Title

A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family: Implications for improving protein design

Author

Friedland, Gregory Daniel

Publication Date

2008

Peer reviewed|Thesis/dissertation

A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family: Implications for improving protein design

by

Gregory D. Friedland

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Graduate Group in Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Dedicated to my parents, who raised me to value curiosity, imagination, skepticism, and the pursuit of knowledge.

Acknowledgements

I would like to thank first my advisor, Tanja Kortemme, for her dedication to my scientific education, encouragement and many great discussions that allowed me to grow as a scientist. I would like to also thank Andrej Sali and Brian Shoichet, the members of my thesis committee, for support and guidance over many years throughout my time at UCSF. I am grateful to my collaborators including Colin Smith, Jens Meiler, Chris McClendon, Anthony Linares, Nils Lakomek, Tanya Freedman, John Kuriyan, and Christian Griesinger, without whom this work would not have been possible. Thanks to all the members of the Kortemme lab for many great discussions. Thanks to Rebecca Brown and Julie Ransom for many years of support. And most of all thanks to Mom, Dad, Aurelia and Alli for unending love and support.

Chapter 2 is a reprint of the material as it appears in The Journal of Molecular Biology. Tanja Kortemme and I designed the experiments. Anthony Linares helped perform the experiments and Colin Smith developed the Monte Carlo Backrub method.

Chapter 3 is a manuscript in preparation. Tanja Kortemme, Jens Meiler, and I designed the experiments and analyzed the data. Nils Lakomek and Christian Griesinger performed many of the RDC experiments.

Abstract

A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family: Implications for improving protein design

Conformational flexibility is key to the function of many proteins and is thus an important focus for effective computational modeling. Sampling side-chain degrees of freedom has been an integral part of many successful computational protein design methods, and backbone flexibility is increasingly being used in these efforts. The predictions of these approaches, however, have not been directly compared to experimental measurements of side-chain and backbone solution-state conformational variability. Here, we describe methods for validating side-chain and backbone flexibility modeling by comparing to two sets of solution state Nuclear Magnetic Resonance (NMR) measurements: side chain relaxation order parameters 17 proteins totaling 530 data points, and backbone amide residual dipolar couplings (RDCs) of ubiquitin in 23 alignment media. The model for backbone flexibility that we use is the "Backrub" method; a Monte Carlo protocol combining rotamer changes with motions inspired by alternative conformations observed in sub-Angstrom resolution crystal structures. First, for modeling side chain conformational variability, we use a Monte Carlo approach comparing sampling of side chains with and without backbone flexibility. Our results indicate that the fixed-backbone model performs reasonably well but including backbone flexibility leads to significant improvements in modeling side-chain order parameters. Second, we focus on modeling backbone flexibility and we present an ensemble of ubiquitin in solution that is created by first sampling conformational space without experimental information using “Backrub” motions, and then refining with residual dipolar coupling measurements (RDCs) to select the final members of the ensemble. We show that the ubiquitin Backrub ensemble is simultaneously consistent with conformational dynamics reflected in the RDCs, the conformational variability present in ubiquitin complex structures, and characteristics of the conformational and sequence diversity of ubiquitin homologs. Our ensemble representation thus supports an overall relation between native-state protein dynamics and evolutionarily sampled sequence space. The presented insights into flexibility and the methods we have developed can be applied to numerous modeling tasks, including improved modeling of sequence diversity in protein design simulations, prediction of correlated motions within proteins, and design of sequence libraries for experimental selection.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

This document is the culmination of my studies between the dates July 1, 2003 to December 12, 2008 towards my Ph.D. Since it is probable that only a small group of people will read this document, and also likely that the number of readers will decrease exponentially at each page, I have decided to organize this manuscript in a corresponding manner. First, I will provide a background of molecular biology for the non-scientific and non-biologist reader. This will include a broad introduction to the role of proteins in the processes of life and a description of protein properties, including the focus of my research: protein flexibility. Second, oriented now towards the molecular biologist reader, I will provide an introduction to the method and applications of computational protein design. This section will include some of the notable successes of design, as well as some of the limitations with regards to modeling protein flexibility. Finally, I will describe the motivation for the approaches used in the published article presented in chapter 2 and the manuscript in preparation presented in chapter 3.

Background for non-scientists and non-biologists

Despite the incredible diversity of life forms on our planet they share some basic common structures. One of these structures is the cell, the smallest living entity (with the debatable exception of viruses). Given the right nutrients, single-celled organisms (such as *E. coli*) and cells from multi-cellular organisms (such as humans) are capable of living on their own and of replicating to form copies of them. Cells vary significantly between the different branches of life and their sizes vary as well: ranging from 1 micron for some bacteria to 1 millimeter for some eukaryotic cells.

These smallest living entities are made of many different components: organelles such as energy producing chlorosomes in some prokaryotes and mitochondria and chloroplasts in eukaryotes, the cytoskeleton which provides structure and enables motion in some cells, the plasma membrane which keeps the integrity of the contents of the cell with respect to the outside environment, and many others too numerous to list here. These functional components are made up of even smaller pieces including small organic molecules, metal ions, and macromolecules. Of the macromolecules there are three main types, and their interactions are concisely summarized with the Central Dogma of molecular biology. DNA stores the genetic code with A's, C's, G's and Ts, and is transcribed into RNA. RNA can perform functions on its own or serve as a working copy of the genetic code for a specific gene and be translated into proteins. Proteins are the end product in this simplified description of the flow of information from the genetic code, and they are the molecules that perform most of the coordinated, complex behaviors occurring within the cell.

Proteins can range in size dramatically with masses ranging from thousands of Daltons (1 Dalton is approximately the mass of 1 Hydrogen atom) to millions of Daltons. They are generally composed of 20 types of amino acids whose identities and order comes from the DNA genetic code. For many proteins, the sequence of amino acids 'folds' like magnetic beads on a string in a collapsed three-dimensional shape, a shape that varies considerable across the universe of proteins. (**Figure 1-1**)

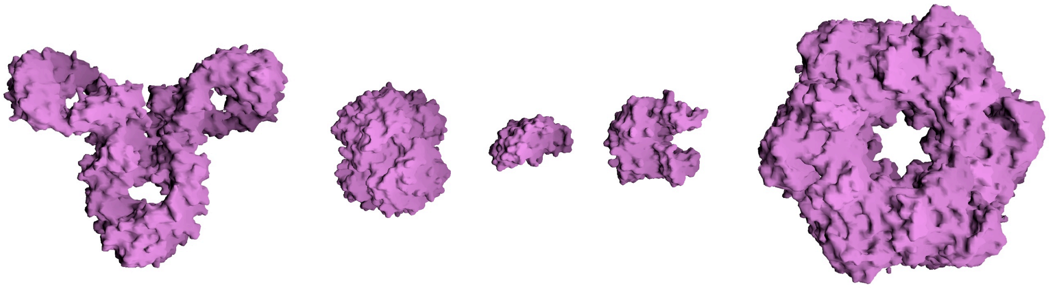


Figure 1-1. The diversity of protein shape and sizes.

Molecular surface representation of an immunoglobulin G (an antibody), hemoglobin, insulin (a hormone), adenylate kinase (an enzyme), and glutamine synthetase (an enzyme). (Licensed under Creative Commons Attribution ShareAlike 2.0)

These arrangements of atoms enable proteins to coordinate an incredible diversity of functions. Some proteins such as kinesin act as miniature motors, converting energy from ATP into a walking-like powered directional movement for the transportation of cargo. (**Figure 1-2**) Actin can generate force by forming filaments that either bundle into larger filaments or cross-link into a web, both of which can change the surface shape of cells, a process important in the pursuit and capture of bacteria by neutrophils. Another class of proteins, the ion channel, is embedded in hydrophobic membranes and provides a conduit for molecules and atoms between the many of the different compartments of the cell and the outside environment. One such ion channel is the voltage-gated potassium channel, which selectively allows the passage of potassium ions into or out of the cell while restricting the passage of very similar sodium ions. As the voltage changes during neuron firing, these channels activate and deactivate in a precise manner to give the action potential the proper shape that is needed for normal neurological activity. Another large class of proteins, including GTPases, are involved in signal transduction pathways of cells, processing information from external or internal stimuli by interacting with other signal transduction proteins and effecting changes in the cell such as the morphological changes performed by actin. Enzymes form yet another diverse class of proteins. These molecules catalyze chemical reactions, facilitating the chemical conversion of substrate molecules to products. Enzymes catalyze an incredible diversity of chemical reactions including the break down of carbohydrates in food into simple sugars and the assembly of very large polymers of sugar molecules for insertion into the outer cell wall of plants and many bacteria.

Introduction to computational protein design and the importance of flexibility

As detailed above, proteins are capable of many varied functions ranging from catalysis, to relaying information, to maintaining control over ion concentrations, to generating force. These processes are made possible because of the diverse three-dimensional structures of proteins, which are enabled by combining amino acids in different linear combinations. These processes are also possible because these three-dimensional structures are not rigid formations, but flexible entities that change conformation in concerted ways, either on their own or as a result of their interaction with different effecting forces such as physical interaction with other molecules.

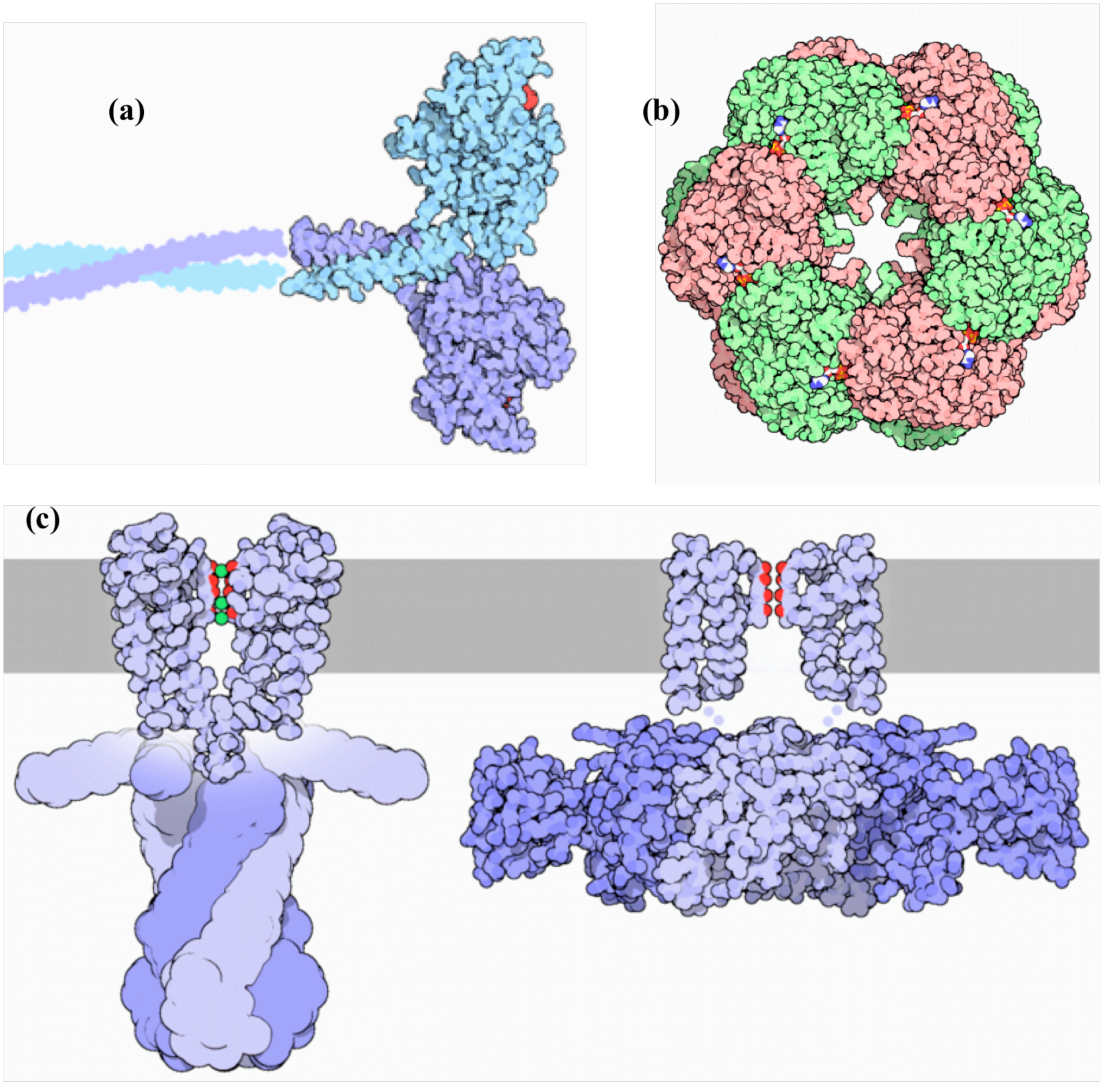


Figure 1-2. Structures of proteins with varied functions.

(a) The two motor domains of a kinesin molecule attached to its stalk. **(b)** 12 glutamine synthetase molecules come together in the active form of the enzyme. **(c)** Open and closed conformations of a voltage-gated bacterial potassium channel. (Images are from the public domain.)

This wide variety of possible protein functions made accessible by changes in the amino acid sequence has been used by evolution to create specific combinations of proteins that work together in a specific organism to achieve specific functions. The diversity of living organisms and the proteins contained within them are evidence of this idea. There are recent examples as well of this evolution, for example, the acquired ability of “superbugs,” such as methicillin-resistant Staphylococcus aureus (MRSA), to defend against antibiotic drugs.

Given that evolution has used this programmable amino acid code to generate a diversity of functions that aid in the survival of organisms, it is not surprising that this programmability can be applied in a rational way to engineer proteins for other purposes, a field known as protein engineering. One example of this is the discovery of a protein found in cold-water fish known to disrupt the formation of ice crystals at very low temperatures. This protein has been modified and applied in commercial products such as ‘Double Churn’ ice cream to increase its ‘creaminess’. Numerous other examples of protein engineering exist and different methods are used to enable the engineering of new functions.

One important tool for protein engineering is computational protein design, which involves using structural modeling to guide in the selection of amino acid sequences that have a target structure or function. This is a relatively new field that began perhaps with the insight of Ponder and Richards in 1987 [1] that there are a limited number of side chain conformations (‘rotamers’) that could be enumerated. Since then there have been numerous successes that highlight the potential of computational protein design. Kuhlman et al. used design methods to create a novel protein fold not seen in nature and verified with X-ray crystallography that the model of the structure was similar within atomic-level accuracy to the experimental structure. [2] Many studies have used design to increase thermodynamic stability or to engineer binding to small molecules such as the nerve agent soman and TNT. [3; 4] Though still a very difficult problem, the redesign of protein-protein interaction specificity has implications for rewiring the control of cellular processes and over the years this process has had more and more successes. In addition, in the past few years there have been several successful designs of catalytic activity into non-catalytic proteins, [5; 6] broadening the applications of computational design into the sphere of engineering of enzymes and biosynthetic pathways.

These successes point towards a future for computational protein design in the engineering of many different biological functions. Using a variety of approaches, enzymes have been engineered with enhanced activity, biosynthetic pathways have been engineered to produce a low-cost Malarial drug, and several projects are engineering the breakdown of cellulosic plant matter to form biofuels. These efforts are not yet guided strongly by computational protein design but there are many potential roles for structure-based "rational" design.

The typical computational design process starts with taking an input structure, defining which residues to allow to change sequence and which to allow to relax their structure in response. Subsequently an optimization procedure is run which allows changes to amino acid sequence and side chain conformations but keeps the peptide backbone fixed. This latter assumption is one of convenience; changing the backbone results in more differences from the input structure, increasing the potential for errors. Optimizing the side chain conformations on a fixed backbone has been shown to be a tractable computational problem whereas changing the backbone results in a large increase in computational complexity, and the correct strategy to choose for making backbone conformational changes is not self-evident. However, ignoring backbone changes misses out on a large component of protein flexibility and an important mechanism for proteins to adapt to amino acid mutations.

Protein flexibility is a key property of proteins that enables many of their diverse functions, and has been increasingly addressed in computational protein design method development in recent years. The methods used include Molecular Dynamics (MD), small dihedral angle changes, normal mode analysis, parameterized structural changes, and various methods inspired by conformational changes observed in X-ray crystal structures. These approaches, have been applied and evaluated by their ability to generate designed proteins with the desired functions; however, they have not been directly compared to measurements of protein flexibility in solution, raising the questions of which of them is the most appropriate for use in design and whether improvements can be gained by modeling protein flexibility as it occurs in solution.

My approach to these questions has been two-fold: to learn what I can about the structural details underpinning the solution-state flexibility of proteins, and to apply this knowledge to improve protein design. I have chosen to use as my mechanism of backbone motion the Backrub mechanism. Backrub conformational changes were frequently observed in ultra-high resolution crystal structures to facilitate correlated changes of backbone and side chain conformations. Subsequently, the method was generalized to allow changes of internal peptide sequences of arbitrary length and incorporated it into a protocol that statistically applied the move across all parts of the protein. [7] Briefly, the move consists of the selection of a random peptide segment in the protein of interest, and a rigid body rotation of atoms in the peptide segment about the endpoint C-alpha atoms.

This method was of interest for at least two reasons. First, it focused the conformational changes to atoms close together in Euclidian space. This is a large advantage from a computational point of view because it means each change is quick to perform and there are no structural consequences in distant regions. Second, it was shown in the original Backrub paper [8] that the move facilitated placement of alternative side chain positions, a property that could prove advantageous during optimization of different amino acids during computational design.

Motivation for the papers in Chapters 2 and 3

The question I pursued in my first publication (Chapter 2) is 'what model of flexibility is necessary to best predict the amplitude and diversity of side chain flexibility?' I tested three models: (i) side chain flexibility near the native rotamer (i.e. the native side chain conformation) with no backbone flexibility, (ii) side chain flexibility in multiple rotamers with no backbone flexibility, and (iii) side chain flexibility in multiple rotamers with backbone variation introduced through application of the Backrub method. I tested these different models of motion against NMR relaxation methyl side-chain order parameters, which measure the amplitude and diversity of side chain motions on a timescale from picoseconds to nanoseconds. The dataset used was 530 methyl groups in 17 proteins, providing a broad description of many different folds of proteins and packing environments.

In the second manuscript (Chapter 3), I focused on improving modeling of backbone conformational variability. Residual dipolar coupling (RDC) experiments have recently been applied to proteins and provide a uniquely detailed window into the orientation and motion of peptide planes on a timescale ranging from picoseconds to milliseconds. I focused on ubiquitin as a model for study because of the amount of data: 23 high-quality datasets of RDC measurements.

This study first proposes and attempts to answer the following hypothesis: conformational changes inspired by observations from X-ray structures can sample the solution state dynamics of a protein. I test this question in two steps; first by generating structures using Backrub conformational sampling and seeing whether ensembles of these structures can match the RDCs. Second, by comparing the pattern of conformational variability of the Backrub-generated ensembles with the pattern of conformational variability of an ensemble of X-ray structures of ubiquitin.

The study in Chapter 3 also pursues a hypothesis motivated by the success in Chapter 2 of using Backrub ensembles to improve modeling of side chain order parameters and motivated by the successful use of Backrub sampling to improve prediction of mutant side chain conformations. This second hypothesis is whether there exists a link between the conformational diversity of the dynamics of a single protein and the dynamics of its natural family. I test this in two steps: First, I compare the variability in a Backrub ensemble fit to the RDC data (used as a proxy for the dynamics of a single sequence) to the conformational diversity of an ensemble of 20 structures from different members of the UBQ subfamily. Second, I use the sequences accessible to a structure or structural ensemble as a proxy for its conformational properties. I compare the sequences compatible with the RDC-fit Backrub ensembles with the sequences in the natural UBQ subfamily, looking to see if the Backrub ensembles fit to RDCs are compatible with the natural subfamily sequence diversity.

The following chapters detail my work investigating the conformational variability of proteins. My aims in these studies were both towards a more complete understanding of protein flexibility and its role in protein function, and the application of these insights to improve modeling of macromolecules. These aims are intertwined, and I urge the reader to keep them in mind.

Chapter 2: A simple model of backbone flexibility improves modeling of side-chain conformational variability

This chapter has been previously published. [9]

ABSTRACT

The considerable flexibility of side-chains in folded proteins is important for protein stability and function, and may have a role in mediating allosteric interactions. While sampling side-chain degrees of freedom has been an integral part of several successful computational protein design methods, the predictions of these approaches have not been directly compared to experimental measurements of side-chain motional amplitudes. In addition, protein design methods frequently keep the backbone fixed, an approximation that may substantially limit the ability to accurately model side-chain flexibility. Here, we describe a Monte Carlo approach to modeling side chain conformational variability and validate our method against a large dataset of methyl relaxation order parameters derived from nuclear magnetic resonance (NMR) experiments (17 proteins and a total of 530 data points). We also evaluate a model of backbone flexibility based on Backrub motions, a type of conformational change frequently observed in ultra-high-resolution X-ray structures that accounts for correlated side chain backbone movements. The fixed-backbone model performs reasonably well with an overall rmsd between computed and predicted side-chain order parameters of 0.26. Notably, including backbone flexibility leads to significant improvements in modeling side-chain order parameters for ten of the 17 proteins in the set. Greater accuracy of the flexible backbone model results from both increases and decreases in side-chain flexibility relative to the fixed-backbone model. This simple flexible-backbone model should be useful for a variety of protein design applications, including improved modeling of protein–protein interactions, design of proteins with desired flexibility or rigidity, and prediction of correlated motions within proteins.

INTRODUCTION

As suggested by Frauenfelder many years ago it is becoming increasingly recognized that representing the "native" state of a protein using a single conformation, while useful for the analysis of many protein properties, is a substantial simplification [10]. A more realistic but also more complex description views proteins as conformational ensembles in both the unfolded [11; 12; 13] and folded states [14]. In particular, the ability of side-chains to adopt several conformations in non-surface positions in folded proteins has received recent attention [14; 15; 16] and it has long been known that aromatic residues are mobile in protein cores [17]. As a consequence of the development of new experimental techniques to characterize side-chain conformational variability, such as nuclear magnetic resonance (NMR) methyl spin relaxation experiments, considerable amounts of data are now available for different types of methyl-group containing side-chains [16].

Interpretation of side-chain methyl relaxation experiments has led to the suggestion that changes in side-chain conformational entropy can contribute substantially to the free energy of binding [18; 19]. More accurate modeling of side-chain conformational flexibility may also be important for structure-based drug design, when a target protein changes its binding site in response to binding different small molecules [20]. Work by Ranganathan and others [21; 22; 23; 24; 25; 26; 27] has provided intriguing evidence for the existence of "communication pathways" in proteins to facilitate the transmission of signals between allosteric and active sites. NMR experiments on several systems suggest that side-chains may play an important role in mediating this conformational coupling [28].

Given the importance of side-chain conformational variability in binding and allostery, modeling this flexibility may lead to considerable improvements in the characterization and design of functional proteins and protein interactions. Several representations of side-chain flexibility incorporating multiple higher-energy conformations have been used in prediction and design[29; 30; 31; 32]. However, the predictions resulting from models commonly used in protein design simulations have not been directly compared to experimental data on the amplitude of side-chain motion. Moreover, even when side-chain flexibility has been explicitly considered, the protein backbone is generally held static in the crystal structure conformation, an approximation that likely leads to inaccuracies modeling side-chain conformational freedom.

In contrast, molecular dynamics (MD) simulations model backbone conformational changes and have been compared to side-chain dynamics data, yielding reasonable agreement with measured side-chain order parameters for several proteins [30; 33; 34; 35]. However, it was noted that estimating order parameters from MD trajectories is difficult for side-chains that make few rotameric transitions during the simulation, although sampling was improved using replica exchange MD [33]. While these MD-based methods achieve considerable accuracy, they are generally computationally prohibitive for use during protein design, which seeks to simultaneously search in sequence and structure space for low energy amino acid combinations and conformations.

In this paper, we describe a Monte Carlo-based approach to model side-chain conformational variability in protein design simulations and validate our method on a dataset of 17 proteins with 530 methyl relaxation order parameters. We find that motions within the native rotamer well are not sufficient to explain the range of experimentally observed side-chain relaxation order parameters. While a multiple-rotamer Monte Carlo model of side-chain conformational variability performs reasonably well for some proteins (with correlation coefficients between calculated and experimental order parameters above 0.6 for 8 out of the 17 proteins in the dataset), prediction accuracy is limited by the fixed backbone approximation and the use of an implicit solvent model. Using the same dataset, we also evaluate a simple representation of backbone flexibility that allows correlated backbone and side-chain motions inspired by conformational changes observed in ultra-high resolution crystal structures [8]. Notably, we find that this simple model of correlated backbone-side-chain ("Backrub" [8]) motions leads to significant overall improvements in modeling side-chain order parameters in our dataset. Incorporating backbone flexibility using Backrub simulations lowers the rmsd between computed and predicted side-chain order parameters for 10 of the 17 proteins, has no significant effect on the rmsd for 5 of the other proteins, and increases the rmsd for 2 proteins. Our results suggest that this flexible backbone protocol is a useful method to sample near-native conformational space. This approach could have substantial impact on many applications, including the computational design of proteins with new functions requiring flexibility.

RESULTS AND DISCUSSION

Rationale and Computational Strategy

We aimed to develop and assess a simple model for fast timescale protein side-chain flexibility that can be applied efficiently in protein structure prediction and design simulations. The major sources of experimental data for model evaluation were methyl relaxation side-chain order parameters. These measurements capture the amplitude of side-chain motions on the picosecond to nanosecond timescale and range from 0 (flexible) to 1 (rigid). Our dataset of 17 proteins containing 530 experimentally characterized methyl groups is shown in **Table 2-1**. We tested three different models with increasing complexity (**Figure 2-1a**): (1) The first model evaluates the extent to which simulations only allowing side-chain motions within the native rotamer well recapitulate experimentally measured side-chain order parameters. (2) The second model allows more extensive side-chain flexibility by using Metropolis Monte Carlo simulations to sample side-chain conformations from multiple rotameric states. (3) The third model tests the effect of including backbone flexibility by calculating side-chain order parameters from side-chain Monte Carlo simulations performed over an ensemble of backbone conformations. These ensembles were generated using backbone motions inspired by conformational variability observed in ultra-high resolution protein structures (**Figure 2-1b** and **c**) [8]. All 3 models sample side-chains by varying chi dihedral angles while side-chain bond lengths and bond angles are unchanged (see Methods).

Table 2-1. The dataset of proteins modeled in this paper.

	# methyls	PDB	Chain	Minimized	Ligand	Class
a3D [a]	15	2A3D	_	y	n	Alpha
albp [b]	28	1LIB	A	n	n	Beta
calmodulin [c]	36	1AHR	A	n	y	Alpha
Cdc42hs	47	1AN0	A	n	y	Alpha/Beta
cytochrome-c2-holo [d]	31	1C2R	A	n	y	Alpha
Eglinc	17	1CSE	I	n	n	Alpha/Beta
flavodoxin-holo [e]	56	1OBO	A	n	y	Alpha/Beta
FNfn10 [f]	35	1FNA	_	n	n	Beta
fyn-sh3 [g]	12	1SHF	A	n	n	Beta
gb1 [h]	10	1PGB	_	n	n	Alpha/Beta
hiv-protease-holo [i]	56	1QBS	dimer	n	y	Beta
mfabp [j]	42	1HMT	_	n	n	Beta
NTnC [k]	26	1AVS	A	n	n	Alpha
plcc-sh2-free [l]	27	2PLD	A	y	n	Alpha/Beta
proteinL [m]	20	1HZ6	A	n	n	Alpha/Beta
TNfn3 [n]	40	1TEN	A	n	n	Beta
ubiquitin	32	1UBQ	A	n	n	Alpha/Beta
Totals	530					

[a] a3D: the *de novo* designed three-helix bundle α3D

[b] alpb, adipocyte lipid-binding protein

[c] calmodulin: Ca^{2+}-loaded calmodulin

[d] cytochrome-c2-holo: cytochrome c2 bound to its heme prosthetic group

[e] flavodoxin-holo: flavodoxin bound to flavin mononucleotide

[f] FNfn10: the tenth fibronectin type III domain from human fibronectin

[g] fyn-sh3: the SH3 domain in human Fyn

[h] gb1: the B1 domain from protein G

[i] hiv-protease-holo: HIV-1 protease homodimer bound to the inhibitor DMP323

[j] mfabp: muscle fatty acid-binding protein

[k] NTnC: the N-terminal domain of chicken skeletal troponin C

[l] plcc-sh2: an SH2 domain of phospholipase C-gamma1

[m] proteinL: the B1 domain of protein L

[n] TNfn3: the third fibronectin type III domain from human tenascin

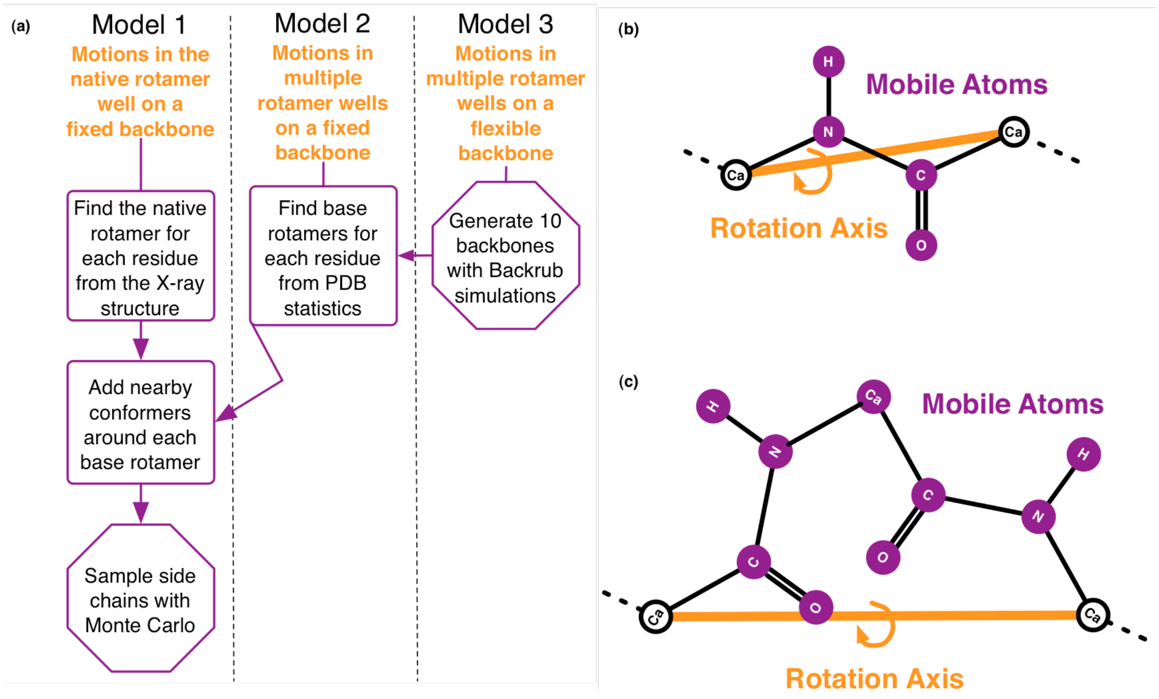

(a)
Model 1
Motions in the native rotamer well on a fixed backbone
Model 2
Motions in multiple rotamer wells on a fixed backbone
Model 3
Motions in multiple rotamer wells on a flexible backbone
Find the native rotamer for each residue from the X-ray structure
Find base rotamers for each residue from PDB statistics
Generate 10 backbones with Backrub simulations
Add nearby conformers around each base rotamer
Sample side chains with Monte Carlo
(b)
H
Mobile Atoms
N
Ca
C
Ca
Rotation Axis
O
(c)
H
Cα
Mobile Atoms
N
C
H
N
O
C
O
Cα
Cα
Rotation Axis

Figure 2-1. Computational strategy and motional models.

(a) Flowchart of the methods used for the 3 models of motion. Schematic of (b) dipeptide and (c) tripeptide “Backrub” conformational changes used to model backbone changes in Model 3. The Backrub motion consists of a rigid body rotation of all atoms between two 2 C atoms, about the axis connecting the C atoms. This rotation is followed by optimization of bond angles involving the endpoint C atoms (see Methods).

Side-chain Monte Carlo simulations allowing motions only around the native rotamer (Model 1)

We first compared experimental (S^2_{exp}) and computed (S^2_{calc}) side-chain order parameters using the simplest approximation of side-chain conformational variability, where motions are only allowed within the rotamer well of the side-chain conformation observed in the crystal structure (Model 1, see Figure 2-1a for a schematic and Methods for details). Figure 2-2 shows that Model 1 fails to recapitulate the range of order parameters observed experimentally. (See Table 2-S1 for statistics on all proteins in the data set.) The methyl groups mostly have high values for S^2_{calc}, whereas S^2_{exp} values cover a larger range with both high and low values. The finding that native rotamer side-chain motions do not represent the range of experimentally observed side-chain motions also holds when sampling on multiple backbones but still restricting side-chain motions to be within one rotameric state (referred to as “Model 1*” in Figure 2-2; white boxes), sampling conformers in wider rotamer wells (3 standard deviations around the Dunbrack mean chi 1 and 2 angles) or increasing the leniency towards steric clashes by ‘softening’ the Lennard-Jones repulsive term in the energy function (see Methods; data not shown). This result agrees with previous work calculating side-chain order parameters from MD methods [14; 30; 33] and from a toy model of side-chain motion [30]. Thus, rotamer transitions, included in Models 2 and 3 below, are necessary to explain measured side-chain order parameters, even for many buried residues.

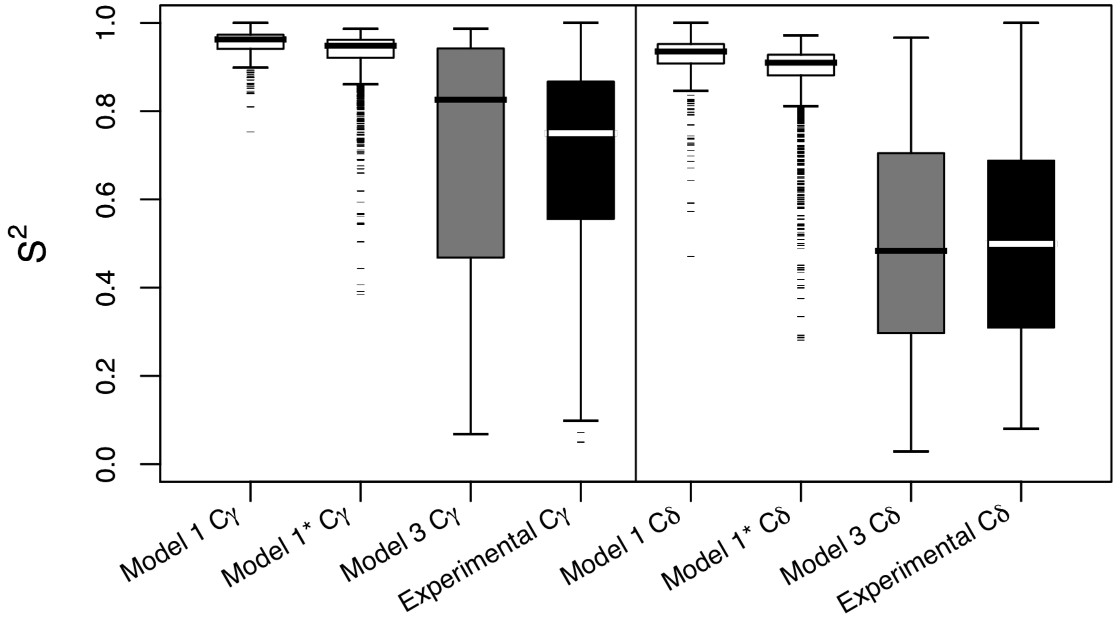


Figure 2-2. Side-chain motions within the native rotamer well do not sample the conformational flexibility observed in methyl relaxation experiments.

Shown are boxplots representing the distributions of order parameters for C and C methyl groups from different models and from the experimental measurements. White boxes: native rotamer motions on a fixed backbone (Model 1) or on an ensemble of backbones (generated using Backrub Monte Carlo simulations that kept the side-chains in their native rotamer well, Model 1*). Grey boxes: results from Model 3 simulations, using an ensemble of backbone conformations and allowing multiple rotameric states. Black boxes: experimental relaxation measurements. The boxes represent the middle 25-75% of the values; the horizontal bar inside the box is the median value; the "whiskers" extending out of the box cover the ~1.5 times the range of the box (or up to the furthest data point); and dashes outside of the whiskers represent outliers.

Side-chain Monte Carlo simulations on a fixed-backbone allowing rotamer transitions (Model 2)

Table 2-2a summarizes the results of side-chain Monte Carlo simulations employing Model 2. As in Model 1, the simulations were carried out on a fixed backbone, but side-chains were allowed to change rotameric conformations during the simulations. The results for each protein are evaluated by the correlation coefficient (r) between experimental (S^2_{exp}) and calculated (S^2_{calc}) side-chain order parameters, and the root mean squared deviation (rmsd) between them. Additionally, we measured how often we correctly model the qualitative rigidity or flexibility of a side-chain dihedral angle. The results from Model 1 (**Figure 2-2**) and the work of others [14; 30; 33] provide a useful distinction between "rigid" and "flexible" side-chain dihedrals, as they indicate that methyl groups with order parameters above 0.7-0.8 are likely to sample a single rotameric well, and methyl groups with order parameters below this threshold are likely to switch between multiple rotameric states. When sampling within the native rotamer well on a fixed backbone, 95% of methyl groups have S^2_{calc} values above 0.9 and 0.8 for chi 1 and chi 2, respectively (**Figure 2-2**). When using an ensemble with backbone variations and no rotamer transitions, the respective S^2_{calc} values are 0.85 and 0.7 for chi 1 and chi 2. Thus, for the purposes of this study, we chose a cutoff value of S^2=0.75 to separate "rigid" and "flexible" methyl groups.

We evaluated several parameters that determine which "conformers" (defined here as side-chain conformational microstates, with rotamers being macrostates containing many conformers within a dihedral energy basin, see Methods) are included in the side-chain Monte Carlo simulations: the number of base rotamers to use for each residue (the size of the base rotamer library), the largest chi angle distance between a conformer and its base rotamer (the rotamer well width), and the chi angle degree increment between adjacent conformers (the conformer resolution). The results described in **Table 2-2a** use the “large” base rotamer library, a rotamer well width of 1.5 times the standard deviation taken from the Dunbrack rotamer library (see Methods) for that base rotamer, and a conformer resolution of 10 degrees for chi 1 and chi 2.

A standard test for the accuracy of side-chain sampling is whether a given method correctly predicts the side-chain conformations (usually within 40 degrees) observed in the crystal structure. To evaluate our conformer library, we used it in repacking simulations of all side-chains in a published test set of 65 high-resolution crystal structures [36] (see Methods for details.) Out of the residues in this set, 87% had correctly assigned chi 1 rotamers and 76% had correctly assigned chi 1 and chi 2 rotamers. These values are similar to repacking results on the same dataset from several other studies [37; 38] and somewhat lower than a study using a library of nearly 50,000 conformers [36].

Table 2-2. Summary of (a) Model 2 and (b) Model 3 results for all proteins.

				(a) Model 2					(b) Model 3					
	# of methyls	# Rigid (exp)	# Flexible (exp)	r	slope	rmsd	Fraction Correct Rigid (calc)	Fraction Correct Flexible (calc)	r	slope	rmsd	Fraction Correct Rigid (calc)	Fraction Correct Flexible (calc)	Mean Cα RMSD of backbone ensembles [a]
a3D	15	0	15	0.64	1.24	0.22	N/A	0.73	0.74	1.35	0.17	N/A	0.86	0.30
albp	28	13	15	0.79	0.98	0.19	0.77	0.73	0.76	0.93	0.19	0.66	0.79	0.19
calmodulin	36	4	32	0.81	1.22	0.21	0.75	0.69	0.78	1.14	0.21	0.78	0.69	0.27
cdc42hs	47	23	24	0.31	0.33	0.33	0.61	0.58	0.41	0.41	0.31	0.46	0.80	0.24
cytochrome-c2-holo	31	21	10	0.37	0.35	0.24	0.67	0.50	0.52	0.51	0.22	0.56	0.83	0.22
eglinc	17	6	11	0.55	0.76	0.25	0.67	0.82	0.59	0.82	0.25	0.83	0.78	0.22
flavodoxin-holo	56	36	20	0.63	0.74	0.21	0.83	0.85	0.62	0.74	0.21	0.79	0.85	0.06
FNfn10	35	12	23	0.40	0.57	0.29	0.67	0.65	0.42	0.49	0.24	0.55	0.66	0.19
fyn-sh3	12	4	8	0.89	1.45	0.15	0.75	0.88	0.62	0.88	0.20	0.55	0.99	0.22
gb1	10	0	10	0.49	0.64	0.26	N/A	0.80	0.68	0.92	0.22	N/A	0.77	0.24
hiv-protease-holo	56	36	20	0.73	0.86	0.22	0.69	0.95	0.62	0.68	0.23	0.57	0.92	0.19
mfabp	42	20	22	0.59	0.69	0.27	0.55	0.82	0.60	0.67	0.26	0.56	0.85	0.15
NTnC	26	2	24	0.77	1.13	0.23	1.00	0.58	0.73	1.05	0.22	1.00	0.59	0.24
plcc-sh2-free	27	3	24	0.57	0.78	0.26	0.33	0.67	0.60	0.74	0.24	0.36	0.84	0.26
protl	20	8	12	-0.07	-0.10	0.35	0.62	0.50	-0.06	-0.09	0.33	0.56	0.50	0.11
TNfn3	40	11	29	0.34	0.40	0.32	0.64	0.55	0.38	0.45	0.31	0.62	0.57	0.12
ubiquitin	32	16	16	0.60	0.82	0.27	0.75	0.75	0.64	0.81	0.24	0.74	0.83	0.22
Total	530	215	315	.	.	0.26 [b]	0.69	0.70			0.25 [b]	0.62	0.76	

The values are: the correlation coefficient and slope for the linear fit between S^2_{exp} and S^2_{calc}; the rmsd between S^2_{exp} and S^2_{calc}; the number of rigid (S^2_{exp} >0.75) and flexible (S^2_{exp}<=0.75) methyl groups based on the experimental order parameters; the fraction of rigid and flexible methyl groups correctly identified as rigid or flexible from the simulations.

[a] All backbone ensembles had standard deviations of these values less than 0.05.

[b] Values calculated by considering all 530 data points together.

Results of Model 2

Introducing rotamer transitions into the simulations results in order parameters spanning the range of the experimental values. For all 17 proteins (**Table 2-2a**), allowing rotamer flips substantially improves the correlation coefficient (r) and rmsd between S^2_{exp} and S^2_{calc} (compare **Table 2-2a** and **Table 2-S1**). (Results for the 439 nonpolar residues, excluding threonines, are given in parentheses; reasons for analyzing threonine residues separately are discussed below; see **Table 2-S2a.**) Out of the 17 proteins in the set, 5(6) have **r** >=0.7 and 8(11) have **r** >=0.6, indicating that Model 2 is a reasonable model of fast-timescale side-chain motion in some proteins. The rmsd over the whole dataset is 0.26, 69% of the 215 rigid methyl groups were correctly modeled as rigid, and 70% of the 315 flexible methyl groups were correctly modeled as flexible. Thus Model 2 had similar success modeling rigid and flexible methyl groups.

Sensitivity analysis

We next tested how the model performance was affected by changes in the strength of the Lennard-Jones repulsive term, the base rotamer library size, the rotamer well width, and the conformer resolution (see Methods for details).

Reduced Lennard-Jones repulsive term

A number of flexible methyl groups were incorrectly modeled as rigid in our Model 2 simulations. Since Model 2 does not allow motion along the backbone degrees of freedom, we tested whether reducing the Lennard-Jones (LJ) repulsive term would help in the cases where a small backbone shift would allow the side-chain to become flexible. We evaluated two different ways of reducing the strength of the LJ repulsive term: in the first we scaled the LJ radii down by 0.95 (the “small radii” LJ repulsive term), and in the second we decreased the slope of the LJ repulsive term (the “soft repulsive” LJ term; see Methods for details on both).

The results in **Table 2-3a** illustrate that there is a clear tradeoff between correctly modeled rigid and correct modeled flexible residues. The original LJ repulsive term gives mostly equivalent modeled fractions of rigid and flexible residues. With the “smaller radii” and the “soft repulsive” LJ terms, the balance shifts in favor of flexible residues, with more methyl groups correctly modeled as flexible and many rigid methyl groups incorrectly modeled as flexible. Thus, these changes to the LJ repulsive term seem to modulate the flexibility or rigidity across all residues in the set, but do not have the environmental specificity needed to increase accuracy.

Base rotamer selection:

The size of the base rotamer library determines the number of rotameric states accessible to the protein’s side-chains. One strategy is to choose a library size that is small and hence allows fast sampling. This approach is useful especially for applications such as *ab initio* structure prediction where large numbers of conformations are generated. Since we are sampling high-energy states rather than trying to find the lowest energy conformation, our strategy here was instead to generate a larger rotamer library including many rotamers that have low but non-negligible probability. As shown in **Table 2-3b**, using the large rotamer set has no effect on the rmsd over the data set, but balances prediction of flexible and rigid methyl groups, whereas the default rotamer set modeled rigid residues better than flexible residues. For the flexible *de novo* designed protein α3D, this improvement in rmsd is substantial (0.35 to 0.22; **Table 2-S3**), as it is for gb1 (0.36 to 0.26). Therefore, the increase in rotamer library size appears useful for modeling some proteins.

Table 2-3. Effect of parameter values of Model 2 on the rmsd and fraction of correctly modeled rigid or flexible methyl groups.

		rmsd	Fraction Correct Rigid (calc)	Fraction Correct Flexible (calc)	Fraction Correct Total (calc)
(a) LJ Repulsive	hard repulsive [a]	0.26	0.69	0.70	0.70
	small radii	0.25	0.60	0.80	0.72
	soft repulsive	0.28	0.36	0.94	0.70
(b) Base Rotamer Library	default	0.26	0.73	0.64	0.68
	large [a]	0.26	0.69	0.70	0.70
(c) Rotamer Well Width	0.5 sd [b]	0.30	0.73	0.58	0.64
	1.0 sd [b]	0.27	0.73	0.64	0.67
	1.5 sd [a c]	0.26	0.69	0.70	0.70
	2.0 sd [c]	0.26	0.69	0.71	0.70
	3.0 sd [c]	0.26	0.69	0.70	0.70
(d) Conformer Resolution	10 degrees [a]	0.26	0.69	0.70	0.70
	15 degrees	0.26	0.68	0.71	0.70
	20 degrees	0.27	0.67	0.68	0.67
	25 degrees	0.27	0.69	0.64	0.66
	1 rotamer	0.33	0.80	0.47	0.60

See the legend of Table 2-2 for a description of column headers.
[a] Indicates parameter values used for the results in Table 2-2
[b] Sampled with 5 degree conformer resolution
[c] Sampled with 10 degree conformer resolution

Width of rotamer wells:

Increasing the width of the rotamer wells has a strong effect on the modeling accuracy (**Table 2-3c**). Sampling in wells with widths of 0.5 or 1.0 standard deviations (sds) around the Dunbrack rotamer chi angles shows improvements in rmsd and prediction of rigidity/flexibility. This trend continues, although weaker, up to 1.5 and 2 sds. There is a very large drop in rmsd for α3D from 1 to 1.5 sds, again highlighting the flexibility of this protein (**Table 2-S3**).

Conformer resolution:

There is not much difference in the performance at 10- and 15-degree conformer resolutions (**Table 2-3d**); however, when using the 20- and 25-degrees conformer resolutions the rmsd increases and the fraction of correct flexible methyl groups drops. As expected, excluding all conformers except for the base rotamers performs poorly overall. The single exception is fyn-sh3, which performs well even with no added conformers per base rotamer (**Table 2-S3**).

Examples of good and bad predictions using Model 2

Figure 2-3 depicts several examples of Model 2 results. Two of the proteins in our dataset, albp and fyn-sh3 (**Figure 2-3a** and **3b)**, are modeled very well with **r** of 0.79 and 0.89, and rmsds below 0.2 (**Table 2-2a**). We correctly classify the flexibility/rigidity of 10 of the 12 methyl groups in fyn-sh3; of the 28 methyl groups in albp, we correctly classify 77% of its rigid and 73% of its flexible residues. Flavodoxin-holo and ubiquitin (**Figure 2-3c** and **3d**) are also modeled well, at least qualitatively: their correlation coefficients are lower (0.63 and 0.6), but 83% and 75% of the methyl groups are correctly classified as rigid/flexible for flavodoxin-holo and ubiquitin, respectively.

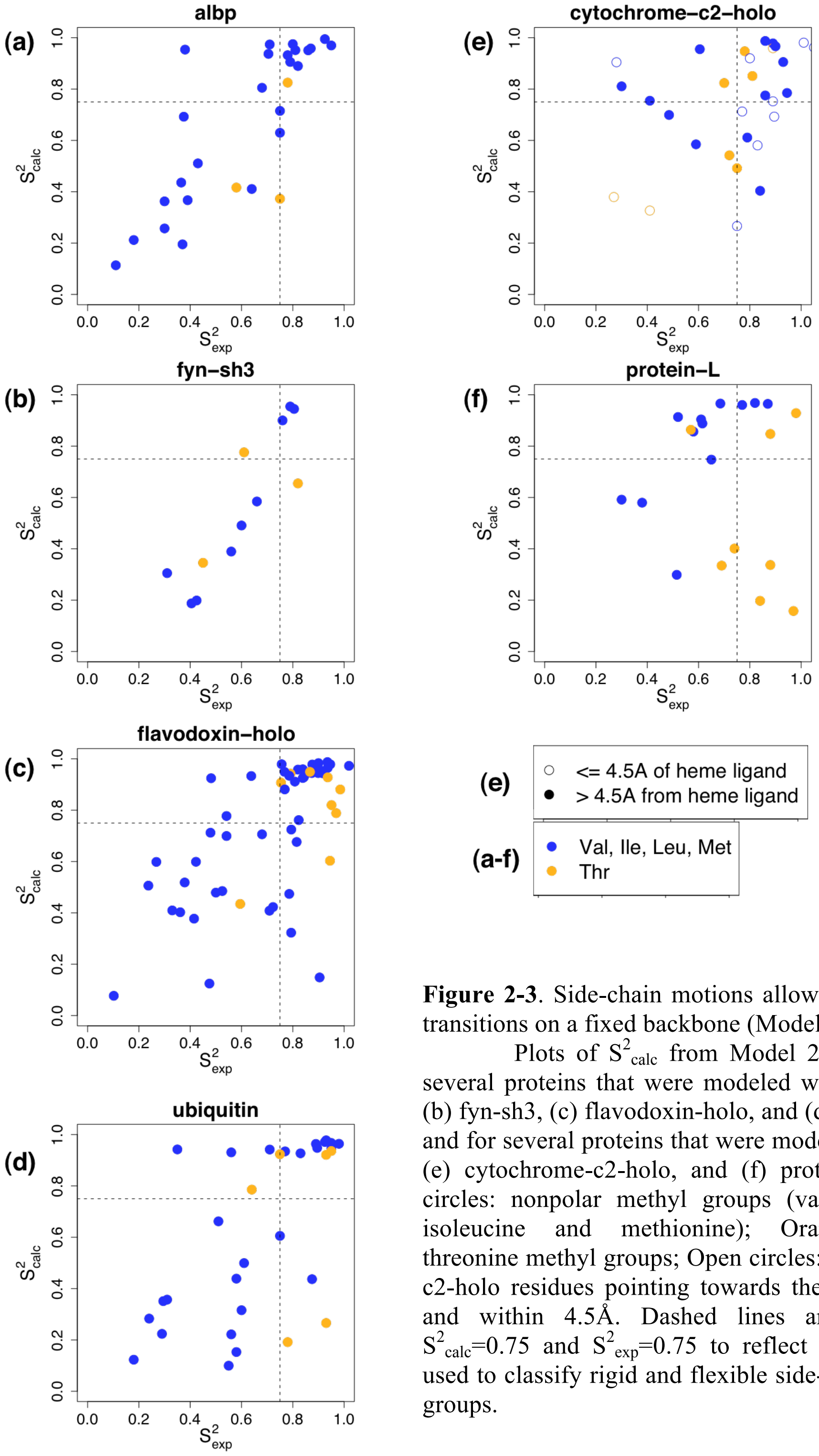


Figure 2-3. Side-chain motions allowing rotameric transitions on a fixed backbone (Model 2).

Plots of S^2_{calc} from Model 2 vs. S^2_{exp} for several proteins that were modeled well—(a) albp, (b) fyn-sh3, (c) flavodoxin-holo, and (d) ubiquitin—and for several proteins that were modeled poorly—(e) cytochrome-c2-holo, and (f) protein-L. Blue circles: nonpolar methyl groups (valine, leucine, isoleucine and methionine); Orange circles: threonine methyl groups; Open circles: cytochrome-c2-holo residues pointing towards the heme group and within 4.5Å. Dashed lines are drawn at S^2_{calc}=0.75 and S^2_{exp}=0.75 to reflect the threshold used to classify rigid and flexible side-chain methyl groups.

For ubiquitin, this leaves 5 nonpolar methyl group outliers: 3 are modeled to be too rigid (I61 Cδ, L50 Cδ, and V70 Cγ) and 2 are modeled to be too flexible (I3 Cδ and I44 Cγ). I61 and I50 are both located on the loop between strands β3 and β4, the longest loop in the protein (13 residues), and V70 is located near the C-terminus. The errors modeling these residues could be the result of backbone flexibility that is not taken into account in Model 2.

Examples of proteins that did not perform well with Model 2 are also shown in **Figure 2-3**. For cytochrome c2 (**Figure 2-3e**), the low fraction of correctly modeled flexible and rigid methyl groups (0.67 and 0.5, respectively) may be related to keeping the large buried heme prosthetic group rigid during the simulations. For example, V107 Cγ, I57 Cδ and L100 Cδ are all less than 4.5Å away from the heme group (**Figure 2-3e)** and have S^2_{calc} (S^2_{exp}) values of 0.28(0.9), 0.27(0.7), and 0.58(0.83). In addition, V114 Cγ, V115 Cγ, I27 Cδ and I20 Cγ are modeled as too rigid. These residues are located near the end of a beta-hairpin or at the C-terminus, which may be flexible in solution, as suggested by the comparatively low backbone amide order parameters of these residues and their neighbors: 0.75 for E26 (which is the residue in register with I20 on the adjacent strand), 0.72 for I27, 0.78 for S113 and 0.81 for V115. In addition, cytochrome c2 has the lowest resolution crystal structure of any protein in the set at 2.5Å.

As illustrated in the examples above, inaccuracies are likely due to a number of simplifications in our model, including the use of a fixed backbone (as discussed above for ubiquitin and cytochrome c) and the approximation of ligand rigidity. Other likely sources of error include the lack of timescale information from the Monte Carlo simulations (flexibility in our model may occur on timescales longer than those reflected in the experimental measurements) and the use of an implicit solvation model, which does not capture effects related to the defined size and properties of water molecules. We expected this latter effect to be most dramatic for solvent-exposed threonine residues, as discussed below.

Slow transitions / Solvent model inaccuracies

Excluding threonines from the correlation and rmsd calculations improved the results significantly for two proteins: ubiquitin and protein L. For protein L (**Figure 2-3f**), excluding the threonine methyl groups (8 of 20 total data points), increased the correlation coefficient from -0.07 to 0.68 and decreased the rmsd from 0.35 to 0.25. These large differences are caused by several surface-exposed threonine side-chains (T5, T19, T25, T39, and T48), which are modeled as flexible but have high S^2_{exp} values (**Figure 2-3f**). A study by Millet et al [39] observed that T19, T25, T39, and T48 have one or two ^3J scalar coupling values inconsistent with a singly populated rotamer (^3J couplings could not be measured for T5 due to signal overlap). Thus, our Monte Carlo simulations may in fact correctly capture the flexibility of these threonines on timescales longer than the picosecond to nanosecond motions reflected in the relaxation order parameters. Millet et al. suggest that these slow rotamer transitions occur because particular backbone conformations change the height of the energy barrier between rotamers, and that these barriers can be altered by relatively modest backbone conformational changes in response to mutation [39].

Alternatively, the slow threonine transitions may be the result of hydrogen bonds to water molecules in the first solvation shell. Inspecting the 5 crystal forms of protein L reveals many potential hydrogen bonds between the side-chain hydroxyl groups of these threonines and nearby water molecules with low temperature factors. As mentioned above, the implicit solvent model used in our simulations does not capture effects due to specific water-mediated hydrogen bonds. Water-mediated hydrogen bonds could restrict the rate of transition between rotameric states for the 5 above-mentioned solvent-exposed threonines in protein L as well as in other proteins, such as ubiquitin (**Figure 2-3c**), which also has several threonine residues near ordered water molecules in the X-ray structure. The idea that missed water interactions could be responsible for modeling inaccuracies is supported by the facts that: (a) predictions for threonine Cγ had the highest rmsd of any methyl type between S^2_{exp} and S^2_{calc} (0.3), and (b) a lower percentage of threonine Cγ methyl groups were correctly modeled as rigid (58%) than either the valine Cγs (75%) or the isoleucine Cγ (91%; data not shown). If the problem modeling threonines is indeed related to the implicit solvent model, it may be ameliorated in future studies by using a "solvated rotamer" approach [40].

Side-chain Monte Carlo simulations on backbone ensembles (Model 3)

We show above that Model 2 is a reasonable approximation capturing the flexibility of side-chains with low methyl relaxation order parameters in some proteins. However, a substantial simplification in Model 2 not present in MD simulations is that the backbone is held fixed. We next asked whether a model of backbone flexibility that is simple enough to be computationally feasible in the context of protein design simulations would improve modeling of side-chain flexibility. For each protein in our set, an ensemble of ten near-native backbone structures was used to represent small backbone variations. The backbones were generated using Backrub Monte Carlo simulations with the Rosetta all-atom scoring function (see **Figure 2-1** and Methods for details), and the resulting structures had Ca rmsds to the crystal structure ranging from 0.01Å to 0.37Å (see **Table 2-2b** for averaged pair-wise Ca rmsds of the ensembles to the crystal structure). This protocol was repeated over ten different ensembles for each protein to estimate the sensitivity to variation in the composition of the ensembles.

We first tested whether inclusion of small backbone variability in this way led to predictions of side-chain order parameter values that were in the experimentally observed range. **Figure 2-2** shows that this is the case, but only when rotameric transitions are allowed in the simulations (compare Model 1*, white boxes, and Model 3, grey boxes, to the experimental data, black boxes). **Table 2-2b** summarizes the results of Model 3 for all methyl groups (results for nonpolar methyl groups only are in parentheses; see **Table 2-S2b**). 4(8) out of the 17 proteins have correlation coefficients between S^2_{exp} and S^2_{calc} >=0.7, and 11(14) proteins have correlation coefficients >=0.6. Although the ability of Model 3 to correctly classify rigid methyl groups is affected (reduced by 7%) by the increased conformational degrees of freedom introduced with the backbone perturbations, Model 3 shows clear improvements over Model 2 with respect to rmsd values.

The boxplots in **Figure 2-4** illustrate the rmsd values resulting from the different backbone ensembles generated for each protein. Including backbone flexibility results in noticeable improvement in the order parameter predictions for 10 out of 17 proteins, with at least 75% of the rmsd values from Model 3 (purple boxes) below the rmsd value for

Model 2 (yellow line). Of the 7 cases where Model 3 does not improve the rmsd relative to Model 2, five proteins have essentially identical results (albp, calmodulin, eglin c, flavodoxin-holo, and NTnC). Of the two cases that worsen substantially (hiv-protease-holo and fyn-sh3), one (fyn-sh3) already performed very well under Model 2, with a correlation coefficient of 0.89 and an rmsd of 0.15.

To assess whether the improvements with Model 3 described above are significant, we performed several statistical tests (see Methods). Tests for the 17 individual proteins confirmed that the results depicted in **Figure 2-4** are statistically significant for 12 proteins, leading to improved rmsds with Model 3 over Model 2 in 10 cases (p-values < 0.005 from the Student's t-test and < 0.01 from the Wilcoxon signed-rank test; asterisks in **Figure 2-4**) and a decrease in agreement with experimental data for 2 proteins (same criteria as above, pound signs in **Figure 2-4**). We also evaluated whether Model 3 errors (defined as the magnitude of the difference between S^2_{calc} and S^2_{exp} for each methyl group) were significantly less than the Model 2 errors when considering results for all 530 methyl groups together. Using the paired Wilcoxon signed-rank test and the paired Student's t-test, we found that the Model 3 errors were indeed smaller than the Model 2 errors with p-values of 8*10-6 and 0.003, respectively. Thus, the overall improvement of Model 3 over the dataset suggests that the ensembles contain relevant conformations that may be populated in the solution experiments at the timescale of interest.

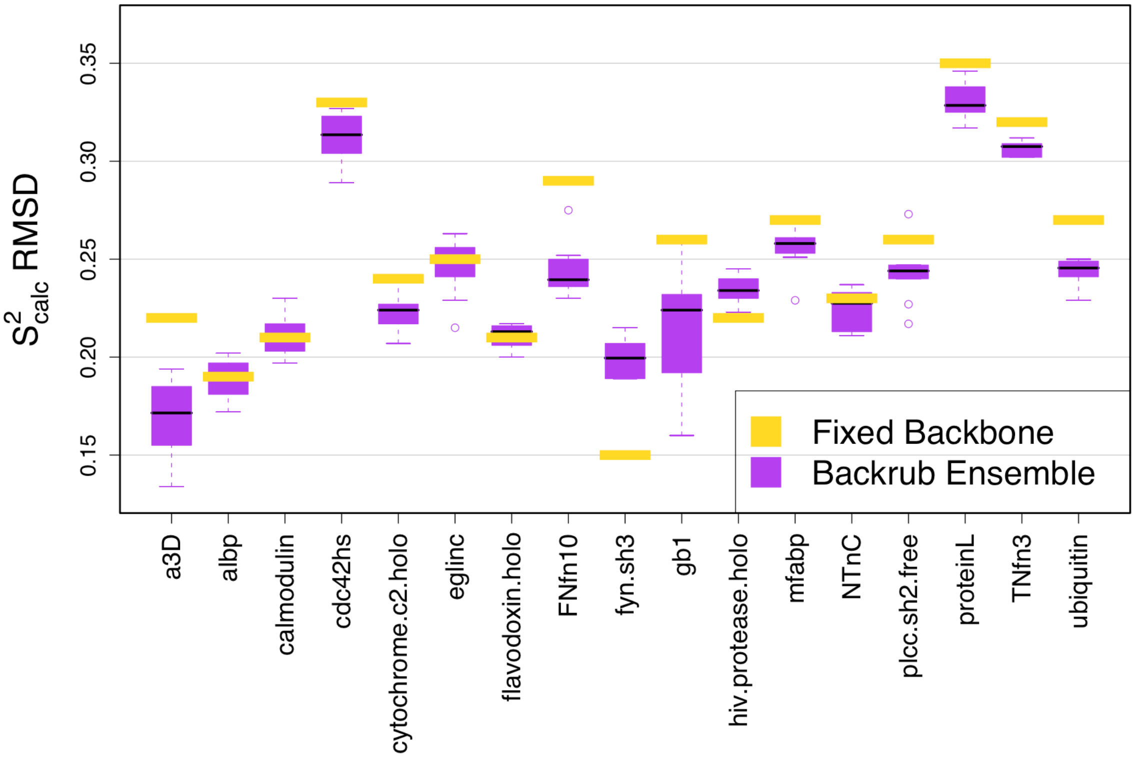


Figure 2-4. A simple model of backbone conformational variability (Model 3) improves modeling of side-chain motions.

Rmsd between experimental order parameters (S^2_{exp}) and S^2_{calc} from Model 2 (yellow lines) and Model 3 (purple boxes). *: the 10 proteins for which the Model 3 ensemble rmsds are significantly lower than the Model 2 rmsd (Student's t-test p-value < 0.005 and Wilcoxon-signed rank test p-value < 0.01). #: the 2 proteins for which the Model 2 rmsd is lower than the Model 3 rmsds (by the same measure). See **Figure 2-2** for an explanation of boxplots. Proteins are depicted in order of increasing rmsd between experimental and computed order parameters from Model 2.

An interesting property of the experimental order parameter measurements is that they suggest the existence of many flexible side-chains in buried positions. We were curious whether our models would be able to capture this core flexibility. On a fixed backbone, buried methyl groups were correctly modeled as flexible in 63% of the cases. With a flexible-backbone this value increased slightly to 66%.

A specific example of the improvements seen for Model 3 over Model 2 is shown in **Figure 2-5** for the B1 domain of protein G (gb1). Fixed-backbone simulations (Model 2) of gb1 yield a correlation coefficient of 0.49 and an rmsd of 0.26. In contrast, incorporating backbone flexibility (Model 3) increases the correlation coefficient to 0.68 and decreases the rmsd (calculated over combined simulations using all 10 ensembles of size 10) to 0.22. As illustrated in **Figure 2-5**, improvements from Model 3 result both from methyl groups becoming more flexible when incorporating backbone flexibility (L7 Cδ and V21 Cγ show approximate S^2_{calc} decreases of -0.25 and -0.1, respectively) and from methyl groups becoming more rigid (residues V29 Cγ and V54 Cγ have S^2_{calc} increases of 0.4 and 0.1, respectively).

To rationalize the observed changes in side-chain dynamics with Model 3, we structurally analyzed the lowest rmsd ensemble (#6 out of 10) of gb1. Intuitively, backbone moves are expected to increase the freedom of protein regions to sample conformational space and hence lead to lower order parameters. Examples consistent with this behavior are V21 Cγ and L7 Cδ. V21 is located in a turn region between a beta sheet and an alpha helix. The averaged Cα rmsd in the ensemble (relative to the crystal structure 1PGB) at this position is 0.61Å (with a standard deviation of 0.26Å) and the largest Ca rmsd to the crystal structure for an individual backbone conformation is 1.23Å.

L7 is located on an extended beta strand, and in the ensemble its Cα and Cβ atoms move up to 0.35Å and 0.49Å from their crystal structure positions, respectively. However, the slight rotation about the backbone resulting from Backrub moves causes larger Cartesian coordinate changes at the Cδ atoms (**Figure 2-6c**) and allows L7 to more extensively sample the chi 2 dihedral degree of freedom. As a result, the chi 2=180 degrees rotamer that was infrequently visited in the fixed-backbone simulations (**Figure 2-6a**) has a much higher population in the flexible-backbone simulations (**Figure 2-6b**), resulting in a lower order parameter that is closer to the experimental value (**Figure 2-5**).

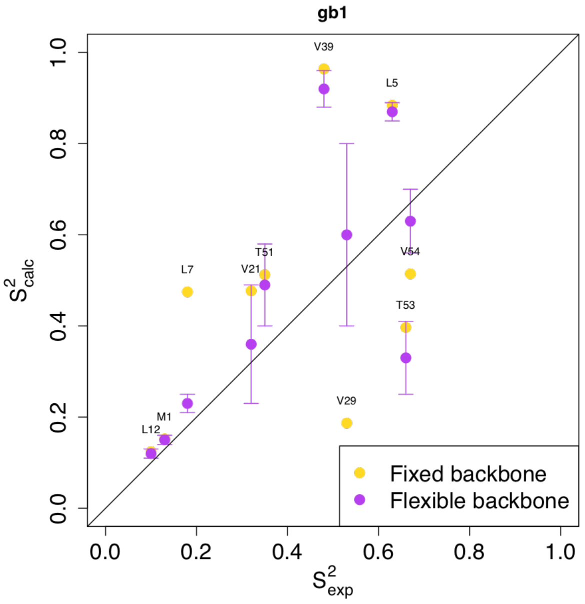


Figure 2-5. Improvement of S^2_{calc} values for Model 3 over Model 2 showing both increased and decreased modeled flexibility of residues in the protein gb1.

S^2_{exp} vs. S^2_{calc} from Model 2 (yellow) and Model 3 (purple). Error bars for Model 3 are the standard deviations of S^2_{calc} over the 10 ensembles.

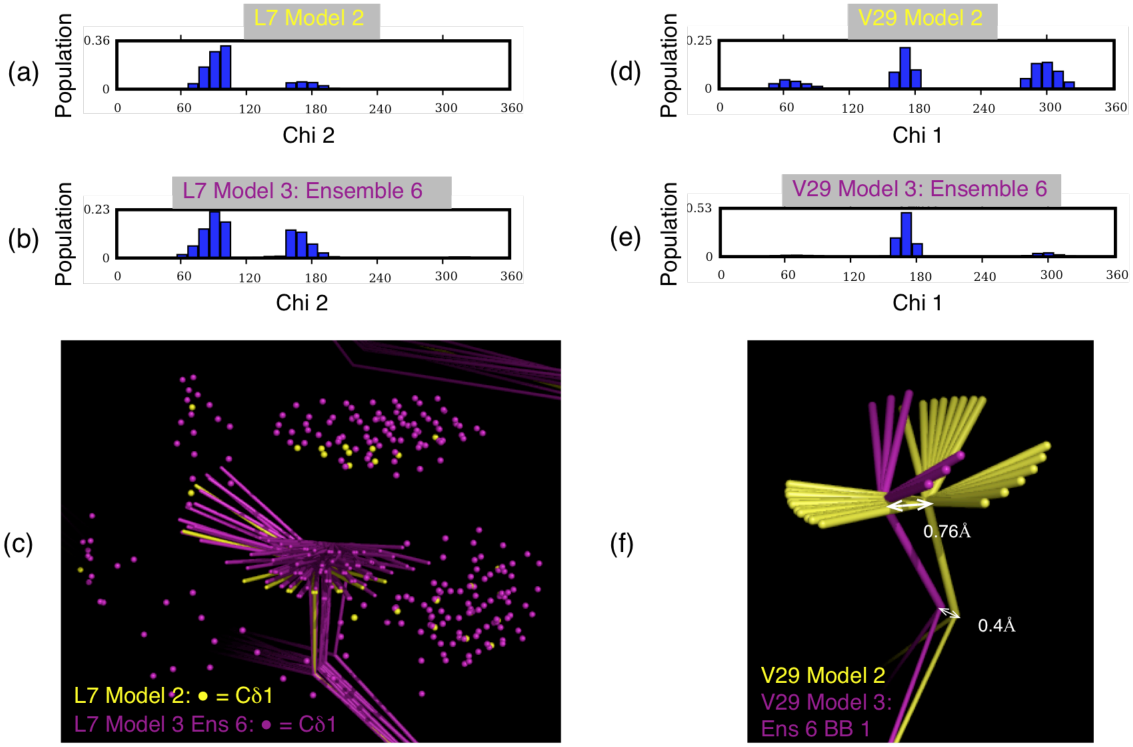


Figure 2-6. Backbone flexibility (Model 3) in gb1 results in the side-chain of L7 becoming more flexible and V29 becoming more rigid.

Probability distributions of L7 chi 2 from **(a)** Model 2, and **(b)** ensemble 6 of Model 3. **(c)** Structures of all conformers for L7 from Model 2 (yellow) and Model 3 (purple). The C-C and C-C bonds are drawn with lines while the C1 atoms are drawn as disconnected spheres for clarity. Probability distributions of V29 chi 1 from **(d)** Model 2, and **(e)** ensemble 6 of Model 3. **(f)** Structures of all V29 conformers from Model 2 (yellow) and a selected backbone from ensemble 6 of Model 3 (purple). This backbone is representative of those that keep V29 predominantly in 1 rotamer.

Notably, backbone and side-chain conformations simulated in the gb1 backrub ensembles also serve to increase some modeled side-chain order parameters. For example, V29 is located in the center of the helix and is facing into solvent with a ~50% solvent-accessible surface-area. Its flexible-backbone S^2_{calc} is substantially higher than the value from Model 2, and is closer to the experimental value (**Figure 2-5**). On the fixed backbone, V29 populates all three rotamers (**Figure 2-6d**) but in the flexible-backbone ensemble only the chi 1=180 degrees rotamer is significantly populated (**Figure 2-6e**). **Figure 2-6f** illustrates a possible mechanism for this increase in modeled rigidity. We observe a hinge motion at residue V29 in a representative backbone in the ensemble, resulting from Backrub rotations that cause a 0.4Å movement of the V29 Cα atom away from its position in the X-ray structure and a 0.76Å movement of the Cβ atom (**Figure 2-6f**). This puts the Cγ atoms in a different environment where they form closer packing interactions with the alpha helix. Thus, small backbone variations may have considerable effects on the rotamer populations of surface-exposed residues (and likely buried residues as well), highlighting the importance of using a flexible-backbone model for prediction and design.

Another interesting case is residues 13 and 15 of ubiquitin. These side-chains, despite being buried in the protein core, were identified as highly flexible in a key study determining ensembles of protein conformations that represent simultaneously the native structure and its associated dynamics [14]. Our dataset contains measured order parameters for the Cγ and Cd methyl groups of I13 and the Cd methyl groups of L15. In all three cases, both Models 2 and 3 predict order parameters below 0.6. Thus, in agreement with reference 5, both side-chains are modeled to populate multiple rotameric states. As described above for V29 in gb1, inclusion of backbone flexibility in ubiquitin increases the S^2_{calc} of two of these methyl groups (I13 Cd and L15 Cd). For I13 Cd, S^2_{calc} rises substantially from 0.1 (Model 2) to 0.49 (Model 3), which is closer to the S^2_{exp} value of 0.55. To rationalize this observation, we analyzed the population distributions of the I13 chi 1 and chi 2 angles predicted from Models 2 and 3. Supplemental **Figure 2-S1a** shows that I13 chi 2 is relatively unaffected by the inclusion of backbone flexibility; however, the chi 1 distribution shifts from one major rotamer with a moderate and a minor rotamer (Model 2) to a different major rotamer with two more equally-populated moderate rotamers. This change in the population distribution for I13 chi 1 will affect the positions of the Cd atoms and is likely responsible for the modeled increase in order parameters for the I13 Cd methyl group. These observations indicate that relatively subtle redistributions in rotamer populations may have substantial effects propagated through the chi dihedral angles to the ends of the side-chain. While this increased order parameter is closer to the experimental value, the order parameter of the I13 Cγ methyl group, which is correctly identified as flexible, is under-predicted by both Models 2 and 3 (with an S^2_{calc} of 0.22 and 0.21, respectively, compared to an S^2_{exp} of 0.6). For L15 Cd, S^2_{calc} changes slightly from 0.32 (Model 2) to 0.4 (Model 3), closer to the S^2_{exp} value of 0.6, with relatively minimal change in the predicted chi angle population distributions between Models 2 and 3 (**Figure 2-S1d**).

Comparison to other methods

Side-chain order parameters are generally difficult to predict using simple models based only on packing density (which work well for backbone amide order parameters) [41] or solvent accessibility [42]. A different method [43] reached correlation coefficients between modeled and observed order parameters of r>0.6 for 4 out of 7 proteins. This model described the number of contacts around each methyl carbon and their distances from the backbone with 4 parameters, which were determined by fitting optimal values to 5 proteins in the dataset [43].

Another approach for modeling side-chain order parameters used the structural variation present in ensembles of crystal structures of the same protein [44]. These ensembles with >98% sequence-identity contain small but significant conformational differences due to different crystallization conditions or small numbers of mutations. From the data available, this method seems to perform similarly to ours in overall average correlation coefficient and rmsd (**Table 2-4)**. An advantage of our method is that it only requires a single structure of the protein in question.

Several other studies performed MD and calculated order parameters from the structures in the trajectory. A thorough comparison with MD studies is difficult because MD statistics with correlation coefficients and/or rmsds have been published for only a few proteins. In one study, Best et al. performed MD in implicit solvent for 5ns on each of 18 proteins and found correlation coefficients between 0.4-0.7; however, the results were not reported per protein [34]. Three proteins simulated with explicit solvent MD are shown in **Table 2-4**; of these, eglin c (run for 80ns) and TNfn3 (run for 3ns) agree better with experimental data than do our simulations, and FNfn10 (run for 3ns) performs similarly [30; 33]. A drawback observed in the MD studies is the difficulty sampling rotamer transitions for some residues on the timescale of the simulation, while our models, which do not directly consider a timescale, do not have this problem. In fact, this difference may contribute to the improved performance of MD in modeling side-chain motions on the relatively short picosecond to nanosecond timescale of the side-chain relaxation measurements, while the rotamer transitions modeled by our method may occur on longer timescales (e.g. 5 surface-exposed threonines of protein L appear to make slow transitions; see above).

Table 2-4. Comparison of the results of Models 2 and 3 to results from other methods as indicated.

	Ensembles of X-ray structures [a]			Explicit Solvent MD [b]		Model 2		Model 3	
	No. of structs	r	rmsd	r	rmsd	r	rmsd	r [c]	rmsd
a3D						0.64	0.22	0.74	0.17
albp	14	0.73	0.19			0.79	0.19	0.76	0.19
calmodulin	28	0.72	0.20			0.81	0.21	0.78	0.21
Cdc42hs	13	0.53	0.30			0.31	0.33	0.41	0.31
cytochrome c2						0.37	0.24	0.52	0.22
eglin c	10	0.37	0.30	0.84		0.55	0.25	0.59	0.25
flavodoxin						0.63	0.21	0.62	0.21
FNfn10				0.51	0.23	0.40	0.29	0.42	0.24
fyn-sh3	12	0.74	0.21			0.89	0.15	0.62	0.20
gb1						0.49	0.26	0.68	0.22
hiv1 protease	330	0.74	0.17			0.73	0.22	0.62	0.23
mfabp						0.59	0.27	0.60	0.26
NTnC	13	0.69	0.19			0.77	0.23	0.73	0.22
plcc sh2						0.57	0.26	0.60	0.24
protein L						-0.07[d]	0.35 [d]	-0.06 [d]	0.33 [d]
TNfn3				0.62	0.23	0.34	0.32	0.38	0.31
ubiquitin	13	0.76	0.18			0.60 [d]	0.27 [d]	0.64 [d]	0.24 [d]
Average (Proteins from [a])		0.66	0.22			0.68 *	0.23 *	0.65 *	0.23 *
Average (All proteins)						0.55 *	0.25 *	0.57 *	0.24 *

[a] Data from [36]

[b] Eglin c data from [21]; FNfn10 and TNfn3 data from [24]

[c] Calculated over all S^2_{calc} values for the 10 backbone ensembles

[d] These proteins exhibited problems modeling threonines (see Results for details)

* Sum of the values for each protein divided by the number of proteins (i.e. different that the metric at the bottom of **Table 2-2**); provided for comparison with values from [36]

The study by Lindorff-Larsen et al [5] mentioned above provides another useful point of comparison. This work derived an ensemble of protein conformations that is simultaneously consistent with both experimentally determined order parameters for the native state of ubiquitin as well as distance information from nuclear Overhauser effect (NOE) data. By incorporating the experimental data as restraints in molecular dynamics simulations, the authors conclude that ubiquitin displays significant conformational heterogeneity in solution, with several side chains populating multiple rotameric conformations. The chi angle probability distributions from our simulations (**Figure 2-S2**) can be compared with those depicted in Figure 3 of reference [5]. The corresponding distributions show substantial agreement, except for the chi 2 angle distribution for Leu 67 that we model to be closer to a previously determined NMR ensemble [45].

Our method has several useful advantages, especially in the absence of extensive experimental data on the structure and dynamics of the protein in question. Our method uses a single crystal structure rather than requiring at least ten crystal structures with high sequence identity. It performs similarly to implicit solvent MD, but not as well as explicit solvent MD in two of the published cases. Nevertheless, our methods model motions (such as slow rotamer transitions) that are difficult to sample with MD. It also runs quickly, taking about 4 minutes to run the fixed-backbone protocol and about 13 hours to run the flexible-backbone protocol (numbers are for modeling ubiquitin on a single AMD Opteron 240 processor). (These time lengths are for the protocol used here in which computational efficiency was not a consideration; we expect that these methods can be sped up significantly.) This lower computational cost provides a significant advantage over MD in the case of protein design methods where both sequence and structure space need to be searched.

CONCLUSIONS

We have compared 3 different models of side-chain flexibility to experimental relaxation side-chain order parameter measurements. While native-rotamer motions (Model 1) do not reproduce the range of S^2_{exp} values, consistent with other studies [14; 30; 33], our fixed-backbone Monte Carlo method to sample rotameric transitions (Model 2) gives reasonable agreement between S^2_{calc} and S^2_{exp} values. Importantly, we expand upon this Monte Carlo model of side-chain motion by incorporating near-native ensembles of backbone structures into our simulations (Model 3). These backbone motions were inspired by Backrub conformational changes observed in ultra-high resolution X-ray structures, and allow correlated backbone-side-chain movements. Using this flexible-backbone model (Model 3), we find statistically significant overall improvements in rmsd (over Model 2), which are consistent over the majority of proteins in our dataset. This result demonstrates that Backrub simulations are a useful method for sampling near-native conformational space. Both Models 2 and 3 achieve these results despite inherent simplifications: (i) ligands are treated as rigid, (ii) the experiments are limited to motions on the picosecond-nanosecond timescale while our simulations are timescale-independent, and (iii) we do not model water molecules explicitly.

Conceptually similar to the dipeptide Backrub move in Model 3 is the 1-D Gaussian Axial Fluctuation (GAF) model. The similarity is worthy of note because this model was shown to explain motions present in Residual Dipolar Coupling experiments, which measure events that can take up to milliseconds to occur [46]. Our model extends the 1-D GAF model by adding both tripeptide moves and rotamer changes, allowing significant anisotropy and structural deviations from the native structure.

The treatment of backbone flexibility here is simple, and leaves room for improvement from more sophisticated models. The dataset used in this study provides a useful benchmark to evaluate such models. Notably, the new degrees of freedom introduced by backbone motions, while producing the expected increase in flexibility of some residues, can also cause increased side-chain rigidity, as was observed in gb1 and ubiquitin.

The models of side-chain flexibility presented here have many uses in prediction or design applications. First, incorporating information about side-chain flexibility changes resulting from macromolecular interactions should improve prediction and design of binding. Specifically, having an improved picture of the high-energy states of a protein will help in prediction and design of binding by conformational selection. Second, we show that modeling backbone flexibility leads to significant differences in sampled side-chain conformations, an effect that is likely to increase the diversity of sequences sampled during protein design. Sequences sampled during fixed-backbone design are strongly biased towards the particular backbone conformation used. Thus, removing the restraint of a fixed backbone (even with the small amount of conformational variability used here) will allow a greater diversity of amino acids at designed positions and may enable the computational design of sequence libraries matched to specific engineering tasks (E. Humphris and T.K., unpublished results). Third, our method can be used to design for flexibility or rigidity of a protein. This can be accomplished by adapting the flexibility prediction technique described here into either a post-processing filter or a score term calculated on the fly during the design protocol. In the latter method, short side-chain Monte Carlo simulations could be used to evaluate whether an amino acid substitution results in the desired flexibility profile. Considering flexibility explicitly may also prove useful in the design of enzymes [47], an idea supported by recent NMR data highlighting the importance of conformational dynamics in the rate of catalytic turnover [48; 49; 50]. Finally, these and similar simulation methods might be useful for investigating energetically connected pathways of residues in proteins, which could lead to the design of proteins with new modes of allosteric regulation.

MATERIALS AND METHODS

Dataset of experimental protein structures and relaxation measurements

The proteins used in this study are: ubiquitin (PDB id: 1UBQ) [51], the third fibronectin type III domain from human tenascin (TNfn3; PDB id: 1TEN) [52], the tenth fibronectin type III domain from human fibronectin (FNfn10; PDB id: 1FNA) [53], the B1 domain from protein G (gb1; PDB id: 1PGB) [54], the SH3 domain in human Fyn (fyn-sh3; PDB id: 1SHF) [42], the *de novo* designed three-helix bundle α3D (a3D; PDB id: 2A3D) [15], eglin c (PDB id: 1CSE) [16; 55], an SH2 domain of phospholipase C-gamma1 (plcc-sh2;

PDB id: 2PLD) [16], the B1 domain of protein L (proteinL; PDB id: 1HZ6) [39], HIV-1 protease homodimer bound to the inhibitor DMP323, (hiv-protease-holo; PDB id: 1QBS), flavodoxin bound to flavin mononucleotide (flavodoxin-holo; PDB id: 1OBO) [56], cytochrome c2 bound to its heme prosthetic group (cytochrome-c2-holo; PDB id: 1C2R) [57], the N-terminal domain of chicken skeletal troponin C (NTnC; PDB id: 1AVS) [58], Cdc42Hs [59], adipocyte lipid-binding protein (albp; PDB id: 1LIB) [60], muscle fatty acid-binding protein (mfabp; PDB id: 1HMT) [60], and Ca^{2+}-loaded calmodulin (PDB id: 1AHR) [61].

For NMR structures (i.e. 2PLD and 2A3D), the first submitted conformation was used after minimizing all atoms with the Protein Local Optimization Program (which uses a variant of the Truncated Newton method with the OPLS force field and Generalized Born solvation) [62; 63]. The highest resolution structure was used for proteins with multiple crystal structures, and the first chain was chosen for structures with multiple chains (except for the homodimer HIV protease). If a protein had relaxation measurements in both apo and holo states, we chose to model the apo state; however if no apo structures were available, the ligands were removed from the structure. If measurements were only available for a ligand-bound protein, we included the ligand atoms held fixed at their crystal structure coordinates.

There are 9 types of methyl groups with relaxation side-chain order parameter (S^2) measurements: alanine β, valine $\gamma1$ & $\gamma2$, threonine γ, isoleucine γ & δ, leucine $\delta1$ & $\delta2$, and methionine ε. We did not analyze alanine methyl group dynamics as alanine side-chains lack non-hydrogen torsional degrees of freedom. Our models are based on idealized bond geometry and thus treat the symmetric methyl groups of valine and leucine identically; for these methyl groups we compare the computed S^2 to the average of the two experimental S^2 values (if both were available). If backbone and side chain motions are anisotropic, then the two methyl groups of leucine and valine residues are not equivalent and will have slightly different order parameters. However, averaging the values can be justified as they are quite close, with differences less than 0.1 for 53 out of the 60 pairs of Leucine or Valine methyl groups in our dataset and less than 0.05 for 37 of these methyl group pairs.

Fixed-backbone Monte Carlo side-chain simulations (Models 1 and 2)

All simulations use the Rosetta program for protein structure modeling and design [64]. Three different models were used to simulate side-chain flexibility. Models 1 and 2 (Model 3 is described in a later section) used Monte Carlo simulations consisting of side-chain conformer changes on a fixed polypeptide backbone evaluated with the Metropolis criterion. This method is similar to side-chain repacking in Rosetta, with the difference here that the temperature is fixed at kT=1 after the initial annealing procedure (see below). A Monte Carlo move in this simulation consisted of randomly choosing a residue in the protein and changing its side-chain conformation to a "conformer" with side-chain chi dihedral angles chosen according to PDB statistics (described in the next paragraph) using idealized bond geometry. (To avoid confusion we use the term "rotamer" to describe the conformational macrostate including all nearby microstates within the same dihedral energy basin. The term "base rotamer" is used to describe the side-chain conformational microstate at the center of this basin, and the term "conformer" is used to describe any side-chain conformational microstate.) An initial simulated annealing procedure was used to equilibrate the protein to the force field; this consisted of starting the Monte Carlo simulations at high temperature (kT=100) and exponentially decreasing the temperature in stages to kT=1. The number of Monte Carlo moves performed in this initial annealing process was 200 times the number of conformers included in the conformer library of a protein. The number of moves performed at fixed temperature was 800 times the number of conformers. Each such simulation on a given protein was performed 10 times (unless otherwise noted) with different seeds for the random number generator. The simulations used the Rosetta all-atom scoring function, which is dominated by Lennard-Jones packing interactions, an orientation-dependent hydrogen bonding potential [65] and an implicit solvation model [66], as described in detail in [2]. The results were somewhat dependent on the simulation temperature, but kT=1 was found optimal overall in the context of the Rosetta all-atom scoring function (data not shown).

Side-chain conformer libraries

The “conformer” library of possible side-chain conformations for a given residue was created by first selecting the base rotamers using Dunbrack’s backbone-dependent rotamer library [67; 68] and then adding conformers around each base rotamer. The library was defined by several attributes: (a) the number of base rotamers to include, (b) the conformer resolution, or chi angle separation between adjacent conformers for chi 1 and 2; and (c) the rotamer well width, or maximum chi 1 and 2 angle distance of conformers from the base rotamer (expressed as the number of standard deviations tabulated in the Dunbrack library). For Model 1, one base rotamer was chosen per residue by finding the base rotamer with the lowest heavy-atom rmsd to the crystal side-chain conformation (with conformers added around it as described). For Model 2, the base rotamer library was either: (i) the "default" library: consisting of 95% or 98% accumulated probability of occurrences in the PDB but restricted to at most 24 or 30 conformers for surface or buried base rotamers, respectively, or (ii) the "large" library: consisting of 99% accumulated probability with at most 45 conformers for a given surface or buried base rotamer. Only the base rotamers were used for chi 3 and 4, without adding neighboring conformers.

Side-chain repacking test

To test side-chain repacking accuracy, we used the same dataset of 65 X-ray structures described in [36]. For each protein, all residues were repacked simultaneously using Rosetta with the large base rotamer library, a rotamer well width of 1.5 times the Dunbrack standard deviation for that base rotamer, and a conformer resolution of 10 degrees. A repacked residue was classified as having a correctly assigned chi 1 or chi 1+2 conformation if the modeled chi values deviated by less than 40 degrees from the corresponding X-ray structure values.

Modified Lennard-Jones terms

Rosetta models the Lennard-Jones (LJ) term using the classical 6-12 potential for attractive contributions and some repulsive contributions; however, at inter-atomic distances in the repulsive regime less than a "switchover" distance, the repulsive potential is modeled as a line with the same slope as the 6-12 potential at this distance [64]. The default ("hard repulsive") value of this switchover distance is $d_{ij}/r_{ij} = 0.6$, which gives a well depth-independent slope of ~ -9000 (where d_{ij} and r_{ij} are the inter-atomic distance and summed van der Waals radii, respectively, for atoms i and j). The LJ radii are derived from fitting atom distances in protein X-ray structures to the 6-12 LJ potential using CHARMm well depths [69]. Two variants of a "reduced" LJ repulsive term were used. The first "small radii" modification used LJ radii values scaled by 0.95 [70]. The second "soft repulsive" modification reduced the linear slope of the LJ repulsive term. The adjusted value of the switchover distance for this "soft repulsive" modification is $d_{ij}/r_{ij} = 0.91$, which gives a well depth-independent slope of ~ -18.

Generation of conformational ensembles using Backrub Monte Carlo simulations

To generate protein conformational ensembles with varying backbone conformations, we ran Metropolis Monte Carlo simulations using two types of moves with equal probability: a) a side-chain conformer change, or b) a backbone and side-chain change resulting from a "Backrub" move. The Backrub move was motivated by a type of conformational variability frequently observed in alternate conformations of the same chain of ultra-high resolution crystal structures [8]. In our implementation, the Backrub move consisted of: (i) choosing a random peptide segment of 2 or 3 successive Cα atoms with endpoint residues a and b, (ii) performing a rigid body rotation of main-chain and side-chain atoms between $C\alpha_a$ and $C\alpha_b$ about the axis connecting $C\alpha_a$ and $C\alpha_b$, (**Figure 2-1b** and **c**) and (iii) optimizing the bond angles extending from $C\alpha_a$ and $C\alpha_b$ using the CHARMm22 [71] bond angle potential. (C.A.S and T.K., unpublished results) This Backrub Monte Carlo simulation was run for 10,000 steps at kT=0.6. Side-chain conformers were taken from the Dunbrack library [67] with conformations around each base rotamer added for chi 1 and chi 2 as described [72].

Simulations including backbone flexibility (Model 3)

Backbone flexibility was included into the side-chain simulations by running fixed-backbone Monte Carlo simulations on an ensemble of backbone conformations generated using Backrub Monte Carlo simulations. Each backbone was generated by selecting the lowest energy structure from a Backrub Monte Carlo simulation (as described above). For each protein, ten ensembles were used. For each ensemble, 100 structures were generated and then pruned down to the ten with the lowest energy. One fixed-backbone side-chain Monte Carlo simulation was then run on each backbone in the 10-member ensemble.

Calculation of order parameters

For each fixed-backbone side-chain Monte Carlo step that a residue was in a particular conformer, the count for that conformer was incremented. At the end of a simulation on a particular backbone, the population of each conformer was calculated as the sum of the conformer counts, divided by the total number of non-annealing steps. For multiple independent simulations (on the same or different backbones) all conformers in the simulations were accumulated and their probabilities were renormalized to a sum of 1. The order parameters were calculated from these conformer populations and the coordinates (x, y, z) of the conformer's relevant methyl carbon (using Cγ1 for valines and Cδ1 for leucines):

$$S^2 = \frac{3}{2}\left[\langle x^2\rangle^2 + \langle y^2\rangle^2 + \langle z^2\rangle^2 + 2\langle xy\rangle^2 + 2\langle xz\rangle^2 + 2\langle yz\rangle^2\right] - \frac{1}{2}$$

Equation 1 [55]

Analysis of goodness-of-fit

The level of agreement between the experimental and simulated order parameters was calculated in three ways: (a) the linear correlation coefficient between the two sets of order parameters, (b) the root mean squared deviation (rmsd) between these two sets, and (c) the percentage of methyl groups that were correctly modeled as "rigid" (defined as S^2 >= 0.75) or "flexible" (defined as $S^2 < 0.75$). The cutoff value for the order parameters of 0.75 is an approximation of the threshold that was observed in multiple studies [30; 34; 67] (including this one) to distinguish qualitatively between side-chains populating one or multiple rotameric states.

Analysis of statistical significance

For each of the 17 proteins in the dataset, the performance of Models 2 and 3 were analyzed by applying the Student's t-test to compare the Model 2 rmsd to the Model 3 ensemble rmsds. The difference between the rmsds of two models was judged significant when the one-tailed p-value was less than 0.005. The Wilcoxon signed-rank test was performed on the same data and judged significant when the p-value was less than 0.01.

The error between the calculated and experimental order parameters for Models 2 and 3 were also compared across each of the 530 methyl groups in the dataset. The errors were calculated as the unsigned distances between S^2_{calc} and S^2_{exp}. (The order parameters for Model 3 were averaged over the 10 ensembles). The difference in the errors between Models 2 and 3 were evaluated with the paired Student's t-test and the paired Wilcoxon signed-rank test against the null hypothesis that there is no difference in the magnitude of the errors).

ACKNOWLEDGEMENTS

We would like to thank Lewis Kay, Anthony Mittermaier, Martin Stone, Robert Oswald, and Adrienne Loh for providing data on side-chain methyl relaxation order parameters and Dan Mandel, Chris McClendon and Ashley Conrad-Saydah for critical reading of the manuscript. Kristian Kaufman provided scripts to model ligands. G.D.F. is the recipient of an NSF Graduate Research Fellowship. A.J.L. was supported by the UCSF Summer Research Training Program and Genentech. C.A.S. received support from an NIH Training Grant GM067547, a DOD NDSEG fellowship and the Genentech Scholars Program. T.K. is an Alfred P. Sloan Foundation Fellow in Molecular Biology. This work was additionally supported by Sandler start-up funding to T.K.

AUTHOR CONTRIBUTIONS

GDF and TK conceived and designed the experiments. GDF and AJL performed the experiments. GDF, TK, and AJL analyzed the data and wrote the paper. CAS contributed reagents/materials/analysis tools.

SUPPORTING INFORMATION

Table 2-S1. Model 1 results for all methyl groups in all proteins.

	No. methyls	r	slope	rmsd	No. Rigid (exp)	No. Flexible (exp)	Fraction Correct Rigid (calc)	Fraction Correct Flexible (calc)
a3D	15	0.30	0.10	0.52	0	15	N/A	0.00
albp	28	0.64	0.11	0.39	13	15	1.00	0.00
calmodulin	36	0.40	0.07	0.51	4	32	1.00	0.00
cdc42hs	47	0.39	0.08	0.37	23	24	1.00	0.04
cytochrome-c2-holo	31	0.21	0.03	0.30	21	10	1.00	0.00
eglinc	17	0.49	0.15	0.39	6	11	1.00	0.09
flavodoxin-holo	56	0.34	0.05	0.31	36	20	1.00	0.00
FNfn10	35	0.43	0.08	0.34	12	23	1.00	0.04
fyn-sh3	12	-0.02	0.00	0.38	4	8	1.00	0.00
gb1	10	0.67	0.38	0.52	0	10	N/A	0.10
hiv-protease-holo	56	0.29	0.07	0.34	36	20	1.00	0.15
mfabp	42	0.43	0.06	0.41	20	22	1.00	0.00
NTnC	26	0.51	0.10	0.47	2	24	1.00	0.00
plcc-sh2-free	27	0.32	0.11	0.48	3	24	0.67	0.04
protl	20	0.77	0.10	0.30	8	12	1.00	0.00
TNfn3	40	0.58	0.12	0.41	11	29	1.00	0.03
ubiquitin	32	0.60	0.23	0.31	16	16	1.00	0.25
Total	530	.	.	0.39	215	315	1.00	0.04

See caption for **Table 2-2** for a description of the column headers.

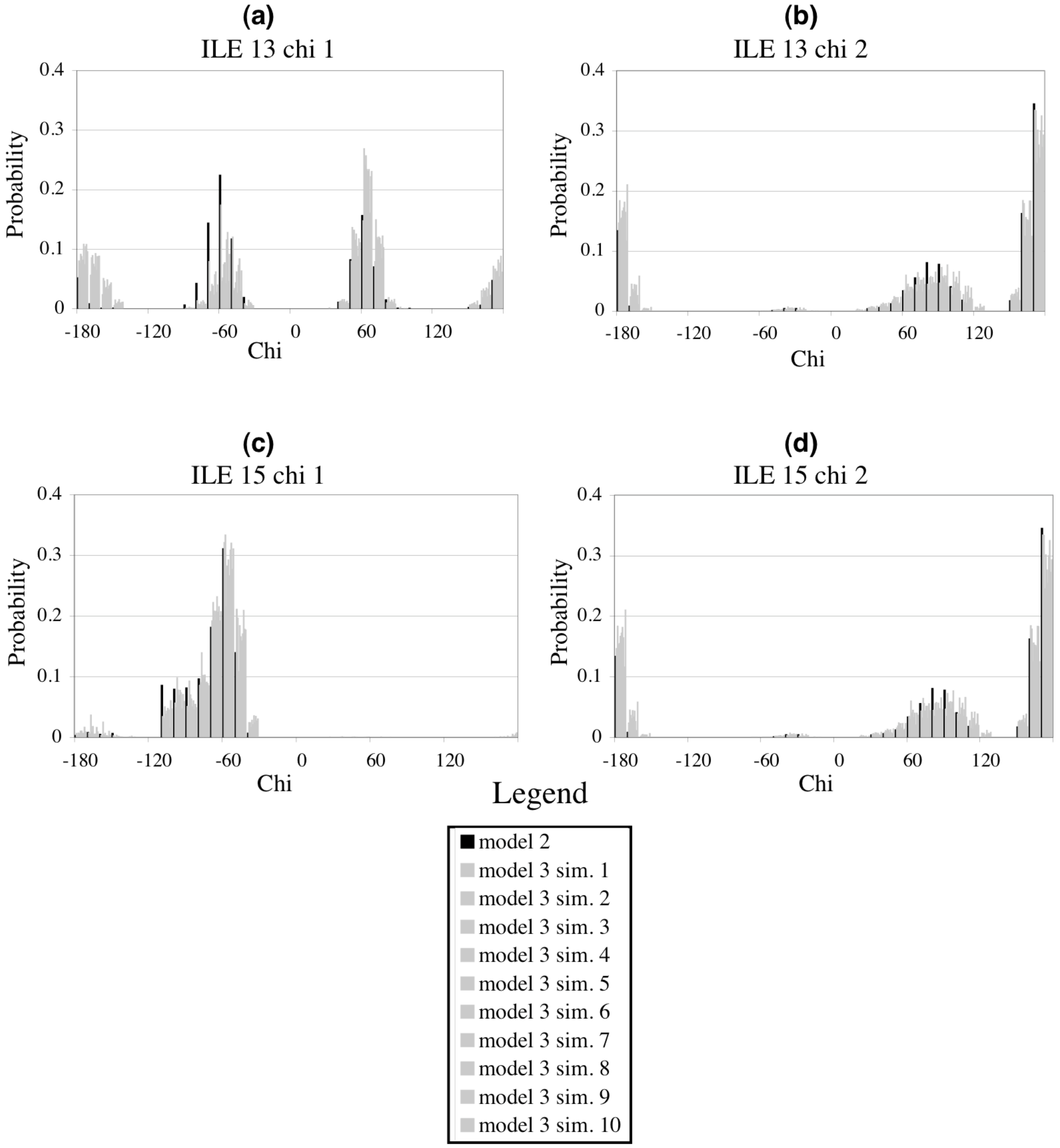
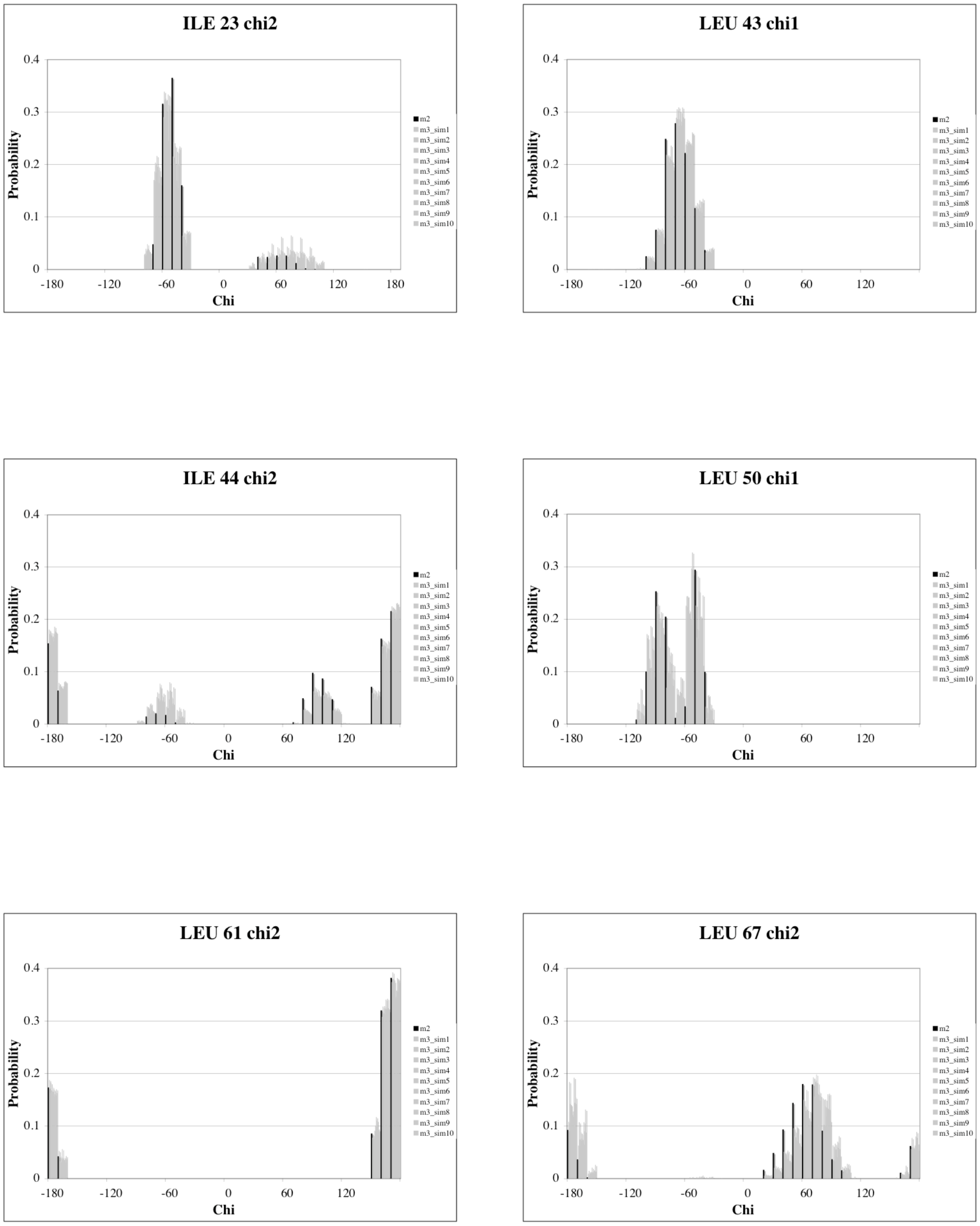


Figure 2-S1. Modeled probability distributions of Isoleucine 13 and Leucine 15 in ubiquitin.

(a) I13 chi 1, (b) I13 chi 2, (c) L15 chi 1, and (d) L15 chi 2 dihedral angles from Model 2 (black lines) and using the 10 different backbone ensembles from Model 3 (grey lines). The standard deviations in calculated order parameters obtained from the different Model 3 ensembles were less than 0.2.



ILE 23 chi2

LEU 43 chi1

ILE 44 chi2

LEU 50 chi1

LEU 61 chi2

LEU 67 chi2

Figure 2-S2. Modeled probability distributions of other selected ubiquitin chi angles.

I23 chi 1, L43 chi 1, I44 chi 2, L50 chi 1, L61 chi 2, and L67 chi 2 dihedral angles from Model 2 (black lines) and using the 10 different backbone ensembles from Model 3 (grey lines). These plots provide an interesting comparison to the probability distributions of these dihedral angles in Figure 3 of reference [5].

Table 2-S2. Results for nonpolar methyl groups only (valine, leucine, isoleucine, and methionine) with (a) Model 2 and (b) Model 3.

		(a) Model 2			(b) Model 3		
	# of methyls	r	slope	rmsd	r	slope	rmsd
a3D	14	0.65	1.28	0.22	0.74	1.36	0.17
albp	25	0.85	1.02	0.18	0.80	0.96	0.18
calmodulin	31	0.86	1.32	0.20	0.86	1.28	0.20
cdc42hs	39	0.35	0.38	0.34	0.45	0.47	0.30
cytochrome-c2-holo	23	0.16	0.13	0.27	0.49	0.50	0.24
eglinc	14	0.56	0.73	0.26	0.61	0.80	0.24
flavodoxin-holo	47	0.64	0.77	0.21	0.65	0.79	0.21
FNfn10	25	0.50	0.72	0.27	0.54	0.64	0.23
fyn-sh3	9	0.94	1.63	0.15	0.77	1.15	0.17
gb1	8	0.59	0.83	0.26	0.85	1.22	0.20
hiv-protease-holo	56	0.73	0.86	0.22	0.62	0.68	0.23
mfabp	34	0.66	0.76	0.25	0.70	0.80	0.24
NTnC	23	0.75	1.10	0.23	0.71	1.05	0.23
plcc-sh2-free	25	0.63	0.88	0.26	0.64	0.81	0.23
protl	12	0.68	0.85	0.25	0.66	0.82	0.23
TNfn3	28	0.40	0.50	0.33	0.50	0.62	0.30
ubiquitin	26	0.70	0.93	0.24	0.75	0.92	0.20
Total	439	.	.	0.25			0.23

See caption for **Table 2-2** for a description of the column headers.

Table 2-S3. Model 2 Parameter sensitivity results for all proteins.

LJ repulsive term
hard repulsive

	No. methyls	r	slope	rmsd	No. Rigid (exp)	No. Flexible (exp)	Fraction Correct Rigid (calc)	Fraction Correct Flexible (calc)
a3D	15	0.64	1.24	0.22	0	15	N/A	0.73
albp	28	0.79	0.98	0.19	13	15	0.77	0.73
calmodulin	36	0.81	1.22	0.21	4	32	0.75	0.69
cdc42hs	47	0.31	0.33	0.33	23	24	0.61	0.58
cytochrome-c2-holo	31	0.37	0.35	0.24	21	10	0.67	0.50
eglinc	17	0.55	0.76	0.25	6	11	0.67	0.82
flavodoxin-holo	56	0.63	0.74	0.21	36	20	0.83	0.85
FNfn10	35	0.40	0.57	0.29	12	23	0.67	0.65
fyn-sh3	12	0.89	1.45	0.15	4	8	0.75	0.88
gb1	10	0.49	0.64	0.26	0	10	N/A	0.80
hiv-protease-holo	56	0.73	0.86	0.22	36	20	0.69	0.95
mfabp	42	0.59	0.69	0.27	20	22	0.55	0.82
NTnC	26	0.77	1.13	0.23	2	24	1.00	0.58

plcc-sh2-free	27	0.57	0.78	0.26	3	24	0.33	0.67
protl	20	-0.07	-0.10	0.35	8	12	0.62	0.50
TNfn3	40	0.34	0.40	0.32	11	29	0.64	0.55
ubiquitin	32	0.60	0.82	0.27	16	16	0.75	0.75
Total	530	.	.	0.26	215	315	0.69	0.70

small radii

a3D	15	0.35	0.54	0.20	0	15	N/A	0.93
albp	28	0.79	0.99	0.18	13	15	0.69	0.80
calmodulin	36	0.82	1.21	0.19	4	32	0.75	0.78
cdc42hs	47	0.34	0.36	0.34	23	24	0.43	0.75
cytochrome-c2-holo	31	0.60	0.57	0.20	21	10	0.52	0.90
eglinc	17	0.56	0.77	0.25	6	11	0.67	0.73
flavodoxin-holo	56	0.64	0.78	0.21	36	20	0.75	0.90
FNfn10	35	0.42	0.59	0.29	12	23	0.58	0.74
fyn-sh3	12	0.88	1.44	0.17	4	8	0.75	1.00
gb1	10	0.61	0.79	0.22	0	10	N/A	0.80
hiv-protease-holo	56	0.73	0.85	0.24	36	20	0.53	0.95
mfabp	42	0.62	0.71	0.28	20	22	0.50	0.91
NTnC	26	0.74	1.10	0.22	2	24	1.00	0.67
plcc-sh2-free	27	0.65	0.85	0.23	3	24	0.67	0.79
protl	20	-0.02	-0.03	0.34	8	12	0.50	0.50
TNfn3	40	0.42	0.48	0.29	11	29	0.64	0.76
ubiquitin	32	0.61	0.85	0.28	16	16	0.75	0.75
Total	530	.	.	0.25	215	315	0.60	0.80

soft repulsive

a3D	15	0.52	0.44	0.21	0	15	N/A	1.00
albp	28	0.73	0.77	0.23	13	15	0.38	1.00
calmodulin	36	0.74	0.88	0.21	4	32	0.75	0.97
cdc42hs	47	0.47	0.42	0.35	23	24	0.17	0.92
cytochrome-c2-holo	31	0.45	0.46	0.33	21	10	0.24	1.00
eglinc	17	0.59	0.78	0.26	6	11	0.50	0.91
flavodoxin-holo	56	0.53	0.62	0.30	36	20	0.44	0.95
FNfn10	35	0.47	0.53	0.31	12	23	0.33	0.96
fyn-sh3	12	0.85	0.96	0.20	4	8	0.25	1.00
gb1	10	0.55	0.64	0.23	0	10	N/A	0.90
hiv-protease-holo	56	0.68	0.73	0.31	36	20	0.25	1.00
mfabp	42	0.64	0.66	0.31	20	22	0.40	1.00
NTnC	26	0.74	0.94	0.19	2	24	1.00	0.83
plcc-sh2-free	27	0.59	0.59	0.24	3	24	0.33	0.96
protl	20	0.07	0.10	0.34	8	12	0.38	0.83
TNfn3	40	0.51	0.50	0.27	11	29	0.18	0.93
ubiquitin	32	0.71	0.90	0.26	16	16	0.69	0.81
Total	530	.	.	0.28	215	315	0.36	0.94

Base rotamer library

default

a3D	15	0.33	0.62	0.35	0	15	N/A	0.33
albp	28	0.81	1.01	0.19	13	15	0.77	0.73
calmodulin	36	0.76	1.10	0.25	4	32	0.75	0.62
cdc42hs	47	0.34	0.32	0.31	23	24	0.61	0.54
cytochrome-c2-holo	31	0.45	0.41	0.22	21	10	0.71	0.60
eglinc	17	0.53	0.70	0.26	6	11	0.83	0.82

flavodoxin-holo	56	0.60	0.68	0.21	36	20	0.86	0.75
FNfn10	35	0.42	0.59	0.28	12	23	0.83	0.61
fyn-sh3	12	0.86	1.24	0.16	4	8	1.00	0.75
gb1	10	0.45	0.70	0.36	0	10	N/A	0.60
hiv-protease-holo	56	0.71	0.82	0.21	36	20	0.67	0.90
mfabp	42	0.55	0.63	0.27	20	22	0.60	0.77
NTnC	26	0.73	1.03	0.25	2	24	1.00	0.58
plcc-sh2-free	27	0.51	0.60	0.25	3	24	0.33	0.71
protl	20	-0.06	-0.09	0.33	8	12	0.62	0.50
TNfn3	40	0.36	0.42	0.32	11	29	0.73	0.52
ubiquitin	32	0.58	0.73	0.25	16	16	0.75	0.69
Total	530	.	.	0.26	215	315	0.73	0.64

large

a3D	15	0.64	1.24	0.22	0	15	N/A	0.73
albp	28	0.79	0.98	0.19	13	15	0.77	0.73
calmodulin	36	0.81	1.22	0.21	4	32	0.75	0.69
cdc42hs	47	0.31	0.33	0.33	23	24	0.61	0.58
cytochrome-c2-holo	31	0.37	0.35	0.24	21	10	0.67	0.50
eglinc	17	0.55	0.76	0.25	6	11	0.67	0.82
flavodoxin-holo	56	0.63	0.74	0.21	36	20	0.83	0.85
FNfn10	35	0.40	0.57	0.29	12	23	0.67	0.65
fyn-sh3	12	0.89	1.45	0.15	4	8	0.75	0.88
gb1	10	0.49	0.64	0.26	0	10	N/A	0.80
hiv-protease-holo	56	0.73	0.86	0.22	36	20	0.69	0.95
mfabp	42	0.59	0.69	0.27	20	22	0.55	0.82
NTnC	26	0.77	1.13	0.23	2	24	1.00	0.58
plcc-sh2-free	27	0.57	0.78	0.26	3	24	0.33	0.67
protl	20	-0.07	-0.10	0.35	8	12	0.62	0.50
TNfn3	40	0.34	0.40	0.32	11	29	0.64	0.55
ubiquitin	32	0.60	0.82	0.27	16	16	0.75	0.75
Total	530	.	.	0.26	215	315	0.69	0.70

Rotamer Well Width
sd 0.5 (5 degrees)

a3D	15	-0.21	-0.39	0.38	0	15	N/A	0.60
albp	28	0.76	0.93	0.22	13	15	0.85	0.67
calmodulin	36	0.71	1.05	0.28	4	32	0.75	0.59
cdc42hs	47	0.26	0.29	0.36	23	24	0.65	0.38
cytochrome-c2-holo	31	0.29	0.32	0.28	21	10	0.67	0.50
eglinc	17	0.55	0.75	0.27	6	11	0.83	0.64
flavodoxin-holo	56	0.57	0.67	0.24	36	20	0.86	0.65
FNfn10	35	0.31	0.45	0.30	12	23	0.67	0.61
fyn-sh3	12	0.86	1.70	0.20	4	8	1.00	0.75
gb1	10	0.36	0.47	0.30	0	10	N/A	0.80
hiv-protease-holo	56	0.56	0.69	0.26	36	20	0.69	0.75
mfabp	42	0.47	0.53	0.30	20	22	0.65	0.68
NTnC	26	0.77	1.11	0.27	2	24	1.00	0.50
plcc-sh2-free	27	0.61	0.87	0.29	3	24	0.67	0.54
protl	20	0.00	0.00	0.37	8	12	0.62	0.42
TNfn3	40	0.24	0.29	0.36	11	29	0.64	0.45
ubiquitin	32	0.54	0.77	0.30	16	16	0.75	0.69

Total	530	.	.	0.30	215	315	0.73	0.58

sd 1 (5 degrees)

a3D	15	-0.01	-0.03	0.33	0	15	N/A	0.67
albp	28	0.78	0.95	0.20	13	15	0.85	0.67
calmodulin	36	0.78	1.15	0.24	4	32	0.75	0.62
cdc42hs	47	0.26	0.28	0.35	23	24	0.61	0.50
cytochrome-c2-holo	31	0.31	0.32	0.27	21	10	0.67	0.50
eglinc	17	0.57	0.78	0.26	6	11	0.83	0.73
flavodoxin-holo	56	0.68	0.78	0.21	36	20	0.94	0.65
FNfn10	35	0.39	0.56	0.29	12	23	0.67	0.65
fyn-sh3	12	0.89	1.60	0.17	4	8	0.75	0.75
gb1	10	0.46	0.59	0.27	0	10	N/A	0.80
hiv-protease-holo	56	0.69	0.82	0.22	36	20	0.69	0.90
mfabp	42	0.55	0.62	0.27	20	22	0.60	0.77
NTnC	26	0.79	1.14	0.24	2	24	1.00	0.58
plcc-sh2-free	27	0.64	0.88	0.25	3	24	0.67	0.58
protl	20	-0.04	-0.07	0.36	8	12	0.62	0.42
TNfn3	40	0.30	0.37	0.34	11	29	0.64	0.52
ubiquitin	32	0.56	0.76	0.28	16	16	0.75	0.69
Total	530	.	.	0.27	215	315	0.73	0.64

sd 1.5 (10 degrees)

a3D	15	0.64	1.24	0.22	0	15	N/A	0.73
albp	28	0.79	0.98	0.19	13	15	0.77	0.73
calmodulin	36	0.81	1.22	0.21	4	32	0.75	0.69
cdc42hs	47	0.31	0.33	0.33	23	24	0.61	0.58
cytochrome-c2-holo	31	0.37	0.35	0.24	21	10	0.67	0.50
eglinc	17	0.55	0.76	0.25	6	11	0.67	0.82
flavodoxin-holo	56	0.63	0.74	0.21	36	20	0.83	0.85
FNfn10	35	0.40	0.57	0.29	12	23	0.67	0.65
fyn-sh3	12	0.89	1.45	0.15	4	8	0.75	0.88
gb1	10	0.49	0.64	0.26	0	10	N/A	0.80
hiv-protease-holo	56	0.73	0.86	0.22	36	20	0.69	0.95
mfabp	42	0.59	0.69	0.27	20	22	0.55	0.82
NTnC	26	0.77	1.13	0.23	2	24	1.00	0.58
plcc-sh2-free	27	0.57	0.78	0.26	3	24	0.33	0.67
protl	20	-0.07	-0.10	0.35	8	12	0.62	0.50
TNfn3	40	0.34	0.40	0.32	11	29	0.64	0.55
ubiquitin	32	0.60	0.82	0.27	16	16	0.75	0.75
Total	530	.	.	0.26	215	315	0.69	0.70

sd 2 (10 degrees)

a3D	15	0.63	1.23	0.21	0	15	N/A	0.73
albp	28	0.80	0.98	0.18	13	15	0.77	0.73
calmodulin	36	0.82	1.22	0.20	4	32	0.75	0.69
cdc42hs	47	0.32	0.34	0.33	23	24	0.61	0.62
cytochrome-c2-holo	31	0.36	0.35	0.25	21	10	0.67	0.50
eglinc	17	0.56	0.76	0.25	6	11	0.67	0.82
flavodoxin-holo	56	0.64	0.76	0.21	36	20	0.81	0.85
FNfn10	35	0.41	0.58	0.29	12	23	0.67	0.65
fyn-sh3	12	0.89	1.45	0.15	4	8	0.75	0.88
gb1	10	0.51	0.68	0.26	0	10	N/A	0.80
hiv-protease-holo	56	0.74	0.88	0.21	36	20	0.72	0.95
mfabp	42	0.59	0.68	0.27	20	22	0.50	0.82
NTnC	26	0.77	1.12	0.23	2	24	1.00	0.58

plcc-sh2-free	27	0.64	0.81	0.22	3	24	0.67	0.79
protl	20	-0.07	-0.11	0.35	8	12	0.62	0.50
TNfn3	40	0.33	0.39	0.32	11	29	0.55	0.55
ubiquitin	32	0.60	0.83	0.27	16	16	0.75	0.75
Total	530	.	.	0.26	215	315	0.69	0.71

sd 3 (10 degrees)

a3D	15	0.70	1.60	0.25	0	15	N/A	0.60
albp	28	0.80	0.98	0.18	13	15	0.77	0.73
calmodulin	36	0.81	1.22	0.21	4	32	0.75	0.69
cdc42hs	47	0.32	0.34	0.33	23	24	0.61	0.58
cytochrome-c2-holo	31	0.39	0.37	0.24	21	10	0.67	0.60
eglinc	17	0.56	0.77	0.25	6	11	0.83	0.82
flavodoxin-holo	56	0.67	0.81	0.20	36	20	0.81	0.80
FNfn10	35	0.41	0.58	0.29	12	23	0.67	0.65
fyn-sh3	12	0.88	1.44	0.15	4	8	0.75	0.88
gb1	10	0.54	0.75	0.26	0	10	N/A	0.80
hiv-protease-holo	56	0.74	0.88	0.21	36	20	0.72	0.95
mfabp	42	0.58	0.68	0.27	20	22	0.50	0.82
NTnC	26	0.78	1.13	0.22	2	24	1.00	0.62
plcc-sh2-free	27	0.65	0.82	0.22	3	24	0.67	0.79
protl	20	-0.07	-0.11	0.35	8	12	0.62	0.42
TNfn3	40	0.33	0.40	0.32	11	29	0.55	0.55
ubiquitin	32	0.61	0.82	0.27	16	16	0.75	0.75
Total	530	.	.	0.26	215	315	0.69	0.70

Rotamer Resolution
10 degrees

a3D	15	0.70	1.60	0.25	0	15	N/A	0.60
albp	28	0.80	0.98	0.18	13	15	0.77	0.73
calmodulin	36	0.81	1.22	0.21	4	32	0.75	0.69
cdc42hs	47	0.32	0.34	0.33	23	24	0.61	0.58
cytochrome-c2-holo	31	0.39	0.37	0.24	21	10	0.67	0.60
eglinc	17	0.56	0.77	0.25	6	11	0.83	0.82
flavodoxin-holo	56	0.67	0.81	0.20	36	20	0.81	0.80
FNfn10	35	0.41	0.58	0.29	12	23	0.67	0.65
fyn-sh3	12	0.88	1.44	0.15	4	8	0.75	0.88
gb1	10	0.54	0.75	0.26	0	10	N/A	0.80
hiv-protease-holo	56	0.74	0.88	0.21	36	20	0.72	0.95
mfabp	42	0.58	0.68	0.27	20	22	0.50	0.82
NTnC	26	0.78	1.13	0.22	2	24	1.00	0.62
plcc-sh2-free	27	0.65	0.82	0.22	3	24	0.67	0.79
protl	20	-0.07	-0.11	0.35	8	12	0.62	0.42
TNfn3	40	0.33	0.40	0.32	11	29	0.55	0.55
ubiquitin	32	0.61	0.82	0.27	16	16	0.75	0.75
Total	530	.	.	0.26	215	315	0.69	0.70

15 degrees

a3D	15	0.71	1.66	0.25	0	15	N/A	0.73
albp	28	0.80	0.98	0.18	13	15	0.85	0.73
calmodulin	36	0.82	1.24	0.21	4	32	0.75	0.69
cdc42hs	47	0.34	0.37	0.33	23	24	0.65	0.62
cytochrome-c2-holo	31	0.36	0.38	0.26	21	10	0.62	0.60

eglinc	17	0.58	0.79	0.25	6	11	0.83	0.82
flavodoxin-holo	56	0.67	0.84	0.21	36	20	0.81	0.85
FNfn10	35	0.39	0.56	0.29	12	23	0.58	0.65
fyn-sh3	12	0.88	1.45	0.16	4	8	0.75	0.88
gb1	10	0.55	0.79	0.27	0	10	N/A	0.70
hiv-protease-holo	56	0.72	0.88	0.22	36	20	0.67	0.95
mfabp	42	0.60	0.69	0.27	20	22	0.50	0.82
NTnC	26	0.78	1.13	0.22	2	24	1.00	0.62
plcc-sh2-free	27	0.63	0.83	0.24	3	24	0.67	0.67
protl	20	-0.05	-0.08	0.35	8	12	0.62	0.50
TNfn3	40	0.33	0.37	0.32	11	29	0.55	0.59
ubiquitin	32	0.62	0.84	0.27	16	16	0.75	0.75
Total	530	.	.	0.26	215	315	0.68	0.71

20 degrees

a3D	15	0.79	2.03	0.28	0	15	N/A	0.60
albp	28	0.80	0.95	0.19	13	15	0.85	0.73
calmodulin	36	0.80	1.20	0.22	4	32	0.75	0.69
cdc42hs	47	0.34	0.38	0.34	23	24	0.57	0.67
cytochrome-c2-holo	31	0.33	0.32	0.25	21	10	0.71	0.50
eglinc	17	0.59	0.81	0.25	6	11	0.83	0.64
flavodoxin-holo	56	0.66	0.83	0.21	36	20	0.78	0.85
FNfn10	35	0.33	0.47	0.30	12	23	0.58	0.61
fyn-sh3	12	0.89	1.52	0.16	4	8	0.75	0.88
gb1	10	0.53	0.76	0.28	0	10	N/A	0.70
hiv-protease-holo	56	0.69	0.88	0.24	36	20	0.61	0.95
mfabp	42	0.59	0.69	0.27	20	22	0.50	0.82
NTnC	26	0.78	1.13	0.24	2	24	1.00	0.54
plcc-sh2-free	27	0.55	0.75	0.27	3	24	0.33	0.67
protl	20	-0.02	-0.03	0.35	8	12	0.62	0.42
TNfn3	40	0.30	0.35	0.32	11	29	0.55	0.55
ubiquitin	32	0.63	0.87	0.26	16	16	0.75	0.75
Total	530	.	.	0.27	215	315	0.67	0.68

25 degrees

a3D	15	0.37	0.78	0.28	0	15	N/A	0.73
albp	28	0.79	0.91	0.21	13	15	0.85	0.73
calmodulin	36	0.78	1.17	0.23	4	32	0.75	0.66
cdc42hs	47	0.41	0.45	0.32	23	24	0.61	0.46
cytochrome-c2-holo	31	0.28	0.27	0.26	21	10	0.62	0.40
eglinc	17	0.61	0.84	0.26	6	11	0.83	0.55
flavodoxin-holo	56	0.65	0.77	0.21	36	20	0.83	0.75
FNfn10	35	0.20	0.28	0.32	12	23	0.67	0.61
fyn-sh3	12	0.92	1.56	0.15	4	8	0.75	0.88
gb1	10	0.54	0.80	0.30	0	10	N/A	0.70
hiv-protease-holo	56	0.71	0.90	0.24	36	20	0.64	0.95
mfabp	42	0.61	0.73	0.26	20	22	0.55	0.77
NTnC	26	0.79	1.13	0.25	2	24	1.00	0.54
plcc-sh2-free	27	0.70	0.91	0.23	3	24	0.67	0.62
protl	20	0.00	0.00	0.36	8	12	0.62	0.42
TNfn3	40	0.26	0.30	0.34	11	29	0.55	0.52
ubiquitin	32	0.64	0.88	0.25	16	16	0.81	0.75
Total	530	.	.	0.27	215	315	0.69	0.64

1 rotamer

a3D	15	-0.12	-0.19	0.48	0	15	N/A	0.27

albp	28	0.70	0.80	0.28	13	15	0.92	0.47
calmodulin	36	0.58	0.89	0.35	4	32	0.75	0.50
cdc42hs	47	0.25	0.26	0.36	23	24	0.74	0.33
cytochrome-c2-holo	31	0.40	0.37	0.26	21	10	0.86	0.30
eglinc	17	0.38	0.51	0.33	6	11	0.83	0.36
flavodoxin-holo	56	0.47	0.55	0.28	36	20	0.92	0.50
FNfn10	35	0.22	0.25	0.30	12	23	0.83	0.57
fyn-sh3	12	0.88	1.49	0.19	4	8	1.00	0.75
gb1	10	0.30	0.41	0.45	0	10	N/A	0.50
hiv-protease-holo	56	0.45	0.50	0.29	36	20	0.81	0.55
mfabp	42	0.39	0.42	0.32	20	22	0.60	0.55
NTnC	26	0.74	1.02	0.32	2	24	1.00	0.46
plcc-sh2-free	27	0.56	0.79	0.34	3	24	0.67	0.50
protl	20	-0.02	-0.03	0.35	8	12	0.62	0.58
TNfn3	40	0.17	0.17	0.39	11	29	0.73	0.28
ubiquitin	32	0.49	0.62	0.29	16	16	0.75	0.62
Total	530	.	.	0.33	215	315	0.80	0.47

Chapter 3: A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family

This chapter was adpated from a manuscript being prepared for submission.

ABSTRACT

Conformational ensembles are increasingly recognized as a useful representation to describe fundamental relationships between protein structure, dynamics and function. Here we present an ensemble of ubiquitin in solution that is created by sampling conformational space without experimental information using "Backrub" motions inspired by alternative conformations observed in sub-Angstrom resolution crystal structures, and then refined with NMR Residual Dipolar Couplings (RDCs) to select the final members of the ensemble. Using this ensemble, we probe two proposed relationships between properties of protein ensembles: (i) a link between native-state dynamics and the conformational heterogeneity observed in crystal structures, and (ii) a relation between dynamics of an individual protein and the conformational variability explored by its natural family. We show that the Backrub motional mechanism can simultaneously explore protein native state dynamics measured by RDCs, encompass the conformational variability present in ubiquitin complex structures and facilitate sampling of conformational and sequence variability matching those occurring in the ubiquitin protein family. Our results thus support an overall relation between native-state protein dynamics and conformational changes enabling sequence changes in evolution. More practically, the presented method can be applied to improve protein design predictions by accounting for intrinsic native-state dynamics.

INTRODUCTION

It has long been know that protein native states are best represented as ensembles of conformations rather than as a single structure. [10] Conformational ensembles provide a detailed structural picture of protein dynamics. As motions are crucial for many aspects of protein function, such as molecular recognition [61; 73; 74] and catalysis [48; 50; 75; 76; 77; 78], an ensemble description of proteins is also useful for improving applications of molecular modeling such as protein-small molecule [79] and protein-protein docking methods [80; 81], as well as protein design [82; 83; 84; 85; 86].

Two related concepts characterizing and interpreting properties of protein conformational ensembles have been proposed: The first suggests a correspondence between the conformational heterogeneity present in crystal structures and the native-state dynamics of proteins observed in simulations and using nuclear magnetic resonance (NMR) measurements. Several studies provide support for this idea. Zoete et al. concluded that the conformational changes present in a large number of crystal structures of HIV-1 protease reflect the inherent flexibility of the protein. [87] Vendruscolo and coworkers showed [44] that side chain relaxation order parameters, reflecting motions on the picosecond to nanosecond time scale [88; 89; 90; 91; 92; 93; 94], could be described using ensembles of crystal structures of the same protein or proteins with high sequence identity. Similarly, modeling "Backrub" motions, a type of conformational change inspired by alternate side chain and backbone conformations observed in high-resolution crystal structures [8], has led to improvements in modeling NMR side chain relaxation order parameters [9] and structural changes upon mutation [7]. Lange et al. [74] showed that ensembles derived from ensemble-averaged restrained molecular dynamics (MD) simulations of ubiquitin, using Residual Dipolar Coupling (RDC) data describing picosecond to millisecond motions [95; 96; 97; 98; 99; 100; 101; 102], encompassed conformations similar to those of ubiquitin in different crystal structures alone and in complex with different partner proteins.

The second concept proposes a link between the dynamics of a single protein and the conformational variability explored within its family of homologous proteins. This link was suggested based on the similar conformational variability observed in an MD simulation of myoglobin and in structures of different members of the globin family [103]. Similarly, Gaussian network models have suggested common dynamical features of proteins in the same family. [104; 105] Several studies extended the notion of a relationship between the dynamics of a single protein and properties of its homologs to the sequence level, showing that modeled sequences consistent with a single protein structure had characteristics in common with a multiple sequence alignment of the protein's natural family [106]. Further investigating the relation between protein dynamics and family sequence variability, other work suggested that sequence diversity and overlap between modeled and evolutionarily observed sequences could be increased by incorporating conformational flexibility of the protein backbone [82; 83; 84; 107; 108].

To combine the two concepts outlined above, here we ask whether conformational ensembles reflecting variability observed in protein crystal structures of a single sequence can be simultaneously related to experimentally determined native-state solution dynamics of an individual protein, and to the conformational and sequence variability of the protein's family. To address these questions, we investigate two related hypotheses using ubiquitin as a model system: First, we test whether ensembles generated using the

Backrub motional mechanism (“Backrub ensembles”), a model inspired by heterogeneity observed in experimental protein structures [8], capture properties of ubiquitin solution state dynamics derived from amide backbone RDC measurements in 23 alignment media. [97] Furthermore, we compare the structural variation in modeled Backrub ensembles to that seen in a set of 46 crystal structures of ubiquitin [74]. Second, we test whether the conformational heterogeneity present in Backrub ensembles that are consistent with the solution dynamics of an individual ubiquitin sequence resembles the structural diversity observed in the UQB subfamily. [109] Furthermore, we predict sequences compatible with ubiquitin Backrub ensembles using computational protein design as implemented in Rosetta [2] and test whether they are similar to the sequences of the UBQ subfamily [109].

Supporting our hypotheses, we find Backrub ensembles that are simultaneously consistent with native-state dynamics reflected in RDC measurements, the conformational variability observed in ubiquitin complex structures, and the conformational and sequence diversity of ubiquitin homologs. As an additional validation of our approach, we show that Backrub ensembles give similar agreement with the RDC data as ensembles generated from RDC-restrained MD simulations [74], and support previous observations of ubiquitin core flexibility [44] and binding by conformational selection [74]. Notably, we discover a common set of Backrub sampling parameters that are simultaneously able to best fit the RDC data and allow sampling of sequences most similar to those of the ubiquitin family. Our method to model Backrub ensembles and sequences consistent with these ensembles may thus be useful for providing insights into the relationship between native state dynamics and sequence diversity and for characterizing evolutionary sequence changes. These results also suggest that Backrub ensembles will be useful for engineering new protein functions through experimental selection from computationally designed libraries [110; 111] that contain sequences accommodated by exploiting intrinsic native-state dynamics.

RESULTS

Overall Computational Strategy

We set out to investigate the hypothesized relations between conformational changes reflecting observed heterogeneity in protein crystal structures, native-state protein dynamics and evolutionarily sampled conformational and sequence diversity in two steps (**Figure 3-1**).

First, to test hypothesis 1, we generated ensemble descriptions of ubiquitin dynamics using the Rosetta scoring function and several parameterizations of the Backrub motional model (described below) without using experimental restraints. Subsequently we selected ensembles according to their agreement with Residual Dipolar Coupling measurements (Test 1). This approach is significantly different from the methods applied earlier to find ensembles compatible with NMR restraints [14; 74; 112], which incorporated experimental data directly in the refinement process. Similar to previous work, we compare the resulting Backrub-generated conformational ensembles with an ensemble of 46 crystal structures of ubiquitin (Test 2).

Second, we use the insight gained from the comparison of Backrub ensembles with characteristics of solution-state dynamics to evaluate hypothesis 2 (**Figure 3-1**). We investigate whether Backrub ensembles that sample the conformational space available on the RDC timescale have similar conformational variability to that explored by ubiquitin homologs (Test 3). Moreover, we test whether sequences consistent with Backrub ensembles fitting RDC measurements of a single ubiquitin sequence, as predicted by computational protein design using Rosetta [2], show overlap with the sequences of the natural UQB subfamily [109] (Test 4).

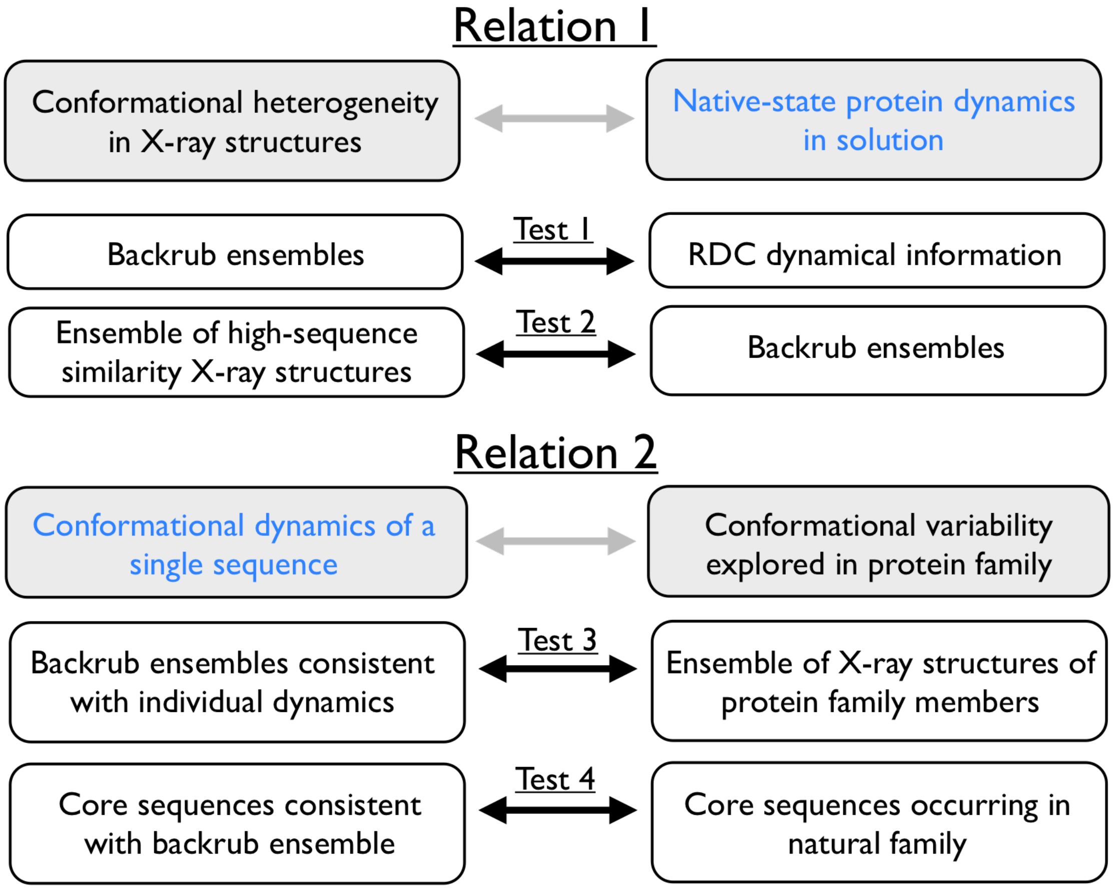

Relation 1
Conformational heterogeneity in X-ray structures
Native-state protein dynamics in solution
Backrub ensembles
Test 1
RDC dynamical information
Ensemble of high-sequence similarity X-ray structures
Test 2
Backrub ensembles
Relation 2
Conformational dynamics of a single sequence
Conformational variability explored in protein family
Backrub ensembles consistent with individual dynamics
Test 3
Ensemble of X-ray structures of protein family members
Core sequences consistent with backrub ensemble
Test 4
Core sequences occurring in natural family

Figure 3-1. Schematic describing the two main hypotheses evaluated in this work and the tests performed.

Strategy to test hypothesis 1

To test hypothesis 1, our approach first uses unrestrained conformational sampling with the Backrub motional model to generate a large set of initial conformations, starting from the ubiquitin crystal structure (pdb code 1UBQ). We use a Monte Carlo protocol consisting of rotamer changes and Backrub moves. Backrub moves involve selection of a random peptide segment, followed by a rigid body rotation of all atoms in that segment about an axis defined by the endpoint C-alpha atoms [7]. The peptide segment length is chosen at random to be either 2-3 residues (denoted in the following as “maximum segment length of 3”; **Figure 3-2a**) or between 2-12 residues (“maximum segment length of 12”; **Figure 3-2b**). 10,000 Backrub-Monte-Carlo simulations are run to generate 10,000 possible conformations in an initial set (see Methods for details). The Backrub motional mechanism thus directly accounts for correlated motions of continuous peptide segments of up to length 3 or 12. Applying these moves repeatedly in randomly chosen parts of the protein using Monte Carlo sampling allows for correlated motions of residues distant in sequence yet close in tertiary structure Correlations between side-chain and backbone dynamics have also been observed in numerous NMR studies, such as for Ribonuclease H on the relaxation time scale [113; 114] and on the RDC time scale for ubiquitin [115] and Protein G [100].

Subsequently we select from the resulting structures to form ensembles based on their agreement to the RDC measurements as measured by the Q-factor (**Figure 3-2c**), defined as:

$$Q = \sqrt{\sum_i (D^i_{\exp} - D^i_{calc})^2 / \sum_i (D^i_{\exp})^2}$$

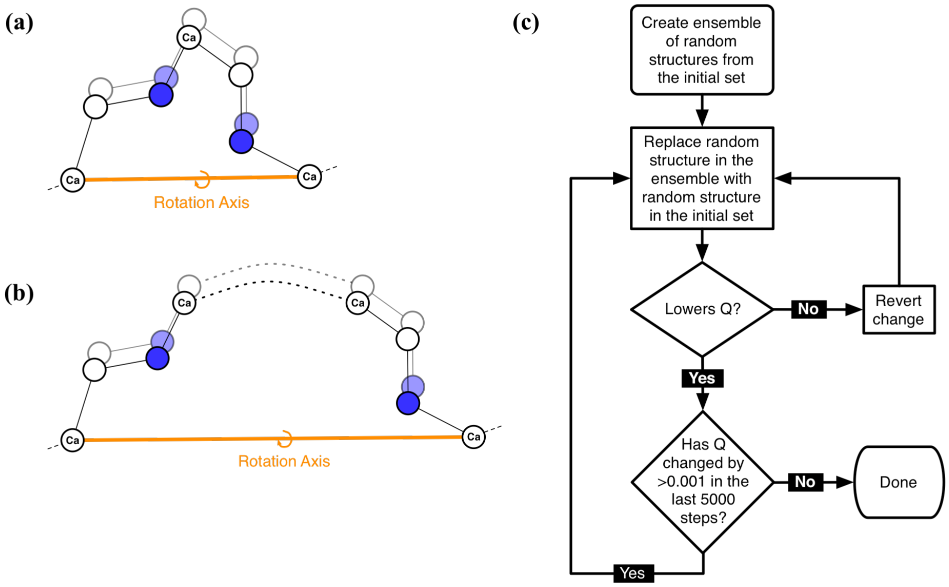

(a)
Ca
Ca
Ca
Rotation Axis
(b)
Ca
Ca
Ca
Ca
Rotation Axis
(c)
Create ensemble of random structures from the initial set
Replace random structure in the ensemble with random structure in the initial set
Lowers Q?
No
Revert change
Yes
Has Q changed by >0.001 in the last 5000 steps?
No
Done
Yes

Figure 3-2. Description of the Backrub motional mechanism and ensemble selection.

Backrub moves for (a) tripeptide segments and (b) segments of arbitrarily length from 2 through 12 residues. (c) Flowchart of the process used to select ensembles to match the RDC measurements.

A similar ensemble selection approach to the one described above has been successfully applied to model relaxation order parameters using snapshots from MD trajectories. [116] In the following sections, “selected” ensembles are defined as those undergoing the Q-factor optimization process described in **Figure 3-2c** and “non-selected” ensembles are generated by choosing random ensembles of 50 structures (without using the RDC information in selecting the ensemble members).

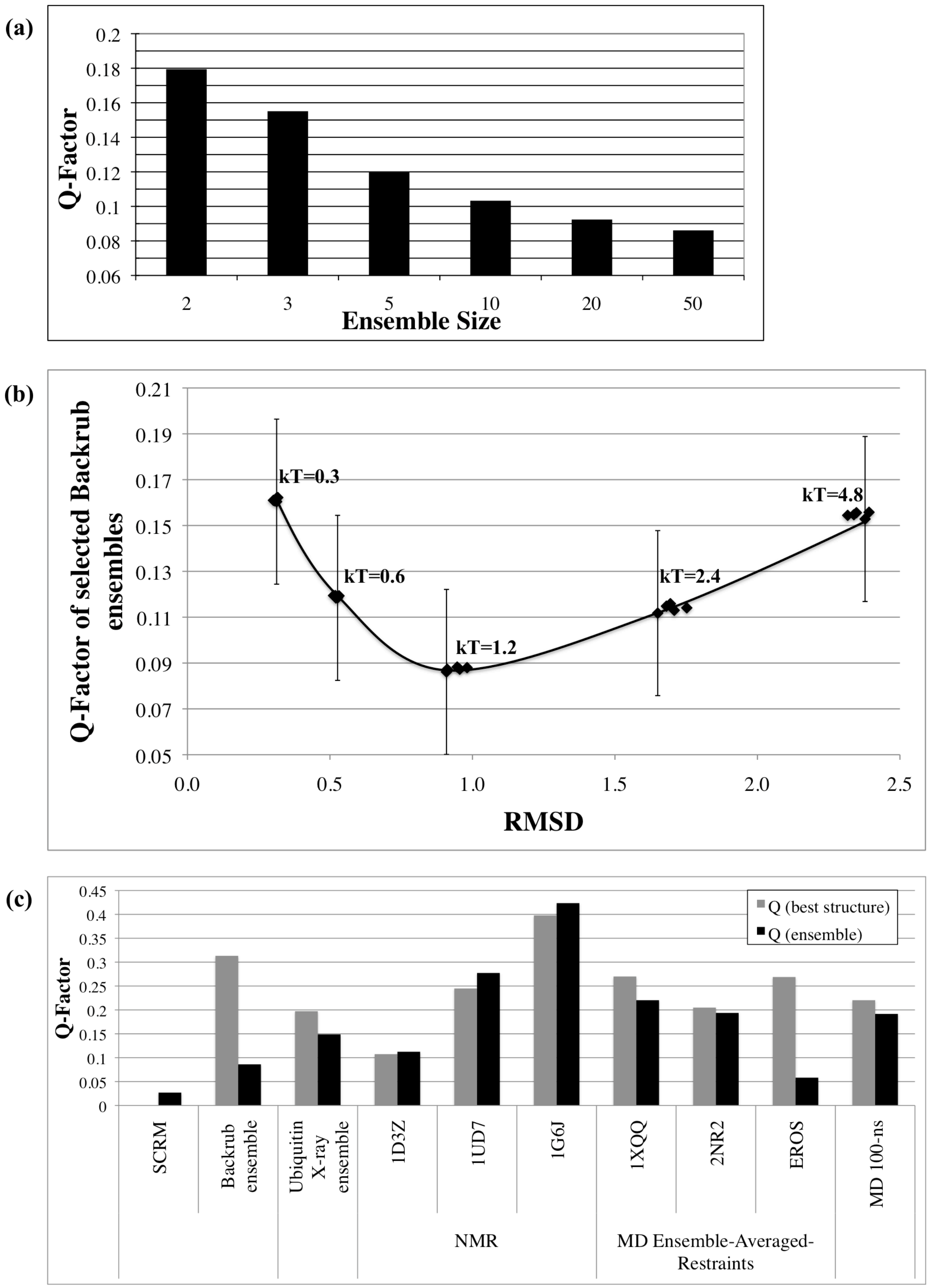
To validate our approach, we compare the Backrub-generated conformational ensembles to reference methods such as snapshots from an MD simulation in explicit solvent [117] and a set of representations of the dynamics commonly used to interpret the motional information present in RDC measurements. One such representation uses the 'model-free' formalism, which provides five parameters describing the movement of each residue. [97; 118; 119; 120] Another approach is ensemble-average-restrained (EAR) Molecular Dynamics, in which an ensemble of molecules (the "EROS" ensemble) is optimized with respect to a Molecular Mechanics force field potential in combination with ensemble-averaged restraints on the NMR measurements, including RDCs. [74] We reason that sampling methods that result in low Q-factors more closely approximate the conformational space relevant to motions on the RDC timescale than other models that describe the experimental data less well.

Correspondence between Backrub conformational ensembles and RDC measurements of ubiquitin dynamics (Test 1)

We first tested whether Q-factors of Backrub ensembles selected according to the strategy described in **Figure 3-2c** decreased as the ensemble size was increased (2, 3, 5, 10, 20 and 50 structures per ensemble). This behavior would be expected if our description captures dynamical information contained in the measurements. **Figure 3-3a** shows the Q-factors of selected ensembles of varying size generated with a Backrub maximum segment length of 12 and a simulation temperature of kT=1.2 (see Methods). There is a clear trend that the Q-factors of selected ensembles decrease as the ensemble size increases. This trend indicates that adding more structures allows a better representation of the RDC measurements and further suggests that these ensembles are representative of conformations that are populated on the timescale of the experiments (even though the Monte Carlo simulations are agnostic to timescale). This result is not simply explained by inclusion of more degrees of freedom and overfitting, as cross-validation analysis supports an ensemble size of 10 or larger (**Table 3-S2**). We use an ensemble size of 50 in the experiments below.

Figure 3-3. Backrub ensembles selected according to agreement with RDCs.

(a) Increasing Backrub ensemble size improves the agreement with the RDCs. Maximum segment length of 12 with kT=1.2. (**b)** Q factors vs. RMSD of the five selected Backrub ensembles with the lowest Q factors at each simulation temperature for maximum segment length=12. Error bars: the maximum of $Q_{\text{experimental\_uncertainty}}$ and $Q_{\text{sampling\_uncertainty}}$ (see Methods). (**c)** Q factors of the SCRM model-free description, the selected Backrub ensemble, the ubiquitin 46-member X-ray ensemble, 3 sets of NMR structures (1G6J, 1UD7, and 1D3Z), 3 Molecular Dynamics simulations with ensemble-averaged NMR restraints (1XQQ, 2NR2, and EROS), and a 100-nanosecond MD simulation. [117] For the X-ray structures, amide hydrogen atoms were added using the Rosetta molecular modeling program with an NH bond length of 1.01 Å.


(a)
Q-Factor
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
2
3
5
10
20
50
Ensemble Size
(b)
Q-Factor of selected Backrub ensembles
0.21
0.19
0.17
0.15
0.13
0.11
0.09
0.07
0.05
kT=0.3
kT=0.6
kT=1.2
kT=2.4
kT=4.8
0.0
0.5
1.0
1.5
2.0
2.5
RMSD
(c)
Q-Factor
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Q (best structure)
Q (ensemble)
SCRM
Backrub ensemble
Ubiquitin X-ray ensemble
1D3Z
1UD7
1G6J
1XQQ
2NR2
EROS
MD 100-ns
NMR
MD Ensemble-Averaged-Restraints

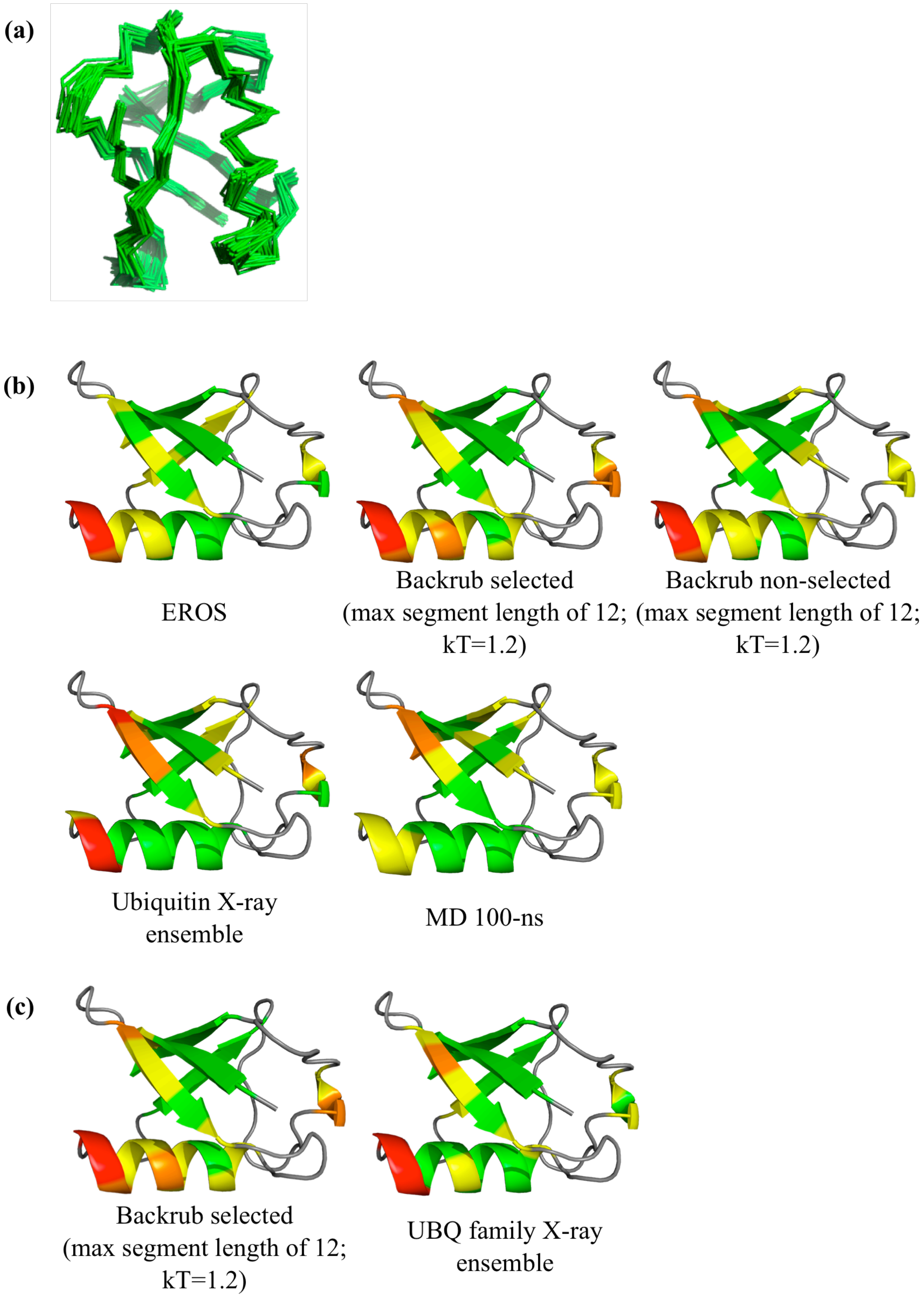
Varying the temperature and the maximum segment length affects the agreement of selected Backrub ensembles with RDC measurements

The selected Backrub ensemble described above has a Q-factor of 0.086 over regions of regular secondary structure (see Methods) and was found by comparing motional models using different Backrub sampling parameters. The first Backrub parameter we varied was the maximum segment length (as described above, the longest peptide segment rotated about an axis defined by the segment endpoint C-alpha atoms). The Backrub conformational change observed in ultra-high resolution X-ray structures consisted of concerted 2- and 3-residue Backrub moves [8]; thus we first tested a maximum segment length of 3. In a previous study [9] we showed that ensembles of structures generated using this maximum segment length improved predictions of side-chain relaxation order parameters. To test the relevance of larger-scale changes, we also tested a maximum segment length of 12 (which included moves of all intermediate segment lengths from 2-12). To measure the effect of varying the amplitude of motion, we tested a range of temperatures for the Metropolis Monte Carlo simulations from kT=0.3 to 4.8. Each simulation was run for 10,000 steps. The resulting mean pair-wise RMSDs to the ubiquitin X-ray structure of the Backrub ensembles spanned the range of 0.18Å to 0.52Å for the maximum segment length of 3 simulations, and spanned the range of 0.3Å to 3.1Å for the maximum segment length of 12 simulations (see Methods for details).

Figure 3b shows the five selected ensembles with lowest Q-factor of size 50 for different initial Backrub starting sets of 10,000 structures with maximum segment length of 12 and different simulation temperatures. For the maximum segment length of 3, the lowest Q factor is 0.089 at kT=2.4 and for the maximum segment length of 12, the lowest

Q factor is 0.086 at kT=1.2. (see **Table 3-S1** for results for all parameters) To compare these two ensembles, we performed cross-validation with four RDC datasets of N-C' couplings and four datasets of H-C' couplings (see Methods for details). The resulting R_{free} values for these ensembles were 21.3% and 18%, respectively. (**Table 3-S2**) Thus the ensemble generated using a maximum segment length of 12 appears to be a better representation of the dynamics in the RDC measurements. The optimal mean pair-wise RMSD for the ensembles with maximum segment length of 12 seems to be between 0.5Å and 1.7Å (**Figure 3-3b**).

Figures 4a illustrates the structural diversity of this ensemble. The average NH order parameter in ordered secondary structure regions is 0.76. The $S_{overall}$ scaling factor of the order parameters determined from model-free analysis of RDCs is a floating parameter that is usually estimated by comparison to relaxation experiments and here we use $S_{overall}$=0.89. [97; 118; 119]. The average NH order parameter in regular secondary structure elements is 0.76, the same as that computed for the model free analysis (0.76) described in Lakomek et al., but lower than for the EROS ensemble (0.83). [74; 97]



EROS

Backrub selected (max segment length of 12; kT=1.2)

Backrub non-selected (max segment length of 12; kT=1.2)

Ubiquitin X-ray ensemble

MD 100-ns

Backrub selected (max segment length of 12; kT=1.2)

UBQ family X-ray ensemble

Figure 3-4. Structural depictions of flexibility in various ubiquitin Backrub ensembles.

(a) Structures of the C-alpha backbone traces of a selected 50-member ensemble of maximum segment length of 12 with kT=1.2. (**b**) and **(c)** Mean C-alpha difference distance values of indicated ensembles mapped onto the 1UBQ X-ray structure. Green: 0-25% of the max value; Yellow: 25-50% of the max; Orange: 50-75% of the max; Red: 75-100% of the max; Grey: loop regions that were not included in the fit to the RDC measurements.

Selected Backrub ensembles match RDC measurements comparably to or better than other methods

Figure 3-3c compares the Q-factors of the selected Backrub ensemble to the Q-factors from various other ubiquitin ensembles: the Self-Consistent RDC-based Model-free (SCRM) description (an analytical description of the RDCs with five parameters per residue that does not provide an explicit all atom structural representation of the motions [97], an ensemble of 46 X-ray structures of ubiquitin alone and in different complexes as used in [74], three sets of NMR structures (1D3Z, 1UD7, and 1G6J), three Molecular Dynamics (MD) Ensemble-Averaged-Restraint (EAR) ensembles (1XQQ, 2NR2, EROS PDB code 2K39) [14; 74; 112] and snapshots from a 100-nanosecond MD simulation [117]. We also examined the root mean squared error in the RDCs as a measure of quality of fit, and the results were similar. (**Figure 3-S1a**) The selected Backrub ensembles have lower Q-factors than ensembles generated using several other methods, except for the SCRM description [97] and the EROS ensemble, both of which were fit with the same dataset of RDC measurements as the Backrub ensembles. Not surprisingly, the SCRM Q-factor is the lowest because it is an analytical description fit to the RDCs. The EROS ensemble was created with an approach where the RDCs are incorporated into the potential function of an ensemble MD simulation and this approach gives very low Q-factors. (See Supplemental Text and **Figure 3-S2** for an analysis of structural quality measures of backrub and other conformational ensembles). The selected Backrub ensembles also have similar R_{free} values from cross-validation: 18%, 16.1%, 20%, 17.8%, and 23.3%, respectively for the selected Backrub ensemble, the EROS ensemble, the 1D3Z structures, the ubiquitin X-ray ensemble and the ensemble of MD snapshots. (**Table 3-S2**)

One important criterion with which the various ensembles of ubiquitin can be assessed, as mentioned above, is whether an ensemble matches the RDCs better than any single structure within it. If this is the case, dynamical information contained in the experimental measurements can be interpreted by analyzing the conformational variability in the ensemble. (**Figure 3-3c**) The selected Backrub ensemble, the MD-EAR ensembles (1XQQ, 2NR2 and EROS PDB code: 2K39) and the ubiquitin X-ray ensemble have improved Q-factors over the best single structure. Of these, the two MD-EAR ensembles that were fit to relaxation NMR measurements (1XQQ and 2NR2) have the smallest fractional improvement in Q-factor, suggesting that the dynamic information present in the RDCs may be different from the information present on the shorter time scale relaxation measurements; this observation is supported by the different pattern of order parameters observed between these two classes of measurements. [97] The Backrub and the EROS ensembles show the largest fractional Q-factor improvement. Note that this does not contradict the fact that Backrub moves were able to improve modeling of faster time-scale picosecond-nanosecond side-chain motions [9]; the Backrub ensembles used in our previous work were not selected for agreement with the RDCs and the simulation temperature used was lower, resulting in smaller motional amplitudes.

The three sets of NMR structures (1D3Z, 1UD7, and 1G6J) do not show an improvement in the Q-factor over the best single structure. For the 1D3Z NMR structures, a subset of the RDCs were used in the refinement and, as a result, the Q-factor (Q=0.107; calculated over all 23 datasets used in this paper) is lower than for the other NMR structures. The Q-factor of the lowest single 1D3Z NMR structure indicates that the 1D3Z NMR structure is a good representation of the average structure, but since these structures were refined with all the restraints applied to each structure, the set of 1D3Z structures do not improve the Q factor.

We also used the strategy described in **Figure 3-2c** to generate selected ensembles consisting of structures from the various ensembles compared in **Figure 3-3c** (see **Figure 3-S1b**). The Q-factor decreased most for the ubiquitin X-ray ensemble (34% lower Q-factor), the MD-EAR ensembles (41%, 49% and 31% decrease in Q-factor for 1XQQ, 2NR2, and EROS, respectively), and the ensemble of snapshots from the 100-ns MD simulation (64% decrease). These results are consistent with the data shown above that all ensemble types except the three sets of NMR structures provide insight into the RDC dynamics. The Q factors of the selected ensembles of ubiquitin X-ray structures (Q=0.0891) and the MD snapshots (Q=0.069) are quite similar to the Q factors of the best selected Backrub ensemble. This latter result suggests that relatively long 100ns explicit water MD simulations, although short compared to the RDC timescale, may allow some parts of ubiquitin to sample conformations in agreement with the RDC measurements even though these different parts do not coincide in the same structures. This idea was suggested by Henzer-Wildman et al. [76] to explain the ability of adenylate kinase to sample substates in nanoseconds along the open-closed trajectory that exchanged on the order of micro- to milliseconds.

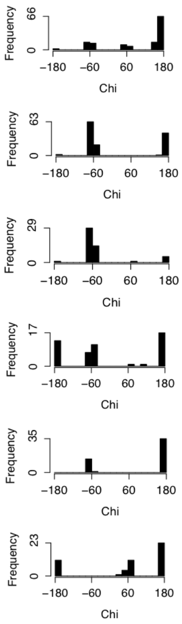
Correspondence of conformational variability in Backrub ensembles and structural heterogeneity of ubiquitin in multiple crystal structures (Test 2)

To characterize the conformational variability of different regions of the protein in our ensembles, we calculated C-alpha difference distance matrices (see Methods). [103] (**Figure 3-S3a**) These matrices show the motion of each residue in an ensemble with respect to all other residues in that ensemble. **(Figure 3-S3b)** For clarity, we collapse these matrices onto a single dimension that represents the average C-alpha difference distance relative to other residues in the protein. This metric is sensitive to motions relative to those of other residues in the ensemble, as opposed to C-alpha RMSD, which is sensitive to changes relative to one conformation in the ensemble. **Figure 3-4b** shows these C-alpha difference distance values mapped onto the ubiquitin structure and colored from green to red depending on the relative amplitude of motion (see Methods).

Supporting hypothesis 1, the pattern of motion of the ubiquitin X-ray ensemble and the selected Backrub ensemble with maximum segment length of 12 with kT=1.2 show substantial similarities. In both these ensembles the C-terminal end of the helix and the N-terminal end of beta strand 2 are the most flexible. These results are consistent with the suggestion of Lange et al. [74] that the native state dynamics of ubiquitin encompass the conformational flexibility found in crystal structures of ubiquitin bound to different partners, supporting a conformational selection model for binding. Moreover, the patterns of motions of the selected Backrub ensemble are similar to the EROS and the MD ensembles despite their different amplitudes (see Supplemental text and **Figure 3-S4** for a more detailed comparison of selected and non-selected conformational ensembles) In addition, selected and non-selected ensembles are similar to each other, at least with respect to the average Ca difference distance matrices shown in **Figure 3-4b**.

Structural and functional insights from ubiquitin conformational ensembles

We showed above that our selected Backrub ensemble (i) gives similar Q-factors to reference ensembles such as an RDC-restrained MD ensemble (EROS) [74], a ubiquitin X-ray ensemble and an ensemble of snapshots from a 100-nanosecond MD trajectory [117] and (ii) has similar regions of structural variability (**Figure 3-4b**). As an additional point of comparison and validation of our approach, we asked whether the selected Backrub ensemble also supports other structural and functional insights derived from previous ensemble descriptions of ubiquitin. Lindorff-Larsen et al. [14] as well as Richter et al. [112] used MD simulations with side chain and backbone relaxation order parameters as restraints. These ensembles displayed liquid-like flexibility of side chains buried in the protein core. The selected Backrub ensemble also has this property, with buried or near buried residues 13, 23, 44, 61, and 67 correctly modeled as flexible with calculated order parameters close to their respective values from NMR relaxation experiments. As shown in **Figure 3-5**, Ile 13 chi2, Ile 44 chi2, and Leu 67 chi2 have modeled order parameters within 0.04 of the experimental values. Ile 13 chi 1 and Ile 61 chi2 have modeled order parameters that are substantially lower than the experimental values but these differences can be due to the short timescale of the relaxation measurements compared to the longer timescale of the RDCs fit by the selected Backrub ensemble. (See **Figure 3-S5** for analysis of all flexible side chains from [14]



	S^2calc, ensemble	S^2exp, relaxation
I13 chi1	0.21	0.56
I13 chi2	0.53	0.55
I23 chi2	0.66	0.51
I44 chi2	0.27	0.31
I61 chi2	0.34	0.56
L67 chi2	0.3	0.3

Figure 3-5. Chi angle distributions of residues in or near the core of ubiquitin.

For the best selected Backrub ensemble with maximum segment length of 12 and kT=1.2, as well as modeled and experimental relaxation order parameters corresponding to these chi angles (chi1 and chi2 correspond to the Cγ and Cδ methyl groups, respectively). The Leucine Cδ methyl group relaxation order parameters were averaged.

Ubiquitin has several hotspots shown to be important in recognition of different binding partners: Ile 44, Asp 58, and His 68. These were identified as rigid in the order parameters of the EROS ensemble. [74] **Figure 3-S4g and 3b** show that these residues are also among the most rigid in the Backrub ensemble according to analysis by order parameter and C-alpha distance difference value. Likewise the secondary structure residues observed to be most flexible by order parameters calculated from the EROS ensemble are those in the N-terminal of strand 2 which our analysis also observes to be quite flexible. We find flexible regions in the C-terminus of the alpha helix that is reflected in the C-alpha distance difference value of the EROS ensemble but not in its order parameter.

Strategy to test hypothesis 2

Our results above provide support for the hypothesis of a correspondence between the properties of Backrub-derived conformational ensembles, solution-state dynamics reflected in NMR measurements and a conformational ensemble of 46 experimental crystal structures of ubiquitin. To broaden this result and shed light more generally on a link between protein dynamics and evolution, we next ask whether there is also a correspondence between the dynamics of a single protein sequence and the conformational variability explored in a protein family to accommodate sequence changes during evolution (hypothesis 2, **Figure 3-1**). In order to test this hypothesis, we first compare the conformational variability present in the selected Backrub ensemble with that observed in a structural alignment of 20 members of the UBQ subfamily (Test 3). Second, we compare sequences modeled on Backrub ensembles to the sequences of the natural UBQ subfamily (Test 4).

Individual and family conformational variation (Test 3)

To test the correspondence of the conformational variability of an individual protein and that of its family, we constructed an ensemble from the available structures of proteins in the multiple sequence alignment of the UBQ subfamily. (See Methods for details) [109] We performed a multiple structure alignment of this 20-member UBQ subfamily ensemble using MAMMOTH-mult [121] resulting in 66 positions that aligned in all proteins (see Methods) which had at most 85% and an average of 21% pair wise sequence identity. We calculated the C-alpha average distance difference matrix for these aligned positions and **Figure 3-4c** shows the average values for each residue in the matrix mapped onto the 1UBQ structure, as described for Test 2.

The resulting UBQ subfamily ensemble shows high variability in the C-terminus of the helix and in the N-terminus of beta strand 2, which is strikingly similar to the regions of high flexibility in the selected Backrub ensemble. Thus, we find similar conformational variability in the structures of the ubiquitin homologs and in an ensemble fit to the solution state dynamics of ubiquitin. This correspondence in pattern of flexibility holds despite the different motional amplitudes of these ensembles: 1.99Å and 0.9 Å pair wise RMSD to the 1UBQ X-ray structure, respectively, for the UBQ subfamily ensemble and the selected Backrub ensemble.

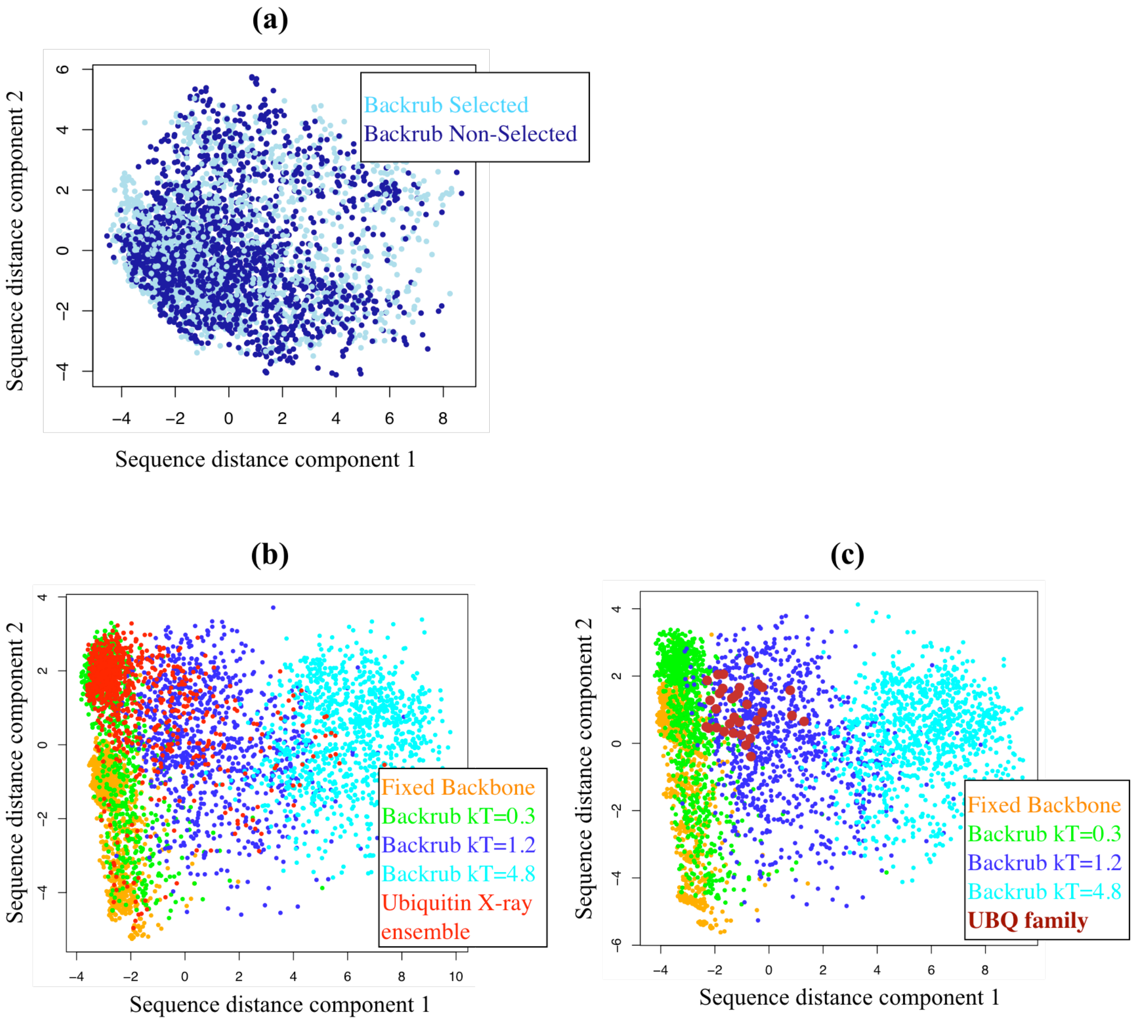
Modeling of Sequence Space (Test 4)

We proposed (hypothesis 2) and showed above that there are similarities in the conformational variability of a single protein and that of its homologs. Here we extend this idea to ask whether the sequences compatible with a structural ensemble describing the dynamics of a single protein are similar to the sequences of the natural family members. As we do not model insertions and deletions, we restrict this analysis to residues in the protein core that are in well-aligned regions in the UBQ subfamily [109]. We first tested whether there is a difference between the sequence spaces sampled by the selected and non-selected Backrub ensembles. We performed computational protein design with Rosetta [2], which used simulated annealing of rotamer conformations and amino acid identities on each backbone in an ensemble to determine the low-scoring sequences compatible with that ensemble. All positions were allowed to vary to any amino acid and 1000 low-energy sequences were generated for each ensemble. In the following, we use the term 'sequence space' to describe the high-dimensional space of possible sequences of a protein.

To compare the sequence space coverage of the various ensembles, we used the BLOSUM62 matrix [122] to calculate the distances between all pairs of sequences considering core residues only. This resulted in a distance matrix of size NxN (where N is the number of sequences compared) representing a sequence space of dimensionality N. To visualize the relative sequence space coverage of different sets of sequences we collapsed this sequence space into two dimensions using multidimensional scaling, retaining the two dimensions containing the most variation in sequence distances (see Methods).

Figure 3-6a shows that the sequence spaces sampled by the selected and non-selected Backrub ensembles with optimal Backrub parameters (maximum segment length of 12 and kT=1.2) are very similar. This is consistent with the idea that the Backrub method captures a significant portion of near-native protein motions, even without directly incorporating the RDC information into the model. In the following, we use results for non-selected ensembles; the results are similar for selected ensembles.

Next we compared the 2-D sequence space of designs on various non-selected Backrub ensembles to the sequence space of designs on the ubiquitin X-ray ensemble. **Figure 3-6b** shows that different non-selected Backrub ensembles of maximum segment length of 12 with varying amplitude (kT=0.3, 1.2 and 4.8) sample overlapping but separate sets of sequences. The lowest amplitude ensemble samples sequences closest to the fixed-backbone sequences; sequences move further away with increasing amplitude of motion in the ensemble. Notably, the Backrub sampling parameters used to generate ensembles which sample sequences most similar to the 46-member ubiquitin X-ray ensemble are the same parameters that gave the lowest Q-factor (maximum segment length of 12 with kT=1.2), supporting the hypothesis that the Backrub ensembles are sampling similar conformational heterogeneity to the ensemble of ubiquitin X-ray structures (Test 2).


(a)
Backrub Selected
Backrub Non-Selected
Sequence distance component 2
Sequence distance component 1
–4
–2
0
2
4
6
8
(b)
Fixed Backbone
Backrub kT=0.3
Backrub kT=1.2
Backrub kT=4.8
Ubiquitin X-ray ensemble
Sequence distance component 2
Sequence distance component 1
(c)
Fixed Backbone
Backrub kT=0.3
Backrub kT=1.2
Backrub kT=4.8
UBQ family
Sequence distance component 2
Sequence distance component 1

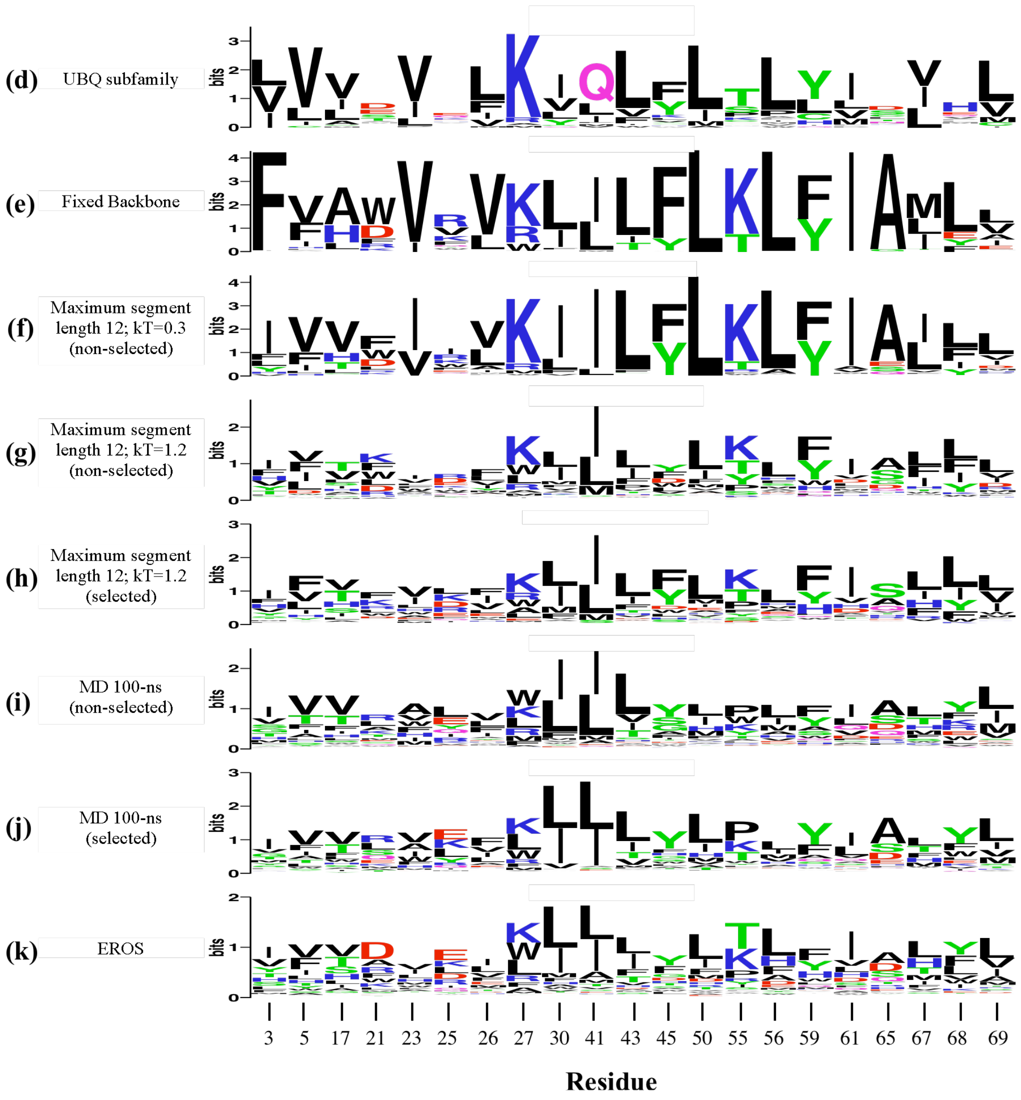


Figure 3-6. Sampling of sequence space by computational design on various backbone models.

(**a**) Designed sequences on non-selected (dark blue), and selected (light blue) Backrub ensembles of maximum segment length of 12 with kT=1.2. (**b**) and (**c**): Low-scoring designed sequences on the fixed backbone of the X-ray structure 1UBQ (orange); on non-selected Backrub ensembles with maximum segment length of 12 with kT=0.3 (green), kT=1.2 (blue), and kT=4.8 (cyan); and **(b)** Low-scoring designed sequences on the ubiquitin X-ray ensemble (red), or **(c)** sequences from the UBQ family (Kiel and Serrano 2006) (brown). (Note that the dimensions shown in the plots are selected to maximize the variation of the points in each plot and will differ between plots) Sequence logo plots for **(d)** the UBQ subfamily, and low-scoring designed sequences on **(e)** the 1UBQ fixed backbone, **(f)** the non-selected ensemble created with maximum segment length of 12 and kT=0.3, and **(g)** the non-selected and **(h)** selected ensembles with maximum segment length of 12 and kT=1.2. Designed sequences on **(i)** non-selected and **(j)** selected ensembles from a Molecular Dynamics trajectory of 100-nanoseconds. **(k)** Designed sequences on the EROS ensemble.

Finally, to test whether there exists a link between the conformational heterogeneity of solution dynamical ensembles and the sequence space compatible with these ensembles (Test 4), we compared the 2-D sequence space of designs on various Backrub ensembles to the sequence space of the UBQ subfamily of the ubiquitin ab roll subfold. (**Figure 3-6c**) The subfamily sequences we used came from a high quality manually curated alignment of 36 homologues created using 3D structural analysis. [109] As shown in **Figure 3-6c**, the sequences of core residues in these naturally occurring proteins represent a subset of the core residue sequence space of the non-selected Backrub ensemble (maximum segment length of 12 with kT=1.2). In contrast, the UBQ subfamily sequences barely overlap with the sequences from the fixed backbone designs or the kT=0.3 designs, and do not overlap with the designs using the backrub ensemble generated with kT=4.8.

The sequence logo representations in **Figure 3-6d-k** support the correspondence between the sequence diversity in Backrub ensembles and the natural family. The predominant amino acid in the UBQ subfamily is recapitulated in the non-selected Backrub ensembles of maximum segment length 12 with kT=0.3 and kT=1.2 (e.g. positions 5, 27, 43, 50, 56, 61, and 69). One notable exception is that the designed sequences fail to recapitulate the conserved glutamine at position 41. Kiel et al. [109] use this position as the main indicator in categorizing subgroups of subfamily UBQ because its presence correlates with the structure of a nearby loop. The side chain amide nitrogen atom of Gln 41 forms a buried hydrogen bond with the backbone of residue 36, which may be responsible for structural specificity of the loop conformation that we are not accounting for in the design simulations. Several positions, such as residues 21, 25, 45, 55, 61, 65, and 68, have high sequence entropy in the natural family. The Backrub ensemble designs recapitulate high sequence entropy for these residues. Especially for residues 45, 55, 61, and 65 the high entropy underscores one of the uses of flexible backbone design, as with a fixed backbone or low temperature Backrub ensemble only a few amino acid types predominate at those positions failing to capture the substantial natural sequence plasticity within the family. We also generated designs compatible with the non-selected and selected ensembles from the trajectory of the 100-ns MD simulation, noting that these sequences showed similar results to the kT=1.2 Backrub ensembles overall, but with higher sequence entropy for several positions.

Taken together, our results thus indicate that the conformational sampling methods we use here to match RDC dynamics produce variability similar to the conformational heterogeneity of X-ray ensembles (both using different ubiquitin structures as well as structures from the UBQ subfamily) and may lead to significant overlap between sequences consistent with modeled ensembles and the sequence space covered by the natural family. Additionally, it appears from the similarity of sequences from selected and non-selected ensembles that the RDCs have led us to determine optimal Backrub sampling parameters (**Figure 3-3b**) that can be used prospectively to make modeling predictions.

DISCUSSION

In this work, we describe the application of the Backrub motional model to create ensembles of structures consistent with RDC measurements and to sample the conformational and sequence space of the UBQ subfamily.

The main new aspect of our work is that we link the conformational dynamics of a single sequence, as reflected by both RDC data and Backrub ensembles, to conformational diversity observed in crystal structures of ubiquitin and its family, and to evolutionary sampled sequence diversity. We achieve this by applying computational protein design to select low-energy sequences consistent with Backrub ensembles. The fact that low-Q factor Backrub ensembles sample a similar sequence space to that of the ubiquitin X-ray ensemble extends results by other groups demonstrating the correspondence of solution-state dynamics and crystallographic heterogeneity. [44; 97] In addition, we find that this designed sequence space consistent with optimal Backrub ensembles encompasses the sequence space of the UBQ subfamily, providing evidence for the idea suggested by Davis et al. [8] that the Backrub motional mechanism may facilitate amino acid changes during evolution.

We find that selected ensembles created with only certain Backrub sampling parameters were able to reach the lowest Q-factors, indicating that the conformational space sampled by these Backrub parameters is the most similar (compared to other parameters) to the conformations giving rise to the RDC measurements. However, while we see significant improvements in Q-factors during the selection protocol, we also find substantial similarities between selected and non-selected Backrub ensembles, in patterns of C-alpha RMSD, order parameters and designed sequence space. This somewhat surprising observation could mean that the selection procedure primarily optimizes for subtle differences in NH-vector orientations, while other dynamical features that are commonly characterized (such as the anisotropy of motions) are essentially indistinguishable between selected and non-selected Backrub ensembles. Analysis by cross-validation shows an improvement in R_{free} for selected over non-selected ensembles, indicating that other aspects of the peptide plane orientation are better represented in the selected ensembles. Notably, there are defined Backrub parameters that simultaneously give the best agreement with the RDC data (after selection) and the best sequence space overlap with the natural family, irrespective of whether we apply selection or not. This could indicate that it is primarily the mechanism and amplitude of motions that is important, and that, as long as the amplitude is in the correct range defined by the appropriate sampling parameters, the Backrub motional model can sample relevant motions without the direct requirement of the RDC data. Hence, the Backrub motional model may be useful (i) to predictively sample conformations similar to ensembles of bound conformations and (ii) to use with design to sample the sequence space of the natural family. Such sampling of sequences likely to be accommodated by a given protein fold may help improve engineering of new protein structures, functions and interactions. For example, coupling backbone ensemble generation and sequence design may be useful to computationally predict sequence libraries enriched in functional members. [111]

There are several potential limitations of the Backrub method, as applied here. As we implement Backrub in a Monte Carlo protocol, the timescale of conformational transitions is not taken into account. Also, the method used here limits the backbone conformational space sampled to those conformations accessible with the Backrub mechanism, a restriction which can be alleviated for example with the addition of small phi/psi changes to the method or by using analytical methods for local loop closure [123], which is a superset of the Backrub move. Nevertheless, Backrub changes have an interesting similarity to the 1D-Gaussian Axial Fluctuation (GAF) analytical model, a simple motional model that has been used with success to model RDCs [46]. A dipeptide Backrub move (a tripeptide Backrub move is shown in **Figure 3-2a**) is essentially the same as the 1D-GAF model. Thus the Backrub Monte Carlo protocol, which includes moves of longer peptide segments incorporated into a Monte Carlo scheme, can be viewed as a generalization of the GAF model.

As necessitated by the scarcity of proteins with sufficient RDC data, we limit our study here to one protein and further work is needed to extend modeling of protein native state dynamics and tolerated sequence space to more proteins. However, the usefulness of the Backrub mechanism for modeling protein motions is supported by several studies [7; 8; 9; 124; 125]. Our studies on ubiquitin provide an interesting benchmark case for future analyses of the correspondence of individual and family variation.

Analysis of the generated ubiquitin Backrub ensembles allows several fundamental insights on the relationship between structure, function, sequence and dynamics. The ubiquitin core flexibility and a binding mechanism by conformational selection have been pointed out previously. [14; 74] Furthermore, our study allows characterization of differences between computationally predicted and evolved protein sequences that may lead to testable hypotheses on effects not modeled in the simulations, such as evolutionary pressures to conserve functional residues. An example is the discrepancy between the predictions and the naturally occurring glutamine residue at position 41 in ubiquitin. A likely explanation why our design simulations fail to predict this preference for glutamine is that we are not taking into account avoidance of certain non-native conformations due to evolutionary pressure enforcing structural specificity.

In conclusion, we have tested a method for sampling conformational diversity using Backrub conformational changes and shown that it can generate ensembles consistent with millisecond-timescale measurements of protein dynamics. This method is computationally more efficient than Molecular Dynamics-based methods, allowing it to be applied to a variety of protein modeling tasks such as sequence design. Notably, we find that the method recapitulated many of the structural properties of the selected Backrub ensembles even when the RDC measurements were not incorporated in the sampling procedure. We additionally find that the sequence diversity of non-selected Backrub ensembles is similar to that of both the ubiquitin X-ray ensemble and the UBQ subfamily X-ray ensemble. This result needs to be tested on more proteins and, if validated, should be useful in making prospective predictions to numerous applications, such as protein-protein or protein-small-molecule docking, protein interface design, and enzyme design.

MATERIALS AND METHODS

Residual Dipolar Coupling Measurements

The dataset of RDCs we use here consist of measurements in 23 alignment media as described in Lakomek et al. [97].

Structure processing

For all X-ray structures, explicit hydrogen atoms were added according to standard geometry using Rosetta, and the positions of hydrogens with rotatable bonds were optimized. [65] The 46-member ubiquitin X-ray ensemble used was the same as that of [74].

Generation of conformational ensembles

To generate protein conformational ensembles, we ran “Backrub” Monte Carlo simulations, as described in [9] and [7]. Briefly, this method randomly makes one of three types of moves: (a) a rotamer change (50% of the time) (b) a local backbone conformational changes (Backrub move) consisting of a rigid body rotation of a random peptide segment about the axis connecting the endpoint C-alpha atoms (25% of the time), or (c) a composite move with a Backrub change and one or two rotamer changes (25% of the time). After each move, the positions of the C-beta and H-alpha atoms are modified to minimize bond angle strain as described. [7] This results in bond angle changes of the main chain atoms of one to four standard deviations. The mean values and standard deviations are very similar to those computed in a set of 240 high-resolution crystal structures (better than 1.3Å) with less than 25% sequence identity culled from the Dunbrack database [126], except for some perturbation to the N-CA-C angle (mean and standard deviations are 111.49 and 4.08 in the Backrub ensembles and 110.98 and 2.46 in the crystal structure set). See **Figure 3-S2** for details on the structural quality analysis for all structures and ensembles used in this study.

We ran a Backrub Monte Carlo simulation at kT=0.1 from the starting PDB conformation (using 1UBQ, which has the highest resolution (1.8Å) of the unbound ubiquitin structures; similar results were obtained for maximum segment length of 3 with PDB entries 1UBI and 1CMX and worse Q factors were obtained for PDB entries 1FXT, 1AAR, 1F9J, and 1TBE) for 10,000 steps with a maximum segment length of 3 or 12, matching the segment length used later. The lowest energy structure from this simulation is used as the starting conformation for 10,000 randomly seeded Backrub simulations at one of 5 different temperatures (kT=0.3, 0.6, 1.2, 2.4, or 4.8) run for an additional 10,000 steps. The last structure from each of these simulations is used to form the starting set of 10,000 structures.

From this initial set of 10,000 structures, ensembles are selected to match the RDCs by minimizing the Q-factor of the ensemble. First, structures are randomly chosen to create a starting ensemble of a given size (2, 3, 5, 10, 20 or 50 structures), and the Q-factor of the ensemble is calculated (see below). Next, a random structure in this ensemble is chosen and replaced with a randomly chosen structure from the initial ensemble of 10,000 structures; then the new Q-factor of the ensemble is calculated. If the new Q-factor is lower than before the replacement, the change is kept, otherwise it is reverted. These structure replacements are repeated until the Q-factor changes by less than 0.001 in 5000 steps. By repeating this method 1000 times, 1000 selected Backrub ensembles are created. There are a very large number of possible subsets of a given size. For example, there are 4*10^61 different sub-ensembles of size 20 from the initial ensemble of size 10,000, too many to be evaluated. The approach described here does not guarantee that the ensemble with the lowest Q-factor will be found, but it starts from many random starting points to broadly sample the space of possible sub-ensembles and the selection process converges to a low Q-factor solution within 10,000 Backrub-generated structures for all Backrub Monte Carlo temperatures (except kT=4.8) (**Supp Figure 3-7**)

Calculating RDCs from a structure or structural ensemble

RDCs are calculated from a single structure and an ensemble of structures as described in [127]. Briefly, we first find the alignment tensor from a structure (or set of structures) and the experimental couplings. This is done using the equation T = A^{-1} D_{exp}, where T is the alignment tensor, A^{-1} is the Moore-Penrose inverted matrix of projection angles for the amide bonds (or averaged projection angles for a set of structures), and D_{exp} is the vector of experimental couplings. The predicted couplings are then calculated with the equation D_{calc} = AT where A is the same matrix of projection angles from above and D_{calc} is the vector of calculated couplings.

Q-factors were calculated for all RDC measurements with the equation:

$$Q = \sqrt{\sum_i (D^i_{\exp} - D^i_{calc})^2 / \sum_i (D^i_{\exp})^2}$$

Errors between experimental and predicted RDCs were calculated with:

$$D_{error} = \sqrt{\sum_i (D^i_{\exp} - D^i_{calc})^2 / N}$$

Loop residues (i.e. those with DSSP [128] secondary structure type not H, E, G or I) are excluded from the analysis in both tensor determination and back-computation of RDCs and Q-values. The non-loop residues used in all analyses in this paper are ubiquitin residues 2-7, 12-16, 23-34, 38-45, 48-49, 57-59, and 66-71.

Sources of error

There are several sources of error in our analysis to consider when assessing the significance of the results. First, there is error in the RDC measurements due to experimental uncertainty. The uncertainty in these values is estimated to be 0.3Hz [97]. To calculate the resulting uncertainty in the Q-factor, we added Gaussian-distributed noise of mean amplitude 0.3Hz to the RDC measurements (see section below) in 1000 Monte Carlo trials. This resulted in a value of $Q_{experimental\_error}$=0.036.

A second source of error results from not finding the ensemble with the lowest possible Q-factor from a given initial structure set. We estimated this error by repeating the selection procedure many times and evaluating the variance in the resulting Q-factors. We take explicit steps to minimize this error by enforcing two convergence criteria on the optimization: 1) ensemble selection is not finished until 5000 steps have passed without a change in Q of more than 0.001, and 2) enough selected ensembles are generated from random starting structures such that the difference in the Q-factors of the best and 10th best selected ensemble is not more than 0.005. Thus, this $Q_{optimization\_error}$ is on the order of 0.005.

A third important source of error is due to insufficient sampling of conformational space with the Backrub Monte Carlo protocol and the 10,000 structures that we use to select ensembles from. We estimated this $Q_{sampling\_error}$ by running the structure generation protocol at each temperature 10 times, thus creating 10 sets of 10,000 Backrub-generated structures at each temperature. The standard deviations of the minimum Q-factors over these 10 sets of 10,000 structures are 0.0151, 0.0104, 0.0025, 0.0039, and 0.0049 for kT=0.3, 0.6, 1.2, 2.4 and 4.8, respectively for a maximum segment length of 12. The standard errors of the mean of these values are 0.0048, 0.0033, 0.0008, 0.0012, and 0.0015, respectively.

Calculation of the experimental uncertainty in the Q-factor ($Q_{experimental\_uncertainty}$)

Gaussian-distributed noise was added to the experimental RDCs with 1000 Monte-Carlo samples. The RDC uncertainty of each measurement was 0.3Hz, [97] which was used as the standard deviation of the Gaussian noise function. The resulting $Q_{experimental\_uncertainty}$ is 0.036 with a standard deviation of 0.001018 over the 1000 samples.

Order parameter calculation

Order parameters were calculated with the equation

$$S^2 = \frac{3}{2}\left[\langle x^2\rangle^2 + \langle y^2\rangle^2 + \langle z^2\rangle^2 + 2\langle xy\rangle^2 + 2\langle xz\rangle^2 + 2\langle yz\rangle^2\right] - \frac{1}{2}$$

where x, y and z are the coordinates of the normalized unit vectors representing the amide bond vector orientations. [55] For the Backrub ensemble, these values were then scaled by 1/ 1.12 = 0.89 to account for librational effects.

Molecular Dynamics trajectory

We used the 100-nanosecond AMBER trajectory of ubiquitin in TIP4Pw/e water from Wong and Case 2008. [117] The protein was allowed to equilibrate over the first 4.32 nanoseconds, and snapshots were taken from the following 100 nanoseconds at 10-picosecond intervals. This resulted in 10,000 structures, which were used to calculate an overall Q-factor for the trajectory. In addition, we applied the selection scheme in **Figure 3-2c** on these 10,000 snapshot structures to select ensembles with optimized Q-factors.

Measurement of sequence space sampling

To estimate the sequence space represented by different structures and ensembles, we used Rosetta computational protein design to generate 1000 low-energy sequences for each single structure or 20 sequences per ensemble member for ensembles of size 50. To find a low-scoring sequence, each design simulation consists of 20 rounds of Monte Carlo simulated annealing with the number of steps in each round equal to the number of rotamers created for the simulation. The backbone of each structure or ensemble member is kept fixed during the design simulations and all positions were allowed to vary to any of the 20 naturally occurring amino acids, adding extra conformers at one standard deviation around the mean rotamer for chi 1 and 2 dihedral angles. The scoring function used was the Rosetta all-atom scoring function [2], which is dominated by a Lennard-Jones potential, an explicit hydrogen-bonding potential [65] and an implicit solvation potential [66].

Distances between sequences were calculated as in [107]. Briefly, these distances were calculated as the sum of the substitution costs (using the BLOSUM62 matrix after normalizing it to range from 0 to 1) [122] for the positions that aligned and were in the core (defined below). After calculating the distances between all pairs of sequences within each ensemble and between pairs of ensembles, we used metric multidimensional scaling in R [129] to reduce the dimensionality of the space to the two dimensions spanning the most sequence distance.

Core residues were defined by counting the number of neighbor residues with C-beta atoms within 10Å of the C-beta atom of the residue of interest (or C-alpha atoms for glycine). The cutoff value used (greater than or equal to 18) was chosen so that approximately one third of the residues fell into the core category (excluding the flexible C-terminus), resulting in 21 buried positions: residues 3, 5, 17, 21, 23, 25, 26, 27, 30, 41, 43, 45, 50, 55, 56, 59, 61, 65, 67, 68, and 69.

C-alpha difference distance matrices

First, for each structure, we calculated the matrix of distances between all C-alpha atoms. Then, for each pair of structures, we calculated the distance difference matrix as the absolute value of the difference of the distance matrices of the structures. These distance difference matrices were averaged to give the C-alpha difference distance matrix of the ensemble. [103]

UBQ subfamily structural alignment

To create a structural ensemble for the UBQ subfamily we took the highest resolution X-ray structure for each protein listed in Table 1 of Kiel et al. [109] (or the first structure of an NMR ensemble if no X-ray structure was available). We removed structures that had 100% sequence identity to other structures in the ensemble. We performed a multiple structural alignment using MAMMOTH-mult [121] and removed PDB id 1WIA because it was missing residues that aligned with part of the helix in the native ubiquitin sequence; all other structures had residues that aligned with all the residues in the secondary structure regions of ubiquitin. The resulting ensemble consisted of 20 structures: 1XD3 chain B, 1BT0 chain A, 1EUV chain B, 1IYF, 1J8C, 1LM8 chain B, 1M94, 1NDD chain A, 1OQY, 1P1A, 1TGZ chain B, 1V5O, 1V5T, 1V86, 1WE6, 1WE7, 1WGD, 1WGG, 1WH3, and 1WM3 chain A. To create the C-alpha distance difference matrix we used the 66 positions that aligned in all 20 structures, which were (using 1UBQ numbering): 1-7, 9-16, 18-34, 36-46, 48-55, 57-64, 66-72.

Cross-Validation

We performed cross-validation by using the alignment tensor calculated from the NH RDC datasets to calculate RDCs for four datasets of NC' RDC couplings and four datasets of HC' couplings. These "free" data were not included in the selection process and are reported as R_{free} factors, as calculated by Lange et al. [74]

$$R_{free} = \sqrt{\sum_{i}^{N} n_i Q_i^2 / (2 \sum_{j}^{N} n_j)}$$

for the N different types of experiments with n_i measurements each and Q-factor Q_i. For selected Backrub ensembles, the R_{free} values are averaged over the five lowest-Q factor ensembles.

ACKNOWLEDGEMENTS

We would like to thank Mariana Babor for computational design simulations, Ryan Ritterson for initial results on individual and family conformational variability, Korvin F.A. Walter for help with the measurements of RDCs, Chris Saunders for sharing data on ubiquitin alignments and modeled sequence diversity, David Case for providing the 100-nanosecond Molecular Dynamics trajectory of ubiquitin, and Ian Davis and Vincent Chen for help with MolProbity. G.D.F. was supported by an NSF Graduate Research Fellowship. T.K. is an Alfred P. Sloan Foundation Fellow in Molecular Biology. This work was supported by an NSF CAREER grant (MCB-0744541) to T.K.

AUTHOR CONTRIBUTIONS

GDF, JM and TK conceived and designed the experiments. GDF performed the experiments. GDF, JM and TK analyzed the data. GDF and TK wrote the paper, in consultation with JM, NAL and CG. NAL and CG contributed reagents/materials/analysis tools.

SUPPORTING INFORMATION

SUPPLEMENTARY RESULTS

Structure quality of Backrub ensembles

In order to evaluate structural quality parameters of the different conformational ensembles, we used MolProbity. [130] The structure quality metrics of the selected Backrub ensemble is generally within the range of values of the X-ray and NMR structures and other ubiquitin ensembles. (**Figure 3-S2**) For example, less than 5% of dihedrals are outside of the favored Ramachandran region in the Backrub ensembles, compared to 0-14% for various X-ray structures, 0-8% for NMR structures, 6-8% for the MD ensemble-averaged-restrained structures, and 4% for the MD structures. Some distortions are observed when the Backrub simulation temperature is increased to kT=4.8 for maximum segment length of 3 and kT=2.4 for maximum segment length of 12, where the number of steric clashes, the occurrence of residues in the non-favored Ramachandran regions and the fractional volume increase over the crystal structure have higher values than typical for the other ensemble types. Nevertheless, the Backrub ensembles that fit the RDC data best (**Figure 3-3c** in the main manuscript) appear to have reasonable geometries.

Differences between selected and non-selected Backrub ensembles

The Q-factors of selected ensembles are substantially lower than the Q-factors of the non-selected ensembles (ranging from Q=0.081 to 0.15 and Q=0.27 to 0.35, respectively, for different Backrub ensembles with maximum segment length of 3; and from Q=0.086 to 0.15 and Q=0.25 to 0.53, respectively, for maximum segment length of 12). Thus, we investigated a range of structural and dynamical parameters to characterize differences between selected and non-selected ensembles. **Supp Figures 4a-f** show the C-alpha mean pair-wise RMSD of selected and non-selected Backrub ensembles of different maximum segment length and amplitude of motion. Interestingly, for the Backrub sampling parameters that yield the lowest Q-factors after selection (kT=2.4 with maximum segment length of 3 and kT=1.2 with maximum segment length of 12; see **Figures 3b**, **4d** and **4g** in the main manuscript), the patterns of C-alpha variation of the selected and non-selected ensembles are comparable even while the amplitude is different. (**Supp Figures 4a-f**) In addition, the pattern of motion using a C-alpha difference distance analysis (see Methods and main manuscript **Figure 3-4c**) is also similar between the selected and the non-selected ensembles. (**Figure 3-S3a & b**)

The order parameters between non-selected and selected ensembles are generally similar as well (**Figure 3-S4g**), apart from some smaller differences in the helix for maximum segment length of 3 with kT=1.2 and differences in beta strand 4 of ensembles from both maximum segment lengths. Nevertheless, the correspondence in C-alpha RMSD and overall order parameter patterns between selected and non-selected ensembles suggest that more subtle differences – i.e. the character of the motion and not its amplitude – account for the significant differences in Q-factors.

Thus we looked in detail at the effect of ensemble selection on the properties of the amide bond vectors. **Supp Figures 6a** and **6b** show the difference in angle of the average amide bond vector orientations of Backrub ensembles relative to the average amide bond vector orientations in the 1D3Z ensemble (which was also fit to subset of the RDC data). Looking at the change in this angular difference from non-selected to selected ensembles (**Figure 3-S6c**) the orientations change in the selected ensembles and move closer to the orientations in the 1D3Z ensemble. These angular differences are also more similar between the two selected ensembles with different maximum segment length (R^2=0.68; **Figure 3-S6d**) than between the two non-selected ensembles (R^2=0.42; **Figure 3-S6e**).

SUPPLEMENTARY METHODS

Structure quality analysis

MolProbity [130] was used for analysis of the following structural quality metrics: the number of clashes greater than 0.4Å, orientation of C-beta atoms, rotamer conformations with less that 1% frequency of occurrence in the PDB, phi/psi dihedral angles in the "core", "allowed" and "outlier" regions (defined as the top 98% of residues, the top 99.5% of residues, and the remaining residues, respectively) [131], and bond lengths and angles of heavy main-chain atoms that are more than 4 standard deviations from their expected values (Vincent Chen, personal communication). The calc-volume tool [132; 133] was used to calculate the packing volume of the various structures. The values shown for ensembles are the means over the structures in the ensemble.

Analysis of amide vector orientation

To plot the amide vector orientations in the Backrub ensembles, we used the NH orientations from the 1D3Z NMR structures as a reference. For each residue we calculated the average orientation of the NH vectors in the 1D3Z structures and the average orientation of the vectors in the Backrub ensemble. We then calculated the angle between these average vectors.

SUPPLEMENTARY FIGURES

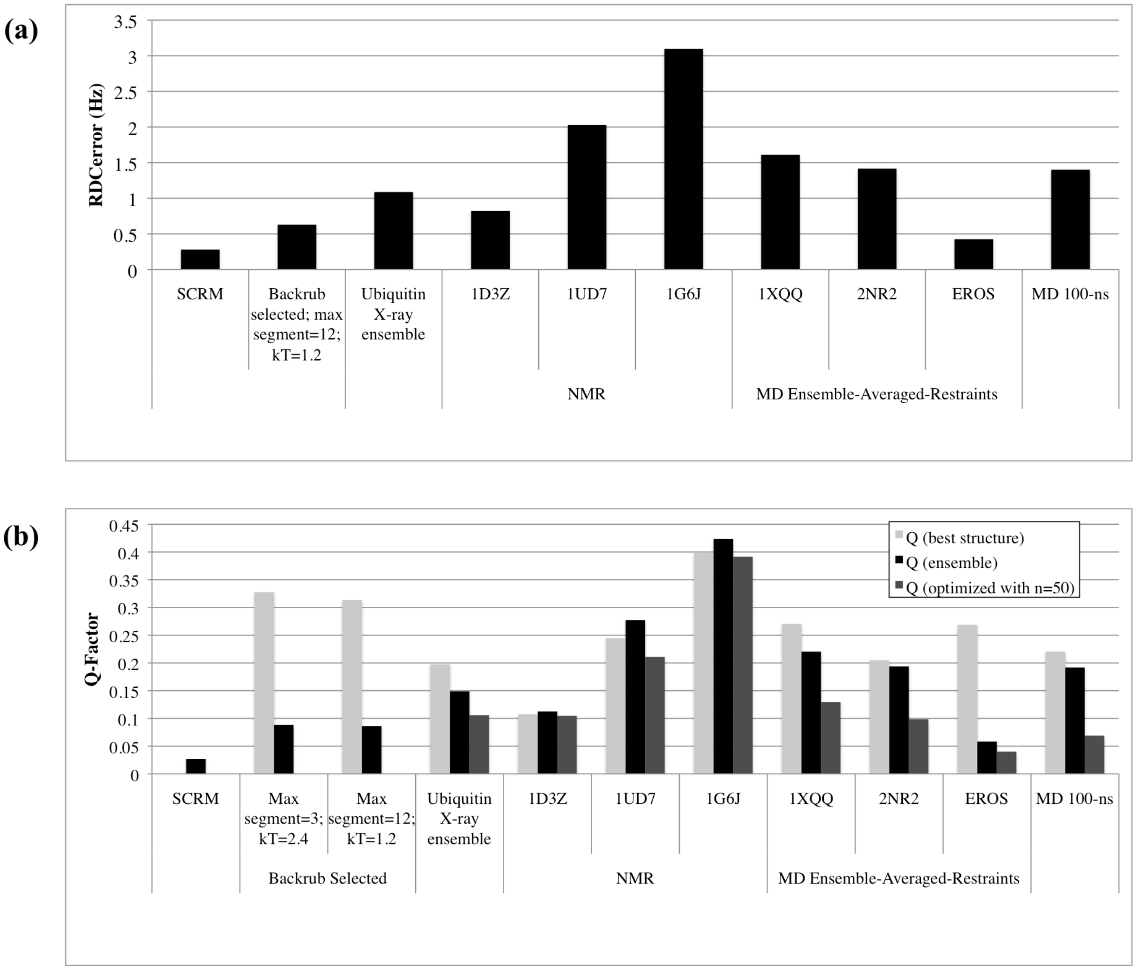


Figure 3-S1. Errors and Q-Factors of various models of ubiquitin flexibility.

(a) Error in the calculated RDCs for various ensembles. **(b)** Backrub ensembles fit the RDCs better than most other computational and experimental ensembles. Same data as **Figure 3-3c** in the main manuscript with the addition of the yellow bars: Minimum Q factors of ensembles of size 50 (allowing multiple instances of the same structure) selected from structures from the given source using the optimization approach in **Figure 3-2c** in the main manuscript.

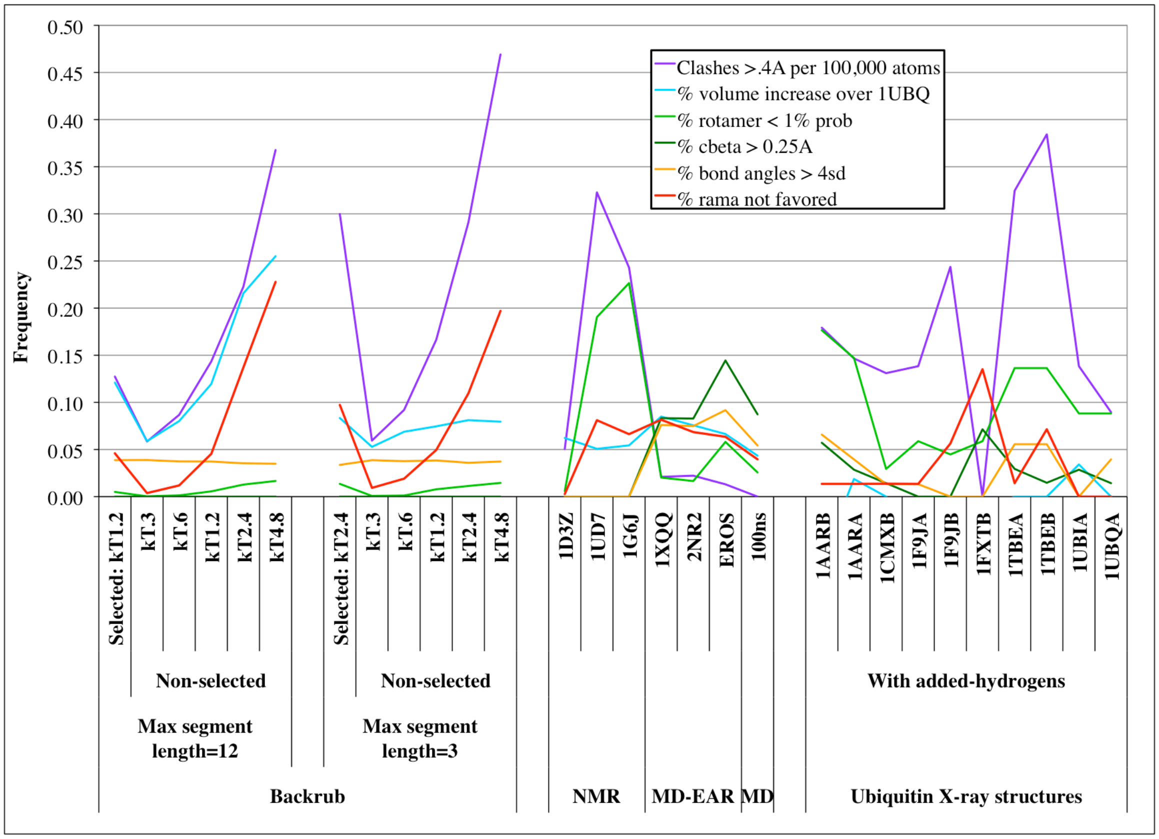
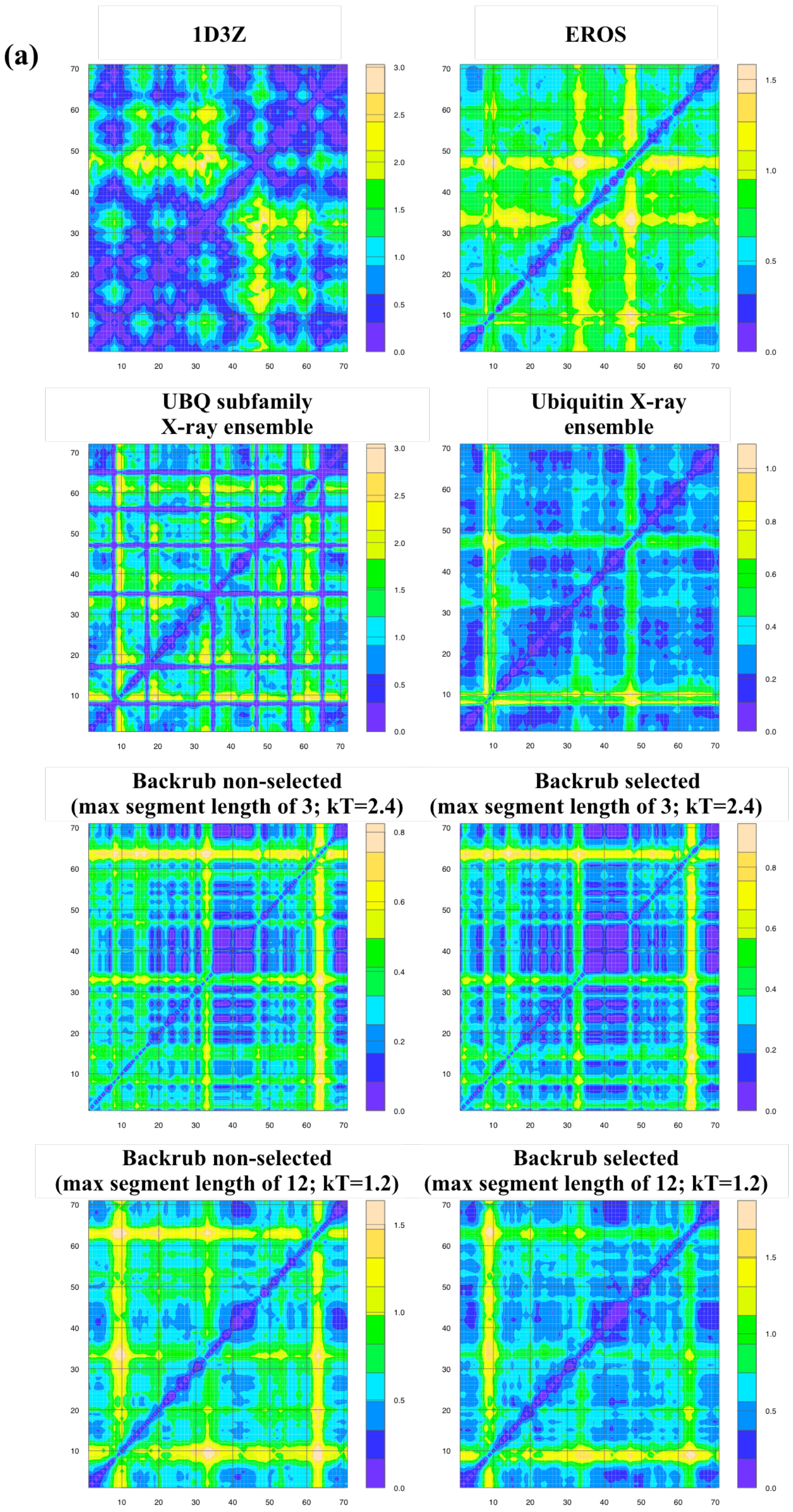


Figure 3-S2. Stereochemistry of Backrub and other ensembles.



(a)

1D3Z

EROS

UBQ subfamily X-ray ensemble

Ubiquitin X-ray ensemble

Backrub non-selected (max segment length of 3; kT=2.4)

Backrub selected (max segment length of 3; kT=2.4)

Backrub non-selected (max segment length of 12; kT=1.2)

Backrub selected (max segment length of 12; kT=1.2)

(b)

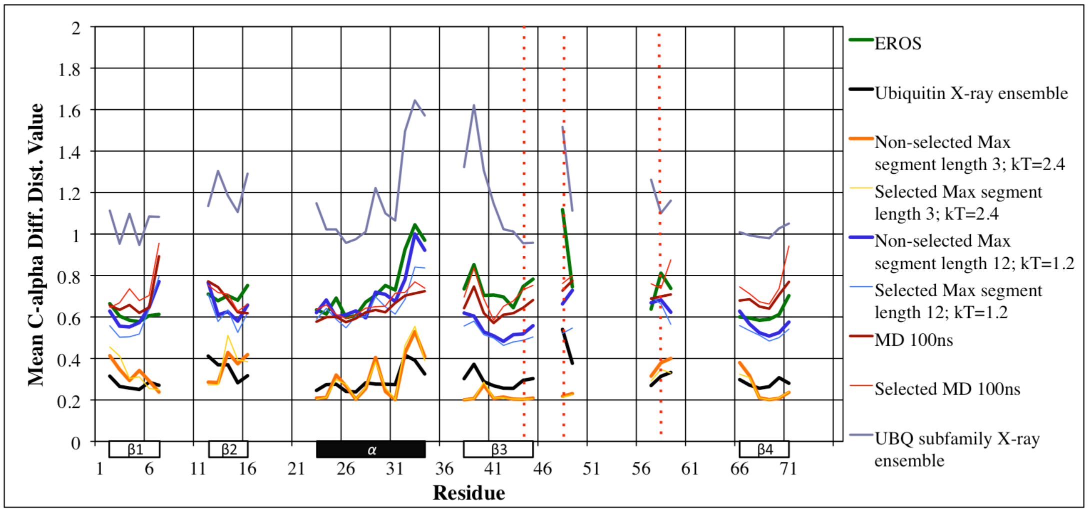


Figure 3-S3. C-alpha difference distance matrices.

(a) C-alpha difference distance matrices of various ensembles. **(b)** Mean C-alpha difference distance values for various ensembles. Red dashed lines: anchor residues 44, 58 and 68.

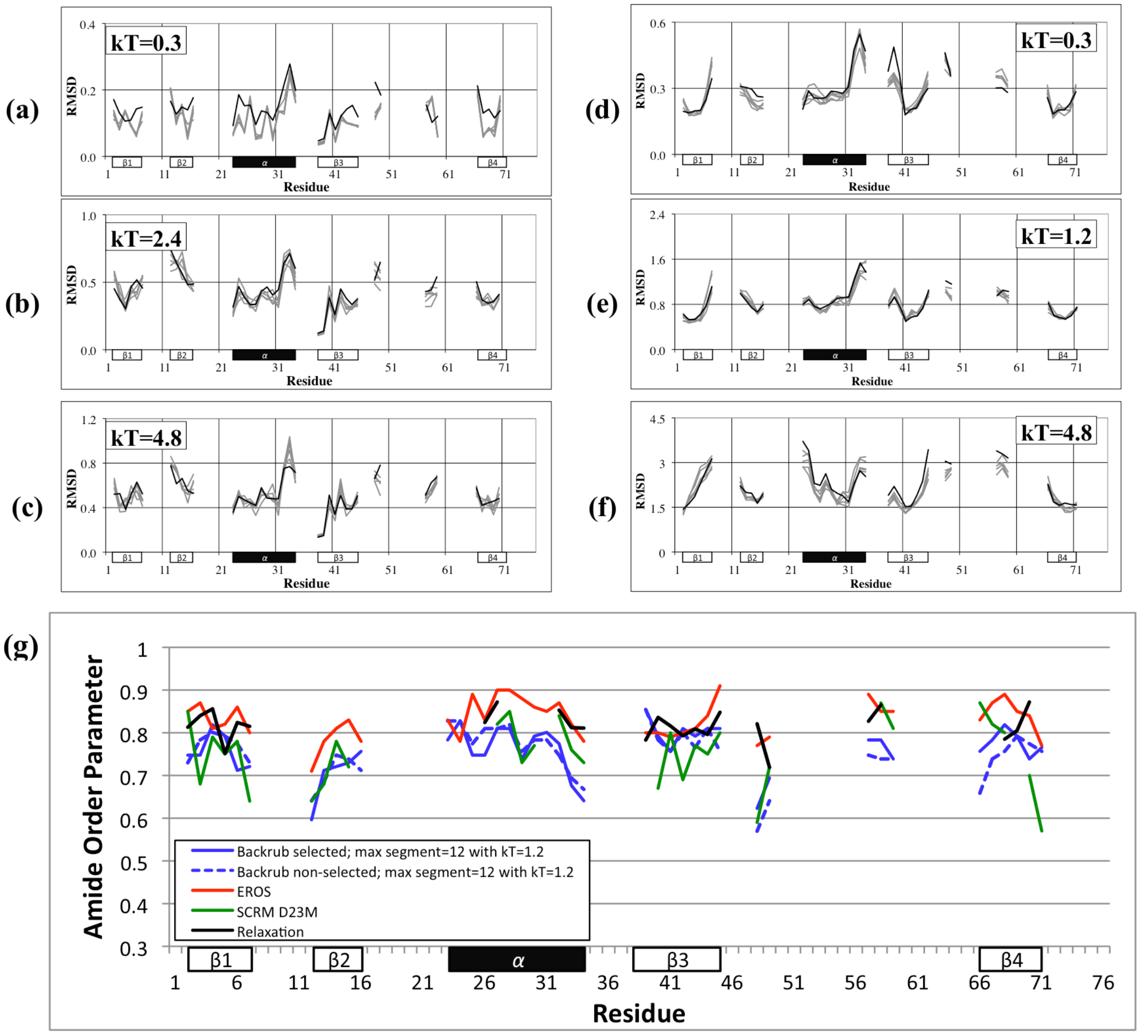


Figure 3-S4. C-alpha RMSD and NH order parameters over sequence.

C-alpha RMSD traces of the best five selected (grey) and one non-selected (black) Backrub ensembles for maximum segment length of 3 with **(a)** kT=0.3, **(b)** kT=2.4, and **(c)** kT=4.8 and maximum segment length of 12 with **(d)** kT=0.3, **(e)** kT=1.2, and **(f)** kT=4.8. **(g)** Amide order parameters for the selected and non-selected Backrub ensembles, the SCRM description, the relaxation experiments, and the EROS ensemble.

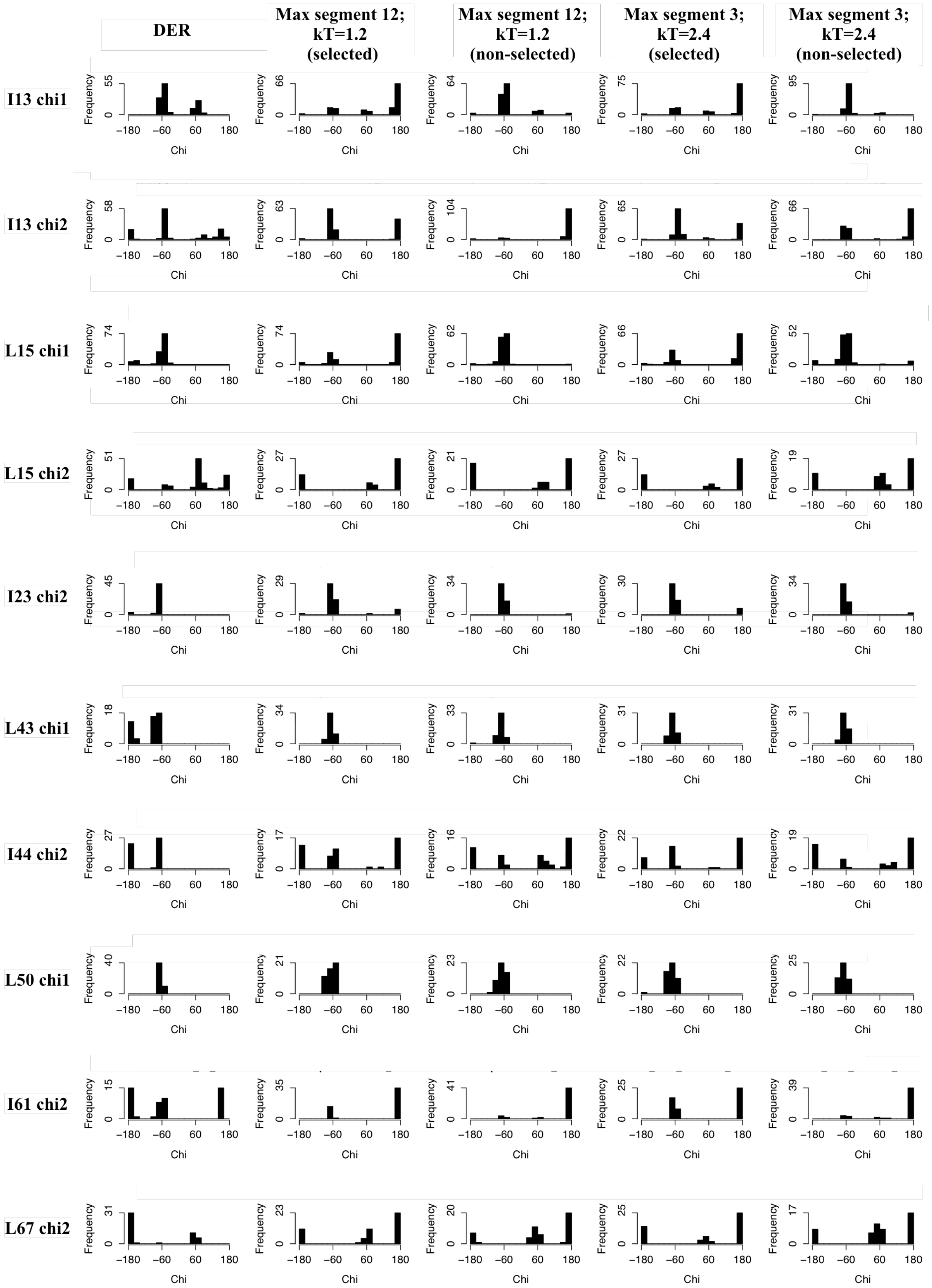

DER
Max segment 12; kT=1.2 (selected)
Max segment 12; kT=1.2 (non-selected)
Max segment 3; kT=2.4 (selected)
Max segment 3; kT=2.4 (non-selected)
I13 chi1
55
66
64
75
95
I13 chi2
58
63
104
65
66
L15 chi1
74
74
62
66
52
L15 chi2
51
27
21
27
19
I23 chi2
45
29
34
30
34
L43 chi1
18
34
33
31
31
I44 chi2
27
17
16
22
19
L50 chi1
40
21
23
22
25
I61 chi2
15
35
41
25
39
L67 chi2
31
23
20
25
17
Frequency
0
−180
−60
60
180
Chi

Figure 3-S5. Chi angle distributions.

Chi angle distributions of various residues in the DER ensemble (1XQQ), selected and non-selected Backrub ensembles with maximum segment length of 12 with kT=1.2, and selected and non-selected Backrub ensembles with maximum segment length of 3 with kT=2.4.

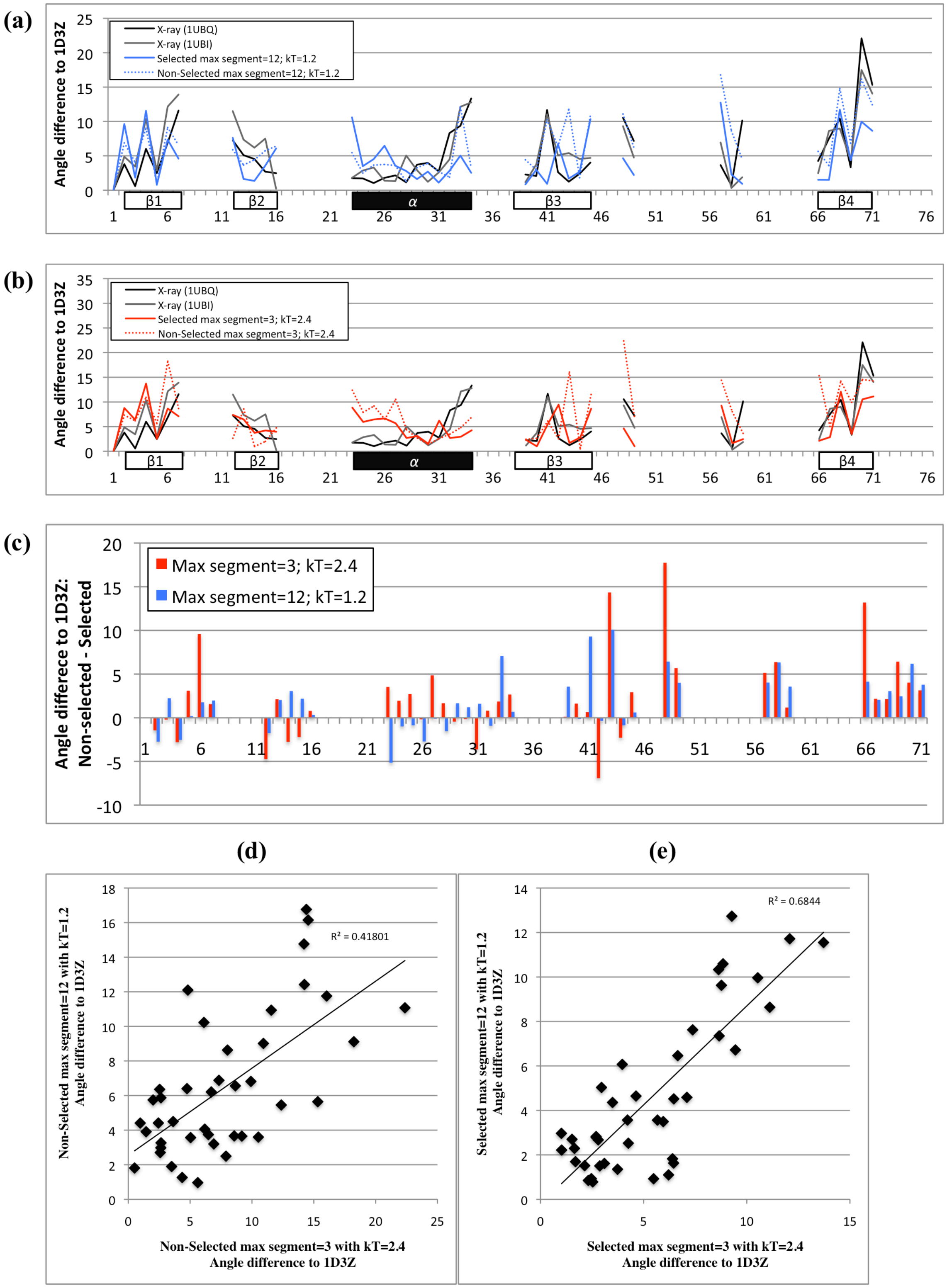

(a)
Angle difference to 1D3Z
X-ray (1UBQ)
X-ray (1UBI)
Selected max segment=12; kT=1.2
Non-Selected max segment=12; kT=1.2
β1
β2
α
β3
β4
(b)
Angle difference to 1D3Z
X-ray (1UBQ)
X-ray (1UBI)
Selected max segment=3; kT=2.4
Non-Selected max segment=3; kT=2.4
β1
β2
α
β3
β4
(c)
Angle differece to 1D3Z: Non-selected - Selected
🟥 Max segment=3; kT=2.4
🟦 Max segment=12; kT=1.2
(d)
Non-Selected max segment=12 with kT=1.2 Angle difference to 1D3Z
R² = 0.41801
Non-Selected max segment=3 with kT=2.4 Angle difference to 1D3Z
(e)
Selected max segment=12 with kT=1.2 Angle difference to 1D3Z
R² = 0.6844
Selected max segment=3 with kT=2.4 Angle difference to 1D3Z

Figure 3-S6. Angle difference between the average amide vector orientation of the 1D3Z NMR ensemble and the average amide vector orientations in selected and non-selected Backrub ensembles.

(a) Maximum segment length of 12 with kT=1.2 and **(b)** maximum segment length of 3 with kT=2,4. The angle difference of the average amide vector orientation of the 1D3Z ensemble is also compared to the orientation of amide vectors in two X-ray structures (with hydrogens added using Rosetta). **(c)** The difference in the angle differences from **(a)** and **(b)** for non-selected minus selected ensembles in secondary structure regions. **(d)** Angle differences of the two **(f)** selected and **(f)** non-selected Backrub ensembles plotted relative to each for residues in secondary structure regions.

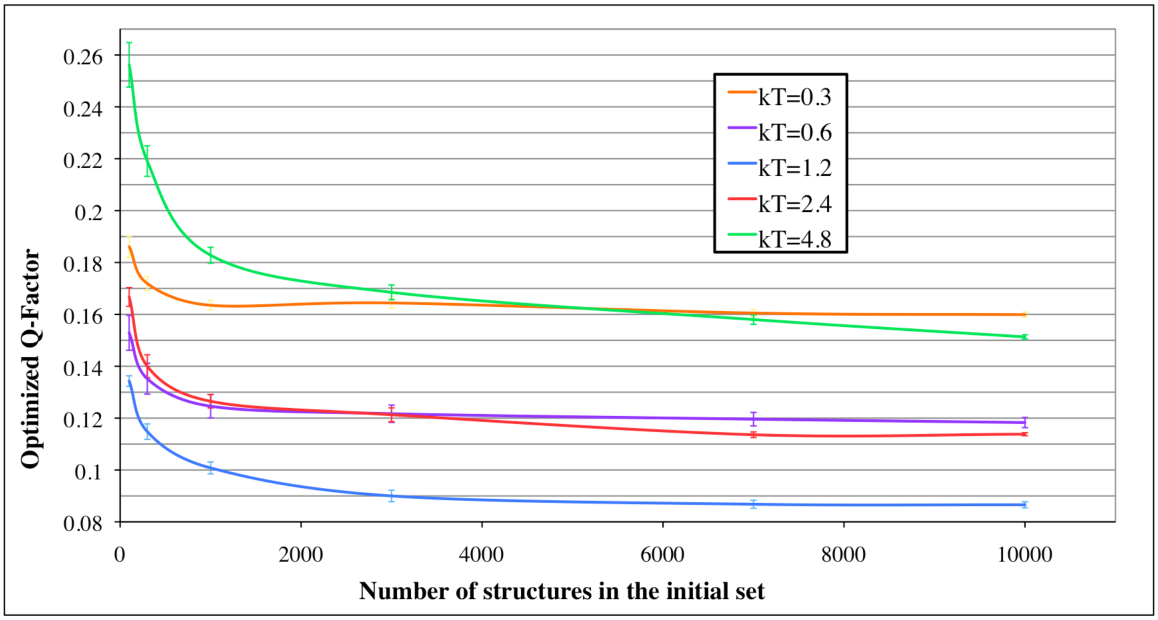


Figure 3-S7. Convergence of Q factors in the optimization protocol.

SUPPLEMENTARY TABLES

Table 3-S1. Q-factors of selected ensembles at various simulation temperatures and maximum segment lengths.

kT	**0.3**	**0.6**	**1.2**	**2.4**	**4.8**
Max segment length=3	0.153	0.108	0.093	0.089	0.098
Max segment length=12	0.160	0.118	0.086	0.112	0.153

Table 3-S2. Cross-validation analysis.

Ensemble	**R_{free}**
1XQQ	23.1%
2NR2	19.5%
2K39	16.1%
1D3Z	20.0%
1G6J	38.1%
1UD7	28.3%
Ubiquitin X-ray ensemble	17.8%
Non-selected Backrub; maximum segment length 12 with kT=1.2	20.5%
Non-selected Backrub; maximum segment length 3 with kT=2.4	26.3%
MD 100ns	23.3%

Ensemble size	**2**	**3**	**5**	**10**	**20**	**50**
Max segment length 3 with kT=2.4	23.5%	25.8%	21.8%	21.9%	21.1%	21.3%
Max segment length 12 with kT=1.2	24.1%	22.0%	20.7%	18.4%	18.6%	18.0%

kT	**0.3**	**0.6**	**1.2**	**2.4**	**4.8**
Max segment length 3 with ensemble size 50	18.8%	18.6%	19.3%	21.3%	24.8%
Max segment length 12 with ensemble size 50	18.8%	17.8%	18.0%	21.6%	27.5%

Chapter 4: Conclusion

The previous chapters describe work I have done with two chief aims in mind: to gain insight into the nature of protein dynamics and to improve methodologies for computationally sampling protein conformational space closer to the ‘real’ solution-state dynamics.

I have chosen to focus on one type of backbone conformational change, although there are many to choose from. The reasons for choosing the Backrub mechanism are numerous. First, it is localized, in that it can be applied to a specified peptide segment without affecting the position of atoms outside of this segment. Second, it has been observed in X-ray structures to facilitate rotamer changes, relatively dramatic effects when you consider that the estimated binding free energy contribution from 1 residue can be 4 kcal/mol [134]. Third, Backrub moves are conceptually similar to the 1D-GAF models, which have been used to explain long timescale protein dynamics [46]. Fourth, Backrub moves have been used to accommodate amino acid changes, [7] which suggests that the mechanism may be used during evolution to adapt to point mutations.

The insights about protein dynamics I have learned in the process include aspects of both side chain and backbone dynamics. The results in Chapter 2 confirm work by other groups underscoring the degree of flexibility of side chains both on the surface and in the core of proteins. The model for side chain motion near the native rotamer does not contain enough degrees of freedom to match the experimental relaxation measurements. Expanding the model to allow sampling in multiple rotamer wells improved the agreement with the experimental results significantly, including for core residues.

Extending the model further to include small backbone variations improved the agreement with the experimental results further still, demonstrating the importance of effects on a scale less than 1 Å in correctly modeling the flexibility of side chains. This subtlety is emphasized by the fact that allowing backbone flexibility can both increase and decrease the flexibility of a side chain.

The flexibility of these core residues in proteins presents somewhat of a surprising result that has multiple implications. Core flexibility is 'residual' entropy left over from folding from the extended or molten globule states. It helps to explain the historical difficulty of predicting protein interaction affinities and specificities, because quantifying the entropy resulting from this flexibility cannot be accomplished with simple models based on burial or the number of rotatable bonds. This entropy also helps to explain the plasticity of some proteins such as calmodulin for binding a variety of ligands. The free energy resulting from the residual entropy of the monomer protein can be 'borrowed' or 'enhanced' during a binding reaction to modulate the relative affinity of different binding partners as Wand and colleagues have shown. [135]

In Chapter 3, I applied the Backrub method to model the solution dynamics of ubiquitin. The resulting dynamic Backrub ensemble shows similar conformational variability as an ensemble of ubiquitin X-ray structures in various bound conformations, supporting work by others [74] that ubiquitin binds primarily by a conformational selection mechanism. I show further that this dynamic Backrub ensemble shows similar conformational variability as the UBQ subfamily X-ray ensemble and it is linked to the sequence variability of the UBQ subfamily.

Taken together, these two chapters demonstrate that the Backrub mechanism, albeit simple, is a useful method for sampling the dynamics of protein side chains and backbones. First, we have developed a method for predicting side chain flexibility of both surface and buried positions. Second, we have developed a method for sampling protein flexibility on the order of up to micro- or milliseconds that is successful for ubiquitin. Third, this latter application has provided a method to sample the conformational space similar to the natural family, either to increase the diversity of sequences beyond those accessible with a fixed backbone or to span the sequence space accessible to the natural family. These latter studies were focused on ubiquitin as it had an extensive dataset of RDC measurements and very few other proteins had such extensive datasets. To validate the generality of the conclusions described above and the methodology, it will be necessary to test the results on a wider set of proteins, focusing on comparing the Backrub ensembles to the protein X-ray ensemble, the natural family X-ray ensemble, and the natural family sequences. This may be possible because the RDC measurements have led us to determine parameters that result in sampling of similar conformational and sequence spaces for ubiquitin with and without RDC restraints, and the differences between these RDC-restrained and non-restrained ensembles are subtle.

The applications of these methods are numerous. First, sampling more diverse sequences will be useful in general for protein design because fixed backbone design limits the sequence diversity of the results. Second, sampling sequences similar to the natural family will be useful for finding new sequence solutions in various design problems. In Chapter 3, the Backrub ensembles were shown to give similar results to using an ensemble of backbones from an MD simulation; the advantages of Backrub are that it is much faster and as a result can likely sample longer timescale conformational changes, although Monte Carlo simulations are independent of timescale. Third, these methods can also be applied to allow for flexibility in protein-small molecule and protein-protein docking to predict binding when conformational selection is applicable. Our work in Chapter 3 and the results of Lange et al. [74] suggest that ubiquitin binds its partners predominantly by conformational selection and this finding may be applicable to other proteins. Fourth, our methods can be used to design flexibility or rigidity at certain sites, which could be used to change the binding specificity. Fifth, the methods may be useful for predicting clusters of correlated residues in proteins. Sixth, Backrub ensemble can be used to design more 'rational' libraries for screening protein functions.

At some point in the future, computational protein design will likely be a simple task, and whole networks or enzymatic pathways will be designed at once. This day may not come tomorrow, but given the rate of progress in protein design, it is within the realm of possibility. Future methodologies may not use the Backrub mechanism specifically, but it is likely that these future methods will sample a conformational space similar to that accessible to the protein in solution. This thesis is a step along that path.

Bibliography

1. 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.

2. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-8.

3. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature* **423**, 185-90.

4. Allert, M., Rizk, S. S., Looger, L. L. & Hellinga, H. W. (2004). Computational design of receptors for an organophosphate surrogate of the nerve agent soman. *Proc Natl Acad Sci U S A* **101**, 7907-12.

5. Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-5.

6. Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., 3rd, Hilvert, D., Houk, K. N., Stoddard, B. L. & Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-91.

7. Smith, C. A. & Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J Mol Biol* **380**, 742-56.

8. Davis, I. W., Arendall, W. B., 3rd, Richardson, D. C. & Richardson, J. S. (2006). The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure* **14**, 265-74.

9. Friedland, G. D., Linares, A. J., Smith, C. A. & Kortemme, T. (2008). A simple model of backbone flexibility improves modeling of side-chain conformational variability. *J Mol Biol* **380**, 757-74.

10. Hartmann, H., Parak, F., Steigemann, W., Petsko, G. A., Ponzi, D. R. & Frauenfelder, H. (1982). Conformational substates in a protein: structure and dynamics of metmyoglobin at 80 K. *Proc Natl Acad Sci U S A* **79**, 4967-71.

11. Dill, K. A. & Shortle, D. (1991). Denatured states of proteins. *Annu Rev Biochem* **60**, 795-825.

12. Kortemme, T., Kelly, M. J., Kay, L. E., Forman-Kay, J. & Serrano, L. (2000). Similarities between the spectrin SH3 domain denatured state and its folding transition state. *J Mol Biol* **297**, 1217-29.

13. Choy, W. Y. & Forman-Kay, J. D. (2001). Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J Mol Biol* **308**, 1011-32.

14. Lindorff-Larsen, K., Best, R. B., Depristo, M. A., Dobson, C. M. & Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128-32.

15. Walsh, S. T., Lee, A. L., DeGrado, W. F. & Wand, A. J. (2001). Dynamics of a de novo designed three-helix bundle protein studied by 15N, 13C, and 2H NMR relaxation methods. *Biochemistry* **40**, 9560-9.

16. Kay, L. E., Muhandiram, D. R., Farrow, N. A., Aubin, Y. & Forman-Kay, J. D. (1996). Correlation between dynamics and high affinity binding in an SH2 domain interaction. *Biochemistry* **35**, 361-8.

17. Wagner, G., DeMarco, A. & Wuthrich, K. (1976). Dynamics of the aromatic amino acid residues in the globular conformation of the basic pancreatic trypsin inhibitor (BPTI). I. 1H NMR studies. *Biophys Struct Mech* **2**, 139-58.

18. Frederick, K. K., Kranz, J. K. & Wand, A. J. (2006). Characterization of the backbone and side chain dynamics of the CaM-CaMKIp complex reveals microscopic contributions to protein conformational entropy. *Biochemistry* **45**, 9841-8.

19. Kay, L. E., Muhandiram, D. R., Wolf, G., Shoelson, S. E. & Forman-Kay, J. D. (1998). Correlation between binding and dynamics at SH2 domain interfaces. *Nat Struct Biol* **5**, 156-63.

20. Thanos, C. D., DeLano, W. L. & Wells, J. A. (2006). Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc Natl Acad Sci U S A* **103**, 15422-7.

21. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295-9.

22. Ota, N. & Agard, D. A. (2005). Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *J Mol Biol* **351**, 345-54.

23. Gunasekaran, K., Ma, B. & Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins* **57**, 433-43.

24. Clarkson, M. W. & Lee, A. L. (2004). Long-range dynamic effects of point mutations propagate through side chains in the serine protease inhibitor eglin c. *Biochemistry* **43**, 12448-58.

25. Igumenova, T. I., Lee, A. L. & Wand, A. J. (2005). Backbone and side chain dynamics of mutant calmodulin-peptide complexes. *Biochemistry* **44**, 12627-39.

26. Fuentes, E. J., Gilmore, S. A., Mauldin, R. V. & Lee, A. L. (2006). Evaluation of energetic and dynamic coupling networks in a PDZ domain protein. *J Mol Biol* **364**, 337-51.

27. Scheer, J. M., Romanowski, M. J. & Wells, J. A. (2006). A common allosteric site and mechanism in caspases. *Proc Natl Acad Sci U S A* **103**, 7595-600.

28. Kern, D. & Zuiderweg, E. R. (2003). The role of dynamics in allosteric regulation. *Curr Opin Struct Biol* **13**, 748-57.

29. Mendes, J., Baptista, A. M., Carrondo, M. A. & Soares, C. M. (1999). Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Proteins* **37**, 530-43.

30. Hu, H., Hermans, J. & Lee, A. L. (2005). Relating side-chain mobility in proteins to rotameric transitions: insights from molecular dynamics simulations and NMR. *J Biomol NMR* **32**, 151-62.

31. Koehl, P. & Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol* **239**, 249-75.

32. Zhang, J. & Liu, J. S. (2006). On side-chain conformational entropy of proteins. *PLoS Comput Biol* **2**, e168.

33. Best, R. B., Clarke, J. & Karplus, M. (2005). What contributions to protein side-chain dynamics are probed by NMR experiments? A molecular dynamics simulation analysis. *J Mol Biol* **349**, 185-203.

34. Best, R. B., Clarke, J. & Karplus, M. (2004). The origin of protein sidechain order parameter distributions. *J Am Chem Soc* **126**, 7734-5.

35. Prabhu, N. V., Lee, A. L., Wand, A. J. & Sharp, K. A. (2003). Dynamics and entropy of a calmodulin-peptide complex studied by NMR and molecular dynamics. *Biochemistry* **42**, 562-70.

36. Peterson, R. W., Dutton, P. L. & Wand, A. J. (2004). Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Protein Sci* **13**, 735-51.

37. Xiang, Z. & Honig, B. (2001). Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol* **311**, 421-30.

38. Liang, S. & Grishin, N. V. (2002). Side-chain modeling with an optimized scoring function. *Protein Sci* **11**, 322-31.

39. Millet, O., Mittermaier, A., Baker, D. & Kay, L. E. (2003). The effects of mutations on motions of side-chains in protein L studied by 2H NMR dynamics and scalar couplings. *J Mol Biol* **329**, 551-63.

40. 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.

41. Mittermaier, A., Kay, L. E. & Forman-Kay, J. D. (1999). Analysis of deuterium relaxation-derived methyl axis order parameters and correlation with local structure. *Journal of Biomolecular NMR* **13**, 181-185.

42. Mittermaier, A., Davidson, A. R. & Kay, L. E. (2003). Correlation between 2H NMR side-chain order parameters and sequence conservation in globular proteins. *J Am Chem Soc* **125**, 9004-5.

43. Ming, D. & Bruschweiler, R. (2004). Prediction of methyl-side chain dynamics in proteins. *J Biomol NMR* **29**, 363-8.

44. Best, R. B., Lindorff-Larsen, K., DePristo, M. A. & Vendruscolo, M. (2006). Relation between native ensembles and experimental structures of proteins. *Proc Natl Acad Sci U S A* **103**, 10901-6.

45. Cornilescu G, M. J., Ottiger M, Bax A. (1998). Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase

. *J Am Chem Soc* **120**, 6836-6837.

46. Bernado, P. & Blackledge, M. (2004). Anisotropic Small Amplitude Peptide Plane Dynamics in Proteins from Residual Dipolar Couplings. *J. Am. Chem. Soc.* **126**, 4907-4920.

47. Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proc Natl Acad Sci U S A* **98**, 14274-9.

48. Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., Skalicky, J. J., Kay, L. E. & Kern, D. (2005). Intrinsic dynamics of an enzyme underlies catalysis. *Nature* **438**, 117-21.

49. Henzler-Wildman, K. A., Lei, M., Thai, V., Kerns, S. J., Karplus, M. & Kern, D. (2007). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **450**, 913-6.

50. Wolf-Watz, M., Thai, V., Henzler-Wildman, K., Hadjipavlou, G., Eisenmesser, E. Z. & Kern, D. (2004). Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat Struct Mol Biol* **11**, 945-9.

51. Lee, A. L., Flynn, P. F. & Wand, A. J. (1999). Comparison of 2H and 13C NMR Relaxation Techniques for the Study of Protein Methyl Group Dynamics in Solution. *J. Am. Chem. Soc.* **121**, 2891-2902.

52. 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.

53. Geierhaas, C. D., Best, R. B., Paci, E., Vendruscolo, M. & Clarke, J. (2006). Structural comparison of the two alternative transition states for folding of TI I27. *Biophys J* **91**, 263-75.

54. Goehlert, V. A., Krupinska, E., Regan, L. & Stone, M. J. (2004). Analysis of side chain mobility among protein G B1 domain mutants with widely varying stabilities. *Protein Sci* **13**, 3322-30.

55. Hu, H., Clarkson, M. W., Hermans, J. & Lee, A. L. (2003). Increased rigidity of eglin c at acidic pH: evidence from NMR spin relaxation and MD simulations. *Biochemistry* **42**, 13856-68.

56. Liu, W., Flynn, P. F., Fuentes, E. J., Kranz, J. K., McCormick, M. & Wand, A. J. (2001). Main chain and side chain dynamics of oxidized flavodoxin from Cyanobacterium anabaena. *Biochemistry* **40**, 14744-53.

57. Flynn, P. F., Bieber Urbauer, R. J., Zhang, H., Lee, A. L. & Wand, A. J. (2001). Main chain and side chain dynamics of a heme protein: 15N and 2H NMR relaxation studies of R. capsulatus ferrocytochrome c2. *Biochemistry* **40**, 6559-69.

58. Gagne, S. M., Tsuda, S., Spyracopoulos, L., Kay, L. E. & Sykes, B. D. (1998). Backbone and methyl dynamics of the regulatory domain of troponin C: anisotropic rotational diffusion and contribution of conformational entropy to calcium affinity. *J Mol Biol* **278**, 667-86.

59. Loh, A. P., Pawley, N., Nicholson, L. K. & Oswald, R. E. (2001). An increase in side chain entropy facilitates effector binding: NMR characterization of the side chain methyl group dynamics in Cdc42Hs. *Biochemistry* **40**, 4590-600.

60. Constantine, K. L., Friedrichs, M. S., Wittekind, M., Jamil, H., Chu, C. H., Parker, R. A., Goldfarb, V., Mueller, L. & Farmer, B. T., 2nd. (1998). Backbone and side chain dynamics of uncomplexed human adipocyte and muscle fatty acid-binding proteins. *Biochemistry* **37**, 7965-80.

61. Lee, A. L., Kinnear, S. A. & Wand, A. J. (2000). Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nat Struct Biol* **7**, 72-7.

62. Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E. & Friesner, R. A. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins* **55**, 351-67.

63. Zhu, K., Shirts, M. R., Friesner, R. A. & Jacobson, M. P. (2007). Multiscale Optimization of a Truncated Newton Minimization Algorithm and Application to Proteins and Protein-Ligand Complexes. *J. Chem. Theory Comput.* **3**, 640-648.

64. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. (2004). Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66-93.

65. Kortemme, T., Morozov, A. V. & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* **326**, 1239-59.

66. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins* **35**, 133-52.

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

68. Dunbrack, R. L., Jr. (2003). The Backbone-Dependent Rotamer Library (May 2002).

69. Neria, E., Fischer, S. & Karplus, M. (1996). Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **105**, 1902-1921.

70. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* **94**, 10172-7.

71. MacKerell Jr, A. D., Bashford, D., Bellott, M., Dunbrack Jr, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H. & Ha, S. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586-3616.

72. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci* **5**, 895-903.

73. Fuentes, E. J., Der, C. J. & Lee, A. L. (2004). Ligand-dependent dynamics and intramolecular signaling in a PDZ domain. *J Mol Biol* **335**, 1105-15.

74. Lange, O. F., Lakomek, N. A., Fares, C., Schroder, G. F., Walter, K. F., Becker, S., Meiler, J., Grubmuller, H., Griesinger, C. & de Groot, B. L. (2008). Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **320**, 1471-5.

75. Eisenmesser, E. Z., Bosco, D. A., Akke, M. & Kern, D. (2002). Enzyme dynamics during catalysis. *Science* **295**, 1520-3.

76. Henzler-Wildman, K. A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M. A., Petsko, G. A., Karplus, M., Hubner, C. G. & Kern, D. (2007). Intrinsic motions along an enzymatic reaction trajectory. *Nature* **450**, 838-44.

77. Boehr, D. D., McElheny, D., Dyson, H. J. & Wright, P. E. (2006). The dynamic energy landscape of dihydrofolate reductase catalysis. *Science* **313**, 1638-42.

78. Schnell, J. R., Dyson, H. J. & Wright, P. E. (2004). Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu Rev Biophys Biomol Struct* **33**, 119-40.

79. Wei, B. Q., Weaver, L. H., Ferrari, A. M., Matthews, B. W. & Shoichet, B. K. (2004). Testing a flexible-receptor docking algorithm in a model binding site. *J Mol Biol* **337**, 1161-82.

80. Chaudhury, S. & Gray, J. J. (2008). Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. *J Mol Biol* **381**, 1068-87.

81. Prasad, J. C., Goldstone, J. V., Camacho, C. J., Vajda, S. & Stegeman, J. J. (2007). Ensemble modeling of substrate binding to cytochromes P450: analysis of catalytic differences between CYP1A orthologs. *Biochemistry* **46**, 2640-54.

82. Fu, X., Apgar, J. R. & Keating, A. E. (2007). Modeling backbone flexibility to achieve sequence diversity: the design of novel alpha-helical ligands for Bcl-xL. *J Mol Biol* **371**, 1099-117.

83. Larson, S. M., England, J. L., Desjarlais, J. R. & Pande, V. S. (2002). Thoroughly sampling sequence space: large-scale protein design of structural ensembles. *Protein Sci* **11**, 2804-13.

84. Ding, F. & Dokholyan, N. V. (2006). Emergence of Protein Fold Families through Rational Design. *PLoS Computational Biology* **2**, e85.

85. Kraemer-Pecore, C. M., Lecomte, J. T. & Desjarlais, J. R. (2003). A de novo redesign of the WW domain. *Protein Sci* **12**, 2194-205.

86. Kono, H. & Saven, J. G. (2001). Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J Mol Biol* **306**, 607-28.

87. Zoete, V., Michielin, O. & Karplus, M. (2002). Relation between sequence and structure of HIV-1 protease inhibitor complexes: a model system for the analysis of protein flexibility. *J Mol Biol* **315**, 21-52.

88. Bremi, T. & Bruschweiler, R. (1997). Locally Anisotropic Internal Polypeptide Backbone Dynamics by NMR Relaxation. *J. Am. Chem. Soc.* **119**, 6672-6673.

89. Muhandiram, D. R., Yamazaki, T., Sykes, B. D. & Kay, L. E. (1995). Measurement of 2H T1 and T1.rho. Relaxation Times in Uniformly 13C-Labeled and Fractionally 2H-Labeled Proteins in Solution. *J. Am. Chem. Soc.* **117**, 11536-11544.

90. Lipari, G. & Szabo, A. (1982). Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J. Am. Chem. Soc. ; Vol/Issue: 104:17*, Pages: 4546-4559.

91. Kay, L. E., Torchia, D. A. & Bax, A. (1989). Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* **28**, 8972-9.

92. Lundstrom, P., Mulder, F. A. & Akke, M. (2005). Correlated dynamics of consecutive residues reveal transient and cooperative unfolding of secondary structure in proteins. *Proc Natl Acad Sci U S A* **102**, 16984-9.

93. Wang, T., Frederick, K. K., Igumenova, T. I., Wand, A. J. & Zuiderweg, E. R. (2005). Changes in calmodulin main-chain dynamics upon ligand binding revealed by cross-correlated NMR relaxation measurements. *J Am Chem Soc* **127**, 828-9.

94. LeMaster, D. M. & Kushlan, D. M. (1996). Dynamical Mapping of E. coli Thioredoxin via 13C NMR Relaxation Analysis. *J. Am. Chem. Soc.* **118**, 9255-9264.

95. Chou, J. J., Case, D. A. & Bax, A. (2003). Insights into the mobility of methyl-bearing side chains in proteins from (3)J(CC) and (3)J(CN) couplings. *J Am Chem Soc* **125**, 8959-66.

96. Clore, G. M. & Schwieters, C. D. (2004). Amplitudes of protein backbone dynamics and correlated motions in a small alpha/beta protein: correspondence of dipolar coupling and heteronuclear relaxation measurements. *Biochemistry* **43**, 10678-91.

97. Lakomek, N. A., Walter, K. F., Fares, C., Lange, O. F., de Groot, B. L., Grubmuller, H., Bruschweiler, R., Munk, A., Becker, S., Meiler, J. & Griesinger, C. (2008). Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics. *J Biomol NMR* **41**, 139-55.

98. Lakomek, N. A., Carlomagno, T., Becker, S., Griesinger, C. & Meiler, J. (2006). A thorough dynamic interpretation of residual dipolar couplings in ubiquitin. *J Biomol NMR* **34**, 101-15.

99. Skrynnikov, N. R., Goto, N. K., Yang, D., Choy, W. Y., Tolman, J. R., Mueller, G. A. & Kay, L. E. (2000). Orienting domains in proteins using dipolar couplings measured by liquid-state NMR: differences in solution and crystal forms of maltodextrin binding protein loaded with beta-cyclodextrin. *J Mol Biol* **295**, 1265-73.

100. Bouvignies, G., Bernado, P., Meier, S., Cho, K., Grzesiek, S., Bruschweiler, R. & Blackledge, M. (2005). Identification of slow correlated motions in proteins using

residual dipolar and hydrogen-bond scalar couplings. *Proc Natl Acad Sci U S A* **102**, 13885-90.

101. Tjandra, N. & Bax, A. (1997). Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278**, 1111-4.

102. Tolman, J. R., Flanagan, J. M., Kennedy, M. A. & Prestegard, J. H. (1997). NMR evidence for slow collective motions in cyanometmyoglobin. *Nat Struct Biol* **4**, 292-7.

103. Elber, R. & Karplus, M. (1987). Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science* **235**, 318-21.

104. Keskin, O., Jernigan, R. L. & Bahar, I. (2000). Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophys J* **78**, 2093-106.

105. Maguid, S., Fernandez-Alberti, S., Ferrelli, L. & Echave, J. (2005). Exploring the common dynamics of homologous proteins. Application to the globin family. *Biophys J* **89**, 3-13.

106. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A* **97**, 10383-8.

107. Saunders, C. T. & Baker, D. (2005). Recapitulation of protein family divergence using flexible backbone protein design. *J Mol Biol* **346**, 631-44.

108. Larson, S. M., Garg, A., Desjarlais, J. R. & Pande, V. S. (2003). Increased detection of structural templates using alignments of designed sequences. *Proteins* **51**, 390-6.

109. Kiel, C. & Serrano, L. (2006). The ubiquitin domain superfold: structure-based sequence alignments and characterization of binding epitopes. *J Mol Biol* **355**, 821-44.

110. Treynor, T. P., Vizcarra, C. L., Nedelcu, D. & Mayo, S. L. (2007). Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc Natl Acad Sci U S A* **104**, 48-53.

111. Humphris, E. L. & Kortemme, T. (in press).

112. Richter, B., Gsponer, J., Varnai, P., Salvatella, X. & Vendruscolo, M. (2007). The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J Biomol NMR* **37**, 117-35.

113. Mandel, A. M., Akke, M. & Palmer, A. G., 3rd. (1995). Backbone dynamics of Escherichia coli ribonuclease HI: correlations with structure and function in an active enzyme. *J Mol Biol* **246**, 144-63.

114. Mandel, A. M., Akke, M. & Palmer, A. G., 3rd. (1996). Dynamics of ribonuclease H: temperature dependence of motions on multiple time scales. *Biochemistry* **35**, 16009-23.

115. Lakomek, N. A., Fares, C., Becker, S., Carlomagno, T., Meiler, J. & Griesinger, C. (2005). Side-chain orientation and hydrogen-bonding imprint supra-Tau(c) motion on the protein backbone of ubiquitin. *Angew Chem Int Ed Engl* **44**, 7776-8.

116. Chen, Y., Campbell, S. L. & Dokholyan, N. V. (2007). Deciphering protein dynamics from NMR data using explicit structure sampling and selection. *Biophys J* **93**, 2300-6.

117. Wong, V. & Case, D. A. (2008). Evaluating rotational diffusion from protein MD simulations. *J Phys Chem B* **112**, 6013-24.

118. Meiler, J., Prompers, J. J., Peti, W., Griesinger, C. & Bruschweiler, R. (2001). Model-free approach to the dynamic interpretation of residual dipolar couplings in globular proteins. *J Am Chem Soc* **123**, 6098-107.

119. Peti, W., Meiler, J., Bruschweiler, R. & Griesinger, C. (2002). Model-free analysis of protein backbone motion from residual dipolar couplings. *J Am Chem Soc* **124**, 5822-33.

120. Tolman, J. R., Al-Hashimi, H. M., Kay, L. E. & Prestegard, J. H. (2001). Structural and Dynamic Analysis of Residual Dipolar Coupling Data for Proteins. *J. Am. Chem. Soc.* **123**, 1416-1424.

121. Lupyan, D., Leo-Macias, A. & Ortiz, A. R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* **21**, 3255-63.

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

123. Coutsias, E. A., Seok, C., Jacobson, M. P. & Dill, K. A. (2004). A kinematic view of loop closure. *J Comput Chem* **25**, 510-28.

124. Georgiev, I., Keedy, D., Richardson, J. S., Richardson, D. C. & Donald, B. R. (2008). Algorithm for backrub motions in protein design. *Bioinformatics* **24**, i196-204.

125. Betancourt, M. R. (2005). Efficient Monte Carlo trial moves for polypeptide simulations. *J Chem Phys* **123**, 174905.

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

127. Meiler, J., Peti, W. & Griesinger, C. (2000). DipoCoup: A versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudocontact shifts. *J Biomol NMR* **17**, 283-94.

128. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637.

129. (2008). R: A Language and Environment for Statistical Computing. Team, R. D. C.

130. Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., 3rd, Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* **35**, W375-83.

131. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* **50**, 437-50.

132. 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.

133. Harpaz, Y., Gerstein, M. & Chothia, C. (1994). Volume changes on protein folding. *Structure* **2**, 641-9.

134. Kortemme, T., Kim, D. E. & Baker, D. (2004). Computational alanine scanning of protein-protein interfaces. *Sci STKE* **2004**, pl2.

135. Frederick, K. K., Marlow, M. S., Valentine, K. G. & Wand, A. J. (2007). Conformational entropy in molecular recognition by proteins. *Nature* **448**, 325-9.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

1/6/09

Author Signature Date