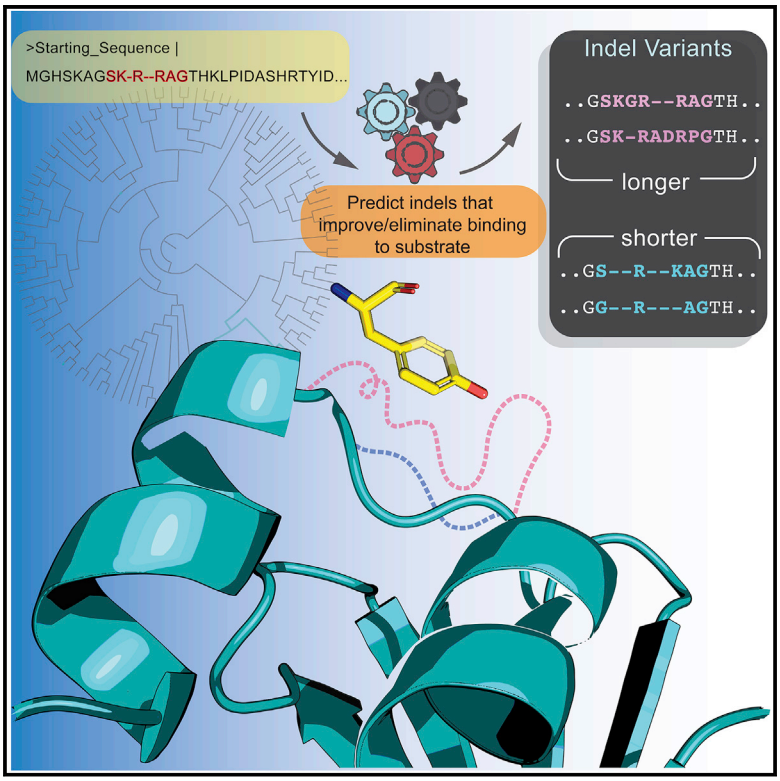
Resource

Structure

IPRO+/–: Computational Protein Design Tool Allowing for Insertions and Deletions

Graphical Abstract



Highlights

- Novel protein design tool to predict amino acid indels (and substitutions)
- Family sequence alignment gives indel probability at a given residue position
- Sampling full sequence space accessible to protein's family by varying chain length
- Guarantees faster recovery of indels than random sequence design ($p < 10^{-15}$)

Authors

Ratul Chowdhury, Matthew J. Grisewood, Veda Sheersh Boorla, Qiang Yan, Brian F. Pfleger, Costas D. Maranas

Correspondence

costas@psu.edu

In Brief

Chowdhury et al. describe an optimization-based protein design tool to identify combinations of protein residue positions, which can allow extra amino acids to be inserted, altered, or even removed from the sequence to arrive at a sequence whose corresponding structure achieves a desired binding affinity with a target molecule.

Chowdhury et al., 2020, Structure *28*, 1–14
December 1, 2020 Published by Elsevier Ltd.
https://doi.org/10.1016/j.str.2020.08.003

CellPress

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

Structure

CellPress

Resource

IPRO+/–: Computational Protein Design Tool Allowing for Insertions and Deletions

Ratul Chowdhury,[1] Matthew J. Grisewood,[1] Veda Sheersh Boorla,[1] Qiang Yan,[2] Brian F. Pfleger,[2] and Costas D. Maranas[1,3,*]

[1]Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802, USA
[2]Department of Chemical and Biological Engineering, University of Wisconsin−Madison, Madison, WI 53706, USA
[3]Lead Contact
*Correspondence: costas@psu.edu
https://doi.org/10.1016/j.str.2020.08.003

SUMMARY

Insertions and deletions (indels) in protein sequences alter the residue spacing along the polypeptide backbone and consequently open up possibilities for tuning protein function in a way that is inaccessible by amino acid substitution alone. We describe an optimization-based computational protein redesign approach centered around predicting beneficial combinations of indels along with substitutions and also obtain putative substrate-docked structures for these protein variants. This modified algorithmic capability would be of interest for enzyme engineering and broadly inform other protein design tasks. We highlight this capability by (1) identifying active variants of a bacterial thioesterase enzyme ('TesA) with experimental corroboration, (2) recapitulating existing active TEM-1 β-Lactamase sequences of different sizes, and (3) identifying shorter 4-Coumarate:CoA ligases with enhanced *in vitro* activities toward non-native substrates. A separate PyRosetta-based open-source tool, Indel-Maker (http://www.maranasgroup.com/software.htm), has also been created to construct computational models of user-defined protein variants with specific indels and substitutions.

INTRODUCTION

Protein design is a core task that underpins many applications, from drug design (Kuhlman and Bradley, 2019) and enzyme engineering for improved or altered substrate specificity (Hernández Lozada et al., 2018) to antibody design for nanomolar affinity for a specific epitope in an antigen (Kumar et al., 2011) or protein pore for (bio)separations (Chowdhury et al., 2018a). At its core, protein design entails the identification of the exact sequence of the amino acids in a polypeptide chain that upon folding leads to the right structure for the desired function. The combinatorial nature of this task arises from the fact that there are 20 amino acid choices for each of the typically hundreds of positions in the polypeptide chain. This implies that exhaustive searches using combinatorial libraries can sample only a tiny fraction of the sequence space. Therefore, it is important to find ways to focus libraries on the most promising combinations of sequence space. Directed evolution (Romero and Arnold, 2009) protocols through a sequence of screens (or preferably selections) have been quite successful in steering libraries toward improved designs (Chowdhury and Maranas, 2019). However, the chances of success are problem specific; limited insight is gained into the molecular mechanism of improvement, and lessons learned from one case study do not always translate to others. At the same time, computational design (Lippow and Tidor, 2007) has emerged as an important tool for homing in on design alternatives that optimize a set of computationally accessible metrics (e.g., binding affinity, overall protein stability, decoy rejection, etc.). A number of success stories for the *de novo* design of enzymes (Hecht et al., 2004; Kaplan and DeGrado, 2004; Khersonsky et al., 2011; Richter et al., 2011), antibodies (Lippow et al., 2007; Kuroda et al., 2012; Lapidoth et al., 2015; Chowdhury et al., 2018b), and inhibitors (Kortemme et al., 2004; Rämisch et al., 2014) have been reported.

Existing computational tools rely on either biophysics-inspired scoring functions (e.g., CHARMM (Brooks et al., 2009), AMBER (Case et al., 2005), GROMACS (Spoel et al., 2005)), to quantify the energetics of the molecular interactions allowing for the *in silico* exploration of the impact of amino acid mutations on binding, or stability metrics. A number of these techniques form the basis of software platforms for protein design, such as Rosetta (Leaver-Fay et al., 2011), Site Directed (Pandurangan et al., 2017), OSPREY (Gainza et al., 2013), and Tinker (Rackers et al., 2018). Alternatively, a number of protein design tools rely on the analysis of the statistics of the amino acid combinations in a protein library that preserve (or enhance) a particular function (Xiong et al., 2014). Often these methods are supplemented by structural motif information (Wu and Zhang, 2007) associated with the desired functionality.

There have been several protein design efforts in the past that allow for insertions or deletions (indels) as a part of grafting or recombination of existing parts chosen from other protein structures in the PDB. The method SEWING (Jacobs et al., 2016) assembles *de novo* protein structures by recombining different

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

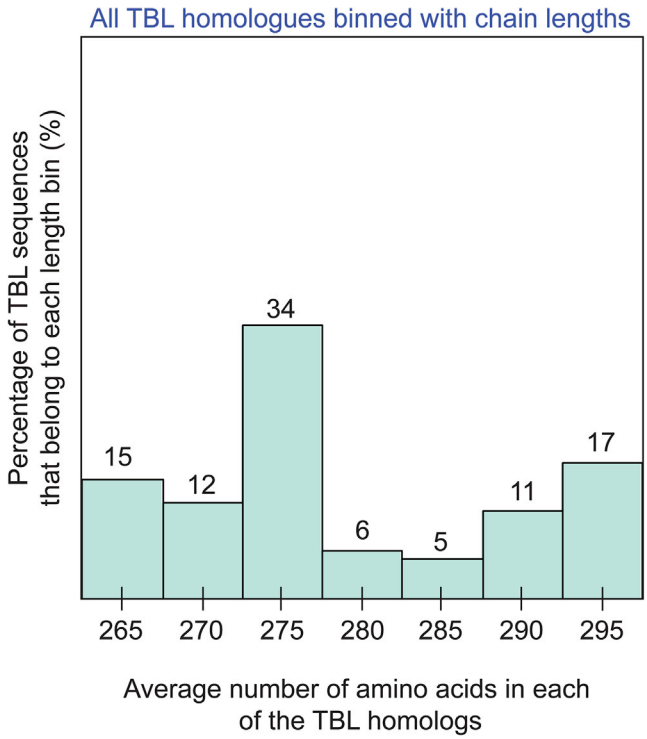


Figure 1. Histogram Showing the Distribution of 156 TEM-1 β-Lactamase Homologs with Respect to the Number of Amino Acids that Constitute the Polypeptides

The sequences have been grouped into bins of five-amino-acid increments, with the mean lengths indicated in the x-axis labels.

structural parts from existing proteins. Alternatively, Netzer et al. designed proteins with altered binding specificities (Netzer et al., 2018) by grafting various loop regions, whereas Lapidoth et al. designed different backbones conforming to a fixed-active-site geometry to generate active enzymes spanning high sequence and structural diversity (Lapidoth et al., 2018). All such methods are mainly focused on combining specific structural parts toward generating *de novo* structures with a desired function. Despite the many success stories and rapid progress, there is still a need for a systematic method that can successively combine amino acid substitutions along with additions and deletions toward achieving a specific binding function without biasing the search based on known structural motifs. Generally, existing computational protein redesign methods require that the original length of the polypeptide chain is prespecified and remains unchanged during the design process.

For example, consider the family of TEM-1 β-Lactamases (TBLs). It catalyzes the formation of the hydroxyl-substituted β-amino acid from its corresponding β-lactam (such as penicillins and cephalosporins) using the conserved Met69, Ser130, and Arg244 catalytic triad. Its high antibiotic resistance makes it a convenient candidate for assessing protein-associated fitness using minimum inhibitory concentration (MIC)-like approaches in high-throughput protein evolution experiments. Multiple studies have shed light onto the effects of indels on ampicillin resistance. Two studies independently assaying 53 and 87 amino acid deletions across the protein reported that more than 26% of variants showed a 99% loss of wild-type MIC score when the deletion occurred in α helices or β sheets, whereas deletions in loops and β-sheet-loop junctions were almost always tolerated. To quantify the prevalence of indels in protein family sequences and demonstrate the problem that this causes for computational protein design, we generated a sequence alignment of a published set of 156 class A β-Lactamases from several bacterial species. Despite the relatively high average sequence similarity (i.e., 86%), there are on average six gaps per sequence in the library and five insertions (see also Figure 1), with respect to *Escherichia coli* TBL (referred to as *Ec*TBL henceforth). The most prevalent backbone size has 275 amino acids, which represents only 34% of the total number of sequences in the family. This means that if a computational protein design algorithm attempted a TBL redesign starting from a member of the aa = 275 grouping, only 34% of the TBL family diversity as encoded in the sequence alignment would be accessible. A random starting sequence would access on average only 15% of the protein family diversity. This becomes even more restrictive for most other protein families that tend to have even more gaps in their alignment due to the lower sequence identity. This means that even though nature seems to extensively use protein length as a "lever" to optimize the function of proteins, existing protein design tools are always restricted to a particular chain length. One could iteratively attempt to apply computational protein design for different lengths, but this is clearly inefficient, as the design goal is not used to guide the search for the most advantageous protein length. This calls for a dedicated method that uses protein size as a design criterion.

In enzyme design, often the objective is to alter the specificity of the enzyme for a new substrate. For example, being able to switch the specificity of an acyl-acyl-carrier protein (ACP) thioesterase from long-chain-acyl acyl-ACP (C14 and higher) to short-chain (C8) acyl-ACPs could unlock microbial octanoate production with implications for the oleochemical industry. Thioesterases hydrolyze the thioester bond in acyl-ACPs and yield the corresponding acid. Thioesterases from different species vary both in length and in substrate size preference. It is reasonable to assume that if substrate size is the only changing factor, then polypeptide chain length should be an important design variable. Deletions in the substrate-binding groove can result in a more compact thioesterase with a remodeled pocket that can accommodate only shorter ACPs. Insertions, on the other hand, could be used to "open up" otherwise smaller pockets for larger substrates. This is observed for two *Cuphea viscosissima* acyl-ACP thioesterase variants (*Cv*FatB1 and *Cv*FatB2) sharing a 70% sequence similarity. The longer variant, *Cv*FatB2 (80 more amino acids), acts on only C14- to C16-ACPs, whereas the shorter variant, *Cv*FatB1, accepts only up to C8-ACPs (Jing et al., 2018). However, it must be noted that the effect of a deletion or insertion could also have a counterintuitive role. For example, if the active site is partially occluded by neighboring loops, a deletion in a loop may enable access to larger substrates. Conversely, an insertion may preclude access to larger substrates, thus changing specificity toward smaller ones.

Motivated by these observations and the lack of computational tools, here we introduce a modified version of the protein design tool IPRO+/− that allows for both residue indels. IPRO+/− builds upon the existing suite of protein design

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

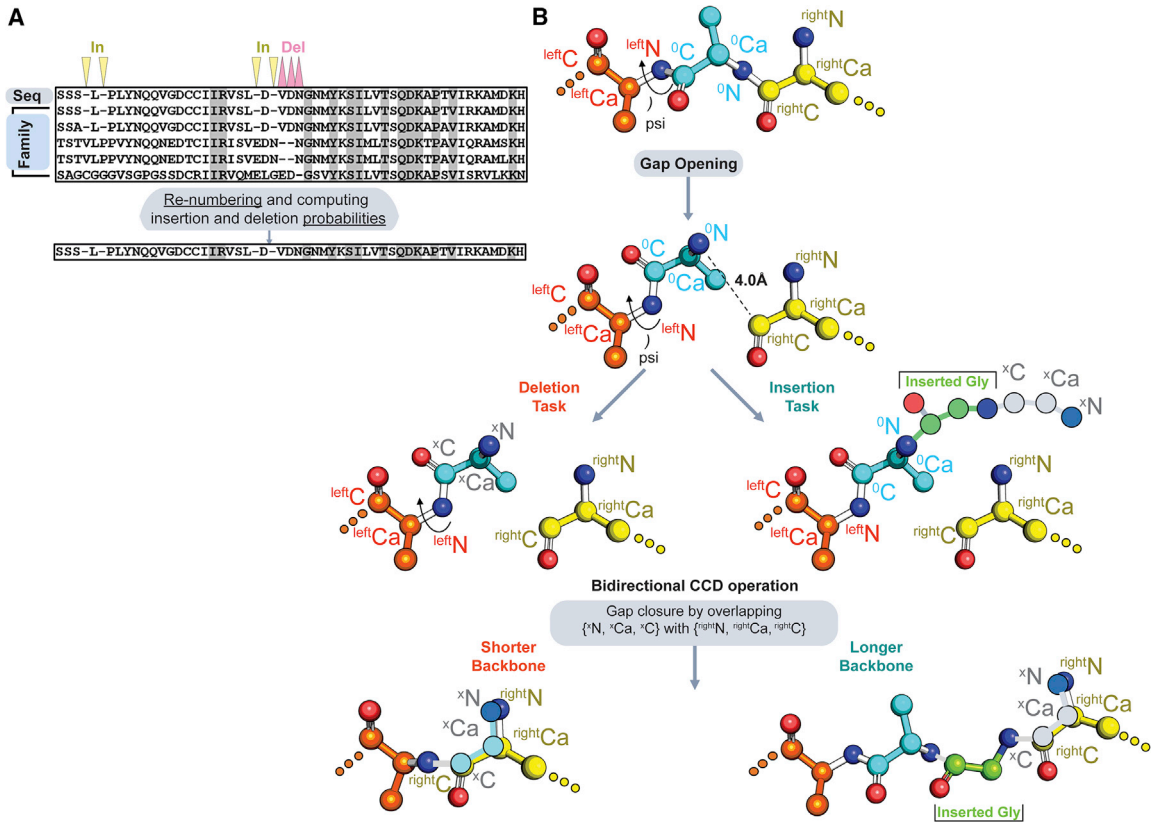


Figure 2. Schematic Overview of Indel Probability Calculation and Protein Backbone Breaking and Re-stitching to Obtain Indel Variant Structure

(A) Family sequence alignment is used as the blueprint to determine allowable amino acid insertion and deletion locations and corresponding probabilities.

(B) Gap opening and gap closure steps for insertion and deletion tasks yield longer and shorter backbone lengths, respectively.

programs IPRO (Pantazes et al., 2015), which use the CHARMM (Brooks et al., 2009) energy function to quantify the energetics of molecular interactions and a mixed-integer linear optimization algorithm to select residue-rotamer combinations that maximize the user-specified design objective (e.g., interaction energy). Conceptually, IPRO+/− allows for indels by allowing every position in the protein sequence to either accept 1 of the 20 amino acids or remain unoccupied (i.e., gap). The family protein sequence alignment provides the blueprint for which positions can remain unoccupied by simply inspecting whether there exists any member that has a gap in the position of interest. In addition, it establishes the maximum number of residue positions and provides a universal numbering scheme for any redesigned protein. This implies that gaps are encoded as the 21st amino acid. Therefore, any chain length contained in the protein family alignment is accessible by IPRO+/− with gaps allowed only at positions where there is already at least one member with a gap in the family protein alignment (see Figure 2A).

The key redesign steps in IPRO+/− are (1) deletion of a residue (aa → _) or (2) or insertion of a residue in a gapped position (i.e., _ → aa). The protein family alignment provides a straightforward way to quantify the probability of occurrence of these two transitions. For example, if, at a given position, A of N proteins involve a gap, then we choose the probability of opening a gap in that position to be A/N (see Figure 2A). Similarly, if B of N proteins in the alignment have a residue at a currently gapped position, then we set the probability of adding a residue at B/N (see Figure 2A). One could envision more elaborate schemes for setting these transition probabilities or allow for direct user-supplied specifications. Residue deletion involves cutting the polypeptide chain, removing the residue in question, and then bringing together the two ends. This task (i.e., end joining) forms a frequently occurring problem in robotics for object retrieval. It arises when a sequence of rotations at different articulated joints needs to be calculated, such as when an articulated mechanical arm needs to grab a stationary target object (Martín, Barrientos and del Cerro, 2018; Kundert and Kortemme, 2019). End joining of the polypeptide chain is accomplished in IPRO+/− using a modified cyclic-coordinate descent method. At each cycle, two rotations around the protein backbones at symmetrical positions from the end-joining locations are carried out with the goal of minimizing the distance between the backbone N-Cα-C triplets of the end-joining segments (see Figure 2B). This sequence of pairwise rotations is initiated five residues away from the end-joining segments. Progressively, pairwise rotations move closer until they meet one another (root-mean-square deviation [RMSD] <0.001Å) at the joining segment during the last cycle (see Figure 2B). By carrying out rotations in a symmetric manner (bidirectional cyclic coordinate descent [CCD]) around the joining end, we avoid any possible direction biases. The same end-joining cyclic

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

procedure is called upon both for deletions and after an insertion of a residue at a given position. A glycine is exclusively introduced at all residue addition events, which can be changed into other residues in follow-up steps of IPRO+/−. In both cases, the torsion angles of the backbone chain are changed, and therefore, new rotamer assignments need to be made using the MILP algorithm in IPRO. Figure 2B illustrates the basic steps of the loop-closure protocol implemented inside the IPRO+/− algorithm. Detailed algorithmic details and implementation information are described in the STAR Methods.

As a demonstration, we tested IPRO+/− for the redesign of a TBL to assess whether IPRO+/− managed to identify designs with a chain length different from the starting point by opening gaps and/or inserting residues in positions consistent with the family sequence alignment. In addition, we carried out the redesign of a 4CL2 (i.e., 4-Coumarate:CoA ligase 2 from *Glycine max*) so that it shows new substrate specificity toward the larger cinnamate, caffeate, and ferulate substrates. We assessed whether IPRO+/− managed to recapitulate the pattern of deletions seen in the protein family alignment for 4CL2 enzyme variants that have activity for larger substrates (such as sinapate and ferulate).

RESULTS

IPRO+/– Algorithmic Description

The traditional IPRO workflow design iteration consists of a backbone perturbation, a rotamer repacking, and amino acid selection using a mixed-integer linear programming approach, target molecule redocking, computing interaction energy metrics, and deciding whether to retain or reject the design, which is followed by a backbone perturbation at a different site. An extended set of decisions and tasks is incorporated within the IPRO workflow such that indels can be used as design choices along with substitutions. The probability of making an indel at each design cycle is guided by the family sequence alignment (see Figure 2A). For either insertion or deletion, the polypeptide backbone must be first opened by performing a ψ-angle rotation on the residue to the left of the break such that the distance between ^{0}N and $^{\text{Right}}$C is at least 4 Å (see Figure 2B). The two new backbone ends are then generated by either appending a new $^{\text{Gly}}$N-$^{\text{Gly}}$Cα-$^{\text{Gly}}$C triplet in case of insertion or removing the N^{0}-Cα^{0}-C^{0} triplet in case of deletion (see Figure 2B). IPRO's loop-closure algorithm is an adaptation of the CCD method, which has been employed in homology modeling (Canutescu and Dunbrack, 2003) and in robotics (Kenwright, 2013) for solving inverse kinematics problems. The objective of the loop-closure algorithm is to minimize the RMSD between the N-Cα-C triplets of the two free ends (see Figure 2B). IPRO+/− first renumbers residues in the protein by adding all gapped positions in the original sequence that accept amino acids for some of the protein family sequence members (see Figure 2A). The insertion step is performed before the rotamer repacking and amino acid selection step, which provides an opportunity for the inserted glycine to be altered to a different amino acid. The deletion step, on the other hand, is executed before the target molecule redocking step of IPRO.

Step 1. Renumbering Amino Acids Based on Family Sequence Alignment

IPRO+/− treats a gap in the sequence alignment as the 21st amino acid (see Figure 2A). The family sequence alignment is used to identify design positions (DPs) on the starting sequence that could accept a different amino acid (aa → aa), insert a glycine (_ → aa), or delete an amino acid (aa → _). Gaps on the starting sequence are assigned residue numbers and the rest of the amino acids are renumbered accordingly (see Figure 2A). Insertion and deletion probabilities for each position are computed as a fraction of sequences in the alignment that have amino acids or gaps in those positions, respectively.

Step 2. Loop Opening

Both insertion and deletion steps start with initiating a break in the polypeptide backbone where an indel needs to be made. As shown in Figure 2B, the break is introduced by performing a ψ-angle rotation on the residue to the left of the intended break point such that the distance between the ^{0}N and the $^{\text{Right}}$C is at least 4 Å. For an insertion task, a glycine ($^{\text{Gly}}$N, $^{\text{Gly}}$Cα, $^{\text{Gly}}$C triplet) is built onto the ^{0}N atom and an extra set of backbone atoms (^{x}N, ^{x}Cα, and ^{x}C) is built onto the $^{\text{Gly}}$N atom (see Figure 2B). Alternatively, in the case of a deletion task, the ^{0}N-^{0}Cα-^{0}C triplet is renamed ^{x}N-^{x}Cα-^{x}C.

Step 3. Loop Closure

The newly generated ends of the polypeptide backbones after an insertion or deletion task are rejoined using bidirectional CCD to obtain the structure of the corresponding indel variant. Loop closure contains a series of φ- and ψ-dihedral rotations with the objective of reducing the RMSD between ^{x}N-^{x}Cα-^{x}C and $^{\text{Right}}$N-$^{\text{Right}}$Cα-$^{\text{Right}}$C triplets to less than 0.001 Å (see Figure 2B). Loop-closure operation involves φ-ψ rotations performed alternately on residues lying to the left and right of the gap, starting with the fifth residue away from the gap and progressively moving toward the gap. The justification for using five residues is described in the Supplemental Information (and Figure S1). Up to 20 rounds of such operations are performed until the polypeptide backbone ends assume the same coordinates (i.e., are rejoined). This bidirectional CCD (Canutescu and Dunbrack, 2003) approach safeguards against any directional bias (toward left or right of the original polypeptide break point). The sequence of φ-ψ rotations is as follows: $^{\text{Left}}_{5}\varphi\psi \rightarrow {}^{\text{Right}}_{5}\varphi\psi \rightarrow {}^{\text{Left}}_{4}\varphi\psi \rightarrow {}^{\text{Right}}_{4}\varphi\psi \rightarrow {}^{\text{Left}}_{3}\varphi\psi \rightarrow {}^{\text{Right}}_{3}\varphi\psi \rightarrow {}^{\text{Left}}_{2}\varphi\psi \rightarrow {}^{\text{Right}}_{2}\varphi\psi \rightarrow {}^{\text{Left}}_{1}\varphi\psi \rightarrow {}^{\text{Right}}_{1}\varphi\psi \rightarrow {}^{x}_{0}\varphi\psi$. The Left/Right superscripts indicate residue location relative to the polypeptide break, and the subscript numbers from 1 to 5 its corresponding distance (see Figure 2B).

Step 4. Integration with Remaining IPRO Workflow

The original IPRO workflow iterates between every five steps for deciding on amino acid substitutions in user-defined positions (i.e., DPs) that optimize a binding assembly metric (such as complex energy for enhanced stability or interaction energy for improved binding with a small molecule or another protein). The algorithmic and implementation details are described by Pantazes et al. (Pantazes et al., 2015). Briefly, the first step of IPRO starts with picking a DP at random, followed by perturbing the backbones of an 11-residue window centered around the DP. Repacking of amino acid side chains of this window is then performed by solving a MILP, where only the DP is allowed to receive rotamers different from the original amino acid type. In IPRO+/−, if the randomly chosen DP is a gap, a glycine is introduced with a probability equal to the insertion probability of that site from the family sequence alignment. In subsequent iterations, this glycine can be replaced and repacked with a different amino acid rotamer depending on its covalently bonded and

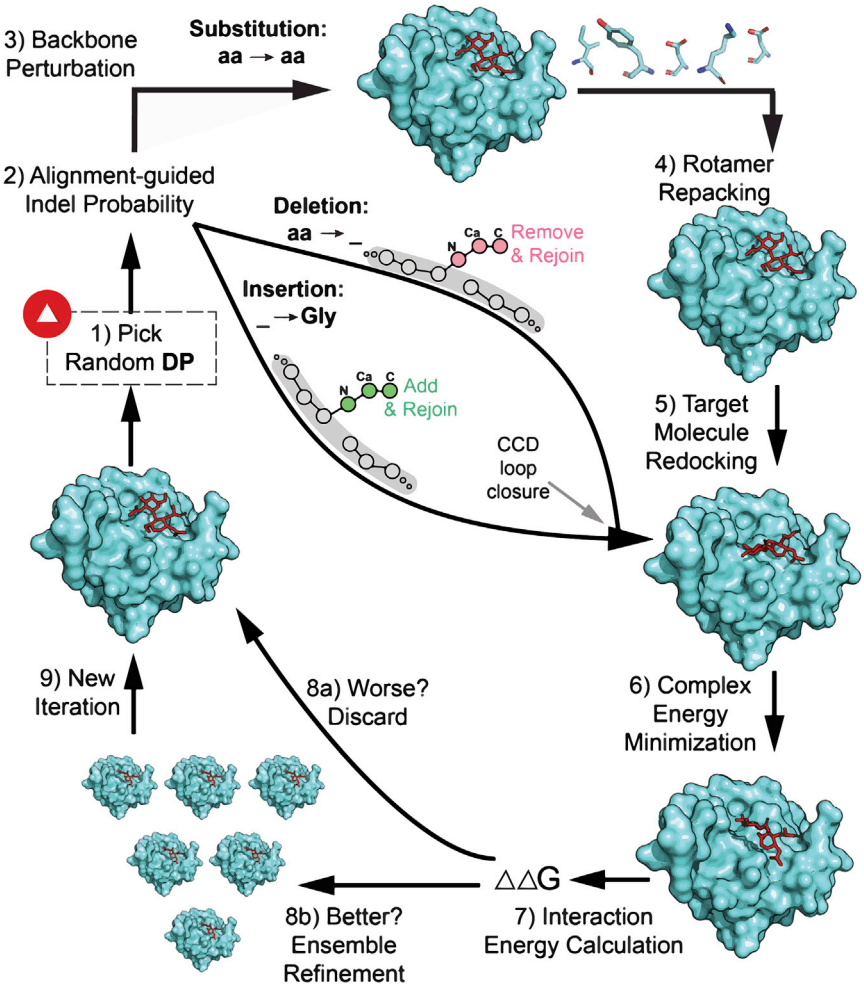


Figure 3. The Steps of the IPRO+/– Design Cycle

Family sequence alignment guides the probabilities of making amino acid insertions or deletions along the polypeptide backbone. After an insertion step, the inserted Gly is allowed to be replaced with a different amino acid from subsequent design cycles. CCD, cyclic coordinate descent; DP, design protein.

non-bonded interaction-energy scores. On the other hand, if the randomly chosen DP is a residue with a non-zero deletion probability, it is deleted with the said probability after the rotamer repacking step of a design cycle. Subsequent target molecule redocking with the protein, energy minimization of the protein (in complex with its binding partner, if present), and CHARMM-interaction energy scores are computed. A design is retained if it performs better than the current best variant for the intended design goal or else rejected with a probability using a Metropolis criterion. IPRO+/− currently uses a default of 3,000 such design runs for a full simulation, on five nodes of 10-core Xeon E7-4830 processors with 4 GB physical memory. A schematic of the IPRO+/− steps is given in Figure 3. The algorithm can be accessed freely from http://www.maranasgroup.com/software.htm and requires the user to have CHARMM (Brooks et al., 2009) and GAMS (Bussieck et al., 2012) licenses.

Experimental Testing of *E. coli* 'TesA Indel Variants Shows up to 30% Activity Improvement

The IPRO+/− algorithm was first used to identify amino acid substitution and deletion combinations at five active-site residues of the thioesterase enzyme ('TesA) from *E. coli* to identify variants that show preferential activity toward short-chain (C8) ACPs with experimental expression of the variants and measurement of product titers. IPRO+/− was then used to identify shorter and longer variants of TBL, with the *Ec*TBL (PDB: 1ERM; Ness et al., 2000) as the input/design sequence. Bacterial resistance to penicillin and cephalosporin-like drugs is primarily governed by β-lactamase-mediated hydrolysis of the drug molecule. We hypothesize that variants that form stable complexes with the penicillin-like boronate (PEB; phenylacetamido-carboxyphenyl ethyl boronate) inhibitor are likely to confer antibiotic resistance. In addition, IPRO+/− was used to design shorter variants of a lignin biosynthesis pathway enzyme, starting with soybean 4CL (*Gm*4CL2) enzyme (homology modeled against *Nt*4CL2, PDB: 5BST; Li and Nair, 2015), that show improved binding to larger substrates (such as sinapate and ferulate) than its native substrate (4-coumarate).

We used IPRO+/− to predict not more than three amino acid substitutions or deletions in only five active-site residues of *E. coli* 'TesA, with the objective of identifying active variants with enhanced turnover of C8:0-ACP. To test the impact of predicted mutations on thioesterase activity, each of the 10 variant sequences (see Figure 4) was cloned into an expression vector, and the resulting fatty acid profiles were quantified by gas chromatography with flame ionization detector. All predicted mutants could be expressed and showed activity. Two of the 10 designed 'TesA variants (N215_ and D268_ G270L) demonstrated marginally higher (~130%), and 8 showed lower, activities compared with the original 'TesA sequence. However, the product portfolio was not significantly different between variants. All variants containing a substitution at I210 demonstrated compromised activity (~35% of 'TesA), while variants containing a mutation at P213 demonstrated activity between 35% and 80% of 'TesA. It is noteworthy that both the high-activity variants contain an amino acid deletion each (N215_ and D268_, respectively) and would be missed if the design procedure relied entirely on substitutions. Subsequent systematic exploration of the sequence space by including additional DPs around the active pocket to improve C8 specificity of the active 'TesA variants has been left for a detailed follow-up study.

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

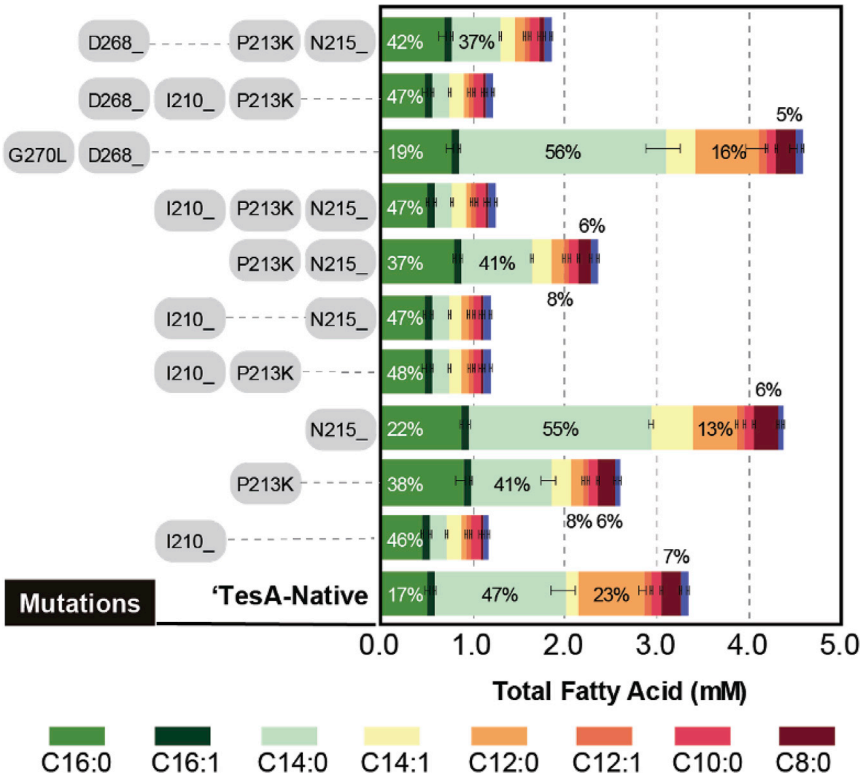


Figure 4. Fatty Acid Production Profiles for the Native 'TesA and Computationally Designed Variants

Data represent the averages of triplicate cultures normalized to an internal standard. The percentage of each chain length for C16, C14, C12, and C8 is indicated along with the stacked bar chart.

Naturally occurring indels in TBLs captured by IPRO +/− TBL are commonly used as a model for protein evolution experiments as they confer resistance to penicillin and cephalosporin-like drugs, which is used as a proxy for protein fitness. A recent study by Gonzalez et al. (Gonzalez et al., 2019) investigated a comprehensive library of 5,270 amino acid insertion and 286 deletion variants of *Ec*TBL (PDB: 1ERM) capable of penicillin resistance. Overall, protein stability was found to be least affected by single-amino-acid indels in loops, followed by indels in tertiary structure-loop junctions, helices, and sheets. We selected as DPs within IPRO+/− 13 positions spanning four of the six loops that involved indels in at least one enzyme family member (see Figures 5A and 5B) from a family sequence alignment with 156 other bacterial β-lactamases (alignment reported by Gonzalez et al., 2019). Enzyme-inhibitor complex energy score was minimized, and only designs with enzyme-inhibitor interaction scores not worse by more than 25% of that of the starting sequence (*Ec*TBL) were accepted. Nine of these 13 positions could be replaced with a gap, whereas 4 gapped positions in the starting sequence *Ec*TBL could be filled with a residue (see Figure 5). In addition, because substitution G120A was co-occurring in 78% of β-lactamase family members that involved deletion D119_, we added Gly120 as a DP that could be substituted but not deleted. The family sequence alignment with 156 homologous TBLs provided the deletion/insertion probabilities at each one of the 13 DPs. The goal was to assess whether IPRO+/− identified backbone length modifications and residue substitutions that mirrored those seen in the natural family of β-lactamases (see Figure 5).

IPRO+/− could in principle sample designs with backbone lengths between 261 and 264 residues through the accumulation of deletions or additions at the 13 DPs by performing 3,000 IPRO+/− redesign runs starting from the 263-residue homolog (*Ec*TBL). Each redesign run was capped at 20 loop-closure iterations for restitching the backbone after an insertion or deletion step. This is a user-defined number and can be adjusted for performing larger contiguous indels. Thirty-five redesigns with at least one indel were identified with interaction energy scores not worse by more than 25% of the wild type (see Table 1). Notably, seven of nine (78%) and five of seven (71%) of the naturally occurring deletions and insertions, respectively, were recovered in these 35 IPRO+/− designs (see Figure 5). Each of the nine possible deletion sites could have 1 of 21 fates (20 amino acids and 1 gap). Thus, the probability of recovering the best design (E64_, K55V, E64K design R1.D1 in Table 1) by a randomly selecting a procedure following a binomial distribution after 3,000 runs is estimated to be 3.8×10^{-19} (computed as $3{,}000 \times (1/21)^9$). This provides confidence that redesigns proposed by IPRO+/− could not have been identified by random chance in 3,000 design cycles. In addition, the identified insertions (_99P, _99V, _115G, _118G, and _118T) match the residue seen in the protein family alignment even though the originally added residue was Gly. Designs with deletions in loops 2 and 3 (positions 62–64, and 99, respectively) enhanced binding to the PEB inhibitor (see designs such as R1.D1, R1.D2, R1.D4 and so on) by reorienting the neighboring polar residue side for electrostatic stabilization of the electronegative carboxyphenyl moiety of the PEB inhibitor, which is corroborative of Gonzalez et al. (2019). Insertions at position 99 in loop 3 yielded the most stable complexes while retaining wild-type substrate binding activity (see designs R1.D5, R1.D6, R1.D11, R1.D14, R1.D17, and R1.D19 indicated with an asterisk in Table 1). This is because even though the location of the inserted residue 99 is approximately 23 Å away from the substrate and does not affect substrate binding, the newly introduced residues (_99P, _99G, and _99V) serve as stabilizing anchors by establishing hydrophobic contacts with the neighboring Leu114 side chain and electrostatic contacts (using the backbone N and O atoms of residue 99) with the side chains of Asp119 and Thr116, from loop 4.

IPRO+/− also predicted six additional insertions in the native *Ec*TBL sequence in the four gapped positions (99, 115, 117, and 118), where the inserted amino acid type in these positions is not observed in any of the family sequences (see Figure 5). Except for an insertion prediction (positively charged _117R) at

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

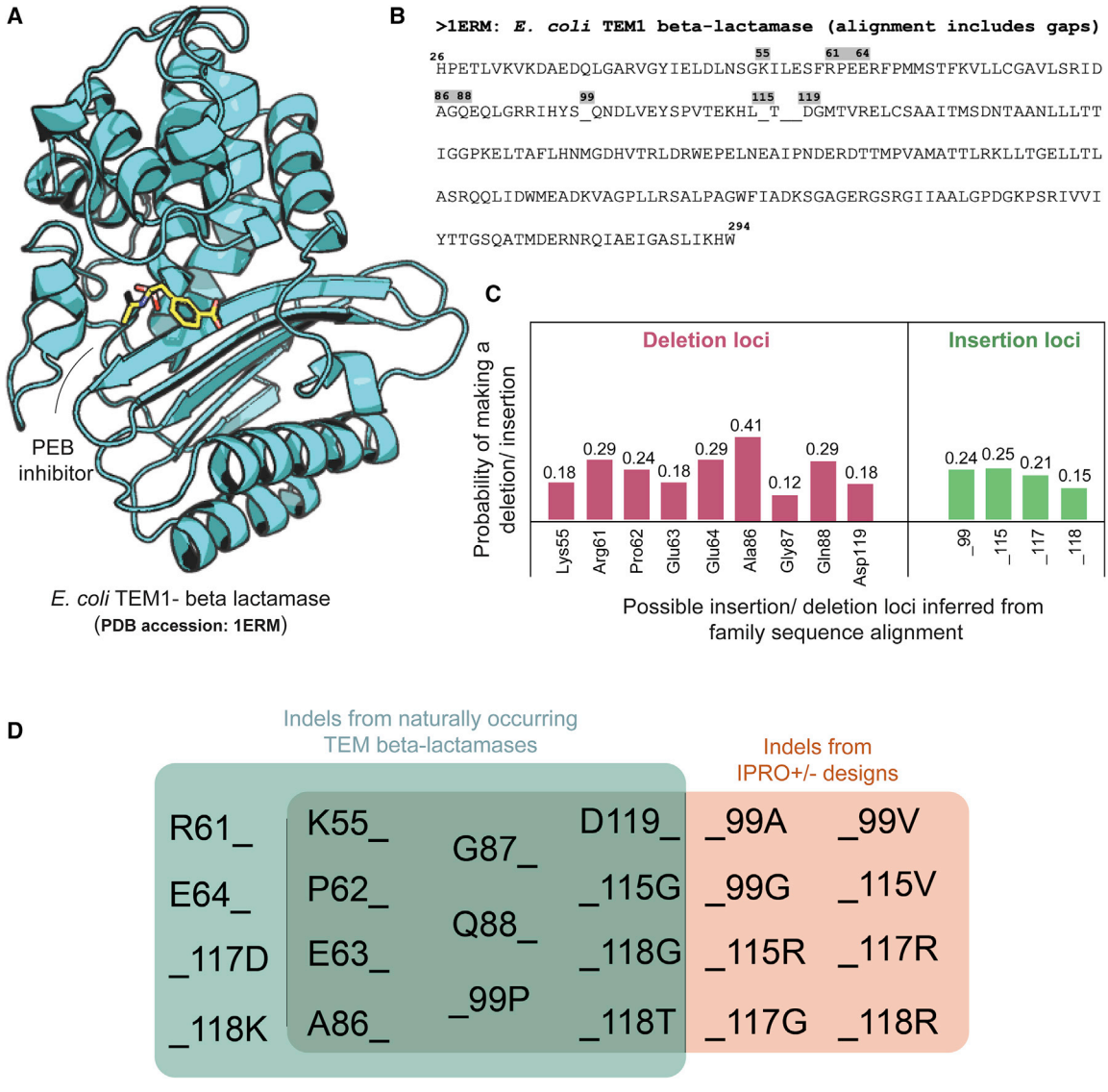


Figure 5. IPRO+/– is Shown to Recapitulate Majority of Naturally Occuring Indel Variants as Stable Complexes with PEB Inhibitor

(A) TEM-1 β-lactamase from *E. coli* (1ERM) in complex with boronate (PEB) inhibitor.

(B) Wild-type TEM-1 β-lactamase aligned with 48 β-lactamase sequences (including gaps) shows the positions that can have an indel.

(C) The insertion and deletion probabilities of each position were selected from the multiple sequence alignment with 48 other natural β-lactamase homologs.

(D) Venn diagram shows the fraction of naturally occurring indels in TBL homologs that were recovered by IPRO+/− simulations on *Ec*TBL.

position 117 (naturally occurring insertion is negatively charged _117D), all other predicted insertions had side chain types consistent with natural sequences (position 99, hydrophobic, and positions 115 and 118, polar). For example, arginine in _118R has a positively charged side chain similar to lysine seen in bacterial TBL from *Bosea lupini* (UniProt: A0A3Q9AU82) in that position. In addition to recovering _99P as seen in *Cobetia sp.* strain: MM1IDA2H-1, IPRO+/− designs predicted alternative hydrophobic amino acids such as glycine and alanine at position 99 with similar interaction and complex energy scores computed using CHARMM.

The computational models of the designed loops were very similar (RMSD <0.24 Å) to those obtained from crystal structures of the loops containing the same indels. Twenty-three of the 156 TBL homologs used for the sequence alignment had reported crystal structures and were used for assessing the quality of structures of *Ec*TBL variants predicted by IPRO+/−. The average Cα-RMSDs of only the loop regions from designs and reported crystals with the same single-amino-acid deletions of Lys55, Pro62, Glu63, Ala86, Gly87, Gln88, and Asp119 were 0.21, 0.17, 0.05, 0.14, 0.15, 0.17, and 0.23 Å, respectively. Furthermore, the φ-ψ angles of the backbone atoms were in the exact same region of the Ramachandran plot for nearly 30% of all the insertions as observed from the crystal or homology-modeled structures. The remaining inserted residues in the designed structures explored different backbone conformations

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

Table 1. CHARMM Interaction and Complex Energy Scores of 35 1ERM Designs from IPRO+/–

Variant ID	Amino Acid Transitions in *Ec*TBL	Variant-Inhibitor Complex Energy Score (CHARMM Energy Units)	Variant-Inhibitor Interaction Energy Score (CHARMM Energy Units)	% Increase (↑)/Decrease (↓) in Interaction Energy Score with Respect to *Ec*TBL
Wild-type *Ec*TBL	None	−10,051.29	−59.88	0
R1.D1	E63_, K55V, E64K	−10,106.01	−93.06	55.41 ↑
R1.D2	P62_, E63G	−10,054.16	−90.89	51.79 ↑
R1.D3	Q88_, E63R, K55G	−10,077.52	−89.78	49.93 ↑
R1.D4	E63_, P62A, E64D	−10,135.22	−89.64	49.70 ↑
R1.D5*	G87_, A86F, _99V	−10,819.16	−88.01	46.98 ↑
R1.D6*	A86_, G87A, _99V	−10,974.09	−87.61	46.31 ↑
R1.D7	Q88_, A86R, K55L	−10,103.79	−85.99	43.60 ↑
R1.D8	E63_, P62V, E64R	−10,101.15	−84.22	40.65 ↑
R1.D9	G87_, A86G, Q88K	−10,059.36	−83.47	39.40 ↑
R1.D10	G87_, A86W, Q88G	−10,063.58	−82.77	38.23 ↑
R1.D11*	Q88_, A86R, _99G	−10,723.98	−82.61	37.96 ↑
R1.D12	P62_	−10,103.02	−81.07	35.39 ↑
R1.D13	Q88_, A86K, K55L	−10,080.03	−78.07	30.38 ↑
R1.D14*	_99G, P62K	−10,904.63	−74.62	24.62 ↑
R1.D15	_117G, G120A	−10,103.06	−72.30	20.74 ↑
R1.D16	_118R, D119G	−10,103.21	−69.99	16.88 ↑
R1.D17*	_99P, P62G, E63K	−10,800.1	−69.69	16.38 ↑
R1.D18	_118G, D119_, G120L	−10,116.84	−68.77	14.85 ↑
R1.D19*	_99G, P62A	−10,884.59	−67.65	12.98 ↑
R1.D20	_115G, T116A	−10,282.3	−67.32	12.42 ↑
R1.D21	_115V, D119S	−10,023.44	−65.88	10.02 ↑
R1.D22	_115D, T116G	−10,090.24	−64.32	7.41 ↑
R1.D23	_115V, T116E	−10,103.52	−63.44	5.95 ↑
R1.D24	_115R, T116A	−10,306.72	−62.92	5.08 ↑
R1.D25	_115R, T116V	−10,090.24	−60.70	1.37 ↑
R1.D26	_117G, D119T	−10,203.52	−60.56	1.14 ↑
R1.D27	K55_, P62D, E63G	−10,186.72	−59.40	0.80 ↓
R1.D28	_118G, D119_, G120K	−10,097.63	−59.05	1.39 ↓
R1.D29	_117R	−10,130.38	−58.73	1.92 ↓
R1.D30	_118G, D119_, G120A	−10,080.11	−57.82	3.44 ↓
R1.D31	D119_, G120P	−10,067.84	−54.98	8.18 ↓
R1.D32	K55_, P62F	−10,073.95	−54.87	8.37 ↓
R1.D33	D119_, G120A	−10,202.92	−54.69	8.67 ↓
R1.D34	D119_	−10,056.59	−50.63	15.45 ↓
R1.D35	K55_	−10,060.07	−46.66	22.08 ↓

The designs are arranged in descending order of variant-inhibitor interaction energy scores (column 4). These designs sample indels that are seen in natural homologs and have complex energy scores (stability metrics) not less than 90% of the wild-type TEM-1 β-lactamase complex energy with PEB inhibitor. Designs that constitute the most stable complexes are indicated with an asterisk (*) (see also Figure S8).

and side-chain orientations, and consequently established new electrostatic and hydrophobic contacts (with Met121 side chain) to stabilize the substrate and the complex as a whole. The local structural variation between wild-type *Ec*TBL and the top six indel variants (R1.D1–R1.D6 in Table 1) is illustrated in Figure S8.

Structural aspects of the active site are preserved between the wild type and the indel designs. Importantly, the catalytic Ser70 residue adopts the same wild-type-like dihedral angles (and side-chain orientation) in all designs. One particular design (design R1.D30 from Table 1) was able to perfectly recapitulate both the loop 4 configuration (Cα-RMSD ~0.02 Å) and the sequence as seen in the TBL from *Vibrio pectenicida* (UniProt: A0A427U5F7), which has a combination of an insertion, a deletion, and a substitution (_118G, D119_, G120A) with respect to the starting sequence, *Ec*TBL. IPRO+/−, therefore, lays the foundation for future computational protein design approaches where predicting indels (along with substitutions) would aid the generation of focused libraries.

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

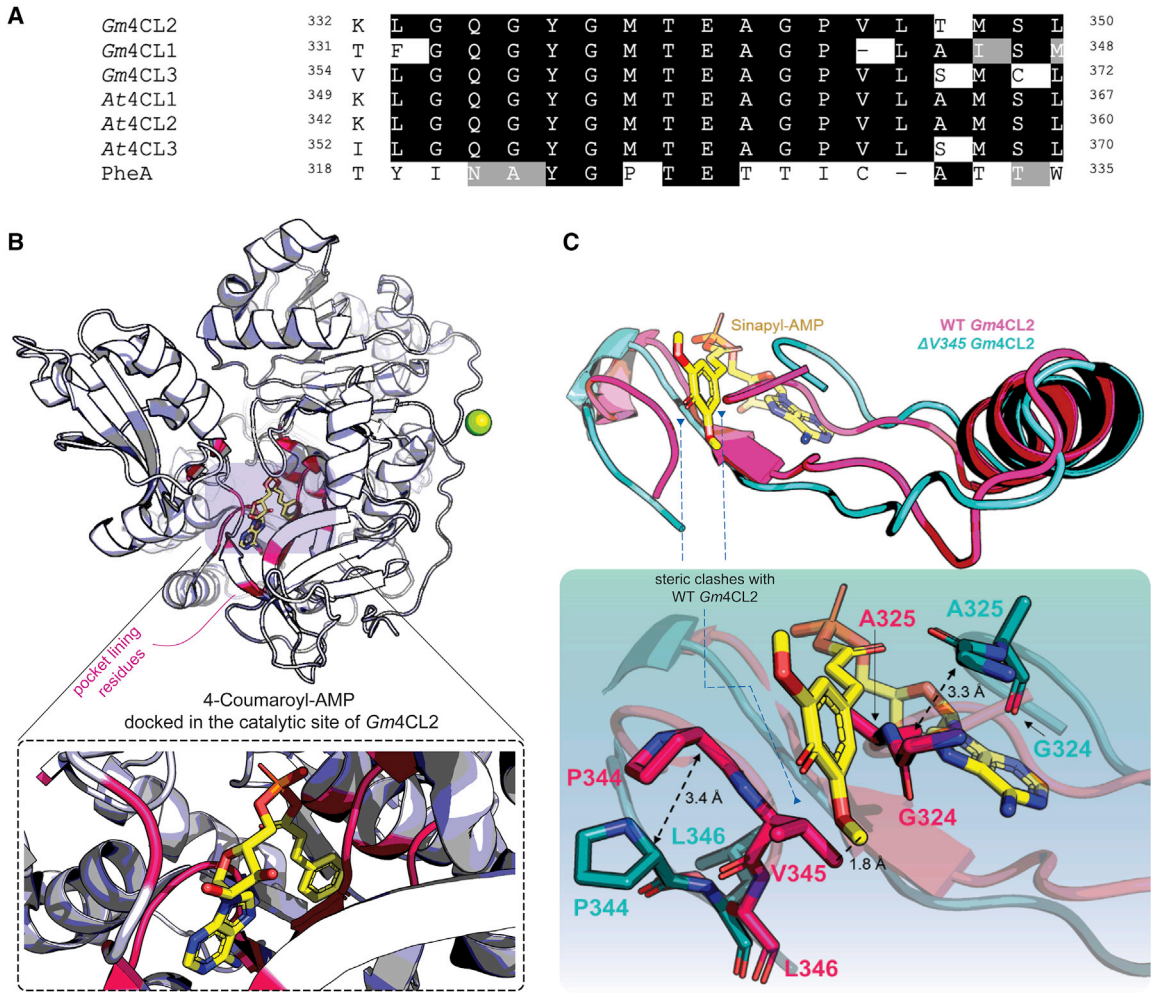


Figure 6. IPRO+/– Corroborates the Experimental Finding that Deletion of V345 Switches Cofactor Specificity of 4CLs from Smaller to Larger Coumaroyl-Substrates Like Sinapate and Ferulate

(A) The sequence alignment of the seven 4CLs with specificities spanning small to large cinnamate derivatives reveals 2 possible deletion sites (V345 and L346) and 13 possible substitution positions in *Gm*4CL2. IPRO+/− redesign aims to enhance binding with larger substrates like sinapate and ferulate by accessing combinations of shorter 4CL sequences (such as *Gm*4CL1 and PheA) or similar-length 4CL sequences with combinations of amino acid substitutions (similar to *Gm*4CL3, *At*4CL1, *At*4CL2, and *At*4CL3).

(B) Docked conformation of reaction intermediate 4-coumaroyl-AMP (yellow) in the catalytic groove of wild-type *Gm*4CL2 along with the 15 binding-pocket residues, indicated in pink.

(C) Deletion of V345 in *Gm*4CL2 opens up the substrate-binding groove, thus favorably accommodating a larger sinapyl group, which otherwise clashes with A235 and V345 alike. The binding pocket expands on either side by more than 3 Å (represented by dotted double arrows) upon V345 deletion.

Indel Variants of 4-Coumarate:CoA Ligase with Altered Substrate Specificity

Plant 4CLs have been characterized from a wide range of species and have exhibited different isoform distribution patterns in terms of folded structure, with substrate specificities spanning several ring-substituted cinnamates. Lindermayr et al. (Lindermayr et al., 2002) reported that soybean (*G. max*) has three *Gm*4CL isoforms with a peptide motif that was functionally linked to the turnover of three cinnamate ring substituents (namely sinapate, ferulate, and caffeate, all of which are bulkier than the native substrate 4-coumarate). Two of three *Gm*4CL isoforms (4CL2 and 4CL3) have an extra amino acid at the center of this motif, which, when deleted, enables these isoforms to show enhanced turnover of the aforementioned larger cinnamic substrates along with compromised native substrate activity. An alignment of six 4CL sequences with high sinapate (and ferulate) activity, comprising two *Gm*4CLs, three *At*4CLs (*Arabidopsis thaliana*), and one PheA (phenylalanine-activating subunit of gramicidin S synthase 1 from *Bacillus brevis*), revealed two possible amino acid deletion sites, Val345 and Leu346, on *Gm*4CL2 (see Figure 6A). Sixteen nonconserved binding-pocket residues on *Gm*4CL2 (with two of them having non-zero deletion probabilities) were chosen as DPs (see Figure 6A). Val285, Lys332, Leu333, Gly334, Gln335, Gly336, Met339, Ala342, Gly343, Pro344, Val345, Leu346, Thr347, Met348, Ser349, and Leu350 constituted the set of DPs. The deletion probabilities at

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

Table 2. Amino Acid Substitutions and Indels from the IPRO+/– Predicted *Gm*4CL2 Variants with Corresponding Improvement or Reduction Over Wild-Type Binding Affinities with Reaction Intermediates of Decreasing Sizes (from Sinapyl-AMP to Cinnamic-AMP)

		CHARMM Interaction Energy Score[a] Reduction (with respect to WT *Gm*4CL2) Between Enzyme Variants and Substrate-AMP Intermediates of Varying Size (= $Score^{WT\ intermediate} - Score^{Variant\ intermediate}$)				
Variant ID	Amino Acid Transitions in *Gm*4CL2	Sinapate $C_{11}H_{12}O_5$ (CHARMM Energy Units)	Ferulate $C_{10}H_{10}O4$ (CHARMM Energy Units)	Caffeate $C_9H_8O_4$ (CHARMM Energy Units)	Coumarate $C_9H_7O_3$ (CHARMM Energy Units)	Cinnamate $C_9H_8O_2$ (CHARMM Energy Units)
WT *Gm*4CL2	None	0	0	0	0	0
R2.D1	K332T, G334V, V345_	87.33 ↑	12.41 ↑	−10.27 ↓	−1.91 ↓	−24.42 ↓
R2.D2	L346_, T347I, M348S, S349T	86.56 ↑	46.00 ↑	27.39 ↑	−4.36 ↓	−49.22 ↓
R2.D3	V285I, L346G	83.63 ↑	13.36 ↑	−14.42 ↓	−49.10 ↓	−11.70 ↓
R2.D4	V345A, T347I	73.99 ↑	30.74 ↑	0.93 ↑	−17.91 ↓	−12.41 ↓
R2.D5	L333F, V345_, L346_	71.04 ↑	52.77 ↑	−5.72 ↓	−39.18 ↓	−63.95 ↓
R2.D6	K332R, V285I, V345G	70.34 ↑	17.28 ↑	16.31 ↑	8.68 ↑	−39.94 ↓
R2.D7	K332R, V345_	68.30 ↑	10.57 ↑	8.00 ↑	−20.31 ↓	−17.60 ↓
R2.D8	V345_, T347A	67.77 ↑	40.72 ↑	36.47 ↑	−21.87 ↓	−57.59 ↓
R2.D9	V285I, K332R, V345_, L346A	64.26 ↑	4.51 ↑	13.71 ↑	4.15 ↑	−3.43 ↓
R2.D10	V345_, M348I	61.90 ↑	15.91 ↑	16.47 ↑	−24.30 ↓	−62.75 ↓
R2.D11	V345C, L346_, T347G	58.88 ↑	30.61 ↑	−11.42 ↓	−13.92 ↓	−70.61 ↓
R2.D12	G336A, V345_	56.67 ↑	22.31 ↑	19.61 ↑	−35.47 ↓	−72.22 ↓
R2.D13	G336A, A342T	52.20 ↑	30.31 ↑	−4.51 ↓	−8.21 ↓	−73.04 ↓
R2.D14	G334A, Q335K, L346G, T347I	51.99 ↑	13.08 ↑	9.44 ↑	−21.15 ↓	−19.06 ↓
R2.D15	L333Y, V345C	47.63 ↑	27.00 ↑	4.32 ↑	−20.06 ↓	−64.62 ↓
R2.D16	P334G, G343S	45.37 ↑	35.76 ↑	21.51 ↑	−28.93 ↓	−25.89 ↓
R2.D17	K332R, L346G, T347G	43.07 ↑	29.66 ↑	−22.38 ↓	−17.43 ↓	−77.61 ↓
R2.D18	K332S, L333G, G334F	38.33 ↑	24.38 ↑	14.53 ↑	−28.75 ↓	−69.40 ↓
R2.D19	T347A, K332T	35.70 ↑	11.28 ↑	9.45 ↑	−14.32 ↓	−23.52 ↓
R2.D20	V285L, K332T, G336A	33.37 ↑	35.10 ↑	23.00 ↑	−10.66 ↓	−69.56 ↓
R2.D21	L346A, M348S, S349G	31.48 ↑	34.86 ↑	1.92 ↑	−19.30 ↓	−28.40 ↓
R2.D22	K332S, L333R, G334A	27.28 ↑	21.12 ↑	9.56 ↑	−39.18 ↓	−62.41 ↓
R2.D23	V285L, K332T, G334I	25.34 ↑	21.11 ↑	14.29 ↑	−11.16 ↓	−72.17 ↓

The differences in interaction energy scores of each of the 23 designs in comparison to wild-type (WT) *Gm*4CL2, along with the amino acid substitutions and deletions with five AMP conjugates of cinnamate-like substrates[a] (sinapate, ferulate, caffeate, 4-coumarate, and cinnamate) in decreasing order of size have been listed. The up arrows indicate improvement and down arrows indicate loss of binding to a substrate in a *Gm*4CL2-variant in comparison to WT[2]. All these variants contain at least one amino acid change (deletion or substitution) seen in other naturally occurring 4CL2 homologs.

[a]CHARMM interaction energy score between *Gm*4CL2 (WT) and (1) sinapate = −8.05, (2) ferulate = −19.31, (3) caffeate = −20.21, (4) coumarate = −79.74, and (5) cinnamate = −109.63 CHARMM energy units. ↑ signifies better-than-WT binding affinity, and ↓ signifies less-than-WT affinity for a certain variant. Lower CHARMM-interaction energy score reflects a stronger enzyme-intermediate binding.

positions Val345 and Leu346 were computed to be 14.3% (one of seven) for both.

Preparation of Enzyme Structure and Substrate Docking

A homology-modeled structure of *Gm*4CL2 (see Figure 6B) using two luciferases from *Photinus pyralis* (PDB: 1BA3; Franks et al., 1998) and one 4CL from *Nicotiana tabacum* (PDB: 5BST; Li and Nair, 2015) as templates was prepared as described in Lindermayr et al. (Lindermayr et al., 2002). The ATP-mediated reaction mechanism proceeds by forming a 4-coumaroyl-AMP intermediate. *Gm*4CL2 variant non-covalent interaction energy scores with 4-coumaroyl-, cinamyl-, caffyl-, ferulyl-, and sinapyl-AMP intermediates were used as *in silico* proxies for *in vitro* enzyme-substrate affinities. The reaction intermediates were first docked onto wild-type *Gm*4CL2 using the torsion geometry of a 4-coumaroyl-AMP co-crystallized with *Nt*4CL2 (PDB: 5BST) as a guide. Very low overall interatomic RMSD (1.8 Å), even lower (0.2 Å) binding-pocket RMSD, and high (>89%) sequence similarity between *Nt*4CL2 and *Gm*4CL2 ensured that the substrate-binding pose and catalytic distances would also be conserved.

Design Runs

IPRO+/− design runs were set up to identify variants that enhance binding with non-native sinapyl-AMP substrate intermediate and check if this improvement comes at the cost of native 4-coumaryl-AMP binding. From the list of the four aforementioned non-native substrates (sinapate, caffeate, cinnamate, and ferulate), the largest substrate, sinapate, was chosen as the target substrate. The 13 (of 15) DPs, with no deletion

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

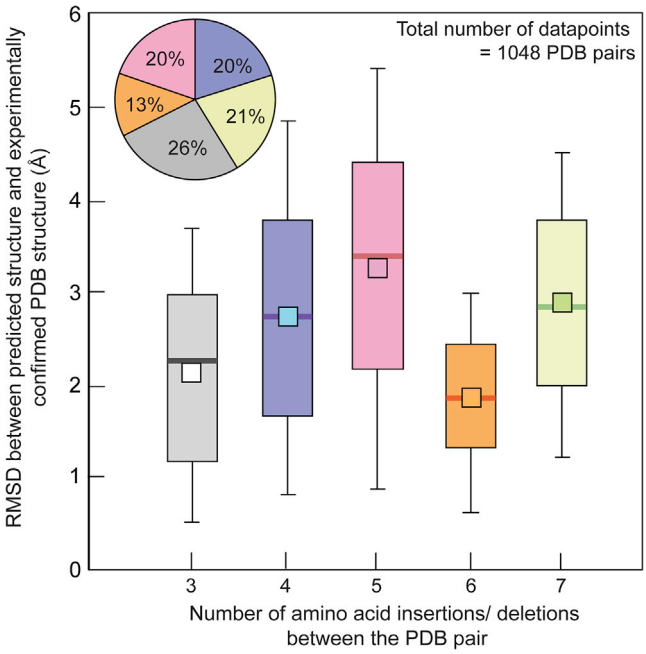


Figure 7. Antibody Variable Domain Pairs Differing by as Many as Seven Amino Acid Indels have been Modeled by Starting with One Any Imposing the Necessary Indels to Obtain the Other

The RMSDs of antibody and indel-variant pairs are indicated as boxplots where each of the five colors represents the number of indels that separate the two sequences that constitute the pair. The length of the box shows the variation in prediction accuracy among each of these five groups with experimentally confirmed structures. The percentage distribution of the 1,048 data points across the five colors has been indicated in the pie chart and the boxplot (see also Figures S9 and S10).

probabilities, were allowed an unconstrained choice of substituent amino acids in order to improve binding to sinapyl-AMP. Overall 298 unique trajectories were sampled and 23 *Gm*4CL2 variants with various combinations of amino acid deletion and substitutions were found (see Table 2). IPRO+/− drove the designs toward (V345 and L346) deletion variants, which opened up more space at the binding pocket, allowing clash-free stabilization of larger substrates. Ten these 23 reflected a site-specific recovery of amino acids as seen in other naturally occurring 4CLs. One of these sequences had both the Val245 and the Leu246 deleted, while at least one deletion was observed in nine sequences. Table 2 shows amino acid substitutions and deletions in each of the 23 successful designs and the corresponding *in silico* interaction energy scores with not only sinapyl-AMP, but also caffeyl-AMP, cinnamyl-AMP, ferulyl-AMP, and native intermediate 4-coumaroyl-AMP.

Lindermayr et al. (Lindermayr et al., 2002) experimentally validated that a *ΔVal345*:*Gm*4GL2 deletion strain was alone sufficient for introducing sinapate turnover, the major distinction between *Gm*4CL1 and *Gm*4CL2. The IPRO+/− interaction energy scores (Table 2) of *Gm*4CL2 variants (such as R2.D1, R2.D5, and so on) containing V345_ show better (−95.38 and −79.09 CHARMM energy units, respectively) sinapyl-AMP stabilization compared with wild-type *Gm*4CL2 (−8.05). The experimentally measured sinapate affinities of *ΔVal345*:*Gm*4GL2 were nearly 250-fold lower (K_m = 1,208 μM) than wild-type *Gm*4GL1 (K_m = 4.7 μM), which, however, could not be captured from simulations. Nevertheless, IPRO+/− was able to make reasonable design decisions on two counts: (1) identify deletion of Val345 and Leu346 as strategy to improve sinapyl-AMP binding and (2) improve sinapyl-AMP binding at the cost of 4-coumaryl-AMP (native substrate) binding (lowered from −79.74 to −57.87 in R2.D7), which is similar to experimental observation (K_m = 42 μM increased to 49 μM for *ΔVal345* + *K332R*:*Gm*4GL2). Structural comparison of R2.D7 variant with wild type revealed that in the wild-type *Gm*4GL2-sinapyl-AMP complex (see Figure 6C), Val345 clashes with one of the methoxy groups of the substrate. Deletion of Val345 opens up the binding pocket by more than 11.2 Å^3 on either side of the methoxy group, thus accommodating the sinapyl moiety and simultaneously rendering the pocket too open for efficient binding of smaller 4-coumaryl-AMP and cinnamyl-AMP (which lack the methoxy groups).

Recapitulating Antibody Variable-Chain Indel Structures

Indels in the complementary determining regions are important for tuning the affinity of broadly neutralizing antibodies targeted against HIV (Kepler et al., 2014). We used the large repository of anti-HIV antibodies to evaluate the efficacy of IPRO+/− to recover experimentally resolved structures containing indels. We extracted 524 antibody variable-chain (heavy/light) pairs where one member differs from the other by no more than seven amino acid indels. Five hundred twenty-four antibody pairs were extracted using Python scripts from the ABG database (Almagro, 2004), which provides a global alignment of the FASTA sequences of heavy and light chains along with their indel variants and corresponding PDB IDs (Almagro, 2004). Indels are not localized, but distributed across the entire length of heavy and light chains.

We applied the Indel-Maker protocol on each of these 524 pairs whereby indel operations were applied on the starting structure to recover a Rosetta energy-minimized structure for the other pair. Subsequently, we computed the RMSD of the Indel-Maker-predicted structure with its experimentally determined coordinates. The same procedure was applied in reverse by starting from the predicted structure and computationally recovering the starting one. Figure 7 illustrates the RMSD between predicted and experimentally resolved structures for all 2 × 524 = 1,048 cases. All Indel-Maker-predicted structures were within 5.5 Å RMSD (median RMSD = 2.46 Å) from the experimentally resolved structures (see Data S1 for details on individual RMSDs and PDB IDs of all variants). One representative case is shown in Figure S9, where three insertions (217L, 218E, and 219C) were imposed on antibody light chain 1MRF.L to obtain the predicted structure of 1NBV.L with RMSD of 2.34Å. Overall, we observe that higher RMSD values mostly arise for sequences where at least one of the indels resides within loops that are longer than 10 amino acids (see Figure S10). Therefore, high RMSD values reflect changes in the structural conformations typically caused by indels.

Indel-Maker for Constructing Enzyme Variants with Desired Indels and Substitutions

We have additionally created a Rosetta-based (open source and freely available from: http://www.maranasgroup.com/

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

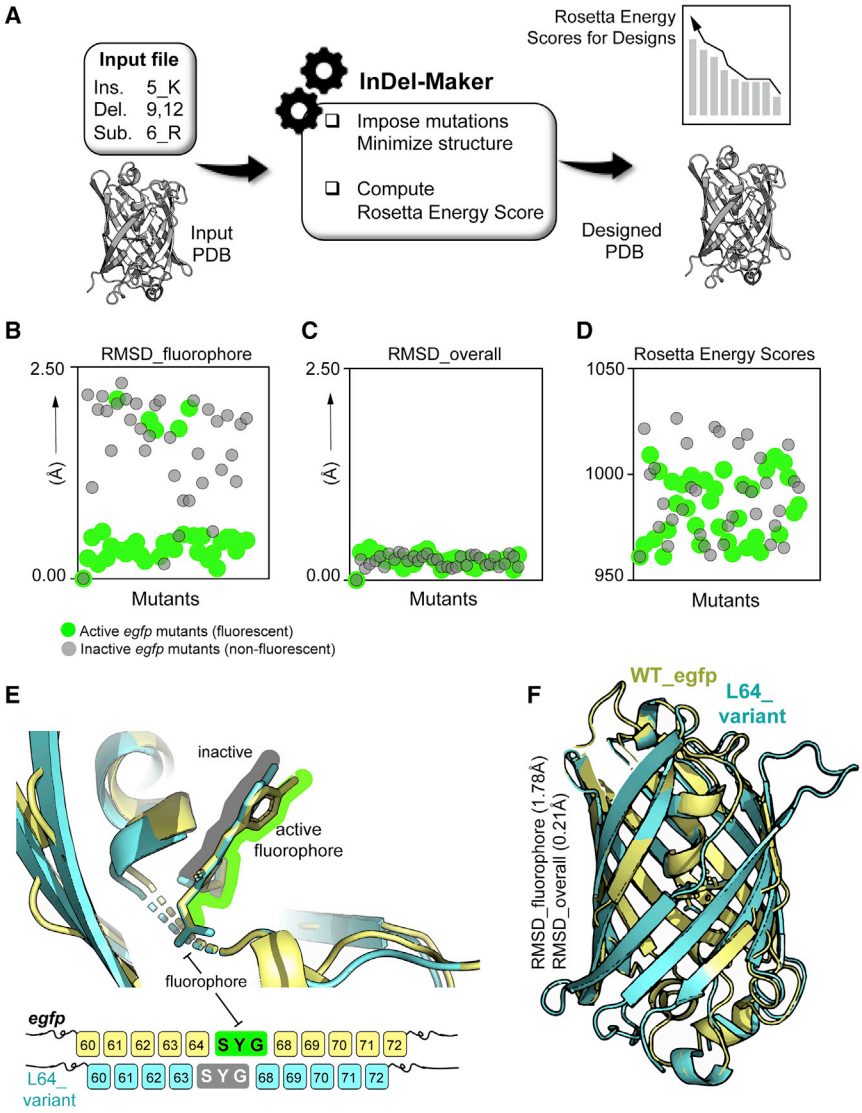


Figure 8. Schematic Overview of Indel-Maker and Results from Test Case Benchmarking on Enhanced Green Fluorescent Protein Variants

(A) Indel-Maker workflow.

(B) Indel-Maker-generated enhanced green fluorescent protein (egfp) variants reveal more than 2 Å deviation of the fluorophore region.

(C) The overall structural RMSD (including side chains) of both the active and the inactive variants was less than 1 Å.

(D) The Rosetta energy scores of all the variants ranged between 950 and 1,025 Rosetta energy units.

(E) Deviation of the SYG fluorophore and the neighboring region in inactive L64_ variant by 2.31 Å.

(F) Overall structural RMSD between wild-type (WT) egfp (yellow) and L64_ variant (cyan) is 0.24 Å.

software.htm) Indel-Maker tool to enable users to construct user-defined variant libraries containing combinations of amino acid insertions, deletions, and substitutions. This would be instrumental in discerning biophysical cues in experimentally tested variants to explain altered substrate/cofactor affinities or mutant stabilities. The workflow (see Figure 8A) requires users to provide an input file specifying insertions, deletions, and modifications to be performed on the input PDB. The required modifications are performed one by one, each followed by the loop closure (in the case of indels) and structure minimization using Rosetta's relax protocol (using the Rosetta all-atom force field). The resulting PDB and its corresponding Rosetta energy score are output at the end of each modification.

In order to benchmark Indel-Maker, we constructed 35 fluorescent and 35 non-fluorescent mutants of enhanced green fluorescent protein (from Arpino et al., 2014). Indel-Maker predicted that overall structural RMSD was less than 1 Å for both sets, but the inactive variants showed more than 2 Å RMSD of the fluorophore region, thus providing structural insights into the inactivity of non-fluorescent variants (see Figures 8B–8F).

DISCUSSION

Structural parts of proteins often are more conserved in evolution compared with sequences ("Mapping the protein universe," Holm and Sander, 1996). This has been exploited using protein design approaches to design non-natural sequences that enhance thermal stabilities/binding affinities of proteins. On the other hand, protein sequences that adopt completely non-natural folds have also been designed (Baker and co-workers: Kuhlman et al., 2003; Richter et al., 2011). Herein we presented a scheme to increase the scope of targeted protein redesign by introduction of indels along with substitutions leading to sampling of a larger sequence space. As demonstrated, integrating indels with protein design can enable the design of versatile protein libraries.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - DNA Synthesis and 'TesA Mutant Construction
  - Bacterial Culturing and Fatty Acid Production
  - Extraction of Adjustable Loop
- MOVING AND FIXED ANCHORS FOR LOOP CLOSURE
  - Establishing Adjustable Loop Size
  - Integration of Loop Closure into IPRO
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.str.2020.08.003.

ACKNOWLEDGMENTS

The United States National Science Foundation grant CBET-1703274 supports this collaboration. We also acknowledge the Center for Bioenergy Innovation of the United States Department of Energy grant DE-SC0018420 for funding our plant enzyme redesign work.

AUTHOR CONTRIBUTIONS

M.J.G. and C.D.M. conceived the study. M.J.G. wrote initial modules for the software. R.C. wrote the rest of the software, automated it, and performed the test cases and wrote the manuscript. V.S.B. wrote the Indel-Maker module and helped edit the manuscript. Q.Y. performed the experiments on 'TesA and wrote the experimental methods. C.D.M. and B.F.P. provided useful insights and helped edit the manuscript.

DECLARATION OF INTERESTS

The authors declare that there are no financial or non-financial competing interests.

Received: January 9, 2020
Revised: July 1, 2020
Accepted: August 7, 2020
Published: August 27, 2020

REFERENCES

Almagro, J.C. (2004). Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. J. Mol. Recognit. *17*, 132–143.

Arpino, J.A., Reddington, S.C., Halliwell, L.M., Rizkallah, P.J., and Jones, D.D. (2014). Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. Structure *22*, 889–898.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235–242.

Brooks, B.R., Brooks, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., et al. (2009). CHARMM: the biomolecular simulation program. J. Comput. Chem. *30*, 1545–1614.

Bussieck, M., Ferris, M., and Lohmann, T. (2012). Algebraic Modeling Systems: Modeling and Solving Real World Optimization Problems (Springer Berlin: Springer Berlin Heildelberg. Heildelberg).

Canutescu, A.A., and Dunbrack, R.L. (2003). Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci. *12*, 963–972.

Case, D.A., Cheatham, T.E., III, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The amber biomolecular simulation programs. J. Comput. Chem. *26*, 1668–1688.

Chaudhury, S., Lyskov, S., and Gray, J.J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics *26*, 689–691.

Chowdhury, R., Ren, T., Shankla, M., Decker, K., Grisewood, M., Prabhakar, J., Baker, C., Golbeck, J.H., Aksimentiev, A., Kumar, M., et al. (2018a). PoreDesigner for tuning solute selectivity in a robust and highly permeable outer membrane pore. Nat. Commun. *9*, 3661.

Chowdhury, R., Allan, M.F., and Maranas, C.D. (2018b). OptMAVEn-2.0: de novo design of variable antibody regions against targeted antigen epitopes. Antibodies *7*, 23.

Chowdhury, R., and Maranas, C.D. (2019). From directed evolution to computational enzyme engineering—a review. AIChE J. *66*, e16847.

Franks, N.P., Jenkins, A., Conti, E., Lieb, W.R., and Brick, P. (1998). Structural basis for the inhibition of firefly luciferase by a general anesthetic. Biophys. J. *75*, 2205–2211.

Gainza, P., Roberts, K.E., Georgiev, I., Lilien, R.H., Keedy, D.A., Chen, C.Y., Reza, F., Anderson, A.C., Richardson, D.C., Richardson, J.S., and Donald, B.R. (2013). Osprey: protein design with ensembles, flexibility, and provable algorithms. Meth. Enzymol. *523*, 87–107.

Gonzalez, C.E., Roberts, P., and Ostermeier, M. (2019). Fitness effects of single amino acid insertions and deletions in TEM-1 β-lactamase. J. Mol. Biol. *431*, 2320–2330.

Grisewood, M.J., Hernandez Lozada, N.J., Thoden, J.B., Gifford, N.P., Mendez-Perez, D., Schoenberger, H.A., Allan, M.F., Floy, M.E., Lai, R.Y., Holden, H.M., et al. (2017). Computational redesign of acyl-ACP thioesterase with improved selectivity toward medium-chain-length fatty acids. ACS Catal. *7*, 3837–3849.

Hecht, M.H., Das, A., Go, A., Bradley, L.H., and Wei, Y. (2004). De novo proteins from designed combinatorial libraries. Protein Sci. *13*, 1711–1723.

Hernández Lozada, N.J., Lai, R.-Y., Simmons, T.R., Thomas, K.A., Chowdhury, R., Maranas, C.D., and Pfleger, B.F. (2018). Highly active C 8 -Acyl-ACP thioesterase variant isolated by a synthetic selection strategy. ACS Synth. Biol. *7*, 2205–2215.

Holm, L., and Sander, C. (1996). Mapping the protein universe. Science *273*, 595–603.

Jacobs, T.M., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J.F., Szyperski, T., and Kuhlman, B. (2016). Design of structurally distinct proteins using strategies inspired by evolution. Science *352*, 687–690.

Jing, F., Zhao, L., Yandeau-Nelson, M.D., and Nikolau, B.J. (2018). Two distinct domains contribute to the substrate acyl chain length selectivity of plant acyl-ACP thioesterase. Nat. Commun. *9*, 860.

Kaplan, J., and DeGrado, W.F. (2004). De novo design of catalytic proteins. Proc. Natl. Acad. Sci. U S A *101*, 11566–11570.

Kenwright, B. (2013). Inverse kinematics – cyclic coordinate descent (CCD). J. Graphics Tools *16*, 177–217.

Kepler, T.B., Liao, H.X., Alam, S.M., Bhaskarabhatla, R., Zhang, R., Yandava, C., Stewart, S., Anasti, K., Kelsoe, G., Parks, R., et al. (2014). Immunoglobulin gene insertions and deletions in the affinity maturation of HIV-1 broadly reactive neutralizing antibodies. Cell Host Microbe *16*, 304–313.

Khersonsky, O., Röthlisberger, D., Wollacott, A.M., Murphy, P., Dym, O., Albeck, S., Kiss, G., Houk, K.N., Baker, D., and Tawfik, D.S. (2011). Optimization of the in-silico-designed Kemp eliminase KE70 by computational design and directed evolution. J. Mol. Biol. *407*, 391–412.

Kortemme, T., Joachimiak, L.A., Bullock, A.N., Schuler, A.D., Stoddard, B.L., and Baker, D. (2004). Computational redesign of protein-protein interaction specificity. Nat. Struct. Mol. Biol. *11*, 371–379.

Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. Nat. Rev. Mol. Cell. Biol. *20*, 681–697.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003). Design of a Novel globular protein fold with atomic-level accuracy. Science *302*, 1364–1368.

Kumar, S., Singh, S.K., Wang, X., Rup, B., and Gill, D. (2011). Coupling of aggregation and immunogenicity in biotherapeutics: T- and B-cell immune epitopes may contain aggregation-prone regions. Pharm. Res. *28*, 949–961.

Kundert, K., and Kortemme, T. (2019). Computational design of structured loops for new protein functions. Biol. Chem. *400*, 275–288.

Kuroda, D., Shirai, H., Jacobson, M.P., and Nakamura, H. (2012). Computer-aided antibody design. Protein Eng. Des. Sel. *25*, 507–521.

Lapidoth, G., Khersonsky, O., Lipsh, R., Dym, O., Albeck, S., Rogotner, S., and Fleishman, S.J. (2018). Highly active enzymes by automated combinatorial backbone assembly and sequence design. Nat. Commun. *9*, 2780–2789.

Lapidoth, G.D., Baran, D., Pszolla, G.M., Norn, C., Alon, A., Tyka, M.D., and Fleishman, S.J. (2015). AbDesign: an algorithm for combinatorial backbone design guided by natural conformations and sequences. Proteins *83*, 1385–1406.

Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P.D., Smith, C.A., Sheffler, W., et al. (2011). Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. Meth Enzymol. *487*, 545–574.

Lennen, R.M., Braden, D.J., West, R.M., Dumesic, J.A., and Pfleger, B.F. (2010). A process for microbial hydrocarbon synthesis: overproduction of fatty acids in Escherichia coli and catalytic conversion to alkanes. Biotechnol. Bioeng. *106*, 193–202.

Li, Z., and Nair, S.K. (2015). Structural basis for specificity and flexibility in a plant 4-coumarate:CoA ligase. Structure *23*, 2032–2042.

Lindermayr, C., Möllers, B., Fliegmann, J., Uhlmann, A., Lottspeich, F., Meimberg, H., and Ebel, J. (2002). Divergent members of a soybean (Glycine max L.) 4-coumarate:coenzyme A ligase gene family. Eur. J. Biochem. *269*, 1304–1315.

Lippow, S.M., and Tidor, B. (2007). Progress in computational protein design. Curr. Opin. Biotechnol. *18*, 305–311.

Lippow, S.M., Wittrup, K.D., and Tidor, B. (2007). Computational design of antibody-affinity improvement beyond in vivo maturation. Nat. Biotechnol. *25*, 1171–1176.

Martín, A., Barrientos, A., and del Cerro, J. (2018). The natural-CCD algorithm, a Novel method to solve the inverse kinematics of hyper-redundant and soft robots. Soft Robot. *5*, 242–257.

Ness, S., Martin, R., Kindler, A.M., Paetzel, M., Gold, M., Jensen, S.E., Jones, J.B., and Strynadka, N.C. (2000). Structure-based design guides the improved efficacy of deacylation transition state analogue inhibitors of TEM-1 β-lactamase. Biochemistry *39*, 5312–5321.

Netzer, R., Listov, D., Lipsh, R., Dym, O., Albeck, S., Knop, O., Kleanthous, C., and Fleishman, S.J. (2018). Ultrahigh specificity in a network of computationally designed protein-interaction pairs. Nat. Commun. *9*, 5286–5313.

Pandurangan, A.P., Ochoa-Montaño, B., Ascher, D.B., and Blundell, T.L. (2017). SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Res. *45*, W229–W235.

Pantazes, R.J., Grisewood, M.J., Li, T., Gifford, N.P., and Maranas, C.D. (2015). The iterative protein redesign and optimization (IPRO) suite of programs. J. Comput. Chem. *36*, 251–263.

Rackers, J.A., Wang, Z., Lu, C., Laury, M.L., Lagardère, L., Schnieders, M.J., Piquemal, Jean-Philip, Ren, P., and Ponder, J.W. (2018). Tinker 8: software tools for molecular design. J. Chem. Theor. Comput.

Rämisch, S., Weininger, U., Martinsson, J., Akke, M., and André, I. (2014). Computational design of a leucine-rich repeat protein with a predefined geometry. Proc. Natl. Acad. Sci. U S A *111*, 17875–17880.

Richter, F., Leaver-Fay, A., Khare, S.D., Bjelic, S., and Baker, D. (2011). De novo enzyme design using Rosetta3. PLoS One *6*, e19230.

Romero, P.A., and Arnold, F.H. (2009). Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. *10*, 866–876.

Spoel, D. Van Der, Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., and Berendsen, H.J.C. (2005). GROMACS: fast, flexible, and free. J. Comput. Chem. *26*, 1701–1718.

Wu, S., and Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. *35*, 3375–3382.

Xiong, P., Wang, M., Zhou, X., Zhang, T., Zhang, J., Chen, Q., and Liu, H. (2014). Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. Nat. Commun. *5*, 5330.

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
Escherichia coli strain RL08ara (K-12 MG1655 Δ*fadD* Δ*araBAD* Δ*araFGH* Φ(Δ*araEp*P_{cp18}–*araE*)	Pfleger Lab (Lennen et al., 2010)	N/A
Chemicals, Peptides, and Recombinant Proteins		
Chloroform	Fisher Scientific	Cat#C298
Hydrogen chloride–methanol solution	Sigma-Aldrich	Cat#17935
Lysogeny broth (LB) medium	Fisher Scientific	Cat#BP1426
Pentadecanoic acid	Sigma-Aldrich	Cat#P6125
Nonanoic acid	Sigma-Aldrich	Cat#W278408
Pentanoic acid	Sigma-Aldrich	Cat#8008210100
Sodium bicarbonate	Fisher Scientific	Cat#BP328
Glycerol	Fisher Scientific	Cat#AC15892-0025
L-arabinose	Acros Organics	Cat#AC365181000
Methanol	Sigma-Aldrich	Cat#17935
Recombinant DNA		
Plasmid: pBAD18-P_{araBAD}-'TesA-Native	•Pfleger Lab (Grisewood et al., 2017)	N/A
Plasmid: pBAD18- P_{araBAD}-'TesA-Mutation	This study, see Figure S2–S7	N/A
Deposited Data		
Photinus pyralis Luciferase	(Franks et al., 1998)	PDB: 1BA3
Nicotina tabacum 4-coumarate:CoA ligase	(Li and Nair, 2015)	PDB: 5BST
Escherichia coli TEM-1 Beta Lactamase	(Ness et al., 2000)	PDB: 1ERM
Escherichia coli eGFP	(Arpino et al., 2014)	PDB: 4KA9
Software and Algorithms		
PyRosetta	(Chaudhury, 2010)	pyrosetta.org
IPRO +/−	This study	http://www.maranasgroup.com/software.htm
IndelMaker	This study	http://www.maranasgroup.com/software.htm

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to Ratul Chowdhury (rchowdhury024@gmail.com).

Materials Availability

This section of study did not generate new unique reagents.

Data and Code Availability

Original data for the Figure 4 in the paper is available as Data S1. Both IPRO +/- package and IndelMaker package are freely available at http://www.maranasgroup.com/software.htm

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Thioesterase activity was tested *in vivo* by expressing each 'TesA variant in *Escherichia coli* strain RL08ara(Lennen et al., 2010) (K-12 MG1655 Δ*fadD* Δ*araBAD* Δ*araFGH* Φ(Δ*araEp*P_{cp18}–*araE*) and quantifying the resulting free fatty acid profile. In these experiments, cultures were inoculated from cells (1:100 dilution) grown overnight in LB containing 100 mg/L carbenicillin. Cells were grown at 37°C

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

and 250 rpm shaking in 250 mL baffled shake flask containing 25 mL LB, 100 mg/L carbenicillin and 4 g/L glycerol. When OD_{600} reached 0.2, 2 g/L L-arabinose was added to induce enzyme production and cultures were incubated for an additional 24 h to produce free fatty acids.

METHOD DETAILS

DNA Synthesis and 'TesA Mutant Construction

All enzyme variants were created from a pBAD18 vector harboring the native 'TesA coding sequence linked to the P_{araBAD} promoter. 'TesA variants were constructed by Gibson assembly of DNA fragments containing the desired mutations. Each variant sequence was verified by Sanger sequencing. The fatty acid titers (mM) using each variant and the corresponding plasmid maps of each 'TesA variant are appended as Table S1 and Figure S2–S7, respectively.

Bacterial Culturing and Fatty Acid Production

Enzyme activity was tested *in vivo* by expressing each 'TesA variant in *Escherichia coli* strain RL08ara(Lennen et al., 2010) (K-12 MG1655 Δ*fadD* Δ*araBAD* Δ*araFGH* Φ(Δ*araEp*P_{cp18}–*araE*) and quantifying the resulting free fatty acid profile. In these experiments, cultures were inoculated from cells (1:100 dilution) grown overnight in LB containing 100 mg/L carbenicillin. Cells were grown at 37°C and 250 rpm shaking in 250 mL baffled shake flask containing 25 mL LB, 100 mg/L carbenicillin and 4 g/L glycerol. When OD_{600} reached 0.2, 2 g/L L-arabinose was added to induce enzyme production and cultures were incubated for an additional 24 h to produce free fatty acids. Each variant was cultured in triplicate and error bars were calculated using standard deviation.

The fatty acid composition of each culture was quantified using protocols described previously(Grisewood et al., 2017; Hernández Lozada et al., 2018). In brief, fatty acids were extracted from 2.5 mL culture samples (~5 OD-mL of cells) using chloroform/methanol (v/v, 1:1). Extracts were converted to methyl esters by incubation with 1.25 M HCl in methanol for 12 h at 50°C. A final concentration of 250 mg/L nonanoic acid and 25 mg/L pentadecanoic acid was used as internal standards to quantify C_8 FFA and C_{10}-C_{18} FFA, respectively. Fatty acid methyl esters were analyzed by GC-FID (Shimadzu, Model GC-2010) equipped with an AOC-20i autoinjector and a 30 m, 0.25 mm ID RTX-5 column. Fatty acid concentrations were inferred from a standard curve created from purchased standards.

Extraction of Adjustable Loop

During the loop closure algorithm, a design position is randomly selected and, based on a subsequent probabilistic decision, will undergo a substitution (aa→aa) or deletion (aa→_) if the starting sequence has an amino acid, or undergo an insertion (_→aa) if the starting sequence has a gap. Next, the loop (termed the adjustable loop) that includes the design position will be extracted and subjected to the adapted cyclic coordinate descent method. Five residues are selected on either side (N-terminal or C-terminal) of the design position, assuming that this position does not reside in an alpha-helix or beta-sheet and does not extend beyond the protein's terminus. The size of the adjustable loop is therefore typically eleven residues, which was established by the superimposition of homologous protein fragments that differ only by the inclusion of a gap in the sequence alignment (see Establishing Adjustable Loop Size *below*). The extracted adjustable loop is next subjected to the adapted cyclic coordinate descent method. The original implementation of the cyclic coordinate descent method identified the optimal rotation of the backbone dihedral that would superimpose backbone atoms from the C-terminus of the hypothetical new conformation (i.e., the 'moving' anchor) of the loop with the real coordinates (i.e., the 'fixed' anchor, which are the atoms' coordinates prior to *re-stitching*) of these atoms. The method described here includes a bidirectional approach (adapted cyclic coordinate descent) to stitching the protein's backbone together. In this way, backbone conformational changes are centered with respect to a design position rather than biasing the modifications to the N-terminal side of the design position. The adapted cyclic coordinate descent method also adds an energy calculation following backbone dihedral angle rotations to ensure effective re-stitching the protein's backbone does not come at the expense of assuming unfavorable loop conformations.

MOVING AND FIXED ANCHORS FOR LOOP CLOSURE

The selection of the appropriate 'moving' and 'fixed' anchors is a function of both the type of backbone modification (i.e., insertion or deletion) and the position of the rotatable bond (i.e., N-terminal or C-terminal) relative to the design position. In the event of an insertion, a fictitious glycine residue (an N-Cα-C triplet) is generated by randomly selecting a set of random psi and psi dihedral angles from a cumulative probability distribution based on the Ramachandran plot. This N-Cα-C triplet is appended to the C-terminal side of the design position *j* as well as the N-terminal side of the *j*+1 residue. In the event of an insertion, these glycine residues alternate serving as the 'moving' or 'fixed' anchors. If the rotatable bond is N-terminal to the design position, then the more N-terminal glycine serves as a 'moving' anchor and the more C-terminal glycine serves as the 'fixed' anchor. On the contrary, if the rotatable bond is C-terminal to the design position, then the more N-terminal alanine serves as the 'fixed'.

Each iteration of the adapted cyclic coordinate descent method begins at the most N-terminal residue of the extracted loop and proceeds – one residue at a time – towards the designated design position. Once all residues on the N-terminal end of the loop have been subjected to the adapted cyclic coordinate descent method, the method restarts at the most C-terminal residue of the extracted loop and proceeds towards the design position. A single adapted cyclic coordinate descent iteration consists of rotations on both the

Please cite this article in press as: Chowdhury et al., IPRO+/−: Computational Protein Design Tool Allowing for Insertions and Deletions, Structure (2020), https://doi.org/10.1016/j.str.2020.08.003

N-terminal and C-terminal sides of the design position. During this process, each residue has its phi and psi dihedral angles optimally rotated such that the 'moving anchor' is optimally superimposed with the 'fixed anchor'. The 'moving anchor' are fictitious atoms that should be as close to the 'fixed anchor', which are the corresponding real atoms, as possible. After rotating both the phi and psi dihedral angles for each residue, the energy of the loop is calculated and compared against the energy of the loop prior to the two rotations. If the energy is more favorable, the move is accepted, and if the energy is unfavorable, the move is accepted using simulated annealing. The initial temperature during simulated annealing is 1000K, and linearly decreases by 10% after each iteration until a final temperature of 50K is reached. This cooling schedule permits more unfavorable conformations during early iterations and is less permissive during later iterations of the adapted cyclic coordinate decent method. This process is repeated until the convergence criteria is met.

The adapted cyclic coordinate descent method is executed repeatedly until the RMSD between the moving and fixed anchors does not change appreciably over many iterations. Specifically, the default parameters for the algorithm dictate that the algorithm has terminated if the following is satisfied:

$$\frac{1}{20} \times \sum_{i=1}^{20} \left| RMSD_{N-i+1} \right| \leq \varepsilon, \quad \forall N \geq 20$$

In this term, i represents a particular completed iteration of the adapted cyclic coordinate descent method, $RMSD_i$ represents the RMSD between the moving anchor and fixed anchor for iteration i, N represents the most recent completed iteration i of the adapted cyclic coordinate descent method, and the convergence tolerance is $\varepsilon = 10^{-3}$. As the optimal angle of rotation for the dihedral angles is deterministic, it remains plausible that the loop may become trapped in a local minimum despite the inclusion of simulated annealing. One example of such a case is the scenario where the first set of rotated residues (i.e., the N-terminus of the loop) is rotated in a manner that the next residue is sterically impeded following the rotation, thereby precluding any successful rotations at the adjacent residue to the loop's N-terminus, irrespective of the total number of adapted cyclic coordinate descent iterations. In such a situation, a reasonably stitched loop may be entirely inaccessible. To alleviate this issue, each of the residue's backbone dihedral angles are randomly perturbed by up to 15°, akin to IPRO's standard Backbone Perturbation[30] step, to escape the local minimum when the final RMSD between the moving and fixed anchors is greater than 0.6 Å. Smaller disparities between the moving and fixed anchors are presumably solvable using a force field energy minimization.

Establishing Adjustable Loop Size

The number of residues that constitute the size of the adjustable loop is a crucial parameter within IPRO+/- loop closure algorithm. If the size of the adjustable loop is too small, then the algorithm will have limited degrees of freedom available due to the smaller number of rotatable dihedral angles and will more likely fail to generate a continuous protein backbone. Alternatively, if the size of the adjustable loop is too large, then an artificially large number of residues will be affected by the presence of an insertion or deletion, compromising the accuracy of the method. We therefore sought to establish a reliable value for this parameter through the extraordinary number of published protein structures (Berman et al., 2000).

The adjustable loop size was determined by first compiling an expansive list of all sequences that were each associated with an experimental structure. Next, protein-protein pairwise sequence alignments were performed for each sequence in the full dataset of 437,094 sequences with the remaining 437,093. Identical amino acid sequences with the exception of 1-5 contiguous gaps in the alignment that consisted of at least 50 residues on either side of the gap were examined. Next, each alignment pair was tested to ensure that (a) all of the backbone atoms for both fragments were included in the structure, (b) the sequences did not originate from the same source, and (c) each structure contained a single protein model for a given sequence. After filtering out these alignment patterns, 3226 protein-protein alignment pairs persisted. For each of these 3226 pairs, the structures were superimposed. The backbone dihedral angles were calculated for each structure, and the RMSD between the two structures was calculated on a per-residue basis. The position of the gap within the fragment structures was treated as position 0. The first residue to the N-terminal side of the gap was treated as position -1, the second residue as -2, and so on. Similarly, the first residue to the C-terminal side of the gap was labeled as position +1, the second as position +2, and so on. This data was compiled and averaged over for each position relative to the gap in the sequence. Analysis of the two datasets (pre-residue backbone RMS data and per-residue RMSD data) showed discrepancies at roughly the same positions, but the per-residue RMSD data was noisier due to propagated errors in the structure superimposition and these results are not discussed further (see Figure S1). The per-residue backbone RMS data was fit to a normal distribution and the standard deviation for the fitted curves ranged from 1.2 residues to 1.7 residues. Therefore, this data suggests that 99.7% of the backbone dihedral difference is accounted for by the five residues ($3\sigma = 5.1$ residues) on either side of the insertion/deletion position. The default adjustable loop size was therefore set to eleven residues- the design position itself in addition to the five residues on either side of the design position.

Integration of Loop Closure into IPRO

The entailed loop was integrated into IPRO to obtain the IPRO+/- cycle. First, a sequence alignment of the starting sequence identifies (a) the design positions on the loop region which can accommodate insertion and deletions (indel-DPs), and (b) the insertion, deletion, and amino acid substitution probabilities for these positions. Based on the overall alignment, the enzyme residues are re-numbered by counting in the alignment gaps. For example, a hypothetical five peptide enzyme with sequence 'AGKLP' if aligned as

'A–G-KL-P' would replace Gly2 with Gly4, Lys3 with Lys5, Pro5 with Pro8 and so on. During an insertion run, a glycine is introduced in the design position and the backbone is stitched using loop closure algorithm. This position has the opportunity to be modified to a different amino acid tailored for the desired biochemical objective, at a later design cycle. In case of a deletion, the backbone and side-chain atoms of the design position (*i*) are deleted and the neighboring residues on either side are joined using loop closure as explained above to attain the lowest energy re-stitched structure. A better-interacting design with lower interaction energy score than the current best is always accepted. Worse designs are accepted with a probability or a Boltzmann factor equal to $e^{\frac{-\Delta(Interaction\ energy)}{kT}}$(using a Metropolis criterion where, *k* is the Boltzmann constant and is $\sim 0.33 \times 10^{-23}$cal/K, and T is temperature in K). A temperature of 3640 K in the Boltzmann factor is used which ensures that there is a 25% probability that a worse design with an interaction energy 10 kcal/mol less negative than the current best will be retained. If retained, this design serves as a starting point for subsequent interactions. Finally, for a substitution run, the loop closure algorithm is side-stepped and only a conventional IPRO run is performed. After an insertion step (or _→aa transition), this position is removed from the list of possible insertion loci and is appended to the list of substitutable loci for the rest of the IPRO+/- sequence design iterations.

QUANTIFICATION AND STATISTICAL ANALYSIS

Fatty acid quantification is described in "Method Details". Each variant was cultured in triplicate and error bars represent standard deviation about the mean recorded value. A two-sample t-test was used to compute the statistical significance (*p-value*) of recovering a certain indel variant using IPRO+/- as opposed to using random sequence design technique.

Structure, Volume *28*

Supplemental Information

IPRO+/−: Computational Protein Design Tool

Allowing for Insertions and Deletions

Ratul Chowdhury, Matthew J. Grisewood, Veda Sheersh Boorla, Qiang Yan, Brian F. Pfleger, and Costas D. Maranas

Supplementary Items

IPRO+/- Computational protein design tool allowing for insertions and deletions

Ratul Chowdhury[1], Matthew J. Grisewood[1], Veda Sheersh Boorla[1], Qiang Yan[2],

Brian F. Pfleger[2], and Costas D. Maranas[1,]*

[1]Department of Chemical Engineering, The Pennsylvania State University, University Park. PA-16802. USA.

[2]Department of Chemical and Biological Engineering, University of Wisconsin−Madison, Wisconsin 53706, USA.

**Corresponding author*

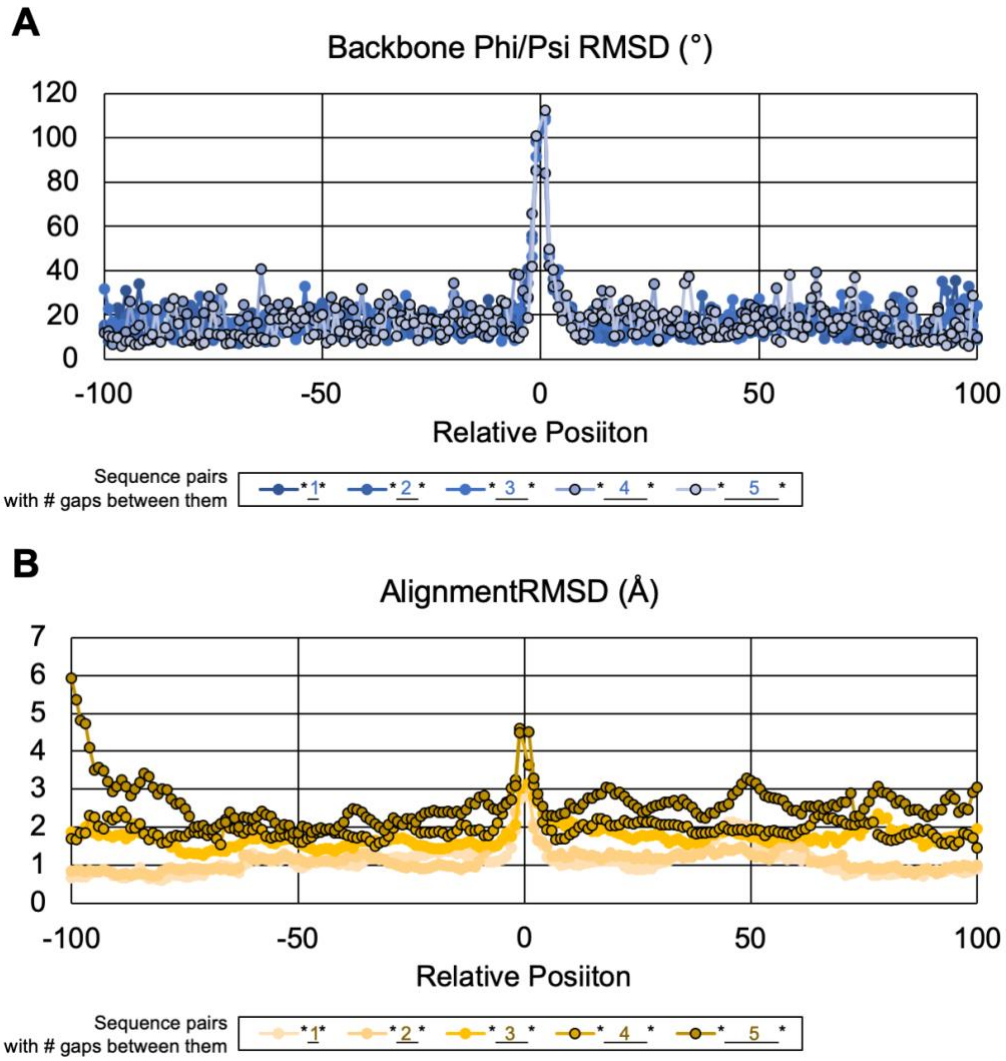
Supplementary Table S1 – Related to STAR Methods 'DNA Synthesis and TesA Mutant Construction'

Free fatty acid titers (mM) in 10 'TesA variants characterized experimentally.

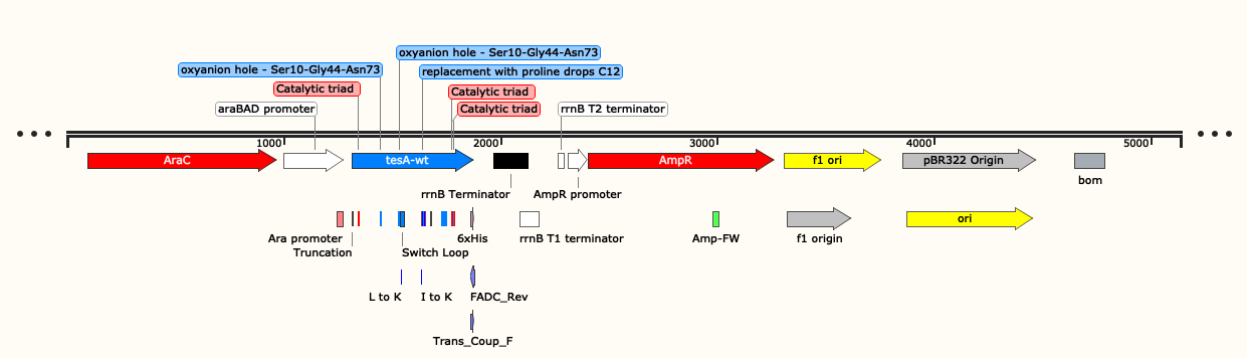
	C_8 FFA (mM)	C_{10} FFA (Mm)	$C_{12:0}$ FFA (mM)	$C_{12:1}$ FFA (mM)	$C_{14:0}$ FFA (mM)	$C_{14:1}$ FFA (mM)	$C_{16:0}$ FFA (mM)	$C_{16:1}$ FFA (mM)	OTHERS (mM)
WT	0.22±0.02	0.12±0.01	0.71±0.04	0.07±0.00	1.43±0.13	0.14±0.01	0.50±0.02	0.08±0.00	0.07±0.00
I210_	0.02±0.00	0.09±0.00	0.06±0.00	0.06±0.00	0.18±0.00	0.14±0.00	0.46±0.00	0.08±0.00	0.08±0.00
P213K	0.16±0.01	0.11±0.01	0.15±0.01	0.06±0.00	0.89±0.08	0.19±0.01	0.91±0.06	0.08±0.00	0.08±0.00
N215_	0.25±0.00	0.12±0.01	0.48±0.01	0.07±0.00	1.99±0.02	0.44±0.01	0.87±0.02	0.09±0.00	0.07±0.00
I210_ P213K	0.02±0.00	0.09±0.00	0.06±0.00	0.05±0.00	0.18±0.01	0.14±0.01	0.49±0.02	0.08±0.01	0.08±0.00
I210_N215_	0.03±0.00	0.09±0.00	0.06±0.00	0.06±0.00	0.18±0.00	0.14±0.00	0.49±0.02	0.08±0.00	0.08±0.01
P213K N215_	0.14±0.00	0.10±0.00	0.14±0.00	0.06±0.00	0.78±0.01	0.19±0.00	0.80±0.00	0.07±0.00	0.07±0.00
I210_ P213K N215_	0.03±0.00	0.09±0.00	0.07±0.00	0.06±0.00	0.19±0.00	0.14±0.00	0.50±0.01	0.09±0.01	0.08±0.00
D268_G270L	0.22±0.03	0.11±0.01	0.69±0.10	0.07±0.00	2.24±0.19	0.33±0.03	0.78±0.04	0.08±0.00	0.06±0.00
I210_P213K D268_	0.02±0.00	0.09±0.01	0.06±0.00	0.06±0.00	0.19±0.00	0.14±0.00	0.49±0.03	0.09±0.01	0.08±0.00
P213K N215_ D268_	0.06±0.00	0.10±0.01	0.10±0.00	0.06±0.00	0.54±0.03	0.16±0.01	0.70±0.04	0.08±0.00	0.08±0.00

Supplementary Figures

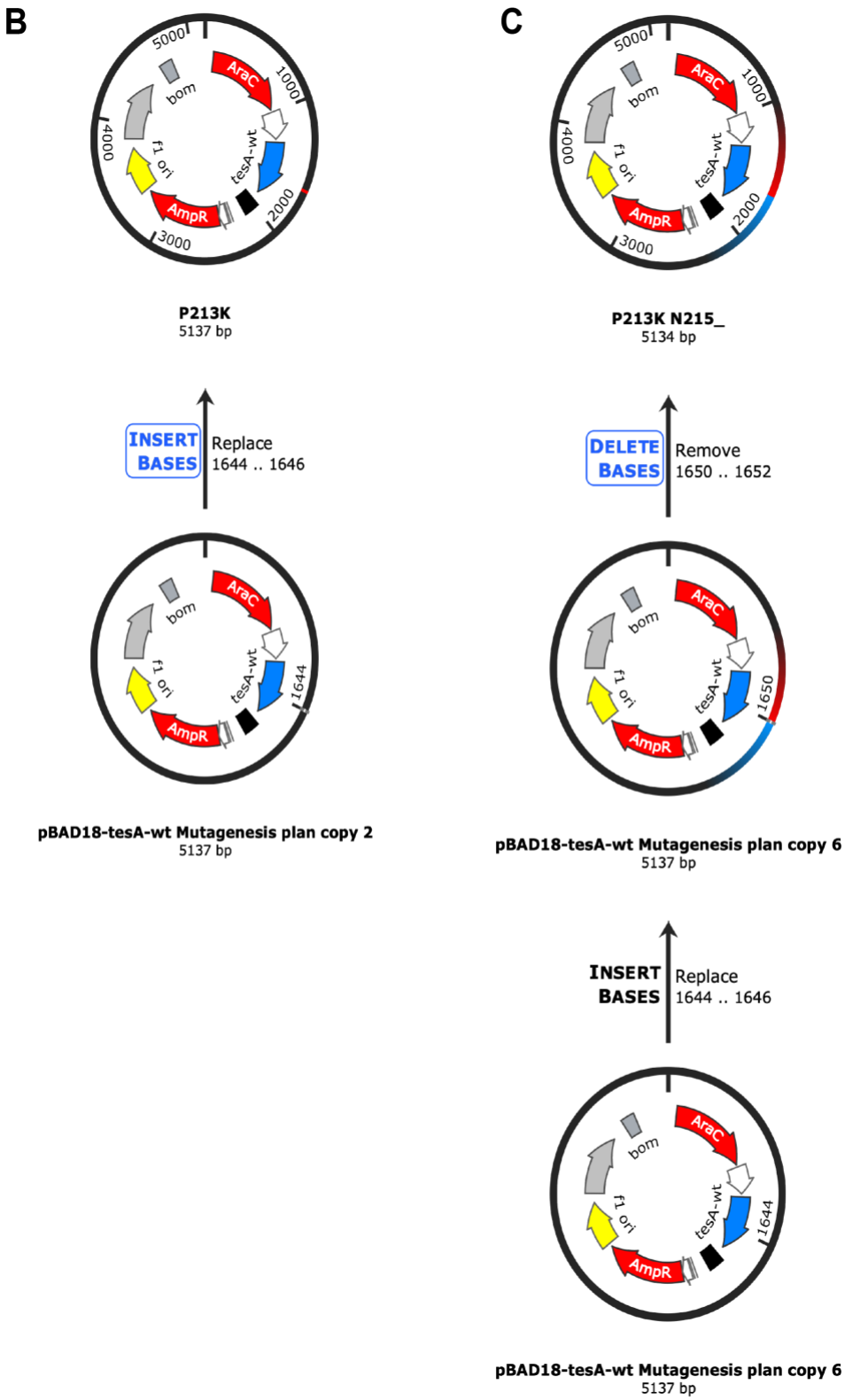
Supplementary Figure S1 – Related to STAR Methods – 'Establishing Adjustable Loop Size'.
RMSD comparison of **(A)** dihedral and **(B)** alignment RMSD between 3226 protein pairs from PDB database with one, two, three, four, and five contiguous aa deletion difference between them reveal that the effect of one (or more) extra residue is felt only up to 5 residues upstream and 5 residues downstream to the point of difference (0_{th} residue). Note: There is some noise in the data with four and five residue differences (see panel B, two darkest yellow lines), but that has not been explored in this article.



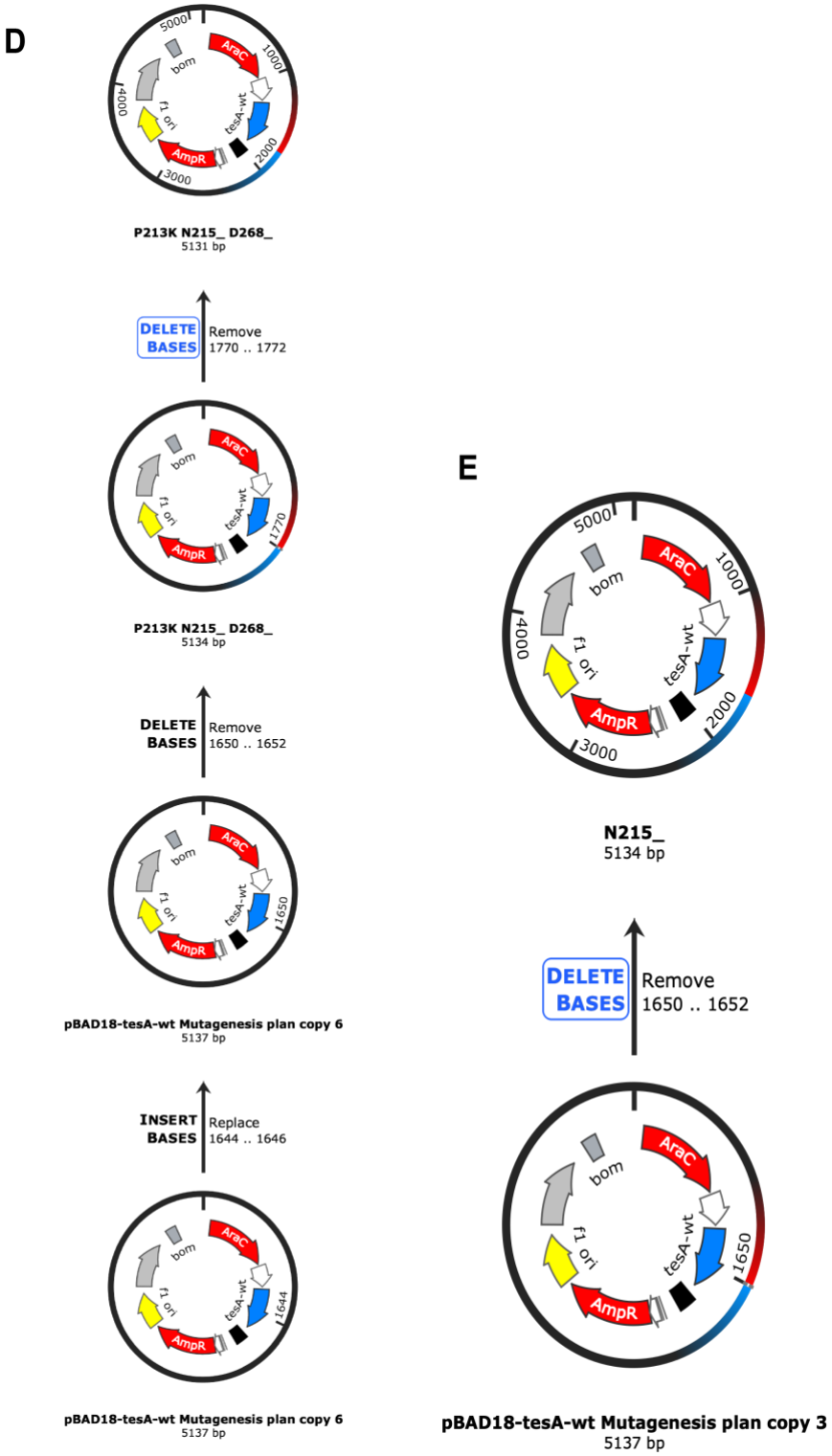
Supplementary Figure S2. Related to STAR Methods 'DNA Synthesis and TesA Mutant Construction'
The plasmid map showing the promoter regions and the nucleotide sequence for 'TesA-wt (indicated in blue). Additional information regarding catalytic triads and Ser-Gly-Asn oxyanion hole are also indicated therein.



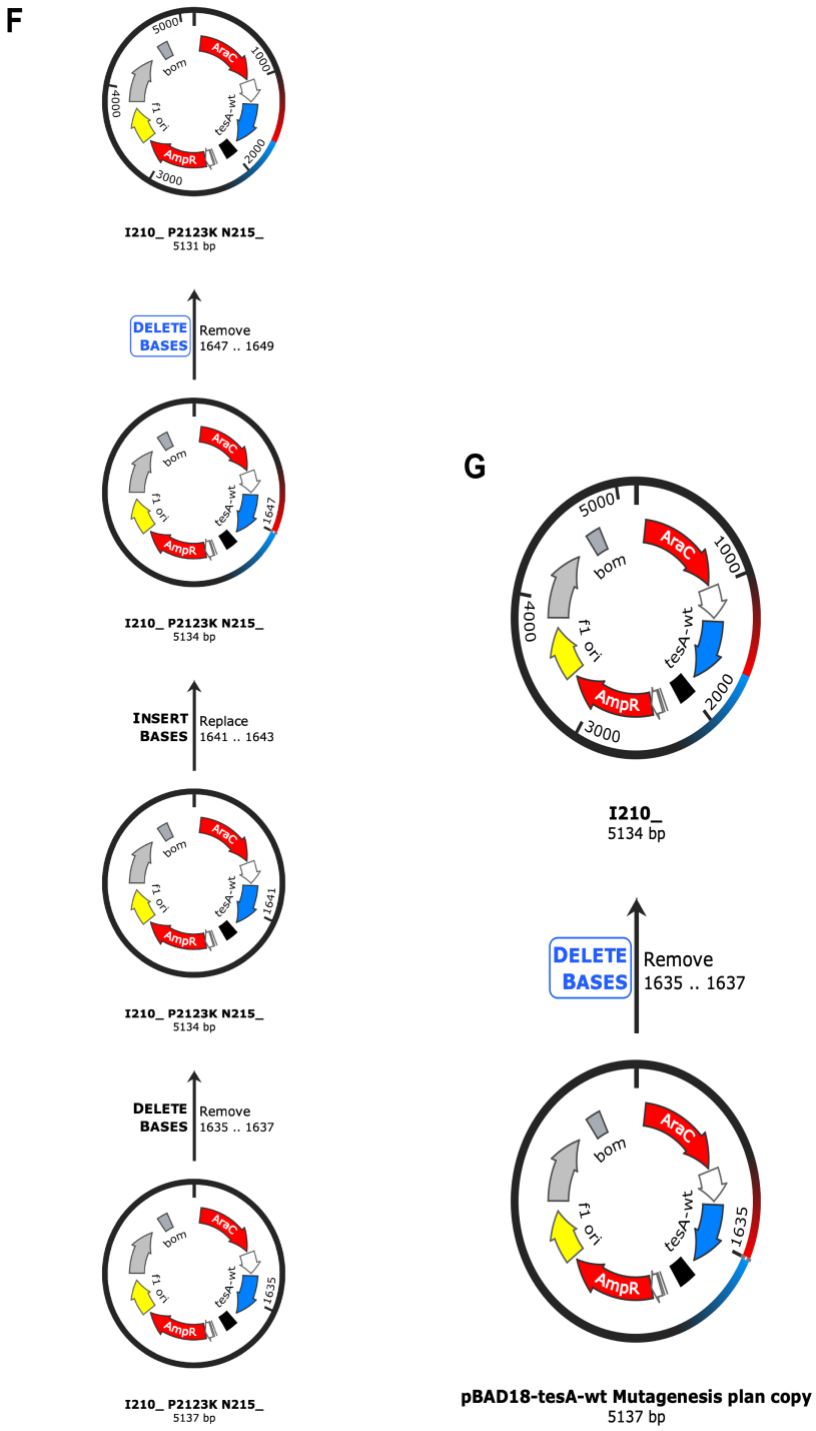
Supplementary Figure S3. Related to STAR Methods 'DNA Synthesis and TesA Mutant Construction' Plasmid map transition from 'TesA-wt to P213K and P213K, N215, respectively, showing the promoter regions and the nucleotide sequence for 'TesA-wt and variant (indicated in blue) is illustrated. Additional information regarding nucleotide changes that effect the mutations have been indicated.



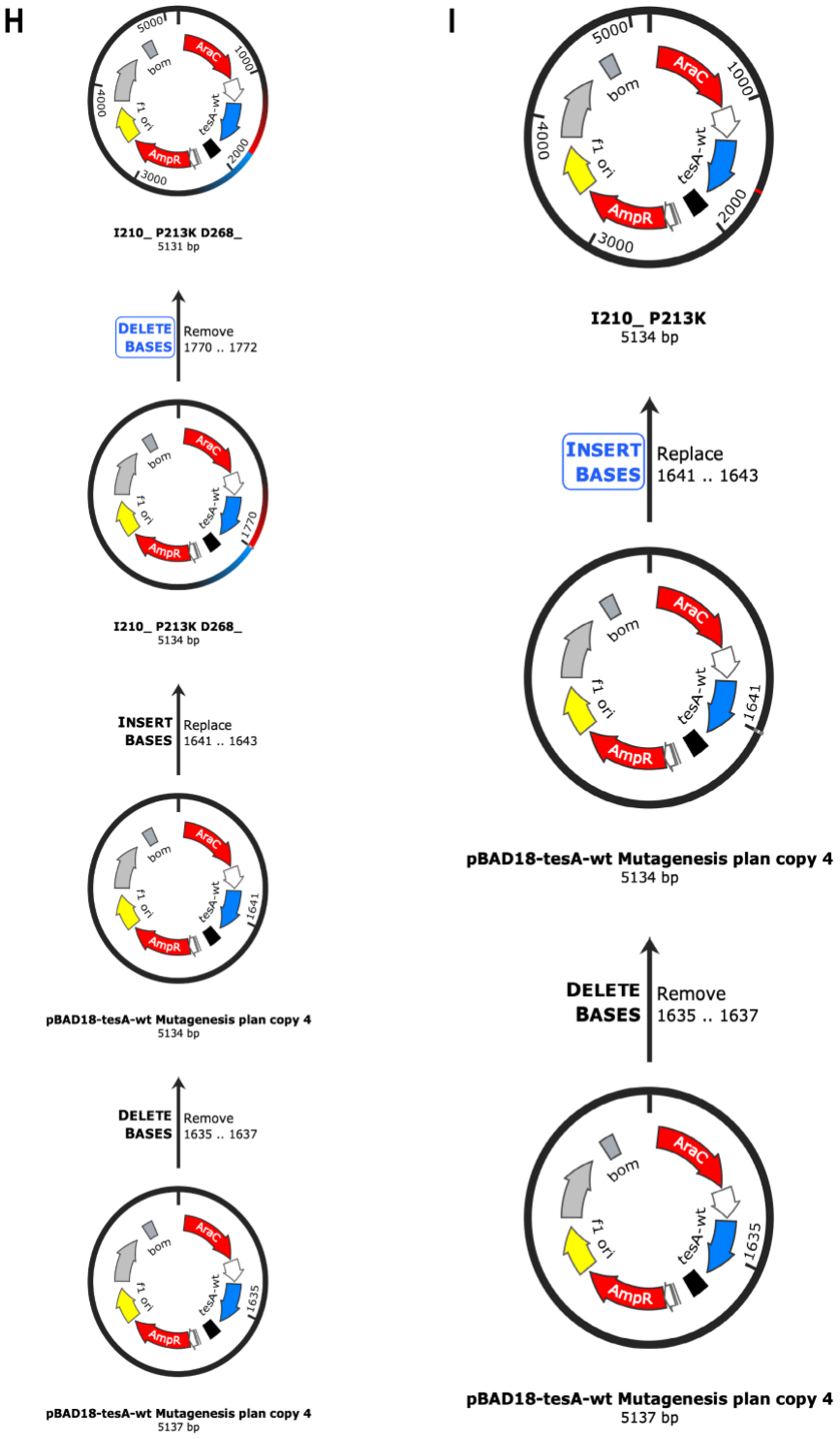
Supplementary Figure S4. Related to STAR Methods 'DNA Synthesis and TesA Mutant Construction' Plasmid map transition from 'TesA-wt to P213K, N215_, D268_ and N215_ respectively, showing the promoter regions and the nucleotide sequence for 'TesA-wt and variant (indicated in blue) is illustrated. Additional information regarding nucleotide changes that effect the mutations have been indicated.



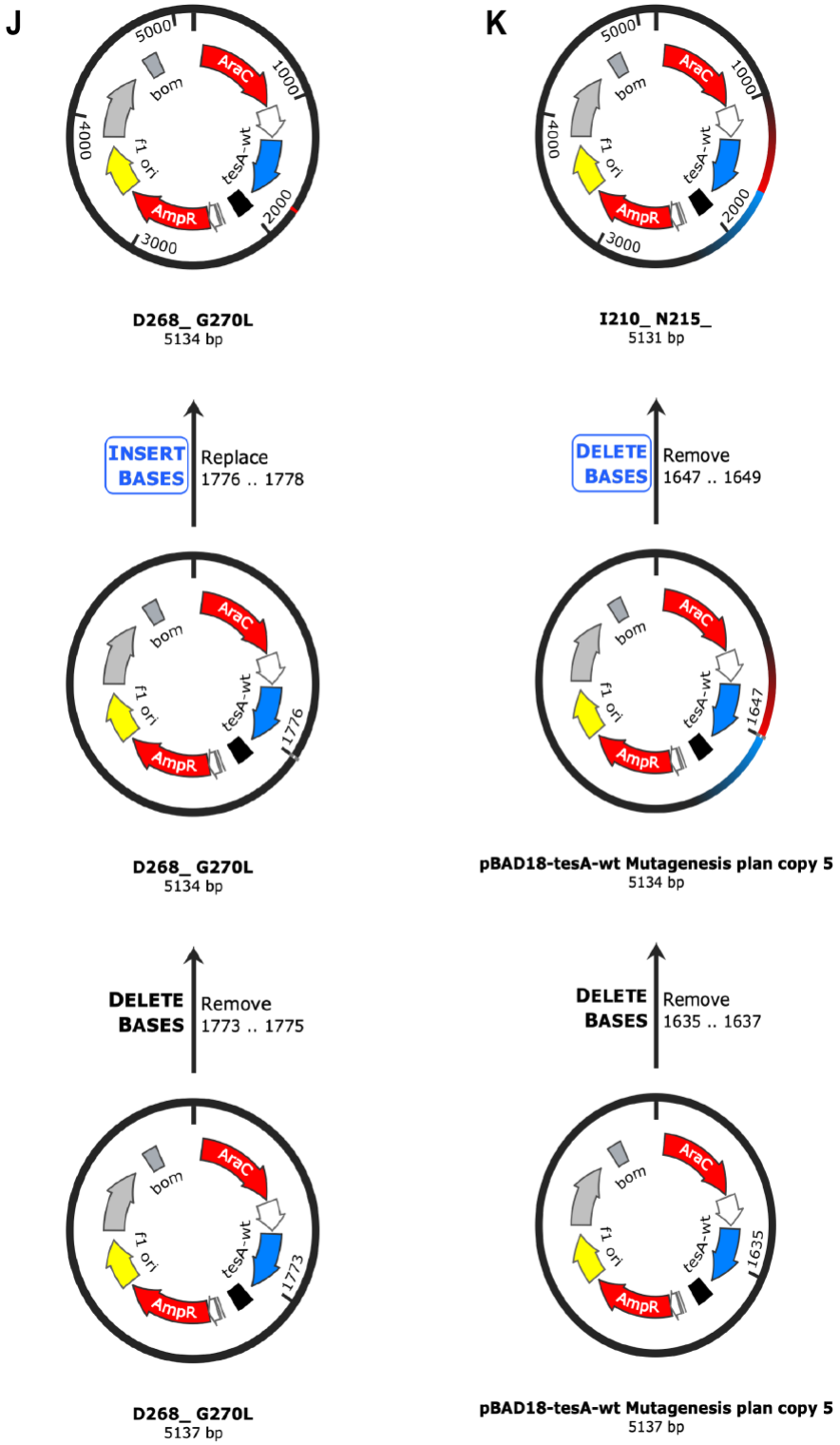
Supplementary Figure S5. Related to STAR Methods 'DNA Synthesis and TesA Mutant Construction' Plasmid map transition from 'TesA-wt to I210_, P213K, N215_ and I210_ respectively, showing the promoter regions and the nucleotide sequence for 'TesA-wt and variant (indicated in blue) is illustrated. Additional information regarding nucleotide changes that effect the mutations have been indicated.



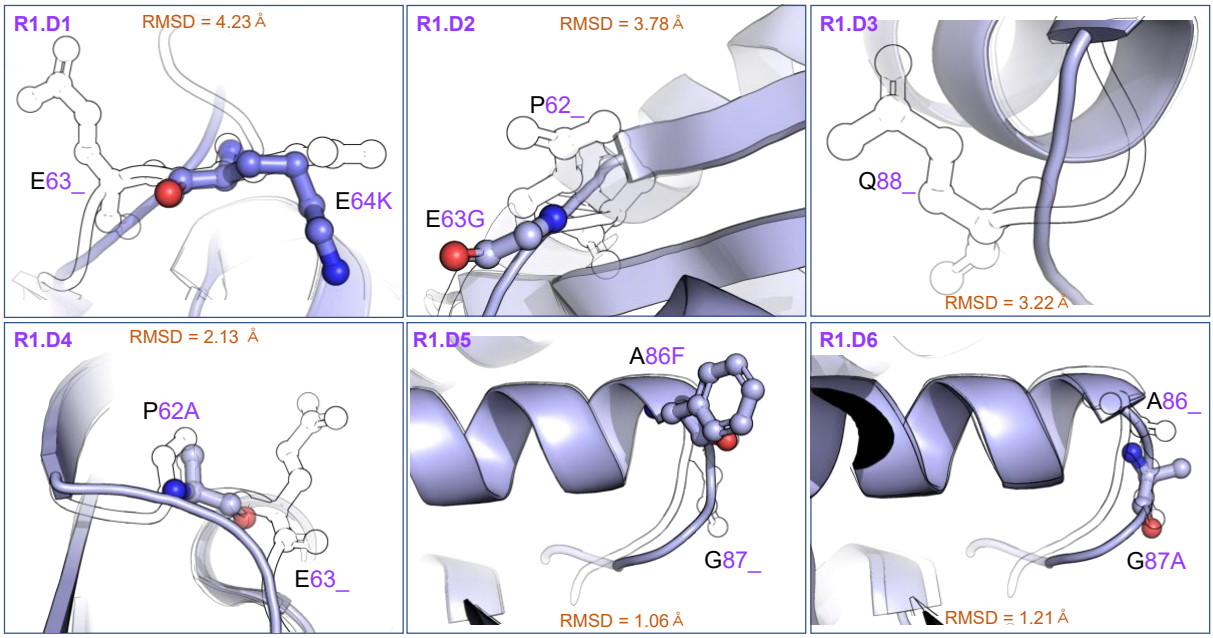
Supplementary Figure S6. Related to STAR Methods 'DNA Synthesis and TesA Mutant Construction'
Plasmid map transition from 'TesA-wt to I210_, P213K, D268_ and I210_, P213K respectively, showing the promoter regions and the nucleotide sequence for 'TesA-wt and variant (indicated in blue) is illustrated. Additional information regarding nucleotide changes that effect the mutations have been indicated.



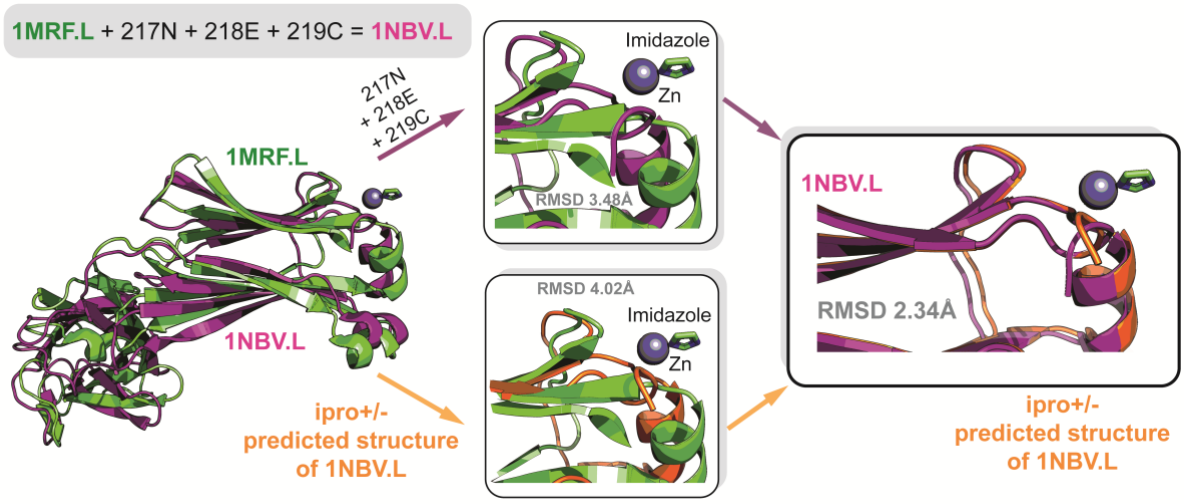
Supplementary Figure S7. Related to STAR Methods 'DNA Synthesis and TesA Mutant Construction' Plasmid map transition from 'TesA-wt to D268_, G270L and I210_, I215_ respectively, showing the promoter regions and the nucleotide sequence for 'TesA-wt and variant (indicated in blue) is illustrated. Additional information regarding nucleotide changes that effect the mutations have been indicated.



Supplementary Figure S8 – Related to Table 1.
Top six TEM-beta lactamase (*Ec*TBL) indel variant structures as predicted by IPRO+/- are indicated in blue while the WT *Ec*TBL structure is shown in black border. The RMSDs with respect to the starting structure has been indicated also.



Supplementary Figure S9 – Related to Figure 7
Overlaid experimentally resolved structures of 1MRF and 1NBV light chains (differ by 3 aa insertions) have an RMSD value of 3.48 Å. The corresponding RMSD value with the IPRO+/-predicted 1NBV.L is 4.02Å. The RMSD between the predicted and experimentally confirmed 1NBV.L structures is 2.34Å.



Supplementary Figure S10 – Related to Figure 7
Higher RMSD values between predicted antibody variable chain structures and their experimentally validated coordinates arise primarily from indels in large loop domains that are flexible and difficult to resolve.