JCTC Journal of Chemical Theory and Computation

Is Alanine Dipeptide a Good Model for Representing the Torsional Preferences of Protein Backbones?

Michael Feig*

Department of Biochemistry and Molecular Biology and Department of Chemistry, Michigan State University, East Lansing, Michigan 48824

Received May 7, 2008

Abstract: The conformational preference for different ϕ/Ψ backbone torsion angles is a key determinant of peptide and protein secondary structure. Often, dipeptides are used as models for understanding protein backbone dynamics and to derive force field parameters. Here, the question is examined to what extent the conformational preferences in dipeptides reflect the backbone dynamics in polypeptides and proteins and to what extent an alanine dipeptide-based backbone torsion parametrization can lead to accurate reproduction of amino acid dependent ϕ/Ψ preferences in protein structures. Results from a comparison of the analysis of Protein Data Bank (PDB) structures with long simulations of selected proteins and amino acid dipeptides suggest that a common alanine dipeptide-based torsion potential does in fact lead to excellent agreement between protein simulations and PDB structures. At the same time, the ϕ/Ψ preferences in the dipeptides are significantly different, suggesting that dipeptides are not good model systems for studying protein backbone dynamics.

Introduction

Protein structure and dynamics are essential determinants of biological function. The structure of most proteins consists of well-defined three-dimensional folds that are built from α-helical and β-sheet secondary structure elements with connecting turns and loops. The ability to form different secondary structure elements is primarily a reflection of the conformational flexibility of the polypeptide backbone. There are essentially two backbone degrees of freedom for each amino acid residue: the torsion angles ϕ (C−N−C_α−C) and Ψ (N−C_α−C−N). For nonglycine and nonproline residues, the well-known Ramachandran plot[1] of Ψ versus ϕ identifies two major minima: α_R ($\phi = -60$, $\Psi = -50$) in the α basin and PPII/C5 ($\phi = -60$ to -170, $\Psi = 120-170$) in the β basin (see Figure 1), which correspond to α-helical and extended β-strand/-sheet secondary structures when repeated over multiple amino acid residues. Secondary minima with higher relative free energies at α_L ($\phi = 50$, $\Psi = 50$) and $C7_{ax}$ ($\phi = 50$, $\Psi = -130$) are relevant in the formation of turns and loops. The conformational preferences of proline are restricted to α_R, PPII ($\phi = -60$, $\Psi = 140$), and $C7_{eq}$ (ϕ

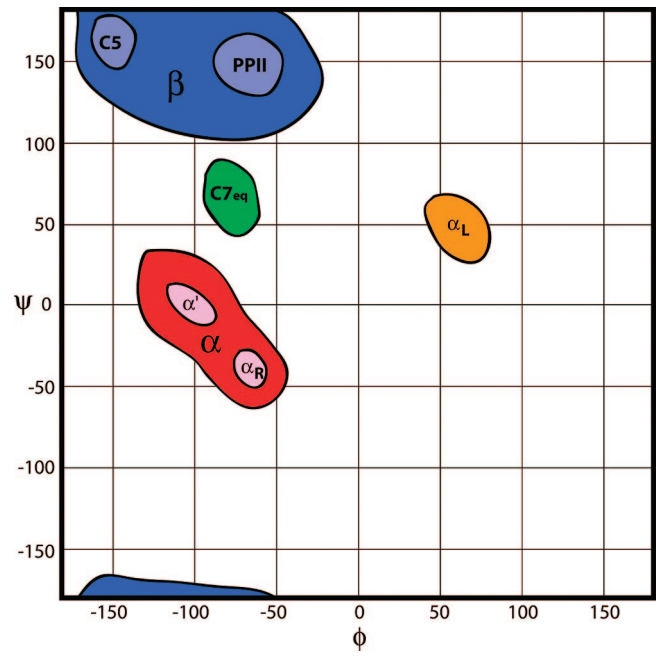

Figure 1. Schematic overview of major conformational basins sampled by ϕ/Ψ backbone torsion angles in nonglycine, nonproline peptide residues.

* Phone: (517) 432-7439. Fax: (517) 353-9334. E-mail: feig@msu.edu.

10.1021/ct800153n CCC: $40.75
Published on Web 07/30/2008

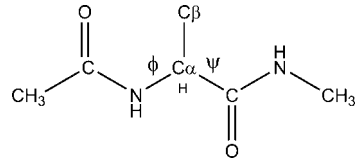

Figure 2. Blocked alanine dipeptide structure.

$= -75, \Psi = 75$) conformations. Glycine does not distinguish between positive and negative ϕ values due to its achiral nature and therefore visits the right side of the Ramachandran plot more frequently compared to the other amino acids. The preference for certain regions of the Ramachandran plot has resulted in the definition of "allowed" versus "disallowed" regions that are often applied in the validation of experimental and theoretical structures.[2,3] However, extensive analysis of high-resolution crystal structures has revealed that a significant fraction of amino acid residues may also exhibit ϕ/Ψ torsion angles outside those canonical regions.[3−5] Furthermore, the preferences for ϕ/Ψ torsions vary not just between proline and glycine but also among nonproline and nonglycine amino acid residues.[6,7] From detailed analyses in previous studies[6,7] and data gathered in this study, it has emerged that, in the β basin, alanine, tryptophan, phenylalanine, tyrosine, serine, glutamine, glutamic acid, arginine, lysine, methionine, and cysteine follow a similar energy landscape with two minima near C5 and PPII that are connected by a very shallow barrier. Valine, isoleucine, and to a lesser extent leucine have instead a single, broad minimum near ($\phi = -120$, $\Psi = 130$) that is intermediate between C5 and PPII and shifted downward to smaller Ψ angles. Aspartic acid, asparagine, and to a lesser degree histidine have an overall much broader β region that extends to $\Psi = 75$ and includes $C7_{eq}$. Finally, threonine exhibits four clearly discernible minima in the extended region, two at $\Psi = 165$ and two at $\Psi = 130$, all of which are connected by very shallow barriers. The energy landscape in the α_R basin also varies between different amino acids. Alanine, tryptophan, leucine, glutamine, arginine, glutamic acid, and methionine predominantly sample α_R ($\phi = -60$, $\Psi = -50$), while phenylalanine, tyrosine, asparagine, serine, threonine, histidine, lysine, aspartic acid, and cysteine also sample a second minimum at ($\phi = -100$, $\Psi = 0$) to a significant degree. Valine and isoleucine stand out by an extended low-energy region that includes ($\phi = -100$, $\Psi = 50$), which corresponds to π-helical conformations. These subtle but significant variations in ϕ/Ψ preferences can be rationalized in part by examining correlations with side-chain torsions.[7] Finally, a special case is given by amino acids that immediately precede proline.[3,8,9] These pre-Pro residues do not significantly populate conformations near ($\phi = -90$, $\Psi = 0$) in the α_R basin but instead populate so-called ζ conformations near ($\phi = -140$, $\Psi = 70$).

Blocked alanine dipeptide (see Figure 2) is commonly studied as a prototype of nonglycine/nonproline protein backbones since it allows full sampling of the ϕ/Ψ conformational space without the additional complexity of side-chain degrees of freedom. Numerous computational and experimental studies of alanine dipeptide have explored its thermodynamic,[10−14] kinetic,[13,15,16] and spectroscopic[17,18] properties. Furthermore, alanine dipeptide and sometimes alanine tri- or tetrapeptides[19] are commonly used for the testing and parametrization of amino acid backbones in molecular mechanics force fields.[19−21] Generally, a modular approach is followed, where alanine dipeptide-derived bonded and nonbonded parameters are used for modeling the backbone of all nonglycine and nonproline residues.[22] However, in some force fields, for example, Amber,[23] the partial charges of the backbone atoms may vary slightly for each amino acid.

The development of amino acid backbone parameters based on alanine dipeptide commonly relies on *ab initio* calculations since sufficiently detailed thermodynamic or kinetic data are not available from experiments. *Ab initio* calculations typically provide reference conformational energies in a vacuum for selected conformers. Recently, it has become possible to obtain conformational energies of alanine dipeptide from high-level theory over the entire range of ϕ/Ψ values on a grid with 15° resolution. This data has allowed much finer control in the parametrization of ϕ/Ψ torsion parameters and challenged the established paradigm of using a combination of univariate Fourier-series torsion potentials to generate the torsion potential.[21] In order to better represent the complex features of the ϕ/Ψ free energy landscape, a map-based spline-interpolated cross-correlation term (CMAP) has been introduced into the CHARMM force field.[21,24] The CMAP term can directly reproduce any given ϕ/Ψ map in alanine dipeptide and has been used in particular to reflect the vacuum conformational energies from the *ab initio* calculations. With the CMAP correction, the torsional preferences in peptides and proteins were found to be substantially improved by reducing an overemphasis on the sampling of π-helical structures[25] and by reducing deviations from crystallographic structures in molecular dynamics simulations.[24]

The possibility to exactly reproduce the *ab initio* ϕ/Ψ map of alanine dipeptide with the CMAP formalism raises the issue of whether parameters derived from alanine dipeptide in a vacuum are appropriate for all of the other (nonglycine and nonproline) amino acids and for condensed phase environments. More specifically, the question is whether the sampling of ϕ/Ψ torsion angles in protein simulations with a common underlying torsion potential reproduces the amino acid type-dependent variations found in crystallographic structures. A secondary point of biophysical interest is to what extent the ϕ/Ψ preferences observed for a given amino acid in the context of protein structures are inherently present at the dipeptide level or are a result of interactions due to the polypeptide and protein environments. In order to probe these questions, long-time molecular dynamics simulations of all amino acid dipeptides and selected proteins were carried out and compared with data extracted from crystallographic structures in the Protein Data Bank.[26] The results demonstrate that a single torsion potential is largely sufficient to reproduce the subtle variations in ϕ/Ψ preferences in the context of proteins. Furthermore, it is found that residue type-dependent variations in ϕ/Ψ preferences are largely absent at the dipeptide level and only fully materialize in the context of protein structures. The results are described and discussed

Table 1. Overview of Simulated Systems

system	starting structure	residues	box length [Å]	simulation length [ns]
dipeptides	extended	1	31.5−35.3	150.0
protein G	3GB1	56	61.0	50.0
ubiquitin	1D3Z	76	69.3	22.0
barnase	1A2P	108	56.9	148.0
barstar	1BTA	89	52.4	142.9
CheY	1CYE	129	56.4	124.7
FKBP12	1FKS	107	62.6	143.5
RNase A	2AAS	124	59.7	148.3
RNase H	2RN2	155	69.5	121.5

in more detail in the following, after a summary of the methodology used in this study.

Methods

Molecular dynamics simulations in explicit solvent were carried out for all amino acid dipeptides and eight small- to medium-size proteins. Dipeptides were blocked with an acetyl group at the N terminus and with N-methylamide at the C terminus. Each amino acid dipeptide was simulated in its standard protonation state at pH = 7. Two simulations were run for histidine, one protonated at N_δ, the other one at N_ε. In all cases except proline, the starting structure was a fully extended peptide with backbone torsions near ($\phi = -160$, $\Psi = 130$). The starting structure for proline was near ($\phi = -60$, $\Psi = 160$). The dipeptides were solvated with explicit water in a cubic box with at least 12 Å from any atom in the dipeptide to the closest edge of the box. Charged amino acids were neutralized with either a single chlorine or sodium ion, placed initially by randomly replacing one of the water molecules. Initial configurations were briefly minimized and then heated up to 298 K during a series of simulations with 1 ps at 50 K, 1 ps at 100 K, 1 ps at 150 K, 2 ps at 200 K, 2 ps at 250 K, 2 ps at 275 K, and 2 ps at 300 K. Further equilibration was carried out with three simulations at 300 K over 4 ps each. For each run, the average water density at the edge of the box was calculated and compared to the expected number density of water of 0.03337/Å^3 at 300 K and 1 atm of pressure. If any deviation was found, the box size was adjusted accordingly. Simulations in the NVT ensemble were then continued for another 150 ns to generate the production trajectories used for analysis. The CHARMM22 force field[20] with the CMAP torsion potential[21] and updated tryptophan parameters[27] was used for the dipeptide. Modified TIP3P[28] parameters from the CHARMM force field were used to model the explicit water. Ion parameters were taken from Roux.[29] Periodic boundaries were employed to avoid solvent boundary artifacts. Electrostatic interactions were calculated with the particle-mesh Ewald method[30] using a 32 × 32 × 32 grid for the discrete fast Fourier transform (FFT) and a 9 Å direct space cutoff. During the simulation, SHAKE[31] was applied to constrain the lengths of bonds involving hydrogen so that an integration time step of 2 fs could be used. The temperature was controlled with the Nosé−Hoover algorithm.[32]

Eight proteins were simulated over 22−148 ns. Table 1 summarizes the simulation details for each protein. Table 2 shows the number of each amino acid from the combined set of proteins. All proteins were started from experimental structures and solvated with sufficient counterions to neutralize each system. The same equilibration protocol and simulation parameters as described above for the dipeptide simulations were applied, except that larger FFT grid sizes were used according to the increased box sizes.

Table 2. Number of Each Amino Acid in Simulated Proteins and Protein Data Bank (PDB) Structures[a]

amino acid	Protein simulations	PDB chains
alanine	66	56716
arginine	35	34832
asparagine	37	28621
aspartic acid	49	39407
cysteine	14	8740
glutamine	38	25817
glutamic acid	57	46199
glycine	62	50196
histidine	10	15522
isoleucine	41	38399
leucine	63	62782
lysine	61	38784
methionine	14	14065
phenylalanine	19	27547
proline	27	32923
serine	43	39330
threonine	63	36648
tryptophan	15	10187
tyrosine	28	24099
valine	50	48320

[a] Pre-proline residues are not included in non-proline amino acid totals.

All of the simulations were run with the CHARMM program[33] in conjunction with the MMTSB Tool Set.[34] A trajectory analysis was also carried out with CHARMM and the MMTSB Tool Set.

The analysis of Protein Data Bank (PDB) structures was performed on the basis of 3326 chains from crystal structures with 2.0 Å resolution or better and not more than 25% sequence identity between any two chains. The list of chains was generated with the protein structure culling server PISCES[35] in June 2007. Table 2 shows the number of each amino acid in the analyzed PDB structures.

Results

Protein Simulations. The conformations sampled during the protein simulations were compared to experimental structures to gauge the degree of realism in the simulations. Experimental structures of monomeric, wild-type apo forms are available from both X-ray crystallography and NMR spectroscopy for all of the systems simulated here with one exception. The crystal structure of barstar was taken from the complex of barstar with ribonuclease Sa (PDB code: 1AY7). Average and final root-mean-square deviation (rmsd) values during the simulation as well as the rmsd of the average structure over the entire trajectory are reported in Table 3. The latter is the most appropriate measure when comparing to the experimental data. In general, the rmsd of the average is lower than the average instantaneous rmsd values. Furthermore, in all cases, the deviation from the crystallographic structures is less than the deviation from

Table 3. Root Mean Square Deviations from Experimental Structures in Protein Simulations (Standard Deviations Are Given in Parentheses)

system	reference	type	avg. C_α rmsd [Å]	C_α rmsd of final structure [Å]	C_α rmsd of avg. structure [Å]
protein G	3GB1	NMR	1.06(0.20)	1.43	0.79
	1PGB	X-ray	0.81(0.21)	0.84	**0.41**
ubiquitin	1D3Z	NMR	1.41(0.20)	1.28	1.25
	1UBQ	X-ray	1.24(0.18)	1.13	**1.04**
barnase	1FW7	NMR	1.71(0.15)	1.67	1.35
	1A2P	X-ray	1.54(0.25)	1.37	**1.15**
barstar	1BTA	NMR	1.34(0.16)	1.21	1.15
	1AY7B	X-ray	0.97(0.12)	0.85	**0.67**
CheY	1CYE	NMR	1.43(0.20)	1.70	1.18
	3CHY	X-ray	1.13(0.17)	1.14	**0.84**
FKBP12	1FKS	NMR	3.58(0.74)	4.77	2.74
	1FKK	X-ray	3.58(0.63)	4.58	**2.68**
RNase A	2AAS	NMR	2.49(0.43)	3.21	2.04
	8RAT	X-ray	2.18(0.34)	2.70	**1.58**
RNase H	1RCH	NMR	2.78(0.17)	2.89	2.54
	2RN2	X-ray	1.98(0.23)	2.01	**1.62**

the NMR structure. The rmsd of the average from the crystallographic structure is less than 1 Å for three of the proteins studied here and between 1 and 2 Å for four other systems. For FKBP12, the deviation is larger, 2.68 Å, due to large fluctuations of residues 32−45 and 80−95, which consist mostly of long loop regions. It is likely that even 150 ns is not sufficient to fully sample the conformational space of those flexible regions and that much longer simulations might be required to improve the agreement with the experimental structures that are averaged over much longer time scales and over a large number of molecules. While the small deviations of the average simulated structures from the experimental structures indicate a high level of realism in the simulations, the larger average instantaneous rmsd values with significant standard deviations indicate broad conformational sampling well beyond the time- and ensemble-averaged experimental structures.

ϕ/Ψ Sampling in Protein Simulations versus PDB Structures. The distribution of ϕ/Ψ backbone torsion angles was analyzed from the protein simulations as a function of the amino acid type and compared to the distributions from PDB structures. Results for selected amino acids representative of major variations in ϕ/Ψ sampling are shown in Figures 3 and 4. Data for all of the other amino acids are given in Figure S1 in the Supporting Information. The agreement between the results from the simulations and from the PDB is remarkably good, especially in the lower-energy regions. A prominent difference is the significant population of high-energy regions in the simulations from instantaneous conformational sampling over very long simulations. Most of these regions are populated only sparsely in the PDB structures.

From a more detailed comparison of the PDB distributions with the simulation results, it can be seen that the subtle variations as a function of amino acid type are reproduced well. In particular, there is good agreement in the following key features: In asparagine, the β region is more extended, the transition region between the α and β basins is lowered, the preference for a second minimum in the α-helical basin at ($\phi = -100$, $\Psi = 0$) is more pronounced, and the sampling of α_L conformations is relatively favorable. In threonine, the β region is split into four distinct minima and the preference for ($\phi = -100$, $\Psi = 0$) is enhanced and extended toward ($\phi = -140$, $\Psi = -30$). Finally, in valine, there is a broad minimum near ($\phi = -120$, $\Psi = 130$) and sampling of fully extended conformations near ($\phi = -180$, $\Psi = 180$) is reduced while the α-helical basin extends to ($\phi = -140$, $\Psi = -20$) and ($\phi = -100$, $\Psi = -50$).

It is remarkable how well the sequence-dependent variations in ϕ/Ψ preferences are reproduced with a single alanine−dipeptide-based CMAP torsion angle term, but there are also some deficiencies that could possibly be addressed through force field adjustments: In general, it appears that fully extended conformations near C5 are slightly too favorable over PPII conformations. Furthermore, sampling of the $C7_{eq}$ conformation near ($\phi = -75$, $\Psi = 75$) in the α/β transition region appears to be too unfavorable, which is especially apparent in alanine and asparagine. In asparagine and to a lesser extent in alanine, there is a third minimum in the simulations near ($\phi = -160$, $\Psi = 0$) which is not seen in the PDB distributions. Finally, valine did not sample the right side of the Ramachandran plot in the simulations. However, it is likely that this may be a reflection of the limited set of simulated structures rather than inherent force field deficiencies since a valine residue not initially found in a conformation with positive ϕ angles is unlikely to be able to assume such a conformation without major structural disruption unless it is located in a flexible loop region.

The backbone conformational preferences of proline and glycine residues are compared in Figure 4. It should be noted that different CMAP torsion potentials are used for those residues in the CHARMM force field to separately reproduce the *ab initio* ϕ/Ψ maps for proline and glycine dipeptide. The overall features of both maps are reproduced well between the simulations and PDB distributions, although there are also some notable differences: In glycine, there appears to be a lack of a clear minimum at α_R in the simulations. Instead, there is a minimum at ($\phi = -80$, $\Psi = 10$). There are also minima at ($\phi = -180$, $\Psi = -25$) and ($\phi = -160$, $\Psi = 30$) next to a high-energy region that do not seem to match the ϕ/Ψ preferences in the experimental structures, while the

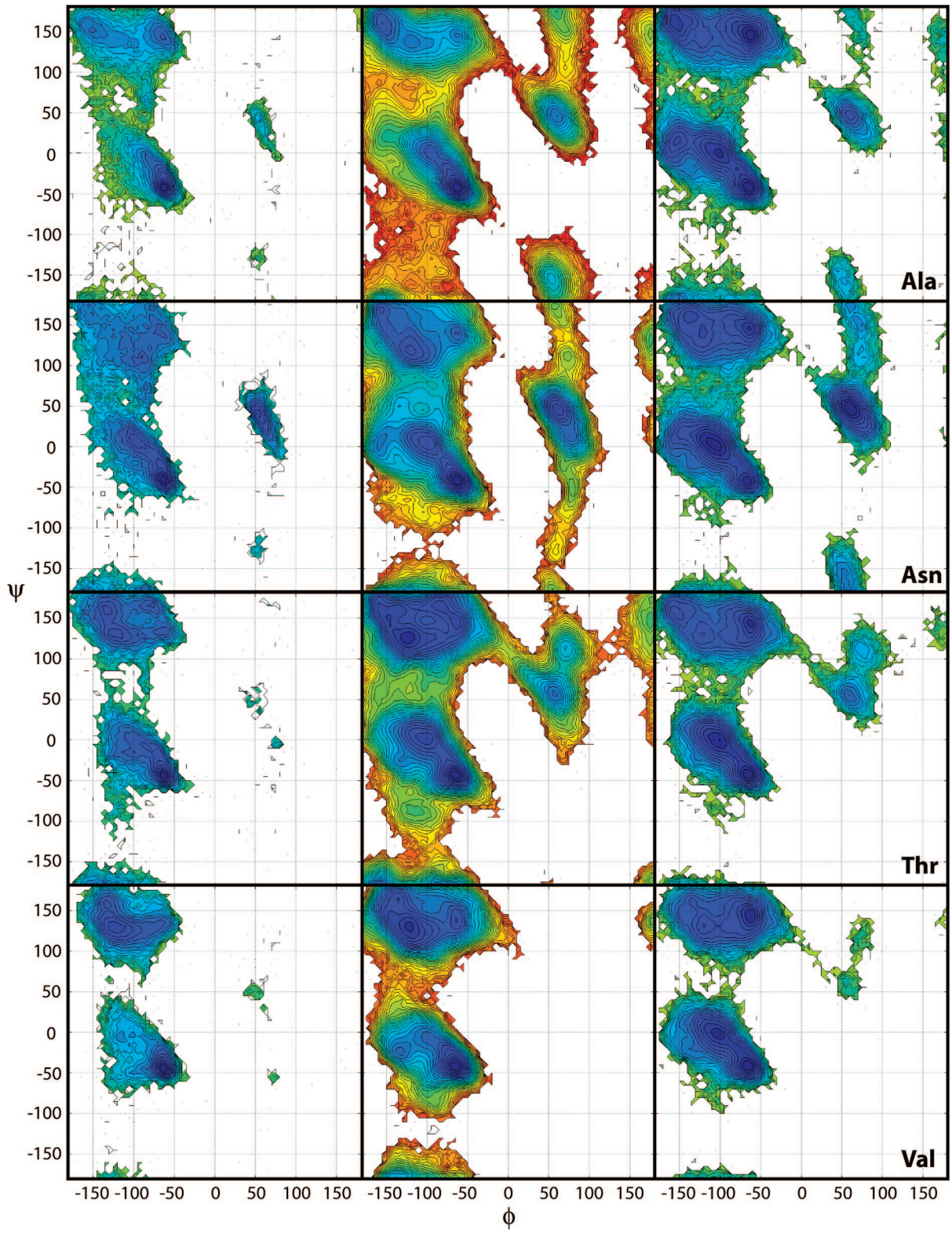

Figure 3. Potentials of mean force for the sampling of ϕ/Ψ backbone torsion angles in selected amino acid residues from PDB structures (left column), protein simulations (center column), and dipeptide simulations (right column). A color bar indicating the energy levels is given in Figure 4.

relative energy of $C7_{eq}$ conformations ($\phi = -75$, $\Psi = 75$) appears to be too high. In proline, the two major minima at α_R and PPII are reproduced reasonably well, but the $C7_{eq}$ conformation is again not favorable enough. Furthermore, the entire transition region is shifted to more negative ϕ angles.

Further analysis was carried out to compare the relative sampling of conformations in the major basins (α, β, and α_L). Table 4 shows the results for nonglycine and nonproline amino acids. The preference for sampling in the α basin versus the β basin matches to a large extent known secondary structure propensities,[36−38] especially for alanine and glutam-

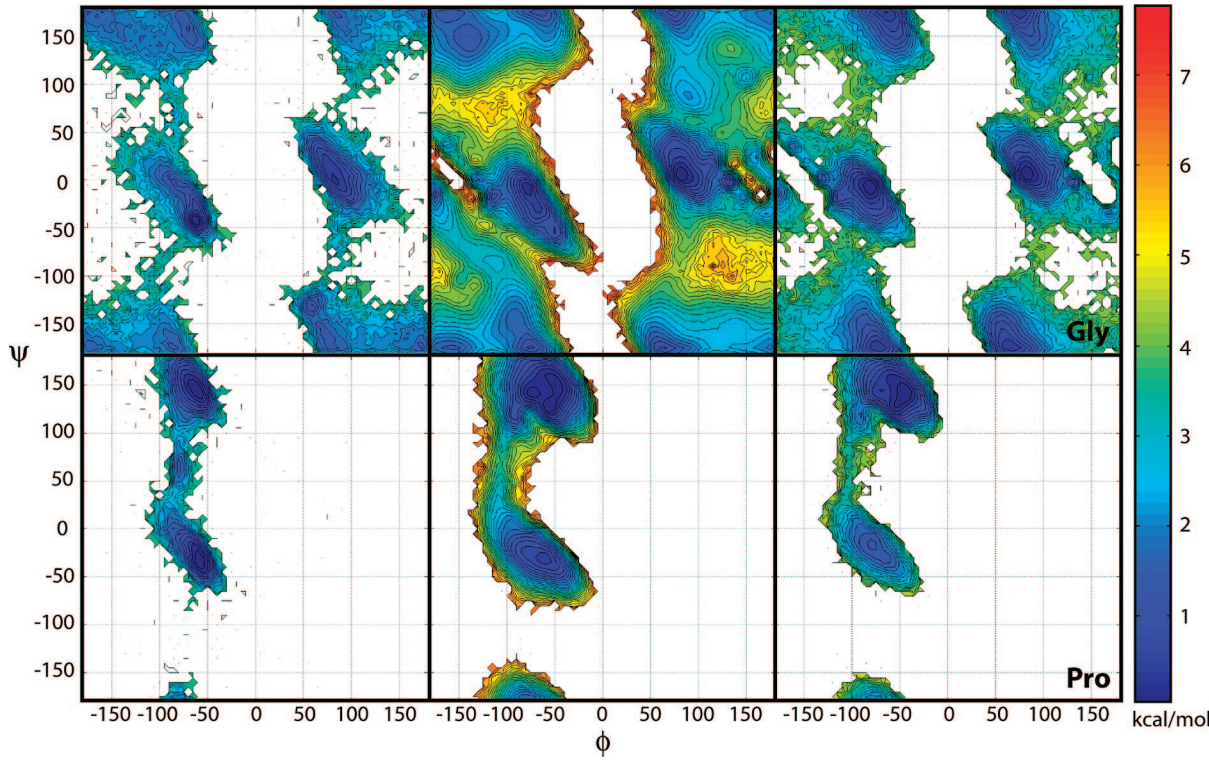

Figure 4. Potentials of mean force for the sampling of ϕ/Ψ backbone torsion angles in glycine and proline as in Figure 3.

Table 4. Relative Sampling (in %) of Different Regions in the Ramachandaran Plot for Each Amino Acid in PDB Structures, Simulated Proteins, and Dipeptides[a]

	PDB chains			protein simulations			dipeptides		
amino acid	α (α_R/α')	β	α_L	α (α_R/α')	β	α_L	α (α_R/α')	β	α_L
Ala	**66** (8.1)	31	1.2	**73** (4.4)	23	**3.6**	48 (0.80)	48	2.7
Arg	**60** (4.8)	36	2.5	**52** (3.3)	43	**4.4**	47 (0.68)	**50**	**3.0**
Asn	**51** (1.4)	33	**12**	**51** (1.7)	32	**16**	46 (0.28)	31	**21**
Asp	**57** (2.1)	34	**5**	**61** (2.2)	37	0.0	48 (1.41)	49	2.7
Cys	45 (3.2)	**51**	2.2	28 (3.4)	**67**	2.1	42 (0.59)	**51**	**6.4**
Gln	**64** (4.9)	31	2.7	**59** (4.3)	34	**6.1**	46 (0.57)	45	**8.4**
Glu	**68** (6.3)	29	1.7	**55** (4.1)	39	**4.7**	38 (1.46)	**60**	2.4
His	**52** (2.4)	41	**4.0**	23 (12)	29	**48**	36 (0.44)	42	**20**
Ile	47 (9.0)	**53**	0.1	33 (6.2)	**67**	0.2	49 (0.40)	**51**	0.2
Leu	**61** (5.5)	37	0.7	**57** (3.9)	42	0.0	46 (1.05)	**51**	2.3
Lys	**62** (5.1)	34	**3.0**	48 (6.4)	45	**5.8**	42 (0.79)	48	**8.8**
Met	**58** (5.2)	39	1.6	48 (6.5)	45	**5.2**	**54** (0.95)	45	1.1
Phe	49 (3.5)	48	1.7	23 (115)	**75**	0.3	42 (0.46)	46	**10**
Ser	**51** (2.6)	45	1.7	44 (1.6)	**50**	**4.9**	33 (0.69)	46	**19**
Thr	49 (2.2)	**50**	0.4	31 (1.3)	**67**	1.2	**57** (0.69)	41	1.9
Trp	**53** (4.5)	45	1.4	46 (12)	39	**14**	**51** (0.54)	46	2.5
Tyr	49 (3.1)	48	1.7	34 (1.2)	**61**	**3.9**	**52** (0.39)	38	**9.1**
Val	42 (7.9)	**58**	0.2	34 (7.8)	**66**	0.0	**55** (0.84)	45	0.1

[a] Pre-proline residues are not included in non-proline amino acid totals. The α basin is defined by the rectangle spanned by ($\phi = -180$, $\Psi = 50$) and ($\phi = 0$, $\Psi = -100$) with the α_R minimum at ($\phi = -80$, $\Psi = -15$) to ($\phi = -35$, $\Psi = -60$) and the secondary minimum (α') at ($\phi = -140$, $\Psi = 40$) to ($\phi = -60$, $\Psi = -15$). The β basin is defined by ($\phi = -210$, $\Psi = 210$) to ($\phi = 0$, $\Psi = 85$) without further subdivision because of a wide variation in sampling in different amino acids. The α_L basin is defined by ($\phi = 0$, $\Psi = 100$) to ($\phi = 110$, $\Psi = -25$). α/β propensities larger than 50% and α_L propensities larger than 3% are highlighted in bold.

ic acid, which have high helical propensities.[36] However, it is interesting that the relative sampling of α_R versus α' conformations within the α basin seems to be an overall better predictor of the propensity to form α helices. All of the amino acids with significant propensities to form α helices[36] (Ala, Gln, Glu, Ile, Leu, Lys, Met, Phe, Trp, and Val) strongly favor α_R sampling over α' compared to the remaining residues, with arginine being the only exception. Conformations on the right-hand side of the Ramachandran plot are important in the formation of turns. The propensity to form α_L conformations is exceptionally high in asparagine and significant in aspartic acid, histidine, and lysine. This correleates with the high frequency of asparagine and aspartic acid residues in turn regions.[36]

The $\alpha/\beta/\alpha_L$ propensities from the protein simulations agree qualitatively with the results from the PDB analysis for most amino acids. Larger deviations are found for amino acids with a small number of representatives in the test sets, in

particular cysteine, tryptophan, histidine, methionine, and phenylalanine, where the results presented here may not be statistically relevant. However, since only amino acids in flexible loop regions and at the termini are able to undergo conformational transitions between the major conformational basins without disrupting the overall structure, the simulations largely reflect the specific distribution of secondary structure elements in the simulated proteins rather than amino acid dependent propensities to form different secondary structure elements according to the underlying force field. The relative sampling of α_R versus α' conformations is also in good qualitative agreement between the simulations and PDB distributions (if amino acids that are rare in the simulations are excluded again). Overall, the ratio of α_R to α' sampling is smaller in the simulations (average, without Cys, Trp, His, Met, and Phe: $\alpha_R/\alpha' = 3.7$) compared to the PDB analysis (average $\alpha_R/\alpha' = 4.8$), suggesting that the sampling of α' might be too favorable in the simulations.

ϕ/Ψ Preferences in Dipeptides versus Proteins. The comparison of ϕ/Ψ preferences between the protein simulations and PDB structures provides an idea of how well the computational methodology can reproduce experimental data. On the other hand, a comparison of ϕ/Ψ preferences between protein simulations or experimental data and dipeptide simulations addresses the more fundamental question of to what extent amino acid dependent variations in ϕ/Ψ preferences found in proteins are already apparent at the dipeptide level. The results in Figure 3 show that the ϕ/Ψ preferences vary only to a small degree between different amino acid dipeptides, suggesting that amino acid dependent variations in ϕ/Ψ preferences do in fact stem predominantly from interactions due to polypeptide and protein environments. Closer inspection reveals some differences between different amino acids. Most significant are variations in the preferences for positive ϕ values. As in the protein simulations (and PDB distributions), asparagine dipeptide samples α_L more frequently, while valine dipeptide samples α_L less frequently than the alanine and threonine dipeptides. Furthermore, alanine and asparagine dipeptides (as well as arginine, cysteine, glutamine, glutamic acid, histidine, methionine, serine, tryptophan, and phenylalanine; see Figure S1, Supporting Information) extend the α basin toward ϕ values near -170, while the other dipeptides do not significantly populate that region. In the β basin, the overall minimum lies at PPII for all dipeptides, but very subtle variations in the conformational landscape of the β basin are apparent. These small differences, for example, diminished sampling near ($\phi = -170$, $\Psi = -170$) for valine, partially mimic the more pronounced variations in the β-basin landscape in the protein simulations and PDB structures but are far from completely reproducing the amino acid dependent variations seen in the protein context. Conformational preferences of glycine and proline dipeptides mostly resemble the preferences within the protein simulations, but differences in sampling fully extended conformations and the transition region near ($\phi = -100$, $\Psi = -100$) are apparent in glycine.

The relative sampling of the major conformational basins in the dipeptides also differs from the protein simulations and PDB structures (see Table 4). The relative sampling of the α basin is generally at or below 50% and less than the relative percentage in the PDB structures for most amino acids. Exceptions are threonine and valine, where conformations in the α basin are sampled more often in the dipeptide than in the protein context. The strong preference for α-helical conformations in alanine, glutamine, glutamic acid, leucine, and lysine found in the PDB structures is not apparent at the dipeptide level. It is particularly remarkable that glutamic acid, which is known to be a strong helix-forming amino acid,[36] actually has the highest propensity for extended structures at the dipeptide level compared to all of the other amino acids. Furthermore, the ratios of α_R to α' sampling are much lower in the dipeptides, mostly below 1, indicating that the sampling of α_R conformations is relatively disfavored in the dipeptides. Therefore, the polypeptide context and, in particular, the ability to form i, $i + 4$ backbone hydrogen bonding is essential in stabilizing α-helical secondary structure elements.

The α_L conformations are sampled at widely varying levels in the dipeptides. Asparagine, histidine, and serine dipeptides spend nearly 20% of the time in the α_L conformation, while isoleucine and valine essentially never sample α_L. This can be understood as a result of attractive intramolecular electrostatics between asparagine, histdine, and serine side chains and the peptide backbone and unfavorable side chain backbone interactions in the case of isoleucine and valine. The amino acid dependent propensities for α_L conformations in the dipeptides do not agree very well with the results from the protein simulations or PDB. However, an overall increased preference for α_L conformations compared to the PDB structures is apparent in both the dipeptide and protein simulations. This finding may suggest a need for raising the energy of α_L conformations in the force field.

Discussion and Conclusion

Previous studies have examined the detailed distribution of ϕ/Ψ torsion angles in experimental structures as a function of the amino acid type.[6,7] Here, these results are compared with torsional preferences from extensive simulations of proteins and dipeptides. The torsional preferences in the protein simulations are in good qualitative and quantitative agreement with the distribution of ϕ/Ψ angles found in PDB structures. Variations as a function of the amino acid type are generally represented well, including subtle features in the detailed energy landscape of the α and β basins. In contrast, ϕ/Ψ preferences in amino acid dipeptides vary much less as a function of the amino acid type. Some of the amino acid dependent variations seen in the context of proteins are also apparent at the dipeptide level, such as the preference for α_L sampling, but other features such as the preference for α-helical conformations in glutamic acid and the strong tendency to sample α_R conformations over α' are not reproduced in the dipeptide simulations. These results suggest that local interactions at the residue level play only a small role in determining the sequence-dependent torsional preferences of peptide backbones, while the more important contributions come from long-range interactions in the context of polypeptide chains and protein structures. An

example is the observation of helix-capping interactions by glutamic acid residues that are not present at the dipeptide level.[39]

The same underlying CMAP torsion potential was used in all of the simulations. One of the main questions prompting this study is whether a common CMAP torsion potential for all nonglycine/nonproline residues is sufficient to accurately reproduce the sequence-dependent variations in ϕ/Ψ preferences. On the basis of the results presented here, this is apparently the case, further supporting the idea that the observed modulation of ϕ/Ψ preferences is largely a function of longer-range (electrostatic and Lennard-Jones type) interactions with neighboring residues and beyond.

Overall, the ϕ/Ψ preferences agree well between the protein simulations and PDB distributions. However, a close inspection suggests that the agreement could be improved further by slight force field adjustments. In particular, it appears that the sampling of positive ϕ values, of ϕ values below -150, and of the α' conformation is too favorable relative to other parts of the energy landscape, while $C7_{eq}$ sampling is underrepresented. There are also differences in the conformational preferences of glycine and proline residues that could be addressed by force field modifications. It is straightforward to adjust the CMAP torsion potentials accordingly, and future studies will examine how simulations with such a modified torsion potential affect the overall sampling of protein structures.

A constant concern with simulation studies is the achievement of converged sampling of all statistically relevant conformational regions. It appears that the dipeptide simulations over 150 ns are sufficient (or at least close to it) since many transitions are observed between the major basins in all of the simulations. However, it is possible that protein simulations of up to 150 ns do not completely sample the conformational space accessible during biological and experimental time scales of milliseconds to minutes. One consequence of the limited test sets is that the simulation results provide little information about the relative sampling of α versus β conformations since the observed α versus β propensities are largely a function of the native secondary structures of the chosen test proteins. Further studies of small helix- and hairpin-forming peptides with the same methodology will be necessary to examine the relative sampling of α versus β conformations in the context of proteins in more detail. However, the sampling of relative conformations within a given basin is expected to be more meaningful since the corresponding structural variations could largely be accommodated without major disruption of a given protein structure.

The dipeptide simulations can be compared to spectroscopic data that indicate that PPII and C5 are the dominant conformations in solution, while α_R is populated only to a small extent in alanine and valine dipeptides.[14,18] The dipeptide simulations presented here show a more equal sampling of α and β basins, suggesting a slight bias toward α-helical conformations. Such a bias has also been suspected in other recent studies with the CHARMM force field in conjunction with the CMAP potential and will require further exploration.[40] A force field that better reproduces experimental data for dipeptides and other small peptides may also alter the torsional preferences in the protein simulations reported here. The hope is that such modifications would improve the agreement with the conformational preferences from the PDB and lead to conformational sampling in even better agreement with crystallographic structures for individual proteins. It is possible, though, that the fixed charged force field used here places limitations on how well experimental data for small peptides and larger proteins can be reproduced simultaneously. In this context, it should be stressed that the simulations reported here only consider the combination of the CHARMM force field with the CMAP torsion potential, and specific results may vary for other force fields. It is likely, however, that the general conclusions are equally valid for other force fields where similar assumptions of a common backbone torsion potential based on alanine dipeptide are made.

Finally, it should be noted that there is a fundamental difference in the way the potentials of mean force are obtained from the simulations and from the experimental structures. The results from the simulations are obtained from a small number of structures over a large number of instantaneous conformations. On the other hand, the results from the experiment are obtained from a large number of structures, each representing an ensemble and time average. The potentials of mean force agree quantitatively very well in the low-energy regions, thereby confirming the validity of the ergodic hypothesis that time averages are equivalent to ensemble averages. In contrast, higher-energy regions are not sampled extensively in the PDB structures, while the simulations show broad conformational sampling well beyond the major conformational basins. This difference is primarily a result of the relatively small sample size used in the analysis of the PDB structures. For example, approximately 57 000 alanine conformations were analyzed from PDB structures (see Table 2) compared to approximately 17 million conformations from the simulations (66 alanine residues over an average simulation length of 130 ns with conformations saved every 0.5 ps). However, it is also possible that experimental structures, except for structures at the very highest resolution, reflect to some extent assumptions about ideal molecular bonding geometries if imposed during molecular refinement. Such constraints would limit the sampling of noncanonical regions of the Ramachandran plot in the experimental structures. It should be mentioned that there are also some theoretical concerns that have been raised about extracting potentials of mean force from PDB structures;[41] however, these arguments may not apply to the present study since we are analyzing simulations and crystallographic structures in an equivalent manner.

We now come back to the central question of this paper: Is alanine dipeptide a good model for representing the torsional preferences of protein backbones? The answer is "yes" and "no". It appears that force field parametrization based on alanine dipeptide along with suitable long-range interactions can accurately reflect amino acid dependent variations in backbone torsional preferences in the context of protein structures. This suggests that a modular approach in the development of the force field is justified, and specific

modifications to bonding terms as a function of amino acid, with the exception of glycine and proline, are probably not necessary. However, the ϕ/Ψ preferences differ significantly between dipeptide and protein environments. As a consequence, dipeptides do not appear to be a suitable model for understanding the backbone torsional preferences of amino acids in proteins.

Acknowledgment. Financial support from NSF CAREER grant 0447799 and the Alfred P. Sloan Foundation and access to computational resources through the High Performance Computing Center at Michigan State University are acknowledged.

Supporting Information Available: Potentials of mean force for sampling of ϕ/ψ backbone torsion angles for remaining amino acid residues. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of Polypeptide Chain Configurations. *J. Mol. Biol.* **1963**, *7*, 95–99.

(2) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.

(3) Lovell, S. C.; Davis, I. W.; Arendall, W. B., III; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. Structure Validation by Cα Geometry: ϕ, Ψ and Cβ Deviation. *Proteins* **2003**, *50*, 437–450.

(4) Karplus, P. A. Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci.* **1996**, *5*, 1406–1420.

(5) Kleywegt, G. J.; Jones, T. A. Phi/Psi-chology: Ramachandran revisited. *Structure* **1996**, *4*, 1395–1400.

(6) Novmöller, S.; Zhou, T.; Ohlson, T. Conformations of amino acids in proteins. *Acta Crystallogr., Sect. D* **2002**, *58*, 768–776.

(7) Chakrabarti, P.; Pal, D. The interrelationships of side-chain and main-chain conformations in proteins. *Prog. Biophys. Mol. Biol.* **2001**, *76*, 1–102.

(8) Ho, B. K.; Brasseur, R. The Ramachandran plots of glycine and pre-proline. *BMC Struct. Biol.* **2005**, *5*, 14–24.

(9) Anderson, R. J.; Weng, Z.; Campbell, R. K.; Jiang, X. Main-Chain Conformational Tendencies of Amino Acids. *Proteins* **2005**, *60*, 679–689.

(10) Smith, P. E. The alanine dipeptide free energy surface in solution. *J. Chem. Phys.* **1999**, *111*, 5568–5579.

(11) Drozdov, A. N.; Grossfield, A.; Pappu, R. V. Role of Solvent in Determining Conformational Preferences of Alanine Dipeptide in Water. *J. Am. Chem. Soc.* **2004**, *126*, 2574–2581.

(12) Wang, Z. X.; Duan, Y. Solvation effects on alanine dipeptide: A MP2/cc-pVTZ//MP2/6-31G** study of (Phi,Psi) energy maps and conformers in the gas phase, ether, and water. *J. Comput. Chem.* **2004**, *25*, 1699–1716.

(13) Feig, M. Kinetics from Implicit Solvent Simulations of Biomolecules as a Function of Viscosity. *J. Chem. Theory Comput.* **2007**, *3*, 1734–1748.

(14) Kwac, K.; Lee, K. K.; Han, J. B.; Oh, K. I.; Cho, M. Classical and quantum mechanical/molecular mechanical molecular dynamics simulations of alanine dipeptide in water: Comparisons with IR and vibrational circular dichroism spectra. *J. Chem. Phys.* **2008**, *128*, 105106.

(15) Chekmarev, D. S.; Ishida, T.; Levy, R. M. Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete-state kinetic models. *J. Phys. Chem. B* **2004**, *108* (50), 19487–19495.

(16) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and beta-hairpin peptide. *J. Phys. Chem. B* **2004**, *108* (21), 6582–6594.

(17) Kim, Y. S.; Wang, J. P.; Hochstrasser, R. M. Two-dimensional infrared spectroscopy of the alanine dipeptide in aqueous solution. *J. Phys. Chem. B* **2005**, *109* (15), 7511–7521.

(18) Grdadolnik, J.; Grdadolnik, S. G.; Avbelj, F. Determination of conformational preferences of dipeptides using vibrational spectroscopy. *J. Phys. Chem. B* **2008**, *112*, 2712–2718.

(19) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. Accurate ab Initio Quantum Chemical Determination of the Relative Energetics of Peptide Conformations and Assessment of Empirical Force Fields. *J. Am. Chem. Soc.* **1997**, *119*, 5908–5920.

(20) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, J. D.; Evanseck, M. J.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(21) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

(22) Mackerell, A. D. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.* **2004**, *25* (13), 1584–1604.

(23) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.

(24) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(25) Feig, M.; MacKerell, A. D.; Brooks, C. L. Force field influence on the observation of pi-helical protein structures in molecular dynamics simulations. *J. Phys. Chem. B* **2003**, *107* (12), 2831–2836.

(26) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyal, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(27) Macias, A. T.; MacKerell, A. D. CH/pi interactions involving aromatic amino acids: Refinement of the CHARMM tryptophan force field. *J. Comput. Chem.* **2005**, *26* (14), 1452–1463.

(28) Jorgensen, W. L. Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.

(29) Roux, B. Valence selectivity of the gramicidin channel: A molecular dynamics free energy perturbation study. *Biophys. J.* **1996**, *71*, 3177–3185.

(30) Darden, T. A.; York, D.; Pedersen, L. G. Particle Mesh Ewald: An Nlog(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(31) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341.

(32) Nose, S. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984**, *52*, 255–268.

(33) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

(34) Feig, M.; Karanicolas, J.; Brooks, C. L., III. MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology. *J. Mol. Graphics Modell.* **2004**, *22*, 377–395.

(35) Wang, G.; Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589–1591.

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

(37) Guzzo, A. The influence of amino-acid sequence on protein structure. *Biophys. J.* **1965**, *5*, 809–822.

(38) Lewis, P. N.; Go, N.; Go, M.; Kotelchuck, D.; Scheraga, H. A. Helix Probability Profiles of Denatured Proteins and Their Correlation with Native Structures. *Proc. Natl. Acad. Sci. U.S.A.* **1970**, *65*, 810–815.

(39) Stellwagen, E.; Shalongo, W. Evidence for Glutamate Self-Capping Within a Peptide Helix. *Biopolymers (Peptide Sci.)* **1997**, *43*, 413–418.

(40) Tanizaki, S.; Clifford, J. W.; Connelly, B. D.; Feig, M. Conformational Sampling of Peptides in Cellular Environments. *Biophys. J.* **2008**, *94*, 747–759.

(41) Ben-Naim, A. Statistical potentials extracted from protein structures: Are these meaningful potentials. *J. Chem. Phys.* **1997**, *107* (9), 3698–3706.

CT800153N