Published in final edited form as:
Proteins. 2011 October ; 79(10): 2844–2860. doi:10.1002/prot.23109.

Computed structures of point deletion mutants and their enzymatic activities

Monica Berrondo and **Jeffrey J. Gray**
Chemical & Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland, 21218

Abstract

Point deletions in enzymes can vary in effect from negligible to complete loss of activity, however, these effects are not generally predictable. Deletions are widely observed in nature and often result in diseases such as cancer, cystic fibrosis, or osteogenesis imperfecta. Here, we have developed an algorithm to model the perturbed structures of deletion mutants with the ultimate goal of predicting their activities. The algorithm works by deleting the specified residue from the wild-type structure, creating a gap that is closed using a combination of local and global moves that change the backbone torsion angles of the protein structure. On a set of five proteins for which both wild-type and deletion mutant x-ray crystal structures are available, the algorithm produces deep, narrow energy funnels within 1.5 Å of the crystal structure for the deletion mutants. To assess the ability of our algorithm to predict activity from the predicted structures, we tested the correlation of experimental activity with several measures of the predicted structure ensemble using a set of 45 point deletions from ricin. Estimates incorporating likely prevalence of active and inactive deletion sites suggest that activity can be predicted correctly over 60% of the time from the active site rmsd of the lowest energy predicted structures. The predictions are stronger than simple sequence organization measures, but more fundamental work is required in structure prediction and enzyme activity determination to allow consistent prediction of activity.

Keywords

prediction algorithm; enzyme activity; loop modeling; backbone relaxation

Introduction

The exploration of mutations in proteins and their effect on protein structure, function, and stability has led to the belief that proteins are remarkably plastic. For many positions in proteins, wide varieties of amino acid substitutions are tolerated without drastic effects on the protein's structure or function. Such changes are often accommodated by small losses in stability, activity, or function. The category of mutations envisioned to lead to the largest effect on protein structure and function is an amino acid deletion because the rearrangement necessary to accommodate the deletion goes beyond a simple perturbation of the surrounding amino acid side chains.[1] However the role of amino acid deletions in proteins is largely unknown due to the difficulty in producing in-phase amino acid deletions on specified target regions of proteins.[2,3] Furthermore, few studies have examined the effects of point deletions on protein structure, and few point deletion mutant structures are recorded in the Protein Data Bank (PDB).[4] While the effect of deletions on residues near the active site

Corresponding author: Jeffrey J. Gray, 3400 N. Charles St., Baltimore MD, 21218, Phone: (410) 516-5313, Fax (410) 516-5510, jgray@jhu.edu.

may be straightforward, deletions far from the active site of a protein can have surprising effects on both activity and stability.[5–7]

Recently, several groups have been able to create directed single-amino-acid deletions in proteins to systematically determine their effects on protein activity and stability.[5–8] Munishkin and Wool[6] and Simm *et al.*[7] deleted single residues in ricin and TEM-1 β-lactamase, respectively, and tested the resulting enzymatic activity after the modification. Ricin is a toxic heterodimeric glycoprotein found in the seeds of *Ricinus communis* (castor bean). The A chain (RTA) is an RNA N-glycosidase that inactivates ribosomes by depurination of a single adenosine of 28S rRNA. We selected ricin as our model system because of the quality of the experimental data; in contrast to the activity assay for β-lactamase which was done in whole cell extracts, the ricin data were collected from an assay in lysate depleted of endogenous mRNA, ensuring isolation of the activity of the ricin protein. The assay for activity assesses the presence or absence of residual activity of ricin mutants and can detect activity up to 10,000-fold reduction from wild-type, providing a clear set of biophysical data for comparison to calculations.

Since the interiors of proteins are packed very tightly, one might expect that removing an amino acid from the core of a protein will destabilize the interactions necessary to maintain an energetically favorable conformation. Furthermore, deletions from regions of secondary structure change the register of those elements and therefore cause downstream effects on the overall structure of the protein.[9] However, 222 of the 267 amino acids of ricin can be deleted in one to twenty residue segments without negative effects on the activity. Of the remaining 45 of those deletions, nine can be deleted individually without affecting activity.[6] Of the individual deletions that retained activity, several of them were near the active site and many were in segments of well-defined secondary structure (Figure 1). Similar results were found for β-lactamase, indicating that proteins are much more tolerant of large perturbations than previously expected.[7]

A particular approach to examining the plasticity of protein structure and function is to computationally predict the effect of a single amino acid mutation on the overall structure of the protein. A possible method for examining the activity of a mutant protein is to create a plausible structure or ensemble of structures that can be subjected to different measures of activity. A structure can be created by *ab initio* folding of the mutant sequence or by starting from the x-ray crystal structure of the wild type protein and retaining some of the structural aspects already present. Since it is expected that a mutation will have a small effect on the overall structure of the protein, we chose to start from the x-ray crystal structure of the wild-type protein. *Ab initio* protein folding has traditionally targeted proteins of 150 residues or less[10] but the average length of enzymatic proteins is 300 residues.[11] The limiting factor for achieving predictions of larger proteins is the conformational space that must be sampled and refined.[10,12–14] Using the crystal structure makes it possible to address proteins that are too large for use with current methods for *ab initio* folding.[10,15,16]

Algorithms for structure refinement seek to overcome the difficulty of optimizing a medium resolution structure to one with sufficient resolution detail to answer relevant biological questions, such as changes in the activity level of the protein. In recent years, structure refinement algorithms have progressed and are now able to refine structures to high accuracy (< 2.0 Å).[10,16] The problem of starting from a wild-type x-ray crystal structure to predict a deletion mutant structure is similar to structure refinement. We sought to leverage available refinement algorithms and optimize them so that they are able to predict the structure of deletion mutants.

The challenge of using the wild-type x-ray crystal structure is that deleting an amino acid from the x-ray crystal structure will create a gap, or hole, in the structure in place of the deleted amino acid (Figure 2B). Closing the resulting gap is not trivial. If backbone torsion angle changes are only made in the nearby residues, global changes throughout the rest of the protein resulting from the deletion will not be captured. Similarly, if backbone torsion angle changes are made globally, throughout the entire protein, without considering the local environment of the mutation, unrealistic geometries may occur at the point of closure and its surrounding area.

Once a structure has been predicted, it could be used to estimate activity. However, despite the fact that computational methods have had recent success in designing enzymes,[17–20] predicting the biochemical activity of an enzyme remains a challenge since there is no clear and well-defined way of computing activity directly from structure. The lack of sufficient progress has made it difficult to relate structural changes in a protein to the changes in enzymatic activity.[21–23] Not surprisingly then, in most cases, correlating predictions of mutant protein activity with biochemical measurements have not been markedly successful and even when deemed successful, it is the *stability* of the mutant protein, not the resulting activity level, that has been examined. [21,24–29] Without structural information, the sequence of a protein as compared to other homologous proteins has also been shown to be valuable in determining the effect of mutating dispersed single residues.[30,31] However, in the successful instances, a very large dataset of mutated residues and homologous sequences was necessary to infer activity.

In this work, our goal was to test whether it is possible to predict the relative level of activity of proteins with single point deletions. To accomplish this goal, we developed structure prediction methods which accommodate both local and global responses of the residues deleted. The algorithm was tested on a small set of proteins for which both wild-type and deletion-mutant x-ray crystal structures were available. The overall algorithm begins with the x-ray crystal structure of the wild-type sequence, deletes the residue of interest, and outputs an ensemble of candidate structures for examination. A variety of metrics are then applied to the ensemble of structures in an attempt to characterize the activity of the mutant protein. Despite the uncertainties in structure and enzymatic activity prediction, examining the active site rmsd of the predicted structures and their rmsd distribution improves prediction of the activity of mutant proteins over simple distance based methods. Nevertheless, we show that further progress is yet needed for this biologically important, but understudied, problem.

Results

Deletion algorithm

At the core of our structure prediction method are combinations of new conformational moves that are used to decrease the amount of time needed to sample the search space of a large protein. The general algorithm for calculating the structure and characterizing the activity of a mutant protein consists of three steps and is shown in Figure 2. (1) Starting from the native, wild-type crystal structure of the desired protein (Figure 2A), the residue of interest is deleted, leaving a hole in the structure (Figure 2B, ligand atoms not shown). (2) The hole left by the deleted residue is repaired and a diverse ensemble of 2000 structures is created through changes in random torsion angles throughout the protein (Figure 2C). (3) The low-energy ensemble of structures is examined for proxies of activity (Figure 2D). All ligand atoms are included during all steps of the algorithm. The ligand is allowed to rotate and translate (±5°/2Å) within the binding pocket, however no ligand flexibility is allowed and the internal conformation is kept fixed.

The deletion algorithm, incorporated into Rosetta, exploits a multi-scale approach combining a low-resolution mode where side chains are represented as pseudo-atom centroids[32] and a high-resolution mode with explicit, all-atom side chains.[33] We developed a number of moves in both the low- and high-resolution modes to increase effectiveness of the algorithm and to allow us to model large proteins. Our algorithmic improvements focused on three main challenges: (1) loop closure, (2) structure diversification, and (3) refinement efficiency. The loop closure pieces were implemented in low resolution where large movements to the backbone are more efficient. Refinement and structure diversification techniques were used in both low- and high-resolution, with the larger moves occurring in low resolution. Once the side chains are reconstructed on the protein backbone, the number of degrees of freedom increases tremendously, therefore most of our efforts for improving speed were targeted to the high-resolution, all-atom, steps.

Loop closure

Deleting a residue from a structure of a protein creates a hole in the polypeptide chain that must be closed to recreate a single chain. The most common method for closing gaps in proteins is through loop building.[34] Loop building typically confines the area of perturbation to within a couple of residues of the gap. While this is a desired approach when examining the local structure of a loop that must be rebuilt, *e.g.* in homology modeling, the presumed effects of deleting a residue suggest that a more global approach to closing the gap is necessary. Therefore, we developed a method combining simple backbone relaxation techniques with loop building to create a global, or holistic, approach to closing the gap.

Holistic method of closing gaps—While traditional loop building in Rosetta uses fragment insertions and backbone perturbations[35] confined to the area around the loop (or in this case the gap), holistic gap closure uses random torsion angle changes made to random residues throughout the protein. Changes to the torsion angles in a residue upstream from the gap propagate downstream, and *vice versa*, allowing for any change to slowly decrease the distance between the residues on either side of the gap. We found that this process can sometimes require many hours to close the gap entirely. To overcome this problem of achieving loop closure, we limit the number of cycles of holistic closure to initiate loop closure (cycles = $n_{res}/10$ where n_{res}, the number of residues, is 290 in ricin; for more details see *Experimental Procedures*). Then closure is completed by applying the cyclic coordinate descent method (CCD), which cycles through each backbone degree of freedom on the ten residues surrounding the gap to minimize the gap.[36] The combined approach is efficient and samples global structural perturbations.

Thread-and-close for secondary sequence space—CCD was not developed for use in secondary structural elements, and when applied to them the helical elements collapse and the strand elements create side chain clashes. Both result in unrealistic backbone torsion angles in disallowed regions of the Ramachandran space. For example, deletion of a residue in the middle of a helix as seen in Figure 3 results in a kinked helix that alters the usual *i* to *i* +4 hydrogen bonding pattern into an unrealistic *i* to *i*+3 pattern that is rarely found in the middle of a helix. To prevent these types of irregularities, we sought a way to realistically repair deletions that occur in elements of secondary structure.

An approach which avoids the collapse of secondary structure is to maintain the original secondary structure by moving the gap into a loop region. To move the gap, the sequence is threaded from either end of the secondary structural element (picked randomly, with a higher probability of starting from the closest end) so that the neighboring residue takes on the backbone position of the deleted residue and so forth until all residues in the element have been moved and the gap is shifted to the middle of the adjoining loop. Figure 4 shows

an example where the deletion was in a helix and the gap is moved four residues into the loop downstream from the deletion, far enough to ensure that the gap is no longer within a secondary structural element. Once the sequence has been threaded, the gap created in the loop is closed using the holistic loop closure approach as described above. We call this method a “thread-and-close” move.

Structure diversification, relaxation, and refinement

After repairing the gap created by deleting a residue, the structure is still close enough to the wild-type structure that it could be trapped in a local energy minimum. Unlike examining a deletion using *ab initio* folding methods, memory of the starting structure is retained and additional sampling is needed to escape the energy barriers. To increase sampling of the conformational space, we used three methods of structure diversification while in low-resolution mode. The simplest method is a forced expansion of the structure by applying random torsion angle changes throughout the protein. The second is a simulated annealing[37] Monte Carlo relaxation procedure where the temperature is increased during Monte-Carlo-plus-minimization cycles to escape local minima. The two approaches follow published protocols and are detailed in *Experimental Procedures.* The third is a new move which begins with a local perturbation and couples it to a final global perturbation, called Expanding Window Relax.

Expanding Window Relax—Similar to simulated annealing, Expanding Window Relax is used in low-resolution mode to overcome barriers between local energy minima. Backbone relaxation consists of a series of backbone torsion angle moves followed by energy minimization and a Boltzmann acceptance or rejection of the move. The key to the Expanding Window Relax procedure is calculating the set of residue positions that are allowed to move.

Expanding Window Relax starts by fixing the movement of all residues except those near the point of deletion. Five residues on either side of the deletion are allowed to move while all other residues remain fixed, and torsion angle motions are propagated towards a cut point at the middle of the loop. After a fixed number of cycles of relaxation using this window of movable residues, the window size is increased on both sides so that more residues are included during relaxation. This process continues until all residues in the protein have been relaxed. The number of relaxation cycles decreases as the number of residues to be relaxed increases. In this manner, the effects of the deletion can be resolved locally if possible, but globally if necessary.

All-atom relaxation and refinement

The goal of relaxation in high-resolution is to move the backbone to the lowest local energy structure for the desired sequence. The computational cost for relaxation of the protein backbone and side chains is dependent on the number of degrees of freedom that are used during minimization. Current algorithms for complete relaxation of the protein backbone have been limited to proteins of 100–150 residues and are known to be slow for large proteins.[10,38,39] As efforts to minimize large proteins become more commonplace, the dependence on number of residues inhibits the ability to produce results in a reasonable amount of time.

To determine how current relaxation algorithms scale with the number of residues, we ran Rosetta’s standard relaxation algorithm[13] ten times on truncations of the ricin protein with lengths 50, 100, 150, 200, 250, and 290 (full size). At each length, the final energy and time were recorded and averaged over the ten runs. Figure 5A shows the time for the ten points as a function of the length of the protein. As the number of residues that need to be relaxed

increases, the time increases dramatically, showing two regions ($n_{res} < 100$, $n_{res} > 100$) with power-law scaling in the upper region. Above 100 residues, the time scales as $n_{res}^{3.6}$. Further empirical analysis of the time dependence of the relaxation protocol suggests that the quasi-Newton energy minimization step alone scales as $d^{2.5}$, where d is the total degrees of freedom in the call to the minimizer. In the standard relaxation algorithm, d is n_{res} times the degrees of freedom of each residue. Therefore, one route for improving the efficiency of relaxation is by narrowing the relevant degrees of freedom and optimizing how often the minimization procedure is called. However, since we are attempting a global energy minimization, all of the residues ultimately need to be included in the relaxation.

Cyclic Interval Relax—To speed up the relaxation procedure and refinement process we adjusted two parts of the relaxation algorithm. 1) We optimized the number of calls to the minimization procedure, which we found to be suboptimal in the existing Rosetta relaxation algorithm. 2) We reduced the number of degrees of freedom used in each minimization step of relaxation. Our solution to minimize the degrees of freedom during relaxation, while still maintaining a global energy minimization strategy, is to reduce the degrees of freedom during minimization to a third by only allowing the angles in every third residue to move in any single call to the quasi-Newton minimizer (see *Experimental Procedures* for details). To ensure that all residues are minimized equally, the algorithm rotates through sets of residues in each call to the minimizer. We call this process Cyclic Interval Relax. Figure 6 shows the selection of residues during the Cyclic Interval Relax protocol.

To compare the time and efficiency of relaxation using Cyclic Interval Relax with Rosetta's standard relaxation protocol, we ran Cyclic Interval Relax on the same subsets of the ricin protein and collected data for the time and final energy of ten runs. Figure 5A shows the times for each length of the protein for Cyclic Interval Relax. Similar to standard Rosetta relaxation, the plot reveals a power-law scaling for proteins above 100 residues so that the time of relaxation scales as $n_{res}^{2.4}$. For the full-size ricin protein, the time required for relaxation is reduced from ~1.6 hours with standard Rosetta relaxation to 4 minutes with Cyclic Interval Relax – a tenfold speedup due to the combined effects of reducing the degrees of freedom and optimizing the number minimization calls during refinement. To ensure that the energy minimization efficiency has not been compromised, Figure 5B shows the final energy of each point, comparing standard Rosetta relax to Cyclic Interval Relax. The Cyclic Interval Relax method achieves equivalent energies in significantly less time. Ensembles of predicted structures were also generated using both relaxation procedures, showing similar final rmsd and energy trends. Similar optimization of the standard Rosetta relax algorithm was recently reported by Tyka *et al.*[40]

Combining the pieces

Each step of the algorithm has a specific purpose towards achieving a structure that can be analyzed for activity (Figure 7). After deleting the residue, the gap is repaired by thread-and-close, holistic loop closure, and CCD methods to give a starting structure. The diversity created by holistic loop closure is augmented by forced expansion, simulated annealing, and Expanding Window Relax. After $n_{res}/10$ cycles of Expanding Window Relax, in which the window size increases each cycle, the side chains are converted to all-atom representations for high-resolution refinement. Finally, the structures are relaxed in high-resolution using Cyclic Interval Relax to find an energy-minimized final structure.

Calculations were performed on a test set of forty-five single residue deletions from the crystal structure of ricin as well as on a structure refinement test set containing five proteins, each with available x-ray crystal structure for both wild-type and point deletion mutants. The combined algorithm requires approximately 30 minutes of computational time per

simulated structure, or decoy on a single Intel 2.4 GHz CPU. The simulation was repeated to create multiple decoys until convergence is reached, where convergence is defined as less than a 0.1 unit fluctuation of the average energy of the ten lowest-energy decoys every time 100 new structures are output. On average, 800 decoys are created by the point of convergence for each deletion mutant, and the total computational time for a prediction is 500 CPU-hours, run on 50 Intel 2.4 GHz CPUs.

Testing the refinement algorithm on known deletion structures

Surprisingly, there are very few examples of proteins with solved structures for wild-type and point-deletion variants (Table 1). We isolated five cases in the PDB (see Experimental Procedures). In all five cases, there is less than a 2.0 Å global rmsd difference between the wild-type and point-deletion mutant structures. In three of the five cases for which the deleted residue is in a helix (122l, 1c7p, and 1eem), most of the structural perturbation in the deletion mutant is in the adjoining loop. For the one case where the deleted residue is at the end of a beta sheet (3ewc), there is a large motion of a loop near the active site due to the loss of contact between the deleted residue and the ligand. As this residue participates in the activity of the protein, its deletion results in loss of activity.[41] In the remaining case (1h8i), the deleted residue is in a surface loop of the protein. The difference between the wild-type and mutant structures is minimal (< 0.2 Å).

We tested the deletion refinement algorithm on the five known cases by starting with the wild-type structure and creating models of the deletion mutants. For the lowest-scoring structures, we calculated the root-mean squared deviation (rmsd) of the models to the wild-type and deletion mutant x-ray crystal structures, and, to assess the variation in the created structures, the standard deviation of the rmsd of the predicted states (Table 1). In all five cases, the algorithm predicts the structure of the deletion mutants to a global rmsd of less than 1.5 Å to the mutant structure. For reference, the x-ray crystal structures of wild-type and mutant sequences typically differ by 1.0 Å and an x-ray crystal structure that is refined using Rosetta typically is perturbed by about 0.5 Å.[10]

Figure 8A shows plots of energy as a function of rmsd to the deletion mutant structure over all C_α atoms in the protein for each of the five proteins in the structure refinement test set. Convergence to a native-like structure would be indicated by the presence of a folding energy funnel, where the lowest-energy structures are those with rmsd values near 0 Å. The plots for four out of the five cases show deep funnels with very little variation in the rmsd of the structures produced. The case of 122l shows a wider distribution of rmsd values, which is reflected in the standard deviation measurement in Table 1. In all five cases, the predicted structure deviates from both the wild-type and mutant structures by ~1.0 Å.

Structure prediction of ricin deletion mutants

Using the same algorithm that was used to determine the structures of proteins with known mutant structures, we predicted the structure of each of the deletion mutants in ricin. Figure 8B shows plots of the energy as a function of the rmsd from C_α atoms in the wild-type structure for the set of decoys created for predictions of three different deletion mutants. (Plots for the remaining mutants can be found in Supp. Fig. 1). Each example shows a different range of resulting rmsd values. For mutant Δ123N (helix), in which residue N123 is deleted, all refined structures are found within 1–2 Å rmsd from the original crystal structure, similar to most of the plots for the known test set proteins. For Δ177S (helix), structures vary from 1 Å to over 5 Å rmsd, and the lowest-energy structure is also the lowest rmsd (1 Å) structure, similar to test case for 122l. For Δ181R (helix), structures vary from 2–10 Å rmsd and the lowest energy structures are 2–3 Å from wild-type. Plots that look similar to that of Δ181 are very different from those encountered while modeling structures

the test case structures above. We hypothesize that such plots with a large deviation in rmsd values may be an indication of the resulting functionality after deletion.

From structure prediction to enzyme activity

To our knowledge, there are no published methods to determine the level of activity of deletion mutants from sequence or structure. For comparison to our structure prediction based methods, a simple determinant of the effect of a deletion on activity could be the distance of a deleted residue from the active site. This distance can be measured on the wild-type structure and does not require any structure predictions. A similar measure is the distance in protein sequence between the deletion site and the closest active site residue, which would only require knowledge of the active site residues and the protein's sequence.

Our hypothesis is that the predicted structures of the mutant proteins could provide estimates of the mutant protein activity. Using the ensemble of predicted structures, we tested several new measures as proxies for enzymatic activity. Structural distortion of the deletion mutant relative to wild-type, as measured by the overall rmsd, could affect activity through disrupted global structure. A large change in energy of the mutant relative to wild-type could affect activity via stability of the overall fold of the protein or of the interface between the protein and the ligand. Distortion of the active site, as measured by a local rmsd, could affect activity by disrupting the catalytic geometry or altering the binding ability of the product or substrate. A low number of fraction of wild-type contacts could affect activity by creating a smaller active pocket into which the ligand can no longer enter and interact with the active site residues. Each of these measures was made on the predicted structures of all 43 deletion mutants. Plots of energy vs. active site rmsd can be found in Supp. Figs. 1 and 2.

A summary of various measurements on the predicted structures of each deletion mutant is shown in Table 2. Experimentally inactive deletion mutants are spread throughout the protein. The stretch of residues between ΔE146 (helix) and ΔI173 (helix) result in prediction of very small active site rmsd, while the stretch between ΔM175 (helix) and ΔY184 (helix) results in higher active site rmsd values. In both ranges the predicted total rmsd values are significantly greater than 1 Å. The disparity between the active site rmsd and total rmsd is likely correlated with the close proximity in sequence to the active site residue (residue 180 is involved in the depurination of the ribosome). Alternatively, ΔG122 (loop) is an example where the deletion results in a drastic predicted effect on the active site rmsd, despite being experimentally active.

Figure 9 shows histograms for measured proxies of activity (wild-type residue-residue contacts, active site rmsd, interface energy, and total rmsd) for all 45 deletion mutants, with the binned frequencies for active and inactive mutants are plotted as a function of each measure. Additionally, the standard deviations of energy, rmsd, and active site rmsd are also shown to represent the funnel height (energy) and width (rmsd and active site rmsd) for the complete ensemble (sequence distance plot not shown). Some measures (e.g. interface energy) show similar distributions for active and inactive mutants, and others (e.g. active site rmsd) show distinct differences.

After choosing an activity threshold, the predictive accuracy of the algorithm can be quantified for each measure of activity by calculating the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Sensitivity, P(act_thresh | active), is the probability of a prediction being within the threshold of activity given an experimentally active mutation. For example, $P(\text{rmsd}_{\text{active}} < 1.0$ Å | active) is the probability that an active mutation will be predicted to have an active site rmsd under 1.0 Å. Conversely, specificity, P(inact_thresh | inactive), is the probability of a prediction being within the threshold of inactivity for an experimentally inactive mutation. Both probabilities

are indicators of the quality of the results; however, they do not indicate the ability of a measure to predict activity blindly. Alternatively, PPV, *P*(active | act_thresh) and NPV, *P*(inactive | inact_thresh), can be used to assess the *predictive* quality of each measure of activity.

Table 3 summarizes the sensitivity, specificity, PPV, and NPV for each of the measurements. Measures of distance in sequence or structure have sensitivity values of 70%, specificity values of 69% and 67%, PPV values of 22% and 23%, and NPV values of 79% and 80% (sequence and structure, respectively). Thus, predictions that a site is inactive are often correct, but predictions that a site is active are often incorrect. For the active site rmsd, two thresholds were used and results in each case are reported. The most predictive measures are active site rmsd (with a threshold value of 1.0 Å), rmsd, and standard deviation of active site rmsd. Despite the high sensitivity of each measure (90%, 70%, and 90% respectively) the PPV for each is low (29%, 33%, and 35% respectively). The NPV values are very high (93%, 88%, and 95%) indicating a high confidence in the ability to predict inactive mutants. All of the measures outperform the sequence and structure distance measures. Wild-type contacts and all energy calculations were less predictive than the rmsd calculations. The energy difference from the wild-type, structure which might be expected to be a strong indicator of activity because of its relation to the protein's stability, showed PPV and NPV values of 24% and 82%, which are barely stronger the values calculated from the distance from the active site.

The low PPV values can be partially attributed to the low prevalence (27%) of active mutations in the point-deletion dataset. For example, for a data set with a prevalence rate of 10%, even if the sensitivity and specificity are over 95%, the PPV can be as low as 68%. The point deletion data set is a subset of the ricin residues and thus is not representative of the overall prevalence of active and inactive deletion sites in the whole protein. As mentioned in the Introduction, in addition to the 45 residues that were deleted individually by Munishkin and Wool,[6] 222 residues could be deleted in large sections resulting in mutant proteins that retain activity. Munishkin and Wool assumed that if the protein is able to tolerate the deletion of a large segment, then individual deletion of each residue in that segment will result in active mutants. Although this assumption may not always stand, it does seem likely that many of the 222 point deletions will remain active and this would alter the prevalence of active and inactive cases.

Therefore to recalculate PPV and NPV values, we followed Munishkin and Wool's assumption by estimating that 80% of the 222 residues that were not individually measured will retain activity when deleted individually, and 20% will result in a loss of activity. That is, we assume a prevalence of 80% active mutants from the unmeasured sites to complement the 27% prevalence of active mutants in the individually measured sites. Assuming that the algorithm for predicting deletion mutant activities will result in similar values for sensitivity and specificity, we determined whole-protein estimates for the statistical measures (Table 4; Experimental Procedures). The increased prevalence for activity alters the PPV and NPV of active site rmsd from 29% to 77% and from 93% to 63%, respectively. If indeed these assumptions hold, then active site rmsd of the predicted mutant structures will indicate the enzyme activity correctly over 60% of the time.

Discussion and Conclusions

Amino acid substitutions result in changes in hydrophobicity, hydrophilicity, charge, and bulk. The consequences of a given mutation depend on the nature of the amino acid that is substituted and the environment in which it occurs. With deletions, the nature of the mutation is more complicated since the surrounding residues are all affected as the protein

backbone must shift to regain connectivity. Munishkin and Wool[6] were able to show that ricin is able to tolerate a wide array of deletions throughout the protein structure and still retain activity. Deletion of one or more amino acids was tolerated in all eight α-helices, all six β-strands, and all of the connecting loops. The work of Munishkin and Wool provides a dramatic illustration of the degree to which proteins can tolerate small deletions (typically one to five amino acids), often involving residues in the hydrophobic core, and yet still be able to assemble an active site and generate measurable catalytic activity.

In this work we present the first attempt to predict the structure of deletion mutants and then the activity of the mutants from the predicted structure. We demonstrated that, on a test set of proteins with available point-deletion mutant x-ray crystal structures, the algorithm is able to predict the structure of a deletion mutant to within 1.5 Å rmsd of the mutant structure, starting from the wild-type structure. From predicted structure of ricin deletion mutations, we were able to show that it is possible to predict, albeit with moderate success, the activity or inactivity of deletions in proteins using an algorithm that couples structural predictions to measures of activity.

Ideally, predictions for each of the deletion mutants would be confirmed by comparing the predicted structures with the corresponding x-ray crystal structures mutant. Since there are no available structures of deletion mutants from which to infer functional properties for ricin, we created a two-step process to predict activity after demonstrating that the structure prediction algorithm can predict the structures of deletion mutants with sufficient accuracy. The first step is to predict structures of the mutant proteins and the second step is to infer activity from the resulting structural information. The two steps are intimately coupled since the structural information is necessary to predict activity and the correlation of these properties is needed to test the accuracy of the structures. The uncertainty of the predicted structures makes it difficult to determine whether negative results arise from inaccuracies in the predicted structures or from the method of inferring activity. Nevertheless, our algorithm shows that structure prediction followed by examination of the active site rmsd and funnel width can improve predictions over simple structure or sequence measures such as the distance between the deletion site and the active site, both in structure and sequence space.

Beyond the ability to predict activity of mutant proteins, a goal of structural biology is to gain understanding about the mechanism by which structure affects function. This work suggests that deletions far from the active site can lead to inactivity through movement of the active site residues, such as in the cases of ΔG141, ΔA200, and ΔS204. While the method may not be able to explain exactly how each deletion leads to the resulting activity, the results from the 45 predicted structures confirm Munishkin and Wool's hypothesis that it is not necessary for the deleted residue to be close to the active site in order to negatively affect activity. Furthermore, there are many reasons why a deletion mutant could be inactive. For example, the protein could become less stable or have residues near the binding site that move to prevent the protein from binding to the ligand or to distort the necessary catalytic geometry. While binding site distortion may not be evident from the overall rmsd, examining the active site rmsd would indicate that relevant resides have moved out of their correct positions.

We examined a number of measures of activity, of which half were shown to be predictive. The two measures of energy – average ensemble energy and average ensemble interface energy – were not found to be predictive measures, probably because the energy calculations are not accurate enough at such high resolution. Alternatively, the geometry of the protein may be more important than the energy. The standard deviation of rmsd and active site rmsd are indications of the shape of the global or local folding energy funnel. The width of the funnel may reflect the mutant protein's flexibility, *i.e.* mutants with a wide funnel are more

likely to sample alternative energy states more frequently. When too many states are available for the protein, it may be less likely to bind the ligand.

Other reasons for inactivity may not be measurable directly from the predicted structure. Aggregation, dimerization, and altered chemical potential are difficult to infer from the structure of a single protein. Inactivation through the destablization of a protein could be indicated by a change in calculated energy,[21,24–29] but the reference energy of the deleted residue would need to be included in the energy calculation to achieve an accurate $\Delta\Delta G$.[21]

Simply examining the residue and atom-level changes in the structure may not be sufficiently precise to consistently predict enzymatic activity. A more accurate indicator of activity might result from performing combined quantum mechanics/molecular mechanics (QM/MM) calculations on the active site to simulate the chemical reaction,[42–45] at considerably increased computational cost.

Methodologically, the creation of faster and more efficient algorithms for relaxing large proteins is a significant improvement in the road towards being able to determine function from structure. Typical structure prediction algorithms target proteins of 150 residues or less,[10] however the average length of enzymatic proteins is 300 residues.[11] The refinement procedure showed a speedup for proteins in the 300-residue range by optimizing the relax algorithm. The ability to relax large proteins efficiently is useful in applications beyond the prediction of protein function.

While high accuracy predictions of the level of activity of deletion mutants are not yet possible, the predicted structures may themselves be useful. Point deletions are important in many diseases and structures may help identify the structural mechanism of the disease. Similarly, homology modeling often includes proteins with small sequence changes due to deletions of amino acids. The deletion algorithm can be used to model these proteins by providing methods to fix the holes that are left behind after threading the desired sequence. The approach—predicting protein structure followed by analysis to estimate activity—is potentially extendable to other types of structural perturbations including insertions, deletions of segments, or even post-translational modifications.

In this work, we set out to test whether algorithms can be created to predict the relative activity of deletion mutants. We have shown that it is possible to predict mutations resulting in loss of activity 93% of the time using active site rmsd. Estimates incorporating likely global prevalence of active and inactive deletion sites suggest that activity can be predicted correctly from predicted structures over 60% of the time. The predictions are stronger than simple sequence organization measures (such as distance to active site), but they emphasize that fundamental work is required in structure prediction and enzyme activity determination to allow consistent prediction of activity. Additional and more complete experimental datasets of point deletions combined with structural and computational analysis will help to further elucidate the protein sequence-structure-function relationship. Improved understanding of determinants of enzyme activity will improve enzyme design algorithms and targeted directed evolution.

Experimental Procedures

Test sets

Structure prediction test set—To find the known wild-type and deletion-mutant structure pairs, we first performed a BLAST of all non-redundant FASTA sequences of all structures in the PDB against all structures in the PDB. The results were filtered to find all structures where BLAST reported no mutations and a single gap and the length of the query

sequence was one residue longer than the length of the subject sequence, indicating a point deletion (32 structures). Any structures with more than three chains were removed (25 structures). We found that sixteen of these structures were not of the full sequence and therefore the deleted residue was not included in the x-ray crystal structure; these were also removed (9 structures). Two of the structures contained mutations in place of the deletions that were not caught by BLAST (leaving 7 structures). Two of the sequences presented technical problems (chain gaps and PDB format issues) and so we used the remaining five structures for the structure prediction set. The structures in the test set were prepared following the steps detailed in the *Initial Conditions* section below, although without any active site rmsd calculations.

Activity prediction test set—The set for predicting activities was taken from Munishkin and Wool.[6]

Initial conditions

All deletions start from the native structure for the A chain of ricin (pdb 1rtc) with the ligand adenyl(3'→5')guanosine (ApG, a transition state analog) placed in the active site (pdb 1apg).[46] Prior to employing the deletion algorithm, the crystal structure was preprocessed as follows. First, the backbone bond lengths and angles were set to ideal values.[32] These small changes resulted in minor clashes between the side chains and the ApG ligand, therefore the ligand rigid-body position was refined by quasi-Newton minimization while allowing small torsion angle changes throughout the backbone of the protein. For the refinement, the surrounding side chain positions were kept fixed, otherwise the side chains required for binding and catalysis would move to allow space for the ligand. The lowest energy idealized and refined ligand structure was then used as the starting structure for all subsequent deletions (there was very little variation, <1%, in rmsd and energy for all the structures created).

Energy function

Rosetta's multi-scale algorithm is based on two energy functions. At low resolution, a fast energy function is used that accounts for the backbone heavy atoms and a pseudo-atom representing the centroid of the side chain atoms. The energy functions, developed for and tested on folding,[33] loop building,[35] and docking problems,[47] include van der Waals clashes, ligand-protein contacts, knowledge-based residue environment and residue-residue pair propensities, a loop-closure measure, as well as a Ramachandran energy.[33,47,48] The deletion algorithm uses the weights from protein folding[15] with an additional chain break term (weight 1.0) from loop building to ensure that the gap created by deleting a residue is negligible.

At high resolution, Rosetta uses an all-atom potential to capture atomic scale physical forces. For the relaxation, the potential includes van der Waals interactions,[47] implicit solvation,[49] orientation-dependent hydrogen bonding,[50–52] and a rotamer probability to capture side chain internal energies.[53,54] Parameters and weights have been published previously.[10] Similar to low resolution, the weights from protein folding are used with the addition of the chain break energy score. During relaxation, the weights of the chain break energy score, Lennard-Jones repulsion, and Ramachandran probability are adjusted as indicated in the algorithm details below. The final structure is scored using published weights plus the chain break energy score (weight 1.0).

Details of low-resolution search

The low-resolution search algorithm starts by deleting the desired residue and, if the deletion is in a secondary structural element, threading the gap to the nearest loop (see *Thread-and-*

close for secondary structure space). Small changes throughout the backbone of the protein are then used in the holistic loop closure approach for five cycles or until the loop is closed and no gap between the two ends is present. Each single cycle of holistic loop closure includes 100 small moves (random perturbations to the ϕ and ψ angles of a random residue) and 100 shear moves (random perturbations to the ψ angle of a random residue and a corresponding opposite perturbation to the ϕ angle of the following residue,[33] alternating between small and shear moves for $n_{res}/10$ cycles. After each series of many small and shear moves, the chain break distance is calculated. If within five cycles the loop has not closed, the algorithm employs a modified CCD[36,55] loop closure, ensuring that the loop closes completely. During holistic loop closure, the Ramachandran weight factor is increased to 1.0 (versus the 0.1 normally used) to force valid sampling of ϕ/ψ space during loop building. This increased weight forces the CCD loop closure to remain in the allowed space.

Once there is no longer a gap in the structure of the protein, a combination of forced expansion to within 0.5 Å of the closed protein, simulated annealing, and Expanding Window Relaxation are used to sample the conformational space of the protein. First, ($n_{res}/5$) cycles of simulated annealing are used to force the structure out of the native conformation and next ($n_{res}/10$) cycles of Expanding Window Relax are used to focus the diversity around the area of deletion. Each relax cycle is composed of 100 small moves and 100 shear moves. After each move, an energy check is used to determine whether to accept or reject the moves based on Boltzmann criteria.

High-resolution search

The purpose of high-resolution refinement is to allow fine changes in structure to produce the most relevant structure for evaluating the energy for candidate structure discrimination. Refinement consists of several cycles, wherein each cycle consists of small and shear moves, minimization in the backbone and side chain space of each residue, and occasional repacking of the side chains. Perturbations at this stage are very small (<5°) to avoid clashes in the highly corrugated all-atom potential function. The periodic repacking is achieved using an embedded Monte Carlo simulated annealing routine to select the best combination of conformations from a discrete rotamer library.[53,54]

The high-resolution search first replaces the centroid atom with all-atom side chains at every residue position and packs them using a rotamer library [53] and quasi-Newton minimization over the χ angles.[56] Since replacing the centroids with all-atom side chains may create clashes, the weight of the van der Waals repulsion term in the energy function is reduced to 10%, thus allowing some overlap between the atoms to occur initially. This term is ramped geometrically back to the normal weight, avoiding initial clashes that result from the addition of the side chains. Next, random small and shear moves[33] are used to perturb the backbone throughout the protein. Finally, the structure undergoes gradient-based minimization following *Cyclic Interval Relax*.

Cyclic Interval Relax uses a map of angles that are allowed to move to define the movable degrees of freedom during minimization. For the first cycle of relaxation, torsion angles ϕ, ψ, and χ for the set of residues with residue numbers $3i+k$ are allowed to move during minimization, where $i \in \{0, n_{res}/3\}$ and k is fixed at 1. Each subsequent cycle increments $k \in \{1,2,3\}$. The choice of k_{max} of 3 showed the best improvement to the algorithm, providing a balance between efficiency and effectiveness. Small values of k_{max} are less efficient, while higher values do not lead to a low energy during minimization.

Statistical measures

The values for sensitivity, specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) were calculated based on the measured True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The formulas for sensitivity, specificity, PPV, and NPV are:

$$\begin{aligned}
&\text{sensitivity} = \frac{TP}{TP+FN} = \frac{N_{\text{act}}^{\text{corr}}}{N_{\text{act}}},\\
&\text{specificity} = \frac{TN}{TN+FP} = \frac{N_{\text{inact}}^{\text{corr}}}{N_{\text{inact}}},\\
&PPV = \frac{TP}{TP+FP} = \frac{N_{\text{act}}^{\text{corr}}}{N_{\text{act}}^{\text{corr}}+N_{\text{inact}}^{\text{inc}}}, \text{ and}\\
&NPV = \frac{TN}{TN+FN} = \frac{N_{\text{inact}}^{\text{corr}}}{N_{\text{inact}}^{\text{corr}}+N_{\text{act}}^{\text{inc}}},
\end{aligned}$$

where N_{act} and N_{inact} are the numbers of experimentally active and inactive mutations, respectively; $N_{\text{act}}^{\text{corr}}$ and $N_{\text{act}}^{\text{inc}}$ are the numbers of experimentally active mutations that were correctly and incorrectly predicted, respectively (TP and FN); and $N_{\text{inact}}^{\text{corr}}$ and $N_{\text{inact}}^{\text{inc}}$ are the numbers of experimentally inactive mutations that were correctly and incorrectly predicted, respectively (TN and FP).

Extrapolation of statistical measures throughout the whole protein

To estimate similar statistical measures over the whole protein, we need to assume a prevalence $\tilde{p}$ of active mutations in the set of mutants corresponding to the $\tilde{N}$ residue positions which were not experimentally tested. Then, the number of active and inactive mutants in that set would be $\tilde{N}_{\text{act}} = \tilde{p}\tilde{N}$ and $\tilde{N}_{\text{inact}} = (1-\tilde{p})\tilde{N}$. The number of correct predictions in the untested mutants is estimated by assuming a success rate the same as in the tested mutants, i.e.,

$\tilde{N}_{\text{act}}^{\text{corr}} = (\tilde{p}\cdot\tilde{N}_T)\frac{N_{\text{act}}^{\text{corr}}}{N_{\text{act}}}$, $\tilde{N}_{\text{act}}^{\text{inc}} = (\tilde{p}\cdot\tilde{N}_T)\frac{N_{\text{act}}^{\text{inc}}}{N_{\text{act}}}$, $\tilde{N}_{\text{inact}}^{\text{corr}} = (\tilde{p}\cdot\tilde{N}_T)\frac{N_{\text{inact}}^{\text{corr}}}{N_{\text{inact}}}$, and $\tilde{N}_{\text{inact}}^{\text{inc}} = (\tilde{p}\cdot\tilde{N}_T)\frac{N_{\text{inact}}^{\text{inc}}}{N_{\text{inact}}}$. The numbers of predictions in each category in the *whole* protein can now be found by summing, $\widehat{N} = N + \tilde{N}$, $\widehat{N}_{\text{act}} = N_{\text{act}} + \tilde{N}_{\text{act}}$, $\widehat{N}_{\text{act}}^{\text{corr}} = N_{\text{act}}^{\text{corr}} + \tilde{N}_{\text{act}}^{\text{corr}}$ etc. Finally, the sensitivity, specificity, PPV, and NPV for the whole protein can be calculated from their definitions in the section above using the whole protein counts. For example, the equation for $\widehat{PPV}$ is (where hats indicate values over the whole protein and tildes indicate values estimated over the untested residue positions):

$$\widehat{PPV} = \frac{\widehat{TP}}{\widehat{TP}+\widehat{FP}} = \frac{\widehat{N}_{\text{act}}^{\text{corr}}}{\widehat{N}_{\text{act}}^{\text{corr}}+\widehat{N}_{\text{inact}}^{\text{inc}}} = \frac{\tilde{N}_{\text{act}}^{\text{corr}}+N_{\text{act}}^{\text{corr}}}{\tilde{N}_{\text{act}}^{\text{corr}}+N_{\text{act}}^{\text{corr}}+\tilde{N}_{\text{inact}}^{\text{inc}}+N_{\text{inact}}^{\text{inc}}} = \frac{(\tilde{p}\cdot\tilde{N}_T)\frac{N_{\text{act}}^{\text{corr}}}{N_{\text{act}}}+N_{\text{act}}^{\text{corr}}}{(\tilde{p}\cdot\tilde{N}_T)\frac{N_{\text{act}}^{\text{corr}}}{N_{\text{act}}}+N_{\text{act}}^{\text{corr}}+(\tilde{p}\cdot\tilde{N}_T)\frac{N_{\text{inact}}^{\text{inc}}}{N_{\text{inact}}}+N_{\text{inact}}^{\text{inc}}}$$

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Shortle D, Sondek J. The emerging role of insertions and deletions in protein engineering. Curr Opin Biotechnol. 1995; 6(4):387–393. [PubMed: 7579648]
2. Kitamura K, Yoshida C, Nishigaki K. GFPs of insertion mutation generated by molecular size-altering block shuffling. FEBS Lett. 2003; 555(3):483–488. [PubMed: 14675760]

3. Hayes RJ, Bentzien J, Ary ML, Hwang MY, Jacinto JM, Vielmetter J, Kundu A, Dahiyat BI. Combining computational and experimental screening for rapid optimization of protein properties. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99(25): 15926–15931. [PubMed: 12446841]

4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1):235–242. [PubMed: 10592235]

5. Flores-Ramirez G, Rivera M, Morales-Pablos A, Osuna J, Soberon X, Gaytan P. The effect of amino acid deletions and substitutions in the longest loop of GFP. BMC Chem Biol. 2007; 7:1. [PubMed: 17594481]

6. Munishkin A, Wool IG. Systematic deletion analysis of ricin A-chain function. Single amino acid deletions. J Biol Chem. 1995; 270(51):30581–30587. [PubMed: 8530493]

7. Simm AM, Baldwin AJ, Busse K, Jones DD. Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 beta-lactamase. FEBS Lett. 2007; 581(21): 3904–3908. [PubMed: 17662719]

8. Osuna J, Yanez J, Soberon X, Gaytan P. Protein evolution by codon-based random deletions. Nucleic Acids Res. 2004; 32(17):e136. [PubMed: 15459282]

9. Lim WA, Farruggio DC, Sauer RT. Structural and energetic consequences of disruptive mutations in a protein core. Biochemistry. 1992; 31(17):4324–4333. [PubMed: 1567879]

10. Bradley P, Misura KMS, Baker D. Toward High-Resolution de Novo Structure Prediction for Small Proteins. Science. 2005; 309(5742):1868–1871. [PubMed: 16166519]

11. Brocchieri L, Karlin S. Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res. 2005; 33(10):3390–3400. [PubMed: 15951512]

12. Das R, Baker D. Prospects for de novo phasing with de novo protein models. Acta Crystallogr D Biol Crystallogr. 2009; 65(Pt 2):169–175. [PubMed: 19171972]

13. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmstrom L, Wollacott AM, Wang C, Andre I, Baker D. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins. 2007; 69 Suppl 8:118–128. [PubMed: 17894356]

14. 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(7167):259–264. [PubMed: 17934447]

15. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D. Free modeling with Rosetta in CASP6. Proteins. 2005; 61 Suppl 7:128–134. [PubMed: 16187354]

16. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. 2007; 5:17. [PubMed: 17488521]

17. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF 3rd, Hilvert D, Houk KN, Stoddard BL, Baker D. De novo computational design of retro-aldol enzymes. Science. 2008; 319(5868):1387–1391. [PubMed: 18323453]

18. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D. Kemp elimination catalysts by computational enzyme design. Nature. 2008; 453(7192):190–195. [PubMed: 18354394]

19. Alvizo O, Allen BD, Mayo SL. Computational protein design promises to revolutionize protein engineering. Biotechniques. 2007; 42(1):31. 33, 35 passim. [PubMed: 17269482]

20. Lassila JK, Privett HK, Allen BD, Mayo SL. Combinatorial methods for small-molecule placement in computational enzyme design. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(45):16710–16715. [PubMed: 17075051]

21. Kortemme T, Kim DE, Baker D. Computational alanine scanning of protein-protein interfaces. Sci STKE 2004. 2004; 219:l2.

22. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. Protein Eng. 1999; 12(7):549–555. [PubMed: 10436080]

23. 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(2):369–387. [PubMed: 12079393]
24. Humphris EL, Kortemme T. Design of multi-specificity in protein interfaces. PLoS Comput Biol. 2007; 3(8):e164. [PubMed: 17722975]
25. Wooll JO, Wrabl JO, Hilser VJ. Ensemble modulation as an origin of denaturant-independent hydrogen exchange in proteins. J Mol Biol. 2000; 301(2):247–256. [PubMed: 10926507]
26. Bueno M, Camacho CJ, Sancho J. SIMPLE estimate of the free energy change due to aliphatic mutations: superior predictions based on first principles. Proteins. 2007; 68(4):850–862. [PubMed: 17523191]
27. Parthiban V, Gromiha MM, Abhinandan M, Schomburg D. Computational modeling of protein mutant stability: analysis and optimization of statistical potentials and structural features reveal insights into prediction model development. BMC Struct Biol. 2007; 7:54. [PubMed: 17705837]
28. Parthiban V, Gromiha MM, Hoppe C, Schomburg D. Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. Proteins. 2007; 66(1):41–52. [PubMed: 17068801]
29. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res. 2006; 34(Web Server issue):W239–W242. [PubMed: 16845001]
30. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol. 2001; 307(2):683–706. [PubMed: 11254390]
31. Lau AY, Chasman DI. Functional classification of proteins and protein variants. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(17):6576–6581. [PubMed: 15087495]
32. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol. 1997; 268(1):209–225. [PubMed: 9149153]
33. Rohl, CA.; Strauss, CEM.; Misura, KMS.; Baker, D.; Ludwig Brand; Michael, LJ. Methods in Enzymology. Vol. Volume 383. Academic Press; 2004. Protein Structure Prediction Using Rosetta; p. 66-93.Volume
34. Moult J. Predicting protein three-dimensional structure. Curr Opin Biotechnol. 1999; 10(6):583–588. [PubMed: 10600698]
35. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins. 2004; 55(3):656–677. [PubMed: 15103629]
36. Canutescu AA, Dunbrack RL Jr. Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci. 2003; 12(5):963–972. [PubMed: 12717019]
37. Chou KC, Carlacci L. Simulated annealing approach to the study of protein structures. Protein Eng. 1991; 4(6):661–667. [PubMed: 1946323]
38. Kim DE, Blum B, Bradley P, Baker D. Sampling bottlenecks in de novo protein structure prediction. J Mol Biol. 2009; 393(1):249–260. [PubMed: 19646450]
39. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 2004; 32(Web Server issue):W526–W531. [PubMed: 15215442]
40. Tyka MD, Keedy DA, Andre I, Dimaio F, Song Y, Richardson DC, Richardson JS, Baker D. Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol. 405(2): 607–618. [PubMed: 21073878]
41. Ho MC, Cassera MB, Madrid DC, Ting LM, Tyler PC, Kim K, Almo SC, Schramm VL. Structural and metabolic specificity of methylthiocoformycin for malarial adenosine deaminases. Biochemistry. 2009; 48(40):9618–9626. [PubMed: 19728741]
42. Mulholland AJ. Modelling enzyme reaction mechanisms, specificity and catalysis. Drug Discov Today. 2005; 10(20):1393–1402. [PubMed: 16253878]
43. Senn HM, Thiel W. QM/MM studies of enzymes. Curr Opin Chem Biol. 2007; 11(2):182–187. [PubMed: 17307018]
44. Senn HM, Thiel W. QM/MM methods for biomolecular systems. Angew Chem Int Ed Engl. 2009; 48(7):1198–1229. [PubMed: 19173328]

45. van der Kamp MW, Mulholland AJ. Computational enzymology: insight into biological catalysts from modelling. Nat Prod Rep. 2008; 25(6):1001–1014. [PubMed: 19030602]

46. Monzingo AF, Robertus JD. X-ray analysis of substrate analogs in the ricin A-chain active site. J Mol Biol. 1992; 227(4):1136–1145. [PubMed: 1433290]

47. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of molecular biology. 2003; 331(1):281–299. [PubMed: 12875852]

48. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins. 1999; 34(1):82–95. [PubMed: 10336385]

49. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins. 1999; 35(2): 133–152. [PubMed: 10223287]

50. Morozov AV, Kortemme T, Tsemekhman K, Baker D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(18):6946–6951. [PubMed: 15118103]

51. Morozov AV, Kortemme T. Potential functions for hydrogen bonds in protein structure prediction and design. Advances in protein chemistry. 2005; 72:1–38. [PubMed: 16581371]

52. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. Journal of molecular biology. 2003; 326(4):1239–1259. [PubMed: 12589766]

53. Dunbrack RL Jr, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 1997; 6(8):1661–1681. [PubMed: 9260279]

54. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. PNAS. 2000; 97(19):10383–10388. [PubMed: 10984534]

55. Wang C, Bradley P, Baker D. Protein-protein docking with backbone flexibility. J Mol Biol. 2007; 373(2):503–519. [PubMed: 17825317]

56. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein-protein docking. Protein Sci. 2005; 14(5):1328–1339. [PubMed: 15802647]

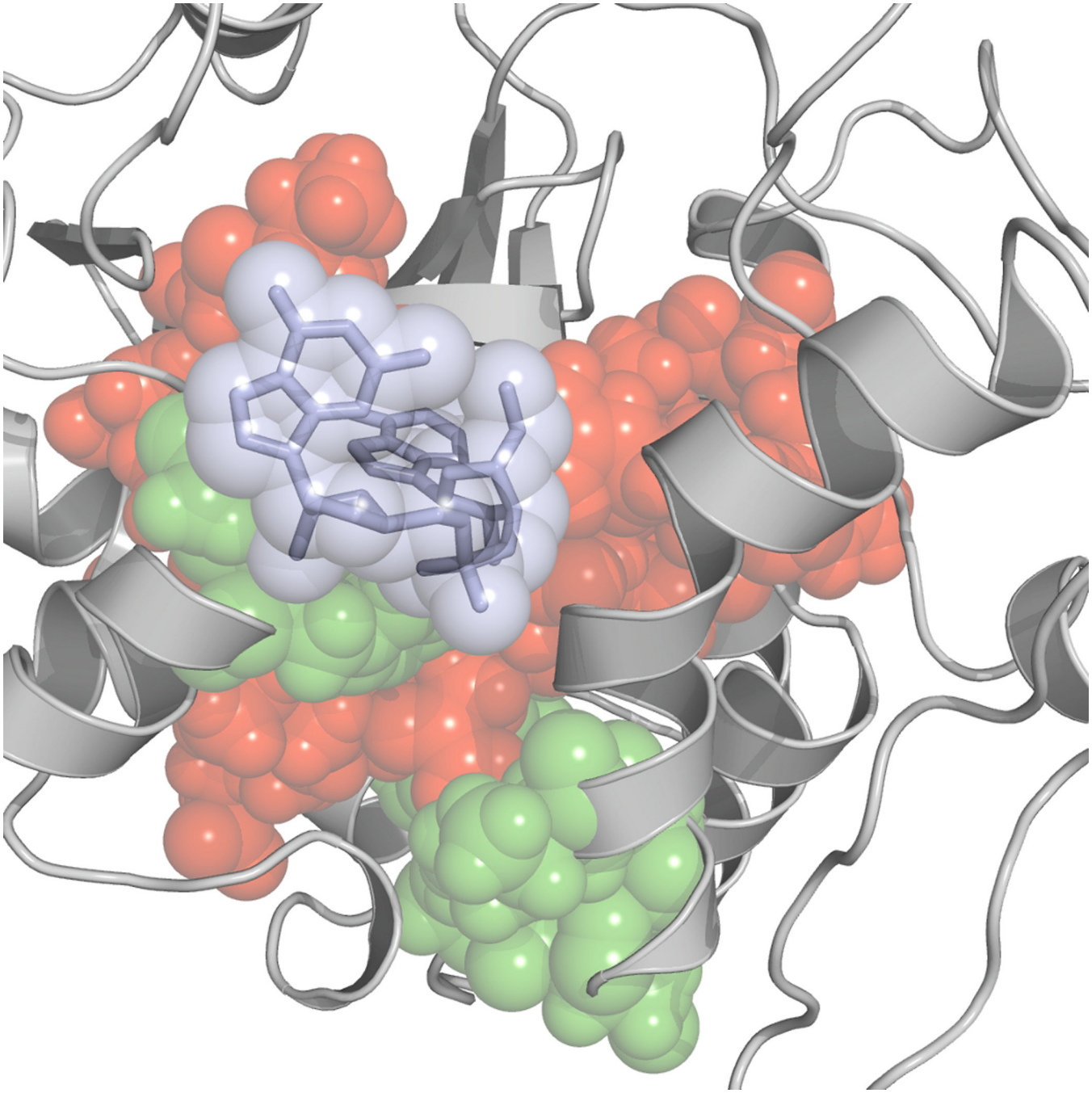

Fig. 1.

Cartoon rendering of the activity of ricin as measured by Munishkin and Wool [6]. Individually deleted residues that retain activity are shown in green spheres and those which cause loss of activity are shown in red spheres. The dinucleotide ligand (substrate analog) is shown in light blue sticks.

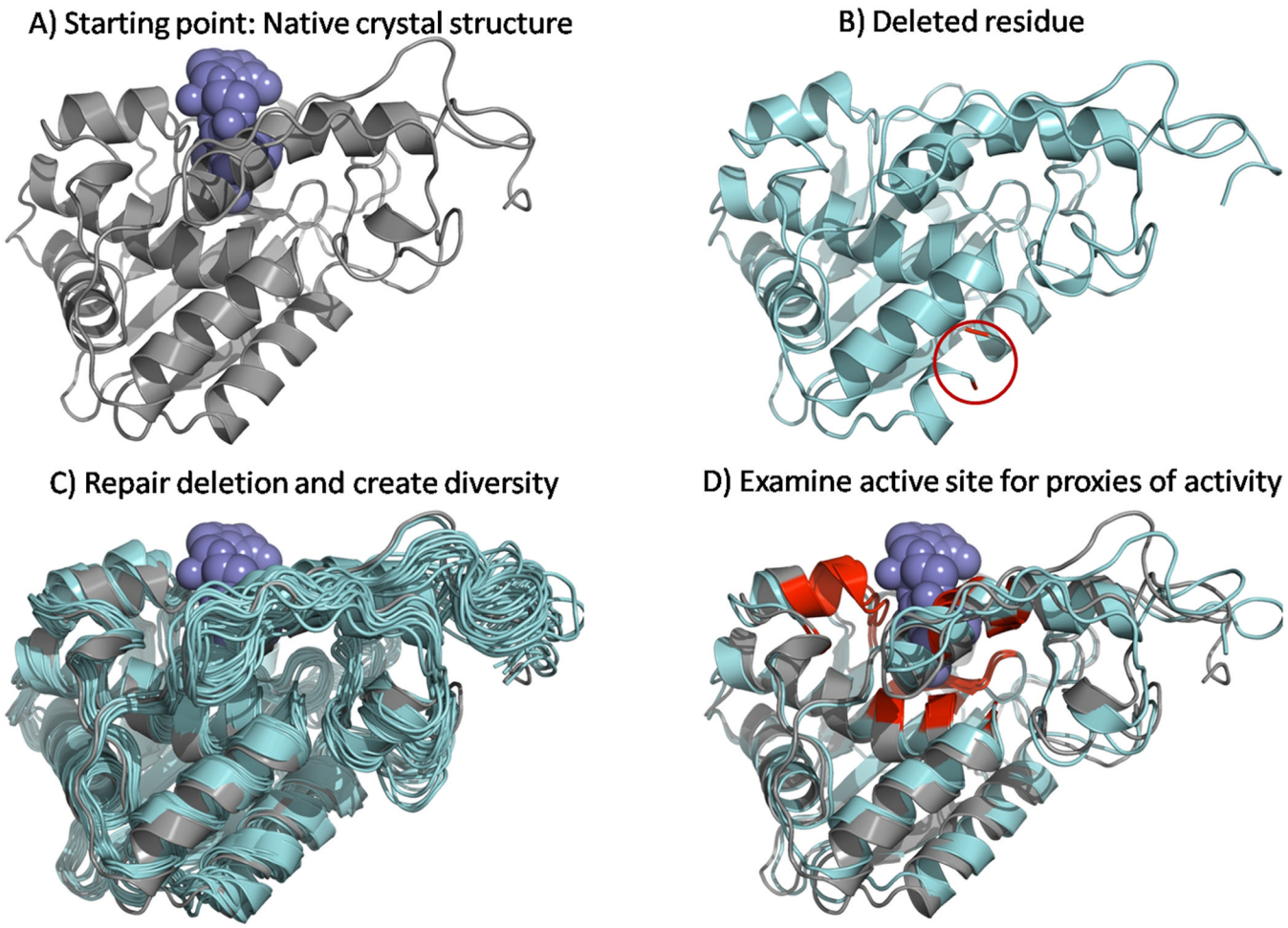

Fig. 2.

General algorithm for predicting mutant activity. (A) Start from native, wild-type structure. (B) Delete residue (Δ24 shown). (C) Repair deletion and create a diversity of structures. (D) Examine active site for proxies of activity (active site residues show in red).

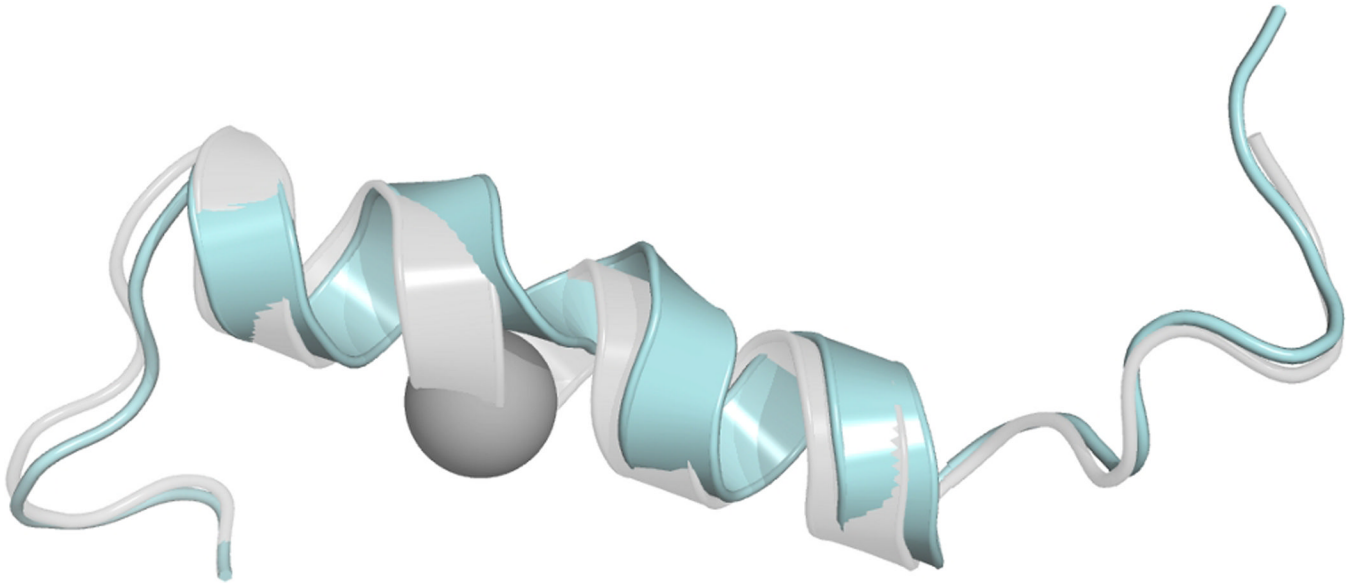

Fig. 3.

Cartoon rendering of a helix before and after deleting residue N24. Disallowed torsion angles resulting from holistic loop building and CCD collapse the helix (cyan). The native crystal structure is shown for reference (gray) with N24 indicated by a sphere.

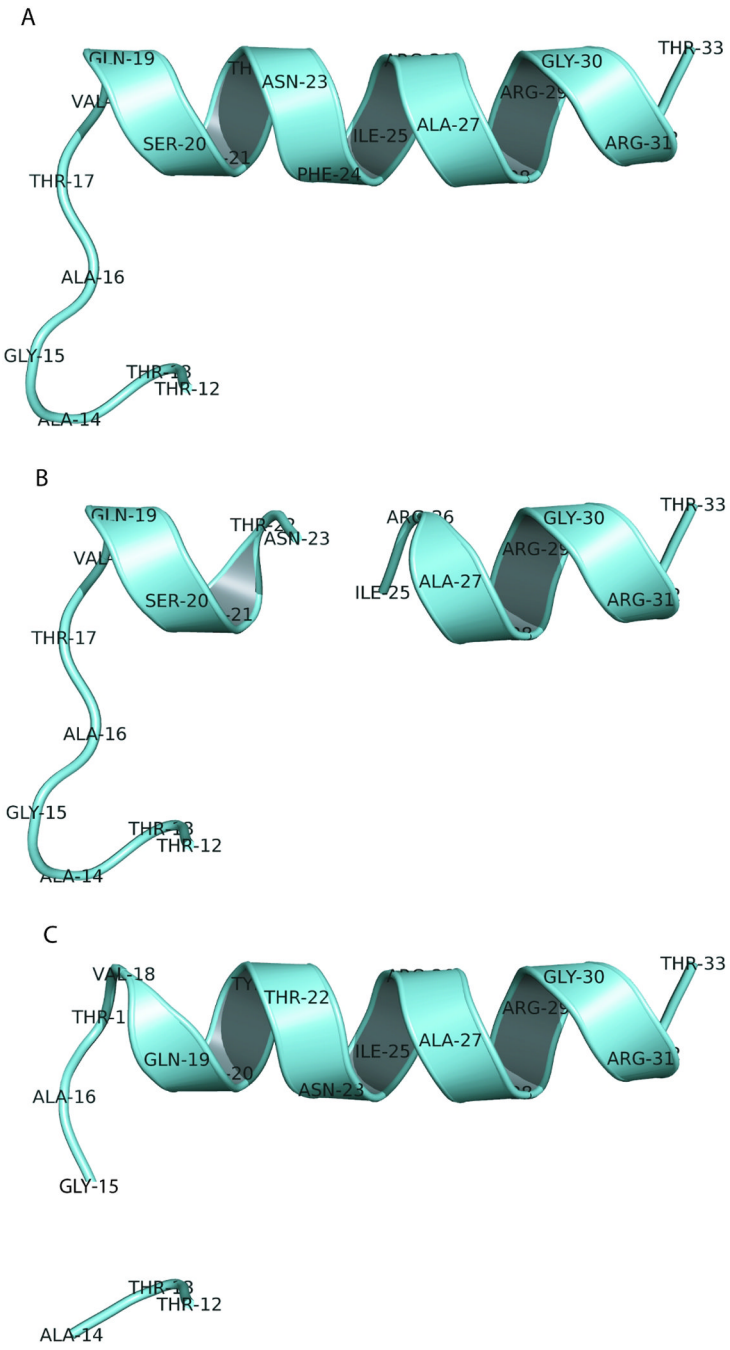

Fig. 4.

A thread-and-close move. (A) The wild-type helix before deleting residue 24, including the residues in the loop upstream. (B) After deleting residue N24 without threading. (C) Threading the gap into the middle of the loop. Notice that residue N24 is missing (see numbering) but the gap has moved to between residues T14 and A15.

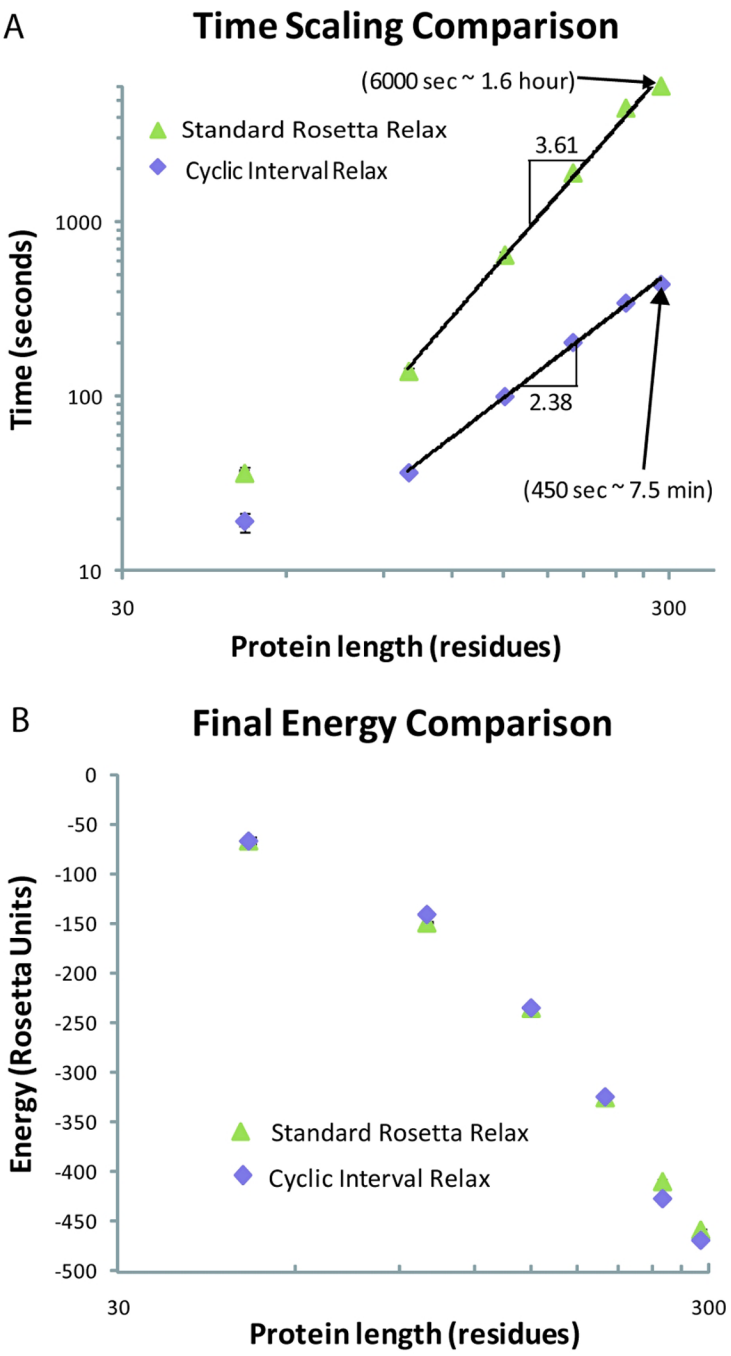

Fig. 5.

Plots comparing the efficiency of Standard Rosetta Relax and Cyclic Interval Relax. (A) Run time vs. the number of residues. (B) Average final energy vs. number of residues. Error bars shown for standard deviation of 10 relaxation runs, only seen in the first point of plot (A) because error bars are smaller than points.

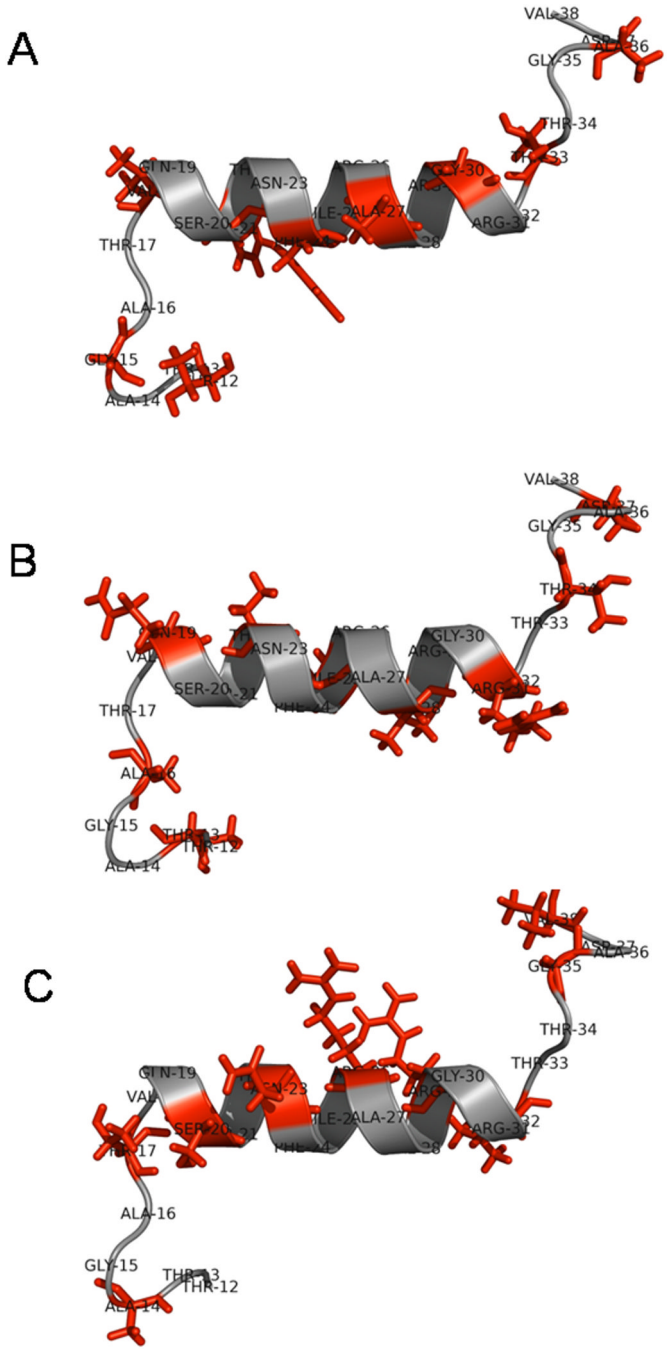

Fig. 6.

Cartoon representation of Cyclic Interval Relax. The ϕ and ψ angles of every third residues are allowed to move (red sticks) while all other residues are held fixed in torsion space (gray). (A) Minimization over residues {3*i*} (B) {3*i*+1} and (C) {3*i*+2}.

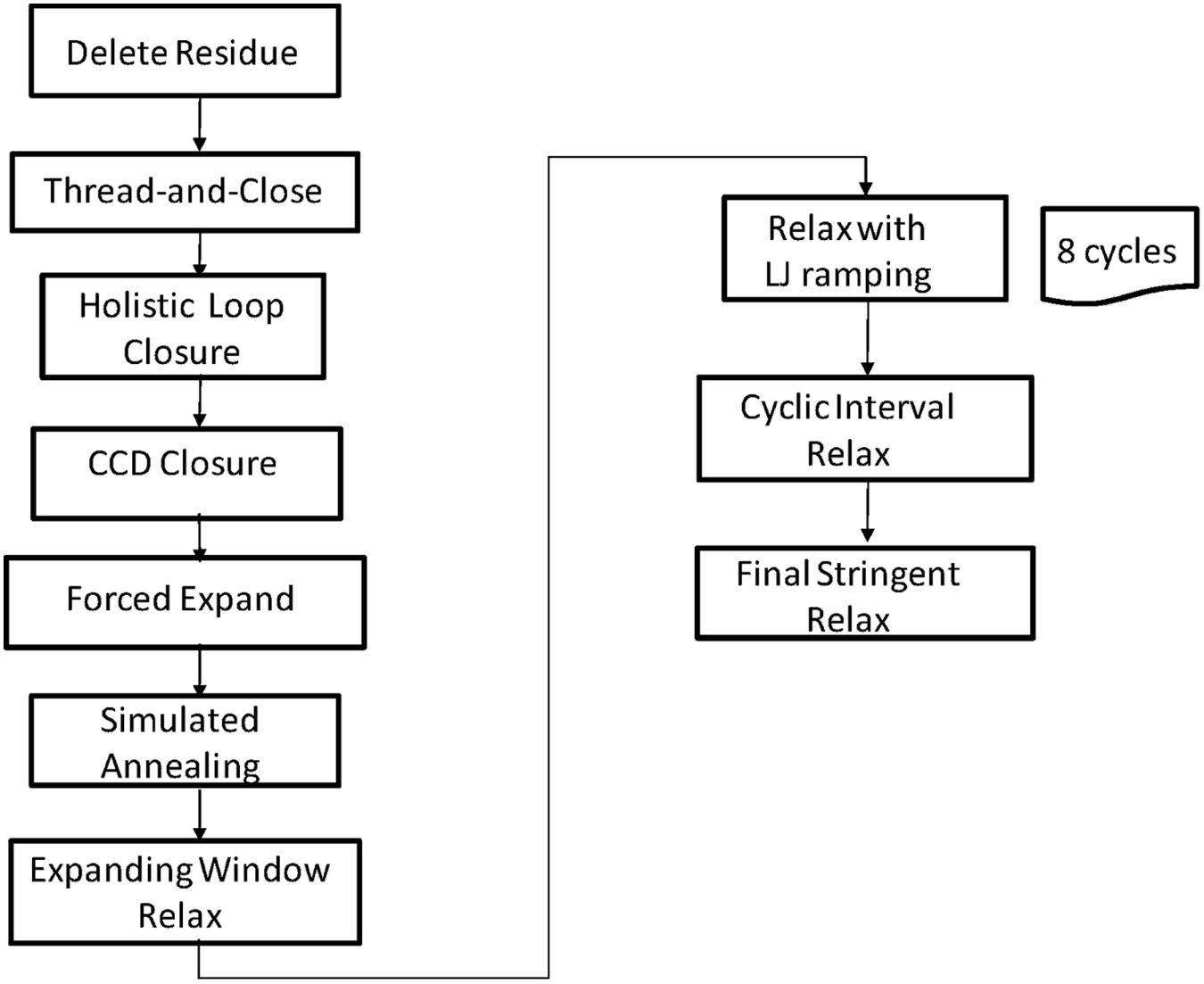

Fig. 7.
Flowchart for deletion algorithm in Rosetta. Left column, low-resolution steps; right column, high-resolution steps.

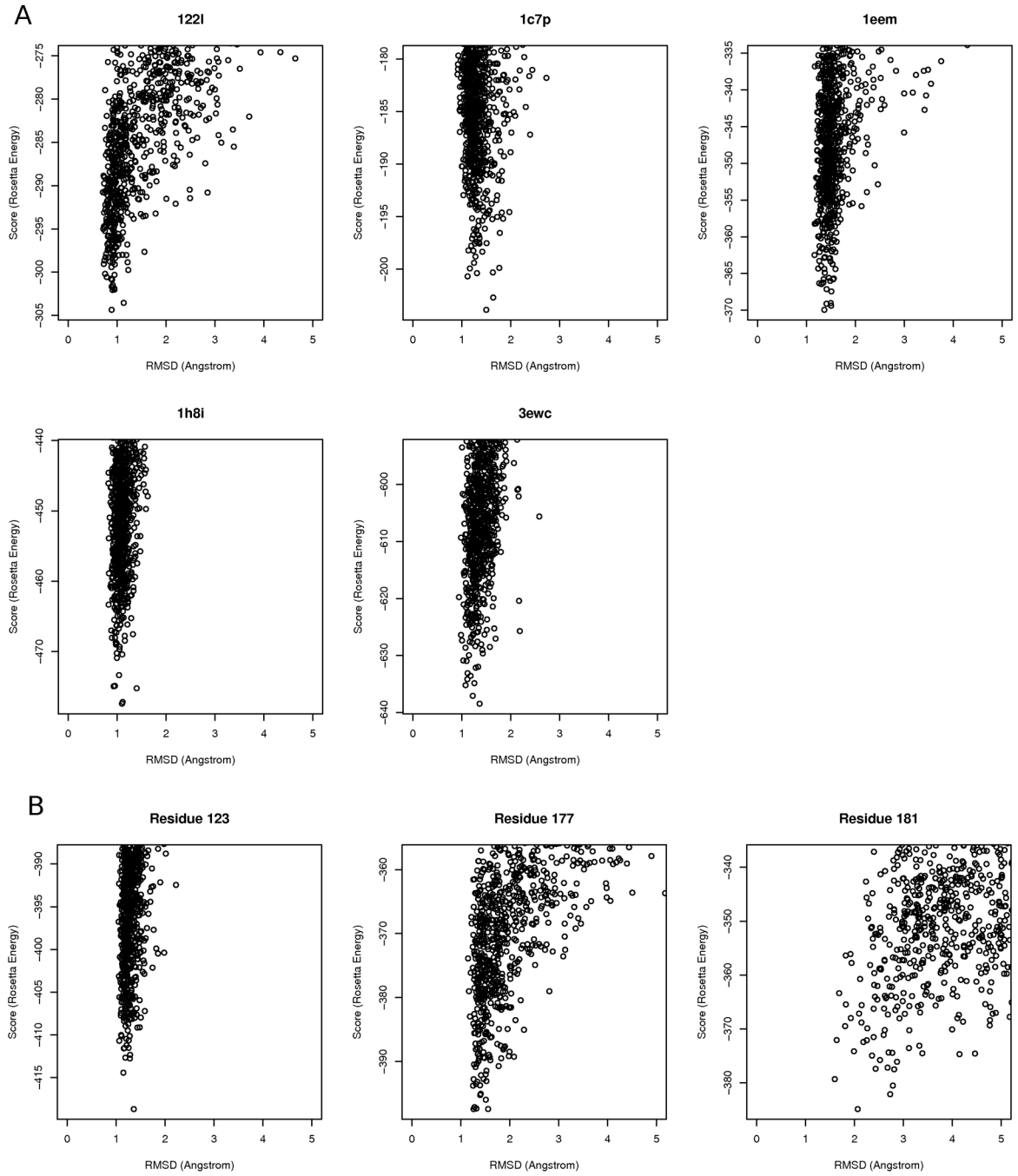

Fig. 8.

Plots of energy versus backbone atom rmsd for 1000 structure predicted decoys for **(A)** all proteins in structure prediction data set and **(B)** three different ricin deletion mutations: Δ123 (experimentally active), Δ177 (inactive), and Δ181 (inactive). All examples are numerically converged, but the shape of the resulting funnel varies.

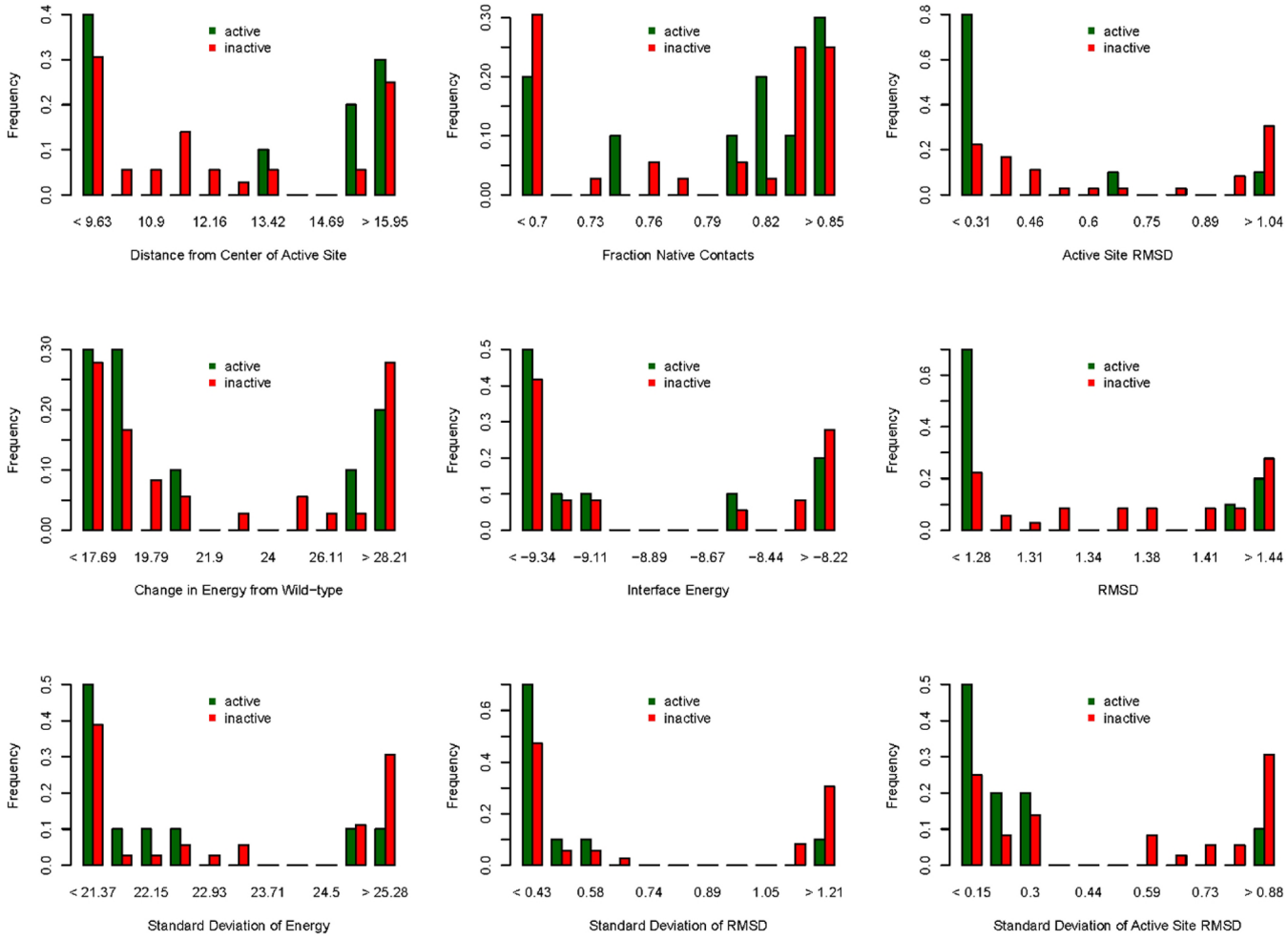

Fig. 9.

Histograms of frequencies of active/inactive mutants as a function of various proxies for activity. (A) Distance from active site center, (B) Fraction native contacts, (C) Active site rmsd, (D) Energy, (E) Interface energy, (F) Rmsd, (G) Standard deviation of energy, (H) Standard deviation of rmsd, (I) Standard deviation of active site rmsd.

Table 1

Deletion structure prediction test set and results

Protein, name of protein; WT PDB, PDB code for wild-type structure; Mutant PDB, PDB code for mutant structure; Mutant, residue number and type of deletion mutant; $RMSD^{ST}$, total rmsd over all atoms of the starting structure for mutant compared to wild-type; RMSDWT, total rmsd of lowest energy predicted structure to wild-type x-ray crystal structure; $RMSD^{MUT}$, total rmsd of lowest energy predicted structure to deletion-mutant x-ray crystal structure; Std Dev RMSD, standard deviation of the total rmsd for all predicted structures (represents funnel width).

Protein	WT PDB	Mutant PDB	Mutant	$RMSD^{ST}$ (Å)	$RMSD^{WT}$ (Å)	$RMSD^{MUT}$ (Å)	Std Dev RMSD
Lysozyme	122lA	210lA	ΔS73	1.37	0.67	0.73	0.74
Lysozyme	1c7pA	1di5A	ΔR101	1.36	1.40	1.50	0.27
Glutothione Transferase	1eemA	3lflA	ΔE155	1.62	1.21	1.36	0.61
Thrombin	1h8iH	1h8dH	ΔW148	0.79	0.76	0.73	0.49
Adenosine Deaminase	3ewcA	3ewdA	ΔD172	1.27	0.99	1.02	0.55

Table 2

Deletion mutant activities and measures from predicted structures

Mutant, residue number and type of deletion mutant; SS, secondary structure type of deleted residue (H-helix, S-sheet, L-loop); Dist, distance of the deleted residue from the active site (Å); Exp Act, experimental activity as measured by Munishkin and Wool; ΔE_{top1}, change in energy from wild-type of lowest energy structure; ΔE_{top10}, average change in energy from wild-type of 10 lowest energy structures; Int E, average change in interface energy of 10 lowest energy structures; RMSD, average total rmsd of 10 lowest energy structures; $RMSD_{active}$, average active site rmsd of 10 lowest energy structures; f_{wt}, average fraction wild-type contacts between ligand and protein of the 10 lowest energy structure;

Mutant	SS	Dist	Exp Act	ΔE_{top1}	ΔE_{top10}	Int E	RMSD (Å)	$RMSD_{active}$ (Å)	f_{wt}
WT	–	–	+	−431.6*	−428.7*	−9.2	0.81	0.19	0.89
ΔN24	H	21.2	–	21.9	22.6	−9.3	1.44	1.05	0.86
ΔF25	H	18.4	–	20.5	19.9	−9.5	1.45	0.98	0.86
ΔA28	H	8.9	–	13.9	15.6	−9.4	1.40	0.97	0.86
ΔV29	H	15.2	–	15.3	14.3	−9.1	1.36	1.01	0.86
ΔY81	S	4.7	–	17.6	17.6	−6.6	1.48	2.17	0.61
ΔV82	S	6.5	–	16.6	24.0	−6.0	1.56	1.83	0.56
ΔV83	S	8.8	–	31.6	35.1	−7.9	1.84	1.26	0.68
ΔG84	S	11.4	–	23.0	25.4	−7.1	1.48	1.40	0.63
ΔG122	L	6.8	+	18.3	18.0	−7.9	1.18	2.20	0.66
ΔN123	H	5.8	+	12.9	15.9	−7.0	1.21	0.28	0.65
ΔL140	L	15.6	+	32.3	37.8	−9.1	1.41	0.35	0.79
ΔG141	L	17.7	+	27.9	28.4	−9.4	1.51	0.31	0.79
ΔE146	H	19.9	–	16.6	20.5	−9.2	1.25	0.37	0.83
ΔE147	H	19.3	–	15.3	19.0	−9.3	1.40	0.42	0.82
ΔA148	H	15.8	–	12.7	16.0	−9.2	1.35	0.34	0.85
ΔI149	H	17.0	–	12.4	12.0	−9.4	1.42	0.34	0.85
ΔS168	H	12.8	–	17.7	18.2	−9.5	1.25	0.34	0.85
ΔF169	H	9.3	–	16.1	18.7	−9.4	1.28	0.32	0.86
ΔI170	H	8.6	–	17.7	17.9	−9.5	1.33	0.37	0.84
ΔI171	H	10.6	–	14.1	14.5	−9.5	1.28	0.38	0.85
ΔC172	H	9.6	–	18.1	19.6	−8.2	1.36	0.42	0.75

Mutant	SS	Dist	Exp Act	ΔE_{top1}	ΔE_{top10}	Int E	RMSD (Å)	$RMSD_{active}$ (Å)	f_{wt}
ΔI173	H	6.0	–	18.3	24.4	−8.3	1.32	0.53	0.75
ΔQ174	H	8.0	+	22.5	27.6	−8.5	1.51	0.64	0.73
ΔM175	H	10.9	–	19.2	26.5	−8.2	1.42	0.46	0.76
ΔI176	H	9.7	–	22.8	29.2	−8.1	1.50	1.74	0.67
ΔS177	H	6.7	–	34.1	33.0	−8.3	1.37	0.64	0.64
ΔE178	H	7.7	–	37.3	39.2	−8.0	1.56	1.87	0.68
ΔA179	H	11.3	–	26.4	38.2	−8.5	1.46	1.05	0.61
ΔA180	H	11.2	–	36.7	34.8	−8.6	1.32	0.79	0.65
ΔR181	H	10.3	–	46.7	49.9	−3.7	2.79	1.57	0.31
ΔF182	H	12.7	–	31.5	31.8	−7.2	1.35	1.80	0.67
ΔQ183	H	16.0	–	35.3	33.5	−7.8	1.40	1.22	0.71
ΔY184	H	16.4	–	29.6	31.7	−9.4	1.42	0.42	0.83
ΔA200	L	16.7	+	14.2	17.0	−9.3	1.20	0.23	0.86
ΔP201	L	14.7	+	21.5	20.4	−9.5	1.14	0.20	0.86
ΔD202	H	17.3	–	17.3	16.2	−9.7	1.13	0.22	0.86
ΔP203	H	17.0	–	19.0	18.6	−9.4	1.26	0.24	0.82
ΔS204	H	16.6	+	17.3	17.7	−9.5	1.22	0.22	0.83
ΔV205	H	13.4	+	17.5	18.3	−9.3	1.27	0.24	0.81
ΔI206	H	11.8	–	12.5	15.3	−9.4	1.30	0.23	0.83
ΔT207	H	13.1	–	18.6	17.8	−9.4	1.26	0.22	0.80
ΔN210	H	9.4	–	13.8	13.1	−10.4	1.15	0.22	0.86
ΔS211	H	11.7	–	15.9	16.3	−9.2	1.27	0.23	0.82
ΔW212	H	10.1	–	16.8	17.8	−9.5	1.22	0.19	0.85
ΔG213	H	11.1	–	17.6	16.5	−9.2	1.35	0.22	0.84

* WT energies used as the reference.

Table 3

Statistical results for predicting the activity of deletion mutants from various measures for the 45 poin deletions of ricin measured by Munishkin and Wool

	Threshold	Sensitivity	Specificity	PPV	NPV
sequence distance (residues)	21.0	0.7	0.69	0.22	0.79
structure distance (Å)	13.4	0.7	0.67	0.23	0.80
f_{wt}	0.8	0.70	0.50	0.28	0.86
Δenergy (*REU)	22	0.70	0.61	0.24	0.82
interface energy (*REU)	−8.7	0.70	0.56	0.26	0.84
std dev energy (*REU)	24	0.80	0.58	0.28	0.88
rmsd (Å)	1.35	0.70	0.39	0.33	0.88
$rmsd_{active}$ (Å)	1.0	0.90	0.61	0.29	0.93
$rmsd_{active}$ (Å)	0.5	0.80	0.53	0.30	0.89
std dev rmsd	1.0	0.90	0.61	0.29	0.93
std dev $rmsd_{active}$	0.5	0.90	0.47	0.35	0.95

PPV, Positive Predictive Value; NPV, Negative Predictive Value.

*REU-Rosetta Energy Units

Table 4

Statistical results for simulated data of the activity of deletion mutants from various measure for simulated data for all the residues in the protein

	Threshold	Sensitivity	Specificity	PPV	NPV
f_{wt}	0.8	0.70	0.51	0.76	0.42
Δenergy (*REU)	22.0	0.70	0.62	0.72	0.36
interface energy (*REU)	−8.7	0.70	0.56	0.74	0.39
std dev energy (*REU)	24	0.80	0.59	0.76	0.47
rmsd (Å)	1.35	0.70	0.40	0.80	0.47
$rmsd_{active}$ (Å)	1.0	0.90	0.62	0.77	0.63
$rmsd_{active}$ (Å)	0.5	0.80	0.53	0.78	0.51
std dev rmsd	1.0	0.90	0.62	0.77	0.63
std dev $rmsd_{active}$	0.5	0.90	0.48	0.81	0.70

PPV, Positive Predictive Value; NPV, Negative Predictive Value.

*REU-Rosetta Energy Units

Proteins. Author manuscript; available in PMC 2012 October 1.