Flavors of Protein Disorder

Slobodan Vucetic[1], Celeste J. Brown[2], A. Keith Dunker[2] and Zoran Obradovic[1,*]

1) Center for Information Science and Technology, Temple University, U.S.A.

2) School of Molecular Biosciences, Washington State University, U.S.A.

*Correspondence to: Zoran Obradovic, Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, U.S.A. Fax: + (215) 204-5082. E-mail: zoran@ist.temple.edu

Running Title: Protein Disorder Flavors

Keywords: secondary structure, structure prediction, genomics, sequence composition, signaling and regulatory proteins, unfolded proteins

ABSTRACT

Intrinsically disordered proteins are characterized by long regions lacking 3-D structure in their native states, yet they have been so far associated with 28 distinguishable functions. Previous studies showed that protein predictors trained on disorder from one type of protein often achieve poor accuracy on disorder of proteins of a different type, thus indicating significant differences in sequence properties among disordered proteins. Important biological problems are identifying different types, or flavors, of disorder and examining their relationships with protein function. Innovative use of computational methods is needed in addressing these problems due to relative scarcity of experimental data and background knowledge related to protein disorder. We developed an algorithm that partitions protein disorder into flavors based on competition among increasing numbers of predictors, with prediction accuracy determining both the number of distinct predictors and the partitioning of the individual proteins. Using 145 variously characterized proteins with long (≥ 30 amino acids) disordered regions, 3 flavors, called V, C, S, were identified by this approach, with the V subset containing 52 segments and 7,743 residues, with C containing 39 segments and 3,402 residues, and with S containing 54 segments and 5,752 residues. The V, C, and S flavors were distinguishable by amino acid compositions, sequence locations, and biological function. For the sequences in SwissProt and 28 genomes, their protein functions exhibit correlations with the commonness and usage of different disordered flavors, suggesting different flavor-function sets across these protein groups. Overall, the results herein support the flavor-function approach as a useful complement to structural genomics as a means for automatically assigning possible functions to sequences.

INTRODUCTION

Although proteins that fail to self-fold into specific 3-D structure are receiving increased attention as evidenced by several recent major reviews[1-6], the concept is not new. That a native protein's function can depend on a structural ensemble rather than a unique 3-D structure was suggested more than 50 years ago[7], and that some proteins don't fold due to an atypical amino acid composition was suggested more than 20 years ago[8, 9]. More recently, such proteins have been called "natively unfolded"[10], "intrinsically unstructured"[1], and "intrinsically disordered"[2]. The failure to self-fold into specific 3-D structure is likely encoded by the amino acid sequence[11] and, furthermore, regions lacking specific 3-D structure have so far been associated with 28 distinguishable functions, ranging from DNA-binding to display of sites for phosphorylation to preventing interactions by means of excluded volume effects (reviewed in[12]).

A disordered protein or a disordered region lacks specific tertiary structure and is comprised of an ensemble made up of members with distinct and usually dynamic Ramachandran angles. In contrast, an ordered protein, even an ordered protein lacking helix or sheet, is comprised of an ensemble with nearly all of the members having the same canonical set of Ramachandran angles. Thus, regions of disorder should not be confused with loops or other regions that lack regular secondary structure. Following previous researchers[13], we suggested that disordered proteins could be either "extended" or "collapsed"[14], while others have pointed out data indicating the possibility of a third disorder class intermediate between extended and collapsed[5]. Collapsed disorder resembles the molten globule[15], a protein form with regular secondary structure but absence of fixed tertiary interactions, and regions of extended disorder can also exhibit secondary structure. Thus, proteins or regions with regular secondary structure are not always ordered.

On the ordered protein side, the structural classes – α-helix, β-sheet, other – can be visualized in protein crystal structures. Statistical analyses of the amino acid compositions of regions with different secondary structure over many proteins were carried out to determine amino acid propensities for α-helix, β-sheet, and other[16, 17]. Four protein folding classes – all α, all β , α/β, and α + β – can be predicted with fairly good accuracy solely from amino acid composition[18, 19]. These predictions of folding classes give reasonably good results because of the differences in amino acid propensities for helix, sheet, and other.

Just as ordered protein is comprised of different types of secondary structure, intrinsic disorder is also made up of distinguishable types as discussed above. Disordered segments, however, do not adopt unique structures and so cannot be visualized in the usual way. An alternative to visualization would be to determine intrinsic flexibility or whether a given disordered region is collapsed, extended, or intermediate, and if collapsed, the type and organization of the semi-stable secondary structure segments. At the present time, however, such data are limited and not organized[5, 20]. Thus, an alternative approach is needed.

Studies using different Predictors of Natural Disordered Regions (PONDRs;[11]) support the suggestion that disordered regions exhibit different types or flavors of disorder. That is, certain PONDRs gave high accuracy for some disordered proteins but low accuracy for others, while certain other PONDRs gave the reverse[21]. To give one example, a protein disorder predictor trained on a set of disordered regions from various calcineurins (PONDR CaN-1)[22] completely missed a long disordered region in the prion protein, but this region was correctly identified by other predictors (PONDR XL-1, VL-1) trained on different sets of disordered proteins[23].

In an initial attempt to identify different flavors of disorder, we developed a novel partitioning algorithm based on differential prediction accuracies. This algorithm uses the notion that a specialized predictor built on a given disorder flavor should have significantly higher same-flavor accuracy than other-flavor predictors or than a global predictor applied to the same given flavor. Moreover, the

algorithm was designed to cope with the following issues: (1) order/disorder information from the available data is noisy, and the representation currently used for prediction is imperfect and likely fails to include all of the important, but so far unexplored attributes; (2) neighboring positions in a disordered sequence share very similar local context and so are likely to be of the same flavor.

We present results of an extensive statistical and qualitative evaluation of the discovered partitioning by comparing the corresponding specialized predictors, as well as amino acid properties and protein functions of different partitions. We also show the results of applying the specialized predictors on the SwissProt database and 28 complete genomes in order to estimate the frequency of the various disorder flavors in nature. Finally, the flavor-specific disordered proteins and the flavor-specific disorder predictions are used to identify possible flavor-function relationships.

METHODS

Data Sets

Our data set of proteins with disordered regions longer then 30 consecutive residues was derived from a previously described data set[24]. Briefly, reports on disordered regions identified by NMR, circular dichroism or protease digestion were located by keyword searches of PubMed (http://www.ncbi.nlm.nih.gov). Additionally, starting from PDB_Select_25[25], a nonredundant subset of the Protein Data Bank (PDB)[26] whose members have less than 25% sequence identity, we identified disordered regions in x-ray crystal structures by searching for residues whose backbone atoms are absent from the ATOMS list of their PDB files[22]. The resulting disordered data set, D_145, contains 145 protein segments longer then 40 consecutive residues with a total of 16,705 disordered residues. Of these proteins, 35% are completely disordered, 16% display disorder at their C-terminal ends, 30% at their N-terminal ends, and 17% at regions internal to the chains.

To provide a representative data set of ordered proteins necessary for training predictors of protein disorder we extracted a set of 130 non-redundant proteins that are completely ordered from their N- to C-termini (O_130). This data set with a total of 32,506 residues was also extracted from PDB_Select_25. To measure the false positive prediction rate of the various predictors, an additional database O_PDB_S25 of ordered protein segments was constructed from PDB_Select_25 by deleting all the residues lacking backbone coordinates.

Several other databases were used for the analysis of the resulting partitions of disordered protein regions and the corresponding specialized predictors. These include the amino acid sequences of known and putative proteins obtained for complete or mostly complete genomes of 28 organisms (http://www.ncbi.nlm.nih.gov/Entrez/Genome/main_genome.html) and the 80,000 amino acid sequences obtained from release 38.00 of the SwissProt database[27].

Data Representation and Attribute Selection

In this study, order/disorder properties at a given position in a sequence are predicted using a subsequence within a window of size W_{in}. Since the amount of information about disordered proteins is rather limited, in accordance with previous work[24], only first-order statistics of the 20 amino acids within a given window were used as attributes to prevent the "curse of dimensionality"[28]. For example, attribute X_A at a given position is calculated as the fraction of amino acid A (ALA or Alanine) within a window. It is worth noting that such representation includes residues at the N- and C- termini of the protein. This simplistic approach is further validated by results known about the incompressibility of protein sequences[29], showing that it can be very difficult to extract relevant higher-order statistics from protein sequences.

A measure of sequence complexity called Shannon's entropy[30] provided another attribute. When measured over a window specified length, Shannon's entropy is given by $K_2 = -\sum_{i=1}^{N} f_i \log_2 f_i$, where $N = 20$, corresponding to the 20 amino acids, and where f_i is the fraction of the i^{th} amino acid in the

window (and where 0 $\log_2 0$ is defined to be zero). Very low complexity sequences are almost always disordered, whereas high complexity sequences can be ordered or disordered[24, 31].

Prediction of Protein Disorder

A measure of prediction accuracy used here is the average between the percent of true positives (disordered positions predicted to be disordered) and true negatives (ordered positions predicted to be ordered). With this convenient measure, ordered and disordered predictions are valued equally regardless of the balance of the data and a higher value indicates a better predictor. If a balanced dataset with the same number of ordered and disordered examples is used for training and testing, then each result above 50% indicates that the predictor is better then a random guess.

Among a range of alternative machine learning algorithms we considered ordinary least squares (OLS), logistic regression (LR) and neural networks (NN) to build disorder predictors. OLS and LR[32] are linear predictors, where OLS is suited for regression and is extremely fast to run, while LR is suited for binary classification and requires a slower (one order of magnitude) iterative parameter fitting procedure. NN are known as universal approximators, capable of discovering highly nonlinear relationships[33], but they require abundant data, have relatively slow training, and are sensitive to the initial conditions and training procedures.

Regardless of the choice of disorder prediction algorithm, the final prediction accuracy could be further improved by exploiting a specific feature of the prediction task, namely, prediction of long disordered regions (>40 residues). If a predictor indicates that most residues in a certain region of a protein are likely to be disordered, then the whole region is likely to be disordered. The opposite holds as well. Therefore, this simple reasoning can facilitate correcting the prediction of apparently misclassified residues. In this study, we used a simple procedure for such correction by averaging predictions over a window of size W_{out}, and we validate this approach in the implementation section.

Algorithm for Discovering Flavors of Protein Disorder

The proposed algorithm uses an assumption that specialized disorder predictors built on different flavors of protein disorder should have significantly higher accuracy on their own flavor then a global or other specialized disorder predictors. Starting from this assumption, we constructed an algorithm that partitions a set of disordered proteins into flavors through a competition of specialized predictors for disordered proteins. If a given set of disordered proteins S_0 is partitioned into L disjoint subsets, the corresponding data subsets are denoted as S_i, i=1, …L, and the whole partitioning as $\mathbf{S}_L = \{S_1, \ldots S_L\}$. By M_i we denote a specialized predictor built using examples of disorder from S_i and the same number of examples taken randomly from the ordered dataset O_130 to provide a balanced training data set.

The algorithm (formally described in Table I) starts by fitting a global predictor on a balanced data set containing examples from all disordered proteins. Next, disordered proteins are divided randomly into two disjoint subsets $S_{1,2}^0$ and $S_{2,2}^0$ of equal size and two separate predictors are trained using each of them. Data sets for both predictors are also balanced with the same number of residues from O_130. The two predictors are then applied on each disordered protein. All disordered proteins where the first predictor achieves higher disorder prediction score are assigned to subset $S_{1,2}^1$ while the rest is assigned to subset $S_{2,2}^1$. The competition procedure is iterated until a stable partitioning $\mathbf{S}_2$ is obtained. The competition procedure is also described in Table I. Subsequently, the quality of the resulting partitioning $\mathbf{S}_2$ is measured as *the overall accuracy* explained below. If there is no accuracy improvement as compared to the global predictor, the algorithm is stopped and it is concluded that all disorder is similar. Otherwise, the algorithm proceeds by partitioning disordered proteins into three subsets in an attempt to further improve the accuracy.

(Table I)

The third predictor is added to the competition in a tree-like manner as described in Table I. Starting from a partitioning into $S_{1,2}$ and $S_{2,2}$, subset $S_{1,2}$ is split into two equal size disjoint subsets, while

$S_{2,2}$ is left intact. The competition of three predictors starting from such a partitioning is performed and after converging to a stable partitioning $\mathbf{S}_{3,1}$, the accuracy is calculated. Also, the same steps are repeated by splitting $S_{2,2}$ into two subsets and leaving $S_{1,2}$ intact to obtain $\mathbf{S}_{3,2}$. The quality of the better of the two partitions (denoted as $\mathbf{S}_3$) is compared to $\mathbf{S}_2$. The procedure continues by adding new predictors into the competition until it is concluded that further partitioning does not improve accuracy. The partition that gives the best overall accuracy is accepted as the final partition of disordered proteins into *the disorder flavors*. The resulting specialized predictors are named *the flavor predictors*.

The quality of each partition **S** is measured as *the overall accuracy* using leave-one-protein-out validation on each subset. Namely, after each competition step, the following procedure is repeated for each disordered protein. If a given protein is assigned to subset S_i, a predictor is learned on a balanced set using the remaining disordered proteins from S_i and the accuracy of that predictor is measured on the selected disordered protein. The overall accuracy is calculated as the average accuracy over all disordered proteins. Therefore, for n disordered proteins in the whole data set, calculation of the overall accuracy required learning n predictors.

Although convergence properties of the proposed algorithm were not analyzed theoretically for partitioning of disordered proteins, our results from similar partitioning algorithms specialized for nonstationary time series[34] and spatial data[35] showed that the competition procedure converges to local optimum and that the convergence is expected to be fast.

Experimental Design

Choice of attributes and data set size. Since the sum of 20 attributes representing amino acid frequencies within a window is one, to prevent co-linearity one attribute should be excluded from the data set. We removed attribute X_M representing the frequency of methionine within a window since our previous studies indicated that it is not important for disorder prediction. The 19 remaining compositional

attributes and sequence complexity were retained because, although some of them might not be important for a global disorder predictor, they could influence some of the specialized disorder predictors.

To improve the speed of partitioning algorithm and to use the fact that neighboring examples tend to be correlated, only 20 examples were randomly included in a training set from each of the D_145 proteins regardless of the protein's length. Thus, the data set used in partitioning had a total of 2,900 examples of protein disorder. Furthermore, such data reduction allowed each disordered protein to have the same importance on the partition procedure.

Choice of specialized predictors. Ordinary least squares (OLS), logistic regression (LR) and neural network predictors were all compared for overall accuracy in predicting order and disorder. Accuracies of the predictors shown in Table II were obtained with leave-one-protein-out validation on the D_145 and O_130 data set using the window of size $W_{in} = 41$. The time presented is that needed to learn one predictor on a complete training data set with 2,900 ordered and 2,900 disordered 20-dimensional examples as implemented in Matlab on a 700MHz NT-based computer with 256MB memory. Neural networks had 5 hidden neurons and were trained for 100 epochs with a resilient backpropagation algorithm[36]. The requirement for the partitioning algorithm was to provide sufficiently accurate partitioning in a reasonable time. Since OLS proved to be the fastest prediction method with overall prediction accuracy close to the other two methods it was chosen for specialized predictors in the partitioning algorithm.

(Table II)

Choice of window size. The window size used for attribute construction in this study was $W_{in} = 41$, because this size has better overall accuracy than windows of size 21 used previously[23] or to windows of size 81 (Table III).

Postprocessing. Averaging the outputs over a window of size $W_{out} = 41$ proved very useful. The overall accuracy of the linear predictor was improved by almost 3% to 80.5% (last row of Table III). Since such postprocessing does not influence the competition procedure of the partitioning algorithm, it was not used in the disorder flavor discovery algorithm. However, it was used later to allow more successful analysis of disorder flavors on a number of out-of-sample protein databases.

(Table III)

Number of experiments. Preliminary analysis revealed that there is a fairly small difference in accuracy between competing predictors on D_145 proteins. As a consequence, the final partitioning is dependent on random initial splits in the competition procedure, and the overall accuracy can vary slightly. The sensitivity to initialization is a common property of clustering algorithms with iterative refinements, and could be explained by the presence of local optima in the space of possible solutions[37]. The standard approach to addressing this problem is repeating the experiment a number of times with different random initializations and selecting the best partitioning according to the defined quality measure. In this study we were interested in partitioning of disordered proteins into the high-quality flavors, without the need for a brute-force exhaustive search over all possible partitions of D_145 proteins. Therefore, we repeated our partitioning algorithm with 20 different initializations, and performed an in-depth analysis of the best resulting partitioning.

RESULTS

Results of the Partitioning

In each of the 20 experiments we apply the proposed algorithm to partition D_145 proteins into 6 subsets. In Table IV we show the minimum and the maximum overall accuracies achieved over 20 experiments

for the number of subsets ranging from two to six. We also report the bootstrap estimate of the 90% confidence interval for the overall accuracy of the best run of the algorithm. It was calculated based on the 1000 bootstrap replicates of 145 accuracies measured on D_145 proteins. Relatively wide confidence intervals are a consequence of a relatively small number of disordered proteins in our data set.

(Table IV)

Partitioning into two or three subsets by the best of 20 experimental runs improved the overall accuracy by 4.4% and 12.9%, respectively, as compared to the accuracy of a global model with 71.5% accuracy. Further partitioning into 4 and 5 subsets resulted in just a slight accuracy improvement by 0.5% and 0.8%, respectively. The accuracy resulting from 6-subset partition decreased slightly as compared to that from a 5-subset partition. Considering the size of the 90% confidence intervals for the overall accuracy, and based on the principle of Occam's Razor, commonly interpreted in machine learning as "the simplest explanation of the observed phenomena is most likely to be a correct one", we concluded that partitioning into 3 subsets by the best run of our algorithm provides the most reasonable solution. We also concluded that at least 3 flavors exist among the 145 disordered proteins in our data set, where the more definite answer should probably await for the significant enlargement of the disordered protein data set.

We denoted the 3 subsets of partitioned proteins as *disordered flavors* V, C, and S (the data is available at DisProrт.wsu.edu/flavors), and the corresponding predictors as *the flavor predictors* VL-2V, VL-2C and VL-2S. Single letter designations were used for convenience, and the choice of V, C, and S was mostly arbitrary. The global predictor was denoted as VL-2 to distinguish it from the first generation predictor VL-1, which was trained on a small number of disordered regions.

An important consideration of our approach is the sensitivity of our algorithm to the random initialization. From Table III it can be seen that the difference between the worst and the best among the 20 runs was less than 3%, indicating a moderate influence of initialization on the overall accuracy of

partitioning. To estimate the stability of the partitioning over N runs of the algorithm we define *a stable protein* as the protein that has the same assignment over all of the N runs with at least another protein. If partition of D_145 proteins into 3 subsets by the algorithm were completely random, an expected number of stable proteins over N runs would be $145\left(1-\left(1-\left(\frac{1}{3}\right)^{N}\right)^{144}\right)$. For N=5, 10, and 20 this number equals 64.9, 0.35, and $6\cdot10^{-6}$, respectively. Partitioning of D_145 proteins into 3 subsets in the first 5, 10 and in all of the 20 runs of our algorithm resulted in 119, 95, and 61 stable proteins, respectively. A sizeable core of stable proteins indicates that the algorithm is fairly robust to random initializations. It is also evident that there exist proteins on which all three specialized predictors achieve comparable accuracy and whose membership into one of the 3 subsets is unstable.

Initial Characterization of the Flavors and Flavor Predictors

To compare the obtained 3-flavor partitioning to chance, we generated 1000 random partitions of D_145 disordered proteins into three subsets. The overall accuracy of the predictors derived from the random partitions was only 68.5 ± 2.1%, which was more than one standard deviation less than the 71.5% (Table IV) observed for the global predictor. Furthermore, since 84.4% (Table IV) is more than 7 standard deviations greater than 68.5%, the obtained 3-subset partitioning is very unlikely to be due to chance. Note further that the random variation of 2.1% is much greater than the 0.5% improvement observed for 4 versus 3 subsets and much less than the 8.5% improvement of 3 versus 2 subsets, which provides further support for the choice of the 3-subset partitioning.

Table V compares the accuracy of the global predictor and the three flavor predictors on the ordered and disordered data sets. The accuracy of each flavor predictor on the ordered data set ranged from 81% to 86%, while accuracy on their corresponding disorder flavors ranged from 83% to 87%. Comparison with global predictor VL-2 shows that partitioning resulted in a significant increase of accuracy on disordered residues while keeping accuracy on ordered residues almost unchanged. The

global predictor does not approach the accuracy of each flavor predictor on its own data set, and each flavor predictor has higher accuracy on its own flavor and substantially lower accuracy on the other flavors.

(Table V)

Compositional Analysis of Disorder Flavors

Our next goal was to determine a set of important attributes for the global and flavor predictors. We used the t-statistics of linear regression parameters obtained using OLS as a measure of attribute importance. In Table VI, we show the most influential attributes (sorted by decreasing importance) for the global and 3 flavor predictors. Only attributes with t-statistics whose absolute values are greater than 7 are shown where bold attributes have t-statistics greater then 14. As evident from the results, there are significant differences between the flavor predictors, with VL-2S exhibiting the most resemblance to the global predictor. In general, the amino acids that are most influential for each predictor are the more hydrophobic ones. K2 entropy is the most important attribute for the VL-2 and VL-2S predictors, moderately important for VL-2C and not very influential for VL-2V.

(Table VI)

In general, the attributes that are most influential to each predictor reflect the characteristics of the various sequences used to train the predictors. The average K_2 entropy for the O_130 proteins was 3.72 and for D_145 was 3.39. The flavor S proteins had the closest average K2, 3.38, to D_145 and the flavor V had the closest average K2, 3.51, to O_130. These values closely reflect the influence of K2 in predicting disorder. The average K2 for flavor C was lowest at 3.24.

All of the predictors are most heavily influenced by the presence of the most hydrophobic residues. The amino acid compositions for each disorder flavor relative to the compositions for order are

shown in Figure 1. Amino acids in the figure are sorted in ascending order by the "flexibility" scale of Vihinen and co-workers[38]. This scale relates more to the tendency of an amino acid to be buried (to the left) or to be exposed (to the right) than it does to inherent flexibilities of the various amino acids. Each bar represents the ratio $(P_j^D - P_j^O) / P_j^O$, where P_j^D is the proportion of amino acid j in the subset of disorder and P_j^O is the same proportion in order. Values higher than 0, for example, indicate that the amino acid is found more often in disordered proteins than in ordered proteins. Similarly, values below 0 indicate the amino acid is less common than in ordered proteins.

Figure 1. Comparison of amino acid compositions of three disorder flavors with the composition of order. Error bars are one standard deviation.

(Figure 1)

Figure 1 clearly elucidates the differences between the disordered protein and the ordered protein, as well as differences among the three flavors. Disordered proteins have fewer of the amino acids that tend to be buried, and more of the most flexible amino acids. The proteins that determine each of the flavors have slight differences in their compositions relative to each other. Thus, Flavor C has more histidine (H), methionine (M), and alanine (A) than is usually found in either ordered protein or the other two flavors. Flavor S has less histidine than the other flavors and ordered protein, and Flavor V has more of the least flexible amino acids (C, F, I, Y) than the other disorder flavors.

The difference among the distributions of the three disorder flavors and order can be compared statistically as well as graphically. Results on the incompressibility of protein sequences[29] indicate that it is very difficult to extract Markovian dependence of higher order in protein sequences. As a consequence, protein sequences can, at least as a first approximation, be considered as random samples taken from some distribution. One commonly used statistic to test the hypothesis that two samples S_1 and S_2 are from

the same distribution is the D statistic defined as $D = \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{20} \frac{(P_j^{(1)} - P_j^{(2)})^2}{P_j^{(12)}}$, where n_1 and n_2 are the number of examples from both samples, $P_j^{(1)}$, $P_j^{(2)}$ and $P_j^{(12)}$ are frequencies of amino acid j in both samples and in the joint sample S_1+S_2. In Table VII we show values of D-statistics between the three flavors of the disordered residues and the ordered residues. To validate that the D-values are significant, we constructed 1000 random partitions of the 145 disordered proteins into 3 subsets and calculated the D-statistic for each partition. Our results showed that 95% of the D-statistics for the random partitions were smaller then 189. From this experiment, we can estimate p-values for the compositional difference between the pairs of disorder flavors; these are given in parentheses in Table VII.

(Table VII)

Note that the D-values indicate that all three of the disorder flavors are very different from ordered proteins. Also, Flavor C seems to be more different than the other two flavors, while flavors V and S are more similar to each other except with regards to sequence complexity, K_2.

Flavors Versus Characterization Method and Location

There are interesting differences among the three flavors in the distribution of disordered regions by characterization method and by location (Table VIII). NMR characterized disorder is fairly evenly distributed among the 3 flavors; there is more CD and proteolysis characterized disorder in Flavor V and more X-ray disorder in Flavor S than if the characterization methods were randomly divided among the 3 flavors. These results are correlated with the location of the disordered regions. Flavor V has more completely disordered proteins than expected due to the greater fraction of CD characterized proteins, which tend to be characterized as complete proteins. The greater number of X-ray derived disordered regions in Flavor S is reflected by the smaller number of completely disordered proteins in this flavor. Interestingly, Flavor C has far fewer internal regions of disorder than the others. The differences in the

distributions of characterization method and disorder location are significantly different among the flavors (df=4, α=0.05, χ2=11.07; df=6, α=0.005, χ2=18.5, respectively).

(Table VIII)

Extended Versus Collapsed Disorder

The protein trinity hypothesis proposes that *native* proteins exist in 3 forms: ordered, extended and collapsed[14]. Limited comparisons suggested that the extended disorder exhibited a lower sequence complexity, a higher net charge, and a reduced overall hydrophobicity compared to collapsed disorder. To determine whether extended versus collapsed disorder might be separated in the V, C, and S partitions, comparisons of these three properties for the ordered protein set and the three disordered subsets were determined (Table IX). Flavors V and C appear to fit the suggested description of extended disorder. However, complexity, net charge, and hydropathy did not exhibit statistically significant separation among V, C, and S, suggesting that the flavors could not clearly distinguish between extended and collapsed disorder.

(Table IX)

Disorder Flavor-Function Relationships

We have carried out a preliminary evaluation of flavor-function relationships for the V, C and S flavors (Table X). Not all regions of disorder have known functions. For instance, the functions of the N-terminal disordered regions of DNA lyase[39] and phosphatidyl-inositol phosphate kinase[40] are not known. This does not mean that these regions have no function, simply that they have not been sufficiently studied. For the disordered regions for which we were able to determine function, some interesting patterns were evident (Table X). Over half of the disordered regions that bind to other proteins partition into flavor S.

Of the ten ribosomal proteins from *E. coli* that were shown to be completely disordered by circular dichroism[41], nine partition into flavor V. These proteins function as structural mortar, interacting with each other and the ribosomal RNA to maintain the structure of the ribosome. Disordered proteins that bind to the genomic RNA of viruses, however, were not found in flavor V. Few of the DNA binding proteins partitioned into flavor V either.

(Table X)

Since the proteins in D_145 with sufficiently confident estimates of disorder function represent a limited sample from the distribution of disorder in nature we also used flavor-specific predictions to assess flavor-function relationships. We extracted predictions for various protein domains from the SwissProt database. The VL-2V predictor gave strong predictions of disorder for nuclear localization signals. VL-2V also gave the strongest disorder predictions of the three predictors for helical regions. We have shown previously that disordered domains that become ordered helices upon binding to other proteins are compositionally distinct from other types of disorder[42]. Both VL-2V and VL-2S strongly predict transactivating domains as disordered. This coincides with the large number of protein-binding regions that are disordered in flavor V and flavor S. The VL-2S predictor strongly predicted disorder in leucine-rich domains, but not leucine zippers. The VL-2C predictor gave strong predictions of disorder for poly- and oligo-saccharide binding domains.

Therefore, although the flavor partition was based exclusively on sequence statistics, we conclude that significant differences exist in disorder function between flavors V, C, and S. This holds both for known functional assignments of D_145 proteins, as well as for the functions of SwissProt proteins predicted to have long regions of disorder

Commonness of Disorder Flavors in Swissprot and Various Genomes

In order to study the prevalence of disorder in various databases and genomes, we developed a method for conservatively estimating the proportion of proteins with long disordered regions. The thresholds of the predictors were adjusted to values that resulted in 5% false positive per-residue error rates on O_PDB_S25 where it is known with reasonable confidence that all proteins are ordered. All of the following results include the postprocessing step, averaging predictions within a window W_{out}=41.

Our first goal was to estimate how common each flavor is in nature, as represented by 80,000 sequences from SwissProt (release 38.0). Using the conservative estimates, VL-2V and VL-2C predict that 22% of proteins in SwissProt have long regions of disorder, and VL-2S predicts 28% (Table XI). The results on the SwissProt database indicate that all three disorder flavors occur frequently in nature, with a similar relative abundance. For comparison, the first two rows of Table XI indicate the maximum false positive rate (O_PDB_S25) and the minimum true positive rate (D_145). With the false positive per-residue error rate set to approximately 5% per predictor, the relatively small true positive rate indicated by the accuracy on the D_145 data set clearly suggests that the abundance of disorder in SwissProt is underestimated.

(Table XI)

Figure 2 shows the overlap among the disorder predictions for the three flavor predictors. Total of 42.5% of proteins in SwissProt are predicted to have disordered regions of length 40 or longer by at least one of the predictors corresponding to 16.2% of the total residues. The majority (66%) of SwissProt residues were predicted to be disordered by only one of the three predictors, which is another indication of important differences between the discovered disorder flavors.

Figure 2. Venn Diagram showing overlap in predictions of flavor predictors on the SwissProt database. Numbers represent the distribution of mono-flavors (V, C, S) and mixed-flavors (VC, VS, CS, VCS) within all SwissProt residues predicted to be disordered by at least one flavor predictor.

(Figure 2)

SwissProt is known to be a biased database, unrepresentative of any individual genome. To obtain an alternative perspective, our analysis was extended to 28 complete genomes, where the disorder is scored as percentages of proteins with having predicted disordered regions of 40 or longer (Table XII). By this measure, the amount of predicted disorder is different for the 3 kingdoms. That is, the percentages of proteins predicted to have long disordered regions by at least one flavor predictor range from: 26-51% in archaea, to 16-45% in eubacteria, and to 52-67% in eukaryotes.

The most common disorder flavor also varies among the genomes. Note that the abundance of predicted disordered flavors varies considerably over different species and kingdoms. For example for 5 bacterial genomes flavor V is predicted on less than 5% of the proteins which is smaller then the 5% lower bound on ordered O_PDB_S25 proteins. This is an indication that flavor V is either extremely underrepresented or even missing in these genomes. At the other extreme, more then 40% of eukaryotic proteins were estimated to have long regions of flavor S.

(Table XII)

Table XII also reveals a large amount of variability in the relative abundance of the three disorder flavors across different species and kingdoms. The archaea, except *A. pernix*, are clearly biased towards flavor V, while all 4 eukaryotes are biased towards flavor S. While eubacteria are mostly biased towards flavor S (with 12 out of the 18 species), 4 eubacteria have affinity towards Flavor C and 2 toward flavor V.

DISCUSSION

Partitioning Sequences by Amino Acid Composition

Since the helix, sheet and other structural classes of ordered protein can be predicted from amino acid compositions, we were motivated to determine whether disorder types, or flavors, could likewise be identified by compositional differences. Unlike ordered protein, however, the membership of a given region of sequence in a specific disordered grouping is not known beforehand.

A straightforward approach for identifying different types of disorder would be to perform hierarchical, k-means, or some other type of clustering[43] that cluster the data based on their representation in the attribute space. However, clustering is effective only on data with a relatively small set of relevant attributes. In case of protein disorder, the attributes constructed on D_145 have different levels of relevance to protein disorder (as could be seen from Figure 1). Applying of-the-shelf clustering algorithms on such data could result in clusters with low biological significance.

In the current work, we developed a novel algorithm that partitions disordered proteins as a result of the competition between specialized disorder predictors for disordered proteins. This algorithm resembles the class of the expectation-maximization algorithms[44] by its iterative improvement of the partitions and their corresponding specialized predictors and by its overall goal of maximizing the accuracy of the specialized predictors. A novel aspect of our algorithm is the use of examples of protein order in partitioning of disordered proteins through construction of specialized disorder predictors. This resolves the problem of irrelevant attributes - they are not used by the specialized disorder predictors and, therefore, do not influence the partitioning.

Our experiments were performed on a relatively small data set of 145 disordered proteins. As a consequence, it was difficult to distinguish the difference between the quality of partition of disordered proteins into 3, 4, 5, or 6 subsets (see Table IV). It is expected that future enlargements of disorder database will lead to a better insight in the variability of protein disorder and allow more accurate

estimation of the number of disorder flavors. Additionally, with enlarged database, using nonlinear predictors in our algorithm is likely to result in higher quality flavors.

In the absence of prior knowledge, we made the simplification that each disordered region contained just one flavor. The results don't fully support this simplification. Many of the disordered regions appear to be better described as having multiple flavors. One example would be a disordered region that could be described as C-V-C, meaning that VL-2V has the most confident prediction in the middle of the region, while VL-2C has the most confident prediction at its ends. Such multi-flavor disordered regions probably contribute to the sensitivity of the partitioning to the initialization. By relaxing the simplification of one flavor per disordered region, improvements in flavor discrimination and prediction accuracy may result. Experiments are in progress to test this possibility.

Commonness of Intrinsic Disorder

Application of the three disorder predictors to a database of protein sequences and to the predicted protein sequences from the complete genomes of 28 species indicates that the prevalence of disorder is likely greater than previously estimated. Our current estimates (Table XII) of the proportion of proteins with disordered regions longer then 40 amino acids are 16-45% in bacteria, 26-51% in archaea, and 52-67% in eucaryotes. These values are larger then the corresponding previous estimates of 7-33%, 9-37%, and 35-51%, respectively[45].

As suggested previously[45,46,47], the increased predictions of disorder in eukaryota compared to the other kingdoms may be related to cell signaling and regulation. The data in Table XII indicate more flavor S proteins in eukaryota and the data in Table X indicate that flavor S is especially associated with protein-protein interactions, which are often involved in signaling and regulation. Thus, the present data support the previous suggestions that intrinsic disorder is important for cell signaling.

Implications for Structural Genomics

The large fraction of proteins that are predicted to have regions of disorder longer than 40 amino acids in length suggests that the field of structural genomics[48-50] must consider prediction of disorder, or lack of structure, a priority. The goal of structural genomics is to ascertain the structures of proteins from complete genomes and then to use these structures as intermediaries for determining function. This venture includes both solving structures experimentally, and predicting structures from sequence. Many experimental determinations of structure are hampered by the presence of disorder. The accuracy of structural predictions will be reduced if the presence of disordered protein is not taken into account.

Structural genomics must also consider disorder in determining protein function. Understanding the relationships between disorder flavor and protein function as described herein shows promise for annotating functions associated with disordered regions. Such an approach would be complementary to the annotation of ordered protein domains. Considerable improvement in the flavor-function approach, however, will be needed before this method becomes truly useful. Substantial enlargement of the set of intrinsically disordered proteins with well-characterized functions has the highest priority. More data would enable several avenues for improving the flavor-function approach. For example, preliminary investigations indicate that long regions of disorder can sometimes be divided into sub-regions that have different flavors, so dropping the mono-flavor simplification used here is likely to foster improvements. Following from the idea of multi-flavored regions, flavor-function relationships could well involve a pattern of flavors rather than individual flavor types, so for example a multi-flavored region such as V-C-S might correlate with a particular function. In addition, for proteins with both ordered and disordered regions, a determination whether certain disorder flavors are found in association with particular folding classes might provide novel insight regarding function for both the ordered and disordered regions. Finally, additional improvements can be expected from the reverse of what was done here, namely to form groups of disordered regions that are associated with common functions and then to determine whether function-specific disorder predictors can be developed from these groups.

We welcome help with the task of enlarging our database of intrinsically disordered proteins. A website, <http://www.DisProt.wsu.edu>, has been established for direct deposit of sequence-function information on proteins that are structurally characterized to have intrinsically disordered regions.

Acknowledgments

Support from NSF-CSE-IIS-9711532 and NSF-IIS-0196237 to Z.O. and A.K.D. and from N.I.H. 1R01 LM06916 to A.K.D. and Z.O is gratefully acknowledged.

References

1. Wright, P.E. and H.J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol., 1999. 293: 321-331.
2. Dunker, A.K., *et al.*, Intrinsically disordered protein. J. Mol. Graph. Model., 2001. 19: 26-59.
3. Namba, K., Roles of partly unfolded conformations in macromolecular self-assembly. Gen. Cells, 2001. 6: 1-12.
4. Demchenko, A.P., Recognition between flexible protein molecules: induced and assisted folding. J. Mol. Recognit., 2001. 14: 42-61.
5. Uversky, V.N., Natively unfolded proteins: a point where biology waits for physics. Protein Sci., 2002. 11: 739-756.
6. Dyson, H.J. and P.E. Wright, Coupling of folding and binding for unstructured proteins. Curr. Opin. Struct. Biol., 2002. 12: 54-60.
7. Karush, F., Heterogeneity of the binding sites of bovine serum albumin. J. Am. Chem. Soc., 1950. 72: 2705-2713.
8. Williams, R.J., Energy states of proteins, enzymes and membranes. Proc. R. Soc. Lond. B Biol. Sci., 1978. 200: 353-389.
9. Williams, R.J., The conformational mobility of proteins and its functional significance. Biochem. Soc. Trans., 1978. 6: 1123-1126.

10. Weinreb, P.H., W. Zhen, A.W. Poon, K.A. Conway, and P.T. Lansbury, Jr., NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. Biochemistry, 1996. 35: 13709-13715.

11. Dunker, A.K., C.J. Brown, and Z. Obradovic, Identification and functions of usefully disordered proteins. Adv. Protein Chem., 2002. 62: 25-49.

12. Dunker, A.K., C.J. Brown, J.D. Lawson, L.M. Iakoucheva, and Z. Obradovic, Intrinsic disorder and protein function. Biochemistry, 2002. 41: 6573 - 6582.

13. Ptitsyn, O.B. and V.N. Uversky, The molten globule is a third thermodynamical state of protein molecules. FEBS Lett.., 1994. 341: 15-18.

14. Dunker, A.K. and Z. Obradovic, The protein trinity - linking function and disorder. Nat. Biotechnol., 2001. 19: 805-806.

15. Dolgikh, D.A., R.I. Gilmanshin, E.V. Brazhnikov, V.E. Bychkova, G.V. Semisotnov, S. Venyaminov, and O.B. Ptitsyn, Alpha-Lactalbumin: compact state with fluctuating tertiary structure? FEBS Lett.., 1981. 136: 311-315.

16. Chou, P.Y. and G.D. Fasman, Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry, 1974. 13: 211-222.

17. Levitt, M., Conformational preferences of amino acids in globular proteins. Biochemistry, 1978. 17: 4277-85.

18. Nakashima, H., K. Nishikawa, and T. Ooi, The folding type of a protein is relevant to the amino acid composition. J. Biochem. (Tokyo), 1986. 99: 153-162.

19. Bahar, I., A.R. Atilgan, R.L. Jernigan, and B. Erman, Understanding the recognition of protein structural classes by amino acid composition. Proteins, 1997. 29: 172-185.

20. Choy, W.Y., F.A. Mulder, K.A. Crowhurst, D.R. Muhandiram, I.S. Millett, S. Doniach, J.D. Forman-Kay, and L.E. Kay, Distribution of molecular size within an unfolded state ensemble using small-angle X-ray scattering and pulse field gradient NMR techniques. J Mol Biol, 2002. 316: 101-12.

21. Wang, J., Family-specific Neural Network Predicters and Different Flavors of Disordered Regions, in School of Electrical Engineering and Computer Sciences. 2000, Washington State University: Pullman, WA.

22. Romero, P., Z. Obradovic, and A.K. Dunker, Sequence data analysis for long disordered regions prediction in the calcineurin family. Genome Inform. Ser. Workshop Genome Inform., 1997. 8: 110-124.

23. Romero, P., Z. Obradovic, and A.K. Dunker, Intelligent data analysis for protein disorder prediction. Artificial Intelligence Rev., 2000. 14: 447-484.

24. Romero, P., Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, and A.K. Dunker, Sequence complexity of disordered protein. Proteins, 2001. 42: 38-48.

25. Hobohm, U. and C. Sander, Enlarged representative set of protein structures. Protein Sci., 1994. 3: 522-524.

26. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, The protein data bank. Nucleic Acids Res., 2000. 28: 235-242.

27. Bairoch, A. and R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res.., 1999. 27: 49-54.

28. Bishop, C.M., Neural Networks for Pattern Recognition. 1995, Oxford, UK: Oxford University Press. xvii, 482 p. : ill. ; 24 cm.

29. Nevill-Manning, C.G. and I.H. Witten. Protein is incompressible. in Data Compression Conference. 1999. Snowbird, Utah.

30. Shannon, C.E., A mathematical theory of communication. Bell Syst. Tech. J., 1948: 379-423, 623-656.

31. Romero, P., Z. Obradovic, and A.K. Dunker, Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Lett.., 1999. 462: 363-367.

32. Davidson, R. and J. MacKinnon, Estimation and Inference in Econometrics. 1993, New York: Oxford University Press. xx, 874 p. : ill. ; 25 cm.

33. Cybenko, G., Approximation by superpositions of a sigmoidal function. MCSS, Math. Control Signals Syst., 1989. 2: 303-314.

34. Vucetic, S., Z. Obradovic, and K. Tomsovic, Price-load relationships in California's electricity market. IEEE Trans. Power Syst., 2001. 16: 280-286.

35. Vucetic, S. and Z. Obradovic, Discovering homogeneous regions in spatial data through competition, in Machine learning: Proceedings of the 17th International Conference (ICML '00), 2000, 1091-1098.

36. Riedmiller, M. and H. Braun, A direct adaptive method for faster backpropogation learning: the RPROP algorithm. Proc. IEEE Internat'l. Conf. on Neural Networks, 1993. 1: 586 - 591.

37. Fayyad, U., Reina, C. and Bradley. P, Initialization of iterative refinement clustering algorithms. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, 1998, 194-198.

38. Vihinen, M., E. Torkkila, and P. Riikonen, Accuracy of protein flexibility predictions. Proteins, 1994. 19: 141-149.

39. Strauss, P.R. and C.M. Holt, Domain mapping of human apurinic/apyrimidinic endonuclease. Structural and functional evidence for a disordered amino terminus and a tight globular carboxyl domain. J. Biol. Chem., 1998. 273: 14435-14441.

40. Rao, V.D., S. Misra, I.V. Boronenkov, R.A. Anderson, and J.H. Hurley, Structure of type IIbeta phosphatidylinositol phosphate kinase: A protein kinase fold flattened for interfacial phosphorylation. Cell, 1998. 94: 829-839.

41. Venyaminov, S., A. Gudkov, Z. Gogia, and L. Tumanova, Absorption and circular dichroism spectra of individual proteins from *Escherichia coli* ribosomes. 1981, Pushchino: AS USSR: Biological Research Center, Institute of Protein Research. 128.

42. Garner, E., P. Romero, A.K. Dunker, C. Brown, and Z. Obradovic, Predicting binding regions within disordered proteins. Genome Inform. Ser. Workshop Genome Inform., 1999. 10: 41-50.

43. Jain, A.K. and R.C. Dubes, Algorithms for Clustering Data. 1988, Englewood Cliffs, NJ, USA: Prentice Hall.

44. Dempster, A.P., N.M. Laird, and D.B. Rubin, Maximum likelihood from data via the EM algorithm. J. R. Stat. Soc., 1977. 39: 1-38.

45. Dunker, A.K., Z. Obradovic, P. Romero, E.C. Garner, and C.J. Brown, Intrinsic protein disorder in complete genomes. Genome Inform. Ser. Workshop Genome Inform., 2000. 11: 161-171.

46. Liu, J. and B. Rost, Comparing function and structure between entire proteomes. Protein Sci., 2001. 10: 1970-1979.

47. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z., and Dunker, A.K., Intrinsic disorder in cell-signaling and cancer-associated proteins, J. Mol. Biol., 2002, 323: 573-854.

48. Frishman, D., K. Albermann, J. Hani, K. Heumann, A. Metanomski, A. Zollner, and H.W. Mewes, Functional and structural genomics using PEDANT. Bioinformatics, 2001. 17: 44-57.

49. Gaasterland, T., Structural genomics: Bioinformatics in the driver's seat. Nat. Biotechnol., 1998. 16: 625-627.

50. Shapiro, L. and C.D. Lima, The argonne structural genomics workshop: Lamaze class for the birth of a new science. Structure, 1998. 6: 265-267.

Table I. Algorithm for discovery of protein disorder flavors and its competition procedure

Algorithm for discovering protein disorder flavors

- Learn a single predictor using all disordered proteins S_0 and calculate its accuracy.
- Split S_0 into $\mathbf{S}_2^0 = \{S_{1,2}^0, S_{2,2}^0\}$, where $S_{1,2}^0$ and $S_{2,2}^0$ are disjoint and of equal size.
- Modify $\mathbf{S}_2^0$ through the competition procedure and calculate the overall accuracy achieved by the obtained partitioning $\mathbf{S}_2 = \{S_{1,2}, S_{2,2}\}$.
- Terminate the algorithm if the accuracy is not improved. There are no disorder "flavors".
- L=2.

repeat

  for i = 1 to L

  - Split $S_{i,\mathrm{L}}$ into two disjoint subsets of the same sizes, $S_{i,\mathrm{L}}'$ and $S_{i,\mathrm{L}}''$, to obtain the initial partitioning $\mathbf{S}_{\mathrm{L}+1,i}^0 = \{\mathbf{S}_\mathrm{L} \backslash S_{i,\mathrm{L}}, S_{i,\mathrm{L}}', S_{i,\mathrm{L}}''\}$.
  - Modify $\mathbf{S}_{\mathrm{L}+1,i}^0$ through *the competition procedure* and calculate the overall accuracy of the resulting partitioning $\mathbf{S}_{\mathrm{L}+1,i}$.

  end

- Out of L options, $\mathbf{S}_{\mathrm{L}+1,i}$, i = 1, …L, choose the partitioning achieving the highest overall accuracy to represent the new partitioning, $\mathbf{S}_{\mathrm{L}+1} = \{S_{1,\mathrm{L}+1}, \ldots S_{\mathrm{L}+1,\mathrm{L}+1}\}$.
- $\mathrm{L} \leftarrow \mathrm{L} + 1$

until there is improvement in accuracy

Output: L–1 "flavors" defined by partitioning $\mathbf{S}_{\mathrm{L}-1}$.

Competition Procedure

Start from a partitioning into L subsets $\mathbf{S}_\mathrm{L}^0$ and set n=0.

repeat

  for each disordered protein j

  - Learn L models M_i, i=1, …L, using the L subsets of $\mathbf{S}_\mathrm{L}^n$, S_i^n, i=1, …L, excluding disordered protein j.
  - If M_i gives the best accuracy on disordered protein j assign this protein to subset S_i^{n+1}.

  end

- Form a new partitioning, $\mathbf{S}_\mathrm{L}^{n+1} = \{S_1^{n+1}, \ldots S_\mathrm{L}^{n+1}\}$.
- $n \leftarrow n+1$

until convergence to a stable solution

Table II. Comparisons of time needed to learn a predictor and predictor accuracy for 3 prediction methods.

Predictor	Time[s]	Accuracy[%]		
		disorder	order	average
OLS	0.66	71.5	84.3	77.8
LR	15.9	74.1	81.9	78.0
NN	45.4	74.0±2.4	82.1±1.5	78.1±2.0

Table III. Accuracies of OLS predictors comparing different windows for data representation ($\mathrm{W_{in}}$) and for output averaging ($\mathrm{W_{out}}$).

$\mathrm{W_{in}}$	$\mathrm{W_{out}}$	Accuracy[%]		
		disorder	order	average
21	1	67.7	80.3	74.0
41	1	71.5	84.3	77.8
81	1	68.7	85.4	77.1
41	41	72.6	88.3	80.5

Table IV. The overall accuracy of the partitioning algorithm as a function of the number of subsets, starting from a global model. *Range over 20 runs* are minimum and maximum overall accuracies over 20 runs of the algorithm. The *Best run* is the bootstrap estimate of 90% confidence interval of the overall accuracy of the best among 20 runs of the algorithm.

Number of subsets	1	2	3	4	5	6
Range over 20 runs [%]	**71.5**	[74.7, **75.9**]	[82.3, **84.4**]	[82.7, **84.9**]	[83.4, **85.2**]	[83.4, **84.8**]
Best run 90% CI [%]	[67.1, 75.9]	[72.3, 79.5]	[81.2, 87.6]	[81.8, 87.9]	[82.2, 88.1]	[81.9, 87.8]

Table V. Accuracy of each of the three flavor predictors (VL-2V, VL-2C, VL-2S) and the global model (VL-2) on the three disorder flavors and on the D_145 and the O_130 data sets.

			Predictor (% accuracy)			
Data Set	# Proteins	# Residues	VL-2	VL-2V	VL-2C	VL-2S
O_130	130	32,506	84.3	81.5	85.7	85.8
D_145	145	16,897*	71.5	64.5	52.9	63.0
Flavor V	52	7,743*	67.1	**83.1**	30.2	49.3
Flavor C	39	3,402*	71.5	51.3	**83.2**	47.4
Flavor S	54	5,752*	75.6	56.1	40.1	**86.9**

*Disordered residues only.

Table VI. Most influential attributes, in decreasing importance, for VL-2 and for each of the three flavor predictors.

PONDR	Most influential attributes[a]
VL-2	$\mathbf{K_2}$, $\mathbf{X_Y}$, $\mathbf{X_W}$, X_I, X_F, X_V, X_L, X_A
VL-2V	$\mathbf{X_W}$, X_A, X_L, X_V, X_Y, X_N
VL-2C	X_F, X_L, X_I, X_D, K_2, X_C, X_Y, X_H
VL-2S	$\mathbf{K_2}$, $\mathbf{X_Y}$, X_I, X_W, X_A, X_F, X_G, X_H, X_V

[a]K_2 is Shannon's entropy and X_A is the frequency of the amino acid A in a window of size 41.

Table VII. D statistics for the difference between amino acid distributions of the three disorder flavors and ordered proteins. P-values in the parentheses correspond to hypothesis tests that the amino acid compositions between flavors are identical.

	Flavor C	Flavor S	Order
Flavor V	476 ($<10^{-3}$)	184 (0.06)	755
Flavor C		428 ($<10^{-3}$)	708
Flavor S			527

Table VIII. Frequency of characterization method and location of disorder in proteins for each of the three flavors.

	Flavor V		Flavor C		Flavor S	
	Number	Percent of Flavor	Number	Percent of Flavor	Number	Percent of Flavor
X-ray	11	21%	15	39%	27	49%
NMR	12	23%	11	29%	12	22%
CD & other	29	56%	12	32%	16	29%
Completely disordered	26	50%	11	29%	10	18%
Internal disorder	12	23%	3	8%	11	20%
N-terminal disorder	10	19%	16	42%	23	42%
C-terminal disorder	4	8%	8	21%	11	20%

Table IX. Comparisons of sequence complexity, net charge, and hydropathy properties

	Sequence Complexity*	Net Charge**	Hydropathy***
Order	3.7 ± 0.2	–2 ± 4	–0.3 ± 0.2
Flavor V	3.6 ± 0.4	3 ± 13	–0.9 ± 0.5
Flavor C	3.3 ± 0.4	0 ± 15	–0.8 ± 0.6
Flavor S	3.4 ± 0.3	–2 ± 16	–0.7 ± 0.5

* Results for K2 entropy calculated over windows of 41 residues

**per 100 residues

***In the hydropathy scale[62], hydrophilic is negative, hydrophobic is positive

Table X. Functions of disordered regions partitioned into flavors V, C, and S. Only functions represented by at least 3 of the D_145 proteins are shown.

Function of disorder	**Flavor**		
	V	**C**	**S**
DNA binding	3	**7**	6
Genomic RNA binding	0	**2**	**1**
Metal binding	5	4	3
Modification sites	4	**8**	3
Protein binding	12	7	**25**
Ribosomal proteins	**9**	1	1

Table XI. Conservative estimates of the percent of proteins with long regions of disorder of each flavor.

Data Set	Proteins predicted to have long disorder[%]			
	VL-2V	VL-2C	VL-2S	At least one predictor
O_PDB_S25	6.8	5.6	5.7	15.6
D_145	58	77	69	72
SwissProt	22	22	28	42.5

Table XII. Conservative estimates of the percent of proteins with long regions of disorder of each flavor for 28 species. Within each kingdom, species are sorted according to increasing commonness of disorder.

Kingdom	Species	Proteins with long disorder[%]			
		VL-2V	VL-2C	VL-2S	At least one predictor
Archaea	*Methanococcus jannaschii*	**20**	5	9	26
	Pyrococcus horikoshii	**19**	6	13	30
	Pyrococcus abyssi	**23**	8	11	31
	Archaeoglobus fulgidus	**20**	9	13	31
	Methanobacterium thermoautotrophicum	**30**	10	22	44
	Aeropyrum pernix	15	**31**	26	51
Average for Archaea		*21*	*12*	*16*	*36*
Bacteria	*Rickettsia prowazekii*	6	2	**11**	16
	Ureaplasma urealyticum	6	3	**15**	20
	Haemophilus influenza	8	9	**11**	21
	Synechocystis sp.	7	10	**14**	23
	Mycoplasma genitalium	7	4	**20**	24
	Campylobacter jejuni	7	3	**19**	24
	Vibrio cholerae	6	13	**14**	25
	Borrelia burgdorferi	7	2	**22**	26
	Bacillus subtilis	**16**	10	12	27
	Xylella fastidiosa	7	**17**	13	28
	Helicobacter pylorl	14	6	**20**	28
	Neisseria meningitidis	12	**17**	10	29
	Chlamydia pneumoniae	13	9	**22**	29
	Chlamydia trachomatis	10	9	**25**	31
	Mycoplasma pneumoniae	9	15	**26**	31
	Thermotoga maritime	**26**	6	18	36
	Treponema pallidum	12	**22**	20	38
	Mycobacterium tuberculosis	7	**38**	14	45
Average for Bacteria		*10*	*11*	*17*	*28*
Eukaryotes	*Caenorhabditis elegans*	32	26	**42**	52
	Saccharomyces cerevisiae	31	24	**51**	58
	Arabidopsis thaliana	32	27	**53**	62
	Drosophila melanogaster	38	46	**53**	67
Average for Eukaryotes		*33*	*31*	*50*	*60*
Overall Average		*16*	*14*	*21*	*34*

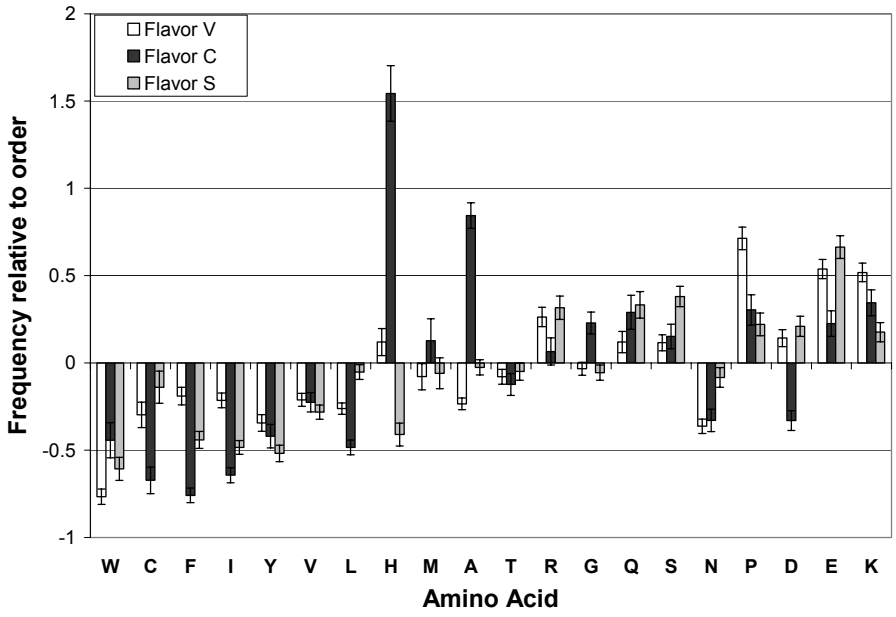

Figure 1. Comparison of amino acid compositions of three disorder flavors with the composition of order. Error bars are one standard deviation.

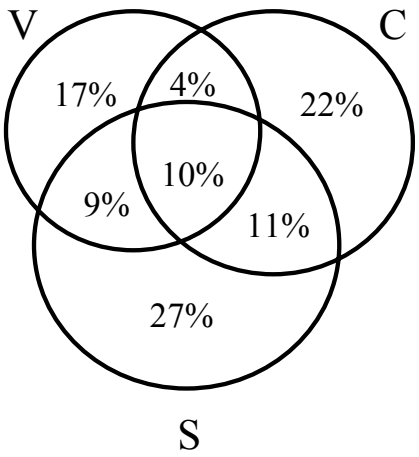

Figure 2. Venn Diagram showing overlap in predictions of flavor predictors on the SwissProt database. Numbers represent the distribution of mono-flavors (V, C, S) and mixed-flavors (VC, VS, CS, VCS) within all SwissProt residues predicted to be disordered by at least one flavor predictor.