Comparing and Combining Predictors of Mostly Disordered Proteins†

Christopher J. Oldfield,[‡,§] Yugong Cheng,[‡] Marc S. Cortese,[‡,#] Celeste J. Brown,[¶] Vladimir N. Uversky,*[,‡,#,||] and A. Keith Dunker*[,‡,#]

Molecular Kinetics, Inc., 6201 La Pas Trail, Suite 160, Indianapolis, Indiana 46268, Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, School of Medicine, Indiana University - Purdue University at Indianapolis, Indianapolis, Indiana 46202, Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844, and Institute for Biological Instrumentation, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia

Received September 17, 2004; Revised Manuscript Received November 18, 2004

ABSTRACT: Intrinsically disordered proteins and regions carry out varied and vital cellular functions. Proteins with disordered regions are especially common in eukaryotic cells, with a subset of these proteins being mostly disordered, e.g., with more disordered than ordered residues. Two distinct methods have been previously described for using amino acid sequences to predict which proteins are likely to be mostly disordered. These methods are based on the net charge−hydropathy distribution and disorder prediction score distribution. Each of these methods is reexamined, and the prediction results are compared herein. A new prediction method based on consensus is described. Application of the consensus method to whole genomes reveals that approximately 4.5% of *Yersinia pestis*, 5% of *Escherichia coli* K12, 6% of *Archaeoglobus fulgidus*, 8% of *Methanobacterium thermoautotrophicum*, 23% of *Arabidopsis thaliana*, and 28% of *Mus musculus* proteins are mostly disordered. The unexpectedly high frequency of mostly disordered proteins in eukaryotes has important implications both for large-scale, high-throughput projects and also for focused experiments aimed at determination of protein structure and function.

Intrinsically disordered proteins do not adopt a stable three-dimensional structure, instead they exist as an ensemble of interchanging conformations in solution. Many examples of these proteins have been described in the literature [see refs *1*−*3* for reviews], that range from wholly disordered to partially structured-partially disordered. In contrast to structured (i.e., ordered) proteins, intrinsically disordered proteins or regions perform their various functions without the prerequisite of a folded conformation (*1*−*3*). Often, disordered proteins are involved in protein−protein (*4*−*7*) or protein−nucleic acid (*8*) interactions, which typically involve coupled binding and folding (*9*). Disorder in these proteins may play a general role by reducing affinity without a corresponding loss of specificity, which may be a particularly important feature for interaction-mediated signal transduction (*10*, *11*). In agreement with this idea, signal transduction proteins are indicated to be highly enriched in disorder (*12*). Also, their conformational heterogeneity may allow disordered proteins to bind to a wide variety of protein or nucleic acid partners. Notably, this low specificity model can explain the function of some chaperones (*13*, *14*). Conversely, the functions of some disordered proteins, such as entropic bristles (*15*), entropic clocks (*16*), or flexible linkers (*17*), rely on the maintenance of a high degree of conformational freedom. Despite a broad range of sequence and functional diversity, disordered proteins have been shown to share common sequence features.

The encoding of intrinsic disorder by amino acid sequences was first investigated by Romero et al. (*18*), who showed that disordered and ordered regions of proteins could be reliably distinguished based solely on local sequence. Since then, several groups have developed sequence-based predictors of intrinsic disorder (*18*−*23*, *53*, *54*). Two of these groups have applied these predictors to make conservative estimates of the proportion of disordered proteins in various genomes (*23*, *24*). Though useful for establishing the relevance of disorder to the study of naturally occurring proteins, all estimates based on per-residue predictions share a few shortcomings. First, they use somewhat arbitrary thresholds for intrinsically disordered proteins. That is, the choice of significance thresholds are highly influenced by the rate of false disorder prediction, rather than the content of disordered residues that is biologically relevant. Second, they indicate an approximate lower bound to the proportion of intrinsically disordered proteins in a given genome by minimization of the rate of false predictions of disordered residues. As a consequence, conservative estimates based on per-residue prediction have a high rate of false prediction of ordered proteins. Third, they summarize per-residue predictions based only on the predicted disordered regions, which overlooks the character of other, possibly ordered, regions of a protein.

† This work was supported in part by NIH Grant R01 LM07688 (A.K.D.), NIH Grant R43 GM06412 (Y.C.), and NLM Grant 5T15LM007359-02 (C.J.O.).

* To whom correspondence should be addressed at Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, School of Medicine, Indiana University − Purdue University at Indianapolis, 714 N. Senate St., Suite 250, Indianapolis, IN 46202. Phone: 317-278-9650; fax: 317-278-9217; e-mail: kedunker@iupui.edu (A.K.D.); vuversky@iupui.edu (V.N.U.).

‡ Molecular Kinetics, Inc.

Indiana University - Purdue University at Indianapolis.

¶ University of Idaho.

|| Russian Academy of Sciences.

§ Present Address: Biophysics Department, Institute for Molecular Virology, 1525 Linden Drive, Madison, Wisconsin 53706.

10.1021/bi047993o CCC: $30.25
Published on Web 01/22/2005

An alternative approach for estimating the content of intrinsically disordered proteins in genomes is binary classification of whole proteins as either mostly disordered or mostly ordered, where mostly ordered indicates proteins that contain more ordered residues than disordered residues and mostly disordered indicates proteins that contain more disordered residues than ordered residues. Of course, the extent to which a sequence is ordered or disordered and the nature of disorder vary widely among proteins (*3*, *11*). However, such a classification scheme would avoid the shortcomings discussed above by implicitly considering protein context and prediction threshold and allowing false prediction rates to be balanced as desired. Therefore, although the binary classification of proteins as ordered or disordered is a gross approximation of the actual biological situation, its potential usefulness in terms of simplified interpretation and broad applicability suggests that such tools would be useful to the biological community.

Two distinct classification methods that use this simplification have been reported previously (*20*, *25*−*27*). One is based on the predictor of natural disordered regions (PONDR) VL-XT[1] (*18*, *19*, *28*), which predicts the order−disorder class for every residue in a protein. Cumulative distribution function (CDF) analysis (*25*) summarizes these per-residue predictions by plotting PONDR scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished based on the distribution of prediction scores. The other established method of order−disorder classification is charge−hydropathy plots (*20*). Ordered and disordered proteins plotted in charge−hydropathy space can be separated to a significant degree by a linear boundary.

Here we reexamine and compare the CDF and charge−hydropathy methods. These methods are also compared to the proportion of residues predicted to be disordered, which was developed for whole protein classification elsewhere (*26*, *27*). Keeping with the binary order−disorder generalization, proteins used to derive the prediction algorithms are extreme examples, where each is, in fact, wholly ordered or wholly disordered. The assumption is that classification of novel proteins, which will commonly be mixtures of order and disorder, can then be made based on whether the protein more closely resembles wholly ordered proteins or wholly disordered proteins. This generalization is partially validated by the use of a third distinct data set that is neither wholly ordered nor wholly disordered. Estimations of the intrinsic disorder content of various protein sequence databases were made and used to illustrate the meaning and significance of predictions of intrinsic disorder to proteome-scale projects.

MATERIALS AND METHODS

Data Sets. Two sets of proteins were used for binary classifier construction. The set of wholly disordered proteins contains only proteins that were shown to be disordered in solution. The set of wholly ordered proteins contains only monomeric proteins whose structures have been well defined by X-ray crystallography. A third set of proteins was used for predictor validation. Denoted partially ordered proteins, this set includes proteins that contain both structured and unstructured regions.

A set of wholly disordered proteins was derived from two previously published data sets (*2*, *20*). These data sets were merged, and all proteins with >25% identity were clustered using the program *blastclust* from NCBI.[2] This program uses single linkage clustering, where the linkages are determined by an all-against-all standard protein BLAST (*29*) search of the data set. The protein in each cluster with the greatest amount of characterized disorder was chosen as the representative of that cluster, which gave a set of 159 sequence-unique proteins. Of these, proteins annotated as disordered over their entire length were selected, which gave a set of 54 proteins containing 10 782 residues.

The set of wholly ordered proteins was derived from the Protein Data Bank (PDB) (*30*) as of November 10, 2002. This set included X-ray crystallography structures that contained a single chain and a unit cell with a primitive space group, which is a subset of the monomers in the database. Monomers were desired for this analysis because monomers preclude the possibility that some chains might in fact be intrinsically disordered but become ordered during complex formation. From this set of monomers, all structures that contained ligands and disulfide bonds were eliminated. Then, to ensure that the chains were wholly ordered, all proteins with missing density were removed. The remaining chains were clustered by sequence identity as described above, and the longest protein of each cluster was chosen as the representative of that cluster. This gave a set of 105 unique, wholly ordered monomers without ligands or disulfide bonds. These 105 proteins contained 22 829 residues.

The set of partially ordered proteins was derived from PDB structures that contained a single chain and a unit cell with a primitive space group. All proteins from this set with a region of missing density of 30 or longer were clustered. The structure with the highest proportion of disorder was chosen as the representative from each cluster, resulting in a set of 64 unique proteins. This set contained 23 785 residues, of which 4074 are disordered. The percentage of disordered residues in these proteins ranged from 4 to 50%.

Predictors were applied to the Swiss-Prot database (*31*) release as of May 2003 with 124 425 sequences, with 9379 human sequences. Genomic annotation data sets were from the National Center for Biotechnology Information (NCBI)[2] and the Ensembl database (*32*).

Cumulative Distribution Function for PONDR Predictions. The CDF, as applied to PONDR VL-XT predictions, is a cumulative histogram of the PONDR scores for a given protein (*25*). At any given point on the CDF curve, the ordinate gives the proportion of residues with a PONDR score less than or equal to the abscissa. CDF curves for PONDR VL-XT predictions always begin at the point (0, 0) and end at the point (1, 1) because PONDR VL-XT predictions are defined only in the range [0, 1] with values less than 0.5 indicating a propensity for order and values greater than or equal to 0.5 indicating a propensity for disorder. Proteins with high PONDR scores will have CDF curves that have low cumulative values over most of the

[1] Abbreviations: CDF, cumulative distribution function; PONDR, predictor of natural disordered regions (PONDR is a registered trademark of Molecular Kinetics, Inc.); VL-XT, variously characterized long disordered regions-X-ray characterized terminal disordered; ROC, receiver operator curves.

[2] http://www.ncbi.nlm.nih.gov.

CDF curve, and proteins with low PONDR scores will have CDF curves that have high cumulative values over most of the CDF curve. Therefore, a boundary may be determined that separates the CDF curves of disordered proteins from the CDF curves of ordered proteins.

For derivation of the order−disorder boundary and subsequent analysis, calculated PONDR scores were divided in 20 evenly spaced bins as an approximation of the CDF curves. The binned CDF curves of the ordered and disordered protein sets were then treated as separate populations with normally distributed cumulative PONDR scores. The optimal boundary point for each bin was determined using the univariate normal probability density function (*33*). This equation is solved to give the point between the means where the probability density was equal. This point is given by

$$x_n = \{\alpha_n - [\alpha_n^2 - (\sigma_{o,n}^{-2} - \sigma_{d,n}^{-2})(\mu_{o,n}^2 \cdot \sigma_{o,n}^{-2} + \mu_{d,n}^2 \cdot \sigma_{d,n}^2 + 2 \ln [\sigma_{o,n}^{1/2} \cdot \sigma_{d,n}^{-1/2}])]^{-1/2}\}/\{\sigma_{o,n}^{-2} - \sigma_{d,n}^{-2}\}$$

$$\alpha_n = \mu_{o,n} \cdot \sigma_{o,n}^{-2} - \mu_{d,n} \cdot \sigma_{d,n}^{-2}$$

where the means of the PONDR scores of ordered and disorder proteins are taken as estimates of $\mu_{o,n}$ and $\mu_{d,n}$, respectively, and the standard deviations of the PONDR scores of ordered and disordered proteins are used as estimates of $\sigma_{o,n}$ and $\sigma_{d,n}$. The subscript n denotes the index of each bin of PONDR scores and takes on discrete values from 1 to 20.

The combination of boundary points that provided the most accurate classification was determined by examining all possible, multipoint boundaries made up of consecutive boundary points. Jack-knifing was used to assess the accuracy of each boundary. In this assessment, each example protein was left out of the set once, the boundary points were calculated from the remaining examples, and the ability of all combinations of boundary points to correctly classify the left out example was determined. The accuracy of a combination of boundary points was then estimated by the average performance over all the left out examples. Jack-knifing gives a relatively unbiased estimation of classifier accuracy (*33*).

Charge−Hydropathy Plots. The mean net charge and the mean normalized Kyte-Doolittle hydropathy (*34*) were calculated for each protein in all three sets. The boundary between the ordered and disordered proteins on a charge−hydropathy plot was determined using a linear discriminant function, assuming normal distributions and equal covariance matrices. The equation for the discriminant line (*33*) is

$$f_x(x; \mu_d, \mu_o, \sum) = [\sum^{-1}(\mu_d - \mu_o)]' \, x - 1/2 \, (\mu_d - \mu_o)' \sum^{-1}(\mu_d + \mu_o)$$

where μ_d and μ_o are the two-dimensional charge−hydropathy vectors of ordered and disordered proteins, respectively, and Σ is the pooled charge−hydropathy covariance matrix. Positive values from this equation indicate that a protein is disordered and negative values indicate order. The solution of this equation for the net charge component of x as a function of the hydropathy component of x yields an equation for the order−disorder boundary. Jack-knifing was used to assess the accuracy of the order−disorder boundary.

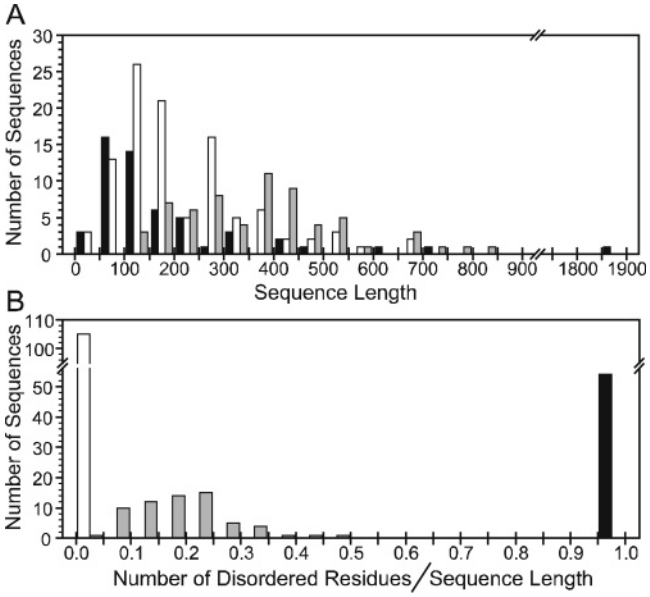

FIGURE 1: Histograms characterizing the length and the disorder content of the three data sets. The wholly disordered protein set (black bars), the wholly ordered protein set (white bars), and the partially ordered proteins (grey bars) are shown. (A) Histogram of the lengths of proteins in each set in bins of 50 residues. Tick labels indicate the boundary between every other bin. (B) Histogram of proportion of disorder in each protein of the three data sets. Proportions of disorder are split into bins of 5%, where the extreme bin boundaries, 0 and 100% are inclusive.

Note that the assumptions of normally distributed data used in this analysis were not necessarily justified. However, trial studies with methods that do not require normally distributed data (e.g., neural networks) and methods that do not require equal covariance matrices (e.g., quadratic discrimination) yielded results nearly equivalent to those obtained using linear discrimination (data not shown).

Proportion of Predicted Disorder Threshold. A previous analysis used the proportion of PONDR VL-XT predictions for the classification of ordered and disordered proteins (*27*). Proteins with >35% of residues predicted to be disordered were classified as disordered, and otherwise were classified as ordered. These classifications are compared to CDF and charge−hydropathy plot analyses.

Relative Composition Plots. From a set of human proteins from the SwissProt database, the compositions for proteins predicted to be disordered by CDF by the entire boundary or charge−hydropathy by a boundary margin of 0.045 were calculated relative to the compositions of proteins that were predicted to be ordered by both algorithms. The relative composition were calculated as describe by Romero et al. (*28*), with the modification that 10 000 bootstrap resampling iterations were used to estimate the relative compositions and 99% confidence intervals.

Entropy Distributions. Shannon's entropy (*35*) was calculated for all windows of 45 residues for the three groups of proteins discussed above. The distribution of minimum entropy windows was calculated by kernel density estimation using a biweight kernel and a sample-based bandwidth.

RESULTS

Data Set Characterization. Two histograms were calculated to give an overview of the three data sets (Figure 1).

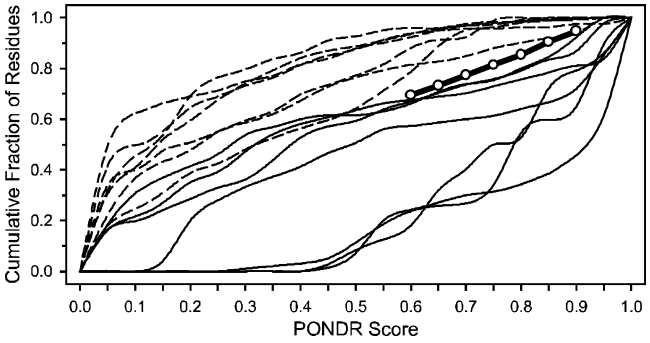

FIGURE 2: Cumulative Distribution Function curves. Seven wholly ordered proteins (- - -), seven wholly disordered proteins (−), and the order−disorder boundary (-○-) are plotted.

The first histogram (Figure 1A) illustrates the distribution of the length of the proteins in the wholly ordered set, wholly disordered set, and the partially ordered set. Wholly ordered and wholly disordered proteins had mean lengths of 217 residues and 200 residues, respectively. The partially ordered proteins were significantly longer on average than the other two sets, with a mean length of 372 residues. Considering that the partially ordered proteins were only selected if they contained a region of missing density of 30 residues or longer and wholly ordered proteins were only selected if they contained no disordered residues, the longer mean length of partially ordered proteins suggests that a large proportion of long monomers contain a disordered region of significant size.

The second histogram (Figure 1B) shows the percent of disordered residues for each of the three sets. The wholly ordered proteins and wholly disordered proteins were selected for no disorder and total disorder, respectively, and necessarily fall completely in the two extreme bins. On the basis of the current partially ordered proteins set, 50% may approximate the upper limit for the amount of intrinsic disorder in a crystallizable protein, although most of the proteins with disorder have less than half this amount.

PONDR Cumulative Distribution Function Analysis. The optimum boundary was calculated as described in Materials and Methods. Examination of all possible combinations of consecutive boundary points showed that seven boundary points located in the 12th through 18th bin provided the optimal separation of the ordered and disordered protein sets. For clarity, the boundary is shown in Figure 2 along with PONDR CDF curves for representative examples from the ordered and the disordered sets. The estimated classification accuracy of this boundary was 88% overall with 87 and 90% of wholly disordered and wholly ordered proteins correctly classified, respectively, where classification was based on whether a CDF curve was above or below a majority of boundary points. As expected, disordered proteins generally fall below this boundary and ordered proteins generally fall above this boundary.

The set of partially ordered proteins were mostly structured but contained short regions of disorder. Because these proteins are crystallizable and ≥50% of their residues are resolved, these proteins are mostly ordered and the order/disorder classification scheme should identify the proteins in this set as ordered proteins. The CDF boundary classified 70% of these proteins as ordered based on having a majority of values above the boundary. This is much lower than the estimated accuracies for the wholly ordered and wholly disordered proteins. It is expected that the higher the disordered residue content of a given protein, the more likely it will be classified as disordered. As expected, the mean fraction of experimentally disordered residues in proteins classified as disordered (0.23) is higher than the fraction of experimentally disordered residues in proteins classified as ordered (0.18). This suggests the classification of some of these proteins as disordered is due to their relatively large disordered residue content.

Table 1: Accuracy of CDF Boundary Classification versus the Number of Boundary Points in Majority[a]

no. of majority points	percent of proteins (%)	disorder accuracy (%)	order accuracy (%)	partially ordered accuracy (%)
4	100	87	90	70
5	92	88	91	80
6	85	88	93	78
7	81	90	94	78

[a] The percent of proteins column gives the proportion of ordered and disordered proteins that have CDF curves that are above or below the boundary by the given number of points. Disorder accuracy, order accuracy, and partially ordered accuracy columns give the percentage of each set predicted correctly.

The CDF curves of many of the example proteins do not fall completely on one side of the boundary or the other. This suggests that the number of boundary points in agreement could be used as an indication of prediction confidence. Calculation of prediction accuracy as a function of the minimum number of boundary points in agreement (Table 1) demonstrates that this is a valid approach. All curves have a minimal boundary majority of four points, which has an average classification accuracy of 88%. For CDF curves that fall completely above or below the boundary (i.e., all seven points), the average accuracy increases to 92%. However the highest accuracy of classification (80%) for the partially ordered protein set is achieved by setting the boundary point criteria to five points in the majority, but then decreases as points are added or removed. This analysis shows that the number of boundary points in the majority is correlated with prediction accuracy, but also that increasing the number of points leads to a reduced proportion of correctly classified proteins. Alternatively, the accuracy can be examined over a range of thresholds, rather than excluding predictions, using receiver operator characteristics (ROC) curves. This assessment, which agrees qualitatively with Table 1, is given in Figure S1 (see Supporting Information).

Charge−Hydropathy Plots. The optimal boundary between the ordered and disordered proteins in charge−hydropathy space was determined (Figure 3) to be

$$\langle \text{charge} \rangle = 2.743 \langle \text{hydropathy} \rangle - 1.109$$

by the procedure described. The classification accuracy based on a jack-knife estimate was 83% overall, 76% for disordered proteins, and 91% for ordered proteins. The boundary was calculated from the complete sets of the wholly disordered and wholly ordered proteins. This boundary is similar to the equation found previously for other data sets (*20*), with the slope and intercept of this line being just 2% less and 4% less, respectively, than those of the previous line.

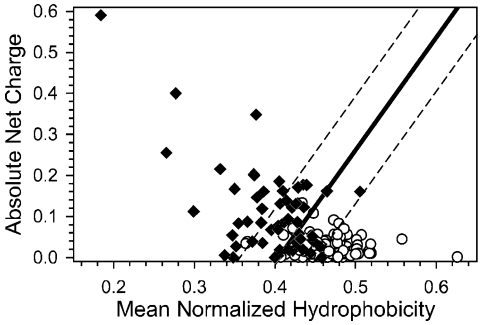

FIGURE 3: Charge−hydropathy plot. The wholly ordered set of proteins (○), wholly disorder set of proteins (◆), the order−disorder boundary (−), and the extents of the 0.045 boundary (- - -) are shown.

Table 2: Accuracy of Charge−Hydropathy Plot Boundary Classification versus the Width of the Boundary Margin[a]

boundary width	percent of proteins (%)	disorder accuracy (%)	order accuracy (%)	partially order accuracy (%)
0.000	100	76	92	94
0.015	88	84	97	98
0.030	70	88	97	98
0.045	50	95	97	97

[a] The boundary margin is the boundary line ± the boundary width as illustrated in Figure 3. The percent of proteins column gives the proportion of proteins that charge and hydropathy place outside the boundary margin. The accuracy estimates are calculated for the proteins that fall outside the margins.

As for the CDF analysis, an indication of prediction accuracy can be derived from the charge−hydropathy plots. The width of the boundary line can be expanded to both sides by applying a fixed Euclidean distance normal to the line, where the distance is obtained by normalizing the boundary equation by the root-sum-square of the ⟨charge⟩ and ⟨hydropathy⟩ coefficients. The margins of this new boundary can then be used to classify the proteins that fall outside of it (Table 2). These results indicate that the more extreme values of hydropathy and net charge that a protein has, the more confidently it can be classified. At a boundary margin of ±0.045, the accuracy reached 95% for disordered proteins and 97% for ordered proteins. However, only 50% of the proteins fall outside these boundary margins. We have also generated ROC curves for charge−hydropathy predictions (Figure S1, Supporting Information) for an alternative view of prediction accuracy.

The accuracy of classification for disordered proteins rose substantially for increasing margin widths, but the accuracy for wholly ordered proteins and partially ordered proteins increased relatively little, and actually decreased slightly for the partially ordered proteins at the largest margin width. However, the overall accuracy for the wholly ordered proteins and partially ordered proteins was much greater than that for the CDF analysis.

As argued in the previous section, the set of partially ordered proteins is expected to be classified as ordered proteins by these predictors. The accuracy of charge−hydropathy plots in the classification of partially ordered proteins as ordered proteins was 94% and increased slightly to 97% at a boundary margin of 0.045, which is a large improvement over the CDF analysis. As for the CDF analysis, the partially ordered proteins that were predicted to be disordered had a higher mean fraction of experimentally disordered residues (0.26) than the proteins that were predicted to be ordered (0.19).

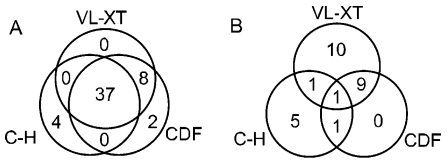

FIGURE 4: Venn diagrams comparing true positives and false positives for three prediction methods. PONDR CDF analysis (CDF), charge−hydropathy plots (C−H) and PONDR VL-XT (VL-XT) classifications are shown. (A) Comparison of the 51 true positive predictions. (B) Comparison of the 27 false positive predictions.

With this analysis, it was apparent that the two approaches examined had relative strengths. CDF analysis was better able to identify all disordered proteins, and charge−hydropathy plots were better at distinguishing partially ordered proteins from wholly disordered proteins. Both analyses had similar accuracy for ordered proteins. These results prompted us to examine predictors based upon a combination of both methods.

Cross Method Accuracy Analysis. PONDR CDF analysis, charge−hydropathy plots, and PONDR VL-XT proportion threshold prediction can all be used to classify proteins as ordered or disordered. A comparison of these three approaches was carried out to investigate whether combining the methods improved the accuracy of two state predictions. For comparison purposes, PONDR VL-XT proportion threshold predictions were made on these data sets using previously derived parameters, which gave accuracies of 83% for wholly disordered proteins and 80% for wholly ordered proteins.

Of the 54 wholly disordered proteins, 51 were predicted to be disordered by at least one of the three methods. The relationships of correct predictions for each method were visualized with a Venn diagram (Figure 4A). The three methods agree well; most of the disordered proteins were predicted correctly by all methods. In addition, charge−hydropathy analysis and CDF contribute unique predictions of disorder. The relatively large overlap of PONDR VL-XT and CDF analysis was expected because they share the same base prediction method.

Of the 105 wholly ordered proteins, 27 were predicted to be disordered by at least one of these methods (Figure 4B). Classification based upon the proportion of predicted disordered residues using 35% as the boundary leads to the highest number of errors (*21*), while the CDF analysis and charge−hydropathy plots make about half as many misclassifications. As observed above for true positives, VL-XT prediction and CDF analysis exhibit the largest number of common mistakes because both methods are based on the same underlying calculations. VL-XT prediction, charge−hydropathy plots, and CDF analysis make 10, 5, and 0 unique mistakes, respectively.

Examination of the true positive and false positive predictions suggests that classification of proteins as wholly disordered or wholly ordered should be made using only the charge−hydropathy plots and the CDF analysis. When considering intersection prediction (i.e., the center regions of Figure 4A,B), there was little improvement in prediction accuracy if VL-XT was used in conjunction with the other

Table 3: Comparison of the Accuracies of Single and Combination Prediction Methods[a]

	single predictions		combination predictions		
data set	CDF analysis (%)	charge−hydropathy plots (%)	unanimous predictions (%)	voting predictions (%)	consensus score (%)
ordered	90	91	98	84	95 (95%)
disordered	87	76	69	94	90 (83%)
partially ordered	70	94	95	69	85 (88%)

[a] The first value in the consensus score column does not include proteins that have ambiguous predictions of disorder, while the numbers in parentheses include ambiguous predictions as order. Ambiguous predictions were obtained for 6% of the order set, 7% of the disorder set, and 14% of the partially ordered set.

two predictors. When considering the union of prediction methods, PONDR VL-XT degraded the classification accuracy by contributing false positives without contributing any true positives. On the basis of these results, charge−hydropathy plots and the CDF analysis were combined into a single classification method. Two straightforward combinations of these predictions were examined first.

The unanimous prediction combination method predicts disorder when both CDF analysis and charge−hydropathy plots predict disorder or otherwise predicts order. The voting prediction combination method predicts disorder when either CDF analysis or charge−hydropathy plots predict disorder and predicts order when neither predicts disorder. Both of these simple predictor combinations perform very well for two of the data sets, but both are relatively poor on one data set (Table 3). Unanimous prediction performs well for ordered proteins and partially ordered proteins but is poor on disordered proteins. Voting prediction performs well for disordered proteins and performs adequately for ordered proteins but is poor on partially ordered proteins. Neither of these straightforward combinations captures the relative strengths of both CDF analysis and charge−hydropathy plots. To improve performance of the combination prediction, a weighted scoring scheme, consensus scoring, was defined.

Consensus Scoring. Consensus scoring focuses on correct classification of proteins for which prediction methods disagree by using a weighted combination of the reliability measures. Here predictions are divided into two types: extreme and mild. CDF curves give extreme predictions when they fall completely above or completely below the boundary but otherwise give mild predictions. Charge−hydropathy plots give extreme predictions when the absolute distance to the boundary is greater than 0.045 but otherwise give mild predictions. These discrete prediction levels reflect the increase in classification confidence as a function of prediction strength.

The consensus score method assigns proteins one of five classifications with numerical equivalents, referred to as the consensus score. The classifications of "confidently disordered" (value of 100) and "confidently ordered" (value of 0) are assigned when both CDF analysis and charge−hydropathy plots are in agreement as to the class of the protein. The classifications of "likely disordered" (value of 75) and "likely ordered" (value of 25) are assigned when one predictor gives an extreme prediction and the other disagrees with a mild prediction. When both prediction methods give mild predictions for opposite classifications or both prediction methods give extreme predictions for opposite classifications, the proteins is classified as "ambiguous" (value of 50). Binary predictions of order−disorder can be obtained from the consensus score in several ways. The accuracy of two classification systems are given in Table 3. One assigns order for values < 50, disorder for values > 50, and does not predicted for values equal to 50. The other (in parentheses) assigns order for values ≤ 50 and disorder for values > 50. As for other predictors, the prediction threshold can be selected as desired (Figure S1, Supporting Information).

Although the consensus method is not the most accurate method on any single data set, the overall performance of the method compares favorably to the overall performance of other prediction methods. For example, the highest average accuracies for the order and disorder sets, 89%, can be achieved by consensus scoring and the voting combination methods. However, the accuracy of the consensus score method for the ordered with disordered regions set is much higher than that of the voting combination method. Thus, the overall accuracy of the consensus score is highest of any of the prediction methods examined.

Predictions of Disorder over Sequence Databases. Large differences in native environment, extent of intercellular coordination, and evolutionary background make comparisons of disorder predictions across genomes inherently interesting. Also, structural and functional genomics projects span a wide range of diverse organisms, so genome-scale prediction is a practical measure to indicate for which organisms disorder predictions will be most relevant. For an initial study, CDF analysis, charge−hydropathy discrimination, and consensus prediction were applied to six genomes, two from each kingdom (Table 4).

The prediction methods vary greatly in their estimates of the number of fully disordered proteins in complete genomes.

Table 4: Summary of Disorder Predictions on Model Genomes from Archaea, Bacteria, and Eukaryotes[a]

kingdom	species	no. of sequences	mean length	CDF (%)	charge−hydropathy (%)	consensus score (%)
Bacteria	*Yersinia pestis* CO92	4078	316	6.0	4.8	4.5
Bacteria	*Escherichia coli* K12	4269	319	5.7	3.3	4.6
Archaea	*Archaeoglobus fulgidus*	2415	277	7.9	4.2	6.3
Archaea	*Methanobacterium thermoautotrophicum*	1873	281	10.8	5.0	8.0
Eukaryota	*Arabidopsis thaliana*	28564	424	25.3	13.1	23.6
Eukaryota	*Mus musculus*	25368	457	31.3	15.5	28.2

[a] The proportions given are the number of proteins predicted to be disordered by each method, relative to the number of sequences predicted. CDF predictions were made conservatively, with only proteins with curves falling completely below the discrimination boundary taken as predictions of disorder. Proteins with consensus score values > 50 were classified as disordered.

Table 5: Top 20 Predictions of Disordered Human Proteins in Swiss-Prot

SwissProt ID	name	length	net charge	hydropathy	distance to CH boundary
hsp1_human	sperm protamine P1	50	0.48	0.24	0.32
thya_human	prothymosin alpha	110	0.40	0.28	0.26
stp1_human	spermatid nuclear transition protein 1	54	0.35	0.27	0.24
thyp_human	parathymosin	101	0.30	0.28	0.22
hsp2_human	sperm protamine P2	102	0.25	0.27	0.22
rl39_human	60S ribosomal protein L39	50	0.36	0.33	0.19
sef2_human	gastric cancer-related protein VRG107	59	0.17	0.27	0.18
trhy_human	trichohyalin	1898	0.03	0.23	0.18
hp20_human	Protein HSPC020	121	0.33	0.34	0.18
nsb1_human	nucleosomal binding protein 1	282	0.15	0.28	0.17
sfr4_human	splicing factor, arginine/serine-rich 4	494	0.19	0.30	0.17
r39l_human	60S ribosomal protein L39-like	50	0.32	0.35	0.16
srch_human	sarcoplasmic reticulum histidine-rich calcium-binding protein	699	0.22	0.31	0.16
rs30_human	40S ribosomal protein S30	59	0.32	0.35	0.16
h173_human	nonhistone chromosomal protein HMG-17-like 3	89	0.19	0.30	0.16
dss1_human	split hand/foot deleted protein 1	70	0.31	0.35	0.16
sfr2_human	splicing factor, arginine/serine-rich 2	221	0.20	0.32	0.15
prpe_human	basic proline-rich peptide P−E	61	0.10	0.28	0.15
dspp_human	dentin sialophosphoprotein	1253	0.20	0.32	0.15
prpm_human	salivary proline-rich protein PO, fragment	234	0.07	0.27	0.15

CDF analysis predicts from 1.2- to 2.2-fold more sequences to be disordered than charge−hydropathy discrimination. This more permissive prediction seems to be independent of kingdom from this limited set. This result is not unexpected, since CDF analysis was shown to be more accurate on characterized disorder (Figure 4), although the magnitude of the difference in frequency is greater than expected. Also as expected, in all but one case consensus prediction provides an intermediate prediction of disorder. That the consensus predicted less disorder in *Yersinia pestis* than either of its component methods indicates that the methods have proportionally fewer classifications in common for *Y. pestis* than for the other organisms.

Despite differing magnitudes of disorder predicted between methods, there was a common trend among genomes. Archaea and bacteria show a similar frequency of predicted disorder by charge−hydropathy discrimination. CDF analysis predicts that Archaea have a somewhat higher frequency of disorder than Bacteria. However, all methods predict Eukaryotes to contain a much higher proportion of disordered proteins than either of the Prokaryotes. These prediction results show kingdom trends similar to other published estimates (*23*−*25*).

To provide illustrative examples of novel predictions of disordered proteins, predictions were made on the Swiss Prot database. This database provides a source of validated proteins compared to proteins predicted by genome annotation that can fluctuate between annotation versions. CDF, charge−hydopathy, and consensus analyses predict 22.2, 9.6, and 14.0% of SwissProt sequences to be disordered. From these predictions, all proteins with CDF curves falling completely below the discrimination boundary were selected and ranked by their distance from the charge−hydropathy boundary. The 20 highest ranked proteins were then selected from a single organism, *Homo sapiens*, to avoid redundant orthologs (Table 5).

As a consequence of ranking proteins by distance from the charge−hydropathy boundary, these proteins represent extremes of high net charge and low hydropathy. For reference, the 9379 human proteins in Swiss-Prot have an average hydropathy of 0.463 ± 0.045 and an average absolute net charge of 0.028 ± 0.032. At least half of the proteins predicted to be highly disordered interact with nucleic acids: seven are involved in chromosome packing

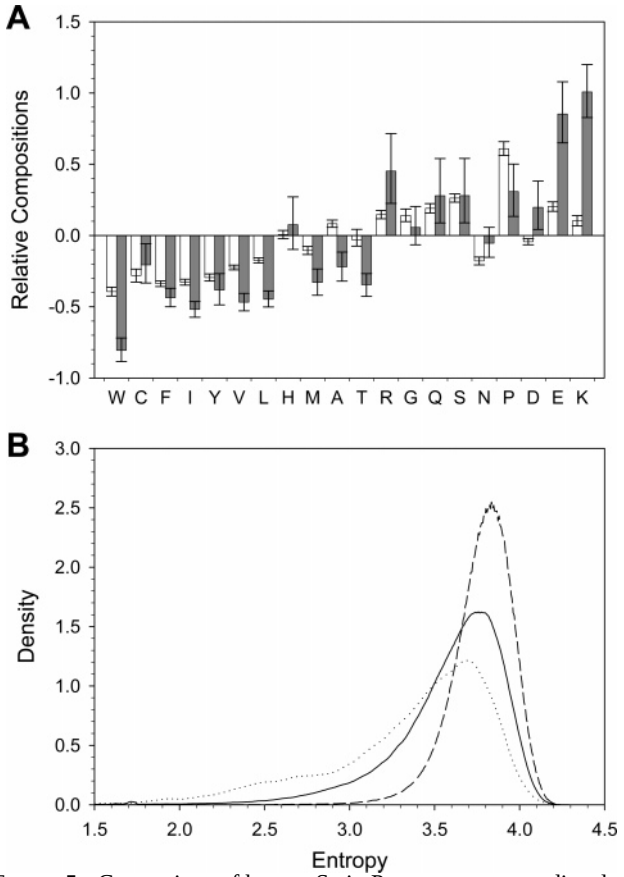

FIGURE 5: Comparison of human SwissProt sequences predicted to be disordered by charge−hydropathy and CDF algorithms. (A) The compositions of proteins predicted to be disordered by CDF by seven boundary points (white bars) and proteins predicted to be disordered by charge−hydropathy by a boundary margin of 0.045 (grey bars) are shown relative to proteins predicted to be ordered by both algorithms. Error bars are the 99% bootstrap confidence intervals. Amino acids are arranged from left to right in the order of increasing flexibility according to Vihenen et al. (*52*) (B) The distribution of the minimum entropy windows from proteins predicted to be ordered by both algorithms (dashed line), predicted to be disordered by a CDF boundary of seven points (solid line), proteins predicted to be disordered by charge−hydropathy by a boundary margin of 0.045 (dotted line).

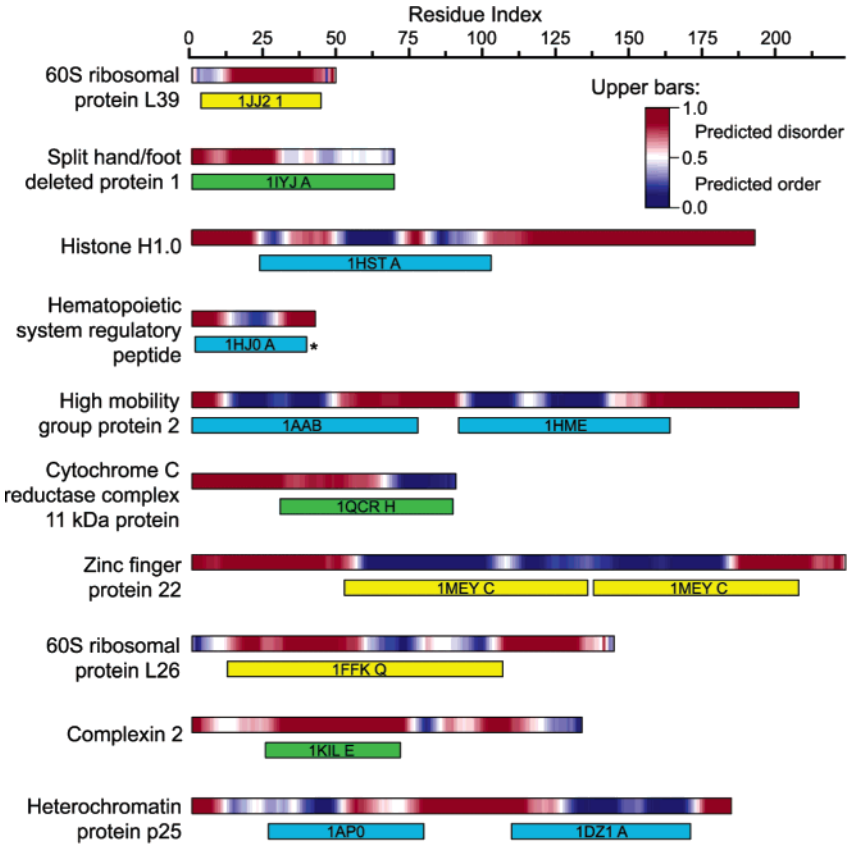

FIGURE 6: The top 10 predictions of wholly disordered human proteins in Swiss Prot that have significant similarity to structures in the PDB. Criteria for similarity was >35% identity for a high scoring pair over at least two-thirds of the PDB chain. Two rows of color-coded bars are shown for each sequence. PONDR VL-XT prediction scores are given by the color gradient of the upper bar. The positions of structure relative to the full-length sequence are given in the lower bars and are color coded according to structural context (blue, monomeric structure; yellow, nucleic acid−protein complex; green, protein−protein complex). PDB IDs are given in each bar.

and maintenance and another three are ribosomal proteins. Other functions of highly disordered proteins include structural scaffolding and ion binding proteins. While some of these proteins have biological functions that are not well understood, at least three have been identified as playing a central role in disease states as diverse as developmental and degenerative disorders.

Comparison of Predictions of Disorder. Comparison of the properties of sequences predicted to be disordered by CDF or charge−hydropathy gave some insight into the relative behaviors of the two algorithms. Specifically, the compositions and sequence entropy were examined for three groups of human sequences from SwissProt: those predicted to be ordered by both algorithms (5673 sequences), those predicted to be disordered by CDF by the entire boundary margin (3434 sequences), and those predicted by charge−hydropathy by a boundary margin of 0.045 (264 sequences). Of the sequences in the charge−hydropathy set, 87% are also in the CDF set. As before, human proteins were selected to avoid redundant related sequences, but the results are qualitatively similar for all SwissProt sequences (data not shown).

Compositions of ordered and disordered sequences have previously been shown to be significantly biased (*24*). Relative to ordered proteins, disordered proteins have decreased compositions of most hydrophobic amino acids and increased compositions of most hydrophilic amino acids. The compositions of predicted ordered sequences and predicted disordered sequences were compared to examine to what extent predictions differ with respect to this finding (Figure 5A). Qualitatively, proteins predicted to be disordered by either algorithm have compositional biases that are consistent with the biases of experimentally verified disordered proteins (*24*). Generally, charge−hydropathy predicted disorder has more extreme compositional biases than CDF predicted disorder, particularly for charged residues.

Globular proteins have been shown to have a minimal requirement for sequence entropy over a window of 45 amino acids, estimated to be approximately 3.10 (*28*). However, high sequence entropy is not a good predictor of order since many disordered proteins have a sequence entropy above this limit. The sequence entropy of predicted disordered proteins was compared to investigate whether prediction methods were biased toward high or low entropy sequences. The entropy of each sequence was calculated over a window of 45 residues, and the distribution was generated by kernel density estimation for each group of proteins (Figure 5B). Proteins predicted to be ordered by both CDF and charge−hydropathy have generally high entropy windows, the vast majority of which are above the minimum for globular protein. CDF predicted disorder is shifted and broadened toward lower entropy windows and charge−hydropathy predicted disorder is shifted and broadened even farther toward low complexity windows.

Comparison of Disorder Predictions to PDB. To illustrate the relationship between disorder predictions on the whole protein level and local structural propensity, proteins predicted to be highly disordered were compared to chains from

the PDB. This comparison is intended to approximate the regions of disorder proteins that may be amenable to structure determination and the conditions that may be necessary to obtain a stable structure. Ten human proteins predicted to be disordered with the furthest distance to the charge−hydropathy boundary, with a CDF curve falling completely below the boundary, and with significant hits to at least one chain in the PDB were selected for illustration (Figure 6). These proteins have charge−hydropathy boundary distances between 0.192 and 0.068, and the percent identity to PDB chains ranged between 43 and 100% over the aligned regions. The only identical hits were observed for human heterochromatin protein p25, for which two domain structures of a closely related murine homologue have been determined.

For the set of proteins that were globally predicted to be disordered, PONDR VL-XT disorder predictions were used to represent their local order−disorder tendency. Comparison of regions of sequences with similarity to known structures and local order−disorder tendency suggest that these proteins fall into two general groups.

One group of proteins has local regions of predicted order that have similarity to regions of known structure. One exceptional case of this is the similarity of hematopoietic system regulatory peptide to thymosin β9 (indicated by an asterisk in Figure 6). Thymosin β9 has no secondary structure in aqueous solvent and requires a 40% 1,1,1,3,3,3-hexafluoro 2-propanol/water solvent for determination of its conformation in solution (*36*). So while this protein has some intrinsic structural propensity indicated by PONDR, it seems unlikely that it will form a stable structure in aqueous solution.

The other group of proteins has no or weak intrinsic structural propensity in regions that have similarity to known structure. However, the similar, structured proteins are involved in protein−protein or nucleic acid complexes. In many cases, disordered proteins have been observed to undergo a disorder-to-order transition when binding to protein or nucleic acid partners (*9*). This suggests that, while these proteins may have regions that can form stable structure, structure determination would first require the identification and inclusion of the appropriate binding partner.

DISCUSSION

Comparison of Disorder Prediction Methods. CDF analysis predicts a much higher frequency of disorder in sequence databases than charge−hydropathy discrimination, particularly for eukaryotes. A large majority of disordered proteins predicted by charge−hydropathy discrimination are also predicted by CDF analysis. Therefore, the latter's higher estimated accuracy for disordered proteins and lower estimated accuracy for ordered proteins account for some of the difference in prediction frequency. However, the details of the algorithms are very different.

Charge−hydropathy discrimination was derived here using linear discrimination based on the distributions of ordered and disordered proteins in charge−hydropathy space. PONDR VL-XT is a neural network, which is a nonlinear classifier, trained to distinguish order and disorder based on a relatively large feature space. Specifically, PONDR VL-XT considers average coordination number (*37*), amino acid compositions (aromatic and charged residues), and net charge (*28*). To a first approximation, one may consider charge−hydropathy feature space to be a subset of PONDR VL-XT feature space, although PONDR VL-XT attributes are more fine-grained in the similar features. This is not to say that charge−hydropathy predictions are equivalent to CDF predictions, but only that PONDR considers the similar attributes as charge−hydropathy prediction in addition to other parameters.

This suggests that differences in predictions by these two classifiers may be physically interpretable, in terms of the protein trinity model (*11*) or the related protein quartet model (*3*). Under the protein trinity model, all soluble proteins fall into one of three classes: one, extended disorder − conformationally dynamic proteins that lack stable secondary and tertiary structure and also lack defined hydrophobic core; two, collapsed disorder − collapsed proteins with defined secondary structure but dynamic tertiary contacts; or three, well-structured − proteins that contain qualities sufficient to fold in isolation. These classes can be characterized by various methods, for example, by examination of the hydrodynamic radius in relation to the protein's molecular weight, which shows clearly that the radius decreases in the order: extended disorder > collapsed disorder > well-structured (*3*). In addition, the related protein quartet model takes into consideration a premolten globule-like state, a distinct conformation with a hydrodynamic radius intermediate to those of extended and collapsed disorder (*3*). These models also postulate that function can arise within each state or from transitions between states.

Under these models, the charge−hydropathy classification is predisposed to discriminate proteins with the extended disorder (coils or premolten globules) from a set of globular conformations (molten globule-like or rigid well-structured proteins). The compositions and entropies of proteins predicted by charge−hydropathy are consistent with this idea. Relative to ordered proteins (Figure 5A), these sequences are highly enriched in charged residues, with twice as much for positively charge residues, and have around half as many of most hydrophobic residues. These polypeptide chains will not be able to fold into a compact globular structure due to weak hydrophobic stabilization and strong electrostatic repulsion (*3*, *20*). The relatively low sequence entropy of these sequences (Figure 5B) also suggests that they will not be able to fold, since it has been shown that there is a minimum required sequence entropy for globular proteins (*28*). In other words, proteins predicted to be disordered by the charge−hydrophobicity approach are likely to belong to the extended disorder class. On the other hand, PONDR-based approaches can discriminate all disordered conformations (coil-like, premolten globules and molten globules) from rigid well-folded proteins, suggesting that charge−hydropathy classification is roughly a subset of PONDR VL-XT, in both predictions of disorder and feature space.

On the basis of the observed behavior of these predictions, we make the following conjectures: (a) Proteins predicted to be disordered by both charge−hydropathy and PONDR (i.e., high net charge and low hydrophobicity) are likely to be in the extended disorder class. (b) Proteins predicted to be disordered by PONDR but predicted to be ordered by charge−hydropathy should have properties consistent with a dynamic, collapsed chain and are likely to be in the collapsed disorder class (i.e., molten globules). (c) Finally, proteins predicted to be ordered by both algorithms are of

course likely to be in the well-structured class. In fact, there is a fourth group of proteins that are predicted to be disordered by charge−hydropathy and predicted to be ordered by PONDR, but this group is small compared to the other three so it is neglected in this interpretation. Importantly, the fact that CDF analysis predicts about 2-fold higher frequency of disorder in sequence databases than charge−hydropathy classification suggests that approximately half of disordered proteins in different proteomes possess extended disorder, whereas another half represents proteins with the collapsed disorder. Currently, work is underway to test this hypothesis, which may lead to a more complete description of the determinants of protein disorder than that provided by charge−hydropathy or PONDR alone.

Potential Data Set Bias. Identification of ordered and disordered regions from crystal structures has the potential to bias the data sets, which is an important consideration here since two of the data sets, wholly ordered proteins and partially ordered proteins, were derived exclusively from crystal structures. It is conceivable that a crystal structure-derived data set may be biased against proteins with characteristics that are unsuitable to crystallization. However, the characteristics that distinguish ordered proteins that crystallize from ordered proteins that do not crystallize are fairly subtle, as evident from the lack of any adequate model that distinguishes these two classes biased on amino acid sequence. While this bias may be an issue for a more detailed sequence−structure analysis, here only gross sequence properties are examined, and so these biases are not likely to have an observable effect on these results. Also, previous results have indicated that sequence determinants of order/disorder are very similar between crystal structures and solution structures (*38*), which further supports the use of crystal structure-based data sets.

Crystallization can cause artificial ordered regions and disordered regions, relative to the solution state of the protein. Large regions of disorder in X-ray structures, as used for this set, can be either true disorder or wobbly domains (*38*). Independent verification of the disorder would be useful for each of these proteins, but such verification would be especially useful for those proteins with >30% disordered residues. As such proteins were not used for classifier development, this distinction is not crucial. However, the presence of wobbly domains rather than disorder will give an overestimation of the usefulness of the classifiers on proteins that contain a mixture of ordered and disordered residues. Conversely, it has been observed that crystallization can induce disorder-to-order transitions thorough crystal contacts (e.g., nucleosome tails, PDB 1EQZ (*39*)). As for disordered regions, this is a fundamental limitation of sampling ordered regions from crystal structures and characterization of ordered and disordered regions by multiple methods (*40*) would be useful to improve the quality of the data set.

The third set of proteins, wholly disordered proteins, differed from the other two sets in that it includes proteins characterized by NMR and CD and not crystallography. The necessity of this is obvious since these techniques are capable of characterizing wholly disordered proteins while crystallography relies on the presence of a unique 3D structure. However, these techniques, as well as structural characterizations in general, are not without caveats. For example, proteins without regular secondary structure, termed NORs by Liu et al. (*41*), may have stable structure but will give a CD spectra similar to disordered proteins. For another example, excluded volume effects, as present in the cellular milieu, have been shown to stabilize protein structure (*42*), and the lack of these effects in vitro may result in false characterization of a protein as wholly disordered (e.g., FlgM (*43*)). However, excluded volume effects are not universal (*44*). The characterization of wholly disordered proteins suffers from as many, if not more, caveats as the characterization of protein order/disorder from crystal structures, so care must be taken in interpreting these results. However, given the parametric nature of the methods used here, some amount of noise in the data set is not likely to significantly bias these results but will reduce the estimated accuracy of these algorithms.

Comparison to Other Studies of Disorder Prediction Methods. At least two groups have also reported use of the distance to the charge−hydropathy boundary as an indication of protein disorder. The FoldIndex[3] measure is similar to the distance value used here, with the caveat that the FoldIndex is not the formal distance to the charge−hydropathy boundary (*45*). Scaling the FoldIndex by 0.338 will give a value equivalent to the present method but will vary somewhat due to the different boundary equations.

Pandey et al. (*46*) compared charge−hydropathy boundary distance to the percent of disordered residues predicted by PONDR. The authors report that charge−hydropathy boundary distance gives predictions equivalent to the fraction of PONDR-predicted disorder. It is apparent from that study and the present study that there is a relationship between these two prediction methods. However, the poor linear fit between PONDR and charge−hydropathy ($R^2 = 0.34$) reported by the authors is more consistent with the complementarity of the prediction methods than with the equivalency of the prediction methods.

Predictions of Ordered and Disordered Proteins. The approach to order−disorder classification used here, as either wholly ordered or wholly disordered, is clearly an approximation of the real physical situation. While some proteins have been characterized as wholly disordered and many wholly ordered proteins are known, many more proteins contain both ordered regions and disordered regions. For practical reasons, however, it is often necessary to summarize local properties across entire proteins. Use of prediction methods intended for prediction of whole protein disorder is therefore useful in this context. Use of whole protein prediction methods prevents the necessity of selecting arbitrary classification criteria based on per-residue disorder prediction methods.

The gray area of classification is of central importance in developing whole protein classification methods. Ideally, structured proteins with disordered loops or termini would be predicted to be structured proteins. Likewise, highly disordered proteins with limited regions of structural propensity would be predicted to be disordered. To test the former, a set of partially ordered proteins was used for predictor validation. Although prediction algorithms differed in performance on this set, the level of performance indicates

[3] http://bioportal.weizmann.ac.il/fldbin/findex.

that the algorithms are not overly biased by a small content of disorder. To investigate the opposite question, proteins predicted to be disordered were compared to structured proteins from PDB (Figure 6). This analysis demonstrates that a strong, but limited, local propensity for order will not cause a highly disordered protein to be classified as ordered.

Intrinsic Protein Disorder in Structural Genomics. Prediction of disorder across genomes suggests that intrinsically disordered proteins will play a significant role in genome-scale projects, particularly those focused on eukaryotic organisms. The high rate of attrition of proteins in high throughput structure determination projects (*47*) may be partially attributable to the failure to adequately account for disordered proteins, particularly in the final stages of structure determination (*48*).

Nonglobular sequences (*49*) are commonly accounted for by filtering low complexity sequences (*50*). Although a minimal sequence complexity is required for the formation of globular structure and many disordered regions have low sequence complexity, this is not an exclusive relationship (*28*, *51*). In fact, only 8% of disordered sequence windows in one study have a complexity lower than that observed for globular proteins (*28*), which means that target prioritization methods that utilize only low complexity filters detect only a small fraction of the disordered proteins in their target lists.

Use of disorder prediction would therefore aid in excluding intrinsically disordered proteins from traditional high throughput structure determination pipelines. This is particularly important because disordered proteins can be the most costly, in time and resources, of all failed structure targets. That is, many disordered proteins may behave reasonably well in solution, but will fail in the final stages of structure determination, which require relatively large quantities of labeled protein. These final stages are the most resource intensive, and therefore greatly increased efficiency could be obtained by avoiding disordered proteins altogether. Indeed, recent experimental results demonstrate that predictions of intrinsic disorder can help improve the yield of structured proteins in structural genomics projects, as assessed by heteronuclear single quantum coherence spectra (*48*).

The high frequency of disordered protein in eukaryotes can be attributable to the increased need for inter- and intracellular signaling and coordination in these organisms. In particular, proteins involved in cellular regulation and related to cancer are highly enriched in disordered residue content (*12*). Since these proteins are among the most biologically interesting targets, in terms of relevance to human health, they are also among the high priority proteins in genome-scale projects. Therefore, disregarding disordered proteins does not contribute to the ultimate goal of understanding cellular processes. Methods for the study of disorder can be integrated with standard structure determination methods (*27*). Disorder prediction can aid the creation of such a system by forming a central fork in the target pipeline that would direct ordered proteins to the conventional pipeline and disordered proteins to an alternative pipeline. The alternative pipeline could include binding partner screens to induce structure in otherwise disordered proteins and empirical domain segmentation to excise ordered domains from disordered domains.

SUPPORTING INFORMATION AVAILABLE

Lists of wholly ordered and partially folded proteins (with corresponding PDB identifications), a list of wholly disordered proteins (with corresponding references), a description of the receiver operator characteristics (ROC) curves, and a figure representing illustrative ROC curves for CDF predictions, charge-hydropathy prediction, and the consensus score. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure−function paradigm, *J. Mol. Biol. 293*, 321−331.
2. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002) Intrinsic Disorder and Protein Function, *Biochemistry 41*, 6573−6582.
3. Uversky, V. N. (2002) Natively unfolded proteins: a point where biology waits for physics, *Protein Sci 11*, 739−756.
4. Liu, D., Ishima, R., Tong, K. I., Bagby, S., Kokubo, T., Muhandiram, D. R., Kay, L. E., Nakatani, Y., and Ikura, M. (1998) Solution structure of a TBP-TAF(II)230 complex: protein mimicry of the minor groove surface of the TATA box unwound by TBP, *Cell 94*, 573−583.
5. Lacy, E. R., Filippov, I., Lewis, W. S., Otieno, S., Xiao, L., Weiss, S., Hengst, L., and Kriwacki, R. W. (2004) p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding, *Nat. Struct. Mol. Biol. 11*, 358−364.
6. Liddington, R. C. (2004) Structural basis of protein−protein interactions, *Methods Mol. Biol. 261*, 3−14.
7. Callaghan, A. J., Aurikko, J. P., Ilag, L. L., Gunter Grossmann, J., Chandran, V., Kuhnel, K., Poljak, L., Carpousis, A. J., Robinson, C. V., Symmons, M. F., and Luisi, B. F. (2004) Studies of the RNA Degradosome-organizing Domain of the *Escherichia coli* Ribonuclease RNase E, *J. Mol. Biol. 340*, 965−979.
8. Spolar, R. S., and Record II, M. T. (1994) Coupling of local folding to site-specific binding of proteins to DNA, *Science 263*, 777−784.
9. Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins, *Curr. Opin. Struct. Biol. 12*, 54−60.
10. Schulz, G. E. (1979) in *Molecular Mechanism of Biological Recognition* (Balaban, M., Ed.) pp 79−94, Elsevier/North-Holland Biomedical Press, New York.
11. Dunker, A. K., and Obradovic, Z. (2001) The protein trinity-linking function and disorder, *Nat. Biotechnol. 19*, 805−806.
12. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins, *J. Mol. Biol. 323*, 573−584.
13. Dobson, C. M. (1993) Flexible friends, *Curr. Biol. 3*, 530−532.
14. Tompa, P., and Csermely, P. (2004) The role of structural disorder in the function of RNA and protein chaperones, *FASEB J. 18*, 1169−1175.
15. Hoh, J. H. (1998) Functional protein domains from the thermally driven motion of polypeptide chains: a proposal, *Proteins 32*, 223−228.
16. Wissmann, R., Baukrowitz, T., Kalbacher, H., Kalbitzer, H. R., Ruppersberg, J. P., Pongs, O., Antz, C., and Fakler, B. (1999) NMR structure and functional characteristics of the hydrophilic N terminus of the potassium channel beta-subunit Kvbeta1.1, *J. Biol. Chem. 274*, 35521−35525.
17. Zhou, H. X. (2001) The affinity-enhancing roles of flexible linkers in two-domain DNA-binding proteins, *Biochemistry 40*, 15069−15073.
18. Romero, P., Obradovic, Z., and Dunker, A. K. (1997) Sequence data analysis for long disordered regions prediction in the calcineurin family, *Genome Informatics 8*, 110−124.
19. Li, X., Romero, P., Rani, M., Dunker, A. K., and Obradovic, Z. (1999) Predicting protein disorder for N-, C-, and internal regions, *Genome Informatics 10*, 30−40.
20. Uversky, V., Gillespie, J., and Fink, A. (2000) Why are "natively unfolded" proteins unstructured under physiological conditions? *Proteins 41*, 415−427.

21. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003) Protein disorder prediction: implications for structural proteomics, *Structure 11*, 1453−1459.
22. Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J. (2003) GlobPlot: Exploring protein sequences for globularity and disorder, *Nucleic Acids Res. 31*, 3701−3708.
23. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J. Mol. Biol. 337*, 635−645.
24. 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) Intrinsically disordered protein, *J. Mol. Graphics Modell. 19*, 26−59.
25. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000) Intrinsic protein disorder in complete genomes, *Genome Informatics 11*, 161−171.
26. Oldfield, C. J. (2003) in *Molecular Biosciences*, p 61, Washington State University, Pullman.
27. Oldfield, C. J., Van, Y. Y., and Dunker, A. K. (2004) in *Structural Genomics* (Kennedy, M. A., Ed.), Humana Press, Totowa, NJ.
28. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein, *Proteins 42*, 38−48.
29. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res. 25*, 3389−3402.
30. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The protein data bank, *Nucleic Acids Res. 28*, 235−242.
31. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The SWISS−PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res. 31*, 365−370.
32. Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Hubbard, T., Kasprzyk, A., Keefe, D., Lehvaslaiho, H., Iyer, V., Melsopp, C., Mongin, E., Pettett, R., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Birney, E. (2003) Ensembl 2002: accommodating comparative genomics, *Nucleic Acids Res. 31*, 38−42.
33. Johnson, D. L. (1998) *Applied Multivariate Methods for Data Analysis*, Duxbury Press, Pacific Grove, CA.
34. Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol. 157*, 105−132.
35. Shannon, C. E. (1948) A mathematical theory of communication, *Bell Syst. Tech. J. 379−423*, 623−656.
36. Stoll, R., Voelter, W., and Holak, T. A. (1997) Conformation of thymosin beta 9 in water/fluoroalcohol solution determined by NMR spectroscopy, *Biopolymers 41*, 623−634.
37. Galaktionov, S. G., and Marshall, G. R. (1996) in *Fourth International Conference on Computational Biology*, pp 42, Washington University Institute for Biomedical Computing, St. Louis, MO.
38. Garner, E., Cannon, P., Romero, P., Obradovic, Z., and Dunker, A. K. (1998) Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization, *Genome Informatics 9*, 201−213.
39. Harp, J. M., Hanson, B. L., Timm, D. E., and Bunick, G. J. (2000) Asymmetries in the nucleosome core particle at 2.5 Å resolution, *Acta Crystallogr. D 56*, 1513−1534.
40. Bracken, C., Iakoucheva, L. M., Romero, P. R., and Dunker, A. K. (2004) Combining prediction, computation and experiment for the characterization of protein disorder, *Curr. Opin. Struct. Biol. 14*, 570−6.
41. Liu, J., Tan, H., and Rost, B. (2002) Loopy proteins appear conserved in evolution, *J. Mol. Biol. 322*, 53−64.
42. Minton, A. P. (2000) Implications of macromolecular crowding for protein assembly, *Curr. Opin. Struct. Biol. 10*, 34−9.
43. Dedmon, M. M., Patel, C. N., Young, G. B., and Pielak, G. J. (2002) FlgM gains structure in living cells, *Proc. Natl. Acad. Sci. U.S.A. 99*, 12681−12684.
44. Flaugh, S. L., and Lumb, K. J. (2001) Effects of macromolecular crowding on the intrinsically disordered proteins c-Fos and p27-(Kip1), *Biomacromolecules 2*, 538−540.
45. Zeev-Ben-Mordehai, T., Rydberg, E. H., Solomon, A., Toker, L., Auld, V. J., Silman, I., Botti, S., and Sussman, J. L. (2003) The intracellular domain of the Drosophila cholinesterase-like neural adhesion protein, gliotactin, is natively unfolded, *Proteins 53*, 758−767.
46. Pandey, N., Ganapathi, M., Kumar, K., Dasgupta, D., Sutar, S. D., and Dash, D. (2004) Comparative analysis of protein unfoldedness in human housekeeping and non-housekeeping proteins, *Bioinformatics 20*, 2904−2910.
47. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., Booth, V., Mackereth, C. D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K. L., Wu, N., McIntosh, L. P., Gehring, K., Kennedy, M. A., Davidson, A. R., Pai, E. F., Gerstein, M., Edwards, A. M., and Arrowsmith, C. H. (2000) Structural proteomics of an archaeon, *Nat. Struct. Biol. 7*, 903−909.
48. Oldfield, C. J., Ulrich, L. E., Cheng, Y., Dunker, A. K., and Markley, J. L. (2004) Addressing the intrinsic disorder bottleneck in structural proteomics, *Proteins*, in press.
49. Wootton, J. C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures, *Comput. Chem. 18*, 269−285.
50. Brenner, S. E. (2000) Target selection for structural genomics, *Nat. Struct. Biol. 7*, 967−969.
51. Romero, P., Obradovic, Z., and Dunker, A. K. (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins, *FEBS Lett. 462*, 363−367.
52. Vihinen, M., Torkkila, E., and Riikonen, P. (1994) Accuracy of protein flexibility predictions, *Proteins 19*, 141−149.
53. Weathers, E. A., Paulaitis, M. E., Woolf, T. B, and Hoh, J. H. (2004) Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered proteins. *FEBS Lett. 576*, 348−352.
54. Garbuzinskiy, S. O., Lobanov, M. Y., and Galzitskaya O. V. (2004) To be folded or to be unfolded? *Protein Sci. 13*, 2871−2877.

BI047993O