Improving The Energy Function Used In Rosetta For Protein Structure Prediction And Design

Patrick Conway

A dissertation submitted in
partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2014

Reading Committee:
David Baker, Chair
Phil Bradley
Frank DiMaio

Program Authorized to Offer Degree:
Biochemistry

University of Washington

Abstract

Improving The Energy Function Used In Rosetta For Protein Structure Prediction And Design

Patrick Conway

Chair of the Supervisory Committee:
Professor David Baker
Department of Biochemistry

Protein structure prediction and design relies on conformational sampling and scoring to uncover a global energy minimum. While it is critical to have an accurate energy function for efficient and precise modeling, the Rosetta energy function trades the accuracy of quantum mechanical modeling for a substantially faster combination of approximated score terms. While the Rosetta energy function has been successfully applied to problems such as *ab initio* folding, novel fold design, catalytic enzyme design, and numerous other projects, recent work utilizing intensive sampling algorithms have identified instances where native structures are not at the global energy minima. Here, we address two known issues in the energy function: nonideal bond geometry modeling and double counting between the statistical sidechain torsion potential and physical score terms. We also introduce rigorous and sensitive methods designed to quantify energy function performance and uncover errors.

Table of Contents

Chapter 1

Introduction: Approaches to Conformational Sampling and Scoring

Life on Earth manifests in an incredible variety of configurations. Even more amazing is that it all can be described by chains of sugar carrying just four different types of nucleobases. These nucleobases in turn dictate how sets of twenty standard building blocks, amino acids, are to be linked to form a polypeptide chain known as a protein. Diversity emerges from the near-infinite number of ways these amino acids can be combined. Function emerges from the final complex three-dimensional conformation adopted by the protein. According to the thermodynamic hypothesis, a given amino acid sequence will always produce the same protein structure due to a descent along an energy gradient to the lowest energy conformation[1].

Protein conformations can be thought to populate an energy landscape where there is one deep energy minimum that is the native structure. In protein structure prediction, a set of thermodynamic equations is used with a conformational sampling method to evaluate candidate conformations until the lowest energy structure is discovered. Implementation details vary greatly depending on requirements of accuracy, speed, and the nature of the experiment. The remainder of this chapter will cover different approaches to sampling and scoring.

Conformational Sampling

To illustrate the challenge of conformational sampling, consider Levinthal's paradox[2]. A protein in an extended conformation has an enormous number of possible conformations due to a very large number of degrees of freedom. Even if backbone bond torsion angles were limited to three values across a protein of 100 residues, there are still 3^{198} possible conformations. Protein folding occurs quickly enough to conclude that not all conformations are sampled and suggests a funnel-like energy landscape.

Thus, the major challenge facing conformational sampling is how to reduce the number of sampled conformations without losing the ability to discover the global minimum. As a first step, some approaches reduce the number of degrees of freedom as suggested by Levinthal. Although what constitutes the necessary allowed degrees of freedom is a matter of debate, limiting degrees of freedom to the torsional ones is common. Furthermore, the allowed torsional degrees of freedom can be reduced to only backbone phi and psi torsions. This is commonly accomplished using an internal coordinate representation of atomic coordinates where each atom is described by its torsional relationship to other atoms, as well as length, angle, and chemical identity. This is in contrast with Cartesian representation where every atom is an additional degree of freedom on each axis of three-dimensional space. While using Cartesian coordinates might be more expensive than simplified internal coordinates, it has the advantage of providing the most conformational flexibility. Another approach to reducing degrees of freedom in

the system lies in eliminating solvent modeling and going with an implicit representation of bulk solvent behavior.

Once a structural model has been determined, the next approach is to perform a sampling algorithm. In the early days of computational protein modeling, it was sufficient to use a straightforward minimization algorithm to find the conformational minimum of a potential energy function[3]. However, energy functions have many local minima and this required the global minimum to be directly accessible from the starting point of minimization. Energy minimization could be used to refine a structure from electron density data, but not to fold a protein *ab initio* from an extended conformation.

To increase the radius of convergence of protein folding algorithm, it has to be able to escape local minima. In 1975, Levitt and Warshel performed alternating cycles of minimization and normal mode thermalization to fold a simplified representation of the bovine pancreatic trypsin inhibitor[4]. Normal mode thermalization explores changes in protein coordinates producing the most significant movements and presents a new conformation to resume minimization from.

About the same time, Martin Karplus was the first to molecular dynamics on protein structures, where the potential energy forcefield is combined with kinetic energy modeled by Newton's laws of motion to model protein dynamics[5]. Variations on the full protein single-trajectory sampling include constraining regions while fully sampling the rest, steering trajectories with dynamic constraints, and umbrella sampling to uncover an ensemble of conformations. A range of integration

algorithms for numerically solving the equations of motion have been developed with an emphasis on limiting accumulation of numerical errors and ensuring energy conservation. While molecular dynamics could potentially arrive at a global minimum from an extended conformation, it is better suited for exploring short timescales due to the huge computational requirements to simulate the femtosecond timescale of bond vibrations.

A fourth method for searching protein conformation space, introduced by Li and Scheraga, relies on Metropolis Monte Carlo[6]. A stochastic sampling method, the Monte Carlo algorithm relies on many repeated runs to produce a probability distribution modeling potential conformations of the given sequence. Within an individual Monte Carlo trajectory, sampling starts with gradient-based minimization of the starting structure, and then samples a random structural perturbation before another minimization. Li and Scheraga sampled single dihedral changes, but other approaches may make more drastic perturbations. If the energy of the next minimization is accepted by the Metropolis criterion, the cycle begins anew. The Metropolis criterion introduces randomness while relying on the log-linear Boltzmann distribution

$$F(\text{state}) \propto e^{-\frac{E}{kT}}$$

to bias sampling toward more energetically favorable conformations. As a result, the same starting point produces a distribution of structures that can be processed to predict a native conformation – usually using a combination of energy function scoring and clustering. Monte Carlo sampling has the advantage over normal mode thermalization since sampling transitions are weighted rather than uniform.

To further reduce conformational sampling space, the Rosetta software package uses a refinement of the Metropolis Monte Carlo algorithm where initial rounds of structure perturbation are performed by sampling homologous backbone conformations of three or nine residues in length[7]. The idea is that local interactions biases secondary structure formation while long-range interactions stabilize them and determine tertiary structures. Final stages limit perturbations to individual backbone torsions that don't disturb the global fold. In addition to backbone conformation searches, Rosetta performs Monte Carlo sampling of sidechain torsions where degrees of freedom are restricted to highly populated discrete conformations. Throughout the entire trajectory, the temperature of the guiding Boltzmann distribution is reduced in order to perform simulated annealing. By permitting sidechain conformations of any residue type regardless of identity, this sampling approach can be extended to perform protein design.

Energy Function

A protein energy function aims to model system energy of a protein conformation relative to other possible conformations. Conformational sampling leverages the thermodynamic hypothesis to describe a native conformation as the one with the lowest energy of all. Thus, an accurate energy function is extremely important. However, it still needs to be fast enough to allow conformational sampling to quickly assess every perturbation it explores. Scheraga found that the energy minimization step of his Monte Carlo protocol using the ECEPP/2 energy function took up 95% of the compute time[6]. There are three approaches when it

comes to developing an energy function – using physical score terms, using knowledge-based score terms, or a combination of both.

Physical score terms are models of chemical interactions parameterized using experimental data. In general, physical terms used in protein structure prediction are optimized for that scale and tend to rely on pairwise-decomposable approximations rather than detailed multibody quantum mechanical calculations. These approximations are often derived as averages of underlying quantum mechanical calculations on model systems such as small peptides. Popular packages that use physical forcefields include CHARMM, AMBER, and GROMACS[8–10].

Physical terms can be categorized as bonded and unbonded. Bonded terms model chemical bonds between atoms. For instance, in the first protein energy minimization published, atomic bond length and angles were modeled using harmonic constraints parameterized on Ramachandran and Sasisekharan's studies[3]. In place of harmonic constraints, the more expensive Morse potential may be used if vibrational spectra need to be accurately modeled. Implementation of dihedral terms varies, but they typically model discrete barriers using a periodic Fourier function.

Unbonded terms describe interactions between neighboring atoms. The Lennard-Jones potential is commonly used to simultaneously describe Pauli repulsion and van der Waals attraction. Since it is expensive to compute interaction of distant atom pairs that have insignificant van der Waals energies, a distance cutoff is often used. More challenging to model are electrostatic interactions. Using Coulomb's Law is a simple approximation, but not an ideal model since atoms do not

have point charges. Delocalization of electrons in molecules cause charges to be shared among many atoms and be influenced by induction and polarization effects. However, Coulomb's Law is still commonly used due to its simplicity. Alternative models include particle mesh Ewald and the multipole algorithm. Water modeling can be treated as a special case of electrostatics by explicitly modeling water molecules, but it is also common to go with an implicit solvent model. These range in computational complexity from the pairwise-decomposable Lazaridis-Karplus solvation model to one of the more expensive multibody Generalized Born or Poisson-Boltzmann variants (which can also take care of electrostatics modeling)[11,12].

Knowledge based potentials are rooted in reproducing distribution of features found in native proteins. The most common justification for deriving energy in this manner lies with the assumption that the native features obey the Boltzmann distribution and can be compared to a reference state, which then allows for a "potential of mean force" to be used[13]. While there is criticism that a Boltzmann distribution only applies to an ensemble of the same sequence at one temperature, numerous theoretical and empirical justifications have been developed to support this approach[14–17]. Advantages to statistical score terms include their simplicity, accuracy, computational speed, and increased radius of convergence due to a smoother energy landscape. Downsides to this approach include uncertainty about an appropriate reference state and issues with double-counting within an energy function where interactions responsible for a feature distribution is already

modeled by another term, physical or statistical[18,19]. Some examples of knowledge-based scoring functions include RAPDF, DOPE, DFIRE, and GOAP[16,20–22].

Statistical potentials can operate on the residue level or the atomic level. Residue-level representations use residue identity in one-body potentials and additionally Cα, Cβ, or sidechain centroid coordinates in distance-based two-body potentials. One-body residue-level potentials often encode torsional distributions, such as a Ramachandran potential or a sidechain conformation potential. They are also used to encode the propensity of the residue occurrence in a given context such as secondary structures. Two-body residue-level potentials include pairwise Ramachandran, sequence-dependent, and hydrophobic burial potentials. Atomic distance-based potentials are preferred in pure knowledge-based scoring functions. For example, the DFIRE potential computes the number of atomic pairs within a given radius[21]. Variations on the type of atoms considered for the DFIRE potential varied from backbone/Cβ-only to residue-specific heavy atom types, but authors found the latter to give best performance. The GOAP energy function builds off of DFIRE, but also includes angle-dependent terms to control bond geometries[22].

Many software programs also opt to combine physical potentials with statistical potentials. This confers the advantages of both provided care is taken in balancing the contributions. The granularity and generalizability of physical score terms is combined with the speed and simplicity of knowledge-based terms. Knowledge-based terms also provide the ability to compensate for inaccuracies in physical terms caused by missing models (such as quantum mechanical or entropic effects) or incorrect parameterization. The Rosetta energy function uses the

physical-based Lennard-Jones atomic distance potential, Laziridis-Karplus implicit solvation model, and a Coulomb electrostatics term[23]. It also used knowledge-based backbone and sidechain torsional terms, hydrogen bonding model, and a term describing the probability of a given amino acid for a backbone phi-psi value[7]. Recent work on the Rosetta energy function has focused on balancing the energy function, eliminating double-counting, and integrating the Coulomb potential with the hydrogen bond model[19,23].

Even though the Rosetta energy function has continued to evolve, recent publications using powerful sampling algorithms have attributed the energy function as the limiting factor in Rosetta structure prediction by identifying instances where native structures are not at the global energy minimum[24–26]. Furthermore, this energy function – when used for protein design – has a low success rate: influenza binder designs only showed binding on 2 out of 88 total designs[27]. Recent comparisons of crystallized design structures to Rosetta predictions consistently showed inaccurate histidine conformations.

As Rosetta pushes toward prediction and design of small molecule binders and enzymes, where sub-angstrom accuracy is required, the energy function plays an increasingly important role in its performance. Areas to be targeted include coupling solvation, hydrogen bonding, and screened electrostatic interactions, properly accounting for partial charge distribution in pairwise scoring of electrostatics, modeling nonideal bond geometries, and addressing double-counted interactions between knowledge-based torsional potentials and physical potentials.

In this thesis, I focus on addressing the latter two targets – modeling nonideal bond geometries and correcting double counting between the sidechain torsional potential and physical score terms. I also introduce rigorous and sensitive methods designed to quantify energy function performance and uncover errors.

Chapter 2

Relaxation of Backbone Bond Geometry Improves Protein Energy Landscape Modeling

Abstract:

A key issue in macromolecular structure modeling is the granularity of the molecular representation. A fine-grained representation can approximate the actual structure more accurately, but may require many more degrees of freedom than a coarse-grained representation and hence make conformational search more challenging. We investigate this tradeoff between the accuracy and the size of protein conformational search space for two frequently used representations: one with fixed bond angles and lengths and one that has full flexibility. We carry out large-scale explorations of the energy landscapes of 82 protein domains under each model, and find that the introduction of bond angle flexibility significantly increases the average energy gap between native and non-native structures. We also find that incorporating bonded geometry flexibility improves low resolution X-ray crystallographic refinement. These results suggest that backbone bond angle relaxation makes an important contribution to native structure energetics, that current energy functions are sufficiently accurate to capture the energetic gain associated with subtle deformations from chain ideality, and more speculatively, that backbone geometry distortions occur late in protein folding to optimize packing in the native state.

Introduction

Macromolecular structure prediction and design efforts are challenged by the vast size of the conformational space available to macromolecules. For example, even a small 100 amino acid protein has thousands of degrees of freedom. While some approaches model this full parameter space explicitly[28,29], most structure prediction and design efforts attempt to reduce the complexity of the problem by reducing the dimensionality[30–32]. For example, Rosetta[33] typically uses an internal coordinate system with fixed bond angles and lengths, as well as idealized aromatic ring structures. Only torsion angles are allowed to vary, reducing the dimensionality of the above 100-residue protein example from thousands to hundreds. This reduces the size of the search space and makes gradient based minimization more efficient[34].

One potential problem with any reduced dimensionality description is that the accuracy of the representation may be compromised, resulting in inaccurate free energy evaluations. The reduced representation may restrict the molecule from accessing low energy states that require relaxation of the constrained variables. Similar problems arise from the common simplification that electron distributions around atoms can be approximated by fixed-point charges centered on the nuclei. The resulting inaccuracies in the electrostatic energy are the subject of current work on polarizable force fields with off-atom charges[35,36].

Determining whether a more detailed, higher dimensionality description warrants the increase in the difficulty of conformational search for the lowest

energy state is a challenging problem in itself. Here we describe a general approach to comparing different dimensionality protein representations, and use it to investigate the tradeoff between the changes in conformational space size and model accuracy associated with ideal versus flexible bond length and angle representations.

Results

Protein structure prediction is a global optimization problem involving a search for the lowest energy structure. Comparing the effectiveness of alternative polypeptide chain representations is not trivial: introduction of additional degrees of freedom will almost always result in lower energy models both close to the native structure and far from the native structure, and the effects of this on conformational search can be quite complex. The most straightforward approach—carrying out *ab-initio* structure prediction calculations using different representations and evaluating the success in prediction—is challenging, as it is very difficult to converge global searches over the vast protein conformational space.

We have taken an approach to tackling the representation granularity problem that reduces the stochastic variation inherent in Monte Carlo global optimization. Large sets of models are generated which span the conformational space, and the lowest energy structures in each RMSD interval are collected. These states collectively represent the low-lying minima in the energy landscape. Changes in model representation and optimization method are then evaluated by relaxing

each model in the set and evaluating the energy gap between models inside and outside of the native energy minimum: representations leading to larger energy gaps (normalized based on the spread in energies among the models) are considered better than those with smaller gaps.

We used this approach to compare fixed and flexible bond angle representations, and optimization in internal coordinates versus Cartesian coordinates (described in Methods and illustrated in Figure 1).

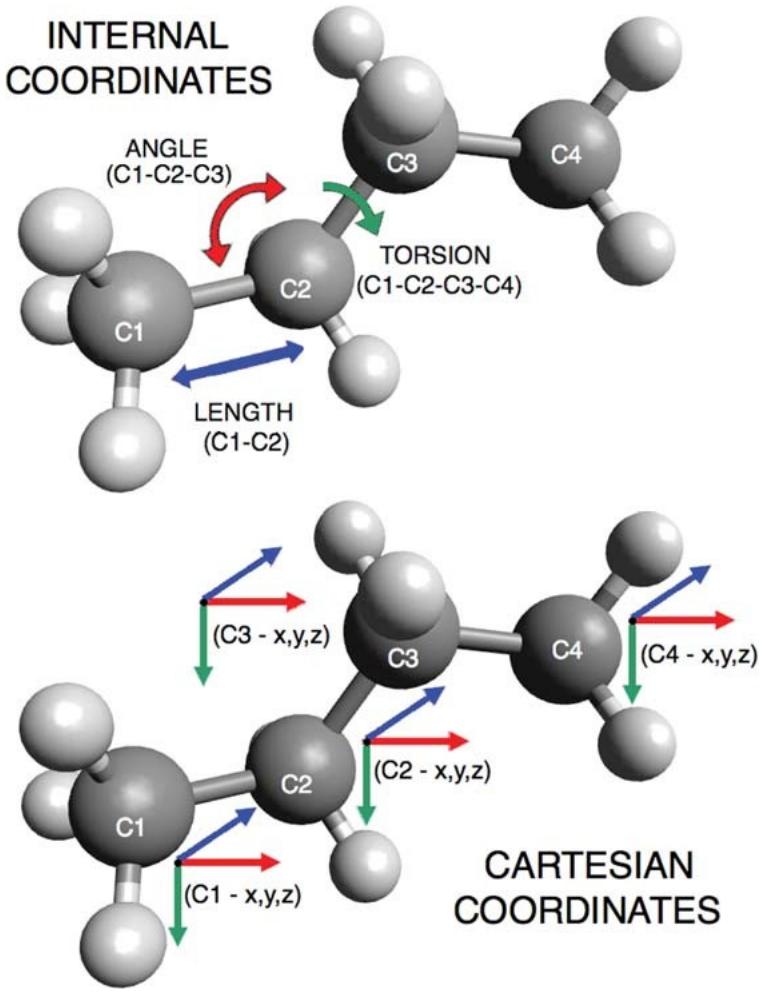

Figure 1. Internal coordinate vs. Cartesian coordinate repre- sentation. Internal coordinates describe a protein structure in terms of angles, lengths, and torsions; Cartesian coordinates describe a protein's conformation using the (x, y, z) position of each atom.

Large numbers of conformations for 82 different proteins were optimized in the different representations for a range of weights on the bonded geometry term. The results are summarized in Figure 2. Figure 2A illustrates some of the energy landscapes for which there was a marked difference between representations. It is evident that the energy gap between close-to-native conformations and far-from-native conformations is larger in the flexible bond angle representation (Figure 2A, right) than in the fixed representation (Figure 2A, left); indeed for protein 2nr7 (bottom row), there is no energy gap in the fixed representation but a clear gap in the flexible representation.

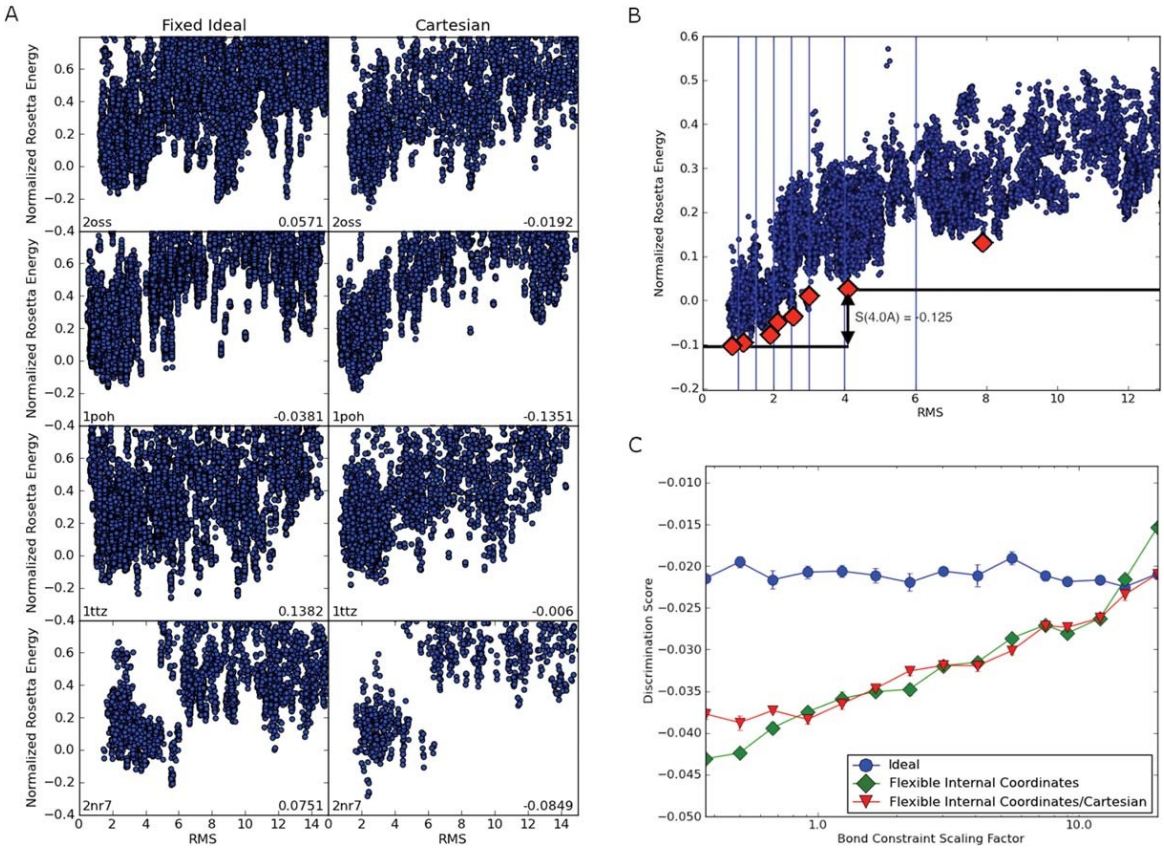

Figure 2. Backbone flexibility increases the native energy gap. (A) Examples of energy landscapes for individual proteins resulting from fixed bond length and angle relaxation (left) and from bond angle relaxation (right). The y-axis is the Rosetta energy normalized by rescaling the energies such that the 95th percentile and fifth percentile fall on 1 and 0, respectively. The x-axis is the RMSD to the native structure. The discrimination measure is provided at the bottom right of each panel; the better the energy funnel, the more negative the value. For these four proteins, the energy gap between the native structure and far from native structures increases with flexible bond relaxation. (B) Illustration of the discrimination calculation. The discrimination measure is the average of energy gaps sampled at seven points on the landscape. The energy gap for each division is computed by finding the difference between the lowest energy structure to the left of the division and the lowest energy structure to the right. The red diamonds represent the lowest energy structure in each bin. In this case, the lowest energy structure to the left of each division will always be the far-left structure. (C) Backbone flexibility increases the native energy gap across the 82 protein benchmark set. For each value of the bond constraint scaling factor on the x-axis, 900 conformations for each of the 82 proteins were relaxed five times. The discrimination measure was computed from the resulting 4500 structures as outlined in panel B, and the values for the 82 proteins were averaged. More negative values indicate larger native energy gaps. At values of the scaling factor less than 0.37 (the lowest value shown on the x-axis) the bonded geometry begins to deviate from that observed in native crystal structures (data not shown). The ideal geometry calculations are not influenced by the scaling factor; the small amount of variation at different values of the scaling factor indicates the amount of noise in the averages. The results for all three protocols converge at high values of the scaling factors as expected since the backbone geometry is near ideal even for the flexible protocols.

It is not feasible to inspect energy landscapes for 82 proteins for many different parameter values; instead, to quantify the magnitude of the energy gap, we developed a discrimination measure described in the methods section and outlined in Figure 2B. We computed the average energy gap over all 82 proteins for fixed internal geometry, flexible internal geometry, and Cartesian geometry. Figure 2C summarizes the dependence of the energy gap on the weight on the bond geometry potential. Consistent with the selected example landscapes in Figure 2A, flexible protocols performed significantly better (more negative discrimination scores indicate larger energy gaps) over the full set compared with the fixed protocol. As expected, all three methods yield similar energy gaps at high covalent restraint weights, where the two flexible models converge on the ideal geometry case. No difference in energy gap is seen between the flexible protocol variants. These plots suggest that lower covalent weights than shown on the figures would improve discrimination, but doing so leads to structures with physically unreasonable bond angles and lengths.

We hypothesized that flexibility in certain bonds may be more important to the energy gap than others. For example, flexibility in backbone geometry may play a much larger role in sharpening the energy landscape than flexibility at the tip of a side chain. To identify which degrees of freedom were responsible for the improved discrimination, we repeated the discrimination experiments with different subsets of the bond angles and lengths free to vary in internal coordinate space; the results are summarized in Figure 3. We found that most – but not all – of the increase in

energy gap is obtained by allowing only the backbone angles to deviate from the ideal values. Allowing only sidechain angles or only τ (N-Cα-C) angles to vary had less effect. Allowing variation in bond lengths alone (Figure 3B) had little effect on energy gap overall, but subtly increased the discrimination score as the constraints increased. Allowing both angles and lengths to vary performed worse than keeping bond lengths fixed (Figure 3A). The poorer performance of angles + lengths flexibility, as well as the slight change in flexible bond length discrimination with higher constraint weight, is likely due to the high bond length spring constants causing the quasi Newton optimization algorithm to take shorter steps and increasing the convergence time[37].

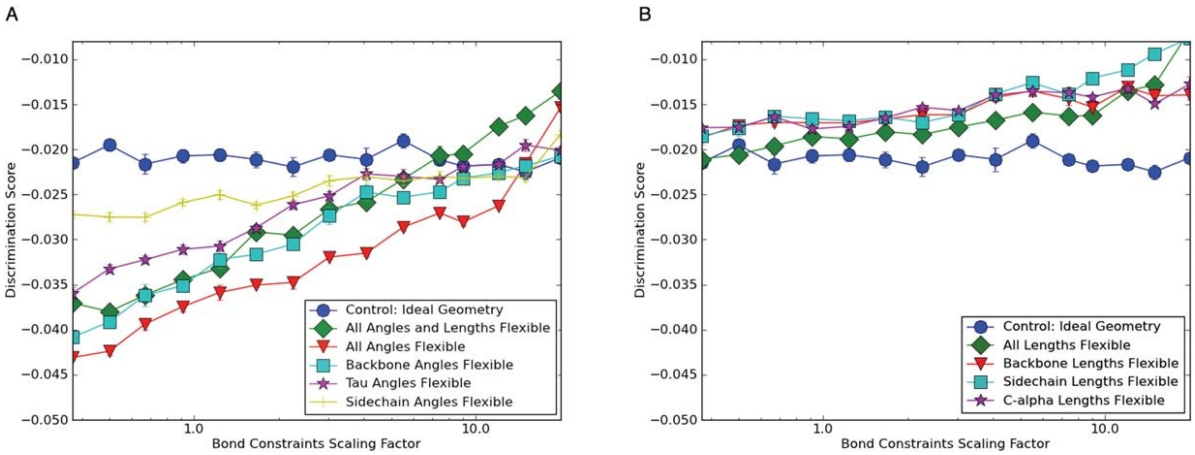

Figure 3. Backbone angle flexibility is the dominant contributor to the increased native energy gap conferred by bonded geometry optimization. (A) The benchmark set calculations described in Figure 1(C) were repeated allowing different subsets of angles to relax during internal coordinate minimization. Keeping bond lengths fixed but allowing all angles to vary led to better discrimination than varying all bond lengths and angles. Most of this improvement resulted from varying all the backbone angles; minimization of sidechain angles or the tau angle (N-Ca-C') had a smaller effect. (B) Varying bond lengths show no effect on discrimination except for a slight decrease in performance at higher weights likely caused by increased convergence time due to shorter steps during quasi Newton optimization.

As described above, because of the computational cost associated with optimizing large numbers of structures, the above comparisons of representations involved relaxation of large numbers of already generated models. For this purpose we used Rosetta FastRelax, an efficient local optimization procedure[33]. However, it is formally possible that new minima exist when the additional degrees of freedom are added, but local optimization fails to identify them. Therefore, we carried out more aggressive global optimization on a subset of targets to see if new local minima emerged due to the additional model flexibility. For 10 of the targets from the original test set, we used the large-scale parallel loop hash (PLS) sampling procedure[26] starting from 200 input structures per target. Either standard ideal geometry FastRelax or FastRelax using flexible bond geometry was used for local optimization in the PLS protocol. As shown in Figure 4, allowing relaxation of backbone-bonded geometry again increased the energy gap between native and non-native structures. Hence, it is unlikely that relaxing bonded geometry creates new minima with energies comparable to the native structure.

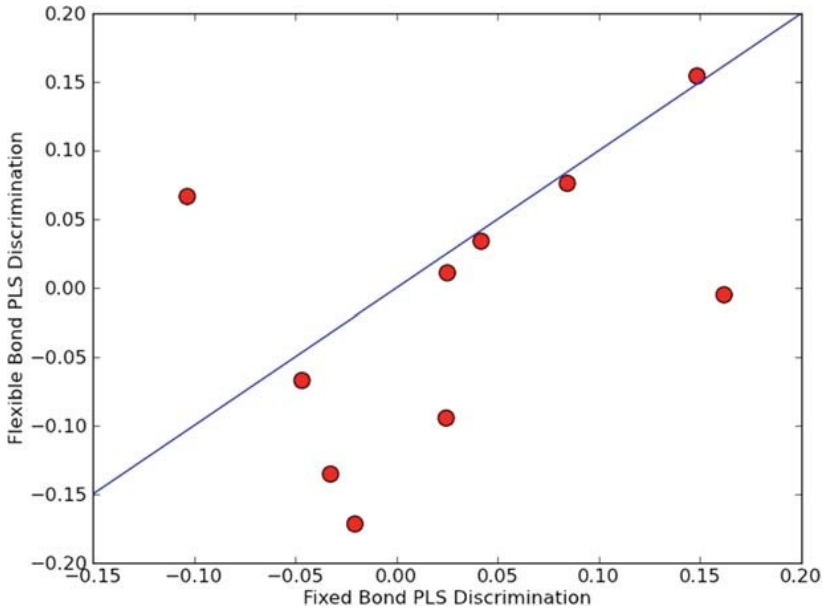

Figure 4. The increase in energy gap with flexible bond optimi- zation is observed even with extensive sampling. Large-scale parallel loophash sampling (PLS) optimization was performed for 10 proteins, and the discrimination score was computed. The increase in energy gap with bond flexibility is similar to that observed with more local optimization in Figure 1.

As a final test of the importance of flexible bond angles in protein structure modeling, we compared X-ray structure refinement with ideal backbone geometry to refinement with flexible backbone geometry. X-ray refinement has recently been implemented in Rosetta[38], and hence this comparison could be readily made using the alternative minimization protocols described above. We refined idealized (bond geometries set to their ideal values) high-resolution structures against truncated (to 4Å) reciprocal space crystal diffraction data; the truncation was to a resolution where bond non-ideality is not specified by the data alone. Idealizing structures provides a common starting point and helps pinpoint the effect of adding bond angle flexibility. The R-free values after refinement with flexible bond angle optimization were consistently lower than after refinement with ideal bond angle optimization

(Figure 5), further highlighting the importance of flexible bond angles in defining native structures.

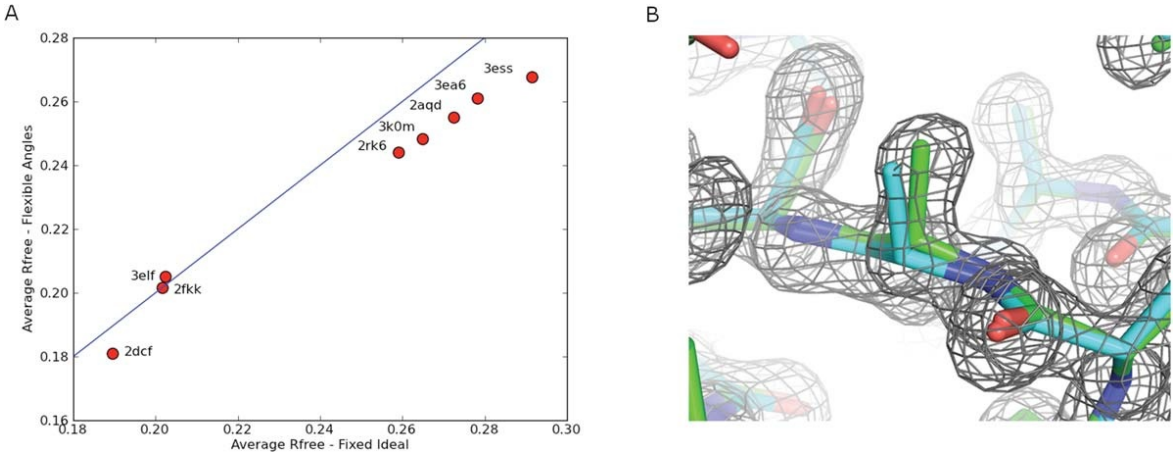

Figure 5. Crystallographic refinement with flexible bonds gives a better fit to low-resolution data than fixed bond refinement. (A) In all but one case, flexible refinement yields a lower average R free. (B) Comparisons of models for 2rk6 after ideal (blue) and flexible (green) refinement. The map is built from high-resolution data - not the low-resolution data used for refinement - to better represent the improved fit.

Having determined that flexible bond angles - particularly backbone angles – yield energy landscapes with the native structure in a deeper minimum, we proceeded to experiment with the FastRelax protocol to identify the most efficient local optimization protocol. Figure 6 shows the results on several different non-ideal FastRelax variants. 270 decoys were optimized under each protocol; the average energy over the set is plotted as a function of average protocol run time. The first observation is that protocols making use of Cartesian optimization were more effective in reducing the energy than protocols optimizing internal coordinates alone. Of the internal coordinate optimization protocols, lowest energies were obtained when all angles but no lengths were allowed to deviate during refinement. Finally, the lowest overall energies were observed by running a two phase protocol,

where the structure was first optimized in internal coordinates with bond lengths fixed, then optimized in Cartesian space. This combined protocol performs considerably better than either internal coordinate torsion angle optimization or Cartesian optimization alone (compare Fig 6 left, red circles with Fig 6 open black).

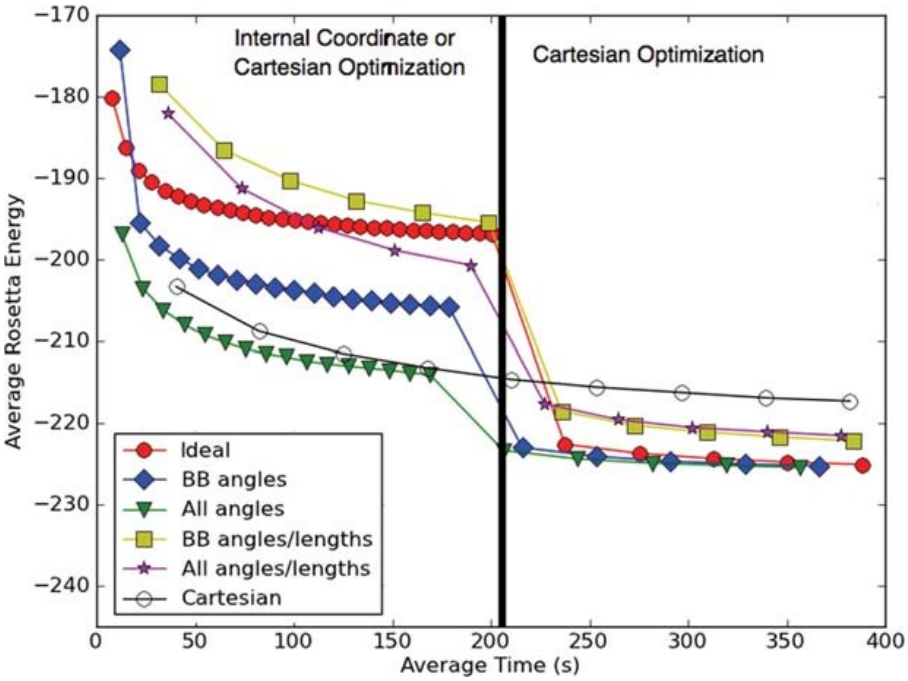

Figure 6. Combined internal coordinate and Cartesian relax is more effective than either one alone. A representative set of starting models was relaxed using different protocols (col- ors) and the average energy over all calculations (y-axis) was determined as a function of run time (x-axis). At 200 s (black vertical bar), the minimization method within the relax proto- col was changed as indicated in the inset box (for example, ideal-Cartesian indicates that the first 200 s used ideal bond internal coordinate minimization, and the second 200 s, Car- tesian minimization (all DOFs variable). The best performance (lowest energies after 400 s) was obtained using protocols that kept bond lengths fixed in the first phase and switched to Cartesian minimization in the second phase.

The improved performance of the protocol with fixed bond length internal coordinate optimization followed by Cartesian optimization may be rationalized as follows. In the initial phases of optimization, restricting degrees of freedom has the advantage of allowing exploration further in conformational space and avoiding

trapping in local minima not accessible with more ideal geometry. Restricting bond lengths has the further advantage of speeding up the internal coordinate optimization. On the other hand, Cartesian minimization is likely to be much more effective in finding the lowest energy structure in the immediate neighborhood and hence is very effective at the late stages of optimization.

Allowing bond flexibility during refinement increases running time somewhat: our combined refinement protocol takes approximately 110 seconds for a 100 residue protein compared to about 40 seconds for the fully ideal protocol. Most of the runtime increase is due to Cartesian optimization, where the large bond length spring constants require significantly more minimization cycles for convergence.

Methods

To assess the accuracy of different dimensionality protein representations, we focus on the energy gap between the native and non-native conformations; for folding to occur, the energy of the native structure must be very much lower than non-native conformations. Identifying the lowest energy non-native conformations from scratch for large numbers of alternative representations is computationally intractable. Therefore, we initially generated a set of very low energy structures spanning conformational space using a large-scale space search procedure[39]. The number of representatives at each RMSD distance from the native structure was normalized to prevent overrepresentation of any particular area of conformational

space. The resulting structures are densely packed and energetically competitive conformations (or "decoys") of the native sequence (Figure 2A). Changes in model representation and local optimization methods are then evaluated by minimizing each pre-calculated structure in the new force field, allowing it to descend to its new local minimum. The difference in energy between models inside and outside of the native energy minimum is then evaluated: model representations leading to larger energy gaps (normalized based on the spread in energies among the models) are considered better than those with smaller gaps (Figure 2B).

We apply this approach to comparing fixed and flexible bond representations in the context of the Rosetta force field. Rosetta modeling calculations generally involve an internal coordinate representation based on a "fold tree[40]". The standard fold tree only allows backbone and sidechain torsions (for rotamers) as degrees of freedom during minimization, bond angles and lengths are kept fixed. Flexible bond geometry was implemented by allowing bond lengths and angles to vary in the internal coordinate space optimization. The implementation allows selective restrictions of subsets of DOFS, for example, letting only the backbone angles vary.

We compared internal coordinate optimization to Cartesian-space optimization (Figure 1); in the former, minimization is guided by gradients with respect to bond and torsion geometries; in the latter, by gradients with respect to the (x,y,z) coordinates of each atom. By definition, Cartesian-space minimization permits flexible bonds and planarity. For both representations, parameters for harmonic bond length and angle force constants were taken from CHARMM32[29]. Additional constraints using CHARMM32 parameterization were also added to

control improper torsions in Cartesian space. A global weight was used to control the scaling of the bonded versus non-bonded terms. MolProbity was used to validate bond angle and length distributions to ensure structure predictions were physically realistic[41].

Local optimization protocol

To compare fixed ideal internal coordinate minimization, flexible internal coordinate minimization, and Cartesian minimization, the standard Rosetta local optimization protocol – FastRelax[42] - was adapted to optionally allow minimization of covalent degrees of freedoms (bond angles and lengths), as well as minimization in Cartesian space. We also modified the FastRelax protocol to carry out first internal coordinate optimization, which may have a larger radius of convergence[34], and then Cartesian minimization.

The original FastRelax protocol uses multiple iterations of repulsive weight annealing with combinatorial rotamer optimization and minimization to optimize energies (ramp-repack-min). There are five cycles each consisting of four iterations of ramp-repack-min starting with repulsive at 2% of full strength, followed by 25%, 55%, and 100% successively. Only minimization permits flexible bonds; the rotamer set used in repacking has ideal bonds.

Fixed ideal internal coordinate relax uses the original FastRelax protocol described above. Flexible internal coordinate relax uses the same protocol, but frees the bond angle degrees of freedom during minimization. Cartesian relax starts with 3 rounds of flexible angle internal coordinate minimization and then performs the

remaining two rounds of FastRelax in Cartesian space. Command lines are provided in the Supplemental Materials.

Discrimination benchmark

The benchmark we use to evaluate the discriminatory power of the Rosetta force field consists of 82 small globular proteins[39,43] covering a diverse set of topologies. All proteins are monomers between 55 and 224 residues in length, have crystal data with <2A resolution, and with crystal-stabilized regions visually identified and removed. For each protein, 40000-200000 decoys were generated using biased and unbiased *ab-initio* sampling runs[39] followed by extensive loop building and relaxation using the Rosetta full-atom energy function and PLS[26]. Additional PLS runs were seeded with the native structure to further increase sampling density near the native state. The resulting decoy structure sets comprise many competitive low-energy non-native conformations, sometimes lower in energy than close-to-native structures. All these conformations were pooled and 1000 representative low-energy structures from each protein were chosen to evenly cover the range of possible RMS values.

To test each set of parameters and flexibility settings, we ran 5 FastRelax trajectories per starting model, producing 369,000 decoys in total. This short refinement balances the need to let each structure optimize against the new parameter set and computational feasibility. Each full test of a parameter set consumed ~50,000 CPU hrs.

Discrimination measure

To quantify the discriminatory ability of a parameter set we used the following procedure, given a large set of structures and their energies. First, energy values are normalized by rescaling the energies such that the 5th percentile and 95th percentile energies take the values of 0 and 1, respectively. Then, for each protein a separate discrimination score **s** is calculated (Figure 2B) at 7 different RMS values **r** = [1.0,1.5,2.0,2.5,3.0,4.0,6.0], by taking the normalized energy difference of the lowest-energy structure below and above the dividing line at each **r**.

$$S = \sum_{r\in\{1,1.5,2,2.5,3,4,6\}} \min_{i,RMS(i)\in[0,r]} E_i - \min_{i,RMS(i)\in(r,\infty]} E_i$$

The total discrimination score is then calculated as the average score over all proteins and all values of r. The score is constructed such as to capture changes in discrimination at various resolutions, with a lower score indicating better overall discrimination.

Backbone Conformational Sampling

To further assess discrimination under more aggressive search, we used the parallel loophash sampling protocol (PLS)[26] on a subset of the 82 proteins used above. In each iteration of PLS, a set of input structures are selected and local structure segments are randomly replaced with segments found in the PDB. Variants are relaxed and the lowest energy variants are accumulated and filtered by

a diversity criterion. PLS is able to generate large backbone conformational changes and samples a significant portion of conformational space around a given topology.

While PLS is a powerful sampling protocol, it was not computationally feasible to run it for all 82 proteins and all parameter sets. Instead, ten proteins without disulfide bonds (which complicate topology sampling) were randomly selected from the set. One PLS run for each structure was performed, starting with 200 low-energy decoys per protein, evenly selected across the range of possible RMS values and excluding any decoys created from native-biased *ab-initio* algorithms. Each run sampled with 8,192 cores for 6 hours on the Intrepid Blue Gene/P supercomputer at Argonne National Laboratory. Upon conclusion of the runs, the energy landscape was well covered over a large range of RMS values.

Crystallographic refinement

A set of eight crystal structures that had been solved using high-resolution crystal data, along with their deposited structure factors, was chosen from the PDB. The structures were "idealized" by forcing ideal geometry and minimizing with constraints on the atom positions of the deposited structure, resulting in a model with ideal geometry and very low RMS deviation (generally less than 0.2Å) from the original model. The crystallographic data was then truncated to 4Å - a resolution too low for the data to identify deviations from ideal geometry - and the structures were refined against the truncated data using Rosetta-Phenix refinement in internal coordinates[38]. Two separate refinement trajectories were run: one where bond

geometry was allowed to deviate from ideality, and one where it was not. After both refinements, the free R factor (using the reflections marked as free in the deposited structure) of the ideal and non-ideal models was calculated.

Optimizing FastRelax for Flexible Geometry

To optimize the FastRelax protocol for the larger search space and dual representations, we used a small benchmark of 270 compact decoys, generated by the Rosetta *ab-initio* protocol described earlier[39], with randomly selected sidechain conformations. Because the centroid structures were minimized with a different energy function than used for full-atom minimization (which is used by FastRelax), the comparison is not biased by the starting minima of the benchmark set. For each FastRelax cycle, we computed the average final energy and elapsed time from start. The relax protocols were performed for 400 seconds. The weights on the energy function score terms were the default Rosetta (score12) weights (ref) with the addition of the *cart_bonded* global bonded term (set to a weight of 0.5). We tested a variety of different FastRelax protocols utilizing various combinations of fixed internal coordinate, flexible internal coordinate and Cartesian minimization.

Chapter 3

Correcting Double Counting In The Rosetta Sidechain Torsion Potential

Abstract

Biomolecular modeling requires a balance between accuracy and speed. To accomplish this, Rosetta uses a combination of statistical and physical potentials in the energy function. The former captures distributions of protein features while the latter directly models the physical interactions responsible for the resulting features. As a result, there can be significant overlap between score terms; leading to double counting of the features that describe a well-folded protein. In particular, the statistical potential used to describe sidechain torsion distribution as a function of backbone torsion angles (Dunbrack rotamer library) overlap with physical terms which score backbone-sidechain and intra-residue interactions. Consequently, rotamer distributions are skewed to those favored by the overlapping physical terms. Using a probability distribution to model sidechain conformation in place of intra-residue physical terms gives better prediction, but still suffers from backbone-sidechain double counting between adjacent residues due to the inter-residue scoring still in place. To address this, we eliminated redundancies from the rotamer library probabilities. In this paper, we describe a method for sampling physical Rosetta potentials over sidechain conformational space and using the observed energies to modify the rotamer library probabilities. We also describe a method of using electron density data to evaluate the impact of energy function changes on a large set of high-resolution protein structures. By training on this benchmark, we

were able to correct 13% of the errors by adding physical intra-residue score terms while adjusting the rotamer library to compensate for double counting.

Introduction

Modeling protein energetics is a challenging problem that has been addressed by a variety of approaches. From the simple Engh-Huber potentials used by X-ray crystallographers to perform density fitting[44] to intensive molecular dynamic trajectories that step through time intervals of 2 femtoseconds, each approach is ultimately a compromise between detail and speed. The Rosetta forcefield uses a combination of physical and statistical potentials to express total system energy as a linear combination of scoring terms[7]. Physical potentials explicitly model chemical interactions while statistical potentials capture distributions of structural features determined by a broad range of underlying chemical interactions. Physical energies are expressed directly by the functional form of the term while statistical energies are computed by taking the negative logarithm of observed probabilities. A significant drawback to this approach is the redundancy that arises when a statistical potential implicitly models the same interaction already described by another score term because it encodes all energetic contributions causing the observed distribution.

In Rosetta, the statistical Dunbrack rotamer library is used to score sidechain conformations[45]. Sidechain torsion distributions taken from the PDB exhibit peaks known as rotamers for each of the chi angles for each residue. In the library, rotamers are described by their mean value and standard deviation. Additionally,

rotamer probabilities depend on the backbone phi and psi torsions, binned every 10 degrees from -180 to 180 degrees. Sidechain conformations are scored by a combination of the probability of their rotameric identity and deviation from the rotamer well center where the former is scored using the negative logarithm of the probability and the latter using a harmonic constraint based on the standard deviation of the rotamer well. Furthermore, the last chi of the aromatic residues and ASN, ASP, GLN, GLU are modeled using asymmetric probability density functions conditioned on the rotamer identities of the preceding chis and backbone torsions and scored accordingly. Using conformational probabilities along with physical score terms is a fast and effective way to predict sidechain conformations.

Due to double counting of intra-residue interactions between the rotamer library and the physical intra-residue score terms, we simply turn off the physical intra-residue score terms in Rosetta and rely solely on the rotamer library. Although the physical score terms have finer granularity, the rotamer library gives better sidechain conformation predictions; presumably due to parameterization issues with the physical score terms or because it captures unmodeled interactions such as quantum mechanical or entropic effects (Table 1). This approach largely eliminates rotamer library double counting. However, it is important to note that the rotamer distributions are not exclusively determined by intra-residue interactions. There is still overlap that arises when rotamer probabilities are influenced by local backbone-sidechain interactions such as hydrogen bonding with an adjacent residue. The rotamer library implicitly scores this hydrogen bond by giving this rotamer a higher probability while the hydrogen bond term in Rosetta also favors it.

This leads to inaccurate evaluation of rotamer energies and incorrect rotamer distributions[19] (Fig 6). Due to performance, the rotamer library should not be removed, but it can be modified to work with inter-residue score terms. Furthermore, it can be modified to work with physical intra-residue score terms to improve modeling resolution.

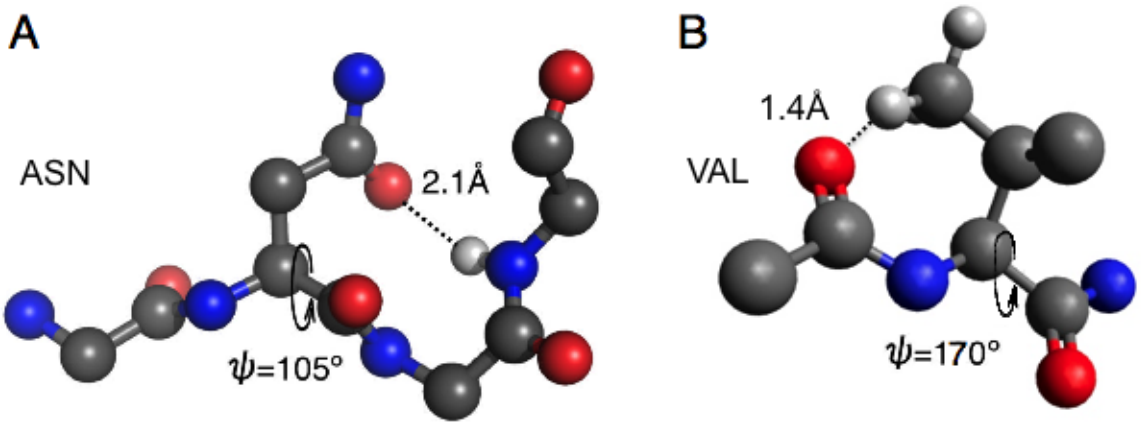

Figure 6. ASN and VAL exhibit backbone-dependent double counting. ASN is able to make a hydrogen bond when the backbone psi torsion is at 105°. The rotamers within hydrogen bonding distance are more probable, thus decreasing the rotamer score relative to others. When the Rosetta hydrogen bond term also decreases the rotamer energy, the sum of the two misrepresents the true rotamer energy relative to other potential rotamers at this backbone conformation. The same situation occurs with VAL and the repulsive score term, except that this rotamer is doubly disfavored when the backbone psi torsion is at 170°. (Fig 6A was created by Song, *et al*[19])

In molecular mechanics, torsional potentials are adjusted so that the new potentials combined with the existing physical forcefield will sum up to give energies that are equal to detailed quantum mechanic calculations on small peptides[36]. A similar approach can be performed for Rosetta where inter-residue energies are subtracted from the rotamer potential energies. In other words, adjust

rotamer probabilities to compensate for the influence of other Rosetta energy terms on rotamer distributions of modeled structures. The result is a new rotamer library that will correctly sum with all score terms to give native rotamer distributions (Fig 7). In this paper, we describe a method to compute corrections to the rotamer library by scoring potential rotamers on a collection of protein fragments and a sensitive benchmark that uses experimental data to verify and guide corrections.

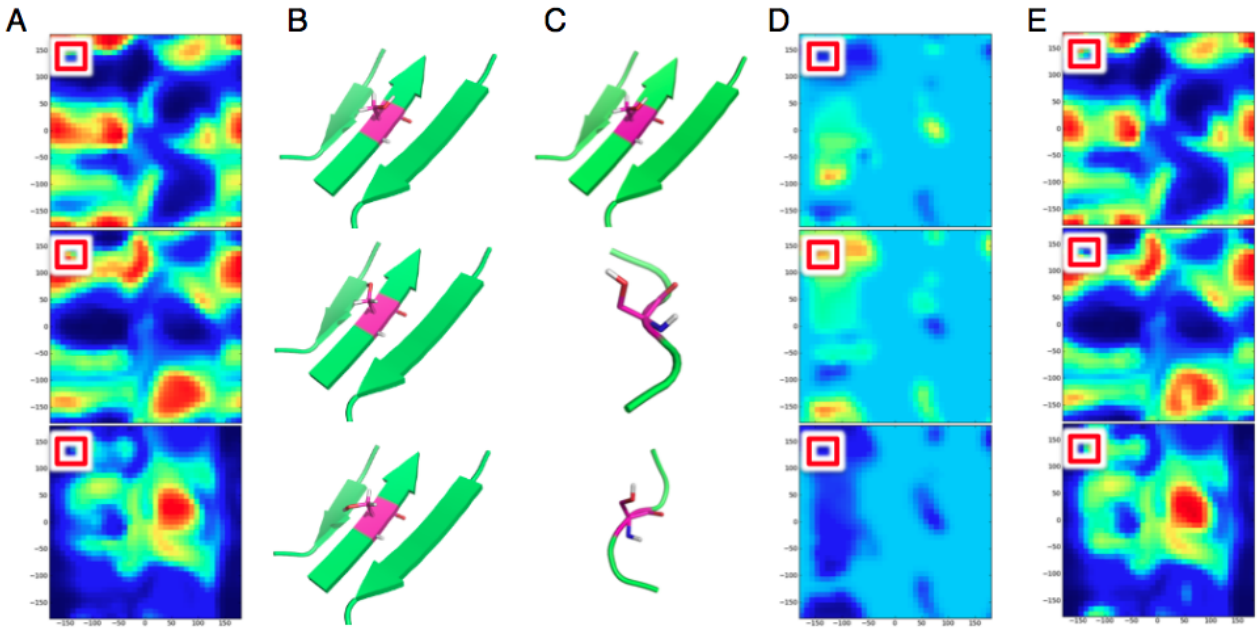

Figure 7. Phi-psi dependent probabilities for a serine rotamer are adjusted based on the influence of the Rosetta energy function. A. Plots showing the phi-psi torsion dependence of probabilities for each of the three serine rotamers. B. A serine fragment representing the red box highlighted in A. The same fragment is shown with each of the three rotamers built in their ideal conformations. Scoring each fragment reveals any rotamer-dependent difference seen by the Rosetta energy function. C. Three fragments from the highlighted region in A with the same rotamer. Average rotamer-dependent scoring difference is obtained by sampling each rotamer on all fragments within the same phi-psi torsion bin. D. Plots showing the phi-psi torsion dependence of probabilities accounted for by the Rosetta energy function. E. The corrected rotamer probabilities are computed by subtracting the probabilities in D from A.

Results

To correct the double-counting issue, we designed a protocol to subtract double-counted interactions from the rotamer library. By using the fragment-based rotamer sampling described in the methods section and a modified Rosetta energy function to consider only double-counted score terms, we were able to compute corresponding Rosetta energies for every rotamer probability in the library. In each phi-psi torsion bin, all rotamer probabilities sums to one. By using observed energies to compute a Boltzmann distribution and then normalizing across all rotamers in a phi-psi bin, we effectively created a rotamer library based on Rosetta energies. To account for undersampling and noise, a Gaussian convolution was applied to smooth the data. From there, we divided each rotamer library probability by the corresponding Rosetta-derived rotamer probability and renormalized. If all rotamers in a given phi-psi bin have equal Rosetta energy, the rotamer library probabilities will all be corrected by the same amount and returned to their original values after normalization. If a rotamer is scored lower than others, the rotamer library probability ends up being lowered to account for the double counting.

Using this protocol, two libraries were made – one correcting double counting interactions present in the default Rosetta energy function and the other correcting double counting from using physical intra-residue terms modeling Lennard-Jones and solvation in addition to the default terms. Including the physical intra-residue terms and accounting for double counting permits us to have the best of both where we benefit from the specificity of the physical terms while retaining the rotamer library's ability to fill in for unmodeled interactions.

To assess the performance of the rotamer probability correction protocol, we needed a sensitive benchmarking protocol that was fast enough to use in an iterative workflow. Exhaustively sampling conformational space to fit new parameters is computationally intractable and confounds efforts to group errors by residue type. Instead of relying on combinatorial sampling of all degrees of freedom to investigate changes guiding the conformation of a single residue, we chose to perform RTMin rotamer recovery on a fixed backbone. RTMin starts with native conformations and allows Rosetta to delete and rebuild sidechains one residue at a time. For each sidechain rebuild, Rosetta samples and minimizes all potential rotamers, then chooses the lowest energy conformation. The benchmark determines recovery by computing the change in electron density correlation between the native and rebuilt sidechain conformations (Fig 8). The sidechain is then returned to native before the next sidechain is rebuilt. This RTMin rotamer recovery protocol is fast and deterministic.

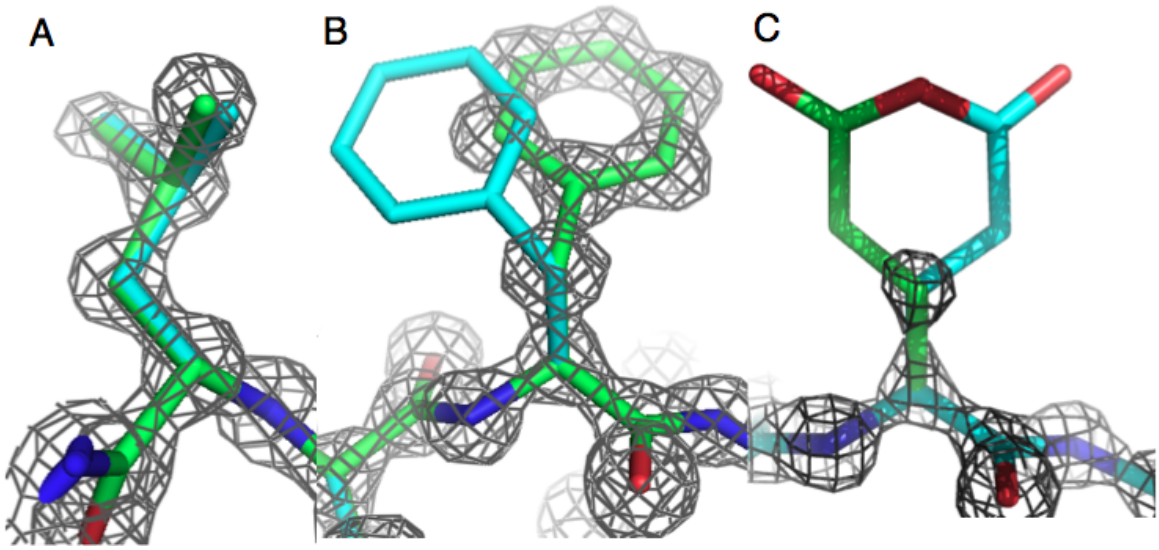

Figure 8. Recovery of native conformation after a sidechain rebuild is determined by electron density data. A. Rosetta chose the correct sidechain conformation. B. Rosetta chose the incorrect sidechain conformation. C. There is not enough electron density data to determine recovery of native conformation. This sidechain is not included in the calculations. Electron density correlation values to determine each of the three scenarios were manually determined by inspecting many structures.

To ensure that the RTMin protocol results are well correlated with energy function issues instead of confounding factors such as undersampling or ambiguous experimental data, we increased sampling and filtered sidechains from the results. We sampled all available rotamers in the rotamer library plus conformations one standard deviation to either side of the two rotameric chi closest to the backbone. Filters on the sidechain results only allowed well-defined sidechains, mainly defined by electron density correlation. The end result is an extremely sensitive benchmark that captures energy function errors and allows one to easily group them by local features.

Table 1 shows rotamer recovery results for several score functions. Results shown in Table 1 were generated using a different data set from the one used for method development, parameter fitting, and rotamer probability corrections. Talaris2013 is the default score function that we are working from. For comparison, the predecessor to talaris2013, score12[46], is shown. The switch to talaris2013, which involved many changes to the energy function, corrected nearly 20% of the errors in the rotamer recovery set. Observing the performance of Rosetta without the rotamer library shows that it is needed, even if physical intra-residue terms are substituted.

Run	Error Rate	% Error Correction
Score12	5.57%	-24.61%
Talaris2013	4.47%	0%
Talaris2013 -rotamer library	7.38%	-65.10%
Talaris2013 + physical intra – rotamer library	6.37%	-42.51%

Corrected Run	Recovery	% Error Correction
Talaris2013	4.28%	4.25%
Talaris2013 + physical intra	4.02%	10.06%
Talaris2013 + physical intra + dun_dev 0.3	3.88%	13.20%

Table 1: Comparison of rotamer recovery error rates for different energy functions. The top four entries all use the uncorrected rotamer library while the bottom three entries show variations on the energy function used for corrections and assessment. Percent error correction is calculated as the amount of errors corrected out of the error count for the default score function, Talaris2013.

Adjusting the rotamer library probabilities using the correction protocol shows a decrease in error rate across all categories. While recovery improves after

correcting for backbone-sidechain double counting, better results are seen when adding physical intra-residue score terms. This shows that sidechain conformation prediction benefits from the granularity in scoring that the rotamer library cannot provide. Loosening the harmonic constraints that restrain sidechains to the center of their rotamer wells shows further improvement. This is presumably due to double counting between the constraints and the intra-repulsive score term leading to unnecessarily tight distributions around rotamer means. Also encouraging is that all residues show improvement in rotamer recovery on the test set (Fig 9). This shows that the correction method is generally applicable and is not subject to tradeoffs.

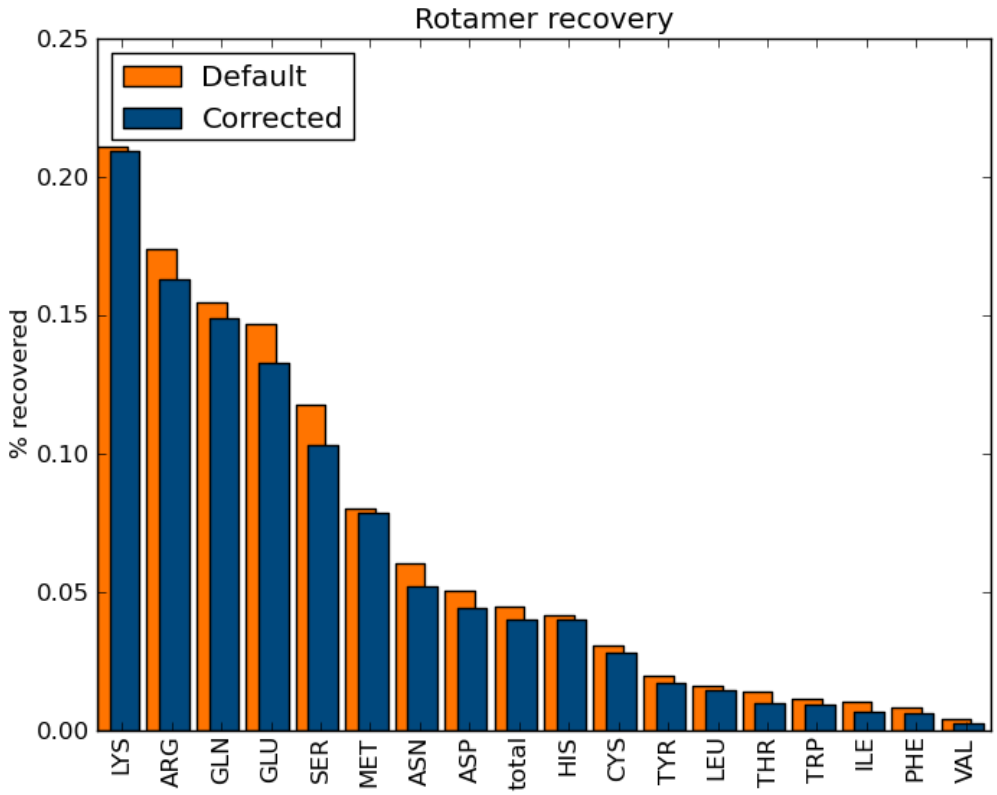

Figure 9. Adding physical intra-residue score terms and adjusting the rotamer library for the double counting improves rotamer recovery for every residue type.

To assess the corrections in a more general manner, we turned to several benchmark protocols (Table 2). A variation on the RTMin rotamer recovery protocol described above is the standard Rosetta rotamer sampling protocol PackRotamers; which performs combinatorial rotamer optimization in place of greedy sampling and minimization. In contrast to RTMin, PackRotamers assesses cooperative influences in sidechain conformation sampling. It also places more emphasis on scoring with rotamer probabilities since the sidechains do not minimize and potentially incur a harmonic penalty for deviation from the rotamer well center.

Finally, we did not consider the native sidechain conformation during PackRotamers sampling. When performed on the same set of structures and filtering results based on the same criteria, we continue to show improved recovery.

	Default Talaris2013	Corrected Talaris2013 + physical intra
RTMin rotamer recovery	95.53%	96.12%
PackRotamers rotamer recovery	85.38%	85.50%
KL Divergence	0.0035	0.0023
Decoy Discrimination	-0.1566	-0.1702

Table 2. Using the corrected rotamer library with physical intra-residue scoring terms shows improvement in all of the benchmarks.

We were also interested in the question of whether rotamer distributions of Rosetta models were more closely aligned with native rotamer distributions. To obtain Rosetta-representative rotamer distributions, we used the PackRotamers protocol. Figure 10 shows the per-residue Kullback-Leibler distances to the native rotamer distribution for default and corrected Rosetta. A distance of zero signifies that the distributions are identical. In this figure, longer sidechains show a higher divergence due to a larger conformational space and lower chance of randomly making a correct prediction. Longer sidechains are also more prone to overfitting by amplifying probabilities of common rotamers at the expense of rare rotamers. Significant improvements in LYS, ARG, and MET show that our RTMin rotamer recovery improvements are due to correctly accounting for energy function influence on rotamer probabilities.

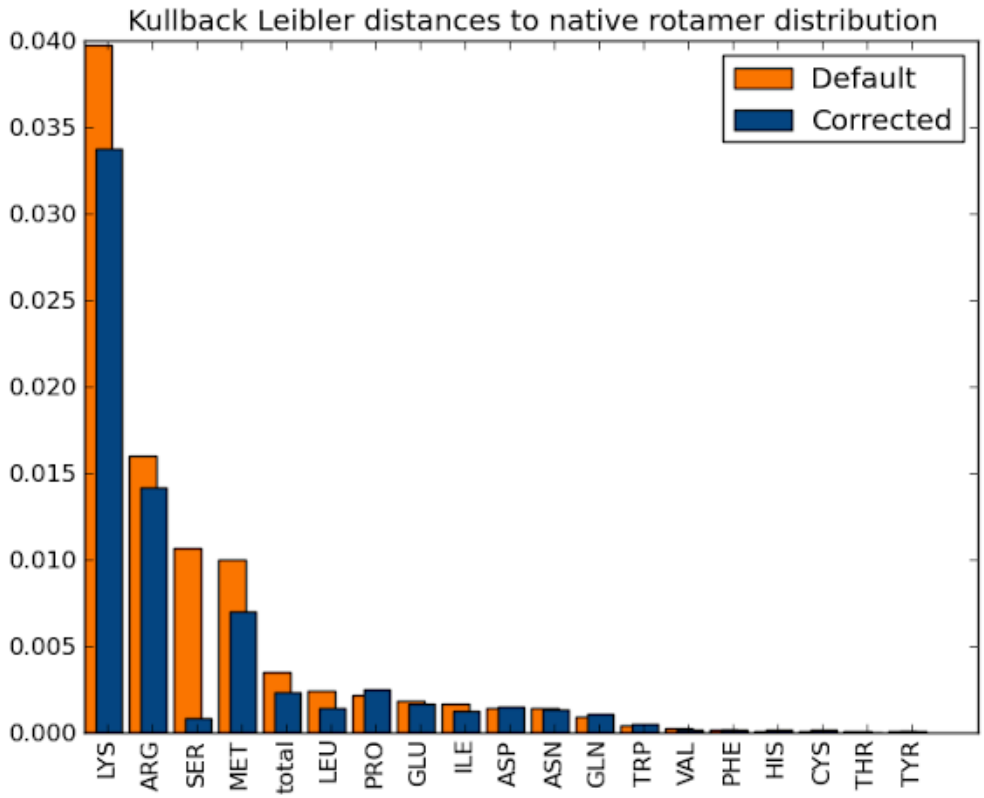

Figure 10. Adding physical intra-residue score terms and adjusting the rotamer library for the double counting improves modeling of the rotamer distribution for most residues when performing the PackRotamers rotamer sampling protocol. Orange compares the native distribution to the default Rosetta rotamer library and score function. Blue shows the same comparison with the use of intra-residue score terms and adjusted rotamer library. Significant improvements in distributions of the longer sidechains show that the rotamer library corrections are correctly accounting for the energy function influence rather than overfitting to high-frequency sidechain conformations.

The gold standard for benchmarking Rosetta energy function changes is the decoy discrimination benchmark using the energy landscape modeling discussed in chapter 1. In contrast to rotamer recovery, it permits all degrees of freedom to be minimized. This time around, we elected to use a modified version of the decoy discrimination protocol previously described. Since the quantitative funnel measure only looks at the energies of the lowest scoring structures across the landscape, we

reduced the energy landscape to 1000 representative structures and performed Batchrelax[26] instead of Fastrelax. Batchrelax performs comparative optimization of a batch of structures, alternating between Fastrelax cycles and culling cycles that reduce the batch size by eliminating a proportion of the highest-energy structures each time. Additionally, the decoy discrimination scoring metric was slightly modified to reduce noise caused by low-population regions. As Table 2 shows, the average normalized energy gap increased by more than 1%. The corrections trained on rotamer recovery without backbone or nonideal degrees of freedom still have the effect of improving energy function performance when using additional degrees of freedom.

Discussion

Double counting of energetic interactions is a known weakness of using knowledge-based potentials in conjunction with other potentials. In Rosetta, the double counting between the sidechain torsion potential and intra-residue score terms was resolved by simply turning off the physical intra-residue score terms. However, the effect of inter-residue score terms was still unaccounted for. Song, *et al.* found specific examples of double counting and manually adjusted rotamer probabilities to address them[19]. This approach is labor-intensive and does not capture unknown errors. In this paper, we describe a method to compute corrections to the rotamer library by quantifying the influence of the Rosetta energy function on rotamer distribution. We show how adjusting the rotamer probabilities

to account for double counting can resolve errors in prediction of sidechain conformation.

Improvement in rotamer recovery between the current energy function and the best performing energy function is not rooted in any one change, but rather it is a cumulative difference over several modifications. Corrections to the default score function resolved backbone-sidechain double counting. The primary driver of these errors is likely the electrostatics term, as demonstrated by the improvements in serine rotamer recovery. It is a smooth long-range potential that falls off slowly enough to have an impact on a large range of situations cases where the energy gap between potential conformations is small.

More significant than correcting the default score function is being able to generally apply this method to correct any modification of the energy function, as well as extend the energy function in ways that was previously impossible due to double counting. For this paper, we demonstrated that we were able to add the physical intra-residue score terms back in and see an improvement in rotamer recovery. Untested is the potential to rebalance the energy function in ways that were not possible due to double counting. For instance, the electrostatics term could potentially be upweighted now that it's influence on rotamer distributions is removed.

This paper demonstrates that it is feasible to quickly make direct corrections to knowledge-based potentials for any given energy function modifications. This method was applied to correcting double counting in the sidechain conformation library, but could be extended to address double counting of the backbone torsion

potential. While physical score terms have finer granularity and more generalizability, they suffer from incorrect parameterization and missing models. By allowing knowledge-based potentials to supplement physical potentials rather than displacing them, the energy function becomes more accurate and useful.

Methods

Dataset

For this paper, we used the Richardson top8000 database of high-resolution crystal structures. For the conformation sampling step of the rotamer library corrections, we used 6832 structures. For the protein core benchmarks, 4000 structures with electron density maps were taken from the top8000 set and split into a training set and a test set.

Dunbrack Rotamer Library Corrections

The rotamer library double counting correction protocol adjusts all rotamer library probabilities to reflect contributions made by the Rosetta energy function. The Dunbrack rotamer library lists rotamer probabilities for phi and psi backbone torsion values at every 10 degrees from -180 to 180 degrees. To account for double-counting influence on rotamer distribution at a given phi and psi torsion pair, we need to compute the corresponding Rosetta energy for each rotamer at this grid point. Initially, we set out to enumerate all backbone conformations on a five-residue fragment and build all rotamers at each step. However, we recognized that we were not accurately accounting for the average local environment seen at each

backbone conformation. By using a large set of high-resolution crystal structures to construct fragments, we can sample local backbone torsions for each residue type. During the rotamer sampling step, averaging over all fragments within a phi-psi bin gives us a typical Rosetta energy for that bin. This has the drawback of undersampling sparsely populated regions of the Ramachandran map, but the benefit of encoding motif-influenced neighboring backbone torsion distributions is well worth it. Local backbone torsions for each residue were sampled according to the DSSP secondary structure type of the sampled residue. Helical fragments are 9 residues in length, loop fragments are 5 residues long, and sheet fragments have up to 15 residues across 3 strands (Fig 11). To reduce noise during energy calculations, all neighboring residues were mutated to alanine. Furthermore, any clashing fragments – defined as having Rosetta repulsive energy greater than 2 – were discarded. Ultimately, we had a set of fragments representing all residue types except ALA, GLY, and PRO.

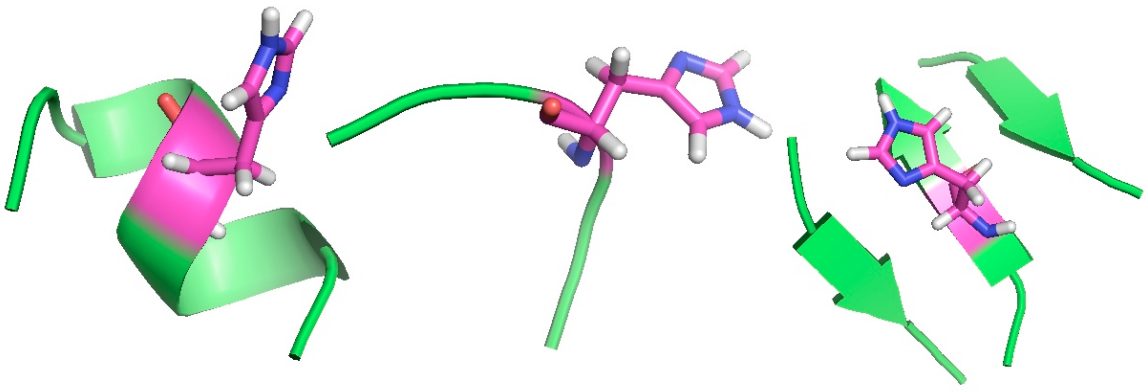

Figure 11. Local backbone torsions for Rosetta energy calculations were sampled as fragments in a secondary-structure dependent manner. All of the green residues are mutated to polyalanine. The rotamer energy sampling protocol samples the pink residue at each rotamer conformation and scores it.

The next step was to build and score all rotamers for each residue so relative differences in rotamer energy could be identified. For the purpose of sampling, the nonrotameric chi probability density function grid points were treated as discrete rotamers. For each residue, every rotamer available in the rotamer library was built and minimized using a modified version of the Rosetta energy function that only scored potential double-counting interactions. Sidechain chis were constrained during minimization using the same harmonic constraints used by the rotamer library scoring term. Given that the rotamers were built without any consideration of local environment, clashes are bound to occur and cause sidechains to minimize out of the rotamer well. Thus, for each minimization, we checked to see if the sidechain is still in the rotamer well – if not, it was reset to center and scored accordingly. To counteract extremely high energy gaps between well-behaved rotamers and clashing rotamers from dominating the corrections, we established a maximum energy gap threshold between the lowest scoring rotamer on the fragment and all others. Any rotamer energies above this relative threshold are changed to the maximum gap value. This threshold was empirically determined by the electron density rotamer recovery benchmark to give best recovery at 3 Rosetta energy units. The final scores were grouped by rotamer, residue type, and phi-psi torsion bin; then averaged to compute Rosetta energies for corresponding probabilities listed in the rotamer library.

Finally, we used the computed Rosetta energies to make adjustments to the probabilities in the rotamer library. First, to account for the downweighting of the rotamer library scoring term by 0.56 during normal Rosetta use, we scaled Rosetta

energies up by 1.785 (1/0.56). We computed a map of Boltzmann probabilities derived from Rosetta energies for all rotamers by normalizing the exponential of the negative rotamer energies across all rotamers in each residue type and phi/psi bin. To account for noise and underpopulated phi/psi bins, a Gaussian convolution was applied to every rotamer across phi/psi space. A Gaussian convolution was also applied to the probability density distributions for each nonrotameric chi across chi values. The extent of smoothing was empirically determined using the electron density rotamer recovery protocol, where rotameric and nonrotameric smoothing were 1.0 and 0.5 standard deviation respectively. After renormalization, we had the final map of Rosetta-energy rotamer probabilities to be applied to the rotamer library as correction factors.

Correcting the rotamer library was a straightforward matter of dividing out the Rosetta-energy rotamer probabilities from the rotamer library probabilities. A larger correction is applied to rotamers that are heavily influenced by the Rosetta energy function. If the Rosetta energy function sees all rotamers as equal, the same correction is applied to all rotamers and renormalization returns the probabilities to their original value. The new probabilities are written over the original probabilities in the rotamer library files and the original rotamer means and standard deviations remain unchanged. Using the corrections is just a matter of loading the new library files in place of the original ones.

Rotamer Recovery

Rotamer recovery is a fixed backbone sidechain rebuild protocol that assesses Rosetta's ability to sample and discriminate among various sidechain conformations. The primary rotamer recovery protocol used in this paper was the RTMin protocol. In combination with the curated data set described below, this protocol gives maximum emphasis on energy function errors while reducing effects of sampling errors. RTMin rebuilds each sidechain one-at-a-time in native context while sampling every available rotamer and the native conformation. Additionally, off-center torsions were sampled at one standard deviation to either side of a rotamer well center for the first two chis of each residue. The lowest scoring conformation is chosen and compared to the native conformation.

Conformation recovery is determined by difference in correlation to electron density data between the native conformation and the chosen conformation. The threshold between recovery and error was determined to be a correlation difference of 0.13 by running RTMin rotamer recovery on a large set of structures and manually inspecting the results in PyMOL. The correlation score is normalized across all residue types, so the same threshold applies equally. When compared to using difference in chi torsions as a metric to determine recovery, use of electron density data corrects many mistakes made when using chi difference and is much more sensitive to subtle errors without increasing false positives.

It is important to have the best signal-to-noise ratio possible when uncovering energy function issues. By allowing Rosetta to sample the entire rotamer library for a residue in native context, undersampling effects are largely mitigated. Sampling

sidechain torsions one standard deviation to either side of rotamer well centers further reduces undersampling effects. Additionally, the input data was filtered so Rosetta only rebuilds well-defined sidechains. Eligible sidechains for rotamer recovery are determined by filtering with minimum electron density correlation of 0.72 and maximum b-factor of 30. As above with the rotamer recovery threshold, the minimum electron density correlation was also determined via manual inspection using PyMOL. Partially defined sidechains were allowed as long as at least one chi torsion was defined by electron density. Sidechains making crystal contacts or with multiple pdb conformations are eliminated as well.

For a closer assessment of normal Rosetta use, the PackRotamers protocol was performed. Instead of greedy sampling and minimization of individual sidechains, PackRotamers does stochastic combinatorial rotamer sampling across all sidechains without minimization. Off-rotamer torsions at one standard deviation to either side of each rotamer chi center were also included. In contrast to RTMin, native sidechain conformations were not considered during sampling. While all sidechains were included in sampling, only the well-defined sidechains used by the RTMin protocol were included in the recovery assessment.

Kullback-Leibler Divergance

The native distribution was taken directly from the same data set used for rotamer recovery. The distribution consisted of counts for each unique rotamer, where a rotamer is defined by the combination of chi angles rounded to the mean of the rotamer well they occupy. Nonrotameric chis are not considered; only the

rotameric chis preceding the nonrotameric chis are counted. For each parameter set, the PackRotamers protocol was performed and counts gathered from the ensuing distribution. Once the counts were obtained for each distribution, the standard Kullback-Leibler divergence calculation was computed for both per-residue counts and overall counts.

Batchrelax Decoy Discrimination

The same decoy discrimination protocol described in Chapter 2 was performed with an emphasis on reducing compute time. First, the input set was reduced from 4500 structures to 1000 where the RMS range was still evenly represented. Next, Batchrelax was performed in place of Fastrelax. Batchrelax is a competitive optimization algorithm that culls the population of structures each time the ramp-repack-min cycle is repeated. Each batch has 20 structures and are grouped by starting energy. The culling operation eliminates the highest-energy 20% of the remaining batch. For 1000 structures per PDB target, we performed Batchrelax on 50 batches and merged the results before computing the discrimination score.

Bibliography

1. Anfinsen CB, Haber E (1961) Studies on the reduction and re-formation of protein disulfide bonds. J. Biol. Chem. [Internet] 236:1361–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/13683523

2. Levinthal C How To Fold Graciously. In: Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois. ; 1969. pp. 22–24.

3. Levitt M, Lifson S (1969) Refinement of protein conformations using a macromolecular energy minimization procedure. J. Mol. Biol. [Internet] 46:269–279. Available from: http://linkinghub.elsevier.com/retrieve/pii/0022283669904215

4. Levitt M, Warshei A Computer simulation of protein folding. 253.

5. McCammon J, Karplus M (1977 [cited 2014]) Dynamics of folded proteins. Nature [Internet]. Available from: http://web.chem.ucsb.edu/~cwu/gallery/PaperPDB/MD1.pdf

6. Li Z, Scheraga H (1987 [cited 2014]) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. … Natl. Acad. Sci. [Internet] 84:6611–6615. Available from: http://www.pnas.org/content/84/19/6611.short

7. Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, Meiler J (2010 [cited 2014]) Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry [Internet] 49:2987–98. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2850155&tool=pmcentrez&rendertype=abstract

8. Brooks BR, Bruccoleri RE, Olafson DJ, States DJ, Swaminathan S, Karplus M (1983) CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J. Comput. Chem. 4:187–217.

9. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman P a. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J. Am. Chem. Soc. [Internet] 117:5179–5197. Available from: http://pubs.acs.org/doi/abs/10.1021/ja00124a002

10. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, et al. (2013 [cited 2014]) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics [Internet] 29:845–54. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3605599&tool=pmcentrez&rendertype=abstract

11. Lazaridis T, Karplus M (1999) Effective Energy Function for Proteins in Solution. 152:133–152.

12. Bashford D, Case D a (2000) Generalized born models of macromolecular solvation effects. Annu. Rev. Phys. Chem. [Internet] 51:129–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15048829

13. Sippl M (1990) Calculation of conformational ensemble from potentials of mean force. J. Mol. Biol 213:859–883.

14. Finkelstein A, Badretdinov Ay, Gutin A (1995) Why do protein architectures have Boltzmann-like statistics? Proteins 2:142–150.

15. Simons KT, 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. [Internet] 268:209–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9149153

16. Shen M, Sali A (2006 [cited 2014]) Statistical potential for assessment and prediction of protein structures. Protein Sci. [Internet]:2507–2524. Available from: http://onlinelibrary.wiley.com/doi/10.1110/ps.062416606/full

17. Skolnick J (2006 [cited 2014]) In quest of an empirical potential for protein structure prediction. Curr. Opin. Struct. Biol. [Internet] 16:166–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16524716

18. Skolnick J, Jaroszewski L, Kolinski a, Godzik a (1997) Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci. [Internet] 6:676–88. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2143667&tool=pmcentrez&rendertype=abstract

19. Song Y, Tyka M (2011 [cited 2013]) Structure-guided forcefield optimization. Proteins Struct. … [Internet] 79:1898–1909. Available from: http://onlinelibrary.wiley.com/doi/10.1002/prot.23013/full

20. Samudrala R, Moult J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. [Internet] 275:895–916. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9480776

21. Zhou H, Zhou Y (2002 [cited 2014]) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection

and stability prediction. Protein Sci. [Internet]:2714–2726. Available from: http://onlinelibrary.wiley.com/doi/10.1110/ps.0217002/full

22. Zhou H, Skolnick J (2011 [cited 2014]) GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys. J. [Internet] 101:2043–52. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3192975&tool=pmcentrez&rendertype=abstract

23. Leaver-Fay A, O'Meara M (2013 [cited 2013]) Scientific benchmarks for guiding macromolecular energy function improvement. Methods … [Internet]:109–143. Available from: http://kinemage.biochem.duke.edu/downloads/PDFs/2013Leaver-Fay_OptE_MethEnz523.pdf

24. Das R (2011 [cited 2013]) Four small puzzles that Rosetta doesn't solve. PLoS One [Internet] 6. Available from: http://dx.plos.org/10.1371/journal.pone.0020044

25. Bowman GR, Pande VS (2009 [cited 2014]) Simulated tempering yields insight into the low-resolution Rosetta scoring functions. Proteins [Internet] 74:777–88. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18767152

26. Tyka MD, Jung K, Baker D (2012 [cited 2012]) Efficient sampling of protein conformational space using fast loop building and batch minimization on highly parallel computers. J. Comput. Chem. [Internet] 33:2483–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22847521

27. Fleishman SJ, Whitehead T a, Ekiert DC, Dreyfus C, Corn JE, Strauch E-M, Wilson I a, Baker D (2011 [cited 2013]) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science [Internet] 332:816–21. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3164876&tool=pmcentrez&rendertype=abstract

28. Wang J, Cieplak P, Kollman P a. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J. Comput. Chem. [Internet] 21:1049–1074. Available from: http://doi.wiley.com/10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F

29. Brooks BR, Brooks III CL, Mackerell Jr. AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, et al. (2009) CHARMM: The Biomolecular Simulation Program. J. Comput. Chem. 30:1545–1614.

30. Abagyan RA, Mazur AK (1989) New Methodology for Computer-Aided Modelling of Biomolecular Structure and Dynamics 1. Non-Cyclic Structures. J. Biomol. Struct.

Dyn. [Internet] 6:815–832. Available from: http://dx.doi.org/10.1080/07391102.1989.10507739

31. Mazur AK, Dorofeev VE, Abagyan RA (1991) Derivation and testing of explicit equations of motion for polymers described by internal coordinates. J. Comput. Phys. [Internet] 92:261–272. Available from: citeulike-article-id:10750684

32. Rodriguez G, Jain A, Vaidehi N (1993) A fast recursive algorithm for molecular dynamics simulation. J. Comput. Phys. [Internet] 106:258–268. Available from: http://www.sciencedirect.com/science/article/pii/S002199918371106X

33. Das R, Baker D (2008 [cited 2013]) Macromolecular modeling with rosetta. Annu. Rev. Biochem. [Internet] 77:363–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18410248

34. Abagyan R (1994 [cited 2013]) ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J. Comput. … [Internet]. Available from: http://onlinelibrary.wiley.com/doi/10.1002/jcc.540150503/full

35. Yu H, van Gunsteren WF (2004 [cited 2013]) Charge-on-spring polarizable water models revisited: from water clusters to liquid water to ice. J. Chem. Phys. [Internet] 121:9549–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15538877

36. Patel S, Brooks CL (2004) CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. J. Comput. Chem. [Internet] 25:1–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14634989

37. Liu D, Nocedal J (1989 [cited 2013]) On the limited memory BFGS method for large scale optimization. Math. Program. [Internet] 45:503–528. Available from: http://link.springer.com/article/10.1007/BF01589116

38. DiMaio F, Echols N, Headd J, Terwilliger T, Adams P BD (2013) Improved protein crystal structures at low resolution by integrated refinement with Phenix and Rosetta. Rev.

39. Tyka MD, Keedy D a, André I, Dimaio F, Song Y, Richardson DC, Richardson JS, Baker D (2011 [cited 2011]) Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol. Biol. [Internet] 405:607–18. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3046547&tool=pmcentrez&rendertype=abstract

40. Rohl CA, Strauss CEM, Misura KMS, Baker D Protein Structure Prediction Using Rosetta. In: Enzymology LB and MLJBT-M in, editor. Numerical Computer Methods,

Part D. Vol. Volume 383. Academic Press; 2004. pp. 66–93. Available from: http://www.sciencedirect.com/science/article/pii/S0076687904830040

41. Chen VB, Arendall WB, Headd JJ, Keedy D a, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010 [cited 2012]) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D. Biol. Crystallogr. [Internet] 66:12–21. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2803126&tool=pmcentrez&rendertype=abstract

42. Khatib F, Cooper S, Tyka MD, Xu K, Makedon I, Popovic Z, Baker D, Players F (2011 [cited 2011]) Algorithm discovery by protein folding game players. Proc. Natl. Acad. Sci. [Internet]:1–5. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1115898108

43. Leaver-fay A, Meara MO, Kuhlman B, Thompson J, Snoeyink J (2012) Scientific Benchmarks for Updating the Rosetta Energy Function. Biomol. Eng.:1–53.

44. Adams PD, Afonine P V, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, et al. (2010 [cited 2012]) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D. Biol. Crystallogr. [Internet] 66:213–21. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2815670&tool=pmcentrez&rendertype=abstract

45. Shapovalov M V, Dunbrack RL (2011 [cited 2013]) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure [Internet] 19:844–58. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3118414&tool=pmcentrez&rendertype=abstract

46. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith C a, Sheffler W, et al. (2011 [cited 2014]) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. [Internet] 487:545–74. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4083816&tool=pmcentrez&rendertype=abstract