

# Helix, Sheet, and Polyproline II Frequencies and Strong Nearest Neighbor Effects in a Restricted Coil Library<sup>†</sup>

Abhishek K. Jha,<sup>‡,§,#</sup> Andres Colubri,<sup>#,§</sup> Muhammad H. Zaman,<sup>‡,§</sup> Shohei Koide,<sup>‡</sup> Tobin R. Sosnick,<sup>\*,§,⊥</sup> and Karl F. Freed<sup>\*,‡,#</sup>

Department of Chemistry, Institute for Biophysical Dynamics, The James Franck Institute, and Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, Illinois 60637

Received November 30, 2004; Revised Manuscript Received April 15, 2005

**ABSTRACT:** A central issue in protein folding is the degree to which each residue's backbone conformational preferences stabilize the native state. We have studied the conformational preferences of each amino acid when the amino acid is not constrained to be in a regular secondary structure. In this large but highly restricted coil library, the backbone preferentially adopts dihedral angles consistent with the polyproline II conformation rather than  $\alpha$  or  $\beta$  conformations. The preference for the polyproline II conformation is independent of the degree of solvation. In conjunction with a new masking procedure, the frequencies in our coil library accurately recapitulate both helix and sheet frequencies for the amino acids in structured regions, as well as polyproline II propensities. Therefore, structural propensities for  $\alpha$ -helices and  $\beta$ -sheets and for polyproline II conformations in unfolded peptides can be rationalized solely by local effects. In addition, these propensities are often strongly affected by both the chemical nature and the conformation of neighboring residues, contrary to the Flory isolated residue hypothesis.

A small energy bias can rationalize the finite time scales observed for protein folding (2), despite the Levinthal-like number of conformations that must be searched (1). One such bias is the preference of the polypeptide backbone to adopt certain  $\phi, \psi$  dihedral angles, as observed in Ramachandran plots for observed dihedral angles (Figure 1) (3). These plots differ for each amino acid because the  $\phi, \psi$  distribution is affected by interactions between the backbone and the side-chains. For example, the Ramachandran plots for proline and glycine are quite distinct due to their unusual side-chains. The other 18 amino acids exhibit a smaller variation and mainly populate the same three regions in the Ramachandran map. These regions are called the extended  $\beta$ -basin, the  $\alpha$ -helical and turn basin, and the polyproline II (PPII)<sup>1</sup> basin (4), which is centered at  $\phi \sim -75^\circ$  and  $\psi \sim 150^\circ$ .

The conformational preferences of a residue are also influenced by the chemical identity and conformation of the neighboring residues (5–11). These observations contradict the zeroth order Flory isolated residue hypothesis (IRH) (12), which states that the conformations adopted by any residue are independent of the chemical identity and conformation

of its neighbors. Therefore, it is important to quantify the influence of nearest neighbors (NNs) on conformational propensities.

Computer simulations of peptides have the potential to identify backbone preferences and delineate NN effects. However, such procedures are computationally challenging because of the large number (8000 for triplets) of distinct combinations of amino acids with their immediate NNs. In addition, the results are very sensitive to the choice of force fields (5, 13–16).

A more fruitful approach is to utilize a subset of residues located outside of regular secondary structures in crystal structures (7, 17–19). Early studies (7, 10, 17, 18, 20–22) assume that the use of the entire database of folded protein structures would average over many environments, thereby largely eliminating the influence of context. Although such an approximation cannot completely account for contextual influences, the database does contain sterically allowed conformations, which also reflect the chemical character of the individual amino acids.

Subsequently, only a subset of the database, termed the “coil library”, has been retained to provide a better reproduction of the intrinsic conformational preferences. Swindells et al. (18) derived their coil library by removing amino acids residing in  $\alpha$ -helices and  $\beta$ -strands from a dataset of 85 high-resolution X-ray structures and used this library to determine intrinsic secondary structure propensities. Hermans et al. (23) compared this type of coil library to computer simulations of “dipeptides” for five residues, finding only modest agreement. Ohlson et al. (24) found a high level of PPII conformations for most residues in their coil library that was constructed from a database of 1024 proteins by deleting the residues in helices and sheets.

<sup>†</sup> This work is supported by grants from the NIH (K.F.F. GM56678, T.R.S. GM55694, and S.K. GM57215). A.J. acknowledges the support of Burroughs Wellcome Fund Interfaces #1001774. A.C. is supported by Postdoctoral Fellowship from The Program in Mathematics and Molecular Biology.

\* To whom correspondence should be addressed. (T.R.S.) E-mail: trsosnic@midway.uchicago.edu. Phone: 773-834-0657. Fax: 773-702-0439. (K.F.F.) E-mail: k-freed@uchicago.edu. Phone: 773-702-7202.

<sup>‡</sup> Department of Chemistry.

<sup>§</sup> Institute for Biophysical Dynamics.

<sup>#</sup> The James Franck Institute.

<sup>⊥</sup> Department of Biochemistry and Molecular Biology.

<sup>1</sup> Abbreviations: C-F, Chou-Fasman; IRH, Flory isolated residue hypothesis; PPII, polyproline II; NN, nearest neighbor.

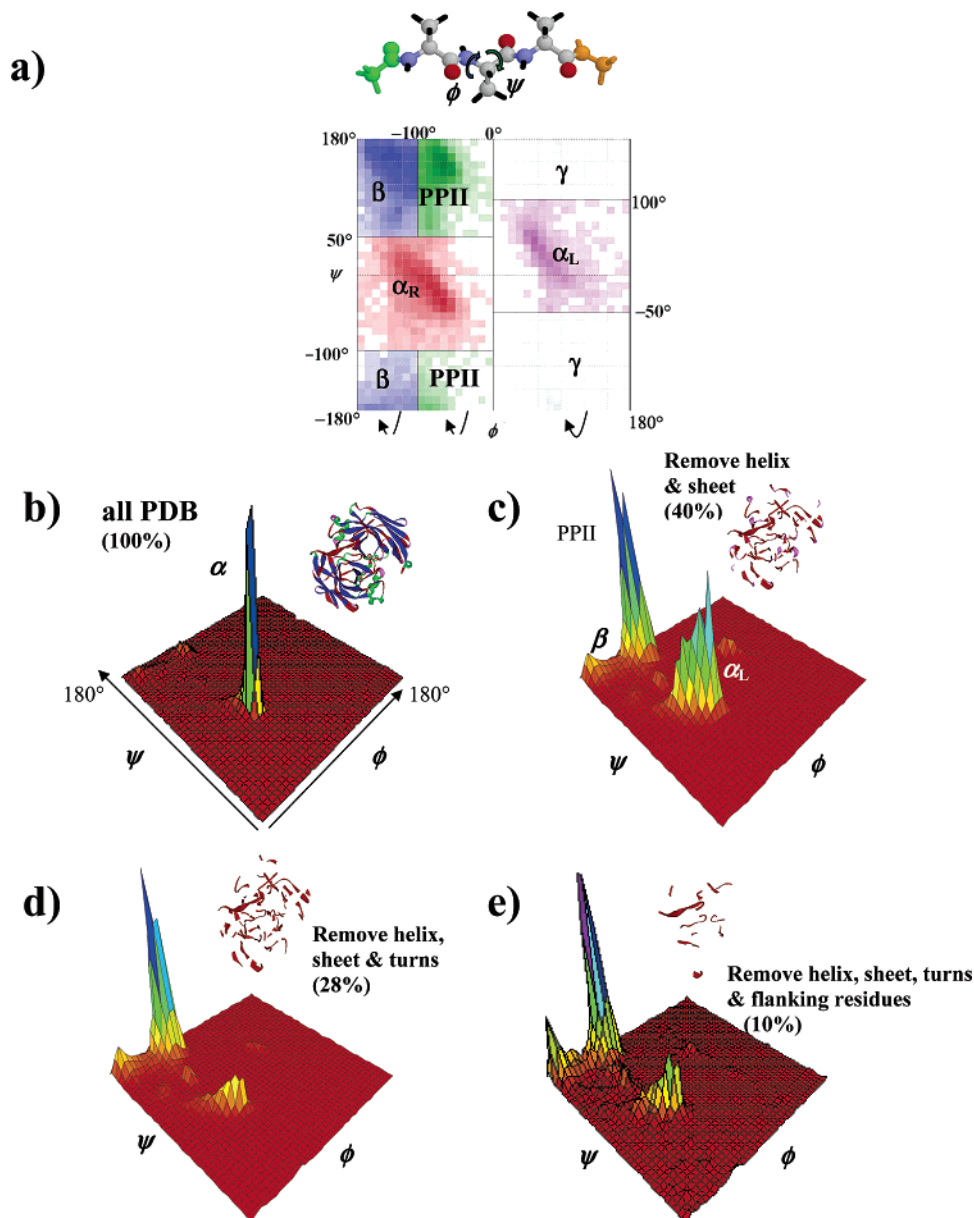


FIGURE 1: Backbone conformational preferences in different libraries. (a) Tri-alanine peptide (Ace-(Ala)<sub>3</sub>-Nme) with central of the three pairs of backbone ( $\phi, \psi$ ) dihedral angles highlighted. The hydrogen atoms are shown in stick representation (black), whereas oxygen (red), nitrogen (blue), and carbon (grey) are depicted in ball-and-stick representation. The Ramachandran map of ( $\phi, \psi$ ) dihedral angles presents the distribution of all residues in the entire PDB of 2020 structures. Rectangles define the  $\phi, \psi$  regions that are termed the “helical basin”, “coil basin”, and “PP II basin” below. Basin occupancy does not necessarily imply secondary structure. Probability distribution in the Ramachandran plane of Ala for the (b)  $C_{\text{all PDB}}$  library, termed the entire PDB, (c)  $C_{\alpha\beta}$ , the library without helices and sheets, (d)  $C_{\alpha\beta}$ , without helices, sheets, and turns, and (e)  $C_{\alpha\beta}^{\text{intern}}$  without helices, sheets, turns, and terminal residues. The structure in the panels depicts the portion of the X-ray structure of a typical protein (PDB code 1GTT-A) that is retained in each of the four coil libraries. The number in parentheses is the fraction of the entire PDB that remains in the respective library.

However, these coil libraries retain all residues in turns, which represent a third class of secondary structure with internal hydrogen bonds. After removing the residues in turns, Serrano derived  $\phi$  angle distributions for each residue from 279 protein structures (25). Smith et al. (7) noted that the use of a coil library that was derived by removing the residues in  $\alpha$ -helices and  $\beta$ -strands exhibits a good correlation coefficient with experimentally determined NMR  $J$ -coupling constants for unstructured peptides. The agreement improved when they accounted for the identity of the amino-terminal neighbor. Overall, these coil libraries have been used to describe structural propensities and NMR parameters for denatured proteins with a reasonable degree of success.

PPII helices are a fourth class of secondary structure that recently has become the focus of a number of studies (26–41), despite having been suggested almost 40 years ago (4). Beyond PPII being the dominant backbone conformation in peptides (4, 26–33), Adzhubei and Sternberg found the presence of an average of over one stretch of four consecutive PPII residues in a library of 80 proteins (42). The origin of the preference for PPII in peptides may be due to preferential solvation (36–38), a suggestion that is consistent with the stretches of PPII tending to occur on the surfaces of these proteins.

The aforementioned coil libraries contain a bias toward the PPII conformation because the inclusion of residues

Table 1: Distribution of Conformations in the Four Libraries<sup>a</sup>

library	helical basin	$\beta$ basin	PPII basin	other
C <sub>all PDB</sub> , all PDB (Figure 1b)	53.0(63.4)	26.7(17.8)	17.7(17.1)	2.6(1.7)
C <sub><math>\alpha\beta</math></sub> , PDB minus (helix & sheet) (Figure 1c)	37.0(38.2)	23.2(14.8)	33.2(42.1)	6.7(4.9)
C <sub><math>\alpha\beta\tau</math></sub> , PDB minus (helix, sheet & turns) (Figure 1d)	21.9(20.7)	31.8(21.2)	42.3(54.7)	4.1(3.6)
C <sub><math>\alpha\beta\tau</math></sub> <sup>intern</sup> , PDB minus (helix, sheet, turns & flanking residues) (Figure 1e)	27.4(23.8)	32.9(24.7)	35.5(48.0)	4.2(3.5)

<sup>a</sup> For all residues except Gly and Proline. Values in parentheses are for alanine only.

flanking secondary structures generally precludes them from being in the same Ramachandran basin as the associated secondary structure. For example, the residue at the end of a helix is less likely to be a helical conformation; otherwise it would be part of the helix. In addition,  $\beta$  conformations do not follow helical structures due to steric clashes (43).

Here we examine the utility of a more stringently defined coil library to describe the backbone conformational preferences. The investigation takes advantage of the 10-fold increase in the size of the PDB database that has occurred in the past decade. Using a nonredundant database (25% maximum sequence homology), we compare the three prior definitions of a coil library to our more stringent library that is composed only of the nonflanking residues in stretches of four or more residues that lie outside of helices, strands, and turns (Figure 1b–e) and that discards other biased residues, such as those preceding prolines and others that are highly exposed and, therefore, are biased toward the  $\alpha$ -basin. This improved coil library has strong to moderate preference for PPII conformations for most residues, and stretches of four or more consecutive PPII residues only account for less than 4% of the coil library. High PPII levels are found even for buried residues, suggesting that the major portion of the preference for PPII is independent of solvation. Nevertheless, the PPII frequencies in the library correlate well with the observed levels in unfolded peptides (Chen et al., personal communication).

After applying a new masking procedure, we find that the backbone conformational preferences in the new coil library for each nonproline residue strongly correlate with the relative frequencies with which the residues appear in regular  $\alpha$ -helices and  $\beta$ -strands in the entire PDB. The method of masking explains why prior attempts at determining structural preferences for  $\alpha$ -helices have been marred by the presence of significant outliers. Finally, NN effects are often found to be of the same magnitude as the spread in structural propensities for the 20 amino acids. These results stress the importance of local effects in determining the native state structure and stability of proteins.

## METHODS

**Coil Libraries.** The protein sequence culling server PISCES (44) is used to obtain X-ray structures of 2020 chains longer than 20 residues with a maximum R-factor of 0.3 and a resolution better than 2 Å. Residues preceding prolines are removed, although with minimal effect. The most exposed ~10% of the residues are excluded when computing monomer propensities, due to the geometric artifact that these surface residues preferentially bend back toward the body of the protein, which places them preferentially in the helical basin. The accessible surface area (ASA) is computed using a probe radius of 1.4 Å (45). Removing residues whose backbone B-factors exceed 30 Å<sup>2</sup> (46) produces negligible

changes in individual residue basin propensities (results not shown), so these are not eliminated. The residues are assigned secondary structures using the DSSP algorithm (47). All residues given the assignments H or G, B or E, and T by the DSSP algorithm are assumed to be in helices, strands, and turns, respectively. The database called the entire PDB library is obtained by excluding only the terminal residues for each of the chains.

**Chou-Fasman Propensities and Ramachandran Masks.** We calculate slightly modified Chou-Fasman (C-F) frequencies. The helical frequency is computed as the frequency that a given amino acid is found in helical fragments once two residues from each end of the helix are excluded. Similarly, the C-F frequencies for  $\beta$ -strands are computed by taking the frequency of occurrence of those residues in  $\beta$ -strands that explicitly lie in the  $\beta$ -basin, but not the PPII or  $\alpha$ -basins. We separately compute the C-F frequency for those residues in  $\beta$ -strands that are in the PPII basin. The basin frequency for an amino acid is computed from the new library as the frequency that the amino acid resides in a given basin as defined in Figure 1a. Our new calculation of helical frequencies uses only a restricted region of the Ramachandran map corresponding to the regions occupied by each residue in regular helical structures. For each residue, all occurrences as internal residues of helical fragments are binned in bins of 10° × 10° to obtain Ramachandran map distributions for the residues in actual helices. We retain contributions only from those helical bins whose frequency of occurrence is greater than 0.15 times the maximum bin frequency for the given residue. The C-F frequencies are converted to an energy scale and are presented relative to alanine.

**Choice of Bin Size for Calculation of Entropy.** The calculation of the backbone entropy depends on the area chosen for the bins. However, the entropy differences are independent of this quantity as long as the bin size is the same for both the cases. To determine the appropriate bin size for these calculations, we test a wide range of bin areas elements in the Ramachandran map varying from 2° × 2° to 20° × 20°. In the extreme limit where each occurrence of a residue in the Ramachandran map lies in a separate bin, the entropy difference then depends only on the number of times each residue occurs in the library. The other limiting case is when the bin size encompasses the entire Ramachandran map, and then all residues have the same backbone entropy. A plot of the entropy difference against the bin area element as well as a function of the size of the monomer library leads to the choice of the optimal bin size of 10° × 10° for calculating backbone entropy differences since, for this case, the computed entropy differences are independent of both the number of occurrences of any residue in the database as well as the bin size (Supporting Information).

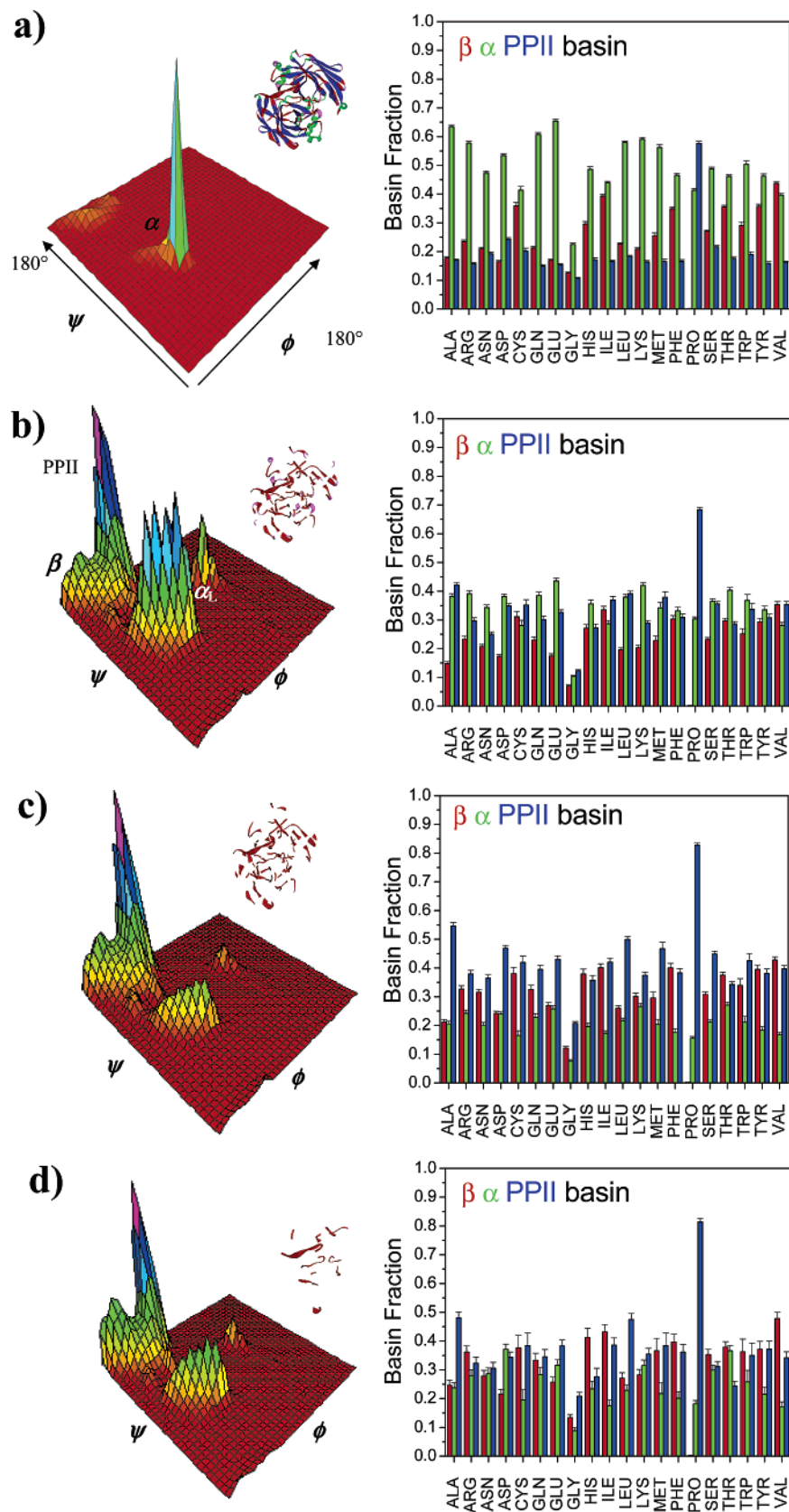


FIGURE 2: Basin preferences in different libraries. Probability distribution in the Ramachandran plane of all residues (except Gly and Pro) for the (a)  $C_{\text{all PDB}}$ , entire PDB, (b)  $C_{\alpha\beta}$ , the library without helices and sheets, (c)  $C_{\alpha\beta\gamma}$ , without helices, sheets, and turns, and (d)  $C_{\alpha\beta\gamma}^{\text{intem}}$  without helices, sheets, turns, and terminal, pre-proline, and most exposed residues described in panels a–d. The cartoon structure shown in the inset of the panels is the same as that in Figure 1b–e. Basin fractions for the 20 amino acids are shown in adjacent panels for the corresponding libraries described in panels a–d. The error bars represent one standard deviation for the respective basin propensities of the amino acids in a given library.

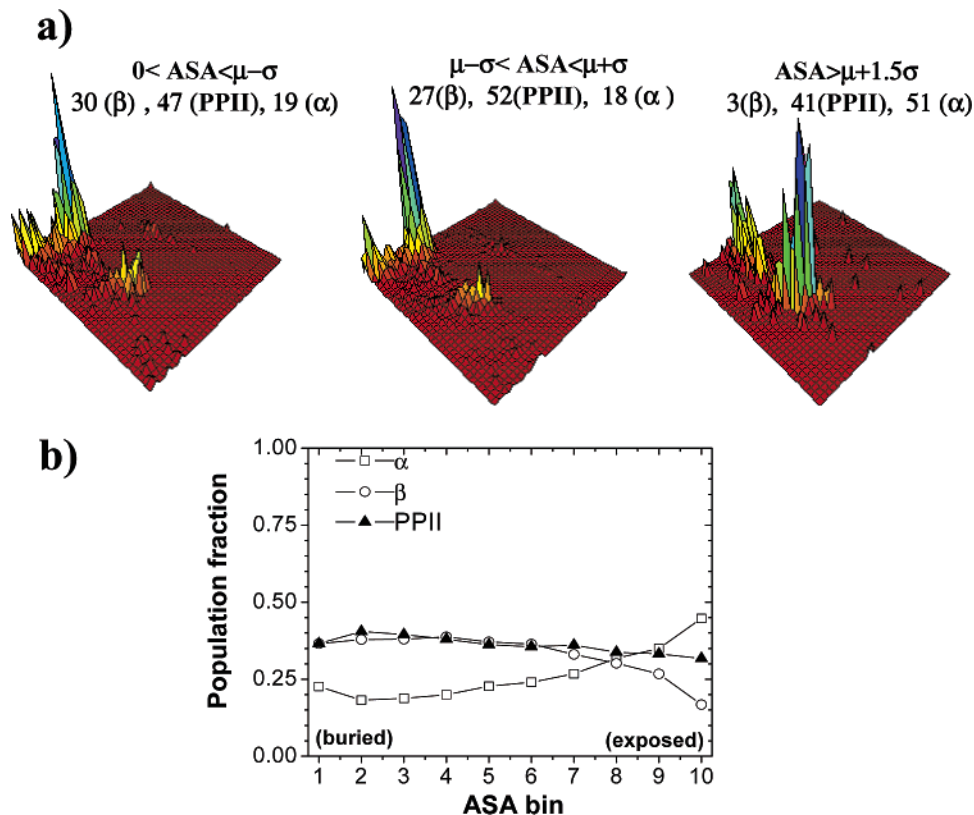


FIGURE 3: Dependence of PPII levels on degree of burial in the coil library. (a) The  $\phi, \psi$  distribution of Ala in the coil library is divided into three parts depending on the accessible surface area (ASA) of the residue in the following manner: The 21% most buried,  $ASA < \mu_{ASA} - \sigma_{ASA}$ , the central 58%,  $\mu_{ASA} - \sigma_{ASA} < ASA < \mu_{ASA} + \sigma_{ASA}$ , and the top 10%,  $ASA > \mu_{ASA} + 1.5\sigma_{ASA}$ . The mean and the standard deviation used ( $\mu_{ASA} = 44 \text{ \AA}^2$  and  $\sigma_{ASA} = 34 \text{ \AA}^2$ ) are computed from the distribution of ASAs evaluated for alanine in the library  $C_{\alpha\beta t}^{\text{intern}}$ . The ASA is computed using a probe radius of  $1.4 \text{ \AA}$  for the entire amino acid (main chain plus side-chain) (45). The values in parentheses are the percentage basin populations for the  $\beta$ -, PPII, and  $\alpha$ -basins, respectively, of alanine for the corresponding distributions. (b) Each residue in the  $C_{\alpha\beta t}^{\text{intern}}$  library is sorted based on its ASA into 10 equal sized bins ( $n = 1, 2, \dots, 10$ ) of increasing ASA. The bins for each amino acid are combined to define the 10 bins shown on the x-axis. The curves present the relative compositions of each of the three major basins, for all residues except Gly and Pro, as a function of bin number. The sharp increase in helical conformers for the most exposed ( $n = 10$ ) part of the library is due to the requirement that at the surface of the protein the chain turns back around toward the body of the protein.

## RESULTS AND DISCUSSION

**PPII Propensity in the Coil Libraries.** We investigate four increasingly stringent PDB-based libraries (Figure 1b–e, Table 1) as possible models for intrinsic backbone conformational preferences. The first three libraries have been used previously (7, 17–19, 23–25), albeit with a much smaller database ( $\sim 2000$  chains of 20 or more amino acids used presently versus only  $\sim 300$  or fewer structures in the prior works, except ref 24, which used 1042 chains). Previous coil libraries use a 50% homology threshold as opposed to the 25% threshold imposed in the present study. For comparison, we also construct analogues of the three previously studied coil libraries using our updated PDB library: (1) the entire PDB (labeled as  $C_{\text{all PDB}}$  Figure 1b); (2) a coil library created by removing helices and sheets, as defined by the DSSP algorithm (47) (denoted as  $C_{\alpha\beta}$  and containing 40% of the residues of the entire library, Figure 1c); and (3) a coil library without the residues in helices, sheets, and turns (denoted as  $C_{\alpha\beta t}$  and containing 28% of library (1), Figure 1d).

Our more stringent coil library  $C_{\alpha\beta t}^{\text{intern}}$  is composed only of a minimally biased subset of the internal residues within stretches of four or more residues in the coil library  $C_{\alpha\beta t}$  (only 10% of all residues in the  $C_{\text{all PDB}}$  library remain, Figure 1e). This new coil library eliminates biases due to the

retention of residues at the ends of structured regions. These terminal residues are inherently predisposed against helical and  $\beta$  (43) conformers and therefore favor PPII conformations. The choice of four residue stretches is a compromise between maintaining a useful size for the library and the elimination of end effects, so that no residue in the new coil library has both neighbors as end residues. An increase in the minimum length of the fragments in the library has a small effect on conformational preferences for individual residues. A library composed of regions with minimum length of 10 residues only increases the PPII propensity from  $49 \pm 2$  to  $56 \pm 4\%$  but reduces the library size by roughly 5-fold (for alanine), a reduction that seriously degrades statistics concerning NN effects. Likewise, the removal of residues whose backbone B-factors (46) exceed  $30 \text{ \AA}^2$  produces negligible changes in the calculated conformational preferences, so these residues are not removed from  $C_{\alpha\beta t}^{\text{intern}}$ . We exclude all the pre-proline residues (46) as well as the most exposed part of the library (discussed below) because these residues also have biased  $\phi$ – $\psi$  distributions. These biased residues are discarded *only* when computing monomer propensities (Figure 2d) but are retained when investigating correlations between the basin propensities of an amino acid and the chemical and conformational identity of its neighbor.

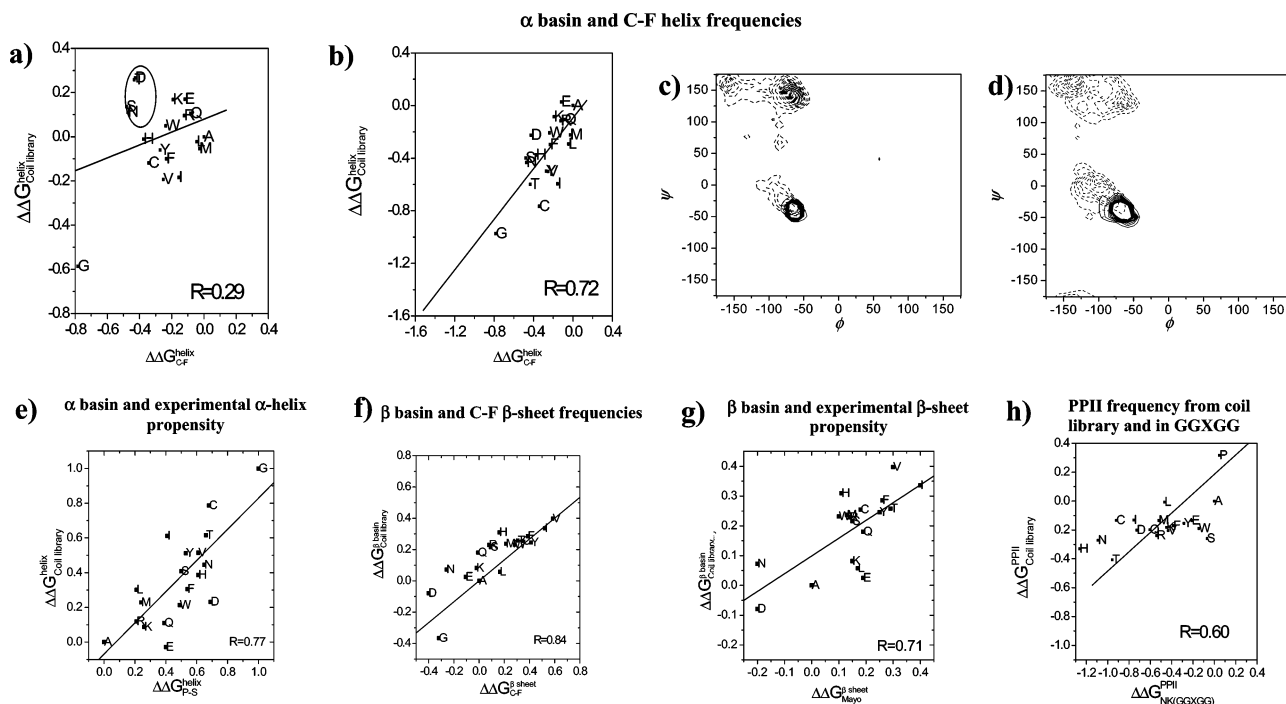


FIGURE 4: Correlation between coil library and structural propensities. Correlation between the observed frequencies in the helical basin, defined in Figure 1a.  $\Delta\Delta G_{i,coil\ library}^{helix} = -RT \ln(\text{Prob}_{i,coil\ library}^{helix}/\text{Prob}_{alanine,coil\ library}^{helix})$  is derived from the coil library, and the C-F helical propensities, normalized to the alanine frequencies, are defined by  $\Delta\Delta G_{i,C-F}^{helix} = -RT \ln(\text{Prob}_{i,authentic\ helix}^{authentic\ helix}/\text{Prob}_{alanine,authentic\ helix}^{authentic\ helix})$ . The index  $i$  refers to the  $i$ th amino acid. (a) When using the entire region of helical basin as shown in Figure 1a, the correlation yields  $\Delta\Delta G_{i,coil\ library}^{helix} = (0.20 \pm 0.23)\Delta\Delta G_{i,C-F}^{helix} - 0.08 \pm 0.08$  ( $R = 0.29$ ). (b) Same as (a) except that the values for the coil library pertain only to the small region occupied by residues in authentic helical structures [represented by solid gray contour lines in (c) and (d)] rather than the entire helical basin which comprises conformations outside of regular helices. The correlation in this case yields  $\Delta\Delta G_{i,coil\ library}^{helix} = (0.96 \pm 0.22)\Delta\Delta G_{i,C-F}^{helix} - 0.09 \pm 0.07$  ( $R = 0.72$ ). (c) Contour lines in the distribution of Ala in the coil library are presented in black dashed lines, while superimposed solid gray contour lines are for the same residue when it lies within a regular helix as defined by DSSP. (d) Similar distributions are displayed for Thr. Note that Ala correlates well in (a), and the peak of both sets of contour lines (solid gray and black dashed) overlaps well in the helical region, unlike the case for Thr, which is one of the four pronounced outliers in (a). (e) Correlation of  $\alpha$ -helical basin propensities derived from the coil library with the experimentally determined propensity scale of Pace and Scholtz:  $\Delta\Delta G_{i,coil\ library}^{helix} = (0.90 \pm 0.18)\Delta\Delta G_{i,P-S}^{helix} - 0.07 \pm 0.10$  ( $R = 0.77$ ). There is high correlation between experimental and C-F helical propensities (49) as well ( $R = 0.89$ , slope = 0.977). The propensities have been divided by that for Gly. (f) Equivalent plot for  $\beta$ -basin and  $\beta$ -sheet propensities:  $\Delta\Delta G_{i,coil\ library}^{\beta\text{basin}} = (0.54 \pm 0.08)\Delta\Delta G_{i,C-F}^{\beta\text{sheet}} + 0.09 \pm 0.03$  ( $R = 0.84$ ). (g) Corresponding plot for  $\beta$ -basin and experimentally determined  $\beta$ -sheet propensities:  $\Delta\Delta G_{i,coil\ library}^{\beta\text{basin}} = (0.59 \pm 0.15)\Delta\Delta G_{i,Mayo}^{\beta\text{sheet}} + (0.09 \pm 0.03)$  ( $R = 0.71$ ). The correlations in (a–g) exclude Pro. (h) Correlation between the PPII propensities computed from the coil library and those obtained from experiments for GGXGG model peptides:  $\Delta\Delta G_{i,coil\ library}^{PPII} = (1.49 \pm 0.48)\Delta\Delta G_{i,NK(GGXGG)}^{PPII} - 0.28 \pm 0.10$  ( $R = 0.60$ ).

The four coil libraries produce quite different Ramachandran basin distributions as illustrated by the  $\phi, \psi$  distribution of all residues types in Figure 2a–d, as well as for Ala alone (Figure 1b–e and Table 1). The  $C_{all\ PDB}$  library (Figures 1b and 2a) is heavily dominated by helical conformers. The exclusion of residues involved in helices and sheets (library  $C_{\alpha\beta}$  in Figures 1c and 2b) strongly reduces the dominance of helical conformers. The fraction of  $\beta$  conformers also decreases, while the PPII population increases. Because turn residues often lie in the helical basin, their removal (to form  $C_{\alpha\beta t}$  in Figures 1d and 2c) further reduces the helical conformer population to the point at which PPII becomes the dominant conformation.

Because the end residues are biased against the helical and sheet conformers, end residues are removed to produce the library  $C_{\alpha\beta t}^{intern}$  (Figures 1e and 2d). Once the end residues are removed, the helical basin population for all residues (except Pro and Gly) increases from 22 to 27%, while the total PPII population is reduced from 43 to 36% of the residues (for alanine, the reduction is from 55 to 49%). The PPII conformation becomes dominant for at least four (Ala,

Glu, Leu, and Pro) residues. PPII is statistically tied for the dominant conformation for seven [Asn, Cys, Gln, Gly (considering only the three main basins), Lys, Met, and Tyr] residues and is subdominant for the remaining nine amino acids. The  $\beta$ -branched residues, however, prefer the  $\beta$ -basin over the PPII basin (Figure 2d).

The exclusion of residues in helices and sheets may be considered as de facto biasing the library toward high PPII levels. This view is unwarranted because being in the coil library does not preclude a given residue from adopting an  $\alpha$  or  $\beta$  conformation. In addition, we find that PPII levels in our library are independent of burial level. This statement applies to all but the most exposed  $\sim 10\%$  of the residues, where helical conformers predominate (Figure 3). There is no evidence to believe that the preponderance of  $\alpha$  conformation for the most exposed part of the coil library is due to solvation. This shift to  $\alpha$  conformations for the most exposed residues is likely a geometric artifact by virtue of being on the surface—residues in the coil library located on the surface preferentially bend back toward the body of the protein, which places them in the helical basin. As a

consequence of the helical bias in the most exposed 10%, these residues are removed *only* for computing single residue intrinsic preferences.

For the remaining residues in the coil library, which are as buried as the average residues in the entire PDB, PPII levels do not change with the degree of burial (Figure 3b). This invariance to burial levels indicates that solvation cannot be used to explain the PPII levels in the coil library. Nevertheless, when the peptide is in water, preferential solvation can explain the high levels of PPII found in the unfolded state of proteins (36–38). Inside a protein, where there is minimal solvent–protein interactions, however, other factors must be at work. Therefore, the propensity of a residue to be in the PPII basin is due to an intrinsic backbone bias (34, 35) and not to either a biased selection criteria or a preferential solvation.

The necessity to form backbone hydrogen bonds within the protein can compensate for the loss of solvation and contributes to the stabilization of PPII geometries. The reduction in steric overlap (35) and the accessibility of the backbone for hydrogen bonding (48) likely are the major operational principles underlying conformational preferences, whether the chain resides in water or inside of a protein. The  $\beta$  conformation generally is considered to be highly extended, but its geometry places the amide N–H and carbonyl C–O bonds in the same plane. As a result, the amide hydrogen and the oxygen of the carbonyl are less accessible to their hydrogen-bonded partners. These partners are likely to interfere partially with each other, making it harder for them to form simultaneous hydrogen bonds to the N–H and C–O. This interference would occur for  $\beta$  geometries regardless of whether the partners are water molecules for the unfolded chain or protein moieties for buried coil regions. On the other hand, the  $45^\circ$ – $60^\circ$  rotation in  $\phi$  from the  $\beta$  to the PPII geometry places the amide proton and carbonyl out of plane. Consequently, the PPII backbone is geometrically better suited to form a pair of hydrogen bonds with water and/or other parts of the protein.

Finally, the improved coil library is in good qualitative agreement with experimental trends in PPII preferences (39–41). A recently derived propensity scale for PPII conformations for GGXGG model peptides, obtained using NMR  $^3J_{\text{NH-C}\alpha\text{H}}$  coupling constants and circular dichroism (Cheng, K.; Kallenbach, N.; et al, unpublished data), correlates well ( $R = 0.60$ ) with the PPII preferences in the  $C_{\alpha\beta\gamma}^{\text{intern}}$  library (Figure 4h). However, the two pairs of glycine residues flanking each residue in the experimental model peptides probably do not truly represent the averaged effect of all neighboring residues' side-chains as is the case for the PPII propensities calculated from the  $C_{\alpha\beta\gamma}^{\text{intern}}$  library. A PPII preference scale derived from a series of longer peptides gives Ala, Gln > Asn > Val (40) and also is consistent with the coil library. Thus, our coil library seemingly captures the essence of PPII preferences in unfolded peptides to the extent that these are known today.

**Secondary Structure Propensities.** We examine whether the preference for an amino acid to reside in a regular helix or sheet in the PDB can be predicted solely from its intrinsic tendency to adopt this conformation in the most stringent coil library  $C_{\alpha\beta\gamma}^{\text{intern}}$ . Accordingly, we compare the helical and  $\beta$ -basin preferences for the amino acids in our coil library

Table 2: Backbone Entropy of All Residues in the Coil Library<sup>a</sup>

residue	$T(S_X - S_{\text{Ala}})$ (kcal mol <sup>-1</sup> )
ALA	0
ARG	0.174 ± 0.060
ASN	0.283 ± 0.056
ASP	0.275 ± 0.061
CYS	-0.127 ± 0.086
GLN	0.104 ± 0.069
GLU	0.109 ± 0.036
GLY	0.475 ± 0.027
HIS	0.110 ± 0.068
ILE	-0.138 ± 0.060
LEU	0.007 ± 0.078
LYS	0.156 ± 0.056
MET	-0.143 ± 0.085
PHE	0.049 ± 0.054
PRO	-0.675 ± 0.056
SER	0.174 ± 0.061
THR	0.134 ± 0.028
TRP	-0.153 ± 0.015
TYR	0.050 ± 0.072
VAL	-0.046 ± 0.050

<sup>a</sup>  $S = -R \sum_{ij}^{36} P_{ij} \ln P_{ij}$ , where  $P_{ij}$  = probability of being in the  $i, j$ th bin. To compute the standard deviations, the coil library is divided into three equal parts for which  $S$  is computed using the above equation. We report the mean and the standard deviations of the three values for  $T\Delta S$  obtained by separating the library into thirds.

$C_{\alpha\beta\gamma}^{\text{intern}}$  to their relative frequencies in authentic helical and sheet structures, as originally done by Chou and Fasman (C-F) (20). The C-F propensity is defined as the frequency with which a residue appears in a given secondary structure. The correlation between the coil library  $C_{\alpha\beta\gamma}^{\text{intern}}$  and our recalculated C-F frequencies using the expanded PDB (see Methods) for helices is poor, even with the exclusion of the helix breakers, glycine and proline (Figure 4a). Four residues are strong outliers (Thr, Ser, Asp, Asn), a result previously noted (17, 18).

This perplexing deviation between the coil and C-F helical frequencies is removed when the former are derived using the much narrower  $\phi, \psi$  region appropriate for regular helical structures [ $\phi \sim -64 \pm 7^\circ$  and  $\psi \sim -42 \pm 7^\circ$ , Figure 5a(i)–e(i) and solid gray contour lines in Figure 4c,d] rather than the larger region employed for the full helical basin, which also contains turn geometries. This feature is displayed in Figure 5a(iii)–e(iii) and the black dashed contour lines in Figure 4c,d for two illustrative residues, Ala (which correlates well in Figure 4a) and Thr (an identified outlier in Figure 4a). With this more restricted helix mask, the correlation becomes strong and accounts for the C-F helical propensity scale as well as for the experimental propensity ( $\Delta G$ ) scale observed in guest–host systems of all non-Pro residues (49) (Figure 4b,e).

The importance of masking is due to the preference of some residues to adopt the explicit dihedral angles associated with helices (solid gray contour lines) when they are in the helical basin in the coil library (black dashed contour lines). For example, Ala strongly prefers to adopt the angles associated with regular helices (Figure 4c), whereas Thr frequently adopts other  $\alpha$ -basin geometries, such as those found in turns (Figure 4d). Therefore, the helical propensity in Figure 4b,e equates to the preference of the chain to adopt a true helical geometry rather than any angle in the helical basin, which includes other geometries.

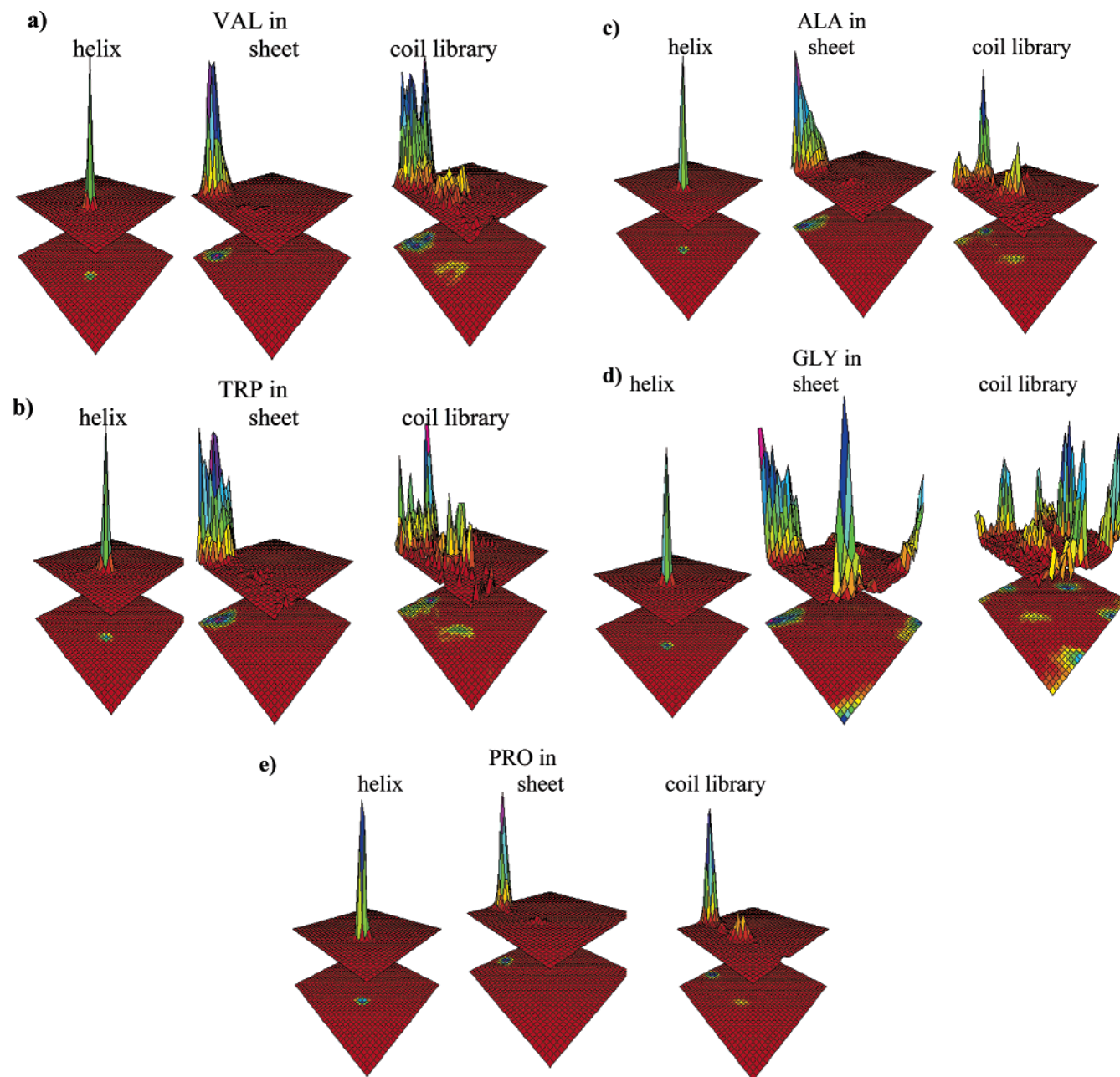


FIGURE 5: Backbone conformations of amino acids in  $\alpha$ -helix,  $\beta$ -sheet, and the coil library. The Ramachandran probability distribution is contrasted for five different amino acids when they are present in a regular helix or sheet structures or in the  $C_{\alpha\beta t}^{\text{intem}}$  library. The representative amino acids are chosen based on the steric properties of their side-chains: (a)  $\beta$ -branched side-chains (Val), (b) aromatic side-chains (Trp), (c) alanine-like (Ala), (d) Gly, and (e) Pro. All amino acids irrespective of their side-chains occupy a very narrow range of  $\phi$ - $\psi$  angles ( $\phi \sim -64 \pm 7^\circ$  and  $\psi \sim -42 \pm 7^\circ$ ) when they are part of regular  $\alpha$ -helices, in agreement with a previous study (24). Compared to the  $\alpha$ -helix distributions, the  $\beta$ -sheet probability distributions are broader, and about 23% of the residues forming  $\beta$ -sheets actually lie in the PPII basin. However, the residues in the coil library  $C_{\alpha\beta t}^{\text{intem}}$  exhibit a mild preference for the PPII basin.

Compared to helical propensities,  $\beta$ -sheet propensities are believed to be less well defined. Guest–host values are so dependent on context, including the identity of structurally adjacent side-chains (50, 51), that the existence of intrinsic  $\beta$  propensities has been questioned (52, 53), although calculations based solely on local, steric effects were able to largely capture  $\beta$ -sheet propensities (54). The correlation between our coil library and C-F frequencies is even stronger for  $\beta$ -sheets (Figure 4f) than for  $\alpha$ -helical frequencies (Figure 4e). The correlation in Figure 4f is between  $\beta$ -basin populations in the coil library  $C_{\alpha\beta t}^{\text{intem}}$  and the frequency of true  $\beta$  conformers in  $\beta$ -sheets, i.e., excluding residues found in  $\beta$ -sheets but that actually occupy the PPII region (23% of

the residues in sheets in the PDB are in the PPII basin) or  $\alpha$  basin. A good correlation also exists between the frequencies derived from the coil library and the corresponding experimental scales (54) (Figure 4g,  $R = 0.71$ ).

Applying a similar masking process for  $\beta$ -sheets (17) as performed for helices has a negligible effect. This insensitivity to masking arises because the angular spread of residues in the  $\beta$ -basin of the coil library  $C_{\alpha\beta t}^{\text{intem}}$  mimics the large spread found in  $\beta$ -sheets [Figure 5a(ii)–e(ii) and 5a(iii)–e(iii)]. The correlation between PPII populations in the coil library and PPII frequencies in  $\beta$ -sheets is moderate ( $R = 0.55$ ).



This strong correlation found for  $\beta$ -sheets does not contradict the importance of context in influencing  $\beta$  propensities. The coil library presumably averages out such contextual effects, thereby unmasking the intrinsic backbone preferences. Furthermore, the coil library represents conformational preferences within the context of a folded protein rather than in a denatured, solvated chain. Likewise, C-F frequencies are based on residues within folded proteins. Hence, desolvation issues, related to side-chain burial in going from the unfolded state to the folded state, are likely to be similar for the residues in both the coil library  $C_{\alpha\beta}^{\text{intern}}$  and in secondary structures. This similarity probably explains the higher degree of success of the coil library in reproducing the C-F values. Since the guest–host systems reflect properties of both the unfolded and the folded state, the basis for the correlation with the propensities derived from the coil library is not as apparent. Potentially, the frequencies in the coil library largely reproduce the conformational distribution of the unfolded state. In effect, the coil library provides a statistical potential for intrinsic backbone conformations.

**Coil Library for Higher Resolution Structures.** We have also analyzed a coil library for which all the retained X-ray structures have resolution better than 1.5 Å, rather than 2.0 Å employed above. The regions in the Ramachandran map depicted in Figure 1a become more restricted, but there is no appreciable change in the basin propensities for the amino acids. The correlation between high and low resolution libraries for monomer  $\alpha$ ,  $\beta$ , and PPII basin frequencies is very high ( $R = 0.93, 0.97,$  and  $0.98,$  respectively). The  $\alpha$ -helical basin propensities of the residues derived from the higher resolution coil library exhibit a marginally better correlation with C-F propensities ( $R = 0.78;$  slope of the fit = 1.38) after applying the masking procedure. The  $\beta$ -basin propensities also yield improved correlation with the corresponding C-F propensities ( $R = 0.80;$  slope of the fit = 0.62). However, the size of the library with higher resolution structures is only one-third as large, which severely restricts its use to study NN effects.

**Backbone Entropy and Glycine's Low Helical Propensity.** Using the basin populations from the coil library, we calculate the backbone conformational entropy for each residue from the relation

$$S = -R \sum_{i=1}^{36} \sum_{j=1}^{36} P_{ij} \ln P_{ij} \quad (1)$$

where  $P_{ij}$  is the normalized probability of the residue being in the  $i,j$ th  $10^\circ \times 10^\circ$  mesh element in the Ramachandran map, and  $R$  is the universal gas constant. Although this calculation implies that  $S$  depends on the mesh size (i.e., the volume per configuration in  $\phi, \psi$  space), entropy differences  $\Delta S$  between residues do not. In addition, these calculations have been performed for different bin sizes, and the final entropy differences reported in Table 2 are independent of the number of occurrences of the residue in the library as well as the size of the bin (Supporting Information). We find that most residues have similar backbone entropies, with proline being lowest as it largely populates just the PPII region (Table 2).

The difference in the helical propensity between Ala and Gly is often attributed to a difference in their backbone

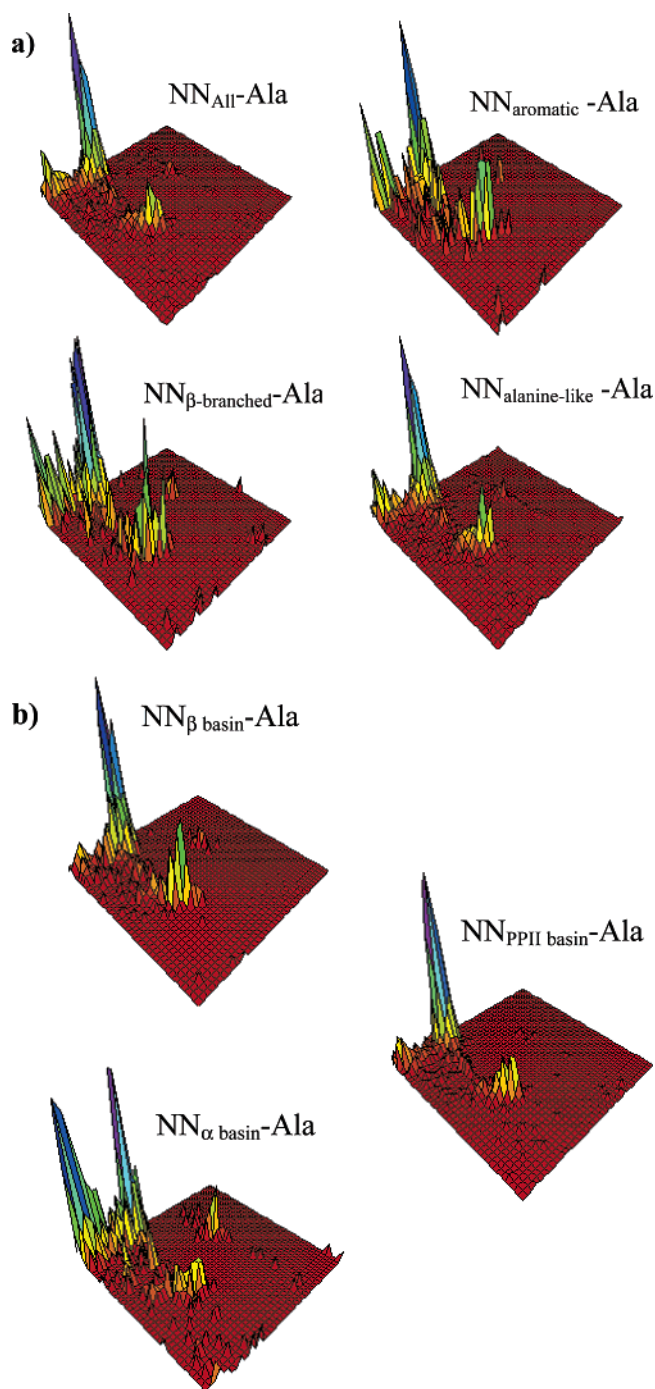


FIGURE 6: Sequence and conformational dependence of NN effects on alanine. (a) Ramachandran map for alanine from the coil library, averaged over all neighbors (except Pro and Gly), for cases in which the upstream residue is  $\beta$ -branched (Val, Ile, Thr), aromatic (His, Trp, Phe, Tyr), or the rest (alanine-like). (b) Same as (a) but the upstream residue is in the  $\beta$ -,  $\alpha$ -, or PPII basins.

entropy in the unfolded state because the folded state has the same backbone entropy and interactions (55, 56). However, glycine also has three strongly preferred, albeit different basins than alanine. As a result, the backbone entropy only is modestly increased compared to alanine ( $T\Delta S_{\text{Gly-Ala}} \sim 0.50 \text{ kcal mol}^{-1}$ ). The primary difference in their entropies is that the area of the major basins for glycine is about 2-fold larger than for alanine.

In contrast, the experimental difference in helical propensity between glycine and alanine is in the range  $\sim 0.7\text{--}1 \text{ kcal mol}^{-1}$  (55, 56), in agreement with the value of  $1.0 \text{ kcal mol}^{-1}$

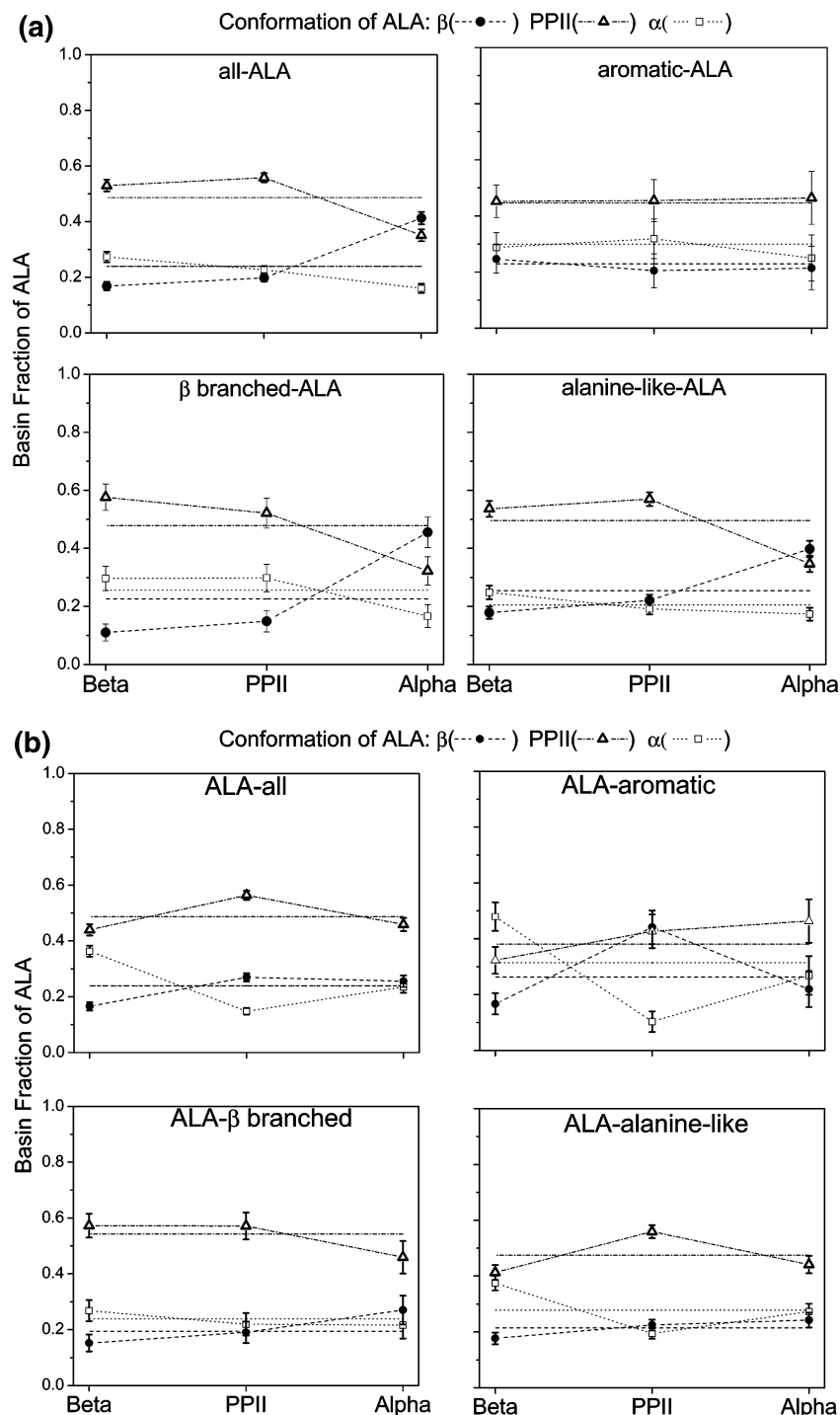


FIGURE 7: Separation of sequence and conformational dependence of NN effects. (a) Only the three most significant conformations for alanine are illustrated. Each panel represents a different amino-terminal neighbor, divided into three classes depending on the steric properties of the amino acids. The abscissa indicates the NN conformation for each of the three classes. The error bars individually denote one standard deviation. The parallel, horizontal lines provide the average fraction for each conformation. The three major basins are shown in the plot:  $\alpha$ -basin fraction (dotted line),  $\beta$  conformation (dashed line), and PPII basin (dashed and dotted line). (b) Same as (a) for the case of carboxy-terminal neighbors.

determined using our coil library (Figure 4b). The backbone entropy is only one component that goes into the experimentally determined entropy. Therefore, the difference in backbone entropies cannot fully explain the difference in helical propensity. The difference, however, can be entirely accounted for by the energy required to move glycine from its preferred locations to the helical basin.

We have assumed that the most stringent coil library mimics the intrinsic conformational preferences of the amino

acids to calculate the backbone entropy. Our all-atom simulations of solvated capped tripeptides using a variety of force fields (5) also indicate that the difference in backbone entropy between alanine and glycine is relatively small,  $T\Delta S_{\text{Ala-Gly}} \sim 0.1 \text{ kcal mol}^{-1}$ . Hence, the small difference in backbone entropy between alanine and glycine is likely to be a robust quantity.

*Nearest Neighbor Effects.* According to Flory's isolated residue hypothesis (IRH) (57), the Ramachandran basin

populations of one residue are independent of the type and conformation of its NN. We investigate this hypothesis by calculating the basin populations of each residue in our improved coil library  $C_{\alpha\beta t}^{\text{intern}}$  as a function of the identity of the NNs (Figures 6 and 7). Our analysis considers the neighboring amino- and carboxy-terminal residues independently, as the number of occurrences of specific triplets is too low to provide adequate sampling for the simultaneous influence of both NNs. To further increase the sample size while still retaining the dominant steric effects (9), we separate NNs into three groups according to the steric properties of their side-chains: (a)  $\beta$ -branched side-chains (Val, Ile, and Thr); (b) aromatic side-chains (Trp, Phe, Tyr, and His); and (c) the remaining called alanine-like (except Gly and Pro).

The breakdown of the IRH is reflected by the marked differences in the Ramachandran maps of, for instance, alanine when it is flanked by different classes of NNs (Figure 6a). Furthermore, the NN's conformation strongly affects the backbone distribution, as illustrated for alanine in Figure 6b. These NN effects can be individually delineated when both the identity and the conformation are considered separately and when the NN is amino- (Figure 7a) or carboxy-terminal (Figure 7b) to an alanine. The  $\alpha$  and  $\beta$  preferences can change by up to 4-fold ( $\sim 0.8$  kcal mol<sup>-1</sup>) in some cases depending upon the NN type and conformation. These results demonstrate that the conformation of a backbone is influenced by the identity, conformation, and relative position of its neighbor (N- or C-terminal), in agreement with other studies (5, 6, 8–10, 58).

Many (59, 60) but not all (10, 61, 62) implementations of helix–coil theories do not include either a dependence on NN identity or conformation. Hence, our analysis presents an opportunity for improving these theories by correcting each residue's helix–coil equilibrium constant according to NN effects.

## CONCLUSIONS

We introduce an improved, more stringent coil library  $C_{\alpha\beta t}^{\text{intern}}$  to extract the intrinsic conformational preferences of residues within a protein. Residues in the three major classes of secondary structures and those flanking such structures are removed, along with other conformationally biased residues. These backbone preferences generate a statistical potential that accurately reproduces helical,  $\beta$ -sheet, and PPII propensities, which are often strongly affected by the identity and conformation of the neighboring residues. Therefore, locally determined backbone conformational preferences exert a strong influence on protein structure, folding, and energetics (63). In combination with tertiary context, these conformational preferences determine the structure and stability of the native state.

## ACKNOWLEDGMENT

We thank Profs. N. Kallenbach, G. Rose, R. S. Berry, S. W. Englander, and R. Baldwin and members of our group for comments and discussions.

## SUPPORTING INFORMATION AVAILABLE

Figure of entropy difference ( $T\Delta S$ ) versus bin size. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## REFERENCES

1. Levinthal, C. (1968) Are there pathways for protein folding? *J. Chim. Phys.* 65, 44–45.
2. Zwanzig, R., Szabo, A., and Bagchi, B. (1992) Levinthal's paradox, *Proc. Natl. Acad. Sci., U.S.A.* 89, 20–22.
3. Ramachandran, G. N., and Sasisekharan, V. (1968) Conformation of polypeptides and proteins, *Adv. Protein Chem.* 23, 283–438.
4. Krimm, S., and Tiffany, M. L. (1974) The circular dichroism spectrum and structure of unordered polypeptides and proteins, *Isr. J. Chem.* 12, 189–200.
5. Zaman, M. H., Shen, M. Y., Berry, R. S., Freed, K. F., and Sosnick, T. R. (2003) Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the Flory isolated-pair hypothesis for peptides, *J. Mol. Biol.* 331, 693–711.
6. Pappu, R. V., Srinivasan, R., and Rose, G. D. (2000) The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding, *Proc. Natl. Acad. Sci., U.S.A.* 97, 12565–12570.
7. Smith, L. J., Bolin, K. A., Schwalbe, H., MacArthur, M. W., Thornton, J. M., and Dobson, C. M. (1996) Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations, *J. Mol. Biol.* 255, 494–506.
8. Keskin, O., Yuret, D., Gursoy, A., Turkay, M., and Erman, B. (2004) Relationships between amino acid sequence and backbone torsion angle preferences, *Proteins* 55, 992–998.
9. Penkett, C. J., Redfield, C., Dodd, I., Hubbard, J., McBay, D. L., Mossakowska, D. E., Smith, R. A., Dobson, C. M., and Smith, L. J. (1997) NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein, *J. Mol. Biol.* 274, 152–159.
10. Gibrat, J. F., Garnier, J., and Robson, B. (1987) Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs, *J. Mol. Biol.* 198, 425–443.
11. Kang, H. S., Kurochkina, N. A., and Lee, B. (1993) Estimation and use of protein backbone angle probabilities, *J. Mol. Biol.* 229, 448–460.
12. Flory, P. J. (1953) *Statistical Mechanics of Chain Molecules*, Cornell University Press, Ithaca, NY.
13. Zaman, M. H., Shen, M. Y., Berry, R. S., and Freed, K. F. (2003) Computer simulation of met-enkephalin using explicit atom and united atom potentials: Similarities, differences, and suggestions for improvement, *J. Phys. Chem. B* 107, 1685–1691.
14. Mu, Y. G., Kosov, D. S., and Stock, G. (2003) Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments, *J. Phys. Chem. B* 107, 5064–5073.
15. Hu, H., Elstner, M., and Hermans, J. (2003) Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution, *Proteins* 50, 451–463.
16. Garcia, A. E., and Sanbonmatsu, K. Y. (2002) Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds, *Proc. Natl. Acad. Sci., U.S.A.* 99, 2782–2787.
17. Munoz, V., and Serrano, L. (1994) Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales, *Proteins* 20, 301–311.
18. Swindells, M. B., MacArthur, M. W., and Thornton, J. M. (1995) Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures, *Nat. Struct. Biol.* 2, 596–603.
19. Avbelj, F., and Baldwin, R. L. (2003) Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: distributions of phi, *Proc. Natl. Acad. Sci., U.S.A.* 100, 5742–5747.

20. Chou, P. Y., and Fasman, G. D. (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins, *Biochemistry* 13, 211–222.
21. Smith, L. J., Fiebig, K. M., Schwalbe, H., and Dobson, C. M. (1996) The concept of a random coil. Residual structure in peptides and denatured proteins, *Fold Des.* 1, R95–106.
22. Stites, W. E., and Pranata, J. (1995) Empirical evaluation of the influence of side chains on the conformational entropy of the polypeptide backbone, *Proteins* 22, 132–140.
23. O'Connell, T. M., Wang, L., Tropsha, A., and Hermans, J. (1999) The "random-coil" state of proteins: comparison of database statistics and molecular simulations, *Proteins* 36, 407–418.
24. Hovmöller, S., Zhou, T., and Ohlson, T. (2002) Conformations of amino acids in proteins, *Acta Crystallogr. D Biol. Crystallogr.* 58, 768–776.
25. Serrano, L. (1995) Comparison between the phi distribution of the amino acids in the protein database and NMR data indicates that amino acids have various phi propensities in the random coil conformation, *J. Mol. Biol.* 254, 322–333.
26. Shi, Z., Olson, C. A., Rose, G. D., Baldwin, R. L., and Kallenbach, N. R. (2002) Polyproline II structure in a sequence of seven alanine residues, *Proc. Natl. Acad. Sci., U.S.A.* 99, 9190–9195.
27. Shi, Z., Woody, R. W., and Kallenbach, N. R. (2002) Is polyproline II a major backbone conformation in unfolded proteins? *Adv. Prot. Chem.* 62, 163–240.
28. Eker, F., Griebenow, K., and Schweitzer-Stenner, R. (2003) Stable conformations of tripeptides in aqueous solution studied by UV circular dichroism spectroscopy, *J. Am. Chem. Soc.* 125, 8178–8185.
29. Pappu, R. V., and Rose, G. D. (2002) A simple model for polyproline II structure in unfolded states of alanine-based peptides, *Protein Sci.* 10, 2437–2455.
30. Creamer, T. P., and Campbell, M. N. (2002) Determinants of the polyproline II helix from modeling studies, *Adv. Protein Chem.* 62, 263–282.
31. Dukor, R. K., and Keiderling, T. A. (1991) Reassessment of the random coil conformation: vibrational CD study of proline oligopeptides and related polypeptides, *Biopolymers* 31, 1747–1761.
32. Wilson, G., Hecht, L., and Barron, L. D. (1996) Residual structure in unfolded proteins revealed by Raman optical activity, *Biochemistry* 35, 12518–12525.
33. Keiderling, T. A., and Xu, Q. (2002) Unfolded peptides and proteins studied with infrared absorption and vibrational circular dichroism spectra, *Adv. Protein Chem.* 62, 111–161.
34. Pappu, R. V., and Rose, G. D. (2002) A simple model for polyproline II structure in unfolded states of alanine-based peptides, *Protein Sci.* 11, 2437–2455.
35. Drozdov, A. N., Grossfield, A., and Pappu, R. V. (2004) Role of solvent in determining conformational preferences of alanine dipeptide in water, *J. Am. Chem. Soc.* 126, 2574–2581.
36. Mezei, M., Fleming, P. J., Srinivasan, R., and Rose, G. D. (2004) Polyproline II helix is the preferred conformation for unfolded polyalanine in water, *Proteins* 55, 502–507.
37. Kentsis, A., Mezei, M., Gindin, T., and Osman, R. (2004) Unfolded state of polyalanine is a segmented polyproline II helix, *Proteins* 55, 493–501.
38. Garcia, A. E. (2004) Characterization of nonalpha helical conformations in Ala peptides, *Polymer* 45, 669–676.
39. Chellgren, B. W., and Creamer, T. P. (2004) Short sequences of nonproline residues can adopt the polyproline II helical conformation, *Biochemistry* 43, 5864–5869.
40. Rucker, A. L., Pagar, C. T., Campbell, M. N., Qualls, J. E., and Creamer, T. P. (2003) Host–guest scale of left-handed polyproline II helix formation, *Proteins* 53, 68–75.
41. Kelly, M. A., Chellgren, B. W., Rucker, A. L., Troutman, J. M., Fried, M. G., Miller, A. F., and Creamer, T. P. (2001) Host–guest study of left-handed polyproline II helix formation, *Biochemistry* 40, 14376–14383.
42. Adzhubei, A. A., and Sternberg, M. J. (1993) Left-handed polyproline II helices commonly occur in globular proteins, *J. Mol. Biol.* 229, 472–493.
43. Fitzkee, N. C., and Rose, G. D. (2004) Steric restrictions in protein folding: an alpha-helix cannot be followed by a contiguous beta-strand, *Protein Sci.* 13, 633–639.
44. Wang, G., and Dunbrack, R. L., Jr. (2003) PISCES: a protein sequence culling server, *Bioinformatics* 19, 1589–1591.
45. Richmond, T. J. (1984) Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect, *J. Mol. Biol.* 178, 63–89.
46. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. (2003) Structure validation by Alpha geometry: phi, psi and Cbeta deviation, *Proteins* 50, 437–450.
47. Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* 22, 2577–2637.
48. Bai, Y., and Englander, S. W. (1994) Hydrogen bond strength and beta-sheet propensities: the role of a side chain blocking effect, *Proteins* 18, 262–266.
49. Pace, C. N., and Scholtz, J. M. (1998) A helix propensity scale based on experimental studies of peptides and proteins, *Biophys. J.* 75, 422–427.
50. Smith, C. K., Withka, J. M., and Regan, L. (1994) A thermodynamic scale for the beta-sheet forming tendencies of the amino acids, *Biochemistry* 33, 5510–5517.
51. Smith, C. K., and Regan, L. (1995) Guidelines for protein design: the energetics of beta sheet side chain interactions, *Science* 270, 980–982.
52. Minor, D. L., Jr., and Kim, P. S. (1994) Context is a major determinant of beta-sheet propensity, *Nature* 371, 264–267.
53. Dill, K. A. (1990) Dominant forces in protein folding, *Biochemistry* 29, 7133–7155.
54. Street, A. G., and Mayo, S. L. (1999) Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone, *Proc. Natl. Acad. Sci., U.S.A.* 96, 9074–9076.
55. D'Aquino, J. A., Gomez, J., Hilsner, V. J., Lee, K. H., Amzel, L. M., and Freire, E. (1996) The magnitude of the backbone conformational entropy change in protein folding, *Proteins* 25, 143–156.
56. Creamer, T. P., and Rose, G. D. (1994) Alpha-helix-forming propensities in peptides and proteins, *Proteins* 19, 85–97.
57. Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, Wiley, New York.
58. Wang, J. Y., Ahmad, S., Gromiha, M. M., and Sarai, A. (2004) Look-up tables for protein solvent accessibility prediction and nearest neighbor effect analysis, *Biopolymers* 75, 209–216.
59. Zimm, G. H., and Bragg, J. K. (1959) Theory of the phase transition between helix and random coil in polypeptide chains, *J. Chem. Phys.* 31, 526–535.
60. Lifson, S., and Roig, A. (1961) On the theory of the helix-coil transition in polypeptides, *J. Chem. Phys.* 34, 1963–1974.
61. Poland, D. C., and Scheraga, H. A. (1965) *Statistical Mechanics of Noncovalent Bonds in Polyamino Acids*. 4. Matrix Treatment of Hydrophobic Bonds in Random Coil and of Helix-Coil Transition for Chains of Arbitrary Length, *Biopolymers* 3, 315.
62. Go, M., Hesselin, F. T., Go, N., and Scheraga, H. A. (1974) *Molecular Theory of Helix-Coil Transition in Poly(Amino Acids)*. 4. Evaluation and Analysis of S for Poly(L-Valine) in Absence and Presence of Water, *Macromolecules* 7, 459–467.
63. Baldwin, R. L., and Rose, G. D. (1999) Is protein folding hierarchic? I. Local structure and peptide folding, *Trends Biochem. Sci.* 24, 26–33.

BI0474822