The Twilight Zone between Protein Order and Disorder

A. Szilágyi, D. Györffy, and P. Závodszky
Institute of Enzymology, BRC, Hungarian Academy of Sciences Karolina út 29, H-1113 Budapest, Hungary

ABSTRACT The amino acid composition of intrinsically disordered proteins and protein segments characteristically differs from that of ordered proteins. This observation forms the basis of several disorder prediction methods. These, however, usually perform worse for smaller proteins (or segments) than for larger ones. We show that the regions of amino acid composition space corresponding to ordered and disordered proteins overlap with each other, and the extent of the overlap (the "twilight zone") is larger for short than for long chains. To explain this finding, we used two-dimensional lattice model proteins containing hydrophobic, polar, and charged monomers and revealed the relation among chain length, amino acid composition, and disorder. Because the number of chain configurations exponentially grows with chain length, a larger fraction of longer chains can reach a low-energy, ordered state than do shorter chains. The amount of information carried by the amino acid composition about whether a protein or segment is (dis)ordered grows with increasing chain length. Smaller proteins rely more on specific interactions for stability, which limits the possible accuracy of disorder prediction methods. For proteins in the "twilight zone", size can determine order, as illustrated by the example of two-state homodimers.

INTRODUCTION

Intrinsically disordered (also called unstructured) proteins are characterized by a lack of stable secondary and tertiary structure under physiological conditions in the absence of a binding partner (1–3). Structural disorder can be assessed by various experimental methods including x-ray crystallography, NMR, circular dichroism, and hydrodynamic measurements (4). Intrinsically disordered proteins, and those having functionally important disordered regions, form a significant fraction of proteomes (3,5); e.g., it has been estimated that ~14% of *Escherichia coli* and 50–60% of yeast proteins contain at least one long (>30 residues) disordered segment (6). Because of their structural malleability (7), intrinsically disordered proteins are often involved in protein-protein interactions (8,9) with multiple binding partners (10–12). They are associated with a wide range of cellular functions (13), the most common being the regulation of transcription and translation, cellular signal transduction, protein phosphorylation, the storage of small molecules, and the regulation of the self-assembly of large multiprotein complexes (14–16).

The amino acid composition of disordered proteins/regions characteristically differs from that of ordered ones, with disorder-promoting residues (A, R, G, Q, S, P, E, K) enriched and order-promoting residues (W, C, F, I, Y, V, L, N) depleted in disordered proteins (3). A number of computational methods have been developed to predict disordered proteins or disordered regions in otherwise ordered proteins (see Ferron et al. (17) for a review). Most disorder-prediction methods rely at least partly on amino acid composition, either directly (18,19) or indirectly. A number of methods employ various scales or scores assigned to each of the 20 amino acids. These scales may be based on the physicochemical properties of amino acids, as in PONDR (3,20), their occurrence in ordered versus disordered segments, as in Globplot (21), etc. The actual algorithm employed for classification may be based on a simple linear combination of the input variables, but it may use sophisticated methods such as a support vector machine, as in Disopred2 (5) or a neural network, as in various versions of PONDR (3,15), Disembl (22) or RONN (23). Regardless of the scale and the algorithm used, the final output from the prediction method often depends only on the amino acid composition of the protein or segment under study; the various methods just provide different mappings of the amino acid composition space to a parameter describing order/disorder.

Some methods only provide predictions for entire protein chains (24,25). Most methods, however, provide a local measure of disorder and try to identify disordered segments of the chain. The local measures are often calculated using a window that slides along the sequence and assigns a score to the middle residue based on the other amino acids within the window. Many methods do not utilize the actual order of residues within the sliding window; the score assigned to the middle residue depends only on the amino acid composition of the segment enclosed by the window.

An earlier comparative study of the accuracy of several disorder prediction methods found an increase in prediction accuracy with increasing length of the disordered regions to be predicted (26). The comparison of the performance of six predictors tested in the CASP5 experiment showed that most predictors were significantly less accurate for short (<31 residues) than for long disordered regions (27,28). Because of the difficulty of predicting short disordered regions, a method was developed that used length-dependent parameters to optimize prediction accuracy for both long and short disordered regions (25,29).

Submitted February 7, 2008, and accepted for publication April 9, 2008.

Address reprint requests to A. Szilágyi, Institute of Enzymology, BRC, Hungarian Academy of Sciences Karolina út 29, H-1113 Budapest, Hungary. E-mail: szia@enzim.hu.

Editor: Ruth Nussinov.

0006-3495/08/08/1612/15 $2.00

doi: 10.1529/biophysj.108.131151

The lower accuracy of prediction of short disordered regions was attributed to the variation of amino acid compositions and sequence properties among disordered regions of different lengths (29). However, the difficulty of predicting disorder on short regions or small proteins may have a more fundamental basis. When disorder was predicted for whole chains by estimating the pairwise energy content from the amino acid composition in the IUPRED method, a greater overlap was found in the distribution of estimated energies for short chains than for long chains (30). This suggests that ordered and disordered sequences occupy overlapping regions in amino acid composition space, and the extent of the overlap varies with chain (or segment) length.

The charge-hydrophobicity plot, introduced by Uversky et al. (31), is a special projection of amino acid composition space (its special property is that the absolute value of the mean net charge is calculated). It was found that small, globular, folded proteins and natively unfolded proteins occupy distinct regions on the charge-hydrophobicity plot, and there is a very sharp boundary between the two regions, leading to an almost perfect separation of the two sets (31). The FoldIndex disorder prediction method uses the charge-hydrophobicity plot and classifies proteins or long segments as ordered or disordered based on which side of the boundary line they appear (32). However, FoldIndex was found to have a 23% false-negative and a 10% false-positive rate, which indicates that the separation is not perfect. In another study by Oldfield et al. (33), similar false-positive and false-negative rates were found when the boundary line on the charge-hydrophobicity plot was used for classification. By excluding proteins in a boundary region where the two classes were found to overlap, the false-negative and false-positive rates dropped to 5% and 3%, respectively, but this came at the price of excluding half of the proteins from the analysis (33). Using a larger set of ordered and disordered proteins, Garbuzynskiy et al. (34) also observed a significant overlap between the two classes on the charge-hydrophobicity plot.

Although the charge-hydrophobicity plot is a projection of amino acid composition space, it is reasonable to assume that the overlap between the classes of ordered and disordered proteins is also present in the full 20-dimensional amino acid composition space. The increasing difficulty of distinguishing ordered from disordered proteins or segments with decreasing chain length suggests that the overlap is stronger for shorter chains or segments. However, to our knowledge, this relation has not yet been investigated.

The existence of two-state homodimers, i.e., proteins that are disordered as monomers but fold (and become ordered) on homodimerization (35–37), makes it abundantly clear that amino acid composition alone does not determine whether a protein is ordered or disordered. Obviously, the amino acid composition of a homodimer is exactly the same as that of the monomer; the switch from the disordered to the ordered state observed with two-state homodimers is a consequence of the doubling of the size of the protein. This fact also demonstrates that protein size (i.e., the total number of amino acids) can have a tremendous influence on order/disorder.

Here, we use amino acid compositions of actual ordered and disordered proteins and segments to describe and characterize the overlap between the corresponding regions of amino acid composition space, and we investigate the dependence of the extent of the overlap on chain/segment length. To identify the possible reasons for the overlap and its dependence on chain length, we used simplified model proteins to see how amino acid composition determines order/disorder in chains of various lengths. A full mapping of amino acid composition space to the structural property of order/disorder is impossible for real proteins because existing proteins represent only a very limited sampling of amino acid composition space. Therefore, we turned to two-dimensional (2D) lattice models with reduced alphabets where amino acid composition space can be fully explored, and the fraction of disordered proteins among proteins with any given amino acid composition can be accurately determined.

The simplest 2D lattice model of proteins is the hydrophobic-polar (HP) model (38). This model contains only two types of residues: H (hydrophobic) and P (polar). The conformation of the chain is restricted to a 2D square lattice, and the only (favorable) interaction is between adjacent H residues. Despite their extreme simplicity, the HP model and other simple exact models display many important properties of real proteins and have been tremendously useful in the theoretical analysis of protein folding (39). For chain lengths for which exhaustive enumeration is possible (up to ~25 residues), 2D lattice models more accurately represent the physically important surface-interior ratios of proteins than do three-dimensional models (38), and the length distributions of helices and sheets that appear in 2D lattice models are also similar to those of real proteins (40).

For the study of protein disorder, the traditional HP model is limited because it does not include the effect of charged residues. A model on a cubic lattice with a three-letter alphabet, i.e., hydrophobic, positively charged, and negatively charged, was used earlier for the study of salt bridges (41), and a 2D model with a four-letter alphabet (hydrophobic, polar, positive, and negative) was used in a few studies of protein evolution and aggregation (42–44). Here, we employ the model with a three-letter alphabet, referred to as the HPN model, on a 2D lattice.

METHODS

Protein sets

To obtain a set of disordered proteins and segments, disordered segments at least 20 residues long were extracted from the DisProt database (45), v3.5. This resulted in 303 sequences. A set of ordered proteins was created by extracting all entries corresponding to single-chain proteins containing no nonstandard residues from the Protein Data Bank (46)). Only single-chain entries were used to avoid including chains that are disordered as monomers and fold only on binding. This resulted in 4041 sequences. The chains in both sets were divided into five bins by length. The size of the ordered set in each bin was significantly larger than the corresponding disordered set. For an unbiased analysis, balanced data sets (sets with the same number of ordered and disordered proteins) were desirable. Therefore, the ordered sets were culled by the following procedure: pairwise Euclidean distances in the amino acid composition space were calculated among all sequences in the bin, and sequences closest to another sequence were removed one by one until the desired target set size was reached. In the end, the bins corresponding to lengths 20–50, 50–99, 100–199, 200–299, and ≥300 contained 194, 184, 126, 64, and 38 proteins, respectively, with an equal number of ordered and disordered proteins in each bin. See Supplementary Material, Data S1, for listings of proteins in each set.

Charge-hydrophobicity plots

The hydrophobicity of a sequence was calculated by summing up the values of the Kyte-Doolittle hydrophobicity scale (47), normalized to the [0,1] interval, for the residues in the sequence. Charge was defined as the absolute value of $n_R + n_K - n_D - n_E$ where n_X denotes the number of X residues in a sequence. For the charge-hydrophobicity plot, both charge and hydrophobicity were divided by the chain length.

Lattice models

Two types of 2D lattice models were used: the traditional HP model (38); and another model having three residue types: hydrophobic, positively charged, and negatively charged (HPN model). The interaction energies between residues were set as follows; in the HP model, $E_{HH} = -1$, $E_{HP} = E_{PP} = 0$. In the HPN model, $E_{HH} = -1$, $E_{PN} = -0.75$, $E_{PP} = E_{NN} = 0.75$, $E_{HP} = E_{HN} = 0$. Changing these values, including the ratio of E_{PN} to E_{HH}, does not appear to change our results qualitatively.

Sequence and structure space were explored with different methods depending on model type and chain length. For HP models, all possible sequences were generated for chains of length 4 to 23, and full enumeration was used to determine the ground state. Maximum contact sets (48) were used to make the search efficient. For chains of length 30, 40, and 50, sequence space was randomly sampled, generating 1000 sequences for every possible H fraction (or all sequences if their number was <1000). A Monte Carlo search with simulated annealing was used to find the minimum energy state for these long chains. For HPN models, all possible sequences were generated for chains up to 14 residues; random sampling was used for chains of length 15 to 20 as well as 30, 40, and 50, generating 1000 sequences for all possible compositions (or all sequences where their number was <1000). To exploit the symmetry of positive and negative charges, only sequences with a nonnegative total charge were considered. To explore structure space, enumeration was used for chains up to 20 residues, using contact sets (48), and Monte Carlo simulated annealing for chains of length 30, 40, and 50.

For the Monte Carlo optimizations, the move set defined by Chan and Dill (49) was used. A simulated annealing protocol was performed; the temperature was reduced from $T = 1.0$ to $\sim T = 0.1$ in 100 steps according to a geometric series. A series of tests were performed for nine 50-residue chains with minimum energies determined by equienergy sampling (50) as well as 200 23-residue sequences where exact enumeration could also be performed. The lowest-energy conformation in a Monte Carlo trajectory of 1 million steps was found to be a good estimate of the actual ground state energy in most cases, e.g., the estimated energy was within two energy units of the actual energy for seven of the nine 50-residue sequences and was accurate for 96.5% of the 23-residue sequences.

Information theoretical analysis of classification problems

Shannon's definitions (51) were used to calculate information contents and mutual information. For a classification problem, we let C denote a random variable whose possible values are assigned to the classes to be predicted. To correctly assign a class to an object, we need $H(C) = -\sum_c p(c) \log_2 p(c)$ bits of information, where p denotes the probability distribution function of C. If we have a predictor variable X, we can then calculate the amount of information about C provided by X using the formula for the mutual information: $I(C;X) = H(C) - H(C|X) = \sum_x \sum_c p(x,c) \log_2(p(x,c)/p(x)p(c))$, where $H(C|X)$ denotes the entropy of C conditional on X; $p(x,c)$ denotes the joint probability distribution function of X and C; and $p(x)$ and $p(c)$ denote the marginal probability distribution function of X and C, respectively. For two predictor variables X and Y, the extra information provided by Y over that provided by X is given by the conditional mutual information $I(C; Y|X)$.

RESULTS

Overlap in amino acid composition space between ordered and disordered proteins

Amino acid composition space is a multidimensional space with each axis representing the fraction of a given amino acid in a given sequence. Because the fractions of the 20 amino acids sum up to 1, any amino acid composition corresponds to a point on a 19-dimensional simplex in 20-dimensional space. We may refer to this simplex as the ‘‘amino acid composition simplex’’.

The points representing the amino acid compositions of a class of proteins can be considered samples from an underlying probability density function in amino acid composition space. For an efficient classification of proteins based on amino acid composition, the density functions corresponding to different classes should have as little overlap as possible.

To determine the extent of overlap between the hypothetical density functions of ordered and disordered proteins, we first created sets of ordered and disordered proteins of various lengths as described in the Methods section. We divided the proteins into five bins by chain (or segment) length and culled the sets so that each bin contained an equal number of ordered and disordered proteins. The bins corresponded to lengths <50, 50–99, 100–199, 200–299, and >299, respectively. The extent of overlap between ordered and disordered proteins in each bin was characterized by calculating the error rate of k-nearest-neighbor (kNN) classification. In kNN classification, the class of an object is predicted to be the same as the most frequent class among its k nearest neighbors in feature space. Here, we used the 19-dimensional amino acid composition simplex and the Euclidean distance metric. Intuitively, the larger the error rate of kNN classification, the more the points from the two sets are mixed in space, i.e., the larger the overlap between the sets. Fig. 1 shows the results for $k = 1$. It is clearly seen that the error rate monotonically decreases as chain length grows. In other words, the overlap between the set of ordered and the set of disordered proteins in amino acid composition space is large for short proteins and decreases with growing chain length.

To get a feeling for the distribution of points corresponding to ordered and disordered proteins in amino acid composition space and how much the two sets overlap, we plotted the points on a charge-hydrophobicity plot, which can be considered a projection of amino acid composition space, apart

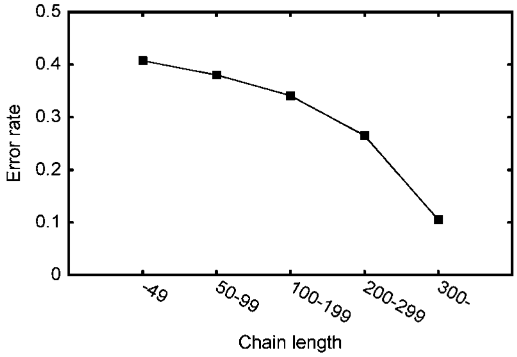

FIGURE 1 The error rate of *k*NN classification of ordered and disordered proteins in the 19-dimensional amino acid composition simplex, with $k = 1$, for proteins in five length bins. Results with $k = 3$ are similar (not shown). The error rate decreases with increasing chain length because of better separation of the two groups.

from the fact that the absolute value of the charge is taken. Charge-hydrophobicity plots have been shown to discriminate well between ordered and disordered proteins (31,33) and are even the basis of the FoldIndex disorder prediction method (32). Fig. 2, *A–E* shows the plots for the proteins in each length bin.

By visual inspection of the plots, the relation between overlap and chain length is clear. Sequences with large hydrophobicity and low charge are mostly ordered, whereas those with small hydrophobicity and high charge are mostly disordered. Between these two extremes, there is a transition region, a ''twilight zone'', where the two types of proteins are mixed. The width of this twilight zone decreases with increasing chain length, changing from very wide for short proteins to very narrow for long ones.

To obtain a more quantitative characterization of the relation between chain length and twilight zone width, we applied logistic regression to each dataset. In logistic regression, a logistic (sigmoid) function, here of the form $1/(1 + \exp(ax + by + c))$, with x being the hydrophobicity and y the charge, is fitted to a binary outcome variable representing each class; we assigned values of 0 to ordered proteins and 1 to disordered ones. The resulting function describes the probability of a protein with the specified hydrophobicity and charge being disordered. In the plots in Fig. 2, *A–E* gray shading indicates the probability function obtained from logistic regression. The medium gray band represents probabilities between 0.2 and 0.8 and is defined as the actual twilight zone. The narrowing of this band is clearly seen as chain length grows. Fig. 2 *F* shows that the percentage of points in the twilight zone also decreases as the chain length grows, demonstrating that the observed sharpening is not a result of a ''shrinking'' of the plot but that the two protein classes indeed separate better when longer chains are considered.

Although the position of the twilight zone also appears to be slightly different for different chain lengths, a clear trend cannot be identified.

Lattice model studies

Definition of disorder for model proteins

To study the relation between amino acid composition and disorder in lattice models, first of all we need to define disorder for 2D lattice models. Intrinsic disorder has usually been defined as the lack of a well-defined, compact native structure under physiological conditions (1,3). In terms of the energy landscape of 2D lattice models, this may be interpreted as a

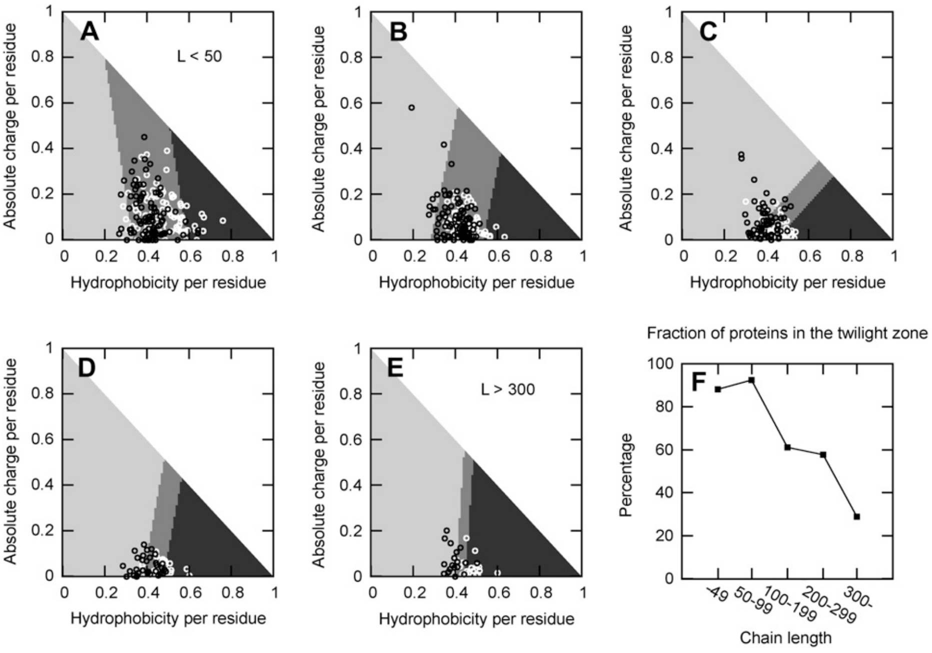

FIGURE 2 (*A–E*) Charge-hydrophobicity plots of proteins in five chain length bins. White circles indicate ordered proteins, and black circles indicate disordered ones. The gray shading represents a bivariate linear logistic function fitted to the data after assigning 1s to disordered proteins and 0s to ordered ones. The grayscale is defined as in Fig. 4. The medium gray region, corresponding to a probability of disorder between 0.2 and 0.8, is defined as the twilight zone. (*F*) The fraction of proteins in the twilight zone in the five chain length bins. With increasing chain length, the twilight zone becomes narrower.

highly degenerate ground state or a large number of low-energy states corresponding to a number of diverse, noncompact structures.

Here, we use a simple definition of disorder that is consistent with this picture. A 2D lattice sequence is considered disordered when the ground state energy per residue E_{ground}/L is higher than a predefined, fixed threshold. The rationale for this definition is that for a protein to be ordered, a significant fraction of its residues should be bound by other residues. Residues that are not bound by other residues are usually free to fluctuate and, therefore, contribute to disorder. The quantity E_{ground}/L is in fact a measure of the thermodynamic stability of the model protein. Namely, the free energy difference between the folded and the unfolded state is $\Delta F = \Delta E - T\Delta S$. If we assume that the unfolded state has an energy $E_{\mathrm{unfolded}} = 0$ and the ground (i.e., folded) state has an entropy $S = 0$, then $\Delta F = E_{\mathrm{ground}} + TS_{\mathrm{unfolded}}$. But S_{unfolded} is approximately proportional to the chain length L, i.e., $S_{\mathrm{unfolded}} = aL$, where a is a proportionality constant. Substituting aL for S_{unfolded} and dividing the formula for ΔF by L, we obtain $\Delta F/L = E_{\mathrm{ground}}/L + aT$. Thus, the sign of ΔF depends on whether E_{ground}/L is below or above $-aT$, a constant at any given temperature. A low value of $-E_{\mathrm{ground}}/L$ therefore indicates a low stability against unfolding; i.e., the protein will be natively unfolded (52). Fig. 3 shows a few examples of ground states of disordered and ordered HP and HPN sequences.

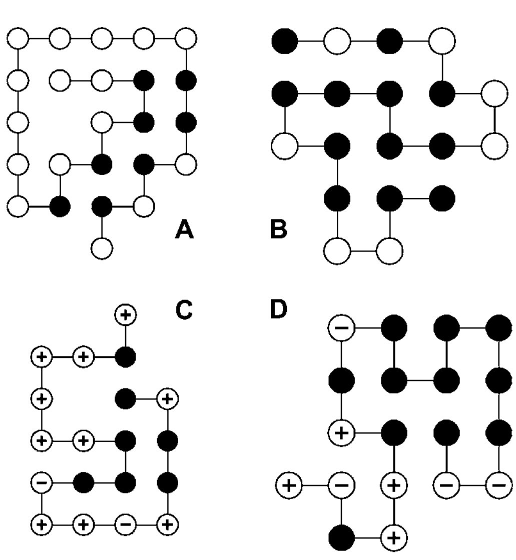

FIGURE 3 Examples of ground states of HP and HPN sequences. Black circles represent H (hydrophobic) monomers; white circles represent P (polar) monomers in the HP model. Circles with + and − signs inside represent P (positively charged) and N (negatively charged) monomers, respectively, in the HPN model. (*A*) A disordered HP model protein ($E_{\mathrm{ground}}/L = -0.25$); (*B*) an ordered HP model protein ($E_{\mathrm{ground}}/L = -0.47$); (*C*) a disordered HPN model protein ($E_{\mathrm{ground}}/L = -0.25$); (*D*) an ordered HPN model protein ($E_{\mathrm{ground}}/L = -0.43$). Although each of these four sequences has a single ground state, the disordered sequences are very unstable because they do not have enough stabilizing contacts in their ground state to ensure a sufficiently low ground state energy for the given chain length.

When the interaction energies specified in the Methods section were used, the threshold for the ground state energy per residue (corresponding to $-aT$ in the derivation above) used to define disorder was set to −0.3 by trial and error. This threshold divides the sequence space of HP models into two roughly equal parts, making the probability ~50% that a random sequence is ordered. Also, by this threshold, most model proteins with >40% hydrophobic residues get classified as ordered, in agreement with the typical fraction of hydrophobic residues in real ordered proteins (53). Different thresholds do not change our results qualitatively; the plots shown in the following figures simply get shifted.

HP model

The traditional HP model contains two types of monomers: the H and the P. The only interaction is between the H monomers: $E_{\mathrm{HH}} = -1$. As described in the Methods section, we generated all possible sequences with lengths 4 to 23 and sampled sequence space for lengths 30, 40, and 50. The ground state of each sequence was found by enumeration for lengths 4 to 23 and estimated by Monte Carlo search for lengths 30, 40, and 50. Each sequence was classified as either ordered or disordered. The amino acid composition simplex of HP sequences is one-dimensional, and is described here by the fraction of H residues.

Fig. 4 *A* shows the fraction of disordered sequences among sequences with a given H fraction and length. A bivariate quadratic logistic function fits the data very well and is visualized in Fig. 4 *A* by shades of gray. All sequences with low H fractions are disordered, and all sequences with high H fractions are ordered. Between the two extremes, there is a twilight zone where a fraction of all sequences with a given H fraction are disordered. In Fig. 4 *A*, a medium gray band, corresponding to fractions 0.2 to 0.8, indicates this twilight zone. In this zone, H fraction alone is not sufficient to tell whether a sequence is ordered; this depends on the specific order of monomers in the sequence. For short chains, the twilight zone is wide (note that the width is measured along vertical lines in Fig. 4 *A*), and as the chain grows, it becomes narrower, and its midpoint shifts to lower H fractions. For longer chains, the midpoint of the twilight zone seems to converge to around H-fraction = 0.4.

HPN model

It is a general observation that disordered proteins contain more charged residues and fewer hydrophobic residues than ordered ones, as often illustrated by charge-hydrophobicity plots. To reproduce this behavior with 2D lattice models, we introduced a model, called the HPN model, with three types of monomers: hydrophobic (H), positively charged (P), and negatively charged (N). As described in the Methods section, we generated all possible sequences for chain lengths 4 to 14 and sampled sequence space for lengths 15 to 20, 30, 40, and

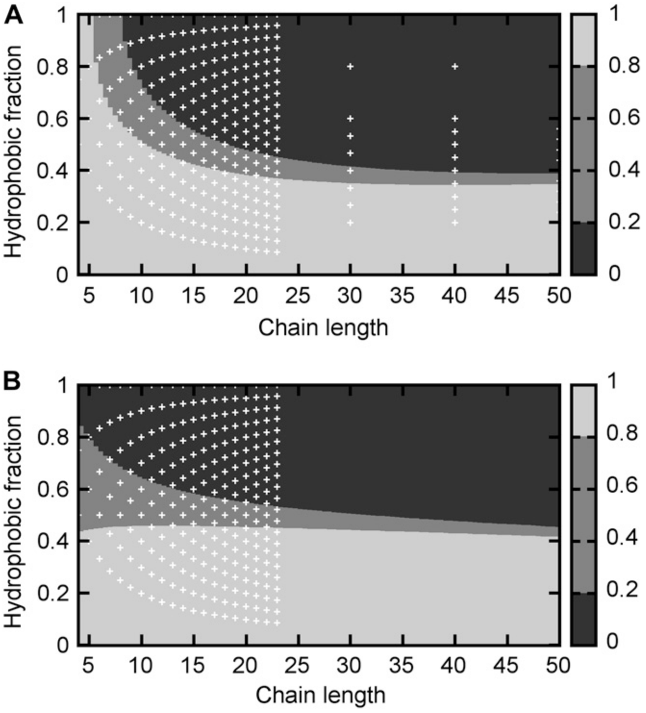

FIGURE 4 (*A*) The fraction of disordered sequences among all HP sequences with a given length and hydrophobic fraction. A bivariate quadratic logistic function, visualized here by shades of gray according to the scale on the right, was fitted to the data. The locations of the data points are indicated by white crosses. The medium gray band, corresponding to disordered fractions between 0.2 and 0.8, is defined as the twilight zone. (*B*) The same as *A* but calculated with length-dependent interaction energies (see text).

50. The ground state was determined by enumeration for chain lengths up to 20 and estimated by Monte Carlo search for lengths 30, 40, and 50. For this model, the amino acid composition simplex is 2D and is described here by the H fraction and the net charge per residue. It should be noted that HPN sequences with low H fractions get primarily stabilized by electrostatic interactions between monomers arranged as in a salt crystal (Fig. 3 *D*). Although globular proteins typically have a hydrophobic core, there are several proteins containing tandem repeats of alternating charges such as the KEKE motif (54). These stretches are thought to form α-helices stabilized by salt bridges.

Fig. 5 shows the fraction of disordered sequences among sequences characterized by a given H fraction and net charge per residue for various chain lengths. A bivariate quadratic logistic function of the form $1/(1 + \exp(ax^2 + by^2 + cxy + dx + ey + f))$ was fitted to the data and shown by gray shading on the H fraction versus net charge per residue plane. The quadratic form was used to allow for curved boundaries between regions of order and disorder. This function fits the data points extremely well, with the root mean-square of residuals ranging from 0.02 to 0.05. Fig. 5 presents the data for different chain lengths. For each length, sequences with high hydrophobicity and low charge are all ordered, and those with low hydrophobicity and high charge are all disordered. Between these two extremes, there is a transition region (twilight zone) with varying width, shape, and position, where the H fraction and the net charge per residue are not sufficient to tell whether a sequence is ordered; it depends on the particular sequence. Generally, as the chain length grows, the region of all-ordered sequences grows and occupies an ever-increasing portion of the plot. The twilight zone shifts toward lower hydrophobicites and higher charges, and it also becomes much narrower.

Lattice models with length-dependent interaction energies

The lattice models presented exhibit a behavior similar to real proteins regarding the decrease of the width of the twilight zone as chain length increases. However, the boundary between regions of order and disorder in the amino acid composition space of the lattice models also shifts to lower hydrophobicities and higher charges as chain length increases, a phenomenon not observed with real proteins. In fact, the hydrophobic fractions of real proteins depend little on chain length (53,55), although a maximum somewhere between 200 and 300 residues was found by Bastolla and Demetrius (56). On the other hand, the native energy per residue also tends to be nearly constant for proteins of various sizes (56,57) despite the fact that the number of contacts per residue increases (56). This apparent contradiction is resolved by the finding that the native energy per contact decreases (in absolute value, i.e., the contacts get weaker) as chain length grows (56).

The ground state energies of sequences with a given chain length (L) and number of hydrophobic residues (H) follow a bell-shaped distribution (Fig. 6). The maximum of this distribution corresponds to the most frequently occurring ground state energy, $E_{\mathrm{ground,m}}$. We calculated the most common ground state energy per contact, $E_{\mathrm{ground,m}}/C$ and the most common ground state energy per residue, $E_{\mathrm{ground,m}}/L$, for all HP sequences used in our HP model calculations. Fig. 7, *A* and *B* shows the results plotted as a function of the chain length L and the hydrophobic fraction $h = H/L$. Clearly, the ground state energy per contact depends very little on either L or h, whereas the ground state energy per residue depends on both: longer chains have lower energies per residue, as do more hydrophobic chains. This behavior is different from that of real proteins, where the stability per residue does not depend much on chain length but the contacts get weaker as chain length grows.

To compensate for the stability increase (i.e., lower ground state energy per residue) of our lattice models as chain length grows, we introduced modified models where the interaction energies depend on chain length. To find the optimum form for the length dependence of interaction energies, we used the observation that the native energy per residue tends to be nearly constant for real proteins (56,57). For any HP model sequence, the ground state energy is equal to $E_{\mathrm{HH}}C_{\mathrm{HH}}$, where

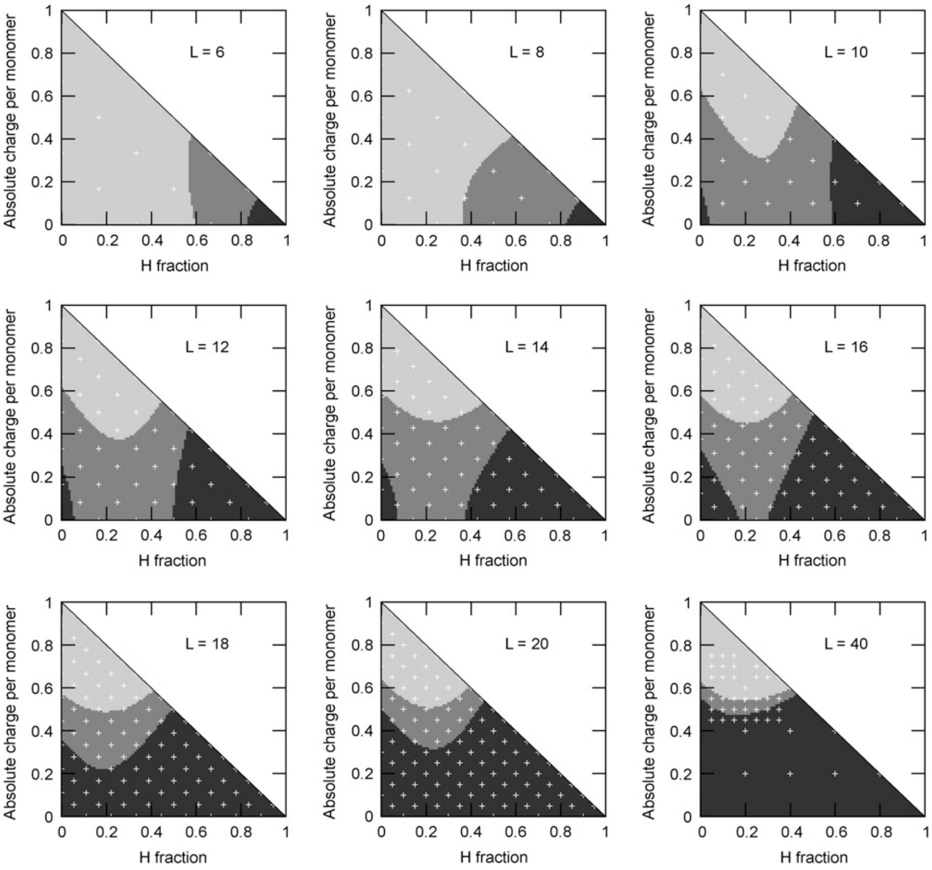

FIGURE 5 The fraction of disordered HPN sequences among all sequences with a given hydrophobic fraction (*horizontal axis*) and absolute charge per monomer (*vertical axis*) for various chain lengths. A bivariate quadratic logistic function, visualized here by shades of gray according to the scale shown in Fig. 4, was fitted to the data. White crosses indicate the locations of actual data points. The medium gray region, representing fractions between 0.2 and 0.8, is the twilight zone.

C_{HH} is the number of H-H contacts in the ground state, and E_{HH} is the H-H contact energy (which was set to -1 in the HP model). For any given chain length L and hydrophobic fraction h, the most frequently occurring value of C_{HH} can be determined. We denote this value by $C_{HH,m}(L,h)$, and let $g(L,h) = C_{HH,m}(L,h)/L$ be the most frequent number of H-H contacts per residue. We found that a quadratic bivariate function fits $g(L,h)$ very well (see Fig. 11 *B*). Because the ground state energy per residue of most HP sequences is equal to $E_{HH}g(L,h)$, we can eliminate the length dependence of the ground state energy per residue by replacing E_{HH} by $E_{HH}(L) = E_{HH}g(L_0,h_0)/g(L,h_0)$, where L_0 is the value of L for which we want $E_{HH}(L)$ to be equal to E_{HH}, and h_0 is the value of h where we want the ground state energies to match those calculated with the length-independent interaction. Fig. 8 shows the resulting length dependence of the H-H interaction energy with $L_0 = 12$ and $h_0 = 0.5$. Using the length-dependent interaction energy, we recalculated the ground state energies of all HP sequences. Fig. 7, *C* and *D*, shows $E_{ground,m}/C$ and $E_{ground,m}/L$ calculated this way. As the figures show, the models now correctly reflect the behavior of real proteins: the native energy per residue depends little on chain length, but the contacts get weaker as the chain grows.

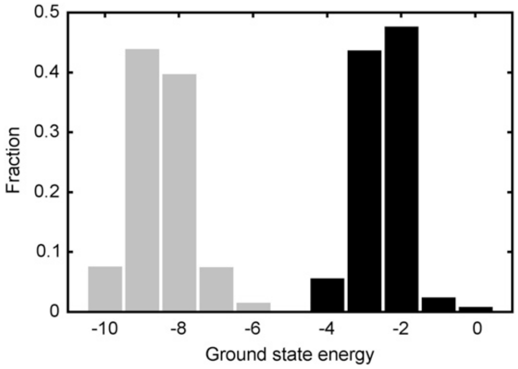

FIGURE 6 The distribution of ground state energies over sequences of length 10 containing five H monomers (*black bars*) and sequences of length 20 containing 12 H monomers (*gray bars*).

For the electrostatic interactions in the HPN model, similar calculations could be carried out and length-dependent forms of E_{PN}, E_{PP}, and E_{NN} could be introduced. However, for the sake of simplicity and consistency, we chose to keep the relative magnitudes of hydrophobic and electrostatic interactions; therefore, we set $E_{PN}(L) = 0.75E_{HH}(L)$ and $E_{PP}(L) = E_{NN}(L) = -E_{PN}(L)$. Because in this way the interaction energies are uniformly scaled, the structures of the ground states remain unchanged.

Using the length-dependent interaction energies, we recalculated the fraction of disordered sequences characterized by a given H fraction (for the HP model) or a given H fraction and net charge per residue (for the HPN model). The results

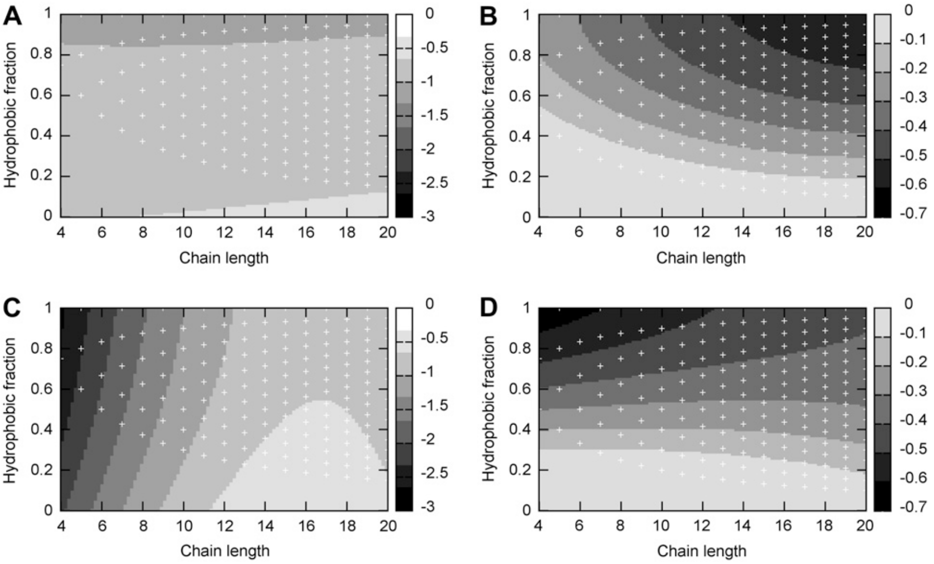

FIGURE 7 The most common ground state energy per contact (*A* and *C*) and per residue (*B* and *D*), calculated with length-independent (*A* and *B*) and with length-dependent (*C* and *D*) H-H interaction energies, as a function of chain length and hydrophobic fraction, for HP models. Bivariate quadratic functions were fitted to the data and are visualized by shades of gray. The locations of the data points are indicated by white crosses.

are presented in Figs. 4 *B* and 9 respectively. We found that the boundary between the ordered and disordered regions in amino acid composition space still gets sharper with increasing chain length (i.e., the twilight zone gets narrower), but it does not move systematically in any direction. Thus, our lattice models with chain-length-dependent interaction energies now correctly reflect the behavior of real proteins.

The amount of information about disorder carried by the amino acid composition

Clearly, the sequence of a (model) protein fully determines its ground state(s) and its entire behavior, including whether it is ordered or disordered. We have seen that for sufficiently low (high) hydrophobicities, all HP model sequences are disordered (ordered), but for medium hydrophobicities, composition alone does not fully determine order/disorder. But the composition still carries some information about disorder. If we were to develop a composition-based disorder prediction method for HP models, we could predict disorder with an accuracy that is limited by the amount of information on disorder that is actually carried by the residue composition. Taking all possible H fractions equally likely, we calculated the amount of information needed to correctly classify a HP sequence with a given length as ordered or disordered (Fig. 10, *circles*). Almost independently from the chain length, this amount is close to 1 bit.

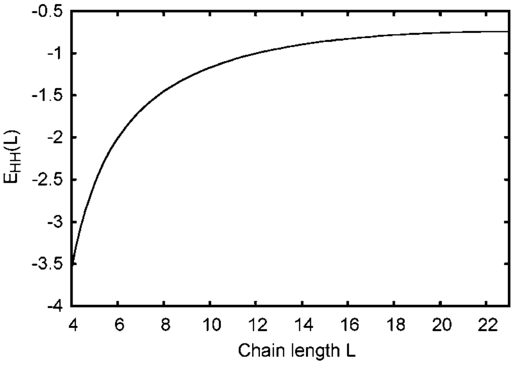

FIGURE 8 The length-dependent H-H interaction energy, defined as described in the text, plotted as a function of chain length.

Next, using the definitions for mutual information, we calculated the amount of information on disorder contained in the residue composition (Fig. 10, *solid squares*). For short chains, this amount is ~0.5 bits; it increases rapidly as chain length grows and is above 0.8 bits for longer chains. The rest of the information needed to correctly classify a sequence is contained in the particular order of residues. Thus, the order of the residues contains ~0.5 bits of information on disorder for short chains but <0.2 bits for long chains.

Some of this information can be extracted by smartly defined sequence-dependent quantities. An interesting property of 2D lattice proteins is that residues i and j ($i < j$) can be in contact only if $j = i + 3 + 2k$, where k is a nonnegative integer. Using this rule, we can calculate the number of potentially interacting H-H pairs for any given sequence. This quantity, which we will denote by Q, is an upper limit for the number of H-H contacts in the ground state and therefore carries information on whether the sequence is ordered. Using the definition of conditional mutual information, we calculated the information on disorder contained in Q on top of that contained in the residue composition (i.e., the H-fraction). Fig. 10 (*open squares*) shows the result: for short chains, Q contains extra information of over 0.2 bits, which reduces to below 0.1 bits for longer chains. Thus, for long chains, disorder is almost fully determined by the residue composition, but for short chains, the particular order of residues must be used in some way for a successful prediction of disorder.

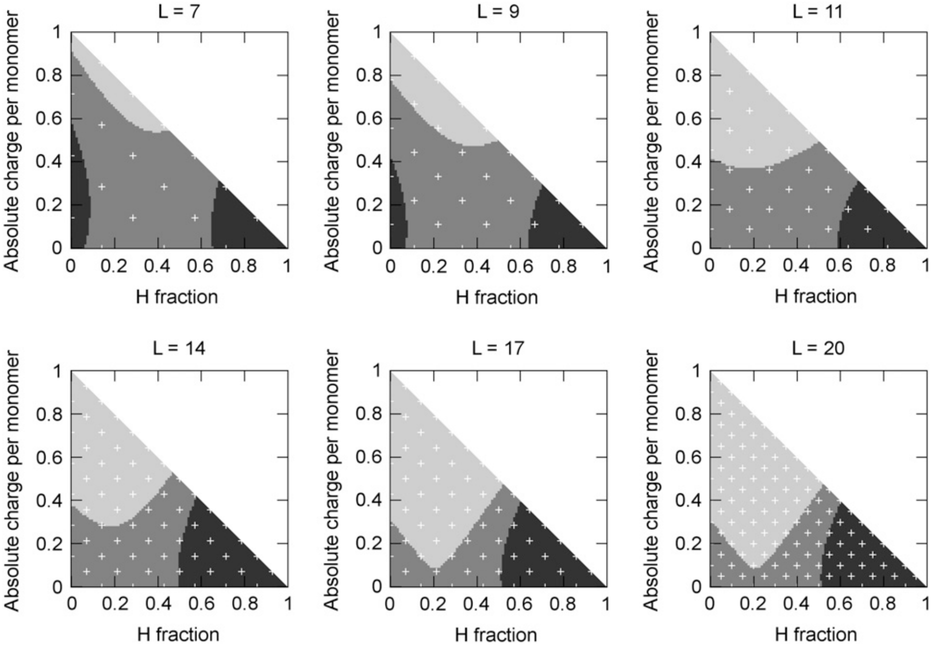

FIGURE 9 The fraction of disordered HPN sequences, calculated with the length-dependent interaction energies (see text), among all sequences with a given hydrophobic fraction (*horizontal axis*) and absolute charge per monomer (*vertical axis*) for various chain lengths. A bivariate quadratic logistic function, visualized here by shades of gray according to the scale shown in Fig. 4, was fitted to the data. White crosses indicate the locations of actual data points. The medium gray region, representing fractions between 0.2 and 0.8, is the twilight zone.

DISCUSSION

Explanations of the observed overlap between protein classes

In this article, we have demonstrated that the sets of ordered and disordered proteins (and protein segments) overlap with each other in amino acid composition space. The extent of the overlap is large for short proteins (and segments) and decreases as chain length increases (Figs. 1 and 2). For long proteins, the boundary between the ordered/disordered regions is quite sharp.

It is important to find out why the overlap is present. Disorder prediction algorithms that use the amino acid composition as their input become inaccurate when the input amino acid composition falls in the twilight zone (33). However, the fact that there is an overlap between the two sets does not mean that an accurate classification is impossible. One may suggest that a very sophisticated, nonlinear method might be able to separate successfully the two classes, provided that amino acid composition actually determines order/disorder. But if there is no such determination, i.e., proteins with exactly the same amino acid composition can be either ordered or disordered, depending on their sequences, then no classification algorithm that uses amino acid composition alone can ever succeed in accurately predicting disorder. Therefore, if we can find the reason for the overlap between the two classes and understand the role of amino acid composition in determining order/disorder, then we can estimate the upper limit for the possible accuracy of disorder prediction methods.

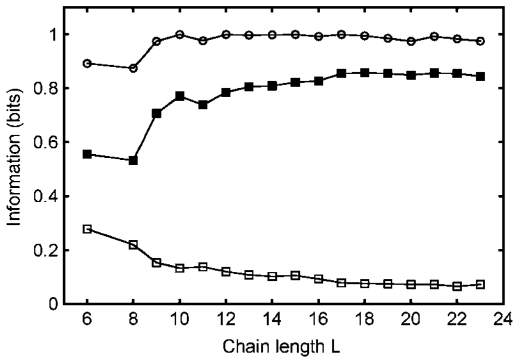

FIGURE 10 Information theoretical quantities related to disorder prediction in HP models, plotted as a function of chain length. The amount of information (in bits) needed to correctly classify an HP sequence as ordered or disordered (*circles*); the information about disorder contained in the residue composition (*solid squares*); the extra information (i.e., on top of that contained by the residue composition) about disorder contained in the quantity Q (the number of potentially interacting pairs of H residues) (*open squares*).

What causes the observed overlap in amino acid composition space between ordered and disordered proteins, and the dependence of the width of the twilight zone on protein length? Several explanations may be proposed.

One explanation could be that the experimental uncertainty in assessing disorder is larger for small proteins than it is for large ones. When the presence of disorder is judged from experimental data, e.g., fluorescence or NMR spectra, the situation may not be clear-cut. But the boundary between order and disorder is usually sharper for large proteins: a 100-residue extended segment should be easy to recognize. However, when a short peptide is investigated, things get fuzzier. ‘‘Dual personality’’ fragments, which seem ordered in some experiments and disordered in others, have a length distribution that is heavily skewed toward shorter fragments (58). Also, smaller proteins have more cysteines per residue, and more disulfide bridges per cysteine, than longer ones (56) and often require cofactors for folding (59). Obviously, the influence of disulfide bridges and cofactors cannot be taken into account when amino acid compositions are considered. These uncertainties may cause some of the observed overlap between ordered and disordered proteins, but we should look for other explanations as well.

Another explanation might point out that the variance of amino acid compositions is larger for small proteins than for large ones. In a naive model of protein sequence generation, amino acids are drawn from an infinite amino acid pool having a particular composition. Clearly, the expected amino acid composition of any sequence generated this way will be equal to the composition of the pool, but the variance will be proportional to $1/L$, where L is the length of the chain. This way, if we assume that ordered proteins come from a pool with a different amino acid composition than disordered ones, the extent of overlap between the two sets will decrease as the variance of amino acid compositions of both sets decreases with increasing chain length. But there is no reason to assume that this model of sequence generation is correct; selective constraints should definitely apply, and the amino acid composition of proteins depends on chain length (53,60).

The explanation we suggest here is that amino acid composition alone does not fully determine order/disorder. The structure of a protein, or the lack of it, is ultimately determined by its sequence and the specific interactions forming between the residues. In short chains, the specific interactions are more important in determining order/disorder than in long chains.

Definition of disorder for model proteins

To study intrinsically disordered proteins using lattice models, we must define intrinsic disorder for these models. A disordered protein is commonly pictured as a mostly extended, highly fluctuating chain. This implies few interactions per residue, and this is the basis of our definition. We define sequences with ground state energy per residue below a specified threshold as ordered. This definition was inspired by the principle of the IUPRED disorder prediction algorithm (30): it estimates the sum of pair interaction energies divided by the length and predicts order when the estimated value is below a threshold. The success of IUPRED lends justification to our definition.

It is worth noting that this definition has an interesting ''side effect'': highly hydrophobic sequences, which have compact but degenerate ground states, will be considered ordered. Although this may seem inappropriate, in fact it agrees well with the intuitive concept of disorder, which implies a mostly extended chain; besides, there are no known proteins that are highly hydrophobic and still considered disordered. By our definition, ordered proteins tend to be more compact than disordered ones, which is consistent with the general idea of protein order/disorder. For example, the average size (the larger of the horizontal and vertical extensions) of the ground states of ordered versus disordered 16-residue HP model sequences is 4.28 vs. 4.88, and the correlation coefficient between ground state energy and size is 0.748. It should be noted that although the ground states of highly hydrophobic sequences are degenerate, their degeneracy is still very low compared with low-hydrophobicity sequences. For example, the 16-residue sequence H_{16} has 69 ground states, whereas $HP_{14}H$ has 11,752. In fact, by our definition, order/disorder is strongly related to a low/high ground state degeneracy; e.g., the average number of ground states (n_{ground}) is 61 vs. 20,543 for ordered versus disordered 16-residue HP model sequences, and the correlation coefficient between ground state energy and log n_{ground} is 0.751.

We considered a number of alternative definitions before deciding in favor of the one we finally adopted. Because intrinsically disordered proteins are thought to have a flexible, fluctuating structure, we may define order/disorder based directly on the degeneracy of the ground state: e.g., sequences with a single ground state (also called ''designing sequences'') could be defined as ordered, and those with multiple ground states as disordered. However, the fraction of designing sequences among all sequences is very small, e.g., $<4\%$ for 22-residue HP sequences with medium hydrophobicity. Therefore, by this definition, using the amino acid composition to predict whether a model protein is ordered or disordered would be impossible, and no meaningful analysis could be carried out. For a meaningful analysis, a definition is needed that classifies a significant fraction of sequences as ordered. Our definition meets this requirement, and although it is not directly based on the number of ground states, it uses a parameter that is strongly correlated with it (see the example above).

Another possible definition of disorder may be based on a thermodynamic property, e.g., the folding temperature (T_f, the temperature above which the unfolded state is dominantly populated) or the folding free energy (ΔF). For example, proteins with a T_f or ΔF above/below a predefined threshold could be defined as ordered/disordered. However, for this sort of definition, sequences with degenerate ground states pose a problem. They are entropically favored and may therefore be more stable than nondegenerate ground states. This would lead to instances where sequences with single ground states would be classified as disordered and other sequences with the same ground state energy but degenerate ground states as ordered, clearly an unacceptable situation.

One solution to this problem would be to restrict our analysis to sequences with unique ground states, as is often done in applications of 2D HP models (39). However, we are modeling disordered proteins and therefore want to include sequences with degenerate ground states. Another solution would be to arbitrarily pick one of the ground states and designate it as ''the native state'' and consider the others misfolded. But in this treatment, the probability of the ''native state'' at any temperature would be <0.5 for sequences with degenerate ground states, and therefore, they would all be considered disordered, and the definition would be equivalent to simply defining order/disorder based on whether the sequence is designing (see above).

Finally, a third solution is to define order/disorder based purely on the ground state energy, thereby accounting for entropy only implicitly, through its correlation with energy. This is the solution we chose for our definition. In fact, our definition is based on a measure of the stability of the folded state, and, as we have shown in the Results section, it is equivalent to a ΔF-based definition at low temperatures.

A limitation of our definition is that it implies a binary classification of proteins. A structural continuum extending from tightly folded single domains, through proteins with long disordered segments to highly extended chains would be a more appropriate description of the diversity of protein structures (16). However, our lattice models are highly simplified and cannot be expected to reflect all those complexities.

What model proteins tell us

Real proteins do not sample amino acid composition space sufficiently to provide us with a full mapping of amino acid compositions to the classes of order and disorder. Therefore, we turned to model proteins for a theoretical approach. We used the well-known HP model to study the influence of hydrophobicity and chain length on disorder, and the new HPN model to also include the influence of charged residues. These models show a narrowing of the twilight zone with increasing chain length, and thereby reproduce the behavior of real proteins. Of course, the length range we studied with model proteins is much shorter than that of real proteins, but because of the highly simplified nature of lattice models, their chain lengths cannot be equated with those of real proteins. Rather, one monomer of a lattice model can be taken to represent a longer segment of a real protein, e.g., a secondary structure element (61). Thus, lattice model chain lengths map to considerably greater chain lengths of real proteins.

What causes the observed shift of the twilight zone toward lower hydrophobicities in the case of HP model proteins? To get a deeper insight into what happens, let us examine the actual ground state energies. We found that a bivariate quadratic function approximates $E_{\mathrm{ground,m}}(L,H)$, i.e., the most frequently occurring ground state energy for chains of length L with H hydrophobic residues, very well and visualized this function by shades of gray in Fig. 11 *A*. As the contour lines show, the same number of hydrophobic residues can usually form a configuration of lower energy in longer chains than in shorter ones. Clearly, longer chains have many more possible configurations than short ones (the number of distinct self-avoiding configurations of chains of length L on a 2D square lattice is proportional to $\sim 2.67^{L}$ (62)), and therefore have a higher probability of bringing the hydrophobic residues together in an energetically favorable arrangement than shorter chains. Also, longer chains have larger core regions relative to the surface, which entails that the fraction of residues with two contacts relative to those with one contact is larger, leading to a lower ground state energy.

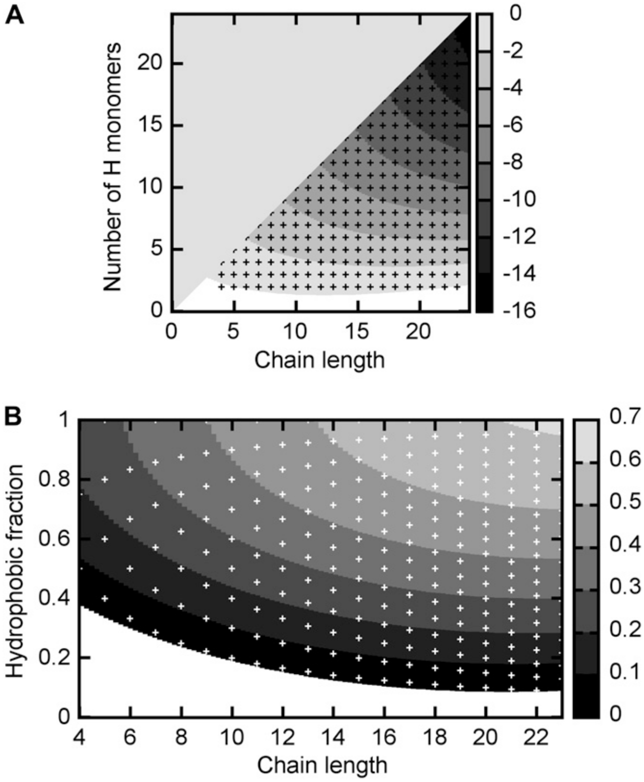

FIGURE 11 (*A*) The most frequent ground state energy of HP sequences with a given chain length and number of H monomers. A bivariate quadratic function, visualized here by shades of gray according to the scale on the right, was fitted to the data. The locations of data points are indicated by black crosses. (*B*) The function $g(L,h) = C_{\mathrm{HH,m}}(L,h)/L$, i.e., the most frequent number of H-H contacts per residue (see text) in ground states of HP sequences, as a function of chain length (L) and hydrophobic fraction (h). A bivariate quadratic function, visualized here by shades of gray according to the scale shown on the right, was fitted to the data. White crosses indicate the locations of data points used for the fitting.

In addition, the same hydrophobic fraction corresponds to a higher absolute number of hydrophobic residues in longer chains, which further lowers the ground state energy attainable by the chain. In the Results section, we defined the function $g(L,h) = C_{\mathrm{HH,m}}(L,h)/L$, i.e., the most frequent number of H-H contacts per residue (Fig. 11 *B*). Looking at the dependence of $g(L,h)$ on the chain length L, with the hydrophobic fraction h held constant, we find that for any given hydrophobic fraction, longer chains tend to form more H-H contacts per residue, and therefore reach a lower ground state energy per residue, than shorter chains. Besides the effect of the growing core region of the ground state structures, this is also related to the fact that a higher number of H residues entails that each H residue has more potential interacting partners, and there is a higher chance that one or two of the partners will be in a position that actually makes the contact possible. The increase in the number of contacts per residue with increasing chain length has been described for real proteins as well (56).

These considerations indicate that for longer model proteins, a lower H-fraction is sufficient to keep the ground state energy low enough to ensure that the protein be ordered. This leads to the observed shift of the twilight zone toward lower hydrophobicities. In accordance with this result, an early theoretical model of protein folding also found that, assuming a constant hydrophobic fraction, longer proteins (up to a large length) should be more stable (in terms of $\Delta F/L$) than shorter proteins (63), which implies that lower hydrophobicity should be sufficient to ensure marginal stability in longer proteins. The fact that experimental results do not support this finding was attributed to the narrow range of lengths of proteins with known experimental stabilities (63). We, however, suggest that the reason lies, at least partly, in the weakening of interactions in real proteins with increasing protein size (see next section).

The same considerations also explain why the twilight zone gets narrower for longer chains. The number of distinct configurations, e.g., different contact maps, of protein chains grows exponentially with chain length. Therefore, for any given residue composition, a longer chain has a greater chance to find a contact map that ensures a low ground state energy for the given sequence. And because each H-monomer has more potential interacting partners, it is easier to find a partner to form an interaction with. The number of contact maps ensuring an energetically favorable arrangement for the chain is higher, and therefore, the chain depends less on specific interactions to attain a low ground state energy. In the end, for longer chains, the fraction of sequences where the ordered state depends on specific interactions, i.e., the particular sequence, is lower than in shorter chains. In other words, the twilight zone is narrower.

The narrowing of the twilight zone with increasing chain length can also be understood by considering the distribution of energy contributions to the total ground state energy of a protein. Let E_i ($i = 1\ldots L$) denote the contribution of residue i to the total energy of a protein of length L; then the criterion for disorder can be written as $E_\mathrm{ground}/L = \langle E_\mathrm{i}\rangle_\mathrm{i} > -0.3$. If the E_is are considered random variables and are assumed to be identically distributed (a plausible assumption when sequences with a given amino acid composition are considered) and independent, then the central limit theorem applies, and it follows that the standard deviation of the mean of E_is is proportional to $1/\sqrt{L}$. In other words, the distribution (over the set of sequences with a given length and amino acid composition) of ground state energies per residue is narrower for longer chains than for shorter chains. Although the average ground state energy per residue of all sequences with a given amino acid composition and length is determined by the amino acid composition (and length), the actual ground state energy per residue of any particular sequence is (trivially) determined by the actual sequence itself. The central limit theorem entails that for longer sequences, large deviations from the average are less likely than for shorter sequences. Consequently, as chain length grows, the amino acid composition becomes more dominant in determining the ground state energy per residue, and the importance of the actual order of residues, i.e., the connectivity of the chain, will diminish.

Although this explanation of the narrowing of the twilight zone is qualitatively correct, it should be noted that the energy contributions of individual residues to the total energy are not independent of each other, and therefore, strictly speaking, the central limit theorem does not apply. Significant correlations among residue energy contributions are found. For example, in 2D lattice models, the energy contributions of residues i and $i + 3$ usually significantly correlate with each other (e.g., for HP sequences of length 10 with five hydrophobic residues, the correlation coefficient between E_2 and E_5 is 0.3, with a p-value $< 10^{-6}$) because residues i and $i + 3$ can easily form a contact. On the other hand, the energy contributions of residues i and $i + 4$ usually have a negative correlation coefficient (-0.2 between E_2 and E_6 ($p = 0.001$) for sequences specified in the previous example) because these two residues can never form a contact for geometric reasons. However, most correlations are small, and as the distance along the sequence between the two residues increases, the correlation diminishes, and its sign alternates (data not shown). Therefore, these correlations matter little as far as the sum of energy contributions is concerned, and the central limit theorem still applies in an approximate sense, entailing that the distribution of E_ground/L will still get narrower with increasing chain length.

The weakening of interactions in real proteins

Our lattice models successfully reproduced the narrowing of the twilight zone with increasing chain length that we observed with real proteins. However, we also found that the twilight zone shifts toward lower hydrophobicities and higher charges. This is not observed with real proteins. To eliminate the shift, we had to introduce length-dependent interaction energies, making the interactions weaker in longer chains.

Although a length-dependent potential may be criticized for being unphysical, it should rather be viewed as a different type of potential. Two main types of potentials are used in computational studies of proteins: semiempirical (physics-based) and statistical (knowledge-based). Physics-based energy functions are typically used for simulations with all-atom models of proteins. Their terms describe well-defined interactions and have a clear physical meaning. Knowledge-based potentials are typically used with simplified protein models and are immensely useful for many types of studies including fold recognition and protein design (64). Being derived from data sets of known protein structures, they are mean-force potentials with less clear-cut physical interpretations. In fact, they often exhibit properties that may be viewed as unphysical. In particular, it was found that knowledge-based potentials derived from small proteins differ from those derived from large proteins, with an overall weakening of residue-residue interactions with increasing protein size (65).

Based on this finding, Dehouck et al. (66) proposed a knowledge-based potential that depends on protein size, scaling roughly as $1/L$, and demonstrated that it performs better in fold recognition tests than a size-independent potential. Our length-dependent potential for 2D lattice models has a similar scaling and can be viewed as a knowledge-based potential as well: we used the known scaling of energetic properties of real proteins to impart similar behavior to our model proteins. It should be kept in mind that 2D lattice models are extremely simplified models of proteins. Our results indicate that, when used with constant interaction energy, they do not correctly reproduce the scaling of energetics of real proteins. The length-dependent energies had to be introduced to compensate for this deficiency.

The finding that knowledge-based potentials derived from proteins of different sizes indicate an overall weakening of interactions with protein size can be explained by considering the partitioning of hydrophobic and polar residues between the protein core and the surface (66). The core region of (globular) proteins grows with protein size (relative to the surface), but this is not accompanied by a corresponding increase in the fraction of hydrophobic residues. As a result, the hydrophobic core gets diluted with an increasing number of buried polar residues as protein size grows, and this leads to an overall weakening of interactions. On the other hand, as the core grows, the number of contacts per residue increases, ensuring sufficient stability for the protein even in the face of weakening per-contact interactions. In addition to the increasing polarity of the core, earlier studies also revealed other factors contributing to the weakening. Bastolla and Demetrius (56) have shown that the sequence of larger proteins is less efficiently optimized to maximize interactions, and Liang and Dill (57) have demonstrated that the overall packing density of proteins decreases with increasing size. This latter effect seems to be a general property of random polymers (67) and results in weaker Van der Waals interactions, also contributing to the weakening effect.

The original 2D HP lattice model with a constant H-H interaction energy does not exhibit these scaling properties. Because the model is on-lattice, it cannot reproduce the lowering of packing density in larger proteins; an off-lattice model would be needed for this purpose, like the one used by Zhang et al. (67). This would have to be combined with a distance-dependent energy function to make the energy dependent on packing. The sampling of sequence space, however, does pose another problem. We sampled the sequence space of HP/HPN models in a uniform way for our studies, but real proteins do not constitute a uniform sample of sequence space; for example, proteins with a high hydrophobicity are rare because of the tradeoff between stability against unfolding and stability against misfolding or aggregation (53).

In fact, most real proteins are only marginally stable, i.e., overly stable proteins are rare (68). Although we could also have progressively biased our sampling of HP/HPN sequences against proteins that are ''too stable'' against unfolding, the shape of the target distribution would have had to be somewhat arbitrarily chosen. Also, biasing the sampling this way would result in many sequences being discarded, leaving too few sequences for analysis unless we use a model with a larger alphabet. A larger alphabet, however, makes sampling computationally more demanding and requires more parameters to set. In summary, these considerations suggest that reproducing the scaling of the energetics of real proteins with a length-independent potential would require an off-lattice protein model with a distance-dependent energy function and a larger alphabet with an appropriately biased sampling of sequence space. Introducing a length-dependent, and therefore knowledge-based, potential can eliminate these complications, and the computationally easily tractable HP/HPN lattice models can be retained.

Earlier studies have shown that the hydrophobic fraction of real (ordered) proteins is roughly independent of protein size (53), and it has been argued (55) that there is an optimum, size-independent, fraction of hydrophobic residues that ensures stability against both unfolding and misfolding or aggregation. Our findings suggest that larger proteins could, in theory, be sufficiently stabilized against unfolding by a lower fraction of hydrophobic residues than smaller ones (a constant fraction would actually lead to overstabilization of large proteins), but in reality, the overall weakening of residue-residue interactions with protein size compensates for this effect, and large proteins still require the same fraction of hydrophobic residues for stability as small ones.

CONCLUSIONS

We have shown that there is a twilight zone in amino acid composition space between ordered and disordered proteins, and the width of this twilight zone decreases as chain length grows. Our model studies suggest that the existence of the twilight zone and the dependence of its width on chain length are a consequence of the intrinsic structural and energetic properties of proteins as heteropolymers and their scaling with protein size. To put it simply, amino acid composition alone does not fully determine whether a protein is ordered or disordered; this essentially depends on the order of residues and the specific interactions among them. However, with increasing chain length, the role of specific interactions diminishes, and the amino acid composition becomes sufficient to correctly classify a protein as ordered or disordered. This finding supports the suggestion that short disordered regions are more context-dependent than long ones (27).

The dependence of order/disorder on protein size also indicates that size can determine order for proteins that are in the twilight zone. More generally, the results suggest that system size is one of the factors that influence the order-disorder state of a given system, and a change in system size may elicit a disorder-to-order transition (or the other way around). The example of two-state homodimers brilliantly illustrates this point: the chains are disordered as monomers but switch to an ordered state on dimerization. Indeed, for a short chain, the twilight zone in amino acid composition space is wide, and the chain can have a sufficiently high probability to be disordered. On dimerization, the effective chain length doubles, and the twilight zone gets narrower. Although the amino acid composition remains the same, the probability of disorder may drop, and transition to an ordered state may ensue. This line of reasoning suggests that two-state behavior is easier to attain with short chains than long ones because short chains have a wider range of amino acid compositions to choose from while retaining the ability to switch states on dimerization. Indeed, two-state dimers tend to have shorter chains than three-state dimers (dimers whose monomers fold before dimerization) (see, e.g., Table 1 in Gunasekaran et al. (69)).

This insight has profound implications for disorder prediction. As we have seen on model proteins, the amount of information carried by the amino acid composition on order/disorder varies with chain length: for short chains, it only contains about half of the information needed for an accurate prediction, but it contains more than 80% of the necessary information for longer chains (Fig. 10). This finding suggests that the regions corresponding to ordered and disordered proteins in amino acid composition space cannot be separated by any method because ordered and disordered proteins in the twilight zone can have exactly the same composition. It also implies that disorder prediction methods relying on amino acid composition alone can never be sufficiently accurate for short chains or segments. The maximum possible accuracy that can be reached by such methods has probably already been reached. The kNN error rates shown in Fig. 1 are indicative of how the error rate of any amino acid composition-based algorithm should depend on chain/segment length. For a more accurate disorder prediction on short sequences, the particular order of amino acids must be taken into consideration. This can be achieved by as simple a method as evaluating dipeptide frequencies. Because specific interactions become more important in short sequences, predicting actual contacts between residues may also be a successful approach (70), although contact prediction, in the absence of a structural template, is a hard problem in itself (71,72). In the end, our results suggest that finding ways to glean more information from the sequence, rather than using ever more sophisticated classification algorithms, is the key to better prediction.

SUPPLEMENTARY MATERIAL

To view all of the supplemental files associated with this article, visit www.biophysj.org.

A. Szilágyi was supported by a Bolyai János fellowship. Funding from the Hungarian Scientific Research Fund (OTKA PD73096, 061915, T046412) and Grant GVOP-3.2.1-2004-0195/3.0 is acknowledged. The authors thank Péter Tompa and Gergely Gyimesi for critically reading the manuscript.

REFERENCES

1. Wright, P. E., and H. J. Dyson. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* 293:321–331.
2. Tompa, P. 2002. Intrinsically unstructured proteins. *Trends Biochem. Sci.* 27:527–533.
3. Dunker, A. K., J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic. 2001. Intrinsically disordered protein. *J. Mol. Graph. Model.* 19:26–59.
4. Receveur-Bréchot, V., J. Bourhis, V. N. Uversky, B. Canard, and S. Longhi. 2006. Assessing protein disorder and induced folding. *Proteins.* 62:24–45.
5. Ward, J. J., J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337:635–645.
6. Tompa, P., Z. Dosztanyi, and I. Simon. 2006. Prevalent structural disorder in *E. coli* and *S. cerevisiae* proteomes. *J. Proteome Res.* 5:1996–2000.
7. Tompa, P. 2005. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* 579:3346–3354.
8. Tompa, P., and M. Fuxreiter. 2008. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem. Sci.* 33:2–8.
9. Mészáros, B., P. Tompa, I. Simon, and Z. Dosztányi. 2007. Molecular principles of the interactions of disordered proteins. *J. Mol. Biol.* 372:549–561.
10. Singh, G. P., M. Ganapathi, and D. Dash. 2007. Role of intrinsic disorder in transient interactions of hub proteins. *Proteins.* 66:761–765.
11. Haynes, C., C. J. Oldfield, F. Ji, N. Klitgord, M. E. Cusick, P. Radivojac, V. N. Uversky, M. Vidal, and L. M. Iakoucheva. 2006. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol.* 2:e100.
12. Dosztányi, Z., J. Chen, A. K. Dunker, I. Simon, and P. Tompa. 2006. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* 5:2985–2995.
13. Xie, H., S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky, and Z. Obradovic. 2007. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* 6:1882–1898.
14. Uversky, V. N., C. J. Oldfield, and A. K. Dunker. 2005. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* 18:343–384.
15. Romero, P., Z. Obradovic, and A. K. Dunker. 2004. Natively disordered proteins: functions and predictions. *Appl. Bioinformatics.* 3:105–113.
16. Dyson, H. J., and P. E. Wright. 2005. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6:197–208.
17. Ferron, F., S. Longhi, B. Canard, and D. Karlin. 2006. A practical overview of protein disorder prediction methods. *Proteins.* 65:1–14.
18. Weathers, E. A., M. E. Paulaitis, T. B. Woolf, and J. H. Hoh. 2004. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett.* 576:348–352.
19. Coeytaux, K., and A. Poupon. 2005. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics.* 21:1891–1900.
20. Romero, P., Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker. 2001. Sequence complexity of disordered protein. *Proteins.* 42:38–48.
21. Linding, R., R. B. Russell, V. Neduva, and T. J. Gibson. 2003. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 31:3701–3708.
22. Linding, R., L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell. 2003. Protein disorder prediction: implications for structural proteomics. *Structure.* 11:1453–1459.
23. Yang, Z. R., R. Thomson, P. McNeil, and R. M. Esnouf. 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics.* 21:3369–3376.
24. Shimizu, K., Y. Muraoka, S. Hirose, K. Tomii, and T. Noguchi. 2007. Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics.* 8:78.

25. Obradovic, Z., K. Peng, S. Vucetic, P. Radivojac, and A. K. Dunker. 2005. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins.* 61(Suppl 7):176–182.

26. Li, X., P. Romero, M. Rani, A. Dunker, and Z. Obradovic. 1999. Predicting protein disorder for N-, C-, and internal regions. *Genome Inform. Ser. Workshop Genome Inform.* 10:30–40.

27. Obradovic, Z., K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, and A. K. Dunker. 2003. Predicting intrinsic disorder from amino acid sequence. *Proteins.* 53(Suppl 6):566–572.

28. Melamud, E., and J. Moult. 2003. Evaluation of disorder predictions in CASP5. *Proteins.* 53(Suppl 6):561–565.

29. Peng, K., P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic. 2006. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics.* 7:208.

30. Dosztányi, Z., V. Csizmók, P. Tompa, and I. Simon. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* 347: 827–839.

31. Uversky, V. N., J. R. Gillespie, and A. L. Fink. 2000. Why are ''natively unfolded'' proteins unstructured under physiologic conditions? *Proteins.* 41:415–427.

32. Prilusky, J., C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman, and J. L. Sussman. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics.* 21:3435–3438.

33. Oldfield, C. J., Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky, and A. K. Dunker. 2005. Comparing and combining predictors of mostly disordered proteins. *Biochemistry.* 44:1989–2000.

34. Garbuzynskiy, S. O., M. Y. Lobanov, and O. V. Galzitskaya. 2004. To be folded or to be unfolded? *Protein Sci.* 13:2871–2877.

35. Shakhnovich, E. I. 1999. Folding by association. *Nat. Struct. Biol.* 6:99–102.

36. Levy, Y., P. G. Wolynes, and J. N. Onuchic. 2004. Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. USA.* 101:511–516.

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

38. Lau, K. F., and K. A. Dill. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules.* 22:3986–3997.

39. Dill, K. A., S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. 1995. Principles of protein folding–a perspective from simple exact models. *Protein Sci.* 4:561–602.

40. Chan, H. S., and K. A. Dill. 1990. Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. USA.* 87:6388–6392.

41. Sindelar, C. V., Z. S. Hendsch, and B. Tidor. 1998. Effects of salt bridges on protein structure and design. *Protein Sci.* 7:1898–1914.

42. Kaffe-Abramovich, T., and R. Unger. 1998. A simple model for evolution of proteins towards the global minimum of free energy. *Fold. Des.* 3:389–399.

43. Harrison, P. M., H. S. Chan, S. B. Prusiner, and F. E. Cohen. 2001. Conformational propagation with prion-like characteristics in a simple model of protein folding. *Protein Sci.* 10:819–835.

44. Dima, R. I., and D. Thirumalai. 2002. Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics. *Protein Sci.* 11:1036–1049.

45. Vucetic, S., Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L. M. Iakoucheva, M. S. Cortese, J. D. Lawson, C. J. Brown, J. G. Sikes, C. D. Newton, and A. K. Dunker. 2005. DisProt: a database of protein disorder. *Bioinformatics.* 21:137–140.

46. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

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

48. Irback, A., and C. Troein. 2002. Enumerating designing sequences in the HP model. *J. Biol. Phys.* 28:1–15.

49. Chan, H. S., and K. A. Dill. 1993. Energy landscapes and the collapse dynamics of homopolymers. *J. Chem. Phys.* 99:2116–2127.

50. Kou, S. C., J. Oh, and W. H. Wong. 2006. A study of density of states and ground states in hydrophobic-hydrophilic protein folding models by equi-energy sampling. *J. Chem. Phys.* 124:244903.

51. Shannon, C. E. 1948. A mathematical theory of communication. *Bell Labs Technical Journal* 27:379–423.

52. Fink, A. L. 2005. Natively unfolded proteins. *Curr. Opin. Struct. Biol.* 15:35–41.

53. Sandelin, E. 2004. On hydrophobicity and conformational specificity in proteins. *Biophys. J.* 86:23–30.

54. Realini, C., S. W. Rogers, and M. Rechsteiner. 1994. KEKE motifs. Proposed roles in protein-protein association and presentation of peptides by MHC class I receptors. *FEBS Lett.* 348:109–113.

55. Miao, J., J. Klein-Seetharaman, and H. Meirovitch. 2004. The optimal fraction of hydrophobic residues required to ensure protein collapse. *J. Mol. Biol.* 344:797–811.

56. Bastolla, U., and L. Demetrius. 2005. Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. *Protein Eng. Des. Sel.* 18:405–415.

57. Liang, J., and K. A. Dill. 2001. Are proteins well-packed? *Biophys. J.* 81:751–766.

58. Zhang, Y., B. Stec, and A. Godzik. 2007. Between order and disorder in protein structures: analysis of ''dual personality'' fragments in proteins. *Structure.* 15:1141–1147.

59. Wittung-Stafshede, P. 2002. Role of cofactors in protein folding. *Acc. Chem. Res.* 35:201–208.

60. White, S. H. 1992. Amino acid preferences of small proteins. Implications for protein stability and evolution. *J. Mol. Biol.* 227:991–995.

61. Pande, V. S., A. Y. Grosberg, and T. Tanaka. 1997. Statistical mechanics of simple models of protein folding and design. *Biophys. J.* 73:3192–3210.

62. Chan, H. S., and K. A. Dill. 1991. Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.* 20:447–490.

63. Dill, K. A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry.* 24:1501–1509.

64. Sippl, M. J. 1995. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 5:229–235.

65. Rooman, M., and D. Gilis. 1998. Different derivations of knowledge-based potentials and analysis of their robustness and context-dependent predictive power. *Eur. J. Biochem.* 254:135–143.

66. Dehouck, Y., D. Gilis, and M. Rooman. 2004. Database-derived potentials dependent on protein size for in silico folding and design. *Biophys. J.* 87:171–181.

67. Zhang, J. F., R. Chen, C. Tang, and J. Liang. 2003. Origin of scaling behavior of protein packing density: A sequential Monte Carlo study of compact long chain polymers. *J. Chem. Phys.* 118:6102–6109.

68. Taverna, D. M., and R. A. Goldstein. 2002. Why are proteins marginally stable? *Proteins.* 46:105–109.

69. Gunasekaran, K., C. Tsai, and R. Nussinov. 2004. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J. Mol. Biol.* 341:1327–1341.

70. Schlessinger, A., M. Punta, and B. Rost. 2007. Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics.* 23:2376–2384.

71. Kundrotas, P. J., and E. G. Alexov. 2006. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. *BMC Bioinformatics.* 7:503.

72. Horner, D. S., W. Pirovano, and G. Pesole. 2008. Correlated substitution analysis and the prediction of amino acid structural contacts. *Brief. Bioinform.* 9:46–56.