doi:10.1016/S0022-2836(03)00765-4 *J. Mol. Biol.* (2003) **331**, 693–711

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Investigations into Sequence and Conformational Dependence of Backbone Entropy, Inter-basin Dynamics and the Flory Isolated-pair Hypothesis for Peptides

Muhammad H. Zaman[1,2], Min-Yi Shen[1,3], R. Stephen Berry[1,3] Karl F. Freed[1,3]* and Tobin R. Sosnick[2,4]*

[1]*Department of Chemistry, The University of Chicago, Chicago IL 60637, USA*

[2]*Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL 60637 USA*

[3]*The James Franck Institute The University of Chicago Chicago, IL 60637, USA*

[4]*Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago IL 60637, USA*

**Corresponding authors*

The populations and transitions between Ramachandran basins are studied for combinations of the standard 20 amino acids in monomers, dimers and trimers using an implicit solvent Langevin dynamics algorithm and employing seven commonly used force-fields. Both the basin populations and inter-conversion rates are influenced by the nearest neighbor's conformation and identity, contrary to the Flory isolated-pair hypothesis. This conclusion is robust to the choice of force-field, even though the use of different force-fields produces large variations in the populations and inter-conversion rates between the dominant helical, extended β, and polyproline II basins. The computed variation of conformational and dynamical properties with different force-fields exceeds the difference between explicit and implicit solvent calculations using the same force-field. For all force-fields, the inter-basin transitions exhibit a directional dependence, with most transitions going through extended β conformation, even when it is the least populated basin. The implications of these results are discussed in the context of estimates for the backbone entropy of single residues, and for the ability of all-atom simulations to reproduce experimental protein folding data.

Keywords: protein folding; Langevin dynamics; backbone conformational entropy; simulation; Ramachandran plot

Introduction

A fundamental descriptor of a polypeptide's conformation is the set of its backbone dihedral or torsional angles. For each residue, these angles specify a location in the Ramachandran plot of Φ, Ψ angles.[1,2] The intrinsic preference for each peptide unit to be in one Ramachandran basin or another and the inter-basin hopping rates directly affect secondary structure preferences and residual structure in the denatured state, as well as the overall thermodynamics and kinetics of protein folding. In spite of this significance, only a few studies have focused on the peptide backbone dynamics using atomic-level force-fields (FFs) in an aqueous environment.[3–5] Furthermore, an analysis of these backbone dynamics and structure is useful to reveal any dependence on context, including the conformation and chemical identity of the nearest-neighbor (NN) residues. Here, we present such a study for amino-acetylated (Ace) and carboxy-amidated (Nme) versions of a mono-alanine "dipeptide" (i.e. Ace-Ala-Nme) and for dimers and trimers (Figure 1(a)) with one, two and three pairs of Φ, Ψ dihedral angles, respectively.

Our analysis tests the applicability of the Flory isolated-pair hypothesis (IPH),[6] which is invoked implicitly in many equilibrium and kinetic treatments of protein folding, including helix-coil theories. According to the IPH, the Ramachandran

Abbreviations used: Ace, acetylated; Nme, amidated; B1, B2 and B3, basin 1, 2 and 3, respectively; $C_i(t)$, autocorrelation function for the ith basin; FF, force-field; G-S-94, Garcia's modified version of Amber 94; IPH, isolated-pair hypothesis; LD, Langevin dynamics; MD, molecular dynamics; NN, nearest-neighbor residue; $P_i(t)$, probability of being in the ith basin; PP-II, polyproline II.

E-mail addresses of the corresponding authors: k-freed@uchicago.edu; trsosnic@midway.uchicago.edu

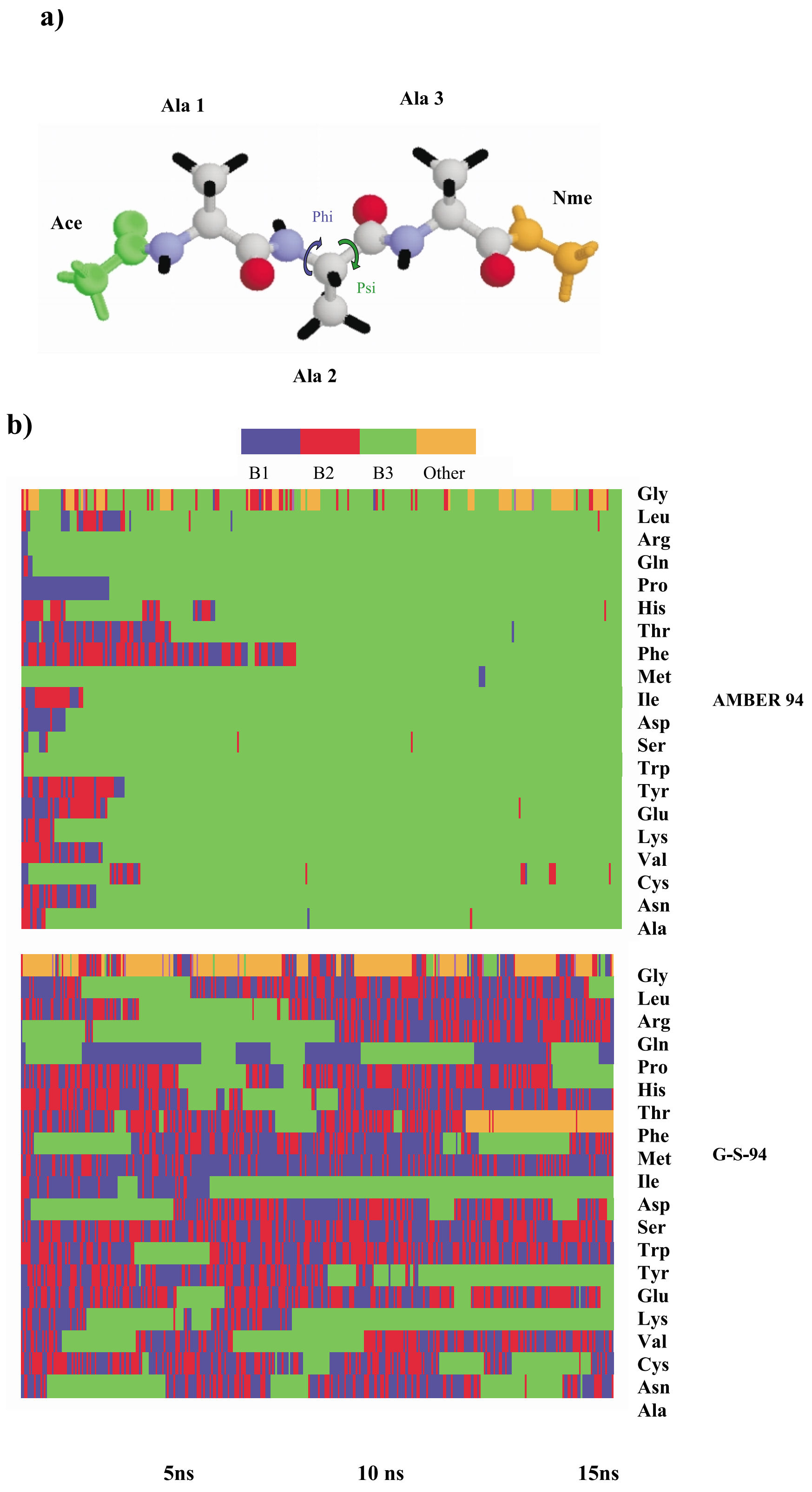

Figure 1 (*legend opposite*)

basin populations of one residue are independent of its neighbors' conformations (except for proline, and residues preceding proline residues): "the interactions associated with rotations of one such independent pair are quite independent of the angles assumed by neighboring pairs."[6] When this pivotal isolated-pair assumption is valid, the backbone entropy of the system can be expressed as

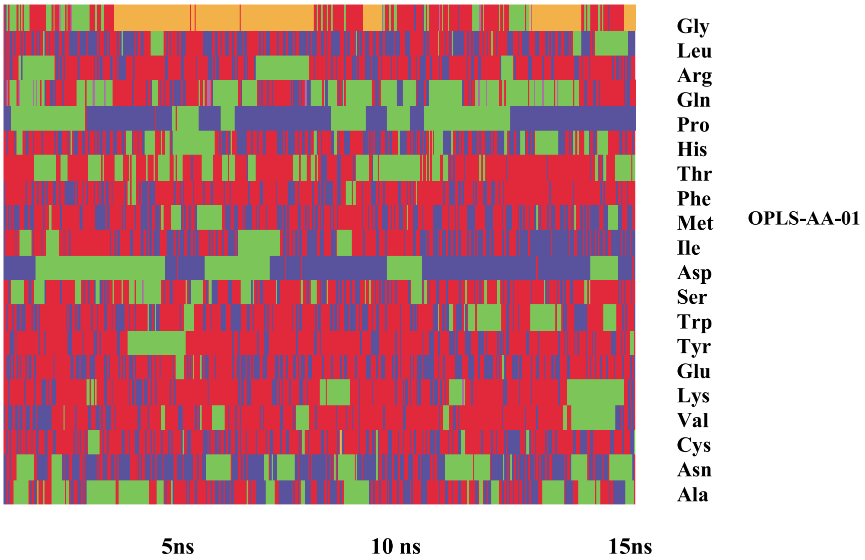

Figure 1. Dynamics of a tri-peptide. (a) Ace-(Ala)$_3$-Nme peptide with center of the three pairs of backbone dihedral angles highlighted. The hydrogen atoms are shown in stick representation (black), whereas oxygen (red), nitrogen (blue) and carbon (grey) atoms are depicted in ball-and-stick representation. (b) Backbone dynamics of different center residues in Ala-X-Ala. The 15 ns time-course is presented for the basin populations, colored according to the legend given at the top of the Figure. Simulations for three representative FFs are provided to demonstrate the wide variation in populations between the FFs. Residues spend considerably more time in basin 3 (helical) when the Amber 94 FF is used compared to the G-S-94 and OPLS-AA-01 FFs, where there is higher probability for extended β structures (basin 2) and PP-II (basin 1).

the sum of individual residues' entropies. Within the IPH, a single helix-coil equilibrium constant can be assigned to each amino acid species without qualification with respect to either its neighbor's configuration or identity, as is done in nearly all analyses of helix-coil transitions.

Pappu *et al.*[7] consider the reduction in sampling due to NN's configuration in polyalanine. In contradiction to the IPH, they find that the central residue, located between two residues with helical geometries, is sterically hindered by these neighbors. However, when the dihedral angles in a polypeptide are chosen according to their relative basin probabilities without restriction to the helical basin, the number of overlapping conformations is minor, for example, only 16% for a 12 residue chain.[8] Hence, steric hard-core type overlap provides only a minor reduction in the total conformational entropy of the unfolded state (in the absence of extensive helical configurations).

Molecular dynamics (MD) simulations have demonstrated recently that different FFs can produce rather large differences in basin populations.[4,9,10] Garcia and co-workers find that the Amber 96 FF must be altered so that a largely alanine-containing peptide is predicted to undergo helix-coil transitions at the experimentally observed temperatures.[11] Their alteration involves the elimination of an additional, backbone dihedral, or torsional potential, which is present with varying topographies in most commonly used FFs.[4] Upon elimination of this added potential, the

Table 1. Basin populations and configurational entropy for different force fields

Force field	PP-II (%)	Extended β (%)	α-Helical (%)	$T\Delta S^{a}$ (kcal mol^{-1} K^{-1})
Amber 94	1.08 (13)	1.5 (3)	96.86 (80)	0
Amber 96	14.15 (41)	76.27 (44)	5.02 (14)	−0.187
Garcia-A94	30.24	17.8	45.31	−0.358
Charmm27	24.20 (55)[b]	18.33	47.62 (45)	−0.365
OPLS-AA-97[c]	82.97[b] (88)[b]		12.57 (12)	−0.355
OPLS-AA-01	31.02 (65)[61.6][d]	41.17 (12)[29.4][d]	20.75 (17)[5.1][d]	−0.372
OPLS-UA[c]	59.31[b]		33.93	−0.427

For Ala2 in an N and C-terminal capped Ala1-Ala2-Ala3 at $T = 300$ K. Values in parentheses are from an explicit solvent MD calculation for tri-alanine with a positively charged (+1) N terminus and neutral C terminus.[10]

[a] Calculated using equation (1) and referenced to value for Amber 94.

[b] Combined values for PP-II and extended β.

[c] PP-II and extended β basins are not distinguished in this FF.

[d] Values in square brackets are obtained with our implicit solvent LD calculation for a tri-alanine with a positively charged (+1) N terminus and neutral C terminus, provided for comparison to the explicit solvent MD calculations[10] performed on a similar molecule.

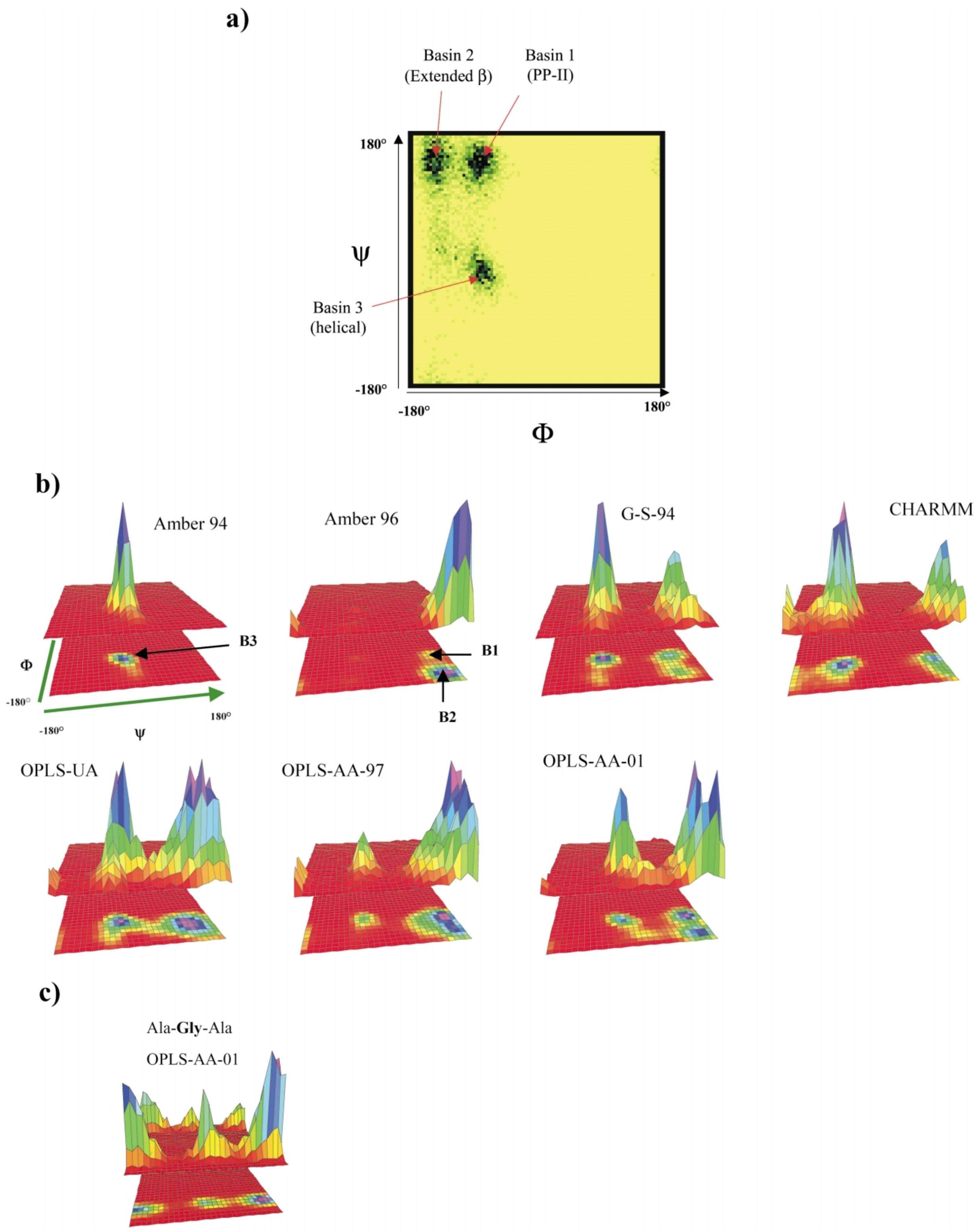

Figure 2. Ramachandran basin populations for difference FFs. (a) Ramachandran plot of Ala2 in Ala1-Ala2-Ala3 computed using the OPLS-AA-01 FF shows the presence of three distinct basins. (b) Basin populations for Ala2 for the seven different FFs, calculated from averages along the time trajectories such as those illustrated in Figure 1(b). The most populated basins are PP-II (basin 1), extended β (basin 2) and α-helical (basin 3). (c) Basin population for Gly in Ala-Gly-Ala using the OPLS-AA-01 FF. An unconventional view for the Ramachandran basins is used to enable visualizing both the population (top) plot and the contour (bottom) plot.

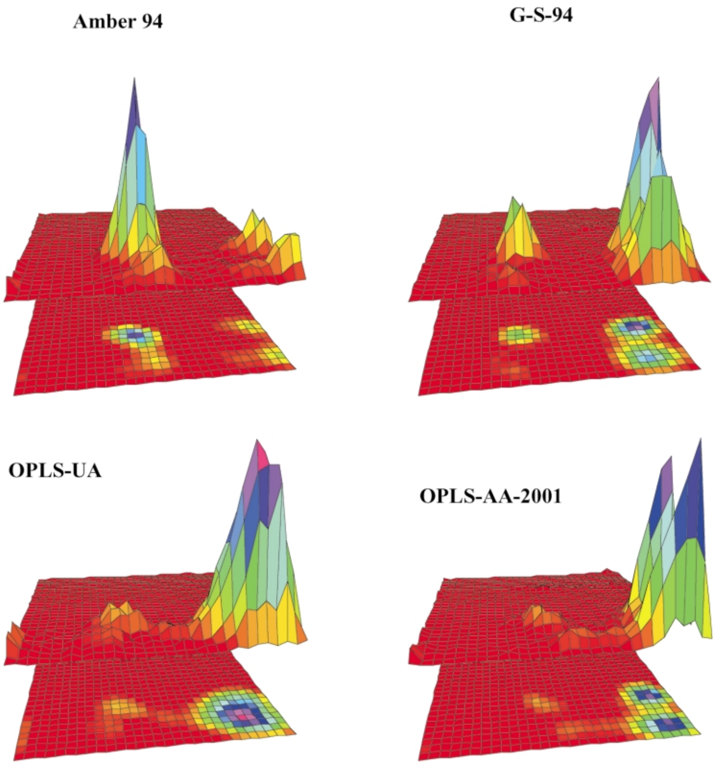

Figure 3. Ramachandran basin populations for Ace-Ala-Nme for different FFs. Populations are obtained from 45 ns LD trajectories.

basin preference in Garcia's FF is determined entirely by backbone, side-chain, electrostatic, and solvent interactions and geometries. Similarly, explicit solvent MD simulations by Hu *et al.* and Mu *et al.* show that the preference for the helical basin ranges from ~10–90% and that individual inter-basin hopping rates can vary up to tenfold when computed using different FFs for the simple examples of di-alanine and di-glycine[4] and tri-alanine.[10]

Because of the considerable influence an FF can exert on predicted inter-basin hopping frequencies, we test the reliability of our conclusions by performing independent calculations employing seven commonly used FFs, namely Amber 94†, Amber 96,[12] Garcia's modified Amber 94[11] (referred to here as G-S-94), Charmm-27,[13] OPLS-united atom,[14] OPLS-AA-97, and the latest OPLS-AA-01 [15] The comparison of predictions obtained from the different FFs is also motivated by the knowledge that they have been optimized to reproduce thermodynamic data (and, in some cases, *ab initio* quantum calculations[15]) and are generally validated by their ability to describe protein structures. Consequently, their suitability for dynamical calculations is unclear, because dynamics is sensitive to the heights of kinetic barriers, whereas thermodynamics and native structures are not.

Our Langevin dynamics (LD) simulations with an implicit solvent model[16] produce nearly the same, strong FF-dependence of basin populations and dynamics obtained from MD calculations with explicit solvent.[10] Moreover, where the same FFs are used for explicit and implicit solvent simulations, good agreement is found, thereby supporting the validity of our computationally far less expensive approach.

In the present extensive study at 300 K, we examine the validity of the IPH using molecular mechanics potentials to construct and analyze the conformational and dynamical properties of peptides composed of many different amino acid combinations (60 different species in all). For all seven FFs considered, the time-course of the LD trajectories reveals that a residue's basin population and dynamics may be influenced strongly by the NN's conformation and chemical identity. We calculate the backbone conformational entropy in the unfolded state for each residue according to its sampling of the Ramachandran plot. This calculation is conducted separately by assuming that the samplings for each residue are independent (the IPH assumption) and by considering the correlated motions in order to quantify the error in IPH. We discuss the implications of the different thermodynamics and dynamics produced by the various widely used FFs upon the ability of all-atom simulations to describe the free energies, folding pathways and time-scales in protein folding.

Results

The Ramachandran basin assignments are derived from the observed time-course of the population distributions (Figure 2(a); see Methods). A common definition is suitable for all

† http://amber.scripps.edu/

Table 2. Alanine conformational preferences as a function of its NN chemical identity

	Ala-Ala-X			X-Ala-Ala		
X	Basin 1	Basin 2	Basin 3	Basin 1	Basin 2	Basin 3
Ala	1.1 ± 1.5/30.02 ± 5.2/31.02 ± 2.1	1.5 ± 2.9/17.9 ± 2.1/41.02 ± 2.0	96.8 ± 2.9/45.75 ± 5.1/20.2 ± 2.6	1.1/30.02/31.02	1.5/17.9/41.02	96.8/45.75/20.2
Trp	6.6/27.25/19.4	11.6/17.83/31.2	80.26/48.41/43.6	2.8/48.0/21.66	9.06/41.16/50.41	87.66/4.50/24.33
Met	1.0/34.08/24.5	1.42/20.22/23.25	97.13/39.91/45.0	2.0/19.33/23.16	4.2/13.58/27.13	93.26/63.68/41.46
Asp	0.4/23.25/12.2	0.4/45.08/40.83	99.0/25.75/44.12	0.2/26.25/28.6	1.04/18.25/23.16	98.53/48.41/41.7
Asn	3.87/39.8/10.0	7.9 /22/14.5	87.0/32.58/17.81	3.16/31.41/36.16	5.00/22.5/35.17	91.20/41.16/20.53
Leu	0.6/51.51/23.65	0.1/25.7/21.81	99.4/14.33/46.33	3.34/22.3/27.83	4.51/28.3/34.25	92.34/39.1/28.50
Gly	6.8/44.75/35.12	12.33/28.25/41.91	79.13/16.41/15.41	4.80/32.0/38.68	4.00/21.75/36.58	90.2/40.48/18.25

Influence on the center alanine's basin populations. Values in percent are given for the Amber 94/G-A96/OPLS-AA-01 FFs, respectively. Errors indicate the difference between a 15 ns and a 45 ns trajectory (omitting an initial 3 ns equilibration period).

seven FFs. The most populated basins are the polyproline II (basin 1, B1), extended β (basin 2, B2), and α-helical (basin 3, B3) conformations (see Table 1). The polyproline II (PP-II) and extended β basins are separated by a free-energy barrier for all the FFs, except the OPLS-UA and OPLS-AA-97 FFs, where only a single basin is present in this region of the Ramachandran plot. The existence of a distinct PP-II basin is established both experimentally in native structures and unfolded fragments[17–19] and in computer simulations.[10,20]

Figure 1(b) presents the time-course of basin occupancies for the central Ala in the tri-Ala peptide as calculated with three different FFs using the color code at the top of the Figure. The color variations between the trajectories from the different FFs strikingly expose the qualitatively different dynamics predicted by the various FFs. The Amber 94 FF populates the helical basin predominantly, whereas the distribution among the three dominant wells is more uniform for G-S-94 and OPLS-AA-01, though G-S-94 yields significantly more helical population than OPLS-AA-01.

Sequence-dependence of NN effects

Underlying the IPH is the assumption of a lack of correlations between the Φ, Ψ dihedral angles of neighboring residues due to the rigidity of the peptide bond. Our first investigation focuses on the importance of the flanking moieties. A series of simulations is performed contrasting the behavior of an alanine dipeptide, i.e. a single alanine molecule capped with acetyl and amide groups (Figure 3), with that of an alanine residue flanked on both sides with alanine residues (Figure 2(b)). The presence of less bulky neighbors in the single alanine molecule increases the fraction of time the alanine spends in the extended β and PPII conformations (basins 1 and 2). For example, using the

Table 3. Influence of NN sequence on alanine's basin population fractions

	AMBER 94 (kcal mol^{-1})			G-S-94 (kcal mol^{-1})			OPLSAA-2001 (kcal mol^{-1})		
X in Ala-X	B1	B2	B3	B1	B2	B3	B1	B2	B3
Ala-Ala	0	0	0	0	0	0	0	0	0
Ala-Trp	0.32187	0.38131	−0.0814	0.2567	0.57102	−0.957	0.28123	0.16389	−0.4616
Ala-Met	0.14866	0.38131	−0.0208	−0.0763	0.36265	−0.4142	0.14119	0.34036	−0.4806
Ala-Asp	−0.3254	−0.8086	0.49557	0.36313	−0.3609	1.11653	0.55953	0.00249	−0.4687
Ala-Asn	−0.1266	0.12945	−0.0073	−0.0157	0.32163	−0.5273	0.67884	0.62365	0.07555
Ala-Leu	0.05046	−0.02	0.00648	−0.1666	0.35422	−0.2238	0.16237	0.37872	−0.4981
Ala-Gly	−0.2749	0.1358	0.00382	−0.0788	0.12601	−0.1303	−0.0749	−0.0132	0.1624
Ala-**Ala**	0	0	0	0	0	0	0	0	0
Trp-**Ala**	−0.1073	−0.2467	0.05502	0.15756	−0.0026	−0.4162	0.21511	−0.124	−0.1116
Met-**Ala**	0.10302	0.25847	−0.0368	−0.053	0.11203	−0.0948	0.17494	0.24776	−0.4314
Asp-**Ala**	0.57189	0.16253	−0.0581	0.13285	0.08293	−0.4892	0.04835	0.34269	−0.4349
Asn-**Ala**	0.05046	−0.0249	5.432×10^{-4}	0.08533	0.04877	−0.3418	−0.0924	0.09203	−0.0097
Leu-**Ala**	0.32554	0.24839	−0.067	0.01402	0.00481	−0.1429	0.06472	0.10793	−0.2065
Gly-**Ala**	−0.8225	−0.233	0.11512	0.0705	0.20779	−0.6366	−0.1328	0.06844	0.06091

For the alanine in bold face, calculated according to *RT* (ln (fractional population in basin Y for alanine when NN is residue X) − ln (fractional population in basin *y* for alanine when NN is alanine)). Only the fractional populations in basins 1, 2 and 3 are presented. The NN conformation is unconstrained. The negative values indicate that the NN effect due to residue X on Ala is less than the NN effect of Ala on Ala.

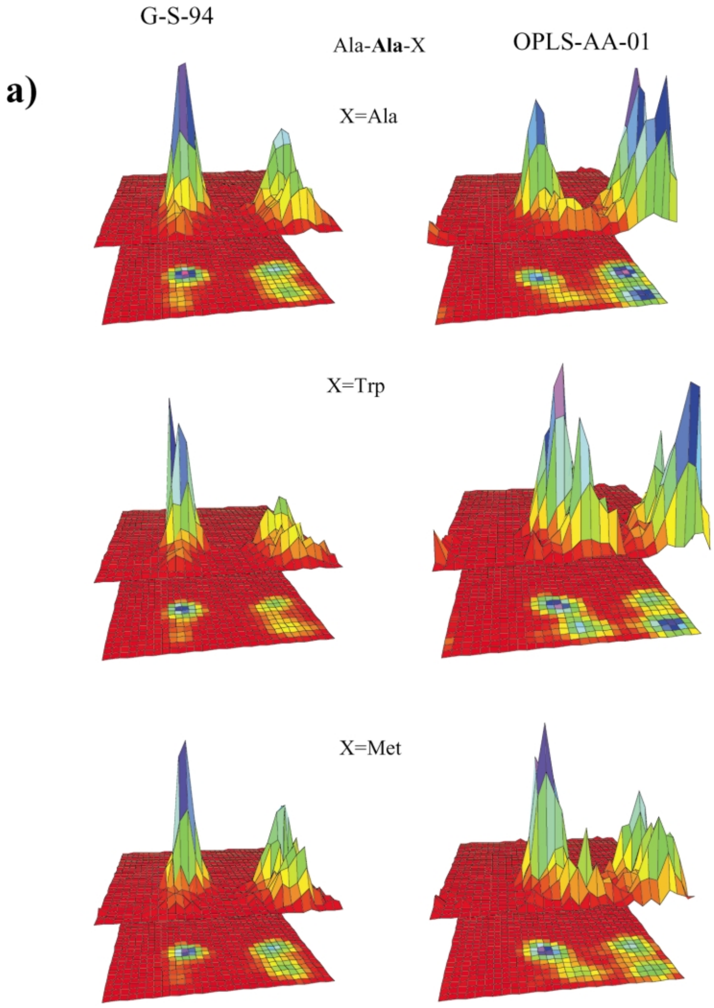

Figure 4 (*caption on page 702*)

Table 4. Sequence-dependence of backbone entropy in Ala-X-Ala with unconstrained neighbors

	$T(S_X - S_{Ala})$ (kcal mol^{-1})			
X	AMBER 94	G-S-94	OPLS-AA-01	OPLS-UA
Ala	0	0	0	0
Asn	0.0045	−0.053	−0.008	−0.0515
Cys	0.06	0.0245	−0.017	−0.049
Val	−0.076	0.067	−0.276	−0.1495
Lys	−0.0815	0.033	−0.111	−0.0455
Glu	−0.041	4.00×10^{-4}	−0.139	−0.1135
Tyr	−0.06765	−0.125	−0.065	−0.02
Trp	0.1015	−0.0945	−0.32	−0.0095
Ser	−0.024	0.025	−0.036	−0.0235
Asp	−0.251	−0.1485	−0.852	−0.12495
Ile	−0.0425	−0.181	−0.674	−0.1565
Met	0.0255	−0.09995	−0.033	−0.0515
Phe	0.3465	−0.2061	0.22	−0.0645
Thr	−0.024	−0.022	−0.164	−0.1005
His	0.1795	0.0105	0.021	−0.0955
Pro	−0.2675	−0.398	−0.473	−0.2505
Gln	−0.02	−0.0385	−0.01	−0.0383
Arg	−0.0415	0.006	0.011	−0.0635
Leu	0.09	−0.032	−0.11	−0.04995
Gly	0.498	0.045	0.22	0.0455

Amber 94 FF, essentially the entire population is in the helical basin 3 for the (capped) tri-alanine molecule, whereas ~20% populates the other two basins in the (capped) mono-alanine molecule. The difference between mono and tri-alanine demonstrates that the rigidity of the peptide backbone does not prevent the neighbor moieties from influencing the backbone configuration of even a small amino acid such as alanine.

These results are similar to those reported by Hu *et al.*, who observe that mono-alanine populates the helical basin 84% of the time with Amber 94.[4,10,21] Our implicit solvent LD simulations for the tri-alanine basin populations agree reasonably with the explicit solvent MD simulations for a similar system reported by Mu *et al.* (Table 1).[5] Mu *et al.* perform their simulations for a tri-alanine with a positively charged (+1) N terminus and a neutral C terminus (Y. Mu, personal communication). Our LD simulations with neutral tri-alanine and the Amber 94 FF yield an α-helical basin population that differs by ~15% from the explicit solvent

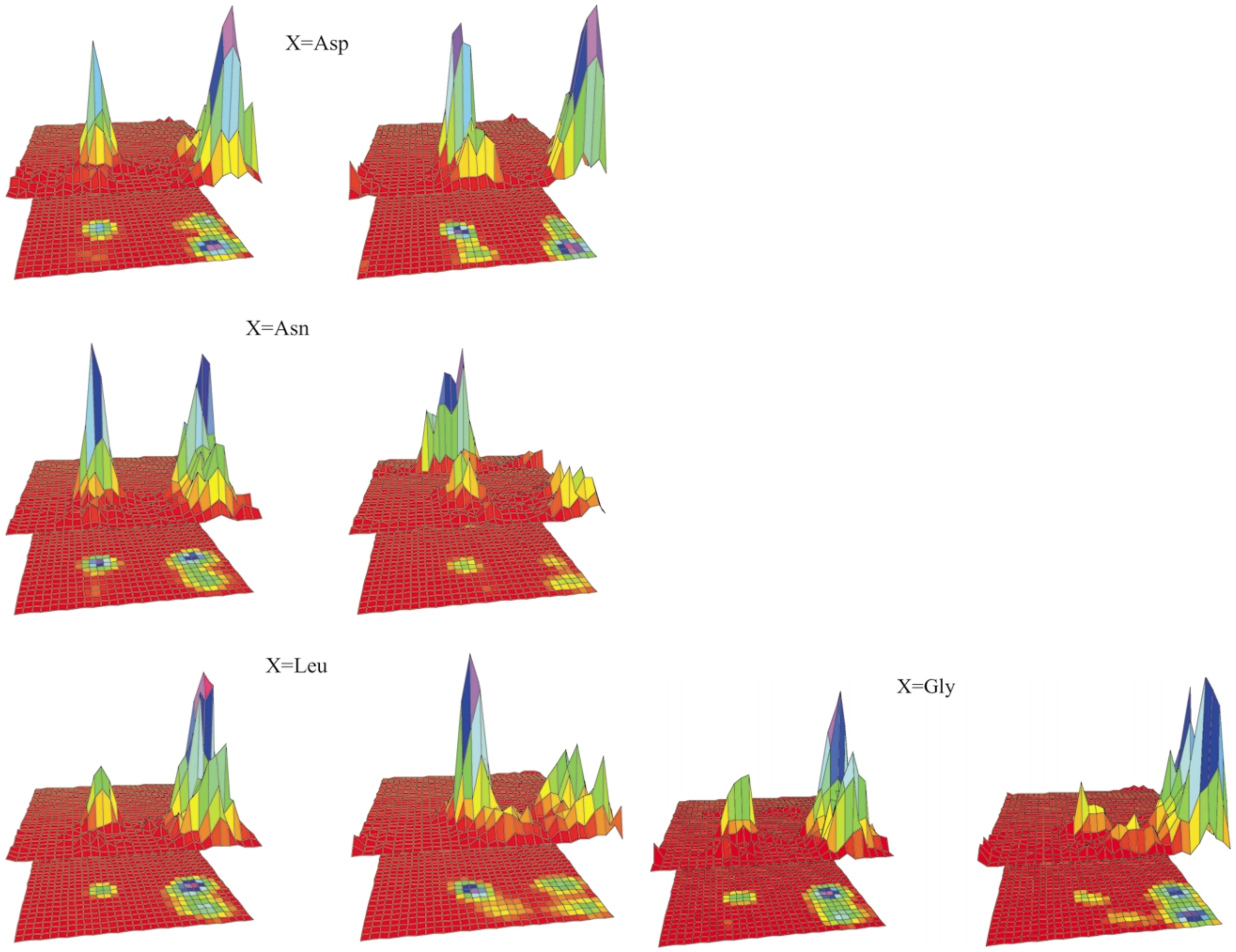

Figure 4 (*caption on page 702*)

calculations. Compared to the explicit solvent calculations with the Amber-96 and OPLS-AA-2001 FFs, our simulations have 27–34% less PP-II population and 29–32% more extended β population, and are within 9% for the helical population. Our simulations for the CHARMM-27 FF produce 12% less combined β and PP-II populations. Very close agreement with the explicit solvent simulations is found for the populations from the OPLS-AA-97 FF (within 6%). This general agreement for similar systems provides strong justification for use of the implicit solvent model in our more extensive study of NN effects that follows.

Moreover, the primary difference between the implicit solvent and the explicit solvent simulations lies in the latter having more PP-II basin population. This enhancement arises, in large part, from the implicit solvent simulations use of a tri-alanine molecule having a charged (+1) N terminus. When a similarly charged molecule is studied with implicit solvent (capped with $-NH_2$ rather than –H), the agreement with the explicit solvent treatment is very good; 65% *versus* 61% for the PPII basin, 12 % *versus* 29% for the extended β basin, 17% *versus* 5% for the helical basin, respectively. Hence, much of the difference in the two types of simulations is eliminated when the identically charged molecule is investigated.

The three FFs (Amber 94, OPLS-AA-01 and G-S-94) generate different basin preferences for the tri-alanine molecule (Figure 1(b) and first row in Table 2). The Amber 94 FF predicts a predominant helix basin population, while the G-S-94 and OPLS-AA-01 yield helix, extended, and PPII basin populations in the ratios roughly of 3:2:1 and 2:4:3, respectively. Table 2 illustrates the NN effect on the central Ala residue in the peptides Ala-Ala-X and X-Ala-Ala for seven different residues X (of varying character), while Table 3 and Figure 4 display the NN influences for X in the seven pairs of di-peptides Ala-X and X-Ala. The G-S-94 and OPLS-AA-01 FFs produce an appreciable NN effect, with the alanine basin populations sometimes changing by a factor of 3 as the neighboring side-chains are varied. For example, the helix basin population for the center alanine residue using the G-S-94 FF ranges from the low of 16.4% when the C-terminal NN is Gly to a high of 48.4% for Trp. The populations are 40.5% and 4.5% when the N-terminal NN is Gly and Trp, respectively.

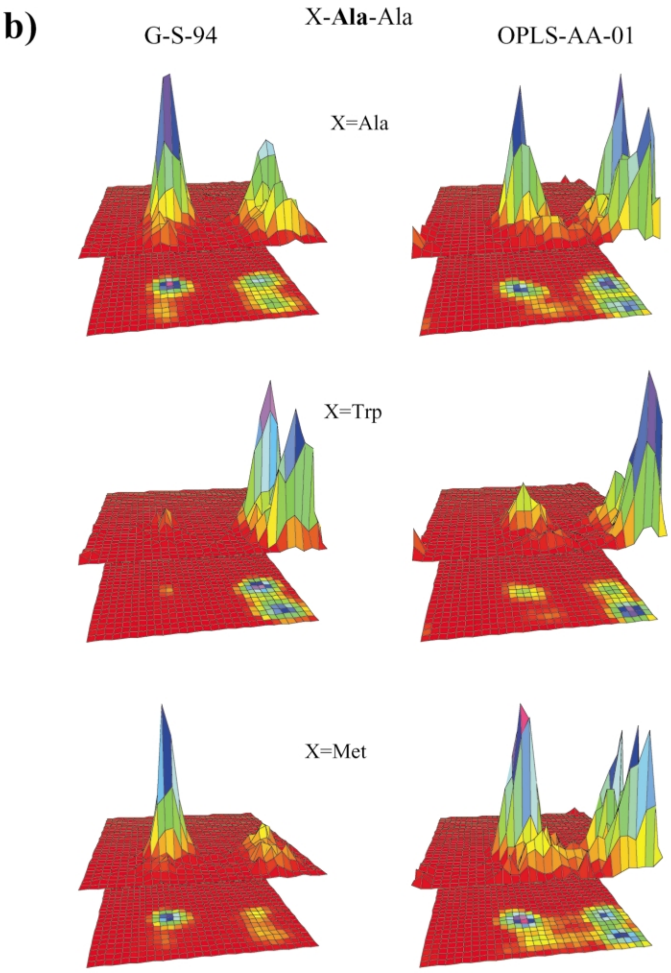

Figure 4 (*caption on page 702*)

Similarly, large NN influences are evident for the G-S-94 FF in Table 2.

The Amber 94 FF yields only a marginal NN effect and only in the di-peptides (data not shown). This difference arises because the Amber 94 FF predicts that the alanine backbone almost always remains in the helical basin, regardless of the NN, whereas the helical basin population varies between 5% and 75% for the other two FFs. Hence, much of our analysis focuses on the two more realistic FFs, G-S-94 and OPLS-AA-01.

The NN effects computed for the dimers are of a magnitude similar to those obtained for the trimers (data not shown), which confirms that the observations concerning NN effects are not artifacts of longer-range $i-1$, $i+1$ side-chain interactions.

Backbone entropy

The influence of NN residues can be quantified in terms of the change in an alanine's backbone entropy due to the presence of different neighbors. Using the basin populations on the Ramachandran map, we calculate the backbone conformational entropy according to the relation (see Methods):

$$S = -R\sum_{i=1}^{120}\sum_{j=1}^{120} P_{ij}\ln P_{ij} \qquad (1)$$

where P_{ij} is the normalized probability of being in the i,jth 3° × 3° mesh element in the Ramachandran map, and R is the gas constant. Although this calculation of S depends on the mesh size (i.e. the volume per configuration in phase space), entropy differences between residues, or between those calculated with different FFs, do not.

The difference in basin populations for the different FFs is manifest in residue-dependent backbone entropies (Table 1 and Figure 5). For example, the entropy is the lowest with the Amber 94 FF, where essentially all the population is in the helical basin. For the center alanine residue in tri-alanine, the backbone entropy $T\Delta S$ calculated using Amber 96 is larger than $T\Delta S$ calculated using Amber 94, by 0.18 kcal mol $^{-1}$ (1 cal = 4.184 J). This increase reflects the binary basin occupancy between the extended and PP-II basins for Amber 96. Because the simulations with the other five FFs yield a more uniform

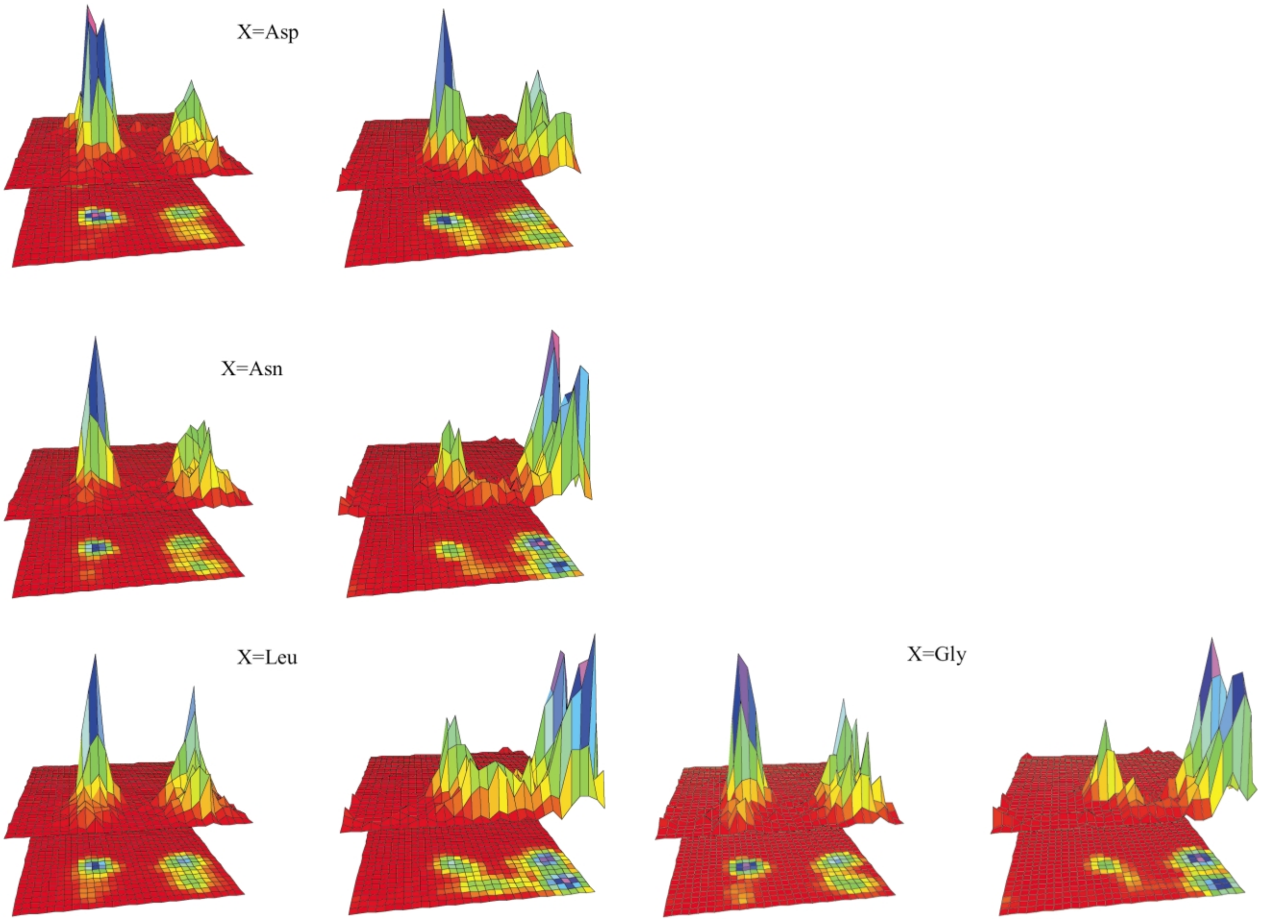

Figure 4. Sequence-dependence of NN effects. The population distribution for the center Ala is presented for two molecules, (a) Ala-Ala-X and (b) X-Ala-Ala calculated using the G-S-94 and OPLS-AA-01 FFs for X = (Ala, Gly, Leu, Trp, Met, Asn, Asp). The fractional population of the center Ala depends upon the neighboring residue type and whether it is N or C-terminal (Table 2).

distribution of these three basins, the backbone entropy of the center residue in tri-alanine for most FFs exceeds the Amber 94 entropy by ~ 0.4 kcal mol $^{-1}$.

The change in an alanine's backbone entropy with different neighbors is of the same order of magnitude as the difference in backbone entropy between different residues (Table 4). On average, the change in backbone entropy of Ala with different neighbors is ~ 0.1 kcal mol $^{-1}$, which is approximately the average difference in the backbone entropy between the different types of residues. This difference in entropy between individual residues is illustrated in Figure 5, where the backbone entropy is presented for each of the three residues in Ala-X-Ala where X ranges over the 20 naturally occurring amino acids. Although the backbone entropies for the G-S-94 and OPLS-AA-01 FFs often differ for individual amino acids, values for the flexible glycine and the highly restricted proline residues lie, as expected, near the extrema in both FFs. The calculations reproduce the known feature that residues preceding *trans*-proline residues are conformationally restricted. This effect is illustrated in Figure 5, where both G-S-94 and OPLS-AA-01 depict a low entropy for Ala1 when it precedes proline.

Geometric dependence of the NN effect

In addition to being sensitive to its NN side-chain identity, a residue's conformation is influenced by its NN's backbone conformation. The helical basin population of residue X in Ala-X-Ala often changes by twofold or more when both flanking alanine residues are in the helical basin. Figure 6 illustrates the influence of the NN conformation by presenting the difference in the backbone entropy:

$$(S_{\text{NN free}} - S_{\text{NN constrained}}) = T\Delta S$$

for each of the 20 amino acids as computed when both the flanking alanine residues are free to occupy all basins according to the equilibrium populations relative to when they are constrained to be in the helical basin. This entropy difference nearly vanishes for six to eight of the residues, depending upon the FF. However, $T\Delta S$ lies in the range of -0.5 to 0.2 kcal mol $^{-1}$ for the majority of

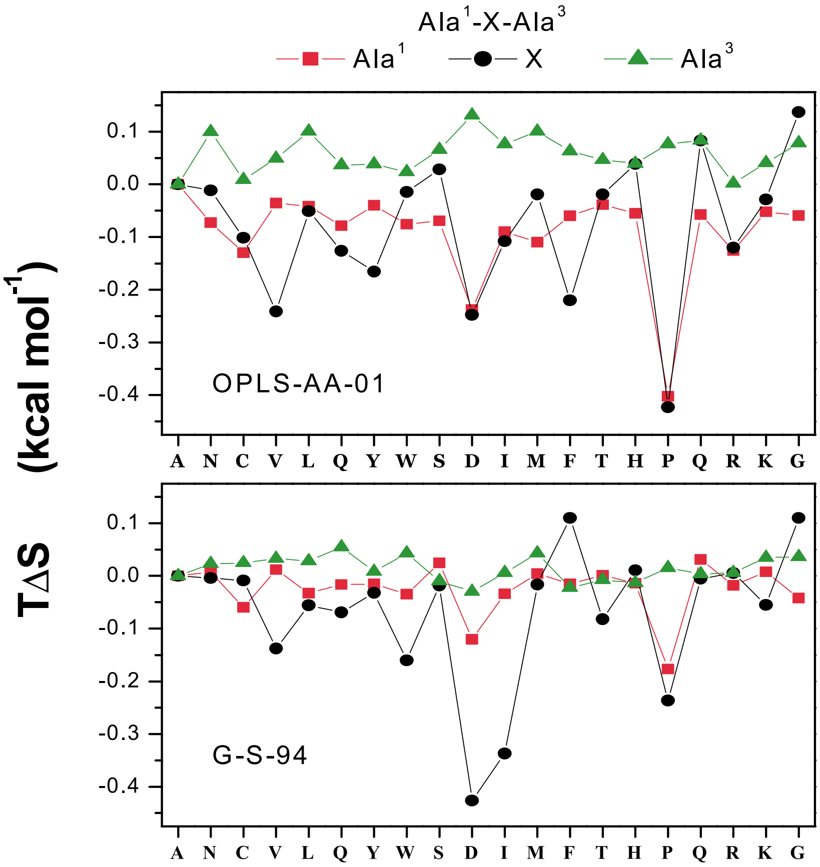

Figure 5. Backbone entropy and sequence-dependence of the NN effect. Entropy for each residue in Ala^1-X-Ala^3, referenced to that of a tri-alanine peptide as calculated with the (a) OPLS-AA-01 and (b) G-S-94 FFs. The value for residue X represents the variation in backbone entropy with amino acid type. Changes in the entropy of Ala1 or Ala2 reflect their dependence on residue X, while their difference is due to being N or C-terminal to the center residue, as well as being at either end. The abscissa is the one-letter code for the amino acids.

residues using either the G-S-94 and OPLS-AA-01 FFs. Thus, a residue's configuration can be affected significantly by its NN conformations.

Because a residue's entropy depends upon its neighbors' conformation, the backbone entropy of the system is not the sum of the individual residues' entropies. To estimate the magnitude of the non-additivity, the entropy of pairs of residues in a trimer molecule are calculated from the location of the pair's configuration in a 4D Ramachandran plot ($((\Phi, \Psi)_{i=1,2})$). This behavior is illustrated for the peptides AAA, LLL, VVV and for a pseudo-random sequence Ala-Glu-Thr-Asn. The difference in the correlated entropy and the sum of the entropies of the individual residues, calculated assuming that they are independent of their NNs' conformation, is in the range of $T\Delta S \sim 0.3$–0.7 kcal mol $^{-1}$ residue^{-1} depending upon the FF employed (Table 5). This range of non-additive contributions is about half the estimated loss of backbone entropy per residue upon unfolding based on experimental data.[22,23] Hence, the non-additive correction is quite significant, and the IPH is inadequate to describe the backbone

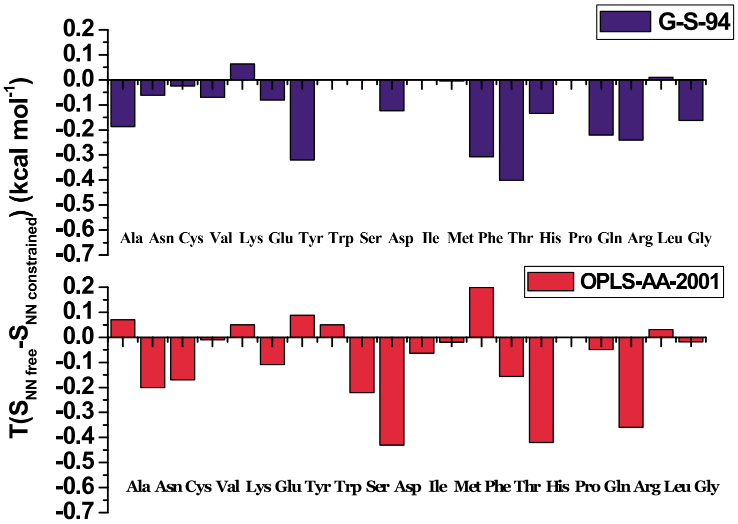

Figure 6. Backbone entropy and conformational dependence of the NN effect. The difference in the backbone entropy for residue X in the tripeptide Ala-X-Ala when the flanking alanine residues are free to be in any basin *versus* when both the flanking residues are in the helical basin (B3).

Table 5. Reduction in backbone entropy due to NN correlations

	$TS_{1,2}$	$T(S_1+S_2)$	$T\Delta S_{1+2}$	$TS_{2,3}$	$T(S_2+S_3)$	$T\Delta S_{2+3}$	$TS_{3,4}$	$T(S_3+S_4)$	$T\Delta S_{3+4}$
A. *Ala-Ala-Ala*									
G-S-94	−3.66	−2.40	1.26	−3.69	−2.57	1.12			
CHARMM	−3.63	−2.30	1.33	−3.77	−2.68	1.09			
OPLS-UA	−3.61	−2.18	1.42	−3.63	−2.34	1.29			
OPLS-AA-01	−3.66	−2.47	1.19	−3.66	−2.45	1.21			
B. *Val-Val-Val*									
G-S-94	−3.77	−2.88	0.88	−3.85	−3.16	0.69			
CHARMM	−3.83	−3.13	0.7	−4.04	−3.41	0.63			
OPLS-UA	−4.01	−3.49	0.52	−3.88	−3.16	0.72			
OPLS-AA-01	−4.00	−3.36	0.64	−4.01	−3.51	0.50			
C. *Leu-Leu-Leu*									
G-S-94	−3.91	−3.34	0.55	−3.76	−2.97	0.78			
CHARMM	−3.84	−3.02	0.82	−4.21	−3.58	0.63			
OPLS-UA	−3.69	−2.57	1.12	−3.71	−2.68	1.02			
OPLS-AA-01	−3.78	−2.95	0.83	−3.79	−2.96	0.92			
D. *Ala-Glu-Thr-Asn*									
G-S-94	−3.68	−2.42	1.25	−3.77	−2.64	1.13	−3.74	−2.68	1.05
CHARMM	−3.68	−2.59	1.08	−3.86	−3.13	0.73	−3.93	−3.36	0.57
OPLS-UA	−3.62	−2.26	1.35	−3.68	−2.41	1.27	−3.68	−2.61	1.07
OPLS-AA-01	−3.78	−3.01	0.77	−4.04	−3.54	0.50	−4.13	−3.63	0.50

Values are in kcal mol $^{-1}$. Reduction in backbone entropy due to NN correlations is obtained according to: ΔS_{i+j} = (entropy of residue i + entropy of residue $i+1$) − (entropy of the system composed of residue i and residue $i+1$, calculated using 4D Ramachandran map) resolved in 10×10 grid elements. As with all calculations of entropy, the value for S depends on the mesh-size, and the numbers listed are relative (see the text). However, entropy differences (ΔS) do not depend on mesh size, and are in absolute terms.

entropy of short peptides. Therefore, an accurate calculation of the unfolded state entropy must include correlations of backbone motions for neighboring residues.

Backbone dynamics

The rates of transitions between basins (or basin escape rates) are studied for each of the seven FFs using the basin auto-correlation function:

$$C_i(t) = \langle P_i(t)\cdot P_i(0)\rangle \qquad (2)$$

where $P_i(t)$ is the probability of being in the ith basin at time t. $P_i(t)$ is defined as unity if the residue is in basin i at time t and is zero otherwise. The long time-limit of the correlation function $C_i(t)$ approaches a constant that equals the equilibrium population of basin i for the FF. The correlation functions for the helical basin 3 are nearly exponential for the different FFs (Figure 7(a)), a behavior consistent with first-order kinetics for the escape from the basins. Poor fits to an exponential arise for transitions out of basins with very low populations because of meager statistics in these cases. This trend of exponential decay kinetics is observed also in the basin escape rates for basins 1 and 2 (data not shown). Inter-basin transition rates k_{ij} are obtained from fitting the correlation functions with an exponential decay towards the constant long time-limit as described in Methods.

The time-constant for escape from the helical basin of an Ala residue exhibits an eightfold dispersion as the FF is varied (Table 1 and Figure 7(b)). As expected, the Amber 94 FF yields the slowest rate due to its overwhelming population in basin 3, while the Amber 96 and OPLS-AA-97 FFs produce the fastest rates due to their negligible populations in the helical basin. The Amber 96 and G-S-94 rates differ by a factor of 5, which arises solely from the flattening of the added torsional potential for the G-S-94 FF (Figure 8). The correlation functions for the other basins exhibit a very similar dispersion in rates, as do those for the alanine residue in an Ala-Ala di-peptide (data not shown). A similar dispersion in rates appears in the explicit solvent calculations reported by Mu *et al.* for the tri-alanine peptide, where the authors suggest that the hopping rates vary by almost an order of magnitude for different FFs.[10]

An interesting aspect of the dynamics is the directional sampling of the Ramachandran basins, i.e. the existence of preferential transitions between certain basins. An analysis of the inter-conversions among the three major basins indicates that transitions are predominantly between basin 2 and either basin 1 or basin 3 (Figure 7(b)) but not between basins 1 and 3. This behavior is common for all FFs (except the OPLS-UA and OPLS-AA-97 FFs, where basins 1 and 2 coalesce into a single basin), indicating that directional basin sampling is general. The origin of the directional sampling can be viewed, for example, as the requirement that the left-handed PP-II conformation (basin 1) tends first to untwist (basin 2) before it can re-twist into the right-handed α-helical conformation (basin 3).

The basin hopping rates depend also on the NN identity. The hopping rate of Ala2 in AAX and XAA changes by almost 50% between X = Ala and X = Gly. Similarly, X = Asn and X = Ala

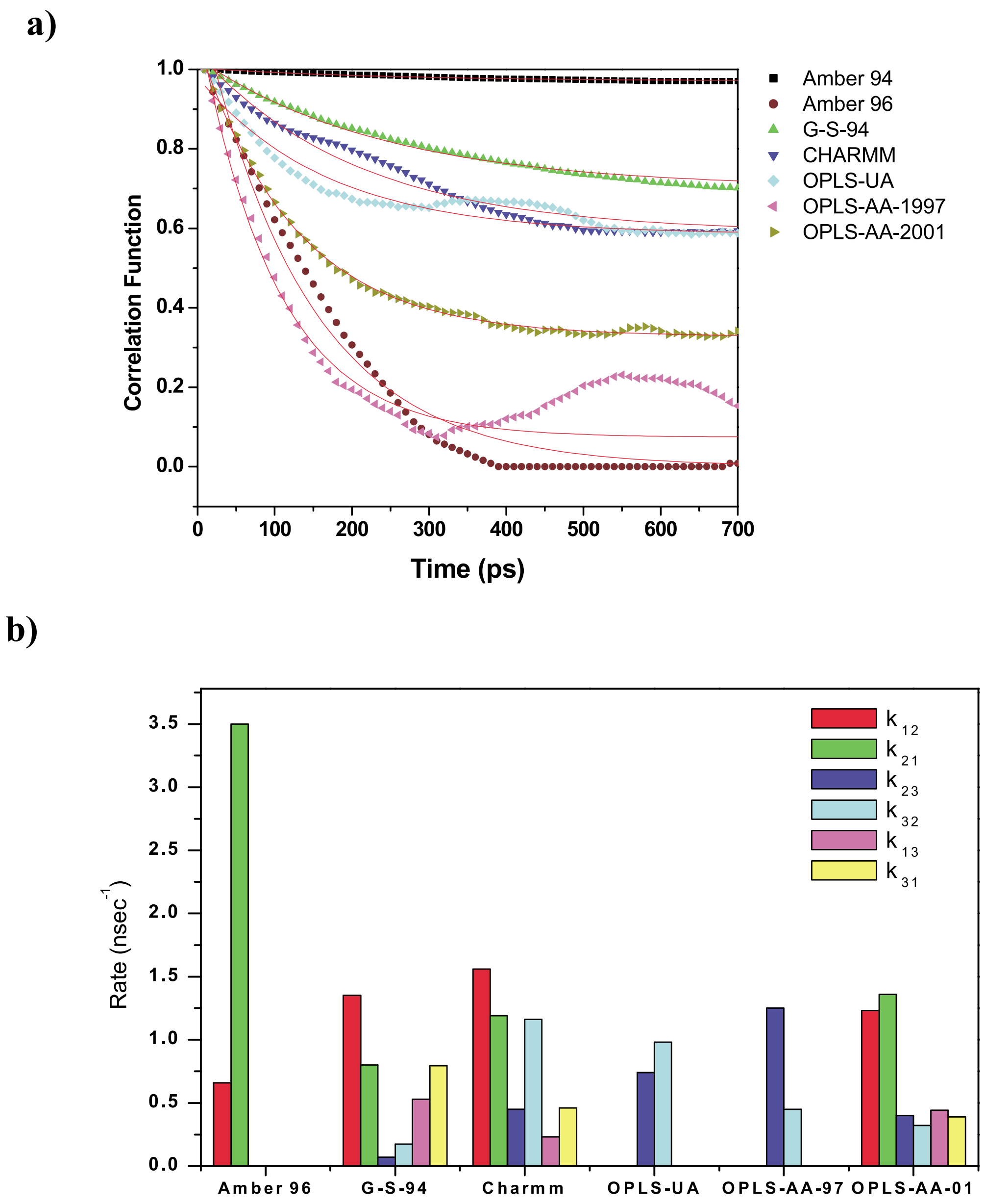

Figure 7. Basin hopping rates and directional sampling. (a) Correlation function for the basin 3 population for the center alanine residue in a tri-alanine peptide. The long time limit of the correlation function is the equilibrium population of basin 3, which varies strongly with the FF. Escape rates are obtained from single-exponential fits to the correlation functions (red lines). The finite duration of the simulations is responsible for some of the noise in the correlation functions. The poorer exponential fits for the Amber 96 and OPLS-AA-97 FFs probably arise because of the small basin 3 populations and because a limited number of transitions occur during a 15 ns trajectory for these two FFs. (b) Interbasin hopping rates for tri-alanine as calculated with several FFs. Rates for the Amber 94 FF are not presented because essentially only one basin is populated and there are very few transitions. Similarly, the rates between basin 2 and basin 3 for Amber 96 are omitted because basin 3 is occupied only rarely. For the OPLS-UA and OPLS-AA-97 FF, the rates are between basin 3 and the combined basins 1 and 2, which are not distinct in these FFs.

display a difference of about 50% in hopping rates (data not shown).

Discussion

We have investigated the backbone dynamics of different peptides using Langevin dynamics simulations with a validated implicit solvent model and employing a variety of commonly used FFs. A residue's conformation, as well as its location in the peptide sequence, can affect its neighbor's Ramachandran basin populations and basin interconversion rates significantly (except with the Amber 94 FF). For example, when the two flanking residues in a trimer are restricted to the helical basin, the residue's backbone entropy may change by the same order of magnitude as the difference in backbone entropy between different amino acids. These results are similar to those reported

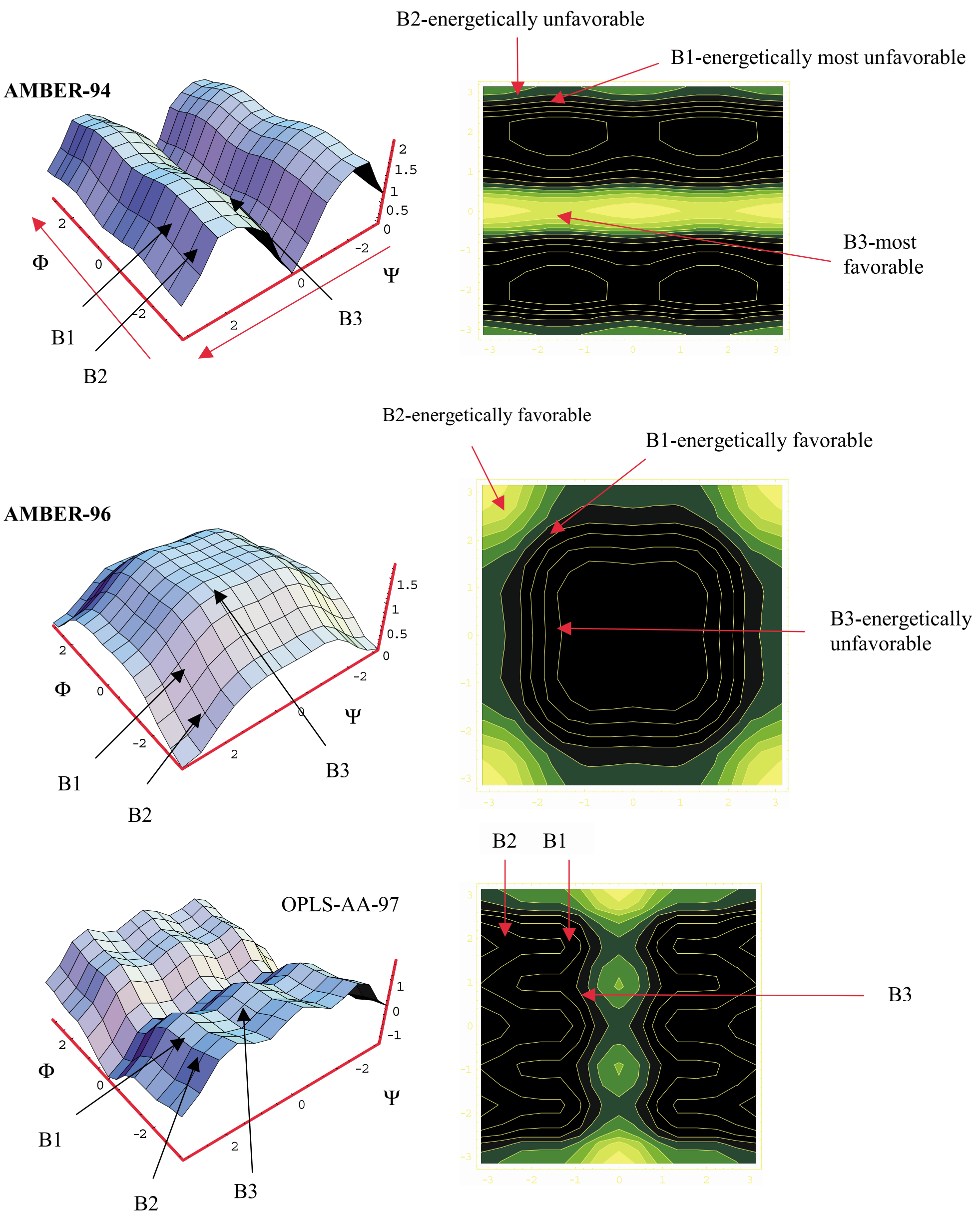

Figure 8 (*legend opposite*)

by Pappu *et al.*, though quantitative differences exist due to their use of a more simplified hard-sphere potential. The influence of either neighboring residues' identity on the backbone entropy of a residue is of the same magnitude. Additionally, the identity of the NN can alter the rate at which an alanine residue leaves, for example, the helical basin by nearly 50%.

Decrease in backbone entropy due to correlated motions

The influence of the NN's conformation on the torsional populations and kinetics of a residue demonstrates the invalidity of the Flory IPH. A similar conclusion is reached by Pappu *et al.*, who also observe a reduction in available conformations for the terminal alanine residue of a helical segment. We quantify the extent to which backbone conformations are coupled by calculating the difference, ΔS, between the sum of the independent entropies of each residue for a bonded pair of amino acid residues and that for the correlated pair. This difference is considerable, $T\Delta S$ ~0.3–0.7 kcal mol $^{-1}$ residue^{-1} (Table 5). Additionally, we observe that these correlations are due to inter-basin motions as illustrated by the following example. For the pseudo-random sequence Ala-Glu-Thr-Asn, the entropy change, ΔS, from correlated motions, as calculated with a coarse, basin-level mesh, *versus* that with a finer mesh in the 4D plot of (Ψ_1, Φ_1, Ψ_2, Φ_2), differs by only ~25% (data not shown). Because the finer mesh entropy is sensitive to intra-basin motions, while the coarser mesh entropy is sensitive only to inter-basin motions, the small difference suggests that

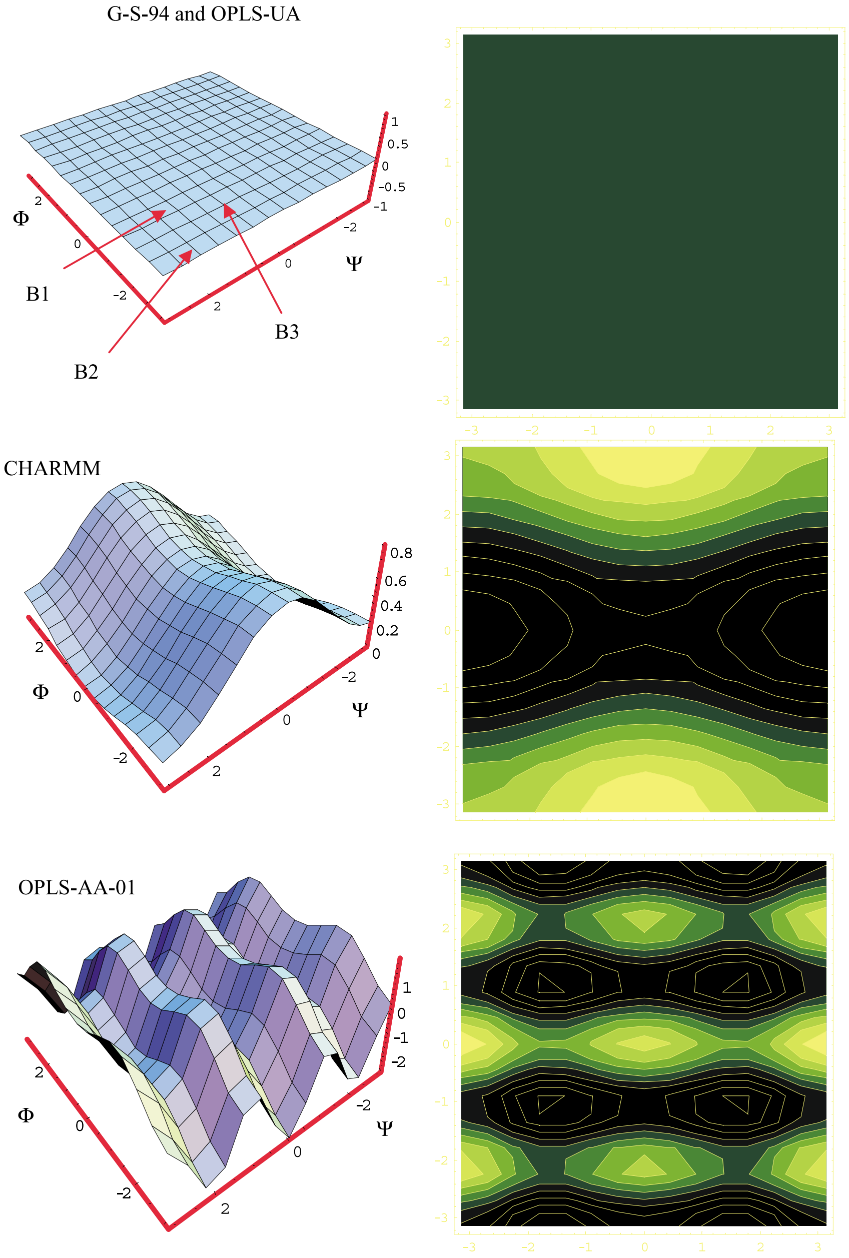

Figure 8. Torsional biases in the FFs: The added backbone torsional potential of the different FFs. The right column shows a contour plot of the surface in the left column. The axes are labeled in radians.

the correlations are primarily at the inter-basin level.

Fortunately, most experiments measure the entropy of the system as a whole and therefore automatically include all contributions from the correlated motions. Future studies will investigate the magnitude of the neighbor effects in an entire protein sequence and whether the entropy of the system is strongly dependent upon sequence order, rather than just composition.

Differences and accuracy of FFs

We have studied the equilibrium populations and inter-basin hopping kinetics with seven widely used FFs to examine the robustness of our conclusions, as well as to address questions concerning the consistency and reliability of the FFs for treating protein dynamics. As noted by Hu *et al.*[4] and Mu *et al.*,[10] an important difference among the FFs is the bias towards certain basins (as exhibited in Figures 2 and 8). The Amber 94 FF describes alanine-like residues as largely populating only the helical basin, while the Amber 96 FF avoids this basin completely. The remaining five FFs lead to the helical, extended and PP-II basins as being populated more equally, although non-helical basins are not distinct in the OPLS-UA and OPLS-AA-97 FFs.

Garcia and co-workers correct for the "helixophobicity" of the Amber 96 FF by completely flattening the added Amber 96 torsional potential, which is 1.5 kcal mol $^{-1}$ unfavorable at the helical basin (Figure 8). The OPLS-UA FF also has a flat added torsional potential, while the added potential varies by 0.5 kcal mol $^{-1}$ for CHARMM and by as much as 2 kcal mol $^{-1}$ for the OPLS-AA-97 and OPLS-AA-01 FFs. However, the added torsional potential determines only a portion of the backbone distribution, because other interactions, such as partial charges, side-chain dihedral potentials, and van der Waals interactions, contribute as well.[4,10,21]

Recent experimental studies have shown that alanine-rich unfolded peptides predominately populate the PP-II basins.[17,19,24–28] Except for the Amber 94 FF and the OPLS-UA FF, all FFs predict significant sampling of PP-II conformations. However, the PP-II basin still is not the most populated for any of the FFs. Thus, there is a disparity between the predictions of the FFs using implicit solvent and experimental observations for very small peptides. However, some of the above-mentioned experiments were carried out on molecules that were either uncapped (charged (+1) N terminus and (−1) C terminus),[28] or partially capped (+1 charged N terminus).[26] Hence, the simulations should exhibit increased PP-II basin populations over that expected for the neutral, capped version, which provides the best mimic of longer polypeptides.

We suggest that the discrepancy between the FFs may reflect the fact that they have been designed on the basis of thermodynamic data (perhaps with some *ab initio* computations centered near potential minima). The protein FFs have generally been validated by the degree to which they can reproduce the structures of folded proteins. However, folded proteins tend to have less PP-II structures than either helical or β-sheet structures, so the underweighting of the PP-II basin by the FFs is, perhaps, not too surprising. Additionally, the dynamics sensitively reflects the heights of the saddle-points connecting the basins, while the thermodynamic and quantum data used to parameterize the FFs are insensitive to these kinetic barriers.

Glycine flexibility and helical propensity

Compared to alanine, the backbone of glycine is more flexible, as it can traverse a larger range of the Ramachandran map (Figure 2(c)). However, glycine still exhibits strongly preferred regions. This preference reduces the overall sampling of configurations, and the backbone entropy is increased only modestly (using realistic FFs), e.g. $T(S_{\text{Ala-Gly-Ala}} - S_{\text{Ala-Ala-Ala}}) \leq 0.11\ \text{kcal mol}^{-1}$ (Table 4).

It is generally believed that the difference in the helical propensities between Ala and Gly at a solvent-exposed position is attributed entirely to differences in the backbone entropy of the unfolded state, because the folded state has the same entropy and interactions[23,29] The difference in helical propensity between glycine and alanine is greater than 0.7 kcal mol $^{-1}$,[23,29] far larger than their difference in backbone entropy for the unfolded state (~0.11 kcal mol $^{-1}$; Table 4). This discrepancy between the known helical propensity and our calculation of backbone entropy implies either that (1) the FFs do not reproduce the sampling of the unfolded state for alanine and/or glycine accurately, or (2) the difference in backbone entropy between Ala and Gly in the unfolded state is not the primary factor determining the difference in helical propensity for these two residues. Additionally, our simulations show that a Gly reduces the preceding Ala's helical population (Figure 4(a)). This NN effect suggests that the low helical propensity of Gly may be due to a combination of its high backbone entropy in the unfolded state, and of the reduction it imposes on its N-terminal neighbor's tendency to be in a helical conformation.

Time-scales and comparisons with experiments

Our results indicate that simulations employing different commonly used FFs can produce basin hopping rates differing by about fivefold as well as populate the major basins differentially. Though our basin hopping rates are dependent on the viscosity of the solvent, we believe that for small peptides, the rates will only rescale with a different viscosity coefficient[16] Our findings agree with recent simulations obtained by Hu *et al.*[4] and Mu *et al.*[10] using explicit solvent MD simulations and several different FFs.

Inter-basin hopping rates and basin sampling affect the folding pathways and, hence, the overall dynamics that are predicted by simulations. Consequently, the folding rate determined from folding simulations may contain further uncertainties. Given these issues, one should expect a factor of at least 2–3 uncertainty in simulated rates[30] because of the uncertainties in the FFs. Furthermore, the extreme bias towards the helical basin from the Amber 94 FF implies that any folding simulation using this FF is unreliable for either dynamics or thermodynamics.

In addition, protein folding simulations with a variety of FFs often tend to exhibit early collapse and the formation of structured intermediates[31–36] In contrast, the folding of small proteins is observed experimentally to be two-state without the accumulation of early intermediates.[37] Potentially, the early intermediates observed in the simulations arise due to inherent limitations of the FFs, which are designed primarily to describe folded structures and not the dynamics of the folding process.

A possible source of the early collapse found in the simulations may lie in an inadequate treatment of the backbone entropy of the unfolded state. Although the backbone entropies are generally within ~1/2 kcal mol $^{-1}$ of each other for the seven FFs, these values are for a single residue. Even a 0.1 kcal mol $^{-1}$ systematic error for a small, 100 residue protein could produce a net error of 10 kcal mol $^{-1}$, or a factor of 10^7 in the equilibrium constant for a fully collapsed species relative to the unfolded state. Thus, small systematic errors in parameters of FFs can easily lead to folding mechanisms that are not observed experimentally.

Conclusions

Our simulations demonstrate that the Flory IPH is invalid because of significant interactions between neighboring amino acid residues. The basin preference and backbone entropy of a residue depends on its neighbor's conformation and identity. We estimate the magnitude of these effects to be $T\Delta S \leq 0.7$ kcal mol $^{-1}$ residue^{-1}. Because most, but not all[38,39] implementations of Zimm–Bragg[40] and Lifson–Roig[41] helix-coil theories do not include either a dependence on the sequence or on the conformation, there is opportunity for improving these theories by correcting for the changes in the entropy of the unfolded state due to NN effects.

Basin populations and inter-conversion rates depend strongly on the choice of FF. This dependence is larger than differences between explicit and implicit solvent calculations using the same FF, suggesting that explicit solvent calculations for the dynamics of small peptides may be unnecessary until the FFs are improved.

The information we obtain concerning the basin hopping rates can be used in coarse-grained folding algorithms that are based solely on torsional dynamics.[42] Moreover, the preference of peptides for certain conformations can help characterize the structure and dynamics of the denatured state, and their influence on the folding pathway.

Methods

The long-time dynamics (15–45 ns) of the di- and tripeptides have been probed using the implicit solvent LD simulation method described by Shen & Freed[16] using seven different FFs at 300 K. The peptides are amino-acetylated and carboxy-amidated in order to model the dynamics of the two or three residues within a larger polypeptide. Similar simulations with uncapped ends lead to very different propensities for individual Ramachandran basins, mainly because the charged ends favor elongated configurations more than in capped systems. Average basin populations and dynamics are accumulated after the first 3 ns of the equilibration simulations.

The Langevin dynamics simulations take the total system energy:

$$U_{\text{total}} = U_{\text{b}} + U_{\text{bend}} + U_{\text{tors}} + U_{\text{imp-tors}} + U_{\text{ch}}(\varepsilon(r)) + U_{\text{vdW}} + U_{\text{solv}}(\sigma)$$

as the sum of the usual types of interaction potentials between the solute atoms, while the solvent contributions are modeled using a distance dependent dielectric "constant" to screen charge–charge interactions $U_{\text{ch}}(\varepsilon(r))$ and a solvation potential $U_{\text{solv}}(\sigma)$. The bonding interactions U_{b}, bond–bond bending interactions U_{bend}, and improper torsional energies $U_{\text{imp-tors}}$ are modeled by harmonic potentials, the regular torsional potentials U_{tors} by standard periodic functions, and the van der Waals interactions by Lennard–Jones 6–12 potentials. The Coulomb interactions:

$$U_{\text{ch}}(\varepsilon) = \sum_{i>j} q_i q_j / \epsilon(r_{ij}) r_{ij}$$

are expressed in terms of atomic partial charges q_i and a Ramstein–Lavery-style[43] distance-dependent dielectric constant $\varepsilon(r)$. The microscopic solvation potential is modeled using the Ooi–Scheraga solvent-accessible surface area (SASA) potential:[44]

$$U_{\text{solv}}(\sigma) = \sum_{i=1}^{N} g_i \sigma_i$$

where σ_i is the accessible surface area of a hypersurface bisecting the first solvent shell surrounding protein atom i, and the empirical surface free energy parameters g_i depend on the atom type. Because the g_i are free energy parameters, the U_{total} generates a temperature-dependent free energy that contains contributions from solvent reorientation within a mean-field approximation.

The LD simulations employ the velocity Verlet algorithm[45] with a time-step of $\Delta t = 2$ fs for integrating the equations of motion for the protein atom positions and velocities. The lengths of all X–H type bonds are constrained using the RATTLE algorithm.[46] The computations are performed using a modified version of the TINKER 3.9 molecular design package† with a faster, non-bonding force evaluation algorithm FAST-LD[47] The frictional forces and corresponding random forces acting on the protein atoms are computed using the Pastor–Karplus accessible surface area model.[48] The solvent-accessible surface areas σ_i for the friction coefficients are calculated from the exposed surface area of solute atoms using a probe of zero radius. The smaller probe size for friction coefficients is used to cancel effectively the results of (more expensive to calculate) hydrodynamic interactions. The accessible surface areas, atomic friction coefficients, and solvation potentials are updated every 100 dynamical steps (0.2 ps), since tests show that this approximation incurs negligible error because significant conformational variations occur on a much longer time-scale.[16]

Identification of basin locations

The entropy calculations (equation (1)) do not depend upon how each basin is defined, as the probability is calculated for each of the 3° × 3° grid elements. However, the populations shown in Tables 1 and 2 and the rates depicted in Figure 7 depend upon the definitions of individual basins. Basins 1, 2 and 3 are defined on the basis of the population of central Ala in tri-Ala (Figure 2(a)). Basin 3 is defined with a circle large enough to encompass the population of that basin for all the FFs (Figure 2(b)). This definition is used to calculate the rate of escape from basin 3 shown in Figure 7. The distinction between basins 1 and 2 is applicable only for G-S-94, OPLS-AA-01 and CHARMM, as other FFs either do not have a clear separation between theses two basins (OPLS-UA and OPLS-AA-97) or have all of its population in only a single basin (Amber 94 and Amber 96). For basins 1 and 2, the G-S-94, OPLS-AA-01 and CHARMM FFs are used to define non-overlapping ellipses that are large enough to accommodate >90% of the populations in each of these basins.

Independence of initial conditions and length of simulation

In order to test the robustness of the computed neighbor effects, simulations have been performed for four

† http://dasher.wustl.edu/tinker/

different di-peptides with varying initial conditions and variable durations of 45 ns and 15 ns. The overall difference in basin populations is less than 3% (due to different initial conditions and longer trajectories), indicating that the basins are sampled adequately within 15 ns, and that the results are not an artifact of the initial conditions or the use of short trajectories.

Calculation of k_{ij} (inter-basin transition rates) from basin auto correlation function

In order to calculate k_{ij}, the rate of transition from basin i to basin j, the escape rate from each basin is calculated. The population decay rate is obtained from an exponential fit to the autocorrelation function $C_i(t)$ for each basin (after having subtracted the long-time basin population). Because transitions from basins 1 and 3 proceed primarily to basin 2, the decay rates of $C_i(t)$ for the basin 1 and 3 correlation functions equal k_{12} and k_{13}, respectively. The decay rate of $C_2(t)$ is the sum $k_{21} + k_{23}$, which can be separated using the equilibrium basin populations and the detailed balance condition for equilibrium, e.g.:

$$[\text{basin 1}]/[\text{basin 2}] = k_{12}/k_{21}$$

Calculation of backbone entropies

Equation (1) is only an approximate relation. The conformational entropy can be computed rigorously only from conformational populations when the latter are obtained from a constant energy simulation. However, both the friction coefficients and the solvation potential are inherently temperature-dependent quantities, so constant energy implicit solvent simulations are not possible. A more rigorous approach would be to follow the far more computationally costly simulation methods of Okamoto and co-workers,[49–51] but this would not be possible for the wide range of dimer and trimer systems and FFs studied here. Hence, the approximate form of equation (1) suffices for our broad study.

Acknowledgements

We thank Drs A. Fernandez, N. Kallenbach, R. Pappu, G. Rose and A. Colubri, and members of our groups for numerous enlightening discussions, and Dr Y. Mu for a preprint of his manuscript and for helpful correspondences. This work was supported by grants from the NIH (K.F.F. GM56678 and T.R.S. GM55694). M.H.Z. is a Burroughs-Wellcome Interdisciplinary Research Fellow.

References

1. Ramachandran, G. N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Advan. Protein Chem.* **23**, 283–438.
2. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99.
3. Bolhuis, P. G., Dellago, C. & Chandler, D. (2000). Reaction coordinates of biomolecular isomerization. *Proc. Natl Acad. Sci. USA*, **97**, 5877–5882.
4. Hu, H., Elstner, M. & 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: Struct. Funct. Genet.* **50**, 451–463.
5. Mu, Y., Kosov, D. S. & Stock, G. (2003). Conformational dynamics of trialanine in water II: comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments. *J. Phys. Chem. ser. B*, **107**, 5064–5073.
6. Flory, P. J. (1969). *Statistical Mechanics of Chain Molecules*, Wiley, New York.
7. Pappu, R. V., Srinivasan, R. & Rose, G. D. (2000). The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc. Natl Acad. Sci. USA*, **97**, 12565–12570.
8. Zaman, M. H., Berry, R. S. & Sosnick, T. R. (2002). Entropic benefit of a cross-link in protein association. *Proteins: Struct. Funct. Genet.* **48**, 341–351.
9. Garcia, A. E. & Sanbonmatsu, K. Y. (2002). Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc. Natl Acad. Sci. USA*, **99**, 2782–2787.
10. Mu, Y., Kosov, D. S. & Stock, G. (2003). Conformational dynamics of trialanine in water II: comparison of AMBER, CHARMM, GROMOS, and OPLS force fileds to NMR and infrared experiments. *J. Phys. Chem. ser. B*, **107**, 5064–5073.
11. Garcia, A. E. & Sanbonmatsu, K. Y. (2002). alpha-Helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc. Natl Acad. Sci. USA*, **99**, 2782–2787.
12. Kollman, P., Dixon, R., Cornell, W., Fox, T., Chipot, C. & Pohorille, A. (1997). The development/application of a “minimalist” organic/biochemical molecular mechanic force field using a combination of *ab initio* calculations and experimental data. In *Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications* (van Gunsteren, W. F. & Wiener, P. K., eds), pp. 83–96, Escom, Dordrecht.
13. MacKerell, A. D., Jr, Bashford, D., Bellott, M., Dunbrack, R. L., Jr, Evanseck, J. D., Field, M. J. *et al.* (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. ser. B*, **102**, 3586–3616.
14. Jorgensen, W. L. T.-R. J. (1988). The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657–1666.
15. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins *via* comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. ser. B*, **105**, 6474–6487.
16. Shen, M. Y. & Freed, K. F. (2002). Long time dynamics of met-enkephalin: comparison of explicit and implicit solvent models. *Biophys. J.* **82**, 1791–1808.
17. Shi, Z., Olson, C. A., Rose, G. D., Baldwin, R. L. & Kallenbach, N. R. (2002). Polyproline II structure in a sequence of seven alanine residues. *Proc. Natl Acad. Sci. USA*, **99**, 9190–9195.
18. Shi, Z., Woody, R. W. & Kallenbach, N. R. (2002). Is polyproline II a major backbone conformation in unfolded proteins? *Advan. Protein Chem.* **62**, 163–240.
19. Woutersen, S., Pfister, R., Hamm, P., Mu, Y., Kosov,

D. S. & Stock, G. (2002). Peptide conformational heterogeneity revealed from nonlinear vibrational spectroscopy and molecular-dynamics simulations. *J. Chem. Phys.* **117**, 6833–6840.

20. Pappu, R. V. & Rose, D. G. (2002). A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci.* **10**, 2437–2455.
21. Zaman, M. H., Shen, M., Berry, R. S. & Freed, K. F. (2003). Computer simulations of Met-Enkephalin using axplicit atom and united atom force-fields: similarities, differences and suggestions for improvement. *J. Phys. Chem. ser. B*, **107**, 1685–1691.
22. Thompson, J. B., Hansma, H. G., Hansma, P. K. & Plaxco, K. W. (2002). The backbone conformational entropy of protein folding: experimental measures from atomic force microscopy. *J. Mol. Biol.* **322**, 645–652.
23. D'Aquino, J. A., Gomez, J., Hilser, V. J., Lee, K. H., Amzel, L. M. & Freire, E. (1996). The magnitude of the backbone conformational entropy change in protein folding. *Proteins: Struct. Funct. Genet.* **25**, 143–156.
24. Dukor, R. K. & Keiderling, T. A. (1991). Reassessment of the random coil conformation: vibrational CD study of proline oligopeptides and related polypeptides. *Biopolymers*, **31**, 1747–1761.
25. Woutersen, S., Mu, Y., Stock, G. & Hamm, P. (2001). Subpicosecond conformational dynamics of small peptides probed by two-dimensional vibrational spectroscopy. *Proc. Natl Acad. Sci. USA*, **98**, 11254–11258.
26. Woutersen, S. & Hamm, P. (2001). Time-resolved two-dimensional vibrational spectroscopy of a short alpha-helix in water. *J. Chem. Phys.* **115**, 7737–7743.
27. Woutersen, S. & Hamm, P. (2001). Isotope-edited two-dimensional vibrational spectroscopy of trialanine in aqueous solution. *J. Chem. Phys.* **114**, 2727–2737.
28. Schweitzer-Stenner, R. (2002). Dihedral angles of tripeptides in solution directly determined by polarized Raman and FTIR spectroscopy. *Biophys. J.* **83**, 523–532.
29. Creamer, T. P. & Rose, D. G. (1994). Alpha-helix-forming propensities in peptides and proteins. *Proteins: Struct. Funct. Genet.* **19**, 85–97.
30. Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. (2002). Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, **420**, 102–106.
31. Zagrovic, B., Snow, C. D., Khaliq, S., Shirts, M. R. & Pande, V. S. (2002). Native-like mean structure in the unfolded ensemble of small proteins. *J. Mol. Biol.* **323**, 153–164.
32. Shimada, J. & Shakhnovich, I. E. (2002). The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc. Natl Acad. Sci. USA*, **99**, 11175–11180.
33. Duan, Y. & Kollman, A. P. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, **282**, 740–744.
34. Shen, M. Y. & Freed, F. K. (2002). All-atom fast protein folding simulations: the villin headpiece. *Proteins: Sturct. Funct. Genet.* **49**, 439–445.
35. Alonso, D. O. V. & Daggett, V. (1998). Molecular dynamics simulations of hydrophobic collapse of ubiquitin. *Protein Sci.* **7**, 860–874.
36. Linhananta, A., Zhou, H. Y. & Zhou, Y. Q. (2002). The dual role of a loop with low loop contact distance in folding and domain swapping. *Protein Sci.* **11**, 1695–1701.
37. Krantz, B. A., Mayne, L., Rumbley, J., Englander, S. W. & Sosnick, T. R. (2002). Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding. *J. Mol. Biol.* **324**, 359–371.
38. Poland, D. C. & 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.
39. Go, M., Hesselink, F. T., Go, N. & 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.
40. Zimm, G. H. & Bragg, K. J. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* **31**, 526–535.
41. Lifson, S. & Roig, A. (1961). On the theory of the helix-coil transion in polypeptides. *J. Chem. Phys.* **34**, 1963–1974.
42. Colubri, A. & Fernandez, A. (2002). Pathway Diversity and Concertedness in Protein Folding, An ab-initio Approach. *J. Biomol. Struct. Dynam.* **19**, 739–764.
43. Ramstein, J. & Lavery, R. (1988). Energetic coupling between DNA bending and base pair opening. *Proc. Natl Acad. Sci. USA*, **85**, 7231–7235.
44. Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl Acad. Sci. USA*, **84**, 3086–3090.
45. Allen, M. P. & Tildesley, D. J. (1987). *Computer Simulation of Liquids*, Oxford University Press, Oxford.
46. Anderson, H. C. (1983). Rattle: a velocity version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* **52**, 24–34.
47. Shen, M. Y. PhD Thesis, Computer simulations of protein dynamics in Chemistry, 2002, University of Chicago: Chicago
48. Pastor, R. W. & Karplus, M. (1988). Parametrization of the friction constant for stochastic simulations of polymers. *J. Phys. Chem.* **92**, 2636–2641.
49. Sugita, Y. & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Letters*, **314**, 141–151.
50. Sugita, Y., Kitao, A. & Okamoto, Y. (2000). Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* **113**, 6042–6051.
51. Mitsutake, A., Kinoshita, M., Okamoto, Y. & Hirata, F. (2000). Multicanonical algorithm combined with the RISM theory for simulating peptides in aqueous solution. *Chem. Phys. Letters*, **329**, 295–303.

Edited by M. Levitt

(Received 7 March 2003; received in revised form 15 May 2003; accepted 2 June 2003)