Simulating Conformational Fluctuations and Designing Switchable Interactions with Linear Protein Recognition Motifs

by

Colin Alexander Smith

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Acknowledgements

First and foremost, I have to thank my primary advisor, Tanja Kortemme, and my secondary advisor, Matt Jacobson for providing a scientific environment where I was able to work independently, yet also to enjoy a great deal of intellectual stimulation and discussion. I also need to thank all the members of both labs for all the times they gave helpful hints and pointed me in the right direction, and also for all the fun times when we relaxed together either during lunch or on the many snowboarding trips. Every lab member has at some time contributed to my scientific endeavors, but I would like to give special mention to Cristina Melero and Ryan Ritterson who helped make my transition from pure computation to a wet lab seamless. I also would like to acknowledge my summer students, Catherine Shi, Matt Chroust, and Thomas Bliska who came into the lab with incredible enthusiasm and made critical progress in my projects during the course of their short time working with me.

In addition to those at UCSF, all the friends that I have gotten to know during my six years in San Francisco have made a profound difference. Getting to know people and living with them through their hardships as well as my own has made my life much fuller. Finally, I would like to thank my wife Rachel, who has been with me through every step of graduate school. After only several months of dating, on many nights she patiently cooked me dinner while I studied for the Biochemistry GRE subject test. Throughout the interview process, she encouraged me to go to the best place for me and then joined me there. During the course of my PhD, the most fun and joy has come in the times I have spent with her. Lastly, she has been incredibly supportive through all the stresses of bringing my PhD to a close. My life would not be the same without her.

The text of this thesis/dissertation/manuscript is a reprint of the material as it appears in the *Journal of Molecular Biology* and *PLoS One*. The coauthor listed in these publications directed and supervised the research that forms the basis for the dissertation/thesis. References to the relevant publications and individual acknowledgements are below.

Chapter 1

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, 742-756.

I thank Andrew Bordner for providing a text version of the point mutant benchmark. Greg Friedland gave helpful feedback about the backrub sampling code and method. Jerome Nilmeier provided useful discussions about detailed balance. Matt Jacobson gave valuable feedback as my co-advisor. Christopher McClendon, Libusha Kelly, and Ian Davis provided useful comments about the manuscript.

Chapter 2

Smith, C. A. & Kortemme, T. (2010) Structure-Based Prediction of the Peptide Sequence Space Recognized by Natural and Synthetic PDZ Domains. *J Mol Biol* 402, 460-474.

I thank Matthew P. Jacobson for comments on the manuscript and Justin Ashworth for collaborating on reimplementation of the genetic algorithm-based sequence optimization in Rosetta 3. The entire Rosetta Commons community provided invaluable software development, testing, and infrastructure. Dev Sidhu and Andreas Ernst graciously gave early access to phage display data and the PDB structure of the PDLIM4 PDZ domain.

Chapter 3

Smith, C. A. & Kortemme, T. (2011) Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design. *PLoS One* 6.

I thank Andrea G. Cochran and Joanne D. Kotz for access to the raw GB1 sequencing data used in their study. Phil Bradley, Brian Kuhlman, and Andrew Leaver-Fay helped with clarifying historical differences in the Rosetta scoring function.

Abstract

Successfully incorporating backbone flexibility into the computational modeling and design of proteins and protein interactions is a key challenge that has yet to be fully solved. Most existing techniques for perturbing the backbone make changes that are not localized but propagate to distant parts of the protein, which can cause inefficiencies when working in a restricted region of the protein. The methods that do make localized changes have mostly been based on complex mathematical formulations that can inhibit their widespread application. In this thesis I describe a simple, automated approach, termed "backrub," for sampling the protein backbone. The method is based on a recurring motif of backbone motion previously observed in ultrahigh resolution (≤1Å) crystal structures, and involves backbone rotations around axes between Cα atoms. It is shown to be useful for a variety of applications, including recapitulating the backbone/side chain bias in known instances of the backrub motif, predicting the conformations of point mutants, and modeling the opening and closing of a loop around an enzyme active site. After these initial results, I undertook a large-scale study in retrospective and prospective prediction of the peptide binding specificities of natural and synthetic PDZ domains. Here, backrub backbone flexibility was shown to significantly improve the accuracy of amino acid frequency prediction. The developed method was able to capture a large fraction of the amino acids frequently observed in phage display experiments both with natural PDZ domains and a large dataset of point mutants. Finally, in an effort to broaden the application and use of the PDZ peptide specificity work, I generalized the method to also predict fold stability, using GB1 phage display as a benchmark, and produced a detailed protocol for others to apply to a wide variety of systems.

Table of Contents

List of Tables

List of Figures

Introduction

If anything has captured my imagination during the course of graduate school, it has been two things: thinking about how proteins move and how those fluctuations affect protein function, and creating better computational tools for simulating and analyzing data related to protein dynamics.

My first project, implementing and evaluating a generalized backrub move for sampling protein conformations, combined both of these interests. The move was inspired by motions observed in high-resolution crystal structures[1] and involved rotations around axes between backbone Cα atoms. One of the important computational advantages of the move was that all changes were localized. In the context of a pairwise decomposable scoring function like is used in Rosetta, this enables backbone perturbations to be evaluated without recomputing all interaction energies in the structure. Purely local moves are not new however, with formulations introduced as early as 1970[2] and further refinements introduced since[3-6]. The distinguishing features of the backrub move are that is inspired by real motions observed in nature, that it leverages bond angle flexibility to make simple geometric perturbations, and that it can make backbone changes as small as a peptide bond rotation. Before I began work on the project, Betancourt[7] had evaluated a similar move with a highly simplified scoring function. Key advances I made were to determine conditions and optimizations under which similar moves could be applied in the context of a detailed, all-atom force field. I also showed that incorporation of such moves enabled sampling of experimentally observed protein dynamics and improved prediction of side chain conformations upon point mutation. See chapter 1 for a detailed analysis of the findings.

A primary focus of Tanja's work was protein design[8-10], while Matt had made a major thrust into understanding how phosphorylation brought about conformational changes and regulated protein function[11,12]. Prior to my joining, they had briefly considered the idea of creating a phosphoswitchable protein, which I then developed into a project designing a PDZ domain whose peptide binding affinity was controlled by phosphorylation. The reverse design goal had previously been accomplished, namely redesigning a peptide to have its PDZ binding affinity regulated by phosphorylation[13]. While incorporation of linear kinase recognition motifs into unstructured peptides was relatively straightforward, the reactivity of a kinase with designed globular domains was unknown. With a visiting summer undergraduate student, Catherine Shi, I began a pilot study to determine how readily incorporation of the protein kinase A (PKA) linear recognition motif into a folded protein domain would yield a competent phosphorylation site. Out of eight candidate phosphosites, two were successfully phosphorylated.

At the same time the phosphoswitchable PDZ project was moving forward, a collaborator, Dev Sidhu, made available a large amount of PDZ-peptide phage display data that were later published[14,15]. Given my interest in regulating PDZ-peptide interactions, I undertook a project to determine how well our flexible backbone protein design methods could recapitulate the large amount of phage display data available. A previous student, Elisabeth Humphris, had developed a method for predicting the set of sequences that could be tolerated at the human growth hormone (hGH)/human growth hormone receptor (hGHR) interface[16]. Her method incorporated the backrub backbone sampling I had developed earlier. A significantly faster reimplementation of her algorithm in a new version of Rosetta allowed the large-scale analysis of hundreds of

natural and synthetic PDZ domains whose peptide phage display profiles were known. Chapter 2 describes the predictive performance in depth. I also took the PDZ-peptide interaction method and made it more generalizable and readily accessible to the greater scientific community using a mechanism known as a “protocol capture”. This work is highlighted in Chapter 3.

In tandem with the PDZ specificity work, I continued the phosphoswitch project by characterizing the mutants to determine factors that may influence phosphorylation. Circular dichroism thermal denaturation showed that the two phosphorylated proteins were the least thermostable, potentially indicating a higher degree of disorder at room temperature. The mutant with the highest rate of phosphorylation was further characterized via NMR. It showed significant chemical shift differences from the wild-type, suggesting a change in structure and/or dynamics. In addition, CLEANEX hydrogen exchange experiments showed an increase in solvent exposure of residues nearby the phosphorylation site. Given these results, we undertook a second round of design, incorporating the destabilizing mutations into previously unsuccessful designs. That strategy worked and rescued three of the original designs. Upon purifying phosphorylated variants, we found that phosphorylation lead to an approximately 10-fold reduction in binding affinity for several of the new designs. While these results showed that we had successfully created a phosphoswitchable PDZ domain, they came too late to incorporate into this thesis beyond the brief description here.

Chapter 1. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction

Abstract

Incorporation of effective backbone sampling into protein simulation and design is an important step in increasing the accuracy of computational protein modeling. Recent analysis of high-resolution crystal structures has suggested a new model, termed backrub, to describe localized, hinge-like alternative backbone and side chain conformations observed in the crystal lattice. The model involves internal backbone rotations about axes between Cα atoms. Based on this observation, we have implemented a backrub-inspired sampling method in the Rosetta structure prediction and design program. We evaluate this model of backbone flexibility using three different tests. First, we show that Rosetta backrub simulations recapitulate the correlation between backbone and side-chain conformations in the high-resolution crystal structures upon which the model was based. As a second test of backrub sampling, we show that backbone flexibility improves the accuracy of predicting point-mutant side chain conformations over fixed backbone rotameric sampling alone. Finally, we show that backrub sampling of triosephosphate isomerase loop 6 can capture the ms/μs oscillation between the open and closed states observed in solution. Our results suggest that backrub sampling captures a sizable fraction of localized conformational changes that occur in natural proteins. Application of this simple model of backbone motions may significantly improve both protein design and atomistic simulations of localized protein flexibility.

Introduction

Proteins undergo conformational fluctuations in response to thermal energy, binding events, and mutation. Understanding and predicting such excursions around the native state of a protein is a key challenge in computational molecular biology. Side chain sampling[17] has been shown to be an extremely useful first-order method for predicting small-scale conformational change. Successful applications include protein-protein docking[18,19], total redesign of protein sequences[20,21], and redesign of both protein-protein[8] and protein-DNA[22] interfaces. However, one key approximation made by many of these applications is keeping the backbone structure fixed. In actual proteins the backbone often undergoes subtle shifts in response to binding events[23] or sequence changes[24]. Successfully capturing such near-native shifts is thus important for many docking and design applications.

Numerous methods have been developed to take backbone flexibility into account for both the whole protein and local subsections. Molecular dynamics is currently one of the most pervasive methods. However, in the absence of a steep energy gradient, dynamics depend on random thermal velocities and a long sequence of time steps to sample motions as simple as a rotamer change. Monte Carlo minimization of backbone torsion angles[25-27] has also been very successful, but can result in highly non-local displacements of the protein backbone and becomes increasingly less efficient with greater protein size. Insertion of peptide fragments has been used for *de novo* protein structure prediction[28] and loop prediction[29], but causes similar propagating changes. Several non-local sampling techniques have been applied to protein design including random torsion angle sampling[30] and more correlated methods such as fragment

insertion[31], parameterized coiled-coils[32], and normal mode analysis[33]. These methods make use of patterns commonly observed in protein structures or a harmonic approximation of intra-protein interactions to increase backbone sampling efficiency. Other methods have addressed the problem of making local perturbations using heuristics to iteratively optimize backbone torsion angles until distortions of covalent geometry are minimized[34-36], but those techniques sometimes leave strained chain junctions that must be relaxed with other algorithms. Another method, called wriggling[37], was developed to make partially local moves in which groups of four torsion angles are changed simultaneously to minimize the displacement of distant atoms.

Deformations of protein backbones are truly local only if all consecutive atoms beyond the perturbed region remain fixed. Several local methods exist, the first being introduced by Go and Scheraga[2] with numerous subsequent refinements and adaptations[3-6]. These methods involve making a random prerotation of one or more backbone angles, followed by solving a geometric constraint equation for six other backbone degrees of freedom to maintain the locality of the move. Several of the methods incorporated bond angle sampling, either as part of the prerotation[3,6], or both the prerotation and the solved constraint equation[5]. The latter work also biased the prerotations towards less perburbed backbone conformations. The implementation of these methods is more complex than other common techniques like rotamer sampling. Another drawback is that such loop closure methods are biased towards proposing moves that satisfy bonded, geometric constraints, whose multiple free rotation axes can lead to radically different conformations, often with substantial steric clashes and unsatisfied hydrogen bonds.

Those non-bonded factors are particularly relevant in highly packed protein cores and interfaces.

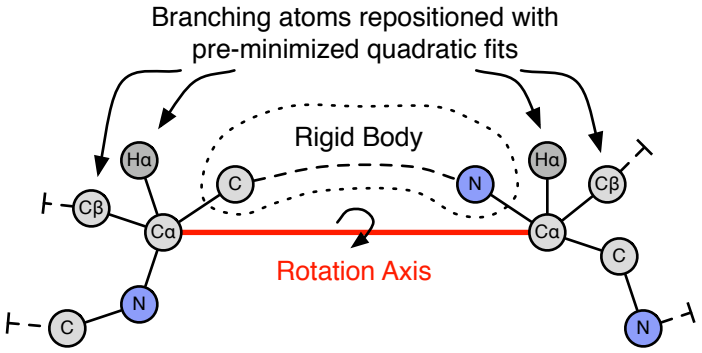
The work described here, instead of being motivated by geometric constraints, derives its motional model from conformational variations observed in high-resolution (≤ 1Å) crystal structures[1]. The fluctuations observed in the crystal lattice motivated Davis et al. to create a simple model, called Backrub, for subtle backbone shifts using just three residues. The core idea in this work is to use that type of motion, observed in nature, to computationally sample backbone configurations in a generalized scheme. A similar move set was recently described[7] in the context of a simplified energy function. Here, we investigate the utility of the backrub move to sample conformations in the context of the Rosetta all-atom force field. Rosetta has been successfully used for protein-protein docking[19], protein-ligand docking[38], redesign of protein cores[31], design of new protein interface specificities[8], and *de novo* prediction of small protein structures[39]. As an initial test, we recapitulate the backbone/side-chain correlations observed in the same high-resolution structures that inspired the Backrub model. We go on to show that backrub backbone flexibility improves side-chain modeling of point mutations. Finally, as a demonstration of the method's potential, we present a proof-of-concept simulation showing efficient sampling of the opening and closing of triosephosphate isomerase loop 6. Our results indicate that the backbone sampling described here captures a sizable fraction of the subtle conformational variability found in folded proteins.

Results

We implemented the backrub sampling protocol inspired by motion observed in protein structures[1] (see Figure 1-1, Figure 1-2, and Methods), and evaluated it using three

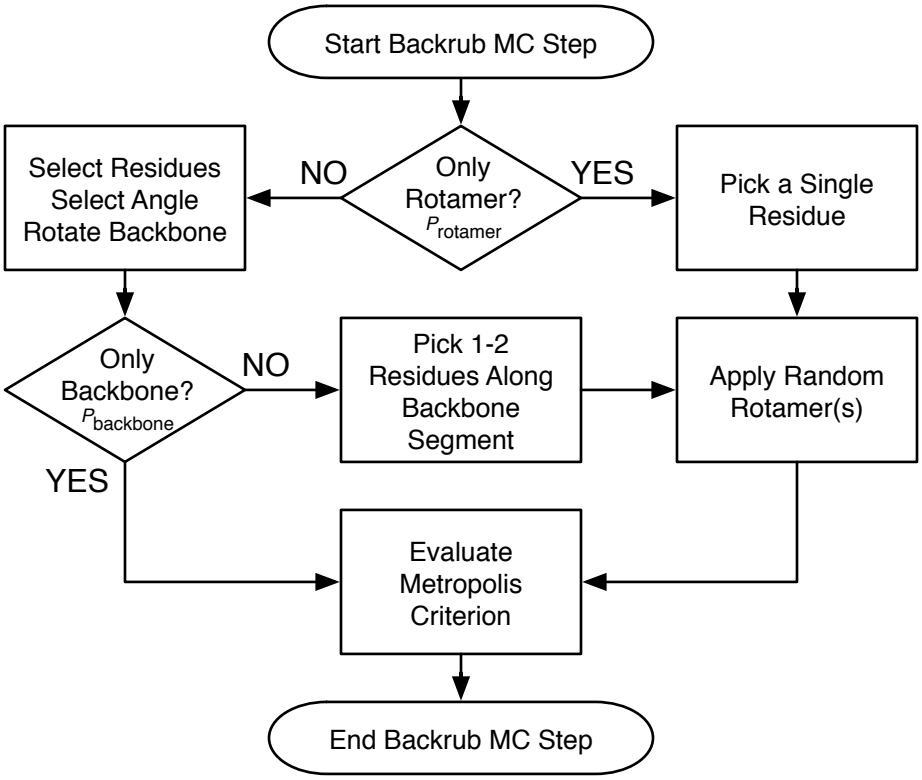
different tests: First, we sought to determine whether the motional model, combined with an all-atom force field, could recapitulate the variation seen in occurrences of a backrub motion in high-resolution crystal structures. Secondly, we test whether backrub sampling can improve the accuracy of modeling small backbone and side chain conformational changes in response to single point mutations in a set of crystal structure pairs. Finally, we show simulations indicating that backrub sampling can capture conformational variability observed in a long time-scale loop motion.

Figure 1-1. Schematic showing the generalized backrub move



Moves are made by first randomly selecting the polypeptide backbone segment size, typically 2-12 residues, then randomly selecting a starting residue compatible with the selected segment size. The Cα atoms of the starting and ending residues define the rotation axis. All atoms between the two Cα atoms are then rotated about that axis by a random angle up to 11-40 degrees, depending on the segment size. To minimize the bond angle penalty imposed by full atom force fields, precise placement of branching Cβ and hydrogen atoms is done using quadratic equations that describe the relationship between the backbone bond angle and branching atom spherical coordinates (see Methods).

Figure 1-2. Flow chart depicting substeps taken during a single Monte Carlo step



The proportion of backbone, rotamer, and backbone + rotamer steps are controlled by two parameters. The first, $P_{rotamer}$, specifies the probability of only making a rotamer move. The second, $P_{backbone}$, specifies the probability that only the backbone is modified, given that a rotamer only move type was not selected.

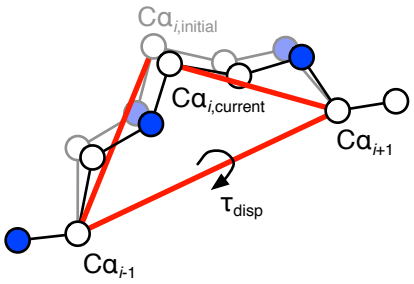
Test 1: Simulation of 3-Residue Backrubs

Davis et al.[1] derived the "Backrub" model of protein backbone motion from examples of three residue segments exhibiting multiple backbone conformations in high-resolution (≤ 1.0 Å) crystal structures. To model those variations, they used the Cα atoms as pivot points and enumerated the three possible rotation axes between them. By manually rotating the backbone around those axes, they were able to model the conformational transitions in a significant number of cases. They catalogued 126 such instances in the PDB that fit their model, the majority of which involved a simultaneous

rotamer change. In those cases, the backbone-determined location of the Cβ atom significantly altered the conformation of attached side chain atoms. As an initial test of our generalized backrub sampling method, we used focused Monte Carlo simulations to determine whether we could detect distinct populations of coupled backbone/side-chain conformations centered on coordinates observed in the PDB.

Out of 161 derived starting structures (see Methods), the majority (105) came from PDB residue entries with χ1 angles of the central side chain, *i*, occupying multiple rotameric bins (-60°, 60°, 180°). In our analysis, we therefore used the χ1 angle as a one-dimensional representation of the side chain conformation. We used the $C\alpha_{i,\text{initial}}$–$C\alpha_{i-1}$–$C\alpha_{i+1}$–$C\alpha_{i,\text{current}}$ pseudo-dihedral angle (τ_{disp}), to represent the backbone conformation of the 3-residue segment. (Figure 1-3) We wanted to determine whether the simulations showed a similar correspondence between side-chain and backbone conformation to that observed in the crystal structures. To answer that question, we calculated τ_{disp} probability distributions for each of the χ1 bins visited during the simulations and compared those distributions to the crystallographic τ_{disp} backbone angles. An example analysis of a simulation showing good agreement with the PDB is given in Figure 1-4.

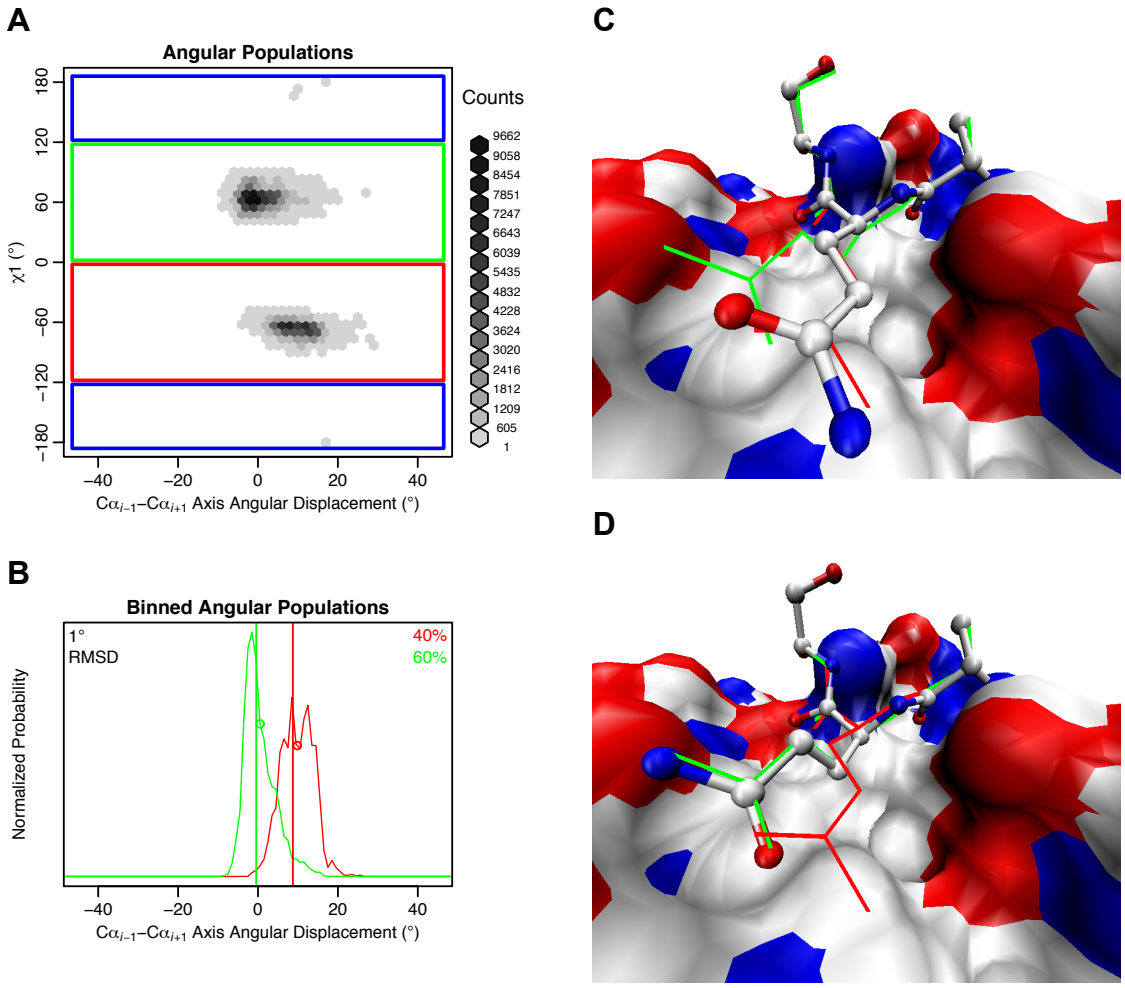
Figure 1-3. Schematic of dihedral angle used for 3-residue backrub analysis



For 3-residue backrub analysis, the τ_{disp} angle was used as a one-dimensional representation of the backbone conformation. In simulations, it was defined as the

$C\alpha_{i,\text{initial}}$–$C\alpha_{i-1}$–$C\alpha_{i+1}$–$C\alpha_{i,\text{current}}$ pseudo dihedral angle (red). In this illustration, the starting atomic coordinates are shown in gray. In some high-resolution PDB structures, alternate Cα coordinates were not provided, so the τ_{disp} angle was instead defined as the $C\beta_{i,\text{initial}}$–$C\alpha_{i-1}$–$C\alpha_{i+1}$–$C\beta_{i,\text{alternate}}$ pseudo dihedral angle (not shown) for all PDB analysis.

Figure 1-4. Example backbone/χ1 populations from a 3-residue backrub simulation



Results are shown for PDB 1PQ7 chain A, residues 62-64, starting from the "B" alternate backbone coordinates. The central side-chain is glutamine. **(A)** To monitor coupled backbone and side-chain conformational changes, we recorded both the $C\alpha_{i,\text{initial}}$–$C\alpha_{i-1}$–$C\alpha_{i+1}$–$C\alpha_{i,\text{current}}$ pseudo-dihedral angle (τ_{disp}) and χ1 angle after every Monte Carlo step. Those angles are shown binned into hexagonal

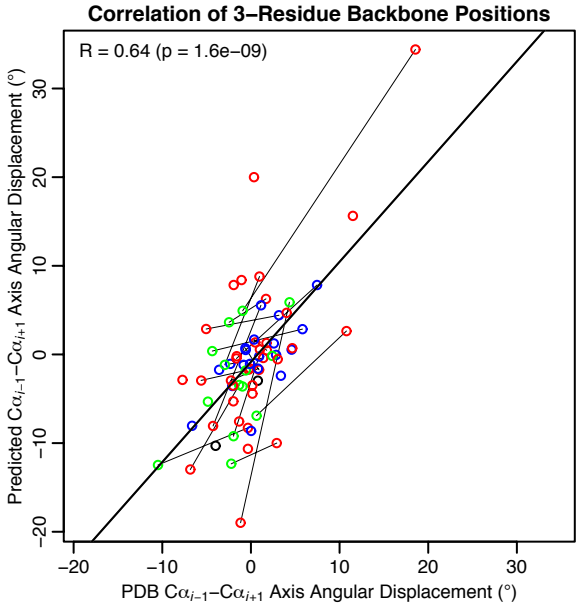
arrays[40]. **(B)** We separated the backbone pseudo-dihedral angles by $\chi 1$ angle and generated normalized histograms for bins -60° (red), 60° (green), and 180° (blue, not shown because the overall population was < 0.05%). Circles indicate the population means ($\langle \tau_{disp} \mid \chi 1 \rangle$). For each alternate Cβ atom position found in the PDB, the $C\beta_{i,\text{initial}}$–$C\alpha_{i-1}$–$C\alpha_{i+1}$–$C\beta_{i,\text{alternate}}$ dihedral angle (also τ_{disp}) is indicated as a vertical line colored according to the $\chi 1$ bin. The overall population of each bin is indicated in the upper right. The RMSD of the population means from the corresponding PDB τ_{disp} angles is shown in the upper left. Representative structures are shown from the **(C)** -60° and **(D)** 60° $\chi 1$ bins. The three simulated residues are shown with a ball and stick representation. Other protein residues are shown using a surface representation. The two crystallographic alternate backbone conformations are shown using a wire representation and colored according to the $\chi 1$ bin. Images were created with VMD.

A simple binary metric indicating if the simulations correctly captured the side-chain/backbone bias is whether the average backbone conformations for each $\chi 1$ bin ($\langle \tau_{disp} \mid \chi 1 \rangle$, circles in Figure 1-4B) were in the same relative orientations found in the PDB (vertical lines in Figure 1-4B). This is easiest to interpret for those residues with PDB side-chain conformations in exactly two $\chi 1$ bins, as in Figure 1-4. There were 98 starting structures where that was the case and of those, in 76 cases the simulations did visit both $\chi 1$ bins observed in the PDB, making the comparison possible. Out of these 76, 55 (73%) showed the correct bias, which is significantly better (chi-square p-value $1 \cdot 10^{-4}$) than would be expected at random (50%). When only buried side chains (SASA <30%, see below) are considered, 15 out of 17 (88%) are correct.

A comparison between the mean τ_{disp} angles from the simulations and those determined from the PDB shows reasonable agreement (Figure 1-5). As the accuracy of

rotamer prediction has been shown to be strongly dependent on the degree of residue burial[41-44], we show results for 24 residues with solvent accessible surface areas (SASA) of <30%, using the surface area of an extended residue flanked by glycines as the reference SASA. Deviations from the diagonal can result from both scoring/sampling problems in our modeling procedure and uncertainty in the crystallographic fitting. However, there is a reasonable positive correlation (R = 0.64). The correlation becomes clearer when individual simulations (connected by lines) are examined. Nearly all such lines show positive slopes, indicating that the simulations capture the direction of correlated side chain and backbone conformational changes correctly in many cases, albeit with some variation in the absolute magnitude.

Figure 1-5. Predicted angular displacements from 3-residue backrub simulations



Backbone angular displacement ($\langle \tau_{disp} \mid \chi 1 \rangle$) is correlated between PDB structures and predicted populations from 3-residue backrub simulations. Points are colored by $\chi 1$ bin: -60° (red), 60° (green), and 180° (blue). Points from the same simulation are connected by thin lines. Any connecting lines with positive slopes represent simulations which show the correct bias between side chain

conformation and backbone conformation. Disconnected points are from simulations where only one of the $\chi 1$ bins seen in the PDB was visited during the simulation. The thick black line shows a least squares linear fit.

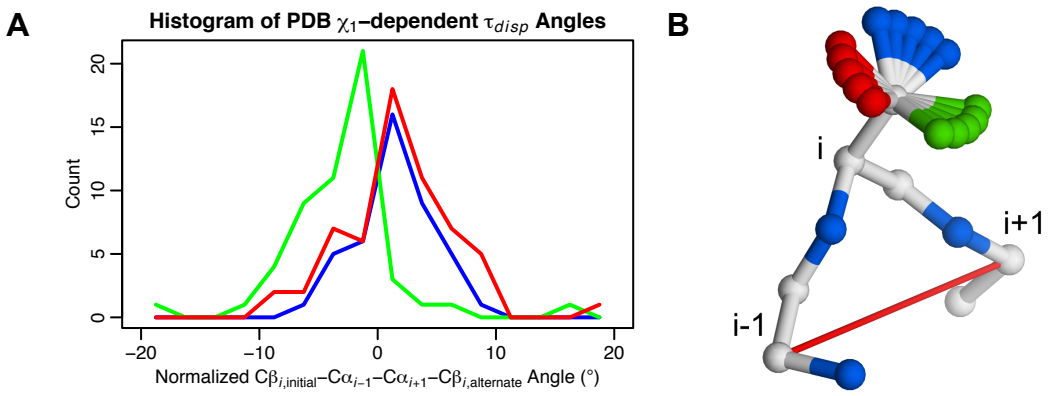
A notable observation is that at certain backbone τ_{disp} angles (i.e. < 3° or > 17° in Figure 1-4), some side chain conformations are completely inaccessible. The intervening backbone conformations form a transitional zone where the rotameric change becomes more and more energetically favorable. In our simulations, there is little evidence for an energetic barrier between the subtle differences in backbone conformations. On the other hand, there can be significant energy barriers involved in side chain transitions, particularly in the protein core. Our data indicate that the side chain rotamers may lock the backbone into slightly different conformations, giving rise to the alternate conformations observed by Davis et al[1]. This mirrors another simulation study, where a side chain transition played a key role in stabilizing a relatively unconstrained backbone conformational transition[45].

Backbone/$\chi 1$ Correlation in Crystal Structures Alone

After observing the correlation of backbone conformation with the side chain $\chi 1$ angle in our simulations, we wanted to determine whether the same biases could be observed at a global level in the Davis et al.[1] dataset, irrespective of the simulation results. To do so, we considered the 68 residues (out of 126) where there were at least two $\chi 1$ bins represented in the PDB. For every alternate backbone conformation, we calculated the τ_{disp} angle, using the first conformation as the reference structure. We then normalized the τ_{disp} angles for each of the 68 residues to make the τ_{disp} weighted mean

(using PDB occupancies as the weights) of each individual residue 0. The distribution of τ_{disp} for each χ1 bin is shown in Figure 1-6.

Figure 1-6. Correlation between crystallographic backbone and sidechain angles



In high-resolution crystal structures, alternate backbone conformations are correlated with the side chain χ1 angle, with a straightforward structural explanation. **(A)** Out of 126 residues in the backrub set, 68 have χ1 angles in multiple rotameric bins. For those residues, the calculated $C\beta_{i,initial}$–$C\alpha_{i-1}$–$C\alpha_{i+1}$–$C\beta_{i,alternate}$ pseudo-dihedral angles (τ_{disp}) described in Figure 1-3B were normalized by the average angle (weighted by PDB occupancy). Histograms of those angles are shown using 2.5° bins and colored by χ1 bin: -60° (red), 60° (green), and 180° (blue). **(B)** The clear difference between the -60°/180° and 60° bins has a straightforward structural explanation, where side chains in the 60° bin push the backbone to the left, and the -60°/180° side chains push the backbone to the right. Hypothetical γ atom positions are colored by χ1 bin.

Interestingly, the distribution of τ_{disp} angles for the 60° χ1 bin was significantly different from the distributions for the -60°/180° χ1 bins. (Figure 1-6A) There was a 5.8° difference in means between the 60° and the joint -60°/180° distributions. The structural explanation for the difference is quite clear when the orientation of the Cβ-C/Oγ bond vector is visualized on a hypothetical backbone with the central side chain pointing up

and the $C\alpha_{i-1}$ atom in front of the $C\alpha_{i+1}$ atom. (Figure 1-6B) In that orientation, the 60° Cβ-C/Oγ vector points nearly perpendicular to the $C\alpha_{i-1}$-$C\alpha_{i+1}$ axis, pushing the backbone in a counter-clockwise direction. That rotation corresponds to a negative τ_{disp} angle. In the -60°/180° bins, the Cβ-C/Oγ vectors point in the opposite direction but are much less perpendicular. The degree of perpendicularity helps explain the negative skew of the 60° distribution in comparison to the relative symmetry of the -60°/180° distributions.

In principle, the dependence of backbone conformation on side-chain conformation could be used to derive coupled moves in sequence and structural optimization algorithms. For example, the differences in backbone distributions could be used to restrict sampling of backbone conformations when switching into or out of the 60° $\chi 1$ rotameric bin.

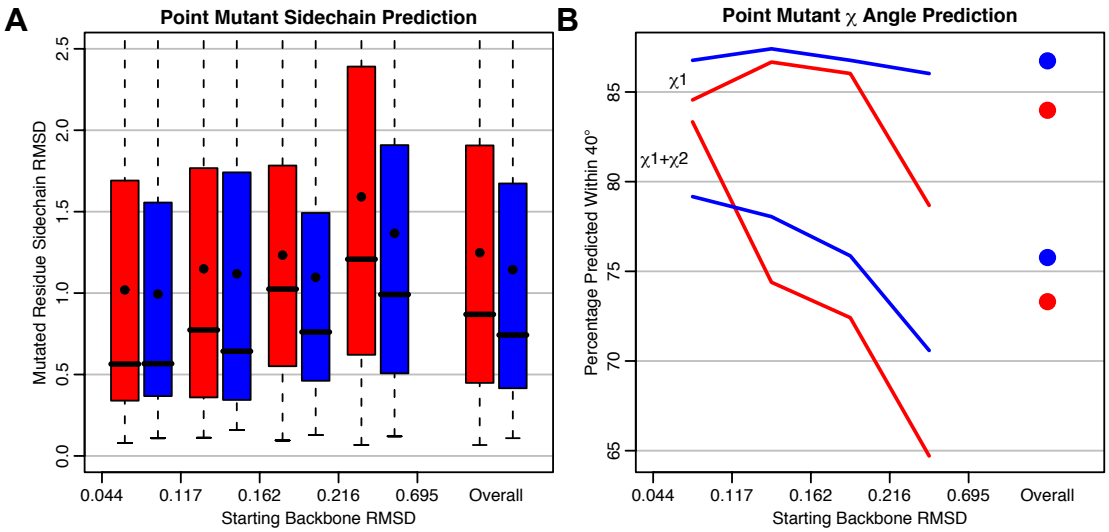
Test 2: Point Mutant Side Chain Prediction

In addition to distinct conformations observed in the high-resolution crystal structure dataset discussed above, another context in which subtle backbone differences may be important are residue point mutation. A single-residue point mutation represents the simplest of increasingly more difficult structural modeling tasks where one is given a template and then must predict the new low energy conformation after a known perturbation. In addition, the ability to accurately predict the conformation of a side-chain upon point mutation has direct bearing on the success of protein sequence design algorithms.

We wanted to determine the extent to which generalized backrub sampling could improve the prediction of point mutant side chains, especially when using a fixed rotamer library as commonly done in computational protein design methods. Recently, Bordner

and Abagyan[46] compiled a large benchmark set of PDB structure pairs differing by a single point mutation. We applied the generalized backrub protocol to locally refine structural models after mutation/fixed backbone rotamer optimization in Rosetta[31]. We found that overall, incorporation of backrub sampling improved both side chain heavy atom RMSD and $\chi 1/\chi 2$ recovery within 40°. (Figure 1-7) We also found that the local backbone RMSD between PDB structure pairs was correlated with prediction difficulty, in terms of both RMSD and $\chi 1/\chi 2$ recovery. The larger the backbone conformational change upon mutation, the larger was the improvement resulting from backrub sampling. In particular, the fraction of pairs with the highest starting RMSD showed the most sizeable improvement. Similar observations were made in a previous study[30] which showed improvements in prediction of side chain conformations after core substitutions in T4 lysozyme when comparing flexible with fixed backbone methods. There backbone flexibility was modeled using a different mechanism employing random continuous adjustments of ±3 degrees to each backbone angle.

Figure 1-7. Point mutant side chain prediction RMSD and chi angle analysis



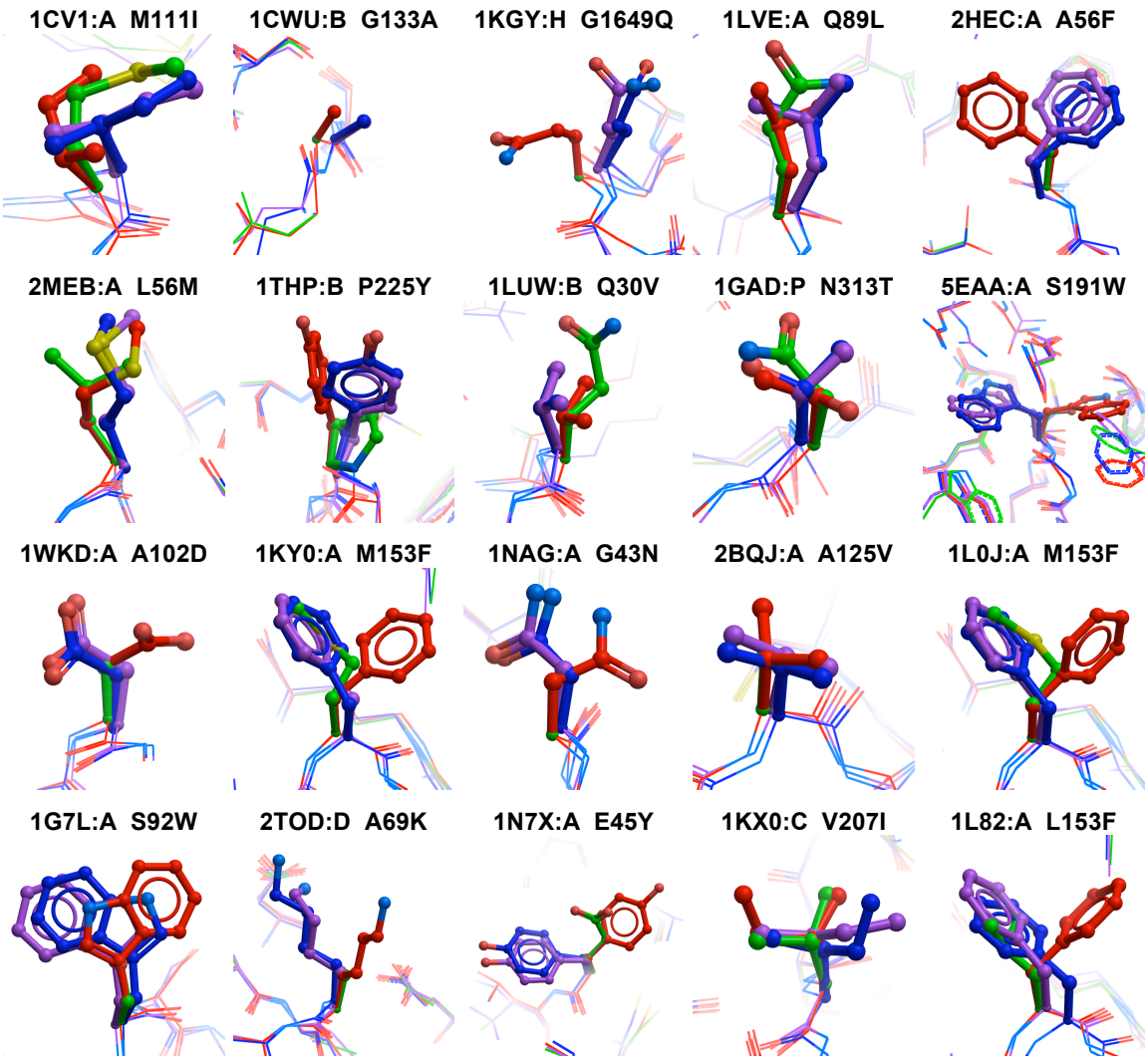
Generalized backrub sampling (blue) improves **(A)** the RMSD and **(B)** $\chi 1/\chi 2$ recovery (within 40°) of predicted point mutant side-chains over fixed backbone sampling alone (red). Prediction results were sorted by increasing starting backbone RMSD and divided into four equally sized groups. The improvement was most distinguished for structure pairs with a larger starting backbone RMSD. Breaks between groups are along the x-axis. Results are shown for 543 non-proline residues for which the solvent accessible surface area of the wild-type residue was < 5%. Residues within a 6Å radius of the wild-type residue were sampled. In the boxplots, boxes indicate the interquartile range (IQR), thick horizontal lines show the median, and dots show the mean. Whiskers extend to the most extreme datapoint within 1.5 times the IQR of the 25th or 75th percentile.

In addition to the dependence on initial backbone RMSD, we also investigated how a number of other factors affected the extent of improvement, including the radius of neighboring residues allowed to change rotameric conformations (4, 5, 6, 7, and 8 Å from the mutated residue) and the degree of burial of the mutated residue (< 5%, < 30%, and ≤ 100% solvent accessible surface area). Figure 1-7 shows results from point mutant predictions that showed the best overall improvement in prediction accuracy, considering

< 5% solvent exposure and a 6Å sampling radius. Between a 4 Å and 8 Å sampling radius, the prediction of side chain RMSD does not change significantly. Backrub sampling gives better $\chi 1/\chi 2$ recovery at 6Å than using any other radius. Considering the amount of residue burial, backrub sampling continues to improve overall RMSD prediction somewhat using a 30% SASA cutoff. When evaluating all residues including those that are largely solvent exposed, backrub sampling still improves predictions for high backbone RMSD pairs, but makes low RMSD pairs slightly worse.

Figure 1-8 shows examples of backrub sampling improving side chain prediction. The improvement can come from two sources, namely better prediction of the side-chain conformation and better prediction of the protein backbone. In some cases, the side-chain improvement comes at the cost of backbone prediction accuracy, as is shown in the last row of images. However, the worsening of backbone ($C\alpha/C\beta$) RMSD is relatively small compared with the improvement in side chain RMSD. (Table 1-1) The source of the error could lie in crystallographic uncertainty, inaccuracies in the scoring function, or compensation for a discretized rotameric side-chain representation.

Figure 1-8. Examples of improved side chain prediction



The fixed backbone prediction is shown in red and the backrub prediction is shown in blue. The starting PDB structure is shown in green and the target mutated PDB structure is shown in purple. Nitrogen and oxygen atoms are shown in light blue and red, respectively. Examples are sorted by the improvement in mutant residue Cα/Cβ RMSD from fixed backbone to backrub protocols. The modeled mutation M153F in 1KYO and 1LOJ is the same, but 1KY0 has a leucine at positions 118 & 121 whereas 1L0J has a methionine at positions 118 & 121. In both cases, backrub sampling correctly shifts the backbone at residue 153 and better recovers the target side chain. In the last five examples, backrub sampling increases the Cα/Cβ RMSD but improves the side chain prediction.

Cα/Cβ and side-chain RMSDs are listed in Table 1-1. Images were created using ICM Browser.

Table 1-1. Cα/Cβ and side-chain RMSD for examples of point mutant predictions

PDB		Mutated Residue Cα/Cβ RMSD			Side Chain RMSD		
Chain	Mutation	Fixed BB	Backrub	Delta	Fixed BB	Backrub	Delta
1CV1:A	M111I	1.13	0.21	-0.91	2.41	0.37	-2.04
1CWU:B	G138A	1.03	0.16	-0.87	NA	NA	NA
1KGY:H	G1649Q	1.22	0.42	-0.80	4.54	1.21	-3.33
1LVE:A	Q89L	0.64	0.09	-0.55	1.14	0.30	-0.84
2HEC:A	A56F	0.70	0.16	-0.55	3.70	0.98	-2.71
2MEB:A	L56M	0.57	0.08	-0.49	1.18	0.59	-0.60
1THP:B	P225Y	0.75	0.31	-0.44	2.24	0.38	-1.86
1LUW:B	Q30V	0.64	0.23	-0.41	1.01	0.32	-0.69
1GAD:P	N313T	0.43	0.13	-0.30	2.61	0.27	-2.34
5EAA:A	S191W	0.68	0.40	-0.28	7.47	0.88	-6.60
1WKD:A	A102D	0.46	0.22	-0.24	2.74	0.57	-2.17
1KY0:A	M153F	0.54	0.32	-0.22	3.66	0.49	-3.17
1NAG:A	G43N	0.45	0.24	-0.21	3.10	0.37	-2.73
2BQJ:A	A125V	0.44	0.24	-0.20	2.50	0.42	-2.08
1L0J:A	M153F	0.34	0.24	-0.10	3.61	0.24	-3.37
1G7L:A	S92W	0.34	0.45	0.11	3.71	0.81	-2.90
2TOD:D	A69K	0.16	0.32	0.16	4.36	1.36	-3.00
1N7X:A	E45Y	0.13	0.31	0.19	6.60	0.65	-5.95
1KX0:C	V207I	0.14	0.44	0.29	3.29	0.76	-2.53
1L82:A	L153F	0.48	0.93	0.45	3.70	0.98	-2.72

Examples were selected from cases where backrub sampling improved side-chain prediction as shown in Figure 1-8. The majority of the selected examples also showed improvement in prediction of the backbone, although this was not always the case.

Test 3: Triosephosphate Isomerase Loop 6 Simulation

As a third, proof-of-principle test of the backrub sampling protocol, we investigated a much larger conformational change. The hinge motion of triosephosphate isomerase (TIM) loop 6 is a well-characterized example of a protein segment undergoing significant conformational change while maintaining a relatively rigid internal conformation. The Cα RMSD between an 11 residue segment (V167-T177) in the closed, (PDB 2YPI[47]), and open (PDB 1YPI[48]) conformations is 4.6 Å. We found that by

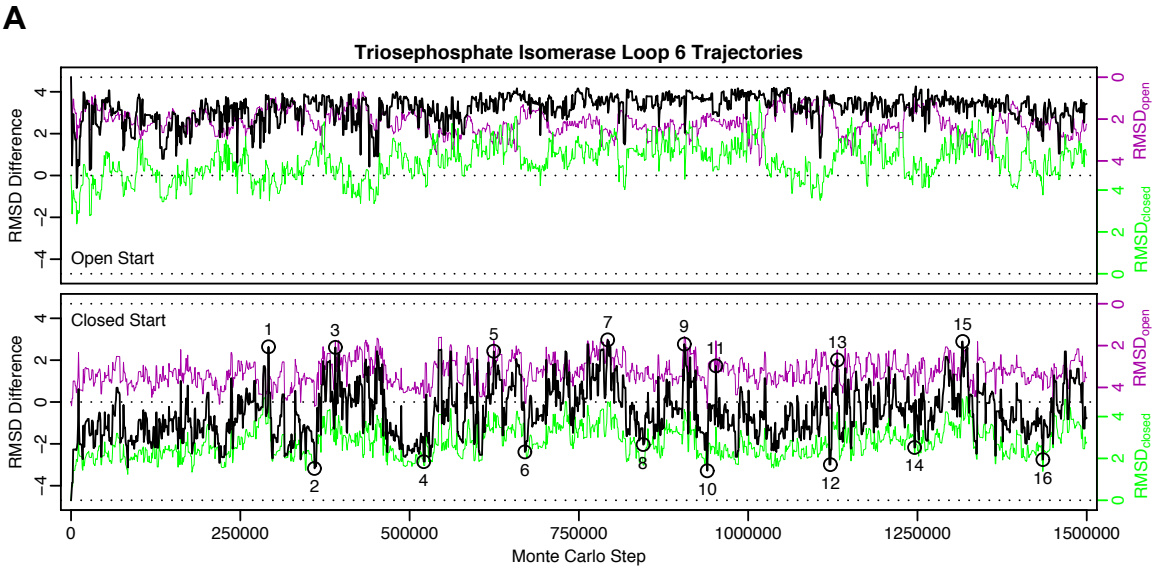
manually rotating the closed segment 50° about the $C\alpha_{167}$–$C\alpha_{177}$ axis, the Cα RMSD to the open form drops from 4.6 to 1.1 Å, indicating that a backrub simulation may capture much of the conformational variability of the TIM loop. However, such a large rotation introduces several significant steric clashes and adds sizable backbone bond angle strain at residue V167.

We wanted to determine whether generalized backrub sampling of the loop region could capture the same degree of conformational variability without producing energetically unreasonable conformations. Previous studies using molecular dynamics have had difficulty capturing the TIM loop 6 conformational transition[49,50]. Notably, the simulations required temperatures from 1000-1200 K to see transitions from one state to the other.

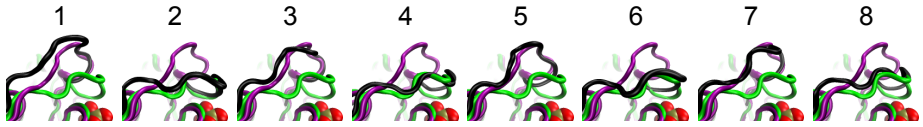
We ran simulations starting from both the open conformation (1YPI) and the closed conformation (2YPI). In each simulation, we allowed backrub moves of size 2-12 on residues 165-179. Residues 128-130 showed small but potentially significant changes between the two conformations, so we allowed backrub moves of size 2-3 for those residues. In addition to all of those residues, rotamer changes were allowed for residues whose side chains were in the vicinity of the loop using a 5 Å cutoff and by visual inspection (3, 7, 95, 96, 131, 134, 139, 164, 180, 183, 208, 211, 216, 219, 220, 223, 230). Each simulation was run without the ligand, making the atomic composition identical. We ran the simulations for 1.5 million Monte Carlo moves, using a temperature of 302 K in the Metropolis criterion. Each simulation took 14 hours to complete on a single 2.0GHz Xeon processor.

To analyze the simulation trajectories, we calculated the Cα RMSD of the loop from both the open (1YPI) and closed (2YPI) conformations. The sign of the difference between those RMSDs indicates whether the loop is closer to the open (positive sign) or closed (negative sign) conformation. Starting from the closed conformation with the ligand removed, the backrub simulations were able to oscillate between the open and closed forms of the loop many times during a 1.5 million step simulation. Eight example transitions are pictured in Figure 1-9, where the minimum RMSD for each approach to the open form ranged 1.57-2.2 Å and the RMSD values for return to the closed form ranged 1.37-2.36 Å. The loop structure (V167-T177) maintained a relatively stable internal conformation over the length of the simulation, with an average aligned Cα RMSD of 1.3 Å (0.3 Å standard deviation) from the starting structure.

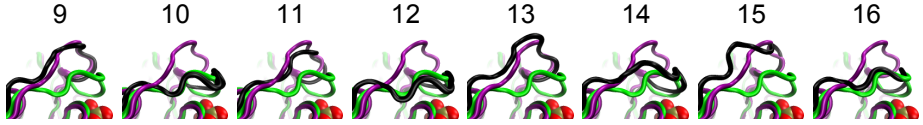
Figure 1-9. Generalized backrub sampling of triosephosphate isomerase loop 6



B



	1	2	3	4	5	6	7	8
$RMSD_{open}$	2.49	4.90	1.74	4.51	1.76	4.75	1.73	3.71
$RMSD_{closed}$	5.12	1.73	4.35	1.66	4.19	2.36	4.71	1.68
Difference	2.64	-3.17	2.61	-2.84	2.42	-2.39	2.97	-2.03



	9	10	11	12	13	14	15	16
$RMSD_{open}$	1.57	5.02	1.77	5.02	2.20	4.39	1.96	4.14
$RMSD_{closed}$	4.34	1.74	3.50	2.04	4.22	2.22	4.85	1.38
Difference	2.77	-3.28	1.73	-2.99	2.01	-2.17	2.89	-2.76

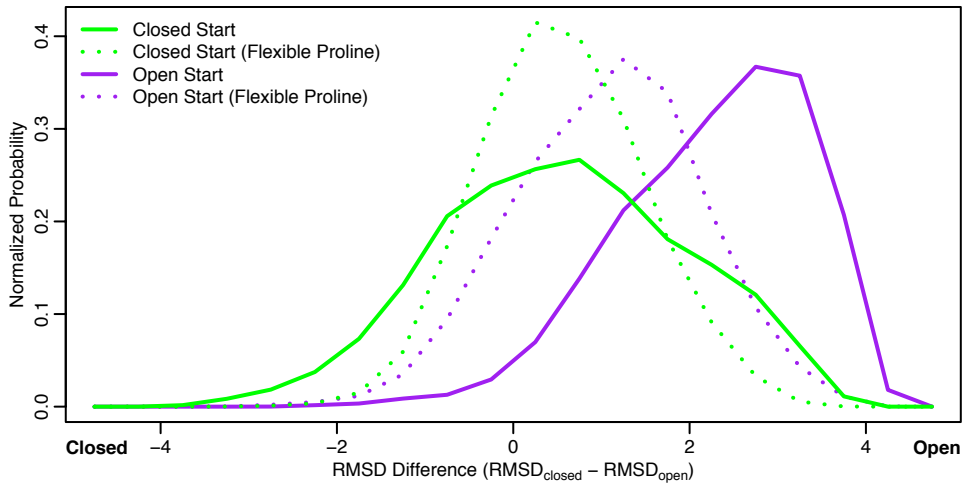
(A) For every 200 accepted moves in the simulations, we calculated the Cα RMSD of the loop from both the open (1YPI, purple lines) and closed (2YPI, green lines) PDB structures. We defined a single reaction coordinate for the simulations as $RMSD_{closed} - RMSD_{open}$ (black lines). The green and purple lines are plotted with modified axes such that the black line is the sum of those component lines. The simulation starting in the open conformation is on top and the simulation starting in the closed conformation is on the bottom. The simulation starting from the open conformation makes an initial excursion closer to the closed conformation but then stays open for the remainder. The simulation

starting in the closed conformation alternates back and forth between the open and closed conformations at least eight times during the simulation. **(B)** The open structure (1YPI) is shown in purple. The closed structure (2YPI) is shown in green. For reference the substrate analogue, 2-phosphoglycolate, is shown using space fill. (It was not present in either simulation.) The conformation at the numbered Monte Carlo step is shown in black.

We found that the motion of the loop depended on the starting structure, not always showing the opening and closing behavior. In the simulation starting from the open conformation, there was a transient excursion closer to the closed form (within 2.38 Å) at the beginning of the simulation. After that, the loop stayed in a predominantly open conformation for the remainder of the simulation, in some cases migrating to a "hyper open" state up to 8.29/4.23 Å from the closed and open structures, respectively.

There are several explanations for the difference in the simulations and lack of convergence. In addition to possibly needing more sampling to equilibrate, it may be that the anchor points or other fixed regions of the different starting protein structures bias the loop motion. Another possible explanation is that the backrub motions are not sufficiently sampling the internal degrees of freedom in the loop. A likely limitation is that proline residues (at TIM residue positions 168 and 176) are not currently allowed as pivot residues in backrub sampling, thus keeping all of their internal angles fixed. This effect may be substantial in the TIM loop case as P168 shows a ψ angle change of 40° between the two conformations. (Figure 1-10)

Figure 1-10. TIM loop 6 simulation starting structure and proline bias

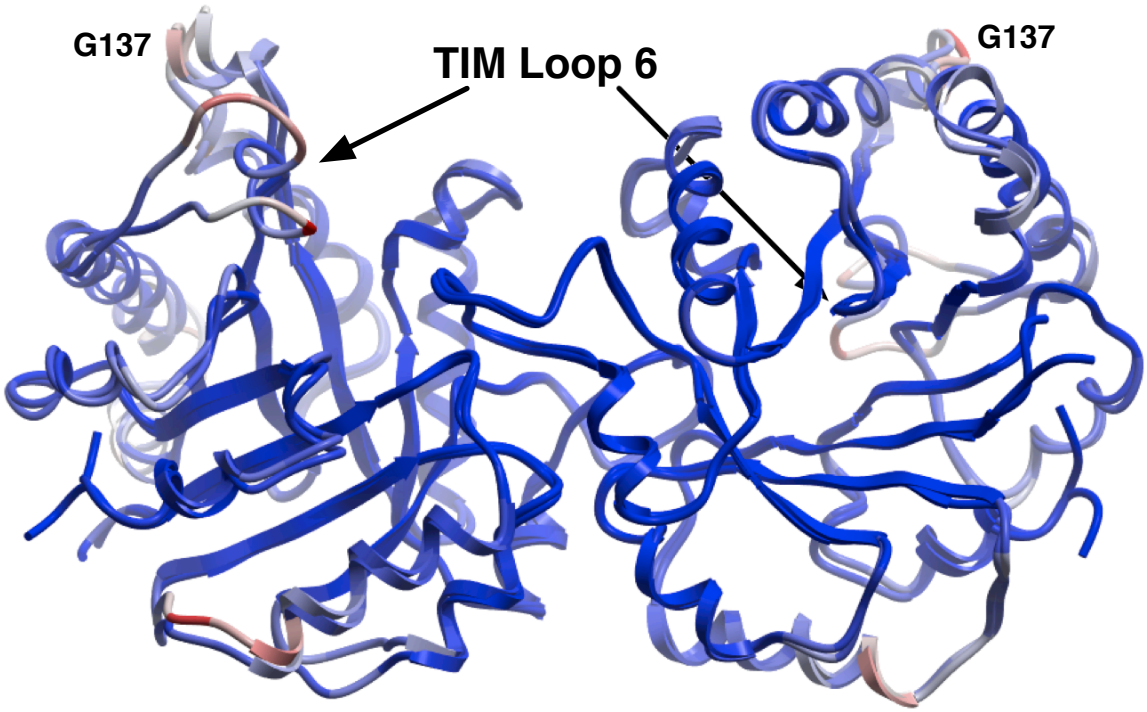


The TIM loop 6 simulations are biased by the starting structure (see main text, Figure 9), but the bias is reduced by including approximate proline flexibility in extended simulations. From each of the open (1YPI) and closed (2YPI) starting structures, 20 extended simulations were run for 20 million moves. A snapshot was recorded every 2,500 moves. The last half of every simulation (4,000 snapshots) was used to calculate probabilities of the loop being closer to the closed or the open form. As is shown in Figure 9 in the main text, simulations starting from the closed form (solid green) cover a range of conformations between the open and closed forms. Simulations starting from the open form (solid purple) are more strongly weighted towards the open conformation. To determine the effect of keeping proline backbones fixed, simulations were run in which the proline backbone was allowed to move, but any strain introduced to the bond lengths or angles of the proline side-chain were not penalized. Simulations starting from the closed form (dotted green) show a sharper distribution, while those starting from the open form (dotted purple) shift towards the closed form distribution. This suggests that the 40° ψ angle difference between the open and closed forms at P168 is significant but does not completely explain the lack of convergence that persists to some extent even after extended simulations. Extending the backrub move set in the future (for example by including shearing

moves, etc.) or combination with other sampling protocols may help overcome some of these limitations.

To assess the degree to which more complete backrub sampling (not limiting sampling to just the loop 6 region) captures the flexibility of the TIM structure, we ran multiple simulations of the complete TIM dimer (with the constraint of fixing the backbone coordinates of 17 core residues in each monomer). The loop 6 region does indeed show the largest conformational variability: three of the four largest calculated B-factors from those simulations are in the tip of loop 6 (G171-G173). (Figure 1-11) In addition to the high calculated B-factors for loop 6, some flexibility was observed in several other regions. These regions also showed structural differences between the open and closed crystal structures.

Figure 1-11. TIM fluctuations observed in whole protein sampling



In backrub simulations of the whole TIM dimer, residues in loop 6 show the highest computed average B-factors. Calculated Cα B-factors are mapped onto the closed (1YPI) and open (2YPI) starting structures with a blue-white-red color scale indicating low to high fluctuations. Three out of the top four most flexible residues (G171, G173, G137, and T172) are in loop 6. For all residues, the B-factors are somewhat correlated (R = 0.62) with the distance between Cα atoms in the two structures. This indicates that the backrub simulations are recapitulating flexibility hinted at by the crystallographic heterogeneity (although alternative conformations in different crystal structures may be caused by factors other than intrinsic flexibility, such as crystal packing). For each starting structure, 20 independent trajectories were run for 250,000 MC steps, with a snapshot recorded every 2,500 steps. 17 core residues (8, 41, 62, 63, 75, 76, 92, 93, 94, 163, 206, 207, 208, 226, 228, 229, and 248) were held fixed during the simulations. Core residues were identified as those having the most heavy atoms within a 10 Å radius of the Cα atom in 1YPI, including at least one residue from each β strand.

B-factors were calculated separately for each starting structure using all 2,000 recorded snapshots from the independent simulations.

Testing for Even Conformational Sampling

If true thermodynamic properties are desired from a Monte Carlo simulation, it is necessary that detailed balance be preserved. An important requirement of detailed balance is that conformational space is evenly sampled, subject to whatever constraints are placed on the simulation. In doing so, one must first define which space one wishes to evenly sample. There are two obvious spaces in which even sampling could be maintained, namely Cartesian space and internal coordinate space. Even sampling of internal coordinates does not necessarily evenly sample Cartesian space, as implied by the spherical coordinate Jacobian determinant.

$$\left|\frac{\partial(x,y,z)}{\partial(r,\theta,\phi)}\right| = r^2 \sin\phi$$

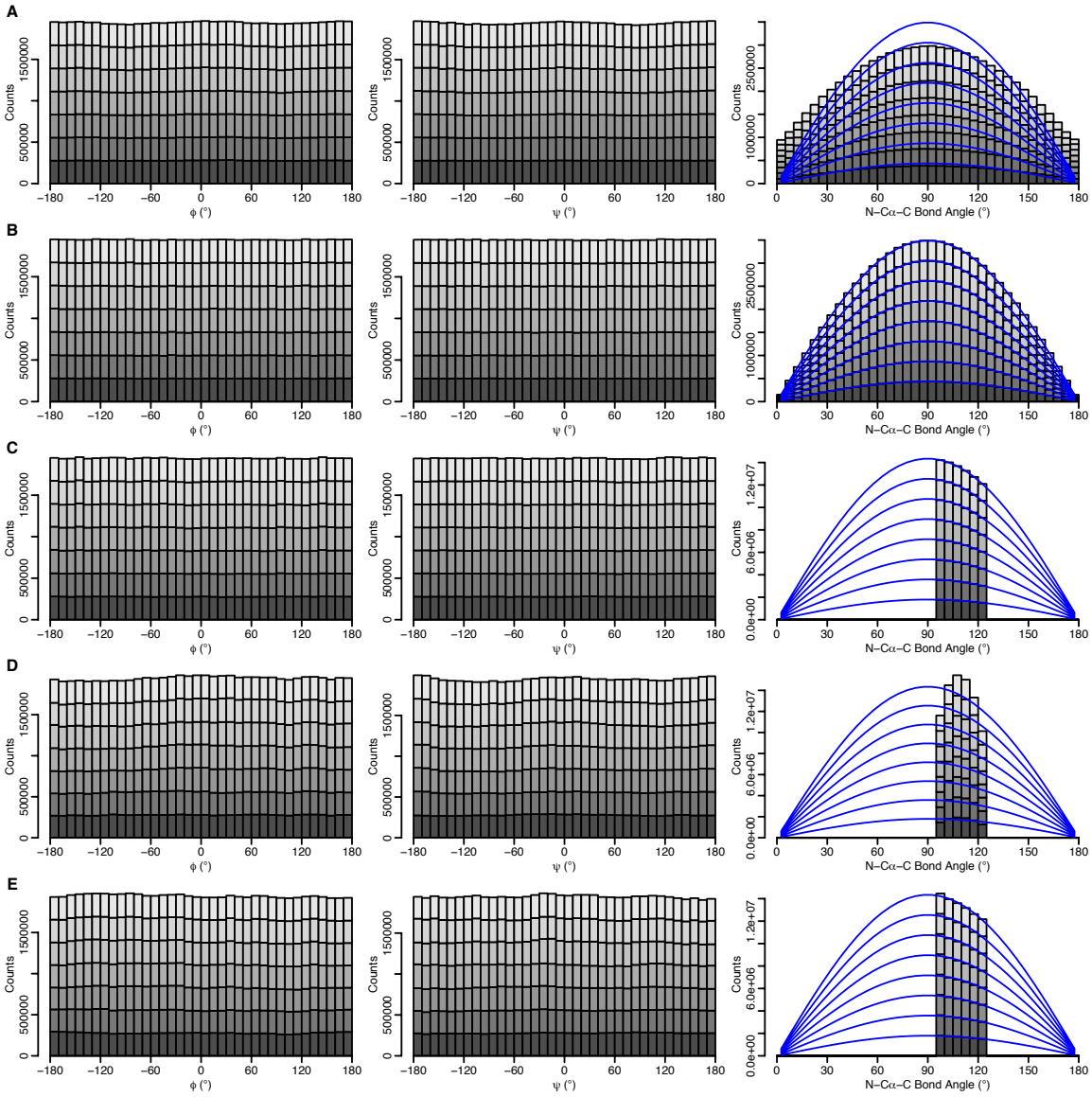
Here *r* represents the radius, θ represents the azimuth angle ranging -180 to 180°, and ϕ represents the zenith angle ranging 0 to 180°. Out of the determinant come expected distributions of each spherical coordinate, the most relevant being an even distribution for azimuth angles and a sin distribution for zenith angles. In internal coordinates, torsion angles are azimuth angles. Thus even sampling of torsion angles implies even sampling in Cartesian space, and vice versa. However, bond angles are instead zenith angles and would be expected be found in a sin distribution.

To determine how backrub sampling affects sampling of conformational space, we ran simulations without any force field on 8-polyalanine, similar to previously described tests[4,5,51]. To avoid gimbal lock, bond angles were constrained to 0.1-179.9°.

To avoid artifacts coming from fixed endpoints, 1% of moves consisted of choosing a random ϕ, ψ, or α angle and selecting a new value from a uniform (ϕ/ψ) or sin (α) distribution. The other 99% of moves consisted of a standard backrub move with τ_{max} set to 180°. Test simulations were run for 10^7 steps.

We implemented the derivatives as described by Betancourt[7], checking each for correctness numerically, and then ran simulations using his acceptance criterion. Histograms of the corresponding backbone degrees of freedom are shown in Figure 1-12A. The bond angle distributions did not follow a sin distribution and instead showed greater sampling at the extremes of the distribution. Sampling α angles from an even distribution during the 1% of moves incurred negligible changes to the bond angle distributions. (data not shown)

Figure 1-12. Test of uniform sampling by backrub move



Backrub moves evenly sample backbone degrees of freedom without the weighting procedure described by Betancourt[7]. Individual residues of 8-polyalanine are shown using different shades of gray. ϕ (C-N-Cα-C torsion) angle distributions are shown for residues 2-8. ψ (N-Cα-C-N torsion) angle distributions are shown for residues 1-7. N-Cα-C bond angle distributions are shown for residues 1-8. Blue lines indicate the theoretical distributions expected from a spherical zenith angle given even Cartesian sampling. **(A)** Betancourt[7] weighting skews N-Cα-C bond angle selection probabilities to the extremes of the

distribution. **(B)** Removing Betancourt weighting results in correct distributions of all sampled backbone internal angles. **(C)** Limiting the N-Cα-C bond angle to the 95-125° interval continues to preserve expected distributions. **(D)** Limiting τ_{max} to 20° leads to over-selection of N-Cα-C bond angles at the center of the allowed interval. **(E)** Incorporation of the correction procedure described in Materials and Methods restores the correct distributions.

To determine whether the Betancourt acceptance criterion was necessary for evenly sampling Cartesian space, we repeated the simulation without it. The internal coordinate distributions exactly matched those expected (Figure 1-12B), indicating that backrub moves inherently produce even sampling of Cartesian space. That result can be explained using the spherical coordinate Jacobian determinant, if one aligns the atoms involved in the backrub move in the proper coordinate frame. By placing $C\alpha_i$ at the origin and $C\alpha_j$ along the positive z-axis (that is, $\phi = 0°$), the backrub move then only results in changes to the θ coordinate for intervening atoms. According to the Jacobian determinant, if θ is sampled evenly, then Cartesian space is sampled evenly. This makes the backrub move a particularly straightforward addition to any Monte Carlo protocol preserving detailed balance in Cartesian space.

We also tested for even sampling when bracketing α angles using $I_{bond\ angle}$. To do so, we limited α angles to the interval 95-125° and repeated the simulation. That too resulted in correct internal coordinate distributions. (Figure 1-12C) We next tested bracketing the τ angle using $I_{rotation\ angle}$ with a τ_{max} value of 20°. Without the weighting described in Materials and Methods, bond angles in the middle of the allowed interval are oversampled because of their increased probability of being within $I_{bond\ angle} \cap I_{rotation\ angle}$.

(Figure 1-12D) However, addition of the weighting protocol restores the correct distributions. (Figure 1-12E)

In addition to backbone degrees of freedom, the branching atom internal coordinates are also modified slightly by the generalized backrub move. As described in Table 1-3 and shown in Figure 1-16, for a given α angle, each branching atom internal coordinate, ω_i, is calculated with the following formula.

$$\omega_i = A_i + B_i\alpha + C_i\alpha^2$$

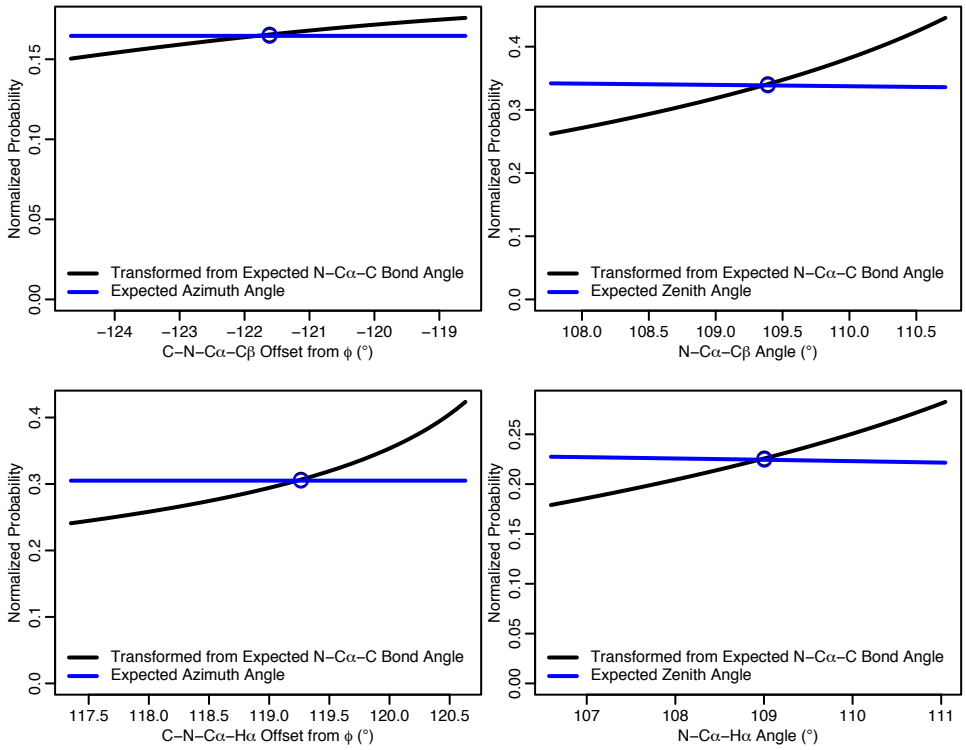
To determine the effect of that transformation on the expected populations of ω_i, we transformed the expected α angle distribution using the following formula.

$$P(\omega_i) = \frac{P(\alpha)}{\left|d\omega_i / d\alpha\right|} = \frac{\sin\alpha}{\left|B_i + 2C_i\alpha\right|}$$

Plots of $P(\omega_i)$ vs. ω_i for each branching atom internal coordinate are shown in Figure 1-13. Importantly, none of the distributions agree with what would be expected from even Cartesian sampling. This is expected because even without using a force field, the branching atom optimization imposes knowledge from the force field that causes greater sampling of Cβ/Hα positions that have favorable energies. Over the 99.7-119.7° range of transformed α angles, the discrepancy is up to 39%. This indicates that if strict observance of detailed balance is desired for side chains, the quadratic functions described in this work should not be used. A detailed balance preserving alternative would involve keeping those degrees of freedom fixed during backbone movement and sampling them separately using the expected distributions. In practice, the impact of the quadratic function update procedure may be negligible compared with errors in the

energy function. In addition, the impact is limited because the discrepancy is much less for those α angles most likely to be accepted by the force field bond angle potential.

Figure 1-13. Nonuniform branching atom sampling using optimized placement



Placement of branching atoms using pre-fit quadratic functions results in an imbalance in the selection probabilities of branching atom internal angles. Using the Amber quadratic coefficients for non-glycine residues, the expected (without a force field) zenith angle distribution for N-Cα-C bond angles along the interval 99.7-119.7° was transformed into the expected distributions for all branching atom internal angles. (black lines) Those distributions disagree with the expected azimuth and zenith angle distributions given even Cartesian sampling. (blue lines) The expected zenith angle distributions follow a sin function but appear flat here due to the limited angular range shown on the x-axis. The location of the overall bond angle minimum (N-Cα-C = 109.7°) is indicated with circles.

Discussion

We have shown that the backrub sampling method is useful for sampling small, high-resolution conformational fluctuations as well as a larger, functionally relevant conformational change. In addition to capturing the structural variability of single sequences, generalized backrub sampling also improves modeling of changes to protein structures upon point mutation. While many of the backbone movements are less than 1 Å, they can result in significant displacements of the attached side chains. In addition, the localized breathing motion that backrub sampling emphasizes can allow otherwise energetically unfavorable rotameric transitions.

This work supports the conclusion advanced by Davis et al[1] that protein backbones are influenced by side-chain conformations in a predictable manner, complementing the accepted notion that side-chain conformations can be backbone dependent. In backbone-dependent rotamer libraries, the side chain conformation is influenced by the ϕ and ψ angles of the residue itself[42]. Our simulations and analysis support the notion of a second-order correlation between a central side-chain and the protein backbone in adjacent residues. The 3-residue simulations indicate that the energetic barriers between the relevant backbone conformations can be significantly less than those typically associated with side-chain rotamer transitions.

As a sampling method, the overall philosophy behind the move used here is somewhat different from other methods (although bearing similarities to local perturbation approaches highlighted earlier[5,6]). First, it is a generalization of movement that is observed in nature at both small and large amplitude. Rotameric sampling was likewise inspired by observations made from crystal structures. Second, instead of

treating bond angles as inviolable, it takes advantage of the small but significant flexibility in the bond angle to move a set of backbone atoms through a single, unified rotation. As indicated by the 3-residue simulations, backrub sampling helps free the protein backbone to explore an ensemble of conformations around the native state. While molecular dynamics could be used to accomplish the same goal, correlated movement of atoms can take a considerable number of time steps, unless there is already a set of forces accelerating the atoms in a concerted direction. Simultaneous, correlated rotation of many atoms is one of the strengths of rotamer sampling. Backrub sampling shares that strength.

Backrub moves are biased towards sampling hinge-like protein motion. Another type of motion sometimes seen in proteins is a shearing move, where a subsection of the protein is translated laterally in relation to the remainder of the protein. That type of motion is almost completely orthogonal to the generalized backrub move described in this work. However, the same philosophy as defined originally for the backrub move[1,7] and described here could be applied to model shearing moves directly. This would require four Cα atoms as pivot points. One embodiment may consist of rotating $C\alpha_2$ about $C\alpha_1$ in the $C\alpha_1$-$C\alpha_2$-$C\alpha_3$ plane, and rotating $C\alpha_3$ about $C\alpha_4$ in the $C\alpha_2$-$C\alpha_3$-$C\alpha_4$ plane, such that the $C\alpha_2$-$C\alpha_3$ distance is preserved. Though somewhat more complex, this type of move may help model subtle shifts of alpha helices and other structural elements by small but significant distances. While development of additional move sets may prove useful in capturing the full range of protein motion, another promising avenue might involve combination of backrub and rotamer sampling with traditional molecular dynamics in a hybrid Monte Carlo approach.

The combination of backrub-inspired backbone flexibility with side chain sampling and protein design protocols has a number of useful practical applications. First, we show here that employing backrub motions in a high-resolution refinement protocol improves mutant side chain predictions in two large datasets, comprising 126 backrub motions in 19 high-resolution structures and 2,023 pairs of protein point mutant structures. These results suggest that backrub sampling may enhance the applicability and accuracy of methods to estimate the change in fold stability or binding affinity of proteins upon point mutation[30,52-55]. Second, we show in related work that incorporation of backbone flexibility using the backrub model significantly increases the agreement between modeled side chain conformational variability in folded proteins and side chain relaxation order parameters measured by NMR (Friedland et al., in press). Such simulations may provide insights into protein dynamics and mechanisms of correlated motions. Finally, the use of near-native backbone ensembles has been shown to broaden the set of sequences identified by computational methods and result in successful designs[33]. Similarly, we find that design simulations employing backrub-generated backbone ensembles predict protein sequence families more similar to those observed by experimental phage display selection methods than predictions using just the crystallographic backbone (Humphris & Kortemme, unpublished data). Given its relative simplicity in implementation and ability to capture relevant conformational changes inspired by observed alternative conformations in high-resolution structures, the backrub method may be generally useful for a broad spectrum of side chain sampling and protein design protocols.

Methods

Generalized Backrub Move

The backrub move (Figure 1-1) is applied to an internal protein segment two or more residues long and consists of a geometric rotation by a random angle, τ, about an axis defined by the flanking Cα atoms. The move simultaneously changes 6 internal backbone degrees of freedom in the protein, namely the ϕ and ψ angles at both pivot points and the N-Cα-C bond angle, α, at both pivots. (Variable names follow the conventions of Betancourt[7] instead of Davis[1], which uses τ for the N-Cα-C bond angle.)

The sampling strategy employed here is similar to the one described by Betancourt[7], in that three types of moves are used, namely backbone only, rotamer only, and rotamer/backbone. However, the move selection is significantly different. We were interested in selectively sampling backbone motion in specified local regions of the protein while keeping other regions fixed. Therefore, we devised a flexible scheme for specifying which parts of the protein structure were variable. At the highest level, the operator indicates for each residue whether to sample the backbone, side chain or both. Backrub moves are only allowed for segments where backbone sampling is enabled for both the beginning (i) and ending (j) residues, and all intervening residues. Because the proline side-chain rejoins to the backbone at the amide nitrogen, it has been excluded as a pivot point. In addition, the minimum and maximum segment size (j-i+1) can be varied. By default, the minimum segment size is 2, corresponding to a rotation of the atoms making up the peptide bond between two consecutive Cα atoms. The default maximum segment size is 12, although higher or lower values may be desired depending on the application.

Given that information, a sparse upper-triangular boolean matrix, B, is created where $B[i,j]$ indicates whether a move starting at residue i and ending at residue j is permissible. B can then be further modified to enable or disable individual residue segments. Before beginning Monte Carlo sampling, a data structure is generated from B that lists each possible segment size, along with all starting residues compatible with that particular segment size. Segment selection then becomes the simple procedure of first selecting a random segment size uniformly from all allowed segment sizes, and then selecting a random segment from all allowed segments with the selected size. As there are fewer long segments than short segments, individual long segments will be selected slightly more often than individual short segments.

Monte Carlo Sampling Protocol

During the course of an actual Monte Carlo simulation, the protocol described in Figure 1-2 is used to perform each move. At the beginning of a step, a decision to make a rotamer only move is made according to the adjustable probability, $P_{rotamer}$. The default value of $P_{rotamer}$ is 0.25. If a rotamer only move is chosen, a single variable side-chain is randomly selected and a rotamer is chosen from a library generated using a backbone-dependent rotamer library[31,56]. The rotamer library is initialized using the ϕ/ψ angles from the starting structure and not updated during the simulation. If a rotamer only move is not made, then a random segment and angle is selected as described previously, and the rotation is applied. At that point, the algorithm decides whether to terminate the move (leaving it as a backbone only move) according to the second adjustable probability, $P_{backbone}$. The default value of $P_{backbone}$ is 0.75 to emphasize the more frequently accepted backbone only moves. If the move is not ended, then one or two residues (respective

probabilities 0.75 and 0.25) are selected from along the length of the perturbed backbone segment and random rotamers are chosen for those residues. After all structural perturbations are complete, the move is evaluated using the Metropolis criterion and the Rosetta energy function. Constraints on the degree of angular perturbation and example acceptance probabilities are given in subsequent sections.

Rosetta Scoring Function

In a previous implementation[7] of the move described here, N-Cα-C bond angles were not energetically scored and were constrained to being within 10° of the median bond angle observed in PDB structures. In this work we used bond angle potentials from the Amber ff94[57] and CHARMM22[58] force fields.

In addition to the added bond angle term, the Rosetta full-atom scoring function[59] uses several bonded terms including a ϕ/ψ angle term based on Ramachandran distributions and a χ angle term based on Dunbrack rotamer statistics. For evaluating non-bonded interactions, Rosetta uses a van der Waals term resembling a Lennard-Jones potential, an explicit geometry-dependent hydrogen bonding term[54], a short-range electrostatics term approximated by a residue-specific pairwise distance potential, and the Lazaridis/Karplus implicit solvation model[60].

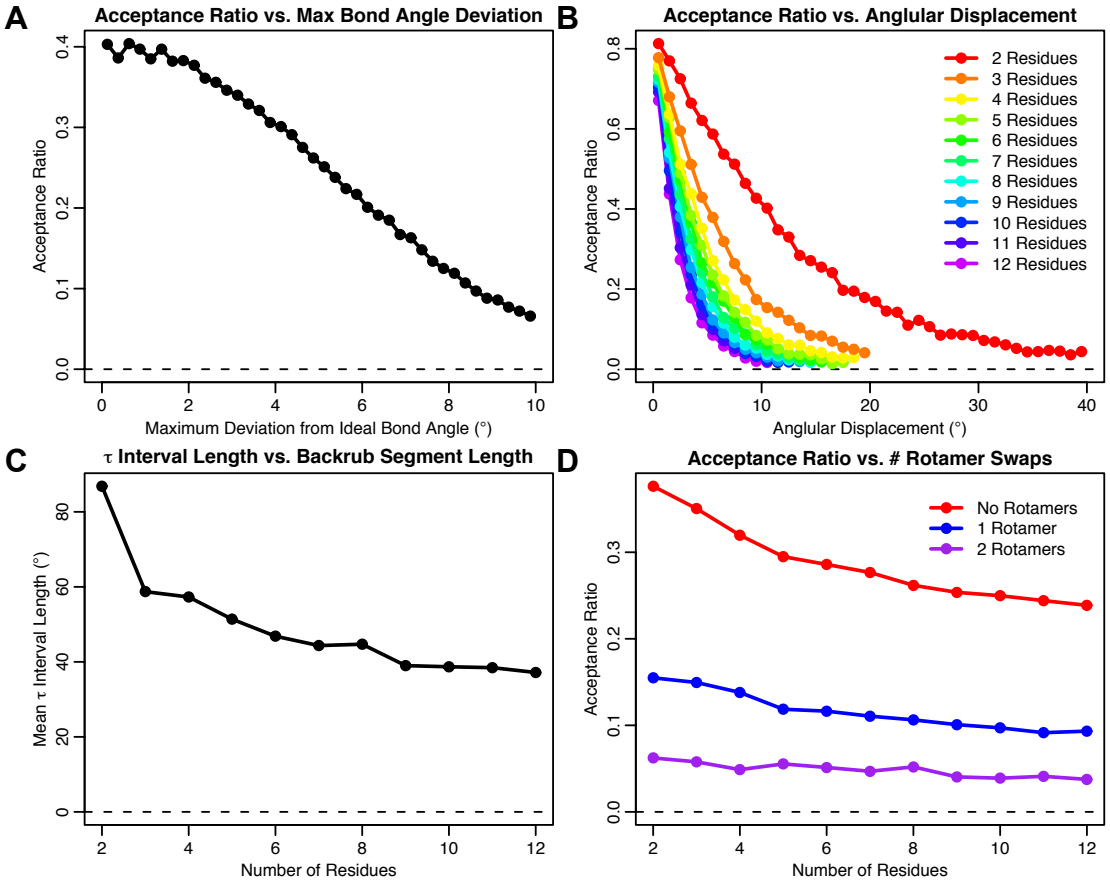
Bond Angle Constraints

In order to reduce the amount of bond angle strain imposed, we sought to bracket the randomly chosen rotation angle such that the bond angle strain never exceeds a threshold value, α_{max}. We used a previously described method[7] to analytically determine the set of τ intervals satisfying that constraint. Briefly, the method involves solving for τ a trigonometric equation that relates α to τ and then plugging in $\alpha_{ideal} - \alpha_{max}$ and $\alpha_{ideal} +$

α_{max} for both the starting and ending residues. The resulting values of τ establish the intervals of allowed τ angles. We term that set of intervals $I_{bond\ angle}$.

To determine how the acceptance ratio decays for increasingly strained bond angles, we performed a long (10^6 step) Rosetta Monte Carlo simulation using a PDZ domain structure (PDB 2H3L[61]), imposing the Amber bond angle potential and limiting bond angles to within 10° of the overall bond angle minimum. Move attempts were binned by the maximum deviation (at either pivot point) from the Amber ideal bond angle and acceptance ratios were calculated (Figure 1-14A). The acceptance rate remained above 20% for all moves where both bond angles remained within 6.25° degrees of ideal. At the extreme, where one of the bond angles reached a 10° deviation from ideal, the acceptance rate dropped to 6.6%. Those rates may initially seem somewhat high, given the severity of the angular strain. However, there are always two bond angles changing during any move. At equilibrium, moves may transfer bond angle strain from one residue to the other, without increasing the total amount of strain in the system.

Figure 1-14. Backrub move acceptance ratios



Constraining both the bond angles and degree of rotation helps maintain relatively high acceptance ratios under many conditions. Monte Carlo acceptance statistics were gathered using a 10^6 step simulation of the Erbin PDZ domain (PDB 2H3L; Appleton 2006) using Amber bond angle parameters at kT = 0.6. **(A)** For every step, the maximum deviation of the N-Cα-C bond angle (α) from ideal (α_{ideal}, composite θ_0 from Table 1-2.) was determined prior to evaluating the acceptance criterion. (For a move about residues *i* and *j*, $\max(|\alpha_i - \alpha_{ideal}|, |\alpha_j - \alpha_{ideal}|)$.) The acceptance remains relatively high (6.6%), even when there is a 10° strain in one of the bond angles. **(B)** The acceptance ratio is highly dependent on both the magnitude of the rotational angular displacement (τ) and segment size. Two residue moves, corresponding to peptide plane rotations, are significantly more flexible than larger moves. **(C)** Simply limiting bond angles to within 10° from ideal and ignoring non-covalent forces, peptide bonds (size = 2) have significantly

greater rotational freedom than other segment sizes. τ interval lengths (the total length of $I_{bond\ angle} \cap [-90°, 90°]$) were calculated for all 2-12 residue segments of PDB 2H3L. **(D)** Limiting the extent of angular displacement for longer segments allows the acceptance ratio to remain reasonably high (>23% for backbone only moves, red), regardless of the segment length.

Table 1-2. Bond angle energy parameters used for simulations

	Amber				CHARMM			
	Non-Glycine		Glycine[1]		Non-Glycine[2]		Glycine	
	K_θ	θ_0	K_θ	θ_0	K_θ	θ_0	K_θ	θ_0
N-Cα-C	63	110.1	63	110.1	50	107.0	50	107.0
N-Cα-Cβ	80	109.7	50	109.5	70	113.5	48	108.0
N-Cα-Hα	50	109.5	50	109.5	48	108.0	48	108.0
C-Cα-Cβ	63	111.1	50	109.5	52	108.0	50	109.5
C-Cα-Hα	50	109.5	50	109.5	50	109.5	50	109.5
Cβ-Cα-Hα	50	109.5	35	109.5	35	111.0	36	115.0
N-Cα-C (composite)	74.7	109.7	72.3	110.0	58.7	107.0	57.4	106.9

[1] For glycine residues, Cβ refers to the position of the second Hα atom.
[2] The CHARMM force field has different bond angle parameters for proline. Because proline residue geometry is currently fixed during backrub sampling, the coefficients are not listed.

Bond angle energy parameters used for simulations were taken from the Amber and CHARMM force fields. K_θ is listed in kcal mol^{-1} $radian^{-2}$. θ_0 is listed in degrees. The total energy for a particular bond is determined using the standard formula $E = K_\theta(\theta - \theta_0)^2$. The composite N-Cα-C parameters were determined from least squares fitting of the total energies described in Table 1-3. The Amber bond angle potential has composite K_θ values 26-27% greater than the CHARMM potential. At an energy of kT = 0.6, the non-glycine composite force constants lead to bond angles of $\theta_0 \pm 5.1°$ and $\theta_0 \pm 5.8°$, respectively. That corresponds to a Metropolis acceptance rate of 37%.

Rotation Angle Constraints

Examining the acceptance statistics further, we made the intuitive observation that as the magnitude of angular displacement increases, the acceptance statistics drop almost

exponentially (Figure 1-14B). This phenomenon is best explained through sterics, where the larger the rotation, the more likely a deleterious steric clash is encountered. We therefore imposed an additional constraint upon moves that restricted the maximum angular rotation to a given threshold, τ_{max}. This was done by generating an additional interval, $I_{rotation\ angle} = [-\tau_{max}, \tau_{max}]$, and then calculating the intersection, $I = I_{bond\ angle} \cap I_{rotation\ angle}$, of that interval with the previously calculated intervals bracketing the bond angle. Importantly, this additional constraint can create an imbalance in the selection probabilities for the possible angles, as the total angular range of the intervals may be different before (l) and after (l') moves. (Figure 1-15) Because the probability of selecting a given angle is inversely proportional to the number of possible values, the following acceptance criterion can be used to produce uniform selection probabilities:

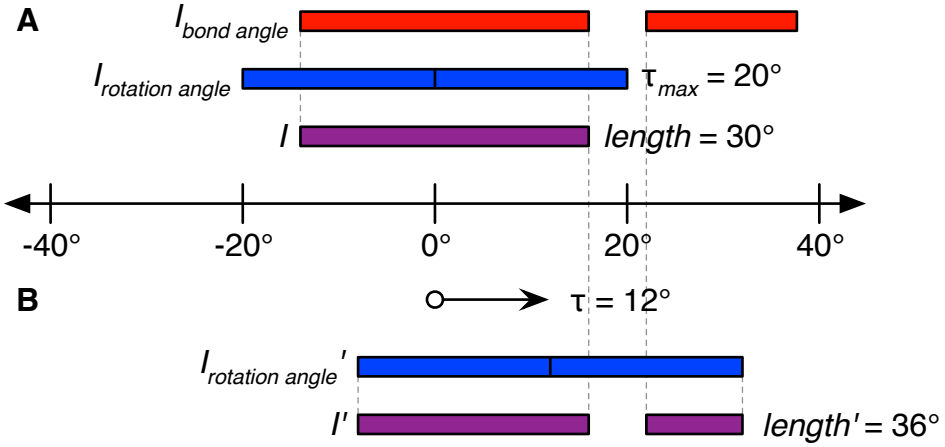
$$P(\tau) = \min\left[1, \frac{l}{l'}\right]$$

Trial move τ angles are generated using this procedure:

1. Calculate the total length, l, of the set of intervals, I.
2. Choose a random threshold, t, uniformly from the interval [0, 1].
3. Choose a random angle, τ, uniformly from I.
4. At angle τ, calculate the new rotational interval, $I_{rotation\ angle}' = [\tau - \tau_{max}, \tau + \tau_{max}]$.
5. Calculate $I' = I_{bond\ angle} \cap I_{rotation\ angle}'$ and the total length, l', of I'.
6. If $l/l' \geq t$, return τ. Otherwise go back to step 3.

Because l and l' are generally quite similar, this procedure rarely iterates more than several times and is considerably less costly than other parts of the simulation.

Figure 1-15. Example of angular constraints



(A) Backrub moves are constrained so that N-Cα-C bond angles remain within a given threshold, α_{max}, of the ideal value. A hypothetical set of constraining intervals ($I_{bond\ angle}$) is shown in red. The set is discontinuous because one of the pivot atom bond angles exceeds the permissible range for τ values between 16° and 22°. Backrub moves are also constrained so that the rotation cannot exceed a given magnitude, τ_{max}, using another interval ($I_{rotation\ angle}$) shown in blue. Moves are made by selecting a random rotation angle, τ, from the intersection of the two intervals (I) shown in purple. **(B)** An example τ angle of 12° is shown, along with the corresponding rotation angle constraints ($I_{rotation\ angle}'$) and overall constraint (I') that would be used on the next move. Because the length of I' is greater than I, the probability of selecting the reverse τ angle (-12°) on the next move is less than going in the forward direction. That results in nonuniform sampling. (Figure 1-12D) To equalize the probabilities, a proposed τ angle is therefore selected with a probability of *length/length'*, which is about 0.83 in this case.

To ameliorate the reduction in acceptance ratio for large segment sizes, the τ_{max} parameter is varied for each possible residue segment. This is distinguished from the Betancourt strategy of making equal magnitude displacements regardless of segment size. Different values of τ_{max} are stored in another sparse upper triangular matrix, T. Based on

empirical observation of the acceptance statistics, we devised the following rule relating τ_{max} to segment size, s:

$$\tau_{max} = \begin{cases} 40; & s = 2 \\ 23 - s; & s \geq 3 \end{cases}$$

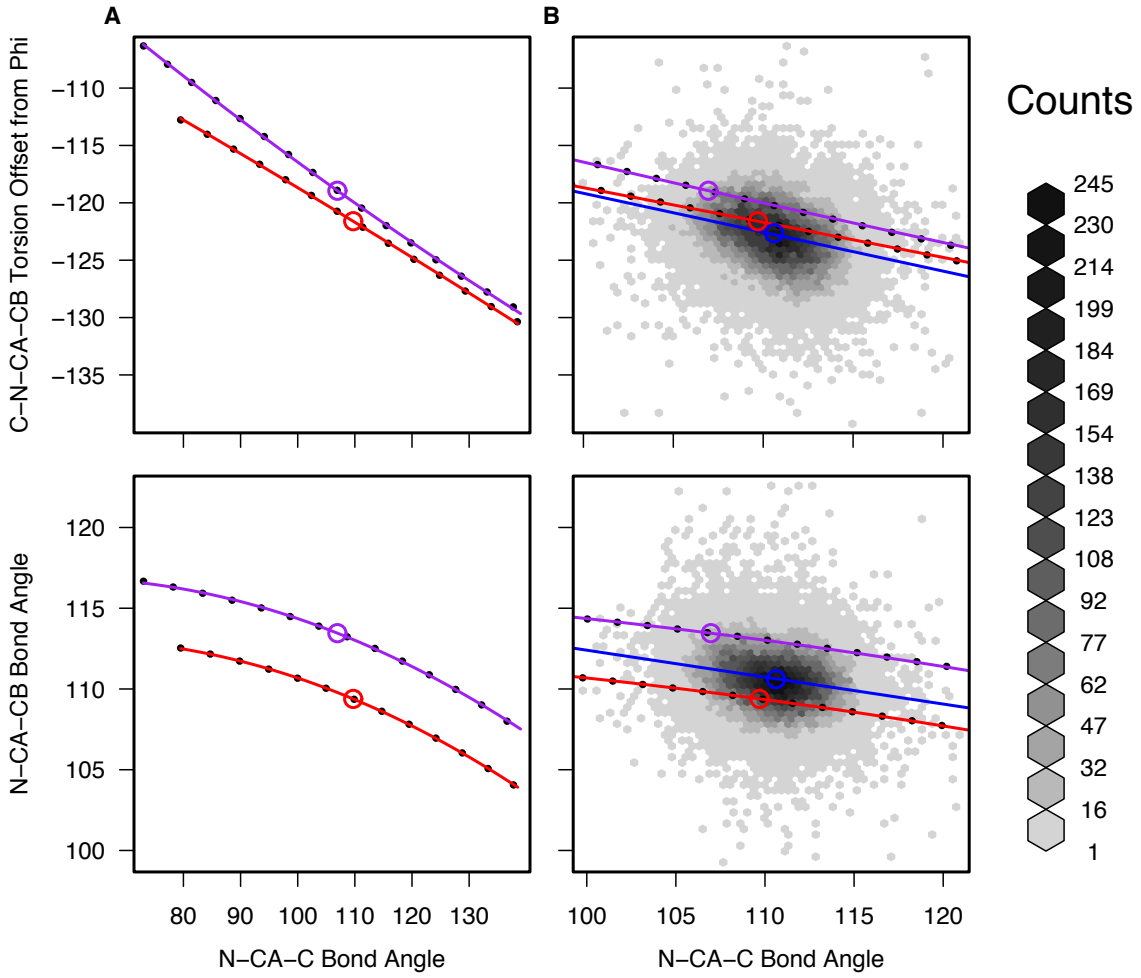
We found peptide bonds (size 2) to be significantly more flexible than other segment sizes. The large increase in flexibility is partially due to peptide bonds lacking the steric constraints of other segment sizes. However, when one looks at the distribution of allowable τ angles, given only a 10° bond angle cutoff, it is also clear that peptide bond segments have significantly more flexibility than larger segments (Figure 1-14C). In addition to Davis et al.[1], a similar type of motion has also been observed in unbiased computational simulations. A recent analysis of correlated ϕ/ψ motions in a large set of molecular dynamics trajectories also observed significant, localized peptide bond fluctuations[62]. Additionally, in pairs of structures of the same protein crystallized multiple times, larger "peptide flips" (involving rotations ~180°) are often observed[63].

As a result of constraining both the N-Cα-C bond angles and maximum angular displacement during a move, the acceptance statistics remain relatively high for segments sizes from 2 to 12 (Figure 1-14D). For the PDZ domain test simulation, backbone only moves showed an average acceptance ratio of 29%, and rotamer only moves showed an acceptance ratio of 34% (data not shown). When combined with the much less accepted simultaneous rotamer/backbone moves, the overall acceptance ratio drops to 26% (weighted mean of all move types). Elimination of simultaneous rotamer/backbone moves would increase the overall acceptance rate to 30%.

Optimized Placement of Cβ and Hα Atoms

An important methodological consideration in a procedure that modulates Cα backbone bond angles is how the positions of the branching Cβ and hydrogen atoms are simultaneously varied. Placement of branching atoms means positioning of Cβ and Hα relative to the N, Cα, and C atoms. A number of angular bisecting heuristics can be applied to place those atoms in positions with acceptable geometries. In this work, to reduce bond angle strain, the branching atoms are placed in positions at the minimum of the force field bond angle potential, given the current N-Cα-C bond angle. Minimization after every Monte Carlo move would be computationally expensive. Fortunately, the minimized internal coordinates of those atoms follow a predictable pattern (Figure 1-16). To enable fast updates of the position of a branching atom X, quadratic functions were fit that related a series of N-Cα-C backbone bond angles to the corresponding fully minimized branching atom internal coordinates, namely the C-N-Cα-X torsion offset from ϕ, and the N-Cα-X bond angle (Table 1-3). These fits were very accurate even to highly unfavorable bond angle energies of 20 kcal/mol.

Figure 1-16. Branching atom internal coordinate optimization



During sampling, branching atom internal coordinates are optimized for each backbone bond angle. **(A)** When the backbone (N-Cα-C) bond angle is varied and branching Cβ positions are minimized solely according to the Amber and CHARMM bond angle potentials, the internal coordinates show a very predictable pattern. (black dots) The minimized coordinates are shown for all backbone angles in which the total bond angle energy after minimization is less than 20 kcal/mol. (see Table 1-3) It is possible to precisely describe each internal coordinate dependency with a quadratic function. (Amber: red lines, CHARMM: purple lines) Colored circles show the overall minima for each force field. Comparable quadratic fits are observed for branching hydrogen positions (not shown). **(B)** In 199 high resolution (≤ 1.0 Å) crystal structures, Cβ atom positions

show a similar dependence on backbone bond angle. Internal coordinate counts for all non-glycine, non-proline residues are shown binned into hexagonal arrays[40]. Linear fits (R = -0.29 & -0.21, respectively) of the PDB coordinate pair data (blue lines) show qualitatively similar slopes to the fits derived from force fields (red and purple lines rescaled from A). This result is expected as crystallographic refinement makes use of a bond angle potential drawn from similar force fields. Blue circles indicate the median values observed in the PDB. Overall, positions determined using the Amber bond angle potential show closer agreement with PDB statistics.

Table 1-3. Quadratic coefficients for optimal placement of Cβ and Hα atoms

		Non-Glycine Residues[1]			Glycine Residues[2]		
		A	*B* (α)	*C* (α^2)	*A*	*B* (α)	*C* (α^2)
Amber	C-N-Cα-Cβ Offset	-1.5886	-0.2540	-0.0130	-1.3072	-0.5629	0.0790
	N-Cα-Cβ Angle	1.8826	0.1753	-0.0843	1.8962	0.2001	-0.1008
	C-N-Cα-Hα Offset	1.1958	0.7613	-0.1560	1.3072	0.5629	-0.0790
	N-Cα-Hα Angle	1.9467	0.1764	-0.1042	1.8962	0.2001	-0.1008
CHARMM	C-N-Cα-Cβ Offset	-1.2906	-0.4873	0.0357	-1.4440	-0.4213	0.0509
	N-Cα-Cβ Angle	1.9570	0.1645	-0.0815	1.8553	0.1969	-0.0974
	C-N-Cα-Hα Offset	1.2681	0.6386	-0.1165	1.4440	0.4213	-0.0509
	N-Cα-Hα Angle	1.9120	0.1652	-0.0966	1.8553	0.1969	-0.0974

[1] The CHARMM force field has different bond angle parameters for proline. Because proline residue geometry is currently fixed during backrub sampling, the coefficients are not listed.
[2] For glycine residues, Cβ refers to the position of the second Hα atom.

Quadratic coefficients were derived for optimal placement of Cβ and Hα atoms with spherical coordinates. The spherical coordinates used are the torsion angle offset from phi (C-N-Cα-C) to C-N-Cα-X and the N-Cα-X bond angle, where X is Cβ or Hα. (There are two sets of equivalent torsion offests and bond angles that can be used to uniquely place the Cβ and Hα atoms using spherical geometry. The difference between the two is whether the spherical coordinate axes are set up N-to-C or C-to-N. Using either set, the optimization procedure we describe will generate identical atomic positions.) Given a N-Cα-C backbone bond angle, α, a particular spherical coordinate, ω_i, can be calculated by applying the formula $\omega_i = A_i + B_i\alpha + C_i\alpha^2$. Coefficients are listed in radians^1, radians^0, and radians^{-1},

respectively. Note the symmetry in coefficients for glycine residues and the slight asymmetry in coefficients for non-glycine residues. The coefficients were produced as follows: N-Cα-C backbone geometries were generated with N-Cα-C bond angles every 0.5 degrees between 70 and 140 degrees. For each fixed backbone bond angle, Cβ and Hα atom positions were calculated by minimizing only the bond angle potential. Spherical coordinates were extracted for all geometries with total energies < 20 kcal/mol. (see Figure 1-16) Coefficients were determined by least squares fitting to the formula above.

After every backrub move, the new branching atom positions are found using those quadratic fits. Subsequently, the coordinates of the side chain prior to the move are rotated about the Cα atom pivot point such that the old Cβ atom is collinear with the new Cα-Cβ axis. Finally, the whole side chain is rotated slightly about the Cα-Cβ axis to restore the $\chi 1$ angle to its original value.

Simulation of 3-Residue Backrubs (Test 1)

Davis et al[1] identified 126 positions in 19 high resolution (≤ 1.0Å) crystal structures where there was evidence for a localized rotation of a 3-residue segment of the protein backbone. In some cases, the conformational variability in the backbone was only implied by alternate Cβ atom positions in the PDB file, with a single set of Cα atom coordinates representing the mean location of multiple Cα atom positions. In those cases, we used a single starting structure with the Cβ atom optimized according to the bond angle potential. If alternate Cα coordinates were present in the PDB file, we generated 2-3 starting structures, one for each variant letter (A, B, or C) in the contiguous set of atoms with alternate backbone coordinates. All other alternate atom coordinates were set to the

A variants. In total, this procedure yielded 161 starting structures in the 3-residue backrub set.

The simulations for each identified backrub were as follows. Given a three residue backbone motion centered on residue *i*, angular perturbations were enabled for residue pairs (*i*-1, *i*+1), (*i*-1, *i*), and (*i*, *i*+1). The side chain of residue *i* was also allowed to sample different rotameric states. 200,000 Monte Carlo steps were run at a temperature of 302 K. 50% of the steps consisted of a random perturbation of the (*i*-1, *i*+1) angle and a simultaneous rotamer swap. The other 50% of steps consisted of a random perturbation of either the (*i*-1, *i*) or (*i*, *i*+1) angle and no rotameric sampling.

Point Mutant Side Chain Prediction (Test 2)

We used a benchmark set of 2,141 pairs of protein structures for which the only difference was a single point mutation, aside from extra or missing residues at the N and C termini[46]. We removed 7 pairs from the set that had, at the mutated residue position, either missing side chain atoms or a non-canonical amino acid. We also removed 8 pairs for which the mutation was duplicated in another pair in the list. Finally we removed 103 pairs that had either missing or zero occupancy backbone atoms in the first structure in the pair. Structures with missing or zero occupancy backbone atoms in the second structure were removed during analysis (see below). That left 2,023 ordered pairs of structures.

During side-chain prediction, we sampled conformations (backbone and side-chain) for both the mutated residue and neighboring residues. Neighboring residues were selected that, prior to mutation, had any atom within a given radius of any atom in the mutated residue. Radial cutoffs of 4Å, 5Å, 6Å, 7Å, and 8Å were tested. At the beginning

of each sampling run, the side-chain in the first PDB structure was mutated and then rotamer optimized along with all the neighboring side-chains using an energy-table based Monte Carlo simulated annealing protocol[31]. Subsequently, the backrub protocol was run for 10^4 steps at a single temperature of kT = 0.6, maintaining either a fixed backbone ($P_{rotamer}$ = 1) or allowing backbone sampling ($P_{rotamer}$ = 0.25). The lowest energy structure found during ten separate executions was used as the prediction.

To facilitate comparison of the prediction with the second PDB structure in the pair, we first superimposed the N, Cα, and C atoms from a set of residues around the mutated residue. The superimposed set was defined as all residues satisfying the following condition in *both* the first and second PDB structures: a heavy atom of the residue must be within 4Å of a heavy atom in the mutated residue. All subsequent RMSD calculations used this fixed superimposition. To compare effects of the mutation on surrounding side-chains, we used a similar set of residues. The set was defined as all non-mutated residues satisfying the following condition in *either* the first or second PDB structures: a non-backbone heavy atom of the residue must be within 4Å of a non-backbone heavy atom in the mutated residue. Any RMSD calculation in which all compared atoms in the second PDB structure had zero occupancy was ignored in calculating overall statistics. All superimposition, RMSD, and chi angle calculations were done using ICM Browser 3.5-1l (Molsoft). Sequence alignments for mapping atom selections from structure to structure were created using ClustalW 1.83[64].

Rosetta Energy Function

Other than the addition of the bond angle term, we used the default Rosetta full-atom energy function and weight set, which is internally referred to as score12. It is very

similar to that described by Kuhlman[59], except as noted. The functional form is as follows:

$$E_{protein} = W_{rot}E_{rot} + W_{aa|phi,psi}E_{aa|phi,psi} + W_{rama}E_{rama} + W_{atr}E_{atr} + W_{rep}E_{rep} + W_{intra}E_{intra} + W_{solv}E_{solv} + W_{pair}E_{pair} + W_{bb\_hbond}E_{bb\_hbond} + W_{sc\_hbond}E_{sc\_hbond} + W_{sc\_bb\_hbond}E_{sc\_bb\_hbond} + W_{bond\_angle}E_{bond\_angle}$$

The weights used were as follows:

W_{rot}	0.56	W_{solv}	0.65
$W_{aa\|phi,psi}$	0.5	W_{pair}	0.49
W_{rama}	0.2	$W_{bb\ hbond}$	1.17
W_{atr}	0.8	$W_{sc\ hbond}$	1.1
W_{rep}	0.44	$W_{sc\ bb\ hbond}$	1.17
W_{intra}	0.004	$W_{bond\ angle}$	1

The E_{intra} term includes intra-residue van der Waals repulsive energies identical in form to the inter-residue E_{rep} term. The bond angle potential was given a weight of unity. The sequence was held fixed during all Monte Carlo simulations so the invariant residue type reference energies are not shown.

Code Availability

Source code for the implemented backrub model is available for download free-of-charge as part of the 2.2 release of the Rosetta molecular modeling software at http://www.rosettacommons.org/.

Chapter 2. Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains

Abstract

Protein-protein recognition, frequently mediated by members of large families of interaction domains, is one of the cornerstones of biological function. Here we present a computational, structure-based method to predict the sequence space of peptides recognized by PDZ domains, one of the largest families of recognition proteins. As a test set, we use the considerable amount of recent phage display data that describe the peptide recognition preferences for 169 naturally occurring and engineered PDZ domains. For both wild-type PDZ domains and single point mutants, we find that 70-80% of the most frequently observed amino acids by phage display are predicted within the top 5 ranked amino acids. Phage display frequently identified recognition preferences for amino acids different from those present in the original crystal structure. Notably, in about half of these cases, our algorithm correctly captures these preferences, indicating that it can predict mutations that increase binding affinity relative to the starting structure. We also find that we can computationally recapitulate specificity changes upon mutation, a key test for successful forward design of protein-protein interface specificity. Across all evaluated datasets, we find that incorporation backbone sampling improves accuracy substantially, irrespective of using a crystal or NMR structure as the starting conformation. Finally, we report successful prediction of several amino acid specificity changes from blind tests in the DREAM4 peptide recognition domain specificity prediction challenge. Because the foundational methods developed here are structure based, these results suggest they can be more generally applied to specificity prediction

and redesign of other protein-protein interfaces that have structural information but lack phage display data.

Keywords: PDZ domain; specificity prediction; tolerated sequence space; protein design; backrub backbone flexibility

Abbreviations

PWM: position weight matrix

ROC: receiver operator characteristic

AAD: average absolute difference

AUC: area under ROC curve

Frobenius: Frobenius (Euclidian) distance

Rank Top: predicted rank of the top amino acid

Introduction

For many proteins, the ability to recognize and bind to other proteins is one of the key determinants of function. For proper cellular behavior, these proteins must have sufficient interaction specificity, that is, discriminate between their true targets and a large number of competing proteins[14]. Determining which partners interact, through experiment or computational prediction, is critical for understanding the roles each protein plays. In addition to characterizing wild type protein interactions, knowing how mutations can affect specificity is similarly important, as perturbed protein-protein interactions are likely to contribute to genetic disorders[65]. Beyond characterizing naturally occurring interactions, the redesign of specificity has important applications,

including the study of cellular functions through perturbation of protein-protein interfaces, or creating proteins with new specificities for use in synthetic biology[66]. In the area of synthetic proteins, computational redesign has had recent success in improving the binding specificity of calmodulin for a single binding partner[67], generating a new pair of DNase-inhibitor proteins that bind to each other but not to the wild type precursors[8,9], and designing peptides that selectively bind to members of the bZIP protein family[68].

In addition to determining the set of sequences that accommodate a given protein-protein interaction, one can more generally consider the problem of predicting the set of sequences that can be tolerated by a protein fold and still maintain function, such as binding or catalysis. Several methods have recapitulated the sequences allowed by a protein family in its core[69,70]. Computational enumeration of sequences tolerating a protein fold has also been used to generate a library of new GFP molecules with altered fluorescent properties[71]. The work presented in this manuscript is motivated by the premise that more accurately predicting sequences that are compatible with a given structure will not only help better characterize similar proteins, but will also improve our ability to design proteins with new functions. Many of the foundational methods necessary for predicting protein-protein interactions, as described here, are relevant to other areas of computational protein design and characterization.

A number of computational methods have been developed specifically to enumerate some or all of the allowed sequences at a protein-protein interface. Wollacott and Desjarlais pioneered protein-peptide interaction prediction for several PDZ domains, several SH3 domains, mdm2, and EVH1[72]. Other studies have predicted protein-peptide interactions, including MHC[73] and SH3 domains[74,75]. Using methods developed for SH3

specificity prediction[75], a large-scale database of predicted protein-linear motif interactions has been released online[76]. Prediction of interaction specificities between several types of globular proteins (which are generally not mediated by linear motifs) has also been demonstrated, including Ras protein interactions[77] and human growth hormone and its receptor[16].

In the present study we focus on PDZ domains, which primarily bind to the C-terminal residue (termed position 0) of other proteins and the residues immediately upstream (positions -1, -2, -3, etc.). A large amount of PDZ-peptide interaction data have been accumulated, beginning with the foundational work of Songyang et al.[78], who synthesized an oriented peptide library and chemically sequenced peptide populations binding to a set of 9 PDZ domains. Their work identified two classes of PDZ domains, with class I binding to a S/T at peptide position -2 and class II binding to a mostly hydrophobic amino acid at the -2 position. Other groups have screened a smaller number of PDZ domains with either a discrete library of peptides using membrane-based synthesis[79] or a set of globular proteins using a yeast two hybrid assay[80]. One of the largest PDZ-peptide interaction studies to date used protein microarrays with subsequent confirmation by florescence anisotropy[81]. While the study examined a large number of PDZ domains (157), the number of genomic peptides used (217) was considerably smaller than the sequence space that can be explored by nature or computation.

To develop and assess our computational methods, we leveraged the considerable amount of PDZ-peptide phage display data that has been collected recently[14,15,82]. In contrast to other techniques, phage display library sizes can exceed 10^{10} peptides and thus sample a considerable fraction of sequence space [82]. Using purely structural information,

we can predict the specificity of wild-type PDZ domains at a majority of the peptide positions. Several studies have developed methods for docking PDZ peptides[83,84]. In contrast, we rely solely on PDZ-peptide complexes from the PDB and aim to predict the space of peptides tolerated by each domain. Different from previous studies, our method also models and captures specificity differences arising from PDZ point mutations[14]. For the more difficult case of PDZ domains with a greater number of mutations (at 4-10 residues)[15], prediction performance decreases but the algorithm is still able to recapitulate the loss and gain in recognition preference for amino acids at some positions. Our results show that incorporation of backbone flexibility into the modeling procedure significantly improves performance. Finally, to assess the predictive capacity of our method, we present results from a recent blind peptide recognition domain specificity prediction challenge.

Results

Human PDZ Prediction

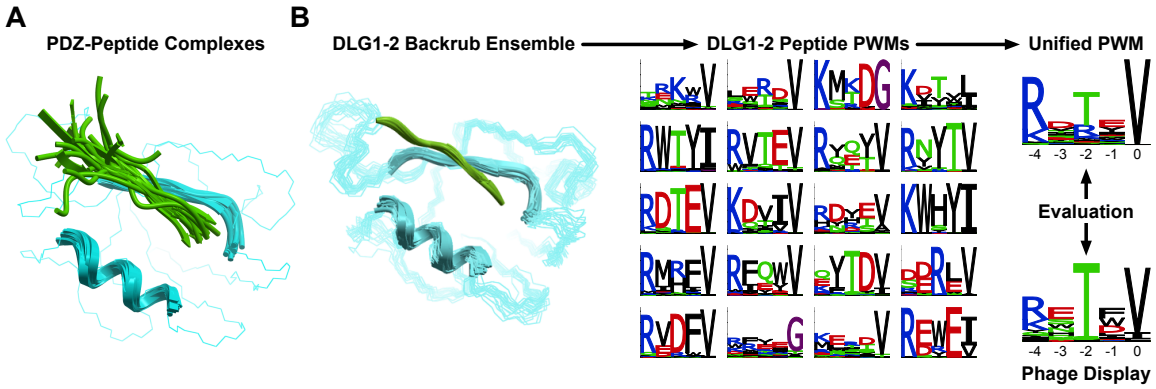
From the 54 human PDZ domains for which phage display specificity profiles had been published[14], we identified 17 that had available PDB structures with a bound peptide (Table 2-1). We truncated any peptide amino acids occurring before position -6 and after position 0. When superimposed, the helix and beta strand forming the peptide-binding site had very similar conformations across all 17 structures, with a mean pairwise Cα-Cα distance less than 0.9 Å for all residues (Figure 2-1A). There was slightly more backbone variation in positions 0 to -3 of the peptide (mean Cα distances of 0.9-1.4 Å), and increasing variation at positions -4, -5 and -6 (mean Cα distances of 2.1, 4.4, and 7.2 Å, respectively).

Table 2-1. Human PDZ domain structures used for PDZ profile prediction

PDZ Domain	PDB Code	Source	PDZ Chain	Peptide Chain	Peptide Residues
CASK-1	1KWA	X-Ray	A	B	568-574
DLG1-2	2I0L	X-Ray	A	C	2001-2006
DLG1-3	2I0I	X-Ray	A	D	2001-2006
DLG2-3	2HE2	X-Ray	A	B	511-517
DLG4-3	1TP5	X-Ray	A	B	420-425
DVL2-1	1L6O	X-Ray	A	D	2-8
ERBB2IP-1-hi	1N7T	NMR	A	B	301-307
MLLT4-1-hi	2AIN	NMR	A	B	99-104
MPDZ-7	2IWQ	X-Ray	A	A	1241-1247
MPDZ-10	2OPG	X-Ray	A	B	1714-1720
MPDZ-12	2IWP	X-Ray	B	A	1921-1927
MPDZ-13	2FNE	X-Ray	A	C	2042-2048
PDLIM4-1	TBD*	X-Ray	TBD*	TBD*	TBD*
PTPN13-2	1D5G	NMR	A	B	9-15
SLC9A3R2-2	2HE4	X-Ray	A	A	228-234
SNTA1-1	1QAV	X-Ray	A	B	1105-1111
TJP1-1	2H2B	X-Ray	A	A	114-120

Where necessary, the listed peptide residues were transformed onto the PDZ domain using crystallographic symmetry. *Structure from Dev Sidhu & co-workers (personal communication).

Figure 2-1. PDZ structures and computational prediction scheme



(A) When the 17 human PDZ domains structures are superimposed onto the PDB 1N7T alpha helix (H79-K87) and beta strand (L23-S28), both the peptide-binding site and the 4-5 C-terminal peptide residues take on very similar conformations. The peptide conformations are shown using a green cartoon representation. The superimposed parts of each PDZ structure are shown using a cyan cartoon representation. The entire 1N7T PDZ backbone trace is shown using cyan wires.

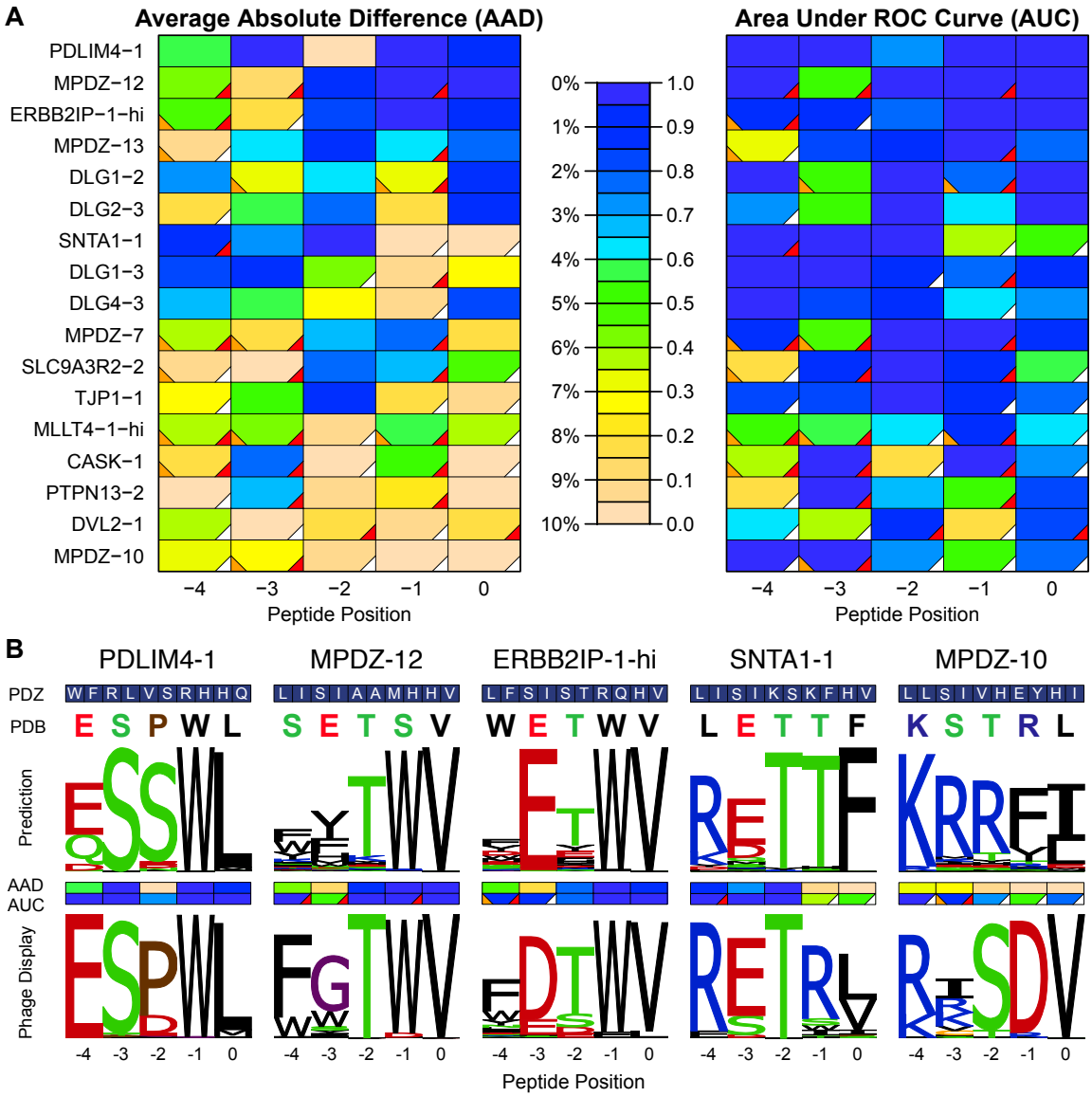
(B) The general scheme for profile prediction is shown using the DLG1-2 PDZ domain as an example. Each member of a backrub ensemble is used to generate a position weight matrix (PWM). The PWMs are combined into a unified PWM for the final prediction and evaluated by comparison with experimental data from phage display[14,15,82].

The computational strategy is summarized in Figure 2-1B. Similar to a previously developed protocol[16], we generated an ensemble of 20 backbone structures using independent Monte Carlo simulations consisting of backrub moves[85], which rotate short segments of the backbone about axes between Cα atoms, and side chain moves. Because peptide positions -5 and -6 were generally observed experimentally to be nonspecific[14], we predicted peptide profiles for only the last 5 positions. We used the design module[31] of the program Rosetta to generate and score approximately 10,000 sequences for each backbone structure. The intramolecular components of the score were downweighted to emphasize intermolecular interactions. For each backbone structure, a position weight matrix (PWM) was generated by Boltzmann weighting the score of each sequence with a given temperature. In the absence of a method for predicting the absolute binding affinity of the highest affinity peptide, we assumed that all PDZ domains had equivalent binding affinities and applied a single Boltzmann factor to all PDZ domains and backbones. For each PDZ domain, PWMs were merged by taking the median frequency (i.e. the 50th percentile of frequencies from all individual backbones) of each amino acid type at each peptide position and renormalizing each position to have a total frequency of 1. The algorithm used three free parameters, whose determination is discussed in Methods.

We evaluated the resulting predictions with several scoring metrics, of which two are graphically depicted in Figure 2-2A. The average absolute difference (AAD)

represents the average amino acid frequency error when predictions were compared with phage display. After visually comparing the predicted and experimental sequence logos at many positions, we defined the cutoff for a good prediction as being < 6%. At this level, one or more of the dominant amino acids will be shared between amino acid profiles, and there will be comparatively few false positives. 43 out of 85 positions displayed such good predictions. One of the PDZ domains, CASK-1, had its specificity previously predicted by Wollacott and Desjarlais[72]. For that domain, our prediction had a slightly better AAD than their prediction (Figure 2-3).

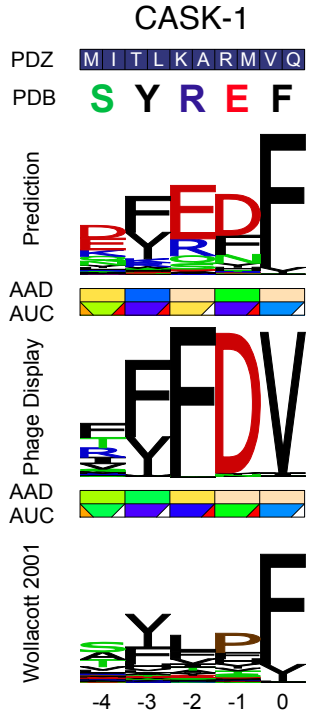
Figure 2-2. Human PDZ peptide profile prediction



(A) For 85 peptide positions corresponding to 17 human PDZ domains, the protocol produces predictions at 43 positions with average absolute difference (AAD) < 6%, and at 54 positions with area under ROC curve (AUC) > 0.75. The PDZ domains are sorted by overall AAD. 12 positions have flat experimental profiles with less than 2 bits of information (orange triangles), which can result in poor AAD scores. 48 positions show a mismatch between the starting PDB sequence and the most frequent amino acid selected in the phage display sequences (white and red triangles). 24 positions have an amino acid with > 10%

experimental frequency that is correctly predicted to be more highly represented than the amino acid in the starting PDB structure (red triangles). **(B)** Sequence logos are shown for five PDZ domain predictions. PDZ domain amino acids aligning with the positions mutated in the Erbin 10 mutation dataset are shown with blue boxes (1N7T positions 23, 25, 26, 27, 28, 48, 49, 51, 79, and 83). The peptide sequence from the PDB, used for generating the backrub ensembles, is shown below the PDZ sequence. Sequence logos were generated with LOLA (University of Toronto).

Figure 2-3. Comparison of CASK-1 prediction with Wollacott & Desjarlais 2001



The Wollacott PWM was generated by Boltzmann weighting the scores listed in Table 8 of their paper (J Mol Biol 313: 317-342). The Boltzmann temperature factor (1.4) was chosen to minimize the average absolute difference (AAD) between the predicted PWM and phage display, whereas the parameters for our prediction were optimized for all predictions. Both predictions used the same PDZ structure, 1KWA. The AAD for our prediction (7.0%) was slightly better than Wollacott (7.7%). Our prediction recapitulates an aspartate at the -1 position and

Wollacott does not. At the -2 position, Wollacott correctly predicts a phenylalanine as being the second most preferred and we do not predict it to have a significant frequency. See Figure 2-2 for a description of the figure elements. Another domain for which they made predictions, the third PDZ domain of PSD-95, has no published phage display data to date.

We also scored the predictions with a metric that depends solely on the relative ranking of amino acids. The area under the ROC curve (AUC) gives an indication of how well the prediction ranks the most populated amino acids, defined as those with phage display frequency ≥ 10%. Another advantage of using AUC is that there is a clearly defined random score of 0.5, unlike for AAD. For the human PDZ dataset 54 out of 85 positions had a good AUC score, defined as > 0.75, halfway between a perfect and random score. Many positions with low AAD scores had high AUC scores, indicating that while frequencies may have been mispredicted, often because of one or two strong false positives, the correct amino acids were still highly ranked.

The predicted and experimentally determined sequence logos for the binding preferences of several representative PDZ domains are shown in Figure 2-2B. In some cases, the prediction correctly captures the preference for a negative or positively charged amino acid, such as ERBB2IP-1 position -3 and MPDZ-10 position -4, but does not recapitulate the preferences for the particular amino acid (i.e. aspartate or glutamate). None of the scoring metrics took amino acid similarity into account, and the absolute frequency based metrics particularly penalized such mistakes. For the SNTA1-1 PDZ domain, the structure we used came from a complex with an internal motif in the nNOS protein. When the nNOS structure was truncated down to 7 amino acids, the C-terminus was not in the canonical conformation, making prediction of positions 0 and -1 much less

accurate. On the other hand, positions -2 to -4 were predicted very well. The prediction at position -4 was particularly notable because, while the crystal structure contained a leucine, the prediction correctly captured preference for arginine.

Having the prediction correctly capture specificity, despite a suboptimal peptide bound in the PDB structure, was often observed in the dataset. At 48 out of 85 of the peptide positions, the peptide amino acid residue from the PDB structure was not the most frequently observed in the phage display (Figure 2-2A). However, at 24 of those mismatched positions, the prediction successfully ranked at least one amino acid (experimentally observed to have ≥ 10% frequency) above the PDB amino acid, thus "overcoming" the starting structure used in the backrub simulations. The predictions were most able to overcome the PDB structure at the -1, -3, and -4 peptide positions. The predictions tended to be much more sensitive to the starting structure at the 0 and -2 positions. This reduced ability to overcome the PDB structure at the 0 and -2 positions corresponded to the overall PDZ domain ranking, where the best predictor of a poor overall score was the mismatch between the PDB sequence and phage display at those positions.

Erbin Single Point Mutant Prediction

In addition to the phage display dataset for wild-type PDZ domains, several datasets with synthetic variants of the Erbin PDZ domain have been published[14,15]. We used these datasets to test our peptide profile prediction method on a set of progressively more difficult modeling cases. The first dataset[14] consisted of 91 Erbin PDZ domain single point mutants at 10 different positions near the binding site. For each of the 10 sites, point mutants were constructed using amino acid residues observed at the same site

in other naturally occurring PDZ domains. With this dataset, we were able to test the prediction of functional changes as the result of mutation.

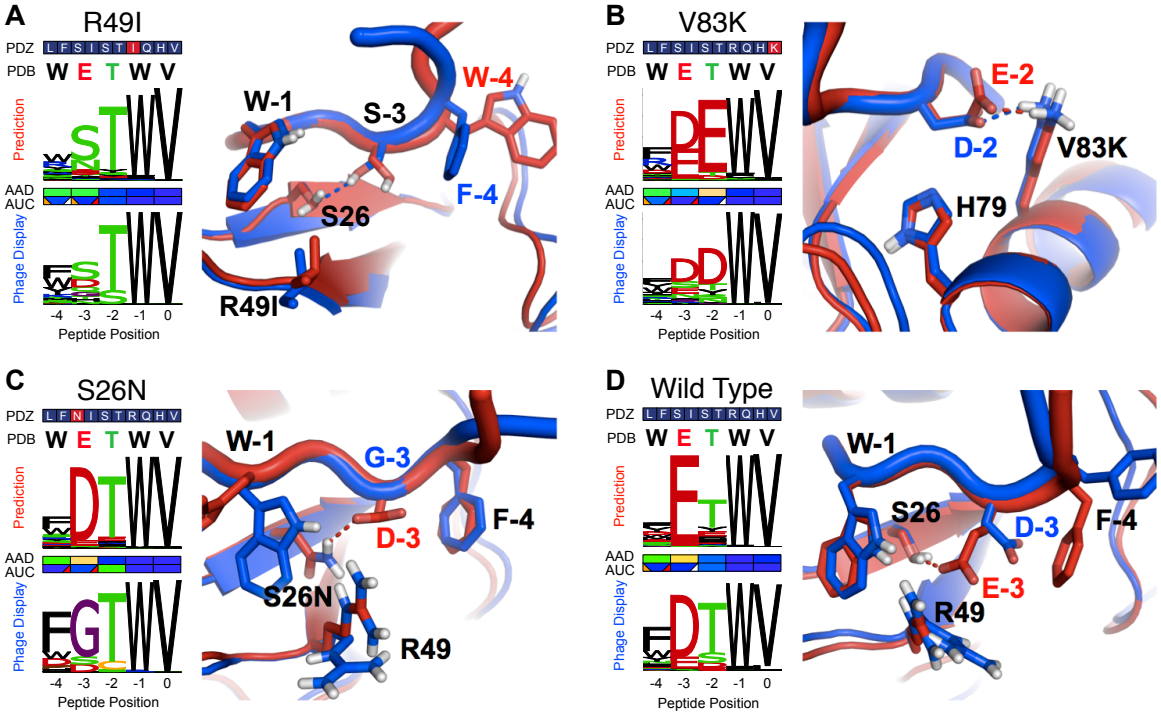
As shown in Table 2-2, prediction performance of Erbin point mutant phage display profiles was better overall than the wild-type human PDZ domains. This difference is likely because the wild-type Erbin phage display profile was already predicted better than most other human PDZ domains and less than half the point mutants had experimentally significant specificity differences from the wild-type[14]. In addition to good overall performance, the predictions were able to capture two of the key specificity changes in response to mutation, the loss of preference for an aspartate or glutamate at the -3 position and loss of serine or threonine at the -2 position. R49 mutants were among the most likely to lead to the D/E loss at the -3 position, an example of which is shown in Figure 2-4A. When compared across the whole dataset, the combined frequency of D/E was slightly overpredicted but still shows reasonable correspondence between phage display and prediction ($R^2 = 0.42$, Figure 2-5). Likewise, the predicted S/T frequency at the -2 position was also observed to correlate with phage display ($R^2 = 0.53$, Figure 2-5). In class I PDZ domains such as Erbin, a S/T preference results from an intermolecular hydrogen bond with the histidine at positions equivalent to position 79 on the PDZ domain[78]. The predictions here captured loss of S/T preference not only for H79 mutants, but also for the V83K mutation (Figure 2-4A) that makes a -2/K83 salt bridge more favorable than a -2/H79 hydrogen bond.

Table 2-2. Summary of performance on the 4 experimental datasets

		Bits of Information					
Dataset	Size	Phage Display	Predicted	AAD	AUC	Frobenius	Rank Top
Human PDZ	17	3.22	2.94	5.49%	0.79	0.59	5.1
Erbin Point Mutant	92	2.93	3.16	4.19%	0.90	0.43	2.5
Erbin 10 Mutation	61	3.22	1.98	6.43%	0.72	0.64	6.1

All data are averaged over positions 0 to -4 on the peptide. The Erbin Point Mutant dataset included one wild-type profile, so that the total number of PDZ domains evaluated is 169.

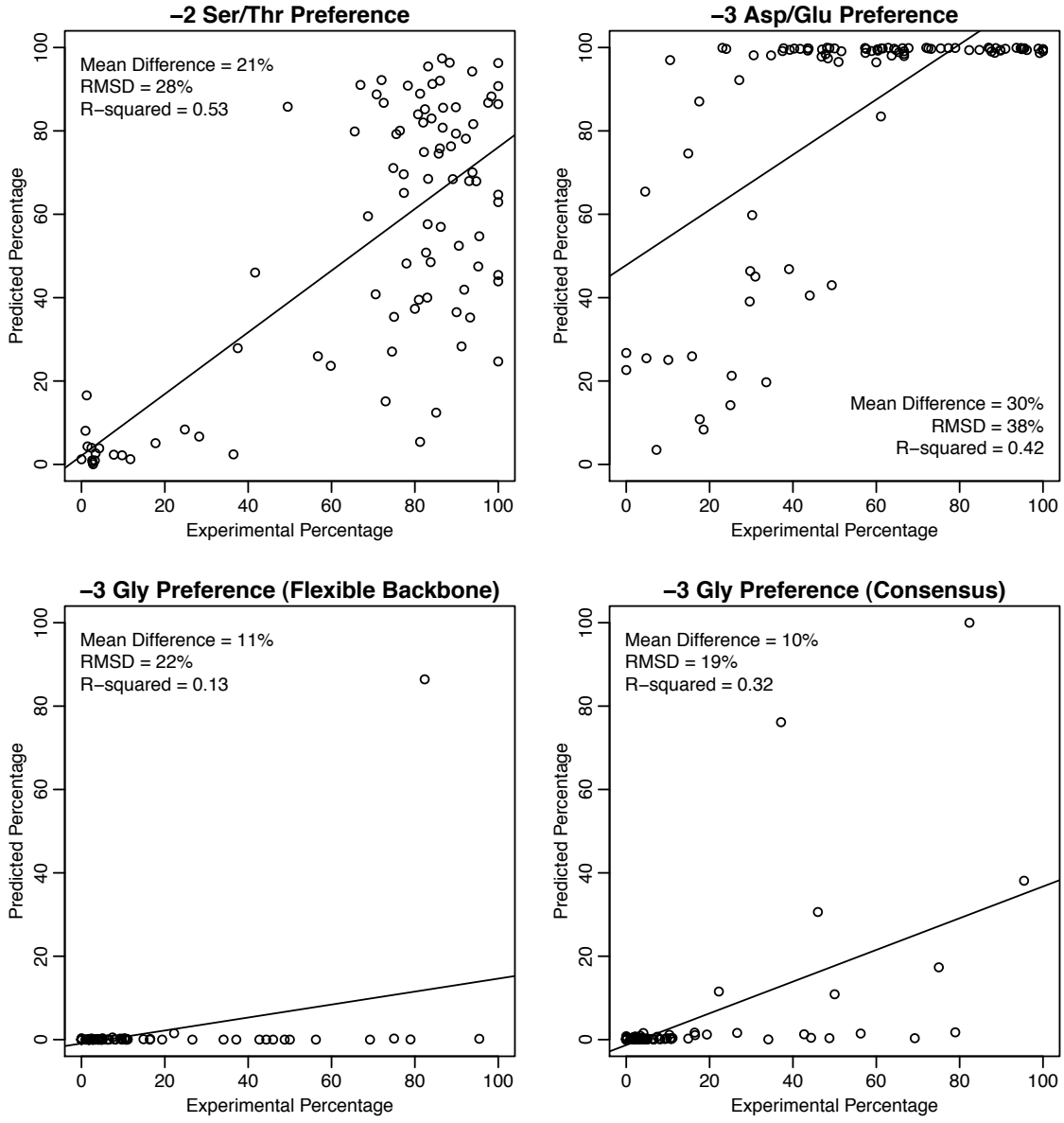
Figure 2-4. Predicted sequence logos and structures for Erbin point mutants



(A)-(C) Three examples from the Erbin Point Mutant dataset are shown, along with the wild type domain **(D)**. Both loss of -3 position aspartate/glutamate and -2 position threonine preferences are captured. The R49I mutation breaks an electrostatic interaction with the peptide -3 position **(A)**. In the case of V83K, a salt bridge interaction between the peptide -2 and K83 is captured, albeit with a glutamate rather than an aspartate **(B)**. Specificity changes to glycine are not well captured because they likely require backbone shifts **(C)**. By overall AAD, the predictions are ranked 1, 22, and 9, respectively (out of 92). See Figure 2-2 for

AAD/AUC color scale. The modeled structures for prediction (red) and phage display (blue) used the backbone for which the consensus peptide was closest in score to the best scoring peptide (i.e. the backbone that most preferred the consensus peptide relative to other peptides).

Figure 2-5. Changes in specificity predicted for the Erbin point mutant dataset



Predictions of the combined serine/threonine frequencies at position -2 correlate with phage display derived frequencies. Aspartate/glutamate frequencies at

position -3 also correlate, but are slightly over-predicted. Glycine frequencies at position -3 correlate more poorly using the WETWV starting sequence (left) than the phage display consensus sequences (right). Linear least squares regression lines are shown.

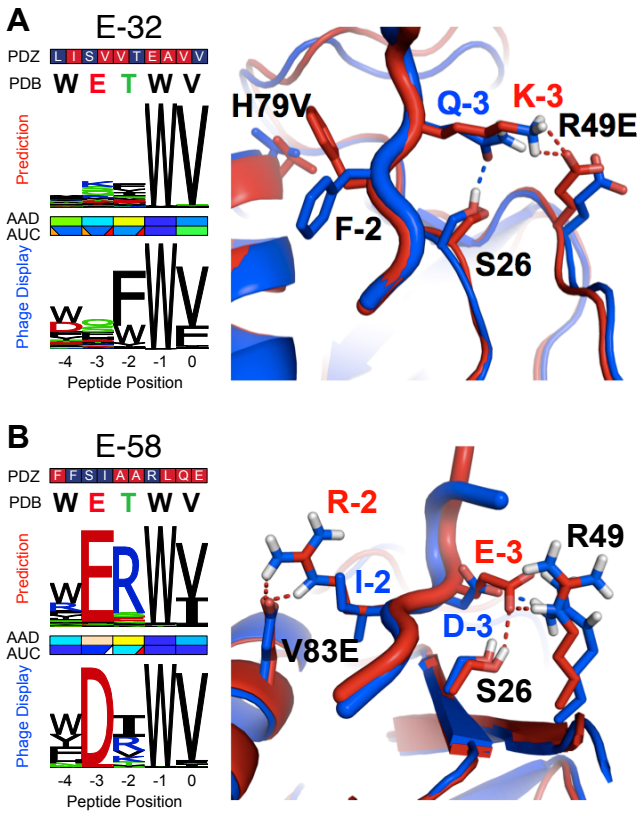
An emerging specificity that was consistently missed in the predictions was glycine at the -3 position, which was associated with mutations at the adjacent S26 and S28 residues. One such mutant, S26N, is shown in Figure 2-4. By phage display, the only S26/S28 mutations that did not switch to a dominant -3 glycine were S26T/S28A/S28G (small amino acids) and S26K/S28R (positively charged amino acids that slightly preferred an aspartate to glycine). The preference for glycine may come because the side chain from any other amino acid sterically clashes with the PDZ backbone or side chains, or from the peptide preferring an area of Ramachandran space not accessible to other amino acids. In either case, without the glycine present during backrub relaxation, the PDZ and peptide backbones will likely be biased away from conformations that prefer glycine, resulting in missed -3 preference for glycine. The one mutation for which we correctly predicted a -3 glycine preference was S26I, which is the largest beta branched amino acid and sterically favors glycine even after backbone relaxation.

Erbin 10 Mutation Prediction

To test our prediction method with a larger number of mutations, we used a recently published dataset[15] that contained 61 random combinations of mutations from the Erbin point mutant dataset. Each synthetic Erbin PDZ domain had 4-10 mutations (with a distribution of 1, 2, 10, 18, 19, 7, and 4 counts, respectively). As shown in Table 2-2, predicting peptide profiles with this number of mutations was more difficult than for either the wild-type or single point mutants. Several good predictions are shown in Figure

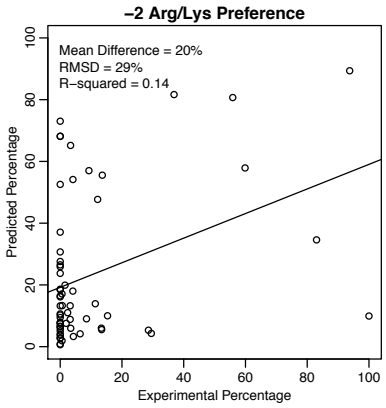
2-6. The best prediction, mutant E-32, captured simultaneous loss of the -3 D/E and -2 S/T preferences. It combined both the R49E and H79V mutations, along with four others. In the Erbin 10 mutation dataset, the new amino acid preference we were able to predict most accurately was -2 R/K. Across the whole dataset, there was slight correspondence in combined -2 R/K frequencies, with the largest errors coming from false positive predictions of a -2 R/K preference.

Figure 2-6. Predicted sequence logos and structures for Erbin 10 mutation domains



Two examples from the Erbin 10 Mutation dataset are shown (identifiers E-32 and E-58[15]). Prediction of the simultaneous loss of -3 aspartate/glutamate and -2 threonine specificities is evident for the E-32 mutant **(A)**. Mutation V83E results in capturing slight preference for an E83/R-3 salt bridge **(B)**. The predictions ranked 1 and 7, respectively (out of 61).

Figure 2-7. Changes in specificity predicted for the Erbin 10 mutation dataset



The general trend best predicted in the Erbin 10 Mutation dataset was the presence of arginine/lysine at the -2 position. Prediction of a F/G/I/V/Y amino acid at -2 or aspartate/glutamate at -3 showed only a very slight upward trend. Prediction of glycine at -2 or -3 showed essentially no positive correlation (data not shown).

The Erbin 10 mutation dataset represents a difficult test case for prediction of specificities, given the number of mutations involved so close to the binding site. Even for point mutations, peptide specificity changes have been reported distant from the site of mutation, which were attributed to ligand orientation effects[14]. While the success we saw here was encouraging, we did not pursue further homology modeling tests in which more residues were changed, inserted, or deleted. Such changes would likely result in further changes to the PDZ domain structure, further decreasing prediction accuracy.

Comparison of Different Protocols

Using the human PDZ, Erbin point mutant, and Erbin 10 mutation datasets, we analyzed whether backbone flexibility improved performance. To do so, we repeated the predictions with a fixed backbone, using either the 1N7T PDB NMR ensemble for Erbin and its variants or a single PDB structure (X-ray structure or model 1 from an NMR

ensemble) for all other human PDZ predictions. By any metric, fixed backbone performance was the same or worse for all three datasets. The most significant change was observed in overall AAD, which increased from 5.37% to 6.69% when the backbone was fixed (Table 2-3).

Table 2-3. Changes in performance with changes to the prediction method

		Bits of Information					
Dataset	Size	Phage Display	Predicted	AAD	AUC	Frobenius	Rank Top
Fixed Backbone	170	3.12	2.09	6.69%	0.78	0.68	5.1
Flexible Backbone	170	3.12	2.69	5.37%	0.80	0.55	4.6
Consensus	170	3.12	2.73	5.02%	0.83	0.52	3.7
Frobenius Optimized	170	3.12	1.89	5.56%	0.80	0.54	4.6

Overall performance is given as the average of the Human PDZ, Erbin Point Mutation, and Erbin 10 Mutation datasets. The flexible backbone method represents the primary method described in this manuscript (i.e. an average of the first three rows of Table 2-2). The fixed backbone method excluded backbone moves from the Monte Carlo simulations. For the consensus variant, the peptide was mutated to the phage display derived consensus sequence prior to generation of the backrub ensemble. The Frobenius optimized method used the average Frobenius distance for parameter optimization instead of the AAD.

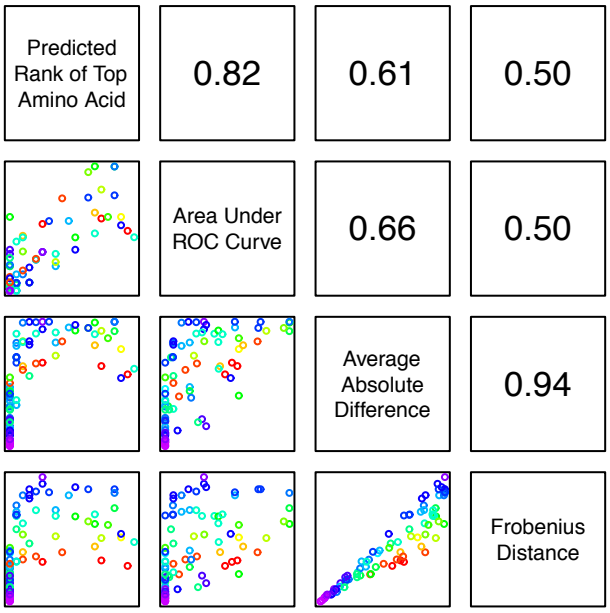
We also analyzed how the predictions were affected by the starting peptide sequence used during backbone ensemble generation. Using phage display data, we determined the most frequent amino acid for all positions. Before starting backrub simulations, we mutated each peptide to use this "consensus" sequence instead of the sequence from the PDB structure. We found that starting from these consensus sequences improved overall performance across the three datasets (Table 2-3). The protocol also improved prediction of glycine at the -3 position in the Erbin point mutant dataset (R^2 = 0.13 for standard flexible backbone vs. R^2 = 0.32 for consensus prediction). While overall performance increased across the datasets, two metrics became slightly worse for the

human PDZ dataset, the AAD (increasing 0.15%) and Frobenius distance (increasing 0.02, see below and Methods). This small increase may be because a somewhat suboptimal sequence can do better job recapitulating backbone flexibility for the human PDZ domains. Another contributing factor may be the protocol used for mutating the PDB peptides to the consensus sequence, which replaces all PDZ and peptide side chains with rotamer optimized, minimized conformations. The resulting side chains may be more poorly packed than the starting PDB conformations.

Comparison of Scoring Metrics

To evaluate the predictions, we used several metrics for scoring profile prediction performance: the area under ROC cure (AUC, see Methods), the predicted rank of the top amino acid (Rank Top), the average absolute difference (AAD), and Frobenius distance. A comparison of the correlations between those four scoring metrics on the human PDZ dataset is shown in Figure 2-8. The AAD and Frobenius distance metrics were strongly correlated, which is expected given the close mathematical relationship between the two (see Methods). Likewise, the Rank Top and AUC metrics were also closely correlated, which stems from each metric similarly scoring the enrichment of at least one of the most frequent amino acids. A comparison between all four metrics reveals that while most positions with good AAD or Frobenius distance scores also receive nearly perfect Rank Top and AUC scores, there is much less correlation for those positions with relatively poor AAD or Frobenius distance scores. From the perspective of method optimization, application of the Frobenius distance was found to result in a substantial reduction in the predicted number of bits (see Table 2-3).

Figure 2-8. Profile evaluation metric correlation



Scatter plots between pairs of metrics are shown in the lower diagonal. Correlation coefficients are shown in the upper diagonal. The mean number of bits in the experimental and predicted profiles is shown with a rainbow color scheme, with red indicating the least number of bits and purple indicating the greatest number of bits. For consistency, area under ROC curve (AUC) values are shown as 1-AUC so that lower scores represent better predictions for all metrics.

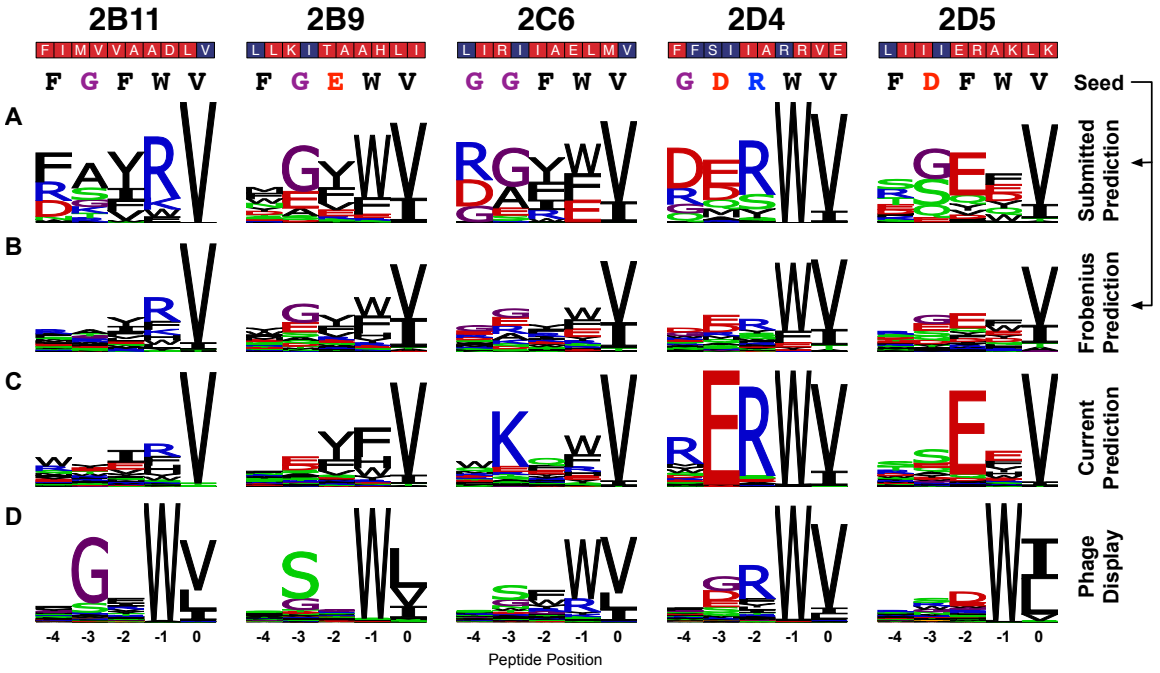
DREAM4 Specificity Prediction Challenge

While developing the current method, we participated in the PDZ specificity prediction part of the DREAM4 challenge. We did so using an earlier iteration of the method described here, with several slight differences. The PDZ section involved blind prediction of 5 Erbin PDZ domain variants that had 6-9 mutations each at the same positions as the Erbin 10 mutation dataset[15]. Because the Erbin 10 mutation dataset was available before the competition deadline, we began by examining the specificities of synthetic PDZ domains with high local sequence identity around each of the individual

peptide positions. Using that information, we made predictions by hand of the amino acid that would be most highly represented at each peptide position, and mutated the peptide to those sequences prior to backrub ensemble generation. After running our prediction method, we noted that several of the computational predictions did not show a high tryptophan frequency at the -1 position (2B11, 2D5, and to a lesser extent 2C6). This contrasted with the Erbin 10 mutation dataset, which preferred tryptophan at the -1 position for all similar PDZ domains. Despite this deficiency, we elected to submit the PWMs unmodified so that our prediction solely reflected the computational results.

The two blind predictions where we performed best, 2C6 and 2D4, are shown in Figure 2-9, along with the three other blind predictions. In both cases, the last four positions (-3 to 0) show good correspondence to experiment. At those positions, at least 2 out of the top 4 preferred amino acids are shared between the predictions and phage display results. At the -4 position, we over predicted the preference for arginine, aspartate and glycine.

Figure 2-9. DREAM4 prediction results for 2 synthetic Erbin variants



Wild type and mutant PDZ domain residues are shown at the top in blue and red, respectively. The last five peptide positions are shown as sequence logos. **(A)** Submitted predictions captured shifts in specificity at the -3 and -2 positions, as well as a slight preference for isoleucine at the 0 position. **(B)** Reoptimization of parameters using the Frobenius distance p-value scoring metric employed in DREAM4 evaluations improves the predictions primarily by flattening out specificity profiles. **(C)** Using the current prediction method, with the unbiased starting peptide sequence of WETWV, similar shifts in specificity at the -3 and -2 positions are seen with the exception of the 2C6 -3 lysine false positive. A slight preference for -1 arginine is also observed. **(D)** Phage display determined specificity profiles released after submission of **(A)**. DREAM4 phage display profiles were never used as training data for any of the predictions shown in A-C.

To determine how several factors affected our performance, we generated altered predictions and scored them using the same method as the competition organizers as well as the other scoring metrics introduced above. The Frobenius distance score used in the

competition was different from the average absolute difference score we used to optimize the parameters on the Erbin 10 mutation training data. Reoptimization of parameters on the training data (excluding the DREAM4 set) using the Frobenius distance improves our score significantly. While much of the improvement comes from flattening out the profiles, there are small increases in the AUC and top rank scores, indicating improvement in the relative frequencies of amino acids. The AAD improves as well, indicating some degree of mismatch between the Erbin 10 mutation training data and the DREAM4 synthetic PDZ domains.

Our current prediction method, which was not seeded with sequences derived from manual analysis, performs more poorly than the prediction we submitted. Whereas the method used for our submitted predictions was optimized using experimental data very similar to what it was predicting, the current prediction method uses parameters simultaneously optimized for wild type structures, point mutants, and larger numbers of mutants. The current method does capture several important changes in specificity from wild-type Erbin. At the -1 position of 2C6, it correctly predicts the emergence of a minor population of arginine. At the -2 position of 2C6, it predicts emergence of phenylalanine and tyrosine. At the -2 position of 2D4, it correctly identifies a switch in preference from threonine to arginine.

Because Frobenius distance scoring appears to reward profiles that are flat and have less information content, we tested how the scoring metric would behave using a relatively naïve PWM with W and V at the last two positions, and a flat profiles at all other positions. Using that PWM for all DREAM4 PDZ proteins yields a score on par with the best prediction submitted to the challenge (Table 2-4). Hence, using the

Frobenius scoring makes it difficult to determine whether shifts in specificity are correctly captured.

Table 2-4. DREAM4 blind prediction challenge performance

Prediction	2B11	2B9	2C6	2D4	2D5	Score	AAD	AUC	Top Rank	Bits
Best Team*	0.90	1.06	0.83	0.98	1.21	47.6				
xxxWV	0.96	1.16	0.76	0.74	1.22	44.5	3.78%	0.63	13.9	1.73
Frobenius	1.38	1.27	0.85	0.63	1.33	27.0	4.96%	0.81	4.9	1.66
Submitted*	1.63	1.28	1.07	0.79	1.57	21.5	5.64%	0.80	5.8	2.73
Current	1.50	1.55	1.08	1.11	1.81	14.2	5.88%	0.73	6.6	2.28

Only starred predictions were submitted to the competition. The "Frobenius" prediction was based on reparameterization on separate training data using the Frobenius score that was used in evaluating the predictions submitted to the DREAM4 challenge. xxxWV was a naïve prediction where all positions in all domains were set to be flat, and the last two positions were set to WV. The Current prediction was based on global parameterization and was not seeded with manual predictions. The Frobenius distance between prediction and experiment is shown for each domain. The score was calculated using the average of the $-\log_{10}$(p-value) of each Frobenius distance. AAD is the average absolute difference. AUC is the average area under the ROC curves. Top Rank is the average predicted rank of the most frequent amino acid from phage display. Bits is the average information content of each prediction, with the value for experiment being 1.97 bits. AAD, AUC, Top Rank, and Bits are calculated only for the last 5 peptide positions.

Relative Biases of Average Absolute Difference and Frobenius Distance

Though highly correlated, there were subtle differences between the AAD and Frobenius distance metrics. Relative to the maximum value of each metric, the Frobenius distance scoring metric was less than or equal to the AAD for all peptide positions. As is shown in Figure 2-8, the degree to which the Frobenius distance was less than the AAD is directly related to the average flatness of the experimental and predicted profiles. The

amount the points fell below the diagonal can be quantified using the ratio between the metrics, which has a Pearson correlation coefficient of 0.84 with the average bits of information. This relationship arises because the shape of either the experimental or predicted amino acid profile sets an upper bound on the Frobenius distance. This is illustrated in Table 2-5. Comparison of AAD and Frobenius distance with hypothetical profiles with a series of hypothetical profiles consisting of one dominant amino acid and increasing levels of background. When compared to the most divergent possible profile, the maximum Frobenius distance drops much more quickly as a function of the bits of information than the AAD. Mathematically, this arises because the Frobenius distance uses the error squared, which strongly penalizes more prominent false positives. From the perspective of method optimization, application of the Frobenius distance can lead to substantial reduction in the predicted number of bits. When used instead of AAD for parameter optimization, the average number of bits shown in Table 2-3 drops from 2.69 to 1.89, substantially below the 3.12 bits present in the phage display profiles.

Table 2-5. Comparison of AAD and Frobenius distance with hypothetical profiles

$p(AA_1)$	$p(AA_2)…p(AA_{20})$	Bits	AAD	Normalized AAD	Frobenius	Normalized Frobenius
1.00	0.00	4.32	10.0%	1.00	1.41	1.00
0.81	0.01	2.81	9.9%	0.99	1.28	0.90
0.62	0.02	1.75	9.8%	0.98	1.16	0.82
0.43	0.03	0.91	9.7%	0.97	1.07	0.76
0.24	0.04	0.30	9.6%	0.96	1.00	0.71
0.05	0.05	0.00	9.5%	0.95	0.97	0.69

The average absolute difference (AAD) is much less constrained by the bits of information content than the Frobenius distance. The hypothetical profiles given in the table are compared to a profile with $p(AA_1)…p(AA_{19}) = 0$ and $p(AA_{20}) = 1$, which gives the maximum possible metric value. The AAD and Frobenius distances shown are normalized by their respective maxima of 10% and $2^{1/2}$.

Discussion

In this study, we have undertaken structure-based prediction of the peptide sequence space recognized by a large set of 169 wild-type and mutant PDZ domains. By visual inspection and evaluation based on multiple scoring metrics, we found that peptide profiles could be successfully predicted at a majority of the positions for wild-type PDZ domains. Incorporation of backbone flexibility was a key aspect of our approach leading to significantly improved performance. A particularly novel aspect of the work presented here is that we capture peptide specificity shifts upon point mutation, which has not been previously reported on a dataset of this size. We observed that though the algorithm can in many cases find better PDZ-peptide interactions than those present in the PDB structure, there is still clear room for improvement, particularly when predicting changes in peptide specificity from large to small amino acids. In the course of evaluating our predictions, we also identified strengths and weaknesses of several different metrics for comparing PWMs. Performance should thus be evaluated using several criteria, where the optimal criterion depends on the intended application, as discussed below. Finally, in a blind test, we were able to predict the gain and loss of preference for several amino acids.

As noted above, the overall predictive performance on wild type and point mutant PDZ domains was quite good, but a fraction of the positions were not correctly predicted. While many of these mispredictions are likely dominated by insufficient sampling or scoring errors, there are also several caveats in using phage display data as the reference. First, while the typical 3-5 rounds of iterative panning and propagation enriches the phage population with high affinity binders, the biases this may introduce into relative

peptide populations are not well understood nor explicitly modeled by the computational predictions. Second, the number of peptides sequenced can limit precision. For the PDZ domains used in this study, 8-193 peptides were sequenced per experiment, with a median of 46 peptides sequenced. Third, for flat profiles, where there are many represented amino acids with small populations, uncertainty in both the absolute experimental frequencies and the computational score function increases, which can lead to high AAD scores. Of the 12 flat human PDZ positions, each with less than 2 bits of information, only 3 had good AAD scores. Nevertheless, we believe phage display remains the best experimental method for comparison with extensive computational sequence sampling.

An advantage of phage display and the computational techniques used here is that they are both in principle capable of detecting covariation in amino acid preferences between positions. Any such covariation will be ignored in the PWM representation used for our analysis. In the phage display data used here, we found few cases where there was evidence of covariation, and the number of peptide sequences reduced the potential significance. This is to be expected given the extended nature of the peptide conformation in the PDZ domain binding site minimizing interactions between peptide positions. While the method presented here is technically capable of capturing correlations between positions, it needs to be evaluated on other systems where covariation is more prevalent.

In the development of protein design algorithms, one of the key metrics used to measure performance is the ability to predict the amino acid originally found in the PDB structure (i.e. native sequence recovery). However, for many of the PDZ structures we examined, the peptide sequence in the structure was not optimal for binding. In those

cases, the more difficult task is to identify those sequences that bind more strongly than the starting sequence. Our results demonstrate the ability to "overcome" the input sequence in half of the cases where phage display showed a stronger preference for amino acids not in the PDB structure. This strongly supports the use of computational design algorithms, like the one presented here, in improving binding affinity between partners of known structure, especially in cases where a technique like phage display is not applicable to a given design scenario, or is not available to a researcher.

One of the important findings in this study is that incorporation of backbone flexibility using the backrub model[85] significantly increases accuracy, especially the ability to predict amino acid frequencies. For both computational protein design and structure based drug design, high-resolution crystal structures are usually preferred. For the human PDZ predictions, almost all of which used crystal structures, using backrub ensembles based on those structures outperformed the original structures themselves. Using NMR ensembles is another means of incorporating backbone flexibility into computational modeling. However, using the Erbin PDZ domain NMR structure as input, backbone flexibility modeled with backrub moves again significantly improved the prediction performance.

Several results from this study highlight the relationship between modeling backbone flexibility and predicting preferences for amino acids not present in the input structure. While important specificity changes were already captured using the default flexible backbone method, using the phage display consensus sequence during backbone generation markedly improved the predicted mutant PDZ profiles. Glycines were notably under-predicted in the Erbin point mutant dataset, but prediction was improved by

seeding glycine into those structures where it was expected. One explanation for the observed improvement is that the ensemble generation procedure optimizes the PDZ domain structure to accept whatever amino acid is input. An ideal method would search both sequence and structure space simultaneously and find the combinations of both that lead to the best binding interaction.

One way to accomplish this simultaneous optimization is through iteration between backbone relaxation and sequence optimization. Such methods have been applied to *de novo* design of new protein folds[59] and loops[86]. However, in those cases the design algorithm only needed to converge on a small number of foldable sequences to be successful. When trying to predict the sequence space of binding peptides, one must evaluate a much larger sequence space. If cysteine is excluded, the combined sequence space for five amino acids is nearly 2.5 million, which presents a challenge when paired with sampling backbone conformational space.

In addition to demonstrating the utility of a model of backbone flexibility in amino acid profile prediction, we also tested and compared a number of different scoring metrics for evaluating performance. Which metric is most applicable depends on the desired application for the computational predictions. For library construction, where one wishes to determine a small set of likely amino acids for combinatorial screening, the area under ROC curve (AUC) is likely the best discriminator. For protein design aimed at testing of individual designs, where one wants to maximize the probability of finding the best amino acid, the predicted rank of the top amino acid (Rank Top) may be a more sensible scoring metric. For binding partner prediction methods that depend on having an

accurate PWM, metrics that directly compare amino acid frequencies, such as the average absolute difference (AAD) or Frobenius distance, are good candidates.

While the Frobenius (or Euclidian) distance, and related metrics like RMSD, are often used for clustering and structural comparison, we found that its use in evaluating and optimizing specificity prediction can have several drawbacks. First, it is biased towards rewarding flat profiles more than specific profiles. When used for algorithm optimization, using a Frobenius distance metric can artificially flatten all profiles, sometimes with comparatively small improvements in performance (Table 2-3). Second, there is no well-defined random model for Frobenius distance. In this study, we found a combination of distance-based metrics, rank-based metrics, and frequency correlations for a subset of important residues the best gauge of overall performance.

One of the best ways to experimentally validate a computational technique is through controlled, blind prediction. In the DREAM4 challenge, we were able to predict both loss of specificity at defined peptide positions as well as a number of new amino acid preferences. Using the Erbin 10 mutation dataset as a guide, we determined that new preference for R/K at the -2 position could be best captured using the method. In the DREAM4 challenge, this observation was borne out, as one of our best performing predictions also contained this specificity. The DREAM challenge also exposed a third drawback related to the application of the Frobenius distance, namely that it alone does not indicate if a given set of predicted amino acid frequencies correlates with the observed frequencies. This was evidenced by the artificial xxxWV prediction, which scored almost as well as the best submitted prediction. To discern whether significant

differences in specificity between PDZ domains are reproduced, direct correlation analysis of the relevant amino acid frequencies is likely much better suited.

Many applications of computational protein design involve making individual mutations, or a library to be screened. Success is measured based on whether any of the limited number of designs were successful. In this study, we have compared computational protein design with experiments that consider nearly all possible binding interactions and can better evaluate the computational effectiveness. The protocol described here predicts a large fraction of the preferred amino acids (> 10% frequency in the phage display sequences) within the top 5 ranked amino acids, both for structures where one of the partners stays fixed (70%) or a single mutation is made (80%). By incorporating backbone flexibility, the predictive power is improved over fixed backbone design, irrespective of whether a crystal or NMR structure is used as a starting point. Applying this method should increase the success rate of computational second-site suppressor designs[8], where a destabilizing mutation on one side of an interface is compensated for by mutations on another side of the interface. Finally, while a significant amount of high throughput binding data has been accumulated for PDZ and other domains, there remain a significant number of other protein binding modules with structures for which exhaustive binding data is not available. This method enables predictive assessment of the amino acid preferences of those domains.

Methods

Structure Preparation

17 PDZ-peptide complex structures (Table 2-1) were used for structure based profile prediction. Bound peptides were N-terminally truncated to at most 7 amino acids.

To prepare the syntrophin PDZ domain structure (1QAV), which is bound to an internal motif on the nNOS PDZ domain, nNOS was C-terminally truncated after the residue that occupies the canonical peptide 0 position. Where necessary, ICM Browser Pro (Molsoft, La Jolla, CA) was used to generate the PDZ domain-peptide complex using crystallographic symmetry. For only the 1N7T Erbin PDZ domain NMR ensemble, all 20 models were used as independent starting structures for backrub ensemble generation. Model 1 was used for the other NMR structures in the set.

Backrub Ensemble Generation

To generate an ensemble of backbone conformations for profile prediction, we ran multiple 10,000-step, 0.6 kT Monte Carlo simulations in the Rosetta protein modeling program with backrub[85] and side rotamer moves (see below), retaining the lowest scoring structure visited during the simulation[16,85]. Erbin PDZ domain mutants were modeled by replacing all side chain conformations using Monte Carlo simulated annealing[31], followed by a two stage minimization of: 1) only side chains and then 2) a combination of side chains and the phi/psi angles of peptide positions -6 to -4. This mutagenesis procedure was repeated prior to every backrub Monte Carlo simulation.

The backrub Monte Carlo protocol we used was implemented in the Rosetta 3 software suite, with several minor differences from the Rosetta 2 algorithm[85]. Because of their low acceptance, combined backbone/side-chain moves were eliminated in favor of backbone only and side chain only moves, each made with respective frequencies of 75% and 25%. Instead of sampling discrete chi angle combinations with equal probabilities, side chain rotamers were sampled according to the frequencies given in the 2002 Dunbrack backbone-dependent rotamer library[87]. Chi angles for a selected rotamer were

sampled continuously from Gaussian distributions with the mean and standard deviation taken the PDB-derived rotamer definition. This chi angle selection scheme is similar to a previously described biased-probability side chain sampling algorithm[88]. In 10% of the side chain moves, the Dunbrack-biased chi angle sampling was bypassed and chi angles were sampled uniformly.

Simulations used the standard Rosetta full-atom energy function, except where noted below. The hydrogen bond potential was not weighted by residue burial. While removal of the hydrogen bond potential environment dependence has been shown to over-estimate the free energy difference between native residues and alanine[53], we found it to better predict specificity for polar and charged peptide residues in the PDZ binding site. Using the standard Rosetta reference energies, we observed an over-prediction of histidine, which may be due in part to the reference energies being parameterized for an environment dependent hydrogen bond potential. To counteract this, we increased the reference energy for histidine by 1.2 score units. Bond angle energies were calculated using CHARMM parameters[58].

Before backrub ensemble generation, the 1N7T NMR ensemble had an average pairwise Cα RMSD of 0.25 for the amino acids in the binding region (residues 23-28 and 79-84). After ensemble generation, the RMSD increased to 0.47. This is slightly more than the structural diversity in a backrub ensemble generated from a crystal structure (2IWP). That ensemble had an average RMSD of 0.34 for the equivalent residues (1845-1850 and 1898-1903).

Profile Prediction

The profile prediction protocol[16] previously implemented in Rosetta 2 was similarly reimplemented in Rosetta 3. Briefly, for the 5 C-terminal peptide residues, a population of 2000 sequences was optimized using a genetic algorithm over the course of 5 generations, leading to evaluation of slightly less than 10^4 non-redundant sequences per ensemble member. Scores for individual sequences were determined using Monte Carlo simulated annealing of the side chain conformations[31] of all residues with a Cα atom within 10 Å of the Cα of one of the 5 predicted peptide residues. Advances in the Rosetta 3 architecture enabled precalculation and caching of all rotamer-rotamer interaction energies, decreasing computation time by a factor of 10-20.

When transforming the evaluated sequence scores for a given backbone into a position weight matrix (PWM), the previous iteration of this algorithm selected those sequences that had intermolecular and intramolecular scores within given deltas of the wild type scores. Instead of treating the intermolecular (sum of all residue-residue scores between chains) and intramolecular (sum of all intraresidue and residue-residue scores within chains) scores separately, we used a linear combination of the two, holding the intermolecular weight fixed and weighting intramolecular scores by a given factor (see below). Additionally, because PDZ domains can interact with many different partners and it is difficult to define what the "wild-type" peptide sequence is, we used Boltzmann weighting to generate a PWM. The temperature was used to vary the contribution of higher scoring sequences to the PWM.

Each backbone in the ensemble was used to produce a different PWM, giving a distribution of frequencies for every amino acid type at every peptide amino acid position

(N values for N structures in the backrub ensemble). To collapse each distribution for a single amino acid at each position into a single frequency, a given percentile of the frequency distribution (a parameter) was calculated. For instance, a percentile of 0.5 (or 50%) would correspond to the median frequency of the distribution. Because this almost always produced, at a given amino acid position, a set of frequencies for all amino acid types that did not sum to 1, the frequencies were renormalized to sum to 1. If at a given sequence position, the percentile cutoff gave a specificity profile with entirely zero frequencies, the percentile cutoff was raised to the minimum necessary to create a non-zero profile.

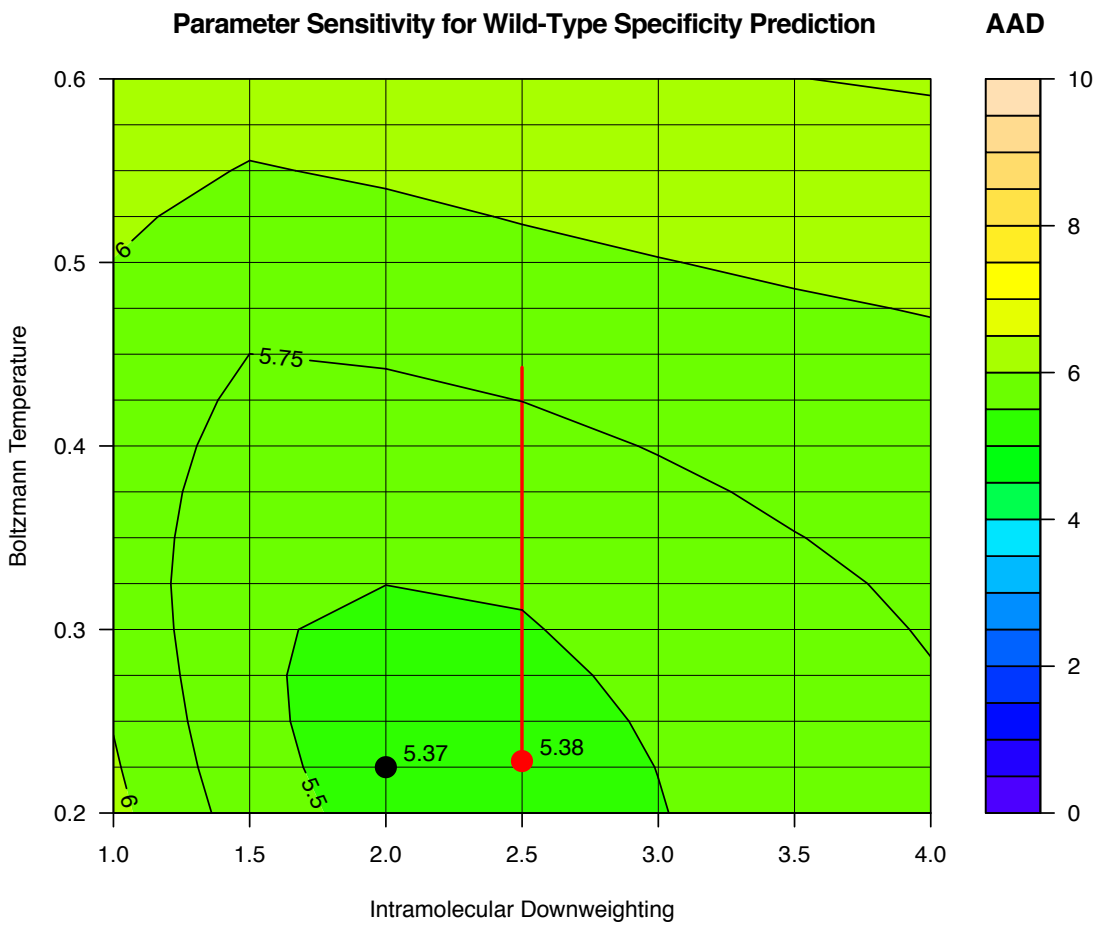
As an alternative to the above approach, we also tried generating a unified PWM by Boltzmann weighting the sequences from all backbones together into a single distribution. Prior to Boltzmann weighting, all scores were normalized such that the lowest score for every backbone was zero. This resulted in worse overall performance than the application of a percentile cutoff. This may come from the percentile cutoff being more resistant to backbones with outlier specificities, several of which are shown in Figure 2-1.

Parameter optimization

To optimize the three free parameters for specificity prediction, we enumerated all combinations of the intramolecular score weight factor (1^{-1}, 1.5^{-1}, …, 4^{-1}), Boltzmann weighting temperature (0.2, 0.225, …, 0.4), and percentile cutoff for PWM unification (0.4, 0.45, …, 0.6) to calculate a 3D grid of the average absolute difference scores for every PDZ domain. Predicted PWMs were compared to phage display using the AAD between all corresponding PWM elements. We observed that the optimal Boltzmann

temperature increased as the number of mutations made to the template structure increased, which produces flatter overall sequence profiles. (See predicted bits of information for Human PDZ and Erbin 10 Mutation datasets in Table 2-2.) This flattening of the profiles implicitly captures the increasing uncertainty as the number of mutations increases. To understand how the number of mutations affected the optimal parameters, we used L-BFGS-B[89] optimization to find the optimal linear parameter fits from 0-10 mutations that produced the lowest mean score across all datasets. The score of each PDZ domain was weighted such that each dataset (Table 2-2) contributed equally to the overall mean. For the flexible backbone predictions (Table 2-3), the only parameter found to change depending on the number of mutations was the Boltzmann temperature, which started at 0.23 for 0 mutations and increased to 0.44 for 10 mutations (Figure 2-10). The optimal intramolecular weight and percentile cutoff for PWM unification did not vary based on the number of mutations made, and were 2.5^{-1} and 0.5 (median frequency), respectively. The increasing uncertainty with more and more mutations is likely transferrable to other systems beyond PDZ domains. Using the parameters derived for this dataset, one would increase the Boltzmann temperature by 0.021 score units for every mutation made to the template.

Figure 2-10. Parameter sensitivity for wild-type PDZ specificity prediction



Two dimensions of the parameter optimization grid are shown for the wild type PDZ domains. The percentile cutoff parameter sensitivity is not shown. Parameters used for PWM generation (intramolecular weighting factor of 2.5^{-1} and Boltzmann temperature of 0.23) are indicated with a red point. The best grid point (at 2.0^{-1} and 0.225, respectively) for the wild-type domains is shown with a black point. The average absolute difference (AAD) did not show a significant change between the two. The parameters used for increasing numbers of mutations were evenly spaced along the red line, with 10 mutation predictions using a Boltzmann temperature of 0.44.

Profile Evaluation Metrics

We used several metrics for evaluating the predictions. The average absolute difference (AAD) is reported as a percentage and was defined as:

$$\frac{1}{N}\sum_{i=1}^{N}\left|E_i - P_i\right|$$

Where E is a vector of experimentally determined amino acid frequencies and P is a vector of predicted frequencies. A perfect prediction is 0% and the worst prediction is 10%. The Frobenius distance was defined as:

$$\sqrt{\sum_{i=1}^{N}\left(E_i - P_i\right)^2}$$

The best score is 0 and the worst score is $2^{1/2}$. The AAD and Frobenius distance are related, as they are both proportional to norms of the difference between vectors. The AAD is directly proportional to the Manhattan norm and the Frobenius distance is identical to the Euclidian norm. Receiver operator characteristic (ROC) curves were generated by calculating the true positive and false positive rates for predicting amino acids represented with a experimental frequency of at least 10%, given different predicted frequency cutoffs. The area under the ROC curve (AUC) was calculated independently for each individual amino acid position in each domain. The top rank was defined as the predicted rank of the most frequent experimentally observed amino acid. In the case of ties in the prediction, the maximum rank was used.

To gauge how flat or peaked a PWM was at a given peptide position, we used the information theoretic bits of information. Given a vector A of amino acid frequencies, the number of bits of information is defined as:

$$\log_2 20 + \sum_{a \in A; a \neq 0} a \log_2 a$$

DREAM4 Specificity Prediction

For the DREAM4 specificity predictions, the last five peptide residues were mutated to expected consensus motifs for each of the five target PDZ domains (2B11: FGFWV, 2B9: FGEWV, 2C6: GGFWV, 2D4: GDRWV, 2D5: FDFWV) derived from manual analysis of similar PDZ sequences from the Erbin 10 mutation dataset prior to backrub ensemble generation. As the DREAM4 predictions used an earlier version of the protocol, the reference energy for histidine was increased 1.5 score units and the reference energy for tryptophan was decreased 0.5 score units. Instead of Boltzmann weighting the peptide sequences, all sequences within a given delta of the lowest score were given equal weight, and those above discarded. The optimized parameters consisted of an intramolecular weight of 1.5^{-1}, score cutoff of 2.5, and percentile cutoff of 0.55. These parameters were determined by grid optimization to produce the best average absolute difference score for the Erbin 10 mutation dataset, irrespective of the number of mutations.

Rosetta Command Lines and Input

Backrub Ensemble Generation (Subversion revision 33982)

backrub_pilot -database **[database location]** -s **[starting structure]**
-resfile **[resfile]** -ex1 -ex2 -ex1aro -ex2aro -extrachi_cutoff 0
-backrub:minimize_movemap **[minimize movemap]**
-backrub:ntrials 10000
-score:weights standard_NO_HB_ENV_DEP.wts
Added for fixed backbone predictions: -backrub:sc_prob 1

Specificity Prediction (Subversion revision 33982)

sequence_tolerance -database **[database location]** -s **[starting structure]**
-resfile **[resfile]** -ex1 -ex2 -ex1aro -ex2aro -extrachi_cutoff 0

-ms:generations 5 -ms:pop_size 2000 -ms:pop_from_ss 1
-ms:checkpoint:prefix **[output prefix]** -ms:checkpoint:interval 200 -ms:checkpoint:gz
-seq_tol:fitness_master_weights 1 1 1 2
-score:weights standard_NO_HB_ENV_DEP.wts -score:ref_offsets HIS 1.2

standard_NO_HB_ENV_DEP.wts File

METHOD_WEIGHTS ref 0.16 1.7 -0.67 -0.81 0.63 -0.17 0.56 0.24 -0.65 -0.1 -0.34 -0.89 0.02 -0.97 -0.98 -0.37 -0.27 0.29 0.91 0.51
fa_atr 0.8
fa_rep 0.44
fa_sol 0.65
fa_intra_rep 0.004
fa_pair 0.49
fa_plane 0
fa_dun 0.56
ref 1
hbond_lr_bb 1.17
hbond_sr_bb 1.17
hbond_bb_sc 1.17
hbond_sc 1.1
p_aa_pp 0.64
dslf_ss_dst 1.0
dslf_cs_ang 1.0
dslf_ss_dih 1.0
dslf_ca_dih 1.0
pro_close 1.0
NO_HB_ENV_DEP

Code Availability

Source code for backrub ensemble generation and genetic algorithm-based specificity prediction will be made freely available as part of the Rosetta 3.2 software release. The tools developed here will also be made available online at https://kortemmelab.ucsf.edu/backrub/.

Chapter 3. Predicting the tolerated sequences for proteins and protein interfaces using Rosetta Backrub flexible backbone design

Abstract

Predicting the set of sequences that are tolerated by a protein or protein interface, while maintaining a desired function, is useful for characterizing protein interaction specificity and for computationally designing sequence libraries to engineer proteins with new functions. Here we provide a general method, a detailed set of protocols, and several benchmarks and analyses for estimating tolerated sequences using flexible backbone protein design implemented in the Rosetta molecular modeling software suite. The input to the method is at least one experimentally determined three-dimensional protein structure or high-quality model. The starting structure(s) are expanded or refined into a conformational ensemble using Monte Carlo simulations consisting of backrub backbone and side chain moves in Rosetta. The method then uses a combination of simulated annealing and genetic algorithm optimization methods to enrich for low-energy sequences for the individual members of the ensemble. To emphasize certain functional requirements (e.g. forming a binding interface), interactions between and within parts of the structure (e.g. domains) can be reweighted in the scoring function. Results from each backbone structure are merged together to create a single estimate for the tolerated sequence space. We provide an extensive description of the protocol and its parameters, all source code, example analysis scripts and three tests applying this method to finding sequences predicted to stabilize proteins or protein interfaces. The generality of this method makes many other applications possible, for example stabilizing interactions with small molecules, DNA, or RNA. Through the use of within-domain reweighting and/or

multistate design, it may also be possible to use this method to find sequences that stabilize particular protein conformations or binding interactions over others.

Introduction

The concept of "tolerated sequence space" – the set of sequences that a given protein can tolerate while still preserving its function at a defined level – has enabled considerable advances in understanding protein sequence-structure relationships and engineering new functions[90]. Knowing which sequences would be tolerated is important for designing for particular functions or inhibiting others[8], optimizing protein stability[91], anticipating drug resistance mutations[92], or characterizing potential evolutionary pathways[93]. Therefore, as illustrated by these examples, the ability to computationally estimate the tolerated sequence space of a protein is of both great scientific interest and practical utility. Even in cases where it is especially difficult to predict sequences optimized for a given function (for example the rate of an enzymatic reaction or the emission spectrum of a fluorescent protein), screening from a pool of predicted tolerated sequences can increase the likelihood of diversifying existing or identifying new functions[71].

To experimentally estimate the tolerated sequence space for a given protein fold, one can either use sequence alignments of orthologous proteins, or a high throughput technique such as phage display. The disadvantage of using evolutionary information is that it represents only a part of the total tolerated sequence space, and may have confounding constraints that have not yet been characterized. Moreover, simply replacing amino acids in one protein with those observed in other members of the protein's family often fails to preserve function[94], because residue interactions in proteins can be

exquisitely interdependent. Phage display has been extensively used to probe the tolerated sequence space of both protein folds[95-97] and protein-protein interactions[14,15,97-101]. Phage display selects for protein binding, but through the use of a binding partner that does not interact directly with the mutated amino acids, binding can be used as a proxy for protein stability. Phage display methods are limited by the number of sequences that can be produced and analyzed. For example, allowing all 20 naturally occurring amino acid types at all positions in a standard-size protein-protein interface is generally not possible in a single screen. Therefore, computational methods that can reduce the enormous number of possible sequences to those that are more likely to be functional are extremely useful, in particular to focus libraries that can then be screened experimentally much more efficiently.

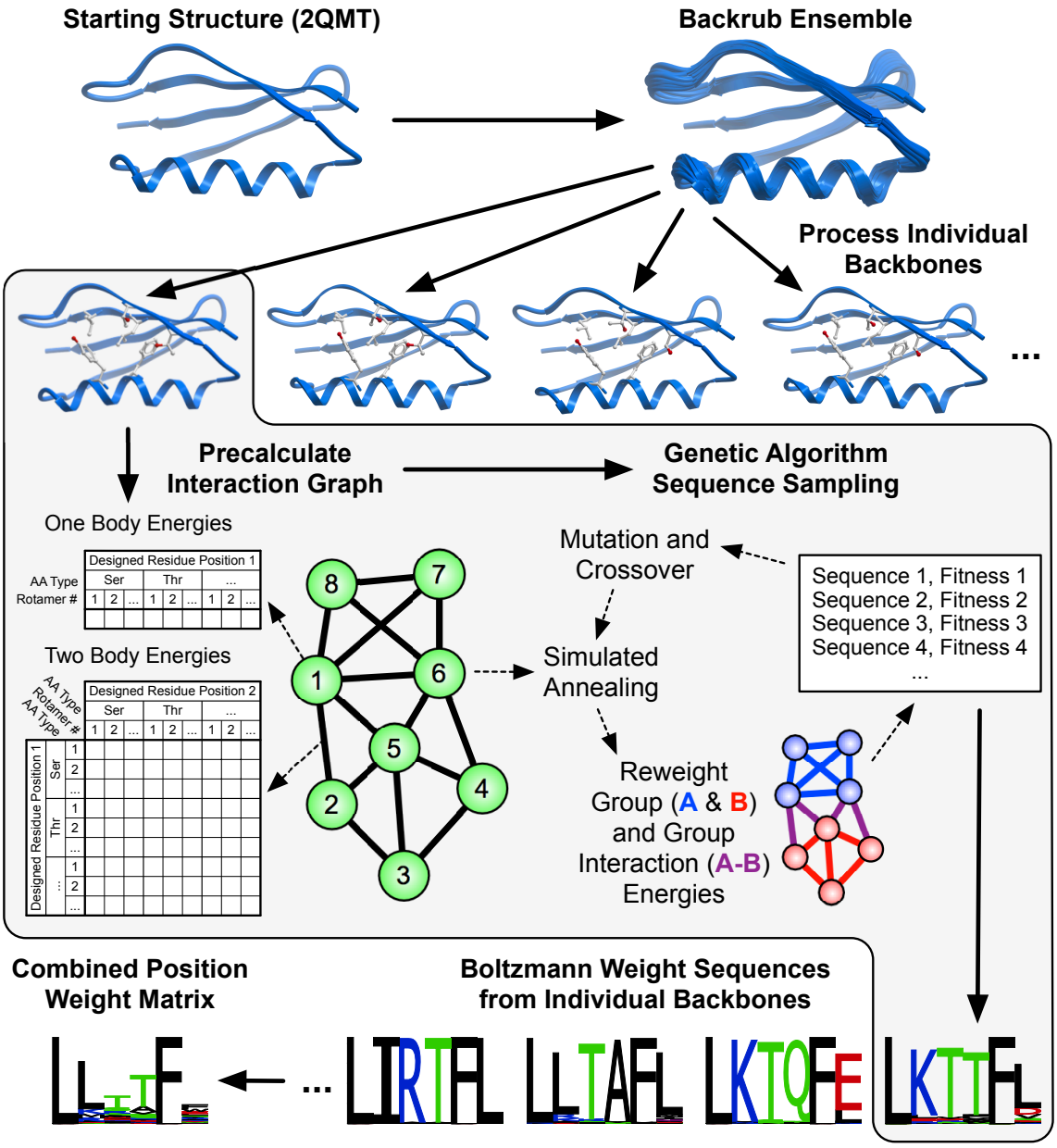
Here we provide a generalized strategy and a set of protocols for using flexible backbone protein design to predict the tolerated sequence space for a given protein fold or interaction, implemented in the Rosetta software suite for molecular modeling. Developing and, importantly, adequately testing flexible backbone protein design approaches has been a long-standing problem ([102] and references therein). Several approaches to considering backbone flexibility in computational protein design have been described. These include sampling small random perturbations of the ψ and φ backbone torsion angles[30], taking backbones from a parametric family of structures[32] or using normal mode analysis[33], utilizing families of crystal structures[103] or computationally generating backbone ensembles[16,70,104], adapting dead end elimination to incorporate backbone changes[105,106], and iterating between sequence and structure optimization[59,69,86,107]. Our protocol utilizes “backrub” conformational moves in

Rosetta[85,108] inspired by observations of conformational heterogeneity in high-resolution crystal structures[1]. We and others[109] have previously shown that backrub moves capture a significant fraction of the conformational variability explored by proteins to enable sequence changes[70].

We first describe the methodology and simulation protocol in-depth. Next we report key benchmarking results using phage display data. These include a new example demonstrating prediction of the tolerated sequence space of the 6 core and boundary residues in GB1, as well as the benchmarks of the generalized protocol for two systems we previously used to test variants of the computational method: the human growth hormone-human growth hormone receptor (hGH-hGHR) interface, for which approximately 1000 tolerated sequences have been determined in six phage display screens[101], and over 8000 sequences from 169 screens of naturally occurring and synthetic PDZ domain-peptide complexes[110]. The main new aspects here are the generalized protocol with a consistent set of parameters tested in several systems, detailed documentation on how to perform the computations (including all necessary source code and analysis tools as well as example input and output as part of this Rosetta collection issue), and the application of this method to the problem of predicting tolerated sequences for fold stability. We hope that providing a well-documented consistent protocol that can be applied to other systems both in a prospective or retrospective manner will stimulate further studies leading to a better understanding of transferability issues as well as scoring and sampling problems. We conclude with a discussion of current limitations as we see them and potential strategies for overcoming them, as well as future applications of the methodology described here.

Methods

Figure 3-1. Scheme for predicting tolerated sequences for a protein fold or interaction



The input is at least one protein structure from the protein structure databank (2QMT in the example). Rosetta first creates an ensemble of backbone conformations using the backrub method[85], then predicts sequences consistent with each conformation in the ensemble, scoring each trial sequence–structure

combination using the Rosetta score12, and finally combines the sequences into a predicted sequence profile. This approach ignores potential covariation between side chains. To speed up calculations, the scoring function is split into one-body terms describing the intrinsic energy of a particular residue conformation, and two-body terms between residues; these residue-residue interaction terms are assumed to be pairwise additive. One- and two-body terms are pre-calculated and stored in an interaction graph[111] such that optimization of sequence–structure combinations for entire proteins only takes seconds using look-up tables of interaction energies. For the interaction graph, vectors of residue self-energies (one body) are stored on the vertices (green circles) and matrices of residue interaction energies (two body) are stored on the edges (thick black lines). Computed interaction energies within proteins, between proteins, or between groups of residues can be reweighted to generate custom fitness functions for specific applications. This flexibility in scoring residue groups allows modeling of separate requirements, such as those to maintain residues required in an interaction interface with a binding partner. Group and group interaction reweighting is typically only done for protein-protein interactions. (For the monomeric GB1 domain shown here, no reweighting was applied.)

Definitions of Sets of Amino Acid Positions

The protocol and methods described here (Figure 3-1) aim to identify the amino acid types that can be tolerated at a given set of positions while still preserving protein fold stability and function (most commonly represented as binding). There are two general stages of the protocol: (1) creation of a set of protein backbone conformations (ensemble generation), and (2) prediction of sequences consistent with the ensemble conformations. The input to the protocol is at least one protein structure in PDB format and a definition of residue positions. There are three sets of sequence positions that can be defined: The first set of amino acids includes those that are mutated prior to ensemble

generation in stage (1) and often remain the same for all subsequent simulations. These positions will be referred to as the "premutated" positions. Definition of premutated positions is optional. If no positions are chosen, the input sequence will be used for ensemble generation. The second, most important set of positions are those that can vary their amino acid type in stage (2); these have to be defined by the user and will be referred to as the "designed" positions. For each designed positions, a set of considered amino acid types can be defined, as described in the "Detailed Workflow" section below. A final set of amino acids includes those whose conformations (but not amino acid types) change during sequence scoring in step (2). This set will be referred to as the "repacked" positions and is often a superset of the "premutated" positions. These positions can be determined by the user or automatically chosen by the protocol. The predicted tolerated amino acid types at the designed positions will depend on how many other positions are allowed to vary simultaneously (for example, allowing residues in a surrounding shell to be repacked may help to accommodate different amino acid choices at designed positions). For all of the results reported here, as well as a in previous study[110], residues chosen for repack included all those with a C-alpha atom with 10 Å of the C-alpha atom of a designed position. This is the current default if repacked positions are chosen automatically by the protocol. Smaller sets of repacked positions can be used to restrict sequence diversity and simulate more conservative changes closer to the starting sequence and conformation, or to reduce the computational time required for the algorithm.

Phage Display Datasets Used for Testing

Our study uses three datasets where a considerable number of tolerated sequences (not just a few) in a given system had been determined experimentally by phage display. The first test dataset investigated effects of sequence variations on the stability of the B1 domain of protein G (GB1) by using phage display to screen a 20 amino acid library for 6 total residues (3 core and 3 boundary)[96]. The second set, one of the largest phage display studies on protein-protein interactions, involved the human growth hormone (hGH) and human growth hormone receptor (hGHR)[101]. Through 6 separate phage display experiments randomizing 5-6 positions each, 35 amino acid positions on hGH were sampled to determine tolerated sequence space for hGHR binding. The third set is taken from a study that has determined the peptide sequence space tolerated for binding to 82 naturally occurring PDZ domains and 91 PDZ single point mutants[14].

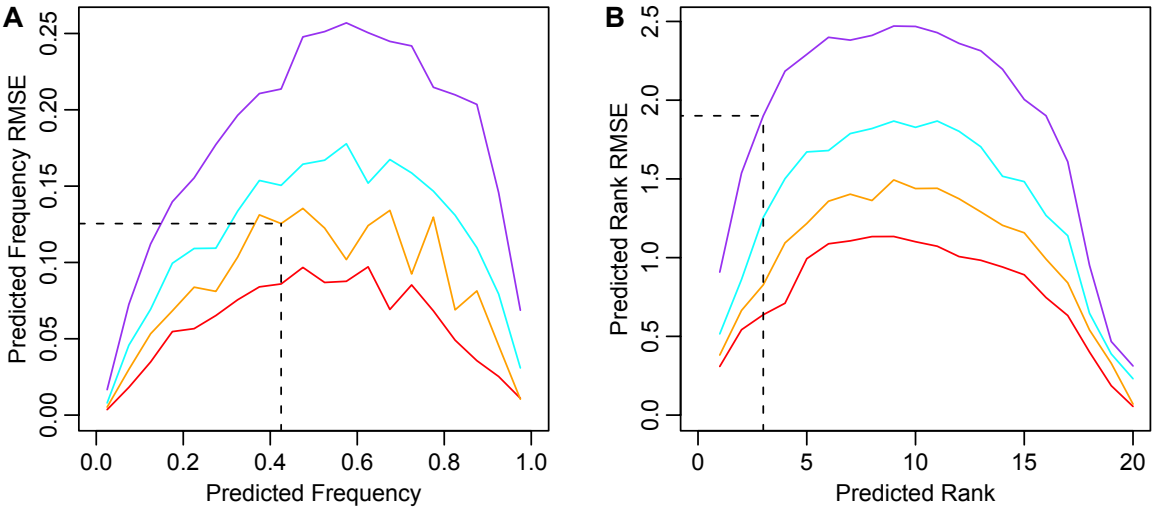
Input Structures

All GB1 simulations were started using PDB code 2QMT[112], which had a resolution of 1.05 Å, the highest available to date. The designed sequence positions were allowed to sample any of the 20 canonical amino acids and included residues 5, 7, 16, 18, 30, and 33. For the 56 residue GB1 domain, the repacked residues included all but 22-24, 40, 42, and 46-49 (i.e. 47 out of 56 residues). All hGH/hGHR simulations used a 2.6 Å resolution structure with PDB code 1A22[113]. PDZ/peptide simulations used the input structures previously reported[110]. For hGH/hGHR and PDZ/peptide simulations, the designed sequence positions were allowed to sample any amino acid but cysteine.

Backrub Ensemble Generation

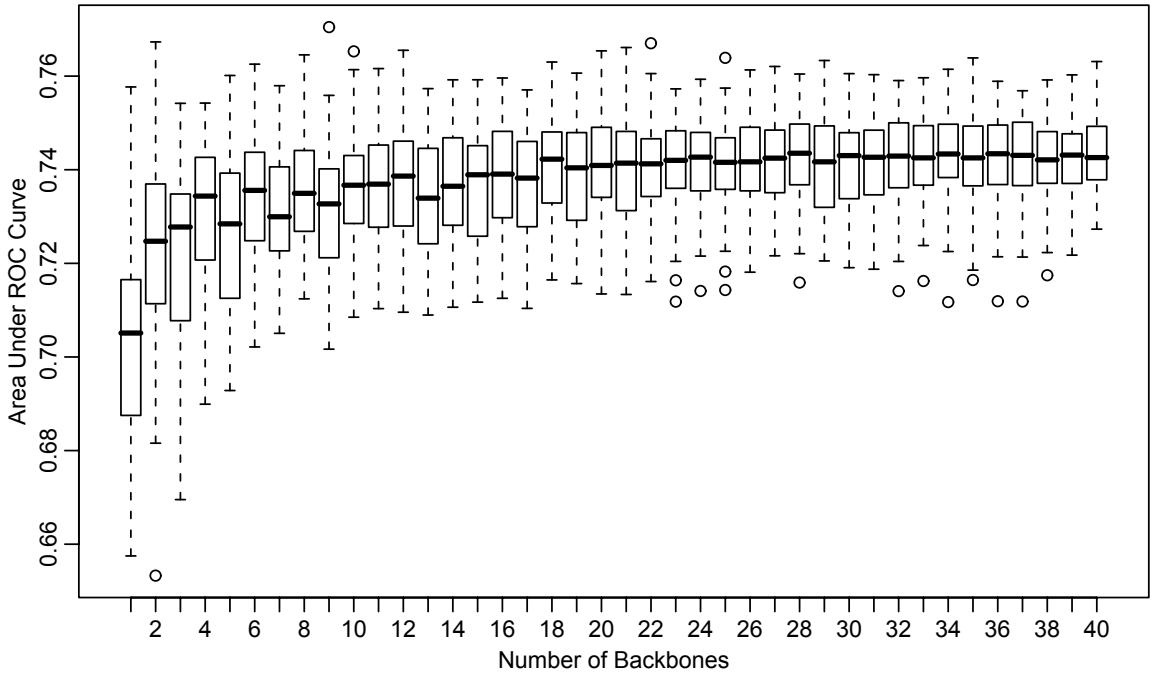
During the first stage of the prediction protocol, an ensemble of backbone structures is generated using backrub Monte Carlo simulations[85,110]. Both the backrub simulations and sequence sampling were implemented in the Rosetta 3 software suite[114]. The move set consists of 75% backrub backbone moves, 22.5% chi angle moves biased by the amino acid rotamer probabilities observed in the protein structure databank[87], and 2.5% uniformly sampled chi angle moves. Moves are accepted or rejected with the Metropolis criterion[115] using a kT of 0.6. After 10,000 moves are applied, the lowest energy structure from the simulation is output for the next stage of sequence sampling. For the results presented here, 200 backbones were generated from independent backrub Monte Carlo simulations for each starting structure. The exception was the hGH/hGHR predictions, which used 100 backbones to match the number of structures used previously[16]. Using fewer backbones will generally produce reasonable results, but exhibit stochastic variation. Figure 3-2 gives estimates of the variation as a function of the number of backbones based on a benchmark using 2000 backbones and approximately 240 million sequence scores. Predicted ranks of selected amino acid types are generally more robust than predicted amino acid frequencies. Figure 3-3 illustrates the dependence of prediction performance on the number of backbones. Predictions using less than 20 backbones show reduced area under ROC curve scores.) If possible, at least 100 backbones are recommended for results more robust to stochastic variation (Figure 3-2). For the scoring metrics summarized in Table 3-1, the average standard deviation over three runs when using 100-200 backbones was between 0.4-1.9% of the dynamic range of each measure.

Figure 3-2. Increasing the number of backbones reduces stochastic variation



2000 backbones were generated for each of the prediction simulations used here, resulting in approximately 240 million sequence scores. The frequencies calculated from the entire dataset ($kT = 0.23$) were treated as the ground truth and used to calculate the root mean squared error (RMSE) for subsets of the data using 200 (red), 100 (orange), 50 (cyan), and 20 (purple) backbones each. **(A)** Frequency data were divided into 20 equally spaced bins and the predicted frequency RMSE was calculated for each bin. For example, if the method is applied using 100 backbones, and an amino acid frequency is predicted to be 0.425, then the estimated error is approximately 0.125 (dashed lines). **(B)** The data were divided by rank and the predicted rank RMSE was calculated for each rank. For example, if this method is applied using 20 backbones, and an amino acid rank is predicted to be 3, then the estimated error is approximately 1.9 (dashed lines). For 20 backbones, the stochastic contribution to the root mean squared error (RMSE) of the predicted frequency can be up to 0.25, which is 25% of the dynamic range. The predicted ranks are more robust, with an RMSE of up to 2.5, or 12.5% of the dynamic range. 100 and 200 backbones reduce the stochastic error by approximately 2-fold and 2.5-fold over 20 backbones.

Figure 3-3. Dependence of prediction performance on number of backbones



Distributions of area under ROC curve (AUC) values are shown for varying numbers of backbones. Prediction performance plateaus at approximately 20 backbones. Each boxplot shows the distribution of mean AUC values for 50 sets of independent backbones (mean AUC values were computed across all datasets, from the equivalent of rows 1, 4, and 5 of Table 3-1). Horizontal lines represent the median, the box spans the interquartile range (IQR), whiskers extend to the furthest data point up to 1.5 times the IQR from the box, and data points outside the range are shown with circles. This figure used the same data that were generated for Figure 3-2).

The conformational variation between different polypeptide backbones modeled by the backrub method is generally small, and using larger variation often leads to flat profiles that do not agree well with experimental data[70]. For all backrub ensembles used here, the average C-alpha atom RMSD from the starting structure was 0.4-0.9 Å.

By default, the starting sequence in the input PDB is used when the entire protein structure is sampled in the fixed-sequence backrub Monte Carlo simulations in stage (1). However, there are several circumstances in which a user may want to change the sequence of the input structure prior to ensemble generation. For example, it may be desirable to mutate residues to more closely represent the experimental system. Also, experimental data may suggest that another amino acid sequence shows greater function than the sequence in the starting structure. As shown in a previous study[110], mutating the starting structure to that sequence prior to ensemble generation improves prediction performance.

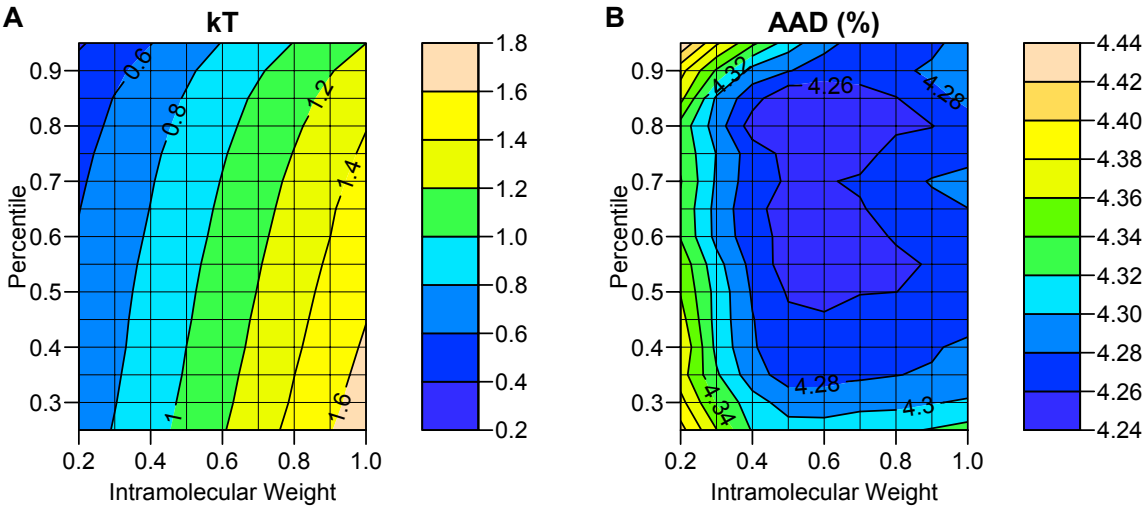
Such mutations can be made manually prior to backrub Monte Carlo or done automatically as a preprocessing step of the simulation. If the automatic option is used, the side chain conformations of the mutated residues and all other residues are optimized using simulated annealing[31]. If desired, iterative minimization can be applied by including progressively more degrees of freedom in three stages (first chi angles only, then chi/phi/psi angles, finally chi/phi/psi angles and rigid body degrees of freedom).

Designed Position Sequence Scoring

Before any sequences are scored, a graph of pairwise interaction energies between all possible conformations of all allowed amino acids is precomputed[111]. The first step of scoring a given sequence is to determine the conformations of side chains that minimize the score of the entire structure. We term this score the "raw Rosetta score". This is done using Monte Carlo simulated annealing[31]. Once that conformation is identified, the interaction energies between and within user-defined groups of residues, often individual protein polypeptide chains, are calculated. The actual total fitness score of a given

sequence is a user-defined linear combination of the self-energies and interaction energies between these groups of residues. We term this score the "reweighted Rosetta fitness score". For the dataset of PDZ domain-peptide complexes, the optimal weights were found to be 1 for the intermolecular PDZ-peptide interaction energies, and 0.4 for the intramolecular score[110]. We used those same weights for the hGH/hGHR interaction energies. Varying these weights in a grid search showed that these parameters are transferable to the hGH system, where they produced nearly optimal fits to the phage display data (Figure 3-4). For the GB1 protein fold stability dataset, only the intramolecular weight was applicable, which was kept at 0.4.

Figure 3-4. hGH/hGHR interface data processing parameter sensitivity



Sequence tolerance prediction for the hGH/hGHR interface is not highly sensitive to data processing parameters. For the 35 designed positions in the human growth hormone (hGH)/human growth hormone receptor (hGHR), position weight matrices (PWM) were generated using a grid of intramolecular weights and percentile cutoffs. **(A)** At each grid point, the value of *kT* was fit such that the average number of bits of information matched that observed in phage display (i.e. 0.89 bits, see Table 3-1). **(B)** In the resulting PWMs, the average absolute difference (AAD) between phage display and prediction shows little sensitivity to

the processing parameters. The point with parameters equivalent to those found in the PDZ/peptide predictions (0.4 intramolecular weight, 0.5 percentile) is only slightly worse (by 0.04% AAD) than the lowest (best) AAD sampled on the grid. The other rank-based metrics also do not change significantly across the same parameter space and are less sensitive to changes in *kT* (data not shown).

The general protocol described here for all three datasets uses the default "score12" energy function in Rosetta 3, with its implementation in the 3.2 release. The only modification to the default score12 energy function was to increase the reference energy of histidine by 1.2 score units, as was done previously for PDZ/peptide specificity prediction[110]. Histidine reweighting was found to improve performance across all three datasets tested here. Other than histidine reweighting, the previous scoring function used for PDZ-peptide specificity prediction[110] differed from score12 in a number of ways: First, the Ramachandran and omega angle energy terms were turned off. (Because omega angles were never varied during the simulations, the omega energy term had no effect.) Second, the short-range backbone-backbone hydrogen bond and the amino acid probability given phi/psi terms were doubled. Third, turning off environment dependent hydrogen bonding was found to improve performance for PDZ-peptide specificity (it is on per default in standard in Rosetta 3). The first two differences to the published method[110] listed above, namely the addition of two terms and the change of two weights, are part of a "score12 patch" that is standard in Rosetta 3 methods using score12, but was not used for the PDZ-specificity prediction[110].

Background on the "standard" and "score12" Rosetta Energy Function Weights

In Rosetta 2, side chain optimization via simulated annealing was done with weights (called "packer" weights) now defined as "standard" in Rosetta 3. Full backbone optimization in Rosetta 2 differed from these packer weights by including several small changes. First, short-range (i.e. alpha helical) hydrogen bond weights were halved. Second, a scoring term based on smoothed Ramachandran plots was added to better restrain phi/psi angles. To compensate for the addition of the Ramachandran score, the weight was halved for the score term taking into account the probability of an amino acid given phi and psi. In Rosetta 3, these three changes were incorporated into the "score12 patch". Rosetta 3 also added an energy term restraining the omega peptide bond angle to the default scoring function, which was similarly incorporated into the score12 patch. Rosetta 3 now generally uses the same scoring function for side chain optimization and full backbone relaxation, which is "score12", including the changes included in the "score 12 patch", by default.

Genetic Algorithm Optimization

Sequence sampling proceeds using a genetic algorithm independently on each backbone in the ensemble. The initial population is generated by selecting random sequences from the user-defined set of allowed amino acids at the designed positions. In addition, a single population member is generated that contains the sequence from a single simulated annealing call where all possible amino acids are allowed (i.e. the sequence with the best raw Rosetta score). The population size for each generation is 2000 sequences and 5 total generations are produced, including the initial population. This results in slightly less than 10,000 sequences scored for each backbone. If 200

backbones are generated, this will result in up to $2*10^6$ sequence scores, which is within an order of magnitude of the theoretical size of the 5 and 6 amino acid libraries ($3.2*10^6$ and $6.4*10^7$ sequences, respectively) used for experimental screening in the GB1, hGH/hGHR, and PDZ systems. In contrast to phage display, however, 4 out of 5 generations of sequences are not selected randomly from all possible combinations, but are increasingly enriched in later generations using an applied fitness function. Changing the number of generations to 30 was previously shown to produce equivalent results[16].

For the genetic algorithm the reweighted Rosetta fitness score is used to determine the fitness for each sequence. For every new generation of the genetic algorithm, the best fitness sequence is automatically propagated to the next generation. The remaining sequences are generated by crossover and mutation of parental sequences from the previous generation. Parental sequences are selected by tournament selection, in which two random sequences are chosen, and the sequence with the best fitness is chosen to be a parent. Half of the new population members are generated by crossover, in which two parents are chosen and the identity of each amino acid is randomly selected between the two parental sequences. Unlike physical DNA crossover, there is no linkage between sequence positions close to one another. The other half of the new population members are generated by mutation, in which a single parent is chosen and each of its amino acids is mutated with a 50% probability.

While our predictions agree reasonably well with experimental data, undersampling of sequence space and trapping in local minima are possible caveats of the applied optimization algorithms. Other sequence optimization methods could be compared to our results, such as approaches that are guaranteed to find the global

minimum energy sequence[116]. Along these lines, we have found that predicted sequences using Rosetta Monte Carlo optimization are similar to results of an approach that finds all low-energy sequences within a given energy threshold of the global minimum of the Rosetta scoring function ([117] & unpublished results). We therefore believe that inaccuracies in scoring and the inability to more accurately sample backbone variation upon sequence changes are more significant contributors to the remaining discrepancies with experimental data than fixed-backbone sequence sampling issues.

Sequence Processing

The sequences output by the genetic algorithm are processed into a single position weight matrix (PWM) by first calculating a PWM for each individual backbone, and then merging the PWMs together. Individual backbone PWMs are calculated by Boltzmann weighting ($w = e^{\Delta G/(kT)}$, w: sequence weight, ΔG: reweighted Rosetta fitness score, kT: Boltzmann factor) each of the individual sequences and calculating residue frequencies. The default Boltzmann factor used here was 0.228, as determined previously[110]. The Boltzmann factor can be changed by the user (see accompanying protocol capture). PWMs are merged together with the assumption that all backbones are equivalent. The contribution of individual backbones is not weighted by their total scores because the total energy of a backbone can be largely determined by structural features distant from the designed region, which could add considerable noise. Instead, to generate a merged PWM, the median frequency for every position/amino acid type element across all backbones is calculated. Taking the median is more robust to outliers than taking the mean or weighted mean. Users can alternatively use any percentile cutoff they wish (in the accompanying protocol capture postprocessing script), with the 50th percentile being

equivalent to the median. While PWM analysis ignores correlations between sequence positions, a similar analysis could be done using the Boltzmann weighted sequences to calculate residue co-occurrence at two or more positions.

Phage Display Data

Raw sequencing data (Andrea G. Cochran, personal communication) from round three of phage display of the Streptococcus GB1 domain using the human IgG Fc domain as bait[96] included 185 sequences. Sequences were excluded that contained ambiguous reads, early stop codons, and mutations at sites other than those explicitly varied, leaving 171 total sequences and 167 unique sequences. For the hGH/hGHR example, phage display frequencies were taken from Figure 2 of the authors' publication[101]. Erbin PDZ frequencies were used as previously described[110].

Detailed Workflow

The following is a detailed description of the steps that need to be taken to apply the described method to another system, or reproduce the results of the analysis done here. The protocol capture contains all the input files, command lines, and postprocessing scripts for replicating the computations, figures, and tables given here. (It can be downloaded at http://kortemmelab.ucsf.edu/data/)

Select and prepare input structure. The input structure should be a crystal structure, NMR structure, or high quality homology model. If multiple structures are available (e.g. an NMR ensemble), the input structures should be placed into separate PDB files for input into the *backrub* application. Input of multiple structures can be facilitated by the *backrub_seqtol.py* script if they are numbered sequentially starting at 1, for instance *PDB_01.pdb*, *PDB_02.pdb*, etc.

Determine which amino acids will be premutated, designed and repacked and create resfiles. Each of these sets of residues is described above. If there are no premutated residues, a *backrub* resfile is unnecessary. If there are, those should be placed as *PIKAA X* (picking the desired amino acid X by one letter code) in the *backrub* resfile, with the default behavior for all other residues specified as *NATAA* (i.e. sample side chain conformations while preserving the native amino acid type).

A resfile is required for the *sequence_tolerance* application and should contain the designed and repacked sets of residues. Designed residues should use either *ALLAA* (all amino acids) or *PIKAA XYZ…* (picking the allowed amino acid residues with one letter codes X, Y, Z, etc.). Repacked residues should use *NATAA* and nonrepacked residues should use *NATRO* (native rotamer). A convenience script, *seqtol_resfile.py*, will generate a resfile for an input structure and a given set of designed residues, automatically determining the repacked residues having C-alpha atoms within 10 Å of the designed residue C-alpha atoms.

Determine whether to minimize after premutation and create movemap file. If premutated residues are specified using the *backrub* resfile, an optional stage of minimization is recommended and can be enabled after the premutation step but before the backrub Monte Carlo simulation. To do so, a movemap file (specified using the *-backrub:minimize_movemap* option) must be created which specifies the sidechain, backbone, and rigid body degrees of freedom to minimize. This was done, for example, in the case of the Erbin mutant V83K to minimize all side chains and the most N-terminal backbone dihedral angles of the peptide. If backbone dihedral angles or rigid body

degrees of freedom are minimized, care should be taken with the fold tree; information on the fold tree is given in the Rosetta 3.2 manual and Leaver-Fay et al[114].

Determine whether to sample phi/psi angles directly and create movemap file. While not used for any results published here or elsewhere to date, it is possible to have the backrub Monte Carlo procedure also make small direct perturbations to phi/psi angles of the protein. To do so, a movemap file (described in the Rosetta 3.2 manual) must be provided using the *-in:file:movemap* option. In addition, the *-sm_prob* option, which gives the probability of making a "small" combined phi/psi move[27], must be given a positive value. The fold-tree warning above about minimizing backbone degrees of freedom applies to backbone perturbations as well.

Create backrub ensemble. The *backrub* application can be run once and produce many different backbones, each starting from the original specified structure. As an alternative, the *backrub* application can be run separately each time a new ensemble member is required. The *backrub_seqtol.py* script does this and renames the resulting structures as if they came from a single execution of the *backrub* application. On a heterogeneous cluster, this stage took 20 seconds to 10 minutes per backbone for the results published here.

Determine appropriate fitness function and score a large number of sequences. The *sequence_tolerance* application is used to score a random selection of sequences that are increasingly enriched in those that conform to the prescribed fitness function, whose coefficients are specified using the *-seq_tol:fitness_master_weights* option, which is fully described in the Rosetta 3.2 manual. The fitness function individually weights interactions between and within sets of residues defined by the PDB chain identifier. The

sequence scoring process took 15 minutes to 5 hours per backbone for the results published here.

Post-process sequence scores. Post processing of the results is done using an *R*[118] script in the *sequence_tolerance.R* file. The function used, *process_specificity()*, takes several parameters. The first parameter, *fitness_coef*, allows the user to specify a vector of coefficients for the fitness function used in postprocessing. The second parameter, *temp_or_thresh*, allows the user to specify the Boltzmann factor (temp) or threshold cutoff value above the minimum fitness (thresh). The third parameter, *type*, determines how sequences are weighted and *temp_or_thresh* is interpreted. Sequences are either weighted using the Boltzmann equation ("boltzmann"), or a binary threshold cutoff ("cutoff"). The final parameter, *percentile*, gives the percentile to use for merging frequencies from multiple backbones together. The default value, 0.5, corresponds to the median frequency across all backbones.

Good results can still be obtained even if the genetic algorithm uses weights for tournament selection that are slightly different from those used for final sequence scoring. For instance, in a previous PDZ peptide specificity study[110] and the results reported here, the genetic algorithm used a ratio of 1:2 between the weights of intramolecular and intermolecular interactions, while the final sequence scoring was done using a ratio of 1:2.5. The user thus has the flexibility to make small perturbations to the weights during post-processing without running the whole algorithm again.

Caveats and Factors Not Taken into Account

For the case of interface optimization, residue-residue interactions across the interface are upweighted in lieu of explicitly calculating the scores of the two partners

separately and in complex. This was done in part for computational efficiency and in part because separate calculation of scores was found to add noise to interface ΔΔG prediction (unpublished results). If the designed residues change their conformations in energetically significant ways when not in complex, the algorithm will neglect those contributions to binding affinity. Also, the contribution of conformational entropy changes is not modeled.

Results

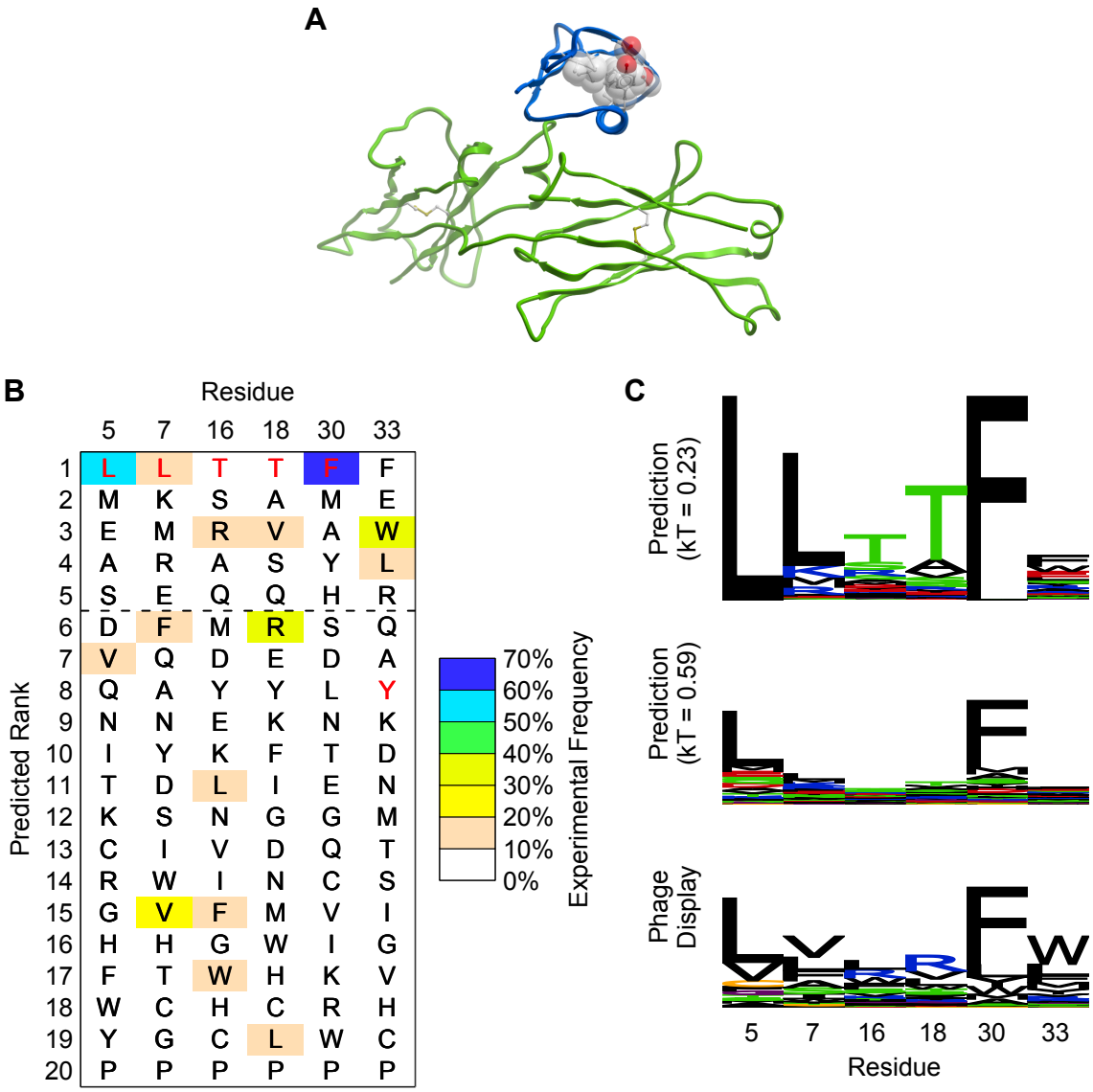
In the following, we show example results that assess the performance of Rosetta Backrub sequence tolerance predictions using three different experimental datasets that determined tolerated sequences for protein fold stability[96] and protein binding[16,110] using phage display. Two of these tests were previously performed with an earlier Rosetta version[16] or scoring function[110]. Here we evaluate the generality of the Rosetta 3 standard protocol described in this Rosetta collection on all three datasets, compare to previous results, present a new test on a dataset of tolerated sequences for fold stability and provide an extensive set of customizable simulation and analysis tools in addition to all source code. Overall, the generalized protocol captures a significant fraction of the observed sequence space in all three datasets (Table 1), with values for the area under a ROC curve between 0.64 and 0.87, and the fraction of sequence space captured by the top 5 ranked amino acid types between 54 and 82%.

GB1 Fold Stability Tolerated Sequence Space Prediction

The fold stability test used a dataset by Kotz et al who determined tolerated sequences for three residues in the core (L5, L7, and F30) of the B1 domain of protein G (GB1) and three residues bordering the core (T16, T18, and Y33)[96]. The authors utilized

the ability of the GB1 domain to bind to the human IgG Fc domain for a phage display screen. The side chains of the six GB1 residues varied in the experiment are at least 7 Å from any heavy atom on the IgG Fc domain in the cocrystal structure between the GB1 and IgG Fc domains[119], as shown in Figure 3-5. Mutating the GB1 residues should thus primarily affect the stability of the GB1 domain and report on sequences tolerated for fold stability, instead of selecting sequences that modify the interaction directly. After three rounds of GB1 display on phage, using IgG as bait, the authors obtained 171 full-length GB1 sequences suitable for analysis.

Figure 3-5. Prediction of tolerated sequences for GB1 fold stability



Frequently observed amino acids in phage display are enriched in the GB1 prediction. **(A)** The structure (PDB code 1FCC) of Streptococcal GB1 (blue) is shown bound to the Fc domain of human IgG (green). The core and peripheral residues that were randomized in phage display are shown with sticks and transparent spheres. The side chain atoms (starting at C-beta) of these amino acids are at least 7 Å away from any atom of the Fc domain, making residues selected at these positions unlikely to interact directly with the Fc domain. **(B)** Amino acids are ranked individually for each sequence position by computationally

predicted frequency (using the Boltzmann factor $kT = 0.23$, as described in the main text). Wild type residues, which were used in protein ensemble generation, are shown in red. The dashed line indicates a typical cutoff of picking the top 5 amino acid choices at each position. **(C)** Sequence logos (LOLA, University of Toronto) are shown for predictions with two different Boltzmann factors. The relative degree of specificity (in terms of bits of information, y-axis) shows good correspondence between prediction and phage display. Increasing the Boltzmann factor lowers the overall specificity and brings the absolute frequencies closer to phage display.

The results of applying the generalized sequence tolerance prediction protocol described in Methods are shown in Figure 3-5. Consistent with previous studies[110], the prediction of sequence rank is often better than the absolute frequencies. Therefore, we compared the predicted ranking of the amino acid types at each position to the experimentally observed frequencies. Averaged over the six positions, 57% of the frequently observed amino acids are found in the top five predicted amino acids. This performance metric, which is helpful for gauging the usefulness of the prediction for library design or other protein engineering applications, is used along with other metrics to compare all three datasets in Table 3-1. For actual protein engineering applications, it is critical to correctly identify at least one "viable" (tolerated) amino acid type at each position. Here, for all six positions, the prediction finds at least one frequently observed amino acid (greater than 10% frequency) within the top five ranked amino acids. (This analysis ignores co-variation between positions, which can be obtained from analysis of the actual predicted sequences).

Table 3-1. Summary of backrub tolerated sequence prediction performance.

	Proteins	Residue positions	Bits of information		Fraction Top 5 (%)	AAD (%)	AUC	Rank Top
			Phage display	Predicted				
GB1 (kT=0.23)	1	6	1.58	2.66	56.9	5.61	0.74	6.17
GB1 (kT=0.59)	1	6	1.58	0.89	54.2	4.05	0.71	7.17
hGH/hGHR[1]	1	16	1.19	3.58	59.3	7.46	0.75	6.00
hGH/hGHR[2]	1	35	0.89	3.24	41.9	7.48	0.64	7.72
PDZ/Peptide	5	25	3.11	2.82	81.7	4.16	0.87	2.84
PDZ/Peptide[3]	5	25	3.11	3.06	82.0	3.67	0.88	2.76

[1] 16 designed hGH amino acid positions as defined in [16] and shown in Figure 3-6.

[2] All designed hGH amino acid positions shown in Figure 3-7.

[3] Performance metrics based on position weight matrices from Smith & Kortemme 2010[110].

Scoring metrics are used as defined previously[110]. Fraction Top 5 gives the average fraction (for every position) of amino acids with phage display frequencies ≥ 10% in the predicted top 5 ranked amino acids. AAD gives the average absolute difference in amino acid frequency between prediction and phage display. AUC gives the area under receiver operator characteristic curve, with true positives defined as those with phage display frequencies ≥ 10%. Rank top gives the average rank of the most frequently observed amino acid in phage display. The table gives results from one set of predictions as described in Methods. To gauge the variability, we repeated the predictions three times and calculated the standard deviation of the scoring metrics. The absolute standard deviations and dynamic ranges are 0.4/4.32 (Bits Predicted), 1.9/100 (Fraction Top 5), 0.4/10 (AAD), 0.006/1 (AUC), and 0.2/19 (Rank Top). As a percentage of the dynamic range of a given metric, the average standard deviations (over the first 5 rows) were: 0.9% (Bits Predicted), 1.9% (Fraction Top 5), 0.4% (AAD), 0.6% (AUC), and 1.1% (Rank Top).

In this example test case, the predictions reveal bias towards the native, input sequence at five positions. Two out of those five positions, core residues L5 and F30, show the wild type sequence to be the most frequent in phage display. Two of the border positions, T16 and T18, are incorrectly biased towards the input sequence. One of those positions is flat, with no single residue having greater than 20% frequency, so it is not surprising that the input bias overwhelms the relatively weak preferences. For residue

Y33, the prediction correctly ranks both frequently observed amino acids in the top five ranked amino acids and above the input wild-type tyrosine.

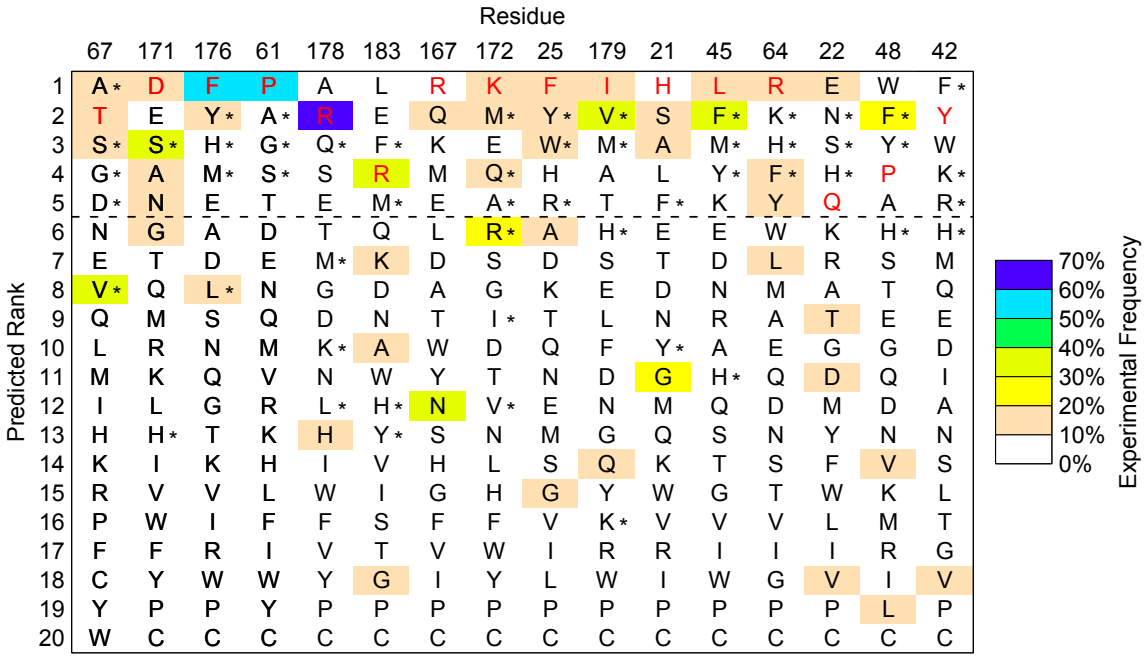
Human Growth Hormone/Human Growth Hormone Receptor Interaction

The first iteration[16] of a sequence tolerance prediction method was implemented in Rosetta 2 and applied to the recapitulation of data from phage display selections of human growth hormone (hGH), using human growth hormone receptor (hGHR) as bait[101]. Besides using an entirely different implementation, which made the present computations approximately 2-20 times faster, there were several algorithmic differences between the previous approach and the generalized protocol presented here.

The main difference lies in the way sequences were scored, filtered and weighted. The earlier protocol used a scoring function parameterized for protein-protein interfaces. In addition, the score of the protein was decomposed into a "binding" score (intermolecular interactions between chains; A-B in Figure 3-1) and a "folding" score (intramolecular interactions, sum of A and B in Figure 3-1). Sequences were allowed to contribute to the calculated frequencies if their binding and folding scores fell below given cutoffs determined using the wild-type sequence scores. The generalized protocol presented here uses the Rosetta 3.2 default all-atom scoring function, including an increased histidine reference energy (see Methods), was designed to work without having a wild-type sequence, and all scores were normalized to the lowest fitness found for a given backbone. Additionally, instead of using two separate scores for weighting, a linear combination of the binding and folding scores was used. Finally, instead of using hard cutoffs, Boltzmann weighting was used to weight the contribution of a given sequence to the final position weight matrix.

The predictions from the generalized protocol were similar to the previous method[16] for the 16 residue positions in which a computationally selected library was described[16] (Figure 3-6, all positions shown in Figure 3-7). Using the residue-specific size of the library as previously defined (Table 2 in reference [16]), the Rosetta 3 protocol has one fewer false negative (and by definition of the fixed-size library one fewer false positive) than the Rosetta 2 protocol. These results thus highlight the transferability of the parameters and protocol used here, while providing a more general prediction framework.

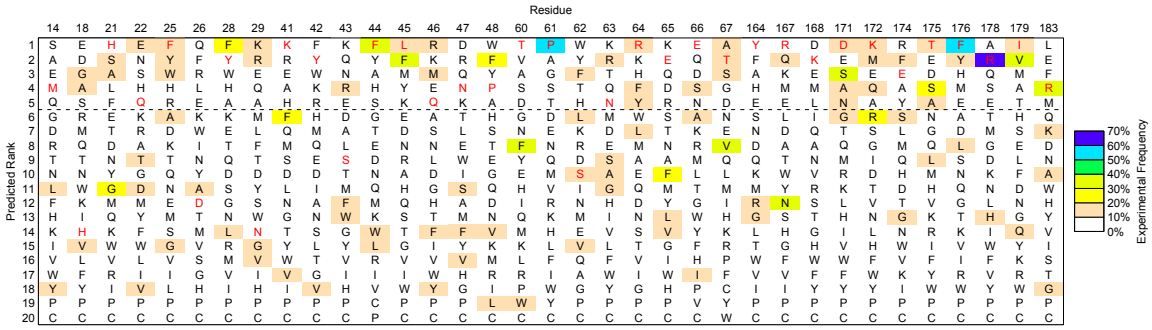
Figure 3-6. hGH/hGHR interface tolerance prediction



The generalized Rosetta 3 protocol described here was applied to rank human growth hormone (hGH) amino acids by computationally predicted frequency. The residue positions shown and their ordering are taken from previously published results using the Rosetta 2 protocol (Humphris & Kortemme, Table 2[16]). Wild type residues, which were used in protein ensemble generation, are shown in red. For each position, an average of 59% of the amino acids observed in phage display (≥ 10% experimental frequency) are predicted within the top five computationally ranked amino acids (above dashed line). Overall performance

was similar to previous results of the Rosetta 2 protocol. Amino acids (other than wild-type) included in the computationally selected library from the Rosetta 2 protocol are indicated with a star. If the same number of amino acids at each position is used as defined in the computational library in[16], Table 2, the Rosetta 3 protocol misses two frequently observed amino acids included by Rosetta 2 (V67 and L176). Conversely, the Rosetta 2 protocol misses three frequently observed amino acids included by Rosetta 3 (S21, A21, and E22). Both protocols share similar false positive predictions. However, the Rosetta 3 histidine reference energy reweighting (see Methods) eliminates 6 out of 8 histidine false positives (H*).

Figure 3-7. hGH/hGHR interface tolerance prediction for all residues



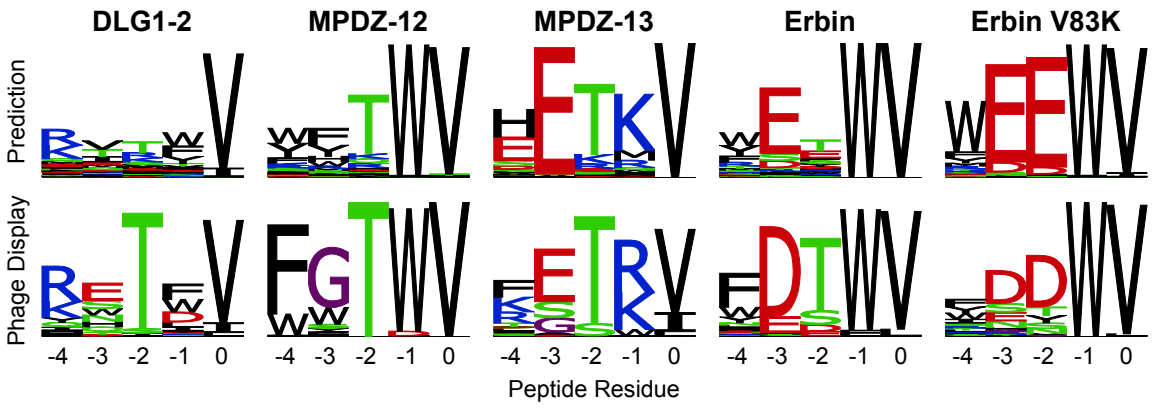
Human growth hormone (hGH) amino acids are ranked by computationally predicted frequency using the generalized Rosetta 3 protocol described here. Wild type residues, which were used in protein ensemble generation, are shown in red. (Representation and color coding is as shown in Figure 3-6).

PDZ/Peptide Interaction

The third test dataset contains peptide sequences selected by phage display to bind to PDZ domains[14]. To determine if the generalized protocol and scripts described here produce similar results to those previously published on the PDZ-peptide dataset[110], we performed 5 representative PDZ/peptide interface specificity predictions. (For details on methodological differences between the published and current protocols, see the Methods

section.) Computational and experimental sequence logos are shown in Figure 3-8. The correspondence to experiment is overall similar to the previous protocol[110], with the largest difference observed in the absolute frequency of amino acids, as shown in Table 3-1. The primary changes are reductions in the preferences for R/K at position -4 and T at position -2 for the DLG1-2 PDZ domain, as well as the preference for T at position -2 for the Erbin PDZ domain. These differences likely come from the restoration of environment dependent hydrogen bonds in the current protocol, which weakens hydrogen bonds in solvent exposed areas.

Figure 3-8. PDZ/peptide interface tolerance predictions



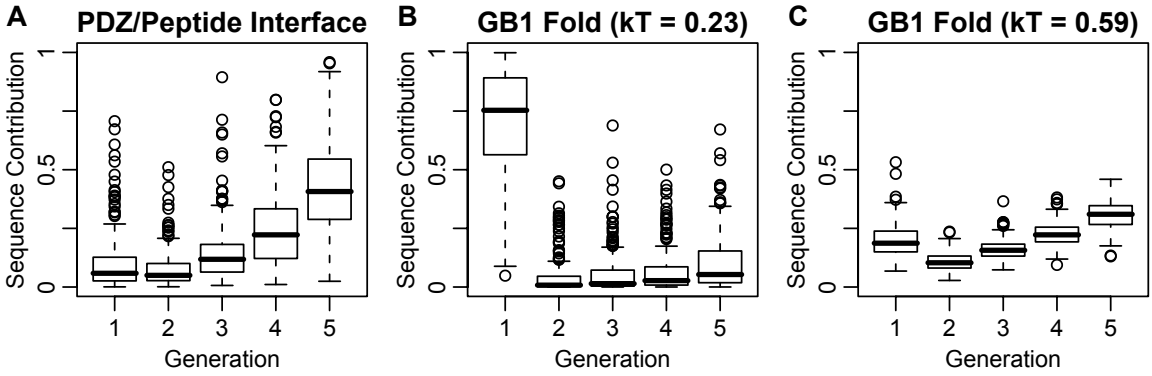
Shown are 5 representative examples of predictions with the generalized protocol, compared to experimental data from phage display. The Erbin V83K interface prediction involved making the indicated point mutant (V83K) to the PDZ domain prior to backrub ensemble generation (an example of a "premutated" position).

Sampling Efficiency and Boltzmann Factors

From an algorithmic point of view, one of the primary differences between the protocols presented here for interface vs. fold stabilization is whether the fitness function is reweighted (interfaces) or not reweighted (fold stabilization) after side chain packing.

The first generation of the genetic algorithm consists of random sequences as well as the sequence with the best raw score as defined by the non-reweighted fitness function. Because the reweighting changes the fitness function, this optimized sequence often does not score as well relative to sequences that evolve in later generations in the case of interface stabilization. This leads to a lower overall contribution of the first generation sequences to the final PWM (Figure 3-9A). However, the reweighted fitness quickly improve, leading to a median fifth generation PWM contribution of 40%.

Figure 3-9. Sequence contribution by genetic algorithm generation



Sequences from later genetic algorithm generations contribute more in interface design prediction than in protein stability design prediction. The total Boltzmann weights in the final PWM for the new sequences sampled in each generation were calculated. The distribution of contributions for each generation across the 200 simulations (one simulation for each backbone in the backrub ensemble) is shown. Boxes span from the first quartile to the third quartile, with the line indicating the median. Whiskers extend to the most extreme data point within 1.5 times the interquartile range of the box. Circles show data points beyond that limit. **(A)** Because the fitness function used for protein-protein interfaces (here shown for a complex between the second PDZ domain of DLG1 and peptides) is different from the fitness function used for optimization of side chain packing, the genetic algorithm is important for enriching the population in sequences predicted to be better binders. **(B)** For optimization of protein fold stability (designing positions

in the GB1 core), the initial full protein design phase is very effective at finding a low energy sequence, which dominates the contribution to the position weight matrix (PWM) when the same Boltzmann factor ($kT = 0.23$) is used. **(C)** When the Boltzmann factor is optimized to minimize the average absolute difference between experiment and computation ($kT = 0.59$), the contribution of the later generations increases significantly.

By contrast, when optimizing sequences to preserve fold stability, the raw Rosetta score for optimization of intramolecular side chain packing and reweighted Rosetta fitness score for Boltzmann weighting are identical. Using the same Boltzmann factor as for interface prediction, the first generation overwhelmingly dominates the contribution to the final PWM (Figure 3-9B). The primary contribution of the first generation comes from the sequence that showed the best overall side chain packing. It typically takes several generations for new sequences to be discovered that score close enough to that sequence to make a significant contribution to the PWM. This imbalance may be partially an artifact of the Boltzmann factor that was not previously assessed for prediction of tolerated sequences for fold stability. The Boltzmann factor increases from 0.23 (taken from the PDZ-peptide study) to 0.59 if it is reoptimized to produce the highest similarity between the predicted and experimental PWMs (Figure 3-9C). Here, the contributions of the different generations are more balanced. Of note, this change in Boltzmann factor does not significantly change the sequence ranks (data not shown), but does make the computational predictions match the relative flatness of the experimental PWM better. If this protocol is applied to other monomeric systems where absolute frequencies matter, the Boltzmann factor of 0.59 may provide a more useful starting point.

Another algorithmic consideration is the influence of introducing backbone flexibility into the prediction method. To determine the effect backbone flexibility had in our simulations, we repeated the predictions without backrub moves and computed overall performance (Table 1-1). The results with the heterogeneous test set used here mirror the previous finding for PDZ-peptide interactions[110], namely that backbone flexibility improves predictions by most metrics. The only place where the fixed backbone method showed better performance was the Fraction Top 5 scores for the GB1 dataset. Overall prediction performance improved with an increasing number of backbones until convergence was reached at about 20 backbones (Figure 3-3) for the three datasets tested here.

Table 3-2. Summary of fixed backbone prediction performance

	Proteins	Residue positions	Bits of information		Fraction Top 5 (%)	AAD (%)	AUC	Rank Top
			Phage display	Predicted				
GB1 (*kT*=0.23)	1	6	1.58	4.25	70.8	7.77	0.73	6.33
GB1 (*kT*=0.66)	1	6	1.58	1.00	76.4	4.53	0.75	6.33
hGH/hGHR[1]	1	16	1.19	3.69	52.8	7.51	0.68	7.38
hGH/hGHR[2]	1	35	0.89	3.56	42.0	7.81	0.62	7.86
PDZ/Peptide	5	25	3.11	2.82	81.0	5.61	0.84	3.36

[1] 16 designed hGH amino acid positions as defined in[16] and shown in Figure 3-6.
[2] All designed hGH amino acid positions shown in Figure 3-7.

As a fraction of the dynamic range of the performance metrics, the predicted bits of information, AAD, AUC, and Rank Top metrics (averaged over all datasets) are better with backrub sampling (see Table 3-1) by 9.4%, 9.1%, 1.6%, and 1.1%, respectively. The only performance metric that was better (by 3.8%) without backrub sampling was Fraction Top 5. This improvement came primarily from the GB1 dataset. Fraction Top 5 was found to be the most variable performance metric across replicated predictions (Table 3-1).

A final point of comparison can be made to a naïve model, in which residues with similar chemical properties to those in the input structure are given equal weight in a

predicted PWM. Using the unmodified *kT* of 0.23, the prediction method presented here also outperforms the naïve model by most performance metrics (Table 3-3).

Table 3-3. Summary of naïve model prediction performance

	Proteins	Residue positions	Bits of information		Fraction Top 5 (%)[3]	AAD (%)	AUC	Rank Top
			Phage display	Predicted				
GB1	1	6	1.58	2.79	52.8	6.39	0.71	9.00
hGH/hGHR[1]	1	16	1.19	2.67	44.3	6.71	0.67	12.63
hGH/hGHR[2]	1	35	0.89	2.67	27.9	7.52	0.57	14.52
PDZ/Peptide	5	25	3.11	2.72	68.7	6.61	0.79	7.96

[1] 16 designed hGH amino acid positions as defined in[16] and shown in Figure 3-6.

[2] All designed hGH amino acid positions shown in Figure 3-7.

[3] Naïve predictions, which rank up to 4 amino acids, do artificially poorly with Fraction Top 5.

Naïve predictions were constructed by generating position weight matrices in which the PDB amino acid and amino acids in its similarity group were given equal weight, and all other amino acids given zero weight. The similarity groups were as follows: DENQ, RKH, LIVM, FYW, PAG, ST, and C[16]. All metrics for the performance of the naïve model (Fraction Top 5, AAD, AUC and Rank Top) were worse than those shown in Table 3-1, with the exception of the hGH/hGHR AAD for the 16-residue set. In addition to performing better than a naïve model, the method described in the main text also does better than random, as evidenced by the area under ROC curves (AUC) being greater than random (0.5) for all datasets (Table 1).

Discussion

One of the key assumptions made in the method described here is that the backbone structures generated with the input sequence will adequately sample backbones that will accommodate other amino acid sequences. While we have shown here and in previous work that incorporation of backbone flexibility improves prediction of tolerated sequence space[16,110], side chain order parameters[108], and residual dipolar couplings[70], this and previous studies indicate that there are limitations to that assumption. To adequately sample both backbone and sequence space, variants of simultaneous or iterative sampling

strategies[59,69] are likely necessary. We have made initial attempts at adding iteration to this method and others, but found that the simulations end up trapped in local minima of sequence space, with the backbones retaining the bias towards the sequence that they start with. Often, the solution to limited sampling is to increase the simulation temperature, which can be done when the backbone is fixed. However, when the backbone is flexible, increasing the temperature can lead to protein unfolding and sampling of unproductive regions of sequence space. Application of constraints, restraints, or other sampling methods may be required to overcome that problem.

While the uses of this protocol to date have been limited to protein-protein interfaces and monomeric protein folds, there are several other applications that it can also be generalized to. For instance, this method could be leveraged in prediction of the amino acid sequences that will bind to a small molecule substrate, cofactor, or inhibitor, as well as for protein-DNA and protein-RNA interfaces. Another potential application would be stabilizing particular conformations of loops or domains. For that purpose, one could place the backbone into a preferred conformation at the outset, and then upweight the interaction energies between the residues that are desired to interact. While many design problems can be described using a single state, adaptation of the code described here could be used to generate a set of sequences that satisfy multiple states or constraints[10,120,121].

References

1. Davis, I. W., Arendall, W. B., Richardson, D. C., & Richardson, J. S. (2006) The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure* 14, 265-274.
2. Go, N. & Scheraga, H. A. (1970) Ring Closure and Local Conformational Deformations of Chain Molecules. *Macromolecules* 3, 178-187.
3. Bruccoleri, R. E. & Karplus, M. (1985) Chain closure with bond angle variations. *Macromolecules* 18, 2767-2773.
4. Dinner, A. R. (2000) Local deformations of polymers with nonplanar rigid main-chain internal coordinates. *J Comput Chem* 21, 1132-1144.
5. Ulmschneider, J. P. & Jorgensen, W. L. (2003) Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias. *J Chem Phys* 118, 4261-4271.
6. Coutsias, E. A., Seok, C., Jacobson, M. P., & Dill, K. A. (2004) A kinematic view of loop closure. *J Comput Chem* 25, 510-528.
7. Betancourt, M. R. (2005) Efficient Monte Carlo trial moves for polypeptide simulations. *J Chem Phys* 123, 174905-174905.
8. Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L., & Baker, D. (2004) Computational redesign of protein-protein interaction specificity. *Nat Struct Mol Biol* 11, 371-379.
9. Joachimiak, L. A., Kortemme, T., Stoddard, B. L., & Baker, D. (2006) Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. *J Mol Biol* 361, 195-208.
10. Humphris, E. L. & Kortemme, T. (2007) Design of multi-specificity in protein interfaces. *PLoS Comput Biol* 3, e164.
11. Groban, E. S., Narayanan, A., & Jacobson, M. P. (2006) Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PLoS Comput Biol* 2.
12. Mandell, D. J., Chorny, I., Groban, E. S., Wong, S. E., Levine, E., Rapp, C. S., & Jacobson, M. P. (2007) Strengths of hydrogen bonds involving phosphorylated amino acid side chains. *J Am Chem Soc* 129, 820-827.
13. Yeh, B. J., Rutigliano, R. J., Deb, A., Bar-Sagi, D., & Lim, W. A. (2007) Rewiring cellular morphology pathways with synthetic guanine nucleotide exchange factors. *Nature* 447, 596-600.
14. Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J. H., Reva, B., Held, H. A., Appleton, B. A., Evangelista, M., Wu, Y., Xin, X., Chan, A. C., Seshagiri, S., Lasky, L. A., Sander, C., Boone, C., Bader, G. D., & Sidhu, S. S. (2008) A specificity map for the PDZ domain family. *PLoS Biol* 6, e239.

15. Ernst, A., Sazinsky, S. L., Hui, S., Currell, B., Dharsee, M., Seshagiri, S., Bader, G. D., & Sidhu, S. S. (2009) Rapid evolution of functional complexity in a domain family. *Sci Signal* 2, ra50.

16. Humphris, E. L. & Kortemme, T. (2008) Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design. *Structure* 16, 1777-1788.

17. 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, 775-791.

18. Fernández-Recio, J., Totrov, M., & Abagyan, R. (2003) ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins* 52, 113-117.

19. Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., & Baker, D. (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* 331, 281-299.

20. Dahiyat, B. I. & Mayo, S. L. (1997) De novo protein design: fully automated sequence selection. *Science* 278, 82-87.

21. Dantas, G., Kuhlman, B., Callender, D., Wong, M., & Baker, D. (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J Mol Biol* 332, 449-460.

22. Ashworth, J., Havranek, J. J., Duarte, C. M., Sussman, D., Monnat, R. J., Stoddard, B. L., & Baker, D. (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* 441, 656-659.

23. Swain, J. F. & Gierasch, L. M. (2006) The changing landscape of protein allostery. *Curr Opin Struct Biol* 16, 102-108.

24. Li, Y., Li, H., Yang, F., Smith-Gill, S. J., & Mariuzza, R. A. (2003) X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nat Struct Biol* 10, 482-488.

25. Li, Z. & Scheraga, H. A. (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci U S A* 84, 6611-6615.

26. Abagyan, R., Totrov, M., & Kuznetsov, D. (1994) ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J Comput Chem* 15, 488-506.

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

28. Simons, K. T., Kooperberg, C., Huang, E., & Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol* 268, 209-225.

29. Rohl, C. A., Strauss, C. E., Chivian, D., & Baker, D. (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* 55, 656-677.

30. Desjarlais, J. R. & Handel, T. M. (1999) Side-chain and backbone flexibility in protein core design. *J Mol Biol* 290, 305-318.

31. Kuhlman, B. & Baker, D. (2000) Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A* 97, 10383-10388.

32. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T., & Kim, P. S. (1998) High-resolution protein design with backbone freedom. *Science* 282, 1462-1467.

33. 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, 1099-1117.

34. Wick, C. D. & Siepmann, J. I. (2000) Self-adapting fixed-end-point configurational-bias Monte Carlo method for the regrowth of interior segments of chain molecules with strong intramolecular interactions. *Macromolecules* 33, 7207--7218.

35. Canutescu, A. A. & Dunbrack, R. L. (2003) Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci* 12, 963-972.

36. 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, 351-367.

37. Cahill, M., Cahill, S., & Cahill, K. (2002) Proteins wriggle. *Biophys J* 82, 2665-2670.

38. Meiler, J. & Baker, D. (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* 65, 538-548.

39. Bradley, P., Misura, K. M., & Baker, D. (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309, 1868-1871.

40. Carr, D. B., Littlefield, R. J., Nicholson, W. L., & Littlefield, J. S. (1987) Scatterplot Matrix Techniques for Large N. *Journal of the American Statistical Association* 82, 424-436.

41. Holm, L. & Sander, C. (1991) Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J Mol Biol* 218, 183-194.

42. Dunbrack, R. L. & Karplus, M. (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* 230, 543-574.

43. Shenkin, P. S., Farid, H., & Fetrow, J. S. (1996) Prediction and evaluation of side-chain conformations for protein backbone structures. *Proteins* 26, 323-352.

44. De Maeyer, M., Desmet, J., & Lasters, I. (1997) All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold Des* 2, 53-66.

45. Formaneck, M. S., Ma, L., & Cui, Q. (2006) Reconciling the "old" and "new" views of protein allostery: a molecular simulation study of chemotaxis Y protein (CheY). *Proteins* 63, 846-867.

46. Bordner, A. J. & Abagyan, R. A. (2004) Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins* 57, 400-413.

47. Lolis, E. & Petsko, G. A. (1990) Crystallographic analysis of the complex between triosephosphate isomerase and 2-phosphoglycolate at 2.5-A resolution: implications for catalysis. *Biochemistry* 29, 6619-6625.

48. Lolis, E., Alber, T., Davenport, R. C., Rose, D., Hartman, F. C., & Petsko, G. A. (1990) Structure of yeast triosephosphate isomerase at 1.9-A resolution. *Biochemistry* 29, 6609-6618.

49. Joseph, D., Petsko, G. A., & Karplus, M. (1990) Anatomy of a conformational change: hinged "lid" motion of the triosephosphate isomerase loop. *Science* 249, 1425-1428.

50. Derreumaux, P. & Schlick, T. (1998) The loop opening/closing motion of the enzyme triosephosphate isomerase. *Biophys J* 74, 72-81.

51. Dodd, L. R., Boone, T. D., & Theodorou, D. N. (1993) A concerted rotation algorithm for atomistic Monte-Carlo simulation of polymer melts and glasses. *Molecular Physics* 78, 961-996.

52. Guerois, R., Nielsen, J. E., & Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* 320, 369-387.

53. Kortemme, T. & Baker, D. (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* 99, 14116-14121.

54. 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, 1239-1259.

55. Yin, S., Ding, F., & Dokholyan, N. V. (2007) Modeling Backbone Flexibility Improves Protein Stability Estimation. *Structure* 15, 1567-1576.

56. Dunbrack, R. L. & Cohen, F. E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* 6, 1661-1681.

57. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc* 117, 5179-5197.

58. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., & Karplus, M. (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J Phys Chem B* 102, 3586-3616.

59. 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, 1364-1368.

60. Lazaridis, T. & Karplus, M. (1999) Effective energy function for proteins in solution. *Proteins* 35, 133--152.

61. Appleton, B. A., Zhang, Y., Wu, P., Yin, J. P., Hunziker, W., Skelton, N. J., Sidhu, S. S., & Wiesmann, C. (2006) Comparative structural analysis of the Erbin PDZ domain and the first PDZ domain of ZO-1. Insights into determinants of PDZ domain specificity. *J Biol Chem* 281, 22312-22320.

62. Fitzgerald, J. E., Jha, A. K., Sosnick, T. R., & Freed, K. F. (2007) Polypeptide motions are dominated by peptide group oscillations resulting from dihedral angle correlations between nearest neighbors. *Biochemistry* 46, 669-682.

63. Hayward, S. (2001) Peptide-plane flipping in proteins. *Protein Sci* 10, 2219-2227.

64. Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-4680.

65. Schuster-Böckler, B. & Bateman, A. (2008) Protein interactions in human genetic diseases. *Genome Biol* 9.

66. Mandell, D. J. & Kortemme, T. (2009) Computer-aided design of functional protein interactions. *Nat Chem Biol* 5, 797-807.

67. Shifman, J. M. & Mayo, S. L. (2003) Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc Natl Acad Sci U S A* 100, 13274-13279.

68. Grigoryan, G., Reinke, A. W., & Keating, A. E. (2009) Design of protein-interaction specificity gives selective bZIP-binding peptides. *Nature* 458, 859-864.

69. Saunders, C. T. & Baker, D. (2005) Recapitulation of protein family divergence using flexible backbone protein design. *J Mol Biol* 346, 631-644.

70. Friedland, G. D., Lakomek, N. A., Griesinger, C., Meiler, J., & Kortemme, T. (2009) A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family. *PLoS Comput Biol* 5, e1000393.

71. 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.

72. Wollacott, A. M. & Desjarlais, J. R. (2001) Virtual interaction profiles of proteins. *J Mol Biol* 313, 317-342.

73. Bordner, A. J. & Abagyan, R. (2006) Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Proteins* 63, 512-526.

74. Hou, T., Chen, K., McLaughlin, W. A., Lu, B., & Wang, W. (2006) Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. *PLoS Comput Biol* 2, 46-55.

75. Fernandez-Ballester, G., Beltrao, P., Gonzalez, J. M., Song, Y. H., Wilmanns, M., Valencia, A., & Serrano, L. (2009) Structure Based Prediction of the S. cerevisiae SH3-Ligand Interactions. *J Mol Biol* 388, 902--916.

76. Encinar, J. A., Fernandez-Ballester, G., Sanchez, I. E., Hurtado-Gomez, E., Stricher, F., Beltrao, P., & Serrano, L. (2009) ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics* 25, 2418-2424.

77. Kiel, C., Wohlgemuth, S., Rousseau, F., Schymkowitz, J., Ferkinghoff-Borg, J., Wittinghofer, F., & Serrano, L. (2005) Recognizing and defining true Ras binding domains II: In silico prediction based on homology modelling and energy calculations. *J Mol Biol* 348, 759-775.

78. Songyang, Z., Fanning, A. S., Fu, C., Xu, J., Marfatia, S. M., Chishti, A. H., Crompton, A., Chan, A. C., Anderson, J. M., & Cantley, L. C. (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science* 275, 73-77.

79. Wiedemann, U., Boisguerin, P., Leben, R., Leitner, D., Krause, G., Moelling, K., Volkmer-Engert, R., & Oschkinat, H. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J Mol Biol* 343, 703-718.

80. Gisler, S. M., Kittanakom, S., Fuster, D., Wong, V., Bertic, M., Radanovic, T., Hall, R. A., Murer, H., Biber, J., Markovich, D., Moe, O. W., & Stagljar, I. (2008) Monitoring protein-protein interactions between the mammalian integral membrane transporters and PDZ-interacting partners using a modified split-ubiquitin membrane yeast two-hybrid system. *Mol Cell Proteomics* 7, 1362-1377.

81. Stiffler, M. A., Chen, J. R., Grantcharova, V. P., Lei, Y., Fuchs, D., Allen, J. E., Zaslavskaia, L. A., & MacBeath, G. (2007) PDZ domain binding selectivity is optimized across the mouse proteome. *Science* 317, 364-369.

82. Zhang, Y., Yeh, S., Appleton, B. A., Held, H. A., Kausalya, P. J., Phua, D. C., Wong, W. L., Lasky, L. A., Wiesmann, C., Hunziker, W., & Sidhu, S. S. (2006) Convergent and divergent ligand specificity among PDZ domains of the LAP and zonula occludens (ZO) families. *J Biol Chem* 281, 22299-22311.

83. Niv, M. Y. & Weinstein, H. (2005) A flexible docking procedure for the exploration of peptide binding selectivity to known structures and homology models of PDZ domains. *J Am Chem Soc* 127, 14072-14079.

84. Raveh, B., London, N., & Schueler-Furman, O. (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins: Structure, Function, and Bioinformatics* .

85. 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, 742-756.

86. Hu, X., Wang, H., Ke, H., & Kuhlman, B. (2007) High-resolution design of a protein loop. *Proc Natl Acad Sci U S A* 104, 17668-17673.

87. Dunbrack, R. L. (2002) Rotamer libraries in the 21st century. *Curr Opin Struct Biol* 12, 431-440.

88. Abagyan, R. & Totrov, M. (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* 235, 983-1002.

89. Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995) A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing* 16, 1190-1208.

90. Friedland, G. D. & Kortemme, T. (2010) Designing ensembles in conformational and sequence space to characterize and engineer proteins. *Curr Opin Struct Biol* 20, 377-384.

91. Pokala, N. & Handel, T. M. (2001) Review: protein design--where we were, where we are, where we're going. *J Struct Biol* 134, 269-281.

92. Frey, K. M., Georgiev, I., Donald, B. R., & Anderson, A. C. (2010) Predicting resistance mutations using protein design algorithms. *Proc Natl Acad Sci U S A* 107, 13707-13712.

93. Bloom, J. D. & Arnold, F. H. (2009) In the light of directed evolution: pathways of adaptive protein evolution. *Proc Natl Acad Sci U S A* 106 Suppl 1, 9995-10000.

94. Marini, N. J., Thomas, P. D., & Rine, J. (2010) The use of orthologous sequences to predict the impact of amino acid substitutions on protein function. *PLoS Genet* 6.

95. Distefano, M. D., Zhong, A., & Cochran, A. G. (2002) Quantifying beta-sheet stability by phage display. *J Mol Biol* 322, 179-188.

96. Kotz, J. D., Bond, C. J., & Cochran, A. G. (2004) Phage-display as a tool for quantifying protein stability determinants. *Eur J Biochem* 271, 1623-1629.

97. Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., & Fields, S. (2010) High-resolution mapping of protein sequence-function relationships. *Nat Methods* 7, 741-746.

98. Smith, G. P. (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* 228, 1315-1317.

99. Fuh, G., Pisabarro, M. T., Li, Y., Quan, C., Lasky, L. A., & Sidhu, S. S. (2000) Analysis of PDZ domain-ligand interactions using carboxyl-terminal phage display. *J Biol Chem* 275, 21486-21491.

100. Laura, R. P., Witt, A. S., Held, H. A., Gerstner, R., Deshayes, K., Koehler, M. F., Kosik, K. S., Sidhu, S. S., & Lasky, L. A. (2002) The Erbin PDZ domain binds with high affinity and specificity to the carboxyl termini of delta-catenin and ARVCF. *J Biol Chem* 277, 12906-12914.

101. Pál, G., Kouadio, J. L., Artis, D. R., Kossiakoff, A. A., & Sidhu, S. S. (2006) Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. *J Biol Chem* 281, 22378-22385.

102. Mandell, D. J. & Kortemme, T. (2009) Backbone flexibility in computational protein design. *Curr Opin Biotechnol* .

103. 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, 2804-2813.

104. Ding, F. & Dokholyan, N. V. (2006) Emergence of protein fold families through rational design.. *PLoS Comput Biol* 2, e85.

105. Georgiev, I. & Donald, B. R. (2007) Dead-end elimination with backbone flexibility. *Bioinformatics* 23, 185-194.

106. Georgiev, I., Lilien, R. H., & Donald, B. R. (2008) The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles.. *J Comput Chem* 29, 1527--1542.

107. Ambroggio, X. I. & Kuhlman, B. (2006) Design of protein conformational switches. *Curr Opin Struct Biol* 16, 525-530.

108. 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, 757-774.

109. 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.

110. Smith, C. A. & Kortemme, T. (2010) Structure-Based Prediction of the Peptide Sequence Space Recognized by Natural and Synthetic PDZ Domains. *J Mol Biol* 402, 460-474.

111. Leaver-Fay, A., Kuhlman, B., & Snoeyink, J. (2005) An adaptive dynamic programming algorithm for the side chain placement problem. *Pac Symp Biocomput* , 16-27.

112. Schmidt, H. L., Sperling, L. J., Gao, Y. G., Wylie, B. J., Boettcher, J. M., Wilson, S. R., & Rienstra, C. M. (2007) Crystal polymorphism of protein GB1 examined by solid-state NMR spectroscopy and X-ray diffraction. *J Phys Chem B* 111, 14362-14369.

113. Clackson, T., Ultsch, M. H., Wells, J. A., & de Vos, A. M. (1998) Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J Mol Biol* 277, 1111-1128.

114. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufmann, K. W., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popovic, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., & Bradley, P. (2011) ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods in Enzymology* 487, 545-574.

115. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953) Equation of State Calculations by Fast Computing Machines. *J Chem Phys* 21, 1087-1092.

116. Voigt, C. A., Gordon, D. B., & Mayo, S. L. (2000) Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* 299, 789-803.

117. Ollikainen, N., Sentovich, E., Coelho, C., Kuehlmann, A., & Kortemme, T. (2009) SAT-based protein design. *Proceedings of the 2009 IEEE/ACM International Conference on Computer-Aided Design (ICCAD 2009)* , 128-35.

118. R Development Core Team (2011) R: A Language and Environment for Statistical Computing. .

119. Sauer-Eriksson, A. E., Kleywegt, G. J., Uhlén, M., & Jones, T. A. (1995) Crystal structure of the C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG. *Structure* 3, 265-278.

120. Havranek, J. J. & Harbury, P. B. (2003) Automated design of specificity in molecular recognition. *Nat Struct Biol* 10, 45-52.

121. Ambroggio, X. I. & Kuhlman, B. (2006) Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J Am Chem Soc* 128, 1154-1161.

Publishing Agreement

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

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

Author Signature

8/8/11
Date