Alchemical Free Energy Calculations

in the Study of Protein Biophysics

A Dissertation Presented

by

Junjie Zou

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Chemistry

Stony Brook University

September 2019

Stony Brook University

The Graduate School

Junjie Zou

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Daniel P. Raleigh, Ph.D. – Dissertation Advisor
Professor, Department of Chemistry, Stony Brook University

Carlos Simmerling, Ph.D. - Dissertation Advisor
Professor, Department of Chemistry, Laufer Center for Physical and Quantitative Biology, Stony Brook University

Jin Wang, Ph.D. - Chairperson of Defense
Professor, Department of Chemistry, Stony Brook University

Peter Tonge, Ph.D. – Third Member
Professor, Department of Chemistry, Stony Brook University

Dima Kozakov, Ph.D. – Outside Member
Professor, Department of Applied Mathematics and Statistics, Stony Brook University

This dissertation is accepted by the Graduate School

Eric Wertheimer

Dean of the Graduate School

Abstract of the Dissertation

Alchemical Free Energy Calculations

in the Study of Protein Biophysics

by

Junjie Zou

Doctor in Philosophy

in

Chemistry

Stony Brook University

2019

Alchemical free energy calculations have been widely used to understand the free energy changes associated with the binding between ligand and protein, the effect of mutations on protein folding, protein-protein interactions and other interactions. Methods for free energy calculations such as thermodynamic integration (TI) are thermodynamically rigorous, which allows the computed free energy to be directly compared with experimentally measured thermodynamic quantities. This makes free energy calculations a popular tool when studying the thermodynamic properties of protein-ligand interaction, protein-protein interactions and protein folding. The accuracy and feasibility of alchemical free energy calculation have been significantly improved recently thanks to the development of force fields and implementation of molecular dynamics (MD) simulation on graphical processing units.

Here, TI calculations were applied to three different scenarios to demonstrate how to combine free energy calculations with experimental results to reveal insights that are difficult to be studied using experimental methods or TI alone.

Two major concerns for computational modeling of protein are whether the force field is accurate enough to describe the energetics of the protein and whether the sampling is sufficient to cover biologically relevant conformations of the protein. The force field and the sampling do not need to be perfect in order to study some properties of protein. Nonetheless, it is important to check whether the accuracy of force field and the amount of sampling are sufficient for probing the questions that are interested at. Free energy calculations are a very good tool to validate the accuracy of the force field and the sampling efficiency because free energies, calculated by using method such as TI, are thermodynamically rigorous and can be directly compared with experimentally measured free energy changes.

Mutagenesis is one of the most popular methods for understanding the energetic contribution of a residue to the folding free energy of protein. Experimentally measured effects of mutations on the folding free energy of proteins is a convoluted effect consisting of the changes caused by the mutation in both the folded and the unfolded state. The mutation effect on the folded state of protein is usually estimated by assuming that the mutation effect on the unfolded state is negligible, which may or may not be true. If the effect on the unfolded state can be ignored, the experimentally measured free energy change $\Delta G°_U$, where $\Delta G°_U$ is the measured free energy of unfolding, can be ascribed to folded state effects. However, such a strategy cannot be applied to study the unfolded state of a protein because the mutation effect on the folded state is unlikely to be negligible. Since TI is thermodynamically rigorous, calculated free energy changes in the folded state of protein can be directly utilized to deconvolute the experimentally measured free energy changes caused by the mutations. This allows the mutation effect on the unfolded state to be characterized quantitatively without direct modelling the unfolded state using MD simulations, which is computationally expensive.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CARL	Subtilisin Carlsberg
CALC	Calculated
C-cap	C-terminal residue of α-helix
CD	Circular dichroism
DSC	Differential scanning calorimetry
DSE	Denatured state ensemble
EXP	Experimental
F	The folded state
GPU	Graphic processing unit
U	The unfolded state
HPLC	High Performance liquid chromatography
HP21	The first 21-residue peptide fragment from HP36
HP36	The C-terminal subdomain of villin headpiece
IDP	Intrinsically disordered protein
ITC	isothermal titration calorimetry
MALDI	Matrix assisted laser desorption and ionization
MM	Molecular Mechanics
MD	Molecular dynamics
NAC	The non-amyloid-component domain in α-synuclein
NTAIL	The C-terminal domain of the nucleoprotein of measles virus
NTL9	The N-terminal domain of protein L9
NMR	Nuclear magnetic resonance

OMTKY3	The turkey ovomucoid inhibitor third binding domain
PDB	Protein data bank
ppm	Parts per million
SASA	Solvent accessible surface area
SAXS	Small angle X-ray scattering
SGPB	*Streptomyces griseus* proteinase B
TFA	Trifluoroacetic acid
TI	Thermodynamic integration
TOCSY	Total correlation spectroscopy
UV	Ultraviolet
VDW	Van der Waals
WT	Wildtype
XD	X domain of measles virus phosphoprotein
αMoRE	α-helical Molecular Recognition Element
ΔG	Free energy change
ΔE	Enthalpy change
Φ, Ψ	Backbone torsion angle of amino acids
3D	Three dimensional

Acknowledgements

I would like to thank my advisors Prof. Daniel Raleigh and Carlos Simmerling for guiding and supporting my research and studies during my Ph.D. candidacy. I am grateful for their encouragement and patience. I am also grateful for their advices on improving my skills of scientific writing, presentation and communication. These advices will certainly benefit my future career.

I would like to thank Prof. Jin Wang, Prof. Peter Tonge and Prof. Dima Kozakov for serving as my dissertation committee members. I thank them for their valuable suggestions on my research projects.

I appreciate the friendship with the past and present members of both the Raleigh's group and Simmerling's group: Dr. Hui Wang, Dr. Cynthia Tsu, Dr. Matthew Watson, Dr. Xiaoxue Zhang, Dr. Amy Wang, Dr. Natalie Stenzoski, Dr. Kyung-Hoon Lee, Dr. Cheng-tsung Lai, Dr. Haoquan Li, Dr. James Maier, Dr. Kevin Hauser, Dr. Hai Nguyen, Dr. He Huang, Dr. Angela Migues, Zack Ridgway, Daeun Noh, Rehana Akter, Ming-Hao Li, Lakshan Manathunga, Matthew Miller, Koushik Kasavajhala, Kenneth Lam, Kellon Belfon, Zachary Fallon, Lauren Raguette, Seungyoun Shin, Yuzhang Wang. I specially want to thank Dr. Bowu Luan and Dr. Ivan Peran for the mentorship I received during my first and second year and thank Chuan Tian for collaborative projects.

Last but not the least I would like to thank my family for their encouragement, time, support and understanding.

List of Publications

1. **Zou, J.**; Song, B.; Simmerling, C.; Raleigh, D., Experimental and computational analysis of protein stabilization by Gly-to-d-Ala substitution: A convolution of native state and unfolded state effects. *J Am Chem Soc* 2016, 138 (48), 15682-15689.
2. **Zou, J.**; Tian, C.; Simmerling, C., Blinded prediction of protein-ligand binding affinity using Amber thermodynamic integration for the 2018 D3R grand challenge 4. *J Comput Aided Mol Des* 2019.
3. **Zou, J.**; Simmerling, C.; Raleigh, D., Dissecting the energetics of intrinsically disordered proteins via a hybrid experimental and computational approach. (*submitted*)
4. **Zou, J.**; Shifeng, X.; Simmerling, C.; Raleigh, D., Molecular Basis of Roughness on The Free Energy Landscape for Protein Folding; Experimental and Computational Studies of a Non-Native Interactions in The Denatured State of a Fast Folding Protein. (*to be submitted*)
5. Watson, M. D.; Peran, I.; **Zou, J**.; Bilsel, O.; Raleigh, D. P., Selenomethionine Quenching of Tryptophan Fluorescence Provides a Simple Probe of Protein Structure. *Biochemistry* 2017, *56* (8), 1085-1094.
6. Neckles, C.; Pschibul, A.; Lai, C. T.; Hirschbeck, M.; Kuper, J.; Davoodi, S.; **Zou, J.**; Liu, N.; Pan, P.; Shah, S.; Daryaee, F.; Bommineni, G. R.; Lai, C.; Simmerling, C.; Kisker, C.; Tonge, P. J., Selectivity of Pyridone- and Diphenyl Ether-Based Inhibitors for the *Yersinia pestis* FabV Enoyl-ACP Reductase. *Biochemistry* 2016, *55* (21), 2992-3006.
7. Zhang, S.; Zhang, Y.; Stenzoski, N. E.; **Zou, J**.; Peran, I.; McCallum, S. A.; Raleigh, D. P.; Royer, C. A., Pressure-Temperature Analysis of the Stability of the CTL9 Domain Reveals Hidden Intermediates. *Biophys J* 2019, 116 (3), 445-453.

8. Holehouse, A. S.; Peran, I.; Stenzoski, N. E.; **Zou, J.**; Piserchio, A.; Ghose, R.; Carrico, I. S.; Bilsel, O.; Raleigh, D. P.; Pappu, R. V., Protein Unfolded States are Characterized by the Duality of Sequence-Specific Conformational Preferences and Ensemble-Averaged Features of Canonical Random Coils. *Biophys J* 2019, 116 (3), 199a-200a.

1. Introduction

1.1 Protein stability

1.1.1 Protein thermodynamics

A protein can adopt many different conformations whose propensities are determined by the potential energies of the conformations. These potential energies are determined by the relative orientation and interactions of two adjacent residues in sequence as well as the interactions between residues and solvent and longer range interactions. As a polymer, proteins have a huge available configuration space which often causes them to remain disordered and unfolded in physiological conditions. Proteins can also have favorable cooperative interactions between amino acids in a single configuration which offsets the entropy lost upon folding and promote them to become structured and folded. The reaction of unfolding and folding can be described by the unfolding free energy, $\Delta G_U = \Delta H_U - T\Delta S_U$. ΔH_U is the enthalpy difference between the unfolded state and the folded state. ΔS_U is the entropy difference between the unfolded state and the folded state. A protein with a random sequence is very likely to be disordered under physiological conditions as a random sequence typically lacks sufficient cooperative favorable interactions between amino acids in a single configuration. Statistics of $\Delta G°_U$, $\Delta H°_U$ and $T\Delta S°_U$ of protein unfolding show that $\Delta H°_U$ and $T\Delta S°_U$ are highly correlated. $\Delta H°_U$ and $T\Delta S°_U$ can have a range of values from -100 kcal/mol to 200 kcal/mol, but the $\Delta G°_U$ are in a range of -10 kcal/mol to 10 kcal/mol (1, 2). Since the unfolding free energy $\Delta G°_U$ is close to 0 kcal/mol, even a small modification of protein can drastically change the relative populations of the folded and the unfolded state.

1.1.2 Importance of protein stability

Enzymes are being widely used in all kinds of industries such as textiles, foods, beverage industry and so on (3, 4). The use of proteins as drugs is also common in the pharmaceutical industry (5). Since the protein functions are highly dependent on the conformations of globular proteins, enhancing the conformational stability of proteins with industrial and pharmaceutical value is very important (6). Enzyme and protein drugs with low stability of the folded state have more populations of conformations that have little or no bioactivity, which impair their values as catalysts and medicines. A protein with low stability of the folded state is also vulnerable to degradation, which can lower the yield of the protein and complicate the production process (7-9). Moreover, a low stability of the folded state increases the chance of aggregation and misfolding of the protein, which can cause a safety hazard in pharmaceutical applications and lose of activity in industrial applications (10, 11).

1.1.3 Contribution of amino acids to protein stability

The enthalpy of proteins is determined by the interaction strength between residues and the interactions between residues and water. The strength of these interactions depends on the properties of the amino acids. Proteins are a heterogenous polymer as the composing monomer, amino acids, are highly diverse in their physical chemical properties. The 20 natural amino acids can be divided into groups based on their physical chemical properties. Ala, Cys, Phe, Ile, Leu, Met, Pro, Val, Trp & Tyr are usually considered hydrophobic residues which are energetically unfavorable when exposed to water, thus, they usually form the core of the folded state of protein. Asp, Glu, Gly, His, Lys, Asn, Gln, Arg, Ser & Thr are considered hydrophilic residues and are energetically favorable when exposed to water. Among these hydrophilic residues, Asp & Glu can be negatively charged and His, Lys & Arg can be positively charged at certain pHs.

In a series of studies on various proteins, it was found that the contribution of hydrophobic residues to the stability of proteins depends on how much the contacting areas between water and hydrophobic residues was removed upon folding. It also depends on the van der Waals interactions that were formed between hydrophobic residues in the folded state of proteins. A statistical study of 22 proteins showed that the average contribution of hydrophobic interactions to the stability protein is 60%, which makes the hydrophobic interactions the predominant stabilizing factor (12). Studies have shown that for a small protein like the 36-residue villin head piece subdomain, HP36, the contribution of a -CH2- group to the stability is around 0.6 kcal/mol in free energy. The contribution can be as high as 1.6 kcal/mol as found for a surface protein from *Borrelia burgdorferi* (VlsE) which has 341 residues. This is because a larger protein is likely to have a more buried interior compared to a small protein, thus, a -CH2- group in the interior of a larger folded protein can have more water contacting areas removed and more van der Waals interactions formed upon folding (12).

Incorporation of salt bridges, which are formed by two residues with opposite charges in close distance, in the folded state of a protein also have shown positive effects on the stability of proteins. The contribution of a surface salt bridge to the stability of protein is usually less than 1 kcal/mol (13). The enhancement of stability that results from the incorporation of a salt bridge is usually less than the value estimated by Coulomb's law. Since the electrostatic interactions are long range interactions, it is possible that the charge-charge interactions also lead to favorable interactions in the unfolded state, which offset the beneficial gain in the folded state (13). The role of buried salt bridges is difficult to accurately estimate because the favorable coulumbic interactions is offset by the unfavorable desolvation effects. The mutation of surface hydrophobic residues to a charged residue was believed to be beneficial to the stability of proteins as the charged residue can form more favorable interactions with water (14, 15). In some cases, however, changing a hydrophobic residue on the surface to a charged residue may disrupt the favorable hydrophobic interactions contributed by the residue. This can lead to a destabilized protein, if the hydrophobic interactions are strong enough (16).

A charged residue in a low dielectric environment like the hydrophobic interior of protein is much more unfavorable than a charged residue in a high dielectric environment like bulk water. Thus, a charged residue strongly prefers being exposed to water than being buried inside protein. Incorporation of charged residues in the interior of the folded state of protein is highly destabilizing.

1.1.4 Enhancing the stability of protein by targeting the unfolded state

An alternative strategy for protein stabilization besides strategies targeting the folded state is to decrease the entropy of the unfolded state. The entropy of a protein in the unfolded state is largely contributed by the available conformation of the backbone of amino acids. The backbone conformations of amino acids can be defined by the two dihedral angles Φ and Ψ (**Fig. 1-1**). The allowed (Φ, Ψ) areas are determined by the steric hindrance caused by the side chains of amino acids. A $\Phi>0$ value can result in steric clash between the CB atom of non-glycine L amino acid and the carboxyl oxygen of the previous residue. The side chain of Gly causes no steric hindrance, so Gly has the largest allowed (Φ, Ψ) areas among all 20 amino acids. On the other hand, Pro has a cyclic side chain which causes constraints in backbone and significantly reduces the available backbone conformation. Thus, Pro has the smallest allowed (Φ, Ψ) areas among the amino acids.

The side chains of other amino acids have milder steric hindrance compared to Pro, but still rule out most areas with $\Phi > 0°$ (**Fig. 1-2**).

The decrease of entropy in the unfolded state destabilizes that state, conversely increasing the relative stability of the folded state **(Fig. 1-3)**. Decrease of the entropy in the unfolded state can be achieved by 1) replacing residues with more backbone flexibility with residues with less backbone flexibility, such as Gly to non-Gly mutations and non-Pro to Pro mutations (17-19); 2) peptide cyclization through formation of disulfide bonds and covalently linking the N and C termini of peptides (20-24).

The modifications to the unfolded state may also cause collateral effects on the folded state and vice versa. These collateral effects may offset the expected beneficial gain of the modifications and result in mixed consequences on ΔG. Thus, it is important to consider all the effects that might be caused by modifications.

1.1.5 Use of Non-natural amino acids in protein stability enhancement

Non-natural amino acids have been widely used in rational protein stabilization. Substitutions of hydrogen with fluorine in Leu and Val significantly enhance the hydrophobicity of the side chain of Leu and Val. Substitution of Leu and Val with 5,5,5-trifluoroleucine and 4,4,4-trifluorovaline have shown success in stabilizing protein NTL9 and increasing the helicity of peptide coil-coil (25-27).

The cis conformation is highly energetically unfavorable for peptide bonds formed by natural amino acids. However, incorporation of 5,5-dimethyl-L-proline allows the peptide bond to adopt cis conformation without paying an energy penalty **(Fig. 1-4)**. Substituting natural amino acids involved in cis conformation found in the folded state with 5,5-dimethyl-l-proline can relieve the peptide bond strain and lower the enthalpy of the folded state (28). Tyrosine derivatives with para-substituted aliphatic thiols of various lengths can form elongated disulfide bonds. Elongation of disulfide bonds grant more freedom in positioning the disulfide bonds in the folded state of protein (29).

1.2 The villin headpiece subdomain

One of the model protein studied in this dissertation is the villin headpiece subdomain (HP36), a 36-residue protein, which is one of the smallest naturally occurring proteins that folds cooperatively on its own (30). The folding kinetics of HP36 is approaching the theoretical speed limit of folding, which makes HP36 a very popular protein model for understanding protein thermodynamics and folding kinetics (31, 32). The structure of HP36 is constructed of a bundle of helices and three Phe residues that contribute to the hydrophobic core of the folded state of HP36 (33) **(Fig. 1-5)**. Residual helicity was found in HP21 (residues 41-63 in full length HP36), which is a segment of HP36 composed of the first two helices of HP36. HP21 was considered as a representation for the denatured state of HP36 (34). However, fragment of helix 1 (residues 41-53 in full length HP36) and fragment of helix 2 (residues 52-61 in full length HP36) are fully unstructured, which indicates that the residual helicity in HP21 and the denatured state of HP36 is stabilized by tertiary contacts formed btween helix 1 and helix 2 (35). Previous studies on the electrostatic interactions in HP36 have shown possible electrostatic interactions in the unfolded state of HP36 (16).

1.3 Intrinsically disordered proteins

Intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) of proteins lack stable secondary and tertiary structures due to their low content of bulky hydrophobic residues and their high content of polar and charged residues (36). Although many IDPs and IDRs were discovered one by one during the last 75 years or so, the abundance of IDPs and IDRs was not recognized until the 1990's. V. Uversky suggested that not having a unified terminology for IDPs and IDRs is part of the reason for the negligence of their abundant existence in the nature (37). This later changed due to bioinformatic studies which showed that IDPs and IDRs form a large portion of the genome of various organisms (38). Since then, the interests in IDPs and IDRs have grown exponentially (37). Many findings about IDPs' important roles in biology emerged during the early 1990s, which caused a paradigm-shift because it was believed that only folded proteins can carry functions (39, 40).

Structural characterization has shown that IDPs are not true random-coils; instead they often contain transient secondary structures and long-range interactions (41-47). For example, the activator for thyroid hormone and retinoid receptors (ACTR) is a pre-molten globular protein with four transient α-helices in the native condition (41). It was found that increasing the helicity in one of the four α-helices leads to the increase of helicity in the other three α-helices due to long-range interactions between them (41). Moreover, long-range interactions in IDPs may also play an important role in protein aggregation diseases. Lewy bodies, which are the hallmark of Parkinson's disease, are fibrils primarily formed by α-synuclein, an IDP in its monomeric state. The non-amyloid-component (NAC) domain in α-synuclein is highly hydrophobic, and isolated NAC segments readily forms aggregates (48). However, in the full length α-synuclein, residues 110-130 and residues 85-95 in the NAC domain form transient hydrophobic interactions, which compete with the self-association of NAC and attenuate the fibrilization of α-synuclein (49).

1.4 The C-terminal domain of the measles virus nucleoprotein, a model IDP

The C-terminal domain of the measles virus nucleoprotein (NTAIL) is used as a model IDP to demonstrate a method described in the dissertation. NTAIL is a 125-residue IDP from residue 401 to 525 in the full-length nucleoprotein (50). The nucleoprotein encapsulates the viral RNA and can interact with the phosphoprotein during the replication of the RNA of *Paramyxoviridae* (51). The interaction is participated by NTAIL and the C-terminal X domain (XD) of the phosphoprotein (52). Upon the transcription and replication of the RNA, the phosphoprotein can incorporate the large protein and enable the large protein to fold into a functional structure **(Fig. 1-6)** (53). The large protein has most of the enzymatic activities necessary for the transcription and replication of the RNA (54). The binding strength between the nucleoprotein and the phosphoprotein has to be carefully tuned. If the interaction between the nucleoprotein and the phosphoprotein is too weak, the replication machinery will not form. On the other hand, too strong interaction will anchor the phosphoprotein on the RNA encapsulated by the nucleoprotein which causes a negative effect on the replication of RNA (55).

The protein XD is a globular protein with three bundles of helices (56). Upon binding to the XD, residues from 489 to 506 of NTAIL fold into a helical conformation and form a complex of four helices with the XD (57). Studies have shown that residues 489-506 have residual helical content in the free state of NTAIL, while the flanking regions of 401-488 and 507-525 are devoid of any structure (58). The role of the flanking regions in the binding between NTAIL and the XD has been examined by measuring the binding strength of the NTAIL/XD complex with various truncation lengths in the region 401-488. A non-monotonic relation between the binding strength and the truncation length was observed. Moreover, the NTAIL variant without the segment of 401-488 showed diffusion limited binding kinetics (59). Single mutations in the region 401-488 indicated that NTAIL may have interactions between region 401-488 and region 489-506 in the free state of NTAIL. These interactions are likely to be a self-inhibitory mechanism which regulates the binding strength between NTAIL and XD (60). This is explained in detail in Chapter 3.

1.5 Molecular mechanics and molecular dynamics simulation

Molecular mechanics (MM) is one of the most popular biomolecule modelling techniques since it is much less computationally expensive compared to methods based on quantum mechanics, yet is accurate enough in most situations. Current force fields can be categorized into fix-charge force fields and polarizable force fields. The major fix-charge force fields being currently used are Amber, CHARMM and OPLS (61-63). Fix-charge force fields treat each atom as a particle with a partial charge which is an averaged description of electron density around the atom. Bonds and angles are described using Hooke’s Law. Dihedral interactions, which determine the torsion angle of four consecutively bonded atoms, are usually described using cosine functions. The dihedral interactions are meant to be corrections for the difference of energies calculated using quantum mechanics and molecular mechanics. Improper interactions are frequently used on sp2 atoms and the atoms bonded to the sp2 atoms, which are meant to keep the four atoms in plane. Van der Waals interactions are calculated using the Lennard-Jones potential energy function. The potential energy of a protein conformation is evaluated by the following terms in Amber force field:

$$U_{potential\ energy} = U_{bonds} + U_{angles} + U_{dihedral} + U_{improper} + U_{VDW} + U_{electrostatic} \tag{1}$$

Each energy term is described as the following:

$$U_{bonds}(\vec{r}) = \sum_{bonds} K_b\,(\vec{r} - r_0)^2 \tag{2}$$

$$U_{angles}(\vec{\theta}) = \sum_{angles} K_\theta\,(\vec{\theta} - \theta_0)^2 \tag{3}$$

$$U_{dihedral}(\gamma) = \sum_{dihedral} K_\gamma\,(1 + \cos(n\gamma - \delta)) \tag{4}$$

$$U_{improper}(\vec{\varphi}) = \sum_{improper} K_\varphi\,(\vec{\varphi} - \varphi_0)^2 \tag{5}$$

$$U_{VDW}(\vec{r}) = \sum_{vdw} \varepsilon_{ij} \left[\left(\frac{Rmin_{ij}}{r_{ij}} \right)^{12} - \left(\frac{Rmin_{ij}}{r_{ij}} \right)^{6} \right] \tag{6}$$

$$U_{electrostatic}(\vec{r}) = \sum_{electrostatic} \frac{q_i q_j}{\varepsilon r_{ij}} \tag{7}$$

All constants that appear in the potential energy function above are parameters defined by a force field (61-63). Given the coordinates of atoms, the potential energy of a microstate can be calculated. The description of proteins using MM allows the application of methods derived from statistical mechanics to better understand the behaviors of a protein.

It is impossible to analytically solve an ensemble of microstates using MM. Numerical solutions such as conformation scannings, Monte Carlo simulations, and molecular dynamics (MD) simulations have been proposed. Conformation scannings and Monte Carlo simulations are widely used in searching conformations of small molecules such as organic liquid simulations and small molecule docking. The efficiency of algorithms, however, like conformation scannings and Monte Carlo, diminishes dramatically for large systems like biomolecules. MD simulation is one of the most popular numerical sampling methods used in studying the conformations and dynamics of biomolecules. In contrast to conformation scannings and Monte Carlo simulations, which spend more effort on high energy microstates, MD simulations focus on sampling conformations around local minima. For biomolecules with rigid structures, such as the folded state of proteins, local minima are more biological relevant and attract more interest. In MD simulations, the propagation of the system is governed by Newton's equation of motion. The conformational ensemble of a protein at a certain temperature can be obtained once the simulation reaches convergence (64).

Graphical processing units (GPUs) are becoming widely used in scientific computing. A GPU is made of many processing cores, each of which can handle relatively simple mathematical operations individually in a parallel fashion. The functions for potential energy used in MM and the functions for propagation used in MD simulations are relatively unsophisticated, but the amount of calculations grow exponentially as the number of atoms in the system. The calculation of potential energy and propagation on each atom can be parallelly conducted simultaneously on the processing cores of a GPU. The use of GPUs has boosted the efficiency of MD simulations by hundreds of times compared to the efficiency in the CPU era (65-67). For a 300 residue protein, it is possible to reach 10 microsecond/month using an affordable GPU card nowadays.

However, despite the huge acceleration given by GPUs, it is still very difficult to obtain a converged trajectory of IDP through typical MD simulation. Due to the relatively flat energy landscape of IDPs, constructing the structural ensemble of IDPs requires sampling through a vast space of configurations (37, 68). Besides the challenge of sufficient sampling, the high computational cost of running MD simulations on IDPs prevents rigorous examination of the accuracy of force fields used for studying IDPs. Currently, force fields still face challenges when reproducing the structural ensembles of IDPs and the unfolded states of proteins (69-73). The experimental observables of IDPs that can be measured are very limited and are not enough to constrain an ensemble of an IDP. When a force field is trained for IDPs, it is difficult to tell if the force field is overfitting the available experimental observables used for training the force field. In summary, studying an IDP through typical MD simulations is still formidable due to insufficiency in sampling and inaccuracy in force fields.

1.6 Thermodynamic integration

Alchemical free energy calculations have been used for decades to study various free energy related problems for biomolecules and drug molecules. Thermodynamic integration (TI) is one of the most fundamental methods for alchemical free energy calculations (74). The statistical mechanics underlying the TI calculations is as follows:

$$U(\lambda, X) = \lambda U_B(X) - (1-\lambda)U_A(X) \tag{8}$$

where λ represents the reaction coordinate between two chemical states A and B. X represents the coordinates of the ligand and the target protein. U is the potential energy defined by MM models. U_A and U_B are the potential energy function for chemical state A and chemical state B, respectively. $U(\lambda)$ will be reduced to U_A when $\lambda = 0$ and U_B when $\lambda = 1$. The partition function Q of the system can be expressed as:

$$Q = \sum_{X,\lambda} e^{-U(\lambda,X)/k_B T} \tag{9}$$

The free energy of this system can be written as:

$$G = -k_B T lnQ \tag{10}$$

The free energy difference between states A and B can be derived as following:

$$
\begin{aligned}
\Delta G(A \rightarrow B) &= \int_0^1 \frac{\partial G(\lambda)}{\partial \lambda} d\lambda \\
&= -\int_0^1 \frac{k_B T}{Q} \frac{\partial Q}{\partial \lambda} d\lambda \\
&= \int_0^1 \frac{k_B T}{Q} \sum_{X,\lambda} \frac{1}{k_B T} e^{-\frac{U(\lambda,X)}{k_B T}} \frac{\partial U(\lambda, X)}{\partial \lambda} d\lambda \\
&= \int_0^1 < \frac{\partial U(\lambda,X)}{\partial \lambda} >_\lambda d\lambda \qquad (11)
\end{aligned}
$$

$<\ldots>_\lambda$ is the ensemble averaged value at a particular λ value. U is the potential energy function defined by a force field.

To calculate the integral, several MD simulations under the alchemical potential energy function $U(\lambda, X)$ with different λ values are conducted. Values of $< \frac{\partial U(\lambda,X)}{\partial \lambda} >$ are then collected from each simulation at different λ values. Trapezoidal integration is usually used to obtain $\Delta G(A \rightarrow B)$. The typical rules for numerical integration should be followed here to decide the number of λ windows. A better resolution for the numerical integration can be obtained when more λ windows are added at where values of $< \frac{\partial U(\lambda,X)}{\partial \lambda} >$ have steep changes.

Methods like TI face two main challenges: inaccuracy in the force fields and the high computational cost of obtaining the ensemble averages. Over the last ten years, both force fields and computing hardware have improved significantly which allows the critical components in $\int_0^1 < \frac{\partial U(\lambda,X)}{\partial \lambda} >_\lambda d\lambda$ to be calculated at high accuracy and relative low cost.

One of the major improvements of force field in recent years is the amino-acid specific corrections on side-chain torsion angles (62, 63, 75). These corrections enable X in **equation 11**, which is the structure of protein, to be stabilized in correct conformations and allow the potential energy defined by U in **equation 11** to be calculated more accurately. The use of GPU brought MD simulations to a new level as it significantly speeds up the sampling and lower the financial cost (65-67). Ensemble averaging, which is represented by the bracket in **equation 11**, is now more easily accessible via MD simulation because of the implementation of TI on GPU platform (76). These improvements dramatically increase the feasibility of using alchemical free energy calculations in studying protein thermodynamics.

Free energy calculations can be used to calculate stability changes and binding affinity changes caused by mutations. To compare calculated free energy changes with experimental results, one can construct thermodynamic cycles as shown in **Fig. 1-7&1-8**. The red arrows represent free energy changes measured by experiments and the blue arrows are free energy changes calculated using free energy calculation. A pair of free energy calculations are conducted in two different states of the protein, for example the folded state versus unfolded state (**Fig. 1-7**) for calculating protein stability or the bound state versus unbound state (**Fig. 1-8**) for calculating protein-protein affinity. The size of the unfolded state of full-length protein is too large for MD simulations and free energy calculations. A tripeptide can be used to approximate the unfolded state of protein for calculating the effect of mutation in the unfolded state (**Fig. 1-7**) provided only local interactions are important. In order to better model the local interactions around the mutation site, the mutation site is in the center of the tripeptide. The flanking residues around the mutation site in the tripeptide are the same as the flanking residues found in the full-length protein. Deviations in the calculated vs experimental $\Delta\Delta G$ values can be due to deviation of the unfolded state from the simple peptide model or because of errors in the native state calculations. If the native state errors are small, the deviation between calculated and experimental $\Delta\Delta G$ values provides information about the unfolded state which cannot be obtained experimentally.

The absolute value obtained from a single free energy calculation, which is the blue arrow in **Fig. 1-7&1-8**, does not have any physical meaning because such value contains a baseline that is force field dependent. This baseline will be cancelled when the difference of free energy changes in two different states are taken.

1.7 Aim of the thesis

This thesis describes three examples in which alchemical free energy calculations were combined with experimentally measured free energy in order to study protein thermodynamics. The examples demonstrate that free energy calculation can contribute to a combined experimental and computational study at different levels. In Chapter 2, a controversy in the effect of Gly-to-D-Ala on the protein stability has been solved by reproducing 8 contradicting experimental stability changes using free energy calculations. Reproducing the experimental values using free energy calculations also validated the computational models used for studying the mutational effects and served as a foundation for more detailed energetic studies. This led to a general method for stabilizing proteins by targeting the unfolded state. In Chapter 3, a novel method which combines free energy calculations with experiments were developed to examine the long-range interactions in NTAIL. The strength of the interactions was quantitively and rigorously measured by using this method. The method and the insights revealed by using this method are described in this chapter. The methodology can be applied to other IDPs.In Chapter 4, free energy calculations were used to deconvolute experimentally measured free energies of HP36 so that the effect of mutations on the unfolded state can be better examined. The experimental free energy changes were measured using double mutant cycles and contain contributions from both the folded state and the unfolded state. The effect of mutation in the folded state can be obtained from the calculations of free energy changes in the folded state. The pure effect of mutation in the unfolded state can be obtained by subtracting the effect in the folded state from the overall effect. In chapter 5, this approach is applied to the N-terminal domain of the ribosomal protein L9 (NTL9). There is a large body of thermodynamic data on NTL9 and prior work has shown there are interactions in the unfolded state thus this system provides an excellent additional test of the methods. Chapter 6 outlines future challenges and outstanding issues.

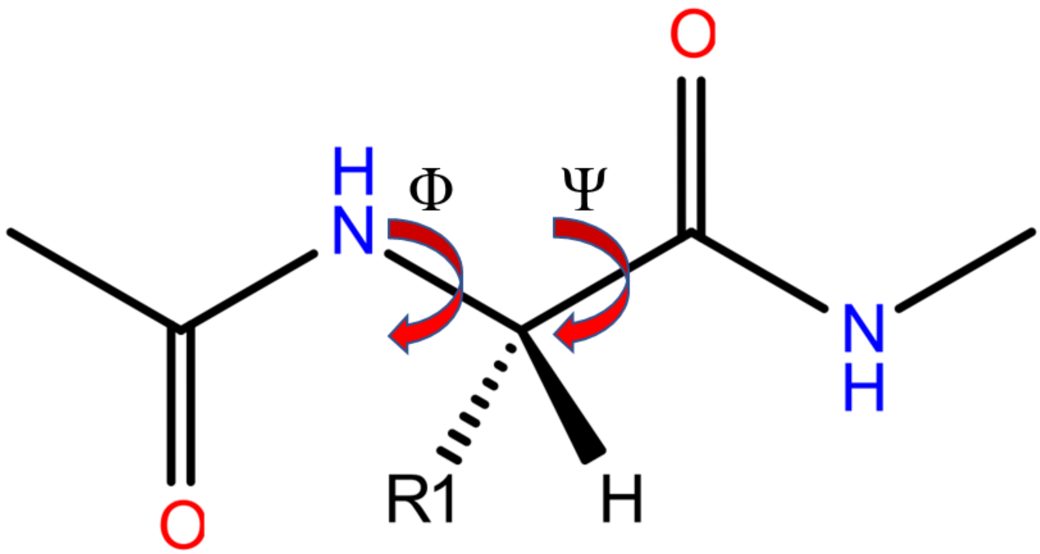


Figure 1-1. The backbone conformation of amino acids. R1 represents the side chain of amino acids. The two dihedral angles Φ and Ψ are show in red arrows.

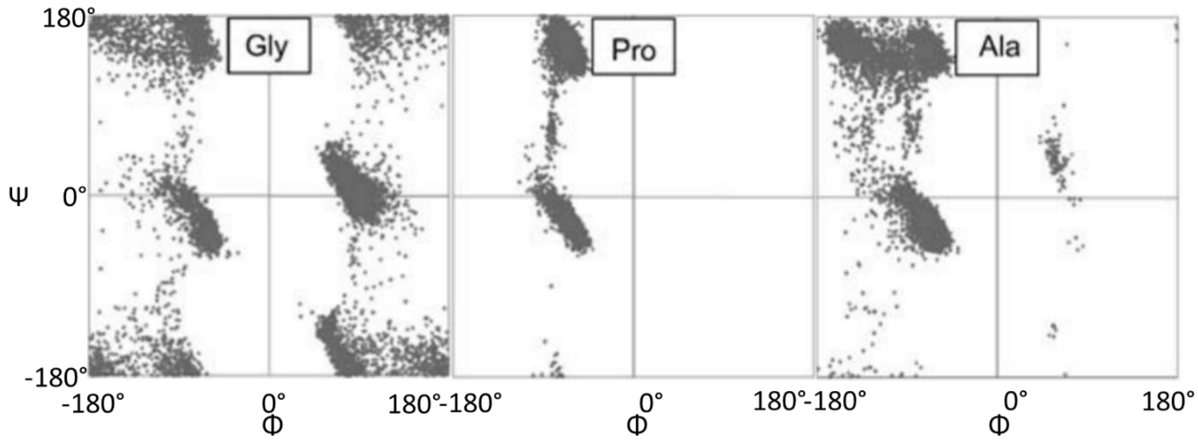


Figure 1-2. The Ramachandran plot of Gly, Pro and Ala. The figures are adapted and modified from ref (77). The black dots represent the occurrences of a (Φ, Ψ) observed in Protein Geometry Database(78), which is a database for high-fidelity X-ray structures of proteins.

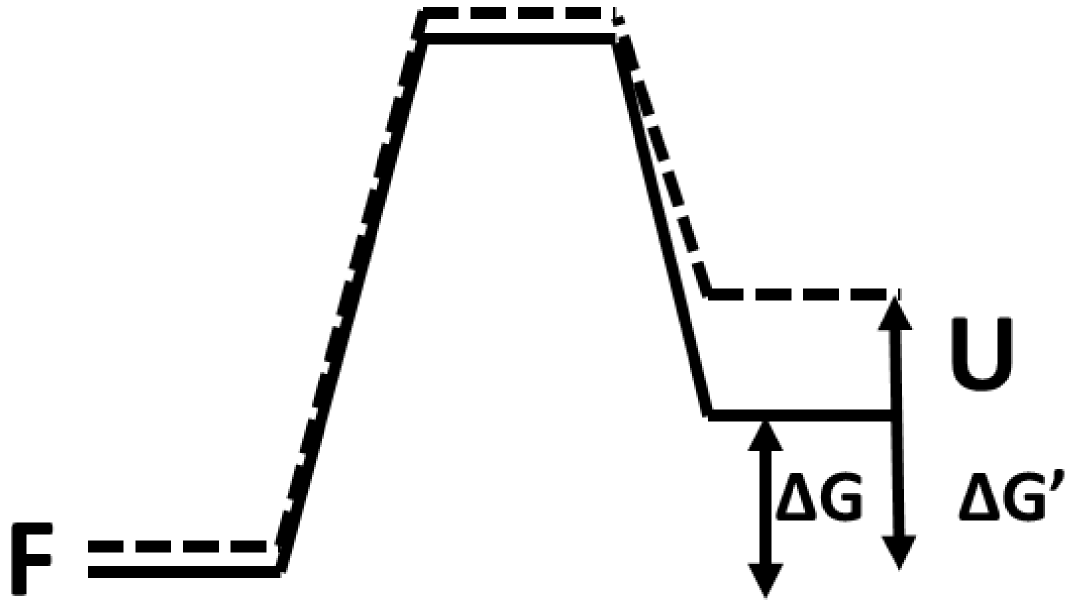


Figure 1-3. The effect on protein stability by targeting the unfolded state. F and U denote the free energy of the folded and unfolded state respectively. The change solid line and dashed line denote the free energy profiles of the protein before and after mutation respectively. A mutation that has minimal effect on the folded state and elevates the free energy of the unfolded state will lead to a net increase of protein stability ($\Delta G' > \Delta G$).

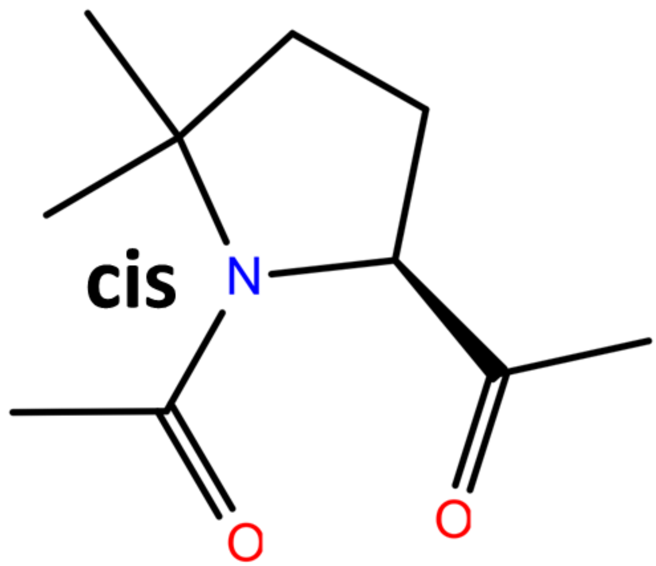


Figure 1-4. 5,5-dimethyl-L-proline in cis conformation.

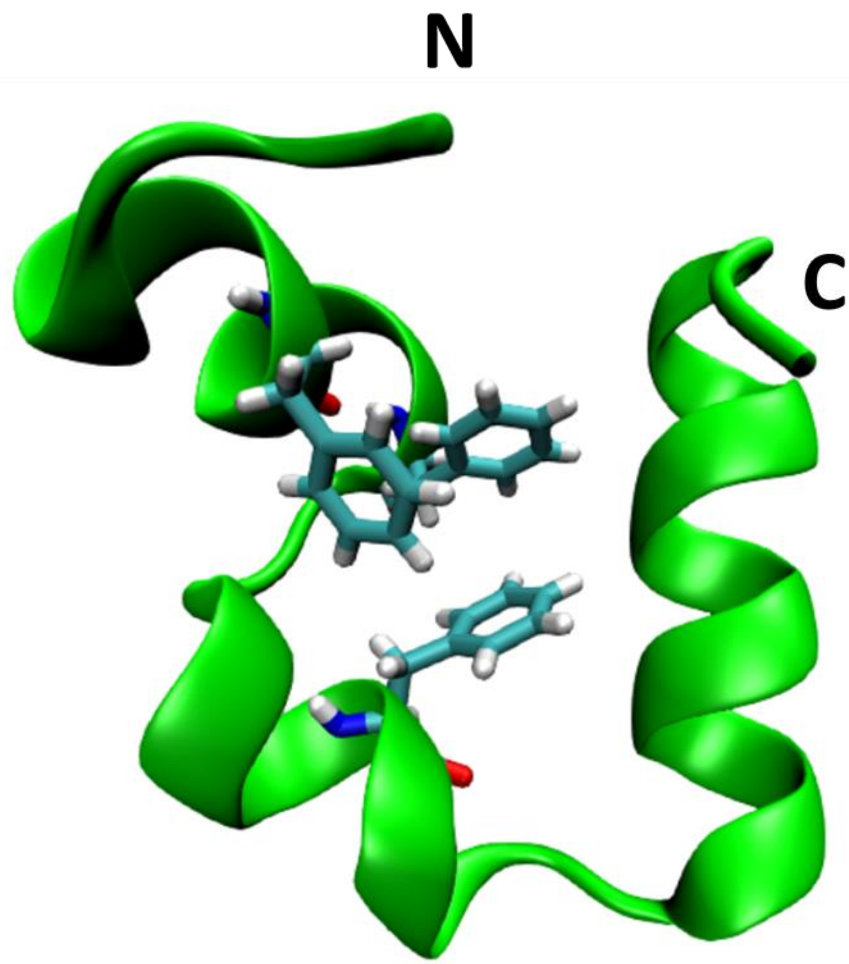


Figure 1-5. The ribbon diagram of HP36. The three Phe contribute to the hydrophobic core of HP36 are shown in licorice.

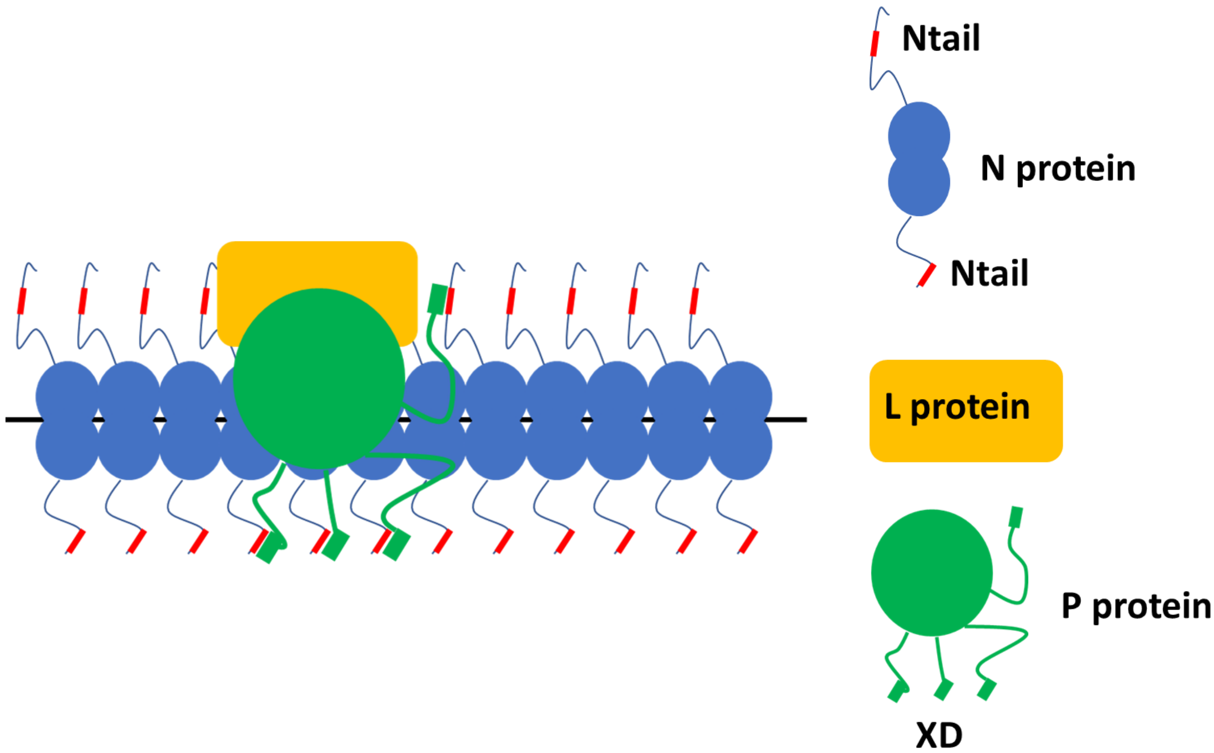


Figure 1-6. The assembly of the transcription and replication machinery of the RNA of measles virus.

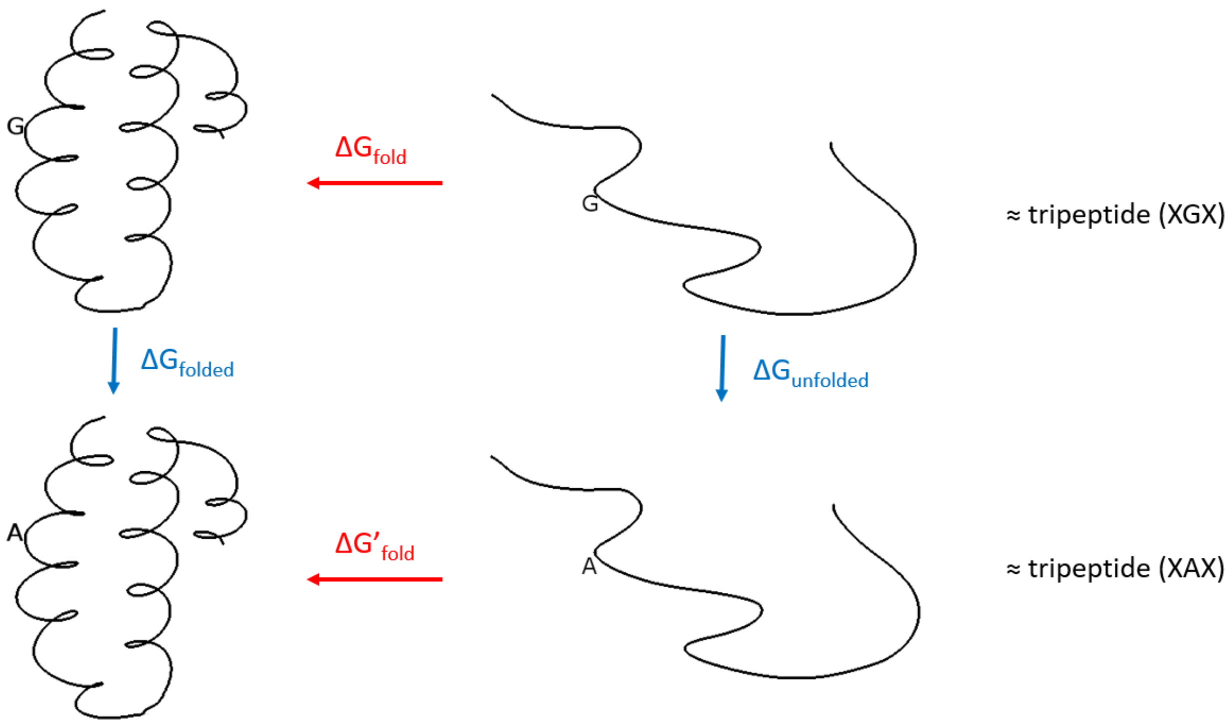


Figure 1-7. The thermodynamic cycle for the unfolding of proteins. ΔG_{fold} and $\Delta G'_{fold}$ (red arrows) are the folding free energies of the Gly variant and Ala variant of the protein respectively. ΔG_{folded} and $\Delta G_{unfolded}$ (blue arrows) are the free energy changes of the Gly-to-Ala mutations in the folded state and unfolded state of protein respectively. A tripeptide can be used to model the unfolded state of protein if the unfolded state is assumed to be highly unstructured. This can avoid direct modelling of the unfolded state of protein in MD simulations which is highly computationally expensive.

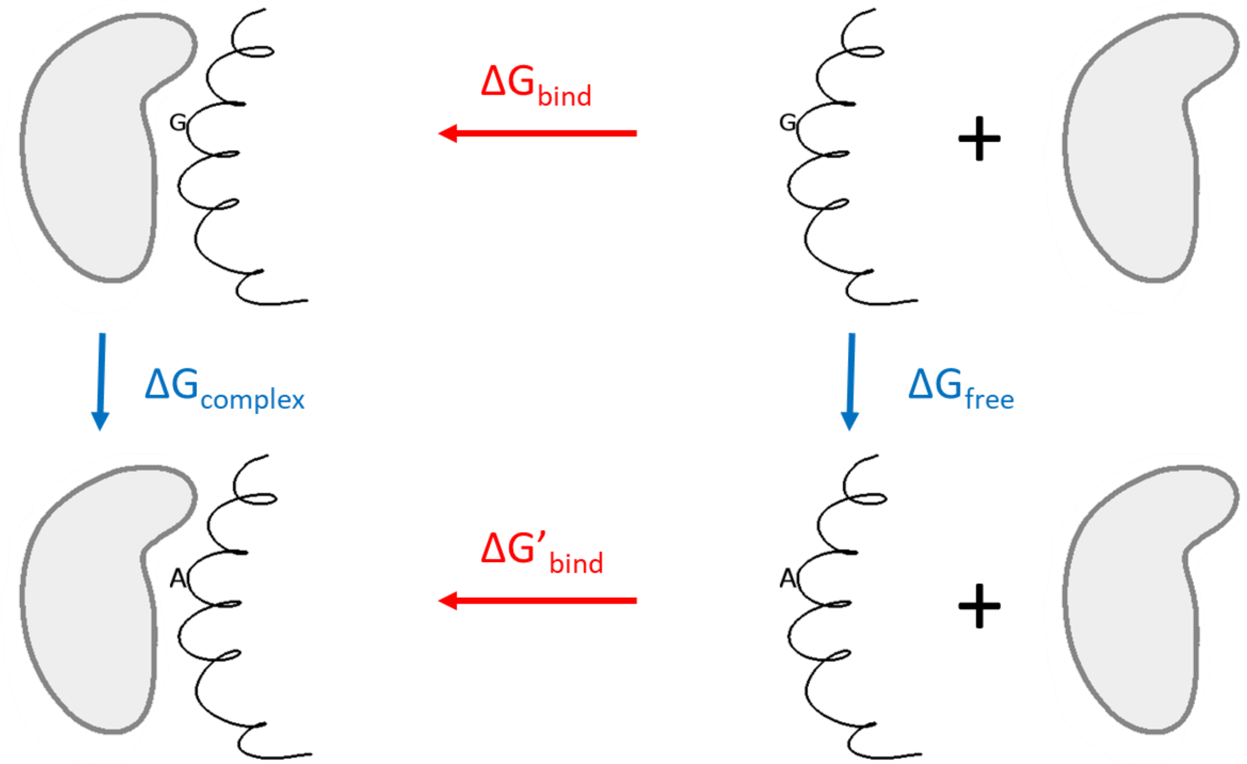


Figure 1-8. The thermodynamic cycle for the protein-protein binding interactions between a helical protein and its binding partner. ΔG_{bind} and $\Delta G'_{bind}$ (red arrows) are the binding free energies of the Gly variant and Ala variant of the helical protein respectively. $\Delta G_{complex}$ and ΔG_{free} (blue arrows) are the free energy changes of the Gly-to-Ala mutations in the complex state and free state respectively.

1.8 References

1. Sharp K (2001) Entropy-enthalpy compensation: fact or artifact? *Protein Sci* 10(3):661-667.

2. Liu L, Yang C, & Guo QX (2000) A study on the enthalpy-entropy compensation in protein unfolding. *Biophys Chem* 84(3):239-251.

3. Schmid A, *et al.* (2001) Industrial biocatalysis today and tomorrow. *Nature* 409(6817):258-268.

4. Chapman J, Ismail AE, & Dinu CZ (2018) Industrial Applications of Enzymes: Recent Advances, Techniques, and Outlooks. *Catalysts* 8(6).

5. Leader B, Baca QJ, & Golan DE (2008) Protein therapeutics: a summary and pharmacological classification. *Nat Rev Drug Discov* 7(1):21-39.

6. Rigoldi F, Donini S, Redaelli A, Parisini E, & Gautieri A (2018) Review: Engineering of thermostable enzymes for industrial applications. *APL Bioeng* 2(1):011501.

7. Daniel RM, Cowan DA, Morgan HW, & Curran MP (1982) A correlation between protein thermostability and resistance to proteolysis. *Biochem J* 207(3):641-644.

8. Parsell DA & Sauer RT (1989) The structural stability of a protein is an important determinant of its proteolytic susceptibility in escherichia-coli. *J Biol Chem* 264(13):7590-7595.

9. McLendon G & Radany E (1978) Is protein turnover thermodynamically controlled? *J Biol Chem* 253(18):6335-6337.

10. Chi EY, Krishnan S, Randolph TW, & Carpenter JF (2003) Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharm Res* 20(9):1325-1336.

11. Chi EY, *et al.* (2003) Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci* 12(5):903-913.

12. Pace CN, *et al.* (2011) Contribution of hydrophobic interactions to protein stability. *J Mol Biol* 408(3):514-528.

13. Pace CN, Grimsley GR, & Scholtz JM (2009) Protein ionizable groups: pK values and their contribution to protein stability and solubility. *J Biol Chem* 284(20):13285-13289.

14. Makhatadze GI, Loladze VV, Ermolenko DN, Chen X, & Thomas ST (2003) Contribution of surface salt bridges to protein stability: guidelines for protein engineering. *J Mol Biol* 327(5):1135-1148.

15. Strickler SS, *et al.* (2006) Protein stability and surface electrostatics: a charged relationship. *Biochemistry* 45(9):2761-2766.

16. Xiao S, *et al.* (2013) Rational modification of protein stability by targeting surface sites leads to complicated results. *Proc Natl Acad Sci U S A* 110(28):11337-11342.

17. Matthews BW, Nicholson H, & Becktel WJ (1987) Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc Natl Acad Sci U S A* 84(19):6663-6667.

18. Anil B, Song B, Tang Y, & Raleigh DP (2004) Exploiting the right side of the Ramachandran plot: substitution of glycines by D-alanine can significantly increase protein stability. *J Am Chem Soc* 126(41):13194-13195.

19. Rodriguez-Granillo A, Annavarapu S, Zhang L, Koder RL, & Nanda V (2011) Computational design of thermostabilizing D-amino acid substitutions. *J Am Chem Soc* 133(46):18750-18759.

20. Matsumura M, Becktel WJ, Levitt M, & Matthews BW (1989) Stabilization of phage T4 lysozyme by engineered disulfide bonds. *Proc Natl Acad Sci U S A* 86(17):6562-6566.

21. Sauer RT, *et al.* (1986) An engineered intersubunit disulfide enhances the stability and DNA binding of the N-terminal domain of lambda repressor. *Biochemistry* 25(20):5992-5998.

22. Wells JA & Powers DB (1986) In vivo formation and stability of engineered disulfide bonds in subtilisin. *J Biol Chem* 261(14):6564-6570.

23. Wetzel R, Perry LJ, Baase WA, & Becktel WJ (1988) Disulfide bonds and thermal stability in T4 lysozyme. *Proc Natl Acad Sci U S A* 85(2):401-405.

24. Tidor B & Karplus M (1993) The contribution of cross-links to protein stability: a normal mode analysis of the configurational entropy of the native state. *Proteins* 15(1):71-79.

25. Horng JC & Raleigh DP (2003) phi-Values beyond the ribosomally encoded amino acids: kinetic and thermodynamic consequences of incorporating trifluoromethyl amino acids in a globular protein. *J Am Chem Soc* 125(31):9286-9287.

26. Tang Y, *et al.* (2001) Stabilization of coiled-coil peptide domains by introduction of trifluoroleucine. *Biochemistry* 40(9):2790-2796.

27. Bilgicer B, Fichera A, & Kumar K (2001) A coiled coil with a fluorous core. *J Am Chem Soc* 123(19):4393-4399.

28. Arnold U, *et al.* (2003) Protein prosthesis: a nonnatural residue accelerates folding and increases stability. *J Am Chem Soc* 125(25):7500-7501.

29. Liu T, *et al.* (2016) Enhancing protein stability with extended disulfide bonds. *Proc Natl Acad Sci U S A* 113(21):5910-5915.

30. McKnight CJ, Matsudaira PT, & Kim PS (1997) NMR structure of the 35-residue villin headpiece subdomain. *Nat Struct Biol* 4(3):180-184.

31. Wang M, *et al.* (2003) Dynamic NMR line-shape analysis demonstrates that the villin headpiece subdomain folds on the microsecond time scale. *J Am Chem Soc* 125(20):6032-6033.

32. Brewer SH, *et al.* (2005) Effect of modulating unfolded state structure on the folding kinetics of the villin headpiece subdomain. *Proc Natl Acad Sci U S A* 102(46):16662-16667.

33. Chiu TK, *et al.* (2005) High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proc Natl Acad Sci U S A* 102(21):7517-7522.

34. Tang Y, Rigotti DJ, Fairman R, & Raleigh DP (2004) Peptide models provide evidence for significant structure in the denatured state of a rapidly folding protein: the villin headpiece subdomain. *Biochemistry* 43(11):3264-3272.

35. Tang Y, Goger MJ, & Raleigh DP (2006) NMR characterization of a peptide model provides evidence for significant structure in the unfolded state of the villin headpiece helical subdomain. *Biochemistry* 45(22):6940-6946.

36. Romero P, *et al.* (2001) Sequence complexity of disordered protein. *Proteins* 42(1):38-48.

37. Uversky VN (2014) Introduction to intrinsically disordered proteins (IDPs). *Chem Rev* 114(13):6557-6560.

38. Dunker AK, *et al.* (2001) Intrinsically disordered protein. *J Mol Graph Model* 19(1):26-59.

39. Oldfield CJ & Dunker AK (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* 83:553-584.

40. Dyson HJ & Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6(3):197-208.

41. Iesmantavicius V, *et al.* (2013) Modulation of the intrinsic helix propensity of an intrinsically disordered protein reveals long-range helix-helix interactions. *J Am Chem Soc* 135(27):10155-10163.

42. Jensen MR, *et al.* (2011) Intrinsic disorder in measles virus nucleocapsids. *Proc Natl Acad Sci U S A* 108(24):9839-9844.

43. Dima RI & Thirumalai D (2004) Asymmetry in the shapes of folded and denatured states of proteins. *J Phys Chem B* 108(21):6564-6570.

44. Muller-Spath S, *et al.* (2010) From the Cover: Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc Natl Acad Sci U S A* 107(33):14609-14614.

45. Tran HT, Mao A, & Pappu RV (2008) Role of backbone-solvent interactions in determining conformational equilibria of intrinsically disordered proteins. *J Am Chem Soc* 130(23):7380-7392.

46. Brucale M, Schuler B, & Samori B (2014) Single-molecule studies of intrinsically disordered proteins. *Chem Rev* 114(6):3281-3317.

47. Warner JBt, *et al.* (2017) Monomeric Huntingtin Exon 1 Has Similar Overall Structural Features for Wild-Type and Pathological Polyglutamine Lengths. *J Am Chem Soc* 139(41):14456-14469.

48. Han H, Weinreb PH, & Lansbury PT, Jr. (1995) The core Alzheimer's peptide NAC forms amyloid fibrils which seed and are seeded by beta-amyloid: is NAC a common trigger or target in neurodegenerative disease? *Chem Biol* 2(3):163-169.

49. Bertoncini CW, *et al.* (2005) Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein. *Proc Natl Acad Sci U S A* 102(5):1430-1435.

50. Bourhis JM, *et al.* (2004) The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Res* 99(2):157-167.

51. Curran J & Kolakofsky D (1999) Replication of paramyxoviruses. *Adv Virus Res* 54:403-422.

52. Kingston RL, Baase WA, & Gay LS (2004) Characterization of nucleocapsid binding by the measles virus and mumps virus phosphoproteins. *J Virol* 78(16):8630-8640.

53. Bloyet LM, *et al.* (2016) HSP90 Chaperoning in Addition to Phosphoprotein Required for Folding but Not for Supporting Enzymatic Activities of Measles and Nipah Virus L Polymerases. *J Virol* 90(15):6642-6656.

54. Morin B, Rahmeh AA, & Whelan SP (2012) Mechanism of RNA synthesis initiation by the vesicular stomatitis virus polymerase. *EMBO J* 31(5):1320-1329.

55. Bloyet LM, *et al.* (2016) Modulation of Re-initiation of Measles Virus Transcription at Intergenic Regions by PXD to NTAIL Binding Strength. *PLoS Pathog* 12(12):e1006058.

56. Kingston RL, Hamel DJ, Gay LS, Dahlquist FW, & Matthews BW (2004) Structural basis for the attachment of a paramyxoviral polymerase to its template. *Proc Natl Acad Sci U S A* 101(22):8301-8306.

57. Dosnon M, *et al.* (2015) Demonstration of a folding after binding mechanism in the recognition between the measles virus NTAIL and X domains. *ACS Chem Biol* 10(3):795-802.

58. Gely S, *et al.* (2010) Solution structure of the C-terminal X domain of the measles virus phosphoprotein and interaction with the intrinsically disordered C-terminal domain of the nucleoprotein. *J Mol Recognit* 23(5):435-447.

59. Gruet A, *et al.* (2016) Fuzzy regions in an intrinsically disordered protein impair protein-protein interactions. *FEBS J* 283(4):576-594.

60. Troilo F, Bonetti D, Bignon C, Longhi S, & Gianni S (2019) Understanding Intramolecular Crosstalk in an Intrinsically Disordered Protein. *ACS Chem Biol* 14(3):337-341.

61. Huang J & MacKerell AD, Jr. (2013) CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem* 34(25):2135-2145.

62. Maier JA, *et al.* (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput.* 11(8):3696-3713.

63. Robertson MJ, Tirado-Rives J, & Jorgensen WL (2015) Improved Peptide and Protein Torsional Energetics with the OPLSAA Force Field. *J Chem Theory Comput* 11(7):3499-3509.

64. Karplus M & McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9(9):646-652.

65. Salomon-Ferrer R, Gotz AW, Poole D, Le Grand S, & Walker RC (2013) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* 9(9):3878-3888.

66. Gotz AW, *et al.* (2012) Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J Chem Theory Comput* 8(5):1542-1555.

67. Le Grand S, Götz AW, & Walker RC (2013) SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Comput Phys Commun* 184(2):374-380.

68. Bottaro S & Lindorff-Larsen K (2018) Biophysical experiments and biomolecular simulations: A perfect match? *Science* 361(6400):355-360.

69. Rauscher S, *et al.* (2015) Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J Chem Theory Comput* 11(11):5513-5524.

70. Jensen MR & Blackledge M (2014) Testing the validity of ensemble descriptions of intrinsically disordered proteins. *Proc Natl Acad Sci U S A* 111(16):E1557-1558.

71. Henriques J, Cragnell C, & Skepo M (2015) Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J Chem Theory Comput* 11(7):3420-3431.

72. Piana S, Donchev AG, Robustelli P, & Shaw DE (2015) Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J Phys Chem B* 119(16):5113-5123.

73. Chan-Yao-Chong M, Durand D, & Ha-Duong T (2019) Molecular Dynamics Simulations Combined with Nuclear Magnetic Resonance and/or Small-Angle X-ray Scattering Data for Characterizing Intrinsically Disordered Protein Conformational Ensembles. *J Chem Inf Model* 59(5):1743-1758.

74. Kirkwood JG (1935) Statistical mechanics of fluid mixtures. *J Chem Phys* 3(5):300-313.

75. Huang J, *et al.* (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* 14(1):71-73.

76. Lee TS, Hu Y, Sherborne B, Guo Z, & York DM (2017) Toward Fast and Accurate Binding Affinity Prediction with pmemdGTI: An Efficient Implementation of GPU-Accelerated Thermodynamic Integration. *J Chem Theory Comput* 13(7):3077-3084.

77. Hollingsworth SA & Karplus PA (2010) A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol Concepts* 1(3-4):271-283.

78. Berkholz DS, Krenesky PB, Davidson JR, & Karplus PA (2010) Protein Geometry Database: a flexible engine to explore backbone conformations and their relationships to covalent geometry. *Nucleic Acids Res* 38(Database issue):D320-325.

2. Experimental and Computational Analysis of Protein Stabilization by Gly-to-D-Ala Substitution: A Convolution of Native State and Unfolded State Effects

Abstract

The rational and predictable enhancement of protein stability is an important goal in protein design. Most efforts target the folded state, however stability is the free energy difference between the folded and unfolded states thus both are suitable targets. Strategies directed at the unfolded state usually seek to decrease chain entropy by introducing cross-links or by replacing glycines. Cross-linking has led to mixed results. Replacement of glycine with an L-amino acid, while reducing the entropy of the unfolded state, can introduce unfavorable steric interactions in the folded state, since glycine is often found in conformations that require a positive φ angle such as helical C–capping motifs or type I' and II" β-turns. L-amino acids are strongly disfavored in these conformations, but D-amino acids are not. However, there are few reported examples and conflicting results have been obtained when glycines are replaced with D-Ala. We critically examine the effect of Gly-to-D-Ala substitutions on protein stability using experimental approaches together with molecular dynamics simulations and free energy calculations. The data, together with a survey of high resolution structures, suggest that the vast majority of proteins can be stabilized by substitution of C-capping glycines with D-Ala. Sites suitable for substitutions can be identified via sequence alignment with a high degree of success. Steric clashes in the native state due to the new sidechain are rarely observed, but are likely responsible for the destabilizing or null effect observed for the small subset of Gly-to-D-Ala substitutions which are not stabilizing. Changes in backbone

solvation play less of a role. Favorable candidates for D-Ala substitution can be identified using a rapid algorithm based on molecular mechanics.

Acknowledgements

The data presented in this chapter has been published (Zou, J., Song, B., Simmerling, C., and Raleigh, D.P. (2016) *J. Am. Chem. Soc.* 138, 15682–15689). This chapter contains direct excertps from the manuscript with a few adjustments. I thank Prof. Rohit Pappu for numerous helpful discussions and Prof. Robert C. Rizzo for helpful suggestions. I also thank the Laufer Center for Physical and Quantitative Biology at Stony Brook University for access to computational resources and support, and Feng Zhang, Dr. James Maier and Koushik Kasavajhala for their administration of computational resources. The work described in this chapter is a combined experimental and computational study. The experimental work was performed by Dr. Benben Song.

2.1 Introduction

A primary goal of protein design is to improve the stability of proteins since marginal stability can lead to loss of function, difficulty in formulating protein based pharmaceuticals, increased aggregation and degradation (1-5). Small stable proteins are of interest as alternative scaffolds for presenting sequences in a defined structural context and as alternatives to antibodies for drug delivery, for targeting and as analytical tools (6, 7). Stabilizing small domains can be a challenge especially if the number of sites which can be targeted is limited by the need to preserve a subset of sites for functional reasons. Stability is dictated by the free energy difference between the unfolded state and the folded state. In order to increase the free energy difference, and thus improve stability, one can stabilize the folded state or destabilize the unfolded state, however the vast majority of approaches to rational design seek to manipulate folded state energetics by exploiting the known three-dimensional structure of the folded state (8-12). The unfolded state is a dynamic ensemble, containing transient as well as longer lived elements of structure that can include both native and non-native interactions. The dynamic nature of the unfolded ensemble has made it difficult to target using rational design. Here we describe a general approach to rational protein design that exploits structurally conserved glycine residues and targets both the unfolded ensemble and the native state.

Folded state stabilization usually involves decreasing native state enthalpy, while unfolded state destabilization usually seeks to decrease its entropy. Increasing stability by decreasing the enthalpy of the folded state is more broadly studied, however, implementation of this strategy requires detailed structural information on the folded state (9, 11). A decrease in the conformational entropy of unfolded states can be achieved by adding disulfide bonds or substituting glycine with non-glycine amino acids (8, 10, 12-17). The former approach also requires tertiary structural information of the folded state, since disulfide bonds can introduce strain into the native state and have strict stereochemical requirements. In theory, the effect of adding a disulfide can be estimated using arguments based on loop entropy; the disulfide introduces a cross link in the chain and thereby reduces the configurational entropy of the unfolded state. However, introduction of a disulfide can stabilize compact conformations in the unfolded state and lead to new unfolded state enthalpic interactions. These effects, together with native state strain, often result in engineered disulfides having only a modest or even unfavorable effect on protein stability (10, 18). Complete cyclization of a protein by covalently linking the N and C termini has been employed in an attempt to enhance protein stability, but the same considerations come into play (19).

Targeting glycine residues is an attractive alternative strategy since introduction of a sidechain is a simple and effective way to decrease configurational entropy owing to the more restricted allowed region of the Ramachandran plot for an L or D amino acid relative to glycine. At room temperature, the entropical effect of restricting the backbone is roughly 0.41 kcal/mol. The approach should be effective provided that the addition of a sidechain does not lead to steric clashes in the folded state and provided the stereochemical constraints introduced by the sidechain are compatible with the native backbone geometry. The latter point is a significant issue since glycine is often located at sites which require a positive value of the backbone dihedral angle φ (20). D-amino acids are the more attractive choice when targeting glycine residues that have positive values of φ, since these conformations are disfavored for L-amino acids, but allowed for D-amino acids (21-23). Glycine residue with positive values of φ are commonly found in α-helical C-capping motifs and in type I' and II" β-turns, where a left-handed conformation (positive φ) is required (22, 24-26). These glycines can often be identified using multiple sequence alignments since they are conserved for structural reasons; helical capping motifs have specific sequence requirements and there are well established sequence rules for type I' and II" β-turns (21-24, 26, 27). Glycines located at C-caps are often solvent exposed, thus any perturbation caused by substituting with a D-amino acid should be minimal since the new side chain is less likely to make steric clashes. This potentially opens the door to rational design in the absence of structural information, however conflicting results have been reported for D-Ala substitutions.

The effect of Gly-to-D-Ala substitutions has been reported for four different proteins: the N-terminal domain of the ribosomal protein L9 (NTL9), the C-terminal Ubiquitin associated domain of HHR23A (UBA), the mini-protein construct TC5b (Trp-cage) and human erythrocytic ubiquitin (ubiquitin) (12, 17, 28). D-amino acids have also been used to stabilize small β-hairpin peptides (25). The limited experimental measurements reveal several apparent contradictions: To first order, the entropic stabilization caused by Gly-to-D-Ala substitution is expected to be system independent, but not all proteins are stabilized by Gly-to-D-Ala substitutions and a significant range of $\Delta\Delta G^{\circ}$ values have been reported for those that are. The stability of NTL9 and UBA are increased by a favorable 1.87 kcal/mol and 0.6 kcal/mol respectively when a C-capping Gly was replaced with D-Ala (12). Note, in this manuscript, we report ΔG° values of unfolding, thus positive values of $\Delta\Delta G^{\circ}$ indicate stabilization. The stability of Trp-cage was improved by 0.9 kcal/mol when G10 was substituted by D-Ala (17). However, a G35D-Ala substitution at a helical C-capping position in ubiquitin was slightly destabilizing at pH=2.5 (28). The lack of an effect was conjectured to be due to unfavorable contributions from backbone desolvation, caused by the introduction of a sidechain, that offset the decreased entropy of the unfolded state (28).

The limited data set indicates that replacement of glycines with positive φ-angles by D-Ala can be stabilizing, but it also leads to important questions: will the trend of an increase in stability be preserved if larger data sets are examined? What causes the range of values of $\Delta\Delta G^{o}$? Why does the replacement lead to no effect in ubiquitin? Can the energetic effects of a D-Ala substitution be quantitatively predicted? From a practical perspective, the key issues are whether or not it is possible to reliably and robustly predict, *a priori* which Gly to D-Ala replacements will be stabilizing, and by how much. This is critical since D-amino-acids must currently be introduced via solid phase synthesis or via chemical ligation methods.

In this study, we use a combined experimental and computational approach to systematically examine the consequences of replacing C-capping glycines with D-amino acids and develop a rapid algorithm for predicting when such substitutions will be stabilizing. Gly-to-D-Ala substitutions at the C-caps of α-helices in four additional proteins were examined, experimentally doubling the number of reported examples: the engrailed homeodomain (EH), the GA albumin-binding module (GA), the peripheral subunit-binding domain (PSBD) and the chicken villin subdomain (HP35) (29-32). These proteins are all α-folds and each contains a glycine C-capping residue with a positive φ angle (**Fig. 2-1**). EH, GA and PSBD were randomly chosen and D-Ala replacements were found to be stabilizing. The small helical protein HP35 was predicted to be destabilized by Gly-to-D-Ala substitutions based on molecular modelling and serves as a negative control. Computational modelling successfully reproduced the experimental stability changes and indicates that intra-molecular van der Waals interactions in the folded state are the reason for the wide range of $\Delta\Delta G^o$ caused by Gly-to-D-Ala substitutions. Screening a database of representative high-resolution X-ray structures shows that 95% of C-capping Gly-to-D-Ala substitutions are predicted to be stabilizing and 80% of all substitutions are predicted to enhance stability by more than 1 kT. This work shows that Gly-to-D-Ala substitutions at C-caps of α-helices, under the guidance of molecular modelling, is a general strategy for rational protein design. This works reveals the rules for stabilizing proteins via D-Ala substitutions. This "mirror image" approach to protein design is widely applicable and sites suitable for substitution can be rapidly predicted.

2.2 Methods

2.2.1 Protein Solid Phase Synthesis

The proteins and their Gly-to-D-Ala variants were chemically synthesized using Fmoc chemistry (33). Sequences of these proteins are provided below. EH, GA and PSBD have a free N-terminus and amidated C-terminus, while HP35 has a free N-terminus and free C-terminus. Peptide identity was confirmed using MALDI or ESI and purity was greater than 95%. EH, observed mass 7453.97, expected mass 7453.52; EH D-Ala, observed mass 7467.75, expected mass 7467.55; GA D-Ala, observed mass 5143.96, expected mass 5143.91; HP35, observed mass 4065.16, expected mass 4064.13; HP35 D-Ala, observed mass 4079.32, expected mass 4078.15. PSBD, observed 4400.72, expected 4402.10.

2.2.2 Sequences of the Proteins Synthesized for This Study

dA refers to D-Ala and L_N refers to nor-leucine.

EH: MDEKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKS

EH-G39D-Ala: MDEKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELdALNEAQIKIWFQN KRAKIKKS

GA: LKNAIEDAIAELKKAGITSDFYFNAINKAKTVEEVNALVNEILKAHA

GA-G16D-Ala: LKNAKEDAIAELKKAdAITSDFYFNAINKAKTVEEVNALVNEILKAHA

HP35: LSDEDFKAVFGMTRSAFANLPLWL$_N$QQHLKKEKGLF

HP35-G11D-Ala: LSDEDFKAVFdAMTRSAFANLPLWL$_N$QQHLKKEKGLF

PSBD: AMPSVRKYAREKGVDIRLVQGTGKNGRVLKEDIDAFLAGGA

PSBD-G15D-Ala: AMPSVRKYAREKdAVDIRLVQGTGKNGRVLKEDIDAFLAGGA

2.2.3 Backbone phi/psi Angles and Calculation of the Solvent Accessibility of the Gly Backbone

The φ/ψ angles of C-capping glycines were calculated by using VMD (34). The same PDB structures used for molecular dynamics simulations were used and missing hydrogen atoms were added using tLeap in Amber (35). The solvent accessible surface area (SASA) of C-capping glycines was calculated by using VMD with a water probe radii of 1.4 Å. The extended tetrapeptides were constructed using tLeap with the same local sequence as the respective full length proteins. The C-termini of the tetrapeptides were amidated and the N-termini were acetylated. Residues in the extended peptides all have φ and ψ angles equal to 180°. Fractional SASA is defined as the ratio between the SASA found for the PDB structure and the SASA found for the extended tetrapeptide.

2.2.4 Thermal and Urea/Guanidine Denaturation

The unfolding free energy of each protein was measured by CD-monitored urea/guanidine hydrochloride denaturation at 222nm under the conditions listed in **Table 2-3**. Thermal denaturation experiments were also conducted at 222nm using the same buffer and pH employed for the urea/guanidine hydrochloride denaturation experiments. The concentration of urea/guanidine was determined by measuring the refractive index on a refractometer. Urea/guanidine denaturation experiments were carried out with a titrator unit interfaced to the CD spectrometer. Unfolding curves for EH, GA, PSBD were recorded using Aviv model 62A DS and 202SF circular dichroism spectrophotometers. Unfolding curves for HP35 were recorded using an Applied Photophysics Chirascan instrument. ΔG^o of unfolding was determined by fitting the urea/guanidine denaturation curves to the following equation:

$$\theta[denaturant] = \frac{(a_n+b_n[denaturant])+(a_d+b_d[denaturant])e^{-\left(\frac{\Delta G^o([denaturant])}{RT}\right)}}{1+e^{-\left(\frac{\Delta G^o([denaturant])}{RT}\right)}} \tag{12}$$

$$\Delta G^o([denaturant]) = \Delta G^o(H_2O) - m[denaturant] \tag{13}$$

where θ is the measured ellipticity, a_n ,b_n ,a_d ,b_d are the parameters that define the signals of the native state and denatured state. $\Delta G^o([denaturant])$ is the free energy change upon unfolding as a function of denaturant and $\Delta G^o(H_2O)$ is the free energy change in the absence of denaturant. Thermal unfolding data was fit using standard methods and the Gibbs-Helmholtz equation to obtain the melting temperature T_m and ΔH^0 at T_m.

$$\theta[T] = \frac{(a_n+b_nT)+(a_d+b_dT)e^{-\left(\frac{\Delta G^o(T)}{RT}\right)}}{1+e^{-\left(\frac{\Delta G^o(T)}{RT}\right)}} \tag{14}$$

$$\Delta G^o(T) = \Delta H^o(T_m)\left(1-\frac{T}{T_m}\right) - \Delta C^o{}_p[T_m - T + Tln(\frac{T}{T_m})] \tag{15}$$

Where T_m is the melting temperature. $\Delta H^o(T_m)$ is the change of enthalpy upon unfolding at the melting temperature. $\Delta C^o{}_p$ is the change of heat capacity upon unfolding.

2.2.5 Molecular Dynamics Simulations Using an Explicit-water Model

The starting structures used for the simulations of EH, GA, HP35, NTL9, PSBD, Trp-cage, UBA and ubiquitin were obtained from the pdb files 1ENH (30) , 1PRB (31), 1WY4 (29), 2HBB (36), 2PDD (32), 1L2Y (37), 1DV0 (38) and 1UBQ (39) respectively. Residues not included in the sequences listed above were deleted from the pdb file and the actual missing residues were added by Swiss PDB (40) and equilibrated by MD simulations with restraints on all other residues. C-Terminal amidation and N-terminal acetylation was added if the studied proteins had these modifications. X-ray structures are available for EH, HP35, NTL9 and ubiquitin, while only NMR structures are available for GA, PSBD, Trp-cage and UBA. For proteins with multiple models from NMR studies, the RMSD of each model was calculated using the average conformation as the reference. The model with the lowest RMSD was chosen as the starting structure for MD simulations. Starting structures for D-Ala mutants were created using tLeap in Amber (35). Four independent MD simulations were run for each protein and for the D-Ala variant with different initial velocities, which results in eight simulations in total. The length of the simulations were 200 ns with the stepsize set to 2 fs. All simulations were performed using the Amber software package with the Amber ff14SB force field (41) and TIP3P water (42). Parameters for nor-leucine were obtained from Forcefield_NCAA (43). No ions were included in the system. All simulations were conducted under constant pressure conditions at 298K using Berendsen barostat to control pressure (44). Temperature was controlled using a weak-coupling algorithm with the coupling constant set to 1 ps (44). Truncated octahedron boxes with periodic boundary condition were used. Particle mesh Ewald methods were used to calculate electrostatic energies (45). Hydrogen atoms were constrained using the SHAKE algorithm (46). The cutoff of non-bonded interactions was set to 8 Å. The N-terminus was acetylated and C-terminus was amidated for proteins which had free termini and in which the termini were calculated to be neutral since deprotonated N-terminus and protonated C-terminus are not currently available in the Amber force field (41). Regular terminal residues defined in the Amber force field (41) were used for cases where the N and C termini were charged.

Local effects in the unfolded state were modeled as blocked tetrapeptides with sequence ACE-Xaa$_1$-Gly/dAla-Xaa$_2$-NH2. Xaa$_1$ and Xaa$_2$ are the two residues adjacent to the C-capping Gly/dAla in the full length protein sequences. This approach provides a model of purely local interactions and is not meant to mimic the actual unfolded chain. In order to enhance sampling, the tetrapeptides were simulated at 500K for 0.4ns, followed by cooling from 500K to 298K in 0.4ns and 0.4ns at 298K. This annealing cycle was repeated 120 times. Only data from 298K was collected for all cycles. These procedures were repeated thrice with different initial velocities which resulted in 3 sets of 4 independent folded state simulations and 3 sets of 120 annealing cycles of unfolded state simulations. A total of 96,000 frames from the folded state simulations and 144,000 frames from the unfolded state simulations at 298K were saved for analysis.

2.2.6 Starting Structures of PSBD, Trp-cage and UBA used for MD Simulations

PSBD, Trp-cage and UBA have multiple models obtained through NMR experiments. For each model, the backbone RMSD was calculated using VMD (34). The reference coordinates are the averaged coordinates of all the models. The models used as starting structures are as follows:

Protein	PDB code	Model number
PSBD	2PDD	Model 32
Trp-cage	1L2Y	Model 32
UBA	1DV0	Model 15

2.2.7 Assignment of Protonation States of Titratable Residues during MD Simulations

Protonation states of titratable residues were set to reflect the pH at which thermodynamic properties of proteins were measured. The H++ server was used to determine the protonation state

(47). Experimental $\Delta\Delta G^{\circ}$ have been reported for the ubiquitin variants over the pH range of 2.5 to 3.5 (28). The value of $\Delta\Delta G^{\circ}$ at pH 2.5 was compared to the calculated value since the TI approach only allows fixed protonation states. By fixing all the acidic residues and the C-terminus to be protonated, the system resembles that expected at pH=2.5.

Protonation states for titratable residues and terminus are listed in the table below. Asp, Glu, and C-termini which are not listed were fixed in the deprotonated state. Lys, Arg and N-termini which are not listed were fixed in the protonated state.

Protein	pH	Asp and Glu	His	C-terminus and N-terminus
EH	5.7			
GA	7.0		52, doubly protonated	
HP35	4.8		68, doubly protonated	
NTL9	5.5			
PSBD	8.0			Deprotonated N-terminus
Trp-cage	7.0			
UBA	6.5			
Ubiquitin	2.5	All Asp and Glu are protonated	68, doubly protonated	Protonated C-terminus

2.2.8 Free Energy Calculations

Free energy calculations were performed using non-softcore thermodynamic integration implemented in Amber (35, 48). Gly was turned into D-Ala in three stages. In the first stage, partial charges on the CA/HA2/HA3 of Gly were turned off. In the second stage, three dummy atoms were added to the disappearing glycine and van der Waals interaction of these dummy atoms were turned on so a D-Ala with no partial charges on the CA/HA/CB/HB1/HB2/HB3 atoms appeared. In the third stage, partial charges on the CA/HA/CB/HB1/HB2/HB3 atoms of D-Ala were turned on. The first and third stages have λ evenly distributed from 0.0 to 1.0 with an interval of 0.1 including 0.0 and 1.0. In order to avoid singularity at λ = 0.0 and λ = 1.0 and have more sampling at where $dV/d\lambda$ has a steep change, the second stage has λ equal to 0.00922, 0.04794, 0.115, 0.20634, 0.316, 0.43738, 0.56262, 0.68392, 0.79366, 0.88495, 0.95206, 0.99078. For the folded state, one set of the TI calculations began with the C-capping glycine in place and used the crystal structures. Dummy atoms were added to the experimental structures to give the starting structures for the second stage of the calculations. Starting structures for the third stage were obtained by changing the Gly in the experimental structures to D-Ala. The alternate set of TI calculations was derived from the last frames of a 50 ns standard MD simulations of the D-Ala mutants. The structures resulting from these simulations were converted back to the Gly containing variants to provide starting structures for the first stage of the calculations.

For the folded state, MD simulations used the same set up as the standard MD simulations described above except that the length of the simulation was set to 12ns for each window. The blocked peptides, which model local interactions in the unfolded state, were converted from Gly to D-Ala in three stages using the same λ values that were used for the folded states. The same sampling enhancement strategy described above was used for all stages and λ windows. Only data from 298K was collected. Numerical integration was performed using trapezoidal integration. Three $\Delta\Delta G^\circ$ values were obtained by dividing simulations of each λ window for the folded states and unfolded states into three blocks. Error bars for the calculated $\Delta\Delta G^\circ$ were the standard deviation of the three $\Delta\Delta G^\circ$ values.

2.2.9 Energy Decomposition and Analysis of First Shell Water Molecules

The van der Waals potential energy between Gly or D-Ala and the rest of protein was calculated by post processing MD simulation trajectories. 1-4 van der Waals interactions were considered as van der Waals interactions with a scaling factor of 0.5. $\Delta\Delta E_{vdw}$ is defined as:

$$\Delta\Delta E(VDW) = [E^{u}_{D-ala}(VDW) - E^{u}_{Gly}(VDW)] - [E^{f}_{D-ala}(VDW) - E^{f}_{Gly}(VDW)] \quad (16)$$

where "u" and "f" indicate unfolded and folded states respectively. For example, $E^{u}_{D-ala}(VDW)$ is the van der Waals interaction between D-Ala residue and the rest of the protein in the unfolded state.

The first shell water molecules were counted by using Cpptraj (49) in Amber, with a cutoff of 3.4 Å. For the folded states, the first shell water molecules around the amide nitrogen, amide proton, carbonyl carbon and carbonyl oxygen of residues i-4 to i+1 (i=Gly/D-Ala) were counted because these atoms are structurally close to the C-capping residues. For the unfolded states, the water molecules around amide nitrogen, amide proton, carbonyl carbon and carbonyl oxygen of residues i-1 to i+1 (i=Gly/D-Ala) were counted.

$$\text{Number of water molecules (unfolded} - \text{folded)} = (n^{u}_{D-ala} - n^{u}_{Gly}) - (n^{f}_{D-ala} - n^{f}_{Gly}) \quad (17)$$

Where n is the number of first shell water molecules. The error bars of $\Delta\Delta E_{vdw}$ and number of water molecules (unfolded-folded) are the standard deviation of the 3 sets of simulations.

The desolvation effect on the backbone was also quantified by using Poisson Boltzmann (PB) equation solved by DelPhi(50). The Amber ff14SB partial charges(41) and Yamagishi, J's radii set(51) were used.

$$\Delta\Delta G(bb\_solvation) = [G^{u}_{D-ala}(bb\_solvation) - G^{u}_{Gly}(bb_{solvation})] - [G^{f}_{D-ala}(bb\_solvation) - G^{f}_{Gly}(bb\_solvation)] \quad (18)$$

Since PB equation is non-linear, the solvation energy of each term on the right side of equation 7 was calculated in two steps. In the first step, we calculated the solvation energy of the whole protein with partial charges on the amide nitrogen, amide proton, carbonyl carbon and carbonyl oxygen of residues i-4 to i+1 (i-1 to i+1 for the unfolded state; i=Gly/D-Ala). In the second step, the partial charges on the amide nitrogen, amide proton, carbonyl carbon and carbonyl oxygen of residues i-4 to i+1 (i-1 to i+1 for the unfolded state; i=Gly/D-Ala) were set to 0 and the solvation energy of the whole protein was calculated again. The difference in the solvation energy obtained from these two step was considered as the solvation energy of the backbone around the Gly/D-Ala.

2.2.10 Calculation of $\Delta\Delta E_{vdw-gb}$ Using an Implicit-solvent Model

The length of the simulations were 5 ns with stepsize set to 1fs. Amber ff14SBonlysc (52) was used and igb was set to 8 which corresponds to GBneck2 implicit solvent model (53). Mbondi3 radii set was used (53). Simulations were conducted under 200K due to low thermostability of proteins in the implicit-solvent model used here(52). Langevin dynamics was employed with the collision frequency set to 1 ps^{-1}. No cutoff of non-bond interactions was used. The salt concentration was set to 0.0 M.

For the experimentally tested proteins (EH, GA, HP35, NTL9, PSBD, Trp-cage, UBA and ubiquitin), the starting structures were prepared in the same way as for the simulations in explicit solvent except no solvent was added. For the 120 proteins and their D-Ala variants listed in **Table S3**, any selenomethionines were converted to methionines and all acidic residues were deprotonated and all basic residues except histidines were protonated. The protonation states of histidines depends on whether the hydrogen on δ or ε nitrogen is resolved by X-ray. If neither of the hydrogens is resolved, the ε nitrogen was protonated. Disulphide bonds were added as indicated by the authors of the structures. All non-protein molecules and ions were deleted. Local effects in the unfolded states of proteins were modeled as blocked tetrapeptides. The tetrapeptides were simulated at 400K for 0.4ns, followed by cooling from 400K to 200K in 0.4ns and 0.4ns at 200K. This annealing cycle was repeated 160 times. The van der Waals potential energy between Gly or D-Ala and the rest of protein was calculated by post processing MD simulation trajectories. 1-4 van der Waals interactions were considered as van der Waals interactions instead of bonded interactions. $\Delta\Delta E_{vdw\_gb}$ is defined as:

$$\Delta\Delta E(VDW\_gb) = [E^{u}_{D-ala}(VDW\_gb) - E^{u}_{Gly}(VDW\_gb)] - [E^{f}_{D-ala}(VDW\_gb) - E^{f}_{Gly}(VDW\_gb)] \quad (19)$$

where "u" and "f" indicate unfolded and folded states respectively. For example, $E^{u}_{D-ala}(VDW\_gb)$ is the van der Waals interaction between the D-Ala residue and the rest of the protein in the unfolded state calculated using the implicit-solvent model.

For the 8 experimentally tested proteins, each $E^{f}_{D-ala}(VDW\_gb)$ value and each $E^{f}_{Gly}(VDW\_gb)$ value is the average over 100,000 frames from 10 independent simulations with different random number seeds for Langevin dynamics. For the 120 target proteins and their variants, $E^{f}_{D-ala}(VDW\_gb)$ values and $E^{f}_{Gly}(VDW\_gb)$ values were averaged over 30,000 frames from 3 independent simulations. For all of the proteins, $E^{u}_{D-ala}(VDW\_gb)$ values and $E^{u}_{Gly}(VDW\_gb)$ values were averaged over 40,000 frames collected from the simulations at 200K.

2.2.11 Protein Chains Dataset and $\Delta\Delta E_{vdw\_gb}$

All protein chains listed here are non-redundant protein chains with BLAST (54) pvalue less than 10e-7. According to the authors of the structures, all of the protein chains are monomeric. All proteins have at least one α-helical C-capping Gly. The criteria for defining a helix was at least 5 sequential residues with $-140^\circ \leqslant \varphi \leqslant -30^\circ$ and $-90^\circ \leqslant \psi \leqslant 45^\circ$. A C-capping Gly is the first non-helical residue at the C-terminus of a helix with $20^\circ \leq \varphi \leq 125^\circ$ and $-45^\circ \leq \psi \leq 90^\circ$ (23). $\Delta\Delta E_{vdw\_gb}$ values were only calculated for proteins with high sequence diversity. In order to do so, a table of sequence redundancy in protein data bank was obtained from Molecular Modelling Database (55). A representative of each non-redundant sequence was chosen according to the ranking provided by this table.

2.3 Results

2.3.1 Proteins are usually stabilized by Gly-to-D-Ala substitution.

Published results on a limited set of proteins indicate a range of effects for Gly-to-D-Ala substitution at C-capping sites. However, the number of systems tested to date is too small to draw general conclusions. In order to gain better insight into the consequences of Gly-to-D-Ala substitutions at C-capping sites, Gly-to-D-Ala substitutions were examined in another four proteins (EH, GA, PSBD and HP35). All of these domains have been shown to fold reversibly in a 2-state fashion (56-59). Like NTL9, UBA and Ubiquitin, these proteins all have a C-capping glycine that is solvent exposed as judged by standard accessible surface area algorithms (**Figure 2-1**). The φ/ψ angles and the solvent accessibility of all of the glycine sites studied are provided in the supporting information (**Table 2-2)**. Thermal and denaturant induced unfolding curves of EH, GA, HP35, PSBD display sigmoidal transitions and all can be fit by standard methods to extract unfolding free energies (**Table 2-1**, **Figures 2-7** and **2-8)**. The stability of EH G39D-Ala, GA G16D-Ala and PSBD G15D-Ala are 0.64 kcal/mol, 0.81 kcal/mol and 1.25 kcal/mol higher than the respective wild-type. HP35 G11D-Ala is 0.38 kcal/mol less stable than wild-type HP35, but HP35 was intentionally selected as a negative control using the computational approach described below. The experimental measurements on these four additional proteins, especially the inclusion of an additional example (HP35) in which D-Ala substitution is destabilizing, provide a more robust test set for the computational studies described in the next several paragraphs.

Five of the six proteins which were randomly chosen without computational guidance exhibit enhanced stability when a C-capping Gly is replaced by D-Ala, suggesting that Gly-to-D-Ala substitutions at C-capping sites are likely to improve protein stability. Left unanswered are the questions why there is a significant range of $\Delta\Delta G^{\circ}$ values and why are HP35 and ubiquitin destabilized?

2.3.2 Gly-to-D-Ala substitutions can modulate $\Delta\Delta G^{\circ}$ via other interactions in addition to entropic stabilization.

Recent computational work reported that Gly-to-L-Ala substitution entropically destabilizes the unfolded state by $-T\Delta S = 0.3$ kcal/mol when the unfolded states are modeled as tri and pentapeptides (60) , while earlier work provide estimates ranging from 0.05 to 0.72 kcal/mol (61-64). The wide range of experimental unfolding free energy changes (0.39 kcal/mol destabilizing to 1.87 kcal/mol stabilizing) argues that interactions beyond entropic destabilization of the unfolded state play an important role in determining the change. A range of effects could counteract or supplement the entropic stabilization of replacing a C-capping Gly. Introduction of a sidechain at a C-capping Gly site can lead to increased desolvation of the polypeptide backbone, a process which is energetically unfavorable (28). All else being equal, desolvation in the native state will destabilize a protein. However, desolvation of the backbone in the folded state is likely compensated by desolvation of the backbone in the unfolded state. Moreover, the desolvation penalty may also be compensated by new favorable intramolecular interactions such as buried hydrogen bonds or favorable van der Waals interactions. Desolvation of the backbone is thus unlikely to be the sole reason for the wide range of experimental $\Delta\Delta G^\circ$ values. On the other hand, unfavorable van der Waals interactions, such as steric clashes between D-Ala and other residues in the folded state can offset the decrease of entropy in the unfolded states. These new folded state interactions will usually be alleviated upon unfolding and are less likely to perturb the unfolded state. We hypothesized that a significant contribution to the difference in $\Delta\Delta G^\circ$ values reflects differences in van der Waals interactions between the C-capping Gly/D-Ala and the rest of the protein in the folded state.

In order to test our hypothesis, molecular dynamics simulations (MD) of wild-type proteins and their D-Ala variants were conducted using the Amber ff14SB force field. MD simulations were also conducted for simplified unfolded state models to account for local unfolded state effects. Per-residue energy decomposition provided an estimate of the intramolecular van der Waals energy (E_{vdw}) contributed by C-capping Gly/D-Ala to the total potential energy of the protein. New unfavorable intramolecular van der Waals interactions in the folded state caused by the D-Ala sidechain lead to a negative value of $\Delta\Delta E_{vdw}$, while new favorable intramolecular van der Waals interactions in the folded state lead to a positive $\Delta\Delta E_{vdw}$ value. A good correlation between $\Delta\Delta E_{vdw}$ and $\Delta\Delta G^\circ$ is expected if the variation in $\Delta\Delta G^\circ$ values is determined by whether or not the D-Ala residue generates new contacts, and how strong these interactions are. $\Delta\Delta E_{vdw}$ can be calculated from snapshots derived from the MD simulations, while the contribution of backbone desolvation to $\Delta\Delta G^\circ$ can be studied by counting the number of water molecules that are blocked from interacting with the peptide backbone at the C-capping site in the folded and unfolded states using snapshots from the MD simulations. The difference provides an estimate of the net desolvation effect. It is important to validate the models used for these analyses and the applicability of the force field employed with more rigorous methods. Consequently, we first tested if our MD simulations were sufficiently converged and our force field accurate enough to reproduce the experimental data using thermodynamic integration (TI) free energy calculations.

2.3.3 Thermodynamic integration validates more approximate computational models and provides further insight into C-capping energetics

The model used for the unfolded states are tetrapeptides with neutral capping groups and the length of the MD simulations can only reach a time scale that is much smaller than the experimental time scale. A recently parametrized force field was chosen in this study, but, like all force fields, is still an approximate description of molecules (41). Therefore, we tested our models by asking if we can reproduce the experimental values of $\Delta\Delta G^\circ$ using TI. 34 λ windows were simulated for 12 ns each. TI is computationally expensive and reaching complete ergodic convergence in each λ window is unlikely, thus two different starting structures of each protein were used for two independent TI calculations in order to evaluate precision. For each protein, one of the starting structures was the PDB structure, while the other one was the last frame of a 50 ns MD simulation. (supplemental information).

Similar values of $\Delta\Delta G^\circ$ were obtained for a given protein independent of the starting structure chosen, suggesting that the TI calculation has reached reasonable convergence during the time-scale of the simulations (**Figure 2-2A)**. The only significant difference between $\Delta\Delta G^\circ_{calc}$ values determined using the different starting structures occurs for EH. We believe the effect is due to the poorly resolved N-terminus of EH in the X-ray structure rather than issues with the computational models implemented here. Residues 1-4 are unresolved and not shown in the crystal structure, while residues 5-7 are resolved, but with low confidence (30). We appended the 4 missing residues as an extended peptide to the crystal structure and conducted a MD simulation with restraints on all resolved residues to relax the four appended residues. The last frame of this restrained MD simulation was used as the starting structure for one of the TI calculations for EH (**Figure 2-2A** red bar). Following the restrained MD simulation, Gly 39 was changed to D-Ala and unrestrained MD simulation was carried out to fully relax the conformation. The last frame of this simulation was used as the starting structure for the other TI calculation of EH (**Figure 2-2A** cyan bar). During the unrestrained MD simulation, residues 1-7 formed contacts with Gly39 or D-Ala39; this was not observed during the restrained MD simulation. The difference in the calculated $\Delta\Delta G^\circ$ of EH may be caused by the difference in the extent of relaxation of the starting structures. Since residues 1-7 in the PDB structures are either unresolved or poorly resolved, the fully relaxed structure is likely a better representation of the structure of EH. The better agreement between $\Delta\Delta G^\circ_{exp}$ and $\Delta\Delta G^\circ_{cal}$ when the fully relaxed structure was used as starting structure is consistent with this hypothesis.

A small root-mean-square error of 0.23 kcal/mol is obtained for the complete set of $\Delta\Delta G^\circ_{exp}$ and $\Delta\Delta G^\circ_{cal}$ values calculated using the last frames of 50 ns MD simulations as the starting structures (**Figure 2-2B**). This indicates that the simplified unfolded state model, sampling sufficiency and choice of force field provide accurate energetics for these systems. The good agreement also argues that the large span in experimental $\Delta\Delta G^{\circ}$ values is neither caused by complexity in the unfolded states nor by the different conditions and methods used for the experimental protein stability measurements since a simplified model for the unfolded states and a consistent computational approach were able to reproduce the experimental trends.

To examine whether the system dependency of experimental $\Delta\Delta G^{\circ}$ is caused by the free energy changes in the folded state or unfolded state. The correlation between the experimental $\Delta\Delta G^{\circ}$ and the free energy change caused by Gly-to-D-Ala mutations in the folded (ΔG_{folded}) and unfolded state ($\Delta G_{unfolded}$) was examined. A strong correlation between experimental $\Delta\Delta G^{\circ}$ and ΔG_{folded} was found. However, there is no correlation between $\Delta\Delta G^{\circ}$ and $\Delta G_{unfolded}$. (**Fig. 2-3**) Thus, the source of system dependency originated from the folded state of the proteins.

2.3.4 The calculated change in van der Waals energy, $\Delta\Delta E_{vdw}$, is strongly correlated with $\Delta\Delta G^{\circ}$, but $\Delta\Delta G^{\circ}$ does not correlate with predicted desolvation effects.

To test our hypothesis that the entropic stabilization is modulated by variation in van der Waals interactions, $\Delta\Delta E_{vdw}$ values were calculated from the MD simulations. There is a strong correlation between $\Delta\Delta E_{vdw}$ and the $\Delta\Delta G^{o}$ values obtained experimentally or computationally with correlation coefficients of 0.89 in both cases (**Figure 2-4**). The results strongly support the hypothesis that van der Waals interactions between the D-Ala/Gly site and the rest of the protein play an important role in determining $\Delta\Delta G^{\circ}$.

In order to examine potential correlations between the extent of backbone desolvation and the $\Delta\Delta G^{\circ}$ values, the first shell water molecules around backbone atoms in both the folded and unfolded states were counted. The difference in the number of water molecules blocked by D-Ala relative to Gly in the unfolded states and folded states (unfolded-folded) provides an estimate of the net desolvation effect of the new sidechain. Since the methyl group in D-Ala is non-polar, the mutation from Gly-to-D-Ala only changes the water accessibility of the backbone and counting the number of water around backbone is a reasonable metric for measuring desolvation effects. The calculations were performed by averaging over the last 160 ns of 12 independent MD simulations for the folded state and 144 ns of MD simulations for the unfolded state of each protein. No significant correlation is observed with $\Delta\Delta G^{\circ}$ values. The correlation coefficient for the number of waters blocked by D-Ala and $\Delta\Delta G^{\circ}_{calc}$ is only 0.16 and is just 0.17 for the correlation with $\Delta\Delta G^{\circ}_{exp}$ (**Figure 2-5**). If the desolvation effects in the unfolded state are disregarded and only the number of blocked waters in the folded state are counted, the correlation between $\Delta\Delta G^{\circ}_{calc}$ or $\Delta\Delta G^{\circ}_{exp}$ and the number of waters blocked by D-Ala relative to Gly is not improved, with correlation coefficients of 0.20 and 0.16 respectively. For three of the proteins (EH, HP35 and GA) the uncertainty, defined here as the standard deviation of the three sets of simulations with 4 independent simulations in each set, in the number of waters blocked by D-Ala in the unfolded and folded states is relatively large. However, this does not affect the conclusion that desolvation effects are not correlated with $\Delta\Delta G^{\circ}$. The good convergence in the $\Delta\Delta G^{\circ}_{calc}$ values in the absence of good convergence in the number of blocked waters reinforces that there is unlikely to be a significant net contribution of desolvation to $\Delta\Delta G^{\circ}$ for the systems studied here.

In principle, Poisson-Boltzmann (PB) based calculations could be used to estimate desolvation effects(65), however we observed during the 200 ns MD simulations of the folded states that subtle changes in conformation can lead to a significant change in the calculated PB desolvation energy of the backbone atoms owing to the long range nature of electrostatic interactions. This results in poor convergence for the PB calculations if the fluctuations in conformation are on the same time scale of the MD simulations and leads to large error bars for PB based calculations of desolvation effects. EH and HP35 showed poor convergence in the PB calculations. The other six proteins have relatively good convergence, but no correlation between the desolvation effects calculated by PB and $\Delta\Delta G^{\circ}_{exp}$ was observed (r=0.28, p=0.58, slope=0.2) (**Figure 2-10**). The small slope indicates that differences in the PB desolvation energy do not make a contribution to the differences in $\Delta\Delta G^{\circ}$. The good convergence in the $\Delta\Delta G^{\circ}_{calc}$ values in the absence of convergence in the PB calculated solvation energy for all proteins further reinforces our conclusion that it is unlikely that desolvation makes a significant contribution to the range of $\Delta\Delta G^{\circ}$ values observed for the systems studied here.

2.3.5 The rapid screening of target proteins for D-Ala substitutions; a designed negative control helps to demonstrate proof of principle

It is prohibitively expensive to generate entire ensembles from an MD trajectory in explicit solvent in order to calculate $\Delta\Delta E_{vdw}$ values for a large set of proteins. Instead, a method which estimates $\Delta\Delta E_{vdw}$ in a time-efficient manner was developed in order to enable rapid screening of proteins for sites suitable for D-Ala substitution. The method was used to identify the HP35 D-Ala11 mutant as a negative control. The approach exploits the strong correlation between $\Delta\Delta E_{vdw}$ and $\Delta\Delta G^{\circ}$ identified above and uses a more rapid method to calculate $\Delta\Delta E_{vdw.}$ We calculated $\Delta\Delta E_{vdw\_gb}$, which like $\Delta\Delta E_{vdw}$, quantifies the contribution of the intramolecular van der Waals energy to $\Delta\Delta G^{\circ}$, but is obtained by running a short implicit-solvent simulation(53) instead of using a large ensemble from a long explicit-solvent MD simulation. The correlation between $\Delta\Delta E_{vdw}$ and $\Delta\Delta E_{vdw\_gb}$ is 0.84 (**Figure 2-11**) for the 8 systems in **Figure 2-4**. Although the implicit-solvent model is more coarse-grained than the explicit-solvent model and the length of simulation is significantly decreased, calculation of $\Delta\Delta E_{vdw\_gb}$ for a range of proteins should allow one to predict trends of $\Delta\Delta G^{\circ}_{exp}$ for hundreds of proteins in a time-efficient manner, provided the correlation between $\Delta\Delta E_{vdw\_gb}$ and the known $\Delta\Delta G^{\circ}_{exp}$ values is good. If desired, one can conduct further analysis of promising sites using longer MD simulations with explicit solvent or TI.

As shown in **Figure 2-6**, $\Delta\Delta E_{vdw\_gb}$ values (positive values represent net stabilization) are strongly correlated with the known values of $\Delta\Delta G^{\circ}_{exp}$ (r=0.94) (**Figure 2-6**). The strong correlation between $\Delta\Delta E_{vdw\_gb}$ and $\Delta\Delta G^{\circ}_{exp}$ further supports our hypothesis that the perturbation of van der Waals interactions are correlated with the effect of Gly-to-D-Ala substitutions on stability.

The strong correlation between $\Delta\Delta E_{vdw\_gb}$ and $\Delta\Delta G^{\circ}_{exp}$ (**Figure 2-6**) indicates that linear regression can be used to predict the $\Delta\Delta G^{\circ}_{exp}$ values from their $\Delta\Delta E_{vdw\_gb}$ values using the empirical function:

$\Delta\Delta G^{\circ}_{exp}$ (kcal/mol) = 1.89 * $\Delta\Delta E_{vdw\text{-}gb}$ + 0.05

We examined a set of 120 monomeric proteins of less than 130 residues, which have structures determined at 2.0 Å resolution or better and at least one helix with a C-capping Gly. $\Delta\Delta E_{vdw\_gb}$ values were calculated for proteins with high sequence diversity. In all, 160 C-capping sites were analyzed (**Table 2-4**) and $\Delta\Delta E_{vdw\_gb}$ values ranging from -0.35 to 1.67 kcal/mol were obtained (**Figure 2-7A**). Here, negative values indicate a net destabilization and positive values reflect a net stabilization. The distribution of predicted $\Delta\Delta G^{\circ}$ values is plotted as a histogram in **Figure 2-7B**. Overall, 95% of the substitutions are predicted to lead to increased stability. Furthermore, ~80% of C-capping Gly-to-D-Ala substitutions in monomeric proteins will result in significant stabilization larger than 1kT.

From this distribution we selected the helical subdomain of the villin headpiece (HP35) as a negative test case, since it was one of the few proteins for which a D-Ala substitution was predicted to be destabilizing (**Figure 2-12**). $\Delta\Delta E_{vdw\_gb}$ for the replacement of Gly by D-Ala in HP35 was -

0.31 kcal/mol (negative values represent net destabilization), which is comparable to the value for ubiquitin (**Figure 2-6**). As noted above, HP35 G11D-Ala has an experimentally measured stability 0.39 kcal/mol lower than wild-type HP35 (**Table 2-1**), confirming the computational prediction made prior to experiments.

2.4 Conclusions

Our analysis indicates that the energetics of C-capping interactions involve an interplay between two competing factors. Glycine residues are selected for such sites because they are able to adopt positive values of φ, but the choice of glycine introduces packing defects in the native state. The extremely high conservation of C-capping sites indicates that the evolutionary pressure to maintain the ability to adopt a positive value of phi at these sites leads to tolerance of packing defects in the structure. This highlights that protein stability includes compromises between competing interactions. Our results clearly show that Gly-to-D-Ala substitutions in C-capping motifs stabilize proteins when the folded state is not perturbed by unfavorable van der Waals interactions. The stability of EH, GA, NTL9, PSBD, Trp-cage and UBA were improved by 0.59 to 1.87 kcal/mol by Gly-to-D-Ala substitutions. Van der Waals interactions make a significant contribution to the observed spread in $\Delta\Delta G^{\circ}$ values. The fact that TI calculations quantitatively reproduced the experimentally observed effects, including the destabilization of HP35 and ubiquitin, argues that the range of reported $\Delta\Delta G^{o}$ values are not caused by variation in experimental protocols or complex effects in the unfolded state. The D-Ala variants of HP35 and ubiquitin were destabilized due to new unfavorable folded state van der Waals interactions that counteract the entropic stabilization. The systems studied here are two state folding but the general principles, unfolded state destabilization via entropic effects and native state stabilization by new favorable Van der Waals interaction also apply to proteins that fold via intermediates.

An important practical observation from this work is that steric clashes may still be generated by D-Ala substitution even if a C-capping glycine is identified as solvent exposed by measuring its solvent accessible surface area (SASA) (**Table 2-3**). The effect arises because the repulsive part of the van der Waals potential energy has a strong distance dependence, with the potential energy increasing rapidly as the distance between two atoms decreases. For example, moving a β-carbon from 3.2 Å to 2.8 Å from a carboxyl oxygen results in an increase in van der Waals potential energy of 1.9 kcal/mol using the Lennard-Jones potential in the Amber ff14SB force field (41). This indicates that a more quantitative method than measuring SASA should be used when predicting the consequence of Gly-to-D-Ala substitutions at C-caps of α-helices.

Does the observation that the effects of the D-Ala substitutions can be predicted accurately using a highly simplified model of the unfolded state imply that the unfolded state is devoid of structure or long range contacts or residual structure? The answer is no; the data simply argues that the substitutions do not significantly impact the energetics of other unfolded state interactions; indeed residual structure has been detected in the unfolded states of several of the proteins studied (36, 66-68).

In this study, experimental values of $\Delta\Delta G^{\circ}$ have been successfully reproduced by using molecular modelling for all proteins tested. These examples show that *in silico* molecular modelling and design serve as an excellent complement to experimental studies, and can allow one to rationally target unfolded state interactions. Predicted $\Delta\Delta G^{\circ}$ values of a large data set of structures indicate that most proteins will be stabilized by Gly-to-D-Ala substitutions at C-capping sites, opening the door to mirror image protein design.

C-capping glycines are strongly conserved in protein structures and can be identified by multiple sequence alignments, thus they can often be identified in the absence of structural information. The analysis presented here demonstrates that the replacement of such glycines is expected to be stabilizing 95% of the cases and to be significantly stabilizing 80% of the cases. This expected success rate is considerably better than has been observed with consensus method based on multiple sequence alignment and is comparable to the most successful consensus method which take into account co-variation, suggesting that rational protein design is possible in the absence of structural information (69, 70).

Table 2-1 Thermodynamic properties of EH, GA, HP35, PSBD and their D-Ala variants.

Protein	ΔG° of unfolding at 25 ^{o}C (kcal/mol)	m (kcal/mol M^{-1})	Tm (^{o}C)	$\Delta H^{\circ}(Tm)$ (kcal/mol)
EH	$1.91 \pm 0.03^{(1)}$	0.61 ± 0.01	55.6 ± 0.18	32.5 ± 0.74
EH G39D-Ala	$2.55 \pm 0.13^{(1)}$	0.66 ± 0.03	60.7 ± 0.38	33.1 ± 1.39
GA	$4.71 \pm 0.16^{(2)}$	1.00 ± 0.03	ND	ND
GA G16D-Ala	$5.52 \pm 0.19^{(2)}$	1.02 ± 0.04	ND	ND
PSBD	$2.75 \pm 0.07^{(1)}$	0.67 ± 0.01	52.5 ± 0.14	29.6 ± 0.51
PSBD G15D-Ala	$4.00 \pm 0.34^{(1)}$	0.73 ± 0.07	61.3 ± 0.23	31.9 ± 0.84
HP35	$2.47 \pm 0.12^{(1)}$	0.38 ± 0.03	76.1 ± 1.78	23.8 ± 0.57
HP35 G11D-Ala	$2.08 \pm 0.13^{(1)}$	0.45 ± 0.03	61.2 ± 1.34	21.7 ± 1.33

(1) Determined by urea denaturation; (2) Determined by GdnHCl denaturation; ND: Not determined.
Uncertainties represent the standard error of the fit.

Table 2-2. Backbone phi/psi and solvent accessibility of Gly

Protein	φ (°)	ψ (°)	SASA (Å^2)	SASA in extended tetrapeptide (Å^2)	Fractional SASA (%)
EH	51.8	35.8	64.0	88.5	72.4
GA	107.8	-21.7	55.5	120.8	45.9
HP35	75.7	19.8	66.9	73.1	91.5
NTL9	70.4	26.9	36.7	98.9	37.1
PSBD	84.0	48.1	63.5	94.0	67.6
Trp-cage	119.9	10.0	31.6	113.0	28.0
UBA	127.0	1.3	64.2	95.9	67.0
Ubiquitin	81.2	5.2	53.7	100.5	53.4

Table 2-3. Conditions for thermal and urea/guanidine denaturation experiments

Protein	Urea/guanidine hydrochloride	Buffer	pH	Temperature (˚C)
EH	Urea	50mM sodium acetate	5.7	5
GA	Guanidine hydrochloride	50mM phosphate	7.0	25
HP35	Urea	100mM sodium chloride and 20mM sodium acetate	4.8	25
PSBD	Guanidine hydrochloride	2mM sodium phosphate, 2mM sodium borate and 50mM sodium chloride	8.0	25

Table 2-4. pH for experimental protein stability and the protonation state used in MD simulations

Protein	pH	Asp and Glu	His	C-terminus and N-terminus
EH	5.7			
GA	7.0		52, doubly protonated	
HP35	4.8		68, doubly protonated	
NTL9	5.5			
PSBD	8.0			Deprotonated N-terminus
Trp-cage	7.0			
UBA	6.5			
Ubiquitin	2.5	All Asp and Glu are protonated	68, doubly protonated	Protonated C-terminus

Table 2-5. Calculated values of $\Delta\Delta E_{vdw\_gb}$ for 160 C-capping sites from 120 non-redundant proteins taken from the pdb bank. Positive $\Delta\Delta E_{vdw\_gb}$ values indicate a stabilizing effect.

pdb code	chain ID	Short description of protein	Organism	Site No.	Calculated $\Delta\Delta E_{vdw\_gb}$ (kcal/mol)
1ABA	A	T4 glutaredoxin	Enterobacteria phage T4 sensu lato	56	0.61
1C44	A	Sterol carrier protein 2	Oryctolagus cuniculus	32	0.15
				86	1.20
				97	0.95
1KAF	A	The DNA Binding Domain Of Phage T4 Transcription Factor MotA	Enterobacteria phage T4 sensu lato	125	0.59
				179	0.36
1KP6	A	Killer toxin kp6 alpha-subunit	Ustilago maydis	9	-0.35
1L8R	A	Dachshund protein	Homo sapiens	255	0.76
1L9L	A	Granulysin from cytolytic T lymphocytes	Homo sapiens	63	0.64
1LWB	A	Phospholipase A2 protein	Streptomyces violaceoruber	75	0.35
1MC2	A	Phospholipase A2 protein	Deinagkistrodon acutus	14	0.34
1MK0	A	The catalytic domain of intron endonuclease I-TevI	Enterobacteria phage T4 sensu lato	38	0.27
1MOL	A	Monellin	Dioscoreophyllum cumminsii	27	0.87
1NWZ	A	Light receptor photoactive yellow protein	Halorhodospira halophila	51	0.31
				86	0.54
1OOH	A	An odorant binding protein LUSH	Drosophila melanogaster	34	1.40
				56	1.08
1ORG	A	A pheromone-binding protein	Rhyparobia maderae	53	1.03
1OSD	A	A mercury-binding protein	Cupriavidus metallidurans	65	0.24
1PBJ	A	A hypothetical protein	Methanothermobacter thermautotrophicus	59	0.40
1Q6V	A	Phospholipase A2 protein	Daboia russelii	14	0.24
1R6J	A	The PDZ2 domain of syntenin	Homo sapiens	231	0.88
1SBX	A	The dachshund-homology domain of Nuclear protooncoprotein SKI	Homo sapiens	165	0.66
1T1J	B	A hypothetical protein	Pseudomonas aeruginosa	43	0.52
				111	0.51
1T8K	A	An apo acyl carrier protein	Escherichia coli	16	0.53
				33	0.37
1TP6	A	A hypothetical protein	Pseudomonas aeruginosa	22	1.21
1TQG	A	CheA phosphotransferase domain	Thermotoga maritima	55	0.20
1U8T	B	CheY protein	Escherichia coli	29	0.37
				102	1.00
1VCD	A	Nudix protein Ndx1	Thermus thermophilus	52	0.33

1VYI	A	The C-terminal domain of a polymerase cofactor	Rabies virus	254	0.70
1WHZ	A	A hypothetical protein	Thermus thermophilus	18	0.60
1WOL	A	An HEPN homologue	Sulfolobus tokodaii	25	0.37
				50	0.67
1WY4	A	A villin headpiece	Gallus gallus	51	-0.31
1XLQ	A	Putidaredoxin	Pseudomonas putida	31	0.44
1XMK	A	The Zβ domain from the RNA editing enzyme ADAR1	Homo sapiens	341	0.78
1YN3	A	An extracellular adherence protein	Staphylococcus aureus	203	0.12
1Z96	A	Mud1 UBA domain	Schizosaccharomyces pombe	307	0.07
1ZMA	A	A bacterocin transport accessory protein	Streptococcus pneumoniae	81	0.38
2ACY	A	An acyl-phosphatase	Bos taurus	34	0.72
2B1L	B	A thiol:disulfide oxidoreductase	Escherichia coli	97	0.37
2B8I	A	A putative bacterial secretion factor	Vibrio vulnificus	56	0.99
2BO1	A	Ribosomal protein L30E	Thermococcus celer	30	0.87
				57	0.65
				75	0.47
2BWF	A	The UBL domain of Dsk2	Saccharomyces cerevisiae	36	-0.05
2CWY	A	A hypothetical protein	Thermus thermophilus	15	0.59
				55	0.48
2CX7	B	Sterol carrier protein 2	Thermus thermophilus	89	0.94
				100	0.65
2D48	A	Interleukin 4	Homo sapiens	95	0.43
2D58	A	An ionized calcium-binding adaptor	Homo sapiens	78	0.55
2FB6	A	A hypothetical protein	Bacteroides thetaiotaomicron	34	0.42
				68	0.59
				82	0.38
				91	0.43
2FC3	A	Ribosomal protein L7Ae	Aeropyrum pernix	46	0.66
				91	0.79
2FE5	A	The second PDZ domain of DLG3	Homo sapiens	270	0.56
2FYG	A	Nsp10	Severe acute respiratory syndrome-related coronavirus	34	0.13
2HC8	A	The actuator domain from Cu^{+}-ATPase	Archaeoglobus fulgidus	277	-0.01
2HL7	A	The periplasmic domain of cytochromes C maturation protein H	Pseudomonas aeruginosa	55	0.72
2HU9	A	A Zn2+ and [2Fe-2S]-containing copper chaperone	Archaeoglobus fulgidus	102	0.78
2I6V	A	Epsc, a crucial component of the type 2 secretion system	Vibrio cholerae	254	0.55
2IAY	A	LP2179, a member of the PF08866 family	Lactobacillus plantarum	31	0.87

2ICT	A	Antitoxin HigA	Escherichia coli	44	0.90
2J5Y	A	An albumin-binding domain	Finegoldia magna	22	0.73
2NT4	A	A response regulator homolog	Myxococcus xanthus	26	0.24
2O0Q	A	A hypothetical protein	Caulobacter vibrioides	20	0.40
				32	0.44
2OGB	A	The C-terminal domain of neuregulin receptor degrading protein 1	Mus musculus	237	0.51
2OY3	A	A macrophage receptor	Mus musculus	463	0.15
2P1H	A	The caspase recruitment domains of apoptotic protease activating factor 1	Homo sapiens	35	0.58
				81	0.22
2P3H	A	The CorC_HlyC domain of a putative hemolysin	Corynebacterium glutamicum	31	0.40
2POS	A	Sylvaticin	Pythium sylvaticum	32	0.30
2PSP	A	A pancreatic spasmolytic polypeptide	Sus scrofa	33	0.39
2PVB	A	Parvalbumin	Esox lucius	34	0.96
2PYQ	C	An uncharacterized protein	uncharacterized protein	20	0.40
				68	0.60
2QJL	A	A ubiquitin-related modifier	Saccharomyces cerevisiae	17	0.30
2RH3	A	The C-terminal domain of VirC2	Agrobacterium tumefaciens	130	0.37
2VB1	A	Triclinic hen egg-white lysozyme	Gallus gallus	16	0.60
				102	0.17
2VSV	A	The PDZ domain of human rhophilin-2	Homo sapiens	55	0.55
2VWR	A	The second PDZ domain of the human numb-binding protein 2	Homo sapiens	379	0.60
2W50	A	The N-terminal domain of human conserved dopamine neurotrophic factor	Homo sapiens	29	0.60
				60	0.71
2WFB	A	The apo Form of the Orange Protein	Desulfovibrio gigas	67	0.54
				88	0.64
2WT8	A	The N-terminal Brct domain of human microcephalin	Homo sapiens	36	0.62
				67	0.59
				83	0.47
2XEV	B	The TPR domain of YbgF	Xanthomonas campestris	15	0.50
				89	0.46
2ZQE	A	The endonuclease domain of an anti-recombination enzyme	Thermus thermophiles	31	0.63
3A0S	A	The PAS domain of histidine kinase ThkA	Thermotoga maritima	448	-0.35
3A0U	A	Response regulator protein TrrA	Thermotoga maritima	25	0.23
3A4R	A	The small ubiquitin-like modifier domain in Nip45	Mus musculus	376	0.66
3B79	A		Vibrio parahaemolyticus	17	0.25

		The N-terminal peptidase C39 like domain of the toxin secretion ATP-binding protein		49	0.73
3BS7	A	The sterile alpha motif domain of hyphen/aveugle	Drosophila melanogaster	71	0.50
3C9P	A	An uncharacterized protein	Streptococcus pneumoniae	25	0.40
				40	0.47
				106	0.41
3CJK	A	Copper transport protein ATOX1	Homo sapiens	59	0.20
3D2Q	B	The tandem zinc finger 3 and 4 domain of muscleblind-like protein 1	Homo sapiens	19	0.57
3E0Z	B	A putative imidazole glycerol phosphate synthase homolog	Agathobacter rectalis	39	0.10
3E11	B	A predicted zincin-like metalloprotease	Acidothermus cellulolyticus	102	0.51
3EZI	B	Histidine kinase NarX sensor domain	Escherichia coli	94	0.75
3FBL	A	An uncharacterized protein	Acidianus filamentous virus 1	66	0.60
3FZ4	A	A possible arsenate reductase	Streptococcus mutans	50	0.26
				68	1.07
3ID4	A	RseP PDZ2 domain	Escherichia coli	239	0.58
3IPJ	A	A domain of the PTS system	Peptoclostridium difficile	83	0.34
3L2A	A	VP35 interferon inhibitory domain	Reston ebolavirus	259	0.31
3LJW	B	The second bromodomain of human polybromo	Homo sapiens	240	1.67
3LLB	A	An uncharacterized protein	Pseudomonas aeruginosa	27	-0.13
				47	-0.06
3M3G	A	An elicitor of plant defense responses	Trichoderma virens	115	0.70
3NIR	A	Crambin	Crambe hispanica	20	0.17
				31	0.27
3NUF	A	A PRD-containing transcription regulator	Lactobacillus paracasei	67	0.59
3O79	B	A Prion protein	Oryctolagus cuniculus	195	0.23
3ODV	A	Kaliotoxin	Androctonus mauritanicus	22	1.28
3PO0	A	A Ubiquitin-like small archaeal modifier proteins	Haloferax volcanii	14	0.27
3QMX	A	Glutaredoxin A	Synechocystis sp. PCC 6803	29	-0.32
3S0A	A	An odorant-binding protein	Apis mellifera	22	0.47
				34	0.90
3SNS	A	The C-terminal domain of lipoprotein BamC	Escherichia coli	263	1.12
				292	0.82
3SVI	A	The Pto-binding domain of HopPmaL	Pseudomonas syringae group genomosp. 3	157	0.48
				173	0.44
3SZS	B	Hellethionin D	Helleborus purpurascens	20	0.33

3T7Z	A	Nop N-terminal domain	Methanocaldococcus jannaschii	60	-0.13
				91	0.56
3UI6	A	Parvulin 14	Homo sapiens	61	0.29
3V1A	A	A Metal interface design	synthetic construct	22	0.61
3W1O	A	A hypothetical protein	Neisseria meningitidis	51	0.86
3WCQ	A	Ferredoxin	Cyanidioschyzon merolae	33	0.31
				73	0.55
3ZR8	X	Rxlr effector AVR3a11	Phytophthora capsici	100	0.65
4CVD	A	A cell wall binding module	Streptococcus phage Cp-1	263	1.08
				279	0.33
4D40	A	Type IV pilin	Shewanella oneidensis	28	0.17
4F55	A	The catalytic Domain of SleB rotein	Bacillus cereus	202	0.56
				222	0.23
4FQN	A	CCM2 C-terminal harmonin homology domain	Homo sapiens	328	0.50
4G9S	A	A goose-type lysozyme	Escherichia coli	60	1.04
4GOQ	A	A hypothetical protein	Caulobacter vibrioides	20	0.41
4HRO	A	Small archaeal modifier protein 1	Haloferax volcanii	14	0.22
4HS5	A	Protein CyaY	Psychromonas ingrahamii	25	0.60
4JIU	A	An uncharacterized protein	Pyrococcus abyssi	14	0.50
4N6X	A	Na(+)/H(+) exchange regulatory cofactor NHE-RF1/Chemokine receptor CXCR2 fusion protein	Homo sapiens	52	0.89
4PXV	A	The LysM domain of chitinase A	Pteris ryukyuensis	32	0.62
4XPX	A	Hemerythrin	Methylococcus capsulatus	69	0.37
				97	0.54

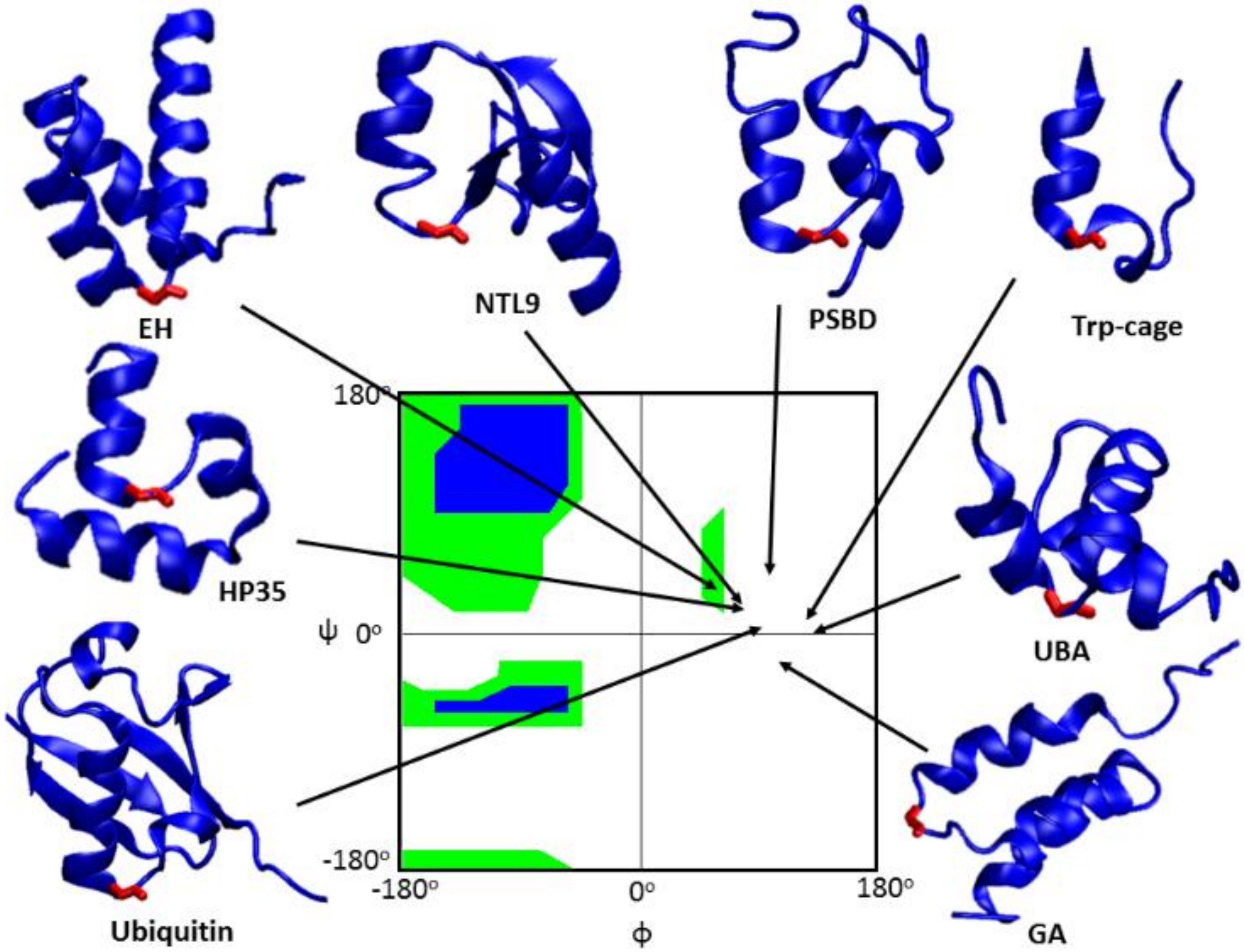


Figure 2-1. Ribbon representation of the proteins studied with the C-capping Gly colored red. φ/ψ angles of the C-capping glycines are indicated by arrows. The Ramachandran plot is colored green for broadly allowed and blue for most favored regions for L-amino acids, which is adopted from Ramaplot in VMD(34). The Ramachandran plot for a D-amino acid is the mirror image about the central point ($\varphi = 0°$ and $\psi = 0°$) of the plot shown above.

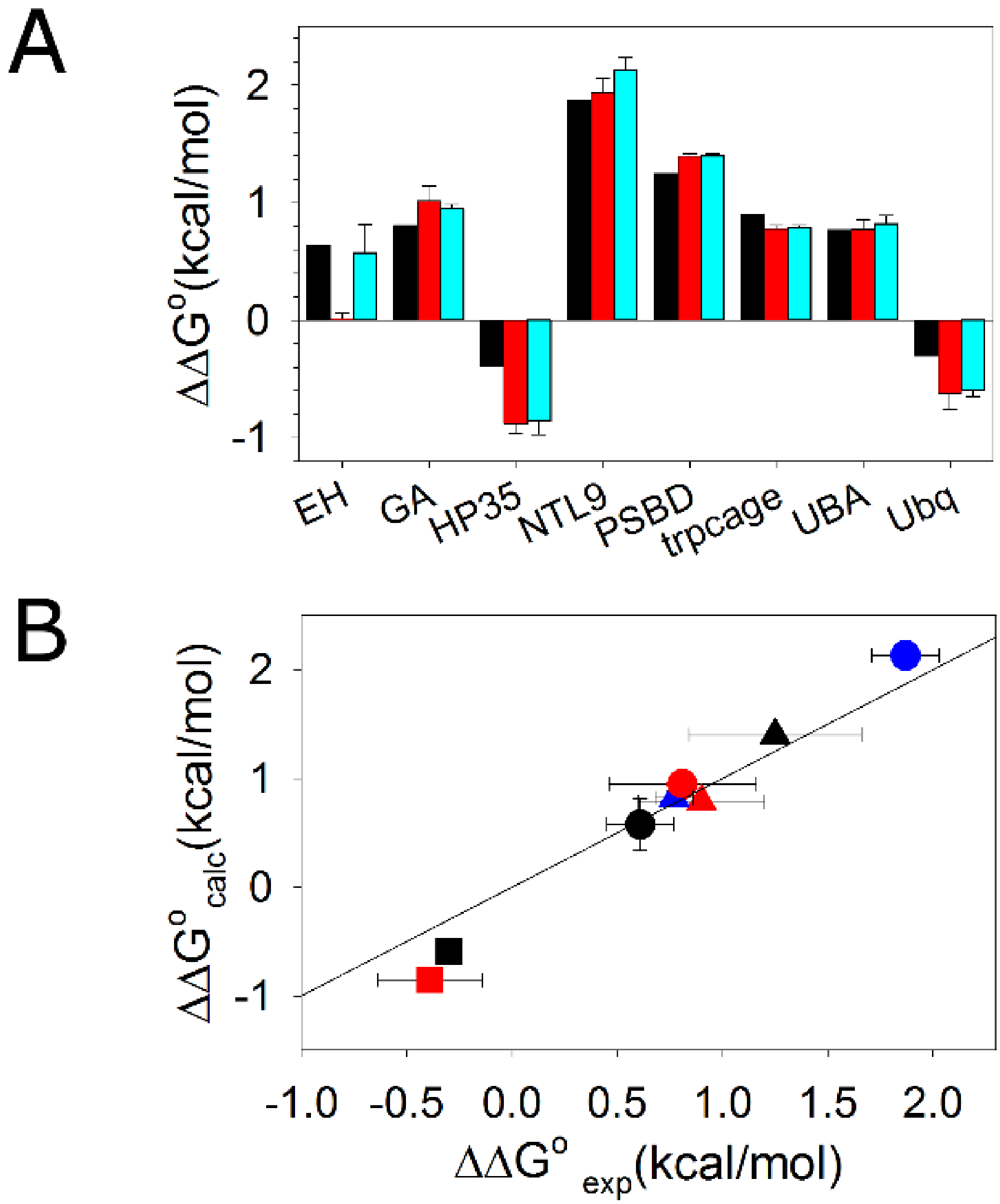


Figure 2-2. Thermodynamic integration reproduces experimental values of ΔΔG˚. (A) Experimental ΔΔG˚ values are shown in black. Calculated ΔΔG˚ values using experimental structures as starting structures are shown in red. Calculated ΔΔG˚ values using the last frames of 50 ns simulations as the starting structures are in cyan. (B) A scatter plot of experimental ΔΔG˚ and calculated ΔΔG˚ values using the last frames of a 50 ns simulation as the starting structure. Solid line represents $\Delta\Delta G^{\circ}_{exp} = \Delta\Delta G^{\circ}_{cal}$. EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■. The calculated value for the EH domain used in the plot was derived by using the unrestrained MD structure as the starting structure for the TI calculation. Positive ΔΔG˚ values indicate stabilization.

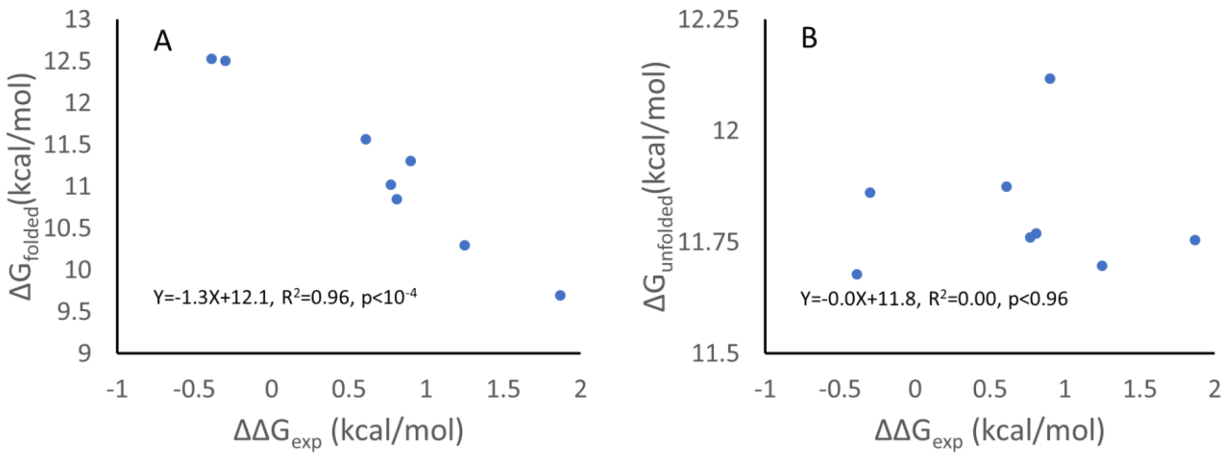


Figure 2-3. Correlation between $\Delta\Delta G_{exp}$ with calculated free energy changes in the folded state (ΔG_{folded}) and unfolded state ($\Delta G_{unfolded}$) using TI.

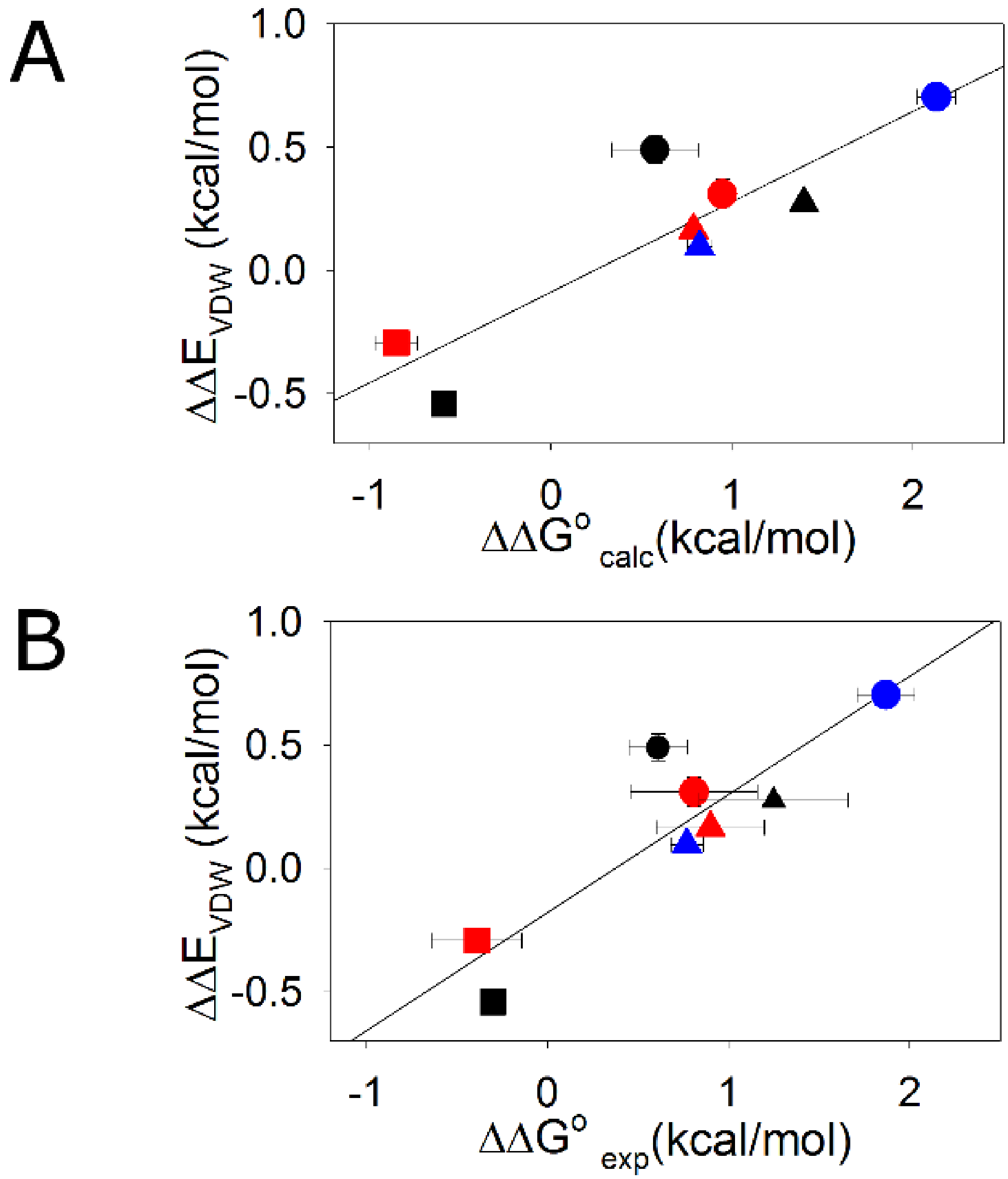


Figure 2-4. Scatter plot of $\Delta\Delta E_{vdw}$ and $\Delta\Delta G^{\circ}$ with solid line showing the linear fit. (A) Correlation of $\Delta\Delta E_{vdw}$ and $\Delta\Delta G^{\circ}$ values calculated by thermodynamic integration. r=0.89, p-value=0.0033 (B) Correlation of $\Delta\Delta E_{vdw}$ and experimental $\Delta\Delta G^{\circ}$ values. r=0.89, p-value =0.0033. EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■. Positive $\Delta\Delta G^{\circ}$ values indicate stabilization.

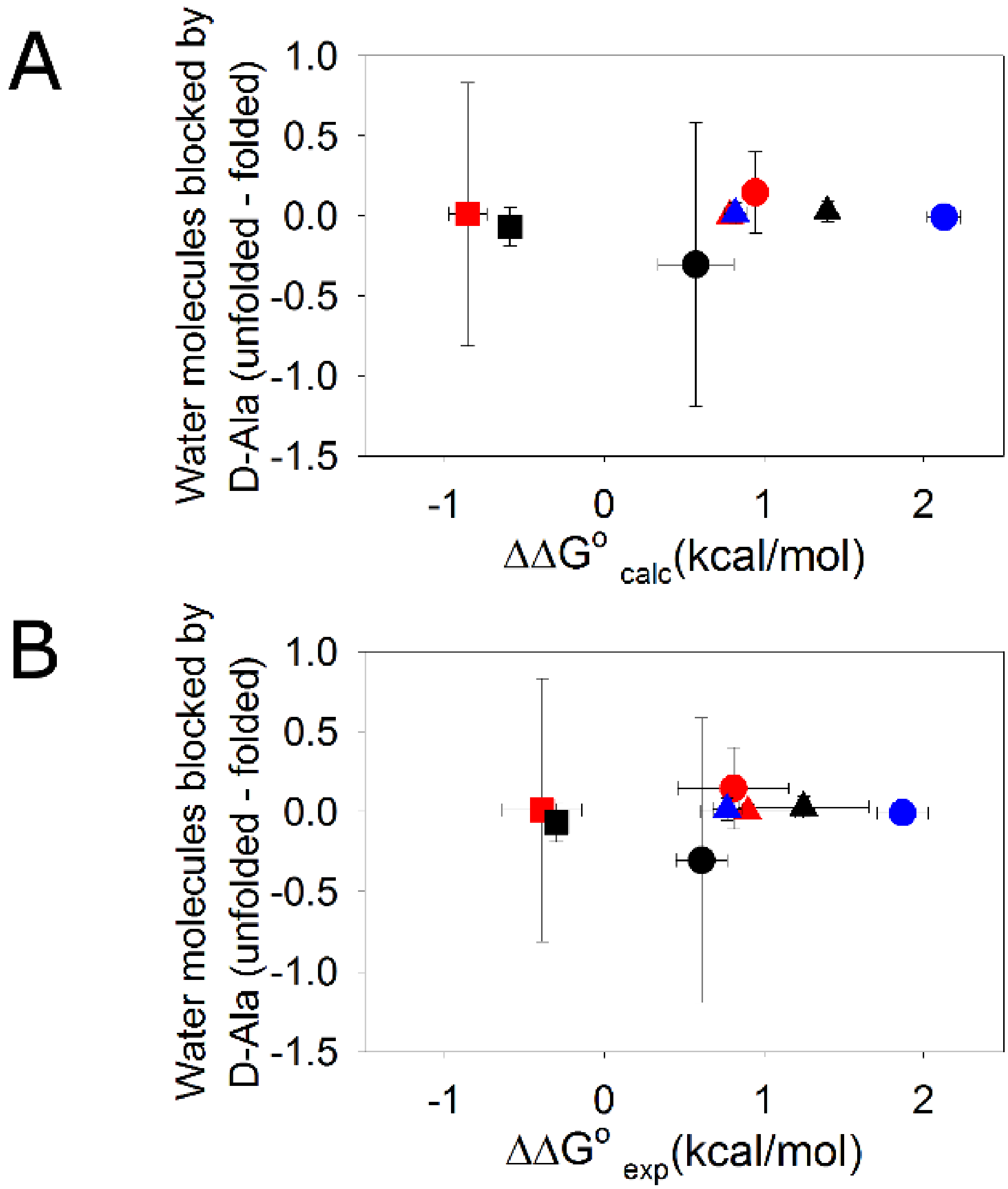


Figure 2-5. Changes in backbone solvation do not correlate with ΔΔG˚. The difference in the number of water molecules blocked by D-Ala relative to Gly (Unfolded-folded) is plotted vs (A) calculated ΔΔG˚ values (r=0.16). (B) experimental ΔΔG˚ values (r=0.17). EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■. Positive ΔΔG˚ values indicate stabilization.

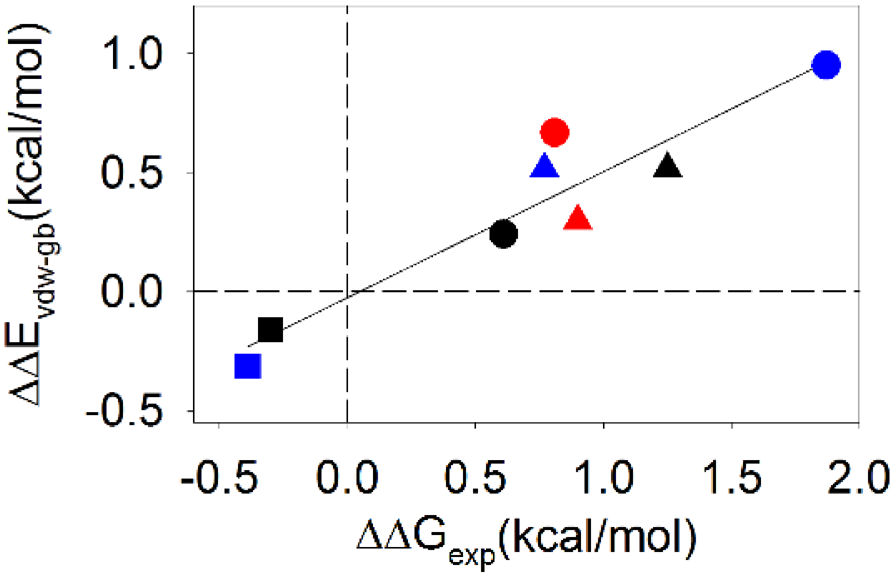


Figure 2-6. There is a strong correlation between $\Delta\Delta E_{vdw\text{-}gb}$ and $\Delta\Delta G^{\circ}_{exp}$. Positive values of $\Delta\Delta G^{\circ}$ indicate stabilizing effects. EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■; r=0.94 and p=0.0004. Positive $\Delta\Delta G^{\circ}$ values indicate stabilization.

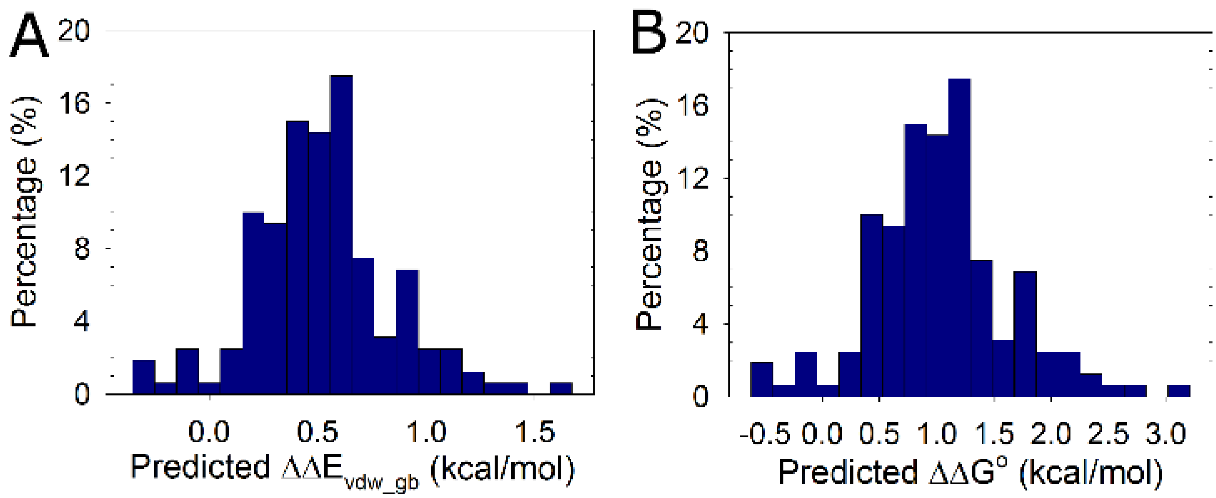


Figure 2-7. Proteins are stabilized by D-Ala substitutions. The distribution of $\Delta\Delta E_{vdw\_gb}$ and $\Delta\Delta G^{\circ}$ values for the 160 C-capping sites in the 120 non-redundant proteins is shown as a histogram. (A) Distribution of $\Delta\Delta E_{vdw\_gb}$ values. (B) Distribution of predicted $\Delta\Delta G^{\circ}$ values. Positive $\Delta\Delta G^{\circ}$ values represent a stabilizing effect.

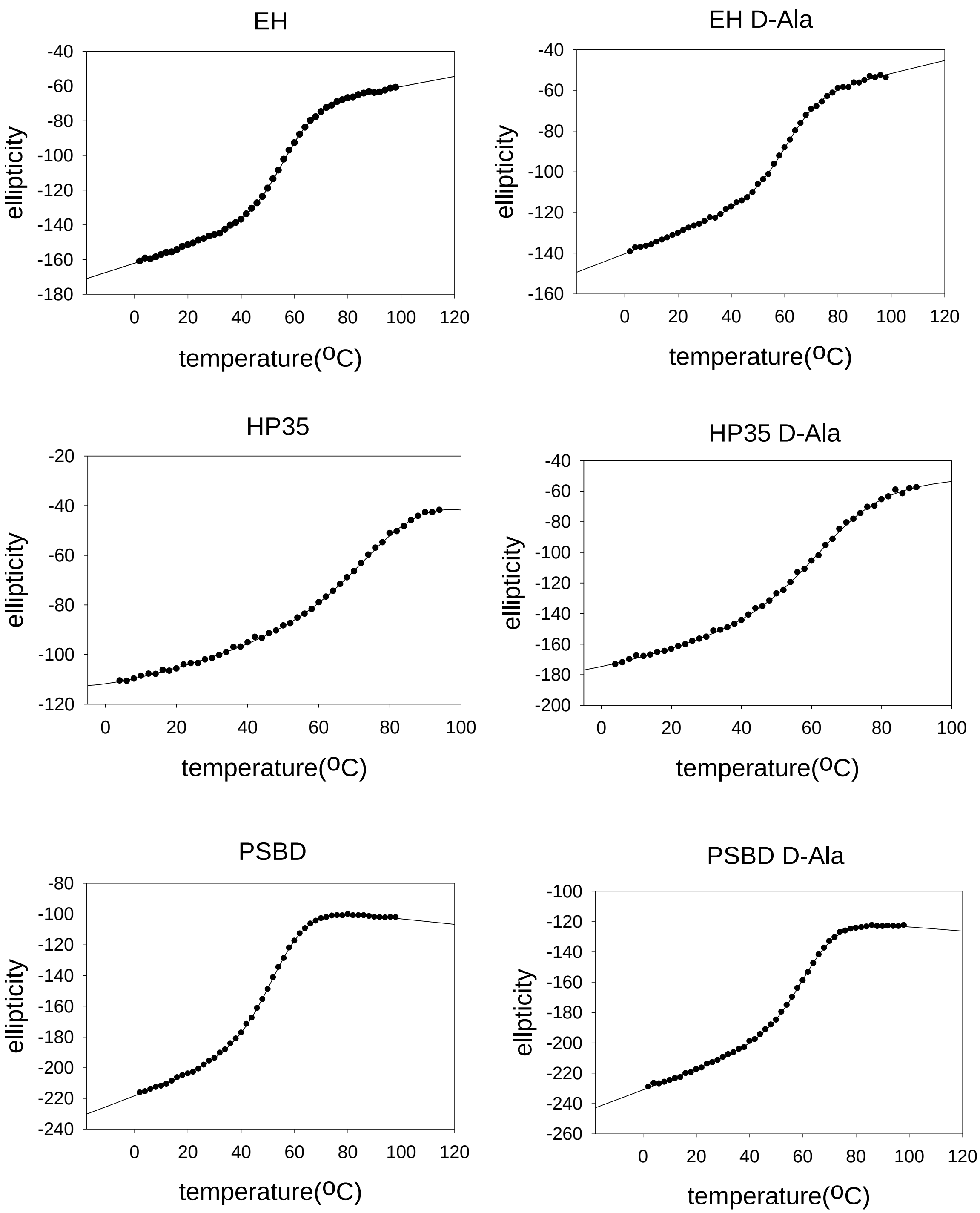
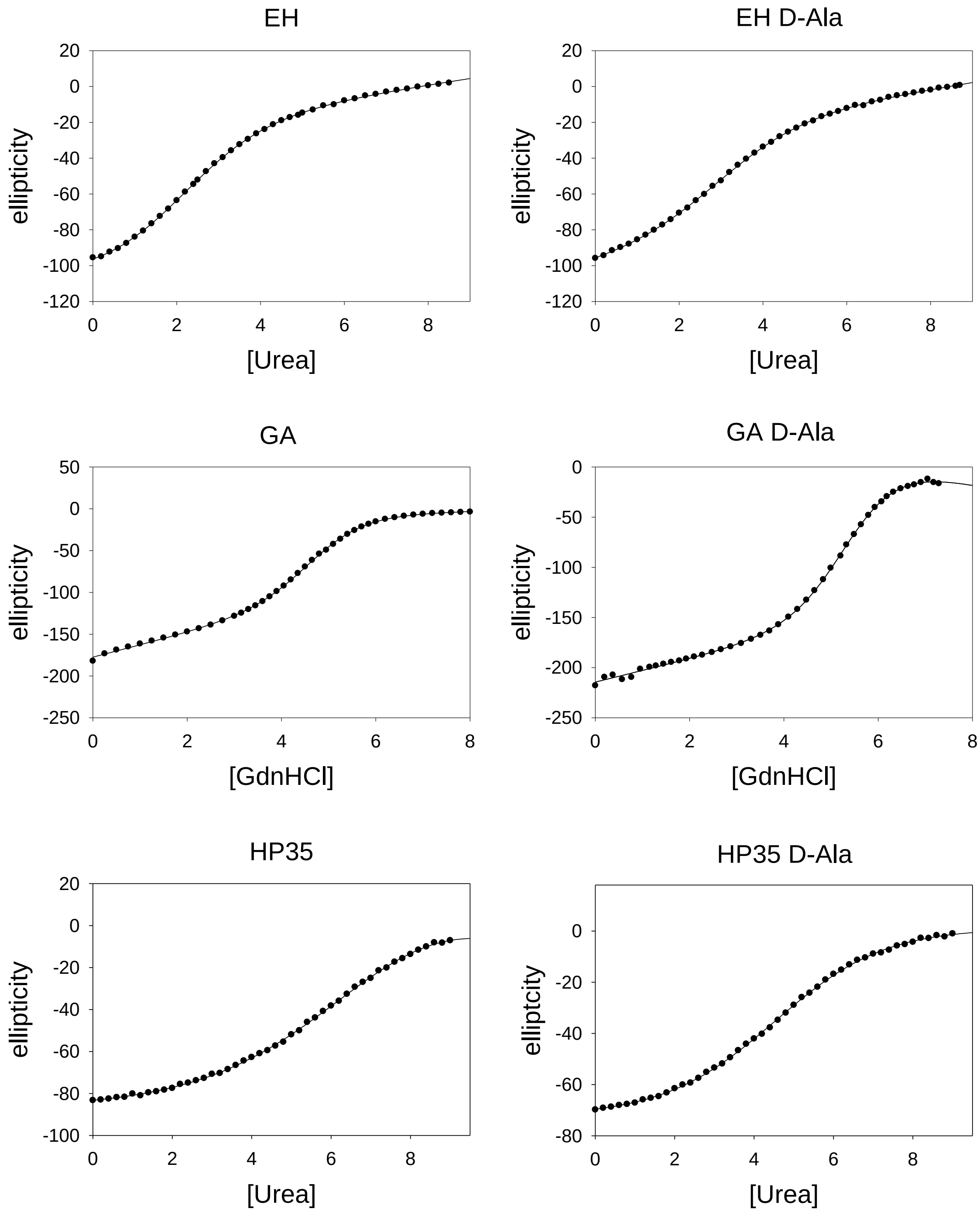


Figure 2-8. Thermal denaturation of EH, HP35, PSBD and their D-Ala variants. The solid line are the fitted curves.


EH
EH D-Ala
GA
GA D-Ala
HP35
HP35 D-Ala
ellipticity
elliptcity
[Urea]
[GdnHCl]

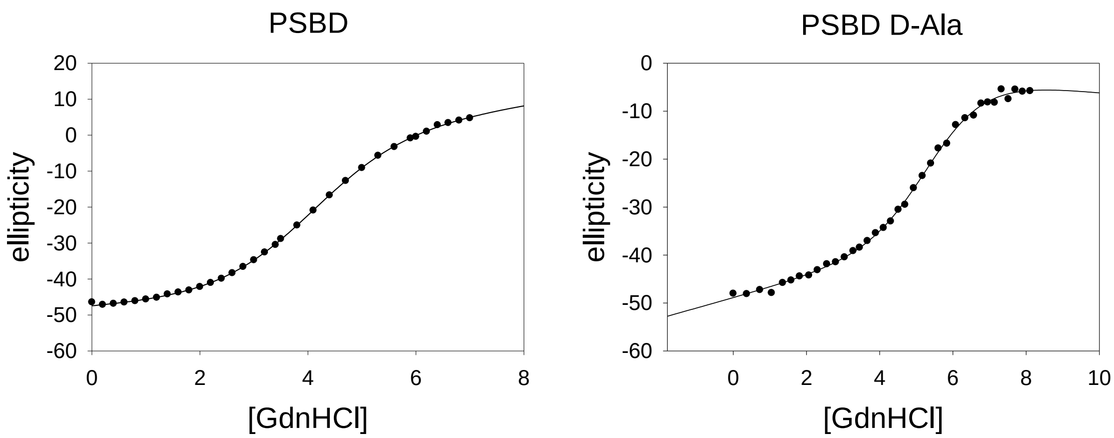


Figure 2-9. Urea/Guanidine hydrochloride denaturation of EH, GA, HP35, PSBD and their D-Ala variants. The solid lines are the fitted curves.

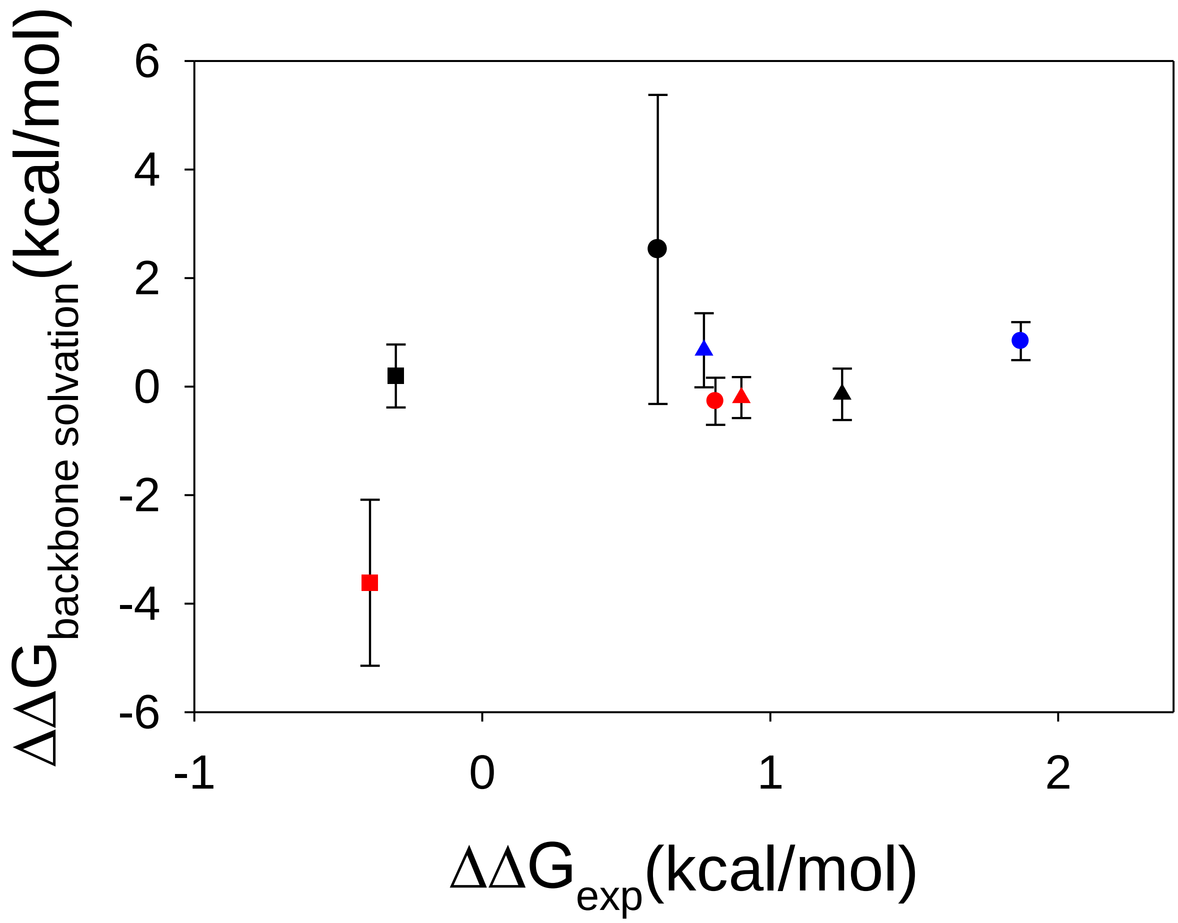


Figure 2-10. Correlation between $\Delta\Delta G_{\text{backbone solvation}}$ and $\Delta\Delta G_{\text{exp}}$. r=0.52, p=0.19. If only proteins with good convergence are included (GA, NTL9, PSBD, Trp-cage, UBA and ubiquitin), r=0.28, p-value=0.58, slope=0.20. EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■;

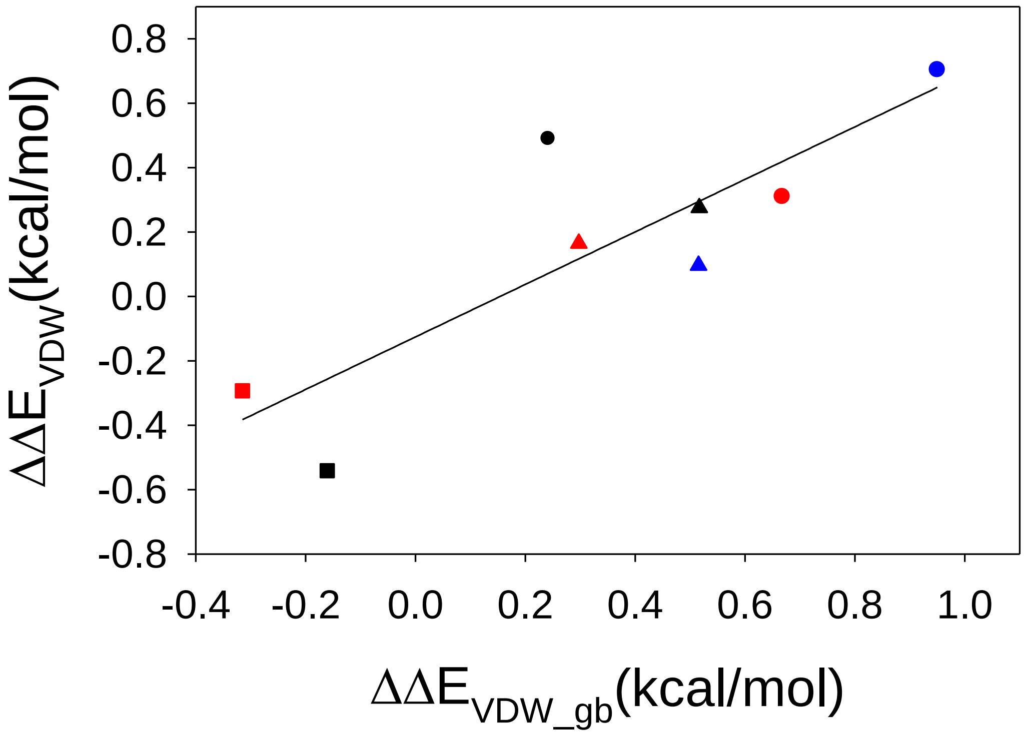


Figure 2-11. Correlation between $\Delta\Delta E_{vdw}$ and $\Delta\Delta E_{vdw\_gb}$. r=0.84, p=0.0079. EH ●; GA ●; HP35 ■; NTL9 ●; PSBD ▲; Trp-cage ▲; UBA ▲; Ubiquitin ■;

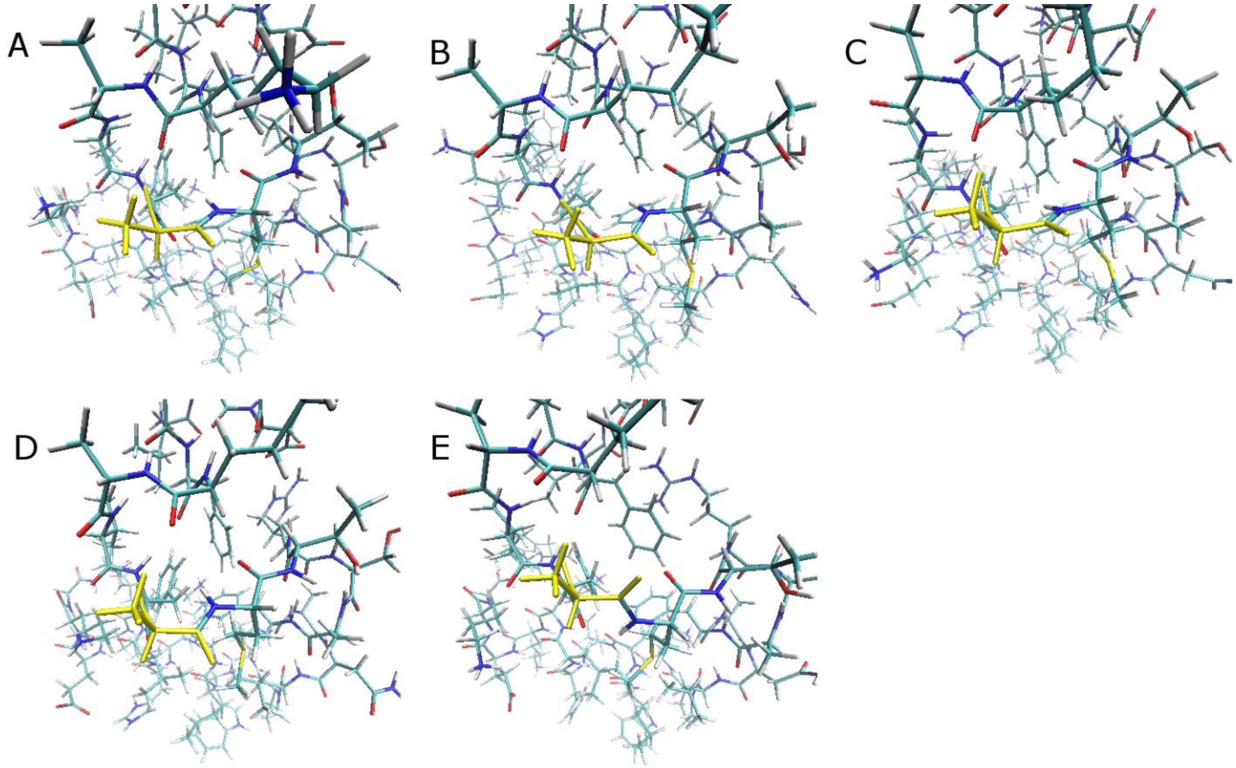


Figure 2-12. Structure of the HP35 G11D-Ala mutant taken from an MD simulation. 5 snapshots at 40 ns (A), 80 ns (B), 120 ns (C), 160 ns (D) and 200 ns (E) are shown with hydrogen included. The D-Ala residues are colored yellow.

2.5 References

1. Chi EY, Krishnan S, Randolph TW, & Carpenter JF (2003) Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharm. Res.* 20(9):1325-1336.
2. Daniel RM, Cowan DA, Morgan HW, & Curran MP (1982) A correlation between protein thermostability and resistance to proteolysis. *Biochem. J.* 207(3):641-644.
3. Parsell DA & Sauer RT (1989) The structural stability of a protein is an important determinant of its proteolytic susceptibility in escherichia-coli. *J. Biol. Chem.* 264(13):7590-7595.
4. Chi EY*, et al.* (2003) Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci.* 12(5):903-913.
5. McLendon G & Radany E (1978) Is protein turnover thermodynamically controlled? *J. Biol. Chem.* 253(18):6335-6337.
6. Binz HK, Amstutz P, & Pluckthun A (2005) Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.* 23(10):1257-1268.
7. Skrlec K, Strukelj B, & Berlec A (2015) Non-immunoglobulin scaffolds: a focus on their targets. *Trends Biotechnol.* 33(7):408-418.
8. Matthews BW, Nicholson H, & Becktel WJ (1987) Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. U. S. A.* 84(19):6663-6667.
9. Nicholson H, Becktel WJ, & Matthews BW (1988) Enhanced protein thermostability from designed mutations that interact with alpha-helix dipoles. *Nature* 336(6200):651-656.

10. Matsumura M, Becktel WJ, Levitt M, & Matthews BW (1989) Stabilization of phage T4 lysozyme by engineered disulfide bonds. *Proc. Natl. Acad. Sci. U. S. A.* 86(17):6562-6566.

11. Spector S, *et al.* (2000) Rational modification of protein stability by the mutation of charged surface residues. *Biochemistry* 39(5):872-879.

12. Anil B, Song B, Tang Y, & Raleigh DP (2004) Exploiting the right side of the Ramachandran plot: substitution of glycines by D-alanine can significantly increase protein stability. *J. Am. Chem. Soc.* 126(41):13194-13195.

13. Sauer RT, *et al.* (1986) An engineered intersubunit disulfide enhances the stability and DNA binding of the N-terminal domain of lambda repressor. *Biochemistry* 25(20):5992-5998.

14. Wells JA & Powers DB (1986) In vivo formation and stability of engineered disulfide bonds in subtilisin. *J. Biol. Chem.* 261(14):6564-6570.

15. Wetzel R, Perry LJ, Baase WA, & Becktel WJ (1988) Disulfide bonds and thermal stability in T4 lysozyme. *Proc. Natl. Acad. Sci. U. S. A.* 85(2):401-405.

16. Tidor B & Karplus M (1993) The contribution of cross-links to protein stability: a normal mode analysis of the configurational entropy of the native state. *Proteins* 15(1):71-79.

17. Rodriguez-Granillo A, Annavarapu S, Zhang L, Koder RL, & Nanda V (2011) Computational design of thermostabilizing D-amino acid substitutions. *J. Am. Chem. Soc.* 133(46):18750-18759.

18. Betz SF & Pielak GJ (1992) Introduction of a disulfide bond into cytochrome-c stabilizes a compact denatured state. *Biochemistry* 31(49):12337-12344.

19. Camarero JA, *et al.* (2001) Rescuing a destabilized protein fold through backbone cyclization. *J. Mol. Biol.* 308(5):1045-1062.

20. Stites WE, Meeker AK, & Shortle D (1994) Evidence for Strained Interactions between Side-Chains and the Polypeptide Backbone. *J. Mol. Biol.* 235(1):27-32.

21. Richardson JS & Richardson DC (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science* 240(4859):1648-1652.

22. Aurora R & Rose GD (1998) Helix capping. *Protein Sci.* 7(1):21-38.

23. Gunasekaran K, Nagarajaram HA, Ramakrishnan C, & Balaram P (1998) Stereochemical punctuation marks in protein structures: glycine and proline containing helix stop signals. *J. Mol. Biol.* 275(5):917-932.

24. Hutchinson EG & Thornton JM (1994) A revised set of potentials for beta-turn formation in proteins. *Protein Sci.* 3(12):2207-2216.

25. Haque TS & Gellman SH (1997) Insights on beta-hairpin stability in aqueous solution from peptides with enforced type I' and type II' beta-turns. *J. Am. Chem. Soc.* 119(9):2303-2304.

26. Sibanda BL & Thornton JM (1985) Beta-hairpin families in globular proteins. *Nature* 316(6024):170-174.

27. Bystroff C & Baker D (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* 281(3):565-577.

28. Bang D, *et al.* (2006) Dissecting the energetics of protein alpha-helix C-cap termination through chemical protein synthesis. *Nat. Chem. Biol.* 2(3):139-143.

29. Chiu TK, *et al.* (2005) High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proc. Natl. Acad. Sci. U. S. A.* 102(21):7517-7522.

30. Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, & Pabo CO (1994) Structural studies of the engrailed homeodomain. *Protein Sci.* 3(10):1779-1787.

31. Johansson MU, *et al.* (1997) Solution structure of the albumin-binding GA module: a versatile bacterial protein domain. *J. Mol. Biol.* 266(5):859-865.

32. Kalia YN, *et al.* (1993) The high-resolution structure of the peripheral subunit-binding domain of dihydrolipoamide acetyltransferase from the pyruvate dehydrogenase multienzyme complex of Bacillus stearothermophilus. *J. Mol. Biol.* 230(1):323-341.

33. Carpino LA & Han GY (1970) 9-fluorenylmethoxycarbonyl function, a new base-sensitive amino-protecting group. *J. Am. Chem. Soc.* 92(19):5748-&.

34. Humphrey W, Dalke A, & Schulten K (1996) VMD: visual molecular dynamics. *J. Mol. Graph.* 14(1):33-38, 27-38.

35. D.A. Case JTB, R.M. Betz, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K.M. Merz, G. Monard, P. Needham, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, R. Salomon-Ferrer, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, D.M. York and P.A. Kollman (2015) *AMBER 2015* (University of California, San Francisco).

36. Cho JH, *et al.* (2014) Energetically significant networks of coupled interactions within an unfolded protein. *Proc. Natl. Acad. Sci. U. S. A.* 111(33):12079-12084.

37. Neidigh JW, Fesinmeyer RM, & Andersen NH (2002) Designing a 20-residue protein. *Nat. Struct. Biol.* 9(6):425-430.

38. Withers-Ward ES, Mueller TD, Chen ISY, & Feigon J (2000) Biochemical and structural analysis of the interaction between the UBA(2) domain of the DNA repair protein HHR23A and HIV-1 Vpr. *Biochemistry* 39(46):14103-14112.

39. Vijaykumar S, Bugg CE, & Cook WJ (1987) Structure of ubiquitin refined at 1.8 a resolution. *J. Mol. Biol.* 194(3):531-544.

40. Guex N & Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18(15):2714-2723.

41. Maier JA, *et al.* (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput.* 11(8):3696-3713.

42. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, & Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* 79(2):926-935.

43. Khoury GA, *et al.* (2014) Forcefield_NCAA: ab initio charge parameters to aid in the discovery and design of therapeutic proteins and peptides with unnatural amino acids and their application to complement inhibitors of the compstatin family. *ACS Synth. Biol.* 3(12):855-869.

44. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, & Haak JR (1984) Molecular-dynamics with coupling to an external bath. *The Journal of chemical physics* 81(8):3684-3690.

45. Darden T, York D, & Pedersen L (1993) Particle mesh ewald - an N.Log(N) method for ewald sums in large systems. *J. Chem. Phys.* 98(12):10089-10092.

46. Ryckaert JP, Ciccotti G, & Berendsen HJC (1977) Numerical-integration of cartesian equations of motion of a system with constraints - molecular-dynamics of N-alkanes. *J Comput Phys* 23(3):327-341.

47. Gordon JC, *et al.* (2005) H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* 33(Web Server issue):W368-371.

48. Kirkwood JG (1935) Statistical mechanics of fluid mixtures. *The Journal of chemical physics* 3(5):300-313.

49. Roe DR & Cheatham TE, 3rd (2013) PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* 9(7):3084-3095.

50. Li L, *et al.* (2012) DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophys.* 5:9.

51. Yamagishi J, Okimoto N, Morimoto G, & Taiji M (2014) A new set of atomic radii for accurate estimation of solvation free energy by Poisson-Boltzmann solvent model. *J. Comput. Chem.* 35(29):2132-2139.

52. Nguyen H, Maier J, Huang H, Perrone V, & Simmerling C (2014) Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J. Am. Chem. Soc.* 136(40):13959-13962.

53. Nguyen H, Roe DR, & Simmerling C (2013) Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* 9(4):2020-2034.

54. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* 215(3):403-410.

55. Madej T, *et al.* (2014) MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Res.* 42(Database issue):D297-303.

56. Religa TL, *et al.* (2007) The helix-turn-helix motif as an ultrafast independently folding domain: the pathway of folding of Engrailed homeodomain. *Proc. Natl. Acad. Sci. U. S. A.* 104(22):9272-9277.

57. Wang T, Zhu YJ, & Gai F (2004) Folding of a three-helix bundle at the folding speed limit. *J. Phys. Chem. B* 108(12):3694-3697.

58. Spector S, *et al.* (1998) Cooperative folding of a protein mini domain: the peripheral subunit-binding domain of the pyruvate dehydrogenase multienzyme complex. *J. Mol. Biol.* 276(2):479-489.

59. Chiu TK, *et al.* (2005) High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proceedings of the National Academy of Sciences of the United States of America* 102(21):7517-7522.

60. Scott KA, Alonso DO, Sato S, Fersht AR, & Daggett V (2007) Conformational entropy of alanine versus glycine in protein denatured states. *Proc. Natl. Acad. Sci. U. S. A.* 104(8):2661-2666.

61. Nemethy G, Leach SJ, & Scheraga HA (1966) Influence of amino acid side chains on free energy of helix-coil transitions. *J. Phys. Chem.* 70(4):998-&.

62. DAquino JA, *et al.* (1996) The magnitude of the backbone conformational entropy change in protein folding. *Proteins: Struct., Funct., Genet.* 25(2):143-156.

63. Zaman MH, Shen MY, Berry RS, Freed KF, & Sosnick TR (2003) Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the flory isolated-pair hypothesis for peptides. *J. Mol. Biol.* 331(3):693-711.

64. Baxa MC, Haddadian EJ, Jumper JM, Freed KF, & Sosnick TR (2014) Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations. *Proc. Natl. Acad. Sci. U. S. A.* 111(43):15396-15401.

65. Kollman PA, *et al.* (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* 33(12):889-897.

66. Meng W, Lyle N, Luan B, Raleigh DP, & Pappu RV (2013) Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* 110(6):2123-2128.

67. Mok KH, *et al.* (2007) A pre-existing hydrophobic collapse in the unfolded state of an ultrafast folding protein. *Nature* 447(7140):106-109.

68. Spector S, Rosconi M, & Raleigh DP (1999) Conformational analysis of peptide fragments derived from the peripheral subunit binding domain from the pyruvate dehydrogenase multienzyme complex of Bacillus stearothermophilus: Evidence for nonrandom structure in the unfolded state. *Biopolymers* 49(1):29-40.

69. Magliery TJ (2015) Protein stability: computation, sequence statistics, and new experimental methods. *Curr. Opin. Struct. Biol.* 33:161-168.

70. Steipe B, Schiller B, Pluckthun A, & Steinbacher S (1994) Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol.* 240(3):188-192.

3. Dissecting the energetics of intrinsically disordered proteins

Abstract

Intrinsically disordered proteins (IDPs) play important roles in biology, but little is known about the energetics of their inter-residue interactions. Methods that have been successfully applied to analyze the energetics of globular proteins are not applicable to the fluctuating partially ordered ensembles populated by IDPs. A general strategy is introduced for analyzing the energetic role of individual residues in IDPs. The approach combines experimental measurements of the binding of wild-type and mutant IDPs to their partners with alchemical free energy calculations of the structured complexes. The approach is validated by the analysis of the effects of mutations upon the binding free energy of the ovomucoid inhibitor third binding domain to its partners and is applied to the C-terminal domain of the measles virus nucleoprotein, a 125-residue IDP involved in the RNA transcription and replication of measles virus. The analysis reveals significant inter-residue interactions in the unbound IDP and suggests a biological role for them.

Acknowledgements

I gratefully acknowledge Koushik Kasavajhala, Chuan Tian and Kellon Belfon for their administration of computational resources. This chapter contains direct excerpts from the manuscript (Dissecting the energetics of intrinsically disordered proteins via a hybrid experimental and computational approach) with a few adjustments.

3.1 Introduction

Intrinsically disordered proteins (IDPs) lack stable secondary and tertiary structure due to their low content of bulky hydrophobic residues and their high content of polar and charged residues(1, 2). While they do not fold into well-defined globular structures in isolation(3, 4), IDPs populate ensembles ranging from expanded states with little residual structure to ensembles which are more compact and contain residual secondary and tertiary interactions and they often fold upon binding to their partners (5-10). IDPs play important roles in biology. Structural characterization of IDPs in their uncomplexed state and of the denatured state of globular proteins show that neither are true random-coils, instead, they often contain transient secondary structure and long-range interactions (11-20). Inter-residue interactions in the unbound state of IDPs could modulate their binding affinity and other interactions, and thus modulate their biological activity. Moreover, long-range interactions in IDPs and in unfolded proteins may play an important role in controlling the propensity to aggregate and thus preventing protein misfolding diseases (12, 13, 21, 22). Progress has been made in characterizing the structural properties of IDPs and in defining relationships between sequence and conformational properties, but much less is known about the energetics of inter-residue interactions in IDPs (2), and quantitative analysis of the inter-residue energetics of IDPs is still absent.

The contribution of a residue to the energetics of the folded state of a protein is traditionally obtained by mutagenesis and unfolding free energy measurements, but such a strategy cannot be applied to IDPs. In principle, it is also possible to estimate the energetics of a residue in a globular protein via molecular modelling if the structure of the protein is resolved in atomic level(23, 24). Structural ensembles of IDPs can be obtained from ensemble fitting using experimental observables such as Rg and NMR as constraints and the ensembles used as input to potential energy functions to estimate energies. However, the ensembles derived are not fully deterministic as the experimental observables are not sufficient to uniquely define highly conformationally heterogeneous IDPs and the energetics obtained are not always reliable. Molecular dynamics (MD) simulations are becoming popular for studying the structural propensities of IDPs as newly developed force fields are being trained not only for globular proteins, but also for IDPs (25-27). While many current force fields have shown good performance at folding proteins (28-30), their accuracy for IDPs still awaits testing and many force fields have difficulty reproducing global properties of protein unfolded states (31). Simulations of IDPs also require considerably more sampling than simulations of folded proteins because of their flat free energy landscape, increasing computational cost. Larger IDPs are currently not suitable for MD simulations in explicit solvent so implicit solvent models must be used, which can be less accurate. This inability to accurately model the disordered state of IDPs has hampered the application of MD to study IDP energetics.

Here, we describe a strategy to quantitatively analyze the energetic effects of mutations upon the free state of IDPs and to define the energetic contribution made by individual residues in the free state. The approach combines experimental structural and energetic information on complexes of IDPs with their binding partners and alchemical free energy calculations of the bound state to deduce the properties of the unbound state via a thermodynamic cycle. The strategy is validated using mutants of the turkey ovomucoid inhibitor third binding domain (OMTKY3) and its binding partners and is then applied to the C-terminal domain of the measles virus nucleoprotein (NTAIL), a 125-residue IDP involved in the RNA transcription and replication of measles virus. NTAIL undergoes binding and folding upon encounter with the X domain (XD) of its phosphoprotein binding partner (32). SAXS studies have shown that NTAIL has a radius of gyration, Rg, of 27.5 ± 0.7 Å which is significantly less than predicted Rg for a random coil with the same length of

NTAIL is 35 – 38 Å (33), suggesting inter-residue interactions lead to compaction in the free state. MD simulations of the αMoRE of NTAIL suggested that there are significant structures in the αMoRE of NTAIL (34). A recent study of the effect of the truncation of the disordered N-terminal region of NTAIL on the binding energetics support this hypothesis. A non-monotonic dependence of binding strength on truncation length was observed (35), indicating that residues in the disordered N-terminal region of NTAIL contribute to the binding affinity of NTAIL and XD. This work also suggested that the folding of NTAIL upon binding is impeded by the disordered regions flanking the αMoRE. However, the molecular basis of these effects is unclear and the mechanisms behind these unexplained observations cannot be determined using conventional methods. We quantify the effect of three mutations in the unbound state of NTAIL and show that the residues participate in long-range interactions in the unbound state and that these interactions can modulate the binding of NTAIL to its partner. Interactions in the free state are predicted to reduce the affinity of NTAIL for XD, thereby providing a mechanism for tuning binding affinity and biological function. Our approach allows quantitative analysis of these interactions without the need to fully model the IDP ensemble.

This work illustrates that, as computational results become more and more reliable due to advances in force fields and computing hardware, it is possible to develop accurate and precise hybrid methods which rely on experiments and calculations simultaneously to reveal insights that cannot be studied by conventional methods.

3.2 Methods

3.2.1 Free Energy Calculations for the Binding Between Ovomucoid Inhibitor Third Domain (OMTKY3) and Its Target Protease

The pdb code for the structures of monomeric OMTKY3, *Streptomyces griseus* proteinase B (SGPB)/OMTKY3 complexes and subtilisin Carlsberg (CARL)/OMTKY3 complex used in the free energy calculations are listed in **Table 3-2**. Hydrogen atoms were added using the MolProbity program(36). Side chain rotamer states for ASN/GLN were corrected based on suggestion provided by MolProbity. LYS, ARG side chains and N-termini were set to be protonated and ASP, GLU side chains and C-termini were set to be unprotonated. All HIS side chains were set to be neutral with protons on Nε except HIS57 in SGPB and HIS64/HIS226 in CARL which have protons on Nδ. Waters present in X-ray structures were kept while all salt ions were deleted. Truncated octahedron boxes were used to solvate the proteins. Free energy calculations were performed using non-softcore thermodynamic integration (TI) implemented in Amber (37, 38). The Amber force field ff14SB and the TIP3P water model were used for the TI calculations (39, 40). Minimization and equilibration under constant pressure (41) were conducted to heat up and relax the X-ray structures. Production runs were conducted using the implementation of GPU-accelerated thermodynamic integration, pmemdGTI (42), under constant volume. Energy minimization was conducted using gradient descent algorithm with 100 kcal/mol position restraints on all heavy atoms of proteins. The maximal number of cycles is 10000. A 0.1 ns constant volume MD simulation was then conducted to slowly heat up the structures from 150K to 294K with 100 kcal/mol position restraints on all heavy atoms of proteins. A 0.1 ns constant pressure MD simulation was then conducted with 100 kcal/mol position restraints on all heavy atoms of proteins. A 0.25 ns constant pressure MD simulation was conducted with 10 kcal/mol position restraints on all heavy atoms of proteins. A 0.1 ns constant pressure MD simulation was conducted with 10 kcal/mol position restraints on all CA,C and N atoms. A 0.1 ns constant pressure MD simulation was conducted with 1 kcal/mol position restraints on all CA,C and N atoms. A 0.1 ns constant pressure MD simulation was conducted with 0.1 kcal/mol position restraints on all CA, C and N atoms. The mutation site (residue 18 of OMTKY3) was excluded in the restraints. In the last step 0.25 ns constant pressure MD simulation was conducted with no restraints. An 1 fs step size was used for the equilibration. The temperature was set to 294K and no salt ions was included. Langevin dynamics was used to control temperature and the collision frequency was set to be 1.0 ps^{-1}. Particle mesh Ewald methods were used to calculate electrostatic energies (43). Hydrogen atoms were constrained using the SHAKE algorithm (44). The cutoff of non-bonded interactions was set to 8 Å. A timestep of 2fs was used. The simulation time length for each λ window is 4ns. The trapezoidal rule was used for the integration of all λ windows.

Intermediate states were created using the program ParmEd, so the overall transition was divided into several steps. Since the single topology approach is used here, the transition always starts from a "large" amino acid to a "smaller" amino acid. The first step in the transition involves the changing of partial charges. The partial charges of atoms in the ligand were changed to the partial charges of the corresponding atoms in the end states and the partial charges on disappearing atoms were set to be zero. No charges were added as the end states always has fewer atoms than the starting states.

The subsequent steps involved removing the LJ interactions of the disappearing atoms. Since non-softcore TI was used in our calculations, for each transition, we designed multiple intermediates to minimize the perturbation caused by removing LJ interactions. For each transition, we trimmed off the heavy atoms furthest away from the common structures by removing the LJ interactions of the heavy atoms and the LJ interactions of any hydrogens attached to the heavy atoms. Based on our experience, if the absolute change of partial charge on the disappearing atom is less than 0.3 units of elementary charge, the removal of the LJ interactions can be done with the removal of partial charges on the atom in the same transition step. Otherwise, the removal of the LJ interaction must be done after the change of partial charges. It is possible to trim multiple heavy atoms in a single transition step as long as the heavy atoms they attached to are not also disappearing in this step. For example, the two CD atoms and their attaching hydrogen atoms can be trimmed off in a single transition step instead of removing one methyl group at a time. In the last transition step, all the bond, angle and dihedral interactions were changed accordingly.

3.2.2 Free Energy Calculations for D-to-A Mutations in SGPB/OMTKY3.

Analysis of the D-to-A mutations requires knowledge of the protonation state of the Asp in the free and complexed state. The free state pKa is available from NMR measurements and the bound state pKa has been estimated from experimental pKa dependent binding free energies. The NMR-based pKa of Asp18 in the free state of OMTKY3-Asp18 is 3.87 (45). The pKa values of Asp18 in the SGPB/OMTKY3-Asp18 complex were estimated to be 9.26 using the experimentally determined pH dependence of association equilibrium constants (46). The experimental pH for the binding affinity measurement was 8.30 (47), which indicates that Asp18 is fully deprotonated in the unbound state but about 90% protonated in the complex state. The thermodynamic cycle of OMTKY3-Asp18 to SGPB can be described as shown in **Fig. 3-5**. To calculate the binding free energy difference between SGPB/OMTKY3-Ala18 and SGPB/OMTKY3-Asp18, which is equivalent to ③ - ⑥, one need to calculate ④ and ⑤ in addition to ② - ① due to the change of protonation state of Asp18 upon binding. The protonation free energy of Asp18 in the free and complex state of OMTKY3 at pH=8.30 can be calculated as:

$$\Delta G = 2.303RT(pH - pKa_{free/com}^{Asp18}) \tag{16}$$

$pKa_{free/com}^{Asp18}$ is the pKa of Asp18 of OMTKY3 in the free or complex state. The binding free energy difference between SGPB/OMTKY3-Ala18 and SGPB/OMTKY3-Asp18 can be obtained by ②+⑤-①-④.

3.2.3 Molecular Dynamics (MD) Simulations of the NTAIL/XD Mutants.

The starting structure for the MD simulations of wildtype NTAIL/XD complexes was obtained from PDB code 1T6O (48). *In silico* mutations of A494G, L495A and L498A were made using Swiss PDB (49). Hydrogen atoms were added using the MolProbity program (36). Residues G484 and S485 which belong to the artificial linker between NTAIL and XD in the X-ray structure were removed and the C-terminus and the N-terminus of NTAIL were amidated and acetylated respectively. LYS, ARG and free N-termini were set to be protonated and ASP, GLU and free C-termini were set to be unprotonated. His498 of XD was set to be neutral with the proton on Nε. The protocol for the simulations was the same as the one used for the TI calculations of SGPB/OMTKY3 except a temperature of 298K was used here. The simulations were 100ns long for each mutant and the last frames of simulations were saved for the backward TI calculations of NTAIL/XD.

3.2.4 Free Energy Calculations for the NTAIL/XD Complexes, Tetrapeptides and Fully Helical NTAIL (486-504)

The procedure for the TI calculations of NTAIL/XD complexes, tetrapeptides and fully helical α-MoRE was the same as the one used for the TI calculations of SGPB/OMTKY3, except that a length of 20ns was used for all windows of TI calculations on tetrapeptides, and a window of 4ns used for the other two calculations.

The fully helical NTAIL (486-504) segment was obtained by stripping off the XD and the artificial linker in the X-ray structure of NTAIL/XD complex (pdb code 1T6O). The C-terminus and the N-terminus were amidated and acetylated respectively.

Three independent TI calculations with different initial velocities were conducted for the tetrapeptides and the fully helical NTAIL (486-504).

3.3 Results

3.3.1 A Thermodynamic Cycle for Analyzing the Energetics of the Free State of IDPs

We calculate the free energy changes caused by mutations in IDPs by taking a detour through a thermodynamic cycle (**Fig. 3-1**). The measurable binding free energy of the wild-type and mutant IDP are combined with high-level thermodynamic integration (TI) calculations on the folded complexes to define the energetics of the free state of the IDP. The approach relies on a thermodynamic cycle, in which the values of three branches of the cycle define the value of the fourth (**Fig. 3-1**). For example, the effect of changing an alanine to a glycine in the unbound state of an IDP, ΔG_{free}, can be obtained by $\Delta G_{free} = \Delta G_{bind} - \Delta G'_{bind} + \Delta G_{com}$. ΔG_{bind} and $\Delta G'_{bind}$ are the experimental binding free energies of the mutant and wildtype respectively. ΔG_{com} is the free energy change caused by mutation of Ala to Gly in the complex and is calculated using alchemical free energy calculations. The effect of the mutation in a capped tripeptide or a small fragment, in which the mutated residue is in the middle of the fragment and flanked by the same amino acids found next to the site in the full-length IDP, also needs to be calculated to provide a reference state and is denoted as ΔG_{frag}. The capped tripeptide is not meant to represent the unbound state of IDP as it ignores secondary structure and long-range interactions. Rather the capped tripeptide is just a necessarily bookkeeping device that accounts for purely local interactions and for the fact that the different sized side chains will make different interactions with water. It is important to reiterate that the capped tripeptide is not used as a model of the IDP, rather the differences in the values of ΔG_{free} and ΔG_{frag} denoted as $\Delta\Delta G_{inter}$, quantitatively defines the energetic effect of non-local interactions that are not present in the peptide model.

In principle, the effect of mutations on interactions in the unbound state of IDPs could be directly calculated using alchemical free energy calculations, but such calculations are practically impossible due to the dynamic nature of the free state which leads to insufficient sampling. The approach developed here circumvents this issue by bridging the complex state and the free state using the binding free energies measured by experiments such as isothermal titration calorimetry (ITC), differential scanning calorimetry (DSC), binding kinetic experiments and others. The experimental binding free energies (ΔG_{bind} and $\Delta G'_{bind}$) provide relationships between the phase space of the bound and free state of the IDP. This allows the calculated free energy change in the complex state (ΔG_{com}) to be combined with experimental binding data to deduce the free energy changes in the free state (ΔG_{free}).

3.3.2 Another Interpretation of the Approach

From another perspective, this approach can be interpreted as a method which obtains the effects of mutations on the free state of IDPs by deconvoluting the experimentally measured binding free energy changes. The values of ΔG_{bind} and $\Delta G'_{bind}$ are the binding affinities of the wild-type and mutant measured experimentally using techniques such as ITC, DSC, binding kinetic experiments and others. The value of ΔG_{bind} - $\Delta G'_{bind}$ represents the effect of the mutation on the binding affinity of the IDP and its binding partner. However, ΔG_{bind} - $\Delta G'_{bind}$ is a convolution of the effect of the mutation on both the complex and free state of the IDP. This is analogous to protein folding where $\Delta G_{mutation}$ contains contributions from the folded and unfolded states. If the effect of the mutation on the free state is to be quantified, ΔG_{bind} - $\Delta G'_{bind}$ must be deconvoluted. In our approach, ΔG_{bind} - $\Delta G'_{bind}$ is deconvoluted by calculating the effect of mutation on the complex state (ΔG_{com}) using free energy calculations. The deconvolution leads to the effect of the mutation on the free state (ΔG_{free}) of the IDP via a thermodynamic cycle. The resulting value of ΔG_{free} is another convolution of the purely local and long-range interactions disrupted by the mutation. The calculation using the capped tripeptide (ΔG_{frag}), which only describes the local interactions, allows further deconvolution of ΔG_{free} so that the long-range interactions disrupted by the mutation ($\Delta\Delta G_{inter}$) can be quantified. One practical consideration is that some IDPs retain disordered tails in the bound state. If necessary, the disordered regions in the bound state can be truncated during the calculation of ΔG_{com} to increase computational efficiency provided that the disordered regions do not alter the conformation or energetics of the structured regions under investigation in the complex. This is likely to be valid for residues which are buried in the binding interface and sequestered from solvent since they are protected from transient contacts with any disordered segments.

3.3.3 IDP Complexes and Mutation Sites that are Suitable for the Approach

Many IDPs retain a certain degree of so-called "fuzziness" upon binding to their partners and are relatively dynamic even in the bound state. These IDPs may remain partially or even fully flexible upon binding and they can adopt multiple binding conformations (50-54). Similar to the free state of IDPs, calculations of ΔG_{com} for this type of IDPs are challenging due to the vast conformational ensembles that need to be sampled. However, some IDPs that form a "fuzzy" complex with one of their partners may form a rigid complex with a different partner (54). Since the energetics of the free state of IDPs are independent of their binding partners and their structures in the complex state, our approach can be applied to any IDP as long as it forms a rigid complex with one of its natural binding partners or an engineered binding partner and the binding free energies can be measured experimentally.

For IDPs that form both structured and disordered regions upon binding, it is easiest to study the residues that belong to the structured regions since large segments of disordered regions are too computationally expensive to model directly. Thus, truncated IDP complexes, with only the structured region, may be used during the calculation of ΔG_{com} provided certain conditions are met. This approximation requires that the disordered regions do not form strong interactions with the structured regions and do not alter the conformations of the structured regions of the complexes. Moreover, analysis of surface mutations should be avoided if using truncated IDP complexes as they may form transient contacts with the disordered regions in experiments that will be missing during the calculations of ΔG_{com} using truncated models. There is a significantly smaller possibility of forming direct contacts between buried residues and residues in the disordered regions, so the effect of transient interactions with the disordered regions on these buried residues can be more safely ignored. For the example of the NTAIL/XD complex studied here, residues 486 to 502 of NTAIL form a so-called α-helical Molecular Recognition Element (αMoRE) and fold into a stable helix upon binding to XD, but the regions preceding (401-485) and following (503-525) αMoRE remain disordered upon binding, do not contact the structured region and make at most a small contribution to binding enerergetics (5, 32, 34, 48, 55-58).

3.3.4 Ovomucoid Inhibitor Protease Interactions Provide an Excellent System to Validate the Approach

The accurate estimation of ΔG_{inter} depends on precise and accurate free energy calculations since they are combined with experimental binding affinities to make predictions about the free state. Thus, it is necessary to evaluate critically the accuracy of alchemical free energy calculations and define their limitations. In a previous study, we successfully reproduced the experimentally measured effects of Gly-to-D-Ala substitutions on the unfolding free energy changes in eight proteins using TI calculations with a root-mean-square error of 0.23 kcal/mol (24). In this study, using the same force field and solvent model, we further validated the accuracy of our free energy calculations on a more relevant system involving calculation of protein-protein binding energetics. We tested the ability of our protocol to reproduce the experimental binding free energy of the turkey ovomucoid inhibitor third binding domain (OMTKY3) to its target proteases using TI calculations (**Fig. 3-2**). This is an excellent model system: high resolution structures of the free and bound states are available and precise thermodynamic binding data has been reported for multiple mutations (47, 59, 60). For the complexes between OMTKY3 and *Streptomyces griseus* proteinase B (SGPB) (**Fig. 3-2**), high resolution crystal structures for all substitutions at position 18 of OMTKY3 have been reported except Met18 and Cys18 (59, 60). The perturbations to the structures of the complexes caused by mutation at position 18 of OMTKY3 are minimal. For the 10 variants studied here (OMTKY3/Leu18, Ala18, Gly18, Asn18, Asp18, Val18, Thr18, Ser18, Phe18 and Tyr18), the root-mean-square deviations for the backbone coordinates are at most 0.158 Å (PDB codes listed in SI). The small differences in structures minimize the complexity of the free energy calculations on the bound state of the SGPB/OMTKY3 complex. Furthermore, the mutation site is only partially buried in the complex interface which allows the efficient exchange of water around the side chain during the free energy calculations of complexes. This avoids any complications that might arise from having different numbers of waters in the interface for different side chains. In the unbound state, OMTKY3 is highly stable and the mutation site is located in a short loop with the side chain fully exposed to solvent. These factors make SGPB/OMTKY3 an ideal system for testing the accuracy of free energy calculations. Calculations were also conducted on a Leu-to-Ala mutation in the complex of a different protease, subtilisin Carlsberg (CARL), with OMTKY3 to check if the calculation is sensitive enough to reproduce context-dependent $\Delta\Delta G$ values (2.95 kcal/mol for SGPB/OMTKY3 and 0.33 kcal/mol for CARL/OMTKY3) for Leu-to-Ala mutations (61). This is important because free energy changes for two Leu-to-Ala mutations in NTAIL were studied later. The difference between $\Delta\Delta G_{exp}$ (= ΔG_{bind} - $\Delta G'_{bind}$) and $\Delta\Delta G_{calc}$ (= ΔG_{free} - ΔG_{com}) (see free energy cycle in **Fig. 3-2**) provides a rigorous test of the accuracy of the free energy calculations.

The Amber force field ff14SB (39) and the TIP3P (40) water model were used for the TI calculations (37, 62), with the implementation of GPU-accelerated thermodynamic integration, using pmemdGTI (42). In order to test the convergence of TI calculations, two independent TI calculations were carried out for each ***X***-to-***Y*** substitution in the SGPB/OMTKY3 complex. One calculation started with the structure of SGPB/OMTKY3-***X***18, and the other calculation started with the structure of SGPB/OMTKY3-***Y***18. A total of 22 independent TI calculations were carried out.

The calculations show excellent agreement with experiments. The root-mean-square error between $\Delta\Delta G_{calc}$ and $\Delta\Delta G_{exp}$ for all mutations was 0.86 kcal/mol (**Fig. 3-3**). V-to-A, Y-to-F and S-to-C have more significant errors, with deviations of 1.30 ~ 2.11 kcal/mol respectively from the experimental results, while all other mutations have errors below 0.45 kcal/mol. The cause of the large errors in V-to-A, S-to-C and Y-to-F mutations are discussed below. Excluding these three apparent outliers reduces the root-mean-square error to 0.27 kcal/mol and gives an even stronger correlation between $\Delta\Delta G_{calc}$ and $\Delta\Delta G_{exp}$ with slope = 1.00 and $R^2 = 0.98$, $p < 10^{-13}$. The precision of the method was tested by carrying out two independent calculations using different starting structures for each mutation (except L-to-A in CARL/OMTKY3 and C-to-S in SGPB/OMTKY3, see caption of **Fig. 3-3**). $\Delta\Delta G_{calc}$ values were essentially identical with an average absolute difference of 0.25 kcal/mol, indicating high precision and good convergence of the calculations. The ΔG_{free} values estimated by $\Delta G_{free} = \Delta G_{bind} - \Delta G'_{bind} + \Delta G_{com}$ and the ΔG_{free} values directly calculated by TI calculations are compared in **Table 3-4**. Overall, the results demonstrate that the TI calculations are accurate and precise enough to be combined with experimental binding free energies as outlined in **Fig. 3-1** with the exception of the mutations which lead to the outliers. One of the outliers, the S-to-C mutation, may be caused by problematic Lennard-Jones (LJ) parameters for sulfur in ff14SB. ε, which defines the minimum of the 6-12 LJ interaction profile, for sulfur in Cys is only 1.2 times that of the hydroxyl oxygen in Ser, but the difference is larger in other force fields: 2.5 times in the OPLS-AA/m (63) and 3.0 times in the CHARMM36m force field(26). Calculations using LJ parameters from OPLS-AA/m reduced the error to 0.8 kcal/mol. The other outlier, the Y-to-F mutation, may due to a buried and structured water forming a hydrogen bond with the hydroxyl group of the Tyr as observed in the X-ray structures (**Fig. 3-6**). This water molecule was included in the TI calculations. However, it is unclear whether this water should be included, since the water may be displaced at the temperature of the experiments. Moreover, the direction dependency of hydrogen bonds is not well described in fixed-charge force fields which may also lead to inaccurate interaction strengths of hydrogen bonds (64, 65). In addition, there may be an issue with the TIP3P water model that leads to inaccuracies in modeling the geometries of ordered waters. These observations suggest caution needs to be employed when conducting TI calculations on systems with bound water that participate in hydrogen bond interactions. It is surprising that the calculated $\Delta\Delta G_{calc}$ for the V-to-A mutation has the largest error because the difference between V and A in terms of size and hydrophobicity is relatively small among the mutations studied here. Moreover, the calculated $\Delta\Delta G$ values of I-to-V and V-to-T are in good agreement with the experimental results as are most non-beta-branched to non-beta-branched mutations. The likely explanation is that the force field has a poor transferability between beta-branched and non-beta-branched amino acids.

3.3.5 Application to the NTAIL Domain: Identification of Long-range Interactions

We next applied the strategy to the NTAIL domain. Residues 486 to 502 of NTAIL, form a so called α-helical Molecular Recognition Element (αMoRE), and fold into a stable helix upon binding to the XD, thereby forming a intermolecular four-helix bundle complex(32, 48, 55). The αMoRE of NTAIL has residual helicity in the unbound state, but the regions preceding (401-485) and following (503-525) αMoRE have much less residual structure (5, 34, 55-58). The regions preceding (401-485) and following (503-525) the αMoRE remain disordered in the bound state and experimental data indicates that they do not form direct contacts with XD in the complex (5, 32, 34, 48, 55-58). In addition, ITC and surface plasmon resonance experiments indicate that the regions following the αMoRE of NTAIL make only minimal contributions to the binding between NTAIL and XD (66, 67). This data suggests that the disordered regions of NTAIL/XD complex do not alter the conformations of the structured region formed by the αMoRE and XD.

We studied the effect of, A494G, L495A and L498A mutations on the free state of NTAIL. A494, L495 and L498 are within the αMoRE region of NTAIL and located in the interface of NTAIL/XD complex (**Fig. 3-4**). The three mutations were chosen for the following reasons. Upon binding to the XD domain, the αMoRE region of NTAIL becomes folded while other regions of NTAIL remains disordered (32, 48, 55). The disordered regions may make transient contacts with surface residues on the NTAIL/XD complex. However, since A494, L495 and L498 are buried in the interface of NTAIL/XD complex, there is less possibility of forming direct contacts with the disordered regions, so the effect of direct interactions with the disordered regions on these residues can be safely ignored. This eliminates the need to model the disordered regions in the TI calculations, which is unfeasible due to insufficient sampling and uncertainty in current force fields as mentioned above.

The binding free energy changes (ΔG_{bind} - $\Delta G'_{bind}$) caused by the A494G, L495A and L498A mutations have been reported (68). The free energy changes caused by the mutations (ΔG_{com}) in the NTAIL/XD complex were calculated using the X-ray structure (PDB code 1T6O) and the protocol that we validated with the SGPB/OMTKY3 complexes. In order to check the convergence of the calculations of ΔG_{com}, two independent TI calculations for each mutation were conducted. One started with the X-ray structure which is the structure of the wild-type NTAIL/XD complex and proceeded toward the mutant. Only the structure of the wild-type NTAIL/XD complex is available, so the second starting structure was a model of the mutant complex. *In silico* mutations of A494G, L495A and L498A were applied to the X-ray structure of NTAIL/XD complex and a 100ns MD simulation for each mutant was conducted to relax any perturbation caused by the mutations. The TI calculations were performed starting with the structure from the last frame of the MD simulation of the mutant complex and proceeded in reverse to the wild-type complex. The forward and reverse calculations were in excellent agreement with differences ranging from only 0.06 to 0.36 kcal/mol (**Table 3-1**).

The free energy changes caused by the mutations in the free state of NTAIL (ΔG_{free}) are all significantly higher than the free energy changes in the tetrapeptide reference state (ΔG_{frag}) with differences ranging from 0.8 to 3.8 kcal/mol (**Table 3-1**). This indicates that the residues are involved in interactions in the free state which are not captured by the tetrapeptide reference state.

In other words, these residues are involved in secondary structure or long-range interactions, or both in the free state.

Because the αMoREs region was experimentally found to have residual helicity in the free state of NTAIL (5, 34, 55-58), we examined whether the difference between ΔG_{free} and ΔG_{frag} could be explained by helicity rather than long-range interactions in the unbound (IDP) state. TI calculations using a fully helical conformation of unbound NTAIL (486-504) monomer (ΔG_{helix}) were conducted. The structure was adopted from the PDB code 1T6O by deleting the XD and the artificial linker. If the contributions from the long-range interactions are negligible, then the free energy changes caused by the mutations in the free state of NTAIL (ΔG_{free}) should be between those of ΔG_{frag} and ΔG_{helix} depending on the fraction of helicity adopted in the free IDP. However, we found that ΔG_{free} values are still higher than the ΔG_{helix} values, especially for the L495A mutations. This argues that any propensity to adopt helical structure cannot explain the ΔG_{free} values, and that A494, L495 and L498 are all involved in favorable long-range interactions in the free state of NTAIL. This is especially notable at position 495, where the long-range interactions favor leucine over alanine.

3.3.6 Analysis of the discrepancy between experimental and calculated ΔΔG of SGPB/OMTKY3-Val18, Thr18 and Ala18

The comparison between the X-ray structures of SGPB/OMTKY3-Val18 and Ala18 (**Fig. A1-1**) clearly indicates that they differ significantly at the CB atoms of Val18 and Ala18 in OMTKY3. This suggested that Val18 may experience steric clashes in the complex between SGPB/OMTKY3. It is plausible by judging the positions of the CB atoms of Val18 and Ala18 that the CG2 atoms of Val18 experiences strong VDW repulsive interactions from its neighboring atoms. However, the distances between the CG2 atom and its imminently adjacent heavy atoms are seemingly fine. This observation can explain the experimental binding affinities which show that Ala-to-2-Abu mutation increases the affinity by 0.8 kcal/mol, but the 2-Abu-to-Val mutation decreases the affinity by 1.1 kal/mol. However, somehow the perturbation observed in the X-ray structures was not captured during the calculations. A TI calculation of Val18-to-Ala was also conducted with 1kcal/mol strong position restraints on all residues except Val18, but an error of 2 kcal/mol still persists (47, 60).

A similar effect was also observed in SGPB/OMTKY3-Thr18. Although in the X-ray structure of SGPB/OMTKY3-Thr18, the CB atoms of Thr18 and Ala18 overlap with each other, the χ1 dihedral of Thr18 adopts a highly unfavorable value. This also indicates that the OG1 atom of Thr18 may experience repulsive interactions from its neighboring atoms. The ΔΔG value for a direct transition from Thr18-to-Ala is 2.4 kcal/mol, which also has a 2.5 kcal/mol discrepancy compared to the experimental ΔΔG value (47, 60).

We focused our efforts on the analysis of Val18 because the Val-to-Ala mutation only involves change of VDW interactions, but the Thr-to-Ala mutation also involves change of electrostatic interactions in addition to VDW interactions. Since the φ/ψ angles of Val18 (-117˚, 49˚) in the SGPB/OMTKY3 complex is significantly different from the φ/ψ angles used in the training of χ1 corrections for Val. Energy scans on different χ1 of Val using ff14SB/ff18SB and QM were performed on a Val dipeptide in which the φ/ψ values of the Val is the same as the φ/ψ values of Val18 observed in X-ray structures of SGPB/OMTKY3-Val18. The energy scans were performed using the same protocol used by Maier, J. (39). The energies for different χ1 calculated by QM/gas phase and ff14SB/gas phase do not show significant in the two set of φ/ψ angles (**Fig. A1-3**). It is important to note that the energies in figure A were zeroed to conformation with χ1/ φ/ψ=-180/-60/-45. Energies in figure B were zeroed to conformation of χ1/φ/ψ=-180/-117/47. The energies were also calculated in implicit solvent. Although the MM energies were calculated using ff18SB, the relative energies of different $\chi1$ in a single backbone conformation should be the same as ff14SB because the additional CMAP term in ff18SB is independent of $\chi1$. The results indicate that gauche$^-$ in the context of φ/ψ=-117/47 is less favored than other rotamer state. Thus, correction on $\chi1$ in the specific backbone conformation with φ/ψ =-117/47 will cause Val18 to be even more favorable than Ala18 in the complex state.

Since a β-branched amino acid has strong energy coupling between $\chi1$ and φ/ψ angles, we also investigated whether two CMAP type correction terms on $\chi1/\varphi$ and $\chi1/\psi$ cause any difference. The exact atoms involved in the two CMAP correction terms are CG1-CB-CA-C-N(i+1) and CG1-CB-CA-N-C(i-1). Note that CB-CA-N-C(i-1) and CB-CA-C-N(i+1) are different from φ and ψ. The CB atom replaces the C atom in φ and replaces the N atom in ψ. This is because the central three atoms must be shared by the two coupled dihedral angles in CMAP. The energy map for these two CMAP corrections can be found in **Fig. A1-5**. However, the corrections lower the energy of Val with a conformation of $\chi1/\varphi/\psi$=-60/-117/47, which is the conformation of Val18 in the complex of SGPB/OMTKY3. The difference between the calculated and experimental ΔΔG of Val18-to-Ala is 4 kcal/mol with the CMAP correction, which is even larger than using the original ff14SB.

OPLS-AA/m force field was also used for the same calculation, but it also gave an error around 2.0 kcal/mol. A calculation using the Amoeba protein force field also gave an error of >2.0 kcal/mol. The software package, TINKER, was used for calculations with Amoeba force field. ΔΔG values for Val18-to-Ala mutations using different force fields are listed in **Table 3-5**.

3.3.7 Analysis of the discrepancy between experimental and calculated ΔΔG of SGPB/OMTKY3-Tyr18-to-Phe, Met18-to-Ala and Cys18-to-Ser

A set of free energy calculations using different force fields and water models were conducted. The results are listed in **Table 3-6**. The software package, TINKER, was used for calculations with Amoeba force field. Although different force fields and water models show different ΔΔG values for Y18-to-F mutations, the difference between them are much smaller than their errors compared to the experimental value of -0.3kcal/mol(47, 60).

Amber and OPLS protein force fields used to have the same LJ parameters for sulfur atoms in Met and Cys, but OPLS has adopted new parameters since OPLS-AA/L (69). Both set of LJ parameters of sulfur can reproduce experimentally measured density and heat of vaporization of CH3-SH and CH3-CH2-SH. However, the new parameters adopted by OPLS have significantly better binding energies of CH3-SH/CH3-SH and CH3-SH/CH3-OH when compared to ab initio values. Replacing the LJ parameters of sulfur atoms in Amber with the ones from OPLS-AA/L force field significantly reduced the error of Met18-to-Ala and Cys18-to-Ser mutations by about 1.0 kcal/mol (**Table 3-7**). Refitting of $\chi 1$ and $\chi 2$ dihedral angle corrections may further reduce the errors.

However, changing the LJ parameters of sulfur cause the Cys to over-favor the left-handed helical conformation in regular MD simulation of Cys dipeptide. Further investigation is required to verify whether refitting the $\chi 1$ dihedral angle corrections in the context of new LJ parameters solve this issue.

3.4 Discussion

In this work, we have validated a hybrid strategy to measure quantitatively the free energy changes caused by mutations in the free state of IDPs, which cannot be measured directly using conventional experimental or computational techniques. A possible limitation of the strategy is that the IDP must form a structured complex with its receptor and many IDPs remain "fuzzy" upon binding (50-54). However, since the energetics of the free state of an IDP are independent of its binding partner, this approach can still be applied to the IDP if it folds upon binding to one a different natural binding partner or an engineered molecule. Another potential limitation is that for IDPs that remain partly disordered upon binding, the truncation of the disordered regions, which is ideal for the calculations of ΔG_{com}, is only rigorous if the disordered regions have minimal interactions with the structured regions. Ideally, the residues of interest will be buried in the complex to avoid transient long-range interactions with disordered regions of the bound complex that complicate the modeling. Fortunately, residues in the binding interface are usually buried and these are often of major interest since residues in the interface normally contribute the most to the binding affinity of a complex. Studying their interactions in the unbound IDP can provide quantitative information about the properties at the free state and about the possible roles of any free state interactions in the regulation of biological activity.

Our free energy calculations on SGPB/OMTKY3 complexes show that the free energy changes caused by most mutations can be reproduced with high precession and within a small error. Although the outliers V-to-A, C-to-S and Y-to-F were identified in our calculations, the validation included L, I, F, and A, which are the four most commonly observed amino acids in the binding interface between IDPs and their binding partners (70-72). On average, about 40% of the residues in the binding interface are one of these four residues (72). This indicates that the approach should be broadly applicable to folding upon binding IDPs. Since the approach relies on accurate calculations of free energy, the sampling convergence of the free energy calculations must be considered carefully (73). For example, buried mutations may cause displacement of interior water which usually has a slow relaxation beyond the timescale of MD simulations (74). However, new algorithms have been developed to address the limitations of sampling in free energy calculations

(75-77). We believe that as the reliability of these calculations continues to improve, the impact of our approach will increase.

Our analysis shows that A494, L495 and L498, which form part of the binding interface of NTAIL and XD, participate in favorable long-range interactions in the free state of NTAIL that modulate the binding affinity of NTAIL to XD. The strength of the long-range interactions made by A494 and L498 in the free state of NTAIL range from 0.3 to 1.6 kcal/mol depending on the helicity of A494 and L498. If NTAIL were to exist as a true random coil in its free state, then L495 would be 4.9 kcal/mol more favorable than A495 in the complex of NTAIL/XD. Even if L495 is in a fully helical structure in the free state of NTAIL, L495 is predicted to be > 4.1 kcal/mol more favorable than A495 in the complex state. However, the experimentally measured binding free energy change indicates that L495 is only 1.1 kcal/mol more favored than A495 in the complex. The difference can be explained by favorable long-range interactions experienced by L495 in the free state of NTAIL, which compete with the binding interactions involving L495. Stated differently, the interactions in the free state modulate the binding energetics and are predicted to reduce affinity. This provides an important mechanism for tuning binding affinity. We constructed a sequence alignment of know NTAIL sequences to test if A494, L495 and L498 are conserved. The sequence consensus of this alignment was visualized in **Fig 3-7**. The insights revealed by our approach correlate well with prior observations on NTAIL and illustrate that long-range interactions between the flanking disordered regions and the αMoRE can modulate the overall dimensions of NTAIL and the NTAIL/XD binding free energy.

It may be surprising that the interactions in NTAIL appear to favor L495 more than A495 by as much as 3.0 ~ 3.8 kcal/mol. However, if such interactions are responsible for inhibiting the aggregation of proteins (22, 78), then they should be reasonably strong. In addition, the full free energy contribution is very unlikely to arise from a pairwise interaction, instead it is highly likely that the residue in question makes interactions with more than one other residue in the free state, or that the mutation alters the unfolded state and thereby modulates multiple other interactions.

The identified long-range interactions identified in NTAIL may play a biological role. Dynamic binding and breaking of the nucleocapsid and polymerase are necessary to ensure the transcription and replication of the RNA encapsulated by nucleocapsids and the affinity needs to be tuned for optimal activity. Increasing and decreasing the binding affinity of NTAIL/XD both lead to a reduction in transcription activity and viral growth (79, 80), so a balanced interaction between NTAIL and XD is crucial. The long-range interactions involving A494, L495 and L498 in the free state attenuate the binding affinity to reach an optimal efficiency of transcription and replication. This provides a clear example of how interactions within the free state of an IDP can tune biological activities.

3.5 Conclusions

This work offers a general methodology for assessing the energetic contributions of individual residues in IDPs to the energetics of the free state of IDPs. The approach allows the identification of residues that participate in long-range interactions and thus may modulate binding affinity, but avoids the difficulties associated with MD simulations of free IDPs. Many IDPs become structured upon binding (81, 82), thus the strategy is expected to be broadly applicable and will become even easier to apply as computing power continues to increase. The work with the OMTKY3 and its binding partners also provides a rigorous evaluation of the accuracy and precision of TI calculations performed using the Amber ff14SB force field. The work also illustrates how IT calculations can be used to help validate force fields. This work demonstrates that advances in force fields and computing hardware have now led to the point where it is possible to develop novel methods which integrate experimental and computational techniques to reveal insights that cannot be studied by using either technique alone. The interactions and effects revealed by the analysis presented here could not be deduced from experiment and computation in isolation.

Table 3-1. Free energy changes (kcal/mol) calculated for A494G, L495A and L498A mutations in the complex, free, tetrapeptide fragment and fully helical state of NTAIL. The calculated ΔG values was referenced to the values of their corresponding ΔG_{frag}. The original values were listed in the parenthesis. The original ΔG values have no physical meaning because they contain a force field dependent baseline. The referenced ΔG does have physical meaning as the force field dependent baseline effect is cancelled.

	Forward ΔG_{com}	Backward ΔG_{com}	ΔG_{inter} [1]	ΔG_{frag} [2]	ΔG_{helix} [2]
A494G	2.36	2.22	1.59±0.13	0	1.22±0.08
	(-6.87)	(-7.01)	(-7.64±0.12)	(-9.23±0.04)	(-8.01±0.07)
L495A	5.09	4.73	3.81±0.21	0	0.81±0.16
	(23.75)	(23.39)	(22.47±0.21)	(18.66±0.04)	(19.47±0.15)
L498A	2.26	2.09	0.77±0.16	0	0.42±0.10
	(20.05)	(19.87)	(18.56±0.13)	(17.79±0.09)	(18.21±0.05)

1. The uncertainties are calculated by combining the standard deviation of ΔG_{com} with the published standard deviation of (ΔG_{bind} - $\Delta G'_{bind}$)(68).
2. The uncertainties are the standard deviation of three independent runs of MD simulations with different starting velocities.

Table 3-2. PDB codes for the structures of SGPB/OMTKY3 and CARL/OMTKY3 studied using TI calculation.

	PDB code
OMTKY3-LEU18	2GKR(83)
SGPB/OMTKY3-LEU18	1SGR(59)
SGPB/OMTKY3-ALA18	1SGP(59)
SGPB/OMTKY3-GLY18	1SGQ(59)
SGPB/OMTKY3-ASN18	1SGN
SGPB/OMTKY3-SER18	1CT0(60)
SGPB/OMTKY3-VAL18	1CT4(60)
SGPB/OMTKY3-ILE18	1CSO(60)
SGPB/OMTKY3-THR18	1CT2(60)
SGPB/OMTKY3-TYR18	1SGY
SGPB/OMTKY3-PHE18	2SGF
SGPB/OMTKY3-ASP18	1SGD
CARL/OMTKY3-LEU18	1R0R(61)

Table 3-3. $\Delta\Delta G_{calc}$ and $\Delta\Delta G_{exp}$ (kcal/mol) of mutations studied in SGPB/OMTKY3 and CARL/OMTKY3

	ΔG_{com} (forward)	ΔG_{com} (backward)	ΔG_{free}	$\Delta\Delta G_{calc}$ (forward)	$\Delta\Delta G_{calc}$ (backward)	$\Delta\Delta G_{exp}$
A to G	-6.64	-6.62	-8.92	2.27	2.31	1.99
S to A	7.56	7.64	8.54	-0.98	-0.90	-1.15
I to V	-21.78	-22.03	-20.21	-1.57	-1.82	-1.42
L to A	17.18	17.19	14.41	2.76	2.78	2.95
L to A (CARL)	14.54		14.41	0.14		0.31
L to N	-51.20	-51.82	-54.79	3.59	2.97	3.35
V to T	-9.60	-9.58	-9.62	0.03	0.05	0.16
L to F	22.72	23.22	21.42	1.31	1.81	1.36
V to A	16.83	16.65	14.82	2.01	1.83	-0.10
C to S		-8.37	-10.72		2.33	4.11
Y to F	22.45	22.25	21.25	1.20	1.00	-0.30
D to A *	54.71	54.02	50.86	-2.24	-2.93	-2.64

*: The ΔG_{com} and ΔG_{free}, which correspond to steps 2 and 1 in Fig S1 respectively, are calculated by changing neutral Asp to Ala. $\Delta\Delta G_{calc}$ is calculated as explained in the previous paragraphs.

Table 3-4. Comparison of ΔGfree obtained through ΔGfree = ΔGbind - ΔG'bind + ΔGcom and the directly calculated ΔGfree (kcal/mol) of mutations studied in SGPB/OMTKY3 and CARL/OMTKY3

	ΔG_{bind} - $\Delta G'_{bind}$ + ΔG_{com} (forward)	ΔG_{bind} - $\Delta G'_{bind}$ + ΔG_{com} (backward)	Directly calculated ΔG_{free}	Error (forward)	Error (backward)
A to G	-8.63	-8.61	-8.92	0.29	0.31
S to A	8.71	8.79	8.54	0.17	0.25
I to V	-20.36	-20.61	-20.21	-0.15	-0.4
L to A	14.23	14.24	14.41	-0.18	-0.17
L to A (CARL)	14.23		14.41	-0.18	
L to N	-54.55	-55.17	-54.79	-0.24	-0.38
V to T	-9.76	-9.74	-9.62	-0.14	-0.12
L to F	21.36	21.86	21.42	-0.06	-0.44
V to A	16.93	16.75	14.82	2.11	1.93
C to S		-12.48	-10.72		-1.76
Y to F	22.75	22.55	21.25	1.50	1.30
D to A *	51.26	50.57	50.86	0.40	-0.29

*: The ΔG_{com} and ΔG_{free}, which are ② and ① in Fig S1 respectively, are calculated by changing neutral Asp to Ala. $\Delta\Delta G_{calc}$ is calculated as explained in the previous paragraphs.

Table 3-5. ΔΔG values for Val18-to-Ala mutations using different force fields.

Calculated ΔΔG (kcal/mol)	Force fields + water model	Experimental ΔΔG (kcal/mol)
2.4	Amber ff14SB+TIP3P	0.10
2.5	Amber ff14SB+OPC	
2.6	Amber ff19SB+OPC	
4.0	Amber ff14SB + χ1/ϕ and χ1/ψ Cmap+TIP3P	
1.8	OPLS-AA/m+TIP4P(63)	
4.2	Amber ff15ipq+SPC-e(84)	
>2.0	Amoeba + Amoeba water model(85)	

Table 3-6. ΔΔG values for Tyr18-to-Phe mutations using different force fields.

Calculated ΔΔG (kcal/mol)	Force fields + water model	Experimental ΔΔG (kcal/mol)
1.3	Amber ff14SB+TIP3P	-0.30
>1.0	Amber ff14SB+OPC	
1.5	Amber ff15ipq+SPC-e(84)	
0.8	Amoeba + Amoeba water model(85)	

Table 3-7. ΔΔG values for Met18-to-Ala and Cys18-to-Ser mutations using different force fields.

	Calculated ΔΔG (kcal/mol)	Force fields + water model	Experimental ΔΔG (kcal/mol)
C18-to-S	2.4	Amber ff14SB+TIP3P	4.1
	3.5	Amber ff14SB+modified SG of C18 (r=2.0 Å, ε=0.425) (69)	
M18-to-A	1.1	Amber ff14SB+TIP3P	2.5
	1.7	Amber ff14SB+modified SD of M18 (r=2.0 Å, ε=0.355) (69)	

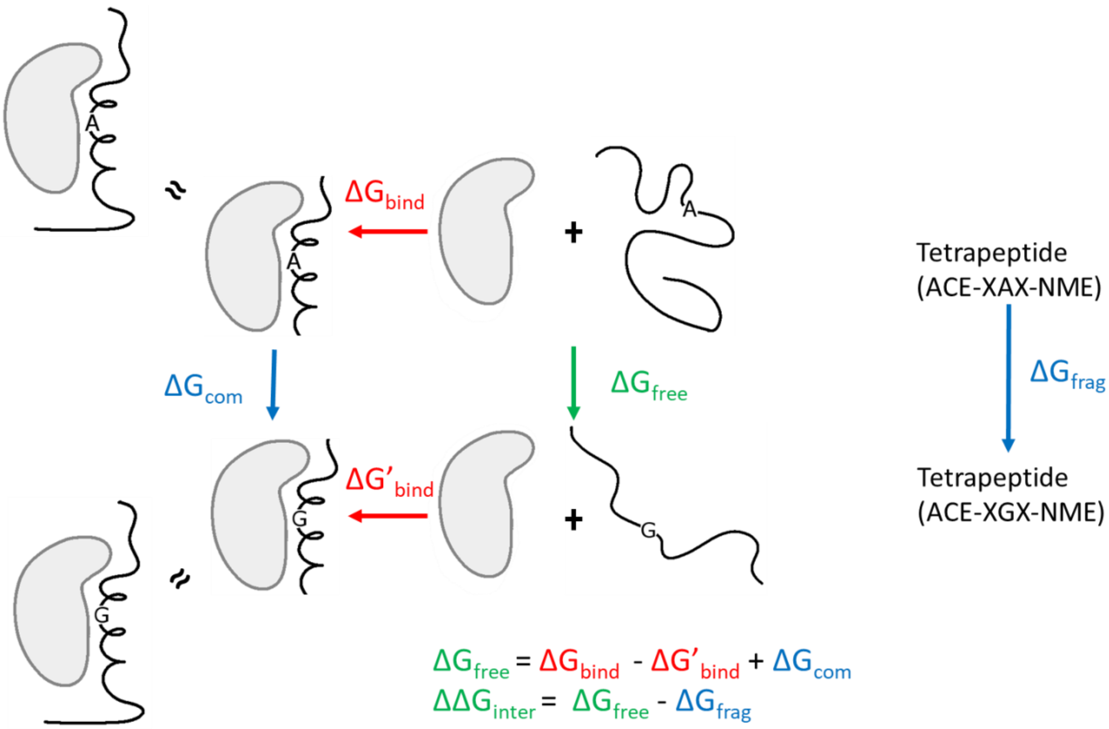


Figure 3-1. Illustration of the approach used to deduce the energetics of the free IDP. The thermodynamic cycle describes the binding of a wild-type IDP and a mutant IDP to its partner. The tetrapeptides have the same residues adjacent to the mutation site as found in the full protein. ΔG_{bind} and $\Delta G'_{bind}$ are the binding free energies measured by experiment for wildtype and mutant respectively (red text). ΔG_{com} and ΔG_{frag} are calculated using alchemical free energy calculations (blue text). The disordered regions in the complex state of the IDP are truncated during the calculation of ΔG_{com}. This approximation is valid if the disordered regions do not alter the conformations of the structured regions in the complex state. The value for ΔG_{free} is obtained from $\Delta G_{free} = \Delta G_{bind} - \Delta G'_{bind} + \Delta G_{com}$. The effect of secondary structure and long-range interactions on the mutations is obtained from $\Delta\Delta G_{inter} = \Delta G_{free} - \Delta G_{frag}$. $\Delta\Delta G_{inter}$ and ΔG_{free} (green text) cannot be measured by either experiments or calculations alone but can be obtained by combining the experimental and computational measurements.

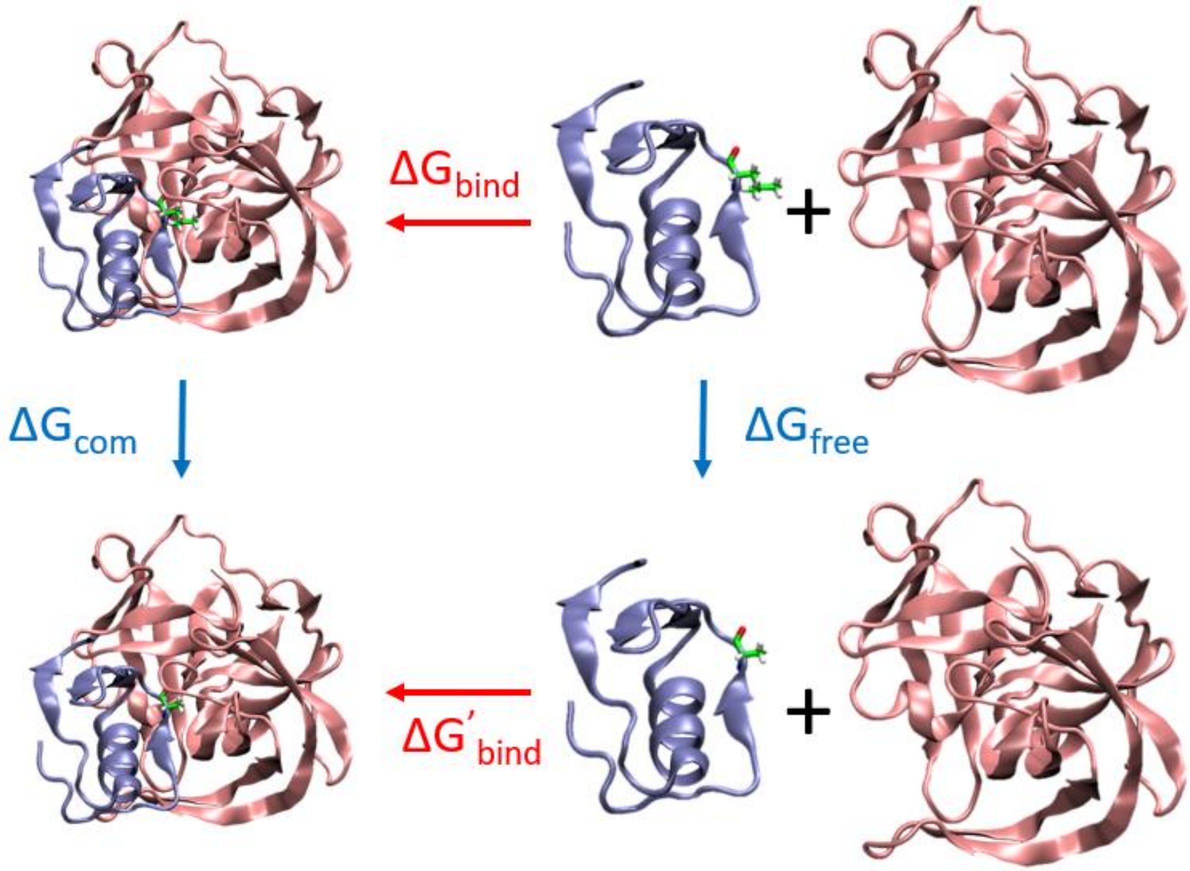


Figure 3-2. Free energy cycle of L-to-A mutations in the binding of SGPB and OMTKY3. Ribbon structures represents SGPB (pink), OMTKY3 (blue) and the SGPB/OMTKY3 complex. Leu18 (top) and Ala18 (bottom) of OMTKY3 are shown in stick format. ΔG values in red are binding free energies measured by experiment(47). ΔG values in blue are free energies calculated by using TI.

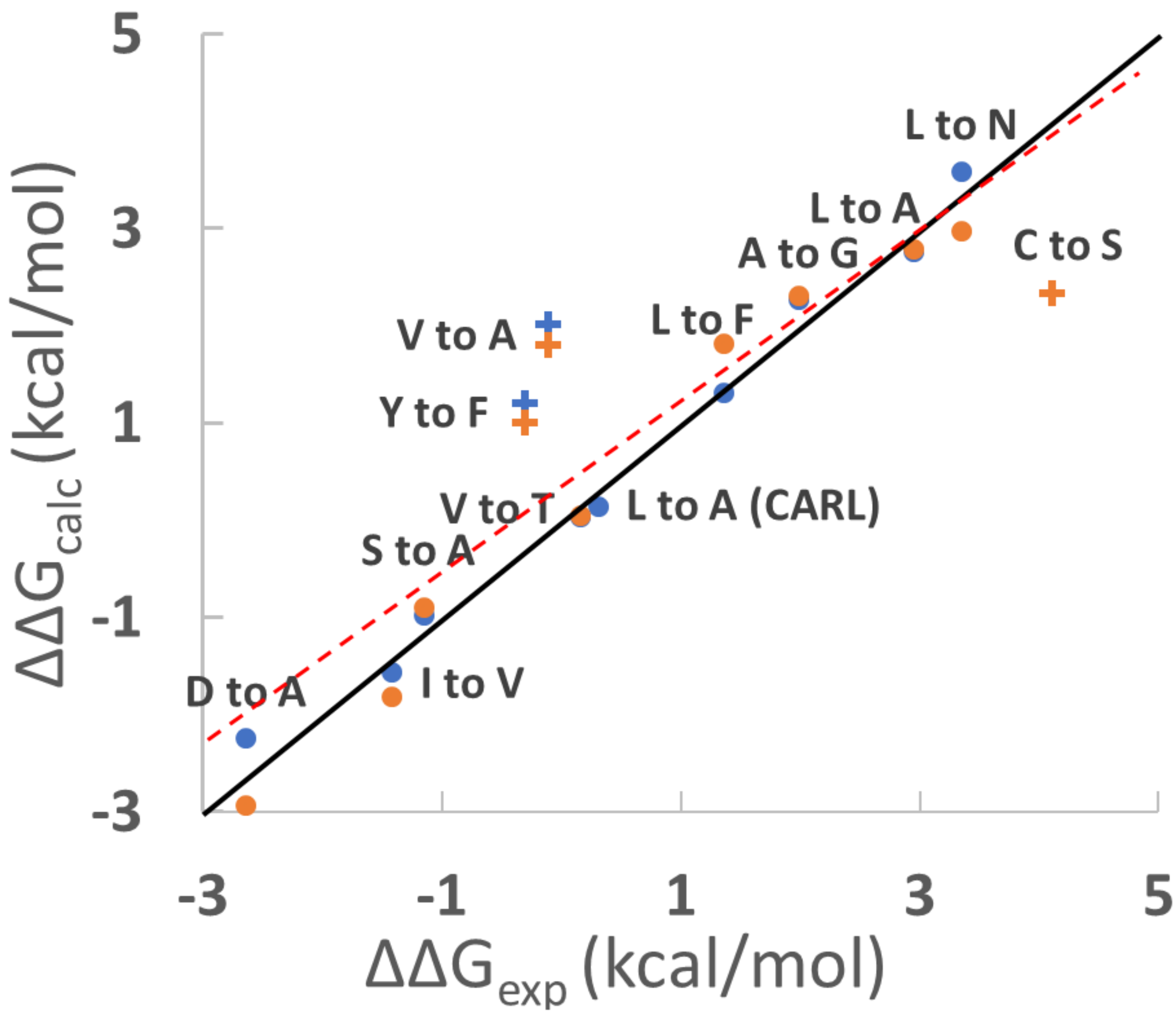


Figure 3-3. Scatter plot of experimental ($\Delta\Delta G_{exp}$) and calculated ($\Delta\Delta G_{calc}$) ΔΔG values for binding between SGPB and different OMTKY3 variants. $\Delta\Delta G_{exp} = \Delta G_{bind} - \Delta G'_{bind}$. $\Delta\Delta G_{calc} = \Delta G_{free} - \Delta G_{com}$. The point labelled with CARL is for the binding between subtilisin Carlsberg and OMTKY3 variants. Blue and orange dots indicate calculated ΔΔG values calculated using different starting structures. For example, the value for the blue dot for A to G was calculated using the structure of SGPB/OMTKY3-Ala18 (PDB code 1SGP) and the value of the orange dot for A to G was calculated using the structure of SGPB/OMTKY3-Gly18 (PDB code 1SGQ). There is no structure for subtilisin Carlsberg/OMTKY3-Ala18 so only one ΔΔG values using the structure of subtilisin Carlsberg/OMTKY3-Leu18 (PDB code 1R0R) was calculated. Similarly, only one ΔΔG for S-to-C mutation in SGPB/OMTKY3 was calculated. Three outliers, V-to-A, S-to-C and Y-to-F are labeled as crosses. All numerical values are provided in **Table S2**. The solid line represents $\Delta\Delta G_{exp} = \Delta\Delta G_{calc}$. The dashed red line indicates the results of linear regression of the data ($y=0.86x+0.31$, $R^2=0.82$ $p<10^{-9}$).

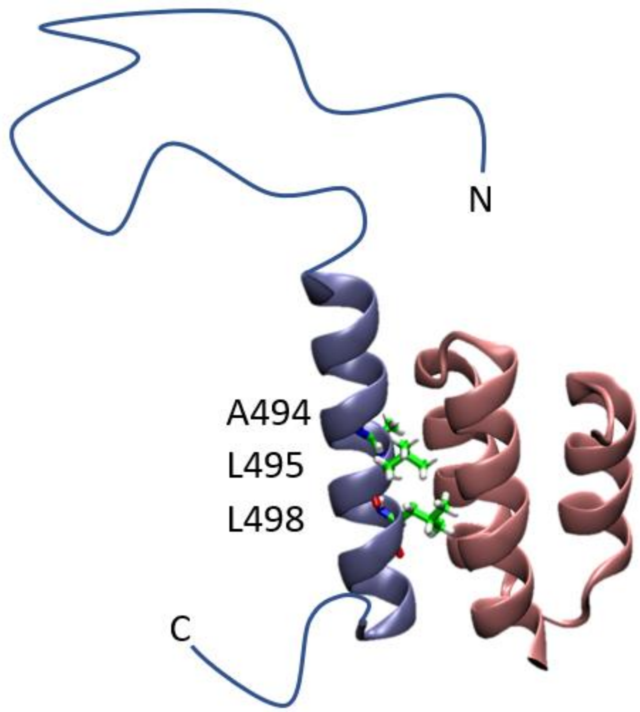


Figure 3-4. Ribbon representation of the NTAIL (486-504)/XD (458-506) complex from the X-ray structure (PDB code 1T6O). The X-ray structure includes the region shown in ribbons, and an artificial linker which was deleted in simulations and is not shown in this figure. Residue A494, L495 and L498 are shown in stick format. The disordered N and C-terminal regions of NTAIL were not included in the X-ray structure and MD simulations and are depicted schematically as thin lines.

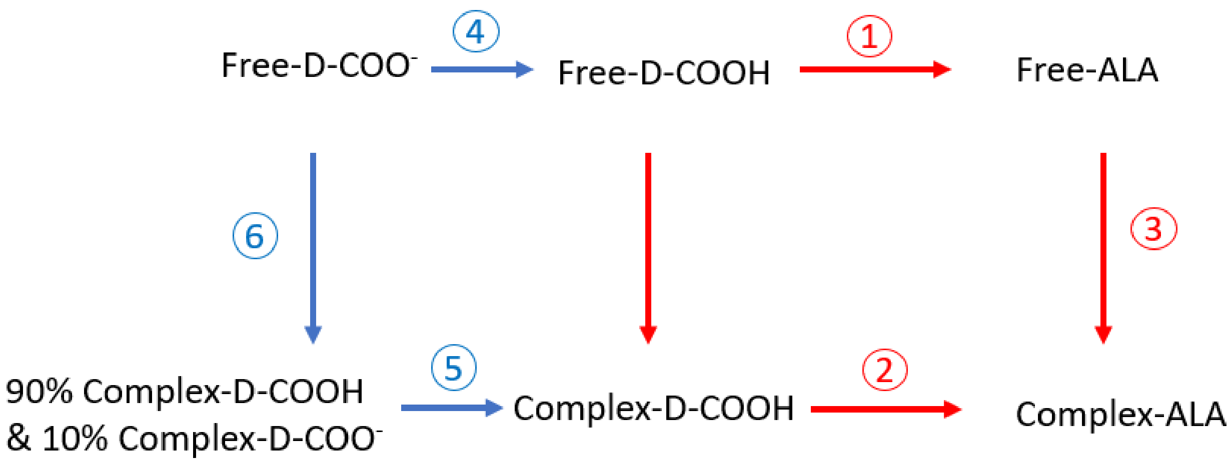


Figure 3-5. Thermodynamic cycle of OMTKY3-ASP18 forming a complex with SGPB with an additional process of protonation. D-COO^- indicates the deprotonated state of Asp18 and D-COOH indicates the protonated state of Asp18. ③ and ⑥ are the binding free energy of OMTKY3-Ala18 to SGPB and OMTKY3-Asp18 to SGPB respectively. ① and ② are the free energy changes of mutation from protonated aspartate to alanine calculated using TI in the unbound state of OMTKY3 and the SGPB/OMTKY3 respectively. ④ and ⑤ are the protonation free energy of Asp18 in the free state and complex state of OMTKY3 at pH 8.30 respectively. The red cycle is the thermodynamic cycle used for non-titratable residues in this study.

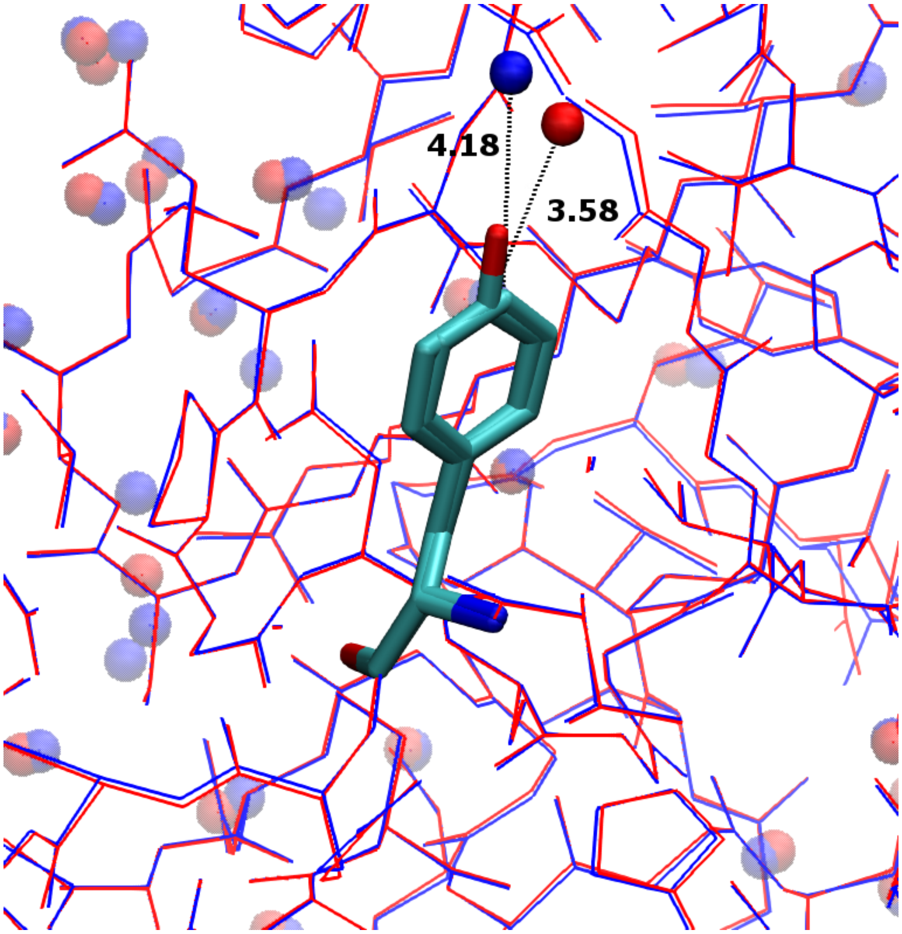


Figure 3-6. The X-ray structures of SGPB/OMYKY3-Phe18 (blue, pdb code 2sgf) and SGPB/OMTKY3-Tyr18 (red, pdb code 1sgy). Phe18 and Tyr18 are shown licorice. The other residues and water molecules are shown as lines and spheres respectively. Two water molecules found adjacent to the rings of Tyr18 and Phe18 are shown as opaque spheres. The distances between these two waters and the ζ-carbons of Phe18 and Tyr18 were measured are 4.18 and 3.58 Å respectively. The shorter distance between the water and the ζ-carbon of Ty18 may be caused by a hydrogen bond involving the hydroxyl group and the water.

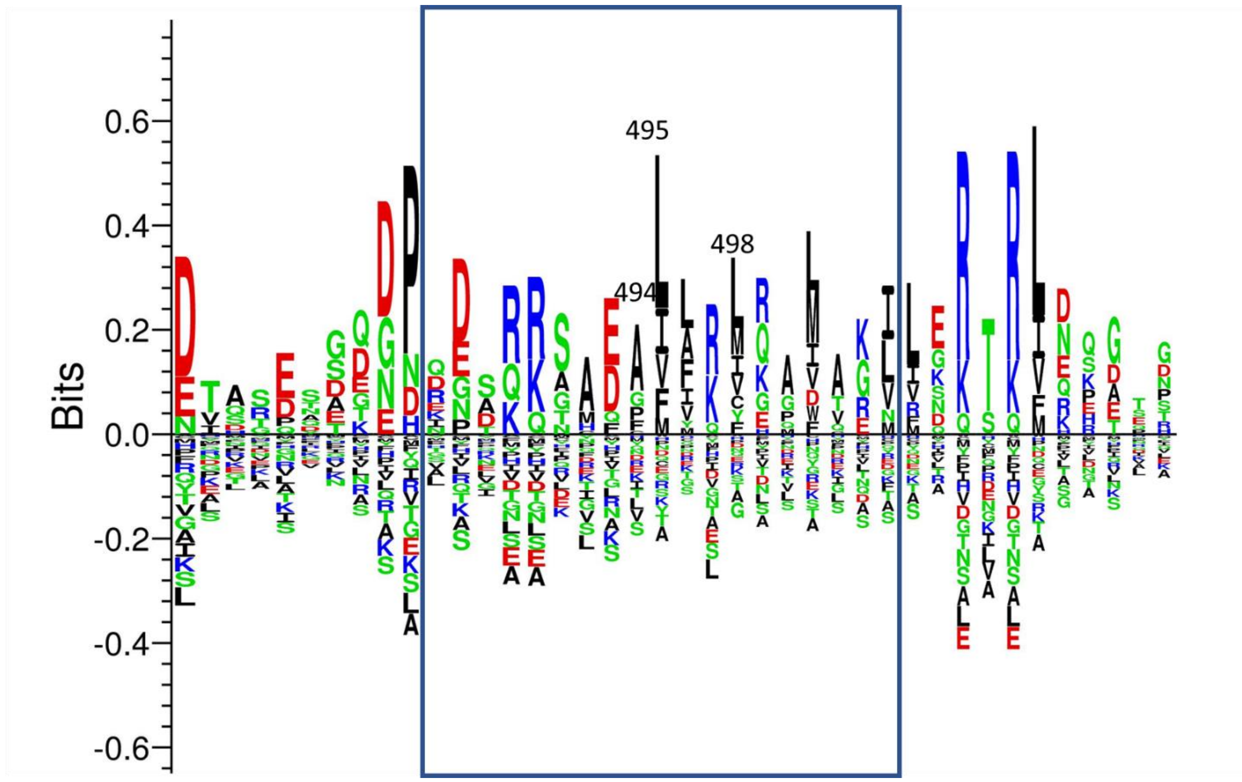


Figure 3-7. Sequence consensus of NTAIL (MeV). Residue 486-504 of NTAIL were encircled by square. The three residues investigated here were labelled with their residue numbers.

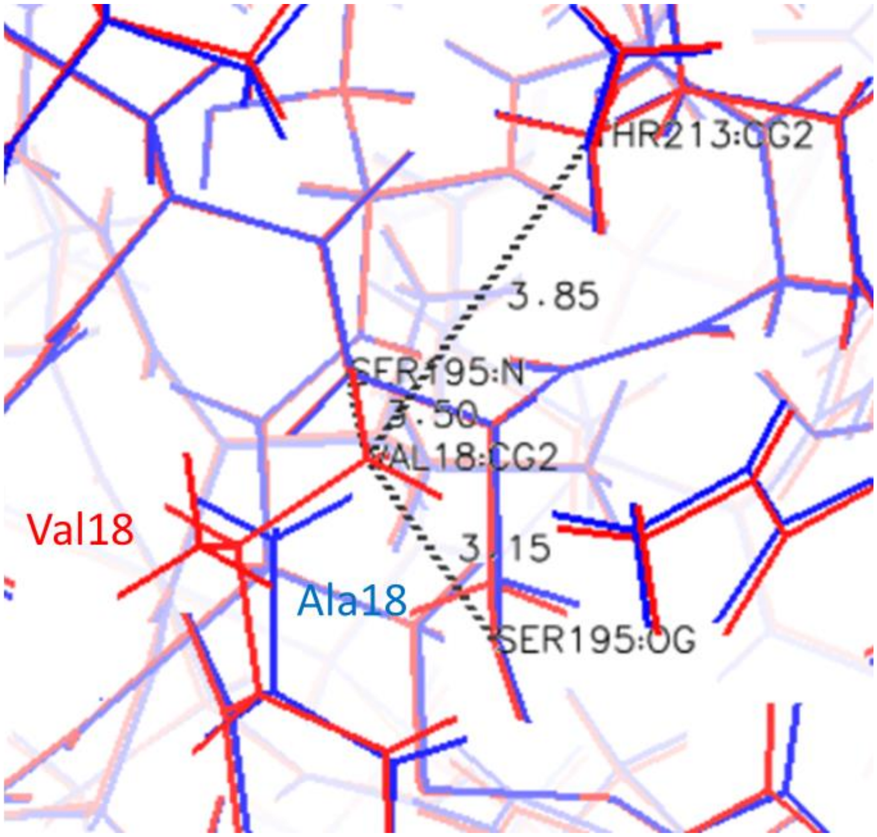


Figure 3-8. Structures of SGPB/OMTKY3-Val18 (red, PDB ID 1CT4) and Ala18 (blue, PDB ID 1SGP). Distances (unit in Å) between CG2 atom of Val18 and some of its surrounding heavy atoms are labelled. The structures were aligned using all backbone atoms excluding the backbone of Val18 and Ala18.

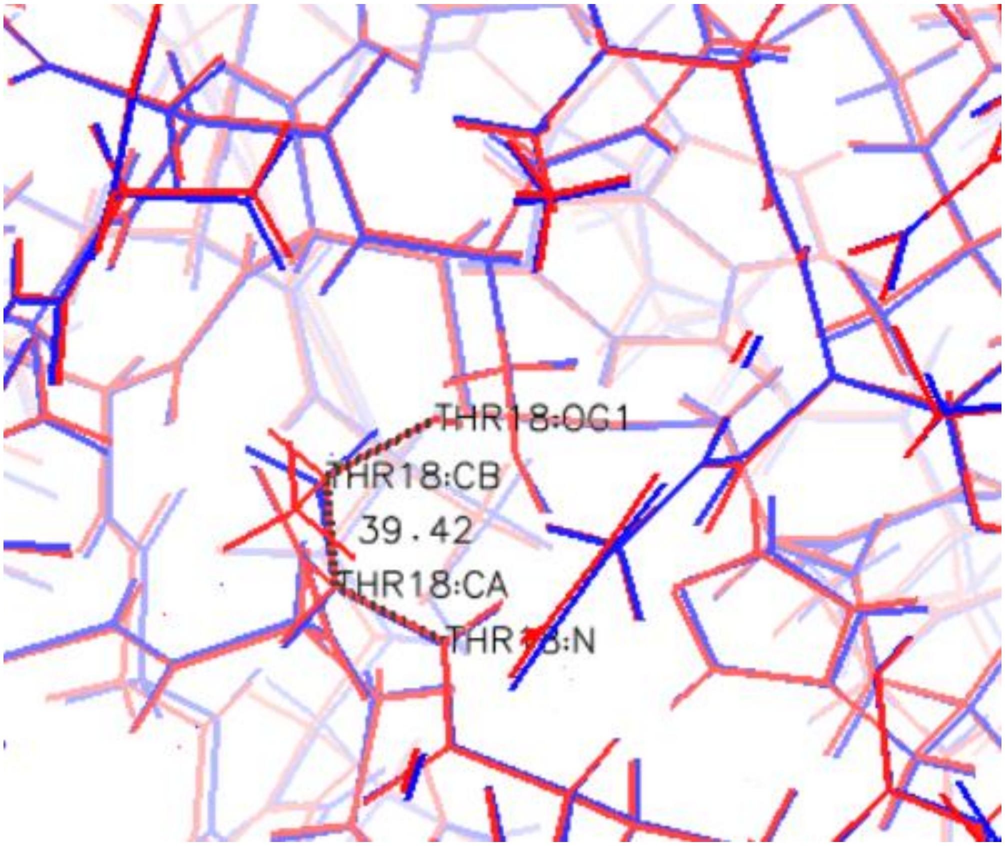


Figure 3-9. Structures of SGPB/OMTKY3-Thr18 (red, PDB ID 1CT2) and Ala18 (blue, PDB ID 1SGP). The χ1 dihedral angle (OG1-CB-CA-N) was labelled. The structures were aligned using all backbone atoms excluding the backbone of Thr18 and Ala18.

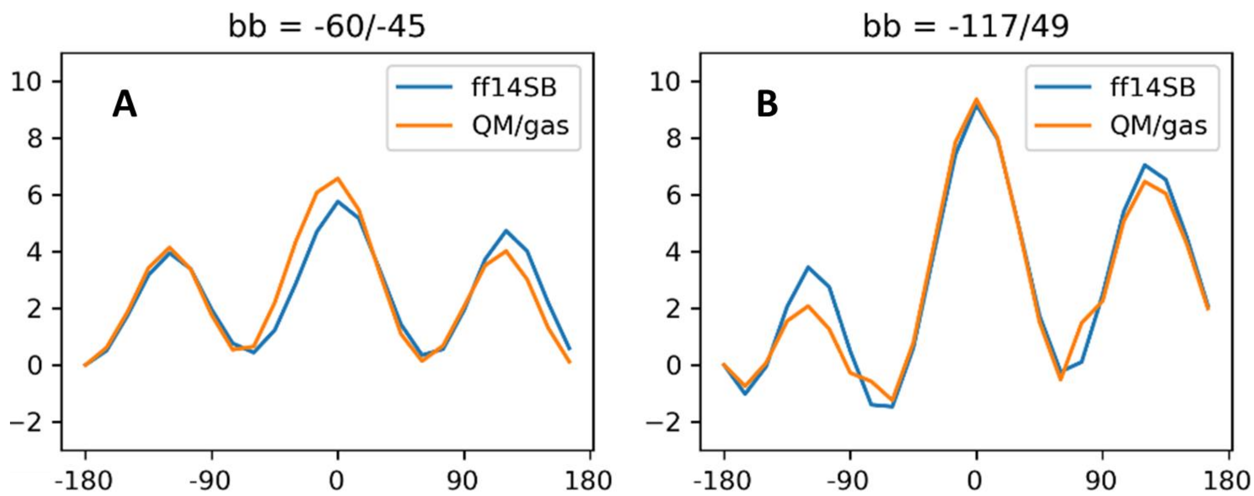


Figure 3-10. QM gas phase energy and ff14SB/implicit solvent energy scans on χ1 of Val in a dipeptide model. A) The φ/ψ of Val equals -60˚/-45˚ (α-helical). B) The φ/ψ of Val equals -117˚/49˚ (Val18 in PDB ID 1CT4). Note: the energies in A and B where zeroed to χ1=-180 separately, which means energies in A and B cannot be compared.

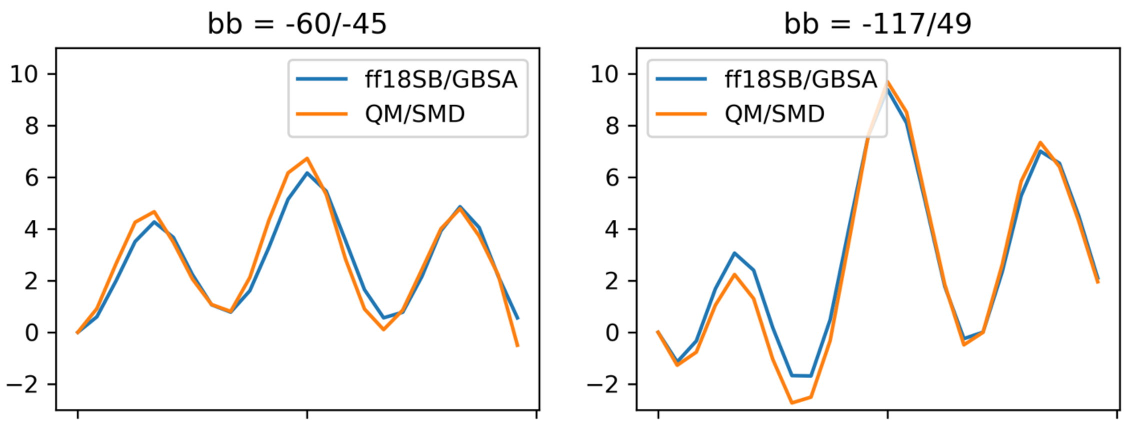


Figure 3-11. QM/implicit solvent energy and ff18SB/implicit solvent energy scans on χ1 of Val in a dipeptide model. A) The ϕ/ψ of Val equals -60˚/-45˚ (α-helical). B) The ϕ/ψ of Val equals -117˚/49˚ (Val18 in PDB ID 1CT4). Note: the energies in A and B where zeroed to χ1=-180 separately, which means energies in A and B cannot be compared.

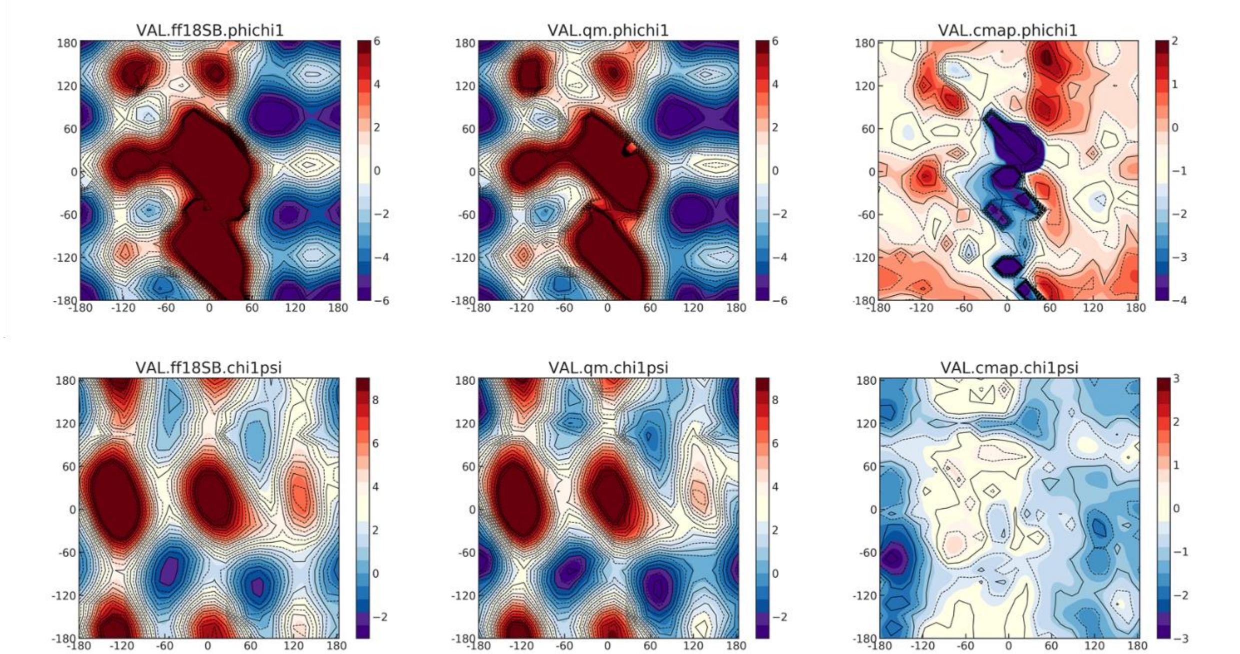


Figure 3-12. QM/implicit solvent energy and ff14SB/implicit solvent energy scans on χ1 of Val in a dipeptide model. Note that here, phi angle refers to CB-CA-N-C(i-1) and psi angle refers to CB-CA-C-N(i+1).

3.6 References

1. Romero P, *et al.* (2001) Sequence complexity of disordered protein. *Proteins* 42(1):38-48.
2. Das RK, Ruff KM, & Pappu RV (2015) Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr Opin Struct Biol* 32:102-112.
3. Oldfield CJ & Dunker AK (2014) Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* 83:553-584.
4. Dyson HJ & Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6(3):197-208.
5. Jensen MR, *et al.* (2011) Intrinsic disorder in measles virus nucleocapsids. *Proc Natl Acad Sci U S A* 108(24):9839-9844.
6. Dima RI & Thirumalai D (2004) Asymmetry in the shapes of folded and denatured states of proteins. *J Phys Chem B* 108(21):6564-6570.
7. Muller-Spath S, *et al.* (2010) Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc Natl Acad Sci U S A* 107(33):14609-14614.
8. Tran HT, Mao A, & Pappu RV (2008) Role of backbone-solvent interactions in determining conformational equilibria of intrinsically disordered proteins. *J Am Chem Soc* 130(23):7380-7392.
9. Brucale M, Schuler B, & Samori B (2014) Single-molecule studies of intrinsically disordered proteins. *Chem Rev* 114(6):3281-3317.
10. Warner JBt, *et al.* (2017) Monomeric huntingtin exon 1 has similar overall structural features for wild-type and pathological polyglutamine lengths. *J Am Chem Soc* 139(41):14456-14469.

11. Iesmantavicius V, *et al.* (2013) Modulation of the intrinsic helix propensity of an intrinsically disordered protein reveals long-range helix-helix interactions. *J Am Chem Soc* 135(27):10155-10163.

12. Meng W, Lyle N, Luan B, Raleigh DP, & Pappu RV (2013) Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc Natl Acad Sci U S A* 110(6):2123-2128.

13. Holehouse AS & Pappu RV (2018) Collapse transitions of proteins and the interplay among backbone, sidechain, and solvent interactions. *Annu Rev Biophys* 47:19-39.

14. Kohn JE, *et al.* (2004) Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc Natl Acad Sci U S A* 101(34):12491-12496.

15. Chan HS & Dill KA (1991) Polymer principles in protein structure and stability. *Annu Rev Biophys Biophys Chem* 20:447-490.

16. Roder H, Maki K, & Cheng H (2006) Early events in protein folding explored by rapid mixing methods. *Chem Rev* 106(5):1836-1861.

17. Borgia A, *et al.* (2016) Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J Am Chem Soc* 138(36):11714-11726.

18. Sherman E & Haran G (2006) Coil-globule transition in the denatured state of a small protein. *Proc Natl Acad Sci U S A* 103(31):11539-11543.

19. Cho JH & Raleigh DP (2005) Mutational analysis demonstrates that specific electrostatic interactions can play a key role in the denatured state ensemble of proteins. *J Mol Biol* 353(1):174-185.

20. Cho JH, *et al.* (2014) Energetically significant networks of coupled interactions within an unfolded protein. *Proc Natl Acad Sci U S A* 111(33):12079-12084.

21. Han H, Weinreb PH, & Lansbury PT, Jr. (1995) The core Alzheimer's peptide NAC forms amyloid fibrils which seed and are seeded by beta-amyloid: is NAC a common trigger or target in neurodegenerative disease? *Chem Biol* 2(3):163-169.

22. Bertoncini CW, *et al.* (2005) Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein. *Proc Natl Acad Sci U S A* 102(5):1430-1435.

23. Dang LX, Merz KM, & Kollman PA (1989) Free-energy calculations on protein stability - Thr-157-] Val-157 mutation of T4 Lysozyme. *J Am Chem Soc* 111(22):8505-8508.

24. Zou J, Song B, Simmerling C, & Raleigh D (2016) Experimental and computational analysis of protein stabilization by Gly-to-d-Ala substitution: A convolution of native state and unfolded state effects. *J Am Chem Soc* 138(48):15682-15689.

25. Wang LP, *et al.* (2017) Building a more predictive protein force field: A systematic and reproducible route to AMBER-FB15. *J Phys Chem B* 121(16):4023-4039.

26. Huang J, *et al.* (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* 14(1):71-73.

27. Robustelli P, Piana S, & Shaw DE (2018) Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci U S A* 115(21):E4758-E4766.

28. Lindorff-Larsen K, Piana S, Dror RO, & Shaw DE (2011) How fast-folding proteins fold. *Science* 334(6055):517-520.

29. Nguyen H, Maier J, Huang H, Perrone V, & Simmerling C (2014) Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J Am Chem Soc* 136(40):13959-13962.

30. Voelz VA, Bowman GR, Beauchamp K, & Pande VS (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J Am Chem Soc* 132(5):1526-1528.

31. Rauscher S*, et al.* (2015) Structural ensembles of intrinsically disordered proteins depend strongly on force field: A comparison to experiment. *J Chem Theory Comput* 11(11):5513-5524.

32. Bourhis JM*, et al.* (2004) The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Res* 99(2):157-167.

33. Longhi S*, et al.* (2003) The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J Biol Chem* 278(20):18638-18648.

34. Wang Y*, et al.* (2013) Multiscaled exploration of coupled folding and binding of an intrinsically disordered molecular recognition element in measles virus nucleoprotein. *Proc Natl Acad Sci U S A* 110(40):E3743-3752.

35. Gruet A*, et al.* (2016) Fuzzy regions in an intrinsically disordered protein impair protein-protein interactions. *FEBS J* 283(4):576-594.

36. Chen VB*, et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66(Pt 1):12-21.

37. D.A. Case IYB-S, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden,*, et al.* (2018) *AMBER 2018* (University of California, San Francisco).

38. Kirkwood JG (1935) Statistical mechanics of fluid mixtures. *J Chem Phys* 3(5):300-313.

39. Maier JA, *et al.* (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput.* 11(8):3696-3713.

40. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, & Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926-935.

41. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, & Haak JR (1984) Molecular-dynamics with coupling to an external bath. *J Chem Phys* 81(8):3684-3690.

42. Lee TS, Hu Y, Sherborne B, Guo Z, & York DM (2017) Toward fast and accurate binding affinity prediction with pmemdGTI: An efficient implementation of GPU-accelerated thermodynamic integration. *J Chem Theory Comput* 13(7):3077-3084.

43. Darden T, York D, & Pedersen L (1993) Particle mesh ewald - an N.Log(N) method for ewald sums in large systems. *J Chem Phys* 98(12):10089-10092.

44. Ryckaert JP, Ciccotti G, & Berendsen HJC (1977) Numerical-integration of cartesian equations of motion of a system with constraints - molecular-dynamics of N-alkanes. *J Comput Phys* 23(3):327-341.

45. Song J, Laskowski M, Jr., Qasim MA, & Markley JL (2003) NMR determination of pKa values for Asp, Glu, His, and Lys mutants at each variable contiguous enzyme-inhibitor contact position of the turkey ovomucoid third domain. *Biochemistry* 42(10):2847-2856.

46. Abul Qasim M, Ranjbar MR, Wynn R, Anderson S, & Laskowski M, Jr. (1995) Ionizable P1 residues in serine proteinase inhibitors undergo large pK shifts on complex formation. *J Biol Chem* 270(46):27419-27422.

47. Lu W, *et al.* (1997) Binding of amino acid side-chains to S1 cavities of serine proteinases. *J Mol Biol* 266(2):441-461.

48. Kingston RL, Hamel DJ, Gay LS, Dahlquist FW, & Matthews BW (2004) Structural basis for the attachment of a paramyxoviral polymerase to its template. *Proc Natl Acad Sci U S A* 101(22):8301-8306.

49. Guex N & Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18(15):2714-2723.

50. Kurzbach D*, et al.* (2014) Compensatory adaptations of structural dynamics in an intrinsically disordered protein complex. *Angew Chem Int Ed Engl* 53(15):3840-3843.

51. Nagulapalli M*, et al.* (2012) Recognition pliability is coupled to structural heterogeneity: a calmodulin intrinsically disordered binding region complex. *Structure* 20(3):522-533.

52. Beier A*, et al.* (2018) Modulation of correlated segment fluctuations in IDPs upon complex formation as an allosteric regulatory mechanism. *J Mol Biol* 430(16):2439-2452.

53. Tuttle LM*, et al.* (2018) Gcn4-mediator specificity is mediated by a large and dynamic fuzzy protein-protein complex. *Cell Rep* 22(12):3251-3264.

54. Bignon C, Troilo F, Gianni S, & Longhi S (2018) Modulation of measles virus NTAIL interactions through fuzziness and sequence features of disordered binding sites. *Biomolecules* 9(1).

55. Gely S*, et al.* (2010) Solution structure of the C-terminal X domain of the measles virus phosphoprotein and interaction with the intrinsically disordered C-terminal domain of the nucleoprotein. *J Mol Recognit* 23(5):435-447.

56. Belle V*, et al.* (2008) Mapping alpha-helical induced folding within the intrinsically disordered C-terminal domain of the measles virus nucleoprotein by site-directed spin-labeling EPR spectroscopy. *Proteins* 73(4):973-988.

57. Bischak CG, *et al.* (2010) Probing structural transitions in the intrinsically disordered C-terminal domain of the measles virus nucleoprotein by vibrational spectroscopy of cyanylated cysteines. *Biophys J* 99(5):1676-1683.

58. Morin B, *et al.* (2006) Assessing induced folding of an intrinsically disordered protein by site-directed spin-labeling electron paramagnetic resonance spectroscopy. *J Phys Chem B* 110(41):20596-20608.

59. Huang K, Lu W, Anderson S, Laskowski M, Jr., & James MN (1995) Water molecules participate in proteinase-inhibitor interactions: crystal structures of Leu18, Ala18, and Gly18 variants of turkey ovomucoid inhibitor third domain complexed with Streptomyces griseus proteinase B. *Protein Sci* 4(10):1985-1997.

60. Bateman KS, *et al.* (2000) Deleterious effects of beta-branched residues in the S1 specificity pocket of Streptomyces griseus proteinase B (SGPB): crystal structures of the turkey ovomucoid third domain variants Ile18I, Val18I, Thr18I, and Ser18I in complex with SGPB. *Protein Sci* 9(1):83-94.

61. Horn JR, Ramaswamy S, & Murphy KP (2003) Structure and energetics of protein-protein interactions: the role of conformational heterogeneity in OMTKY3 binding to serine proteases. *J Mol Biol* 331(2):497-508.

62. Kaus JW, Pierce LT, Walker RC, & McCammont JA (2013) Improving the efficiency of free energy calculations in the Amber molecular dynamics package. *J Chem Theory Comput* 9(9).

63. Robertson MJ, Tirado-Rives J, & Jorgensen WL (2015) Improved peptide and protein torsional energetics with the OPLSAA force field. *J Chem Theory Comput* 11(7):3499-3509.

64. Oroguchi T & Nakasako M (2017) Influences of lone-pair electrons on directionality of hydrogen bonds formed by hydrophilic amino acid side chains in molecular dynamics simulation. *Sci Rep* 7(1):15859.

65. Ren P, Wu C, & Ponder JW (2011) Polarizable atomic multipole-based molecular mechanics for organic molecules. *J Chem Theory Comput* 7(10):3143-3161.

66. Blocquel D, *et al.* (2012) Interaction between the C-terminal domains of measles virus nucleoprotein and phosphoprotein: a tight complex implying one binding site. *Protein Sci* 21(10):1577-1585.

67. Yegambaram K & Kingston RL (2010) The feet of the measles virus polymerase bind the viral nucleocapsid protein at a single site. *Protein Sci* 19(4):893-899.

68. Bonetti D, *et al.* (2017) Analyzing the Folding and Binding Steps of an Intrinsically Disordered Protein by Protein Engineering. *Biochemistry* 56(29):3780-3786.

69. Kaminski GA, Friesner RA, Tirado-Rives J, & Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* 105(28):6474-6487.

70. Teilum K, Olsen JG, & Kragelund BB (2015) Globular and disordered-the non-identical twins in protein-protein interactions. *Front Mol Biosci* 2:40.

71. Meszaros B, Tompa P, Simon I, & Dosztanyi Z (2007) Molecular principles of the interactions of disordered proteins. *J Mol Biol* 372(2):549-561.

72. Wong ET, Na D, & Gsponer J (2013) On the importance of polar interactions for complexes containing intrinsically disordered proteins. *PLoS Comput Biol* 9(8):e1003192.

73. Chodera JD, *et al.* (2011) Alchemical free energy methods for drug discovery: progress and challenges. *Curr Opin Struct Biol* 21(2):150-160.

74. Maurer M, de Beer SB, & Oostenbrink C (2016) Calculation of relative binding free energy in the water-filled active site of oligopeptide-binding protein A. *Molecules* 21(4):499.

75. Deng Y & Roux B (2008) Computation of binding free energy with molecular dynamics and grand canonical Monte Carlo simulations. *J Chem Phys* 128(11):115103.

76. Luccarelli J, Michel J, Tirado-Rives J, & Jorgensen WL (2010) Effects of water placement on predictions of binding affinities for p38alpha MAP kinase Inhibitors. *J Chem Theory Comput* 6(12):3850-3856.

77. Ben-Shalom IY, Lin C, Kurtzman T, Walker RC, & Gilson MK (2019) Simulating water exchange to buried binding sites. *J Chem Theory Comput*.

78. Klein-Seetharaman J*, et al.* (2002) Long-range interactions within a nonnative protein. *Science* 295(5560):1719-1722.

79. Brunel J*, et al.* (2014) Sequence of events in measles virus replication: role of phosphoprotein-nucleocapsid interactions. *J Virol* 88(18):10851-10863.

80. Bloyet LM*, et al.* (2016) Modulation of re-initiation of Measles virus transcription at intergenic regions by PXD to NTAIL binding strength. *PLoS Pathog* 12(12):e1006058.

81. Wright PE & Dyson HJ (2009) Linking folding and binding. *Curr Opin Struct Biol* 19(1):31-38.

82. Dyson HJ & Wright PE (2002) Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* 12(1):54-60.

83. Lee TW, Qasim MA, Laskowski M, Jr., & James MN (2007) Structural insights into the non-additivity effects in the sequence-to-reactivity algorithm for serine peptidases and their inhibitors. *J Mol Biol* 367(2):527-546.

84. Debiec KT, *et al.* (2016) Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *Journal of Chemical Theory and Computation* 12(8):3926-3947.

85. Shi Y, *et al.* (2013) The Polarizable Atomic Multipole-based AMOEBA Force Field for Proteins. *J Chem Theory Comput* 9(9):4046-4063.

4. Molecular Basis of Roughness on The Free Energy Landscape for Protein Folding; Experimental and Computational Studies of a Non-Native Interactions in The Denatured State of a Fast Folding Protein.

Abstracts

Proteins fold on relatively smooth free energy landscapes which are biased towards the native state, but even simple topologies which fold rapidly, in an apparent two-state fashion can experience roughness on their free energy landscape. The atomistic details of the interactions which lead to roughness are difficult to address experimentally. Closely related to the problem of deciphering the details of the free energy landscape is the problem of defining the interactions in the denatured state ensemble (DSE) which is populated under native conditions, i.e. under conditions where the native state is stable. The DSE of many proteins clearly deviates from classic random coil models, but quantifying and defining the transiently populated interactions in the ensemble is extremely challenging. Here we use experimental thermodynamic and pKa measurements in combination with computational thermodynamic integration to show that non-native sidechain interactions can stabilize native backbone structure in the DSE of the Villin headpiece helical domain. A non-native salt bridge stabilizes native like backbone structure in the DSE. The salt bridge is broken during the transition to the native state and the sidechains transition to form a tertiarry interaction in the folded state.

Acknowledgements

I gratefully acknowledge Koushik Kasavajhala, Chuan Tian and Kellon Belfon for their administration of computational resources.

4.1 Introduction

The villin headpiece subdomain (HP36), a 36-residue (or 35-residue depending on the construct used) protein, is one of the smallest naturally occurring proteins that folds cooperatively (**Fig 4-1.**) (1). The folding rate of HP36 approaches the theoretical limit for folding. Its small size, simple three-helix topology and rapid folding have made HP36 a very popular protein model for understanding protein thermodynamics and folding kinetics (2-9). The structure of HP36 is constructed of three α-helices. Three Phe residues make significant contribution to the hydrophobic core of the folded state of HP36 (10). Proteins typically folded on smooth free energy landscapes which are biased towards the native state. An atomic level characterization of interactions that contribute to roughness on the free energy landscape is challenging. A closely related, and also challenging, issue is the characterization of interactions formed in the DSE populated under native conditions. Here we characterized non-native interactions in the DSE of HP36 that involve the acidic residue.

Extensive experimental and computational studies have been conducted on the DSE of HP36 and indicate that native-like structure exists in the DSE (4, 7, 8, 11-15). HP21, constructed from the helix 1 and helix 2 of HP36, has been used as a model for studying the DSE of HP36. HP21 lacks tertiary structures but has considerable helical content and nativelike structures (11, 12, 15). The nuclear Overhauser effect spectroscopy on HP21 found both native and non-native contacts between residues, which indicates that the native structure in the HP21 can be stabilized by non-native interactions as well (11, 12). Temperature-jump infrared spectroscopy, in combination with site-specific labels, indicates that helix 3 is the most stable followed by helix 1 and helix 2 is the least stable in the full length HP36 (16). In several computational studies, residual native and non-native structures were found in the DSE of HP36 (4, 7, 8, 14), including non-native salt bridge between the D44 and K48 in the denatured state of HP36 that was observed in one of the computational studies (4). Experimental evidence for the interactions observed in MD simulations is scarce.

The pKa of titratable residues reflects the electrostatic interactions they experienced under the circumstance they are measured (17, 18). The titratable residues may experience either favorable or unfavorable electrostatic interactions in the denatured state, if the pKa values significantly deviate from the pKa of model compound or the pKa measured in peptide fragment. The pKa of titratable residues in the denatured state of various proteins have been studied by using either direct NMR measurements on the denatured state of proteins (19-21) or the Tanford-Wyman linkage relationship (22-24).

In a previous study, mutagenesis and pKa measurement indicate that K48 and E45 form a salt bridge in the native state of HP36 (6) (**Fig. 4-1**). Note the notation used here corresponds to the numbering of this region in full villin head piece plus an additional Met. Thus, the first residue is designated as M41 and the second residue is L42. This notation is adopted to facilitate comparison with earlier work. However, K48M mutation increases the stability of HP36 which was proposed to be caused by the disruption of favorable interactions involving K48 in the denatured state of HP36 (6), but the origin of these effects is not known. Here, we show D44 has a suppressed pKa in the denatured state of HP36 indicating that D44 is also involved in favorable electrostatic interactions in the DSE of HP36. A double mutant thermodynamic cycle analysis combined with pKa analysis and alchemical free energy calculations were conducted and showed that D44 and K48 make favorable electrostatic interactions in the denatured state of HP36. However, these residues do not form a salt bridge in the native state of HP36, rather D44 forms a tertiary salt bridge with R55 and K48 forms a {i to i+3} salt bridge with E45 (6). This study reveals that non-native salt bridges with significant strength can exist in the DSE and illustrate how they are compatible with native-like backbone structure.

4.2 Methods

4.2.1 Protein expression and purification

HP36 wildtype and mutants were expressed as a fusion protein with NTL9 as described (25). The factor Xa cleavage was carried out at 23 ^{o}C for 16-20 h. The proteins were purified by ion exchange chromatography and reverse-phase HPLC using a gradient of 30-65% buffer B in 70 minutes. The identity was confirmed by matrix-assisted laser desorption and ionization time-of-flight mass spectrometry (MALDI).

4.2.2 Protein stability measurements

Protein stability was measured by CD monitored urea and thermal denaturation experiments. Urea induced unfolding was performed at 25 ^{o}C and 222 nm with samples of 15-30 μM protein in 10 mM sodium acetate and 150 mM sodium chloride at pH 6.0 on an AVIV 202SF spectrophotometer. The concentration of urea was increased from 0 to about 10 M in ~0.25 M steps. Urea concentrations were determined by measuring the refractive index. Urea unfolding curves were analyzed by a non-linear least squares fit. Thermal unfolding experiments were performed on an Applied Photophysics Chirascan CD instrument at 222 nm over the range of 2 ^{o}C to 94 ^{o}C in 2 ^{o}C intervals. The buffer and protein concentration were the same as used in the urea denaturation experiments. Measurements were made at pH 3.0 and 6.0. The reversibility of unfolding was confirmed by comparing the initial CD signal at the start of the run to the signal measured after the run was completed and the sample was cooled to the starting temperature.

4.2.3 Protein pKa measurement

The pKa values of Asp, Glu and C-terminus in the HP36D44N and HP36D44NK48M mutant were measured by using NMR. The chemical shifts of the Hβ and Hγ protons were collected. The Hα chemical shifts of the C-terminus residue, F76, were used to measure the pKa of the C-terminus. The chemical shift data collected over pH=2 to pH=7 were fit to the Henderson–Hasselbalch equation to yield the pKa values.

4.2.4 Unfolded state pKa calculations

The method has been illustrated by J. Shen (22). The unfolding free energy as a function of the pKa values in the folded and unfolded state and the pH can be expressed as:

$$\Delta\Delta G = \Delta G^{pH1} - \Delta G^{pH2} = RT\sum_i ln\frac{(1+10^{(pK_a^F(i)-pH2)})(1+10^{(pK_a^U(i)-pH1)})}{(1+10^{(pK_a^U(i)-pH2)})(1+10^{(pK_a^F(i)-pH1)})} \quad (17)$$

where ΔG^{pH1} and ΔG^{pH2} are the unfolding free energy measured at pH1 and pH2 respectively. For this study, pH1 = 6.0 and pH2 = 3.0 were chosen since the thermodynamic data was measured at these two pH (**Table 4-1**). i represents all titratable residues in wildtype HP36 and mutants.

If we assume the mutation has negligible effect on pKa of other titratable residues in the denatured state.

$$\Delta\Delta G^{WT} - \Delta\Delta G^{MU} = \Delta G^{pH1,WT} - \Delta G^{pH2,WT} - (\Delta G^{pH1,MU} - \Delta G^{pH2,MU})$$

$$= RTln\frac{\left(1+10^{\left(pK_a^{F,WT}(j)-pH2\right)}\right)\left(1+10^{\left(pK_a^{U,WT}(j)-pH1\right)}\right)}{\left(1+10^{\left(pK_a^{U,WT}(j)-pH2\right)}\right)\left(1+10^{\left(pK_a^{F,WT}(j)-pH1\right)}\right)} + RT\sum_{i\neq j} ln\frac{\left(1+10^{\left(pK_a^{F,WT}(i)-pH2\right)}\right)}{\left(1+10^{\left(pK_a^{F,WT}(i)-pH1\right)}\right)} -$$

$$RT\sum_{i\neq j} ln\frac{\left(1+10^{\left(pK_a^{F,MU}(i)-pH2\right)}\right)}{\left(1+10^{\left(pK_a^{F,MU}(i)-pH1\right)}\right)} \quad (18)$$

where j is the residue of interest. WT and MU stand for wildtype HP36 and HP36 mutant respectively. The mutant has a neutral analog in the place of the residue j. For example, HP36D44N was used as the mutant to determine the pKa of D44 in the denatured state of HP36 using this method. The latter two terms on the right side of equation 18 will be 0, if we assume that the mutation has negligible effect on the pKa of other residues in the native state.

4.2.5 Free Energy Calculations for the Asp44-to-Asn44 and K48-to-M48 mutations in HP36

The pdb code 1YRF was used for the structure of HP36. Hydrogen atoms were added using the MolProbity program (26). Side chain rotamer states for ASN/GLN were corrected based on suggestion provided by MolProbity. LYS, ARG side chains and N-termini were set to be protonated and ASP, GLU side chains and C-termini were set to be unprotonated. Waters present in X-ray structures were kept while all salt ions were deleted. Truncated octahedron boxes were used to solvate the proteins. Free energy calculations were performed using non-softcore thermodynamic integration (TI) implemented in Amber (27, 28). The Amber force field ff14SB and the TIP3P water model were used for the TI calculations (29, 30). Minimization and equilibration under constant pressure(31) were conducted to heat up and relax the X-ray structures. Production runs were conducted under constant pressure (31). Energy minimization was conducted using gradient descent algorithm with 100 kcal/mol position restraints on all heavy atoms of proteins. The maximal number of cycles is 10000. A 0.1 ns constant volume MD simulation was then conducted to slowly heat up the structures from 150K to 298K with 100 kcal/mol position restraints on all heavy atoms of proteins. A 0.1 ns constant pressure MD simulation was then conducted with 100 kcal/mol position restraints on all heavy atoms of proteins. A 0.25 ns constant pressure MD simulation was conducted with 10 kcal/mol position restraints on all heavy atoms of proteins. A 0.1 ns constant pressure MD simulation was conducted with 10 kcal/mol position restraints on all CA,C and N atoms. A 0.1 ns constant pressure MD simulation was conducted with 1 kcal/mol position restraints on all CA,C and N atoms. A 0.1 ns constant pressure MD simulation was conducted with 0.1 kcal/mol position restraints on all CA, C and N atoms. The mutation site (residue 44 or 48) was excluded in the restraints. In the last step 0.25 ns constant pressure MD simulation was conducted with no restraints. An 1 fs step size was used for the equilibration. The temperature was set to 298K and no salt ions was included. Langevin dynamics was used to control temperature and the collision frequency was set to be 1.0 ps^{-1}. The pressure relaxation time was set to 0.1 ns. Particle mesh Ewald methods were used to calculate electrostatic energies (32). Hydrogen atoms were constrained using the SHAKE algorithm (33). The cutoff of non-bonded interactions was set to 8 Å. A timestep of 2fs was used for the production runs. The simulation time length for each λ window is 4ns. The trapezoidal rule was used for the integration of all λ windows.

Mutation of N-to-D was conducted in one transition step with equally spaced λ from 0 to 1 with an interval of 0.1. Mutation of K-to-M was conducted in two transition steps. In the first step, the charges were changed with equally spaced λ from 0 to 1 with an interval of 0.1. In the second step, the VDW interactions were changed with a series of λ = 0.00922, 0.04794, 0.11505, 0.20634, 0.31608, 0.43738, 0.56262, 0.68392, 0.79366 0.88495, 0.95206 and 0.99078. Amber library files were built for Asp and Met with dummy atoms to match the disappearing atoms in Asn and Lys.

4.2.6 Propagation of errors

The accumulation of errors is calculated using the following equation.

$$\Delta = \sqrt{\left(uncertainty\ of\ ①\right)^2 + \left(uncertainty\ of\ ②\right)^2 + \left(uncertainty\ of\ ③\right)^2 + \left(uncertainty\ of\ ④\right)^2} \tag{19}$$

4.3 Result

4.3.1 D44 has a suppressed pKa in the denatured state of HP36.

To confirm that K48 has favorable interactions with D44 in the DSE of HP36, the pKa of D44 in the DSE was estimated. The pKa values of acidic residues in the denatured state of HP36 were estimated following the procedure described in the methods section. The stability measurements for HP36WT and its mutant were conducted at pH = 3.0 and pH = 6.0 respectively. The previous measured native state pKa values indicate that the acid residues in HP36 are partially protonated at pH = 3.0 and are fully deprotonated at pH = 6.0 in the native state (6). This ensures that the stability difference of HP36 at pH=3.0 and pH=6.0 includes the protonation free energies contributed by the acidic residues. The stabilities were measured using thermal unfolding experiments at pH = 3.0 since lowering the pH value in the presence of urea requires addition of large amounts of acid due to the protonation of urea and hence involves a change in ionic strength. At pH = 6.0, the stabilities were measured by using urea unfolding experiments. The thermodynamic data are listed in **Table** 4-**1**.

Since the basic residues and the N-terminus of HP36 are expected to have pKa values much higher than 6.0, only acidic residues need to be considered in equation 18. The pKa values of the acidic residues and the C-terminus of wildtype HP36 in the native state are listed in **Table 4-2**. The pKa in the denatured state of HP36 were first estimated by assuming that the mutation of residue j has no effect on the pKa of other residues in both the native and denatured states (**Table 4-3)**. To provide a reference, these pKa values were compared to the pKa values for these residues in a set of peptide fragments. E45 has a similar DSE pKa value compared to its counterpart in the peptide fragment(6). D46 and E72 have DSE pKa values 0.23 and 0.26 higher than their counterparts in the peptide fragments respectively. However, the pKa of D44 in the DSE is 0.42 lower than that of the peptide fragment, which indicates that D44 makes favorable electrostatic interactions in the denatured state of HP36, beyond those which are captured in the fragment peptides.

The pKa of E45, D46, E72 and the carboxyl group of C-terminus of HP36 D44N were measured to examine the assumption that the D44-to-N44 mutation has no effect on the pKa of other acidic residues in the native state of HP36. The results are listed in the **Table 4-2**. For completion, the pKa of E45, D46, E72 and carboxyl group of HP36K48M and HP36D44NK48M are also listed. These experiments indicate that the D44N mutation does not significantly alter the native state pKa's of other acidic residues. The largest shift is only 0.09 pKa units. Notably, a K48M mutant only perturb the pKa of E45 in the native state, which is consistent with the formation of a K48-E45 salt bridge. We repeated the calculations and using native state pKa's determined for HP36D44N (**Table 4-2**) and obtained DSE pKa of 3.65 for D44, which is very close to the value of 3.58 deduced using the simpler method. This indicates that D44 experiences favorable electrostatic interactions in the denatured state.

4.3.2 Double mutant cycle analysis indicates that there are favorable interactions between D44 and K48 in the unfolded state

K48 has been suggested to make favorable interactions in the denatured state of HP36 (6) and D44 has a suppressed DSE pKa, hence it is likely that they make favorable interactions with each other in the denatured state. In order to determine whether they form a favorable interaction in the denatured state as well as to estimate the strength of the interaction, a double mutant thermodynamic cycle analysis was designed (**Fig. 4-2**).

In these thermodynamic cycles, ⑥-⑤=②-① and ⑦-⑧=④-③. These two equations lead to ⑥-⑦-(⑤-⑧) = ②-①-④+③ in which ①, ②, ③ and ④ are the standard unfolding free energy listed in **Table 4-4**. If any non-electrostatic perturbations are negligible in this thermodynamic cycle, ⑤-⑧ and ⑥-⑦ can be interpreted as the electrostatic interaction strength between D44 and K48 in the native state and in the denatured state respectively. Asn and Met are neutral analogs of Asp and Lys respectively which are expected to cause little perturbations to the structure and energetics of HP36 in the native and denatured state besides the loss of electrostatic interactions. However, it is possible that N44 and M48 introduce unexpected non-electrostatic perturbations into HP36.

Urea unfolding experiments were conducted to obtain the unfolding free energies of HP36WT (①), HP36D44N (②), HP36K48M (③) and HP36D44NK48M (④) (**Table 4-4)**. This leads to ⑥-⑦-(⑤-⑧) = ②-①-④+③ = -0.75 ± 0.17 kcal mol^{-1}. The uncertainty is calculated from propagation of errors (method).

This indicates that the interaction of D44 and K48 is -0.75± 0.17 kcal/mol more favorable in the denatured state than in the native state. This value is a net interaction strength from both the native state and the denatured state and need to be deconvoluted to get the interaction strength in the denatured state.

In order to estimate the interaction strength between D44 and K48 in the denatured state (⑥-⑦), the interaction strength between D44 and K48 (⑤-⑧) in the native state was estimated by using:

$$⑤\text{-}⑧ = 2.303RT\ \Delta pKa \tag{20}$$

Where ΔpKa is the pKa shift of D44 in the native state caused by K48 to M48 mutations. By using the pKa shift published in our previous study (6), the electrostatic interaction strength between D44 and K48 in the native state was calculated to be -0.27 kcal/mol. Thus, if the non-electrostatic perturbations caused by the mutations are negligible, D44 and K48 have a favorable electrostatic interaction of -1.02 kcal/mol in the denatured state of HP36. This value is obtained from (⑤-⑧)-(②-①-④+③). Note that the value of ②-①-④+③ obtained from the double mutants cycle also includes any non-electrostatic perturbations caused by the D44-to-N44 and K48-to-M48 mutations. Moreover, an error of 0.1 unit of pKa translates to an error of 0.14 kcal/mol in free energy at 25 °C. Thus, we used alchemical free energy calculations to independently investigate the effect of the D44-to-N44 and K48-to-M48 mutations in the native state, and to estimate any non-electrostatic contribution that may need to be deconvoluted to more accurately estimate the DSE interactions.

4.3.3 Computational study confirmed the favorable electrostatic interaction of D44 and K48 in the denatured state

In contrast to the estimation using ΔpKa, which ignores any non-electrostatic perturbation caused by the mutations and only considers the electrostatic interactions between D44 and K48, ⑤-⑧ calculated by free energy calculations includes all interactions introduced by the mutations. Two thermodynamic integration (TI) calculations of D44-to-N44 mutations were conducted to give estimates of ⑤-⑧ (**Fig. 4-3**). One in the context of HP36WT and the other one in the context of HP36K48M. The results show that ⑤-⑧ = -0.23 ± 0.21 kcal/mol. This leads to a refined estimate of ⑥-⑦ = -0.52 ± 0.27 kcal/mol. ⑤-⑧ can also be obtained by conducting TI calculations of K48-to-M48 mutations in the context of HP36WT and HP36D44N (**Fig. 4-3**). TI calculations of K48-to-M48 mutations gave a value of ⑤-⑧ of 0.06 ± 0.10 kcal/mol, which leads to ⑥-⑦ = -

0.69 ± 0.20 kcal/mol. Note that ⑥-⑦ includes any interactions newly formed by N44 and M48 as well as contributions from the removal of interactions between D44 and K48 in the DSE of HP36. If N44 and M48 are assumed to cause little to no perturbation to the DSE besides their effect on electrostatics, the electrostatic interaction strength between D44 and K48 in the denatured state is -0.69 ± 0.20 to -0.52 ± 0.27 kcal/mol.

4.4 Conclusion

In this study, the pKa of D44 was found to be suppressed in the denatured state of HP36. Considering that K48 was found to have favorable electrostatic interactions in the denatured state of HP36, a salt bridge between D44 and K48 in the denatured state was proposed here. A double mutant cycle showed that D44 and K48 have an electrostatic interaction in the denatured state that is 0.75 kcal/mol more favorable than the electrostatic interaction between D44 and K48 in the native state of HP36. The pKa measurements of D44 in both HP36WT and HP36K48M showed that the electrostatic interactions between D44 and K48 is about -0.27 kcal/mol in the native state. The alchemical free energy calculations, which also consider non-electrostatic perturbations caused by mutations, indicate that the strength of interactions between D44 and K48 is 0.06~0.23 kcal/mol in the native state. Thus, the predicted interaction strength between D44 and K48 in the denatured state of HP36 is 0.52 to 1.02 kcal/mol.

The direct pKa measurements showed pKa of D44 is not suppressed in the fragment peptide (residue 41-53) of HP36 (6). An isolated helix 1 of HP36 is largely unstructured as indicated by an NMR study, however, the helix 1 is significantly structured in a peptide containing both helix 1 and helix 2 of HP36 due to tertiary contacts (15). We proposed that the salt bridge between D44 and K48 is formed due to residual helical conformations in the DSE of full-length HP36 which stabilize the non-native interactions between D44 and K48.

In principle, a direct modelling of the DSE of HP36 using MD simulation can also probe the interaction between D44 and K48 in the DSE. However, such simulations with explicit waters are impractical due to high computational cost. The thermodynamic cycles in **Fig. 4-3**, which combined both the computationally and experimentally measured free energies, provide a physically rigorous detour to probe interactions in the DSE without the direct modelling of the DSE.

Table 4-1. Thermodynamic parameters for the unfolding of HP36 wildtype and the mutants at pH=3.0 and pH=6.0.

Protein	pH 6.0	pH 3.0		
	ΔG^o (urea) (kcal mol^{-1})	T_m (^oC)	$\Delta H^o(T_m)$ (kcal mol^{-1})	ΔG^o (ΔCp=0.38) (kcal mol^{-1})
HP36 WT	3.39±0.06	52.8±0.2	25.4±0.5	1.70
D44N	2.93±0.12	51.3±0.1	26.5±0.4	1.73
E45Q	2.99±0.11	52.7±0.1	27.8±0.3	1.90
D46N	2.51±0.09	49.7±0.1	25.9±0.4	1.61
E72Q	3.28±0.10	50.4±0.1	25.4±0.3	1.60

The uncertainties are the standard errors to the fit. The ΔCp used in the fitting of thermal unfolding curve was previously determined.(25)

Table 4-2. pKa of acidic residues in the native state of wildtype HP36 and HP36 mutants

Protein	D44	E45	D46	E72	C-terminus
HP36WT(6)	3.04±0.12	3.95±0.02	3.44±0.11	4.37±0.03	2.91±0.08
HP36D44N	N/A	3.86±0.02	3.43±0.05	4.32±0.02	2.85±0.10
HP36K48M(6)	3.23±0.07	4.68±0.09	3.38±0.05	4.41±0.03	2.90±0.06
HP36D44NK48M	N/A	4.39±0.05	3.49±0.08	4.37±0.03	2.96±0.09

pKa values of HP36WT and HP36K48M were taken from reference (6).

Table 4-3. Estimated pKa of acidic residues in the denatured state HP36 and measured pKa of acidic residues in the peptide fragments. The fragments were residue 41 to 53 for D44, E45, D46 and residue 70 to 76 for E72.

residue	estimated wildtype denatured state $pK_a^{U,WT}(j)$	pKa in peptide fragment(6)
D44	3.58	4.00
E45	4.42	4.50
D46	4.12	3.86
E72	4.38	4.15

The pKa of the residues in peptide fragments were taken from reference (6).

Table 4-4. Thermodynamic parameters for the unfolding of HP36 WT and the mutants.

Protein	ΔG^o (kcal mol^{-1})	m-value (kcal mol^{-1} K^{-1})	C_M (M)
HP36 WT	3.39 ± 0.06	0.52 ± 0.01	6.5
D44N	2.93 ± 0.12	0.63 ± 0.02	4.7
K48M	4.26 ± 0.07	0.52 ± 0.01	8.2
D44NK48M	3.05 ± 0.07	0.49 ± 0.02	6.3

The uncertainties are the standard error to the fits. Experiments were performed at 25 ^oC with 10 mM sodium acetate and 150 mM sodium chloride, pH 6.0.

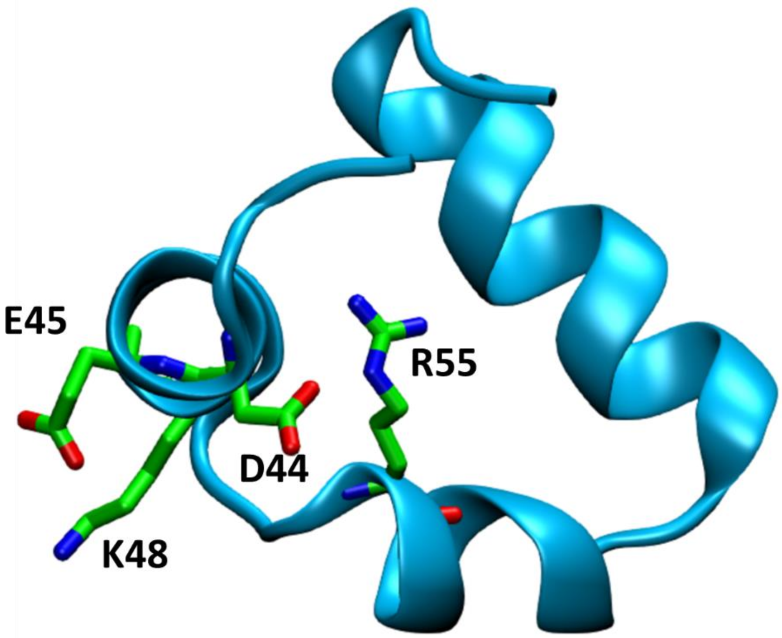


Figure 4-1. Cartoon structure of HP36 (pdb code 1yrf). D44, E45, K48 and R55 are shown as licorice.

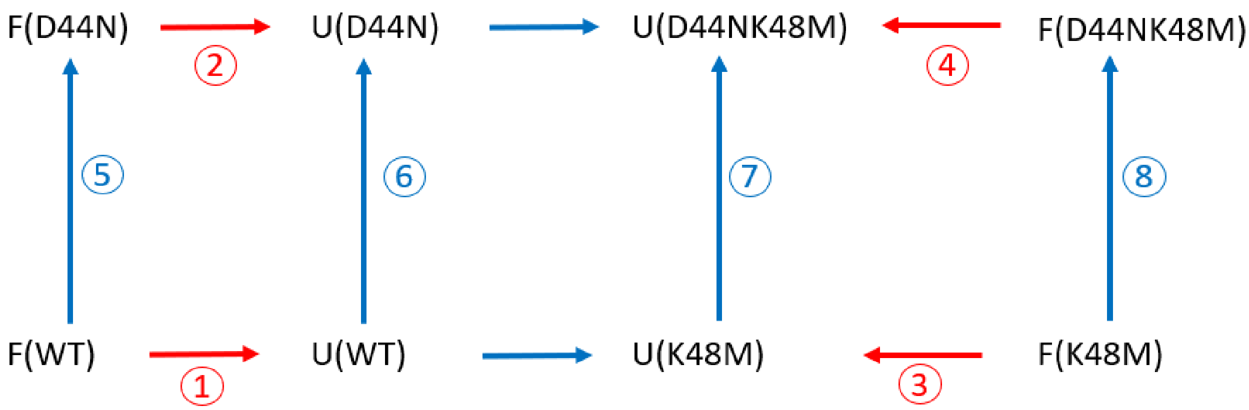


Figure 4-2. Thermodynamic cycles for the unfolding of wildtype HP36 and its mutants, HP36D44N, HP36K48M and HP36D44NK48M. The red arrows are the unfolding processes of HP36 and its mutants. Blue arrows are the mutation processes in the context of native/folded state (F) and denatured/unfolded state (U). Circled numbers represent the free energy associated with the processes. The values denoted in red can be experimentally measured; those in blue cannot.

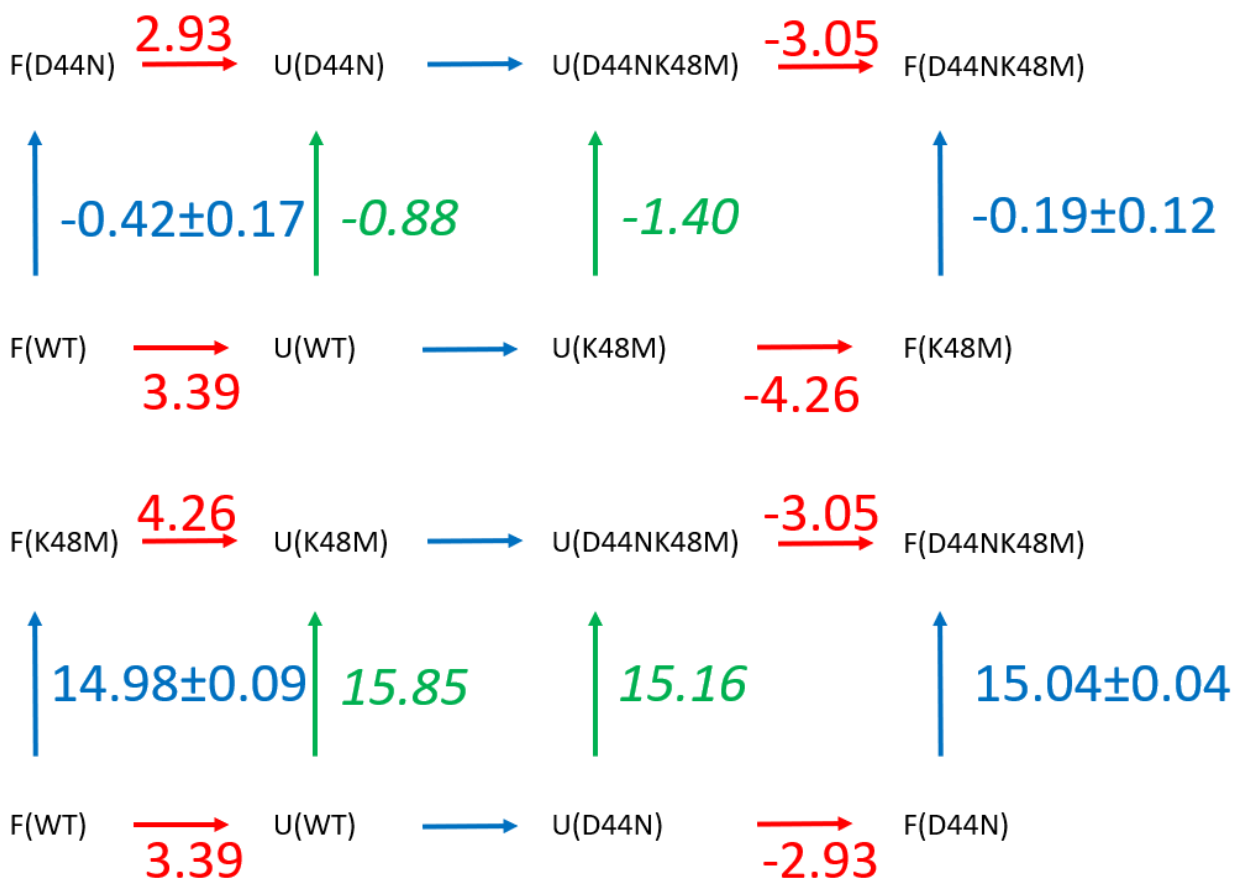


Figure 4-3. Thermodynamic cycles for the transitions among HP36WT, HP36D44N, HP36K48M and HPD44NK48M. The red numbers are unfolding free energy measured by experiments. The blue numbers are calculated free energy changes in the native/folded (F). Green numbers in *italics* are calculated using two unfolding free energies and one calculated free energy in the same thermodynamic cycles.

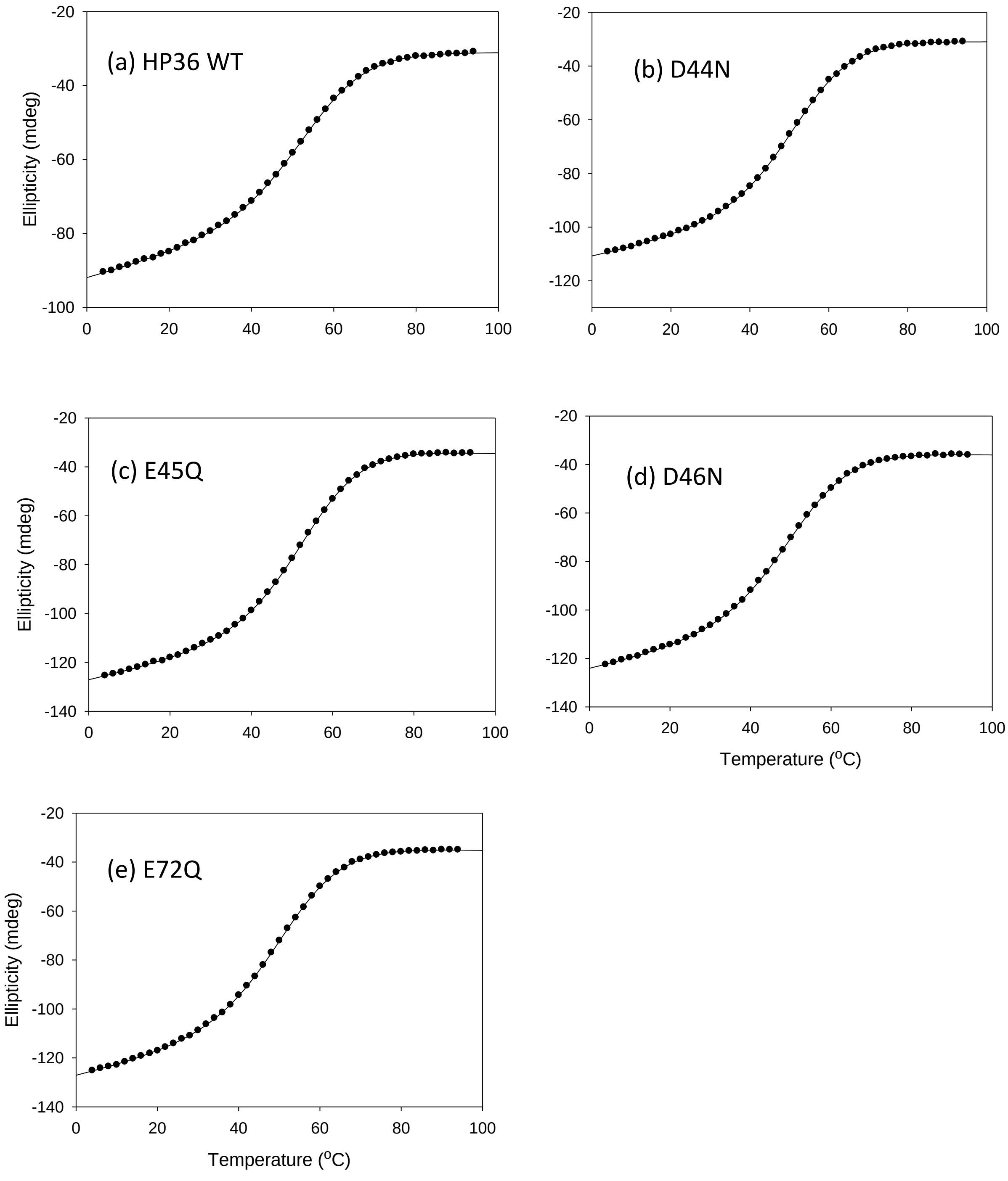


Figure 4-4. Temperature induced unfolding transitions of HP36 wildtype and the mutants. (a)HP36 WT, (b)D44N, (c)E45Q, (d)D46N, (e)E72Q. The solid line represents the best fit to a two-state folding transition. Signals were recorded at 222 nm. All spectra were collected in 10 mM sodium acetate, 150 mM sodium chloride, pH 3.0 using a 1.0 cm cuvette.

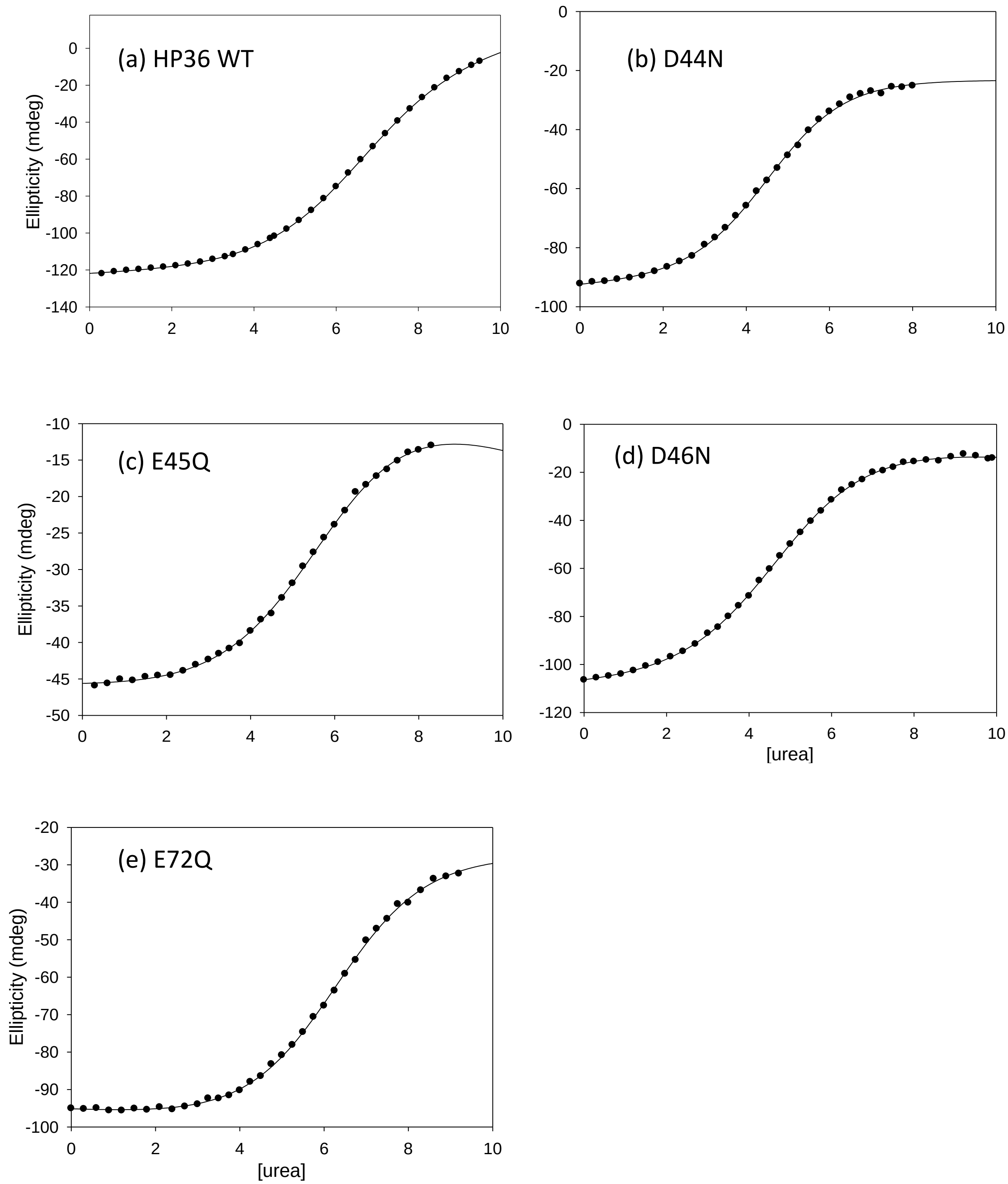


Figure 4-5. Urea induced unfolding transitions of HP36 wildtype and the mutants. (a)HP36 WT, (b)D44N, (c)E45Q, (d)D46N, (e)E72Q. The solid line represents the best fit to a two-state folding transition. Signals were recorded at 222 nm. All spectra were collected at 25 ^{o}C in 10 mM sodium acetate, 150 mM sodium chloride, pH 6.0 using a 1.0 cm cuvette.

4.5 Reference

1. McKnight CJ, Matsudaira PT, & Kim PS (1997) NMR structure of the 35-residue villin headpiece subdomain. *Nat Struct Biol* 4(3):180-184.
2. Wang M, *et al.* (2003) Dynamic NMR line-shape analysis demonstrates that the villin headpiece subdomain folds on the microsecond time scale. *J Am Chem Soc* 125(20):6032-6033.
3. Brewer SH, *et al.* (2005) Effect of modulating unfolded state structure on the folding kinetics of the villin headpiece subdomain. *Proc Natl Acad Sci U S A* 102(46):16662-16667.
4. Freddolino PL & Schulten K (2009) Common structural transitions in explicit-solvent simulations of villin headpiece folding. *Biophys J* 97(8):2338-2347.
5. Bi Y, *et al.* (2007) Rational design, structural and thermodynamic characterization of a hyperstable variant of the villin headpiece helical subdomain. *Biochemistry* 46(25):7497-7505.
6. Xiao S, *et al.* (2013) Rational modification of protein stability by targeting surface sites leads to complicated results. *Proc Natl Acad Sci U S A* 110(28):11337-11342.
7. Zagrovic B & Pande VS (2006) Simulated unfolded-state ensemble and the experimental NMR structures of villin headpiece yield similar wide-angle solution X-ray scattering profiles. *J Am Chem Soc* 128(36):11742-11743.
8. Wickstrom L, *et al.* (2006) The unfolded state of the villin headpiece helical subdomain: computational studies of the role of locally stabilized structure. *J Mol Biol* 360(5):1094-1107.

9. Reiner A, Henklein P, & Kiefhaber T (2010) An unlocking/relocking barrier in conformational fluctuations of villin headpiece subdomain. *Proc Natl Acad Sci U S A* 107(11):4955-4960.

10. Chiu TK, *et al.* (2005) High-resolution x-ray crystal structures of the villin headpiece subdomain, an ultrafast folding protein. *Proc Natl Acad Sci U S A* 102(21):7517-7522.

11. Meng W, Shan B, Tang Y, & Raleigh DP (2009) Native like structure in the unfolded state of the villin headpiece helical subdomain, an ultrafast folding protein. *Protein Sci* 18(8):1692-1701.

12. Tang Y, Goger MJ, & Raleigh DP (2006) NMR characterization of a peptide model provides evidence for significant structure in the unfolded state of the villin headpiece helical subdomain. *Biochemistry* 45(22):6940-6946.

13. Havlin RH & Tycko R (2005) Probing site-specific conformational distributions in protein folding with solid-state NMR. *Proc Natl Acad Sci U S A* 102(9):3284-3289.

14. Lindorff-Larsen K, Piana S, Dror RO, & Shaw DE (2011) How Fast-Folding Proteins Fold. *Science* 334(6055):517-520.

15. Tang Y, Rigotti DJ, Fairman R, & Raleigh DP (2004) Peptide models provide evidence for significant structure in the denatured state of a rapidly folding protein: the villin headpiece subdomain. *Biochemistry* 43(11):3264-3272.

16. Nagarajan S, Xiao S, Raleigh DP, & Dyer RB (2018) Heterogeneity in the folding of villin headpiece subdomain HP36. *J Phys Chem B* 122(49):11640-11648.

17. Tanford C (1970) Protein denaturation. C. Theoretical models for the mechanism of denaturation. *Adv Protein Chem* 24:1-95.

18. Wyman J, Jr. (1964) Linked functions and reciprocal effects in hemoglobin: A second Look. *Adv Protein Chem* 19:223-286.

19. Lindman S, *et al.* (2010) pK(a) values for the unfolded state under native conditions explain the pH-dependent stability of PGB1. *Biophys J* 99(10):3365-3373.

20. Tollinger M, Forman-Kay JD, & Kay LE (2002) Measurement of side-chain carboxyl pK(a) values of glutamate and aspartate residues in an unfolded protein by multinuclear NMR spectroscopy. *J Am Chem Soc* 124(20):5714-5717.

21. Meng W & Raleigh DP (2011) Analysis of electrostatic interactions in the denatured state ensemble of the N-terminal domain of L9 under native conditions. *Proteins* 79(12):3500-3510.

22. Shen JK (2010) A method to determine residue-specific unfolded-state pKa values from analysis of stability changes in single mutant cycles. *J Am Chem Soc* 132(21):7258-7259.

23. Oliveberg M, Arcus VL, & Fersht AR (1995) pKA values of carboxyl groups in the native and denatured states of barnase: the pKA values of the denatured state are on average 0.4 units lower than those of model compounds. *Biochemistry* 34(29):9424-9433.

24. Tan YJ, Oliveberg M, Davis B, & Fersht AR (1995) Perturbed pKA-values in the denatured states of proteins. *J Mol Biol* 254(5):980-992.

25. Bi Y (Studies of the folding and stability of the villin headpiece subdomain.

26. Chen VB, *et al.* (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66(Pt 1):12-21.

27. D.A. Case IYB-S, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden,, *et al.* (2018) *AMBER 2018* (University of California, San Francisco).

28. Kirkwood JG (1935) Statistical mechanics of fluid mixtures. *J Chem Phys* 3(5):300-313.

29. Maier JA, *et al.* (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput.* 11(8):3696-3713.

30. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, & Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926-935.

31. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, & Haak JR (1984) Molecular-dynamics with coupling to an external bath. *J Chem Phys* 81(8):3684-3690.

32. Darden T, York D, & Pedersen L (1993) Particle mesh ewald - an N.Log(N) method for ewald sums in large systems. *J Chem Phys* 98(12):10089-10092.

33. Ryckaert JP, Ciccotti G, & Berendsen HJC (1977) Numerical-integration of cartesian equations of motion of a system with constraints - molecular-dynamics of N-alkanes. *J Comput Phys* 23(3):327-341.

5. Probing long-range interactions in the denatured state of NTL9

Abstract

The denatured state ensemble (DSE) of NTL9 has residual secondary structure and is significantly more compact than predicted by the random coil model. Previously reported double mutants cycle studies indicate that energetically significant networks of coupled interactions exist in the DSE of NTL9. It was proposed that several hydrophobic residues may form transient hydrophobic clusters in the DSE of NTL9, which stabilize favorable electrostatic interactions between D8 and K12 in the DSE. In this study, the long-range interactions of the hydrophobic residues in the DSE of NTL9 were examined using the combined experimental and computational method described in Chapter 3. Out of the 7 hydrophobic residues that were proposed to be coupled to K12 in the DSE, 6 of them experience favorable long-range interactions in the DSE. This is consistent with the hypothesis that hydrophobic clusters coupled to K12 are present in the DSE of NTL9. However, more experiments need to be conducted to understand the couplings between hydrophobic residues in the DSE of NTL9 and the role of K12 on the thermodynamics of the hydrophobic clusters.

Acknowledgements

I gratefully acknowledge Koushik Kasavajhala, Chuan Tian and Kellon Belfon for their administration of computational resources.

5.1 Introduction

The N-terminal domain of the ribosomal protein L9 (NTL9) is a popular model protein for studying folding kinetics, protein stability and denatured state ensemble (DSE) conformations (1-6). Evidence has shown that the DSE of NTL9 cannot be described by using a simple random coil model (2). The native state of NTL9 consists of both α-helices and β-sheets and folds via a two-state mechanism (1, 5, 6). A destabilized mutant F5A NTL9 has shown that the DSE of NTL9 under native condition possesses significant compactness and residual secondary structure (7). In 8.3M urea, paramagnetic relaxation data for NTL9 indicates that the DSE of NTL9 has residual contacts (2). FRET studies conducted with ultrafast microfluidic methods have shown that there are significant long-range interactions in the DSE under native conditions (8). It was proposed that hydrophobic clusters are formed in the DSE of NTL9, which protect NTL9 from deleterious intermolecular interactions between unfolded NTL9 that lead to aggregation (2).

The mutation of K12M enhances the stability of NTL9 by 1.9 kcal/mol, which is believed to be mainly caused by removing favorable interactions experienced by K12 in the DSE of NTL9 including favorable electrostatic interactions with D8 (9, 10). In other words, the mutation is believed to stabilize the protein by destabilizing the unfolded state. Stability data collected from double mutant cycles revealed that hydrophobic residues are strongly coupled to K12 in term of stability. Since many of the hydrophobic residues coupled to K12 are distant from K12 in the folded state of NTL9, it is believed that these couplings are mainly in the DSE of NTL9. It was proposed that formation of transient hydrophobic clusters allows D8 and K12 to come closer in the DSE, which leads to a non-native salt bridge between D8 and K12 (**Fig. 5-1**). Truncating mutations of the hydrophobic residues have been shown to weaken the hydrophobic clusters and reduce the favorable electrostatic interactions between D8 and K12 in the DSE, which leads to a non-additive effect of double mutations on the stability of NTL9 (1). The double mutant cycles, the coupling strengths between K12M and each truncating mutation of a hydrophobic residue were measured, and values ranges from -1.28 to -0.01 kcal/mol were detected (1). These values can include contributions from the native state and the DSE. The coupling free energy between two residues X and Y is defined as $\Delta\Delta G_{coupling} = \Delta G(X\text{->}0) - \Delta G(Y\text{->}0) - \Delta G(X\text{->}0, Y\text{->}0) + \Delta G(X, Y)$, where X->0 and Y->0 represent the mutation of X and Y. ΔG(X, Y) is the folding free energy of the wildtype protein.

Experimentally, it is impossible to distinguish between native and DSE effects from stability measurements. However, native state effects can be calculated using thermodynamic integration. In Chapter 3, a combined experimental and computational method which can quantitatively estimate the strength of long-range interactions disrupted by mutations in IDPs was described. This method can be applied to estimate the strength of long-range interactions disrupted by mutations in the DSE as well, provided the experimentally measured stability changes caused by mutations are given. Here, the experimental stability changes caused by truncating mutations of the hydrophobic residues in NTL9 were combined with free energy changes computed by using TI on the folded state to probe long-range interactions made by the hydrophobic residues and their strengths in the DSE of NTL9.

5.2 Results

5.2.1 Hydrophobic residues experience favorable long-range interactions in the DSE of NTL9

The effect of truncation mutation of hydrophobic residues was calculated using TI for L-to-A, I-to-V and A-to-G. I-to-A and V-to-A mutations were not conducted as TI fails to reproduce the binding free energy changes caused by these mutations in SGPB/OMTKY3. The thermodynamic cycle used for calculating the long-range interactions disrupted by the truncating mutations is illustrated in **Fig. 5-2**. The long-range interactions disrupted by the mutations in the DSE of NTL9, $\Delta G°_{long\text{-}range}$, can be calculated as $\Delta\Delta G°_{long\text{-}range} = \Delta G°_{unfolded} - \Delta G°_{fragment}$, where $\Delta G°_{fragment}$ is the free energy changes caused by the mutations in tripeptides. The structure of the mutants is required for the calculations. Crystal structures are not available for the mutants. Consequently, the mutant structures were modelled by modifying the PDB file of the wildtype protein (PDB code 2HBB). Since all mutants only involve truncation of the side chains, the structures for the mutants can be obtained by deleting atoms in the PDB file of the wildtype protein. The experimental unfolding free energies of NTL9-wildtype and NTL9 mutants are adopted from a previous publication (3). The values of $\Delta G°_{folded}$, $\Delta G°_{fragment}$ and $\Delta\Delta G°_{long\text{-}range}$ are listed in **table 5-1**. A negative $\Delta\Delta G°_{long\text{-}range}$ value indicates that the residue is more favorable than its less hydrophobic counterpart in the DSE of NTL9, which suggests that the residue participates in favorable hydrophobic interactions in the DSE. The $\Delta\Delta G°_{long\text{-}range}$ for A-to-G mutations contain effects from decrease of hydrophobicity as well as the increase of backbone entropy in the DSE as explained in Chapter 2. Since the increase of backbone entropy will make Gly more favorable in the DSE, the long-range hydrophobic interactions disrupted by A-to-G mutations should be more negative than $\Delta\Delta G°_{long\text{-}range}$. The values of $\Delta\Delta G°_{long\text{-}range}$ suggests that A22, A36, A39, I4, L35 and L47 make strong favorable long-range hydrophobic interactions in the DSE of NTL9. I18V appears to be involved in moderate favorable long-range hydrophobic interactions. It is possible that A42 has favorable long-range hydrophobic interaction as well, considering the potential entropic effect of the A-to-G mutation in the DSE, and its contribution to $\Delta G°$.

5.2.2 Residues that are energetically coupled to K12 have significant long-range interactions in the DSE

If the residues coupled to K12 form transient hydrophobic clusters in the DSE, they should experience favorable long-range hydrophobic interactions with the residues in the DSE. The experimental coupling strength between the hydrophobic residues and K12 ($\Delta\Delta G°_{coupling}$) are compared with $\Delta\Delta G°_{long\text{-}range}$ in **table 5-2**.

The trend of $\Delta\Delta G°_{long\text{-}range}$ values matches the trend of $\Delta\Delta G°_{coupling}$ values for all three L-to-A mutations. For the I-to-V mutations, $\Delta\Delta G°_{long\text{-}range}$ and $\Delta\Delta G°_{coupling}$ have matched trends for I4V and I37V mutations. However, a strong coupling between I18 and K12 in the DSE was observed, while the calculated $\Delta\Delta G°_{long\text{-}range}$ of I18V indicates a more moderate hydrophobic interaction in the DSE. For the A-to-G mutations, A39 has a strong hydrophobic interaction as described by the calculated $\Delta\Delta G°_{long\text{-}range}$, but the coupling between A39 and K12 has a moderate strength in the DSE. A42 makes little to no hydrophobic interactions as judged by $\Delta\Delta G°_{long\text{-}range}$, but its coupling with K12 is strong in the DSE, as judged by $\Delta\Delta G°_{coupling}$.

Overall, for the seven residues (A22, A36, A42, I4, I18, L35 and L47) that were suggested to strongly couple with K12 ($\Delta\Delta G°_{coupling} < -0.9$ kcal/mol) and form transient hydrophobic clusters in the DSE, five of them (A22, A36, I4, L35 and L47) are calculated to have significant favorable hydrophobic interactions, which are disrupted by the truncating mutations and one of them (I18) has a moderate value of $\Delta\Delta G°_{long\text{-}range}$. For the three residues (A39, I37 and L6) that have moderate coupling strengths as estimated by the experimental double mutant cycle analysis with K12 ($\Delta\Delta G°_{coupling} > -0.6$ kcal/mol), only A39 has strong favorable hydrophobic interactions. We plotted the calculated values of $\Delta\Delta G°_{long\text{-}range}$ versus the experimental double mutant cycle coupling free energies $\Delta\Delta G°_{coupling}$ (**Fig. 5-3**).

5.3 Discussion

The comparison between $\Delta\Delta G°_{long\text{-}range}$ and $\Delta\Delta G°_{coupling}$ indicates that most of the residues that are strongly coupled with K12 experience strong hydrophobic interactions in the DSE, which is consistent with the hypothesis that there are transient hydrophobic clusters stabilizing the favorable interactions of K12 in the DSE (1). In previous work, double mutants cycle studies showed that there is minimal coupling between A36 and A42 as well as between I4 and L35 in the DSE (1). However, the calculated $\Delta\Delta G°_{long\text{-}range}$ of these residues indicate that they all have significant favorable hydrophobic interactions in the DSE. In addition, these residues are all strongly coupled with K12 in DSE. An explanation is that, for example, the A36G mutation disrupts the hydrophobic clusters that stabilize the interactions of K12, but A42 is still involved in hydrophobic clusters that do not couple with K12. However, this hypothesis cannot explain why in the background of K12M, unfavorable couplings are observed for A36 with A42, I4 with L35 and A22 with L47 in the DSE. More double mutants cycle experiments, which involve two hydrophobic residues, and more triple mutants cycle experiments, which involve K12 and two hydrophobic residues, need to be conducted to examine the coupling between the hydrophobic residues in the DSE and the role of K12 in the formation and stability of the hydrophobic clusters.

It would also be attractive to examine proteins for which structures of mutants are available. T4 lysozyme is one example where a large number of mutant structures are known.

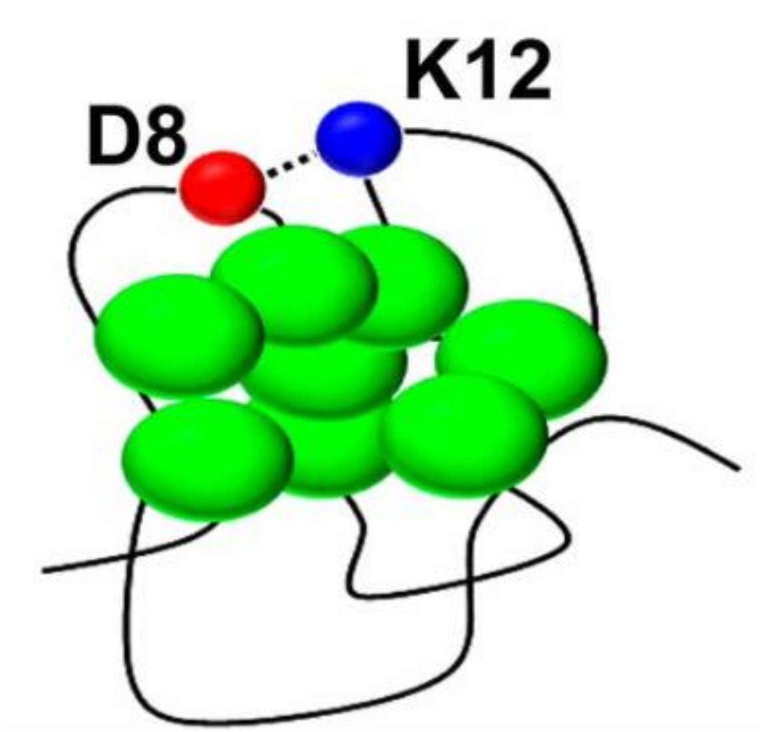


Figure 5-1. A hypothetical cartoon model for the DSE of NTL9 wild-type. Hydrophobic residues are shown as green beads. This figure was adapted from reference (1).

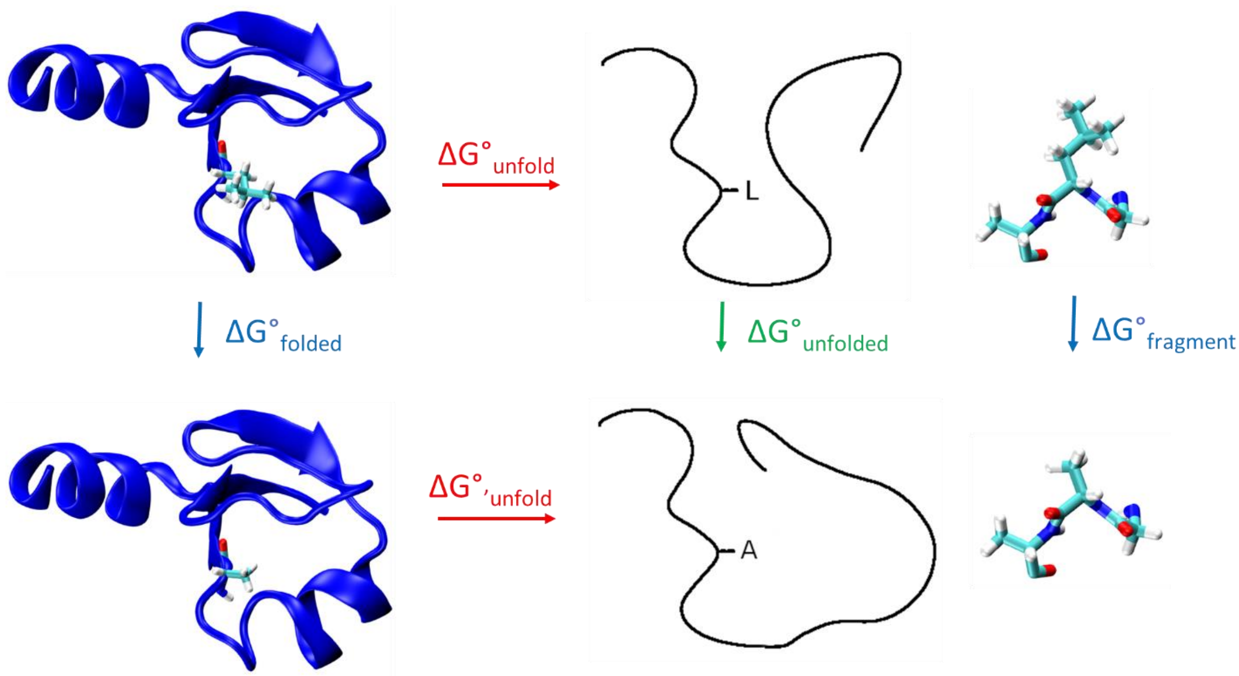


Figure 5-2. The thermodynamic cycle for calculating the effect of mutations on the long-range interactions in the DSE of NTL9. The L35A mutation is used as an example. $\Delta G°_{unfold}$ and $\Delta G°'_{unfold}$ are the experimentally measured unfolding free energy of NTL9-wildtype and NTL9-L35A respectively. $\Delta G°_{folded}$ and $\Delta G°_{fragment}$ are the free energy changes caused by L35A mutation calculated using TI in the folded state of NTL9 and the tripeptide reference state respectively. The effect of mutation on the unfolded state of NTL9 can be estimated as $\Delta G°_{unfold} = \Delta G°'_{unfold} - \Delta G°_{unfold} + \Delta G°_{folded}$.

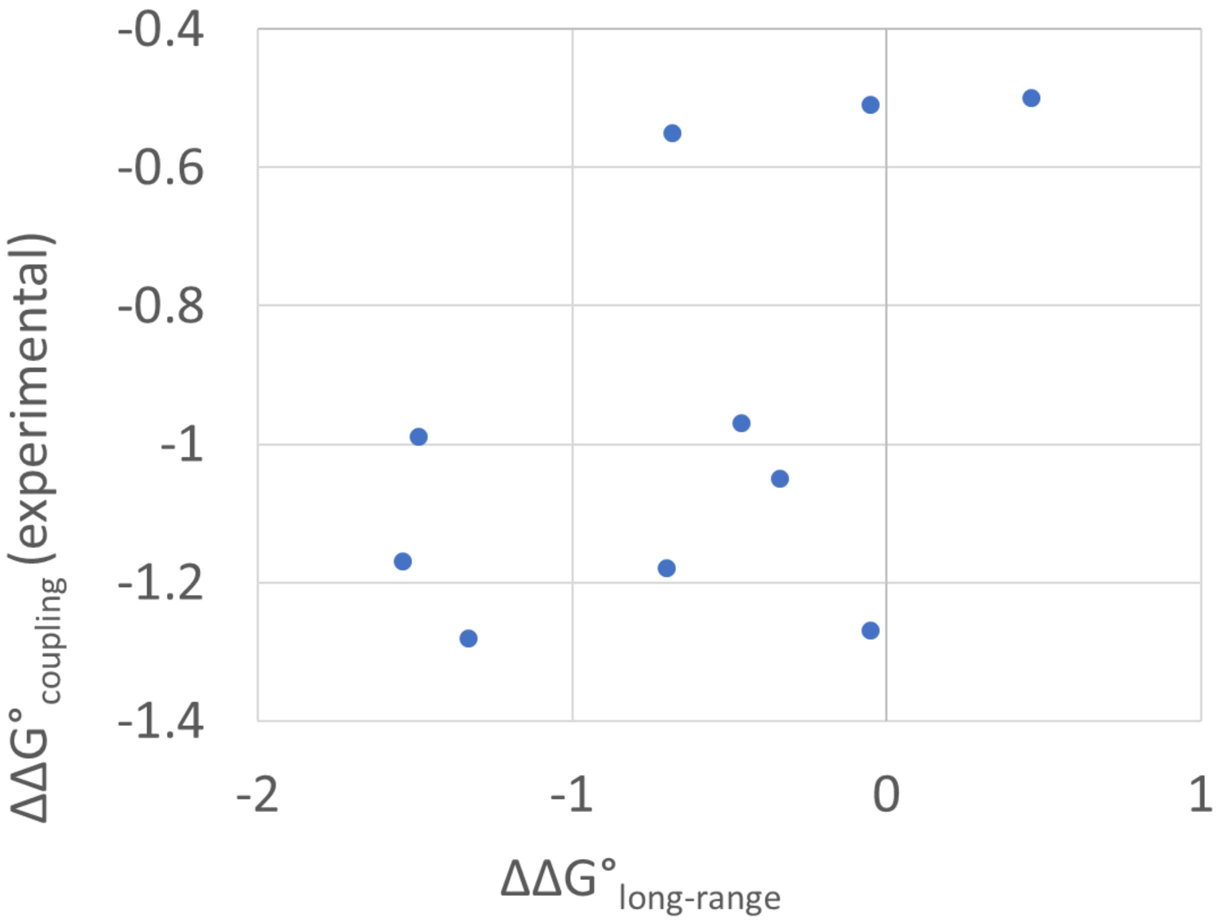


Figure 5-3. A scatter plot for the calculated values of $\Delta\Delta G°_{long\text{-}range}$ versus the experimental double mutant cycle coupling free energies $\Delta\Delta G°_{coupling}$ (experimental).

Table 5-1. The values of $\Delta G°_{folded}$, $\Delta G°_{fragment}$, $\Delta G°_{unfolding}$, $\Delta G°'_{unfolding}$ and $\Delta\Delta G°_{long\text{-}range}$ for the truncating mutations. Units are kcal/mol.

	$\Delta G°_{folded}$	$\Delta G°_{fragment}$	$\Delta G°_{unfolding}$ (WT)	$\Delta G°'_{unfolding}$	$\Delta\Delta G°_{long\text{-}range}$
A22G	-9.04	-10.06	4.30	3.98	-0.70
A36G	-7.62	-9.63		2.75	-0.46
A39G	-7.86	-9.77		2.71	-0.68
A42G	-9.34	-9.72		3.97	-0.05
I4V	-17.30	-19.43		3.71	-1.54
I18V	-19.44	-19.85		4.23	-0.34
I37V	-19.40	-19.54		4.39	-0.05
L6A	21.08	18.44		1.20	0.46
L35A	21.37	18.67		3.09	-1.49
L47A	21.55	19.38		3.46	-1.33

The values of $\Delta G°_{unfolding}$ and $\Delta G°'_{unfolding}$ were adapted from reference and were measured in 20 mM sodium acetate and 100 mM NaCl at pH 5.5 and 25°C (3).

A negative value of ΔΔG°long-range indicates a favorable long-range interaction in the DSE is disrupted by the mutation.

Table 5-2. The values of $\Delta\Delta G°_{long\text{-}range}$ and $\Delta\Delta G°_{coupling}$ for the truncating mutations. Units in kcal/mol.

	$\Delta\Delta G°_{long\text{-}range}$	$\Delta\Delta G°_{coupling}$
A22G	-0.70	-1.18
A36G	-0.46	-0.97
A39G	-0.68	-0.55
A42G	-0.05	-1.27
I4V*	-1.54	
I18V*	-0.34	
I37V*	-0.05	
L6A	0.46	-0.50
L35A	-1.49	-0.99
L47A	-1.33	-1.28
I4A*		-1.17
I18A*		-1.05
I37A*		-0.51

$\Delta\Delta G°_{coupling}$ represents the coupling strength between the residue and K12 in the DSE. The values were adopted from reference and were measured in 20 mM sodium acetate and 100 mM NaCl at pH 5.5 and 25°C (1).

*:$\Delta\Delta G°_{coupling}$ between I4/I18/I37 and K12 were measured using I-to-A mutations, which differs from the I-to-V mutations used for $\Delta\Delta G°_{long\text{-}range}$.

5.4 References

1. Cho JH, *et al.* (2014) Energetically significant networks of coupled interactions within an unfolded protein. *Proc Natl Acad Sci U S A* 111(33):12079-12084.
2. Meng W, Lyle N, Luan B, Raleigh DP, & Pappu RV (2013) Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc Natl Acad Sci U S A* 110(6):2123-2128.
3. Sato S, Cho JH, Peran I, Soydaner-Azeloglu RG, & Raleigh DP (2017) The N-Terminal Domain of Ribosomal Protein L9 Folds via a Diffuse and Delocalized Transition State. *Biophys J* 112(9):1797-1806.
4. Taskent H, Cho JH, & Raleigh DP (2008) Temperature-dependent Hammond behavior in a protein-folding reaction: analysis of transition-state movement and ground-state effects. *J Mol Biol* 378(3):699-706.
5. Kuhlman B, Boice JA, Fairman R, & Raleigh DP (1998) Structure and stability of the N-terminal domain of the ribosomal protein L9: evidence for rapid two-state folding. *Biochemistry* 37(4):1025-1032.
6. Kuhlman B & Raleigh DP (1998) Global analysis of the thermal and chemical denaturation of the N-terminal domain of the ribosomal protein L9 in H_2O and D_2O. Determination of the thermodynamic parameters, deltaH(o), deltaS(o), and deltaC(o)p and evaluation of solvent isotope effects. *Protein Sci* 7(11):2405-2412.
7. Anil B, Li Y, Cho JH, & Raleigh DP (2006) The unfolded state of NTL9 is compact in the absence of denaturant. *Biochemistry* 45(33):10110-10116.

8. Peran I, *et al.* (2019) Unfolded states under folding conditions accommodate sequence-specific conformational preferences with random coil-like dimensions. *Proc Natl Acad Sci U S A* 116(25):12301-12310.

9. Cho JH, Sato S, & Raleigh DP (2004) Thermodynamics and kinetics of non-native interactions in protein folding: a single point mutant significantly stabilizes the N-terminal domain of L9 by modulating non-native interactions in the denatured state. *J Mol Biol* 338(4):827-837.

10. Cho JH & Raleigh DP (2005) Mutational analysis demonstrates that specific electrostatic interactions can play a key role in the denatured state ensemble of proteins. *J Mol Biol* 353(1):174-185.

6. Current challenges for the application of alchemical free energy calculations to protein biophysics

In this dissertation, alchemical free energy calculations have shown successes in studying the thermodynamics of protein folding, protein-protein binding interactions and long-range interactions in unfolded proteins as well as IDPs. This was made possible thanks to the development of more accurate force fields and advances in computing hardware. However, alchemical free energy calculations still face a number of challenges in the aspects of system preparation, inaccuracy in current force fields and sampling insufficiency. These issues have been discussed in a recent publication on the applications of alchemical free energy calculations to drug design (1). Here, some of the issues will be elaborated in the context of calculating the free energy changes caused by mutations.

6.1 System preparation

Some amino acids can have different chemical states in physiological condition, for example the protonation states of titratable residues, the tautomeric states of His, the redox states of Cys and others. The occurrence of a certain chemical state is determined by its intrinsic preference as well as the surrounding chemical environment. When calculating the free energy changes caused by mutations, it is important to know the chemical states of the residues involved in mutations and the surrounding residues of mutation sites. In regular MD simulations, the chemical states of residues must be preset and fixed during the whole simulations. Fixing the chemical states will not cause issues if the chosen chemical states are the predominant states observed in experiments. However, in case that the residue involved in mutation and its surrounding residues have mixed chemical states or change their chemical states from one conformation to another, the free energy changes in each microscopic chemical state must be calculated. A simple example is provided by changes in protonation state between different conformations. In addition, the relative free energies for each microscopic chemical state must be measured through experiments or calculated using free energy calculations. An example is the calculation of the binding free energy change of SGPB/OMTKY3 caused by the Asp18-to-Ala mutations in OMTKY3 (**Fig. 3-5**), in which the relative free energies for protonating Asp18 were estimated using experimentally measured pKa values. However, this approach is practically infeasible when the system has too many microscopic chemical states that are important for free energy changes. To tackle this issue, constant pH and constant redox potential MD simulations have been implemented in major MD simulation software, which allows dynamic chemical states during the course of MD simulations (2-6).

Another challenge is the design of intermediate states for the transformations of functional groups. To maximize the precision of free energy calculation, a series of intermediate states, which provides sufficient overlap of phase space between the two functional groups need to be carefully designed. Moreever, the number of intermediate states need to be minimized to avoid accumulation of errors. Theoretically, the optimal intermediate states should have low and uniform variance of $U_B(\lambda, X) - U_A(\lambda, X)$ in **equation 1-11** (7, 8). However, optimal intermediate states are system dependent and may require modifying the code of MD simulations (8). Thus, usually a series of sub-optimal intermediate states are designed through trial and error. Designing the intermediate states for transformations that only involve deletion or addition of atoms on the ends of side chains is trivial. Transformations between Pro and non-Pro residues are less straightforward as they involve bond breaking. This is an example of a large problem known as “scaffold hopping” in free energy calculations (9, 10). A method using λ-dependent scaling of bond interactions, called soft potential bond, has been introduced. This method allows the users to construct intermediate states with low and uniform variance of $U_B(\lambda, X) - U_A(\lambda, X)$ for transformations involving bond breaking (9, 10). However, this method is not implemented in most major MD simulation software packages.

The calculations of free energy changes involving net charge change require additional cautions. Countering alchemical ions or correction terms can be used to correctly model the effect caused by changing the net charge of the system (11-13).

6.2 Force fields

Inaccuracy of force fields has always been a major challenge in molecular modelling. One of the major concerns about the widely-used fixed-charge force fields is the negligent of the anisotropy in electron distributions. For some interactions such as pi-pi, pi-cation and hydrogen bond interactions, the potential energies of interactions are highly sensitive to the relative orientations between the chemical groups due to the anisotropic distributions of electrons. However, since the fixed-charge force fields treat atoms as particles emitting isotropic electrostatic fields from all angles, the fixed-charge force fields are usually insensitive to the orientation between two atoms.

Several approaches have been proposed to better describe the electron distributions in molecules. In the Drude polarizable force field, a classical Drude oscillator attaches the atoms with additional dummy particles, which carry partial charges and masses, using harmonic bonds (11). The positions of the dummy particles respond to the changes of electric fields. In addition, the Drude polarizable force field also adopts dummy particles for lone pairs to better describe directionality of hydrogen bonds (12). The Amoeba force field treats polarization by using induced dipoles based on Thole's damped interaction method (13). Atoms in the Amoeba force field also have high order multipoles to account for the directionality of electron distributions. Since the Amoeba force field directly solves the function of electric field and multiple moments, no dummy particles are required. The Amoeba force field is considered to be more rigorous than the Drude force field.

The higher level of physics comes with a price of speed. The Drude force field requires 4 times more computational cost than fixed-charge force field (12) and the Amoeba force field requires 20 times more (14). Although both force fields have shown significant improvement in reproducing inter- and intra-molecular energies calculated by high level quantum mechanics, the high computational cost prevents the broad applications of these two force fields. Even with recent advances in hardware these methods are still too slow to tackle most protein problems. Only a few studies regarding the use of these two force fields in the calculation of free energy changes caused by mutations have been published (15). Another limitation of the Drude and the Amoeba force fields is that not all major MD simulation software is compatible with them.

Specific issues we faced with the Amber force field employed in the work of this thesis were problems with parameterization of Cys and Met. We also noted issues involving changes of β-branched to non-β-branched amino acids and vice versa. An important future direction for Amber users is to correct these problems.

6.3 Sampling

For mutations that do not significantly alter the structure of a protein, a simulation length of about 5ns is typically used for each λ window in free energy calculations (10, 16). However, if the mutations cause significant structural changes, longer simulations are required so that the structures can be fully equilibrated. A typical example is the different rotamer states of Val111 observed in structures of T4 lysozyme bound to different compounds (17). Different rotamer states of Val111 significantly affect the calculated binding free energies because the exchange of the rotamer states of Val111 is slow. Thus, it is important for the conformations of Val111 to be fully equilibrated during MD simulations, which requires extending the MD simulations. However, prolonging the simulations may cause the protein to drift away from its correct structures in MD simulations due to inaccuracy in force fields. Thus, methods which facilitate the structural changes during free energy calculations have been developed to tackle this issue (18-21).

Replica exchange with solute tempering (REST) is inspired by the traditional replica exchange method, however, REST allows enhanced sampling of specified regions which requires less replicas (22). Later, a method called REST/FEP was developed, in which REST was combined with free energy calculations (21). In REST/FEP, REST is applied on the neighboring atoms of the mutation sites to facilitate the structural changes caused by mutations. The temperature replica exchange in the original REST was replaced by Hamiltonian exchange and the Hamiltonian exchange is accompanied by the exchange of λ, the alchemical states of mutations. In the middle replicas (λ≈0.5), the Hamiltonian of the surrounding atoms is weakened, thus any λ windows passing through the middle replicas will receive a boost on the dynamics of the surrounding atoms. It has been shown that this method significantly facilitates the sampling of Val111 in T4 Lysozyme during free energy calculations, which eliminate the dependency of calculated binding free energies on the initial conformation of Val111 (21). The method does not require prior structural knowledge regarding the effect of mutations. However, this method requires sufficient exchanging between each replica, which may not be true for some systems. Moreover, this method has not been implemented in most major MD simulations packages.

Another method called confine-and-release has also been used to tackle the dependency of calculated binding free energies on the conformations of Val111 in T4 lysozyme. In this method, the binding free energies of different compounds were calculated with auxiliary restraints on Val111 which induced the change of the rotamer states. The free energy changes associated with the changes of rotamer states were calculated and decoupled from the binding free energies (20). This approach could be used in many major MD simulations packages and this would be a useful development. However, this method requires prior knowledges about the structural changes caused by mutations.

Modelling slow exchanging solvent molecules near the mutation sites is another challenge faced by alchemical free energy calculations. The mutations from larger side chains to smaller ones may create buried cavities that can accommodate water molecules and ions, which alter the partitioning of solvent molecules. However, the amount of time required for solvent molecules to enter and leave buried cavities usually exceeds the allowed time-scale of MD simulations. One approach to tackle this difficulty is to calculate the free energy change caused by the repartitioning of solvent and the free energy change caused by mutation without repartitioning of solvent. The two calculated free energy changes can be combined to obtain the free energy change caused by the mutation with repartitioning of solvent (23). Another approach is to use dynamic repositioning of solvent molecules which allows accelerated repartitioning of solvent molecules during the course of MD simulations (24, 25). Nevertheless, both approaches require expert knowledge of free energy calculations and may involve significant modifications to the existing MD simulation code.

6.4 References

1. Chodera JD, *et al.* (2011) Alchemical free energy methods for drug discovery: progress and challenges. *Curr Opin Struct Biol* 21(2):150-160.
2. Goh GB, Hulbert BS, Zhou H, & Brooks CL, 3rd (2014) Constant pH molecular dynamics of proteins in explicit solvent with proton tautomerism. *Proteins* 82(7):1319-1331.
3. Mongan J, Case DA, & McCammon JA (2004) Constant pH molecular dynamics in generalized Born implicit solvent. *J Comput Chem* 25(16):2038-2048.
4. Machuqueiro M & Baptista AM (2006) Constant-pH molecular dynamics with ionic strength effects: protonation-conformation coupling in decalysine. *J Phys Chem B* 110(6):2927-2933.
5. Radak BK, *et al.* (2017) Constant-pH Molecular Dynamics Simulations for Large Biomolecular Systems. *J Chem Theory Comput* 13(12):5933-5944.
6. Cruzeiro VWD, Amaral MS, & Roitberg AE (2018) Redox potential replica exchange molecular dynamics at constant pH in AMBER: Implementation and validation. *J Chem Phys* 149(7):072338.
7. Shenfeld DK, Xu H, Eastwood MP, Dror RO, & Shaw DE (2009) Minimizing thermodynamic length to select intermediate states for free-energy calculations and replica-exchange simulations. *Phys Rev E Stat Nonlin Soft Matter Phys* 80(4 Pt 2):046705.
8. Pham TT & Shirts MR (2011) Identifying low variance pathways for free energy calculations of molecular transformations in solution phase. *J Chem Phys* 135(3):034114.
9. Wang L, *et al.* (2017) Accurate Modeling of Scaffold Hopping Transformations in Drug Discovery. *J Chem Theory Comput* 13(1):42-54.

10. Yu HS, *et al.* (2017) Accurate and Reliable Prediction of the Binding Affinities of Macrocycles to Their Protein Targets. *J Chem Theory Comput* 13(12):6290-6300.

11. Chen W, *et al.* (2018) Accurate Calculation of Relative Binding Free Energies between Ligands with Different Net Charges. *J Chem Theory Comput* 14(12):6346-6358.

12. Reif MM & Oostenbrink C (2014) Net charge changes in the calculation of relative ligand-binding free energies via classical atomistic molecular dynamics simulation. *J Comput Chem* 35(3):227-243.

13. Rocklin GJ, Mobley DL, Dill KA, & Hunenberger PH (2013) Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: an accurate correction scheme for electrostatic finite-size effects. *J Chem Phys* 139(18):184103.

14. Lamoureux G & Roux B (2003) Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm. *J Chem Phys* 119(6):3025-3039.

15. Lopes PE, *et al.* (2013) Force Field for Peptides and Proteins based on the Classical Drude Oscillator. *J Chem Theory Comput* 9(12):5430-5449.

16. Shi Y, *et al.* (2013) The Polarizable Atomic Multipole-based AMOEBA Force Field for Proteins. *J Chem Theory Comput* 9(9):4046-4063.

17. Eastman P, *et al.* (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* 13(7):e1005659.

18. Panel N, Villa F, Fuentes EJ, & Simonson T (2018) Accurate PDZ/Peptide Binding Specificity with Additive and Polarizable Free Energy Simulations. *Biophys J* 114(5):1091-1102.

19. Zou J, Song B, Simmerling C, & Raleigh D (2016) Experimental and computational analysis of protein stabilization by Gly-to-d-Ala substitution: A convolution of native state and unfolded state effects. *J Am Chem Soc* 138(48):15682-15689.

20. Deng Y & Roux B (2006) Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J Chem Theory Comput* 2(5):1255-1273.

21. Jiang W & Roux B (2010) Free Energy Perturbation Hamiltonian Replica-Exchange Molecular Dynamics (FEP/H-REMD) for Absolute Ligand Binding Free Energy Calculations. *J Chem Theory Comput* 6(9):2559-2565.

22. Jiang W, Thirman J, Jo S, & Roux B (2018) Reduced Free Energy Perturbation/Hamiltonian Replica Exchange Molecular Dynamics Method with Unbiased Alchemical Thermodynamic Axis. *J Phys Chem B* 122(41):9435-9442.

23. Mobley DL, Chodera JD, & Dill KA (2007) The Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *J Chem Theory Comput* 3(4):1231-1235.

24. Wang L, Berne BJ, & Friesner RA (2012) On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities. *Proc Natl Acad Sci U S A* 109(6):1937-1942.

25. Liu P, Kim B, Friesner RA, & Berne BJ (2005) Replica exchange with solute tempering: a method for sampling biological systems in explicit water. *Proc Natl Acad Sci U S A* 102(39):13749-13754.

26. Hamelberg D & McCammon JA (2004) Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. *J Am Chem Soc* 126(24):7683-7689.

27. Ben-Shalom IY, Lin C, Kurtzman T, Walker RC, & Gilson MK (2019) Simulating Water Exchange to Buried Binding Sites. *J Chem Theory Comput* 15(4):2684-2691.

28. Ross GA, Bruce Macdonald HE, Cave-Ayland C, Cabedo Martinez AI, & Essex JW (2017) Replica-Exchange and Standard State Binding Free Energies with Grand Canonical Monte Carlo. *J Chem Theory Comput* 13(12):6373-6381.