# IDENTIFICATION AND FUNCTIONS OF USEFULLY DISORDERED PROTEINS

BY A. KEITH DUNKER, CELESTE J. BROWN AND
School of Molecular Biosciences
Washington State University, Pullman, WA 99164-4660

ZORAN OBRADOVIĆ
Center for Information Science and Technology,
Temple University, Philadelphia, PA 19122



Numerous apparently native proteins have disordered regions and some are wholly disordered, yet both types of protein often utilize their unordered amino acids to carry out function. How do such disordered proteins fit into the view that amino acid sequence codes for protein structure? Our hypothesis was that disorder, like order, is encoded by the amino acid sequence. To test this hypothesis, intrinsically disordered protein was compared with ordered protein. Initially, small sets of ordered and disordered sequences were compared by prediction of order and disorder. Success rates much greater than expected by chance indicated that disorder is encoded by the sequence. Once larger datasets of ordered and disordered proteins were collected, direct sequence comparisons could be made. The amino acid compositions, sequence attributes, and evolutionary characteristics of disordered sequences differed from the corresponding features of ordered sequences in ways commensurate with our hypothesis that disorder is encoded by the sequence. The differences between ordered and disordered sequences enabled the development of predictors of natural disordered regions (PONDRs). Since sequence codes for structure and since sequence codes for disorder, it follows that disorder ought to be considered a category of native protein structure. Two questions arise from this categorization: how common is native protein disorder and what functions are carried out by this category of structure? Application of a particular PONDR to genomic sets of sequences indicated that disordered regions are extremely common, especially in eukaryotic cells. Since function depends on structure and since disorder is a type of structure, disorder-function relationships were examined. A wide variety of functions were found to be associated with disordered structure, with the common use of disorder in signaling and information networks being especially interesting.

## I. Testing Whether Intrinsic Disorder is Encoded by the Amino Acid Sequence

In our usage, an ordered protein contains a single canonical set of Ramachandran angles, whereas a disordered protein or region contains an ensemble of divergent angles at any instant and these angles interconvert over time. Intrinsically disordered protein can be extended (random coil-like) or collapsed (molten globule-like). The latter type of disorder typically includes regions of fluctuating secondary structure, so disorder does not mean absence of helix or sheet. Both types of disorder have been observed in apparently native proteins (Wright and Dyson, 1999; Dunker and Obradovic, 2001)

Intrinsic disorder might not be encoded by the sequence, but rather might be the result of the absence of suitable tertiary interactions. If this were the general cause of intrinsic disorder, then any subset of ordered sequences and any subset of disordered sequences would likely be the same within the statistical uncertainty of the sampling. On the other hand, if intrinsic disorder were encoded by the amino acid sequence, then any subset of disordered sequences would likely differ significantly from samples of ordered protein sequences. Thus, to test the hypothesis that disorder is encoded by the sequence, we collected examples of intrinsically ordered and examples of intrinsically disordered proteins and then determined whether and how their sequences were distinguishable.

A. Using Prediction

In our first attempts to determine whether sequence codes for disorder, lack of resources limited us to the collection of only a small number of disorder examples that contained only about 1,200 residues in total. For such small numbers, differences between ordered and disordered sequences could not be discerned with statistical reliability. Thus, we turned to prediction as a way to estimate whether ordered and disordered sequences are the same or different (Romero *et al.*, 1997b).

In these initial studies, disordered regions were sorted according to length: short = 7 – 21, medium = 22-39, and long = 40 or more residues. As described in more detail previously (Romero et al., 1997b) and below in Section II,A, predictors were developed for each length class and for the three length classes merged together. For predictor training, an initialization is required: 5 independent initializations were used for each predictor. Also, 5-cross validation on disjoint training and testing sets was used, so each result is based on 5 x 5 = 25 sets of predictions.

Since equal numbers of disordered and ordered residues were used for training and testing, prediction success would be about 50% if disordered and ordered sequences were the same. In contrast to this 50% value, prediction success rates for the short, medium, long, and merged datasets were 69% ± 3%, 74 ± 2%, 73% ± 2%, and 60% ± 3%, respectively (Romero et al., 1997b), where the standard errors were determined over about 2,200, 2,600, 2,000 and 6,800 individual predictions, respectively.

The success rate of every prediction set was greater than the value of 50% expected by chance. Specifically, the various sets of predictions differed from the 50% value by about 3 standard deviations (for the lowest success rate, which was for the merged data) to about 12 standard deviations (for the highest success rates, which were for the medium and long regions of disorder). Overall, these data provided very strong support for our hypothesis that disorder is encoded by the amino acid sequence (Romero et al., 1997b).

B. Database Comparisons

To obtain statistically significant comparisons of ordered and disordered sequences, much larger datasets were needed. To this end, disordered regions of proteins or wholly disordered proteins were identified by literature searches to find examples with structural

characterizations that employed one or more of the following methods: 1. x-ray crystallography, where absence of coordinates indicates a region of disorder; 2. nuclear magnetic resonance (NMR), where several different features of the NMR spectra have been used to identify disorder; and 3. circular dichroism (CD) spectroscopy, where whole-protein disorder is identified by a random-coil type CD spectrum.

Once sufficient numbers of intrinsically disordered and ordered protein sequences were collected, it became possible to compare them directly. The sequences in these databases were examined for differences in amino acid composition, sequence attributes, and evolutionary characteristics.

1. Databases of Ordered and Disordered Proteins

Three groups of disordered proteins have been assembled, with the groups defined by the experimental method used to characterize the lack of ordered structure. Because the focus has been on long regions of disorder, an identified disordered protein or region was not included in these groups if it failed to contain 40 or more consecutive residues. Disordered regions from otherwise ordered proteins as well as wholly disordered proteins were identified. Table I summarizes the collection of sequences in this database.

Three groups of ordered sequences have been developed from the Protein Data Bank (PDB) for various purposes. The first group, called Globular 3-D, was formed from NRL 3-D (Pattabiraman *et al.*, 1990) by deletion of the nonglobular proteins. NRL-3D contains essentially all of the residues with backbone coordinates in PDB. Globular 3-D has the advantage of containing the largest number of ordered chains and residues, but has the disadvantage of also containing many proteins with high sequence similarity or even identity. The second group was constructed by deleting the unobserved residues and keeping the observed residues from a non-redundant subset of proteins called PDB_Select_25 (Hobohm and Sander, 1994), yielding the collection called O_PDBS25. Since PDB_Select_25 was formed by grouping PDB proteins into sets having 25% or more sequence identity, the sequence identity between any two proteins in O_PDBS25 is less than 25%. The third group was assembled from the proteins in PDB_Select_25 having no unobserved residues, giving the completely ordered subset or CO_PDBS25. These three groups of ordered proteins are described in Table II.

2. Comparing Amino Acid Compositions

Since different protein folding classes can be identified by differences in their amino acid compositions (Nakashima *et al.*, 1986), we reasoned that, if disorder were encoded by the sequence, then regions of disorder would be analogous to a new folding class and so should be distinguishable by amino acid compositional differences compared to ordered protein.

Figure 1 shows the amino acid compositions and compositional differences of the various protein groups versus amino acid type, where the amino acids are arranged using the "flexibility" scale of Vihinen and co-workers. In this arrangement, the tendency to be buried increases to the left and the tendency to be exposed increases to the right (Vihinen *et al.*, 1994). The compositions of the three disordered sets are very similar to each other (Fig. 1, top), and the compositions of the three ordered sets are even more similar to each other (Fig. 1, middle). The compositional differences show systematic distinctions between ordered and disordered protein (Fig. 1, bottom). Since the differences are calculated as (disorder – order) / (order), positive peaks represent fractional enrichments and negative peaks represent fractional depletions of amino acids in disordered as compared to ordered protein.

Although the disordered proteins were characterized by three different methods, all three datasets of disordered amino acids showed semi-quantitatively similar changes for 16 of the 20 amino acids (Fig. 1, top). Since the three methods rely on completely different underlying biophysical principles for determining disorder, substantial compositional differences among the datasets were expected but surprisingly not observed. Thus, the compositions of the disordered proteins (Fig. 1, top) likely indicate inherent tendencies of this type of protein.

A few proteins are extremely over-represented in PDB and CO_PDBS25 does not contain very many proteins, so Globular 3-D, O_PDBS25 and CO_PDBS25 might exhibit compositional differences. The three groups, however, have nearly the same composition for every amino acid (Fig. 1, middle), so over-representation of some proteins in Globular 3-D and the smaller number of sequences in CO_PDBS25 did not lead to significant amino acid biases.

The compositional differences (Fig. 1, bottom) show that, compared to ordered protein, the three disordered datasets exhibit large and significant depletions of 8 amino acids, namely W, C, F, I, Y, V, L, and N, enrichments in 7, namely K, E, P, S, Q, R and A, and inconsistent changes for 5, namely H, M, T, G, and D.

All the disorder-specific depletions except one are from the leftmost, typically buried amino acids. The one exception, N, is out-of-place in being both a surface-preferring residue and an order-promoting residue. Perhaps the short side chain with its propensity to hydrogen bond to the backbone (Presta and Rose, 1988; Richardson and Richardson, 1988) tends to induce local structure. This order-inducing tendency might explain the out-of-place behavior of N.

The disorder-specific enrichments are mainly from the rightmost, typically exposed amino acids, with the exceptions of N, T, G, and D. Like N as discussed above, T and D also have groups with hydrogen-bonding potential attached to the β-carbon and so can readily form hydrogen bonds with the backbone. Thus, the tendencies of T and D to be less disorder-promoting than their neighbors might also be due to the ordering effects of such hydrogen bonding.

Here G is classified as order-disorder neutral, whereas in a previous study G was classified as disorder-promoting (Williams *et al.*, 2001). This change in classification arises from very small differences between the disordered data in the two studies. The data then and now are not statistically different from each other, neither are the data significantly different between order and disorder for this residue.

During the development of the "flexibility scale" a dual behavior was noticed for G. Specifically, when flanked by residues with high flexibility indices, G exhibited a high average flexibility index, but when flanked by residues with low flexibility indices, G exhibited a low average flexibility. The context dependence of the flexibility index of G was much larger than the context dependence of any other residue. To explain these data, it was suggested that, when in a flexible region, G enhances the flexibility by being able to adopt many conformations, but when in a rigid, buried region, G enhances the rigidity by facilitating tight packing (Vihinen *et al.*, 1994). This dual structural role for G may account for its neutrality with respect to the promotion of order or disorder.

The last two amino acids, H and M, are inconsistent across the three datasets, but both are found between the order-promoting and disorder-promoting sets. Thus, these amino acids exhibit intermediate tendencies to be buried or exposed and like-wise exhibit intermediate tendencies to promote order or disorder, respectively.

The enrichments and depletions displayed in Figure 1 are concordant with what would be expected if disorder were encoded by the sequence (Williams et al., 2001). Disordered regions are depleted in the hydrophobic amino acids that tend to be buried and enriched in the hydrophilic amino acids that tend to be exposed. Such sequences would be expected to lack the ability to form the hydrophobic cores that stabilize ordered protein structure. Thus, these data strongly support the conjecture that intrinsic disorder is encoded by local amino acid sequence information, and not by a more complex code involving, for example, lack of suitable tertiary interactions.

Others have studied the relationship of amino acid composition and protein structure from different points of view. Karlin identified sequences with unusual compositions (Karlin and Brendel, 1992), while Wootton used Shannon's entropy to estimate sequence complexity, showed that nonglobularity was associated with low complexity, and found that sequence databases were much richer than PDB in proteins with regions of low complexity (Wootton, 1993; Wootton, 1994b; Wootton, 1994a; Wootton and Federhen, 1996). Our extensions of Wootton's work revealed that not one of the more than $2.6 \times 10^6$ overlapping 45-residue segments in Globular 3-D contains fewer than 10 different amino acids nor a Shannon's entropy value less than 2.9 (Romero *et al.*, 1999). Attempts to select a folded but simplified SH3 domain by phage display (Riddle *et al.*, 1997) yielded a protein with a greatly reduced average value for Shannon's entropy and a reduced average amino acid alphabet size, yet the lowest of the resulting reduced complexity values were very similar to the lowest observed in Globular 3-D. The near coincidence of the lowest complexity values for both laboratory and natural selection suggested the possibility of a lower bound for the sequence complexity of ordered, globular protein structure (Romero et al., 1999).

3. Comparing Sequence Attributes

Another way of studying amino acid sequences is by means of sequence attributes such as hydropathy, net charge, side chain volume, bulkiness, etc. If P (S | x) is the conditional probability of observing structure type S in a region of sequence having an attribute value of x, graphs of P (S | x) versus x were previously shown to provide useful insight regarding sequence-structure relationships (Arnold *et al.*, 1992).

For S = order or disorder, and for x = average Sweet & Eisenberg hydropathy (Sweet and Eisenberg, 1983) over a window of 21 amino acids, plots of P (S | x) versus x derived from a balanced set of ordered and disordered 21-residue segments gave the data of Figure 2. As expected from the compositional biases displayed in Figure 1, increasing hydropathy values correlate with the formation of ordered structure; e.g. the more hydrophilic the sequence the greater the tendency to be disordered, whereas the more hydrophobic the sequence the greater the tendency to be ordered.

Dividing the area between the two curves in Figure 2 by the total area gives the area ratio (AR), which provides a means for ranking different attributes (Xie *et al.*, 1998). AR values were used to rank 265 attributes using balanced numbers of ordered and disordered segments having 21 residues each. When this approach was used to compare the x-ray, NMR, and CD-characterized sets of disorder with the same (randomly selected) set of ordered segments, the rankings were very similar but not quite identical across the three disorder datasets, suggesting only very slight differences among the disorder characterized by the three different methods (Williams et al., 2001). These results are consistent with the small compositional differences between the differently characterized disordered regions shown in Figure 1.

Uversky and co-workers recently used a pair of sequence attributes, specifically the Kyte and Doolittle hydropathy scale and net charge, to distinguish between folded and "natively unfolded" proteins (Uversky *et al.*, 1999). With an AR of 0.42, the Kyte & Doolittle scale ranked 79$^{th}$ among the 265 scales, well below 5 other hydropathy scales. The Sweet & Eisenberg scale, which ranked 3$^{rd}$ overall with an AR value of 0.538 (Williams et al., 2001), was the best of this type for discriminating between the order and disorder in our datasets. Since the Sweet & Eisenberg scale was developed by determining each amino acid's average degree of exposure in a set of structures from PDB, it is interesting that this scale out-performs other hydropathy scales with regard to discriminating between ordered and disordered segments. Furthermore net charge, with an AR value of 0.236, ranked 174$^{th}$. Thus, the hydropathy-net charge pair is almost certainly not the best pair of attributes for discriminating ordered and disordered sequences. Nevertheless, the work of Uversky and co-workers was a very important contribution due to its simplicity and the insight it provided. Also, this pair of attributes could be optimal for some particular groups of natively unfolded proteins.

Like the differences in composition, the sequence attribute differences between ordered and disordered sequences are exactly as would be expected if disorder were encoded by the sequence. Attribute values of course depend on amino acid compositions, so the composition and attribute results are not independent. However, the attribute analysis is nevertheless useful because it provides a biophysical perspective, allowing insight into the amino acid properties that are important for promoting order or disorder.

4. Comparing Evolutionary Characteristics

A third way of comparing ordered and disordered sequences is by their relative changes over evolutionary time. Constraints are imposed on ordered protein by the requirement to form well-packed protein cores, and these constraints are absent in disordered regions. Disordered regions are generally expected, therefore, to exhibit higher rates and different patterns of amino acid substitution over evolutionary time as compared to ordered regions. However, since function, not structure, is the property subjected to natural selection, special circumstances could mitigate against the expected general trend. Such exceptions could potentially provide additional insight regarding the importance of intrinsic disorder. On a related matter, if differences in evolutionary characteristics are found between ordered and disordered regions, such a result would be a strong indicator that disorder exists *in vivo*.

a. Construction of Disordered Protein Families

In order to test the hypothesis that disordered and ordered proteins differ both in the quantity and the quality of their evolutionary change, families of disordered proteins were developed. Families were constructed for proteins from each of the three disordered groups in Table I and include both proteins that are wholly disordered and proteins with regions of disorder and order. The entire protein sequence was used to identify homologous members of the protein family by BLASTP searches (Altschul *et al.*, 1990; Altschul *et al.*, 1997) on the nonredundant protein database at NCBI (www.ncbi.nlm.nih.gov). Homologues were aligned using the default settings of CLUSTALW (Thompson *et al.*, 1994) at the Baylor College of Medicine website (Smith *et al.*, 1996). Except for the wholly disordered proteins, aligned sequences were then partitioned into ordered sequences and disordered sequences based on alignment with the structurally-characterized sequence. This procedure is outlined in Figure 3.

b. Comparing Substitution Patterns

To test whether disordered and ordered proteins differ in the pattern of their evolutionary change, substitution matrices were constructed based upon 55 aligned disordered protein families. The substitution matrix based upon disordered protein families was then compared to the commonly used substitution matrices, which are based upon mostly ordered protein families. To develop the disordered matrix, initial alignments were performed (Fig. 3) using the BLOSUM62 substitution matrix and the usual first-gap/gap-extension penalties. From the aligned disordered sequences, a new substitution matrix was built. New sequence alignments were developed using the new matrix for disorder. This new-alignment/new-matrix cycle was continued until the change of the matrix in two successive iterations dropped below a pre-specified threshold, thus yielding a substitution matrix specifically for regions of intrinsic disorder. The relative improvement of the final matrix over published substitution matrices was confirmed by tests that used Hidden Markov Models of aligned family and non-family sequences (Radivojac *et al.*, 2002).

This procedure led to a substitution matrix for aligning disordered protein that was different from the commonly used substitution matrices, such as BLOSUM62 (Fig. 4). The matrix for disordered protein is generally better than order-based matrices for aligning disordered proteins whose sequence identities are between 20 - 50%. These results indicate that disordered and ordered protein can be distinguished by their patterns of evolutionary change.

c. Comparing Substitution Rates

To test whether disordered protein evolves more rapidly than ordered protein, comparisons were made between the ordered and disordered regions of 26 protein families with both order and long regions of disorder (Brown *et al.*, 2002). Twenty-four of the families had been structurally characterized by NMR or X-ray crystallography. The ordered and disordered regions in the two CD-characterized proteins had been dissected by limited proteolysis. The pair-wise genetic distance for each ordered region was compared to the pair-wise genetic distance for the corresponding disordered region from the same protein pairs. The average difference between these pair-wise genetic distances ($\Delta$) was calculated for each protein family (Table III). An appropriate statistical test was designed to determine whether $\Delta$ was significantly different from zero (ie. if ordered and disordered regions differed in their rates of evolution). For five families, there were no significant differences in pair-wise genetic distances between ordered and disordered sequences. The disordered region evolved significantly faster than the ordered region for 19 of the 26 families. The functions of these disordered regions are diverse, including binding sites for protein, DNA, or RNA and also including flexible linkers. The functions of some of these regions are unknown. The disordered regions evolved significantly slower than the ordered regions for the two remaining families. The functions of these more slowly evolving disordered regions include sites for DNA binding. These results indicate that, in general, disordered regions of proteins evolve more rapidly than their ordered regions. Understanding the exceptions to these rules may help to further characterize the roles of disorder in protein function.

II. Prediction of Order and Disorder from the Amino Acid Sequence

The results described in Section I suggest that amino acid sequence codes for intrinsic protein disorder. In this circumstance, constructing a predictor of order and disorder would be useful as a means to extend and generalize from the current experimental results.

The steps involved in building disordered predictors are the following: 1. Develop datasets of ordered and disordered protein; 2. Identify a set of features or attributes for

discriminating between order and disorder; and 3. Use an appropriate set of features or attributes as the basis for predictors of intrinsic order and disorder, while taking care to use disjoint subsets of sequences for training and testing. Each of these steps has been investigated in some detail as described above in Section I,A or as reported previously (Romero *et al.*, 1997a; Romero et al., 1997b; Li *et al.*, 1999; Romero *et al.*, 2000; Vucetic *et al.*, 2001).

A. Predictors of Natural Disordered Regions (PONDRs)

Applying standard machine learning algorithms and approaches to databases and sequence information as described above in Section I, a series of predictors of intrinsic disorder and order called PONDRs have been developed. These predictors use amino acid sequence as inputs and give numerical outputs, with 0.0 to < 0.5 indicating order and with ≥ 0.5 to 1.0 indicating disorder. Various data representations for the inputs have been tried, such as the compositions of selected amino acids or the averaged values of selected sequence attributes including hydropathy, net charge, aromaticity, etc. In addition, the data representation typically involved calculating the inputs over sliding windows, and some experimentation to optimize window size has been carried out. Finally, various operations on the input data have been tried, including both linear data modeling, such as logistic regression, and non-linear modeling, such as artificial neural networks, to yield predictors of intrinsic order and disorder (Li et al., 1999; Vucetic et al., 2001). The various experiments suggested that data representation, window size, and linear or nonlinear data modeling make relative small differences in prediction accuracies.

To distinguish among the several PONDRs that resulted from the various experiments described above, a two-letter extension and a version number were added. The first letter, X, N, or V, indicates the method of structural characterization of the training data, namely x-ray diffraction, NMR, or various means, respectively. The second letter, S, M, L, N, C or T, indicates the length or position of the disordered regions in the training data, with S for short (length = 9 to 20 residues), M for medium (length = 21 to 39 residues), L for long (length ≥ 40 residues), N for amino termini, C for carboxy termini, and T for both termini (length = 5 to 14 residues). Although predictors were initially developed for the three indicated length classes (Romero et al., 1997b), since then the focus has been on long regions of disorder (Romero *et al.*, 1998). PONDRs have also been developed that use disorder information from a single family of proteins; the extension for these PONDRs is the abbreviation of the name of the representative family member. For example, the putative disordered regions from 13 calcineurin (CaN) proteins were used to construct PONDR CaN (Romero et al., 1997a). Finally, as stated above, for each type of predictor, a version number is specified.

One recent predictor is called PONDR VL-XT (Romero *et al.*, 2001), because it is a merger of PONDR VL1 (training set = variously characterized, long regions of disorder) with PONDR XN and PONDR XC, i.e. PONDR XT (training set = x-ray characterized chain termini) (Li *et al.*, 2000), while another is called PONDR VL2 (training set = variously characterized, long regions of disorder, version 2) (Vucetic *et al.*, 2002).

B. PONDR Accuracies

The success of the initial PONDRs based on small databases of disordered protein motivated attempts to improve predictor accuracy. The main limitation for such attempts has been and continues to be the lack of low-noise structural data for both ordered and disordered protein, where noise means ordered regions misclassified as disordered and *vice-versa*.

The accuracies of the various PONDRs were estimated (Table IV) by applying them to the ordered sequences in O_PDBS25 as summarized in Table II and to the merged set of disordered proteins described in Table I. Overall, the prediction accuracy of each PONDR was much better on the 222,116 ordered residues of O_PDBS25 than on the 18,833 residues of the merged disorder set. Thus, prediction of order generalized much better than prediction of disorder.

Even when as few as 502 ordered and 502 disordered residues were used for predictor development, the accuracy of 73 ± 4% estimated from 5-cross validation during training matched within 1 standard deviation the accuracy of 71% observed when the predictor was applied to O_PDBS25, which had 222,116 ordered residues. Since O_PDBS25 contains one representative from each protein family in PDB, this database spans the information on ordered protein structure. Such a good generalization from such a small training set was totally unexpected. Every PONDR so far tested, including some others not among the 5 examples given in Table IV, show comparably good generalization for prediction of protein order.

The poorer generalization of the prediction of disorder probably arises from two or perhaps three sources. These relate to differences in the training sets, possible differences in the volumes of sequence space occupied by ordered and disordered protein, and differences in the levels of noise in the ordered and disordered testing data.

The ordered training data involved non-overlapping ordered segments randomly selected from different proteins, whereas the disordered training data involved overlapping windows that were generated by sliding windows in single-residue steps. This strategy was chosen to maximize use of the limited amount of disordered data. Thus, the disordered training data spanned a smaller volume of sequence space than did the ordered data.

The volume of sequence space occupied by natively disordered proteins might simply be much larger than the corresponding region for ordered proteins. Lattice-model approximations of protein folding suggest that only a small fraction of random sequences fold into specific structure (Abkevich *et al.*, 1996). A random library containing 80-residue sequences designed to have helical periodicities with an average hydrophobicity level similar to that of natural proteins yielded only a small fraction that exhibited cooperative folding behavior (Davidson *et al.*, 1995), providing further evidence that folding sequences represent a small proportion of all sequences. If the ordered and disordered sequences utilized in nature reflect their statistical abundances, then disordered protein sequence space would be much larger than ordered protein sequence space. In this case, a greater number of disorder examples would be needed to achieve the same level of generalization as observed for the ordered data. Note the especially poor performance of PONDR CaN on the disordered set; the poor performance probably relates substantially to the small region of disordered sequence space sampled by the CaN regions of disorder.

A third possible origin of the reduced performance of disorder prediction is that the disordered data may simply be noisier than the ordered data, i.e. the disordered data might simply contain more ordered residues that are misclassified as disordered than *vice versa*. NMR probably yields the best characterization of disorder, yet NMR-characterized regions of disorder very likely contain significant regions of misclassified order (Garner *et al.*, 1999). The misclassification problem is likely to be worse for both the CD- and x-ray characterized data as compared to the NMR data. Since CD spectra provide estimates of the structure averaged over the entire sequence, an ordered domain within a sea of disorder could easily be

missed in the CD spectrum. Long regions of missing coordinates in x-ray structures could be structured but wobbly domains rather than true disorder.

Despite obtaining all the ordered data from crystal structures, which gives accurate structural information and despite the removal of the unobserved (disordered) residues, it still cannot be assumed that the ordered data in Globular 3-D, O_PDBS25, and CO_PDBS25 are noiseless. For example, examination of segments from O_PDBS25 having the longest consecutive (putatively false positive) predictions of disorder shows many of these segments to be associated with DNA, with other large ligands including other proteins, or with the contacts that form the crystal lattice. Since disorder-to-order transitions upon complex formation or upon crystallization can occur, it is difficult to know whether such segments are ordered or disordered in the absence of the inter-molecular association. Thus, these segments very possibly correspond to segments that are actually disordered when not ligand-associated or when not in the crystal, in which case these segments would represent noise in the ordered data.

The next phase of testing PONDR accuracy will be the careful comparison of experiment and prediction on individual proteins. Figure 5 shows results from one such analysis, in which proteolysis was used to test PONDR VL-XT accuracy. This test was based on the knowledge that disordered protein is orders of magnitude more sensitive to protease digestion than is ordered protein (Fontana *et al.*, 1997). The sequence of *Xeroderma pigmentosum* group A (XPA), which is a DNA damage-recognition protein, was the test case. The predictions indicated that the XPA sequence has long regions of intrinsic disorder at both ends. Identification of hypersensitive trypsin digestion sites by mass spectrometry revealed a remarkable agreement between predicted disorder and protease sensitivity, showing that the PONDR indications of disorder were very accurate for this protein. These results suggested that the combination of PONDR predictions + protease digestion + mass spectrometry offers a useful approach for the analysis of protein disorder (Iakoucheva *et al.*, 2001).

III. PONDR Estimations of the Commonness of Intrinsically Disordered Proteins

More than 30 proteomes have been PONDRed to estimate the commonness of putative disorder. We have used 40 or more consecutive predictions of disorder (e.g. a putative long disordered region, or LDR) as a convenient indicator, and then scored each proteome by estimating the fraction of proteins predicted to contain at least one LDR. As shown in Table V, proteins with putative LDRs are quite common. The wide range of putative LDRs among the proteomes from eubacteria and archaea was quite surprising. A second surprise was the large jump in putative LDRs in the eukaryota compared to the eubacteria and archaea.

Further discussion of prediction errors provides more insight into these estimates of the commonness of disorder. Based on O_PDBS25, VL-XT gave a false positive prediction of disorder of ~ 20% on a per-residue basis (Table IV); this error decreases to ~ 0.4% for consecutive predictions of 40 or longer (Romero et al., 2001). These error rates lead to ~6% of the non-redundant proteins from PDB having consecutive false positive predictions of disorder $\geq$ 40 residues in length. As discussed above in Section II,B, this may be an over-estimate of the false positive error rate because many of the apparent consecutive errors correspond to regions of disorder that are ordered in the crystal due to ligand binding or crystal contacts. Also, since disordered regions of length $\geq$ 40 residues are often missed due to false negative predictions of order, the data in Table V probably represent lower bounds on the amount of disorder per genome.

VI. Functions of Intrinsically Disordered Regions

We are attempting to understand the biological significance of the large variations in frequency of putative LDRs, whether between different types of bacteria or archaea, or between pro- and eukaryota. We have carefully studied the literature of more than 90 example proteins selected from our disordered protein databases and found reports on the functions most of the disordered regions (Dunker *et al.*, 2002). The observed functions and the number of examples in each functional class are given in Table VI. As indicated, four major functional classes were found: molecular recognition, molecular assembly or disassembly, protein modification, and entropic chains.

For two of the categories, molecular recognition and protein modification, the proteins in Table VI are mostly involved in signaling, control or regulation. Thus, our current hypothesis is that an increased requirement for signaling, regulation and control is the underlying cause for the significantly larger fractions of proteins with long regions of intrinsic disorder in some organisms as compared to others (Dunker & Obradovic, 2001). Consistent with our current hypothesis but developed without regard to the role of protein disorder, recent proteomic comparisons indicate that eukaryotes are much richer in regulatory proteins than are the prokaryotes and archaea (Liu and Rost, 2001). Tests are underway to determine whether the increased amount of predicted disorder is indeed associated with an increased number of regulatory proteins in the various proteomes.

## V. Conclusions

The experiments and data presented herein support the proposals that protein disorder is encoded by the amino acid sequence and that protein disorder is essential for many important biological functions. Thus, intrinsic disorder ought to be considered a distinct category of *native* protein structure. Categorization is of course an essential step in the development of knowledge (Lakoff, 1987), so the concept that intrinsic disorder represents a *category of native protein structure*, rather than being simply an intermediate on the way to the native structure, has important implications (Dunker *et al.*, 1997; Dunker *et al.*, 2001; Dunker et al., 2002). The association of protein disorder with function is not a new idea. Experiments reported more than 50 years ago gave strong indications that an ensemble of structures enabled serum albumin to bind a structurally-diverse set of ligands (Karush, 1950), and numerous additional papers were published on disorder/function relationships about 20 years ago for several other proteins (Stubbs *et al.*, 1977; Bloomer *et al.*, 1978; Bode *et al.*, 1978; Jardetzky *et al.*, 1978; Huber, 1979; Schulz, 1979; Blow, 1982; Holmes, 1983; Bennett and Huber, 1984). Our extensions of current disorder knowledge by predictions upon whole proteomes leads to the conclusion that functions carried out by disordered regions are likely to be very common, especially with regard to signaling and regulatory functions in eukaryotic cells.

References


Abkevich, V. I., Gutin, A. M., and Shakhnovich, E. I. (1996). *Proc. Natl. Acad. Sci. USA* **93**, 839-844.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389-3402.

Arnold, G. E., Dunker, A. K., Johns, S. J., and Douthart, R. J. (1992). *Proteins: Struct., Func. Gen.* **12**, 382-399.

Aviles, F. J., Chapman, G. E., Kneale, G. G., Crane-Robinson, C., and Bradbury, E. M. (1978). *Eur. J. Biochem.* **88**, 363-371.

Bennett, W. S., and Huber, R. (1984). *Crit. Rev. Biochem.* **15**, 291-384.

Berger, J. M., Gamblin, S. J., Harrison, S. C., and Wang, J. C. (1996). *Nature* **379**, 225-232.

Bidwell, L. M., McManus, M. E., Gaedigk, A., Kakuta, Y., Negishi, M., Pedersen, L., and Martin, J. L. (1999). *J. Mol. Biol.* **293**, 521-530.

Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R., and Klug, A. (1978). *Nature* **276**, 362-368.

Blow, D. M. (1982). *Nature* **297**, 454.

Bode, W., Schwager, P., and Huber, R. (1978). *J. Mol. Biol.* **118**, 99-112.

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T., Oldfield, C. J., Williams, C. J., and Dunker, A. K. (2002). *J. Mol. Evol.* in press.

Campbell, K. M., Terrell, A. R., Laybourn, P. J., and Lumb, K. J. (2000). *Biochemistry* **39**, 2708-2713.

Choi, H. K., Tong, L., Minor, W., Dumas, P., Boege, U., Rossmann, M. G., and Wengler, G. (1991). *Nature* **354**, 37-43.

Davidson, A. R., Lumb, K. J., and Sauer, R. T. (1995). *Nat. Struct. Biol.* **2**, 856-864.

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002). *Biochemistry* in press.

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001). *J. Mol. Graph. Model.* **19**, 26-59.

Dunker, A. K., and Obradovic, Z. (2001). *Nat. Biotech.* **19**, 805-806.

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000). *Genome Informatics* **11**, 161-171.

Dunker, A. K., Obradovic, Z., Romero, P., Kissinger, C., and Villafranca, E. J. (1997). *PDB Newsletter* **81**, 3-5.

Fontana, A., Zambonin, M., Polverino de Laureto, P., De Filippis, V., Clementi, A., and Scaramella, E. (1997). *J. Mol. Biol.* **266**, 223-230.

Garner, E., Romero, P., Dunker, A. K., Brown, C. J., and Obradovic, Z. (1999). *Genome Informatics.* **10**, 41-50.

Gorman, M. A., Morera, S., Rothwell, D. G., de La Fortelle, E., Mol, C. D., Tainer, J. A., Hickson, I. D., and Freemont, P. S. (1997). *EMBO J.* **16**, 6548-6558.

Hobohm, U., and Sander, C. (1994). *Prot. Sci.* **3**, 522-524.

Holmes, K. C. (1983). *CIBA Found. Symp.* **93**, 116-138.

Hopper, P., Harrison, S. C., and Sauer, R. T. (1984). *J. Mol. Biol.* **177**, 701-713.

Horvath, M. P., Schweiker, V. L., Bevilacqua, J. M., Ruggles, J. A., and Schultz, S. C. (1998). *Cell* **95**, 963-974.
Huang, Y., Komoto, J., Konishi, K., Takata, Y., Ogawa, H., Gomi, T., Fujioka, M., and Takusagawa, F. (2000). *J. Mol. Biol.* **298**, 149-162.
Huber, R. (1979). *Nature* **280**, 538-539.
Iakoucheva, L. M., Kimzey, A. L., Masselon, C. D., Bruce, J. E., Garner, E. C., Brown, C. J., Dunker, A. K., Smith, R. D., and Ackerman, E. J. (2001). *Prot. Sci.* **10**, 560-571.
Iwata, S., Lee, J. W., Okada, K., Lee, J. K., Iwata, M., Rasmussen, B., Link, T. A., Ramaswamy, S., and Jap, B. K. (1998). *Science* **281**, 64-71.
Jacobs, D. M., Lipton, A. S., Isern, N. G., Daughdrill, G. W., Lowry, D. F., Gomes, X., and Wold, M. S. (1999). *J. Biomol. NMR* **14**, 321-331.
Jardetzky, O., Akasaka, K., Vogel, D., Morris, S., and Holmes, K. C. (1978). *Nature* **273**, 564-566.
Jimenez, M. A., Evangelio, J. A., Aranda, C., Lopez-Brauet, A., Andreu, D., Rico, M., Lagos, R., Andreu, J. M., and Monasterio, O. (1999). *Prot. Sci.* **8**, 788-799.
Karlin, S., and Brendel, V. (1992). *Science* **257**, 39-49.
Karush, F. (1950). *J. Am. Chem. Soc.* **72**, 2705-2713.
Kim, K. K., Kim, R., and Kim, S. H. (1998). *Nature* **394**, 595-599.
Kissinger, C. R., Parge, H. E., Knighton, D. R., Lewis, C. T., Pelletier, L. A., Tempczyk, A., Kalish, V. J., Tucker, K. D., Showalter, R. E., Moomaw, E. W., Gastinel, L. N., Habuka, N., Chen, X., Maldanado, F., Barker, J. E., Bacquet, R., and Villafranca, J. E. (1995). *Nature* **378**, 641-644.
Lakoff, G. (1987). "Women, Fire and Dangerous Things: What categories Reveal About the Human Mind", University of Chicago Press, Chicago.
Lapthorn, A. J., Harris, D. C., Littlejohn, A., Lustbader, J. W., Canfield, R. E., Machin, K. J., Morgan, F. J., and Isaacs, N. W. (1994). *Nature* **369**, 455-461.
Li, X., Obradovic, Z., Brown, C. J., Garner, E. C., and Dunker, A. K. (2000). *Genome Informatics* **11**, 172-184.
Li, X., Romero, P., Rani, M., Dunker, A. K., and Obradovic, Z. (1999). *Genome Informatics* **10**, 30-40.
Liu, J., and Rost, B. (2001). *Protein Sci.* **10**, 1970-1979.
Logan, D. T., Mazauric, M. H., Kern, D., and Moras, D. (1995). *EMBO J.* **14**, 4156-4167.
Louie, G. V., Yang, W., Bowman, M. E., and Choe, S. (1997). *Mol. Cell* **1**, 67-78.
Mosyak, L., Reshetnikova, L., Goldgur, Y., Delarue, M., and Safro, M. G. (1995). *Nat. Struct. Biol.* **2**, 537-547.
Muchmore, S. W., Sattler, M., Liang, H., Meadows, R. P., Harlan, J. E., Yoon, H. S., Nettesheim, D., Chang, B. S., Thompson, C. B., Wong, S. L., Ng, S. L., and Fesik, S. W. (1996). *Nature* **381**, 335-341.
Nakashima, H., Nishikawa, K., and Ooi, T. (1986). *J. Biochem. (Tokyo)* **99**, 153-162.
Pattabiraman, N., Namboodiri, K., Lowrey, A., and Gaber, B. P. (1990). *Prot. Seq. Data Anal.* **3**, 387-405.
Presta, L. G., and Rose, G. D. (1988). *Science* **240**, 1632-1641.
Radivojac, P., Obradovic, Z., Brown, C. J., and Dunker, A. K. (2002). *PSB 2002* **7**, 589-600.
Richardson, J. S., and Richardson, D. C. (1988). *Science* **240**, 1648-1652.
Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q., and Baker, D. (1997). *Nat. Struct. Biol.* **4**, 805-809.

Riek, R., Hornemann, S., Wider, G., Glockshuber, R., and Wuthrich, K. (1997). *FEBS Lett.* **413**, 282-288.
Romero, P., Obradovic, Z., and Dunker, A. K. (1997a). *Genome Informatics* **8**, 110-124.
Romero, P., Obradovic, Z., and Dunker, A. K. (1999). *FEBS Lett.* **462**, 363-367.
Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., and Dunker, A. K. (1997b). *Proc. IEEE Intl. Conf. Neural Networks* **1**, 90-95.
Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., Guilliot, S., Garner, E., and Dunker, A. K. (1998). *PSB 1998* **3**, 437-448.
Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001). *Proteins: Struc., Funct., Gen.* **42**, 38-48.
Romero, P., Obradovic, Z., and Dunker, A.K. (2000). *Artificial Intelligence Rev.* **14**, 447-484.
Schmitz, M. L., dos Santos Silva, M. A., Altmann, H., Czisch, M., Holak, T. A., and Baeuerle, P. A. (1994). *J. Biol. Chem.* **269**, 25613-25620.
Schulz, G. E. (1979). In "Molecular Mechanism of Biological Recognition" (Balaban, M., ed.), pp 79-94. Elsevier/North-Holland Biomedical Press, New York.
Silva, A. M., and Rossmann, M. G. (1985). *ACTA Cryst.* **B41**, 147-157.
Smith, R. F., Wiese, B. A., Wojzynski, M. K., Davison, D. B., and Worley, K. C. (1996). *Genome Res.* **6**, 454-462.
Stubbs, G., Warren, S., and Holmes, K. (1977). *Nature* **267**, 216-221.
Sweet, R. M., and Eisenberg, D. (1983). *J. Mol. Biol.* **171**, 479-488.
Tell, G., Perrone, L., Fabbro, D., Pellizzari, L., Pucillo, C., De Felice, M., Acquaviva, R., Formisano, S., and Damante, G. (1998). *Biochem. J.* **329**, 395-403.
Tesmer, J. J., Berman, D. M., Gilman, A. G., and Sprang, S. R. (1997). *Cell* **89**, 251-261.
Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). *Nuc. Acids Res.* **22**, 4673-4680.
Tucker, P. A., Tsernoglou, D., Tucker, A. D., Coenjaerts, F. E., Leenders, H., and van der Vliet, P. C. (1994). *EMBO J.* **13**, 2994-3002.
Uversky, V. N., Gillespie, J. R., Millett, I. S., Khodyakova, A. V., Vasiliev, A. M., Chernovskaya, T. V., Vasilenko, R. N., Kozlovskaya, G. D., Dolgikh, D. A., Fink, A. L., Doniach, S., and Abramov, V. M. (1999). *Biochemistry* **38**, 15009-15016.
Vihinen, M., Torkkila, E., and Riikonen, P. (1994). *Proteins* **19**, 141-149.
Vonderviszt, F., Kanto, S., Aizawa, S., and Namba, K. (1989). *J. Mol. Biol.* **209**, 127-133.
Vucetic, S., Brown, C.J., Dunker, A.K., and Obradovic, Z. (2002). *Bioinformatics,* submitted.
Vucetic, S., Radivojac, P., Obradovic, Z., Brown, C. J., and Dunker, A. K. (2001). *Intl. Joint INNS-IEEE Conf. Neural Networks* **4**, 2718-2723.
Williams, R. M., Obradovic, Z., Mathura, V., Braun, W., Garner, E. C., Young, J., Takayama, S., Brown, C. J., and Dunker, A. K. (2001). *PSB 2001* **6**, 89-100.
Wootton, J. C. (1993). *Comput. Chem.* **17**, 149-163.
Wootton, J. C. (1994a). *Comput. Chem.* **18**, 269-285.
Wootton, J. C. (1994b). *Curr. Opin. Struct. Biol.* **4**, 413-421.
Wootton, J. C., and Federhen, S. (1996). *Meth. in Enzymol.* **266**, 554-571.
Wright, P. E., and Dyson, H. J. (1999). *J. Mol. Biol.* **293**, 321-331.
Xie, Q., Arnold, G. E., Romero, P., Obradovic, Z., Garner, E., and Dunker, A. K. (1998). *Genome Informatics* **9**, 193-200.

Table I. Number of proteins and residues in databases of intrinsically disordered protein characterized by various methods.

| Detection Method | Number of Proteins | Number of Residues |
|---|---|---|
| X-ray | 59 | 3,907 |
| NMR | 43 | 4,108 |
| CD | 55 | 10,818 |
| Merged | 157 | 18,833 |

Table II. Number of proteins and residues in databases of intrinsically ordered protein.

| Name | Number of Proteins | Number of Residues |
|---|---|---|
| Globular 3-D | 14,540 | 2,610,197 |
| O_PDBS25 | 1,021 | 222,116 |
| CO_PDBS25 | 130 | 32,509 |

Table III. Average difference in genetic distance, **Δ**, between ordered and disordered regions of 26 protein families.

| Protein Family | Reference | Detection Method[a] | # Seq. | $\Delta$[b] | p-value[c] |
|---|---|---|---|---|---|
| Replication protein A | (Jacobs *et al.*, 1999) | NMR | 7 | 1.92 | 0.001 |
| NF-KB p65 | (Schmitz *et al.*, 1994) | NMR | 4 | 1.18 | 0.001 |
| Glycyl-tRNA synthetase | (Logan *et al.*, 1995) | X-ray | 24 | 1.69 | 0.002 |
| Regulator of G-protein signaling 4 | (Tesmer *et al.*, 1997) | X-ray | 17 | 0.96 | 0.001 |
| Topoisomerase II | (Berger *et al.*, 1996) | X-ray | 28 | 0.87 | 0.001 |
| Calcineurin | (Kissinger *et al.*, 1995) | X-ray | 23 | 0.84 | 0.001 |
| c-Fos | (Campbell *et al.*, 2000) | NMR | 23 | 0.82 | 0.001 |
| Thyroid transcription factor | (Tell *et al.*, 1998) | CD, LP | 12 | 0.76 | 0.001 |
| Sulfotransferase | (Bidwell *et al.*, 1999) | X-ray | 12 | 0.74 | 0.013 |
| Phenylalanine-tRNA synthetase | (Mosyak *et al.*, 1995) | X-ray | 14 | 0.69 | 0.001 |
| Coat protein, tomato bushy stunt virus | (Hopper *et al.*, 1984) | X-ray | 7 | 0.63 | 0.001 |
| Gonadotropin | (Lapthorn *et al.*, 1994) | X-ray | 9 | 0.61 | 0.001 |
| Coat protein, sindbis virus | (Choi *et al.*, 1991) | X-ray | 6 | 0.60 | 0.025 |
| Histone H5 | (Aviles *et al.*, 1978) | NMR | 9 | 0.41 | 0.001 |
| Small heat shock protein | (Kim *et al.*, 1998) | X-ray | 6 | 0.36 | 0.457 |
| Telomere binding protein | (Horvath *et al.*, 1998) | X-ray | 8 | 0.29 | 0.001 |
| Cytochrome BC1 | (Iwata *et al.*, 1998) | X-ray | 7 | 0.27 | 0.034 |
| DNA-lyase | (Gorman *et al.*, 1997) | X-ray[d] | 8 | 0.18 | 0.001 |
| Bcl-$x_L$ | (Muchmore *et al.*, 1996) | X-ray, NMR | 7 | 0.13 | 0.001 |
| Coat protein, southern bean mosaic virus | (Silva and Rossmann, 1985) | X-ray | 6 | 0.09 | 0.100 |
| α-Tubulin | (Jimenez *et al.*, 1999) | NMR | 80 | 0.06 | 0.034 |
| Epidermal growth factor | (Louie *et al.*, 1997) | X-ray | 10 | 0.03 | 0.736 |
| Prion | (Riek *et al.*, 1997) | NMR | 72 | 0.03 | 0.636 |
| Glycine N-methyltransferase | (Huang *et al.*, 2000) | X-ray | 11 | 0.09 | 0.095 |
| ssDNA binding protein | (Tucker *et al.*, 1994) | X-ray | 20 | 0.37 | 0.010 |
| Flagellin | (Vonderviszt *et al.*, 1989) | LP | 34 | 0.66 | 0.023 |

[a]Disordered state detected by NMR=nuclear magnetic resonance, X-ray=X-ray crystallography, CD=circular dichroism, LP=limited proteolysis

[b]Negative values of **Δ** indicate disordered regions are evolving faster than ordered.

[c] P-values for a two-sided test of the null hypothesis.

[d] Useful crystallization only in the absence of most of the disordered region.

Table IV. Accuracies of neural network predictors of natural disordered regions (PONDR).

| Name | Training Set | # Disordered Residues | Accuracy % | | |
|---|---|---|---|---|---|
| | | | Train | Order | Disorder |
| XL1 | 7 X-ray | 502 | 73 ± 4 | 71 | 47 |
| CaN | 13 CaN | 1,175 | 83 ± 5 | 84 | 29 |
| VL1 | 7 NMR<br>8 X-ray | 1,366 | 83 ± 2 | 83 | 45 |
| VL-XT | Merger of VL1, XN and XC | | | 80 | 60 |
| VL2 | 35 NMR<br>52 X-ray<br>56 CD | 16,854 | 79 ± 4 | 83 | 75.5 |

Table V. Prevalence of predicted disorder in genomes of various species[a].

| Kingdom | Species | # seqs | Disorder | Lengths $\geq 40$[b] |
|---|---|---|---|---|
| Archaea | *Methanococcus jannaschii* | 1714 | | 9% |
| Archaea | *Pyrococcus horikoshii* | 2062 | | 16% |
| Archaea | *Pyrococcus abyssi* | 1764 | | 19% |
| Archaea | *Archaeoglobus fulgidus* | 2402 | | 20% |
| Archaea | *Methanobacterium thermoautotrophicum* | 1869 | | 34% |
| Archaea | *Halobacterium sp.*NRC-1 | 2057 | | 35% |
| Archaea | *Aeropyrum pernix* K1 | 2694 | | 37% |
| Bacteria | *Ureaplasma urealyticum* | 611 | | 7% |
| Bacteria | *Rickettsia prowazekii* | 834 | | 6% |
| Bacteria | *Borrelia burgdorferi* | 845 | | 7% |
| Bacteria | *Campylobacter jejuni* | 2309 | | 6% |
| Bacteria | *Mycoplasma genitalium* | 480 | | 8% |
| Bacteria | *Helicobacter pylori* | 1532 | | 9% |
| Bacteria | *Aquifex aeolicus* | 1522 | | 15% |
| Bacteria | *Haemophilus influenzae* | 1708 | | 13% |
| Bacteria | *Bacillus subtilis* | 4093 | | 15% |
| Bacteria | *Escherichia coli* | 4281 | | 17% |
| Bacteria | *Vibrio cholerae* | 3815 | | 16% |
| Bacteria | *Mycoplasma pneumoniae* | 675 | | 14% |
| Bacteria | *Xylella fastidiosa* | 2761 | | 17% |
| Bacteria | *Thermotoga maritime* | 1842 | | 18% |
| Bacteria | *Neisseria meningitidis MC58* | 2015 | | 17% |
| Bacteria | *Chlamydia pneumoniae* | 1052 | | 18% |
| Bacteria | *Synechocystis sp* | 3167 | | 20% |
| Bacteria | *Chlamydia trachomatis* | 894 | | 19% |
| Bacteria | *Treponema pallidum* | 1028 | | 11% |
| Bacteria | *Pseudomonas aeruginosa* | 5562 | | 24% |
| Bacteria | *Mycobacterium tuberculosis* | 3916 | | 31% |
| Bacteria | *Deinococcus radiodurans* chr 1 | 2580 | | 33% |
| Eukaryota | *Plasmodium falciparum* chr II, III | 422 | | 35% |
| Eukaryota | *Caenorhabditis elegans* | 17049 | | 36% |
| Eukaryota | *Arabodiopsis thaliana* | 7849 | | 41% |
| Eukaryota | *Saccharomyces cerevisiae* | 6264 | | 40% |
| Eukaryota | *Drosophila melanogaster* | 13885 | | 51% |

[a]From (Dunker *et al.*, 2000)

[b]The percentages of proteins in the indicated genomes predicted to have at least one region of disorder of $\geq 40$ amino acids. Predictions were made by PONDR VL-XT (Li et al., 1999; Romero et al., 2001).

Table VI. Functional categories for disordered regions of proteins

| Category | Number | Transition | Description |
|---|---|---|---|
| Molecular Recognition | 113 | D → O | *Protein, ssDNA, dsDNA, tRNA, rRNA, mRNA, nRNA, bilayers, ligands, co-factors, metals, autoregulatory* |
| Molecular Assembly / Disassembly | >13 | D → O<br>O → D | *Hetero complexes, linear polymers, phages, viruses* |
| Protein Modification | 36 | Variable | *Acetylation, fatty acylation, glycosylaion, methylation, phosphorylation, ADP-ribosylation, ubiquitination, proteolytic digestion* |
| Entropic Chains | 17 | None | *Linkers, spacers, bristles, clocks, springs, detergents, self-transport* |

**Figure 1.** Comparisons of amino acid compositions of ordered protein and disordered protein. (Top) Amino acid compositions of three disordered datasets. (Middle) Amino acid compositions of three ordered datasets. (Bottom) Compositions of disordered datasets relative to the Globular 3-D dataset. The ordinates are (% amino acid in disordered dataset - % amino acid in Globular 3-D) / (% amino acid in Globular 3-D) = (D-O)/O. Negative values indicate that the disordered database has less than the ordered, positive indicates more than the ordered. Error bars are one standard deviation.

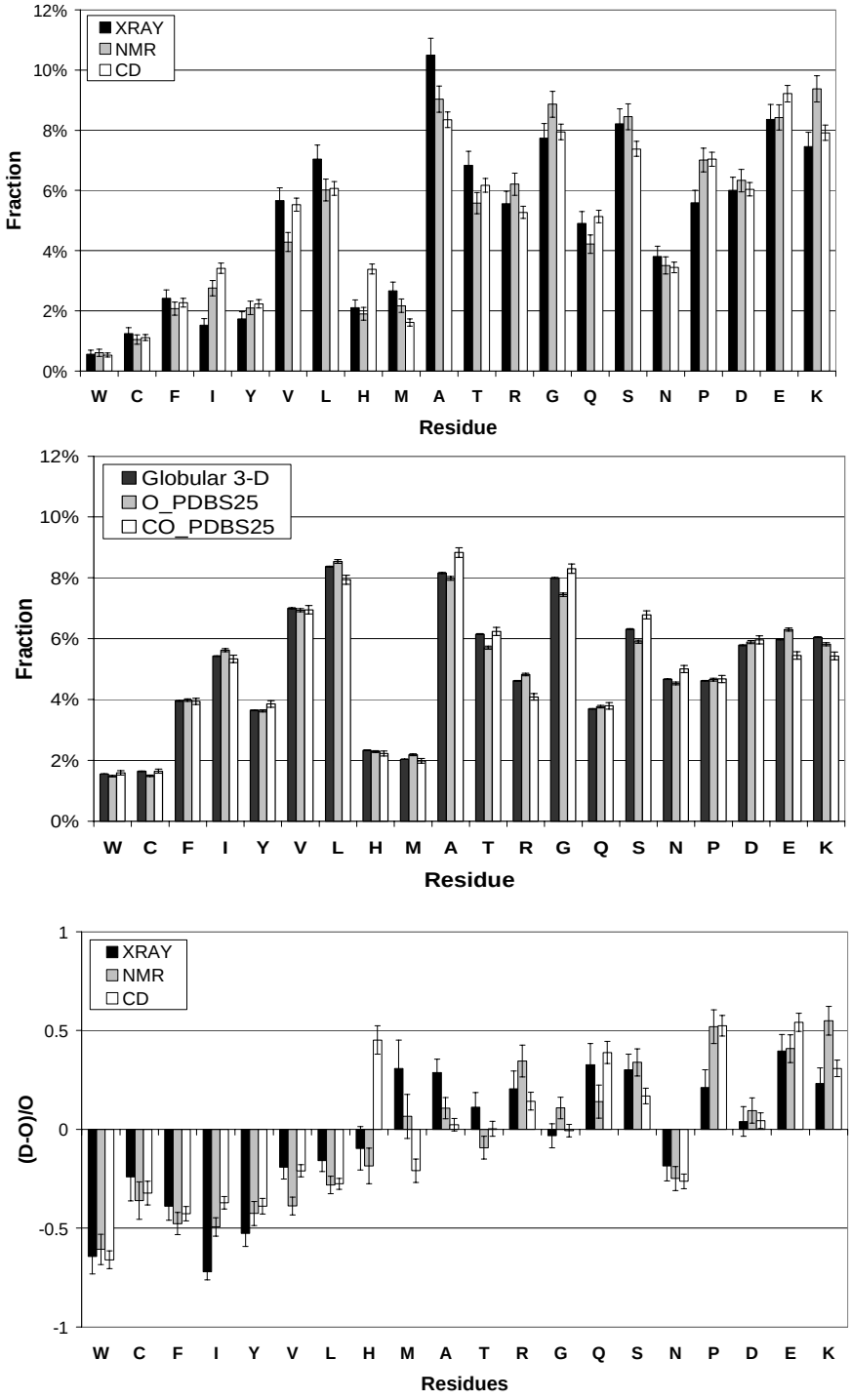

**Figure 2.** Conditional probability plot for Sweet and Eisenberg's (1983) hydropathy scale. The black line is the probability (y-axis) that a residue is ordered given the hydropathy score indicated on the x-axis. The dashed line is the probability of disorder. Negative values for hydropathy indicate hydrophilicity, positive values indicate hydrophobicity. The area between the two curves is divided by the total area of the graph to obtain the area ratio.

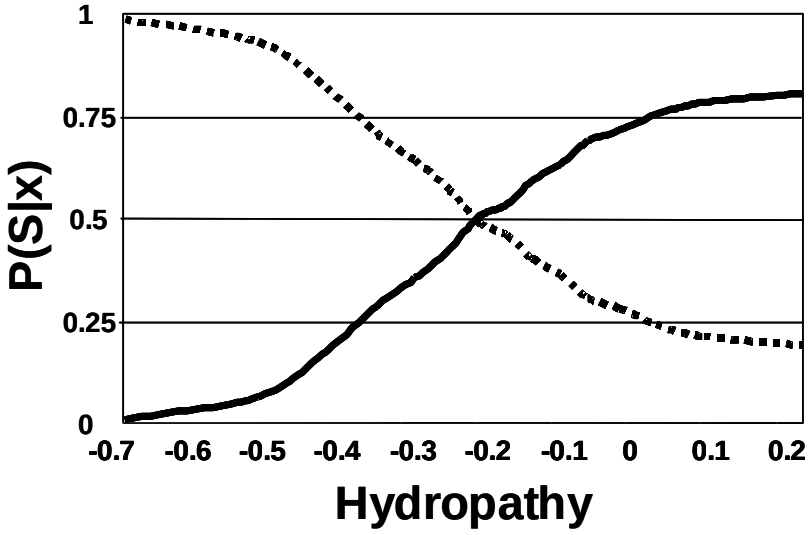

**Figure 3.** Procedures for identifying order and disorder in protein families. Black boxes indicate ordered sequences, gray boxes indicate disordered sequences, and open boxes indicate insertions relative to the starting sequence.

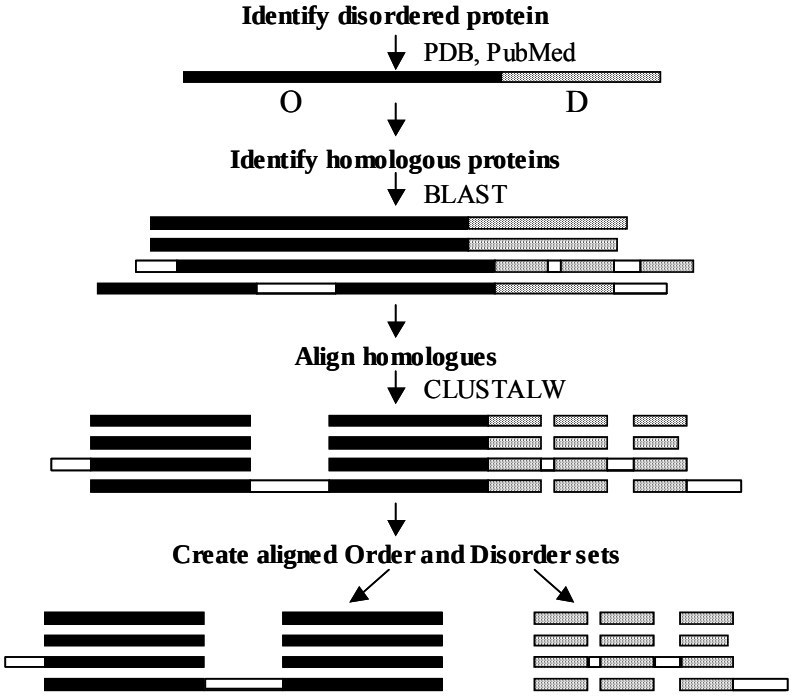

**Figure 4.** Substitution matrix based upon disordered protein families. Below the diagonal are the scores for each amino acid substitution. Above the diagonal are the differences between BLOSUM 62 and the disorder matrix. On the diagonal are the scores/differences.

| | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W | B | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | 10/-1 | -1 | -2 | -1 | 1 | 0 | -2 | 0 | 0 | 0 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -2 | 3 | 0 | 0 |
| S | 0 | 3/1 | 0 | -1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| T | 1 | 1 | 4/1 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -2 | 0 | -1 | 0 | 0 | 1 | 0 | 0 | -1 | 3 | 0 | 0 |
| P | -2 | 0 | -1 | 6/1 | 0 | -1 | -1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -2 | -1 | -1 | 0 | -3 | 1 | 0 |
| A | -1 | 1 | 0 | -1 | 3/1 | 0 | -1 | -1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | -1 | 0 |
| G | -3 | 0 | -2 | -1 | 0 | 5/1 | 0 | 0 | 0 | 0 | -1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 2 | 0 | 0 |
| N | -1 | 1 | 0 | -1 | -1 | 0 | 4/2 | 0 | 0 | -1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | 0 | 0 |
| D | -3 | 0 | -1 | -2 | -1 | -1 | 1 | 4/2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 2 | 0 |
| E | -4 | -1 | -1 | -1 | -1 | -2 | 0 | 2 | 4/1 | 2 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |
| Q | -3 | 0 | 0 | -1 | -1 | -2 | 1 | 0 | 0 | 5/0 | -1 | 0 | 1 | 1 | -1 | 0 | 0 | -1 | -1 | -1 | 0 | 2 |
| H | -1 | -1 | 0 | -2 | -2 | -1 | 2 | -1 | -1 | 1 | 8/0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | 0 | 0 | 0 | 1 |
| R | -1 | -1 | -1 | -2 | -2 | -2 | 0 | -2 | -1 | 1 | 0 | 5/0 | 0 | 0 | -1 | 0 | -1 | 0 | 0 | -3 | 0 | 1 |
| K | -3 | -1 | 0 | -1 | -1 | -2 | 0 | -1 | 0 | 0 | -1 | 2 | 4/1 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| M | 0 | -2 | -1 | -2 | -1 | -4 | -2 | -4 | -3 | -1 | -2 | -1 | -2 | 7/-2 | 0 | 0 | 0 | -1 | 0 | 0 | 1 | 1 |
| I | 0 | -2 | -1 | -2 | -1 | -5 | -3 | -4 | -3 | -2 | -2 | -2 | -2 | 1 | 4/0 | 0 | 0 | -1 | -1 | -1 | 1 | 0 |
| L | -1 | -2 | -2 | -1 | -1 | -4 | -3 | -4 | -3 | -2 | -2 | -2 | -2 | 2 | 2 | 4/0 | 0 | -1 | -1 | 0 | 0 | 0 |
| V | 1 | -2 | 0 | -1 | 0 | -4 | -3 | -4 | -2 | -2 | -2 | -2 | -2 | 1 | 3 | 1 | 4/0 | -1 | 0 | 1 | 1 | 0 |
| F | -1 | -2 | -2 | -3 | -2 | -4 | -2 | -4 | -4 | -2 | 0 | -3 | -3 | 1 | 1 | 1 | 0 | 7/-1 | -1 | 2 | 1 | 1 |
| Y | 0 | -2 | -1 | -3 | -2 | -3 | -1 | -4 | -3 | 0 | 2 | -2 | -2 | -1 | 0 | 0 | -1 | 4 | 8/-1 | -1 | 1 | 1 |
| W | -5 | -3 | -5 | -1 | -5 | -4 | -3 | -4 | -4 | -1 | -2 | 0 | -3 | -1 | -2 | -2 | -4 | -1 | 3 | 13/-2 | 0 | 1 |
| B | -3 | 0 | -1 | -2 | -1 | -1 | 1 | 4 | 2 | 0 | -1 | -2 | -1 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | 4/2 | 0 |
| Z | -4 | -1 | -1 | -1 | -1 | -2 | 0 | 2 | 4 | 0 | -1 | -1 | 0 | -3 | -3 | -3 | -2 | -4 | -3 | -4 | 2 | 4/1 |

**Figure 5.** Comparison of proteolysis data for *Xeroderma pigmentosum* group A (XPA) with PONDR predictions and NMR structure. (Top) Full-length *Xenopus laevis* XPA is depicted as a bar, with all possible trypsin sites indicated by white vertical lines (*X. laevis* numbering). The line below represents the human Minimal Binding Domain of XPA in the same format, the structure of which has been determined by NMR. Four regions with low certainty of assignment or high flexibility are indicated in gray. Each of the unique experimentally observed trypsin proteolysis fragments are drawn as horizontal lines below; the end points of these lines indicate the trypsin sensitive sites. (Below) Disorder prediction for *X. laevis* XPA. Each residue (x-axis) is assigned a disorder score (y-axis) by PONDR VL-XT. Disorder scores ≥ 0.5 signify disorder. Note the coincidence between predictions of disorder and the observed cut sites, and note also the coincidence between predictions of order and lack of observed cut sites.