PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Role of conformational sampling in computing mutation-induced changes in protein structure and stability

Elizabeth H. Kellogg,[1] Andrew Leaver-Fay,[2] and David Baker[1]*

[1] Department of Biochemistry, University of Washington, Seattle, WA 98105
[2] Department of Biochemistry, University of North Carolina, Chapel Hill, NC 27599

ABSTRACT

The prediction of changes in protein stability and structure resulting from single amino acid substitutions is both a fundamental test of macromolecular modeling methodology and an important current problem as high throughput sequencing reveals sequence polymorphisms at an increasing rate. In principle, given the structure of a wild-type protein and a point mutation whose effects are to be predicted, an accurate method should recapitulate both the structural changes and the change in the folding-free energy. Here, we explore the performance of protocols which sample an increasing diversity of conformations. We find that surprisingly similar performances in predicting changes in stability are achieved using protocols that involve very different amounts of conformational sampling, provided that the resolution of the force field is matched to the resolution of the sampling method. Methods involving backbone sampling can in some cases closely recapitulate the structural changes accompanying mutations but not surprisingly tend to do more harm than good in cases where structural changes are negligible. Analysis of the outliers in the stability change calculations suggests areas needing particular improvement; these include the balance between desolvation and the formation of favorable buried polar interactions, and unfolded state modeling.

Proteins 2011; 79:830–838.

Key words: $\Delta\Delta G$ prediction; protein stability; backbone flexibility; free energy change.

INTRODUCTION

Accurate modeling of the impact of a mutation in a protein must recapitulate both the structural change associated with a mutation as well as the change in the free energy of the folded state. As with most other macromolecular structure prediction problems,[1] accurately predicting the structural changes associated with a point mutation requires, first, an efficient method for conformational sampling, and second, an accurate free energy function. Once the structure of the mutant protein has been computed, the change in the free energy of folding can be estimated from the difference in the free energies of the folded wild-type and mutant structures, assuming the change in the unfolded state free energies depends only on the identities of the amino acids at the substituted positions. Previous studies have used conservative sampling procedures to predict differences in free-energies, $\Delta\Delta G$s, allowing only the mutated residue to reconfigure within a fixed environment,[2–4] as well as methods incorporating increased protein flexibility.[5,6] Although the above studies all report impressive correlations with experimental values, they use quite different energy functions and sampling strategies, hence it is not clear which features of the approaches are sufficient and necessary for good performance.

Prompted by a recent study reporting poor performance of the Rosetta methodology in predicting the free energy changes associated with mutations,[7] we present here a detailed analysis of the tradeoff between the resolution of the energy function and the extent of conformational sampling in $\Delta\Delta G$ prediction. We go beyond previous work by systematically evaluating a wide range of sampling methodologies (Fig. 1) in the context of the same forcefield, separating the contribution of the forcefield from that of the sampling methodology. We show that roughly equivalent overall performance can be achieved using a wide range of sampling techniques, ranging from an entirely fixed backbone approximation to full-protein flexibility, provided that the resolution of the energy function is matched to the granularity of the sampling technique. The poor results obtained by Potapov *et al.* are shown to be the result of inappropriate combination of limited sampling with an undamped potential function. By studying the distributions of prediction failures, we identify areas of modeling which need to be improved

Additional Supporting Information may be found in the online version of this article.
*Correspondence to: David Baker, Department of Biochemistry, University of Washington, Seattle, WA 98105. E-mail: dabaker@u.washington.edu.
Received 20 April 2010; Revised 8 September 2010; Accepted 13 October 2010
Published online 19 October 2010 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.22921

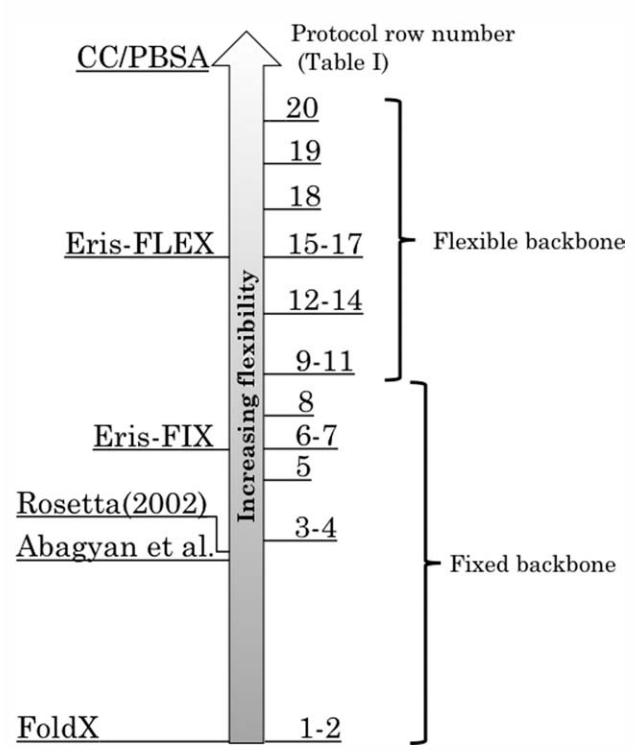

Figure 1

Extent of conformational sampling in the $\Delta\Delta G$ prediction protocols. Protocols considered here are on the right, and previously described methods (Refs. 1–5) on the left.

for higher accuracy prediction of the changes in stability and structure brought about by point mutations.

METHODS

Data-set

Except for comparison to the results of Potapov *et al.* mentioned in the discussion, all tests reported in this article utilized a benchmark set comprised of 1210 single mutations obtained from Protherm.[8] Duplicate entries were resolved by taking the highest resolution structure, and if multiple experimental measurements were recorded, the mean of all reported measurements was used. Structures greater than 350 residues were eliminated due to the computational intensiveness of some of the protocols tested. A representative set of 771 mutations was used to assess the most computationally intensive protocols, including the proteins barnase (1a2p), apomyoglobin (1bvc), FK506 binding protein (1fkj), staphylococcal nuclease (1stn), α-spectrin (1u5p), chymotrypsin inhibitor II (2ci2), and T4-lysozyme (2lzm).

Description of protocols

The first set of protocols (Table I) we considered relax the sidechains but keep the backbone fixed. Sidechains are optimized in two steps—first, discrete combinatorial rotamer optimization and second, continuous optimization of the sidechain torsion angles. The combinatorial rotamer optimization (referred to as repacking throughout the remainder of the text) is carried out using Monte Carlo simulated annealing with the Dunbrack backbone dependent rotamer library.[9] The continuous optimization is carried out using quasi-Newton minimization and is referred to as minimization throughout the remainder of the text.

We experimented with two energy functions at both the repacking and minimization steps. The first is the standard Rosetta all atom energy function used in prediction and design calculations;[10] we refer to this as "hard-rep" because the Lennard-Jones repulsive interactions are not damped, thus atomic clashes incur very large energetic penalties. The second has the repulsive interactions at short atomic separations damped as described in the Supporting Information but is otherwise identical; we refer to this as "soft-rep" because small atomic overlaps are not heavily penalized.

We also experimented with allowing different numbers of residues surrounding the site of mutation to be

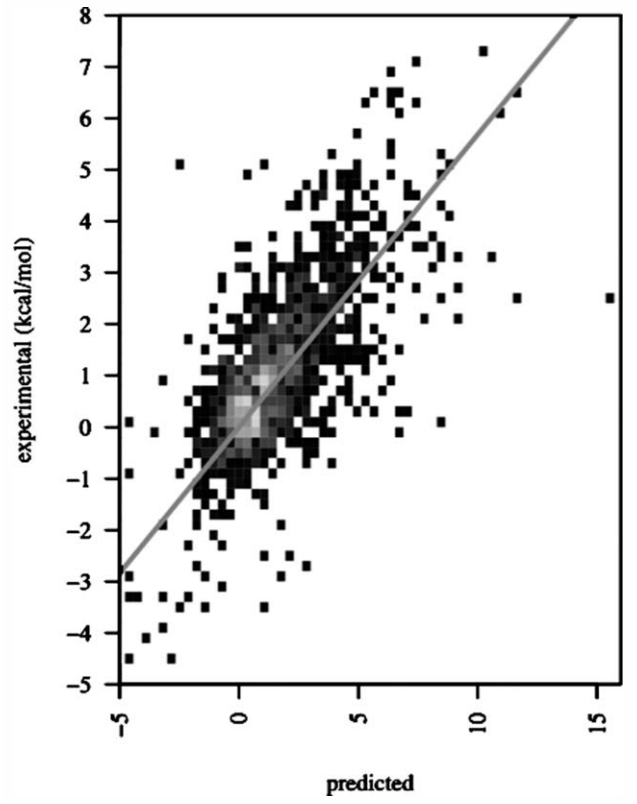

Figure 2

Comparison of predicted to experimentally observed $\Delta\Delta G$s. Method 16 (Table I) which employs backbone minimization after repacking all sidechains was used in this calculation. The correlation is 0.69 on the full set of 1210 mutations. Predicted values along the *x*-axis versus experimental values (kcal/mol) on the *y*-axis. The equation of the best-fit line is $y = 0.57x$. All results are for the 1,210 mutation test set except for those in the last row, which are on a reduced set of 771 mutations, due to the computational cost of the ensemble method.

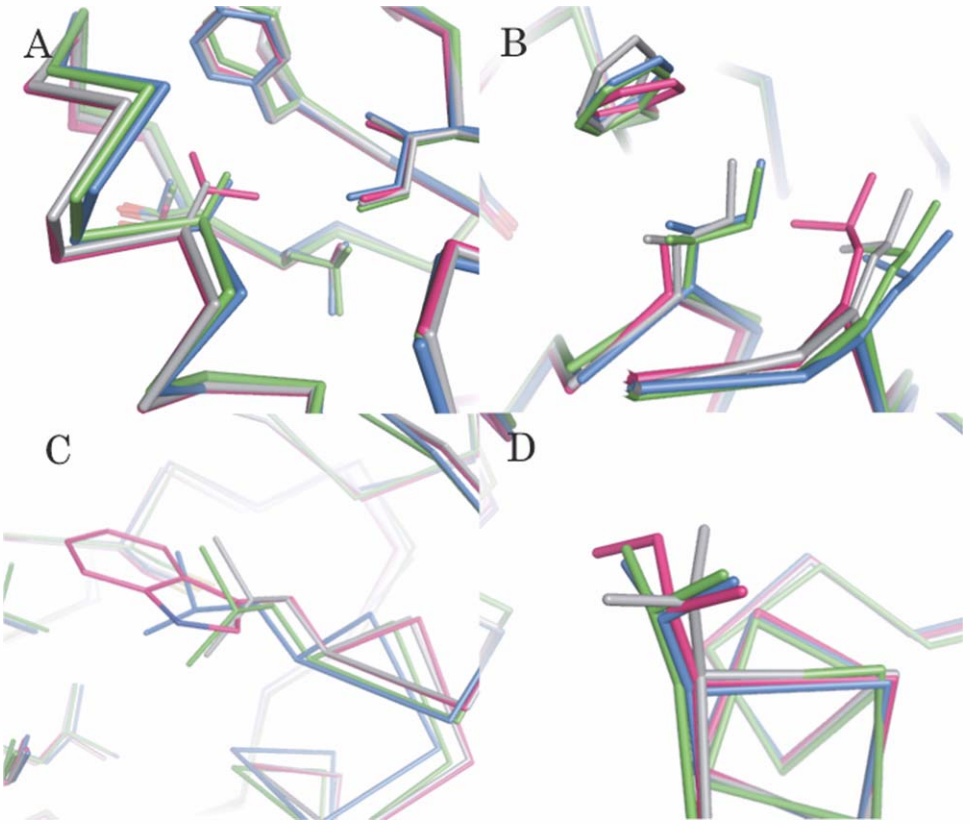

Figure 3

Examples for which modeling backbone flexibility improves structural recapitulation. (**A**) T4-lysozyme mutant (1qtb), V 42 A; (**B**) T4-lysozyme mutant (241l) A 29 I; (**C**) FK506 binding protein (1fkj) W 59 L; and (**D**) T4-lysozyme (2lzm) I 3 V. Pink, starting wild-type crystal structure; blue, mutant crystal structure; gray, structural prediction with limited backbone minimization; and green, structure produced with less stringent constraints around the site of mutation and uniform harmonic constraints outside this region (row 18, Table I). In (D), green is the structure produced from perturbed backbone protocol (row 20, Table I).

repacked. As indicated in the Table I protocol summary, we considered three possibilities: first, only repacking the mutated residue, second, only residues within 8 Å of the mutated residue, and third, all residues.

We also explored protocols which carry out backbone torsion angle minimization following sidechain repacking in attempts to more accurately model the structural consequences of mutations. To prevent the backbone from moving too much from the native structure, in some protocols, we included distance constraints during the backbone minimization as described in the Supporting Information.

Finally, we explored protocols which more extensively search through alternative backbone conformations. We developed a Monte Carlo simulated annealing protocol that generates backbone conformations with ideal bond lengths and bond angles that uniformly sample the space of conformations surrounding any given native structure. The protocol carries out 100,000 moves each consisting of a small random perturbation of the backbone torsion angles; the scoring function prevents sampling from deviating by more than a specified tolerance from the starting structure. Single side chain rotamer flips are attempted at one-tenth the frequency of backbone moves. The resulting structures have small and partially compensating changes in nearly all the backbone torsion angles. The lowest energy structure sampled during each trajectory is subjected to backbone and sidechain minimization using the hard-rep energy function. Full details are provided in the Supplementary Information.

RESULTS

As described in detail in the Methods section, we experimented with a range of different protocols for computing free energy changes accompanying mutations. In all of these protocols, the calculations focus on energy changes in the native state—changes in free energy of the unfolded state are assumed to be context independent for computational tractability.

Sidechain-only optimization

In the first set of protocols, the sidechains, but not the backbone, are allowed to relax following introduction of the amino acid sequence change. Several trends are evident in the comparison of the performance of the different fixed backbone protocols in Table I. First, the

Table I
ΔΔ*G* Prediction Accuracies for All Tested Protocols

	extent of sidechain repacking	energy function used in repacking	extent of minimization	constraints used during minimization	energy function used in minimization	performance		
						all mutations	large-to-small mutations	small-to-large mutations
1	1 residue	soft-rep				0.66 / 0.73	0.67 / 0.73	0.55 / 0.65
2		hard-rep				0.02 / 0.53	0.25 / 0.57	0.10 / 0.52
3	within 8 Å	soft-rep				0.68 / 0.72	0.68 / 0.73	0.56 / 0.64
4		hard-rep				0.04 / 0.54	0.46 / 0.64	0.14 / 0.53
5		hard-rep	sidechain	no	hard-rep	0.26 / 0.55	0.50 / 0.66	0.38 / 0.55
6	all-residues	soft-rep				0.67 / 0.71	0.67 / 0.72	0.57 / 0.64
7		hard-rep				0.10 / 0.54	0.46 / 0.63	0.22 / 0.55
8		hard-rep	sidechain	no	hard-rep	0.25 / 0.55	0.58 / 0.72	0.32 / 0.54
9	1 residue	soft-rep	backbone and sidechain	uniform	soft-rep	0.65 / 0.69	0.63 / 0.70	0.63 / 0.66
10		soft-rep			hard-rep	0.51 / 0.69	0.65 / 0.72	0.27 / 0.55
11		hard-rep			hard-rep	0.58 / 0.72	0.66 / 0.73	0.40 / 0.58
12	within 8 Å	soft-rep			soft-rep	0.65 / 0.70	0.62 / 0.71	0.61 / 0.66
13		soft-rep			hard-rep	0.66 / 0.72	0.65 / 0.72	0.58 / 0.65
14		hard-rep			hard-rep	0.64 / 0.73	0.65 / 0.74	0.50 / 0.63
15	all-residues	soft-rep			soft-rep	0.57 / 0.69	0.53 / 0.68	0.65 / 0.67
16		soft-rep			hard-rep	0.69 / 0.72	0.67 / 0.72	0.66 / 0.67
17		hard-rep			hard-rep	0.63 / 0.73	0.67 / 0.74	0.45 / 0.62
18		soft-rep		position-specific	hard-rep	0.67 / 0.70	0.64 / 0.71	0.64 / 0.67
19		soft-rep		no	hard-rep	0.63 / 0.69	0.60 / 0.69	0.61 / 0.67
20	all-residues	Monte Carlo ensemble generation (see methods)				0.65 / 0.68*	0.66 / 0.68*	0.54 / 0.73*

Values reported for each method are the correlation/stability-classification accuracy with respect to experimental data. While overall correlations are very similar among different methods, the small-to-large class shows improved performance with the addition of more protein flexibility.
*Values correspond to a reduced set of 771 mutations.

performance of protocols using the hard-rep potential improved with increasing conformation freedom (rows 2, 4, and 7, Table I), but for all of the fixed backbone sampling strategies, better performance was achieved with the soft-rep potential (rows 1 and 2, rows 3 and 4, rows 6 and 7, Table I). Atomic clashes in models of the mutant structures that cannot be fully resolved with sidechain–only optimization are likely to account for both of these trends; indeed, filtering the data by removal of ΔΔ*G*s with large clashes (>7 E_{rep}) after repacking all residues with the hard-rep energy function (row 7, Table I) increased the correlation to 0.62 (1117 mutations) from 0.10 (1210 mutations). The atomic clashes that remained following sidechain repacking were not resolvable with sidechain minimization (r = 0.25; row 8, Table I). Second, the performance of protocols using the soft-rep energy function was insensitive to the amount of conformational freedom (rows 1, 3, and 6, Table I); very similar performance was obtained whether only the mutated residue was repacked, a subset of residues were repacked or all residues were repacked (correlations of 0.67 and stability-classification accuracies of 0.73; rows 1, 3, and 6, Table I). This result is consistent with the earlier observation[2] that the soft-rep potential is well suited to recapitulating ΔΔ*G*s with a fixed backbone, and does not require the optimization of neighboring sidechains to obtain a significant correlation.

The reason for the poor results obtained by Potapov *et al.* is evident from the above analysis. Perhaps because of unclear documentation, Potapov *et al.*[7] used the hard-rep potential with a limited sidechain repacking protocol followed by sidechain minimization (similar to row 5 in Table I), and found very little correlation between predicted and observed ΔΔ*G*s (0.26 on 1913 mutations). As we have discussed previously,[11] if a fixed backbone representation with discrete rotamer optimization is carried out, the repulsive interactions must be damped, otherwise they dominate the computed energies.

Limited backbone minimization

The set of sidechain-and-backbone protocols (rows 9–19 in Table I), extends the set of sidechain-only protocols by applying a restrained quasi-Newton minimization step to backbone and sidechain degrees of freedom starting from sidechain optimized structures while tethering the structure to the initial starting model (see Supporting Information). Correlations were higher when the soft-rep energy function was used during the sidechain optimiza-

tion step than when the hard-rep energy function was used ($R = 0.63$ with hard-rep, 0.69 with soft-rep when backbone and sidechain minimization is performed after repacking all-residues, see rows 16 and 17, Table I). However, if the soft-rep energy function was used during minimization, increased conformational freedom yielded worse correlations (row 15, Table I, $R = 0.57$) highlighting the incompatibility of the soft-rep energy function with flexible backbone modeling.

The performance of these protocols improved as more conformational freedom was introduced to the system—to a point. As more sidechains were allowed to repack before the backbone minimization step, the correlations improved (rows 10, 13, and 16, Table I). Repacking all residues before minimization ($r = 0.69$, row 16, Table I) performed marginally better than repacking residues with 8A ($r = 0.66$, row 13 Table I), which performed better than repacking the mutant residue only (0.51, row 10 Table I). This improvement is likely due to sensitivity of the hard-rep scoring function to residual atomic clashes; excluding mutations for which models contained high-repulsive energies (>7 E_{rep}) restored the correlations of the protocols to 0.68 (1,207 mutations).

The minimization step in the above calculation is constrained using crystal structure derived restriants. Allowing increased conformational freedom in the neighborhood of the mutation during the minimization stage (weakening the crystal structure based distance restraints in the neighborhood of the mutation site) yielded a slightly worse correlation (from $r = 0.69$, row 16, Table I to $r = 0.67$, row 18, Table I), and complete removal of restraints during minimization (row 19, Table I), yielded a worse correlation still (0.63, Table I).

Monte Carlo ensembles

Sampling is quite limited with quasi-Newton minimization of the backbone; it locates the nearest minimum but is unable to cross barriers into lower-energy minima nearby. To increase the exploration of the energy landscape close to the native structure, we developed a protocol that generates an ensemble of structures centered on the native structure. Other methods have recently been developed using "back-rub" motions that have proven quite powerful.[12,13] Our goal was to generate ensembles with levels of structural perturbation similar to those generated with back rub while restricting bond lengths and angles to ideal values (see Supporting Information), since the addition of bond length and bond angle degrees of freedom and associated potential terms can introduce noise. The new protocol was tested with the hard-rep energy function and yields ensembles with uniform deviations from the starting native structure both in Cartesian coordinates and in the individual torsion angles (see Supporting Information, Fig. 2).

Although significant correlations can be produced with stochastic sampling of backbone conformations close to the starting structure, these correlations are not as high as those obtained using limited backbone minimization ($r = 0.65$, row 20 in Table I vs. $r = 0.69$, row 16 in Table I, both evaluated on a set of 771 mutations). As previously observed by Benedix *et al.*,[6] the correlations increase as more models in the ensemble are produced (Supporting Information Fig. 3). This is likely due to reduction in the noise associated with stochastic sampling of the protein backbone. The considerable improvement obtained by Benedix *et al.* with conformational sampling compared to using static crystal structure likely reflects the undamped potential they used.

Comparison of sampling techniques

No one protocol significantly outperforms the others; among the best combinations of energy function and optimization method for each of the sampling regimes, the correlations ranged from 0.65 to 0.69 (rows 6 and 16, Table I). However, if mutations are divided according to the change in van der Waals volume, clear trends are observed. In particular, the best protocol that relaxed the backbone (row 16, Table I) showed a significant improvement over the best sidechain-only protocol (row 6, Table I) for the small-to-large class of mutations ($r = 0.66$ vs. $r = 0.57$ on a set of 164 mutations, rows 6 and 16, Table I) and also on mutations involving only hydrophobic residues ($r = 0.68$ vs. $r = 0.57$ on a set of 365 mutations; see Supporting Information).

The inclusion of restrained backbone minimization (row 16, Table I) did not compromise the correlation on large-to-small mutations; the correlation is equivalent to the maximum obtained by other methods, 0.67 (row 16, Table I). A similar result was reported by Yin *et al.*[5] The protocol involving extensive backbone movement (row 20, Table I) has correlations similar to the fixed backbone methods in all size categories—improvements in modeling mutations that induce significant backbone changes are offset by the introduction of noise in modeling the remaining mutations. The stability-classification accuracies for the best methods were 0.73 for large-to-small mutations (rows 1, 3, and 11, Table I) and 0.67 for small-to-large mutations (rows 15, 16, 18, and 19, Table I); no protocol significantly outperforms the others using this metric.

Structure recapitulation

Overall, the variation in the protein backbones produced by the methods increases with increasing conformational searching. Constrained minimization protocols (rows 9–17, Table I) on an average produce structures 0.08 Cα RMSD from the starting structure, whereas minimization with no constraints (row 19, Table I) produces structures on an average 0.57 Cα RMSD from the start-

Table II
The Free Energy Changes Associated with Surface Substitutions and Polarity Changing Substitutions Are Relatively Poorly Predicted. Results Shown Are for the Best Performing Method (row 16, Table I), Involving Limited Backbone Minimization After Repacking All Sidechains

Category	Correlation	Fraction correct	Number of Mutations
All	0.69	0.72	1210
Low B-factor	0.69	0.75	596
High B-factor	0.67	0.7	606
Buried	0.66	0.78	397
Partially exposed	0.63	0.71	421
Exposed	0.54	0.72	384
Nonpolar	0.68	0.76	365
Polar-to-nonpolar	0.58	0.68	456
Polar	0.79	0.7	81

Exposed mutations and polarity changes are relatively poorly predicted. Results shown are for the best performing method, involving limited backbone minimization after repacking all sidechains.

ing structure. The Monte Carlo ensemble method (row 20, Table I) is more aggressive than limited backbone flexibility but somewhat constrained compared with free minimization, producing backbones of 0.44 Cα RMSD on an average.

We evaluated the performance of the different protocols described above in recapitulating structural changes accompanying point mutations observed in crystal structures of mutant proteins, using a set of 154 pairs for which the crystal structures of both wild-type and mutant proteins were available (see Supporting Information). The more aggressive flexible backbone methods produced quite striking recapitulations of structural changes in a number of cases (Fig. 3) but overall did not result in improved predictions over the more conservative methods (Supporting Information Fig. 5). Overall, loosening constraints around the site of mutation yielded better predictions than the uniform constraint minimization method in 62 cases (of 154), whereas the more aggressive backbone perturbation method yielded better predictions than the best limited-backbone minimization protocol in 44 cases, as assessed by comparing the all-atom RMSD of the mutant sidechain to the crystal structure (see Supporting Information). Prediction accuracy for small-to-large, buried mutations increases slightly with increasing structural variability, but only when the backbone is known to shift ($\geq$0.4 Å Cα shift). When the backbone is essentially correct to begin with, the all-atom RMSD prediction accuracy for the flexible backbone methods is not surprisingly worse than for the fixed backbone methods (Supporting Information Fig. 5). The failure of the flexible backbone methods to give an overall improvement reflects in part the large fraction of cases where very little backbone movement actually occurs. This lack of consistent improvement in structural recapitulation also in part explains why the flexible backbone methods do not do better overall in $\Delta\Delta G$ prediction.

$\Delta\Delta G$ prediction performance with empirical structural knowledge

To determine if improved structural models necessarily lead to improved energetic predictions, we computed predicted $\Delta\Delta G$s based on the solved crystal structures (data not shown). Not surprisingly, naively taking the difference in total computed energy between the wild-type and mutant crystal structures resulted in zero correlation with the experimental $\Delta\Delta G$ data, since small differences throughout the independently solved structures drown out the energy differences due the sequence change itself. To reduce this noise, we computed the difference not in the total energies of the wild type and mutant crystal structures but of the total interaction energies of residues at the mutation site. The correlation of this computed interaction energy difference with the experimental $\Delta\Delta G$ data, 0.77, is the same as that of the best limited-backbone minimization protocol over this set of mutations, a finding corroborated by other studies.[4]

Energy function training incorporating both $\Delta\Delta G$ and sequence recovery data

The Rosetta energy function contains "reference energies" for each of the 20 amino acids, which represent the average energy of the residue in the unfolded state. The parameters in the standard energy function used in the calculations described, thus, far in this article, were determined by maximizing sequence recovery in comprehensive sequence design calculations for a large set of proteins.[9] In this weight optimization, the reference energies are influenced by the overall frequencies of the amino acids, and, hence, will also incorporate effects related to the metabolic cost of making amino acids, their effects on solubility, and so forth. Hence, we reasoned that better performance might be achieved if these reference energies were fit directly on $\Delta\Delta G$ data where overall amino acid composition biases are absent. We fit the 20 reference energies, using 20-fold cross-validation, keeping all other weights fixed except for a constant term to adjust the energies to a kcal/mol scale, obtaining an overall correlation of 0.73 (Supporting Information Table II). Optimization of weights on other forcefield terms did not improve the correlation sufficiently to be justified (Supporting Information Table II). Although the increase in performance resulting from fitting on $\Delta\Delta G$s was not large, a notable advantage is that this puts the overall energy function on a kcal/mol scale matched to experimental $\Delta\Delta G$ measurements.

For design calculations, reference energies trained on sequence recovery are likely to be desirable, whereas for $\Delta\Delta G$ calculations, training on thermodynamic data is more appropriate. To obtain a compromise reference weight set, we trained on both datasets at the same time using the opt-E weight-optimization suite (Leaver-Fay *et al.*, in preparation), yielding a weight set with a $\Delta\Delta G$

Table III
Classes of Mutation Enriched in the Outlier Population

	Outlier mutations			All mutations		
Category	Number	Total	Percentage	Number	Total	Percentage
Unfolded state significantly affected by mutation	13	38	34	23	305	8
Buried hydrogen bonds	16	85	19	106	1210	9
Buried polar–polar hydrogen bonds	11	85	13	67	1210	6
Buried charged-polar hydrogen bonds	7	85	8	59	1210	5
Introduction of buried unsatisfied hydrogen bonding partner	7	85	8	52	1210	4
Putative conformational change	6	85	7	34	1210	3
Buried, hydrogen bonded to water	8	85	9	45	1210	4
Ligand contacts	5	85	6	18	1210	1
Buried, mobile region	5	85	6	69	1210	6

Residues making buried hydrogen bonds, hydrogen bonds to buried water molecules, or contacting ligands are enriched in the outlier population, as well as mutations affecting the unfolded state.

correlation 0.69 and a sequence recovery rate of 29% (Supporting Information Table II; parameters in Supporting Information).

The correlation after weight-training on $\Delta\Delta G$ experimental data, $r = 0.73$, is essentially equivalent to correlations obtained by other algorithms, ranging from 0.59 to 0.76. Why do such widely different conformational sampling protocols and energy functions have such similar prediction accuracies? A likely explanation is that the remaining variance in the experimental data is due to factors not represented in any of the models. The first of these is experimental error in the measurements themselves—it was recently estimated based on differences in the free energy changes determined in different groups for the same mutation that the maximum correlation possible is 0.86.[7] The second missing contribution is likely due to errors/missing features in the energy function. We survey these potential missing contributions in the following paragraphs.

Contributions to failures in prediction accuracy

To investigate potential systematic problems, mutations were categorized according to polarity, burial, and B-factor (see Supporting Information for category definitions; Table II and Supporting Information Table I). We also compared the enrichment of specific structural features in mutations systematically mispredicted by all of our methods (Table III). The outlier set is defined as the consensus of 10% worst predictions for all protocols; removal of these outliers improved correlations ($r = 0.71$–0.75). Features we examined included the unfolded state, hydrogen bond characteristics, and interactions with buried, bound water molecules or ligands.

The largest errors in accuracy are for cases where polar residues are swapped for hydrophobic residues or vice versa, with correlations ranging from 0.55 to 0.6 (Table II and Supporting Information Table I), which suggests the largest areas for improvement involving the delicate trade-off between polar desolvation and the formation of favorable buried polar interactions. Consistent with this, buried hydrogen bonds are two-fold enriched in the outlier population (19% vs. 9%). Cases in which an unsatisfied hydrogen bonding group is introduced in a buried hydrophobic environment are also enriched in the outlier category (8% vs. 4%). Finally, buried residues making hydrogen bonds to water molecules, an interaction absent in our implicit solvation model, are somewhat enriched in the outlier class as well (9% vs. 4%). The development of polarizable electrostatics models and the inclusion of explicit water molecules[14] may help better recapitulate the energetics of these interactions.

Buried residues are in general predicted better than exposed ones, as has been reported in previous studies.[3,5,6,15] Although the correlation within the category of exposed residues is poor ($r = 0.47$), the stability-classification accuracy is very similar (0.71 for exposed mutations and 0.78 for buried residues, Table II). Because mutations to exposed residues are mostly neutral, they are easy to categorize even if their $\Delta\Delta G$s are challenging to predict (see Supporting Information for definitions).

To examine the potential contribution of the unfolded state, we collected 305 *m*-values for mutations from staphylococcal nuclease.[16–19] Mutations whose *m*-values significantly affected the energy of the unfolded state ($\geq$20% difference from the wild-type *m*-values) were enriched four times more than average in the outlier class (34% vs. 8%). Previous studies have also noted difficulties in modeling this class of mutation accurately.[4,5] Improved modeling of such mutations may require explicit modeling of context-dependent unfolded state effects.

We observe only a marginal decrease in performance for mutations in high-B-factor regions when compared with low-B-factor regions (0.68 vs. 0.64, fixed-backbone, all sidechains repacked, row 6 Table I; Supporting Information Table I). Inclusion of backbone flexibility reduces the discrepancy further (0.69 vs. 0.67, limited backbone

minimization after sidechain repacking, row 16, Table I) (see Table II). Entropic effects may contribute to prediction inaccuracy overall but are not as evident as might be expected in this subset of the data.

Conformational sampling appears to be still in part limiting. The outlier class includes a number of mutations of large to small hydrophobic residues. The free energy changes in these cases are predicted to be extremely destabilizing due to the creation of a large hydrophobic cavity, whereas the effect of the mutation is near neutral, indicating significant conformational rearrangements. Comparison of our predictions to the mutant crystal structure[20,21] suggests that some failures are due to the inability to sample correct conformations.

DISCUSSION

Previous studies have shown that free energy changes accompanying point mutations can be reasonably well predicted, but the features contributing to this success are not evident as the different methods use very different sampling procedures and energy functions. Here, we demonstrate that the free energy changes associated with point mutations can be predicted equally well by protocols that involve widely varying amounts of conformational sampling, provided that the resolution of the energy function matches the coarseness of the sampling. As found in previous studies,[22] protocols involving coarse conformational sampling perform well when repulsive interactions are damped, whereas protocols involving aggressive conformational sampling perform well when repulsive interactions are not damped. We find that protocols that incorporate backbone flexibility are better suited than fixed-backbone protocols for modeling small-to-large mutations, but the preponderance of large-to-small mutations masks this improvement on the overall dataset. Expanding on the results of Dantas *et al.*, we show that the best methods for modeling small-to-large mutations utilize a damped energy function for sidechain optimization followed by an undamped potential during constrained, gradient-based minimization (row 16, Table I). When used during optimization, the hard-rep energy function can select incorrect rotamer conformations (row 17, Table I), which are often not rescued during the subsequent round of gradient-based minimization because of the difficulty in crossing high-energy barriers.

Our calculations model the contributions of mutations to the free energy of folding in different ways. The change in enthalpy resulting from the mutations is calculated explicitly through Lennard Jones interactions, hydrogen bonding, and so forth. Interactions with solvent, both enthalpic and entropic, are modeled using an implicit solvent model.[23] The changes in the entropy and enthalpy of the unfolded state are assumed to be context independent: for example, the change in unfolded state free energy for all leucine to alanine substitutions are assumed to be identical. This assumption clearly breaks down when the residue is making specific interactions and/or has restricted conformational freedom in the unfolded state.[16,24] The reasonable success rate in predicting $\Delta\Delta G$s with this rather drastic assumption suggests that unfolded state effects are not major contributors to the $\Delta\Delta G$, but, as noted in the results, they could well be responsible for some of the deviations between the computations and experiments.

The poor results of Potapov *et al.* resulted from use of a limited sampling protocol without dampening the repulsive interactions. Consistent with the previously reported results, a protocol analogous to that of Potapov, (row 4, Table I), produced a correlation near 0 for our benchmark set of 1210 mutations. On the dataset used by Popatov, our best performing method using limited backbone minimization (row 16, Table I) yields an overall correlation of 0.57 on 1937 mutations, and 0.62 on 1920 mutations of the Potapov set (excluding as in the Potapov study mutations with repulsive interactions of $\geq$7 units); this is equivalent to the performance of the best algorithms tested by Potapov. For comparison, the EGAD method had a correlation of 0.59 on a set of 1065 mutations, FoldX had a correlation of 0.50 on a set of 1200 mutations, and CC/PBSA had a correlation of 0.56 on a set of 478 mutations.

In conclusion, our best-performing method for $\Delta\Delta G$ prediction involves limited backbone minimization; with training and 20-fold cross-validation, it produces a correlation of $r = 0.73$ on a comprehensive set of 1210 mutations (Supporting Information Table II), matching that of previously published algorithms but on a larger test set. Although addition of protein flexibility in some cases improves the modeling of structural response to mutation, we find that more often than not, more aggressive remodeling can decrease the ability of a method to recapitulate mutant structure and can have correspondingly negative impact on $\Delta\Delta G$ prediction. More extensive sampling with more accurate potential functions hopefully will reverse this dissapointing fall off in predictions in the not too distant future. Analyses of consistently badly predicted mutations among all methods reveal that improvements in modeling the unfolded state, buried polar networks, and explicit water or ligand contacts may be the key to further improvements in performance. There is clearly much room for improvement in $\Delta\Delta G$ prediction methodology.

REFERENCES

1. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450:259–264.
2. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci USA 2002;99:14116–14121.
3. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–387.

4. Bordner AJ, Abagyan RA. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins 2004;57:400–413.
5. Yin S, Ding F, Dokholyan NV. Eris: an automated estimator of protein stability. Nat Methods 2007;4:466–467.
6. Benedix A, Becker CM, de Groot BL, Caflisch A, Bockmann RA. Predicting free energy changes using structural ensembles. Nat Methods 2009;6:3–4.
7. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel 2009;22:553–560.
8. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 2006;34(Database issue):D204–D206.
9. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97:10383–10388.
10. Das R, Baker D. Macromolecular modeling with rosetta. Annu Rev Biochem 2008;77:363–382.
11. Dantas G, Corrent C, Reichow SL, Havranek JJ, Eletr ZM, Isern NG, Kuhlman B, Varani G, Merritt EA, Baker D. High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. J Mol Biol 2007;366:1209–1221.
12. Davis IW, Arendall WB, III; Richardson DC, Richardson JS. The backrub motion: how protein backbone shrugs when a sidechain dances. Structure 2006;14:265–274.
13. Smith CA, Kortemme T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J Mol Biol 2008;380:742–756.
14. Jiang L, Kuhlman B, Kortemme T, Baker D. A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. Proteins 2005;58:893–904.
15. Gilis D, Rooman M. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. J Mol Biol 1997;272:276–290.
16. Green SM, Meeker AK, Shortle D. Contributions of the polar, uncharged amino acids to the stability of staphylococcal nuclease: evidence for mutational effects on the free energy of the denatured state. Biochemistry 1992;31:5717–5728.
17. Shortle D, Stites WE, Meeker AK. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. Biochemistry 1990;29:8033–8041.
18. Byrne MP, Manuel RL, Lowe LG, Stites WE. Energetic contribution of side chain hydrogen bonding to the stability of staphylococcal nuclease. Biochemistry 1995;34:13949–13960.
19. Meeker AK, Garcia-Moreno B, Shortle D. Contributions of the ionizable amino acids to the stability of staphylococcal nuclease. Biochemistry 1996;35:6443–6449.
20. Consonni R, Santomo L, Fusi P, Tortora P, Zetta L. A single-point mutation in the extreme heat- and pressure-resistant sso7d protein from sulfolobus solfataricus leads to a major rearrangement of the hydrophobic core. Biochemistry 1999;38:12709–12717.
21. Fulton KF, Jackson SE, Buckle AM. Energetic and structural analysis of the role of tryptophan 59 in FKBP12. Biochemistry 2003;42:2364–2372.
22. Pappu RV, Marshall GR, Ponder JW. A potential smoothing algorithm accurately predicts transmembrane helix packing. Nat Struct Biol 1999;6:50–55.
23. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
24. Shortle D, Meeker AK. Residual structure in large fragments of staphylococcal nuclease: effects of amino acid substitutions. Biochemistry 1989;28:936–944.