Protein flexibility and intrinsic disorder

Predrag Radivojac, [1] Zoran Obradovic, [1] David K. Smith, [2] Guang Zhu, [3]

Slobodan Vucetic, [1] Celeste J. Brown, [4#] J. David Lawson, [4†] A. Keith Dunker [4*¤]

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

2) Department of Biochemistry, University of Hong Kong, Hong Kong

3) Department of Biochemistry, Hong Kong University of Science and Technology, Hong Kong

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

*Correspondence to: A. Keith Dunker, Center for Computational Biology and Bioinformatics, Indiana University, Indianapolis, IN 46202. Telephone: (317) 278-9650. Facsimile: (317) 274-4686. E-mail: kedunker@iupui.edu

Present Address: IBEST, Department of Biological Sciences, University of Idaho, ID 83844

† Present Address: Concurrent Pharmaceuticals, 502 W. Office Center Drive, Fort Washington, PA 19034

¤ Present Address: Center for Computational Biology and Bioinformatics, Indiana University, Indianapolis, IN 46202.

Running Title**:** Protein flexibility and intrinsic disorder

28 pages, 4 tables, 1 figure.

ABSTRACT

Comparisons were made among four categories of protein flexibility: 1. low-B-factor ordered regions; 2. high-B-factor ordered regions; 3. short disordered regions; and 4. long disordered regions. Amino acid compositions of the four sets were found to be statistically different from each other, with high-B-factor ordered and short disordered regions being the most similar pair. The high-B-factor (flexible) ordered regions are characterized by a higher average flexibility index, higher average hydrophilicity, higher average absolute net charge, and higher total charge than disordered regions. The low-B-factor regions are significantly enriched in hydrophobic residues and depleted in the total number of charged residues as compared to the other three classes. We examined the predictability of the high B-factor regions and developed a predictor that discriminates between regions of low and high B-factors. This predictor achieved an accuracy of 70% and a correlation of 0.43 with experimental data, outperforming the 64% accuracy and 0.32 correlation of predictors based solely on flexibility indices. To further clarify the differences between short disordered regions and ordered regions, a predictor of short disordered regions was developed. Its relatively high accuracy of 81% indicates considerable differences between ordered and disordered regions. The distinctive amino acid biases of high-B-factor ordered regions, short disordered regions, and long disordered regions indicate that the sequence determinants for these flexibility categories differ from one another, while the significantly-greater-than-chance predictability of these categories from sequence suggest that flexible ordered regions, short disorder, and long disorder are, to a significant degree, encoded at the primary structure level.

KEY WORDS: temperature factor, natively unfolded, intrinsically unstructured, flexibility prediction

INTRODUCTION

The B-factor of the α-carbon and the B-factor averaged over the four backbone atoms have both been used as measures of residue flexibility of folded proteins (Karplus and Schulz 1985; Vihinen et al. 1994; Kundu et al. 2002). In crystal structures of macromolecules, the B-factor reflects the uncertainty in atom positions in the model and often represents the combined effects of thermal vibrations and static disorder (Rhodes 1993).

B-factors have been studied from a variety of viewpoints. Karplus and Schulz (1985) determined normalized α-carbon B-factors for each amino acid from which flexibility indices were calculated and subsequently used in a sliding-window prediction of the B-factor. Vihinen et al. (1994) and Smith et al. (2003) further developed the method of Karplus and Schulz and improved the correlation between predicted and experimentally determined B-factors. These flexibility indices do not indicate inherent amino acid plasticity, but rather correlate with the tendency of the side chain to be buried or exposed (Sheriff et al. 1985), which can explain, among other behaviors, the mid-range index value for glycine and the high value for proline (Vihinen 1987). Indeed, Halle (2002) showed that the B-factor is inversely proportional to the atomic packing density and argued that little information on polypeptide chains is contained in B-factors apart from the atom coordinates. This theory was supported by Kundu et al. (2002) who achieved significant improvement in predicting experimental B-factors when atomic coordinates were known. Other researchers studied statistical properties of the B-factor (Altman et al. 1994; Wampler 1997) or aspects such as reliability of B-factors (Carugo and Argos 1999), use of B-factors for predicting biologically active sites (Ragone et al. 1989; Carugo and Argos 1998), and use of B-factors for characterizing protein regions (Carugo 2001).

Intrinsically disordered proteins

In addition to regions with high B-factors, crystallized proteins often contain disordered regions characterized by a lack of associated electron density. Some missing density may correspond to wobbly, ordered domains rather than to intrinsically disordered ensembles. However, the amino acid compositions of long

regions of missing electron density are very similar to the amino acid compositions of disordered ensembles characterized by NMR; furthermore, predictors based on NMR-characterized disorder for the most part predict disorder for the long regions of missing electron density. Thus, as an explanation of long regions of missing electron density, wobbly, ordered domains are probably the exception rather than the rule (Garner et al. 1998).

Many other apparently non-crystallizable proteins are mostly comprised of similar disordered regions, with some of these proteins lacking persistent 3-D structure along their entire lengths. Following the work of Ptitsyn and Uversky (1994), we proposed that native proteins may exist in ordered (folded, structured) and/or disordered (unfolded, unstructured) form, where the existence of disorder is determined by overall protein dynamics rather than by local secondary structure. Thus, α-helix, β-sheet, and coil, the three types of secondary structure that are characteristic for ordered chains, may also occur in regions of intrinsic disorder.

Given the strong association of disorder with function (Dunker et al. 2002a), disordered proteins are becoming the subject of increased interest (Wright and Dyson 1999; Dunker et al. 2002a; Dyson and Wright 2002; Uversky 2002b). The predictability of disordered regions from amino acid sequence (Obradovic et al. 2003), the observed compositional biases of such regions (Romero et al. 2001), the typically faster rates of evolution (Brown et al. 2002), and the distinctive amino acid substitution patterns during evolution (Radivojac et al. 2002) combine to strongly indicate that intrinsic protein disorder is generally encoded by the amino acid sequence (Dunker et al. 2002b).

Flexible ordered regions versus intrinsically disordered regions

We and others previously found significant amino acid compositional differences between regions of order and long regions of intrinsic disorder. However, regions of intrinsic disorder and regions of high B-factors (Ringe and Petsko 1986; Smith et al. 1986; Rhodes 1993) could both be associated with large thermal vibrations of individual atoms and with high intramolecular flexibility, so it is important to examine whether high B-factor regions more closely resemble disorder or low B-factor regions. Here, we have extended our studies to four flexibility categories: 1. low-B-factor ordered regions; 2. high-B-factor or-

dered regions; 3. short disordered regions; and 4. long disordered regions. In addition to comparing the local amino acid compositions, we also developed predictors of high versus low B-factor regions and short disordered versus ordered regions. These two predictors were compared with a predictor of long disordered regions (Vucetic et al. 2003). The results of our study indicate that the high B-factor regions are more similar to disorder than to low B-factor regions. Sequence determinants for the high-B-factor regions and intrinsically disordered regions are correlated, but significant differences exist between them as well.

RESULTS

Comparing ordered and intrinsically disordered regions

In this study, an ordered residue is considered to have a high B-factor if its normalized B-factor (Materials and Methods) is 2.0 or higher; otherwise a residue is considered to have a low B-factor. Residues of both low- and high-B-factor ordered sets were extracted from DATASET-O (Materials and Methods). Short disordered residues, i.e. the disordered residues occurring in short stretches, were extracted from DATASET-SD, while long disordered residues were extracted from the previously collected DATASET-LD (Vucetic et al. 2003). The short disordered set was assembled to be similar in its length distribution to the high-B-factor ordered set, while the long disordered set was formed from unrelated proteins having disordered regions of length ≥30 residues.

The amino acid compositions of the low-B-factor ordered, the high-B-factor ordered and the two intrinsically disordered sets were compared to the compositions of a reference ordered set, GLOBULAR-3D (Romero et al. 2001), in order to gain insight into the differences among these datasets (Figure 1). Since the low- and high-B-factor sets contain about 91% and 9% of the ordered amino acids, low-B-factor order has amino acid compositions very similar to those of the reference ordered set. However, the differences from the reference ordered set, although small, are not random: low-B-factor order is slightly enriched in almost all of the more buried residues (left) and slightly depleted in three particular surface residues (right), serine, glutamic acid, and lysine.

(Figure 1)

The high B-factor, short disorder, and long disorder sets exhibit similar depletions of the typically buried tryptophan, phenylalanine, tyrosine, and isoleucine, and similar enrichments in the typically exposed glutamine, glutamic acid, and lysine. The long disorder set shows much less depletion compared to the high B-factor and short disorder sets for cysteine, valine, and leucine. The high B-factor order set is especially enriched in asparagine and aspartic acid, the short disorder set is slightly enriched in these two residues, while the long disorder set is significantly depleted in asparagine, but not in aspartic acid. The high B-factor and short disorder sets are both enriched in glycine, while the long disorder set is not. Finally, the long disorder set is more enriched in proline compared to the high-B-factor order and short disorder sets.

The four distributions can also be compared using a more rigorous statistical approach. Due to the result that there is little higher-order Markov dependence in proteins (Nevill-Manning and Witten 1999), all segments from each group can be concatenated to form four distinct samples S_k (k = 1…4). Each sample S_k can be considered as a realization of an independent and identically distributed random process that emits symbols from an alphabet of 20 amino acid codes. To compare the four amino acid frequency distributions, we calculated the Kullback-Leibler (KL) distance between each pair of distributions p_1 and p_2 as

$$d_{KL}(S_1, S_2) = \sum_{i=1}^{20} p_1(i) \cdot \log_2 \frac{p_1(i)}{p_2(i)},$$

where $p_1(i)$ and $p_2(i)$ represent relative frequencies of amino acid i in samples S_1 and S_2. In all cases, the reference distribution p_2 was chosen to be the one with fewer observations. Table 1 presents the six non-zero KL-distances among these four distributions.

(Table 1)

KL-distance was also used as a test statistic in order to evaluate the statistical significance of the differences between the pairs of underlying sample distributions. Using bootstrapping, we tested the null hypothesis that each pair of samples was generated from the same distribution (also given in Table 1). For the pair with the smallest KL-distance, high B-factor regions and short disordered regions, we rejected the

null hypothesis with a p-value of 0.0053, while the p-values for rejection of the null hypotheses for all other pairs of distributions were significantly lower. Consequently, the estimated probability distributions from Fig. 1 between all four datasets are different with high confidence. Furthermore, the distances suggest that the two most similar sets are high-B-factor order and short disorder, but that these two, together with long disorder, are all closer to one another than any is to the low-B-factor order set.

To further understand the distinctions among the sets, five averages were determined: segment length, flexibility index value, hydropathy, net charge, and total charge (Table 2). Flexibility indices were compared because these are the focus of the current study, while average hydropathy and charge were compared because these two properties have been shown to be an indicator of natively unfolded proteins (Williams 1979; Uversky 2002b). The results from Table 2 indicate, surprisingly, that high-B-factor ordered regions have a higher average flexibility index, a higher average hydrophilicity, a higher average absolute net charge, and a higher total charge than do either short or long disordered regions. The low-B-factor ordered regions are significantly enriched in hydrophobic residues and depleted in the total number of charged residues as compared to the other three classes. Finally, long disordered regions noticeably differ from both short disordered and high-B-factor ordered regions as their total charge is relatively high, but their (absolute) net charge is low with high variance. This indicates an overall balance of positively and negatively charged residues in the set of long disordered segments. Further analysis, however, indicates that individual segments often have significant net positive or negative charge, which contributes to the large variance in the bootstrapping experiment, with a slightly greater occurrence of negatively charged regions.

(Table 2)

Correlation between B-factor values

We investigated the correlation of B-factors between aligned pairs (without gaps) of highly similar protein sequences from DATASET-EO (Materials and Methods). In each iteration of our bootstrap resampling strategy, we randomly selected a set of 195 clusters of homologous sequences and drew no more than 3 protein pairs from each cluster. Correlation coefficients between the B-factor data for each selected pair

were calculated and then averaged over all pairs classified into three ranges of sequence identity. The final estimate of correlation was obtained as the average over all bootstrap iterations within each range (Table 3). The correlation between B-factor values at aligned residues clearly decreases as sequence identity decreases, which is expected. Table 3 also illustrates the extent to which experimental conditions and crystal packing may influence B-factor values. Homologous pairs crystallized within the same space group have more highly correlated B-values than homologous pairs crystallized in different space groups.

(Table 3)

In the next experiment, we studied the effect of normalization on discrimination between low- and high-B-factor residues and approximated the upper limit on predictability of the high B-values. Raw data and data normalized using a method by Smith et al. (2003) were dichotomized into class high, if the B-values were at least 32Å^2 (2.0), and class low. These thresholds provided equal class ratios in both cases. For all pairs of identical sequences selected from DATASET-EO, we then compared the proportion of superimposed residues with the same class and confirmed that the normalization process significantly improves agreement between the residues (data not shown). Since experimental reproducibility limits our ability to predict B-factors, we believe that the average of the agreement between class high (65.2%) and class low (96.8%) sets the upper limit on predictability of the B-factor only from amino acid sequence to approximately 81%.

Predicting B-factor values

Despite the problems that arise from differences in crystal environments, B-factors show correlation with amino acid sequence, which suggests that they should be predictable from amino acid sequence. To test this hypothesis, three logistic regression models based on different attribute sets were trained to discriminate between high and low B-factors. The models were systematically evaluated for various window sizes, w_{in} and w_{out}, and the best results were in all cases obtained for $w_{in} = 1$ for structural attributes, $w_{in} = 5$ for non-structural attributes and $w_{out} = 5$. The three models are called the NS predictor, which uses no structural information, the KS predictor, which uses known secondary structure, and the PS predictor, which uses predicted secondary structure.

The NS predictor reached 64.5% accuracy (sn = 62.8 ± 0.9, sp = 66.1 ± 0.3), the PS predictor reached 67.0% accuracy (sn = 66.8 ± 0.9, sp = 67.2 ± 0.4), while the KS predictor reached 67.8% accuracy (sn = 65.3 ± 0.8, sp = 70.3 ± 0.3). The disparity in confidence intervals is due to the difference in sizes between the two classes. Construction of non-linear models only marginally improved prediction accuracy (64.5% for the NS, 67.2% for the PS and 68.3% for the KS predictor). Although the models were trained only to discriminate between high and low B-factor regions, we found that the approximated probability that the residue has a high B-factor is well correlated with the experimental B-values. The observed correlation coefficients for the experimental data versus the raw outputs of the NS, PS, and KS predictors reached 0.34 ± 0.02, 0.38 ± 0.02, and 0.41 ± 0.02, respectively.

The prediction accuracies and correlation coefficients of our B-factor predictors were compared with a predictor based only on flexibility indices by Vihinen et al. (1994), which was previously found to outperform other similar methods. The method by Vihinen et al. achieved 63.8% accuracy, while the correlation coefficient with the experimental data was 0.32 ± 0.02. Thus, our PS single-sequence predictor attained an improvement of 3.4 percentage points (5.3%) in prediction accuracy and 0.06 (19%) in correlation coefficient as compared to that of Vihinen et al.

B-factor predictor with evolutionary modeling

It is well-known that adding evolutionary information in the form of sequence alignments leads to improved secondary structure prediction (Benner et al. 1992; Levin et al. 1993; Rost 2001). In recent examples of this principle, Jones (1999) and Przybylski and Rost (2002) improved single sequence prediction accuracy by 2-4 percentage points. Using a similar reasoning for B-factor prediction, we constructed protein families using PSI-BLAST and enhanced the performance of our models (Materials and Methods). The average improvement of the prediction results was 2.0 percentage points for the NS predictor and 2.5 percentage points for the PS predictor. Thus, the overall prediction accuracy reached 69.7%. We note that, the higher the number of available homologs, the higher the prediction accuracy. For example, in the case when 30 or more non-redundant homologs can be found, the average prediction accuracy reaches 70.8%.

In terms of average correlation coefficients, PSI-BLAST enhanced NS and PS predictors reached 0.36 ± 0.02 and 0.43 ± 0.02, respectively. Thus, the overall improvement over a predictor based only on flexibility indices by Vihinen et al. reached 5.9 percentage points (9.2%) in prediction accuracy and 0.11 (34.4%) in correlation coefficient. The quality of our predictions can be verified from the figure presented in supplemental data.

Predictor-based analysis of the ordered and disordered data

To further explore the relationship between the ordered and disordered datasets that was suggested by the amino acid frequency data, we used two predictors of intrinsic disorder: (i) previously constructed predictor of long disordered regions, VL2 (Vucetic et al. 2003), and (ii) a logistic-regression based predictor developed here to discriminate between short disordered regions and ordered regions. The short disorder predictor, named XS1 according to our conventions (Obradovic et al. 2003), was developed from DATASET-SD and used the same set of attributes as our PS high B-factor predictor. The maximum performance of 80.6% was obtained using $w_{in} = 9$, $w_{out} = 7$, while the structural attributes were averaged in a window of 5.

The high B-factor predictor, short disorder predictor and long disorder predictor were all applied to three datasets (DATASET-O, DATASET-SD, and DATASET-LD) and the prediction results are shown in Table 4. This experiment confirmed that high B-factors and short disorder are the most similar phenomena among the three. On the other hand, VL2 performance on both B-factor and short disorder datasets was weak, in part caused by longer averaging ($w_{in} = w_{out} = 41$). Correlation coefficients between predictor outputs were: 0.26 ± 0.02 between VL2 and the high B-factor predictor, 0.31 ± 0.02 between VL2 and the short disorder predictor, and 0.88 ± 0.02 between high B-factor and the short disorder predictor.

(Table 4)

DISCUSSION

Properties of flexibility data

Comparing the B-factor values from highly similar pairs of crystallized chains provides evidence that flexibility is encoded at the amino acid sequence level to a significant degree and therefore should be predictable, at some level, from the amino acid sequence (Table 3). However, because of variations that result from experimental conditions, crystal contacts, or refinement procedures, the B-factor data are noisy.

Crystal packing effects can be viewed as a special case of non-local interactions. Given the dependence of the B-factor on packing density (Halle 2002) and hence on non-local interactions, crystal packing would be expected to exert large effects on B-factor values. In agreement with this, previous comparisons indicated that different crystal forms of myohemerythrin (Sheriff et al. 1985) and myoglobin (Phillips 1990) exhibited rather low correlations in their B-values, with further confirmation on additional protein pairs (Kundu et al. 2002). Our comparisons of many similar and identical proteins in the same and different space groups show that crystal packing effects generally perturb B-factor values and the effects can be very significant (Table 3). Overall, the B-factor perturbations arising from crystal packing effects are probably the largest source of noise in the B-factor data.

Prediction accuracy

Prediction of B-factors cannot exceed the accuracy with which B-factors can be experimentally reproduced; thus, the noise in the B-factor data sets an upper limit to prediction of flexibility. To estimate this upper limit, we collected pairs of B-factor sets from identical proteins and subjected the data to the same analysis used to compare the predicted and observed B-factor values. The results suggest that the upper limit on prediction accuracy is approximately 81%. In terms of the agreement between raw predictions and experimental values, the upper limit on the correlation coefficient is about 0.8 (Table 3). From this perspective, our achievement of about 70% accuracy and correlation coefficient of 0.43 seems quite reasonable.

Our predictor of high B-factors joins many other machine learning tools that attempt to predict protein features from amino acid sequence (Lund et al. 1997; Blom et al. 1999; Jones 1999; Pollastri et al.

2002; Obradovic et al. 2003). Its prediction accuracy is comparable to the 64-77% accuracies for coordination number, two-class inter-residue distances or relative solvent accessibility and lower than the 75-80% prediction accuracy of secondary structure or long regions of intrinsic disorder. Since flexible ordered and short disordered protein regions are frequently involved in important biological functions and they were not previously predictable from the sequence using our old predictors, we expect this B-factor predictor to be an advanced practical tool to aid in the automated discovery of short molecular recognition regions and possibly even the active sites. Moreover, the raw outputs of this predictor can be utilized in semi-automated detection of flexible ordered regions (see supplemental material). The correlation of the high-B-factor regions with short disordered regions may prove important in high-throughput genome-wide characterization of novel proteins with unknown structure and function.

The improvement in B-factor prediction from adding either known (KS predictor) or predicted (PS predictor) secondary structure is small but significant. This improvement is related to the differences in average flexibility observed over the three structural categories (data not shown). Addition of evolutionary information obtained by PSI-BLAST alignments improves prediction of B-factors, both for NS and PS predictors. The improvement of about 3 percentage points matches the increase in secondary structure prediction (Przybylski and Rost 2002). The fact that the evolutionary information improved prediction results and that the PSI-BLAST enhanced PS predictor outperformed the KS predictor is further support for the predictability of B-factor values from amino acid sequence.

In terms of correlation coefficients, results achieved in this study exceed other methods from the literature. Predictors by Karplus and Schulz (1985), Vihinen et al. (1994) and Smith et al. (2003) reach correlation coefficients between 0.30 and 0.33, while some earlier methods (Bhaskaran and Ponnuswamy 1988; Ragone et al. 1989) cannot surpass 0.3. On the other hand, our PS predictor reached 0.38 without the presence of evolutionary information, while, on average, homologous sequences boost the correlation coefficient to 0.43. However, a gap of 0.23 between sequence based methods and the 0.66 found using the methods of Kundu et al. (2002), which includes known atom coordinates, is still significant.

The gap between sequence-based approaches and approaches based on atomic coordinates is likely to be further decreased in future. An immediate route is noise reduction, which can be effectively achieved by determining residues that are involved in crystal contacts and excluding them from model training. We believe that the improvement similar to that in methods based on atomic coordinates can result (Kundu et al. 2002). Additionally, due to the imbalance between sizes of low vs. high B-factor classes, our model was constructed using balanced data that, in turn, lead to a significant over-prediction of the high B-factors. In our future research, we will study ways to detect locally flexible regions based on their local and non-local neighborhoods and thus reduce the number of false positives outputted by our model.

Comparing compositions of high-B-factor ordered and intrinsically disordered proteins

Our original hypothesis was that amino acid composition determines whether a protein folds into specific 3-D structure or not. While early indications of this idea were developed from structural studies on protein sequences (Williams 1978), we missed this original work and developed our version of this hypothesis from prior studies on lattice models of protein structure by Shakhnovich and Gutin (1993). In those lattice studies, the determination whether a lattice-model protein folds or not depended on the polar/nonpolar ratio, which corresponds to the amino acid composition in real proteins. Given a folding polar/nonpolar ratio (composition), the detailed arrangement of the amino acids indicated which fold was stabilized. Here we suggest that, not only foldability, but also flexibility is determined, to a significant degree, by the amino acid composition.

Comparison of the amino acid compositions of experimentally characterized regions of protein disorder with regions of order (Romero et al. 2001) showed that disordered proteins generally have more of the flexible amino acids as defined by the scale of Vihinen et al. (1994), suggesting that disordered regions and high B-factor regions might be quite similar to each other. Furthermore, Romero et al. (1997) also indicated that disordered regions of different lengths might have different amino acid compositions, but the original datasets were quite small. Here, comparisons of the amino acid compositions of low and high B-factor regions and short and long disordered regions indicate that all four categories are distinct

(Fig. 1, Tables 1-2). While the compositional distinctions among the high-B-factor order, short disorder and long disorder sets might change as more data are added, we expect the overall trends indicated in Tables 1-2 to be maintained. This expectation is based on the observation that the current datasets are large enough already to show statistically significant distinctions.

Just as amino acid compositions vary for different types of secondary structure (Nakashima et al. 1986; Liu and Chou 1999; Cai et al. 2002), compositional differences might distinguish different types of intrinsic disorder or different types of flexible regions. For example, regions of extended disorder might be expected to be more hydrophilic than either regions of collapsed disorder or regions corresponding to the premolten-globule, if indeed this form is distinctive (Uversky 2002a). Also, there could be compositional biases in subsets of intrinsically disordered proteins that correlate with function such as enrichments in lysine and arginine for nucleic acid binding regions. Indeed, recently published work provides some support for this conjecture (Vucetic et al. 2003).

Previously we found significant amino acid compositional differences between ordered protein and long regions of intrinsic disorder. If structure-sequence relationships existed on a continuum, then one would expect to observe monotonic increases or decreases in the various amino acid compositions as the set of interest is changed from low B-factor regions, to high B-factor regions, to short disordered regions and to long disordered regions. However, almost none of the amino acids exhibit monotonic changes in the order indicated. Even the global averages of Table 2 do not exhibit monotonic changes across the different flexibility/disorder classes in the order indicated. Thus, the amino acid compositions that specify flexibility and intrinsic disorder are evidently distinct and not merely quantitative differences on a continuum.

MATERIALS AND METHODS

Datasets

The first set of protein chains, DATASET-O, consists of 290 non-redundant sequences from the PDB (Berman et al. 2000) selected in the study of Smith et al. (2003). All crystallized chains, consisting of at least 80 amino acids, were required to have a resolution of ≤2Å, and an R-factor ≤20%. Sequence identity

within the set was limited to 25% and only chains without non-standard residues and missing backbone or side chain atoms were chosen, making a database of 67,552 residues in total.

The second set of protein chains, DATASET-EO, contains 1,287 sequences from the PDB divided into 195 disjoint clusters of similar sequences. For each chain in a cluster there is at least one other chain with ≥50% sequence identity. Minimum and maximum cluster sizes are 2 and 205, while the total number of residues is 238,133. All proteins in the dataset were required to have at least 50 residues and a resolution of ≤2Å.

The third dataset, DATASET-SD, was extracted from the PDB and it contains non-redundant chains with stretches of missing coordinates no longer than 10 consecutive residues. The length limitation of 10 residues was chosen in order to make the average segment length and standard deviation comparable to the high B-factor regions from DATASET-O. All chains from DATASET-SD were required to be at least 80 residues in length while the maximum sequence identity between any two chains was limited to 25%. DATASET-SD contains 511 sequences with 3,216 disordered residues in short stretches out of 174,301 total residues.

All datasets are publicly available at our website www.ist.temple.edu.

Data representation and types of predictors

In order to construct a predictor, a machine-learning example (data point) was constructed for each residue where the corresponding C-α atom B-factor was quantized into classes high and low, according to a threshold, and included as a binary target designation. To compensate for the large variability of averages over proteins, C-α B-factors were normalized using the method of Smith et al. (2003) prior to quantization.

An attribute vector for each position in a protein was constructed considering neighboring amino acids within a symmetric input window of size w_{in}. The window was centered at a given position except near N-/C-terminus where it was allowed to expand/collapse. The first twenty one attributes were the twenty relative frequencies of each amino acid within w_{in} and K_2 entropy, a measure of sequence com-

plexity (Wootton and Federhen 1996). The last set of attributes used in this study exploits secondary structure information. Since each residue may belong to structure forms α-helix, β-sheet, and coil, we included three structural attributes, constructed in the same way as compositional attributes. The NMR or X-ray determined structures of a query sequence were used for the KS predictor (known structure), the first of the three models built in this work. For proteins whose structure was unknown, the raw PHD secondary structure predictions (Rost et al. 1997) on the query amino acid sequence were used. We refer to the predictor using PHD scores as the PS predictor (predicted structure). Finally, the NS predictor (no structure), which does not exploit secondary structure information, was used for comparison purposes. It is possible to optimize the size of the input window for each attribute, however, due to the high computational requirements, the window size for the structural attributes was optimized separately from the remaining attributes.

After predictions were made for each residue in a protein, the raw outputs were smoothed using a moving average post-filtering. The size of the smoothing (output) window w_{out} was also subject to optimization.

Model choice, training, and evaluation criteria

We use logistic regression for linear modeling and bagged neural networks (Breiman 1996) for non-linear modeling. To train a predictor we applied the following procedure: the original set of 290 proteins was first randomly split into training and testing sets in the ratio 75%-25%. From the set of training proteins we constructed examples for all available residues and then fed them to the model, which learned from a class balanced data set. After the model was trained, we evaluated its performance on all examples from the test set. The whole process of splitting, training, and testing was repeated 30 times in all experiments.

In order to evaluate the performance of the predictors, we measured sensitivity (*sn*) and specificity (*sp*) for a given set of parameters. Sensitivity is defined as the percentage of high B-factors correctly predicted, while specificity is the percentage of low B-factors correctly predicted (Hastie et al. 2001). This type of model evaluation is commonly used in cases of class imbalance (Kubat et al. 1998). Assum-

ing the class sizes are equal, the accuracy of prediction (*acc*) is expressed as the arithmetic mean of sensitivity and specificity. Therefore, random predictors or models that always output only one class will have an accuracy of 50%. Together with sensitivity and specificity, we also report their 95% confidence intervals calculated as $\pm 2 \cdot s / \sqrt{n}$, where s is the standard deviation of the estimate (*sn* or *sp*) and n the number of experimental repetitions.

Prediction averaging over evolutionary data

Families of homologous proteins were built using PSI-BLAST queries of GenBank (Benson et al. 1999). The conditions for the PSI-BLAST queries included using the BLOSUM62 scoring matrix (Henikoff and Henikoff 1992) with 11/1 gap penalties and E-values of 0.0002 to include a sequence in a profile and of 0.01 to accept it as a family member. The maximum number of iterations was limited to three in order to constrain the influence of potential false positives. Construction of profiles usually incorporates some form of weight assignment in order to avoid the influence of very similar hits, but also sequences from the "twilight zone". As noted in the study of Altschul et al. (1997), several intuitive weighting schemes usually yield similar results. Based on these previous studies, the following simple scheme was devised: all sequences with sequence identity above 70% or below 30% in the region of the local alignment to the query sequence were discarded from the family. Additionally, no pair of homologs within a family was allowed to exceed the 70% sequence identity threshold. Pairwise sequence alignments were performed using the Smith-Waterman algorithm (1981) with the BLOSUM62 scoring matrix and 11/1 gap penalties. The remaining sequences in each family were all assigned equal weights and prediction of the B-factor for the query sequence at position i was formed as an average over all proteins in a family that do not have a gap at that position.

ACKNOWLEDGEMENTS

This work was supported by the following grants: NIH grant 1R01 LM06916 awarded to AKD and ZO; NSF grants NSF-CSE9711532 and NSF-11S-0196237 awarded to ZO and AKD; University of Hong Kong CRCG grant 10202779 awarded to DKS; Research Grants Council of HK grants HKUST6208/00M

and 6124/02M awarded to GZ. We also acknowledge the help of T. R. O'Connor and C. J. Oldfield, both undergraduate students at Washington State University. Finally, we acknowledge the many detailed suggestions of two anonymous reviewers.

REFERENCES

Altman, R.B., Hughes, C., and Jardetzky, O. 1994. Compositional characteristics of disordered regions in proteins. *Prot. Peptide Letters* **2:** 120-127.

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.

Benner, S.A., Cohen, M.A., and Gerloff, D. 1992. Correct structure prediction? *Nature* **359:** 781.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A., and Wheeler, D.L. 1999. GenBank. *Nucleic Acids Res.* **27:** 12-17.

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.

Bhaskaran, R., and Ponnuswamy, K.P. 1988. Positional flexibilites of amino acid residues in globular proteins. *Int. J. Pept. Protein Res.* **32:** 241-255.

Blom, N., Gammeltoft, S., and Brunak, S. 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294:** 1351-1362.

Breiman, L. 1996. Bagging predictors. *Mach. Learn.* **24:** 123-140.

Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T., Oldfield, C.J., Williams, C.J., and Dunker, A.K. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **55:** 104-110.

Cai, Y.D., Liu, X.J., Xu, X.B., and Chou, K.C. 2002. Artificial neural network method for predicting protein secondary structure content. *Comput Chem* **26:** 347-350.

Carugo, O. 2001. Detection of breaking points in helices linking separate domains. *Proteins* **42:** 390-398.

Carugo, O., and Argos, P. 1998. Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. *Proteins* **31:** 201-213.

Carugo, O., and Argos, P. 1999. Reliability of atomic displacement parameters in protein crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **55:** 473-478.

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

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

Dyson, H.J., and Wright, P.E. 2002. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12:** 54-60.

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 Inform. Ser. Workshop Genome Inform.* **9:** 201-213.

Halle, B. 2002. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA.* **99:** 1274-1279.

Hastie, T., Tibshirani, R., and Friedman, J.H. 2001. *The elements of statistical learning: data mining, inference, and prediction*. Springer Verlag, New York, NY.

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89:** 10915-10919.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292:** 195-202.

Karplus, P.A., and Schulz, G.E. 1985. Prediction of chain flexibility in proteins. *Naturwissenschaften* **72:** 212-213.

Kubat, M., Holte, R.C., and Matwin, S. 1998. Detection of oil spills in satellite radar images of sea surface. *Mach. Learn.* **30:** 195-215.

Kundu, S., Melton, J.S., Sorensen, D.C., and Phillips, G.N., Jr. 2002. Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys. J.* **83:** 723-732.

Kyte, J., and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157:** 105-132.

Levin, J.M., Pascarella, S., Argos, P., and Garnier, J. 1993. Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng.* **6:** 849-854.

Liu, W., and Chou, K.C. 1999. Prediction of protein secondary structure content. *Protein Eng* **12:** 1041-1050.

Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J., and Brunak, S. 1997. Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng.* **10:** 1241-1248.

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

Nevill-Manning, C.G., and Witten, I.H. 1999. Protein is incompressible. In *Data Compression Conference*, pp. 257-266, Snowbird, Utah.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., and Dunker, A.K. 2003. Predicting intrinsic disorder from amino acid sequence. *Proteins*, **in press**.

Phillips, G.N., Jr. 1990. Comparisons of the dynamics of myoglobin in different crystal forms. *Biophys. J.* **57:** 381-383.

Pollastri, G., Baldi, P., Fariselli, P., and Casadio, R. 2002. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* **47:** 142-153.

Przybylski, D., and Rost, B. 2002. Alignments grow, secondary structure prediction improves. *Proteins* **46:** 197-205.

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

Radivojac, P., Obradovic, Z., Brown, C.J., and Dunker, A.K. 2002. Improving sequence alignments for intrinsically disordered proteins. *Pac. Symp. Biocomput.* **7:** 589-600.

Ragone, R., Facchiano, F., Facchiano, A., Facchiano, A.M., and Colonna, G. 1989. Flexibility plot of proteins. *Protein Eng.* **2:** 497-504.

Rhodes, G. 1993. *Crystallography made crystal clear: a guide for users of macromolecular models*. Academic Press, San Diego.

Ringe, D., and Petsko, G.A. 1986. Study of protein dynamics by X-ray diffraction. *Methods Enzymol.* **131:** 389-433.

Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., and Dunker, A.K. 1997. Identifying disordered regions in proteins from amino acid sequences. *IEEE Int. Conf. Neural Netw.* **1:** 90-95.

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.

Rost, B. 2001. Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134:** 204-218.

Rost, B., Schneider, R., and Sander, C. 1997. Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270:** 471-480.

Shakhnovich, E.I., and Gutin, A.M. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. USA.* **90:** 7195-7199.

Sheriff, S., Hendrickson, W.A., Stenkamp, R.E., Sieker, L.C., and Jensen, L.H. 1985. Influence of solvent accessibility and intermolecular contacts on atomic mobilities in hemerythrins. *Proc. Natl. Acad. Sci. USA* **83:** 1104-1107.

Smith, D.K., Radivojac, P., Obradovic, Z., Dunker, A.K., and Zhu, G. 2003. Improved amino acid flexibility parameters. *Protein Sci.* **12:** 1060-1072.

Smith, J.L., Hendrickson, W.A., Honzatko, R.B., and Sheriff, S. 1986. Structural heterogenity in protein crystals. *Biochemistry* **25:** 5018-5027.

Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195-197.

Uversky, V.N. 2002a. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11:** 739-756.

Uversky, V.N. 2002b. What does it mean to be natively unfolded? *Eur. J. Biochem.* **269:** 2-12.

Vihinen, M. 1987. Relationship of protein flexibility to thermostability. *Protein Eng.* **1:** 477-480.

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

Vucetic, S., Brown, C.J., Dunker, A.K., and Obradovic, Z. 2003. Flavors of protein disorder. *Proteins* **52:** 573-584.

Wampler, J.E. 1997. Distribution analysis of the variation of B-factors of X-ray crystal structures; temperature and structural variations in lysozyme. *J. Chem. Inf. Comput. Sci.* **37:** 1171-1180.

Williams, R.J. 1978. The conformational mobility of proteins and its functional significance. *Biochem. Soc. Trans.* **6:** 1123-1126.

Williams, R.J.P. 1979. The conformational properties of proteins in solution. *Biol. Rev. Camb. Philos. Soc.* **54:** 389-437.

Wootton, J.C., and Federhen, S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol* **266:** 554-571.

Wright, P.E., and Dyson, H.J. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293:** 321-331.

Tables

Table 1. Kullback-Leibler distance (p-value*) between estimated probability distributions of four data sets.

	High-B-factor order	Short disorder	Long disorder
Low-B-factor order	0.181 ($p < 10^{-4}$)	0.142 ($p < 10^{-4}$)	0.102 ($p < 10^{-4}$)
High-B-factor order		0.012 ($p = 0.0053$)	0.051 ($p < 10^{-4}$)
Short disorder			0.033 ($p < 10^{-4}$)

*Estimates of p-values were calculated in $N = 50{,}000$ bootstrap iterations. As a reference, KL-distance between the four distributions and the uniform distribution are: 0.16 for low-B-factor order, 0.42 for high-B-factor order, 0.40 for short disorder, and 0.32 for long disorder.

Table 2. Properties of proteins from four data sets

	Segment length (st. dev)	Flexibility*,**	Hydropathy†	Net charge	Total charge††
Low B-factor order	34.2 (35.4)	0.996 ± 0.001	–0.125 ± 0.041	–0.008 ± 0.006	0.207 ± 0.007
High B-factor order	4.3 (3.5)	1.027 ± 0.002	–1.310 ± 0.084	–0.059 ± 0.018	0.326 ± 0.016
Short disorder	4.6 (2.2)	1.024 ± 0.002	–1.175 ± 0.106	–0.038 ± 0.023	0.310 ± 0.019
Long disorder	127.8 (231.7)	1.015 ± 0.002	–0.853 ± 0.091	–0.005 ± 0.024	0.294 ± 0.017

*The per segment means and 95% confidence intervals for flexibility, hydropathy, and charge were calculated using bootstrapping. All regions of length 1, methionine at the N-terminus, and His-tags were excluded from each data set.

**Vihinen et al. (1994)

†Kyte and Doolittle (1982)

††Calculated as the fraction of charged residues in each segment.

Table 3. Relationship between B-factors of highly similar sequences

as a function of sequence identity and space groups

	Correlation coefficients*		
Sequence identity (*si*)	All Pairs	Same Space Group	Different Space Group
si ∈ [70, 90) %	0.59 ± 0.07; 0.56; 23	0.63 ± 0.09; 0.65; 10	0.59 ± 0.06; 0.56; 21
si ∈ [90, 100) %	0.76 ± 0.04; 0.81; 122	0.82 ± 0.03; 0.88; 93	0.61 ± 0.04; 0.61; 66
si = 100 %	0.79 ± 0.02; 0.86; 290	0.81 ± 0.02; 0.86; 286	0.63 ± 0.05; 0.66; 50

*Average correlation coefficient ± 95% confidence intervals; median; average number of pairs.

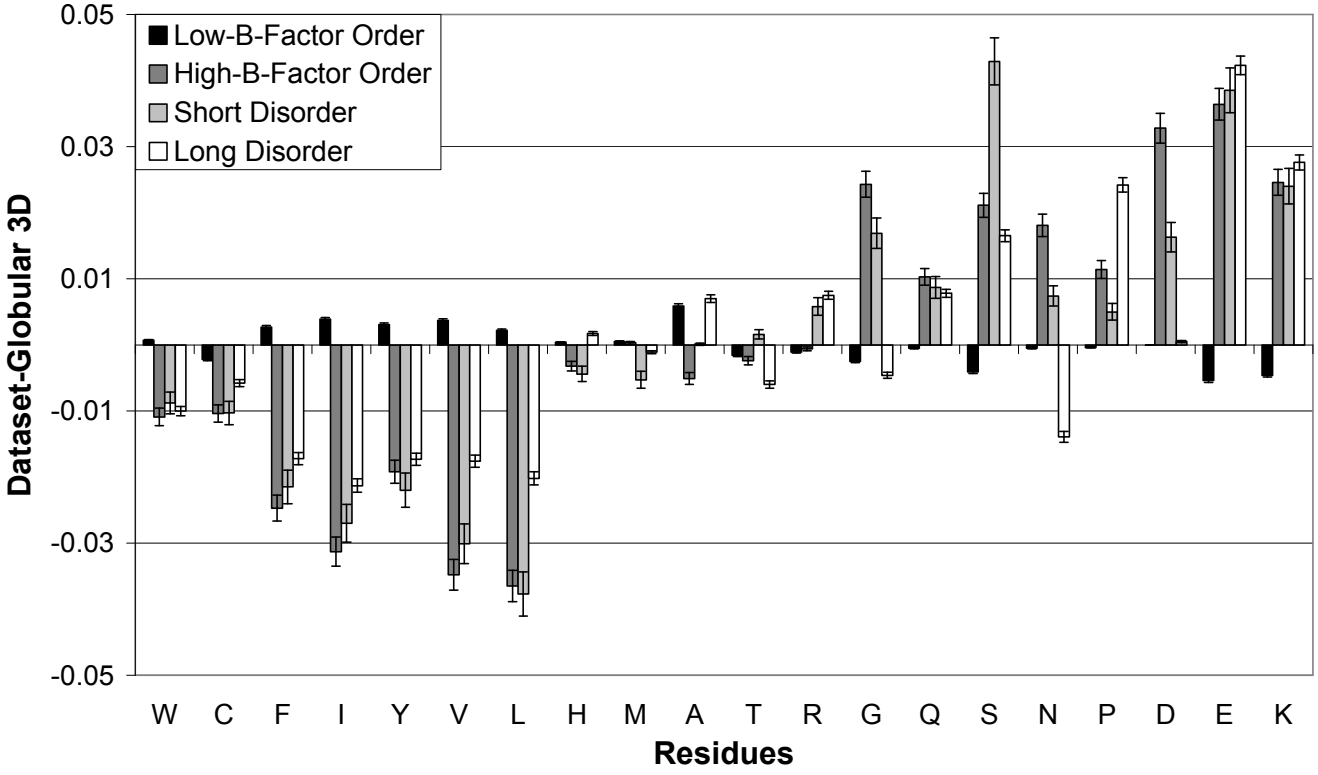
Table 4. Prediction accuracies ±95% confidence intervals for the PS B-factor predictor, PS predictor of short disorder, and VL2 long disorder predictor on 3 datasets.

	Prediction accuracy [%]								
	DATASET-O			DATASET-SD			DATASET-LD		
	sensitivity	specificity	accu-racy	sensitivity	specificity	accu-racy	sensitivity	specificity	accu-racy
B-factor predictor	65.1 ± 2.3	68.7 ± 1.0	66.9	67.4 ± 2.2	85.2 ± 0.7	76.3	59.6 ± 3.4	64.9 ± 1.0	62.3
Short disorder predictor	41.6 ± 2.4	83.8 ± 0.9	62.7	78.1 ± 2.8	83.0 ± 0.6	80.6	51.5 ± 4.0	80.3 ± 1.1	65.9
Long disorder predictor	17.7 ± 3.0	87.2 ± 2.2	52.6	33.6 ± 3.8	82.2 ± 1.8	57.9	76.2 ± 5.0	83.8 ± 2.4	80.0

All accuracies were estimated on a per protein basis, i.e. sensitivity and specificity were calculated for all proteins and then averaged. Prediction accuracy was obtained as an average of estimated sensitivity and specificity.

Figure legends

Figure 1. Amino acid compositions of various data sets. The composition of each amino acid of a reference data set of ordered proteins, GLOBULAR-3D, is subtracted from the composition of the four sets described herein; thus, negative peaks indicate depletions compared to the ordered reference set and positive peaks represent enrichments. The order of the amino acids along the x-axis is from the most buried (left) to the most exposed (right) in typical globular proteins. Error bars indicate one standard deviation. Methionine at the N-terminus and His-tags were not included in calculations.


0.05
0.03
0.01
-0.01
-0.03
-0.05
Dataset-Globular 3D
Low-B-Factor Order
High-B-Factor Order
Short Disorder
Long Disorder
W C F I Y V L H M A T R G Q S N P D E K
Residues