A statistical analysis of the PPII propensity of amino acid guests in proline-rich peptides (Supporting Material)

Mahmoud Moradi, Volodymyr Babin, Celeste Sagui*, and Christopher Roland*
Center for High Performance Simulations (CHiPS) and Department of Physics, North Carolina State University, Raleigh, NC 27695-8202

I. SAMPLING PROTOCOL

The conformational equilibrium of the proline-rich peptides is characterized, among other things, by different cis/trans patterns of the prolyl bonds. The free energy barriers separating these cis and trans states are relatively high (of the order of 15 kcal/mol[1]), so that regular molecular dynamics (MD) simulations at room temperatures are unable to overcome the barriers in a reasonable amount of time, and therefore cannot be used to study the conformational equilibria directly. To deal with this sampling issue, we therefore made use of the *replica exchange*[2] scheme.

In the replica exchange method[2], one considers several copies (replicas) of a system subject to some sort of ergodic dynamics based on different Hamiltonians, and attempts to exchange the trajectories of these replicas at some predetermined rate. Care must be taken with respect to the choice of the Hamiltonians, since these determine the performance of the method. In this regard, it is convenient to consider the following two aspects: (a) the details of the so-called "hot" replica that facilitates the crossing of barriers, and (b) the random walk between the replicas. The latter is typically described in terms of an exchange rate between pairs of replicas. Let us assume, for a moment, that these rates are sufficiently high, so that the random walk in replica space is efficient. The purpose of the "hot" replica is to increase the barrier crossing rates (or, more formally, to decrease the ergodic time scale). One possibility for this is to run the hot replica at high temperature. Another possibility[3] is to construct the hot replica by adding a biasing potential to the original Hamiltonian that acts on some collective variable that (presumably) describes one of the slow modes of the system that need "acceleration". A combination of such Hamiltonian and Temperature based replica exchange molecular dynamics (HT-REMD[3–5]) provides for a practical way to reduce the computational costs associated with REMD sampling, since it facilitates the sampling in the "hottest" replica by both means, and therefore also allows for a better "tuning" of the entire setup. Finally, the mixing properties of the hottest replica must be assessed separately (by simulating it alone), in order to make sure it is ergodic, and that its ergodic time scale is sufficiently short as to generate enough independent samples within a feasible runtime.

While elevated temperatures provide for a generic way to promote barrier crossing events, the use of biasing potentials (U) allows one to directly focus on specific slow modes of the system. The latter can often be identified on the grounds of chemical and physical intuition, and are usually described in terms of a *collective variable* $\sigma = \sigma(\mathbf{r})$ defined as a smooth function of the atomic positions $\mathbf{r} = \mathbf{r}_1, \ldots, \mathbf{r}_N$. The corresponding free energy or potential of mean force (PMF)[6]:

$$f(\xi) = -k_B T \ln \left\langle \delta \left[\xi - \sigma(\mathbf{r}) \right] \right\rangle,$$

(the angular brackets denote an ensemble average) provides for an ideal biasing potential to be used for the hottest replica. Indeed, for the system biased with $U(\mathbf{r}) = -f\left[\sigma(\mathbf{r})\right]$, the probabilities of different values of the collective variable would all be equal, since there are no barriers present. Unfortunately, the true free energy $f(\xi)$ is typically unknown in advance. However, even an approximate $f(\xi)$, accurate within a few $k_B T$, is often sufficient. The latter can be computed in a variety of ways[6]. In this work, we make use of the Adaptively Biased Molecular Dynamics (ABMD)[7] method in order to calculate suitable biasing potentials for HT-REMD simulations.

The ABMD[7] method is an umbrella sampling method with a time-dependent biasing potential modifying the potential energy Φ of the system

$$\Phi_{ABMD}\left(\mathbf{r}, t\right) = \Phi\left(\mathbf{r}\right) + U\big[\sigma\left(\mathbf{r}\right), t\big].$$

This biasing potential "floods" the true free energy landscape as it evolves in time according to:

$$\frac{\partial U(\xi, t)}{\partial t} = \frac{k_B T}{\tau_F} G\big[\xi - \sigma\left(\mathbf{r}\right)\big].$$

*Electronic address: sagui@ncsu.edu, cmroland@ncsu.edu

Here, $G(\xi)$ is a positive definite, symmetric kernel (in analogy with the *kernel density estimator* widely used in statistics [8]), which may be thought of as a smoothed Dirac delta function. For large enough τ_F (the flooding timescale) and small enough kernel width, the biasing potential $U(\xi, t)$ converges towards $-f(\xi)$ as $t \to \infty$[9, 10]. ABMD can be used in conjugation with the replica exchange protocol. In this case it is possible to use different collective variables and/or temperatures on a per-replica basis (please see Refs.3, 7 for a more detailed exposition). Currently, the ABMD method has been implemented into the AMBER v.10, 11 simulation packages[11], and is freely available to the simulation community. To date, the method has been successfully used to investigate a variety of biomolecular systems such as various peptides and sugars[1, 3, 5, 7, 12–14].

Our simulations proceeded in stages, and first computed the approximate free energy associated with a chosen collective variable that "captures" the cis/trans transitions of the prolyl bonds at different temperatures using a combination of ABMD and parallel tempering. In the next step, several additional replicas running at the lowest temperature T_0 were introduced into the setup. One of these replicas is completely unbiased, and therefore samples the *unbiased* Boltzmann distribution at $T = T_0$. The other replicas, also at $T = T_0$, are subject to a reduced bias (*i.e.*, these biasing potentials are scaled down by a constant factor). The purpose of these "proxy"-replicas is to ensure adequate exchange rates, and thereby enhance the mixing. A more detailed presentation of this point can be found in Ref.3.

II. EMPIRICAL POTENTIAL

There is reasonable consistency between the AMBER ff99SB force field and *ab initio* data for the cis/trans isomerization of the prolyl bond. This issue has previously been discussed by us and others[14–19]. In brief, the energies associated with the equilibrium state have been shown to be in very good agreement with the *ab initio* results, although there are some differences associated with the transition barriers.

As an illustration, Fig.S2 shows a comparion of the energy of the Ace-Pro-Pro-NMe molecule as a function of the Pro-Pro ω angle as obtained from the AMBER ff99SB force field and *ab initio* data. The structures were optimized in the gas phase for $\omega = -180^\circ, -177^\circ, \ldots, 180^\circ$ (every 3°) carefully following the various meta-stable branches, and choosing the smallest values of energy. The *ab initio* data was calculated using the HF/6-31G* model with the latest version of Gaussian. The relative depths of the minima compare very well (≈ 0.6 kcal/mol for ff99SB versus ≈ 0.1 kcal/mol for ab initio data). The *ab initio* transition barriers are somewhat larger, and differ by ≈ 2 kcal/mol. It is important to note, however, that the precise values of the potentials away from the minima are not critically important for our study, as we focus on equilibrium properties only. The latter are, of course, determined by the depths and the shapes of the minima. Finally, apart from the empirical potential itself, there is several other sourves of error with the next largest one being the GB/SA solvation model. The total error in computed properties of the molecules is therefore very difficult to estimate a priori. Hence, we believe that the ultimate judgement is to be made on the basis of comparison against experimental data.

III. SIMULATION DETAILS

The ABMD simulations were carried out using 20 replicas with temperatures ranging from 300 to 1200 K. These simulations were performed for PPP and PAP peptides. Both PPP and PAP simulations involved long ABMD runs with Ω as the collective variable using replica exchange with the temperatures distributed as: 300, 322, 347, 373, 401, 432, 464, 499, 537, 578, 622, 669, 720, 774, 833, 896, 964, 1037, 1115, 1200 K. Each replica had its own biasing potential. A kernel width of $4\Delta\xi = 0.2$ was used with a flooding timescale of $\tau_F = 25$, 50, 100, or 200 *ps* at different stages of the 100 *ns* runtime. We then repeated the simulations using the calculated biasing potentials, with PAP results providing the initial estimate for the biasing potentials for the rest of the single-guest peptides (except for PPP', for which already available PPP results were used). A flooding timescale of τ_F =100,200 *ps* was used for the different stages of these 10 *ns* long runs. The resulting *one-dimensional* free energy maps formed the basis of the HT-REMD runs for enhanced equilibrium sampling. Prior to starting the production runs, we assessed the ergodicity of the hottest replica (the one at the highest temperature biased by the approximate free energy associated with the Ω at that temperature) by simulating it alone. It turned out that all possible cis/trans states were visited on a timescale of a few hundred picoseconds with the Ω's autocorrelation time being less than 2 ns. This translates into a reasonable number of independent samples over the subsequent 100 *ns* production runtime. We note here that our choice of the highest temperature in the temperature ladder, and the total number of replicas were also influenced by the peculiarities of our local computer setup. We used 20 replicas with their full biasing potentials with the same temperature distribution as used for the ABMD runs. Four more replicas were then added, all at $T = 300$ K: one with no biasing potential, and three with the ABMD generated biasing potential scaled down by a factor of 0.49, 0.76

and 0.9, respectively. The choice of temperatures, the scaling factors, and the ratio of temperature-varying versus Hamiltonian-varying replicas (*i.e.*, 20 versus 4) was to ensure a similar rate of exchange, which varied between 55 to 60%, between all neighboring replicas. We then ran 100 ns HT-REMD simulations for all the host-guest peptides, with coordinates of the equilibrium $T = 300$ K replica sampled every picosecond, so as to provide 100,000 equilibrium configurations.

IV. QUANTIFYING PPII CONTENT: FITTING TECHNIQUE AND PUCKERING

We use the (ϕ, ψ) dihedral angles (see Fig.S1 for their definition) to identify different regions[20] in the Ramachandran plot. According to this scheme, both PPII and PPI helices are found in the F region, with traditional boundaries defined by $-110^\circ < \phi < -40^\circ$ and $130^\circ < \psi < 180^\circ$. In order to include all possible fluctuations of the ϕ and ψ angles for both PPI and PPII conformations, we extended this region to cover $-110^\circ < \phi < -20^\circ$ and ($50^\circ < \psi < 180^\circ$ or $-180^\circ < \psi < -120^\circ$). The β region consists of two parts: ($-180^\circ < \phi < -110^\circ$, $50^\circ < \psi < 180^\circ$ or $-180^\circ < \psi < -120^\circ$) and ($160^\circ < \phi < 180^\circ$ and $120^\circ < \psi < 180^\circ$) regions. The α region is also divided into two parts: α_R defined by $-120^\circ < \psi < 50^\circ$ and $-160^\circ < \phi < -20^\circ$ and α_L defined by $-50^\circ < \psi < 110^\circ$ and $20^\circ < \phi < 160^\circ$. These regions are all clearly marked on the Ramachandran plots shown in Fig.1.

While this defines clear regions for the dihedrals of most residues, it turns out that for the non-proline guests, there is considerable overlap between the populations that fall into the F and β regions of the Ramachandran plots, as illustrated in Fig.2. In order to handle this situation, we model the distribution of the ϕ angle as consisting of the sum of three Gaussians, and identify the area of the Gaussian centered in the F regions as defining the F content of the residue (the technical details are provided in the Supporting Material).

For non-proline amino acids, the PPII content is simply given by its F population. However, both the cis and trans isomers of proline fall in the F region. Since the cis isomer is a signature of the PPI structure, one should not count these isomers if one is *only* interested in the PPII content of the peptide (however one should not forget that PPI structures are also a signature of the proline host peptide, that could become quite important with other solvents). The cis/trans pattern is determined by the ω torsion angles (see Fig.S1). We therefore classify a residue as PPII if its dihedrals fall into the F region, with ω being in trans configuration. For non-proline amino acids (which always have ω in trans), this definition based on dihedrals falling in the F region is unchanged.

Typical sample of the ϕ distributions, along with the fitted curves are shown in Fig.S3. The fitting was carried out as follows. First, a histogram for each distribution was constructed from all the data in the F/β regions using 40 bins, with a width of 5° in the $-200^\circ < \phi < -20^\circ$ region. The data was then fitted by the sum of three Gaussians using the Marquardt-Levenberg algorithm[21]. The resulting width, center, and weight of each Gaussian was used as the initial guess for the latent variables for the Expectation-Maximization (EM) algorithm[22], which makes use of all ϕ angle data in the F/β region directly. The resulting optimized values are therefore completely independent of the histogram construction. Since the EM algorithm is iterative, we repeated the calculations until the desired convergence was reached. Typically, only a few iterations were required. The area of the Gaussian centered in the F region, multiplied by the percentage of samples in the F/β region defines the F content of the residue.

Finally, in addition to the cis/trans conformation, a proline residue is also characterized by its pyrrolidine ring puckering. The latter may be described in terms of the endocyclic torsion angle χ^1 (defined by atoms N-C^α-C^β-C^γ, as shown in Fig.S1). Here χ^1 takes on values of approximately 30° and -21° for the down- and up-puckered states, respectively. In this study, we have identified the puckering patterns of the different conformers for completeness only, since this structural feature does not play any significant role in the definition of PPII, and also does not appear to correlate with either the F or PPII content.

V. PROLINE-GUEST CORRELATION

To gain a qualitative insight into the Pro-X correlation issue, consider Fig.S4 which shows the ϕ_X distribution of the guest amino acid for select PXP peptides, given that the preceding P_3 residue has its ϕ angle in the F (red solid line) or α (dashed blue line) region of the Ramachandran plot. Each conditional probability distribution has two peaks, one centered in the β and one in the F region. Comparing the red and blue distributions, it is clear that the peak associated with the F region (centered about -75^o) is more pronounced when P_3 is in the F-region. In other words, the red curve is always higher than the blue curve for that region. The peak heights are not the same for the different PXP peptides and, roughly speaking, they track the P_3-X correlations: the larger the difference in peak heights, the stronger the expected correlation. Thus, the peak height potentially defines a useful measure of this correlation. Figure S4 shows typical results ranging from PAP (characterized by a large difference in peak heights and a high experimental PPII content) to PVP (small difference in peak height and low experimental PPII content).

For illustrative purposes, this data can also be translated into free energy language, by converting the conditional probability for ϕ_X into a PMF, as shown in Fig.S5. Since only free energy differences are meaningful, we have shifted the resulting PMFs such that the minima associated with the β region coincide. This then defines Δf_F and Δf_α, as shown in the left panels of Fig.S5 for PAP, PGP and PIP. One can further define the free energy difference $\Delta\Delta f = \Delta f_F - \Delta f_\alpha$, which may be considered to be another representation of the P_3-X correlation. Note that guest data for Fig.S4 and the left panels of Fig.S5 are based on the state of P_3, *i.e.*, the state of X is conditioned on the state of P_3. It is of course possible to study the reverse condition, and ask how P_3 is conditioned on the state of X. The right panels of Fig.S5 illustrate such results: specifically, the PMFs (*i.e.*, Δf_F and Δf_β) as a function of the ψ angle of the P_3 residue. This leads to a definition of $\Delta\Delta f = \Delta f_F - \Delta f_\beta$, which may be interpreted as a measure of the influence of X on P_3.

While the correlation between P_3 and X amino acids can also be expressed in other ways (see for instance Ref.23 for an interesting approach), we ultimately found that the Odds Ratio statistic gave us the cleanest results. Most likely, this is because of the binary nature of the characteristics associated with the states of the $P_3 - X$ correlation, which in turn are well represented with the Odds Ratio analysis.

Finally, in terms of results, Table S1 provides for a sequence-based analysis of the PXP and PXP' peptides, while Table S2 presents a full population analysis for the same. Also, Fig.S6 shows two typical two-dimensional free energy plots, whose collective variables distinguish between trans and cis rich conformers. The plot shows that in the presence of a non-proline guest, there is a slight preference for trans configurations, so that the overall cis content of the peptide is reduced.

1. Moradi, M., V. Babin, C. Roland, T. Darden, and C. Sagui. 2009. Conformations and free energy landscapes of polyproline peptides. *Proc. Natl. Aca. Sci. USA*. 106:20746.
2. Geyer, C. J. 1991. Markov chain monte carlo maximum likelihood. *In* Computing Science and Statistics: The 23rd symposium on the interface. Interface Foundation, Fairfax. 156 – 163.
3. Babin, V., and C. Sagui. 2010. Conformational free energies of methyl-α-l-iduronic and methyl-β-d-glucronic acids in water. *J. Chem. Phys.* 132:104108.
4. Laghaei, R., N. Mousseau, and G. Wei. 2010. Effect of the disulfide bond on the monomeric structure of human amylin studied by combined hamiltonian and temperature replica exchange molecular dynamics simulations. *J. Phys. Chem. B.* 114:7071–7077.
5. Moradi, M., V. Babin, C. Roland, and C. Sagui. 2010. A classical molecular dynamics investigation of the free energy and structure of short polyproline conformers. *J. Chem. Phys.* 133:125104.
6. Frenkel, D., and B. Smit. 2002. Understanding Molecular Simulation. Computational Science Series, Academic Press.
7. Babin, V., C. Roland, and C. Sagui. 2008. Adaptively biased molecular dynamics for free energy calculations. *J. Chem. Phys.* 128:134101.
8. Silverman, B. W. 1986. Density Estimation for Statistics and Data Analysis. Monographs on statistics and applied probability, Chapman and Hall.
9. Lelievre, T., M. Rousset, and G. Stoltz. 2007. Computation of free energy profiles with parallel adaptive dynamics. *J. Chem. Phys.* 126:134111.
10. Bussi, G., A. Laio, and M. Parrinello. 2006. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.* 96:090601.
11. Case, D. A., T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvary, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. T. J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin, and P. A. Kollman. 2008. "AMBER 10". University of California, San Francisco.
12. Babin, V., C. Roland, T. A. Darden, and C. Sagui. 2006. The free energy landscape of small peptides as obtained from metadynamics with umbrella sampling corrections. *J. Chem. Phys.* 125:2049096.
13. Babin, V., V. Karpusenka, M. Moradi, C. Roland, and C. Sagui. 2009. Adaptively biased molecualr dynamics: An umbrella sampling method with a time-dependent potential. *Int. J. of Q. Chem.* 109:3666–3678.
14. Moradi, M., J.-G. Lee, V. Babin, C. Roland, and C. Sagui. 2010. Free energy and structure of polyproline peptides: an ab initio and classical molecualr dynamics investigation. *Int. J. Quantum Chem.* 110:2865 – 2879.
15. Doshi, U., and D. Hamelberg. 2009. Reoptimization of the amber force field parameters for peptide bond (omega) torsions using accelerated molecular dynamics. *J. Phys. Chem. B.* 113:16590 – 16595.
16. Jhon, J. S., and Y. K. Kang. 1999. Imide cis–trans isomerization of n-acetyl-n'-methylprolineamide and solvent effects. *J. Phys. Chem. A.* 103:5436–5439.
17. Kang, Y. K., and H. Y. Choi. 2004. Cistrans isomerization and puckering of proline residue. *Biophysical Chemistry.* 111:135–142.
18. Kang, Y. K., J. S. Jhon, and H. S. Park. 2006. Conformational preferences of proline oligopeptides. *J. Phys. Chem. B.* 110:17645–17655.

19. Kang, Y. K., and N. S. Kang. 2006. Conformational preferences of n-methoxycarbonyl proline dipeptide. *J. Phys. Chem. B.* 110:21338–21348.
20. Zimmerman, S. S., M. S. Pottle, G. Némethy, and H. A. Scheraga. 1977. Conformational analysis of the 20 naturally occurring amino acid residues using ecepp. *Macromolecules.* 10:1–9.
21. Levenberg, K. 1944. A method for the solution of certain non-linear problems in least squares. *Q. App. Math.* 2:164 – 168.
22. Dempster, A. P., N. M. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. Ser. B.* 39:1 – 38.
23. Cossio, P., F. Marinelli, A. Laio, and F. Pietrucci. 2010. Optimizing the performance of bias-exchange metadynamics: Folding a 48-residue lysm domain using a coarse-grained model. *J. Phys. Chem. B.* 114:3259 – 3265.

peptide	(a) number of t					(b) number of F					(c) secondary structure	
	0	1	2	3	>3	7	6	5	4	<4	first	second
PPP	10	29	27	22	12	44	51	5	0	0	FFFFFFα (40)	FFFFFFF (34)
PPP'	13	28	31	16	12	54	43	3	0	0	FFFFFFF (42)	FFFFFFα (33)
PQP	21	44	26	8	1	4	16	35	40	5	FF$\alpha\beta$FFα (33)	FF$\alpha\beta$FFF (18)
PDP	14	43	28	12	3	1	15	46	37	1	FF$\alpha\beta$FFα (31)	FF$\alpha\beta$FFF (28)
PGP	24	38	22	11	5	3	25	39	30	3	FF$\alpha\beta$FFα (12)	FFFNFFF (11)*
PAP	17	45	28	9	1	4	14	38	41	3	FF$\alpha\beta$FFα (35)	FF$\alpha\beta$FFF (23)
PAP'	19	41	25	11	4	6	18	40	24	12	FF$\alpha\beta$FFF (26)	FF$\alpha\beta$FFα (20)
PKP	15	42	28	15	0	4	21	36	34	5	FF$\alpha\beta$FFα (27)	FF$\alpha\beta$FFF (19)
PSP	20	48	24	7	1	4	16	38	38	4	FF$\alpha\beta$FFα (33)	FF$\alpha\beta$FFF (22)
PEP	20	43	24	10	3	3	13	39	41	4	FF$\alpha\beta$FFα (35)	FF$\alpha\beta$FFF (25)
PHP†	16	42	30	11	1	4	18	42	34	2	FF$\alpha\beta$FFα (29)	FF$\alpha\beta$FFF (24)
PFP	12	41	28	16	3	3	19	41	35	2	FF$\alpha\beta$FFα (31)	FF$\alpha\beta$FFF (21)
PCP	20	42	26	10	2	4	20	41	32	3	FF$\alpha\beta$FFα (27)	FF$\alpha\beta$FFF (20)
PNP	20	43	26	10	1	2	14	42	40	2	FF$\alpha\beta$FFα (35)	FF$\alpha\beta$FFF (25)
PRP	12	40	34	12	2	4	22	42	29	3	FF$\alpha\beta$FFα (24)	FF$\alpha\beta$FFF (21)
PMP	13	41	31	13	2	3	17	36	40	4	FF$\alpha\beta$FFα (34)	FF$\alpha\beta$FFF (17)
PLP	18	42	26	12	2	3	15	39	38	5	FF$\alpha\beta$FFα (32)	FF$\alpha\beta$FFF (21)
PHP‡	13	39	30	15	3	3	16	40	40	1	FF$\alpha\beta$FFα (35)	FF$\alpha\beta$FFF (23)
PTP	25	41	26	7	1	4	19	40	35	2	FF$\alpha\beta$FFα (31)	FF$\alpha\beta$FFF (20)
PWP	10	36	33	18	3	5	23	39	30	3	FF$\alpha\beta$FFα (26)	FF$\alpha\beta$FFF (16)
PIP	22	44	23	9	2	3	18	38	36	5	FF$\alpha\beta$FFα (30)	FF$\alpha\beta$FFF (19)
PVP'	21	43	26	8	2	4	19	48	28	1	FF$\alpha\beta$FFF (32)	FF$\alpha\beta$FFα (25)
PVP	20	43	27	9	1	2	14	37	41	6	FF$\alpha\beta$FFα (36)	FF$\alpha\beta$FFF (21)
PYP	9	39	32	16	4	4	19	38	35	4	FF$\alpha\beta$FFα (30)	FF$\alpha\beta$FFF (19)

TABLE S1: Sequence based analysis of PXP and PXP' peptides: (a) The populations of the structures based on the number of c isomers. (b) The populations of the structures based on the number of residues in F region (ending Gly-Tyr is ignored). (c) The two most probable patterns and their populations (as a percentage), considering the regions in the Ramachandran plot *i.e.*, F, α or β.

*Gly in this structure is not in α, β, or F region (denoted as N).

†His is assumed to be protonated.

‡His is assumed to be neutral.

peptide	F_tF	$F_t\beta$	F_cF	$F_c\beta$	α_tF	$\alpha_t\beta$	α_cF	$\alpha_c\beta$	other
PQP	9	9	3	3	7	36	3	29	1
PDP	4	14	1	8	5	49	0	16	3
PGP	7	4	1	2	9	11	2	10	55
PAP	5	7	2	2	11	33	3	35	1
PAP'	11	9	3	3	18	28	4	24	1
PKP	8	9	3	9	17	29	2	23	1
PSP	5	9	2	2	11	33	4	32	2
PEP	3	7	2	3	12	40	1	29	1
PHP*	5	12	3	3	9	28	5	34	1
PFP	5	9	5	7	14	37	3	22	0
PCP	5	10	3	3	14	37	5	21	2
PNP	3	13	1	4	8	43	1	24	3
PRP	4	13	5	11	6	31	3	25	1
PMP	6	11	4	6	18	30	2	23	0
PLP	3	12	2	5	10	40	3	24	1
PHP†	3	11	3	7	10	38	2	27	1
PTP	4	9	2	2	19	42	2	18	1
PWP	6	8	9	11	20	25	3	17	0
PIP	3	10	3	3	14	47	3	18	0
PVP'	4	12	2	3	15	41	2	21	0
PVP	2	10	2	3	14	43	2	25	0
PYP	2	11	5	9	20	26	2	22	0

TABLE S2: Populations (as a percentage) of different patterns of the P_3X sequence in PXP and PXP' peptides considering the c/t conformations of P_3, as well as the Zimmerman regions of both residues.
*His is assumed to be protonated.
†His is assumed to be neutral.

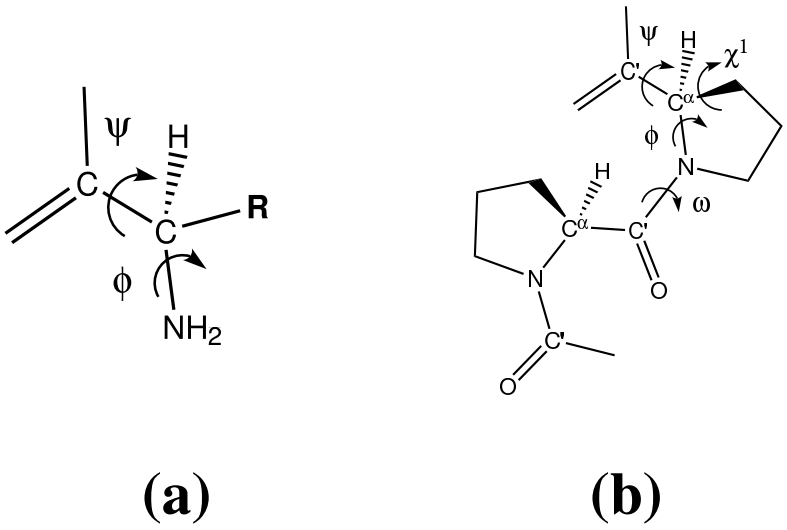

FIGURE S1: Schematic diagrams illustrating the dihedral angles: (a) ϕ and ψ in guest residues (b) ω, ϕ, ψ, and χ^1 in proline residues.

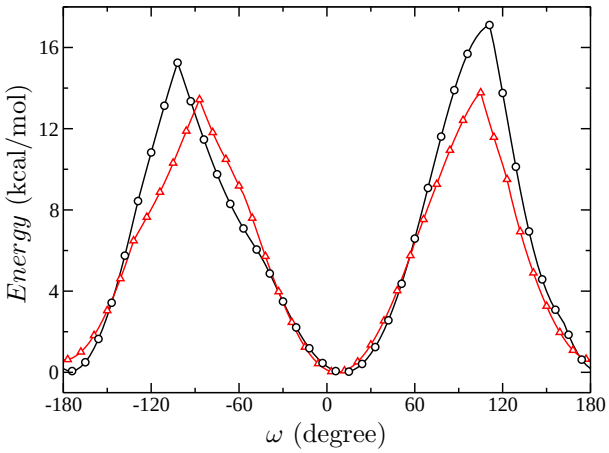

FIGURE S2: Ground state energy for the Ace-Pro-Pro-NMe molecule in gas phase as a function of the Pro-Pro ω angle: HF/6-31G* (circles) and AMBER ff99SB (triangles). The computations were done for 121 values of the ω angle (every 3°) with symbols shown every fourth point to avoid clutter.

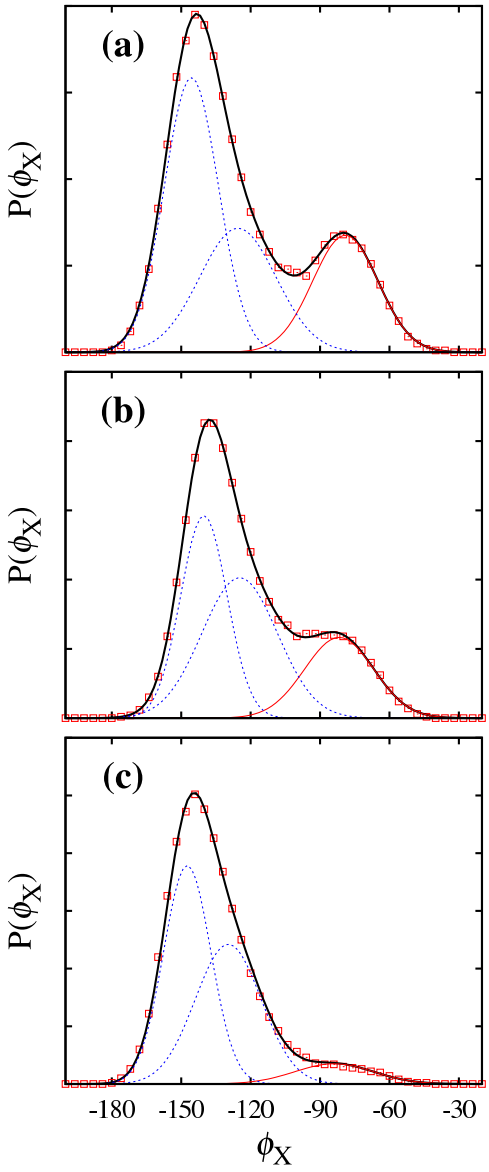

FIGURE S3: The probability distribution (area normalized to unity) of the ϕ angle for the guest residue, *i.e.*, ϕ_X, for (a) PQP, (b) PVP, and (c) PDP. The raw data (red squares) is obtained from histograms, and the fitted curve to this data is marked as a thick solid black line. This fitted curve may readily be decomposed into the sum of three Gaussians, as shown. The solid red line belongs to the F region, while the two dashed blue lines fall into the β region of the Ramachandran plots.

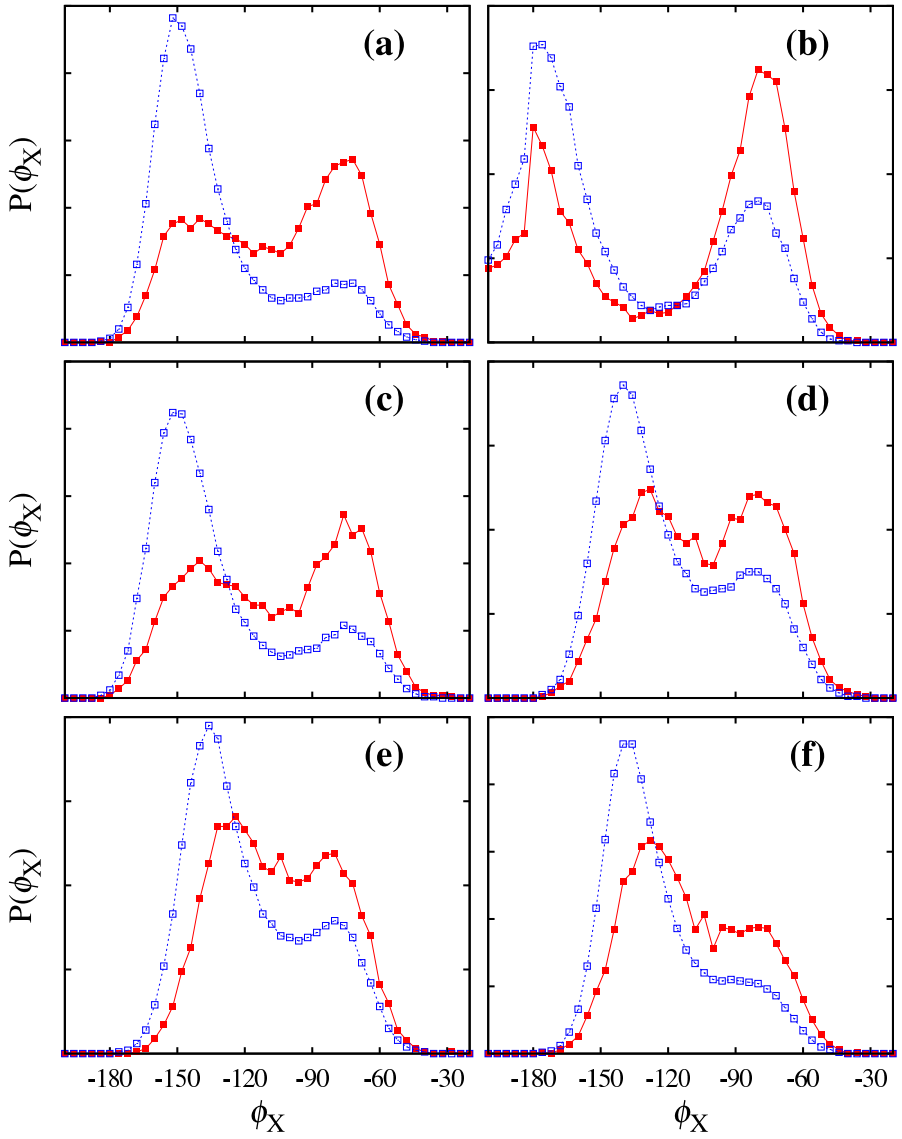

FIGURE S4: The probability distribution (area normalized to unity) of the ϕ_X angle of the guest residue for (a) PAP, (b) PHP (with a neutral His), (c) PSP, (d) PTP, (e) PIP, and (f) PVP. The red solid (blue dashed) lines are obtained from samples that had their P_3 residue in F (α) region of the Ramachandran plots.

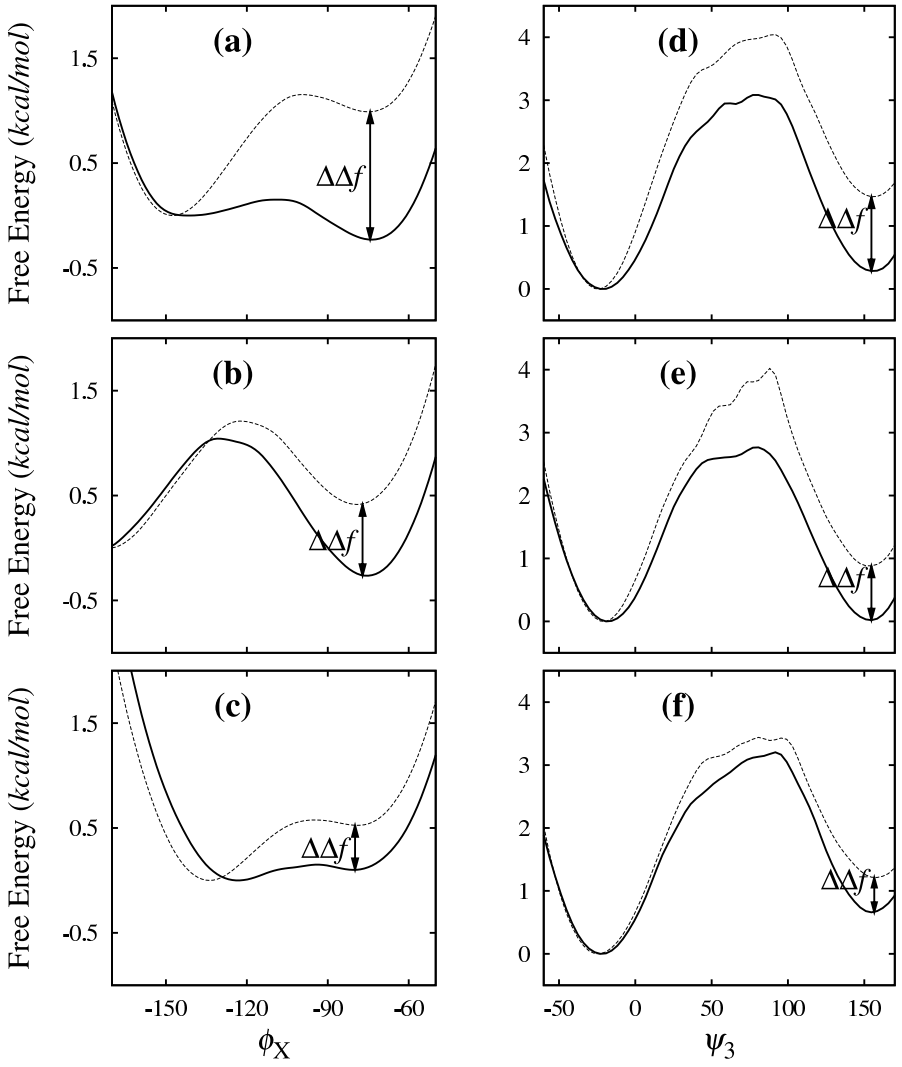

FIGURE S5: "Conditional" free energies obtained from the conditional probability distributions. Left panels: PMF in terms of the dihedral angle ϕ_X of the guest X for (a) PAP; (b) PGP; and (c) PIP. The solid (dashed) lines represent PMFs as obtained from samples with the P_3 residues in the F (α) region of the Ramachandran plots. The curves have been shifted so that minima associated with the β region coincide. $\Delta\Delta f = \Delta f_F - \Delta f_\alpha$ is indicated. Right panels: same as for left panels, except that the PMFs are plotted as a function of the ψ angle of P_3 (*i.e.,* ψ_3) and solid (dashed) lines denote samples for which the guest residue falls in the F (β) region of Ramachandran plot. These PMFs have been displaced so that the minima corresponding to the α region of P_3 coincide, and now $\Delta\Delta f = \Delta f_F - \Delta f_\beta$.

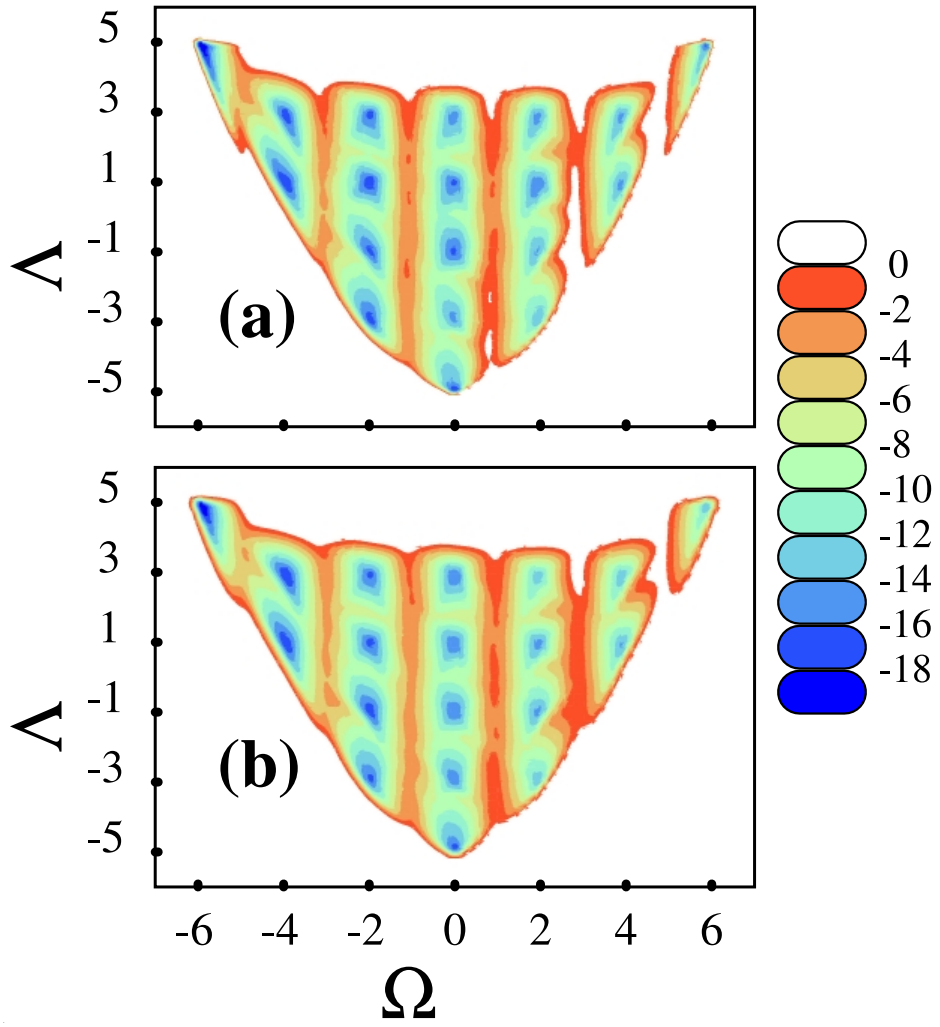

FIGURE S6: Two-dimensional (Ω,Λ) free energy landscapes for (a) PPP and (b) PAP peptides in implicit water (see Refs.(1,5) for a definition and discussion of these collective variables). Here, trans (cis) rich conformers are associated with Ω values less (greater) than zero. Also, the larger the numerical value of Λ, the more the structures conform to an ideal PPII or PPI helix. Comparing (a) and (b), one sees that in the presence of a guest amino acid, there is decrease in the overall cis content of the conformers signalled by a slightly greater preference for structures with Ω less than zero.