derived from “Interactivity”, “Buriability”, and “Transfer energy”.

As a result, the IDP-Hydropathy scale is most closely related to AAIndex scales that are derived using “Interactivity”, “Buriablility”, and “Transfer energy”. A similar correlation study of the 19 hydropathy scales obtained from ExPAsy reveals that they are also highly related to AAIndex scales derived using these three approaches (Table 9). These comparisons show that the IDP-Hydropathy scale is very similar to the experimentally derived hydropathy scales.

2.7 Heat map and values of IDP-Hydropathy versus other scales

The IDP-Hydropathy scale is compared to four of its most correlated scales from AAIndex, four disorder propensity scale, and Kyte-Doolittle scale by means of a heat map (Figure 8). Compared to Kyte-Doolittle scale, IDP-Hydropathy assigns Trp and Tyr

as hydrophobic, the same as other four hydrophobicity scales and four disorder propensity scales.

Table 9 19 hydropathy scales in ExPAsy and their top 3 most correlated AAIndex scale methods.

Scale names	**Top 3 most correlated AAIndex Methods**		
Miyazawa & Jernigen (1985)	Interactivity	Interactivity	Interactivity
Rao & Argos (1986)	Others	Interactivity	Others
Manavalan & Ponnuswamy (1978)	Buriability	Interactivity	Buriability
Rose et al. (1985)	Others	Buriability	Others
Sweet & Eisenberg (1983)	Others	Others	Interactivity
Black & Mould (1991)	Buriability	Others	Transfer energy
Janin (1979)	Transfer energy	Interactivity	Buriability
Wolfenden et al. (1981)	Others	Transfer energy	Others
Guy (1985)	Interactivity	Others	Others
Hopp & Woods (1981)	Others	Others	Transfer energy
Abraham & Leo (1987)	Others	Others	Interactivity
^Kyte & Doolittle (1982)	Others	Buriability	Buriability
Welling et al. (1985)	Others	Others	Transfer energy
Tanford (1962)	Hydropathy	Interactivity	Transfer energy
Chothia (1976)	Buriability	Buriability	Others
Eisenberg et al. (1984)	Others	Transfer energy	Interactivity
Fauchere & Pliska (1983)	Others	Others	Others
Bull & Breese (1974)	Transfer energy	Buriability	Others
Roseman (1988)	Transfer energy	Others	Buriability

Note that IDP-Hydropathy is highly correlated with AAIndex scale methods “Interactivity”, “Buriability”, and “Transfer energy”, which are all observed many times in this table.

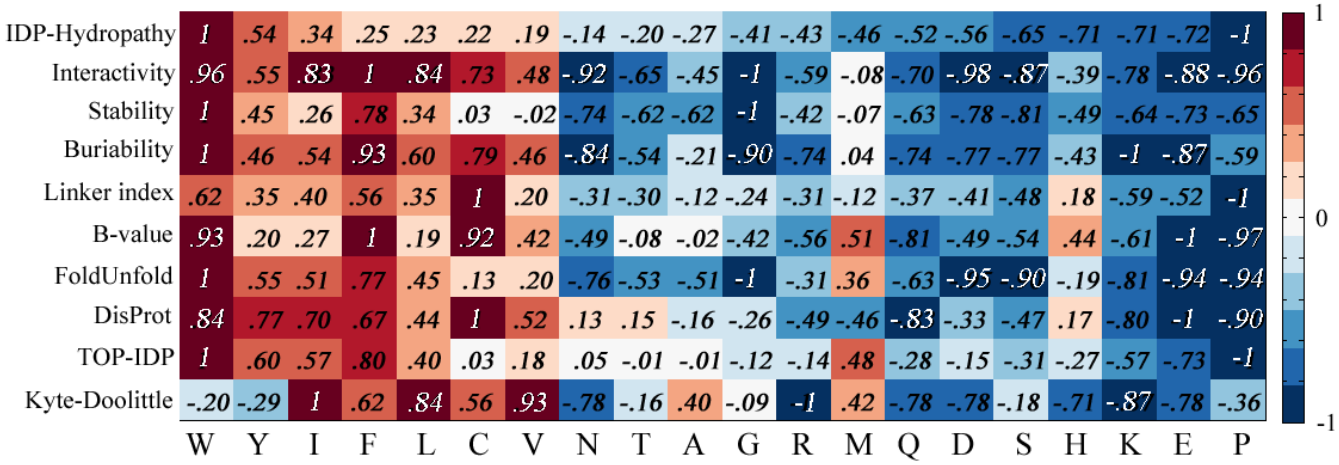
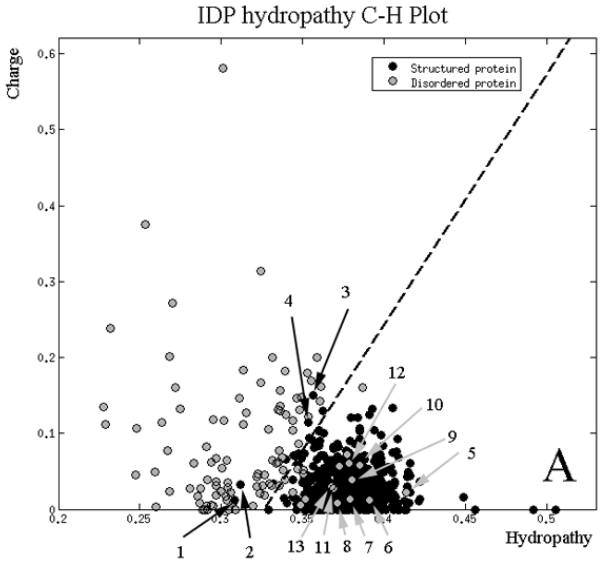


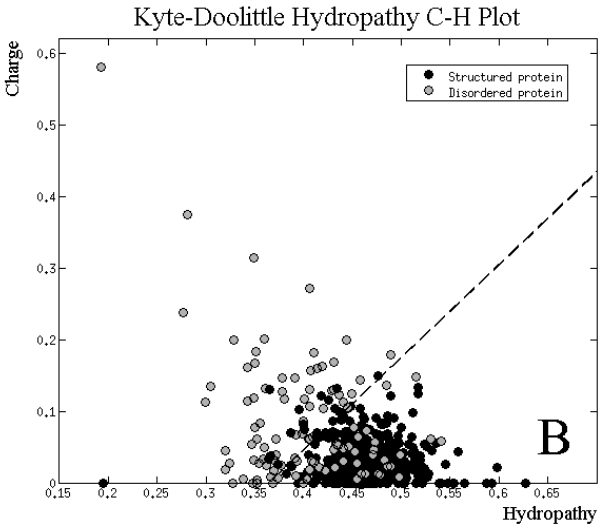
Figure 8 Comparing hydropathy scales

To visually compare several of the hydropathy scales, a heat map was constructed for the following scales (from top to bottom): IDP-Hydropathy scale, the 4 most correlated scales from AAIndex (interactivity scale, stability scale, buriability scale and linker index), 4 disorder propensity scales (B-value, FoldUnfold, Disprot, and TOP-IDP), and Kyte-Doolittle scale. Each scale is annotated with their values normalized from -1 to 1.

2.8 Comparing C-H plots from different hydropathy scales

In Figure 9, the C-H plots constructed using 3 different hydropathy scales are compared: 1. The IDP-Hydropathy Scale (4A), the Kyte-Doolittle Hydropathy Scale (4B) and the Guy Hydropathy Scale (4C). In terms of classification accuracy, 4A is better than 4C which is better than 4B. Notice that the net charge of each protein remains constant, so changes in the hydropathy scales correspond to horizontal movements of the proteins on the C-H plot. In effect, IDP-hydropathy leads to the greatest leftward shifts for the disordered proteins and the greatest rightward shifts for the structured proteins.


IDP hydropathy C-H Plot
Structured protein
Disordered protein
Charge
Hydropathy
A
1
2
3
4
5
6
7
8
9
10
11
12
13


Kyte-Doolittle Hydropathy C-H Plot
Structured protein
Disordered protein
Charge
Hydropathy
B

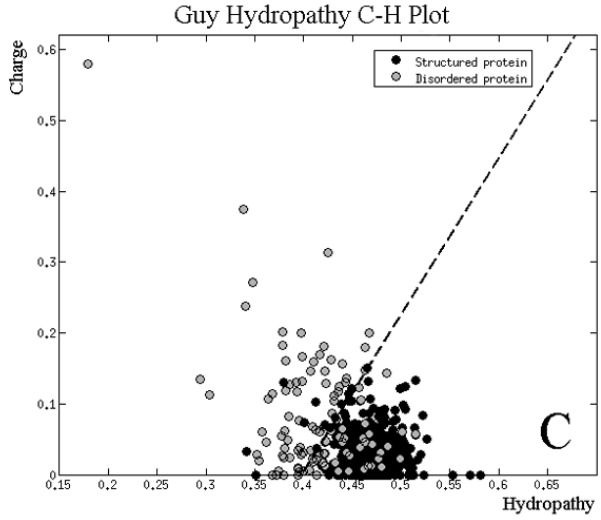
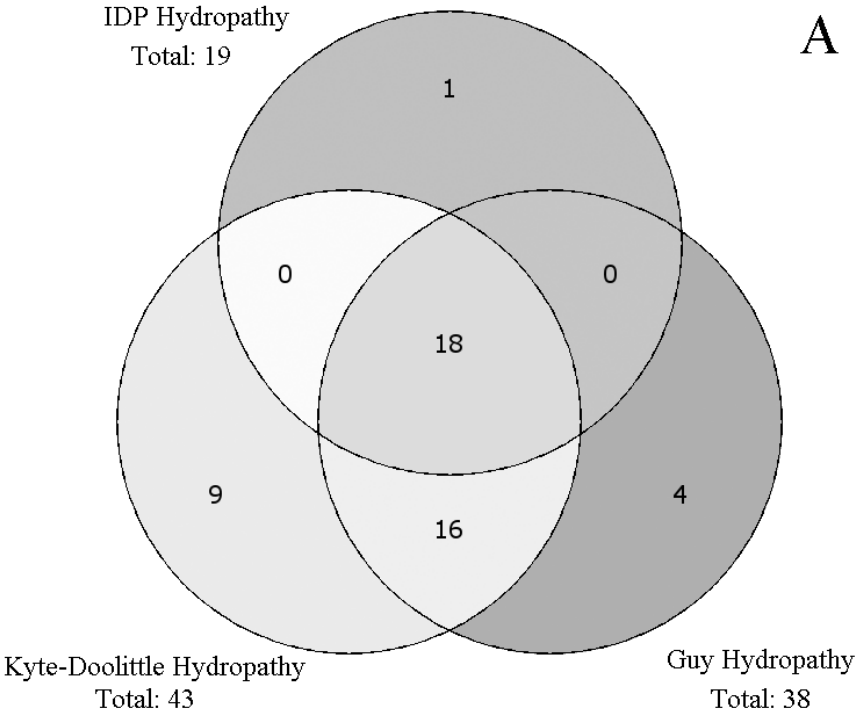

Guy Hydropathy C-H Plot
Structured protein
Disordered protein
Charge
Hydropathy
C

Figure 9 Charge-Hydropathy plots.

In (A) the IDP-Hydropathy scale was used, in (B) the Kyte−Doolittle (1981) Hydropathy scale was used, and in (C) the Guy (1985) hydropathy scale was used. Gray circles indicate disordered proteins, black circles indicate structured proteins. For these plots, each scale was normalized to be in the interval of 0 to 1. In (A) the numbers are indices for misclassified examples given in Table 10, and the function describing the boundary is: <charge> = 3.35<hydropathy> − 1.09. In (B) the function describing the boundary is: <charge> =1.31<hydropathy> − 0.48. In (C), the function describing the boundary is: <charge> = 2.25<hydropathy> − 0.90.

The relationships among the misclassified proteins observed for the three hydropathy scales are shown in the Venn diagram of Figure 10, with misclassified disordered proteins in 5A and misclassified ordered proteins in 5B.

Mis-classified disordered proteins Venn diagram



Mis-classified ordered proteins Venn diagram

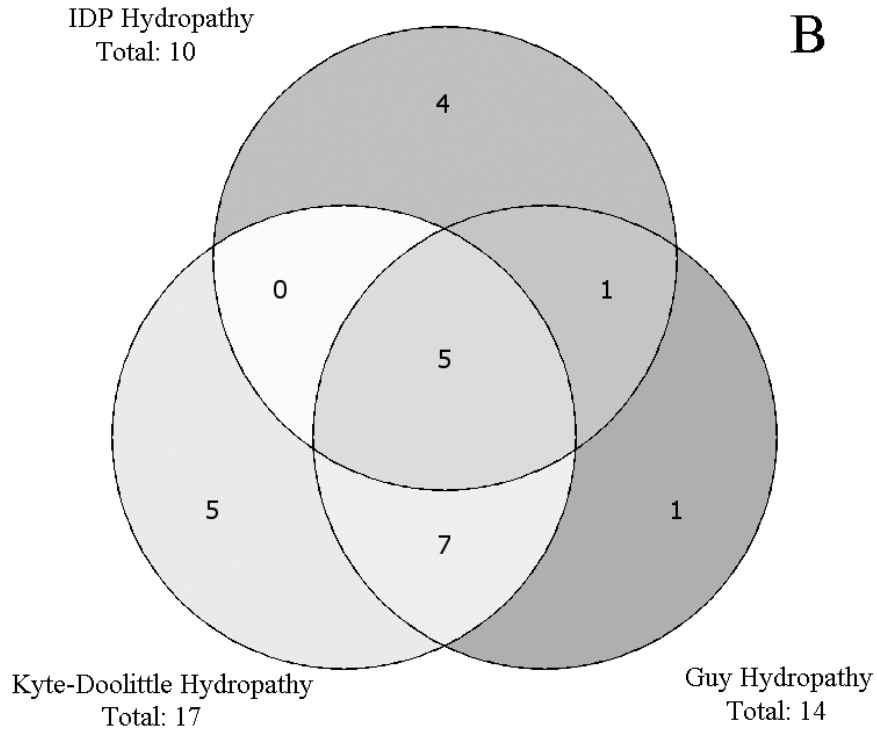


Figure 10 Venn diagrams.

In (A), the intersections are given for sets of misclassified disordered proteins that result from use of IDP hydropathy, Kyte–Doolittle (1981) hydropathy, or Guy (1985) hydropathy. In (B), the intersections are given for the sets of misclassified ordered proteins that result from use of the same three hydropathy scales.

In our dataset of disordered proteins (109 in total, Figure 10A), IDP-Hydropathy made 19 mistakes (17%), while the Kyte-Doolittle scale made 43 (39%), and the Guy scale made 38 (35%). Almost all of the disordered proteins predicted to ordered by the IDP-Hydropathy (18 out of 19) are also misclassified by the other two scales. On the other hand, 29 of the mistakes made by either the Kyte-Doolittle scale or by the Guy scale are correctly classified by IDP-Hydropathy scale.

Of the 563 ordered proteins, there are only 10 (2%), 17 (3%), and 14 (3%) misclassifications by the IDP-hydropathy, Kyte-Doolittle, and Guy scales, respectively. The IDP-Hydropathy scale committed 4 errors not committed by either of the other two scales, and 6 errors shared by one of other two scales. On the other hand, IDP-hydropathy avoided 13 errors made by at least one of the other two predictors.

Consider the C-H plot based on IDP-Hydropathy (Figure 9A). For this plot there are only 10 ordered proteins misclassified as disordered, and 19 disordered proteins misclassified as ordered. Several of these misclassified proteins are very close to the boundary and so were ignored, leaving 4 of the misclassified ordered proteins (indexed 1-4), and 9 of the misclassified disordered proteins (indexed 5-13). These individual misclassified proteins were subjected to detailed analyses to determine, if the possible, the reasons for the misclassification.

Protein 1 (PDB ID: 2PNE) is a structured protein that was predicted to be disordered. This protein adopts a highly unusual six-stranded polyproline II helix bundle, and contains two structure-stabilizing disulfide bonds that somehow escaped the detection of our disulfide bonds filter, which tested the distances between cysteine sulfur atoms to detect disulfide bonds. This failure of our filter is being studied further to

improve future work. The amino acid composition of this protein certainly looks like that of an unstructured protein, and this protein is strongly predicted to be disordered by our various PONDR® algorithms. This highly unusual structured protein is simply misclassified by the C-H plot as well as by various other order / disorder predictors.

Protein 2 (PDB ID: 1L2P) is a coiled-coil, which in fact should not have been classified as globular, structured protein. We need to check our filter that is supposed to yield globular, monomeric proteins. Coiled-coil proteins often exhibit lower complexity than globular proteins[184], as is the case for this protein, and many coiled-coil proteins are predicted to be disordered[185]. Indeed, much this protein is predicted to be markedly disordered by several of our PONDR® predictors.

Protein 3 (PDB ID: 1X3O) was crystallized using dioxane as the precipitant. Adding low dielectric co-solvents often causes disorder-to-order shifts in protein structure, likely by promoting backbone hydrogen bond formation and by reducing the ionization of the acidic and basic side chains. We cannot find information whether, for example, this protein contains substantial disorder that becomes structured as dioxane or another dieclectric-lowering, water miscible co-solvent is added. Furthermore, protein 3 contains a serine modified with a large moiety that is rather hydrophobic, uncharged, and contains several hydrogen bond donors and acceptors, and thus could very likely induce formation of structure. Interestingly, different PONDR® prediction methods behave quite differently with protein 3. In particular, a large part of this protein, approximately residue 20 to 65 is predicted to be disordered by VLXT and VL3, while this same region is predicted structured by VSL2 and PONDR®-FIT. It is not surprising that our C-H plot predictor is in agreement with VLXT and VL3, since they both rely heavily on

hydropathy. Another interesting observation is that regions predicted to be ordered by VSL2 and PONDR®-FIT have scores fairly close to the order/disorder boundary region, suggesting very modest stability. Given such lower confidence scores and the disagreement among predictors, it is quite possible that dioxane and the modified serine assist in the process of folding as suggested above.

Protein 4 (PDB ID: 1CEI) is misclassified as disordered by the C-H plot method. However, all 4 PONDR® predictors correctly predict this protein to have substantial structured regions. We cannot find any reason to explain this classification error made by the C-H plot.

For protein 5, the first of the disordered proteins predicted to be structured, and its circular dichroism (CD) spectrum shows that this protein contains a substantial amount of secondary structure. Thus, it is not likely to be entirely disordered as indicated by DisProt. Indeed, PONDR® predictors indicate substantial regions of structure for this protein. Its classification as disordered in DisProt was due mainly to author comments. Based on these observations, its fully disordered status is being re-evaluated.

Protein 6 is classified as fully disordered in DisProt on the basis of its sensitivity to proteolysis along with author comments. Unlike other experimental procedures such as CD, sensitivity to proteolysis is a less confident approach in the determination of disordered proteins. PONDR® predictions also suggest that it has abundant structured regions.

Protein 7 is sensitive to proteolysis only in the absence of its scaffold protein *Agrobacterium* autoinducer (AAI). When AAI is present, protein 7 folds onto the scaffold and becomes insensitive to proteolysis. This protein has the characteristics of

structured proteins.All four PONDR® methods predict large structured regions for this protein. In possibly related studies[186,187], the fully structured thioredoxin protein was cleaved by proteases in separate experiments at two different single sites; for each single cleavage, the two fragments were separated. When alone, three of the four fragments did not aggregate significantly and behaved like IDPs. When the pairs of fragments from the same single cleavage were combined, the fragments mutually and correctly folded into an intact structure.[89,90] These thioredoxin (Trx) fragments have the appropriate balance of hydrophobic and hydrophilic amino acids for folding, as they located at the structural side of C-H plot, but they evidently lack an appropriately folded state with steric fitting of the various side chains, a shortfall that could potentially retard folding. From the folding funnel perspective, when alone these fragments evidently lack a deep well. From these earlier experiments and theories, we speculate that Protein 7 has an appropriate sequence for folding, but does not have a low-energy, self-folding structure. However, Protein 7 folds quite well when it encounters the surface of its scaffold protein.

For protein 8, a sub-region, namely residues 1-96, is reported to be disordered. However, the entire sequence of protein 8 was mistakenly put into the DisProt dataset. Trimming protein 8 and leaving only residue 1-96 yields a correct disordered prediction for this region. Thus, this error was due to an annotation error in DisProt, and this annotation error will be corrected.

Proteins 9 and 13 appear to be molten globules in their native states. Molten globules are characterized by backbones having semi-stable secondary structure but with loss of rigid packing by the side chains thus leading to fluctuating, unstable tertiary structure[188,189]. Since native molten globules would likely have values for their net

charge and hydropathy similar to those for folded proteins, elsewhere we have previously suggested that native molten globules would likely appear to be structured proteins on the C-H plot[5,9]. These data support those earlier suggestions.

Proteins 10-12 were all indicated to be disordered by random-coil-type CD spectra, while at the same time these proteins were observed to contain relatively high fractions of hydrophobic residues[190,191]. As would be expected for such amino acid compositions, all of these proteins are predicted to be mostly structured by the various PONDR®s. While having sufficient numbers of hydrophobic groups is necessary for protein folding, these results suggest that this feature alone is not sufficient, perhaps for the same reasons discussed above for Protein 7. Therefore, we speculate that, like Protein 7, each of these proteins would readily fold into 3D structure in the presence of the appropriate partner.

2.9 Discussion

In our work, we show that the performance of C-H plot can be improved significantly by introducing a new hydropathy scale. This new IDP-Hydropathy scale boosts the predictor's F-score from an original value of 0.68 to the 26% higher value of 0.86. This new scale also performs considerably better than four existing disorder propensity-based scales and other 531 amino acid scales obtained from the AAIndex. A correlation study and a heat map show that this scale is indeed highly associated with amino acid hydropathy.

2.9.1 Disorder is harder to predict

In all of our tested scales, including IDP-Hydropathy, disorder prediction accuracy is much lower than the order prediction accuracy. We hypothesize that this results from the existence of many small regions with increased order propensity that are located inside larger disordered regions. Despite of these short structure-prone elements, these regions are still experimentally shown to be mostly disordered. These regions with increased order propensity are likely to be functional domains within the disordered proteins. Molecular recognition features (MoRFs) that bind to specific protein or nucleic acid partners are one type of disorder-based functional regions. When not bound to a partner, such MoRF segments remain disordered and flexible. Upon binding, they typically become structured, adopting an ordered conformation that depends on the template provided by the binding partner. Their flexibility in the unbound state allows them change their shape as needed to fit onto the surfaces of different and distinct partners[75,81,162,192].

2.9.2 Error analysis

Judging from our error analysis in Table 10 and in supplementary information, the underlining theory for C-H plot is simple yet powerful. Many of the mis-classified examples in this study resulted from errors of the process of dataset collection and annotations. Only 2 of the strongly misclassified structured proteins is clearly not related to annotation errors or potential disorder-to-order transitions induced by experimental conditions, and one of these proteins is an extremely unusual 6-stranded PPII helical bundle. Furthermore, only three of the misclassified disordered proteins strongly violate

the overall hypothesis. These three are significantly hydrophobic with modest net charges, yet are shown to be disordered by CD. We speculate that these proteins do in fact become structured upon association with the appropriate protein partners as is observed for one protein having a similarly modest net charge along with a significant hydrophobicity.

Table 10 Missclassified examples and brief comments

Index	PDB/DisProt Entry	Brief comments
1	2PNE[193]	Antifreeze protein. 6 Polyproline II helix with 2 disulfide bonds to stabilize
2	1L2P[194]	Coil-coil structure.
3	1X3O	Dioxane used as precipitant. Contains modified O-(pantetheine 4'-phosphoryl)serine
4	1CEI[195]	High net charge.
5	DP00714[196]	CD shows substantial secondary structures.
6	DP00723[197]	Sensitivity to proteolysis is the sole evidence of disorder.
7	DP00198[198]	This protein is sensitive to proteolysis in the absence of a scaffold protein.
8	DP00069[199]	Only residue 1-96 is used for CD. Our predictor correctly predicts this sub-region.
9	DP00193[200]	Molten globule protein with substantial secondary structures.
10	DP00626[190]	High proportions of hydrophobic amino acids(Val, Ile, Leu). CD supports high amount of disorder
11	DP00288[191]	High hydrophobicity. Shown to be highly disordered by CD.
12	DP00626_C001[190]	High proportion of hydrophobic amino acids. Disorder supported by CD
13	DP00465[201]	Molten globule protein with substantial secondary structures.

The index of each example corresponds to the index in Figure. 8, with 1-4 being structured proteins misclassified as disordered and 5-13 being disordered proteins misclassified as structured.
CD: circular dichroism
The citation for each protein is indicated by its superscript number. A manuscript describing the 1X3O structure is yet to be published.

2.9.3 Limitations of composition based IDP predictors

Apart from the hydrophobic effect, hydrogen bonding within the protein is equally important to protein stability[171]. The interactions of neighboring residues are not explicitly studied in the pure compositional based predictor. Moreover, in the study of protein secondary structure prediction, long-range residue interaction is considered as one of the accuracy limitation[202]. Also, the 'chameleon' sequences, which do not have a strong structural preference, adopt different secondary structures depending on their surrounding sequences[203]. It is possible that the above features for globule proteins are also applicable to disordered proteins or regions, all of which are not fully addressed in the simple compositional based approach.

2.9.4 Application and future work

This new scale, IDP-Hydropathy derived from entirely disordered and structured proteins, is a very handy tool because of its simplicity and prediction power. This new scale should improve other disorder predictors that use hydropathy as one of the input features. We are looking forward to the incorporation of this new scale into a per-residue predictor based on these same principles.

The original hydrophobicity scale of Nozaki and Tanford[174] was developed with the purpose of understanding the relative importance of different amino acids to protein folding. The IDP-hydropathy scale developed here is based on sets of sequences that fold into 3D structure as compared to collections of sequence that don't fold, using the C-H plot as the classifier. Thus, to a very significant degree, IDP-hydropathy fulfills the

intent of the original scale by providing a measure of how the various amino acids contribute to protein folding by means of their hydropathy values.

2.10 Summary

The earliest whole protein order / disorder predictor (Uversky et al., Proteins, 41: 415-427 (2000)), herein called the charge-hydropathy (C-H) plot, was originally developed using the Kyte & Doolittle (1982) hydropathy scale, which was chosen due to its frequent use. In an effort to improve the performance of the CH-plot, we tested alternative hydropathy scales, with the finding that the Guy (1985) hydropathy scale was the best of the tested hydropathy scales for separating structured proteins and intrinsically disordered proteins (IDPs) on the C-H plot. Next, we developed a new scale, named IDP-hydropathy, which further improves the discrimination between structured proteins and IDPs. Applying the C-H plot to one particular dataset containing 109 IDPs and 563 non-homologous fully structured proteins, the Kyte & Doolittle (1982) hydropathy scale, the Guy (1985) hydropathy scale, and the IDP-hydropathy scale gave balanced two-state classification accuracies of 79%, 83%, and 91%, respectively, indicating a very substantial overall improvement is obtained by using different hydropathy scales. A study comparing the IDP-hydropathy with 554 amino acid indices shows that IDP-hydropathy is strongly correlated with other hydropathy scales or with scales that are in turn highly correlated with hydropathy scales, thus suggesting that IDP-hydropathy has only small contributions from amino acid properties other than hydropathy. We suggest that this scale would likely be the best one generally to use for predictors of protein disorder.

3. Per-residue Charge-Hydropathy Disorder Prediction

Our previously developed IDP-Hydropathy scale optimizes disorder prediction for entire protein sequences. However, many proteins have dual disorder/order states, in which one protein has both disordered regions and ordered regions[9]. In addition, experimentally identified IDP regions seldom have uniform sequence characteristics regarding their disorder tendencies. MoRFs, for example, typically display hydrophobic residues inside a stretch of mostly hydrophilic plus proline residues having a strong tendency to be a disordered region[74,75]. The intermixed disordered and ordered regions and the intermixed order and disorder tendencies in disordered regions are both crucial to the protein's function. The hydrophobic regions can readily bind to the surface of another protein while the disorder region can adapt to the conformation of that partner, as described in the introduction and as illustrated by a MoRF's one-to-many binding. The typical example of such flexible binding is provided by the N and C termini of the p53 protein[192]. Both disordered termini of p53 use multiple conformations to bind to a large number of specific partners.

Hence, predicting the per-residue disordered or ordered state of a protein or predicting changes in the order or disorder tendencies can reveal important information regarding protein function. Many per-residue disorder predictors have been developed, including the PONDR family (VLXT[61], VSL3[63], VSL2[54], PONDR-FIT[59]), IUPred[52], Disopred[167], SPINE-D[53] and many more. They have been built based on different hypothesis and by means of different algorithms.

Just as for protein function and structure predictions, evolutionary comparisons, such as by means of multiple sequence alignments, make very powerful contributions to

many of these methods[53,54,167]. However, such sequence comparisons are very time consuming, with computations over entire proteomes taking weeks or more to complete. Therefore, many methods (such as VSL2B) have been developed to be both accurate and fast without the usage of evolutionary information.

Most of the per-residue disorder predictors have been built using supervised machine learning models with many inputs. In recent years, many complex and integrated methods have been developed that use very large feature spaces. However, as shown in recent the CASP exercises[66–68], the accuracy of disorder prediction is more likely to be dependent on the difficulty of the examples being predicted than on the power of the predictors being used. The feature information used for prediction is often so complex as to not provide any useful insight regarding the IDP regions under study.

Going back to the basics, among the most intuitive and fundament features underlying the folding of a given protein are its hydropathy and charge[4,9]. In this study, we built a simple, fast yet accurate and informative linear model that uses these hydropathy and charge features as the only inputs. More specifically, we optimized the protein hydropathy scale on protein local regions with linear Support Vector Machine (SVM)[146,177]. Then we used this scale to calculate local hydropathy. Together with charge, the local hydropathy are used to build the linear model to predict protein disorder.

Three different hydropathy scales were optimized individually for the N-terminus, C-terminus, and internal regions of regions of sequence. Such an approach was found empirically to be superior to the use of a single hydropathy scale. Each scale then provides the basis for per-residue charge- and hydrpathy-based disorder prediction for the residues in the three respective N-, C- and internal regions.

Our predictor, which we are calling FoldIndex II, performs comparably to much more complex recently-developed per-residue disorder predictors. More importantly, in addition to order or disorder prediction, our algorithm also outputs regional hydropathy and net charge. Comparing to the original FoldIndex, which also uses linear function of charge and hydropathy for prediction, our predictor is much more accurate. In addition, our predictor includes predictions to the N- and C- protein termini, regions that they are skipped in the original FoldIndex.

3.1 Disorder is not evenly distributed alone the sequence

To examine the effect of residue position on residue disorder or order status, we need to first define the number of residues relative to the N- or C-terminus. To compare these predictors with our previous work[61], we chose 21 residues to be the window size for the order-disorder prediction with the prediction applied to the residue at the center of the window. In our training data, there are 6839 partially overlapped disordered windows at N terminus, 83268 at internal region, and 6839 at the C terminus.

3.1.1 Disorder is enriched at the N/C terminus

In Figure 11, the normalized number of disordered residues at N, C, and internal regions are calculated and their distributions are plotted as a boxplot. Disorder is not evenly distributed at these three regions, with a significant p-value of 3.79e-69 from ANOVA analysis. In fact, the N and C termini are much more enriched of disordered residues.

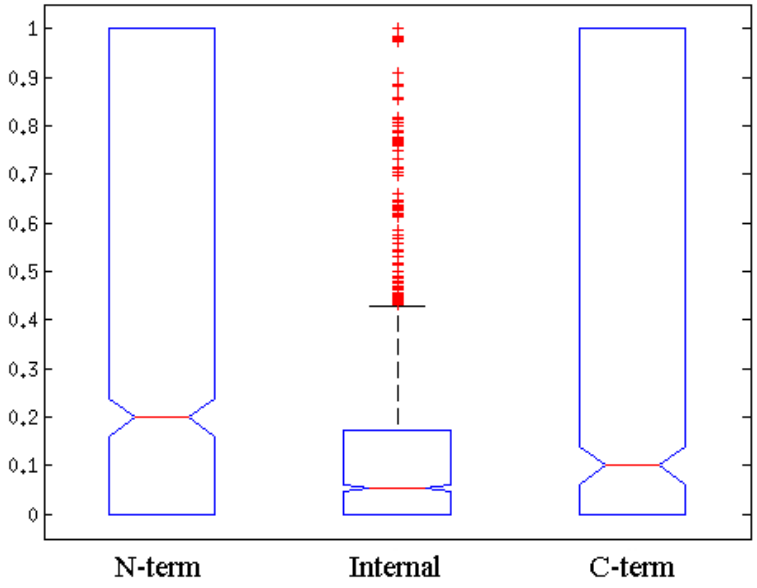


Figure 11 Boxplot for normalized number of disordered at N/C terminus and internal regions. The red lines are the median, meaning 50% of data are greater than this value. The upper and lower blue lines are the upper and lower quartile, respectively. There are 25% of data are greater or less than them, respectively. The black line is the maximum, which is the greatest value excluding outliers. The short red lines are the outliers, meaning that their values are more than 3/2 times of the value at upper quartile.

3.1.2 Disorder composition different significantly among the N and C termini and the internal regions

Even though disorder is more enriched at the termini, the amino acid composition indicating disorder could still remain the same for these regions. So we summarized the composition for disordered residues at N and C termini and internal region, respectively, and carried out correlation analysis as shown in Table 11. Interestingly, in Table 11A, the correlation coefficient is low between internal regions and N-terminus, and their p-value indicates that they are not correlated. On the other hand, the N-terminus disorder-indicating amino acid composition correlates to that of the C-terminus with an r of 0.63. Its p-value, 0.0086, is at the edge of being significant. Meanwhile, the C-terminus is related to internal region with an r of 0.68, and its p-value is 0.0027. Again, this p-value is close to the significance threshold.

In the composition bar plot, which uses the composition of internal region as the baseline, the content of many residues in N-terminus is much higher or lower than the corresponding contents observed for internal regions. Because the initiation codon encodes methionine, methionine is particularly enriched at the N terminus. On the otherhand, the common use of polyhistidine-tag on the N or C terminus of protein sequences for purification likely contributes to the sharp peaks of histidine composition. Moreover, there are residues, such as P, T, and V, that are enriched or depleted in both N and C terminus; and there are also residues, such as D, E, Q, and K, that are enriched or depleted only in N-terminus. Interestingly, most residues in N and C terminus that vary significantly from internal region are either both enriched or both depleted.

In general, the amino acid compositions of some residues in N/C terminus and internal regions for disordered proteins are significantly different. However, some residues are consistently enriched or depleted across all three locations. The N-terminus is very different from the internal regions, with high p-value indicating no correlation. The correlation coefficient between N-term/C-term, and the correlation between C-term/internal are relatively high, but with interesting p-value. This might be explained by the fact that mostly, residue contents among these regions are similar, causing r as high as over 0.6, but with some exceptions and raises p-value.

To account for the differences in the compositions of some residues among three regions, we decided to divide the sequence into three regions as N/C terminus and internal regions, and optimizes the hydropathy scale individually.

Table 11 Amino acid composition correlation coefficient (A) and p-values (B) at N/C terminus and internal regions

	N term	Internal	C term
N term			
Internal	0.31		
C term	0.63	0.68	

Table 11 A. Correlation coefficient

	N term	Internal	C term
N term			
Internal	0.56		
C term	0.0086	0.0027	

Table 11 B. p-value after Bonferroni correction

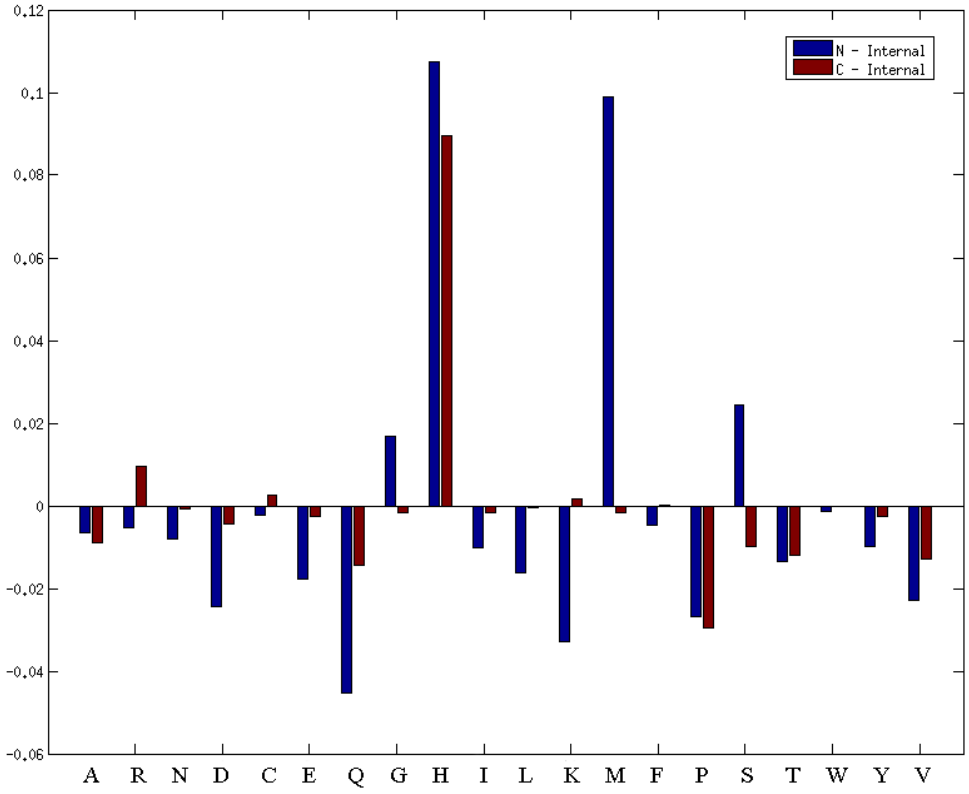


Figure 12 Amino acid compositions bar graph at N-terminus, internal regions, and C-terminus

3.2 Optimizing hydropathy scale for improved disorder prediction

We used linear SVM to optimize the hydropathy scale for disorder prediction. In particular, protein sequences are divided into three regions, N-termnius, C-terminus and internal regions. Amino acid composition was calculated for protein sequence at a window size of 21. Spacer characters are added at the N and C ends of protein sequences to fill any window that extends outside of the protein chain. Charge of the protein sequence at a specific window is calculated as the sum of number of arginine and lysine, minus the sum of number of aspartic acid and glutamic acid. Then the charge is normalized by dividing the length of window size. Then we take the absolute value of the normalized charge, which also make the charge attribute non-redundant with the composition attributes.

To obtain the hydropathy scale that optimize disorder prediction, linear SVM were applied to N-terminus, C-terminus and internal regions separately. We take the coefficient for each amino acid composition, and normalize them to be between -1 and 1. In the end, the prediction accuracies are shown in table 12.

Table 12 Accuracy matrices for disorder prediction on N/C terminus and internal regions.

	Balanced Accuracy	Sensitivity	Specificity	F-score	MCC	AUC
N-term	0.75	0.74	0.76	0.75	0.49	0.82
Internal	0.74	0.73	0.75	0.61	0.44	0.81
C-term	0.72	0.72	0.72	0.70	0.44	0.80

The predictor is optimized to achieve the best balanced sensitivity and specificity, and minimize the difference between sensitivity and specificity. Note that the data are heavily biased towards ordered proteins, especially for the internal regions. This is the likely explanation for the observation that, while sensitivity and specificity are similar for all three regions, the F-score for the internal region is much lower than both the N and C termini. The predictor can compensate the disorder prediction power by over-predicting disordered residues, therefore boost its sensitivity. Meanwhile, the number of over-predicted disorder mistakes is small compared to the large number of ordered residues, and thus maintaining the specificity. The calculation of F-score involves positive predictive value (PPV), which measures the proportion of true positives versus predicted positives. The over-prediction of disorder in the internal region is reflected in PPV, and thus lowering its F-score. Despite of this, sensitivity and specificity is a common measurement of prediction power, and they are straightforward and meaningful. So we still use the traditional measurement.

The receiver operating characteristic (ROC) plot (Figure 13) plots the true positive rate against the false positive rate at varying thresholds of the linear function output. The dashed diagonal line represents random guesses. Predictors for all three regions perform fairly well above the diagonal line, and have similar AUC (area under curve).

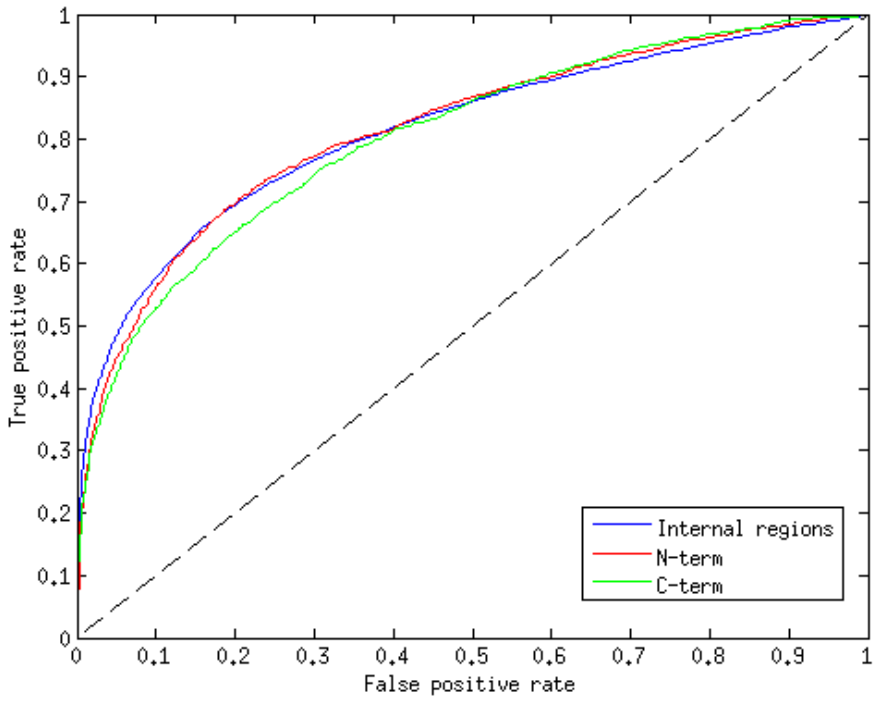


Figure 13 ROC curve for disorder prediction on internal regions, N-terminus and C-terminus, respectively.

3.3 Benchmark the per-residue IDP scale

The performance of the optimized scale is tested against three different predictors as shown in Figure 14. FoldIndex is also a per-residue disorder predictor, which also uses hydropathy and charge as input features. Note that since FoldIndex do not have spacer characters, it starts prediction at the residues that are within half of the window size from the beginning, and ends prediction half of the window size from the end of the sequence. Therefore, we applied the same procedure in this test. FoldIndex uses Kyte-Doolittle hydropathy scale to calculate hydropathy. Its ROC curve is shown as magenta in Figure 14. We also trained a second predictor using the Guy hydropathy scale, since it showed best prediction performance for entirely disordered proteins (green dashed line in Figure 14). Our optimized predictor outperforms both of them (blue line in Figure 14).

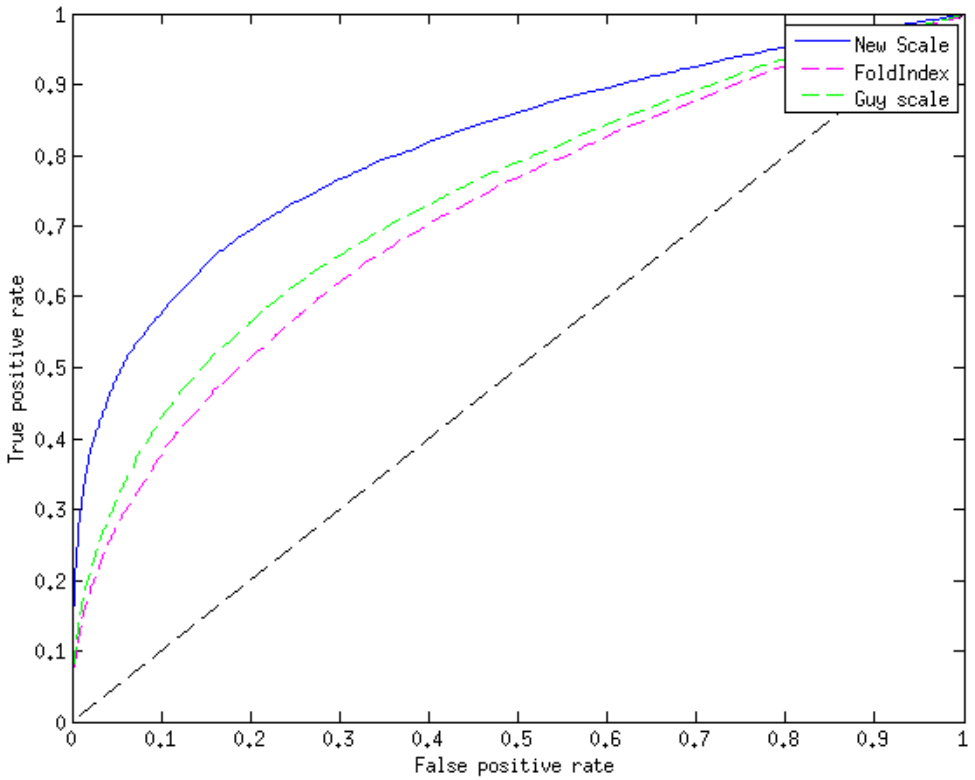


Figure 14 ROC curve for different predictors. Blue line is the new optimized predictor. Magenta line represents the ROC curve for FoldIndex predictor, red line is the predictor optimized with Kyte-Doolittle hydropathy scale, and the green line is the predictor optimized with Guy hydropathy scale.

3.4 Further improvements

During our development of the per-residue IDP-hydropathy scale, we found that different window sizes drastically affect the prediction power. We also found that smoothing the output by a typical window size could improve the prediction. However, our dataset is dominated with ordered protein examples. Moreover, one part of the dataset is comprised of protein examples from Disport, which contains consecutive long disordered regions. Therefore, using longer input window size and an output smoothing window is likely to enforce ordered predictions in general, and disordered predictions over long disordered regions. Short disordered segments are likely to be missed, even though they may bear important biological meanings.

The impact of the likely explanation for the observation that long and short disorder regions prediction has been studied and resulted in the development of VSL2 predictor. However, at that time, there were only 153 sequences from Disprot. Now, there are 694 disordered sequences from Disprot alone. The conclusion that short and long disordered residues have different amino acid preferences for order and disorder is likely still valid, but the construction of length-dependent disorder predictors needs to be re- addressed the the larger datasets currently available.

Most current disorder predictors use compositional and evolutionary features. As shown in recent CASP, the disorder prediction is more likely dependent on the difficulty of examples than their prediction power. As more and more data become available, new attributes, such as adding inputs based on amino acid sequence patterns, should further improve the prediction power.

3.5 Summary

Here we describe a per-residue C-H IDP predictor called FoldIndex II that used a linear Support Vector Machine with just normalized net charge and normalized hydropathy as the only inputs. Using our recently developed method for optimizing a hydropathy scale for prediction of disorder using just normalized net charge and normailized hydropathy, three hydropathy scales are optimized individually, one for the amino terminal region, one for the internal region and one for the carboxyl terminal region. These three scales along with normalized net charge then provide the inputs for the Support Vector Machine. FoldIndex II substantially outperforms the original FoldIndex algorithm as shown by receiver operating characteristic curve analysis.

REFERENCES

1. Dunker AK, Brown CJ, Obradovic Z. Identification and functions of usefully disordered proteins. *Adv. Protein Chem.* 2002;62:25–49.
2. Dunker AK, Garner E, Guilliot S, et al. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 1998;473–484.
3. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 1999;293(2):321–331.
4. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins Struct. Funct. Bioinforma.* 2000;41(3):415–427.
5. Dunker AK, Lawson JD, Brown CJ, et al. Intrinsically disordered protein. *J. Mol. Graph. Model.* 2001;19(1):26–59.
6. Dunker AK, Babu MM, Barbar E, et al. What’s in a name? Why these proteins are intrinsically disordered. *Intrinsically Disord. Proteins*. 2013;1(1):e24157.
7. Jirgensons B. Classification of proteins according to conformation. *Makromol. Chem.* 1966;91(1):74–86.
8. Baranger M. Chaos, Complexity, and Entropy: a Physics Talk for Non-Physicists. *Wesley. Univ. Phys. Dept Colloq.* 2001;
9. Uversky VN, Dunker AK. Understanding protein non-folding. *Biochim. Biophys. Acta*. 2010;1804(6):1231–1264.
10. Chouard T. Structural biology: Breaking the protein rules. *Nature*. 2011;471(7337):151–153.
11. Fischer E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte Dtsch. Chem. Ges.* 1894;27(3):2985–2993.
12. Edsall JT. Hsien Wu and the First Theory of Protein Denaturation (1931). *Adv. PROTEIN Chem. 46 1 Protein Stab.* 1994;
13. Blake CC, Koenig DF, Mair GA, et al. Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. *Nature*. 1965;206(4986):757–761.
14. Bernstein FC, Koetzle TF, Williams GJ, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977;112(3):535–542.
15. Holt C, Sawyer L. Caseins as rheomorphic proteins: interpretation of primary and secondary structures of the αS1-, β- and κ-caseins. *J. Chem. Soc. Faraday Trans.* 1993;89(15):2683–2692.
16. Holt C, Wahlgren NM, Drakenberg T. Ability of a beta-casein phosphopeptide to modulate the precipitation of calcium phosphate by forming amorphous dicalcium phosphate nanoclusters. *Biochem. J.* 1996;314 (Pt 3):1035–1039.
17. Sigler PB. Transcriptional activation. Acid blobs and negative noodles. *Nature*. 1988;333(6170):210–212.
18. Cavanagh J, Fairbrother WJ, III AGP, Skelton NJ, Rance M. Protein NMR Spectroscopy: Principles and Practice. Academic Press; 2010.

19. Muchmore SW, Sattler M, Liang H, et al. X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. *Nature*. 1996;381(6580):335–341.
20. Jensen MR, Ruigrok RW, Blackledge M. Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr. Opin. Struct. Biol.* 2013;23(3):426–435.
21. Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci. U. S. A.* 1996;93(21):11504–11509.
22. Daughdrill GW, Chadsey MS, Karlinsey JE, Hughes KT, Dahlquist FW. The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28. *Nat. Struct. Biol.* 1997;4(4):285–291.
23. Allison JR, Varnai P, Dobson CM, Vendruscolo M. Determination of the free energy landscape of alpha-synuclein using spin label nuclear magnetic resonance measurements. *J. Am. Chem. Soc.* 2009;131(51):18314–18326.
24. Choy WY, Forman-Kay JD. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* 2001;308(5):1011–1032.
25. Gehring WJ, Qian YQ, Billeter M, et al. Homeodomain-DNA recognition. *Cell*. 1994;78(2):211–223.
26. Martin AJM, Walsh I, Tosatto SCE. MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. *Bioinforma. Oxf. Engl.* 2010;26(22):2916–2917.
27. Larsson G, Martinez G, Schleucher J, Wijmenga SS. Detection of nano-second internal motion and determination of overall tumbling times independent of the time scale of internal motion in proteins from NMR relaxation data. *J. Biomol. NMR*. 2003;27(4):291–312.
28. Dedmon MM, Patel CN, Young GB, Pielak GJ. FlgM gains structure in living cells. *Proc. Natl. Acad. Sci. U. S. A.* 2002;99(20):12681–12684.
29. Selenko P, Wagner G. Looking into live cells with in-cell NMR spectroscopy. *J. Struct. Biol.* 2007;158(2):244–253.
30. Binolfi A, Theillet F-X, Selenko P. Bacterial in-cell NMR of human α-synuclein: a disordered monomer by nature? *Biochem. Soc. Trans.* 2012;40(5):950–954.
31. Li C, Charlton LM, Lakkavaram A, et al. Differential dynamical effects of macromolecular crowding on an intrinsically disordered protein and a globular protein: implications for in-cell NMR spectroscopy. *J. Am. Chem. Soc.* 2008;130(20):6310–6311.
32. Barnes CO, Pielak GJ. In-cell protein NMR and protein leakage. *Proteins*. 2011;79(2):347–351.
33. Hura GL, Budworth H, Dyer KN, et al. Comprehensive macromolecular conformations mapped by quantitative SAXS analyses. *Nat. Methods*. 2013;10(6):453–454.
34. Hegde ML, Tsutakawa SE, Hegde PM, et al. The disordered C-terminal domain of human DNA glycosylase NEIL1 contributes to its stability via intramolecular interactions. *J. Mol. Biol.* 2013;425(13):2359–2371.
35. Varadi M, Kosol S, Lebrun P, et al. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.* 2013;gkt960.

36. Smyth E, Syme CD, Blanch EW, et al. Solution structure of native proteins with irregular folds from Raman optical activity. *Biopolymers*. 2001;58(2):138–151.
37. Adler AJ, Greenfield NJ, Fasman GD. Circular dichroism and optical rotatory dispersion of proteins and polypeptides. *Methods Enzymol.* 1973;27:675–735.
38. Johnson WC. Secondary Structure of Proteins Through Circular Dichroism Spectroscopy. *Annu. Rev. Biophys. Biophys. Chem.* 1988;17(1):145–166.
39. Kelly SM, Price NC. The application of circular dichroism to studies of protein folding and unfolding. *Biochim. Biophys. Acta*. 1997;1338(2):161–185.
40. Provencher SW, Glöckner J. Estimation of globular protein secondary structure from circular dichroism. *Biochemistry (Mosc.)*. 1981;20(1):33–37.
41. Fontana A, Zambonin M, Polverino de Laureto P, et al. Probing the conformational state of apomyoglobin by limited proteolysis. *J. Mol. Biol.* 1997;266(2):223–230.
42. Vucetic S, Obradovic Z, Vacic V, et al. DisProt: a database of protein disorder. *Bioinforma. Oxf. Engl.* 2005;21(1):137–140.
43. Xie, Arnold, Romero, et al. The Sequence Attribute Method for Determining Relationships Between Sequence and Protein Disorder. *Genome Inform. Workshop Genome Inform.* 1998;9:193–200.
44. Romero P, Obradovic Z, Kissinger C, Villafranca JE, Dunker AK. Identifying disordered regions in proteins from amino acid sequence. *Int. Conf. Neural Networks1997*. 1997;1:90–95 vol.1.
45. Romero, Obradovic, Dunker. Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. *Genome Inform. Workshop Genome Inform.* 1997;8:110–124.
46. Dosztányi Z, Csizmók 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(4):827–839.
47. Jacob C, Giles GI, Giles NM, Sies H. Sulfur and selenium: the role of oxidation state in protein structure and function. *Angew. Chem. Int. Ed Engl.* 2003;42(39):4742–4758.
48. Gallogly MM, Mieyal JJ. Mechanisms of reversible protein glutathionylation in redox signaling and oxidative stress. *Curr. Opin. Pharmacol.* 2007;7(4):381–391.
49. Campen A, Williams RM, Brown CJ, et al. TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. *Protein Pept. Lett.* 2008;15(9):956–963.
50. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of Machine Learning. The MIT Press; .
51. Romero P, Obradovic Z, Li X, et al. Sequence complexity of disordered protein. *Proteins Struct. Funct. Bioinforma.* 2001;42(1):38–48.
52. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*. 2005;21(16):3433–3434.
53. Zhang T, Faraggi E, Xue B, et al. SPINE-D: Accurate Prediction of Short and Long Disordered Regions by a Single Neural-Network Based Method. *J. Biomol. Struct. X00026 Dyn.* 2012;29(4):799–813.
54. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*. 2006;7:208.

55. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*. 2005;61 Suppl 7:176–182.
56. Williams RJ. The conformational mobility of proteins and its functional significance. *Biochem. Soc. Trans.* 1978;6(6):1123–1126.
57. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 1982;157(1):105–132.
58. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, et al. FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*. 2005;21(16):3435–3438.
59. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: A meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta BBA - Proteins Proteomics*. 2010;1804(4):996–1010.
60. Xue B, Oldfield CJ, Dunker AK, Uversky VN. CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Lett.* 2009;583(9):1469–1474.
61. Peng K, Vucetic S, Radivojac P, et al. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinform. Comput. Biol.* 2005;3(1):35–60.
62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol.* 1990;215(3):403–410.
63. Obradovic Z, Peng K, Vucetic S, et al. Predicting intrinsic disorder from amino acid sequence. *Proteins*. 2003;53 Suppl 6:566–572.
64. Sickmeier M, Hamilton JA, LeGall T, et al. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.* 2007;35(Database):D786–D793.
65. CASP5. Proceedings of the 5th Meeting on the Critical Assessment of Techniques for Protein Structure Prediction. 1-5 December 2002, Asilomar, California, USA. *Proteins*. 2003;53 Suppl 6:333–595.
66. Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A. Evaluation of disorder predictions in CASP9. *Proteins Struct. Funct. Bioinforma.* 2011;79(S10):107–118.
67. Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K. Assessment of protein disorder region predictions in CASP10. *Proteins*. 2013;
68. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8. *Proteins Struct. Funct. Bioinforma.* 2009;77(S9):210–216.
69. Di Domenico T, Walsh I, Martin AJM, Tosatto SCE. MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinforma. Oxf. Engl.* 2012;28(15):2080–2081.
70. Oates ME, Romero P, Ishida T, et al. D^2P^2: database of disordered protein predictions. *Nucleic Acids Res.* 2013;41(Database issue):D508–516.
71. Xue B, Dunker AK, Uversky VN. 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(2):137–149.
72. Cheng Y, Oldfield CJ, Meng J, et al. Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry (Mosc.).* 2007;46(47):13468–13477.

73. Sandhu KS, Dash D. Conformational flexibility may explain multiple cellular roles of PEST motifs. *Proteins*. 2006;63(4):727–732.
74. Vacic V, Oldfield CJ, Mohan A, et al. Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.* 2007;6(6):2351–2366.
75. Mohan A, Oldfield CJ, Radivojac P, et al. Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* 2006;362(5):1043–1059.
76. Wong ETC, Na D, Gsponer J. On the importance of polar interactions for complexes containing intrinsically disordered proteins. *PLoS Comput. Biol.* 2013;9(8):e1003192.
77. Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* 2008;33(1):2–8.
78. Mittag T, Kay LE, Forman-Kay JD. Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit. JMR*. 2010;23(2):105–116.
79. Onnela J-P, Saramäki J, Hyvönen J, et al. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U. S. A.* 2007;104(18):7332–7336.
80. Choromański K, Matuszak M, Miękisz J. Scale-Free Graph with Preferential Attachment and Evolving Internal Vertex Structure. *J. Stat. Phys.* 2013;151(6):1175–1183.
81. Hsu W-L, Oldfield CJ, Xue B, et al. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Sci.* 2013;22(3):258–273.
82. Hasty J, Collins JJ. Protein interactions. Unspinning the web. *Nature*. 2001;411(6833):30–31.
83. Oldfield CJ, Meng J, Yang JY, et al. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics*. 2008;9(Suppl 1):S1.
84. Hsu W, Oldfield C, Faraggi E, Xue B, Huang F. In Press. Characterizing the binding profiles of many-to-one intrinsically disordered protein complexes. *FEBS J.* .
85. Buljan M, Chalancon G, Dunker AK, et al. Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. *Curr. Opin. Struct. Biol.* 2013;23(3):443–450.
86. Colak R, Kim T, Michaut M, et al. Distinct types of disorder in the human proteome: functional implications for alternative splicing. *PLoS Comput. Biol.* 2013;9(4):e1003030.
87. Ellis JD, Barrios-Rodiles M, Colak R, et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell*. 2012;46(6):884–892.
88. Vuzman D, Levy Y. Intrinsically disordered regions as affinity tuners in protein-DNA interactions. *Mol. Biosyst.* 2012;8(1):47–57.
89. Nadassy K, Wodak SJ, Janin J. Structural features of protein-nucleic acid recognition sites. *Biochemistry (Mosc.).* 1999;38(7):1999–2017.
90. Passner JM, Ryoo HD, Shen L, Mann RS, Aggarwal AK. Structure of a DNA-bound Ultrabithorax–Extradenticle homeodomain complex. *Nature*. 1999;397(6721):714–719.
91. Liu Y, Matthews KS, Bondos SE. Multiple intrinsically disordered sequences alter DNA binding by the homeodomain of the Drosophila hox protein ultrabithorax. *J. Biol. Chem.* 2008;283(30):20874–20887.

92. Wahl MC, Will CL, Lührmann R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell*. 2009;136(4):701–718.
93. Moore PB. How should we think about the ribosome? *Annu. Rev. Biophys.* 2012;41:1–19.
94. Lecompte O, Ripp R, Thierry J-C, Moras D, Poch O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.* 2002;30(24):5382–5390.
95. Timsit Y, Acosta Z, Allemand F, Chiaruttini C, Springer M. The role of disordered ribosomal protein extensions in the early steps of eubacterial 50 S ribosomal subunit assembly. *Int. J. Mol. Sci.* 2009;10(3):817–834.
96. Coelho Ribeiro M de L, Espinosa J, Islam S, et al. Malleable ribonucleoprotein machine: protein intrinsic disorder in the Saccharomyces cerevisiae spliceosome. *PeerJ*. 2013;1:e2.
97. Monod J, Wyman J, Changeux J-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* 1965;12(1):88–118.
98. Koshland DE, Némethy G, Filmer D. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits*. *Biochemistry (Mosc.)*. 1966;5(1):365–385.
99. Hilser VJ. An Ensemble View of Allostery. *Science*. 2010;327(5966):653–654.
100. Beckett D. Regulating transcription regulators via allostery and flexibility. *Proc. Natl. Acad. Sci.* 2009;106(52):22035–22036.
101. Motlagh HN, Hilser VJ. Agonism/antagonism switching in allosteric ensembles. *Proc. Natl. Acad. Sci.* 2012;
102. Kumar R, McEwan IJ. Allosteric modulators of steroid hormone receptors: structural dynamics and gene regulation. *Endocr. Rev.* 2012;33(2):271–299.
103. Ferreon ACM, Ferreon JC, Wright PE, Deniz AA. Modulation of allostery by protein intrinsic disorder. *Nature*. 2013;498(7454):390–394.
104. Ivanyi-Nagy R, Davidovic L, Khandjian EW, Darlix J-L. Disordered RNA chaperone proteins: from functions to disease. *Cell. Mol. Life Sci. CMLS*. 2005;62(13):1409–1417.
105. Tompa P, Csermely P. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* 2004;18(11):1169–1175.
106. Shewmaker F, Kerner MJ, Hayer-Hartl M, et al. A mobile loop order-disorder transition modulates the speed of chaperonin cycling. *Protein Sci. Publ. Protein Soc.* 2004;13(8):2139–2148.
107. Tompa P, Kovacs D. Intrinsically disordered chaperones in plants and animals. *Biochem. Cell Biol. Biochim. Biol. Cell.* 2010;88(2):167–174.
108. Bordo D, Argos P. Evolution of protein cores. Constraints in point mutations as observed in globin tertiary structures. *J. Mol. Biol.* 1990;211(4):975–988.
109. Jernigan RL, Kloczkowski A. Packing regularities in biological structures relate to their dynamics. *Methods Mol. Biol. Clifton NJ*. 2007;350:251–276.
110. Guy HR. Amino acid side-chain partition energies and distribution of residues in soluble proteins. *Biophys. J.* 1985;47(1):61–70.

111. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*. 1985;18(3):534–552.
112. Manavalan P, Ponnuswamy PK. Hydrophobic character of amino acid residues in globular proteins. *Nature*. 1978;275(5681):673–674.
113. Fauchere J-L, Pliska VE. Hydrophobic parameters pi of amino acid side chains from partitioning of N-acetyl-amino-acid amides. *Eur J Med Chem*. 1983;18:369–357.
114. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. *Science*. 1985;229(4716):834–838.
115. Sweet RM, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.* 1983;171(4):479–488.
116. Black SD, Mould DR. Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Anal. Biochem.* 1991;193(1):72–82.
117. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U. S. A.* 1981;78(6):3824–3828.
118. Bull HB, Breese K. Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys.* 1974;161(2):665–670.
119. Abraham DJ, Leo AJ. Extension of the fragment method to calculate amino acid zwitterion and side chain partition coefficients. *Proteins Struct. Funct. Bioinforma.* 1987;2(2):130–152.
120. Chothia C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 1976;105(1):1–12.
121. Roseman MA. Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J. Mol. Biol.* 1988;200(3):513–522.
122. J K Mohana Rao PA. A conformational preference parameter to predict helices in integral membrane proteins. *Biochim. Biophys. Acta*. 1986;869(2):197–214.
123. Janin J. Surface and inside volumes in globular proteins. *Nature*. 1979;277(5696):491–492.
124. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* 1984;179(1):125–142.
125. Tanford C. Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins. *J. Am. Chem. Soc.* 1962;84(22):4240–4247.
126. Welling GW, Weijer WJ, van der Zee R, Welling-Wester S. Prediction of sequential antigenic regions in proteins. *FEBS Lett.* 1985;188(2):215–218.
127. Wolfenden R, Andersson L, Cullis PM, Southgate CCB. Affinities of amino acid side chains for solvent water. *Biochemistry (Mosc.)*. 1981;20(4):849–855.
128. Kawashima S, Ogata H, Kanehisa M. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* 1999;27(1):368–369.
129. Kawashima S, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res.* 2000;28(1):374.
130. Kawashima S, Pokarowski P, Pokarowska M, et al. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 2008;36(Database issue):D202–205.

131. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. To be folded or to be unfolded? *Protein Sci. Publ. Protein Soc.* 2004;13(11):2871–2877.
132. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. *Proteins*. 1994;19(2):141–149.
133. Vacic V, Uversky VN, Dunker AK, Lonardi S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics*. 2007;8:211.
134. He H, Garcia EA. Learning from Imbalanced Data. *IEEE Trans Knowl Data Eng*. 2009;21(9):1263–1284.
135. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. *BMJ*. 1994;309(6947):102.
136. Heston TF. Standardizing predictive values in diagnostic imaging research. *J. Magn. Reson. Imaging JMRI*. 2011;33(2):505; author reply 506–507.
137. Gunnarsson RK, Lanke J. The predictive value of microbiologic diagnostic tests if asymptomatic carriers are present. *Stat. Med.* 2002;21(12):1773–1785.
138. Rao RB, Krishnan S, Niculescu RS. Data mining for improved cardiac care. *SIGKDD Explor Newsl*. 2006;8(1):3–10.
139. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000;16(5):412–424.
140. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* 1993;39(4):561–577.
141. Weiss G, Provost F. The Effect of Class Distribution on Classifier Learning: An Empirical Study. 2001.
142. Laurikkala J. Improving Identification of Difficult Small Classes by Balancing Class Distribution. *Proc. 8th Conf. AI Med. Eur. Artif. Intell. Med.* 2001;63–66.
143. Estabrooks A, Jo T, Japkowicz N. A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Comput. Intell.* 2004;20(1):18–36.
144. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* 2002;16:321–357.
145. Han H, Wang W-Y, Mao B-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *Adv. Intell. Comput.* 2005;878–887.
146. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2(3):27:1–27:27.
147. Shen L, Kim S, Qi Y, et al. Identifying Neuroimaging and Proteomic Biomarkers for MCI and AD via the Elastic Net. *Multimodal Brain Image Anal.* 2011;27–34.
148. Pearson K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philos. Trans. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character*. 1896;187:253–318.
149. Oldfield CJ, Cheng Y, Cortese MS, et al. Comparing and combining predictors of mostly disordered proteins. *Biochemistry (Mosc.)*. 2005;44(6):1989–2000.
150. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* 2002;323(3):573–584.

151. Xie H, Vucetic S, Iakoucheva LM, et al. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* 2007;6(5):1882–1898.
152. Vucetic S, Xie H, Iakoucheva LM, et al. 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.* 2007;6(5):1899–1916.
153. Xie H, Vucetic S, Iakoucheva LM, et al. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J. Proteome Res.* 2007;6(5):1917–1932.
154. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 1995;247(4):536–540.
155. Orengo CA, Michie AD, Jones S, et al. CATH--a hierarchic classification of protein domain structures. *Struct. Lond. Engl. 1993*. 1997;5(8):1093–1108.
156. Reeck GR, de Haën C, Teller DC, et al. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*. 1987;50(5):667.
157. Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. *Proteins Struct. Funct. Bioinforma.* 2003;52(4):573–584.
158. Le Gall T, Romero PR, Cortese MS, Uversky VN, Dunker AK. Intrinsic disorder in the Protein Data Bank. *J. Biomol. Struct. Dyn.* 2007;24(4):325–342.
159. Mohan A, Sullivan WJ Jr, Radivojac P, Dunker AK, Uversky VN. Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. *Mol. Biosyst.* 2008;4(4):328–340.
160. Xue B, Oldfield CJ, Van Y-Y, Dunker AK, Uversky VN. Protein intrinsic disorder and induced pluripotent stem cells. *Mol. Biosyst.* 2011;
161. Sun X, Xue B, Jones WT, et al. 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.* 2011;77(3):205–223.
162. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* 2002;12(1):54–60.
163. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 2004;337(3):635–645.
164. Huang F, Oldfield C, Meng J, et al. Subclassifying disordered proteins by the CH-CDF plot method. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 2012;128–139.
165. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 2005;6(3):197–208.
166. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. Intrinsic disorder and protein function. *Biochemistry (Mosc.)*. 2002;41(21):6573–6582.
167. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. The DISOPRED server for the prediction of protein disorder. *Bioinformatics*. 2004;20(13):2138–2139.
168. He B, Wang K, Liu Y, et al. Predicting intrinsic disorder in proteins: an overview. *Cell Res.* 2009;19(8):929–949.

169. Deng X, Eickholt J, Cheng J. A comprehensive overview of computational protein disorder prediction methods. *Mol. Biosyst.* 2011;8(1):114–121.
170. Peng Z-L, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr. Protein Pept. Sci.* 2012;13(1):6–18.
171. Pace CN, Shirley BA, McNutt M, Gajiwala K. Forces contributing to the conformational stability of proteins. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* 1996;10(1):75–83.
172. Rose GD, Fleming PJ, Banavar JR, Maritan A. A backbone-based theory of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* 2006;103(45):16623–16633.
173. Krimm S. The hydrophobic effect: Formation of micelles and biological membranes, Charles Tanford, Wiley-Interscience, New York, 1980, 233 pp. price: $18.50. *J. Polym. Sci. Polym. Lett. Ed.* 1980;18(10):687–687.
174. Nozaki Y, Tanford C. The Solubility of Amino Acids and Two Glycine Peptides in Aqueous Ethanol and Dioxane Solutions ESTABLISHMENT OF A HYDROPHOBICITY SCALE. *J. Biol. Chem.* 1971;246(7):2211–2217.
175. Artimo P, Jonnalagedda M, Arnold K, et al. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* 2012;40(W1):W597–W603.
176. Wilkins MR, Gasteiger E, Bairoch A, et al. Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol. Clifton NJ.* 1999;112:531–552.
177. Cortes C, Vapnik V. Support-vector networks. *Mach. Learn.* 1995;20(3):273–297.
178. Holladay NB, Kinch LN, Grishin NV. Optimization of linear disorder predictors yields tight association between crystallographic disorder and hydrophobicity. *Protein Sci. Publ. Protein Soc.* 2007;16(10):2140–2152.
179. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A Library for Large Linear Classification. *J Mach Learn Res*. 2008;9:1871–1874.
180. Whitley D. A genetic algorithm tutorial. *Stat. Comput.* 1994;4(2):65–85.
181. Radivojac P, Obradovic Z, Smith DK, et al. Protein flexibility and intrinsic disorder. *Protein Sci. Publ. Protein Soc.* 2004;13(1):71–80.
182. Bae K, Mallick BK, Elsik CG. Prediction of protein interdomain linker regions by a hidden Markov model. *Bioinformatics*. 2005;21(10):2264–2270.
183. Zhou H, Zhou Y. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins*. 2004;54(2):315–322.
184. Romero P, Obradovic Z, Dunker AK. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett.* 1999;462(3):363–367.
185. Anurag M, Singh GP, Dash D. Location of disorder in coiled coil proteins is influenced by its biological role and subcellular localization: a GO-based study on human proteome. *Mol. Biosyst.* 2012;8(1):346–352.
186. Yu W-F, Tasayco ML, Tung C-S, Wang H. NMR analysis of cleaved Escherichia coli thioredoxin (1–73/74–108) and its P76A variant: Cis/trans peptide isomerization. *Protein Sci.* 2000;9(1):20–28.
187. Yang XM, Georgescu RE, Li JH, et al. Recognition between disordered polypeptide chains from cleavage of an alpha/beta domain: self-versus non-self-association. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 1999;590–600.
188. Dolgikh DA, Gilmanshin RI, Brazhnikov EV, et al. Alpha-Lactalbumin: compact state with fluctuating tertiary structure? *FEBS Lett.* 1981;136(2):311–315.

189. Baldwin RL, Rose GD. Molten globules, entropy-driven conformational change and protein folding. *Curr. Opin. Struct. Biol.* 2013;23(1):4–10.
190. Simon SM, Sousa FJR, Mohana-Borges R, Walker GC. Regulation of Escherichia coli SOS mutagenesis by dimeric intrinsically disordered umuD gene products. *Proc. Natl. Acad. Sci. U. S. A.* 2008;105(4):1152–1157.
191. Gazit E, Sauer RT. Stability and DNA binding of the phd protein of the phage P1 plasmid addiction system. *J. Biol. Chem.* 1999;274(5):2652–2657.
192. Oldfield CJ, Cheng Y, Cortese MS, et al. Coupled Folding and Binding with α-Helix-Forming Molecular Recognition Elements†. *Biochemistry (Mosc.).* 2005;44(37):12454–12470.
193. Pentelute BL, Gates ZP, Tereshko V, et al. X-ray structure of snow flea antifreeze protein determined by racemic crystallization of synthetic protein enantiomers. *J. Am. Chem. Soc.* 2008;130(30):9695–9701.
194. Del Rizzo PA, Bi Y, Dunn SD, Shilton BH. The "second stalk" of Escherichia coli ATP synthase: structure of the isolated dimerization domain. *Biochemistry (Mosc.).* 2002;41(21):6875–6884.
195. Chak KF, Safo MK, Ku WY, Hsieh SY, Yuan HS. The crystal structure of the immunity protein of colicin E7 suggests a possible colicin-interacting surface. *Proc. Natl. Acad. Sci. U. S. A.* 1996;93(13):6437–6442.
196. Amos FF, Evans JS. AP7, a Partially Disordered Pseudo C-RING Protein, Is Capable of Forming Stabilized Aragonite in Vitro†. *Biochemistry (Mosc.).* 2009;48(6):1332–1339.
197. Murase K, Hirano Y, Sun T, Hakoshima T. Gibberellin-induced DELLA recognition by the gibberellin receptor GID1. *Nature*. 2008;456(7221):459–463.
198. Zhu J, Winans SC. The quorum-sensing transcriptional regulator TraR requires its cognate signaling ligand for protein folding, protease resistance, and dimerization. *Proc. Natl. Acad. Sci. U. S. A.* 2001;98(4):1507–1512.
199. Fasshauer D, Otto H, Eliason WK, Jahn R, Brünger AT. Structural changes are associated with soluble N-ethylmaleimide-sensitive fusion protein attachment protein receptor complex formation. *J. Biol. Chem.* 1997;272(44):28036–28041.
200. Ikeguchi M, Kato S, Shimizu A, Sugai S. Molten globule state of equine beta-lactoglobulin. *Proteins*. 1997;27(4):567–575.
201. Vamvaca K, Vögeli B, Kast P, Pervushin K, Hilvert D. An enzymatic molten globule: efficient coupling of folding and catalysis. *Proc. Natl. Acad. Sci. U. S. A.* 2004;101(35):12860–12864.
202. Kihara D. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci. Publ. Protein Soc.* 2005;14(8):1955–1963.
203. Ikeda K, Higo J. Free-energy landscape of a chameleon sequence in explicit water and its inherent α/β bifacial property. *Protein Sci.* 2003;12(11):2542–2548.

CURRICULUM VITAE

Fei Huang

Education

Indiana University Aug 2008-Jan 2014
PhD in Bioinformatics/Biochemistry
GPA: 3.80

University of Maryland-College Park Sep 2006-Jun 2007
Sichuan University Sep 2004-Jul 2008
Bachelor of Science in Biology

Research

Jul 2009- Jan 2014

PhD Thesis

Bioinformatics: Mentor: Dr. Keith Dunker

Past Experience

Biostatistics/Clinical-Pharmacology	Mentor: Dr. Lang Li
Genetics	Mentor: Dr. Nuria Morral
Cell and Physiology	Mentor: Dr. Jeffrey Elmendorf
Plant Genetics	Mentor: Dr. Zhongchi Liu
Immunology	Mentor: D. Jingqiu Chen

Publications

1. Fei Huang, Christopher Oldfield, Jingwei Meng, Wei-lun Hsu, Bin Xue, Vladimir N. Uversky, Pedro Romero, and A. Keith Dunker; Subclassifying Disordered Proteins by the CH-CDF Plot Method, Pacific Symposium on Biocomputing 17:128-139(2012)
2. Wei-Lun Hsu1, Christopher J. Oldfield1, Bin Xue, Jingwei Meng, Fei Huang, Pedro Romero1, Vladimir N. Uversky, A. Keith Dunker: Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding, Protein Science. 22(3):258-73(2013)
3. Wei-Lun Hsu, Christopher Oldfield, Jingwei Meng, Fei Huang, Bin Xue, Vladimir N. Uversky, Pedro Romero, and A. Keith Dunker; Intrinsic Protein Disorder and Protein-Protein Interactions, Pacific Symposium on Biocomputing 17:116-127(2012)

Conferences

Speaker at the Pacific Symposium on Biocomputing (PSB) 2012, Big Island, Hawaii
Poster presentation, Biophysical society 2011,Baltimore, Maryland
Poster presentation, Pacific Symposium on Biocomputing 2011, Big Island, Hawaii
Poster presentation, Gordon Conference 2010, Davidson, North Carolina

Programming skills

Matlab, Perl, Python, SQL, C/C++, Java, and R.