Protein Science (1993), 2, 2050–2065. Cambridge University Press. Printed in the USA.

Local and nonlocal interactions in globular proteins and mechanisms of alcohol denaturation

PAUL D. THOMAS[1] AND KEN A. DILL[2]

[1] Graduate Group in Biophysics, University of California, San Francisco, California 94143-0448
[2] Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143-1204

(RECEIVED May 15, 1993; ACCEPTED October 1, 1993)

Abstract

How important are helical propensities in determining the conformations of globular proteins? Using the two-dimensional lattice model and two monomer types, H (hydrophobic) and P (polar), we explore both nonlocal interactions, through an HH contact energy, ϵ, as developed in earlier work, and local interactions, through a helix energy, σ. By computer enumeration, the partition functions for short chains are obtained without approximation for the full range of both types of energy. When nonlocal interactions dominate, some sequences undergo coil–globule collapse to a unique native structure. When local interactions dominate, all sequences undergo helix–coil transitions. For two different conformational properties, the closest correspondence between the lattice model and proteins in the Protein Data Bank is obtained if the model local interactions are made small compared to the HH contact interaction, suggesting that helical propensities may be only weak determinants of globular protein structures in water. For some HP sequences, varying σ/ϵ leads to additional sharp transitions (sometimes several) and to "conformational switching" between unique conformations. This behavior resembles the transitions of globular proteins in water to helical states in alcohols. In particular, comparison with experiments shows that whereas urea as a denaturant is best modeled as weakening both local and nonlocal interactions, trifluoroethanol is best modeled as mainly weakening HH interactions and slightly enhancing local helical interactions.

Keywords: alcohol denaturation; compact conformations; helical propensities; HP lattice model; hydrophobic interaction

What is the relative importance of local interactions (helical propensities) compared to nonlocal interactions (mainly solvent-mediated and hydrophobic interactions) in determining the native structures of globular proteins? One view has held that the local interactions are the main determinants of structure, while the nonlocal interactions just provide nonspecific stability to the compact state (Anfinsen & Scheraga, 1975). That is, interactions among adjacent and near-neighbor peptide units tend to drive proteins to configure into helices at certain points in the sequence, which ultimately end up as helices in the native structure. In this view, helices should be identifiable by factors that are local within the sequence, and this should strongly specify how they can pack to form the native structure. Recently, an alternative view has developed that nonlocal interactions may be a major determinant not only of the stabilities but also of the structures of globular proteins (Dill, 1990; Chan & Dill, 1991b). In this view, a major factor in determining where helices form in native structures is the sequence positions of hydrophobic monomers. That is, predicting helices (and also sheets) in native structures is more a matter of finding the correct global hydrophobicity patterns than of finding good local helical propensity patterns, even though the latter are known to be non-negligible. The present work aims to explore this question in more detail by using a simplified model that has an exact partition function, and to study the consequences of different balances between local and nonlocal interactions. We refer to this model as the "helical-HP model."

We then apply the model to the denaturation of proteins by alcohols and other agents. The ability of alcohols to induce α-helical conformations in proteins was first noted in optical rotatory dispersion experiments by Tanford et al. (1960) on β-lactoglobulin. Tamburro et al. (1968) studied the effects of trifluoroethanol (TFE) on the conformations of the ribonuclease *S*-peptide; TFE was found to stabilize the small peptide in the same (α-helical) confor-

Reprint requests to: Ken A. Dill, Department of Pharmaceutical Chemistry, Box 1204, University of California, San Francisco, California 94143-1204.

mation that it adopts in the native protein. Since then, alcohols have been used widely to examine the conformational (particularly helical) propensities of peptides (Nelson & Kallenbach, 1986; Lehrman et al., 1990; Segawa et al., 1991; Sönnichsen et al., 1992) and to induce conformational changes in intact proteins (Stone et al., 1985; Wilkinson & Mayer, 1986; Dufour & Haertlé, 1990; Jackson & Mantsch, 1992; Buck et al., 1993; Fan et al., 1993). The primary physical mechanisms by which alcohols effect these changes are still unresolved. The helical-HP model reproduces some basic characteristics of experimental denaturation of proteins by alcohols and other denaturants, and offers an interpretation of their mechanisms of action.

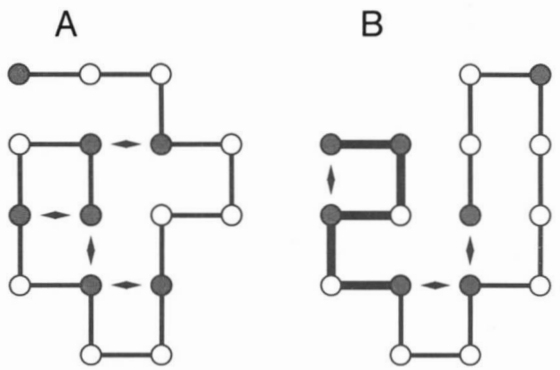

Fig. 1. 2D helical-HP model conformations, for a sample sequence (HHPHPHPPHPPPHPPH). H (hydrophobic) residues are filled; P (polar) residues are open. Conformation A contains four HH contacts (black diamonds) and no helical bonds, $\Delta G = 4\epsilon$. Conformation B contains three HH contacts and two helical bonds (bold lines), $\Delta G = 3\epsilon + 2\sigma$.

The helical-HP model

The helical-HP model is of short chains, which are specific sequences of H (hydrophobic) and P (polar) monomers, configured as self-avoiding walks on two-dimensional (2D) lattices. A monomer represents a single amino acid residue. Figure 1 shows two possible conformations of a 16-monomer chain on the lattice. Each monomer occupies one site on the lattice, and monomers that are consecutive in the sequence must be next to each other on the lattice (chain connectivity constraint). No two monomers may occupy the same lattice site (excluded volume constraint). We use a search-tree algorithm to enumerate all of the possible conformations of a chain of given length (Chan & Dill, 1989). For the 16-monomer chains modeled in this paper, there are 802,075 possible conformations on the 2D square lattice. The relative populations, $p(\mathrm{A})/p(\mathrm{B})$, of any two conformations A and B are given by the Boltzmann distribution:

$$\frac{p(\mathrm{A})}{p(\mathrm{B})} = \frac{e^{-\Delta G_\mathrm{A}/kT}}{e^{-\Delta G_\mathrm{B}/kT}}, \qquad (1)$$

where ΔG_A and ΔG_B represent the Gibbs free energies of conformations A and B, respectively, relative to a common reference, k is Boltzmann's constant, and T is the absolute temperature.

In a previous treatment, referred to as the HP model (Lau & Dill, 1989; Chan & Dill, 1991a; Shortle et al., 1992), a single type of energy was considered, accounting for the favorability of each HH contact in a given chain conformation. This interaction is nonlocal in the sense that HH contacts involve two H monomers that need not be near neighbors in the sequence; they are mediated through space rather than being mediated along neighboring covalent bonds. Conformational properties of the HP model have been studied in detail. Even though the model chains are short and configured in two dimensions, nevertheless the HP model shows certain features of globular proteins (Lau & Dill, 1989, 1990; Chan & Dill, 1991a; Lipman & Wilbur, 1991; Miller et al., 1992; Shortle et al., 1992; Camacho & Thirumalai, 1993a,b; O'Toole & Panagiotopoulos, 1993; Unger & Moult, 1993). For example, increasing the strength of the HH interaction in the HP model leads to a relatively sharp transition from a denatured ensemble of open conformations to a small number of compact conformations (often only one or a few) that are composed of secondary structures and have cores mainly composed of H monomers. The conformations of lowest energy in the HP model are different for different sequences. Native states are relatively insensitive to mutations, particularly on the surface; there is high "convergence," i.e., many different sequences fold to a given native state; and there are favored folding pathways. Thus, the simple HP lattice model offers a useful framework for studying the physical bases of protein properties that are currently difficult to study by other types of models.

We now consider a more general model, the helical-HP model, in which the free energy is a sum of two types of energy. First, as before, we consider a contact interaction, $\epsilon \leq 0$, for each HH contact. The net exposed surface of H monomers in the 2D model protein (i.e., contacting either P monomers or solvent) is decreased by two perimeter units (the 2D equivalent of surface area) for every contact made between two H monomers. Physically, the HH contact primarily represents the burial, or desolvation, of nonpolar amino acid side chains. HH or "hydrophobic" contacts is our shorthand notation for a more complex reality, and represents all the factors involved in the transfer of each nonpolar side chain from water to its specific contact environment, including water ordering, dispersion interactions, and hydrogen bonding, and thus differs from some other definitions. If m is the number of HH contacts in a model conformation, then the total nonlocal contact energy of that conformation is $m\epsilon$. (Because we are not concerned in this paper with the temper-

ature dependence of the model, we interchangeably refer to contact energies as contact free energies.) It should also be noted that although the HH contacts need not be local in sequence, they may be. In this way, a 2D helix can be stabilized by both intrahelix $(i, i+3)$ HH contacts and HH contacts between monomers in helices and monomers in other parts of the chain (e.g., in Fig. 1B).

Second, we introduce an additional energy term to account for local interactions. Local interactions are those among near-neighboring residues in the sequence. Local interactions drive turns (Dyson & Wright, 1991) and α-helices in peptides (Zimm & Bragg, 1959), and are the predominant sites of hydrogen bonding (Stickle et al., 1992). The helical-HP model represents local interactions with an energy that favors helix formation. The model helical interaction also represents more than a single type of atomic interaction: several factors are involved in helix formation, including intramolecular $(i, i+4)$ hydrogen bonds and steric and dispersion interactions. In this regard, the distinction between local and nonlocal interactions is not a distinction between hydrogen bonding and water ordering effects; it is a distinction mainly between helical propensities and nonpolar transfers. Two turns of a 2D lattice helix are shown in Figure 1B. The reason this particular lattice conformation is considered to be helical is because it is the only conformation on the 2D square lattice that has the same topology (i.e., contact map) as a three-dimensional α-helix (Chan & Dill, 1989). When six consecutive monomers of a 2D chain are in such a structure, a helical energy of 2σ is assigned to the conformation, one contribution of σ for each of the two $(i, i+3)$ contacts. The minimum length of a helix is defined to be six residues, i.e., two helical bonds, since a single $(i, i+3)$ contact is more properly defined as a turn. We refer to each helical $(i, i+3)$ contact in the model as a "helical bond." A helical bond is assumed to be favorable, i.e., $\sigma \leq 0$, independently of the HP sequence. The sequence independence of the model helical interaction is an approximation based on the experimental observation that, over the 20 amino acids, helical propensities vary relatively little in energy compared to hydrophobicity. Helical propensities in peptides vary by approximately a factor of 8, ranging over the amino acids, excluding proline (Lyu et al., 1990; O'Neil & DeGrado, 1990; Scholtz & Baldwin, 1992), whereas hydrophobic interactions vary by a factor of 100 or more (Nozaki & Tanford, 1971; Chothia, 1976; Fauchère & Pliska, 1983; Rose et al., 1985). It would be straightforward to include a sequence dependence of helicity in this model, but it only adds additional parameters that would obscure the purpose of the present study. Every added helical unit (two added residues) adds σ to the energy sum; the local energy is then $n\sigma$ for a conformation having n helical bonds.

Thus, the total energy of a conformation is:

$$\Delta G = m\epsilon + n\sigma \tag{2}$$

relative to an open reference conformation that has no HH contacts and no helical structure. This energy accounts for the intrachain interactions; the conformational entropy is treated through the enumeration process. For a given sequence of 16 H and P monomers, we evaluate the free energy of each of the 802,075 conformations according to Equation 2. The "native state" of a sequence is defined as the conformation (or conformations) that has the lowest free energy for particular values of σ and ϵ. To illustrate the energy contributions, consider the conformations shown in Figure 1. Conformation A contains four HH contacts but no helical bonds; it has energy $\Delta G_A = 4\epsilon$. Conformation B contains three HH contacts and two helical bonds; it has energy $\Delta G_B = 3\epsilon + 2\sigma$. The population ratio of conformer A relative to conformer B will be

$$\frac{\exp(4\epsilon/kT)}{\exp[(3\epsilon + 2\sigma)/kT]}. \tag{3}$$

The strengths of the two types of interaction are varied by changing the values of σ and ϵ. Conformations A and B will be present in equal populations, in this example, when $\Delta G_A = \Delta G_B$, i.e., when $\sigma/\epsilon = 1/2$. Relative to this value, if σ/ϵ is decreased, conformation A will be favored; if σ/ϵ is increased, B will be more populated. Thus, in general, the native conformation(s) of any sequence will depend on the ratio σ/ϵ, i.e., on the balance between local and nonlocal interactions. Likewise, different native states for the same sequence will be present in equal populations at integral fractional values of σ/ϵ.

The fractional population of any single conformation can be calculated for given values of σ and ϵ as the Boltzmann weight for that conformation, divided by the partition function, Z (the sum of the Boltzmann factors of all possible conformations):

$$f(\text{conf}) = \frac{e^{-[(m_{\text{conf}}-m_{\text{native}})\epsilon+(n_{\text{conf}}-n_{\text{native}})\sigma]}}{\sum_{m=0}^{m_{\max}}\sum_{n=0}^{n_{\max}} g(m,n)\cdot e^{-[(m-m_{\text{native}})\epsilon+(n-n_{\text{native}})\sigma]}}, \tag{4}$$

where the subscripts "native" and "conf" refer to the properties of the native conformation (for that particular ratio of σ/ϵ) and the conformation of interest, respectively; m is the number of HH contacts; n is the number of helical bonds; $m_{\max}$ is the maximum possible number of HH contacts for the given sequence; and $n_{\max}$ is the maximum possible number of helical turns for a given chain length. The degeneracy function $g(m,n)$ specifies the number of conformations in the ensemble having particular values of m and n. Our convention is that the native state has a Boltzmann weight of 1. The fractional population, and therefore the stability, of any conformation can be calculated exactly for any helical-HP sequence at any values of σ and ϵ.

In physical terms, the quantities ϵ and σ represent the degrees to which external factors, such as temperature and solvents, determine: (1) the affinity between two nonpolar amino acids, and (2) the propensity to form helices, respectively. For example, as noted in more detail below, alcohols such as ethanol and TFE in water weaken the hydrophobic interaction and strengthen the helical propensity. Increasing concentrations of guanidine hydrochloride (GuHCl) and urea in water weaken both types of interaction (Robinson & Jencks, 1965; Nandi & Robinson, 1984). GuHCl can denature helical peptides (Shoemaker et al., 1988), implying that increasing its concentration in water would correspond, in our model, to a decrease in both ϵ and σ.

Predictions of the model

First, consider two limiting cases. When the contact interactions are weak (ϵ small), the model describes helix–coil processes: all sequences become helical as σ becomes large. For $\epsilon = 0$ when $\sigma < 0$, the helix is the conformation of minimum energy and therefore defines the native state (i.e., the conformation of minimum free energy) for all sequences. On the other hand, when there are no local interactions ($\sigma = 0$), the helical-HP model reduces to the HP model explored in previous studies (Lau & Dill, 1989, 1990; Chan & Dill, 1991a; Lipman & Wilbur, 1991; Miller et al., 1992; Shortle et al., 1992; Camacho & Thirumalai, 1993a,b; O'Toole & Panagiotopoulos, 1993; Unger & Moult, 1993).

The two types of interaction, local and nonlocal, can cooperate to stabilize a given conformation. Figure 2 shows, for two different HP sequences, the effect of strengthening the HH attraction (i.e., making ϵ more negative) for a constant small helical energy (σ). The HH contact interaction helps to increase the stability (i.e., the fractional population) of the 16-residue helix conformation of the sequence labeled A, but decreases the helicity for sequence B. For sequence A, the helical conformation contains several HH contacts, and increasing the energy associated with these contacts stabilizes the helical conformation. Sequence B in a helix makes only one HH contact, so increasing ϵ only destabilizes the helix in favor of the many other conformations that have more HH contacts. Experiments on helical peptides show a similar result: helices are stabilized if their formation results in the burial of nonpolar surface (Tanford, 1968; Chou et al., 1972; Richards & Richmond, 1978). Curve A also shows that when a helix is stabilized by HH interactions, further strengthening of the HH interactions can ultimately lead to transitions to even more stable conformations having more HH contacts and fewer helical bonds.

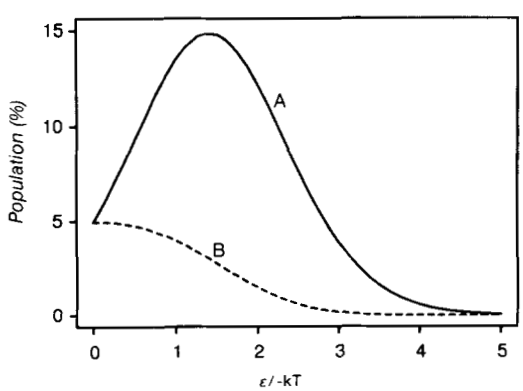

Fig. 2. Population of helical conformer versus HH interaction, ϵ, for sequences A (HPPHHPHHPHPPHHHH) and B (HHPHPPPHPHH PHPHH). For sequence A, the helix is first stabilized by favorable HH interactions, but then is destabilized as chain compactness becomes more important. Sequence B is destabilized by HH attractions. Populations are calculated by Equation 4; for both cases, $\sigma = -2.25kT$.

For a given HP sequence, the ratio σ/ϵ determines which conformation(s) are native. Figure 3 shows examples of "native state phase diagrams," i.e., maps of the minimum-energy states, and the transitions among them, for different values of σ and ϵ. The conformations shown on the phase diagrams are the most populated (native) conformations, but other conformations are also present in lesser concentrations, depending on their Boltzmann weights. In the lower left corner of each diagram, which corresponds to both interactions being weak, all conformations have nearly the same populations. This region corresponds to the denatured state of the chain. As the interaction energies are increased (by increasing the distance from the origin), the native conformation(s) become more populated. Increasing both interaction energies, from the origin along a line at angle θ with respect to the x-axis of these phase diagrams, corresponds to a fixed balance of energies $\sigma/\epsilon = \tan\theta$. Transition points (dotted lines) are points of equal populations of different native states, and occur at integral fractions of σ/ϵ in the lattice model.

The model predicts "conformational switching." This follows from the existence of transition lines in the phase diagrams. That is, in some particular solvent (i.e., particular values of σ and ϵ) the sequence has one native state, whereas changing the solvent causes a transition to a different native state. Thus, the model can flip-flop between two different states with changes in external conditions. At any given value of σ/ϵ there may be more than one native conformation with the same number, m, of HH contacts and the same number, n, of helical bonds (e.g., in Fig. 3C).

How many HP sequences fold to a *unique* native conformation? It depends on the relative interaction strengths, σ/ϵ. By "unique" we mean that only one conformation, out of 802,075 possible, has the minimum free energy. We find the native conformations of each possible sequence of HP copolymers of length 16, over all possible values of σ/ϵ. The number of distinct sequences (i.e., the size of

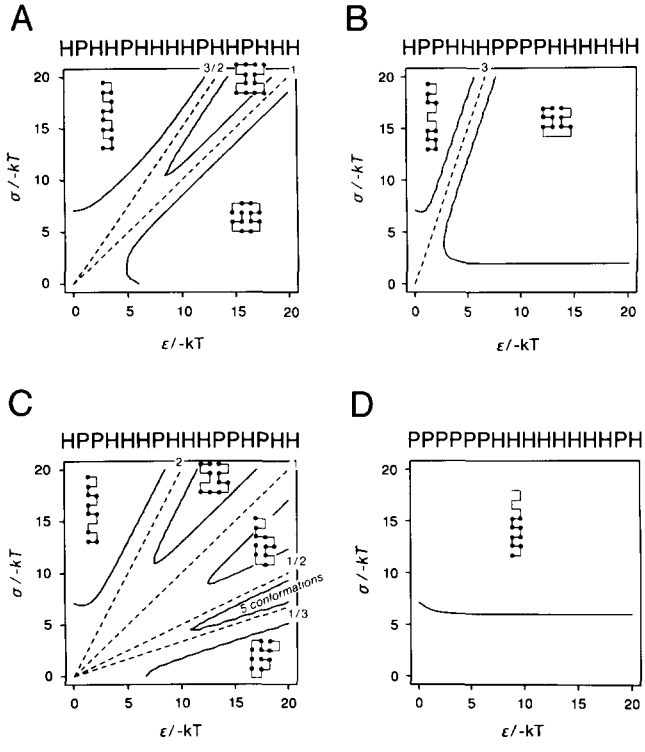

Fig. 3. Native state phase diagrams versus σ and ϵ. Open regions show which 2D lattice conformations are native for the given sequence for particular ranges of σ and ϵ. H residues in the sequence are shown as black dots. Dashed lines indicate transitions, and are labeled with the transition values of σ/ϵ; solid lines enclose the regions where the conformation(s) shown comprise over 95% of the total population (by Equation 4). For small σ and ϵ, there is a large ensemble of denatured conformations. For large σ and ϵ, sequence A undergoes two conformational switching transitions as σ/ϵ is increased. Sequence B shows only a single transition (at high σ and ϵ) between helix and helix bundle; this sequence has no unique native conformation at $\sigma/\epsilon = 0$ for any value of ϵ. Sequence C has many transitions, one to a nonunique native state favored for $1/3 < \sigma/\epsilon < 1/2$. Sequence D has only a helix-coil transition, because it cannot form a core of contacting H residues. All sequences undergo helix-coil transitions. Some undergo coil-globule, globule-globule, or globule-helix transitions.

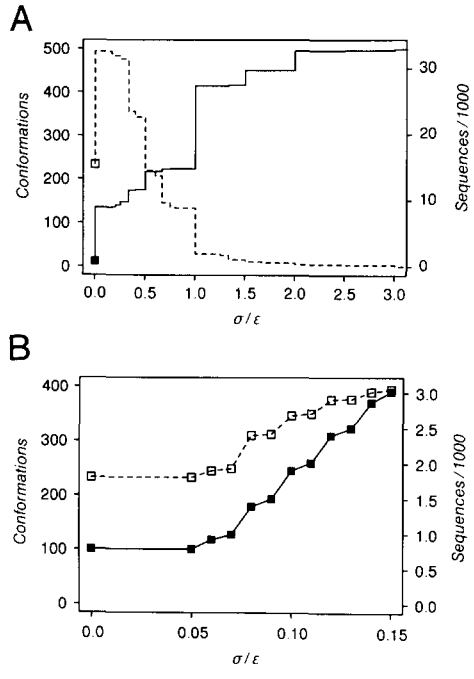

Fig. 4. Uniqueness of native structure versus σ/ϵ. **A:** Solid line (right scale) is the number of sequences in all of sequence-space that fold to unique native states (size of single-sequence set). It increases with σ/ϵ because all sequences fold to a helix for large σ. Dashed line (left scale) is the number of different conformations to which sequences fold (size of single-conformation set), which decreases to 1 for large σ. These curves are discontinuous because transitions between native states occur at a finite number of integral fractions of σ/ϵ. **B:** With "minimum stability" criteria, which make the low-σ/ϵ end of each curve more continuous. Note that these criteria are applied only for $\sigma/\epsilon \leq 0.15$. New sequences are added to the list of proteinlike sequences when the unique native conformation becomes stable (see text), so for this range both curves are increasing.

"sequence-space") is 32,896 if we take into account mirror-image equivalence (HP chains have no polarity, e.g., the tetramer sequence PHPH is equivalent to HPHP). Figure 4A shows the number of sequences that fold to a unique native structure. When local forces are dominant (σ/ϵ is large), all sequences fold to a unique conformation, namely the 16-residue helix. When HH contact interactions are dominant, fewer sequences map to a greater number of different unique native states.

When helical interactions are large, but not so large that they completely overwhelm the HH interactions, many helical-HP model sequences will form "helical bundles": two helices packed side by side antiparallel to each other with a common "core" of contacting H residues. (The two-helix bundle is the two-dimensional equivalent of a bundle of multiple helices in three dimensions.) Physically, these intermediate values of σ/ϵ correspond to the conditions present in the interior of cell membranes. The nonpolar environment of the membrane interior would lead to a reduced hydrophobic driving force and increased intramolecular hydrogen-bonding driving force (because there will be no solvent hydrogen bonds available in the membrane interior) relative to water. The membrane-spanning regions of many integral membrane proteins are α-helical. These helices associate into helical bundles, with the more hydrophobic groups exposed to the membrane (Rees et al., 1989), suggesting that the hydrophobic interactions between side chains are less favorable than interactions with the acyl chains of membrane molecules. In our model, this situation corresponds to a contact interaction between P monomers, which is equivalent to the HH interaction (by symmetry). Thus, the helical-HP model finds a helical bundle conformation to be the native state for many sequences under "membrane-like" conditions.

This result, that many helical-HP sequences fold to helical bundles for intermediate values of σ/ϵ, also has an

implication for designing protein folding algorithms. It implies that by choosing helical propensities too large, not only will helical bundle proteins fold into helical bundles but also many other sequences will incorrectly fold to helical bundles. Conversely, our results suggest that if a folding algorithm incorrectly folds most sequences to helical bundles, the helical propensities have been chosen to be too large.

How often does conformational switching involve a transition between unique native conformations, i.e., from one unique conformation to another unique conformation? The sequence shown in Figure 3A has a unique native conformation at all values of σ/ϵ (except exactly at transition points) when both types of interaction are strong. This type of behavior turns out to be quite general: more than half of the sequences that have a unique native conformation at one value of σ/ϵ will also have a unique but different native conformation if σ/ϵ is increased through a transition point. The chain does not pass through a disordered ensemble of conformations between the two states, but rather undergoes a simple two-state transition, toggling between states in response to appropriate changes in external conditions. These sequences are "conformational switches." For other sequences, the behavior can be more complex. Experimentally, peptides have been designed that can act as conformational switches, changing conformation in response to changes in external conditions (Mutter & Hersperger, 1990). The polypeptide gramicidin A has been shown to undergo a conformational change from double-helical dimer to helical monomer upon insertion into a cell membrane (Killian, 1992). For many HP sequences, the helical-HP model predicts a globule–helix transition for such a change in external conditions, i.e., a large increase in σ and a decrease in ϵ.

Comparison of the helical-HP model with properties of proteins

Our aim in this section is to compare properties of the helical-HP lattice model with properties of native protein structures from the Brookhaven Protein Data Bank (PDB), in order to determine which relative strengths of local (helical) and nonlocal (HH contact) interactions cause the model to resemble real proteins most closely. To do so, we must first choose which model sequences are most "proteinlike." In this regard, we consider only unique native states predicted by the model, since native states of real proteins are unique, at least to low resolution. In other words, if over a certain range of σ/ϵ a given helical-HP sequence has only a single conformation having the minimum energy (according to Equation 2), it will be considered to be proteinlike over that range; if it has multiple (degenerate) native conformations, it will not. We construct a set of all proteinlike sequences for each range of σ/ϵ (see Fig. 4). Because the native states of each helical-HP sequence vary with σ/ϵ, different balances of local and nonlocal interactions yield different sets of unique native conformations. We aim to find the balance of σ/ϵ in our model proteins that most closely mimics aqueous solvent conditions for real proteins.

The native state of a helical-HP sequence is defined as the conformation(s) of lowest energy. On either side of a transition point, there will be a different native state. But only in the limit of infinite interaction energies will there be an exact transition point between stable states. As is evident in Figure 3, for finite σ and ϵ there is a range of σ/ϵ surrounding each transition point in which the native state is not highly stable; the size of this range depends on the magnitude of σ and ϵ relative to kT. In addition to the requirement that a particular helical-HP sequence have a unique native conformation to be considered proteinlike, we add a stability requirement. We choose $\epsilon = -9.5kT$ and a fractional population of 95% (by Equation 4), which we refer to as our "minimum stability" criteria. Thus, for example, the helical-HP sequence shown in Figure 3B would not meet the minimum stability requirement for approximately $\sigma/\epsilon < 0.2$. Figure 4B plots the number of sequences that are proteinlike on this basis, for small σ/ϵ.

For comparison with real proteins, we choose a representative sample of 113 proteins from the PDB, derived from Appendix 3 of *Protein Architecture* (Lesk, 1991). Because we do not know how the known proteins "sample" all the possible protein sequences and structures, we consider the two possible unbiased ways to represent sets of model conformations. The "single-conformation" set contains each unique native conformation once. That is, no matter how many sequences fold to a given conformation, it is represented once. By contrast, in the "single-sequence" set, each sequence that folds to a unique native conformation is represented once; so a given conformation is represented in proportion to the number of sequences that have that same conformation. Hence, the single-sequence set is larger than the single-conformation set, and grows with increasing σ/ϵ (Fig. 4). Probably neither of these sequence ensembles truly represents the proteins in the PDB sample, but they provide two unbiased limiting cases. Our PDB sample is not a single-conformation set because nearly identical native structures often appear more than once, such as the globins, due to different sequences that have essentially the same structures. It is therefore probably closer to being a single-sequence set, but it too may be a biased sample of all proteins, because the PDB contains mostly relatively small, crystallizable molecules.

To compare the model to real protein structures, we examine two properties: (1) the ratio of helix to sheet secondary structure in each set, and (2) the distribution of tertiary structural similarities, according to a property defined below. We choose these two properties because they are the most sensitive tests of the balance between the two

types of energies, and because their distributions do not involve artifacts that might arise from cross comparisons of lattice model and real protein structures. Rather, we obtain a distribution function for the lattice model and one for real proteins, and then compare the two distribution functions.

Secondary structures in proteins

Lattice model studies show that the formation of helices and sheets can be driven by compactness, for example, as a consequence of HH interactions (Chan & Dill, 1990). For many sequences, the HP model leads to native states with much secondary structure. However, while the 2D lattice model predicts amounts of helix and sheet similar to those found in the Protein Data Bank (Chan & Dill, 1990), recent studies that are not confined to the discrete bond angles of a lattice show that the secondary structures driven by compactness alone are structurally diverse. Without hydrogen bonding, only a small fraction of the conformations that have the helix topology ($(i, i+3)$ contact repeats) are specifically α-helices (Gregoret & Cohen, 1991; Hao et al., 1992; D. Yee, H.S. Chan, T. Havel, & K.A. Dill, in prep.). Compactness is predicted to give stability, but not structural specificity, to secondary structures in globular proteins. It is not yet known exactly how much of the free energy stabilizing α-helices derives from compactness versus hydrogen bonding; it depends strongly on what criteria are used to distinguish an α-helix from a nonhelical conformation. The problem is substantial: the amount of α-helix predicted by different published helical criteria can vary from 3 to 50% for a given chain conformation (Yee et al., in prep.). If loose criteria are used to define helices, then compactness can account for most of their driving force, but if helices are defined by stringent criteria, then hydrogen bonding must also be invoked to account for their observed populations.

In the present study we compare secondary structures from the 2D lattice model with those from proteins in the Protein Data Bank. Because of the difficulties noted above, we focus here on the shapes and relative areas of distribution functions rather than on absolute numbers of secondary structures, because the latter are so sensitive to subjective criteria. Whereas there are ambiguities in defining real protein secondary structures, there is no ambiguity in defining them on 2D square lattices. The definitions of the 2D secondary structure types follow from properties of the contact map and are given in Figure 5 (Chan & Dill, 1990).

What balance between local and nonlocal interactions is needed to bring the model into closest correspondence with real proteins? This is addressed in Figures 6 and 7. For comparison, Figure 6A shows the length distributions of protein helices and sheets derived from the PDB by Kabsch and Sander (1983). The corresponding model distributions are shown in Figure 6B and C, for different values of σ/ϵ. It has been noted before (Dill, 1990) that if helical propensities were the dominant forces responsible for the structures in globular proteins, proteins would have more long helices and fewer short ones: helix stability should increase with length. This prediction is seen in the model distributions shown in Figure 6B: when local helical interactions are significant ($\sigma/\epsilon > 1/2$), more helices are longer and fewer are shorter in the native states of the model. But when there is less helical driving force ($\sigma/\epsilon < 1/6$), the model shows a monotonically decreasing helicity with length, as is observed in proteins in the PDB. Hence, the distribution functions in Figure 6 have shapes most similar to those for real proteins if the free energy to form one helical bond is from approximately 0 to 15% of the free energy involved in forming one HH contact.

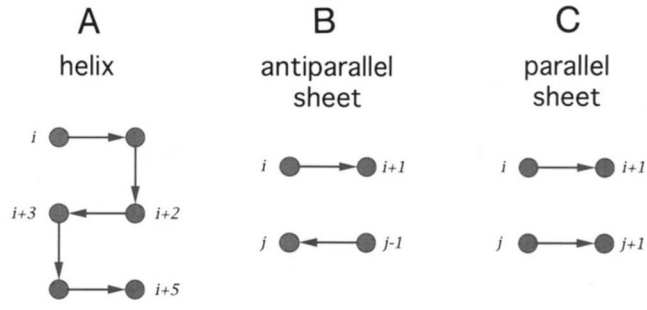

Fig. 5. Secondary structures on the 2D square lattice. **A:** Helix, at least two sequential noncovalent contacts between residues $[(i, i+3), (i+2, i+5), \ldots, (i+2n, i+2n+3)]$. **B:** Antiparallel sheet $[(i, j), (i+1, j-1), \ldots, (i+n, j-n)]$. **C:** Parallel sheet $[(i, j), (i+1, j+1), \ldots, (i+n, j+n)]$.

Figure 7 shows how the weighted areas under these distribution functions, which equal the total number of residues participating in each type of secondary structure, depend on σ/ϵ. The use of the area under the curve helps reduce arbitrariness of decisions about whether helices observed in real proteins are continuous or broken, and it helps compensate for the limited data due to the shortness of the chains in the lattice model. Figure 7A and B shows the helix/antiparallel sheet ratio and the helix/parallel sheet ratio, respectively, as functions of σ/ϵ in the model. The value for the Kabsch and Sander (1983) PDB study is shown as a horizontal line. Both the single-sequence and the single-conformation sets are shown in each figure; they show significant agreement, and mainly differ only in magnitude.

From Figure 7 it is evident that $0.05 < \sigma/\epsilon < 0.1$ is the range for which the model most accurately reproduces the helix/parallel sheet and the helix/antiparallel sheet ratios of proteins in the PDB. One notable feature of the model curves is the large step increase in the helix/sheet ratio for even the smallest increment in σ/ϵ above zero. This arises because when even a small nonzero local term is added to the free energy, it breaks the degeneracy between conformations with a given number of HH contacts that

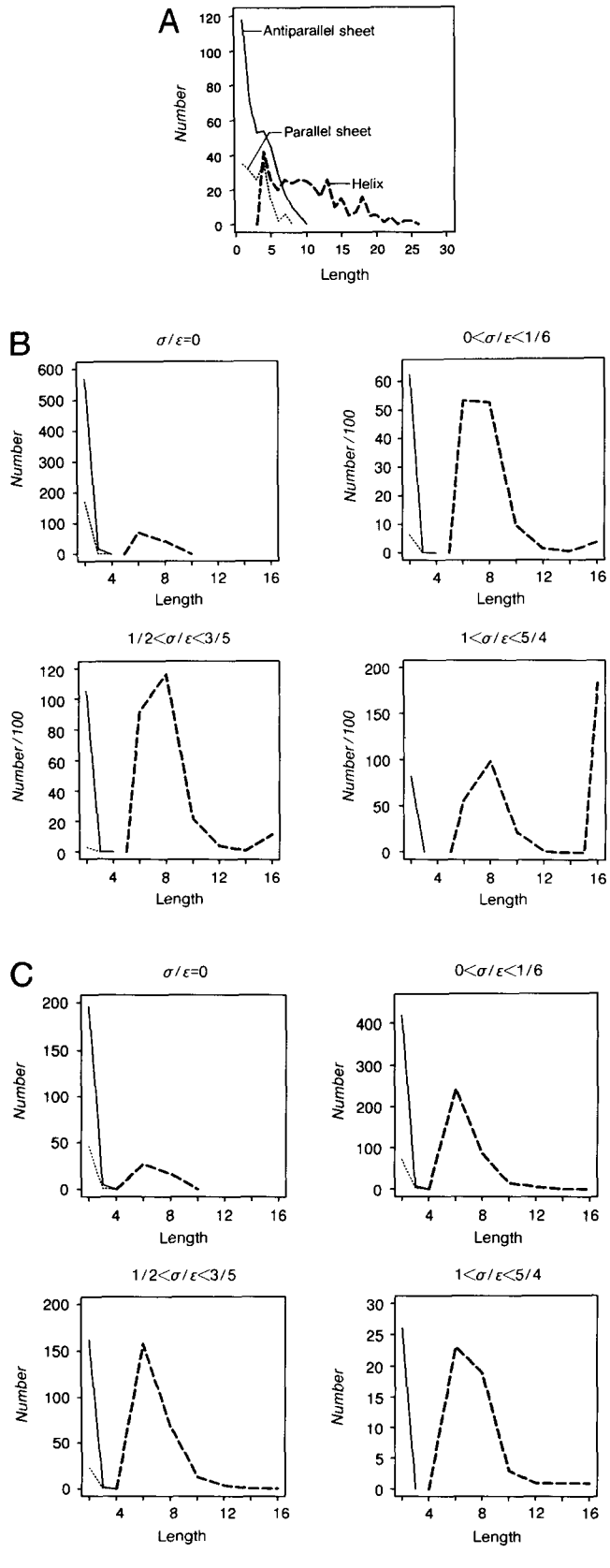

Fig. 6. Distributions of secondary structures versus length. **A:** Kabsch and Sander (1983) distributions obtained from the PDB. **B:** Model using single-sequence sets over four sample intervals of σ/ϵ. **C:** Model using single-conformation sets.

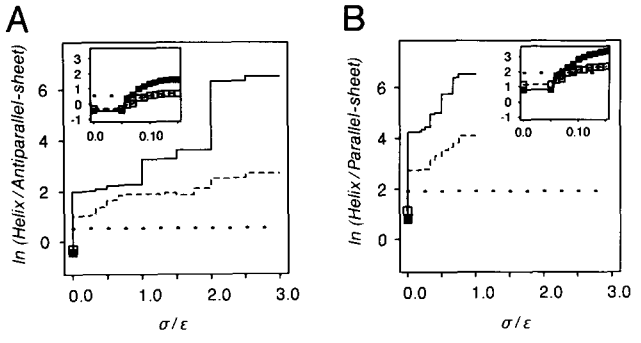

Fig. 7. Ratios of secondary structure types versus σ/ϵ. **A:** The ratio of (total number of residues in helices)/(total number in antiparallel sheets) for the model and the PDB (dotted flat line). **B:** Helices/parallel sheets. The solid line corresponds to the single-sequence sets, and the dashed line represents the single-conformation sets. The insets show small σ/ϵ using the additional "minimum stability" requirement for proteinlike conformational states to smooth the low-σ/ϵ end of the curves. In all cases, the model values intersect the PDB values for $\sigma/\epsilon < 0.1$.

have different numbers of helical bonds. As a result, all conformations in the $\sigma/\epsilon > 0$ sets that are not present in the $\sigma/\epsilon = 0$ set *must* contain at least one helical unit (six residues). Using the minimum stability requirement, these helix-containing conformations are added to the set gradually with increasing σ/ϵ, and we obtain smooth curves for $\sigma/\epsilon < 0.15$ (Fig. 7, insets).

The helix/parallel sheet ratio (Fig. 7B) is even more sensitive than the helix/antiparallel sheet ratio to the helical propensities. Because of the shortness of the model chains, there are relatively few ways to form parallel sheets when 6 of the 16 residues are already involved in a helix. That there is so little freedom to form such sheets accounts in part for the large increase in the helix/parallel sheet ratio for $\sigma/\epsilon > 0$. However, because of the 2D definitions of secondary structure (Fig. 5), a helical residue can also be counted as participating in a sheet when another part of the chain folds back to contact the helix. For example, in Figure 1B, residues 1-6 are in a helix and residues 2 and 3 also form a parallel sheet with residues 15 and 16.

We conclude that the model native secondary structure distributions are most similar, in both shapes and relative areas, to those in the PDB when the model local interaction is assumed to be much smaller than the nonlocal interaction ($\sigma/\epsilon < 0.1$). This is a statement about a model and about how it can be brought into closest correspondence to real proteins. It is a prediction that might be profitably tested in other models, including off-lattice models.

Tertiary structures and the QQ plot

Having compared secondary structure distributions between the model and real proteins, we now examine a property involving tertiary structure comparisons. We first briefly describe a method for obtaining a distribu-

tion of pairwise dissimilarities between tertiary structures, and then a general method for comparing different distributions. Finally, we use these methods to compare our model conformation distributions to the PDB. For this tertiary structure comparison as well, the model bears closest resemblance to real proteins when the helical interactions in the model are set to be near zero.

The pairwise structural dissimilarity distribution, described in more detail elsewhere (Yee & Dill, 1993), is based on a measure, $d(R, S)$, of pairwise dissimilarities of two polymer or protein conformations, R and S. $d(R, S)$ is a number that ranges from 0 to 1, 0 indicating the structures are identical and 1 indicating they have the greatest possible structural dissimilarity. The dissimilarity of two chain conformations is computed from their weighted distance maps. For a chain of length N, the weighted distance map is an $N \times N$ matrix in which each element $w(i, j)$ equals the distance between the positions of residues i and j (Cα coordinates are used for protein comparisons; lattice monomer sites are used for the model) raised to the inverse power, p ($p > 0$):

$$w(i, j) = d(i, j)^{-p}. \tag{5}$$

The weighted distance map has the property that residues that are close together in space are weighted more heavily than residues that are distant in space. Here we use $p = 2$, so the weights include contributions from residues other than nearest neighbors. The comparison of two distance maps, to get the score $d(R, S)$, is made by sliding one map across the other and finding the alignment of highest similarity.

Using this measure, $d(R, S)$, of dissimilarity between conformations R and S, we find the pairwise dissimilarities among all the conformations in a given set. The $d(R, S)$ distribution is shown in Figure 8 for $n(n-1)/2 =$ 6,328 pairwise comparisons of 113 representative proteins in the Protein Data Bank. It represents the relatedness among the tertiary structures of known proteins. Correspondingly, we obtain a $d(R, S)$ distribution for the native conformations predicted by the 2D helical-HP model for different intervals of σ/ϵ (Fig. 9, insets). For each interval of σ/ϵ, distributions were generated for both the single-conformation and single-sequence sets.

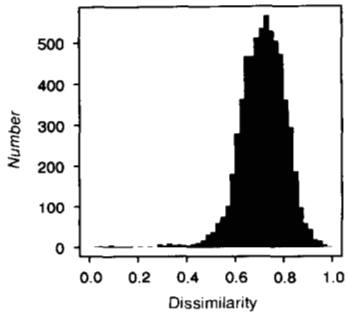

Fig. 8. Distribution of pairwise tertiary structure dissimilarities for 113 proteins from the PDB (see text for details).

How similar are the two distribution functions: of the pairwise dissimilarities among proteins and of the pairwise dissimilarities among native states in the helical-HP model? We make this comparison using a quantile–quantile (QQ) plot (Chambers et al., 1983). This type of plot compares the shapes of two distribution functions. The range of values on the x- and y-axes of the QQ plot is 0 to 1. Point (x, y) on the QQ plot is the pair ($d(R, S)$ value of one distribution, $d(R', S')$ value of the other distribution) at which the areas under the two distribution functions are equal. That is, a point (x, y) on a QQ plot is represented by:

$$\int_0^x A\, ds = \int_0^y B\, ds, \tag{6}$$

where A is the distribution function represented on the x-axis of the plot, B on the y-axis, and ds is a small score interval.

The QQ plot has several properties that make it useful for comparing distribution functions. First, if two distributions have identical shape, the QQ plot is linear. For example, the comparison of two Gaussian distributions will yield a linear QQ plot irrespective of their relative widths and displacements. If the QQ plot is linear, the slopes and the intercepts have a simple interpretation. The slope corresponds to the relative widths of the distributions; e.g., a slope of 3 indicates that distribution B is wider than distribution A by a factor of 3. The y-intercept corresponds to the shift of the mean of one distribution relative to the other: it is the point on distribution A that corresponds to 0 on distribution B. Figure 9 also shows QQ plots for the 2D helical-HP lattice model versus the PDB for the given intervals of σ/ϵ.

We focus on three properties of the QQ plot: (1) linearity (resemblance of the shapes), (2) slope (resemblance of the widths), and (3) slope-intercept (resemblance of the mean values). These three properties of the QQ plot most quantitatively express the resemblance of the pairwise structural dissimilarity distribution of the lattice model native conformations to the PDB distribution, and are plotted versus σ/ϵ in Figure 10. The root mean square deviation (RMSD) of the QQ plots from linearity (Fig. 10A) is probably the best measure of deviation between distributions, because if linearity does not hold, the other measures lose their simple interpretations.

Figure 10A shows that the model most nearly resembles the PDB for $\sigma/\epsilon = 0$ for the single-sequence set and for $\sigma/\epsilon < 1$ for the single-conformation set. This result holds true even when the more stringent requirement of minimum stability is used to define the conformation sets (Fig. 10A, inset). As noted above, the single-sequence set may be more representative of the sequence distribution in the PDB. The single-conformation set is much less sensitive to small changes in σ/ϵ, since conformations are lost gradually with increasing σ/ϵ (Fig. 4A). The fall of the

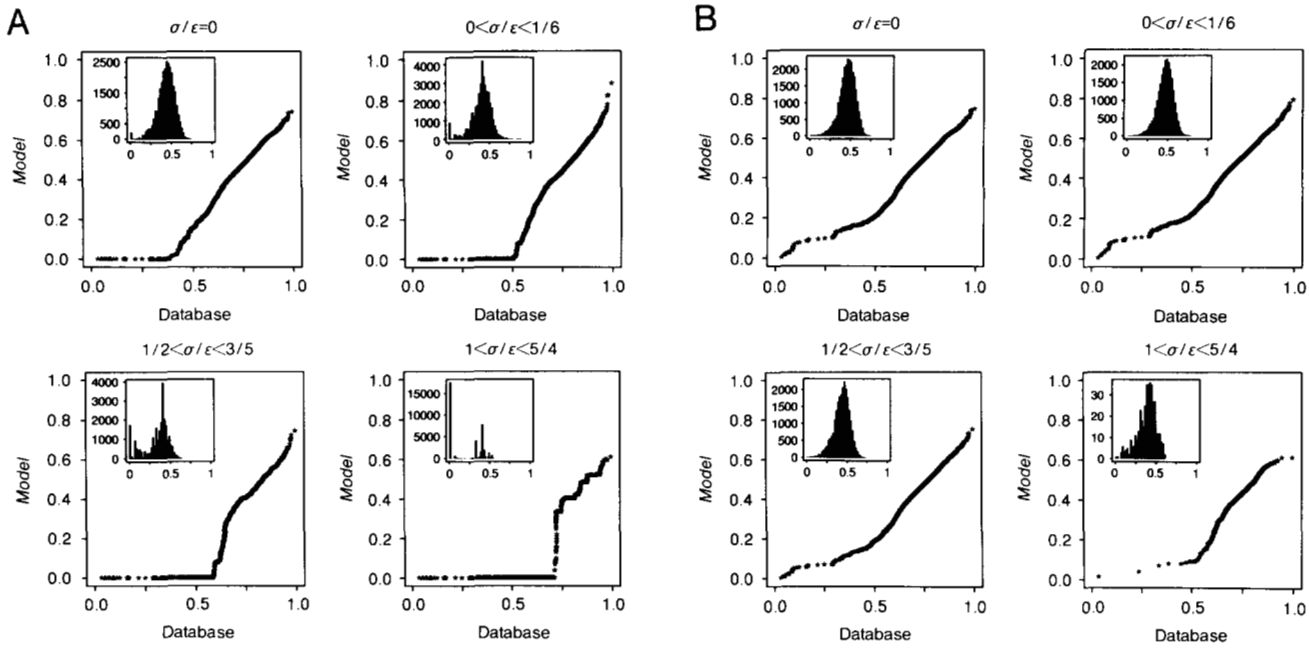

Fig. 9. Model pairwise dissimilarity distributions (insets). Comparison with the PDB distribution is shown as QQ plots. The more linear the plot, the more similar the model and PDB distributions. Single-sequence (**A**) and single-conformation (**B**) sets. The greatest similarity is for $\sigma/\epsilon = 0$ in A, and $\sigma/\epsilon < 3/5$ in B. These distributions are constructed by comparing each conformation in the set over a given range of σ/ϵ pairwise with every other conformation in that set. The peak at score 0 in A reflects different sequences folding to identical conformations.

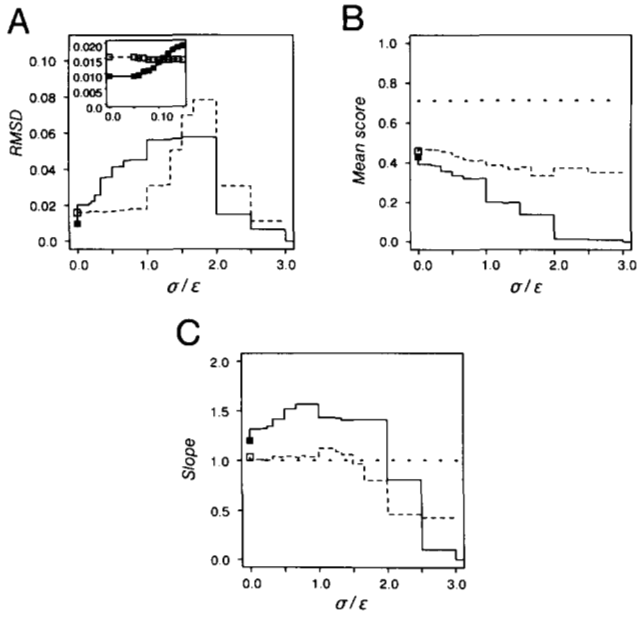

Fig. 10. Comparing the QQ plots. **A:** RMSD of QQ plots from linearity. The inset uses the "minimum stability" requirement for defining proteinlike conformational states, which yields a "continuous" curve for $\sigma/\epsilon \leq 0.15$. An RMSD of 0 indicates that two distribution shapes are identical; larger RMSDs indicate less similarity between distribution shapes. **B:** Mean dissimilarity scores of the distributions. The PDB distribution has a mean of 0.71. **C:** Slopes of linear fits to the QQ plots. A slope of 1 indicates that the two distributions have identical widths; smaller slopes indicate narrower distributions. Solid and dashed lines represent single-sequence and single-conformation sets, respectively.

RMSD to zero for large σ/ϵ does not indicate an increase in similarity of the distributions; it becomes a pathological comparison at that point and simply reflects that most sequences fold to a single conformation, the helix. The single-sequence dissimilarity distribution collapses to a peak at 0 and the QQ plot is fit well by a line of slope 0, while the single-conformation distribution becomes undefined because there is only one conformation.

The mean scores graphed in Figure 10B simply confirm the expected result that sets of 16-residue conformations on a 2D lattice do not have as much structural variation as 3D proteins of lengths from about 50 to 200 residues. The mean score for the model decreases with increasing σ/ϵ, as native conformations become increasingly helical and tend to resemble each other more closely. What is more surprising, however, is that the model dissimilarity distributions predicted by the model for $\sigma/\epsilon < 1$ have approximately the same width as the PDB (Fig. 10C). This result indicates that the model displays a *range* of structural variation similar to the PDB.

Thus, the dissimilarity distribution for the single-sequence set at $\sigma/\epsilon = 0$ differs from the PDB distribution primarily in having a lower mean score. If we were to shift the PDB distribution to match this lower mean score, the low-score tail of the distribution would show negative dissimilarity scores. Significantly, the relative area of this low-score region almost exactly matches the relative area of the zero-score peak of the single-sequence distribution

for $\sigma/\epsilon = 0$. This tail indicates some clustering of the proteins in the PDB into families (Yee & Dill, 1993), which the low-resolution model reflects as different sequences having identical conformations. In this way, the single-sequence set for $\sigma/\epsilon = 0$ captures a prominent feature of the PDB that is absent from the single-conformation sets.

We conclude that the closest resemblance of the model to real proteins is obtained when the nonlocal interactions dominate and the local interactions in the model are small. Thus, adding a helical interaction to the HP model does not improve its ability to resemble properties of proteins in the PDB. Even if the 2D model intrinsically overestimates the amounts of helix, then it would artifactually account for why no helical propensities need be added to the HP model to reach the amounts of helix observed in real proteins. But this explanation would fail to account for the agreement of the distributions of secondary structural types, the decreasing frequency of helices with length, and the pairwise tertiary structural similarities. Apart from the disagreement of the chain-length-dependent mean score, the distribution of model native structures for $\sigma/\epsilon = 0$ resembles closely that of proteins, including some clustering into families seen in the single-sequence set. Because the model has the low resolution of a square lattice, is restricted to short chains, and is two dimensional, it is obviously not a microscopically accurate representation of real proteins. Nevertheless, the model reproduces the general features of these distributions, and two different measures, of secondary and tertiary structures, show that the best correspondence with real proteins is obtained when local interactions are set to be very small compared to nonlocal interactions. The implications of the model for proteins are that the relative distributions of helix and sheet, and the general distribution of tertiary structural topologies, of globular proteins in aqueous solution are dictated more strongly by the nonlocal hydrophobic interactions than by helical propensities.

The mechanism of alcohol denaturation

Globular proteins can be denatured or stabilized by agents in solution that mediate nonlocal or local interactions or both. For example, whereas urea and guanidine hydrochloride disrupt both hydrophobic clustering and secondary structure, denaturation by alcohols is more complex. Also, peptides adopt different secondary structures depending on the nature of the solvent (Zhong & Johnson, 1992). In this section, we use the helical-HP model to explore the different classes of structures that could arise when different types of agents, particularly alcohols, act on globular proteins.

How might alcohols affect peptide and protein conformations? Several physical mechanisms have been proposed to explain these conformational effects, but the relative importance of each has not yet been resolved. Focusing on helix induction by TFE, Nelson and Kallenbach (1986) have compared charge and hydrogen-bonding mechanisms. The dielectric constant of TFE is approximately one-third that of water, so charge interactions should be more important in TFE. By varying the dielectric constant, Nelson and Kallenbach found a negligible increase in peptide helix stabilization in TFE and concluded that hydrogen bonding may be the more important mechanism of the two. Nuclear magnetic resonance studies by Llinás and Klein (1975) have shown that TFE is a slightly stronger proton donor than water, but is a much weaker proton acceptor. Polypeptide backbone groups have both donors and acceptors. Because the effect of TFE on proton acceptance dominates, adding TFE to aqueous solution will primarily decrease the solvent's ability to compete with peptide carbonyl acceptors. In TFE, the peptide amide donors should therefore favor making hydrogen bonds with peptide carbonyls. Therefore, Nelson and Kallenbach (1986) concluded that intramolecular hydrogen bonds should be strengthened by the addition of TFE to an aqueous solution. Because the α-helical conformation forms backbone–backbone hydrogen bonds, it will become increasingly favored by addition of TFE to an aqueous solution.

The effects of alcohols on polypeptides have been found to be strongly sequence-dependent (Lehrman et al., 1990; Segawa et al., 1991; Sönnichsen et al., 1992). The tendency to induce helical structure has been found in nearly all studies, but the amount of alcohol needed varies widely depending on the sequence and the length of the polypeptide. Three peptides corresponding to β-sheet-containing regions of plastocyanin show no appreciable increase in helical content in up to 90% TFE (Dyson et al., 1992). For some peptides, the amount of TFE required for the coil-to-helix transition correlates with the helical propensity according to secondary-structure prediction methods such as those of Chou and Fasman (1974). For proteins, the effects of alcohol appear to be more complex. Myelin basic protein (MBP) in 92% TFE shows approximately 47% helical structure as judged from far-UV circular dichroism data (Stone et al., 1985), whereas ubiquitin shows almost 100% helical structure under similar conditions (Wilkinson & Mayer, 1986). More recent NMR studies of ubiquitin in 60% methanol (dielectric constant = 51), however, indicate that much of the native secondary structure is preserved despite the strongly helical CD results (Harding et al., 1991). In additional studies on the fragments of MBP (Stone et al., 1985) and other proteins (Lehrman et al., 1990; Segawa et al., 1991), the sum of CD spectra for proteolytic fragments of a protein is found to differ from the spectrum for the intact protein. Hence, nonlocal interactions appear to play some role in determining the conformations that are induced by alcohols.

The stability of ribonuclease is decreased by addition of alcohols and tetraalkylammonium salts; the magnitude of this decrease depends only on the number of effective

methylene groups of the compound, i.e., on the hydrophobicity of the solvent (von Hippel & Wong, 1965). Further, the effect of ethanol on the thermodynamics of thermal ribonuclease denaturation displays the same temperature dependence as partition experiments on small hydrophobic solutes (Brandts & Hunt, 1967). NMR and near-UV CD studies of alcohol-induced states of proteins indicate a disruption of well-defined tertiary structure (Dufour & Haertlé, 1990; Harding et al., 1991; Buck et al., 1993; Fan et al., 1993; Alexandrescu et al., 1994). The alcohol denaturation of ubiquitin (Wilkinson & Mayer, 1986) and myoglobin appears to be well described by a "preferential solvation model" in which alcohols are proposed to interact most strongly with hydrophobic groups on the protein (Arakawa & Goddette, 1985). Therefore, in addition to its effects on hydrogen bonding, TFE and other alcohols may also affect the strength of nonlocal interactions in proteins.

Thus, it is known how the helicity of some proteins changes with concentration of agents such as TFE and urea. Because these agents affect both local and nonlocal interactions, our model provides a way to explore their actions. The helical-HP model contains both local and nonlocal mechanisms in their simplest forms. The strengthening of intramolecular helical bonds of proteins in TFE is modeled as an increase in the helical bond energy, σ. The weakening of hydrophobic interactions is modeled as a decrease in the HH contact interaction energy, ϵ. As noted earlier, we have found that globular proteins in aqueous solutions are best represented in our model by taking the ratio of local to nonlocal interaction strength, σ/ϵ, to be approximately zero. Now, let us suppose the effect of increasing the concentration of some agent, such as alcohol, is to strengthen helical bonds (σ more negative) and weaken HH interactions (ϵ less negative) in the ratio μ. For example, for an agent that has $\mu = 1$, every increment of its concentration leads to 1 unit of free energy stabilization of helical interactions and 1 unit of free energy destabilization of the HH interaction. An agent that has $\mu = 0$ is one that acts entirely to weaken HH interactions. An agent with $\mu \gg 1$ is a pure helical bond stabilizer. An agent with $\mu \ll -1$ is a pure helical bond destabilizer. Figure 11A shows a native state phase diagram for a helical-HP sequence, and the lines with different slopes, μ, correspond to the ways that different agents will lead to the traversal of different stable conformations.

The average helicity, $\langle h/N \rangle$, for any helical-HP sequence can be calculated as a function of the energies σ and ϵ, using

$$\langle \text{helicity} \rangle = \left\langle \frac{h}{N} \right\rangle = Z^{-1} \sum_{m=0}^{m_{\max}} \sum_{n=0}^{n_{\max}} \sum_{h=0}^{N} \frac{h}{N} \times g(m,n,h) \cdot e^{-[(m-m_{\text{native}})\epsilon + (n-n_{\text{native}})\sigma]}, \quad (7)$$

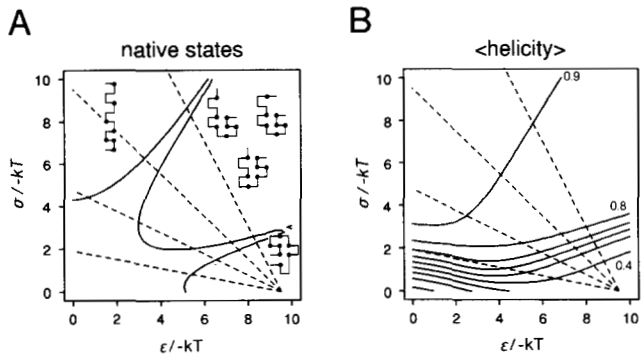

Fig. 11. Model solvent effects on σ and ϵ. **A:** The dashed lines show different possible models for how an incremental change in solvent concentration affects σ and ϵ, in the ratios (from bottom to top) $\mu = 1/5$, 1/2, 1, 2 for the native state phase diagram of the sequence (PHPPPHPPHPHPHHHP). The alcohol-induced conformations are less stable than the aqueous native state for $\mu < 1$. **B:** The corresponding ⟨helicity⟩ phase diagram, calculated from Equation 7.

where h is the number of residues participating in helices, N is the chain length, and all other symbols are as defined for Equation 4. Figure 11B shows a sample ⟨helicity⟩ phase diagram for a helical-HP sequence.

Figure 12A shows the model denaturation curves for the sequence in Figure 11. The various curves represent different balances (μ) in the increase in helical bond strength relative to the decrease in HH interaction strength. These curves are "slices" through the ⟨helicity⟩ phase diagram (Fig. 11B). For comparison, experimental helicities measured by far-UV CD during alcohol and urea denaturation of different proteins are shown in Figure 13. The alcohol-denaturation data (Fig. 13A) are best fit by approximately $1/2 < \mu < 1/5$, suggesting that the primary effect of alcohol is on the nonlocal interactions rather than on the local helical interactions. That is, in the action of alcohols on the conformations of globular proteins, our model predicts that the effect on the hydrophobic contribution to the free energy is two- to fivefold more important than the effect on helical propensities. We refer to agents with these values of μ as "model alcohols."

Let us compare denaturation by urea with denaturation by alcohols. Figure 13B shows the experimental CD data for lysozyme and FK-binding protein (FKBP). The model best fits these curves for a range of values $\mu \leq 0$. Thus, whereas TFE and other alcohols strengthen helical propensities, urea weakens them. According to this model comparison, urea acts on proteins mainly by weakening the intrachain hydrophobic interactions and to a lesser degree by weakening helical propensities.

The shapes of the transition curves depend not only on μ but also on the value of ϵ chosen to represent the pure "aqueous" solution. The value of ϵ controls the sharpness of the curves (see Fig. 12B). We also find that different HP sequences yield different denaturation curves. How-

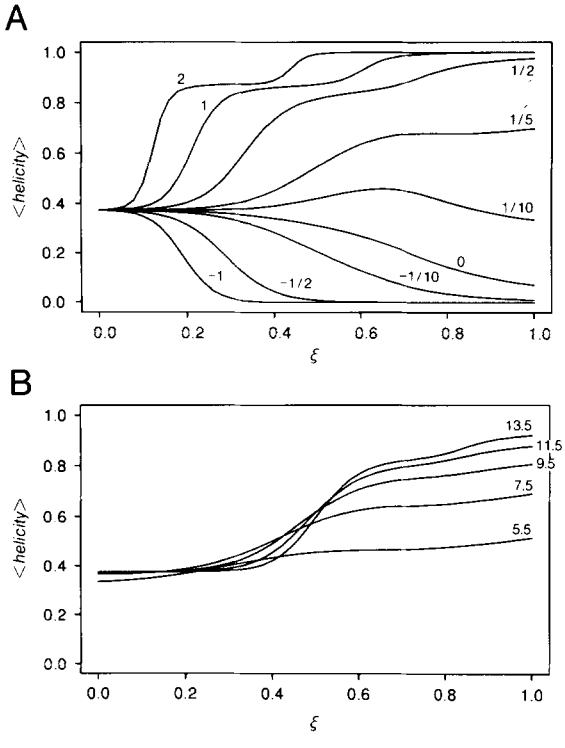

Fig. 12. Model denaturation curves for different solvents. Same sequence as in Figure 11. **A:** Model solvent μs are shown for $\epsilon = -9.5kT$, and range from -1 (weakens helix and HH in equal proportions) to $+2$ (helix is strengthened twice as much as HH is weakened). The reaction coordinate ranges from $\xi = 0$ at $\epsilon = -9.5kT$ to $\xi = 1$ at $\epsilon = 0$, following the dotted lines in Figure 11 from lower right to upper left. **B:** The effect of different values of ϵ, in units of $-kT$, in model aqueous solution ($\sigma/\epsilon = 0$), i.e., of changing the location of the starting point, but not the angle μ ($=1/4$), for the traces.

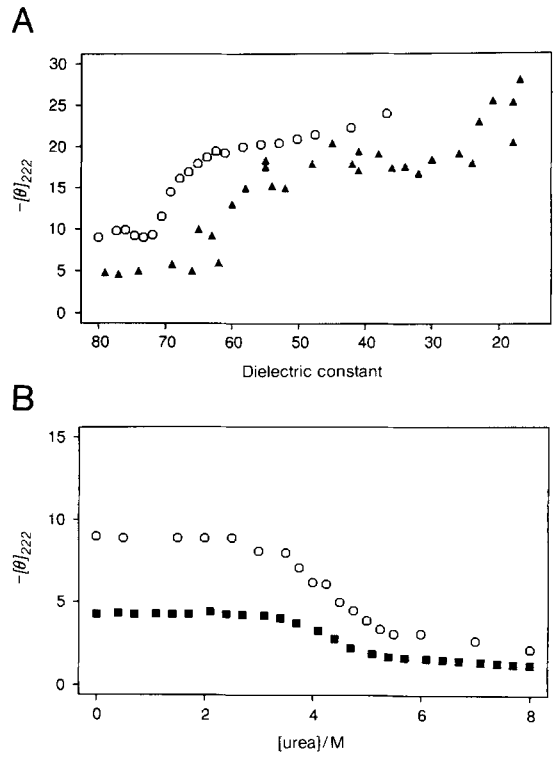

Fig. 13. Experimental protein denaturation monitored by CD. **A:** The helicity, molar ellipticity at 222 nm, as a function of the dielectric constant of the solvent, for different alcohols (Wilkinson & Mayer, 1986). The TFE denaturation of hen egg-white lysozyme is shown with circles, and the denaturation of ubiquitin using methanol, ethanol, isopropanol, and butanol is shown with triangles. **B:** The molar ellipticity at 222 nm as a function of urea concentration for hen egg-white lysozyme (circles) and FK-binding protein (squares). Data for ubiquitin are adapted from Wilkinson and Mayer (1986), for lysozyme from Buck et al. (1993), and for FKBP from Egan et al. (1993). Note that the units for $[\theta]_{222}$ are different for the different proteins.

ever, for nearly all sequences having a unique native conformation at $\sigma/\epsilon = 0$, the qualitative behavior is similar to Figure 12, depending mainly on the amount of helix in the native conformation.

We find that different native structures respond differently to alcohol. That is, effects are sequence-dependent. Adding model alcohol to a sheet lattice protein causes it to become helical (Fig. 14A). Even though the helical assistance from the model alcohol is weak, it is sufficient to stabilize helices because the alcohol disrupts the HH interactions that hold the sheet together. The sheet-to-helix transition has been observed experimentally for several β-sheet proteins, including concanavalin A (Jackson & Mantsch, 1992) and β-lactoglobulin (Tanford et al., 1960; Dufour & Haertlé, 1990). This mechanism resolves a puzzle. If TFE caused helix formation primarily by its action on hydrogen bonding alone, then it would be difficult to explain why β-sheet proteins should become helical, because the total number of hydrogen bonds (intramolecular and peptide-solvent) should not change substantially in the sheet-to-helix transition. For a helical protein, on the other hand, the model predicts that alcohol will have essentially no effect on helical content, up to the very highest alcohol concentrations (see Fig. 14B), with perhaps some slight decrease at the very highest concentrations. Consistent with this prediction, myoglobin, which is mostly α-helix, is found to have little change in helical content in 0–100% chloroethanol (Jackson & Mantsch, 1992).

Figure 12A also shows that for some HP sequences, model alcohol denaturation can induce a two-state conformational transition to a relatively stable state that has different conformational properties than the "aqueous" state. This model alcohol state still contains a significant population of conformations other than the native for approximately $\mu \leq 1/2$; for larger values of μ, the helical states are of comparable stability to the model "aqueous" native state ($\sigma/\epsilon = 0$). A putative folding intermediate for hen egg-white lysozyme in 50% TFE has recently been characterized by CD, 2D NMR, and NMR deuterium exchange experiments (Buck et al., 1993). The transition to this intermediate appears to be two-state, and NMR in-

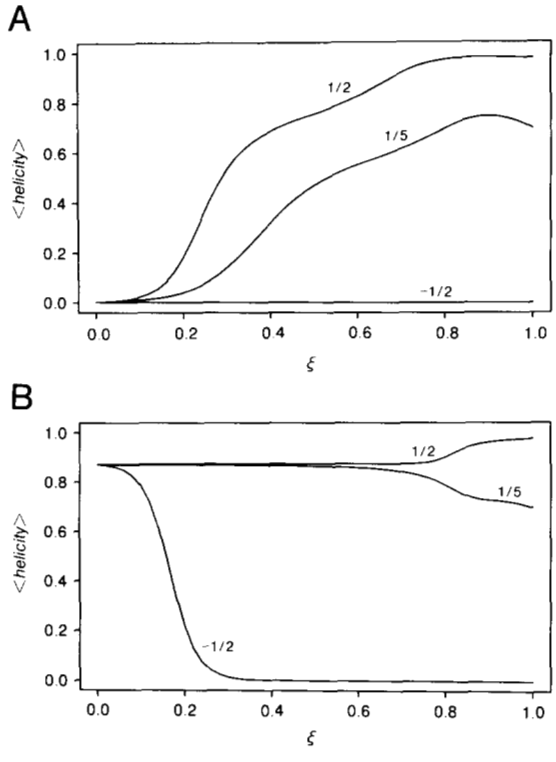

Fig. 14. A: A model sheet protein becomes helical in model alcohol ($1/5 < \mu < 1/2$), but denatures in model urea ($\mu = -1/2$). Sequence HPHHPHHHHHHPHHHH has a sheet conformation in aqueous conditions ($\sigma/\epsilon = 0$). **B:** A model helical bundle protein, sequence HHPH HHHPHHPHHHHH, does not change helical content in model alcohol, but denatures in model urea.

dicates significant conformational averaging in the intermediate state. A novel state of β-lactoglobulin has also been observed at about 20% ethanol that shows a different binding stoichiometry and only a small change in ellipticity at 222 nm relative to the aqueous protein (Dufour & Haertlé, 1990). Monellin adopts a relatively stable nonnative conformation in both 50% ethanol and 50% TFE, in which one of the β-strands converts to an α-helix while the native α-helix remains intact (Fan et al., 1993). Two helical regions have been identified from the 2D NMR spectrum of α-lactalbumin in 50% TFE at low pH (Alexandrescu et al., 1994); one of these regions forms an α-helix in the aqueous conformation as well, while the other does so only in the presence of TFE. We find similar behavior for the model helical-HP sequence shown in Figure 11A. A conformational transition occurs at $\sigma/\epsilon = 1/3$ in which the N-terminal residues become helical in model alcohol. This state contains three native conformations, all of which preserve the C-terminal helix found in the native state as well as the N-terminal model alcohol-induced helix. The three conformations differ in the relative orientation of the two helices, i.e., the secondary structure is well defined but the tertiary structure is not.

Despite its simplicity, the helical-HP model displays features of real proteins in solutions with alcohol cosolvents, including CD curves and conformational transitions for some sequences. These data are qualitatively fit for a range of values $1/5 < \mu < 1/2$, which suggests that the dominant effect of alcohols on aqueous protein solutions is on the hydrophobic interaction. The model predicts that all sequences will ultimately become helical in alcohol solutions. The observation that some peptides do not adopt a helical conformation in TFE (Dyson et al., 1992) would require a generalization of the model in which different residues would have different helical propensities.

Conclusions

We describe the helical-HP lattice model of protein conformations. It contains two interaction parameters: ϵ, representing the nonlocal contact interactions that favor HH contacts, and σ, representing a local propensity for helix formation. For all the possible sequences for chains of length $n = 16$, we vary these two energy parameters over their full ranges. When the HH contact energy is the dominant interaction, different sequences collapse to different unique native conformations that are compact and globular. When the helix interaction is dominant, the lowest energy state of all sequences is the helical conformation. What relative strength of helical and HH contact energies causes the lattice model to have structural properties resembling those of real proteins? We consider two structural properties. First, we consider the relative amounts of different types of secondary structure. HH-driven compactness alone, in the absence of significant local interactions, leads to model distributions of secondary structures that most closely resemble those of proteins in the PDB. Even a small increase in the helical interactions in the model leads to much more helix than is observed in real proteins. The second property we consider is the distribution of tertiary structural similarities among all proteins. The 2D lattice model gives distributions that closely resemble those of real proteins. For this property, too, the model most closely resembles real proteins if the helical interaction is very small. Of course, there are important caveats due to the simplicity of the model—it is two dimensional, restricted to lattice conformations, low resolution, and it involves only short chains. Nevertheless, the similarities we find, for different properties, between real proteins and our model at $\sigma/\epsilon = 0$ suggest that intrinsic helical propensities may not be a strong driving force for the folding of globular proteins.

When both HH and helical interactions are present, chains are found to undergo *conformational switching transitions* often from one unique conformation to another in response to changes in external solvent conditions. Conformational switching involves transitions between different native states, and results from a trade-off between helix formation and HH contacts. This model

describes the effects of different types of denaturing agents on protein conformations, including the transitions of sheet proteins and others to helical conformations in alcohols. Comparisons with experiments suggest that whereas urea weakens nonlocal hydrophobic interactions and helical propensities, alcohols such as TFE act mainly by weakening hydrophobic interactions but with some small strengthening of helical propensities.

Acknowledgments

We thank David Yee and Dr. Hue Sun Chan for helpful discussions and for providing some computer programs, and Dr. Chris Dobson for helpful discussions and for sending preprints prior to publication. Support was provided by the NIH. P.D. Thomas is a Howard Hughes Medical Institute Predoctoral Fellow.

References

Alexandrescu, A.T., Ng, Y.-L., & Dobson, C.M. (1994). Characterization of a TFE-induced partially folded state of α-lactalbumin. *J. Mol. Biol.*, in press.

Anfinsen, C.B. & Scheraga, H.A. (1975). Experimental and theoretical aspects of protein folding. *Adv. Protein Chem. 29*, 205–300.

Arakawa, T. & Goddette, D. (1985). The mechanism of helical transition of proteins by organic solvents. *Arch. Biochem. Biophys. 240*, 21–32.

Brandts, J.F. & Hunt, L. (1967). The thermodynamics of protein denaturation. III. The denaturation of ribonuclease in water and in aqueous urea and aqueous ethanol mixtures. *J. Am. Chem. Soc. 89*, 4826–4838.

Buck, M., Radford, S.E., & Dobson, C.M. (1993). A partially folded state of hen egg white lysozyme in trifluoroethanol: Structural characterization and implications for protein folding. *Biochemistry 32*, 669–678.

Camacho, C.J. & Thirumalai, D. (1993a). Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. USA 90*, 6369–6372.

Camacho, C.J. & Thirumalai, D. (1993b). Minimum energy compact structures of random sequences of heteropolymers. *Phys. Rev. Lett.*, in press.

Chambers, J.M., Cleveland, W.S., Kleiner, B., & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Duxbury Press, Boston.

Chan, H.S. & Dill, K.A. (1989). Compact polymers. *Macromolecules 22*, 4559–4573.

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

Chan, H.S. & Dill, K.A. (1991a). Sequence-space soup of proteins and copolymers. *J. Chem. Phys. 95*, 3775–3787.

Chan, H.S. & Dill, K.A. (1991b). Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem. 20*, 447–449.

Chothia, C. (1976). The nature of accessible and buried surfaces in proteins. *J. Mol. Biol. 105*, 1–14.

Chou, P.Y. & Fasman, G.D. (1974). Prediction of protein conformation. *Biochemistry 13*, 222–245.

Chou, P.Y., Wells, M., & Fasman, G.D. (1972). Conformational studies on copolymers of hydroxylpropyl-L-glutamine and L-leucine. Circular dichroism studies. *Biochemistry 11*, 3028–3043.

Dill, K.A. (1990). Dominant forces in protein folding. *Biochemistry 29*, 7133–7155.

Dufour, E. & Haertlé, T. (1990). Alcohol-induced changes of β-lactoglobulin-retinol binding stoichiometry. *Protein Eng. 4*, 185–190.

Dyson, H.J., Sayre, J.R., Merutka, G., Shin, H.-C., Lerner, R.A., & Wright, P.E. (1992). Folding of peptide fragments comprising the complete sequence of proteins: Models for initiation of protein folding II. plastocyanin. *J. Mol. Biol. 226*, 818–835.

Dyson, H.J. & Wright, P.E. (1991). Defining solution conformations of small linear peptides. *Annu. Rev. Biophys. Biophys. Chem. 20*, 519–538.

Egan, D.A., Logan, T.M., Liang, H., Matayoshi, E., Fesik, S.W., & Holzman, T.F. (1993). Equilibrium denaturation of recombinant human FK binding protein in urea. *Biochemistry 32*, 1920–1927.

Fan, P., Bracken, C., & Baum, J. (1993). Structural characterization of monellin in the alcohol-denatured state by NMR: Evidence for β-sheet to α-helix conversion. *Biochemistry 32*, 1573–1582.

Fauchère, J.-L. & Pliska, V. (1983). Hydrophobic parameters Π of amino acid side chains from the partitioning of *N*-acetyl-amino acid amides. *Eur. J. Med. Chem. Ther. Chem. 18*, 369–375.

Gregoret, L.M. & Cohen, F.E. (1991). Protein folding—Effect of packing density on chain conformation. *J. Mol. Biol. 219*, 109–122.

Harding, M.M., Williams, D.H., & Woolfson, D.N. (1991). Characterization of a partially denatured state of a protein by two-dimensional NMR: Reduction of the hydrophobic interactions in ubiquitin. *Biochemistry 30*, 3120–3128.

Hao, M.H., Rackovsky, S., Liwo, A., Pincus, M.R., & Scheraga, H.A. (1992). Effects of compact volume and chain stiffness on the conformations of native proteins. *Proc. Natl. Acad. Sci. USA 89*, 6614–6618.

Jackson, M. & Mantsch, H.H. (1992). Halogenated alcohols as solvents for proteins: FTIR spectroscopic results. *Biochim. Biophys. Acta 1118*, 139–143.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*, 2577–2637.

Killian, J.A. (1992). Gramicidin and gramicidin-lipid interactions. *Biochim. Biophys. Acta 1113*, 391–425.

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

Lau, K.F. & Dill, K.A. (1990). Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA 87*, 638–642.

Lehrman, S.R., Tuls, J.L., & Lund, M. (1990). Peptide α-helicity in aqueous TFE: Correlations with predicted α-helicity and the secondary structure of the corresponding regions of bovine growth hormone. *Biochemistry 29*, 5590–5596.

Lesk, A.M. (1991). *Protein Architecture [Practical Approach Series]*. IRL Press, New York.

Lipman, D.J. & Wilbur, W.J. (1991). Modelling neutral and selective evolution of protein folding. *Proc. R. Soc. Lond. B 245*, 7–11.

Llinás, M. & Klein, M.P. (1975). Charge relay at the peptide bond: a protein magnetic resonance study of solvation effects on the amide electron density distribution. *J. Am. Chem. Soc. 97*, 4731–4737.

Lyu, P.C., Liff, M.I., Marky, L.A., & Kallenbach, N.R. (1990). Side chain contributions to the stability of alpha-helical structure in peptides. *Science 250*, 669–673.

Miller, R., Danko, C.A., Fasolka, M.J., Balazs, A.C., Chan, H.S., & Dill, K.A. (1992). Folding kinetics of proteins and copolymers. *J. Chem. Phys. 96*, 768–790.

Mutter, M. & Hersperger, R. (1990). Peptides as conformational switch: Medium-induced conformational transitions of designed peptides. *Ang. Chem. 29*, 185–187.

Nandi, P.K. & Robinson, D.R. (1984). Effects of urea and guanidine hydrochloride on peptide and nonpolar groups. *Biochemistry 23*, 6661–6668.

Nelson, J.W. & Kallenbach, N.R. (1986). Stabilization of the Ribonuclease *S*-peptide α-helix by trifluoroethanol. *Proteins Struct. Funct. Genet. 1*, 211–217.

Nozaki, Y. & Tanford, C. (1971). The solubility of amino acids and two glycine polypeptides in aqueous ethanol and dioxane solutions. *J. Biol. Chem. 246*, 2211–2217.

O'Neil, K.T. & DeGrado, W.F. (1990). A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science 250*, 646–651.

O'Toole, E.M. & Panagiotopoulos, A.Z. (1993). Effect of sequence and intermolecular interactions on the number and nature of low-energy states for simple model proteins. *J. Chem. Phys. 90*, 3185–3190.

Rees, D.C., DeAntonio, L., & Eisenberg, D. (1989). Hydrophobic organization of membrane proteins. *Science 245*, 510–513.

Richards, F.M. & Richmond, T. (1978). Solvents, interfaces and protein structure. In *Molecular Interactions and Activity in Proteins*

(Wolstenholme, G.E., Ed.), pp. 23–45. Ciba Foundation Symposium 60. Excerpta Medica, Amsterdam.

Robinson, D.R. & Jencks, W.P. (1965). The effect of compounds of the urea-guanidinium class on the activity coefficient of the acetyltetraglycine ethyl ester and related compounds. *J. Am. Chem. Soc. 87*, 2462–2470.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., & Zehfus, M.H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science 229*, 834–838.

Scholtz, J.M. & Baldwin, R.L. (1992). The mechanism of alpha-helix formation by peptides. *Annu. Rev. Biophys. Biomol. Struct. 21*, 95–118.

Segawa, S.-I., Fukono, T., Fujiwara, K., & Noda, Y. (1991). Local structures in unfolded lysozyme and correlation with secondary structures in the native conformation: Helix-forming and -breaking propensity of peptide segments. *Biopolymers 31*, 497–509.

Shoemaker, K.R., Fairman, R., York, E.J., Stewart, J.M., & Baldwin, R.L. (1988). Circular dichroism measurement of peptide helix unfolding. In *Peptides: Proceedings of the Tenth American Peptide Symposium* (Marshall, G.R., Ed.), pp. 15–20. ESCOM, Leiden.

Shortle, D., Chan, H.S., & Dill, K.A. (1992). Modeling the effects of mutations on the denatured states of proteins. *Protein Sci. 1*, 201–215.

Sönnichsen, F.D., van Eyk, J.E., Hodges, R.S., & Sykes, B.D. (1992). Effect of TFE on protein secondary structure: An NMR and CD study using a synthetic actin peptide. *Biochemistry 31*, 8790–8798.

Stickle, D.F., Presta, L.G., Dill, K.A., & Rose, G.D. (1992). Hydrogen bonding in globular proteins. *J. Mol. Biol. 226*, 1143–1159.

Stone, A.L., Park, J.Y., & Martenson, R.E. (1985). Low-ultraviolet CD spectra of oligopeptides 1–95 and 96–168 derived from myelin basic protein of rabbit. *Biochemistry 24*, 6666–6673.

Tamburro, A.M., Scatturin, A., Rocchi, R., Marchiori, F., Borin, G., & Scoffone, E. (1968). Conformational transitions of bovine pancreatic ribonuclease *S*-peptide. *FEBS Lett. 1*, 298–300.

Tanford, C. (1968). Protein denaturation. *Adv. Protein Chem. 23*, 121–282.

Tanford, C., De, P.K., & Taggart, V.G. (1960). The role of the α-helix in the structure of proteins: Optical rotatory dispersion of β-lactoglobulin. *J. Am. Chem. Soc. 82*, 6028–6034.

Unger, R. & Moult, J. (1993). Genetic algorithms for protein folding simulations. *J. Mol. Biol. 231*, 75–81.

von Hippel, P.H. & Wong, K.-Y. (1965). On the conformational stability of globular proteins: The effects of various electrolytes and nonelectrolytes on the thermal ribonuclease transition. *J. Biol. Chem. 240*, 3909–3923.

Wilkinson, K.D. & Mayer, A.N. (1986). Alcohol-induced conformational changes of ubiquitin. *Arch. Biochem. Biophys. 250*, 390–399.

Yee, D. & Dill, K.A. (1993). Families and the structural relatedness among globular proteins. *Protein Sci. 2*, 884–889.

Zhong, L. & Johnson, W.C., Jr. (1992). Environment affects amino acid preference for secondary structure. *Proc. Natl. Acad. Sci. 89*, 4462–4465.

Zimm, B.H. & Bragg, J.K. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys. 31*, 526–535.