이학박사 학위논문

Site-directed Thermodynamic Analysis Method: Its Application for Protein Folding Studies and Assessment of Solvation Models

사이트 지정 열역학 분석 방법: 단백질 폴딩 연구 및 용매화 모델 평가

2023 년 08 월

서울대학교 대학원

화학부 물리화학 전공

조 명 근

Site-directed Thermodynamic Analysis Method: Its Application for Protein Folding Studies and Assessment of Solvation Models -

지도 교수 신 석 민

이 논문을 이학박사 학위논문으로 제출함

2023 년 08 월

서울대학교 대학원

화학부 물리화학 전공

조 명 근

조명근의 이학박사 학위논문을 인준함

2023 년 08 월

위 원 장	석 차 옥	(인)
부위원장	신 석 민	(인)
위 원	장 순 민	(인)
위 원	정 연 준	(인)
위 원	이 주 용	(인)

ABSTRACT

Site-directed Thermodynamic Analysis Method: Its Application for Protein Folding Studies and Assessment of Solvation Models

Myung Keun Cho

Department of Chemistry

The Graduate School

Seoul National University

Folding of a protein depends heavily on its aqueous environment. How solvation affects protein folding has been widely studied, but the extent to which folding stability is controlled by the solvation is unclear at the individual amino acid level. In this dissertation, we report the results of protein folding studies that employ the site-directed thermodynamic analysis method to assess the folding free energy components for each backbone and side chain of proteins. Thermodynamic results from tens of μs-length molecular dynamics simulations of the folding phenomenon of each of the representative β-sheet and α-helical proteins, human Pin WW domain protein and the villin headpiece subdomain are reported, respectively. We provide a quantitative measure of folding stability contributions from each of the critical sites of two model proteins, without introducing physical modifications to the system as in site-directed mutagenesis methods. Moreover, the resulting folding free energy of Pin WW was –4.9 kcal/mol, within the error bound of experimental reporting of –3.4 kcal/mol. By incorporating the decomposition method of solvation free energy and

gas-phase potential energy into single amino acid resolution, we determine the energetic consequence of basic molecular interactions such as hydrogen bonding and hydrophobic interaction that govern protein stability.

The application of the site-directed thermodynamic method is then extended to compare the influence of explicit and implicit solvation models on thermodynamic stability of the two model proteins. Thermodynamic analysis is often carried out by sampling a large number of atomistic conformations using molecular dynamics simulations that use either an explicit or implicit water model. However, it remains unclear to what extent thermodynamic results from different solvation models are reliable at the molecular level. Here, we quantify the influence of both solvation models on folding stability at single backbone and side chain resolution. Using simulation trajectories resultant from TIP3P solvent and the generalized Born/surface area solvent models, we assess the residue-specific folding free energy components of the two proteins described above. We find that the thermodynamic discrepancy from the generalized Born solvent mostly originates from positive side chains, followed by under-stabilized hydrophobic ones. In contrast, the backbone residue contributions in both proteins were comparable. Our study lays out the foundation for a detailed thermodynamic assessment of solvent models in the context of protein simulation.

Keyword: Molecular dynamics simulation, free energy decomposition, beta-sheet, generalized born solvent model, solvation free energy, 3D-RISM, intraprotein potential energy.

Student Number: 2016-20364

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. Introduction

Thermodynamic stability of biomolecules is governed by the driving forces of their folding such that the traces of these forces are rooted within the energetic components.[1,2] The stability of a native protein is described by the change in Gibbs free energy ΔG upon folding.[3,4] While the separation of this thermodynamic variable typically involves enthalpy and entropy, one can also decompose it into "favorable" and "unfavorable" terms, $\Delta G = \Delta f - T\Delta S_{conf}$.[5–7] Here, the favorable forces are captured by the change in the solvent-averaged effective energy ($\Delta f = \Delta E_u + \Delta G_{solv} < 0$),[8,9] consisting of the intra-protein energy (ΔE_u) and solvation (ΔG_{solv}) components, which are counteracted by unfavorable forces from the configurational entropy change ($\Delta S_{conf} < 0$).[10] The inequalities $\Delta f < 0$ and $\Delta S_{conf} < 0$ are expected to hold generally in protein folding since the folding must be an energetically downhill process (i.e., $\Delta f < 0$) to resolve Levinthal's paradox[9] and since protein structures in the unfolded state are more disordered than those in the folded state (hence, $\Delta S_{conf} < 0$). A representative contributor to Δf is the net effect of forming a hydrogen-bond (H-bond), comprising the gain in the intra-protein potential energy ($\Delta E_u < 0$) and the dehydration penalty ($\Delta G_{solv} > 0$); van der Waals contacts ($\Delta E_u < 0$); and the solvent-induced interaction ($\Delta G_{solv} < 0$). It is therefore essential to investigate thermodynamic stability using Δf for elucidating the strength and nature of stabilizing interactions underlying protein folding.

The position of amino acid in the primary sequence is a well-known determinant of thermodynamic stability of protein folding, let alone the identity of the amino acid residue. Site-directed mutagenesis methods have now become standard tools to investigate impacts of individual amino acid residues on protein

stability. The amide-to-ester or amide-to-olefin mutagenesis has been typically employed to understand the role of site-specific backbones, and the alanine mutagenesis to analyze the effect of side-chains.[11,12] However, these mutation-based methods inevitably introduce physical modifications to systems, which may induce unexpected effects and complicate the interpretation of the energetic consequences.[12–15] For example, physical modifications caused by mutations may propagate intricate structural perturbations to various extents in mutants, which make a systematic comparison between critical amino acid residues difficult. Moreover, distinct conformational ensembles could be constructed depending both on the number of mutants and experimental methods, which undermines the reliability of a quantitative comparison of thermodynamic properties for every residue. In particular, the unfolded-state ensembles of the same mutants have been reported to differ by ~50% structurally between the chemical and the thermal denaturation methods.[16]

Recently, a computational method has been developed that partitions thermodynamic functions of biomolecules into contributions from constituent atoms.[17,18] This allows us to investigate the thermodynamic stability at the individual residue level without introducing any physical perturbations to the system. In **Chapter 2**, we apply this method to the folding of the human Pin WW domain and villin headpiece subdomain (HP-36) proteins. The WW domain is the shortest (~33 residue long) naturally occurring β-sheet protein identified to date, and the HP-36 is a 36-residue helical protein.[19,20] They have served as exceptional models for protein folding studies both experimentally[11–13,19,21–24] and theoretically[25–33] due to its brevity and the folding time of ~100 μs. We performed all-atom molecular dynamics (MD) simulations of the WW domain and the HP-36 with an explicit-water treatment on both the folded- and unfolded-states. Then, the intra-protein energy and the solvation

energy were analyzed to acquire the site-resolved effective energy change upon folding for every backbone and side-chain. The solvation free energy is obtained by solving the integral-equation theory using 3-dimensional reference interaction site models (3D-RISM).[34] By combining structural and thermodynamic analyses, we investigate how the site-specific stability is determined from underlying main-chain and side-chain interactions in a protein. Thereby, we would like to provide insights into fundamental structural elements that stabilize the native structures of the β-sheet and the α-helix proteins.

In **Chapter 3**, our comparison study of influence of solvation models on site-resolved folding stability is introduced. Implicit solvent model is a practical alternative to explicit water in atomistic protein simulation. Recognized as a prime implicit solvent model for protein simulation, the generalized Born/surface area (GBSA) solvent speeds up the protein simulation by up to 100-fold compared to explicit solvents and facilitates exploration of the conformational ensemble.[35–38] By resolving total G_{solv} into polar and nonpolar terms, one facilitates the computation of the polar term using the GB approach that provides an analytical approximation of the Poisson-Boltzmann equation. The nonpolar term is estimated by the SA dependent approach. Despite its wide use, it remains unclear how accurately the GBSA solvent model reproduces thermodynamic results of protein folding. In particular, it is important to understand to what extent GBSA solvent artifacts influence the protein conformational sampling and to quantify its underlying consequences on the folding stability of a given protein. This is because obtaining reliable results of protein stability requires the proper representations of both folded and unfolded state conformations.[16,39]

Reproducing the structural characteristics and protein folding landscape from a nascent peptide chain has been a standard practice in evaluating the reliability of implicit solvent simulation. Recent efforts into assessing several GBSA solvent models showed that key attributes of native proteins such as the root-mean-square difference (RMSD)[36], α-helix[40,41], and β-sheet populations[42,43] are reproduced with sufficient accuracy. In contrast, a systematic bias in non-native characters such as salt-bridge has been reported in using GBSA solvents due to overstabilized electrostatic[41,44,45] and nonpolar interactions.[46–48] Such dichotomy between native and non-native characters in using implicit solvent models can be well-captured by the protein stability, as defined in ref. 49 by Lazaridis and Karplus, whose thermodynamic function takes small and large values for the folded and unfolded states, respectively. This thermodynamic function, called the solvent-mediated effective energy f, is obtained from a sum of the gas-phase potential energy E_u and the solvation free energy G_{solv}. A reliable descriptor of protein stability, the effective energy f primarily responds to the native structural character of protein and not to nonnative ones.[7]

Evaluating the influence of GBSA solvent artifacts on the folding stability is important in thermodynamic assessment of the solvent model. It is then useful to determine the folding stability at the single amino acid resolution to quantify how sensitive each residue is to the solvent model. To this end, Chong and Ham spearheaded the development of a computational method, called the site-directed thermodynamic analysis method, that provides an exact decomposition of a thermodynamic function into contributions from constituent amino acid residues.[17,50] The application of this method in partitioning solvation free energy G_{solv} and effective energy f allowed a systematic comparison among individual residues

quantitatively, elucidating the key role of certain residues in dictating the protein solubility[18,51] and stability[7,52]. This site-specific effective energy f can be separated into either corresponding residues[53], or corresponding main and side chains.[54] From the determination of site-specific folding stability contributions, the structural origins of the site-specific stability from critical residues can be identified. For example, the folding stability contributions from the entire backbones and side chains of WW domain was found to be comparable to each other.[54]

Here, we propose a study that assesses the structural and thermodynamic differences between explicit and implicit solvent simulation results of two model proteins. We adopted widely used explicit and implicit solvation models, TIP3P and GBSA-OBC(II), respectively.[55,56] The proteins of interest are representative β-sheet and α-helix proteins, WW domain and HP-36, respectively. These short proteins have often been the subject of protein folding simulations in both explicit solvent[25,54,57–63] and implicit solvent[40,46,64–67]. Then, we identify the critical residues that lead to the differences in solvation models using a recently modified site-directed thermodynamic analysis method that resolves the solvation free energy G_{solv} and the gas-phase potential energy E_{u} into contributions from the main and side chains.[54] This allows us to evaluate residue-level folding stability $\Delta f = f\,(\mathrm{folded}) - f\,(\mathrm{unfolded})$ in protein conformations from each of explicit and implicit solvent simulations. To the best of our knowledge, thermodynamic comparison between explicit and implicit solvent simulations at the individual residue level has not been carried out. Thus, we would like to understand the scope of limitations in employing the GBSA simulation to capture the folding stability behavior.

Chapter 2. Site-Specific Backbone and Side-Chain Contributions to Thermodynamic Stabilizing Forces

It has been challenging to determine how hydrogen bonds and hydrophobic contacts contribute to protein stability at single amino acid resolution. Here, site-specific thermodynamic stability was quantified at the molecular level to extend our understanding of the stabilizing forces in protein folding. A decomposition of the thermodynamic properties into contributions from main- and side-chain groups enabled us to identify the key residues in the secondary structure and hydrophobic core formation, without introducing physical modifications to the system as in site-directed mutagenesis methods. By relating the structural and thermodynamic changes upon folding for each residue, we find that the simultaneous formation of the backbone hydrogen bonds and side-chain contacts cooperatively stabilize the WW domain protein, as shown in **Figure 2.1**

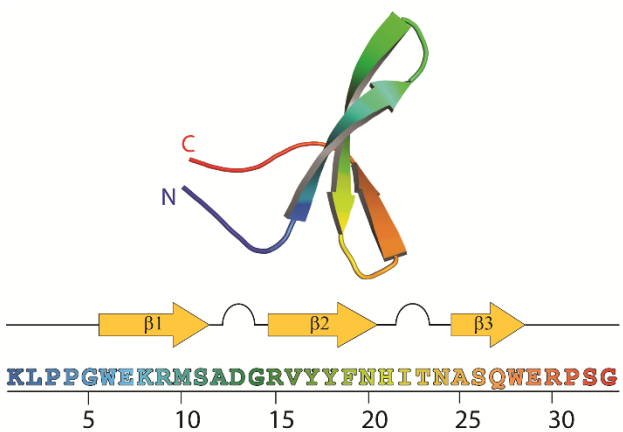


Figure 2.1. Pin WW structure and sequence. The structural cartoon of Pin WW is colored with respect to the corresponding amino acid residue sequence. The secondary structure propensity of the residue is indicated in the cartoon figure below the structure figure. The yellow arrows represent β-strands, curved segments represent the loops, and the rest are the terminal regions.

2.1. Methods

2.1.1. Folded-State Simulations.

The initial structure of Pin WW (PDB ID: 2F21),[19] determined from X-ray crystallography, was taken and modified as the sequence in **Figure 2.1**. The protonation states were designated as appropriate for the physiological pH. The AMBER20 package [68] was used to perform MD simulations with the ff99SB-ILDN protein force field[69] and TIP3P water model.[55] The system consisted of a protein solvated in 5957 water molecules and two chloride ions in a cubic periodic box. The steepest descent and the conjugate gradient minimization algorithms were each applied for 500 steps under 500 kcal/(mol Å^2), followed by another round of minimization of 1000 steps and 1500 steps without harmonic constraints, respectively. 20 ps of *NVT* ensemble equilibration was performed to raise the temperature gradually from T = 1 K to 300 K. *NPT* ensemble equilibration was then carried out for 200 ps at T = 300 K and P = 1 bar with a 2 fs time step. 1 μs production MD was performed at T = 300K and P = 1 bar. The procedure was repeated to produce six independent trajectories in total with a different initial velocity. The temperature was controlled by Berendsen thermostat with the coupling constant of 1.0 ps and the pressure by Berendsen barostat with the coupling constant of 2.0 ps.[70] The PME method was invoked in treating long-range electrostatic interactions, and the remaining nonbonded interactions were cut off at 10 Å. The hydrogen atoms were treated with the SHAKE algorithm.[71]

2.1.2. Unfolded-State Simulations.

The unfolded-state simulations were conducted using the configuration from *NPT* equilibration process in the folded-state. The system was first heated to 600 K under *NVT* ensemble. The simulated annealing simulation was performed with

a gradual decrease of temperature by 50 K for every 1 ns *NVT* ensemble simulation to reach the final temperature T = 300 K.[72] After an *NPT* ensemble equilibration simulation for 5 ns at T = 300 K and P = 1 bar, a 2 μs production run was conducted. Eight separate trajectories were obtained with random initial velocities. Structural analysis and thermodynamic calculations were performed using the last 1 μs trajectories.

2.1.3. Structural Analysis.

From each of the folded- and unfolded-state trajectories of 1 μs length, 20,000 protein conformations were extracted with a 50 ps time interval. Structural analyses of trajectories were carried out using CPPTRAJ.[73] Using the DSSP algorithm,[74] the anti-parallel β-strands region was defined by residues ^{6}WEKRM^{10}, ^{16}VYYFN^{20}, ^{26}SQ^{27}. The residues that form hydrophobic cluster 1 (HC1) are L2, P3, W6, Y18, and P31; that of hydrophobic cluster 2 (HC2) are R9, Y17 and F19.[19] The *k*-means clustering algorithm was used to find the representative protein structure for each trajectory. The geometrical criteria for a hydrogen bond (H-bond) formation were set to a separation by at most 3.5 Å between a hydrogen acceptor (e.g. C=O group) and a donor (e.g. N−H group) and at least 135° angle cutoff. The H-bonds were characterized into those formed between main-chains, between side-chains, and between a main-chain and a side-chain. The number of H-bonds was then separated into per amino acid residue according to participating donor and acceptor groups. To evaluate the extent of hydrophobic association of the protein for each residue, the side-chain contact analyses were also carried out. The side-chain contacts were defined with a distance cutoff of 5.4 Å between a side-chain group and the corresponding site at least three residues apart.[75] Both Cα for the main-chain and the farthest side-chain carbon atom from the peptide for the side-chain were selected

as the specific sites of interest.

2.1.4. Thermodynamic Analysis.

The Gibbs free energy of folding, $\Delta G = \Delta f - \Delta TS_{\mathrm{conf}}$ ($\Delta X = X_{\mathrm{folded}} - X_{\mathrm{unfolded}}$), was computed using the simulated protein structures in the folded and unfolded states based on the computational method developed in ref. 76. We first computed the solvent-averaged effective energy $f(\boldsymbol{r}_{\mathrm{u}}) = E_{\mathrm{u}}(\boldsymbol{r}_{\mathrm{u}}) + G_{\mathrm{solv}}(\boldsymbol{r}_{\mathrm{u}})$ for each simulated protein structure $\boldsymbol{r}_{\mathrm{u}}$. The molecular mechanics force field was used for the computation of $E_{\mathrm{u}}(\boldsymbol{r}_{\mathrm{u}})$, whereas the three-dimensional reference interaction site model (3D-RISM) theory[77,78] was employed for obtaining $G_{\mathrm{solv}}(\boldsymbol{r}_{\mathrm{u}})$. The change Δf in the effective energy upon folding can be computed as the difference of average $f(\boldsymbol{r}_{\mathrm{u}})$ values for the folded and unfolded states. We then constructed the probability distribution $W(f)$ of $f(\boldsymbol{r}_{\mathrm{u}})$ values sampled in each of the folded and unfolded states. If the distribution $W(f)$ is well approximated by Gaussian, which will be verified below, the configurational entropy is given by $TS_{\mathrm{conf}} = (1/2k_BT)\overline{\delta f^2}$ in terms of the variance $\overline{\delta f^2}$.[79,80]

2.1.5. Site-directed Thermodynamic Analysis

An exact partitioning of the solvation free energy G_{solv} into contributions from constituent atoms (labeled by α), $G_{\mathrm{solv}} = \sum_\alpha G_{\mathrm{solv},\alpha}$, was derived in ref. 17 based on the Kirkwood charging formula. By an appropriate grouping of those atomic contributions, one obtains the backbone, side-chain and individual residue contributions to G_{solv}. In the present work, the backbone group is defined as -C-O-NH-C_α-, and the rest are treated as the side-chain group. A corresponding partitioning of the intra-protein potential energy E_{u} into groups (labeled by i and j) can be obtained as follows:

$$E_{\mathrm{u},i} = E_i^{\mathrm{intra}} + \frac{1}{2}\sum_{j\neq i}^{N}\left(E_{\mathrm{bond},ij}^{\mathrm{inter}} + E_{\mathrm{nonbond},ij}^{\mathrm{inter}}\right) + \sum_{j\neq i}^{N}\sum_{\alpha=1}^{2}\frac{\alpha}{3}E_{\mathrm{ang},ij}^{\mathrm{inter},\alpha} + \sum_{j\neq i}^{N}\sum_{\alpha=1}^{3}\frac{\alpha}{4}E_{\mathrm{dih},ij}^{\mathrm{inter},\alpha}$$

eq. 1

Here, E_i^{intra} denotes the potential energy within the *i*-th group, and $E_{\mathrm{bond},ij}^{\mathrm{inter}}$ and $E_{\mathrm{nonbond},ij}^{\mathrm{inter}}$ refer to the bonded and nonbonded interactions between *i*-th and *j*-th groups, respectively. The angle and dihedral angles terms, $E_{\mathrm{ang},ij}^{\mathrm{inter},\alpha}$ and $E_{\mathrm{dih},ij}^{\mathrm{inter},\alpha}$, are treated such that the associated energies are divided according the number of atoms α that belong to the *i*-th group. These partitioning methods allow us to obtain the site-specific E_{u}, G_{solv}, and $f = E_{\mathrm{u}} + G_{\mathrm{solv}}$.

2.1.6. Error Analysis

The standard error (SE) of the mean for the folded state was obtained based on average values of six independent folded-state trajectories: the standard deviation (SD) of the average values was first computed, followed by the division by square root of the number of trajectories (n_{traj})

$$\mathbf{SE} = \frac{\mathbf{SD}}{\sqrt{\boldsymbol{n}_{\mathbf{traj}}}} \qquad \textbf{eq. 2}$$

The standard error for the unfolded was estimated from the eight independent trajectory average values. After obtaining the standard errors for both the folded- and unfolded-states, the standard error of the difference, $\Delta X = X_{\mathrm{folded}} - X_{\mathrm{unfolded}}$, was evaluated as follows:[81]

$$\mathbf{SE}(\Delta \boldsymbol{X}) = \sqrt{\mathbf{SE}(\boldsymbol{X}_{\mathbf{folded}})^2 + \mathbf{SE}(\boldsymbol{X}_{\mathbf{unfolded}})^2} \qquad \textbf{eq. 3}$$

2.2. Results and Discussions

2.2.1. Structural Differences between the Folded and Unfolded States of WW Domain

We investigated how thermodynamic energy of a protein shapes its structure, focusing on the WW domain, one of the shortest β-sheet proteins. We carried out both folded- and unfolded-state MD simulations of the Pin WW domain under physiological conditions (T = 300 K and P = 1 bar). The folded-state simulations were initiated from an X-ray structure, whereas the unfolded-state simulations were preceded by heat-denaturation. We characterized various native structural features, including RMSD, secondary structure content, and hydrophobic cluster contact properties, summarized in **Table 2.1**. The protein conformations in the folded-state remain stable over the simulation timescales with the average Cα RMSD of 0.9 Å to the native structure, whereas those in the unfolded-state fluctuate heavily with the value of 7.2 Å. The secondary structure contents show that the folded-state retained significant β-sheet conformations, while coil/turn conformations are dominant in the unfolded-state (see **Figure 2.2** for per-residue analyses of the secondary structures). The native structure of Pin WW domain is packed with two hydrophobic clusters (HC1 and HC2) that protrude from either side normal to the β-sheet surface (**Figure 2.3**). Most of the HC1 and HC2 contacts are retained in the folded-state simulations, while they are lost in the unfolded-state simulations (**Table 2.1**) The native contacts fraction, Q, serves as a validation criterion of the unbiased MD simulation that discriminates the folded- and the unfolded-states from misfolded structures, as displayed in **Table S2.1**. The average Q_f and Q_u cutoff values were chosen to be 0.92 and 0.20 for WW domain, respectively. This corresponded to the defined criterion for other folding studies as well.[82,83] In addition, the average Q_f and Q_u cutoff values

for HP-36 were reported to be 0.75 and 0.18, which are less conservative for the folded state than the previous study with 0.89 and 0.20, respectively. Thus, there are distinctive differences in protein conformations if we focus on the native contacts.

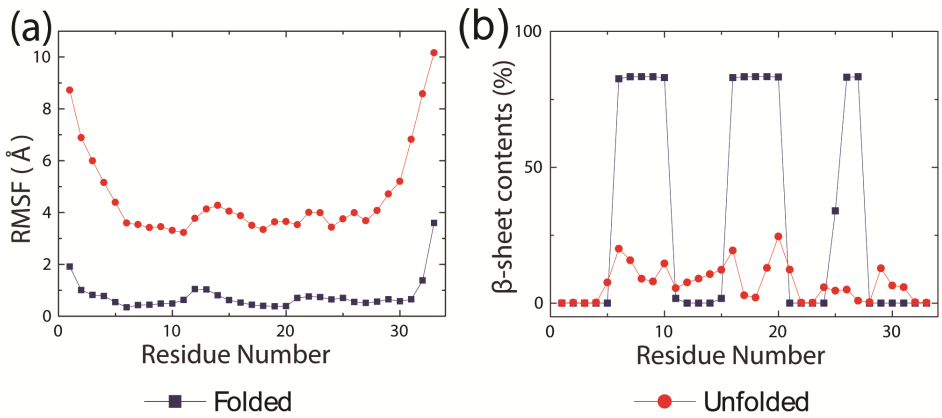


Figure 2.2. Structural characteristics of Pin WW per amino acid residue averaged from the folded and unfolded states MD simulations at 300 K. (A) Root-mean-squared-fluctuation per amino acid residue number is shown. (B) β-sheet contents per amino acid residue as obtained from DSSP analysis is indicated.

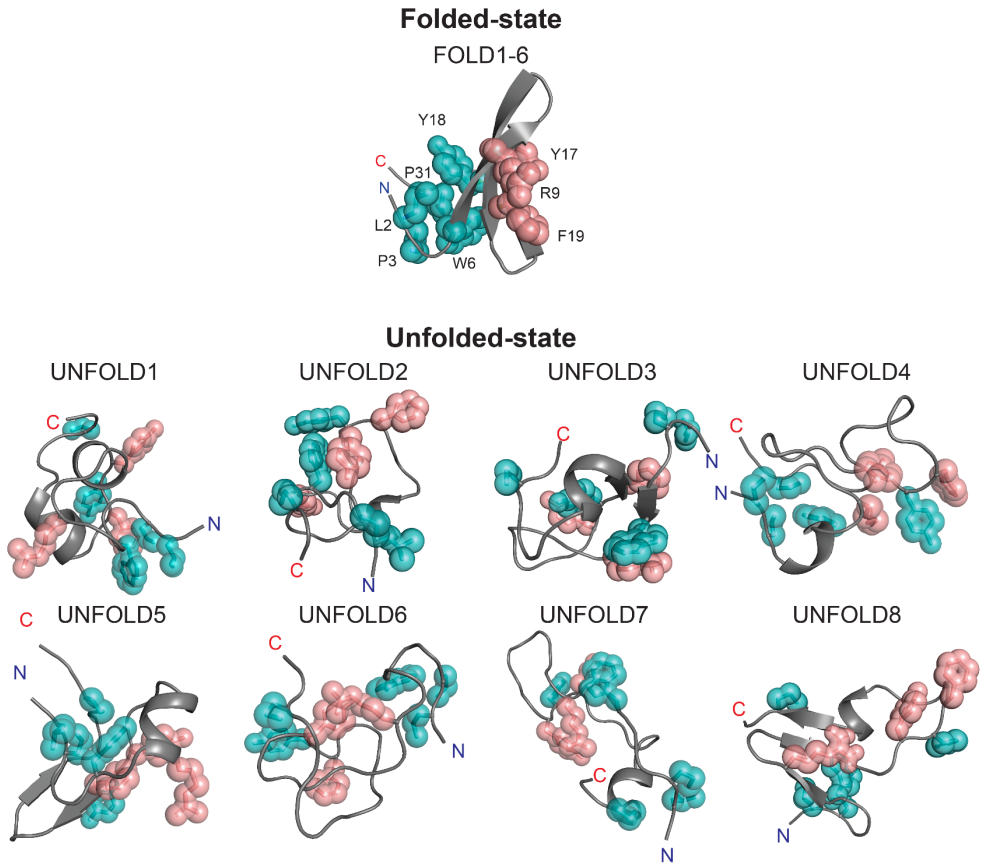


Figure 2.3. The MD simulation snapshot for the folded and the unfolded trajectories. Representative folded-state conformations selected using a k-means clustering method are from the trajectories FOLD1 and FOLD6. Representative unfolded-state conformations are from the trajectories UNFOLD1-UNFOLD8.

Table 2.1. Native structural characteristics in the folded and unfolded states of Pin WW

		Secondary Structure Contents (%) [b]			Heavy atom contacts (%) [c]	
	C_α RMSD [a]	β-1	β-2	β-3	HC 1	HC 2
X-ray Structure PDB ID: 2F21	0	100	100	100	100	100
folded-state trajectories [d]						
FOLD1	0.90 ± 0.20	99.7 ± 2.5	99.9 ± 1.6	99.9 ± 2.6	83.3 ± 6.9	86.5 ± 10.6
FOLD2	0.90 ± 0.20	99.7 ± 2.6	99.9 ± 1.7	99.9 ± 2.6	83.0 ± 7.0	86.3 ± 10.2
FOLD3	0.89 ± 0.20	99.7 ± 2.4	99.9 ± 1.6	99.9 ± 2.6	83.0 ± 7.0	87.6 ± 8.5
FOLD4	0.90 ± 0.20	99.8 ± 2.2	99.9 ± 1.3	99.8 ± 3.3	83.2 ± 7.2	87.3 ± 10.8
FOLD5	0.91 ± 0.22	99.7 ± 2.6	99.9 ± 1.8	99.9 ± 2.7	83.3 ± 7.1	86.5 ± 13.9
FOLD6	0.91 ± 0.20	99.8 ± 2.4	99.9 ± 1.4	99.9 ± 2.4	83.0 ± 7.2	86.3 ± 7.7
Average [e]	0.90 ± 0.01	99.7 ± 0.1	99.9 ± 0.1	99.9 ± 0.1	83.1 ± 0.1	86.8 ± 0.2
unfolded-state trajectories [d]						
UNFOLD1	7.98 ± 0.07	0.0 ± 1.0	0.0 ± 0.0	0.0 ± 0.0	4.2 ± 2.5	0.0 ± 0.0
UNFOLD2	6.74 ± 0.19	0.0 ± 0.3	34.3 ± 14.0	0.0 ± 1.6	1.9 ± 1.1	0.0 ± 0.0
UNFOLD3	6.97 ± 0.56	13.2 ± 17.8	24.2 ± 17.6	0.1 ± 2.4	0.9 ± 3.5	0.0 ± 0.0
UNFOLD4	6.40 ± 1.35	7.6 ± 11.2	12.0 ± 16.7	2.4 ± 10.8	16.0 ± 14.0	6.7 ± 11.1
UNFOLD5	5.85 ± 0.46	14.0 ± 9.7	0.9 ± 4.3	1.0 ± 6.8	37.3 ± 7.9	0.3 ± 3.3
UNFOLD6	7.74 ± 0.79	0.7 ± 5.0	0.6 ± 3.4	7.0 ± 17.3	5.6 ± 6.0	42.9 ± 21.3
UNFOLD7	8.03 ± 0.59	0.3 ± 2.5	0.3 ± 2.5	0.0 ± 0.4	4.5 ± 3.6	20.6 ± 22.2
UNFOLD8	7.88 ± 0.43	56.5 ± 26.0	0.0 ± 0.6	0.0 ± 1.1	5.5 ± 2.3	1.3 ± 3.8
Average [e]	7.20 ± 0.27	11.5 ± 6.3	7.8 ± 4.0	1.3 ± 0.8	9.5 ± 4.0	9.0 ± 5.1

[a] Root-mean-square deviations (Å) for C_α atoms; [b] Average population (%) of the β-strand formations in β-1 (W6-M10), β-2 (V16-N20) and β-3 (S26-Q27); [c] Average population (%) of side-chain heavy atom contacts in hydrophobic cluster 1 (HC 1; L2, P3, W6, Y18, and P31) and hydrophobic cluster 2 (HC 2; R9, Y17, and F19); [d] Average ± standard deviation; [e] Average ± standard error.

However, taking into account non-native contacts as well raises a different view. **Table 2.2** compares the number of H-bonds and side-chain contacts in the folded- and unfolded-state simulations (see **Table S2.2** for the per-residue analyses). The average number of H-bonds is 22.2 for the folded state, which is only slightly larger than the one (20.1) for the unfolded state. Similarly, the average number of side-chain contacts does not differ significantly between the folded- (46.0) and unfolded-state (36.6). In **Figure S2.1**, the backbone H-bonds formation in β-sheet regions of the folded state is contrasted with their nearly uniform distribution throughout the backbones in the unfolded state. One of the main issues we address in the following is why the folded state is more stabilized than the unfolded state, even though the number of intra-protein contacts is comparable.

Table 2.2: Non-native structural characteristics in the folded and unfolded states of Pin WW

	Number of H-bonds [a]				Number of side-chain contacts [b]		
	Total	MC−MC	MC−SC	SC−SC	Total	MC-SC	SC-SC
folded-state trajectories [c]							
FOLD1	22.3 ± 2.5	10.7 ± 1.1	5.2 ± 1.5	6.3 ± 1.7	46.2 ± 3.0	22.9 ± 1.8	23.2 ± 1.8
FOLD2	22.4 ± 2.4	10.8 ± 1.1	5.2 ± 1.5	6.4 ± 1.6	45.9 ± 3.0	22.8 ± 1.8	23.1 ± 1.9
FOLD3	22.3 ± 2.4	10.7 ± 1.1	5.2 ± 1.5	6.4 ± 1.6	46.0 ± 2.9	22.9 ± 1.8	23.1 ± 1.8
FOLD4	22.2 ± 2.4	10.7 ± 1.1	5.1 ± 1.4	6.4 ± 1.6	46.3 ± 3.0	22.9 ± 1.8	23.4 ± 1.9
FOLD5	22.1 ± 2.5	10.7 ± 1.1	5.1 ± 1.5	6.2 ± 1.7	46.1 ± 3.0	23.0 ± 1.8	23.2 ± 1.9
FOLD6	22.0 ± 2.4	10.7 ± 1.1	5.1 ± 1.5	6.2 ± 1.6	45.5 ± 3.0	22.8 ± 1.8	22.7 ± 1.9
Average [d]	22.2 ± 0.1	10.7 ± 0.1	5.1 ± 0.1	6.3 ± 0.1	46.0 ± 0.7	22.9 ± 0.1	23.1 ± 0.1
unfolded-state trajectories [c]							
UNFOLD1	24.1 ± 3.1	8.8 ± 1.5	11.4 ± 1.9	4.0 ± 1.6	37.9 ± 2.7	21.6 ± 1.7	16.3 ± 1.7
UNFOLD2	21.6 ± 2.7	10.4 ± 1.7	8.1 ± 1.9	3.2 ± 1.8	44.0 ± 3.3	24.7 ± 2.0	19.3 ± 2.1
UNFOLD3	17.8 ± 3.2	7.9 ± 1.7	6.8 ± 2.0	3.3 ± 1.8	38.1 ± 3.9	19.1 ± 2.1	19.0 ± 2.7
UNFOLD4	19.1 ± 3.7	7.0 ± 1.9	7.9 ± 2.5	4.3 ± 1.6	34.5 ± 4.2	18.1 ± 2.6	16.4 ± 2.5
UNFOLD5	18.3 ± 2.7	7.1 ± 1.8	4.8 ± 2.2	6.4 ± 1.7	34.8 ± 3.9	16.7 ± 2.5	18.1 ± 2.1
UNFOLD6	17.8 ± 3.4	6.4 ± 1.7	8.2 ± 1.9	3.2 ± 1.9	36.1 ± 6.1	18.8 ± 3.2	17.4 ± 3.5
UNFOLD7	21.9 ± 3.2	8.0 ± 1.6	8.2 ± 2.6	5.8 ± 2.2	35.5 ± 4.5	18.9 ± 2.8	16.6 ± 2.4
UNFOLD8	20.0 ± 3.6	8.0 ± 3.1	6.3 ± 3.1	5.8 ± 2.1	31.6 ± 5.5	17.0 ± 3.0	14.6 ± 2.9
Average [d]	20.1 ± 0.8	7.9 ± 0.4	7.7 ± 0.6	4.5 ± 0.4	36.6 ± 0.7	19.4 ± 0.8	17.2 ± 0.5

[a] The number of intra-protein H-bonds in total, between main-chains (MC−MC), between a main-chain and a side-chain (MC−SC) and between side-chains (SC−SC); [b] The number intra-protein heavy-atom contacts involving side-chains in total, main-chain−side-chain contacts (MC−SC) and side-chain-side-chain contacts (SC−SC); [c] Average ± standard deviation; [d] Average ± standard error.

2.2.2. Folding Free Energy of Pin WW.

We computed the folding free energy $\Delta G = \Delta f - T\Delta S_{\text{conf}}$ of the WW domain based on the solvent-averaged effective energy $f(\boldsymbol{r}_{\text{u}}) = E_{\text{u}}(\boldsymbol{r}_{\text{u}}) + G_{\text{solv}}(\boldsymbol{r}_{\text{u}})$ evaluated for each protein conformation $\boldsymbol{r}_u$ sampled in the folded- and unfolded-state simulations (see Methods). The change in effective energy upon folding provides the thermodynamic driving force of folding ($\Delta f < 0$), and is the key descriptor of thermodynamic stability. We computed the average $f(\boldsymbol{r}_{\text{u}})$ values for all the individual trajectories (**Tables 2.3** and **S2.3**), from which we obtain $\Delta f = -29.2 \pm 2.0$ kcal/mol (average ± standard error). For the estimation of configurational entropy, we constructed the probability distribution $W(f)$ of the sampled $f(\boldsymbol{r}_{\text{u}})$ values (**Figure 2.4**). We find that $W(f)$ for both the folded and unfolded states is well-described by the Gaussian distribution as can be verified by the small skewness and excess kurtosis (**Table S2.4**) which approach zero when the distribution assumes a perfect Gaussian. As derived in ref. 76 (see also the Methods), the configurational entropy can be

estimated from the variance of the sampled $f(\boldsymbol{r}_u)$ values when $W(f)$ is Gaussian (the convergence of the configurational entropy estimation from our simulations is demonstrated in **Figure S2.2**, and the fact that the tail of $W(f)$ is also well captured by the Gaussian statistics is shown in **Figure S2.3**). From the estimated $T\Delta S_{conf}$ values for all the individual trajectories (**Tables 2.3** and **S2.3**), we obtain $-T\Delta S_{conf}$ = 24.3 ± 5.2 kcal/mol. The resulting folding free energy is ΔG = –4.9 kcal/mol with a standard error of 4.9 kcal/mol.

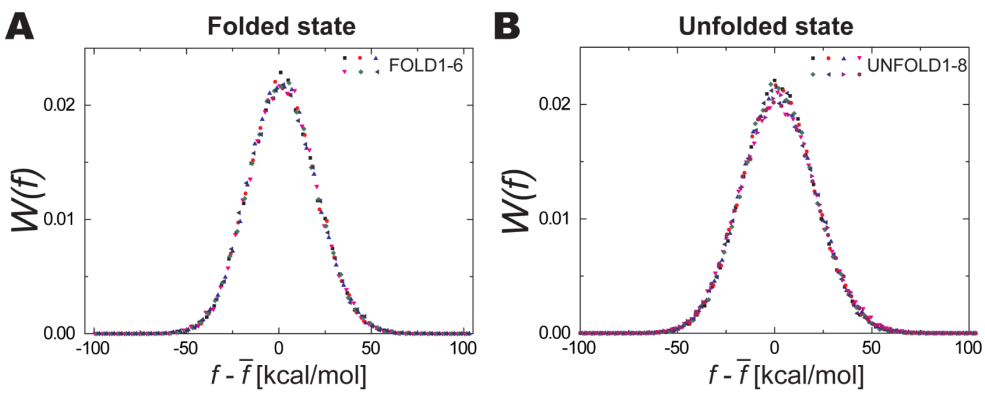


Figure 2.4. The distribution of the effective free energy. (A) Probability distribution function $W(f)$ of the effective energy f as a function of its deviation from the mean $\bar{f}$ for the six independent trajectories of the folded-state (FOLD1-FOLD6). (B) Corresponding result for the eight independent unfolded-state trajectories (UNFOLD1-8).

This value is in accord with the experimental values of –3.4 kcal/mol from the temperature-jump experiment on the identical sequence[19] and –2.2 kcal/mol from urea-denaturation on hYAP WW domain.[84] The large standard error of ΔG mainly stems from the configurational entropy term, $-T\Delta S_{conf}$. Specifically, we note that the magnitude of the standard error from $-TS_{conf}$ of the unfolded protein is only 1.7% of

the average value, but the error is significantly enlarged by the large cancellation between the folded- and unfolded-state contributions.

Table 2.2. Tabulated data of folding free energy

State	f [a]	$-TS_{conf}$ [b]	$G = f - TS_{conf}$ [c]
folded-state trajectories			
FOLD1	–303.2	–278.5	–581.7
FOLD2	–302.7	–274.1	–576.8
FOLD3	–303.3	–273.8	–577.1
FOLD4	–303.2	–277.8	–581.0
FOLD5	–302.9	–276.2	–579.1
FOLD6	–301.7	–282.0	–583.7
Average [d]	–302.8 ± 0.2	–277.1 ± 1.1	–579.9 ± 1.0
unfolded-state trajectories			
UNFOLD1	–283.3	–282.8	–566.1
UNFOLD2	–275.5	–289.3	–564.8
UNFOLD3	–274.0	–296.0	–570.0
UNFOLD4	–265.6	–323.9	–589.5
UNFOLD5	–272.3	–289.8	–562.1
UNFOLD6	–264.9	–296.2	–561.1
UNFOLD7	–276.3	–316.0	–592.3
UNFOLD8	–278.0	–316.6	–594.6
Average [d]	–273.7 ± 2.0	–301.3 ± 5.1	–575.0 ± 4.8
	Δf	$-T\Delta S_{conf}$	$\Delta G = \Delta f - T\Delta S_{conf}$
Difference [d]	–29.2 ± 2.0	24.3 ± 5.2	–4.9 ± 4.9

[a] Effective energy [kcal/mol]; [b] Configurational entropy multiplied by $-T$ [kcal/mol]; [c] Gibbs free energy [kcal/mol]; [d] Average ± standard error.

Here, some comments might be appropriate concerning our estimate of the configurational entropy (TS_{conf}). As is apparent from **Table 2.3**, we first computed TS_{conf} for individual trajectories, which were then averaged in obtaining the folding free energy. The convergence of TS_{conf} of individual trajectories was also confirmed (**Figures S2.2** and **S2.3**). On the other hand, another plausible approach would be to combine all the independent trajectories together (separately for the folded- and unfolded-states) and estimate TS_{conf} for such an ensemble of protein configurations. If individual trajectories were long enough to fully explore the configuration space,

both approaches would yield the same value of TS_{conf}. However, we found that this is not the case with our short unfolded-state simulations of 2 μs length. In this sense, our estimate of $-T\Delta S_{conf}$ may still be unreliable. We would like to examine this issue using much longer simulations[25] in our future study. We notice that this convergence issue of TS_{conf} does not affect our subsequent argument that is primarily concerned with the effective energy (f).

2.2.3. Site-specific Stabilizing Forces

To provide a detailed thermodynamic characterization of free energy contributions from the individual constituents of protein, we resolve the descriptor of folding stability Δf into individual residue contributions (**Figure 2.5A**) and further into the respective backbone and side-chain terms (**Figure 2.5C** and **D**; corresponding results for ΔE_u and ΔG_{solv} are shown in **Figure S2.4**, and numerical values with standard errors are provided in **Tables S2.5** and **S2.6**). The site-resolved contributions to Δf from each backbone and side-chain residue exhibit a wide range of values, which underscores the position-dependence in thermodynamic stability of the polypeptide. We find that the majority of β-sheet regions contribute favorably to folding, whereas the turn and terminal regions are mostly destabilizing (**Figure 2.5A** and **C**). Furthermore, the significance of the side-chain hydrophobic core formation in thermal stability is apparent from large favorable contributions arising from HC1 (colored dark cyan) and HC2 (dark pink) regions (**Figure 2.5D**).

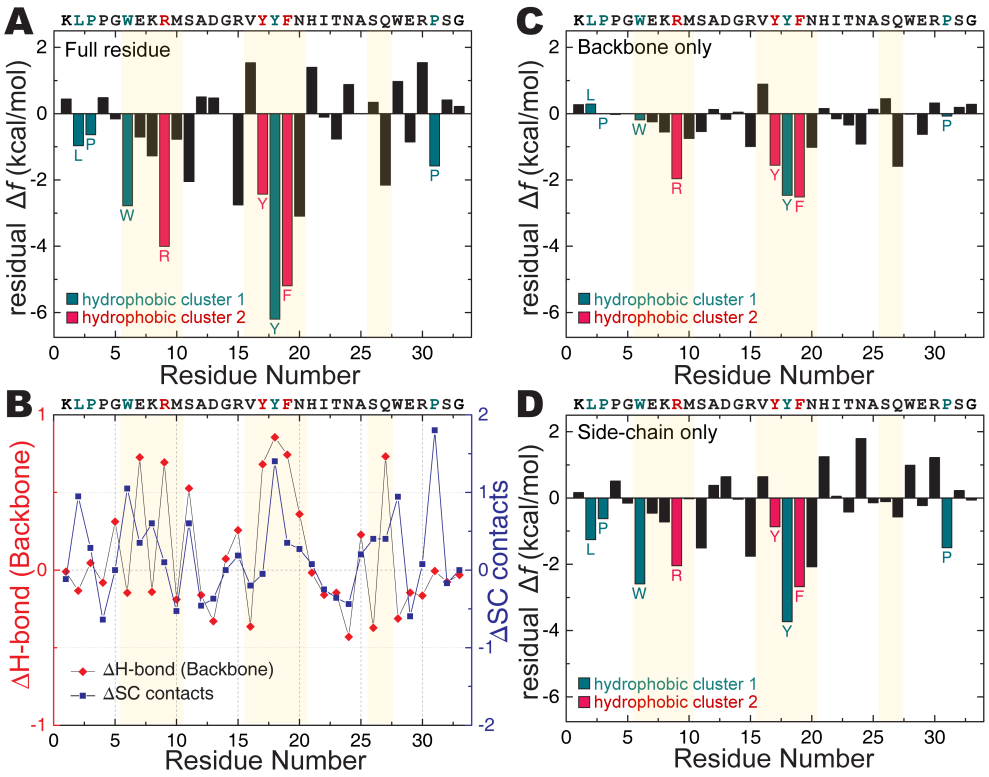


Figure 2.5. Thermodynamic and structural analyses for each backbone and side-chain of the WW domain. (A) The folding effective energy contributions $\Delta f = f$ (folded) – f (unfolded) for every constituent residue. (B) The number of H-bond between main-chains upon folding, indicated by red diamond symbol, corresponding to the scale on the left y-axis and the number of side-chain–side-chain contacts upon folding, indicated by blue square symbol, corresponding to the scale on the right y-axis are shown. (C, D) The site-directed thermodynamic analysis results of Δf are obtained for each backbone (C) and for each side-chain (D). The residues that comprise hydrophobic cluster 1 and hydrophobic cluster 2 are indicated by dark cyan and dark pink, respectively. The yellow strips demarcate the three β-strand regions.

The total backbone and side-chain contributions to Δf are found to be –13.6 and –15.7 kcal/mol, respectively. Thus, the backbone and side-chain contributions to the thermal stability of Pin WW domain are comparable (larger contribution from side-chains). To further connect how these contributions originate from underlying interactions, we show in **Figure 2.5B** the changes in backbone H-bonds (red diamonds) and side-chain contacts (blue squares) upon folding. It is seen that both

the backbone H-bonds and side-chain contacts are simultaneously formed upon folding, in particular, in the central β-2 sheet region. As we stated above, the total number of H-bonds and side-chain contacts are comparable between the folded and unfolded states. On the other hand, there is a significant thermodynamic difference (reflected in a large negative Δf) between these two states. This indicates that, whereas H-bonds and side-chain contacts are formed independently in the unfolded state, their simultaneous formation cooperatively stabilizes the folded structure. This notion of an enhanced stability gained from the interplay between H-bonds and side-chain contacts is in accord with the experimental observations that a secondary structure is in general not stable by itself[l] and that H-bond strength is amplified when sequestered in more hydrophobic environment.[21,85–87]

Such cooperative nature of the stabilizing forces also accounts for the irrelevance of non-native contacts. For example, if one carefully examines Figure S4, residues such as R30 are found to exhibit large positive ΔE_u values, indicating that certain intra-protein contacts are stabilizing the unfolded-state. Indeed, in our simulations, the side chain of R30 forms a salt-bridge with the carboxyl group of the C-terminus (G33), and its fraction is higher in the unfolded-state than in the folded-state. This explains why ΔE_u of R30 is positive, i.e., the intra-protein energy E_u of R30 is lower in the unfolded-state than in the folded-state. Similarly, we confirmed that other residues that exhibit positive ΔE_u values in **Figure S2.4**, such as K1, K8, and E29, all have higher salt-bridge contents in the unfolded state. However, the formation of these salt-bridges occurs independently, i.e., does not involve the simultaneous formation of nearby contacts, and such positive changes in ΔE_u are simply compensated by negative changes in ΔG_{solv} (i.e., the dehydration penalty is

larger in the unfolded state), resulting in small net variations in $\Delta f = \Delta E_u + \Delta G_{solv}$ (**Table S2.6**). Therefore, those residues involved in the formation of independent, non-native contacts do not significantly contribute to the folding stability.

The site-resolved results shown in **Figure 2.5** also enable us to identify critical residues to the thermal stability. For example, the top five residues that contribute most to Δf of amino acid residue (**Figure 2.5A**) are W6, R9, Y18, F19 and N20. This is in good agreement with the key residues (W6, Y18, F19, and N20) whose mutation to alanine was reported to render the protein to unfold.[12] The site-resolved backbone (**Figure 2.5C**) and side-chain contributions (**Figure 2.5D**) allow us to carry out more detailed comparison with the stability changes ($\Delta\Delta G$) from mutagenesis studies. We find that lists of both the backbone and side-chain Δf in increasing order of the stabilizing residues are in fair agreement with the corresponding lists of reported $\Delta\Delta G$ from the amide-to-ester[11] (backbone) and alanine mutagenesis[12] (side-chain) studies, respectively (**Tables S2.5** and **S2.6**, respectively). Indeed, Spearman's correlation coefficients, which measure the strength and direction of association between two ranked quantities, were computed to be 0.73 and 0.76 for backbones and side-chains, respectively, indicating that our computational and experimental results are strongly correlated. In this regard, it is worthwhile to emphasize that our results were obtained solely from the analysis of the wild type protein, i.e., without introducing any mutations as in mutagenesis studies, and this is the major advantage of our site-specific analysis method.

We further explain what the effective energy f entails in the context of protein folding of HP-36. Using the site-directed thermodynamic analysis method mentioned previously, we resolve Δf into individual backbone and side chain contributions to identify the stabilizing residues in HP-36 as shown in **Figure 2.6**.

The significance of the approach is established in that a large favorable increase in folding stability, indicated by negative Δf, arising from the H-bonding backbones and hydrophobic cluster side chains cooperatively are well-captured, while the overall contributions from the charged ones are destabilizing ($\Delta f > 0$). This finding agrees well with the experimental study that also found the H-bonding backbones and hydrophobic cluster of HP-35 to be stabilizing significantly.[88]

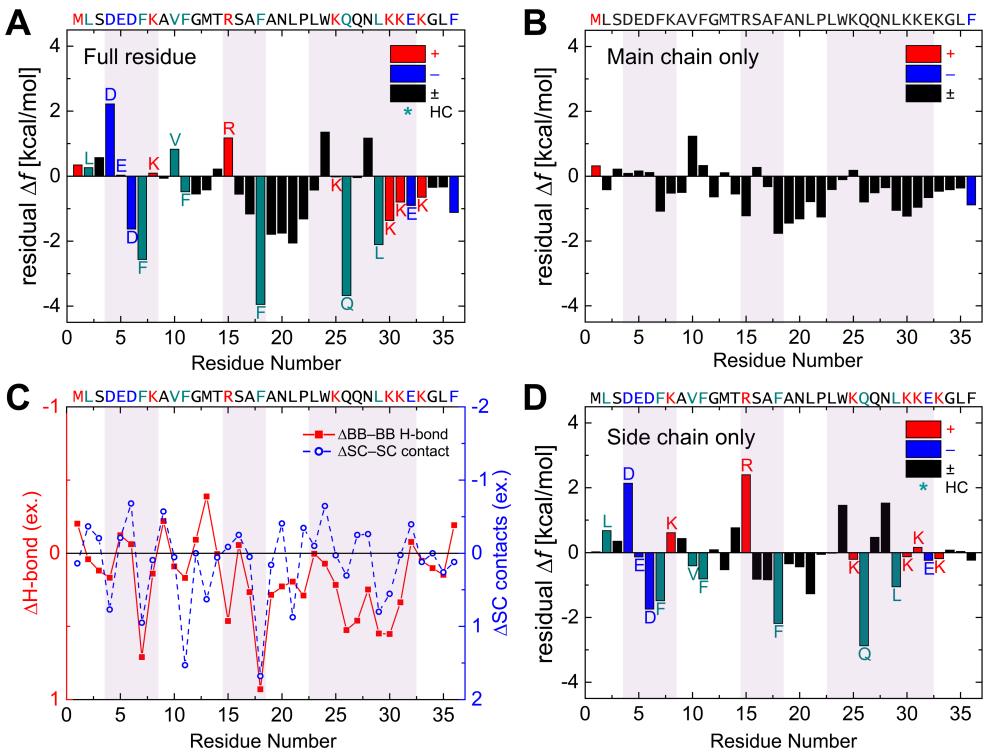


Figure 2.6. Thermodynamic and structural analyses for each backbone and side-chain of the HP-36. The folding effective energy contributions $\Delta f = f$ (folded) $- f$ (unfolded) are resolved for every constituent (A) amino acid residue, (B) backbone, and (D) side chain using the site-directed thermodynamic analysis of HP-36. The amino acid residue sequence is provided at the top of each plot, for which positively charged, negatively charged, neutral, and hydrophobic cluster residues are colored as red, blue, black, and cyan, respectively. The purple strips demarcate three α-helix regions. (C) The differences in the number of backbone H-bond formation upon folding and the number of side chain contacts upon folding are represented by a filled square connected to solid lines and an empty circle connected to dashed lines, respectively.

We would like to extend comments on the convergence of configurational entropy calculation for the unfolded-state trajectories. **Figure S2.3B** does not exhibit a constant increase in configurational entropy, signifying its well converged result in each trajectory. However, a comparable magnitude of standard error in the unfolded-state configurational entropy to the overall folding free energy poses a concern that makes our thermodynamic calculation less significant. The reason for a sizeable standard error stems from a statistical deviation in some of unfolded-state trajectories, UNFOLD1 (maximum) and UNFOLD4 (minimum), that contribute to the average value with an equal weight. For statistical purposes, adding more unfolded-state trajectory results that explore wider phase space can reduce the effect of such deviating portions on the average value and the standard error. While the convergence of configurational entropy is still an important issue, its thorough discussion is beyond the scope of this study and can be addressed in future studies.

Simple native-centric models have been suggested for estimating Φ values of individual amino acid residues in a protein:[89–91] in these models, Φ values are approximated in terms of the fraction of native contacts at the folded, unfolded, and transition states. The estimated Φ values generally correlate well with the experimental values and can be used for identifying residues critical to protein folding. Whereas the estimation of Φ values can be efficiently done with Gō-type model simulations,[91,92] such an estimation is difficult with physics-based potentials since it requires an ensemble of transition-state protein configurations, which is difficult to obtain: the transition-state ensemble in the previous applications was typically generated by the use of the special-purpose supercomputer.[89,90] On the other hand, our thermodynamic decomposition method works solely with the unfolded-

and folded-state simulations of microsecond timescales, which can nowadays be done routinely by normal computers equipped with GPU (graphical processing unit) cards.[93] This is another advantage of our site-specific analysis method.

Chapter 3. Comparing the Influence of Explicit and Implicit Solvation Models on Site-Specific Thermodynamic Stability of Two Model Proteins

This work is a comparative study of thermodynamic stability between explicit and implicit solvent simulations of representative alpha- and beta-sheet proteins. The site-directed thermodynamic analysis method is then used to decompose the folding stability into contributions from individual backbones and side chains of protein. From a systematic comparison among residues, the key structural origins of thermodynamic discrepancy from GBSA solvent are identified. Using the trajectories containing the TIP3P and the generalized Born/Surface Area solvent models from molecular dynamics simulation, we assess the residue-specific folding free energy components of WW domain and HP-36, as shown in **Figure 3.1**.

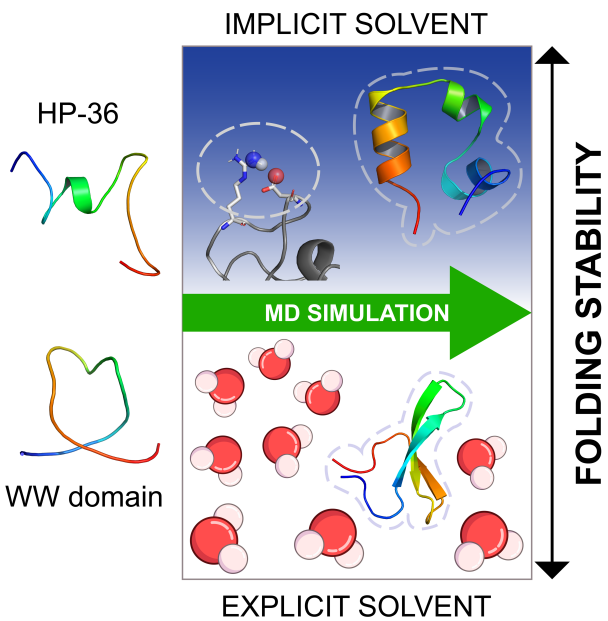


Figure 3.1. Thermodynamic comparison of HP-36 and WW domain proteins in explicit and implicit solvents, regarding folding stability contributions from backbones and side chains.

3.1 Methods

3.1.1. GBSA Solvent MD Simulation of Folded and Unfolded States

The structures of WW domain (PDB ID: 2F21) and HP-36 (PDB ID: 1VII) were taken from the RCSB Protein Data Bank,[19,94] and truncated such that the first five residues and the ones trailing G39 were removed. The protonation states of residues were assigned based on the physiological pH, and HIS21 was Nε. The AMBER ff14SB[95] as protein force field and the GBSA model[56] of Onufriev, Bashford, and Case ($GB^{OBC}II$), also denoted as igb5 in AMBER, as the implicit solvent model were used with the Debye-screening parameter of 1 nm^{-1}. A fully extended peptide was prepared for the unfolded state simulation, for which the following procedures were identical as the folded state. The OpenMM software package[96], accelerated by CUDA-enabled graphics processing units (GPUs), was used to perform the implicit solvent MD simulation of 1 μs-long production run. The system was minimized using an energy tolerance level of 10kJ/mol. The last 900 ns with 1 ns time interval was subjected to structural and thermodynamic analyses. The Langevin integrator is used with the collision frequency of 2 $picosecond^{-1}$. Four independent folded state simulations in equilibrium were carried out for each WW domain and HP-36.

A fully extended peptide was prepared for the unfolded-state simulation with parameters identical to those of the folded state, as shown in **Figure 3.2**. A 1 μs *NVT* production run was performed with the same Langevin integrator. Eight independent trajectories were obtained for the WW domain protein, and six for the HP-36. For each trajectory of both folded and unfolded states, the last 900 ns with a 1 ns time interval were subjected to structural and thermodynamic analyses.

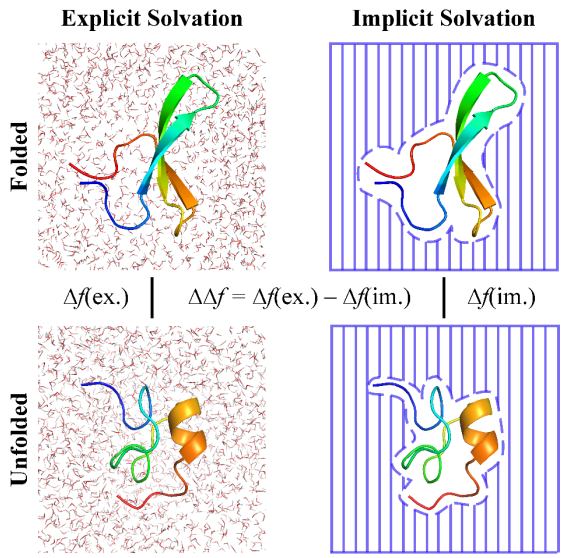


Figure 3.2. Comparative schematic of the protein–solvent systems in explicit and implicit solvents. The illustration juxtaposes the protein systems solvated in explicit water molecules, represented by licorice structures, and those in implicit solvents, shown in blue stick representations. The thermodynamic stability difference between the folded and unfolded protein states is denoted by the effective energy upon folding Δf. A comparative analysis between explicit (ex.) and implicit (im.) solvent results is illustrated by the difference in folding effective energy $\Delta\Delta f$.

3.1.2. Explicit Water MD Simulation of Folded and Unfolded States

The Amber program packages, accelerated by CUDA-enabled GPUs, were used to perform explicit solvent all atom MD simulations.[97] The previously reported simulation trajectories of the WW domain[54] and the HP-36[76] were taken, for which the simulation setup is briefly described here. The Amber ff99SB-ILDN protein force field[98] and the TIP3P water model[55] were used to perform the WW domain simulation. Six independent folded-state production runs of 1 μs-long each were performed at T = 300 K and P =1 bar. For each of the eight independent unfolded-state simulations, the simulated annealing approach was implemented on the folded protein initiated with heating at T = 600 K, which was followed by a gradual decrease in the temperature. At T = 300 K, 2 μs-long *NPT* production runs were then

performed, for which only the last 1 μs was subjected to structural and thermodynamic analyses. For the HP-36 simulation, the ff99SB force field[69] and the TIP3P model[55] were used. Three independent folded-state trajectories were obtained with a procedure identical to that of the WW domain, while nine independent unfolded-state trajectories out of ten reported trajectories (corresponding to UNFOLD 1,2, and 4–10) from the HP-36 simulations as described in **Chapter 2** were taken, each of which had a 5 μs-long *NPT* production run.

3.1.3. Structural Analyses

From each of the 1-μs long folded and unfolded state trajectories produced by the GBSA implicit (im.) solvent, the last 900 ns were subjected to analysis, for which 900 protein conformations were extracted using a 1 ns time interval. Likewise, each of the last 1-μs TIP3P explicit (ex.) water simulations was used to extract 20,000 conformations of the WW domain with a 50 ps interval and 4,000 conformations of the HP-36 with a 250 ps interval, which were subjected to structural and thermodynamic analyses. The CPPTRAJ software[73] was used for the root-mean-square deviation (RMSD), secondary structure, and hydrogen bond analyses. The DSSP algorithm[74] was used to characterize the secondary structures, for which both α- and 3_{10}-helical residues were designated as helical contents. The native contacts fraction Q was computed with the equation provided in ref. 99. The three β-strand regions of the WW domain were defined as ^{6}WEKRM^{10},^{16}VYYFN^{20}, and ^{26}SQ^{27}. The α-helix regions of the HP-36 were defined as ^{4}DEDFK^{8},^{15}RSAF^{18}, and ^{23}LWKQQNLKKE^{32}. The WW domain side chains that form hydrophobic cluster 1 (HC1) are L2, P3, W6, Y18, and P31, whereas those of hydrophobic cluster 2 (HC2)

are R9, Y17, and F19. Those of the HP-36 side chains for the hydrophobic cluster are L2, F7, V10, F11, F18, K25, Q26 and L29. The RMSD-based k-means clustering method was used to determine representative structures. Hydrogen bond (H-bond) formation was assumed to have occurred when the heavy atom distance between the H-bond donor and acceptor was less than 3.5 Å, and the donor-hydrogen-acceptor was greater than 135°. The H-bonds were classified into three types: those formed between main chains (MC–MC), between side chains (SC–SC), and between a main chain and a side chain (MC–SC). The number of side chain contacts was counted to evaluate the strength of the hydrophobic interaction with 5.4 Å cutoff between a side chain group and the site of interest at least three residues apart. The side chain contacts were categorized as those formed between side chains (SC–SC) and between a main and a side chain (MC–SC). The numbers of H-bonds and side chain contacts were then separated per main chain and per side chain. The overall and per-residue salt-bridge (SB) contents were also computed. We also used the ColabFold program[100] that predicts the experimental protein structures from the amino acid sequence using the AlphaFold algorithm to include another set of structures for a comparison purpose.

3.1.4. Site-Specific Thermodynamic Analyses

For each simulated protein conformation $\boldsymbol{r}_u$, the intraprotein potential energy E_u was calculated using the physics-based molecular mechanics force field, and the decomposition of E_u into the backbone and side chain contributions was further carried out using the site-directed thermodynamic analysis method in **Chapter 2** and ref. 54. To obtain the protein solvation free energy G_{solv}, we applied the three-dimensional reference interaction site model theory for computing the protein-solvent distribution function.[78,101] Then, an atomic decomposition method of G_{solv},

based on the Kirkwood charging formula, was utilized to obtain the site-specific contributions from each of the main and side chains to G_{solv}.[17] These two quantities simplified into the site-specific effective energy $f(r_u) = E_u(r_u) + G_{solv}(r_u)$. The site-specific thermodynamic analysis was performed for the implicit solvent simulation trajectories and the previously reported simulation trajectories of the WW domain[54] and HP-36[76] in standard TIP3P water. From these trajectories, 20,000 conformations of the WW domain and 4,000 conformations of the HP-36 were subjected to the atomic decomposition of G_{solv} and thermodynamic analysis.

3.2 Results and Discussions

We conducted 1 μs implicit solvent MD simulations for each folded and unfolded trajectory of the 33-residue WW domain and the 36-residue HP-36. For the initial structures of the folded state simulation, the X-ray structure of WW domain and the solution NMR structure of HP-36 were used, whereas fully extended amino acid chains were used for the unfolded state simulation. The TIP3P solvent simulation trajectories of WW domain[54] and HP-36[76] were taken from previous studies and **Chapter 2**.

To compare conformational sampling between the distinct simulation trajectories for the two different solvent models, we constructed the two-dimensional probability distribution plots of α-carbon RMSD and the native contacts fraction Q using trajectories with either explicit or implicit water in **Figure 3.3**. As the GBSA simulation is designed to facilitate conformational sampling in general, the distribution profiles from the implicit solvent display greater conformational fluctuations in both the folded and unfolded states. While the major clusters in the folded state are similar between the solvent models, a noticeable variation in the

unfolded protein conformations emerges because the most sampled regions differ. Interestingly, the larger fluctuation does not affect the trajectory-averaged values, as indicated by comparable native contacts fraction Q and Cα RMSD values between the solvent models as shown in **Tables S3.1 and S3.2**. To understand the origin of this similarity in native structural contents, we characterized the formation of secondary structures and native hydrophobic clusters from the side chains. These native structural features were then compared with the non-native ones. The results illustrate that GBSA simulations reproduce the native structures relatively well, while the non-native features show differences between these solvent models. Still, the native salt-bridge, as characterized by the ion-pair distance distribution, exhibited differences in the explicit and implicit solvents in the unfolded state.

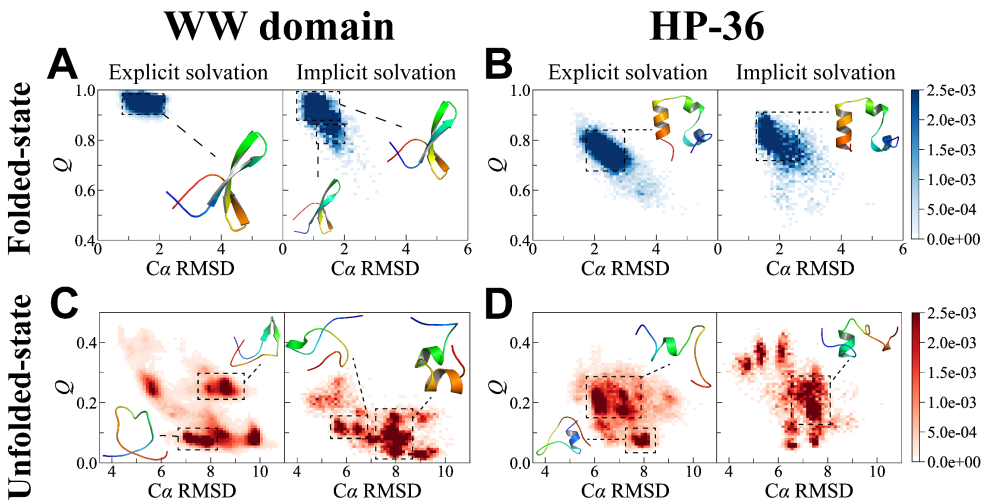


Figure 3.3. Probability distribution plots of the WW domain and HP-36 in both explicit and implicit solvent models. Two-dimensional probability distribution plots of native contacts fraction Q and Cα RMSD from the explicit and implicit solvent simulations of the folded state for (A) WW domain and (B) HP-36 are shown. The representative protein conformations, selected using the k-clustering algorithm, correspond to the most sampled regions (black dashed rectangles). The same distribution plots of the unfolded-state for (C) WW domain and (D) HP-36 are also shown. The color bars that indicate the probability scale are displayed at the rightmost side.

For the thermodynamic comparison of solvent models, we obtained the overall folding contributions to Δf from each trajectory of the folded and unfolded states. By taking the difference of the folding effective energy between the explicit and implicit solvents, $\Delta\Delta f = \Delta f$ (ex.) – Δf (im.), the solvent preference of protein folding can be estimated. As shown in **Table 3.1**, we observed that the explicit solvent stabilizes the folding of the WW domain more where $\Delta\Delta f = -8.9 \pm 3.3$ kcal/mol. In contrast, HP-36 folding is more stable in the GBSA simulation where $\Delta\Delta f = +12.6 \pm 2.9$ kcal/mol. To identify the key residues that contribute to the thermodynamic discrepancy between the two solvation models, we resolved the folding effective energy difference $\Delta\Delta f$ into the contributions of each main and side chain of the protein. Thermodynamic analyses of $\Delta\Delta E_u$ and $\Delta\Delta G_{solv}$ indicated whether the $\Delta\Delta f$ contribution arises from a discrepancy in either energetic or solvent interactions.

Table 3.1. Tabulated Data of Thermodynamic Values for WW domain and HP-36

	WW domain			HP-36		
	ΔE_u	ΔG_{solv}	Δf	ΔE_u	ΔG_{solv}	Δf
Ex[a]	4.9 ± 17.0	–34.1 ± 16.3	–29.2 ± 2.0	–14.2 ± 22.7	–7.6 ± 21.5	–21.8 ± 2.1
Im[a]	33.9 ± 17.8	–54.1 ± 17.1	–20.3 ± 2.6	15.8 ± 20.9	–50.2 ± 20.4	–34.4 ± 1.9
	$\Delta\Delta E_u$	$\Delta\Delta G_{solv}$	$\Delta\Delta f$	$\Delta\Delta E_u$	$\Delta\Delta G_{solv}$	$\Delta\Delta f$
Δ [a]	–29.0 ± 24.6	20.0 ± 23.6	–8.9 ± 3.3	–30.0 ± 30.8	42.6 ± 29.7	12.6 ± 2.9

[a] Average ± standard error

3.2.1 Secondary Structure Preference

To assess the difference in secondary structure preference between the TIP3P and GBSA models, we analyzed the β-strand propensity of the WW domain and the number of hydrogen bonds between main chains (MC–MC). Both solvation models exhibit highly comparable β-strand contents, where the secondary structure content differences in the β-1 strand between the two models are 1.2 ± 0.2 % and 7.1 ± 6.6 % in the folded and unfolded states, respectively (**Table S3.2**). The similarity in the

secondary structure preference between solvation models is further highlighted by comparing the number of H-bonds between backbones (**Figure 3.4**, **Tables S3.3**, and **S3.4**). The number of H-bonds remains at approximately 10 bonds in the folded state for both solvation models and seven to eight in the unfolded state.

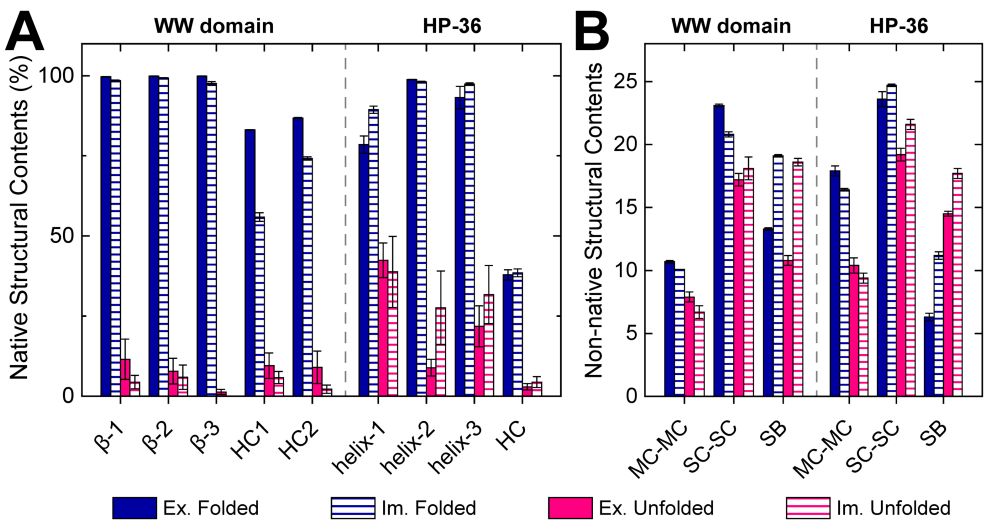


Figure 3.4. Native and (B) non-native structural contents in the folded and unfolded states of the representative β-sheet (WW domain) and α-helix (HP-36) proteins from explicit and implicit solvent simulations. (A Populations of native structure features, including β-strand, helix, and hydrophobic cluster (HC) contents, are shown with a dashed grey line to split the WW domain and HP-36 panels. The explicit (Ex.) and implicit (Im.) solvent simulation results are indicated by filled and sparsely stripped boxes, respectively. (B) The non-native features correspond to the number of H-bonds between main chains (MC–MC), the number of side chain contacts (SC–SC), and the salt-bridge (SB) contents. The salt-bridge content is presented as a percentage. The salt-bridge content alone is in the unit of percentage. The standard error bars are placed at the top of each bar.

The α-helical content comparison in the HP-36 also revealed that the implicit solvent simulation mildly overpopulates the secondary structure consistently in both the folded and unfolded states such that the difference in the largest helix α-3 are found to be 4.2 ± 3.5 % and 10 ± 11.1 %, respectively (See **Table S3.2**). In contrast, the H-bonds between backbones in the GBSA simulation were

underestimated by approximately one bond in each of the folded and unfolded states, as listed in **Tables S3.3** and **S3.5**. It can be observed that fewer H-bonds between main chains are replaced by higher H-bonds between a main chain and a side chain (MC–SC). Understanding how the marginal helical bias in the HP-36 and comparable β-strand content in the WW domain translate to thermodynamic differences is important.

The overall backbone stability of the WW domain was found to be solvent-independent because the overall folding effective energy of backbones between the two different solvation models was highly comparable with the total difference of $\Delta\Delta f_{\text{backbone}} = +1.1 \pm 1.8$ kcal/mol, according to **Figure 3.5** and **Table S3.6**. A closer examination of the individual residues revealed that the folding of β-strand backbones is slightly more stabilized in the TIP3P solvent ($\Delta\Delta f_{\text{backbone}} < 0$), which is counteracted by relatively destabilized turns and termini backbones ($\Delta\Delta f_{\text{backbone}} > 0$), resulting in a small net $\Delta\Delta f_{\text{backbone}}$. This thermodynamic similarity in explicit and implicit solvent simulations from the WW domain backbones resonates with comparable β-strand propensities. In contrast, most of the HP-36 backbone residues were found to be more stabilized in the GBSA simulation where $\Delta\Delta f_{\text{backbone}} = +6.4 \pm 2.6$ kcal/mol (**Table S3.8**). A considerable amount of thermodynamic deviation from HP-36 backbones can be attributed to slightly larger α-helical propensity and larger backbone H-bond propensity in N-terminal and loop regions. While the total $\Delta\Delta f_{\text{backbone}}$ suggests a thermodynamic discrepancy, the per-residue analysis indicates that no single backbone residue contributes to significant thermodynamic differences, where $\Delta\Delta f_{\text{backbone}} < 1.0$ kcal/mol in all but one backbone residue. This result suggests that the backbone folding stability behaviors of both the WW domain and HP-36 are represented with reasonable accuracy by the GBSA solvent.

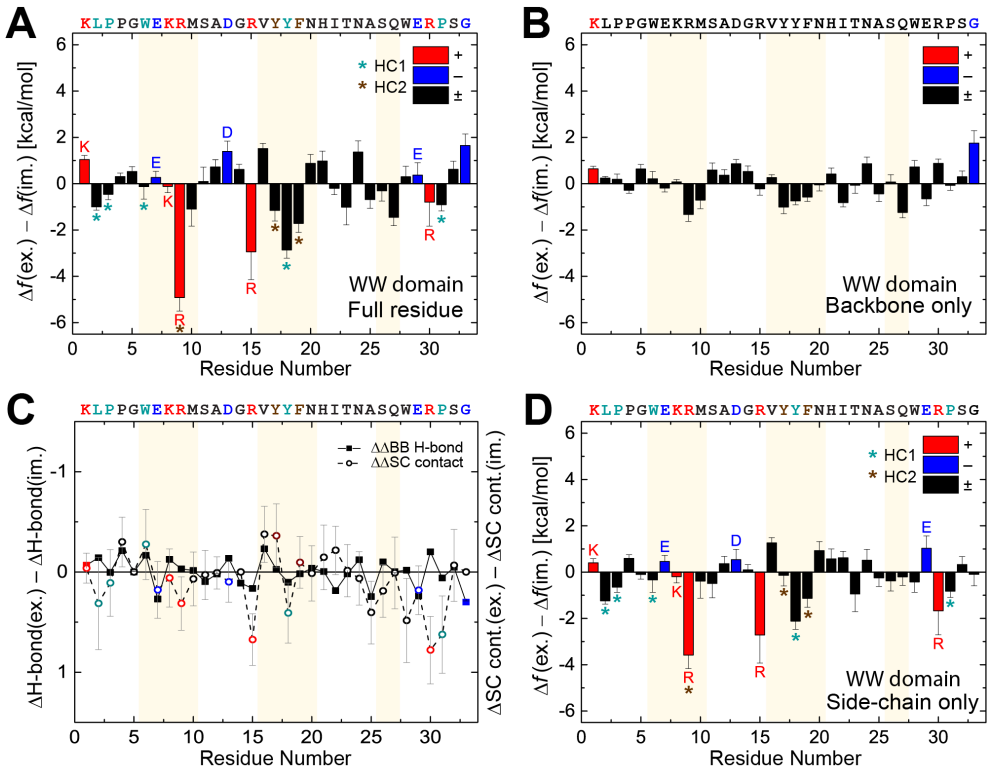


Figure 3.5. Thermodynamic and structural differences in solvation models upon folding of the WW domain. Between explicit (ex.) and implicit (im.) solvation models, the folding effective energy differences $\Delta\Delta f = \Delta f(\text{ex.}) - \Delta f(\text{im.})$ are resolved for (A) total, (B) backbone, and (D) side chain contributions of every constituent residue using the site-directed thermodynamic analysis. The amino acid residue sequence is provided at the top of each plot, for which positively charged, negatively charged, neutral, hydrophobic cluster 1, and hydrophobic cluster 2 residues are colored red, blue, black, cyan, and brown, respectively. The yellow strips demarcate three β-strand regions. (C) The differences in the number of backbone–backbone (BB–BB) H-bond formation (filled square) and side chain (SC–SC) contacts (empty circle) upon folding.

3.2.2 Hydrophobic Cluster Preference

The strength of hydrophobic interaction is often measured by the number of pairwise contacts between side chain atoms. To determine the sensitivity of hydrophobic interactions in the explicit and implicit water simulations, we compared the native heavy atom contacts within each of the two segregated hydrophobic clusters (HCs) in the WW domain and a single hydrophobic cluster in HP-36. In the folded WW domain, a sizable discrepancy in the formation of two native hydrophobic clusters was observed between the two solvent simulations. Notably, folded hydrophobic cluster 1 (HC1) was largely preferred in the TIP3P solvent, with its population maintained at 83.1 ± 0.1 % compared to 55.9 ± 1.3 % in the GBSA solvent (**Figure 3.4A** and **Table S3.2**). In **Figure 3.5C** and **Table S3.4**, the number of side chain contacts per residue (SC cont.) is presented to estimate the contribution of van der Waals interaction on a residue-by-residue basis. Here, we found that four (L2, P3, Y18, and P31) out of five residues exhibit a lower tendency to form hydrophobic cluster 1 in the GBSA solvent ($\Delta\Delta$SC cont. > 0).

The structural difference in the WW domain hydrophobic clusters between TIP3P and GBSA water translates to thermodynamic deviation in the two solvent models. All the side chains comprising HC1 and HC2 are shown to be under-stabilized in the GBSA solvent simulation, as evident from the uniformly negative $\Delta\Delta f_{\text{side chain}}$ values. The largest contribution to $\Delta\Delta f_{\text{side chain}}$ among uncharged side chains arises from Y18 of HC1 at –2.1 kcal/mol alone (**Table S3.7**). Here, a limitation of the GBSA solvent simulation is highlighted in that the key residues critical to folding are under-stabilized, as exemplified in Y18. To investigate the under-stabilization of the HCs in the GBSA solvent simulation, we examined both the folding intraprotein potential energy difference $\Delta\Delta E_u = \Delta E_u$ (ex.) – ΔE_u (im.) and

the folding solvation free energy difference $\Delta\Delta G_{solv} = \Delta G_{solv}$ (ex.) – ΔG_{solv} (im.). Y18 formed fewer side chain contacts ($\Delta\Delta$SC cont. > 0) in the GBSA solvent. This under-representation can be attributed to weaker van der Waals contacts, which are indicated by more positive ΔE_u(im.) than ΔE_u(ex.) of TIP3P water simulation, resulting in the net negative change in folding intraprotein potential energy ($\Delta\Delta E_u <$ 0), as depicted in **Figure S3.1**. Specifically, a residue–residue interaction of Y18 in the folded state is less favorable in the GBSA solvent.

We investigated further how the key residues that contribute to folding stability are represented with respect to water-mediated interactions in the GBSA solvent simulation. It has been shown that these key residues are the ones that simultaneously form the secondary structure H-bonds and the side chain hydrophobic cluster.[54] This is because the dehydration penalty is alleviated when those contacts cooperatively stabilize. Other residues that only form either H-bonds or side chain contacts gain a large negative intraprotein potential energy, only to find it cancelled by equally large positive solvation free energy ($\Delta f = \Delta E_u + \Delta G_{solv} \sim 0$).[54] Thus, it is also important to rely on the role of solvation free energy G_{solv} in interpreting the thermodynamic discrepancy in folding stability, particularly for the key hydrophobic residues. As shown in **Figure S3.1D**, the positive values in $\Delta\Delta G_{solv}$ in most HC1 side chains (P3, W6, Y18, and P31) indicate that the GBSA solvent assumes a higher affinity of water upon folding, which is reflected by less positive values of ΔG_{solv}(im.). This result can be understood as follows: the surface area approximation of key residues to folding stability underestimates the favorable van der Waals contacts (less negative ΔE_u(im.)) and weakens the strength of hydrophobic interactions (less positive ΔG_{solv}(im.)). In consequence, the hydrophobic cluster residues in the GBSA solvent suffer from reduced folding stability ($\Delta\Delta f < 0$), which can be attributed to the solvation effect or

the difficulty in capturing dehydration penalty appropriately.

In the HP-36 simulations, it was noticed that the native structural contents in the folded HP-36 were comparable at 37.9 ± 1.6 % and 38.5 ± 1.2 % in the TIP3P and GBSA solvents, respectively (**Figure 3.4A**). This suggests that the surface area estimation of nonpolar contributions in MD simulations closely emulates the influence of the explicit treatment of water on the hydrophobic interaction in the HP-36. This protein structural similarity between the explicit and implicit solvent simulations translates into comparable thermodynamic contributions from the side chains that form the hydrophobic cluster, as the subtotal $\Delta\Delta f_{\text{side chain}}$ value of the entire eight HC side chains was merely +1.5 ± 1.0 kcal/mol (**Table S3.9**). However, the key residue for folding stability, F18, is again shown to be under-stabilized in the GBSA simulation, indicating that the thermodynamic discrepancy behind key neutral residues in GBSA is due to a dehydration penalty or a solvent effect (**Figure 3.6** and **S3.2**).

However, the absolute quantities of the heavy atom contacts in the folded state are found to be low at 38 % and 39% in the TIP3P and GBSA solvents. To understand the limited preservation of hydrophobic contacts, we obtained the HP-36 structure prediction using the AlphaFold program.[100] With the structure prediction reporting 53 %, it seems that some of the native hydrophobic contacts are overly represented.

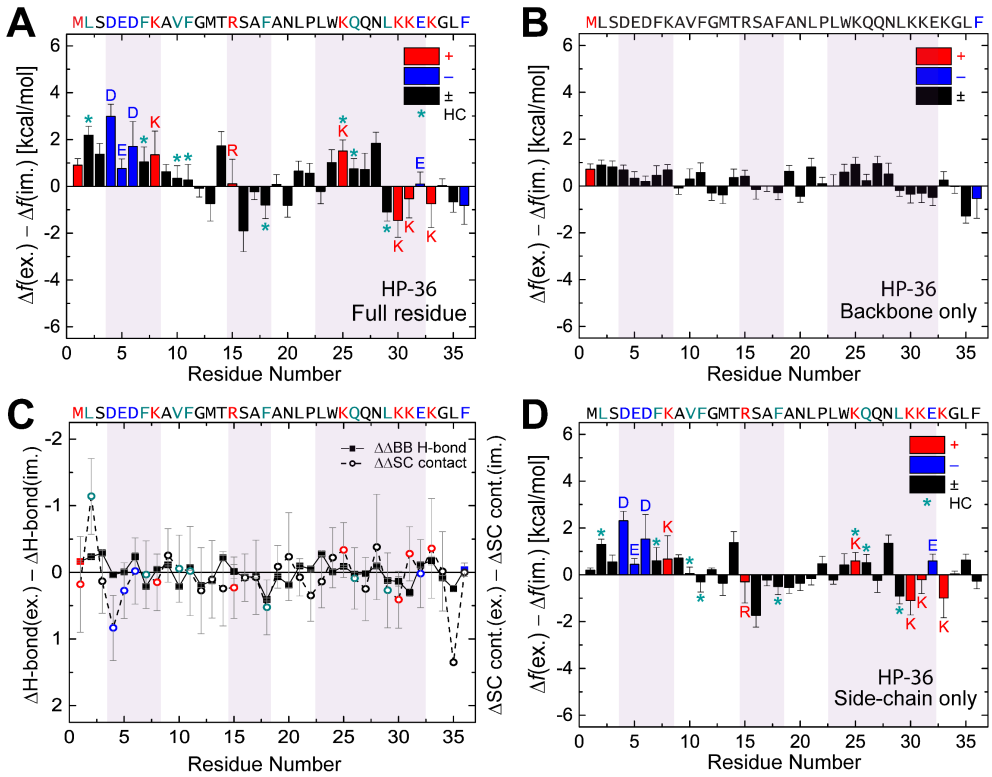
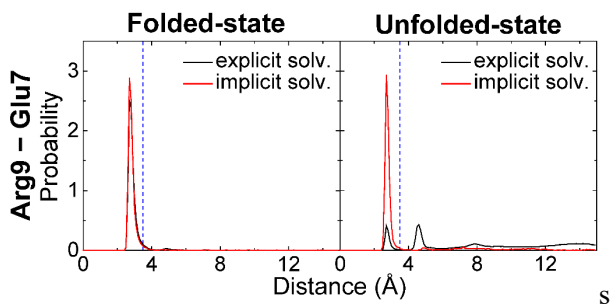


Figure 3.6. Thermodynamic and structural differences in solvation models upon folding of the HP-36. Between explicit (ex.) and implicit (im.) solvation models, the folding effective energy differences $\Delta\Delta f = \Delta f$ (ex.) – Δf (im.) are resolved for (A) total, (B) backbone, and (D) side chain contributions of every constituent residue using the site-directed thermodynamic analysis. The amino acid residue sequence is provided at the top of each plot, for which positively charged, negatively charged, neutral, and hydrophobic cluster residues are colored red, blue, black, and cyan, respectively. The purple strips demarcate three α-helix regions. (C) The differences in the number of backbone–backbone (BB–BB) H-bond formation (filled square) and side chain (SC–SC) contacts (empty circle) upon folding.

3.2.3 Salt Bridge Preference

The crystal structure of the WW domain contains a salt-bridge, R9–E7, for which the structural properties were characterized by the ion-pair distance distribution to compare the electrostatic strength shaped by the explicit and implicit solvation models, as shown in **Figure 3.7**. The main structural difference stems from the formation of a salt-bridge in the unfolded state in the GBSA solvent. Specifically, the salt-bridge populations of R9 in the unfolded state were found to be 51 ± 10 % in the TIP3P solvent and 95 ± 2 % in the GBSA solvent, while those of the folded R9 in both solvents were 97 ± 0.3 % (See **Figure S3.3**). Furthermore, the overall non-native salt-bridge contents nearly doubled in the GBSA solvent in all cases, indicating biased electrostatic interactions (**Figure 3.4B**). It is important to recognize that there are two consequences of the excess formation of salt-bridges. In the folded state, the formation of an extra non-native salt-bridge can interrupt the formation of stabilizing native hydrophobic interactions. Moreover, excess salt-bridge formation in the unfolded state mitigates the stability gain arising from protein folding, as observed in R9–E7.



s

Figure 3.7. Ion-pair distance distribution of a native salt-bridge, Arg9–Glu7, of the WW domain for the explicit (black) and implicit (red) solvent simulations obtained from compiled simulation trajectories of both the folded and unfolded states. The salt-bridge cutoff distance at 3.5 Å is indicated (blue dashed line).

The thermodynamic consequence of stronger electrostatic interactions in the GBSA solvent was clearly recognizable because the large contribution to negative $\Delta\Delta f$ arises from the positive side chains of the WW domain, suggesting that the folding stability of these residues was underestimated in the GBSA solvent. In particular, the R9 and R15 side chains exhibited the most significant contributions to the latter of the two consequences of more salt-bridges described above, where the $\Delta\Delta f$ contribution from R9 side chain amounts to –3.6 ± 0.5 kcal/mol. Indeed, sizable contributions to $\Delta\Delta f$ from the two positive residues originate from a large negative value of $\Delta\Delta E_u$, indicating a more favorable intraprotein contact formation upon folding in the TIP3P solvent. In contrast, the remaining positively charged residues, including K1, K8, and R30, contributed to relatively insignificant $\Delta\Delta f$. This behavior of positive residues can be understood by a change in salt-bridge formation upon folding, that is, ΔSB (ex.) – ΔSB (im.). For example, both R9 and R15 that contribute significantly to $\Delta\Delta f$ achieve $\Delta\Delta SB > 0$, as opposed to the rest of the positive residues with $\Delta\Delta SB < 0$. This tendency is also captured by observing $\Delta\Delta E_u$ contributions from K1, K8, and R30 that gain more SB upon folding in the GBSA solvent simulation ($\Delta\Delta E_u > 0$). The resulting low values in this $\Delta\Delta f$ can be attributed to a large mismatch in $\Delta\Delta E_u$ and $\Delta\Delta G_{solv}$, which are inversely directed, as shown in **Figure S3.1**.

For HP-36, over-stabilization of the salt-bridge was again observed because the GBSA simulation roughly generated an extra salt-bridge in each of the folded and unfolded states compared to the TIP3P simulation. However, the positive side chains contribute to a small amount of $\Delta\Delta f$ because their overall contributions mostly cancel each other. The reduced influence of the positively charged side chains on $\Delta\Delta f$ can be attributed to the lack of native salt-bridges in HP-36. The side chains of K31 and

K33 acquired more salt-bridge contents upon folding in the TIP3P solvent ($\Delta\Delta SB > 0$), again resulting in a large favorable change in $\Delta\Delta E_u$. However, their limited contributions to $\Delta\Delta f$ were found to be less than –1 kcal/mol each, indicating a minor role of non-native salt-bridges in folding stability. This minor contributions from non-native salt-bridges were again observed in the negative side chains. Both D4 and D6 that were over-stabilized in the GBSA solvent ($\Delta\Delta f > 0$) exhibited contrasting behavior in salt-bridge formation, where D4 acquired fewer salt-bridge upon folding in the TIP3P solvent ($\Delta\Delta SB > 0$), as opposed to D6. The overall side chain contributions to $\Delta\Delta f$ were found to be +6.0 ± 2.6 kcal/mol, which can be partially attributed to the salt-bridge over-stabilization in the GBSA solvent. As the per-residue analysis revealed that all side chains do not exceed $\Delta\Delta f$ by 2.0 kcal/mol except for a single side chain, it can be argued that the folding behavior of the HP-36 protein using the GBSA solvent simulation is generally in agreement with that of TIP3P solvent simulation except for those over-stabilized negative residues ($\Delta\Delta f > 0$).

To compare the folding effective energy values Δf obtained from each of the explicit and implicit solvent simulations, we calculated the Pearson correlation coefficients R for the backbone and side chain Δf values. As shown in **Figure 3.8**, a linear relationship was evident between Δf values obtained from TIP3P and GBSA solvent simulations in both the backbones and side chains of WW domain and HP-36. The Pearson coefficients R of the WW domain were found to be 0.59 for backbones and 0.60 for side chains and those of HP-36 were 0.70 and 0.72. A strong correlation arising from the HP-36 thermodynamic results suggests that the folding stability of HP-36 on a residue-by-residue basis is mostly captured by the GBSA solvent simulation. However, the WW domain results suggest that proteins with

native salt-bridges should be addressed carefully when using the GBSA solvent simulation. In this regard, it can be argued that the determination of residue-specific $\Delta\Delta f$ using the site-directed thermodynamic analysis enabled us to quantify the influence of GBSA solvent artifacts in each protein.

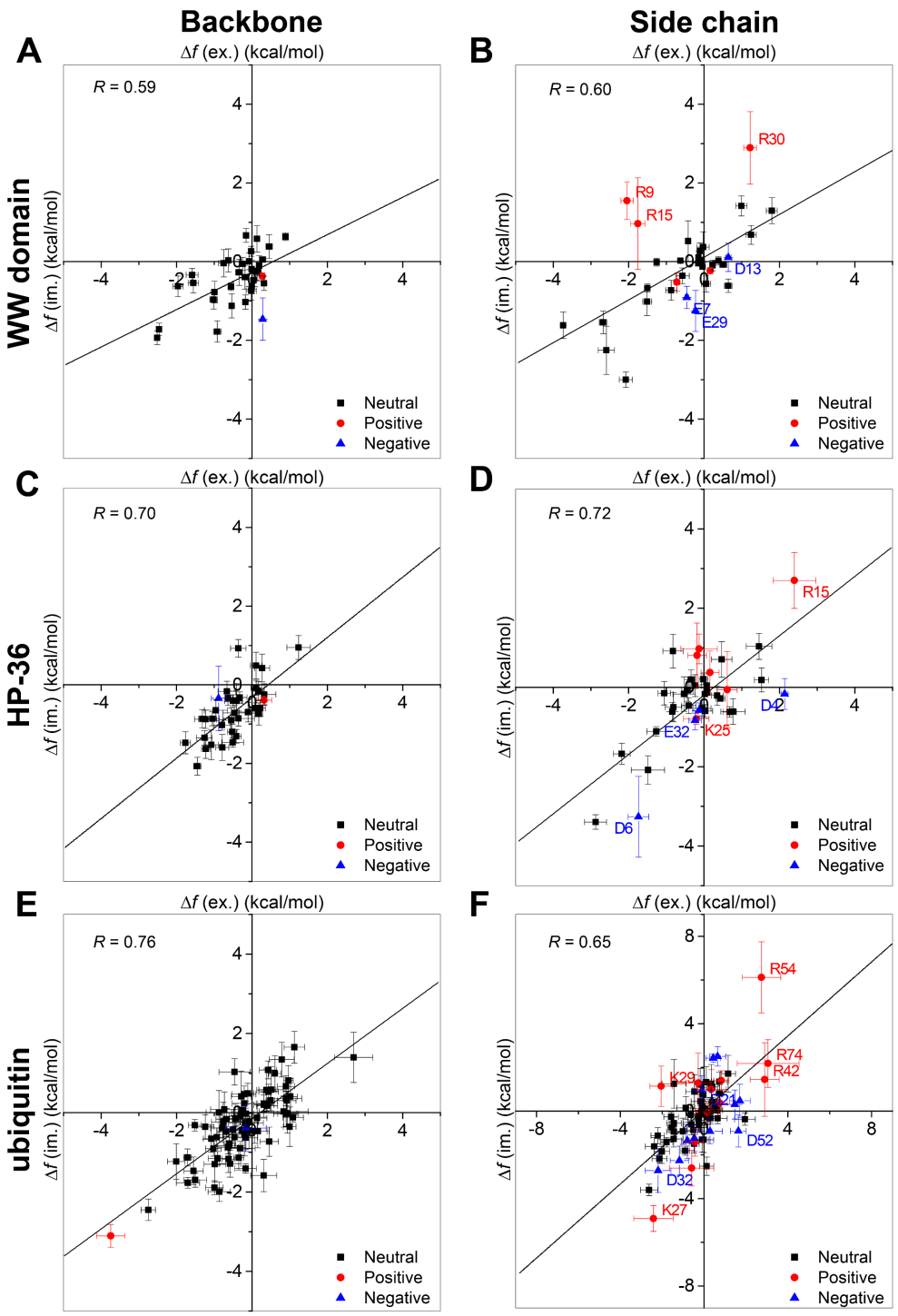


Figure 3.8. Correlation plots between the folding effective energy from explicit (ex.) and implicit (im.) solvent simulations, separated by backbones and side chains of (A, B) WW domain, (C, D) HP-36, and (E, F) ubiquitin. Solid lines are based on the linear fits to the respective data points, along with the corresponding Pearson correlation coefficient *R*.

3.2.4. Generalization to Other Proteins

Here, we would like to discuss the extent to which our findings of the folding stability behavior can be applied to other proteins and acknowledge whether our results are dependent on the choice of protein. We performed the same site-directed thermodynamic analysis on the folding of ubiquitin,[102] a 76-residue protein composed of two helices and five β-sheets (See **Figures S3.4**, and **S3.5**). It was found that the Pearson coefficients of ubiquitin were 0.76 and 0.65 for the backbone and side chain Δf values, respectively (See **Figure 3.8E** and **3.8F**). The difference Δf(ex.) – Δf(im.) for the backbone was 6.8 ± 6.9 kcal/mol and that of the side chain was – 0.3 ± 6.2 kcal/mol. Despite the notable difference in backbone $\Delta\Delta f$, the high correlation observed in the Pearson coefficient suggests that the folding stability behavior of most residues is accounted for well in the GBSA simulation. As shown in **Figure S3.4**, most helical main chains within I23 to E34 and L56 to Y59 exhibited positive $\Delta\Delta E_u$, counteracted by negative $\Delta\Delta G_{solv}$, which is the same thermodynamic behavior observed in the helical regions of the HP-36 protein.

More prominent discrepancies arose from charged side chains and bulky hydrophobic residues. These positive side chains, as they are under-stabilized in the GBSA solvent, tend to form fewer side chain contacts ($\Delta\Delta$SC Contacts > 1), as evident in K8, K29, and R54 in **Figure 3.9**. In addition, K11, K27, and R42 form more side chain contacts, resulting in positive $\Delta\Delta f$ values. In contrast, the thermodynamics of negative side chains is less responsive to the structural differences, as observed in E24, E34, and D52. T9 and Y59 have been identified as the major contributors to the observed discrepancy among neutral amino acid residues, with the former leading to over-stabilized contacts and the latter to under-stabilized contacts in the implicit solvent. We highlight that over-stabilized H-bonds

accompanied by over-stabilized side chain contacts in T9 cooperatively lead to the most significant thermodynamic discrepancy on a residue-by-residue basis, and vice versa. This cooperative nature of folding stability is not fully accounted for in the GBSA solvent and lead to under-stabilization, resulting in discrepancies in the folding thermodynamics analysis of protein folding.

By performing both explicit and implicit solvent simulations, we assessed how accurately implicit solvent simulations reproduce the structural and thermodynamic characteristics at the individual residue level. We found that some of the native structure features such as the RMSD and secondary structure propensities in both representative α-helix and β-sheet proteins are reproduced with sufficient accuracy, which agrees well with the previous GBSA simulation study of WW domain that obtained an RMSD value as low as 0.5.[43] In contrast, the non-native characteristics such as the salt-bridge show sizable discrepancies between the two solvent models for most residues of both the WW domain and HP-36. This over-stabilization of salt bridges in implicit solvent models has been previously documented.[103,104] Meanwhile, an underestimation of the folding stability contributions from native hydrophobic clusters in the GBSA model is found to be protein-dependent, where the hydrophobic clusters are underestimated only in the WW domain, unlike in HP-36. Our result in ref. 105 of over-stabilized α-helix and under-stabilized β-sheet proteins is consistent with the previous folding study that reported the GBSA solvent's tendency to favor the folding of α-helix over β-sheet proteins.[41]

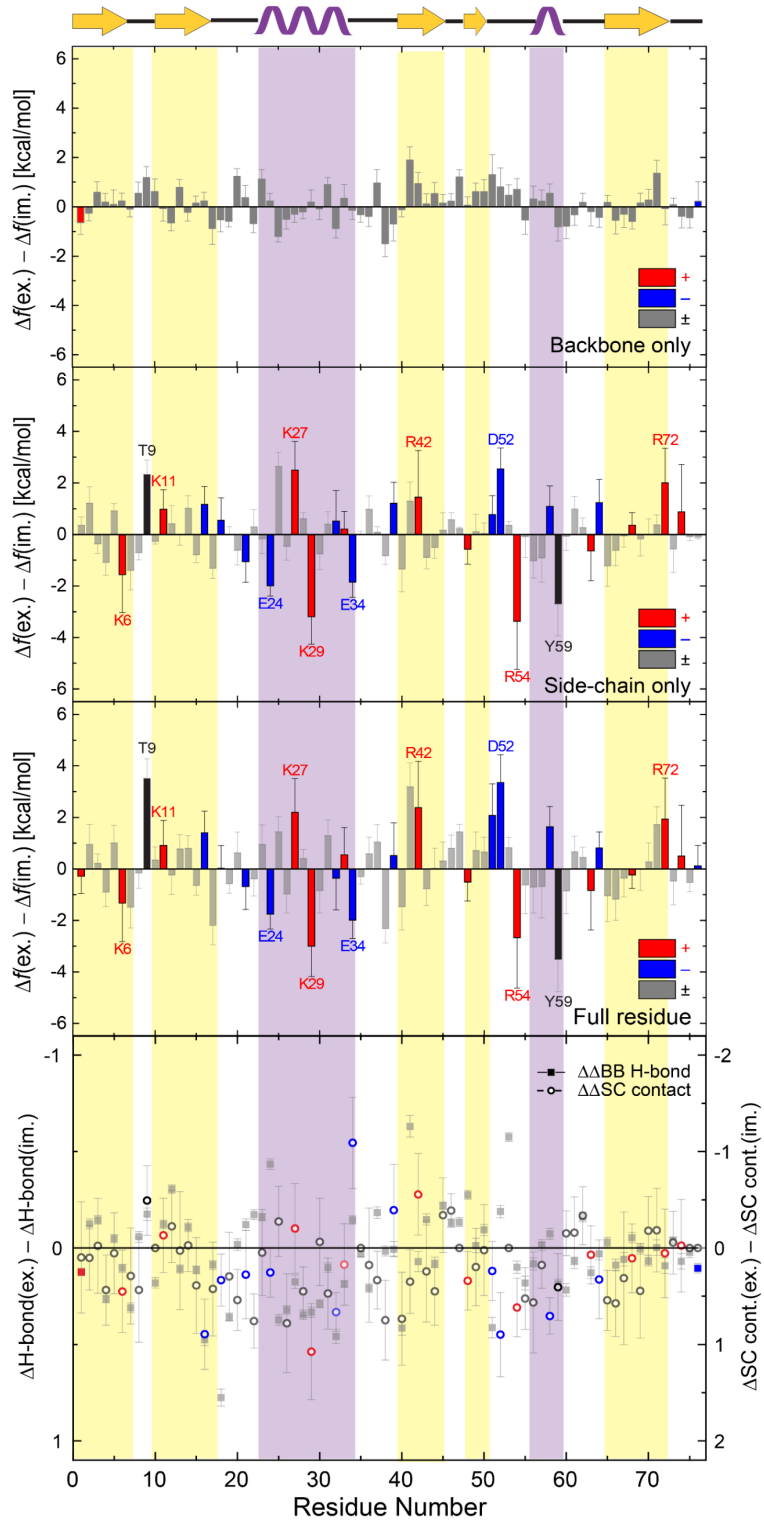


Figure 3.9. Thermodynamic analyses for each main and side chain residues of ubiquitin. The main and side chain contributions to the folding effective energy difference Δf (ex.) – Δf(im.).

Chapter 4. Conclusion

Discovering a structure-thermodynamics relationship of macromolecules like protein carries a significant implication in the protein studies. In the present work, we applied the site-specific thermodynamic analysis method to provide a quantitative comparison of critical sites in the WW domain folding. Inspired by the previous studies that resolved the extent of stability in terms of constituent amino acid residues, we took the decomposition scheme a step further to break down the solvent-averaged effective energy into respective main- and side-chain contributions. The key advantage of this method is that the same set of conformational ensembles is used to characterize all the sites without the introduction of physical modifications of chemical groups. Therefore, the stability contribution from each group is quantified concurrently in the presence of unperturbed interactions, such that the numerical comparison between the stabilizing sites is more consistent. This method enabled us to identify specific residues whose backbone and side-chain interactions are critical to folding stability, which are in agreement with the previous mutagenesis studies. Moreover, our analysis elucidates how the backbone hydrogen bonds and the side-chain packing of the hydrophobic clusters cooperatively determine the folding stability. We were able to analyze whether protein-protein or protein-water interaction dominated the folding stability of the folding from a determination of folding free energy. Successful determination of folding free energy was accomplished by the explicit treatment of the solvation model in protein simulations and the adaptation of a validated solvation model. This study analyzed the thermodynamic origin with a decomposition into the residue-level of the WW domain such that the research laid a foundation to a further understanding of the intrinsic molecular factors and protein engineering.

A comparative study of the influence of two popular solvation models on protein structure and thermodynamics was conducted. We utilized a recently developed site-specific thermodynamic analysis method to identify the critical sites that lead to the folding stability discrepancy of TIP3P and GBSA water models by decomposing the free energy component at a single amino acid resolution. A key advantage of this method is that stability contributions can be decomposed into main and side chains without the introduction of perturbation, allowing a systematic comparison among amino acid residues to be made more consistent. A detailed analysis of the structure-thermodynamics relationship revealed that the structural origin of the under-stabilized WW domain in the GBSA solvent simulation is mainly due to the presence of native salt-bridges, followed by hydrophobic clusters, instead of β-sheet backbones. In contrast, the folding stability tendency of HP-36 is relatively accurately represented by the GBSA solvent simulation, owing to the lack of the native salt-bridge, supported by a strong correlation according to the Pearson correlation analysis. The reason for thermodynamic discrepancies in key neutral residues critical to folding is an underestimation of solvent effect or the dehydration penalty in the GBSA solvent simulation. This study investigated the structural and thermodynamic influence of solvation models on the folding of representative α-helix and β-hairpin proteins, which lays the foundation for developing more accurate methods for protein folding simulations.

SUPPLEMENTARY INFORMATION

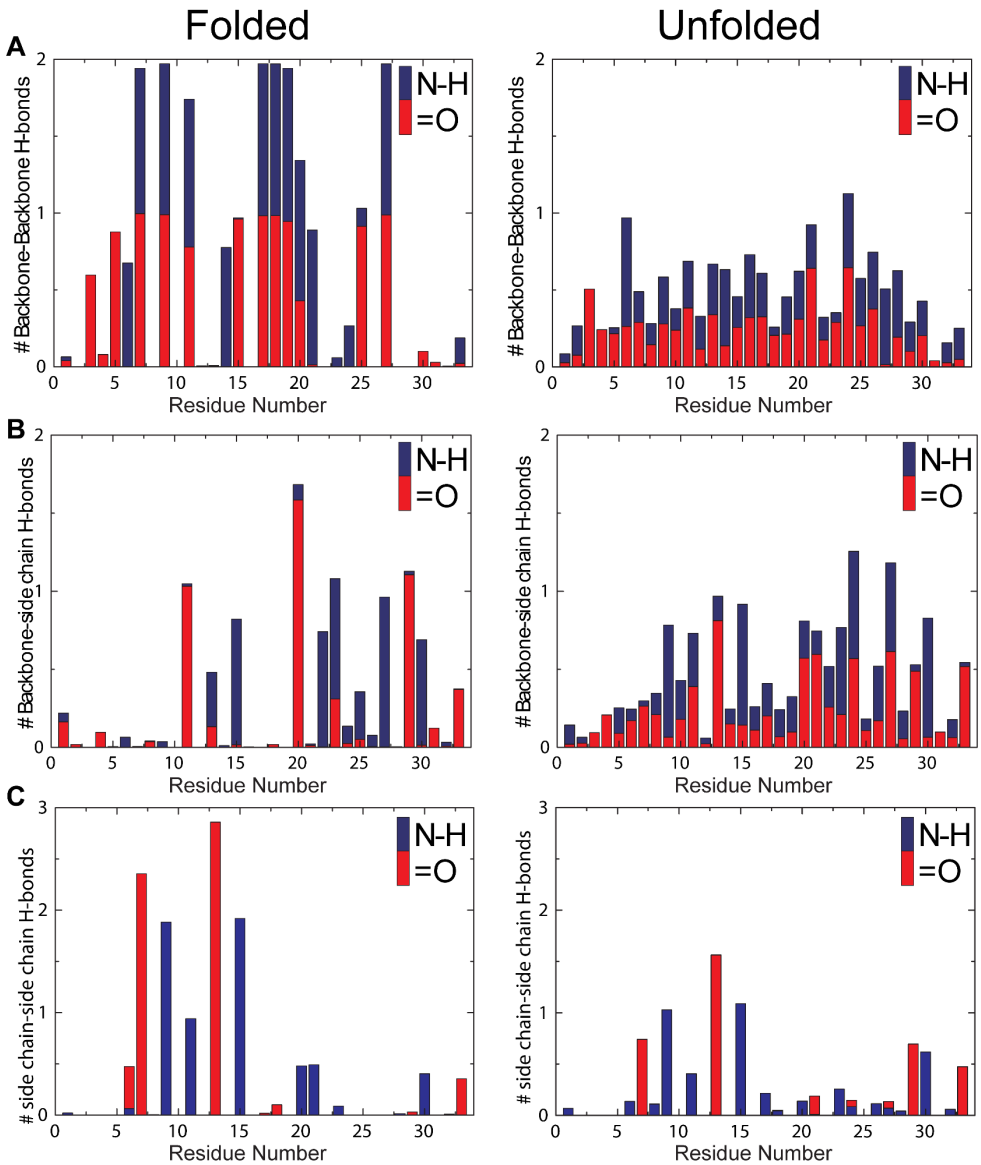


Figure S2.1: The number of hydrogen bonds (H-bonds) per residue. Three types of H-bonds are (A) between backbones, (B) between a backbone and a side-chain, and (C) between side-chains. The H-bond donor (blue) and the H-bond acceptor (red) are stacked within an amino acid residue.

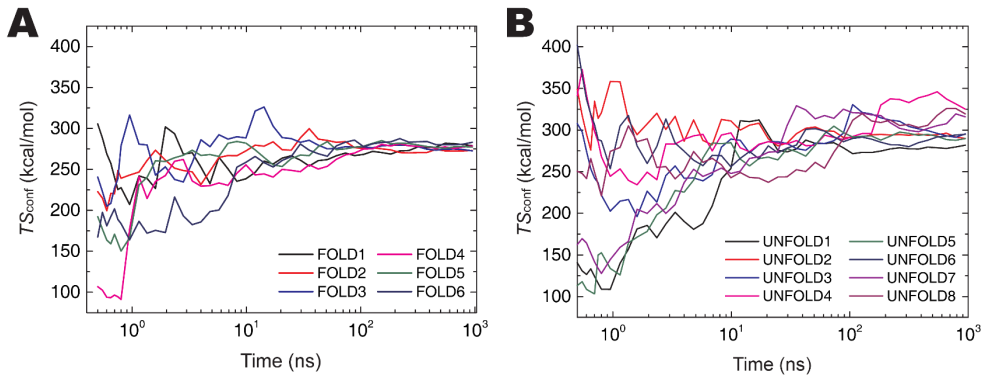


Figure S2.2: The running averages of configurational entropy TS_{conf} over time in log scale. (A) Six independent folded-state trajectories based on protein conformations taken from 1 μs simulation and (B) eight independent unfolded-state trajectories based on conformations taken from the last 1 μs specified by different colors.

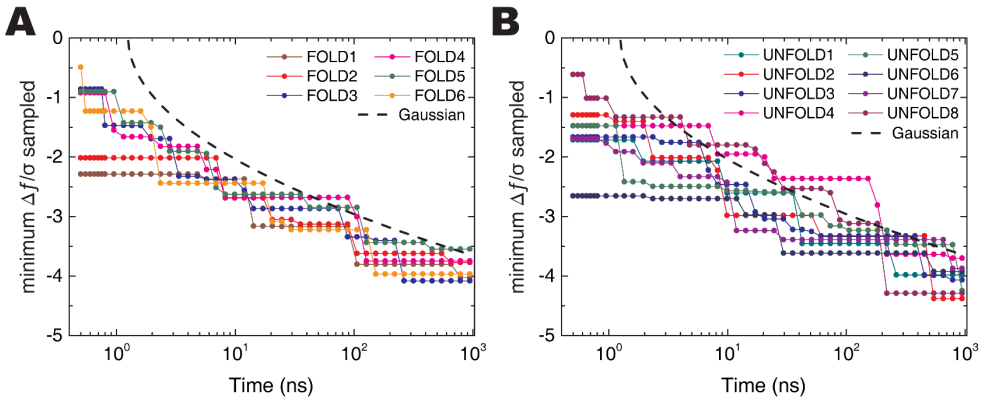


Figure S2.3: The running minimums of the adjusted effective energy $\Delta f = f - \bar{f}$ divided by trajectory standard deviation (σ) vs. time in each of (A) the folded- and (B) unfolded-state trajectories. The dashed line denotes the minimums that are sampled from ideal Gaussian statistics. A total of 46 minimums (circle) per trajectory is generated.

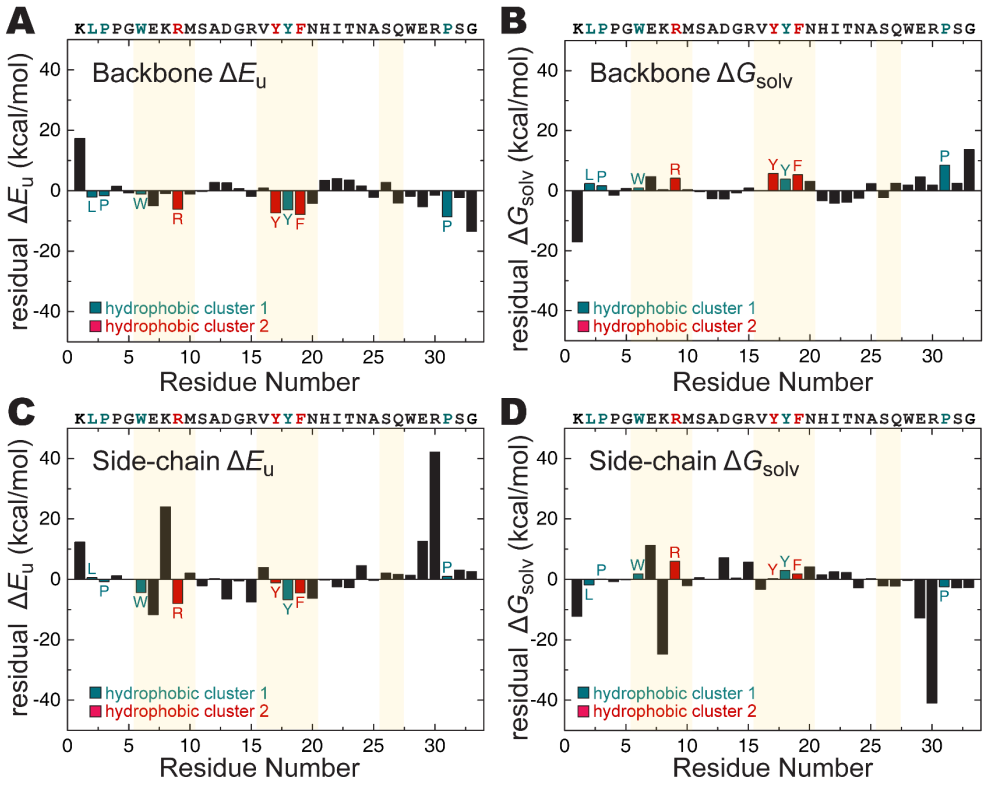


Figure S2.4: Thermodynamic analyses for each backbone and side-chain of the WW domain. (A, B) The intra-protein potential energy upon folding ΔE_u = E_u (folded) – E_u (unfolded) and the solvation free energy upon folding ΔG_{solv} = G_{solv} (folded) – G_{solv} (unfolded) for every constituent backbone (C, D) and for every side-chain residues. The residues that comprise hydrophobic cluster 1 and hydrophobic cluster 2 are indicated by dark cyan and dark pink, respectively. The yellow strips demarcate the three β-strand regions.

Table S2.1: Native contacts fraction in the folded and unfolded states of WW domain and HP-36

	Native Contacts Fraction Q	
	WW domain	HP36
X-ray/NMR	1	1
	folded-state trajectories	
FOLD1	0.962	0.765
FOLD2	0.959	0.697
FOLD3	0.960	0.771
FOLD4	0.958	N/A
FOLD5	0.955	N/A
FOLD6	0.955	N/A
Average[a]	0.958 ± 0.001	0.745 ± 0.019
	unfolded-state trajectories	
UNFOLD1	0.086	0.220
UNFOLD2	0.080	0.171
UNFOLD3	0.200	0.195
UNFOLD4	0.197	0.166
UNFOLD5	0.190	0.255
UNFOLD6	0.092	0.073
UNFOLD7	0.167	0.176
UNFOLD8	0.079	0.184
UNFOLD9	N/A	0.215
Average[a]	0.137 ± 0.019	0.184 ± 0.016

[a] Average ± standard error

Table S2.2: Per-residue structural data of Pin WW upon folding

Res. Number	ΔNumber of H-bonds [a]			ΔNumber of SC contacts [b]	
	MC–MC	MC–SC	SC–SC	MC–SC	SC–SC
K1	0.0	0.1	0.0	0.0	–0.1
L2	–0.1	0.0	0.0	–0.1	1.0
P3	0.0	–0.1	0.0	–0.5	0.3
P4	–0.1	–0.1	0.0	–0.6	–0.6
G5	0.3	–0.2	0.0	0.4	0.0
W6	–0.1	–0.2	0.2	0.4	1.1
E7	0.7	–0.1	0.8	0.5	0.4
K8	–0.1	–0.2	–0.1	0.4	0.6
R9	0.7	–0.2	0.4	0.0	0.1
M10	–0.2	–0.4	0.0	–0.2	–0.5
S11	0.5	–0.3	0.2	0.0	0.6
A12	–0.2	–0.1	0.0	–0.5	–0.5
D13	–0.3	0.1	0.6	–0.6	–0.4
G14	0.1	–0.2	0.0	–0.4	0.0
R15	0.3	0.2	0.4	0.5	0.2
V16	–0.4	–0.3	0.0	0.1	–0.2
Y17	0.7	–0.3	–0.1	–0.1	0.0
Y18	0.9	–0.2	0.0	0.8	1.4
F19	0.7	–0.3	0.0	1.2	0.4
N20	0.4	0.0	0.2	0.2	0.3
H21	0.0	–0.1	0.1	–0.4	0.1
I22	–0.2	0.2	0.0	–0.2	–0.3
T23	–0.1	0.1	–0.1	–0.4	–0.4
N24	–0.4	–0.1	–0.1	0.2	–0.4
A25	0.2	0.2	0.0	0.4	0.2
S26	–0.4	–0.2	–0.1	0.7	0.4
Q27	0.7	–0.2	–0.1	0.1	0.4
W28	–0.3	–0.1	0.0	0.4	0.9
E29	–0.2	0.9	–0.3	0.0	–0.6
R30	–0.2	0.0	–0.1	0.1	0.1
P31	0.0	0.0	0.0	1.0	1.8
S32	–0.1	–0.1	–0.1	0.1	–0.2
G33	0.0	–0.1	–0.1	0.1	0.0
Total [c]	2.8 ± 0.4	–2.6 ± 0.6	1.8 ± 0.4	3.4 ± 0.9	5.9 ± 0.5

[a] The number of intra-protein H-bonds upon folding between main-chains (MC−MC), between a main-chain and a side-chain (MC−SC) and between side-chains (SC−SC); [b] The number intra-protein heavy-atom contacts involving side-chains upon folding, main-chain−side-chain contacts (MC−SC) and side-chain-side-chain contacts (SC−SC); [c] Residue total ± standard error.

Table S2.3: Tabulated thermodynamic quantities summary table obtained for both folded and unfolded state trajectories

State	E_u [a]	G_{solv} [b]	f [c]	$-TS_{conf}$ [d]	$G=f-TS_{conf}$ [e]
FOLD1	–397.7	94.5	–303.2	–278.5	–581.7
FOLD2	–400.3	97.6	–302.7	–274.1	–576.8
FOLD3	–399.3	96.0	–303.3	–273.8	–577.1
FOLD4	–394.0	90.8	–303.2	–277.8	–581.0
FOLD5	–389.5	86.6	–302.9	–276.2	–579.1
FOLD6	–392.3	90.6	–301.7	–282.0	–583.7
Average [f]	–395.5 ± 1.6	92.7 ± 1.5	–302.8 ± 0.2	–277.1 ± 1.1	–579.9 ± 1.0
UNFOLD1	–455.9	172.7	–283.3	–282.8	–566.1
UNFOLD2	–445.9	170.4	–275.5	–289.3	–564.8
UNFOLD3	–307.8	33.8	–274.0	–296.0	–570.0
UNFOLD4	–401.1	135.5	–265.6	–323.9	–589.5
UNFOLD5	–414.8	142.5	–272.3	–289.8	–562.1
UNFOLD6	–357.9	93.0	–264.9	–296.2	–561.1
UNFOLD7	–445.0	168.7	–276.3	–316.0	–592.3
UNFOLD8	–375.1	97.1	–278.0	–316.6	–594.6
Average [f]	–400.4 ± 16.9	126.7 ± 16.2	–273.7 ± 2.0	–301.3 ± 5.1	–575.0 ± 4.8
Difference [f]	4.9 ± 17.0	–34.1 ± 16.3	–29.2 ± 2.0	24.3 ± 5.2	–4.9 ± 4.9

[a] Intra-protein potential energy [kcal/mol]; [b] Solvation free energy [kcal/mol]; [c] Effective energy [kcal/mol]; [d] Configurational entropy multiplied by $-T$ [kcal/mol]; [e] Gibbs free energy [kcal/mol]; [f] Average ± standard error.

Table S2.4: Statistical analyses of the distribution function of effective energy

State	skewness	excess kurtosis
FOLD1	0.067	0.009
FOLD2	0.088	0.031
FOLD3	0.081	0.011
FOLD4	0.041	–0.029
FOLD5	0.084	–0.038
FOLD6	0.072	0.011
UNFOLD1	0.080	0.018
UNFOLD2	0.063	0.009
UNFOLD3	0.077	0.003
UNFOLD4	0.107	0.031
UNFOLD5	0.059	0.013
UNFOLD6	0.039	–0.031
UNFOLD7	0.047	–0.050
UNFOLD8	0.086	0.017

Table S2.5: Tabulated effective free energy by individual amino acid residue of Pin WW separated by the backbone atoms

Res. Number	ΔE_u [a]	ΔG_{solv} [b]	Δf Backbone only [c]	Δf order	$\Delta\Delta G$ [d]
1	17.32 ± 0.99	–17.05 ± 0.95	0.27 ± 0.07	-	N/A
2	–2.10 ± 0.16	2.39 ± 0.12	0.29 ± 0.05	16	18
3	–1.62 ± 0.09	1.62 ± 0.09	–0.01 ± 0.04	-	N/A
4	1.51 ± 0.28	–1.54 ± 0.27	–0.02 ± 0.04	-	N/A
5	–0.74 ± 0.31	0.73 ± 0.25	0.00 ± 0.08	-	N/A
6	–1.10 ± 0.23	0.91 ± 0.19	–0.19 ± 0.08	12	9
7	–4.95 ± 0.23	4.70 ± 0.22	–0.25 ± 0.05	11	7
8	–0.93 ± 0.21	0.38 ± 0.21	–0.56 ± 0.03	9	13
9	–6.12 ± 0.23	4.16 ± 0.20	–1.96 ± 0.13	3	3
10	–1.12 ± 0.35	0.37 ± 0.31	–0.75 ± 0.14	8	17
11	–0.25 ± 0.31	–0.29 ± 0.30	–0.54 ± 0.04	10	11
12	2.74 ± 0.32	–2.62 ± 0.28	0.13 ± 0.08	-	N/A
13	2.57 ± 0.35	–2.74 ± 0.37	–0.17 ± 0.05	-	N/A
14	0.72 ± 0.21	–0.68 ± 0.16	0.04 ± 0.08	-	N/A
15	–1.87 ± 0.19	0.88 ± 0.20	–0.99 ± 0.07	-	N/A
16	0.91 ± 0.33	–0.01 ± 0.30	0.89 ± 0.07	18	15
17	–7.31 ± 0.29	5.75 ± 0.19	–1.56 ± 0.14	5	5*
18	–6.34 ± 0.18	3.88 ± 0.16	–2.46 ± 0.07	2	8
19	–7.83 ± 0.23	5.32 ± 0.21	–2.52 ± 0.06	1	2
20	–4.15 ± 0.38	3.13 ± 0.31	–1.02 ± 0.10	6	1
21	3.47 ± 0.27	–3.31 ± 0.21	0.16 ± 0.09	15	10
22	4.02 ± 0.29	–4.18 ± 0.24	–0.16 ± 0.06	-	N/A
23	3.52 ± 0.47	–3.87 ± 0.38	–0.34 ± 0.09	-	N/A
24	1.58 ± 0.17	–2.50 ± 0.20	–0.92 ± 0.11	7	5*
25	–2.18 ± 0.28	2.31 ± 0.28	0.14 ± 0.04	14	14
26	2.74 ± 0.59	–2.28 ± 0.51	0.46 ± 0.09	17	12
27	–4.09 ± 0.44	2.50 ± 0.44	–1.59 ± 0.16	4	4
28	–1.83 ± 0.38	1.81 ± 0.36	–0.02 ± 0.14	13	16
29	–5.25 ± 0.41	4.62 ± 0.39	–0.63 ± 0.07	-	N/A
30	–1.51 ± 0.28	1.83 ± 0.28	0.32 ± 0.04	-	N/A
31	–8.60 ± 0.25	8.53 ± 0.26	–0.07 ± 0.03	-	N/A
32	–2.27 ± 0.22	2.46 ± 0.21	0.19 ± 0.02	-	N/A
33	–13.44 ± 1.60	13.73 ± 1.62	0.28 ± 0.03	-	N/A
Total [e]	–44.49 ± 13.5	30.93 ± 13.5	–13.6 ± 1.4		

[a] Intra-protein potential energy contribution from each backbone upon folding [kcal/mol]; [b] Solvation free energy change upon folding [kcal/mol]; [c] Effective energy change upon folding [kcal/mol]; [d] From ref. 11; [e] Average ± standard error; * Identical values.

Table S2.6: Tabulated effective free energy by individual amino acid residue of Pin WW separated by the side-chain atoms.

Res. Num.	ΔE_u [a]	ΔG_{solv} [b]	Δf Side-chain Only [c]	Δf order	$\Delta\Delta G$ [d]
K1	12.38 ± 1.17	–12.22 ± 1.12	0.17 ± 0.06	21	28
L2	0.56 ± 0.16	–1.82 ± 0.12	–1.26 ± 0.07	9	7
P3	–0.77 ± 0.25	0.14 ± 0.19	–0.63 ± 0.07	12	10
P4	1.22 ± 0.29	–0.71 ± 0.24	0.51 ± 0.10	23	25
G5	–0.01 ± 0.09	–0.14 ± 0.10	–0.15 ± 0.05	17	12
W6	–4.42 ± 0.38	1.82 ± 0.17	–2.59 ± 0.21	3	1*
E7	–11.68 ± 0.80	11.23 ± 0.79	–0.46 ± 0.05	14	20
K8	24.01 ± 1.02	–24.74 ± 0.97	–0.72 ± 0.07	11	23
R9	–8.04 ± 1.32	6.00 ± 1.18	–2.04 ± 0.16	5	8
M10	2.03 ± 0.34	–2.06 ± 0.20	–0.02 ± 0.16	19	14
S11	–2.10 ± 0.14	0.59 ± 0.08	–1.51 ± 0.12	7	16
A12	0.30 ± 0.15	0.09 ± 0.11	0.38 ± 0.08	-	N/A
D13	–6.50 ± 1.80	7.15 ± 1.84	0.64 ± 0.07	-	N/A
G14	–0.51 ± 0.10	0.46 ± 0.09	–0.04 ± 0.03	18	11
R15	–7.48 ± 1.18	5.72 ± 1.06	–1.76 ± 0.19	6	17
V16	3.90 ± 0.23	–3.26 ± 0.23	0.64 ± 0.10	24	26
Y17	–1.19 ± 0.38	0.32 ± 0.25	–0.87 ± 0.15	10	6
Y18	–6.72 ± 0.28	2.99 ± 0.24	–3.74 ± 0.05	1	1*
F19	–4.50 ± 0.25	1.82 ± 0.17	–2.68 ± 0.10	2	5
N20	–6.25 ± 0.52	4.17 ± 0.35	–2.08 ± 0.17	4	1*
H21	–0.26 ± 0.21	1.51 ± 0.14	1.25 ± 0.14	27	22
I22	–2.46 ± 0.23	2.51 ± 0.13	0.05 ± 0.12	20	21
T23	–2.73 ± 0.19	2.31 ± 0.16	–0.42 ± 0.05	15	9
N24	4.60 ± 0.32	–2.80 ± 0.21	1.80 ± 0.14	28	18
A25	–0.35 ± 0.10	0.21 ± 0.08	–0.15 ± 0.06	-	N/A
S26	2.05 ± 0.33	–2.17 ± 0.26	–0.11 ± 0.07	-	N/A
Q27	1.63 ± 0.59	–2.20 ± 0.42	–0.57 ± 0.18	13	15
W28	1.35 ± 0.44	–0.35 ± 0.32	0.99 ± 0.15	25	24
E29	12.59 ± 1.61	–12.81 ± 1.65	–0.23 ± 0.05	16	13
R30	42.15 ± 1.06	–40.93 ± 0.95	1.22 ± 0.17	26	19
P31	0.98 ± 0.23	–2.49 ± 0.25	–1.50 ± 0.07	8	1*
S32	3.05 ± 0.15	–2.79 ± 0.12	0.22 ± 0.04	22	27
G33	2.60 ± 0.24	–2.66 ± 0.25	–0.06 ± 0.01	-	N/A
Total [e]	49.40 ± 9.64	–65.11 ± 8.91	–15.7 ± 1.6		

[a] Intra-protein potential energy contribution from each side-chain upon folding [kcal/mol]; [b] Solvation free energy change upon folding [kcal/mol]; [c] Effective energy change upon folding [kcal/mol]; [d] From ref. 12; [e] Average ± standard error; * Remain unfolded upon alanine mutation.

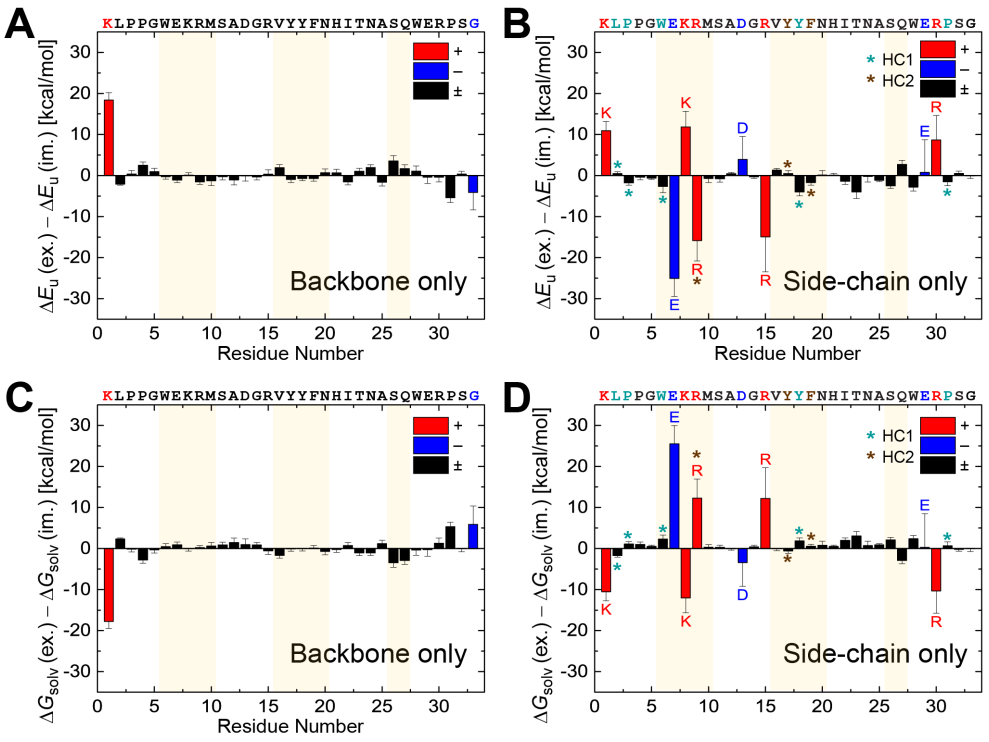


Figure S3.1: Thermodynamic analyses for each main and side chain of the WW domain. The main and side chain contributions to (A, B) the folding intra-protein potential energy difference ΔE_u(ex.) – ΔE_u(im.), (C, D) the folding solvation free energy difference ΔG_{solv}(ex.) – ΔG_{solv}(im.).

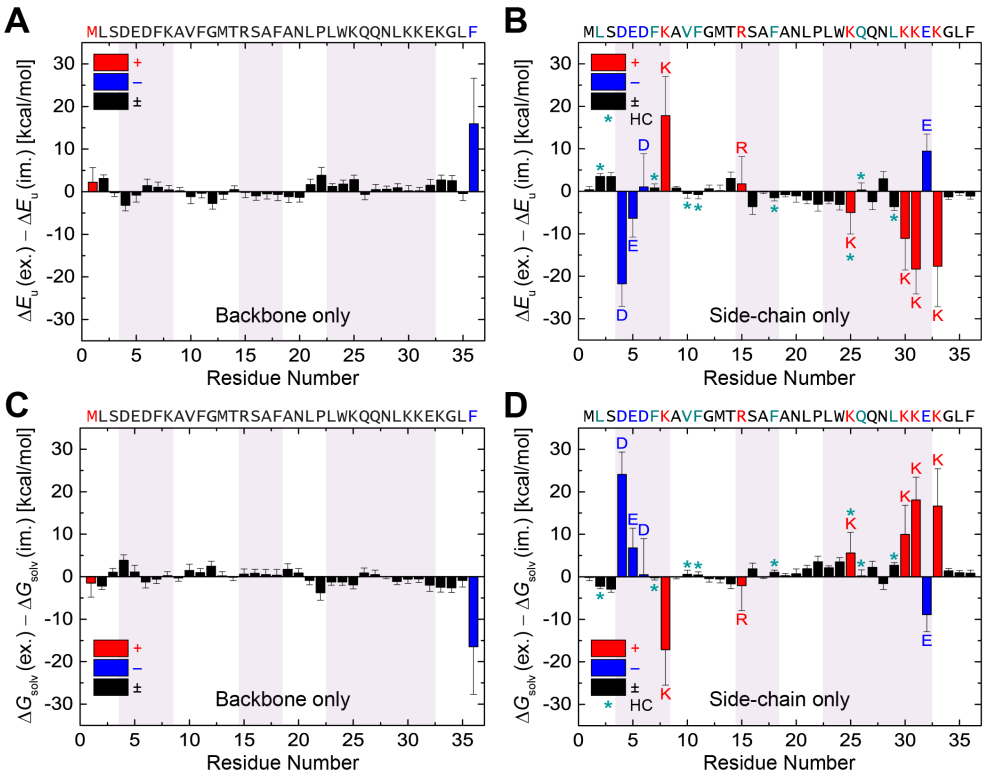


Figure S3.2: Thermodynamic analyses for each main and side chain of the HP-36. The main and side chain contributions to (A, B) the folding intra-protein potential energy difference ΔE_u(ex.) – ΔE_u(im.), (C, D) the folding solvation free energy difference ΔG_{solv}(ex.) – ΔG_{solv}(im.).

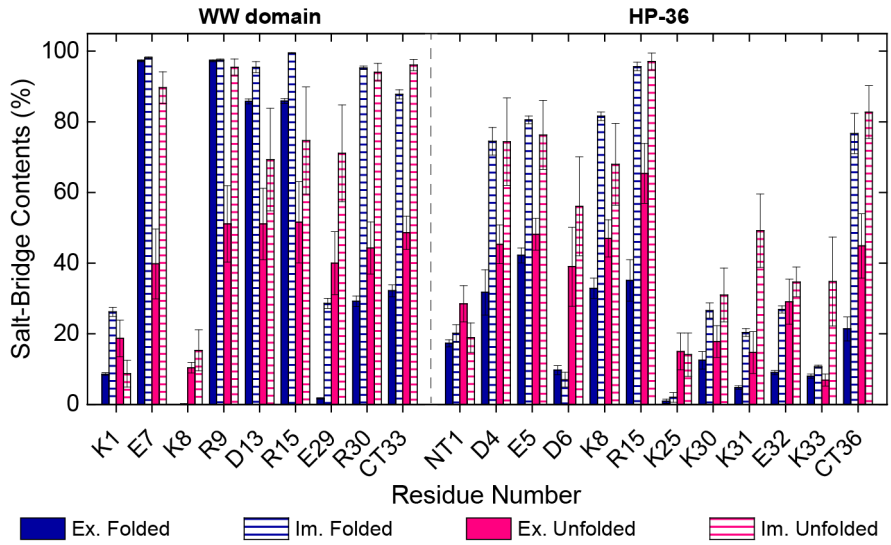


Figure S3.3. Per-residue analysis of salt-bridge contents of WW domain and HP-36.

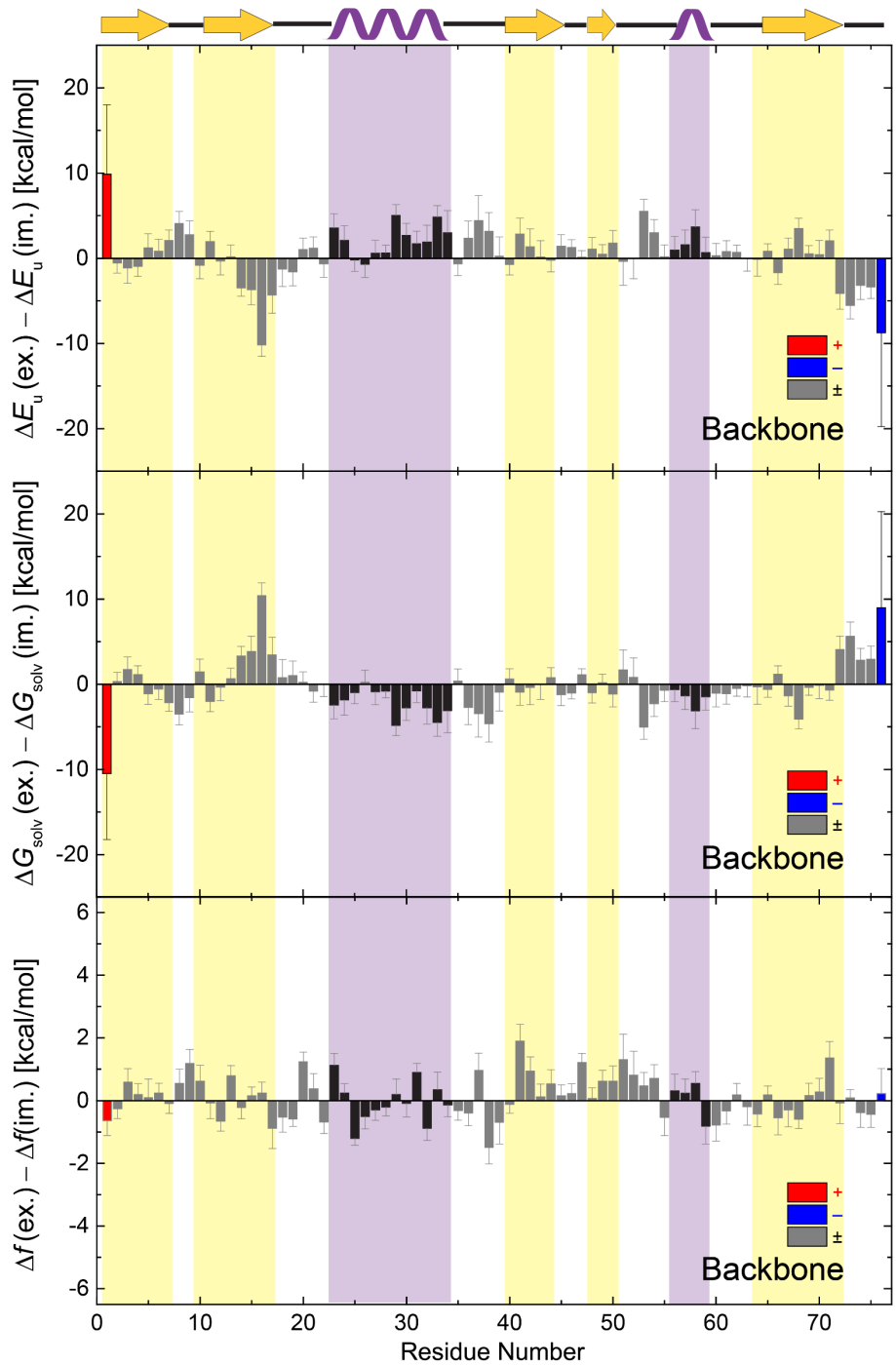


Figure S3.4. Thermodynamic analyses for each main chain residues of ubiquitin. The main and side chain contributions to the folding intra-protein potential energy difference ΔE_u(ex.) – ΔE_u(im.), the folding solvation free energy difference ΔG_{solv}(ex.) – ΔG_{solv}(im.), and the folding effective energy Δf (ex.) – Δf(im.).

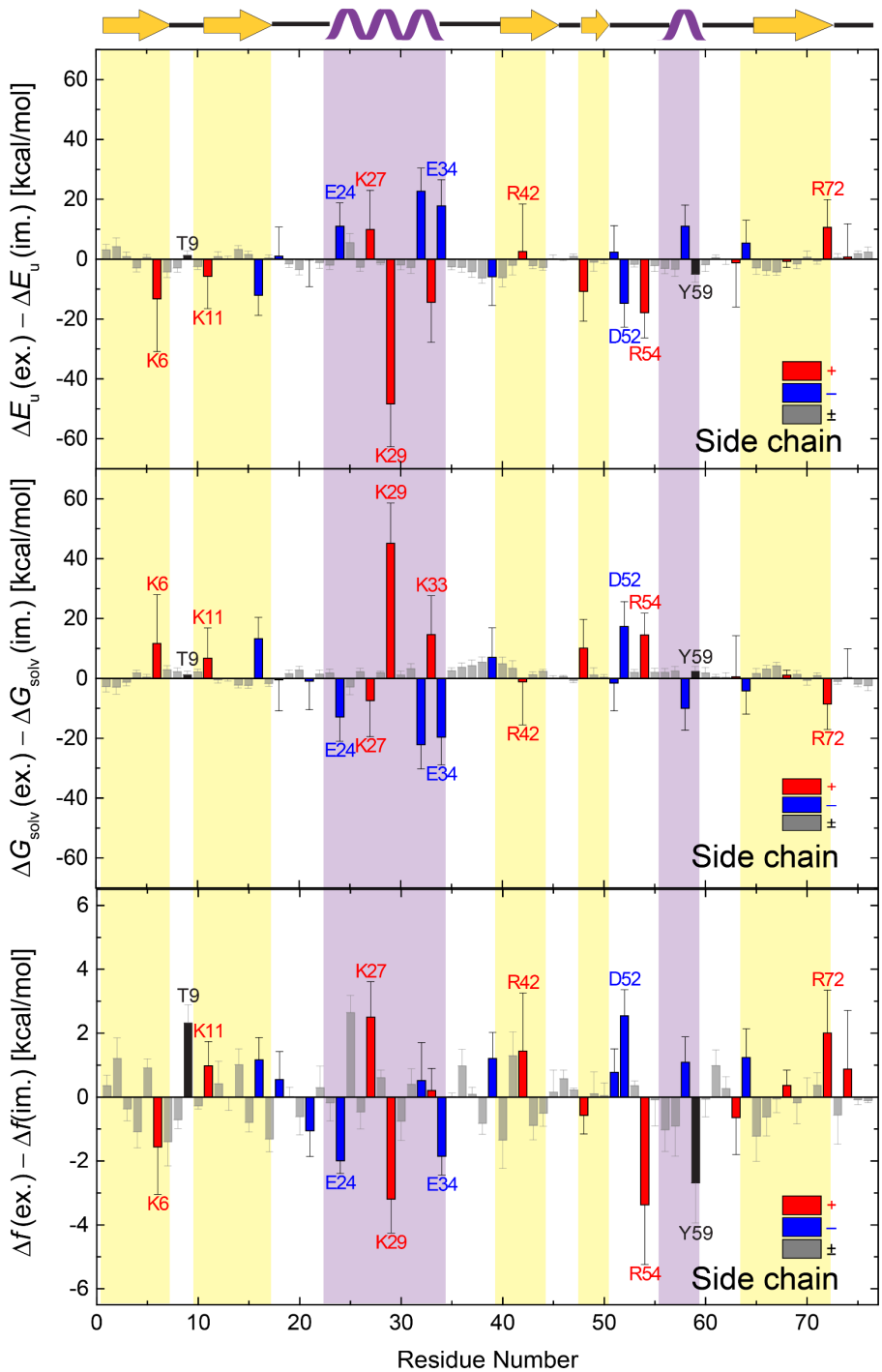


Figure S3.5. Thermodynamic analyses for each side chain residues of the ubiquitin. The main and side chain contributions to the folding intra-protein potential energy difference ΔE_u(ex.) – ΔE_u(im.), the folding solvation free energy difference ΔG_{solv}(ex.) – ΔG_{solv}(im.), and the folding effective energy Δf (ex.) – Δf(im.).

Table S3.1: Native contacts fraction in the folded and unfolded states of WW domain and HP-36

	Native Contacts Fraction Q			
	WW domain		HP36	
X-ray/NMR	1		1	
AlphaFold	0.965		0.868	
	Explicit	Implicit	Explicit	Implicit
	folded-state trajectories			
FOLD1	0.962	0.896	0.765	0.796
FOLD2	0.959	0.914	0.697	0.761
FOLD3	0.960	0.886	0.771	0.791
FOLD4	0.958	0.900	N/A	0.777
FOLD5	0.955	N/A	N/A	N/A
FOLD6	0.955	N/A	N/A	N/A
Average[a]	0.958 ± 0.001	0.899 ± 0.005	0.745 ± 0.019	0.781 ± 0.007
	unfolded-state trajectories			
UNFOLD1	0.086	0.219	0.220	0.211
UNFOLD2	0.080	0.093	0.171	0.333
UNFOLD3	0.200	0.116	0.195	0.181
UNFOLD4	0.197	0.150	0.166	0.121
UNFOLD5	0.190	0.095	0.255	0.318
UNFOLD6	0.092	0.032	0.073	0.189
UNFOLD7	0.167	0.059	0.176	N/A
UNFOLD8	0.079	0.107	0.184	N/A
UNFOLD9	N/A	N/A	0.215	N/A
Average[a]	0.137 ± 0.019	0.107 ± 0.019	0.184 ± 0.016	0.225 ± 0.031

[a] Average ± standard error

Table S3.2: Native structural characteristics in the folded and unfolded states of WW domain and HP-36

			Secondary Structure Contents (%) [b]			Heavy atom contacts (%) [c]	
		C_α RMSD [a]	β-1	β-2	β-3	HC 1	HC 2
WW domain	X-ray Structure PDB ID: 2F21	0	100	100	100	100	100
	AlphaFold	0.59	100	100	100	86.8	88.2
	Ex. Folded [d]	0.90 ± 0.01	99.7 ± 0.1	99.9 ± 0.1	99.9 ± 0.1	83.1 ± 0.1	86.8 ± 0.2
	Im. Folded [d]	1.30 ± 0.01	98.5 ± 0.2	99.3 ± 0.1	97.6 ± 0.6	55.9 ± 1.3	74.2 ± 0.5
	Ex. Unfolded [d]	7.20 ± 0.27	11.5 ± 6.3	7.8 ± 4.0	1.3 ± 0.8	9.5 ± 4.0	9.0 ± 5.1
	Im. Unfolded [d]	7.46 ± 0.41	4.4 ± 2.1	5.9 ± 3.8	0.0 ± 0.0	5.8 ± 1.9	2.2 ± 1.3
		C_α RMSD [a]	helix-1	helix-2	helix-3	HC	
HP-36	NMR Structure PDB ID: 1VII	0	100	100	100	100	
	AlphaFold	1.20	100	100	100	53.0	
	Ex. Folded [d]	2.47 ± 0.12	78.5 ± 2.7	98.8 ± 0.1	93.2 ± 3.5	37.9 ± 1.6	
	Im. Folded [d]	2.03 ± 0.04	89.4 ± 1.1	98.1 ± 0.2	97.4 ± 0.4	38.5 ± 1.2	
	Ex. Unfolded [d]	7.11 ± 0.21	42.4 ± 5.4	8.9 ± 2.6	21.8 ± 6.4	2.9 ± 1.0	
	Im. Unfolded [d]	6.76 ± 0.32	38.8 ± 11.1	27.6 ± 11.5	31.7 ± 9.1	4.4 ± 1.7	

[a] Root-mean-square deviations (Å) for C_α atoms; [b] Average population (%) of the β-strand formations in β-1 (W6-M10), β-2 (V16-N20) and β-3 (S26-Q27); [c] Average population (%) of side-chain heavy atom contacts in hydrophobic cluster 1 (HC 1; L2, P3, W6, Y18, and P31) and hydrophobic cluster 2 (HC 2; R9, Y17, and F19); [d] Average ± standard error. For the HP-36, [b] Average population (%) of the α-helix formation in helix-1 (D4-F8), helix-2 (R15-F18), and helix-3 (L23-E32); [c] hydrophobic cluster residues (HC; L2, F8, V10, F11, F18, and L29)

Table S3.3: Non-native structural characteristics in the folded and unfolded states of Pin WW and HP-36

		Number of H-bonds [a]			Number of side-chain contacts [b]		Salt-bridge contents (%)
		MC−MC	MC−SC	SC−SC	MC−SC	SC−SC	
WW domain	X-ray Structure PDB ID: 2F21	11	8	2	27	25	6.1
	AlphaFold	11	2	4	26	22	4.2
	Ex. Folded [d]	10.7 ± 0.1	5.1 ± 0.1	6.3 ± 0.1	22.9 ± 0.1	23.1 ± 0.1	13.3 ± 0.1
	Im. Folded [d]	10.0 ± 0.0	6.3 ± 0.1	7.5 ± 0.1	20.4 ± 0.0	20.8 ± 0.2	19.1 ± 0.1
	Ex. Unfolded [d]	7.9 ± 0.4	7.7 ± 0.6	4.5 ± 0.4	19.4 ± 0.8	17.2 ± 0.5	10.8 ± 0.4
	Im. Unfolded [d]	6.7 ± 0.5	10.1 ± 0.7	9.0 ± 0.9	20.1 ± 0.7	18.1 ± 0.9	18.6 ± 0.3
HP-36	NMR Structure PDB ID: 1VII	12	1	0	20	21	0
	AlphaFold	19.4	5.2	0.8	20.8	29.9	5.5
	Ex. Folded [d]	17.9 ± 0.4	3.5 ± 0.1	2.1 ± 0.1	18.0 ± 0.1	23.6 ± 0.6	6.3 ± 0.3
	Im. Folded [d]	16.4 ± 0.1	5.5 ± 0.2	4.1 ± 0.1	18.2 ± 0.2	24.7 ± 0.1	14.5 ± 0.2
	Ex. Unfolded [d]	10.4 ± 0.6	6.7 ± 0.4	3.4 ± 0.5	19.5 ± 0.8	19.2 ± 0.5	11.2 ± 0.3
	Im. Unfolded [d]	9.4 ± 0.4	7.9 ± 0.5	6.5 ± 0.6	20.1 ± 0.9	21.6 ± 0.4	17.7 ± 0.4

[a] The number of intra-protein H-bonds between main-chains (MC−MC), between a main-chain and a side-chain (MC−SC) and between side-chains (SC−SC); [b] The number intra-protein heavy-atom contacts involving side-chains in main-chain−side-chain contacts (MC−SC) and side-chain-side-chain contacts (SC−SC); [d] Average ± standard error.

Table S3.4: Per-residue structural analysis data of Pin WW upon folding

Res. Number	ΔΔNumber of H-bonds [a]			ΔΔNumber of SC contacts [b]	
	MC–MC	MC–SC, SC-MC	SC–SC	MC–SC	SC–SC
K1	0.0	0.0	0.0	0.0	0.0
L2	–0.1	0.0	0.0	–0.2	0.3
P3	0.0	0.0	0.0	–0.3	0.1
P4	–0.1	–0.1	0.0	–0.3	–0.3
G5	0.0	0.0	0.0	–0.1	0.0
W6	–0.2	0.1	0.0	0.2	–0.3
E7	0.1	0.6	0.9	0.2	0.2
K8	–0.1	0.0	0.0	0.0	0.1
R9	–0.2	0.1	1.0	–0.3	0.3
M10	0.0	0.0	0.0	–0.1	0.1
S11	0.1	0.3	–0.1	0.1	0.0
A12	–0.1	0.1	0.0	–0.1	0.0
D13	–0.2	–0.2	0.2	–0.2	0.1
G14	0.1	0.0	0.0	0.2	0.0
R15	0.1	0.5	0.4	0.3	0.7
V16	–0.2	0.0	0.0	0.0	–0.4
Y17	0.0	0.0	–0.1	–0.2	–0.4
Y18	0.2	–0.1	0.0	0.4	0.4
F19	0.1	–0.1	0.0	0.2	–0.1
N20	–0.2	–0.2	0.0	0.1	0.0
H21	0.0	–0.1	0.1	0.0	–0.1
I22	0.1	0.1	0.0	0.1	–0.2
T23	0.1	0.0	0.0	0.1	0.0
N24	–0.2	–0.2	0.1	0.1	0.1
A25	0.3	0.0	0.0	0.4	0.4
S26	–0.1	0.2	0.0	0.2	0.2
Q27	0.0	–0.3	–0.1	–0.1	0.0
W28	–0.1	0.1	0.0	0.1	0.5
E29	0.0	0.1	0.5	0.2	0.2
R30	–0.1	0.2	0.2	0.2	0.8
P31	0.0	0.1	0.0	0.6	0.6
S32	0.0	0.0	0.0	0.4	–0.1
G33	0.0	0.3	0.3	0.8	0.0
Total [c]	–0.5	1.2	3.3	3.1	3.2

[a] The number of intra-protein H-bonds upon folding between main-chains (MC−MC), between a main-chain and a side-chain (MC−SC/SC−MC) and between side-chains (SC−SC); [b] The number intra-protein heavy-atom contacts involving side-chains upon folding, main-chain−side-chain contacts (MC−SC) and side-chain-side-chain contacts (SC−SC)

Table S3.5: Per-residue structural analysis data of HP-36 upon folding

Res. Number	ΔΔNumber of H-bonds [a]			ΔΔNumber of SC contacts [b]	
	MC–MC	MC–SC, SC-MC	SC–SC	MC–SC	SC–SC
M1	–0.1	0.0	0.0	–0.1	0.2
L2	0.0	–0.2	0.0	–0.5	–1.1
S3	–0.3	0.0	–0.1	0.0	0.1
D4	0.1	0.2	0.6	0.4	0.8
E5	0.0	0.0	0.1	0.3	0.3
D6	–0.2	–0.4	0.0	–0.1	0.0
F7	0.2	0.0	0.0	–0.1	0.0
K8	–0.1	0.0	–0.1	0.3	0.2
A9	0.0	–0.1	0.0	–0.2	–0.3
V10	0.1	0.1	0.0	0.0	–0.1
F11	0.0	–0.1	0.0	0.3	0.0
G12	0.1	0.1	0.0	0.4	0.0
M13	0.0	0.1	0.0	–0.2	0.3
T14	–0.1	0.0	–0.3	–0.2	0.1
R15	0.1	0.0	0.1	0.1	0.2
S16	0.1	0.2	0.3	0.1	0.2
A17	–0.1	0.1	0.0	0.1	0.1
F18	0.2	0.2	0.0	0.3	0.1
A19	0.1	0.0	0.0	0.4	0.5
N20	0.1	–0.1	0.1	0.0	–0.1
L21	0.1	–0.2	0.0	0.0	–0.2
P22	–0.2	0.1	0.0	0.2	0.1
L23	–0.4	0.1	0.0	0.6	0.4
W24	0.2	–0.2	0.0	0.0	0.1
K25	–0.1	0.0	0.0	–0.3	–0.2
Q26	0.0	0.0	0.0	–0.1	–0.3
Q27	0.0	0.1	0.1	–0.1	0.1
N28	0.0	–0.6	0.1	–0.2	0.3
L29	0.1	0.1	0.0	–0.3	–0.4
K30	0.2	–0.2	0.1	0.4	0.3
K31	0.4	0.0	0.1	0.2	0.4
E32	–0.1	0.0	–0.2	–0.4	–0.3
K33	–0.1	0.0	0.1	–0.5	0.0
G34	0.1	0.0	0.0	–0.2	0.0
L35	0.3	–0.1	0.0	–0.1	–0.4
F36	–0.2	0.1	0.0	0.0	0.0
Total	0.5	–0.7	1.1	0.3	1.4

[a] The number of intra-protein H-bonds upon folding between main-chains (MC− MC), between a main-chain and a side-chain (MC−SC/SC−MC) and between side-chains (SC−SC);
[b] The number intra-protein heavy-atom contacts involving side-chains upon folding, main-chain−side-chain contacts (MC−SC) and side-chain-side-chain contacts (SC−SC)

Table S3.6: Tabulated effective free energy by individual amino acid residue of WW domain separated by the backbone atoms

Res. Number	$\Delta\Delta E_u$ [a]	$\Delta\Delta G_{solv}$ [b]	$\Delta\Delta f$ Backbone only [c]
1	18.4	–17.8	0.6
2	–2.1	2.3	0.2
3	0.3	–0.2	0.2
4	2.5	–2.8	–0.3
5	0.9	–0.3	0.6
6	–0.3	0.5	0.2
7	–1.1	0.9	–0.2
8	0.1	0.0	0.1
9	–1.6	0.2	–1.3
10	–1.3	0.6	–0.7
11	–0.3	0.9	0.6
12	–1.1	1.5	0.4
13	–0.1	1.0	0.9
14	–0.3	0.9	0.5
15	0.3	–0.5	–0.2
16	1.9	–1.7	0.3
17	–0.9	–0.1	–1.0
18	–0.7	0.0	–0.7
19	–0.7	0.1	–0.6
20	0.7	–0.7	–0.1
21	0.6	–0.2	0.4
22	–1.5	0.7	–0.8
23	1.0	–1.1	–0.1
24	1.9	–1.1	0.9
25	–1.6	1.2	–0.4
26	3.5	–3.5	0.1
27	1.7	–2.9	–1.2
28	1.1	–0.4	0.7
29	–0.4	–0.3	–0.7
30	–0.4	1.3	0.9
31	–5.4	5.3	–0.1
32	0.3	0.0	0.3
33	–4.1	5.9	1.7
Total [e]	11.4	–10.3	1.1

[a] Intra-protein potential energy contribution from each backbone upon folding [kcal/mol]; [b] Solvation free energy change upon folding [kcal/mol]; [c] Effective energy change upon folding [kcal/mol]

Table S3.7: Tabulated effective free energy by individual amino acid residue of Pin WW separated by the side-chain atoms.

Res. Number	$\Delta\Delta E_u$ [a]	$\Delta\Delta G_{solv}$ [b]	$\Delta\Delta f$ Side-chain Only [c]
K1	10.9	–10.5	0.4
L2	0.5	–1.7	–1.2
P3	–1.8	1.1	–0.7
P4	–0.4	1.0	0.6
G5	–0.6	0.5	–0.1
W6	–2.7	2.3	–0.3
E7	–25.1	25.5	0.5
K8	11.8	–12.0	–0.2
R9	–15.9	12.3	–3.6
M10	–0.7	0.3	–0.4
S11	–0.8	0.3	–0.5
A12	0.4	–0.1	0.4
D13	3.9	–3.4	0.5
G14	–0.3	0.4	0.1
R15	–14.9	12.2	–2.7
V16	1.3	–0.1	1.3
Y17	0.5	–0.6	–0.1
Y18	–4.0	1.9	–2.1
F19	–1.7	0.6	–1.1
N20	0.1	0.8	0.9
H21	0.1	0.5	0.6
I22	–1.4	2.0	0.6
T23	–4.0	3.1	–0.9
N24	–0.2	0.7	0.5
A25	–1.2	0.9	–0.2
S26	–2.5	2.1	–0.4
Q27	2.7	–2.9	–0.2
W28	–2.8	2.4	–0.4
E29	0.8	0.3	1.0
R30	8.7	–10.3	–1.7
P31	–1.5	0.7	–0.8
S32	0.5	–0.2	0.3
G33	0.0	–0.1	–0.1
Total [e]	–40.3	30.0	–10.3

[a] Intra-protein potential energy contribution from each side-chain upon folding [kcal/mol]; [b] Solvation free energy change upon folding [kcal/mol]; [c] Effective energy change upon folding [kcal/mol]

Table S3.8: Tabulated effective free energy by individual amino acid residue of HP-36 separated by the backbone atoms

Res. Number	$\Delta\Delta E_u$ [a]	$\Delta\Delta G_{solv}$ [b]	$\Delta\Delta f$ Backbone only [c]
1	2.2	–1.5	0.7
2	3.1	–2.2	0.9
3	–0.2	1.0	0.8
4	–3.2	3.8	0.7
5	–0.8	1.1	0.3
6	1.4	–1.2	0.2
7	1.0	–0.6	0.4
8	0.4	0.2	0.7
9	0.2	–0.3	–0.1
10	–1.2	1.5	0.3
11	–0.4	0.9	0.6
12	–2.8	2.5	–0.3
13	–0.6	0.2	–0.4
14	0.5	–0.1	0.4
15	–0.2	0.6	0.4
16	–1.0	0.8	–0.2
17	–0.5	0.5	0.0
18	–0.6	0.3	–0.3
19	–1.1	1.8	0.6
20	–1.3	0.9	–0.4
21	1.7	–0.9	0.8
22	3.8	–3.8	0.1
23	1.2	–1.2	0.0
24	1.8	–1.2	0.6
25	2.8	–1.9	0.9
26	–0.6	0.9	0.2
27	0.5	0.5	1.0
28	0.6	–0.1	0.5
29	0.9	–1.1	–0.2
30	0.2	–0.5	–0.4
31	0.2	–0.5	–0.3
32	1.5	–2.0	–0.5
33	2.7	–2.5	0.2
34	2.6	–2.6	0.0
35	–0.4	–0.9	–1.3
36	15.9	–16.5	–0.5
Total [e]	30.4	–24.0	6.4

[a] Intra-protein potential energy contribution from each backbone upon folding [kcal/mol]; [b] Solvation free energy change upon folding [kcal/mol]; [c] Effective energy change upon folding [kcal/mol]

Table S3.9: Tabulated effective free energy by individual amino acid residue of HP-36 separated by the side-chain atoms.

Res. Number	$\Delta\Delta E_u$ [a]	$\Delta\Delta G_{solv}$ [b]	$\Delta\Delta f$ Side-chain Only [c]
M1	0.3	–0.2	0.2
L2	3.5	–2.2	1.3
S3	3.4	–2.9	0.6
D4	–21.8	24.1	2.3
E5	–6.3	6.8	0.4
D6	1.0	0.5	1.5
F7	0.8	–0.2	0.6
K8	17.8	–17.1	0.7
A9	0.8	–0.1	0.7
V10	–0.5	0.6	0.1
F11	–0.8	0.4	–0.3
G12	0.6	–0.4	0.2
M13	0.2	–0.5	–0.4
T14	3.0	–1.6	1.4
R15	1.8	–2.1	–0.3
S16	–3.6	1.9	–1.7
A17	–0.2	–0.1	–0.2
F18	–1.5	1.0	–0.5
A19	–0.8	0.3	–0.5
N20	–1.1	0.7	–0.4
L21	–2.1	1.9	–0.2
P22	–3.1	3.5	0.5
L23	–2.3	2.1	–0.2
W24	–3.1	3.5	0.4
K25	–5.0	5.6	0.6
Q26	0.3	0.2	0.5
Q27	–2.4	2.2	–0.2
N28	2.9	–1.6	1.3
L29	–3.6	2.7	–0.9
K30	–11.1	10.0	–1.1
K31	–18.3	18.1	–0.2
E32	9.5	–8.9	0.6
K33	–17.6	16.6	–1.0
G34	–1.3	1.4	0.0
L35	–0.3	1.0	0.6
F36	–1.1	0.8	–0.3
Total [e]	–62.0	68.1	6.0

[a] Intra-protein potential energy contribution from each side-chain upon folding [kcal/mol]; [b] Solvation free energy change upon folding [kcal/mol]; [c] Effective energy change upon folding [kcal/mol]

BIBLIOGRAPHY

(1) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **1990**, *29* (31), 7133–7155.

(2) Shakhnovich, E. Protein Folding Thermodynamics and Dynamics: Where Physics, Chemistry, and Biology Meet. *Chem. Rev.* **2006**, *106* (5), 1559–1588.

(3) Pace, C. N. Conformational Stability of Globular Proteins. *Trends Biochem. Sci.* **1990**, *15* (1), 14–17.

(4) Freddolino, P. L.; Park, S.; Roux, B.; Schulten, K. Force Field Bias in Protein Folding Simulations. *Biophys. J.* **2009**, *96* (9), 3772–3780.

(5) Lazaridis, T.; Karplus, M. "New View" of Protein Folding Reconciled with the Old through Multiple Unfolding Simulations. *Science* **1997**, *278* (5345), 1928–1931.

(6) Chong, S. H.; Ham, S. Dynamics of Hydration Water Plays a Key Role in Determining the Binding Thermodynamics of Protein Complexes. *Sci. Rep.* **2017**, *7* (1), 1–10.

(7) Chong, S. H.; Im, H.; Ham, S. Explicit Characterization of the Free Energy Landscape of PKID–KIX Coupled Folding and Binding. *ACS Cent. Sci.* **2019**, *5* (8), 1342–1351.

(8) Lazaridis, T.; Karplus, M. Effective Energy Functions for Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2000**, *10* (2), 139–145.

(9) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis. *Proteins Struct. Funct. Bioinforma.* **1995**, *21* (3), 167–195.

(10) Chong, S. H.; Ham, S. Impact of Chemical Heterogeneity on Protein Self-Assembly in Water. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (20), 7636–7641.

(11) Deechongkit, S.; Dawson, P. E.; Kelly, J. W. Toward Assessing the Position-Dependent Contributions of Backbone Hydrogen Bonding to β-Sheet Folding Thermodynamics Employing Amide-to-Ester Perturbations. *J. Am. Chem. Soc.* **2004**, *126* (51), 16762–16771.

(12) Dave, K.; Jäger, M.; Nguyen, H.; Kelly, J. W.; Gruebele, M. High-Resolution Mapping of the Folding Transition State of a WW Domain. *J. Mol. Biol.* **2016**, *428* (8), 1617–1636.

(13) Ardejani, M. S.; Powers, E. T.; Kelly, J. W. Using Cooperatively Folded Peptides to Measure Interaction Energies and Conformational Propensities. *Acc. Chem. Res.* **2017**, *50* (8), 1875–1882.

(14) Nesloney, C. L.; Kelly, J. W. Progress towards Understanding β-Sheet Structure. *Bioorganic Med. Chem.* **1996**, *4* (6), 739–766.

(15) Cheng, P. N.; Pham, J. D.; Nowick, J. S. The Supramolecular Chemistry of β-Sheets. *J. Am. Chem. Soc.* **2013**, *135* (15), 5477–5492.

(16) Narayan, A.; Bhattacharjee, K.; Naganathan, A. N. Thermally versus Chemically Denatured Protein States. *Biochemistry* **2019**, *58* (21), 2519–2523.

(17) Chong, S. H.; Ham, S. Atomic Decomposition of the Protein Solvation Free Energy and Its Application to Amyloid-Beta Protein in Water. *J. Chem. Phys.* **2011**, *135* (3), 034506.

(18) Chong, S. H.; Ham, S. Distinct Role of Hydration Ater in Protein Misfolding and Aggregation Revealed by Fluctuating Thermodynamics Analysis. *Acc. Chem. Res.* **2015**, *48* (4), 956–965.
(19) Jäger, M.; Zhang, Y.; Bieschke, J.; Nguyen, H.; Dendle, M.; Bowman, M. E.; Noel, J. P.; Gruebele, M.; Kelly, J. W. Structure-Function-Folding Relationship in a WW Domain. *Proc. Natl. Acad. Sci.* **2006**, *103* (28), 10648–10653.
(20) Wang, L.; O'Connell, T.; Tropsha, A.; Hermans, J. Molecular Simulations of β-Sheet Twisting. *J. Mol. Biol.* **1996**, *262* (2), 283–293.
(21) Gao, J.; Bosco, D. A.; Powers, E. T.; Kelly, J. W. Localized Thermodynamic Coupling between Hydrogen Bonding and Microenvironment Polarity Substantially Stabilizes Proteins. *Nat. Struct. Mol. Biol.* **2009**, *16* (7), 684–690.
(22) Jäger, M.; Nguyen, H.; Dendle, M.; Gruebele, M.; Kelly, J. W. Influence of HPin1 WW N-Terminal Domain Boundaries on Function, Protein Stability, and Folding. *Protein Sci.* **2007**, *16* (7), 1495–1501.
(23) Szczepaniak, M.; Iglesias-Bexiga, M.; Cerminara, M.; Sadqi, M.; Sanchez de Medina, C.; Martinez, J. C.; Luque, I.; Muñoz, V. Ultrafast Folding Kinetics of WW Domains Reveal How the Amino Acid Sequence Determines the Speed Limit to Protein Folding. *Proc. Natl. Acad. Sci.* **2019**, *116* (17), 8137–8142.
(24) Liu, F.; Du, D.; Fuller, A. A.; Davoren, J. E.; Wipf, P.; Kelly, J. W.; Gruebele, M. An Experimental Survey of the Transition between Two-State and Downhill Protein Folding Scenarios. *Proc. Natl. Acad. Sci.* **2008**, *105* (7), 2369–2374.
(25) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334* (6055), 517–520.
(26) Piana, S.; Klepeis, J. L.; Shaw, D. E. Assessing the Accuracy of Physical Models Used in Protein-Folding Simulations: Quantitative Evidence from Long Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98–105.
(27) Kachlishvili, K.; Korneev, A.; Maisuradze, L.; Liu, J.; Scheraga, H. A.; Molochkov, A.; Senet, P.; Niemi, A. J.; Maisuradze, G. G. New Insights into Folding, Misfolding, and Nonfolding Dynamics of a WW Domain. *J. Phys. Chem. B* **2020**, *124* (19), 3855–3872.
(28) a Beccara, S.; Skrbic, T.; Covino, R.; Faccioli, P. Dominant Folding Pathways of a WW Domain. *Proc. Natl. Acad. Sci.* **2012**, *109* (7), 2330–2335.
(29) Zhang, M.; Case, D. A.; Peng, J. W. Propagated Perturbations from a Peripheral Mutation Show Interactions Supporting WW Domain Thermostability. *Structure* **2018**, *26* (11), 1474-1485.e5.
(30) Culka, M.; Rulíšek, L. Factors Stabilizing β-Sheets in Protein Structures from a Quantum-Chemical Perspective. *J. Phys. Chem. B* **2019**, *123* (30), 6453–6461.
(31) Rickard, M. M.; Zhang, Y.; Pogorelov, T. V.; Pogorelov, T. V.; Pogorelov, T. V.; Pogorelov, T. V.; Gruebele, M.; Gruebele, M.; Gruebele, M.; Gruebele, M. Crowding, Sticking, and Partial Folding of GTT WW Domain in a Small Cytoplasm Model. *J. Phys. Chem. B* **2020**, *124* (23), 4732–4740.
(32) Heilmann, N.; Wolf, M.; Kozlowska, M.; Sedghamiz, E.; Setzler, J.; Brieg,

M.; Wenzel, W. Sampling of the Conformational Landscape of Small Proteins with Monte Carlo Methods. *Sci. Rep.* **2020**, *10* (1), 1–13.
(33) Markthaler, D.; Kraus, H.; Hansen, N. Overcoming Convergence Issues in Free-Energy Calculations of Amide-to-Ester Mutations in the Pin1-WW Domain. *J. Chem. Inf. Model.* **2018**, *58* (11), 2305–2318.
(34) Imai, T.; Harano, Y.; Kinoshita, M.; Kovalenko, A.; Hirata, F. Theoretical Analysis on Changes in Thermodynamic Quantities upon Protein Folding: Essential Role of Hydration. *J. Chem. Phys.* **2007**, *126* (225102), 1–9.
(35) Anandakrishnan, R.; Drozdetski, A.; Walker, R. C.; Onufriev, A. V. Speed of Conformational Change: Comparing Explicit and Implicit Solvent Molecular Dynamics Simulations. *Biophys. J.* **2015**, *108* (5), 1153–1164.
(36) Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C. Folding Simulations for Proteins with Diverse Topologies Are Accessible in Days with a Physics-Based Force Field and Implicit Solvent. *J. Am. Chem. Soc.* **2014**, *136* (40), 13959–13962.
(37) Feig, M. Kinetics from Implicit Solvent Simulations of Biomolecules as a Function of Viscosity. *J. Chem. Theory Comput.* **2007**, *3* (5), 1734–1748.
(38) Im, W.; Lee, M. S.; Brooks, C. L. Generalized Born Model with a Simple Smoothing Function. *J. Comput. Chem.* **2003**, *24* (14), 1691–1702.
(39) Eichenberger, A. P.; Van Gunsteren, W. F.; Riniker, S.; Von Ziegler, L.; Hansen, N. The Key to Predicting the Stability of Protein Mutants Lies in an Accurate Description and Proper Configurational Sampling of the Folded and Denatured States. *Biochim. Biophys. Acta* **2015**, *1850* (5), 983–995.
(40) Shao, Q.; Zhu, W. Assessing AMBER Force Fields for Protein Folding in an Implicit Solvent. *Phys. Chem. Chem. Phys.* **2018**, *20* (10), 7206–7216.
(41) Robinson, M. K.; Monroe, J. I.; Shell, M. S. Are AMBER Force Fields and Implicit Solvation Models Additive? A Folding Study with a Balanced Peptide Test Set. *J. Chem. Theory Comput.* **2016**, *12* (11), 5631–5642.
(42) Roe, D. R.; Hornak, V.; Simmerling, C. Folding Cooperativity in a Three-Stranded β-Sheet Model. *J. Mol. Biol.* **2005**, *352* (2), 370–381.
(43) Ensign, D. L.; Pande, V. S. The Fip35 WW Domain Folds with Structural and Mechanistic Heterogeneity in Molecular Dynamics Simulations. *Biophys. J.* **2009**, *96* (8), 53–55.
(44) Hua, D. P.; Huang, H.; Roy, A.; Post, C. B. Evaluating the Dynamics and Electrostatic Interactions of Folded Proteins in Implicit Solvents. *Protein Sci.* **2016**, *25* (1), 204–218.
(45) Lang, E. J. M.; Baker, E. G.; Woolfson, D. N.; Mulholland, A. J. Generalized Born Implicit Solvent Models Do Not Reproduce Secondary Structures of De Novo Designed Glu/Lys Peptides . *J. Chem. Theory Comput.* **2022**.
(46) Shao, Q.; Zhu, W. The Effects of Implicit Modeling of Nonpolar Solvation on Protein Folding Simulations. *Phys. Chem. Chem. Phys.* **2018**, *20* (27), 18410–18419.
(47) Chen, J.; Brooks, C. L. Implicit Modeling of Nonpolar Solvation for Simulating Protein Folding and Conformational Transitions. *Phys. Chem. Chem. Phys.* **2008**, *10* (4), 471–481.
(48) Yang, C.; Pak, Y. Comparative Study of Implicit and Explicit Solvation Models for Probing Tryptophan Side Chain Packing in Proteins. *Bull. Korean Chem. Soc.* **2012**, *33* (3), 828–832.

(49) Lazaridis, T.; Karplus, M. Thermodynamics of Protein Folding: A Microscopic View. *Biophys. Chem.* **2003**, *100* (1–3), 367–395.
(50) Chong, S. H.; Ham, S. Site-Directed Analysis on Protein Hydrophobicity. *J. Comput. Chem.* **2014**, *35* (18), 1364–1370.
(51) Chong, S. H.; Ham, S. Interaction with the Surrounding Water Plays a Key Role in Determining the Aggregation Propensity of Proteins. *Angew. Chemie* **2014**, *53* (15), 3961–3964.
(52) Chong, S. H.; Ham, S. Examining a Thermodynamic Order Parameter of Protein Folding. *Sci. Rep.* **2018**, *8* (1), 1–9.
(53) Chong, S. H.; Ham, S. Folding Free Energy Landscape of Ordered and Intrinsically Disordered Proteins. *Sci. Rep.* **2019**, *9* (1), 14927.
(54) Cho, M. K.; Chong, S. H.; Shin, S.; Ham, S. Site-Specific Backbone and Side-Chain Contributions to Thermodynamic Stabilizing Forces of the WW Domain. *J. Phys. Chem. B* **2021**, *125* (26), 7108–7116.
(55) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.
(56) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins Struct. Funct. Genet.* **2004**, *55* (2), 383–394.
(57) Mittal, J.; Best, R. B. Tackling Force-Field Bias in Protein Folding Simulations: Folding of Villin HP35 and Pin WW Domains in Explicit Water. *Biophys. J.* **2010**, *99* (3), L26–L28.
(58) Zhou, R.; Berne, B. J.; Germain, R. The Free Energy Landscape for β Hairpin Folding in Explicit Water. *Proc. Natl. Acad. Sci. U. S. A.* **2001**.
(59) Lucent, D.; Vishal, V.; Pande, V. S. Protein Folding under Confinement: A Role for Solvent. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (25), 10430–10434.
(60) Yoda, T.; Sugita, Y.; Okamoto, Y. Hydrophobic Core Formation and Dehydration in Protein Folding Studied by Generalized-Ensemble Simulations. *Biophys. J.* **2010**, *99* (5), 1637–1644.
(61) Markthaler, D.; Fleck, M.; Stankiewicz, B.; Hansen, N. Exploring the Effect of Enhanced Sampling on Protein Stability Prediction. *J. Chem. Theory Comput.* **2022**, *18* (4), 2569–2583.
(62) Jiang, F.; Wu, Y. D. Folding of Fourteen Small Proteins with a Residue-Specific Force Field and Replica-Exchange Molecular Dynamics. *J. Am. Chem. Soc.* **2014**, *136* (27), 9536–9539.
(63) Andrews, B.; Long, K.; Urbanc, B. Soluble State of Villin Headpiece Protein as a Tool in the Assessment of MD Force Fields. *J. Phys. Chem. B* **2021**, *125* (25), 6897–6911.
(64) Singh, P.; Sarkar, S. K.; Bandyopadhyay, P. Folding-Unfolding Transition in the Mini-Protein Villin Headpiece (HP35): An Equilibrium Study Using the Wang-Landau Algorithm. *Chem. Phys.* **2016**, *468*, 1–8.
(65) Cumberworth, A.; Bui, J. M.; Gsponer, J. Free Energies of Solvation in the Context of Protein Folding: Implications for Implicit and Explicit Solvent Models. *J. Comput. Chem.* **2016**, *37* (7), 629–640.
(66) Lei, H.; Wu, C.; Liu, H.; Duan, Y. Folding Free-Energy Landscape of Villin Headpiece Subdomain from Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (12), 4925–4930.

(67) Jang, S.; Kim, E.; Shin, S.; Pak, Y. Ab Initio Folding of Helix Bundle Proteins Using Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2003**, *125* (48), 14841–14846.
(68) Case, D. A.; Darden, T.; Iii, T. E. C.; Simmerling, C.; Brook, S.; Roitberg, A.; Wang, J.; Southwestern, U. T.; Duke, R. E.; Hill, U.; et al. Amber 14 Manual. University of California: San Francisco, CA, 2014.
(69) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins Struct. Funct. Bioinforma.* **2006**, *65* (3), 712–725.
(70) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.
(71) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341.
(72) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220* (4598), 671–680.
(73) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.
(74) Frishman, D.; Argos, P. Knowledge-Based Protein Secondary Structure Assignment. *Proteins Struct. Funct. Bioinforma.* **1995**, *23* (4), 566–579.
(75) Yuan, C.; Chen, H.; Kihara, D. Effective Inter-Residue Contact Definitions for Accurate Protein Fold Recognition. *BMC Bioinformatics* **2012**, *13* (292).
(76) Chong, S. H.; Ham, S. Protein Folding Thermodynamics: A New Computational Approach. *J. Phys. Chem. B* **2014**, *118* (19), 5017–5025.
(77) Kovalenko, A. Three-Dimensional RISM Theory for Molecular Liquids and Solid-Solid Interfaces. In *Molecular Theory of Solvation*; Hirata, F., Ed.; Kluwer Academic: Dordrecht, The Netherlands, 2003; pp 169–275.
(78) Imai, T.; Harano, Y.; Kinoshita, M.; Kovalenko, A.; Hirata, F. A Theoretical Analysis on Hydration Thermodynamics of Proteins. *J. Chem. Phys.* **2006**, *125* (2), 024911.
(79) Chong, S. H.; Ham, S. Configurational Entropy of Protein: A Combined Approach Based on Molecular Simulation and Integral-Equation Theory of Liquids. *Chem. Phys. Lett.* **2011**, *504* (4–6), 225–229.
(80) Chong, S. H.; Ham, S. Dissecting Protein Configurational Entropy into Conformational and Vibrational Contributions. *J. Phys. Chem. B* **2015**, *119* (39), 12623–12631.
(81) Grossfield, A.; Patrone, P. N.; Roe, D. R.; Schultz, A. J.; Siderius, D.; Zuckerman, D. M. Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations. *Living J. Comput. Mol. Sci.* **2019**, *1* (1), 1–24.
(82) Best, R. B.; Hummer, G.; Eaton, W. A. Native Contacts Determine Protein Folding Mechanisms in Atomistic Simulations. *Proc. Natl. Acad. Sci.* **2013**, *110* (44), 17874–17879.
(83) Zhou, R.; Maisuradze, G. G.; Suñol, D.; Todorovski, T.; Macias, M. J.; Xiao, Y.; Scheraga, H. A.; Czaplewski, C.; Liwo, A. Folding Kinetics of

WW Domains with the United Residue Force Field for Bridging Microscopic Motions and Experimental Measurements. *Proc. Natl. Acad. Sci.* **2014**, *111* (51), 18243–18248.

(84) Koepf, E. K.; Petrassi, H. M.; Sudol, M.; Kelly, J. W. WW: An Isolated Three-Stranded Antiparallel β-Sheet Domain That Unfolds and Refolds Reversibly; Evidence for a Structured Hydrophobic Cluster in Urea and GdnHCl and a Disordered Thermal Unfolded State. *Protein Sci.* **1999**, *8* (4), 841–853.

(85) Sheu, S. Y.; Yang, D. Y.; Selzle, H. L.; Schlag, E. W. Energetics of Hydrogen Bonds in Peptides. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (22), 12683–12687.

(86) Pace, C. N.; Fu, H.; Fryar, K. L.; Landua, J.; Trevino, S. R.; Schell, D.; Thurlkill, R. L.; Imura, S.; Scholtz, J. M.; Gajiwala, K.; et al. Contribution of Hydrogen Bonds to Protein Stability. *Protein Sci.* **2014**, *23* (5), 652–661.

(87) Fernández, A.; Scott, R. Dehydron: A Structurally Encoded Signal for Protein Interaction. *Biophys. J.* **2003**, *85* (3), 1914–1928.

(88) Bunagan, M. R.; Gao, J.; Kelly, J. W.; Gai, F. Probing the Folding Transition State Structure of the Villin Headpiece Subdomain via Side Chain and Backbone Mutagenesis. *J. Am. Chem. Soc.* **2009**, *131* (21), 7470–7476.

(89) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; et al. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330* (6002), 341–346.

(90) Best, R. B.; Hummer, G. Microscopic Interpretation of Folding φ-Values Using the Transition Path Ensemble. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113* (12), 3263–3268.

(91) Clementi, C.; García, A. E.; Onuchic, J. N. Interplay among Tertiary Contacts, Secondary Structure Formation and Side-Chain Packing in the Protein Folding Mechanism: All-Atom Representation Study of Protein L. *J. Mol. Biol.* **2003**, *326* (3), 933–954.

(92) Naganathan, A. N.; De Sancho, D. Bridging Experiments and Native-Centric Simulations of a Downhill Folding Protein. *J. Phys. Chem. B* **2015**, *119* (47), 14925–14933.

(93) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9* (9), 3878–3888.

(94) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. NMR Structure of the 35-Residue Villin Headpiece Subdomain. *Nat. Struct. Biol.* **1997**, *4* (3), 180–184.

(95) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713.

(96) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; et al. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13* (7), 1–

17.
(97) Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D.; Cheatham, T. E. I.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Gilson, M. K. AMBER18. University of California: San Francisco, CA, 2018.
(98) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (8), 1950–1958.
(99) Best, R. B.; Hummer, G.; Eaton, W. A. Native Contacts Determine Protein Folding Mechanisms in Atomistic Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (44), 17874–17879.
(100) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat. Methods* **2022**, *19* (6), 679–682.
(101) Kovalenko, A. *Molecular Theory of Solvation; Kluwer Academic: Dordrecht, The Netherlands*; 2003.
(102) Vijay-kumar, S.; Bugg, C. E.; Cook, W. J. Structure of Ubiquitin Refined at 1.8 Å Resolution. *J. Mol. Biol.* **1987**, *194* (3), 531–544.
(103) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. Investigation of Salt Bridge Stability in a Generalized Born Solvent Model. *J. Chem. Theory Comput.* **2006**, *2* (1), 115–127.
(104) Lwin, T. Z.; Luo, R. Force Field Influences in β-Hairpin Folding Simulations. *Protein Sci.* **2006**, *15* (11), 2642–2655.
(105) Cho, M.; Chong, S.-H.; Ham, S.; Shin, S. Comparing the Influence of Explicit and Implicit Solvation Models on Site-specific Thermodynamic Stability of Proteins. *J. Comput. Chem.* **2023**, *44*(25), 1976–1986.

국문초록

조명근

화학부 물리화학 전공

자연과학대학

서울대학교 대학원

단백질 접힘은 수성 환경에 크게 의존합니다. 용매화가 단백질 접힘에 미치는 영향은 널리 연구되어 왔지만 접힘 안정성이 용매화에 의해 제어되는 정도는 개별 아미노산 수준에서 명확하지 않습니다. 여기에서 우리는 단백질의 각 백본과 측쇄에 대한 접힘 자유 에너지 요소를 평가하기 위해 사이트 지정 열역학 분석 방법을 사용합니다. 따라서 대표적인 β-시트 및 α-나선 단백질의 각 중요 부위에서 시스템에 물리적 변형을 도입하지 않고 접힘 안정성 기여도를 정량적으로 측정합니다. 인간 Pin WW 도메인 단백질 및 빌린 헤드피스 서브도메인 단백질 각각의 접힘 현상에 대한 수십 μs 길이의 분자 역학 시뮬레이션으로부터의 열역학적 결과가 보고됩니다. 결과로는 Pin WW의 접힘 자유 에너지는 -4.9 kcal/mol이었으며, 이는 기존 실험 결과 보고와 흡사했습니다. 용매화 자유 에너지 및 진공 상태의 단백질 에너지의 분해 방법을 단일 아미노산 분해능에 통합함으로써 단백질 안정성을 지배하는 수소 결합 및 소수성 상호 작용과 같은 기본 분자 상호 작용의 에너지 결과를 결정합니다.

사이트 지정 열역학 방법의 적용은 두 모델 단백질의 열역학적 안정성에 대한 명시적 및 암시적 용매화 모델의 영향을 비교하기 위해 확장됩니다. 열역학 분석은 종종 명시적 또는 암시적 물 모델을 사용하는 분자 역학 시뮬레이션을 사용하여 많은 수의 원자적 형태를 샘플링하여 수행됩니다. 서로 다른 용매화 모델의 열역학적 결과가 분자 수준에서 어느 정도 신뢰할 수 있는지는 불확실합니다. 여기서 우리는 단일 백본 및 측쇄 분해능에서 폴딩 안정성에 대한 두 용매화 모델의 영향을 정량화합니다. TIP3P 용매 및 일반화된 Born/표면적 용매 모델에서 생성된 시뮬레이션 궤적을 사용하여 위에서 설명한 두 단백질의 잔류물 특정 폴딩 자유 에너지 구성 요소를 평가합니다. 일반화 된 Born 용매의 열역학적 불일치는 대부분 양성 측쇄에

서 비롯된 다음 불안정한 소수성 측쇄에서 비롯된 것으로 나타났습니다. 대조적으로, 두 단백질의 백본 잔기 기여도는 비슷했다. 우리의 연구는 단백질 시뮬레이션의 맥락에서 용매 모델의 상세한 열역학적 평가의 토대를 마련합니다.

주요어: 분자동역학 시뮬레이션, 자유에너지 분해, 베타-시트, generalized born solvent model, 용매화 자유 에너지, 3D-RISM, 단백질 내부 포텐셜 에너지
학생번호: 2016-20364