Protein Flexibility and Rigidity Predicted From Sequence

Avner Schlessinger[1,2] and Burkhard Rost[1–3]*
[1]*CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*
[2]*Columbia University Center for Computational Biology and Bioinformatics, New York, New York*
[3]*Northeast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*

ABSTRACT **Structural flexibility has been associated with various biological processes such as molecular recognition and catalytic activity. In silico studies of protein flexibility have attempted to characterize and predict flexible regions based on simple principles. B-values derived from experimental data are widely used to measure residue flexibility. Here, we present the most comprehensive large-scale analysis of B-values. We used this analysis to develop a neural network–based method that predicts flexible–rigid residues from amino acid sequence. The system uses both global and local information (i.e., features from the entire protein such as secondary structure composition, protein length, and fraction of surface residues, and features from a local window of sequence-consecutive residues). The most important local feature was the evolutionary exchange profile reflecting sequence conservation in a family of related proteins. To illustrate its potential, we applied our method to 4 different case studies, each of which related our predictions to aspects of function. The first 2 were the prediction of regions that undergo conformational switches upon environmental changes (switch II region in Ras) and the prediction of surface regions, the rigidity of which is crucial for their function (tunnel in propeller folds). Both were correctly captured by our method. The third study established that residues in active sites of enzymes are predicted by our method to have unexpectedly low B-values. The final study demonstrated how well our predictions correlated with NMR order parameters to reflect motion. Our method had not been set up to address any of the tasks in those 4 case studies. Therefore, we expect that this method will assist in many attempts at inferring aspects of function.**
Proteins 2005;61:115–126.

Key words: flexibility prediction; protein dynamics; protein motion; protein structure prediction; solvent accessibility; multiple alignments; secondary structure prediction; protein function prediction; enzyme active sites; conformational switch

INTRODUCTION

Protein flexibility is related to function. Proteins are dynamic molecules that are in constant motion. The structural flexibility that enables this motion has been associated with various biological processes such as molecular recognition and catalytic activity.[1–21] In fact, even such a coarse-grained aspect of protein structure as the secondary structure assigned from X-ray crystals of proteins captures flexibility relevant for protein function.[22]

Flexible regions can be predicted from sequence. In silico studies have attempted to characterize and predict flexible regions from the amino acid sequence. Different groups used different definitions for flexibility. On a very coarse-grained level, all regions with high net charge and low hydrophobicity were considered to be natively unfolded.[23] The rationale for this assumption is that repulsion from equal charge–charge interactions and the reduced "folding driving force" in regions of low hydrophobicity account for flexibility. Dunker and his group introduced another radical approach that considers all regions with missing coordinates in X-ray structures as "disordered" and applied neural networks to predict such regions.[1,24,25] Other groups have used to same definition to develop related methods to predict such "disorder."[26–28] Our group took a much simpler angle to identify long regions with NORS (i.e., stretches of 70 or more sequence-consecutive residues depleted of helices and strands).[2,29] Analyzing all proteins

Abbreviations: 1D structure, 1-dimensional (e.g., sequence or string of residue secondary structure or numbers for residue solvent accessibility); 3D structure, 3-dimensional structure (i.e., coordinates of all residues/atoms in a protein; B-value/B-factor, "temperature"—or "Debye–Waller"—factor that describes the degree to which the electron density in an X-ray image of a residue is dispersed; B_{norm}, normalized B-value [Eq. (1) for experimental and Eq. (2) for predictions]; DSSP, automatic assignment of secondary structure and solvent accessibility from 3D coordinates; EB, describes strand states according to DSSP assignments; GHI, describes helical states according to DSSP assignments; GTP, guanosine triphosphate; HSSP, describes database of protein structure-sequence alignments; NORS, long regions with no regular secondary structure; PDB, Protein Data Bank of protein structures; PROFacc, profile-based neural network prediction of solvent accessibility; PROFsec, profile-based neural network prediction of secondary structure; RNase H, ribonuclease H.

Grant sponsor: National Institutes of Health; Grant number: RO1-GM63029-01. Grant sponsor: National Library of Medicine; Grant number: R01-LM07329-01.

*Correspondence to: Burkhard Rost, CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, BB217, New York, NY 10032. E-mail: rost@cubic.bioc.columbia.edu

Received 5 January 2005; Revised 2 March 2005; Accepted 4 March 2005

Published online 3 August 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20587

in entirely sequenced organisms (i.e., full proteomes), we found[2] that many more proteins have such regions in full proteomes than in the proteins of known 3D structure deposited in today's PDB,[30,31] that eukaryotes have at least 5 times more such proteins than other organisms, and that NORS regions are over-represented in regulatory and promiscuously interacting proteins. Similar findings were reported based on the Dunker disorder.[27,32]

The Debye–Waller factor (B-value) measures local residue flexibility. B-values (also referred to as B-factors) reported in experimental atomic resolution structures provide information about local mobility. They represent the decrease of intensity in diffraction due to both the dynamic disorder caused by the temperature-dependent vibration of the atoms and the static disorder, which is related to the orientation of the molecule.[33] The B-value is defined by $8\pi^2\langle u^2\rangle$ to the unidirectional mean-square displacement, u^2, averaged over the lattice.[34] B-values of Cα atoms are commonly used to represent motion of the backbone.[35,36] However, the experimentally determined B-value is not an absolute quantity; instead it depends on other factors such as the overall resolution of the structure, crystal contacts, and importantly, on the particular refinement procedures.[37,38] B-values from different structures can therefore not be reasonably compared without some normalization.[37,39,40] Typically, the following normalization is applied[39]:

$$B_{\text{norm}} = (B - \langle B\rangle)/\sigma \quad (1)$$

where σ is the standard deviation and $\langle B\rangle$ is the average of B-values of a given structure. Generally, data of higher resolution provides more reliable B-values. We mainly analyzed structures with resolutions below 2 Å (0.2 nm); however, in order to significantly increase the size of our data set and to cover a larger fraction of sequence space, we also included structures down to 2.5 Å (0.25 nm). At this resolution limit, the correlation between resolution and the log of the mean diffraction intensity is close to linear and B-value assignment is still quite reliable.[34] Furthermore, our results were qualitatively rather similar for data sets with different cutoffs (1.5 Å, 2.0 Å, and 2.5 Å).

Flexibility is informative about protein structure and function. Experimental data about B-values and predictions of flexibility were shown to be useful for predicting residues that cannot be crystallized.[28,41] More importantly, packing density is inversely proportional to thermal motion[42]; therefore, the prediction of flexibility may help to unravel protein function. While our manuscript was in review, a recent study showed that B-values can be predicted from sequence.[43] However, that study contained no direct evidence as to whether or not the functionally important residues were predicted correctly. In addition, Andersen et al.[22] suggested that secondary structure assignments can be continuous. For instance, not all helices are the same; some are "fuzzy" and some are well defined. Therefore, combining flexibility and secondary structure prediction can be a useful tool for predicting these local structures.

Aims of this work. First, we hoped that an analysis of B-values based on a recent comprehensive and unbiased data set of high-resolution structures might unravel correlations between sequence and flexibility that had previously been overlooked. Second, we wanted to develop a method that predicts normalized B-values as accurately as possible from sequence. Third, we wanted to demonstrate that this method could be applied to solving 2 different yet related problems, namely, the prediction of conformational switches and related folds.

METHODS

Data set. All proteins were taken from the PDB.[31] Sequence redundancy was reduced at HSSP-values[44–46] of 0 (corresponding to less than 22% pairwise sequence identity for long alignments). In other words, for no pair of proteins used for training and testing could we predict structural similarity from sequence. Note that we also removed any pair between the training and testing set that could be aligned at PSI-BLAST[47] *E*-values $< 10^{-3}$ according to our standard procedure of 3 automated iterations.[48] Technically, the largest sequence-unique subset was taken from the EVA server,[49,50] which maintains a sequence-unique subset of the PDB that is updated every week. We included only structures with resolutions ≤ 0.25nm (i.e., better than 2.5 Å) and normalized the B-values according to Eq. (1). Our final sequence-unique subset contained 1513 X-ray structures. We split this set into 3 sets: the first used for training (i.e., for optimizing the bulk of all free parameters), the second for validation (i.e., for choosing additional free parameters such as the number of hidden units and the stop of training), and the third for testing. We then repeated this experiment 3 times, such that each protein in our data set was used for testing exactly once. Note that all estimates for performance that we report are valid for the testing set; in particular, we never reported any values that had been subject to any optimization. We argued that no prediction method will ever be as accurate as experiments; therefore, we compiled a data set of 716 unique proteins, each of which had more than 1 experimental structure (i.e., more than 1 answer for what the real B-values are). We considered the difference between these alternative experimental solutions to be the "upper limit" of prediction.

Data set of enzymes. In order to study whether our B-value predictions were correlated somehow with active sites in enzymes, we included 69 high-resolution (≤ 2.5 Å) X-ray structures of apo-enzymes; the set was sequence-unique in the sense that no pair of enzymes had > 25% pairwise percentage sequence identity. Additionally, these structures had *R*-factors below 0.2 and did not contain any disordered regions (except for the termini by "Dunker disorder"). The active sites of these enzymes were taken from the lines annotated with SITE at their corresponding PDB files. If there was no annotation of the active site, the information was obtained from a different structure of the same protein from the PDB. This data set had originally been used to establish the observation that active sites in enzymes are unexpectedly rigid.[5]

Data set of NMR order parameters. NMR spin relaxation spectroscopy experiments are widely used to detect and characterize internal motions in proteins.[8,21,51–53] Specifi-

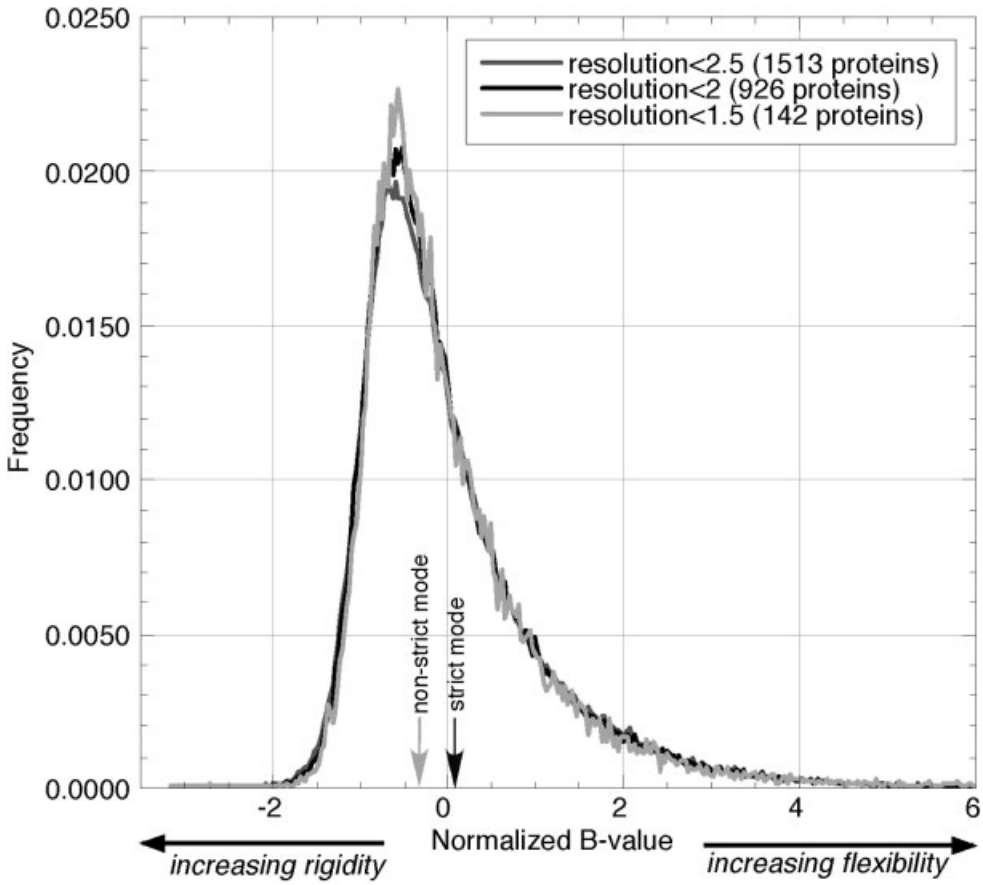

Fig. 1. Distribution of normalized B-values. Normalized B-values compiled from PDB according to Eq. (1). Note that high values (to the right) indicate more flexible residues; lower values (to the left) indicate more rigid residues. The distributions given for 3 different sequence-unique subsets, each of which corresponded to different thresholds in the maximal resolution allowed to include structures (≤0.15 nm, ≤0.2 nm, and ≤0.25 nm) were almost indistinguishable. The arrows mark the 2 different thresholds that we used to classify normalized B-values into 2 states, namely, flexible and rigid. We refer to the classification originating from the left arrow as "nonstrict," and that from the right arrow as "strict."

cally, S^2 is the square of the generalized order parameter that represents motions on the pico- to nanosecond timescales. S^2 values range from 0 to 1; lower values indicate larger amplitudes of internal motions. For this study, we obtained a very refined experimental data set of the order parameter values for *Escherichia coli* RNase H from a study done by the Palmer group.[54] Due to resonance overlap, internal motions and chemical exchange broadening (caused by slower motions), reliable data could not be collected for all residues; the final data set included order parameter values for 81 of the 149 residues in RNase H.[54]

Prediction method. A straightforward back-propagated feed-forward artificial neural network, similar to networks used to predict secondary structure, solvent accessibility, nuclear localization, and protein–protein interaction sites,[55–59] was used to predict residue flexibility. We sampled the "space of network parameters" by testing networks with 5–35 hidden nodes. The input nodes used both local and global information in the following way. Local information: for a window of w sequence-consecutive residues, we used the evolutionary profile (below) for each residue (w = 9, i.e., 9 × 21 input units), the 3-state secondary structure predicted by PROFsec[55,60,61] (w = 5, i.e., 5 × 3 units), and the 2-state solvent accessibility predicted by PROFacc[55,60,61] (w = 5, i.e., 5 × 2 units). In terms of global information, the entire protein was represented by its predicted secondary structure content (3 units), its predicted fraction of surface residues (2 units), and by its length (3 units). To facilitate "learning," a 10-state description that corresponds to different B-values was assigned to the output nodes. The 10-state description reduced the problem of identical samples with similar, but not identical, B_{norm} values. The problem that we tried to address here was that the distribution of normalized B-values has no clear cutoff between "flexible" and "rigid." Instead, most residues have normalized B-values somewhere between the 2 extremes [i.e., the distribution has a peak (Fig. 1)]. Deciding on a cutoff right inside of this peak raises the problem that against the physical reality, the neural networks have to learn that 1 residue with a value b1 is rigid and another with a value b1 + ϵ is flexible. In analogy to our work on predicting solvent accessibility,[57] we tried to reduce this problem by introducing 10 states by portioning the observed distribution with 10 identical

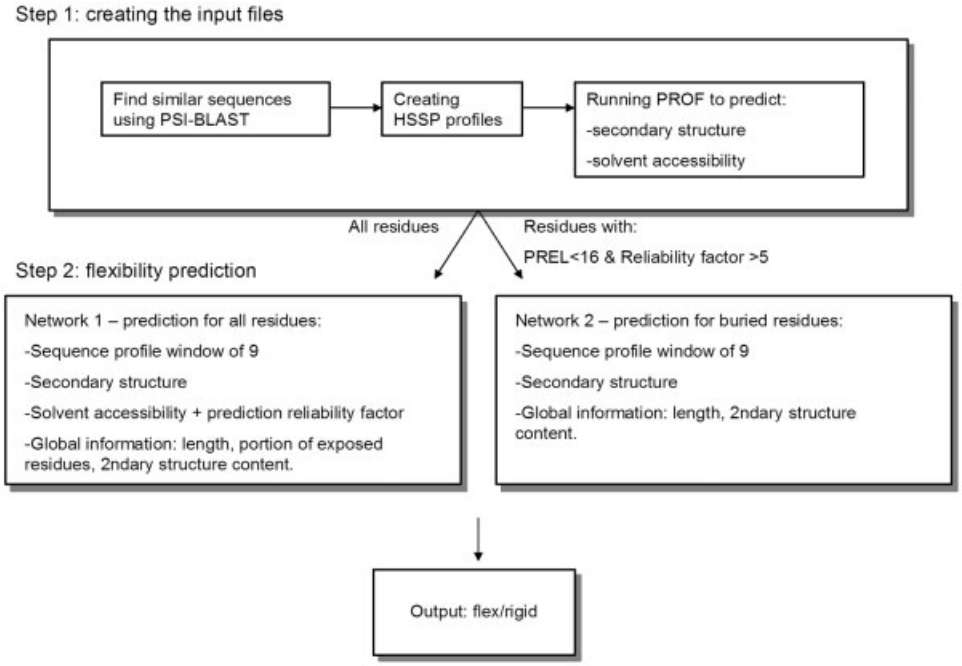

Fig. 2. Prediction system. Step 1: Compile information used for neural network input. HSSP profiles were created using PSI-BLAST; these profiles are used to predict 1D structure (secondary structure and solvent accessibility) by PROFphd. Step 2: System of neural networks. Network 1 was trained on all residues with all input features, while network 2 was trained exclusively on reliably predicted buried residues. Residues that PROFacc predicted as buried with high reliability were passed to network 2; all others to network 1.

intervals. Furthermore, we trained 2 different networks: one on all residues, the other trained on the subset of residues predicted a high reliability by PROFacc (cutoff = 6 in Fig. 2).

Evolutionary information. We obtained multiple sequence alignments by searching with PSI-BLAST[47] against all known sequences contained in SWISS-PROT,[62] TrEMBL,[62] and PDB.[30,31] All hits below a PSI-BLAST *E*-value of 10^{-3} were subsequently filtered[44,48] and included in the sequence profile. The profiles were further filtered to remove sequences that were extremely similar to each other by simply removing all proteins with levels of pairwise sequence identity $> 80\%$ to any previously added sequence (this value was chosen by intuition instead of by optimization). This was done in order to maximize the number of aligned sequences that are not nearly identical to the query sequence and therefore obtain evolutionary information from more distant sequences. Note that while we know that this level of sequence similarity implies that all proteins included in the profile have similar structures, the thresholds did not guarantee that all members of the sequence-structure family also had similar B-values.

Conversion of neural network output into values for predicted normalized B-values. The raw neural networks have 2 output values: one coding for flexible residues, and the other for rigid residues. Mostly, we evaluated the accuracy in the prediction of the extremes (i.e., flexible or rigid residues). However, our predictions contained information beyond this. In order to convert the raw output into values similar to the experimental normalized B-values [Eq. (1)], we simply applied this formula to the difference between the two output units:

$$B_{\text{norm}}^{\text{predicted}} = \frac{(\Delta - \langle\Delta\rangle)}{\sigma}, \text{ with } \Delta = out_1 - out_2, \qquad (2)$$

where out_1 and out_2 were the values for the raw network output for the units coding for flexible and rigid, respectively; $\langle\Delta\rangle$ was the average overall residue predictions in 1 protein, and σ was the standard deviation of the corresponding distribution of predicted Δ.

Evaluating the results. In order to simplify the prediction task from one of continuous values (normalized B-values range from -3.13 to 12.46) to a 2-state problem (flexible–rigid), we defined a residue to be flexible according to the following thresholds: (1) Strict: if the normalized B_{norm} value ≥ 0.03, and (2) Nonstrict: if the normalized $B_{\text{norm}} \geq -0.3$. All residues in between these 2 extreme states were ignored for both training and testing. This training method turned out to be suboptimal; the best performance was obtained using balanced training[56] [i.e., the neural network "saw" the samples from the extremes about twice as often as the samples from the peak in the distribution (Fig. 1)]. Note that for testing we always used all values (residues at the extremes were predicted much more accurately). In the nonstrict threshold, most of the residues are flexible; therefore, if we find a rigid residue on the surface, it is likely to have a functional role. On the other hand, in the strict mode, only about a third of the residues are flexible; therefore, if we find a flexible stretch of residues, this stretch might be functionally important. We evaluated our results by calculating the accuracy/specificity and coverage/sensitivity according to

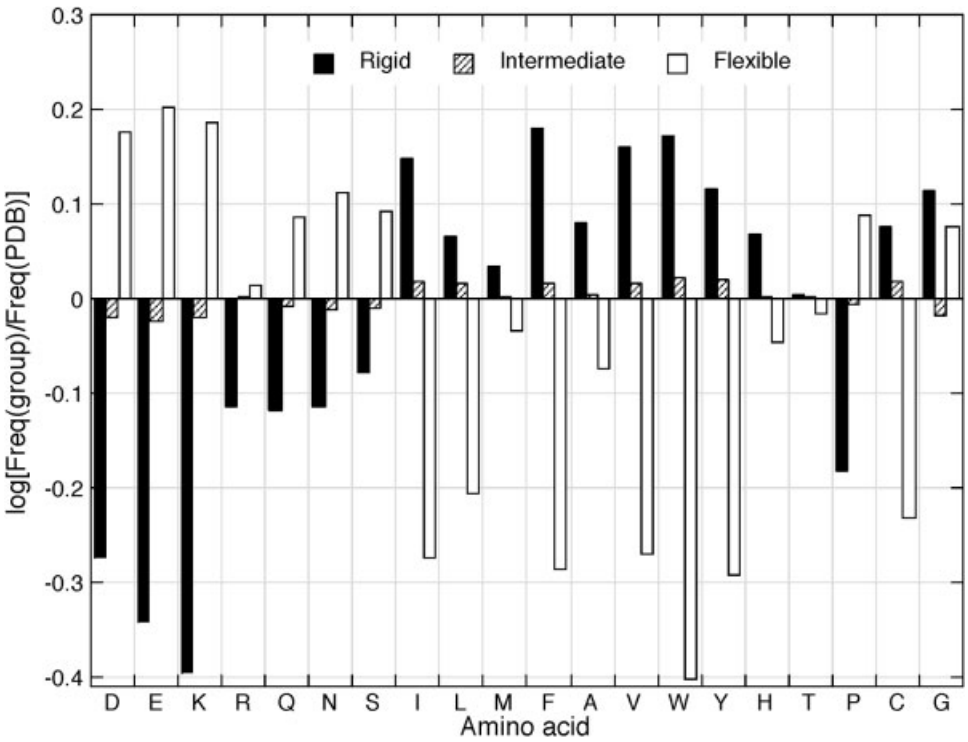

Fig. 3. Amino acid preferences of for rigid and flexible residues. Are particular amino acids more frequently flexible, or rigid than all others? Here, we split our data into 3 states: rigid, intermediate, and flexible, chosen according to the normalized B-values (Fig. 1). Amino acids are given by their 1-letter code; they are roughly grouped by their biophysical features. The propensities of amino acids found in a nonredundant set of 1513 X-ray structures from the PDB were used as background (i.e., to compile the deviation of the observed distribution from the background). The sign of the bar corresponds to over- (positive) or under-representation (negative) of amino acid in a group (rigid, intermediate, and flexible); the 0 level describes counts that did not deviate from the background. This is particularly common for residues that were neither rigid nor flexible (intermediate in stippled bars). In general, charged and hydrophilic residues are over-represented in the flexible group (open bars, positive) and under-represented in the rigid group (black bars, negative), while residues with hydrophobic and bulky side-chains follow the opposite trend. Cysteine, proline, and glycine clearly stand out from these trends.

$$ACC = \frac{TP}{TP + FP} \quad COV = \frac{TP}{TP + FN}, \tag{3}$$

where *TP* (true positives) is the number of residues correctly predicted to be flexible, *TN* (true negatives) is the number of residues correctly predicted not to be flexible, *FP* (false positives) is the number of residues predicted to be flexible and observed to be rigid, and *FN* (false negatives) is the number of residues predicted to be rigid and observed to be flexible. We combined these 2 values through their harmonic mean:

$$F = \frac{2 \cdot \mathrm{ACC} \cdot \mathrm{COV}}{\mathrm{ACC} + \mathrm{COV}}. \tag{4}$$

RESULTS AND DISCUSSION

Large-Scale Analysis of Normalized B-Values

Different amino acids have preferences for different B-values. Previous studies have shown that hydrophobic residues, which are usually buried, tend to be more rigid whereas charged residues tend to be more flexible.[23,35,36,41] Here, we performed a large-scale analysis of residues with different B_{norm} values. In order to prove the significance of our findings, we used the find-self test that is applicable to related entities such as proteins.[63] This test basically measures the degree of consistency between data sets with different labels (here: flexible, rigid, other). As expected, prolines are significantly over-represented in regions of high flexibility (Fig. 3), while cysteines are over-represented in nonflexible regions (supposedly due to forming rigid disulfide bridges). Previous studies confirmed the intuition that glycines are flexible.[35,41] In contrast, we find in more detail that glycines are over-represented in both extremes [i.e., very flexible and very rigid regions (Fig. 3)]. Supposedly, this preference is explained by hydrophobic glycines embedded into the protein cores. This is not unexpected, because as the only amino acid without side-chains, glycines can adopt very unusual conformations that may be both rather flexible and rigid, and that are often structurally important.[64,65]

Secondary structure and B_{norm} values correlate. It is well known that long loops tend to be more flexible than regular secondary structures such as helices and strands. Here, we verified this difference in a quantitative way on a large set of proteins (Fig. 4). Interestingly, residue mobility is correlated in a different way with β-strand-related structures (states E and B in DSSP[66]) than α-helix-related

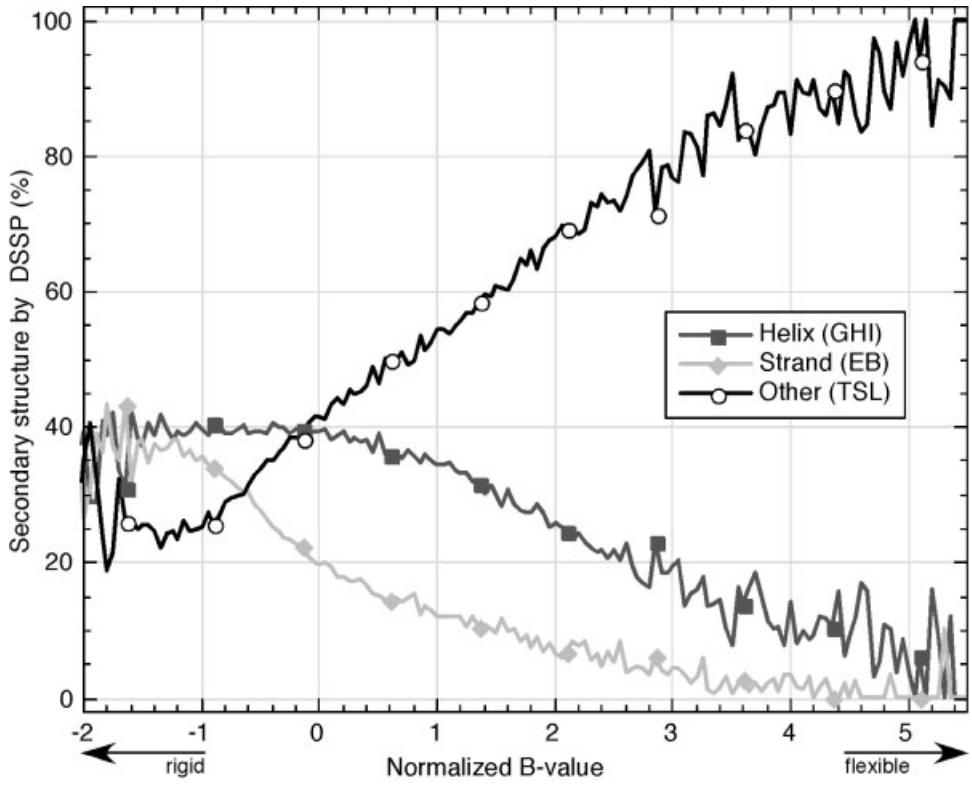

Fig. 4. Correlation between secondary structure state and normalized B-values. The normalized B-values are plotted against secondary structure assignments from DSSP[46] with the following conversions: DSSP state GHI → helix, DSSP state EB → strand, all other DSSP states to other. Each point represents the percentage of secondary structure assignments for a given bin in normalized B-values. As expected, residues in regular secondary structure (helix and strand) tend to occupy regions of lower normalized B-values, while residues in nonregular secondary structure occupy the bins with higher normalized B-values. However, the functions are significantly different for helices and strands.

structures (states H, G, and I in DSSP[66]; Fig. 4). In particular, the following observations were remarkable. First, strands were, on average, more rigid than helices: While both had similar percentages for B-values < -1 (most rigid, Fig. 4), only 20% of all residues are in strands, while about 30% are in helices (i.e., the same overall fraction for very rigid residues over-represented residues in strands). Second, the transition from rigid to flexible was very different for helix (smooth) and strand (almost sigmoid). In other words, very few strands are partially flexible. This may originate from the simple fact that strands are only half as long as helices on average (i.e., breaking a few hydrogen bonds is statistically more likely to break strands than helices). It may also originate from the particular DSSP definition that we used (Fig. 4): Two overlapping helices that fulfill the minimum hydrogen bond requirements can overlap forming a longer helix that is missing hydrogen bonds.[66] The correlation of irregular helices with normalized B-values might be different from the correlation of perfect helices. Finally, our observation might also be explained by the fact that strands are more often found in the protein core than helices.

Prediction of B-Values

Upper boundary for predictions of extreme B-values flexibility: 77–83%. B-values are influenced by crystal contacts,[38] by the experimental resolution,[34,67] by interactions with ligands,[68] by packing density,[42] and by intrinsic properties of the crystal that result from particular refinement methods.[37] Arguably then, the upper boundary on methods that predict B-values is given by the ranges of differences in B-values between different experimental results for the structures of the same proteins. We compared 716 such pairs with resolutions better than 2.5 Å and found that they agreed for 77–83% of their residues in the assignment of extreme B-values (flexible–rigid). The interval for this upper boundary is explained by using different thresholds to decide which residues are flexible, in particular, 0.03 yielded 77%, while −0.3 yielded 83%.

Lower boundary provided by predicted solvent accessibility: 46–65%. Trivially, solvent accessibility is correlated with flexibility[38,39] (i.e., flexible residues are found more often on the surface than in the core of proteins). Surprisingly, no group that developed methods to predict flexibility compared their results to simple accessibility predictions. Given the correlation between accessibility and flexibility, and the accuracy in predicting accessibility, we could consider predictions of solvent accessibility to constitute a lower threshold for predicting flexibility. We found that relative solvent accessibility predicted by PROFacc correlates rather well with normalized B-values (Table I). When we used solvent accessibility prediction information

TABLE I. Prediction of Residue Flexibility on a Sequence-Unique Set

	Nonstrict[a]			Strict[a]		
Prediction method[b]	ACC[c]	COV[c]	F[d]	ACC[c]	COV[c]	F[d]
2 experiments	83.0 ± 0.18	83.9 ± 0.18	83.4 ± 0.14	77.5 ± 0.24	78.2 ± 0.23	77.9 ± 0.19
PROFacc	69.7 ± 0.32	61.3 ± 0.34	65.3 ± 0.27	62.9 ± 0.55	36.7 ± 0.40	46.4 ± 0.41
Sequence + 1D	70.0 ± 0.18	66.4 ± 0.19	68.1 ± 0.15	63.1 ± 0.32	37.5 ± 0.24	47.0 ± 0.25
Profile + 1D	70.0 ± 0.17	73.3 ± 0.17	71.5 ± 0.14	63.0 ± 0.25	45.7 ± 0.18	52.9 ± 0.20
PROFboval	70.1 ± 0.18	73.9 ± 0.17	71.9 ± 0.14	63.1 ± 0.25	46.2 ± 0.18	53.3 ± 0.20

[a]Nonstrict and strict refer to different thresholds in the classification of normalized B- values into the 2 classes, flexible and rigid (arrows in Fig. 1).
[b]Methods: *2 experiments* marked the agreement between proteins for which B-factors were determined by more than 1 experiment (*Note.* in this case, we consider one of the experiments as the "truth," the other as the prediction); *PROFacc* marked the success in simply considering all residues that are predicted as solvent accessible (by PROFacc[57]) as flexible; *Sequence + 1D* marked a simple neural network that uses only single sequences and all the predicted 1D structure and sequence features described in methods; *Profile + 1D* marks a network that uses profiles instead of single sequence, otherwise as "Sequence + 1D"; and PROFbval marked our final prediction system using 2 layers of networks and profiles (see Methods section).
[c]Accuracy (ACC) and Coverage (COV) as defined in Eq. (3); note that the ± values mark the standard errors compiled over our data set.
[d]F-measure as defined in Eq. (4).

only, we achieved almost 63% accuracy at over 37% coverage in the strict mode (corresponding to $F \sim 46\%$, Table I). In fact, when analyzing the results from networks that explicitly used predicted solvent accessibility as input in detail, we found that these networks basically predicted all buried residues to be rigid, and all exposed residues to be flexible. Additionally, we found that extreme incorrect predictions for flexibility often resulted from mistakes in the accessibility predictions. We reduced the seriousness of the second problem by including the reliability of the accessibility prediction as explicit input units. We addressed the first problem by training a separate set of networks trained only on reliably predicted buried residues (Fig. 2). The rationale for this solution is that the factors that determine a buried residue to be flexible might differ from those that determine the flexibility of an exposed residue. For example, a buried charged residue usually will be in a salt bridge and adopt a rigid conformation, while when exposed, it forms contacts with the water and is more flexible.

Sequence + 1D structure improves marginally over accessibility-based predictions. The local features from sequence and predicted 1D structure that we found to be correlated with flexibility, together with global information, improved the performance of our neural network-based prediction method by 3 percentage points (F-measure in nonstrict mode, Table I) over predictions that use only predicted solvent accessibility to predict flexibility. However, evaluated in the strict mode performance was almost identical.

Evolutionary information significantly improves prediction. The use of alignment profiles instead of raw sequence

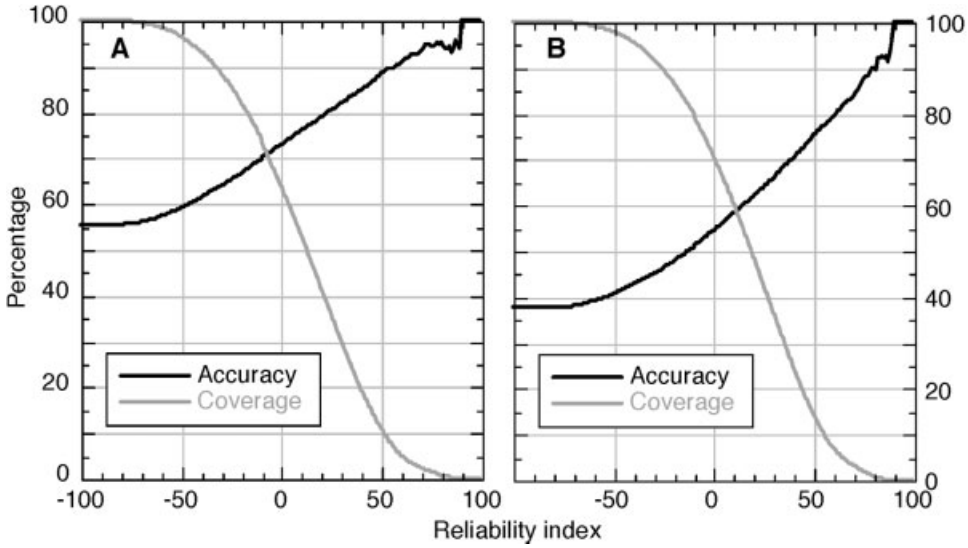

Fig. 5. Reliability of B-value predictions. The raw output from the neural networks was converted into a reliability index that reflected the strength of the prediction. In particular, we defined the reliability index as the difference between the output unit coding for flexible and that coding for rigid residues. The 2 panels distinguished between the 2 different modes of converting B-values into the 2 states, flexible–rigid [arrows in Fig. 1; (**A**) nonstrict mode, (**B**) strict mode]. Note that for any accuracy, the coverage is much higher for the nonstrict than for the strict mode.

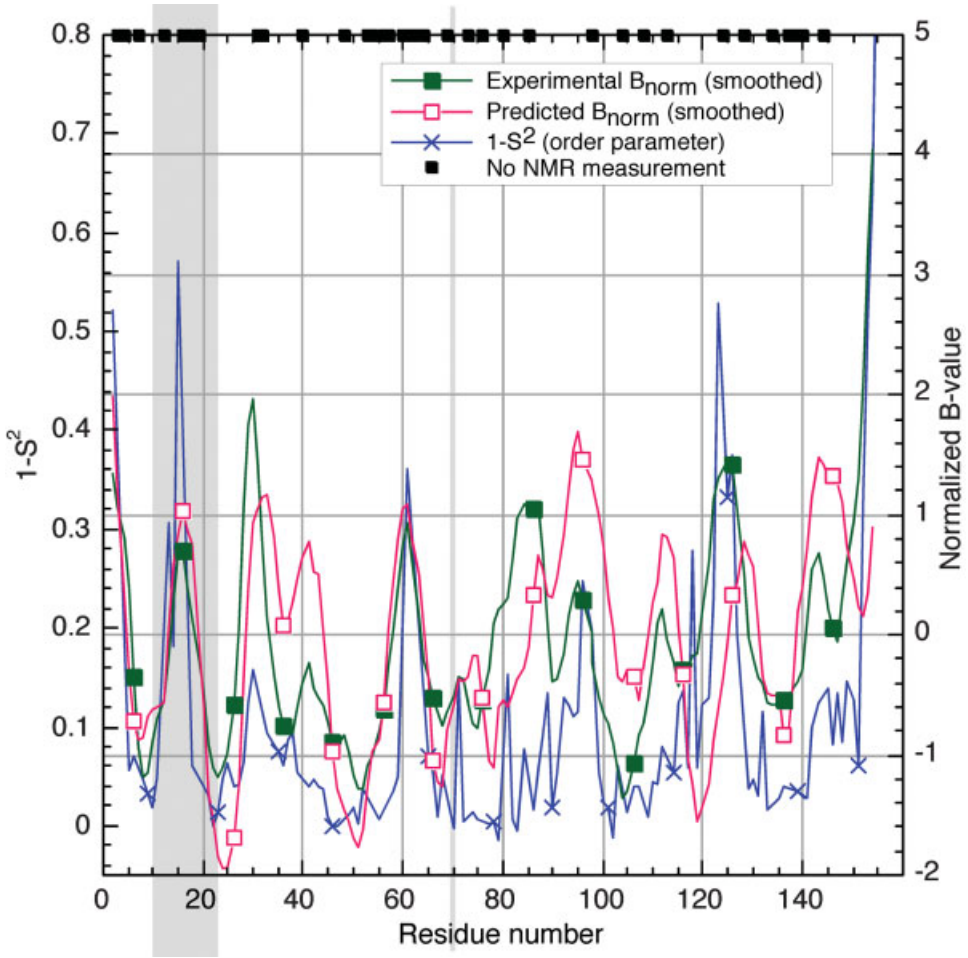

Figure 6.

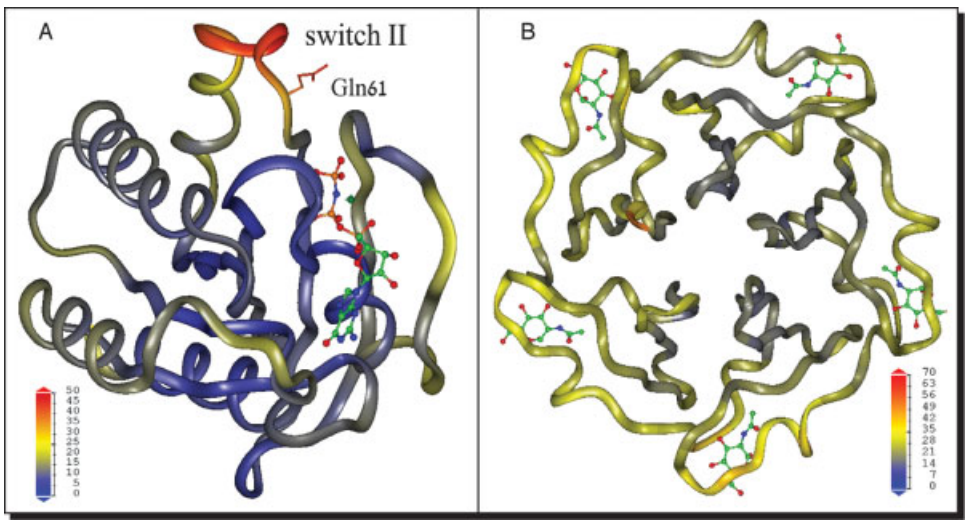

Figure 7.

significantly improves predictions of secondary structure,[55,56,60] solvent accessibility,[57] and nuclear localization.[59] To improve our predictions, we used profiles using alignments produced by PSI-BLAST[47] (Fig. 2). These profiles, along with all other features, significantly improved performance in both nonstrict and strict modes (Table I). To put these value into perspective: The upper boundary (different experiments) was $F = 83$ and $F = 78$ (nonstrict and strict, respectively); the lower boundary (prediction based on solvent accessibility) was $F = 65$ and $F = 46$, respectively. The difference between these 2 extremes was $\Delta F = 18$ (83–65, nonstrict) and $\Delta F = 32$ (78–46, strict). On this scale, our prediction method (PROFbval in Table I) covered about 7 percentage points.

Very high accuracy for most strongly predicted residues. The strength of our final prediction system (compiled as the difference between the two output units) correlated very well with the reliability of the prediction (Fig. 5). This feature will enable users to focus on more reliably predicted regions. For example, users can focus on the subset of all residues predicted at 90% accuracy; these corresponded to about 10% of all residues in the nonstrict and to about 1% of all residues in the strict mode.

Correlation between observed and predicted normalized B-values. Eq. (2) describes a simple way of transferring the output of our final network into "real-valued" predictions for B-values. The Pearson correlation coefficient between these predicted and the experimentally observed normalized B-values over all the proteins in our test sets was 0.44. The best prediction method published by Yuan et al. was reported to reach a correlation of 0.53.[43] On the same data set, our method reached an unusually high level of 0.50. This was still slightly below the method from Yuan et al. that, in contrast to our method, had been optimized on that data set and on the objective of optimizing the correlation coefficient rather than the binary accuracy (as our method had). The results in Figure 6 illustrate how high a correlation above 0.4 is.

Application to Problems Relating Function and Flexibility–Rigidity

Case study I: Switch II region in Ras. Structure and function are closely linked (i.e., the 3D structure of a protein determines its function). However, for some biological functions, the ability of a protein or a region within a protein not to adopt a rigid structure, but instead to be in motion, is crucial.[4,11–15,17] Switch II regions in the Ras protein are known to be critical for the GTPase activity of Ras; this region is defined by 11 consecutive residues of the following sequence GQEEYSAMRDQ (residues 60 to 70 in SWISS-PROT file RASH_HUMAN). When bound to GTP, the switch II region is rigid; upon GTP hydrolysis, the switch II region becomes flexible. The key residue in this 9-residue segment is Gln61 which is the catalytic residue of the reaction.[71,72] A recent study showed that the "hyperflexibility" of this conserved residue is critical for the GTPase reaction.[6] Although the whole switch II region is exposed to solvent, our prediction method singled out the crucial residue Gln61 and its 3 sequence neighbors to be the most flexible residues; this prediction is confirmed by experimental results [Fig. 7(A)]. We did not select Ras because our method worked for this protein; instead, we chose it because it is a common oncogene that has been studied in great detail. In particular, the biochemical significance of the Ras flexibility has been proven not just by crystallographic B-values. Methods such as heteronuclear NMR,[74] time-resolved crystallography,[75] molecular dynamics simulations,[76–79] and the use of fluorescent GTP analogue[80] have shown how important the flexibility of the switch II region is for the function of Ras.

Case study II: Identification of functionally important rigid regions in propeller folds. β-propeller folds [Fig. 7(B)] are characterized by 4–8 symmetrical repeats of 4-stranded antiparallel and twisted β-sheets that are arranged around a central "tunnel."[81,82] The majority of these proteins use the tunnel or the entrance to it for the coordination of a ligand or as the site of catalytic activity. The structural rigidity of the propeller domain has been suggested as crucial for the function of these proteins.[82] A combination of repeat detection, secondary structure prediction, and fold recognition can be exploited to predict β-propeller folds from sequence.[81,82] When we applied our prediction method to propeller folds, we correctly detected the regions around the tunnel as rigid, although this region is exposed (and correctly predicted as such). Again, we did not choose the propeller folds because our method worked on these—in fact, we only looked at the 2 cases of Ras and propeller folds in more detail; instead this class of proteins is currently under investigation by our colleagues in the Wayne Hendrickson group (Columbia). However, this choice is also rationalized by the fact that a PubMed search with the keywords "protein structural rigidity" brought propeller folds up as the first relevant hit.

Case study III: Active sites in enzymes predicted as more rigid. Over the past decades the number of enzyme structures in the PDB has increased significantly. These data resulted in several attempts at the characterization of residues that participate in the catalytic reactions (i.e., the active site residues).[5,83–86] Specifically, 2 groups showed that active site residues have, on average, lower normalized B-values than do non-active-site residues and are therefore more rigid.[5,86] Yuan et al.[5] investigated the normalized B-value differences between active- and non-

Fig. 6. Predicted and observed flexibility in RNase H. The residues from RNase H are numbered consecutively on the *x* axis (1…149). The **y** axis on the right shows normalized B-values from the experimental X-ray structure [2RN2,[69] green; Eq. (1)] and from our predictions [pink; Eq. (2)]. The *y* axis on the left superimposes the order parameters from NMR measurements (blue). Three major points stood out: (1) Experimental (green) and predicted (pink) B-values correlated very well (note that these values constituted the average performance of our method). (2) The NMR order parameters (blue) and the B-values (green) correlated somehow but clearly less than did observed (green) and predicted (pink) B-values. (3) The correlation between NMR order parameters and B-values was much higher than average for the functionally important residues (residues 11–23, 10, 48, and 70).[69,70]

Fig. 7. Our flexibility method can identify conformational switches and related folds. Colors: the more red, the higher the B-value (flexible); the more blue, the lower (rigid); the GTP analogue is colored in CPK. (**A**) PROFbval correctly identified the key residues in the switch II region of ras as highly flexible. (**B**) PROFbval also correctly predicted the residues in the tunnel of a β-propeller fold to be rigid, although these residues are surface loops and strands (structure from Beisel et al.[73]).

TABLE II. Prediction of Residue Flexibility on a Sequence-Unique Set of Enzymes

	Average normalized B-values[a]			Average predicted normalized B-values[b]		
Structure subset[c]	All residues	REL ≥ 5%	REL ≥ 16%	All residues	PREL ≥ 5%	PREL ≥ 16%
Active-site residues	−0.39	−0.21	−0.04	−0.48	0.18	0.09
Non-active-site residues	0.00	0.23	0.35	0.01	0.36	0.39

[a]Normalized B-values were calculated using Eq. (1) and then averaged over the residues from all 69 enzyme structures used in a previous study[5]: REL marked the relative solvent accessibility from X-ray structures (in percentage as compiled from DSSP[66] using the normalization with maximal values observed in isolation.[57]
[b]Predicted normalized B-values were calculated through Eq. (2); here, PREL marked the relative solvent accessibility predicted by PROFacc.[57]
[c]Data set of enzymes: *Active-site residues* marked the residues from the lines annotated with SITE in the PDB files. If there was no annotation of the active site, the information was obtained from a different structure of identical protein from the PDB; *Non-active- site residues* marked all the residues that are not annotated on the SITE line.

active-site residues in 69 sequence nonidentical enzyme apo-structures. They confirmed that active site residues are less flexible than non-active-site residues. Using the same data set as Yuan et al. (Table II), we compared active- and non-active-site residues according to the observed normalized B-values (as pioneered by Yuan et al.) as well as by our predicted B-values [Eq. (2), Table II). We found that the difference "normalized B-values of active site residues–normalized B-values of non-active-site residues" was similar for the observed and predicted B-values (−0.39 and −0.48, respectively). In other words, our method correctly predicted that the active site residues were considerably more rigid than all other residues. Interestingly, our method also correctly solved a more difficult task, namely, the correct prediction that also the exposed active site residues are more rigid than the non-active-site surface residues (Table II).

Case study IV: Predicted B-values correlate with NMR order parameters. Solution NMR spectroscopy methods are commonly used to characterize the dynamic properties of proteins.[21,51,52] It has been previously shown that order parameter data obtained from NMR spin relaxation experiments are, to an extent, correlated with experimental B-value data.[87,88] However, unlike B-values from X-ray structures, order parameter data are independent of crystal packing; they probe dynamics on timescales that are relevant for biological function.[88] Here, we compared (15)N nuclear magnetic spin relaxation order parameters of ribonuclease (RNase HI) to experimental normalized B-values taken from the apo X-ray structure (PDB identifier 2RN2[69]) and to the normalized output from our network of the same protein [Eq. (2), Fig. 6). This example also illustrated more explicitly to what extent our predictions correlated with the experimental values. Note that RNase H was representative of the overall performance. The correlation between experimental and predicted B-values was higher than that between experimental B-values and order parameters, and between predicted B-values and order parameters (Fig. 6). However, the correlations were clearly higher for the functionally most important residues. These include the segment of residues 11–23 that is known to bind a DNA–RNA hybrid and the 3 active site residues (D10, E48, D70) that bind Mg^{2+} ions.[69,70] Interestingly, the active site residues are harder to predict to be rigid due to the fact that they are partly accessible to solvent. In fact, when ranking all the residues that are predicted to be on the surface (≥ 5%) and by the strength of our prediction of rigidity, E48 was ranked first. This was probably due to the fact that we used evolutionary conservation as an input to the neural network, since E48 is very conserved. Our prediction method "misplaced" 1 peak, namely, that between residues 120–130 that was shifted by 5 residues in our prediction. This was probably due to an incorrect prediction for accessibility: W118 and W120 were incorrectly predicted as completely buried (note that our prediction method used the predicted solvent accessibility of the residue and its neighbors as input).

How well can we predict flexibility? Our final flexibility predictions achieved F values [Eq. (4)] between 53% (strict) and 72% (nonstrict). The addition of all the attributes that were found to be correlated with high B-values such as predicted solvent accessibility, secondary structure, protein length, the use of multiple sequence alignment as an input, together with our unique training method, made the final method, PROFbval, become a rather complex new prediction method; the orchestration of all features was required to make the method become significantly more accurate than predictions based on predicted accessibility. The performance was clearly worse than the agreement between different experimental determinations of B-values. In the absence of comparable methods, we cannot conclude whether or not the performance of PROFbval will suffice to make the method become an important milestone on the way toward predicting protein function. However, our 4 case studies suggested that the method could contribute important information that was not available through other means. Combining our predictions with data (e.g., from NMR experiments) could substantially increase the fraction of the sequence space covered and, hence, the performance of the prediction system.

CONCLUSIONS

Recent studies have shown that protein flexibility and protein function are strongly linked.[1,2,4,9,11–15,17,19,27,89] Numerous proteins have regions that adopt different conformation under different conditions, allowing them to take part in cellular and molecular regulation.[2] In this study, we first showed that flexible residues (i.e., residues with high normalized B-values) differ from regular and rigid residues in local features such as secondary struc-

ture, solvent accessibility, and amino acid preferences. For our analysis of B-values, we used the largest unbiased data set that had so far been explored. Possibly due to this representative data set, the only results that could be phrased in the form of simple rules—such as glycines populate the extremes of very high and very low B-values—were not very surprising. Interestingly, we showed that local sequence features alone did not suffice to develop a good prediction method. Instead, global features and evolutionary profiles significantly improved performance. We improved further by adding another neural network, specialized for reliably predicted buried residues. Last, we presented 2 applications of this method, namely, to the flexible switch II region in Ras and to the rigid region in propeller folds. Both the flexible region in Ras and the rigid region in propeller folds are indicative of function and were correctly predicted by our method. Lacking a large data set of such examples, we cannot guarantee that these 2 case studies are fully representative. However, the 2 cases were not chosen because "they worked" but because of the readily available experimental data for both. Therefore, we hypothesized that the flexibility–rigidity prediction method, together with other methods, can serve both as a tool to identify functional residues in protein and to identify specific folds. Our results for a large set of enzymes and for the order parameter measurements of RNAse H added evidence to strengthen this hypothesis.

ACKNOWLEDGMENTS

Thanks to Jinfeng Liu and Megan Restuccia (both at Columbia) for computer assistance, and to Sven Mika, Andrew Kernytsky, and Dariusz Przybylski (all at Columbia) for providing preliminary information and programs; particular thanks go to Mickey Kosloff, Marco Punta, and Yanay Ofran (all at Columbia) for very valuable suggestions. Particular thanks also go to Art Palmer and Joel Butterwick (both at Columbia) for providing the data set of order parameters for RNase H. Thanks to both anonymous reviewers for their helpful suggestions. Last but not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

REFERENCES

1. Dunker AK, Obradovic Z. The protein trinity-linking function and disorder. Nat Biotechnol 2001;19:805–806.
2. Liu J, Tan H, Rost B. Loopy proteins appear conserved in evolution. J Mol Biol 2002;322:53—64.
3. Tainer JA, Getzoff ED, Alexander H, Houghten RA, Olson AJ, Lerner RA, Hendrickson WA. The reactivity of anti-peptide antibodies is a function of the atomic mobility of sites in a protein. Nature 1984;312:127–134.
4. Wright PE, Dyson HJ. Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. J Mol Biol 1999;293:321–331.
5. Yuan Z, Zhao J, Wang ZX. Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng 2003;16:109–114.
6. Kosloff M, Selinger Z. GTPase catalysis by Ras and other G-proteins: insights from substrate directed superimposition. J Mol Biol 2003;331:1157–1170.
7. Carr PA, Erickson HP, Palmer AG III. Backbone dynamics of homologous fibronectin type III cell adhesion domains from fibronectin and tenascin. Structure 1997;5:949–959.
8. Akke M, Liu J, Cavanagh J, Erickson HP, Palmer AG III. Pervasive conformational fluctuations on microsecond time scales in a fibronectin type III domain. Nat Struct Biol 1998;5:55–59.
9. Teague SJ. Implications of protein flexibility for drug discovery. Nat Rev Drug Discov 2003;2:527—541.
10. Daniel RM, Dunn RV, Finney JL, Smith JC. The role of dynamics in enzyme activity. Annu Rev Biophys Biomol Struct 2003;32:69–92.
11. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573–6582.
12. Dunker AK, Brown CJ, Obradovic Z. Identification and functions of usefully disordered proteins. Adv Protein Chem 2002;62:25–49.
13. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 2002;323:573–584.
14. Demchenko AP. Recognition between flexible protein molecules: induced and assisted folding. J Mol Recognit 2001;14:42–61.
15. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–533.
16. Uversky VN. What does it mean to be natively unfolded? Eur J Biochem 2002;269:2–12.
17. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. Protein Sci 2002;11:739–756.
18. Uversky VN. Protein folding revisited: a polypeptide chain at the folding–misfolding–nonfolding cross-roads: which way to go? Cell Mol Life Sci 2003;60:1852–1871.
19. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 2002;12:54–60.
20. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
21. Dyson HJ, Wright PE. Unfolded proteins and protein folding studied by NMR. Chem Rev 2004;104:3607–3622.
22. Andersen CAF, Palmer AG, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. Structure 2002;10:175–185.
23. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 2000;41:415–427.
24. Romero P, Obradovic Z, Dunker AK. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Lett 1999;462:363–367.
25. Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. Predicting intrinsic disorder from amino acid sequence. Proteins 2003;53(Suppl 6):566–572.
26. Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. Proteins 2003;53(Suppl 6):573–578.
27. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337:635–645.
28. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure 2003;11:1453–1459.
29. Liu J, Rost B. NORSp: predictions of long regions without regular secondary structure. Nucleic Acids Res 2003;31:3833–3835.
30. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein DataBank: a computer based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
31. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 2002;58:899–907.
32. Romero P, Obradovic Z, Kissinger C, Villafranca JE, Garner E, Guilliot S, Dunker AK. Thousands of proteins likely to have long disordered regions. Pac Symp Biocomput 1998;3:437–448.
33. Creighton T. Proteins: structures and molecular properties. New York: W. H. Freeman; 1993.
34. Blow D. Outline of crystallography for biologists. New York: Oxford University Press; 2002. 237 p.
35. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins 1994;19:141–149.
36. Karplus PA, Schultz GE. Prediction of chain flexibility of peptide antigens. Naturwissenchaften 1985;72:212–213.
37. Tronrud DE. Knowledge-based B-factor restraints for the refinement of proteins. J Appl Crystallogr 1996;29:100–104.
38. Sheriff S, Hendrickson WA, Stenkamp RE, Sieker LC, Jensen LH.

Influence of solvent accessibility and intermolecular contacts on atomic mobilities in hemerythrins. Proc Natl Acad Sci USA 1985;82:1104–1107.
39. Carugo O, Argos P. Correlation between side chain mobility and conformation in protein structures. Protein Eng 1997;10:777–787.
40. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci 2003;12: 1060–1072.
41. Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. Protein flexibility and intrinsic disorder. Protein Sci 2004;13:71–80.
42. Halle B. Flexibility and packing in proteins. Proc Natl Acad Sci USA 2002;99:1274–1279.
43. Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. Proteins 2005;58:905–912.
44. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
45. Mika S, Rost B. UniqueProt: creating representative protein sequence sets. Nucleic Acids Res 2003;31:3789–3791.
46. Sander C, Schneider R. Database of homology-derived structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
47. Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
48. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. Proteins 2002;46:195–205.
49. Koh IYY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Narayanan E, Graña O, Valencia A, Sali A, Rost B. EVA: evaluation of protein structure prediction servers. Nucleic Acids Res 2003;31:3311–3315.
50. Eyrich V, Martí-Renom MA, Przybylski D, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001;17: 1242–1243.
51. Palmer AG III, Kroenke CD, Loria JP. Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. Methods Enzymol 2001;339: 204–238.
52. Palmer AG III. NMR characterization of the dynamics of biomacromolecules. Chem Rev 2004;104:3623–3640.
53. Palmer AG III. NMR probes of molecular dynamics: overview and comparison with other techniques. Annu Rev Biophys Biomol Struct 2001;30:129–155.
54. Kroenke CD, Rance M, Palmer AG III. Variability of the 15N chemical shift anisotropy in *Escherichia coli* ribonuclease H in solution. J Am Chem Soc 1999;121:10119–10125.
55. Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. Meth Enzymol 1996;266:525–539.
56. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
57. Rost B. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
58. Ofran Y, Rost B. Predicted protein–protein interaction sites from local sequence information. FEBS Lett 2003;544:236–239.
59. Nair R, Rost B. Better prediction of sub-cellular localization through evolution and structure. Proteins 2003;53:917–930.
60. Rost B. How to use protein 1D structure predicted by PROFphd. In: Walker JE, editor. The proteomics protocols handbook: methods in molecular biology. Totowa, NJ: Humana; 2005. p 875–901.
61. Rost B. Protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–218.
62. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
63. Ofran Y, Rost B. Analysing six types of protein–protein interfaces. J Mol Biol 2003;325:377–387.
64. Brändén C, Tooze J. Introduction to protein structure. New York/London: Garland; 1991.
65. Lesk AM. Introduction to protein architecture—the structural biology of proteins. Oxford, UK: Oxford University Press; 2004.
66. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;12:2577–2637.
67. Drenth J. Principles of protein X-ray crystallography. New York: Springer-Verlag; 1999. 342 p.
68. Carugo O, Argos P. Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. Proteins 1998;31:201–213.
69. Katayanagi K, Miyagawa M, Matsushima M, Ishikawa M, Kanaya S, Nakamura H, Ikehara M, Matsuzaki T, Morikawa K. Structural details of ribonuclease H from *Escherichia coli* as refined to an atomic resolution. J Mol Biol 1992;223:1029–1052.
70. Goedken ER, Keck JL, Berger JM, Marqusee S. Divalent metal cofactor binding in the kinetic folding trajectory of *Escherichia coli* ribonuclease HI. Protein Sci 2000;9:1914–1921.
71. Vetter IR, Wittinghofer A. The guanine nucleotide-binding switch in three dimensions. Science 2001;294:1299–1304.
72. Sprang SR. G proteins, effectors and GAPs: structure and mechanism. Curr Opin Struct Biol 1997;7:849–856.
73. Beisel HG, Kawabata S, Iwanaga S, Huber R, Bode W. Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab *Tachypleus tridentatus*. EMBO J 1999;18:2313–2322.
74. Ito Y, Yamasaki K, Iwahara J, Terada T, Kamiya A, Shirouzu M, Muto Y, Kawai G, Yokoyama S, Laue ED, Walchli M, Shibata T, Nishimura S, Miyazawa T. Regional polysterism in the GTP-bound form of the human c-Ha-Ras protein. Biochemistry 1997;36: 9109–9119.
75. Schlichting I, Almo SC, Rapp G, Wilson K, Petratos K, Lentfer A, Wittinghofer A, Kabsch W, Pai EF, Petsko GA, Goody RS. Time-resolved X-ray crystallographic study of the conformational change in Ha-Ras p21 protein on GTP hydrolysis. Nature 1990;345: 309–315.
76. Diaz JF, Wroblowski B, Engelborghs Y. Molecular dynamics simulation of the solution structures of Ha-ras-p21 GDP and GTP complexes: flexibility, possible hinges, and levers of the conformational transition. Biochemistry 1995;34:12038–12047.
77. Ma J, Karplus M. Ligand-induced conformational changes in ras p21: a normal mode and energy minimization analysis. J Mol Biol 1997;274:114–131.
78. Kosztin I, Bruinsma R, O'Lague P, Schulten K. Mechanical force generation by G proteins. Proc Natl Acad Sci USA 2002;99:3575–3580.
79. Farrar CT, Ma J, Singel DJ, Halkides CJ. Structural changes induced in p21Ras upon GAP-334 complexation as probed by ESEEM spectroscopy and molecular-dynamics simulation. Structure 2000;8:1279–1287.
80. Moore KJ, Webb MR, Eccleston JF. Mechanism of GTP hydrolysis by p21N-ras catalyzed by GAP: studies with a fluorescent GTP analogue. Biochemistry 1993;32:7451–7459.
81. Springer TA. Folding of the N-terminal, ligand-binding region of integrin α-subunits into a β-propeller domain. Proc Natl Acad Sci USA 1997;94:65–72.
82. Fulop V, Jones, D.T. Beta propellers: structural rigidity and functional diversity. Curr Opin Struct Biol 1999;9:715–721.
83. Todd AE, Orengo CA, Thornton JM. Plasticity of enzyme active sites. Trends Biochem Sci 2002;27:419–426.
84. Todd AE, Orengo CA, Thornton JM. Sequence and structural differences between enzyme and nonenzyme homologs. Structure 2002;10:1435–1451.
85. Zvelebil MJ, Sternberg MJ. Analysis and prediction of the location of catalytic residues in enzymes. Protein Eng 1988;2:127–138.
86. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324:105–121.
87. Haliloglu T, Bahar I. Structure-based analysis of protein dynamics: comparison of theoretical results for hen lysozyme with X-ray diffraction and NMR relaxation data. Proteins 1999;37:654–667.
88. Wang C, Karpowich N, Hunt JF, Rance M, Palmer AG. Dynamics of ATP-binding cassette contribute to allosteric control, nucleotide binding and energy transduction in ABC transporters. J Mol Biol 2004;342:525–537.
89. Namba K. Roles of partly unfolded conformations in macromolecular self-assembly. Genes Cells 2001;6:1–12.