INVESTIGATION OF THE ROLE OF PEPTIDE BOND CONFORMATIONS IN THE BINDING OF INTRINSICALLY DISORDERED PROTEINS

A Thesis

Presented by Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Bioinformatics and Computational Biology

in the

College of Graduate Studies

University of Idaho

by

Yingqian (Ada) Zhan

Major Professor: F. Marty Ytreberg, Ph.D.

Committee Members: Stephen Krone, Ph.D.; Robert Heckendorn, Ph.D.

Department Administrator: Eva Top, Ph.D.

July 2016

Authorization to Submit Thesis

This thesis of Yingqian Zhan, submitted for the degree of Master of Science with a Major in Bioinformatics and Computational Biology and titled, "INVESTIGATION OF THE ROLE OF PEPTIDE BOND CONFORMATIONS IN THE BINDING OF INTRINSICALLY DISORDERED PROTEINS," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ______________________________ Date: __________________
F. Marty Ytreberg, Ph.D.

Committee Members: ______________________________ Date: __________________
Stephen Krone, Ph.D.

______________________________ Date: __________________
Robert Heckendorn, Ph.D.

______________________________ Date: __________________
Jill L. Johnson, Ph.D.

Department
Administrator: ______________________________ Date: ___________________
Eva Top, Ph.D.

Abstract

Peptide bonds in proteins are predominantly found in the trans conformation. The cis conformation is typically found associated with prolines in intrinsically disordered proteins (IDPs), less so in structured proteins. It is not currently well understood how the cis-trans isomerization of a proline amino acid modifies protein-protein binding in IDPs. In this thesis, computer simulations were used to study how the cis and trans conformations of a proline in the IDP p53 modify its affinity for MDM2. Results show that the cis isomer of p53(17-29) binds more weakly to MDM2 as compared to the trans isomer, and that this is primarily due to the difference in the free energy cost associated with the loss of conformational entropy of p53(17-29) when it binds to MDM2. In addition, a survey was conducted analyzing the frequencies of both cis and trans conformations in a database containing membrane protein molecular recognition features (mpMoRFs). These mpMoRFs are a class of IDPs in membrane that become structured when they bind to their partners. Analysis of amino acid composition showed that mpMoRFs consist both order- and disorder-promoting amino acids and that the distributions of peptide bonds for Xaa-Pro mpMoRFs are distinct from natively structured proteins. In mpMoRFs, only 0.11%/0.75% of peptide bonds are in cis for non-proline/proline, in contrast to natively structured proteins where 0.03%/5.2% are in cis for non-proline/proline. These results suggest that cis-trans isomerization in mpMoRFs are important for function.

Acknowledgements

Grateful thanks should be given to Dr. F. Marty Ytreberg for his wise guidance and consistent support that led me to the accomplishment of a master degree in this subject. He always allows the research and paper to be my own work, but steered me in the right direction.

I would like to acknowledge my other Committee Members, Dr. Celeste Brown, Dr. Stephen Krone and Dr. Robert Heckendorn for their helpful discussion at committee meetings, and for reviewing this thesis. I am gratefully indebted to them for their valuable comments on this thesis.

I am also thankful to the Statistical Consulting Center at the Department of Statistical Science at University of Idaho for the advice on statistical analyses.

I gratefully appreciate Dr. François-Xavier Theillet for providing NMR data and analysis (In-cell NMR Laboratory, Department of NMR-supported Structural Biology, Leibniz Institute of Molecular Pharmacology, Berlin, Germany).

I am also grateful to the BCB director, Dr. Eva Top and the BCB program coordinator, Ms. Lisha Abendroth for escorting through every change in my program.

Dedication

This thesis is dedicated to my husband, Xinran Du, for his endless support and trust. It is also dedicated to my two daughters, Zumi and Mila, who are the source of love and strength. Without their sacrifices this would have not been possible.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Protein structure

Proteins typically have a specific three-dimensional structure that determines how they will function in a cell. They are built from 20 types of amino acid residues that only differ by their side chains (*1*) (Figure 1.1). Proline is unique due to its cyclic structure linking back to the backbone (Figure 1.1 (b)). A chain of residues are covalently linked together by peptide bonds making up the primary structure (Figure 1.2). Depending on the arrangement of residues in the sequence, a chain of residues can fold into regular secondary structures, such as alpha helix, beta strand, turns and loops (*1*). When the chain of residues gets longer, complex structures emerge. Different parts of a protein can have distinct secondary structures and they interact by physical forces such as Van der Waals and electrostatic (*2*) forming a tertiary structure. A higher level of complexity in protein structure, called quaternary structure, can occur when chains of residues assemble into multi-subunit structures.

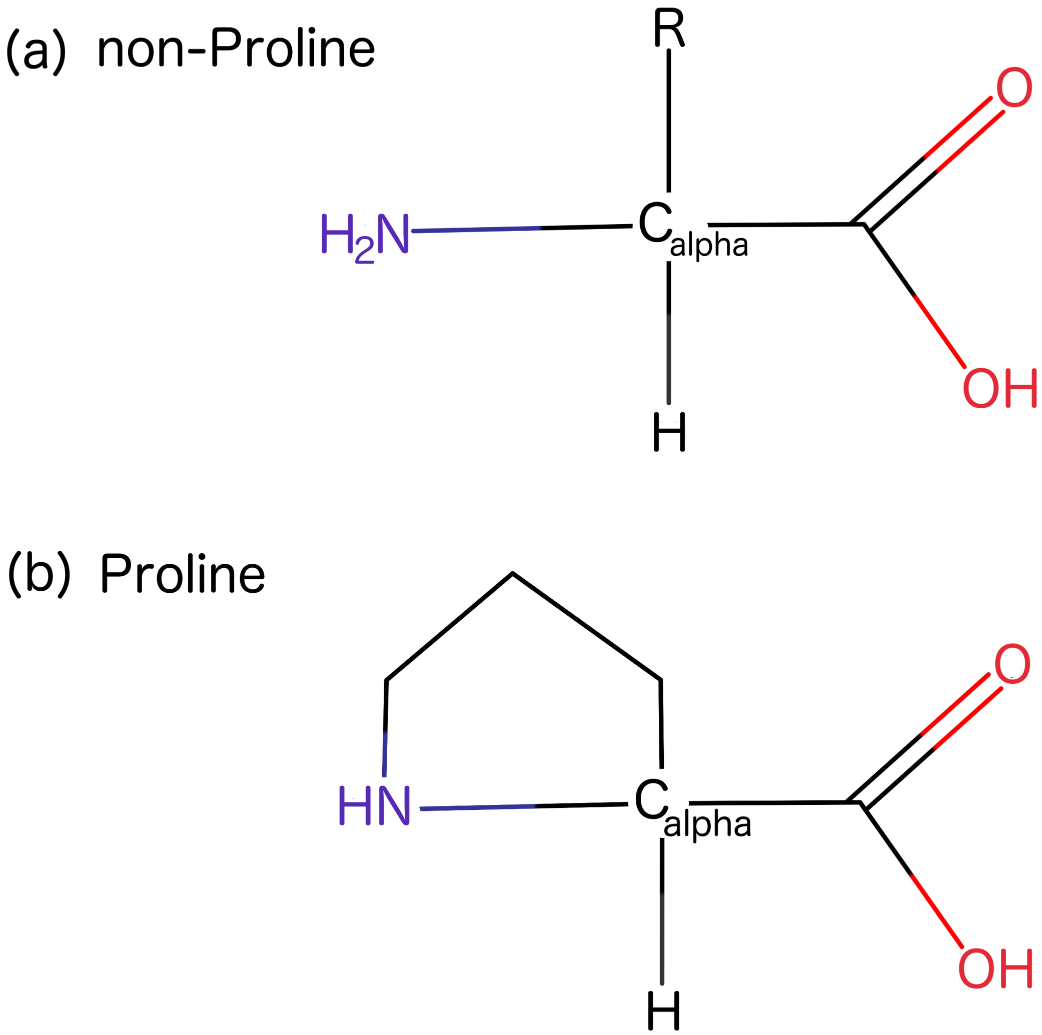

Figure 1.1: The structure of (a) one amino acid and (b) proline amino acid. R: the side chain that can have 20 different types, C_{alpha}: the carbon linked to the side chain, H: Hydrogen, O: Oxygen, N: Nitrogen. Proline has a cyclic side chain that is unique.

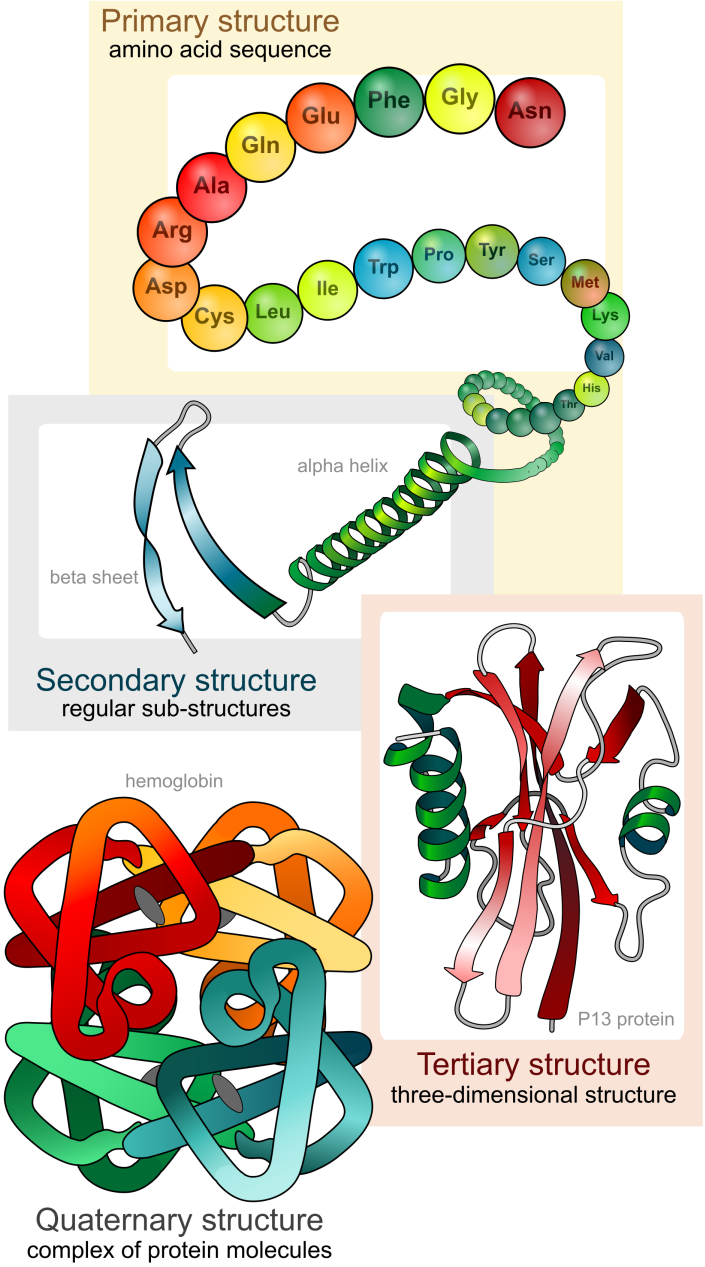

Figure 1.2: Protein structure level: from primary to quaternary structure. (This figure was reproduced from Wikimedia Commons and permitted by the original author LadyofHats (A professional illustrator based in Germany).)

The peptide bond is a covalent chemical bond that joins two amino acids together. It can have two conformations (also called isomers): trans and cis (Figure 1.3 (a)). Under certain conditions a peptide bond can interchange between trans and cis isomers; this process is called isomerization. As more and more protein structures have been resolved it has become clear that peptide bonds in protein are predominantly found in trans (*3*). The cis isomer occurs with very small probability because the there is a large steric clash between atoms in the amino acids as compared to the trans isomer. Due to its special structure, Xaa-Pro (Xaa: any amino acid, Pro: proline) peptide bonds are much more likely to be found in the cis isomer as compared to other amino acid peptide bonds (Figure 1.3 (b)). In a nonredundant set of 571 natively structured proteins, 0.03% of cis conformations are observed in Xaa-nonPro and 5.2% for Xaa-Pro (nonPro: any amino acid but proline) (*4, 5*).

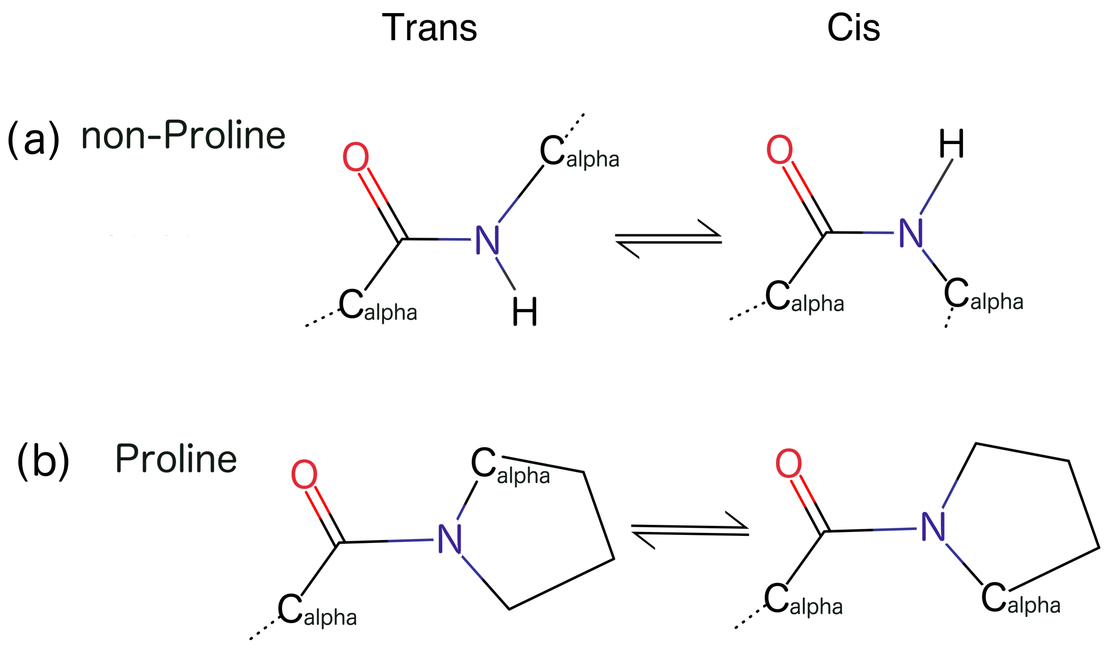

Figure 1.3: Chemical structures of trans and cis in (a) non-Proline and (b) Proline peptide bonds. The equilibrium arrow in each case indicates the interchange between the two isomers, i.e. isomerization. O (red): Oxygen, N (purple): Nitrogen, C_{alpha} (black): Alpha carbon. Alpha carbons link to side chains that differentiate residues.

The structure of the backbone of a protein (amino acids minus the side chains) is defined by three dihedral angles, phi, psi and omega as shown in Figure 1.4 (*6, 7*). Omega is the angle determining isomer of the peptide bond. Peptide bonds are partial double bonds that restrict omega to either ~ 180° (trans) or ~ 0° (cis).

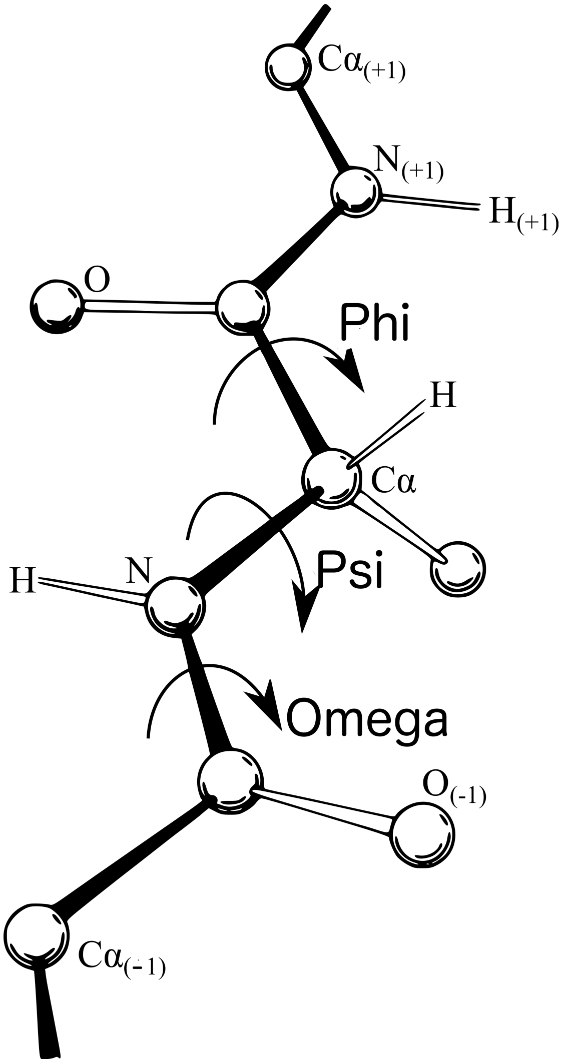

Figure 1.4: Protein backbone dihedral angles. (This figure was revised from Wikimedia Commons and permitted by the original author Dcrjsr (Dr. Jane Shelby Richardson at Duke University).)

1.2 Cis and trans isomerization and intrinsically disordered proteins

Cis-trans isomerization plays an important role in a variety of biological processes. Many studies have stressed the importance of the cis-trans isomerization of peptide bonds for protein folding processes (*8-11*). The isomerization process can be a timing mechanism in cell signaling (*12, 13*), ion channel gating (*14*), and gene expression (*15*). In addition, cis prolyl residues are more conserved through evolution (*16*) and the transition between cis and trans are believed to assist the emergence of new function among structural homologous proteins (*17*).

Cis isomers are more frequently associated with prolines in intrinsically disordered proteins (IDPs) as compared to natively structured proteins. Unlike natively structured proteins, IDPs do not fold into ordered three-dimensional structures. Instead, IDPs exhibit a broad range of conformations, and thus are often multi-functional and can act as hubs in the networks of biological processes (*18*). A recent report revealed that IDPs tend to have more prolines in their sequences on average (*19*) compared to natively structure proteins. Prolines may play multiple roles in IDPs since cis-trans isomerization in prolines exhibit biological significance and IDPs are highly populated with prolines. For example, it is known that the cis and trans proline isomers of the adaptor protein Crk turn the inhibitory functions on and off (*20*).

The goal of this thesis is to investigate the role of peptide bond cis-trans isomerization in the interactions of IDPs with other proteins. To accomplish this goal, we used computer simulations to determine how a single cis-trans isomerization of proline in the human tumor suppressor p53 affects its binding to the E3 ubiquitin ligase murine double minute clone 2 (MDM2), and we studied the properties of the cis-trans peptide bonds in binding regions for a large database of IDPs. This study is important since to date, there have been very few studies to investigate the functional implications of proline cis-trans isomerization in IDPs.

1.3 p53 and MDM2

The p53 protein is termed the "guardian of the genome" and helps maintain genomic integrity of the cell (*21*). It is stabilized and activated in response to a variety of cellular stresses leading to cell cycle arrest and the subsequent transcription of target genes to revive the cell (*21*). If the damage is irreversible, p53 will initiate cell death.

Normally, MDM2 binds and ubiquinates p53 to trigger the degradation process (*22*). However, some p53 proteins escape, thereby enabling the transcription of the MDM2 gene, which maintains the feedback loop (*22*). Therefore, the levels of p53 are kept low by the interaction with MDM2 in non-stressed cells. This low level of p53 is able to ensure a rapid response to stresses. Under stressed conditions, the binding between p53 and MDM2 is abrogated by phosphorylation, leading to the activation of p53.

An ordered N-terminal domain of MDM2 that is comprised of residues 1 to 109 binds to a short, disordered segment of the p53 transactivation domain (p53TAD) (*23, 24*). Residues 15-30 make up the short, disordered segment of p53TAD that binds to MDM2 (*23-27*). The disordered MDM2 binding region of p53TAD undergoes coupled folding and binding with MDM2 (*25, 28*), that is, the disordered segment of p53TAD folds into an ordered helical structure concomitant with binding to MDM2. The binding site for MDM2 contains residues 18 to 26 that are transiently helical in the unbound state (*23*). The proline at position 27 (P27) in p53 is in the C-terminal flanking region and is adjacent to L26, one of three critical residues for binding MDM2 (F19, W23, and L26) (*27*). The residue P27 has been established as a disrupter to the MDM2-binding motif of p53 as confirmed in recent studies (*29, 30*).

Molecular dynamics simulations are an effective tool to study how p53 and MDM2 bind or interact. In chapter 2, absolute binding affinities were calculated for p53 and MDM2 when a proline in p53 was in both cis and trans conformations. Results show that the cis isomer of p53(17-29) binds more weakly to MDM2 than the trans isomer, and that this is primarily due to the difference in the free energy cost associated with the loss of conformational entropy of p53(17-29) when it binds to MDM2.

1.4 Membrane protein molecular recognition features (mpMoRFs)

Molecular recognition features (MoRFs) are short, intrinsically disordered regions in proteins that undergo a disorder-to-order transition upon binding to their partners (*31*). MoRFs are capable of binding multiple partners. MoRFs play important roles in modulating the binding of IDPs and hence in regulating molecular recognition and cell signaling (*32, 33*). The database of MoRFs in membrane proteins (mpMoRFs) is a publically available database specialized on MoRFs (*34*).

Membrane proteins are vital in cell signaling (*35*). They are divided into three classes depending on their positions relative to the membrane: transmembrane proteins that span across the lipid bilayer, integral monotopic proteins that are attached to one side of the membrane, and peripheral proteins that are temporarily bound either to the membrane or to the other two types of membrane protein (*36*). A previous study revealed that around 20% of the proteins containing MoRFs are transmembrane (*37*). Two other studies demonstrated that mpMoRFs have unique amino acid composition compared to other proteins and thus MoRF predictors do not have good accuracy for transmembrane MoRFs (*37, 38*).

In chapter 3, we conducted a survey on the database of mpMoRFs and analyzed the peptide bond distributions. Analysis of amino acid composition showed that mpMoRFs consist both order- and disorder-promoting amino acids. The peptide bonds for Xaa-Pro mpMoRFs were found to be different than natively structured proteins. Predictions of proline peptide bonds showed that many proline bonds are predicted to be cis but are actually found in trans. These results suggest that cis-trans isomerization plays an important role in mpMorF function. More studies will be required to understand just how this isomerization relates to mpMoRF function and to improve peptide bond isomer prediction algorithms.

Chapter 2

The cis conformation of proline leads to weaker binding of a p53 peptide to MDM2 compared to trans

Note: This chapter has been published as Zhan Y, Ytreberg FM, Arch Biochem Biophys, 575:22-29 (2015).

The cis and trans conformations of the Xaa-Pro (Xaa: any amino acid) peptide bond are thermodynamically stable while other peptide bonds dominate in trans. The effect of proline cis-trans isomerization on protein binding has not been thoroughly investigated. In this study, computer simulations were used to calculate the absolute binding affinity for a p53 peptide (residues 17-29) to MDM2 for both cis and trans isomers of the p53 proline in position 27. Results show that the cis isomer of p53(17-29) binds more weakly to MDM2 than the trans isomer, and that this is primarily due to the difference in the free energy cost associated with the loss of conformational entropy of p53(17-29) when it binds to MDM2. The stronger binding of trans p53(17-29) to MDM2 compared to cis may leave a minimal level of p53 available to respond to cellular stress. The population of cis p53(17-29) was estimated to be 0.8% of the total population in the bound state. This study demonstrates that it is feasible to estimate the absolute binding affinity for an intrinsically disordered protein fragment binding to an ordered protein that are in good agreement with experimental results.

2.1 Introduction

Although the cis conformation of proline residues represents a very small fraction of peptide bonds, it is still biologically important. The vast majority of peptide bonds in proteins are observed in trans conformation (omega ~ 180°) due to favored interactions between the amide hydrogen and the preceding alpha carbon (*39*). However, peptide bonds in cis conformation (omega ~ 0°) are also found in some cases (*40*). In a nonredundant set of 571 proteins, a very small fraction (0.03%) of cis conformation are observed in Xaa-nonPro and this increases to 5.2% for Xaa-Pro (Xaa: any amino acid, nonPro: any amino acid but proline) (*3, 4, 41, 42*). Many studies have stressed the importance of the cis-trans isomerization of peptide bonds for protein folding processes (*8-11*). It has been shown that the isomerization processes is likely to play roles in cell signaling, ion channel gating, and gene expression (*12-15*). The unique structure of proline allows for a smaller entropic loss than other amino acids when undergoing isomerization from trans to cis (*11, 43*). The slow inter-conversion between cis and trans isomers of Xaa-Pro peptide bonds can be catalyzed by peptidyl-prolyl cis-trans isomerase to regulate biological processes (*44-47*). Dysfunction of the isomerization process may result in diseases such as cancer and Alzheimer's (*48-51*).

Prolines play important roles in the structure and dynamics of intrinsically disordered proteins (IDPs). The frequency of prolines in IDP sequences is twice that of ordered proteins (*19*). The ring structure of proline that links to the peptide backbone tends to disrupt the alpha helical structure of proteins if it is not at a capping position (*52-54*). Prolines in N-terminal flanking regions of pre-structured motifs have been predicted to promote helical structure whereas prolines in C-terminal flanking regions tend to terminate helix formation in IDPs (*55*). Mutations on prolines that cause increased helicity may enhance the binding between an IDP and its partner, and affect signaling in cells (*30, 56*).

The accurate estimation of binding affinities for IDPs could be valuable for designing therapeutic drugs (*57*), or in protein engineering since IDPs play important roles in cell signaling and transcription (*33, 58-60*). The networks of protein-protein interactions regulate a wide range of biological activities from cellular metabolism to signal transduction (*61*). The functions of IDPs are carefully tuned by the structures, dynamics and binding affinities (*18, 30, 62*).

The estimation of absolute binding affinities for protein-protein systems is a key challenge in computational biology (*63*). Various methods with differing levels of complexity and accuracy have been used to calculate protein-protein binding affinities. Empirical energy functions and scoring schemes are used to screen large protein databases in the search of a good binding partner (*64-67*). This class of approaches is designed to handle a large amount of molecules with high throughput, but tend to be inaccurate due to the simplicity of the scoring functions. Other methods such as linear interaction energy method (*68*) and the molecular mechanical and continuum solvent approach (*69*), combine the use of conformations from molecular dynamics (MD) simulations in explicit solvent with binding affinity functions. This class of methods is widely used but suffers from inaccuracy in some cases due to insufficient sampling of MD simulations and /or functions that are not general enough. Another group of methods, for example free energy perturbation (*70*) and thermodynamic integration (*71*), are based on statistical mechanics principles and depend entirely on simulations, typically with explicit solvent. These methods provide the most accurate binding affinity estimates, in principle, but can be hampered by insufficient sampling and/or very long simulation times. Another class of methods for calculating absolute binding affinities that also provides information about the binding/unbinding pathway is to estimate the potential of mean force (PMF) using restraining potentials to enhance convergence (*72-75*). The slope of the PMF provides information about the average force over all conformations along a defined reaction coordinate (*71*). The PMF can be integrated to estimate the free energy difference between two states. Restraints on the degrees of freedom of the system reduce the conformational space available enhancing the convergence of simulations. The free energies associated with the restraints are rigorously accounted for in order to generate an unbiased estimate of the binding affinity (*73, 76*). Some specific examples of this approach have been reported for AcpYEEI peptide binding with the human p56^{lck} SH2 domain (*73*), KID protein in association with KIX protein (*77*), and peptide APSYSPPPPP interacting with the SH3 domain of the Abl kinase (*76*).

The model system used in the current study is a disordered fragment of p53 (residues 17 to 29) binding to the E3 ubiquitin ligase murine double minute clone 2 (MDM2). Protein p53 activates the expression of MDM2 (*78-80*). In turn, MDM2 binds p53 for ubiquitination causing p53 to be transported out of the nucleus for degradation by the proteasome (*81-83*). This elegant feedback loop maintains low levels of p53 in non-stressed cells. Under stressed conditions, the binding between p53 and MDM2 is abrogated by post-translational modifications, resulting in increased levels of p53 (*84, 85*). The activated p53 then leads to cell cycle arrest and the subsequent transcription of target genes to revive the cell (*22*). The binding site for MDM2 contains residues 18 to 26 that are transiently helical in the unbound state (*28*). The proline at position 27 (P27) in p53 is in the C-terminal flanking region and is adjacent to L26, one of three critical residues for binding MDM2 (F19, W23, and L26) (*27, 69*). The residue P27 was established as a disrupter to the MDM2-binding motif of p53 as confirmed in recent studies (*29, 30, 55*).

In this study, we used computer simulations to calculate PMFs and corresponding binding affinities to understand how the cis and trans conformations of P27 in a p53 fragment (residue 17 to 29) affect the binding with MDM2. Nuclear magnetic resonance (NMR) spectroscopy revealed that around 5.5 % of the L26-P27 peptide bonds are in the cis conformation for the unbound p53(1-63), but a cis signal could not be resolved for the same peptide bond in the p53-MDM2 complex (unpublished NMR data from Dr. François-Xavier Theillet). A PMF-based approach was used to compute the absolute binding affinity for both trans and cis isomers binding with MDM2 and found to be -11.8 (1.0) kcal/mol and -8.9 (0.8) kcal/mol, respectively. Based on these affinity calculations the cis isomer was estimated to be 0.8 % of the total bound state population with the rest in trans. It was found that N29 of the trans isomer contributes to the binding by having stronger electrostatic attraction to MDM2 than the cis isomer. In addition, the cis isomer has more flexibility in the unbound state compared to trans that decreases the binding affinity for cis. The stronger binding of trans p53(17-29) to MDM2 compared to cis may suggest a mechanism to help maintain minimal levels of p53 in unstressed cells and allow for rapid response to cellular stress. Our results suggest that around 5 % of the p53 proteins in the cell may not be targeted for degradation because they are in cis and thus essentially unavailable for binding to MDM2.

2.2 Methods

Thermodynamic Cycle

The binding free energy, or binding affinity, is the free energy difference between bound and unbound states of a system. When restraints are added to the system the unbiased free energy can be calculated by accounting for the free energies of releasing these restraints. This process can be illustrated by a thermodynamic cycle shown in Figure 2.1 (*74*). In this study, restraining potentials were applied to the p53 to limit the freedom of the system and allow the simulations to converge more quickly. Conformational, axial, and orientational restraints were applied to p53 as shown in Figure 2.2. Since free energy is a state function, the change from unbound to bound state is independent of the path taken (*86*) and so the unbiased binding affinity ΔG_{bind} was calculated as

$$\Delta G_{\mathrm{bind}} = \Delta G^{\mathrm{u}}_{\mathrm{conf}} + \Delta G^{\mathrm{u}}_{\mathrm{axial}} + \Delta G^{\mathrm{u}}_{\mathrm{orient}} + \Delta G^{\mathrm{res}}_{\mathrm{bind}} + \Delta G^{\mathrm{b}}_{\mathrm{conf}} + \Delta G^{\mathrm{b}}_{\mathrm{axial}} + \Delta G^{\mathrm{b}}_{\mathrm{orient}} \quad (2.1)$$

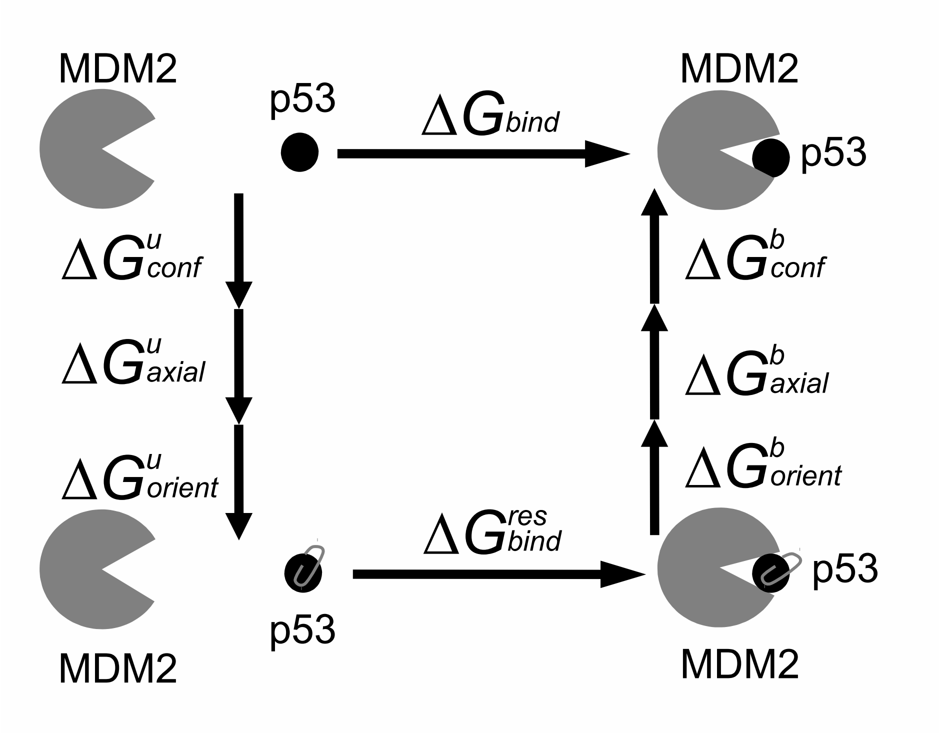

Figure 2.1: Thermodynamic cycle used to compute absolute binding affinities ΔG_{bind}. Conformational (conf), axial, and orientational (orient) restraints were applied to the unbound state (u) and bound state (b) of p53. The absolute binding free energy ΔG_{bind} was calculated by adding the restrained binding free energy ΔG^{res}_{bind} to the free energy differences associated with all the restraints via Eq. (2.1).

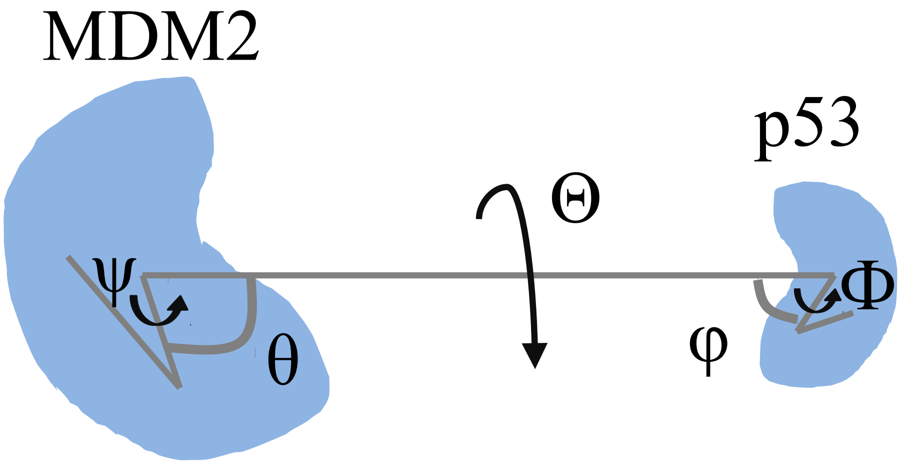

Figure 2.2: Schematic representation of the internal coordinates used to define the position and orientation of p53 relative to MDM2 and construct the restraining potentials.

Binding affinity with restraints

The binding affinity with restraints, $\Delta G_{\mathrm{bind}}^{\mathrm{res}}$, was calculated using the PMF w(r), where r the center of mass distance between p53 and MDM2. The formula used in this study was reported previously (*77, 87*) and given by

$$e^{-\beta \Delta G_{\mathrm{bind}}^{\mathrm{res}}} = 4\pi r^{*2} C^0 \int_0^{r^*} e^{-\beta[w(r)-w(r^*)]}\, dr \quad (2.2)$$

where $\beta = 1/k_bT$ (k_b is the Boltzmann constant.); C^0 is the standard state concentration 1.0 mol/liter (i.e. 1/1661 Å^3); r* is an arbitrary reference distance where the interaction between the two molecules is negligible.

Contributions from conformational restraints

The free energy cost to impose the conformational restraint for either the bound or unbound state was computed from the PMF w(ξ) , where ξ is the root mean square deviation (RMSD) of p53 relative to an equilibrated conformation. The conformation of p53 is restrained by a harmonic potential $u_c(\xi)$.

$$e^{\beta \Delta G^{b}_{conf}} = \frac{\int d\xi e^{-\beta(w^{b}(\xi)+u_c(\xi))}}{\int d\xi e^{-\beta w^{b}(\xi)}} \text{ , } e^{-\beta \Delta G^{u}_{conf}} = \frac{\int d\xi e^{-\beta(w^{u}(\xi)+u_c(\xi))}}{\int d\xi e^{-\beta w^{u}(\xi)}} \qquad (2.3)$$

Contributions from axial and orientational restraints

For the axial and orientational restraints in the bound state the free energy change was calculated using the Bennett acceptance ratio approach (*88*).

For the unbound state, the free energy cost associated with imposing axial and orientational restraints were calculated from numerical integration over the spherical angles, θ and φ, and Euler angles, ψ, Θ, and Φ. The formulas for the axial and orientational restraints are:

$$e^{-\beta \Delta G^{u}_{axial}} = \frac{\int d\tau e^{-\beta u_a}}{\int d\tau} = \frac{\int_0^{\pi} d\theta \sin\theta \int_0^{2\pi} d\varphi e^{-\beta u_a}}{4\pi} \qquad (2.4)$$

$$e^{-\beta \Delta G^{u}_{orient}} = \frac{\int d\Omega e^{-\beta u_o}}{\int d\Omega} = \frac{\int_0^{\pi} d\Phi \sin\Phi \int_0^{2\pi} d\Theta \int_0^{2\pi} d\Psi e^{-\beta u_o}}{8\pi^2} \qquad (2.5)$$

where u_a and u_o are respectively the axial and orientational restraint potentials.

Population of p53^{cis}-MDM2

The ratio of the population of p53^{cis}-MDM2 and p53^{trans}-MDM2 were derived from their binding affinities. According to Boltzmann statistics, the ratio of probabilities of two states in a system is equal to a ratio of their Boltzmann factors (*89*). The probability of p53^{cis}-MDM2 is the population of this complex divided by the total population of p53-MDM2, similarly for p53^{trans}-MDM2. The ratio of population of p53^{cis}-MDM2 and p53^{trans}-MDM2 was calculated using

$$R = \frac{\text{Population of p53}^{\text{cis}}-\text{MDM2}}{\text{Population of p53}^{\text{trans}}-\text{MDM2}} = \frac{e^{-\beta\Delta G_{\text{bind}}^{\text{cis}}}}{e^{-\beta\Delta G_{\text{bind}}^{\text{trans}}}} = e^{-\beta(\Delta G_{\text{bind}}^{\text{cis}}-\Delta G_{\text{bind}}^{\text{trans}})} \quad (2.6)$$

Assuming no other proline conformation besides cis and trans are present, the population of p53^{cis}-MDM2 is given by $\frac{R}{R+1}$.

Computational details

The initial structure of the p53^{trans}-MDM2 complex was obtained from the protein databank (PDB ID: 1YCR) (*23*) and includes residues 17-29 of p53 and residues 25-109 of MDM2. The coordinates for the missing hydrogen atoms in the crystal structure were guessed by VMD psfgen package (*90*). The simulations of unbound p53^{trans} were also initiated from PDB ID: 1YCR in the absence of MDM2.

The initial structure of the p53^{cis}-MDM2 complex was obtained by rotating the dihedral angle omega of L26-P27 peptide bond in p53 (PDB ID: 1YCR) from 180° to 0°

using VMD (*90*). This did not produce any steric clashes. Similarly, the unbound p53^{cis} was transformed from unbound p53^{trans}.

Each system was solvated in a TIP3P (*91*) water box that extended 17 Å in each direction from the solute and was given a neutral charge by adding 150 mM sodium chloride (NaCl). All simulations were carried out at 300 K using the CHARMM22 force field (*92*) as implemented in NAMD 2.9 (*93*). The long range electrostatic interactions were calculated using particle mesh Ewald (*94*) with a real-space cutoff of 12 Å. Nonbonded van der Waals interactions were smoothly switched to zero between 10 and 12 Å. The vibration of the bonds involving hydrogen atoms were constrained using SHAKE (*95*) to allow for a time step of 2 fs. Trajectory snapshots were saved every 2 ps. Each system was initially minimized for 1000 steps with the backbone atoms fixed and then minimized again for additional 1000 steps with all constraints removed. The minimized structures were gradually heated up to 300 K in 300 ps followed by equilibration at constant pressure 1 atm (Langevin piston) and temperature 300 K (Langevin dynamics) for 100 ps. (*96*) Independent 100 ns simulations were performed three times for p53^{trans}-MDM2 and p53^{trans}, three times for p53^{cis}, and six times for p53^{cis}-MDM2.

For p53^{trans}-MDM2 and p53^{cis}-MDM2 systems PMF calculations were performed using umbrella sampling (*97*) with WHAM (weighted histogram analysis method) (*98*). Umbrella sampling introduces a biasing potential energy that keeps the simulation near a particular value of the reaction coordinate (*97*). The WHAM package version 2.0 from Alan Grossfield was used for processing the umbrella sampling simulation results with error estimated by bootstrapping (*99*).

We calculated the PMF w(r) for p53 binding to MDM2 using various restraints to enhance the convergence of the simulations. A harmonic biasing potential $u_r = k_r (r\text{-}r_0)^2/2$ was applied to each window centered at different distances r_0 using a force constant $k_r = 10$ kcal/mol·Å^2. We used 37 windows with inter-window spacing 0.5 Å for $r_0 \leq 26$ Å and spacing of 1 Å for $r_0 > 26$ Å. For each window, 2 ns simulations were performed and the last 1 ns was used for analysis. The reference distance, r*, was set to be 35 Å, where p53 and MDM2 were not interacting.

The PMFs for the conformational restraints $w^u(\xi)$ and $w^b(\xi)$ were calculated from 21 umbrella sampling simulations separated by 0.5 Å for the unbound system and 0.4 Å for the bound system. For each window, 2 ns simulations were performed and the last 1 ns was used for analysis. A value of 1 kcal/mol·Å^2 was set to the force constant in the conformational restraining potential $u_c(\xi) = k_c\xi^2/2$. Harmonic biasing potentials $k_\xi(\xi\text{-}\xi_0)^2/2$ were used with force constants of 15 kcal/mol·Å^2.

The free energy differences, ΔG^{b}_{axial} and ΔG^{b}_{orient}, associated with the axial and orientational restraints were calculated by the Bennett acceptance ratio method. The axial restraining potential was $u_a = \frac{1}{2}k_a[(\theta - \theta^{ref})^2 + (\varphi - \varphi^{ref})^2]$ and the orientational potential was

$u_o = \frac{1}{2}k_o[(\Phi - \Phi^{ref})^2 + (\Theta - \Theta^{ref})^2 + (\Psi - \Psi^{ref})^2]$ with $k_a = k_o = 0.03$ kcal/mol·degree^2. For both axial and orientational restraint calculations, simulations of the complex were 2 ns for each value of k_a or k_o = 0.03, 0.021, 0.015, 0.01, 0.006, 0.003, 0.001, 0. The first 1 ns was discarded for equilibration and the remaining 1 ns was used to compute the corresponding free energy difference.

2.3 Results

In this study, a PMF-based approach with restraints was used to calculate the absolute binding affinity of the trans and cis isomers of p53(17-29) (p53$^{\text{trans/cis}}$) to MDM2. The binding affinities were calculated via Eq. (2.1) as depicted in the thermodynamic cycle (Figure 2.1). All the contributions to the affinity are discussed below and summarized in Table 2.1.

Table 2.1: Computation of the free-energy contributions to MDM2 and p53 binding.

Component	trans (kcal/mol)	cis (kcal/mol)
ΔG_{bind}^{res}	-23.89 ± 0.84	-23.92 ± 1.42
ΔG_{conf}^{u}	11.28 ± 1.01	13.89 ± 0.82
ΔG_{axial}^{u}	3.48 ± 0.00	3.50 ± 0.00
ΔG_{orient}^{u}	5.97 ± 0.08	5.82 ± 0.06
ΔG_{conf}^{b}	-7.03 ± 0.80	-7.33 ± 0.57
ΔG_{axial}^{b}	-0.90 ± 0.25	-0.45 ± 0.15
ΔG_{orient}^{b}	-0.73 ± 0.20	-0.44 ± 0.07
ΔG_{bind}(this study, p53(17-29))	**-11.83 ± 1.02**	**-8.93 ± 0.78**
ΔG_{bind} (experiment, p53(16-29))	-9.3	
ΔG_{bind} (MM-PBSA, p53(17-29))	-16.3	

The value of ΔG_{bind} was calculated by summing all the free energy components; see Eq. (2.1). Uncertainties are given by the standard errors and were estimated from six independent trials for cis and three trials for trans. Experimental and MM-PBSA values are obtained from references (*100*) and (*56*) respectively.

Binding affinity with restraints

Figure 2.3 shows PMFs for the transition from the p53$^{\mathrm{trans/cis}}$-MDM2 complex to the unbound state as a function of the center of mass separation between MDM2 and p53$^{\mathrm{trans/cis}}$ with all restraints present. All simulations were initiated from long independent MD simulations of the protein complexes leading to differences in the most favorable center of mass separations (PMF minima). All PMF curves are very flat for r > 30 Å where the interaction between the proteins is negligible. The corresponding free energy differences $\Delta G_{\mathrm{bind}}^{\mathrm{res}}$ are calculated from the PMFs using Eq. (2.2) using a reference distance of $r^{*} = 35$ Å. Table 2.1 shows that the average $\Delta G_{\mathrm{bind}}^{\mathrm{res}}$ values are very similar for the two isomers: -23.89 (0.84) kcal/mol for trans and -23.92 (1.42) kcal/mol for cis.

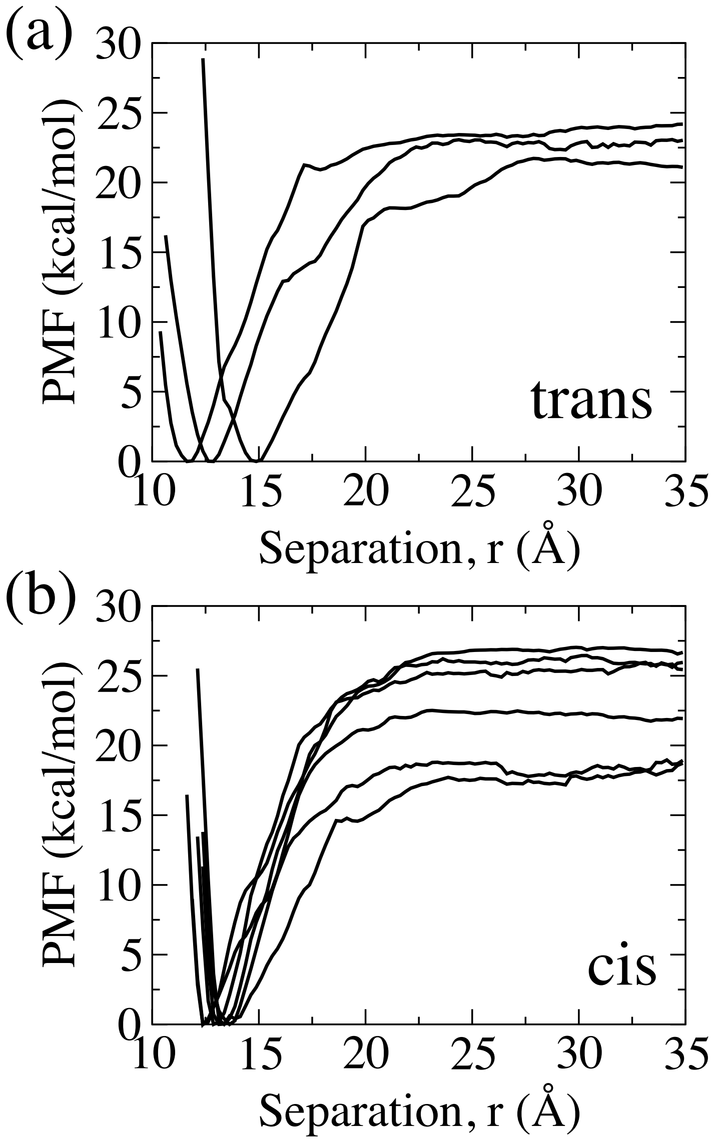

Figure 2.3: Potential of mean force (PMF) as a function of center of mass separation for the restrained (a) trans and (b) cis conformations of p53 binding to MDM2. Three independent trials were performed for the trans isomer and six trials were performed for the cis. Independent trials were initiated from the last frames of 100 ns simulations of the p53-MDM2 complex that resulted in the differences in the location of the PMF minima.

Contributions from conformational restraints

Figure 2.4 shows the PMFs as a function of RMSD calculated for p53$^{trans/cis}$ in the bound state and unbound states. The conformational restraints are based on the RMSD of p53$^{trans/cis}$ relative to an equilibrated structure of the complex for trans and cis isomers respectively. The width of the PMF is an indication of the range of conformational states. As one would expect, the PMFs for the unbound p53 (Figure 2.4 (b)(d)) are wider than the bound p53 (Figure 2.4 (a)(c)) for both trans and cis, indicating the larger conformational freedom of the unbound p53. The RMSD PMFs for the bound state p53^{trans} in Figure 2.4 (a) exhibit packed profiles overall, peaking sharply around 1 – 1.5 Å. On the other hand, the RMSD PMFs for the bound state p53^{cis} in Figure 2.4 (c) have more diverse profiles. For the unbound state the RMSD PMFs for p53^{cis} in Figure 2.4 (d) have broader profiles than p53^{trans} in Figure 2.4 (b) in general. These observations suggest a wider range of accessible conformational states is allowed for p53^{cis} compared to p53^{trans}. It is possible that broader PMFs would be seen if more than three trials of p53^{trans} were performed, however, we believe that this is not likely given the consistency of the p53^{trans} trials compared to p53^{cis}.

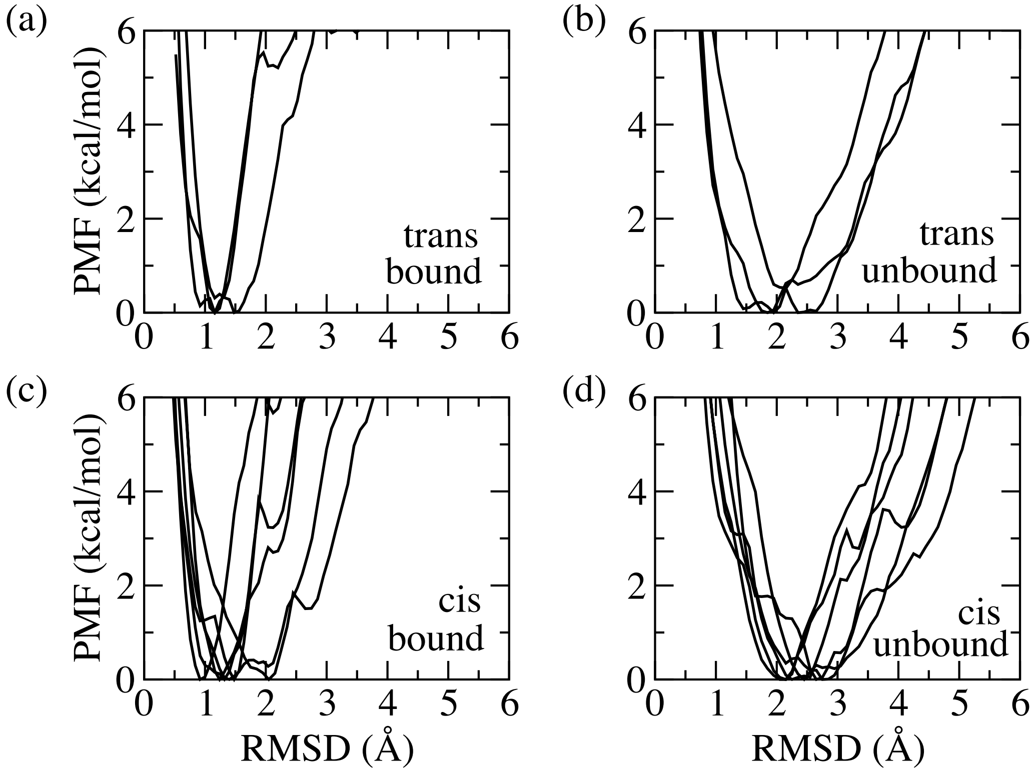

Figure 2.4: Conformational potential of mean force (PMF) results as a function of the root mean square distance (RMSD) in the bound and unbound states. Three independent trials were performed for the (a) bound state trans isomer and (b) for the unbound state trans isomer. Six independent trials were performed for the (c) bound state cis isomer and (d) unbound state cis isomer. Independent trials were initiated from the last frames of 100 ns simulations of the p53-MDM2 complex.

The free energy changes due to the restraints on the RMSD were calculated using Eq. 2.3. Taken together, $\Delta G^{u}_{conf}$and ΔG^{b}_{conf} determine the free energy cost associated with the loss of conformational freedom of p53 when adopting a specific conformation in the bound state. As shown in Table 2.1, this free energy cost, $\Delta G^{u}_{conf} + \Delta G^{b}_{conf}$, is calculated as 4.25 (1.29) kcal/mol for p53^{trans} and 6.56 (1.00) kcal/mol for p53^{cis}. As a result, the difference of the free energy cost due to conformational restraints between p53^{trans} and p53^{cis} is 2.31 (1.63) kcal/mol. This accounts for most of the 2.90 (1.28) kcal/mol difference between cis and trans ΔG_{bind} suggesting that the conformational flexibility is a major contributor to how p53^{cis} binds differently to MDM2 than p53^{trans}.

Contributions from axial and orientational restraints

Table 2.1 shows the free energy costs for restraining the axial and orientational degrees of freedom of p53(17-29) for both trans and cis conformations (see Appendix I for more detailed results). The axial restraint cost $\Delta G^{\mathrm{u}}_{\mathrm{axial}} + \Delta G^{\mathrm{b}}_{\mathrm{axial}}$ is 2.58 (0.25) kcal/mol for p53^{trans} and 3.05 (0.15) kcal/mol for p53^{cis} and suggests that axial degrees of freedom are not contributing much to the difference between cis and trans p53(17-29) binding to MDM2. Similarly, the orientational restraint cost $\Delta G^{\mathrm{u}}_{\mathrm{orient}} + \Delta G^{\mathrm{b}}_{\mathrm{orient}}$ is 5.24 (0.22) kcal/mol for p53^{trans} and 5.38 (0.09) kcal/mol for p53^{cis} suggesting that orientational degrees of freedom are also not contributing much to the difference between cis and trans p53(17-29) binding to MDM2.

Calculation of binding affinity

Table 2.1 shows the binding affinities for p53$^{trans/cis}$-MDM2, ΔG_{bind}, calculated using Eq. (2.1). The affinities are -11.83 (1.02) kcal/mol for p53^{trans}-MDM2 and -8.93 (0.78) kcal/mol for p53^{cis}-MDM2. The p53^{trans}-MDM2 affinity estimate compares favorably (within several kcal/mol) to the reported experimental result of -9.3 kcal/mol for peptide p53(16-29) binding with MDM2(17-125) (*100*). For comparison, we also provide the value calculated using the MM-PBSA/GBSA approach of -16.3 kcal/mol (*56*).

Equilibrium simulations

To gain insight into the structures and interactions for trans and cis equilibrium simulations were performed of the unbound p53(17-29) and p53(17-29)-MDM2 complexes (see Methods). The structures of trans and cis at L26-P27 in p53(17-29) stayed in trans and cis respectively during the course of all simulations for both bound and unbound p53(17-29). RMSD-based clustering analysis was performed on all of the bound state p53(17-29) structures using the VMD clustering plugin (*90, 101*) with cutoff 1.0 Å. Figure 2.5 shows representative structures from the largest clusters for trans (1,040 out of 15,000 structures) and cis (1,165 out of 30,000 structures) p53(17-29) while bound to MDM2. These structures are representative of the most common structures seen in the simulations. Figure 2.5 shows that the tail region (residues 27 to 29) of p53^{cis} in the complex tended to point away from the MDM2 binding pocket while the tail of p53^{trans} remained close to the binding pocket. This trend was observed for all clusters (data not shown). Unfortunately, there is no experimental evidence to support this conclusion, however, we believe that the 100 ns simulations are sufficiently long for the C-terminus to have found energetically favorable structures.

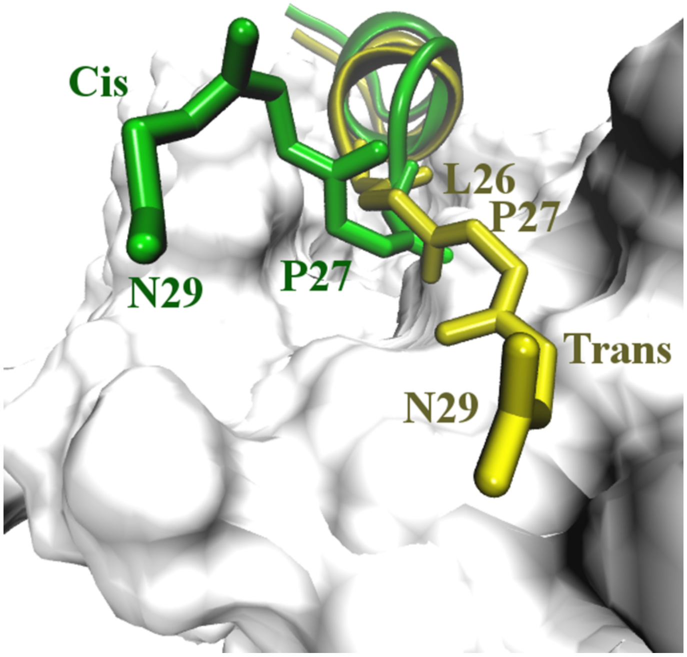

Figure 2.5: Most common trans (yellow) and cis (green) conformations of p53 while bound to MDM2 (white surface). These structures are representative of the largest cluster obtained from clustering analyses using all equilibrium simulations of the trans and cis complexes of lengths 300 ns and 600 ns respectively. Residues 26-29 of p53 are shown in licorice and the other p53 residues are shown in cartoon.

2.4 Discussion

Our results show that the binding affinity for p53^{cis}-MDM2 is around 2.90 (1.28) kcal/mol weaker than p53^{trans}-MDM2. Based on this binding affinity difference, for the bound state, the cis conformation of L26-P27 in p53(17-29) is estimated to be 0.8 % of the total population. NMR spectroscopy has shown that, for the unbound state, around 5.5 % of the L26-P27 peptide bonds in p53(1-63) are in the cis conformation (unpublished NMR data from Dr. François-Xavier Theillet).

The most significant contributor to the binding affinity difference between cis and trans is the free energy change due to the conformational restraints for unbound p53(17-29), ΔG^{u}_{conf}, that was found to be 2.61 (1.30) kcal/mol larger for cis than for trans. This term is the free energy cost associated with adding the conformational restraint on p53 in the unbound state. The larger cost for cis indicates that $p53^{cis}$ in the unbound state has more accessible conformations than unbound $p53^{trans}$. Figure 2.6 shows the backbone root mean square fluctuations (RMSF) for $p53^{trans}$ and $p53^{cis}$ in the unbound state and suggests that cis is more flexible than trans for the helical region (residues 19 to 24) and the region around the L26-P27 peptide bond (residues 26 and 27). Taken together, Figure 2.6 and the ΔG^{u}_{conf} values in Table 2.1 suggest that the primary reason that $p53^{cis}$ binds more weakly to MDM2 than $p53^{trans}$ is because the cis conformation of p53(17-29) is more flexible in the unbound state leading to a greater conformational entropy loss upon binding.

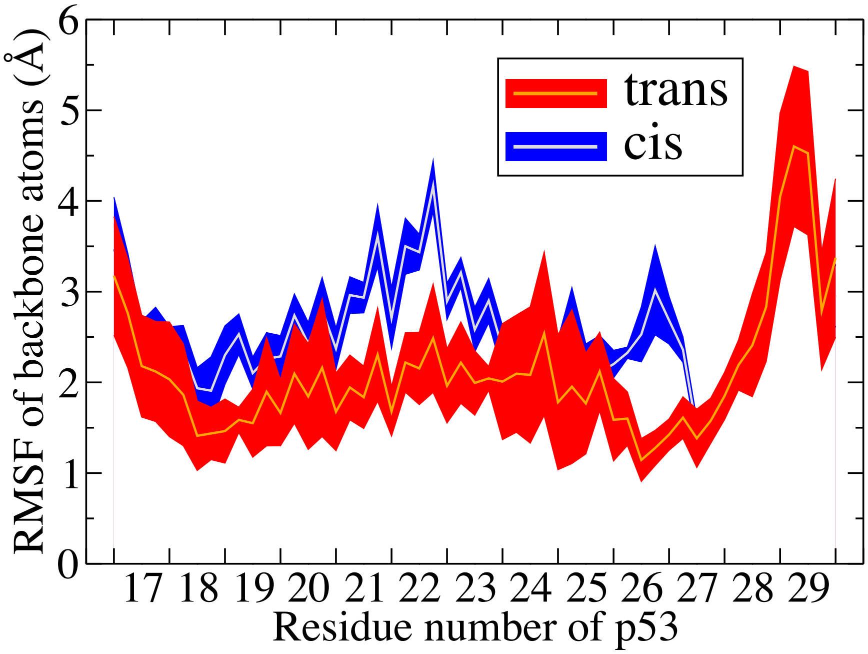

Figure 2.6: Backbone root mean square fluctuations (RMSF) for the unbound p53 fragment. Each curve is an average obtained from three independent 100 ns molecular dynamics simulations (last 50 ns used to compute averages) for cis (green line) and trans (cyan line) isomers. The shading shows the standard errors for the independent simulations.

To gain insight into how the interaction energies differ between p53^{trans}-MDM2 and p53^{cis}-MDM2, we studied intra- and inter-molecular interactions of p53$^{trans/cis}$-MDM2 from 300/600 ns molecular dynamic simulations. The intra-molecular interactions of bound and unbound p53 and MDM2 are indistinguishable between the trans and cis isomers. The inter-molecular interaction energy of p53^{trans}-MDM2 is stronger than that of p53^{cis}-MDM2 primarily due to the electrostatic interaction. (Shown in Figure 2.7) A detailed analysis of individual contributions to the electrostatic energy for every p53 residue (Figure 2.8) shows that the electrostatic interaction between N29 of p53^{trans} and MDM2 is ~ 33.2 kcal/mol stronger than for p53^{cis} and MDM2. Although the error bars are partially overlapping, the difference is appreciable. We think this difference is consistent with the observation that

N29 of p53^{cis} tends to point away from MDM2 (Figure 2.5). Note that both the protein fragments used in this study and the fragments used in the experimental comparison study in Table 2.1 were not capped, that is, the C-termini were negatively charged. The N-termini of the protein fragments used in this study were positively charged and the N-terminus of the fragments used in the experiments were negatively charged due to conjugating fluorescein isothiocyanate at the N-terminus (*100*).

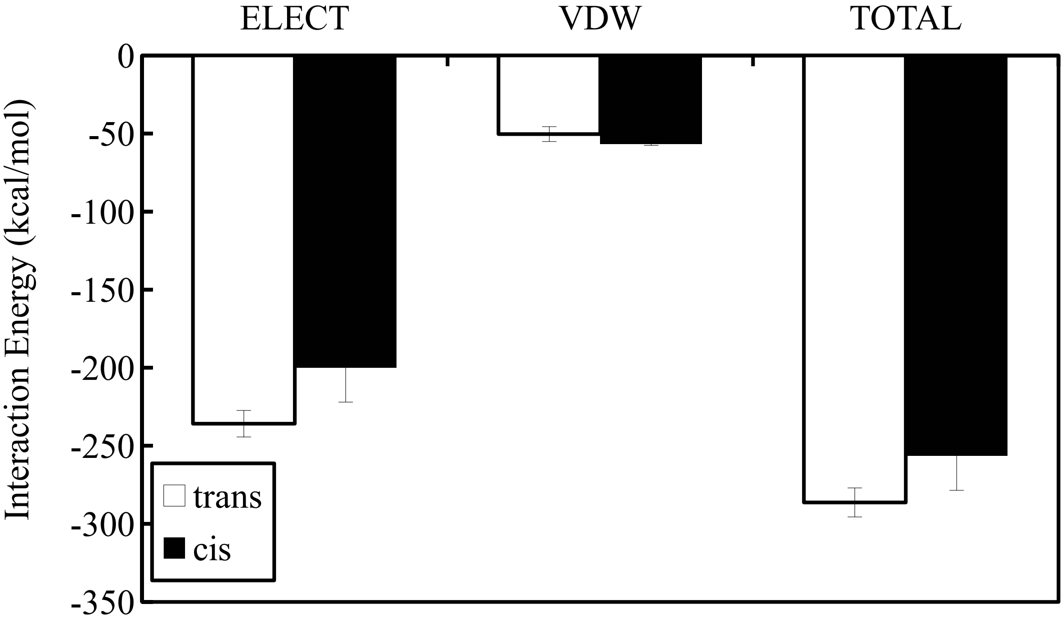

Figure 2.7: Van der Waals (VDW) and electrostatic (ELECT) interaction energies for the p53-MDM2 complex. Each bar is an average obtained from three/six independent 100 ns molecular dynamics simulations (last 50 ns used to compute averages) for the trans/cis isomers (white/black bars). The error bars are the standard errors for the independent simulations.

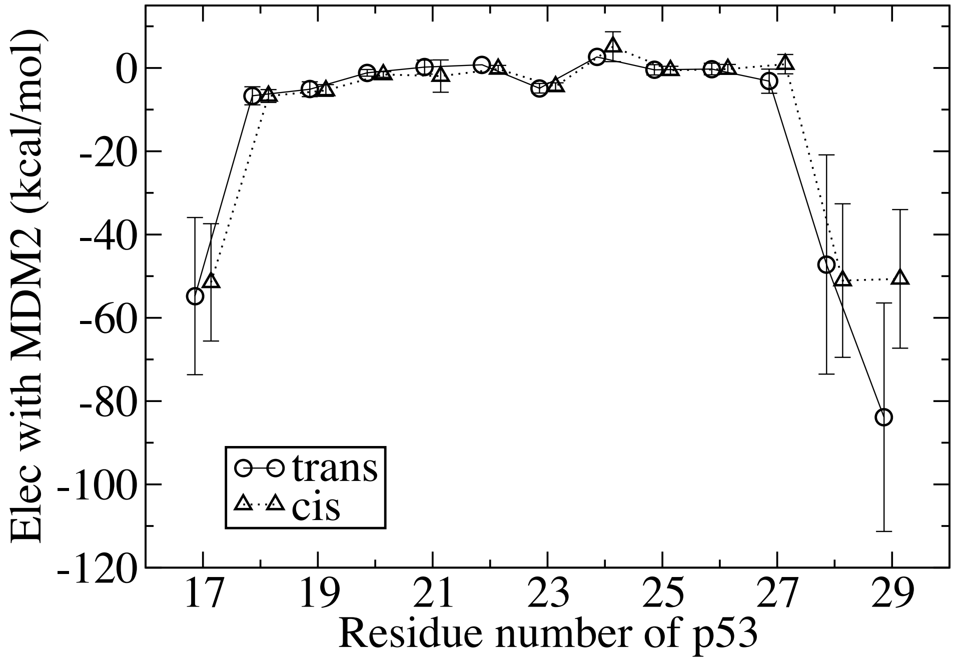

Figure 2.8: Electrostatic interaction energy between p53 and MDM2. Each data point is an average obtained from three/six independent 100 ns molecular dynamics simulations (last 50 ns used to compute averages) for the trans/cis isomers (solid/dashed lines). The error bars are the standard error for the independent simulations.

The stronger binding of p53^{trans} to MDM2 compared to p53^{cis} suggests a possible mechanism to help maintain a minimal level of p53 in unstressed cells. While the isomerization and binding rates of p53 in vitro have not been determined to our knowledge, it is known that the timescale for the conversion of cis to trans is minutes without any catalyzer at room temperature (*47, 102*), and that typical timescales for binding are in the range of nanoseconds to milliseconds (*103, 104*) .This separation of timescales, together with the NMR data (unpublished data from Dr. François-Xavier Theillet) and our binding affinity results suggest that up to 5.5 % of the p53 proteins in the cell may be essentially unavailable for binding MDM2 because they are in the cis conformation at that moment. Since MDM2 binding signals the ubiquitination process that leads to nuclear export and degradation of p53 (*81-83*), the preferential binding of the trans conformation may leave a small population of unbound p53 molecules that remain available to respond to cellular stress. Given the complex network of interactions in vivo, experimental evidence will be required to justify the above claim.

This study demonstrates that it is feasible to estimate absolute binding affinities that are in good agreement with experimental data for a system involving an IDP binding to an ordered protein. We found that it was necessary to restrain the axial, orientational, and conformational degrees of freedom to enhance the convergence of the simulations, as has been discussed previously (*72-77*). The final (unbiased) binding affinities were calculated by rigorously accounting for all the restraints used in the simulations. The affinities are estimated -11.83 (1.02) kcal/mol for p53^{trans}-MDM2 and -8.93 (0.78) kcal/mol for p53^{cis}-MDM2. The p53^{trans}-MDM2 affinity estimate compares favorably (within several kcal/mol) with the experimental binding affinity of -9.3 kcal/mol on a fragment of p53 (residues 16-29) binding to MDM2 (residues 17-125) (*100*). Our estimate is also similar to experimental affinities measured on the transactivation domain of p53, -8.9 kcal/mol (*105*), and full-length p53, -8.8 kcal/mol (*106*).

2.5 Conclusions

In this study we used a PMF approach to compute the absolute binding affinity for both trans and cis isomers of the L26-P27 peptide bond of p53(17-29) binding to MDM2. We find that the trans conformation of p53(17-29) binds more strongly by around 2 kcal/mol. Based on the binding affinity difference between p53^{trans}-MDM2 and p53^{cis}-MDM2, the cis isomer was estimated to be ~0.8 % of the total population. A more detailed analysis revealed that N29 at C-terminal of the trans isomer contributes to the binding by having stronger electrostatic attraction to MDM2 than the cis isomer. In addition, the cis isomer exhibits higher flexibility at unbound state and lowers the binding affinity. NMR spectroscopy showed that when p53(1-63) is unbound around 5.5 % of the L26-P27 peptide bonds are in the cis conformation (unpublished NMR data from Dr. François-Xavier Theillet). NMR could not determine the cis population for p53(1-63) when bound to MDM2 due to possible peak overlapping. The stronger binding of $p53^{trans}$ to MDM2 compared to $p53^{cis}$ may suggest a mechanism to help maintain minimal levels of p53 in unstressed cells, which could help for rapid response to cellular stress. Our results suggest that around 5.5 % of the p53 proteins in the cell will not be targeted for degradation because they are in cis and thus essentially unavailable for binding to MDM2. Finally, this study demonstrates that it is feasible to estimate absolute binding affinities that are in good agreement with experimental data for a system involving an IDP binding to an ordered protein, provided that restraints are used to enhance convergence of the simulations.

2.6 Acknowledgement

The authors would like to thank Dr. François-Xavier Theillet for providing NMR data and analysis (In-cell NMR Laboratory, Department of NMR-supported Structural Biology, Leibniz Institute of Molecular Pharmacology, Berlin, Germany). This project was supported by the National Institutes of Health (R21 GM83827 to FMY), the INBRE Program (NIH grant P20 GM103408), and the IBEST Computing Core (NIH grant P30 GM103324).

Chapter 3

Analysis of omega dihedrals in molecular recognition features in membrane proteins

Molecular recognition features (MoRFs) in membrane proteins are short intrinsically disordered proteins that become structured when they bind to their partners. MoRFs often initiate molecular recognition and in membrane proteins are responsible for a wide range of cellular functions. The role of peptide bond isomers (defined by the omega dihedral) in the binding of mpMoRFs to their partners is not well understood. In this study, we conducted a statistical survey of a database of mpMoRFs and analyzed peptide bonds. Analysis of amino acid composition showed that mpMoRFs consisted of both order- and disorder-promoting amino acids. It was also found that the peptide bonds for Xaa-Pro mpMoRFs are different from natively structured proteins. In mpMoRFs, only 0.11%/0.75% of peptide bonds are in cis for non-proline/proline, in contrast to natively structured proteins where 0.03%/5.2% are in cis for non-proline/proline. Predictions of proline peptide bond isomers were also performed and it was found that many proline bonds are predicted to be cis but actually found in trans. These results suggest that cis-trans isomerization plays an important role in mpMoRFs function. More studies are required to understand how cis-trans isomerization is important for mpMoRF function and to improve existing prediction algorithms for proline peptide bonds.

3.1 Introduction

Intrinsically disordered proteins (IDPs) and disordered regions in proteins do not form stable structures as do most proteins that are natively structured. This structural flexibility comes from their amino acid composition; the lack of hydrophobic amino acids means that IDPs do not form hydrophobic cores like natively structured proteins (*107, 108*). The net charge and proline content of a protein are positively correlated with the structural instability (*109*). IDPs are common from virus to vertebrates and play important roles in cell signaling and regulation (*32, 33, 110, 111*). Often a single IDP is able to bind many different partners and thus often act as hubs in biological interaction networks (*18, 111, 112*).

Molecular recognition features (MoRFs) are short, intrinsically disordered regions in proteins that undergo a disorder-to-order transition upon binding to their partners. Molecular recognition is the initial step of protein-protein interaction and subsequent biological functions and MoRFs play an important role in modulating the binding of IDPs and hence in regulating molecular recognition and cell signaling (*31, 111*). Upon binding with their partners MoRFs can form alpha helices, beta strands, and irregular structures with both alpha helices and beta strands.

Membrane proteins are vital in cell signaling (*35*). They are categorized into three classes depending on their positions relative to the membrane: transmembrane proteins that span across the lipid bilayer, integral monotopic proteins that are attached to one side of the membrane, and peripheral proteins that are temporarily associated to either the membrane or to the two other types of membrane proteins (*35, 36*). A previous study revealed that around 20% of the MoRFs containing proteins are transmembrane (*31*). Two other studies demonstrated that mpMoRFs have a distinct amino acid composition and that current MoRFs predictors are not very accurate for transmembrane MoRFs (*37, 38*).

Constrained by a partial double bond, a peptide bond can be in either the trans (omega ~ 180°) or cis (omega ~ 0°) conformation (also termed isomer) (Figure 1.3). Under certain conditions these conformations can interchange. The change between trans and cis conformations is called isomerization and plays an important role in a variety of biological processes. Many studies have stressed the importance of the cis-trans isomerization of peptide bonds for the protein folding processes (*8-11*). As more and more protein structures have been resolved it has become clear that peptide bonds in protein are predominantly found in trans (*3*). Cis isomers are more frequently associated with prolines in intrinsically disordered proteins (IDPs) as compared to natively structured proteins and a recent report revealed that IDPs tend to have more prolines in their sequences on average compared to natively structured proteins (*19*). We believe prolines may play multiple roles in IDPs since cis-trans isomerization in prolines have biological significance, and IDPs are highly populated with prolines (*113*). For example, it is known that the cis and trans proline isomers of the adaptor protein Crk turn the inhibitory functions on and off (*20*).

Interaction with other proteins is critical for the function of mpMoRFs, but it is not well understood how peptide bond isomerization might modify the binding of mpMoRFs to their partner proteins. In this study, we conducted a statistical survey of a database of mpMoRFs and analyzed the peptide bonds for all amino acid types. Analysis of amino acid composition showed that mpMoRFs consist of both order- and disorder-promoting amino acids. The peptide bonds for Xaa-Pro mpMoRFs were found to be distinct from natively structured proteins. In mpMoRFs, only 0.11%/0.75% of peptide bonds are in cis for non-proline/proline, in contrast to natively structured proteins where 0.03%/5.2% are in cis for non-proline/proline. Predictions of proline peptide bond isomers were performed and it was found that many proline bonds are predicted to be cis but actually found in trans. These results suggest that cis-trans isomerization plays an important role in mpMoRFs function. More studies will be required to understand how this isomerization relates to mpMoRF function and to improve algorithms to predict peptide bond isomers.

3.2 Methods

Dataset

The database mpMoRFsDB is a collection of molecular recognition features in membrane proteins. The current version of the mpMoRFsDB is 1.1 and it contains 172 proteins with 233 MoRFs. We extracted the structural files for the MoRFs in mpMoRFsDB from protein data bank (*114*) via the R package Bio3D (*115*).

To reduce the redundancy in sequences and keep the high quality structural files, we culled the protein structures solved by X-ray crystallography, with resolution better than 3.0 angstrom, R-factor < 0.25, and sequence percentage identity smaller than 30%. The resulted dataset is called non-redundant mpMoRFs or nrmpMoRFs containing 123 MoRFs from 102 proteins.

Software

Torsion angles were calculated using Bio3D (*115*) in R. Statistical analyses were carried out using the R stats package. Plots were created with ggplot2 in R. The omega value predictions were performed on the CisPEPred sever (*116*).

Relative difference

In order to compare the frequency of each residue between alpha MoRFs and irregular MoRFs, we normalized their occurrences to the count of that residue for the respective structures using the following expression

$$Freq_{alpha}^{aa} = \frac{Count_{aa}^{alpha}}{\sum_{aa} Count_{aa}^{alpha}}, Freq_{irr}^{aa} = \frac{Count_{aa}^{irr}}{\sum_{aa} Count_{aa}^{irr}}$$

where *aa* represents any amino acid, *alpha* is for alpha helix structures, *irr* is for irregular structures, and *Freq* is the normalized frequency.

Since the amino acids are not evenly distributed, in order to compare the frequency differences for different amino acids, $Freq_{alpha}^{aa} - Freq_{irr}^{aa}$, we calculated relative frequency differences (RF) as plotted in Figure 3.1 (d):

$$RF = \frac{Freq_{alpha}^{aa} - Freq_{irr}^{aa}}{Freq_{alpha}^{aa} + Freq_{irr}^{aa}}$$

RF values range from -1 to 1; a negative RF value for a residue indicates that this residue is more prevalent in irregular structures, and a positive RF indicates that this residue is found more frequently in alpha helices. If RF is close to zero, the residue is found with around the same probability in both alpha and irregular structures.

3.3 Results and discussion

Overview of the Dataset

Some basic information about the MoRFs pool is shown in the Table 3.1. From the mpMoRFs database we isolated non-redundant 123 MoRFs that have high-resolution crystal structures. This culled set of MoRFs is denoted as nrmpMoRFs (non-redundant membrane protein MoRFs). The MoRFs in our study belong to 102 proteins from 30 organisms. There are 61 MoRFs derived from Homo Sapiens, 13 from Mus Musculus, 10 from Saccharomyces Cerevisiae, and 8 from Rattus Norvegicus. Other organisms contribute less than 3 MoRFs to our dataset. As membrane proteins, they are classified into three types according to their relative location to the membrane: peripheral, single spanning, and multi spanning proteins, where the latter two are transmembrane proteins. We found 49 peripheral, 50 single spanning, and 24 multi spanning mpMoRFs in the dataset. Among the 61 nrmpMoRFs just from Homo Sapiens, there were 20 peripheral, 31 single spanning, and 10 multi spanning mpMoRFs. More than 50% of single spanning mpMoRFs were contributed by Homo Sapiens. Given that almost half of the dataset was derived from Homo Sapiens our results should not be considered general for all eukaryotes and prokaryotes.

Table 3.1: An overview of the culled mpMoRFs database.

		Occurrences
MoRFs		123
Protein		102
Protein Type	Peripheral	49
	Single Spanning	50
	Multi Spanning	24
Secondary Structure	Alpha	41
	Beta	5
	Complex	1
	Irregular	76
Organism		30

Number of MoRFs, and proteins they belong to, in the nrmpMoRFs dataset. MoRFs were categorized by protein types (peripheral, single spanning, and multi spanning membrane proteins) and secondary structures (alpha helix, beta strand, complex structure and irregular structure). The occurrence of MoRFs in the nrmpMoRFs dataset by each category is provided in this table. The number of organism that the MoRFs belong to was summarized.

The length of nrmpMoRFs ranges from 7 to 68 amino acids with median length of 24, however, not every MoRF has its full structure resolved by X-ray crystallography. Only considering nrmpMoRFs with fully-resolved structures, the length ranges from 3 to 66 amino acids with a median value of 17.

Secondary structure analysis

Secondary structure types are defined as alpha helix, beta strand, complex structures that contain both alpha and beta, and irregular structure that don’t contain any alpha or beta. We identified 41 alpha helices, 5 beta strands, 1 complex structure and 76 irregular structures (Table 3.1). The complex structure comes from chain P of 2OSL that contains two helices and two turns by STRIDE assignment (*117, 118*). There are 18 alpha helices and 31 irregular structures in the peripheral mpMoRFs, and 23 alpha helices, 5 beta strands, 1 complex structure and 45 irregular structures in the transmembrane MoRFs. For our non-redundant dataset, irregular structures are thus the majority, beta strands are rare and only found in transmembrane MoRFs, and complex structures are nearly absent. It is possible that no beta strand structures were found in peripheral membrane MoRFs due to the size of our dataset, but the result still indicates that the beta strand structure is not highly favored in nrmpMoRFs. A previous study on MoRFs, in agreement with our results, showed that the irregular structure is the most abundant type and beta strand is the least (*31, 38*).

Amino acid composition

In the dataset there are 2793 amino acids in total and they are not evenly distributed among the twenty types of amino acids. Figure 3.1 (a) shows the amino acid composition in the nrmpMoRFs dataset. Their occurrences are shown on the top of the bars. L (8.6%), A (8.0%), K (7.2%), E (7.1%), S (6.9%) and R (6.3%) are the top six most common residues, while W (1.2%), H (2.2%), C (2.6%), M (2.7%), Y (3.4%), and F (3.7%) are the least common residues.

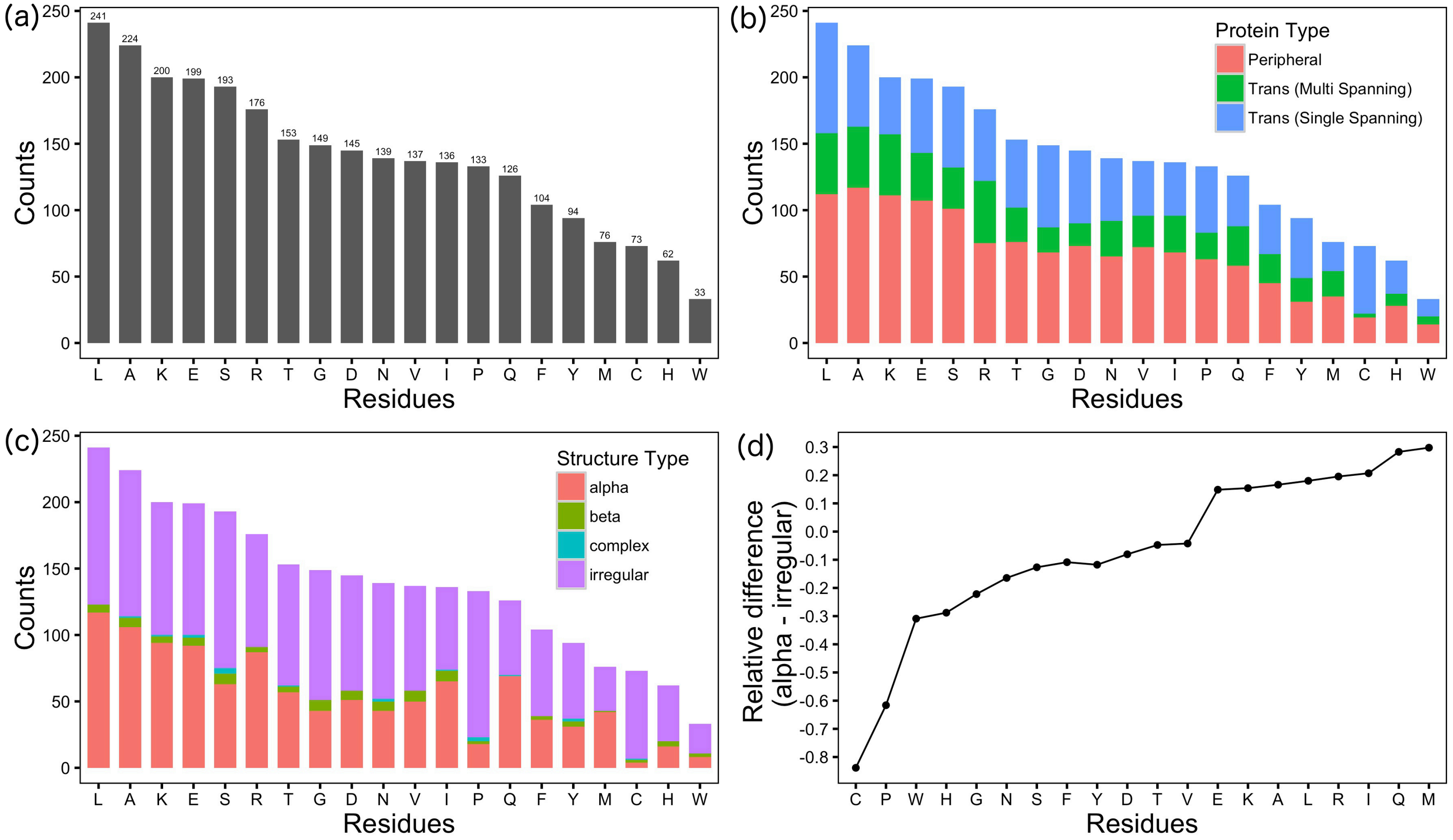

Figure 3.1: The occurrences of amino acids in the culled mpMoRFs dataset.

MoRFs are disordered when not bound with their partners but need to be able to fold upon binding. This property requires both disorder- and order-promoting residues in MoRFs and is consistent with our results from nrmpMoRFs (Figure 3.1). Previous studies have shown that A, R, G, Q, S, P, E and K are disorder-promoting residues and W, C, F, I, Y, V, L and N are order-promoting (*107, 108*). Among the top populated residues, A, K, E, S and R tend to break ordered structures, and residue L promotes structure. Other order-promoting residues W, C, Y, and F are not favored in the nrmpMoRFs dataset. Interestingly, all aromatic residues W, H, Y and F occur at low frequency in the dataset. When compared with the overall amino acid composition of proteins, an enrichment of charged residues (D, E, K and R) and a depletion of the most hydrophobic residues (I, V and L) were observed in mpMoRFs previously (*37*). However, inside the nrmpMoRFs dataset it is enriched in disorder-promoting residues and residue L is the major order-promoting residue that contributes to the folding and binding process.

Peripheral and transmembrane (multi-spanning and single-spanning) MoRFs have similar amino acid compositions as they do in combination. Subtle differences stem from the ranking on the most and least populated six residues separately. (Figure 3.1 (b)) Similar to the previous study, trans membrane MoRFs contain more C than peripheral MoRFs (*37*).

Figure 3.1 (c) shows the amino acid distributions in different secondary structures: alpha, beta, complex and irregular structures. There are not enough beta and complex structures to infer statistical impact for our dataset. Alpha helices and irregular structures have different amino acid compositions, e.g., P and C are more popular in irregular structures. To get a quantitative view of the differences between alpha helices and irregular structures, in Figure 3.1 (d) we plotted the relative difference (RF) for each residue which is defined in the method part. RF values range from -1 to 1. A negative RF value of a residue suggests this residue is more prevalent in irregular structures, and a positive RF means this residue is more in alpha helices. If RF is close to zero, the residue is similarly popular in alpha and irregular structures. Compared to other residues, C and P are predominantly found in irregular structures with RF values less than -0.6. The other residues have RFs in the range of -0.3 to 0.3. When we only consider the absolute relative difference, RF values of residues except C and P are averaged at 0.17 (± 0.08). Therefore, other than C and P, all residues have a similar distribution in alpha and irregular structures. It is also interesting that all aromatic residues have negative RF values suggesting that irregular structures have increased aromatic content. Cysteine (C) are thought to support ordered structures whereas P tends to disrupt ordered structure (*107, 108*). MoRFs in general possess a high content of C that is attributed to the formation of disulfide bonds (*31*). Aromatic residues are largely hydrophobic and are found buried in the core of structured proteins. In the case of MoRFs, it is believed that aromatic residues contribute to protein-protein interactions (*37*). Residue C and aromatic residues may help stabilize regions of irregular structures. Residue P disrupts the formation of the ordered structure and provide the protein an opportunity to more easily switch to a different structure. In this way, the complexity of irregular structures is maintained.

Cis/trans isomer statistics

By calculating the omega dihedrals we have determined the peptide bond isomers. When the value of an omega dihedral is close to 180° or -180°, it is trans conformation and it is cis conformation when its value is near 0°. As more and more protein structures have been resolved it has become clear that peptide bonds in protein are predominantly found in trans conformation (*3*). The cis conformation occurs with very small probability because there is a larger steric clash between atoms in the amino acids as compared to the trans conformation. In a nonredundant set of 571 proteins, a very small fraction (0.03%) of cis conformation was observed in Xaa-nonPro and this increased to 5.2% for Xaa-Pro (Xaa: any amino acid, nonPro: any amino acid but proline, Pro: proline) (*4, 5*). And cis-trans isomerization of Xaa-Pro bonds may play important roles in regulating transport channel opening and closing in cell membrane (*113, 119*).

Figure 3.2 shows the omega dihedral distribution for each residue in nrMoRFs. The number of counts for a specific omega value is coded by the ln(count). If the color is closer to purple in the rainbow (high frequency color) then the count is larger. As shown in Figure 3.2, high frequency colors tend to cluster around ±180°, indicating majority of peptide bonds are in trans that is consistent with previous findings (*3*). There are also some peptide bonds with omega values close to 0° and are thus in the cis conformation.

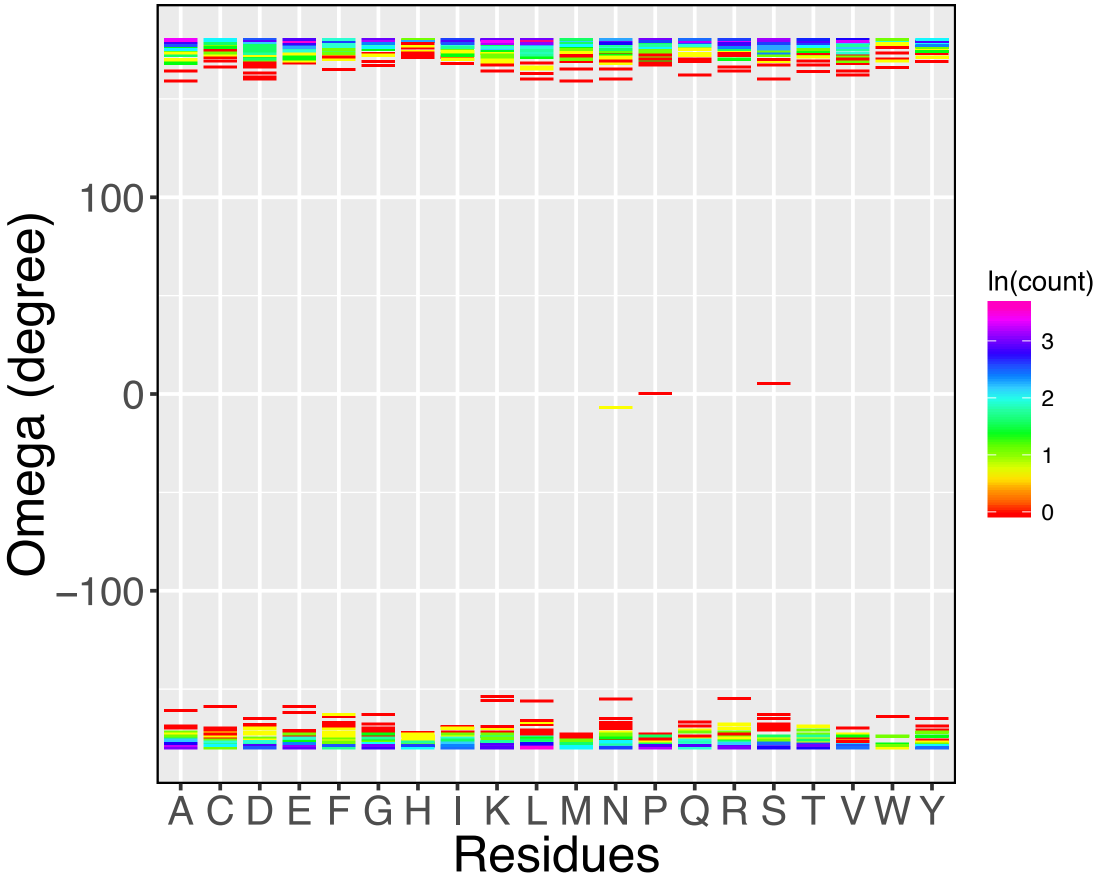

Figure 3.2: The distribution of omega dihedral values for each residue in the culled dataset. The rainbow color codes for ln(count), where count is the total number of occurrences of a residue for a specific omega value.

Figure 3.3 shows the distribution of peptide bond isomers for MoRFs in the nrmpMoRFs dataset. The plot shows that 99.86% of the peptide bonds are trans, 0.14% are cis. For cis, we found 2/1/1 peptide bonds from residues N/P/S. In our nrmpMoRFs set, a small fraction (0.11%) of cis conformation was observed in Xaa-nonPro and this slightly increased to 0.75% for Xaa-Pro. In overall, the population of trans conformation in nrmpMoRFs is similar to previous studies (*3, 4*). Xaa-Pro bonds in the nrmpMoRFs dataset present less cis conformations than the previous study (*4*) while Xaa-nonPro bonds have more cis conformations. The Xaa-nonPro bonds in our dataset come from residue N that does not have a regular secondary structure in the X-ray, and also from S that exhibits a turn structure. As pointed in previous studies (*120, 121*) on cis Xaa-nonPro bonds, nonPro peptide bonds do occur and may be more with higher resolution structures. Further investigation will be required to understand the roles of cis conformations found in this survey.

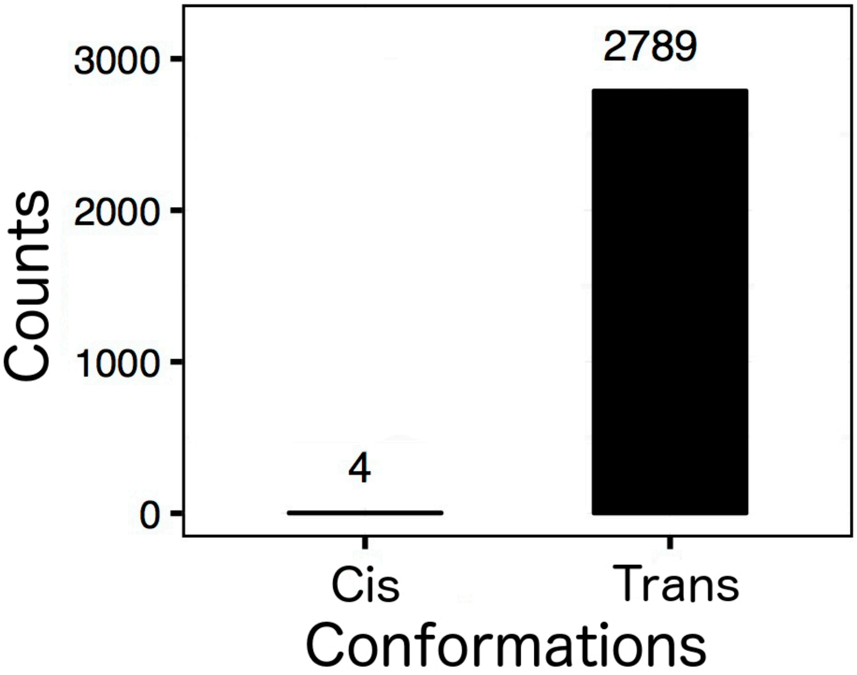

Figure 3.3: The occurrences of omega conformations for peptide bonds in the culled dataset. Number of occurrences is shown on the top of each conformation bar.

Deviations from a perfect cis/trans conformation

The perfect trans conformation is the planar form of a peptide bond with omega =180° and a perfect cis conformation is omega = 0°. With the development of techniques such as high-resolution NMR and crystallography, researchers have found that omega angles can deviate from their perfect value by as much as 24° (*122, 123*). Below, we discuss the distribution of trans omega angles for each residue and for all residues. This analysis is not completed for cis since there are only four observations. For the discussion below the omega value for a residue Xaa' represents the peptide bond angle for Xaa-Xaa'.

Figure 3.4 is the probability distribution for trans omega dihedrals of all peptide bonds and also for just proline peptide bonds. The figure shows that the distribution for all peptide bonds is broader than that of proline. This is consistent with the fact that proline residues have the constraint of a double bond that is not present for other residues.

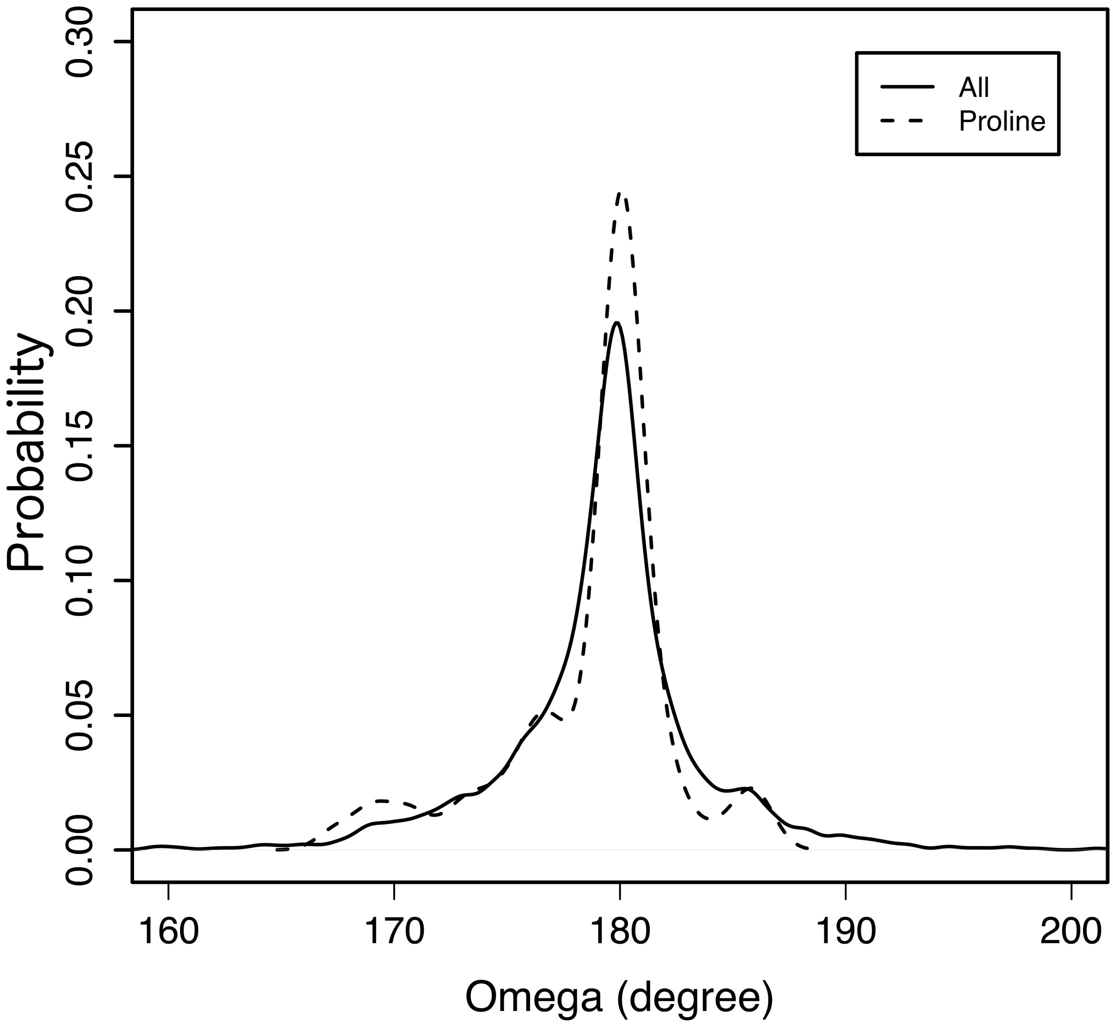

Figure 3.4: The probability distribution of trans omega values for all peptide bonds (solid) and Xaa-Pro bonds (dash).

Figure 3.5 shows a box plot of the trans omega dihedrals for individual residues. The width of box represents the sample size. The plot shows that omega values for some residues can deviate from 180° more than 20°. More detailed statistics of the trans omega angles for each residue are provided in Table 3.2. The omega value for proline is 178.8° with a standard deviation 3.8° and variance 14.1°. Compared with other residues, proline has smaller standard deviation and variance suggesting proline is relatively rigid when it is already in trans conformation in nrmpMoRFs.

Table 3.2: Statistics of trans omega values for each residue and all.

Residues	Mean	Median	Sd	Var	Min	Max
A	179.42	179.83	4.25	18.07	159.49	198.65
C	180.09	179.81	5.14	26.46	166.36	201.00
D	178.37	179.48	5.52	30.49	159.51	194.56
E	179.19	179.41	4.19	17.55	168.20	201.28
F	179.87	179.46	5.59	31.23	164.85	197.21
G	180.12	179.81	3.69	13.60	167.43	197.38
H	180.54	180.71	3.37	11.35	170.92	187.91
I	179.82	179.75	4.00	16.01	168.41	190.54
K	179.49	179.64	4.79	22.98	163.78	206.31
L	179.18	179.50	4.66	21.68	159.69	204.12
M	178.18	179.32	4.58	20.94	159.10	186.65
N	179.47	179.43	5.43	29.50	160.49	204.85
P	178.79	179.89	3.75	14.09	167.22	186.52
Q	179.89	179.64	3.98	15.82	162.30	193.14
R	179.54	179.49	4.31	18.57	164.46	205.40
S	179.23	179.69	4.35	18.94	160.09	197.37
T	179.77	179.96	4.05	16.44	163.97	191.25
V	178.46	179.33	4.18	17.44	161.88	189.54
W	179.17	179.42	5.53	30.58	166.04	196.10
Y	179.61	179.45	4.06	16.50	168.83	194.60
All	179.38	179.65	4.49	20.12	159.10	206.31

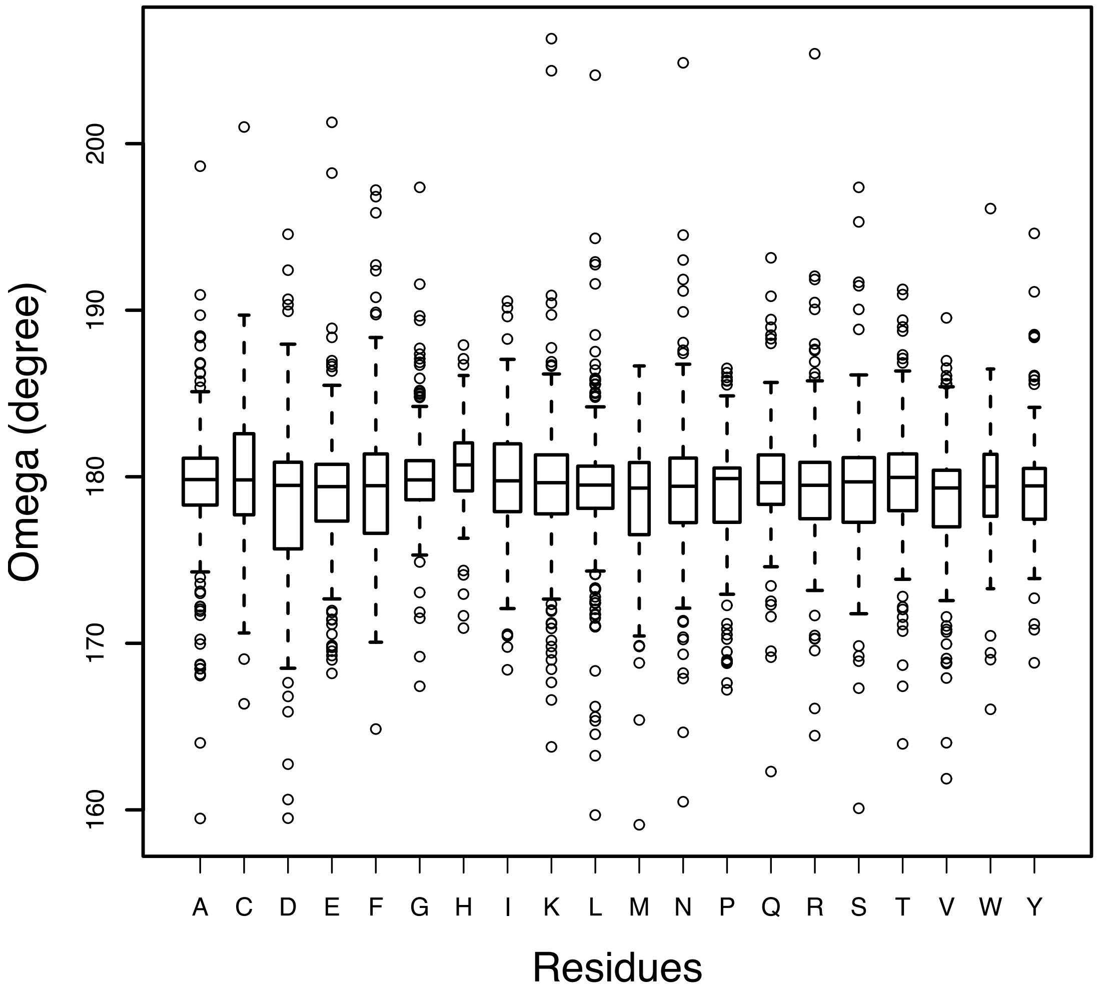

Figure 3.5: The box plot of trans omega values for each residue. The bottom and the top of the box are the first and third quartiles. The band inside the box is the second quartile (the median). The ends of the whiskers represent 1.5 interquartile range of the lower/upper quartile. The width of box indicates the relative the sample size. The circles are outliers.

Ramachandran plot

A Ramachandran plot shows the backbone conformations of proteins in (phi, psi) space (see Figure 1.4). Figure 3.6 shows a Ramachandran plots for the nrmpMoRFs dataset and for natively structured proteins. Figure 3.6 (a) is the Ramachandran plot built from proteins in the Top500 database (*124*). This database contains 500 high resolution, low homology, and high quality protein structures. It has more than 100 thousand residues. Our data are presented in Figure 3.6 (b). As shown in the figure, the Ramachandran plot for nrmpMoRFs is similar to that for structured proteins (*39, 124*). This suggests that after mpMoRFs fold, their backbone conformations converge to structured proteins. One difference is that beta strand (in the black box) structures are not that populated in mpMoRFs compared with structured proteins. It indicates that beta strands are not favored for molecular recognition in mpMoRFs that fits the statistical analysis on the beta strands in nrmpMoRFs above and a previous study on MoRFs in general (*31*).

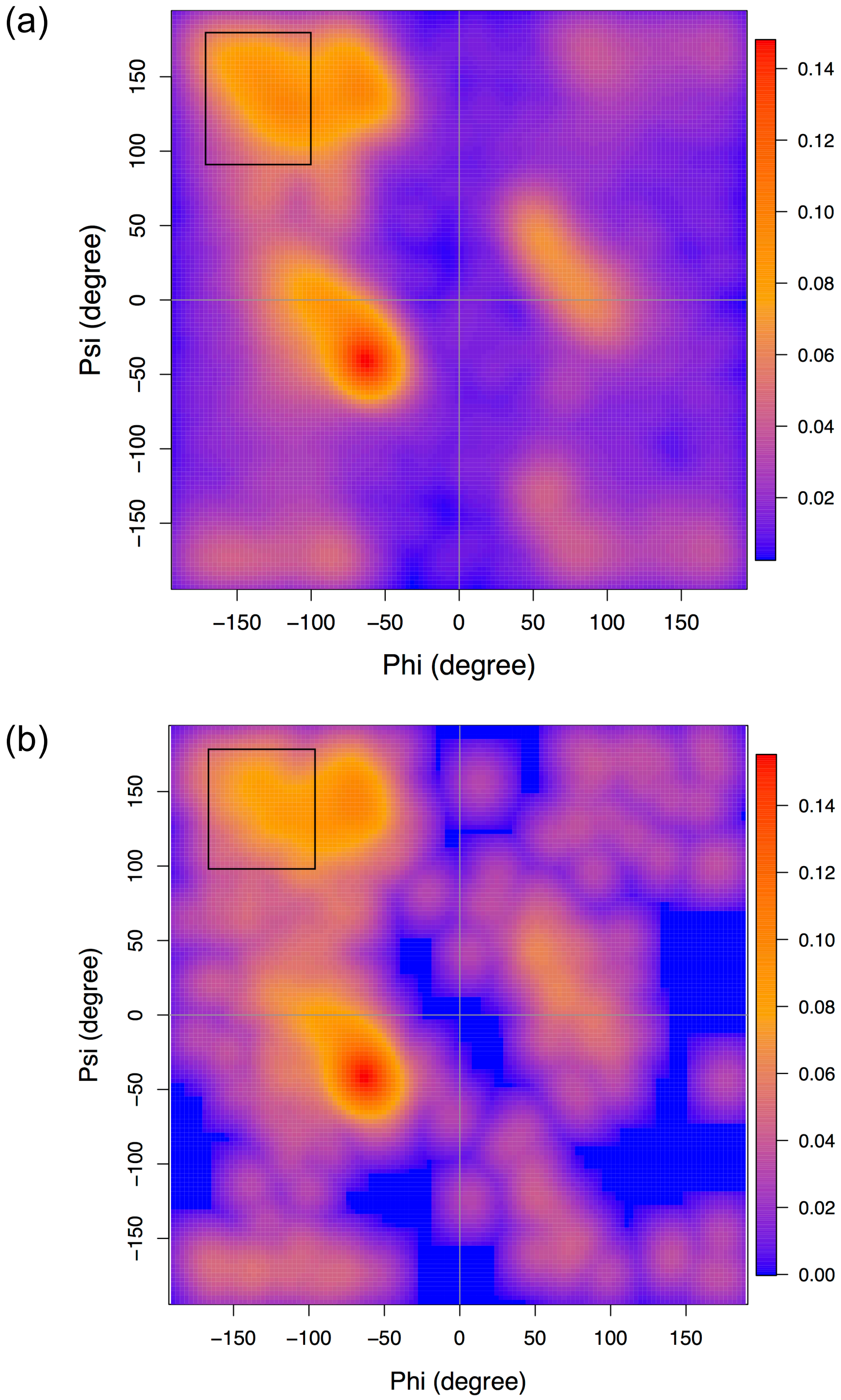

Figure 3.6: Ramachandran plots for (a) the structured proteins and (b) the culled mpMoRFs database. Data in case (a) was from Top500 database (*124*). Case (b) represents the (phi, psi) space of the culled mpMoRFs dataset in this study. More red the color, more populated the region on the map. Blue color means no such conformation was detected. Black boxes enclose approximate regions for beta strands (*7*).

Individual Ramachandran plots for each of the 20 amino acids are shown in Figure 3.7 for the nrmpMoRFs dataset and for natively structured proteins from Top500 database (*124*). Every plot contains contour lines for omegas greater (red) or smaller (green) than 180°. Notice that most amino acids have two distinct maxima in the beta strand region (upper left) and alpha helix region (lower left). Residue G has the most complicated plot followed by residues N and H. Glycine (G) is the smallest of the 20 amino acids. Without any restraint from the side chain, glycine has the most flexible (phi, psi). Asparagine (N) and histidine (H) have some left hand helix structures (upper right). A previous study observed the similar phenomenon for residue N, and explained it as the result of terminating proteins (*125*). However, asparagine does not dominate either terminus of proteins in general according to a survey on protein termini (*126*) so being at a terminus does not explain the emergency of left hand helices for asparagine. For mpMoRFs, histidine is more sporadic than in the case of structured proteins (*125*). It is also interesting to notice that (phi, psi) are more diffusive in the case of omegas greater than 180° for all residues except lysine (K) for mpMoRFs but it is not observed for structured proteins.

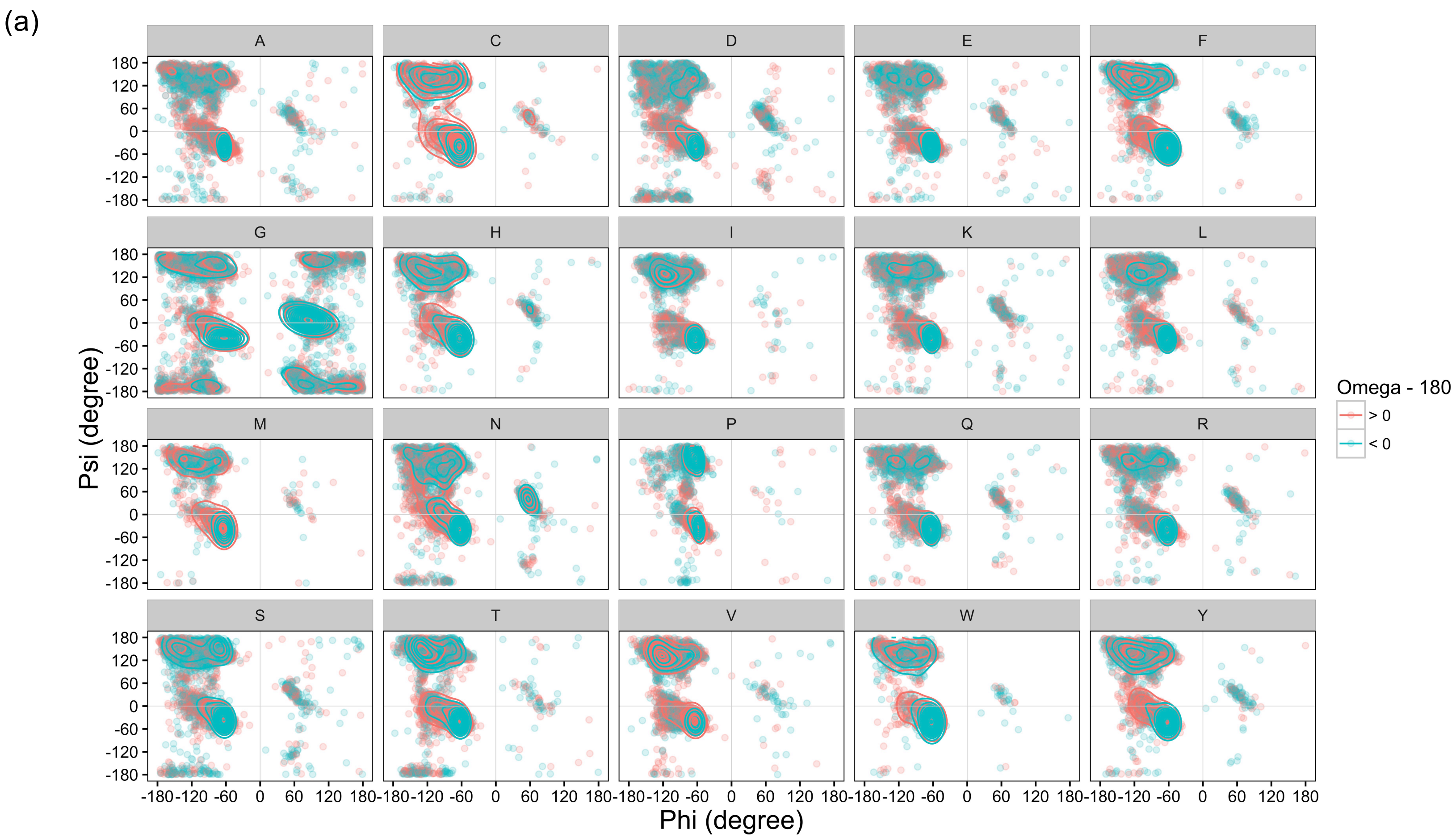

Figure 3.7: Ramachadran plots for 20 residues with distinct sets of omega values in cases of (a) structured proteins and (b) the culled mpMoRFs dataset. Contour lines are on top of scattered data points. Red color is for the omega values greater than 180° and green color is for the omega values smaller than 180°.

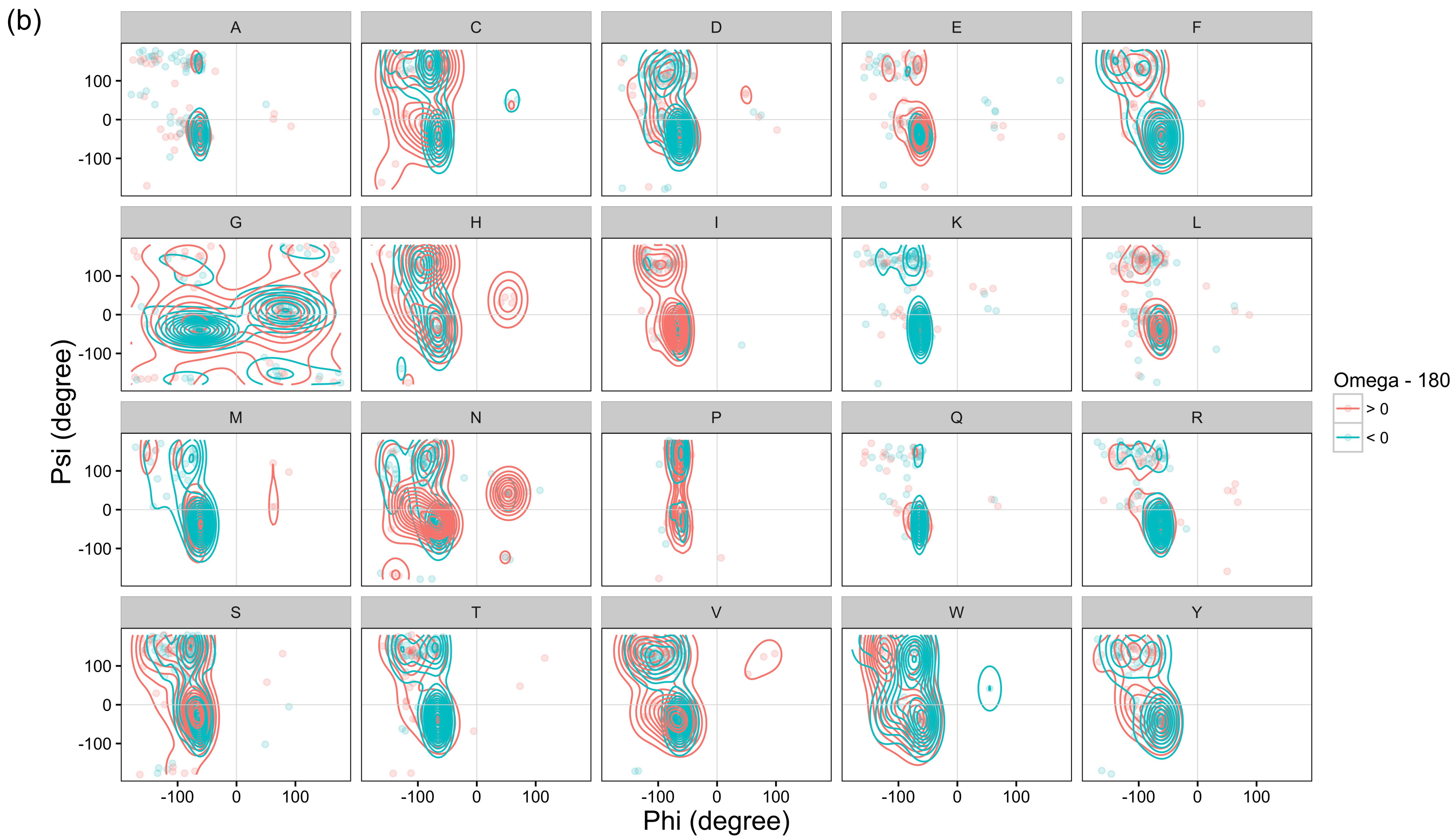

Figure 3.7: Ramachadran plots for 20 residues with distinct sets of omega values in cases of (a) structured proteins and (b) the culled mpMoRFs dataset. Contour lines are on top of scattered data points. Red color is for the omega values greater than 180° and green color is for the omega values smaller than 180°.

Cis/trans isomer prediction

Figure 3.8 shows predictions of the isomers for omega dihedrals in the nrmpMoRFs dataset using the online software CISPEPpred (*116*). There are 21 Xaa-Pro bonds predicted to be cis that are actually found in trans. Seven Pro-Pro bonds are predicted to be cis and this prediction exceeds all other nonPro-Pro bonds. Only one Xaa-Pro bond was predicted to be trans but was actually in cis. This cis proline bond was found in the sequence of APD, where the A-P (Ala-Pro) bond was predicted to be trans. The predicted cis population of Xaa-Pro peptide bonds is 15.8% that is much higher than a general case (5%) (*4*). This result suggests that proline, especially Pro-Pro bonds, may have a different structural preference in mpMoRFs than in the training dataset used to generate prediction by CISPEPpred. Larger training set including more mpMoRFs or specialized training dataset for mpMoRFs may be required to produce more reliable predictions on peptide bond conformations of mpMoRFs.

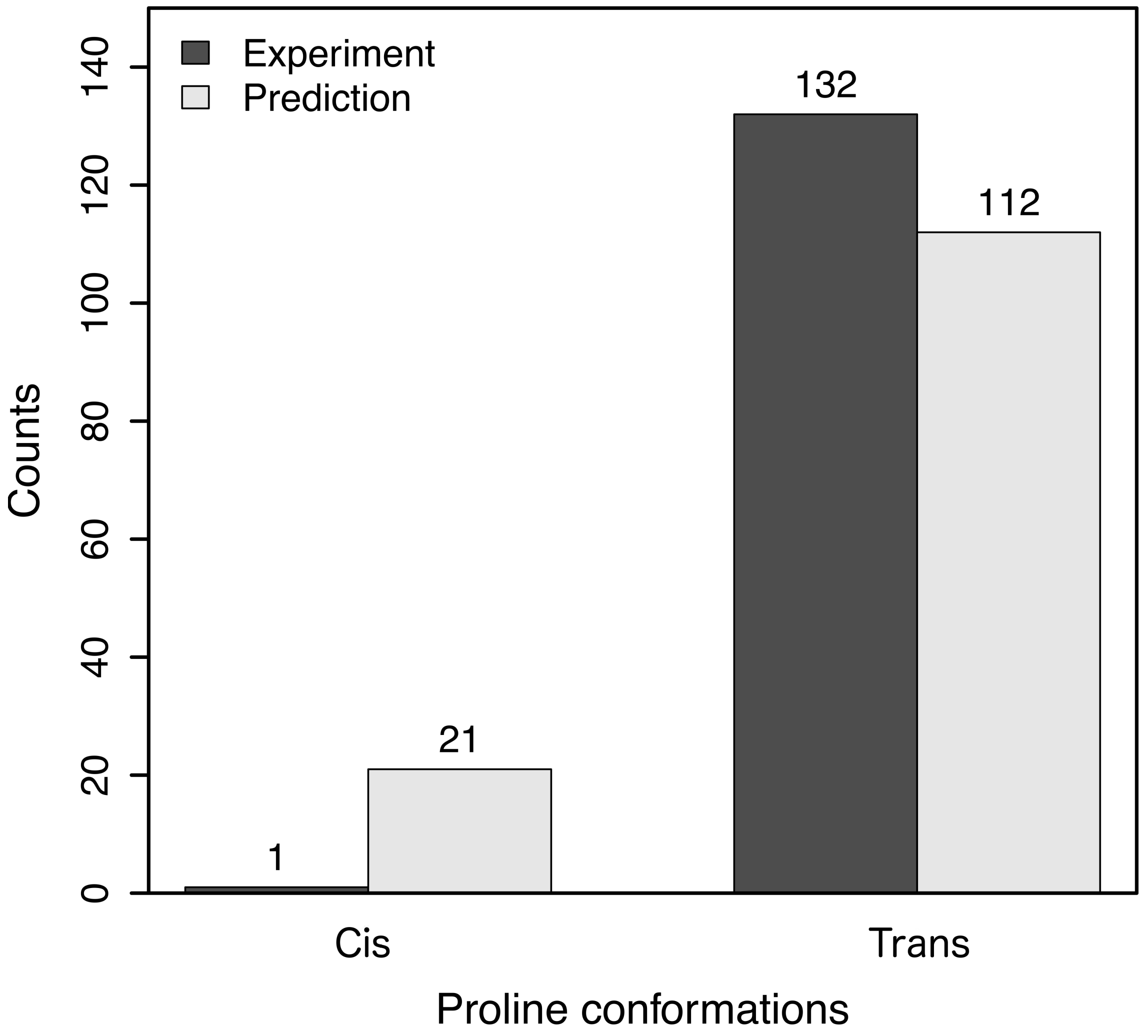

Figure 3.8: The occurrences of omega conformations for Xaa-Pro peptide bonds from experiments and predictions. Experimental values were calculated from crystal structures. Predictions were performed by CISPEPpred. Number of occurrences is shown on the top of each conformation bar.

Effect of neighboring residues

Figure 3.9 shows the normalized frequency for every neighboring residue of proline. The nearest neighbors are positioned in a sequence with proline as follows: ZaaXaaProYaa. To account for the bias from the non-uniform distribution of residues in the dataset, we normalized the frequency of the residues to their respective occurrences. This figure provides the information for the neighbor preference of the proline in mpMoRFs. Residues P, S, T, V and C are the top 5 residues that were found following a proline. Residues T, P, N, C and L are the top 5 residues that are likely to precede a proline. Residues W, N, Y, T and D (C is very close to D) are the top 5 residues that were found at the next nearest neighbor position, Zaa. Every residue has its own preference on the position relative to a proline. Proline is often in conjunction with the other proline. Asparagine (N) prone to precede proline one or two residues. Cysteine (C) were found almost equally at the three neighboring site. It is interesting that tryptophan (W) was never found preceding a proline but it favors the next nearest neighbor position prior to a proline.

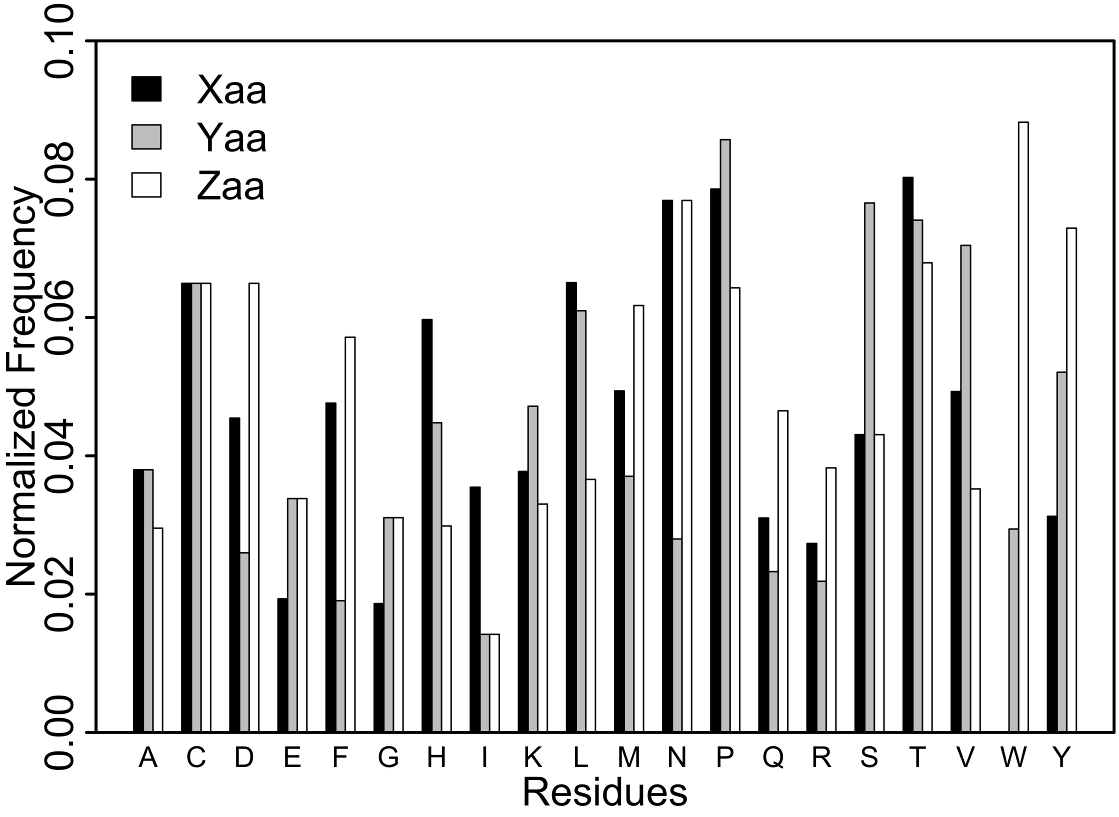

Figure 3.9: The distribution of the nearest neighbors of prolines in the culled dataset. The relation between the neighbors and proline is shown as the following: ZaaXaaProYaa. (aa: any amino acid)

Figure 3.10 shows the normalized frequencies of the nearest neighbors of proline that were predicted to have cis peptide bonds. Aside from Pro-Pro bonds, Xaa-Pro bonds followed by residue Y or F are more likely to be predicted as cis. Zaa may also exert some influence on the prediction. Also, the Xaa-Pro peptide bond in the sequence of ProXaaPro is more likely to be predicted as cis. These apparent biases may be true for proteins in the training set of CISPEPpred, but they do not appear to apply to mpMoRFs. Removal of these biases will be important for future peptide bond prediction algorithms.

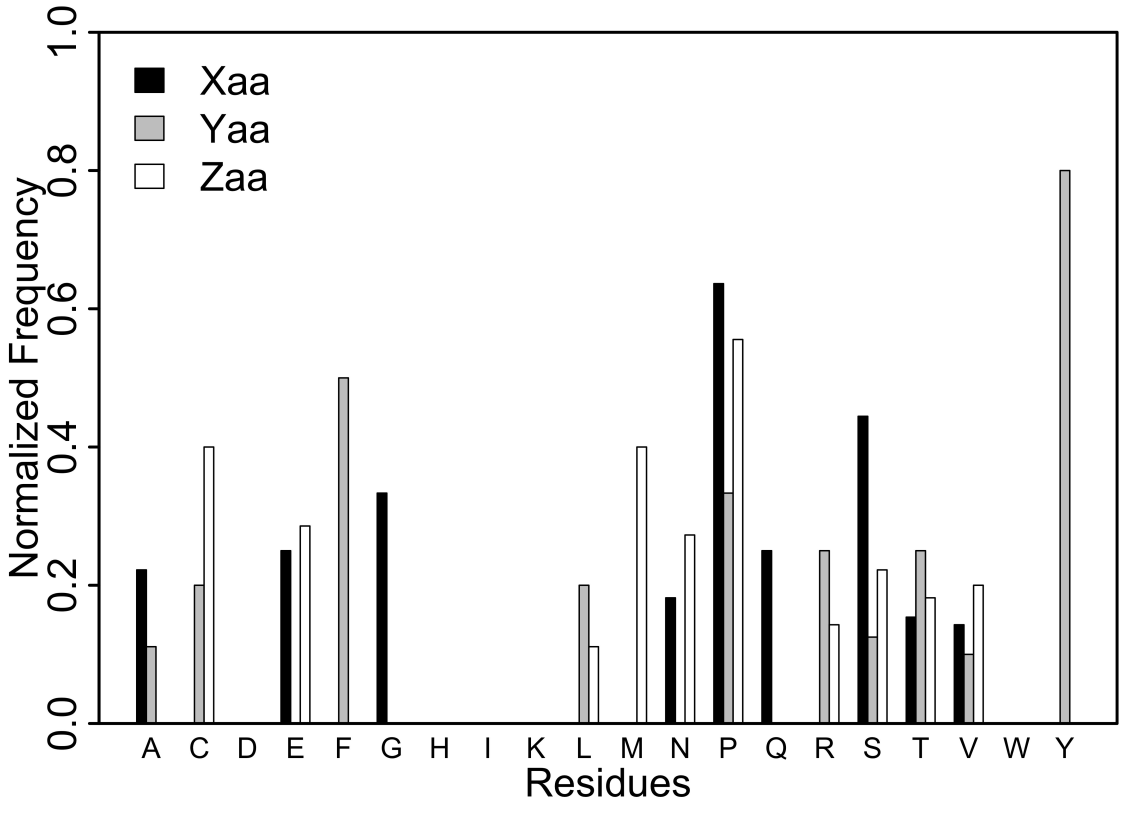

Figure 3.10: The distribution of the nearest neighbors of prolines that have Xaa-Pro bonds predicted to be cis. The relation between the neighbors and proline is shown as the following: ZaaXaaProZaa. (aa: any amino acid)

3.4 Conclusions

In this chapter, we studied the amino acid composition and peptide bond statistics in a database of MoRFs for membrane proteins. Analysis of amino acid composition showed that mpMoRFs consist both order- and disorder-promoting amino acids. This property ensures that mpMoRFs do not fold stably when unbound but can undergo disorder-to-order transition upon binding with their partners. The peptide bonds for Xaa-Pro mpMoRFs are distinct from natively structured proteins. In mpMoRFs, only 0.11%/0.75% of peptide bonds are in cis for non-proline/proline, in contrast to natively structured proteins where 0.03%/5.2% are in cis for non-proline/proline. Predictions of proline peptide bonds were performed and it was found that many proline bonds are predicted to be cis but actually found in trans. These results suggest that cis-trans isomerization plays an important role for mpMoRFs function. More studies will be required to understand how cis-trans isomerization relates to mpMoRFs function and to improve existing prediction algorithms for proline peptide bonds.

Chapter 4

Conclusions

Due to their disordered nature, many IDPs are capable of folding into different shapes upon binding with their partners and thus participate in a wide range of biological functions (*18, 127*). Cis-trans isomerization has been found to play critical roles in protein folding (*8-11*), cell signaling (*12, 13*), ion channel gating (*14*), and gene expression (*15*). MoRFs are a class of IDPs that undergo a disorder-to-order transition upon binding to their partners. They are thought to be the initial step in molecular recognition, which is important for subsequent biological processes. However, there have been very few studies to investigate the functional implications of proline cis-trans isomerization in the binding of IDPs to date. With our studies, we have gained a better understanding on the roles of cis-trans isomerization in IDP binding.

In this thesis, we have used computer simulations to study how the cis-trans isomerization of a proline in an intrinsically disordered region of p53 modifies binding to its partner MDM2, and we have conducted statistical analyses of cis and trans conformations in a MoRFs database for membrane proteins.

In chapter 2, we hypothesized that the cis conformation of proline in p53 would bind more weakly to MDM2 than the trans conformation. We used computer simulations to test this hypothesis by calculating the absolute binding affinity between p53 and MDM2 for both cis and trans isomers of the p53 proline in position 27. Results showed that the cis isomer of p53(17-29) binds more weakly to MDM2 than the trans isomer, and that this is primarily due to the difference in the free energy cost associated with the loss of conformational entropy of p53(17-29) when it binds to MDM2. The stronger binding of trans p53(17-29) to MDM2 compared to cis may leave a minimal level of p53 available to respond to cellular stress. The population of cis p53(17-29) was estimated to be 0.8% of the total population in the bound state. This study also demonstrates that it is feasible to estimate the absolute binding affinity for an intrinsically disordered protein fragment binding to an ordered protein that are in good agreement with experimental results.

In chapter 3, we hypothesize that cis-trans isomerization plays an important role in the function of MoRFs in membrane proteins. To test this hypothesis, a statistical survey was conducted to analyze the cis and trans dihedrals in mpMoRFs. Analysis of amino acid composition showed that mpMoRFs consists of both order- and disorder-promoting amino acids. This property supports the flexible structures of mpMoRFs and at the same time retains their ability to form stable structures upon binding. It was also found that the peptide bonds for Xaa-Pro mpMoRFs are different from natively structured proteins. In mpMoRFs, only 0.11%/0.75% of peptide bonds are in cis for non-proline/proline, in contrast to natively structured proteins where 0.03%/5.2% are in cis for non-proline/proline. Predictions of proline peptide bonds were performed and it was found that many proline bonds are predicted to be cis but actually found in trans. These results suggest that, consistent with our hypothesis, cis-trans isomerization plays an important role in mpMoRFs function. More studies are required to understand how cis-trans isomerization is important for mpMoRF function and to improve current prediction algorithms for proline peptide bonds.

A common observation in our studies is that the trans conformation of the proline peptide bonds is almost always seen in complexes between MoRFs and their protein partners. This finding may be indicative that cis-trans isomerization plays a role in attenuating MoRF binding.

This thesis paves the way for several future projects. As mentioned in chapter 3, some proline peptide bonds are predicted to be cis but are in fact found in the trans isomer. To improve prediction on proline peptide bonds in mpMoRFs, a machine learning algorithm can be developed by dividing the culled database created by this study into training sets and prediction sets. Since the database is small, building a larger MoRFs database of this kind will further help improve the prediction accuracy. Chapter 3 also briefly touched the neighboring effects of proline. More analyses relating neighbor residues to phi, psi and omega values can be conducted. It may provide valuable information on how neighbor residues are associated with the backbone and peptide bond conformation. In addition, extensive molecular dynamic simulations may be involved to study how the cis-trans isomerization affects the binding affinity of those MoRFs that have diverged predicted and actual conformations in proline peptide bonds. Also, there are several non-proline peptide bonds found in cis. It is interesting to dig deep on those cases by computer simulations. To understand the biological impact of cit-trans isomerization, peptide bonds locked in cis or trans can be invented and incorporated into biological systems to test the consequences.

Appendix I

Contributions to binding affinity calculation for MDM2 and p53^{trans}.

Component	trans1	trans2	trans3	Mean	Error
ΔG_{bind}^{res}	-24.15	-22.33	-25.20	-23.89	0.84
ΔG_{conf}^{u}	10.01	13.27	10.54	11.28	1.01
ΔG_{axial}^{u}	3.48	3.48	3.47	3.48	0.00
ΔG_{orient}^{u}	6.11	5.96	5.84	5.97	0.08
ΔG_{conf}^{b}	-6.07	-8.62	-6.39	-7.03	0.80
ΔG_{axial}^{b}	-0.39	-1.18	-1.12	-0.90	0.25
ΔG_{orient}^{b}	-1.13	-0.51	-0.55	0.73	0.20
ΔG_{bind}	-12.14	-9.92	-13.41	-11.83	1.02

The values are all in kcal/mol.

References

1. J. M. Berg, J. L. Tymoczko, L. Stryer, in *Biochemistry*. (W H Freeman, New York, 2002), chap. 3.
2. G. A. Petsko, D. Ringe, *Protein Structure and Function*. (New Science Press, 2004), vol. 3, pp. 195.
3. M. S. Weiss, A. Jabs, R. Hilgenfeld, Peptide bonds revisited. *Nature Structural & Molecular Biology* **5**, 676-676 (1998).
4. D. E. Stewart, A. Sarkar, J. E. Wampler, Occurrence and role of cis peptide bonds in protein structures. *Journal of molecular biology* **214**, 253-260 (1990).
5. D. P. Raleigh, P. A. Evans, M. Pitkeathly, C. M. Dobson, A peptide model for proline isomerism in the unfolded state of staphylococcal nuclease. *Journal of molecular biology* **228**, 338-342 (1992).
6. J. S. Richardson, The anatomy and taxonomy of protein structure. *Adv Protein Chem* **34**, 167-339 (1981).
7. G. N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95-99 (1963).
8. F. X. Schmid, R. L. Baldwin, Acid catalysis of the formation of the slow-folding species of RNase A: evidence that the reaction is proline isomerization. *Proceedings of the National Academy of Sciences of the United States of America* **75**, 4764-4768 (1978).
9. M. Levitt, Effect of proline residues on protein folding. *Journal of molecular biology* **145**, 251-263 (1981).
10. T. Kiefhaber, H. P. Grunert, U. Hahn, F. X. Schmid, Replacement of a cis proline simplifies the mechanism of ribonuclease T1 folding. *Biochemistry* **29**, 6475-6480 (1990).
11. W. J. Wedemeyer, E. Welker, H. A. Scheraga, Proline cis-trans isomerization and protein folding. *Biochemistry* **41**, 14637-14644 (2002).
12. G. Wulf, G. Finn, F. Suizu, K. P. Lu, Phosphorylation-specific prolyl isomerization: is there an underlying theme? *Nature cell biology* **7**, 435-441 (2005).
13. P. Sarkar, C. Reichman, T. Saleh, R. B. Birge, C. G. Kalodimos, Proline cis-trans isomerization controls autoinhibition of a signaling protein. *Molecular cell* **25**, 413-426 (2007).
14. S. C. R. Lummis *et al.*, Cis-trans isomerization at a proline opens the pore of a neurotransmitter-gated ion channel. *Nature* **438**, 248-252 (2005).
15. C. J. Nelson, H. Santos-Rosa, T. Kouzarides, Proline isomerization of histone H3 regulates lysine methylation and gene expression. *Cell* **126**, 905-916 (2006).
16. S. Lorenzen, B. Peters, A. Goede, R. Preissner, C. Frömmel, Conservation of cis prolyl bonds in proteins during evolution. *Proteins* **58**, 589-595 (2005).
17. A. P. Joseph, N. Srinivasan, A. G. de Brevern, Cis-trans peptide variations in structurally similar proteins. *Amino acids* **43**, 1369-1381 (2012).
18. H. J. Dyson, P. E. Wright, Intrinsically unstructured proteins and their functions. *Nature reviews. Molecular cell biology* **6**, 197-208 (2005).

19. F.-X. Theillet *et al.*, The alphabet of intrinsic disorder I. Act like a Pro: on the abundance and roles of proline residues in intrinsically disordered proteins. *Intrinsically Disordered Proteins* **1**, e24360 (2013).

20. P. Sarkar, T. Saleh, S.-R. Tzeng, R. B. Birge, C. G. Kalodimos, Structural basis for regulation of the Crk signaling protein by a proline switch. *Nature chemical biology* **7**, 51-57 (2011).

21. W. S. el-Deiry, Regulation of p53 downstream genes. *Seminars in cancer biology* **8**, 345-357 (1998).

22. J. Momand, H. H. Wu, G. Dasgupta, MDM2--master regulator of the p53 tumor suppressor protein. *Gene* **242**, 15-29 (2000).

23. P. H. Kussie *et al.*, Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science (New York, N.Y.)* **274**, 948-953 (1996).

24. J. Chen, V. Marechal, A. J. Levine, Mapping of the p53 and mdm-2 interaction domains. *Molecular and cellular biology* **13**, 4107-4114 (1993).

25. S. Bell, C. Klein, L. Müller, S. Hansen, J. Buchner, p53 contains large unstructured regions in its native state. *Journal of molecular biology* **322**, 917-927 (2002).

26. J. D. Oliner *et al.*, Oncoprotein MDM2 conceals the activation domain of tumour suppressor p53. *Nature* **362**, 857-860 (1993).

27. A. Böttger *et al.*, Molecular characterization of the hdm2-p53 interaction. *Journal of molecular biology* **269**, 744-756 (1997).

28. H. Lee *et al.*, Local structural elements in the mostly unstructured transcriptional activation domain of human p53. *The Journal of biological chemistry* **275**, 29426-29432 (2000).

29. S. C. Zondlo, A. E. Lee, N. J. Zondlo, Determinants of specificity of MDM2 for the activation domains of p53 and p65: proline27 disrupts the MDM2-binding motif of p53. *Biochemistry* **45**, 11945-11957 (2006).

30. W. Borcherds *et al.*, Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells. *Nature chemical biology* **10**, 1000-1002 (2014).

31. A. Mohan *et al.*, Analysis of molecular recognition features (MoRFs). *Journal of molecular biology* **362**, 1043-1059 (2006).

32. P. Radivojac *et al.*, Intrinsic disorder and functional proteomics. *Biophysical journal* **92**, 1439-1456 (2007).

33. L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradović, A. K. Dunker, Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of molecular biology* **323**, 573-584 (2002).

34. F. Gypas, G. N. Tsaousis, S. J. Hamodrakas, mpMoRFsDB: a database of molecular recognition features in membrane proteins. *mpMoRFsDB: a database of molecular recognition features in membrane proteins*, (2013).

35. J. N. Sachs, D. M. Engelman, Introduction to the membrane protein reviews: the interplay of structure, dynamics, and environment in membrane protein function. *Annu Rev Biochem* **75**, 707-712 (2006).

36. G. von Heijne, Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225**, 487-494 (1992).

37. I. Kotta-Loizou, G. N. Tsaousis, S. J. Hamodrakas, Analysis of Molecular Recognition Features (MoRFs) in membrane proteins. *Biochim Biophys Acta* **1834**, 798-807 (2013).

38. B. Xue, L. Li, S. O. Meroueh, V. N. Uversky, A. K. Dunker, Analysis of structured and intrinsically disordered regions of transmembrane proteins. *Mol Biosyst* **5**, 1688-1702 (2009).

39. G. N. Ramachandran, V. Sasisekharan, Conformation of polypeptides and proteins. *Adv Protein Chem* **23**, 283-438 (1968).

40. G. N. Ramachandran, A. K. Mitra, An explanation for the rare occurrence of cis peptide units in proteins and polypeptides. *Journal of molecular biology* **107**, 85-92 (1976).

41. P. Craveur, A. P. Joseph, P. Poulain, A. G. de Brevern, J. Rebehmed, Cis-trans isomerization of omega dihedrals in proteins. *Amino acids* **45**, 279-289 (2013).

42. M. W. MacArthur, J. M. Thornton, Influence of proline residues on protein conformation. *Journal of molecular biology* **218**, 397-412 (1991).

43. S. S. Zimmerman, H. A. Scheraga, Stability of cis, trans, and nonplanar peptide groups. *Macromolecules* **9**, 408-416 (1976).

44. S. F. Gothel, M. A. Marahiel, Peptidyl-prolyl cis-trans isomerases, a superfamily of ubiquitous folding catalysts. *Cell Mol Life Sci* **55**, 423-436 (1999).

45. X. Z. Zhou, P. J. Lu, G. Wulf, K. P. Lu, Phosphorylation-dependent prolyl isomerization: a novel signaling regulatory mechanism. *Cell Mol Life Sci* **56**, 788-806 (1999).

46. G. Fischer, T. Aumüller, Regulation of peptide bond cis/trans isomerization by enzyme catalysis and its implication in physiological processes. *Reviews of physiology, biochemistry and pharmacology* **148**, 105-150 (2003).

47. K. P. Lu, G. Finn, T. H. Lee, L. K. Nicholson, Prolyl cis-trans isomerization as a molecular timer. *Nature chemical biology* **3**, 619-629 (2007).

48. F. Suizu, A. Ryo, G. Wulf, J. Lim, K. P. Lu, Pin1 regulates centrosome duplication, and its overexpression induces centrosome amplification, chromosome instability, and oncogenesis. *Molecular and cellular biology* **26**, 1463-1479 (2006).

49. C. M. Eakin, A. J. Berman, A. D. Miranker, A native to amyloidogenic transition regulated by a backbone trigger. *Nature structural & molecular biology* **13**, 202-208 (2006).

50. L. Pastorino *et al.*, The prolyl isomerase Pin1 regulates amyloid precursor protein processing and amyloid-beta production. *Nature* **440**, 528-534 (2006).

51. V. Y. Torbeev, D. Hilvert, Both the cis-trans equilibrium and isomerization dynamics of a single proline amide modulate β2-microglobulin amyloid assembly. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 20051-20056 (2013).

52. M. K. Kim, Y. K. Kang, Positional preference of proline in alpha-helices. *Protein science : a publication of the Protein Society* **8**, 1492-1499 (1999).

53. D. S. Kemp, T. P. Curran, (2S,5S,8S,11S)-1-Acetyl-1,4-diaza-3-keto-5-carboxy-10-thia-tricyclo-[2.8.04,8]-tridecane, 1 synthesis of prolyl-proline-derived, peptide-functionalized templates for α-helix formation. *Tetrahedron Letters* **29**, 4931-4934 (1988).

54. R. H. Yun, A. Anderson, J. Hermans, Proline in alpha-helix: stability and conformation studied by dynamics simulation. *Proteins* **10**, 219-228 (1991).

55. C. Lee *et al.*, Contribution of proline to the pre-structuring tendency of transient helical secondary structure elements in intrinsically disordered proteins. *Biochimica et biophysica acta* **1840**, 993-1003 (2014).
56. S. G. Dastidar, D. P. Lane, C. S. Verma, Multiple peptide conformations give rise to similar binding affinities: molecular simulations of p53-MDM2. *Journal of the American Chemical Society* **130**, 13514-13515 (2008).
57. A. H. Elcock, D. Sept, J. A. McCammon, Computer Simulation of Protein−Protein Interactions. *The Journal of Physical Chemistry B* **105**, 1504-1518 (2001).
58. K. S. Sandhu, Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins. *Journal of molecular recognition : JMR* **22**, 1-8 (2009).
59. A. Wlodawer, Rational approach to AIDS drug design through structural biology. *Annual review of medicine* **53**, 595-614 (2002).
60. T. Kortemme, D. E. Kim, D. Baker, Computational alanine scanning of protein-protein interfaces. *Science's STKE : signal transduction knowledge environment* **2004**, pl2 (2004).
61. T. Pawson, P. Nash, Assembly of cell regulatory systems through protein interaction domains. *Science (New York, N.Y.)* **300**, 445-452 (2003).
62. D. Vuzman, Y. Levy, Intrinsically disordered regions as affinity tuners in protein-DNA interactions. *Molecular bioSystems* **8**, 47-57 (2012).
63. C. Chipot, A. Pohorille, C. Chipot, A. Pohorille, *Free Energy Calculations*. (Springer Science & Business Media, 2007), pp. 518.
64. M. K. Gilson, H.-X. Zhou, Calculation of protein-ligand binding affinities. *Annual review of biophysics and biomolecular structure* **36**, 21-42 (2007).
65. B. K. Shoichet, Virtual screening of chemical libraries. *Nature* **432**, 862-865 (2004).
66. C. Zhang, S. Liu, Q. Zhu, Y. Zhou, A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *Journal of medicinal chemistry* **48**, 2325-2335 (2005).
67. Z. Weng, C. Delisi, S. Vajda, Empirical free energy calculation: comparison to calorimetric data. *Protein science : a publication of the Protein Society* **6**, 1976-1984 (1997).
68. J. Aqvist, C. Medina, J. E. Samuelsson, A new method for predicting binding affinity in computer-aided drug design. *Protein engineering* **7**, 385-391 (1994).
69. I. Massova, P. A. Kollman, Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspectives in drug discovery and design* **18**, 113-135 (2000).
70. R. W. Zwanzig, High ‐ Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **22**, 1420-1426 (1954).
71. J. G. Kirkwood, Statistical Mechanics of Fluid Mixtures. *The Journal of Chemical Physics* **3**, 300-313 (1935).
72. D. L. Mobley, J. D. Chodera, K. A. Dill, The Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *Journal of chemical theory and computation* **3**, 1231-1235 (2007).
73. H.-J. Woo, B. Roux, Calculation of absolute protein-ligand binding free energy from computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 6825-6830 (2005).

74. Y. Deng, B. Roux, Computations of standard binding free energies with molecular dynamics simulations. *The journal of physical chemistry. B* **113**, 2234-2246 (2009).

75. J. Wang, Y. Deng, B. Roux, Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophysical journal* **91**, 2798-2814 (2006).

76. J. C. Gumbart, B. Roux, C. Chipot, Standard binding free energies from computer simulations: What is the best strategy? *Journal of chemical theory and computation* **9**, 794-802 (2013).

77. V. M. Dadarlat, R. D. Skeel, Dual role of protein phosphorylation in DNA activator/coactivator binding. *Biophysical journal* **100**, 469-477 (2011).

78. K. H. Vousden, C. Prives, Blinded by the Light: The Growing Complexity of p53. *Cell* **137**, 413-431 (2009).

79. S. Shangary, S. Wang, Small-molecule inhibitors of the MDM2-p53 protein-protein interaction to reactivate p53 function: a novel approach for cancer therapy. *Annual review of pharmacology and toxicology* **49**, 223-241 (2009).

80. Y. Barak, E. Gottlieb, T. Juven-Gershon, M. Oren, Regulation of mdm2 expression by p alternative promoters produce transcripts with non-identical translation protential. *Genes & Development* **8**, 1739-1749 (1994).

81. Y. Haupt, R. Maya, A. Kazaz, M. Oren, Mdm2 promotes the rapid degradation of p53. *Nature* **387**, 296-299 (1997).

82. R. Honda, H. Tanaka, H. Yasuda, Oncoprotein MDM2 is a ubiquitin ligase E3 for tumor suppressor p53. *FEBS letters* **420**, 25-27 (1997).

83. M. H. Kubbutat, S. N. Jones, K. H. Vousden, Regulation of p53 stability by Mdm2. *Nature* **387**, 299-303 (1997).

84. Y. Xu, Regulation of p53 responses by post-translational modifications. *Cell Death & Differentiation* **10**, 400-403 (2003).

85. J. Chen, The Roles of MDM2 and MDMX Phosphorylation in Stress Signaling to p53. *Genes & cancer* **3**, 274-282 (2012).

86. P. Kollman, Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews* **93**, 2395-2417 (1993).

87. F. M. Ytreberg, Absolute FKBP binding affinities obtained via nonequilibrium unbinding simulations. *The Journal of chemical physics* **130**, 164906 (2009).

88. C. H. Bennett, Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **22**, 245-268 (1976).

89. D. V. Schroeder, *An introduction to Thermal Physics*. (Addison Wesley, ed. 2, 2000).

90. W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics. *Journal of molecular graphics* **14**, 33-38, 27 (1996).

91. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926-935 (1983).

92. A. D. MacKerell *et al.*, All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry. B* **102**, 3586-3616 (1998).

93. J. C. Phillips *et al.*, Scalable molecular dynamics with NAMD. *Journal of computational chemistry* **26**, 1781-1802 (2005).

94. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N · log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98**, 10089-10092 (1993).

95. W. F. van Gunsteren, H. J. C. Berendsen, Algorithms for macromolecular dynamics and constraint dynamics. *Molecular Physics* **34**, 1311-1327 (1977).

96. S. E. Feller, Y. Zhang, R. W. Pastor, B. R. Brooks, Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics* **103**, 4613-4621 (1995).

97. G. M. Torrie, J. P. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. **23**, 187-199 (1977).

98. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, J. M. Rosenberg, The weighted histogram analysis method for free-energy calculations on biomolecules I The method. *Journal of Computational Chemistry* **13**, 1011-1021 (1992).

99. A. Grossfield.

100. J. Phan *et al.*, Structure-based design of high affinity peptides inhibiting the interaction of p53 with MDM2 and MDMX. *The Journal of biological chemistry* **285**, 2174-2183 (2010).

101. L. Gracia.

102. C. Grathwohl, K. Wüthrich, NMR studies of the rates of proline cis-trans isomerization in oligopeptides. *Biopolymers* **20**, 2623-2633 (1981).

103. H. Deng, N. Zhadin, R. Callender, Dynamics of protein ligand binding on multiple time scales: NADH binding to lactate dehydrogenase. *Biochemistry* **40**, 3767-3773 (2001).

104. S. McClendon, D. M. Vu, K. Clinch, R. Callender, R. B. Dyer, Structural transformations in the dynamics of Michaelis complex formation in lactate dehydrogenase. *Biophysical journal* **89**, L07-09 (2005).

105. Y. A. Zhan, H. Wu, A. T. Powell, G. W. Daughdrill, F. M. Ytreberg, Impact of the K24N mutation on the transactivation domain of p53 and its binding to murine double-minute clone 2. *Proteins* **81**, 1738-1747 (2013).

106. G. W. Yu *et al.*, The central region of HDM2 provides a second binding site for p53. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 1227-1232 (2006).

107. P. Romero *et al.*, Sequence Complexity of Disordered Protein. *Proteins: Structure, Function, and Genetics* **42**, (2001).

108. R. M. Williams *et al.*, The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 89-100 (2001).

109. J. A. Marsh, J. D. Forman-Kay, Sequence determinants of compaction in intrinsically disordered proteins. *Biophysical journal* **98**, 2383-2390 (2010).

110. A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, Z. Obradovic, Intrinsic disorder and protein function. *Biochemistry* **41**, 6573-6582 (2002).

111. G. P. Singh, M. Ganapathi, D. Dash, Role of intrinsic disorder in transient interactions of hub proteins. *Proteins* **66**, 761-765 (2007).

112. A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, V. N. Uversky, Flexible nets. The roles of intrinsic disorder in protein interaction networks. *The FEBS journal* **272**, 5129-5148 (2005).

113. C. J. Brandl, C. M. Deber, Hypothesis about the function of membrane-buried proline residues in transport proteins. *Proc Natl Acad Sci U S A* **83**, 917-921 (1986).

114. H. M. Berman *et al.*, The Protein Data Bank. *Nucleic acids research* **28**, 235-242 (2000).

115. B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon, L. S. D. Caves, Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics (Oxford, England)* **22**, 2695-2696 (2006).

116. J. Song, K. Burrage, Z. Yuan, T. Huber, Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. *BMC bioinformatics* **7**, 124 (2006).

117. M. Heinig, D. Frishman, STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic acids research* **32**, 2 (2004).

118. D. Frishman, P. Argos, Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566-579 (1995).

119. C. R. L. Sarah *et al.*, Cis–trans isomerization at a proline opens the pore of a neurotransmitter-gated ion channel. *Nature*, (2005).

120. D. Pal, P. Chakrabarti, Cis peptide bonds in proteins: residues involved, their conformations, interactions and locations. *Journal of molecular biology* **294**, 271-288 (1999).

121. A. Jabs, M. S. Weiss, R. Hilgenfeld, Non-proline cis peptide bonds in proteins. *J Mol Biol* **286**, 291-304 (1999).

122. M. W. MacArthur, J. M. Thornton, Deviations from planarity of the peptide bond in peptides and proteins. *Journal of molecular biology* **264**, 1180-1195 (1996).

123. D. S. Berkholz, C. M. Driggers, M. V. Shapovalov, R. L. Dunbrack, P. A. Karplus, Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 449-453 (2012).

124. S. C. Lovell *et al.*, Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* **50**, 437-450 (2003).

125. S. Hovmoller, T. Zhou, T. Ohlson, Conformations of amino acids in proteins. *Acta Crystallogr D Biol Crystallogr* **58**, 768-776 (2002).

126. I. N. Berezovsky, G. T. Kilosanidze, V. G. Tumanyan, L. L. Kisselev, Amino acid composition of protein termini are biased in different manners. *Protein Eng* **12**, 23-30 (1999).

127. H. J. Dyson, P. E. Wright, Coupling of folding and binding for unstructured proteins. *Current opinion in structural biology* **12**, 54-60 (2002).