A MOLECULAR MECHANICS KNOWLEDGE BASE

APPLIED TO TEMPLATE BASED STRUCTURE PREDICTION

A Dissertation

by

XIAOTAO QU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2009

Major Subject: Biochemistry

A MOLECULAR MECHANICS KNOWLEDGE BASE

APPLIED TO TEMPLATE BASED STRUCTURE PREDICTION

A Dissertation

by

XIAOTAO QU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Jerry Tsai
	J. Martin Scholtz
Committee Members,	James Sacchettini
	Michael Polymenis
Head of Department,	Gregory D. Reinhart

December 2009

Major Subject: Biochemistry

ABSTRACT

A Molecular Mechanics Knowledge Base
Applied to Template Based Structure Prediction.
(December 2009)

Xiaotao Qu, B.S., Fudan University, Shanghai, China

Co-Chair of Advisory Committee: Dr. Jerry Tsai
Dr. J. Martin Scholtz

Predicting protein structure using its primary sequence has always been a challenging topic in biochemistry. Although it seems as simple as finding the minimal energy conformation, it has been quite difficult to provide an accurate yet reliable solution for the problem. On the one hand, the lack of understanding of the hydrophobic effect as well as the relationship between different stabilizing forces, such as hydrophobic interaction, hydrogen bonding and electronic static interaction prevent the scientist from developing potential functions to estimate free energy. On the other hand, structure databases are limited with redundant structures, which represent a non-continuous, sparsely-sampled conformational space, and preventing the development of a method suitable for high-resolution, high-accuracy structure prediction that can be applied for functional annotation of an unknown protein sequence. Thus, in this study, we use molecular dynamics simulation as a tool to sample conformational space. Structures were generated with physically realistic conformations that represented the properties of ensembles of native structures. First, we focused our study on the relationship among different factors that stabilize protein structure. Using a well-characterized mutation system of the β-hairpin, a fundamental building block of protein, we were able to identify the effect of terminal ion-pairs (salt-bridges) on the stability of the β-hairpin, and its relationship with hydrophobic interactions and hydrogen bonds. In the same study, we also correlated our theoretical simulations qualitatively with experimental results. Such analysis provides us a better understanding of beta-hairpin stability and helps us to improve the protein engineering method to design more stable hairpins. Second, with large-scale simulations of different representative protein folds, we were able to conduct a fine-grained analysis by sampling the continuous conformational space to characterize the relationship among backbone conformation, side-chain conformation and side-chain packing. Such information is valuable for improving high-resolution structure prediction. Last, with this information, we developed a new prediction algorithm using packing information derived from the conserved relative packing groups. Based on its performance in CASP7, we were able to draw the conclusion that our simulated dataset as well as our packing–oriented prediction method are useful for template based structure prediction.

DEDICATION

To my wife Miao

whom I love with all my heart

and my parents

who support me all the way here

ACKNOWLEDGEMENTS

I would like to thank my committee co-chairs, Dr. Tsai and Dr. Scholtz, my committee members, Dr. Sacchettini and Dr Polymenis, and my previous committee members, Dr. Liwang and Dr Guarino for their guidance and support throughout the course of my study.

I would also like to thank Dr. Swanson for her help and all the members of Jerry's Lab, who have provided help, friendship, and simulating conversation over the many years.

Thanks also go to the department faculty and staff for making my time at Texas A&M University a great experience.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

It has been a principal goal for biochemists to predict protein structure from its sequence since Anfinsen's work showing that a protein is able to refold on its own.[1] With the completion of many genome projects, the application of more accurate protein structure prediction methods would have a broad impact on function determination, genome annotation and even comprehending cellular processes by helping to defined protein pathways. The most direct way to predict the protein structure is energy minimization, based on the fact that the native conformation always has the lowest free energy in the folding funnel. By following the folding pathway towards energy minimal on a given sequence, we should be able to identify the native conformation in the end. But the problem, better known as "Levinthal Paradox",[2] which indicates that with too many degrees of freedom, it is impossible to sample all available conformations for a given sequence of 100 residues. More practically, the energy function used to evaluate the structure is not robust enough to discriminate native conformation from decoys (near native conformation). Furthermore, the forces that keep protein structure folded, hydrophobic effect, hydrogen bond, salt-bridges and etc, are not well understood, as well as their contribution and relationship during the folding process.

Thus, the most accurate way to predict a protein structure now is to use a known homologue structure as a template based on the concept that similar sequences usually

This dissertation follows the style *of Journal of Molecular Biology*.

have similar structures thus similar functions. Methods have been developed for years to predict the protein structure using such approach. And it has been improved dramatically since the exponential explosion of protein structure database, PDB.[3] Even though such increase in the number of available structures provides more possible templates for sequences with unknown structures, the structure database sill sparsely samples the native conformational space and partially because of such non-continuous sampling, the current prediction methods fail to provide high resolution and high accuracy structure models that is required for biological meaningful interpretation of the predicted structures. There is no guarantee that whether or not there will be any valuable information can be derived from the predicted structure as well as how much information derived from the predicted structure is accurate.

To have a better understanding of protein folding and the relationship between the primary sequence and the tertiary structure, we use molecular dynamics simulation as a tool and make use of the molecular mechanics potential within it to sample the near native conformational space for ensemble of structures with physical realistic conformation. We first study one of the simplest structure elements, β-hairpin, to try to understand the relationship among different forces that can stabilize protein: hydrophobic interaction, hydrogen bond and electrostatic interaction between ion-pairs. We found that the terminal ion pairs could stabilize hairpin structure with reduced hydrophobic core interaction at the strand region. Second, we conducted a large-scale analysis on ensemble structures of different protein folds that represent the native conformation. We focus on understanding the relationship among side-chain packing, sidechain conformation, and backbone conformation, in another word, on understanding what are the characteristics of the native conformation in terms of packing (residue's volume), backbone conformation (torsion angles) and rotamer conformation (rotamer angles). We find that the backbone conformation and the packing of the residue intervene with each other and can be characterized using backbone torsion angles (ϕ and ψ). We also provide a detailed, steric based explanation of such relationship that summarizes previous studies with more details. At last, we developed a method to predict structure using the conserved packing information derived from the packing of template structures and a scoring function by analyzing the backbone dependent side-chain packing and conformation in the dynameome dataset. Using CASP7 targets as a double-blinded test set, we were able to show that this method is able to improve the template structure and provide a different approach of refining the template structure. Furthermore, the large-scale molecular simulation, on which our scoring function is based, is also proved to provide a fine-grained, continuous ensemble for high-resolution study of native conformation and properties of native structure.

Molecular Dynamics

Molecular dynamics simulation uses the molecular mechanics potential function based on Newton's second law to simulate movement of atoms over time. Based on the Ergodic Hypothesis,[4] time averaged properties (simulation of one single molecule over long time) are equal to ensemble averaged properties (experimental observation of the equilibrium system). Molecular dynamics simulation holds a great potential to represent real folding environment especially when the protein flexibility is concerned.[5-9] Protein flexibility is important for protein function, such as molecular recognition, enzyme activity and protein turnover. Because it is very difficult to determine the protein flexibility experimentally, the atomistic simulation becomes the only viable alternation for such kind of study.[9] Besides studying protein flexibility, molecular dynamics simulations can also be used to study the interactions between protein and water molecules as well as ions, and to generate model structures in the protein structure prediction. Also, molecular dynamics simulations are commonly used in the NMR structure determination to determine the protein structure based on experimental data because it is the only way to sample the conformational space with physical realistic conformations.

More recently, beyond studying individual molecular system, a lot of effort has been applied to create large scaled, systematic molecular dynamics simulations on different protein folds to study the overall dynamics of native conformations. A recent study using different package of molecular dynamics simulations on 30 protein meta-folds[10] shows that even though the force field applied in different package is different, the overall dynamics and stability of all simulations are in consensus,[9] which indicates the ability of molecular dynamics simulation to sample the near native conformation space consensually. Database collecting different molecular dynamics simulation are generated and can be used as a knowledge base for understanding native structures.

Template Based Structure Prediction

Template based modeling/prediction (TBM) of protein structure, was commonly known as homology modeling and more recently, comparative modeling.[11,12] It creates a prediction of an unknown structure using a close structural homolog. The new designation, TBM, which is started to be used since CASP7,[13] is more general and allows for the distinct contrast to template free structure prediction, more commonly known as *ab initio* or *de novo* modeling.[14-16] Because it is believed that a representative of every protein fold will eventually be solved,[17-23] template based structure prediction holds a great deal of promise. The availability of a representative fold as a starting template for a sequence of unknown structure offers the quickest path to generating a model of the real structure. Furthermore, template based methods produce the most reliable and accurate predictions of the protein structure aside from experimental determination[24,25] and have been used successfully in a variety of applications, such as studying the effect of mutations, designing site-directed mutagenesis, predicting binding sites and docking small molecules in structure-based drug discovery. Unfortunately, the imprecise variations between the close template and the real structure produce the major source of challenges facing in this field today.

In general, any TBM method can be classified into four steps[26-28] as outlined in Figure 1,[29] although the delineation between steps is somewhat arbitrary since many methods combine steps. First, the parent structure(s) are identified using sequence searches against the known structure database (the Protein Data Bank[3]). Second, the initial template structure(s) are constructed by aligning the target sequence to the parent structure and by identifying conserved and variable regions. Third, the template structure(s) are refined through a combination of backbone moves, side-chain packing, and loop modeling of highly variable regions. This step is an attempt to sample the

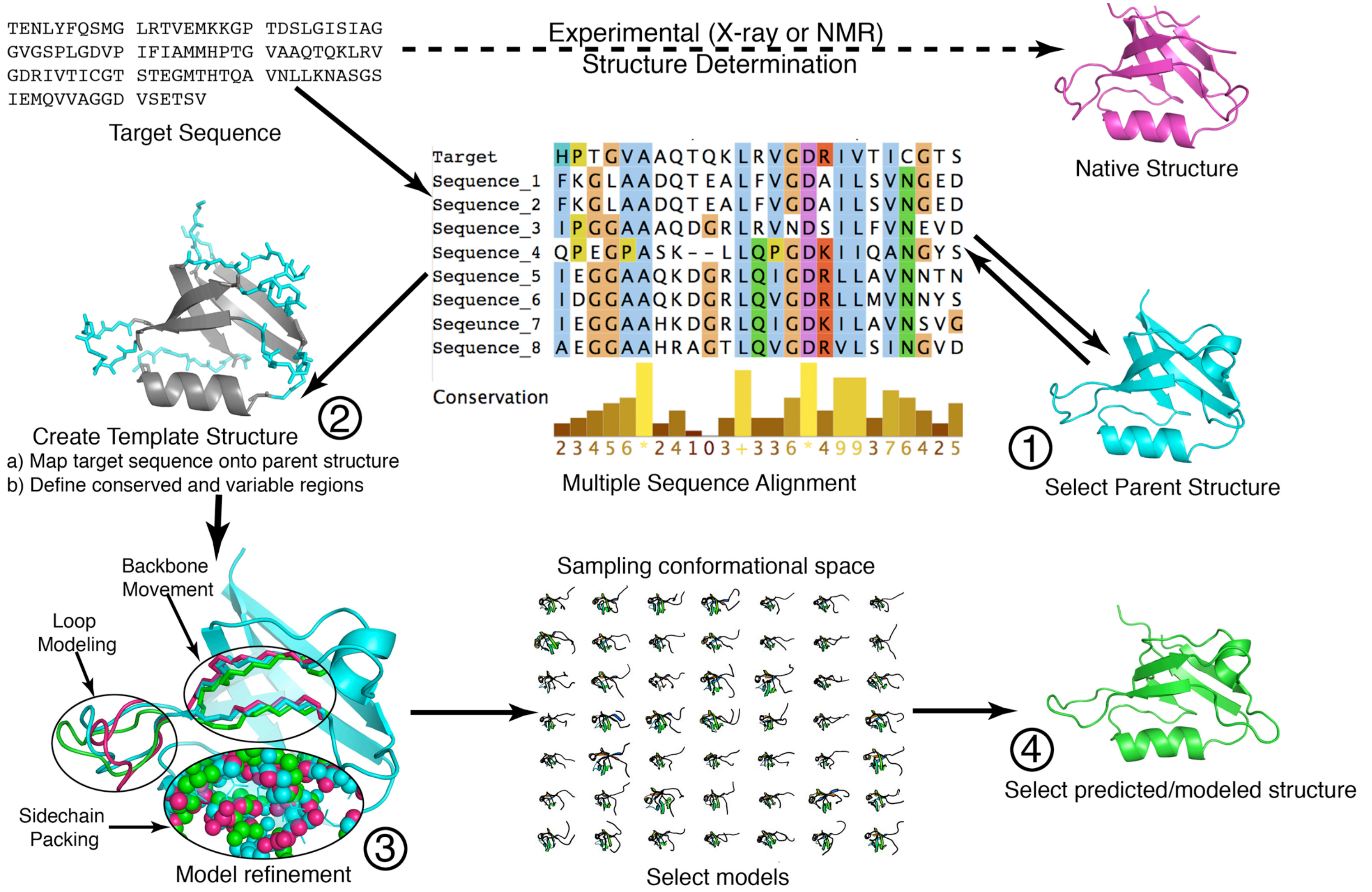

TENLYFQSMG LRTVEMKKGP TDSLGISIAG
GVGSPLGDVP IFIAMMHPTG VAAQTQKLRV
GDRIVTICGT STEGMTHTQA VNLLKNASGS
IEMQVVAGGD VSETSV
Target Sequence
Experimental (X-ray or NMR)
Structure Determination
Native Structure
Target HPTGVAAQTQKLRVGDRIVTICGTS
Sequence_1 FKGLAADQTEALFVGDAILSVNGED
Sequence_2 FKGLAADQTEALFVGDAILSVNGED
Sequence_3 IPGGAAAQDGRLRVNDSILFVNEVD
Sequence_4 QPEGPASK--LLQPGDKIIQANGYS
Sequence_5 IEGGAAQKDGRLQIGDRLLAVNNTN
Sequence_6 IDGGAAQKDGRLQVGDRLLMVNNYS
Seqeunce_7 IEGGAAHKDGRLQIGDKILAVNSVG
Sequence_8 AEGGAAHRAGTLQVGDRVLSINGVD
Conservation
23456*24103+336*499376425
Multiple Sequence Alignment
1
Select Parent Structure
2
Create Template Structure
a) Map target sequence onto parent structure
b) Define conserved and variable regions
Backbone
Movement
Loop
Modeling
Sidechain
Packing
3
Model refinement
Sampling conformational space
Select models
4
Select predicted/modeled structure

Figure 1: Outline of steps for template based structure prediction.
Outline of steps involved in the template based structure prediction are shown. The goal is to use the target sequence to predict the native structure that can be solved experimentally.[29]

conformational space where the native conformation exists and usually a number (on the order of hundreds to thousands) of potential models are created. Thus, the last step is to evaluate these models and choose the best one whose conformation is closest to the native conformation. Usually, the first and second steps occur concurrently, and the same can be said for the third and fourth steps. Newer approaches include *de novo* prediction of variable regions during the refinement as well as procedures that iterate within different steps to optimize the final structure.[29] For the last 14 years, template based structure prediction has evolved steadily during 7 CASP experiments[13,30-36] and now it was able to provide "added value" to the best template structure generated from best homologues.[37,38]

Packing of Protein

Packing of the protein is usually addressed using two different models based on experimental analysis of protein structure. The "jig-saw puzzle" model states that the principle behind the protein folding was the stereo-specific packing of the protein side-chain and was supported by the evidence that side-chain to side-chain contacts in the protein core was highly complimentary.[39] Based on it, it is not surprise to find that the packing densities in protein cores are close to those of small organic molecules crystals.[40] On the other hand, "nuts and bolts" model based its theory on the experimental finding that protein structure has amazing capability to withstand changes by repacking the core as mutations necessitate. For example, it has been found that 7 methionine mutation of T4 lysozyme can still retain its overall fold[41] and two proteins with only low sequence identity (24%) have very similar overall folds but different core flexibility.[42] A more details review and comparison of two models can be found elsewhere.[43,44]

Knowledge Based Scoring Function

Comparing to physical based potential function as used in molecular dynamics simulations. Knowledge based scoring function or sometimes called statistical based potential function in terms of their different applications has become a standard in the structure prediction, especially as the PDB database grows exponentially to provide more structural diversity. Knowledge based potentials are derived from statistics of native protein structures, where statistics refers to the frequency distribution of a calculated property in a set of protein structures.[45,46] These frequency distributions can be transformed into energy terms as a potential of mean force.[47] The focus of these statistical functions is to categorize the general features of the native structure. Most of these features are simple chemical properties and in effect discriminate based on what is more native protein-like. Examples include bond lengths, atom contact distances (directly or indirectly based on the analysis of contacts, either inter residue contacts, inter atom contacts, or contacts with solvent[48]). WHAT_CHECK[49] and Verify3D[50] are example of standalone programs that provide such scores and are widely used as a quick approach to discriminate non-native conformation from native structure. The greatest strengths of statistical based scores are that they are computationally more efficient, and they provide reasonable estimates of the characteristics that make up a folded protein.

CASP

CASP (Critical Assessment of Structure prediction) is a large-scale community experiment held biannually.[38,51] From CASP1 (1994) to CASP7 (2006),[13,52-57] prediction teams from all over the world applied their prediction algorithms in a double blind test against unknown, but solved protein structures. After each CASP, a meeting is held, where the performance of all the groups is examined based on their model quality compared to the experimentally determined structures. Also, an important aspect of the CASP meeting is the sharing/dissemination of successful methods, which has speeded the development in the protein prediction field. After the meeting, a special issue of Proteins is published, where progress, drawbacks, and future goals are discussed. Given its blind prediction feature, CASP has become a standard for testing the merits of any structure prediction algorithm.

Besides CASP, CAFASP[58-60] (Critical Assessment of Fully Automated Structure Prediction experiment) was used to evaluate automatic structure prediction methods and has run in parallel with CASP since 1998 on the same target set. More continuous assessment has been done by LiveBench[61] and EVA,[62] which is also done on a relatively large number of prediction targets compiled every week from newly released PDB structure. The evaluation from all of these endeavors has been vital to pushing the structure prediction field forward. In addition, they have exposed the exact areas where resources and effort need to be placed for advancement of the structure prediction. As a byproduct of sharing knowledge and information, much productive collaboration has grown from the interactions promoted by this meeting and its assessment of TBM.

Beta Hairpin

Beta hairpin is one of the fundamental structure elements, which are considered as building block of the protein structure as alpha helix. But unlike alpha helix,[63-65] there is relative less study on the conformation and stability of the beta hairpin due to it is intrinsic unstableness in solution. Fortunately, with the discovery of self-folded hairpin in solution, for example, the second hairpin of protein G, which is one of the first hairpins that can form solution structures,[66-68] more studies have been done on this structure element. Using it as a model system, specific non-covalent interactions that are crucial to the secondary structure formation as well as those are important at early events of protein folding can be studied.

The second hairpin of protein G is an anti-parallel, 16-residue hairpin. It contains a 4:4 type IV turn[69,70] making up by residues 47-50 (sequence DATK), seven possible main-chain-to-main-chain hydrogen bonds, a hydrophobic core involves residue Trp43, Tyr45, Phe52 and Val 54 and a possible ion pair at the termini.[71] These four component represent 4 major factors that contribute to the folding and stability of the G-hairpin, the intrinsic β-turn propensity,[72-74] the hydrophobic interaction of side-chain across two strands,[75-80] the hydrogen bonds that define and maintain hairpin architecture,[81-83] and the favored electrostatic interaction.[71,84-86] Besides G-hairpin, a lot of other hairpin systems have been studied as well as a lot of well-folded "designed beta-hairpins" using existing information have also be developed. Our understanding of beta-hairpin formation and stability has been progressing in recent years.[87]

CHAPTER II

MODELING THE PERTURBATIONS OF A TERMINAL ION PAIR ON β-HAIRPIN GEOMETRY AND STABILITY

Overview

β-hairpin is one of the fundamental secondary structure elements, which are considered as building blocks of protein structure as alpha helix. To understand the relationship between β-hairpin stability and its conformational geometry, a series of peptides with mutations based on the second β-hairpin from the B1 domain of protein G were studied. In total, eight peptides differing at their N-termini are studied with added or deleted potential terminal ion pair interactions. The same set of peptides was used to experimentally show that ion pairing between the termini increases β-hairpin stability. While each peptide exhibits different ion pair orientations, we find that the ion pair interactions between the termini are significant sampling over 60% of our ensemble structures. The effects of such terminal interactions were correlated to other β-hairpin regions (turn, hydrophobic core, and main-chain hydrogen bonds) in terms of persistence and geometry. All the structures in the ensemble maintain a stable turn and significant hydrophobic contacts across the strands, which is consistent with previous experimental and theoretical studies that the turn and hydrophobic core are important for hairpin stability. While the changes in terminal ion-pair interactions do not significantly affect the overall hairpin geometry, the results show that terminal interactions increase the stability of hairpin by fixing the termini with a pseudo hydrogen bond ring as well as by cooperating with hydrophobic interactions between the strands. Also, the results show that the hydrophobic core can accommodate small perturbations and still keep its conformational stability.

Background

The hairpin we are interested in this study is the second β-hairpin from the B1 domain of protein G and we denoted it as G-hairpin throughout this discussion. The G-hairpin consists of 3 regions defining 4 areas of important interactions: the turn, the strand's anti-parallel main-chain hydrogen bonding, the strand's hydrophobic core and the termini. The G-hairpin's turn sequence of Asp47-Ala48-Thr49-Lys50 produces a type IV turn that shows a high level of stability in both experiments[88,89] and simulation.[90-93] It has also been shown that a well formed turn can increase the stability of β-hairpin. For example, changing the turn sequence to D-Pro-Gly to form a Type II' turn has been shown to increase the stability of the G-hairpin.[94] Between the β-hairpin's anti-parallel strands, two types of interactions occur: main-chain hydrogen bonds and side-chain hydrophobic contacts. Which one is more important to β-hairpin stability is under debate since different β-hairpins inconsistently exhibits a preference for either one. In the G-hairpin, the zipper model, which is supported by experiments[81] and simulations,[83] emphasizes the importance of main-chain hydrogen bond formation. In contrast, a body of theoretical work suggests that the side-chain hydrophobic contacts occur early in folding[95] and are important in maintaining the β-hairpin conformation.[75-80] By comparison, relatively few studies on the role of interactions between the β-hairpin's termini have been performed because of the belief that fraying of the termini doesn't permit significant contribution to the β -hairpin stability. In contrast, a MD study suggested that a disulfide bond connecting the termini of a 19-residue β-hairpin from tendamistat was necessary for the stability.[96] For the G-hairpin, in a previous MD study shows that ion-pair interactions can form across the termini of the G-hairpin[92] between the free N-terminal amino group of Gly41 and one of the two carboxyl groups of C-terminal Glu56 (Figure 2) and these interactions help prevent G-hairpin from unfolding. This initial theoretical work has been followed by a number of NMR studies showing that ion-pair and aromatic-aromatic contacts placed across the termini exhibit notable contributions to β-hairpin stability.[97-99] These studies provide a clear picture of the nature and contribution of the various interactions to β-hairpin stability.

Overall, the stability of hairpin is in equilibrium among different types of interactions. In this study, by performing MD simulations on the same set of peptides that have experimentally shown ion pair stabilization of the G-hairpin structure. We are interested in characterizing the peptide geometry necessary to properly present these interactions and how/whether these conformations allow the interactions to cooperate in the stabilization of β-hairpin structure. Table 1 shows the eight G-hairpin based peptides that have been characterized experimentally as well as the number of simulations and the

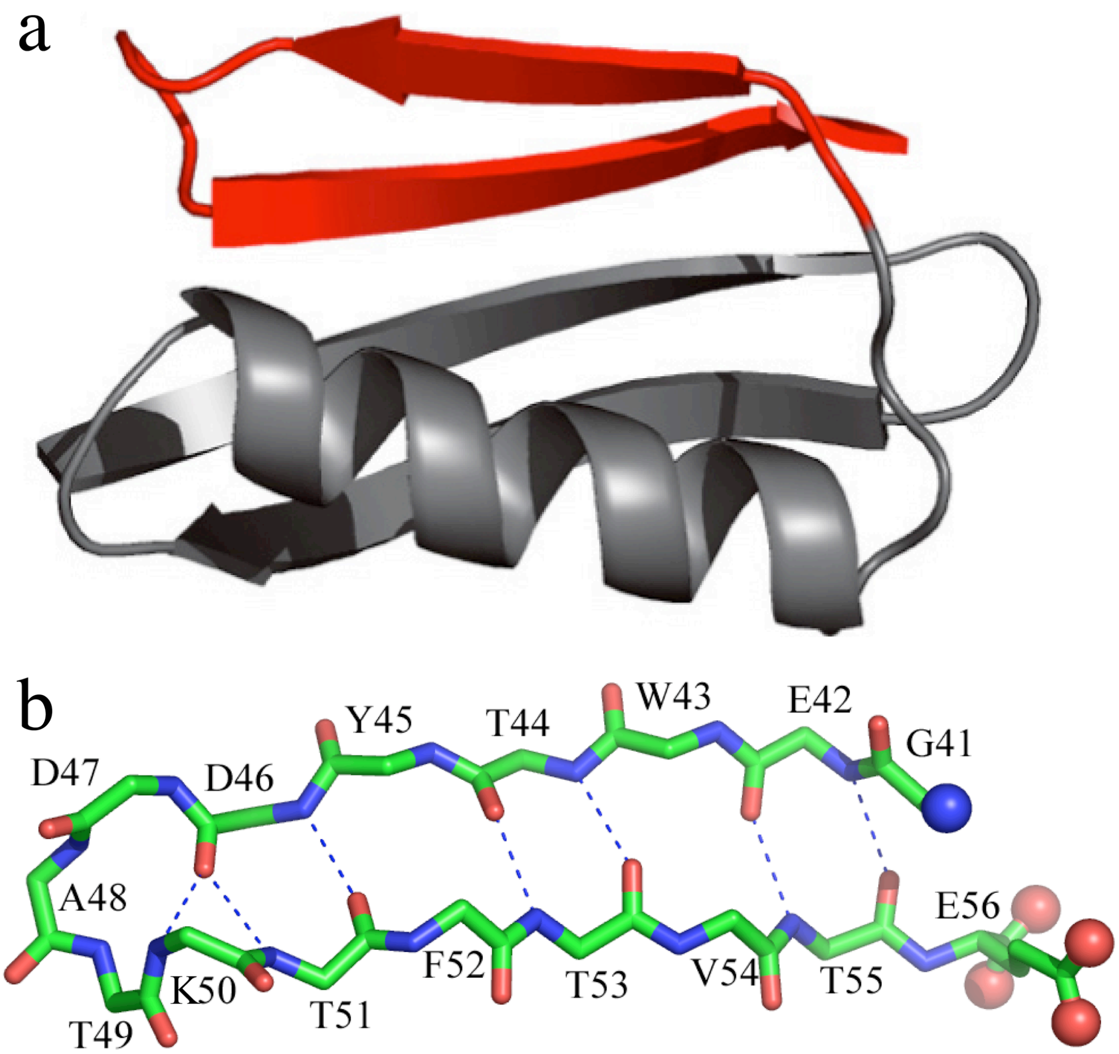


Figure 2: The second hairpin of protein G (G-hairpin).
Diagrams of the β-hairpin were shown. a) A ribbon diagram of the structure of the B1 domain of protein G (PGB1[100]) produced from the crystal structure coordinates (PDB: 1PGB) using MOLSCRIPT.[101] The G-hairpin, residues 41 to 56, is highlighted in red. b) Main-chain backbone atoms of the G-hairpin taken from the crystal structure. The dashed lines represent the hydrogen bonds formed between backbone atoms. The N-terminal amino nitrogen (blue ball) and the carboxyl oxygen atoms of Glu56 (red ball) are also highlighted.

Table 1: β-Hairpin peptide variants

Name	Sequence[a]	Trajectories[b]		Box Size[c] (Å)	Number of[d] Water Atoms
	A-D-D-Y-T-W-E-G		X	28.4	
G41	|	20	Y	43.8	1058
	T- K-T- F-T-V-T-E		Z	27.4	
	A-D-D-Y-T-W-E-G**-Ac**		X	28.3	
Ac-G41	|	10	Y	43.6	1043
	T- K-T- F-T-V-T-E		Z	27.3	
	A-D-D-Y-T-W-E-**K**		X	28.3	
K41	|	20	Y	43.6	1040
	T- K-T- F-T-V-T-E		Z	27.2	
	A-D-D-Y-T-W-E-**K-Ac**		X	28.2	
Ac-K41	|	10	Y	43.5	1028
	T- K-T- F-T-V-T-E		Z	27.1	
	A-D-D-Y-T-W-E-G**-G**		X	27.8	
G40	|	18	Y	27.8	1043
	T- K-T- F-T-V-T-E		Z	43.6	
	A-D-D-Y-T-W-E-G**-G-Ac**		X	26.9	
Ac-G40	|	10	Y	43.7	1082
	T- K-T- F-T-V-T-E		Z	29.7	
	A-D-D-Y-T-W-**E**		X	28.3	
E42	|	10	Y	43.5	1040
	T- K-T- F-T-V-T-E		Z	27.2	
	A-D-D-Y-T-W-**E-Ac**		X	28.4	
Ac-E42	|	10	Y	43.8	1057
	T- K-T- F-T-V-T-E		Z	27.4	

[a.] Modifications are in bold. Ac stands for acetylation.
[b.] Number of simulations performed for each peptide.
[c.] Size of the simulation box.
[d.] Total number of water molecules surrounding protein in the simulation box.

average box size of each simulation. This set of peptides perturbs the interaction between the G-hairpin's termini by differing only in their N-termini. Such modifications change the potential for the terminal salt-bridge (hydrogen bonded ion pair interactions) involving N termini amide group and C termini carboxyl groups (main chain and side chain), while keeping the interactions from other parts of the β-hairpin untouched. The following shorthand notations were used to refer to each peptide as well as a brief description of each peptide. G41 is the native G-hairpin consisting of residues 41 to 56 from Protein G (see Figure 1), which can make one of two possible ion pairs between the N-terminus amino group and the C-terminus carboxyl group or side-chain carboxyl group of the Glu56. K41 is the mutation G41K with increased potential of forming terminal ion pairs by adding the positive amino group of the lysine side chain to the N-terminus, where now extra ion pairs can form. G40 is one residue longer towards the N-terminus, which moves the potential salt-bridging interactions out of symmetry. E42 does the same by eliminating one residue from the N-terminus. For all of these peptides, we also study the corresponding N-terminal acetylated versions (Ac-G41, Ac-K41, Ac-G40, and Ac-E42) that prevent any ion pairs from forming at the N-terminus. The peptides Ac-G41, Ac-G40, and Ac-E42 cannot form any ion-pair interactions between the termini, while Ac-K41 can possibly form one between the Lys41 side-chain's amino group and the carboxyl groups of the C-terminus or the side-chain of Glu56.

Results and Discussion

Theoretically, the time-averaged properties of structures in ensembles can approximate the ensemble equilibrium state and by extension, the properties observed in experiments. In this study, the ensembles generated by MD simulations are not only used to characterize the interactions important to stabilize the β-hairpin conformation but also to provide a detailed molecular view of specific interactions. Referring to the experimental results, the ensembles generated by MD simulation qualitatively represent the experimental observed properties. As a first step, the idealized conformation of the β-hairpin was used to point out the conformational adjustments caused by the β-hairpin's stabilizing interactions. Then, three regions: turn, strands and termini were analyzed for their contribution to β-hairpin stability in various measures of properties, such as conformation, hydrogen bonding, ion pair and hydrophobic contact surface area (HCSA). Previously, it have been showed that the simulated G41 ensemble produces many features similar to the experimental results,[92] especially the stability measurement of 50% folded for the G41 peptide.[102] Therefore, we use the G41 ensemble as a reference point for the β-hairpin. In general, we find that the non-acetylated hairpins are more stable than their acetylated counterparts are. The terminal interaction doesn't perturb the β-hairpin's turn conformation and overall geometry, which is consistent with the belief that the turn is the folding core of the β-hairpin. However, our results indicate that the terminal interaction shows strong salt-bridge features as hydrogen bonded ion pair and stabilizes the β-hairpin conformation by acting like a β-sheet hydrogen-bond ring in a concerted manner that closes off the hydrophobic core.

The ideal β-hairpin structure

Figure 3 shows the ideal β-hairpin structure as the perfect planer structure with optimized inter-strand hydrogen bonds (Figure 3a). To highlight the geometric

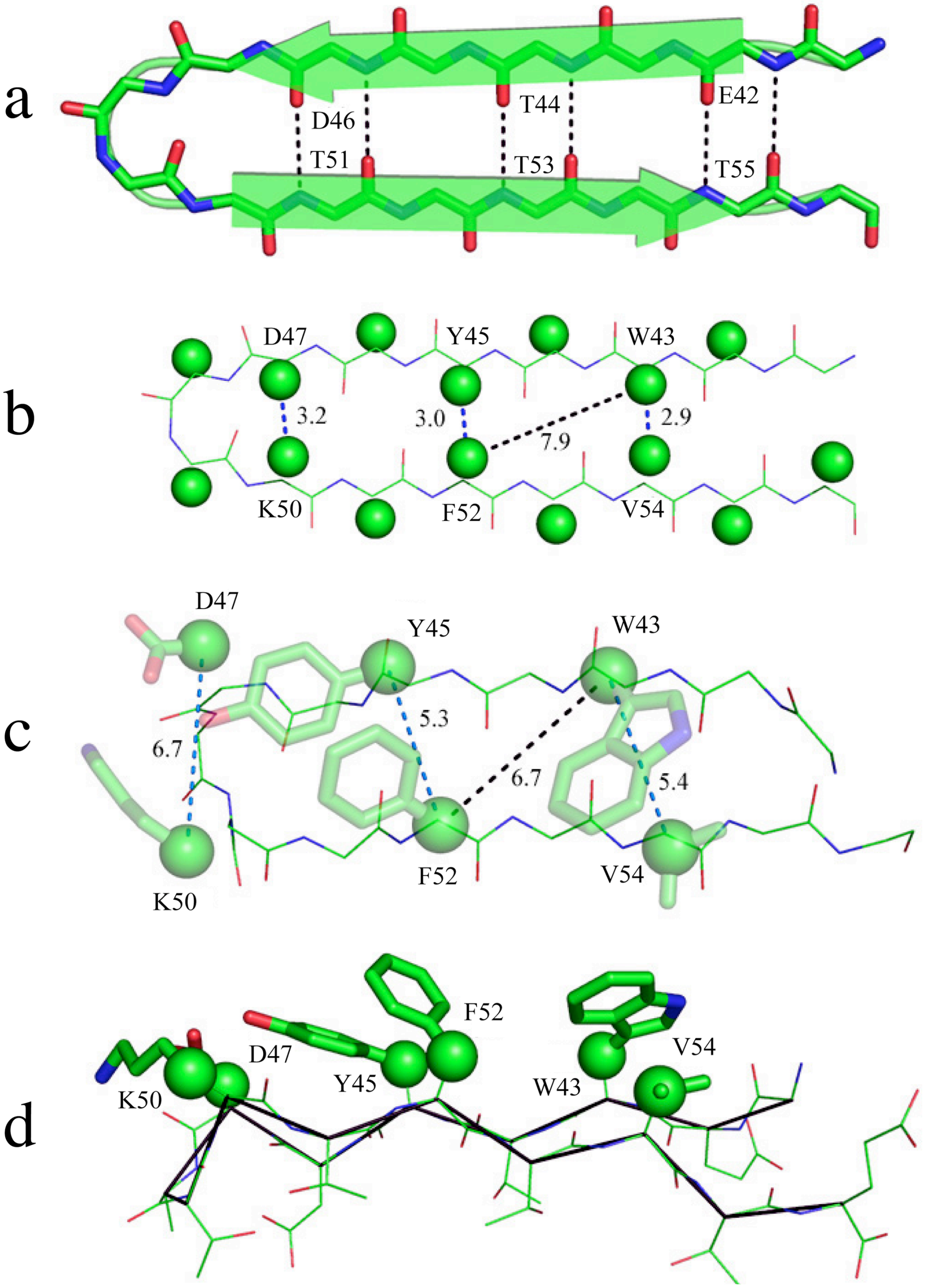


Figure 3: Conformational difference between ideal hairpin and G-hairpin.

The conformation of G-hairpin in its theoretical, ideal geometry is compared to its native conformation. a) The transparency cartoon view shows the β-hairpin conformation in its "ideal/perfect" geometry. All main-chain atoms are in the same plan and the black dash line shows the well-organized hydrogen bonds between 2 strands. b) View of the ideal hairpin perpendicular to the hairpin plane. Green spheres show the position of C_β atoms and distance between selected C_β atoms are measured as dash line. c) View of the hairpin in its native conformation at the same angle as b). Side-chains are shown for residues involved in hydrophobic interactions. Distance between C_β atoms corresponding to those in b) are also shown. d) View of the hairpin in c) as the hairpin rotating 90 degree into the paper. The size of the sphere for the C_β atoms is reduced for better illustration.

perturbations of the G41 β-hairpin structure, this ideal conformation is used as a reference point (Figure 3b). On the hydrophilic side of the ideal β-hairpin, all side-chain C_β atoms point away from each other. On the hydrophobic side, the C_β atoms point towards each other and result in a highly unfavorable Van der Waals clashes. The planarity of the backbone atoms as well as the perfect backbone hydrogen bonds places the C_β atoms of the hydrophobic core (residues Trp43, Tyr45, Asp47, Lys50, Phe52 and Val54) in such position that they overlap with each other: the 3.1 Å distance between C_β atoms, which is averaged over three pairs of distance shown in Figure 3b, is smaller than the sum of radii of two carbon atoms 3.6 Å (Figure 3b). Because they occur between C_β atoms, changing side-chain rotamers cannot relieve such clashes. Furthermore, the hydrophobic residues only pack in overlapping pairs and are restrained from forming a cluster with each other. To favorably accommodate the C_β atoms of the hydrophobic core, the G41 β-hairpin makes a number of deviations from this ideal. In a comparison of the native G41 that is part of the protein G B1 domain to the ideal β-hairpin (Figure 3c and 3d), it shows that the changes in the torsion angle rotate the strands along their own axes so that the C_β atoms on both sides of the β-hairpin are nearly perpendicular to the hairpin plane. Also, the non-planar turn, which bends over to the hydrophilic side, is stabilized by interactions involving Asp46 and it allows for both horizontal and vertical shearing between the strands. As seen in Figure 3c, the horizontal shear displaces the residues within the β-hairpin plane. Such shearing not only relieves the C_β clashes between residues 52 and 45, but also makes it possible for residues 52 and 43 to packing against each other, which completes the hydrophobic core. As seen from Figure 3d, the vertical shearing of 8° relieves the clash between residues 54 and 43 but keeps them packed. In summary, these conformational "rearrangement" not only increase the distance between C_{β} atoms on the hydrophobic side from 3.1Å to up to 6.6 Å, but the twisting of the strands from the ideal planer conformation also improves hydrophobic packing by increasing the interactions diagonally across the strands.[103]

Turn

Experimentally, the stability of the four-residue, type IV turn is crucial for the formation of the isolated β-hairpin[104,105] as well as the overall stability of protein G.[88] Theoretical studies support such findings by verifying that the turn formation is the early step during the β-hairpin folding.[106,107] It has also been shown that a well formed turn can increase the stability of the β-hairpin.[72-74] Table 2 lists a number of averaged properties of the turn for each β-hairpin variant. Overall, terminal ion pairs effect this portion of the structure the least. For the turn, all the ensemble structures have a stable turn with CαRMSD values no larger than 0.3 Å and very little variation. We find that the side-chain carboxyl group of Asp47 and the side-chain amino group of Lys50 forms an ion pair with its occupancy nearly 100% for all the variants. Also, all the structures increase the bending of the turn from 70º to around 110° towards the hydrophilic side and the torsion angles of such bending are very similar with comparable variations. As mentioned above, such bending enables the turn residues to form more interactions, especially hydrogen bonds. In particular, a number of hydrogen bonds with high frequency listed in Table 2 involve the side-chain carboxyl of Asp46 with other turn

Table 2: The properties of interactions in the turn region

Property			G41	K41	E42	G40	Ac-G41	Ac-K41	Ac-E42	Ac-G40
CαRMSD (Å)[a]		*Turn*	0.2(0.1)	0.2(0.1)	0.3(0.2)	0.2(0.1)	0.3(0.1)	0.3(0.1)	0.2(0.1)	0.2(0.1)
Ion Pairs (%)[b]		*IP*	97.6	99.8	100	99.9	96.8	99.5	96.2	97.8
Conformation (°)[c]	**Bend**	*Turn*	110(8)	109(10)	109(8)	110(8)	110(10)	109(13)	107(12)	112(7)
	Phi	*Asp47*	-91(14)	-95(14)	-96(16)	-94(14)	-92(15)	-96(15)	-90(14)	-92(13)
		Ala48	-85(26)	-91(23)	-83(22)	-92(24)	-86(23)	-89(22)	-83(25)	-92(21)
		Thr49	-100(14)	-99(13)	-101(14)	-99(13)	-98(14)	-98(13)	-101(15)	-97(13)
		Lys50	75(13)	76(13)	78(15)	76(13)	73(13)	73(16)	71(16)	73(12)
	Psi	*Asp47*	3(20)	5(20)	0(17)	6(20)	1(19)	3(19)	-1(19)	6(18)
		Ala48	-47(13)	-45(12)	-47(12)	-44(12)	-46(12)	-46(12)	-49(13)	-43(11)
		Thr49	-20(14)	-25(11)	-22(13)	-24(12)	-23(12)	-27(11)	-24(17)	-24(11)
		Lys50	60(14)	57(13)	57(14)	56(13)	60(13)	61(15)	65(18)	60(13)
Hydrogen Bonds (%)[d]		*Asp48m to Asp46s*	57	70	56	73	63	66	49	77
		Thr49m to Asp46s	82	99	99	99	91	99	92	99
		Thr49s to Asp46s	88	99	99	84	99	99	99	99
		Lys50m to Asp46s	29	57	57	51	36	59	54	26
		Thr51m to Asp46s	24	44	50	43	28	36	31	17
		Thr51s to Asp46s	26	49	53	44	30	37	36	16
		Lys50s to Asp47m	11	11	12	13	11	12	13	11
		Lys50s to Asp47s	99	99	99	99	98	99	99	99

[a] The average CαRMSD for the four-residue turn using G41 turn as reference structure. The numbers in parentheses are the standard deviation of the CαRMSD.

[b] The percentage of structures that have the ion pair formed at the turn region for each hairpin respectively. The ion pair forms between the sidechain carboxyl group of Glu50 and the sidechain amino group of Lys47.

[c] Conformation of turn in torsion angles. The Bend angle is measures as the torsion angle by pairing Cα atoms of residue 47 to 48 and Cα atoms of 49 to 50. The numbers in parentheses are the standard deviation of the torsion angle.

[d] The percentage of structures that have hydrogen bonds formed at the turn region. Different types of hydrogen bonds are labeled as follows: the amino group of donor and carboxyl group of acceptor is shown using 6 letters respectively. Three-letter residue code is followed by the residue number and then followed by the atom type, "m" indicates mainchain atom while "s" indicates sidechain atom. The amino group of the donor is always shown first. Only hydrogen bonds with occupancy bigger than 5% are shown.

residues that occur across all the variants. Asp46's carboxyl group interacts predominantly with both the backbone and side-chain of Thr49. The significance of Asp46 has been noted experimentally.[88] The high hydrogen bonding frequency between the side-chains of Asp47 and Lys50 indicates that the ion pair formed between these two groups is a true salt-bridge. In terms of water penetration, the side chains of the four turn residues prefer forming hydrogen bonds within the protein to the water molecules. For all of the eight hairpin variants, the turn shows similar conformation, with similar torsion angles, small CαRMSD values compared to the native turn, and similar occupancies of all the important interactions. As a result, the turn is insulated from changes at the termini by the interactions between the strands.

Strands

1. Backbone conformation

As pointed out above, the native G41 structure deviates from an ideal planar conformation to allow for better hydrophobic packing. We also expect more deviation from our analysis as the β-hairpins are isolated in the solution. Therefore, we used a broad range of torsion angles to define the β-sheet conformation: the backbone torsion angles of $\phi = -180$ to $-30°$ and $\psi = 60$ to $180°$ and -180 to $-150°$.[108,109] Table 3 summarizes the variation of the strand conformation for each β-hairpin variant. No patterns are found across all the variants, which suggest that the ensembles of these peptides are rather heterogeneous within the isolated β-hairpin region. The only

Table 3: The properties of strands

Property	Type	G41	K41	E42	G40	Ac-G41	Ac-K41	Ac-E42	Ac-G40
Hairpin conformation (%)[a]	Thr55	32	38	36	28	46	52	26	52
	Val54	61	33	42	55	60	32	62	36
	Thr53	40	73	66	50	30	35	56	40
	Phe52	53	58	57	61	28	49	43	64
	Thr51	99	95	82	99	99	80	80	97
	Asp46	99	96	99	99	95	99	99	199
	Tyr45	38	43	44	62	21	57	54	51
	Thr44	20	41	42	48	30	38	43	22
	Trp43	67	67	37	51	71	69	71	45
	Glu42	67	34	–—	31	15	23	5	33
	Averag	58	58	51	58	50	56	50	54
Shear (°)[b]	*Strand*	18(12)	21(16)	12(18)	16(8)	30(15)	8(21)	8(24)	28(10)
Hydrogen Bonds[c]	*Native*	3.7(1.1)	2.5(1.6)	2.9(1.6)	3.6(1.3)	3.4(1.6)	2.4(1.8)	2.9(1.9)	3.8(1.0)
	Total	5.7(1.6)	3.7(2.0)	4.4(1.8)	5.4(1.9)	5.4(2.3)	3.2(2.5)	4.8(2.9)	5.6(1.4)
Native Hydrogen Bonds (%)[d]	Glu42 - Thr55	*	-	-	*	*	-	-	*
	Thr55 - Glu42	39	-	5	45	39	12	56	39
	Thr44 - Thr53	65	44	71	53	69	20	54	53
	Thr53 - Thr44	33	25	34	48	21	40	28	50
	Asp46 - Thr51	90	69	73	85	79	57	61	88
	Thr51 - Asp46	79	59	60	65	67	53	43	81
	Lys50 - Asp46	75	66	59	71	71	64	60	83
Non-Native Hydrogen Bonds (%)[e]	*Glu56 - Gly41*	8	*	-	26	13	*	43	*
	Thr55 - Gly41	9	*	-	21	26	*	47	*
	Glu42 - Val54	8	*	-	*	*	*	-	*
	Trp43 - Val54	-	-	28	-	-	*	-	-
	Val54 - Trp43	13	-	-	11	-	14	6	25
	Thr53 - Trp43	9	*	*	10	*	13	7	18
	Thr44 - Phe52	27	13	14	22	46	5	16	22
	Tyr45 - Thr53	17	22	21	*	9	*	*	14
	Tyr45 - Phe52	34	15	22	25	51	6	20	29
	Tyr45 - Thr51	*	*	10	*	*	*	*	*
	Ala48 - Asp46	13	14	7	14	12	11	9	17
	Thr53 - Thr51	20	24	23	5	10	*	6	16

Table 3: Continued

Property	Type	G41	K41	E42	G40	Ac-G41	Ac-K41	Ac-E42	Ac-G40
HCSA (Å^2)[f]	*Total*	91(14)	93(18)	88(18)	84(17)	85(16)	95(19)	79(21)	86(15)
	Cluster	72(15)	75(19)	69(17)	68(16)	68(16)	79(19)	64(19)	66(14)

[a] The percentage of ensemble structures that have hairpin conformation.

[b] The shearing angle between two strands. Standard deviations are shown in parentheses.

[c] The total number of all hydrogen bonds formed in the ensemble structure. Native strands for hydrogen bonds can be observed in the native structure while Total stands for all hydrogen bonds in the ensemble structure.

[d] The average number of native hydrogen bonds formed in the ensemble structures. (*) indicates that the hydrogen bond is not significantly formed (below 5%) while (-) shows that it is not formed at all. The hydrogen bonds are labeled from the amino group of the donor to the carbonyl group of the acceptor. Standard deviations are shown in parentheses.

[e] The average number of non-native hydrogen bonds formed in the ensemble of structures.

[f] The hydrophobic contact surface area (HCSA) between strands. The Total averages over the sum of all the side-chain hydrophobic contacts, while the Cluster averages over only the largest hydrophobic cluster. Standard deviations are shown in parentheses. (see Methods)

observation across all variants is that residues closest to the turn region (Asp46 and Thr51) are most consistently in the β-sheet conformation across all the variants. The remaining residues are less often in the β-sheet conformation and the occupancy decreases as the residue become closer to the termini. Similarly, the residues on the hydrophobic side that makes up the hydrophobic core (Trp43, Tyr45, PHe52, and Val54) sample the β-sheet conformation more often than those do on the hydrophilic side (Glu42, Thr44, Thr53, and Thr55). In Table 3, the vertical shear angle between the two strands of the β-hairpin is also measured. If we assume the G41 ensemble as a reflective of the native conformation, then the non-acetylated peptides remain closest to the G41 ensemble's average value of 18°. Of these variants, the K41 ensemble shows the highest average of shearing, which again indicates some perturbation on the conformation to accommodate the K41 side chain. The acetylated ensembles' averages differ substantially from the G41 ensemble average. Ac-G41 and Ac-G40 are upwards 30° and 28° respectively, while Ac-K41 and Ac-E42 have a low value of 8° with large deviations extending into values of negative shearing. Overall, these results show that the ion pair at the termini does not restrict conformational sampling by the backbone, but instead inhibits the strands from separating.

2. Main-chain hydrogen bonds

There are seven main-chain to main-chain hydrogen bonds observed in the native G41 from the PGB1 structure as shown in Figure 2b. These hydrogen bonds represent the signature interactions of the anti-parallel strands in the β-hairpin. The frequency is shown in Table 3 and a schematic representation is shown in Figure 4, where broken

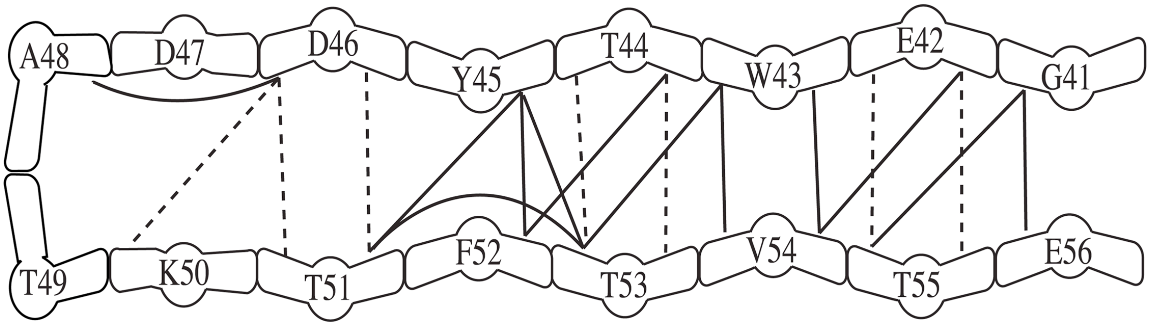


Figure 4: A schematic representation of the backbone hydrogen bonds.
A schematic representation of the backbone hydrogen bonds observed in the ensemble structures generated from the MD simulations. Residue is represented as two arms (amino group and carboxyl group) with their residue number in the middle. Lines are drawn between two arms indicating hydrogen bonds. The dashed lines represent possible hydrogen bonds in the crystal structure of PGB1 (native hydrogen bonds in Table 3) and the solid lines represent hydrogen bonds only observed in the ensemble structures (non-native hydrogen bonds in Table 3).

lines represent native hydrogen bonds. In all of the variants, the native hydrogen bond closest to the termini (amino of Glu42 to carbonyl of Thr55) is virtually never made, which is direct evidence that the ends have frayed. As with the backbone conformation, no strong patterns of hydrogen bonding are observable across β-hairpin variants, with or without ion pairs. In general, stronger hydrogen bonding is found nearer the turn. Also, an interesting periodicity based on strand direction is noticeable. Native main-chain hydrogen bonds are more stable between backbone amino groups of the incoming strand (residue Gly41 to Asp46) to the carbonyl groups of the outgoing strand (Thr51 to Glu56). Following these observations, the most consistent backbone hydrogen bonds occur between the amino group of Asp46 and the carbonyl group of Thr51. Table 3 also shows frequencies of non-native hydrogen bonding, which are represented by solid lines in Figure 4. All of these are low in frequency, which is less than 50% for all the β-hairpin variants. It suggests that the variants can sample more conformations that may deviate from conformation of the native β-hairpin but still keep that general hairpin structure. For instance, the hydrogen bond between the amino group of Thr44 and the carbonyl group of Phe52 involves the wrapping of the two strands around each other, which effectively causes a flipping of the carbonyl group of Phe52 from the outside to the inside of the β-hairpin. On the other hand, it is also interesting to note that on average the K41, which makes a very consistent terminal ion pair, makes fewer native and non-native backbone hydrogen bonds than other β-hairpin variants do. Again, the terminal interaction involving the long side-chain of Lys41 causes a distortion that appears to prevent proper backbone hydrogen bonds. As a measure of strand separation, we also analyzed the hydrogen bonding of main-chain groups to the surrounding water molecules (data not shown). Because of the fraying at the ends, it is not surprising that the polar groups at the termini are more exposed to the water molecules. Of all the variants, the Ac-G40 ensemble shows a tendency for more exposure to water, which suggests a relative open structure or less stability. Consistent with the analysis on the backbone conformation of the β-hairpin, the terminal ion pairs help to stabilize the β-hairpin structure by preventing the strands from opening up, but do not significantly restrict the structural conformation, except in the case of the K41 ensemble.

3. Hydrophobic contacts

Experimental studies supported by the theoretical work showed that moving the hydrophobic cluster closer to the termini on the hydrophobic side is destabilizing while moving it closer to the turn is stabilizing. Furthermore, G-hairpin's four middle hydrophobic residues (Trp43, Tyr45, Phe52 and Val54) are necessary for its stability. In experiments, when either Tyr45 or Phe52 are replaced with Ala, the stability of the G-hairpin substantially decreased. Molecular dynamics (MD) simulations indicate that at least three of these four hydrophobic residues are required to maintain the β-hairpin conformation.[90] As shown in Table 3, more than 75% of the total HCSA (hydrophobic contact surface area) in all the peptide variants is made up of the clustered hydrophobic interactions (see Methods). Consistent with previous studies, all the hydrophobic clusters involves the 4 interior hydrophobic residues: Trp43, Tyr45, Phe52 and Val54. The overall averages in Table 3 show that K41 and Ac-K41 produce the largest clustered HCSA, while Ac-E42 has the least. Even so, only small differences separate the peptide ensembles based only on the overall side-chain contribution to hydrophobic burial. Providing in greater details, Figure 5 diagrams the contribution of different residues to the HCSA that is averaged for each β-hairpin variant, where the line thickness relates to the frequency in the ensemble population. For all the ensemble structures, residues Asp47 and Lys50 also contribute to the hydrophobic cluster and thereby connect the β-hairpin turn with the hydrophobic core. The variation among different variants can be seen as an interchange between the residues at the termini and the four residues that make up the hydrophobic core. For G41, the ion pair at the termini is tenuously joined to the hydrophobic cluster through a hydrophobic contact between Val54 and Glu56. The acetylated form of G41 exhibits the same cluster, but it is less prevalent. Including all the interactions of the G41 ensembles, the K41 ensemble strongly links the terminal ion pair from Lys41 through both Trp43 and Val54. This capping of the hydrophobic core may explain the rigidity of the K41 ensemble. The Ac-K41 ensemble also shows a similar capping pattern of its terminal ion pair. The remaining four peptides exhibit different but interesting behavior. The acetylated versions of G40 and E42 produce interactions within their hydrophobic core with higher frequency than their non-acetylated counterparts do. This result reveals a trend supported generally by all the β-hairpin variants as well as previously noted:[92] the stability relies on the hydrophobic core much more without terminal ion-pair interactions.

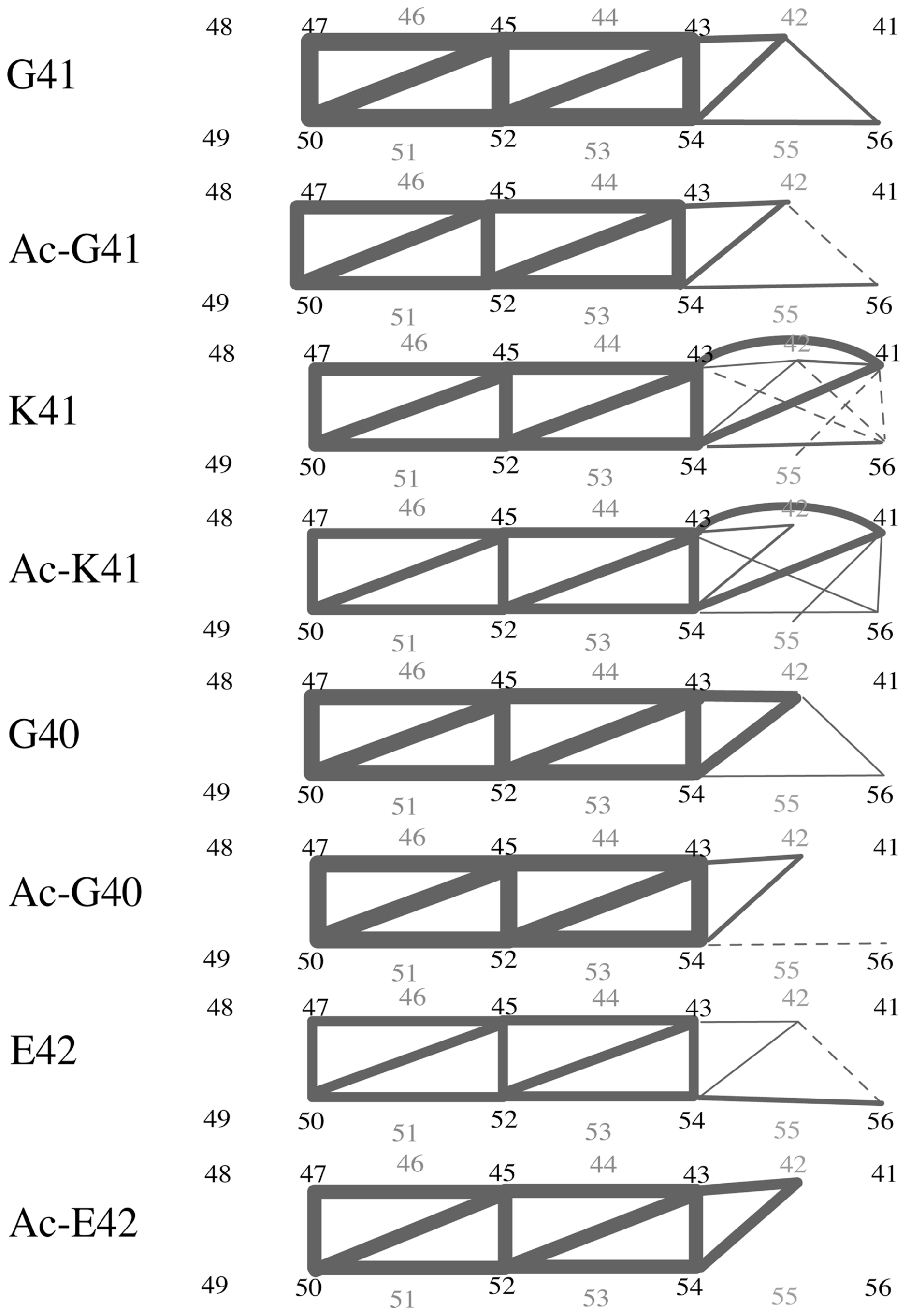


Figure 5: The contribution of different hydrophobic contacts.

The contributions of different hydrophobic contacts to the hydrophobic core are shown. Residue numbers are used to indicate position of residues in the hairpin conformation and lines connecting two residues are used to represent hydrophobic interactions. The thickness of the line is proportional to the occurrence of the hydrophobic contact in the corresponding ensemble structures. Residue numbers in gray have side chains that locate in the hydrophilic side of the hairpin.

To further investigate this phenomenon, we analyzed the HCSA distribution in the G41 based on whether or not the terminal ion pair forms (Figure 6). The distributions clearly show that the terminal ion-pair interaction compensates for a certain amount of hydrophobic interaction, such that formation of the terminal ion pair requires less HCSA to form a stable hairpin.

4. Termini

Table 4 shows how often and what type of terminal ion-pair interaction is formed for each β-hairpin variant with the total number of possible ion pairs and a schematic representation of different chemical groups that can form ion pairs. The frequency of ion-pair interaction can be loosely classified based on the peptide variant's potential to make ion pairs, which begins from K41 with two of four possible ion pairs to G40, Ac-K41, G41, and E42 with one out of two, and lastly to Ac-G41, Ac-E42, and Ac-G410 with no ion pairs. K41 and G40 form ion pairs close to 100% during their simulations. Since K41 could possibly form four different types of ion pairs (2 at same time), we observe that more than one is populated (average of 1.1 per structure for K41). The observed NOE contact between the Cα proton of Glu56 and the methylene protons of Lys41[71] corroborate these findings that an ion pair is well formed in the K41 peptide. Because G40 can form only one out of two possible ion pairs, the result for G40 is surprising and suggests that such extension of one residue favors ion pair between the G40 amino group and the E56 side-chain carboxyl group.

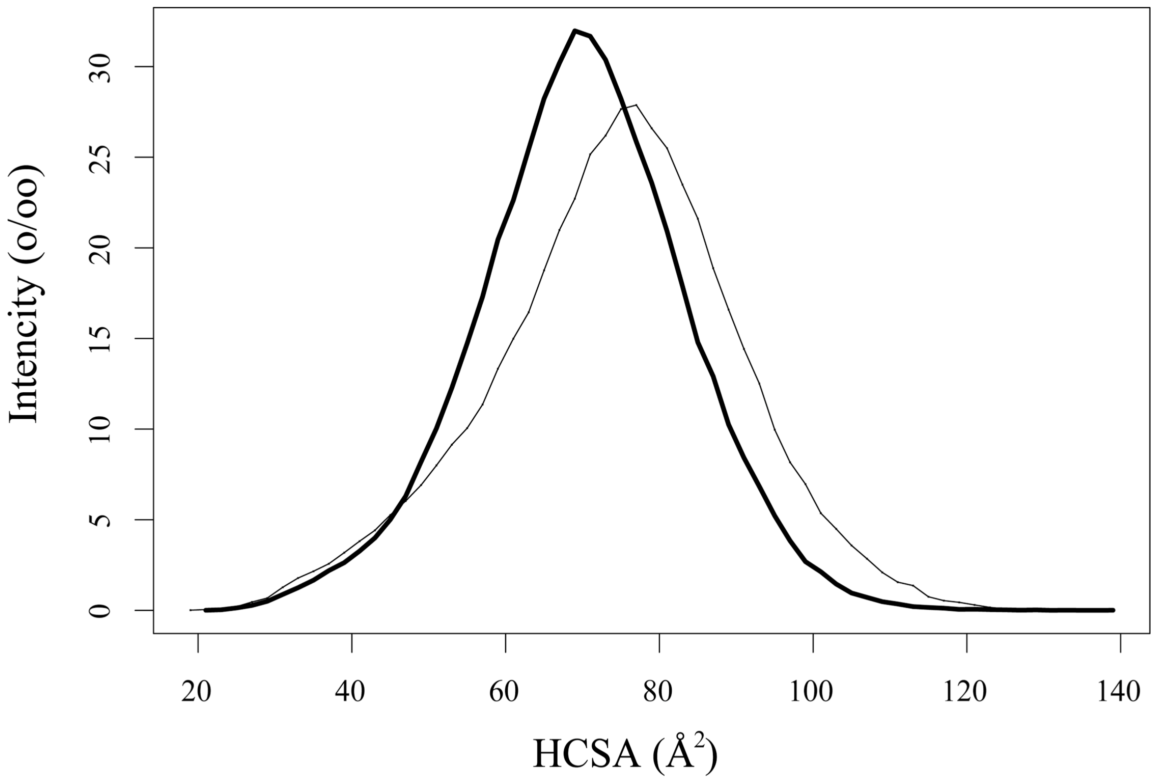
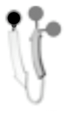
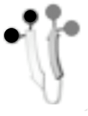
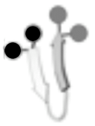
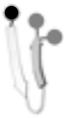
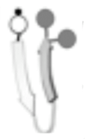
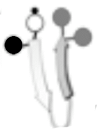
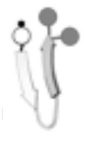
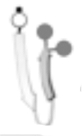


Figure 6: HCSA distribution of G41 for structure with/without terminal interaction.

HCSA (hydrophobic contact surface area) distribution of ensemble structures with and without terminal ion pairs is shown for the hairpin G41. Large HCSA values indicate strong hydrophobic interaction. The thin line is the distribution of structures without terminal ion pairs formed while the thick line is the distribution of structures with terminal ion pairs formed.

Table 4: Terminal ion pair (IP) and CαRMSD for the MD structures

Property	Category	G41	K41	E42	G40	Ac-G41	Ac-K41	Ac-E42	Ac-G40
Terminal IP	*Possible IP*[a]	2	4	2	2	0	2	0	0
	IP[b]	0.6	1.1	0.6	1	–	0.9	–	–
	SB[c]*(%)*	92	78	84	70		65		
Type of Terminal IP[d]	*Nm to Os*	0.2	0	0	0.9	–	–	–	–
	Nm to Om	0.4	0.1	0.6	0.1	–	–	–	–
	Ns to Os	–	0	–	–	–	0.1	–	–
	Ns to Om	–	1	–	–	–	0.8	–	–
	Schematic[e]								
CαRMSD (Å)[f]	*Hairpin*	1.8(0.6)	2.1(0.5)	1.7(0.7)	1.8(0.5)	2.7(0.9)	2.7(0.7)	2.5(1.6)	3.0(0.8)
	IP	1.6(0.4)	2.1(0.5)	1.5(0.3)	1.8(0.5)	–	2.8(0.7)	–	–
	No IP	2.1(0.7)	3.1(0.2)	2.1(0.8)	2.5(0.4)	2.7(0.9)	1.7(0.3)	2.5(1.6)	3.0(0.8)
	Ideal	3.1(0.4)	3.2(0.5)	2.8(0.6)	3.0(0.4)	3.6(0.6)	3.2(0.6)	3.2(0.9)	3.5(0.4)

[a] The total number of possible ion pairs involving the N- and C- terminal residues.

[b] The average number of ion pairs in the ensemble of structures. The dashed lines indicate no ion pairs.

[c] The percent occurrence that the ion pairs also form salt bridges, defined by distance and angle restraints for hydrogen bonds.

[d] The type of possible terminal ion pairs include N-terminal (Nm) or Lys41 side chain amino group (Ns) to C-terminal (Om) or Glu56 side-chain carboxyl (Os).

[e] Representation of possible terminal ion pairs between the backbone and side chains with amino groups in black, carboxyl groups in grey, and the acetyl group in white.

[f] The average CαRMSD from the structure of G-hairpin in the crystal structure of PGB1. The CαRMSD values for the hairpin were subdivided into structures that have or do not have terminal ion pairs, respectively and the CαRMSD value using the ideal hairpin as reference structure is also listed. The numbers in parentheses are the standard deviations of the CαRMSD.

For all the remaining three variants that can form one out of two possible ion pairs, the Ac-K41 forms the ion pair between the side-chain amino group of Lys41 and main-chain carboxyl group of Glu56 with frequency of 90%, which is very similar to what is observed in the K41 variant. G41 and E42, the remain two variants that can form one out of two ion pairs, show reduced frequencies at about 60%. G41 splits its ion pair between two types while E42 favors the ion pair forming between the main-chain groups. Because the majority (at least >65%) of the ion pairs can also be classified as hydrogen bonds, we could also consider these interactions as salt-bridges.

The importance of ion pairs in sustaining low CαRMSD values during the simulations is shown in Figure 7 where three individual simulations of G41 are shown with CαRMSD values (Figure 7a) and the distance between the two charged terminal groups that can form ion pair, Gly41 and Glu56 (Figure 7b). When ion pairs are formed, which is defined as the distance of two opposite charged groups is smaller than 3.5 Å, (the blue curve in Figure 7b), lower and stabilized CαRMSD values are observed (blue curve in Figure 7a). Whereas, when ion pairs are not formed (the black curve in Figure 7b), larger and increasing CαRMSD values are observed (black curve in Figure 7a), that increase over the entire trajectory. In particular, the appearance of an ion pair forming in the middle of simulation (red curve in Figure 7b), the CαRMSD value decreased and become stable for the rest of the simulation (red curve in Figure 7a). Similar trends can be observed for other β-hairpin, which indicates that the forming of terminal ion pair can stabilize the hairpin.

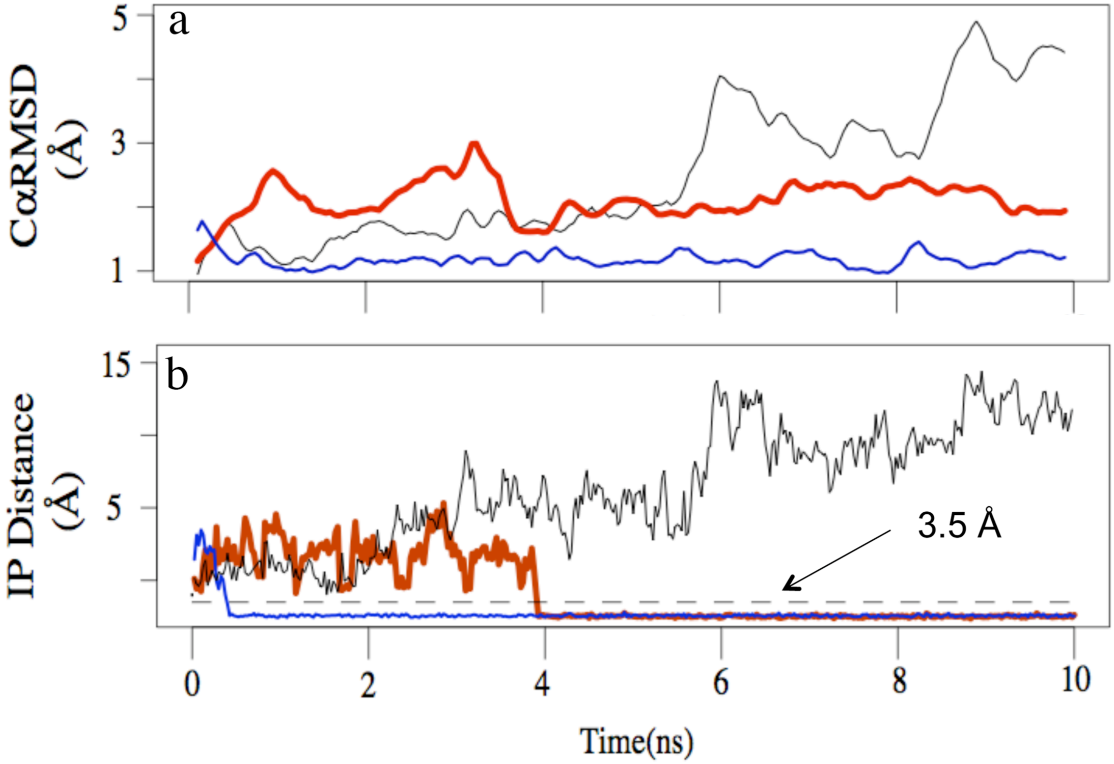


Figure 7: The CαRMSD value and distance of terminal ion pair over time. CαRMSD values and distance of terminal ion pair over time are shown. Data were plotted for three individual simulations of G41 over 10 ns. a) shows the CαRMSD values over time for 3 simulation. b) shows the distance between two terminal groups (amino group and carboxyl group) that can form ion pair over time. The distance was calculated as the shortest between two terminal groups and whenever the distance is smaller than 3.5 Å (dash line), the ion pair is formed. The blue line indicates the ion pair is formed at the very beginning of the simulation, red line indicates the ion pair is formed in the middle of simulation and the black line indicates no ion pair is formed during the simulation. (Fig 7 in Huyghues-Despointes, B. M. and et al. (2006) *Proteins* **63**, 1005-17[71]).

In Table 4, CαRMSD values of the eight G-hairpin variants are shown with different classifications, which can be used to indicate the relative stability of the hairpin. The CαRMSD values were measured against the coordinates of the G41 in X-ray structure for each variant. We assume that larger CαRMSD values with larger deviations indicate less stability. In general, the ensemble-measured stability is consistent with the experimental results that the acetylated hairpins are less stable than the non-acetylated ones.[71] Of the non-acetylated peptides, it is surprising that the K41 peptide ensemble exhibits the largest average CαRMSD value, suggesting that the β-hairpin conformation have to be changed to maintain its ion-pair interactions.

To better compare to the experimental determined stability of eight G-hairpin variants, the distribution of the CαRMSD values for each hairpin variant are shown in Figure 8. For all the variants, there is an initial of ~1.2 Å CaRMSD change in the value because the G41 was modified for each particular variant and minimized in the presence of water before each simulation (see Methods). Although, the same CαRMSD value doesn't directly imply the same structure, a single peak in the CαRMSD distribution suggests a single population of hairpin conformation with less variation in the structure, in another word, more stable structures. Based on this, the difference between non-acetylated and acetylated β-hairpin variants is clear: the non-acetylated variants show populations generally exhibit narrow CaRMSD distributions with mean values below 2.5 Å, while the acetylated distributions are wider with mean values above 6 Å.

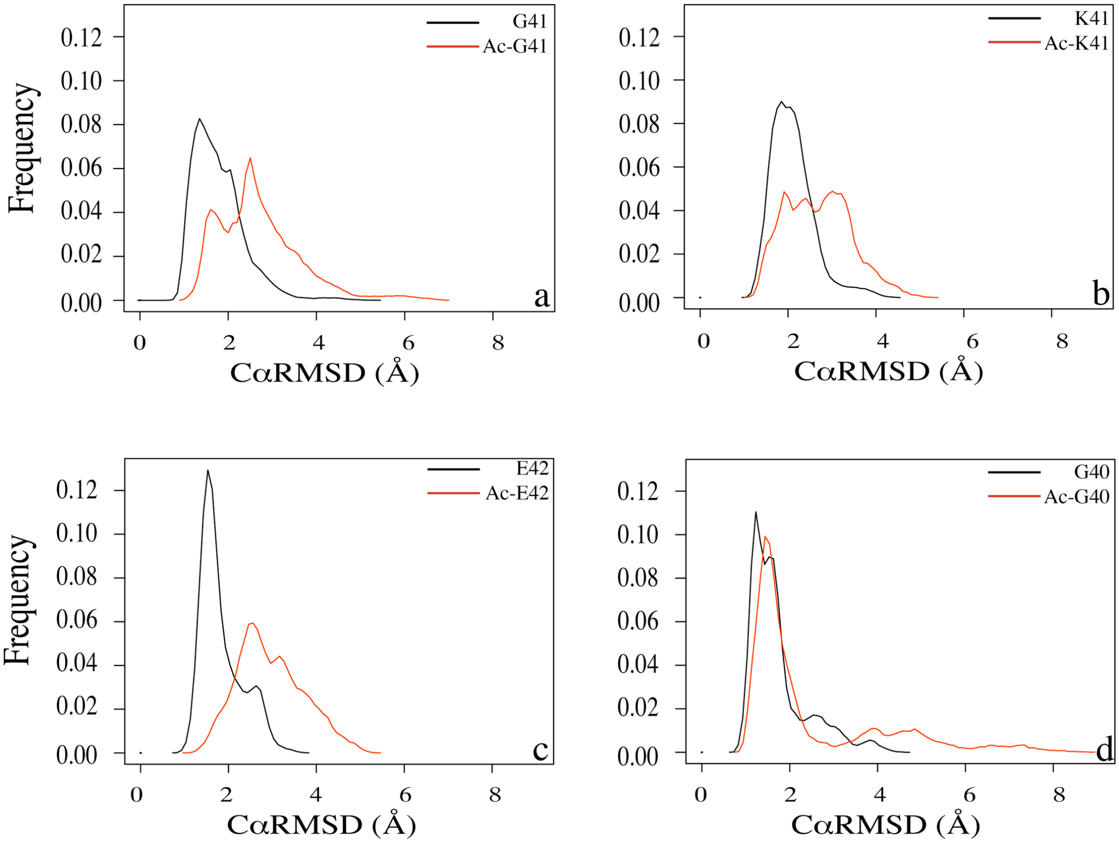


Figure 8: The distribution of CαRMSD for each β-hairpin.

The probability distribution of the average CαRMSD in the MD-generated ensemble of structures of a) G41/Ac-G41, b) K41/Ac-K41, c) G40/Ac-G40, d) E42/Ac-E42. Non-acetylated peptides are in black, and acetylated peptides are in red. (Fig 8 in Huyghues-Despointes, B. M. and et al. (2006) *Proteins* **63**, 1005-17[71]).

Conclusion

Although the contribution of weak interactions on β-hairpin stability, such as an ion pair, is believed to be hard to observe, a previous study[110] showed that the total effect of 2 ion pairs (one near the turn, the other near the end) is bigger than the sum of each single ion pair. Given the fact that there is one ion pair interaction observed in the G41 structure between the side-chain carboxyl of Asp47 and the side-chain amino group of Lys50, which is with 100% occupancy in all eight β-hairpin variants in our study, we expected to be able to observe the effects of terminal ion-pair interactions in our study. We have studied the changes in CαRMSD, backbone torsion angles, hydrogen bonds, hydrophobic interactions and turn conformation for β-hairpin structures with different ion pair interactions at the termini. Our results indicate that the presence of the terminal ion pair(s) doesn't fix a structure into a more β-hairpin like conformation. Instead, the ion pair improves the hairpin stability by preventing the β-hairpin's strands from fraying. In a survey of protein structures, it was found that anti-parallel β-sheets more often finish with non-hydrogen bonded residues.[111] Thus the presence of a terminal interaction can act like another pseudo hydrogen bond ring that caps the open end of the β-hairpin. Such a pseudo hydrogen bond ring has the potential to prevent the end from opening up, which is believed to be the first event of the unfolding of the β-hairpin. Also, terminal ion pair can cooperate with the hydrophobic core interactions in two ways: first, by directly take part in the hydrophobic cluster, as seen in the K41 variant, to increase the overall strength of hydrophobic interaction. Second, while fixing the ends with a terminal ion pair decreases the entropy, such entropy loss can be compensated by the added entropy gain in the hydrophobic core as seen in Figure 5 and Figure 6, which suggests that by forming a terminal ion pair, it reduce the requirement of forming a tight-pacing hydrophobic core as well as increase number of conformation with different combination of hydrophobic interactions. In other words, the forming of the terminal ion pair can potentially broaden the energy well to allow more flexibility in the β-hairpin conformation.

Materials and Methods

Generating the ensembles

We ran a total of 108 simulations lasting 10 ns each using the potential energies and the F3C water model in the ENCAD program.[112-114] The coordinates of G41 (C-terminal residues 41 to 56) were taken from the crystal structure of 1PGB,[100] placed in a box of water, and minimized. The box of water was trimmed so that the edges were at least 8 Å away from the closest protein atom. All water molecules within 1.67 Å of the protein were removed and the box sides were corrected to match the density of water (0.997 g/ml) at 298 K.[115,116] Box sizes and number of water molecules for each variant are given in Table 1. Sodium or chloride ions were used to replace water molecules at random positions to yield an electrically neutral system. This system was relaxed by performing 3,000 conjugate gradient energy minimization steps in the following order. First, the protein was fixed, and the water molecules were minimized over 1,000 steps. The protein was then minimized in 1,000 steps, holding the water molecules fixed. Finally, the whole system was relaxed in the last 1,000 steps. To obtain different runs, the system was equilibrated to 298 K with different random seed numbers. During the simulation, the coordinates of the structure were updated at two femtosecond intervals and sampled every picosecond (or 500 steps). Therefore, each 10 ns simulation generated 10,000 structures. Modifications were made by modifying the structure file of the G41. The coordinates of backbone atoms remained unchanged while the side-chain atoms that were to be changed were deleted and regenerated within ENCAD using standard residue conformations before the water was added.

Ideal "planer" hairpin

The "ideal" β-hairpin structure was built based on the following criteria. First, all the backbone atoms were placed in the same plane. Second, for proper anti-parallel structure, main-chain hydrogen bonding rings should be made between residues 46 to 51, 44 to 53, and 42 to 55, respectively. Hydrogen bond distances needed to be below 3.5 Å and hydrogen bond angle greater than 120° but less than 180º between atoms N, O and C. Third, the turn beginning at residue 47 and ending at residue 50 was placed in the same plane as the strands are. Based on these requirements and a standard residue conformation,[117] the backbone was first defined. The angles between each bond formed by backbone atoms are all set to 120º and the bond length for N-Cα, Cα-C and C-N is set to 1.44 Å, while it is 1.24 Å for C=O. Then 2 anti-parallel strands with 8 residues were aligned with each other to allow proper hydrogen bonding between strands. Finally, keeping Cα atoms of residue 46, 47, 48 and residue 49,50,51 in a line respectively with Cα atoms of residue 46 and 51 fixed, the turn is made by rotating these two lines towards each other simultaneously in the same plane of strands until the Cα atoms of residue 48 and 49 are within 3.8 Å, which is the distance between Cα atoms for

2 peptide bonded standard residues. Finally, the side chains are built on using the native conformation from the X-ray structure of G-hairpin.

Analysis

Programs written in C and PERL were used to analyze the structures resulting from the various MD simulations. The structures were viewed using Pymol.[118] Only the structures within the last 9 nanoseconds were used to analyze the β-hairpin structural features and patterns of stabilizing interactions (hydrogen bonds, hydrophobic contacts, and ion pairs). For each β-hairpin variant, properties were averaged over all ensemble structures. CαRMSD values were calculated using the method of Kabsch and Sander.[119] The frequency were calculated and represented using the R package.[120]

Hydrogen bonds were defined as those between the donor hydrogen and the acceptor oxygen, where the distance between hydrogen and oxygen was less than 2.6 Å and the angle formed by the acceptor oxygen, hydrogen, and the donor atom had to be greater than 120º. The definition of an ion pair was based on a simple distance cutoff, which is 3.5 Å between the positively charged nitrogen of the amino group and the negatively charged oxygen of the carboxyl group. Because a salt-bridge is a hydrogen-bonded ion pair, its definition required satisfying both the hydrogen bond and ion pair.

The hydrophobic contact surface area or HCSA was calculated using the Voronoi Polyhedra method.[121] Two carbon atoms sharing a polyhedron face are considered as a contact and the area of that face is defined as HCSA of such contact. Finally, hydrophobic clusters were defined by the biggest side-chain to side-chain contact network. For example, if residue 41 contacts residue 42 and residue 42 contacts residue 43, then all residues were considered to belong to one cluster.

The geometry of the equilibrium ensemble β-hairpin variants was compared to the "ideal" planar conformation in the following features in the turn bend angle and the strand shear. The bend angle of the turn was defined as the torsion angle between Cα atoms of residue Asp47, Ala48, Thr49, and Lys50. The strand-shearing angle is measured between the lines formed by the Cα atoms of residues Gly41 to Asp46 and Lys50 to Thr55, respectively.

CHAPTER III

CHARACTERIZING THE BACKBONE CONFORMATIONAL DEPENDENCIES OF RESIDUES USING MOLECULAR DYNAMICS SIMULATION

Overview

In a fine-grained analysis of protein structure, we investigated the relationship that a residue's backbone conformation has with its side-chain packing as well as conformation. To produce continuous distributions for each amino acid, we ran molecular dynamics simulations over a set of protein folds (dynameome). In effect, this dynameome samples the near-native conformational space of protein structures. As an extensive set of data, the dynameome has the advantage of representing known conformations that are not well represented in the structure database (PDB). In our analysis, we characterized the mutual influence that the backbone ϕ,ψ angles have with the first side-chain torsion angle χ_1 and the volume occupied by a side chain, respectively. Furthermore, we explored the dependency of these relationships on side-chain environment and amino acid identity. Generally, our results imply somewhat counterintuitively that side-chains pack more densely in regions where extended β-sheet backbone conformation is preferred and less densely in regions where α-helical is preferred. As expected, residue volumes on the protein surface were larger than those in the interior. For the first side-chain torsion angle χ_1, our results are consistent with previous studies of known protein structures, but with higher resolution. We found that the *gauche$^-$, gauche$^+$,* and *trans* rotamer conformation show ψ dependent patterns, and variations in the χ_1 value are only skewed to one side of their canonical values. By demonstrating the utility on dynameomic modeling of the native state ensemble, this study reveals the interplay among backbone conformation, residue volume and side-chain conformation.

Background

During the past 14 years, the progress made in predicting protein structure from amino acid sequence has been accomplished with simple representations of side chains, for example, as a single centroid. As a step to improve prediction accuracy, we pursue a higher resolution description of the relationship between backbone and side-chain conformation. In particular, our study seeks to better understand the determinants of this relationship.

It has been proposed that the local main-chain conformation has the greatest influence in determining the side-chain conformation.[122] Given the native backbone conformation, accurate packing of side chains can be achieved.[123] At the same time, side-chain conformation also effects the backbone conformation.[124] The most widely used description relating backbone and side-chain conformation are rotamer libraries.[125-130] A rotamer library clusters the observed conformations of side-chains into groups, from which Bayesian distributions can be derived. Populated rotamers are thought to reflect local minima on a potential energy map or to represent an average conformation over some region of dihedral angle space.[128] Even though recent rotamer libraries have benefited from the increased number of structures (especially high-resolution structures) in the PDB,[3] these libraries' coverage of conformational space is still limited due to the sparse sampling in the PDB and the fact that PDB structures are closely clustered around the minimum-energy X-ray crystal structure. Furthermore, structures deposited in the PDB seldom reflect the simplified states of side-chain conformation in rotamer libraries.[131] Broad distributions of side-chain dihedral angles are often observed.[132,133] Many rotamer conformations that can be accommodated by residues, such as those on the surface, are highly under-represented in crystallographic structures. Thus, sampling side-chain conformations from a continuous conformational space would provide higher accuracy.

Second, to reduce system complexity as well as to address the inadequate sampling, rotamer libraries bin side-chain conformations based on the three most populated rotamer conformations around each bond between heavy atoms: *gauche*$^{+}$*,* *gauche*$^{-}$ and *trans*. In addition, the backbone conformation is also binned to discrete areas of secondary structure space. By defining side-chain conformations in this way, rotamer libraries decrease the combinatorial complexity of packing/placing side chains in protein structure prediction. The result of this approximation is that rotamer libraries are a low-resolution description of the relationship between backbone and side-chain conformation. Suggested library improvements include adding extra information, such as a side-chain-orientation-dependent term[134,135] or the addition of solvated rotamers, in which several water molecules accompany the rotamer.[136] Moreover, a refined rotamer library, in which only high resolution, non-clashed side-chains are included with smaller and more continuous bins has greatly improved the accuracy over other rotamer libraries.[129] However, current approaches using rotamer libraries are reaching their limits.[137,138] In template-based modeling, starting templates cannot be refined towards the native structure because current methods cannot resolve over- or mis-packed side chains.

As a step towards improving the refinement of protein models, we have undertaken a study that provides a more detailed description of the relationship between a residue's backbone conformation and its side chain conformation. To produce a more complete view of the native state in the conformation space, we follow the approach of previous work[139] and generate a dataset of molecular dynamics (MD) simulations over a set of protein folds (dynameome).[140] These dynameomic approaches have been shown to accurately sample the structures in the near native conformational space across different protein folds and reproduce the ensemble properties of the native state environment.[9,141-143] Therefore, the purpose of our dynameome dataset is to model a more continuous set of native conformations, as opposed to the classic use of molecular dynamics simulations for time dependent information. Containing over 4 million structures, this dynameome allows a more refined view of protein structure. Specifically, we investigate the mutual dependence of backbone conformation (ϕ,ψ), the volume occupied by the residue and the first side-chain rotamer angle (χ_1). Our analysis finds that side-chain volumes exhibit a somewhat counterintuitive dependence on secondary structure. In addition to previous analysis of the PDB database, we detail the backbone's influence on each of the 3 χ_1 rotamer angles. We also investigate the effect that χ_1 has upon residue volume. Furthermore, we discuss the physical basis for each of this analysis.

Results and Discussion

The dynameome dataset

In this study, the purpose of the dynameome dataset is to provide a more complete sampling of the native conformational space instead of the usual kinetic properties measured in MD simulations. As a first step, we chose a set of structures that broadly represents all protein folds. Using the SCOP[144] classification (Table 5), the set of 85 starting structures consists of 25 α-helical proteins, 19 β-sheet; 27 are mixed α/β, and 14 belong to the "other" classification. The largest structure (1AKR[145]) is an α/β protein with 147 residues, while the smallest one (1G7A[146]) has 21 residues and is classified as a small protein in SCOP. To insure that the MD simulations not only sampled the near native conformations but also sample as many as rotamer configurations, an averaged CαRMSD of 4 Å from their starting structures was used as cutoff. This cutoff reduces artifacts from non-native conformations but ensures plenty of conformation sampling. The dynameome drifts on average 2.6 Å CαRMSD from the native structure with a standard deviation 0.5 Å per fold. Such a small deviation demonstrates that our dynameome dataset only samples conformational space close to the native conformation. Table 6 summarizes some simple properties for each protein fold. With each simulation

Table 5: SCOP classification of simulated folds

SCOP Class	Number of folds	Smallest[a]	Largest[a]	Member
alpha	25	31	131	1aie,2erl,2cpgA,1utg,1i27A,1dp7P,1g8qA,1cy5A,1fk5A,1lriA,1psrA,3caoA,1jr8A,256bA,1bkrA,1i8oA,1dlwA,1elwA,2a0b,2mhr,1fazA,1ijyA,1e85A,1c52,1kr7A
beta	19	64	135	1c8cA,1c9oA,1gutA,1c4qA,1gvp,1g9oA,1c5eA,3vub,3chbD,1qauA,2mcm,1f86A,1flmA,1whi,2cuaA,1bfg,1jb3A,1rie,1c1lA
alpha/beta	6	87	147	1aba,1thx,1jf8A,1ccwA,1i5gA,1akr
alpha+beta	21	61	138	2igd,1cseI,1b3aA,1cc8A,1vcc,1euvB,1fm0D,1iqzA,1opd,1cyo,1rgeA,1t1dA,4ubpA,1lkkA,1ew4A,1bkf,1kafA,1kpf,1qtoA,1c7kA,1gmuA
other[b]	14	21	83	1g7aA,1sgpI,1isuA,1nxb,1f94A,1vfyA,1i71A,1jekB,1jekA,1svfA,1et1A,1ppt,1wfbA,1g6uA

[a] Number of residues.

[b] SCOP classification of "other" contains classification of small proteins: 1g7aA, 1sgpI, 1isuA, 1nxb, 1f94A, 1vfyA and 1i71A; coiled coil proteins: 1jekB, 1jekA and 1svfA; peptides: 1et1A, 1ppt and 1wfbA; designed proteins:1g6uA

Table 6: Summary of MD simulation

PDB	Num of	Num of [a]	Box Size [b]	RMSD (Å) [c]	
ID	Residue	Water	(Å^3)*1000	Mean	STD
1aba	87	3437	115	1.8	0.4
1aie	31	2256	72	2.4	0.7
1akr	147	4017	139	2.3	0.5
1b3aA	67	3286	108	3.8	1.5
1bfg	126	3848	133	2.0	0.3
1bkf	107	3744	127	2.0	0.3
1bkrA	108	3561	122	2.4	0.5
1c1lA	135	3919	136	2.7	0.3
1c4qA	69	2284	78	3.4	0.8
1c52	131	3883	134	2.5	0.3
1c5eA	95	3323	112	2.0	0.4
1c7kA	131	3857	133	2.8	0.5
1c8cA	64	2823	93	2.5	0.6
1c9oA	66	2717	90	2.4	0.5
1cc8A	72	2374	81	2.4	0.5
1ccwA	137	4048	139	2.5	0.3
1cseI	63	2389	81	1.3	0.4
1cy5A	92	2937	101	2.4	0.7
1cyo	88	3576	121	2.4	0.4
1dlwA	116	3489	119	2.8	0.9
1dp7P	76	3772	126	1.5	0.2
1e85A	124	3791	130	1.9	0.2
1elwA	117	3493	127	1.7	0.3
1et1A	34	1901	62	2.1	0.6
1euvB	79	3347	111	2.9	0.5
1ew4A	106	3507	120	2.2	0.4
1f86A	115	3562	122	2.3	0.4
1f94A	63	2533	85	2.8	0.6
1fazA	122	4008	137	2.5	0.4
1fk5A	93	2823	96	2.8	0.5
1flmA	122	4116	139	2.8	0.6
1fm0D	81	2706	92	2.9	0.8
1g6uA	47	2534	89	1.6	0.3

Table 6: Continued

PDB	Num of	Num of [a]	Box Size [b]	RMSD (Å) [c]	
ID	Residue	Water	(Å^3)*1000	Mean	STD
1g7aA	21	1211	39	3.1	1.3
1g8qA	90	3879	128	2.3	0.4
1g9oA	91	3943	134	3.7	0.6
1gmuA	138	4295	147	3.5	0.7
1gutA	67	3166	107	3.9	0.6
1gvp	87	4102	148	3.5	0.5
1i27A	73	2956	99	2.8	1.3
1i5gA	144	4174	145	2.2	0.3
1i71A	83	3497	117	2.8	0.7
1i8oA	113	3554	121	2.8	0.4
1ijyA	122	3990	137	2.5	0.5
1iqzA	81	2740	93	2.6	0.6
1isuA	62	2314	78	3.0	0.9
1jb3A	127	4284	146	3.5	0.5
1jekA	40	2420	88	2.2	0.6
1jekB	34	2188	71	1.8	0.7
1jf8A	130	3984	137	2.2	0.3
1jr8A	105	3663	125	2.2	0.3
1kafA	108	3656	125	2.4	0.5
1kpf	111	4053	139	3.2	0.9
1kr7A	110	3350	115	2.0	0.3
1lkkA	105	3864	130	2.2	0.4
1lriA	98	3190	108	2.3	0.5
1nxb	62	2449	82	3.3	0.6
1opd	85	2829	96	2.0	0.3
1ppt	36	2035	66	3.1	0.8
1psrA	100	4173	142	2.8	0.7
1qauA	112	4352	145	2.2	0.7
1qtoA	122	4590	167	3.8	0.8
1rgeA	96	3226	110	2.3	0.3
1rie	127	3997	137	2.4	0.5
1sgpI	51	2116	70	3.3	0.7
1svfA	64	4225	163	3.0	1.0

Table 6: Continued

PDB ID	Num of Residue	Num of Water[a]	Box Size[b] (Å^3)*1000	RMSD (Å)[c] Mean	STD
1t1dA	100	3636	124	2.4	0.3
1thx	108	3559	121	2.3	0.3
1utg	70	3536	116	2.6	0.8
1vcc	77	3150	105	2.2	0.3
1vfyA	67	2677	90	3.7	0.6
1wfbA	37	2126	68	1.5	0.5
1whi	122	4127	140	1.8	0.3
256bA	106	3560	121	2.0	0.3
2a0b	118	3859	132	2.4	0.5
2cpgA	43	2490	81	3.6	0.8
2cuaA	122	3855	132	2.7	0.5
2erl	40	1900	62	2.5	0.5
2igd	61	2598	86	3.5	0.9
2mcm	112	3234	110	3.0	0.4
2mhr	118	3943	135	2.8	0.9
3caoA	102	4297	143	3.8	0.5
3chbD	103	3851	130	2.8	0.4
3vub	101	3908	131	1.7	0.2
4ubpA	100	3671	131	2.3	0.4
AVG[d]	**92**	**3345**	**114**	**2.6**	**0.5**

[a] The number of water molecules surrounding protein molecules in the simulation box.

[b] The size of simulation box.

[c] The averaged CαRMSD value for all the simulations when comparing to the native structure.

[d] The average of all the simulations in the dynameome dataset

resulting in 9,000 structures (see Methods),our dynameome dataset contains near 4 million structures for analyzing the ensemble-averaged properties of the native state. Our analysis focused on the relationship among the backbone conformation, residue packing, and side-chain conformation**.**

Residue volumes of the 20 amino acids

In Table 7, the average residue volumes were measured over the dynameome dataset for each of the 20 amino acids and compared to the residue volumes calculated from the ProtOr, a standard set of protein atom volumes representing well-packed residues.[147] As expected, the average residue volumes calculated from the dynameome are larger than the ProtOr set on average by about 3%. It has been shown that residues are more regularly packed the deeper they are buried in the protein, which results in smaller volumes, as opposed to the heterogeneous packing at the protein/water interface, which results in larger volumes.[147,148] As contrasted in the middle columns of Table 7, buried residues exhibit smaller volumes by about 4% on average than their respective exposed residues. For example, the difference between buried and exposed GLY is 3 Å^3 or about 5% of its calculated volume. Comparing the volumes of buried residues calculated from the dynameome dataset to the volumes from the ProtOr, they are more similar to each other, but there are some notable differences between the two sets. The CYS, TRP, and MET residues are significantly larger, whereas the charged ASP, GLU, and LYS are smaller. The largest volume difference comes from the ProtOr's CYS volume, which is 19 Å^3 smaller than ours, corresponding to 15% of its average volume.

Table 7: Residue volumes for the 20 amino acids

	Average Volume (Å^3)					Secondary Structure[c]			
	Ave[a]	Std	ProtOr[b]	Exposed	Buried	E	H	C	T
GLY	66	6	65	66	63	64	65	66	66
ALA	92	8	90	93	91	91	92	92	92
VAL	143	12	139	145	141	141	145	144	143
LEU	173	13	164	174	171	170	174	173	173
ILE	171	13	164	172	169	168	172	171	171
PHE	203	16	192	204	200	198	206	202	202
TYR	205	14	197	206	199	200	208	205	205
TRP	242	16	228	243	240	237	244	242	242
MET	182	16	167	183	179	180	183	181	181
PRO	129	9	123	129	124	128	128	129	129
SER	95	6	95	95	91	94	95	96	95
THR	122	9	126	123	119	121	123	123	122
CYS	122	12	103	123	120	119	124	121	121
ASN	128	8	125	128	121	125	127	128	128
GLN	156	10	149	156	147	154	156	157	156
HIS	163	11	160	163	159	160	164	162	163
LYS	174	11	167	175	154	172	175	174	175
ARG	200	11	194	201	188	198	200	200	202
ASP	105	6	117	105	102	106	104	106	106
GLU	133	7	142	133	127	132	132	133	133

[a] Residue's volume averaged over all the structures in the dynameome dataset

[b] Residue's volume based on previous calculation of ProtOr dataset of buried residue.

[c] Residue's volume in different secondary structure conformation, E: strand, H: helix, C:coil and T:turn.

The primary factor for this difference is that the CYS residues used to define the ProtOr set were mostly disulfide bonded,[147] which significantly reduces a CYS residue's volume. The dynameome possessed mostly reduced CYS residues, which are expected to show larger volume values. The volumes for buried charged groups ASP, GLU, and LYS from dynameome dataset are 13%, 11%, and 8% smaller respectively, than those from the ProtOr set. Since these are all buried, we find that they are forming salt-bridges. Due to the strong electro-constriction in a salt-bridge,[149] the overall residue volumes for these are smaller. However, the remaining 14 residues deviate by an average of less than 3% from the ProtOr values. Therefore, volume values of buried residue calculated using the dynameome dataset are generally consistent with the ProtOr volumes calculated from crystal structure data. Furthermore, this result supports the idea that our dataset is a good approximation of the native conformation.

Table 7 also shows the averaged volume of 4 classifications of secondary structure. When comparing an individual residue's volume across the secondary structure, two interesting features are observed. First, there is no large difference between the volumes associated with different kinds of secondary structure. It is commonly assumed that residues in α-helices and β-strands pack well; in turns moderately well; and in coils more loosely. However, Table 7 shows a maximum residue volume variation within secondary structure of only about 10% (data not shown), and the average difference between secondary structures is only 1%. Such small volume differences suggest that packing is not optimized for helices and sheets over other secondary structures. The second feature is that these small differences show a different order to how well the residue can be packed into secondary structure. Even though the volume differences reported in Table 7 among all the secondary structure elements are small, from 1 to 2 Å^3, the large size of our dataset strongly supports that these minor differences are meaningful properties. There is a general trend that residues in strands exhibit the smallest volumes followed by coils/turns and then those attached to helices usually show the largest residue volumes. On average, comparing strand to helix, a residue in a strand occupies only 98% as much volume as the same residue in a helix. If we assume that smaller volume indicates denser packing, our results demonstrate that sheets pack best, followed by turns/coils and lastly by helices. This ordering is somewhat counterintuitive since the helical and coil backbones pack the tightest, whereas sheet and coil backbones less well. Yet, when including the full residue's side chain, it makes sense that regular sheet conformations allow tighter residue packing than helices do with side-chains extended radially from a helical cylinder.

Volume variation with backbone conformation (ϕ,ψ)

To show the dependence of side-chain volume on the backbone conformation in more detail, the residue volume dependence on backbone torsion angles ϕ,ψ is plotted in Figure 9. Residue volumes were "normalized" for comparisons by expressing them as the percent of the corresponding amino acid's mean volume or vol% (see Methods for details). Using the color scale with bluer indicates larger than average volumes (looser packing) and redder indicates smaller than average volumes (tighter packing). Figure 9a plots the vol% versus ϕ,ψ for 408 experimentally determined structures selected from the PISCES.[150] This distribution from the PDB data is somewhat irregular even with the

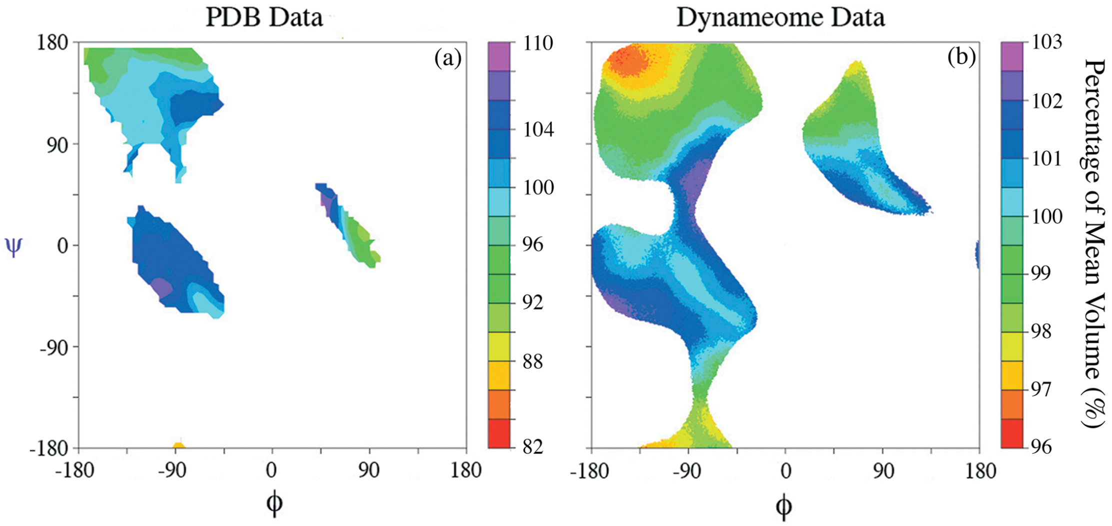


Figure 9: Residue volume versus backbone conformation.
The percentage of the mean volume (vol%) is shown as contour plots in backbone torsion angle spaces (ϕ,ψ). A color scale is used, where blue indicates larger volumes and red indicates smaller volumes. a) Analysis over a set of 408 PDB structures. Using a 5º resolution for ϕ,ψ values, vol% ranges from 82% to 110%. b) Results from our dynameome dataset with 1º resolution. The vol% ranges from 96% to 103%. Comparing the two plots, both a) and b) share similar patterns. The extra area in left-handed helical region in Figure 9b when comparing to the PDB plots in Figure 9a is due to the increased conformational sampling of exposed residues.

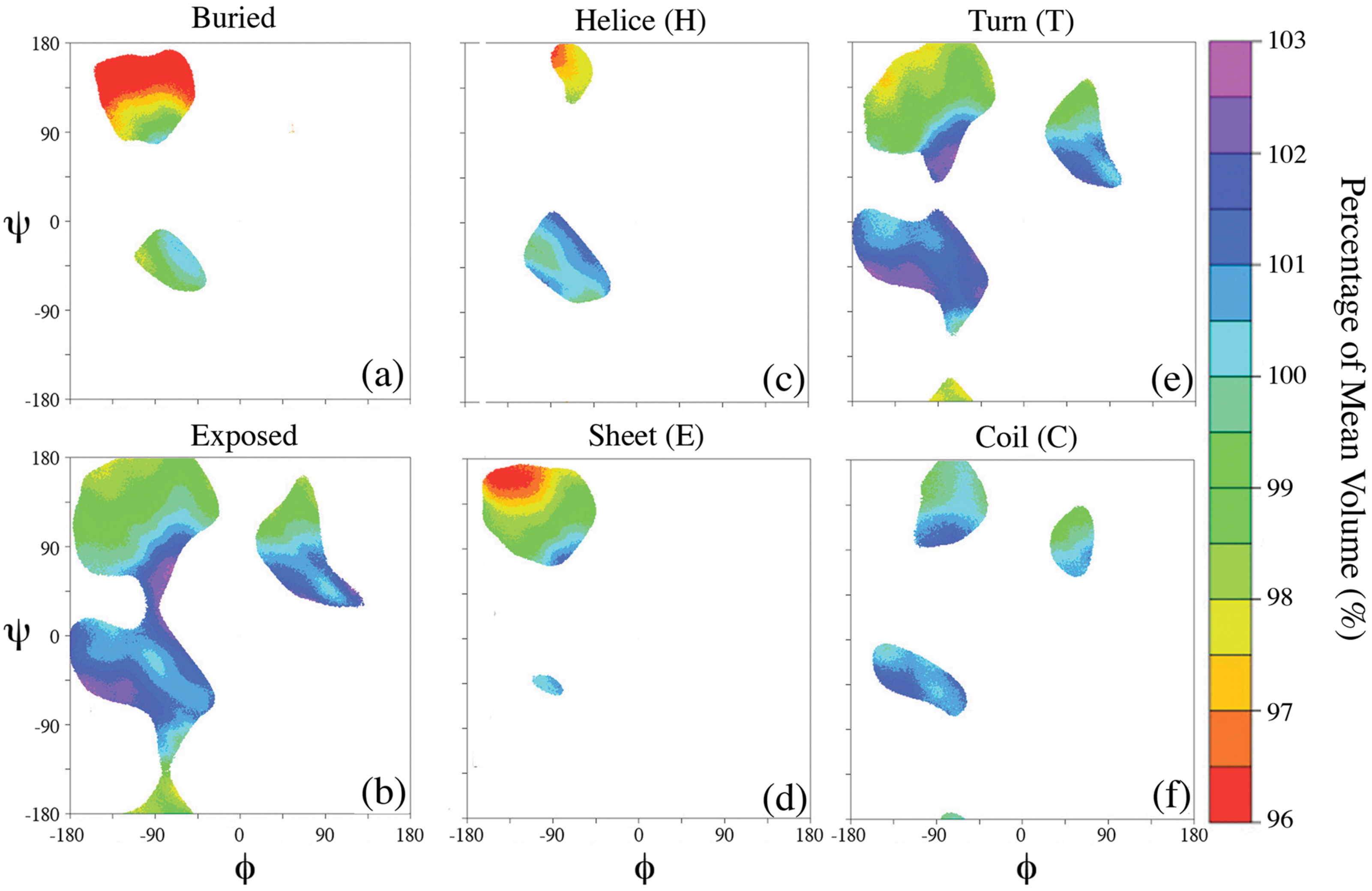
interpolation performed by the R statistical package.[120] This is especially true in the right-handed helical region that shows non-uniform patches of residue volumes. Even so, the distribution shows certain areas of the Ramachandran plot pack differently than others, where smaller residue volumes are seen ϕ,ψ values of -180°, 180° and larger volumes can also be seen in the right handed helical region (discussed in more detail below). Our dynameome dataset (Figure 9b) was able to reproduce this distribution exhibited by the experimental PDB data. The dynameome results are also consistent with other Ramachandran analysis[151-154] by showing no values in strongly disallowed regions. For example, the blank region around $\phi = 0$ in Figure 9b represents steric clashes between Oi-1...Ni+1, Oi-1...O and Oi-1...C. An increase population in the α_L conformation was also observed as expected from other studies using MD simulation to sample the conformational space, corresponding to the residues in the exposed portions of the protein that lacking regular secondary structure[155,156] and these residues are believed to be critical for β-structure.[157] The range of residue volumes from the dynameome is not as broad as those from the PDB. Since the PDB structures don't include all water molecules that surrounding the protein, we have to put protein molecule into fixed water box (see Methods) to calculate the volume, such approximation could cause cavities at the water-protein interface, produces larger volumes for the surface residues, thus increase the variation of volume values in the PDB dataset (Figure 9a). Basically, the dynameome data plotted in Figure 9b can be seen as representing an energy landscape hosting a collection of conformations with free energy close and slightly higher than the native conformation, supporting the idea that our dynameome dataset is an approximation of the near native ensembles. Comparing to the sparsely sampled conformational space using the experimental structures from the PDB, the dynameome's broad sampling of protein structures is seen in recent MD simulation of studying backbone conformation propensities[156,158] and it produces a far smoother distribution. The plot of this data in Figure 9b clearly depicts the smooth dependency of residue volumes on backbone conformation. The connected region between right handed helical and sheet region can also be seen as well as the expected increment of population in region where the α_L conformation is preferred. This ability to sample over many possible conformations in the native ensemble allows us greater detail in the characteristics of native structure.

Figure 9b and Table 7 both show that packing volumes don't vary greatly. In Figure 9b, the range of variation is about 7%, from 96% to 103% of the mean volume of each residue; in Table 7, it is about 1% for different secondary structure. The reason that the volume differences associated with different types of secondary structure in Table 7 are smaller than the volume differences seen as a function of ϕ and ψ is that the values in Table 7 are averaged over large regions of ϕ,ψ space. However, when considering a dense packing environment like the native conformation, even this limited amount of variation, which ranges from about 5 to 25 Å^3 in volume or 1 to 2 Å in radius, is significant. It is about the same size as the static variability seen among structures with clear sequence homology, and corresponds roughly to the volume of one atom. In another word, even a small backbone or side-chain conformational changes in one residue can cause significant changes in its packing environment. Meanwhile, the average standard deviation is 7% of the mean volume and shows no dependency on the backbone torsion angles (data not shown), indicating that the flexibility of side-chain or the flexibility of the protein packing is not determine by the backbone conformation.

In keeping with Table 7, Figure 10a and 10b show the differences in residue volumes and therefore packing between buried and exposed residues, respectively. All volumes in Figure 10a below the mean volume of the residue (vol% < 100), which indicates that buried residues generally occupy smaller volumes and pack tighter than exposed residues in Figure 10b. Furthermore, buried residues in regions where β-sheet is preferred pack even tighter than Figure 9b indicates. The vol% values of up to 4% smaller or 92% were found consistently in this region for buried residues (data not shown). In order to keep a consistent scale without losing details, data point with volume value lower than 96% in Figure 10a were rounded as 96% in Figure 10. By optimizing the main-chain hydrogen bonds, this region of ϕ,ψ space has the potential to promote the tightest packing of side-chains, and even more so for buried residues. Figure 10a and 10b also reveal dramatic differences in the backbone conformational freedom between buried and exposed residues. For buried residues, Figure 10a shows very limited sampling of ϕ,ψ space, populating only the regions of well-defined secondary structure near the center of the sheet region and right handed helical regions. Therefore, residues on the protein interior are conformationally restricted. On the other hand, exposed residues in Figure 10b exhibit the same range of sampling as seen in the dynameome dataset in Figure 9b. Figure 10a and 10b shows the different influences that non-loc

Figure 10: Residues’ volume versus backbone conformation in different environment.

The percentage of the mean volume in different packing environment is shown as contour plots in the backbone torsion angle space (ϕ,ψ). A color scale is used, where blue indicates larger volumes and red indicates smaller volumes. a) Buried residues from the dynameome. Buried residues only occupy limited ϕ, ψ space. b) Exposed residues from the dynameome. c) Residues classified as in α-helix (H) conformation. d) Residues classified as in β-sheet (E) conformation. e) Residues classified as in Turn (T) conformation. f) Residues classified as in Coil (C) conformation.


Buried
Helice (H)
Turn (T)
Exposed
Sheet (E)
Coil (C)
(a)
(b)
(c)
(d)
(e)
(f)
ψ
φ
180
90
0
-90
-180
Percentage of Mean Volume (%)
103
102
101
100
99
98
97
96

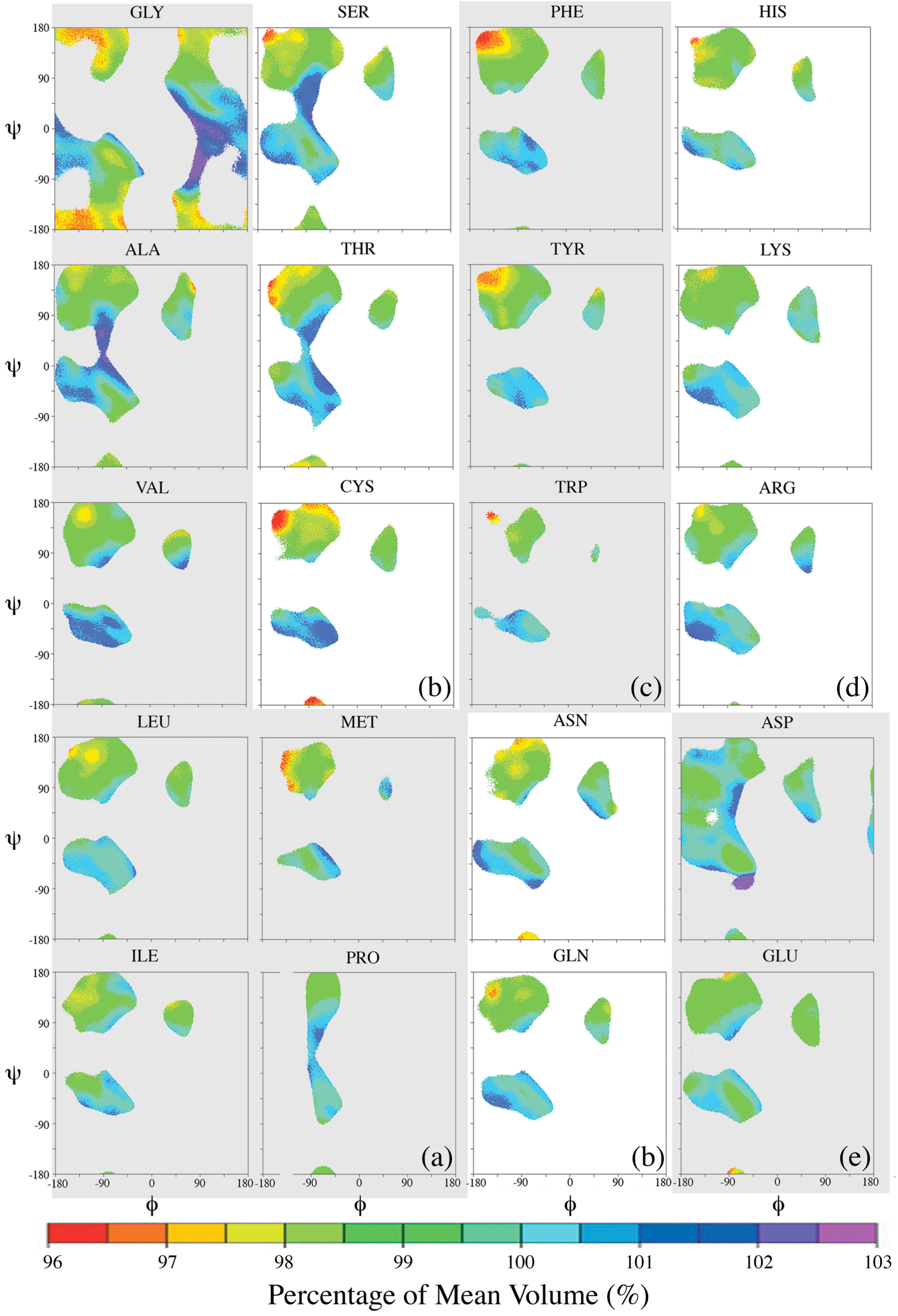
environment has on protein volumes and packing. In addition to the tighter interior packing, the very restricted conformational space sampled by buried residues in this study suggests that theoretical studies of protein folding (as well as structural verification procedures) which use crystal structures from the PDB could benefit from integrating information of solution-like structures.

As mentioned earlier, the most striking feature of Figure 9 is that they show a clear dependence of the residue volumes on the backbone conformation. In other words, different ϕ,ψ regions foster different packing environments. To further investigate this relationship, we split up Figure 9b into the four classes of the secondary structure as seen in Figure 10c to 10f. Because PROMOTIF defines secondary structure primarily on hydrogen bond patterns instead of torsion angles,[159] secondary structure classes aren't necessarily restricted to the backbone torsion angle space as the Ramachandran plot normally suggests, we find residues classified as α helices have backbone conformation in the region where β-sheet is preferred (Figure 10c), and vice versa (Figure 10d). Regardless of these inconsistencies, Figure 10c to 10f show that the pattern of residue volumes over ϕ,ψ is consistent across and therefore independent of different secondary structure classifications. Therefore, the backbone conformation only can be used to discuss the plots in terms of the dependence. Overall, Figure 10 confirms that residues pack more loosely in regions where right-handed helical is preferred than they do in regions where β-sheet is preferred. The helical region shows a saddle-like pattern, where residues pack more loosely toward the saddle's edges, in the H, C, & T classes of secondary structure where it is populated (Figure 10c, 10d, and 10f, respectively). For an α-helix, this less dense packing corresponding to the conformational requirement for the sidechain to point radically away from the cylinder formed by the backbone. In contrast, the sheet region in the upper-left corner defined by $-180 < \phi < -125^{\circ}$ and $125^{\circ} < \psi < 180^{\circ}$ exhibits tighter packing. Structurally, this region corresponds to an alignment of the CO…HN dipole-dipole interaction between two strands,[154] indicating that strand to strand main-chain hydrogen bonds promote/permit tighter packing.. In addition, this region is dominated by the *gauche*$^{+}$ conformation of the side-chain χ_1 rotamer, which has a relatively small packing environment (discussed later). Interestingly, these figures all show some dependency on ψ values. A "belt" shape ($-100^{\circ} < \psi < 80^{\circ}$), including regions where the left handed helical conformation is preferred, prefer an overall looser packing of residue, while outside the "belt", tighter packing that extends into sheet region is observed.

Figure 11 shows the variation of vol% with respect to ϕ,ψ for individual amino acids. We will discuss them in terms of their distribution of vol% and population. In general, the vol% patterns show that the residue occupies more volume when its backbone conformation falls into the right-handed helical region than it does in the β-sheet region. There are some interesting consistencies among residues very similar to what we have seen in Figure 9: a shape of saddle is usually observable in the region where right-handed helical is preferred and the packing is less dense towards the edges. CYS and VAL are exceptions that they pack consistently less densely through the middle of this helical region. In region where β-sheet is preferred, all amino acids pack a

Figure 11: Residues' volume of 20 amino acids versus the backbone conformation.

The percentage of the mean volume (vol%) is shown as contour plots in the backbone torsion angle spaces (ϕ,ψ) for each amino acid. A color scale is used, where blue indicates larger volumes and red indicates smaller volumes. The scale is kept consistent across all plots for easy comparison. The order of residues is the same as it is in Table 7, starting from GLY and ending up with GLU from top to bottom and left to right. Gray background was used to distinguish different physical properties of amino acid's R group. From left to right, top to bottom, area a) consists amino acids with nonpolar, aliphatic R groups, area b) consist amino acids with polar but uncharged R groups, area c) are amino acids with aromatic R groups and area d) and e) are amino acid with charged R groups, in which d) are positively charged and e) are negatively charged.


GLY
SER
PHE
HIS
ALA
THR
TYR
LYS
VAL
CYS
TRP
ARG
LEU
MET
ASN
ASP
ILE
PRO
GLN
GLU
(a)
(b)
(c)
(d)
(b)
(e)
180
90
0
-90
-180
ψ
-180
-90
0
90
180
ϕ
96
97
98
99
100
101
102
103
Percentage of Mean Volume (%)

little more densely than average. The amino acids HIS, GLY, MET, PHE, SER, THR, TRP, TYR, CYS all show smaller volumes or pack more tightly packed in the ϕ,ψ region toward -180°, 180°. Residues LEU, VAL, and GLN exhibit an island of tight packing (though the volumes are not the smallest vol% when comparing to other residues) in a region centered around $\phi = -125°$ and $\psi = 160°$. ASP shows an interesting spur of larger than average volumes in region where $-45 < \phi < -90$ and $-45 < \psi <-90$. Closer inspection of these conformations reveals that these larger than average ASP volumes belong to residues in the turn conformation, which also have contacts with water molecules and these very small populations are considered marginal significance with the cutoff values we used to eliminate extreme observations. For all residues, bridging areas between right-handed helical and sheet regions are packed less densely.

For sampling of Ramachandran space shown in Figure 11, the 20 residues exhibit the expected distributions, where GLY samples the most conformational space and PRO samples the most restricted one. Surprisingly, GLY does not populated extensively in regions where β-sheet is preferred, probably due to the lack of side chain interactions. If we assume that the more ϕ,ψ space a residue can populate, the easier it can replace other residues or be replaced, the clear difference among the amino acids in their populated regions may be clues as to which amino acids are least responsible for maintaining the folded state of a protein: namely, GLY, ALA, SER, THR, and ASP. In contrast, TRP and MET show quite restricted conformational possibilities (as does of course, PRO). HIS, CYS, PHE, and TYR also have relatively limited backbone conformational freedom. These 7 conformation-restricting amino acids represent 20% of the residues in our dynameome data set, while the 5 least-restricting amino acids represent 34% of the dynameome data set (data not show). In the BLOSUM62 matrix,[160] which represents how well amino acids are conserved during evolution, as well as the likelihood that each will substitute for another, the most conserved amino acids, in order of conservation, are TRP, CYS, HIS, TYR, PRO (BLOSUM62 diagonal elements of 11,9,8,7,7). The least-conserved are ALA, SER, THR, VAL, LEU, ILE (BLOSUM62 diagonal elements all equal 4). Thus, with the "full-range" sampling over the conformational space for each residue provided by our dynameome dataset, further analysis can be done to understand whether the more conformational restrictive amino acids are responsible for determining the fold. These results match well to an in-depth statistical analysis of Ramachandran distributions of the 20 amino acids.[161]

Backbone dependency of side-chain conformation

Figure 12 shows the population and value distribution of the first side-chain torsion angle χ_1 plotted against backbone torsion angles. Similar studies have been done using several rotamer libraries.[125-127,129,162-164] As mentioned before, due to the limited sampling of the PDB data, such libraries are usually studied by clustering the observed conformations or by dividing the torsion angle space into bins and determining the average conformations in each bin. Rare side-chain conformations that sparsely populate the Ramachandran space are underestimated even with a continuous statistical approximation. As shown above, our dynameome dataset exhibits a continuous sampling

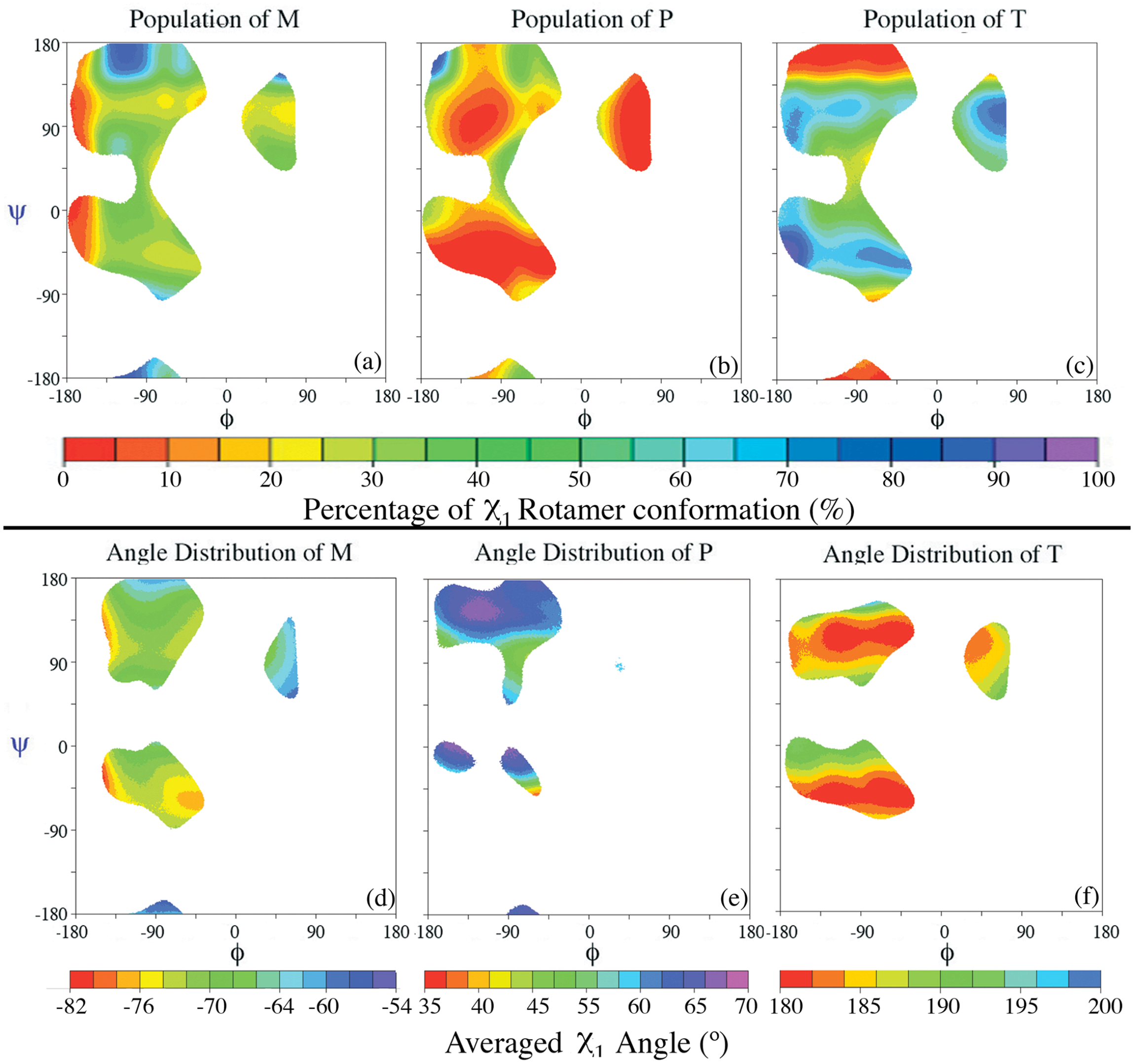


Figure 12: Population and angle distribution of 3 χ_1 rotamer versus backbone conformation.

The population and angle distribution of 3 χ_1 rotamer are plotted respectively for different backbone conformation. PRO, ALA and GLY are excluded from plots. a), b) and c): χ_1 rotamer populations in percentages are plotted against backbone ϕ, ψ torsion angles. A color scale is used from blue (higher occupancy) to red (lower occupancy). a) Population of rotamer conformation *gauche*$^-$: M. b) Population of rotamer conformation *gauche*$^+$: P. c) Population of rotamer *trans*: T. At any given ϕ, ψ angle, the percentage of the population in each of the 3 rotamer conformations sums to 100%. d), e) and f): average χ_1 rotamer angles are plotted against backbone ϕ,ψ torsion angles. d) Angle distribution for rotamer M. e) Angle distribution for rotamer P. f) Angle distribution for rotamer T. A color scale is used from red (smaller angles) to blue (larger angles) for each rotamer respectively. Rotamer M has angles range from -82° to -54°, rotamer P has angles range from 35° to 70° and rotamer T has angles range from 180° to 200° (-160° to -180° before conversion).

of the native conformational space that allows us to highlight some unique features of the native state that are less clear when data size is limited.

The plots in Figure 12 are split based on the 3 rotamer conformations: *gauche$^-$,* *gauche$^+$* and *trans*, which we designate with M, P and T, respectively. To begin with, we discuss the first rotamer angle χ_1 in broad terms, namely, population. The three panels of Figure 12a, 12b and 12c show the population of side-chains found in each of these three χ1 rotamers (M, P, and T, respectively) as a function of ϕ and ψ (except residue PRO, ALA and GLY). At any given ϕ,ψ angle, the population percentages from each of the three rotamers sums to 100%. Figure 12a shows that the M rotamer is highly populated in the β-sheet region where $-135^\circ < \phi < -90^\circ$ and $\psi > 135^\circ$ and the fringe of the two helical regions. In Figure 12b, the P rotamer only populates limited regions due to its nudged conformation and mostly where both T and M rotamers are not favored ($-180^\circ < \phi < -150^\circ$, $150^\circ < \psi < 180^\circ$). Interestingly, this region is where the packing is tightest (Figure 9b). In contrast, T is scarce in the region where $\psi > 150^\circ$ but becomes the preferred rotamer in the sheet region where $\phi < -135^\circ$ (far left hand side), where the M rotamer rarely populates. In the remaining portions of the sheet region, Figure 12a and 12c show that T and M both populate equally with T being slightly favored at where $90^\circ < \psi < 135^\circ$ and around $\psi = -45^\circ$. Consistent with other analysis of the rotamer dependence on backbone conformation, rotamer P is only favored by SER (data not shown) due to the hydrogen bond interaction with C=O of residue i-1.[127] Also, the high population of M seen in Figure 12a where $180<\phi<-150$ and $150<\psi<180$ was previously observed as the results of rotamer distributions from the experimental data.[126]

In Figure 12d, 12e and 12f, plots of the first side-chain torsion angle χ_1 reveals that its preferred value depends strongly on the backbone conformation. This has also been found previously, but defined more coarsely due to the sparse sampling of the PDB. The advantage of our dynameome data is the continuous sampling of the conformational space, which produces smooth transition in each χ_1 rotamer class. Figure 12d clearly shows that the M rotamer is dependent on ϕ in most of its region except the region with extreme ψ values ($\psi > 150^o$) where left handed helical is preferred. Its optimal value of -60° is shown around area where the M population is the highest (Figure 12a). However, the most preferred value of χ_1 for the M rotamer is -70°, lower than its optimal value (see below for discussion). The value of χ_1 angle for the M rotamer ranges from -55° to -85°. The P rotamer in Figure 12e is dependent on both ϕ,ψ with larger χ_1 values towards 70° centered approximately at $\phi = -125°$ and $\psi = 145°$ or $\phi = -120°$ and $\psi = 0°$. Again, the P rotamer has it optimal χ_1 value of 60° in its most populated area at $\phi = -180°$ and $\psi = 180°$ (Figure 12b). However, the χ_1 value of the P rotamer ranges from 40° to 70°. The T rotamer (Figure 12f) shows a dependency on ψ in regions where β-sheet and right-handed helical are preferred and such dependency is stronger that what is seen with the M rotamer. The T rotamer also exhibits weak dependency on ϕ in the region where left-handed helix is preferred. Optimal χ_1 values for the T rotamer occur in bands where $-90° < \psi < -40°$ or $90° < \psi < -135°$. The χ_1 distribution for the T rotamer is also skewed and ranges from 175° to 205°.

Packing in different χ_1 rotamer conformation

For completeness, we plotted the percentage of the mean volume (%) against the rotamer χ_1 angles in Figure 13. The plot does not include data from the amino acid PRO, since the residue also restricts the χ_1 value in the P rotamer around 0º. In general, there is no dependency observable between the percentage of the mean volume and χ_1 angles, which suggests that these two values are independent of each other. Consistent with Figure 12, the M, P, and T rotamers show their most populated χ_1 angles of -70º, 65° and 180º, respectively at vol% value of 100. As χ_1 moves away from its mean in the 3 rotamers, residue volumes still peak around their mean volume, but with a drop off in the population. Also, we see that the M and T rotamer distributions are connected and that the P rotamer is isolated. This is expected as well since the P rotamer is bounded on both sides by the N-Cα and the Cα-C bonds, whereas the connection between M and T is not impeded by the hydrogen attached to the Cα atom. The primary difference between three χ_1 rotamers is their range of volumes. With vol% extending up to 132%, the T rotamer has relatively more volumes to sample than either M or P does. P samples the least amount of residue volumes, which is consistent with the fact that P is only favored in limited backbone conformations (Figure 12b). These distributions suggest different flexibility of different χ_1 rotamers. For the T rotamer, especially in regions where right-handed helix and β-sheet are preferred, the ϕ value stays negative, which makes the backbone bend away from the side-chain and allows side-chain of residues in the T conformation be able to occupy more space, about 10% of their volume. For each given residue, different volume can be treated as different conformations. Thus, residues with

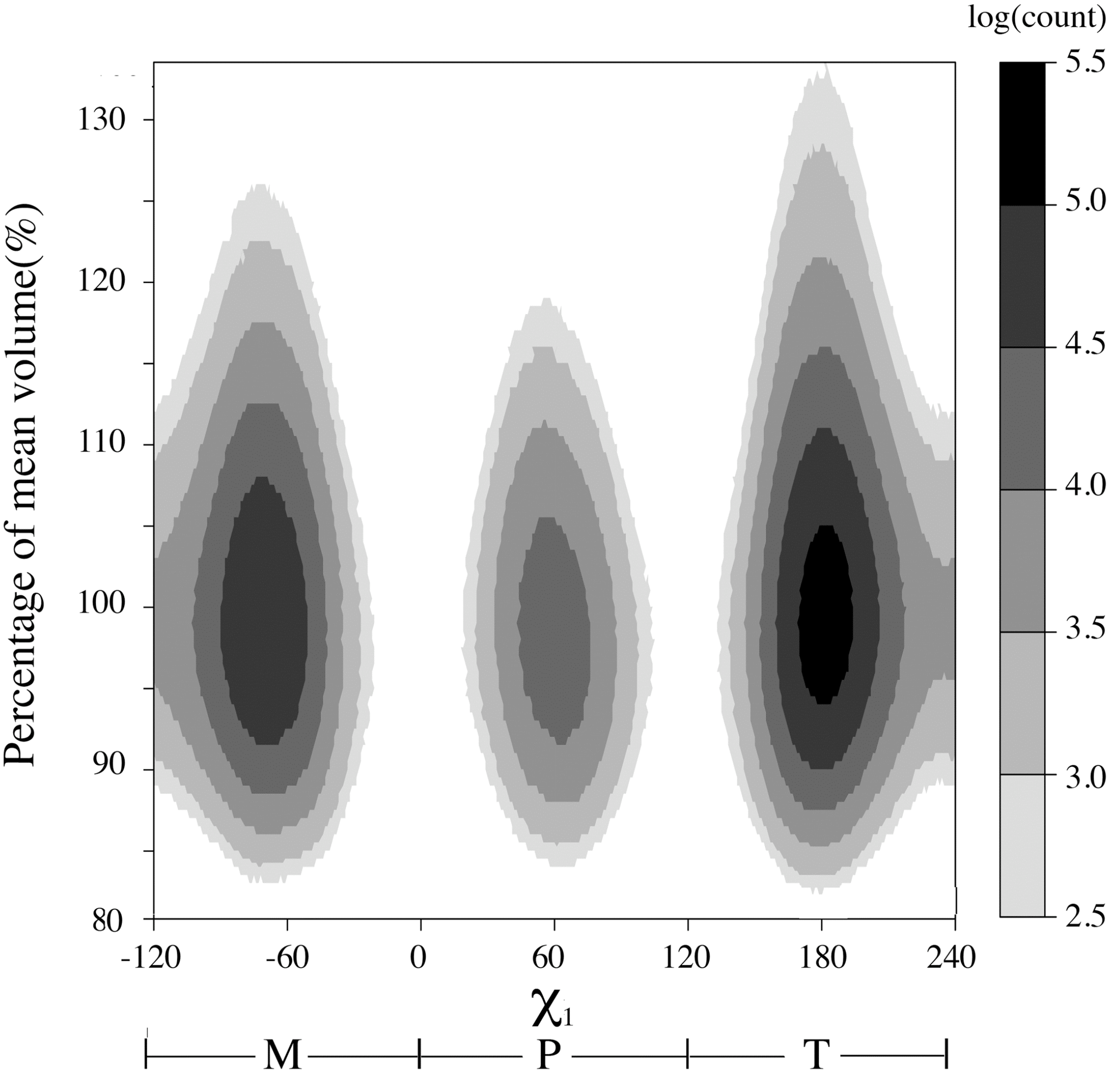


Figure 13: Residues' volume distribution versus χ_1 rotamer.
Distribution of vol% (percentage of mean volume) for all residues except GLY, ALA, and PRO was plotted against their χ_1 value. The counts are shown on a log scale and a total count cutoff of 500 was used to eliminate extreme values. The darker the color is, the more the observations, and the lighter the color is, the fewer. Three peaks were observed around percentage of mean volume of 100% at χ_1 value of -70º, 65º, and 180º for rotamer M, P and T, respectively.

their sidechain in the conformation of the T rotamer tends to be more flexible than the same residue with their sidechain not in the conformation of the T rotamer. Figure 13 shows that all rotamer conformations can sample all of the packing environments (volume) available no matter what its sidechain conformation is. This suggests that packing is not determined by different χ_1 rotamers or in another words, excluded volume is not sufficient enough to define the explicit rotamer conformations.[165]

Rationalization of the interdependence between χ_1 and ϕ,ψ

The relationship between the χ_1 angle and the backbone conformation can be explained in detail using physically based steric interactions as diagrammed in Figure 14. Similar explanations have been made using butane and syn-pentane interactions as well as similar Newman projections.[125-128] Here, because our dynameome provides a more continuous sampling of the protein conformational space, our description of the Newman projection samples at the single degree resolution over all the allowable χ_1 angles. Furthermore, the dynameome also allows us to visualize the interdependence of the backbone and χ_1 angle by directly modeling the steric repulsions between a residue's with its main-chain N-Cα or Cα-O bonds and atoms over various conformations. For simplicity, disallowed conformations caused by the backbone clashes are not discussed in detail and the influence of backbone conformation on side-chain conformation is only discussed for observed backbone conformations. In the following section, we base our discussion on clashes from the Ramachandran map as well as syn-pentane interaction between Cγ and corresponding backbone atoms. Also, to simplify the discussion, we

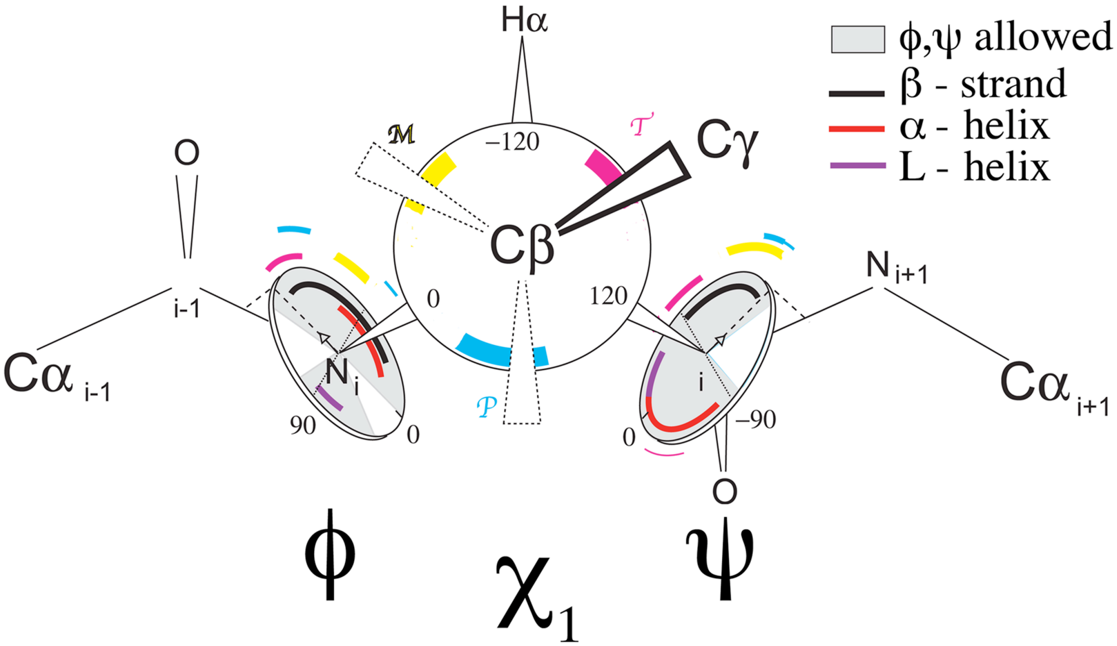


Figure 14: Schematic view of backbone influence on side-chain conformation. A portion of a peptide chain (from $C\alpha_{i-1}$ to $C\alpha_{i+1}$) is shown in extended conformation. Because of the extended conformation, the main chain atoms all lie in a single plane. However, the planarity of the diagram has been "polluted" by the addition of a side-chain Cβ atom to the central Cα. The single plane defined by the extended chain has been tilted toward the viewer by 55 degrees necessary to let the viewer look straight down the Cβ – Cα bond. The individual peptide planes can be rotated out of the plane of the extended conformation. φ plane is shown at the left side as well as ψ plane at the right side. The arrow and the dashed lines in each plane minimally indicate the plane. Generously allowed regions (as determined in this work) from the Ramachandran plots are indicated by the very light gray shading. Completely forbidden regions are white. Heavy black arc is used to indicate where β-sheet is preferred, heavy red arc is used for α-helix and short purple arc for left-handed helix. In the center "dial", the preferred regions for side-chain rotamers (χ_1) found in this work are shown. The "dial indicator" on the χ dial is the Cβ– Cγ bond, which is placed in "T" conformation (the magenta arc). "M" (yellow) and "P" (cyan) conformation positions are indicated with dashed-outline Newman-projection-style bonds.

do not use the i subscript for atoms on the reference residue, but do use it to refer to atoms preceding or adjacent to the reference residue. In Figure 14, the χ_1 angle of three rotamers is indicated by the Cγ position where the "dial indicator" on the χ_1 dial is the Cβ – Cγ bond, and the positions of "T" (crossing the magenta arc), "M" (yellow) and "P" (cyan) conformation are indicated with dashed-outline Newman-projection-style bonds.

Figure 14 shows a schematic view of the allowable φ, ψ, and χ_1 torsion angle values. The position of the χ_1 rotamer relative to the φ or ψ angles helps to explain its dependence. Rotamer T is closer to the "ψ side" and is influenced by the next residue, i+1. Rotamer M is closer to the "φ side" and is influenced by the previous residue, i-1. Rotamer P is between φ and ψ, thus it can be influenced by both residues i-1 and i+1. However, all 3 χ_1 rotamers show dependency on ψ. The ψ torsion angle involves the atoms N, Cα, C and N_{i+1} atoms and can occupy any angle from cis to trans conformations. This flexibility brings two heavy atoms N_{i+1} and O to pack against the Cβ and Cγ atoms. For the M and T rotamers, this pushes the Cγ atom towards the Hα atom, so that χ_1 of M rotamers become more negative and that of T rotamers becomes more positive as the ψ changes. For the P rotamer, these interactions move the Cγ towards the N atom and lower values of χ_1. The φ torsion angle involves the atoms C_{i-1}, N, C_α and C, but only C_{i-1} can form syn-pentane interactions that can affect the Cγ atom and the χ_1 angle. Furthermore, the φ angle is negative except in the region where left-handed helix is preferred, which means the atoms C_{i-1} and O_{i-1} are positioned mostly in a trans conformation relative to Cγ. As shown in Figures 12d and f, the φ takes on more of a cis interaction with Cγ in the left handed helical region and therefore, has stronger influence on χ_1 in this region.

Besides the above general effects, each rotamer has certain unique properties of their χ_1 dependence on the backbone conformation. Facing toward the N atom, the M rotamers is expected to depend on ϕ and is influence by the O_{i-1} clashes with the Cγ. As explained above, even the M rotamer facing away from the C atom shows some dependence on ψ. It can be explained as the influence of other rotamers, we find the M rotamer's dependence on ψ angle is due to the packing of the Cβ atom with the O atom as ψ changes. The major deviation is in the regions where right-handed helix is preferred around ϕ,ψ value of -45°,-45°, where χ_1 angle are highly skewed (Cγ closer to Hα). This can be attributed to the hydrogen bond made by H_N with the O_{i-4}. This hydrogen bond packs an O_{i-4} atom against the Cγ and forces the Cγ towards Hα for a more negative $\chi_1$value.

For rotamer P, due to the clashes of Cγ atoms "pinched" between N-Cα and Cα-O bonds, it rarely populates in the left-handed helical region. For the same reason, rotamer P shows clear dependency on both ϕ and ψ. The P rotamer's ϕ dependence is due to the packing of the Cγ with the N and O_{i-1} atoms. The ψ dependence of the P rotamer is due to the C and O atoms. At around $\psi = 0°$ or $\psi = 180°$, χ_1 values stay around 65°. As the ψ angle changes, atom N_{i+1} (when $\psi < 0°$) or atom O (when $\psi > 0°$), moves closer to the Cγ atom and causes the χ_1 angle to decrease. Also worth noticing, the lack of the P rotamer in the region where right-handed helix is preferred could be attributed to the requirement of forming hydrogen bonds in the α helix, which prevents the Cγ from occupying this rotameric conformation.

As expected, the rotamer T is ψ dependent in regions where right-handed helix and β-sheet are preferred. Rotamer T stays in its optimal conformation around ψ = 120º and ψ = -60º where atoms O and N_{i+1} both have the least interaction with residue's Cγ atom. From either point, when the Cα–C bond rotates and either atom N_{i+1} or atom O approaches atom Cγ, the Cγ is pushed towards the Hα, which increases the χ_1 value (χ_1 angle of -160º corresponds to 200º in Figure 12f). In the region where left-handed helix is preferred, rotamer T shows a slight ϕ dependency due to possible interaction with the O_{i-1} atom.

Conclusion

In this study, we took advantage of MD simulations to generate near 4 million structures that sample the native conformational space. In contrast to the sparse data provided by the PDB, we were able to sample from a continuous conformational space and to better characterize the dependency of the side-chain packing and conformation upon the backbone conformation. We were able to determine the contribution of the local environment (backbone conformation) and non-local environment (solvent exposure) on the volume of residues with implications about the side-chain packing. A comparison between buried and exposed residues shows that buried residues (protein core) prefer tight packing and are found only in a rather limited conformational space (Figure 10a). We also found that the packing is only slightly but noticeably different for different secondary structures, where strands promote tighter packing while α helices promote looser packing (Table 7 & Figure 9). In addition, the packing has a strong dependency on the backbone conformation regardless of different side-chain conformations (Figure 13). Because the dynameome dataset allows more fine-grained analysis, we were also able to more precisely define the relationship between the first side-chain rotamer χ_1 and the backbone conformation. First, all rotamers show dependence on the ψ torsion angle due to clashes of the O atom with the Cγ, while the influence of the ϕ torsion angle is less so due to weaker interactions of the C_{i-1} and O_{i-1} with the Cγ. Second, the variance of all 3 χ_1 rotamers from their canonical conformation are skewed to one side due to syn-pentane interactions. Third, "non-local" interactions, such as hydrogen bond from i-4 residues in α helices play important role in the side-chain conformation. These results help to define the exact role that the backbone conformation plays in the determination of protein folds. Although we have couched our discussion in terms of the dependence of side-chain characteristics on the backbone conformation, in fact it is a two-way street. While the backbone conformation sets the placement for the side chain, the packing of side chains determines the position of the backbone atoms.

Materials and Methods

Dataset

We ran 5 independent 10 ns MD simulations on each protein using the ENCAD program[112] and the F3C explicit water model.[113] The ENCAD program and the associate force-field provide a useful means to approach this problem, as it does not suffer from some of the problems that the CHARMM and Amber force-fields exhibited (and which have since been corrected. The ENCAD suite has been used successfully and recently in many applications including folding/unfolding studies and replica-exchange studies, In addition, some comparisons have been made between different force fields.

For each simulation, the coordinates of each structure were placed in a box of water and then energy-minimized. Each box of water was trimmed so that the edges were at least 8 Å away from the closest protein atom. All water molecules within 1.67 Å of the protein were removed, and the box sides were corrected to match the density of water (0.997 g/ml) at 298 K. Sodium or chloride ions were used to replace water molecules at random positions to yield an electrically neutral system. Conjugate gradient energy minimization was performed in the following order: The protein was fixed while the water molecules were minimized over 1,000 steps. The protein was then minimized in the next 1,000 steps, holding the water molecules fixed. Finally, the entire system was minimized over 1,000 steps. To begin each of the simulations from a unique starting point, the system was equilibrated to 298 K using a different random-seed number to assign initial velocities. During the calculations, the coordinates of the structure were updated at two femtosecond intervals and sampled every picosecond (500 steps), such that each 10 ns simulation generated 10,000 steps. All simulations are summarized in Table 6. The largest simulation has a water box of 60.5 Å by 55.1 Å by 50.0 Å in size and 4590 water molecules around a 122-residue protein 1QTO[166] while the smallest simulation has a water box of 40.5 Å by 28.0 Å by 34.4 Å in size and 1211 water molecules around a 21-residue protein 1G7A.[146]

Data analysis

Because the initial steps in the simulation equilibrate the system to 298 K, we decided to disregard the first nanosecond (ns) of the simulation and performed the analysis using only the last 9 ns of the simulation (1-10 ns). Programs written in C and PERL were created to analyze the native ensemble of structures. Coordinates were viewed using PyMol.[118] Figures were generated using the R statistical package.[120]

1. Secondary structure assignment

For each structure generated from the simulation, the secondary structure of the protein was defined using PROMOTIF[159] and categorized in the following manner. Residues without any assignment were assigned to the random coil (C) class. Both β-turns and G-turns were combined as turn (T). All the helices were classified as (H). Strand and β-bulges were combined as extended strand (E).

2. Volume calculations

The volume were calculated using the Voronoi Polyhedra method[167] for heavy atoms, which is explained in more detail in a previous study.[121] Only contacts with surface area larger than 1 Å^2 were considered. An exposed residue is defined as directly contacting the water molecules, while buried residues were those that only contacted protein. For each residue, the total volume is the sum of each atom's volume.

To compare volumes of different residues in different sizes in the dynameomic dataset, we normalized all 20 residues' volumes to a common scale: percentage of mean volume or vol%. First, mean volumes, <*vol*>, for each of the 20 amino acids were calculated over the whole dynameome dataset. As shown in equation (1), the vol% is derived by dividing the volume, *vol,* of a residue in a particular structure at a particular timestep by the respective residue's <*vol*>.

$$\text{vol\%} = \frac{\text{vol}}{<\text{vol}>} \qquad (1)$$

The residue volume plots in the Figure 9 and 10 show the average value of all the vol% values at a particular backbone conformation (ϕ,ψ) over certain sets of residues and/or conditions like secondary structure, buried or exposed residues. Only values with more than 1,000 observations at each backbone conformation (ϕ,ψ) were plotted.

For the PDB dataset, each structure was placed in the center of a water box to mimic the protein solution environment. This water box was taken from a MD simulation of pure water using the same parameters as in the protein simulations. Duplication of water box is applied if necessary to generate large enough box for protein. Any water atom within a distance of 1.8Å of protein atoms was removed. Volumes, torsion angles are calculated using the same method for simulated structures as described above. The same approach was used to plot the PDB data as it has been done on the dynameomic data (described above), mean volumes were calculated for each type of residue and the average percentage of mean volume was plotted accordingly. An observation cutoff of 250 was used and a different backbone bin size was used (see below).

3. Calculation of torsion angles

The ϕ, ψ and χ_1 values for each residue in every structure were calculated using PROMOTIF and values were rounded up to the next whole number. In effect, we used

1° bins for the dynameomic data and a 5° bins for the data from the PDB. The values for the first and last residue were omitted from the calculation. For $\chi_1$value analysis, ALA, GLY and PRO are excluded.

The χ_1 values (except those from PRO, ALA and GLY) were classified using similar nomenclature to a previous study.[129] As a simplification, M, P and T are used to refer the 3 χ_1 rotamers, respectively. M stands for *gauche*$^-$ where $-120^\circ < \chi_1 < 0^\circ$, P stands for *gauche*$^+$ conformation where $0^\circ < \chi_1 < 120^\circ$, and T stands for *trans* conformation where $120^\circ < \chi_1 < 240^\circ$. Since torsion angles are calculated from -180° to 180°, the χ_1 values in the T rotamer from -180° to -120° needed to be converted to their positive values by adding 360º to insure a continuous transition between -180º and 180º in various plot. For ILE and THR, since χ_1 is defined differently than other residues, the calculated χ_1 values were translated to reflect corresponding C_γ atoms in other residues by subtracting 120º. These translated values were then evaluated as M, P, or T as defined above.

The plots in Figure 12 were made using certain criteria. For the χ_1 rotamer population plots, a count cutoff of 1000 was used. At each backbone conformation, the values over the 3 χ_1 rotamers add up to 1 or 100%. For example, at the ϕ,ψ value of -60°, -40° in the α-helical region, M population is 28%, P is 2% and T is 70%, which adds up to 1. Since P populations are often small, a count cutoff of 500 was used for the P rotamer instead of using the 1000 used on T and M rotamer value.

The distribution of vol% versus χ_1 angle (Figure 13) required that the counts be on a log scale. The count cutoff was 500. While the bin size for χ_1 values was 1° (as explained above), the bin size for vol% is 0.5%.

CHAPTER IV

USING CONSERVED PACKING INFORMATION FOR TEMPLATE BASED STRUCTURE PREDICTION

Overview

Template based structure prediction is becoming more important in the structure prediction field because it is believed that a representative of every available protein fold can be obtained, and all structure predictions will eventually become template based. Currently, most template based structure prediction methods concentrate on finding the right backbone conformation of the target sequence, and refine it using various refinement procedures. For the past decade, template based structure prediction methods have always suffered from the same problem: compared to the template structure derived from close homologues, the refined structures are less accurate or further away from their native conformation. In this study, we test our template based structure prediction method using 52 prediction units from CASP7 experiments. Our packing orientated method predicts structure using spatial constraints derived from the conserved relative packing groups, which were obtained from available multiple template structures. By mimicking the experimentally determined NMR data, but with longer constraints reflecting conserved packing environments at the sequence level, we were able to provide “added value” to the starting structure. The long-range spatial constraints (>8Å) derived from the relative packing groups were important to improve the starting structure. We believe that our method provides a new angle on template based structure prediction.

Background

Template based structure prediction, in contrast to template free structure prediction, creates a prediction of an unknown structure using close structural homologue(s). It holds a great deal of promise due to the belief that a representative of every protein fold will eventually be solved.[17-23] Template based methods produce the most reliable and accurate predictions of protein structures aside from experimental determinations.[24,25] The availability of a representative fold as a starting template offers the quickest path to generating a model of the real structure. Because its capability of providing conformational information about proteins that lack experimental structures, template based structure predictions have been used successfully in a variety of applications, such as studying the effects of mutations, designing site-directed mutagenesis, predicting binding sites, and docking small molecules in structure-based drug discovery. For the past 14 years, template based structure prediction has evolved steadily during seven CASP experiments, and now is able to provide “added value” to the best template structures (starting structure) generated from the best homologue(s).

Template based structure prediction usually involves 4 different steps.[26-29] First, the parent structure(s) are identified using sequence searches against the known structure database (the Protein Data Bank[3]). Second, the initial template structure(s) (starting structure) are constructed by aligning the target sequence to the parent structure(s) with identification of conserved and variable regions. Third, the starting structure(s) are refined through a combination of backbone moves, side-chain packing, and loop modeling of highly variable regions and during this refinement, hundreds or thousands of model structure are generated. The last step is to evaluate these models and choose the one that is closest to the native conformation. In general, template based structure prediction tries to find the closest backbone conformation first, and then refine the backbone conformation towards the native conformation. Unfortunately, the imprecise variations between the close template(s) and the real structure produce the major source of challenges that have existed in this field. To be more specific, while template structure(s) possess backbone conformations that were well packed for the template sequence, these backbone conformations may represent an over-packed environment for side chains of the target sequence.[33] Thus, current refining protocols tend to move the starting structure (derived from templates) away from the native conformation.[168,169] In many cases, just submitting the starting structure without perturbations from the refining protocols provides the best model. On the other hand, as compared to the best available template structure (a known structure with a high sequence identity), even though the final prediction may be closer to the native conformation, it often lacks the resolution and accuracy necessary for practical applications, such as molecular replacement.[170]

One of the problems with template structure prediction is the mapping of sequence changes in the multiple sequence alignment to structural changes among homologues. A recent study on relative packing groups (RPGs) suggests a new angle to approach this problem.[171] A relative packing group, which can represent the smallest tertiary packing unit in a protein structure, contains a simplified packing environment within a tertiary conformation. The unique pair-wise contacts derived from these groups can be used to accurately and automatically classify folds within a super family. Furthermore, these contacts correlate well with the sequence identity, and thus provide a direct relationship between the changes in the sequence and the changes in the structure; for example, a minor 10% change in the sequence causes the packing of the globin fold to change by up to 50 contacts. These contacts are also useful for understanding the structural quality of the sequence or the structure alignments, and for providing the context necessary for calculating a value for structural randomness, both of which are important for properly accessing the quality of a structural alignment. Most importantly, the information that a relative packing group contains can be used to predict unknown folds.

To utilize the packing information of the relative packing groups, pair-wise contacts can be expressed in terms of distance or spatial constraints. Spatial constraints are widely used in NMR structure determination[172] where the chemical shifts that reflect interactions between atoms are translated into spatial constraints. By solving the distance matrix to satisfy the maximum number of spatial constraints with conformation with lowest energy, an ensemble of native structures can be defined. Several structure-predicting methods have already made use of spatial constraints in order to improve the prediction accuracy. MODELLER[173] uses spatial constraints in various steps, including homology-derived distance (Cα-Cα and backbone N-O distances), dihedral angles (backbone and side-chain dihedral angles), non-bonded inter-atomic distances from representative sets of known protein structures, and optional manually curated restraints.

All constraints are expressed as probability density functions and combined with CHARM22 force-field terms[174] in order to enforce proper stereochemistry. TASSER and TASSER-based methods,[36,175,176] on the other hand, use spatial restraints (only Cα and side-chain centers of mass) extracted from threading alignments in their Monte Carlo simulation to generate candidate structures with native-like conformations. By mimicking the experimentally determined data in NMR, spatial constraints provide a reliable way to reproduce native conformations, and have been found to be less susceptible to alignment errors.[177-179]

In this study, in an effort to refine the starting structure, we use spatial constraints derived from the conserved relative packing groups based upon multiple sequence and structure alignment. Using 52 prediction units (single domain proteins) in the CASP7 experiment, including both structures selected for submission and all the other structures generated during the structure refinement, we were able to test our packing orientated structure prediction method in a double-blind test and drew the conclusions that 1): constraints derived from the conserved relative packing groups are able to provide "added value" to the starting structure; 2) correct constraints (true positives), especially long-range constraints (>8Å), are important for improving predictions; and 3) wrong constraints (false positives) don't affect the quality of the refinement upon the starting structure. With the right constraints, a template structure derived from multiple homologues can be improved as much as 2Å in CαRMSD from 4Å away from the native conformation. Meanwhile, 4) scoring functions based on an extensive dynameome dataset, which is orthogonal to our structure refinement method, show great potential for selecting most native-like conformations.

Results and Discussion

In order to compare our results with the results from other CASP7 experiments, we based our analysis on assessment units (AU), or single domains of protein in CASP7. For simplicity, we used "target" to identify each assessment unit, "template" for the homologue structure, "starting structure" for the structure we try to refine (which is generated by MODELLER based on multiple sequence and structure alignment), "candidate structure" for any structure that was refined from the starting structure, and "submitted structure" or "submission" for structures we selected to submit to the CASP (up to five per target).

Table 8 shows the general description of all CASP7 targets used in this study. There are a total of 52 targets, 51 of which belong to a TBM (template based modeling) classification. T0309 is an FM (free modeling) target. 19 out of the 52 targets are high accuracy targets (HA). In Table 8, CαRMSD value of the starting structure, the best candidate structure and the difference between these two are shown. Targets are ordered based on their CαRMSD difference between the starting and best candidate structure (column 8: "CαRMSD Improvement"). Table 8 also shows the relative ranks of the best submissions of each target as compared to all of the submissions as percentages. Out of these 52 targets, two ranked #1 (T0340 and T0339D1), six ranked within the top 50

Table 8: Description of prediction units used in CASP7

Target[a]	Num[b] of Residue	Prediction[c] Class	Num[d] of Candidate	Relative[e] Rank (%)	CαRMSD of[f] Starting Structure (Å)	CαRMSD of[f] Best Structure (Å)	CαRMSD[g] Improvement (Å)	True[h] Positive (%)	False[h] Positive (%)	False[h] Negative (%)	Num of[i] Native Constraints
T0331	149	TBM	274	88.5	18.4	14.8	3.6	29.5	42.7	70.5	543
T0382	123	TBM	163	93.1	17.4	14.1	3.4	46.2	34.0	53.8	494
T0309	76	FM	471	51.4	14.1	10.9	3.2	60.0	114.8	40.0	135
T0327	102	TBM	560	92.8	13.1	10.4	2.7	43.6	23.2	56.4	298
T0305	297	HA	232	78.7	5.3	2.8	2.6	75.3	32.9	24.7	1296
T0302	132	HA	701	3.9	4.1	1.7	2.3	81.0	30.5	19.0	573
T0317	163	HA	440	62.7	5.8	3.5	2.3	66.0	16.9	34.0	708
T0323D1	101	TBM	327	83.4	11.4	9.2	2.2	39.2	32.7	60.8	416
T0373	147	TBM	270	74.7	6.5	4.5	2.0	76.1	36.0	23.9	506
T0329D2	92	TBM	424	67.8	8.4	6.4	1.9	66.9	91.7	33.1	384
T0359	97	HA	177	36.8	3.8	2.2	1.7	64.4	27.5	35.6	371
T0332	159	HA	292	45.9	4.1	2.5	1.6	60.5	18.5	39.5	681
T0288	93	HA	199	24.1	4.0	2.4	1.6	72.4	32.7	27.6	352
T0338D2	113	TBM	90	74.7	8.2	6.7	1.5	54.7	18.7	45.3	492
T0375	296	TBM	210	74.1	6.4	5.4	0.9	48.3	34.4	51.7	1378
T0384	325	TBM	114	73.0	7.1	6.2	0.9	55.8	43.6	44.2	1460
T0374	160	TBM	290	83.6	10.2	9.3	0.9	49.2	26.4	50.8	655
T0341D1	148	TBM	259	5.5	2.8	1.9	0.9	61.6	19.1	38.4	638
T0362	151	TBM	210	68.1	4.5	3.6	0.8	40.9	18.8	59.1	607
T0376	321	TBM	140	71.2	5.1	4.3	0.8	62.4	35.7	37.6	1444
T0338D1	143	TBM	90	55.8	10.0	9.4	0.6	58.2	13.6	41.8	625
T0379D1	140	TBM	219	59.1	3.9	3.3	0.6	62.0	46.7	38.0	608
T0297	211	TBM	150	63.4	6.0	5.4	0.5	66.2	34.6	33.8	969
T0330D2	72	TBM	226	82.9	3.2	2.7	0.5	44.1	15.7	55.9	261
T0303D2	77	TBM	444	75.9	7.4	7.1	0.4	53.4	17.0	46.6	311
T0366	106	HA	264	47.3	2.4	2.1	0.3	70.8	22.4	29.2	370
T0339D2	267	HA	65	22.1	2.6	2.3	0.3	61.3	16.0	38.7	1255
T0308	165	HA	359	19.2	2.4	2.2	0.2	73.6	23.5	26.4	758
T0339D1	136	TBM	65	0.2	2.8	2.6	0.2	59.0	15.0	41.0	588
T0315	257	HA	292	4.9	1.3	1.1	0.2	73.9	16.1	26.1	1195
T0329D1	66	TBM	424	11.2	1.5	1.3	0.2	87.5	34.7	12.5	216
T0379D2	64	TBM	219	38.6	4.1	3.9	0.2	66.5	28.0	33.5	239
T0381D2	176	TBM	216	2.9	1.9	1.8	0.1	71.2	19.2	28.8	798
T0340	90	HA	248	0.2	2.2	2.2	0.1	84.1	23.5	15.9	358
T0326	304	HA	177	59.8	11.7	11.6	0.1	77.1	18.6	22.9	1385
T0371D1	162	TBM	216	29.8	3.7	3.6	0.1	60.7	25.3	39.3	697
T0323D2	116	TBM	327	80.5	4.1	4.1	0.0	56.8	39.6	43.2	470
T0341D2	104	TBM	259	49.2	2.5	2.5	0.0	67.6	30.3	32.4	413
T0378D2	142	TBM	216	22.3	2.0	2.0	-0.1	57.9	21.9	42.1	636
T0345	185	HA	201	8.1	1.0	1.1	-0.1	89.2	23.2	10.8	883
T0313	322	HA	320	11.8	2.6	2.7	-0.1	70.8	26.4	29.2	1488
T0330D1	153	TBM	226	20.3	3.2	3.4	-0.1	49.5	32.1	50.5	691
T0324D1	142	HA	324	62.2	2.0	2.3	-0.2	57.8	28.9	42.2	612
T0303D1	147	HA	444	33.3	2.0	2.2	-0.3	57.1	24.3	42.9	671
T0371D2	121	TBM	216	42.7	2.4	2.7	-0.3	68.7	32.8	31.3	482
T0346	172	HA	240	47.7	0.5	0.8	-0.3	82.0	45.6	18.0	807
T0318D2	335	TBM	70	40.5	2.5	2.9	-0.4	76.8	26.9	23.2	1661

Table 8: Continued

Target[a]	Num[b] of Residue	Prediction[c] Class	Num[d] of Candidate	Relative[e] Rank (%)	CαRMSD of[f] Starting Structure (Å)	CαRMSD of[f] Best Structure (Å)	CαRMSD[g] Improvement (Å)	True[h] Positive (%)	False[h] Positive (%)	False[h] Negative (%)	Num of[i] Native Constraints
T0380	145	TBM	180	47.4	2.9	3.4	-0.5	41.6	17.6	58.4	546
T0381D1	61	TBM	216	77.6	2.2	2.8	-0.5	41.6	14.2	58.4	226
T0324D2	65	HA	324	54.3	2.1	2.9	-0.8	65.2	15.0	34.8	233
T0318D1	154	TBM	70	8.9	8.5	9.6	-1.0	45.2	63.0	54.8	659
T0334	530	HA	223	68.5	2.5	3.9	-1.3	38.2	12.4	61.8	2684

[a.] Target name as single domain prediction unit, D1 and D2 indicate different domains of the same target.

[b.] Number of residues

[c.] CASP7 official classification of prediction type. HA: high resolution. TBM: template based modeling. FM: free modeling. (HA is part of TBM but has higher resolution)

[d.] Total number of candidate structures generated during structure refinement

[e.] Rank of the best submitted structure as percentile among all prediction groups (average 500 per target)

[f.] CαRMSD value of the starting structure and the best candidate structure comparing to native structure.

[g] CαRMSD value difference between the best candidate structure and the starting structure. Rows are ordered by the value. Positive values mean improvement over the starting structure while negative value means no improvement.

[h] Number of constraints derived from the conserved relative packing groups as percentage of total constraints calculated based on the native structure. TP: True Positive (right constraints), FP: False Positive (wrong constraints, FN: False Negative (missing constraints)

[i.] Total number of constraints calculated based on the native structure

(T0302, T0315, T0341D1, T0381D2, T0318D1 and T0345), and more than half of the submissions ranked within the top 50%. Our results suggest that the accuracy of the predicted structure when using our method was above average among the over 500 prediction groups. With a relatively smaller number of constraints, our method provides a reliable way to sample the conformational space near the native conformation.

The "added value"

It has been shown that including information from multiple templates provides "added value"[37,38] to the starting structure, which moves the starting structure closer to its native conformation. Unfortunately, no group has consistently performed better than the "virtual prediction group," which simply replaces the corresponding residues using the best template found in the PDB database. In order to evaluate the capabilities of our method to provide "added value", in another word to improve the starting structure, we only compared our candidate structures to the starting structure we generated for each target. Since the native structure was not available when the candidate structures were generated, we assumed that our starting structure was the best representation of native conformation at the time of refinement, even though it may not have been the best available template structure identified after the CASP7 experiments.

Figure 15 shows the CαRMSD values of the starting structure and the best candidate structure comparing to the native structure for each target. The best candidate structure is the structure generated during the refinement, but not necessarily selected for submission after the scoring. The smaller the CαRMSD value, the closer the structure is

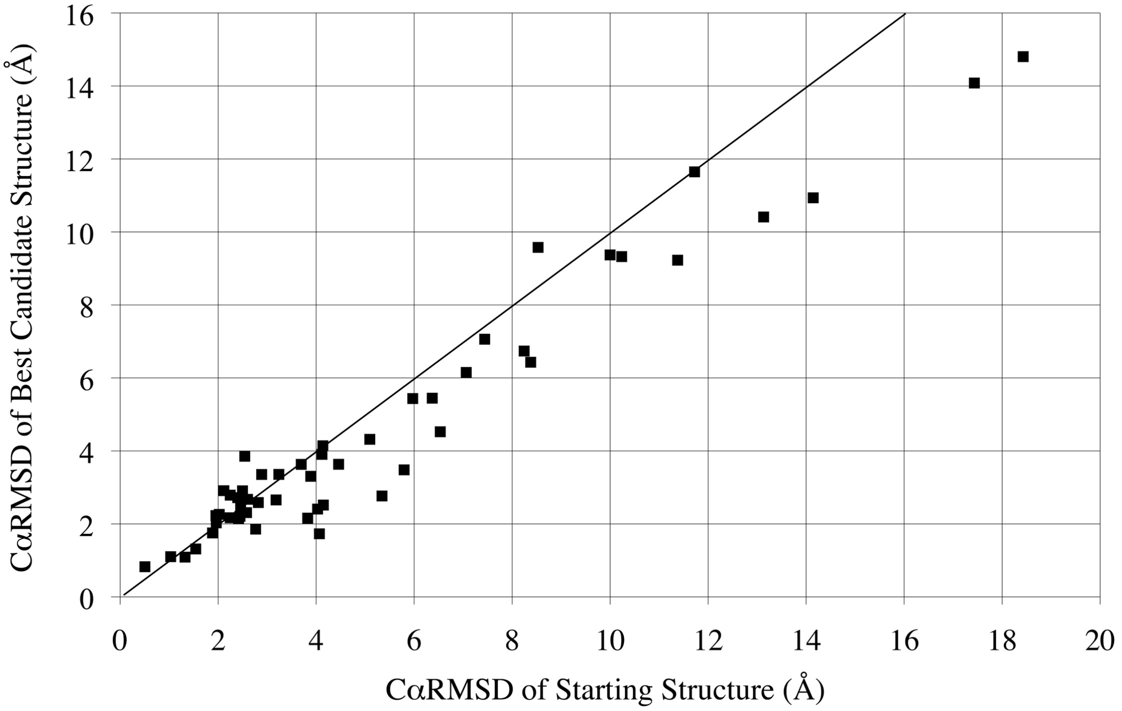


Figure 15: CαRMSD value of best candidate structure and starting structure. CαRMSD values of the best candidate structure and the starting structure were plotted when compared to the native structure for each target. A line was draw where these two values were equal to each other. Any data points below this line indicate a smaller value for the best candidate structure, which means an improvement on the starting structure. About two thirds of the targets have a candidate structure with smaller CαRMSD value and the majority of the rest of the targets are concentrated in an area where the starting structure has CαRMSD values range between 2 and 3.5 Å.

to the native conformation, and thus the better the prediction is. In Figure 15, any data point below the diagonal line indicates that there is at least one candidate structure better than the starting structure. About two thirds of the targets have a better candidate structure, and it is clear that we were able to sample the conformational spaces towards the native conformation from the starting structure. In other words, our prediction method is able to refine the starting structure towards the native conformation. Targets without any better candidate structures cluster into regions where the CαRMSD is smaller than 3.5Å. Structure improvement in this region is relatively moderate (except T0315, which shows 15% improvement, about 0.2 over 1.3 Å on CαRMSD value) compared to that in regions where larger CαRMSD values were observed. Such a decrease in the level of improvement is well-known in the field of template based structure prediction: the closer the template structure is to the native conformation, the harder the refinement becomes, and the more the refining procedure tends to push the template away from the native conformation. At the same time, the magnitude of the structure improvement (the difference between the starting structure and the best candidate structure) depends upon the quality of the starting structure; our method can only sample the conformational space close to the starting structure towards the native conformation and it will be trapped in the local minima if the starting structure is too far away from the native conformation. For a few targets, structure refinement was started using only the primary sequence as an extended chain instead of using the starting structure. With the same set of constraints derived from the conserved relative packing groups, we were hoping to sample the conformational space more freely without being trapped at the local minima. Candidate structures with comparable quality were generated using such method, but due to the time limit of CASP7 experiments and the limited computer resources, there was not enough data for a reliable analysis and further investigation is expected to make full us of this approach. In summary, with a starting structure based on close homologues and constraints based on conversed packing information, our method was able to refine the starting structure towards the native conformation. In other words, we provided “added value” to the template structure.

RPG derived constraints

One of the differences of our method from other methods using spatial constraints is that the constraints used in our method were derived from the conserved relative packing groups. It is believed that these relative packing groups are able to characterize the packing environment of the protein tertiary conformation and reflect the sequence changes onto structural changes when compared within the same folding family,[171] or in other words, among homologue structures. To study the strengths and weakness of these constraints, we classified our constraints after converting them as the percentage of the total number of constraints calculated based on the input constraints (Table 8) as follows: TP - True Positive (correctly predicted constraints), FP - False Positive (incorrectly predicted), and FN - False Negative (missing constraints). Figure 16 shows the averaged percentage of TP and FP for residues with different levels of structure deviation. For all the residues of all the candidate structures, the residues were

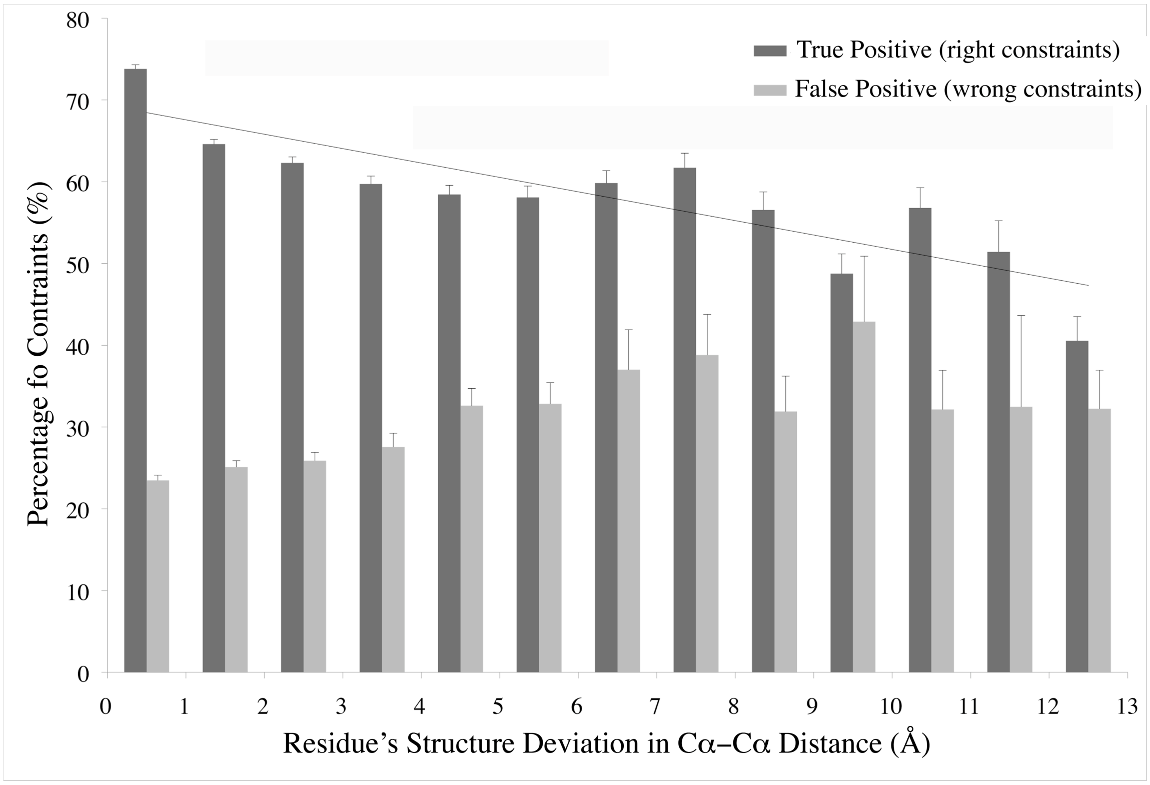


Figure 16: Percentage of constraints of residues with different structure deviation.

The averaged percentages of two types of constraints were plotted for residues with different level of structure deviation in terms of Cα-Cα distance. The Cα-Cα distance was calculated between matching residues of the best candidate structure and the native structure after they were supposed on each other. Each histogram bar represents the average value of the percentage of constraints for residues with the same level of deviation. Black bar represents values for right constraints: TP or True Positive. They are the constraints we used as input and can be also identified in the native structure. Grey bar repents values for wrong constraints: FP or False Positive. They are the constraints we used but that can't be identified in the native structure. A trend line was also shown for TP data to demonstrate that the quality of prediction decreases (increase in residue's deviation) when number of right constraints (TP%) decrease.

grouped base on their structure deviation, which is measured as Cα-Cα distance between matching residues after superposing the candidate structure with the native structure. Different level of structure deviation can be seen to represent different accuracy of the structure prediction. The smaller the distance is, the better the prediction. Within each level, averaged percentage of two different constraints (TP and FP) was calculated for all the residues in the same level. The larger the TP percentile is, the better the input constraints used in the refinement represent the native conformation.

Figure 16 clearly shows that larger TP values correspond to better predictions. As the TP percentage decreases, the Cα-Cα distance increases, indicating that the conformation of the predicted structure was moving away from the native conformation. Since the value of TP and FN added up to 100% or 1, the prediction quality decreases as the FN value increases (data points not shown), which means that lacking of the constraints that represent the native conformations will prevent our method from improving the starting structure. Wrong constraints (FP) showed no obvious correlation with the prediction quality. The FP value doesn't change consistently as the Cα-Cα distance changes. Such lack of correlation between the FP and the prediction quality suggests that our method is able to retain starting structure conformations even though some of the constraints are wrong. For example, one of the number one ranked targets, T0340, has a starting structure of CαRMSD value 1.0 Å, which is very close to the native structure. Even though 24% (see Table 8) wrongly predicted constraints (FP) were used, only a few were satisfied in the candidate structure, and none of our submitted structures show obvious deviations from the starting structure. The error (FP) can arise from various sources when the constraints are being calculated. For example, the “over-packed” template structures may contain conserved secondary structure elements or packing environments that are not suitable for the target sequence. One of the explanations of such behavior of FP constraints relies on the use of the starting structure. Since the starting structure already contains a certain level of native structural information from the multiple sequence alignment, satisfying wrongly predicted constraints will result in higher energy conformations.

For relative packing groups, members with two residues contacting each other are most abundant (data not shown). To be considered as RPG, all residues within the RPG group have to contact one another. Figure 17 shows the schematic view of all contacts among residues. Example of residues forming RPG with 4 members is colored blue and residues not forming RPG of 4 members are colored red, which instead form two RPGs with 3 members each. For one RPG with four members, average constraints between Cα atoms are labeled. These constraints are averaged among residues that align together in the multiple sequence alignment. There are only a few groups with five members and rarely any groups with six members due to the strict criteria for defining such relative packing groups. Thus, constraints derived from these groups won’t exceed 20Å, and a lot of these constraints are short distance constraints that define secondary structure (which are the major components of FP constraints). Compared to the constraints derived from NOSEY experimental data, which are between 1.8 and 6.0 Å,[172] constraints derived from the RPGs are more sensitive to weak interactions (constraints longer than 8Å). Based on our analysis, these long-range distance constraints that

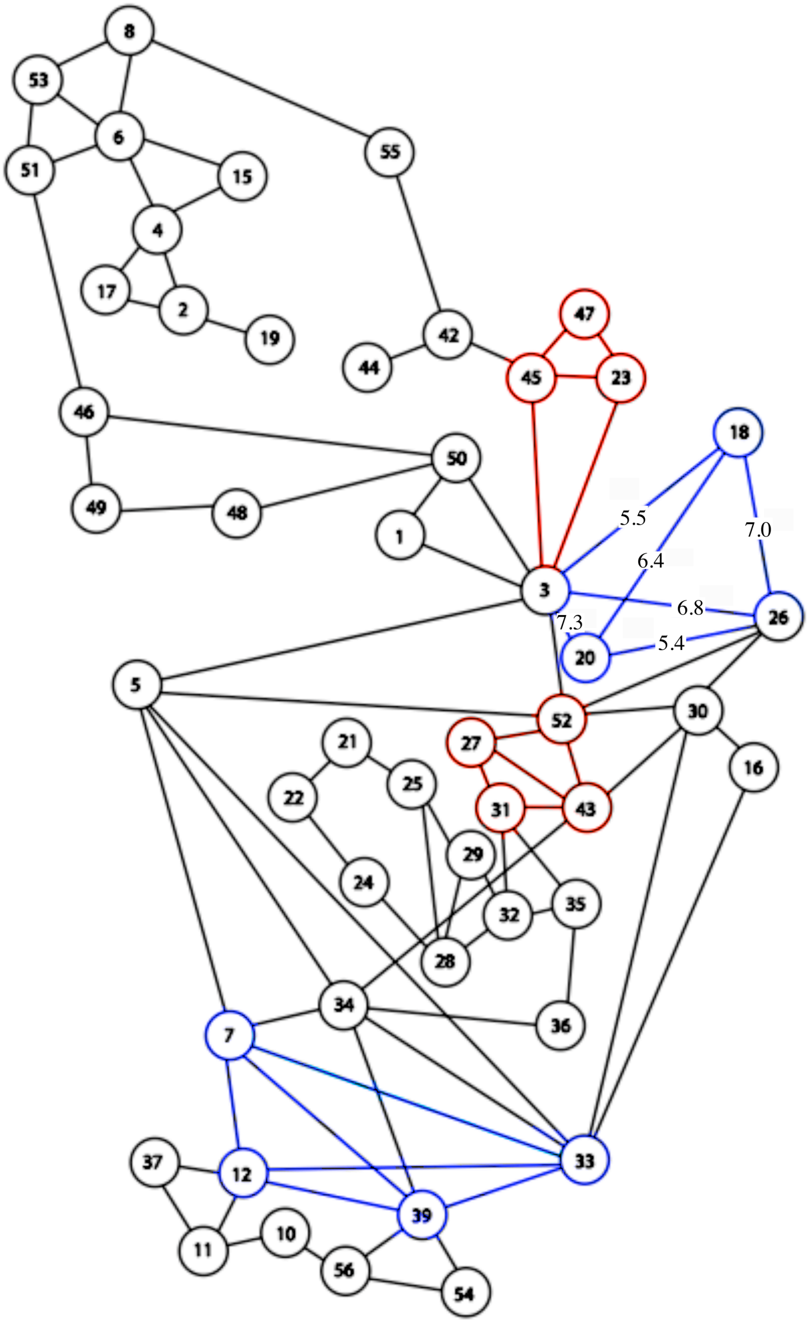


Figure 17: Schematic view of RPG and derived constraints.
Schematic view of all contacts among residues for protein 1PGB[100] are shown. Examples of residues forming RPG with 4 members are colored blue and residues not forming RPG with 4 members are colored red. Example of constraints between Cα atoms in Å are labeled for one of the blue RPGs.

beyond the NMR data range are of great importance for improving the starting structure.

Figure 18 shows the relationship between the quality of the prediction (GDT score from the CASP7 official assessment for the best submitted structure) and the percentage of the correct long-range constraints belonging to the TP class, which stands for the percentage of constraints that are longer than 8Å and can be observed in the native structure. The GDT score measures how well the candidate structure and the native structure can be superimposed on each other. A perfect match will give a score of 100, and the higher the score is, the better the prediction is. Figure 18 clearly shows that the better we reproduce the long-range constraints, the closer the final structure is to the native conformation. Target T0326 is one exception, which shows a relatively high TP percentage (50%) but a lower GDT_TS score, around 50, as compared to other predictions with the same TP percentage. Such exception is due to the fact that we chose to include residues 1 to 33 in the model for refinement. Since no conserved RPGs were found for these residues, no long-range constraints were used to refine the starting structure. As a result, these residues became an extended chain in our submitted structure, thus lowered the GDT score even though the rest of the protein had more than 70% of the correct long-range constraints in it.

By combining multiple sequence and structure alignment, the conserved relative packing groups across the homolog structures represent a set of residues close in space, or in other words, a conserved packing environment. These packing groups have the

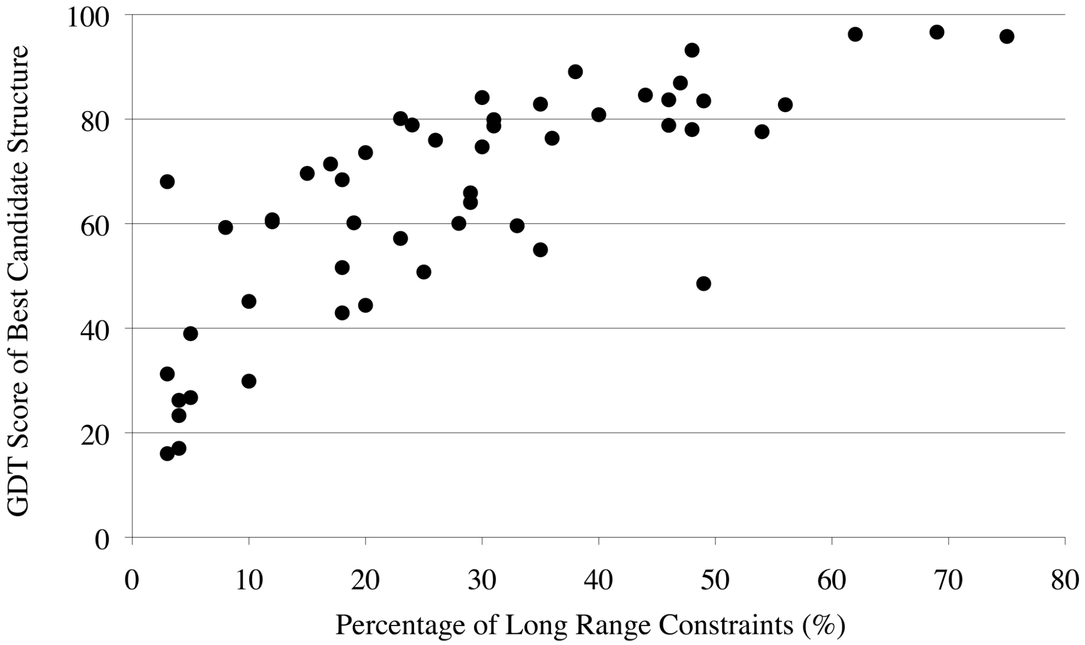


Figure 18: GDT score of best candidate versus the percentage of long range constraints.

The GDT score of the best candidate structure was plotted against the percentage of the long-range constraints (distance > 8Å) that can be identified as TP constraints (True Positive or right constraints) for every target. The higher the GDT score, the closer the candidate structure is to native conformation. As shown above, larger values in the percentages correspond to a better structure, which indicates that the long-range constraints are important in our structure predictions.

potential to represent the packing conservation (rather than the sequence conservation) over the process of evolution. With constraints similar to the experimental NMR data and the long-range constraints representing weak interactions, our method provides an alternative way to sample conformational space and was able to refine the starting structure towards the native conformation as shown in Figure 15. Moreover, a recent study shows that the relative orientation of protein sidechain can be used to improve the template based structure prediction.[180] The constraints in our method not only contain such information (by defining constraints between every atom) but also offer higher resolution by providing the relative position of side-chain C_β atoms.

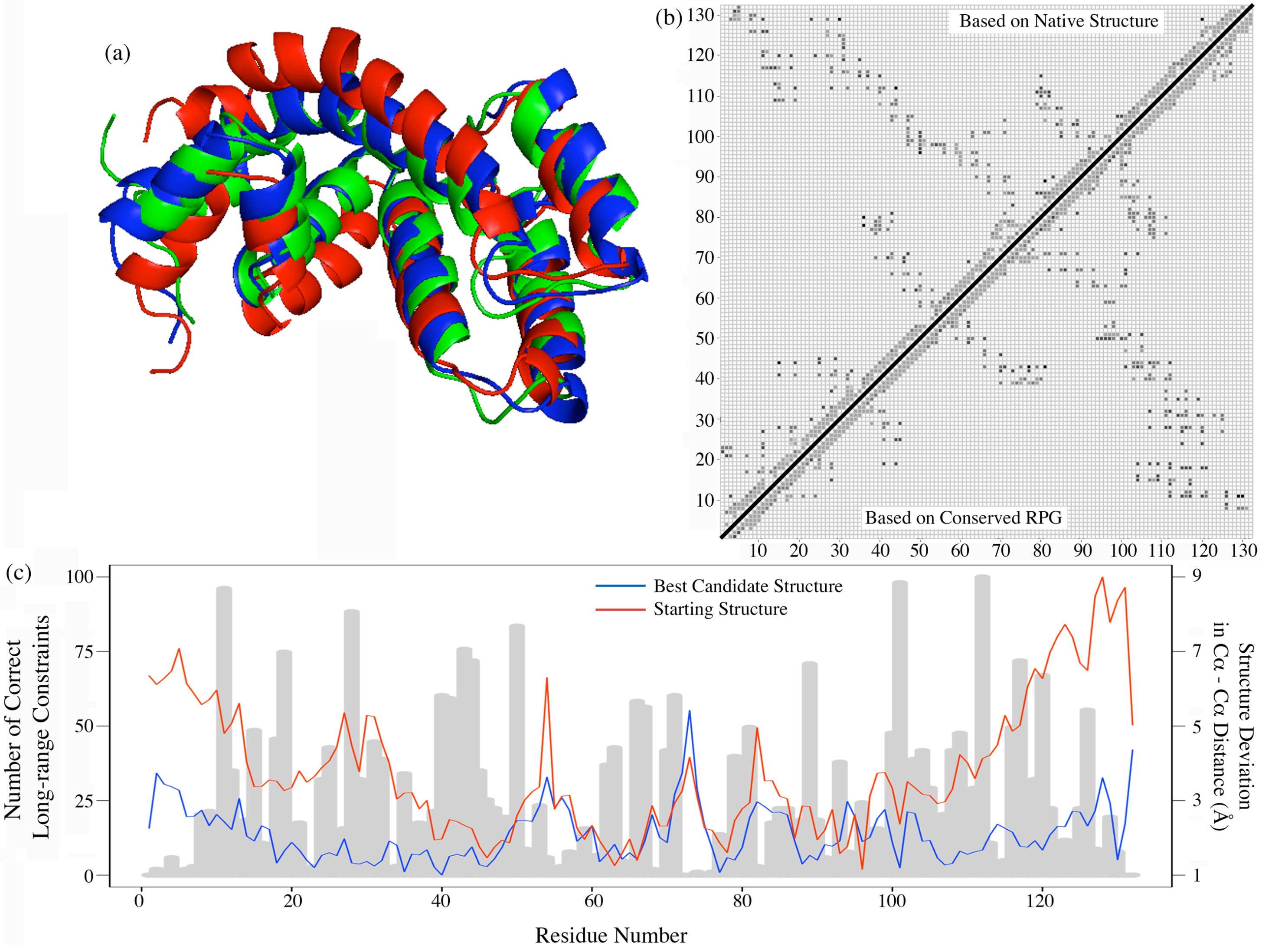
Successful refinement

There are several elements in our method that contribute to the success of improving the starting structure towards the native conformation. First, similar to other successful methods in CASP7, our method makes full use of multiple sequence alignment instead of only one template. By collecting pair-wise constraints derived from the conserved relative packing groups, we mapped the structural conservation to the sequence changes. Second, by using MODELLAR to generate a starting structure close to the native conformation, we were able to narrow our conformation space search. Third, with long-range constraints representing weak interactions in the specific folding environment, we were able to sample conformational space more efficiently towards the native conformation, and provide more candidate structures in a limited time for scoring, especially for larger proteins with well-defined long-range constraints. Last, by using a scoring function that is independent of our structure refinement method and derived from a huge dynameome dataset containing structures with near-native conformation, we were able to reinforce the selection of structures that best represent the native conformation.

Figure 19 shows a case study of our capability of providing “added value” to the starting structure for target T0302. The starting structure is generated based on 14 templates, including the best template identified by CASP7 assessors (1argE/H[181]). 426 out of a total of 701 candidate structures had a smaller value of CαRMSD than the one the starting structure had, which is 4.1 Å when compared to the native structure. Our best submitted structure ranked at 20 and has a CαRMSD value of 1.7 Å, which is about a 60% improvement over the starting structure. Figure 19a shows the superposition of the starting structure (red), the best candidate structure (blue), and the native structure (green). Target T0302 is an all-alpha protein with 132 residues; the starting structure had all the secondary structure elements correct due to the use of the multiple sequence alignment, but the distance between the three termini alpha helices is clearly different from their native value, which results in the larger structure deviation observed in the starting structure. Figure 19b shows the map of the pair-wise constraints derived from the native structure (the upper left region above the diagonal line) and constraints derived from the conserved relative packing groups (the bottom right region below the diagonal line), which we used to refine the starting structure. The data points parallel to the diagonal line indicate the local interaction of the alpha helix and the data points perpendicular to the diagonal line indicate the interactions between two helices. There

Figure 19: Case study of a successful structure prediction.

Case study of target T0302 as an example of successful prediction. a) Superposition of three structures: native structure (green), best candidate structure (blue), and starting structure (red). The helices at the end of the starting structure are further away, as compared to the best candidate structure and the native structure. b) Constraints map for every residue pair. The constraints map was divided into two sections by a solid line. The top left section represents the pair-wise constraints derived from the native structure. The bottom right section represents the pair-wise constraints derived from the conserved relative packing groups. c) The number of correct long-range constraints (right constraints with a distance longer than 8Å) in relation to the residue's structure deviation in terms of Cα-Cα distance for both of the starting structure (red) and the best candidate structure (blue) when compared to the native structure for every residue. The gray histogram indicates the number of correct long-range constraints at left y-axis while the Cα-Cα distance was plotted at right y-axis.

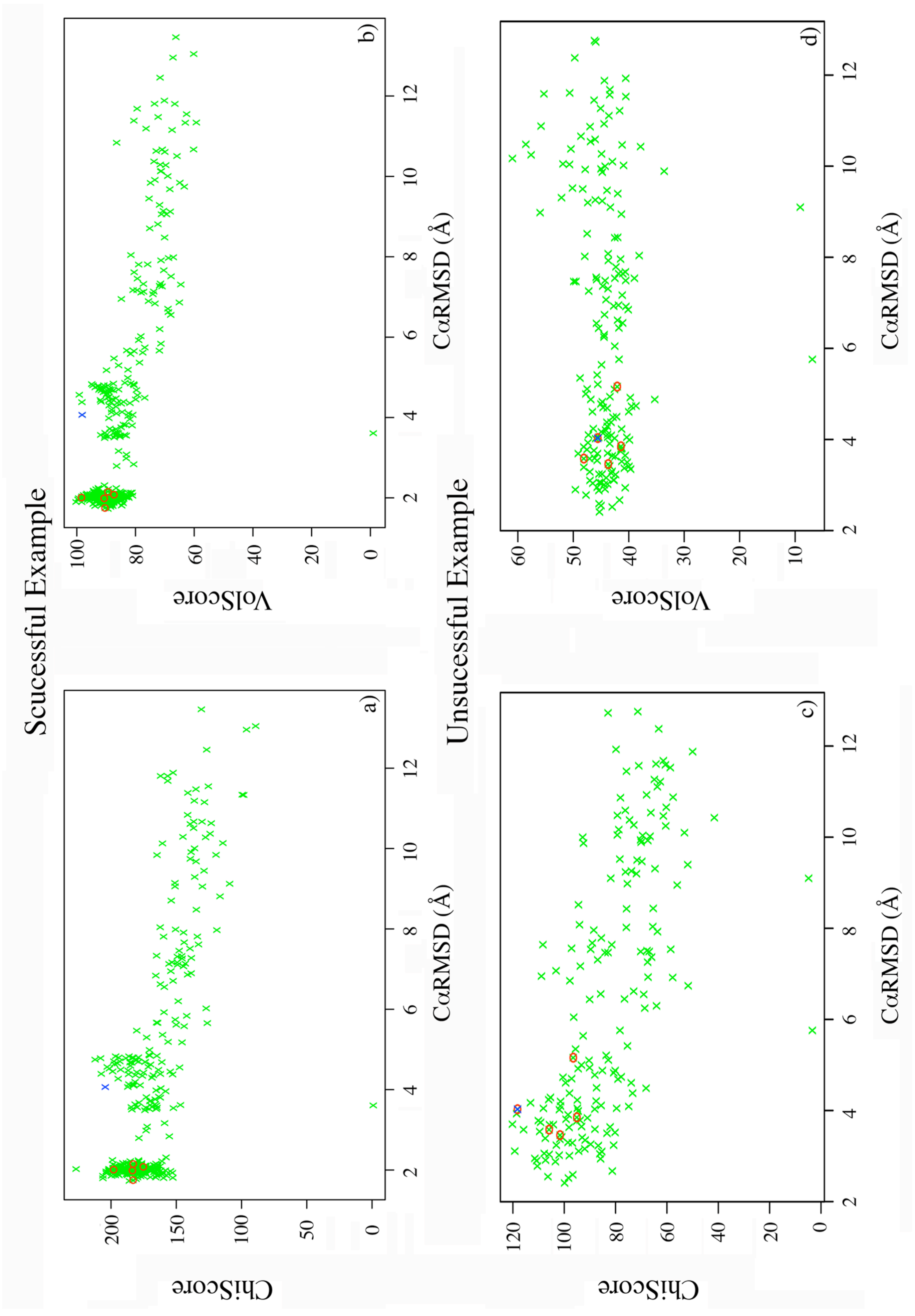

(a)
(b)
Based on Native Structure
Based on Conserved RPG
130
120
110
100
90
80
70
60
50
40
30
20
10
10 20 30 40 50 60 70 80 90 100 110 120 130
(c)
Number of Correct
Long-range Constraints
100
75
50
25
0
Best Candidate Structure
Starting Structure
Structure Deviation
in Cα - Cα Distance (Å)
9
7
5
3
1
0 20 40 60 80 100 120
Residue Number

were several conserved relative packing groups (in the bottom right of Figure 19b) can be identified by the constraints as the conserved packing environment, such as the region defined by constraints between residues 60-80 and 40-50. The termini helices mentioned above in the starting structure with larger deviations were defined by the constraints between residues 10-30 and 100-130. Compared to the native constraints, our input constraints captured most of the recognizable packing regions belong to the native structure, especially those involve termini helices. The bottom-right section in Figure 19b is almost a mirror image of the top-left section. Figure 19c shows the relationship between residue's structure deviation for each residue of the starting structure and the best candidate structure (right y-axis) and the number of correctly predicted long-range constraints (left y-axis). It shows that the biggest structure improvement from the starting structure (blue) to the best candidate structure (red) comes from the two termini regions. It also shows that when the number of correct long-range constraints decreases (the region around residue 55 and residue 75), the improvement is minimized or doesn't exist at all. All together, with the correct constraints derived from the conserved relative packing groups, our method is able to refine the starting structure towards the native conformation. It is worth noticing that T0302 is the only target we predicted that uses NMR structures as the native structure, which indicates that our method may have the potential to expedite the structure determination procedure with significantly reduced constraints from the experimental NMR data.

Another unique element that makes our method successful is the use of a scoring function that scoring candidate structures based on their packing efficiency and side-chain conformation preferences. These two properties were measured as the volume of the residue and the first rotamer angle χ_1 of the side-chain, which is not directly modeled by the spatial constraints we used, but rather as a result of proper packing. In other words, our scoring function is totally independent of our refinement method, and a higher score will reflect a truly better packing environment as the result of structure refinement. Furthermore, the volume and χ_1 score are based on a large dynameome dataset generated by the molecular dynamics simulations. Such datasets provide a fine-grained, high-resolution estimation of the native ensembles, thus gave us a chance to develop a scoring function that could sample structures more continuously into the conformational space. Figure 20 shows the example of performance of our scoring function. Figure 20a and b show χ_1 value score and volume score for target T0302 as an example of successful scoring. Figure 20c and d show the score for target T0288 as an example of unsuccessful scoring. We define successful scoring as the ability to find structures with lowest CαRMSD values and unsuccessful scoring as it fails to find such structure when it exists among the candidate structures. In Figure 20a and 20b, both χ_1 score and volume score are able to identify relative better structures for submission (red circle) from all the candidate structure (green cross) and those submission structures are better than the starting structure (blue cross). While in Figure 20c and d, even though some submitted structures are better than the starting structure, the scoring function fail to find the best candidate structure. During scoring, in order to achieve structural diversity for submissions in CASP7, candidate structures are first subject to cluster

Figure 20: Examples of successful and unsuccessful scoring.

Examples of successful and unsuccessful scoring of two targets are shown. a) and b) are successful scoring of both ChiScore (score based on sidechain rotamer conformation) and VolScore (score based on residue’s volume) on target T0302, respectively. c) and d) are unsuccessful scoring of both ChiScore and VolScore on target T0288, respectively. The overall score for starting structure (blue cross), submitted structures (red circle) and every candidate structure (green cross) are plotted against CαRMSD values of each target. Large score and small CαRMSD value are expected for native-like conformation.


Scucessful Example
Unsucessful Example
a)
b)
c)
d)
ChiScore
VolScore
CαRMSD (Å)
0
50
100
150
200
0
20
40
60
80
100
120
10
30
50
60
2
4
6
8
10
12

program based on their similarity. From the top five largest clusters, each cluster center and four other closest members in the same cluster were selected using the score function. As a drawback, structures that don't belong to the five largest clusters are omitted from the selection procedure. As noticed by other groups,[182] there are relatively few structures being sampled that very close to native conformation at the near-native conformational spaces when the sampling time is limited. As seen in Figure 20 c and d, though our method is able to sample towards the native conformation, our scoring function was unable to select the best structure available because they don't belong to the 5 largest clusters. At the same time, better structures (low CαRMSD value) are not always corresponding to higher scores, especially in Figure 20d, where structures with larger CαRMSD value are having higher score. This is due to the use of static water box to calculate volume, in which cavity can be expected at the water-protein interface (see method for details).

As a first attempt to utilize such a large dataset, our scoring function is based on simple statistics from the dynameome dataset on the frequency of packing (residue volume) and side-chain conformation (χ_1 angles) at given backbone conformation. For each volume and χ_1 angle in the candidate structure, the score was calculated as the log value of the frequency that the same value can be observed in the dynameome dataset with same conformation and residue type. Though our scoring function is in its primitive state, it shows great potential for selecting the most available "best" structures. We believe that a fine-grained scoring function on a more statistically sound basis will show improved selecting power.

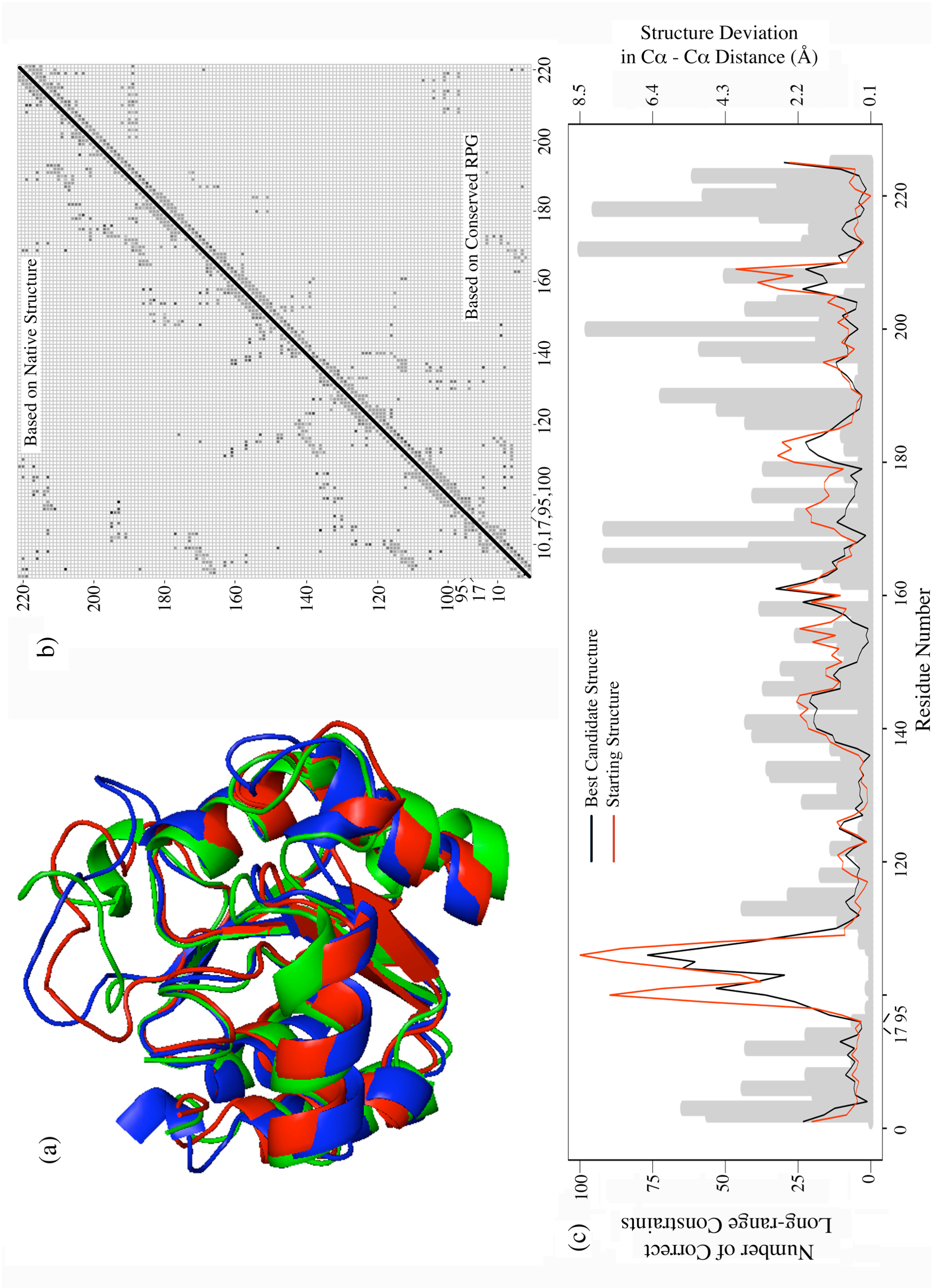
Limitation of constraints derived from the RPGs

During our prediction, relatively few and more specific constraints were used to generate the candidate structures. On one hand, the use of close template structures makes our method capable of maintaining the native conformation even with wrongly predicted constraints (FP in Figure 16). On the other hand, since the number of constraints is limited, our method shows its limitations when inadequate constraints are obtained from the template structures.

Based on our analysis, a lack of proper constraints is the dominant fact for not being able to improve the starting structure. Target T0334, for example, has a starting structure with a CαRMSD value of 2.5 Å, while the CαRMSD value of the best candidate structure is 3.9 Å (Table 8). For this target, we got most of the secondary structure elements right, but the orientation and the position of these secondary structures are wrong. Since T0334 is the largest target (530 residues) we used in this study, and only one template structure was available at the time of prediction, it has the smallest constraint to residue ratio among all the targets and most of them are short range constraints that define secondary structures (data not shown). Such lack of long-range constraints defining conserved packing environments makes our structure refinement procedure inefficient to improve the starting structure. Figure 21 shows a case study for another target: T0303D1, as an example of unsuccessful prediction. Figure 21a, b and c are generated in the same way as Figure 19. Since T0303D1 has only 147 residues and is defined as the residue of 1-17 and 95-224, residues 18 to 94 were

Figure 21: Case study of an unsuccessful prediction.

Case study of target T0303D1 as an example of an unsuccessful prediction. a) Superposition of three structures: native structure (green), best candidate structure (blue) and starting structure (red). b) Constraints map for every residue pair. The constraints map was divided into two sections by a solid line. The top left section represents pair-wise constraints derived from the native structure. The bottom right section represents the pair-wise constraints we derived from the conserved relative packing groups. c) The number of correct long-range constraints (right constraints with a distance longer than 8Å) in relation to the residue's structure deviation in terms of Cα-Cα distance for both of the starting structure (red) and the best candidate structure (black) when compared to the native structure for every residue. The gray histogram indicates the number of correct long-range constraints at left y-axis while the Cα-Cα distance was plotted at right y-axis.


(a)
b)
Based on Native Structure
Based on Conserved RPG
220
200
180
160
140
120
100
95
17
10
10,17,95,100
(c)
Number of Correct
Long-range Constraints
100
75
50
25
0
Best Candidate Structure
Starting Structure
Residue Number
0
17 95
120
140
160
180
200
220
Structure Deviation
in Cα - Cα Distance (Å)
8.5
6.4
4.3
2.2
0.1

omitted from the plot. In Figure 21a, the starting structure (red) and the best candidate structure (blue) were superposed on the native structure (green), respectively. In Figure 21b, constraints derived from the native structure are in the top left region, and constraints derived from the conserved relative packing groups are in the bottom right region. Similar to Figure 19b, the data points parallel to the diagonal line indicate the local interaction of the alpha helix, and the data points perpendicular to the diagonal line indicate interactions between helices. Contrary to Figure 19b, relative packing groups derived from the native structure did not have their counterparts in the input constraints, especially in the bottom right region from residues 160 to 220. At the corresponding region in Figure 19c, which shows the number of correct long-range constraints at the left y-axis and residue's structure deviation in terms of Cα-Cα distance at the right y-axis for each residue, an increase in the structure deviation in the candidate structure (black) can be observed consistently along the sequence when compared to those values of the starting structure (red). For example, the native structure has a small helix centered on residue 206, and forms several packing groups between residues 190-210 and residues 210-220 (top-left section in Figure 21b). These interactions were missing in the bottom right region. Thus the Cα-Cα distance in this region increased in our candidate structure. Also, due to the lack of information at the end of residue 17 and residue 95, both the starting structure and the candidate structure show large Cα-Cα distance around these residues.

There are many factors that affect the quality of the prediction, such as the sequence identity, the resolution of the template structure, the efficiency of the conformational sampling, and the selectivity and sensitivity of the scoring function. It becomes very hard to decide what went wrong if a structure is far away from its native conformation. Based on our analysis, a lack of constraints in our input played an important role when other factors were considered the same. In other words, in order to improve the starting structure, proper constraints are required. Without these constraints, the starting structure will have more flexibility when the conformational space is sampled. Under such condition, either the structure refinement procedures won't have enough time to sample the right conformation space or the energy function used in the refinement becomes inefficient to distinguish the low energy conformation from the high-energy conformation. Again, from another point of view, our method has the ability to predict important constraints and use these constraints to improve the starting structure.

Conclusion

In this study, by using 52 assessment units from the CASP7 experiment, we analyzed the ability of our packing orientated structure prediction algorithm on improving the starting structure that are generated based on homologues. Using spatial constraints derived from the conserved relative packing groups, we were able to provide "added value" to the starting structures for a majority of the CASP7 targets. TP (correctly predicted) constraints are important to the quality of the prediction, and FP (wrongly predicted) constraints have little effect on the prediction quality. With a relatively small number of constraints representing the conserved packing environment over multiple homologues, especially long-range constraints, we were able to improve the starting structure up to 2Å in CαRMSD value. Meanwhile, the scoring function we developed based on the large dynameome dataset also shows its strength on selecting native-like conformations. We believe that our sidechain packing orientated structure prediction algorithm provides a new angle of mapping the sequence changes to the structure changes, and has the potential to improve the template-based structure predictions, as well as experimental NMR structure determinations.

Materials and Methods

Starting structure

The starting structure was generated using the sequence alignment BLAST,[183] structure alignment MUSTANG,[184] and structure modeling MODELLER.[173] First, a target sequence was searched against the PDB database using BLAST, and template structures were selected based on the BLAST E-value. At least one template structure with a significant E-value was selected as the parent structure, and up to 43 templates were used for one target prediction (T0305). All template structural files were downloaded from the local PDB database and cleaned using in-house programs to remove heteroatoms and alternative conformations. MUSTANG was used to apply a structure alignment in order to calculate the conserved relative packing groups. Using structure-based alignment from MUSTANG as the guideline, up to 20 template structures were selected based on their similarity, and inputted to MODELLER to generate starting structures using the quick build algorithm. Several possible starting structures were generated, verified and compared to the parent structure by a human expert before being selected as the final starting structure, the starting point of structure refinement.

Candidate structure generation

Candidate structures were built using the NIH-Xplor[185] package with spatial constraints derived from the reserved relative packing groups, based upon multiple sequence and structure alignment. Three types of constraints were used: torsion angle constraints, distance constraints between main-chain atoms, and distance constraints between side chain C_β atoms. Torsion angle constraints and main-chain distance constrains were applied to all targets, while side chain constraints were only applied to a few targets with the parent structure of high sequence identity. Two structure modeling methods were used: with and without a starting structure. For all targets, a starting structure was generated using MODELLER, and the constraints were applied to refine the starting structure towards the native conformation. For a few selected targets, which were either small or well packed or had only one template structure, an extended chain of target sequence was used as an alternative starting point. All generated structures were collected and scored by the same scoring function for conformations closest to the native structure. The average time for building 100 residues was about 10 hours on an Athon 1800+ CPU. In order to sample the conformational space more efficiently, different seed numbers were used for different XPLOR jobs. Around 300 candidate structures were generated for each target on average using an 84-node cluster, as well as 11 desktops with Pentium IV 2.4G CPUs.

Torsion angel constraints

Torsion angle constraints (ϕ, ψ and χ) were derived from MUSTANG-based multiple structural alignments of all template structures. For each target residue, the angle value was calculated as the average value of all the corresponding residues at the same aligned position, regardless of their residue type. For allowed variations,[186] the value was calculated as the standard deviation from all the corresponding residues. If the standard deviation was not available, the variation was set to 30º. The ϕ value of the first residue and ψ value of the last residue were omitted, even when their alignment position was in the middle of the template sequence.

Spatial constraints

First, relative packing groups of every template structure were calculated separately, as described previously.[171] Those relative packing groups were then annotated using target sequence numbers based on the multiple sequence alignments, regardless of their residue compositions. If any member of the relative packing groups is in non-aligned region, such relative packing groups were omitted from later calculations. A conservation cutoff (a default of 50%) was used to select conserved relative packing groups, for example, to be considered as a conserved RPGs, besides align together, five out of ten template structures had to form the same RPGs, which means not only their sequence is conserved but also the contacts between them have to be conserved, in another word, the same packing environment. Within each conserved relative packing group, for each pair of mainchain atoms between different residues (members), the average distance among different templates as well as the standard deviation of the distance was used to define the spatial constraints and allowed variation of such interaction. Similar procedures were used to derive the spatial distance constraints between the C_β atoms in the relative packing groups.

PDB constraints

For structures with only a few homologues, an alternative method was used to derive the spatial distance constraints. First, a database of relative packing groups based upon non-redundant, high resolution PDB structures[150] was generated. All the relative packing groups in the database were clustered based on their number of members, secondary structure composition, and the structure similarity of all four main chain atoms (main-chain RMSD). The cluster center was used to represent the relative packing groups. In order to calculate the spatial distance constraints using the PDB-derived database, each relative packing group derived from the template structure was compared to the PDB generated database. The closest cluster center with the same number or members, the same secondary structure composition and the smallest main-chain RMSD value as chosen by aligning the calculate relative packing group to the center group. Up to 100 randomly selected cluster members were then used to calculate the average distance and the standard deviation as spatial distance constraints and allowed variations, respectively.

Scoring function

The scoring function is based on simple statistics of the dynameome dataset on the frequency of packing (residue volume) and side-chain conformation (χ_1 angles) at given backbone conformation. For each volume and χ_1 angle in the candidate structure, the score was calculated as the log value of the frequency that the same conformation can be observed in the dynameome dataset. In order to calculate the volume, each structure was placed in the center of a water box to mimic the protein solution environment. This water box was taken from a MD simulation of pure water using the same parameters as in the protein simulations. Duplication of water box is applied if necessary to generate large enough box for protein. Any water atom within a distance of 1.8Å of protein atoms was removed.

Analysis

All the structural comparisons between the native structure and structure of interest, such as the starting structure, best candidate structure and submitted structure, use the same sets of residues consistent with the official CASP assessment. Residues in the target sequence that didn’t correspond to any native structure were omitted.

Programs written in C and PERL were created to analyze the native structures. A MySQL database was used to store and analyze different properties for every structure and distance constraint. 3D structure models were viewed using MacPyMol.[118] Figures were generated using the R[120] and Microsoft Excel program. CαRMSD values were calculated using the Kabsch and Sander method.[119] Using the same method, the structure deviation of each residue observed in Figure 19 and Figure 21 can be calculated as the Cα-Cα distance of corresponding residues after superpose two structures. Relative packing groups was calculated the same way as they were done in the structure generation procedure for both the native structure and the structure of interest.

To generate Figure 16, residues were first grouped by their structure deviation distance every 1 Å. For every residue belongs to the same group, the percentage of three different constraints were calculated when compared to the number of constraints observed in the native structure for each residue. The percentage of true positive (TP), false positive (FP) and false native (FN) was calculated by comparing to the number of native constraints, based on which, the sum of percentage of TP and FN will always equal to 1. To distinguish the true positive (TP), false positive (FP) and false negative (FN) constraints, the distance constraints from the native structure were first aligned to the target sequence. Constraints were considered true positives if the constraints exist between two residues that can both be aligned. False positives were those constraints that only existed as input, and false negatives were those constraints that only existed in the native structure. The number of long-range constraints was defined as those constraints that are longer than the distance cutoff (8 Å) in the native structure. The number of constraints was converted into the percentage of either the total number of native constraints or the total number of input constraints so that cross target comparison become possible.

CHAPTER V

SUMMARY AND CONCLUSION

In this study, molecular dynamics simulations were used to generate ensembles of different protein structures to sample the native conformational space, in order to understand the stability and dynamics of protein structures. First β-hairpin, as a simple secondary structure element, was studied with an emphasis on the specific ion pair interactions at the ends of two free termini. Results show that termini ion-pairs are able to stabilize β-hairpin as “pseudo hydrogen bond” by directly keeping the termini from opening up and indirectly cooperating with the main-chain hydrogen bond network, as well as the core hydrophobic interactions. Such analysis helps us to have a better understanding of the hairpin conformation and give us the opportunity to optimize β-hairpin stability. Second, near 4 million structures were generated and used to represent the ensemble of near-native conformations, providing us with a continuous conformational space to better characterize the side-chain packing and conformation upon the influence of backbone conformations. We were able to determine the contribution of the local and non-local environments on the residue volume with implications about side-chain packing. Because such dynameome datasets allow for more fine-grained analysis, we were also able to more precisely define the relationship between the first side-chain rotamer conformation and the backbone conformation. These results help us to define the exact role that the backbone conformation plays on the determination of the protein fold. Last, based on the information we obtained from the molecular dynamics simulations, we developed a side-chain packing orientated method for template-based protein structure prediction. Using spatial constraints derived from the conserved relative packing groups to mimic the experimental NMR data, we were able to provide "added value" to the starting structure, which is derived from the multiple sequence and structure alignment. Our method depends upon the correct prediction of long-range constraints using homologous, but was not significantly affected by the wrongly predicted constraints (false positives). The scoring function we used base on the dynameome dataset also show its selective power against native-like conformations.

In conclusion, MD simulations provide a way to sample conformations in the native conformational space with a realistic level of physical accuracy. Structures generated using MD simulations can be used to provide detailed, high-resolution representation and a better understating of protein structure at the molecular level. The information derived from the large-scale MD simulations (dynameome dataset) can be used as a knowledge base to improve current methods of protein structure prediction and structure refinement.

REFERENCES

1. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* **181**, 223-30.

2. Levinthal, C. (1968). Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique* **65**, 44-45.

3. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42.

4. Neumann, J. V. (1932). Proof of the Quasi-Ergodic Hypothesis. *Proc Natl Acad Sci U S A* **18**, 70-82.

5. Warshel, A. (1976). Bicycle-pedal model for the first step in the vision process. *Nature* **260**, 679-83.

6. McCammon, J. A., Gelin, B. R. & Karplus, M. (1977). Dynamics of folded proteins. *Nature* **267**, 585-90.

7. Tirado-Rives, J. & Jorgensen, W. L. (1991). Molecular dynamics simulations of the unfolding of an alpha-helical analogue of ribonuclease A S-peptide in water. *Biochemistry* **30**, 3864-71.

8. Karplus, M. & Kuriyan, J. (2005). Molecular dynamics and protein function. *Proc Natl Acad Sci U S A* **102**, 6679-85.

9. Rueda, M., Ferrer-Costa, C., Meyer, T., Perez, A., Camps, J., Hospital, A., Gelpi, J. L. & Orozco, M. (2007). A consensus view of protein dynamics. *Proc Natl Acad Sci U S A* **104**, 796-801.

10. Day, R., Beck, D. A., Armen, R. S. & Daggett, V. (2003). A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci* **12**, 2150-60.

11. Rost, B., Fariselli, P. & Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci* **5**, 1704-18.

12. Krieger, E., Nabuurs, S. B. & Vriend, G. (2003). Homology modeling. *Methods Biochem Anal* **44**, 509-23.

13. Kopp, J., Bordoli, L., Battey, J. N., Kiefer, F. & Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. *Proteins* **69**, 38-56.

14. Zemla, A., Venclovas, C., Reinhardt, A., Fidelis, K. & Hubbard, T. J. (1997). Numerical criteria for the evaluation of *ab initio* predictions of protein structure. *Proteins* **1** Suppl, 140-50.

15. Nanias, M., Chinchio, M., Oldziej, S., Czaplewski, C. & Scheraga, H. A. (2005). Protein structure prediction with the UNRES force-field using Replica-Exchange Monte Carlo-with-Minimization; Comparison with MCM, CSA, and CFMC. *J Comput Chem* **26**, 1472-86.

16. Bradley, P., Misura, K. M. & Baker, D. (2005). Toward high-resolution *de novo* structure prediction for small proteins. *Science* **309**, 1868-71.

17. Goldsmith-Fischman, S. & Honig, B. (2003). Structural genomics: computational methods for structure analysis. *Protein Sci* **12**, 1813-21.

18. Liu, X., Fan, K. & Wang, W. (2004). The number of protein folds and their distribution over families in nature. *Proteins* **54**, 491-9.

19. Brenner, S. E., Chothia, C. & Hubbard, T. J. (1997). Population statistics of protein structures: lessons from structural classifications. *Curr Opin Struct Biol* **7**, 369-76.

20. Du, P., Andrec, M. & Levy, R. M. (2003). Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Eng* **16**, 407-14.

21. Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature* **357**, 543-4.

22. Grant, A., Lee, D. & Orengo, C. (2004). Progress towards mapping the universe of protein folds. *Genome Biol* **5**, 107.

23. Yan, Y. & Moult, J. (2005). Protein family clustering for structural genomics. *J Mol Biol* **353**, 744-59.

24. Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science* **294**, 93-6.

25. Ginalski, K. (2006). Comparative modeling for protein structure prediction. *Curr Opin Struct Biol* **16**, 172-7.

26. Xiang, Z. (2006). Advances in homology protein structure modeling. *Curr Protein Pept Sci* **7**, 217-27.

27. Sanchez, R. & Sali, A. (1997). Advances in comparative protein-structure modelling. *Curr Opin Struct Biol* **7**, 206-14.

28. Floudas, C. A., Fung, H. K., McAllister, S. R., Monnigmann, M. & Rajgaria, R. (2006). Advances in protein structure prediction and *de novo* protein design: A review. *Chemical Engineering Science* **61**, 966-988.

29. Qu, X., Swanson, R., Day, R. & Tsai, J. (2009). A guide to template based structure prediction. *Curr Protein Pept Sci* **10**, 270-85.

30. Tramontano, A. & Morea, V. (2003). Assessment of homology-based predictions in CASP5. *Proteins* **53** Suppl 6, 352-68.

31. Tress, M., Ezkurdia, I., Grana, O., Lopez, G. & Valencia, A. (2005). Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins* **61** Suppl 7, 27-45.

32. Venclovas, C. (2003). Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. *Proteins* **53** Suppl 6, 380-8.

33. Venclovas, C. & Margelevicius, M. (2005). Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins* **61** Suppl 7, 99-105.

34. Dunbrack, R. L., Jr. (1999). Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. *Proteins* **3** Suppl, 81-7.

35. Ginalski, K. & Rychlewski, L. (2003). Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins* **53** Suppl 6, 410-7.

36. Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69** Suppl 8, 108-17.

37. Kryshtafovych, A., Fidelis, K. & Moult, J. (2007). Progress from CASP6 to CASP7. *Proteins* **69** Suppl 8, 194-207.

38. Kryshtafovych, A., Venclovas, C., Fidelis, K. & Moult, J. (2005). Progress over the first decade of CASP experiments. *Proteins* **61** Suppl 7, 225-36.

39. Banerjee, R., Sen, M., Bhattacharya, D. & Saha, P. (2003). The jigsaw puzzle model: search for conformational specificity in protein interiors. *J. Mol. Biol* **333**, 211-226.

40. Richards, F. M. & Lim, W. A. (1993). An analysis of packing in the protein folding problem. *Q Rev Biophys* **26**, 423-98.

41. Gassner, N. C., Baase, W. A. & Matthews, B. W. (1996). A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proc Natl Acad Sci U S A* **93**, 12155-8.

42. Best, R. B., Rutherford, T. J., Freund, S. M. & Clarke, J. (2004). Hydrophobic core fluidity of homologous protein domains: relation of side-chain dynamics to core composition and packing. *Biochemistry* **43**, 1145-55.

43. Beasley, J. R. & Hecht, M. H. (1997). Protein design: the choice of *de novo* sequences. *J Biol Chem* **272**, 2031-4.

44. Bromberg, S. & Dill, K. A. (1994). Side-chain entropy and packing in proteins. *Protein Sci* **3**, 997-1009.

45. Skolnick, J. (2006). In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol* **16**, 166-71.

46. Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr Opin Struct Biol* **5**, 229-35.

47. Muegge, I. (2006). PMF scoring revisited. *J Med Chem* **49**, 5895-902.

48. Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34**, 82-95.

49. Hooft, R. W., Vriend, G., Sander, C. & Abola, E. E. (1996). Errors in protein structures. *Nature* **381**, 272.

50. Eisenberg, D., Luthy, R. & Bowie, J. U. (1997). VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol* **277**, 396-404.

51. Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* **15**, 285-9.

52. Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T. & Tramontano, A. (2007). Critical assessment of methods of protein structure prediction-Round VII. *Proteins* **69** Suppl 8, 3-9.

53. Moult, J., Fidelis, K., Rost, B., Hubbard, T. & Tramontano, A. (2005). Critical assessment of methods of protein structure prediction (CASP)--round 6. *Proteins* **61** Suppl 7, 3-7.

54. Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. (2003). Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins* **53** Suppl 6, 334-9.

55. Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K. & Pedersen, J. T. (1997). Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins* **1** Suppl, 2-6.

56. Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins* **3** Suppl, 2-6.

57. Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. (2001). Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins* **5** Suppl, 2-7.

58. Fischer, D., Rychlewski, L., Dunbrack, R. L., Jr., Ortiz, A. R. & Elofsson, A. (2003). CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins* **53** Suppl 6, 503-16.

59. Fischer, D., Elofsson, A., Rychlewski, L., Pazos, F., Valencia, A., Rost, B., Ortiz, A. R. & Dunbrack, R. L., Jr. (2001). CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins* **5** Suppl, 171-83.

60. Fischer, D., Barret, C., Bryson, K., Elofsson, A., Godzik, A., Jones, D., Karplus, K. J., Kelley, L. A., MacCallum, R. M., Pawowski, K., Rost, B., Rychlewski, L. & Sternberg, M. (1999). CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins* **3** Suppl, 209-17.

61. Rychlewski, L. & Fischer, D. (2005). LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci* **14**, 240-5.

62. Koh, I. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A. & Rost, B. (2003). EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res* **31**, 3311-5.

63. Chakrabartty, A. & Baldwin, R. L. (1995). Stability of alpha-helices. *Adv Protein Chem* **46**, 141-76.

64. Williams, S., Causgrove, T. P., Gilmanshin, R., Fang, K. S., Callender, R. H., Woodruff, W. H. & Dyer, R. B. (1996). Fast events in protein folding: helix melting and formation in a small peptide. *Biochemistry* **35**, 691-7.

65. Scholtz, J. M., Marqusee, S., Baldwin, R. L., York, E. J., Stewart, J. M., Santoro, M. & Bolen, D. W. (1991). Calorimetric determination of the enthalpy change for the alpha-helix to coil transition of an alanine peptide in water. *Proc Natl Acad Sci U S A* **88**, 2854-8.

66. Blanco, F. J., Jimenez, M. A., Pineda, A., Rico, M., Santoro, J. & Nieto, J. L. (1994). NMR solution structure of the isolated N-terminal fragment of protein-G B1 domain. Evidence of trifluoroethanol induced native-like beta-hairpin formation. *Biochemistry* **33**, 6004-14.

67. Blanco, F. J., Rivas, G. & Serrano, L. (1994). A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nat Struct Biol* **1**, 584-90.

68. Blanco, F. J. J., M.A.; Herrantz, J.; Rico, M.; Santoro, J. and Nieto, J.L. (1993). NMR evidence of a short linear peptide that folds into a .beta.-hairpin in aqueous solution. *J. Am. Chem. Soc.* **115**, 5887-5888.

69. Sibanda, B. L., Blundell, T. L. & Thornton, J. M. (1989). Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J Mol Biol* **206**, 759-77.

70. Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv Protein Chem* **34**, 167-339.

71. Huyghues-Despointes, B. M., Qu, X., Tsai, J. & Scholtz, J. M. (2006). Terminal ion pairs stabilize the second beta-hairpin of the B1 domain of protein G. *Proteins* **63**, 1005-17.

72. Dhanasekaran, M., Prakash, O., Gong, Y. X. & Baures, P. W. (2004). Expected and unexpected results from combined beta-hairpin design elements. *Org Biomol Chem* **2**, 2071-82.

73. Du, D., Zhu, Y., Huang, C. Y. & Gai, F. (2004). Understanding the key factors that control the rate of beta-hairpin folding. *Proc Natl Acad Sci U S A* **101**, 15915-20.

74. Chen, R. P., Huang, J. J., Chen, H. L., Jan, H., Velusamy, M., Lee, C. T., Fann, W., Larsen, R. W. & Chan, S. I. (2004). Measuring the refolding of beta-sheets with different turn sequences on a nanosecond time scale. *Proc Natl Acad Sci U S A* **101**, 7305-10.

75. Kolinski, A., Ilkowski, B. & Skolnick, J. (1999). Dynamics and thermodynamics of beta-hairpin assembly: insights from various simulation techniques. *Biophys J* **77**, 2942-52.

76. Roccatano, D., Amadei, A., Di Nola, A. & Berendsen, H. J. (1999). A molecular dynamics study of the 41-56 beta-hairpin from B1 domain of protein G. *Protein Sci* **8**, 2130-43.

77. Pande, V. S. & Rokhsar, D. S. (1999). Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G. *Proc Natl Acad Sci U S A* **96**, 9062-7.

78. Ma, B. & Nussinov, R. (1999). Explicit and implicit water simulations of a beta-hairpin peptide. *Proteins* **37**, 73-87.

79. Zagrovic, B., Sorin, E. J. & Pande, V. (2001). Beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *J Mol Biol* **313**, 151-69.

80. Lee, J. & Shin, S. (2001). Understanding beta-hairpin formation by molecular dynamics simulations of unfolding. *Biophys J* **81**, 2507-16.

81. Munoz, V., Thompson, P. A., Hofrichter, J. & Eaton, W. A. (1997). Folding dynamics and mechanism of beta-hairpin formation. *Nature* **390**, 196-9.

82. Munoz, V., Henry, E. R., Hofrichter, J. & Eaton, W. A. (1998). A statistical mechanical model for beta-hairpin kinetics. *Proc Natl Acad Sci U S A* **95**, 5872-9.

83. Klimov, D. K. & Thirumalai, D. (2000). Mechanisms and kinetics of beta-hairpin formation. *Proc Natl Acad Sci U S A* **97**, 2544-9.

84. Kobayashi, N., Honda, S., Yoshii, H. & Munekata, E. (2000). Role of side-chains in the cooperative beta-hairpin folding of the short C-terminal fragment derived from streptococcal protein G. *Biochemistry* **39**, 6564-71.

85. Ciani, B., Jourdan, M. & Searle, M. S. (2003). Stabilization of beta-hairpin peptides by salt bridges: role of preorganization in the energetic contribution of weak interactions. *J Am Chem Soc* **125**, 9038-47.

86. de Alba, E., Rico, M. & Jimenez, M. A. (1997). Cross-strand side-chain interactions versus turn conformation in beta-hairpins. *Protein Sci* **6**, 2548-60.

87. Hughes, R. M. & Waters, M. L. (2006). Model systems for beta-hairpins and beta-sheets. *Curr Opin Struct Biol* **16**, 514-24.

88. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of beta-hairpin formation in protein G folding. *Nat Struct Biol* **7**, 669-73.

89. Frank, M. K., Clore, G. M. & Gronenborn, A. M. (1995). Structural and dynamic characterization of the urea denatured state of the immunoglobulin binding domain of streptococcal protein G by multidimensional heteronuclear NMR spectroscopy. *Protein Sci* **4**, 2605-15.

90. Ma, B. & Nussinov, R. (2000). Molecular dynamics simulations of a beta-hairpin fragment of protein G: balance between side-chain and backbone forces. *J Mol Biol* **296**, 1091-104.

91. Sheinerman, F. B. & Brooks, C. L., 3rd. (1998). Calculations on folding of segment B1 of streptococcal protein G. *J Mol Biol* **278**, 439-56.

92. Tsai, J. & Levitt, M. (2002). Evidence of turn and salt bridge contributions to [beta]-hairpin stability: MD simulations of C-terminal fragment from the B1 domain of protein G. *Biophysical Chemistry* **101-102**, 187-201.

93. Bolhuis, P. G. (2003). Transition-path sampling of beta-hairpin folding. *Proc Natl Acad Sci U S A* **100**, 12129-34.

94. Colombo, G., De Mori, G. M. & Roccatano, D. (2003). Interplay between hydrophobic cluster and loop propensity in beta-hairpin formation: a mechanistic study. *Protein Sci* **12**, 538-50.

95. Dinner, A. R., Lazaridis, T. & Karplus, M. (1999). Understanding beta-hairpin formation. *Proc Natl Acad Sci U S A* **96**, 9068-73.

96. Bonvin, A. M. & van Gunsteren, W. F. (2000). beta-hairpin stability and folding: molecular dynamics studies of the first beta-hairpin of tendamistat. *J Mol Biol* **296**, 255-68.

97. Kiehna, S. E. & Waters, M. L. (2003). Sequence dependence of beta-hairpin structure: comparison of a salt bridge and an aromatic interaction. *Protein Sci* **12**, 2657-67.

98. Ciani, B., Jourdan, M. & Searle, M. S. (2003). Stabilization of β-Hairpin peptides by salt bridges: role of preorganization in the energetic contribution of weak interactions. *J. Am. Chem. Soc.* **125**, 9038-9047.

99. Fesinmeyer, R. M., Hudson, F. M. & Andersen, N. H. (2004). Enhanced hairpin stability through loop design: the case of the protein G B1 domain hairpin. *J Am Chem Soc* **126**, 7238-43.

100. Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G. L. (1994). Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry* **33**, 4721-9.

101. Kraulis, P. J. (1991). MOLSCRIPT: a program to product both detailed and schematic plots of protein structures. *Journal of Applied Crystallography* **24**, 946-950.

102. Honda, S., Kobayashi, N. & Munekata, E. (2000). Thermodynamics of a beta-hairpin structure: evidence for cooperative formation of folding nucleus. *J Mol Biol* **295**, 269-78.

103. Santiveri, C. M., Santoro, J., Rico, M. & Jimenez, M. A. (2004). Factors involved in the stability of isolated beta-sheets: turn sequence, beta-sheet twisting, and hydrophobic surface burial. *Protein Sci* **13**, 1134-47.

104. Struthers, M. D., Cheng, R. P. & Imperiali, B. (1996). Design of a monomeric 23-residue polypeptide with defined tertiary structure. *Science* **271**, 342-5.

105. Eva de Alba, M. Angeles Jime¡änez & Rico, M. (1997). Turn residue sequence determines beta-hairpin conformation in designed peptides. *J. Am. Chem. Soc.* **119**, 175-183.

106. Zhou, R. (2003). Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins* **53**, 148-61.

107. Santiveri, C. M., Jimenez, M. A., Rico, M., Van Gunsteren, W. F. & Daura, X. (2004). Beta-hairpin folding and stability: molecular dynamics simulations of designed peptides in aqueous solution. *J Pept Sci* **10**, 546-65.

108. Bystroff, C., Thorsson, V. & Baker, D. (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol* **301**, 173-90.

109. Karplus, P. A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci* **5**, 1406-20.

110. Mark S. Searle, S. R. G.-J., and Henry Skinner-Smith. (1999). Energetics of weak interactions in a β-hairpin peptide: electrostatic and hydrophobic contributions to stability from lysine salt bridges. *J. Am. Chem. Soc.* **121**, 11615-11620.

111. Penel, S., Morrison, R. G., Dobson, P. D., Mortishire-Smith, R. J. & Doig, A. J. (2003). Length preferences and periodicity in beta-strands. Antiparallel edge beta-sheets are more likely to finish in non-hydrogen bonded rings. *Protein Eng* **16**, 957-61.

112. Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995). Potential-energy function and parameters for simulations of the molecular-dynamics of proteins and nucleic-acids in solution. *Computer Physics Communications* **91**, 215-231.

113. Levitt, M., Hirshberg, M., Sharon, R., Laidig, K. E. & Daggett, V. (1997). Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. *Journal of Physical Chemistry B* **101**, 5051-5061.

114. Daggett, V. & Levitt, M. (1992). A model of the molten globule state from molecular dynamics simulations. *Proc Natl Acad Sci U S A* **89**, 5142-6.

115. Grindley, T. & Lind, J. E. (1971). PVT properties of water and mercury. *Journal of Chemical Physics* **54**, 3983-3989.

116. Vedam, R. & Holton, G. (1968). Specific volumes of water at high pressures obtained from ultrasonic-propagation measurements. *Journal of the Acoustical Society of America* **43**, 108-116.

117. Engh, R. A. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystall*, 392-400.

118. DeLano, W. L. (2002). The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA.

119. Kabsch, W. (1978). Discussion of solution for best rotation to relate 2 sets of vectors. *Acta Crystallographica Section A* **34**, 827-828.

120. Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988). *The New S Language*, Chapman & Hall, New York.

121. Gerstein, M., Tsai, J. & Levitt, M. (1995). The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J Mol Biol* **249**, 955-66.

122. Samudrala, R. & Moult, J. (1998). Determinants of side chain conformational preferences in protein structures. *Protein Eng* **11**, 991-7.

123. Chung, S. Y. & Subbiah, S. (1995). The use of side-chain packing methods in modeling bacteriophage repressor and cro proteins. *Protein Sci* **4**, 2300-9.

124. Chakrabarti, P. & Pal, D. (1998). Main-chain conformational features at different conformations of the side-chains in proteins. *Protein Eng* **11**, 631-47.

125. Dunbrack, R. L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* **230**, 543-74.

126. Dunbrack, R. L., Jr. & Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat Struct Biol* **1**, 334-40.

127. Dunbrack, R. L., Jr. & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* **6**, 1661-81.

128. Dunbrack, R. L., Jr. (2002). Rotamer libraries in the 21st century. *Curr Opin Struct Biol* **12**, 431-40.

129. Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). The penultimate rotamer library. *Proteins* **40**, 389-408.

130. Chakrabarti, P. & Pal, D. (2001). The interrelationships of side-chain and main-chain conformations in proteins. *Prog Biophys Mol Biol* **76**, 1-102.

131. MacArthur, M. W. & Thornton, J. M. (1999). Protein side-chain conformation: a systematic variation of chi 1 mean values with resolution - a consequence of multiple rotameric states? *Acta Crystallogr D Biol Crystallogr* **55**, 994-1004.

132. Zhao, S., Goodsell, D. S. & Olson, A. J. (2001). Analysis of a data set of paired uncomplexed protein structures: new metrics for side-chain flexibility and model evaluation. *Proteins* **43**, 271-9.

133. West, N. J. & Smith, L. J. (1998). Side-chains in native and random coil protein conformations. Analysis of NMR coupling constants and chi1 torsion angle preferences. *J Mol Biol* **280**, 867-77.

134. Yan, A. & Jernigan, R. L. (2005). How do side chains orient globally in protein structures? *Proteins* **61**, 513-22.

135. Buchete, N. V., Straub, J. E. & Thirumalai, D. (2004). Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci* **13**, 862-74.

136. Jiang, L., Kuhlman, B., Kortemme, T. & Baker, D. (2005). A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins* **58**, 893-904.

137. Shapovalov, M. V. & Dunbrack, R. L., Jr. (2007). Statistical and conformational analysis of the electron density of protein side chains. *Proteins* **66**, 279-303.

138. Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. (1997). Protein folding: the endgame. *Annu Rev Biochem* **66**, 549-79.

139. Day, R. & Daggett, V. (2003). All-atom simulations of protein folding and unfolding. *Adv Protein Chem* **66**, 373-403.

140. Beck, D. A., Jonsson, A. L., Schaeffer, R. D., Scott, K. A., Day, R., Toofanny, R. D., Alonso, D. O. & Daggett, V. (2008). Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations. *Protein Eng Des Sel* **21**, 353-68.

141. Rueda, M., Chacon, P. & Orozco, M. (2007). Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure* **15**, 565-75.

142. Kehl, C., Simms, A. M., Toofanny, R. D. & Daggett, V. (2008). Dynameomics: a multi-dimensional analysis-optimized database for dynamic protein data. *Protein Eng Des Sel* **21**, 379-86.

143. Simms, A. M., Toofanny, R. D., Kehl, C., Benson, N. C. & Daggett, V. (2008). Dynameomics: design of a computational lab workflow and scientific data repository for protein simulations. *Protein Eng Des Sel* **21**, 369-77.

144. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536-40.

145. O'Farrell, P. A., Walsh, M. A., McCarthy, A. A., Higgins, T. M., Voordouw, G. & Mayhew, S. G. (1998). Modulation of the redox potentials of FMN in *Desulfovibrio vulgaris* flavodoxin: thermodynamic properties and crystal structures of glycine-61 mutants. *Biochemistry* **37**, 8405-16.

146. Smith, G. D., Pangborn, W. A. & Blessing, R. H. (2001). Phase changes in T(3)R(3)(f) human insulin: temperature or pressure induced? *Acta Crystallogr D Biol Crystallogr* **57**, 1091-100.

147. Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *J Mol Biol* **290**, 253-66.

148. Tsai, J. & Gerstein, M. (2002). Calculations of protein volumes: sensitivity analysis and parameter database. *Bioinformatics* **18**, 985-95.

149. Albeck, S., Unger, R. & Schreiber, G. (2000). Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges. *J Mol Biol* **298**, 503-20.

150. Wang, G. & Dunbrack, R. L., Jr. (2003). PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589-91.

151. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95-9.

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

153. Mandel, N., Mandel, G., Trus, B. L., Rosenberg, J., Carlson, G. & Dickerson, R. E. (1977). Tuna cytochrome c at 2.0 A resolution. III. Coordinate optimization and comparison of structures. *J Biol Chem* **252**, 4619-36.

154. Ho, B. K., Thomas, A. & Brasseur, R. (2003). Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. *Protein Sci* **12**, 2508-22.

155. Griffiths-Jones, S. R., Sharman, G. J., Maynard, A. J. & Searle, M. S. (1998). Modulation of intrinsic phi,psi propensities of amino acids by neighbouring residues in the coil regions of protein structures: NMR analysis and dissection of a beta-hairpin peptide. *J Mol Biol* **284**, 1597-609.

156. Feig, M. (2008). Is alanine dipeptide a good model for representing the torsional preferences of protein backbones? *Journal of Chemical Theory and Computation* **4**, 1555-1564.

157. Minor, D. L., Jr. & Kim, P. S. (1994). Context is a major determinant of beta-sheet propensity. *Nature* **371**, 264-7.

158. Beck, D. A., Alonso, D. O., Inoyama, D. & Daggett, V. (2008). The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. *Proc Natl Acad Sci U S A* **105**, 12259-64.

159. Hutchinson, E. G. & Thornton, J. M. (1996). PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci* **5**, 212-20.

160. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-9.

161. Dahl, D. B., Bohannan, Z., Mo, Q., Vannucci, M. & Tsai, J. (2008). Assessing side-chain perturbations of the protein backbone: a knowledge-based classification of residue Ramachandran space. *J Mol Biol* **378**, 749-58.

162. Chandrasekaran, R. & Ramachandran, G. N. (1970). Studies on the conformation of amino acids. XI. Analysis of the observed side group conformation in proteins. *Int J Protein Res* **2**, 223-33.

163. Benedetti, E., Morelli, G., Nemethy, G. & Scheraga, H. A. (1983). Statistical and energetic analysis of side-chain conformations in oligopeptides. *Int J Pept Protein Res* **22**, 1-15.

164. Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* **193**, 775-91.

165. Kussell, E., Shimada, J. & Shakhnovich, E. I. (2001). Excluded volume in protein side-chain packing. *J Mol Biol* **311**, 183-93.

166. Kawano, Y., Kumagai, T., Muta, K., Matoba, Y., Davies, J. & Sugiyama, M. (2000). The 1.5 A crystal structure of a bleomycin resistance determinant from bleomycin-producing Streptomyces verticillus. *J Mol Biol* **295**, 915-25.

167. Voronoi, G. (1908). New parametric applications concerning the theory of quadratic forms - Second announcement. *Journal Fur Die Reine Und Angewandte Mathematik* **134**, 198-287.

168. Tramontano, A. (2003). Of men and machines. *Nat Struct Biol* **10**, 87-90.

169. Misura, K. M. & Baker, D. (2005). Progress and challenges in high-resolution refinement of protein structure models. *Proteins* **59**, 15-29.

170. Read, R. J. & Chavali, G. (2007). Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins* **69** Suppl 8, 27-37.

171. Holmes, J. B. & Tsai, J. (2005). Characterizing conserved structural contacts by pair-wise relative contacts and relative packing groups. *J Mol Biol* **354**, 706-21.

172. Fuentes, G., van Dijk, A. D. & Bonvin, A. M. (2008). Nuclear magnetic resonance-based modeling and refinement of protein three-dimensional structures and their complexes. *Methods Mol Biol* **443**, 229-55.

173. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815.

174. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir,

L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B* **102**, 3586-3616.

175. Zhou, H., Pandit, S. B., Lee, S. Y., Borreguero, J., Chen, H., Wroblewska, L. & Skolnick, J. (2007). Analysis of TASSER-based CASP7 protein structure prediction results. *Proteins* **69** Suppl 8, 90-7.

176. Zhang, Y., Arakaki, A. K. & Skolnick, J. (2005). TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* **61** Suppl 7, 91-8.

177. Wallner, B. & Elofsson, A. (2005). All are not equal: a benchmark of different homology modeling programs. *Protein Sci* **14**, 1315-27.

178. Wallner, B., Fang, H., Ohlson, T., Frey-Skott, J. & Elofsson, A. (2004). Using evolutionary information for the query and target improves fold recognition. *Proteins* **54**, 342-50.

179. Wang, G. & Dunbrack, R. L., Jr. (2004). Scoring profile-to-profile sequence alignments. *Protein Sci* **13**, 1612-26.

180. Lu, M., Dousis, A. D. & Ma, J. (2008). OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *J Mol Biol* **376**, 288-301.

181. Graber, R., Kasper, P., Malashkevich, V. N., Sandmeier, E., Berger, P., Gehring, H., Jansonius, J. N. & Christen, P. (1995). Changing the reaction specificity of a pyridoxal-5'-phosphate-dependent enzyme. *Eur J Biochem* **232**, 686-90.

182. Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M. D., Bhat, D., Chivian, D., Kim, D. E., Sheffler, W. H., Malmstrom, L., Wollacott, A. M., Wang, C., Andre, I. & Baker, D. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69** Suppl 8, 118-28.

183. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402.

184. Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. & Lesk, A. M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins* **64**, 559-74.

185. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. (2003). The Xplor-NIH NMR molecular structure determination package. *J Magn Reson* **160**, 65-73.

186. Schwieters, C. D., Kuszewski, J. J. & Clore, G. M. (2006). Using Xplor-NIH for NMR molecular structure determination. *Progress in Nuclear Magnetic Resonance Spectroscopy* **48**, 47-62.

VITA

Xiaotao Qu was born in Shenyang, China. He lived in Shenyang from 1978 to 1997 before he went to Shanghai, China and received his Bachelor of Science degree in biology from Fudan University, Shanghai, China in 2001. He entered the graduate school at Texas A&M University in September 2001 and received his Doctor of Philosophy in biochemistry in December 2009.

Mr. Qu may be reached at 14211 Les Palms Cir Apt 202, Tampa, Florida 33613. His email is xiaotaoqu@gmail.com.