Title
A Monte Carlo Method for Generating Side Chain Structural Ensembles

Permalink
https://escholarship.org/uc/item/1g16z2db

Journal
Structure, 23(1)

ISSN
1359-0278

Authors
Bhowmick, Asmit
Head-Gordon, Teresa

Publication Date
2015

DOI
10.1016/j.str.2014.10.011

Peer reviewed

A Monte Carlo Method for Generating Side Chain Structural Ensembles

Asmit Bhowmick[1] and Teresa Head-Gordon[1,2,3,4*]
[1]Department of Chemical and Biomolecular Engineering, [2]Department of Chemistry, [3]Department of Bioengineering, University of California Berkeley
[4]Chemical Sciences Division, Lawrence Berkeley National Labs
Berkeley, California 94720, USA

We present a new Monte Carlo side chain entropy (MC-SCE) method that uses a physical energy function inclusive of long-range electrostatics and hydrophobic potential of mean force, coupled with both backbone variations and a backbone dependent side chain rotamer library, to describe protein conformational ensembles. Using the MC-SCE method in conjunction with backbone variability, we can reliably determine the side chain rotamer populations derived from both room temperature and cryogenically cooled X-ray crystallographic structures for CypA and H-Ras and NMR J-coupling constants for CypA, Eglin-C, and the DHFR product binary complexes E:THF and E:FOL. Furthermore, we obtain near perfect discrimination between a protein's native state ensemble and ensembles of misfolded structures for 55 different proteins, thereby generating far more competitive side chain packings for all of these proteins and their misfolded states.

***Corresponding author**
Stanley Hall 274
510-666-2744 (V)
thg@berkeley.edu

Structure, in press 2014

INTRODUCTION

Anfinsen’s thermodynamic hypothesis(Anfinsen, 1973) states that the native protein ensemble resides in a global minimum free energy basin that defines its functional state whether it be binding, catalysis, or signaling. This has been traditionally interpreted as a free energy basin dominated by O(~1) unique conformations, an interpretation heavily influenced by X-ray crystallographic protein structures that have proven to be invaluable for providing functional insight. Nonetheless, the perspective of considering just one native conformation opposes evidence that proteins are highly flexible(Kohn et al., 2010), especially at the level of backbone displacements(Friedland et al., 2008) that aid side chain packing rearrangements(Fenwick et al., 2014; Moorman et al., 2012; Schnell et al., 2004; Tzeng and Kalodimos, 2012). For example, new analysis of weak electron density features in X-ray crystallographic data has shown that a large percentage of PDB structures have alternate rotameric side chains(Lang et al., 2014; Lang et al., 2010). Furthermore, X-ray crystallographic structures that are cryogenically cooled also tend to overemphasize a level of uniqueness in native state structures that are too small and overpacked, and miss important catalytic side chain conformers that are present in room temperature crystallographic data (Fraser et al., 2009; Fraser et al., 2011).

The thermodynamic manifestation of conformational flexibility is encompassed in entropic effects(Baldwin and Rose, 2013), with statistical fluctuations of side chain packing arrangements playing a dominant role. NMR groups have made quantitative progress on equating Lipari-Szabo order parameters, S^2, to conformational entropy for both the backbone and side chains(Fenwick et al., 2014; Lee et al., 2000; Mittermaier et al., 1999; Stone, 2001). For example, NMR experiments on calmodulin(Frederick et al., 2007) and CAP(Tzeng and Kalodimos, 2012) proteins have shed light on this ‘residual’ free energy arising from the alternate conformations a side chain can take. A good percentage of side chains were found to have the side chain order parameter in the range $0.3<S^2<0.7$ which indicates that these side chains may be populating alternate rotameric wells on the nanosecond-microsecond timescale, although the fast motions measured by S^2 are not always probing side chain rotamer transitions(Tuttle et al., 2013). Instead, three bond J-coupling constants $^3J_{C\gamma N}$ and $^3J_{C\gamma CO}$ that report on χ_1 dihedral angle fluctuations in the broad picosecond-millisecond timescale have enabled quantitative estimation of different rotamer populations in solution(Tuttle et al., 2013). In addition, recent work using relaxation experiments have also highlighted the dynamic nature of side chains up to the millisecond, and longer, timescale (Farès et al., 2009; Henzler-Wildman et al., 2007).

Although it is true that the conformational flexibility of an unfolded protein compared to a folded protein is increased, numerical studies have shown that the number of possible ways of packing side chains on the backbone of a folded protein is by no means small or unique. Zhang and Liu reported that the total number of self-avoiding (i.e. with just hard sphere interactions) side chain conformations for the 17-residue protein 1ebx is of the order of 10^{11} (Zhang and Liu, 2006), and this number would be expected to be larger for larger proteins. However theoretical approaches for sampling the low energy alternative side chain arrangements of a protein is a difficult problem, and while molecular dynamics (MD) simulations give a good description of side chain conformational change on the nanosecond to sub-microsecond level(Li and Brüschweiler, 2009), the experimental estimates indicate that the timescales are much longer. While it is true that distributed computing paradigms such as Folding@Home (Shirts and Pande, 2000) and special purpose hardware like the Anton computer (Shaw et al., 2008) can reach the millisecond timescale for MD, we assert that computing the side chain populations and the thermodynamic entropy for tens to hundreds of native proteins and hundreds of their misfolded states, ad we have done in this study, is well beyond a comfortable scale for MD even using these two powerhouse computing platforms.

Therefore to circumvent the sampling issues imposed by MD, many groups have resorted to advanced Monte Carlo (MC) schemes(DuBay and Geissler, 2009; Friedland et al., 2008; Zhang and Liu, 2006) which are designed to more exhaustively sample the Boltzmann weighted populations of side chain conformations of the protein on the NMR timescale of ns to ms or even longer. In this work we develop a new MC approach for calculating side chain entropy (SCE) by introducing several new features that make our MC-SCE method more quantitative compared to past efforts, including a better convergent Rosenbluth sampling scheme(Rosenbluth and Rosenbluth, 1955), the use of an augmented Dunbrack library(Shapovalov and Dunbrack, 2011), a very robust physics-based energy function(Lin et al., 2007; Lin and Head-Gordon, 2008, 2011), and side chain rotamer sampling on an ensemble of backbone structures.

Here we use our MC-SCE algorithm to generate ~20,000 different side chain packings for native X-ray crystal backbones, and the same number for perturbations to the backbone using short MD simulations and so-called “backrub motions” by Friedland et al.(Friedland et al., 2008), for 60 different proteins. As a first test of our MC-SCE algorithm, we use it to quantify the side chain rotamer populations on backbones derived from cryogenically cooled (CC) and room temperature (RT) X-ray crystallographic structures for CypA, and the Ser99Thr mutant (Fraser et al., 2009) and for H-Ras(Fraser et al., 2011). We also compare directly to NMR J-coupling data for CypA(Fraser et al., 2009), Eglin-C(Clarkson et al., 2006), and the DHFR binary complexes of E:THF and E:FOL (Tuttle et al., 2013). We find overall excellent agreement across the full range of X-ray and NMR data. Finally we consider alternative rotamer packings for 55 native proteins and each of the hundreds of misfolded structures from a difficult Rosetta set that exhibit near-native features in their backbone fold(Tsai et al., 2003). We use our MC-SCE approach to provide the thermodynamic functions of energy (enthalpy), side chain entropy, and free energy to discriminate the native state of a protein from its misfolded states. This large validation suite shows that we can nearly perfectly discriminate between a protein’s native state ensemble and ensembles of misfolded structures, and provide for an even more competitive decoy set with better optimized side chain packings.

RESULTS

Overview. We present results below based on a new and more robust MC side chain growth method to evaluate side chain entropy, MC-SCE, to estimate structural ensemble properties of proteins. Details are given in the Methods section, but the important points are highlighted here to better present the following results. Backbone structures are provided by either an X-ray crystal structure or a given backbone from a misfolded decoy library. Additional backbone variability on these starting structures is introduced in two independent ways: through so-called “backrub motions” (Friedland et al., 2008), which lead to small backbone RMSD with respect to the crystal structure of ~0.1-0.7 Å, and from snapshots generated from a thermalized molecular dynamics simulation with explicit solvent that lead to slightly larger RMSD changes of ~0.6-1.3 Å.

Given these different backbones, the side chains atoms beyond the C_β position are stripped away, and then all are regrown using the MC-SCE algorithm to generate an *ensemble* of ~20,000 different side chain packings, allowing us to evaluate both the side chain entropy at each residue position and rotamer populations. Table S1 provides the definition of the side chain dihedral angles sampled. One of the key features of this work is the use of well-tested physics-based energy function based on Generalized-Born electrostatics and a hydrophobic potential of mean force (Lin et al., 2007; Lin and Head-Gordon, 2008, 2011) to perform the Boltzmann weighting, and which is used to define the potential energy rank of all 20,000 structures. Here we demonstrate our ability to reliably reproduce and predict the side chain rotamer ensembles of the following class of problems: (1) cryo-cooled vs. room temperature X-ray crystallography for CypA and H-Ras, (2) both X-ray and NMR data taken on CypA, Eglin-C and the product binary complexes of DHFR, E:THF and E:FOL, and finally (3) native vs. misfolded state discrimination using a difficult Rosetta decoy set. All components of the MC-SCE approach (including the energy function) have been implemented into our in-house version of the TINKER(Ponder, 2009) software package. As an example of the cost of the MC-SCE method, we can generate a side chain ensemble of CypA (164 residues) with 20,000 structures in ~12 hours using an MPI implementation that distributes work across 16 cores; this timing uses our in-house computing cluster with the AMD Opteron(TM) Processor 6274 (2.2 Ghz) cores.

Comparison with X-ray crystallography and NMR for CypA and H-Ras. Recently, Fraser et al. found population shifts in side chain rotamer states when comparing X-ray structures obtained under cyro-cooling vs. room temperature crystallization conditions for the proteins CypA (Fraser et al., 2009; Fraser et al., 2011). Given that the backbone differences between the CC and RT structures are negligible (RMSD ~ 0.1 Å), a good test of our MC-SCE algorithm would be to determine if we can predict the major and minor side chain rotamer populations that are reported in the CC and RT crystallographic data for CypA and H-Ras.

Experiments on CypA showed that alternate side chain conformations for Arg55 and Met61 were found with RINGER in the CC data, and additional side chain rotamer changes were evident for Leu98, Ser99 and Phe113 in the RT data, helping to explain the catalytically competent and incompetent conformations of the active site residues (Fraser et al., 2009). Table 1 reports the CC and RT X-ray experimental χ rotamers and their populations and the corresponding MC-SCE values for WT CypA and the Ser99Thr mutant. The MC-SCE calculations were done on the CC backbone, as well as an average over two RT backbones based on so-called major and minor conformers reported for the room temperature crystal structure (RT-M or RT-m). The 20,000 structures of the generated side chain packing ensemble for each backbone allow us to report MC-SCE population percentages. We also averaged over the 20,000 side chain ensembles generated for each backbone relevant to RT backbone variations: two backrub ensembles of 10 structures each based on the starting RT-M and RT-m backbones, and 3 backbones generated from MD snapshots at 0.2 ns, 2.0 ns and 4.0 ns. For side chain conformations predicted from the MC-SCE algorithm, the χ rotamers were binned as is done conventionally with bin centers on 60°, 180° and -60.

When performed on the CC X-ray backbone, our MC-SCE method predicts the same major conformer for residues Leu98(χ_1), Phe113(χ_1), Arg55(χ_3), and Met61(χ_2), as well as detecting the minor rotamer states for the latter two residues that was found from the RINGER analysis of weak electron density features. When performed on the RT X-ray backbone, our MC-SCE method also predicts the major and minor conformer for all four same residues. Furthermore, we determined an increase in SCE (using Eq. 7) when going from the CC to RT backbone as was observed in (Fraser et al., 2011), in which the RT backbone allows for greater conformational flexibility of the side chains. Even better agreement with reported X-ray rotamer populations is found with a thermalized backbone (i.e. side chains grown on backrub and MD backbone ensembles) for these same residues as shown in Table 1. We also perform our MC-SCE calculations on the Ser99Thr mutant, which through active site interactions stabilizes the minor rotamers for Phe113(χ_1), Arg55(χ_3), and Met61(χ_2) compared to the WT form, which is exactly what we observe in our simulations (Table 1).

In all cases, regardless of method for creating the backbone, we do not predict the 180° rotamer for Ser99(χ_1), and we do not find the same 180° dominant rotamer for the Thr99(χ_1) mutant (although we do predict it as a minor conformation). One possibility is that the energy function, and possibly the use of an implicit solvent model for water, accounts for this discrepancy, although our energy function with implicit solvent has been extensively validated(Lin et al., 2007; Lin and Head-Gordon, 2008, 2011). When we perform MD with explicit solvent using the CC crystal as the start state, the 180° rotamer flips to the 60° rotamer and maintains that value for the entirety of the simulation run. Hence the very different energy functions and sampling methods (implicit vs. explicit solvent and MC vs. MD) favors an alternate rotamer to the major rotamer seen experimentally. Therefore we believe that overall the energy function used with MC-SCE is performing well. The fact that we are able to correctly predict the change in rotameric states for Phe113(χ_1), Arg55(χ_3), and Met61(χ_2) when going from WT to the SerThr99 mutant for CypA, indicates that the adoption of the 180° rotamer at position 99 for WT and mutant CypA is not necessary, suggesting that we are seeing a similar cooperative network effect among residues that were analyzed in the NMR relaxation experiments(Fraser et al., 2009).

To provide for better contact with NMR solution data for CypA and the Ser99Thr mutant, Table 2 reports $^3J_{C\gamma N}$ and $^3J_{C\gamma C}$ values evaluated from our MC-SCE ensemble populations and compared to the same experimentally measured values for various aromatic residues(Fraser et al., 2009). We evaluated our J-couplings using the Karplus equation parameterization values found in both (Schmidt 2007) and (Tuttle et al., 2013), and they are also reported in Table 2. To put the comparison in some context, we also calculate the difference between the experimental J-coupling, $\bar{J}_{XY}^{(i,\exp)}$ and the average scalar coupling calculated from a given MC-SCE structural ensemble, $\bar{J}_{XY}^{(i)}$ for each residue (Eq. (9)), normalizing it by the uncertainty of the Karplus parameters and any experimental error, to generate χ_J^2 values (Eq. (10)). We use a conservative uncertainty value due to the Karplus equation of σ_J = 0.5 Hz for both types of scalar couplings, estimated from the difference in calculated J-couplings using the two Karplus equation parameterizations that use the same underlying structural ensemble. Any dominant error due to the underlying structural ensembles themselves would then correspond to values of $\chi_J^2 > 1$. Our calculated deviations are $\chi^2_{JC\gamma N}$ =0.53 and $\chi^2_{JC\gamma C}$ =0.99 indicating that the underlying structural ensembles are sound. As such, we also observe a change in J-coupling values for Phe113 which confirms a switch in rotameric state from 60° to -60° as per the experiment.

Table 3 reports the CC and RT X-ray crystallographic and MC-SCE generated side chain χ rotamers and their populations for H-Ras. Again, the MC-SCE calculations were done on the CC backbone and its backrub variation, and the 20,000 structures of the generated side chain packing ensembles allow us to report MC-SCE population percentages. However the RT variations of the reported

9 individual side chains involved more than two rotameric states, and in combination would result in a large combinatorial number of RT crystal backbones that are inconvenient for performing the backrub motions. Instead we represent backbone variability using 3 MD snapshots at 0.2 ns, 2.0 ns and 4.0 ns to analyze the higher temperature data, given its consistency with backrub motions for CypA and for the Rosetta data sets described further below.

When performed on the CC X-ray backbone, or its backrub variant, our MC-SCE method predicts the same major conformer for residues Asp30(χ1), Glu62(χ1), Ser65(χ1), His94(χ1), Val103(χ1), Arg97(χ3), Glu98(χ2), and Gln99(χ2), with the only exception being Gln61(χ2) in which the MC-SCE algorithm predicts it to be a minor (up to 25%) population. What is most interesting is that the MC-SCE method using the CC backbone can also determine the minor side chain conformation detected in the RT crystal structure for Glu62(χ1), His94(χ1), Val103(χ1), Arg97(χ3), Glu98(χ2), as well as Gln61(χ2) which samples all three rotameric states. This suggests that the cryogenic backbones are not completely deficient for accommodating alternate rotamers, but apparently their electron density features are either far too weak to detect, or possibly that crystalline contact interactions favor certain rotamers. The MD results are also interesting, showing the time evolution of the rotamer populations for these residues as the backbone varies, flipping between the major and minor rotamer states.

However, although the MC-SCE does predict the major conformer, it does not predict the alternative rotamer preference observed in the RT X-ray data for either Asp30(χ1) or Ser65(χ1) on any backbone. Given that these residues are surface residues, they may be more prone to crystal packing artifacts that bias the populations of a particular rotamer class. Figure 1 shows that there are stabilizing interactions for these two residues with the surrounding lattice that favor the RT major rotamer that is experimentally observed; in particular Asp30 interacts with arginine and Ser65 shows very close approach to glutamic acid. Since we do not represent the crystal lattice, these favorable hydrogen-bonding interactions would not be present, and thus would not preferentially stabilize the experimentally observed RT major rotamer.

Comparison to NMR data for Eglin-C, E:THF, and E:FOL. We next analyze the MC-SCE approach against solution-based NMR scalar coupling constants $^3J_{C\gamma N}$ and $^3J_{C\gamma C}$ generated by Clarkson et al on Eglin-C (Clarkson et al., 2006) and by Tuttle and co-workers for the DHFR binary product complexes E:THF and E:FOL (Tuttle et al., 2013). To calculate the scalar coupling constants, we again use the standard Karplus equation, Eq. (8), with Karplus parameters from (Tuttle et al., 2013). Figures 2 shows the agreement between the experimental coupling values and the $\overline{J}^{(i)}_{XY}$ generated from the MC-SCE ensembles, taken on both the CC and MD backbones, for C-Eglin. The overall χ_J^2 values for $^3J_{C\gamma N}$ is 0.36 and for $^3J_{C\gamma C}$ is 0.84 on the CC backbone, and these values change to $\chi^2_{JC\gamma N}$=0.20 and $\chi^2_{JC\gamma C}$=1.44 on the averaged molecular dynamics backbones, indicating that the structural ensembles are in overall good agreement with the rotamer populations for the 12 residues. Table S2 in the Supplementary materials provides a more detailed rotamer assignment for Eglin-C, and we note that although our J-coupling values for two of the residues, Thr 17 and Thr 26, are in excellent agreement with the experimental measurements, we do not agree with the experimental study in the assigned rotameric populations, suggesting that the experimental rotameric populations may be flawed.

We next consider the J-coupling constants for E:THF and E:FOL, requiring us to develop parameters for the bound ligand on which the NMR data was taken; the introduction of the ligand means that we can’t generate backrub ensembles from the server(Friedland et al., 2008), and hence we use MD data to provide for backbone variations. Figures 3 and 4 show the agreement between the experimental coupling values and the $\overline{J}^{(i)}_{XY}$ generated from the MC-SCE ensembles, taken on the averaged MD backbones, for the E:THF complex and E:FOL complex, respectively. The overall χ_J^2 values on the CC backbone is small ($\chi^2_{JC\gamma N}$ = 0.45 to 0.64) for both proteins, while the deviation in $^3J_{C\gamma C}$ is larger when measured on the MD generated backbones ($\chi^2_{JC\gamma C}$ = 2.71 to 3.06). The large χ_J^2 value for the $^3J_{C\gamma C}$ coupling for E:THF is due to genuine disagreement for what is the major rotamer for only three residues: Val40, His114, Thr123, although for His114 we find it to be a minor rotamer instead (Table S3). For the DHFR complex E:FOL we again find disagreement for the major rotamer for two residues: Val10, Val40, and Thr123. It is noteworthy that Val40 and Thr123 are among one of the few residues that have different major rotamers in the multiple DHFR complexes studied in (Tuttle et al., 2013).

For E:THF the MC-SCE structural ensembles show overall very good agreement across 46 of the 49 residue NMR measurements, with $\chi^2_{JC\gamma N}$ = 0.65 and $\chi^2_{JC\gamma C}$ = 1.62, in which the major rotamer is correctly selected for all of these residues. For E:FOL the MC-SCE structural ensembles show overall very good agreement across 20 of the 22 residue NMR measurements, with $\chi^2_{JC\gamma N}$ = 0.23 and $\chi^2_{JC\gamma C}$ = 1.33, in which the major rotamer is correctly selected for all of these residues. Problems in the structural ensembles that gives rise to disagreement with the $^3J_{C\gamma C}$ measurement for the two protein complexes are due to differences in the assignment of the minor rotamer for a smaller subset of residues, i.e. no minor rotamer detected, detected with a much smaller population, or assignment of a different minor rotamer state (Table S4).

Discrimination of native folded vs. misfolded states. Given the very good agreement between X-ray and NMR data and the MC-SCE ensembles, we next test our ability for selecting native state structures when compared to the 2007 Rosetta decoy set(Qian et al., 2007) generated by the popular fragment assembly folding program ROSETTA(Leaver-Fay et al., 2011). We define a traditional energy rank, E_{single}, as the energy of the *provided* side chain packing on a given backbone which is either the native X-ray PDB structure or the Rosetta decoy structures. Next we consider a free energy rank, F, based on ensembles of alternative side chain packings for the given backbones for all 55 native proteins and misfolded structures using our MC-SCE method that evaluates the side chain entropy. This free energy function is defined as

$$F = E_{best} - TS_{SC} \tag{1}$$

where E_{best} is the structure whose side chain packing for a given backbone (native or decoy) is the lowest energy in the generated ensemble, and $-TS_{SC}$ is the temperature weighted side chain entropy (Eq. (7) in Methods). We also judge the quality of these thermodynamic metrics through calculation of a Z-score, the free energy (or E_{single} or E_{best}) difference between the native state quantity and the same quantity averaged over the misfolded states. A larger value of the Z-score signals better separation of the native structure from the misfolded conformers. All detailed data is reported in Table S4 in the SI material, in order for us to highlight the important points here.

The traditional rank based on our physics-based energy function, E_{single}, does a very good job of discrimination of the given native state from all of the members of a given decoy set, in which 40/55 proteins are ranked 1st with a Z-score of -3.76 for this subset (-2.95 over all proteins). Our energy function comfortably outperforms many recent popular statistical potentials like DFIRE(Zhou and Zhou, 2002) (21/58), DOPE(Shen and Sali, 2006) (21/58), and EPAD(Zhao and Xu, 2012) (34/58), and is competitive with other reported energy functions like EPAD2(Zhao and Xu, 2012) (46/58) and PM6 (Faver et al., 2011) (49/49).

However, the free energy is the true thermodynamic quantity, and given that our MC-SCE algorithm can generate an ensemble of side chain conformer packings, we compare the native side chain ensembles and the respective decoy ensembles, based on the evaluation of the free energy, F, using Eq. (1). Using the free energy thermodynamic metric, the absolute native state discrimination improves modestly to 42/55 natives identified (Z-score for natives of -3.62), with the Z-score over all proteins improving slightly to -3.07. Even so, for 8 proteins whose native states were not selected, the ensemble F rank improved native state ranking, significantly in most cases, compared to using E_{single}: 1ail (rank 62 to

3), 1c8c (rank 47 to 2), 1enh (rank 81 to 13), 1hz6 (rank 7 to 3), 1rnb (rank 93 to 89), 1utg (rank 94 to 75), 1vcc (rank 4 to 2), 1ubi (rank 9 to 5), while 1pgx and 1dhn were 2nd ranked by either single or thermodynamic ensemble metrics (Table S4). A breakdown of the free energy shows that selection of the native conformation using the F rank is largely driven by E_{best}, since the Z-score based on the side chain ensemble best energy alone is lowered to -3.94 for all native states selected, and -3.27 over all proteins. This clearly indicates that the original native PDB structure and provided Rosetta misfolded structures have not optimized side chain arrangements for the given backbone. Furthermore, these lower energy side chain packings are providing sharper discrimination of folded vs. misfolded states. These results are consistent with a number of recent studies that have shown that weak features in the electron density maps from X-ray protein crystallography support alternate side chain packings that differ from the original reported side-chain rotamers (Fenwick et al., 2014; Fraser et al., 2011; Lang et al., 2014; Lang et al., 2010; Tyka et al., 2011).

In order to push toward better native state discrimination, we also considered additional ensemble characterizations involving the native state backbone, with the expectation that small perturbations to the backbone might allow for new side chain rotamer packings. These backbone changes may remove overly unique side chain rotamer states that arise from cryo-cooling (Fraser et al., 2009; Fraser et al., 2011), as well as crystal contacts, oligomeric packing, or ligand-binding interactions(Tyka et al., 2011). For example, 1ail has been crystallized as a dimer, while 1c8c has a bound peptide, and thus are illustrative of perhaps why many of their decoys, generated independently from the original crystallization conditions but with near-native features, are energetically better than the crystallized native state (Tyka et al., 2011).

Therefore to test how the backbone perturbations influence the free energy ranking, we used backrub motions (Friedland et al., 2008) that minimize repositioning of the backbone, but which can drastically affect side chain rotamer populations due to reorientation of the C_β atoms, for the 13 proteins in which the native state was not selected or very poorly predicted by the free energy function. We performed backrub motions on the X-ray backbone for each of these proteins, generating 10 different backrub structures. We then removed the side chains atoms beyond the C_β position for each, and then used our MC-SCE approach to generate side chain packing ensembles, in order to calculate thermodynamic rankings using E_{best}, and the free energy F, and their corresponding Z-scores (Table 4). In addition we also do the same MC-SCE procedure for backbones derived at the end of a short molecular dynamics simulation in explicit water at ambient temperature and pressure as an independent way to relax the crystalline constraints of the X-ray native structure. In both cases the native backbones were found to change by a little less than 1.0Å RMSD compared to their PDB structure, on average. The relative RMSD of the final thermalized native backbone with respect to the decoy set was unchanged on average, i.e. making the decoys no more or no less competitive for determining the native state ensemble.

The resulting drastic improvement in ranking– 53/55 proteins native states are now well distinguished from the misfolds– suggests that the initial failure of the free energy to identify the native state cannot be attributed primarily to the limitation of the energy function or MC-SCE sampling protocol. Instead, the small changes in backbone flexibility, consequences of which were also examined by Tyka and co-workers(Tyka et al., 2011), highlights the sensitivity of SCE to subtle effects of the backbone configuration, which improved the discrimination for 11 of the 13 problematic proteins. Since the native state is selected for in ~96% proteins of the best available Rosetta decoy set, considered to be a challenging test of any new sampling method, statistical potential, physical force field or scoring function, MC-SCE appears to provide an excellent standard for native state prediction. Two proteins for which we did not discriminate for the native were 1hz6 (whose rank remained 3rd whether using the PDB or MD backbone) and 1utg (whose 75th native rank with the PDB structure rose to 10th with the backrub motions), and would require more careful consideration of available NMR data.

In order to check the similarity between the best energy native structure in our free energy ensemble with the deposited PDB crystal structure, the χ_1 torsional angles between the 2 structures were compared for each of the 55 proteins we analyzed. A residue was said to have had a change in torsional angle if the absolute value of their difference exceeded 40°, which is similar to the convention adopted by Bower and co-workers(Bower et al., 1997), and the fraction of the total residues that changed the χ_1 angle is listed in the final column of Table S4. On an average, our MC-SCE algorithm found an alternate χ_1 dihedral angle in the best free energy native structure compared to the crystal structure 25% of the time, consistent with the ~18% of alternate side chain rotamers on reexamination of electron density from 402 high resolution X-ray crystal structures(Lang et al., 2010). Since Lang and co-workers only considered unbranched side chains in their electron density analysis, as well as ignoring density fitting with combinations of $\chi_{1,}\chi_{2,}\chi_3$ etc., it would likely explain the quantitative discrepancy with what we have found since we considered all residues and the full rotameric set of χ values for any given amino acid. Figure 5 shows the typical distribution of side chain entropy on the 2CHF PDB backbone, where the side chain conformations that showed most variability did not exclusively select surface residues, but core positions as well.

CONCLUSIONS

In summary, we have introduced a new MC-SCE algorithm for generating side chain packing ensembles, allowing us to predict side chain rotamer populations and side chain entropy, which we have compared to extensive data sets from both NMR and X-ray crystallographic experiments. We have validated our approach by making direct contact with X-ray crystallography and NMR data on side chain rotamer populations for CypA and its Ser99Thr mutant(Fraser et al., 2009), HRas(Fraser et al., 2011), Eglin-C(Clarkson et al., 2006), and the DHFR complexes E:THF and E:FOL (Tuttle et al., 2013). For all proteins we find overall excellent agreement of rotamer values, their populations, and calculated J-couplings when compared to crystallographic data and with NMR experimental J-couplings.

We have shown that the side chain populations measured depend significantly on the given backbone structure, and hence our MC-SCE technique is aided by introducing small deviations (~1.0 Å

RMSD) from the crystallographic backbone structures using both backrub motions and thermalized explicit solvent molecular dynamics simulations. For the case of CypA and its Ser99Thr mutant we found all of the major and minor rotamers of all reported residues except for Ser(Thr)99. However, it had no discernable influence on our successful ability to predict the Phe113 catalytic rotameric state for WT as well as stabilizing the minor rotamers for Phe113(χ_1), Arg55(χ_3), and Met61(χ_2) in the active site of the mutant form(Fraser et al., 2009).

For the protein H-Ras, we found that we can detect the minor or alternate rotamer state of a sidechain when the ensemble is generated on the CC backbone and its backrub variants, although the experimental density is only evident in the RT X-ray data(Fraser et al., 2011). In addition, we do not observe the same minor rotameric states that are experimentally found for Asp30 and Ser65. In both cases, i.e. our ability to detect new rotamers on the CC backbone or observing alternate minor rotamers to that found in the RT data, can be explained by the fact that the surrounding crystal lattice is not present in our approach. Previous work has shown that stabilizing packing interactions often arise from polar-polar interactions with the surrounding crystal lattice, and thus can influence the experimentally observed rotamer populations(Dasgupta et al., 1997). We found that such specific interactions with the surrounding lattice are present for Asp30 and Ser65, for example, and hence would not be predicted with our MC-SCE approach that instead represents aqueous solution conditions.

We have also compared our MC-SCE rotamer populations to those estimated from solution phase NMR data. Our calculated agreement with scalar coupling measurements for CypA, C-Elgin, and the two DHFR complexes E:THF and E:FOL were found to be overall excellent. The calculated scalar couplings using our MC-SCE method was well within experimental and Karplus parameter uncertainty for ${}^{3}J_{C\gamma N}$ for all four proteins, and for 85-100% of residues for the ${}^{3}J_{C\gamma C}$ measurement across the four data sets. The primary error for the E:THF and E:FOL complexes was the failure to predict the major rotamer for Val40 and Thr123, although these same residues were found to sample alternate rotameric states in the full series of DHFR complexes(Tuttle et al., 2013).

Finally, we have a highly reliable method for discrimination of native states from misfolded structures based on a difficult Rosetta decoy set. One consequence of our MC-SCE algorithm is that we find better side chain rotamer and packing representations of both the native state *and* the decoy set. This can be quantified for the decoy set by the Z-score between the PDB structure, i.e. the single backbone and side chain rotamers of the X-ray structure, and the E_{best} from the decoy set ensembles, which shrinks to –2.77. We have provided this new decoy set, Berkeley-SC-Ensemble, which we have made available at our web site http://thglab.berkeley.edu. It also includes the ensemble of new side chain packing arrangements on native PDB backbones that will be of interest to X-ray crystallographers and NMR groups.

ACKNOWLEDGMENTS

This work was supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under US Department of Energy Contract No. DE-AC02-05CH11231. We thank Dr. Po Hu for an early implementation of the MC-SCE algorithm and Dr. Alexander Scouras for helpful discussions.

MATERIALS AND METHODS

We introduce a new and more robust MC chain growth method to evaluate side chain entropy, MC-SCE, to estimate structural ensemble properties of proteins. We use an augmented Rosenbluth chain growth algorithm(Batoulis and Kremer, 1988; Rosenbluth and Rosenbluth, 1955; Zhang and Liu, 2006) to generate an ensemble of side chain packings for a given (and fixed) protein backbone. The algorithm starts with a PDB file of the enzyme, and all the side chain atoms, except the C_β atom, and any existing water molecules are eliminated. Backbone mobility is provided by a decoy library, backrub motions, or captured during a MD simulation. The side chain ensemble that can populate a provided bare backbone is then realized by growing side chains of each residue in a sequential manner with dihedral angle inputs from a backbone dependent rotamer library(Shapovalov and Dunbrack, 2011) to approximate the continuous nature of side chain dihedrals. We have augmented the rotamer library selection based on probabilities of occurrence in the PDB and by allowing for dihedral angle variations that are Gaussian distributed about a given rotamer value. All of the χ_1 and χ_2 torsional angles of all residues, except for arginine and lysine, were expanded by including a standard deviation, resulting in 3 values, χ_i and $\chi_i \pm \sigma$. After expansion, all the rotamers were further perturbed by about 0.5° to place them optimally with respect to the backbone. This is necessary because of the sensitivity of the energy function to slight changes in the protein that could distort statistics and increase the number of dead end chain growths. In our model, alanine and glycine have no dihedral degrees of freedom and hence no side chain entropy, and all residues are grown with ideal bond lengths and angles.

From the initial condition (step 0) of a bare backbone conformation *m*, for subsequent steps *i*, we develop a MC scheme whereby the residue *k* that has the lowest side chain partition function

$$Q_k = \sum_{\{v_k\}} e^{-\beta E_k^{(m,v_k)}} \tag{2}$$

is considered for the next side chain growth. For residue *k*, a side chain conformation v_k is defined by the resulting set of dihedral angles selected from the rotamer library, i.e. (χ_1, χ_2, ….). Each side chain rotamer r_k is selected according to the following probability

$$\rho_k^{(m,r_k)} = \frac{P_{r_k}^{(pdb)} e^{-\beta E_k^{(m,r_k)}}}{\sum_{\{v_k\}} P_{v_k}^{(pdb)} e^{-\beta E_k^{(m,v_k)}}} \tag{3}$$

where *{v_k}* are the possible side chain conformations for residue *k*, $E_k^{(m,r_k)}$ is the energy of interaction of side chain *k* with the backbone and all protein side chains grown so far using Eq. (3) only, and $P_{r_k}^{(pdb)}$ is the probability of the side chain conformation calculated using the values reported in the recent backbone-dependent Dunbrack library(Shapovalov and Dunbrack, 2011). The reason for including this knowledge based $P_{r_k}^{(pdb)}$ is to guide the growth process especially early on when very few side chains have been placed and to minimize picking rotamers which are known to occur infrequently in the PDB database; conformations with probability less than 0.001 in the library were ignored. Once the side chain of a residue is placed, the process is repeated until all the side chains are grown, thereby creating one complete protein structure. This complete chain growth procedure for one N-residue enzyme structure is then repeated ~20,000 times to give an ensemble of structures. Each structure *m* is then assigned a weight *W(m)* in order to get correct statistics in the canonical ensemble.

$$W(m) = e^{-\beta F_{solv}} \prod_{k=1}^{N} \frac{\sum_{\{v_k\}} P_{v_k}^{(pdb)} e^{-\beta E_k^{(m v_k)}}}{P_{r_k}^{(pdb)}} \tag{4}$$

This is defined on the basis of our chain growth probabilities as well as now including the Boltzmann factor using the GB-HPMF implicit solvent model(Lin et al., 2007; Lin and Head-Gordon, 2008, 2011). When the chain growth is unsuccessful because of unresolvable clashes, the partially grown structure is considered dead and its weight is set to zero.

The side chain entropy of a given residue *k* is evaluated using the Gibbs probabilistic definition of entropy.

$$S^{(k)} = -k_B \sum_{\{v_k\}} p_{v_k}^{(k)} \log p_{v_k}^{(k)} \tag{5}$$

where the probability $p_{v_k}^{(k)}$ of a conformational state v_k of residue *k* is calculated using the weights of the structures in the ensemble

$$p_{v_k}^{(k)} = \frac{\sum_{m=1}^{M} W(m) \delta_{r_k, v_k}^{(m)}}{\sum_{m=1}^{M} W(m)} \tag{6}$$

The sum in Eq. (6) is over all of the successful structures grown by the Rosenbluth procedure. The Kronecker delta is 1 if the side chain conformation r_k that was picked for the residue k in the m-th structure is v_k and 0 otherwise. The weights of each structure ensure that the probabilities are Boltzmann weighted. The total side chain entropy of a protein is calculated by summing over the individual entropy values

$$S_{SC} = \sum_{k}^{\#residues} S^{(k)} \tag{7}$$

NMR J-coupling calculations: Three-bond J-coupling values between the C_γ atom and the backbone carbonyl carbon ($^3J_{C\gamma CO}$) and amide nitrogen ($^3J_{C\gamma N}$) of the same residue can be calculated using

$$J_{XY} = A\cos^2(\theta+\delta) + B\cos(\theta+\delta) + C \tag{8}$$

where θ represents the dihedral angle between atoms (Y-C_α-C_β-X). The Karplus parameters (A,B,C,δ) are amino-acid specific and were taken from the original experimental sources. For Valine, ^3J values for only $C_{\gamma 1}$ have been reported in this paper.

The J-coupling value, $J_{XY}^{(k)}(m)$ for residue k in the m-th structure of our side chain ensemble was calculated from Eq. (8). These values were then used to calculate the average J-coupling value with

$$\bar{J}_{XY}^{(k)} = \frac{\sum_{m=1}^{M} W(m) J_{XY}^{(k)}(m)}{\sum_{m=1}^{M} W(m)} \tag{9}$$

where W(m) are the weights given in Eq. (4). We also calculated χ^2 values defined as

$$\chi_{XY}^2 = \frac{1}{N} \sum_{i\in\{k\}}^{N} \frac{(\bar{J}_{XY}^{(i)} - J_{XY}^{(i,\exp)})^2}{\sigma_i^2} \tag{10}$$

where N is the number of residue measurements taken. We have assumed that the primary source of experimental uncertainty is the Karplus parameters themselves; we assume an average uncertainty of σ=0.5 Hz given the differences found for these same scalar couplings for CypA.

Rosetta decoy set calculations. The single side chain native structure (the PDB) and the provided Rosetta decoy structures (with a given side chain arrangement) undergo local optimization, and are sorted in ascending order based on their energy in order to determine the E_{single} rankings. These minimized structures are then stripped of their side chains beyond the C_β position, and 20,000 alternate side chain packings with no steric clashes (which signals a failed chain growth) are generated on the native backbone and each Rosetta decoy backbone. The lowest energy structure for each ensemble is then minimized (to relax residual geometric artifacts arising from the fixed bond and bond angles assumed in the MC-SCE sampling using the rotamer library), and these minimized native and decoy structure for each protein are sorted in ascending order based on their energy in order to determine the E_{best} rankings. The side chain ensemble of structures generated for each backbone, native or decoy, shows a Gaussian distribution of energies, and we define the side chain entropy of the protein, S_{SC} in Eq. (7), based on Boltzmann weighted structures, Eq. (6), with energies below two standard deviations from the mean energy. We find that this subset of ~200 structures typically underestimates the entropy by ~5-10%, but since it is systematically applied across the protein and decoy sets, it suffices for this study.

REFERENCES

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science *181*, 223-230.

Baldwin, R.L., and Rose, G.D. (2013). Molten globules, entropy-driven conformational change and protein folding. Curr. Opin. in Struct. Bio. *23*, 4-10.

Batoulis, J., and Kremer, K. (1988). Statistical properties of biased sampling methods for long polymer chains. J. Phys. A *21*, 127.

Bower, M.J., Cohen, F.E., and Dunbrack, R.L. (1997). Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J. Mol. Biol. *267*, 1268-1282.

Clarkson, M.W., Gilmore, S.A., Edgell, M.H., and Lee, A.L. (2006). Dynamic coupling and allosteric behavior in a non-allosteric protein. Biochemistry *45*, 7693-7699.

Dasgupta, S., Iyer, G.H., Bryant, S.H., Lawrence, C.E., and Bell, J.A. (1997). Extent and Nature of Contacts Between Protein Molecules in Crystal Lattices and Between Subunits of Protein Oligomers. Proteins: Structure, Function and Genetics *28*, 494-514.

DuBay, K.H., and Geissler, P.L. (2009). Calculation of proteins' total side-chain torsional entropy and its influence on protein-ligand interactions. J. Mol. Biol. *391*, 484-497.

Farès, C., Lakomek, N.-A., Walter, K.F.a., Frank, B.T.C., Meiler, J., Becker, S., and Griesinger, C. (2009). Accessing ns-micros side chain dynamics in ubiquitin with methyl RDCs. J. Biomol. NMR *45*, 23-44.

Faver, J.C., Benson, M.L., He, X., Roberts, B.P., Wang, B., Marshall, M.S., Sherrill, C.D., and Merz, K.M. (2011). The Energy Computation Paradox and ab initio Protein Folding. Plos One *6*.

Fenwick, R.B., van den Bedem, H., Fraser, J.S., and Wright, P.E. (2014). Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. Proc Natl Acad Sci USA *111*, E445-454.

Fraser, J.S., Clarkson, M.W., Degnan, S.C., Erion, R., Kern, D., and Alber, T. (2009). Hidden alternative structures of proline isomerase essential for catalysis. Nature *462*, 669-673.

Fraser, J.S., van den Bedem, H., Samelson, A.J., Lang, P.T., Holton, J.M., Echols, N., and Alber, T. (2011). Accessing protein conformational ensembles using room-temperature X-ray crystallography. Proc. Natl. Acad. Sci. USA *108*, 16247-16252.

Frederick, K.K., Marlow, M.S., Valentine, K.G., and Wand, a.J. (2007). Conformational entropy in molecular recognition by proteins. Nature *448*, 325-329.

Friedland, G.D., Linares, A.J., Smith, C.a., and Kortemme, T. (2008). A simple model of backbone flexibility improves modeling of side-chain conformational variability. J. Mol. Biol. *380*, 757-774.

Henzler-Wildman, K.a., Lei, M., Thai, V., Kerns, S.J., Karplus, M., and Kern, D. (2007). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature *450*, 913-916.

Kohn, J.E., Afonine, P.V., Ruscio, J.Z., Adams, P.D., and Head-Gordon, T. (2010). Evidence of functional protein dynamics from X-ray crystallographic ensembles. PLoS Comput Biol *6*.

Lang, P.T., Holton, J.M., Fraser, J.S., and Alber, T. (2014). Protein structural ensembles are revealed by redefining X-ray electron density noise. Proc. Natl. Acad. Sci. USA *111*, 237-242.

Lang, P.T., Ng, H.L., Fraser, J.S., Corn, J.E., Echols, N., Sales, M., Holton, J.M., and Alber, T. (2010). Automated electron-density sampling reveals widespread conformational polymorphism in proteins. Protein Sci *19*, 1420-1431.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., *et al.* (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol *487*, 545-574.

Lee, A.L., Kinnear, S.A., and Wand, A.J. (2000). Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. Nat Struct Biol *7*, 72-77.

Li, D.-W., and Brüschweiler, R. (2009). A dictionary for protein side-chain entropies from NMR order parameters. J. Amer. Chem. Soc. *131*, 7226-7227.

Lin, M.S., Fawzi, N.L., and Head-Gordon, T. (2007). Hydrophobic potential of mean force as a solvation function for protein structure prediction. Structure *15*, 727-740.

Lin, M.S., and Head-Gordon, T. (2008). Improved energy selection of nativelike protein loops from loop decoys. J Chem Theory Comput *4*, 515-521.

Lin, M.S., and Head-Gordon, T. (2011). Reliable protein structure refinement using a physical energy function. J. Comp. Chem. *32*, 709-717.

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

Moorman, V.R., Valentine, K.G., and Wand, a.J. (2012). The dynamical response of hen egg white lysozyme to the binding of a carbohydrate ligand. Protein science *21*, 1066-1073.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comp. Chem. *25*, 1605-1612.

Ponder, J.W. (2009). TINKER: Software Tools for Molecular Design. (Saint Louis, Washington University School of Medicine).

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A.J., Read, R.J., and Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. Nature *450*, 259-264.

Rosenbluth, M.N., and Rosenbluth, A.W. (1955). Monte Carlo Calculation of the Average Extension of Molecular Chains. J. Chem. Phys. *23*, 356.

Scheidig, A.J., Burmester, C., and Goody, R.S. (1999). The pre-hydrolysis state of p21(ras) in complex with GTP: new insights into the role of water molecules in the GTP hydrolysis reaction of ras-like proteins. Structure Fold. Des. *7*, 1311-1324.

Schmidt , J.M. (2007). Asymmetric Karplus curves for the protein side-chain 3 J couplings. J. Biomol. NMR *37*, 287-301.

Schnell, J.R., Dyson, H.J., and Wright, P.E. (2004). Effect of cofactor binding and loop conformation on side chain methyl dynamics in dihydrofolate reductase. Biochemistry *43*, 374-383.

Shapovalov, M.V., and Dunbrack, R.L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure (London, England : 1993) *19*, 844-858.

Shaw, D.E., Deneroff, M.M., Dror, R.O., Kuskin, J.S., Larson, R.H., Salmon, J.K., Young, C., Batson, B., Bowers, K.J., Chao, J.C., *et al.* (2008). Anton, A Special-Purpose Machine for Molecular Dynamics Simulation. Comm. ACM *51*, 91-97.

Shen, M.Y., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein Sci *15*, 2507-2524.

Shirts, M., and Pande, V.S. (2000). COMPUTING: Screen Savers of the World Unite! Science *290*, 1903-1904.

Stock, A.M., Martinez-Hackert, E., Rasmussen, B.F., West, A.H., Stock, J.B., Ringe, D., and Petsko, G.A. (1993). Structure of the Mg(2+)-bound form of CheY and mechanism of phosphoryl transfer in bacterial chemotaxis. Biochemistry *32*, 13375-13380.

Stone, M.J. (2001). NMR relaxation studies of the role of conformational entropy in protein stability and ligand binding. Acc. Chem Res *34*, 379-388.

Tsai, J., Bonneau, R., Morozov, A.V., Kuhlman, B., Rohl, C.A., and Baker, D. (2003). An improved protein decoy set for testing energy functions for protein structure prediction. Proteins-Structure Function and Genetics *53*, 76-87.

Tuttle, L.M., Dyson, H.J., and Wright, P.E. (2013). Side-Chain Conformational Heterogeneity of Intermediates in the Escherichia coli Dihydrofolate Reductase Catalytic Cycle. Biochemistry *52*, 3464-3477.

Tyka, M.D., Keedy, D.a., André, I., Dimaio, F., Song, Y., Richardson, D.C., Richardson, J.S., and Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. J. Mol. Biol. *405*, 607-618.

Tzeng, S.-R., and Kalodimos, C.G. (2012). Protein activity regulation by conformational entropy. Nature *488*, 236-240.

Zhang, J., and Liu, J.S. (2006). On side-chain conformational entropy of proteins. PLoS Comp. Bio. *2*, e168.

Zhao, F., and Xu, J. (2012). A position-specific distance-dependent statistical potential for protein structure and functional study. Structure (London, England : 1993) *20*, 1118-1126.

Zhou, H., and Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci *11*, 2714-2726.

ACKNOWLEDGEMENTS. This work was supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231. We thank Dr. Po Hu for an early implementation of the MC-SCE algorithm. We thank so-so for insightful discussions about RINGER calculations. We thank the reviewers for many helpful suggestions.

AUTHOR CONTRIBUTIONS. T.H-G. conceived and designed the experiments; A.B. performed the experiments; A. B. and T. H-G. analyzed the data; A. B. and T. H-G. co-wrote the paper.

TABLES

Table 1. *X-ray crystallographic and MC-SCE generated side chain χ rotamers for active site residues of CypA.* Experimental rotamer populations (Fraser et al., 2009) are the occupancies reported in the deposited PDB files for CypA and mutant (CC: 3k0m, RT: 3k0n, Ser99Thr: 3k0o). In certain cases, the minor rotamer was identified in the CC structure using the software Ringer (Lang et al., 2010). MC-SCE calculations were done on the backbone of the cryo-cooled structure (CC) as well as an average over the backbone conformers M and m reported for the room temperature crystal structure (RT-M and RT-m). MC-SCE calculations were also performed on a RT backbone ensemble comprised of backrub motions and MD simulations (RT ensemble).

CypA		**X-ray Population**		**MC-SCE population using CC, RT, and Ensemble backbones**			**CypA Mutant Ser99Thr**	
Res	*χ Class*	*CC*	*RT*	*CC backbone*	*RT (M, m) backbone*	*RT ensemble*	*X-ray RT*	*MC-SCE RT*
Leu98 (χ_1)	60							
	180	**100.0**	**63.0**	100.0	50.0	57.5	**100.0**	19.0
	-60		**37.0**		50.0	42.5	**Ringer**	74.7
Ser99 (Thr99) (χ_1)	60		**37.0**		50.0	22.4	**Ringer**	44.3
	180	**100.0**	**63.0**				**100.0**	3.8
	-60			100.0	50.0	77.6		51.9
Phe113 (χ_1)	60	**100.0**	**63.0**	100.0	50.0	75.0		
	180							
	-60		**37.0**		50.0	25.0	**100.0**	100.0
Arg55 (χ_3)	60				3.5	17.3		
	180	**100.0**	**63.0**	78.6	45.2	53.0	**Ringer**	25.3
	-60	**Ringer**	**37.0**	21.4	51.3	29.7	**100.0**	74.7
Met61 (χ_2)	60	**40.0**	**37.0**	1.8	0.7	9.0	**100.0**	83.5
	180	**60.0**	**63.0**	98.2	99.3	91.0		1.3
	-60							15.2

Table 2: *J-coupling data for CYPA calculated using MC-SCE on the cryo-cooled backbone of wild type CYPA (3k0m) and Ser99Thr mutant (3k0p)*. The experimental values are taken from(Fraser et al., 2009). Two sets of Karplus parameters have been used to generate the MC-SCE scalar couplings: (Schmidt 2007) CC(S), and (Tuttle et al., 2013), CC(T). Using either parameters, we observed a change in J-coupling values for Phe113 which confirms a switch in rotameric state from 60° to -60°.

	WT						**Ser99Thr**					
Residue	***$^3J_{CC}$(Hz)***			***$^3J_{NC}$(Hz)***			***$^3J_{CC}$(Hz)***			***$^3J_{NC}$(Hz)***		
	Expt	*CC (S)*	*CC (T)*	*Expt*	*CC (S)*	*CC (T)*	*Expt*	*CC (S)*	*CC (T)*	*Expt*	*CC (S)*	*CC (T)*
Phe25	3.6	3.91	4.36	0.8	0.36	0.37	3.7	3.91	4.37	0.7	0.36	0.37
Tyr79	3.3	3.78	4.39	1.0	0.44	0.42	3.4	3.79	4.40	0.7	0.41	0.41
Phe88	3.6	3.91	4.41	0.8	0.39	0.37	3.3	3.92	4.39	0.5	0.39	0.37
His92	3.6	4.22	5.06	1.0	0.57	0.47	3.9	4.22	5.06	1.2	0.57	0.51
Phe113	**1.1**	**0.43**	**0.38**	**0.9**	**0.51**	**0.41**	**2.8**	**3.9**	**4.41**	**0.8**	**0.46**	**0.39**
Phe145	3.3	3.91	4.39	0.9	0.36	0.37	3.5	3.91	4.38	0.2	0.44	0.37

Table 3. *X-ray crystallographic and MC-SCE generated side chain χ rotamers for active site residues of H-Ras.* Experimental rotamer populations are the occupancies reported in the deposited PDB files (CC: 1ctq, RT: 3TGP). In certain cases, the minor rotamer was identified using the software Ringer (Lang et al., 2014; Lang et al., 2010). MC-SCE calculations were done on the cryo-cooled backbone (CC), backrub ensemble of the cryo-cooled backbone (BR-CC) as well as on RT MD snapshots generated at 0.2, 2 and 4 ns time points to incorporate backbone flexibility.

H-Ras		**X-ray Populations**		**MC-SCE using CC backbone**		**MC-SCE using RT MD backbone**		
Res	*χ Class*	*CC χ*	*RT χ*	*CC*	*BR-CC*	*0.2ns*	*2.0 ns*	*4.0 ns*
Asp 30 (χ_1)	60		**55.0**					
	180	**100.0**	**45.0**	100.0	58.3	90.0	100.0	50.0
	-60				41.7	10.0		50.0
Glu 62 (χ_1)	60						1.5	
	180		**100.0**	7.7	41.6		24.2	66.7
	-60	**100.0**		92.3	58.4	100.0	74.3	33.3
Ser 65 (χ_1)	60	**100.0**		88.5	50.0	65.2	9.1	66.7
	180		**100.0**					
	-60			11.5	50.0	34.8	90.9	33.3
His 94 (χ_1)	60	**100.0**	**48.0**	61.6	50.0	11.6	7.6	100.0
	180		**52.0**	34.6	50.0	88.4	22.7	
	-60			3.8			69.7	
Val 103 (χ_1)	60	**Ringer**					1.5	
	180		**38.0**	38.5	58.3			
	-60	**100.0**	**62.0**	61.5	41.7	100.0	98.5	100.0
Gln 61 (χ_2)	60		**66.0**	23.1	25.0	1.5		
	180	**100.0**	**34.0**	3.8	25.0	30.4	90.9	66.7
	-60		**Ringer**	73.1	50.0	68.1	9.1	33.3
Arg 97 (χ_3)	60		**Ringer**	3.8		97.1	1.5	
	180	**100.0**	**100.0**	96.2	83.3	0.0	98.5	100.0
	-60				16.7	2.9		
Glu 98 (χ_2)	60				16.7	10.2	7.6	
	180	**100.0**		80.8	75.0	73.9	1.5	83.3
	-60		**100.0**	19.2	8.3	15.9	90.9	16.7
Gln99 (χ_2)	60		**Ringer**		8.3		39.4	
	180	**100.0**	**100.0**	84.6	83.3	92.7	54.6	66.7
	-60			15.4	8.3	7.3	6.1	33.3

Table 4. *Thermodynamic rankings and Z-scores of the native X-ray structure and MD and Backrub(Friedland et al., 2008) relaxed backbones.*

Protein	RMSD (Å)	E_{best} Rank	E_{best} Z-score	F Rank	F Z-score	Protein	RMSD (Å)	E_{best} Rank	E_{best} Z-score	F Rank	F Z-score
1ail	**0.00**	**9**	**-1.22**	**3**	**-1.73**	**1pgx**	**0.00**	**3**	**-2.29**	**2**	**-2.23**
Backrub	0.26	1	-2.01	1	-2.33	Backrub	0.70	1	-3.08	1	-3.22
MD	1.13	1	-3.58	1	-3.36	MD	1.13	1	-2.72	1	-2.71
1c8c	**0.00**	**3**	**-2.01**	**2**	**-2.34**	**1rnb**	**0.00**	**90**	**1.18**	**89**	**1.17**
Backrub	0.33	3	-2.22	1	-2.59	Backrub	0.73	1	-3.87	1	-3.87
MD	0.56	3	-2.35	1	-2.96	MD	1.23	1	-4.07	1	-4.01
1dhn	**0.00**	**2**	**-2.40**	**2**	**-2.01**	**1ubi**	**0.00**	**10**	**-1.23**	**5**	**-1.38**
Backrub	0.41	1	-3.47	1	-2.82	Backrub	0.35	1	-2.66	1	-2.46
MD	0.94	1	-3.70	1	-3.34	MD	0.68	1	-2.92	1	-2.66
1enh	**0.00**	**14**	**-1.03**	**13**	**-1.02**	**1utg**	**0.00**	**81**	**0.94**	**75**	**0.73**
Backrub	0.31	1	-2.66	1	-2.47	Backrub	0.56	10	-1.29	10	-1.24
MD	0.60	1	-2.44	1	-2.32	MD	1.29	15	-0.99	26	-0.78
1gvp	**0.00**	**21**	**-0.81**	**18**	**-1.04**	**1vcc**	**0.00**	**3**	**-1.99**	**2**	**-2.06**
Backrub	0.65	1	-3.57	1	-3.67	Backrub	0.26	1	-3.38	1	-3.20
MD	1.14	1	-4.04	1	-3.86	MD	0.85	1	-3.77	1	-3.60
1hz6	**0.00**	**3**	**-2.03**	**3**	**-2.22**	**1vls**	**0.00**	**75**	**0.45**	**98**	**2.11**
Backrub	0.04	7	-1.76	6	-1.86	Backrub	0.93	1	-2.16	1	-1.81
MD	0.68	3	-2.11	3	-2.27	MD	1.02	1	-3.39	1	-2.58
256b	**0.00**	**1**	**-2.07**	**3**	**-1.65**						
Backrub	0.42	1	-2.81	1	-2.32						
MD	1.31	1	-4.02	1	-3.57						

FIGURE CAPTIONS

Figure 1. *The PDB backbone and crystallization conditions for H-Ras.* The residues represented are Asp30 (olive) and Ser65 (orange). They are important catalytic residues studied for H-Ras in which both the cryogenically cooled structure (1CTQ)(Scheidig et al., 1999) and the room temperature structure (3TGP)(Fraser et al., 2011) were crystallized with a bound GTP ligand bound (purple). MC-SCE could not predict the major rotamer reported in the room temperature crystal structure for these 2 residues. The meshes represent the crystal elements nearby as reported in the room temperature crystal structure(3TGP). The figure was generated using the PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.

Figure 2. *J-coupling constants (a) $^3J_{C\gamma N}$ and (b) $^3J_{C\gamma CO}$ for Eglin-C.* The red symbols are the experimental data from (Clarkson et al., 2006). The blue symbols are calculated from the MC-SCE ensemble using backbones from molecular dynamics and the Karplus parameterization from (Tuttle et al., 2013).

Figure 3. *J-coupling constants (a) $^3J_{C\gamma N}$ and (b) $^3J_{C\gamma CO}$ for the DHFR binary product complex E:THF.* The red symbols are the experimental data from (Tuttle et al., 2013). The blue symbols are calculated from the MC-SCE ensemble using backbones from molecular dynamics and the Karplus parameterization from (Tuttle et al., 2013).

Figure 4. *J-coupling constants (a) $^3J_{C\gamma N}$ and (b) $^3J_{C\gamma CO}$ for the DHFR binary product complex E:FOL.* The red symbols are the experimental data from (Tuttle et al., 2013). The blue symbols are calculated from the MC-SCE ensemble using backbones from molecular dynamics and the Karplus parameterization from (Tuttle et al., 2013).

Figure 5. *The PDB native backbone of the Mg^{2+}-bound form of CheY(Stock et al., 1993) (2CHF).* The regions colored red are the side chain positions where alternate rotameric states was found by the MC-SCE algorithm compared to the PDB side chain packing. It is notable that side chain repacking occurs for both interior and surface residues. The Figure was generated using Chimera(Pettersen et al., 2004).