Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing

Bharat V. Adkar,1,4 Arti Tripathi,1,4 Anusmita Sahoo,1 Kanika Bajaj,1 Devrishi Goswami,1 Purbani Chakrabarti,1 Mohit K. Swarnkar,2 Rajesh S. Gokhale,2 and Raghavan Varadarajan1,3,*

1Molecular Biophysics Unit, Indian Institute of Science, Bangalore-560012, India
2Institute of Genomics and Integrative Biology, Delhi-110007, India
3Chemical Biology Unit, Jawaharlal Nehru Center for Advanced Scientific Research, Jakkur PO, Bangalore-560064, India
4These authors contributed equally to this work
*Correspondence: varadar@mbu.iisc.ernet.in
DOI 10.1016/j.str.2011.11.021

SUMMARY

A major bottleneck in protein structure prediction is the selection of correct models from a pool of decoys. Relative activities of ~1,200 individual single-site mutants in a saturation library of the bacterial toxin CcdB were estimated by determining their relative populations using deep sequencing. This phenotypic information was used to define an empirical score for each residue (RankScore), which correlated with the residue depth, and identify active-site residues. Using these correlations, ~98% of correct models of CcdB (RMSD ≤ 4Å) were identified from a large set of decoys. The model-discrimination methodology was further validated on eleven different monomeric proteins using simulated RankScore values. The methodology is also a rapid, accurate way to obtain relative activities of each mutant in a large pool and derive sequence-structure-function relationships without protein isolation or characterization. It can be applied to any system in which mutational effects can be monitored by a phenotypic readout.

INTRODUCTION

Protein structure prediction is a difficult problem that has been addressed by various approaches. Homology-based modeling and threading are two widely used approaches to derive structural models for a given protein sequence (Eswar et al., 2006; Kiefer et al., 2009; Lobley et al., 2009; Meller and Elber, 2001; Sali and Blundell, 1993; Wu and Zhang, 2007). An alternative approach of ab initio structure prediction is used when no reliable homologs can be identified. Ab initio techniques have been successfully used to predict structures of small proteins (Rohl et al., 2004; Roy et al., 2010). The two major issues in ab initio structure prediction are generation of accurate models and selection of correct models from a large dataset of decoy models. In the present work, we explore the use of a residue-depth-based parameter for model ranking. Residue depth is the average minimum distance of the constituent atoms of a given residue from the nearest bulk water molecule (Chakravarty and Varadarajan, 1999; Tan et al., 2011). We estimate this parameter using mutant phenotype data obtained from deep sequencing and use this information to guide protein structure prediction. Deep sequencing has been successfully used for single-nucleotide polymorphism (SNP) detection (Van Tassell et al., 2008), genome sequencing, and genetic analysis (Bentley et al., 2008; Margulies et al., 2005). The tremendous power of this highly parallelized technique has been recently explored in deriving relative fitness estimates and sequence-function correlations for a subset of residues in proteins and nucleic acids (Fowler et al., 2010; Hietpas et al., 2011), determining protein:DNA interacting residues (Kinney et al., 2010), studying codon bias effects on fitness (Hietpas et al., 2011), and for RNA structural characterization (Lucks et al., 2011). We extend the application of deep sequencing combined with saturation mutagenesis to the field of protein structure prediction.

The experimental system used is the Controller of Cell Division or Death B (CcdB) protein. CcdB is an F plasmid encoded Escherichia coli toxin responsible for plasmid maintenance and is an inhibitor of DNA gyrase (Bernard and Couturier, 1992; Bernard et al., 1993; Dao-Thi et al., 2005). It is a homodimeric protein with 101 residues per protomer. The ccdB gene was cloned under the PBAD promoter so as to achieve tunable protein expression by varying the concentrations of arabinose (inducer) or glucose (repressor) (Chakshusmathi et al., 2004). Since the protein is a toxin, the cells bearing active CcdB mutants die, and the cells with inactive mutants survive. Since activity depends on the protein expression level, mutants that show an inactive phenotype at low expression levels typically show an active phenotype at higher levels (Bajaj et al., 2008). We have previously constructed a large library of ~1,500 single-site mutants of CcdB by site-directed mutagenesis (Bajaj et al., 2008). In the present study, we characterized each mutant in terms of its relative activity using next generation sequencing. This, in turn, was derived from its relative population in cells transformed with the library and plated in the presence of various repressor and inducer concentrations. From this data, we derived an empirical parameter, RankScore, which correlated with a structural parameter, residue depth. RankScore was further used to discriminate good structural models from a pool of CcdB decoys. The approach was extended to other proteins to show that RankScore information obtained from
such mutagenesis experiments can be used as a general tool to guide protein structure prediction.

RESULTS

Pooled Library Preparation and Data Statistics

The plasmid library of 1,430 site-directed mutants was transformed into the CcdB sensitive E. coli strain Top10pJAT. Transformed cells were plated on various LB plates containing seven different concentrations of glucose (repressor) and arabinose (inducer) arranged in order of increasing protein expression level. At each condition, only cells containing mutants that display an inactive phenotype will survive. The ccdb gene from plasmid pools isolated at each condition, as well as the original library, was amplified by PCR using primers containing a 10 base long Multiplex IDentifier sequence tag (MID) to encode the growth conditions. Tagged libraries were then pooled together and sequenced with the 454 GS FLX sequencer (Roche Inc., Basel, Switzerland). Reads at each condition (hereafter designated by the MID number) were processed and analyzed. Overall, ~250 megabases of data was obtained from all the eight libraries. Data statistics of mapped reads and bases are shown in Table S1 (available online). The overall read length distribution showed that most reads were >400 bases long (Figure 1), which is greater than the ccdb gene length (306 bases). The numbers of amino acid mutations per read are shown in Table S2. Fifty to sixty percent of the reads have only a single amino acid mutation, whereas 4%–9% reads showed more than one amino acid mutation per read. A large fraction of reads (35%–40%) showed zero detectable amino acid mutations per read. These reads do not necessarily correspond to wild-type (WT) reads, and the observation of zero mutations may be due to the stringent criteria used for mutation detection. Only reads with a single mutation were analyzed further. The total number of such reads is normalized to $10^5$ (arbitrary reference) at each MID. Normalized read numbers for all mutants at each MID are listed in Table S3. The fraction of inactive mutants decreases with increasing expression level from 32% at MID 2 to 7.4% at MID 8. Thus, at high expression levels, most single-site mutations were tolerated.

Mutational Sensitivity Score

WT CcdB shows an active phenotype and kills cells, even at the lowest expression level achievable here (MID 2, 0.2% glucose). The activity of the various CcdB mutants can be measured in terms of their effect on cell growth as a function of expression level/MID. A mutational sensitivity score ($MS_{seq}$) was derived as the MID at which cells transformed with the mutant are killed. In terms of the number of sequencing reads, the $MS_{seq}$ can be derived as the MID at which the number of reads for a particular mutant decreases by 5-fold or more relative to its previous MID. Interestingly, for most mutants the number of reads as a function of MID show a step-function-like behavior as opposed to a gradual decrease, suggesting that the observed phenotypes are a result of threshold effects. At low expression levels, a mutant with lowered activity does not affect cell growth and results in detectable reads. Once a threshold expression level is reached, all cells containing the mutant are killed, and hence the number of reads drops to zero. The $MS_{seq}$ ranges from two (mutants that show activity indistinguishable from WT) to nine (mutants that were still inactive at the highest expression level achievable in this study). Overall, $MS_{seq}$ was assigned to 1,176 mutants (Table S3). To validate the phenotypes and scores derived from sequencing data, 100 mutants spanning the complete $MS_{seq}$ range were chosen and individually transformed. Individual mutant phenotypes as a function of expression level were monitored by plating transformed cells for each mutant at multiple repressor and inducer concentrations to derive a plate-based mutational sensitivity score ($MS_{plate}$). These scores were compared with those derived from sequencing data (Figure 2). Both scores were highly correlated (correlation coefficient ~0.95 and slope ~0.96) validating the phenotypes derived from deep sequencing. In addition, the sequences of the mutants obtained from deep sequencing were in good agreement with the known sequences of the site-directed mutants that had been pooled to create the library. Of the 1,176 mutants detected in the pooled library by deep sequencing, the sequences of 1,035 mutants were identical to the corresponding mutants that had been created by site-directed mutagenesis. Though we have not performed rigorous analysis for error estimation, there was highly correlation between $MS_{seq}$ and $MS_{plate}$ and also detection of 88% of the true positive mutations (1,035 of 1,176) are indicative of low error rates.

RankScore

A parameter called RankScore was defined, which combines activity data at various expression levels. A rank was assigned to every mutant based on the population distribution of $MS_{seq}$ values (Table 1; see Experimental Procedures). The average rank of all mutants at a particular residue position was termed as the RankScore of that position. A lower RankScore indicates higher mutational tolerance. It was found that surface amino acids exhibited low RankScore values, whereas buried ones had high RankScore (Figure 3). The hot-spot active-site residues had very high RankScores. The RankScore was better correlated with residue depth (averaged over all atoms in the residue) than with accessibility (correlation coefficients of ~0.61 and ~0.53,
respectively). When the top 10% of residues with the highest RankScores are excluded, the RankScore showed enhanced correlation ($r^2 = 0.65$) with residue depth, rather than with accessibility ($r^2 = 0.52$). This is because such potential active-site residues have low depth but are highly sensitive to mutation and therefore have high RankScores. The correlation with depth improves if average values within a bin are used ($r^2 = 0.91$ and 0.41 for depth and accessibility, respectively; Figure 4), suggesting that mutational effects at positions buried just below the solvent accessible surface of the protein are different than those at positions buried deep within the core of the protein.

**Model Discrimination Based on RankScore**

If a structural model is similar to the native structure, residue depth from the native structure and models should be highly correlated. Since we have shown that RankScore is correlated to residue depth from the native structure, in principle, the correlation coefficient between RankScore and model residue depth ($r_{\text{score}}(\text{depth}(e))$), where $e$ stands for experimental, can be used for model discrimination. To test this hypothesis, a decoy set was generated for CcdB by refining models obtained from threading. 27 threading alignments (from eight different dimeric templates with different RMSD values from the CcdB crystal structure, Protein Data Bank [PDB] id 3VUB; Loris et al., 1999) were used for model generation. Each of these models was further refined using the Rosetta refinement protocol (Raman et al., 2009) to generate 300–500 models per alignment (see Experimental Procedures). Of these, 50 randomly chosen models per alignment were used for model discrimination (Figure 5A). The performance of $r_{\text{score}}(\text{depth}(e))$ was compared with the model discrimination ability of the Rosetta energy function (Figures 5C and 5D). The Rosetta energy function has been used to rank models obtained from ab initio simulations, and it has been shown that for proteins

![Figure 2. Correlation of MS Scores from Deep Sequencing with Those Obtained from Plating Individual Clones](image)

(A) Correlation of MS scores for 100 CcdB mutants obtained from plating experiments of individual mutants and derived from 454 sequencing data. MS\text{plate} score for each mutant was assigned as the MID at which the mutant colonies disappear on plates. The mutants were chosen so to span the complete MS\text{seq} range. In the case of 454 deep sequencing data, the MS\text{seq} score was the MID at which the number of reads decrease by 5-fold or more relative to the previous MID. The MID here represents various glucose/arabinose concentrations arranged in order of increasing expression level. The regression fit is shown as a solid line. The slope was 0.96 and correlation coefficient is 0.95.

(B) Since several data points overlapped (i.e., different mutants had identical values of MS\text{seq} and MS\text{plate}), the x-coordinates for such points were spread randomly over ± 0.25 units to show all the data points.

(C) An experimental derivation of the mutational sensitivity score on plates (MS\text{plate}). 40 representative CcdB mutants along with the wild-type and a negative control (thioredoxin, Trx) were spotted on LB agar plates containing various glucose/arabinose concentrations corresponding to that used at various MIDs. MS\text{seq} was assigned as the MID at which the colonies on the plate disappear and corresponds to the inducer concentration at which a given mutant shows an active phenotype. Cells transformed with WT CcdB do not survive under any conditions, whereas those transformed with a nontoxic control, E. coli thioredoxin (Trx), grow under all conditions.

### Table 1. Rank Assignment Based on the Population Distribution of Relative Activities in the Mutant Pool

<table>
<thead>
<tr>
<th>MS\text{seq}</th>
<th>Incremental No. of Active Mutants</th>
<th>Cumulative No. of Active Mutants</th>
<th>Population (%)</th>
<th>Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>796</td>
<td>796</td>
<td>68</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>908</td>
<td>69</td>
<td>69</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>822</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>852</td>
<td>72</td>
<td>71</td>
</tr>
<tr>
<td>6</td>
<td>122</td>
<td>974</td>
<td>83</td>
<td>73</td>
</tr>
<tr>
<td>7</td>
<td>46</td>
<td>1,020</td>
<td>87</td>
<td>84</td>
</tr>
<tr>
<td>8</td>
<td>69</td>
<td>1,089</td>
<td>93</td>
<td>88</td>
</tr>
<tr>
<td>9</td>
<td>87</td>
<td>1,176</td>
<td>100</td>
<td>94</td>
</tr>
</tbody>
</table>

*Increasing values of MS\text{seq} are associated with increasing levels of CcdB expression.
less than about 100 residues, the models with the lowest energy are typically close to the actual native structure (DiMaio et al., 2011; Kuhlman et al., 2003; Rohl et al., 2004). The number of models selected having backbone RMSD ≤ 4Å in the top 10, 25, 50, or 75 models based on rscore depth ðeÞ were 50%–130% better (p value = 0.014) than were the corresponding numbers in respective RMSD ranges when selected using Rosetta energy (Figure S1).

Simulated RankScore
To examine the applicability of our approach for model discrimination to other proteins for which we did not have experimental RankScore values, we assigned a set of random RankScore values (in the range of 1–100) to residues in the target protein subject to the constraint that this set of values should have a rscore depth ðsÞ value of ~0.6 with native residue depth values (derived from the crystal structure of the target protein). The generated values are termed as simulated RankScore (rscore depth ðsÞ). Ten sets of simulated RankScores were generated and used for model discrimination on the CcdB decoys (Figures 5B and S2). CcdB decoys were generated as described in the Experimental Procedures. Ten models from each set with the highest rscore depth ðsÞ were selected and sorted to obtain the unique models (81 models). The RMSD distribution of these models was compared with the same number of models selected based on either experimental RankScore or Rosetta energy (Figures 5C and 5D). 79 out of 81 selected models had backbone RMSD ≤ 4Å when experimental RankScore was used for model discrimination as compared to 64 and 35 models when simulated RankScore or Rosetta energy respectively were used for model discrimination. Furthermore, structural clustering of the selected 81 models was performed, and a maximum of the top five clusters were considered. Clusters were ranked based on the average correlation coefficient of cluster members. The backbone RMSD of cluster centers, with respect to the crystal structure of CcdB, selected based on rscore depth ðeÞ (2.51, 3.06, 3.34, and 2.74) or rscore depth ðsÞ (4.99, 2.60, 3.02, 2.86, and 2.51) were found to lie at lower values compared to when Rosetta energy was used for model discrimination (5.39, 2.36, 4.16, 4.48, and 2.8). This suggests that rscore depth ðsÞ can be used as an alternative metric to energy-based parameters for model discrimination. Although the simulated RankScore performs worse than the experimental RankScore in terms of model discrimination, it is still sufficiently accurate and hence can be used to test the discriminatory ability of rscore depth ðsÞ on other proteins of known structure for which experimental RankScore values are not yet available.

Alternative Parameter for Model Discrimination
The hot-spot gyrase binding residues of CcdB were observed to have high RankScore values. In general it is likely that binding-site/active-site residues will be spatially clustered in a single surface patch (Guharoy and Chakrabarti, 2010; Landgraf et al., 2001; Schueler-Furman and Baker, 2003). A spatial proximity parameter (R) was defined as the radius of gyration of CcdB.
residues (excluding Gly and Pro) having RankScore $\geq$ \(\text{RankScore(\text{mean})} + 1.5\sigma\). A large fraction of selected residues with high RankScore are expected to be part of the binding/active site. Consequently, if a model is correctly folded, these residues should be spatially clustered and hence should have a lower R value relative to incorrect models. Gly and Pro were excluded because these have specific stereochemical properties and may be unusually sensitive to mutation, even when not part of the active-site. In the case of CcdB, nine residues satisfy this criterion, of which three are part of the active site, three are within 5Å of one of the active-site residues, and the remaining three are at buried sites. As shown in Figure 6A, the models with low R values lie in the lower RMSD range for the CcdB decoy set. For the lowest 5% of R values (67 models), 88% of models have an RMSD of $\leq 4\text{Å}$ from the native structure.

Furthermore, simultaneous use of both parameters, R and \(\text{rscore\_depth}\) in case of CcdB was found to be helpful in filtering out some of the higher RMSD models that were selected when one of the two parameters was used (Figures 6B and 6C). The inset of Figure 6C features those models that have the highest 5% of \(\text{rscore\_depth}\) and the lowest 5% of R values. These models are highly enriched in low RMSD structures. Hence, in addition to \(\text{rscore\_depth}\), R and/or a combination of both parameters is also a useful metric to enrich native-like models from a pool of decoys.

**Extension to Other Proteins**

The proposed approach for model discrimination based on \(\text{rscore\_depth}\) was initially extended to seven other proteins for which crystal structures are known. Approximately two monomeric proteins each, from all four structural classes were chosen with size in the range of 100–150 amino acids. For each of the seven proteins, $10^5$ ab initio models were generated by fragment assembly using the Rosetta Abinitio module. For five of these seven proteins, where ab initio modeling did not generate any models with backbone RMSD $\leq 4\text{Å}$, the decoy sets were generated by threading as discussed in the Experimental Procedures. Briefly, models were built for 7–9 threading alignments and further refined using Rosetta to generate 300–500 models per alignment. Of these, 50 randomly chosen models per alignment constituted the decoy set (Table 2). Ten sets of randomly assigned RankScores were simulated for every protein based on the crystal structure depth with the constraint that the correlation coefficient between RankScore and native depth was $0.6 \pm 0.01$. The model selection protocol was the same as that followed for CcdB. The backbone RMSD distribution of models was compared with the threading alignment used. Similar plots for nine other similarly randomly assigned sets of RankScore values are shown in Figure S2. The top panel shows the distribution of Rosetta Energy values as a function of RMSD from the crystal structure for the dataset of decoys. The dashed line separates the 81 lowest energy models from the rest. The middle and lower panels show the 81 best models as a function of RMSD from the CcdB crystal structure. Models are selected based on \(\text{rscore\_depth}\) when experimental (middle) or simulated (lower) RankScores were used (see text for details).

(D) Backbone RMSD distribution of selected models in (C) when using RankScore derived from deep sequencing (black), simulated RankScores (red), and Rosetta energy (green). The total RMSD range was divided into 1Å bins. Both experimental and simulated RankScore showed improved discrimination between good and bad models relative to the Rosetta energy function. See also Figure S1 and S2.
models selected based on \( r_{s} \) was compared with the same number of models selected based on Rosetta energy. The distribution of \( r_{s} \) and Rosetta energy as a function of backbone RMSD is shown in Figures 7A, 7B, and S3 for a specific case of the Hpt domain of the anaerobic sensor kinase ArcB (PDB id 2A0B; Kato et al., 1999). When \( r_{s} \) was used for model discrimination, the number of selected models with backbone RMSD \( \leq 4 \) Å were nine as compared to two with Rosetta energy (Figure 7C). Also, a large number of higher RMSD models (>10 Å) were selected by Rosetta energy (89 models) as compared to \( r_{s} \) (50 models; Figure 7C).

In an alternative set of evaluations, four additional proteins were chosen from a set of targets previously evaluated by I-TASSER (Roy et al., 2010). Three of these were considered to be hard targets, and the remaining one was considered a medium target by I-TASSER based on statistical significance of alignments (Roy et al., 2010). For these four proteins (70–90 residues length), \( 2 \times 10^{6} \) decoys were generated individually using I-TASSER. A similar model selection protocol based on \( r_{s} \) was followed as discussed previously. The selected models were structurally clustered, and the clusters were ranked based on the average correlation coefficient of the members. The backbone RMSDs’ of the top ten cluster centers were found to lie in a significantly lower range when compared with the RMSD of the top ten models predicted by I-TASSER (Table 2, bottom 4 rows).

The overall performance of \( r_{s} \) in terms of model discrimination is summarized in Table 2. The \( r_{s} \) is able to select low RMSD models from large pools of decoys for all of the cases examined. The CcdB data suggests that a substantial further improvement in the predictions is likely if experimental, rather than simulated, \( r_{s} \) and \( r \) values are used.

**DISCUSSION**

Reliable protein models are sometimes difficult to obtain when there is low sequence identity between the target and proteins of known structure. The present studies demonstrate that Saturation Mutagenesis phenotypes Analyzed by Deep Sequencing (abbreviated as SMADS) can guide protein structure prediction. A flowchart of the SMADS methodology is shown in Figure 8. The RankScore and \( r \) parameters described previously, used either individually or in combination, are able to extract good models from a large pool of decoys. The Rosetta energy-based ab initio modeling approach works remarkably well for small proteins but is less efficient for proteins greater than 125 residues. In contrast, the RankScore performs better for proteins having a large number of deep residues (approximately >10 residues with depth \( \geq 6 \) Å) and on average proteins larger than \( \sim 75 \) residues satisfy this condition. In the present work, we observe that mutational tolerance is more closely related to residue depth than to accessibility, as all residues below the surface have accessibilities close to zero but vary in depth. In several previous studies, including our own, residues have been categorized as either buried or exposed, distinguished by arbitrary accessibility cut-off typically ranging from 5%–15% (Ahmad and Gromiha, 2002; Bajaj et al., 2005; Miller et al., 1987). One of the initial goals of this work was an attempt to derive an appropriate burial cut-off based on mutational sensitivity. Instead, it was observed that mutational sensitivity and hence RankScore correlate with residue depth. Hence, using this biologically relevant readout, there is no appropriate “cut-off” with which to distinguish buried and exposed residues. Instead, precisely because of such a correlation, it is possible to use mutational data to guide model discrimination, in an energy independent manner. This is a useful complement to existing energy-dependent computational methods.

The current definition of RankScore takes into account only the population distribution of \( MS_{eq} \) scores and might be improved by inclusion of terms like change in polarity, size, and/or hydrophobicity upon mutation. Furthermore, if experimental RankScore information can be incorporated into ab initio structure building in a manner similar to X-ray diffraction data (DiMaio et al., 2011) or NMR chemical shifts (Shen et al., 2008), this might further guide the generation of correct models.

The present work also outlines a rapid and accurate methodology for obtaining relative populations of different mutants in a library under varying conditions. We observe that the largest decreases in fitness result from mutations at active site residues. Since these residues are likely to be spatially clustered, this provides an additional metric to discriminate correctly folded models in a pool of decoys. Additionally, it can also be used for active-site identification/prediction and functional annotation.
Table 2. Backbone RMSD Distribution of Models Generated by Either Ab Initio Fragment Assembly or Threading for Various Proteins and Overall Performance of roscore depth ($\delta$) and Rosetta Energy / I-TASSER in Terms of Model Discrimination.

<table>
<thead>
<tr>
<th>PDB id</th>
<th>Length</th>
<th>Decoy Set Secondary Str b</th>
<th>No. of Selected Models c</th>
<th>No. of Selected Models with RMSD (\leq 4) Å d</th>
<th>Mann-Whitney Test p Value f</th>
<th>RMSD of the Top 5 Models after Clustering at Radius 2 Å (Minimum Cluster Size = 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size</td>
<td>RMSD Range</td>
<td>%H</td>
<td>%S</td>
<td>%L</td>
<td>No. with RMSD (\leq 4) Å</td>
</tr>
<tr>
<td>1LU4g</td>
<td>134</td>
<td>2-17.4</td>
<td>45</td>
<td>23</td>
<td>54</td>
<td>47</td>
</tr>
<tr>
<td>3MYTg</td>
<td>146</td>
<td>2-17.1</td>
<td>100</td>
<td>68</td>
<td>32</td>
<td>48</td>
</tr>
<tr>
<td>1TP6g</td>
<td>126</td>
<td>2-9.7</td>
<td>236</td>
<td>32</td>
<td>47</td>
<td>21</td>
</tr>
<tr>
<td>1VM9g</td>
<td>109</td>
<td>2-11.9</td>
<td>82</td>
<td>7</td>
<td>39</td>
<td>53</td>
</tr>
<tr>
<td>1WOUg</td>
<td>119</td>
<td>3.5-10.7</td>
<td>79</td>
<td>23</td>
<td>28</td>
<td>78</td>
</tr>
<tr>
<td>1VYFg</td>
<td>135</td>
<td>2-20.0</td>
<td>123</td>
<td>17</td>
<td>56</td>
<td>96</td>
</tr>
<tr>
<td>2A0Bt</td>
<td>118</td>
<td>2-17.3</td>
<td>123</td>
<td>17</td>
<td>56</td>
<td>96</td>
</tr>
<tr>
<td>1kjs_k</td>
<td>74</td>
<td>3.5-26.4</td>
<td>37</td>
<td>25</td>
<td>38</td>
<td>99</td>
</tr>
<tr>
<td>1mkyA3k</td>
<td>81</td>
<td>3.2-18.3</td>
<td>132</td>
<td>17</td>
<td>56</td>
<td>96</td>
</tr>
<tr>
<td>1o2fB_k</td>
<td>77</td>
<td>3.3-17.4</td>
<td>2216</td>
<td>62</td>
<td>27</td>
<td>51</td>
</tr>
<tr>
<td>1tig_k</td>
<td>88</td>
<td>2-10.7</td>
<td>2216</td>
<td>62</td>
<td>27</td>
<td>51</td>
</tr>
</tbody>
</table>

a Backbone RMSD range (in Å) observed in the decoy set.
b Secondary structure content from the respective crystal structures.
c Total number of unique models selected based on RankScore or Rosetta energy.
d Number of models with backbone RMSD \(\leq 4\) Å selected based on RankScore or Rosetta energy.
e \(\text{rscore depth}(\delta)\) is the correlation coefficient of simulated RankScore versus model residue depth.
f Nonparametric Mann-Whitney test (one-tailed) to assess if the median of the RMSD distribution of the selected models based on \(\text{rscore depth}(\delta)\) is significantly lower as compared to that based on Rosetta energy.
g Very few low RMSD \(\leq 4\) Å models were generated for these proteins by the ab initio approach. Hence, a decoy set of 300–450 models were generated by refining threading derived models as described in the Experimental Procedures and used for model discrimination. The median of the RMSD distribution of models selected based on \(\text{rscore depth}(\delta)\) is lower than that based on the Rosetta energy.
h The median of the RMSD distribution of models selected based on \(\text{rscore depth}(\delta)\) is higher than that based on the Rosetta energy.
i Decoy datasets comprise exclusively of 10 models derived using the Rosetta Abinitio module.
j Clustering done using radius 3 Å.
k Decoy set generated using I-TASSER. The targets were selected as described in the Experimental Procedures.
l Test of significance was performed for RMSD values of the top ten cluster centers selected based on \(\text{rscore depth}(\delta)\) and the top models predicted by I-TASSER.
There have been recent attempts to use deep-sequencing to evaluate activities/fitness of mutant proteins. These have been typically restricted to a small stretch of sequences (Fowler et al., 2010; Hietpas et al., 2011). In the present study, we show that using a parallelized site-directed mutagenesis strategy, it is possible to accurately determine phenotypic information for all possible single-site mutants of a protein from a pooled library. The fact that each clone in the library is mutated at a single codon, greatly simplifies interpretation of the data. The demonstrated methodology can be applied, in principle, to any protein system in which structural modulations due to

Figure 7. Model Discrimination for a Decoy Set of $10^5$ Ab Initio Models Generated for PDB id 2A0B

(A) Model discrimination using $r_{\text{score}}$ and $\text{depth}$. A representative set of $r_{\text{score}}$ values are shown here. Similar plots for nine other sets of $r_{\text{score}}$ versus RMSD are shown in Figure S3. Models with ten highest $r_{\text{score}}$ values are shown above the dotted line.

(B) Model discrimination based on Rosetta energy. Models with the 95 lowest Rosetta energy values are shown below the dotted line.

(C) Backbone RMSD distribution of 95 selected models based on $r_{\text{score}}$ from ten independent simulations (filled bars, see text of Figure S3 for details) and Rosetta energy (empty bars, data from B). The total RMSD range was divided into 1Å bins. $r_{\text{score}}$ shows improved discrimination between good and bad models relative to the Rosetta energy function (see also Figure S3).

Figure 8. A Flowchart of the SMADS Methodology

(A) Experimental method.

(B) Data analysis.
mutations are reflected in a phenotypic readout. With advancements in high-throughput mutant generation and screening, and next-generation sequencing techniques, such studies can be completed in a relatively short time span for any protein/nucleic acid system for which a phenotypic readout is possible. For many proteins, a binding partner or ligand can be used to construct such a readout. Even in the absence of any binding partner, cell-surface expression of an epitope-tagged protein can be used as a convenient readout (Chao et al., 2006). In contrast to conventional methods for structure determination, such as crystallography or NMR, such readouts do not require proteins to be purified. Hence, this may prove useful for proteins, such as glycoproteins, membrane proteins, and other macromolecular systems, that are difficult to structurally characterize by crystallography or NMR. In addition, the validated saturation mutagenesis phenotypes and RankScore data shown in Table S3 provide a rich resource for understanding effects of mutations on protein stability, folding, and evolution.

EXPERIMENTAL PROCEDURES

Plasmids and Host Strains

The ccdB gene was cloned under the control of the PBAD promoter in the vector pBAD24 to achieve tunable expression levels using arabinose (inducer) or glucose (repressor) (Chakshumathi et al., 2004). Two E. coli host strains were used: Top10pJAT and CSH501. Top10pJAT is a CcdB-sensitive strain and was used for screening the phenotypes. The pJAT8araE plasmid encoding the arabinose transporter araE was introduced into Top10 strains to ensure uptake of approximately equal amounts of arabinose in all cells (Bajaj et al., 2008). The plasmid pJAT8araE was kindly provided by Dr J.D. Keasling (University of California, Berkeley, CA, USA). CSH501 is a CcdB-resistant strain due to a mutation in the chromosomal copy of gyrA, which abolishes DNA gyrase-CcdB binding. CSH501 was kindly provided by Dr M. Couturier (Universite Libre de Bruxelles, Brussels, Belgium). Individual mutant plasmids transformed into CSH501 were stored as glycerol stabs (Bajaj et al., 2008).

Mutant Library Generation

1,438 single-site mutants of CcdB (Bajaj et al., 2008) from E. coli CSH501 were inoculated individually into 96 deep well plates containing 1 ml Terrific Broth per well with 100 μg/mL ampicillin as antibiotic marker, with the aid of a Biomek 3000 liquid handling system (Beckman Coulter Inc). The cultures were grown until saturation (~36 hr) at 180 rpm and 37°C. 500 μl of the saturated culture from each well was pooled together and harvested at 6000 rpm for 15 min at 4°C. The master pool of plasmids was purified from the pooled culture using a QiaGen kit and the presence of the CcdB insert was confirmed by sequencing. The pJAT and CSH501 Top10 plasmids were recovered from the master pool was amplified using a set of forward and reverse primers, which contained multiplex identifier (MID) sequences (10 bases long) unique for each condition. ~122 ng for MID 1 and ~61 ng each for MID 2–8 of purified PCR products (as estimated from a Bioanalyzer, Agilent Technologies) were mixed together, and the pooled library was sequenced using the 454 Genome Sequencer FLX platform (Roche Inc., Basel, Switzerland).

Data Processing

All reads with a minimum read-length of 200 bases were aligned with the WT CcdB sequence using ClustalW (Chenna et al., 2003). A mutation at a codon level was assigned if (1) the quality score (Q) of all bases of the codon was ≥20 and (2) no base in ± 5 bases had Q < 20 around the mutant codon. All instances of the putative mutant codons per read were subjected to these criteria. The reads with only a single amino acid mutant codon were considered for further analysis (~23% of total reads).

Mutational Sensitivity and RankScore

For a given mutant, a mutational sensitivity (MSseq) score is assigned as the MID at which the normalized number of reads (normalized to 105 total reads for each MID, an arbitrary reference) decrease by 5-fold or more compared to its previous MID. This score ranges from 2−9 with two representing the most active and nine the least active, that is, the most sensitive to mutation. A parameter called RankScore was defined for each residue based on the overall distribution of mutational sensitivity scores. All mutants with a MSseq score of two were assigned a rank of one (say a% of total mutants). The next rank was assigned as (a + 1) to all mutants (say b%) having a MSseq score of three and thus accounting for a cumulative total of (a + b)% of mutants. Furthermore, all mutants (say c%) with a MSseq score of four were assigned a rank of (a + b + 1) and so on. The ranks are thus related to the cumulative distribution of mutants with the most active mutants having the lowest ranks and the least active having the highest. The RankScore for a position was defined as the numerical average of the assigned ranks to all mutants at that position.

Ab Initio Model Generation

Ab initio models for monomeric proteins were generated by using the Rosetta (version 2.3) ab initio protocol (Kuhlman et al., 2003; Rohl, 2005; Rohl et al., 2004). A database of proteins with <30% sequence identity to the target protein was used for fragment generation. ~105 models were generated with the following command line options: rosetta -fa_output -new_centroid_packaging -vary_omega -omega_weight 0.5. Ab initio models for the CcdB homodimer were generated using the fold-and-dock protocol of Rosetta as described previously (Das et al., 2009).

Model Generation by Threading and Refinement

The query protein sequence was submitted to the LOMETS metaserver (Wu and Zhang, 2007). The search database contained protein sequences with <30% identity to the query sequence. In the case of CcdB, models were generated using alignments only with dimeric templates. For the remaining monomeric proteins, the models were generated from LOMETS-derived template alignments using MODELLER (Eswar et al., 2006; Sali and Blundell, 1993). 7–9 templates were selected from various backbone RMSD ranges with respect to the crystal structure of the target (care was taken not to overpopulate any specific RMSD range) and further refined using Rosetta with the following options: rosetta -relax -farx -1 -ex2 -short_range_hb_weight 0.5 -long_range_hb_weight 1.0 -farx_cycle_ratio 1.0 to generate ~300–500 models for each template. Dimeric models of CcdB generated by threading were refined using following options of Rosetta (version 3.2): rosetta -relax: thorough -relax:chi_move -symmetry: symmetry_definition. 50 randomly chosen models from each of these refined sets were used for further analysis. The backbone RMSD was calculated with respect to the native crystal structure using the g_confrms tool of Gromacs (version 4.0.5; Hess et al., 2008; Van Der Spoel et al., 2005).

Model Generation by I-TASSER

Four targets were selected from the target list used by I-TASSER (Zhang, 2008), of which three were hard targets (1mkvA3, 1o2tb.., and 1tj..), and one was of medium difficulty (1kjs..). Decoy sets of 2 × 104 models were generated for each target using a stand-alone version of I-TASSER (Roy et al., 2010). A database of proteins with <20% sequence identity to the target protein was used as the search dataset. Clustering and subsequent refinement protocols of I-TASSER were used to obtain ten predictions for each target from
Structural Clustering and Ranking
The selected models were structurally clustered using g_cluster tool of Gromacs (version 4.0.5) with a clustering radius of 2 Å. The minimum cluster size was set to two. Clusters were ranked based on the average correlation coefficient of the cluster members. The backbone RMSD of the cluster center with the native structure was reported.

Tests of Significance
A nonparametric, one-tailed Mann-Whitney test was performed using GraphPad Prism (version 5.01 for Windows, GraphPad Software, San Diego, CA, USA, http://www.graphpad.com) to assess if models selected based on RankScore have significantly lower RMSD from the native structure compared to models selected based on Rosetta energy.

SUPPLEMENTAL INFORMATION
Supplemental Information includes three figures and three tables and can be found with this article online at doi:10.1016/j.str.2011.11.021.

REFERENCES
Deep Sequencing-based Protein Model Discrimination


